범주형자료분석팀

2팀
이지연
심예진
조장희
조혜현
진효주

INDEX

0

1주차 리뷰

1주차 리뷰

분할표

여러 개의 범주형 변수를 기준으로 관측치를 기록하는 표

2차원 분할표(I*J)

	Y		합계
X	n_{11}	n_{12}	n_{1+}
	n_{21}	n_{22}	n_{2+}
합계	n_{+1}	n_{+2}	n_{++}

3차원 분할표(I*J*K)

	Y		합계
X	n_{11+}	n_{12+}	n_{1++}
	n_{21+}	n_{22+}	n_{2++}
합계	n_{+1+}	n_{+2+}	n_{+++}

「주변분할표」

		Y		합계
Z	X	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{++2}

「부분분할표」

1주차 리뷰

독립성 검정의 종류

2차원 분할표 독립성 검정

- 대표본
  - 명목형
    - 피어슨 카이제곱 검정
    - 가능도비 검정
  - 순서형
    - MH 검정
- 소표본
  - 피셔의 정확검정

1주차 리뷰

오즈비

각 오즈(odds)의 비

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

성별	연인 유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
	오즈 = 4.388	
남성	398 (0.793)	104 (0.207)
	오즈 = 3.826	

$$\frac{4.388}{3.826} = 1.1468$$

여성이 연인이 있을 오즈가 남성이 연인이 있을 오즈보다 1.1468배 더 높다!

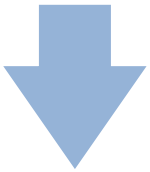
1

GLM

GLM(일반화선형모형,Generalized Linear Model)이란?

일반 선형회귀 모형

최소제곱법(LSE)으로 연속형 변수 사이의 회귀식 추정



반응변수가 범주형 자료이거나 count data인 경우?

이항자료	다항 자료	도수 자료
Yes or No	*강아지/고양이/다람쥐*	*물 마신 횟수*

GLM

GLM(일반화선형모형,Generalized Linear Model)이란?

일반 선형회귀 모형

최소제곱법(OLS)을 이용한 회귀식 추정

일반선형회귀모형 불가능!

일반선형모형에서 확장된 GLM(일반화선형모형) 사용

이항자료	다항 자료	도수 자료
Yes or No	*강아지/고양이/다람쥐*	*물 마신 횟수*

GLM

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…

GLM(일반화선형모형,Generalized Linear Model)이란?

GLM의 필요성

1. 범주형 자료분석은 오차항의 확률분포가 정규분포가 아니기 때문에
선형회귀모형 사용 불가

2. GLM은 **ML 방법(최대우도법)** 사용하여 모형을 적합하기 때문에
LSE와 같은 정규성 조건 필요 없음

GLM(일반화선형모형,Generalized Linear Model)이란?

분할표와의 비교

분할표	GLM
변수 간의 효과 파악만 가능 (독립성 검정)	변수 간 연관성 파악과 반응 변수에 대한 예측 가능
범주형 자료만 표현 가능	연속형 설명변수 사용 가능

GLM(일반화선형모형,Generalized Linear Model)이란?

GLM의 특징

오차항의 다양한 분포를 가정

정규성 가정을 만족해야 하는 일반선형회귀와 다르게 GLM은 다양한 분포를 가정

선형 관계식 유지

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

회귀계수 β에 대한 선형성을 유지하여 해석 용이

GLM(일반화선형모형,Generalized Linear Model)이란?

GLM의 특징

범위가 제한되는 반응변수도 사용 가능

연결함수 $g(\mu)$를 통해 범위 조정

독립성 가정만 필요

기존의 회귀 가정(정규성, 등분산성, 독립성, 선형성) 중에서 독립성 가정만 만족하면 됨

자기상관성 검정 필요!

GLM(일반화선형모형,Generalized Linear Model)이란?

GLM의 특징

범위가 제한되는 반응변수도 사용 가능

연결함수 $g(\mu)$를 통해 선형성 가정

독립성 가정만 필요

기존의 회귀 가정(정규성, 등분산성, 독립성, 선형성) 중에서 독립성 가정만 만족하면 됨

자기상관성 검정 필요!

자기상관성이란?

시간 또는 공간적으로 연속된

일련의 관측치들 간에 존재하는 상관관계

더빈-왓슨 검정(DW Test)을 통해 오차의 독립성 검정

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

랜덤 성분 | **체계적 성분** | 연결 함수

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

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

설명변수 X들의 선형 결합

(교호작용항이나 곡선효과를 나타내는 항을 넣을 수 있음)

교호작용: $x_i = x_a x_b$

곡선효과: $x_i = x_a^2$

GLM의 구성성분

랜덤 성분 | 체계적 성분 | **연결 함수**

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

$g(\,)$

좌변(랜덤성분)과 우변(체계적 성분)을 연결하며

둘의 범위를 맞춰주는 역할

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 종류

GLM	랜덤성분	연결함수	체계적 성분
일반 회귀 분석	정규 분포	항등	연속형
분산 분석			범주형
공분산 분석			혼합형
선형 확률 모형	이항 자료	항등	혼합형
로지스틱 회귀 모형		로짓	
프로빗 회귀 모형		프로빗	
기준범주 로짓 모형	다항 자료	로짓	
누적 로짓 모형			
이웃범주 로짓 모형			
연속비 로짓 모형			
로그 선형 모형	도수 자료	로그	범주형
포아송 회귀 모형			혼합형
음이항 회귀 모형			
카우시 모형			
율자료 포아송 회귀 모형	비율 자료		

선택과 집-중의 감성으로 알아보자!

GLM의 모형 적합

모형 적합(model fitting)

주어진 데이터를 근거로 모형의 모수를 추정하는 것

최소제곱법
Least Square Method

오차의 제곱합을 최소화하는

모수를 찾는 방법

회귀의 기본 가정들을 모두 만족해야 함

최대가능도추정법
Maximum Likelihood Method

가능도를 최대화하는

모수를 찾는 방법

독립성 가정만 만족하면 됨!

GLM은 최대가능도추정법(MLE)를 통해 모형을 적합!

GLM의 모형 적합

모형 적합(model fitting)

주어진 데이터를 분석하고 모형(모수)를 추정하는 것

가능도(Likelihood)란?

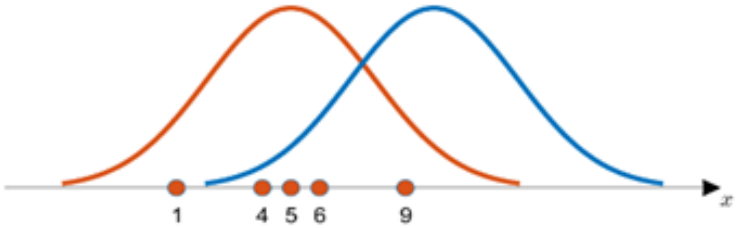
관측값이 고정되었을 때, 그 관측값이 어떤 확률분포를 따를 가능성

LSE는 오차의 제곱합을 최소화하는 방법으로
회귀의 기본 가정들을 만족해야 함

cf) 확률은 분포가 고정되었을 때 값이 관측될 가능성

GLM은 그러한 가정들이 없기에 LSE사용 불가

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!

GLM의 모형 적합

가능도 함수

각 데이터 샘플에서 후보 분포에 대한 높이를 계산하여 모두 곱한 식

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

로그 가능도 함수

계산의 편의성을 위해 가능도 함수에 로그를 취한 함수

$$L(\theta|x) = \log P(x|\theta) = \sum_{i=1}^{n} \log P(x|\theta)$$

MLE의 기본 아이디어

가능도 함수가 최대가 되는 모수 θ를 찾는다!

로그 가능도 함수의 도함수=0을 만족하는 모수 θ를 찾음으로써

MLE(최대가능도추정량)를 구할 수 있음!

2

유의성 검정

유의성 검정이란?

유의성 검정

모형의 모수 추정값이 유의한지에 대한 검정

축소 모형의 적합도가 좋은지에 대한 검정

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

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

$H_1 :$ 적어도 하나의 β는 0이 아니다.

가능도비 검정

가능도비 검정

도비는 자유에요...

귀무가설 하에서 계산되는 가능도 함수 l_0 와

MLE에 의해 계산되는 가능도 함수 l_1 의 차이 이용

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

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

자유도는 귀무가설과 대립가설 간의 모수의 개수 차이

유의성 검정

가능도비 검정

l_0 와 l_1의 차이가 커짐 → 검정통계량 $-2(l_0 - l_1)$이 커짐 → p-value가 작아짐 → 귀무가설 기각 → 적어도 하나의 β는 0이 아님 → 모형의 모수 추정값은 유의하구나!

가능도비 검정은 가장 많은 양의 정보를 사용하기 때문에 좋은 검정력을 가짐

이탈도 차이를 통한 모형비교에서도 사용!

유의성 검정

이탈도

관심모형(M)

유의성 검정을 진행하고자 하는 모형

L_M = 모형 M에서 얻은 로그 가능도 함수의 최댓값

범주팀 복지 (Y) =
$\beta_0 + \beta_1 \times$ 스터디 시간 (x_1) +
$\beta_2 \times$ 교안 페이지 수 (x_2)

포화모형(S)

각 관측값에 대하여
완벽하게 자료를 적합하는 모형

L_s = 모형 S에서 얻은 로그 가능도 함수의 최댓값

범주팀 복지 (Y) =
$\beta_0 + \beta_1 \times$ 스터디 시간 (x_1) +
$\beta_2 \times$ 교안 페이지 수 (x_2) +
$\beta_3 \times$ 스터디시간 $\times$ 교안페이지수$(x_1 x_2)$

유의성 검정

이탈도

이탈도

포화 모형 S와 관심 모형 M을 비교하기 위한 가능도비 통계량

이탈도(deviance) = $-2\ (L_M\ -\ L_S)$

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)**

유의성 검정

이탈도

이탈도가 작음
(검정통계량이 작음)

P-value가 큼

귀무가설 기각 X

관심모형에 포함되지
않은 계수들이 모두 0

간단한 관심모형으로
적합 가능!

유의성 검정

이탈도

두 모형의 이탈도의 차이는 **가능도비 통계량**과 같다!

M_0 은 단순 모형 , M_1은 복잡모형, L_0 , L_1은 로그 가능도함수의 최댓값일 때

$$\underbrace{-2(L_0 - L_s) - \{-2(L_1 - L_s)\}}_{M_0\text{의 이탈도 - }M_1\text{의 이탈도}} = \underbrace{-2(L_0 - L_1)}_{\text{가능도비 통계량}}$$

만약 M_0 가 M_1에 내포되지 않은 경우,

AIC와 같은 측도를 이용해서 모형 비교

유의성 검정

이탈도

이탈도의 차이가 작음 → 검정통계량이 작음 → p-value가 큼 → 귀무가설 기각 X → 관심모형에 포함되지 않은 계수들이 모두 0 → 간단한 관심모형으로 적합 가능!

반대로 이탈도 차이가 크면 단순모형이 복잡모형에 비해

잘 적합되지 않는다는 뜻!

3

로지스틱 회귀 모형

로지스틱 회귀 모형

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 모형

로지스틱 회귀 모형이란?

종속변수가 범주 1이 될 확률로 가정한 식

$$\pi(x) = P(Y = 1|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

좌변의 범위 = $(0,1)$ 우변의 범위 = $(-\infty, \infty)$

좌변에 오즈를 설정

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

좌변의 범위 = $(0, \infty)$ 우변의 범위 = $(-\infty, \infty)$

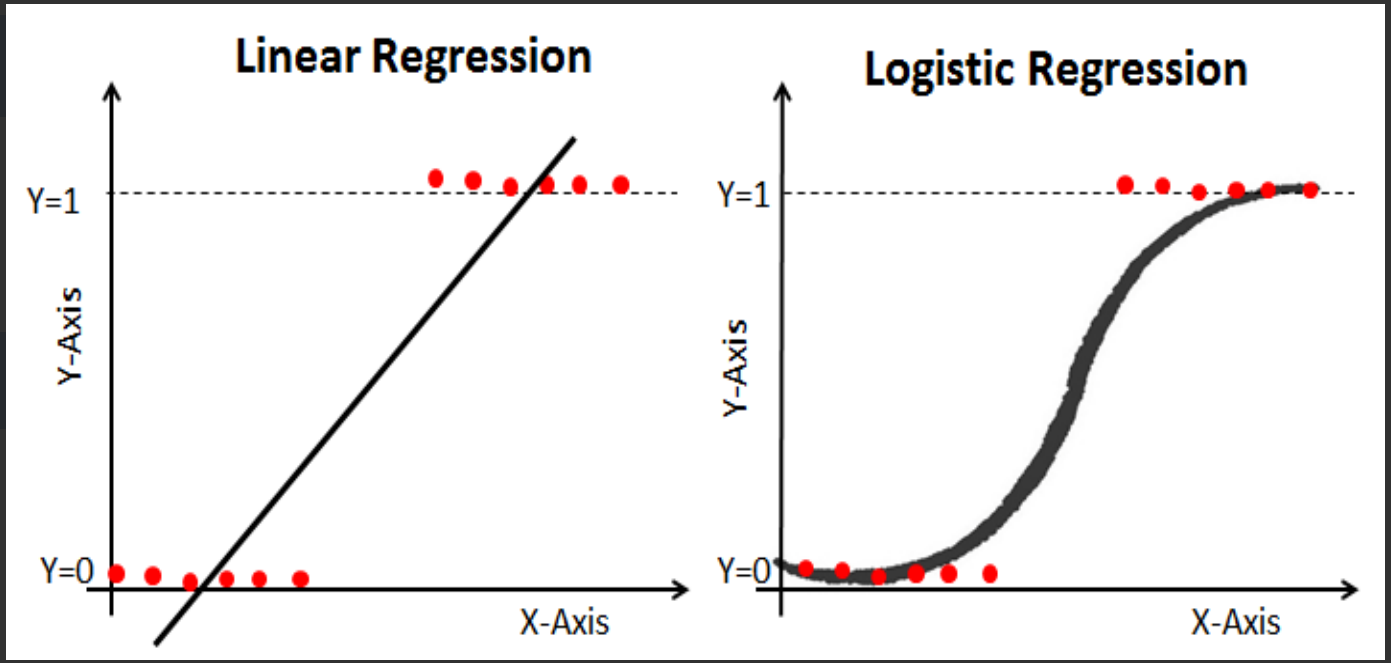
오즈에 로그를 취해주어 로지스틱 회귀 모형 완성

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

좌변의 범위 = $(-\infty, \infty)$ 우변의 범위 = $(-\infty, \infty)$

3 로지스틱 회귀 모형

로지스틱 회귀 모형이란?



오즈에 로그를 취해주어 로지스틱 회귀 모형 완성

이렇게 범위를 맞춰 줄 수 있다-!

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

좌변의 범위 = $(-\infty, \infty)$ 우변의 범위 = $(-\infty, \infty)$

로지스틱 회귀 모형

로지스틱 회귀 모형이란?

로지스틱 회귀 모형의 장점

1. 로짓연결함수를 통해 범위 문제 해결
2. 가정으로부터 자유로움(독립성 가정만 만족하면 됨)

   반응변수 Y가 이항분포를 따르기 때문에 정규성과 등분산성을 만족할 수 없음

로지스틱 회귀 모형의 해석

로지스틱 회귀 모형 식을 확률에 대한 식으로 변형

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

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

모수 β를 갖는 로지스틱 회귀 모형의 접선의 기울기

$$\beta\pi(x)[1 - \pi(x)]$$

β가 양수 = 상향 곡선

β가 음수 = 하향 곡선

$|\beta|$이 증가함에 따라 변화율이 증가

로지스틱 회귀 모형의 해석

로지스틱 회귀모형의 연결함수가 로그 오즈 함수이므로
오즈와 오즈비를 이용하여 해석가능

$$\log\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = [\beta_0 + \beta(x+1)] - [\beta_0 + \beta x]$$

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

$$\frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta}$$

X가 한 단위 증가하면 Y = 1일 오즈가 e^{β}배만큼 증가한다고 해석

로지스틱 회귀 모형

로지스틱 회귀 모형의 해석

예시

로지스틱 회귀모형의 연결함수가 로그 오즈 함수이

오즈와 오즈비를 이용하여 해석가능

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 4 + 3x,$$

Y = 1 (합격), Y = 0 (불합격), x (학점)

$$\log\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = [\beta_0 + \beta(x+1)] - [\beta_0 + \beta x]$$

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

$$\frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta}$$

x가 한 단위 증가할 때 Y = 1 (합격)일 오즈가 $e^3 = 20.086$배 증가

X가 한 단위 증가하면 Y = 1일 오즈가 e^{β}배만큼 증가한다고 해석

로지스틱 회귀 모형의 해석

추정된 $\hat{\pi}(x_1)$과 $\hat{\pi}(x_2)$를 통해 $\hat{\pi}(x_2) - \hat{\pi}(x_1)$를 구하여

x_1에서 x_2로 증가할 때의 확률 변화 추정 가능

예시) 학점이 2.5에서 4.5로 증가하는 경우

$$\frac{\exp(4 + 3 \times 4.5)}{1 + \exp(4 + 3 \times 4.5)} - \frac{\exp(4 + 3 \times 2.5)}{1 + \exp(4 + 3 \times 2.5)} = 0.00001$$

만큼 Y = 1 (합격)일 확률이 증가

4

다범주 로짓 모형

다범주 로짓 모형 (Multicategory Logit Model)

다범주 로짓 모형 반응변수의 범주가 3개 이상인 모형

즉, 반응변수가 다항 분포를 따르는 다항자료

기존 로짓 모형

다범주 로짓 모형

다범주 로짓 모형

다범주 로짓 모형의 종류

다범주 로짓 모형	명목형	기존 범주 로짓 모형(Baseline-Category Logit Model)
	순서형	이웃 범주 로짓 모형 (Adjacent-Categories Model)
		연속비 로짓 모형 (Continuation-ratio Logit Model)
		누적 로짓 모형 (Cumulative Logit Model)

반응변수가 명목형 자료인지 순서형 자료인지에 따라
사용할 수 있는 모형이 다름

기준 범주 로짓 모형 (Baseline-Category Logit Model)

명목형 변수

범주 하나를 기준 범주로 정한 후, 나머지 범주들을 짝지어 로짓을 정의

다범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

기준 범주 로짓

$$\log\left(\frac{\pi_j}{\pi_J}\right), j = 1, \ldots, J-1$$

- 기준 범주 : 범주J
- 나머지 범주: 범주1, 범주2,…,범주J-1
- 반응변수가 J범주에서 일어났다는 조건 하에서 반응이 j범주일 로그 오

왜 하필 J인가? 모든 팀원 이름에 J가 들어가기 때문이지.. 후후

다범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

모형

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

- 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!

다범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

확률

$$\pi_j = \frac{e^{\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K}}{\sum_{i=1}^{J} e^{\alpha_i + \beta_i^A x_1 + \cdots + \beta_i^K x_K}} \quad , j = 1, \ldots, (J-1)$$

- π_j : 범주j에 속할 확률
- 앞 슬라이드의 모형 공식 변형하여 나타낼 수 있음

기준 범주 로짓 모형 (Baseline-Category Logit Model)

회귀 계수 β 해석

$$log\left(\frac{\pi_j}{\pi_J}\right) = log\left(\frac{P(Y=j|X=x)}{P(Y=j|X=x)}\right)$$

$$= \alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K, j = 1, \ldots, (J-1)$$

$$\rightarrow \quad \frac{\pi_j}{\pi_J} = \mathrm{e}^{\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K}, j = 1, \ldots, (J-1)$$

- 기준 범주 범주J에 비해 범주j일 로그 오즈를 통해 해석
- (다른 설명 변수가 고정되어 있을 때)

  x 가 1단위 증가하면 범주J 대신 범주j일 오즈가 e^{β}배 증가한다!

기준 범주 로짓 모형 (Baseline-Category Logit Model)

회귀 계수 β 해석

$$\log\left(\frac{\pi_2}{\pi_1}\right) - \log\left(\frac{\pi_2/\pi_J}{\pi_1/\pi_J}\right) = \log\left(\frac{\pi_2}{\pi_J}\right) - \log\left(\frac{\pi_1}{\pi_J}\right)$$
$$= [\alpha_2 - \alpha_1] + [(\beta_2^A - \beta_1^A)x_1 + \cdots + (\beta_2^K - \beta_1^K)x_K]$$

- 기준 범주 외의 또 다른 범주끼리 로짓을 빼서 그 범주의 관계로 해석
- (다른 설명 변수가 고정되어 있을 때)

  x 가 1단위 증가하면 범주1 대신 범주2일 오즈가 $e^{\beta_2 - \beta_1}$배 증가한다!

다범주 로짓 모형

누적 로짓 모형 (Cumulative Logit Model)

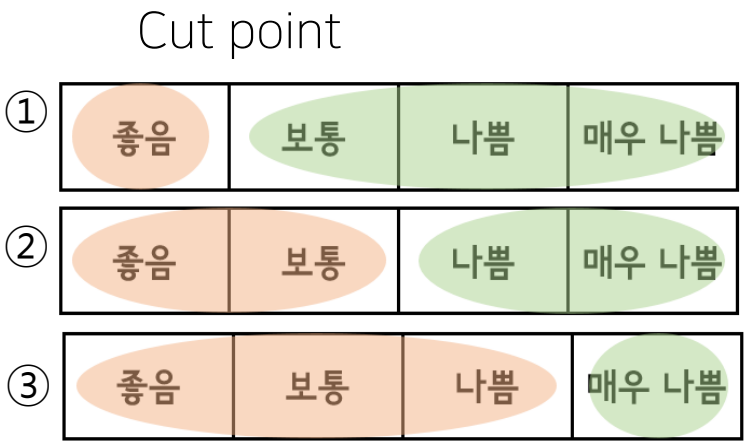
순서형 범주일 때 사용!

범주(카테고리)를 순서대로 정렬한 뒤 collapse

collapse

정렬된 범주들을 두 부분으로 나누는 과정

Cut point를 기준으로 나눔



누적 로짓 모형 (Cumulative Logit Model)

누적확률

$$P(Y \leq j|X = x) = \pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)\ , j = 1, \ldots, J$$

범주1부터 범주j까지의 확률을 모두 더하면 누적확률!

누적 로짓 모형 (Cumulative Logit Model)

누적 로짓 모형

$$
\begin{aligned}
& logit[P(\mathrm{Y} \leq \mathrm{j}|X = x)] \\
&= \log\left(\frac{P(Y \leq j|X = x)}{1 - P(Y \leq j|X = x)}\right) \\
&= \log\left(\frac{P(Y \leq j|X = x)}{P(Y > j|X = x)}\right) \\
&= \log\left(\frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_J(x)}\right) \\
&= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \,, j = 1, \ldots, (J-1)
\end{aligned}
$$

누적확률에 로짓 연결함수 씌우면~ 누적 로짓 모형!

누적 로짓 모형 (Cumulative Logit Model)

누적 로짓 모형

$$logit[P(\mathrm{Y} \leq \mathrm{j}|X = x)]$$
$$= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \, , j = 1, \ldots, (J-1)$$

- α_j가 다른 $J-1$개의 로짓 방정식이 만들어짐
- 근데 회귀계수 β에는 j첨자가 존재하지 않음
- $J-1$개의 로짓 방정식에서의 회귀계수 β의 효과가 동일하기 때문!

  → 비례 오즈 가정

누적 로짓 모형 (Cumulative Logit Model)

비례 오즈 가정

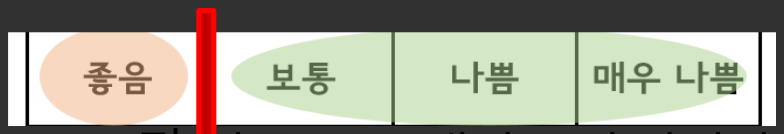
누적 로짓 모형

collapse 과정에서 cut point를 어디로 지정하든

회귀계수 β의 효과는 동일해야 한다

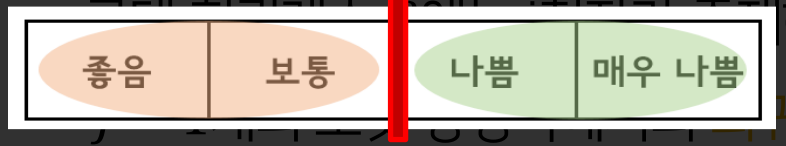
$$= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \, , j = 1, \ldots, (J-1)$$

Cut point



- α_j가 다른 $J-1$개의 로짓 방정식이 만들어짐

Cut point

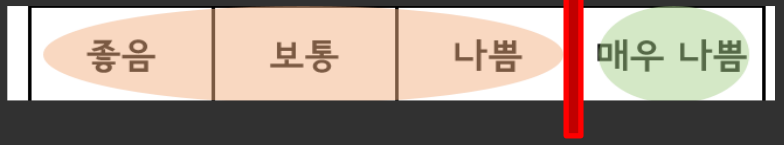


비례 오즈 가정이 충족되지 않는다면,

순서형 범주이더라도 명목형 로짓 모형 씀

-> 비례 오즈 가정

Cut point



다범주 로짓 모형

누적 로짓 모형 (Cumulative Logit Model)

회귀 계수 β 해석

$$\log\left(\frac{P(Y \le j|X=x)}{P(Y>j|X=x)}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \,, j = 1, \ldots, (J-1)$$

$$\rightarrow \frac{P(Y \le j|X=x)}{P(Y>j|X=x)} = e^{\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p}, j = 1, \ldots, (J-1)$$

- 누적확률의 로그 오즈 보고 해석
- x 가 1단위 증가하면 $Y > j$ 대신 $Y \le j$ 일 오즈가 e^{β}배 증가한다!

  (다른 설명 변수가 고정되어 있을 때)

5

포아송 회귀 모형

포아송 회귀 모형

포아송 회귀 모형(Poisson Regression Model)

포아송 분포

단위 시간 동안 어떤 사건이 일어난 건수 또는 횟수를 표현하는 이산확률분포

예를 들면 뭐랄까,,
1분 동안 꼬마아가씨가 나에게 반한 횟수?

포아송 회귀 모형(Poisson Regression Model)

포아송 회귀 모형

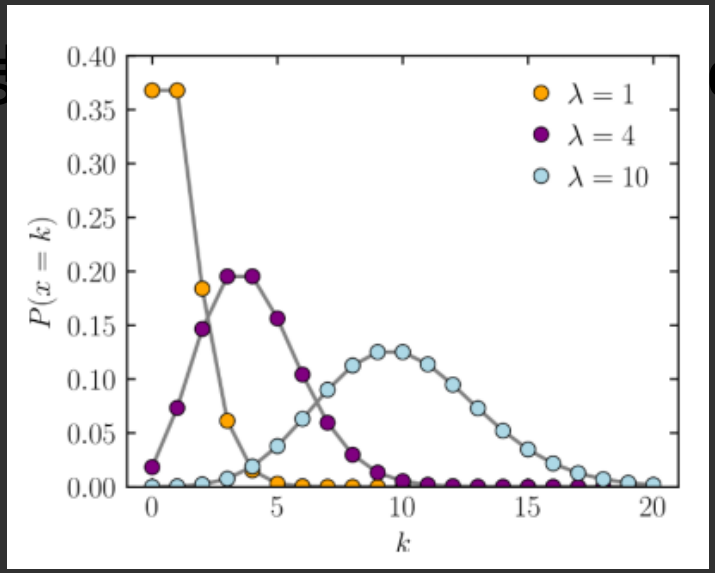
$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

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

포아송 회귀 모형, 왜 쓸까?

포아송 회귀 모형

포아송 회귀 모형



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

포아송 분포는 $\lambda < 10$일 때, 그래프 꼬리가 오른쪽으로 길게 치우침

만약 OLS 모형을 적합하면,

① 표준 오차와 유의수준의 편향문제 발생!

② 자연수여야 하는 count data가 음수로 반환되는 일이 발생

포아송 회귀 모형(Poisson Regression Model)

회귀 계수 β 해석

$$\log\left(\frac{\mu(x+1)}{\mu(x)}\right) = \beta\,, \qquad \frac{\mu(x+1)}{\mu(x)} = e^{\beta}$$

- $x+1$과 x 대입해서 빼주는 방식으로 해석
- (다른 설명 변수가 고정되어 있을 때)

  x 가 1단위 증가하면 기대도수 μ가 e^{β}배 증가한다!

포아송 회귀 모형(Poisson Regression Model)

$\hat{\mu}$ 값의 추정

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

$$\rightarrow \mu = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$$

추정된 회귀계수와 주어진 데이터를 대입하면

기대도수의 예측값 $\hat{\mu}$ 구할 수 있음!

포아송 회귀 모형

포아송 회귀 모형(Poisson Regression Model)

과대산포
(Overdispersion)

포아송 분포인데 분산이 평균보다 크게 나타나는 경우
회귀계수 추정량의 표준오차가 편향되어 작아짐

대안 모델 음이항 회귀 모형 (Negative Binomial Regression)

포아송 회귀 모형

포아송 회귀 모형(Poisson Regression Model)

음이항 회귀 모형

포아송 회귀 모형의 문제 ①

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

- 포아송 분포의 분산이 평균보다 크게 나타나는 경우

과대산포
(Overdispersion)

- 회귀계수 추정량의 표준오차가 편향되어 작아짐

대안 모델

- 분산이 평균보다 큰 값을 갖도록 하는 산포모수 D를 가짐

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

- 검정해서 D가 0이면 , $E(Y) = Var(Y) = \mu$ 이므로 과대산포 해결!

- 음이항 회귀모형(Negative Binomial Regression)

  → 포아송 회귀 모형과 동일

포아송 회귀 모형(Poisson Regession Model)

과대영
(Excess Zeros)

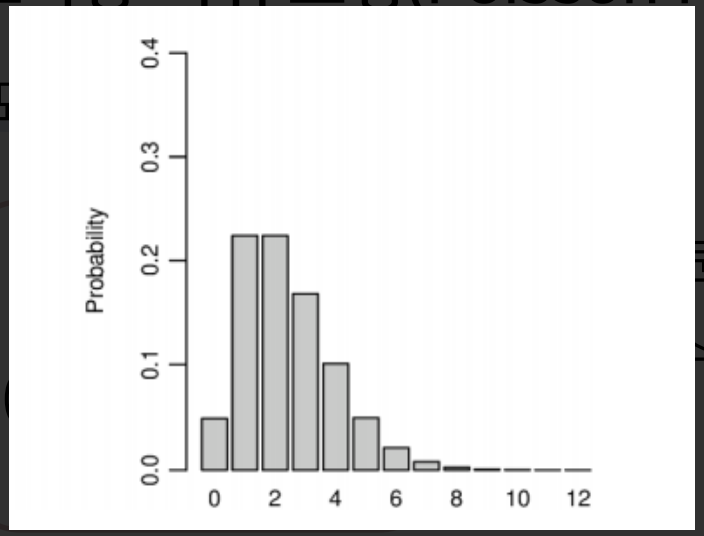
포아송 분포보다 0이 많이 나타나는 경우

Ex) 출석부 데이터

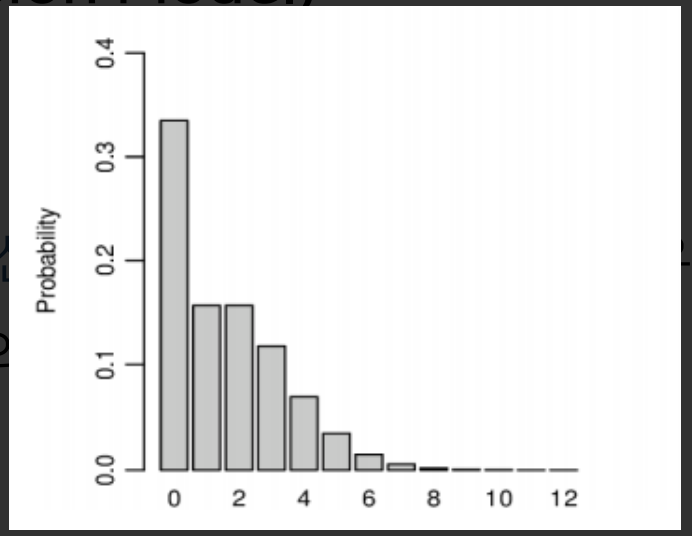
대안 모델 ZIP 회귀 모형(Zero-Inflated Regression)

포아송 회귀 모형
ZIP 회귀 모형

정상 포아송 분포



과대영 발생



포아송 회귀 모형(Poisson Regession Model)

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

다음 주 예고

1. 혼동 행렬
2. ROC / AUC
3. Sampling
4. Encoding