14강. 경시적자료분석: LMM 변량계수 모형

■ 주요용어

용어	해설
변량계수모형	회귀계수가 개체에 따라 변화하는 경우, 이를 모형화 는 한 방법은 회귀계수에 개체효과를 더하여 모형화 하는 것. 예를 들면 아래와 같은 모형을 말한다. Y_{ij}: j번째 개체의 i번째 관측값 x_{ij}: 공변량 $Y_{ij}=(\beta_0+u_{0j})+(\beta_1+u_{1j})x_{ij}+\epsilon_{ij}$ 여기서 $U_j=(u_{0j},\ u_{1j})'\sim^{iid} N(0,D)$, $\epsilon_{ij}\sim^{iid}N(0,\sigma^2)$이고 서로 독립. $D=\begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}$.
변량효과의 유의성 검정	두 개 변향효과 u_{0j}, u_{1j} 으로 구성된 선형혼합모형에서 아래의 가설을 검정한다 귀무가설(H_0): 변량효과 u_{1j}은 유의하지 않다. 대립가설(H_1): 변량효과 u_{1j}은 유의하다. 검정통계량: 제한가능도비검정통계량(LR) $LR=-2\log\left(\frac{L_{H_0}}{L_{H_1}}\right)$, 여기서 L_{H_0}와 L_{H_1}는 각각 귀무가설과 대립가설에서의 제한최대가능도 값이다. 검정통계량의 분포는 $LR\sim^{H}_{0}\ 0.5\chi^2(1)+0.5\chi^2(2)$.
가능도비검정통계량 (LR)	$LR=-2\log\left(\frac{L_{H_0}}{L_{H_1}}\right)$, 여기서 L_{H_0}와 L_{H_1}는 각각 귀무가설과 대립가설에서의 최대가능도 값이다. 단 $H_0\subset H_1$성립한다. 고정효과의 유의성을 검정할 때 사용한다. 검정통계량의 분포는 $LR\sim^{H}_{0}\ \chi^2(df)$ 이며 df는 카이제곱 분포의 자유도로 $df=(H_1$에서 추정된 모수의 수$)-(H_0$에서 추정된 모수의 수$)$.

정리하기

■ 요약하기

1. 변량계수 모형이란?

1) 회귀계수가 개체에 따라 변화하는 경우, 이를 모형화 는 한 방법은 회귀계수에 개체효과를 더하여 모형화 하는 것. 예를 들면 아래와 같은 모형을 말한다.

Y_{ij}: j번째 개체의 i번째 관측값

X_{ij}: 공변량

$Y_{ij}=(\beta_0+u_{0j})+(\beta_1+u_{1j})x_{ij}+\epsilon_{ij}$

여기서 $U_j=\ (u_{0j},\ u_{1j})' \sim^{iid}\ N(0,D),\ D=\begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix},\ \epsilon_{ij}\sim^{iid}N(0,\sigma^2)$이고 U_j와 ϵ_{ij} 서로 독립.

2) 변량계수모형의 특성

① 모평균: $E(Y_{ij})=\beta_0+\beta_1x_{ij}$

② 조건부 평균: $E(Y_{ij}|U_j)=\beta_0+u_{0j}+(\beta_1+u_{1j})x_{ij}$

③ 분산: $Var(Y_{ij})=Var(u_{0j})+x_{ij}^2Var(u_{1j})+Cov(u_{0j},u_{1j})+Var(\epsilon_{ij})$
$=\sigma_0^2+x_{ij}^2\sigma_1^2+2*x_{ij}\sigma_{01}+\sigma^2$

2. 자폐아 자료(Autism data, Oti et al. 2006)

1) 연구목적

- 자폐아에서 초기 소통능력이 성장하면서 사회화정도에 어떻게 영향을 주는가?

2) 자료수집

자폐(ASD: Autism Spectrum Disorder) 또는 전반적발달장애(PDD: Pervasive Developmental Disorder)가 있는 158명의 어린 아이를 관찰한 연구. 각 어린이는 두 살 때 소통발달정도 (1=low, 2=medium, 3=high)을 측정. 2, 3, 5, 9, 13세 때에 사회화의 정도를 관측

3) 관측 변수

- 개체 간 특성 변수[Subject(Level 2) variables]

① childid: 아이의 고유 번호

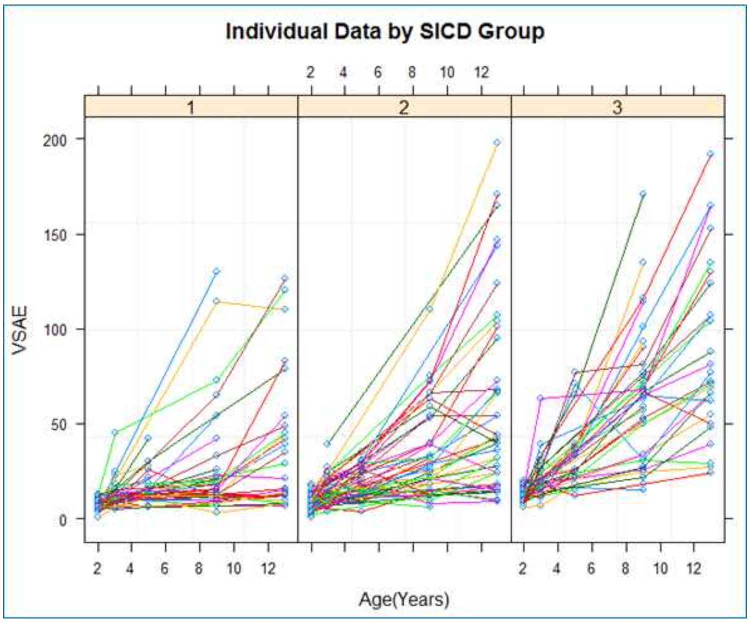
② sicdegp: 2세 때의 소통발달정도(Sequenced Inventory of Communication Development Expressive Group: 1=low, 2=medium, 3=high)

- 개체 내 특성 변수[Time-varing(Level 1) variables]

③ age: 관측 나이(2, 3, 5, 9, 13)

④ vsae: 사회화정도(Vineland Socialization Age Equivalent: patient reported socialization)

4) 각 아이의 나이에 따른 vsae 점수 분포



3. 자폐아 자료분석을 위한 변량계수 모형

1) 모형구축 전략

- 하향식 모형구축: 자료탐색에서 발견된 자료의 특성을 잘 반영할 것으로 예상되는 충분히 복잡한 모형에서 출발

2) 평균모형의 설정

- sicdegp의 수준 2에서 age의 2차 곡선 경향성이 있는 듯이 보임. 수준1과 2에서는 직선 경향성 보임.

① age^2항과 sicdegp 항과 교호작용 항을 포함시킴. 즉, age, age^2, sicdegp, sicdegp*age, sicdegp*,age^2 항 평균모형에 포함시킴.

3) 변량모형의 설정

- age가 증가함에 따라 vsae가 증가하는 형태는 각 아이에서 다르게 나타남. age가 증가함에 따라 vsae의 분산이 증가하고 있음.

① 절편, age, age^2 항의 회귀계수에 변량효과를 추가.

4) 최초 설정한 변량계수 모형

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})age.2_{ij} + (\beta_2 + u_{2j})age.2_{ij}^2 + \beta_3 sic2_j + \beta_4 sic3_j$$
$$+ \beta_5 age.2_{ij} \cdot sic2_j + \beta_6 age.2_{ij} \cdot sic3_j$$
$$+ \beta_7 age.2_{ij}^2 \cdot sic2_j + \beta_8 age.2_{ij}^2 \cdot sic3_j + \epsilon_{ij}$$

여기서 $age.2_{ij} = age_{ij} - 2$, $sic2_j = \begin{cases} 1 & \text{if } sicdegp = 2 \\ 0 & otherwise \end{cases}$, $sic3_j = \begin{cases} 1 & \text{if } sicdegp = 3 \\ 0 & otherwise \end{cases}$

$U_j = (u_{0j},\ u_{1j}, u_{2j})' \sim^{iid} N(0, D)$, $D = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}$, $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$이고 U_j와 ϵ_{ij}서로 독립.

4. 모형구축

1) 변량효과의 구조 선택

- 최초 설정한 모형에 모수 추정치 계산에 문제가 발생
- 변량절편 u_{0j}를 최초 모형에서 제거한 후 모형 적합.
- $age.2_{ij}^2$ 항의 변량계수 u_{2j}에 대한 유의성 검정 결과 유의하여, 최종적으로 $age.2_{ij}$항의 변량계수 u_{1j}와 $age.2_{ij}^2$ 항의 변량계수 u_{2j}가 있는 변량계수 모형을 선택.

2) 고정효과의 모형 선택

- 최초 설정한 고정효과 모형에서 가장 고차항인 sicdegp*,age^2 항에 대한 유의성 결과 유의하지 않아 모형에서 제거.
- sicdegp*age 항은 유의하여 모형에 포함됨.

3) 최종분석 모형

$$Y_{ij} = \beta_0 + (\beta_1 + u_{1j})age.2_{ij} + (\beta_2 + u_{2j})age.2_{ij}^2 + \beta_3 sic2_j + \beta_4 sic3_j$$
$$+ \beta_5 age.2_{ij} \cdot sic2_j + \beta_6 age.2_{ij} \cdot sic3_j + \epsilon_{ij}$$

여기서 $age.2_{ij} = age_{ij} - 2$, $sic2_j = \begin{cases} 1 & \text{if } sicdegp = 2 \\ 0 & otherwise \end{cases}$, $sic3_j = \begin{cases} 1 & \text{if } sicdegp = 3 \\ 0 & otherwise \end{cases}$

$U_j = (u_{1j}, u_{2j})' \sim^{iid} N(0, D)$, $D = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$이고 U_j와 ϵ_{ij}서로 독립.

5. 모수추정치의 해석

1) 모형 적합결과

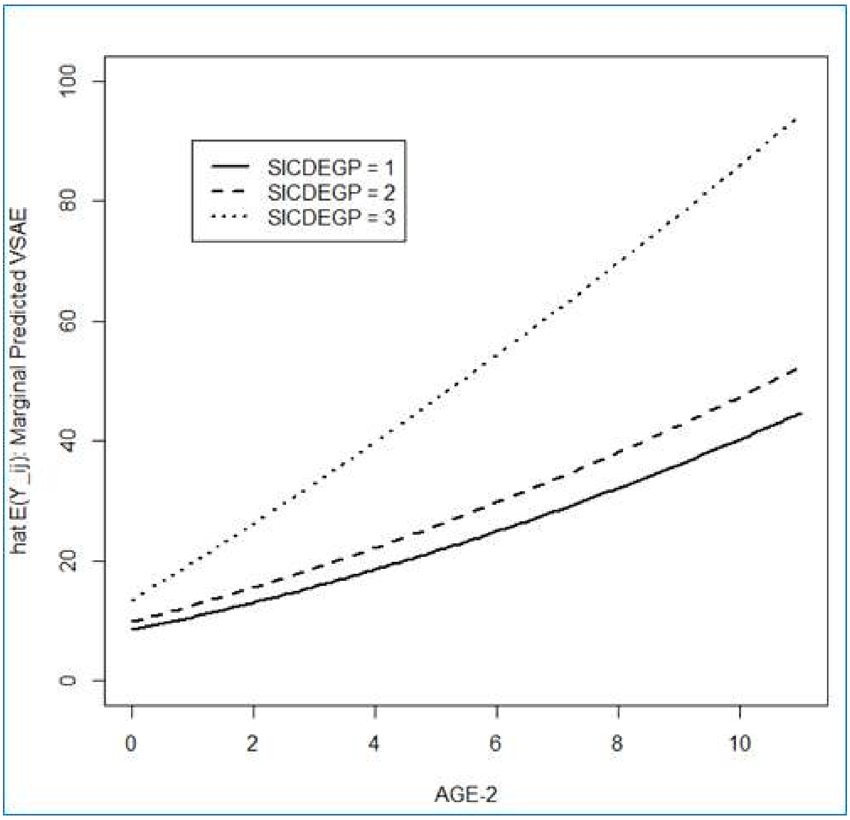
```
> summary(model.3.fit)
Linear mixed-effects model fit by REML
 Data: autism.grouped
      AIC      BIC    logLik
  4633.57 4681.991 -2305.785

Random effects:
 Formula: ~age.2 + I(age.2^2) - 1 | childid
 Structure: General positive-definite, Log-Cholesky parametrization
           StdDev    Corr
age.2      3.8110274 age.2
I(age.2^2) 0.3556805 -0.306
Residual   6.2281389

Fixed effects: vsae ~ age.2 + I(age.2^2) + sicdegp.f + age.2:sicdegp.f
                    Value Std.Error  DF   t-value p-value
(Intercept)      8.475894 0.7094431 448 11.947249  0.0000
age.2            2.080709 0.6482319 448  3.209822  0.0014
I(age.2^2)       0.109008 0.0427795 448  2.548125  0.0112
sicdegp.f2       1.364819 0.9215857 155  1.480946  0.1407
sicdegp.f3       4.987639 1.0379064 155  4.805480  0.0000
age.2:sicdegp.f2 0.572512 0.7960151 448  0.719222  0.4724
age.2:sicdegp.f3 4.068041 0.8797676 448  4.623995  0.0000
```

2) 고정효과 모형의 추정식

구분	age=2	age=x
sicdegp=1	$\hat{\beta}_0 = 8.48$	$\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = 8.48 + 2.08x + 0.11x^2$
sicdegp=2	$\hat{\beta}_0 + \hat{\beta}_3 = 8.48 + 1.36$	$\hat{\beta}_0 + \hat{\beta}_3 + (\hat{\beta}_1 + \hat{\beta}_5)x + \hat{\beta}_2 x^2$ $= 8.48 + 1.36 + (2.08 + 0.57)x + 0.11x^2$
sicdegp=3	$\hat{\beta}_0 + \hat{\beta}_4 = 8.48 + 4.99$	$\hat{\beta}_0 + \hat{\beta}_4 + (\hat{\beta}_1 + \hat{\beta}_6)x + \hat{\beta}_2 x^2$ $= 8.48 + 4.99 + (2.08 + 4.07)x + 0.11x^2$



3) 주변부 공분산 $V = Var(Y_{ij})$의 추정

```
> ## 주변부 공분산 행렬 추정치
> getVarCov(model.3.fit, individual="1", type="marginal")
childid 1
Marginal variance covariance matrix
      1       2       3        4       5
1 38.79   0.000   0.000    0.000    0.00
2  0.00  52.610  39.728   84.617  120.27
3  0.00  39.728 157.330  273.610  425.25
4  0.00  84.617 273.610  769.400 1293.00
5  0.00 120.270 425.250 1293.000 2543.20
```

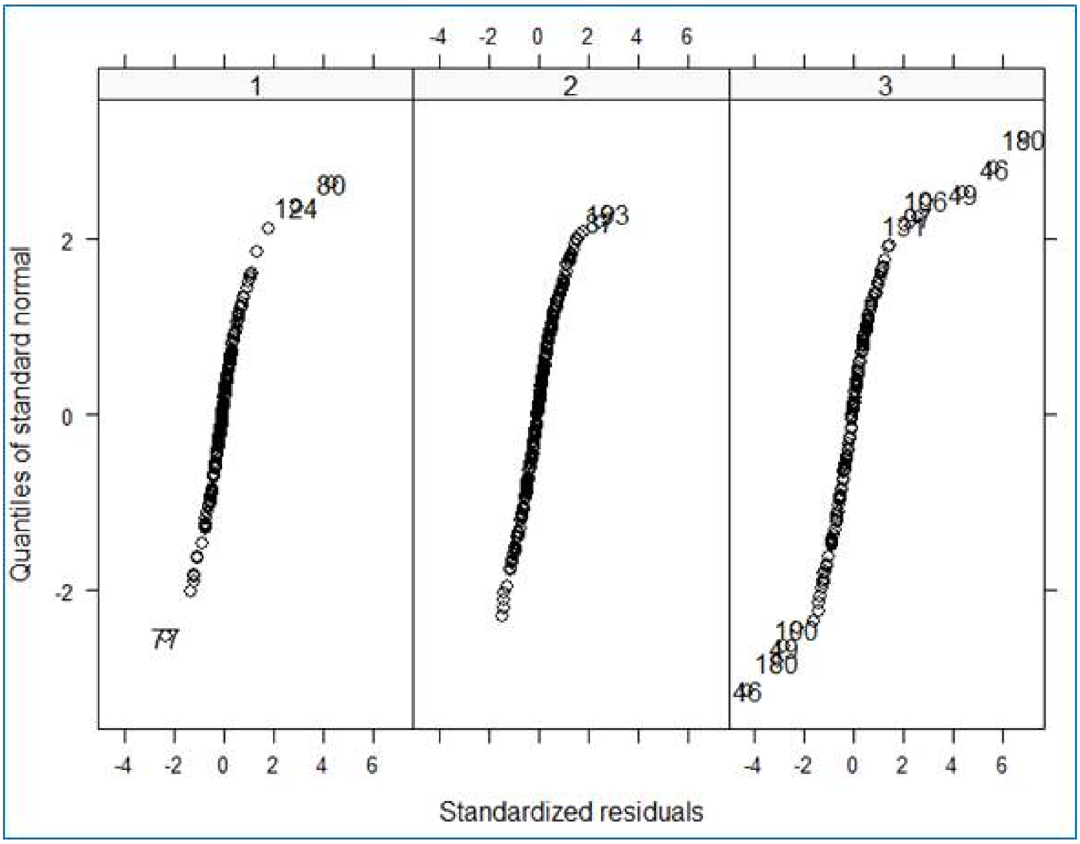
- 해석

① 현재 최종 모형에서는 2세 때($age.2_{ij} = 0$) vsae 값과 나머지 나이 때의 vsae 값 사이에는 상관이 없다고 가정한다(이는 변량절편 u_{0j}가 모형에서 제거되었기 때문. 주변부 공분산에 대한 gls 모형으로 추가 분석해 보는 것도 의미 있음).

② 나이가 증가할 수록 분산이 급속히 증가한다. 이는 자료에서 나타난 현상을 타당하게 모형화 한 것으로 볼 수 있음.

6. 모형진단

1) 잔차진단



- 해석

① 오차에 대한 등분산성과 정규성 가정이 타당한지에 대한 의심이 생김. 오른쪽으로 긴 꼬리를 가지는 분포형태가 나타남

2) 새로운 분석 시도 방향

① 오차분산에 대한 이분산성 (heteroscedasticity) 가정 또는,

② Y_{ij} 에 대한 새로운 분포 가정(예: Gamma 분포) 또는

③ Y_{ij} 에 대한 변수 변환(예: log 변환)

과제하기

구분	내용
과제 주제	1. Y_{ij}: j번째 개체의 i번째 관측값, x_{ij}: 공변량 일 때 $Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{ij} + \epsilon_{ij}$ 이고, 여기서 $U_j = (u_{0j},\ u_{1j})' \sim^{iid} N(0,D)$, $D = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$, $\epsilon_{ij} \sim {}^{iid}N(0,\sigma^2)$ 이며, U_j와 ϵ_{ij}서로 독립이다. 다음을 구하라. ① $E(Y_{ij})$ ② $E(Y_{ij} \mid U_j)$ ③ $Var(Y_{ij})$ ④ $Cov(Y_{ij}, Y_{i'j})$ 단, $i \neq i'$ ⑤ $Cov(Y_{ij}, Y_{i'j'})$ 단, $i \neq i'$ $j \neq j'$ 2. 최종분석모형(Model3) 적합결과로부터 ① 아래 각 가설 $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$, $i = 1,2,\cdots,6$ 에 대하여 유의수준 5%에서 t-검정을 실시하시오. ② β_i $i = 1,2,\cdots,6$들의 의미하는 바를 기술하고 이들에 대한 95% 신뢰구간을 제시하시오.
목적	14주차 강의 내용을 복습하고, 변량계수모형의 구축과 모형의 선택에 대한 이해를 심화 하기 위함. 또한 모수 추정치의 의미와 설명법에 대한 이해도를 높이기 위함.
제출 기간	14주차 강의 후 1주 후 토요일 밤 10시까지
참고 자료	교재와 강의자료를 참고하기 바람
기타 유의사항	