

범주형자료분석팀

2팀

조장희
위재성
김지현
조수미
송지현
김민지

INDEX

1. GLM

2. 유의성 검정

3. 로지스틱 회귀 모형

4. 다범주 로짓 모형

5. 포아송 회귀 모형

1

GLM

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(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

랜덤 성분

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

체계적 성분

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

연결 함수

$$g()$$

- GLM의 모양

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

삼단합체!




GLM 구성 성분

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

랜덤 성분

체계적 성분

연결 함수

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


Y 의 확률분포를 정해줌으로써 **반응변수 Y** 정의
가정한 확률분포의 **기댓값**인 μ 로 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

 $\mu(=E(Y))$

“랜덤성분의 분포를 일반화”

GLM은 랜덤성분의 분포로 아무 분포나 가질 수 있음

Y 의 확률분포를 정해줌으로써 **반응변수 Y** 정의
 가정한 확률분포의 기댓값인 μ 로 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

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

설명변수 X 를 명시하는 성분 X 의 선형결합으로 구성

GLM 구성 성분

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

랜덤 성분

체계적 성분

연결 함수

 $g()$

$g()$ 는 랜덤 성분과 체계적 성분을 연결하는 성분
두 성분의 범위를 맞춰주는 역할

GLM 구성 성분

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

랜덤 성분

체계적 성분

연결 함수

 $g()$

$g()$ 는 랜덤 성분과 체계적 성분을 연결하는 성분
두 성분의 범위를 맞춰주는 역할

만약 랜덤성분이 범주형 & 체계적 성분이 연속형이라면?

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

범주형

연속형 $(-\infty, \infty)$

범위가 맞지 않음

-> 연결함수가 이를 해결

GLM 구성 성분

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

랜덤 성분

체계적 성분

연결 함수

“랜덤성분의 함수 일반화”

연결함수 $g(\mu)$ 를 사용해서 아무 함수나 쓸 수 있음 $g()$ $g()$ 는 랜덤 성분과 체계적 성분을 연결하는 성분

두 성분의 범위를 맞춰주는 역할

만약 랜덤성분이 범주형 & 체계적 성분이 연속형이라면?

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

범주형

연속형 $(-\infty, \infty)$

범위가 맞지 않음

→ 연결함수가 이를 해결

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 Y 가 **연속형**일 때 사용
ex) 일반선형회귀모형

로그 연결함수

$$g(\mu) = \log(\mu)$$

반응변수 Y 가 **도수자료(count data)**일 때 사용
ex) 포아송 분포 / 음이항 분포

로짓 연결함수

$$g(\mu) = \log[\mu/(1-\mu)]$$

반응변수 Y 가 **이항분포**를 따를 때 사용
ex) 로지스틱 회귀

GLM의 모형 적합 : 주어진 데이터를 근거로 모형의 모수를 추정하는 것

GLM은 일반선형회귀처럼 LSE(최소제곱추정법)를 사용해 모형 적합 X
랜덤성분이 정규분포가 아닌 경우도 있기 때문!

최대가능도 추정법 사용
(method of maximum likelihood estimation)

GLM의 모형 적합 : 주어진 데이터를 근거로 모형의 모수를 추정하는 것

GLM은 일반선형회귀처럼 LSE(최소제곱추정법)를 사용해 모형 적합 X
랜덤성분이 정규분포가 아닌 경우도 있기 때문!

최대가능도 추정법 사용
(method of maximum likelihood estimation)



GLM의 모형 적합

가능도 (Likelihood)

관측값이 고정됐을 때, 그 관측값이 어떤 확률분포를 따를 가능성

가능도 함수 (Likelihood function)

결합확률밀도(질량)함수를 모수에 대해 정의한 함수

결합 확률밀도함수 $f(x;p)$ 와 같다고 생각하면 됨

(자세한 내용은 통계적추론입문 시간에..)

GLM의 모형 적합

*** 최대가능도 추정법**

가능도 (Likelihood)
(Maximum Likelihood Estimation)

관측값이 고정됐을 때, 그 관측값이 어떤 확률분포를 따를 가능성

→ 이 가능도함수가 최대가 되는 추정량 θ 를 찾는 방법

가능도 함수 (Likelihood function)

결합확률밀도(질량)함수를 모수에 대해 정의한 함수

결합 확률밀도함수 $f(x;p)$ 와 같다고 생각하면 됨

(자세한 내용은 통계적추론입문 시간에..)

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

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

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

GLM 특징

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

- (1) 선형 관계식 유지
- (2) 범위가 제한된 반응변수도 사용 가능
- (3) 독립성 가정만 필요
- (4) 오차항의 다양한 분포를 가정

2

유의성 검정

유의성 검정이란

유의성 검정

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

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

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

H_1 : 적어도 하나의 β 는 0이 아니다.

ML을 이용한 검정

왈드 검정

$$\text{검정 통계량 : } Z = \frac{\hat{\beta}}{\text{S.E}} \sim N(0,1) \text{ 또는 } Z^2 = \left(\frac{\hat{\beta}}{\text{S.E}}\right)^2 \sim \chi^2_1$$

$$\text{기각역 : } Z \geq |z_\alpha| \text{ 또는 } Z^2 \geq \chi^2_{\alpha,1}$$

추정값($\hat{\beta}$) 과 표준오차(S.E)만 사용하기 때문에 간단
범주형이나 소표본인 경우 가능도비 검정보다 검정력 ↓

ML을 이용한 검정

가능도비 검정

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

$$\text{기각역} : G^2 \geq \chi^2_{a,df}$$

가능도 함수의 **최댓값**을 이용해 비교

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

ML을 이용한 검정

가능도비 검정

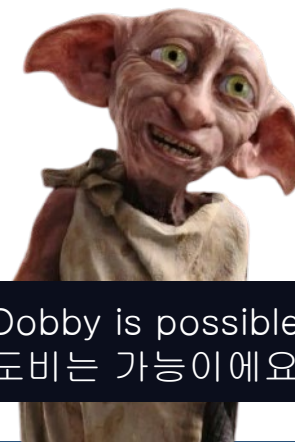
$$\text{검정 통계량} : G^2 = -2 \log \left(\frac{l_0}{l_1} \right) = -2(L_0 - L_1) \sim \chi^2_{df}$$

$$\text{기각역} : G^2 \geq \chi^2_{a, df}$$

가능도 함수의 **최댓값**을 이용해 비교

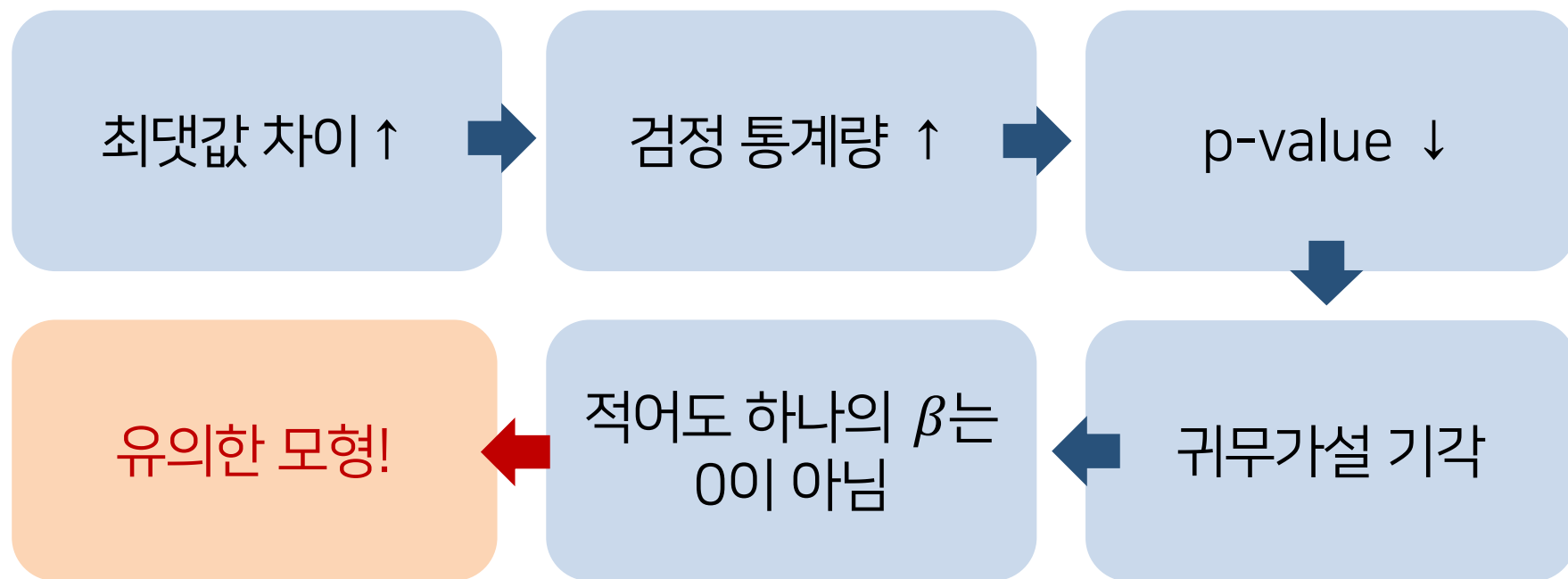
- l_0 : 귀무가설 하에서의 가능도함수
- l_1 : 전체공간 하에서의 가능도함수
- df : 귀무가설과 대립가설 모수 개수의 차이

(짹도비 조마루 씨)



Dobby is possible!
도비는 가능이에요!

ML을 이용한 검정



귀무가설 하(l_0) & 전체공간 하(l_1)에서의 가능도 함수에 대한 정보 사용

=> 가장 많은 정보 사용하기 때문에 **왈드 통계량보다 검정력 ↑**

이탈도

관심모형(M)

우리가 관심 있는 모형, 유의성을 검정할 모형

$$\text{범주팀의 행복정도 (Y)} = \beta_0 + \beta_1 \times \text{지현이의 귀여움}(x_1) + \beta_2 \times \text{패키지 난이도}(x_2)$$

포화모형(S)

모든 관측값에 대해 모수를 갖는 모형

$$\begin{aligned} \text{범주팀의 행복정도 (Y)} = & \beta_0 + \beta_1 \times \text{지현이의 귀여움}(x_1) + \beta_2 \times \text{패키지 난이도}(x_2) \\ & + \beta_3 \times \text{스터디 시간}(x_3) + \beta_4 \times \text{차농남의 드립력}(x_4) \end{aligned}$$

이탈도

관심모형(M)

우리가 관심 있는 모형, 유의성을 검정할 모형

$$\text{범주팀의 행복정도 (Y)} = \beta_0 + \beta_1 \times \text{지현이의 귀여움}(x_1) + \beta_2 \times \text{패키지 난이도}(x_2)$$

포화모형(S)

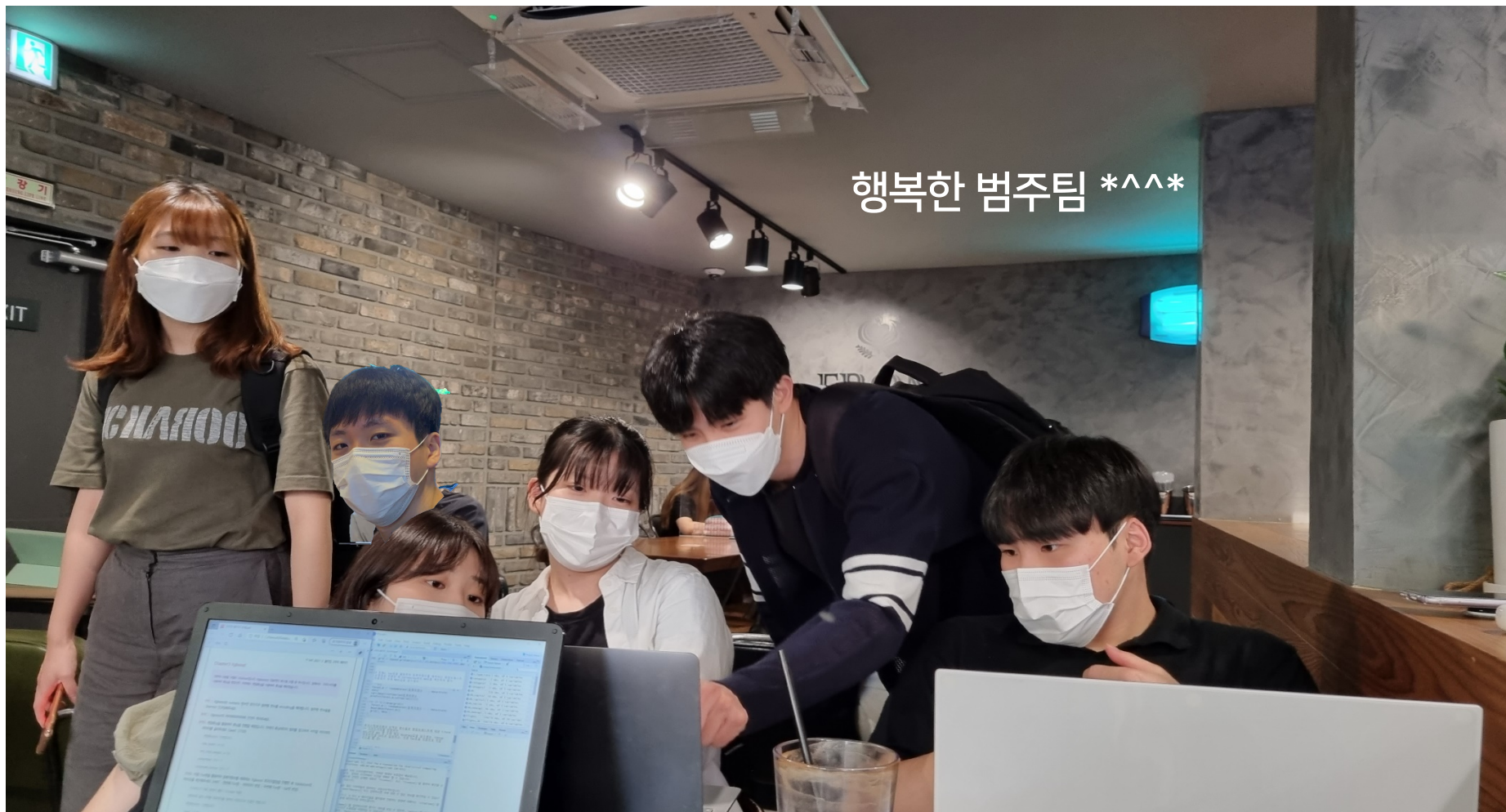
모든 관측값에 대해 모수를 갖는 모형



관심 포화

$$\begin{aligned} \text{범주팀의 행복정도 (Y)} = & \beta_0 + \beta_1 \times \text{지현이의 귀여움}(x_1) + \beta_2 \times \text{패키지 난이도}(x_2) \\ & + \beta_3 \times \text{스터디 시간}(x_3) + \beta_4 \times \text{차농남의 드립력}(x_4) \end{aligned}$$

이탈도



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 ($M \subset S$)

이탈도의 검정 flow



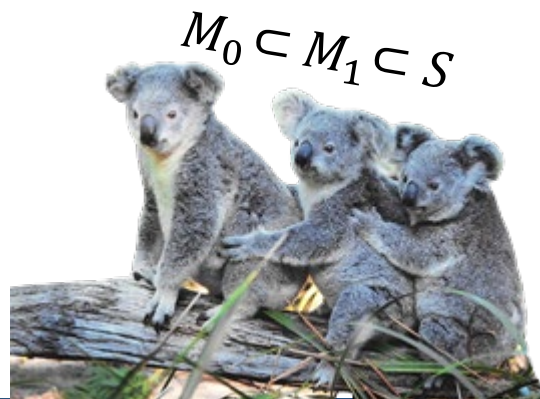
이탈도와 가능도비 검정의 관계

모형 간의 이탈도 차 = 가능도비 검정 통계량

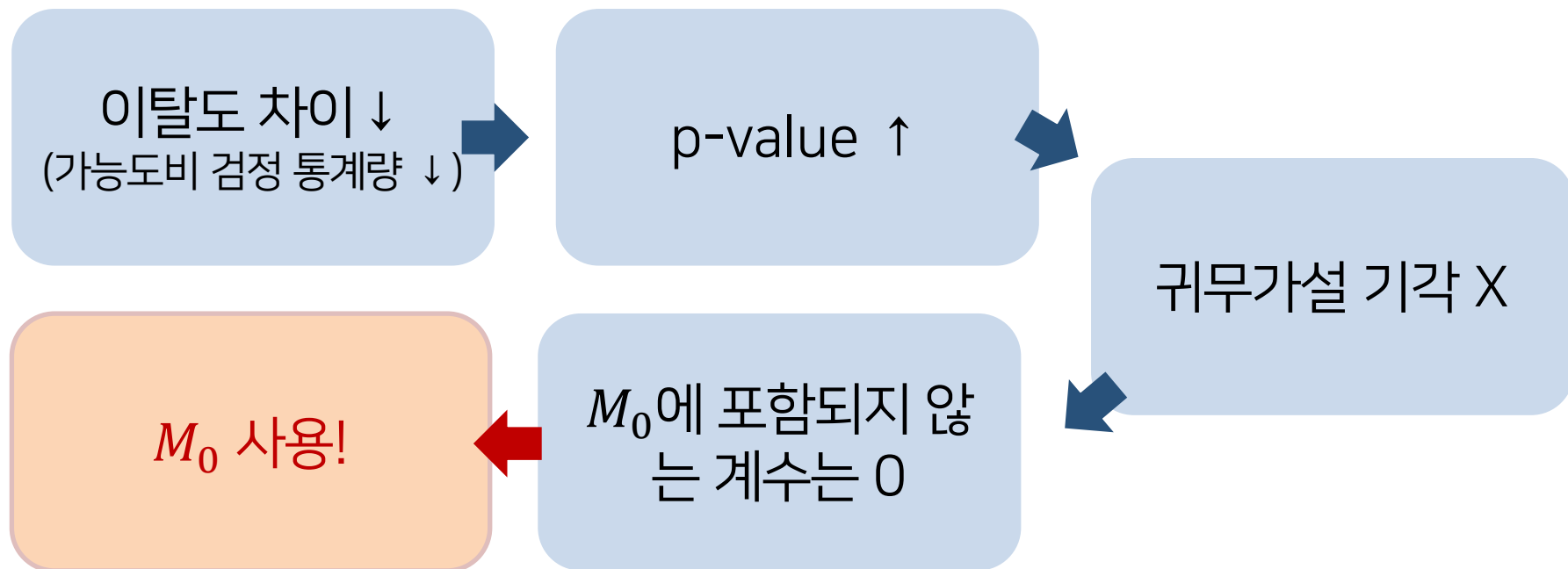
$$-2(L_0 - L_S) - \{-2(L_1 - L_S)\} = -2(L_0 - L_1)$$

모형이 내포(nested)될 때만 사용 가능 ($M_0 \subset M_1$)

- M_0 : 간단한 관심 모형
- M_1 : 복잡한 관심 모형
- S: 포화모형



이탈도의 검정 flow



3

로지스틱 회귀 모형

로지스틱 회귀 모형이란?

반응변수 Y가 이항자료일때 사용

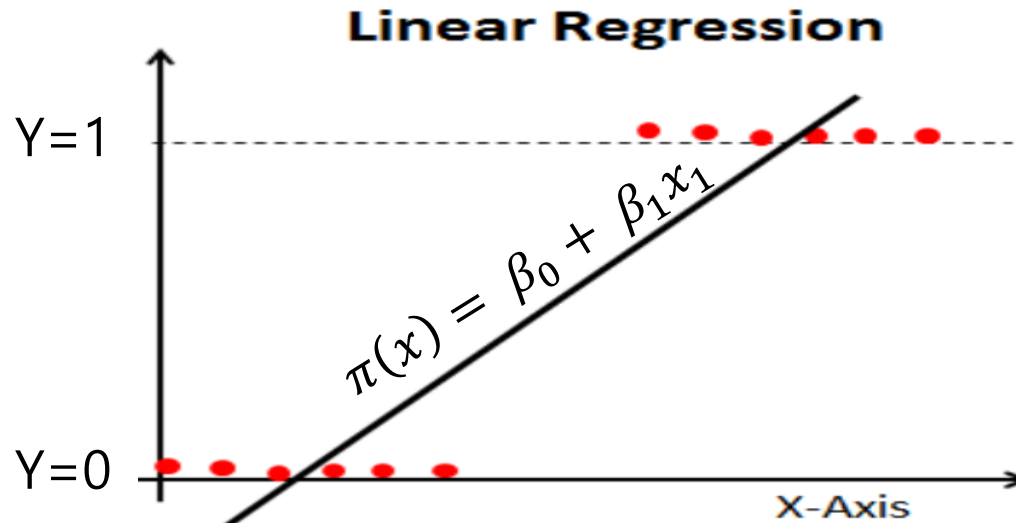
$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

반응 변수 Y가 성공 또는 실패의

이항분포를 따르는 변수이기에

일반 선형회귀는 사용할 수 없음 ...why?

로지스틱 회귀 모형이란?



단순선형회귀를 예시로,

$$\pi(x) = P(Y = 1|X = x) = \beta_0 + \beta_1 x_1$$

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

=> 범위가 일치하지 않음

로지스틱 회귀 모형이란?

STEP 1

종속변수가 범주 1이 될 확률로 가정한 식

⋮

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

좌변의 범위 = (0,1), 우변의 범위=(-∞, ∞)

STEP 2

좌변에 오즈를 설정

⋮

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

좌변의 범위 = (0,1), 우변의 범위=(-∞, ∞)

로지스틱 회귀 모형이란?

STEP 1

종속변수가 범주 1이 될 확률로 가정한 식

⋮

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

좌변의 범위 = (0,1), 우변의 범위=(-∞, ∞)

STEP 2

좌변에 오즈를 설정

⋮

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

좌변의 범위 = (0, ∞), 우변의 범위=(-∞, ∞)

로지스틱 회귀 모형이란?

종속변수가 범주 1이 될 확률로 가정한 식

STEP 3

오즈에 로그를 취해주어 로지스틱 회귀 모형 완성

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

좌변에 오즈를 설정

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

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

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

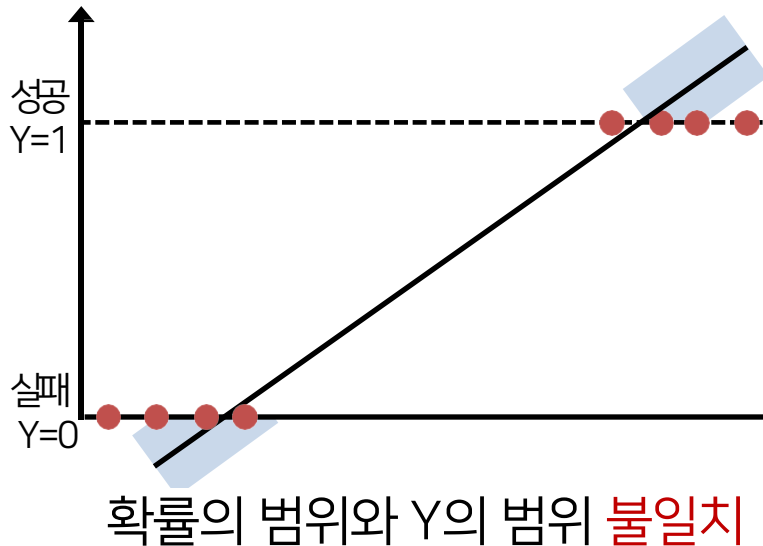
좌변의 범위 = (0, 1), 우변의 범위 = $(-\infty, \infty)$

로지스틱 회귀 모형이란?

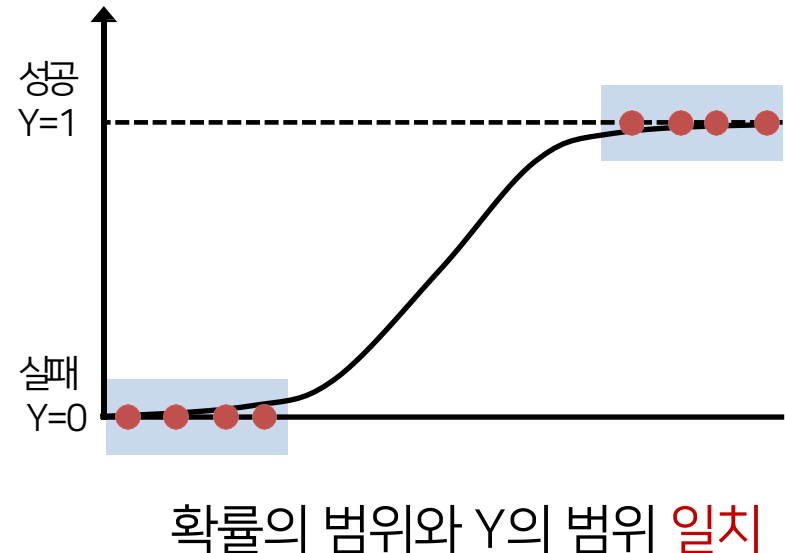
$$\pi(x) = P(Y = 1 | X = x) = \text{성공}$$

$$\pi(x) = P(Y = 0 | X = x) = \text{실패인 경우}$$

<일반회귀 모형>




<로지스틱 회귀모형>



로지스틱 회귀 모형이란?

로지스틱 회귀모형의 장점



로짓연결함수를 통해
범위 문제 해결

가정으로부터 자유로움
(독립성 가정만 만족하면 OK)

후향적 연구에도 사용 가능
(오즈, 오즈비)

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 β 의 해석

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

$|\beta|$ 가 증가함에 따라 변화율이 증가

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 $Y=0$

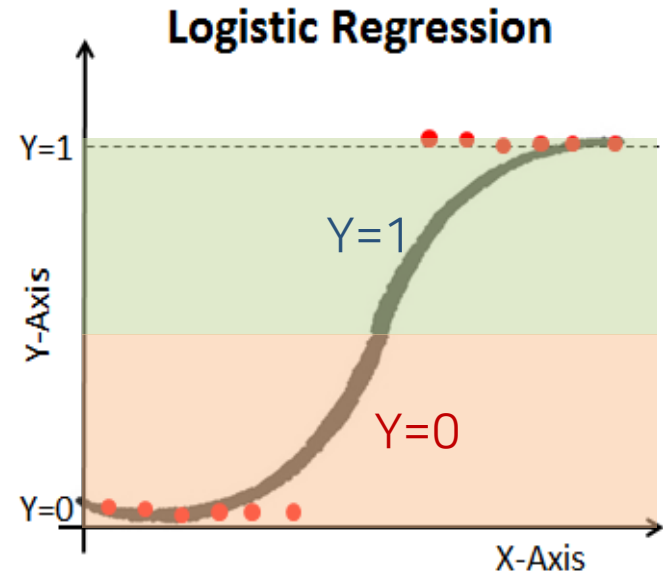
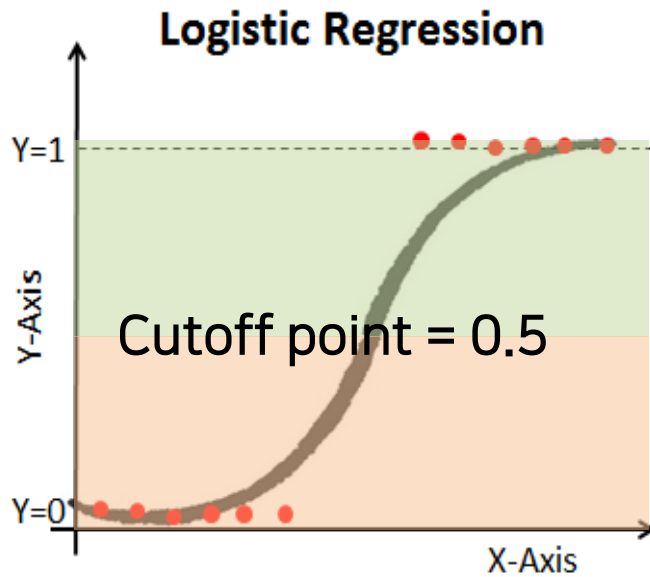
모수 β 의 해석

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

$|\beta|$ 가 증가함에 따라 변화율이 증가

로지스틱 회귀 모형의 해석

Cutoff point가 0.5일 때



로지스틱 회귀 모형의 해석

오즈로 해석

STEP 1

로지스틱 회귀식에 x 와 $x + 1$ 대입

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

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

STEP 2

대입한 식들의 차를 구함

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

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

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



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

로지스틱 회귀 모형의 해석

오즈로 해석

STEP 1

로지스틱 회귀식에 x 와 $x + 1$ 대입

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

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

STEP 2

대입한 식들의 차를 구함

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

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

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



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

로지스틱 회귀 모형의 해석

오즈로 해석

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

x 가 한 단위 증가할 때
 $Y=1$ 일 오즈가 e^{β} 배 증가함



오즈 잊은거 아니지,,?

로지스틱 회귀모형의 해석 (예시)

로지스틱 회귀 모형의 해석
EX)

오즈로 해석

$$\log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = 4 + 0.08x$$

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

$Y=1$ (연애성공), $Y=0$ (연애실패), $x = \text{키}$

x 가 한 단위 증가할 때

$Y=1$ 일 오즈가 e^{β} 배 증가함

$Y=1$ (연애성공)일 오즈가 $e^{0.08} = 1.08$ 배 증가

오즈 잊은거 아니
자?

4

다범주 로짓 모형

다범주 로짓 모형(Multicategory Logit Model)

다범주 로짓 모형

반응변수가 **다항분포**를 따르고,
연결함수가 로짓 연결 함수인 GLM

즉, 반응변수의 범주가 **3개 이상**인 모형

<기존 로짓 모형>

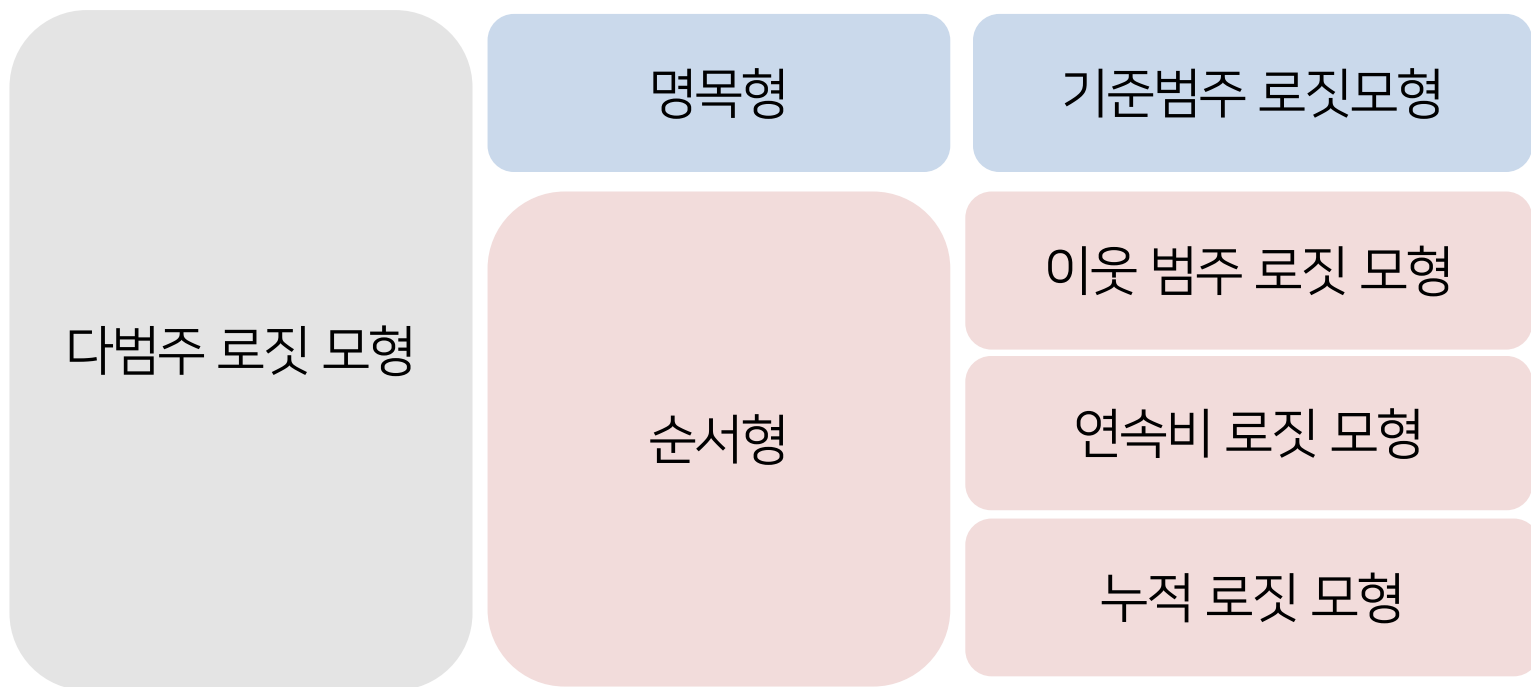


<다범주 로짓 모형>



다범주 로짓모형의 종류

<반응변수의 자료유형>



다범주 로짓모형의 종류

<반응변수의 자료유형>



기준범주 로짓모형(Baseline-Category Logit Model)

명목형 변수일때)

반응변수 Y의 여러 범주 중 하나를 **기준범주**로 선택



기준범주와 나머지 범주를 짝지어 로짓을 정의

나머지 범주 (제덕쿵야) 나머지 범주
기준 범주



=



기준범주 로짓모형(Baseline-Category Logit Model)

기준 범주 로짓 모형

범주 j일 때 x_1 의 회귀계수

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

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

J=2 라면, 로지스틱 회귀모형

기본범주 로짓모형 (예시) 다범주 로짓 모형



제니 기준!

기본범주 선택도 (Base-Category Logit Model)

Ex) 반응변수 Y: 최애 블랙핑크 멤버

범주=> 제니, 로제, 지수, 리사

$$\log\left(\frac{\pi_{\text{로제}}}{\pi_{\text{제니}}}\right) = \log\left(\frac{\pi_j}{\pi_1}\right) = \alpha_j + \beta_j^1 x_1 + \dots + \beta_j^p x_p, j = 1, \dots, (J - 1)$$

$$= 8 + 0.7x_1 + \dots - 0.2x_p$$

$$\log\left(\frac{\pi_{\text{지수}}}{\pi_{\text{제니}}}\right) = 4 + 0.02x_1 + \dots + 3x_p$$

• 기준 범주: 범주 J

• 나머지 범주: 범주 1, 범주 2, ..., 범주 J-1 => 계수의 값이 모두 다름

J=2 라면, 로지스틱 회귀모형

$$\log\left(\frac{\pi_{\text{리사}}}{\pi_{\text{제니}}}\right) = -0.6 + 11x_1 + \dots + 14x_p$$

기준범주 로짓모형(Baseline-Category Logit Model)

확률

$$\pi_j = \frac{e^{\alpha_j + \beta_j^1 x_1 + \dots + \beta_j^p x_p}}{\sum_{i=1}^J e^{\alpha_i + \beta_i^1 x_1 + \dots + \beta_i^p x_p}}, j = 1, \dots, (J - 1)$$

- π_j : 범주 j에 속할 확률
- 로지스틱 회귀 모형 공식을 변형!

기준범주 로짓모형(Baseline-Category Logit Model)

회귀계수 β 해석

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

- 오즈와 기준범주를 통해 해석!
- 해석: (다른 설명변수가 고정되어 있을 때)
 x_i 가 한 단위 증가하면 \rightarrow J범주 대신 j범주일 오즈가 e^{β^i} 만큼 증가

기준범주 로짓모형(Baseline-Category Logit Model)

회귀계수 β 해석

$$\begin{aligned} & \log\left(\frac{\pi_2}{\pi_J}\right) - \log\left(\frac{\pi_1}{\pi_J}\right) \\ &= (\alpha_2 + \beta_2^1 x_1 + \cdots + \beta_2^p x_p) - (\alpha_1 + \beta_1^1 x_1 + \cdots + \beta_1^p x_p) \\ &= [\alpha_2 - \alpha_1] + [(\beta_2^1 - \beta_1^1)x_1 + \cdots + (\beta_2^p - \beta_1^p)x_p] \end{aligned}$$

- 기준 범주와의 로짓끼리 빼서 해석!

- 해석: (다른 설명변수가 고정되어 있을 때)

x_i 가 한 단위 증가하면-> 1범주 대신 2범주일 오즈가 $e^{\beta_2^i - \beta_1^i}$ 만큼 증가

누적 로짓모형(Cumulative Logit Model)

순서형 범주일때)

범주를 순서대로 정렬한 뒤 **collapse** 과정 거침
=> 순서정보 고려!

collapse

순서대로 정렬된 범주들을 두 부분으로 나누는 과정

Cut point를 기준으로 나눔



누적 로짓모형(Cumulative Logit Model)

누적확률

누적확률에 로짓 연결함수를 씌운 모형

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

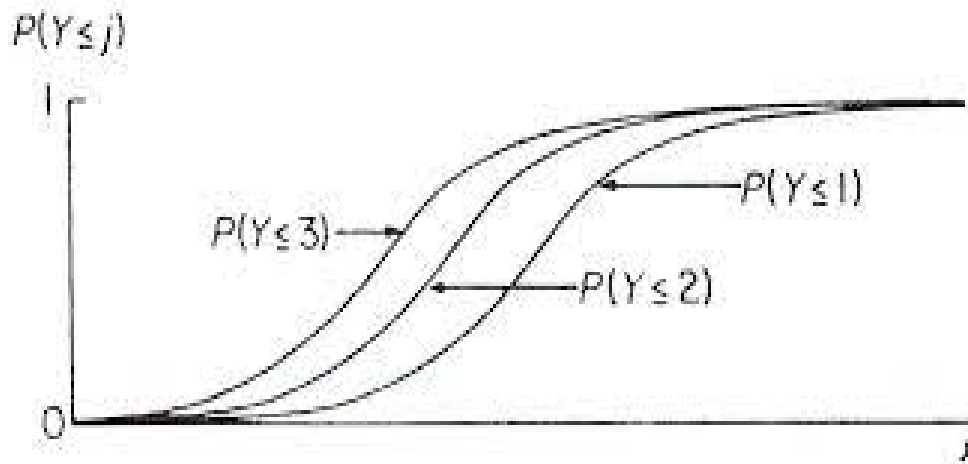
- α_j 가 다른 J-1 개의 로짓 방정식이 생김
- 회귀계수 β 에는 j 첨자X
- J-1개의 로짓 방정식에서의 회귀계수 β 의 효과가 동일하기 때문!
=> '비례 오즈 가정'

누적 로짓모형(Cumulative Logit Model)

누적확률

누적확률에 로짓 연결함수를 씌운 모형

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



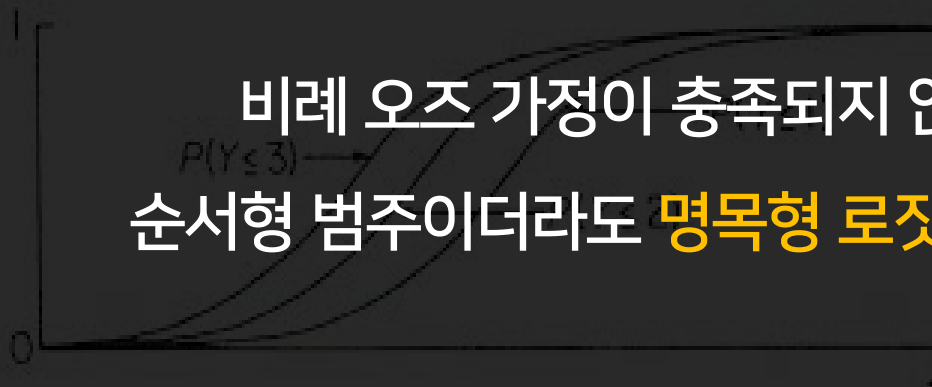
=> 일종의 평행

누적 비례 오즈 가정(Cumulative Logit Model)

누적화률 누적화률에 로짓 연결함수를 씌운 모형

Collapse 과정에서 cut point를 어디로 지정하든
 $\text{logit}[P(Y \leq j)] = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta_1 x_1 + \dots + \beta_p x_p,$
 회귀계수 β 의 효과는 동일하다.

$P(Y \leq j)$



비례 오즈 가정이 충족되지 않으면,
 \Rightarrow 일종의 평행
 순서형 범주이더라도 **명목형 로짓 모형**을 씀

비례 오즈 가정 (예시)

다범주 로짓 모형

누적 로짓 모형 (Cumulative Logit Model)

Ex) 반응변수 Y : 회귀팀 심OO씨의 매력으로 인한 지현의 팬심 정도이고,

누적확률

누적확률에 근거하여 범주를 사용 모형

소/중/대/극대의 순서형 범주를 가질 때

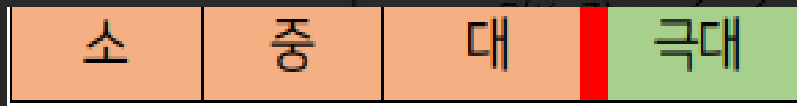
Cut point



Cut point



Cut point



0

x

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

$j = 1, \dots, (J-1)$

$$\text{logit}[P(Y \leq \text{소})] = 8 + 0.07x_1 + \dots + 0.6x_p$$

$$\text{logit}[P(Y \leq \text{중})] = -5 + 0.07x_1 + \dots + 0.6x_p$$

$$\text{logit}[P(Y \leq \text{대})] = 12 + 0.07x_1 + \dots + 0.6x_p$$

=> 일종의 평행

$P(Y \leq 1)$

누적 로짓모형(Cumulative Logit Model)

회귀계수 β 해석

$$\begin{aligned}\text{logit}[P(Y \leq j)] &= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p, j = 1, \dots, (J - 1) \\ &= e^{\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p}, j = 1, \dots, (J - 1)\end{aligned}$$

- 누적 확률의 로그 오즈 보고 해석
- 해석: (다른 설명 변수가 고정되어 있을 때)
 x 가 1단위 증가하면, $Y > j$ 에 비해 $Y \leq j$ 일 오즈가 e^β 만큼 증가한다!

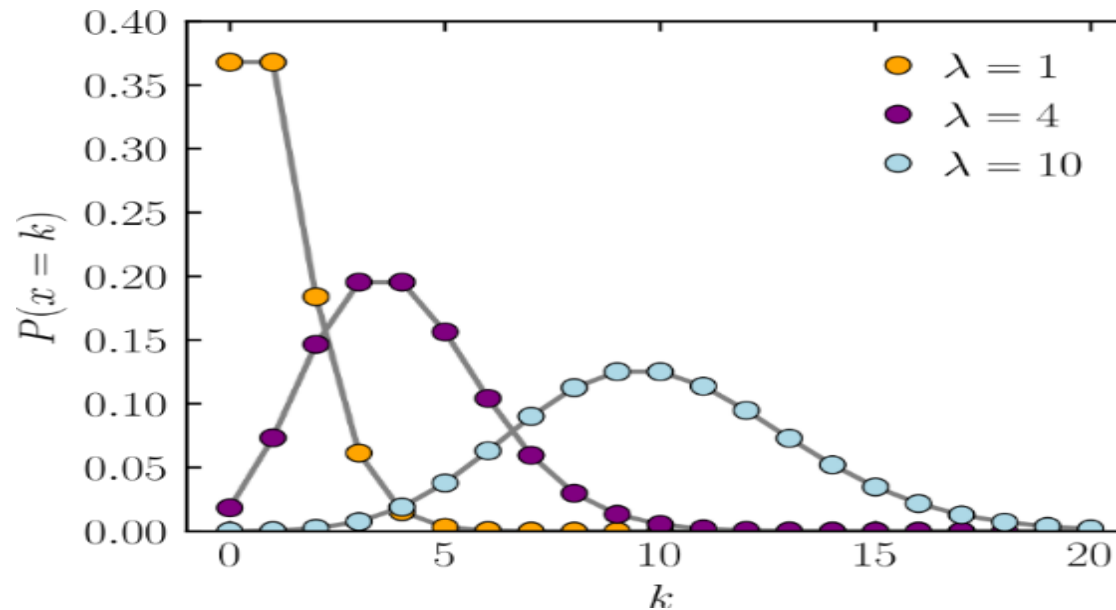
5

포아송 회귀모형

포아송 회귀 모형 (Poisson Regression Model)

포아송 분포

주어진 시간 또는 영역에서 어떤 사건의 발생횟수에 대한 확률모형.



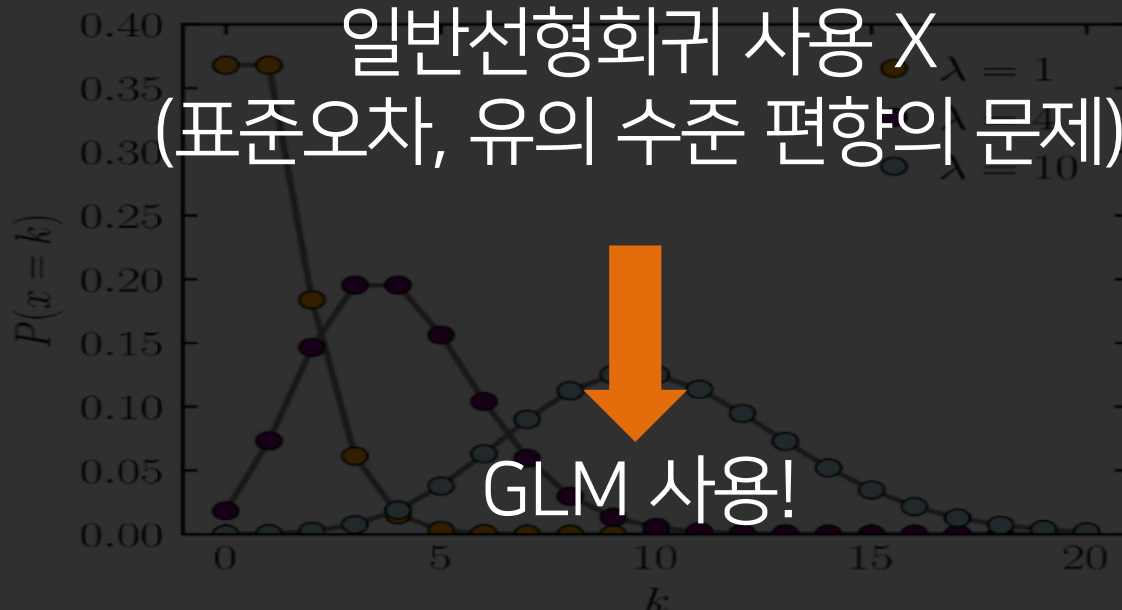
포아송 회귀 모형 (Poisson Regression Model)

포아송 분포는 정규성, 등분산성 만족 X

포아송 분포

주어진 시간 또는 영역에서 어떤 사건의 발생횟수에 대한 확률모형.

일반선형회귀 사용 X
(표준오차, 유의 수준 편향의 문제)



포아송 회귀 모형 (Poisson Regression Model)

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

- 반응변수 Y가 도수자료인 경우의 회귀모형
- 랜덤성분 : 포아송 분포
- 연결함수 : 로그 연결함수

※ 도수자료의 범위($0 \sim \infty$)와 체계적 성분의 범위($-\infty \sim \infty$)를 맞추기 위해 사용

포아송 회귀 모형 (Poisson Regression Model)

- ① 도수로 나타내서 해석

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

- ② 차이를 이용해서 해석

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



포아송 회귀 모형 식에서 $x+1$ 과 x 를 대입해서 뺀

포아송 회귀 모형 (Poisson Regression Model)

- ① 도수로 나타내서 해석

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

- ② 차이를 이용해서 해석

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



포아송 회귀 모형 식에서 $x+1$ 과 x 를 대입해서 뺀

x 가 한 단위 증가할 때 기대도수 μ 가 e^β 배만큼 증가

포아송 회귀 모형 (Poisson Regression Model)

만약 반응변수 Y : 연애 횟수, X : 학점 이고,

① 도수로 나타내서 해석

회귀모형 식 : $\log(\mu) = -2 + 0.01x$ 이라면

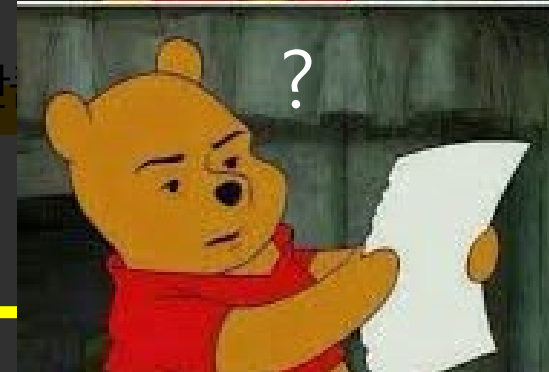
$$\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

② 차이를 이용해서 해석

학점(x)이 1점씩 증가할 때,

기대도수 μ 가 $\frac{\mu(x+1)}{\mu(x)} = e^{0.01} \approx 1.01$ 배만큼 증가.

x 가 한 단위 증가할 때 기대도수 μ 가 e^β 배만



포아송 회귀 모형 (Poisson Regression Model)

과대 산포 (Overdispersion) 문제

- 포아송 분포는 평균과 분산이 같다는 **등산포**를 가정
- 그러나 일반적으로 데이터는 분산이 평균보다 더 큼



표준오차를
더 작게 왜곡

대안

Quasi-Poisson
모형

음이항 회귀 모형

포아송 회귀 모형 (Poisson Regression Model)

Quasi-Poisson 모형

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

- 포아송과 비슷함. : 포아송 랜덤성분, 로그연결함수
- 분산이 평균보다 클 수 있도록 산포모수 θ 를 추가
- 분산이 평균과 선형관계임을 가정

$$E(Y) = \mu, \quad Var(Y) = \theta\mu$$

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

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

포아송 회귀 모형 (Poisson Regression Model)

음이항 회귀모형

Quasi-Poisson 모형과 음이항 모형은

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

포아송 모형과 회귀계수 β 값 **동일**

그러나 표준오차 \uparrow , 검정이 정확해짐

- 음이항 랜덤효과, 로그-링크함수
- 음이항 분포는 이미 분산이 평균보다 큰 상태
- 분산이 평균과 비선형관계임을 가정, 산포모수 D 사용

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

포아송 회귀 모형 (Poisson Regression Model)

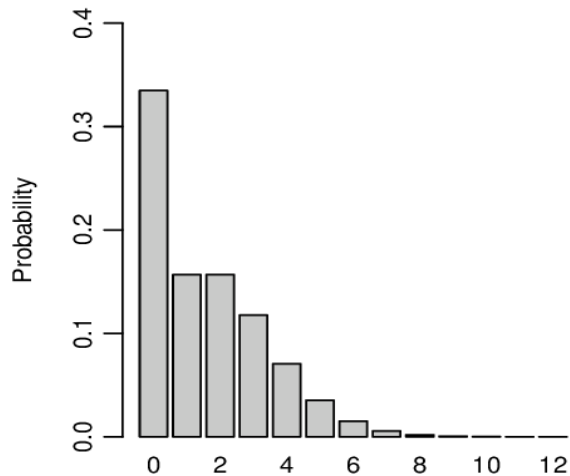
과대영 문제

: "0" 이 많이 나타나는 문제

Ex) 로또 당첨자 수 데이터

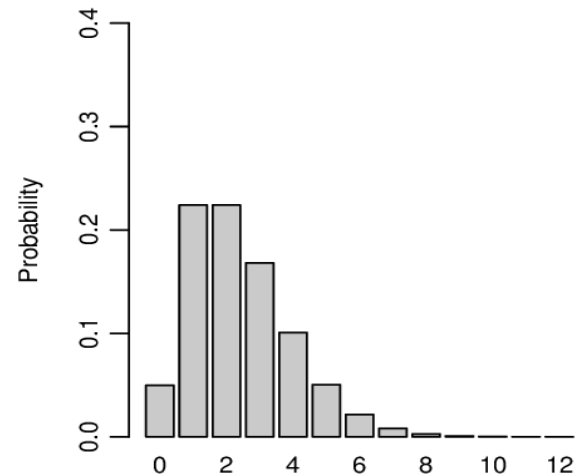


ZIP($\pi = 0.3, \lambda = 3$)



<과대영 문제 발생 그래프>

ZIP($\pi = 0, \lambda = 3$) = Poi($\lambda = 3$)



<일반 포아송 분포 그래프>

영과잉 포아송 모형(ZIP)으로 해결!

포아송 회귀 모형 (Poisson Regression Model)

ZIP의 반응 변수 Y 는 0의 값이 발생하는 점확률분포와

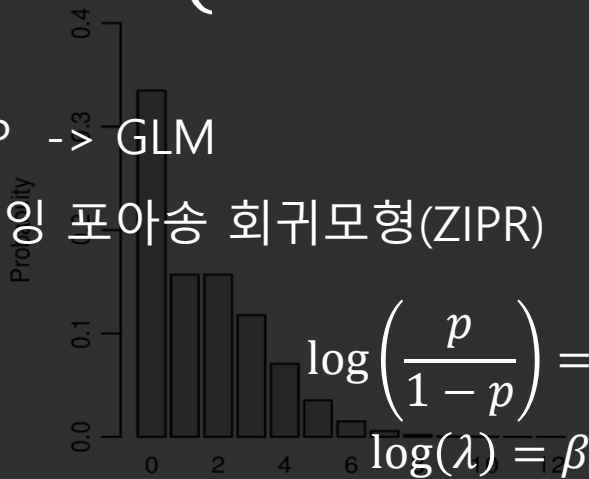
과대영 문제 0보다 큰 정수값을 갖는 포아송 분포의 혼합구조

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

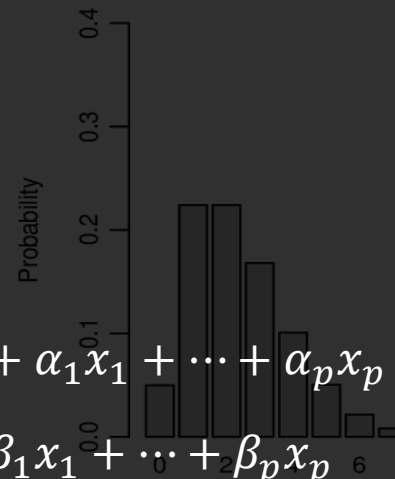
$\text{ZIP}(\pi=0.5, \lambda=3) = \text{Poi}(\lambda=3)$

※ZIP -> GLM

영과잉 포아송 회귀모형(ZIPR)



<과대영 문제 발생 그래프>



<일반 포아송 분포 그래프>

로짓연결함수

로그연결함수

다음 주 예고

1. 혼동행렬
2. 평가지표
3. 샘플링
4. 인코딩