BIOS 755: Introductory concepts

Alexander McLain

January 9, 2024

Outline

Introduction

Lecture goals:

- Give an overview of the objectives of longitudinal analysis and discuss features in the data.
- Longitudinal analysis is concerned with estimating how individuals change throughout the study.
- Examine the factors that are related to differences among individuals over time.
- Review features of longitudinal study designs.
- Introduce notation.

Treatment of Lead-Exposed Chidren (TLC) Trial

Recall the TLC study (a balanced design)

- 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

Objectives of Longitudinal Analysis

- Defining feature of longitudinal study: two or more observations taken on (at least) some subjects.
- Multiple measurements over time allow assessment of within-individual changes in the response variable.
- Thus, some main goals are to:
  - characterize (a loaded word) the change in the response over time.
  - determine whether changes are related to exposures of interest

Longitudinal Analysis

- One way to achieve this is to look at *"change scores"* or *"difference scores,"* between pre- and post-treatment.
- There are different ways to form such a question
  1. are the *change scores* related to group.
  2. is the change in the scores related to covariates.
  3. is the final score related to covariates (path analysis).
- 1 could be answered using a paired t-test.
- 2 could be answered using standard linear regression techniques (possibly adjusting for the pre-score and differences in time).
- 3 also could use linear regression (adjusting for the pre-score and differences in time).

Longitudinal Analysis

- The usefulness of change score analysis is limited to situations with two measurements, with more than two measurements you will have more than 1 change score.
- Analyses with change scores should be used with caution as they
  - have been shown to lead to bias in observational (non-randomized) studies (reference), and
  - require complete data or multiple imputation.

Number of measurements

- The number of measurements can vary greatly from study to study and can be equally or unequally separated in time.
  - Pain measured every 15 minutes for 2 hours.
  - Height measured every 3-months until 3 yr old, then every year thereafter.
- Both of the above are considered to be **balanced**.
- **Unbalanced** data are very common in the health sciences.
  - individuals will miss schedule visits,
  - visits are not made exactly at the scheduled date,
  - timings are relative to a benchmark event (e.g. relative to sero-conversion), or
  - visits are themselves random (common in retrospective data).
- Missing data are the rule, not the exception, so unbalanced data are the rule.

Data Notation

- Consider the following

$$Y_{ij} = \text{the } j\text{th measurement taken on unit } i.$$

  where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n$
- In the TLC data each child had four measurements at baseline, 1, 4 and 6 weeks.
- Y_{ij} represents the **random** response of the ith child at measurement j, for $j = 1, 2, 3, 4$.
- y_{ij} denotes the realized value of Y_{ij}.
- t_{ij} is the time of the observation of the ith child at measurement j, for all i

$$t_{i1} = 0, \quad t_{i2} = 1, \quad t_{i3} = 4, \quad t_4 = 6$$

Random Vectors

- It is convenient to represent all observations for a specific unit as a **random vector**.
- For the TLC data each child has,

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix}$$

  the random vector for child i.
- Also use $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$.
- In general, we have $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$.

Expectations and mean

- The mean, average, or "expectation" of each response is

$$E(Y_{ij}) = \mu_{ij}$$

μ_{ij} is the *conditional* mean at the jth occasion (i.e., conditional on covariate values).

Dependence and Correlation Introduction

- Two variables are said to be independent if the behavior of one variable does not depend on the value of another variable.
- For example, LDL cholesterol and sex are independent if the distribution of LDL cholesterol is the same for males and females.
- Longitudinal data methods do not make the assumption that the observations are independent.
- For example, if I tell you my LDL cholesterol at baseline was 80, what is a plausible range for this value at 4 weeks?
- What if it was 165 at baseline?

Variance and Covariance

- Along with the expectation we'll use variance

$$\text{var}(Y_{ij}) = E(Y_{ij} - \mu_{ij})^2 = \sigma_j^2$$

- The standard deviation is $\sqrt{\sigma_j^2} = \sigma_j$.
- **Covariance:** a measure of how two random variables vary together.
- Mathematically this is expressed as,

$$\text{cov}(Y_{ij}, Y_{ik}) = E\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\} = \sigma_{jk}$$

Covariance Matix

- The covariance matrix of $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$

$$\text{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{pmatrix} = \Sigma$$

- You will sometimes see $\sigma_j^2 = \text{var}(Y_{ij}) = \text{cov}(Y_{ij}, Y_{ij}) = \sigma_{jj}$.

Covariance to correlation

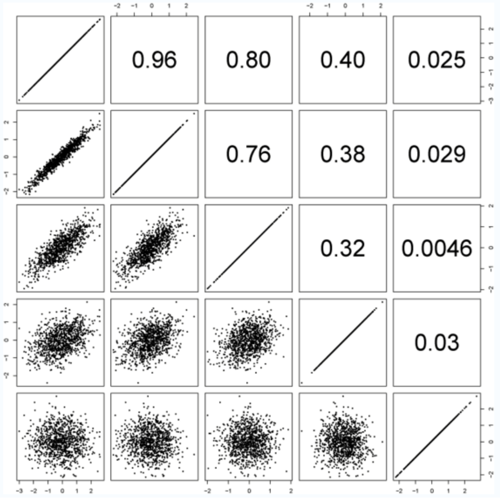
- The covariance values are hard to interpret since their magnitude is dependent on the variance of the variables.
- Covariance is commonly standardized to correlation.
- The population correlation of two elements is

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2 \sigma_k^2}}$$

- Which gives the correlation matrix

$$\mathrm{Corr}(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix}$$

Correlation Matix Example



Objectives of TLC Trial

- Goal: determine whether the new treatment reduces blood lead levels over time relative to placebo.
- Let μ_{jS} and μ_{jP} are the mean levels at occasion j for succimer and placebo groups, respectively.
- Different ways to answer this question:
  1. $H_0 : \mu_{jS} = \mu_{jP}$ for all $j = 1, \ldots, 4$.
  2. $H_0 : \mu_{jS} - \mu_{1S} = \mu_{jP} - \mu_{1P}$ for $j = 2, 3, 4$.

Correlation in TLC

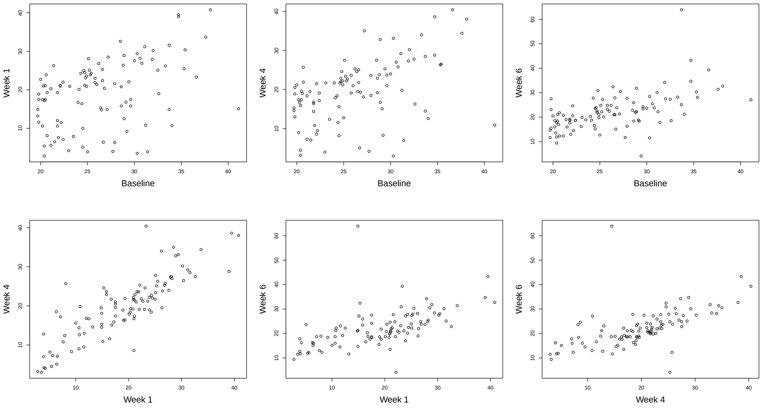


Figure: Correlation for 50 children in the placebo group

Estimated covariance & correlation matrices

- The estimated covariance matrix for the TLC

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} 25.2 & 22.8 & 24.3 & 21.4 \\ 22.8 & 29.8 & 27.0 & 23.4 \\ 24.3 & 27.0 & 33.1 & 28.2 \\ 21.4 & 23.4 & 28.2 & 31.8 \end{pmatrix}$$

- The estimated correlation matrix for the TLC

$$\mathrm{Corr}(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & 0.83 & 0.84 & 0.76 \\ 0.83 & 1 & 0.86 & 0.76 \\ 0.84 & 0.86 & 1 & 0.87 \\ 0.76 & 0.76 & 0.87 & 1 \end{pmatrix}$$

Time plot for TLC

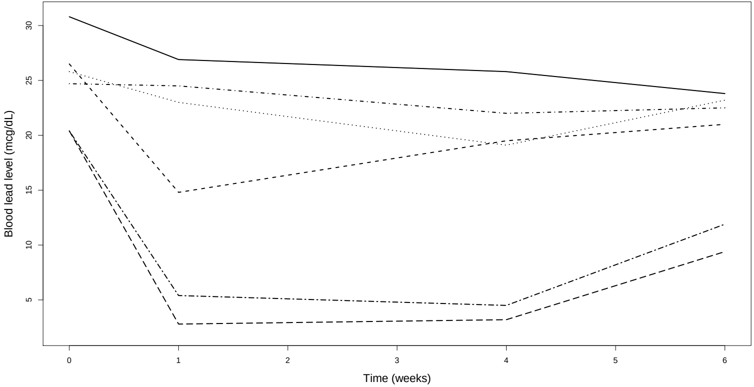


Figure: Time plot for 6 subjects

Correlation "truths"

1. the correlations are positive
2. the correlations often decrease with increasing time separation
3. the correlations between repeated measures rarely ever approach zero
4. the correlation between a pair of repeated measures taken very closely rarely approaches one.

Sources of variability

Three potential sources of variability

- between-individual heterogeneity
- within-individual biological variation
- measuement error