Introduction

Biostatistics 653

Applied Statistics III: Longitudinal Analysis

Academic Integrity

- All assignments you submit for evaluation must represent your own work
- If you copy or paraphrase other work, you must clearly mark these sections and indicate sources
- Cheating, plagiarism, and aiding and abetting these acts constitutes academic misconduct and is a serious offense
- See also the School policy on academic conduct
  - http://www.sph.umich.edu/academics/policies/conduct.html
- **In a set of assignments or exams that are broadly identical, each will be scored as zero and referred to Department or School.**

Longitudinal Study

- What is a longitudinal study?
- Why do we do longitudinal studies?
- Why do we need new statistical methods for longitudinal studies?

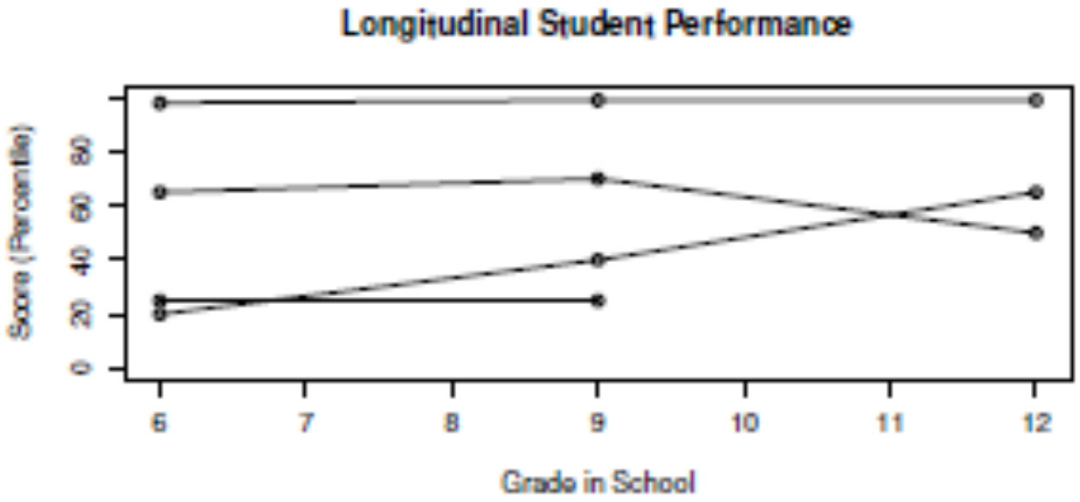
Example I

- To evaluate teachers' performance, we can:

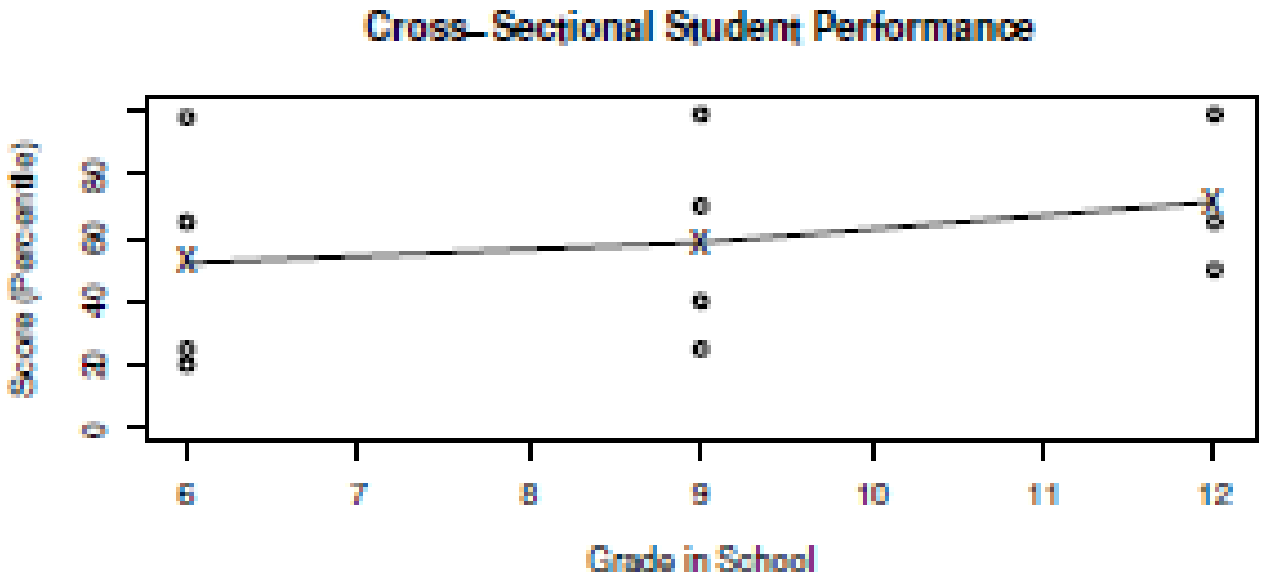
-- perform a longitudinal study, recruit students from 6^{th} grade, follow them until 12^{th} grade (except in the case of dropouts), and record their class performance (score) at 6^{th}, 9^{th} and 12^{th} grades.

-- perform a cross-section study, which is completed at one point in time, with the 6^{th}, 9^{th} and 12^{th} cohorts tested concurrently.

Longitudinal Study



Cross-sectional Study



Questions of Interest

- Question: are the 12^{th} grade teachers better than the 6^{th} or 9^{th} grade teachers

- or are test scores higher due to other factors (e.g. better teachers earlier in school, dropout of the poorest students, or a particularly bad group of 6^{th} graders?)

Generally

- In the cross-sectional design, we do not have any repeated scores from the same student. All observations may be treated as independent.

- In a longitudinal study, we can investigate

  -- changes over time within individuals

  -- differences among individuals in their response levels

  -- factors associated with such change or difference

Example II: Dental Study

- Changes in the distance (measured in mm) from the center of the pituitary gland to the pterygomaxillary fissure are important in orthodontic therapy

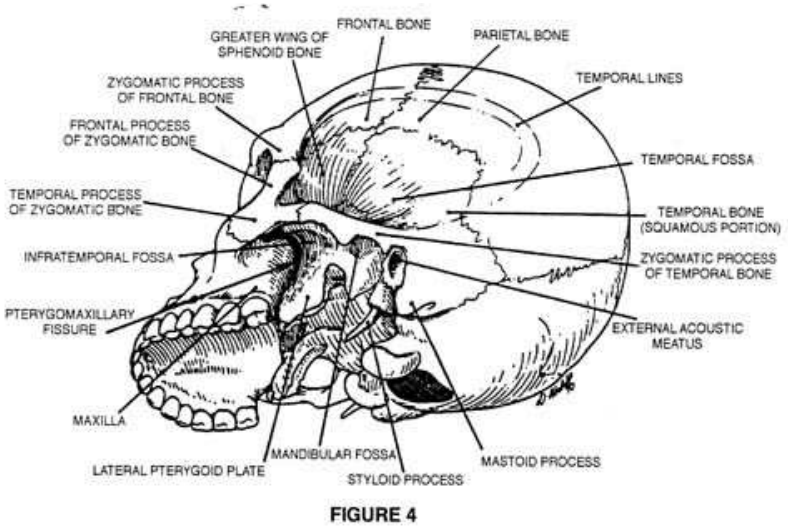
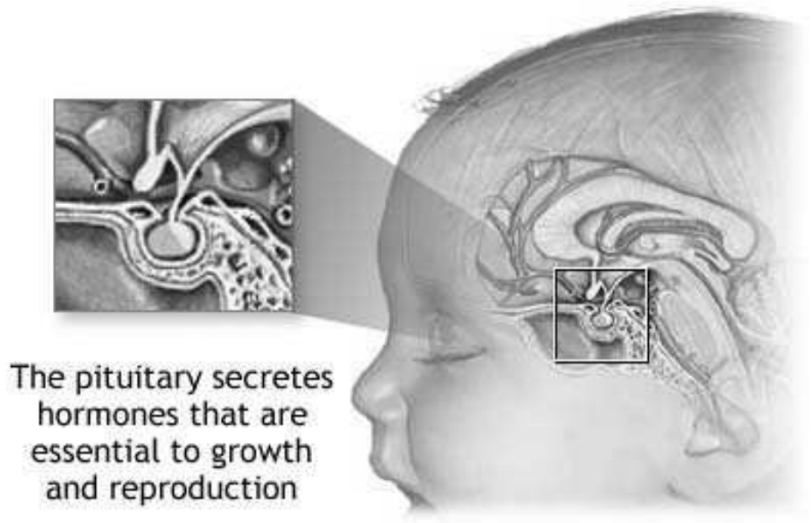


FIGURE 4

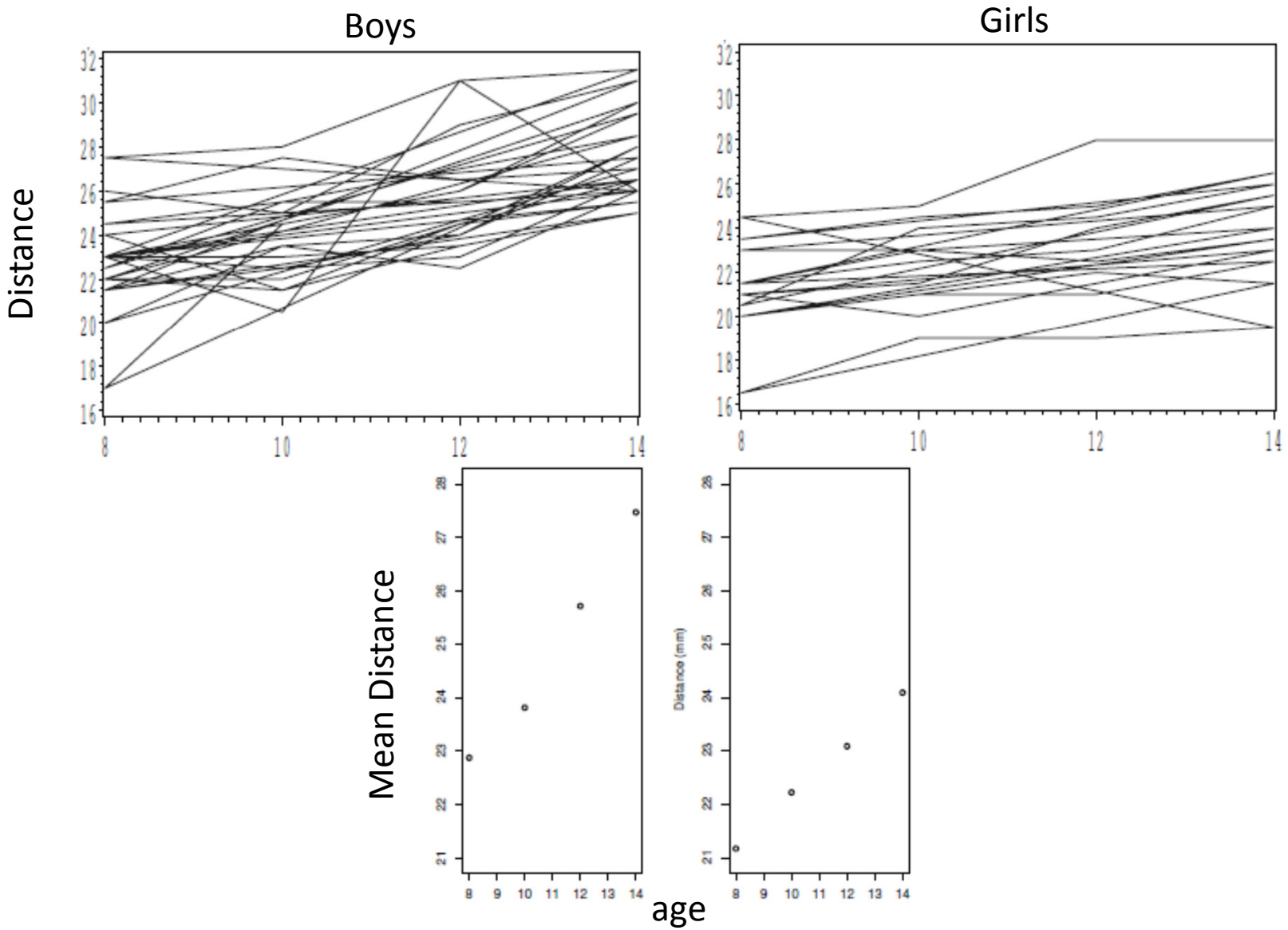


- Measure this distance at ages 8, 10, 12, and 14 in 27 children (16 boys and 11 girls

Questions of Interest

- Does distance change over time?
- What is the pattern of the change?
- Is the pattern of change different for boys and girls? How?

Data



Properties of the Data

- Data are *balanced*: all children have 4 measurements at the same time points (ages)
- Tracking is apparent: children who start small tend to stay small
- Most individual patterns may roughly be described using a straight line
- Average distance also follows an approximate straight line pattern
- Data are “ideal” in the sense that measurements are equally spaced and at a common set of times for all subjects, and no data are missing

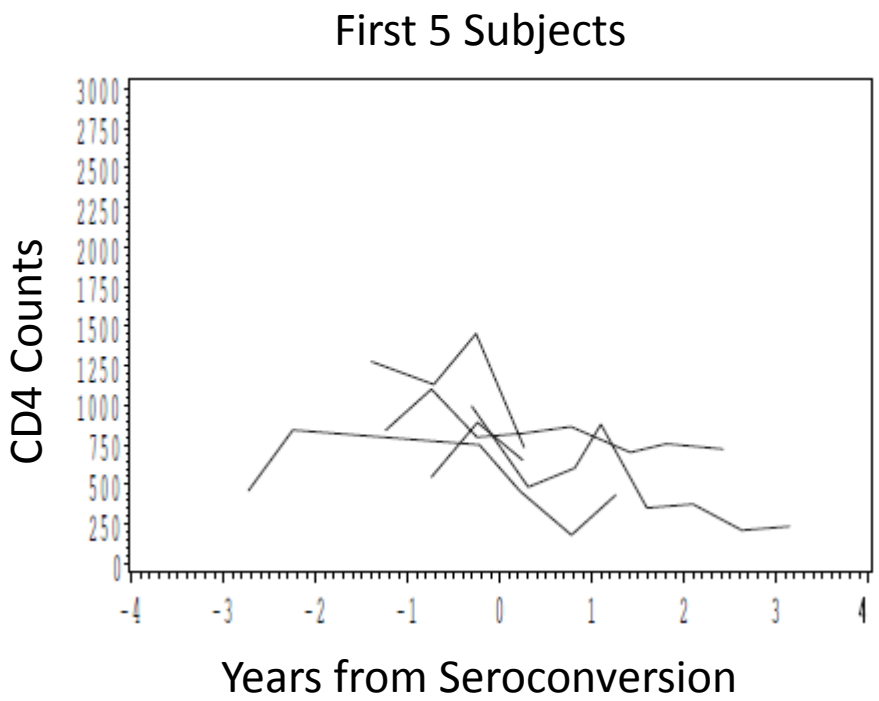
Example III: HIV Study

- HIV attacks CD4+ cells, which regulate the body's immunoresponse to infectious agents. (Uninfected subjects have around 1,100 cells/ml blood.)
- Investigators measured CD4+ counts over time for 369 infected men in the Multicenter AIDS Cohort Study.

Questions of Interest

- What is the typical time course of CD4+ cell depletion?
- What factors, if any, predict CD4+ cell count changes?
- Is there heterogeneity across men in the rate of progression (measured by decline in CD4+ cell counts)?

Data


First 5 Subjects
CD4 Counts
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
-4
-3
-2
-1
0
1
2
3
4
Years from Seroconversion

Properties of the Data

- Data are *unbalanced*: subjects have different numbers of observations taken at different times
- Levels lower after seroconversion in general?
- Much within-subject variability

Benefits of a Longitudinal Study

- A longitudinal study is a cohort study (follow-up study), in which repeated measurements are taken over time for each individual.
- A longitudinal study allows us to study the change of the outcome variable over time
- A longitudinal study is more powerful to detect an association of interest compared to a cross-sectional study

  -- sample size

  -- each subject serves as his/her own control
- A longitudinal study allows us to estimate the between-subject variation and the within-subject variation.

Challenges in Analyzing Longitudinal Data

- In classical linear/logistic regression, the key assumption is the independence between observations
- In a longitudinal study, the observations over time from the same subject are likely to be correlated
- Thus, a valid statistical inference must take the within-subject variation into account
- In a longitudinal study, we usually still assume subjects are independent

If we ignore the correlation…

- Inference may be incorrect (variance estimates too small or too large).
- Estimates of β are inefficient
- Correlation itself may be of interest (e.g. does a treatment work equally well for all patients, or is there heterogeneity in the response to treatment?)

Sources of Correlation: Empirical Observation

- The correlations are positive.
- The correlations often decrease with increasing time separation.
- The correlations between repeated measures rarely ever approaches zero, even in cases where they are taken many years apart.
- The correlation between a pair of repeated measures taken very closely together in time rarely approaches one.

Sources of Correlation

- **Between-Individual Heterogeneity**: reflects natural variation in individuals' propensity to respond. The individuals can vary in average response as well as the response trajectory. The response propensity can be attributed to subject-level characteristics, be it demographic, environmental, or treatment-induced. The between-subject variability in response trajectory essentially categorizes individuals in different classes such as "high" or "low". The level of response for an individual is sustained across the multiple measurements obtained on the individual which in turn induces positive correlation among the repeated measures within individual.

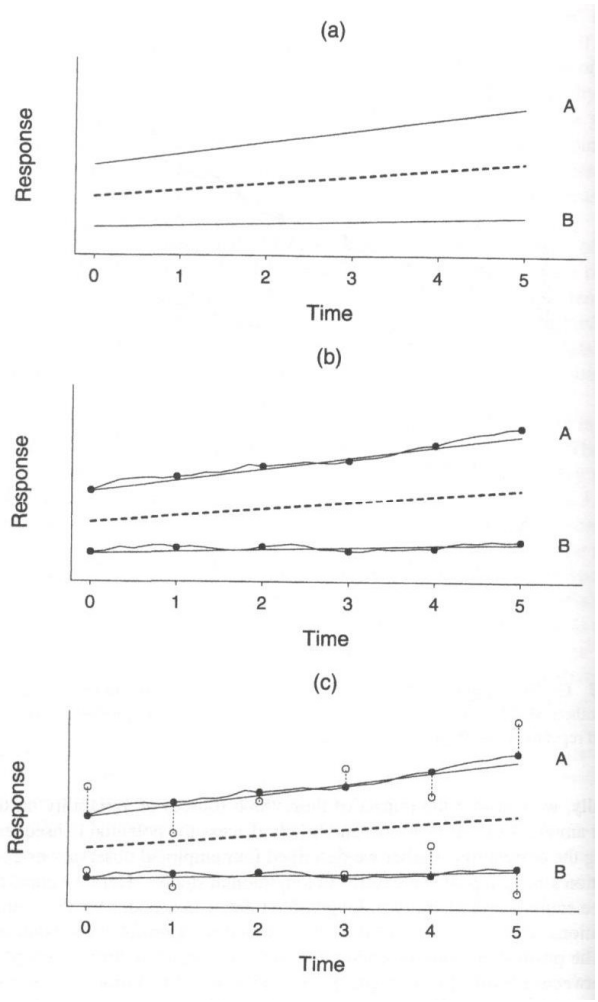
Sources of Correlation (Contd.)

- **Within-Individual Biological Variation**: reflects the biological variation across the repeated measurements within an individual. The variability could be natural variation such as diurnal cyclic patterns of variation or could be induced by external factors. Represents random deviations from an individual's true underlying response trajectory, that are likely to be more similar when observed at short intervals of time. The within subject variation introduces serial correlation which diminishes as the separation between repeat observation points grows.

Sources of Correlation (Contd.)

- **Measurement Error**: reflects the imprecision in the process of measurement. This variability is estimable only when replicate measurements on the same unit is available. In general, the effect of measurement error is to attenuate the correlation of repeated measures.

Sources of Correlation



Sources of Correlation: Empirical Observation

- The correlations are positive: **Between-Individual Heterogeneity + Within-Individual Biological Variation** in response**.**
- The correlations often decrease with increasing time separation: **Within-Individual Biological Variation** in response **+ Between-Individual Heterogeneity** in trajectory.
- The correlations between repeated measures rarely ever approaches zero, even in cases where they are taken many years apart: **Between-Individual Heterogeneity** in the propensity to respond.
- The correlation between a pair of repeated measures taken very closely together in time rarely approaches one: **Measurement Error.**

Acknowledgement

- Slides are based on lecture notes from Amy Herring.