05강. 로지스틱회귀모형 [1]

■ 주요용어

용어	해설
포화모형(saturated model)	각 관측값에 대하여 각각 모수를 갖는 경우로 $\hat{\mu}_i = y_i$을 만족하는 모형을 말함. 포화모형은 주어진 데이터에 대해서 가정할 수 있는 가장 복잡한(가장 모수가 많은) 형태의 모형으로 가능도함수(로그가능도 함수)의 최대값을 가짐
이탈도(deviance)	L_M을 모형 M에 대하서 구한 로그가능도 함수의 최대값이라고 하고, L_S를 포화모형에 대해서 구한 로그가능도 함수의 최대값이라고 할 때 이탈도(Deviance)$=-2[L_M - L_S]$로 정의함
피어슨 잔차	$e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}}$로 정의됨
표준화잔차	표준화잔차$= \dfrac{y_i - \hat{\mu}_i}{SE\ of\ (y_i - \hat{\mu}_i)} = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)(1-h_i)}}$
그룹화된 자료	데이터가 분할표와 같은 형식으로 요약된 자료를 말함
비그룹화된 자료	분할표 형식으로 요약되지 않은 원시 데이터를 말함

정리하기

1. GLM 모수에 대한 추정
   - 범주형 자료에 대한 대부분의 GLM에 대해서 모수추정은 ML 추정법을 통해서 이루어짐
   - 모수 β에 대한 95% 신뢰구간(Wald 방법): $\hat{\beta} \pm Z_{\alpha/2}(SE)$
   - $H_0 : \beta = 0$에 대한 가설검정
     ▸ Wald 방법: $Z = \dfrac{\hat{\beta}}{SE} \sim N(0,1)$을 따름을 이용하거나, $Z^2 = \left(\dfrac{\hat{\beta}}{SE}\right)^2$이 근사적으로 χ_1^2을 따름을 이용하여 검정함
     ▸ 가능도비를 이용한 방법: 가능도비(likelihood ratio) 검정통계량 이용

2. 이탈도(Deviance)
   - 포화모형: 주어진 관측 데이터에 대하여 적용 가능한 가장 복잡한(가장 모수가 많은) 형태의 모형으로 가능도함수(또는 로그가능도함수)의 최대값을 가짐
   - 데이터 분석에서 고려하고 있는 모형을 M이고, 포화모형을 S라고 할 때 이탈도는 다음과 같이 정의함
     - ▸ L_M : 모형 M에 대해서 구한 로그가능도함수의 최대값
     - ▸ L_S : 포화모형 S에 대해서 구한 로그가능도함수의 최대값
     - ▸ 검정할 귀무가설과 대립가설
       H_0 : 모형 M을 따름 vs H_1 : 포화모형 S를 따름
       "귀무가설 H_0는 포화모형에 있는 모수 중 모형 M에 포함되지 않은 모수들은 모두 0임을 뜻함"
     - ▸ 이탈도(Deviance)$=-2[L_M - L_S]$
     - ▸ 이탈도는 자유도 $df=$(자료 수 - 모형 모수의 개수)인 χ^2 분포로 근사됨
     - ▸ 검정통계량 값이 크고 p-값이 작을수록 모형 M의 적합결여에 대한 강한 증거임

3. 관측치와 모형 적합값을 비교하는 잔차
   - 잔차 분석: 관측도수와 적합값을 비교해 봄으로써 더 자세한 정보를 알 수 있음
     - ▸ 잔차 $=y_i-\hat{\mu}_i$ (대개 μ_i가 커질수록 잔차도 커져 표준화가 필요함)
     - ▸ 피어슨 잔차: $e_i=\dfrac{y_i-\hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}}$
     - ▸ 표준화잔차$=\dfrac{y_i-\hat{\mu}_i}{SE\ of\ (y_i-\hat{\mu}_i)}=\dfrac{y_i-\hat{\mu}_i}{\sqrt{\widehat{var}(y_i)(1-h_i)}}$
       "표준화 잔차가 2나 3정도로 크면 주의 깊게 살펴볼 필요가 있음"

4. 로지스틱회귀모형의 해석
   - 이항분포를 따르는 Y에 대하여서 로짓 연결함수를 사용함

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right]=\alpha+\beta x \Leftrightarrow \pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

   - β의 부호 해석
   - ▸ $\beta>0 \Leftrightarrow \pi(x)\uparrow\ as\, x\uparrow$
   - ▸ $\beta<0 \Leftrightarrow \pi(x)\downarrow\ as\, x\uparrow$

- $\beta=0 \Leftrightarrow \pi(x)=\dfrac{e^{\alpha}}{1+e^{\alpha}}=$ 상수 $as\, x \uparrow$

- 오즈비 해석

- $\dfrac{\pi(x+1)}{1-\pi(x+1)}=e^{\beta}\dfrac{\pi(x)}{1-\pi(x)}$

  "$x+1$에서의 오즈는 x에서 오즈의 e^{β}를 곱한 것과 같음"

5. 후향적 연구에서의 로지스틱회귀모형
   - 반응변수 Y가 랜덤이 아니고 설명변수 X가 랜덤인 경우(예: 사례-대조 연구)에도 로지스틱회귀모형을 적용할 수 있음
   - 사례-대조 연구에서 $Y=1$("사례")과 $Y=0$("대조")인 개체로 구성된 표본으로부터 X값이 관측되는 경우에 적용할 수 있으며 로지스틱회귀모형을 통해서 설명변수의 효과를 추정할 수 있음
   - 적합된 로지스틱회귀모형에서 절편을 나타내는 α는 의미가 없음

6. 로지스틱회귀모형에 대한 추론
   - 효과 β에 대한 신뢰구간: $\hat{\beta} \pm Z_{\alpha/2}(SE)$ (Wald 방법)
   - e^{β}에 대한 신뢰구간: $\left(e^{\hat{\beta}-Z_{\alpha/2}(SE)},\ e^{\hat{\beta}+Z_{\alpha/2}(SE)}\right)$

     "설명변수 x가 한 단위 증가할 때 오즈에 미치는 효과(e^{β})에 대한 신뢰구간"
   - β에 대한 유의성 검정
     - $H_0 : \beta=0 \Leftrightarrow Y$와 X는 서로 독립
     - Wald 방법: $Z=\dfrac{\hat{\beta}}{SE}$이 귀무가설 하에서 $N(0,1)$을 따르는 점을 이용함
     - 가능도비 검정: $-2(L_0-L_1) \sim \chi^2(1)$을 이용함

과제하기

구분	내용
과제 주제	- 박태성 & 이승연 (2020) 117쪽 문제 3.13 - 박태성 & 이승연 (2020) 151쪽 문제 4.5 (선형근사식 관련 설명: 교재 120-121쪽 참조) - 사례-대조 연구관련 논문을 읽고 A4용지 1매로 정리 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1726024/pdf/v020p00054.pdf
목적	5주차 강의 내용을 복습하고, 로지스틱회귀모형을 실제 데이터에 적용함으로써 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	5주차 강의 후 1주 후 일요일 밤 12시까지
참고 자료	
기타 유의사항	