09강. 로지스틱회귀모형 [4]

■ 주요용어

용어	해설
다범주 로짓모형	명목형 반응변수 Y가 범주 $1, 2, \cdots, J$ 를 갖는 경우 임의로 하나의 기준범주(baseline-category)를 선택한 후 이 범주와 나머지 각 반응범주와 짝을 지어 로짓을 정의함
비례오즈 누적 로짓모형	예측변수 X에 대하여 다음 모형을 가정함. $\log[P(Y \le j)] = \alpha_j + \beta x,\ j = 1, 2, \cdots, J-1$ 각각의 누적 로짓에 대해서 서로 다른 절편 α_j와 같은 기울기 β를 가정함
조건부 독립성	제3의 변수(Z)를 통제한 상태에서 두 변수 X, Y가 서로 독립일 때 조건부 독립이라고 함
맥니마 검정(Mcnemar test)	대응쌍 자료에서 주변동질성$(\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21})$ 을 만족하는가를 검정하기 위한 검정법

정리하기

1. 명목형 반응의 로짓모형(다범주 로짓모형)

- 명목형 반응변수 Y가 범주 $1, 2, \cdots, c$ 를 갖는 경우($c > 2$)
- 각 범주에 대응하는 반응확률 : $\{\pi_1, \pi_2, \cdots, \pi_c\}$, $\sum_j \pi_j = 1,\ \pi_j = P(Y = j)$
- n명의 관측치를 c개 범주에 할당시키는 표본모형은 다항분포(multinomial distribution)를 따름
- 기준범주를 이용한 로짓모형
  ① 임의로 하나의 기준범주(baseline-category)를 선택한 후 이 범주와 나머지 각 반응범주와 짝을 지어 로짓을 정의함
  ② 기준범주 로짓 (마지막 범주 c가 기준일 때)
  ⇨ 예측변수 x를 가진 기준범주 로짓모형

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x,\ j = 1, 2, \cdots, c-1$$

  "각 로짓에 대해서 서로 다른 모수 (α_j, β_j) 가정"

  ③ $\exp(\hat{\beta})$는 반응범주 c에 대한 반응범주 j의 오즈에 예측변수 x가 1단위 증가함으로써 나타나는 승법효과임

2. 순서형 반응변수들에 대한 누적 로짓 모형

- 반응범주들이 순서형인 경우는 순서를 고려한 로짓을 정의할 수 있음

  ⇨ 순서를 고려한 로짓 모형은 해석이 간단하고 보통의 다범주 로짓 모형보다 더 좋은 검정력을 갖게 됨

- 누적확률(cumulative probability): $P(Y \le j) = \pi_1 + \cdots + \pi_j,\ j = 1, 2, \cdots, c$
- 누적 로짓(cumulative logit)

$$\mathrm{log}it[P(Y \le j)] = \log\left[\frac{P(Y \le j)}{1 - P(Y \le j)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right]$$

3. 비례오즈 누적 로짓 모형

- 예측변수 X에 대하여 $\log[P(Y \le j)] = \alpha_j + \beta x,\ j = 1, 2, \cdots, c-1$
- "$1, \cdots, j$ 의 범주들을 하나의 범주로 합하고, $j+1$부터 c까지의 범주를 다른 하나의 범주로 보는 로지스틱 회귀모형과 비슷함"
- 각각의 누적 로짓에 대해서 서로 다른 절편 α_j와 같은 기울기 β를 가정함
- $\dfrac{odds\,of\,(Y \le j)\,at\,a}{odds\,of\,(Y \le j)\,at\,b} = e^{\beta(a-b)}$

⇨ 어떤 주어진 범주 이하의 반응에 대한 오즈는 x가 한 단위 증가하면 e^{β}배만큼 증가함

⇨ 비례오즈 모형(proportional odds model)

- $\beta = 0$ 만족 ⇔ X와 Y는 통계적으로 독립

4. 조건부 독립성의 개념

- 조건부 독립성 : 제3의 변수(Z)를 통제한 상태에서 두 변수 X, Y가 서로 독립일 때 조건부 독립이라고 함

5. 대응쌍 자료의 주변동질성 개념

- 동일한 대상에 대해서 두 번의 조사를 한 경우나 한 표본의 개체와 다른 표본의 개체간에 자연스러운 짝 관계(pairing, 쌍)가 있는 경우에 만들어진 대응쌍 자료의 분석방법
- 주변동질성 검정의 문제를 다룸

6. 맥니마 검정(Mcnemar test)

- H_0 : 주변동질성 만족 $(\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21})$ $\Leftrightarrow$ $\dfrac{\pi_{12}}{\pi_{12}+\pi_{21}} = \dfrac{1}{2}$

  ⇨ "대응쌍 자료에서 귀무가설을 만족하는 경우에는 n_{12}와 n_{21}은 같은 기대도수를 갖게 됨"

- $n^* = n_{12} + n_{21}$으로 정의하면 귀무가설이 성립할 때

  $$n_{12} \sim B(n^*, \frac{1}{2}), \quad E(n_{12}) = n^*/2, \quad \sqrt{Var(n_{12})} = \sqrt{n^*(\frac{1}{2})(\frac{1}{2})}$$

- 검정통계량(맥니마 검정)

  $$Z = \frac{n_{12} - n^*/2}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} \sim N(0,1) \quad \text{또는} \quad Z^2 = \frac{(n_{12}-n_{21})}{n_{12}+n_{21}} \sim \chi_1^2$$

  $$= \frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}}$$

- $\pi_{1+} - \pi_{+1}$ 에 대한 신뢰구간 작성

  · $\pi_{1+} - \pi_{+1}$ 의 추정량 : $p_{1+} - p_{+1}$

  · $SE = \sqrt{\widehat{Var}(p_{1+} - p_{+1})}$
  $$= \frac{1}{n}\sqrt{(n_{12}+n_{21}) - \frac{(n_{12}-n_{21})^2}{n}}$$

  · 95% 신뢰구간 : $(p_{1+} - p_{+1}) \pm 1.96 \times SE$

과제하기

구분	내용
과제 주제	- 박태성 & 이승연 (2020) 246쪽 문제 6.16
목적	9주차 강의 내용을 복습하고, 로지스틱회귀모형을 실제 데이터에 적용함으로써 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	11주차 강의 후 1주 후 일요일 밤 12시까지
참고 자료	
기타 유의사항	