Covariance Pattern Model

Reference

- 연세대학교 강승호 교수님의 일반화 혼합 모형 수업.
- Applie Mixed Models in Medicine, Chapter 6
- Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, Chapter 7
- https://www.rdocumentation.org/packages/nlme/versions/3.1-144/topics/gls

Covariance pattern model

이름은 거창하지만 말 그대로 covariance을 모델링하는 방법이다. 하지만 이때 random effect을 지정하지는 않는다. 간단한 toy example로 아래의 데이터를 생각해보자.

patient	visit
1	1
1	2
1	3
2	1
2	2
2	3
2	4
3	1
3	2

표 1

위 표는 세 환자가 각각 3번, 4번, 2번 반복측정한 데이터의 한 예시이다. 경시적 자료에서는 기본적으로 각 subject간에는 독립, subject 내의 반복 측정 자료에 대해서는 correlate 되어 있다고 가정한다. 즉, 표 1에서 patient 1의 세 개의 반복 측정은 서로 correlate 되어 있지만 patient 1과 patient 2의 반복 측정은 서로 독립이라고 가정하는 것이다. 이는 reasonable한 가정이고 이러한 가정을 통해서 현상을 최대한 간단하게 설명하는 것이 바로 통계학의 목적이기도 하다.

표 1에는 총 9개의 데이터가 있는데, 경시적 자료 하에서 이들의 covariance structure을 살펴보자.

$$\begin{bmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21} & \sigma_2^2 & \theta_{23} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{13} & \theta_{23} & \sigma_3^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} & 0 & 0 \\ 0 & 0 & 0 & \theta_{21} & \sigma_2^2 & \theta_{23} & \theta_{24} & 0 & 0 \\ 0 & 0 & 0 & \theta_{13} & \theta_{23} & \sigma_3^2 & \theta_{34} & 0 & 0 \\ 0 & 0 & 0 & \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_1^2 & \theta_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{21} & \sigma_2^2 \end{bmatrix} := \begin{bmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ 0 & 0 & R_3 \end{bmatrix} \tag{1}$$

가장 일반적인 covariance structure이다. R_i는 i번째 환자의 block covariance이다. 이들의 off diagonal은 0으로, 서로 독립임을 가정한다. R_i을 자세하게 살펴보자. 이들의 off diagonal은 0이 아니므로 서로 correlate되어 있음을 가정한다. R_i의 차원은 i번째 환자의 반복 측정 횟수에 의존한다. 즉, patient 1의 block covariance는 R_1으로, 3×3이다.

선형 회귀에서 covariance structure와 비교해보자. 선형 회귀에서는 등분산을 가정하고 관측치들 간의 독립을 가정한다. 즉, 아래의 covariance structure을 가정하는 것이다.

$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} := \begin{bmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ 0 & 0 & R_3 \end{bmatrix} \tag{2}$$

경시적 자료에 대해서 (2)의 가정을 하는 것은 무리가 있다. 경시적 자료에서 한 subject 내의 자료들은 서로 correlate 될 수밖에 없다. 예를 들어, 한 학생의 여러 학기 성적은 독립이 아니라고 가정하는 것이 reasonable하다.

(1)의 covariance는 가장 일반적인 상황을 가정한다. 즉, 각 시점별로 서로 correlate 된 정도가 다 다르다는 것이다. 예를 들어 첫 번째 관측치와 두 번째 관측치, 세 번째 관측치가가 correlate 된 정도를 나타내는 모수 θ_{12}, θ_{13}는 서로 다르다. 물론 이렇게 가정하면 다양한 상황을 대비한다는 점에서 좋겠지만 모수가 늘어난다는 단점이 있다. 통계학에서는 모수를 최대한 줄여서 simple한 모델을 가정하고, 이 모델에 데이터를 적합하여 model fit이 적절한지를 본다. gerneral form 이외에 다양한 covariance structure을 살펴보자.

Simple Covariance Patterns

1. General

$$R_i = \begin{bmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} \\ \theta_{21} & \sigma_2^2 & \theta_{23} \\ \theta_{13} & \theta_{23} & \sigma_3^2 \end{bmatrix}$$

   앞서 살펴본 가장 일반적인 형태이다.

2. First order autoregressive

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

   시점이 늘어날 때마다 correlate되는 정도가 ρ만큼 곱해지는데, 이 값은 -1에서 1 사이이므로 시점이 멀어질수록 correlate되는 정도가 약해진다는 의미이다. 즉, 첫 번째 시점과 두 번째 시점간의 correlation은 첫 번째 시점과 세 번째 시점간의 그것보다 더 약함을 가정한다.

3. Compound symmetry

$$R_i = \begin{bmatrix} \sigma^2 & \theta & \theta \\ \theta & \sigma^2 & \theta \\ \theta & \theta & \sigma^2 \end{bmatrix}$$

   (ii)의 가정과는 달리 시점이 멀어져도 동일한 correlation을 가정한다.

4. Toeplitz

$$R_i = \begin{bmatrix} \sigma^2 & \theta_1 & \theta_2 \\ \theta_1 & \sigma^2 & \theta_1 \\ \theta_2 & \theta_1 & \sigma^2 \end{bmatrix}$$

   동일한 시점이 떨어져있다면 동일한 correlation을 가정한다. 예를 들어 첫 번째, 두 번째 시점간의 correlation과 두 번째, 세 번째 correlation은 동일하다고 본다.

(ii)에서 (iv)의 diagonal은 모두 σ^2로 동일하다. 그런데 가끔 몇 번째 측정했는지에 따라서 variability가 달라질 수 있다. 즉, diagonal을 모두 σ^2으로 통일하는 것이 적절하지 않을 수도 있다. 그럴 때는 diagonal을 σ_i^2로 다르게 지정하여 Heterogeneity을 가정할 수도 있다.

Separate covariance patterns for each treatment group

가끔 서로 다른 treatment group의 관측치들은 다른 covariance structure을 가지는 경우가 있다. 예를 들어서, 신약을 투여한 그룹의 변수가 위약을 투여한 그룹보다 더 active할 수도 있다. 이러한 경우에, treatment group별로 서로 다른 covariance structure을 지정한다.

예를 들어 표 1의 세 환자가 각각 treatment A, B, A를 받았다고 해보자. 이런 상황에서 separate

compund symmetry는 아래와 같다.

$$\begin{bmatrix} \sigma_A^2 & \theta_A & \theta_A & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_A & \sigma_A^2 & \theta_A & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_A & \theta_A & \sigma_A^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_B^2 & \theta_B & \theta_B & \theta_B & 0 & 0 \\ 0 & 0 & 0 & \theta_B & \sigma_B^2 & \theta_{23} & \theta_B & 0 & 0 \\ 0 & 0 & 0 & \theta_B & \theta_B & \sigma_B^2 & \theta_B & 0 & 0 \\ 0 & 0 & 0 & \theta_B & \theta_B & \theta_B & \sigma_B^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_A^2 & \theta_A \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_A & \sigma_A^2 \end{bmatrix} := R$$

Banded covariances

시점이 먼 관측치들 사이에 correlation이 작을 수도 있다. 특히, 반복 측정을 여러번 실행한 실험에서 먼 시점 간의 correlation을 가정하는 것은 fit에는 영향을 주지 않고 모수만 늘릴 위험이 있다. 이럴 때, 시점이 먼 관측치들 사이의 correlation은 0으로 가정하는 것이 Banded covariances이다.

$$R_i = \begin{bmatrix} \sigma_1^2 & \theta_{12} & 0 \\ \theta_{21} & \sigma_2^2 & \theta_{23} \\ 0 & \theta_{23} & \sigma_3^2 \end{bmatrix}$$

Choice of covariance pattern

그러면 위와 같이 선택지가 꽤 많은데, 그 중에 어떤 covariance pattern을 선택해야 할까? 이 질문은 모수를 최소로 하며 fit을 최대로 해주는, 이상적인 covariance pattern이 무엇인지 물어보는 것과 동일하다. 통계학에서는 모수를 고려하여 fit을 비교하는 개념으로 model fit과 likelihood ratio tests가 있다.

$$AIC = log(L) - q$$

$$SIC = log(L) - (qlog(N - p))/2$$

주의 할점은 likelihood ratio test을 쓰기 위해, 두 모델은 nested 관계에 있어야 한다는 점이다.

$$2(log(L_1) - log(L_2)) \sim \chi^2_{DF}$$

즉, L_1을 계산한 모델이 L_2를 계산한 모델을 nest해야 한다.

Which covariance patterns to consider?

책에서 제시하는 전략으로, 처음에는 간단한 covariance pattern부터 적합해보라고 조언한다. 그리고 점점 복잡한 pattern을 적합하며 fit을 비교해보는 것이다. 그리고 반복 측정 갯수가 그렇게 많지 않으면 사실 compund symmetry에서 대부분 좋은 fit을 낸다고 하니 참고하면 될 것 같다.

GLS and Covariance pattern model

GLS는 generalized least squares의 약자로, OLS처럼 등분산이 아닌, 일반적인 covariance을 가정한다. 근데 만약 그 형태가, 위에서 살펴본 covariance 형태, 즉 subject 간에는 독립이고 subject 내에는 correlation을 가정한다면 바로 Covariance pattern model이 되는 것이다. 즉, GLS가 Covariance pattern model을 포함한다고 볼 수 있다. 따라서 Covariance pattern model을 적합하는 것은, GLS에서 원하는 covariance structure로 바꿔주면 된다. R에서도 이와 동일하게 수행한다.

Data Analysis in R

위에서 살펴본 것처럼 covariance pattern model은 gls의 nested model이므로 R에서 gls 함수를 통해 수행한다. R의 gls 함수에서 correlation 인자를 변화시키며 covariance pattern model을 적합하면 된다. R에서는 아래의 covariance structure을 제공한다.

1. corAR1: autoregressive process of order 1
2. corCompSymm: compund symmetry structure corresponding to a constant correlation
3. corSymm: general correlation matrix, with no dditional structure

```
library(nlme)
opposites<-read.table("https://stats.idre.ucla.edu/stat/r/examples/alda/data/opposites_pp.txt",hea
result_ar1 = gls(opp~time*ccog, opposites, correlation=corAR1(0.2, form = ~1 | id))
result_comsym = gls(opp~time*ccog, opposites, correlation=corCompSymm(0.2, form = ~1 | id))
result_sym = gls(opp~time*ccog, opposites, correlation=corSymm(form = ~1 | id))
```

AIC, log likelihood, 사용된 모수를 정리하면 아래와 같다.

	ar1	comsym	sym	OLS
AIC	1277	1299	1278	1391
BIC	1295	1316	1310	1405
covariance의 모수 갯수	2개	2개	10개	1개
log likelihood	-632	-643	-628	-690

표 2

ar1과 comsym은 모수의 갯수가 같기 때문에 log likelihood만 비교하면 ar1이 더 좋은 모형이다. sym과 likelihood ratio test를 해보자.

$$2(log_{sym} - log_{ar1}) = 8 < \chi^2_{10-2} = 15$$

즉 유의하지 않으므로 귀무가설인 ar1 모형을 채택한다. 추가로 covariance structure을 지정하지 않고, 등분산을 가정한 OLS 모형도 적합하였다. covariance pattern model과 LRT을 해보면, 모두

귀무가설을 기각한다. 즉, reduced model인 OLS 모형을 기각한다는 것이다. 다시 말해서 경시적 자료에 대해 covariance pattern model을 사용하는 것이 적절함을 알 수 있다.
ar1 모형 결과를 살펴보자.

	value	Std.Eror	p-value
time	27.19	1.8	0
ccog	-0.03	0.48	0.9455
time:ccog	0.41	0.15	0.0065

표 3

표 3로부터 time과 time & ccog의 interaction term이 유의함을 알 수 있다. 즉, ccog에 따라서 time이 종속변수 opp에 미치는 영향은 달라지고, ccog 변수는 opp에 유의한 영향을 미치지 않는다고 해석할 수 있다.

그런데 표 2를 보면 log likelihood와 AIC, BIC의 방향성이 상충한다. log likelihood 기준으로는 ar1이 가장 좋지만 AIC, BIC 기준으로는 comsym이 좋기 때문이다. 이에 대해서는 관련 내용을 더 찾아봐야 할 것 같다.

위의 모형은 subject 별로 Homogeneous variance 가정을 한 것인데, Heterogeneous variance 가정을 할 수도 있다. 이는 gls 함수의 weights 옵션을 통해 할 수 있다.