BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 4

Statistical Basis of Longitudinal Analysis

Statistical Basis of Longitudinal Analysis (Part 1)

Overview:

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

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

We introduce some additional vector and matrix notation.

We present a general linear regression model for longitudinal data.

Single-Group Repeated Measures Design

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

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

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

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

Listing each observation at the n occasions:

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

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

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

where

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

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

Example: Treatment of Lead-Exposed Children Trial

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

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

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

Then the null hypothesis of interest is

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

How can we test this hypothesis?

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

This approach allows only pairwise comparisons.

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

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

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

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

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

Later we will consider the statistical basis for this analysis.

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

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

PROC MIXED in SAS

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

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

Multivariate (or Wide) Form of Succimer Data

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

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

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

Selected Output from PROC MIXED

```
                    The Mixed Procedure

               Estimated R Matrix for id 1

Row         Col1         Col2         Col3         Col4

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

```
Covariance Parameter Estimates

Cov Parm     Subject    Estimate

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

```
                   Fit Statistics

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


              Null Model Likelihood Ratio Test

                DF    Chi-Square      Pr > ChiSq

                 9         86.73          <.0001
```

```
                   The Mixed Procedure

                Solution for Fixed Effects

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

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

```
                           The Mixed Procedure

                     Type 3 Tests of Fixed Effects

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

time           3      49       163.72     54.57         <.0001   <.0001


                               Contrasts

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

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

Covariance Structure

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

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

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

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

PROC GLM versus PROC MIXED in SAS

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

Selected Output from PROC GLM

```
                         The GLM Procedure

Dependent Variable: y

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

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



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

 time                        3  5104.418150   1701.472717    29.43  <.0001
```

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

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

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

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

Selected Output from PROC MIXED

```
                        Solution for Fixed Effects

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

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

```
                          The Mixed Procedure

                    Type 3 Tests of Fixed Effects

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

time           3      49        163.72      54.57         <.0001    <.0001



                               Estimates

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

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

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

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

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

Notation of General Linear Model

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

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

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

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

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

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

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

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

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

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

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

Linear Models for Longitudinal Data

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

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

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

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

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

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

Vector and Matrix Representation

Note that the linear model

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

describes the mean response at all n_i occasions.

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

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

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

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

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

Note that the model

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

is simply a shorthand representation for

$$E\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}.$$

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

Illustration: *Treatment of Lead-Exposed Children*

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

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

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

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

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

Assume two treatment groups have different intercepts and slopes:

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

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

Thus, for children in the placebo group

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

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

Similarly, for children in the succimer group

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

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

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

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

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

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

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

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

This individual was assigned to treatment with placebo.

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

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

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

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

This individual was assigned to treatment with succimer.

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

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

where, for example,

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

for children in the placebo group.

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

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

for children in the placebo group, and

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

for children in the succimer group.