Generalized Linear Mixed Models
일반화선형혼합모형

이소연, 황서진

고려대학교 통계학과

April 8, 2021

Contents

1. Introduction
2. The Salamander Mating Experiments
3. Inference

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)

Introduction III

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

  관측개체별 공유되는 특성을 개체별 랜덤 효과 U_i로 표현한다. 따라서, U_i가 주어져 있다면, 각 개체 내 관측치 $y_1, ..., y_n$는 서로 **독립** 이라고 가정한다. 즉, **조건부 독립(conditional independence)**를 가정한다.

  각 관측치의 평균을 $E(y_i|U_i) = \mu_i$ 라고 할 때, **연결함수**(link function) g에 의해 다음과 같이 표현될 수 있다.

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

  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.

Introduction VI

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 사례례로 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)

The Salamander Mating Experiments III

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?

Inference 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 form이 존재하지 않고 복잡해 일반적으로 랜덤 효과에 대해서 다변량 정규분포를 적용한다.

  이 경우 크게 (1) **구적법**(Quadrature), (2) **몬테카를로**(Monte Carlo) 방법, (3) **우도함수의 근사화** 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 함
- SAS의 NLMIXED 모형의 경우, AGQ 방법 또는 1차 테일러 시리즈 근사를 통해 적분을 근사시킴

우도함수의 근사화

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

$$\int_{-\infty}^{\infty} exp(f(x))dx \approx \int_{-\infty}^{\infty} exp[f(\tilde{x}) - (x-\tilde{x})^2/2\sigma^2]dx \tag{1}$$

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

$$= exp(f(\tilde{x}))\sqrt{2\pi}\sigma \tag{3}$$

$$= c|f''(x)|^{-1/2}exp(f(\tilde{x})) \tag{4}$$

- (3)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 적용함
- 랜덤 효과가 정규분포를 따를 때만 적용할 수 있다.

Reference

[1] Alan Agresti. *Categorical Data Analysis,Third Edition*. Wiley. 2013.

[2] Breslow, N. and Clayton, D. *Approximate inference in generalized linear mixed models*. Journal of the American Statistical Association. 88:9-25. 1993.

[3] Christina Knudson. *Likelihood-Based Inference for Generalized Linear Mixed Models: Inference with R Package glmm*. University of St. Thomas. 2017.

[4] Jiming Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer. 2007.

[5] M. Ataharul Islam. *Analysis of Repeated Measures Data*. Springer. 2017.

[6] 김양진. *R과 SAS를 이용한 경시적 자료분석*. 자유아카데미. 2017.