

Outline

1. Introduction: longitudinal studies, advantage and challenge.
2. Exploratory data analysis and summary statistics.
3. Methods based on (general) linear models, for continuous outcomes.
 - Weighted least-square (WLS).
 - Maximum likelihood (ML) and restricted maximum likelihood (REML).
 - Linear mixed models (LMM).
4. Methods based on generalized linear models, for binary, count and categorical outcomes.
 - Generalized linear model (GLM), quasi-likelihood (QL).
 - Generalized estimating equations (GEE).
 - Generalized linear mixed models (GLMM).
 - Transition models.
5. Special topics:
 - Time-dependent covariates and causal inference.
 - Missing data.
 - Semi- and non-parametric methods.
 - Non-linear regression models.
 - Bayesian methods.

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.

Orthodontic Measurements

- Orthodontic measurements were taken from 27 children (16 boys and 11 girls) every two year from age 8 to 14.
- Note that here the data is *balanced*, that is, the subjects were measured at the same times with no missing data. Unbalanced data (due to design or missing data) is more common in biomedical studies and introduces extra technical difficulties.
- The following table presents the data in the “wide” form, i.e., one subject per row, with multiple columns representing multiple measurements of the same variable. Alternatively the data can also be presented in the “long” form, that is one row for each time point at which one measurement is taken.

Table 1: Orthodontic measurements over time for 11 girls.

Subject \ Age	8	10	12	14
01	21.0	20.0	21.5	23.0
02	21.0	21.5	24.0	25.5
03	20.5	24.0	24.5	26.0
04	23.5	24.5	25.0	26.5
05	21.5	23.0	22.5	23.5
06	20.0	21.0	21.0	22.5
07	21.5	22.5	23.0	25.0
08	23.0	23.0	23.5	24.0
09	20.0	21.0	22.0	21.5
10	16.5	19.0	19.0	19.5
11	24.5	25.0	28.0	28.0

MACS: CD4+ Cell Number

- HIV attacks CD4+ cell which regulates the body's immune-response to infectious agents; An uninfected individual has around 1100 cells per milliliter of blood.
- In Figure 1, 2376 values of CD4+ cell counts are plotted against time since seroconversion (detectable HIV antibodies) for 369 infected men enrolled in the Multicenter AIDS Cohort Study (MACS).
- Highly unbalanced with irregular observation times and numbers.
- 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.

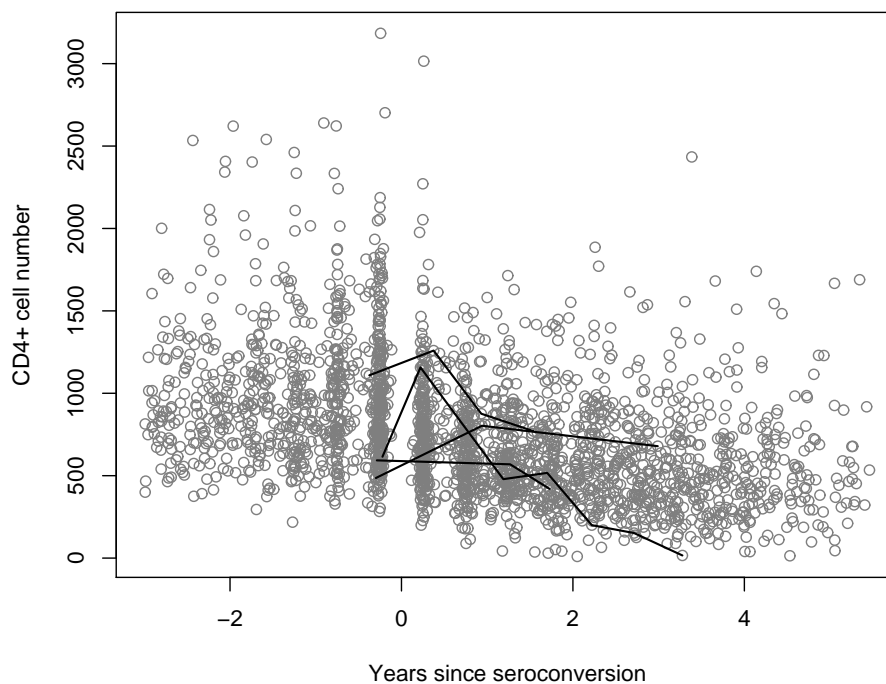


Figure 1: CD4+ cell numbers and time since seroconversion.

Indonesian Children’s Health Study (ICHS)

- Study effects of Vitamin A deficiency (as indicated by xerophthalmia, “dry eyes”) in pre-school children in Indonesia.
- Over 3000 children examined for up to six visits to assess whether they suffer from respiratory infection (RI). Weight and height are also measured.
- Binary responses (infection yes/no).
- Goals:
 - Whether vitamin A deficient children are at increased risk of respiratory infection.
 - Estimate the degree of heterogeneity in the risk of disease among children

Table 2: Summary of 1200 observations of respiratory infection (RI), xerophthalmia and age on 275 children from the ICHS

Xerophthalmia	RI	Age						
		1	2	3	4	5	6	7
No	No	90	236	330	176	143	65	5
	Yes	8	36	39	9	7	1	0
Yes	No	0	2	18	15	8	4	1
	Yes	0	0	7	0	0	0	0

Analgesic Crossover Trial

- Three-period crossover trial of an analgesic drug for pain relieving.
- Three 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.
- Ignoring the order of treatment, pain was relieved for 26% with placebo, 71% with low dose, and 80% with high dose
- A cross-over study, the treatment changes each time.
- Binary outcome.
- Need assessment of carry-over effects.

Table 3: Number of patients for each treatment and response sequence in three-period cross-over trial of analgesic treatment for pain from primary dysmenorrhoea.

	Response sequence in periods 1, 2, 3 (0= no relief; 1= relief)								Total
	000	100	010	001	110	101	011	111	
PLH	0	0	2	2	1	0	9	1	15
PHL	2	1	0	0	0	0	9	4	16
LPH	0	1	1	1	0	8	3	1	15
LHP	0	1	1	1	8	0	0	1	12
HPL	3	0	0	0	1	7	2	1	14
HLP	1	5	0	0	4	3	1	0	14
Total	6	8	4	4	14	18	24	8	86

Epileptic Seizures

- Clinical trial of 59 epileptics.
- For each patients, the number of epileptic seizures was recorded during a baseline period of eight weeks.
- Patients were randomized to be treated with the anti-epileptic drug progabide or placebo.
- Number of seizures was then recorded in four consecutive two-week intervals.
- Does progabide treatment reduce the rate of epileptic seizures?
- Count data.

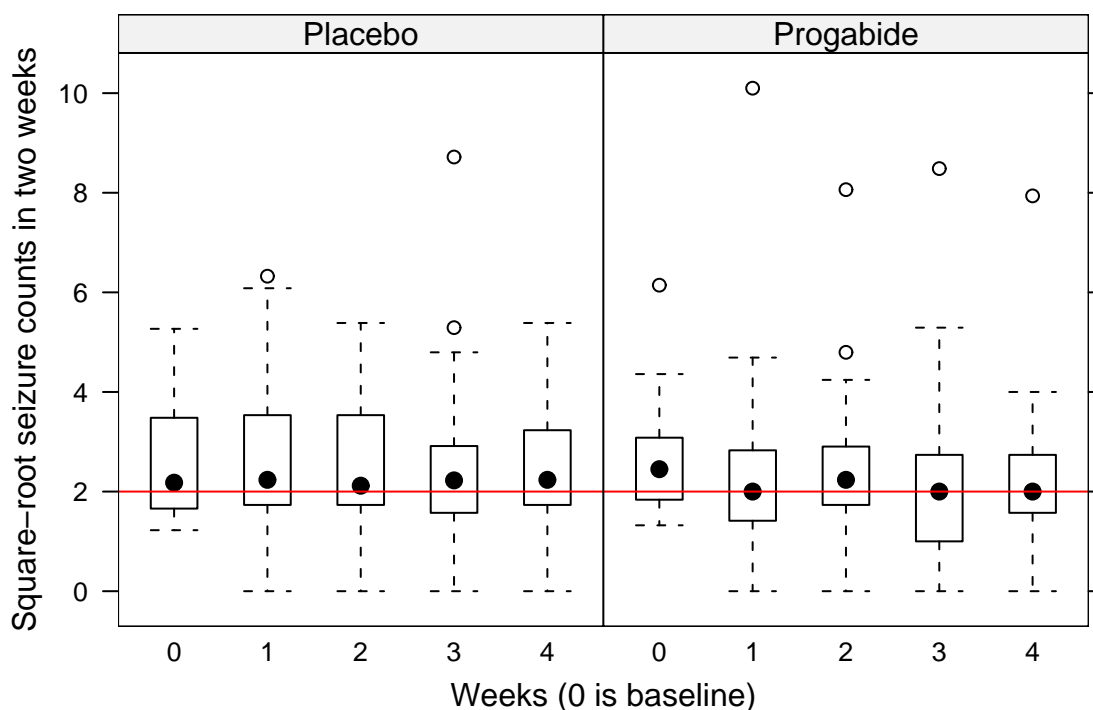


Figure 2: Boxplots of square-root transformed seizure counts per two weeks for epileptics at baseline and for four subsequent two-week periods after the patients were randomized to either placebo or progabide treatment.

Longitudinal vs. Cross-sectional Studies

- 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.
- Longitudinal studies can distinguish the effect due to aging (*i.e.*, changes over time within subject) from cohort effects (*i.e.*, difference between subjects at baseline).

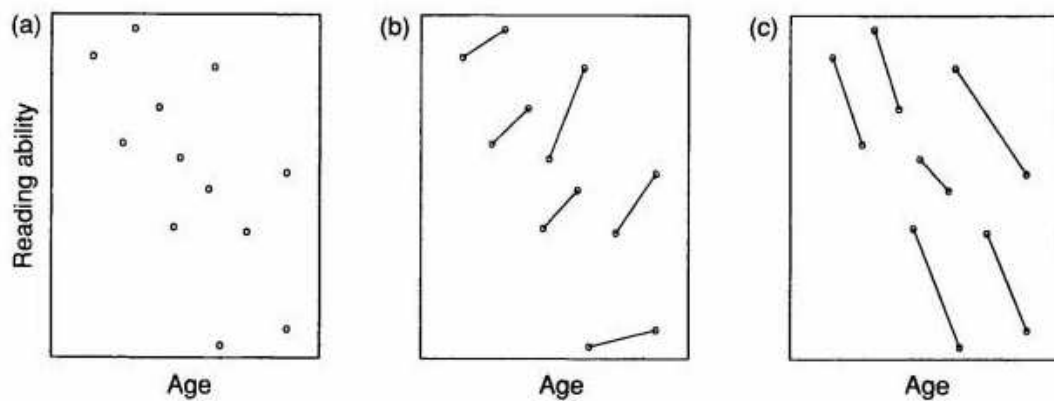


Figure 3: Hypothetical data (Fig 1.1. Diggle et al, 2002).

Bias in cross-sectional studies of change

Consider a model of the form:

$$Y_{ij} = \beta_0 + \beta_C x_{ij} + \epsilon_{ij}, j = 1, \dots, n, i = 1, \dots, m, \quad (1)$$

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

Model (1) can also be written as:

$$Y_{ij} = \beta_0 + \beta_C x_{i1} + \beta_C(x_{ij} - x_{i1}) + \epsilon_{ij}. \quad (2)$$

Note that this model assumes that the baseline effect is the same as the longitudinal effect.

To relax this assumption, we can assume different parameters for the two effects:

$$Y_{ij} = \beta_0 + \beta_C x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij}. \quad (3)$$

Using (3) we can estimate both β_C and β_L and test if $\beta_C = \beta_L$. we can also estimate β_L and interpret it from

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

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

The OLS estimate of β_C ($= \beta_L$) from model (2) is:

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

If the model (3) is true, i.e., $\beta_C \neq \beta_L$, then the expectation of the naive estimator $\hat{\beta}_C$ is:

$$\begin{aligned} E(y_{ij} - \bar{y}) &= \beta_0 + \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) \\ &\quad - (\beta_0 + \beta_c \bar{x}_1 + \beta_L(\bar{x} - \bar{x}_1)) \\ &= \beta_L(x_{ij} - \bar{x}) + (\beta_c - \beta_L)(x_{i1} - \bar{x}_1) \\ E(\hat{\beta}_C) &= \beta_L \frac{\sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \\ &\quad + (\beta_C - \beta_L) \frac{\sum_i \sum_j (x_{ij} - \bar{x})(x_{i1} - \bar{x}_1)}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \\ &= \beta_L + (\beta_C - \beta_L) \frac{\sum_i n(\bar{x}_i - \bar{x})(x_{i1} - \bar{x}_1)}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \end{aligned}$$

where $\bar{x}_i = \sum_j x_{ij}/n$ and $\bar{x}_1 = \sum_i x_{i1}/m$,

- If $\beta_C = \beta_L$ or $\{x_{i1}\}$ and $\{\bar{x}_i\}$ are orthogonal to (independent of) each other, $\hat{\beta}_C$ (based on model 2) 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 .

Longitudinal vs. Cross-sectional Studies (cont.)

- Distinguish the degree of variation in Y across time within a subject from the variation in Y between subjects.
- 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 for a fixed number of subjects, N .
- 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.
- Investigation of individual-level changes.

Correlated Data

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

$$Pr(Y_1, \dots, Y_n | x_1, \dots, x_n, \boldsymbol{\beta}) = Pr(Y_1 | x_1, \boldsymbol{\beta}) \cdots Pr(Y_n | x_n, \boldsymbol{\beta}).$$

That is, the Y 's are *conditionally independent* given the covariates \boldsymbol{x} (and the parameters $\boldsymbol{\beta}$). However, in general, the Y 's are *not* independent marginally, i.e., $\text{Cor}(Y_1, Y_2) \neq 0$ or $Pr(Y_2 | Y_1) \neq Pr(Y_2)$. Longitudinal data is a special case of *correlated data* where $\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta}$ are *not* independently distributed.

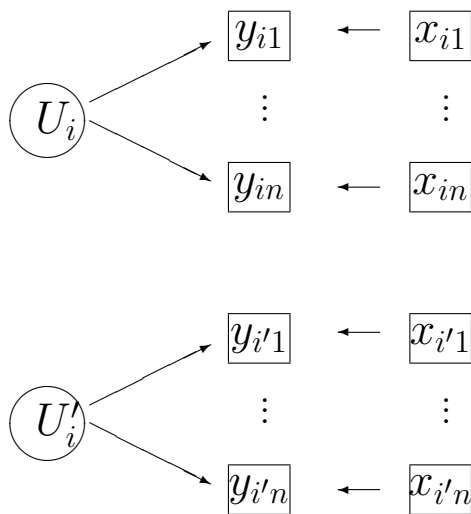
Examples of correlated data

- 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 (cyclic?) 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 the outcome variables. For univariate survival data, the event can only occur once.

Characteristic of Longitudinal Data

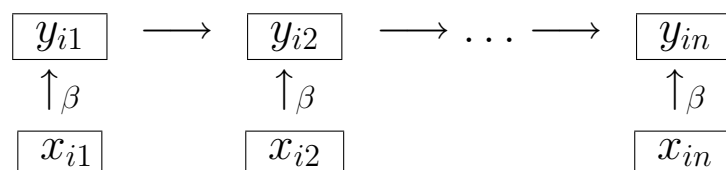
- When the measurements are taken is not of primary interest and is considered fixed by design. The fact that people have different number of observations and at different times is often ignored under certain assumptions.
- Small number of observations per subject but relatively large number of subjects.
- The variability can be divided into three components:
 1. Heterogeneity between individuals.
 2. Serial correlation, measurements closely spaced are more similar.
 3. Measurement error.

Random effects/latent variable



$$Y_{ij} = x_{ij}\beta + U_i + \epsilon_{ij}; \quad U_i : \text{unobserved}$$

Serial correlation



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

Notations

We will mostly follow the notation in our textbook (DHLZ).

- Vectors: \mathbf{x} , \mathbf{Y} , $\boldsymbol{\beta}$. When we say a n -vector, it is an array with n rows and 1 column (i.e., a column vector):

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

- Matrices (upper-case): \mathbf{X} , $\boldsymbol{\Sigma}$.
- *Parameters* are represented by Greek letters.
- Random variables are upper-cased: Y_i , \mathbf{Y} , for scalar and vectors, respectively.
- Observed (non-random) variables are lower-cased: x_i , \mathbf{x} , for scalar and vectors, respectively.
- Let $i = 1, \dots, m$ index *subjects*.
- For each subject i , there are n_i observations at times t_{ij} , $j = 1, \dots, n_i$.
- \mathbf{x}_{ij} is a p -vector that are the covariates for observation j of subject i . \mathbf{X}_i is a $n_i \times p$ matrix of all the covariates for i and \mathbf{X} is all the covariates.
- The outcome for subject i is denoted by the n_i -vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ with mean $\boldsymbol{\mu}_i$ and $n_i \times n_i$ covariance matrix \mathbf{V}_i where $v_{ijk} = \text{Cov}(Y_{ij}, Y_{ik})$. The correlation matrix is \mathbf{R}_i .
- $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_m)^T$ is an N -vector with $N = \sum_{i=1}^m n_i$.

Analysis of Longitudinal Data

The main challenge involves how to take into account the correlation structure.

- Summary statistics based approach: calculate a univariate summary statistic for the multiple measurements and then it can be analyzed as a function of covariates.
 - 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:

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad (4)$$

$$\text{Var}(\mathbf{Y}_i) = \mathbf{V}_i(\boldsymbol{\alpha}), \quad (5)$$

where both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ must be estimated, and \mathbf{V}_i may also depend on some covariates.

- Conditional approach (random effects model, hierarchical model, multi-level model): assumes correlation arises because of heterogeneity in subjects, that is,

$$\mathbf{Y}_i | \mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \sigma^2\mathbf{I}) \quad (6)$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \quad (7)$$

- Transition model:

$$E(Y_{ij} | Y_{ij-1}, \dots, Y_{i1}, \mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha Y_{ij-1}. \quad (8)$$

Further Reading

- Chapter 1 of textbook (Diggle et al 2002).
- Optional: Chapter 1 of Hedeker and Gibbons (2006).