반복측정자료분석

-일반화선형혼합모형-

2020021218 이소연
2020020316 황서진

고려대학교 통계학과

2020.04.13

Contents

Introduction I

경시적 자료에 대한 전통적인 ANOVA 방법의 한계

- 모든 개체들이 같은 시점과 같은 반복 횟수를 가정한다.
- 시간 가변 공변량(time-varying covariate)의 적용에 한계를 가진다.
- 결측자료의 효과를 반영하는 방법이 많지 않다.

$\Longrightarrow$ 보다 일반화된 반복측정 자료 분석 모형이 필요하다.

Introduction II

선형모형의 확장

종속변수가 정규분포를 따르지 않는다면?

$\Longrightarrow$ **Generalized Linear Model**(GLM)

관측치 간 상관관계가 있다면?

$\Longrightarrow$ **Linear Mixed Model**(LMM)

관측치가 정규성을 만족하지도 않고, 상관관계까지 있다면?

$\Longrightarrow$ **Generalized Linear Mixed Model**(GLMM)

Defining GLMM

- Linear Mixed Model: LMM

선형 혼합효과모형(LMM)은 $y_1, ..., y_n$ 에 대하여 다음과 같이 표현된다.

$$y_i = X_i\beta + Z_iU_i + \epsilon_i$$

X와 Z는 각 고정 효과, 랜덤 효과와 관련된 공변량을 나타내며, β는 고정 효과, U는 랜덤 효과를 나타낸다. ϵ은 오차항을 의미한다.

이들은 다음과 같은 가정을 만족한다.

$$E(U_i) = 0, Var(U_i) = G,$$
$$E(\epsilon_i) = 0, Var(\epsilon_i) = R_i, Cov(U_i, \epsilon_i) = 0$$

$\Rightarrow$ 선형모형 하에서 경시적 자료의 모형화를 위해선 적절한 G와 R에 대한 적용과 **오차항의 공분산구조의 설정**이 중요하다.

e.g., 균일상관모형, AR(1), ...

Introduction IV

Defining GLMM

- Generalized Linear Mixed Model: GLMM

  선형혼합모형을 이항, 순서형, 이산형 경시적 자료로 확장한다.

  - **GLM**과의 차이: 랜덤효과를 추가, 오차항의 공분산구조를 통해 반복치들 간의 연관관계 모형화
  - **GEE**와의 차이: 우도함수의 적용 여부와 추정된 회귀계수의 해석

Defining GLMM

- Generalized Linear Mixed Model: GLMM

  n명의 독립적인 관측개체의 반응변수가 $Y = (y_1, ..., y_n)$라고 할 때, 관측개체별 공유되는 특성을 개체별 랜덤 효과 U_i로 표현한다. 이때, U_i가 주어져 있다면, 각 개체 내 관측치 $y_i = (y_{i1}, ..., y_{im_i})$는 서로 **독립** 이라고 가정한다.
  즉, **조건부 독립(conditional independence)**를 가정한다.

  각 관측치의 평균을 $E(y_{ij}|U_i) = \mu_{ij}$ 라고 할 때, **연결함수**(link function) g에 의해 다음과 같이 표현될 수 있다.

$$g(\mu_{ij}) = X_{ij}\beta + Z_{ij}U_i$$

  여기서 X_{ij}, Z_{ij}는 각 고정 효과 및 랜덤 효과와 관련된 공변량을 의미하며, 일반적으로 랜덤 효과 U_i는 평균벡터가 0이고 공분산행렬 $G(\theta)$를 가정한다. 가장 일반적인 선택은 정규분포이다.

Defining GLMM

- Generalized Linear Mixed Model: GLMM

자료의 타입에 따라 다음과 같은 예시를 들 수 있다.

- **이항 경시적 자료**(binary): $Y_{ij}|U_i \sim B(1, \mu_{ij}), logit\mu_{ij} = X'_{ij}\beta + Z_{ij}U_i$
- **이산형 경시적 자료**(count) $Y_{ij}|U_i \sim Poisson(1, \mu_{ij}), log\mu_{ij} = X'_{ij}\beta + Z_{ij}U_i$
- 가우시안 선형혼합모델 또한 GLMM의 특별한 사례로 간주할 수 있다.

$$Y_{ij}|U_i \sim Normal(\mu_{ij}, \tau^2), \mu_{ij} = X'_{ij}\beta + Z_{ij}U_i$$

공분산구조는 그룹 간 변동(among-group variation)과 그룹 내 자기 상관 잔차(autocorrelated residuals)로 구분될 수 있다.

GLMM은 보다 다양한 경시적 자료 분석 모형에 랜덤효과를 추가하여, 오차항의 공분산구조 Σ (e.g., $\Sigma_{AR(1)}$)를 통해 반복치들 간의 연관관계 모형화할 수 있다.

The Salamander Mating Experiments I

두 타입의 도롱뇽 모집단: Rough Butt (R) and White Side (W)

$\surd$ Do salamanders prefer mating with their own population?

실험계획

- 각 도롱뇽은 두 타입의 파트너와 모두 매칭됨 (반복 측정)
- 각 도롱뇽은 짝짓기에 대한 개별적인 성향을 가지고 있으며, 이는 측정할 수 없음
- 각 도롱뇽의 성향은 독립적이라고 가정
- 도롱뇽이 짝짓기를 할 확률에 영향을 미치는 효과:
  - 페어링 타입 (RR, RW, WR, WW) (고정 효과)
  - 암컷의 개별 짝짓기 성향 (랜덤 효과)
  - 수컷의 개별 짝짓기 성향 (랜덤 효과)

실험계획

- 반응변수: 짝짓기 유무 (Binary)
- 고정효과: $\beta_{RR}, \beta_{RW}, \beta_{WR}, \beta_{WW}$ (짝짓기 확률의 로그 오즈)
- 랜덤효과: 개별 성향 반영, 서로 독립이고 정규분포를 따른다고 가정
- 추정 분산: σ_F^2, σ_M^2

▷ How?

Estimation of Parameters I

Likelihood based Inference

- $y_i = (y_{i1}, ..., y_{im_i}, U_i)$의 결합 분포(Joint):

$$f(y_i, U_i; \beta, \theta) = f(U_i; \theta) f(y_i | U_i, \beta) = f(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij} | U_i, \beta)$$

- y_i의 주변분포(Marginal):

$$f(y_i, U_i; \beta, \theta) = \int f(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij} | U_i, \beta) dU_i$$

- 우도함수(Likelihood):

$$L(\beta, \theta; y) = \prod_{i=1}^{n} \int f(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij} | U_i, \beta) dU_i$$

  y_{ij}가 정규분포라면 위 적분의 계산이 비교적 간단한 반면, 대부분의 경우 closed form 이 존재하지 않고 복잡하다.

  이 경우 크게 (1) **구적법**(Quadrature), (2) **우도함수의 근사화**, (3) **몬테카를로**(Monte Carlo) 방법을 사용한다.

수치적 방법: 구적법

구적에 의한 근사화는 피적분함수의 가중치합(weighted sum)이다. 이때 여러가지 구적법 중 **Gauss-Hermite**(GHQ)와 **adaptive Gaussian**(AGQ) 구적법이 가장 많이 쓰인다.

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{k=1}^{R} w_k f(a_k)$$

- 다항함수로 근사화될 수 있는 f에 대한 적분은 R개의 가중치합으로 근사화될 수 있으며, 여기서 a_k는 **구적점**(quadrature point), w_i는 **구적 가중치**(quadrature weight)라고 한다.
- GHQ 방법에서는 구적점이 고정된 반면, AGQ 방법에서는 구적접의 위치를 피적분함수의 형태에 따라 변화시킴으로써 적분의 정확성을 향상시키고자 한다.
- SAS의 GLIMMIX 모형의 경우, GHQ 방법 또는 1차 테일러 시리즈 근사를 통해 적분을 근사시킨다.

Estimation of Parameters III

수치적 방법: 구적법

예를 들어, 다음의 모형을 가정하였을 때,

$$g(\mu_{ij}) = x'_{ij}\beta + U_i, U_i \sim N(0, \sigma^2)$$

GHQ 방법에 의해 $L(\beta, \theta; y)$의 적분구간은 다음과 같이 R개의 가중합으로 표현된다.

$f(U_i; \sigma) \to \phi \sim N(0, 1)$이 되도록 $U_i^* = U_i/\sigma$로 표준화 한다면,

$$\int f(U_i;\sigma)\prod_{j=1}^{m_i} f(y_{ij}|U_i)dU_i = \int \phi(U_i^*)\prod_{j=1}^{m_i} f(y_{ij}|\sigma, U_i^*)dU_i^* \tag{1}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-U_i^{*2}/2} \prod_{j=1}^{m_i} f(y_{ij}|\sigma, U_i^*)dU_i^* \tag{2}$$

$$\approx \sum_{k=1}^{R} w_k^* \prod_{j=1}^{m_i} f(y_{ij}|\sigma, a_k^*) \tag{3}$$

여기서 $w_k^* = w_k/\sqrt{2\pi}, a_k^* = \sqrt{2}a_k$이다.

우도함수의 근사화

적분을 포함한 우도함수에 대해 **Laplace** 근사화를 적용할 수 있다.

$$\text{Taylor expansion: } q(x) = q(\tilde{x}) + \tfrac{1}{2}q''(\tilde{x})(x-\tilde{x})^2$$

$$\int_{-\infty}^{\infty} exp(f(x))dx \approx \int_{-\infty}^{\infty} exp[f(\tilde{x}) - (x-\tilde{x})^2/2\sigma^2]dx \tag{4}$$

$$= \int_{-\infty}^{\infty} exp(f(\tilde{x}))\sqrt{2\pi}\sigma\phi(x;\tilde{x},\sigma^2)dx \tag{5}$$

$$= exp(f(\tilde{x}))\sqrt{2\pi}\sigma \tag{6}$$

$$= c|f''(x)|^{-1/2}exp(f(\tilde{x})) \tag{7}$$

- (4)-(6)은 $\phi(x;\tilde{x},\sigma^2)$는 평균이 $\tilde{x}$이고 분산이 σ^2인 정규분포일때, $\tilde{x}$가 f(x)의 최빈값으로 $f'(\tilde{x}) = 0$이고 $f''(\tilde{x}) = 1/\sigma^2$임을 적용했다.

Estimation of Parameters V

우도함수의 근사화

Laplace 근사화를 GLMM의 적분에 적용한다.

$$L(\beta, \theta; y) = \prod_{i=1}^{n} \int f(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij}|U_i, \beta) dU_i$$

$$\Rightarrow f(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij}|U_i) = exp\{log[(U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij}|U_i)]\}$$

- 라플라스 근사화에서 정의된 최빈값($\tilde{x}$)은 다음과 같다.

$$\tilde{U}_i = argmax f((U_i; \theta) \prod_{j=1}^{m_i} f(y_{ij}|U_i)$$

우도함수의 근사화

- (1)-(3)의 정규분포 예시에 Laplace 근사화를 적용해보면, 다음과 같다.

$$logf(y_i;\beta,\sigma) \approx log(\sqrt{2\pi}\sigma_i) + log(\phi(\tilde{U}_i;\sigma)) + \sum logf(y_{ij}|\tilde{U}_i) \tag{8}$$

$$= log(\sigma_i/\sqrt{\sigma}) - \tilde{U}_i^2/2\sigma + \sum_{j=1}^{m_i} logf(y_{ij}|U_i) \tag{9}$$

- MLE를 보장하진 않는다.
- 랜덤 효과가 정규분포를 따를 때만 적용할 수 있다.
- 데이터가 sparse하면 성능이 좋지 않을 수 있다.

우도함수의 근사화

Penalized Quasi-Likelihood(PQL) 근사화

- Laplace 근사화의 일반화된 방법이다.
- $U_i \sim N(0, G)$가 주어져 있을 때, 각 개체 내 관측치 $y_i = (y_{i1}, ..., y_{im_i})$가 서로 **조건부 독립** 이라는 가정 하에 PQL 근사화 식은 다음과 같이 정의된다.

$$L_Q \propto |G|^{-1/2} \int exp\{-\tfrac{1}{2}\sum_{j=1}^{m_i} d_{ij} - \tfrac{1}{2} U_i' G^{-1} U_i\} dU_i$$

  여기서 $E(y_{ij}|U_i) = \mu_{ij} = g^{-1}(X_{ij}\beta + Z_{ij}U_i)$, $var(y_{ij}|U_i) = a_{ij}(\phi)v(\mu_{ij})$일 때 ϕ는 산포모수(dispersion parameter), $v(\mu_{ij})$는 알려진 분산 함수라면,

$$d_{ij} = -2\int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij}-u}{a_{ij}(\phi)v(u)}$$는 Quasi 분산이다.

- 위 식의 적분 구간을 Laplace 근사화를 통해 근사한다.

우도함수의 근사화

Penalized Quasi-Likelihood(PQL) 근사화

- 우도함수를 근사화하는 방법이기 때문에 MLE를 보장하지 않으며, 추정량이 consistent 하지 않다.
- Laplace 근사법의 한계로 인해 데이터의 분산이 크다면 성능이 좋지 않을 수 있다.
- R의 glmmPQL 패키지를 사용하거나, lme4의 glmer 함수에서 method ="PQL"로 지정하여 사용

Monte Carlo Integration

1. 단순 몬테칼로 적분법 (Simple Monte Carlo integration)

   f_X로부터 추출된 표본을 이용하여 표본 평균을 구해 기댓값을 근사한다.

$$E(h(X)) = \int h(x) f_X(x) dx \approx \bar{h} = \frac{1}{M} \sum_{\ell=1}^{M} h(\tilde{x}_\ell)$$

$$E(f(y_{ij}|\sigma, U_i^*)) = \int \prod_j f(y_{ij}|\sigma, U_i^*) dU_i^* \approx \frac{1}{M} \sum_{\ell=1}^{M} f(y_{ij}|\sigma, \tilde{U}_\ell^*)$$

   표본에 의해 발생되는 분산의 증가를 가져오는 단점을 가지게 된다.
   또한 f_X로부터 직접적으로 표본을 추출하는 것이 불가능한 경우 적용불가능하다.

2. 중요 샘플링 (Importance sampling)

   단순 몬테칼로 적분법의 단점을 극복하기 위한 대안으로, 적절하게 선택된 중요 함수 (importance function) g를 이용한다. g의 조건은 다음과 같다.

   1. f_X와 같은 받침(support)을 가진다.
   2. $h(x)f_X(x)/g(x)$가 x의 smooth 함수이며 제한된 값을 가진다.
   3. $h(x)f_X(x)/g(x)$의 계산이 용이하다.

   이 때 $h(x)$의 평균은

$$E(h(X)) = \int g(x)\frac{h(x)}{g(x)}f_X(x)dx \approx \frac{1}{M}\sum_{\ell=1}^{M}\frac{h(\tilde{x}_\ell^*)f_X(\tilde{x}_\ell^*)}{g(\tilde{x}_\ell^*)}$$

$$E(f(y_{ij}|\sigma, U_i^*)) \approx \sum_{\ell=1}^{M}\frac{\phi(\tilde{U}_\ell; \mu_i, \tau_i^2)f(y_{ij}|\sigma, \tilde{U}_\ell)}{\phi(\tilde{U}_\ell; \mu_i, \tau_i^2)}$$

Estimation of Parameters XI

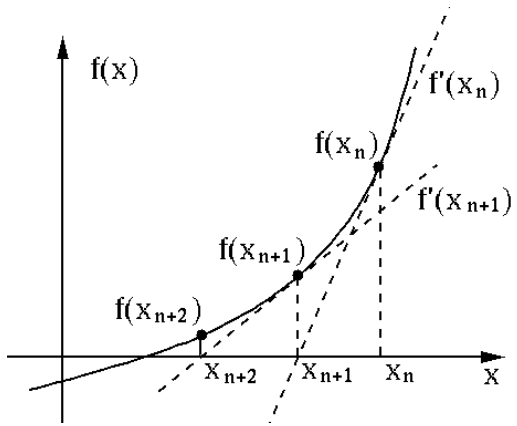
Monte Carlo EM Algorithm

$$g(\mu_{ij}) = X'_{ij}\beta + Z_{ij}U_i, \quad U_i \sim N(0, \sigma^2)$$

1. E-step:
   미지의 랜덤 효과와 그 함수를 구하기 위해 조건부 평균 계산.
   분포 가정이 없을 경우 Markov chain Monte Carlo (MCMC) 알고리즘 사용.

$$Q\{\theta|\theta^{(k)}\} = E\{\log f(y|\theta)|y, \theta^{(k)}\}$$

2. M-step:
   Newton-Raphson (NR) 기법을 적용해 최대우도추정량 계산.



Modelling the Covariance I

공분산구조

- 자료의 종속 구조를 반영하여 연관 관계 파악 가능
- 평균 모형을 구성하는 회귀계수의 표준 오차를 줄임으로써 효율성 향상

1) 균일상관모형(Uniform correlation model, compound symemetric, exchangeable)

- 관측개체 i의 모든 반복치 쌍들의 상관관계가 동일

$$Corr(Y_{ij}, Y_{ik}) = \rho, (i \neq k)$$
$$V_0 = \sigma^2|(1-\rho)I + \rho J|$$

- 모든 시간 간격에 대해 동일한 상관계수 가정은 현실적으로 적절하지 않을 수 있음

Modelling the Covariance II

2) 지수상관모형(Exponential correlation model)

- 근접한 관측치의 상관관계는 거리가 먼 관측치와의 상관관계보다 강할 것으로 예상됨
- 관측 시점이 서로 다른 경우 시간 간격을 고려한 상관계수 고려

$$Cov(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma^2 \exp(-\phi|t_j - t_k|) \quad \phi > 0$$

- 시간 간격이 일정한 경우

$$\sigma_{jk} = \sigma^2 \rho_{jk}$$
$$\rho_{jk} = \exp(-\phi|j-k|), \quad 0 \le \rho \le 1$$

3) 무구조 상관모형(Unstructured correlation model)

- 분산 공분산의 형태에 제한 없음
- 자료에 가장 충실하나 추정해야 할 모수가 많기에 계산적 부담 존재

$$V_0 = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

Modelling the Covariance III

Structure	Example	Parameters
Linear random coefficients (RCL)	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}' + \begin{bmatrix} \sigma^2 & & & & \\ & \sigma^2 & & & \\ & & \sigma^2 & & \\ & & & \sigma^2 & \\ & & & & \sigma^2 \end{bmatrix}$	4
Independent increments (I-I)	$\begin{bmatrix} \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ & \sigma_1^2+\sigma_2^2 & \sigma_1^2+\sigma_2^2 & \sigma_1^2+\sigma_2^2 & \sigma_1^2+\sigma_2^2 \\ & & \sigma_1^2+\sigma_2^2+\sigma_3^2 & \sigma_1^2+\sigma_2^2+\sigma_3^2 & \sigma_1^2+\sigma_2^2+\sigma_3^2 \\ & & & \sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2 & \sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2 \\ & & & & \sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2+\sigma_5^2 \end{bmatrix}$	5
Heterogeneous compound symmetry (CSH)	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho & \sigma_1\sigma_5\rho \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho & \sigma_2\sigma_5\rho \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho \\ & & & & \sigma_5^2 \end{bmatrix}$	6
Heterogeneous first-order autoregressive [ARH(1)]	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 & \sigma_1\sigma_5\rho^4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 & \sigma_2\sigma_5\rho^3 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho^2 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho \\ & & & & \sigma_5^2 \end{bmatrix}$	6
Huynh–Feldt (HF)	$\begin{bmatrix} \sigma_1^2 & (\sigma_1^2+\sigma_2^2)/2-\lambda & (\sigma_1^2+\sigma_3^2)/2-\lambda & (\sigma_1^2+\sigma_4^2)/2-\lambda & (\sigma_1^2+\sigma_5^2)/2-\lambda \\ & \sigma_2^2 & (\sigma_2^2+\sigma_3^2)/2-\lambda & (\sigma_2^2+\sigma_4^2)/2-\lambda & (\sigma_2^2+\sigma_5^2)/2-\lambda \\ & & \sigma_3^2 & (\sigma_3^2+\sigma_4^2)/2-\lambda & (\sigma_3^2+\sigma_5^2)/2-\lambda \\ & & & \sigma_4^2 & (\sigma_4^2+\sigma_5^2)/2-\lambda \\ & & & & \sigma_5^2 \end{bmatrix}$	6

Structure	Example	Parameters
First-order factor analytic with Constant Diagonal [FA1(1)]	$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}' + \begin{bmatrix} \sigma^2 & & & & \\ & \sigma^2 & & & \\ & & \sigma^2 & & \\ & & & \sigma^2 & \\ & & & & \sigma^2 \end{bmatrix}$	6
Quadratic random coefficients (RCQ)	$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}' + \begin{bmatrix} \sigma^2 & & & & \\ & \sigma^2 & & & \\ & & \sigma^2 & & \\ & & & \sigma^2 & \\ & & & & \sigma^2 \end{bmatrix}$	7
First-order antedependence [ANTE(1)]	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_1\sigma_5\rho_1\rho_2\rho_3\rho_4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_4\rho_2\rho_3 & \sigma_2\sigma_5\rho_2\rho_3\rho_4 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho_3 & \sigma_3\sigma_5\rho_3\rho_4 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho_4 \\ & & & & \sigma_5^2 \end{bmatrix}$	9
Heterogeneous Toeplitz (TOEPH)	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 & \sigma_1\sigma_5\rho_4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 & \sigma_2\sigma_5\rho_3 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho_1 & \sigma_3\sigma_5\rho_2 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho_1 \\ & & & & \sigma_5^2 \end{bmatrix}$	9
First-order factor analytic	$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}' + \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \sigma_3^2 \end{bmatrix}$	10

Applications

- biological and medical research
- longitudinal data analysis
- small area estimation

Examples I

Indonesian Children Health Study data

미취학 인도네시아 아동을 대상으로 비타민 A 결핍 및 다른 공변량과 호흡기 질환과의 관계를 알아보고자 함

표본: 250명

시점: 6개 (0, 3, 6, 9, 12, 15주, time)

반응변수: 호흡기 질환 여부 (response: 0,1)

설명변수: 성별(gender), 나이(age), 비타민 결핍여부(vita: 0,1)

	id	response	time	gender	vita	age
1	1	0	0	1	0	4
2	1	0	3	1	0	4
3	1	0	6	1	0	4
4	1	1	9	1	0	4
5	1	0	12	1	0	4
6	1	0	15	1	0	4
7	2	0	0	0	0	3
8	2	0	3	0	0	3

Examples II

Model

$$\text{logit } \Pr(Y_{ij} = 1|U_i) = \Pr(Y_{ij} = 1|U_i)/\Pr(Y_{ij} = 0|U_i)$$

1) Random intercept

$$\text{logit } \Pr(Y_{ij} = 1|U_i) = \beta_0 + \beta_1 gender_{ij} + \beta_2 age_{ij} + \beta_3 vita_{ij} + \beta_4 t_{ij} + U_i,$$
$$U_i \sim N(0, \sigma_u^2)$$

2) Random slope

$$\text{logit } \Pr(Y_{ij} = 1|U_i) = \beta_0 + \beta_1 gender_{ij} + \beta_2 age_{ij} + \beta_3 vita_{ij} + \beta_4 time_{ij} + U_{i0} + U_{i1} t_{ij},$$
$$\begin{pmatrix} U_{i0} \\ U_{i1} \end{pmatrix} \sim BVN\left(0, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$$

Examples III

Interpretation

- 랜덤효과 모수 추정치
  - 분산: 클수록 해당 변수의 변이가 큰 것을 의미
  - 공분산: 양수이면 두 변수가 양의 상관관계, 음수이면 음의 상관관계
- 고정효과 모수 추정치
  - i번째 사람이 비타민 결핍이 아닐 때

$$\exp(\beta_0 + U_i) = \frac{\Pr(Y_{ij} = 1 | x_{ij} = 0, U_i)}{\Pr(Y_{ij} = 0 | x_{ij} = 0, U_i)}$$

  - i번째 사람이 비타민 결핍일 때

$$\exp(\beta_0 + \beta_1 + U_i) = \frac{\Pr(Y_{ij} = 1 | x_{ij} = 1, U_i)}{\Pr(Y_{ij} = 0 | x_{ij} = 1, U_i)}$$

  - 비타민 결핍 여부에 따른 호흡기 질환에 대한 오즈비

$$\exp(\beta_1) = \frac{\exp(\beta_0 + \beta_1 + U_i)}{\exp(\beta_0 + U_i)}$$

Code I

R

```r
rm(list=ls())
library(lme4) # `glmm`, `nmle` is also available

ichs = read.table('/Users/hsj/Desktop/통계상담/ichs.txt')
colnames(ichs) = c("id", "response", "time", "gender", "vita", "age")

fit1 = glmer(response ~ time + vita + gender + age + (1|id),
             data = ichs, family = 'binomial')
      # default for intergral: Laplace
summary(fit1)

fit2 = glmer(response ~ time + vita + gender + age + (time|id),
             data = ichs, family = 'binomial')
summary(fit2)
```

Code II

```
> summary(fit1)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite
  Quadrature, nAGQ = 7) [glmerMod]
 Family: binomial  ( logit )
Formula: response ~ time + vita + gender + age + (1 | id)
   Data: ichs

     AIC      BIC   logLik deviance df.resid
  1345.3   1377.1   -666.6   1333.3     1494

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1761 -0.3864 -0.1750  0.2967  2.6900

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 7.806    2.794
Number of obs: 1500, groups:  id, 250

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.15209    0.53337  -2.160  0.03077 *
time         0.03410    0.01599   2.133  0.03296 *
vita         0.59914    0.43218   1.386  0.16565
gender      -1.09267    0.42083  -2.596  0.00942 **
age         -0.14040    0.10908  -1.287  0.19802
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr) time   vita   gender
time   -0.235
vita   -0.300  0.008
gender -0.367 -0.014  0.010
age    -0.717 -0.007 -0.030 -0.028
```

Code III

```
> summary(fit2)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: response ~ time + vita + gender + age + (time | id)
   Data: ichs

     AIC      BIC   logLik deviance df.resid
  1360.4   1402.9   -672.2   1344.4     1492

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0868 -0.3699 -0.1718  0.3088  2.5999

Random effects:
 Groups Name        Variance Std.Dev. Corr
 id     (Intercept) 8.46162  2.90889
        time        0.00125  0.03535  -0.66
Number of obs: 1500, groups:  id, 250

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.22593    0.54798  -2.237  0.02527 *
time         0.04390    0.02566   1.711  0.08714 .
vita         0.61209    0.43244   1.415  0.15694
gender      -1.14447    0.42228  -2.710  0.00672 **
age         -0.15822    0.10950  -1.445  0.14847
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr) time   vita   gender
time   -0.327
vita   -0.272 -0.089
gender -0.357  0.064 -0.005
age    -0.645 -0.129 -0.014 -0.035
```

Code IV

SAS

```
/* Model 1*/
proc glimmix data=ichs method=quad(qpoints=7);
class id;
model response(desc) = vita time age gender/s dist=binary link=logit;
/* desc: targetting y=1  */
random intercept / subject=id type=un;
/* type: covariance type */
run;

/* Model 2*/
proc glimmix data=ichs method=quad;
class id;
model response(desc) = vita time age gender/s dist=binary link=logit;
random intercept time/ subject=id type=un;
run;
```

Code V

The GLIMMIX Procedure

Model Information	
Data Set	WORK.ICHS
Response Variable	response
Response Distribution	Binary
Link Function	Logit
Variance Function	Default
Variance Matrix Blocked By	id
Estimation Technique	Maximum Likelihood
Likelihood Approximation	Gauss-Hermite Quadrature
Degrees of Freedom Method	Containment

Fit Statistics	
-2 Log Likelihood	1333.25
AIC (smaller is better)	1345.25
AICC (smaller is better)	1345.31
BIC (smaller is better)	1366.38
CAIC (smaller is better)	1372.38
HQIC (smaller is better)	1353.76

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1.1521	0.5333	246	2.16	0.0317
vita	-0.5991	0.4321	1249	-1.39	0.1659
time	-0.03409	0.01599	1249	-2.13	0.0332
age	0.1404	0.1091	1249	1.29	0.1982
gender	1.0926	0.4208	1249	2.60	0.0095

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	id	7.8052	1.3560

Code VI

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	id	9.3114	2.5751
UN(2,1)	id	-0.08294	0.1319
UN(2,2)	id	0.005452	0.005996

Fit Statistics	
-2 Log Likelihood	1332.14
AIC (smaller is better)	1348.14
AICC (smaller is better)	1348.23
BIC (smaller is better)	1376.31
CAIC (smaller is better)	1384.31
HQIC (smaller is better)	1359.48

Solutions for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1.2218	0.5622	246	2.17	0.0307
vita	-0.6027	0.4474	1000	-1.35	0.1782
time	-0.04199	0.02448	249	-1.71	0.0876
age	0.1506	0.1135	1000	1.33	0.1850
gender	1.1147	0.4352	1000	2.56	0.0106

Reference I

[1] Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

[2] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

[3] Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Patrick J Heagerty, Scott Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002.

[4] M Ataharul Islam and Rafiqul I Chowdhury. *Analysis of repeated measures data*. Springer, 2017.

[5] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007.

[6] Christina Knudson, Sydney Benson, Charles Geyer, and Galin Jones. Likelihood-based inference for generalized linear mixed models: Inference with the r package glmm. *Stat*, 10(1):e339, 2021.

[7] Shonosuke Sugasawa and Tatsuya Kubokawa. Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*, 3(2):693–720, 2020.

[8] 김양진. *R과 SAS를 이용한 경시적 자료분석*. 자유아카데미, 2020.