STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

March 27, 2018

Linear Mixed Models

- Learning Objectives
  - Define fixed and random effects
  - Write out the models used and the assumptions for inference
  - Develop a statistical toolbox for analyzing linear mixed models
  - Interpret the respective R outputs
- *Reference: SJS, Chapter 10*

Example I: Orthodontics Growth Data

- Study conducted at Department of Orthodontics from North Carolina Dental School
- Followed growth of 27 children (16 males, 11 females)
- Measured at ages 8, 10, 12 and 14
- Response: Distance (in mm) from the centre of the pituitary to the pterygomaxillary fissure
- Interest: Model distances in terms of age and sex
- What are the fixed effects?
- What are the random effects?

Example I: Orthodontics Growth Data

```
Grouped Data: distance ~ age | Subject
  distance age Subject  Sex
1      26.0   8     M01 Male
2      25.0  10     M01 Male
3      29.0  12     M01 Male
4      31.0  14     M01 Male
5      21.5   8     M02 Male
6      22.5  10     M02 Male
Grouped Data: distance ~ age | Subject
   distance age Subject    Sex
65      21.0   8     F01 Female
66      20.0  10     F01 Female
67      21.5  12     F01 Female
68      23.0  14     F01 Female
69      21.0   8     F02 Female
70      21.5  10     F02 Female
```

Example I: Mixed Model

$$
\begin{aligned}
Distance_{ijk} = \; & \beta_0 \\
& + \beta_1 \mathbf{I}_{[sex=male],j} \\
& + \beta_2 \mathbf{I}_{[age=10],k} + \beta_3 \mathbf{I}_{[age=12],k} + \beta_4 \mathbf{I}_{[age=14],k} \\
& + \beta_5 \mathbf{I}_{[sex=male],j} * \mathbf{I}_{[age=10],k} \\
& + \beta_6 \mathbf{I}_{[sex=male],j} * \mathbf{I}_{[age=12],k} \\
& + \beta_7 \mathbf{I}_{[sex=male],j} * \mathbf{I}_{[age=14],k} \\
& + u_{ij} \\
& + \epsilon_{ijk}
\end{aligned}
$$

where

- $Distance_{ijk}$: distance at time k on subject i in treatment j
- u_{ij}: random effect due to subject i of sex j
- ϵ_{ijk}: random error

Example II: Carbohydrates in Diabetes

- Diet study on $n=71$ persons with Type 2 diabetes
- Each person was assigned to 1 of 3 treatment (diet) groups:
  I) HG: high GI (glycemic index)
  II) LG: low GI
  III) HM: high in monosaturated fats
- Traced for 6 months: measurements taken at 0, 3 and 6 months

Example II: Data

- The first 10 observations:

Obs	id	diet	hdl1	hdl2	hdl3
1	1	LG	0.97	1.14	1.45
2	3	HG	0.85	0.85	0.84
3	5	HM	1.11	1.36	1.25
4	6	HM	0.95	0.99	0.96
5	8	LG	0.78	0.80	0.72
6	9	HG	0.71	0.76	0.78
7	10	LG	0.58	0.69	0.72
8	11	HM	1.24	1.24	1.31
9	12	HM	0.93	1.18	0.98
10	13	HG	1.65	1.23	1.24

- Variables of interest: ID#, Diet, Time
- Outcome of interest: HDL- level of "good" cholesterol
- Aim: Is there a diet$*$ time interaction? Do differences among diets change over time?

Example II: Mixed Model

$$
\begin{aligned}
Y_{ijk} = \; & \beta_0 \\
& + \beta_1 \mathbf{I}_{[diet=HG],j} + \beta_2 \mathbf{I}_{[diet=HM],j} \\
& + \beta_3 \mathbf{I}_{[time=1],k} + \beta_4 \mathbf{I}_{[time=2],k} \\
& + \beta_6 \mathbf{I}_{[diet=HG],j} * \mathbf{I}_{[time=1],k} + \beta_7 \mathbf{I}_{[diet=HG],j} * \mathbf{I}_{[time=2],k} \\
& + \beta_8 \mathbf{I}_{[diet=HM],j} * \mathbf{I}_{[time=1],k} + \beta_9 \mathbf{I}_{[diet=HM],j} * \mathbf{I}_{[time=2],k} \\
& + u_{ij} \\
& + \epsilon_{ijk}
\end{aligned}
$$

Fixed

Random

Random error

where

- Y_{ijk}: response at time k on subject i in treatment j
- u_{ij}: random effect due to subject i under diet j
- ϵ_{ijk}: random error

Variance-Covariance Structure

Assume:

- Subjects are independent and
- u_{ij} are $i.i.d.$ $N(0, \sigma_u^2)$
- ϵ_{ijk} are $i.i.d.$ $N(0, \sigma_e^2)$
- u_{ij} and ϵ_{ijk} are independent

within

Therefore:

- Different subjects, $i \neq l$:

$$Cov(Y_{ijk}, Y_{lmn}) = 0$$

Subj. $\boxed{T1 \quad T2 \ldots T_k}$

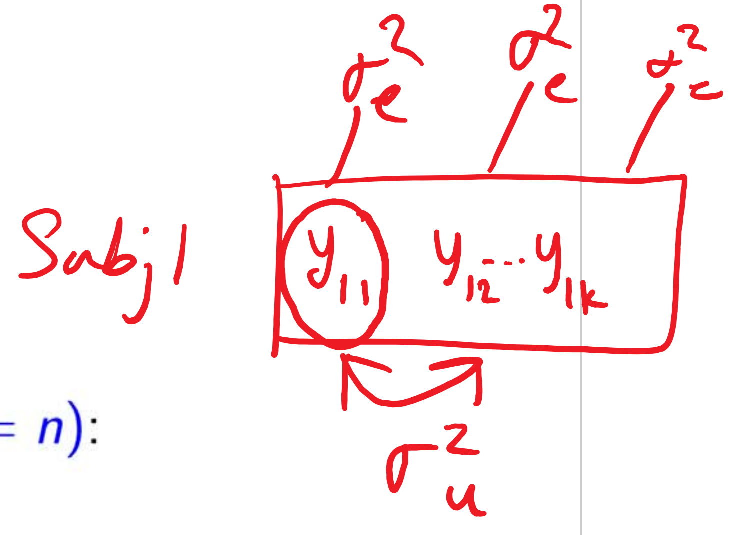
$i = 1, \ldots, n$

n subjects, k obs per subj.

Total: nk observations

Eg 1	Eg 2
n = 27	n = 71
k = 4	k = 3
↓	↓
108	213

Variance-Covariance Structure



- Same subject at the same time ($i = l, j = m, k = n$):
  $Var(Y_{ijk}) = Cov(Y_{ijk}, Y_{lmn}) = \sigma_e^2 + \sigma_u^2$

- Same subject at different time ($k \neq n$):

$$
\begin{aligned}
Cov(Y_{ijk}, Y_{ijn}) = &\, Cov(u_{ij}, u_{ij}) \longrightarrow Var(u_{ij}). \\
&+ Cov(u_{ij}, e_{ijn}) \to 0 + Cov(e_{ijk}, u_{ij}) \to 0 \\
&+ Cov(e_{ijk}, e_{ijn}) \to 0 \\
= &\, \sigma_u^2
\end{aligned}
$$

u_{ij}, ϵ_{ijk}

$\leftarrow u \perp e$

$$
D_{k\times k} = \begin{array}{c}
\begin{array}{cccc} T_1 & T_2 & \cdots & T_k \end{array} \\
\begin{pmatrix}
\sigma_e^2+\sigma_u^2 & \sigma_u^2 & & \sigma_u^2 \\
\sigma_u^2 & \sigma_e^2+\sigma_u^2 & & \sigma^2 \\
\vdots & & \ddots & \\
\sigma_u^2 & \sigma_u^2 & \cdots & \sigma_e^2+\sigma_u^2
\end{pmatrix}
\end{array}
$$

Compound Symmetry Variance-Covariance Structure

Intraclass Correlation Coefficient:

$$\rho_{IC} = \frac{Cov(Y_{ijk}, Y_{ijn})}{\sqrt{Var(Y_{ijk})Var(Y_{ijn})}} = \underbrace{\frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}}$$

$\sigma_e^2 I$

$$\begin{pmatrix} (\sigma_e^2+\sigma_u^2) & & & \\ & (\sigma_e^2+\sigma_u^2) & & \\ & & \ddots & \\ & & & (\sigma_e^2+\sigma_u^2) \end{pmatrix}$$

- σ_u^2 the correlation is the same for every pair of observations on the same subject
- $\sigma_e^2+\sigma_u^2$ the variance is the same for all observations

Var(**Y**)=

$$\begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_n \\ \hline S_1 & D_{k\times k} & (0)_{k\times k} & \cdots & 0_{k\times k} \\ S_2 & 0_{k\times k} & D_{k\times k} & \cdots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ S_n & 0_{k\times k} & 0_{k\times k} & \cdots & D_{k\times k} \end{array} \quad nk\times nk$$

Eg, 108×108

213×213

$$D_{k\times k} = \begin{pmatrix} \sigma_e^2+\sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots \\ \sigma_u^2 & \ddots & \cdots & \sigma_u^2 \\ \vdots & & \ddots & \vdots \\ \sigma_u^2 & \cdots & \cdots & \sigma_e^2+\sigma_u^2 \end{pmatrix}_{k\times k}$$

Linear Mixed model for repeated measures data

MIXED Model:

nk

$$\mathbf{Y} = \underbrace{f(\mathbf{X}, \beta)}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \underbrace{\epsilon}_{\text{random noise}}$$

$\epsilon_{ijk} \sim N(\quad)$

Part	Properties	*Eg.I*	*Eg.II*
Response, Y	continuous, $Y \sim$ *Normal*	*HDL*	*Distance*
Fixed effects, X	continuous or categorical	*diet, time*	*sex, age*
Random effect, u	$\mathbf{u} \sim N(0, \sigma_u^2)$	*subject*	*subject*
Random error, ϵ	$\mathbf{e} \sim N(0, \sigma_e^2)$, $u \perp e$		

- Random effect induces a random intercept (different intercept for different subjects)

Restricted Maximum Likelihood Estimation

- Restricted Maximum Likelihood (or Residual ML): gives unbiased estimators of variance-covariance parameters
- based on the notion of separating the likelihood used for estimating Σ from that used for estimating $\boldsymbol{\beta}$
- Steps:
  (1) Estimate variance-covariance parameters using maximum likelihood estimation
  (2) Use the estimates from (1) and estimate β's of fixed effects using generalized least squares (GLS) where:
    - GLS is least squares optimization with adjustment since $\text{Var}(\mathbf{Y}) = V \neq \sigma^2 \mathbf{I}$
    - We get $\widehat{\beta} = (X^\top V^{-1} X)^{-1} X^\top V^{-1} Y$ and $\text{Var}(\widehat{\beta}) = (X^\top V^{-1} X)^{-1}$
  (3) Repeat steps (1) and (2) until convergence

$$\begin{pmatrix} (\sigma^2) & & 0 \\ & (\sigma^2) & \\ 0 & & (\ ) \end{pmatrix}$$

Within-subject Covariance structures

- Between-subject: Recall that we assumed that subjects are independent, so between-subject covariance is 0, i.e.,
$$Cov(Y_{ijk}, Y_{lmn}) = 0 \text{ if } i \neq l$$
- Within-subject:

$\sigma^2_e + \sigma^2_u$

Type	Interpretation	# cov. parameters
CS	-same variances and common covariance — σ^2_u	2
UN	-different variances and difference covariances	(t*(t+1))/2
AR(1)	-same variances, covariance decrease exponentially with distance	2

Eg, $k=4$

$\frac{k(k+1)}{2} = \frac{4(5)}{2} = 10$

Within-subject Covariance structures

- Compound Symmetry [2]: same variance and common covariances

σ_e^2, σ_u^2

$$D_{CS} = \begin{bmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 \end{bmatrix}$$

- Unstructured $[t(t+1)/2]$: different variances and different covariances

Eg, $t=3$.

$$D_{UN} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{12}^2 & \sigma_2^2 & \sigma_{23}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_3^2 \end{bmatrix}$$

- Auto-Regressive, lag1 [2]: same variances, covariances decrease exponentially

σ^2, ρ

$-1 \le \rho \le 1$

$$D_{AR(1)} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Comparing models

- Using likelihood-based criteria (due to MLE): compare models with same Y and X's but different covariance structures
  - AIC$= -2 \,\text{Res} \log \mathcal{L} + 2$(# of covariance parameters)
  - BIC$= -2 \,\text{Res} \log \mathcal{L} + $(# of covariance parameters)$\log(n)$,
    where n=# of subjects
  - $G^2 = -2 \,\text{Res} \log(\frac{\mathcal{L}_{\mathcal{R}}}{\mathcal{L}_{\mathcal{F}}})$
- Using t and F tests (due to GLS): check relevance of fixed effects (test β's)