Biostatistics Short Course
Introduction to Longitudinal Studies

Zhangsheng Yu

Division of Biostatistics
Department of Medicine
Indiana University School of Medicine

Outline

Course objectives

- Familiarize with basics of longitudinal studies.
- Know the differences between longitudinal studies and cross-sectional studies.
- Learn how to perform simple analysis.
- Learn the basics of random effects models for longitudinal data.

What are longitudinal studies?

Measures collected repeatedly on the same individuals over time.

- Examples
  - heights or weights in growth studies;
  - numbers of tumors in cancer studies during the follow up period;
  - depression scores in a mental health treatment study during the follow up period.
- In epidemiology, *cohort* study is a longitudinal study . A *cohort* is a set of individuals sharing a common characteristic (e.g. same baseline age) or experience in a particular time period.

Why longitudinal studies?

- In longitudinal data analysis:
  - Interest is on the behavior of a response variable over time.
  - Require special statistical methods to address intra-individual correlation and inter-individual variability.
- They can tease out changes over time within individuals (age effects) from differences among groups of people bonded by time or common life experience (cohort effects).
- They will increase the ”sample size".

Cross-sectional studies

- They aim to describe the relationship between diseases and potential risk factors in a specified population at a particular time.
- They must be done on representative samples of the population for valid generalization.

Longitudinal studies vs. cross-sectional studies

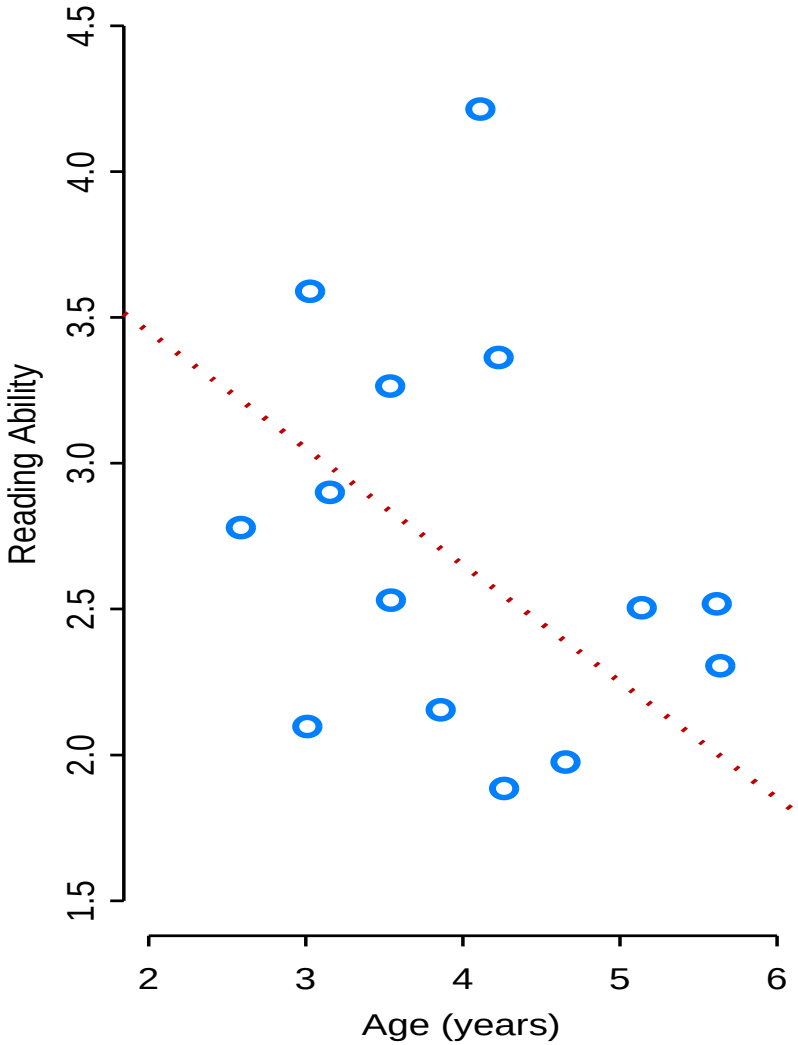
Longitudinal

- Advantage
  - Age vs. cohort effects
  - Efficiency improvement
  - Time to event analysis
- Disadvantage
  - Expensive
  - Analysis more complicated
  - Missing values

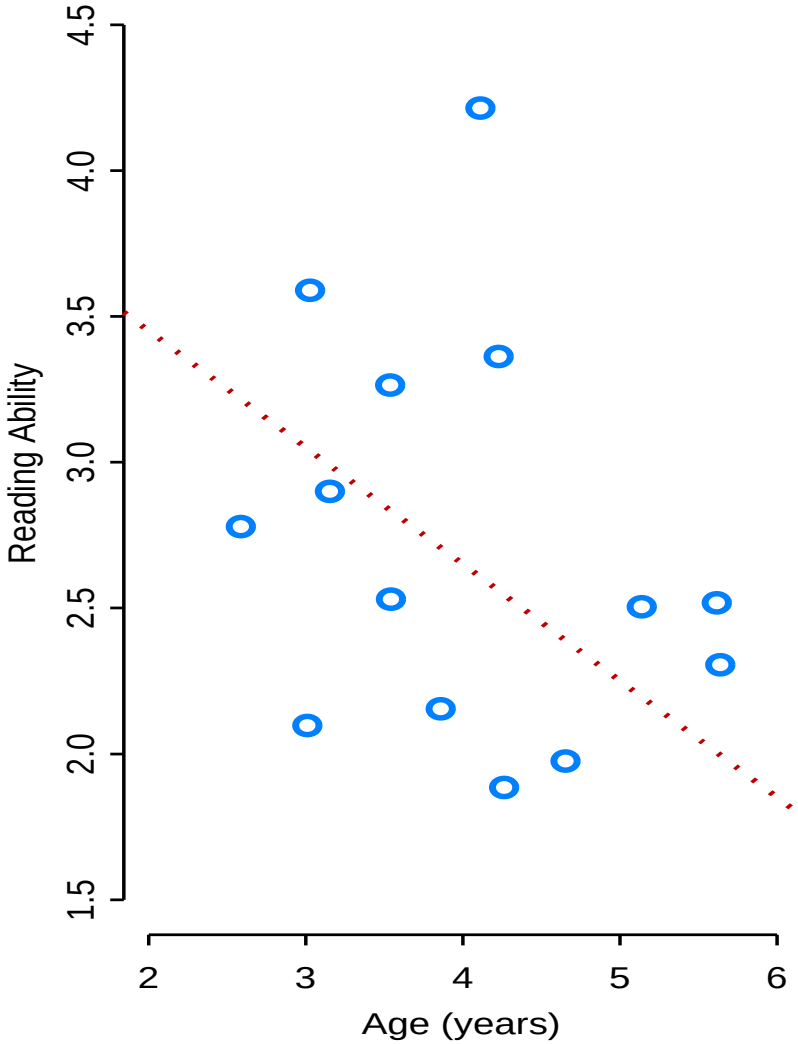
Cross-sectional

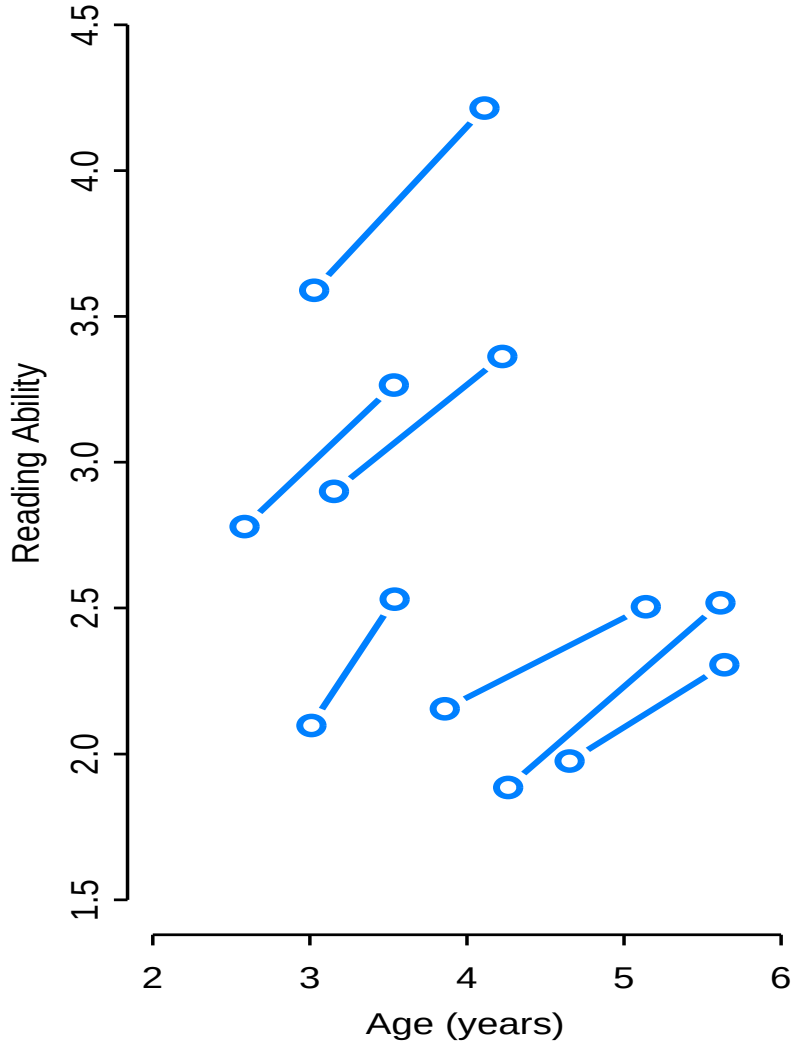
- Advantage
  - Less expensive
  - Simple data structure
  - Many analysis tools
- Disadvantage
  - Unable to detect age effects
  - more difficult to generalize

Reading ability example: cohort vs. age effects



Reading ability example: cohort vs. age effects





Cohort vs. age effects

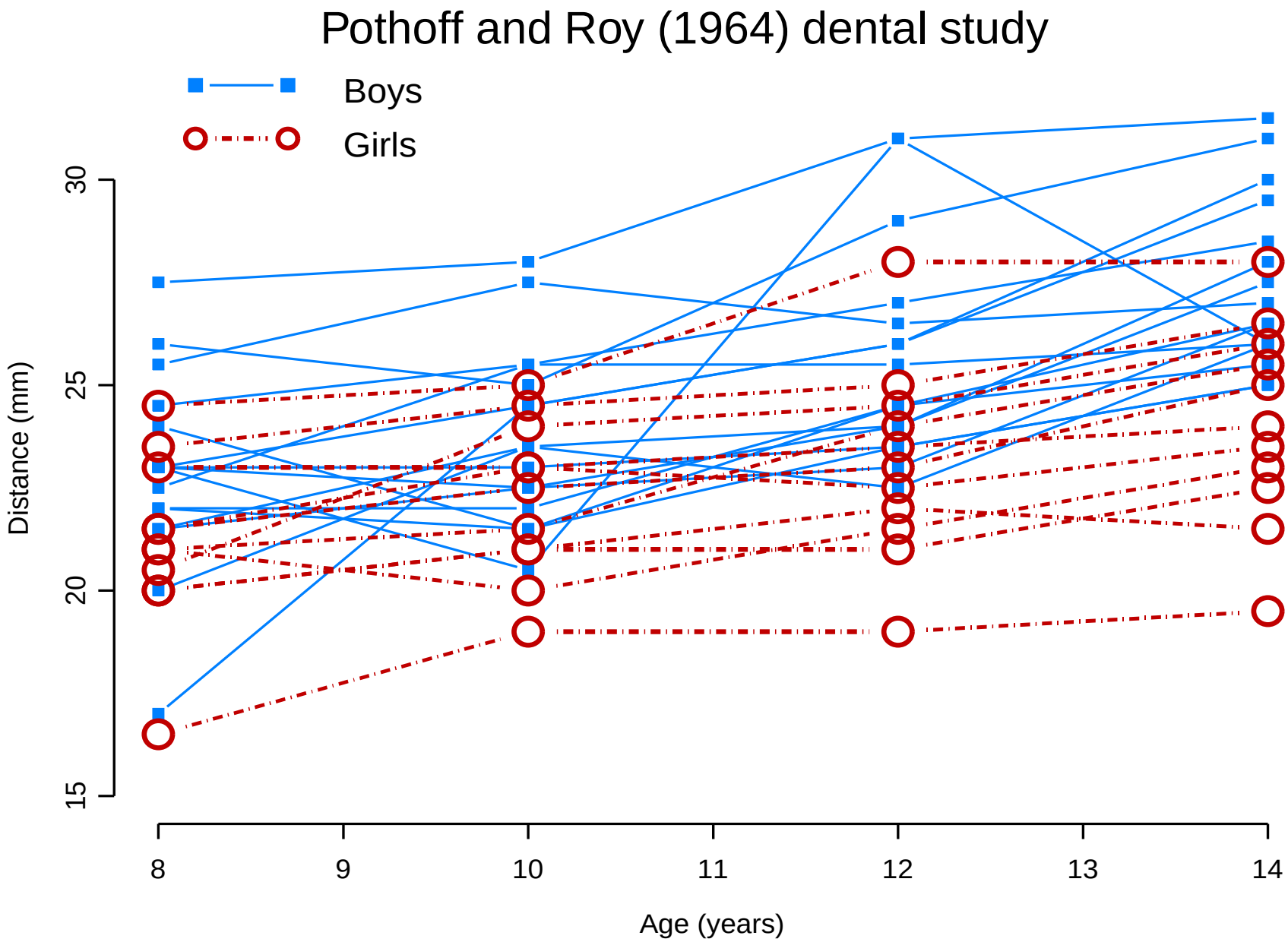
- Age effects: changes over time within subject.
- Cohort effects: differences between subjects at baseline.
- Cross-sectional data can NOT separate age from cohort effects.

Pothoff and Roy dental study (1964)

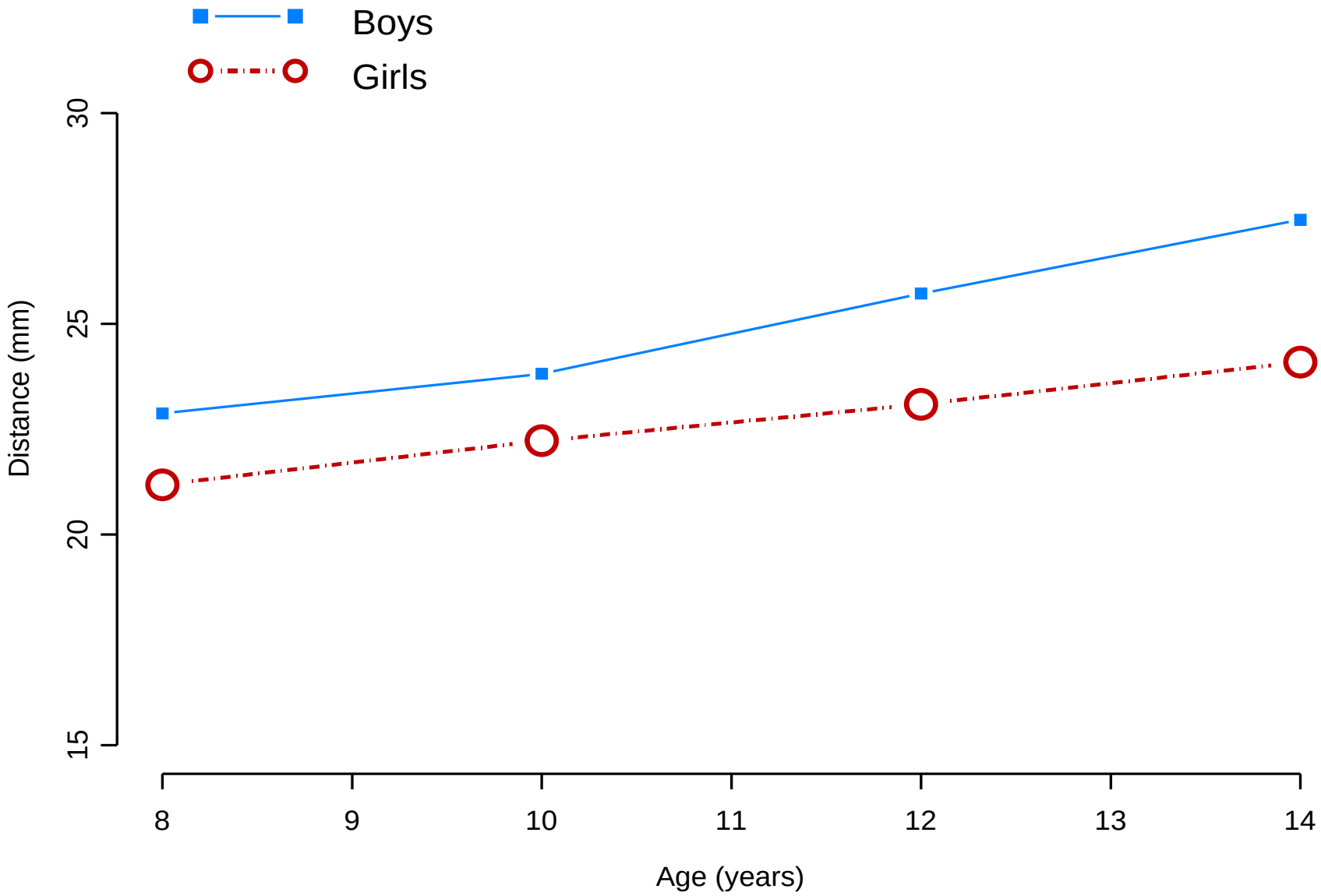
- 27 children, 16 boys, 11 girls;
- On each child, distance (mm) from the center of the pituitary to the pteryomaxillary fissure measured on each child at ages 8, 10, 12, and 14 years of age;
- A continuous measure of growth.

Questions of interest:

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


Pothoff and Roy (1964) dental study
Boys
Girls
Distance (mm)
30
25
20
15
8
9
10
11
12
13
14
Age (years)

Sample average dental distance (Pothoff and Roy, 1964)



Some observations

- All children have all 4 measurements at the same time points (balanced).
- Children who start high or low tend to stay high or low.
- The individual pattern for most children follows a rough straight line increase (with some jitter).
- Average distance (across boys and across girls) follows an approximate straight line pattern.

Statistical issues

- Scientific objectives can be formulated as regression problems: Dependence of a response variable on explanatory variables.
- Repeated observations on each experimental unit: observations from one unit to the next are independent; multiple observations on the same unit are dependent.
- Multiple observations from each subject makes longitudinal data powerful. It also makes it a challenge to analyze.

Analysis of longitudinal data

- Exploratory analysis comprises techniques to visualize patterns of data.
- Confirmatory analysis is judicial work, weighing evidence in data for or against hypotheses.

Exploratory analysis to longitudinal data

- Highlight aggregate patterns of potential scientific interest;
- Identify both cross-sectional and longitudinal patterns;
- Make easy the identification of unusual individuals or unusual observations.

Confirmatory analysis to longitudinal data

- ad hoc analysis:
  - Reduce the repeated values into one or two summary variables;
  - Treat repeatedly observed outcomes from the same subjects as if they are independent and perform regular regression.
- formal statistical modeling:
  - Random effects models;
  - Marginal models;
  - Markov transition models.

Pothoff and Roy dental study (1964)

GROUP=Boys

Variable	N	Mean	Std Dev	minimum	Maximum
AGE8	16	22.875	2.452	17.0	27.5
AGE10	16	23.812	2.136	20.5	28.0
AGE12	16	25.718	2.651	22.5	31.0
AGE14	16	27.468	2.085	25.0	31.5

GROUP=Girls

Variable	N	Mean	Std Dev	minimum	Maximum
AGE8	11	21.182	2.124	16.5	24.5
AGE10	11	22.227	1.902	19.0	25.0
AGE12	11	23.091	2.364	19.0	28.0
AGE14	11	24.091	2.437	19.5	28.0

ad hoc analysis of the dental study

Do things change over time?

- For each gender, compare the mean at age 8 to the means at subsequent ages (i.e. age 10, 12, and 14) using paired t-tests;
- P-values for boys: 0.15, 0.0003, 0.0001;
- Conclusion? Multiple comparisons?
- How to characterize “change"?

ad hoc approach

Are the patterns different for boys and girls?

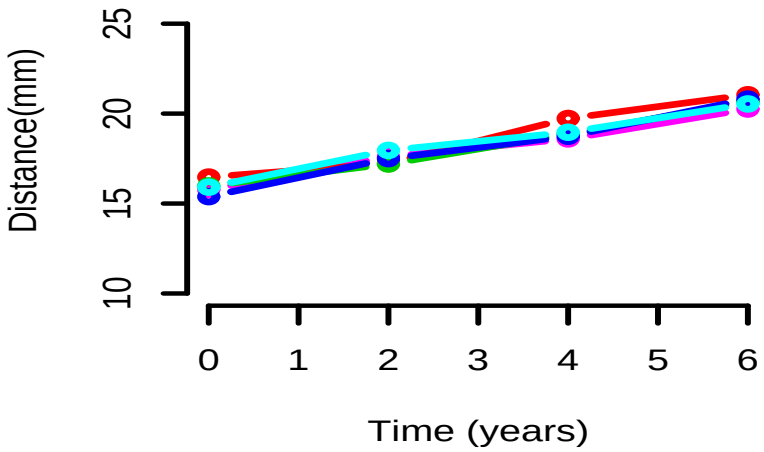
- Compare mean distances between boys and girls at each ages 8, 10, 12, 14 using two-sample t-tests;
- P-values: 0.08, 0.06, 0.01, 0.001;
- Conclusion? Multiple comparisons?
- How to “put this together " to say something about the differences in patterns and how they differ? What are the patterns, anyway?

Random effects model: a formal statistical approach

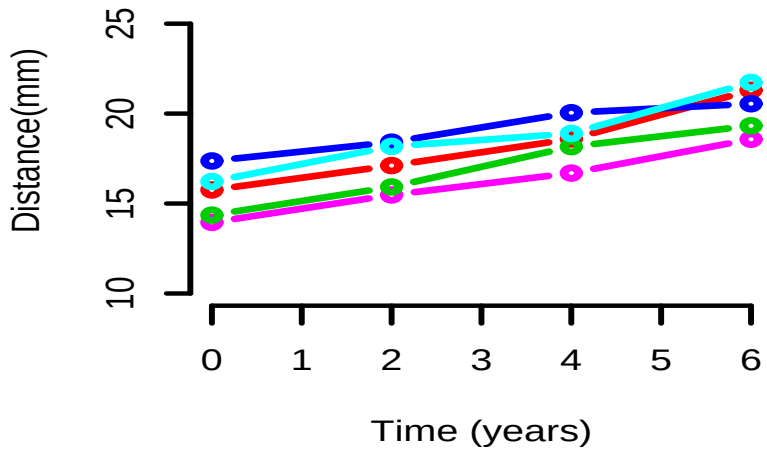
- The model describes patterns of change as trends; e.g., a straight line.
- The model allows starting value (intercept) and trend vary among subject.
- The model acknowledges associations among observations on the same subject by formally modeling the correlation.

Graphical explication of the random effects model

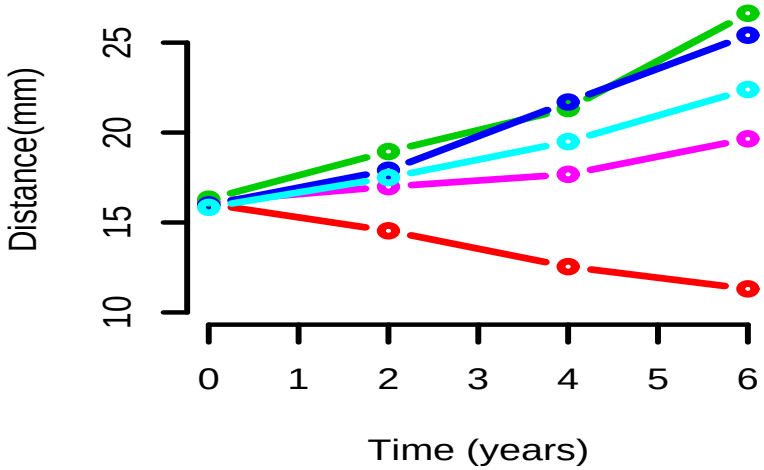
Y(t) = 16 + 0.8t + e(t)



Y(t) = (16+noise) + 0.8t + e(t)



Y(t) = 16 + (0.8+noise)t + e(t)



Y(t) = (16+nois) + (0.8+nois)t + e(t)

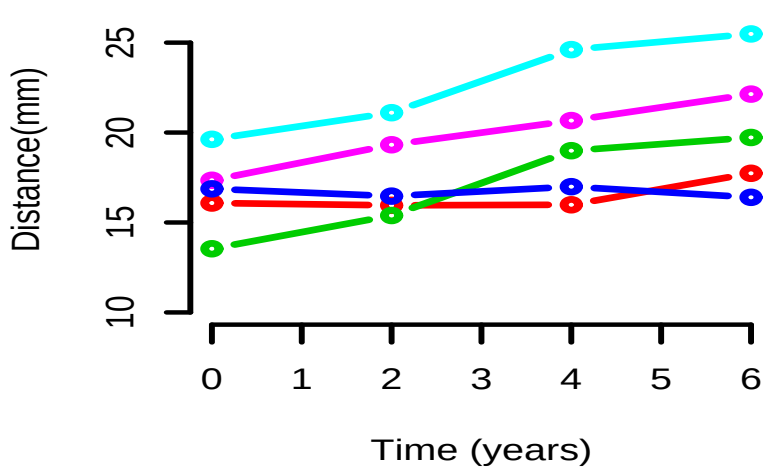


Illustration from the dental study

- Distance measurements for subject i at occasion j is denoted y_{ij}. Note that $t_{ij} = 8, 10, 12, 14$.
- Errors in measuring distance are likely committed at each time point. The term e_{ij} represents the “error" in measurement.
- Each child has its own growth trajectory in the form of a straight line with intercept β_{0i} and slope β_{1i}:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}.$$

- In a linear regression model with gender, age, and gender age interaction as covariates:

$$\textit{Boys}: \quad \beta_{0i} = \beta_{0B}, \beta_{1i} = \beta_{1B}$$

$$\textit{Girls}: \quad \beta_{0i} = \beta_{0G}, \beta_{1i} = \beta_{1G}$$

Illustration from the dental study

Random effect model:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}.$$

$$\textit{Boys}: \quad \beta_{0i} = \beta_{0B} + b_{0i}, \beta_{1i} = \beta_{1B} + b_{1i}$$

$$\textit{Girls}: \quad \beta_{0i} = \beta_{0G} + b_{0i}, \beta_{1i} = \beta_{1G} + b_{1i}$$

- There are population average intercept and slope across boys and across girls. Individual intercepts and slopes deviate from these.
- So if β_{0i} is “high" relative to β_{0B}, then i’s trajectory will be “high" relative to other boys.
- β_{1i} may be steeper or shallower than the average increase (β_{1B}).
- y_{ij} for different j will depend β_{0i} and β_{1i}, so they are correlated.

Analysis results of the dental study

Solution for Fixed Effects

Effect	gender	Estimate	S.E.	DF	t Value	Pr > |t|
Intcpt		16.341	1.019	25	16.04	<.0001
gender	F	1.032	1.596	54	0.65	0.5205
gender	M	0	.	.	.	.
age		0.784	0.086	25	9.12	<.0001
age*gender	F	-0.305	0.135	54	-2.26	0.0277
age*gender	M	0	.	.	.	.

Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
gender	1	54	0.42	0.5205
age	1	25	88.00	<.0001
age*gender	1	54	5.12	0.0277

Analysis results of the dental study (continued)

- Is there a change in distance over time for boys and girls? $\beta_{1B} = 0$? $\beta_{1G} = 0$?
  - estimates: $\hat{\beta}_{1B} = 0.784, \hat{\beta}_{1G} = 0.784 - 0.305 = 0.479$;
  - p-values: < 0.0001.
- Is the pattern of change the same? $\beta_{1B} = \beta_{1G}$? (p-value = 0.027).

Summary

- Information in longitudinal data is often not fully exploited because ad hoc methods are used.
- By conceptualizing longitudinal data, we gain a framework to “make the most".
- Focused here on continuous measurements, but similar methods exist for binary (yes/no) and categorical outcomes.

Acknowledgement

Slides courtesy of Dr. Menggang Yu.