범주형자료분석팀

2팀
조장희
위재성
김지현
조수미
송지현
김민지

INDEX

GLM(일반화 선형모형, Generalized Linear Model)

연속형 반응변수에 대한 모형과 범주형 반응변수에 대한 모형

모두를 포함하는 모형의 집합

* 선형회귀모형: GLM 중 하나

모형을 일반화할 때, 두 가지를 일반화

1) 랜덤성분의 분포 일반화
2) 랜덤성분의 함수 일반화

" GLM = 기존의 회귀모형을 포함한 더욱 넓은 범위의 모형! "

자세한 설명은 뒤에서 계속 …

GLM 구성 성분

$$g(\boldsymbol{\mu}) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

랜덤 성분 | 체계적 성분 | 연결 함수

$\boldsymbol{\mu}(=E(Y))$

Y의 확률분포를 정해줌으로써 반응변수 Y 정의

가정한 확률분포의 **기댓값**인 μ로 랜덤성분을 표기

이진형 자료 | 이항분포의 평균인 $\pi(x)$로 랜덤성분 표기

GLM 구성 성분

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

랜덤 성분	체계적 성분	연결 함수
		$g()$

- 연결 함수의 종류

항등 연결함수 $g(\mu) = \mu$	반응변수 Y가 연속형일 때 사용 ex) 일반선형회귀모형
로그 연결함수 $g(\mu) = \log(\mu)$	반응변수 Y가 도수자료(count data)일 때 사용 ex) 포아송 분포 / 음이항 분포
로짓 연결함수 $g(\mu) = \log[\mu/(1-\mu)]$	반응변수 Y가 이항분포를 따를 때 사용 ex) 로지스틱 회귀

최대가능도 추정법 (Maximum Likelihood Method)

LSE를 사용한 일반선형회귀와는 달리
GLM은 최대가능도법 (Maximum Likelihood Method)을 사용해 적합된 모형

↓

정규성 조건을 맞출 필요 없음

: 오차항이 정규분포를 따라야 한다는 가정!

↓

GLM은 보다 더 포괄적인 범위의 반응변수를 다룰 수 있다는 특징

유의성 검정이란

유의성 검정

- 모형의 모수 추정값이 유의한지 검정
- 축소 모형의 적합도가 좋은지 검정

$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$일 때,

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

H_1: 적어도 하나의 β는 0이 아니다.

ML을 이용한 검정

가능도비 검정

검정 통계량 : $G^2 = -2\log\left(\frac{l_0}{l_1}\right) = -2(L_0 - L_1) \sim \chi^2{}_{df}$

기각역 : $G^2 \geq \chi^2{}_{a,df}$

가능도 함수의 최댓값을 이용해 비교

- l_0: 귀무가설 하에서의 가능도함수
- l_1: 전체공간 하에서의 가능도함수
- df : 귀무가설과 대립가설 모수 개수의 차이

이탈도

포화모형 S와 관심모형 M을 비교하기 위한 가능도비 통계량

$$\text{이탈도} = -2\log\left(\frac{l_M}{l_S}\right) = -2(L_M - L_S)$$

H_0: 관심모형 M에 포함되지 않는 모수는 모두 0이다.

H_1: 적어도 하나는 0이 아니다.

가능도 함수의 최댓값의 차이 사용

모형이 내포(nested)될 때만 사용 가능 $(\mathrm{M} \subset S)$

로지스틱 회귀 모형이란?

반응변수 Y가 이항자료일때 사용

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

반응 변수 Y가 성공 또는 실패의

이항분포를 따르는 변수이기에

일반 선형회귀는 사용할 수 없음 ...why?

로지스틱 회귀 모형의 해석

확률로 해석

로지스틱 회귀 모형 식을 확률에 대한 식으로 변형

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

확률 값 $\pi(x)$가 cutoff point보다 크면 Y=1, 작으면 Y=0

모수 β의 해석

$\beta > 0$: 곡선이 상향, $\beta < 0$: 곡선이 하향

$|\beta|$가 증가함에 따라 변화율이 증가

기준범주 로짓모형(Baseline-Category Logit Model)

기준 범주 로짓 모형

범주 j일 때 x_1의 회귀계수

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^p x_p, j = 1, \ldots, (J-1)$$

- 기준 범주: 범주 J
- 나머지 범주: 범주1, 범주2, ⋯, 범주 J-1

J=2 라면, 로지스틱 회귀모형

누적 로짓모형(Cumulative Logit Model)

누적확률 누적확률에 로짓 연결함수를 씌운 모형

$$logit[P(Y \le j)] = \log\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p ,$$
$$j = 1, \ldots, (J-1)$$

- α_j 가 다른 J-1 개의 로짓 방정식이 생김

- 회귀계수 β 에는 j 첨자X
- J-1개의 로짓 방정식에서의 회귀계수 β의 효과가 동일하기 때문!

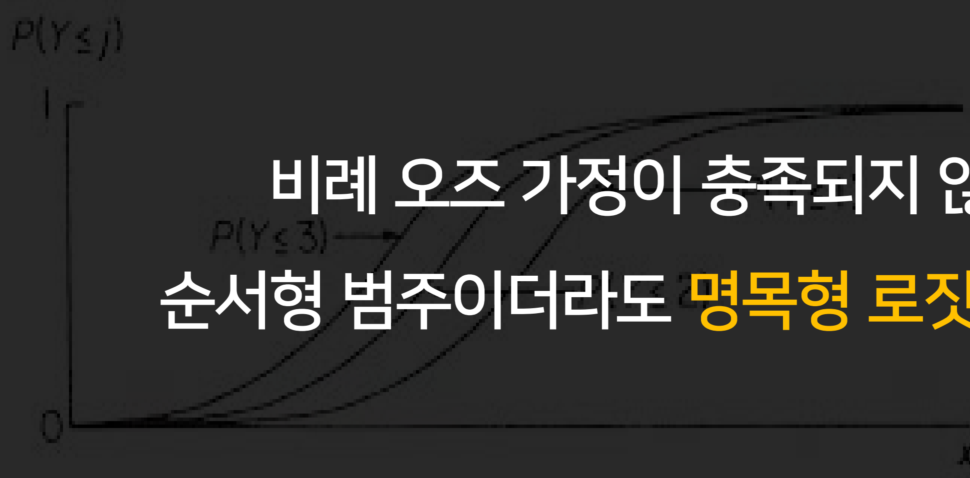
=> '비례 오즈 가정 '

비례 오즈 가정(proportional odds)

Collapse 과정에서 cut point를 어디로 지정하든

회귀계수 β 의 효과는 동일하다.

$$logit[P(Y \le j)] = \log\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p ,$$



=> 일종의 평행

비례 오즈 가정이 충족되지 않으면,

순서형 범주이더라도 명목형 로짓 모형을 씀

포아송 회귀 모형 (Poisson Regression Model)

음이항 회귀모형

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- 음이항 랜덤성분, 로그연결함수
- 음이항 분포는 이미 분산이 평균보다 큰 상태
- 분산이 평균과 비선형관계임을 가정, 산포모수 D 사용

$$(Y) = \mu\,, \qquad Var(Y) = \mu + D\mu^2$$

포아송 회귀 모형 (Poisson Regression Model)

영과잉 포아송 모형(ZIP)으로 해결!

ZIP의 반응 변수 Y는 0의 값이 발생하는 점확률분포와 0보다 큰 정수값을 갖는 포아송 분포의 혼합구조

$$Y = \begin{cases} 0\ , & with\ probability\ p \\ \text{포아송 분포(평균 } \lambda) , & with\ probability\ 1-p \end{cases}$$

과대영 문제 : 0이 많아 나타나는 문제

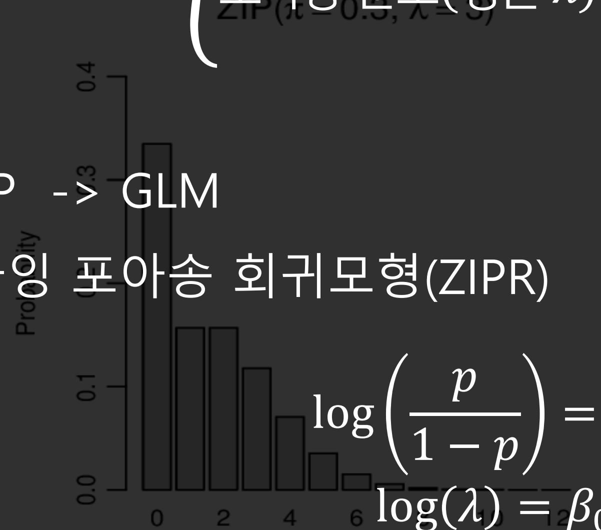
예) 로또 당첨자 수 데이터

※ZIP -> GLM

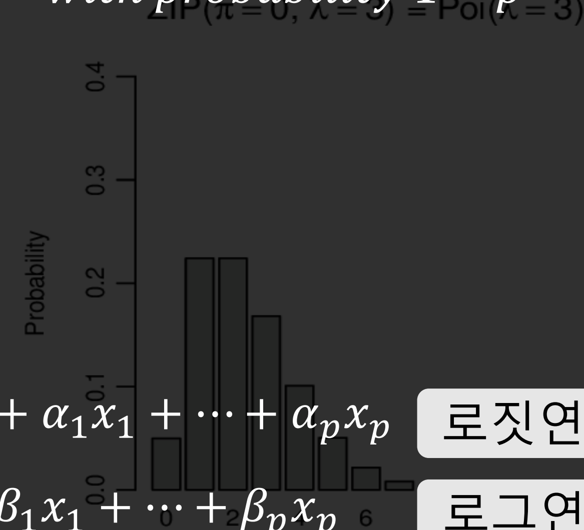
영과잉 포아송 회귀모형(ZIPR)

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p$$ 로짓연결함수

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$ 로그연결함수



<과대영 문제 발생 그래프>



<일반 포아송 분포 그래프>