

Mixed effect model and GEE in multiple observation study



이 지 성

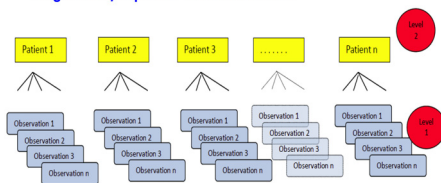
서울아산병원 의학통계학회

Ji Sung Lee, PhD

Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Introduction

- 집락 자료 (Clustered data)
 - 계층적(hierarchical) 또는 내포된(nested) 형태로 이루어진 자료
 - Example:
 - Clustered/multilevel studies
 - Longitudinal/Repeated Measures studies



즉, 관측치가 집락(각 환자) 내에 반복측정(repeated measures) 되어 있다.

Introduction

- 경시적 자료 (longitudinal data)
 - 특별히 같은 **subject**에서 시간에 따라 여러 번 (2번 이상) 반복 측정된 자료를 경시적 자료(longitudinal data)라고 한다.
 - 이 때 **subject**가 집락이 된다.
 - 집락 내 (within-cluster or within-subject)에 반복 측정된 자료들은 서로 **상관되어(correlated)** 있다.
 - 개체들 간(between cluster or between-subject)에는 서로 **독립적(independent)**이다.
 - 여기서 독립적이란 한 개체의 반복측정치들이 다른 개체의 측정치에 관한 정보를 담고 있지 않다는 의미이다.

경시적 연구 혹은 반복측정된 자료 분석시 고려해야 할 사항

- 통계분석 방법에서 **집락 혹은 개체 내 상관(within-cluster or within-subject correlation)**을 고려해 주어야 한다.
 - 지금까지 배운 통계분석법은 모든 측정치들이 **서로 독립**이라는 가정을 하였음.
 - 개체 내 상관을 고려하지 않은 분석을 한 경우
 - 관심 추정치에 대한 표준오차(standard error)가 일반적으로 과소추정(underestimate) 됨
 - 제 1 종의 오류 (Type I error) 증가
 - P-value 는 작아지고 신뢰구간은 지나치게 좁아짐.
 - 따라서, 부적절한 결과를 바탕으로 잘못된 해석을 내릴 수 있음.
- 결측치 발생
- Baseline 값의 처리
- Time-dependent 변수의 처리

집단 간 비교시 부적절한 분석방법

- 집단 별로 모든 값을 이용하여 단일 선형회귀분석을 하거나 집단 간 비교를 위해 일원 분산분석을 시행
 - 동일한 개체로부터 반복적인 측정값을 얻었다는 점이 고려되지 않았기 때문에 부적절한 분석이다. (개체 내 상관 고려되지 않았음).
- 각 시점 별로 two sample t-test 또는 분산분석을 시행하여 집단 간 평균 비교를 하는 것은 잘못된 분석법이다.
 - 개체 내 상관 무시
 - 다중 비교의 문제
 - 개체 내 변화의 정보를 잃음.

Profile Plots

▪ Example: Hypothetical Cholesterol Study

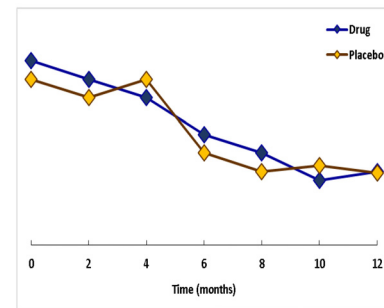
Group	Initial	2 mo.	4 mo.	6 mo.	8 mo.	10 mo.	12 mo.
Drug	피험자들은 각각의 방문시점에서 콜레스테롤 수치를 측정함						
Placebo							

– Questions of Interest:

- 1) 시간이 지남에 따라 피험자의 콜레스테롤 수치에 변화가 있는가?
• 즉, Time effect (within-subjects effect)가 있는가?
- 2) Treatment(or Group) effect (between-subjects effect)가 있는가?
- 3) Drug와 Placebo 모두 time effect가 동일한가?
• 이것이 primary interest임.

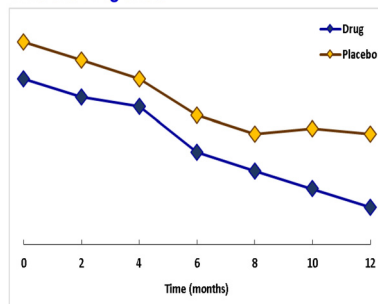
Profile Plots Illustrating the Questions of Interest

Time effect ONLY



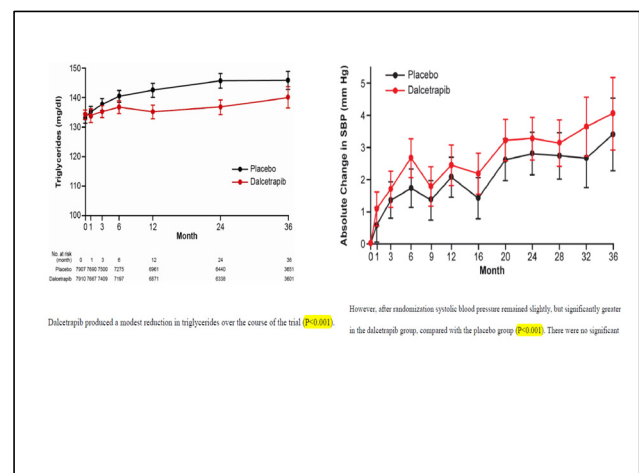
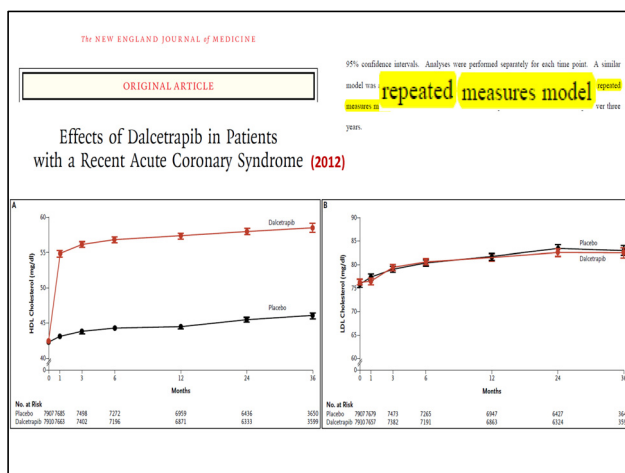
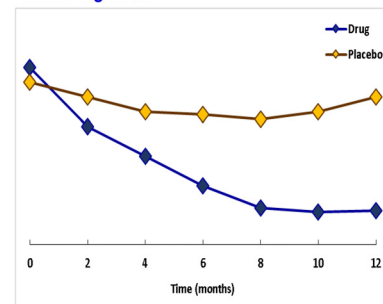
Profile Plots Illustrating the Questions of Interest

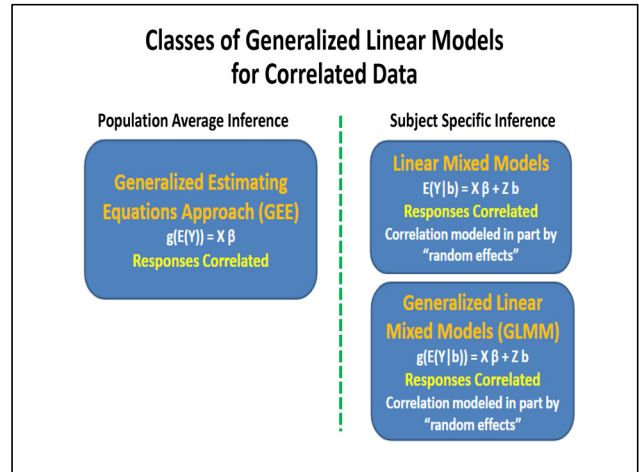
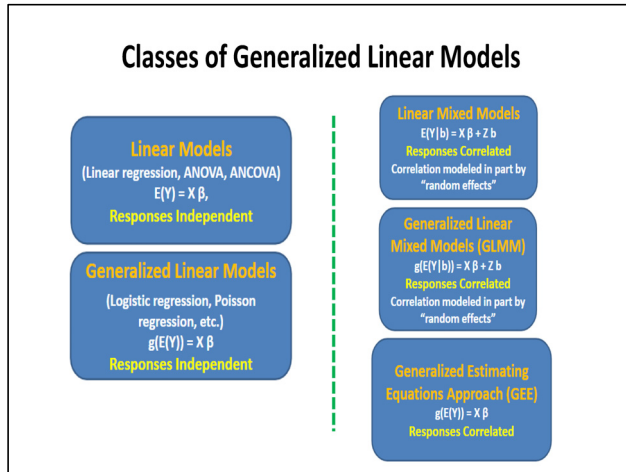
Time and Drug effect



Profile Plots Illustrating the Questions of Interest

Time*Drug effect





Linear Mixed Model (선형혼합모형)

Linear Mixed Model (LMM)

- 반복측정자료의 특성인 개체내 상관을 고려하여 분석하는 방법이 선형 혼합 모형 (linear mixed model: LMM)이다.
 - Random Effects model
 - Hierarchical linear effects model
 - Multilevel models
- LMM은 고정효과(fixed effect)와 랜덤효과(random effect)로 구성되어 있음.
 - 개인이 평균(fixed) 결과로부터 랜덤으로 편차가 있음을 가정
 - Fixed effect model
 - 모든 변동은 모형에 포함되는 독립변수로 설명가능하다고 생각될 때
 - Mixed effect model (or Random coefficient model)
 - 독립변수에 의해 설명되지 않은 변동이 있다고 판단될 때

Ex) 100명의 환자를 대상으로 피험자마다 혈압을 1주일 간격(fixed level)으로 3번 반복 측정할 경우, 3시점 간에 혈압이 다른지 보고자 한다면, **시점을 고정효과**로, **피험자를 랜덤효과**로 고려하여 분석할 수 있음.

✓ 선형회귀모형

$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, R)$$

$$Var(y) = R$$

where
 y = dependent variable
 ε = residual error
 X = design matrix for fixed effects
 β = fixed effects parameters
 R = residual variance matrix

✓ 선형혼합모형

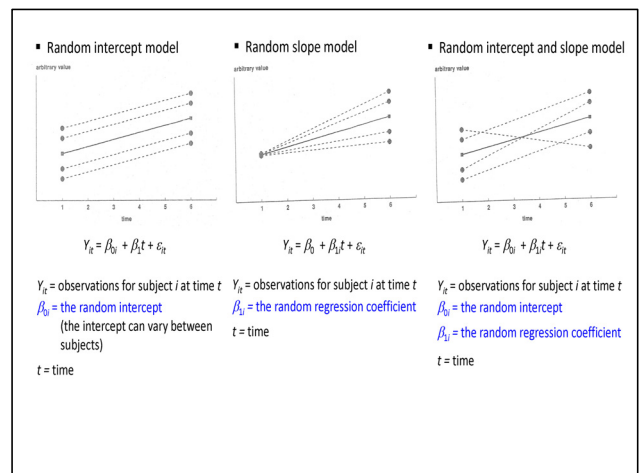
$$y = X\beta + Z\gamma + \varepsilon$$

$$\gamma \sim N(0, G)$$

$$\varepsilon \sim N(0, R)$$

$$V = Var(y) = ZGZ' + R$$

where
 y = dependent variable
 ε = residual error
 X = design matrix for fixed effects
 Z = design matrix for random effects
 β = fixed effects parameters
 γ = random effects parameters
 V = the covariance matrix



Covariance Structure Type

▪ Example of type of covariance structure

① AR(1)

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- Correlation between responses depends on the interval of time between responses

② Compound Symmetry

$$\begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix}$$

- Any responses within the same subject has the same correlation

③ Toeplitz

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

④ Unstructured

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix}$$

- All correlation coefficients free to take any value

LMM의 특징

- 측정간격과 측정회수가 같을 필요가 없음

- 모든 관측된 값을 이용할 수 있음.

- GEE보다 LMM이 선호되는 상황

- 군집의 개수가 작을 때
- 연구 초점이 개별 대상자에서 결과를 예측하는 경우
 - 담배를 피우고, 술을 마시는 66세 남성의 골밀도를 예측 → LMM
 - 연령, 흡연, 음주, 성별이 평균 골밀도에 미치는 영향확인 → GEE

Generalized Estimating Equation (GEE)
(일반화추정방정식)

- Linear Mixed Model은 반복측정된 결과변수가 정규분포를 따르는 **연속형 변수인 경우**에만 가능하다.

✓ Generalized linear mixed model (일반화 선형 혼합 모형, GLIMMIX)

- 연속형 자료 뿐만 아니라 이분형, 순서형, 계수형 자료에 대한 반복측정 자료분석을 하는 경우



▪ 일반화 추정 방정식(generalized estimating equations, GEE)

- 기본적으로 주변모형은 일반화 선형 모형 (generalized linear model)과 형태가 같다.
 - 이 주변모형을 적합하기 위해서는 추정방법이 필요한데 그 추정방법을 GEE라 부름.
- GEE에서는 개체 내 반복측정으로 인한 상관을 고려하기 위해 **working correlation**을 이용함.
- GEE의 장점은 가정된 working correlation이 틀릴지라도 표본이 크면 GLM의 회귀계수의 추정 및 검정이 유효하다.
 - 물론 working correlation이 맞으면 더 좋은 결과를 얻을 것이다.
- 상관관계의 극적인 변화를 기대하지 않는다면, **exchangeable working correlation structure**를 추천한다.
- GEE로 얻어진 로지스틱 회귀의 추정치의 해석은 일반화 로지스틱 회귀계수의 해석과 동일하다.

Examples of Generalized Linear Models

▪ Traditional Linear Model

- response variable: a continuous variable
- distribution: normal
- link function: identity $g(\mu) = \mu$

▪ Poisson Regression in Log Linear Model

- response variable: a count
- distribution: Poisson
- link function: $\log$ $g(\mu) = \log(\mu)$

▪ Logistic Regression

- response variable: a proportion
- distribution: binomial
- link function: logit

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

▪ Multinomial Models (ordinal model)

- distribution: multinomial
- link function: cumulative logit

Fitting Algorithm

1. Compute an initial estimate of β with an ordinary generalized linear model assuming independence.
2. Compute the working correlations R based on the standardized residuals, the current β , and the assumed structure of R .
3. Compute an estimate of the covariance: $V_i = \phi A_i^{-1} \hat{R}(\alpha) A_i^{-1}$
4. Update β :

$$\beta_{r+1} = \beta_r + \left[\sum_{i=1}^K \frac{\partial \mu_i}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \right]^{-1} \left[\sum_{i=1}^K \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (Y_i - \mu_i) \right]$$
5. Iterate steps 2-4 until convergence

※Working correlation matrix(R): 각 집락내에 있는 관찰값들 간의 상관구조에 대한 working structure를 가정한 상태로 분석을 진행

Working Correlation

For subject i

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & & \\ \vdots & & \ddots & \rho_{n-1} \\ \rho_{n1} & \rho_{n-1} & & 1 \end{bmatrix}$$

For repeated measures ρ_{lm} = correlation between times l and m
 For clustered data ρ_{lm} = correlation between measures l and m
 For all models considered here V_i is assumed to be same for all subjects

Types of working correlation matrix

Independent

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Exchangeable

$$R = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

Stationary M-dependent

$$R = \begin{bmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 \\ 0 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}$$

AR(1)

$$R = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

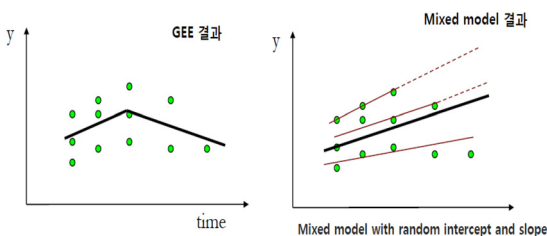
Unstructured

$$R = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{bmatrix}$$

GEE vs MM(Mixed effects model)

GEE	MM
Robust SE(β)와 Model-based SE(β) 모두 제공	Model-based SE(β)만 제공 Recent proposals for robust variance estimates under MM
Output: Coefficients(β) relating Y to X	Output: Coefficients(β) relating Y to X conditional on b Estimates of subject specific random effect Variance of subject specific random effects
β 는 population-averaged (or marginal) 추정치	β 는 subject-specific (or conditional) 추정치
Missing mechanism 가정 = MCAR	Missing mechanism 가정 = MAR
→ Missing이 MCAR가 아닌 경우는 GEE분석결과와는 biased 일 수 있음	

EX: 반복측정자료분석결과(GEE vs Mixed model)



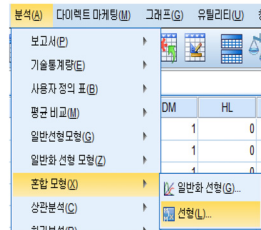
Data Format is Long (for MM and GEE)

SPSS: 데이터 → 구조변환 → 맵시 이용

	RD	FU	INTERVAL	FU INTERVAL	INTERVAL	INTERVAL	mmse0	mmse1	mmse2
1	1	0	832	1186	26	22	25		
2	2	0	361	1019	24	25	25		
3	3	0	349	1165	22	25	22		
4	4	0	398	679	17	20	24		
5	5	0	1105	1076	27	25	30		
6	6	0	410	790	25	25	25		
7	7	0	1218	2136	27	27	29		
8	8	0	177	644	19	20	21		
9	9	0	330	1579	28	29	29		
10	10	0	99	364	26	29	27		
11	11	0	1174	1807	25	28	26		
12	12	0	307	614	17	15	15		
13	13	0	227	568	27	27	28		
14	14	0	257	753	22	22	22		
15	15	0	452	672	9	10	6		
16	16	0	333	457	22	25	24		

Using SPSS

MM(Mixed effects model)



GEE

