3. Application to Longitudinal Data: Part I

3.1. Preliminary

- Longitudinal studies

The defining feature of longitudinal study is that individuals are measured repeatedly over time.

Example 3.1. A hypothetical example.

Example 3.2 (CD4+ cell numbers).
- Kaslow *et al.* (1987, *Amer. J. Epidem.*).
- HIV attacks CD4+ - an immune cell.
- 2376 CD4+ numbers plotted against time since seroconversion for 369 infected men.
- Uninfected individual has about 1,100 cells/millilitre.
- The number decreases over time as disease progresses.

- Exploring longitudinal data

Unlike time-series, longitudinal data is both cross-sectional and longitudinal - both patterns should be explored.

Graphical tools

a. Scatter plots against time

b. Residual plots

c. Smooth curves

e. g., smoothing splines, kernel est., lowess (local weighted LS lines; Cleveland 1979).

Time series analysis

e. g., Autocorrelation function

$$\widehat{\rho}_i(s) = \text{cor}(y_{it}, y_{i\underline{t+s}}).$$

3.2. A mixed linear model for longitudinal data

- Motivation

Modelling the mean response as well as the correlation structure

Example 3.3. Let y_{ij} be the observation from subject i at time t_j, $i = 1, \dots, m$, $j \in J_i \subset J = \{1, \dots, b\}$. Suppose

$$y_{ij} = x'_{ij}\beta + u_i + e_{ij},$$

where x_{ij} is a vector of subject/time dependent covariates, β is a vector of regression coefficients, u_i is a subject-specific random effect, and e_{ij} is an (unexplained) error.

If u_i's and e_{ij}'s are indep., $u_i \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, we have $\text{cor}(y_{ij}, y_{i'j'}) = 0$ if $i \neq i'$,

$$\text{cor}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}, \quad j \neq k.$$

• Serial correlation?

Example 3.4. Suppose that the observation times are equally spaced. W.l.o.g., let $t_j = j$. Consider the model

$$y_{ij} = x'_{ij}\beta + u_i + w_{ij} + e_{ij},$$

where u_i and e_{ij} are the same as before, and w_{ij} corresponds to a serial correlation, which follows an AR(1) process:

$$w_{ij} = \phi w_{i\underline{j-1}} + z_{ij},$$

where $\phi \in (0, 1)$ is a constant, and z_{ij}'s are indep. $\sim N\{0, \sigma_w^2(1-\phi^2)\}$.

Assume that u, w and e are indep. Then,

$$\text{cov}(y_{ij}, y_{ik}) = \sigma_u^2 + \sigma_w^2\phi^{|j-k|} + \sigma_e^2\delta_{j,k},$$

where $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise.

• A general linear mixed model

Let y_i denote the vector of observations from the ith subject. Suppose $y_i, \ldots, y_m$ are indep.,

$$y_i = X_i\beta + Z_i v_i + e_i,$$

where y_i is $n_i \times 1$, X_i is $n_i \times p$, Z_i is $n_i \times r_i$; $\mathrm{E}(v_i) = 0$, $\mathrm{Var}(v_i) = G_i$; $\mathrm{E}(e_i) = 0$, $\mathrm{Var}(e_i) = R_i$; and $\mathrm{cov}(v_i, e_i) = 0$.

• Parametric covariance model

$G_i = G_i(\theta)$, $R_i = R_i(\theta)$, where θ is a vector of variance components.

• Nonparametric covariance model

G_i, R_i, or $V_i = \mathrm{Var}(y_i)$ unspecified.

3.3. Estimation of the mean

- Weighted least squares (WLS)

Write $y = (y_i)_{1\le i\le m}$, $X = (X_i)_{1\le i\le m}$.

$$\min_\beta\{(y - X\beta)'W(y - X\beta)\},$$

where W is a given weighting matrix.

If $X'WX$ is non-singular, the solution is

$$\hat{\beta}_W = (X'WX)^{-1}X'Wy.$$

- Some properties of WLS

(i) Unbiased: $\mathrm{E}(\hat{\beta}_W) = \beta$.

(ii)
$$\begin{aligned}&\mathrm{Var}(\hat{\beta}_W)\\ &= (X'WX)^{-1}X'WVWX(X'WX)^{-1},\end{aligned}$$

where $V = \mathrm{Var}(y) = \mathrm{diag}(V_1, \dots, V_m)$.

(iii) $W = I$ gives the ordinary least squares (OLS) estimator.

Note. Although OLS est. is the same as LS est. in linear regression in terms of point estimation, its covariance matrix is different:

$$\text{Var}(\widehat{\beta}_{\text{OLS}}) = (X'X)^{-1}X'VX(X'X)^{-1}$$

(not $\sigma^2(X'X)^{-1}$ as in linear regression).

- What W?

The optimal W is $W = V^{-1}$. However, V is unknown.

3.4. Estimation of covariances under normality

- Balanced data: $V_i = V_0$, $1 \leq i \leq m$.

V_0 can be estimated by ML or REML method in linear mixed models.

Example 3.5 (Growth of trees).
- Diggle *et al.* (1996).
- 79 Sitka spruce trees; 54 grown with ozone (70 ppb), 25 control.
- 4 growth chambers: ozone (27, 27), control (12,13).
- Two growth periods: 1988 (five obs. times), 1989 (eight obs. times).

<u>Models for the mean</u>

For 1988 data:

$$\begin{aligned}\mu_1(t_j) &= \beta_j, \quad i = 1, \ldots, 5; \\ \mu_2(t_j) &= \beta_j + \tau + \gamma t_j, i = 1, \ldots, 5.\end{aligned}$$

For 1989 data:

$$\begin{aligned}\mu_1(t_j) &= \beta_j, \quad i = 6, \ldots, 13; \\ \mu_2(t_j) &= \beta_j + \tau, \quad i = 6, \ldots, 13.\end{aligned}$$

<u>REML est. of the covariances</u>

• In general (not necessarily balanced data), if a parametric model is assumed for the covariances, say, $G_i = G_i(\theta)$, $R_i = R_i(\theta)$. Estimate θ by ML or REML.

Example 3.6 (Effect of air pollution episodes on children).
- Laird & Ware (1982).
- Approx. 200 school children examined under normal conditions, then during an air pollution alert and three successive weeks following the alert.
- Objective was to determine whether FEV_1 (volume of air exhaled in the 1st sec. of a forced exhalation) was depressed during the alert.

Model I

$$y_{ij} = \beta_j + u_i + e_{ij},$$

$i = 1, \ldots, m$, $j = 1, \ldots, 5$. It is assumed that u_i's and e_{ij}'s are indep. with $u_i \sim N(0, \sigma^2)$, $e_{ij} \sim N(0, \tau^2)$.

Then, we have a parametric covariance model with $\theta = (\sigma^2, \tau^2)'$.

Model II

Another random effet is introduced to quantify the mean decline in FEV_1 for each child. The model can be expressed as

$$y_i = X_i\beta + Z_i v_i + e_i,$$

where $X_i = I_5$, $v_i = (v_{i1}, v_{i2})' \sim N(0, D)$, where D is a 2×2 unknown covariance matrix, and

$$Z' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Again, we have a parametric covariances with

$$\theta = (\sigma_1^2, \sigma_2^2, \rho, \tau)',$$

where $\sigma_j^2 = \text{var}(v_{ij})$, $j = 1, 2$, $\rho = \text{cor}(b_{i1}, b_{i2})$, and $\tau^2 = \text{var}(e_{ij})$; $G_i(\theta) = Z_i D Z_i'$ and $R_i(\theta) = \tau^2 I_5$.

REML was used to estimate θ.