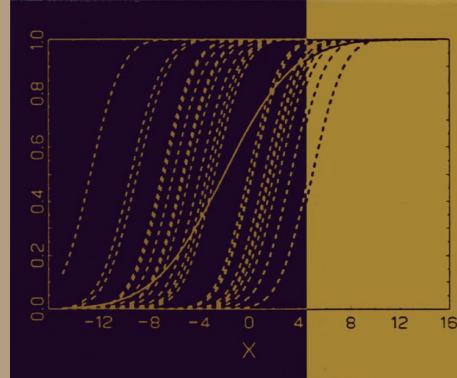

Wiley Series in Probability and Statistics

Generalized, Linear, and Mixed Models

Charles E. McCulloch, Shayle R. Searle



Chapter 1.

Introduction

1. Models

a. Linear models (LM) and Linear Mixed models (LMM)

• LM : Linear Model

- 통계학 관심 : average & variation average

- ANOVA

관측치들을 평균합 + 평균 간 차이로 표현

특정 상황 하, 평균/변동성의 특성에 대한 추론

- 선형 모형의 필수적 조건

① 각 자료의 평균이 알려져 있지 않은 모수 (상수)의 선형결합

선형 요구 조건은 관측치 y_{ij} 의 기댓값 (평균) : μ_{ij}

$$\text{ex) } \mu_{ij} = \mu + \alpha_i + \beta_j \leftarrow \mu, \alpha_i, \beta_j \text{는 마지막의 상수, 추정관심}$$

② 데이터는 정규분포 가정.

: y_{ij} 는 평균 μ_{ij} 를 갖는 정규분포

③ 선형 모형 : 모형이 모수에 대해 선형적.

$$\mu_{ij} = b_0 + b_1 x_{ij} + b_2 x_{ij}^2 \text{ 도 선형성 有.}$$

• LMM : Linear Mixed Model

- LM의 변형

: LM내 모수들이 상수로 다뤄지지 X. 확률 변수로 다룬다.

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

→ 랜덤으로 취급되는 α_i

① 평균 = 0 ② 등분산 (σ^2_α) ③ Un correlation

④ 정규성

b. Generalized models (GLMs and GLMMs)

LM, LMMs $\xrightarrow{\text{extended!}}$ GLM, GLMMs

• generalization

- 데이터가 정규분포 따르는 가정 불필요

- 평균은 모수들의 선형조합일 필요X

But, 모수들의 선형조합은 평균의 함수

ex) Poisson distribution 따르는 data

: 평균 λ , 모수의 선형조합 $\log(\lambda)$

- GLM : 모든 모수를 fixed 된 상수로 고려

GLMM : 일부 모수를 random 으로 고려

2. Factors, Levels, Cells, Effects & DATA

* 데이터의 변동성을 다양한 범주 분류의 결과로 본다!

<예제 통해 확인!>

: basal cell epithelioma site 연구에서, 환자들을 성별,

나이대, 햇빛 노출 정도에 따라 분류

Table 1.1: A Format for Summarizing Data

Gender	Low Exposure to Sunshine			High Exposure to Sunshine		
	Age Group			Age Group		
	A	B	C	A	B	C
Male						
Female						

- factor : 각 자료의 source 확인하는 성별, 나이대, 햇빛 노출 정도

- level : 각 factor가 가질 수 있는 값

ex) 성별 factor의 level : Male, Female

- cell

: 모든 factor의 한 level의 intersection에서의 데이터

부분집합 → 2(성별 level 수) × 2(햇빛 노출) × 3(나이) = 12개 cell

- crossed

: 3개의 나이대 그룹은 햇빛 노출 (high/low)에 상관없이

같은 범주 의미 → Age와 Sunshine factor는 crossed

- nested

Table 1.2: Summarizing Exam Grades

Gender	English Section			Geology Section		
	A			B		
	A	B	C	A	B	C
Male						
Female						

: English와 Geology에서의 3 section에 대한 시험 성적 조사

→ English의 'A' section과 Geology의 'A' section 사이의 connection X

→ Section factor는 과목 factor 내 nested 됨

• effect

: factor와 level 관점에서 데이터 분류할 때, 관심있는 feature은 다른 level들이 관심 변수에 영향을 미치는 범위

① Fixed effect

: 모수가 고정된 상수로 간주. 우리의 관심사

데이터에서의 한 factor의 유한한 level의 정합

② Random effect

: 모수를 random으로! 한 factor에 대한 무한한 level 정합

why? 랜덤 샘플은 데이터에서 발생한 것으로 간주.

cf. Fixed Effect vs Random Effect

ex) 3개 다른 온도에서 구워진 6덩이 빵 중 4덩이 추출

이 때, 특정 온도에 관심있다고 하면 통계적 관심 온도효과 추정 → fixed effect!

"온도"는 연속값 → 450° 온도효과와 405° 온도효과 비슷 → 온도효과 랜덤 X

ex) batches는 연속값으로 정의 X → 실제 개체들, 사람, 냉동…

batch effect를 랜덤으로 가정하고 그 효과의 variation 추정 관심

→ Variance Components

: random variation은 batch variance & error variance

→ 선형 모델에서, 관측된 변수의 variance는 Variance Components 합.



• fixed effect model

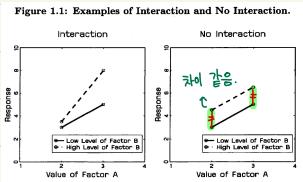
: 오직 fixed effect만 존재

• random effect model

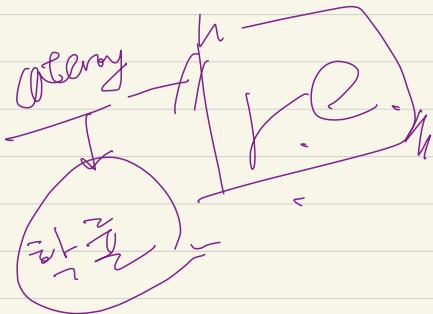
: 모든 관측치에서의 공통된 하나의 일반적 평균과 별도로, 랜덤 효과 갖는 것

③ Interaction effect

: 여러개의 factor가 있을 때, 2개 이상의 조합 효과



→ factor A의 2개 레벨 사이 반응변수 평균 차이가 factor B의 다른 레벨들과 같으면 interaction X



④ main effect : 하나의 factor 효과

* LM&LMM의 분산분석 기술은 balanced data에 제한.

- balanced data

: 데이터의 모든 케이스가 같은 수의 관측치를 갖는다.

- unbalanced data

: 같은 수 관측치 X

① all-cells - filled data

② Some-cells - empty data

3. Fixed Effects Model

<Example 1: Placebo and a drug >

• y_{ij} : i 치료 받은 j 환자의 발작 횟수 ($i=1$: 풀라시보, $i=2$: 약물투여)

• y_{ij} 모델링

$$E(y_{ij}) = \mu_i = \mu + \alpha_i$$

μ : general mean, α_i : i 치료 동안 발작 횟수 효과.

- y_{ij} 의 기대값 모델링은 각 μ_i ($=\mu$ 와 α_i)를 고정된 미지수

상수로 고려한다. 추정하고자 하는 것은 $\mu_1, \mu_2, \mu_2 - \mu_1$

- 두 처치의 차이 고려하기 위해서, $\mu_2 - \mu_1 < 0$ 인지 test

→ 약이 발작 횟수를 줄이는지 검증

이 때 그 처치 이외의 다른 처지는 없다고 고려함. (Fixed effect)

- 데이터에 관한 추론은 데이터 샘플링 방법에 기인

fixed effects 모델은 적절한 샘플링 과정 보장 가능

→ 반복된 일상 시험에서 도출될 수 있는 그거지 치료와 관련된

가능한 데이터 세트 중 하나, 각 치료를 받는 다른 사람들의

샘플의 반복 사용

⇒ 2개의 랜덤 샘플 데이터 합침,

데이터의 한 모집단 평균은 μ_1 , 다른 하나의 모집단 μ_2

• Fixed effect의 특징

: 관심있는 factor의 다양한 level의 반응 변수 y 의

effect를 상수로 나타낸다.

→ Example 1에서, 처리는 특정 관심 factor의 level이며,

이 치료법의 관심 때문에 선택됨.

<Example 2. Comprehension humor>

• 상황 설명

- 만화 3종류 : 시작, 언어, 시각 + 언어

- 사춘기 나이대 그룹 : 평범, 학습 장애 有

- \bar{y}_{ij} : i 만화책에 따른 j 그룹내 사람들의 평균 이해력 점수

(이해X) 1점 ~ 9점 (이해O)

• \bar{y}_{ij} 모델링

$$E(\bar{y}_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j$$

μ : general mean, α_i : 만화 종류에 따른 이해도 effect ($i=1, 2$)

β_j : 사춘기 Group 내 응답자에 따른 이해도 effect ($j=1, 2, 3$)

- 2-Crossed factors : 3 종류의 만화와 그 그룹의 청소년들

- α_i 와 β_j 는 fixed effect

: 더 많은 만화 종류와 청소년 그룹에서 임의로 선택되었다고 할 수 X

<Example 3 : Four dose levels of drug >

• 상황 설명

- 약을 4가지의 다른 복용량 수준에서 관찰하는 일상시험

- y_{ij} : i 복용량을 받은 j 번째 사람의 발작 횟수.

- y_{ij} 모델링

$$E(y_{ij}) = \mu_{ij} = \mu + \alpha_i \quad (i=1, 2, 3, 4) \rightarrow \mu_i (\alpha_i) \text{는 fixed effects}$$

- 우리가 관심 있는 복용량 고정.

관심사는 발작 낮추는 효과가 있는 복용량 level 사이의 차이.

• Fixed effect 모델링

: factor의 개수가 얼마나 많은 지에 상관 없이, 모두

fixed effect라면, 선형 조합 가능

$$\mu_i = \mu + \alpha_i + \beta_i + \dots$$

④ 비선형 모형에 관한 fixed effect도 가능

4. Random effect Models

<Example 4 : Clinics>

- 뉴욕 도시의 20개 다른 병원에서 임상시험. 복용량은 수준 하나.

- Y_{ij} : i 병원 내 j 환자의 밀작 횟수

$$- E(Y_{ij}) = \mu + \alpha_i \quad (i=1, 2, \dots, 20)$$

이 때, 병원들은 뉴욕시 모든 병원 분포에서의 랜덤 샘플로 고려.

• α_i : fixed effect. 밀작 횟수에 대한 복용량 i 레벨의 effect

복용량 i 레벨은 관심에 따라 미리 정해짐. (ex3)

• α_i : i 병원 내 있는 환자의 밀작 수에 대한 effect

뉴욕의 모든 병원 모집단 대표. (임의 선택) \rightarrow random effect

- random effect 특징 : 추론은 모집단에 대해서!

\rightarrow random variable 들의 Variance 추론에 유용함

ex) 병원 간 Variation 크기,

어떤 병원이 밀작을 가장 잘 줄이는지 예측

$$\alpha_i \sim \text{IID}(\mu, \sigma^2_\alpha) \quad \forall i$$

• Notation

① Properties of random effects in LMMs

- α_i 를 random variable로 취급하려면 학술적 특성 가정.

$\rightarrow \alpha_i$ 들은 independent, identical distribution (iid)

$\rightarrow \alpha_i$ 의 평균 = 0, 분산 = σ^2_α (등분산)

$$\Rightarrow \alpha_i \sim \text{iid}(\mu, \sigma^2_\alpha) \quad \forall i$$

$$E(\alpha_i) = 0 \quad \forall i, \quad \text{Var}(\alpha_i) = E[(\alpha_i - E(\alpha_i))^2] = E(\alpha_i^2) = \sigma^2_\alpha$$

$$\text{Cov}(\alpha_i, \alpha_k) = 0 \quad \text{for } i \neq k$$

② The Notation of mathematical statistics

- α_i 를 random variable로 가정할 때

$$E(Y_{ij}) = \mu + \alpha_i \quad \leftarrow \alpha_i \text{값에 대해 계산된 조건부 평균}$$

$$\Rightarrow E(Y_{ij} | A_i = \alpha_i) = \mu + \alpha_i \quad \rightarrow E(Y_{ij} | A_i) = \mu + A_i$$

$$- E(Y_{ij})$$

$$= E_A(E(Y_{ij} | A_i)) = E_A(\mu + A_i) = \mu + E_A(A_i) = \mu$$

\because A 분포에서 평균 = 0 가정, $E_A(A_i) = 0$

(증명) goal : $E_A(A_i) = 0 \rightarrow E_A(A_i) = C$ 가정.

$$E(Y_{ij}) = E_A[E(Y_{ij} | A_i)] = E_A(\mu + A_i) = \mu + E_A(A_i) = \mu + \underline{C} = \underline{\mu}$$

$$E(Y_{ij} | A_i) = \mu + A_i = \mu + C + A_i - C = \mu' + A_i'$$

$$A_i' = A_i - C \rightarrow E(A_i') = E(A_i) - C = C - C = 0$$

$$E(Y_{ij} | A_i = \alpha_i) = \mu + \alpha_i = \mu + C + \alpha_i - C = \mu' + \alpha_i'$$

$$E(\alpha_i') = E(\alpha_i) - C = C - C = 0$$

- Random effect의 특성 level에 관심 : α_i 의 실제값 예측

ex) 1 번째 병원이 평균값과 얼마나 다른지

- A_i 가 추출되는 모집단에 관심 : $\text{Var}(A_i) = \sigma^2_\alpha$ 추정

③ Variance of y

- $\text{Var}(a_i) = \delta_a^2$ 로 정의하면, random factor 정한 뒤

데이터에 남아있는 variation 고려하여 $\text{Var}(y_{ij})$

- 데이터 정규분포일 때

$$\text{residual error} = y_{ij} - E(y_{ij}|a_i)$$

$$\rightarrow y_{ij}|a_i \stackrel{\text{iid}}{\sim} N(E(y_{ij}|a_i), \delta^2)$$

정규성이 없다면 ...?

ex) y_{ij} : 빙작수, $y_{ij} \stackrel{\text{iid}}{\sim} \text{Poisson}(E(y_{ij}|a_i))$

$$\rightarrow y_{ij}|a_i \stackrel{\text{iid}}{\sim} \text{Poisson}(E(y_{ij}|a_i))$$

"residual" variation도 표아송 분포. $\leftarrow y_{ij} - E(y_{ij}|a_i)$

\Rightarrow 하지만 잔차가 정수가 아닌 횟수(표아송 분포)는 어색!

④ Variance and Conditional expected values

- $\text{Var}(y) = \text{Var}(E(y|u)) + E(\text{Var}(y|u))$ 식 이용해 $\text{Var}(y)$ 구하기

$$(\text{증명}) \text{Var}(y) = \text{Var}(E(y|u)) + E(\text{Var}(y|u)) * \text{Var}(x) = E(x^2) - E(x)^2$$

$$\text{Var}(E(y|u)) = E(E(y|u)^2) - E(E(y|u))^2$$

$$E(\text{Var}(y|u)) = E(E(y^2|u) - E(y|u)^2) = E(E(y^2|u)) - E(E(y|u)^2)$$

$$\therefore \text{Var}(y) = E(E(y|u)^2) - E(E(y|u))^2 + E(E(y^2|u)) - E(E(y|u)^2)$$

$$= E(E(y^2|u)) - E(E(y|u))^2 = E(y^2) - [E(y)]^2$$

The law of Iterated expectation

- 등분산 선형 모형의 경우, 일반적인 components of variance 성립.

$$\delta_y^2 = \text{Var}(y_{ij}) = \text{Var}(E(y_{ij}|a_i)) + E(\text{Var}(y_{ij}|a_i)) = \text{Var}(M+a_i) + E(\delta^2)$$

$$= \delta_a^2 + \delta^2$$

- 공분산

$$\text{Cov}(y, w) = \text{Cov}(E(y|u), E(w|u)) + E(\text{Cov}(y, w|u))$$

$$\text{Cov}(y_{ij}, y_{ij'}) = \text{Cov}(E(y_{ij}|a_i), E(y_{ij'}|a_i)) + E(\delta^2)$$

$$= \text{Cov}(M+a_i, M+a_i') = \delta_a^2$$

$\rightarrow \delta_a^2$: 클래스 내 Covariance, 같은 클래스 내 모든 쌍의 관측치 간 Cov

$\rightarrow \frac{\delta_a^2}{\delta_a^2 + \delta^2}$: intra-class correlation coefficient, 굽내 상관 계수

〈Example 5 : Ball bearings and calipers〉

• 상황 설명

- 특정 지름의 공 100개 생산, 20 아이크로미터 캘리퍼로 측정

- y_{ij} : j 번째 캘리퍼로 측정한 i 번째 공의 지름

$$\Rightarrow E(y_{ij}) = M + a_i + b_j.$$

- 2-Crossed factor

a_i : 생산라인으로 부터 100개공 랜덤 샘플

b_j : 가능한 캘리퍼 모집단 중 20개 캘리퍼 랜덤 샘플

$$\rightarrow \text{Cov}(a_i, b_j) = 0 ; a_i \text{와 } b_j \text{ 독립.}$$

- 추론의 대상

(캘리퍼들 사이에의 분산 크기

(생산된 공 사이의 분산 크기

c.f. The law of Iterated expectation, 반복 평균 법칙

: 두 확률 변수 X, Y 에 대해 $E_X[E(Y|X)] = E(Y)$

5. Linear Mixed Models (LMMs)

<Example 6 : Medications and Clinics >

- 2-crossed factors

: 20개 병원에서 복용량 4수준 치료.

병원 내 각 환자는 복용량 수준 임의로 배정

- y_{ijk} : i 병원내 j수준 복용량에 대한 k 환자 자료

- Mixed model

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + C_{ij}$$

α_i : i 병원 효과, random effect, β_j : j 복용량 효과, fixed effect.

C_{ij} : 병원과 복용량 상호작용 효과, random effect

→ fixed effect 와 random effect 사이 상호작용이니까 random

⇒ random + fixed effect

: fixed effect 추출과 Variance Components 문제 합침

- 관심대상

: 다른 복용량에 대한 효과성, 병원 사이에서 변동성 (분산)

- Mixed model

: 설명되지 않는 변동 포함하는 써있는 모든 모형은 mixed model

따라서 자동적으로 fixed effect + random effect.

But, 실제로 Mixed model은 일반적으로 설명되지 않는 변동성

뿐만 아니라 미지의 fixed effect 와 random effect 모두

갖는 모형

데이
↓
상황
13) 14

✓<Example 7 : Drying methods and fabrics>

• 상황 설명

- 드라이 이후 세척된 섬유 부드러움 평가

- 9개의 다른 섬유 5개의 드라이 방

- 드라이 방법은 fixed effect

- 섬유

Pop
Fixed Factor
건조 (자연)
기계 펴질 건조
섬유 속
ir movement.

① fixed factor : 9개 섬유가 고려중인 유일한 섬유로 선택되었을 때.

② random factor : 9개 섬유가 모집단에서 온 랜덤 샘플이었을 때.

⇒ factor 가 fixed / random 결정하는 것은 데이터의 level과

factor의 특성을 생각하는 것.

⇒ 추론 방법 : fixed effect를 사이 차이,

random effect를 사이 분산 크기

<Example 8 : Potomac River Fever>

- 뉴욕주 522개 농장 내 말의 사육적 그룹들로부터 말 샘플링

- 사육적 그룹 : 동물들이 같은 첫간 / 초원 내 함께 있게 앉기를 원하는지로 정의

- fixed factor

: 품종, 성별, 동물 케어 유형 (마구간 청소, 파리 스프레이 빈도) 등

- random factor

: 농장과 농장 내 nested 된 말의 사육적 그룹

• Regression models

- 회귀분석의 기초 모형식 : $E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$

- β : 고정된 상수로 고려 (fixed effect), y 와 x 의 데이터로 추정함.

- Random Coefficients Model

: 몇 β 들을 random으로 생각.

◦ Longitudinal data

- 각 개체에 대해 2번 이상 수집된 데이터

- Mixed Model 의 적용 많이 함

- longitudinal data에서 실험하는 이유

① 개체 내 비교함으로써 민감도 ↑

② 시간에 따른 변화 연구

③ 연구에 등록된 개체들 호흡적 사용

- fixed vs random

: longitudinal 연구에서 effect가 개체에 따라 달라지는지를

보고 fixed / random 결정

→ random : 개체들은 더 큰 모집단에서의 랜덤 샘플이고,

개체에 대해 상수가 아닌 모든 effect 들

- ex) 혈압 악 test

: 각 개체에 대해 2회분 / 통제용량 (여러분)

✓ 개인들은 평균 혈압이 다를 것 → 모형은 개인에 대해 각 낭절편 有

✓ 액을 용량 증가에 대한 반응도 개체마다 다를

→ 각 개체에 따라 복용량에 대해 따로 X절편 모델링

✓ 개체 나이에 따라 혈압의 절진적 변화 가능성

→ $y_{ij} = \beta_0 + d_{ij}$, X_i 나이의 i번째 개체가 축정한 혈압

$$E(y_{ij}) = \alpha_i + b_i d_{ij} + \gamma X_i.$$

α_i, b_i : i번째 개체에 따라 구체화됨. random factor

γ : 모든 개체에 대해 같음. fixed

→ 전반적인 모집단 반응에 관심 있을 때

$$E(y_{ij}) = (\alpha + \alpha'_i) + (\beta + b'_i) d_{ij} + \gamma X_i.$$

$$\alpha'_i = \alpha_i - \alpha, b'_i = b_i - \beta \rightarrow \alpha, \beta \text{는 개체의 모집단 평균. fixed}$$

α'_i, b'_i : 전체 평균에서 온 개체별 편차, random effect.

$$\text{평균} = 0, \text{분산} = \sigma^2_a, \sigma^2_b$$

◦ Model equations

: $E(y_{ij}) = \mu + \alpha_i$ 와 $E(y_{ij}) = \mu + \alpha_i$ 는 본질적으로 동일

But 동일한 모형 X. α_i 는 fixed, α_i 는 random effect

→ 데이터 분석도 방향 달라짐.

⊕ α_i, α_i 는 fixed / random effect 이외 다른 모형 적용 가능

ex) $\mu + \alpha_i$: 분산 스타일 모형 분석, α_i 는 이항 모델에서 발생

cf 민감도

		Condition (실제)	
		Positive	Negative
Prediction (예측)	Positive	TP = 20 True Positive	FP = 90 False Positive
	Negative	FN = 10 False Negative	TN = 880 True Negative
		Sensitivity (민감도)	Specificity (특이도)

- 정확도 (Accuracy) : $\frac{TP+TN}{total}$

- 민감도 (sensitivity) : 실제 positive 중에서 positive로 예측한 비율.

$$\frac{TP}{TP+FN}$$

- 특이도 (specificity) : 실제 negative 중에서 negative라고 예측한 비율.

$$\frac{TN}{TN+FP}$$

6. Fixed or Random?

• $E(y_{ij}) = \mu + \alpha_i + \beta_j$ 와 $E(y_{ij}) = \mu + \alpha_i + b_j$ 는 구분 X

→ fixed / random 효과의 해석 때문에 모형 어려움

<Example 9: Clinic effects >

• 상황 설명

- 새로운 수술 절차 효율성 판단.

① 절경 더 많은 병원에서 시행되는 절차라면, 병원 대표 collection

선택해 병원을 random effect로 간주.

② 매우 전문적 절차여서 적은 수의 병원에서 시행한다는 가정

병원의 더 큰 그룹에서 샘플 선택 불가능.

현재의 병원 추론, 병원을 fixed effect로 간주.

→ 모형 적용 상황은 effect가 fixed / random 고려하는데 있어

factor를 결정하는 것.

EY

y_e

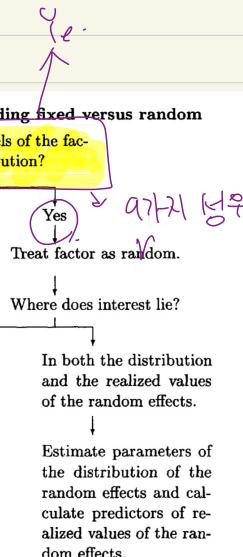


Figure 1.2: Decision tree for deciding fixed versus random

Is it reasonable to assume that levels of the factor come from a probability distribution?

No
5/7
기초방법

Yes
9가지 경우
V

Treat factor as fixed.

Fix

Only in the distribution of the random effects.

Estimate parameters of the distribution of the random effects.

In both the distribution and the realized values of the random effects.

Estimate parameters of the distribution of the random effects and calculate predictors of realized values of the random effects.

Making a decision

- effect들이 fixed / random 인지 결정할 때,

데이터 맥락, 수집방법, 데이터 추출 환경들은 factor 결정

- factor의 level들은 분포를 갖는 모집단에서 온 랜덤샘플로 고려?

① Yes! : effect들은 random effects로 고려해야 함

② No! : effect는 고정된 상수, fixed effect

→ 데이터 내 effects가 랜덤 샘플로 간주되는 effects의

분포에 관해 추론한다면 랜덤으로 고려.

추론이 모델 내 효과에 국한되면 fixed effect

- 확률 분포에서 factor의 level들이 있나?

데이터 내 factor의 level들이 랜덤샘플과 같다 라고 결론을 충분한 정보 있나?

① Yes! :

: factor를 random effect factor로 고려. 그 factor로 인한

variance component 추정. 데이터의 random effect 실현값에도

관심 있으면 특정 값에 대한 예측값 사용 가능

② No! :

: factor를 fixed effects factor로 고려. level들의 효과를

추정함. 추론의 범위 제한

- 랜덤 가정은 정규성 가정 수반 X

: 정규성 가정은 random effect에 대해 아워지지만, effect가

random 가정하는 것 다음에 만들어진 벌도의 가정이다.

→ Variance Components의 많은 추정 절차가 정규성 요구 X

But 추정 결과의 분포적 특징 조사하면 종종 random effects의

정규성 가정

7. Inference

통계분석 : 데이터 수집, 데이터 요약, 추론

데이터는 모집단에서의 샘플이고, 이 데이터로 모집단 추론

→ 확률에 의해 지지되는 추론의 결과

통계 사용의 목표

: 추론 (선회구간 포함), 검정, 예측

전통적인 선형 모형에서 추론 기초

- fixed effect : 최소 제곱 추정법

- random effect : 분산 추정 위해 분산 제곱합 분석

- 가설 검정, 신뢰구간 계산, 최선의 예측값과 예측구간 계산

: 정규성 가정 → LM, LMM들만 사용

- GLM / GLMM : 정규성과 다른 분포 가정해 기존 선형 모형보다

더 넓은 범위의 결과로 사용.

a. Estimation

- fixed와 Mixed Model에서, fixed effect와 선형함수 모두 추정

특히, 주어진 factor의 level들 사이의 차이.

[ex1]에서 풀라시보/아닌 편자들 사이에서의 밸런스 평균 차 추정

[ex2] 학습장애 有 / 無 사람들 사이 유무 이해 차이 추정

⊕ 사속 유형 만화 / 언어 유형 만화 / 시각 + 언어 차이

[ex3] 건조방법들 간 설포 부드러움 차이 추정

- 모형 일부 내 random effect 존재

: 요인 내 effect들의 분산, 공분산 추정

why? random effect의 설명 일부분으로 포함시키기 위해!

[ex4] 병원 effect의 분산 추정

: 연구 내 20개 병원 뿐만 아니라 전체 병원 모집단 내 분산추정치

- Mixed model

: fixed effect 추정 뿐만 아니라 random effect의 분산 추정

[ex6] 다양한 약물 복용 level 사이에서 밸런스 차이 추정

⊕ 병원들 사이 분산 추정

① Maximum Likelihood (ML)

- 주된 추정 방법

- y : 데이터 벡터, θ : y 분포 함수의 파라미터 일 때,

$f(y|\theta)$: θ 의 주어진 값에 대한 y 의 밀도 함수

→ θ 가 특정 값 갖는다면, $L(\theta|y) = f(y|\theta)$

이 때, $L(\theta|y)$ 는 우도 (likelihood)

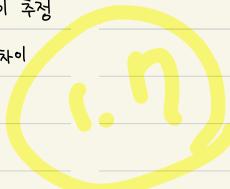
- ML은 θ 의 함수 $L(\theta|y)$ 를 최대화하는 θ 값 찾는 과정.

- 다루기 쉬운 밀도 함수라면, 간단한 과정이고, y 의 함수로서 θ 를

최대화 하는 단일 대수식을 산출

- But, 어려운 함수일 때 반복되는 수치 미분 요구,

θ 를 최대화하는 단일 값을 항상 산출 X :



☆ ??? ☆

fixed effect 브리?

② Restricted maximum likelihood (REML)

- ML과 관련된 방법은 Restricted (Residual) Maximum Likelihood

- y 의 선형 함수에 ML 적용한다는 idea. ($K'y$)

→ $K'y$ 가 y 의 모델 내 fixed effect 포함하지 않도록 설계? 무기 피상 T

ML에서 y 를 $K'y$ 로 바꾸면 REML

- REML은 선형 모형, 비선형 모형 모두 일반화 가능

- REML 사용 결과

1. Variance component & fixed effect에 의해 영향 받지

않고 추정됨. → 분산 추정치는 fixed effect 값에 따라 변화X

2. Fixed effect의 자유도 계산됨

ML은 불가능

ex) y_i 정규분포 데이터에서 σ^2 추정

$$y_i \sim N(\mu, \sigma^2) \quad i=1, \dots, n$$

$$\bar{y} = \frac{\sum y_i}{n}, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

→ σ^2 추정

$$\cdot \text{REML} = \frac{S_{yy}}{n-1} \quad \cdot \text{ML: } \frac{S_{yy}}{n}$$