다중 수준 분석
Multi-Level Analysis

울산의대 예방의학교실 이무송
2004년 7월 9일

순서

- 개요
  - 정의 / 개념 및 수식
  - 적용 영역 / 장단점
  - 특기 사항
  - 주요 자료
    - Statistical and Substantive Inferences in Public Health: Issues in the Application of Multilevel Models. Bingenheimer JB, Raudenbush SW. Ann Rev Public Health 2004;25:53-77. 등
- 응용 예: 상관성 연구 등
  - 프로그램 사용 예: SAS, MLWin

배경

- 교육학 등 사회학적 연구에서 활용됨
  - 지난 10년간 의학연구에 소개되고 발전, 주로 사회역학연구
- 초기 채택자의 지나친 열성 *vs.* 전혀 새로운 것이 없다는 비판
- 전통적인 단일 수준(single-level) 모델보다 유용한 통계적 추론
  - 기존 통계분석(random-effects model 등)과 동일
  - 개념적 이해 및 해석이 용이
- 활용 영역
  - 서로 다른 수준에서 중재조치의 효과
    - 군집 무작위배정 시험(cluster randomized trial)
  - 질병 위험요인의 다중 수준(multilevel) 인과성
  - 의료 제공자의 상대적 수행 능력을 평가

단순 다중 수준 모델

- 독립변수 x_{ij}
  - 개인의 성별
  - 해당 개인 거주지역(neighborhood) 내 패스트푸드점의 존재 여부
- 종속변수 y_{ij}
  - 개인의 체질량지수, j번째 거주지역에 사는 i번째 개인의 체질량지수
    - 각 거주지역에서 체질량지수: 지역특수 평균 β_{0j} 분산 σ^2인 정규분포
    - 지역 특수 평균은 평균 γ_{00} 분산 τ_{00}인 정규분포

2-수준 모델

$$Y_{ij} = \beta_{0j} + r_{ij} \quad 1a$$
$$\beta_{0j} = \gamma_{00} + u_{0j} \quad 1b$$
$$r_{ij} \sim (i)\,N(0,\sigma^2),\quad u_{0j} \sim (i)\,N(0,\tau_{00}),\quad \mathrm{cov}(r_{ij},u_{0j}) = 0 \quad 1c$$

- 1*a*: 각 지역 내 체질량지수의 변이, 수준-1 모델
- 1*b*: 지역 간 체질량지수의 변이, 수준-2 모델
- 1*c*: 분산-공분산 구조

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad 2$$

- γ_{00}: 전체 평균, u_{0j}: 지역간 변이, r_{0j}: 개인의 확률 변이
- 공변량이 없음
- 일원 확률효과 분산분석(random-effects ANOVA)과 동일

변이의 분할

- Y_{ij}의 전체 변이 = $\sigma^2 + \tau_{00}$
  - τ_{00} : 지역간 변이에 기인한 변이
- ICC(intracluster correlation coefficient) ρ
  - 체질량지수의 변이 중 (지역 내 변이가 아닌) 지역간 변이가 차지하는 분율

$$\rho = \frac{\tau_{00}}{\sigma^2 + \tau_{00}} \quad 3$$

확장: 수준-1 공변량 X_{ij}

- X_{ij}: 0(여자), 1(남자)

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad 4a$$
$$\beta_{0j} = \gamma_{00} + u_{0j} \quad 4b$$
$$\beta_{1j} = \gamma_{10} + u_{1j} \quad 4c$$
$$r_{ij} \sim (i)\,N(0,\sigma^2),\quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim (i)\,N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right] \quad 4d$$

- 수준-2 모형: 4*b*, 4*c*
- 한 지역에서 여자의 평균 체질량지수= β_{0j}
- 한 지역에서 남자의 평균 체질량지수= $\beta_{0j} + \beta_{1j}$
  남녀 차이의 평균= β_{1j}

확장: 수준-1 공변량 X_{ij}

- 지역간으로 넓혀 보면, 평균 체질량지수
  - 여자= γ_{00}, 남자= $\gamma_{00} + \gamma_{10}$, 평균 차이= γ_{10}
  - 남녀 차이는 지역에 따라 일정하지 않다.
    - 차이는 평균 γ_{10}, 분산이 τ_{11} 인 정규분포
    - τ_{11} 인=1인 경우 남녀 차이는 모든 지역에서 일정

확장: 수준-2 공변량 W_j

- W_j: 지시 변수(indicator)
  - 1(j번째 지역에 패스트푸드점이 있음, 1(없음)

$$Y_{ij} = \beta_{0j} + r_{ij} \quad 5a$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad 5b$$
$$r_{ij} \sim (i)\, N(0, \sigma^2), \quad u_{0j} \sim (i)\, N(0, \tau_{00}), \quad \mathrm{cov}(r_{ij}, u_{0j}) = 0 \quad 1c$$

  - 5a: 각 지역 내에서 체질량지수는 평균 β_{0j}, 분산 σ^2인 정규분포
  - 5b: 패스트푸드점이 없는 지역에서 체질량지수의 평균은 γ_{00} 주변
  - 있는 지역에서는 $\gamma_{00} + \gamma_{01}$ 주변에 위치

확장: 수준-1 + 수준-2 공변량

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad 6a$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad 6b$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad 6c$$
$$r_{ij} \sim (i)\, N(0, \sigma^2), \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim (i)\, N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right] \quad 4d$$

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j + u_{0j} + u_{1j} X_{ij} + r_{ij} \quad 7$$

- 두 수준간 상호작용: $\gamma_{11} X_{ij} W_j$, 식 6c에서 W_j에 해당

상호작용의 해석

- 남녀의 평균 (체질량지수의) 차이는 해당 지역에 패스트푸드점이 있는지에 따라 다르다.
  - 즉 없는 지역에서는 γ_{10}, 있는 지역에서는 $\gamma_{10} + \gamma_{11}$
- 패스트푸드점의 존재 여부에 따른 지역간 차이는 개인의 성별에 따라 다르다.
  - 즉 여성의 경우 존재하는 지역에서 없는 지역에 비하여, γ_{01}의 차이가 있다.
  - 남자는 $\gamma_{01} + \gamma_{11}$의 차이가 있다.
- $\gamma_{11} W_j$를 없애면 상호작용이 없어진다.
  - random slope model이 아니다.
- u_{0j}를 없애면 절편이 일정해진다.
  - random intercepts model이 아니다.

다른 상황에로의 일반화

- 여러 사람(수준 2)에서 시간(수준 1)에 따른 반복측정 데이터
  - 개인 성장곡선의 추정, 경시적 자료
  - 시간에 따라 변화하는 공변량
- Y_{ij}가 정규분포 이외의 분포를 따를 때
  - 이분성, 범주형, 순서형, 횟수(count), 사건 발생까지의 시간
  - 이분성 분포, 포아송 분포, 생존시간의 분포
  - 예) 비만 여부: Bernoulli 분포 + 로지스틱 link
- 수준이 세 개 이상
  - 예) 여러 지역에서 여러 개인에 대해 반복측정: 지역, 개인, 시간

모수의 종류

- Microparameter: β_{0j}, β_{1j} - 수준-1 계수
- Macroparameter: γ_{00}, γ_{01}, γ_{10}, γ_{11} - 수준-2 계수
- Random effects: u_{0j}, u_{1j}
- Variance components: σ^2, τ_{00}, τ_{11}, $\tau_{01}= \tau_{10}$

구체적 활용 영역
cluster randomized trial

- 집단-수준 중재조치의 효과
  - 무작위배정: 집단(도시, 학교, 학급, 작업장) 수준
  - 결과변수: 개인 수준
- 질문: 중재조치에 의해 결과변수에 차이가 있는가?
- 분석: 식 5*a*, 5*b* - γ_{01}로 평가
- 다중 수준 분석의 필요성
  - 개인 수준으로만 분석: 1종 오류가 증가
    - ← (동일한 군집 내에서 개인들의 의존성을 무시 → γ_{01} 추정치의 정밀도 과장)
  - 집단 수준으로만 분석
    - *군집별 평균을 각각 구한* 후, 이들을 종속변수로 간주하여 분석
    - 이때 개인 수준의 공변량을 보정: 예) 연령, 성별을 표준화한 SMR

cluster randomized trial에서 ML의 장점

- 두 단계를 동시에 수행
  - 개인-수준 공변량은 수준-1 모델에서 적절 보정
- 개인별 공변량에 의해 정의된 소집단 간의 치료효과를 검정
  - 예) 남자와 여자의 치료 효과가 동일한가? - 상호작용 평가
- 유연한 분석이 가능하고, 검정력이 증가

다중 수준 인과성의 평가

- 지역과 같은 집단 특성 변수가 건강에 미치는 영향은?
  - 과거 개인 특성 변수(생활습관 등)이 건강에 미치는 영향을 주로 평가
  - 집단 특성 변수의 평가가 쉽지 않다.
    - 주로 생태학적 관련성을 평가하는 문제이기 때문
    - 통계적인 처리가 개념적으로 쉽지 않다.
      - random effects model, nested structure
- 분산의 분할
  - 주요 건강 지표의 지역에 따른 변동이 어느 정도인가?
    - 식 1에서 variance components τ_{00}로 정량화
    - 귀무가설(τ_{00}=0, 지역간 변이가 없다.)의 검정: 검정력은 낮다.
    - 연속형 결과변수의 경우, σ^2의 추정이 가능하며, ICC의 추정이 가능

Context and Composition

- 관찰된 지역간 변이 중 거주자의 특성에 의한 부분은?
  - Context and composition
  - 개인-수준 공변량 $X_{1ij}, \ldots, X_{kij}$ 을 모델에 포함
  - 회귀계수를 고정: $\beta_{1j}=\gamma_{10}, \ldots, \beta_{kj}=\gamma_{k0}, \tau_{11}=\ldots=\tau_{kk}$
  - 얻어진 수준-2 variance component 추정치인 τ^*_{00}
    - 개인 수준 공변량에 의한 변이가 제거된 추정치
  - 귀무가설 $\tau^*_{00}=0$을 검정하거나, 비조건부 모델에서 얻어진 τ_{00} 와 비교
    - 유의하다면, 지역간 변이의 일부는 지역의 composition(개인별 특성)이 아니라 지역 특성에 기인하였다는 의미
  - τ^*_{00}/τ_{00}: 지역간 변이 중 지역-수준 요인에 기인한 분율

Context and Composition의 상호 교란: 해결책

- 2단계 분석
  - 식 4a-4c를 사용 + 기울기를 고정($u_{1j}=0$) + 개인-수준 공변량을 집단-평균에 대해 중앙화(group-mean centering the individual-level covariates)

$$Y_{ij}=\beta_{0j}+\beta_{1j}(X_{ij}-\bar{X}_{.j})+r_{ij} \qquad 4a$$

$$\beta_{0j}=\gamma_{00}+u_{0j} \qquad 4b$$

$$\beta_{1j}=\gamma_{10} \qquad 4c$$

$$r_{ij}\sim(i)\,N(0,\sigma^2);\ u_{0j}\sim(i)\,N(0,\tau_{00});\ \mathrm{cov}(r_{ij},u_{0j})=0 \qquad 1c$$

  - $\bar{X}_{.j}$ 는 공변량 X_{ij}의 지역-특수 평균
    - 이 공변량은 지역-수준 공변량 W_j와 독립이므로, γ_{10}의 추정치는 X_{ij}의 지역내 효과의 평균에 해당

Context and Composition의 상호 교란: 해결책

- 2단계 분석
  - 보정된 종속변수를 구성 $Y^*_{ij}=Y_{ij}-\hat{\gamma}_{00}-\hat{\gamma}_{10}X_{ij}$
  - 보정된 종속변수로 모델을 적합화
  - γ^*_{00}은 1단계에 포함된 모든 개인-수준 공변량을 보정한 후, 지역-수준 공변량에 기인한 변이.

지역-수준 변수 W_j의 평가

- 주의
  - 수준-1 모델에서 개인-수준 공변량의 처리
    - 교란(confounding)의 문제
      - 특정 지역에 거주하게 되는 이유 + 건강에 영향
      - 예: 개인의 경제상태 / 모델에 포함하는 것이 바람직
    - 중재(mediation)의 문제
      - 지역-수준 변수가 건강에 영향을 주는 기전(경로) 상에 위치
      - 예: 흡연이 많은 지역 → 개인의 음주 수준 → 간장 질환
      - 모델에 포함하면 과보정의 문제
  - 모델에 포함할 지역-수준 변수의 선택
    - 센서스로 평가한 지역-수준 변수 등은 상호 상관성이 매우 높음
    - 1개 변수만 사용 / 가능한 여러 변수를 사용: 추정치의 정밀도가 감소
    - 결합 지표를 사용: 해석이 어려워짐
      - dishonest specificity, honest ambiguity

Microparameters

- j 군집 각각에서의 microparameter, β_{0j}, β_{1j}가 관심사인 경우
  - 각 군집의 평균 체질량지수, 발생률 등
- 실증적 베이즈 추정치(empirical Bayes estimator) 사용
  - 식 1*a*-1*c*에서
  - j 번째 지역에서의 두 가지 β_{0j} 추정치
    - 지역-특수 표본 평균 [illegible]
    - 전체 평균 [illegible]
    - 최적의 추정치: 이 두 개의 가중 평균 [illegible]
      - λ_j 신뢰성 지수, 지역-특수 표본평균에서 설계 점수와 전체 점수 분산의 비
      - 지역의 크기가 클수록 크다
      - 지역의 크기가 작으면 이 추정치는 다른 지역에서 얻어진 정보를 빌려 쓴다.
      - 데이터에서 신뢰성 지수를 추정하므로, 실증적 베이즈 추정법이라 한다.

실증적 베이즈 추정치의 사용

1) 소규모 인구집단의 발생률 추정: 발생이 드문 사건
   - 관찰된 발생률만으로 추정하면 불안정한 추정치가 얻어짐
   - 다른 모든 군집의 평균율(population average rate)를 사용하면, 특정 지역의 발생 양상을 왜곡할 가능성이 있음

2) 의료제공자(health service provider)의 상대적 수행능력 평가
   - 수행능력 지표: 특정 수술의 의사- 병원- 수준의 사망률
   - 신뢰도: 체계적 요소 + 우연적 요소
     - 관찰된 사망률은 두 요소를 모두 반영
     - 우연적 요소가 너무 큰 경우 추론이 불가: 실증적 추정치 사용
   - 위험도 보정: patient mix - 다층 수준 모델에 포함하여 보정

3) SMR이 높은 지역을 찾아냄

실증적 베이즈 추정치의 제한점

1) 적절한 개인-수준 공변량의 파악
2) 안정성이 충분히 확보되지 않을 수도 있음
3) 바이어스: 불편 추정치 [illegible]
   - 신뢰도 계수가 큰 지역에서는 바이어스가 적다.
   - 표본 크기에 따라 바이어스의 정도가 다르다.
     - 순위가 변화할 수 있다.
     - 규모가 큰 + 수행능력이 높은 제공자 vs.
     - 규모가 작은 + 수행능력이 낮은 제공자

사용 가능한 통계 패키지

- Random effects (+ fixed effects ≡ mixed effects) 모델의 분석이 가능한 패키지
  - SAS, SPSS, SPLUS, GENSTAT, ...
- 전용 패키지
  - MLWin
  - Flexibility가 떨어짐
  - 대규모 자료 분석 능력이 떨어짐
  - 사용이 용이

예제

- 정규분포를 따르는 결과(종속)변수의 다중-수준 분석 예
  - MLWin을 이용한 실제 자료 분석 demonstration
  - 해석 등
    - ICC 추정치, 모형 체크 등
  - SAS 프로그램과 비교
- 포아송 분포를 따르는 결과변수의 경우
  - SMR을 이용한 정규분포 근사: MLWin 및 SAS
- 사회역학 자료 분석에 사용한 경우
  - 자료 분석 결과 및 해석