STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

March 22, 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*

What are repeated measures? y_i

- more than one observation per subject or experimental unit
- Clustered Data
  - members belong to groups (clusters) (Eg. a family, a class)
  - response is measured for each subject
  - Egs, STA scores of students by classrooms, birthweight of rats by litter

Cluster S1 S2 S3 S4...

- Cross-over study
  - experiment where each subject gets each treatment
  - Egs, weight loss of subjects on 2 diets, 2 glaucoma treatments applied to eyes of dogs

Sub1 Trt 1 Trt 2

- Longitudinal data
  - subject followed over time
  - observations at a few, regular intervals (in contrast to time series)

frequent, regular

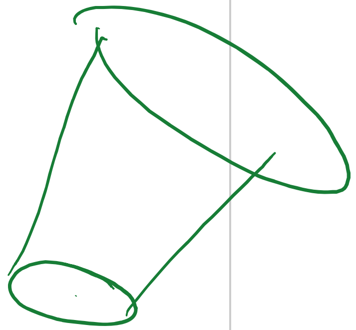
Sub1 Time1 Time2...

- What's a key property of repeated measures?

Fixed and Random Effects

MIXED.

- **Fixed effect**: levels of each variable are of specific interest
  - e.g., gender, age group, treatment, age, height, race
- **Random effect**: levels of variable are randomly sampled from a large population being studied; not of specific interest
  - e.g.,randomly selected subjects, size of randomly selected supermarkets



Aim:

- Comparison within subject/ unit
- Comparison across (between) subjects/ units or groups of subjects/ units

Between and Within-subject Effects

Fixed. Random effect of subject

EFFECT	"BETWEEN-SUBJECT"	"WITHIN-SUBJECT"
Factors on which there are	-no repeated measures -constant within a subject -a single level	-repeated measures -several varying values
Example	race, sex	time, treatment

- Key issue: repeated observations are not independent; expect observations on one subject to be correlated

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
- y_i 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? Age, Sex, Sex * Age
- What are the random effects? Subject

Example I: Orthodontics Growth Data

y x_1 z_1 x_2

```
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
```

Longitudinal data study setting

- Typical:
  - n subjects, $i = 1, \ldots, n$
  - J treatments, $j = i, \ldots, J$
  - Measured K times, $k = 1, \ldots, K$
  - Outcome: Y_{ijk}- continuous

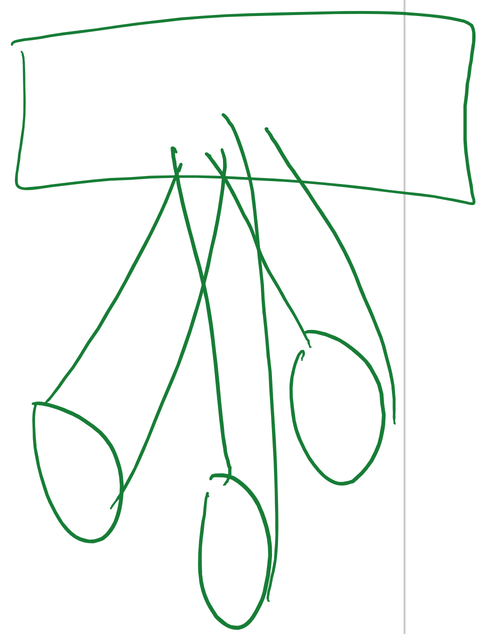
- Example I:
  - $n = 27$
  - $J = 2$ treatments: Sex- males, females
  - $K = 4$ times: Ages 8, 10, 12 and 14
  - Y_{ijk}- Distance for ith subject of sex$j$ at time k

Model for longitudinally repeated data

- Assume $Y_{ijk} \sim$ a Normal distribution → Linear model
- "FIXED" Effects: unknown β's are constant parameters; levels of each variable are of specific interest
  - $J-1$ indicator variables for treatments
  - $K-1$ indicator variables for time
  - $(J-1)(K-1)$ time*treatment interaction terms

Issue: The K observations on each subject are not independent!

- Hence, treat subject as a "RANDOM" Effect
  - subjects are selected at random from a large population
  - not of specific interest
  - model as a random variable $U \sim N(0, \sigma_u^2)$



Model for longitudinally repeated data

- Benefits:
  - Can test $H_0 : \sigma_u^2 = 0$ to see if subjects differ
  - Inference extends beyond subjects measured to entire population of interest
  - Key: Account for "within-subject" correlation
- Called: MIXED model since it includes both *fixed* and *random* effects

Example I: Mixed Model

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

Age

Age $* \; 1_{sex}$

where

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

$i=1,\dots,n$

- The first 10 observations:

j k y

ID #	Diet	Season	Time	Weight	Hemo	Gluc	Chol	HDL	TG	LDL
1	LG	1	1	76.1	0.0770	7.9	4.98	0.97	3.54	2.40
1	LG	1	2	76.5	0.077	7.5	5.63	1.14	1.96	3.60
1	LG	1	3	76.7	0.0790	6.7	5.96	1.45	4.61	2.41
3	HG	1	1	73.6	0.0725	11.9	5.58	0.85	1.42	4.08
3	HG	1	2	73.1	0.065	10.6	5.61	0.85	2.09	3.81
3	HG	1	3	72.9	0.0680	11.9	5.21	0.84	1.34	3.76
5	HM	1	1	78.2	0.0690	7.8	5.88	1.11	2.13	3.80
5	HM	1	2	75.5	0.056	7.2	6.27	1.36	1.19	4.37
5	HM	1	3	76.7	0.0585	7.4	6.84	1.25	1.88	4.74
6	HM	1	1	64.6	0.0750	4.1	5.66	0.95	1.31	4.11

- Variables of interest: ID#, Diet, Time

y_{ijk}

- 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

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

where

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