13 강

데이터분석방법론2

LMM
3-수준 군집자료분석

대전대학교 빅데이터인공지능학과 강위창 교수

학습목차

학습개요 및 목표

이번 강의에서는 3-수준 군집자료를 분석하는 LMM을 예제자료 중심으로 소개합니다.

1. 3-수준 군집자료의 통계적 특성을 이해하고 해당 자료에 적합한 LMM 분석모형을 제시할 수 있다.
2. 3-수준 군집자료에 LMM을 올바르게 적합시키고 분석결과를 해석할 수 있다.

01

제 13강. LMM 3-수준 군집자료분석

3-수준 군집자료

1. 예제자료

수업개선연구자료(Anderson et al., 2009)

연구목적

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

1. 예제자료

수업개선연구자료(Anderson et al., 2009)

자료의 생성과 특성

- U.S 초등학교 모집단에서 107개의 학교(school)를 확률 추출함
- 추출된 각 학교에서 1학년의 일부 학급을 랜덤 추출: 312개 학급
- 추출된 각 학급에서 학생 일부를 랜덤 추출: 1190명 학생
- 유치원 때 대비 수학성적변화(MATHGAIN:)을 측정함.

3-수준 군집자료(three-level clustered data)

- **Level 1(student-level): 관측 자료의 가장 아래 수준(학생)**
- **Level 2(classroom-level): Level 1의 바로 위 수준(학급)**
- **Level 3(school-level) Level 2의 바로 위 수준(학교)**

1. 예제자료(2)

수업개선연구자료(Anderson et al., 2009)

자료 구조(https://websites.umich.edu/~bwest/chapter4.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)
  - ✓ 수학성적변화(mathgain), 성별(sex), 소수자여부(minority), 유치원수학성적(mathkind), 사회경제적상태(ses), 학생번호(childid)
- Level 2(classroom-level)
  - ✓ 교사의교육경험년수(yeartea), 교사의 수학교육준비수준(mathprep), 교사의수학지식(mathknow), 교실번호(classid)
- Level 3(school-level)
  - ✓ 학교주변의 주거환경(housepov), 학교번호(schoolid)

2. 자료 준비

- https://websites.umich.edu/~bwest/chapter4.html 에서 자료를 다운받아 classroom.csv 파일로 저장.

- classroom.csv 파일이 아래 폴더 [C:₩강위창₩방통대₩데이터분석방법론2₩강의노트_2024₩예제자료]에 있다고 가정

02

제 13강. LMM 3-수준 군집자료분석

LMM : 군집자료의 특성 탐색

1. 요인(변수)의 구분 : 고정요인 vs. 변량요인

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
0	0	504	65	0.74	2	-0.72	0.082	2.33	211	2	15
1	0	527	29	0.31	2	-0.72	0.082	2.33	211	2	16
1	0	462	152	1.1	2	-0.72	0.082	2.33	211	2	17
0	0	483	50	0.52	12.54		0.082	2.3	307	2	18
1	1	516	60	-1.15	12.54		0.082	2.3	307	2	19
0	1	476	74	0	12.54		0.082	2.3	307	2	20
0	1	453	91	0.39	12.54		0.082	2.3	307	2	21

- 고정요인 (fixed factors)
  - Level 1 : 성별(sex: 0=남, 1=여), 소수자여부(minority: 0=,no, 1=yes), 유치원수학성적(mathkind), 사회경제적상태(ses)
  - Level 2 : 교사교육경험년수(yeartea), 교사수학교육준비수준(mathprep), 교사수학지식(mathknow)
  - Level 3 : 학교주변 주거환경(housepov)
- 변량요인 (random factors)
  - Level 1 : 학생
  - Level 2 : 교실
  - Level 3 : 학교

2. 탐색적 자료분석 : 기술 통계량

Level 1 변수들의 기술 통계량

```
> ##### 탐색적 분석 : Classroom data
> class <- read.csv("C:/강위창/방통대/데이터분석방법론2/강의노트_2024/예제자료/classroom.csv", h = T)
> ## Level 1 기술통계량
> attach(class)
> level1 <- data.frame(mathgain, sex,minority,mathkind,ses)
> summary(level1)
   mathgain            sex            minority        mathkind          ses          
 Min.   :-110.00   Min.   :0.0000   Min.   :0.0000   Min.   :290.0   Min.   :-1.61000
 1st Qu.:  35.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:439.2   1st Qu.:-0.49000
 Median :  56.00   Median :1.0000   Median :1.0000   Median :466.0   Median :-0.03000
 Mean   :  57.57   Mean   :0.5059   Mean   :0.6773   Mean   :466.7   Mean   :-0.01298
 3rd Qu.:  77.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:495.0   3rd Qu.: 0.39750
 Max.   : 253.00   Max.   :1.0000   Max.   :1.0000   Max.   :629.0   Max.   : 3.21000
```

- mathgain : 중앙값 < 평균값
- sex : 여학생 50.6%
- minority : 소수자 67.7%
- ses : 중앙값 < 평균값

2. 탐색적 자료분석 : 기술 통계량

Level 2 변수들의 기술 통계량

```
> level.a2 <- aggregate(class,list(classid = class$classid),mean)
> level2 <- data.frame(level.a2$yearstea, level.a2$mathprep, level.a2$mathknow)
> summary(level2)
 level.a2.yearstea level.a2.mathprep level.a2.mathknow
 Min.   : 0.00     Min.   :1.000     Min.   :-2.50000
 1st Qu.: 4.00     1st Qu.:2.000     1st Qu.:-0.76000
 Median :10.00     Median :2.300     Median :-0.19000
 Mean   :12.28     Mean   :2.577     Mean   :-0.08025
 3rd Qu.:20.00     3rd Qu.:3.000     3rd Qu.: 0.62000
 Max.   :40.00     Max.   :6.000     Max.   : 2.61000
                                     NA's   :27
```

- yeartea : 평균 12.28년 (중앙값 < 평균값)
- mathknow : 27 개의 classroom(교사)에서 결측치 존재

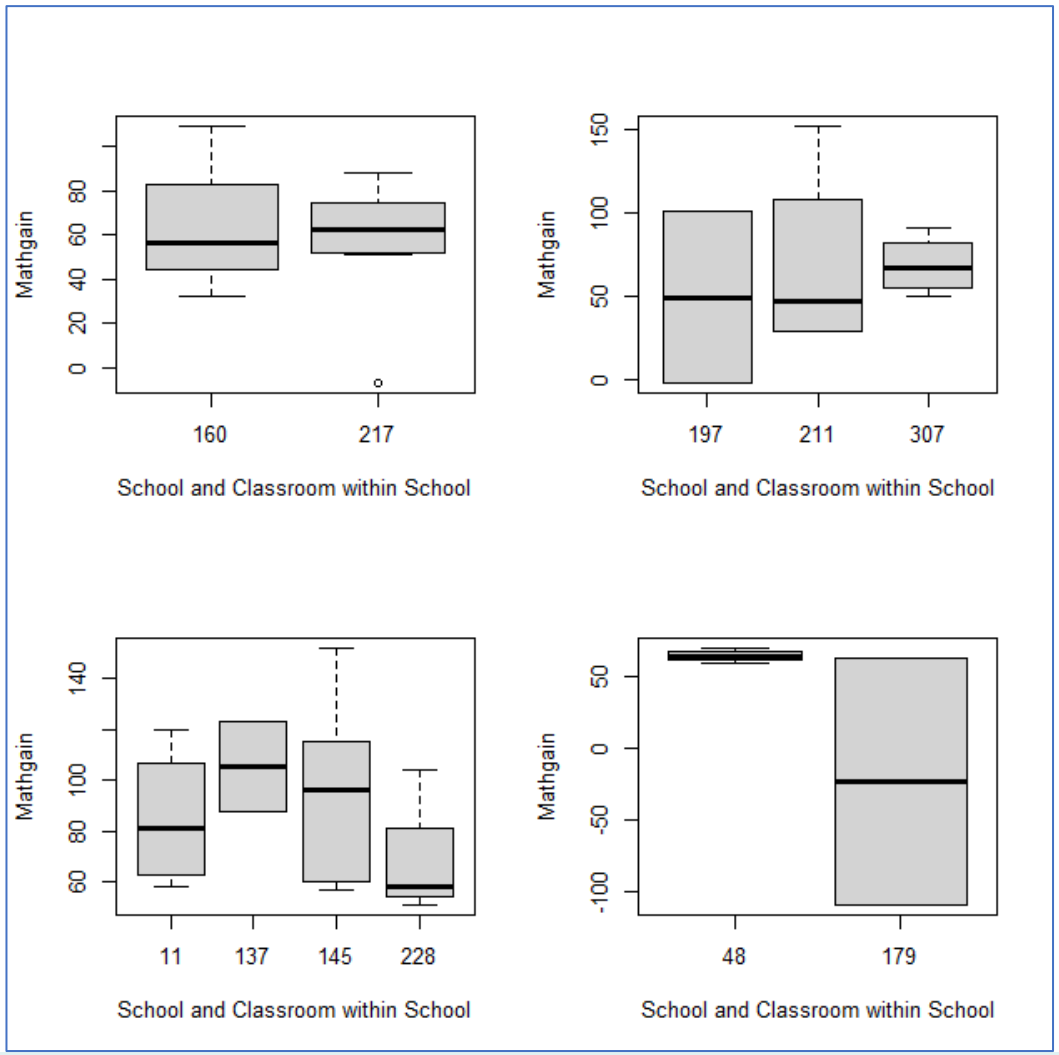
Level 3 변수의 기술 통계량

```
> level3 <- aggregate(class,list(schoolid = class$schoolid),mean)
> summary(level3$housepov)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0120  0.0855  0.1480  0.1941  0.2645  0.5640
```

- housepov : 평균 이하 재산 가구 비율의 평균 19.41%

2. 탐색적 자료분석 : 학교와 학급에 따른 박스그림

4개 학교와 각 학교의 학급에서의 종속변수 박스그림



- 종속변수(mathgain)의 분포(평균과 분산 등)는
  - ➢ 학교(school)에 따라서 변동성이 존재
  - ➢ 각 학교내의 학급에 따라서도 변동성이 존재

- 학교와 학급을 수준 별 변량요인으로 설정하는 Multilevel models (LMM) 분석 시도
  - ✓ 변량절편모형
  - ✓ 변량계수모형 등

2. 탐색적 자료분석 : 학교와 학급에 따른 박스그림

4개 학교와 각 학교의 학급에서의 종속변수 박스그림

R code

```
class.fst4 <- class[class$schoolid <=4,]
par(mfrow=c(2,2))
for (i in 1:4)
{boxplot(class.fst4$mathgain[class.fst4$schoolid==i]
          ~ class.fst4$classid[class.fst4$schoolid==i],
        ylab="Mathgain", xlab="School and Classroom within School")}
```

03

제 13강. LMM 3-수준 군집자료분석

LMM : 군집자료의 분석모형 구축

1. Classroom 연구자료 모형구축 : Three-level 변량절편모형

상향식(Step-Up) 모형구축 전략

Step1 : "mean-only"(variance components) 모형 적합

- 고정효과는 "상수항"만 포함
- 2-수준 및 3-수준 변량요인(각각 학급, 학교) 포함
- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma^2_{cl}),\ u_k \sim^{iid} N(0, \sigma^2_{sch})$$

이고 $\varepsilon_{ijk}, u_{j|k}, u_k$ 는 서로 독립.

- 분산성분모형(variance components models)
- 3-수준 변량절편모형

1. Classroom 연구자료 모형구축 : Three-level 변량절편모형

상향식(Step-Up) 모형구축 전략

Step1 : "mean-only"(variance components) 모형 적합

- "mean-only" 모형

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

  - "unconditional" 모형이라고도 함.
- Step1 에서 모형 검정의 내용
  - 변량효과의 유의성 검정(가능도비 검정)
    - ✓ 학급변량효과 $u_{j|k}$의 유의성
    - ✓ 학교변량효과 u_k의 유의성

1. Classroom 연구자료 모형구축 : Three-level 변량절편모형

상향식(Step-Up) 모형구축 전략

Step2 : Level 1 공변량 추가한 모형 적합

- Step1 에서 선택된 모형에 Level 1 공변량 추가 모형

$$Mathgain_{ijk} = \beta_0 + X_{ijk}\beta_1 + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서 X_{ijk}는 $(1 \times p)$ Level 1 공변량 벡터, β_1는 $(p \times 1)$ 미지의 모수벡터

- Step2 에서 모형 검정의 내용(가능도비 검정, F-검정, t-검정)
  - ✓ β_1 의 유의성

1. Classroom 연구자료 모형구축 : Three-level 변량절편모형

상향식(Step-Up) 모형구축 전략

Step3 : Level 2 공변량 추가한 모형 적합

- Step2 에서 선택된 모형에 Level 2 공변량 추가 모형

$$Mathgain_{ijk} = \beta_0 + X_{ijk}\beta_1 + X_{jk}\beta_2 + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서 X_{jk}는 $(1 \times q)$ Level 2 공변량 벡터, β_2는 $(q \times 1)$ 미지의 모수벡터

- Step3 에서 모형 검정의 내용(가능도비 검정, F-검정, t-검정)
  - ✓ β_2 의 유의성

1. Classroom 연구자료 모형구축 : Three-level 변량절편모형

상향식(Step-Up) 모형구축 전략

Step4 : Level 3 공변량 추가한 모형 적합

- Step3 에서 선택된 모형에 Level 3 공변량 추가 모형

$$Mathgain_{ijk} = \beta_0 + X_{ijk}\beta_1 + X_{jk}\beta_2 + X_k\beta_3 + u_{j|k} + u_k + \varepsilon_{ijk}$$

  여기서 X_k는 $(1 \times l)$ Level 3 공변량 벡터, β_3는 $(l \times 1)$ 미지의 모수벡터

- Step4 에서 모형 검정의 내용(가능도비 검정, F-검정, t-검정)
  - ✓ β_3 의 유의성

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step1

"mean-only" 모형(모형4.1) 적합

- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

이고 $\varepsilon_{ijk}, u_{j|k}, u_k$ 는 서로 독립.

모형4.1 적합

```
> library(nlme)
> # "mean-only" 모형(variance components models)(Model 4.1)
> model4.1.fit <- lme(mathgain~1,random = ~1|schoolid/classid,data=class, method = "REML")
```

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step1

"mean-only" 모형(모형4.1) 적합

- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

```
> summary(model4.1.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.802955

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.961301 32.06609
```

- $\widehat{\sigma_{sch}} = 8.80,\ \widehat{\sigma_{sch}^2} = 77.44$
- $\widehat{\sigma_{cl}} = 9.96,\ \widehat{\sigma_{cl}^2} = 99.20$
- $\hat{\sigma} = 32.07,\ \widehat{\sigma^2} = 1028.49$

- 오차분산(3-수준) 추정값 큼
- 3-수준 공변량의 추가 시사

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step1

변량효과 유의성 검정 : 학교 내 학급의 유의성 검정

- **가설의 설정**

$$H_0 : \sigma_{cl}^2 = 0 \quad vs. \quad H_1 : \sigma_{cl}^2 > 0$$

귀무가설 모형

$$Mathgain_{ijk} = \beta_0 + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

대립가설 모형

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step1

변량효과 유의성 검정 : 학교 내 학급의 유의성 검정

- 가능도비 검정통계량 LR과 분포

$$LR = -2log\left(\frac{L_{H_0}}{L_{H_1}}\right) \approx^{H_0} 0.5 \times x^2(0) + 0.5 \times x^2(1)$$

여기서 L_{H_0}과 L_{H_1} 은 각각 귀무가설과 대립가설에서 구한 REML값

귀무가설 모형 적합

```
> model4.1A.fit <- lme(mathgain~1, random=~1|schoolid, data=class, method = "REML")
```

대립가설 모형 적합

```
> model4.1.fit <- lme(mathgain~1,random = ~1|schoolid/classid,data=class, method = "REML")
```

검정 결과

```
> anova(model4.1.fit, model4.1A.fit)
              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model4.1.fit      1  4 11776.76 11797.09 -5884.382
model4.1A.fit     2  3 11782.67 11797.91 -5888.335 1 vs 2 7.904762  0.0049
```

p-value= $0.5 * 0.0049 = 0.0025$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step2

Level 1 공변량 추가한 모형(모형4.2) 적합

- Level 1 공변량 : sex(x_{1ijk}: 0=남, 1=여), minority (x_{2ijk}: 0=,no, 1=yes), mathkind(x_{3ijk}), ses(x_{4ijk})을 모형4.1 에 추가
- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk} + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma^2_{cl}),\ u_k \sim^{iid} N(0, \sigma^2_{sch})$$

모형4.2 적합

```
# 모형2: Level 1 공변량 추가한 모형: Model 4.2.
model4.2.fit <- lme(mathgain~sex+minority+mathkind+ses, random=~1|schoolid/classid,
                    data=class, na.action = "na.omit", method = "REML")
summary(model4.2.fit)
```

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step2

Level 1 공변량 추가한 모형(모형4.2) 적합

```
> summary(model4.2.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.671991

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:     9.12604 27.10286

Fixed effects:  mathgain ~ sex + minority + mathkind + ses
                Value Std.Error  DF    t-value p-value
(Intercept) 282.79034 10.853234 874  26.055860  0.0000
sex          -1.25119  1.657730 874  -0.754762  0.4506
minority     -8.26213  2.340113 874  -3.530655  0.0004
mathkind     -0.46980  0.022266 874 -21.099524  0.0000
ses           5.34638  1.241094 874   4.307794  0.0000
```

모수 추정치

- $\widehat{\beta_1} = -1.25$
- $\widehat{\beta_2} = -8.26$
- $\widehat{\beta_3} = -0.47$
- $\widehat{\beta_4} = 5.35$
- $\widehat{\sigma_{sch}} = 8.67$
- $\widehat{\sigma_{cl}} = 9.12$
- $\hat{\sigma} = 27.10$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step2

Level 1 공변량 유의성 검정

- **가설의 설정**

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad vs. \quad H_1 : not\ H_0$$

귀무가설 모형

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

대립가설 모형

$$Mathgain_{ijk} = \beta_0 + \underline{\beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk}} + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step2

Level 1 공변량 유의성 검정 : 가능도비 검정

- 가능도비 검정통계량 LR과 분포

$$LR = -2log\left(\frac{L_{H_0}}{L_{H_1}}\right) \approx^{H_0} x^2(4)$$

여기서 L_{H_0}과 L_{H_1} 은 각각 귀무가설과 대립가설에서 구한 ML값

유의성 검정과 결과

```
> ## Level 1 공변량 유의성 검정: 가능도비 검정
> # 귀무가설: "means-only"모형(모형4.1): ML estimation.
> model4.1.ml.fit <- lme(mathgain~1,random=~1|schoolid/classid, data=class, method="ML")
> # 대립가설: 모형4.2: ML estimation.
> model4.2.ml.fit <- lme(mathgain~sex+minority+mathkind+ses, random=~1|schoolid/classid,
+                    data=class, na.action="na.omit", method="ML")
> anova(model4.1.ml.fit, model4.2.ml.fit)
                Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model4.1.ml.fit     1  4 11779.33 11799.66 -5885.666
model4.2.ml.fit     2  8 11406.96 11447.62 -5695.481 1 vs 2 380.3684  <.0001
```

p-value< 0.0001

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step3

모형4.2에 Level 2 공변량 추가한 모형(모형4.3) 적합

- Level 2 공변량 : yeartea(x_{5jk}), mathknow(x_{6jk}), mathprep(x_{7jk})을 모형4.2에 추가
- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk}$$
$$+\beta_5 * x_{5jk} + \beta_6 * x_{6jk} + \beta_6 * x_{7jk} + u_{j|k} + u_k + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma^2_{cl}),\ u_k \sim^{iid} N(0, \sigma^2_{sch})$$

모형4.3 적합

```
# 모형3: 모형4.2에 Level 2 공변량 추가한 모형: Model 4.3.
model4.3.fit <- lme(mathgain~sex+minority+mathkind+ses+yearstea+mathknow+mathprep,
                    random=~1|schoolid/classid,data=class, na.action="na.omit", method="REML")
```

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step3

모형4.2에 Level 2 공변량 추가한 모형(모형4.3) 적합

```
> summary(model4.3.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.671285

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.310153 26.71766

Fixed effects:  mathgain ~ sex + minority + mathkind +
                Value Std.Error  DF    t-value p-value
(Intercept) 282.02452 11.701687 792  24.101185  0.0000
sex          -1.33950  1.718580 792  -0.779423  0.4360
minority     -7.86886  2.418081 792  -3.254177  0.0012
mathkind     -0.47501  0.022747 792 -20.882471  0.0000
ses           5.41925  1.275995 792   4.247078  0.0000
yearstea      0.03974  0.117070 177   0.339435  0.7347
mathknow      1.91448  1.147015 177   1.669094  0.0969
mathprep      1.09485  1.148493 177   0.953296  0.3417
```

모수 추정치

- $\widehat{\beta_1} = -1.34$
- $\widehat{\beta_2} = -7.87$
- $\widehat{\beta_3} = -0.48$
- $\widehat{\beta_4} = 5.42$
- $\widehat{\beta_5} = 0.04$
- $\widehat{\beta_6} = 1.91$
- $\widehat{\beta_7} = 1.09$
- $\widehat{\sigma_{sch}} = 8.67$
- $\widehat{\sigma_{cl}} = 9.31$
- $\hat{\sigma} = 26.72$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step3

Level 2 공변량 유의성 검정 : t-검정

- "mathknow" 에 27개의 결측치 존재
  - 전체 자료를 사용하여 적합한 모형4.2와 결측치를 제외하고 적합한 모형4.3에 대하여 가능도비 검정을 실시하는 부 적절
  - 각 개별 공변량에 대하여 t-검정 실시

가설의 설정

- 교사교육경험년수(yeartea)에 대한 유의성 검정

$$H_0 : \beta_5 = 0 \quad vs. \quad H_1 : \beta_5 \neq 0$$

- 교사수학지식(mathknow)에 대한 유의성 검정

$$H_0 : \beta_6 = 0 \quad vs. \quad H_1 : \beta_6 \neq 0$$

- 수학교육준비수준(mathprep)에 대한 유의성 검정

$$H_0 : \beta_7 = 0 \quad vs. \quad H_1 : \beta_7 \neq 0$$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step3

Level 2 공변량 유의성 검정 : t-검정

```
> summary(model4.3.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.671285

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.310153 26.71766

Fixed effects:  mathgain ~ sex + minority + mathkind +
                Value Std.Error  DF    t-value p-value
(Intercept) 282.02452 11.701687 792  24.101185  0.0000
sex          -1.33950  1.718580 792  -0.779423  0.4360
minority     -7.86886  2.418081 792  -3.254177  0.0012
mathkind     -0.47501  0.022747 792 -20.882471  0.0000
ses           5.41925  1.275995 792   4.247078  0.0000
yearstea      0.03974  0.117070 177   0.339435  0.7347
mathknow      1.91448  1.147015 177   1.669094  0.0969
mathprep      1.09485  1.148493 177   0.953296  0.3417
```

개별 t-검정 결과

- Level 1 공변량이 모형에 있을 때 Level 2 공변량을 추가하는 것은 통계적으로 유의하지 않음
  - ➢ 모형4.2를 분석모형으로 유지

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step4

모형4.2에 Level 3 공변량 추가한 모형(모형4.4) 적합

- Level 3 공변량 : housepov(x_{8k})을 모형4.2에 추가
- k번째 학교 내 j 번째 학급의 i번째 학생에 대하여

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk}$$
$$+\beta_8 * x_{8k} + u_{j|k} + u_k + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma^2_{cl}),\ u_k \sim^{iid} N(0, \sigma^2_{sch})$$

모형4.4 적합

```
# 모형4: 모형4.2에 Level 3 공변량 추가한 모형: Model 4.4.
model4.4.fit <- update(model4.2.fit, fixed = ~sex+minority+mathkind+ses+housepov)
```

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step4

모형4.2에 Level 3 공변량 추가한 모형(모형4.4) 적합

```
> summary(model4.4.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.818243

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.030822 27.10018

Fixed effects:  mathgain ~ sex + minority + mathkind + ses + housepov
                Value Std.Error  DF    t-value p-value
(Intercept) 285.05800 11.020766 874  25.865534  0.0000
sex          -1.23460  1.657434 874  -0.744884  0.4565
minority     -7.75588  2.384993 874  -3.251950  0.0012
mathkind     -0.47086  0.022281 874 -21.132931  0.0000
ses           5.23971  1.244971 874   4.208703  0.0000
housepov    -11.43923  9.937384 105  -1.151131  0.2523
```

모수 추정치

- $\widehat{\beta_1} = -1.23$
- $\widehat{\beta_2} = -7.76$
- $\widehat{\beta_3} = -0.47$
- $\widehat{\beta_4} = 5.24$
- $\widehat{\beta_8} = -11.44$
- $\widehat{\sigma_{sch}} = 8.82$
- $\widehat{\sigma_{cl}} = 9.03$
- $\hat{\sigma} = 27.10$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step4

Level 3 공변량 유의성 검정 : t-검정 또는 가능도비 검정

◆ **가설의 설정**

$$H_0 : \beta_8 = 0 \quad vs. \quad H_1 : \beta_8 \neq 0$$

귀무가설 모형

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk} + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

대립가설 모형

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk}$$

$$+\beta_8 * x_{8k} + u_{j|k} + u_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step4

Level 3 공변량 유의성 검정 : t-검정

```
> summary(model4.4.fit)
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.818243

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.030822 27.10018

Fixed effects:  mathgain ~ sex + minority + mathkind + ses + housepov
                Value Std.Error  DF    t-value p-value
(Intercept) 285.05800 11.020766 874  25.865534  0.0000
sex          -1.23460  1.657434 874  -0.744884  0.4565
minority     -7.75588  2.384993 874  -3.251950  0.0012
mathkind     -0.47086  0.022281 874 -21.132931  0.0000
ses           5.23971  1.244971 874   4.208703  0.0000
housepov    -11.43923  9.937384 105  -1.151131  0.2523
```

t-검정 결과

- Level 1 공변량이 모형에 있을 때 Level 3 공변량을 추가하는 것은 통계적으로 유의하지 않음
  - ➢ 모형4.2를 최종 분석모형으로 선택

2. Classroom 연구자료 LMM 적합

Three-level 변량절편모형 적합 : Step4

Level 3 공변량 유의성 검정 : 가능도비 검정

- 가능도비 검정통계량 LR과 분포

$$LR = -2log\left(\frac{L_{H_0}}{L_{H_1}}\right) \approx^{H_0} x^2(1)$$

여기서 L_{H_0}과 L_{H_1} 은 각각 귀무가설과 대립가설에서 구한 ML값

유의성 검정과 결과

```
> ## Level 3 공변량 유의성 검정: 가능도비 검정
> # 귀무가설: 모형4.2: ML estimation.
> model4.2.ml.fit <- lme(mathgain~sex+minority+mathkind+ses, random=~1|schoolid/classid,
+                        data=class, na.action="na.omit", method="ML")
> # 대립가설: 모형4.4: ML estimation.
> model4.4.ml.fit <- update(model4.2.ml.fit, fixed = ~sex+minority+mathkind+ses+housepov)
> anova(model4.2.ml.fit, model4.4.ml.fit)
                Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model4.2.ml.fit     1  8 11406.96 11447.62 -5695.481
model4.4.ml.fit     2  9 11407.64 11453.38 -5694.822 1 vs 2 1.318652  0.2508
```

✓ 모형4.2를 최종 분석모형으로 선택

p-value= 0.25

04

제 13강. LMM 3-수준 군집자료분석

LMM : 군집자료 모형적합 결과 해석

1. 최종 분석모형과 적합결과 해석

Classroom 연구자료에 대한 최종분석 모형

분석모형 기술

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk} + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서

$$\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2),\ u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2),\ u_k \sim^{iid} N(0, \sigma_{sch}^2)$$

이고 $\varepsilon_{ijk}, u_{j|k}, u_k$ 는 서로 독립

모형 적합

최종모형 적합

```
# 최종 분석모형: Model 4.2.
model4.2.fit <- lme(mathgain~sex+minority+mathkind+ses, random=~1|schoolid/classid,
                    data=class, na.action="na.omit", method="REML")
```

1. 최종 분석모형과 적합결과 해석

최종 모형 적합 결과

```
> summary(model4.2.fit)
```

고정효과 모형 적합결과

	Value	Std.Error	DF	t-value	p-value
(Intercept)	282.79034	10.853234	874	26.055860	0.0000
sex	-1.25119	1.657730	874	-0.754762	0.4506
minority	-8.26213	2.340113	874	-3.530655	0.0004
mathkind	-0.46980	0.022266	874	-21.099524	0.0000
ses	5.34638	1.241094	874	4.307794	0.0000

공변량 효과 추정치 해석 : 예

- "사회경제적상태(ses)" 추정치 해석
  - "나머지 공변량들(sex, minority, mathkind)이 보정되었을 때 ses 가 1 단위 높을 수록 수학성적성취점수는 평균적으로 5.35 (SE=1.24) 높다 ($p<0.0001$)."
- ✓ 과제: 다른 공변량 효과 추정치에 대한 해석

고정효과 모형의 적합 결과 기술 : $\widehat{E(Y_{ij})}$

$$\widehat{E(Y_{ijk})} = \widehat{\beta_0} + \widehat{\beta_1} * x_{1ijk} + \widehat{\beta_2} * x_{2ijk} + \widehat{\beta_3} * x_{3ijk} + \widehat{\beta_4} * x_{4ijk}$$
$$= 282.79 - 1.25 * x_{1ijk} - 8.26 * x_{2ijk} - 0.47 * x_{3ijk} + 5.35 * x_{4ijk}$$

1. 최종 분석모형과 적합결과 해석

최종 모형 적합 결과

분산 모수$[\sigma^2_{sch}, \sigma^2_{cl}, \sigma^2]$의 추정결과와 해석

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk} + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서 $\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2)$, $u_{j|k} \sim^{iid} N(0, \sigma^2_{cl})$,

$u_k \sim^{iid} N(0, \sigma^2_{sch})$ 이고 ε_{ijk}, $u_{j|k}$, u_k 는 서로 독립

- $\widehat{\sigma^2_{sch}} = (8.67)^2 = 75.17$ (77.44)*
- $\widehat{\sigma^2_{cl}} = (9.13)^2 = 83.36$ (99.20)
- $\widehat{\sigma^2} = (27.10)^2 = 734.41$ (1028.49)
  - * 모형4.1 추정값

분산모수 추정 결과

```
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.671991

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:     9.12604 27.10286
```

- 분산성분모형에 Level 1 공변량을 추가하면
  - ✓ 오차분산은 1028.49 ⇒ 734.41 (29%↓)
  - ✓ 학교 내 학급분산 99.20 ⇒ 83.36 (16%↓)
  - ✓ 학교 분산 77.44 ⇒ 75.17 (3%↓)

감소. 즉 Level 1 공변량, 자료 변동을 효과적으로 설명

1. 최종 분석모형과 적합결과 해석

최종 모형 적합 결과

변량효과($u_{j|k}, u_k$)예측결과와 해석

$$Mathgain_{ijk} = \beta_0 + \beta_1 * x_{1ijk} + \beta_2 * x_{2ijk} + \beta_3 * x_{3ijk} + \beta_4 * x_{4ijk} + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서 $\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2)$, $u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2)$,

$u_k \sim^{iid} N(0, \sigma_{sch}^2)$ 이고 ε_{ijk}, $u_{j|k}$, u_k 는 서로 독립

변량효과 $u_k, u_{j|k}$의 예측치(BLUPs)

Level: schoolid	(Intercept)	Level: classid %in% schoolid	(Intercept)
1	0.49814218	1/160	3.40246448
2	5.60559955	1/217	-2.85079317
3	12.80151595	2/197	-2.98964788
4	-7.51320961	2/211	4.94523708
5	-0.54047662	2/307	4.25237421
6	7.98976163	3/11	0.77710766
7	-6.31932542	3/137	3.78299534
8	3.25485950	3/145	10.44351923
9	-5.06060416	3/228	-0.82648709

$\widehat{u_k}, \widehat{u_{j|k}}$의 해석

- 3^{rd} 학교의 변량효과가 큰 값으로 예측
- 3^{rd} 학교의 145^{th} 학급의 변량효과가 큰값으로 예측

2. Three-level 군집자료의 급내상관계수(ICC) 추정

분산성분모형과 급내상관계수(동질성의 측도) 추정

분산성분모형과 분산모수 추정 그리고 ICC

[Three-level variance components model]

$$Mathgain_{ijk} = \beta_0 + u_{j|k} + u_k + \varepsilon_{ijk}$$

여기서 $\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2)$, $u_{j|k} \sim^{iid} N(0, \sigma_{cl}^2)$,

$u_k \sim^{iid} N(0, \sigma_{sch}^2)$ 이고 ε_{ijk}, $u_{j|k}$, u_k 는 서로 독립

- $\widehat{\sigma_{sch}^2} = (8.80)^2 = 77.44$
- $\widehat{\sigma_{cl}^2} = (9.96)^2 = 99.20$
- $\widehat{\sigma^2} = (32.07)^2 = 1028.49$

분산모수 추정 결과

```
Random effects:
 Formula: ~1 | schoolid
        (Intercept)
StdDev:    8.802955

 Formula: ~1 | classid %in% schoolid
        (Intercept) Residual
StdDev:    9.961301 32.06609
```

급내상관계수: ICC

- 학교의 ICC
  - ✓ $\widehat{ICC}_{sch} = \frac{\hat{\sigma}_{sch}^2}{\hat{\sigma}_{sch}^2 + \hat{\sigma}_{cl}^2 + \hat{\sigma}^2} = \frac{77.44}{77.44+99.20+1028.49} = \mathbf{0.06}$
- 학교 내 학급의 ICC
  - ✓ $\widehat{ICC}_{sch} = \frac{\hat{\sigma}_{sch}^2 + \hat{\sigma}_{cl}^2}{\hat{\sigma}_{sch}^2 + \hat{\sigma}_{cl}^2 + \hat{\sigma}^2} = \frac{77.44+99.20}{77.44+99.20+1028.49} = \mathbf{0.15}$

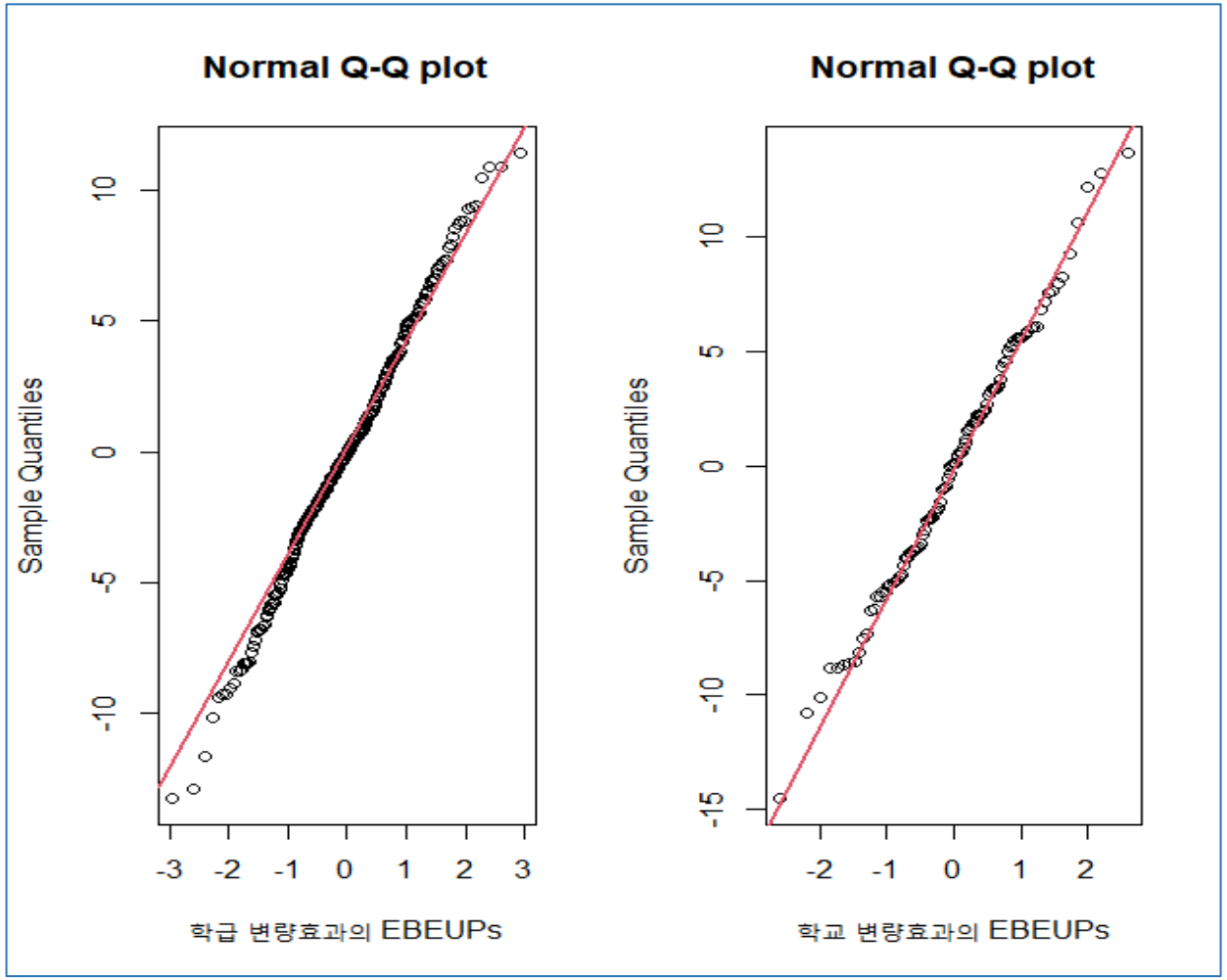
05

제 13강. LMM 3-수준 군집자료분석

LMM : 모형진단

1. EBLUPs 그림 : 변량효과의 분포 가정 진단

학급 및 학교 변량효과의 EBLUPs 그림



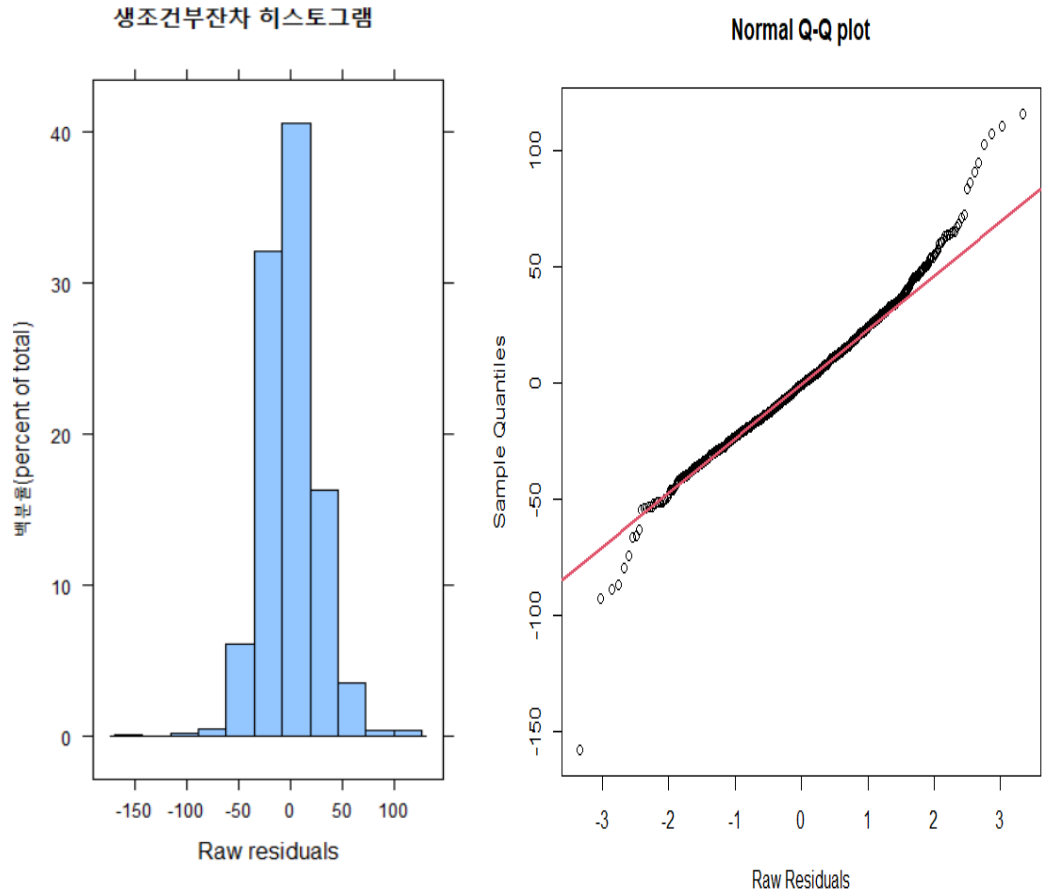
➢ 정규분포로부터 이탈하는 뚜렷한 징후는 보이지 않음

```
par(mfrow=c(1,2))
# 학급에 대한 EBEUPs Q-Q plot
cl <- ranef(model4.2.fit,level=2)
qqnorm(cl[,1], main="Normal Q-Q plot",xlab="학급 변량효과의 EBEUPs")
qqline(cl[,1], col=2, lwd=2, lty=1)

# 학교에 대한 EBEUPs Q-Q plot
sch <- ranef(model4.2.fit,level=1)
qqnorm(sch[,1], main="Normal Q-Q plot",xlab="학교 변량효과의 EBEUPs")
qqline(sch[,1], col=2, lwd=2, lty=1)
```

2. 잔차 분석 : 조건부 생잔차 그림

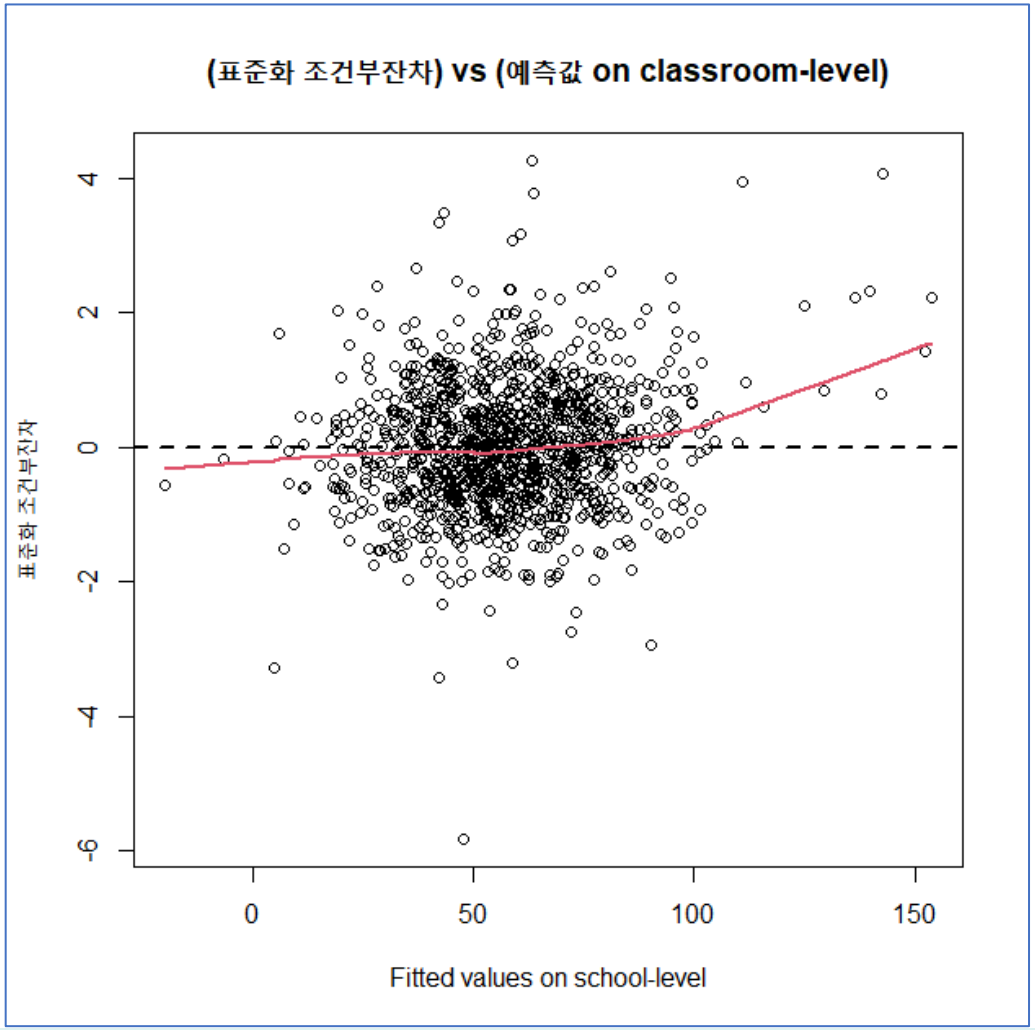
Histogram: ε_{ij}의 정규성 검토 & 이상치 탐색



- 양쪽(특히 왼쪽) 긴 꼬리를 가지는 분포
  - 정규분포에서의 이탈 시사
  - 이들에 대한 추가적인 탐구 시사
    - ✓ 변수변환(?)
    - ✓ 변량요인(예: 변량계수 등) 추가 검토
    - ✓ 오차분산구조(예: 이분산 등) 탐구

2. 잔차 분석 : 예측값 vs. 잔차

예측값 vs. 잔차 산점도 : ε_{ijk}의 등분산성 검토



❖ Loess 곡선 큰 예측값에서 증가하는 경향

➢ 큰(작은) 예측값을 설명할 수 있는 공변량 또는 변량요인의 결여 시사

- 각 공변량 별로 산점도 탐구
  - ✓ 비선형항 검토
  - ✓ 변량계수모형 검토

2. 잔차 분석

R code

```
## 조건부 잔차진단
par(mfrow=c(1,1))
# 생조건부잔차 histogram
rrsid <- data.frame(resid(model4.2.fit))
histogram(~rrsid[,1],aspect=2, xlab="Raw residuals", main="생조건부잔차 히스토그램")

# 생조건부잔차 Q-Q plot
qqnorm(rrsid[,1], main="Normal Q-Q plot", xlab="Raw Residuals")
qqline(rrsid[,1], col=2, lwd=2, lty=1)

# (표준화 조건부잔차) vs (예측값 on classroom-level)
plot(resid(model4.2.fit, type="normalized") ~ fitted(model4.2.fit,level=2),
     main="(표준화 조건부잔차) vs (예측값 on classroom-level)",
     xlab="Fitted values on school-level", ylab="표준화 조건부잔차")
abline(h = 0, lty = 2, lwd=2)
lines(lowess(resid(model4.2.fit, type="normalized") ~ fitted(model4.2.fit,level=2)),
     col=2, lwd=2)
```

14 강

다음시간안내

경시적자료분석
LMM 변량계수모형

수고하셨습니다.