04강. 일반화선형모형

■ 주요용어

용어	해설
랜덤성분	반응변수 Y의 확률분포
체계적 성분	설명변수 $\{x_j\}$의 선형식으로 선형예측(linear predictor) 구성
연결함수	랜덤성분과 체계적 성분간의 연결함수를 말함
프로빗 회귀모형	연결함수가 표준정규분포 CDF의 역함수로 주어진 일반화선형모형을 말함
과대산포	도수 데이터에 대해서 포아송분포를 적용하여 분석할 때 $\sigma^2 > \mu$인 경우를 볼 수 있는 데 이를 과대산포(overdispersion)라고 함
포아송 회귀모형	확률변수 Y가 포아송분포를 따른다고 가정할 수 있을 때, 항등연결함수 또는 로그연결함수를 통해서 분석하는 방법

정리하기

1. 모형을 이용한 분석의 장점
   - 모형의 구조에 의해 연관성 및 교호작용 유형을 설명할 수 있음
   - 모수의 추론을 통해 반응에 대한 설명변수의 영향 평가 가능
     ⇒ 추정 모수의 크기는 효과의 강도 및 그 중요성을 결정함
   - 모형 예측값은 자료를 평활(smoothing)하여 반응 평균에 대한 개선된 추정치 제공

2. GLM(Generalized Linear Model)
   - 전통적인 회귀모형을 반응변수가 정규분포를 따르지 않는 경우로 확장
   - GLM을 적용하기 위해서는 반응변수가 지수족 분포(exponential family of distributions)를 따라야만 함
   - 랜덤성분, 체계적 성분, 연결함수로 구성

3. GLM의 구성 요소
   - 랜덤성분 (random component)
   ▸ 반응변수 Y를 결정함

- $Y_1, Y_2, \cdots, Y_n$을 정규분포, 포아송분포, 이항분포 등에서 추출된 랜덤표본으로 가정

- 체계적 성분(systematic component)
  - 설명변수 $\{x_j\}$의 선형식으로 선형예측(linear predictor) 구성 $\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$
- 연결함수(link function)
  - 랜덤성분과 체계적 성분간의 연결함수를 말함
  - μ의 단조함수 $g(\mu)$를 모형 $g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$와 같이 표현할 때, 함수 $g(\cdot)$를 연결함수라고 함
  - $g(\mu)=\mu$(항등연결), $g(\mu)=\log(\mu)$(로그연결), $g(\mu)=\log\left(\frac{\mu}{1-\mu}\right)$(로짓연결) 등

4. 이항자료에 대한 일반화선형모형
   - 선형확률모형
     - $\pi(x)=\alpha+\beta x$: 성공확률이 x에 따라 선형적으로 변함
     - 이항확률분포에 대해서 항등연결함수를 갖는 GLM
     - x의 값이 대단히 크거나 작은 경우에는 $\pi(x)<0$이나 $\pi(x)>1$인 경우가 발생할 수 있다는 단점이 있음
   - 로지스틱 회귀모형
     - $\pi(x)$와 x간의 관계는 비선형 형태로 볼 수 있음
     - 연결함수와 체계적 성분을 $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x$으로 정의하여 적용함
   - 프로빗모형(probit model)
     - 연결함수와 체계적 성분을 $probit[\pi(x)]=\Phi^{-1}(\pi(x))=\alpha+\beta x$으로 정의하여 적용함. 여기서 $\Phi(\cdot)$는 표준정규분포의 CDF임

5. 포아송 회귀모형
   - 포아송 회귀모형
     - Y는 포아송 분포를 따르고, x를 설명변수로 가정함
     - 항등연결함수: $\mu=\alpha+\beta x$
     - 로그연결함수: $\log(\mu)=\alpha+\beta x$
   - 과대산포
     - 포아송 분포를 따를 경우 평균과 분산은 같음($E(Y)=Var(Y)=\mu$)
     - 모형의 랜덤성분에 의해 예측되는 분산보다 더 큰 분산을 갖는 현상을 과대산포(overdispersion)라고 함

과제하기

구분	내용
과제 주제	- 박태성 & 이승연 (2020) 114쪽 문제 3.5 - 박태성 & 이승연 (2020) 114쪽 문제 3.6 - 박태성 & 이승연 (2020) 117쪽 문제 3.11, 3.12
목적	4주차 강의 내용을 복습하고, GLM 모형을 실제 데이터에 적용함으로써 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	4주차 강의 후 1주 후 토요일 밤 12시까지
참고 자료	
기타 유의사항	