STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

April 4, 2018

Repeated measures / Mixed Model Diagnostics

- What procedures do we use to:
  - Estimate parameters in a general linear mixed model?
    - Restricted Maximum likelihood estimation for variance and covariance parameters
    - Generalized least squares for fixed effects
  - Carry out inference (significance tests and C.I.s)
    - t and F tests based on the Normal distribution for fixed effects

What are the conditions for inference to be valid?

- Observations on different subjects are independent but observations on the same subject are correlated
- Correct form of the model:
  - fixed between Y and X's
  - random covariance structure for observations on same subject
- Error variance can be modelled to vary across X's, e.g., across different sexes

  Eg. Model 10.5 $\hat{\sigma}_{e,M}$ $\hat{\sigma}_{e,F}$
- Normally distributed error terms and random effects: this implies no outliers
- Large enough sample sizes for LR tests to compare nested models (with same Y and X's but different var-cov structure)

Within-subject Covariance structures

- CS [2]: same variance and common covariances

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

$\hat{\sigma}_{e,F}$

$\hat{\sigma}_{e,M}$

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

\# of parameters

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

- AR1 [2]: same variances, covariances decrease exponentially

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

Comparing models

- Using likelihood-based criteria: compare models with same Y and X's but different covariance structures
  - AIC$= -2$ Res $\log \mathcal{L} + 2(\#$ of parameters)
  - BIC$= -2$ Res $\log \mathcal{L} + (\#$ of parameters)$\log(n)$,
    where $n = \#$ of subjects
  - $\boxed{G^2 = -2 \text{ Res} \log(\frac{\mathcal{L}_{\mathcal{R}}}{\mathcal{L}_{\mathcal{F}}})} \sim \chi^2_{df_0}$
- Using t and F tests: check relevance of fixed effects

— Interaction btw. Age & Sex is significant.
($p = 0.0057$).

Checking Residuals

$$\boxed{y_i = f(x_i;\beta) + u_i + \epsilon_i}$$

$$\epsilon_i = y_i - f(x_i;\beta)$$

1, 2 4

- Marginal residuals- interested in quantities averaged over all levels of the random effect. If predictor, X is quantitative, use plot of residuals vs X to see if linear form is appropriate.
- Conditional residuals- interested in effects for a particular subject (a level of the random effect). Use to check for normality.
- Cholesky (Studentized) residuals- standardized residuals (zero correlation, variance is one); helpful for identifying outliers

$$\epsilon_i = y_i - (f(X_i;\beta) + u_i).$$

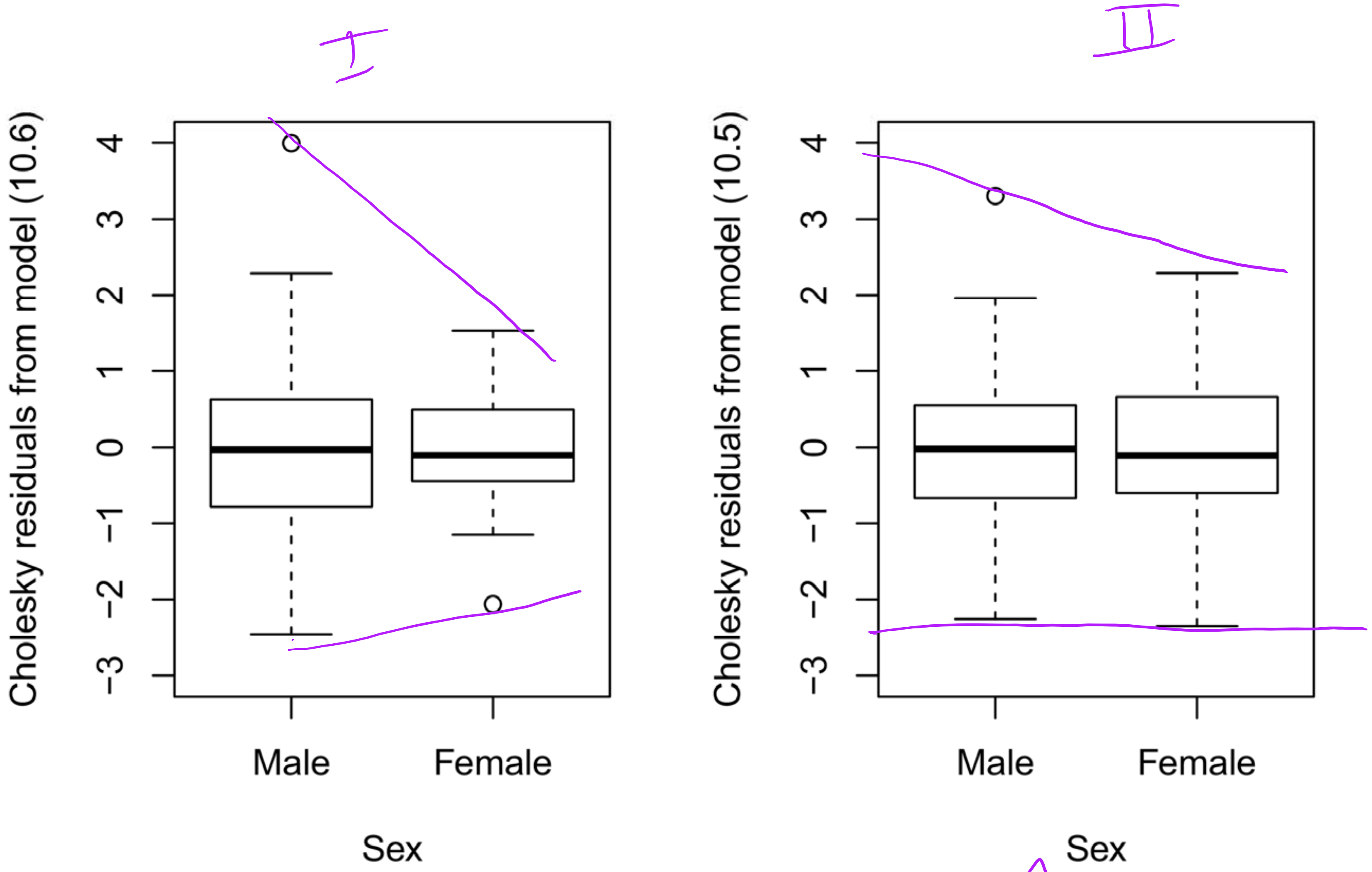
3 4

$$e_i = \epsilon_i + u_i$$

$$= y_i - f(x_i;\beta)$$

1 2

Plot of Cholesky Residuals



$\hat{\sigma}_e$

$\hat{\sigma}_{e,F}$

$\hat{\sigma}_{e,M}$

Summary of MM Example 1 Analysis

- Significant interaction between Sex and Age (p=0.0057, model 10.5)
- Final model established has compound symmetry (CS) structure for within-subject effect, varying with sex, that is, same variance on each subject with same sex/ age but different variances for each sex.
- (10-5) Final model has smaller AIC. Residuals are proper.
- In general, choose a parsimonious model that makes sense and is easier to interpret.

Course Summary

What did we cover in our course-
STA 303/1002: Methods of Data Analysis II?

Cases and Methods

Case	Title	Method(s)
I	Spock's Trial	Two-sample T-test
		General Linear Model (1-way)
		Multiple comparisons (Bonf., Tukey).
II	The Pygmalion Effect	General Linear Model (2-way)
III	The Donner Party	Binary Logistic Regression
IV	The Krunnit Islands	Binomial Logistic Regression
V	Mating Elephants	Poisson Regression
VI	Heart Study	Difference in proportions (2-way CT).
		Pearson's TOA or LRT (2-way CT)
		Multinomial Logistic model (2-way CT)
		Log-linear model (2-way CT)
VII	Three Drugs	Log-linear model (3-way CT)
VIII	Orthodontic Growth	General Linear Mixed Model
IX	Carbs in Diabetes	General Linear Mixed Model (PP. / HW).

Outcomes / Responses

STA 302 LM cts any (mostly cts).

Method	Y	X	Dist. of Y
General LM	continuous^{1}	categorical	Normal
Binary Logit	binary	any	Bernoulli
Binomial Logit	counts	any	Binomial
Poisson	counts	any	Poisson
Contingency (2-way)	counts	categorical	Multinomial
(2-way and 3-way)	counts	categorical	Poisson
Mixed	continuous$^{>1}$	any	Normal

>1 observation per person

General Linear Model and Assumptions

$$Y = f(\mathbf{X}; \beta) + \epsilon$$

OR

$$g(E(Y)) = f(\mathbf{X}; \beta), \text{ with } g(\cdot) = 1$$

- Y is a linear function of β's
- Correct form of the model along with:
  - Observations are independent
  - $\epsilon_i \sim N(0, \sigma^2)$: errors have/are
    - zero expectation
    - constant variance
    - uncorrelated
    - jointly normal

  So no outliers or heavy/light tails, or additional X's

Logistic Regression and Assumptions

$$\log\left(\frac{\pi}{1-\pi}\right) = f(\mathbf{X};\beta) \cancel{+ \epsilon}$$

OR

$$g(E(Y)) = f(\mathbf{X};\beta), \quad \text{with } g(\cdot) = \mathit{logit}$$

- Y is a linear function of β's
- Correct form of model along with:
  - Observations are independent
  - Variance follows Bernoulli / Binomial distribution form
  - No outliers
  - Sample size is large

Poisson Regression/ Log-linear Model and Assumptions

$$\log(\mu) = f(\mathbf{X}; \boldsymbol{\beta}) \cancel{+ \boldsymbol{\epsilon}}$$

OR

$$g(E(Y) = \mu) = f(\mathbf{X}; \boldsymbol{\beta}), \text{ with } g(\cdot) = \log$$

- Y is a linear function of β's
- Correct form of model along with:
  - Observations are independent
  - Variance= Mean
  - No outliers
  - Sample size is large

General Linear Mixed Model

$$Y = f(\mathbf{X}; \beta) + \underline{u} + \epsilon$$

- Y is a linear function of β's, u is the random effect; identity link
- Correct form of model including:
  - Observations on different subjects are independent but observations on the same subject are correlated
  - Error variance can be modelled to vary across X's
  - Normal error and random effects (so no outliers)
  - Large sample sizes for LR tests to compare nested models (with same Y and X's but different var-cov structure)

Estimation and Inference Procedures

GLM

Regression	Estimation	Inference
General LM	Least Squares (LS)	F, t
Logistic	MLE	LRT, Wald
Poisson	MLE	LRT, Wald
Mixed (random)	ML	LRT
(fixed)	Generalized LS	F, t

Main R procedures

1. `~~proc~~ lm( . )`
2. `~~proc~~ anova`
3. `~~proc~~ glm`
4. `~~proc~~ lme`

Model Extensions

- Other link functions- e.g., log-log, gamma
- Penalized Regression- for model selection with high-dimensional data STA 414
- Principal Component Analysis- for correlated observations STA 437
- Markov Chain Monte Carlo Methods (MCMC)- conditioning on the past STA 447.
- Non-parametric density estimation- eg. kernels, polynomial smoothers STA 355
- Quantile Regression- to obtain conditional response quantiles
- Generalized Linear Mixed Model- eg. Binomial Logistic Mixed Model STA 490

Theory: 347, 452, 453.

All the best on your Exam!

- When: Wednesday, April 25 at 9am to 12noon
- Where: EX 300, 310, 320 (Exam Centre)
- What's covered: All topics - with emphasis on latter half
- Why: for academic evaluation!
- Who'll be there: Us (Classmates, TAs, Instructor, Invigilators)

OH : April 19-24

Thanks for being a great class!