11 강

데이터분석방법론2

선형혼합모형 소개

대전대학교 빅데이터인공지능학과 강위창 교수

학습목차

학습개요 및 목표

이번 강의에서는 상관된 연속형 반응변수에 대한 대표적 분석모형인 선형혼합모형(LMM: Linear Mixed Models)에 대해 소개합니다.

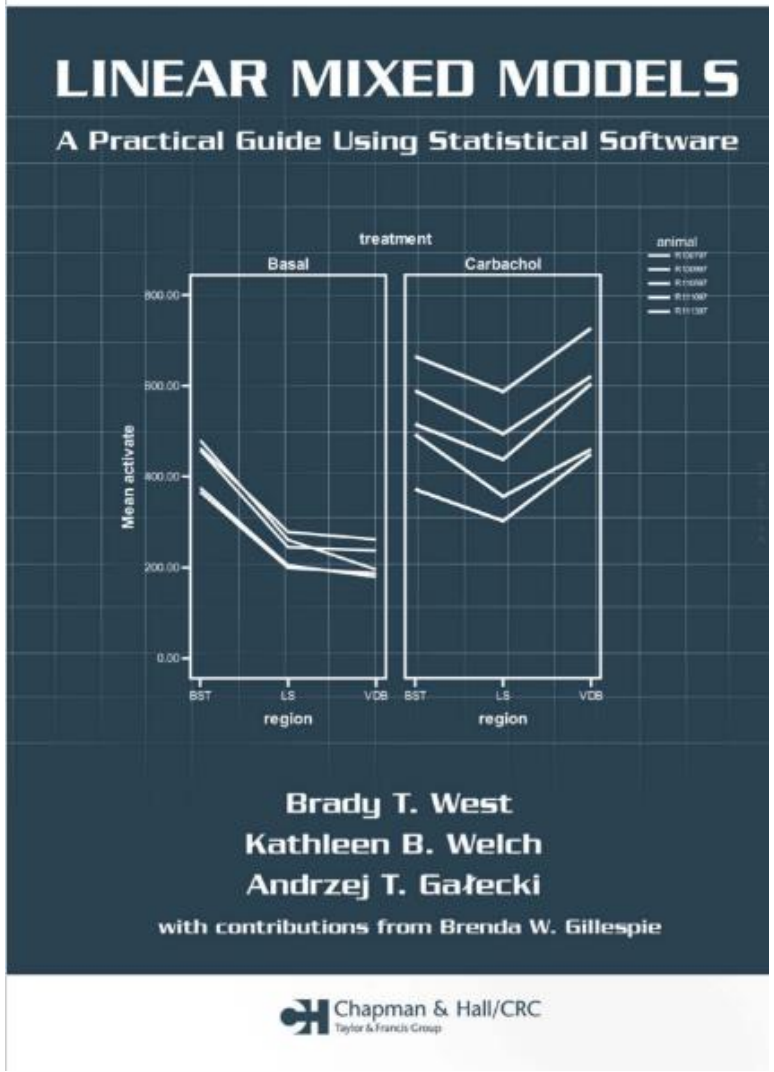
1. LMM의 구성 성분을 이해하고 분석하는 자료의 형태와 구조를 설명할 수 있다.
2. LMM의 구체적 기술을 개별 반응변수 또는 벡터(행렬)의 형태로 기술할 수 있다.
3. LMM 모수추정과 변량효과 예측 방법에 대해 요약할 수 있다.
4. LMM의 모형 구축 전략에 대해 개략할 수 있다.

01

제 11강. 선형혼합모형 소개

교재 소개

1. 교재



◆ LMM 분석에 소개되는 통계팩키지

- ✓ SAS
- ✓ SPSS
- ✓ R
  - ➢ lme() from nlme package
  - ➢ lmer() form lme4 package
- ✓ Stata
- ✓ HLM

◆ 자료와 프로그램들

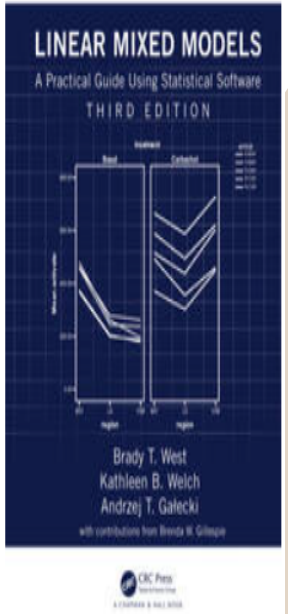
- ✓ http://www.umich.edu/~bwest/almmussp.html

2. 교재의 예제자료와 분석 프로그램(1)

Linear Mixed Models: A Practical Guide Using Statistical Software (Third Edition)

Brady T. West, Ph.D.
Kathleen B. Welch, MS, MPH
Andrzej T. Galecki, M.D., Ph.D.

Note: The third edition is now available via online retailers (e.g., crcpress.com, amazon.com).



This book provides readers with a practical introduction to the theory and applications of linear mixed models, and introduces the fitting and interpretation of several types of linear mixed models using the statistical software packages SAS (PROC MIXED / PROC GLIMMIX), SPSS (the MIXED and GENLINMIXED procedures), Stata (mixed), R (the lme() and lmer() functions), and HLM (Hierarchical Linear Models).

The book focuses on the statistical meaning behind linear mixed models. Why fit them? Why are they important? When are they applicable? What do they mean for research conclusions? The book also presents and compares practical, step-by-step analyses of real-world data sets in all of the aforementioned software packages, allowing readers to compare and contrast the packages in terms of their syntax/code, ease of use, available methods and options, and relative advantages.

Click on any of the following chapters for links to the data sets, updates to the software code in the book, and miscellaneous additional information:

Chapter 3 -> Two-level Models for Clustered Data: The Rat Pup Example
Chapter 4 -> Three-level Models for Clustered Data: The Classroom Example
Chapter 5 -> Models for Repeated Measures Data: The Rat Brain Example
Chapter 6 -> Random Coefficient Models for Longitudinal Data: The Autism Example
Chapter 7 -> Models for Clustered Longitudinal Data: The Dental Veneer Example
Chapter 8 -> Models for Data with Crossed Random Factors: The Sat Score Example

2. 교재의 예제자료와 분석 프로그램(2)

Chapter 3: Two-level Models for Clustered Data

Note: If given the option, right-click on the files, and choose "Save Link/Target As".

Data Sets
The Rat Pup Data
Level 1 SPSS Data Set for HLM
Level 2 SPSS Data Set for HLM
MDM Data File for HLM

Syntax for Mixed Model Analyses
SAS Syntax (includes syntax for descriptives and diagnostics)
SPSS Syntax (Version 28)
R Syntax: lme()
R Syntax: lmer()
Stata .do file (Version 17)

Syntax for Descriptive Statistics
SPSS Syntax for Descriptive Statistics
R Syntax for Descriptive Statistics
Stata Do-file for Descriptive Statistics
HLM Steps for obtaining Descriptive Statistics

Syntax for Final Model Diagnostics
SPSS Syntax for Final Model Diagnostics
R Syntax for Final Model Diagnostics
Stata Do-File for Final Model Diagnostics
Chapter 4 in the book describes how residual files can be saved and processed using the HLM software.

02

제 11강. 선형혼합모형 소개

LMM 자료의 형태와 구조

1. 예제자료(1)

생쥐 출생체중 자료(Pinheiro and Bates, 2000)

연구목적

- 생쥐의 출생 시 무게가 어미 쥐에게 폭로된 위험물질 용량군(세가지 군)에 따라 영향을 받는가?

1. 예제자료(1)

생쥐 출생체중 자료(Pinheiro and Bates, 2000)

자료의 생성과 특성

- 30마리 암컷 쥐를 세 가지 용량군(고용량군, 저용량군, 대조군)에 10마리 씩 랜덤 배정
- 각 암컷 쥐에서 태어난 생쥐의 무게를 측정
- 고용량군 배정된 세 마리 암컷 쥐는 새끼를 생산하기 전에 죽음
- 한 배에 생산된 생쥐들의 수 : 2~18

- 2-수준 군집자료(two-level clustered data)
  - Level 1: 관측 자료의 가장 아래 수준에서 관측되는 것
  - Level 2: Level 1의 바로 위 수준에서 관측되는 것

1. 예제자료(1)

생쥐 출생체중 자료(Pinheiro and Bates, 2000)

자료 구조(http://www.umich.edu/~bwest/almmussp.html)

pup_id	weight	sex	litter	litsize	treatment
1	6.6	Male	1	12	Control
2	7.4	Male	1	12	Control
3	7.15	Male	1	12	Control
4	7.24	Male	1	12	Control
5	7.1	Male	1	12	Control
6	6.04	Male	1	12	Control
7	6.98	Male	1	12	Control
8	7.05	Male	1	12	Control
9	6.95	Female	1	12	Control
10	6.29	Female	1	12	Control
11	6.77	Female	1	12	Control
12	6.57	Female	1	12	Control
13	6.37	Male	2	14	Control
14	6.37	Male	2	14	Control
15	6.9	Male	2	14	Control
310	6.21	Female	26	9	High
311	6.42	Female	26	9	High
312	6.42	Female	26	9	High
313	6.3	Female	26	9	High
314	5.64	Male	27	9	High
315	6.06	Male	27	9	High
316	6.56	Male	27	9	High
317	6.29	Male	27	9	High
318	5.69	Male	27	9	High
319	6.36	Male	27	9	High
320	5.93	Female	27	9	High
321	5.74	Female	27	9	High
322	5.74	Female	27	9	High

- **2-수준 군집자료(two-level clustered data)**
  - **Level 1 : 생쥐에서 관측되는 값**
    - ✓ **생쥐번호, 체중, 성별**
  - **Level 2 : 어미쥐에서 관측되는 값**
    - ✓ **어미쥐 번호, 배의 크기, 처리의 종류**

1. 예제자료(2)

수업개선연구자료(Anderson et al., 2009)

연구목적

- 교육환경(학교의 주변환경, 교사의 특성, 사회경제적 상태 등)은 초등학교 학생들의 수학성적 성취에 영향을 주는가?

1. 예제자료(2)

수업개선연구자료(Anderson et al., 2009)

자료의 생성과 특성

- U.S 초등학교 모집단에서 107개의 학교(school)를 확률 추출함
- 이들 107개의 학교에서 312개의 학급(classroom)을 확률 추출함.
- 312개의 학급에 있는 1,190명 초등학교 1학년 학생(student)들에 대해 유치원 때 대비 수학성적변화(MATHGAIN:)을 측정함.

- 3-수준 군집자료(three-level clustered data)
  - Level 1(student-level): 관측자료의 가장 아래 수준에서 관측되는 것
  - Level 2(classroom-level): Level 1의 바로 위 수준에서 관측되는 것
  - Level 3(school-level) Level 2의 바로 위 수준에서 관측되는 것

1. 예제자료(2)

수업개선연구자료(Anderson et al., 2009)

자료 구조(http://www.umich.edu/~bwest/almmussp.html)

sex	minority	mathkind	mathgain	ses	yearstea	mathknow	housepov	mathprep	classid	schoolid	childid
1	1	448	32	0.46	1		0.082	2	160	1	1
0	1	460	109	-0.27	1		0.082	2	160	1	2
1	1	511	56	-0.03	1		0.082	2	160	1	3
0	1	449	83	-0.38	2	-0.11	0.082	3.25	217	1	4
0	1	425	53	-0.03	2	-0.11	0.082	3.25	217	1	5
1	1	450	65	0.76	2	-0.11	0.082	3.25	217	1	6
0	1	452	51	-0.03	2	-0.11	0.082	3.25	217	1	7
0	1	443	66	0.2	2	-0.11	0.082	3.25	217	1	8
1	1	422	88	0.64	2	-0.11	0.082	3.25	217	1	9
0	1	480	-7	0.13	2	-0.11	0.082	3.25	217	1	10
0	1	502	60	0.83	2	-0.11	0.082	3.25	217	1	11
1	1	502	-2	0.06	1	-1.25	0.082	2.5	197	2	12
0	0	430	101	0.3	1	-1.25	0.082	2.5	197	2	13
0	0	526	30	-0.27	2	-0.72	0.082	2.33	211	2	14

◆ 3-수준 군집자료(three-level clustered data)

- Level 1(student-level)
  - ✓ 성별(sex), 소수자여부(minority), 유치원수학성적(mathkind), 수학성적변화(mathgain), 사회경제적상태(ses), 학생번호(childid)
- Level 2(classroom-level)
  - ✓ 교사의교육경험연수(yeartea), 교사의 수학교육준비수준(mathprep), 교사의수학지식(mathknow), 교실번호(classid)
- Level 3(school-level)
  - ✓ 학교주변의 주거환경(housepov), 학교번호(schoolid)

1. 예제자료(3)

쥐의 뇌 연구자료(Douglas et al., 2004)

연구목적

- 특정 실험 조건에서 쥐의 뇌 부위에 따라 뉴클레오타이드 활성화 정도에 차이가 있는가?

1. 예제자료(3)

쥐의 뇌 연구자료(Douglas et al., 2004)

자료의 생성과 특성

- 다섯 마리의 성체 쥐를 대상으로 뇌의 다른 영역 7곳에서 실험조건에 따른 뉴클레오타이드 활성화 정도를 방사능사진촬영(autoradiography)으로 측정함
- 각 뇌 영역에서 두 종류의 처리(①식염수(saline solution) ②카르바콜(carbachol))를 한 후 각 처리에서 흡광도(optical density)를 측정함.

- 반복측정자료(repeated-measures data)
  - 반복측정치(repeated measures)
    - 종속변수(흡광도)
    - 반복측정요인(repeated-measures factors)
      - ✓ 뇌의 영역(brain regions)
      - ✓ 처리(treatments)

1. 예제자료(3)

쥐의 뇌 연구자료(Douglas et al., 2004)

자료 구조(http://www.umich.edu/~bwest/almmussp.html)

animal	treatment	region	activate
R111097	1	1	366.19
R111097	1	2	199.31
R111097	1	3	187.11
R111097	2	1	371.71
R111097	2	2	302.02
R111097	2	3	449.7
R111397	1	1	375.58
R111397	1	2	204.85
R111397	1	3	179.38
R111397	2	1	492.58
R111397	2	2	355.74
R111397	2	3	459.58
R100797	1	1	458.16
R100797	1	2	245.04
R100797	1	3	237.42
R100797	2	1	664.72
R100797	2	2	587.1
R100797	2	3	726.96

◆ 반복측정자료 (repeated-measures data)

- 반복측정치(level 1)
  - ✓ 반복측정요인: 처리 종류(treatment), 뇌의 영역(region)
  - ✓ 종속변수: 흡광도(activate)
- 측정(실험) 개체(level 2)
  - ✓ 쥐(animal)

1. 예제자료(4)

자폐아 연구자료(Anderson et al. 2009)

연구목적

- 자폐아에서 장애의 유형, 초기 언어 능력이 사회화 정도에 영향을 주는가?

1. 예제자료(4)

자폐아 연구자료(Anderson et al. 2009)

자료의 생성과 특성

- **자폐**(ASD: autism spectrum disorder) **또는 전반적발달장애**(PDD: pervasive developmental disorder)**가 있는** 158**명의 어린 아이를 관찰한 연구**
- **각 어린이는 두 살 때 언어능력**(1=low, 2=medium, 3=high)**을 측정**
- 2, 3, 5, 9, 13**세 때에 사회화의 정도를 관측**

- **경시적 자료(longitudinal data)**
  - **경시적측정치(longitudial measures)**
    - **종속변수(사회화 정도)**
    - **경시적 요인(logitudinal factors)**
      - ✓ **나이(age)**

1. 예제자료(4)

자폐아 연구자료(Anderson et al. 2009)

자료 구조(http://www.umich.edu/~bwest/almmussp.html)

age	vsae	sicdegp	childid
2	6	3	1
3	7	3	1
5	18	3	1
9	25	3	1
13	27	3	1
2	17	3	3
3	18	3	3
5	12	3	3
9	18	3	3
13	24	3	3
2	12	3	4
3	14	3	4
5	38	3	4
9	114	3	4

◆ 경시적 자료 (longitudinal data)

- 경시적 측정치(level 1)
  - ✓ 경시적 요인: 나이(age)
  - ✓ 종속변수: 사회화 정도(vsae)
- 측정개체 (level 2)
  - ✓ 아이(childid)
  - ✓ 2세 때 언어능력(sicdegp)

2. 예제자료들의 특성

- 종속변수는 연속형
- 종속변수에 대한 독립성 ?
- 종속변수에 대한 등분산성 가정 ?

예: 자폐아 연구자료

단순 통계량

변수	N	평균	표준편차 값
vsae2	156	9.08974	3.85607
vsae3	149	15.25503	7.97814

피어슨 상관 계수
H0: Rho=0 가정하에서 Prob > |r|
관측치 개수

	vsae2	vsae3
vsae2	1.00000	0.50826 <.0001 156
		147
vsae3	0.50826 <.0001 147	1.00000 149

→ 등분산성 의심!

→ 독립성 의심!

3. 선형혼합모형(LMM: Linear Mixed Models) 이란?

LMM의 주요 특성

◆ 모수적 선형모형의 하나

◆ 종속변수의 형태
- ✓ 연속형 종속변수(continuous dependent variable)

◆ 독립변수 또는 예측변수
- ✓ 연속형/범주형 공변량: 고정효과(fixed effect)를 모형화
- ✓ 변량요인(random factors): 변량효과(random effect)를 모형화
  - ➢ 고정효과의 변동성 유입 → 등분산성 가정을 이분산성 모형으로 확장
  - ➢ 종속변수의 상관성 유입 → 독립성 가정을 상관된 자료분석 모형으로 확장

◆ 적용하는 자료의 대표적인 형태
- ✓ 군집자료(clustered data)
- ✓ 반복측정자료(repeated-measures data)
- ✓ 경시적 자료(longitudinal data) 등

4. 요인의 유형과 관련효과

요인(factor) 이란?

◆ 종속변수의 변동성을 유발하는(할 것으로 예측되는) 변수

- ✓ 생쥐 출생체중자료
  - ➢ 위험물질 용량, 성별, 배의 크기, 어미 쥐 등
- ✓ 자폐아 연구자료
  - ➢ 나이, 2세 때 언어능력, 아이 등

4. 요인의 유형과 관련효과

요인의 구분: 고정요인 vs. 변동요인

◆ 고정요인(fixed factors)

✓ 성별, 처리조건, 나이집단 등과 같이 변수의 수준이 연구의 관심 영역에서 고정되는 변수

✓ 연속형 공변량, 범주형 변수, 분류변수 등

- 생쥐 출생체중자료
  - ➢ 위험물질 용량, 성별, 배의 크기
- 자폐아 연구자료
  - ➢ 나이, 2세 때 언어능력

4. 요인의 유형과 관련효과

요인의 구분: 고정요인 vs. 변동요인

◆ 변량요인(random factors)

✓ 변수의 수준이 모집단에서 확률적으로 추출(또는 선택)된 것으로 판단되는(될 수 있는) 변수

✓ 변수가 취할 수 있는 모든 수준(또는 값이)이 자료에 실현되지는 않는다.

- 생쥐 출생체중자료
  - 어미 쥐
- 자폐아 연구자료
  - 자폐아 환아

4. 요인의 유형과 관련효과

요인의 유형에 따른 효과: 고정효과 vs. 변동효과

◆ 고정효과(fixed effects)
- ✓ 고정요인의 영향
- ✓ 종속변수와 공변량의 관계를 모형화 하는 모수
- ✓ 회귀분석에서는 회귀계수 또는 고정효과 모수
- ✓ 미지의 상수

◆ 변량효과(random effects)
- ✓ 변량요인의 영향
- ✓ 확률변수가 실현될(또는 예측) 값으로 설정
- ✓ 고정효과의 변동성을 모형화 하는 확률변수
  - ➢ 변량절편(random intercepts), 변량계수(random coefficients) 등

5. 선형모합모형(LMM: Linear Mixed Models) 이란?

- 군집자료, 반복측정자료, 경시적자료 등에서 측정되는 연속형 종속변수에 적용되는 회귀모형
- 종속변수의 분포는 정규분포를 가정
- 고정요인과 변량요인(random factors)를 선형으로 혼합하는 모형

03

제 11강. 선형혼합모형 소개

LMM의 구체적 기술

1. 개별 관측값에 대한 기술

모형의 기술

$$Y_{ti} = \beta_1 X_{ti}^{(1)} + \cdots + \beta_p X_{ti}^{(p)} \quad \} \text{ 고정요인 모형화}$$
$$+ u_{1i} Z_{ti}^{(1)} + \cdots + u_{qi} Z_{ti}^{(q)} + \varepsilon_{ti} \} \text{ 변량요인 모형화}$$

✓ 기호에 대한 정의와 모형 가정

- Y_{ti}: i번째 개체에서 t 시점에 관측된 반응변수$i = 1, \cdots m,\ t = 1, \cdots n_i$
- $\beta_1, \cdots, \beta_p$: 고정효과를 나타내는 모수
- $X_{ti}^{(1)}, \cdots X_{ti}^{(p)}$: 1를 포함하여 고정인자를 나타내는 공변량(covariates)
- $u_{1i}, \cdots, u_{qi}$: 변량효과를 나타내는 확률변수로 정규분포를 가정
- $Z_{ti}^{(1)}, \cdots Z_{ti}^{(q)}$: 변량효과 관련된 공변량
- ε_{ti} : 오차항으로 정규분포를 가정
- 변량효과와 오차항은 서로 독립이라고 가정

2. 행렬을 사용한 기술

모형의 기술

- $Y_i = X_i\beta + Z_i u_i + \varepsilon_i \ , \ \mathrm{i} = 1, \cdots, m$

  고정효과 모형화 변량효과 모형화

여기서

$$Y_i = \begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{1i}^{(1)} & \cdots & X_{1i}^{(p)} \\ \vdots & \ddots & \vdots \\ X_{n_i i}^{(p)} & \cdots & X_{n_i i}^{(p)} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$Z_i = \begin{pmatrix} Z_{1i}^{(1)} & \cdots & Z_{1i}^{(q)} \\ \vdots & \ddots & \vdots \\ Z_{n_i i}^{(p)} & \cdots & Z_{n_i i}^{(q)} \end{pmatrix}, \quad u_i = \begin{pmatrix} u_{1i} \\ \vdots \\ u_{qi} \end{pmatrix},$$

2. 행렬을 사용한 기술

모형의 기술

$u_i \sim N(0, D)$,
$\varepsilon_i \sim N(0, R_i)$,

$$D = Var(u_i) = \begin{pmatrix} Var(u_{1i}) & \cdots & cov(u_{1i}, u_{qi}) \\ \vdots & \ddots & \vdots \\ cov(u_{1i}, u_{qi}) & \cdots & Var(u_{qi}) \end{pmatrix},$$

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} Var(\varepsilon_{1i}) & \cdots & cov(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & \cdots & Var(\varepsilon_{n_i i}) \end{pmatrix},$$

$Cov(u_i, \varepsilon_i) = 0$, 즉 u_i와 ε_i 는 서로 독립.

2. 행렬을 사용한 기술

변량효과의 분산-공분산 행렬 D 의 구조

- 변량 효과 2개 이상일 때, 분산-공분산 행렬 D 의 구조를 가정 필요
- 자주 사용되는 대표적인 D 의 구조

▪ **변량효과가 두 개일 때**

➢ **무구조 행렬(unstructured matrix)**

$$D = Var(u_i) = \begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1u_2} \\ \sigma_{u_1u_2} & \sigma_{u_2}^2 \end{pmatrix}$$

➢ **분산성분 구조(variance component structure)**

$$D = Var(u_i) = \begin{pmatrix} \sigma_{u_1}^2 & 0 \\ 0 & \sigma_{u_2}^2 \end{pmatrix}$$

2. 행렬을 사용한 기술

오차항의 분산-공분산 행렬 R_i의 구조

대표적인 R_i 구조(1)

- **대각행렬(diagonal matrix)**

$$R_i = Var(\varepsilon_i) = I\sigma^2 = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

- **합성대칭행렬(compound symmetry matrix)**

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2 + \sigma_1 \end{pmatrix}$$

2. 행렬을 사용한 기술

오차항의 분산-공분산 행렬 R_i의 구조

대표적인 R_i 구조(2)

- **1차 자기상관행렬(first-order autoregressive matrix)**

$$Var(\varepsilon_i) = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \cdots & \sigma^2 \end{pmatrix}$$

04

제 11강. 선형혼합모형 소개

LMM 모수 추정

1. 모수 추정의 대상

고정효과 모수(fixed effect parameters)

- β : 미지의 회귀모수

분산-공분산 모수(variance-covariance parameters)

- θ_D : 행렬 D에 있는 미지의 모수
- θ_R : 행렬 R_i에 있는 미지의 모수

2. 최대가능도(ML : Maximum Likelihood) 추정

가능도함수(likelihood function)란?

- 가정한 모형의 확률밀도함수에서 표본관측값을 대입한 모수의 함수
- i 번째 개체의 가능도함수

$$L_i(\beta,\theta) = \det(2\pi V_i)^{-1/2} exp\left[-0.5(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta)\right]$$

여기서 $\det(\cdot)$는 행렬식을 나타내며 y_i는 Y_i의 관측값, $V_i = Z_i D Z_i' + R_i$

2. ML 추정

로그가능도함수(log-likelihood function)

전체 자료의 로그가능도함수

$$l(\beta,\theta) = log\left[\prod_{i=1}^{m} L_i(\beta,\theta)\right]$$
$$= -\frac{1}{2}\sum_{i=1}^{m}\{(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta) + log[\det(2\pi V_i)]\}$$

최대가능도 추정값 (maximum likelihood estimate)

- 로그가능도함수 $l(\beta,\theta)$를 최대가 되게 하는 값

2. ML 추정

분산-공분산 모수 θ의 최대가능도 추정

θ의 단면로그가능도함수(profile log-likelihood function)

$$l_{ML}(\theta) = -\frac{1}{2}\sum_{i=1}^{m}\{r_i' V_i^{-1} r_i + log[\det(2\pi V_i)]\}$$

여기서 $r_i = y_i - X_i'[(\sum_i X_i' V_i^{-1} X_i)^{-1} \sum_i X_i' V_i^{-1} y_i]$

θ의 최대가능도 추정값

- $\hat{\theta}$: $l_{ML}(\theta$를 최대가 되게 하는 θ 값

2. ML 추정

회귀모수 β의 최대가능도 추정

β의 최대가능도 추정값

$$\hat{\beta} = \left(\sum_i X_i' \widehat{V_i^{-1}} X_i\right)^{-1} \sum_i X_i' \widehat{V_i^{-1}} y_i$$

여기서 $\widehat{V_i^{-1}}$은 V_i^{-1}에 있는 θ 대신에 $\hat{\theta}$으로 대체한 것

$\hat{\beta}$의 분산-공분산 행렬

$$Var(\hat{\beta}) = \left(\sum_i X_i' \widehat{V_i^{-1}} X_i\right)^{-1}$$

3. 제한최대가능도(REML : Restricted Maximum

REML 추정이란?

- 분산-공분산 모수 θ 추정에 가능도함수를 사용하는 대신에 제한가능도함수(restricted likelihood function)를 사용하는 것
- REML 가능도 함수
  - 가능도함수에서 회귀모수 β의 추정으로 인한 정보의 손실(자유도)을 보정한 함수 (Harville, 1977; Cooper and Thompson, 1977; Verbyla, 1990)
  - 분산-공분산 모수 θ의 REML 추정은 β의 불편추정량을 제공
  - 잔차가능도함수(residual likelihood function)라고도 함

3. REML 추정

θ의 REML 추정

θ 의 REML 로그가능도함수(REML log-likelihood function)

$$l_{REML}(\theta) = -\frac{1}{2}\sum_{i=1}^{m}\{r_i' V_i^{-1} r_i + log[\det(2\pi V_i)]\} - \underbrace{\frac{1}{2} log\left\{det\left[\frac{1}{2\pi}\sum_i X_i' V_i^{-1} X_i\right]\right\}}$$

- β의 추정으로 인한 정보손실(자유도) 보정

θ 의 REML 추정값

$\widehat{\theta^{REML}}$: l_{REML}을 최대가 되게 하는 값 θ

05

제 11강. 선형혼합모형 소개

LMM 변량효과 예측

1. 변량효과 예측

최량선형불편예측량(BLUP : Best Linear Unbiased Predictors)

u_i의 EBLUPs (emprical BLUPs) : $\widehat{u_i}$

$$\hat{u}_i = E(u_i|Y_i = y_i) = \hat{D}Z_i'\widehat{V_i^{-1}}(y_i - X_i\hat{\beta})$$

$\widehat{u_i}$ 의 분산-공분산 행렬

$$Var(\hat{u}_i) = \hat{D}Z_i'\left(\widehat{V_i^{-1}} - \widehat{V_i^{-1}}X_i\left(\sum X_i\widehat{V_i^{-1}}X_i\right)^{-1}X_i\widehat{V_i^{-1}}\right)Z_i\hat{D}$$

$\widehat{u_i}$ 의 특성

축소추정량(shrinkage estimators) : u_i를 고정효과로 추정했을 때 보다 0으로 축소됨

06

제 11강. 선형혼합모형 소개

LMM 모형 구축 전략

1. 하향식 전략 (Top-Down Strategy)

하향식 모형구축전략(Verbeke & Molenberghs 2000)

(1) 가능한 꽉 찬 평균모형 ($E(Y_{ij}) = X_{ij}\beta$)에서 출발

- 교호작용 효과를 포함하여 가능한 많은 수의 고정효과를 모형에 포함
  - 공분산 구조를 탐구하기 전에 종속변수에 내포된 체계적인 변동성이 최대한 설명되도록 함

(2) 변량효과와 공분산 구조(D) 를 선택

- **모형에 포함될 변량효과와 공분산 구조를 선택**
  - **변량효과와 공분산 구조에 대하여** REML **기반 가능비 검정으로 유의성을 검정**

1. 하향식 전략 (Top-Down Strategy)

하향식 모형구축전략(Verbeke and Molenberghs 2000)

(3) 오차 공분산 구조 (R_i)를 선택

- 적합해 보이는 몇 가지 후보 공분산 구조를 가정
  - REML 기반 가능도비 검정(검정 모형 간에 지분관계가 성립할 때) 또는 정보량기준 (검정 모형 간에 지분관계가 성립하지 않을 때)으로 가장 적절한 공분산 구조를 선택

(4) 평균 모형의 축소

- 적절한 검정법(근사적인 $F-$검정 또는 $T-$검정, 가능도비 검정 등)으로 고정효과의 유의성을 검정

2. 상향식 전략 (Step-Up Strategy)

상향식 모형구축전략(Raundenbush & Bryk 2002, Snijders & Bosker 1999)

HLMs(Heirarchical Linear Models)에서 개발된 모형구축전략

(1) 가장 단순한 ($E(Y_{ij}) = \beta_0$)에서 출발

- Level 1 의 고정효과로 절편항 만 포함. Level 2, Level 3 의 변량인자와 관련된 변량효과는 모형에 포함
  - 공변량의 보정없이 level 에 따른 종속변수의 변동성 평가

(2) Level 1 모형 구축과 Level 2 변량 효과 추가

- Level 1 **공변량을 고정효과 모형에 추가**
- Level 2 **변량효과를** Level 1 **공변량의 고정효과에 추가하는 시도**

2. 상향식 전략 (Step-Up Strategy)

상향식 모형구축전략(Raundenbush & Bryk 2002, Snijders & Bosker 1999)

(3) Level 2 모형 구축과 Level 3 변량 효과 추가

- Level 2 **공변량을 고정효과 모형에 추가**
  - Level 2 **모형에서** Level 1 **공변량 모형이 설정되면** Level 2 **모형의 변량효과에 대한 모형 가정(정규성, 등분산성 등)을 평가**
- 3-level **자료인 경우**, Level 3 **변량효과를** Level 2 **공변량의 고정효과에 추가하는 시도**

12 강

다음시간안내

LMM
2-수준 군집자료분석

수고하셨습니다.