데이터분석방법론(2) 제11강 과제

1. 다음 기술이 타당하면 O 틀리면 X로 표시하시오.

① 선형모형을 사용하여 상관된 정규분포 자료를 올바르게 분석할 수 있다.
→ O
② 선형혼합모형에서는 고정효과와 변량효과를 선형관계식으로 표현한다.
→ O
③ 변량효과는 미지의 상수값으로 일반적으로 추론의 대상이 된다.
→ X
④ 선형혼합모형의 분산-공분산 모수의 추정은 편의 측면에서 최대가능도추정보다 제한최대가능도추정이 우수하다.
→ O
⑤ 변량효과에 대한 가능도비 검정은 제한가능도 함수를 사용한다.
→ O

2. 선형혼합모형을 개별 관측값 표현식 및 행렬 표현식으로 기술하라.

개별 관측값 표현식

- $Y_{ti} = \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \cdots + \beta_p X_{ti}^{(p)} + u_{1i} Z_{ti}^{(1)} + u_{2i} Z_{ti}^{(2)} + \cdots + u_{qi} Z_{ti}^{(q)} + \epsilon_{ti}$
- 기호에 대한 정의 및 가정
  - Y_{ti} : i번째 개체에서 t시점에 관측된 반응변수
  - $\beta_1, \cdots, \beta_p$: 고정효과를 나타내는 모수
  - $X_{ti}^{(1)}, \cdots, X_{ti}^{(p)}$: 1을 포함하여 고정인자를 나타내는 공변량
  - $u_{1i}, \cdots, u_{qi}$: 변량효과를 나타내는 확률변수(정규분포 가정)
  - $Z_{ti}^{(1)}, \cdots, Z_{ti}^{(q)}$: 변량효과와 관련된 공변량
  - ϵ_{ti}: 오차항(정규분포 가정)
  - 변량효과와 오차항은 서로 독립
- 이 표현식에서
  - $\beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \cdots + \beta_p X_{ti}^{(p)}$ 은 고정요인을 모형화한 부분이고,
  - $u_{1i} Z_{ti}^{(1)} + u_{2i} Z_{ti}^{(2)} + \cdots + u_{qi} Z_{ti}^{(q)} + \epsilon_{ti}$ 은 변량요인을 모형화한 부분이다.

행렬 표현식

- $Y_i = X_i \beta + Z_i u + \epsilon_i, \quad i = 1, 2, \cdots, m$
- 기호에 대한 정의
  - $Y_i = \begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}$
  - $X_i = \begin{pmatrix} X_{1i}^{(1)} & \cdots & X_{1i}^{(p)} \\ \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & \cdots & X_{n_i i}^{(p)} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$
  - $Z_i = \begin{pmatrix} Z_{1i}^{(1)} & \cdots & X_{1i}^{(q)} \\ \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & \cdots & Z_{n_i i}^{(q)} \end{pmatrix}, u = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix}$
- 이 행렬 표현식에서
  - 변량효과 u_i의 분산-공분산 행렬 D는 다음과 같다.

      - $D = Var(u_i) = \begin{pmatrix} Var(u_{1i}) & \cdots & Cov(u_{1i}, u_{qi}) \\ \vdots & \ddots & \vdots \\ Cov(u_{1i}, u_{qi}) & \cdots & Var(u_{qi}) \end{pmatrix}$
    - 오차항 ϵ_i의 분산-공분산 행렬 R_i는 다음과 같다.
      - $R_i = Var(\epsilon_i) = \begin{pmatrix} Var(\epsilon_{1i}) & \cdots & Cov(\epsilon_{1i}, \epsilon_{n_i i}) \\ \vdots & \ddots & \vdots \\ Cov(\epsilon_{1i}, \epsilon_{n_i i}) & \cdots & Var(\epsilon_{n_i i}) \end{pmatrix}$

        $Cov(u_i, \epsilon_i) = 0$ (즉, u_i와 ϵ_i는 서로 독립)

3. 변량효과 검정법을 요약하라.

- 하향식 모형구축전략은 다음의 네 가지 단계를 통해 모형을 구축하는 전략이다.
  - [1단계] 가능한 꽉 찬 평균모형 $E(Y_{ij}) = X_{ij}\beta$ 에서 출발
    - 교호작용 효과를 포함하여 가능한 많은 수의 고정효과를 모형에 포함
      - 공분산 구조를 탐구하기 전에 종속변수에 내포된 체계적인 변동성이 최대한 설명되도록 함
  - [2단계] 변량효과와 공분산 구조(D)를 선택
    - 모형에 포함될 변량효과와 공분산 구조를 선택
      - 변량효과와 공분산 구조에 대하여 REML 기반 가능도비 검정으로 유의성을 검정
  - [3단계] 오차 공분산 구조(R_i)를 선택
    - 적합해 보이는 몇 가지 후보 공분산 구조를 가정
      - REML 기반 가능도비 검정(검정 모형 간에 지분관계가 성립할 때) 또는 정보량 기준(검정 모형 간에 지분관계가 성립하지 않을 때)으로 가장 적절한 공분산 구조를 선택
  - [4단계] 평균 모형의 축소
    - 적절한 검정법(근사적인 F-검정 또는 T-검정, 가능도비 검정 등)으로 고정효과의 유의성을 검정

- 문제에서 이야기하는 변량효과 검정법은 하향식 모형구축 전략 2단계에서 언급된 변량효과와 공분산 구조에 대한 REML 기반 가능도비 검정을 의미한다고 생각되며, REML 기반 가능도비 검정에 대한 내용은 다음과 같다.
- 변량효과의 가능도비 검정을 위해서는 두 모형을 설정해야 한다.
  - 모형1: 변량효과를 포함한 모형
  - 모형2: 변량효과를 포함하지 않은 모형
  - 모형1과 모형2는 중첩된 모델이어야 하며, 모형1이 모형2보다 더 복잡한 모형이다.
- 그 다음으로 각각의 모형에 대해서 REML 가능도함수를 사용해서 가능도 값(L_1, L_2)를 구한다.
  - 가능도비 검정 통계량 $G^2 = -2 \cdot log(\frac{L_2}{L_1})$
  - 이 검정통계량(G^2)은 카이제곱 분포를 따르며, 자유도는 두 모델의 파라미터 수의 차이이다.
- 검정 결과, 검정통계량 값이 충분히 크고, 해당 값이 카이제곱 분포에서 유의한 값을 가지면 변량효과가 유의하다고 판단할 수 있다. 만약 그렇지 못하다면 변량효과는 유의하지 않다고 판단한다.