Introduction

This course describes statistical methods for the analysis of longitudinal data, with a strong emphasis on applications in the biological and health sciences

- **Univariate statistics:** each subject gives rise to a single measurement, termed *response*.
- **Multivariate statistics:** each subject gives rise to to a vector of measurements, or *different responses*.
- **Longitudinal data:** each subject gives rise to a vector of measurements, but these represent the same response measured at a sequence of observation times.
- Repeated responses over time on independent units (persons)

Topics

- **Basic issues and exploratory analyses**
  Definition and examples of LDA
  Approaches to LDA
  Exploring correlation
- **Statistical methods for continuous measurements**
- General Linear model with correlated data
  Weighted Least Squares estimation
  Maximum Likelihood estimation
  Parametric models for covariance structure
- **Generalized linear models for continuous/discrete responses**
- Marginal Models
- Random Effects Models
- Transition Models
  Log Linear Model and Poisson Model for count responses
  Logistic model for binary responses
  GEE estimation methods
  Estimation techniques

Introduction

- **Longitudinal study**: people are measured repeatedly over time;
- **Cross-sectional study**: a single outcome is measured for each individual
- In a LDA we can investigate:
  - changes over time within individuals (*age effects*)
  - differences among people in their baseline levels (*cohort effects*)
- LDA requires special statistical methods because the set of observations on one subject tends to be intercorrelated.

Characteristics

- Repeated observations on individuals
- Scientific questions $\rightarrow$ regression methods
- response $= f$(predictors)
- discrete/continuous responses and predictors

Why special methods?

Repeated observations $y_{i1}, y_{i2}, \ldots, y_{in_i}$ are likely correlated, assumption of independence is violated

What if we use standard regression methods anyway (ignore correlation)?

- Correlation may be of scientific focus
- Incorrect Inference
- Inefficient estimates of β

What is special about longitudinal data?

- Opportunities
- Distinguish "longitudinal" from "cross sectional" effects
- Choose several targets of estimation

Challenges

- Repeated observations tend to be autocorrelated (Y_{ij} more like $Y_{i'j}$ than like $Y_{i'j'}$)
- Correlation must be modeled

Examples

1. CD4 + cell numbers (continuous)
2. Growth of Sitka spruce - tree size (continuous)
3. Protein contents of milk (continuous)
4. Indonesian Children's health study (binary)
5. Analgesic Crossover trial (binary)
6. Epileptic seizures (count)
7. Health Effects of air Pollution

CD4+ cell numbers

- HIV attack CD4+ cell which regulates the body's immunosreponse to infectious agent
- 2376 values of CD4+ cell number plotted against time since sieroconversion for 369 infected men enrolled in the MACS
- **Q:** What is the impact of HIV infection on CD4 counts over time?
- **Goals:**

1. characterize the typical time course of CD4+ cell depletion
2. identify factors which predict CD4+ cell changes
3. estimate the average time course of CD4+ cell depletion
4. characterize the degree of heterogeneity across men in the rate of progression

Growth of tree

- Data for 79 trees over two growing seasons
- 54 trees were grown with ozone exposure at 70 ppb
- 25 trees were grown under control conditions
- **Goal:** compare the growth patterns of trees under the two conditions

Protein content of milk

- Milk was collected weekly for 79 Australian cows and analyzed for its protein content
- Cows were maintained on one of three diets
- **Goal:** to determine how diet affects the protein milk
- **Problem**: about half of the 79 sequences are incomplete – missing data

Indonesian Children's Health Study

- Dr. Sommer conducted a study to determine effects of vitamin A deficiency in pre-school children
- Over 3000 children examined for up to six visits to assess whether they suffered from respiratory infection, an ocular manifestation of vitamin A deficiency. Weight and height are also measured.
- **Q:** predictors of infection?
- **Goals:**

1. Estimate the increase in risk of respiratory infection for children who are vitamin A deficient while controlling for other demographic factors
2. Estimate the degree of heterogeneity in the risk of disease among children

Analgesic Cross over trial

- 3 period crossover trial of an analgesic drug for relieving pain for primary dysmenorrhea
- 3 levels of analgesic (control, low, and high) were given to each of the 86 women
- Women were randomized to one of the six possible orders for administering the three treatment levels
- Pain was relieved for 26% with placebo, 71% with low dose, and 80% with high dose
- **Q:** treatment effect?

Epileptic seizures

- Clinical Trail of 59 epileptics
- For each patient, the number of epileptic seizures was recorded during a baseline period of eight weeks
- patient were randomized to treatment with the anti-epileptic drug progabide or placebo
- Number of seizures was then recorded in four consecutive two weeks intervals
- **Question:** is progabide reduces the rate of epileptic seizures?

Health Effects of Air Pollution

- daily time series data for Baltimore
- primary outcome: daily mortality
- covariates: time, season, PM_{10}
- **Q:** association between mortality and air pollution?

What these examples have in common?

- there are repeated observations on each experimental unit;
- units can be assumed independent of one other;
- multiple responses within each unit are likely to be correlated;
- the objectives can be formulate 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.

Course Overview

- Scientific objectives include

  Characterize change

  Component of variation

  Hypothesis testing
- We will focus on regression methods
- We will consider (up to 6) cases-studies in detail.
- Computing using **Stata** will be introduced

Notation

- Y_{ij} = response variable
- x_{ij} = explanatory variable observed at time t_{ij}
- $j = 1, \ldots, n_i$ observations
- $i = 1, \ldots, m$ subject
- $E(Y_{ij}) = \mu_{ij}, \ V(Y_{ij}) = v_{ij}$
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$
- $E(\boldsymbol{Y}_i) = \mu_i, \ V(\boldsymbol{Y}_i) = V_i, \ [V_i]_{jk} = cov(Y_{ij}, Y_{ik})$
- **Regression Model**
- $\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$

Cross sectional versus longitudinal study

- Cross-sectional study ($n_i = 1$)

1. $Y_{i1} = \beta_c x_{i1} + \epsilon_{i1},\ i = 1, \ldots, m$

- β_c represents the difference in average Y across two sub-populations which differ by one unit in x.
- With repeated observations we can extend the model

2. $Y_{ij} = \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij}$
   $j = 1, \ldots, n_i,\ i = 1, \ldots, m$
3. $(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij} - \epsilon_{i1},\ (2. - 1.)$

- β_L represents the expected change in Y over time per one unit change in x for a given subject.
- if $n = 1 \rightarrow$ model 1. = model 2.

Cross sectional versus longitudinal study

- In CS the basis is a comparison of individuals with a particular value of x to others with a different value
- in LDA each person is his or her control. β_L is estimated by comparing a person's response at two times assuming that x changes over time
- in LS we can distinguish the degree of variation of Y *across time for one individual* from the variation of Y *across people.*

Example

- Suppose that we want to estimate a man's immune status as reflected in his $CD4+$ level
- in CS, one man's estimate must draw upon the data from others to overcome measurement error. But averaging across people ignores the natural differences in $CD4+$ among persons
- in LS we can borrow strength across time for the persons of interest as well as across people
- little variability among people, then one man's estimate can rely on data for others as in the CS case
- large variability among people, we might prefer use only the data for the individuals

Approaches to LDA

- If we have one observation on each experimental unit, we are confined to modeling the population average of Y called *marginal mean response*
- If we have repeated measurements, there are several approaches that can be adopted

1. reduce the repeated values into one or two summary variables
2. analyze each summary variable as a function of covariates $\boldsymbol{x}_i$

Where does correlation come from?

- Past causing present

$$\text{logit} Pr(Y_{ij} = 1 \mid \text{past}) = x_{ij}\beta + \alpha y_{ij-1}$$

- Latent variables

$$\log \frac{Pr(Y_{ij} = 1 \mid U_i)}{Pr(Y_{ij} = 0 \mid U_i)} = \beta_0 + U_i + \beta_1 x_{ij}$$

- U_i are unobserved

Approaches to LDA

- **Marginal Model**

  $E(\boldsymbol{Y}_i) = X_i\boldsymbol{\beta},\ V(\boldsymbol{Y}_i) = V_i(\boldsymbol{\alpha})$

- **Random Effects Model**

$$\begin{aligned} E(Y_{ij} \mid \boldsymbol{\beta}_i) &= \boldsymbol{x}'_{ij}\boldsymbol{\beta}_i \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \boldsymbol{U}_i \end{aligned}$$

- **Transition Models**

  $E(Y_{ij} \mid Y_{ij-1}, \ldots, Y_{i1}, \boldsymbol{x}_{ij})$

Marginal model - ICHS example

- $Y_{ij} = \begin{cases} 1 & \text{Resp infection} \\ 0 & \text{Not} \end{cases}$
- $x_{ij} = \begin{cases} 1 & \text{Vit A deficiency} \\ 0 & \text{Not} \end{cases}$
- Mean model

$$\log \frac{Pr(Y_{ij} = 1)}{Pr(Y_{ij} = 0)} = \beta_0 + \beta_1 x_{ij}$$

- coefficients describe/compare subpopulations
- $\exp(\beta_1)$ = ratio of odds of RI for two vitamin A groups
- with binary responses, models for odds ratios preferred to correlations

Random Effects models

- Idea - correlations among Y_{ij} caused by a latent variable U_i

$$\log \frac{Pr(Y_{ij} = 1 \mid U_i)}{Pr(Y_{ij} = 0 \mid U_i)} = \beta_0 + U_i + \beta_1 x_{ij}$$

- $\beta_0 + U_i$ = child i intercept
- $\beta_1 x_{ij}$ = common vitamin A effect
- β_1 is the log odds of RI for given child when he is vitamin A deficient versus when he is not

Transition models

- Idea - past responses have an effect on current responses

$$\log \frac{Pr(Y_{ij} = 1 \mid \text{past}_i)}{Pr(Y_{ij} = 0 \mid \text{past}_i)} = x_{ij}\beta + \alpha y_{ij-1}$$