Abstruct

정렬 시퀀스 데이터 셋에 fitting 하는 demographic model.

→ regions, loci 내에서 재조합 없이 coalescent history 를 갖는 데이터 가정

↓

가정에 대한 실패효과 완화.

random effect vs Fixed effect

→ 모수치가 고정되어 있느냐, 측정단위에 따라 변하느냐

Simple Regression

$y_i = b_0 + b_1 x_i + e_i \quad e_i \sim N(0, \delta^2)$ -①

여기서 b_0, b_1은 측정단위, i에 따라 다르다는 가정X

⇒고정 효과. (Fixed effect)

Random Effect

측정단위마다 b_0, b_1 다르다는 가정

= b_0, b_1에 대한 상위 분포 가정

ex) $b_{i,0} = m_0 + e_{i,0}$, $e_{i,0} \sim N(0, \tau_0^2)$ -②

i에 따라 b_0 달라짐.

⇒ $b_{i,0} \sim N(m_0, \tau_0^2)$ -③

정규분포일 필요는 없지만, 관습적으로 다루기 편함.

②를 ①에 대입

: $y_i = m_0 + e_{i,0} + b_1 x_i + e_i$ -④

⇒이는 절편이 모든 측정단위에 대해 같지X

→ Random Intercept model

b_1이 측정 단위별로 달라질 때,

$b_{i,1} = m_1 + e_{i,1} \quad e_{i,1} \sim N(0, \tau_1^2)$

$b_{i,1} \sim N(m_1, \tau_1^2)$

→ $y_i = b_0 + (m_1 + e_{i,1}) x_i + e_i$

이는 slope가 모든 측정단위에 대해 같지 않다

Random slope model

Random Intercept, random slope

$y_i = (m_0 + e_{i,0}) + (m_1 + e_{i,1}) x_i + e_i$

Mixed effect Model

: Fixed effect를 갖는 또 다른 독립변수 z_i 존재

$y_i = (m_0 + e_{i,0}) + (m_1 + e_{i,1}) x_i + b_2 \cdot z_i + e_i$

→ random effect와 Fixed effect 같이 있음.

Random Effect Model

기업마다 우연가 통제되지 않는 기업별효과가 있는데 이 효과를 하나의 확률변수로 모형화 하여 각 기업은 동일한 분포로 부터 기업별 효과를 추출해 그 값에 따라 기업의 성과가 결정된다고 보는 시각

Fixed Effect Model

확률변수가 따르는 분포가 워낙 지지대가 넓어서 그 확률분포로 부터 추출한 기업별 효과를 상호 비교 한 것이 전혀 의미없다고 생각하여 기업별 효과를 각각 별개의 더미변수로 처리

Within Variation 과 Between Variation

: TSS = ESS + RSS 의 관계식과 관련됨

TSS : 종속변수의 총 변동
ESS : 설명계수에 의해 설명되는 종속변수의 변동
RSS : 설명되지 않은 종속변수의 변동

• Within Variation

: 한 Entity 내에서 설명되는 변동 의미.

어떤 하나의 고정된 entity 내에서 설명변수 값이 변했을 때 한 entity 내 종속변수 값은 얼마나 변하나? 설명변수 변동으로 설명되지 않는 부분은 얼마나 되는가?

$\Rightarrow \sum_i \sum_t (y_{it} - \bar{y}_i)^2 = \hat{\beta} \sum_i \sum_t (x_{it} - \bar{x}_i)^2 + \sum_i \sum_t (u_{it} - \bar{u}_i)^2$

↳ $\bar{y}_i$는 개체들 사이의 차이 의미.
평차 형태로 표현하는 것은 개체 간 차이 제거.
개체 안에서의 시간에 따라 변화하는 차이만 바라보겠다.

• Between Variation

: entity들 사이의 변동.

A, B, C 사장의 평균 노동시간 차이는 평균 소득 차이를 얼마나 설명하는가?

즉, 각 entity들의 평균값들의 차이를 가지고 변동을 찾아내는 것이 between variation

$\Rightarrow \sum (\bar{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum (\bar{x}_i - \bar{x})^2 + \sum (\bar{u}_i - \bar{u})$

각 개체의 시간에 따른 변화값들을 평균하면 시간에 따른 요인 사라지고, 시간에 따라 변하지 않는 그 개체의 특수한 요인들만 남는다.

∴ 평균들 사이의 차이는 between variation 의미

Random effects Model

= Variance components model.

: 구조적 선형모형(hierarchical linear model) 일종.
population average를 fixed effect로 여기고, subject-specific effects를 random effect로 본다.

ex)
각 나라로부터 수많은 초등학교로부터 m개 선택.
선택된 학교에서 같은 나이 학생들 n명 임의 선택.

$$Y_{ij} = \mu + U_i + W_{ij}$$

Y_{ij} : i번째 학교에서 j번째 학생 점수
μ : 전체 학생들의 평균 시험점수 (학교 무관)
W_{ij} : 개인 특성(차이)에 따른 오차.
i번째 학교에 대한 평균으로 부터 j번째 학생의 점수에 대한 편차
→ 랜덤으로 추출가능.
(why? 어떤 학생에 대한 고정된 값이라 해도 학교안에서 학생들 또한 임의로 선택되기 때문이다.)

U_i : 학교별로 어떤 특징이 있다는 학교 특성 random effect로 볼 수 있다.
→ 전체 국가에서 평균 점수와 ith 학교의 평균 점수간의 차이로 볼 수 있다.
→ 이차이가 왜 random?
그 학교가 전체국가에서 학교가 랜덤하게 선택 되었기 때문

확장 : 관심변수 추가

$$Y_{ij} = \mu + \beta_1 Sex_{ij} + \beta_2 Race_{ij} + \beta_3 ParentsEduc_{ij} + U_i + W_{ij}$$

성별 : 남/여의 가변수.
인종 : 아시아/흑인/백인 의 가변수.
부모 교육 수준이 어느정도 영향 있지 않을까 수준
⇒ 관심변수 뿐과 개인특성, 학교특성에 관련된 혼합모형.

왜 랜덤효과 모형을 분산요소 모형 이라 부르나?

$Var(U_i) = \tau^2$
$Var(W_{ij}) = \sigma^2$
> ⊕ : 각 학교에서 각 학생들의 점수에 대한 분산

$\bar{Y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$, ith 학교에서 임의로 선택된 학생들의 평균 점수
이는 i번째 학교 전체 평균 (χ). Why? 랜덤 샘플.

$\bar{Y}_{\cdot\cdot} = \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}$. 전체 평균

그룹 내 차이에 대한 합의 제곱 (SSW, sum of Square within groups)
그룹 간 차이에서 발생되는 합의 제곱
↳ (SSB, sum of square between groups)

$\frac{1}{m(n-1)} E(SSW) = \sigma^2$
$\frac{1}{(m-1)n} E(SSB) = \frac{\sigma^2}{n} + \tau^2$

⇒ 이러한 기대 평균제곱 (Expected Mean Square)은 분산 요소인 σ^2, τ^2의 추정치의 근거로서 사용된것