

chapter 13

Linear Mixed - Effects Model

2018/00 84 고영희

다 이해해야 함!

파이팅 92



0. 기본 개념 정립.

: 고정 효과 모형 (Fixed Effect Model), 임의 효과 모형 (Random Effect Model)

요인 (factor) : 실험에서 결과에 영향을 미칠 것이라고 고려되는 독립 변수

수준 (level) : 실험에서 사용되는 요인의 값.

ex) 굴나무에서 가장 많은 꽂을 수확하기 해주는 토질의 종류가 무엇인지 알아보는 실험

이때, 굴의 수확량에 영향을 미치는 토질 → 요인,

토질의 종류인 모래흙, 일반흙, 진흙 → 요인의 수준

이 실험을 통해, 관심 요인인 토질의 수준 간 효과 차이를 검증하게 된다.

고정 효과 (Fixed Effect)

- 요인 (factor)의 수준을 실험차가 직접 지정한 경우, 실험자는 오직 이 수준들의 비교에만 관심이 있다.

- 실험된 요인의 수준에 대해서만 비교가 가능하여 통계 추론이 실험에 사용된 수준에 제한된다.

- 동일한 개체 (subject)에서 반복 측정된 자료에 적용하기에 부적절.

- 고정 요인이 포함된 모형을 고정 효과 모형 (Fixed Effects model)

ex1] 여학생과 남학생 시험성적 비교할 때, 성별 → 고정효과.

ex 2] 폐경한 카로아 종류 A, B, C에 따른 개선 정도를 비교.

약의 종류 → 고정 효과.

고정 효과 모형 (Fixed Effects Model)

$$y_{ij} = \mu + T_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad i=1, 2, \dots, t; j=1, 2, \dots, r$$

- y_{ij} : i 번째 수준에서 j 번째 관측값.

- μ : 모평균, ε_{ij} : 오차항.

- T_i : i 번째 수준의 효과.

이 모형에서 관심은 수준들마다 관측값에 차이가 있는지이다.

가설 $H_0: T_1 = T_2 = \dots = T_t = 0$ (처리 효과 차이가 X)

$H_1: \text{not } H_0$ (처리 효과 차이가 있다)

랜덤 효과함으로 대상의 과도한 변동 설명

임의 효과 (Random Effects)

- 요인의 수준이 임의 추출된 경우. ex) 집단 내 키 여러번 재기

- 각 개체에서 반복적으로 관측된 자료가 있는 연구에서 어떤 특정한 개체 종, 관측된 번복 측정 자료가 하나의 군집을 이루는 경우에 모형은 각 개체에 대한 랜덤 효과함을 포함하게 된다. 이 랜덤 효과함은 하나의 표본 군집이 모든 가능한 군집에서 추출되었다는 것 의미.

- 해당 요인의 수준을 넘어 표본단으로 해석 확대 가능. 요인의 모든 수준을 고려하는 경우, 그만큼 많은 모수를 모형에 포함하게 되며, 모형이 복잡해지는 문제가 발생하게 된다.

→ 고정효과가 아닌 임의 효과로 취급하면 이들의 분포를 설명하는 하나의 모수 (시그마)만을 모형에 포함하게 되어 훨씬 단순하게 모형설정 가능.

- 수준의 효과는 학회 변수로 간주되어, 분포를 가정한다

: 일반적으로, 평균 0, 분산 σ^2 인 정규분포 가설. → 분산의 추정값: 군집 간의 변동성

- 임의 요인이 포함된 모형을 임의 효과 모형 (Random Effects Model)

- 고정효과+임의 효과 모두 존재 → 혼합효과 모형 (Mixed effects Model)

임의 효과 모형 (Random Effects Model)

$$y_{ij} = \mu + T_i + \varepsilon_{ij}, \quad T_i \sim N(0, \sigma^2_T), \quad \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

- ε_{ij} : i 번째 수준의 j 번째 관측값

- μ : 모평균, ε_{ij} : 오차항

- T_i : i 번째 수준의 효과

→ 고정효과 모형에서와 다르게 T_i 는 상수 X. 학회 변수 0 → 오차와 indep.

Why? 실험 전에 어떤 수준이 선택될지 알 수 X → 학회 변수 !!

가설 $H_0: \sigma^2_T = 0$ (어느 수준에서든 효과 동일 = 처리효과 값 = 분산은 0)

$H_1: \sigma^2_T > 0$ (수준마다 효과 다른 = 처리효과 여러 값 = 분산 0보다 큼)

I. Introduction

Fixed Effect model

상관 데이터에서 fixed effects를 가진 모델

→ population averaged model

why? mean - structure parameters는 전체 모집단에서 종속 변수의 평균값의 공변량 (covariates) 으로 해석될 수 있다.

관측값들 사이의 관계는 factors 그룹의 같은 level을 공유하는 관측값들의 그룹핑 결과.

그룹화된 데이터 → longitudinal data (종단 자료)

Single level of grouping. (같은 개인의 level에서 그룹화)

- 계층 → ARMD data: 각 환자의 다중 시력 측정값
- 임상 실험 내에서 그룹화된 환자 데이터

multilevel hierarchy

- 학생 점수: 학생으로 grouping, 학생은 학급으로 grouping, 학교, 지역으로 grouping
↳ 점수의 전체 변동성을 학생내, 학생간, 학급간, 학교 간의 변동성의 결과

grouping 때문에 관측 데이터의 복잡한 연관구조 예상 할 수 있다.

(ex) 개별 학생의 점수 상관관계가 같은 학급(=학교) 내 다른 학생의 점수 상관관계 기대

LMM (Linear Mixed-Effects Model)

: 연속적이고 계층적인 데이터 분석 시 사용하는 모델 종류.

데이터 셋에 포함된 관측치들의 상관관계 고려

↳ 종속 변수의 전체 변동을 데이터 계층의 다른 레벨에 해당하는 요소로 효과적 분할

→ subject-specific model

Why? subject-specific 계수 포함

* population average model vs subject-specific model

↳ fixed model

↳ random effect model

즉 random effect

인 시점이 있음

가변수로 써드리는 데 그걸 high dim

개념 추가

여러 학교에서 학생들 성적 조사한다고 할 때,

“학교”가 공변량

* 공변량 (covariates) 연속형 변수를 통제 변수로 이용.

: 여러 변수들이 공통적으로 함께 공유하고 있는 변량. 종속 변수에 대해 독립변수와

기타 값들에 종속 변수에 영향을 미치는 것을 통제함으로써, 독립 변수

자체의 순수한 영향을 측정하는데 목적 만약, 공변량이 종속 변수에

영향을 주지 않는다면, 이것은 통계적으로 통제할 필요가 없으며,

이 경우에는 공변량을 분석 모형에서 제거한다 해도 결과 변함 X.

그러나, 종속 변수에 영향을 준다면, 실험할 경우 공변량은 통계적으로 통제.

통계적으로 통제를 하는 공분산 분석의 경우, 공변량 통제 → 종속 변수에 주는 영향적 제거.

⇒ 독립 변수가 종속 변수에 주는 영향을 좀 더 명확하게 규명하고 검출

ex] 커피 광고 효과 얻기 위한 3가지 광고 (맛, 향, 불위기)

공변량: 카페에 사는 커피의 간수

두 범인 X (광고 여부 사는 커피 간수), Y (평균에 사는 커피 간수)의 개별 값들이 평균치 (M_x, M_y)로부터 떨어져 있는 개인 평균 (x, y)를 구하고

X 와 Y 의 서로 대응하는 벙주의 X 와 Y 값을 곱하여 일어나는 $X\cdot Y$ 를 교적(cross product).

이 교적을 모두 합한 것을 교적합(sum of cross product)

x^*, y^* 을 X 와 Y 의 합성평균(sum of averages)

$X\cdot Y$ 를 사례수 (N)으로 나눈 값이 공변량.

⇒ Y (평균 커피 간)의 영향을 제거하고 각 실험 단위의 X 의 평균 구하기.

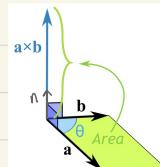
→ 정평균치 (corrected mean)

→ 같은 범인에 종속 변수에 미치는 영향이 제거되고, 솔직히 독립변인이 종속 변수에 대하여 잊어버리는 집단 간 차이.

* longitudinal data (종단 자료)

: 개인에 대해 시간이 지남에 따라 수집된 다중 값을 갖는 data

* (④) 교적 (cross product)



$$a \times b = |a| |b| \sin(\theta) n$$

$|a|$: vec a 의 길이, $|b|$: vec b 의 길이

θ : a 와 b 의 사이각,

n : a 와 b 모두에 대해 각각의 단위 vec

④ 첫걸리는 개념 정리

#1. covariate (공변량) VS Fixed effect (고정효과)

A factor is categorical variable → 고정효과는 범주형 변수. ex) 성별, ...

A covariate is a continuous variable → 공변량은 연속형 변수.

내가 이해했을 때,

점수변수		독립변수	
성적	학고	교사	
36	A	김	
78	B		
90	C		
60	D	최	
75			

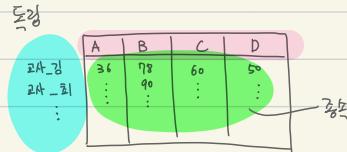
총속		학교				총점	
성적	학고_A	학고_B	학고_C	학고_김			
36	1	0	0	1			
78	0	1	0	0			
90	0	0	1	0			
60	0	0	0	0			
75	:	:	:	0			

→ 이건 고정효과

→ 가변수화 → high dimension

학교-D 변수는 나머지 학교 변수 모두 0일 때로 추론 가능
정보 충분 → 제거.

보통 변수는 데이터에 영로서 들어온데,
공변량은 행으로 들어온다고 생각하면 될듯!



→ 공변량은 “학교”: 각 집단의 평차 Vec 고집 후
행의 개수 n으로 나눔

#2. covariate (공변량) VS independent variable (독립 변수)

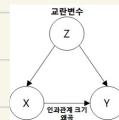
covariate (공변량)은 독립변수가 될 수 있다. 만약 독립변수로 원하지 않으면, confounding variable (교란 변수)가 된다.

covariate를 모델에 추가하는 것은 경학도 향상 (약간 노이즈 중에는 느낌...?)

④ 교란 변수 (confounding variable)

X와 Y 두 변수 모두 영향을 미치나, X와 Y 사이에도 인과관계가 존재할 경우 의미.

교란 변수의 존재는 X와 Y 사이의 인과관계 크기를 실제보다 크거나 작은 것으로 보아야 한다.



II. The classical Linear Mixed-Effects Model

1. Specification at a Level of a Grouping Factor.

grouping의 single level의 계층데이터에 대한 classical LMM

$$y_{\bar{i}} = X_{\bar{i}} \beta + b_{\bar{i}} + \varepsilon_{\bar{i}}$$

[변수 설명]

$\bar{i} = 1, \dots, N \leftarrow$ group factor에 대한 level (i번쨰의 그룹)
 $\bar{j} = 1, \dots, n_{\bar{i}} \leftarrow$ 그룹에 대한 $n_{\bar{i}}$ 개 관측 값.

- $y_{\bar{i}} = \begin{pmatrix} y_{\bar{i}1} \\ y_{\bar{i}2} \\ \vdots \\ y_{\bar{i}n_{\bar{i}}} \end{pmatrix}$, 연속형 종속변수

- $X_{\bar{i}} = \begin{pmatrix} x_{\bar{i}1}^{(1)} & x_{\bar{i}1}^{(2)} & \cdots & x_{\bar{i}1}^{(p)} \\ x_{\bar{i}2}^{(1)} & x_{\bar{i}2}^{(2)} & \cdots & x_{\bar{i}2}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\bar{i}n_{\bar{i}}}^{(1)} & x_{\bar{i}n_{\bar{i}}}^{(2)} & \cdots & x_{\bar{i}n_{\bar{i}}}^{(p)} \end{pmatrix} = (x_{\bar{i}1}^{(1)} \ x_{\bar{i}1}^{(2)} \ \cdots \ x_{\bar{i}1}^{(p)})$
독립변수를 용어로 험열
i번쨰 그룹에 대한 설계행렬 ($p < n_{\bar{i}}$)

- $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, 알려지지 않은 회귀 파라미터

- $\varepsilon_{\bar{i}} = \begin{pmatrix} \varepsilon_{\bar{i}1} \\ \varepsilon_{\bar{i}2} \\ \vdots \\ \varepsilon_{\bar{i}n_{\bar{i}}} \end{pmatrix}$, \bar{i} 그룹내 residual error의 vector

- $Z_{\bar{i}} = \begin{pmatrix} z_{\bar{i}1}^{(1)} & z_{\bar{i}1}^{(2)} & \cdots & z_{\bar{i}1}^{(q)} \\ z_{\bar{i}2}^{(1)} & z_{\bar{i}2}^{(2)} & \cdots & z_{\bar{i}2}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{\bar{i}n_{\bar{i}}}^{(1)} & z_{\bar{i}n_{\bar{i}}}^{(2)} & \cdots & z_{\bar{i}n_{\bar{i}}}^{(q)} \end{pmatrix} = (z_{\bar{i}1}^{(1)} \ z_{\bar{i}1}^{(2)} \ \cdots \ z_{\bar{i}1}^{(q)})$

q: 공변량 (covariates) 행렬. 알려진 값.

- $b_{\bar{i}} = \begin{pmatrix} b_{\bar{i}1} \\ \vdots \\ b_{\bar{i}q} \end{pmatrix}$, random effects vec, 관측 X.

여기서 q는 random effects의 개수

$$\Rightarrow b_{\bar{i}} \sim N_q(0, D), \quad \varepsilon_{\bar{i}} \sim N_n(0, R_{\bar{i}}), \quad b_{\bar{i}} \perp \varepsilon_{\bar{i}}$$

[변수 설명 _ continue]

① 같은 그룹 내 residual error $\varepsilon_{\bar{i}}$ 는 random effect $b_{\bar{i}}$ 와 독립

→ classical LMM 와 extended LMM 구별할 수 o

② 다른 그룹에 대해서 random effect vec 와 residual error 독립

→ $\bar{i} \neq \bar{i}'$ 일 때, $b_{\bar{i}}$ 와 $\varepsilon_{\bar{i}'}$ 독립

$$D = \sigma^2 D, \quad R_{\bar{i}} = \sigma^2 R_{\bar{i}}$$

- σ^2 : unknown scalar parameter

- $D, R_{\bar{i}}$: Positive - definite (별도로 영시 X 일 때)

→ D 와 $R_{\bar{i}}$ 는 일반적으로 unique하지 않다. 여러개 나올 수 있음!

식별 위해, 분산 함수 (variance function)과 상관행렬 (correlation matrix)의 parameter set의 한정에서 $R_{\bar{i}}$ 구조 구체화. → $R_{\bar{i}}$ 식별 가능

설계행렬 $X_{\bar{i}}$ 구성하는데 사용되는 fixed effects parameter β 외에도

모델 ($y_{\bar{i}} = X_{\bar{i}}\beta + b_{\bar{i}} + \varepsilon_{\bar{i}}$)은 2개의 random components (회) 포함

→ 그룹내 residual error $\varepsilon_{\bar{i}}$ 와 설계행렬에 포함되는 공변량의 random effect $b_{\bar{i}}$

알려진 변수의 fixed 와 random 효과의 존재는 모델의 이름 알 수 있게 한다.

대부분, $b_{\bar{i}}$ 에 포함된 random effects는 β 에 있는 fixed effects에 상응함.

→ $Z_{\bar{i}}$ 행렬은 종종 $X_{\bar{i}}$ 행렬의 적절한 열의 부분집합을 택함으로써 만들어진다.

상용하는 fixed effects와 random effects는 "coupled"라고 불림.

이런 모델은 two-stage (level) model로 언급.

But, single-level LMM 이라 부르는 사람 있.

↳ Why? grouping의 single level에 의해 정의되는 데이터 계층 적용.

개념 추가

classical LMM은 multilevel 그룹화 데이터에도 적용 가능.

예] two-levels 그룹화된 데이터의 모델.

first-level : 관측값들 N개의 그룹화. ($i=1, \dots, N$)

2nd-level (sub)group : n_j 관측치 포함. ($j=1, \dots, n_i$)

$$y_{ij} = X_{ij}\beta + z_{1,ij}b_i + z_{2,ij}b_i + \varepsilon_{ij} \quad \text{random variable}$$

$$y_{ij} = X_{ij}\beta + \varepsilon_{ij} \rightarrow E(y_{ij}) = X_{ij}\beta, \text{Var}(y_{ij}) = \text{Var}(\varepsilon_{ij}) \rightarrow \text{Scalar}$$

$\oplus b_i \sim N_p(0, D)$, $b_i \sim N_p(0, I)$, $\varepsilon_{ij} \sim N_{n_i}(0, R_{ij})$

vector. $\rightarrow g \times g$ matrix.

$$\text{Var}(y_{ij}) = \text{Var}(X_{ij}\beta + \varepsilon_{ij}) = R_{ij} \leftarrow \text{fixed effect term } X$$

$$\downarrow \text{y은 } n_i \text{ 개별 범주 } \times$$

- random vector인 $b_i, b_{ij}, \varepsilon_{ij}$ 각각 서로 독립.

- b_i : 1st level 그룹화 관련된 random effects.

- b_{ij} : 1st level의 random effects와 동일하고 2nd level 그룹화 관련된 random effects.

- 설계 행렬 (design matrix) : $Z_{1,ij}, Z_{2,ij}$ 같을 수 있음

(무조건 같아야 하는거 X)

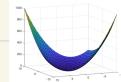
이 모델은 two-level LMM으로 불린다.

* 양의 정부호 행렬 (positive definite matrix)

$$1) \text{정의} : f(x) = x^T A x > 0 \quad \forall x \neq 0$$

이 함수는 하나의 실수로 나타나며, x 의 각 component들에 대한 2차 합수 형태로

(대칭 행렬에 대해서만 적용)



2) multivariate 함수로서의 이해.

$$2 \times 2 \text{ 대칭 행렬} \rightarrow f(x,y) = (x \ y) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2bxy + cy^2$$

이 함수는 원점을 무조건 지나가게 되는데, 위(아래)로 불록한 그릇 형태이거나

암간 형태의 함수. x, y 평면상에 contour 그려보면 타원형 / 포물선 형태.

이 중, 위로 불록한 그릇 형태로 모든 (x,y) 에 대해 양수가 될 때,

positive definite matrix $\rightarrow f(x,y)$ 을 두 component 계산으로 나타낼 수 있다.

어떤 힘의 이계도 함수 행렬 표현

vector calculus의 Hessian 이용해 나타낼 수 있다

: 원점의 second derivative들이 양수 (=Hessian matrix의 determinant 양수)임 때

원점이 극소값이 되면서 위로 불록한 함수 의미.

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}, \quad \det(H(f)|_{x=0, y=0}) > 0$$

3x3 행렬

: $f(x,y,z)$ 의 contour surface + 3차원상 타원. 원점에서 3×3 Hessian 행렬 암시

이 matrix의 eigen vector들이 contour 타원들의 axis들을 이룬다.

\rightarrow 이 함수를 가장 빨리 / 천천히 증가시키는 방향 나타냄
대칭 행렬 아래에 eigen vector들은 서로 orthogonal.

3) 필요 충분 조건

① 모든 eigen value는 양의 실수

② 모든 sub-determinant는 양수: $|1 \times 1$ 부분 행렬 ... $|n \times n$ 부분 행렬의 det 양수.

③ elimination 한 후 모든 pivot 들이 양수

: 일반적으로 Gaussian elimination하고 남은 pivot 들의 부호들, eigen value 부호 일치.

4) 성질

① positive definite한 행렬의 역행렬도 positive definite

② $A+B$ 가 positive definite하면 $A+B$ 도 positive definite

$$\begin{aligned} \mathbb{E}(y_{ij}|b_i) &= \sum_j y_{ij} \cdot p_{y_{ij}|b_i} (y_{ij}|b_i) \\ &\quad \text{marginal은 random effect term 생각!} \\ &\quad \text{@ covariance 계산 때} \\ &\rightarrow \mathbb{E}(X_{ij}\beta + z_{1,ij}b_i + z_{2,ij}b_i + \varepsilon_{ij}|b_i) \\ &\quad \text{정리} \rightarrow \mathbb{E}(y_{ij}|b_i) = 0. \end{aligned}$$

$$\begin{aligned} &= \mathbb{E}(X_{ij}\beta|b_i) + \mathbb{E}(z_{1,ij}b_i|b_i) + \mathbb{E}(\varepsilon_{ij}|b_i) \\ &= X_{ij}\beta + z_{1,ij} \cdot \mathbb{E}(b_i|b_i) + 0. \quad \text{Condition} \\ &= X_{ij}\beta + z_{1,ij} \cdot b_i \quad \text{random variable 알고 있음} \end{aligned}$$

2. Specification for All Data.

single-level LMM을 모든 데이터에서 설명.

→ multi-level LMM로 일반화 가능!

- $y = (y'_1, y'_2, \dots, y'_N)'$, 종속 변수의 관측값 $n = \sum_{i=1}^N n_i$ 포함하는 vec

- $b = (b'_1, b'_2, \dots, b'_N)$, 모든 N_g (random effects) 포함하는 vec

- $\varepsilon = (\varepsilon'_1, \varepsilon'_2, \dots, \varepsilon'_N)'$ n 의 residual error 포함하는 vec

- $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$, 일반적으로 dimension은 $n \times p$

- $Z = \begin{bmatrix} z & 0 & \cdots & 0 \\ 0 & z & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & z_N \end{bmatrix}$, 0은 영행렬. 일반적인 dim: $n \times N_g$

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

$$\oplus b \sim N_{N_g}(0, \sigma^2 D), \varepsilon \sim N_n(0, \sigma^2 R)$$

$$D = I_N \otimes D = \begin{bmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & D \end{bmatrix}, \quad R = \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & R_N \end{bmatrix}$$

↓ : kronecker product

- block-diagonal form (블록 대각 행렬)

: Z, D, R

- Single-level LMM이 데이터의 특정 계층과 random effect를 가정한다는 사실에서 거친함 $b_i \sim N_{N_g}(0, D)$
 $b_i \perp \varepsilon_i \sim N_n(0, R_i)$ 처럼 명백하게 보여줌.
- 특히, 모델은 특정한 factor의 level에 의해 정의된 다른 그룹에서의 random effects는 독립이라 가정.
- 계층화는 factor들을 그룹핑 함으로써 생님되고 하나의 factor라는 다른 factor 내 nested(내포) 된다.

개념 추가

* 크로네커 곱 (kronecker product)

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{bmatrix}_{n \times p}$$

행렬 A 와 행렬 B 의 크로네커 곱셈은 $A \otimes B$ 로 표시.

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}$$

$A \otimes B$ 를 계산한 후 행렬의 크기는 $m \times np$ 행렬

크로네커 곱셈을 이용하면 행렬의 크기는 크되, 일정한 규칙이 존재하는

행렬을 쉽게 만들 수 있다

* Block-Diagonal Matrix (블록 대각 행렬)

: 대각 선을 따라 암역행렬, 즉 더 작은 행렬을 배치해 만든 행렬

* crossed effects

: 10명의 환자를 5명씩 나눠 A약과 B약(treatment)을 비교한다고 하자.

이 때, treatment는 환자 - 치아 각 level에 영향 줌.

이 treatment를 crossed factor를 가진다고 한다.

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

⇒ non-block-diagonal matrix로 Z, D, R 사용해

random effect 뿐 정령화 가능

이 때, crossed random effects를 가진 모델.

ch. 15.

III. The Extended Linear Mixed-Effects Model

$$b_i \sim N_g(0, D) \quad \varepsilon_i \sim N_{n_i}(0, R_i) \quad b_i \perp \varepsilon_i \quad \text{에서}$$

residual error $[\varepsilon_i]$ 와 random effect $[b_i]$ 독립 가정 → 제한적!

예) mean-variance model에서 처럼, residual error의 분산은

subject-specific한 평균값에 의존한다는 가정.

ε_i 와 b_i 독립 가정 완화 \Rightarrow extended LMM.

$$y_i = X_i \beta + Z_i b_i + \varepsilon_i$$

$$Z_i = \begin{bmatrix} z_i^{(1)} & z_i^{(2)} & \cdots & z_i^{(g)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{i1}^{(1)} & z_{i1}^{(2)} & \cdots & z_{i1}^{(g)} \end{bmatrix} = [z_i^{(1)} \ z_i^{(2)} \ \cdots \ z_i^{(g)}]$$

$$b_i = \begin{bmatrix} b_{i1} \\ \vdots \\ b_{ig} \end{bmatrix}$$

$$b_i \sim N_g(0, D), \quad \varepsilon_i | b_i \sim N_{n_i}(0, R_i)$$

$$\hookrightarrow D = \delta^2 D, \quad R_i = \delta^2 R_i$$

\Rightarrow "hierarchical specification!"

만약, ε_i 가 random effects와 독립임 가정 하면, classical LMM.

따라서 extended LMM은 더 일반적인 모델링이다.

extended LMM에서, two-level의 hierarchical specification

$$y_{ij} = X_{ij} \beta + Z_{1,ij} b_i + Z_{2,ij} b_{ij} + \varepsilon_{ij}$$

$$\oplus b_i \sim N_{g_1}(0, D_1), \quad b_{ij} | b_i \sim N_{g_2}(0, D_2)$$

$$\varepsilon_{ij} | b_i, b_{ij} \sim N_{n_{ij}}(0, R_{ij})$$

IV. Distributions Defined by the y and b Random Variables.

classical LMM과 extended LMM은 2개의 연속형 학률 변수 $\Rightarrow b, y$
이들은 2개의 확률 밀도 함수 (pdf)로 설명되고, LMM을 정의하는데 중요.

b: 관측 x , random effects 의 unconditional distribution

$$b \sim N_{nq}(0, \delta^2 D)$$

y: (random) 종속 변수의 조건부 분포, random effects는 알려진 값 가정

1. Unconditional Distribution of Random Effects

$b_i \sim N_8(0, D)$
 $f_b(b_i)$: random effects (b_i)의 unconditional distribution

평균 0과 분산-공분산 행렬 (D)를 갖는 다변량 정규분포

$$\Rightarrow D(\delta^2, \theta_D) = \delta^2 D(\theta_D)$$

[변수 설명]

- θ_D : 파라미터 vec, δ^2 에 의해 스케일 된 b_i 원소의 분산과 공분산 나타냄

- D 행렬 : random effects b_i 의 분산-공분산 행렬 정의에 사용됨

θ_D 파라미터 벡터를 사용하여 매개 변수화 된다.

많은 경우, b_i 벡터의 어떤 2개의 원소는 상관관계 있을 수 있고

D 행렬의 제한 X (단, 양의 정부른 행렬과 Symmetric 은 제한 有)

D가 positive-definite 일 때

D는 q 분산에 해당되는 $\frac{q(q+1)}{2}$ 개의 별개의 원소와

b_i 에 있는 random effects의 $\frac{q(q-1)}{2}$ 개의 공분산을 갖는 일반적인

양의 정부른 행렬 구조. $\rightarrow \theta_D$ 는 $\frac{q(q+1)}{2}$ 개의 별개 파라미터 포함.

증명이 작을지라도, 모든 파라미터 초성은 이어줄 것 (예: 평균의 크기 제한적)

이 때, 행렬 D의 단순화된 구조를 선택할 수 있다.

예시] b_i 벡터의 모든 원소들이 독립성을 가정하는 것은, diagonal 형태로
초성하는 것과 같다. $\rightarrow \theta_D$ 는 오직 q개의 파라미터 포함.

④ 가정에 대한 타당성은 데이터에 따라 달라짐.

2. Conditional Distribution of y Given the Random Effects

classical LMM은 조건부 분포 $f_{y|b}(y_i|b_i)$ 를 따른다:

$y_i|b_i$ 는 다변량 정규분포.

$$E(y_i|b_i) = M_i = X_i \beta + Z_i b_i, \quad \text{Var}(y_i|b_i) = \sigma^2 R_i$$

$$M_i = (M_{i1}, \dots, M_{in_i})'$$

$$E(y_{ij}|b_i) = M_{ij} = X'_{ij} \beta + Z'_{ij} b_i$$

$$\rightarrow X_{ij} = (x_{ij}^{(1)}, \dots, x_{ij}^{(r)})', \quad Z_{ij} = (z_{ij}^{(1)}, \dots, z_{ij}^{(q)})'$$

: i 번째 그룹에서 j 번째 관측값들에 대한 예측값 X, Z 포함

random effects b_i 의 미지수 값에 대해 조건부인 증속 변수 b_i 의 평균값은

random effects b_i 와 fixed effects β 에 각각 상응하는

group 별 설계행렬 X_i 와 Z_i 내에 열로서 포함된 X 공변량과 Z 공변량 벡터의

선형 결합으로 정의.

y_i 의 조건부 분산 - 공분산 행렬은 residual error ε_i 의 분산 - 공분산 행렬과 같다.

대부분 일반적인 형태에서, LMM은 식별할 수 없다.

→ why? $D = \delta^2 D, R_i = \delta^2 R_i$ 식이 유일 X. 미지수 모수 多

D 행렬과 유사하게, θ_R 와 구불되는 계획된 모수 θ_R 의 합수로써, R_i 의 모수 고려해 LMM 식별.

R_i 행렬에 대해, $R_i(M_{ij}, \theta_R; V_{ij}) = \Lambda_i(M_{ij}, \delta; V_{ij}) C_i(s) \Lambda_i(M_{ij}, \delta; V_{ij})$ 이용해

행렬 분해하고, 이를 variance function 과 correlation structure 과 결합 가능.

R_i 는 variance function 과 correlation function의 파라미터로

매개 변수화 된다.

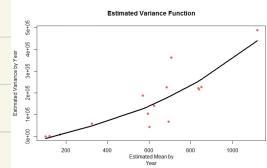
⇒ 모델의 파라미터 수 줄어든다. $D = \delta^2 D, R_i = \delta^2 R_i$ 식별화 가능해짐

* 분산 함수 (Variance function)

: 통계에서 분산 함수는 그 평균의 함수로 간접 수량의 분산을 나타내는 부드러운 함수이다. 통계 모델링의 많은 설정에서 큰 역할을 한다. 일반적인 선형모델

프레임워크의 주된 성분이며, 비모수 회귀분석, 반 모수 회귀 분석 및 가능 데이터 분석에서 사용되는 둘이다. 모수 모델링에서 분산함수는 모수 형식을 취하여, 분산과 간접 수량의 평균 사이의 관계를 명시적으로 설명한다.

비모수 설정에서는 분산 함수를 평활 함수로 가정한다.



* Mean dependence vs Mean independence

: 확률론에서, 임의 변수 Y 는 조건부 평균 $E(Y|X=x)$ 가 모든 x 에 대해

평균 $E(Y)$ 와 같을 경우에만 Y 는 간접 변수 X 와 평균독립적 (Mean Indep.)

Y 는 확률이 '0'이 아닌 모든 x 에 대해 $E(Y|X=x)$ 가 일정하지 않으면,

Y 는 X 에 대해 평균 종속적 (Mean dependent) 라 한다.

2. conditional distribution of y Given the Random Effects_계속

$\langle \delta, M \rangle$. $\langle M \rangle$ group으로부터의 variance function을 사용하기 위해서

행렬 D 와 R_T 의 구조를 선택하는 것 필요X.

hierarchical specification을 사용하고, $\epsilon_i | b_i$ 의 조건부 분포의 분위적용.

D 와 R_T 는 냄비벡터의 marginal 분산-공분산 행렬 형태의 결과를 갖는다.

$$b_i \sim N_p(0, D) \quad \text{Var}(\epsilon_{ij} | b_i) = \delta^2 \lambda^2(M_{ij}, \delta; V_{ij})$$

$\hookrightarrow V_{ij}$ 가 M_{ij} , δ 에 종속된다.

$$\text{단. } M_{ij} = E(y_{ij} | b_i) = X'_{ij} \beta + Z'_{ij} b_i$$

correlation structure와 variance function 합치면

$$\text{Var}(\epsilon_i | b_i) = \delta^2 R_T(M_i, \theta_R; V_i)$$

$$-\theta_R \equiv (\delta, \gamma)$$

δ : 분산 함수 $\lambda(\cdot)$ 에 의한 분산 파라미터 vec.

γ : 행렬 R_T 에 대해 선택된 correlation structure와 관련있는 파라미터 vec.

$$-V_i \equiv (V_{i1}, \dots, V_{in_i})'$$

1번째 그룹의 관측값에서의 분산-공분산 vec.

$$\text{Var}(\epsilon_{ij} | b_i) = \delta^2 \lambda(M_{ij}, \delta; V_{ij}) \text{ 와 } \text{Var}(\epsilon_i | b_i) = \delta^2 R_T(M_i, \theta_R; V_i) \text{는}$$

$\langle \delta, M \rangle$ 와 $\langle M \rangle$ 그룹에서 평균 종속 분산 함수를 갖는 모델에 대해,

ch. 13. 8

ϵ_i 는 $b_i \sim M_i$ 에 종속이다.

$\langle \delta, M \rangle$ 와 $\langle M \rangle$ 그룹들에서의 분산함수를 사용하여 정의되는 Extended LMMs은

이론적, 계산적으로 어려움 \rightarrow classical LMMs에 대해서 집중!

classical LMMs에서, R_T 는 평균 독립 분산 함수의 구체화됨

$\langle \delta \rangle$ -그룹의 variance function처럼 mean-independent 함수일 경우,

$$\text{Var}(\epsilon_i | b_i) = \delta^2 R_T(M_i, \theta_R; V_i) \text{식을}$$

$$\hookrightarrow \text{Var}(\epsilon_i | b_i) = \text{Var}(\epsilon_i) = \delta^2 R_T(\theta_R; V_i) \text{로 바꿀 수 있다.}$$

\rightarrow 잔차 ϵ_i 는 random effects b_i 와 독립이라는 가정과 일치.

평균 독립 분산 함수를 갖는 hierarchical model specification은

$R_T = \delta^2 R_T(\theta_R; V_i)$ 인 classical LMMs 유도.

3. Additional Distributions Defined by y and b

1) Joint Distribution of y and b .

: classical LMMs 의 y 와 b 의 결합 분포 $f_{y,b}(y_i, b_i)$ 는
random effects b 의 unconditional distribution과

y 의 조건부 분포의 곱이다.

$$f_{y,b}(y_i, b_i) = f_{y|b}(y_i | b_i) f_b(b_i)$$

→ $f_b(b)$ 와 $f_{y|b}(y|b)$ $f_b(b)$ 는 다변량 정규분포.

즉, 결합 분포 $f_{y,b}(y, b)$ 또한 정규분포.

(주변 확률 분포)

2) Marginal Distribution of y .

y_i 의 marginal distribution $f_y(y_i)$

→ y_i 와 b_i 의 결합 분포에서 random effects b_i 적분

$$f_y(y_i) = \int f_{y,b}(y_i, b_i) db_i = \int f_{y|b}(y_i | b_i) f_b(b_i) db_i$$

- $f_{y,b}$: y_i 와 b_i 의 결합 분포의 밀도 함수

- $f_{y|b}$: b_i 가 주어졌을 때 y_i 의 조건부 분포

- f_b : b_i 의 unconditional distribution의 밀도함수.

→ $f_{y,b}$ 와 f_b 가 다변량 정규분포의 밀도함수를 주어지면,

y 의 주변 확률 분포 또한 다변량 정규분포.

$$y_i \sim N_{n_i}(X_i\beta, \sigma^2 Z_i D Z_i' + \sigma^2 R_i)$$

3) Posterior Distribution of b Given y is known.

◀ 새로 안 내용

이는 x_i 의 sub-column 일 수로 있고

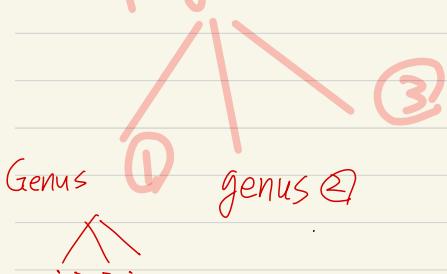
아닐 수도 O

“일반화 분형 모형”
GLM?
GML?
무지?

	OTU1	OTU2	
Sub 1	Z_1	0.75	0.01	0.03	...
Z_2					

proportion

phy ~ 예측



phy ~



$$\text{logit}(x) = \log \frac{P}{1-P}$$

↳ 학률값 반영

$$\text{logit}(x) = x_i \beta_i + \gamma_i b_i + \varepsilon_i$$

Random effect 이슈

① 가변수로 하면 고차원.

(X_1)

Category : 텁, 웃, 음

② 실험전에는 어떤 수준이 될지 모르기 때문에 확률 변수로 고려

$X_1 \dots X_n$.