07강. 로지스틱회귀모형 [3]

■ 주요용어

용어	해설
전진선택법(forward selection procedure)	더 이상 적합이 개선되지 않을 때까지 항(예측변수)을 추가해 모형을 적합하는 방법
후진제거법(backward elimination precedure)	복잡한 모형에서 시작해서 항을 제거하면서 모형을 적합하는 방법
AIC(Akaike Information Criterion)	'AIC = - 2(로그가능도 - 모형에 있는 모수 개수)'로 정의되며 여러 모형 중에서 AIC 값이 최소인 모형을 선택함
Hosmer-Lemeshow 검정	예측된 확률$(\hat{\pi_i})$을 크기 순에 따라 나열하여 자료를 그룹화하여 관찰값과 적합값을 구하여 검정하는 방법
Pearson 및 표준화 잔차	- 피어슨 잔차: $e_i = \frac{y_i - n_i\hat{\pi_i}}{\sqrt{n_i\hat{\pi_i}(1-\hat{\pi_i})}}$ - 표준화 잔차: $r_i = \frac{y_i - n_i\hat{\pi_i}}{SE} = \frac{e_i}{\sqrt{1-h_i}}$
표집영(sampling zero)	이론상으로 그 칸에 속한 관측값을 가능하지만 현재 data 상으로는 해당 칸의 도수가 0인 경우

정리하기

1. 모형의 선택 과정에서 고려사항
   - 자료에 대한 적합성(모형이 복잡해질수록 유리)
   - 적합된 모형의 해석의 용이성(모형이 간단할수록 유리)

2. 모형 선택 방법
   - 전진선택법(forward selection procedure) : 더 이상 적합이 개선되지 않을 때까지 항(예측변수)을 추가해 모형을 적합하는 방법
   - 후진제거법(backward elimination precedure) : 복잡한 모형에서 시작해서 항을 제거하면서 모형을 적합하는 방법

3. 모형선택의 기준
   - use theory, other research as guide
   - parsimony(simplicity) is good
   - 모형선택의 기준(AIC, Akaike Information Criterion)을 이용할 수 있음
     - ▸ AIC = - 2(로그가능도 - 모형에 있는 모수 개수)
     - ▸ AIC 값이 최소인 모형을 선택
   - 탐색적 연구라면 후진제거법과 같은 자동화 방법을 사용할 수 있음
   - 각 예측변수에 대하여 반응변수의 각 수준에서 적어도 10개의 관측치가 있는 것이 바람직함

4. 가능도비 모형비교
   - 해당 모형과 더 복잡한 모형을 비교하는 가능도비 검정을 통해서 적합결여 여부를 검증하는 방법

   ⇨ 더 복잡한 모형을 적합하더라도 현재 고려하고 있는 모형과 비교하여 적합 정도가 개선되지 않는다면 이미 선택된 모형이 적합하다고 할 수 있음

5. 그룹화된 자료, 비그룹화된 자료와 연속형 예측변수에 대한 적합도 검정
   - 예측변수가 범주형인 경우 자료 파일의 구분
     - · 그룹화된 자료 : 분할표 형식으로 요약된 경우
     - · 비그룹화된 자료: 분할표 등으로 요약되기 전의 원자료
   - 모수의 ML추정값은 위의 두 가지 형태의 자료에 대해서 동일하게 되지만 적합도 검정은 그룹화된 자료의 경우에만 적용할 수 있음 (X^2과 G^2는 적합도수가

5이상인 분할표에 대해서 적용)

- 연속형 또는 연속형에 가까운 예측변수를 갖는 경우의 로지스틱 회귀모형의 적합도 검정방법

① 각 예측변수를 범주화하여(예: 사분위수를 이용하여 4개의 범주로 구분) 그룹화된 자료의 관찰도수와 적합도수에 대해 X^2과 G^2를 적용함

② 예측된 확률$(\hat{\pi_i})$을 크기 순에 따라 나열하여 자료를 그룹화 하여 관찰값과 적합값을 구하여 검정하는 방법 : Hosmer-Lemeshow 검정

6. 로짓모형의 잔차

- 범주형 예측변수에 대해서, 관측도수와 적합도수를 비교하기 위하여 잔차를 사용함
- Pearson 잔차: $e_i = \dfrac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$ (SAS GENMOD에서 Reschi로 표현)

  y_i: "성공"한 도수, n_i: 전체 시행횟수, $\hat{\pi}_i$: 적합 모형으로부터 구한 π_i의 예측값

  ⇨ $e_i \sim N(0, v),\ v < 1$

- 표준화 잔차(Standardized Pearson Residual): $r_i = \dfrac{y_i - n_i\hat{\pi}_i}{SE} = \dfrac{e_i}{\sqrt{1-h_i}}$

  h_i : 관측값의 레버리지(leverage)를 나타내며 첫 행렬(hat matrix)의 대각원소

  ⇨ 근사적으로 $r_i \sim N(0,1)$

  ⇨ $|r_i| > 2 \text{ or } 3$ 이면 모형의 적합결여를 시사함

7. 희박자료(sparse data)효과

- 희박한 자료(sparse data)는 작은 도수를 갖는 칸들이 많은 분할표의 경우를 말함
- 예측변수가 많거나 여러 수준 수로 분류된 분할표에서 흔히 발생함
- 표집영(sampling zero) : 이론상으로 그 칸에 속한 관측값을 가능하지만 현재 data 상으로는 해당 칸의 도수가 0인 경우
- 모형에 따라 표집영은 모형의 모수에 대한 ML추정값이 무한대가 되는 원인이 됨

과제하기

구분	내용
과제 주제	- 박태성 & 이승연 (2020) 156쪽 문제 4.16 - 박태성 & 이승연 (2020) 202쪽 문제 5.14
목적	7주차 강의 내용을 복습하고, 로지스틱회귀모형을 실제 데이터에 적용함으로써 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	9주차 강의 후 1주 후 일요일 밤 12시까지
참고 자료	
기타 유의사항	