1. Introduction

Longitudinal study

In a longitudinal study, each subject is measured multiple times often over a considerable time interval, as opposed to cross-sectional data where a single outcome is measured for each individual.

Examples

1. Orthodontic measurements

2. Multicenter AIDS Cohort Study (MACS)

3. Indonesian Children's Health Study (ICHS)

4. Analgesic crossover trial (Crossover trial)

5. Epileptic seizures

Longitudinal vs. Cross-sectional Studies

- A cross-sectional study found that older people smoke more.
  Possible explanations:
  - People tend to smoke more as they get older.
  - Older people grew up in an environment where the harm of smoking was less widely accepted.

- Longitudinal studies can distinguish between the effect due to 'age' at measurement and birth date (cohort). Together they determine the date of measurement (period).

Advantage of Longitudinal Studies

- Increased power, by repeated measurements, and separating measurement errors and sampling (in time) errors.

- Reducing bias. e.g., length-bias sampling.

- Investigation of individual-level changes.

Correlated Data

In a regression analysis, we model the mean of a response $(Y_1, \cdots, Y_n)$ as a function of covariates $(x_1, \cdots, x_n)$, where the subscripts $1, \cdots, n$ denote study units. We assume that

$$P(Y_1, \cdots, Y_n | x_1, \cdots, x_n, \beta) = P(Y_1 | x_1, \beta) \cdots P(Y_n | x_n, \beta).$$

i.e., the Y's are conditionally independent given covariates x.

However, the Y's are not independent marginally (i.e., $P(Y_2 | Y_1) \neq P(Y_2)$). Longitudinal data is a special case of correlated data where $Y | X, \beta$ are not independently distributed.

Examples of correlated data

- Clustered data: multi-center studies, kids in the same classroom. Subjects in the same cluster are correlated.

- Split-plot design: nested factors.

- Familial data and social networks: complex correlation patterns.

- Time-series data: typically a few subjects with many observations over a long period of time. The emphasis is typically on prediction (i.e., using past time-course pattern to predict the future).

- Spatial data: There is in essence only one subject (the earth). Similar to time-series, only with a higher dimension (2D or 3D) and without the directionality.

- Recurrent-event data: the observation times are random and are typically the variables of interest. For survival data, the event can only happen once.

Characteristic of Longitudinal Data

- Small number of observations per subject but relatively large number of subjects.

- The emphasis is on <u>inference</u> in comparing subjects or subject groups.

- Longitudinal data can also arise from clustered, familial or spatial data.

- Longitudinal data often require (allow) the most elaborate modeling of the correlation structure.

- The variability can be divided into three components:

  1. Heterogeneity between individuals (random effects).
  2. Serial correlation, measurements closely spaced are more similar.
  3. Measurement error.

  By virtual of replication, in subject and in time, it is possible to distinguish between them.

What if we ignore the correlation?

There are at least three consequences:

- Incorrect inferences about regression coefficients β.

- Inefficient estimates of β (i.e., less precise than possible).

- Sub-optimal protection against biases causes by missing data.

Sources of Correlation

Random effects/latent variable

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

where $u_i =$ unobserved.

Serial correlation

Notations We will mostly follow the notation in our textbook.

- Vectors: x, Y, β
  For example,

$$x = (x_1, \cdots, x_n)^T = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

- Matrices (uppercase): X, Σ

- Parameters are represented by Greek letters.

- For scalars and vectors, random variables are uppercased: Y_i, Y.

- For scalars and vectors, observed (non-random) variables are lowercased: x_i, y.

- Let $i = 1, \cdots, m$ index subjects.

- For each subject i, we have n_i observations at time t_{ij}, $j = 1, \cdots, n_i$.

- x_{ij} is a p-vector that are the covariates for observation j of subject i. X_i is a $n_i \times p$ matrix of all the covariates for i and X is all the covariates.

- The outcome for subject i is denoted by the n_i-vector $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ with mean μ_i and $n_i \times n_i$ covariance matrix V_i where $v_{ijk} = cov(Y_{ij}, Y_{ik})$. The correlation matrix is R_i.

- $Y = (Y_1^T, \cdots, Y_m^T)^T$ is an N-vector with $N = \sum_{i=1}^m n_i$.

Review of Linear Model Theory

- A classic linear model can be written as

$$E(Y|X) = \mu = X\beta, \tag{1}$$

$$var(Y|X) = \Sigma = \sigma^2 I, \tag{2}$$

where Y is a m-vector, β is a p-vector (p is the number of regression parameters) and X is a $m \times p$ design matrix, I is the identity matrix.

- The method of least squares aims to minimize the quadratic loss function (sum of squared errors):

$$(Y - X\beta)^T (Y - X\beta).$$

- The OLS (ordinary least squares) solution is

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

- It is also the MLE of β if we assume Y has a multivariate normal distribution

$$Y|X \sim N(X\beta, \sigma^2 I).$$

- It follows then that $\hat{\beta}$ is also multivariate normal with mean β and variance $\sigma^2 (X^T X)^{-1}$.

- An unbiased estimator of σ^2 is

$$\hat{\sigma^2} = \frac{RSS}{m-p} = \frac{1}{m-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta}).$$

Note that it is not the MLE.

Analysis of Longitudinal Data

Several possible approaches to analyze longitudinal data. The main challenge involves how to take into account the correlation structure.

- Summary statistics based on approach: calculate a univariate summary statistic for the multiple measurements which can be used in the second step of the analysis.
  - Simple and especially useful for exploratory analysis.
  - Lost of information, underestimation of uncertainty.
  - Cannot deal with time-dependent covariates.

- Marginal approach: models the marginal mean responses

$$E(Y_i) = X_i\beta, \tag{3}$$

$$var(Y_i) = V_i(\alpha), \tag{4}$$

where both β and α must be estimated, and V_i may also depend on some covariates.

- Conditional approach (random effect model, hierachical model, multi-level model): assumes correlation arises becuase of heterogeneity in subjects. i.e.,

$$Y_i|b_i \sim N(X_i\beta + Z_i b_i, \sigma^2 I), \tag{5}$$

$$b_i \sim N(0, \tau^2 I). \tag{6}$$

- Transition model:

$$E(Y_{ij}|Y_{ij-1}, x_{ij}) = x_{ij}^T\beta + \alpha Y_{ij-1}. \tag{7}$$

Bias of Naive Analysis

For longitudinal data, if we ignore the correlation we can have a model of the form

$$Y_{ij} = \beta_0 + \beta_c x_{ij} + \epsilon_{ij}, \quad j = 1, \cdots, n; \quad i = 1, \cdots, m,$$

where β_c represents the difference in average Y across two subpopulations which differ by one unit in x.

Model (8) can also be written as

$$Y_{ij} = \beta_0 + \beta_c x_{i1} + \beta_c(x_{ij} - x_{i1}) + \epsilon_{ij}. \tag{8}$$

Note that this model assumes that the baseline effect is the same as the longitudinal effect. To relax this assumption, we have different parameters for the two effects

$$Y_{ij} = \beta_0 + \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij}. \tag{9}$$

- When $j = 1$, the two models (8) and (9) are equivalent.

- Using (9) we can estimate β_c from the longitudinal data.

- β_c retains the same cross-sectional interpretation.

- We can also estimate β_L and interpret it from

$$E(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}),$$

where β_L represents the expected change in Y over time per unit change in x for a subject.

The OLS estimate of β_c from model (8) is

$$\hat{\beta}_c = \frac{\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})}{\sum_i \sum_j (x_{ij} - \bar{x})^2}. \tag{10}$$

If model (9) is true, then

$$E(y_{ij} - \bar{y}) = \beta_L(x_{ij} - \bar{x}) + (\beta_c - \beta_L)(x_{i1} - \bar{x}_1).$$

In (10),

$$E(\hat{\beta}_c) = \beta_L + (\beta_c - \beta_L)\frac{\sum_i (\bar{x}_{i\cdot} - \bar{x})(x_{i1} - \bar{x}_{\cdot 1})}{\sum_i \sum_j (x_{ij} - \bar{x})^2},$$

where $\bar{x}_{i\cdot} = \sum_j x_{ij}/n$ and $\bar{x}_{\cdot 1} = \sum_i x_{i1}/n$.

- If $\beta_c = \beta_L$ or $\{x_{i1}\}$ and $\{\bar{x}_{i\cdot}\}$ are orthogonal to (independent of) each other, $\hat{\beta}_c$ (based on model (8)) is unbiased.

- In general $\hat{\beta}_c$ is biased for β_L by an amount depending on the correlation between x_{i1} and $\bar{x}_{i\cdot}$.

Exploratory Data Analysis

Gaols of EDA

- Relationship between mean response and covariates (including time).
- Variance, correlation structure, individual-level heterogeneity.

Guidelines for graphical displays of longitudinal data

- Show relevant raw data, not just summaries.
- Highlight aggregate patterns of scientific interest.
- Identify both cross-sectional and longitudinal patterns.
- Identify unusual individuals and observations.

General Techniques

- Scatter plots, use connected lines to reveal individual profiles.
  - Displays of the responses against time
  - Displays of the responses against a covariate (with/without time trend)

- Use smooth curves to reveal mean response profile at the population level.
  - Kernel estimation
  - Smooth spline
  - Lowess

- Variograms for checking variance/covariance structure