범주형자료분석팀

2팀
이지연
심예진
조장희
조혜현
진효주

INDEX

GLM(일반화선형모형,Generalized Linear Model)이란?

GLM에서 일반화

일반회귀모형을 두 가지로 일반화

1) 랜덤성분이 정규분포를 포함한 다른 분포를 갖도록 일반화

2) 랜덤성분의 함수인 연결함수 g( )로 모형화하여 일반화

일반선형회귀모형은 GLM의 한 종류!

당신이 알던 OLS회귀는 빙산의 일각에 불과하다…

GLM의 구성성분

랜덤 성분 | 체계적 성분 | 연결 함수

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

$\mu\ (= E(Y))$

반응변수 Y를 정의하고 반응변수에 대한 확률분포를 가정하는 데,

가정한 확률분포의 기대값인 평균 μ를 랜덤성분으로 표기

이항분포를 따르는 경우: $\pi(x)$

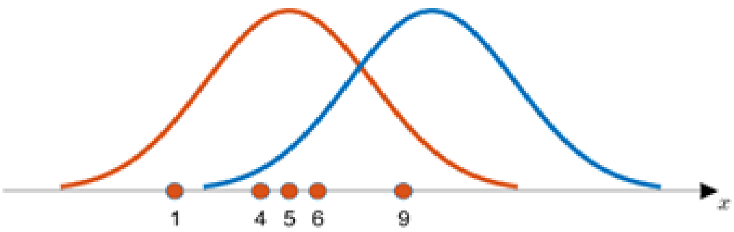
포아송분포를 따르는 경우: μ (또는 λ)

GLM의 구성성분

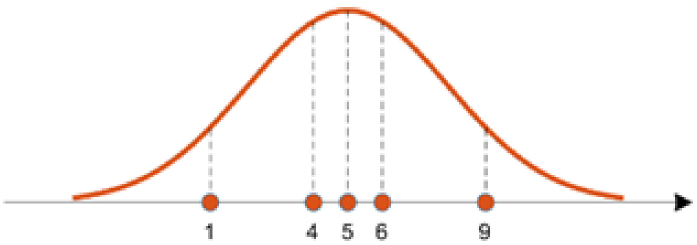
연결 함수

항등 연결 함수 $g(\mu) = \mu$	반응변수가 연속형일 때 사용 Ex) 정규분포
로그 연결 함수 $g(\mu) = \log(\mu)$	반응변수가 도수자료일 때 사용 Ex) 포아송 분포/음이항 분포
로짓 연결 함수 $g(\mu) = \log[\mu/(1-\mu)]$	반응변수가 이항자료일때 사용 로짓(Logit)? 오즈에 로그를 씌운 값

GLM의 모형 적합



데이터들의 분포가 주황색 곡선의 중심에 더 일치



점선의 높이는 각 관측치가 확률분포를 따를 가능도를 표현

데이터를 관찰함으로써
데이터가 추출되었을 것으로 생각되는
분포의 특성을 추정할 수 있음

독립인 각 관측치들의 가능도를
모두 곱한 것이 가능도 함수!

가능도 함수를 통해 MLE를 찾을 수 있다!

유의성 검정이란?

유의성 검정

모형의 모수 추정값이 유의한지에 대한 검정

축소 모형의 적합도가 좋은지에 대한 검정

$\boldsymbol{g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$ 일 때,

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

H_1 : 적어도 하나의 β는 0이 아니다.

가능도비 검정

도비는 자유에요...

가능도비 검정

귀무가설 하에서 계산되는 가능도 함수 l_0 와

MLE에 의해 계산되는 가능도 함수 l_1 의 차이 이용

검정통계량 : $-2\log\left(\frac{l_0}{l_1}\right) = -2(L_0 - L_1) \sim X^2$

$$-2\log\left(\frac{\text{모수가 귀무가설 } H_0\text{을 만족할때 }(\beta=0\text{일 때) 가능도 함수의 최댓값}}{\text{모수에 대한 아무 제한 조건이 없을 때의 가능도 함수의 최댓값}}\right)$$

자유도는 귀무가설과 대립가설 간의 모수의 개수 차이

이탈도

이탈도

포화 모형 S와 관심 모형 M을 비교하기 위한 가능도비 통계량

이탈도(deviance) = $-\mathbf{2}\ (\boldsymbol{L_M} - \boldsymbol{L_S})$

이탈도는 M의 계수가 S의 계수에 포함이 되어있는 경우(nested)에만 사용 가능

H_0 : 관심 모형에 속하지 않는 모수는 모두 0이다. **(관심모형 사용)**

H_1 : 관심 모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다. **(관심모형 사용불가)**

로지스틱 회귀 모형이란?

로지스틱 회귀 모형

반응변수 Y가 성공 혹은 실패를 나타내는 이항 자료인 회귀 모형

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

반응 변수 Y가 성공 혹은 실패의
이항분포를 따르는 변수이기때문에
기존의 회귀모델을 그대로 적용할 수 없음

로지스틱 회귀 모형의 해석

로지스틱 회귀 모형 식을 확률에 대한 식으로 변형

$$\boldsymbol{\pi(x)} = \frac{\mathbf{exp}(\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_p x_p})}{\mathbf{1} + \mathbf{exp}(\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_p x_p})}$$

확률 값이 cutoff point 보다 크면 $Y = 1$, 작으면 $Y = 0$ 으로 예측 가능

모수 β를 갖는 로지스틱 회귀 모형의 접선의 기울기

$$\boldsymbol{\beta\pi(x)[1 - \pi(x)]}$$

β가 양수 = 상향 곡선

β가 음수 = 하향 곡선

$|\beta|$이 증가함에 따라 변화율이 증가

기준 범주 로짓 모형 (Baseline-Category Logit Model)

모형

$$log\left(\frac{\pi_j}{\pi_J}\right) = log\left(\frac{P(Y=j|X=x)}{P(Y=J|X=x)}\right) = \alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K, j = 1, \dots, (J-1)$$

- 기준 범주 : 범주J
- 나머지 범주 : 범주1, 범주2,⋯,범주J-1
- A~K : 설명변수 x에 대한 첨자 (A제곱 아님) *β_j^A : 범주 j일 때 x_1 의 회귀계수*
- J=2이면 보통의 로지스틱 회귀!

누적 로짓 모형 (Cumulative Logit Model)

순서형 범주일 때 사용!

범주(카테고리)를 순서대로 정렬한 뒤 collapse

collapse

정렬된 범주들을 두 부분으로 나누는 과정

Cut point를 기준으로 나눔

Cut point

포아송 회귀 모형(Poisson Regression Model)

포아송 회귀 모형

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

- 포아송 분포를 따르는 도수 자료(count data)를 반응변수로 갖는 GLM
- 양변의 범위 맞춰 주기 위해 로그연결함수 사용 like 로지스틱의 로짓 연결 모형

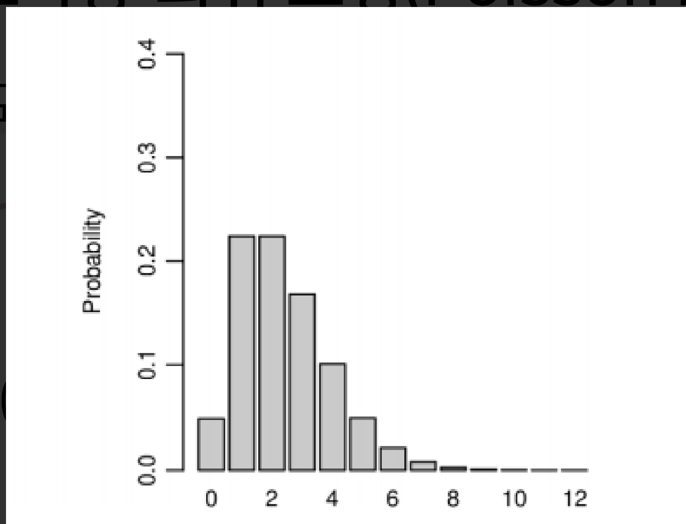
5

포아송 회귀 모형

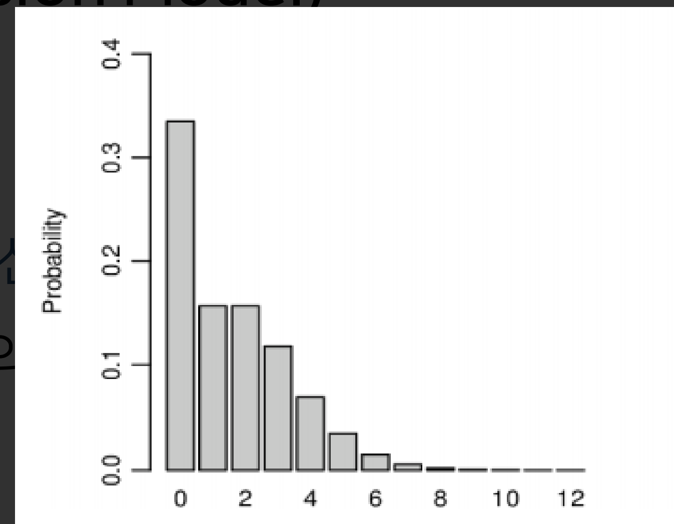
ZIP 회귀 모형

포아송 회귀 모형(Poisson Regession Model)

정상 포아송 분포



과대영 발생



대안 모델

$$y_i = \begin{cases} 0 \ , & with\ probability\ \phi i \\ g(y_i)\ , & with\ probability\ 1 - \phi i \end{cases}$$

- 음이항 회귀모형(Negative Binomial Regression)

① 항상 0인 집단 vs 0이 아닌 집단으로 나눔

② 0이 아닌 집단에 대해서 포아송 회귀 모형 적합