BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 3

LONGITUDINAL DATA - BASIC CONCEPTS

Longitudinal Data - Basic Concepts

Features of Longitudinal Data:

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

Longitudinal studies allow direct study of change over time.

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

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

Terminology

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

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

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

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

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

Correlation

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

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

Why are longitudinal data correlated?

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

Variability

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

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

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

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

Notation

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

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

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

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

Table 1: Tabular representation of longitudinal data, with n repeated observations on N individuals.

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

Vector Notation

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

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

Alternatively, we can denote the response vectors Y_i as

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

Correlation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Example: Treatment of Lead-Exposed Children Trial

We restrict attention to the data from placebo group.

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

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

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

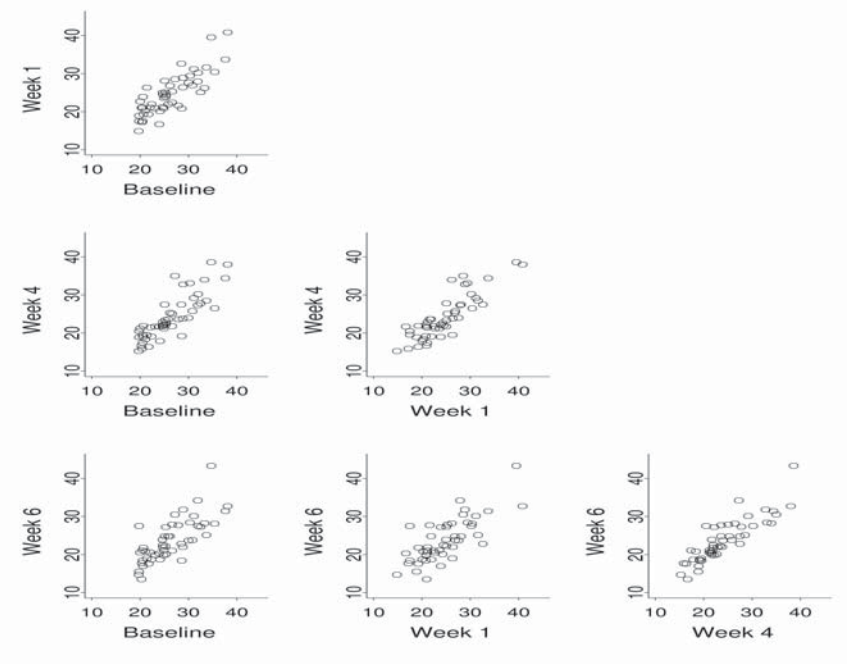

Figure 1: Pairwise scatter-plots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group.

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

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

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

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

Observations about Correlation in Longitudinal Data

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

(i) correlations are positive,

(ii) correlations decrease with increasing time separation,

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

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

Consequences of Ignoring Correlation

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

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

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

An estimate of change is given by

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

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

In the TLC trial, the estimate of change in the succimer group is -1.6 (or $24.7 - 26.3$).

For inferences, we also need a standard error.

Variance of $\widehat{\delta}$ is

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

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

Substituting estimates of the variances and covariance into this expression:

$$\widehat{\text{Var}}(\hat{\delta}) = \frac{1}{50}\{25.2 + 29.8 - 2(22.8)\} = 0.19.$$

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

Independence $\Longrightarrow$ zero covariance.

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

$$\frac{1}{50}(25.2 + 29.8) = 1.10,$$

which is almost six times larger.

In this illustration, ignoring the correlation results in:

- standard errors that are too large (1.05 versus 0.43)
- confidence intervals that are too wide
- p-values for the test of H_0: $\delta = 0$ that are too large

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

Summary

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

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

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