INTRODUCTION

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

- This course will cover the design, analysis and interpretation of longitudinal studies.

- The course will emphasize model development, use of statistical software, and interpretation of results.

- The theoretical basis for results will be mentioned but not developed.

- No calculus or matrix algebra is assumed.

FEATURES OF LONGITUDINAL DATA

- Defining feature: repeated observations on individuals, allowing the direct study of change.
- Note that the measurements are commensurate, i.e. the same variable is measured repeatedly.
- Longitudinal data require sophisticated statistical techniques because the repeated observations are usually (positively) correlated.
- Sequential nature of the measures implies that certain types of correlation structures are likely to arise.
- Correlation must be accounted for to obtain valid inferences.

EXAMPLE

Auranofin Therapy for Rheumatoid Arthritis (Bombardier et al, *Am. J. Med*, 1986).

Randomized, placebo-controlled study of auranofin treatment of rheumatoid arthritis.

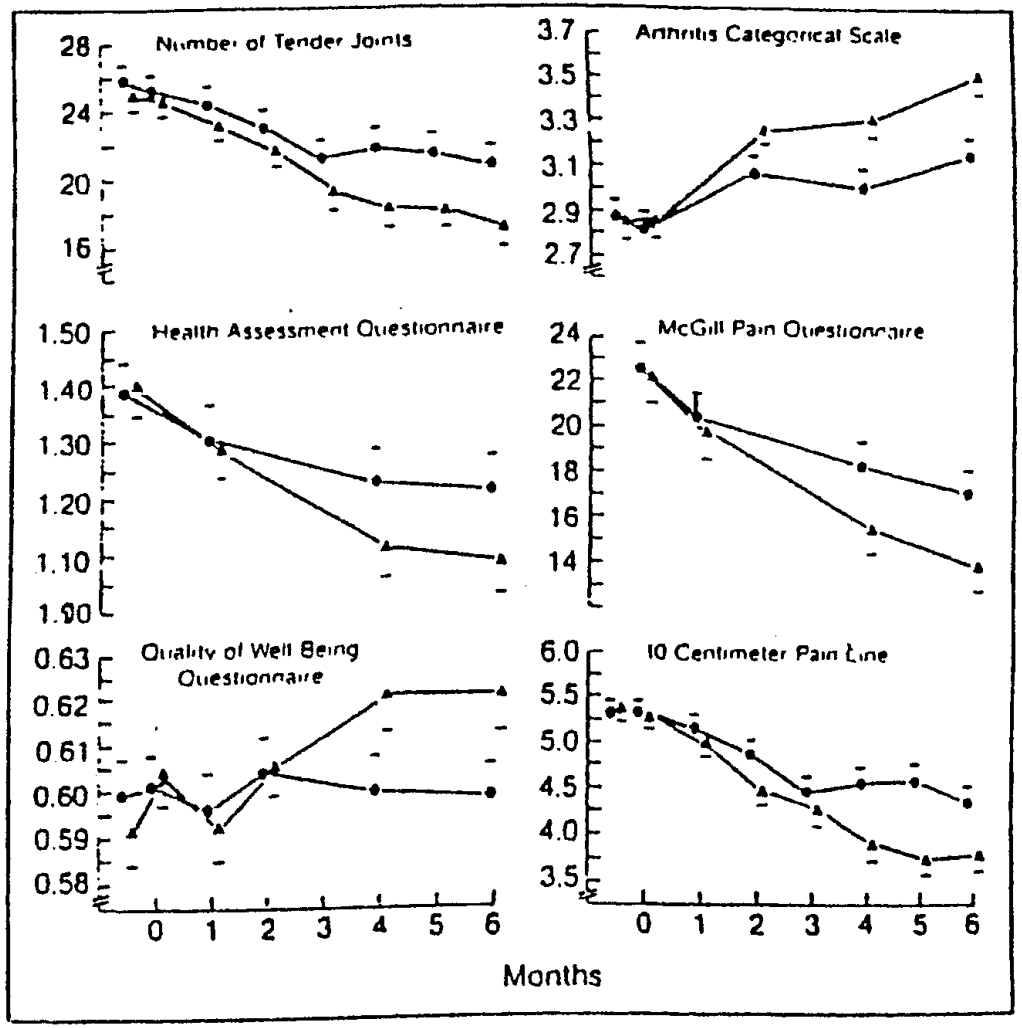
Outcome variables: More than 20 measures of pain, function, global health, utility.

Measurement schedule: Two baseline measurements and monthly measurements for six months.

Sample size:

303 patients with classic/definite rheumatoid arthritis

154 patients on auranofin
149 on placebo


Number of Tender Joints
28
24
20
16
Arthritis Categorical Scale
3.7
3.5
3.3
3.1
2.9
2.7
Health Assessment Questionnaire
1.50
1.40
1.30
1.20
1.10
1.00
McGill Pain Questionnaire
24
22
20
18
16
14
Quality of Well Being Questionnaire
0.63
0.62
0.61
0.60
0.59
0.58
10 Centimeter Pain Line
6.0
5.5
5.0
4.5
4.0
3.5
0 1 2 3 4 5 6
0 1 2 3 4 5 6
Months

EXAMPLE

Erythropoietin treatment of pruritus in hemodialysis patients (De Marchi et al, *NEJM*, 1992).

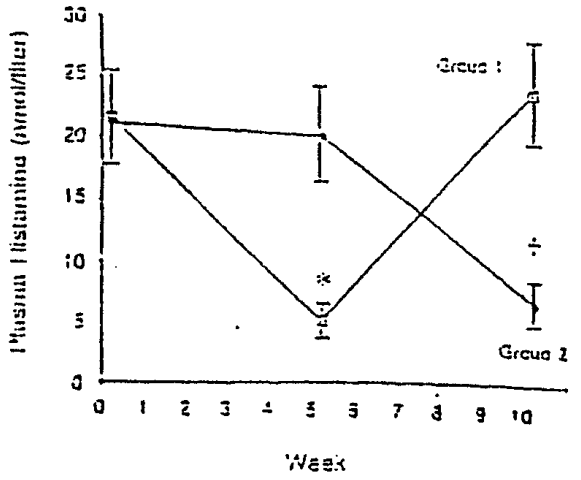
Randomized, placebo-controlled crossover study.

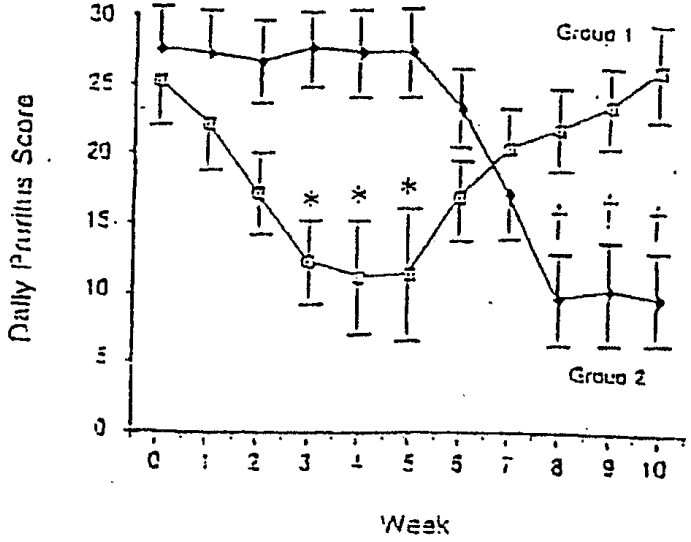
Outcome variables: Severity of pruritus, plasma histamine level.

Treatment and measurement schedule: 5 weeks of placebo and 5 weeks of erythropoietin in random order. Weekly pruritus score.

Sample size:

10 patients with pruritus


Plasma Histamine (nmol/liter)
30
25
20
15
10
5
0
Group 1
Group 2
0 1 2 3 4 5 6 7 8 9 10
Week


Daily Pruritus Score
30
25
20
15
10
5
0
Group 1
Group 2
0 1 2 3 4 5 6 7 8 9 10
Week

EXAMPLE

Six Cities Study: Respiratory Illness in Children
(Laird, Beck and Ware, 1984)

A non-randomized longitudinal study of the health effects of air pollution.

Subset of data from one of the participating cities: Steubenville, Ohio

Outcome variable: Binary indicator of respiratory illness in child

Measurement schedule: Four annual measurements at ages 7, 8, 9, and 10.

Interested in the influence of maternal smoking on children's respiratory illness.

Sample size:

537 children

Example: Presence and absence of respiratory infection

Mother Did Not Smoke					Mother Smoked				
Age of child				frequency	Age of child				frequency
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

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 chemotherapy.

Outcome variable: Count of number of seizures

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

Sample size: 59 epileptics

28 patients on placebo
31 patients on progabide

[Reproduced from Diggle, Liang & Zeger (1994)]

Table 1.5. Four successive two-week seizure counts for each of 59 epileptics. Covariates are adjuvant treatment (0=placebo, 1=progabide), eight-week baseline seizure counts, and age (in years)

Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age	Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age
5	3	3	3	0	11	31	0	4	3	0	1	19	20
3	5	3	3	0	11	30	3	6	1	3	1	10	20
2	4	0	5	0	6	25	2	6	7	4	1	19	18
4	4	1	4	0	8	36	4	3	1	3	1	24	24
7	18	9	21	0	66	22	22	17	19	16	1	31	30
5	2	8	7	0	27	29	5	4	7	4	1	14	35
6	4	0	2	0	12	31	2	4	0	4	1	11	57
40	20	23	12	0	52	42	3	7	7	7	1	67	20
5	6	6	5	0	23	37	4	18	2	5	1	41	22
14	13	6	0	0	10	28	2	1	1	0	1	7	28
26	12	6	22	0	52	36	0	2	4	0	1	22	23
12	6	8	5	0	33	24	5	4	0	3	1	13	40
4	4	6	2	0	18	23	11	14	25	15	1	46	43
7	9	12	14	0	42	36	10	5	3	8	1	36	21
16	24	10	9	0	87	26	19	7	6	7	1	38	35
11	0	0	5	0	50	26	1	1	2	4	1	7	25
0	0	3	3	0	18	28	6	10	8	8	1	36	26
37	29	28	29	0	111	31	2	1	0	0	1	11	25
3	5	2	5	0	18	32	102	65	72	63	1	151	22
3	0	6	7	0	20	21	4	3	2	4	1	22	32
3	4	3	4	0	12	29	8	6	5	7	1	42	25
3	4	3	4	0	9	21	1	3	1	5	1	32	35
2	3	3	5	0	17	32	18	11	28	13	1	56	21
8	12	2	8	0	28	25	6	3	4	0	1	24	41
18	24	76	25	0	55	30	3	5	4	3	1	16	32
2	1	2	1	0	9	40	1	23	19	8	1	22	26
3	1	4	2	0	10	19	2	3	0	1	1	25	21
13	15	13	12	0	47	22	0	0	0	0	1	13	36
11	14	9	8	1	76	18	1	4	3	2	1	12	37
8	7	9	4	1	38	32							

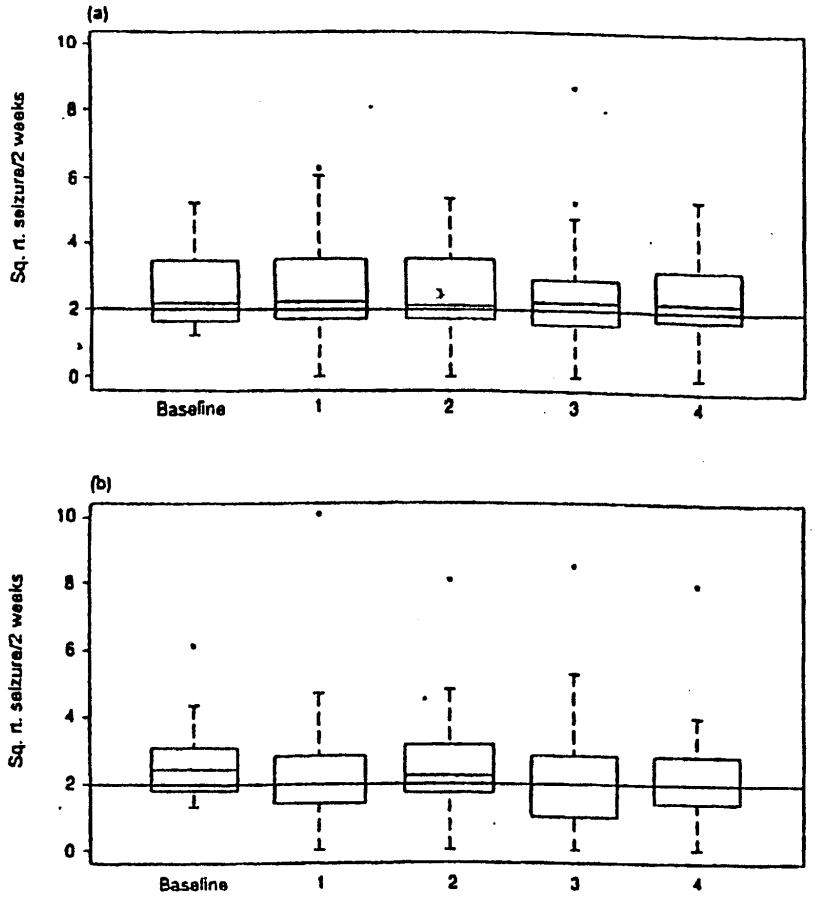


Fig. 1.5. Boxplots of square-root-transformed seizure rates for epileptics at baseline and for four subsequent two-week periods: (a) placebo; (b) progabide-treated

Some features of the studies:

Repeated measurements of study participants.

Two general types of design:

Parallel Design:

Groups of subjects defined by treatment or exposure category are followed over time. The main objective is to compare the trajectories of outcome variables between groups.

Crossover Design:

Subjects are exposed to multiple treatments or exposures. The objective is to compare the responses of the same subjects to different conditions.

GENERAL DATA STRUCTURE

$y_{ij} = j^{th}$ observation on the i^{th} subject

Observations

	Time				
	1	2	3	...	p
Subjects					
1	y_{11}	y_{12}	y_{13}	...	y_{1p}
2	y_{21}	y_{22}	y_{23}	...	y_{2p}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
n	y_{n1}	y_{n2}	y_{n3}	...	y_{np}

TWO SPECIAL CASES

Two-Groups Parallel Design:

		Time				
		1	2	3	...	p
	Subjects					
Tx 1						
	1	y_{11}	y_{12}	y_{13}	...	y_{1p}
	2	y_{21}	y_{22}	y_{23}	...	y_{2p}
	.	.	.	.	...	.
	.	.	.	.	...	.
	.	.	.	.	...	.
	m	y_{m1}	y_{m2}	y_{m3}	...	y_{mp}
Tx 2						
	$m+1$	$y_{m+1,1}$	$y_{m+1,2}$	$y_{m+1,3}$	...	$y_{m+1,p}$
	$m+2$	$y_{m+2,1}$	$y_{m+2,2}$	$y_{m+2,3}$	...	$y_{m+2,p}$
	.	.	.	.	...	.
	.	.	.	.	...	.
	.	.	.	.	...	.
	n	y_{n1}	y_{n2}	y_{n3}	...	y_{np}

Crossover Design:

	Treatment		
	Ctrl	T_1	T_2
Subjects			
1	y_{11}	y_{12}	y_{13}
2	y_{21}	y_{22}	y_{23}
.	.	.	.
.	.	.	.
.	.	.	.
n	y_{n1}	y_{n2}	y_{n3}

In longitudinal studies the outcome variable can be:

- continuous
- binary
- count

The data set can be incomplete.

Subjects may be measured at different occasions.

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) *Repeated Measures ANOVA*

Review of One-way ANOVA

Repeated Measures ANOVA

- Outcome: Continuous
- Balanced and complete data
- Software: PROC GLM/MIXED in SAS

2) *General Linear Model 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) *Nonlinear Models for Longitudinal Data*

Generalizations and extensions to allow fitting of nonlinear models to 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 the RANDOM STATEMENT

BACKGROUND ASSUMED

1) Samples and populations

2) Sample and population values

   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
   Multiple logistic regression
   ANOVA

ANALYSIS OF VARIANCE

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

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

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

$H_A : \mu_i$'s not all equal

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

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

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

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

Data Structure:

Pop1	Pop2	...	Pop$k$	
y_{11}	y_{21}	$\cdots$	y_{k1}	
y_{12}	y_{22}	$\cdots$	y_{k2}	Samples
$\vdots$	$\vdots$		$\vdots$	
y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}	

$$
\begin{aligned}
y_{ij} &= \text{value of j}^{th} \text{ observation in } i^{th} \text{ sample} \\
i &= 1, \ldots, k \text{ (number of samples)} \\
j &= 1, \ldots, n_i \text{ (\# obs. in } i^{th} \text{ sample)}
\end{aligned}
$$

NB: Number of observations in each sample is not necessarily the same (i.e., the n_i can differ)

Terminology

Factor: Any nominal (or categorical) variable, e.g., treatment $(trt_1, \ldots, trt_k)$

Level: The categories of a factor are called levels, e.g., trt_2 is one particular level of the "treatment" factor

Effects: Differences in the mean of the response variable among different levels of a factor, e.g., the "effect" of "treatment" are the extent to which the mean response differs

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:

1. Does the mean of Y differ among levels of a factor?

2. 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.

Assumptions of One-Way ANOVA

1. Independence of observations:

   Y_{ij} are independent random variables for all i, j.

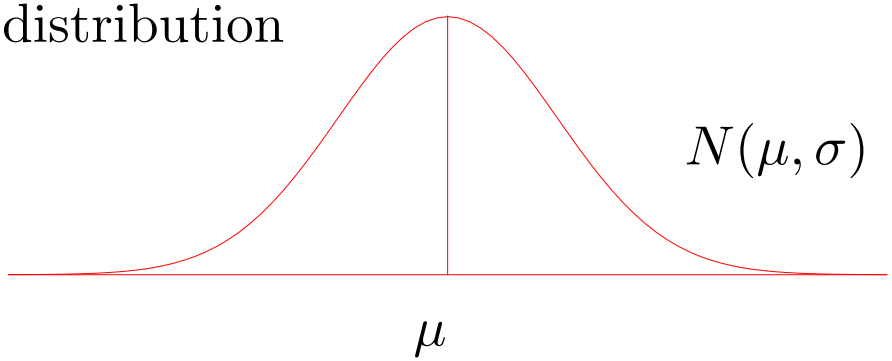
   $\Longrightarrow$ independent random samples from k populations

2. Y_{ij} have Normal dist^n with mean μ_i
3. Common variance for all populations, $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$

Recall: the Normal probability density of Y_{ij} is:

$$f\left(Y_{ij}\right) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\left(Y_{ij} - \mu_i\right)^2 / 2\sigma^2\right].$$

The normal distribution



Model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

grand mean over all groups	+	effect of being in i^{th} group	+	unexplained random error of j^{th} obs'n in i^{th} group

Under our assumptions

$$Y_{ij} \overset{d}{\sim} N(\mu_i, \sigma^2)$$

- $\mu_i = \mu + \alpha_i$
- $\varepsilon_{ij} \overset{d}{\sim} N(0, \sigma^2)$

Constraint:

- Note that we have the following parameters:

  $\mu, \alpha_1, \alpha_2, \ldots, \alpha_k$

- However, there are only k population means to estimate

- We have to "constrain" α's in some way

One common constraint is:

$$\left.\begin{array}{rcl} \alpha_k & = & 0 \end{array}\right\} \begin{array}{l} \mu \text{ is the mean of pop } k \\ \text{and } \alpha_i \ (i \neq k) \text{ measures} \\ \text{difference from pop } k \end{array}$$

Hypothesis Testing

Test: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
H_A : not all μ_i equal

Basic Idea:

- quantify variability between sample means

  $\longrightarrow$ Between groups variability

- quantify error variability or variability of observations in the same group

  $\longrightarrow$ Error or within groups variability

Between >> Within (Error) $\Longrightarrow \mu_i$'s vary
$\Longrightarrow$ reject H_0

Otherwise cannot reject H_0.

Summary Statistics:

	Group			
	1	2	...	k
	y_{11}	y_{21}	$\cdots$	y_{k1}
	$\vdots$	$\vdots$	...	$\vdots$
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}
Means	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$		$\bar{y}_{k\bullet}$

Notation:

$$N = \sum_{i=1}^{k} n_i = \text{ total number of observations}$$

$$\overline{y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = i^{th} \text{ sample mean}$$

(sometimes written as $\overline{y}_i$)

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \text{ sample mean of all observations}$$

Total Variability in Y's measured by sum of squares

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{\bullet\bullet})^2$$

Using $(y_{ij} - \overline{y}_{\bullet\bullet}) = (y_{ij} - \overline{y}_{i\bullet}) + (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})$,We obtain

$$\begin{aligned} SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \\ &= \sum_{i=1}^{k} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \\ &= SSB + SSW \end{aligned}$$

Total Variation	=	Between Groups Variation	+	Within Groups Variation

ANOVA Table

The ANOVA table provides a summary of these sources of variation

Source	SS	df	MS	F
Between Groups	SSB	$k-1$	MSB	$\frac{\text{MSB}}{\text{MSE}}$
Within Groups	SSW (=SSE)	$N-k$	MSE	
Total	SST			

$$\text{MSB} = \frac{\text{SSB}}{k-1}$$

$$\text{MSE} = \frac{\text{SSE}}{\text{N-}k} = \frac{\text{SSW}}{\text{N-}k}$$

Sources of Variability

Between groups variability

$$MSB = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}\right)^2$$

Analogous to the sample variance of the group means.

Within groups variability

$$\begin{aligned} MSW = MSE &= \tfrac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{i\bullet}\right)^2 \\ &= \tfrac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2 \end{aligned}$$

where S_i^2 is the sample variance of the i^{th} group.

MSW represents:

- A weighted average of sample variances
- Average variability within the groups

Test Statistic: $F = MSB/MSE$

Interpretation of F Test

The F test is the ratio of two variances.

The denominator

$$\text{MSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\bullet}\right)^2 / n - k$$

represents the within-group mean square and has expected value

$$E\left(\text{MSE}\right) = \sigma^2$$

(Note: $E(\cdot)$ can be thought of as denoting the average over all data sets that might have arisen).

The numerator

$$\text{MSB} = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2 / \left(k-1\right)$$

is the between-group sum of squares. It has expected value

$$E\left(\text{MSB}\right) = \sigma^2 + \sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu}\right)^2 / \left(k-1\right)$$

where $\bar{\mu} = \frac{\sum_{i=1}^{k} n_i \mu_i}{\sum_{i=1}^{k} n_i}$

When $H_0 : \mu_1 = \ldots = \mu_k$ is true,

$$E\left(\text{MSB}\right) = \sigma^2$$

When H_0 is not true,

$$E\,(\text{MSB}) > \sigma^2$$

When H_0 is true, F will take values close to 1. When H_0 is not true, F will tend to be large.

Null distribution of $F \sim F_{k-1,N-k}$
$\Longrightarrow$ reject $H_0 : \mu_1 = \ldots = \mu_k$ if $F > F_{k-1,N-k,(1-\alpha)}$

- This is a "global" test.
- It does not specify which μ_i's differ.

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

  ie. differences in mean dosage required to kill animal

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

ANOVA Table

Source	df	SS	MS	F
between (drug)	3	249.9	83.3	8.5
within	36	350.9	9.7	
Total	39	600.8		

$F = 83.3/9.7 = 8.5 \quad (p < 0.001)$

Estimated (common) standard deviation of the distribution of scores in each group

$$\begin{aligned} &= \sqrt{9.7} \\ &= 3.11 \end{aligned}$$

Estimation

A single mean:

$$\hat{\mu}_i = \bar{y}_i$$

$$Var\left(\bar{y}_i\right) = \sigma^2/n_i \text{ and } \hat{\sigma}^2 = \text{MSE}$$

Thus, a 95% C.I. for μ_i is

$$\bar{y}_i \pm t_{N-k,\,0.975}\sqrt{MSE/n_i}$$

Linear combinations of means:

Interest could focus on any of, e.g.,

μ_1

$\mu_1 - \mu_2$

$\mu_1 - \left(\frac{\mu_2+\mu_3+\mu_4}{3}\right)$

A unified estimation theory can be developed using the linear form

$$\sum_{i=1}^{k} w_i \mu_i$$

We estimate all such linear combinations using

$$\sum_{i=1}^{k} w_i \hat{\mu}_i = \sum_{i=1}^{k} w_i \bar{y}_i$$

$$Var\left(\sum_{i=1}^{k} w_i \bar{y}_i\right) = \left(\sum_{i=1}^{k} w_i^2/n_i\right)\sigma^2$$

Same estimation theory applies for any weighted sum, e.g. for guinea pig data:

$$\widehat{\mu}_1 - \widehat{\mu}_2 = \bar{y}_1 - \bar{y}_2 = 25.9 - 22.2 = 3.7$$

with variance $\left(\frac{1}{10} + \frac{1}{10}\right)\sigma^2 = .2 * 9.7 = 1.9$

A linear combination of means whose weights sum to 0 is called a *contrast*:

$$\sum_{i=1}^{k} c_i \mu_i, \text{ with } \sum_{i=1}^{k} c_i = 0$$

Thus, the second and third examples above are contrasts, whereas the first is not.

Multiple Comparison Procedures

When the initial F test in an ANOVA is significant, we would like to identify the specific pairs of means that differ (and contrasts that differ from 0).

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

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

Source	**df**	**SS**	**MS**	**F**
Drug	3	249.9	83.3	8.5
Error	36	350.9	9.7	
($p < 0.001$)				

Which pairs of means differ?

Method 1: Form a 95% C.I. for each contrast between groups. Declare a difference if the C.I. does not include 0.

Problem: Many tests $\Rightarrow$ high rate of false positive results. Increased type 1 error rate.

Method 2: Bonferroni Adjustment

Determine Q, the number of contrasts to be tested. Test each at level $0.05/Q$.

For example, with four groups there are $Q = 6$ possible pairwise comparisons. A good method if we want to focus on a few contrasts. Very conservative if Q is large.

Method 3: Scheffe's Method

To obtain simultaneous confidence intervals for all possible contrasts,

$$\sum_{i=1}^{k} c_i \mu_i \text{ with } \sum_{i=1}^{k} c_i = 0$$

such that we are 95% confident that *all* of the C.I.'s cover the population value:

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm \delta \sqrt{\text{MSE} \left(\sum_{i=1}^{k} \frac{c_i^2}{n_i} \right)}$$

where $\delta^2 = (k-1) F_{k-1,\, N-k,\, 1-\alpha}$ is a larger critical value.

For pairwise comparisons, this reduces to

$$\bar{y}_i - \bar{y}_j \pm \delta \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Example: Data on guinea pigs

$$\delta = \sqrt{3 \times 2.886} = 2.94$$

$$\delta \sqrt{\text{MSE}\left(\tfrac{1}{n_i} + \tfrac{1}{n_j}\right)} = 2.94 \sqrt{9.75\left(\tfrac{1}{10} + \tfrac{1}{10}\right)}$$

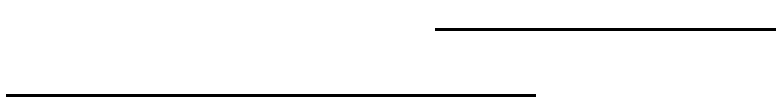
$$= 4.11$$

Hence the 95% C.I. for $\mu_1 - \mu_4$ is

$$25.9 - 19.6 \pm 4.1 = (2.2,\ 10.4)$$

Means that differ by more that 4.1 are significantly different.

A graphical representation:



19.6	20.0	22.2	25.9
$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\mu}_3$	$\widehat{\mu}_4$

$$\mu_1 \neq \mu_4 \text{ and } \mu_1 \neq \mu_3$$

For Scheffe's Method, the ANOVA F test is significant if and only if at least one contrast is significantly different from 0.

SAS Syntax for One-Way ANOVA

```
data toxic;
      infile 'g:\shared\bio226\tox.dat';
      input drug y;
run;
proc glm data=toxic;
      class drug;
      model y=drug;
      means drug / scheffe;
      title 'Scheffe Pairwise Comparisons';
run;
proc glm data=toxic;
      class drug;
      model y=drug;
      contrast 'Group 1 versus 2' drug 1 -1 0 0;
      title 'Testing contrast of means';
run;
```

TWO WAY ANOVAs

Recall: One-Way ANOVA

- one factor with k different levels
- compare mean response between factor levels

In two- (or more) way ANOVA:

- ≥ 2 factors observed
- compare mean response across levels of factor 1, and factor 2, ...

Questions of Interest

1. Does mean outcome differ between the levels of factor 1?

2. Does mean outcome differ between the levels of factor 2?

3. Is there an "interaction" between factors 1 and 2, i.e., do differences between the levels of factor 2 depend on the levels of factor 1? (or vice versa)

Relationship between ANOVA and Multiple 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 k levels:

Define $X_1 = 1$ if subject belongs to level 1, and 0 otherwise;
and define $X_2, ..., X_{k-1}$ similarly.

Note: Here the last level of the factor is selected as a "reference".

For example, for a subject at level 2:

X_1	X_2	X_3	$\cdots$	X_{k-1}
0	1	0	$\cdots$	0

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

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

as

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots \beta_{k-1} X_{k-1,ij} + \epsilon_{ij}$$

Note:

$$\begin{aligned} \mu_1 &= \beta_0 + \beta_1 \\ \mu_2 &= \beta_0 + \beta_2 \\ &\vdots \\ \mu_k &= \beta_0 \end{aligned}$$

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.

ANALYSIS OF REPEATED MEASURES

Longitudinal Studies: Designs in which the outcome variable is measured repeatedly over time (for at least some study participants).

Repeated Measures: Older term applied to a special set of longitudinal designs characterized by measurement at a common set of occasions, usually in an experimental setting.

Initially, we will consider methods for analyzing longitudinal data collected in two of the most important designs: single-group and parallel group repeated-measures designs.

We will focus on linear models for continuous responses. Later in the course we will discuss methods for categorical responses and count data.

POTENTIAL SCIENTIFIC ADVANTAGES OF LONGITUDINAL DESIGNS

1. They allow investigation of events that occur in time. Essential to the study of normal growth and aging, and the effects of individual characteristics, treatments, or environmental exposures on those changes. Also essential to the study of the temporal pattern of response to treatment.

2. Can study the order of events.

3. Permit more complete ascertainment of exposure histories in epidemiologic studies.

4. Reduce unexplained variability in the response by using the subject as his or her own control.

THE SINGLE-GROUP REPEATED MEASURES DESIGN

Each subject receives each of p treatments at p different occasions.

We assume initially that each subject receives the treatments in the same order. Later, we will relax that assumption.

Listing each observation under the appropriate treatment columns:

	TREATMENT					
	T_1	T_2	.	.	.	T_p
Subject						
1	Y_{11}	Y_{12}	.	.	.	Y_{1p}
2	Y_{21}	Y_{22}	.	.	.	Y_{2p}
.						
.						
.						
n	Y_{n1}	Y_{n2}	.	.	.	Y_{np}

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

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,\,p-1} + e_i \\ &= \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + e_i \end{aligned}$$

where

$$X_{ij} = \begin{cases} 1, & \text{if observation } i \text{ was obtained} \\ & \text{while receiving treatment } j; \\ 0, & \text{otherwise.} \end{cases}$$

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

EXAMPLE: PLACEBO-CONTROLLED STUDY OF TWO ANTIHYPERTENSIVE TREATMENTS

Subject	Baseline (Trt 0)	Placebo (Trt 1)	Trt A (Trt 2)	Trt B (Trt 3)
01	113	108	98	103
02	108	110	96	104
03	110	106	110	107
04	99	98	78	88
05	114	114	112	105
06	109	103	88	97
07	113	110	106	103
08	112	108	99	108
09	121	111	96	115
10	98	103	98	107
11	107	104	98	100
12	123	117	97	103
13	112	109	108	114
14	96	94	99	97
15	108	106	101	101
16	111	99	102	87
17	124	130	121	124
18	113	119	101	114
19	106	99	89	90
20	99	99	94	85

Denote the population means at the p occasions by $\mu_1,\ \mu_2,\ \ldots,\ \mu_p$.

Then the null hypothesis of interest is

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

How can we test this hypothesis?

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

This approach allows only pairwise comparisons.

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

DEPENDENCE AND CORRELATION

INDEPENDENCE

Two variables, X and Y, are said to be independent if the conditional distribution of Y given X does not depend on X.

Example: Blood pressure would be independent of age if the distribution of blood pressures were the same for every age group.

CORRELATION

Two variables, Y and Z, are uncorrelated if

$$E\left[\left(Y-\mu_Y\right)\left(Z-\mu_Z\right)\right]=0$$

Note: Independent variables are uncorrelated, but variables can be uncorrelated without being independent. Independence is a stronger condition.

CORRELATED OBSERVATIONS

Two variables, Y and Z, are correlated if

$$E\left[\left(Y-\mu_Y\right)\left(Z-\mu_Z\right)\right] \neq 0$$

The quantity, $E\left[\left(Y-\mu_Y\right)\left(Z-\mu_Z\right)\right]$, is called the covariance. Notice that the covariance of a variable with itself is the variance.

Covariance can take any positive or negative value and its value depends on the units of the variables. To make it independent of units, we divide by the standard deviations of the two variables:

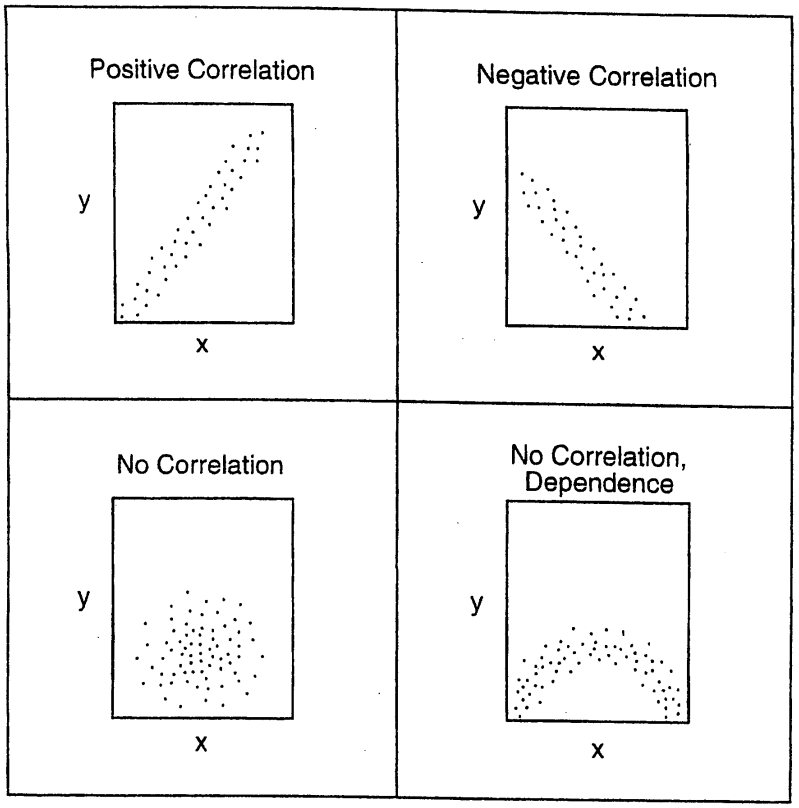
$$\text{Corr}\left(Y,Z\right) = E\left[\left(Y-\mu_Y\right)\left(Z-\mu_Z\right)\right]/\sigma_Y\sigma_Z$$

Correlations must lie between -1 and 1.

Repeated measures obtained from the same person are usually positively correlated.

FROM SAS, INC. (2000)
PROC CORR DOCUMENTATION.

Figure 12.1 Examining Correlations Using Scatterplots



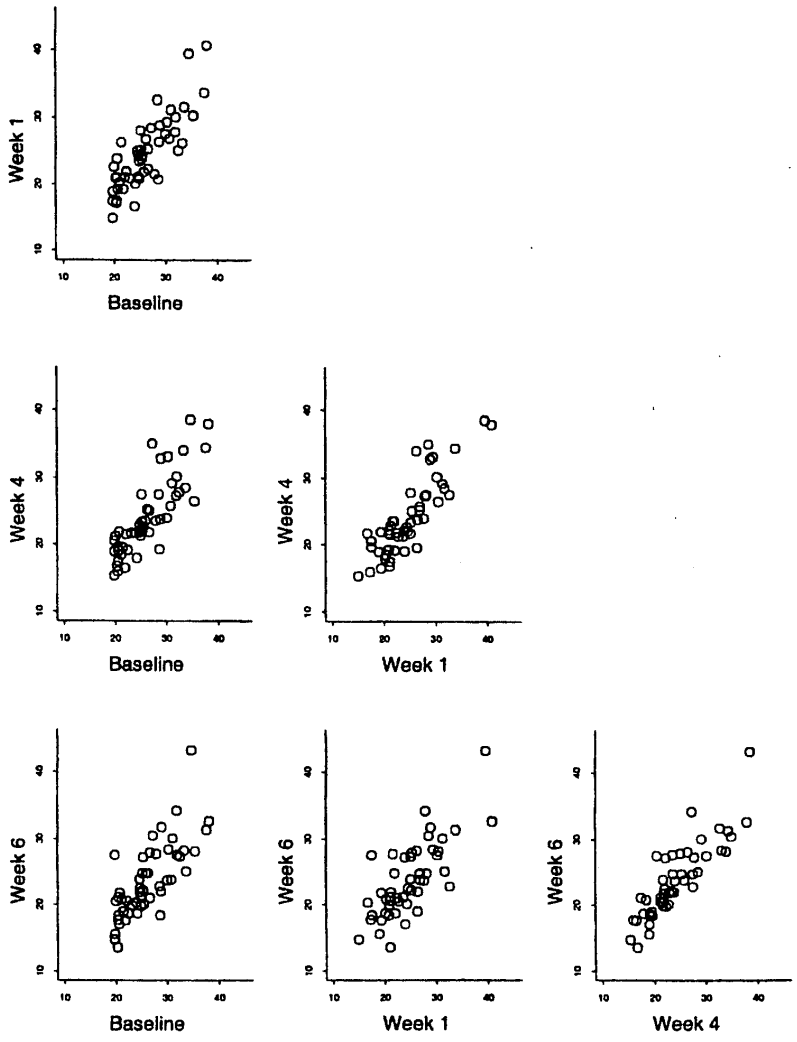


Fig. 2.1 Pairwise scatter-plots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Given vectors of observations $(Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ we define the covariance matrix as the following array of variances and covariances:

$$\text{Cov}\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ Y_{ip} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

where $\text{Cov}\,(Y_{ir}, Y_{is}) = \sigma_{rs}$

We can use SAS to calculate the sample covariance matrix for the diastolic blood pressure measurements as follows.

SAS CODE

```
data dbp;
      infile 'g:\shared\bio226\dbp.dat';
      input id y0 y1 y2 y3;
run;

proc sort data = dbp;
      by id;
run;

title "Listing of the Raw Data";
proc print data = dbp;
run;

title "Means, Covariances, and Correlations";
proc corr data = dbp cov;
      var y0 - - y3;
run;
```

Listing of the Raw Data

OBS	ID	Y0	Y1	Y2	Y3
1	1	113	108	98	103
2	2	108	110	96	104
3	3	110	106	110	107
4	4	99	98	78	88
5	5	114	114	112	105
6	6	109	103	88	97
7	7	113	110	106	103
8	8	112	108	99	108
9	9	121	111	96	115
10	10	98	103	98	107
11	11	107	104	98	100
12	12	123	117	97	103
13	13	112	109	108	114
14	14	96	94	99	97
15	15	108	106	101	101
16	16	111	99	102	87
17	17	124	130	121	124
18	18	113	119	101	114
19	19	106	99	89	90
20	20	99	99	94	85

Means, Covariances, and Correlations

Correlation Analysis

4 Variables: Y0 Y1 Y2 Y3

Covariance Matrix DF = 19

	Y0	Y1	Y2	Y3
Y0	60.69	54.18	36.96	49.97
Y1	54.18	70.77	49.32	69.15
Y2	36.96	49.32	85.63	61.39
Y3	49.97	69.15	61.39	102.36

Simple Statistics

Variable	N	Mean	Std Dev
Y0	20	109.8000	7.7907
Y1	20	107.3500	8.4122
Y2	20	99.5500	9.2536
Y3	20	102.6000	10.1172

Pearson Correlation Coefficients, N = 20
Prob > $|r|$ under H_0: Rho = 0

	Y0	Y1	Y2	Y3
Y0	1.0000	0.8266	0.5126	0.6339
	0.0	0.0001	0.0208	0.0027
Y1	0.8266	1.0000	0.6336	0.8124
	0.0001	0.0	0.0027	0.0001
Y2	0.5126	0.6336	1.0000	0.6557
	0.0208	0.0027	0.0	0.0017
Y3	0.6339	0.8124	0.6557	1.0000
	0.0027	0.0001	0.0017	0.0

Thus, the assumption of independence is inappropriate.

One approach to analyzing repeated measures 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, 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 placebo-controlled study of two antihypertensive treatments; later we will consider the statistical basis for the analysis.

Note: PROC MIXED requires the data to be in a univariate form. Often it will be necessary to transform the data from a "multivariate mode" to a "univariate mode".

SAS CODE

```
data dbp;
      infile 'g:\shared\bio226\dbp.dat';
      input id y0 y1 y2 y3;
            y=y0; trt=0; output;
            y=y1; trt=1; output;
            y=y2; trt=2; output;
            y=y3; trt=3; output;
      drop y0-y3;
run;

proc mixed data=dbp covtest;
      class id trt;
      model y = trt /s chisq;
      repeated /type=un subject=id r;
      contrast 'T2 - T1'
            trt 0 -1 1 0 / chisq;
run;
```

Univariate Form of DBP Data (1st 5 subjects)

OBS	ID	Y	trt
1	1	113	0
2	1	108	1
3	1	98	2
4	1	103	3
5	2	108	0
6	2	110	1
7	2	96	2
8	2	104	3
9	3	110	0
10	3	106	1
11	3	110	2
12	3	107	3
13	4	99	0
14	4	98	1
15	4	78	2
16	4	88	3
17	5	114	0
18	5	114	1
19	5	112	2
20	5	105	3

SELECTED OUTPUT FROM PROC MIXED

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > |Z|
UN(1,1)	60.6947	19.6920	3.08	0.0010
UN(2,1)	54.1789	19.5077	2.78	0.0055
UN(2,2)	70.7658	22.9595	3.08	0.0010
UN(3,1)	36.9579	18.5857	1.99	0.0468
UN(3,2)	49.3237	21.1417	2.33	0.0196
UN(3,3)	85.6289	27.7817	3.08	0.0010
UN(4,1)	49.9684	21.4101	2.33	0.0196
UN(4,2)	69.1474	25.1572	2.75	0.0060
UN(4,3)	61.3895	25.6838	2.39	0.0168
UN(4,4)	102.36	33.2093	3.08	0.0010

Estimated R Matrix for id 1

Row	COL1	COL2	COL3	COL4
1	60.6947	54.1789	36.9579	49.9684
2	54.1789	70.7658	49.3237	69.1474
3	36.9579	49.3237	85.6289	61.3895
4	49.9684	69.1474	61.3895	102.36

Fit Statistics

-2 Res Log Likelihood	504.8
AIC (smaller is better)	524.8
AICC (smaller is better)	528.2
BIC (smaller is better)	534.7

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > Chi Sq
9	55.79	<.0001

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		102.60	2.2623	19	45.35	0.0001
trt	0	7.2000	1.7765	19	4.05	0.0007
trt	1	4.7500	1.3196	19	3.60	0.0019
trt	2	-3.050	1.8057	19	-1.69	0.1075
trt	3	0.000	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
trt	3	19	32.87	< .0001

Contrasts

Label	Num DF	Den DF	Chi-Square	Pr > ChiSq
T2 - T1	1	19	21.07	< .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 correlation among repeated measures?

We can address that question by re-analyzing the diastolic blood pressure data under the assumption of independence and comparing the results to those provided by PROC MIXED.

```
data dbp;
      infile 'g:\shared\bio226\dbp.dat';
      input id y0 y1 y2 y3;
      y=y0; trt=0; t=0; output;
      y=y1; trt=1; t=1; output;
      y=y2; trt=2; t=2; output;
      y=y3; trt=3; t=3; output;
      drop y0-y3;
run;

proc glm data=dbp;
      class trt;
      model y=trt/solution;
      estimate 'T2 - T1'
            trt 0 -1 1 0;
run;

proc mixed data=dbp noclprint;
      class id trt t;
      model y=trt/s chisq;
      repeated t/type=un subject=id;
      estimate 'T2 - T1'
            trt 0 -1 1 0;
run;
```

RESULTS USING PROC GLM

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1278.05	426.02	5.33	0.0022
Error	76	6069.50	79.86		
Total	79	7347.55			

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		102.60	2.00	51.34	0.0001
trt	0	7.20	2.83	2.55	0.0129
trt	1	4.75	2.83	1.68	0.0969
trt	2	-3.05	2.83	-1.08	0.2839
trt	3	0.00	.	.	.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
T2-T1	-7.80	2.83	-2.76	0.0072

RESULTS USING PROC MIXED

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		102.60	2.2623	19	45.35	< .0001
trt	0	7.2000	1.7765	19	4.05	0.0007
trt	1	4.7500	1.3196	19	3.60	0.0019
trt	2	-3.0500	1.8057	19	-1.69	0.1075
trt	3	0	.	.	.	.

Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
TRT	3	19	32.87	< .0001

Estimates

Label	Estimate	Standard Error	t Value	Pr > |t|
T2 - T1	-7.8000	1.6992	-4.59	0.0002

Note that the estimates of the treatment contrast are the same in both analyses, i.e., -7.8; but the standard errors are discernibly different.

The standard error yielded by PROC GLM, 2.83, is not valid since the procedure has incorrectly assumed that all of the observations are independent.

The standard error yielded by PROC MIXED, 1.70, is valid since the procedure has accounted for the correlation among repeated measures in the analysis.

INTRODUCING A COVARIATE FOR TIME OF MEASUREMENT

So far, we have assumed that every subject receives the treatments in the same order.

More often, the order of administration is varied. Also, some observations may be missing.

To allow for this possibility, the data set should always contain a variable representing the time period of measurement (and defined as a class variable when the number of occasions is fixed and shared).

Suppose that the data for the first subject is

Time Period	Treatment	Response
Baseline	Baseline	113
Time 1	Trt A	108
Time 2	Trt B	98
Time 3	Placebo	103

The analysis of these data requires the inclusion of a variable defining the time period in which the treatment occurred.

Subj	y	Time	Treatment
1	113	0	0
1	108	1	2
1	98	2	3
1	103	3	1

The expected value of the response is assumed to depend only on the treatment received (or, in general, on the covariates).

The variances and covariances of the residuals are assumed to depend on the time periods in which the observations were obtained.

If an observation is missing, as in

Time Period	Treatment	Response
Baseline	Baseline	113
Time 1	Trt A	.
Time 2	Trt B	98
Time 3	Placebo	103

this can be expressed in the data by including a row with a missing value indicator for y.

Subj	y	Time	Treatment
1	113	0	0
1	.	1	2
1	98	2	3
1	103	3	1

Note: If the data set includes a class variable designating the time of measurement, the data record with the missing response need not be present in the data set.

SAS CODE

```
proc mixed data=dbp noclprint;
      class id trt time;
      model y=trt/s chisq;
      repeated time/type=un subject=id;
      estimate 'T2 - T1'
            trt 0 -1 1 0;
run;
```

STATISTICAL BASIS FOR REPEATED MEASURES ANALYSIS

In this lecture we introduce the general linear model for repeated measurements and discuss inference based on maximum likelihood (ML).

Example:

To motivate the theory underlying the general linear model for repeated measures consider the following example. A study was designed to compare diastolic blood pressure levels of 20 subjects at baseline and after two weeks of treatment on each of three treatment regimens: Placebo, Treatment A, and Treatment B.

Initially, we want to test the null hypothesis that blood pressure levels are unrelated to the treatments.

EXAMPLE: PLACEBO-CONTROLLED STUDY OF TWO ANTIHYPERTENSIVE TREATMENTS

Subject	Baseline (Trt 0)	Placebo (Trt 1)	Trt A (Trt 2)	Trt B (Trt 3)
01	113	108	98	103
02	108	110	96	104
03	110	106	110	107
04	99	98	78	88
05	114	114	112	105
06	109	103	88	97
07	113	110	106	103
08	112	108	99	108
09	121	111	96	115
10	98	103	98	107
11	107	104	98	100
12	123	117	97	103
13	112	109	108	114
14	96	94	99	97
15	108	106	101	101
16	111	99	102	87
17	124	130	121	124
18	113	199	101	114
19	106	99	89	90
20	99	99	94	85

THE GENERAL LINEAR MODEL FOR THE SINGLE-GROUP REPEATED MEASURES DESIGN

Let $Y_{ij}, j = 1, \ldots, p$, be the sequence of observed measurements for the i^{th} subject, $i = 1, \ldots, n$.

We assume initially that each subject is observed at the same time points or under the same treatment conditions (balanced) and that no observations are missing (complete data).

Each observation, Y_{ij}, has an associated set of covariates $X_{ij0}, X_{ij1}, X_{ij2}, \ldots, X_{ijk-1}$, where typically $X_{ij0} = 1$.

The linear model for Y_{ij} can be written as

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_{k-1} X_{ijk-1} + e_{ij} \\ &= \mathbf{X}_{ij}\boldsymbol{\beta} + e_{ij} \end{aligned}$$

where

$\mathbf{X}'_{ij}$ denotes the $(k \times 1)$ vector of covariates,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1})'$ is a $(k \times 1)$ vector of regression parameters.

With repeated measures, we expect the e_{ij} to be correlated within individuals.

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

Assumptions:

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

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

$$\mu_{ij} = E(Y_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$$

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

The variance-covariance matrix of the vector of observations, $Y_{i1}, \ldots, Y_{ip}$, is denoted $\boldsymbol{\Sigma}$ and its elements are $\sigma_{jj'}$.

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 one-way ANOVA model is:

$$f(Y_i) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\left(Y_i - \mu_i\right)^2 / 2\sigma^2\right]$$

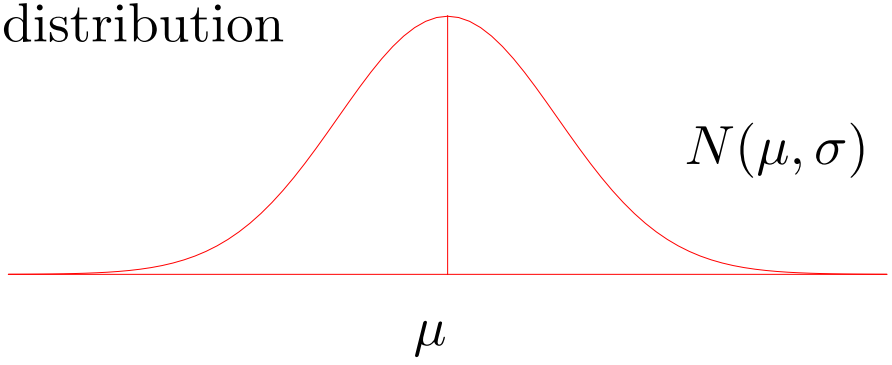
or

$$f(Y_i) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\left(Y_i - \mathbf{X}_i\boldsymbol{\beta}\right)^2 / 2\sigma^2\right].$$

Note that $f(Y_i)$ describes the probability or relative frequency of occurrence of particular values of Y_i.

Specifically, $f\left(Y_i\right)$ describes a bell-shaped curve.

The normal distribution



Recall that the area under the curve between any two values represents the probability of Y_i taking a value within that range.

Notable Features:

- $f(Y_i)$ is completely determined by (μ_i, σ^2)
- $f(Y_i)$ depends to a very large extent on

$$\frac{(Y_i - \mu_i)^2}{\sigma^2} = (Y_i - \mu_i)(\sigma^2)^{-1}(Y_i - \mu_i)$$

- The latter has interpretation in terms of a standardized distance of Y_i from μ_i, relative to the spread of values around μ_i

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 response.

Multivariate Normal Distribution

For the repeated measures design, we need to introduce additional vector and matrix notation to describe the multivariate normal density for the set of observations on the i^{th} subject.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$ denote the $(p \times 1)$ vector of responses, and

$$\mathbf{X}_i = \begin{bmatrix} 1 & X_{i11} & X_{i12} & \ldots & X_{i1,k-1} \\ 1 & X_{i21} & X_{i22} & \ldots & X_{i2,k-1} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & X_{ip1} & X_{ip2} & \ldots & X_{ip,k-1} \end{bmatrix}$$

denote the $(p \times k)$ matrix of covariates.

Then, we assume $\mathbf{Y}_i$ has a multivariate normal distribution with mean

$$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

and covariance matrix $\boldsymbol{\Sigma}$.

The multivariate normal probability density function has the following representation:

$$f\left(\mathbf{Y}_i\right) = f\left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right) =$$

$$\left(2\pi\right)^{-p/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left[-\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)/2\right]$$

where $|\boldsymbol{\Sigma}|$ is the *determinant* of $\boldsymbol{\Sigma}$ (also known as the *generalized variance*).
Note that $f\left(\mathbf{Y}_i\right)$ describes the probability or relative frequency of occurrence of a particular set of values of $(Y_{i1}, Y_{i2}, \ldots, Y_{ip})$.

Notable Features:

- $f(\mathbf{Y}_i)$ is completely determined by $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$
- $f(\mathbf{Y}_i)$ depends to a very large extent on

$$(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

- 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

$$(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) =$$

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

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, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

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

The main idea behind ML is really quite simple and conveyed by its name: use as estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ the values that are most probable (or "likely") for the data that we have observed.

That is, choose values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that maximize the probability of the response variables evaluated at their observed values (or that best predict the observed data).

The resulting values, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are called the *maximum likelihood estimates* of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

Independent observations:

To obtain maximum likelihood estimates of $\boldsymbol{\beta}$ in the ordinary one-way ANOVA model, we find the values of the regression parameters that maximize the probability density function.

With independent observations, the joint density is simply the product of the individual univariate normal densities for Y_{ij}.

Hence, we wish to maximize

$$f(Y) = \left(2\pi\sigma^2\right)^{-np/2} \exp\left[-\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}\right)^2 / 2\sigma^2\right],$$

evaluated at the observed values of the data, with respect to the regression parameters, $\boldsymbol{\beta}$.

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

$$-\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij}-\mathbf{X}_{ij}\boldsymbol{\beta}\right)^2/2\sigma^2$$

by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij}-\mathbf{X}_{ij}\boldsymbol{\beta}\right)^2/2\sigma^2$$

Note: This is equivalent to finding the least squares estimates of $\boldsymbol{\beta}$, i.e., the values that minimize the sum of the squares of the residuals.

The least squares solution can be written as

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\mathbf{X}_{ij}'\mathbf{X}_{ij}\right)\right]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\mathbf{X}_{ij}'Y_{ij}\right)$$

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

Next we consider how to extend these ideas to the setting of correlated data.

GENERALIZED LEAST SQUARES

To find the maximum likelihood estimate of $\boldsymbol{\beta}$ in the repeated measures setting we first assume that $\boldsymbol{\Sigma}$ is known (later, we will relax this unrealistic assumption).

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

$$\begin{aligned}
&\ln \{ (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \\
&\exp \left[-\sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \Sigma^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) /2 \right] \} \\
&= -\tfrac{np}{2} \ln (2\pi) - \tfrac{n}{2} \ln |\boldsymbol{\Sigma}| \\
&- \left[\sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) /2 \right]
\end{aligned}$$

or minimize

$$\sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

The estimate of $\boldsymbol{\beta}$ that minimizes this expression is known as the generalized least squares estimate and can be written as

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \left(\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)\right]^{-1} \sum_{i=1}^{n} \left(\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Y}_i\right)$$

This is the estimate that PROC MIXED provides.

Properties of GLS:

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

2. $Cov(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \left(\mathbf{X}_i' \mathbf{\Sigma}^{-1} \mathbf{X}_i\right)\right]^{-1}$

3. Sampling Distribution of $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left[\sum_{i=1}^{n} \left(\mathbf{X}_i' \mathbf{\Sigma}^{-1} \mathbf{X}_i\right)\right]^{-1}\right)$$

4. If $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$, GLS estimate reduces to the ordinary least squares estimate.

5. The most efficient generalized least squares estimate is the one that uses the true value of $\mathbf{\Sigma}$.

Since we usually do not know $\boldsymbol{\Sigma}$, we typically estimate it from the data.

In general, it is not possible to write down simple expressions for the ML estimate of $\boldsymbol{\Sigma}$. The ML estimate of $\boldsymbol{\Sigma}$ has to be found by using numerical algorithms that maximize the likelihood.

Once the ML estimate of $\boldsymbol{\Sigma}$ has been obtained, we simply substitute the estimate of $\boldsymbol{\Sigma}$, say $\widehat{\boldsymbol{\Sigma}}$, in the generalized least squares estimator to obtain the ML estimate of $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \left(\mathbf{X}_i' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_i\right)\right]^{-1} \sum_{i=1}^{n} \left(\mathbf{X}_i' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{Y}_i\right)$$

In large samples, the resulting estimator of $\boldsymbol{\beta}$ will have all the same properties as when $\boldsymbol{\Sigma}$ is known.

Statistical Inference

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

$$\left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}.$$

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

Note. Though we usually signify row vectors by the transpose symbol, e.g, $\mathbf{L}'$, we assume here that $\mathbf{L}$ is a matrix whose rows represent different linear combinations for a single linear combination, $\mathbf{L}$ is then a row vector.

Example: Suppose $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and let $L = (0, 0, 1)$, then $H_0 : L\boldsymbol{\beta} = 0$ is equivalent to $H_0 : \beta_2 = 0$.

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

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

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

Suppose that L is a single row vector.

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

Thus an approximate 95% confidence interval is given by:

$$L\widehat{\boldsymbol{\beta}} \pm 1.96\sqrt{LCov(\widehat{\boldsymbol{\beta}})L'}$$

Wald Test

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

$$Z = \frac{L\widehat{\boldsymbol{\beta}}}{\sqrt{LCov(\widehat{\boldsymbol{\beta}})L'}}$$

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

$$W = (L\widehat{\boldsymbol{\beta}})(LCov(\widehat{\boldsymbol{\beta}})L')^{-1}(L\widehat{\boldsymbol{\beta}})$$

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.

Suppose that L has r rows, then a simultaneous test of the r contrasts is given by

$$W = (L\widehat{\boldsymbol{\beta}})'(LCov(\widehat{\boldsymbol{\beta}})L')^{-1}(L\widehat{\boldsymbol{\beta}})$$

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.

Aside:

Suppose that one model (the "reduced" model) is a special case of the other (the "full" model). That is, the reduced model is simpler than the full model, so that when the reduced model holds the full model must necessarily hold. The reduced model is then 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*.

We can use likelihood ratio tests 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 that the multivariate normal probability density has the following representation:

$$(2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)/2\right]$$

where the subscript i refers to a subject rather than a single observation (multivariate versus univariate representation).

To obtain the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ we maximize the likelihood, which is the product of this expression over the n subjects.

The solutions to this maximization problem are the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

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

To see this problem, consider ordinary regression with independent errors.

If the model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{k-1} X_{ik-1} + e_i \\ &= \mathbf{X}_i \boldsymbol{\beta} + e_i \end{aligned}$$

and the n observations are independent, we can maximize the likelihood

$$\left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i\boldsymbol{\beta}\right)^2 / 2\sigma^2\right].$$

This gives the usual least squares estimator of $\boldsymbol{\beta}$, but the ML estimator of σ^2 is

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right)^2 / n$$

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

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

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 $\boldsymbol{\beta}$, also, is estimated. That is, in the estimator of σ^2 we have replaced $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$.

It should not be too surprising that similar problems arise in the estimation of $\boldsymbol{\Sigma}$.

Recall: An unbiased estimator is given by using $n - k$ 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 $\boldsymbol{\beta}$ from the likelihood so that it is defined only in terms of $\boldsymbol{\Sigma}$. This can be done in a number of ways.

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 $\boldsymbol{\beta}$.

For example, the residuals after estimating $\boldsymbol{\beta}$ by ordinary least squares can be used.

The likelihood for these residuals will depend only on $\boldsymbol{\Sigma}$, and not on $\boldsymbol{\beta}$.

Thus, rather than maximizing

$$-\frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)$$

REML maximizes the following slightly modified log-likelihood

$$\begin{aligned} -\frac{n}{2}\ln|\boldsymbol{\Sigma}| \quad &- \quad \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right) \\ &- \quad \frac{1}{2}\ln\left|\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right| \end{aligned}$$

When the residual likelihood is maximized, we obtain estimates of $\boldsymbol{\Sigma}$ whose degrees of freedom are corrected for the reduction in degrees of freedom due to estimating $\boldsymbol{\beta}$.

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 generalized least squares estimates of $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}_i\right)$$

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

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;

ASSUMPTIONS ABOUT THE COVARIANCE MATRIX

In the example, we allowed the covariance matrix to be "unrestricted" or "unstructured", allowing any valid pattern of variances and covariances.

The model was

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + e_{ij}$$

where the p errors for each subject have an unstructured covariance matrix.

Some widely-used methods for analysis of repeated measurements make special assumptions about the covariance matrix.

Historically, one of the most popular methods is known as "univariate" or "mixed-model" analysis of variance.

This model assumes the correlation between repeated measurements arises because each subject has an underlying (latent) level of response which persists over time or treatment.

This subject effect is treated as a random variable in the mixed-model ANOVA.

Thus, if the expected response to treatment is given by

$$E\left(Y_{ij}\right) = \mathbf{X}_{ij}\boldsymbol{\beta}$$

the response for subject i is assumed to differ from the population mean by a subject effect, b_i, and a within-subject measurement error, w_{ij}.

The mixed-model for repeated measurements is

$$\begin{aligned} Y_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta} + b_i + w_{ij} \\ &\quad (\beta_0 + b_i) + \beta_1 X_{ij1} + \ldots \beta_p X_{ijp} + w_{ij} \end{aligned}$$

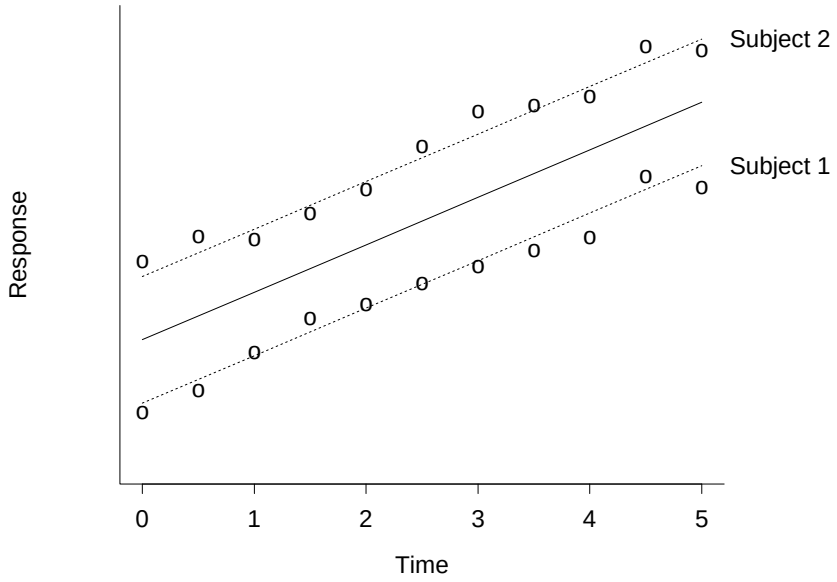
Note: The b_i and w_{ij} are assumed to be independent.

That is, we distinguish two sources of variation that account for differences in the response of subjects measured at the same occasion:

1. *Between-Subject Variation*: Different subjects simply respond differently; some are "high" responders, some are "low" responders, and some are "medium" responders.

2. *Within-Subject Variation*: Random variation arising from the process of measurement; e.g. due to measurement error and/or sampling variability.

Random Intercepts Model:



If we let $var\,(b_i) = \sigma_b^2$ and $var\,(w_{ij}) = \sigma_w^2$ the covariance matrix of the repeated measurement can be shown to have the following *compound symmetry* form:

$$\begin{bmatrix} \sigma_b^2 + \sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_w^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_w^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_w^2 \end{bmatrix}$$

The compound symmetry assumption is usually inappropriate for longitudinal data. Why?

The validity of the assumption of compound symmetry can be checked in the data.

To fit the model under the assumption of compound symmetry, simply change the ***repeated*** statement in PROC MIXED to read:

```
repeated time/type=cs subject=id r;
```

SAS Output from Analysis of the Blood Pressure Data
Under the Assumption of Compound Symmetry

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	79.8618	53.4943	53.4943	53.4943
2	53.4943	79.8618	53.4943	53.4943
3	53.4943	53.4943	79.8618	53.4943
4	53.4943	53.4943	53.4943	79.8618

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > |Z|
CS	id	53.4943	19.5336	2.74	0.0062
Residual		26.3675	4.9391	5.34	<0.0001

Fit Statistics

-2 Res Log Likelihood	518.3
AIC (smaller is better)	522.3
AICC (smaller is better)	522.5
BIC (smaller is better)	524.3

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > Chi Sq
1	42.23	<.0001

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		102.600	1.9983	19	51.34	<0.0001
trt	0	7.200	1.6238	57	4.43	<0.0001
trt	1	4.750	1.6238	57	2.93	0.0049
trt	2	-3.050	1.6238	57	-1.88	0.0655
trt	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > Chi Sq
trt	3	57	48.47	<0.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
T2 - T1	-7.8000	1.6238	57	-4.80	<0.0001

Assessing the Adequacy of Compound Symmetry

The adequacy of the compound symmetry assumption can be formally tested by comparing the goodness of fit of the compound symmetry to the unrestricted or unstructured model.

Recall: Fit Statistics for **Unstructured Covariance**

Fit Statistics

-2 Res Log Likelihood	504.8
AIC (smaller is better)	524.8
AICC (smaller is better)	528.2
BIC (smaller is better)	534.7

To check whether the assumption of compound symmetry is appropriate, we use the information that is provided in the *Fit Statistics* panel of the SAS output.

The Res Log Likelihood measures the goodness of fit of the assumed covariance structure.

To calculate the likelihood ratio test comparing the compound symmetry to the unstructured model, take twice the difference in log likelihoods and compare to the chi-squared distribution.

Example:

	CS	UN
Res LL	-259.2	-252.4
-2 LL	518.3	504.8
No. of Covariance Parameters	2	10

-2 Log Likelihood Ratio: 518.3 - 504.8 = 13.5 with 8 (i.e., $10 - 2$) degrees of freedom, $p \approx 0.097$.

Therefore, at the 0.05 significance level, we fail to reject the assumption of compound symmetry.

Thus, the compound symmetry assumption appears to be adequate for these data.

In general, however, the compound symmetry assumption is inappropriate for longitudinal data.

Later in the course we will consider alternative models for the covariance.

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 $\mathbf{Y}^{(o)}$ denote the measurements observed and $\mathbf{Y}^{(m)}$ 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.

A hierarchy of three different types of missing data mechanisms can be distinguished:

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

3) Missing data are nonignorable when the probability that an individual value will be missing depends on $\mathbf{Y}^{(m)}$. If missing values are nonignorable, standard methods of analysis are not valid. Usually, a sensitivity analysis is recommended.

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

CROSSOVER DESIGNS

So far, we have considered the single-group repeated measures design where each subject receives each of p treatments at p different occasions.

Next we consider a variant of the single-group repeated measures design known as the crossover design.

In the simplest version of the cross-over design, two treatments, say A and B, are to be compared. Subjects are randomly assigned to the two treatment orders: A→B and B→A.

Example: Placebo-controlled study of the effect of erythropoietin on plasma histamine levels and pruritus scores of 10 dialysis patients.

Treatment schedule was 5 weeks of placebo and 5 weeks of erythropoietin in random order.

Designs in which subjects are randomly assigned to either P→T (placebo, treatment) or T→P are called two-period crossover designs.

If we assume that there is no carryover[2] of treatment effects from period 1 to period 2, we can write the basic model as

$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_{ij} + e_{ij}$$

where Y_{ij} is the response of subject i at time j, and time_{ij} and trt_{ij} are the values of the time and treatment variable associated with Y_{ij}.
If there is a carryover of the effect of treatment (e.g. erythropoietin) from period 1 to period 2, we need to define a new indicator variable:

$$CO_{ij} = \begin{array}{l} 1, \text{if } T \text{ given in the previous period;} \\ 0, \text{otherwise.} \end{array}$$

This indicator variable will equal 1 only in the second period for the group assigned to T→P.

[2] ***Carryover*** is the persistence of a treatment effect applied in one period in a subsequent period of treatment.

Example: Placebo-controlled study of the effect of erythropoietin on plasma histamine levels.

id	time	trt	co	y
1	1	1	0	24
1	2	2	0	5
2	1	1	0	23
2	2	2	0	8
3	1	1	0	19
3	2	2	0	3
4	1	1	0	26
4	2	2	0	8
5	1	1	0	16
5	2	2	0	3
6	1	2	0	2
6	2	1	1	18
7	1	2	0	8
7	2	1	1	29
8	1	2	0	5
8	2	1	1	26
9	1	2	0	6
9	2	1	1	28
10	1	2	0	4
10	2	1	1	19

SAS CODE

```
data hist;
      infile 'g:\shared\bio226\histam.dat';
      input id time trt co y;
run;

proc sort data=hist;
      by time trt;
run;

proc means;
      var y;
      by time trt;
run;

proc mixed data=hist;
      class id time trt co;
      model y=time trt co /s chisq;
      repeated time / type=un subject=id r;
run;
```

OUTPUT OF PROC MEANS

Analysis Variable: Y

time=1 trt=1

N	Mean	Std Dev	Minimum	Maximum
5	21.6000000	4.0373258	16.0000000	26.0000000

time=1 trt=2

N	Mean	Std Dev	Minimum	Maximum
5	5.0000000	2.2360680	2.0000000	8.0000000

time=2 trt=1

N	Mean	Std Dev	Minimum	Maximum
5	24.0000000	5.1478151	18.0000000	29.0000000

time=2 trt=2

N	Mean	Std Dev	Minimum	Maximum
5	5.4000000	2.5099801	3.0000000	8.0000000

The MIXED Procedure
Unstructured Covariance - With Carryover Effects

Estimated R Matrix for id 1

Row	Col1	Col2
1	10.6500	9.4750
2	9.4750	16.4000

Fit Statistics

-2 Res Log Likelihood	87.4
AIC (smaller is better)	93.4
AICC (smaller is better)	95.4
BIC (smaller is better)	94.3

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	7.4000	3.3660	9	2.20	0.0555
time 1	-0.4000	2.3259	9	-0.17	0.8673
time 2	0	.	.	.	.
trt 1	16.6000	2.0640	9	8.04	<0.0001
trt 2	0	.	.	.	.
co 0	-2.0000	4.2895	9	-0.47	0.6521
co 1	0	.	.	.	.

The MIXED Procedure
Compound Symmetry - With Carryover Effects

Estimated R Matrix for id 1

Row	Col1	Col2
1	13.5250	9.4750
2	9.4750	13.5250

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	id	9.4750
Residual		4.0500

Fit Statistics

-2 Res Log Likelihood	88.1

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |T|
Intercept	7.4000	3.4504	9	2.14	0.0606
time 1	-0.4000	2.3259	7	-0.17	0.8683
time 2	0	.	.	.	.
trt 1	16.6000	2.3259	7	7.14	0.0002
trt 2	0	.	.	.	.
co 0	-2.0000	4.2895	7	-0.47	0.6552
co 1	0	.	.	.	.

Test for compound symmetry compares -2 Res Log Likelihood from the unstructured and compound symmetry models:

	-2 Res Log L
Compound Symmetry	88.1
Unstructured	87.4

$\Rightarrow$ $-2*$Res Log Likelihood Ratio $= 0.7$ with 1 df.

Assumption of compound symmetry is defensible.

Note: With carryover effects, the estimated treatment effect, 16.6, is based entirely on the data in the first period.

That is, the estimate of the treatment effect is not based on within-subject comparisons, and has standard error of 2.33.

Next, consider model without carryover effects.

The MIXED Procedure
Compound Symmetry - Without Carryover Effects
Estimated R Matrix for ID 1

Row	Col1	Col2
1	12.5250	8.4750
2	8.4750	12.5250

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	id	8.4750
Residual		4.0500

Fit Statistics

-2 Res Log Likelihood	93.0

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |T|
Intercept	5.9000	1.2062	9	4.89	0.0009
time 1	-1.4000	0.9000	8	-1.56	0.1584
time 2	0	.	.	.	.
trt 1	17.6000	0.9000	8	19.56	<0.0001
trt 2	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
time	1	8	2.42	0.1198
trt	1	8	382.42	0.0001

The pooled estimate of the treatment effect, combining the responses from period 1 and period 2, is 17.6.

Note: Assuming no carryover effects, the treatment effect is based on within-subject comparisons, and the standard error has decreased from 2.33 to 0.90.

STATISTICAL PROPERTIES OF THE CROSSOVER DESIGN

The compound symmetry model can be written

$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_{ij} + \beta_3 \text{CO}_{ij} + b_i + w_{ij}$$

where b_i is the subject effect, with

$$\begin{aligned} var\,(b_i) &= \sigma_b^2 \\ var\,(w_{ij}) &= \sigma_w^2 \end{aligned}$$

Then

$$\begin{aligned} var\,(Y_{ij}) &= \sigma_b^2 + \sigma_w^2, \\ cov\,(Y_{i1}, Y_{i2}) &= \sigma_b^2 \end{aligned}$$

Let

$$\begin{array}{ll} \bar{Y}_{P1} & \bar{Y}_{T2} \\ \bar{Y}_{T1} & \bar{Y}_{P2} \end{array}$$

be the mean responses at each combination of treatment and time.

Then, if $\bar{d}_1 = \bar{Y}_{P1} - \bar{Y}_{T2}$ and $\bar{d}_2 = \bar{Y}_{P2} - \bar{Y}_{T1}$,

$$Var\left(\bar{d}_j\right) = 2\sigma_w^2/n,$$

where n is the number of subjects receiving each sequence.

With no carryover effects, the treatment effect is estimated by

$$\left(\bar{d}_1 + \bar{d}_2\right)/2, \text{ which has variance } \sigma_w^2/n.$$

In contrast, with two independent groups, and $2n$ subjects in each group, an estimate of the treatment effect has variance $\left(\sigma_b^2 + \sigma_w^2\right)/n$.

Thus, the crossover design has the potential to substantially increase the precision of the estimate of the treatment effect.

The difficulty is that the carryover effect is estimated by

$$\left(\bar{Y}_{T2} - \bar{Y}_{P2}\right) - \left(\bar{Y}_{T1} - \bar{Y}_{P1}\right) = \left(\bar{Y}_{T2} + \bar{Y}_{P1}\right) - \left(\bar{Y}_{T1} + \bar{Y}_{P2}\right)$$

which has variance

$$\left(8\sigma_b^2 + 4\sigma_w^2\right)/n.$$

Thus, the test for carryover will be much less powerful than the test for treatment effects. It can fail to detect interactions that are substantially larger than the putative treatment effects.

Dilemma: Choice of estimator for treatment effect is between an efficient but potentially biased estimator (using within-subject comparisons) and an unbiased but inefficient estimator (using between-subject comparisons).

This problem is an intractable feature of the simple crossover design. Thus, it should be used only when carryover is biologically implausible.

Carryover can often be avoided by having a sufficiently long wash-out time between the two periods.

Textbooks by Senn (1993), Jones and Kenward (1989) and Crowder and Hand (1989) describe a wide variety of more complex crossover designs that are less vulnerable to the problem of carryover.

For example, Crowder and Hand (1989) recommend the design

Group 1	PTT
Group 2	TPP

SUMMARY

When the crossover design is used, it is important to avoid carryover.

Designing a crossover study requires knowledge and consideration of the disease and the likely effects of treatment:

- The disease should be chronic and stable
- The effects of treatment should develop fully within the treatment period

Washout periods should be sufficiently long for complete reversibility of treatment effect.

The crossover design may be useful for demonstrating the bio-equivalence of two formulations of the same drug.

PARALLEL GROUPS REPEATED MEASURES DESIGN

In the parallel groups design, two or more groups of subjects are defined and measured repeatedly over time.

The groups can be defined by characteristics of the study subjects, such as age, gender, or baseline blood pressure level. Studies based on such categories are observational.

Alternatively, groups can be defined by randomization to alternative treatments.

We consider designs in which all subjects are intended to be measured at the same set of follow up times (*balanced*). However, the methods will allow for missing data (*incomplete*).

The main goal of the analysis will be to characterize the *patterns of change* over time in the several groups and to determine whether those patterns differ in the groups.

Example:

In a randomized clinical trial of HIV infected patients, we may wish to study whether alternative treatments have differential effects on the pattern of decline of CD4 count.

We focus initially on the two-groups design.

Generalizations to more than two groups are straightforward.

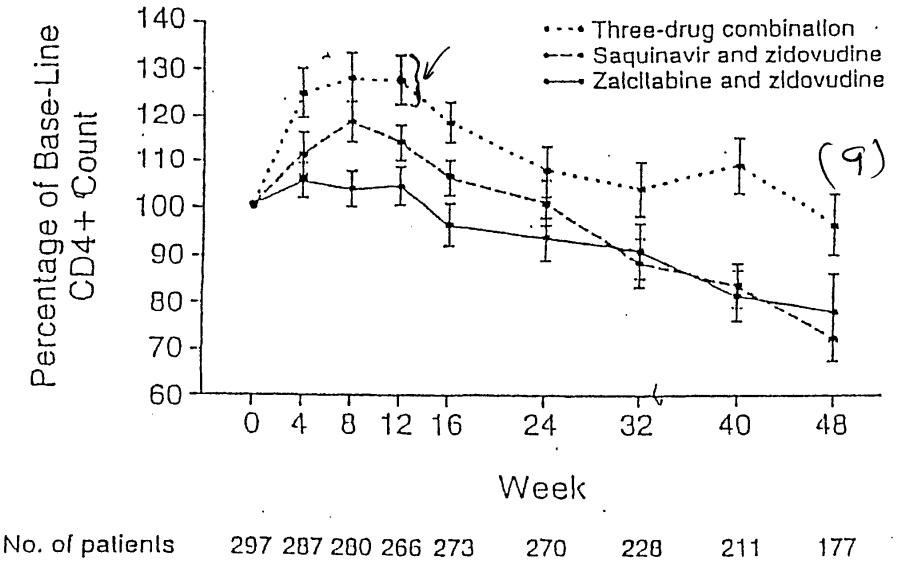


Figure 1. Mean (±SE) CD4+ Cell Counts, Expressed as Percentages of the Base-Line Count, According to Study Week.

The numbers below the graph are the numbers of patients studied in the weeks shown.

DATA STRUCTURE FOR THE BALANCED TWO-GROUPS REPEATED MEASURES DESIGN

		1	2	3	...	p
Group 1						
Subj	1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1p}
	2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2p}
	.	.	.	.	...	.
	.	.	.	.	...	.
	m	Y_{m1}	Y_{m2}	Y_{m3}	...	Y_{mp}
Group 2						
Subj	$m+1$	$Y_{m+1,1}$	$Y_{m+1,2}$	$Y_{m+1,3}$	...	$Y_{m+1,p}$
	$m+2$	$Y_{m+2,1}$	$Y_{m+2,2}$	$Y_{m+2,3}$	...	$Y_{m+2,p}$
	.	.	.	.	...	.
	.	.	.	.	...	.
	n	Y_{n1}	Y_{n2}	Y_{n3}	...	Y_{np}

Models for the Mean Response

Next we discuss different choices for modelling the mean response over time and emphasize that they can all be described by the general linear model

$$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

for appropriate choices of $\mathbf{X}_i$.

We can distinguish two basic strategies for modelling the time trend:

I. **Arbitrary Means (Profile Analysis)**

II. **Parametric Curves**

In the following, we will assume that each subject is measured on all p occasions, later we will show how to modify models to account for less than p responses due to missed examinations or dropouts.

I. Profile Analysis

Consider the following example from a two group randomized trial comparing *Treated* and *Control*.

	Occasion			
Group	1	2	...	p
Treated	μ_{T1}	μ_{T2}	...	μ_{Tp}
Control	μ_{C1}	μ_{C2}	...	μ_{Cp}
Difference	Δ_1	Δ_2	...	Δ_p

In this study we are primarily interested in testing the null hypothesis of no treatment effect.

The appropriate test of no treatment effect will depend on whether $\mathbf{Y}_i$ includes a baseline response or only post-randomization responses.

Null Hypothesis of No Treatment Effect

(a) $\mathbf{Y}_i$ includes baseline response:

$H_0 : \Delta_1 = ... = \Delta_p$

H_0 : no time-by-group interaction

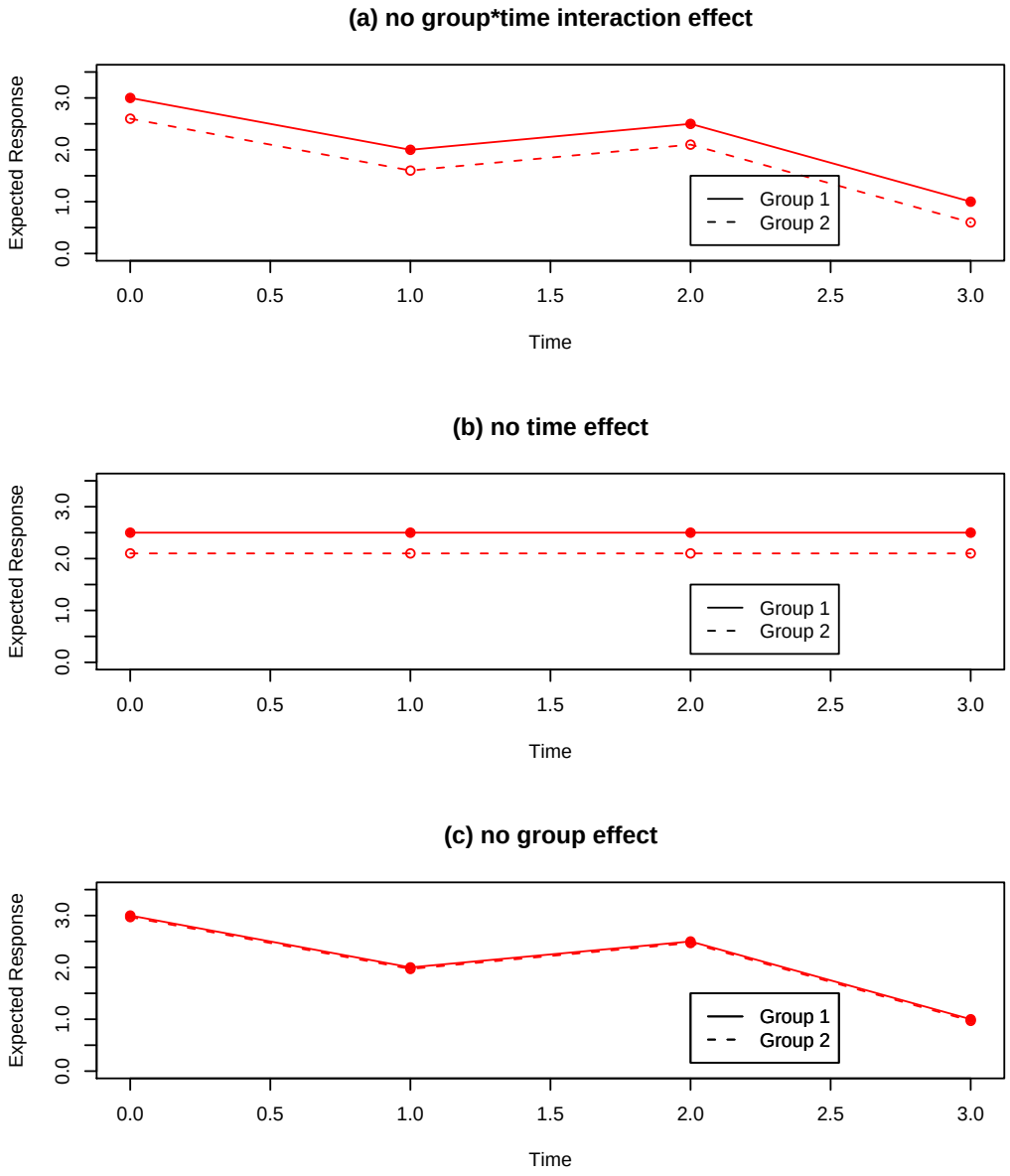
Are the "profiles of means" similar in the two treatment groups, in the sense that the line segments between adjacent occasions are parallel?

(b) $\mathbf{Y}_i$ is post-randomization response only:

$H_0 : \Delta_1 = \ldots = \Delta_p = 0$

H_0 : no group effect

Are the profiles of means parallel and also at the same level, i.e. do the profiles of means coincide?



Graphical representation of the null hypothesis of (a) no group $\times$ time interaction effect, (b) no time effect, and (c) no group effect.

In profile analysis we can distinguish three hypotheses that may be of scientific interest.

Scientific hypotheses:

H_{10}: Are the profiles of means similar in the groups, in the sense that the line segments between adjacent occasions are parallel? This is the hypothesis of *no group by time interaction*.

H_{20}: If the population profiles are parallel, are they also at the same level? This is the hypothesis of *no group effect*.

H_{30}: If the population profiles are parallel, are the means constant over time? This is the hypothesis of *no time effect*.

Although these general formulations of the study hypotheses are a good place to begin, the appropriate hypotheses in a particular study must be derived from the relevant scientific issues in that investigation.

General Linear Model Formulation

Let n be the number of subjects and N be the total number of observations. Consider the 'univariate representation' of the data, with one row for each observation of the dependent variable.

To write the model for the two-group repeated measures design with p occasions of measurement, we must define $p-1$ indicator variables.

For the i^{th} observation in the transformed data set ($i = 1, \ldots, N$), let

$$X_{ij} = \begin{array}{l} 1, \text{observation taken at time } j; \\ 0, \text{otherwise.} \end{array}$$

for $j = 1, \ldots, p-1$.

We can let $X_{i,p}$ be the indicator variable for group. That is

$$X_{i,p} = \begin{array}{l} 1, \text{observation in group 1;} \\ 0, \text{observation in group 2.} \end{array}$$

The interaction variables can be thought of as products of the time and group indicators,

$$X_{i,p+j} = X_{ij}X_{ip} \qquad j = 1, \ldots, p-1$$

When interaction terms are included, the model has $2p$ regression parameters.

For example, if the referent occasion is time p, the mean value in group 1 at time 1 is

$$\beta_0 + \beta_1 + \beta_p + \beta_{p+1}$$

while the mean value in group 2 is

$$\beta_0 + \beta_1$$

	Time			
	1	2	...	p
Group 1	$\beta_0 + \beta_1 + \beta_p + \beta_{p+1}$	$\beta_0 + \beta_2 + \beta_p + \beta_{p+2}$	...	$\beta_0 + \beta_p$
Group 2	$\beta_0 + \beta_1$	$\beta_0 + \beta_2$	...	β_0

Thus, profile analysis can be expressed in terms of the general linear model

$$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

To conduct a profile analysis of data from two or more treatment groups measured repeatedly over time, we can use the following SAS code:

```
proc mixed;
      class id trt time t;
      model y=trt time trt*time /s chisq;
      repeated t / type=un subject=id r;
run;
```

This model assumes an unstructured covariance matrix. However, alternative assumptions about the covariance structure can be considered.

Next, we consider a profile analysis of the blood lead data of 100 children from the treatment and placebo groups of the Treatment of Lead-Exposed Children (TLC) trial to illustrate these ideas.

Example: *Treatment of Lead-Exposed Children Trial*

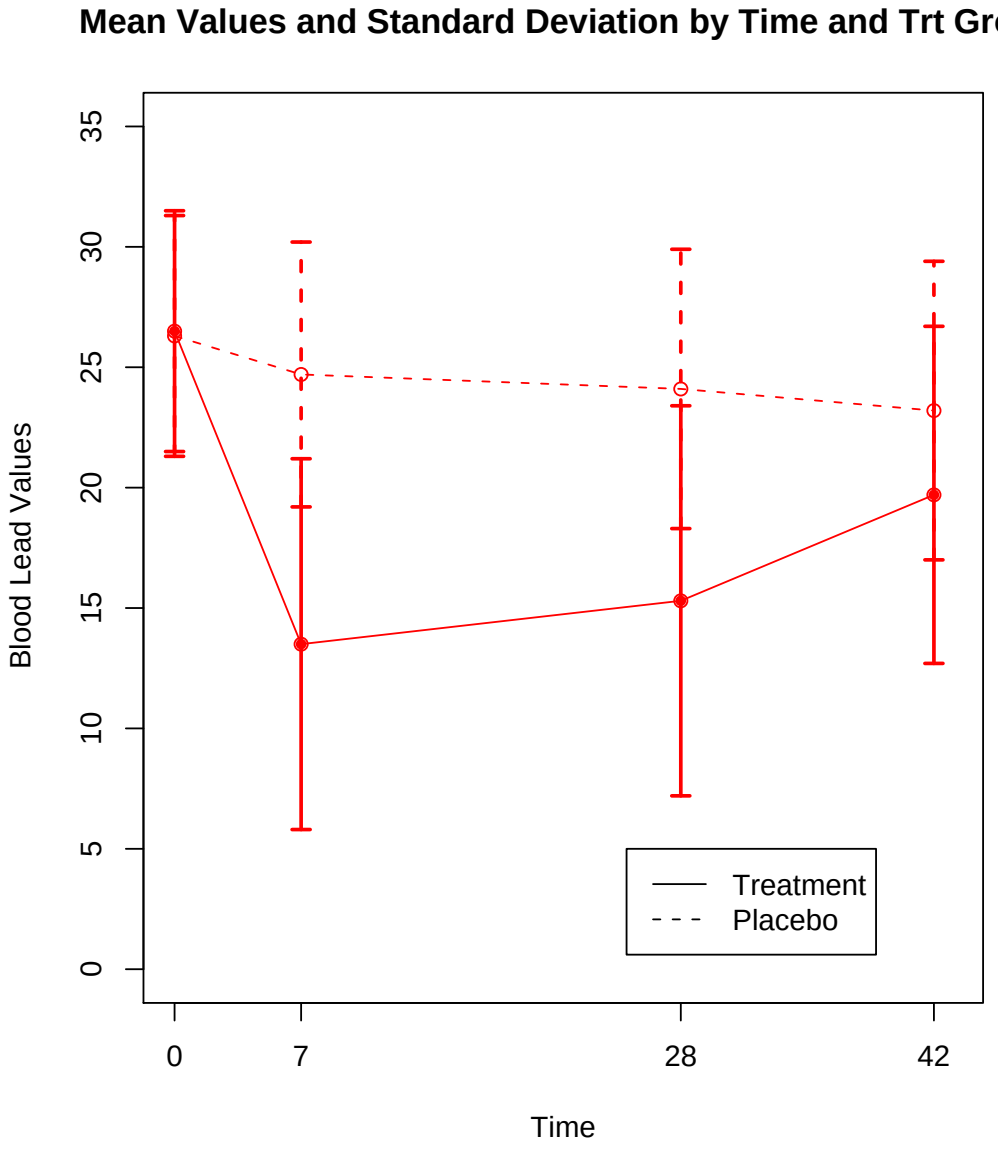
- Exposure to lead during infancy is associated with substantial deficits in tests of cognitive ability
- Chelation treatment of children with high lead levels usually requires injections and hospitalization
- A new agent, *Succimer*, can be given orally
- Randomized trial examining changes in blood 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

BLOOD LEAD VALUES FROM THE FIRST TEN OF 100 CHILDREN TREATED WITH EITHER ACTIVE THERAPY (TRT = A) OR PLACEBO (TRT = P) AND MEASURED AT BASELINE AND 7, 28, AND 42 DAYS

ID	TRT	PbB_1	PbB_2	PbB_3	PbB_4
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
256	P	19.7	14.9	15.3	14.7
416	P	31.1	31.2	29.2	30.1

MEAN VALUES (SD) BY TIME AND TRT GROUP

	T1	T2	T3	T4
Treatment	26.5	13.5	15.3	19.7
	(5.0)	(7.7)	(8.1)	(7.0)
Placebo	26.3	24.7	24.1	23.2
	(5.0)	(5.5)	(5.8)	(6.2)


Mean Values and Standard Deviation by Time and Trt Group
Blood Lead Values
0
5
10
15
20
25
30
35
0
7
28
42
Time
Treatment
Placebo

SAS CODE

```
data lead;
      infile 'g:\shared\bio226\leadnew.dat';
      input id trt $ y1 y2 y3 y4;
      y=y1; time=1; t=1; output;
      y=y2; time=2; t=2; output;
      y=y3; time=3; t=3; output;
      y=y4; time=4; t=4; output;
run;

proc mixed data = lead noclprint;
      class id trt time t;
      model y=trt time time*trt / s chisq;
      repeated t / type=un subject=id r;
      estimate 'Trt*Time 2'
            trt*time -1 1 0 0 1 -1 0 0 / e;
      estimate 'Trt*Time 3'
            trt*time -1 0 1 0 1 0 -1 0;
      estimate 'Trt*Time 4'
            trt*time -1 0 0 1 1 0 0 -1;
run;
```

Summary of SAS Output

Estimated R Matrix for id 3200046

Row	Col1	Col2	Col3	Col4
1	25.2257	19.0504	19.4257	17.1781
2	19.0504	44.3065	35.3026	27.4906
3	19.4257	35.3026	48.9190	31.4383
4	17.1781	27.4906	31.4323	43.5820

Fit Statistics

-2 Res Log Likelihood	2385.8
AIC (smaller is better)	2405.8
AICC (smaller is better)	2406.4
BIC (smaller is better)	2431.8

Solution for Fixed Effects

Effect	trt	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			23.2440	0.9336	98	24.90	<.0001
trt	A		-3.5700	1.3203	98	-2.70	0.0081
trt	P		0	.	.	.	.
time		1	3.0280	0.8301	98	3.65	0.0004
time		2	1.4146	0.8113	98	1.75	0.0840
time		3	0.8160	0.7699	98	1.06	0.2918
time		4	0	.	.	.	.
trt*time	A	1	3.8380	1.1739	98	3.27	0.0015
trt*time	A	2	-7.5940	1.1473	98	-6.62	<.0001
trt*time	A	3	-5.2380	1.0888	98	4.81	<.0001
trt*time	A	4	0	.	.	.	.
trt*time	P	1	0	.	.	.	.
trt*time	P	2	0	.	.	.	.
trt*time	P	3	0	.	.	.	.
trt*time	P	4	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trt	1	98	29.32	<.0001
time	3	98	61.56	<.0001
trt*time	3	98	37.68	<.0001

Coefficients for Trt*Time 2

Effect	trt	time	Row 1
Intercept			
trt	A		
trt	P		
time		1	
time		2	
time		3	
time		4	
trt*time	A	1	-1
trt*time	A	2	1
trt*time	A	3	
trt*time	A	4	
trt*time	P	1	1
trt*time	P	2	-1
trt*time	P	3	
trt*time	P	4	

Estimates

Label	Estimate	Standard Error	DF	t	Pr > |t|
Trt*Time 2	-11.4320	1.1213	98	-10.20	<0.0001
Trt*Time 3	-9.0760	1.1882	98	-7.64	<0.0001
Trt*Time 4	-3.8380	1.1739	98	-3.27	0.0015

For illustrative purposes, consider the model with compound symmetry covariance.

Fit Statistics

-2 Res Log Likelihood	2415.5
AIC (smaller is better)	2419.5
AICC (smaller is better)	2419.5
BIC (smaller is better)	2424.7

A likelihood ratio test for compound symmetry

	-2 Res Log L
Compound Symmetry	2415.5
Unstructured	2385.8

$\Rightarrow$ -2*Res log likelihood ratio = 29.7, 8 df. ($p < 0.0005$)

Clearly, the assumption of compound symmetry is not defensible.

Summary of Features of Profile Analysis

- Does not assume any specific time trend
- May have low power to detect specific trends; e.g., linear trends
- Can be used to accommodate "area under the curve" (AUC) analyses or other linear combinations of response vector
- How to incorporate mistimed measurements?

II. Parametric Curves

An alternative approach for analyzing the parallel-groups repeated measures design is to consider parametric curves for the time trends.

In this approach we model the means as an explicit function of time.

(a) Linear Trend

If the means tend to change linearly over time we can fit the following model:

$$E(Y_{ij}) = \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Trt}_i + \beta_3 \text{Time}_j \times \text{Trt}_i$$

Then, for subjects in treatment group 2,

$$E(Y_{ij}) = \beta_0 + \beta_1 \text{Time}_j$$

While for subjects in treatment group 1,

$$E(Y_{ij}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Time}_j$$

Thus, each group's mean is assumed to change linearly over time.

SAS CODE FOR MODELING A LINEAR TREND IN TIME

```
proc mixed;
      class id trt t;
      model y=trt time time*trt / s chisq;
      repeated t / type=un subject=id r;
run;
```

Note: t is a copy of the variable *time*.

In this model *time* is no longer included in the list of class variables.

This model yields the following expected values:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Trt}_i + \beta_3 \text{Time}_j \times \text{Trt}_i$$

where Time_j is modeled as a continuous variable and Trt_i is an indicator variable which takes the value 1 if the subject receives treatment 1, and zero otherwise.

Example: *Treatment of Lead-Exposed Children Trial*

We return now to the analysis of the Treatment of Lead-Exposed Children (TLC) trial and consider a linear trend in time.

MEAN VALUES BY TIME AND TRT GROUP

	T1	T2	T3	T4
Treatment	26.5	13.5	15.3	19.7
	(5.0)	(7.7)	(8.1)	(7.0)
Placebo	26.3	24.7	24.1	23.2
	(5.0)	(5.5)	(5.8)	(6.2)

Note: It would appear that the mean values in the Placebo group only can be described by a linear trend. We will proceed with the analysis for illustrative purposes only.

SAS CODE

```
data lead;
      infile 'g:\shared\bio226\leadnew.dat';
      input id trt $ y1 y2 y3 y4;
      y=y1; time=0; t=1; output;
      y=y2; time=1; t=2; output;
      y=y3; time=4; t=3; output;
      y=y4; time=6; t=4; output;
run;

proc mixed data=lead method=reml;
      class id trt t;
      model y=trt time time*trt / s chisq;
      repeated t / type=un subject=id r;
run;
```

MODELING LINEAR TREND IN TIME, UNSTRUCTURED COVARIANCE MATRIX

ML Estimation

Estimated R Matrix for ID 3200046

Row	col1	col2	col3	col4
1	27.2306	8.4399	11.9547	15.0550
2	8.4399	100.39	74.6766	38.7560
3	11.9547	74.6766	76.3701	39.4166
4	15.0550	38.7560	39.4166	45.8154

Note that the estimated covariance matrix is discernibly different from that obtained in the profile analysis. Why?

Fit Statistics

-2 Res Log Likelihood	2530.4
AIC (smaller is better)	2550.4
AICC (smaller is better)	2551.0
BIC (smaller is better)	2576.5

Solution for Fixed Effects

Effect		Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		26.0485	0.6959	98	37.43	<0.0001
trt	A	-1.5086	0.9842	98	-1.53	0.1285
trt	P	0	.	.	.	.
time		-0.4122	0.1228	98	-3.36	0.0011
trt*time	A	-0.0459	0.1737	98	-0.26	0.7922
trt*time	P	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trt	1	98	2.35	0.1285
time	1	98	25.10	<.0001
trt*time	1	98	0.07	0.7922

For participants in the placebo group,

$$E\left(Y_{ij}\right) = 26.05 - 0.41 \text{ Time}_j$$

while for participants in the active treatment group,

$$\begin{aligned} E\left(Y_{ij}\right) &= (26.05 - 1.51) - (0.41 + 0.05) \text{ Time}_j \\ &= 24.54 - 0.46 \text{ Time}_j \end{aligned}$$

Is the model with linear trend in time appropriate for these data?

Recall: The linear trend model and profile analysis (where time is treated as a categorical variable) are *nested* models.

Assessing Adequacy of Linear Trend Model

Compare the ML log likelihood of the linear trend model to the ML log likelihood of the model with time treated as a categorical variable (i.e., profile analysis).

Note: We must re-fit both models using ML rather than REML (the default):

proc mixed data=lead method=ml;

	-2 (ML) Log L
Profile Analysis	2394.4
Linear Trend Model	2527.8

-2*log likelihood ratio = 133.4, 4 df.
($p < 0.0001$)

$\Rightarrow$ Linear trend model is clearly not defensible.

(b) Quadratic Trend

If the means tend to change over time in a quadratic manner, we can fit the following model:

$$E(Y_{ij}) = \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Time}_j^2 + \beta_3 \text{Trt}_i + \beta_4 \text{Time}_j \times \text{Trt}_i + \beta_5 \text{Time}_j^2 \times \text{Trt}_i$$

Then, for subjects in treatment group 2,

$$E(Y_{ij}) = \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Time}_j^2$$

While for subjects in treatment group 1,

$$E(Y_{ij}) = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) \text{Time}_j + (\beta_2 + \beta_5) \text{Time}_j^2$$

Aside: To avoid problems of collinearity in the quadratic (or in any higher-order polynomial) trend model, should always "center" Time_j on its mean prior to the analysis (i.e. replace Time_j by its deviation from the mean).

For example, suppose $\text{Time}_j = (1, 2, ..., 10)$.

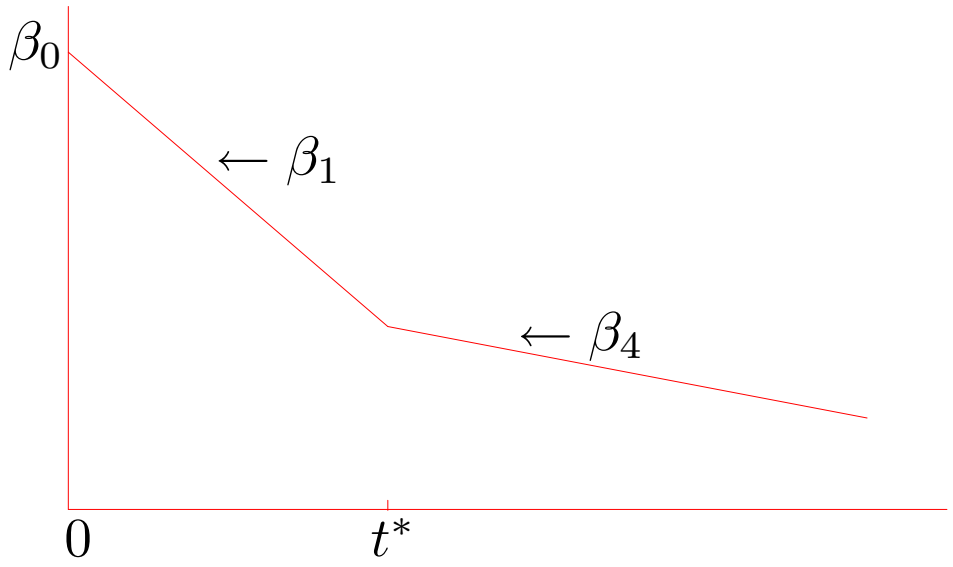
The correlation between Time_j and Time_j^2 is 0.975.

However, if we create a "centered" variable, say $\text{Time(C)}_j = \left(\text{Time}_j - \overline{\text{Time}}\right)$, then the correlation between Time(C)_j and Time(C)_j^2 is zero.

(c) Linear Spline

If the means change over time in a piecewise linear manner, we can fit the following linear spline model with knot at t^*:

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Trt}_i + \beta_3 \text{Time}_j \times \text{Trt}_i \quad \text{Time}_j \leq t^* \\ E(Y_{ij}) &= \beta_0 + \beta_1 t^* + \beta_2 \text{Trt}_i + \beta_3 t^* \times \text{Trt}_i \\ &\quad + \beta_4(\text{Time}_j - t^*) + \beta_5(\text{Time}_j - t^*) \times \text{Trt}_i \quad \text{Time}_j > t^* \end{aligned}$$



Then, for subjects in treatment group 2,

$$\begin{array}{rcll} E\left(Y_{ij}\right) & = & \beta_0 + \beta_1 \text{Time}_j & \text{Time}_j \leq t^* \\ E\left(Y_{ij}\right) & = & \beta_0 + \beta_1 t^* + \beta_4(\text{Time}_j - t^*) & \text{Time}_j > t^* \end{array}$$

While for subjects in treatment group 1,

$$\begin{array}{rcll} E\left(Y_{ij}\right) & = & (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Time}_j & \text{Time}_j \leq t^* \\ E\left(Y_{ij}\right) & = & (\beta_0 + \beta_2) + (\beta_1 + \beta_3)t^* & \\ & & +(\beta_4 + \beta_5)(\text{Time}_j - t^*) & \text{Time}_j > t^* \end{array}$$

Note that models with more than one knot can be considered.

Example: *Treatment of Lead-Exposed Children Trial*

In the Treatment of Lead-Exposed Children (TLC) trial it would appear that a piecewise linear model with knot at week 1 (T2) might be appropriate.

MEAN VALUES BY TIME AND TRT GROUP

	T1	T2	T3	T4
Treatment	26.5	13.5	15.3	19.7
	(5.0)	(7.7)	(8.1)	(7.0)
Placebo	26.3	24.7	24.1	23.2
	(5.0)	(5.5)	(5.8)	(6.2)

SAS CODE

```
data lead;
      infile 'g:\shared\bio226\leadnew.dat';
      input id trt $ y1 y2 y3 y4;
      y=y1; time=0; t=1; output;
      y=y2; time=1; t=2; output;
      y=y3; time=4; t=3; output;
      y=y4; time=6; t=4; output;
run;

data spline;
      set lead;

 st1=min(time,1);
 st2=max(0,time-1);



 proc mixed data=spline;
      class id trt t;
      model y=trt st1 st2 st1*trt st2*trt / s chisq;
      repeated t / type=un subject=id r;
run;
```

Fit Statistics

-2 Res Log Likelihood	2403.8
AIC (smaller is better)	2423.8
AICC (smaller is better)	2424.4
BIC (smaller is better)	2449.9

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		26.2522	0.7086	98	37.05	<.0001
trt	A	0.4422	1.0021	98	0.44	0.6600
trt	P	0	.	.	.	.
st1		-1.6066	0.7927	98	-2.03	0.0454
st2		-0.2653	0.1561	98	-1.70	0.0925
st1*trt	A	-11.4800	1.1211	98	-10.24	<.0001
st1*trt	P	0	.	.	.	.
st2*trt	A	1.3608	0.2208	98	6.16	<.0001
st2*trt	P	0	.	.	.	.

Then, for subjects in the placebo group,

$$\begin{aligned} E(Y_{ij}) &= 26.25 - 1.61\text{Week}_j && \text{Week}_j \leq 1 \\ E(Y_{ij}) &= (26.25 - 1.61) - 0.27(\text{Week}_j - 1) && \text{Week}_j > 1 \end{aligned}$$

While for subjects in the succimer group,

$$\begin{aligned} E(Y_{ij}) &= (26.25 + 0.44) - (1.61 + 11.48)\text{Week}_j && \text{Week}_j \leq 1 \\ E(Y_{ij}) &= (26.25 + 0.44 - 1.61 - 11.48) \\ &\quad +(-0.27 + 1.36)(\text{Week}_j - 1) && \text{Week}_j > 1 \end{aligned}$$

PREDICTED MEAN VALUES

	T1	T2	T3	T4
Treatment	26.7	13.6	16.9	19.1
Placebo	26.3	24.6	23.8	23.3

Summary of Features of Parametric Curve Models

1. Allows one to model time trend and treatment effect(s) as a function of a small number of parameters. That is, the treatment effect can be captured in one or two parameters, leading to more powerful tests when these models fit the data.

2. Since $E(Y_{ij})$ is defined as an explicit function of the time of measurement, Time_j, there is no reason to require all subjects to have the same set of measurement times, nor even the same number of measurements.

3. May not always be possible to fit data adequately.

Finally, note that the parametric curve analyses in (a)-(c) can all be expressed in terms of the general linear model

$$E\left(\mathbf{Y}_i\right) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

GENERAL LINEAR MODEL FOR PARALLEL GROUPS REPEATED MEASURES DESIGN

In the general linear model formulation, information about treatment group and time of observation will be expressed through a set of covariates rather than through subscripting.

Thus, associated with each vector of responses $\mathbf{Y}_i$, there is a matrix of covariates $\mathbf{X}_i$.

The model for $\mathbf{Y}_i$ can then be written as

$$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

Assumptions

1. The subjects represent random samples from each of the study groups.

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

3. The vector of observations, $\mathbf{Y}_i$ for a given subject has a multivariate normal distribution with

   - mean given by the linear regression model,

$$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

   - covariance matrix, $\boldsymbol{\Sigma}$

4. If observations are missing, they are missing at random (MAR) or missing completely at random (MCAR).

Handling Baseline Measurements

There are a number of possible ways to handle baseline measurements in the analysis.

1. Base analysis on change scores only, say $Y_{ij} - Y_{i1}$ (difference between the response at times 2 through p and the baseline response).

   - This approach has broad intuitive appeal
   - Loss of efficiency
   - May have missing baseline measures

2. Include baseline measure as a covariate; $\mathbf{Y}_i$ is the post-baseline responses only

   - Usually only appropriate for randomized studies
   - More efficient than change score analysis
   - There can be a proliferation of parameters

$$E\left(Y_{ij}|Y_{i1}\right) = \mu_{ij} + \gamma_j Y_{i1}; \quad j = 2, ..., p.$$

3. Include baseline as part of the response vector $\mathbf{Y}_i$, and if appropriate, adjust the model for treatment effect to exclude differences at baseline

   - Usually only appropriate for randomized studies
   - As efficient as analysis that treats baseline as a covariate
   - Can handle missing baseline measures

Issues in Adjusting for Baseline in Non-Randomized Studies

Adjusting for baseline is no substitute for randomization.

Adjusting for baseline in non-randomized studies can potentially lead to erroneous conclusions.

Example 1: Consider an observational study of pulmonary function decline in adults. Suppose that asthmatics have lower pulmonary function at all ages, but that the rates of decline are equal for asthmatics and non-asthmatics.

Suppose the model that describes the data is:

$$Y_{ij} = \beta_0 + \beta_1 \text{Asthma}_i \ + \beta_2 \text{Age}_{ij} + e_{ij}$$

Thus the model for the non-asthmatics is,

$$E(Y_{ij}) = \beta_0 + \beta_2 \text{Age}_{ij}$$

and the model for the asthmatics is,

$$E(Y_{ij}) = (\beta_0 + \beta_1) + \beta_2 \text{Age}_{ij}$$

Clearly, the rate of change or decline, expressed by β_2, 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:

$$Y_{ij} = \beta_0 + \beta_1 \text{Asthma}_i + \beta_2 \text{Age}_{ij} + \beta_3 Y_{i0} + e_{ij}$$

This model corrects the predicted values for asthmatics and non-asthmatics 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, for 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 will 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 expected to increase or regress to the mean level for non-asthmatics

As a result, the rates of decline are not the same in the two groups (given they have the same initial level of pulmonary function).

Example 2: The early studies of Operation Head Start were designed to select children at a similar developmental level for the Head Start and control programs.

Children were selected from low income communities for the Head Start program.

Because randomization was considered impractical, a control group was selected from neighboring communities.

Unfortunately, the children from the neighboring communities had higher income and greater access to educational resources.

The children were matched on initial or baseline developmental status, e.g. by the Bayley Scales of Infant Development.

Because the children from the low income community were selected from the upper end of the distribution of developmental status, their post-test scores tended to regress to the means in their communities.

Similarly, children from the neighboring communities were selected from the lower end of the distribution of developmental status, and the post-test scores in the control group regressed to their means.

Thus, despite the favorable impact of Head Start on development, the initial studies found a larger increase in developmental scores during the study period in the control children.

These results spuriously caused Head Start to appear to be ineffective.

MODELING THE MEAN

We have discussed several models for the mean response for the k group repeated measures design:

1. ***Profile Analysis*:**
   This model may be chosen a priori or used to describe treatment effects when the treatment effects do not have a simple form. The hypothesis of no treatment effect corresponds to the test for no time by treatment interaction and has $(k-1)(p-1)$ d.f.
2. ***Parametric Curves*:**
   These models may be chosen a priori or used descriptively when, for example, the differences in expected response increase linearly with time. In the simplest form, a linear trend in time is assumed and we fit the following model (in SAS notation):

   $$y = \text{time trt time*trt}$$

   where time is treated as a continuous variable.
   In this model, the hypothesis of no longitudinal treatment effect corresponds to the test for no time by treatment interaction and has $(k-1)$ d.f.

This model can be extended in a natural way to include quadratic or cubic trends; alternatively a piecewise linear model can be considered.

Note that these models are nested within the "saturated" model for the mean assumed in profile analysis. As a result, it is possible to test the adequacy of these models.

3. ***The Baseline or 'Constant Effect' Model*:**
   In some studies the exposure or treatment might be expected to cause a shift in the mean response that remains constant across measurement occasions.

   To fit a model describing a treatment effect that is constant over the measurement occasions after baseline, we can create a new variable for time:

$$\begin{aligned} \text{posttime} \;=\; & 0 \text{ if baseline } (\text{time} = 0) \\ & 1 \text{ if post-baseline } (\text{time} > 0) \end{aligned}$$

Then, we can fit the following model (in SAS notation):

$$y = \text{posttime trt posttime*trt}$$

This model tests whether the differences between the treatment group means, averaged over the $(p-1)$ post-baseline time periods, are significantly different from the corresponding differences at baseline.

That is, the hypothesis of no longitudinal treatment effect corresponds to the test of no posttime by trt interaction and has $(k-1)$ d.f.

Note that this model is nested within the saturated model, so that the adequacy of the model relative to the saturated model can be tested.

Also, this model is valid for both randomized and non-randomized studies.

However, when there is randomization, the analysis of covariance test of the constant effect of treatment is more powerful.

To see this, suppose that we have two repeated measures, Y_{i1} and Y_{i2}, with a compound symmetric covariance matrix.

Then, for an analysis based on the change score, $Y_{i2} - Y_{i1}$,

$$\begin{aligned} var\left(Y_{i2} - Y_{i1}\right) &= \sigma^2 - 2\rho\sigma^2 + \sigma^2 \\ &= 2\sigma^2\left(1 - \rho\right). \end{aligned}$$

In contrast, the analysis of covariance (with Y_{i1} treated as a covariate) has residual variance

$$var\left(Y_{i2}|Y_{i1}\right) = \sigma^2\left(1 - \rho^2\right).$$

In this case, the residual variance of the analysis of covariance model is always smaller than the residual variance of the repeated measures (or change score) model.

Consider the ratio of the two variances,

$$\frac{\sigma^2\left(1-\rho^2\right)}{2\sigma^2\left(1-\rho\right)} = \frac{1+\rho}{2}.$$

Thus,

- when $\rho = 1$, the analyses are equally efficient
- when $\rho = 0$, the analysis of the change score is only half as efficient as the analysis of covariance (which in turn is equivalent to an analysis of Y_{i2} only)
- As ρ approaches -1, the relative efficiency of the analysis of the change score approaches zero

SUMMARY

To analyze data from the parallel groups repeated measures design:

1. Choose a "working" covariance structure.

   Note that the choice of model for the mean and covariance are interdependent.

   For designed experiments: assume a saturated treatment-by-time model. Use the REML log likelihood as the criterion to guide choice of covariance structure.

   Ordinarily, use the unstructured model unless p is large and/or a simpler model is clearly satisfactory.

   When p is relatively large and/or there are mistimed measurements, alternative models for the covariance will need to be considered.

2. Decide a priori whether to model effect of treatment on patterns of change by:

   a) time by treatment interaction, where time is regarded as a categorical variable (profile analysis)
   b) time trend(s) by treatment interaction(s), where the means are modelled as an explicit function of continuous time (parametric curves)
   c) treatment effects in an analysis that includes the baseline measure as a covariate
   d) treatment effects in an analysis that includes the baseline measure as part of the response vector but assumes no treatment differences at baseline
   e) post-baseline time (posttime) by treatment interaction 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. Re-fit the final model using REML.

GENERAL LINEAR MODEL FOR LONGITUDINAL DATA

We have stressed that, in fitting linear models to longitudinal data, we have two modeling tasks:

1. We must choose a covariance model that provides a good fit to the observed variances and covariances.

2. We must fit a linear regression model that provides a good fit to the mean of the outcome variable.

So far, the focus has been on balanced designs, where every individual is measured at the same set of occasions.

Next, we consider general linear models that

a) can handle mixed discrete and continuous covariates

b) allow a wider class of covariance structures

c) permit individuals to be measured on different number of occasions and at different times

Later, we will describe how to obtain so-called robust variances that yield valid standard errors when the assumed covariance matrix does not provide a good fit to the observed covariance matrix.

GENERAL LINEAR MODEL

Let $Y_{ij}, j = 1, \ldots, p$, be the sequence of observed measurements for the i^{th} subject, $i = 1, \ldots, n$.

(Later we will relax the assumption that each subject is observed at the same time points).

Information about the time of observation, treatment group, and other predictor variables can be expressed through a vector of covariates.

Each observation, Y_{ij}, has an associated set of covariates $X_{ij0}, X_{ij1}, X_{ij2}, \ldots, X_{ijk-1}$, where typically $X_{ij0} = 1$.

The general linear model for Y_{ij} can be written

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_{k-1} X_{ijk-1} + e_{ij} \\ &= \mathbf{X}_{ij}\boldsymbol{\beta} + e_{ij} \end{aligned}$$

where

$\mathbf{X}'_{ij}$ denotes the $(k \times 1)$ vector of covariates,
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1})'$ is
a $(k \times 1)$ vector of
regression parameters.

With longitudinal data, we expect the e_{ij} to be correlated within individuals. That is,

$$Cov\,(e_{ij}, e_{ik}) \neq 0$$

Since the general linear model for each Y_{ij} is

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_{k-1} X_{ijk-1} + e_{ij} \\ &= \mathbf{X}_{ij}\boldsymbol{\beta} + e_{ij} \end{aligned}$$

if we introduce additional vector and matrix notation, the general linear model for the set (or vector) of responses, $\mathbf{Y}_i$, can be written

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i$$

where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$ denotes the $(p \times 1)$ vector of responses, and

$$\mathbf{X}_i = \begin{bmatrix} 1 & X_{i11} & X_{i12} & \ldots & X_{i1,k-1} \\ 1 & X_{i21} & X_{i22} & \ldots & X_{i2,k-1} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & X_{ip1} & X_{ip2} & \ldots & X_{ip,k-1} \end{bmatrix}$$

denotes the $(p \times k)$ matrix of covariates.

Assumptions:

1. The individuals are a random sample from the population of interest.

2. The values of the dependent variable have a multivariate normal distribution, with mean $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$.

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

4. If there are missing data they are assumed to be 'ignorable', i.e. MAR or MCAR.

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 $\boldsymbol{\beta}$

$\Rightarrow$ weaker inferences concerning $\boldsymbol{\beta}$.

With too much structure (e.g compound symmetry), there is more information available for estimating $\boldsymbol{\beta}$. However, there is a potential risk of model misspecification

$\Rightarrow$ apparently stronger, but potentially biased, inferences concerning $\boldsymbol{\beta}$.
Thus far, we have encountered three covariance structures:

1) independence

2) compound symmetry

3) unstructured

Next, we consider a number of additional covariance models suitable for longitudinal data.

Autoregressive: AR(1)

The first-order autoregressive model has covariances of the form:

$$\sigma_{jk} = \sigma^2 \rho^{|j-k|}$$

For example, with 4 occasions of measurement, the AR(1) covariance matrix is as follows:

$$\begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix}$$

Note that it has homogeneous variances and correlations that decline over time.

This structure is best suited to equally-spaced measurements.

Theoretical Justification:

The AR(1) covariance structure arises when the errors, e_{ij} are thought of as coming from the following autoregressive process:

$$e_{ij} = \rho e_{ij-1} + w_{ij}$$

where $w_{ij} \sim N\left(0, \sigma^2\left\{1-\rho^2\right\}\right)$.

The process is initiated by $e_{i0} \sim N\left(0, \sigma^2\right)$.

Then,

$$Var\left(e_{ij}\right) = \sigma^2$$

and

$$Cov\left(e_{ij}, e_{ik}\right) = \sigma^2 \rho^{|j-k|}$$

The generic SAS code for fitting the AR(1) covariance model is as follows:

```
proc mixed;
      class id trt time t;
      model y=trt time time*trt / s chisq;
      repeated t / type=ar(1) subject=id r rcorr;
```

Note: If the variances are not homogeneous, we can consider a generalization of AR(1) that has the same correlation structure but allows for heterogeneous variances: ARH(1):

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$$

```
proc mixed;
      class id trt time t;
      model y=trt time time*trt / s chisq;
      repeated t / type=arh(1) subject=id r rcorr;
```

Example: *Exercise Therapy Study*

Subjects in an exercise therapy study were assigned to one of two weightlifting programs.

In the first program (treatment 1), the number of repetitions was increased as subjects became stronger. In the second program (treatment 2), the amount of weight was increased as subjects became stronger.

For illustration, we focus only on measures of strength taken at baseline (day 0) and on days 4, 6, 8, and 12.

```
data stren;
      infile 'g:\shared\bio226\strentwo.dat';
      input id trt t1 t2 t3 t4 t5 t6 t7;
      time=0; y=t1; t=1; output;
      time=4; y=t3; t=2; output;
      time=6; y=t4; t=3; output;
      time=8; y=t5; t=4; output;
      time=12; y=t7; t=5; output;
run;

proc mixed data = stren;
      class id trt time t;
      model y=trt time trt*time / s chisq;
      repeated t / type=un subject=id r rcorr;
run;

proc mixed data = stren;
      class id trt time t;
      model y=trt time trt*time / s chisq;
      repeated t / type=ar(1) subject=id r rcorr;
run;
```

Unstructured Covariance Model:

Estimated R Matrix for id 1

Row	col1	col2	col3	col4	col5
1	9.6683	10.1752	8.9741	9.8125	9.4070
2	10.1752	12.5501	11.0912	12.5801	11.9284
3	8.9741	11.0912	10.6417	11.6857	11.1007
4	9.8125	12.5801	11.686	13.9905	13.1213
5	9.4070	11.9284	11.1017	13.1213	13.9444

Estimated R Correlation Matrix for id 1

Row	col1	col2	col3	col4	col5
1	1.0000	0.9237	0.8847	0.8437	0.8102
2	0.9237	1.0000	0.9597	0.9494	0.9017
3	0.8847	0.9597	1.0000	0.9577	0.9113
4	0.8437	0.9494	0.9577	1.0000	0.9394
5	0.8102	0.9017	0.9113	0.9394	1.0000

Fit Statistics

-2 Res Log Likelihood	597.3
AIC (smaller is better)	627.3
AICC (smaller is better)	630.6
BIC (smaller is better)	651.5

AR(1) Covariance Model:

Estimated R Matrix for id 1

Row	col1	col2	col3	col4	col5
1	11.8673	11.1573	10.4899	9.8623	9.2723
2	11.1573	11.8673	11.1573	10.4899	9.8623
3	10.4899	11.1573	11.8673	11.1573	10.4899
4	9.8623	10.4899	11.1573	11.8673	11.1573
5	9.2723	9.8623	10.4899	11.1573	11.8673

Estimated R Correlation Matrix for id 1

Row	col1	col2	col3	col4	col5
1	1.0000	0.9402	0.8839	0.8311	0.7813
2	0.9402	1.0000	0.9402	0.8839	0.8311
3	0.8839	0.9402	1.0000	0.9402	0.8839
4	0.8311	0.8839	0.9402	1.0000	0.9402
5	0.7813	0.8311	0.8839	0.9402	1.0000

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
AR(1)	id	0.9402
Residual		11.8673

Fit Statistics

-2 Res Log Likelihood	621.1
AIC (smaller is better)	625.1
AICC (smaller is better)	625.1
BIC (smaller is better)	628.3

Test for AR(1) versus Unstructured Covariance:

	-2 Res Log L
AR(1)	621.1
UN	597.3

$\Rightarrow$ -2*Res log likelihood ratio = 23.8, 13 d.f.

($p \approx 0.036$)

AR(1) model is not defensible.

When the measurements are unequally spaced over time, with the measurement occasions being at times t_j, the exponentially decreasing correlation form can be incorporated by taking

$$\sigma_{jk} = \sigma^2 \rho^{|t_j - t_k|}$$

Note: This form is invariant under linear transformation of the time scale. That is, if we replace t_j by $(a + bt_j)$, the same form for the covariance matrix holds.

This is sometimes referred to as the 'exponential correlation model', since

$$\sigma_{jk} = \sigma^2 \rho^{|t_j - t_k|} = \sigma^2 \exp\left(-\theta \left|t_j - t_k\right|\right)$$

where

$$\rho = \exp\left(-\theta\right) \quad (\text{for } \theta > 0)$$

The generic SAS code for fitting the 'exponential correlation model' is as follows:

```
proc mixed;
      class id trt time t;
      model y=trt time time*trt / s chisq;
      repeated t / type=sp(exp)(ctime)
            subject=id r rcorr;
```

The option:

$$\text{ype=sp(exp)}(ctime)$$

is used to specify the exponential correlation structure, with *ctime* as the variable used to calculate the time-separation between the measurement occasions.

Example: *Exercise Therapy Study*

Recall that the 5 measurement occasions (0, 4, 6, 8, and 12) are unequally spaced.

```
data stren;
      infile 'g:\shared\bio226\strentwo.dat';
      input id trt t1 t2 t3 t4 t5 t6 t7;
      time=0; ctime=0; y=t1; t=1; output;
      time=4; ctime=4; y=t3; t=2; output;
      time=6; ctime=6; y=t4; t=3; output;
      time=8; ctime=8; y=t5; t=4; output;
      time=12; ctime=12; y=t7; t=5; output;
run;

proc mixed data = stren;
      class id trt time t;
      model y=trt time trt*time / s chisq;
      repeated t / type=sp(exp)(ctime) subject=id r rcorr;
run;
```

Exponential Covariance Model:

Estimated R Matrix for id 1

Row	col1	col2	col3	col4	col5
1	11.8738	10.8875	10.4255	9.9832	9.1539
2	10.8875	11.8738	11.3700	10.8875	9.9832
3	10.4255	11.3700	11.8738	11.3700	10.4255
4	9.9832	10.8875	11.3700	11.8738	10.8875
5	9.1539	9.9832	10.4255	10.8875	11.8738

Estimated R Correlation Matrix for id 1

Row	col1	col2	col3	col4	col5
1	1.0000	0.9169	0.8780	0.8408	0.7709
2	0.9169	1.0000	0.9576	0.9169	0.8408
3	0.8780	0.9576	1.0000	0.9576	0.8780
4	0.8408	0.9169	0.9576	1.0000	0.9169
5	0.7709	0.8408	0.8780	0.9169	1.0000

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
SP(EXP)	id	46.1262
Residual		11.8738

Note: SAS estimates $1/\theta$ rather than θ. Thus, estimate of $\rho = \exp\{-1/(46.13)\} = 0.97856$

Fit Statistics

-2 Res Log Likelihood	618.5
AIC (smaller is better)	622.5
AICC (smaller is better)	622.6
BIC (smaller is better)	625.8

Test for Exponential versus Unstructured Covariance:

	-2 Res Log L
EXP	618.5
UN	597.3

$\Rightarrow$ -2*Res log likelihood ratio = 21.2, 13 d.f.

($p \approx 0.086$)

Exponential correlation model is defensible.

Results for Exponential Correlation Model:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
trt	1	35	1.70	0.1977
time	4	128	28.18	<.0001
trt*time	4	128	3.57	0.4679

We cannot reject the null hypothesis of no treatment by time interaction.
$\Longrightarrow$ profiles of means are similar in the two groups.

GENERAL LINEAR MIXED EFFECTS MODEL

Next, we consider mixed models for longitudinal data.

Note: A mixed model is one that contains both fixed and random effects.

Mixed models for longitudinal data explicitly identify individual (random effects) and population characteristics (fixed effects).

Mixed models are very flexible since they can accommodate any degree of imbalance in the data. That is, we do not necessarily require the same number of observations on each subject or that the measurements be taken at the same times.

Also, the use of random effects allows us to model the covariance structure as a continuous function of time.

Recall: Compound Symmetry Model

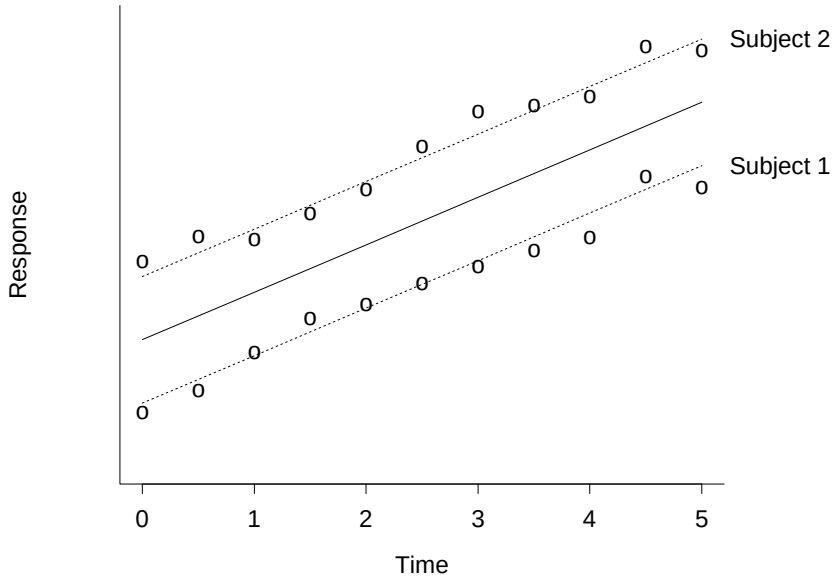
Assumes the correlation between repeated measurements arises because each subject has an underlying level of response which persists over time.

This subject effect is treated as random and the mixed model is

$$\begin{aligned} Y_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta} + b_i + e_{ij} \\ &= (\beta_0 + b_i) + \beta_1 X_{ij1} + \ldots \beta_p X_{ijp} + e_{ij} \end{aligned}$$

The response for the i^{th} subject is assumed to differ from the population mean, $\mathbf{X}_{ij}\boldsymbol{\beta}$, by a subject effect, b_i, and a within-subject measurement error, e_{ij}.

Compound Symmetry (or Random Intercepts) Model:



If $var\,(b_i) = \sigma_b^2$ and $var\,(e_{ij}) = \sigma^2$ the covariance matrix of the repeated measurements has the compound symmetry form:

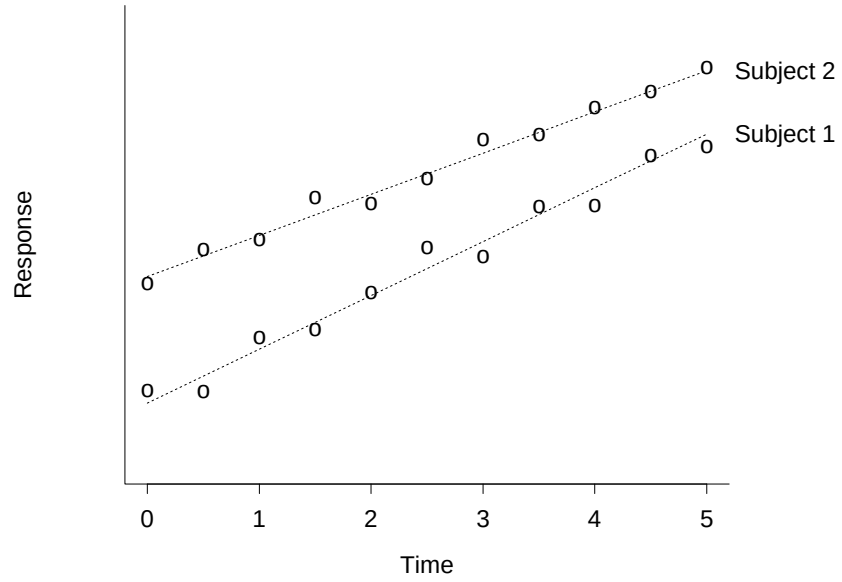
$$\begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 & \dots & \sigma_b^2 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma^2 \end{bmatrix}$$

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 model.

However, we can easily generalize these ideas.

Random Intercepts and Slopes Model:



Random Intercepts and Slopes Model:

Consider the following model with intercepts and slopes that vary randomly among subjects.

For the i^{th} subject at the j^{th} measurement occasion

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + e_{ij}$$

(Note: we are using double subscripting here)

Linear Mixed Model

Notation: Suppose we have n individuals on which we have collected p_i repeated observations at times t_{ij}.

Consider the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

where $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of fixed effects;

$\mathbf{b}_i$ is a $(q \times 1)$ vector of random effects and
$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$;

$\mathbf{X}_i$ is a $(p_i \times k)$ matrix of covariates;

$\mathbf{Z}_i$ is a $(p_i \times q)$ matrix of covariates (usually the columns of $\mathbf{Z}_i$ are a subset of the columns of $\mathbf{X}_i$ and $q < k$);

$\mathbf{e}_i$ is a $(p_i \times 1)$ vector of errors and $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$.

Example

Consider again the random intercepts and slopes model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + e_{ij}$$

In matrix form this can be represented as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i$$

where

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ . & . \\ . & . \\ 1 & t_{ip_i} \end{bmatrix}$$

Let $var\,(b_{i0}) = g_{11}$, $var\,(b_{i1}) = g_{22}$, and $cov\,(b_{i0}, b_{i1}) = g_{12}$.

These are the three unique elements of the (2×2) covariance matrix $\mathbf{G}$.

We assume that $var\,(e_{ij}) = \sigma^2$. Thus, $\mathbf{R}_i = \sigma^2\mathbf{I}$.

Then it can be shown that

$$var\,(Y_{ij}) = g_{11} + 2t_{ij}g_{12} + g_{22}t_{ij}^2 + \sigma^2$$

and

$$cov\,(Y_{ij}, Y_{ik}) = g_{11} + (t_{ij} + t_{ik})\,g_{12} + g_{22}t_{ij}t_{ik}$$

Note: Covariance is expressed as a function of time.

Covariate effects (e.g. due to treatment) can be expressed by allowing the mean values of the intercept and slope to depend upon the covariates (e.g. by allowing them to differ across the treatment groups).

In the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

$\mathbf{R}_i = var\left(\mathbf{e}_i\right)$ describes the covariance among observations when we focus on the response profile of a specific individual.

That is, it is the covariance of the i^{th} subject's deviations from his/her mean profile $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$.

Usually, it is assumed that $\mathbf{R}_i = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is a $(p_i \times p_i)$ identity matrix

$\Rightarrow$ 'conditional independence assumption'

Alternatively, a structured model for $\mathbf{R}_i$ could be assumed, e.g. AR(1).

In the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

The vector of regression parameters $\boldsymbol{\beta}$ are the *fixed effects*, which are assumed to be the same for all individuals.

These regression parameters have population-averaged interpretation (e.g. in terms of changes in the mean response, averaged over individuals).

Although the *conditional* mean of $\mathbf{Y}_i$, given $\mathbf{b}_i$, is

$$E\left(\mathbf{Y}_i|\mathbf{b}_i\right) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

note that the *marginal* or population-averaged mean of $\mathbf{Y}_i$ is

$$E\left(\mathbf{Y}_i\right) = \mathbf{X}_i\boldsymbol{\beta}$$

In contrast to $\boldsymbol{\beta}$, the vector $\mathbf{b}_i$ is comprised of subject-specific regression coefficients.

These are the *random effects* and the $\mathbf{b}_i$ have a distribution (usually, but not necessarily, assumed to be normal).

Combined with the fixed effects, these describe the mean response profile of a specific individual.

That is, the mean response profile for the i^{th} individual is

$$\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i.$$

In the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

recall that

$$E\left(\mathbf{Y}_i|\mathbf{b}_i\right) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

and

$$E\left(\mathbf{Y}_i\right) = \mathbf{X}_i\boldsymbol{\beta}.$$

Similarly,

$$var\left(\mathbf{Y}_i|\mathbf{b}_i\right) = var\left(\mathbf{e}_i\right) = \mathbf{R}_i$$

and

$$\begin{aligned} var\left(\mathbf{Y}_i\right) &= var\left(\mathbf{Z}_i\mathbf{b}_i\right) + var\left(\mathbf{e}_i\right) \\ &= \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i \end{aligned}$$

Of note, even if $\mathbf{R}_i = \sigma^2\mathbf{I}$,

$$var\left(\mathbf{Y}_i\right) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \sigma^2\mathbf{I}$$

is <u>not</u> a diagonal matrix.

Thus, the introduction of random effects, $\mathbf{b}_i$, induces correlation (marginally) among the $\mathbf{Y}_i$.

That is,

$$var\left(\mathbf{Y}_i\right) = \mathbf{\Sigma}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$$

which, in general, has non-zero off-diagonal elements.

Finally, note that $var\left(\mathbf{Y}_i\right)$ is described in terms of a set of covariance parameters, some defining $\mathbf{G}$ and some defining $\mathbf{R}_i$.

Example: *Exercise Therapy Study*

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 SAS code:

```
proc mixed data = stren;
      class id trt;
      model y=trt time time*trt / s chisq;
      random intercept time / type=un sub=id g;
```

Random Intercepts and Slopes Model:

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

Effect	trt	Estimate	Standard 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:

$$\begin{aligned} var\left(\mathbf{Y}_i\right) &= var\left(\mathbf{Z}_i\mathbf{b}_i\right) + var\left(\mathbf{e}_i\right) \\ &= \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i \end{aligned}$$

Given estimates of $\mathbf{G}$:

$$\begin{bmatrix} 9.54695 & 0.05331 \\ 0.05331 & 0.02665 \end{bmatrix}$$

and of $\mathbf{R}_i = \sigma^2\mathbf{I}$: $(0.6862)\mathbf{I}$,

and with

$$\mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 12 \end{bmatrix}$$

We can obtain the following estimate of $var\left(\mathbf{Y}_i\right)$:

$$\begin{bmatrix} 10.23 & 9.76 & 9.87 & 9.97 & 10.19 \\ 9.76 & 11.09 & 10.72 & 11.04 & 11.68 \\ 9.87 & 10.72 & 11.83 & 11.57 & 12.43 \\ 9.97 & 11.04 & 11.57 & 12.79 & 13.17 \\ 10.19 & 11.68 & 12.43 & 13.17 & 15.35 \end{bmatrix}$$

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 option in PROC MIXED:
random intercept time / type=un sub=id g v vcorr;

Next, consider the model with random intercepts only (equivalent to compound symmetry).

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 / type=un sub=id g;
```

Alternatively, we could fit this model by specifying a compound symmetry model for $\mathbf{R}_i$ and assume no random effects:

```
proc mixed data = stren;
      class id trt t;
      model y=trt time time*trt / s chisq;
      repeated t / type=cs sub=id r;
```

Random Intercepts Model:

Estimated G Matrix

Effect	id	col1
Intercept	1	10.8506

Residual: 1.1579

Fit Statistics

-2 Res Log Likelihood	660.4
AIC (smaller is better)	664.4
AICC (smaller is better)	664.5
BIC (smaller is better)	667.6

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		81.2895	0.7445	35	109.18	<.0001
trt	1	-1.2805	1.1314	134	-1.13	0.2598
trt	2	0	.	.	.	.
time		0.1605	0.02887	134	5.56	<.0001
time*trt	1	-0.0267	0.04212	134	-0.63	0.5274
time*trt	2	0	.	.	.	.

However, in current setting,

	-2 Res Log L
Random Intercepts	660.4
Random Intercepts & Slopes	632.0

$\Rightarrow$ -2*Res log likelihood ratio = 28.4, 2 d.f. ($p < 0.0001$)

So, in this case, there is no doubt that the Random Intercepts (or compound symmetry) model is not defensible.

We will revisit this test later when we discuss AIC.

Prediction of Random Effects

In most applications, inference is focused on the fixed effects, $\boldsymbol{\beta}$.

However, in some studies we may want to predict (or "estimate") subject-specific response profiles.

Technically, because the $\mathbf{b}_i$ are random, we customarily talk of "predicting" the random effects rather than "estimating" them.

Using maximum likelihood, the prediction of $\mathbf{b}_i$, say $\widehat{\mathbf{b}}_i$, is given by:

$$\mathbf{G}\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma}_i = var\,(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$.

Aside: This is known as the Best Linear Unbiased Predictor (or BLUP).

When the unknown covariance parameters are replaced by their ML or REML estimates, the resulting predictor,

$$\widehat{\mathbf{b}}_i = \widehat{\mathbf{G}}\mathbf{Z}_i'\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}),$$

is often referred to as the "Empirical BLUP" or the "Empirical Bayes" (EB) estimator.

Furthermore, it can be shown that

$$var(\widehat{\mathbf{b}}_i) = \mathbf{G}\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\mathbf{G} - \mathbf{G}\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i(\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\mathbf{G}$$

Finally, the i^{th} subject's predicted response profile is,

$$\begin{aligned}
\widehat{\mathbf{Y}}_i &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i \\
&= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{G}}\mathbf{Z}_i'\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \\
&= (\widehat{\mathbf{R}}_i\widehat{\boldsymbol{\Sigma}}_i^{-1})\mathbf{X}_i\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \widehat{\mathbf{R}}_i\widehat{\boldsymbol{\Sigma}}_i^{-1})\mathbf{Y}_i
\end{aligned}$$

That is, the i^{th} subject's predicted response profile is a weighted combination of the population-averaged mean response profile, $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$, and the i^{th} subject's observed response profile $\mathbf{Y}_i$.

Note that the subject's predicted response profile is "shrunk" towards the population-averaged mean response profile.

The amount of "shrinkage" depends on the relative magnitude of $\mathbf{R}_i$ and $\boldsymbol{\Sigma}_i$.

Note that $\mathbf{R}_i$ characterizes the within-subject variability, while $\boldsymbol{\Sigma}_i$ incorporates both within-subject and between-subject sources of variability.

As a result, when $\mathbf{R}_i$ is "large", and the within-subject variability is greater than the between-subject variability, more weight is given to $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$, the population-averaged mean response profile.

When the between-subject variability is greater than the within-subject variability, more weight is given to the i^{th} subject's observed data $\mathbf{Y}_i$.

SAS CODE

The Empirical Bayes (EB) estimates, $\widehat{\mathbf{b}}_i$, can be obtained by using the following option on the RANDOM statement in PROC MIXED:

random intercept time / type=un sub=id s;

Alternatively, a subject's predicted response profile,

$$\widehat{\mathbf{Y}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i,$$

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 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 SAS code:

```
proc mixed data = stren;
      class id trt;
      model y=trt time time*trt / s chisq;
      random intercept time / type=un sub=id g s;
```

Empirical Bayes Estimates of $\mathbf{b}_i$:

Solution for Random Effects

Effect	id	Estimate	Std Err Pred	t Value	Pr > |t|
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 sub=id g s;

proc print data = predict;
      var id trt time y Pred StdErrPred Resid;
```

Predicted Response Profiles

id	trt	time	y	Pred	StdErr 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
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

GROWTH CURVE MODELS

In this lecture we discuss growth curve models. These are simply random coefficient (e.g. random intercepts and slopes) models that also allow for the possibility that subjects may be drawn from different groups.

As we shall see later, growth curve models are simply a special case of the mixed effects models.

In order to motivate the methods, 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.

Two-Stage Model

Growth curve models can be motivated in terms of a two-stage model. In the two-stage 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**)

More formally, if Y_{ij} is the response of the i^{th} individual measured at time t_{ij}, we assume

Stage 1:

$$Y_{ij} = v_{i1} + v_{i2}t_{ij} + e_{ij}$$

where v_{i1} and v_{i2} are parameters specific to the i^{th} subject and the errors, e_{ij}, are implicitly assumed to be independent within a subject.

Stage 2: In the second stage, the intercepts and the slopes are regressed on other covariates:

$$v_{i1} = \alpha_1 + \mathbf{X}_i\boldsymbol{\beta}_1 + \varepsilon_{i1}$$
$$v_{i2} = \alpha_2 + \mathbf{X}_i\boldsymbol{\beta}_2 + \varepsilon_{i2}$$

Two-Stage Analysis - "NIH Method"

One classic approach to the analysis of such 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 the two-stage method, we simply fit a straight line (or curve) to the response data for each subject (first stage), and then regress the estimates of the individual intercepts and slopes on subject-specific covariates (second stage).

One of the attractions of this method is that it is very easy to perform using existing statistical software for linear regression.

We can illustrate the method by considering a two-stage analysis of Feldman's clearance data.

STRUCTURE OF THE DATASET

ORGAN	ID	DAYS	CFP	LOGCFP
lung	1	3	102	2.00860
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

SAS CODE

```
filename rats 'g:\shared\bio226\rat.dat';

data rats;
      infile rats;
      input organ $ id days cfp logcfp;
      if (days=0) then delete;
run;
```

Two-Stage Analysis

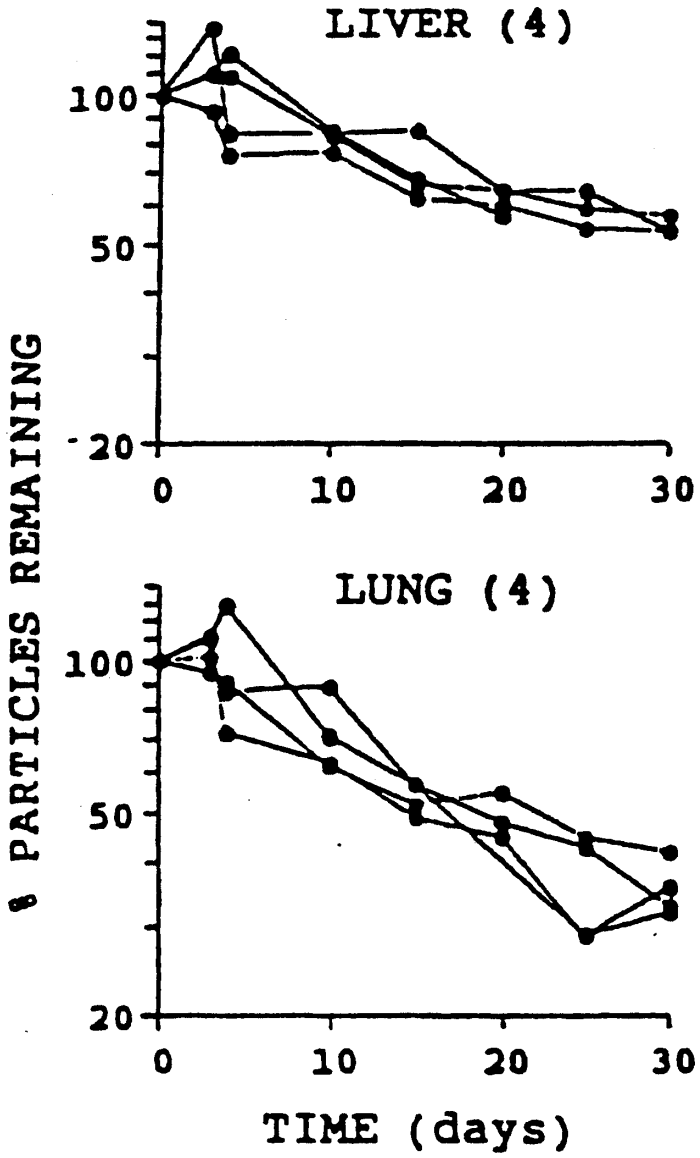
Stage 1:

```
proc reg data=rats 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


LIVER (4)
100
50
20
0
10
20
30
% PARTICLES REMAINING
LUNG (4)
100
50
20
0
10
20
30
TIME (days)

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

Source	DF	Sum of Squares	Mean 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			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.037476922	0.03533167	57.67	<.0001
Organ liver	0.010363771	0.04996652	0.21	0.8425
Organ lung	0.000000000	.	.	.

ANOVA for the Slopes

General Linear Models Procedure

Dependent Variable: days

Source	DF	Sum of Squares	Mean 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			

Parameter	Estimate	Standard 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 of Growth Curve Model

We can develop a mixed effects model in two stages corresponding to the two-stage model:

Stage 1: $$Y_{ij} = v_{i1} + v_{i2} t_{ij} + e_{ij}$$

where v_{i1} is the intercept for the i^{th} subject, v_{i2} is the slope for the i^{th} subject, and errors, e_{ij}, are assumed to be independent and normally distributed around the individual's regression line, that is, $e_{ij} \sim N\left(0, \sigma^2\right)$.

Stage 2:

Assume that the intercept and slope, v_{i1} and v_{i2}, are random and have a joint multivariate normal distribution, with mean dependent on covariates (e.g. the organ studied):

$$v_{i1} = \beta_0 + \beta_1 \text{ Organ } + \varepsilon_{i1}$$

$$v_{i2} = \beta_2 + \beta_3 \text{ Organ } + \varepsilon_{i2}$$

Also, let $var\left(\varepsilon_{i1}\right) = g_{11}$, $cov\left(\varepsilon_{i1}, \varepsilon_{i2}\right) = g_{12}$, $var\left(\varepsilon_{i2}\right) = g_{22}$. Typically, it is assumed that the variances and covariance do not depend on covariates.

If we substitute the expressions for v_{i1} and v_{i2} into the equation in stage 1, we obtain

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 \text{ Organ } + \beta_2 t_{ij} + \beta_3 \text{ Organ } \times t_{ij} \\ &\quad + \varepsilon_{i1} + \varepsilon_{i2} t_{ij} + e_{ij} \end{aligned}$$

The first four terms give the regression model implied by the two-stage model.

The covariates are organ, days, and organ*days.

The last three terms are the error terms in the growth curve model, and it can be shown that

$$\begin{aligned} var\left(Y_{ij}\right) &= g_{11} + 2t_{ij}g_{12} + t_{ij}^2 g_{22} + \sigma^2 \\ &= \left(1 \; t_{ij}\right) \mathbf{G} \left(1 \; t_{ij}\right)' + \sigma^2 \mathbf{I} \end{aligned}$$

This model can be fit using the ***random*** statement in PROC MIXED.

SAS CODE FOR GROWTH CURVE MODEL

```
filename rats 'g:\shared\bio226\rat.dat';

data rats;
      infile rats;
      input organ $ id days cfp logcfp;
      if (days=0) then delete;
run;

proc mixed data = rats;
      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

Fit Statistics

-2 Res Log Likelihood	-111.2
AIC (smaller is better)	-103.2
AICC (smaller is better)	-102.3
BIC (smaller is better)	-102.9

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

Results suggest that mean clearance of foreign particles is faster from the lung.

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 columns of Z_i are a subset of the columns of X_i.

SELECTION OF MODEL FOR COVARIANCE

For a given linear model, how can we decide which model for the covariance to use in the 'final' analysis?

There are two general approaches for comparing models for the covariance matrix:

1. Restricted ML (REML) when the models are *nested*.

2. *Information criteria* when they are not *nested*:

   - Akaike's Information Criterion (AIC)
   - Schwarz's Bayesian Information Criterion (BIC)

Comparing Nested Models for the Covariance

The REML likelihood provides a measure of the goodness of fit of an assumed model for the covariance.

A standard approach for comparing two nested models is via the likelihood-ratio test.

Take twice the difference in maximized log likelihoods and compare to the chi-squared distribution (with df equal to the difference in number of covariance parameters).

In many settings the likelihood-ratio test is a valid method for comparing nested model.

However, here it may not always be valid due to the nature of the null hypothesis.

Technically, the reason for the problem is that the LRT may be testing a null hypothesis that is "on the boundary of the parameter space" (e.g., testing that a variance component is zero).

As a result, the usual conditions required for classical likelihood theory are no longer met.

As a consequence, the usual null distribution for the LRT may no longer be valid.

Instead, the null distribution for the LRT may be a mixture of chi-squared distributions.

Illustration

Suppose: $Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{i0} + \epsilon_{ij}$; $\text{var}(b_{i0}) = g_{11}$.
To test $H_0 : g_{11} = 0$ versus $H_A : g_{11} > 0$, the asymptotic null distribution of the standard LRT is <u>not</u> a chi-squared distribution with 1 df.

Instead, it is an equally-weighted mixture of chi-squared distributions with 0 and 1 df.

Similarly, to test
H_0: random intercepts model
H_0: random intercepts and slopes model,

the asymptotic null distribution of the standard LRT is <u>not</u> a chi-squared distribution with 2 df.

Instead, it is an equally-weighted mixture of chi-squared distributions with 1 and 2 df.

Note that if the classical null distribution is used instead, the resulting p-value will be overestimated $\Longrightarrow$ Model selected for covariance is too parsimonious.

What is the lesson to be learned here?

Comparing nested models for covariance can be a non-standard problem.

The reason is that the null hypothesis is often on the boundary of parameter space.

As a consequence, the usual null distributions may no longer be valid.

If the usual null distribution is used the resulting p-value will be overestimated.

Thus, in general, ignoring this problem can lead to selection of model for covariance that is too simple.

Comparing Non-Nested, or Non-standard Nested, Models for the Covariance

For non-standard nested comparisons or when the models are non-nested, they can be compared in terms of *Information Criteria* that effectively extract a penalty for the estimation of an additional parameter.

The two most widely used criteria are Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC).

Akaike's Information Criterion (AIC) is defined as

$$\text{AIC} = \text{Log } L - c$$

where log L is either the maximized ML or REML log likelihood and c is the number of covariance parameters.

With this definition of AIC, it can be used to compare models with the same fixed effects but different covariance structures.

The model with the largest AIC is deemed best.

Schwarz's Bayesian Information Criterion (BIC) is defined as

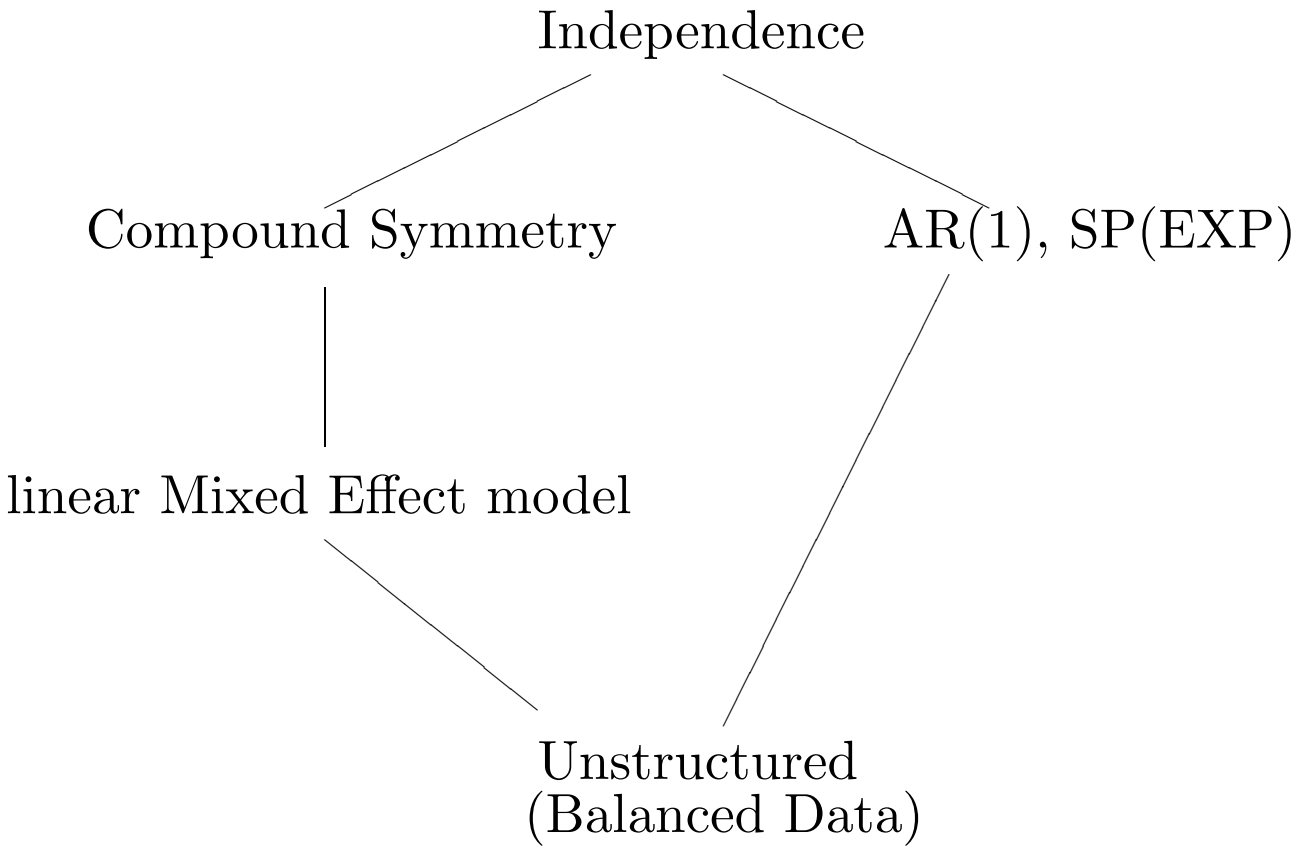
$$\text{BIC} = \text{Log } L - \tfrac{c}{2} \ln (n^*)$$

where log L is either the maximized ML or REML log likelihood, c is the number of covariance parameters, and n^* is the number of effective subjects, n, in the case of ML and $n - k$ in the case of REML estimation.

When $n*$ is relatively large, BIC extracts a substantial penalty for the estimation of each additional parameter.

In general, comparing non-nested models using BIC entails a high risk of selecting a model that is too simple for the data at hand.

Note that the models for the covariance structure have the following hierarchical relationship:



Thus, AR(1) and SP(EXP) have hierarchical relationships to independence and unstructured models, but not to the CS or mixed effects models.

Since the exponential correlation model and the random intercepts and slopes model do not have a hierarchical relationship (i.e. they are not *nested* models), they cannot be compared in terms of a likelihood ratio test.

Instead we can compare the models in terms of their AIC.

Note: SAS prints out a slightly different definition of AIC to make it comparable to -2 Res Log L:

$$\text{AIC} = -2\text{Log}L + 2c$$

and reminds us of this by printing "(smaller is better)".

Example: *Exercise Therapy Study*

Consider a model for the mean that assumes a linear trend, a treatment effect, and their interaction. We can fit a variety of models for the covariance. SAS gives the following AIC's (smaller is better):

Model	**AIC**
CS (Random Intercepts)	664.4
AR(1)	639.7
SP(EXP)	636.7
Random Intercepts and Slopes	640.0
Unstructured	642.2

Likelihood ratio tests (versus unstructured covariance) suggest using either the exponential correlation or the random intercepts and slopes model.

On the basis of AIC, the exponential correlation model is the preferred model.

EMPIRICAL VARIANCE ESTIMATION

We have focussed on regression models for longitudinal data where the primary interest is in making inference about the regression parameters $\boldsymbol{\beta}$.

For statistical inference about $\boldsymbol{\beta}$ we need

(i) an estimate, $\widehat{\boldsymbol{\beta}}$

(ii) estimated standard error, se($\widehat{\boldsymbol{\beta}}$)

So far, we have made inferences about $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}_i\right)$$

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

It has variance matrix,

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}var\left(\mathbf{Y}_i\right)\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}$$

If $\text{var}(\mathbf{Y}_i)$ is replaced by $\widehat{\boldsymbol{\Sigma}}$, the REML estimate of $\boldsymbol{\Sigma}$, $\text{var}(\widehat{\boldsymbol{\beta}})$ can be estimated by

$$\left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}$$

However, if the covariance has been mis-specified then an alternative estimator for $\text{var}\left(\mathbf{Y}_i\right)$ is needed.

The empirical or so-called robust variance of $\widehat{\boldsymbol{\beta}}$ is obtained by using

$$\widehat{\mathbf{V}}_i = \left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'$$

as an estimate of $\text{var}\left(\mathbf{Y}_i\right)$.

Thus, the empirical variance of $\widehat{\boldsymbol{\beta}}$ is estimated by

$$\left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{V}}_i\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\left[\sum_{i=1}^{n}\left(\mathbf{X}_i'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right)\right]^{-1}$$

This empirical variance estimator is also known as the 'sandwich estimator'.

The remarkable thing about the empirical estimator of $var(\widehat{\boldsymbol{\beta}})$ 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 p (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 the potential correlation among repeated measures (i.e., put head in sand) and assume an independence 'working' covariance matrix. Then, obtain correct standard errors using the empirical variance estimator.

Why should we bother to explicitly model the covariance?

Reasons:

1. Efficiency: The optimal (most precise) estimator of $\boldsymbol{\beta}$ uses the true $var\left(\mathbf{Y}_i\right)$. Given sufficient data, we can attempt to estimate $var\left(\mathbf{Y}_i\right)$.
2. When n (number of individuals) is not large relative to p (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.

Example: *Exercise Therapy Study*

Recall that the 5 measurement occasions (0, 4, 6, 8, and 12) are unequally spaced.

The SAS code for fitting the 'exponential correlation model' with empirical variances is as follows:

```
proc mixed data=stren empirical;
      class id trt t;
      model y=trt time trt*time / s chisq;
      repeated t / type=sp(exp)(ctime)
            subject=id r rcorr;
run;
```

Exponential Covariance Model with Empirical Variances

Fit Statistics

-2 Res Log Likelihood	632.7
AIC (smaller is better)	636.7
AICC (smaller is better)	636.8
BIC (smaller is better)	639.9

Note: Estimated covariance parameters and fixed effects (and Fit Statistics) will be identical to the analysis without empirical variances.

Solution for Fixed Effects

Effect	trt	Estimate	Standard Error	t Value	Pr > |t|
Intercept		81.0709	0.6624	122.39	<.0001
trt	1	-1.3425	1.0043	-1.34	0.1899
time		0.1694	0.0359	4.72	<.0001
time*trt	1	-0.0330	0.0625	-0.53	0.5976

Comparing model-based and empirical standard errors

Effect	Estimate	Model-based Std Error	Empirical Std Error
INTERCEPT	81.0709	0.7542	0.6624
TRT	-1.3425	1.1468	1.0043
TIME	0.1694	0.0456	0.0359
TIME*TRT	-0.0330	0.0676	0.0625

There are some discernible differences between the model-based and empirical standard errors.

These differences suggest that the exponential correlation model may not be the best possible approximation to $var\left(\mathbf{Y}_i\right)$.

MISSING DATA AND DROPOUT

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

Let $\mathbf{Y}$ denote the complete response vector which can be partitioned into two sub-vectors:

(i) $\mathbf{Y}^{(o)}$ the measurements observed

(ii) $\mathbf{Y}^{(m)}$ the measurements that are missing

If there were no missing data, we would have observed the complete response vector $\mathbf{Y}$.

Instead, we get to observe $\mathbf{Y}^{(o)}$.

The main problem that arises with missing data is that the distribution of the observed data may not be the same as the 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 for measurement at 12 months because they are not well (say, values of Y_2 are low), then the distribution of observed Y_2's will be positively skewed compared to the distribution of Y_2's in the population of interest.

In general, the situation may often be quite complex, with some missingness unrelated to either the observed or unobserved response, some related to the observed, some related to the unobserved, and some to both.

A particular pattern of missingness that is common in longitudinal studies is 'dropout' or 'attrition'. This is where an individual is observed from baseline up until a certain point in time, thereafter no more measurements are made.

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, 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.

In order to obtain valid inferences from incomplete data the mechanism (probability model) producing the missing observations must be considered.

A hierarchy of three different types of missing data mechanisms can be distinguished:

1) Data are missing completely at random (MCAR) when the probability that an individual value will be missing is independent of $\mathbf{Y}_{(o)}$ and $\mathbf{Y}_{(m)}$.
2) Data are missing at random (MAR) when the probability that an individual value will be missing is independent of $\mathbf{Y}_{(m)}$ (but may depend on $\mathbf{Y}_{(o)}$).
3) Missing data are nonignorable when the probability that an individual value will be missing depends on $\mathbf{Y}_{(m)}$.

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

If missingness depends only on X, then technically it is MCAR. However, sometimes this is referred to as *covariate dependent* non-response.

Thus, in general, if non-response depends on covariates, X, it is harmless and the 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: 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: 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).

Methods of Handling Missing Data

1) ***Complete Case Methods*:** These methods omit all cases with missing values at any measurement occasion.

   Drawbacks:

   (i) Can results in a very substantial loss of information which has an impact on precision and power.
   (ii) Can give severely biased results if complete cases are not a random sample of population of interest, i.e. complete case methods require MCAR assumption.

2) ***All Available Case Methods*:** This is a general term for a variety of different methods that use the available information to estimate means and covariances (the latter based on all available *pairs* of cases).

   In general, these methods are more efficient than complete case methods (and can be fully efficient in some cases).

Drawbacks:

(i) Sample base of cases changes over measurement occasions.

(ii) Pairwise available case estimates of correlations can lie outside (-1, 1).

(iii) Available case methods require MCAR assumption.

3. ***Imputation Methods*:** These are methods that fill in the missing values. Once imputation is done, the analysis is straightforward.
   (a) Stratification: Stratify into homogeneous subsets or classes; impute mean value of strata, or randomly draw data value from those in the strata.
   (b) Regression imputation: Estimate response via some appropriately chosen regression model.

Drawbacks:

(i) Systematically underestimate the variance and covariance.

(ii) Treating imputed data as real data leads to standard errors that are too small (multiple imputation addresses this problem).

(iii) Their performance can be unreliable and usually require MCAR assumption.

(iv) Can be fairly ad hoc, e.g. LVCF.

Last Value Carried Forward: Set the response equal to the last observed value (or sometimes the 'worst' observed value).

This method of imputation is only valid under very strong assumptions. In general, LVCF is not recommended!

4. ***Likelihood-based Methods*:** At least in principle, maximum likelihood estimation for incomplete data is the same as for complete data and provides valid estimates and standard errors for more general circumstances than methods 1), 2), or 3).

   That is, under clearly stated assumptions likelihood-based methods have optimal statistical properties.

   For example, if missing data are 'ignorable' (MCAR/MAR), likelihood-based methods (e.g. PROC MIXED) simply maximize the marginal distribution of the observed responses.

   If missing data are 'non-ignorable', likelihood-based inference must also explicitly (and correctly) model the non-response process. However, with 'non-ignorable' missingness the methods are very sensitive to unverifiable assumptions.

5. ***Weighting Methods*:** Base estimation on observed data, but weight the data to account for missing data.

   Basic idea: some sub-groups of the population are under-represented in the observed data, therefore weight these up to compensate for under-representation.

For example, with dropout, can estimate the weights as a function of the individual's covariates and responses up until the time of dropout.

This approach is valid provided the model for dropout is correct, i.e. provided the correct weights are available.

REVIEW: LOGISTIC AND POISSON REGRESSION

In this lecture we consider Logistic and Poisson Regression for a single response variable.

Logistic Regression:

So far, we have considered linear regression models for a continuous response, Y, of the following form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

The response variable, Y, is assumed to have a normal distribution with mean

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

and with variance, σ^2.

Recall that the population intercept, β_0, has interpretation as the mean value of the response when all of the covariates take on the value zero.

The population slope, say β_1, has interpretation in terms of the expected change in the mean response for a single-unit change in x_1 given that all of the other covariates remain constant.

In many studies, however, we are interested in a response variable that is dichotomous 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

$$\pi = E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

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.

$$var(Y) = \pi\,(1 - \pi)$$

Instead, we can consider a logistic regression model where

$$\ln\left[\pi/\,(1 - \pi)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

This model accommodates the constraint that π is restricted to values between 0 and 1.

Recall that $\pi/(1-\pi)$ 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.

We can use ML estimation to obtain estimates of the logistic regression parameters, under the assumption that the binary responses are *Bernoulli* random variables.

Given the logistic regression model

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

the population intercept, β_0, has interpretation as the log odds of success when all of the covariates take on the value zero.

The population slope, say β_1, has interpretation in terms of the change in log odds of success for a single-unit change in x_1 given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say x_1, then β_1 has a special interpretation:

$\exp\left(\beta_1\right)$ is the *odds ratio* or ratio of odds of success for the two possible levels of x_1 (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

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_0 + \beta_1 \text{Weight}$$

where $\pi = E(Y) = \Pr(Y = 1) = \Pr(\text{BPD})$

SAS CODE

```
proc genmod data=infant;
    model y=weight / d=bin link=logit;
run;
```

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_1 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:

$$0.3658/\left(1 + 0.3658\right) = 0.268$$

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) and 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. $\ln(\lambda/t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

   Note that since $\ln(\lambda/t) = \ln(\lambda) - \ln(t)$, the Poisson regression model can also be considered as

$$\ln(\lambda) = \ln(t) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

   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

$$\ln(\lambda/t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

the population intercept, β_0, has interpretation as the log expected rate when all the covariates take on the value zero.

The population slope, say β_1, has interpretation in terms of the change in log expected rate for a single-unit change in x_1 given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say x_1, then β_1 has a special interpretation:

$\exp(\beta_1)$ is the rate ratio for the two possible levels of x_1 (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:

$$\ln(\lambda/t) = \ln(\textit{rate of CHD}) = \beta_0 + \beta_1 \, \boldsymbol{Smoke}$$

or

$$\ln(\lambda) = \ln(t) + \beta_0 + \beta_1 \, \boldsymbol{Smoke}$$

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_1 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:

$$\ln(\lambda/t) = \beta_0 + \beta_1 \text{ Smoke} + \beta_2 \text{ Type} + \beta_1 \text{BP}$$

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, the adjusted rate of CHD (controlling for blood pressure and behavior type) increases by a factor of $\exp(0.027) = 1.028$ for every cigarette smoked.

Thus, the adjusted rate of CHD in smokers of one pack per day (20 cigs) is estimated to be $(1.027)^{20} = 1.704$ times higher than the 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

$$\ln(\text{rate}) \approx \ln(p) \approx \ln[p/(1-p)]$$

Therefore, both the dependent variables and the parameters for Poisson regression and logistic regression models are approximately equal when the event being studied is rare.

In that case, the results from a Poisson and logistic regression will not give discernibly different results.

INTRODUCTION TO GENERALIZED LINEAR MODELS

In the first part of the course, we have focused on methods for analyzing longitudinal data where the dependent variable is continuous and the vector of responses is assumed to have a multivariate normal distribution.

We have also focused on fitting a linear model to the repeated measurements.

We now turn to a much wider class of regression problems; namely those in which we wish to fit a generalized linear model to the repeated measurements.

The generalized linear model is actually a class of regression models, one that includes the linear regression model but also many of the important nonlinear models used in biomedical research:

- Linear regression for continuous data

- Logistic regression for binary data

- Poisson models for counts

In the next few lectures, we will review the generalized linear model and its properties, and show how we can apply generalized linear models in the longitudinal data setting.

Before beginning a discussion of the theory, we will describe a data set illustrating some of the analytic goals.

EFFECTIVENESS OF SUCCIMER IN REDUCING BLOOD LEAD LEVELS

In the Treatment of Lead-Exposed Children Trial, 100 children were randomized equally to succimer and placebo.

The percentages of children with blood lead levels below 20 μg/dL at the three examinations after treatment were as follows:

	Succimer	Placebo	Total
Time (Days)			
7	78	16	47
28	76	26	51
42	54	26	40

How can we quantify the effect of treatment with succimer on the probability of having a blood lead level below 20 μg/dL at each occasion?

How can we test the hypothesis that succimer has no effect on these probabilities?

If we had observations at only a single time point, we could model the relative odds using logistic regression.

Here, we have to carefully consider the goals of the analysis and deal with the problem of correlation among the repeated observations.

Generalized Linear Model

The generalized linear model is actually a family of probability models that includes the normal, Bernoulli, Poisson, and Gamma distributions.

Generalized linear models extend the methods of regression analysis to settings where the outcome variable can be a dichotomous (binary) variable, an ordered categorical variable, or a count.

The generalized linear model has some of the properties of the linear model.

Most importantly, a parameter related to the expected value is assumed to depend on a linear function of the covariates.

However, the generalized linear model also differs in important ways from the linear model.

Because the underlying probability distribution may not be normal, we need new methods for parameter estimation and a new theoretical basis for the properties of estimates and test statistics.

Next we consider the main properties of the generalized linear model.

Let $Y_i, i = 1, \ldots, n$, be independent observations from a probability distribution that belongs to the family of statistical models known as generalized linear models.

The probability model for Y_i has a three-part specification:

1. The distributional assumption.

2. The systematic component.

3. The link function.

1. The distributional assumption.

Y_i is assumed to have a probability distribution that belongs to the exponential family.

The general form for the exponential family of distributions is

$$f(Y_i) = \exp\left[\left\{Y_i\theta_i - a\left(\theta_i\right)\right\}/\phi + b\left(Y_i, \phi\right)\right].$$

where θ_i is the 'canonical' parameter and ϕ is a 'scale' parameter.

Note: $a(\cdot)$ and $b(\cdot)$ are specific functions that distinguish distributions belonging to the exponential family.

When ϕ is known, this is a one-parameter exponential distribution.

The exponential family of distributions include the normal, Bernoulli, and Poisson distributions.

Normal Distribution

$$\begin{aligned} f\left(Y_i;\mu_i,\sigma^2\right) &= \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\left(Y_i-\mu_i\right)^2/2\sigma^2\right\} \\ &= \exp\left\{-1/2\ln\left(2\pi\sigma^2\right)\right\}\exp\left\{-\left(Y_i-\mu_i\right)^2/2\sigma^2\right\} \\ &= \exp\left\{-\left(Y_i^2-2Y_i\mu_i+\mu_i^2\right)/2\sigma^2-1/2\ln\left(2\pi\sigma^2\right)\right\} \\ &= \exp\left[\left\{Y_i\mu_i-\mu_i^2/2\right\}/\sigma^2-1/2\left\{Y_i^2/\sigma^2+\ln\left(2\pi\sigma^2\right)\right\}\right] \end{aligned}$$

is an exponential family distribution with $\theta_i=\mu_i$ and $\phi=\sigma^2$

Thus, θ_i is the 'location' parameter (mean) and ϕ the 'scale' parameter.

Note that $a\left(\theta_i\right)=\theta_i^2/2$ and

$$\begin{aligned} b\left(Y_i,\phi\right) &= -1/2\left\{Y_i^2/\sigma^2+\ln\left(2\pi\sigma^2\right)\right\} \\ &= -1/2\left\{Y_i^2/\phi+\ln\left(2\pi\phi\right)\right\} \end{aligned}$$

Two other important exponential family distributions are the Bernoulli and the Poisson distributions.

Bernoulli Distribution

$$f\left(Y_i;\pi_i\right)=\pi_i^{Y_i}\left(1-\pi_i\right)^{1-Y_i}$$

$$\text{where } \pi_i=\Pr\left(Y_i=1\right)$$

Note that

$$\begin{aligned} f\left(Y_i;\pi_i\right) &= \pi_i^{Y_i}\left(1-\pi_i\right)^{(1-Y_i)} \\ &= \exp\left\{Y_i\ln\left(\pi_i\right)+\left(1-Y_i\right)\ln\left(1-\pi_i\right)\right\} \\ &= \exp\left\{Y_i\ln\left[\pi_i/\left(1-\pi_i\right)\right]+\ln\left(1-\pi_i\right)\right\} \end{aligned}$$

Since

$$\begin{aligned} f\left(Y_i;\pi_i\right) &= \exp\left\{Y_i\ln\left[\pi_i/\left(1-\pi_i\right)\right]+\ln\left(1-\pi_i\right)\right\} \\ \Rightarrow \quad \theta_i &= \ln\left[\pi_i/\left(1-\pi_i\right)\right] \\ &= \text{logit}\left(\pi_i\right) \end{aligned}$$

and $\phi=1.$

Poisson Distribution

$$\begin{aligned} f\left(Y_i;\lambda_i\right) &= e^{-\lambda_i}\lambda^{Y_i}/Y_i! \\ &= \exp\left\{Y_i \ln\lambda_i - \lambda_i - \ln\left(Y_i!\right)\right\} \\ \Rightarrow \quad \theta_i &= \ln\left(\lambda_i\right) \end{aligned}$$

and

$$\phi = 1$$

Both the Bernoulli and the Poisson distributions are one-parameter distributions.

Mean and Variance of Exponential Family Distribution

The representation of the exponential family distributions makes it easy to see that the log likelihood is

$$\log L\left(Y_i;\theta_i,\phi\right)=\left\{Y_i\theta_i-a\left(\theta_i\right)\right\}/\phi+b\left(y_i,\phi\right)$$

Using calculus, various properties of the exponential family distributions can be derived.

In particular, it can be shown that

$$\begin{aligned} E\left(Y_i\right) &= a'\left(\theta_i\right) \\ Var\left(Y_i\right) &= a''\left(\theta_i\right)\phi \end{aligned}$$

$$\begin{aligned} \text{where } a'\left(\theta_i\right) &= \partial a\left(\theta_i\right)/\partial\theta_i \text{ and} \\ a''\left(\theta_i\right) &= \partial^2 a\left(\theta_i\right)/\partial\theta_i^2. \end{aligned}$$

$$\log L\left(Y_i;\theta_i,\phi\right)=\left\{Y_i\theta_i-a\left(\theta_i\right)\right\}/\phi+b\left(y_i,\phi\right)$$

For the Bernoulli distribution, we have

$$\log L\left(Y_i;\theta_i,\phi\right)=Y_i\ln\left[\pi_i/\left(1-\pi_i\right)\right]+\ln\left(1-\pi_i\right)$$

so that $\theta_i=\ln[\pi_i/\left(1-\pi_i\right)]$ and $\pi_i=e^{\theta_i}/\left(1+e^{\theta_i}\right)$.
Thus, the term $\ln\left(1-\pi_i\right)$, when expressed as a function of θ_i, is

$$\ln\left(1/\left(1+e^{\theta_i}\right)\right)=-\ln\left(1+e^{\theta_i}\right)$$

Hence, $a\left(\theta_i\right)=\ln\left(1+e^{\theta_i}\right)$, so that

$$a'\left(\theta_i\right)=e^{\theta_i}/\left(1+e^{\theta_i}\right)=\pi_i$$

Furthermore

$$\begin{aligned}a''\left(\theta_i\right) &= \left\{e^{\theta_i}\left(1+e^{\theta_i}\right)-e^{2\theta_i}\right\}/\left(1+e^{\theta_i}\right)^2\\ &= e^{\theta_i}/\left(1+e^{\theta_i}\right)^2\\ &= \pi_i\left(1-\pi_i\right)\end{aligned}$$

is the variance of the Bernoulli distribution.

2. The systematic component

Given covariates $X_{i1}, \ldots, X_{ik}$, the effect of the covariates on the expected value of Y_i is expressed through the 'linear predictor'

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} = \mathbf{X}_i \boldsymbol{\beta}$$

3. The link function, $g(\cdot)$, describes the relation between the linear predictor, η_i, and the expected value of Y_i (denoted by μ_i),

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

Note: When $\theta_i = \eta_i$, then we say that we are using the 'canonical' link function.

Common Examples of Link Functions

Normal distribution:

If we assume that $g(\cdot)$ is the identity function,

$$g(\mu) = \mu$$

then

$$\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

gives the standard linear regression model.

We can, however, choose other link functions when they seem appropriate to the application.

Bernoulli distribution:

For the Bernoulli distribution, the mean μ_i is π_i, with $0 < \pi_i < 1$, so we would prefer a link function that transforms the interval $[0, 1]$ on to the entire real line $[-\infty, \infty]$.

There are several possibilities:

$$\begin{aligned} \text{logit}: \quad \eta_i &= \ln\left[\left(\pi_i / \left(1 - \pi_i\right)\right)\right] \\ \text{probit}: \quad \eta_i &= \Phi^{-1}\left(\pi_i\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Another possibility is the complementary log-log link function:

$$\eta_i = \ln\left[-\ln\left(1 - \pi_i\right)\right]$$

Poisson Distribution:
For count data, we must have $\mu_i = \lambda_i > 0$.
The Poisson distribution is often used to model count data with the log link function

$$\eta_i = \ln(\lambda_i)$$

with inverse $\lambda_i = e^{\eta_i}$.
With the log link function, additive effects contributing to η_i act multiplicatively on λ_i.

Maximizing the Likelihood

Maximum likelihood (ML) estimation is used for making inferences about $\boldsymbol{\beta}$.

We maximize the log likelihood with respect to $\boldsymbol{\beta}$ by taking the derivative of the log likelihood with respect to $\boldsymbol{\beta}$, and then finding the values of $\boldsymbol{\beta}$ that make those derivatives equal to 0.

Given

$$\ln L = \sum_{i=1}^{n} \left(\left\{ Y_i \theta_i - a\left(\theta_i\right) \right\} / \phi + b\left(Y_i, \phi\right) \right)$$

the derivative of the log likelihood with respect to $\boldsymbol{\beta}$ is,

$$\begin{aligned} \partial \ln L / \partial \boldsymbol{\beta} &= \sum_{i=1}^{n} \left(\partial \theta_i / \partial \boldsymbol{\beta}\right) \left\{ Y_i - a'\left(\theta_i\right) \right\} / \phi \\ &= \sum_{i=1}^{n} \left(\partial \theta_i / \partial \boldsymbol{\beta}\right) \left\{ Y_i - \mu_i \right\} / \phi \end{aligned}$$

When a 'canonical' link function, $\theta_i = \eta_i$, has been assumed

$$\partial \ln L / \partial \boldsymbol{\beta} = \sum_{i=1}^{n} \mathbf{X}_i' \left\{ Y_i - \mu_i \right\} / \phi$$

Note: This is a vector equality because there is one equation for each element of $\boldsymbol{\beta}$.

Solving this set of simultaneous equations,

$$\sum_{i=1}^{n} \mathbf{X}_i' \left\{ Y_i - \mu_i \right\} = 0$$

yields the maximum likelihood estimates of $\boldsymbol{\beta}$.

GENERALIZED LINEAR MODELS FOR LONGITUDINAL DATA

In this lecture, we will briefly survey a number of general approaches for analyzing longitudinal data. These approaches can be considered extensions of generalized linear models to correlated data.

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

Recall that in linear 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 assumptions about the source of the correlation can lead to regression coefficients with distinct interpretations.

We will return to this issue later in the course.

We will consider three main extensions of generalized linear models:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

Suppose that $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ 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,

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right)$$

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:

$$f\left(Y_{ij}\right) \text{ for } j = 1, 2, \ldots, p$$

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).

Consider the Treatment of Lead-Exposed Children Trial where 100 children were randomized equally to succimer and placebo.

The percentages of children with blood lead levels below 20 μg/dL at the three examinations after treatment were as follows:

	Succimer	Placebo	Total
Time (Days)			
7	78	16	47
28	76	26	51
42	54	26	40

We might let x_i be an indicator variable denoting treatment assignment. If μ_{ij} is the probability of having blood lead below 20 in treatment group i at time j, we might assume

$$\text{logit}\left(\mu_{ij}\right) = \beta_0 + \beta_1 \text{time}_{1ij} + \beta_2 \text{time}_{2ij} + \beta_3 X_i$$

where time_{1ij} and time_{2ij} are indicator variables for days 7 and 28 respectively.

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 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 according to a random effects distribution, denoted by F.

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.

Suppose, for example, that the probability of a blood lead < 20 for participants in the TLC trial is described by a logistic model, but that the risk for an individual child depends on that child's latent (perhaps environmentally and genetically determined) 'random response level'.

Then we might consider a model where

$$\text{logit Pr}\left(Y_{ij} = 1 | b_i\right) = \beta_0 + \beta_1 \text{time}_{1ij} + \beta_2 \text{time}_{2ij} + \beta_3 X_i + b_i$$

Note that such a model also requires specification of $F\left(b_i\right)$.

Frequently, it is assumed that b_i is normally distributed with mean 0 and some unknown variance, σ_b^2.

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,

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right) = f\left(Y_{i1}\right) f\left(Y_{i2}|Y_{i1}\right) \cdots f\left(Y_{ip}|Y_{i1}, \ldots, Y_{i,p-1}\right)$$

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.

Thus, for the blood lead data in the TLC trial, we could write the probability model as

$$f\left(Y_{i1}|x_i\right) f\left(Y_{i2}|Y_{i1}, x_i\right) f\left(Y_{i3}|Y_{i1}, Y_{i2}, x_i\right)$$

That is, the probability of a blood lead value below 20 at time 2 is modeled conditional on whether blood lead was below 20 at time 1, and so on.

As noted above, transitional models potentially provide a complete description of the joint distribution.

For the linear model, the regression coefficients of the marginal, mixed effects and transitional models can be directly related to one another.

For example, coefficients from random effects and transitional models can be given marginal interpretations.

However, with non-linear link functions, such as the logit or log link functions, this is not the case.

We will return to this point later.

For now, we describe the development and application of marginal models for the analysis of repeated responses.

MARGINAL MODELS

A feature of marginal models is that the models for the mean and the covariance are specified separately.

We assume that the marginal density of Y_{ij} is given by

$$f\left(Y_{ij}\right) = \exp\left[\left\{Y_{ij}\theta_{ij} - a\left(\theta_{ij}\right)\right\}/\phi + b\left(Y_{ij}, \phi\right)\right]$$

That is, each Y_{ij} is assume to have a distribution from the exponential family.

The marginal expectation, $E\left(Y_{ij}\right)$, of each response is then modelled as a function of covariates.

Specifically, with marginal models we make the following assumptions:

- the marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on explanatory variables, $\mathbf{X}_{ij}$, through a known link function

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$$

- the marginal variance of Y_{ij} depends on the marginal mean according to

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

  where $\upsilon(\mu_{ij})$ is a known 'variance function' and ϕ is a scale parameter that may need to be estimated.

- the covariance between Y_{ij} and Y_{ik} is a function of the means and perhaps of additional parameters, say $\boldsymbol{\alpha}$, that may also need to be estimated.

Marginal models are considered a natural approach when we wish to extend the generalized linear models methods of analysis of independent observations to the setting of correlated responses.

The crucial point to keep in mind is that with marginal models the mean and within-subject correlation are modelled *separately*.

Examples of Marginal Models:

Continuous responses:

1. $\mu_{ij} = \eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$
   (i.e. linear regression)

2. $Var\left(Y_{ij}\right) = \phi$
   (i.e. homogeneous variance)

3. $Corr\left(Y_{ij}, Y_{ik}\right) = \alpha^{|k-j|}$ $(0 \leq \alpha \leq 1)$
   (i.e. autoregressive correlation)

Binary responses:

1. $\text{Logit}\left(\mu_{ij}\right) = \eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$
   (i.e. logistic regression)
2. $Var\left(Y_{ij}\right) = \mu_{ij}\left(1 - \mu_{ij}\right)$
   (i.e. Bernoulli variance)
3. $Corr\left(Y_{ij}, Y_{ik}\right) = \alpha_{jk}$
   (i.e. unstructured correlation)

Count data:

1. $\text{Log}\left(\mu_{ij}\right) = \eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$
   (i.e. Poisson regression)
2. $Var\left(Y_{ij}\right) = \mu_{ij}\phi$
   (i.e. extra-Poisson variance)
3. $Corr\left(Y_{ij}, Y_{ik}\right) = \alpha$
   (i.e. compound symmetry correlation)

Interpretation of Marginal Models

The regression parameters, $\boldsymbol{\beta}$, have 'population-averaged' interpretations:

- describe the effect of covariates on the marginal expectations or average responses

- contrast the means in sub-populations that share common covariate values

The regression parameters, $\boldsymbol{\beta}$, have the same interpretation as in cross-sectional analyses.

The nature or magnitude of the correlation does not alter the interpretation of $\boldsymbol{\beta}$.

Statistical Inference for Marginal Models

Maximum Likelihood (ML):

Unfortunately, with discrete response data there is no 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.

Recall: In linear models for normal responses, specifying the means and the covariance matrix fully determines the likelihood.

This is not the case with discrete response data.

To specify the likelihood for multivariate discrete data, additional assumptions about 'higher-order moments' are required.

Even when additional assumptions are made, the likelihood is often intractable.

To illustrate some of the problems, consider a discrete response, Y_{ij}, having C categories.

The joint distribution of $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ is multinomial, with a C^p joint probability vector.

For example, when $C = 5$ and $p = 10$, the joint probability vector is of length 10,000,000.

Problem: Computations grow exponentially with p and ML quickly becomes impractical.

Solution: We will consider an alternative approach to estimation - *Generalized Estimating Equations* (GEE).

GENERALIZED ESTIMATING EQUATIONS

Since there is no 'convenient' or natural specification of the joint multivariate distribution of $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ for marginal models when the responses are discrete, we need an alternative to maximum likelihood (ML) estimation.

Liang and Zeger (1986) proposed such a method based on the concept of 'estimating equations'.

This provides a general and unified approach for analyzing discrete and continuous responses with marginal models.

The essential idea was to generalize the usual univariate likelihood equations by introducing the covariance matrix of the vector of responses, $\mathbf{Y}_i$.

For linear models, generalized least squares (GLS) can be considered a special case of this 'estimating equations' approach.

For non-linear models, this approach is called 'generalized estimating equations' (or GEE).

Fitting Marginal Models

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ be a vector of correlated responses for the i^{th} subject $(i = 1, \ldots, n)$.

Suppose that the following marginal model has been assumed:

- the marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on explanatory variables, $\mathbf{X}_{ij}$, through a known link function

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$$

- the marginal variance of Y_{ij} depends on the marginal mean according to

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

  where $\upsilon(\mu_{ij})$ is known and ϕ may have to be estimated.

- the correlation between Y_{ij} and Y_{ik} is a function of some additional parameters, $\boldsymbol{\alpha}$, and may also depend on μ_{ij} and μ_{ik}.

Then, an estimate of $\boldsymbol{\beta}$ can be obtained as the solution to the following 'generalized estimating equations'

$$\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i \right) = 0$$

where $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ and $\mathbf{V}_i$ is a 'working' covariance matrix, i.e. $\mathbf{V}_i \approx Cov\left(\mathbf{Y}_i\right)$.

That is,

$$\mathbf{D}_i = \begin{bmatrix} \partial \mu_{i1}/\partial \beta_1 & \partial \mu_{i1}/\partial \beta_2 & \dots & \partial \mu_{i1}/\partial \beta_k \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ \partial \mu_{ip}/\partial \beta_1 & \partial \mu_{ip}/\partial \beta_2 & \dots & \partial \mu_{ip}/\partial \beta_k \end{bmatrix}$$

Note that $\mathbf{D}_i$ is a function of $\boldsymbol{\beta}$, while $\mathbf{V}_i$ is a function of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Recall that we can express $\mathbf{V}_i$ as a function of the variances and correlations,

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}\left(\boldsymbol{\alpha}\right) \mathbf{A}_i^{1/2}$$

where $\mathbf{A}_i$ is a diagonal matrix with $\upsilon\left(\mu_{ij}\right)$ as the j^{th} diagonal element.

That is,

$$\mathbf{A}_i = \begin{bmatrix} \upsilon\left(\mu_{i1}\right) & 0 & 0 & \dots & 0 & 0 \\ 0 & \upsilon\left(\mu_{i2}\right) & 0 & \dots & 0 & 0 \\ . & . & . & \dots & . & . \\ . & . & . & \dots & . & . \\ . & . & . & \dots & . & . \\ 0 & 0 & 0 & \dots & 0 & \upsilon\left(\mu_{ip}\right) \end{bmatrix}$$

Therefore the generalized estimating equations depend on both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Because the generalized estimating equations depend on both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, an iterative two-stage estimation procedure is required:

1. Given current estimates of $\boldsymbol{\alpha}$ and ϕ, an estimate of $\boldsymbol{\beta}$ is obtained as the solution to the 'generalized estimating equations'

2. Given current estimate of $\boldsymbol{\beta}$, estimates of $\boldsymbol{\alpha}$ and ϕ are obtained based on the standardized residuals,

$$r_{ij} = \left(Y_{ij} - \widehat{\mu}_{ij}\right) / \upsilon \left(\widehat{\mu}_{ij}\right)^{1/2}$$

Properties of GEE estimators

Assuming that the estimators of $\boldsymbol{\alpha}$ and ϕ are consistent, $\widehat{\boldsymbol{\beta}}$, the solution to the generalized estimating equations has the following properties:

1. $\widehat{\boldsymbol{\beta}}$ is a consistent estimate of $\boldsymbol{\beta}$ (with high probability, it is close to $\boldsymbol{\beta}$)

2. In large samples, $\widehat{\boldsymbol{\beta}}$ has a multivariate normal distribution

3. $Cov\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{F}^{-1}\mathbf{G}\mathbf{F}^{-1}$, where

$$\begin{aligned} \mathbf{F} &= \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \\ \mathbf{G} &= \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} Cov\left(\mathbf{Y}_i\right) \mathbf{V}_i^{-1} \mathbf{D}_i \end{aligned}$$

Note that $\mathbf{F}$ and $\mathbf{G}$ can be estimated by replacing $\boldsymbol{\alpha}$, ϕ, and $\boldsymbol{\beta}$ by their estimates, and replacing $Cov\left(\mathbf{Y}_i\right)$ by $\left(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i\right)\left(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i\right)'$.
That is, we can use the empirical or so-called 'sandwich' variance estimator.

In summary, the GEE estimators have the following attractive properties:

1. In many cases $\widehat{\boldsymbol{\beta}}$ 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. $\widehat{\boldsymbol{\beta}}$ is consistent even if the covariance of $\mathbf{Y}_i$ has been misspecified

3. Standard errors for $\widehat{\boldsymbol{\beta}}$ can be obtained using the empirical or so-called 'sandwich' estimator

Example 1: Estimating the Effect of Succimer Therapy on $Pr\,(\text{Blood lead } < 20\mu g/dL)$.

Consider the Treatment of Lead-Exposed Children Trial where 100 children were randomized equally to succimer and placebo.

The percentages of children with blood lead levels below 20 μg/dL at the three examinations after treatment were as follows:

	Succimer	Placebo	Total
Time (Days)			
7	78	16	47
28	76	26	51
42	56	28	42

Suppose that we wish to fit a logistic model to the 3 by 2 table of response rates in the blood lead study.

We could begin by fitting the model

$$\text{logit}\left(\mu_{ij}\right) = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_i$$

where time is initially assumed to be a continuous variable taking the values 1, 4, and 6 (weeks).

However, we must also make some assumptions about the variances and the correlations. For example, we could assume that

$$var\left(Y_{ij}\right) = \mu_{ij}\left(1 - \mu_{ij}\right)$$

and

$$corr\left(Y_{ij}, Y_{ik}\right) = \alpha \text{ ('exchangeable')}.$$

We can perform the analysis using the GEE option is SAS PROC GENMOD.

```
data lead;
      input id trt $ week y;
cards ;
001 P 1 0
001 P 4 0
001 P 6 0
002 A 1 1
002 A 4 1
002 A 6 0
  .   .  .  .
  .   .  .  .
  .   .  .  .
;

proc genmod data=lead descending;
      class id trt;
      model y=week trt / d=bin;
      repeated subject=id / type=exch corrw modelse;
      output out=pprobs p=pred xbeta=xbeta;

proc print data=pprobs;
run;
```

GEE Model Information

Description	Value
Correlation Structure	Exchangeable
Subject Effect	id (100 levels)
Number of Clusters	100
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Working Correlation Matrix

	col1	col2	col3
ROW1	1.0000	0.4571	0.4571
ROW2	0.4571	1.0000	0.4571
ROW3	0.4571	0.4571	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	Z	Pr> $\|Z\|$
Intercept		-1.0591	0.2823	-3.750	0.0002
week		-0.0420	0.0517	-0.814	0.4158
trt	A	2.0487	0.3661	5.596	<.0001

Comparison of observed and predicted probabilities from the model with continuous time and treatment.

	Succimer	Placebo
Time (Days)		
7	.78 (.72)	.16 (.25)
28	.76 (.69)	.26 (.23)
42	.56 (.68)	.28 (.21)

(Predicted probabilities in parentheses)

Next, consider fitting the model with treatment*time interaction,

$$\text{logit}\left(\mu_{ij}\right) = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_i + \beta_3 \text{time}_{ij} * \text{trt}_i$$

We can perform the analysis using the following commands in SAS PROC GENMOD:

```
proc genmod data=lead descending;
      class id trt;
      model y=week trt week*trt/ d=bin;
      repeated subject=id / type=exch corrw;
run;
```

GEE Model Information

Description	Value
Correlation Structure	Exchangeable
Subject Effect	id (100 levels)
Number of Clusters	100
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Working Correlation Matrix

	col1	col2	col3
ROW1	1.0000	0.4758	0.4758
ROW2	0.4758	1.0000	0.4758
ROW3	0.4758	0.4758	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	Z	Pr> $\|Z\|$
INTERCEPT		-1.7343	0.4020	-4.31	<.0001
WEEK		0.1428	0.0786	1.82	0.0692
TRT	A	3.3788	0.5714	5.91	<.0001
WEEK*TRT	A	-0.3478	0.1043	-3.33	0.0009

Placebo:

$$\text{logit}\left(\widehat{\mu}_{ij}\right) = -1.7343 + 0.1428\ \text{time}_{ij}$$

Succimer:

$$\begin{aligned}
\text{logit}\left(\widehat{\mu}_{ij}\right) &= (-1.7343 + 3.3788) \\
&+ (0.1428 - 0.3478)\,\text{time}_{ij} \\
&= 1.6445 - 0.2050\ \text{time}_{ij}
\end{aligned}$$

Thus, in the placebo group the odds of blood lead $< 20\mu$g/dL is increasing over time, while in the succimer group the odds is decreasing.

$$\text{Recall: Odds} = \frac{\Pr\left(\text{Blood lead} < 20\mu g/dL\right)}{\Pr\left(\text{Blood lead} \geq 20\mu g/dL\right)}$$

Comparison of observed and predicted probabilities from the model with interaction of continuous time and treatment.

Time (Days)	Succimer	Placebo
7	.78 (.81)	.16 (.17)
28	.76 (.70)	.26 (.24)
42	.56 (.60)	.28 (.29)

(Predicted probabilities in parentheses)

Generalized Estimating Equations

Example 2: Six Cities Study of Respiratory Illness in Children.

A non-randomized longitudinal study of the health effects of air pollution.

Subset of data from one of the participating cities: Steubenville, Ohio

Outcome variable: Binary indicator of respiratory infections in child.

Measurements on the children were taken annually at ages 7, 8, 9, and 10.

Interested in changes in the rates of respiratory illness and the influence of maternal smoking.

Proportion of children reporting respiratory illness, by age and mother's smoking status
Six Cities Study of Air Pollution and Health

	Age			
	7	8	9	10
Smoking Mother	0.17	0.21	0.19	0.14
Non-Smoking Mother	0.16	0.15	0.14	0.11

Assume marginal probability of infection follows the following logistic model,

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_i + \beta_3 \text{age}_{ij} * \text{smoke}_i$$

where age_{ij} = child's age - 9; and $\text{smoke}_i = 1$ if child's mother smokes, 0 otherwise.
Also, we assume that

$var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$

Need to make assumptions about the association, e.g., $corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$ ('unstructured').

However, with binary responses correlations are not the best choice for modelling the association.

Measures of Association for Binary Responses

A drawback of using correlations as a measure of association with binary responses is that the correlations 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.

$\Longrightarrow$ Use modified GEE with association modelled in terms of odds ratios rather than correlations.

For binary responses, PROC GENMOD has options that allow modelling of the log odds ratios.

SAS Commands for PROC GENMOD

```
data child;
   input id smoke age y;
cards;
001 1 7  0
001 1 8  0
001 1 9  0
001 1 10 1
002 0 7  1
002 0 8  0
002 0 9  1
002 0 10 0
     .
     .
     .
;
proc genmod data=child descending;
    class id;
    model y = age smoke age*smoke / d=bin;
    repeated  subject=id /  covb logor=fullclust;
run;
```

SAS Output from PROC GENMOD

GEE Model Information

Log Odds Ratio Structure	Fully Parameterized Clusters
Subject Effect	id (537 levels)
Number of Clusters	537
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	4

Log Odds Ratio Parameter Information

Parameter	Group
Alpha1	(1, 2)
Alpha2	(1, 3)
Alpha3	(1, 4)
Alpha4	(2, 3)
Alpha5	(2, 4)
Alpha6	(3, 4)

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	Z	Pr > |Z|
Intercept	-1.9052	0.1191	-16.00	<.0001
age	-0.1434	0.0583	-2.46	0.0139
smoke	0.3061	0.1882	1.63	0.1038
age*smoke	0.0685	0.0887	0.77	0.4399
Alpha1	1.9460	0.2630	7.40	<.0001
Alpha2	1.7344	0.2688	6.45	<.0001
Alpha3	1.9889	0.2817	7.06	<.0001
Alpha4	2.4268	0.2752	8.82	<.0001
Alpha5	1.9358	0.2817	6.87	<.0001
Alpha6	2.2250	0.2885	7.71	<.0001

Next, consider the model without age*smoke interaction.

```
proc genmod data=child descending;
    class id;
    model y = age smoke / d=bin;
    repeated  subject=id /  covb logor=fullclust;
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	Z	Pr > |Z|
Intercept	-1.8847	0.1138	-16.56	<.0001
age	-0.1158	0.0439	-2.63	0.0084
smoke	0.2561	0.1779	1.44	0.1501
Alpha1	1.9393	0.2634	7.36	<.0001
Alpha2	1.7275	0.2690	6.42	<.0001
Alpha3	1.9788	0.2814	7.03	<.0001
Alpha4	2.4302	0.2752	8.83	<.0001
Alpha5	1.9448	0.2815	6.91	<.0001
Alpha6	2.2271	0.2881	7.73	<.0001

Thus, there is evidence that the rates of respiratory infection decline with age, but the rates do not appear to depend on whether a child's mother smokes.

For a child whose mother does not smoke, the predicted probability of infection at ages 7, 8, 9, and 10 is 0.170, 0.150, 0.132, and 0.116 respectively.

While for a child whose mother does smoke, the predicted probability of infection at ages 7, 8, 9, and 10 is 0.205, 0.182, 0.161, and 0.142 respectively.

GEE with Missing Data

PROC GENMOD will allow you to have missing data that is either intermittent or due to dropouts, provided that the data are missing completely at random (MCAR).

If there is imbalance due to missing data, it may be necessary to include a within-subject effect when using PROC GENMOD. That is, it may be necessary to define an effect specifying the order of measurements within individuals:

```
proc genmod data=child descending;
   class id time;
   model y=age smoke / d=bin;
   repeated  subject=id /
     within=time logor=fullclust covb;
```

Note: The variable defining the within-subject effect must be listed on the CLASS statement.

GENERALIZED LINEAR MIXED MODELS

So far, we have discussed marginal models for longitudinal data and the use of generalized estimating equations to fit these models.

To fit marginal models, we made some assumptions about the marginal distribution at each time point, and estimated a matrix of correlation coefficients linking repeated observations of the same subject.

In specifying the marginal expectations and variances and the covariance matrices, we were not fully specifying the joint distribution of the repeated measurements.

Thus, estimation using GEE is not likelihood-based. Nevertheless, we described methods for estimating and forming confidence intervals for the regression parameters.

Next, we consider a second type of extension of the generalized linear model, the *generalized linear mixed model*.

We describe how these models extend the conceptual approach represented by the linear mixed model [3].

We also highlight their greater degree of conceptual and analytic complexity relative to the marginal models.

[3] Recall: A mixed model is one that contains both fixed and random effects

Incorporating Random Effects into Generalized Linear Models

The basic premise is that we assume that there is natural heterogeneity across individuals in a subset of the regression coefficients.

That is, a subset of the regression coefficients (e.g. intercepts) 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.

Before discussing generalized linear mixed models, lets consider the simplest case: the *linear mixed effects model*.

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

First Step: Assumes $\mathbf{Y}_i$ has a normal distribution that depends on population-specific effects, $\boldsymbol{\beta}$, and individual-specific effects, $\mathbf{b}_i$,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

with $\mathbf{e}_i$ being a vector of errors, and $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$.

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

That is, the response for the i^{th} subject at the j^{th} occasion is assumed to differ from the population mean, $\mathbf{X}_{ij}\boldsymbol{\beta}$, by a subject effect, $\mathbf{b}_i$, and a within-subject measurement error, e_{ij}.

Note 1: $\mathbf{R}_i = var(\mathbf{e}_i)$ describes the covariance among observations when we focus on the response profile of a *specific individual*.

That is, it is the covariance of the i^{th} subject's deviations from his/her mean profile $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$

Usually, it is assumed that $\mathbf{R}_i = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix $\Longrightarrow$ 'conditional independence assumption'.

Note 2: In the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

the vector of regression parameters $\boldsymbol{\beta}$ are the fixed effects, which are assumed to be the same for all individuals.

In contrast to $\boldsymbol{\beta}$, the $\mathbf{b}_i$ are subject-specific regression coefficients and describe the mean response profile of a specific individual (when combined with the fixed effects) .

Finally, in the mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

recall that

$$E(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

and

$$E\left(\mathbf{Y}_i\right) = \mathbf{X}_i\boldsymbol{\beta}$$

Similarly,

$$var\left(\mathbf{Y}_i|\mathbf{b}_i\right) = var\left(\mathbf{e}_i\right) = \mathbf{R}_i = \sigma^2\mathbf{I}$$

and

$$\begin{aligned} var\left(\mathbf{Y}_i\right) &= var\left(\mathbf{Z}_i\mathbf{b}_i\right) + var\left(\mathbf{e}_i\right) \\ &= \mathbf{Z}_i\mathbf{G}Z_i' + \mathbf{R}_i = \mathbf{Z}_i\mathbf{G}Z_i' + \sigma^2\mathbf{I} \end{aligned}$$

Thus, the introduction of random effects, $\mathbf{b}_i$, induces correlation (marginally) among the $\mathbf{Y}_i$.

Generalized Linear Mixed Model

For non-Normal responses, $\mathbf{Y}_i$, the generalized linear mixed model can also be considered in two steps:

First Step: Assumes that the conditional distribution of each Y_{ij}, given $\mathbf{b}_i$, belongs to the exponential family with conditional mean,

$$g(E[Y_{ij}|\mathbf{b}_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i$$

where $g(\cdot)$ is a known link function.

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

Note: There is an additional assumption of 'conditional independence', i.e., given $\mathbf{b}_i$, the responses $Y_{i1}, Y_{i2}, ..., Y_{ip}$ are mutually independent.

Example 1:

Binary logistic model with random intercepts:

$$\text{logit}(E[Y_{ij}|b_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} + b_i$$

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

Example 2:

Random coefficients Poisson regression model:

$$\log(E[Y_{ij}|\mathbf{b}_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i$$

with $\mathbf{X}_{ij} = \mathbf{Z}_{ij} = [1, t_{ij}]$ (i.e. random intercepts and random slopes) and $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$.

Interpretation

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

Main focus is on the individual and the influence of covariates on the individual.

Regression parameters, $\boldsymbol{\beta}$, measure the direct influence of covariates on the responses of heterogeneous individuals.

For example, in the following logistic model,

$$\text{logit}(E[Y_{ij}|b_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} + b_i$$

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 a specific individual having an underlying propensity to respond positively, b_i.

Estimation

The joint probability density function is given by:

$$f(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{b}_i)f(\mathbf{b}_i)$$

Inferences are based on the marginal or integrated likelihood function:

$$\prod_{i=1}^{n} \int f(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i$$

That is, ML estimation of $\boldsymbol{\beta}$ and $\mathbf{G}$ is based on the marginal or integrated likelihood of the data (obtained by averaging over the distribution of the unobserved random effects, $\mathbf{b}_i$.

Estimation using maximum likelihood involves a two-step procedure:

1. ML estimation of $\boldsymbol{\beta}$ and $\mathbf{G}$ is based on the marginal or integrated likelihood of the data (obtained by averaging over the distribution of the unobserved random effects, $\mathbf{b}_i$).

   However, simple analytic solutions are rarely available and numerical or Monte Carlo integration techniques are required.

2. Given estimates of $\boldsymbol{\beta}$ and $\mathbf{G}$, the random effects can be predicted as follows,

$$\hat{\mathbf{b}}_i = E(\mathbf{b}_i|\mathbf{Y}_i; \hat{\boldsymbol{\beta}}, \hat{\mathbf{G}})$$

   (Posterior mean)

   Note that $E(\mathbf{b}_i|\mathbf{Y}_i; \hat{\boldsymbol{\beta}}, \hat{\mathbf{G}})$ involves integrating (or averaging) over the distribution of the unobserved random effects, $\mathbf{b}_i$).

   However, simple analytic solutions are rarely available and numerical or Monte Carlo integration techniques are required.

Statistical Software: PROC NLMIXED in SAS

A potential limitation of generalized linear mixed models is their computational burden. Because, in general, there is no simple closed-form solution for the marginal likelihood, numerical integration techniques are required.

Maximum (marginal) likelihood estimation has only recently been implemented in standard statistical software, e.g., PROC NLMIXED in SAS.

PROC NLMIXED directly maximizes an approximate integrated likelihood (using numerical quadrature).

Example: Six Cities Study of Respiratory Illness in Children.

A non-randomized longitudinal study of the health effects of air pollution.

Subset of data from one of the participating cities: Steubenville, Ohio

Outcome variable: Binary indicator of respiratory infections in child.

Measurements on the children were taken annually at ages 7, 8, 9, and 10.

Interested in changes in an individual's rate of respiratory illness and the influence of maternal smoking?

Assume conditional probability of infection follows the mixed effects logistic regression model,

$$\text{logit}(E[Y_{ij}|b_i]) = \beta_0 + b_i + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_i$$

where age_{ij} = child's age - 9, and smoke_i =1 if child's mother smokes, 0 otherwise;
and $b_i \sim N(0, \sigma^2)$.
Also, we assume that

$$var(Y_{ij}|b_i) = E[Y_{ij}|b_i](1 - E[Y_{ij}|b_i]).$$

SAS Commands for PROC NLMIXED

```
proc nlmixed data=child qpoint=50;
   parms beta0=-1.88 beta1=-.1158
         beta2=.2561 s2u=2.0;
   eta=beta0 + beta1*age + beta2*smoke + u;
   expeta=exp(eta);
   p=expeta/(1+expeta);
   model y ~ binary(p);
   random u ~ normal(0, s2u) subject=id;
run;
```

SAS Output from PROC NLMIXED

The NLMIXED Procedure

Specifications

Data Set	WORK.CHILD
Dependent Variable	Y
Distribution for Dependent Variable	Binary
Random Effects	u
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	2148
Observations Not Used	0
Total Observations	2148
Subjects	537
Max Obs Per Subject	4
Parameters	4
Quadrature Points	50

Fitting Information

-2 Log Likelihood	1595.3
AIC (smaller is better)	1603.3
AICC (smaller is better)	1603.3
BIC (smaller is better)	1620.4

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
beta0	-3.1015	0.2190	536	-14.16	<.0001
beta1	-0.1756	0.0677	536	-2.59	0.0097
beta2	0.3985	0.2731	536	1.46	0.1450
s2u	4.6866	0.8005	536	5.85	<.0001

Results of the analysis suggest:

1. Estimates of the fixed effects in the mixed effects logistic model are larger than in the marginal model

2. β_2 has interpretation in terms of the log odds of infection for a particular child.
   That is, the ratio of odds of infection for a given child whose mother smokes, versus the same child (or a child with identical latent, underlying risk) whose mother does not smoke, is 1.49 ($e^{0.399}$).

3. Estimated variance of the random intercepts in relatively large

4. Heterogeneity should not be ignored

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 chemotherapy.

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?

[Reproduced from Diggle, Liang & Zeger (1994)]

Table 1.5. Four successive two-week seizure counts for each of 59 epileptics. Covariates are adjuvant treatment (0=placebo, 1=progabide), eight-week baseline seizure counts, and age (in years)

Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age	Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age
5	3	3	3	0	11	31	0	4	3	0	1	19	20
3	5	3	3	0	11	30	3	6	1	3	1	10	20
2	4	0	5	0	6	25	2	6	7	4	1	19	18
4	4	1	4	0	8	36	4	3	1	3	1	24	24
7	18	9	21	0	66	22	22	17	19	16	1	31	30
5	2	8	7	0	27	29	5	4	7	4	1	14	35
6	4	0	2	0	12	31	2	4	0	4	1	11	57
40	20	23	12	0	52	42	3	7	7	7	1	67	20
5	6	6	5	0	23	37	4	18	2	5	1	41	22
14	13	6	0	0	10	28	2	1	1	0	1	7	28
26	12	6	22	0	52	36	0	2	4	0	1	22	23
12	6	8	5	0	33	24	5	4	0	3	1	13	40
4	4	6	2	0	18	23	11	14	25	15	1	46	43
7	9	12	14	0	42	36	10	5	3	8	1	36	21
16	24	10	9	0	87	26	19	7	6	7	1	38	35
11	0	0	5	0	50	26	1	1	2	4	1	7	25
0	0	3	3	0	18	28	6	10	8	8	1	36	26
37	29	28	29	0	111	31	2	1	0	0	1	11	25
3	5	2	5	0	18	32	102	65	72	63	1	151	22
3	0	6	7	0	20	21	4	3	2	4	1	22	32
3	4	3	4	0	12	29	8	6	5	7	1	42	25
3	4	3	4	0	9	21	1	3	1	5	1	32	35
2	3	3	5	0	17	32	18	11	28	13	1	56	21
8	12	2	8	0	28	25	6	3	4	0	1	24	41
18	24	76	25	0	55	30	3	5	4	3	1	16	32
2	1	2	1	0	9	40	1	23	19	8	1	22	26
3	1	4	2	0	10	19	2	3	0	1	1	25	21
13	15	13	12	0	47	22	0	0	0	0	1	13	36
11	14	9	8	1	76	18	1	4	3	2	1	12	37
8	7	9	4	1	38	32							

Assume conditional rate of seizures follows the following mixed effects loglinear model,

$$\log(E[Y_{ij}|\mathbf{b}_i]) = \log(t_{ij}) + \beta_0 + b_{i0} + \beta_1 \text{time}_{ij} + b_{i1}\text{time}_{ij} + \beta_2 \text{trt}_{ij} + \beta_3 \text{time}_{ij} * \text{trt}_{ij}$$

where t_{ij} = length of period; $\text{time}_{ij} = 1$ if periods 1-4, 0 if baseline; $\text{trt}_{ij} = 1$ if progabide, 0 if placebo.

(b_{i0}, b_{i1}) are assumed to have a bivariate normal distribution with zero mean and covariance $\mathbf{G}$.

Also, we assume that

$$var(Y_{ij}|\mathbf{b}_i) = E[Y_{ij}|\mathbf{b}_i].$$

SAS Commands for PROC NLMIXED

```
proc nlmixed data=new qpoints=50;
   parms beta0=1.346  beta1=.1118 beta2=-.1068
         beta3=-.3024 s2u1=0.5 s2u2=0.25 cu12=0.01;
   eta = ltime + beta0 + beta1*time + beta2*trt
         + beta3*time*trt + u1 + u2*time;
   mu=exp(eta);
   model y ~ poisson(mu);
   random u1 u2 ~ normal([0,0], [s2u1,cu12,s2u2])
   subject=id;
run;
```

Output from PROC NLMIXED

The NLMIXED Procedure

Specifications

Data Set	WORK.NEW
Dependent Variable	y
Distribution for Dependent Variable	Poisson
Random Effects	u1 u2
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	290
Observations Not Used	0
Total Observations	290
Subjects	58
Max Obs Per Subject	5
Parameters	7
Quadrature Points	50

Fitting Information

-2 Log Likelihood	1787.1
AIC (smaller is better)	1801.1
BIC (smaller is better)	1815.5

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
beta0	1.0692	0.1344	56	7.96	<.0001
beta1	0.0078	0.1070	56	0.07	0.9421
beta2	-0.0079	0.1860	56	-0.04	0.9661
beta3	-0.3461	0.1489	56	-2.33	0.0237
s2u1	0.4529	0.0935	56	4.84	<.0001
s2u2	0.2163	0.0587	56	3.68	0.0005
cu12	0.0151	0.0529	56	0.29	0.7762

Results of the analysis suggest:

1. A patient treated with placebo has the same expected seizure rate before and after randomization [$exp(0.010) \approx 1$].

2. A patient treated with progabide has expected seizure rate reduced after treatment by approximately 28% [$1 - exp(0.010 - 0.345) \approx 0.28$].

3. Estimated variance of the random intercepts and slopes in relatively large

4. Heterogeneity should not be ignored

TRANSITIONAL MODELS

In transitional models the conditional distribution of each response is expressed as an explicit function of the past responses and covariates (the 'history').

Therefore the correlation among the repeated responses can be thought of as arising due to past values of the responses explicitly influencing the present observation (i.e. the 'present' depends on the 'past')

Additional notation:

Let the 'history' of the past responses at the j^{th} occasion be denoted by,

$$H_{ij} = \{Y_{i1}, ..., Y_{i,j-1}\}$$

Next, we consider a broad class of transitional models known as generalized autoregressive models.

Generalized Autoregressive Models

We assume that $f(y_{ij}|H_{ij})$ has a distribution belonging to the exponential family, with conditional mean

$$g(E[y_{ij}|H_{ij}]) = \mathbf{X}_{ij}\boldsymbol{\beta} + \sum_{r=1}^{s} \alpha_r f_r(H_{ij}),$$

where $g(\cdot)$ is a known link function and the $f_r(H_{ij})$ are known functions of previous observations.

For example,

$$f_1(H_{ij}) = Y_{i,j-1};\ f_2(H_{ij}) = Y_{i,j-2};\ f_3(H_{ij}) = Y_{i,j-3}.$$

Note: Initial values $Y_{i0}, Y_{i,-1}, ...,$ are assumed part of the covariates, X_{ij}.

In generalized autoregressive models the assumption that $\boldsymbol{\beta}$ are homogeneous with respect to time and in the population can be relaxed; e.g., homogeneity of parameters in sub-populations.

If only a finite number of past responses are included in H_{ij}, so-called Markov models are obtained.

Note: When the response is discrete, these models are often called *Markov chain models*.

Models in which the conditional distribution of Y_{ij} given H_{ij} depends only on the q immediately prior observations are known as
Markov models of order q.

Examples:

Continuous responses

1. $f(Y_{ij}|H_{ij})$ has a normal distribution
2. $E(Y_{ij}|H_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_1 Y_{i,j-1} + \alpha_2 Y_{i,j-2}$
3. $var(Y_{ij}|H_{ij}) = \sigma^2$

This is a second order Markov model.

Binary responses

1. $f(Y_{ij}|H_{ij})$ has a Bernoulli distribution
2. $\text{logit}(E(Y_{ij}|H_{ij})) = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_1 Y_{i,j-1}$
3. $var(Y_{ij}|H_{ij}) = E[Y_{ij}|H_{ij}](1 - E[Y_{ij}|H_{ij}])$

This is a first order Markov chain model.

Count data

1. $f(Y_{ij}|H_{ij})$ has a Poisson distribution

2. $\log(E(Y_{ij}|H_{ij})) = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_1\{\log(Y^*_{i,j-1}) - \mathbf{X}_{i,j-1}\boldsymbol{\beta}\}$

   where $Y^*_{ij} = max(Y_{ij}, k), 0 < k < 1.$

   The constant k prevents Y_{ij} becoming an 'absorbing state', with $Y_{i,j-1} = 0$ forcing all future responses to be 0.

3. $var(Y_{ij}|H_{ij}) = E(Y_{ij}|H_{ij})$

Note that if we assume

$$\log(E(Y_{ij}|H_{ij})) = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_1 Y_{i,j-1}$$

then

$$E[Y_{ij}|H_{ij}]) = \exp(\mathbf{X}_{ij}\boldsymbol{\beta})\exp(\alpha_1 Y_{i,j-1})$$

and the conditional mean grows exponentially over time when $\alpha_1 > 0$.

On the other hand, when $\alpha_1 < 0$ the model describes negative correlation among the repeated responses.

Therefore, this model has limited use for analyzing longitudinal count data and the alternative specification of the time dependence is required.

Interactions in Transitional Models

Note that we could allow completely separate logistic regressions for children with and without infections at the previous occasion:

$$\text{logit}(Pr[Y_{ij} = 1|Y_{i,j-1} = 0]) = \mathbf{X}_{ij}\boldsymbol{\beta}_0$$

$$\text{logit}(Pr[Y_{ij} = 1|Y_{i,j-1} = 1]) = \mathbf{X}_{ij}\boldsymbol{\beta}_1$$

This model can be written as

$$\text{logit}(Pr(Y_{ij} = 1|Y_{i,j-1}) = \mathbf{X}_{ij}\boldsymbol{\beta}_0 + Y_{i,j-1} * (\mathbf{X}_{ij}\boldsymbol{\alpha})$$

where $\boldsymbol{\alpha} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$.

From this perspective, the model

$$\text{logit}(Pr(Y_{ij} = 1|Y_{i,j-1}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha Y_{i,j-1}$$

can be seen to make a fairly strong assumption.

That is, it asserts that the effect of other covariates on risk is the same for those with and without infections at the previous occasion.

In most practical applications of transitional models, potential interactions between covariates and the lagged responses should be examined.

Statistical Inference

Recall that the joint distribution of $(Y_{i1}, ..., Y_{ip})$ can always be expressed as a series of conditional distributions,

$$f(Y_{i1}, Y_{i2}, ..., Y_{ip}) = f(Y_{i1}) f(Y_{i2}|Y_{i1}) \cdots f(Y_{ip}|Y_{i1}, ..., Y_{i,p-1}).$$

This provides a complete representation of the joint distribution.

Note that when the conditional distributions satisfy the first order Markov assumption, this reduces to

$$f(Y_{i1}) f(Y_{i2}|Y_{i1}) f(Y_{i3}|Y_{i2}) \cdots f(Y_{ip}|Y_{i,p-1}) = f(Y_{i1}) \prod_{j=2}^{p} f(Y_{ij}|Y_{i,j-1}).$$

Thus, the joint distribution is the product of the marginal distribution at time 1 and $p-1$ conditional distributions, all of which have the same form.

Since the marginal model at time 1 cannot be derived from the conditional model, we maximize only the second part involving the conditional distributions.

That is, statistical inference is based on the *conditional likelihood*.

For example, with the first order Markov model inference is based on

$$f(Y_{i2}, ..., Y_{ip}|Y_{i1}) = \prod_{j=2}^{p} f(Y_{ij}|Y_{i,j-1}).$$

Conditional MLEs may be less efficient, but do not require additional assumptions concerning the distribution of 'initial responses'.

When maximizing the conditional likelihood, it can be shown that estimation proceeds as in ordinary generalized linear models for independent responses.

Thus, we can use existing statistical software for generalized linear models by simply regressing Y_{ij} on an extended vector of covariates,

$$\{\mathbf{X}_{ij}, f_1(H_{ij}), f_2(H_{ij}), ..., f_s(H_{ij})\}.$$

Note: When $g(E[Y_{ij}|H_{ij}])$ is not a linear function of β and α then a slightly modified algorithm is required.

Finally, note that the empirical or so-called robust variance has a potential role in transitional models.

For example, if the conditional mean has been correctly specified but the conditional variance has not, fitting the conditional model with empirical variances will guard against such departures from the model assumptions.

Example: Six Cities Study of Respiratory Illness in Children.

A non-randomized longitudinal study of the health effects of air pollution.

Subset of data from one of the participating cities: Steubenville, Ohio

Outcome variable: Binary indicator of respiratory infections in child.

Measurements on the children were taken annually at ages 7, 8, 9, and 10.

Assume conditional probability of infection follows a logistic autoregressive model,

$$\text{logit}(E[Y_{ij}|H_{ij}]) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_i + \alpha_1 Y_{i,j-1}$$

where age_{ij} = child's age - 9, and $\text{smoke}_i = 1$ if child's mother smokes, 0 otherwise.

Preliminary analyses allowed for additional interactions between the lagged response and covariates.

None of these interactions were found to be statistically significant.

SAMPLE SAS CODE FOR FITTING TRANSITIONAL MODELS

```
data trans;
   set child;
   y=y2;
   age=-1;
   lagy=y1;

   output;
   y=y3;
   age=0;
   lagy=y2;

   output;
   y=y4;
   age=1;
   lagy=y3;
   output;

proc genmod data=trans;
   model y=age smoke lagy / d=bin;
run;
```

Transitional Model Parameter Estimates

Parameter	Estimate	SE	Z
Intercept	-2.478	0.117	-21.18
Age	-0.243	0.090	-2.70
Smoke	0.296	0.155	1.91
$Y_{i,j-1}$	2.211	0.187	11.82

The parameter $\exp(\alpha_1)$ is the ratio of odds of infection among those children who did and did not have an infection on the previous occasion.

The estimate of $\exp(2.211) = 9.21$ suggests strong time dependence.

Interpretation of β_2:

Given a child's infection status at the previous occasion, the conditional odds of infection among children of mothers who smoke is 1.34 ($e^{0.296}$) times that of children whose mothers do not smoke.

Note: Interpretation of β_2 depends on the first order Markov assumption.

If a second order Markov chain model is assumed, where

$$\text{logit}(E[Y_{ij}|H_{ij}]) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_i + \alpha_1 Y_{i,j-1} + \alpha_2 Y_{i,j-2}$$

the estimate of β_2 is 0.174.

Note, however, that the interpretation of β_2 has also changed.

Given a child's infection status at the *previous two occasions*, the conditional odds of infection among children of mothers who smoke is 1.20 ($e^{0.174}$) times that of children whose mothers do not smoke.

Caution: When we treat previous responses as explanatory variables, inferences about the covariates can be quite sensitive to the time-dependency model.

Hence, the sensitivity of the analysis to the time-dependency model should be checked.

The conditional model can also be potentially misleading if a treatment or exposure changes risk throughout the follow-up period, so that the conditional risk, given previous health status, is altered somewhat less strikingly.

Contrasting Models for Longitudinal Data

In the first part of the course, we focused on methods for analyzing longitudinal data where the dependent variable was continuous and the vector of responses was assumed to have a multivariate normal distribution.

We also focused on fitting a linear model to the repeated measurements.

For the remainder of the course we have considered a much wider class of regression models.

These models can be thought of as extensions of the generalized linear model to longitudinal data.

The main focus has been on discrete response data, e.g. count data or binary responses.

We have considered three main extensions of generalized linear models:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

Recall that these three quite different analytic approaches arise from somewhat different specifications of, or assumptions about, the joint distribution,

$$f(Y_{i1}, Y_{i2}, ..., Y_{ip}).$$

Unlike linear models for continuous responses, with non-linear models for discrete data different assumptions about the source of the correlation can lead to regression coefficients with quite distinct interpretations.

Marginal Models

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

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

In the marginal model, we model the regression of the response on covariates and the covariance structure separately.

That is, the mean and within-subject correlation are modelled separately.

Mixed Effects Models

The basic premise is that we assume that there is natural heterogeneity across individuals in a subset of the regression parameters.

That is, a subset of the regression parameters (e.g. intercepts) 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.

These models extend the conceptual approach of the linear mixed effects model and are most useful when the scientific objective is to make inferences about individuals rather than the population averages.

That is, the main focus is on the individual and the influence of covariates on the individual.

Transitional (Markov) Models

In transitional models the conditional distribution of each response is expressed as an explicit function of the past responses (the 'history') and covariates.

Therefore the correlation among the repeated responses can be thought of as arising due to past values of the responses explicitly influencing the present observation (i.e. the 'present' depends on the 'past').

In transitional models, the covariates and lagged responses are treated on an equal footing.

We can compare and contrast these three analytic approaches in a variety of different ways:

1. **Likelihood-based inference:**

   - traditional ML methods are (in principle) straightforward for the mixed effects and transitional models
   - there is no unified likelihood-based approach for marginal models, partly due to the absence of a 'convenient' likelihood function for discrete responses

2. **Sensitivity to time-dependence assumption:**

   - in marginal models, $\hat{\boldsymbol{\beta}}$ is relatively robust
   - in mixed effects models, $\hat{\boldsymbol{\beta}}$ is relatively insensitive to assumptions concerning the distribution of the random effects
   - in transitional models, $\hat{\boldsymbol{\beta}}$ is not robust since covariates and lagged responses are treated symmetrically

3. **Interpretation of $\boldsymbol{\beta}$:** Consider β_2 in the Six Cities example.

- marginal model: β_2 describes the ratio of population odds.
  '...prevalence or odds of infection is 1.3 times higher among children whose mothers smoke...'

- mixed effects model: β_2 describes the ratio of a specific individual's odds.
  '...odds of infection is 1.5 times higher for a child whose mother starts smoking...'

- transitional model: β_2 describes the ratio of conditional odds.
  '...given the history of infections at the previous two occasions, the odds of infection is 1.2 times higher among children whose mothers smoke...'

Consider the interpretation of β_2 from the mixed effects model:

'...odds of infection is 1.5 times higher for a child whose mother starts smoking...'

However, because smoking is a between-subject covariate, there is no information in the data that directly measures the effect a within-subject change in smoking would have on the odds of infection.

So how is this numerical estimate calculated?

Answer: It's a function of both the

- marginal (population-averaged) effect
  (which is directly observable from the data)

- the distributional form assumed for the random effects, $\mathbf{b}_i$.

$\Rightarrow \beta$ for a between-subject covariate can be sensitive to the assumptions for the random effects.

Example: Random Intercept Logistic Regression Model

$$\widehat{\beta}_{MIXED} \approx \left[\left\{16(\sqrt{(}3))/(15\pi)\right\}^2 \sigma^2 + 1\right]^{-1/2} \widehat{\beta}_{PA}$$

Estimates of 'maternal smoking effect' using the Six Cities data.

	Marginal	**Mixed Effect**	**Transitional**	
β_2	0.256	0.399	$0.296^{\dagger}$	$0.174^{\dagger\dagger}$
$\exp(\beta_2)$	(1.30)	(1.50)	(1.34)	(1.20)

† : Based on first order Markov model

†† : Based on second order Markov model

Choice among models?

- should be guided by specific scientific question of interest:

- Marginal Model: Population-averaged effect
  - Interest focuses on differences between sub-groups in the study population
  - Applicable for either type of covariate (Both time-varying and time-invariant).

- Mixed Model: Within-subject effect
  - Interest focuses on estimating intercepts or slopes for each subject
  - Interest focuses on controlling for unmeasured subject characteristics.
  - Best used when interest focuses on a time-varying covariate.

- Transitional Model: How does the "present" depend on the "past"?

MULTILEVEL MODELS

Until today, this course has focused on the analysis of longitudinal data.

Mixed models can also be used to analyze multilevel data.

Longitudinal data are clustered. Individual subjects, or units, provide multiple observations.

Multilevel data are also clustered.

> Randomized trials study patients clustered within practices.
> In studies of social determinants of health, we may study families of individuals clustered by neighborhood within a community.

The first example is a two-level setting, patients within practices. The second is a three-level setting, individuals within families within neighborhoods.

Multilevel Data

Multilevel data can arise from the study design or a natural hierarchy in the target population, or sometimes both.

In developmental toxicity studies, mothers are dosed with the study chemical and outcomes are measured in the offspring. Observations are naturally clustered within litter.

In studies of clustering of disease within families, we measure the disease status of each family member.

Other naturally occurring clusters include schools, households, hospital wards, medical practices, and neighborhoods.

Multi-stage sampling design.
NHANES III chose a random sample of primary sampling units, a random sample of areas within each PSU, a random sample of households in each area, and a random individual within each household (four-level sampling).

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.

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.

We introduce the ideas with the two-level model, then 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 (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. (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, $\mathbf{X}_{ij}$

Consider the following model for the mean:

$$E(Y_{ij}) = \mathbf{X}_{ij}\beta$$

In a multi-center clinical trial comparing two treatments, we might assume that:

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

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

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + b_j + e_{ij}$$

Or, more generally,

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_j + e_{ij}$$

with more than 1 random effect.

Features of the Two-Level Linear Model

1. The model defines two sources of variation. The magnitudes of the within- and between-cluster correlation determine the degree of clustering.

2. For a given level 2 unit, the random effects are assumed constant across level 1 units.

3. The conditional expection of Y_{ij}, given the identity of the level 2 group, is
$$\mathbf{X}_{ij}\beta + \mathbf{Z}_{ij}\mathbf{b}_j$$

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.

The Three-Level Linear Model

Now 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. Denote the response at the i^{th} observation of the j^{th} patient in the k^{th} clinic by

$$Y_{ijk}$$

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:

$$Y_{ijk} = \mathbf{X}_{ijk}\beta + \mathbf{Z}^{(3)}_{ijk}\mathbf{b}^{(3)}_{k} + \mathbf{Z}^{(2)}_{ijk}\mathbf{b}^{(2)}_{jk} + e_{ijk}$$

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.

Then a hierarchical three-level model for the response is given by

$$Y_{ijk} = \beta_1 + \beta_2 t_{ijk} + \beta_3(\text{Trt}_{ij} \times t_{ijk}) + \mathbf{b}_k^{(3)} + \mathbf{b}_{jk}^{(2)} + e_{ijk}$$

This model assumes a common intercept and separate linear trends over time in the two treatment groups.

If

$$\text{Var}(\mathbf{b}_k^{(3)}) = \mathbf{G}^{(3)}, \text{Var}(\mathbf{b}_{jk}^{(2)}) = \mathbf{G}^{(2)}, \text{ and } \text{Var}(e_{ijk}) = \sigma^2,$$

and all random effects are assumed to be independent, then

$$\text{Var}(Y_{ijk}) = \mathbf{G}^{(2)} + \mathbf{G}^{(3)} + \sigma^2$$

and the covariance between two observations from the same patient is

$$\mathbf{G}^{(2)} + \mathbf{G}^{(3)}$$

Thus, the observations for a given patient have an intraclass correlation structure, with

$$\text{Corr}(Y_{ijk}, Y_{ijl}) = \frac{\mathbf{G}^{(2)} + \mathbf{G}^{(3)}}{\mathbf{G}^{(2)} + \mathbf{G}^{(3)} + \sigma^2}.$$

Because this is a linear mixed model,

$$E(Y_{ijk}) = \beta_1 + \beta_2 t_{ijk} + \beta_3(\text{Trt}_{ij} \times t_{ijk})$$

Estimation

For the three-level linear model, the standard distributional assumptions are that:

$$\mathbf{b}_k^{(3)} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}^{(3)}), \mathbf{b}_{jk}^{(2)} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}^{(2)}), \text{ and } e_{ijk} \sim \mathbf{N}(\mathbf{0}, \sigma^2)$$

Given these assumptions, estimation of the model parameters is relatively straightforward. The GLS estimate of β is given by

$$\widehat{\beta} = \left\{ \sum_{k=1}^{n_3} (\mathbf{X}_k' V_k^{-1} \mathbf{X}_k) \right\}^{-1} \sum_{k=1}^{n_3} (\mathbf{X}_k' V_k^{-1} \mathbf{Y}_k)$$

where $\mathbf{Y}_k$ is a column vector of length $\sum_{j=1}^{n_{2k}} n_{1jk}$, the number of observations in the k^{th} cluster. $\mathbf{X}_k$ is the corresponding matrix of covariates, and V_k is the covariance matrix of $\mathbf{Y}_k$.

Estimation (Continued)

As before, we use REML (or ML) to obtain estimates of $\mathbf{G}^{(3)}$, $\mathbf{G}^{(2)}$, 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 SAS PROC MIXED.

It is also available in MLwiN and HLM, two stand-alone programs developed for multi-level modeling.

Example: Developmental Toxicity of Ethylene Glycol

In a classic developmental study, ethylene glycol at doses of 0, 750, 1,500, or 3,000 mg/kg/day was administered to 94 pregnant mice. The crude results were as follows:

Dose				Weight (gm)	
(mg/kg)	Sqrt(Dose/750)	Dams	Fetuses	Mean	St. Dev.
0	0	25	297	0.97	0.10
750	1	24	276	0.90	0.10
1500	1.4	22	229	0.76	0.11
3000	2	23	226	0.70	0.12

Based on experience and these data, the investigators modeled the response as linear in sqrt(dose).

Because the observations are clustered within dam, the analysis must take account of clustering. If it does not, the sample size for comparisons between doses will be exaggerated.

To fit a two-level model that is linear in sqrt(dose), use the following SAS code:

```
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
<u>Fixed Effects</u>			
Intercept	0.98	0.02	61.3
Newdose	-0.13	0.01	-10.9
<u>Random Effects</u>			
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.

Example: 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		THKS	
CC	TV	n	Mean	Std Dev	Mean	Std Dev
No	No	421	2.15	1.18	2.34	1.09
No	Yes	380	2.05	1.29	2.82	1.09
Yes	No	416	2.09	1.29	2.48	1.14
Yes	Yes	383	1.98	1.29	2.74	1.07

The mean value of Pre-THKS does not differ significantly among treatment groups.

The Model

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + b_k^{(3)} + b_{jk}^{(2)} + e_{ijk}$$

where we list fixed and random effects on separate lines for clarity.

In a slightly modified notation, assume

$$\begin{aligned} e_{ijk} &\sim N(0, \sigma_1^2) \\ b_{jk}^{(2)} &\sim N(0, \sigma_2^2) \\ b_k^{(3)} &\sim N(0, \sigma_3^2) \end{aligned}$$

This is the standard hierarchical (or multi-level) linear model with random effects at each level to introduce correlation within clusters, The fixed effects model has both main effects and interactions for CC and TV.

SAS Code

```
data tvandcc;
  infile 'c:\bio226\datasets\tv.txt';
  input sid cid cc tv baseline THKS;
  level=cc+2*tv;
run;
proc mixed data=tvandcc covtest;
  class sid cid;
  model y2 = y1 cc tv cctv / s;
  random intercept / subject=sid  g ;
  random intercept / subject=cid  g ;
run;
```

Fixed and Random Effects Estimates

Variable	Estimate	Standard Error	Z
<u>Fixed Effects</u>			
Intercept	1.70	0.13	13.6
Pre-THKS	0.31	0.03	11.8
CC	0.64	0.16	4.0
TV	0.18	0.16	1.2
CC $\times$ TV	-0.33	0.22	-1.5
<u>Random Effects</u>			
Level 3 Variance (σ_3^2)	0.04	0.03	1.5
Level 2 Variance (σ_2^2)	0.07	0.03	2.3
Level 1 Variance (σ_1^2)	1.60	0.06	27.1

Comments on the Estimates of Fixed Effects

Pre-THKS is an important predictor of knowledge after the intervention.

CC had a clear effect on knowledge, but TV did not.

Comments on the Estimates of Random Effects

There is relatively little clustering as measured by the small values for the level 2 and 3 variances as compared to the level 1 variance.

The variability among classrooms is twice as large as the variability among schools.

The correlation among children in the same classroom is

$$(0.04 + 0.07)/(0.04 + 0.07 + 1.60) = 0.06$$

The Effect of Ignoring Clustering

Because the correlations are small, we might conclude that the clustering is unimportant. But consider an analysis that treats the observations as independent, ignoring clustering.

Variable	Estimate	Standard Error	Z
Intercept	1.66	0.08	19.7
Pre-THKS	0.33	0.03	12.6
CC	0.64	0.09	7.0
TV	0.20	0.09	2.2
CC $\times$ TV	-0.32	0.13	-2.5

The estimates change little but the model-based standard errors are too small, leading to erroneous conclusions.

If we assume independence, we are implicitly assuming a larger sample size for comparisons between classrooms.

Generalizations

The multi-level model can be generalized to an arbitrary number of levels.

Generalized Linear Mixed Effects models have also been developed for the analysis of binary outcomes and counts in the multi-level 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 these limitations, multi-level models are now widely used.

In both designed experiments and studies of the effects of family and community factors that influence health and well-being, multi-level models provide a usually effective approach to data analysis that accounts for correlations induced by clustering.

Multi-level models are, in one sense, no different than longitudinal models. Unlike logistic regression and survival analysis, where the concept of regression analysis can be applied quite robustly and with few choices, longitudinal and multi-level analysis require more careful thought about the choice and meaning of models.

This is both their challenge and their reward.

APPENDIX

Some comments on denominator df in PROC MIXED

PROC MIXED reports both t and F statistics for testing hypotheses concerning the fixed effects.

The t statistics are Wald test statistics (estimate/s.e.). In large samples, these have a standard normal distribution.

The F test for a class variable, reported in PROC MIXED, is based on a multivariate Wald-type test statistic divided by its numerator df. In large samples, the multivariate Wald statistic has a chi-squared distribution with numerator df.

To obtain p-values based on the chi-squared distribution, simply add the option CHISQ to the MODEL statement.

(Recall: If Z has a $N(0,1)$ distribution; Z^2 has a $\chi^2_{(1)}$ distribution).

Review: Multivariate Wald Test Statistics

To test a null hypothesis that involves more than one parameter, a multivariate Wald-type test statistic can be constructed.

To test $H_0 : \boldsymbol{\beta} = 0$, where $\boldsymbol{\beta}$ is a vector of k regression parameters, the following statistic can be computed

$$W = \widehat{\boldsymbol{\beta}}' Cov(\widehat{\boldsymbol{\beta}})^{-1} \widehat{\boldsymbol{\beta}}.$$

Under $H_0 : \boldsymbol{\beta} = 0$, W has, approximately, a chi-squared distribution with k df.

PROC MIXED constructs t and F tests for the fixed effects using approximate denominator df.

Except for certain special cases, there is no general theory to justify the use of the t and F distributions for tests of the fixed effects. Because of this, there is also no obvious way to obtain the required denominator degrees of freedom for the t and F tests.

PROC MIXED provides 5 options for computing denominator degrees of freedom (see pages 2117-2119 of the manual for a detailed description):

1. Residual method (DDFM=RESIDUAL)

2. Between-Within method (DDFM=BETWITHIN or DDFM=BW)

3. Containment method (DDFM=CONTAIN)

4. Satterthwaite's approximation (DDFM=SATTERTH)

5. Kenward and Roger approximation (DDFM=KENWARDROGER)

Of the 5 methods, (1) is by far the least appealing; (2) and (3) are the default methods depending on whether the model uses a REPEATED or RANDOM statement respectively; and (4) and (5) are computationally intensive methods.

Caveat: Reliance on the default methods can often lead to different denominator df for two identical models that have been formalized in somewhat different ways.

For example, the denominator df in a model with a compound symmetry covariance matrix will differ depending on whether the REPEATED statement (with TYPE=CS) or the RANDOM statement (with RANDOM INTERCEPT) has been used.

Example 1: Exercise Therapy Study (Compound Symmetry: REPEATED/TYPE=CS)

Fit Statistics

Res Log Likelihood	-330.2
Akaike's Information Criterion	-332.2
-2 Res Log Likelihood	660.4

Tests of Fixed Effects

Source	NDF	DDF	ChiSq	F	Pr > ChiSq	Pr > F
TRT	1	35	1.28	1.28	0.2577	0.2654
TIME	1	134	48.80	48.80	0.0001	0.0001
TIME*TRT	1	134	0.40	0.40	0.5263	0.5274

Example 2: Exercise Therapy Study (Compound Symmetry: RANDOM INTERCEPT)

Fit Statistics

Res Log Likelihood	-330.2
Akaike's Information Criterion	-332.2
-2 Res Log Likelihood	660.4

Tests of Fixed Effects

Source	NDF	DDF	ChiSq	F	Pr > ChiSq	Pr > F
TRT	1	134	1.28	1.28	0.2577	0.2598
TIME	1	134	48.80	48.80	0.0001	0.0001
TIME*TRT	1	134	0.40	0.40	0.5263	0.5274

Summary

PROC MIXED constructs t and F tests for the fixed effects using approximate denominator df.

There is no general theory to justify the use of the t and F distributions for tests of the fixed effects.

In general, when the sample size is moderately large (say, greater than 100), the denominator df computed using any of the 5 methods will be sufficiently large that the p-values obtained by comparing the Wald statistics to t and F distributions will not differ discernibly from the p-values obtained from the corresponding standard normal and chi-squared distributions.