

## 10강. 분할표 및 도수자료에 대한 로그선형 모형

### ■ 주요용어

용어	해설
동질연관성 모형	어떤 두 변수 사이의 조건부오즈비는 나머지 변수의 모든 수준에서 동일한 경우를 말함
독립성 그래프 (independence graph)	독립성 그래프의 각 꼭지점은 변수를 나타내며, 분할표의 차원의 수만큼 꼭지점들이 있고 두 꼭지점은 선분으로 연결되거나 연결되지 않음. 선분으로 연결되지 않은 경우는 꼭지점에 대응하는 두 변수간의 조건부 독립성을 나타냄
붕괴가능성 조건	삼차원 분할표에서 $Z$ 와 $X$ 가 조건부 독립이거나 $Z$ 와 $Y$ 가 조건부 독립이면 $XY$ 의 주변 오즈비와 부분 오즈비는 서로 같음

### 정리하기

#### 1. 분할표 자료에 대한 로그 선형모형

- 로그 선형모형은 모든 변수들을 반응변수로 간주하여 분석함
- $n$ 개 관측값을 두 개 범주형 변수에 따라 교차분류하여  $I \times J$  분할표를 작성한 경우
  - ⇒ 행변수  $X$ 와 열변수  $Y$ 가 서로 독립
    - ⇒  $\pi_{ij} = \pi_{i+} + \pi_{+j}$ ,  $i = 1, 2, \dots, I$ ,  $j = 1, 2, \dots, J$
    - ⇒ 칸 확률  $\{\pi_{ij}\}$ 는 다항분포(multinomial distribution)의 모수
    - ⇒  $\{\pi_{ij}\}$  대신 기대도수  $\{\mu_{ij} = n\pi_{ij}\}$ 를 사용 (포아송추출모형)
    - ( $X$ 와  $Y$ 의 독립성 가정  $\Leftrightarrow$  모든  $i, j$ 에 대하여  $\mu_{ij} = n\pi_{i+}\pi_{+j}$ )

## 2. 2차원 분할표에 대한 독립성 로그선형모형

- 행 변수를  $X$ , 열 변수를  $Y$ 로 표시할 때

$$\cdot X \text{와 } Y \text{가 독립} \Leftrightarrow \mu_{ij} = n\pi_{i+}\pi_{+j} \quad (\mu_{ij} = n\pi_{ij})$$

$$\Leftrightarrow \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \quad (\text{독립성로그선형모형})$$

$\lambda_i^X$  : 변수  $X$ 의  $i$ 번째 수준의 효과

$\lambda_j^Y$  : 변수  $Y$ 의  $j$ 번째 수준의 효과

- 두 범주형 변수의 독립성  $\Leftrightarrow$  로그 선형모형의 적합성

- $\cdot X^2$ 와  $G^2$ (독립성검정)를 이용하여 로그 선형모형의 적합도 판단

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \quad G^2 = 2 \sum n_{ij} \log \left( \frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

$\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ ,  $X^2$  와  $G^2$ 는 자유도  $(I-1)(J-1)$  인 카이제곱 분포로 근사

## 3. 이차원 분할표에 대한 포화모형

- 독립성을 만족하지 않는 경우의 로그 선형모형 적합

$$\cdot \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

여기서  $\{\lambda_{ij}^{XY}\}$  는 독립성에서 벗어난 정도를 반영하는 연관성을 나타냄

- $\cdot$  이차원 분할표에 대한 포화 로그 선형모형(saturated loglinear model)이라고 함

## 4. 삼차원 분할표에 대한 로그 선형모형

- $\{\mu_{ijk}\}$  : 삼차원 분할표에서 칸 기대도수(cell expected frequency)

- 포화모형: 삼차 교호작용항을 포함하며 자료를 완전하게 적합함

$$\cdot \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} : (XYZ) \text{로 표시함}$$

- 동질연관성모형

$$\cdot \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} : (XY, XZ, YZ) \text{로 표시함}$$

- $\cdot$  어떤 변수 사이의 조건부 오즈비는 나머지 변수의 모든 수준에서 동일함

- $X$ 와  $Y$  사이의 조건부 독립성 모형

$$\cdot \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} : (XZ, YZ) \text{로 표시함}$$

- 상호독립모형

$$\cdot \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z : (X, Y, Z) \text{로 표시함}$$

## 5. 카이제곱 적합성 검정

- 칸의 적합값과 관측값을 비교함으로써 모형 적합도를 검정함
  - 가능도비 통계량(모형의 이탈도: Deviance):  $G^2 = 2 \sum n_{ijk} \log \left( \frac{n_{ijk}}{\widehat{\mu_{ijk}}} \right)$
  - 피어슨 통계량:  $X^2 = \sum \frac{(n_{ijk} - \widehat{\mu_{ijk}})^2}{\widehat{\mu_{ijk}}}$
  - 자유도 = (총 칸의 개수) - (추정할 모형의 모수 개수)
  - 포화모형의 자유도는 0이고, 자유도는 모형이 복잡할수록 줄어듦

## 6. 로그 선형모형의 해석을 위한 로지스틱모형의 사용

- 삼차원 분할표에서 동질 연관성 모형:
$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$
- $Y$ 를 이항형 반응변수로 가정하고,  $X$ 와  $Z$ 를 설명변수로 가정
  - $X$ 의 수준이  $i$ 이고  $Z$ 의 수준이  $k$ 라고 할 때  $\text{logit}[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z$
- 로그 선형모형은 적어도 두 변수들이 반응변수일 때 사용하는 것이 바람직하고, 반응변수가 하나인 경우는 로지스틱 모형을 적합하는 것이 좋음

## 7. 독립성 그래프와 붕괴가능성 조건

- 로그선형모형에서 그래프를 이용하여 조건부 독립성을 표현하는 방법
- 로그선형모형의 독립성 그래프는 꼭지점의 집합이며 각 꼭지점은 변수를 나타냄
- 선분으로 연결되지 않는 경우는 꼭지점에 대응하는 두 변수의 조건부 독립성을 뜻함
- 삼차원 분할표에서 붕괴가능성 조건(Collapsibility condition):  
삼차원 분할표에서  $Z$ 와  $X$ 가 조건부 독립이거나  $Z$ 와  $Y$ 가 조건부 독립이면  $XY$ 의 주변 오즈비와 부분 오즈비는 서로 같음

[과목명] 10강. 분할표 및 도수자료에 대한 로그 선형모형

	<b>과제하기</b>
--	-------------

구분	내용
과제 주제	- 박태성 & 이승연 (2020) 285쪽 문제 7.2 - 박태성 & 이승연 (2020) 286쪽 문제 7.3
목적	10주차 강의 내용을 복습하고, 로그 선형모형을 실제 데이터에 적용함으로써 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	12주차 강의 후 1주 후 일요일 밤 12시까지
참고 자료	
기타 유의사항	