BIO 226: APPLIED LONGITUDINAL ANALYSIS

INSTRUCTOR: BRENT COULL

Department of Biostatistics

Harvard School of Public Health

Considerable Thanks to Garrett Fitzmaurice for the Course Notes,

and Earlier Instructors Notably Jim Ware and Michael Hughes

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 1

INTRODUCTION TO THE COURSE

Introduction

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

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

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

Theoretical basis for results mentioned but not developed.

No calculus or matrix algebra is assumed.

Features of Longitudinal Data

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

Longitudinal studies allow direct study of change over time.

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

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

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

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

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

Complications:

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

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

Correlation arises due to repeated measures on the same individuals.

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

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

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

Relation to Correlated Data

Correlated data commonly arise in many applications.

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

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

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

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

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

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

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

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

Table 2: Mean blood lead levels (and standard deviation) at baseline, week 1, week 4, and week 6.

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

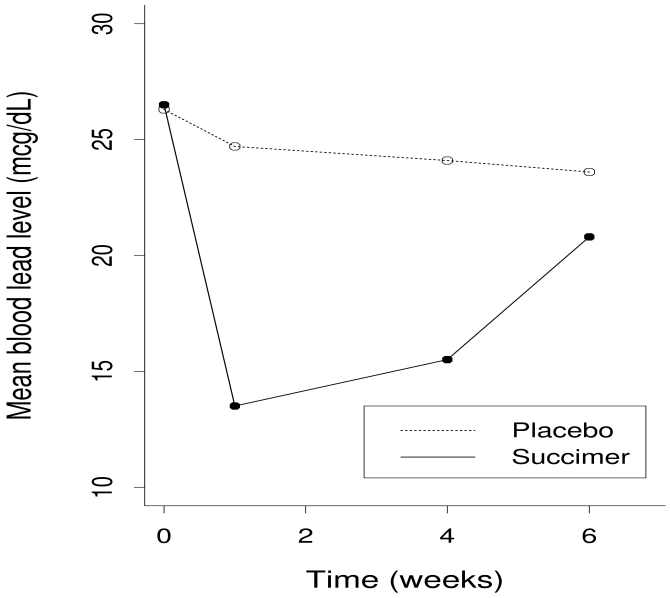

Figure 1: Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

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

- Longitudinal study designed to characterize lung function growth in children and adolescents.
- Most children were enrolled between the ages of six and seven and measurements were obtained annually until graduation from high school.
- Focus on a randomly selected subset of the 300 female participants living in Topeka, Kansas.
- Response variable: Volume of air exhaled in the first second of spirometry manoeuvre, FEV_1.

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

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

Note: Time represents time since entry to study.

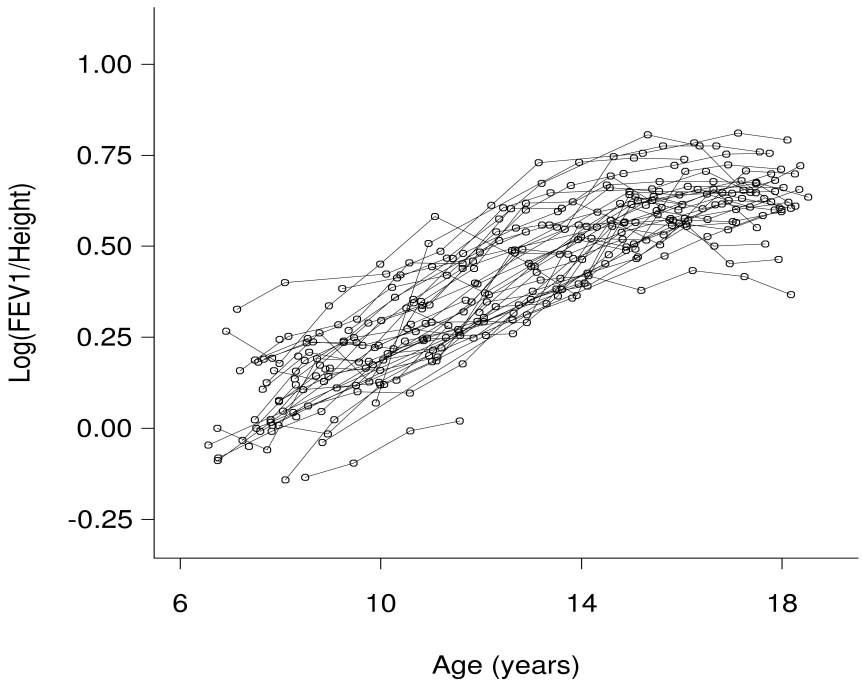

Figure 2: Timeplot of $\log(FEV_1/\text{height})$ versus age for 50 randomly selected girls from the Topeka data set.

Example 3: *Influence of Menarche on Changes in Body Fat*

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

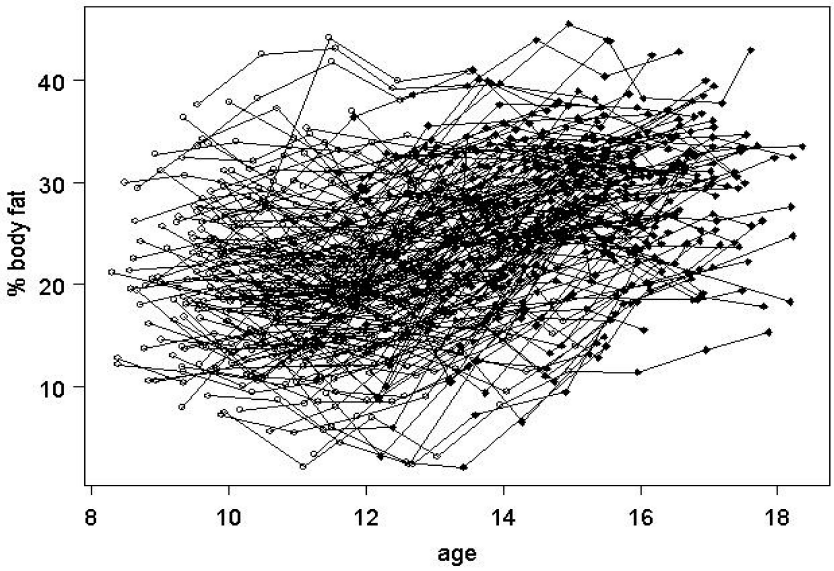

Figure 3: Timeplot of percent body fat against age (in years).

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

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

Note: measurement protocol is the same for all girls.

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

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

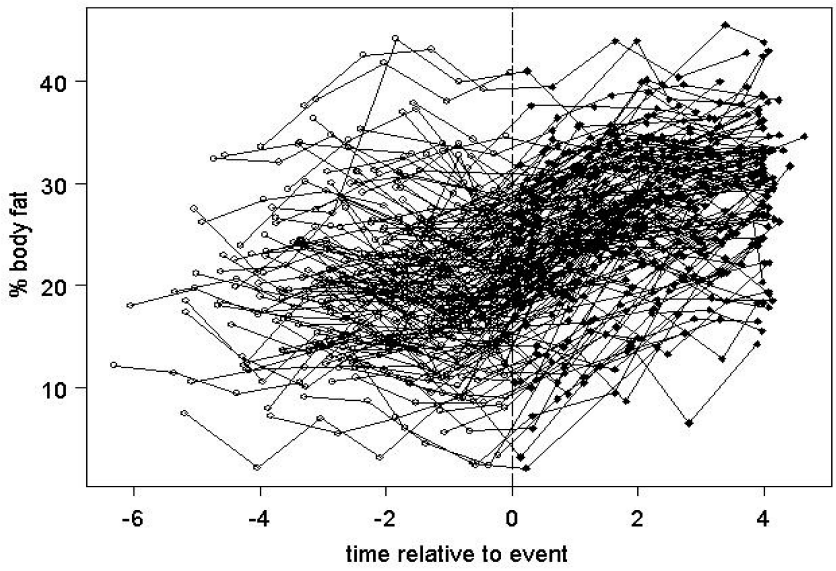

Figure 4: Timeplot of percent body fat against time, relative to age of menarche (in years).

Example 4: *Oral Treatment of Toenail Infection*

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

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

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

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

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

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

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

Outcome variable: Count of number of seizures.

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

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

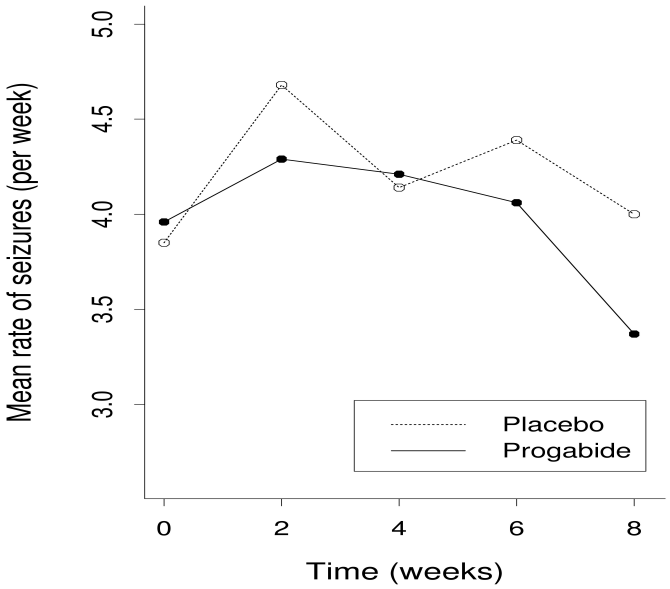

Figure 5: Mean rate of seizures (per week) at baseline, week 2, week 4, week 6, and week 8 in the progabide and placebo groups.

Terminology

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

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

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

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

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

Features of Longitudinal Data

In longitudinal studies the outcome variable can be:

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

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

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

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

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

Organization of Course

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

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

Organization of Course (cont.)

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

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

Background Assumed

1) Samples and populations

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

3) Variables:

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

4) Regression Models

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

5) Inference

Estimation, testing, and confidence intervals

6) Multiple linear regression/ANOVA
Multiple logistic regression