Part II: Longitudinal Data Analysis

Outline

- Repeated measurements
- General linear regression
- Linear mixed-effects model basics
- Generalized estimating equation
- Generalized linear mixed-effects model

Data Types

- **Cross-sectional data**
  Outcome variable(s) and covariates that are measured at a single time point
- **Longitudinal data**
  Each subject gives rise to a vector of measurements, but these represent the same response measured at a sequence of observation times

Characteristics of Longitudinal Data

- Individuals are measured repeatedly over time
- The time when the measurements are taken is not of primary interest and is often considered fixed by design.
- Small number of observations per subject but relatively large number of subjects.
- The variability of observed data can be divided into three components:
  1. Heterogeneity between individuals.
  2. Serial correlation, measurements closely spaced are more similar.
  3. Measurement error.

Longitudinal Data Analysis

Longitudinal data analysis (LDA) focuses on

- changes over time within individuals
- differences among people in their baseline levels

Types of LDA

- Time series studies
- Panel studies (sociology and economics)
- Prospective studies (clinical trials)

Longitudinal Study vs Cross-Sectional Study

Example: A cross-sectional study found that older people smoke more. Possible explanations:

- People tend to smoke more when they get older.
- Older people grew up in an environment where the harm of smoking was less widely accepted.

LDA can distinguish the effect due to aging (*i.e.*, changes over time within subject) from cohort effects (*i.e.*, difference between subjects at baseline). Cross-sectional study cannot.

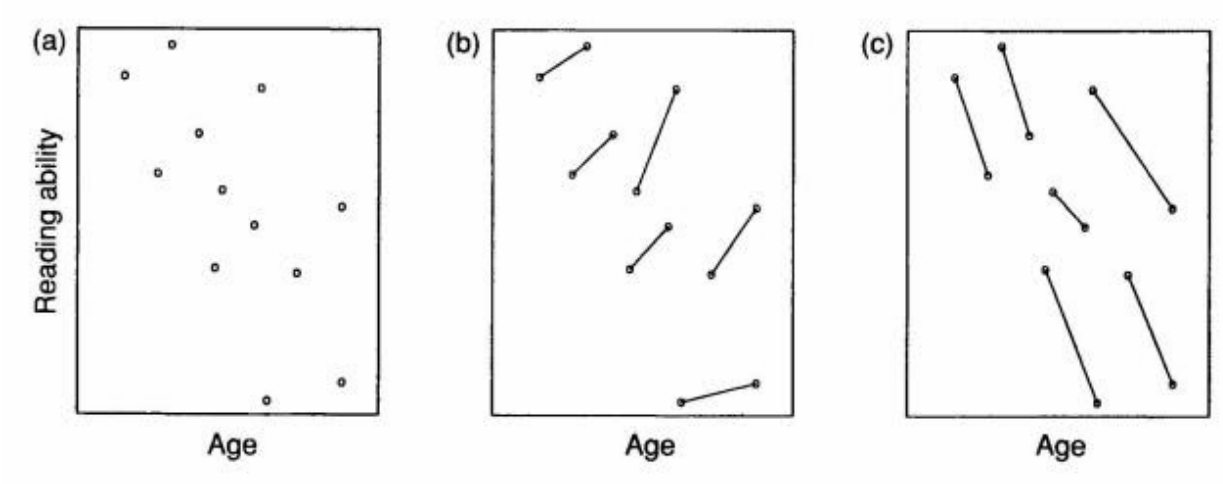
Advantages of Longitudinal Study

- Each subject can serve as his/her own control. Influence of genetic make-up, environmental exposures, and maybe unmeasured characteristics tend to persist over time.
- Distinguish the degree of variation in Y across time within a subject from the variation in Y between subjects. With repeated values, one can borrow strength across time for the person of interest as well as across people.
- Increased power, by repeated measurements. The repeated measurements from the same subject are rarely perfectly correlated. Hence, longitudinal studies are more powerful than cross-sectional studies.

Why Special Methods?

LDA requires special statistical methods because the set of observations on one subject tends to be inter-correlated.

Example: Reading Ability (hypothetical data)



- Assume this is a longitudinal study with two measurements per child.
- The two measurements per subject may be highly correlated.
- If we use cross-sectional methods to analyze the data, we may not be able to distinguish changes over time within individual and difference among people in their baseline levels.

In general, repeated observations $y_{i1}, \cdots, y_{in_i}$ for subject i are likely to be correlated, so the independence assumption is violated.
The standard regression methods (ignoring correlation) may lead to

- Incorrect inference
- Inefficient estimates of β
- Oversight of important correlation structure

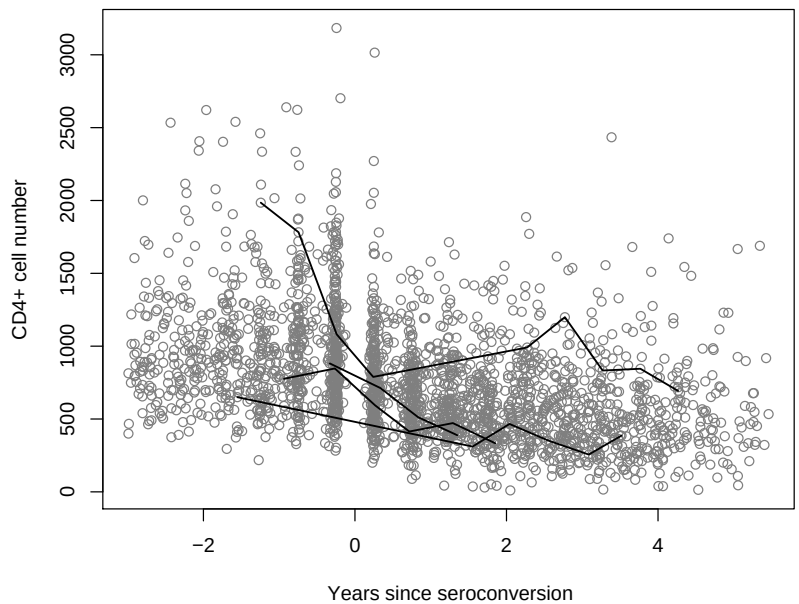
Example

CD4+ Cell Numbers:

- HIV attacks CD4+ cells which regulate the body's immune response to infectious agents. The number of CD4+ cells predicts AIDS-related events. An uninfected individual has around 1100 cells per milliliter of blood.
- 2376 values of CD4+ cell counts plotted against time since seroconversion (detectable HIV antibodies) for 369 infected men enrolled in the MACS.
- Question: What is the impact of HIV infection on CD4+ counts over time?

Goals:

- Characterize the time course of CD4+ cell depletion.
- Identify factors which predict CD4+ cell changes.
- Characterize the degree of heterogeneity across men in the rate of progression.



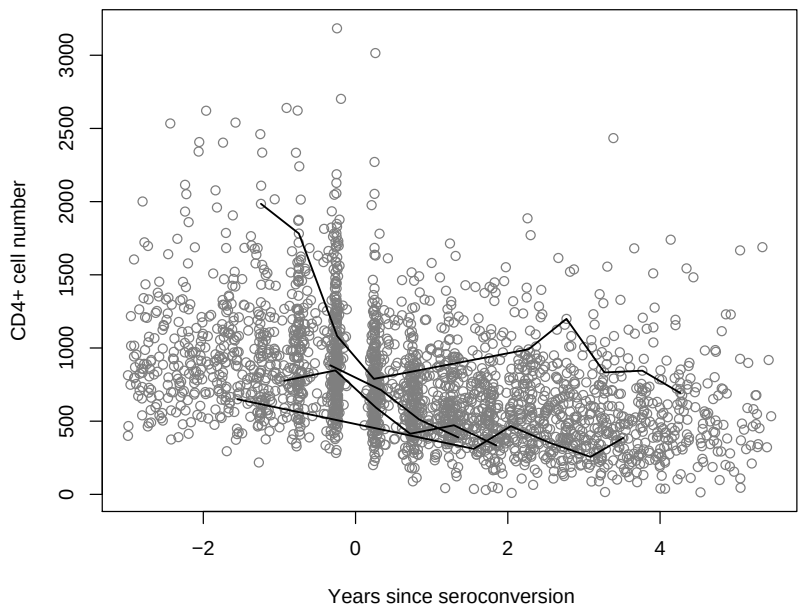
Data are highly unbalanced with irregular observation times and numbers.

- In a usual regression analysis,

$$Y_{ij} = x_{ij}\beta + \epsilon_{ij}$$

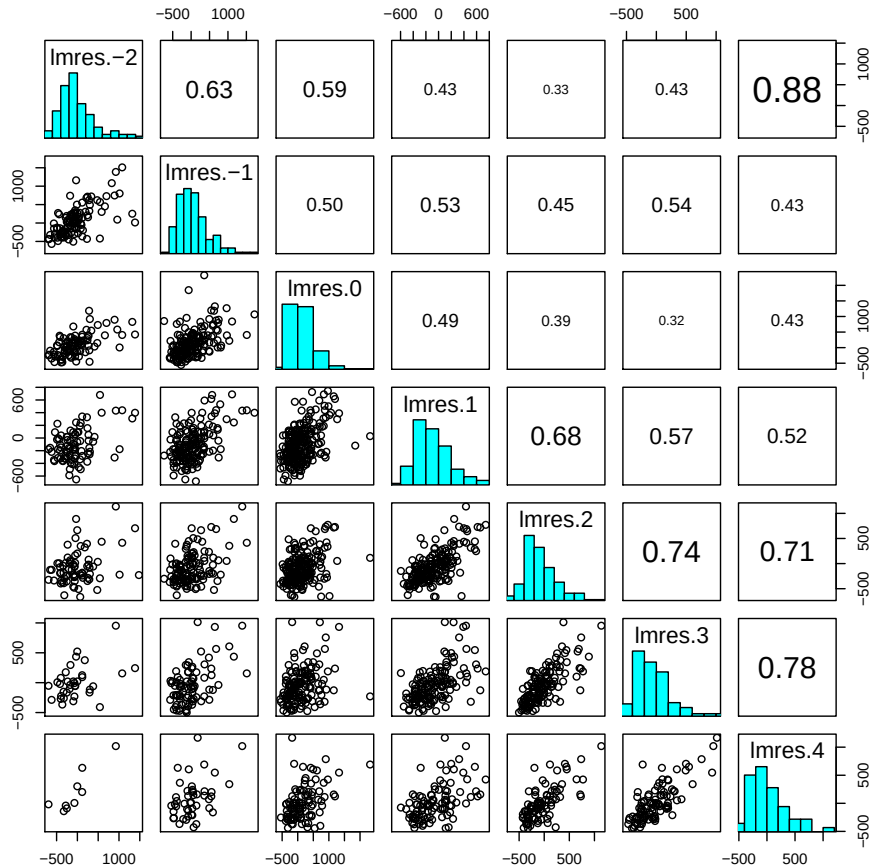
  i=person, j= observation; ϵ_{ij}'s are independent

- Repeated measurements on a subject are likely correlated; assumption of independence is violated.



To see the correlation structure within each subject

- First regress y_{ij} onto x_{ij} using the ordinary least-squares (OLS) and obtain the residuals $r_{ij} = y_{ij} - x_{ij}^T \hat{\beta}$.
- Then create the scatter plots of residuals at different time points (or time intervals if the time points are not regular).



Example

Respiratory Infection:

- Determine effects of vitamin A deficiency in pre-school children
- Over 3000 children were examined for up to 6 visits to assess whether they suffered from respiratory infection.
- Weight and height are also measured.
- Question: What are the predictors of infection?

- Estimate the increase in risk of respiratory infection for children who are VA deficient, while controlling for other demographic factors
- Estimate the degree of heterogeneity in the risk of disease among children
- Responses are binary (i.e., $i \rightarrow (0, 1, 1, 0, 0)$)
- Data are irregularly measured

Example

Epileptic Seizures:

- Clinical trial of 59 epileptics
- For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks
- Patients were randomized to treatment with the antiepileptic drug progabide or placebo
- Number of seizures was then recorded in 4 consecutive 2-week intervals
- Question: Does progabide reduce the epileptic seizure rate?

- Identify whether treatment is related to the change of seizure rate.
- Responses are counts.
- Correlations within a subject are high.
- Data are regularly measured.

Table 1.5. Four successive two-week seizure counts for each of 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = progabide), eight-week baseline seizure counts, and age (in years).

Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age	Y_1	Y_2	Y_3	Y_4	Trt.	Base	Age
5	3	3	3	0	11	31	0	4	3	0	1	19	20
3	5	3	3	0	11	30	3	6	1	3	1	10	20
2	4	0	5	0	6	25	2	6	7	4	1	19	18
4	4	1	4	0	8	36	4	3	1	3	1	24	24
7	18	9	21	0	66	22	22	17	19	16	1	31	30
5	2	8	7	0	27	29	5	4	7	4	1	14	35
6	4	0	2	0	12	31	2	4	0	4	1	11	57
40	20	23	12	0	52	42	3	7	7	7	1	67	20
5	6	6	5	0	23	37	4	18	2	5	1	41	22
14	13	6	0	0	10	28	2	1	1	0	1	7	28
26	12	6	22	0	52	36	0	2	4	0	1	22	23
12	6	8	5	0	33	24	5	4	0	3	1	13	40
4	4	6	2	0	18	23	11	14	25	15	1	46	43
7	9	12	14	0	42	36	10	5	3	8	1	36	21
16	24	10	9	0	87	26	19	7	6	7	1	38	35
11	0	0	5	0	50	26	1	1	2	4	1	7	25
0	0	3	3	0	18	28	6	10	8	8	1	36	26
37	29	28	29	0	111	31	2	1	0	0	1	11	25
3	5	2	5	0	18	32	102	65	72	63	1	151	22
3	0	6	7	0	20	21	4	3	2	4	1	22	32
3	4	3	4	0	12	29	8	6	5	7	1	42	25
3	4	3	4	0	9	21	1	3	1	5	1	32	35
2	3	3	5	0	17	32	18	11	28	13	1	56	21
8	12	2	8	0	28	25	6	3	4	0	1	24	41
18	24	76	25	0	55	30	3	5	4	3	1	16	32
2	1	2	1	0	9	40	1	23	19	8	1	22	26
3	1	4	2	0	10	19	2	3	0	1	1	25	21
13	15	13	12	0	47	22	0	0	0	0	1	13	36
11	14	9	8	1	76	18	1	4	3	2	1	12	37
8	7	9	4	1	38	32							

Recap of LDA

- Repeated observations on each experimental unit
- Units can be assumed independent of one another
- Multiple responses within each unit are likely to be correlated
- The objectives can be formulated as regression problems whose purpose is to describe the dependence of the response on explanatory variables
- The choice of the statistical model must depend on the type of the outcome variable