범주형자료분석팀

2팀
임지훈
안은선
강세현
심현구
하희나

INDEX

1

GLM이란?

1 GLM이란?

GLM의 정의

GLM (일반화 선형 모형)

연속형 반응변수들에 대한 모형을 확장시켜

다양한 형태의 반응 변수에 대한 모형들을 **모두 포함한 모형의 집합**

일반 선형회귀모형의 한계를 극복해

반응변수가 **범주형 자료**이거나 **도수 자료**인 경우에도

설명변수와 반응변수 간의 최적의 관계식 추정 가능

1 GLM이란?

GLM의 필요성

자료의 오차항이 **정규성**을 **만족하지 않는 경우**

일반 선형회귀모형처럼 **LSE으로 모형 적합 불가능**

GLM은 MLE로 모형을 적합하기 때문에

정규분포 외의 다른 확률분포를 따르는 반응변수에 대한 분석 가능

1 GLM이란?

GLM의 필요성

분할표는 **주어진 범주형** 변수들 간의 **연관성**만 파악 가능

GLM은 **범주형** 자료와 **연속형** 자료 간의 **연관성**도 **파악** 가능하며

새로운 설명변수 값에 따른 **반응변수의 값 예측** 가능

1 GLM이란?

GLM의 구성성분

GLM

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

GLM의 구성성분

랜덤 성분	체계적 성분	연결 함수
$\mu\,(= E(Y))$	$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$	$g(\cdot)$

1 GLM이란?

GLM의 구성성분

랜덤 성분 (Random Component)

가정한 반응변수 Y의 확률분포의 기대값

반응변수 Y의 확률분포를 정해줌으로써 Y를 정의

반응변수	확률분포	표기
이진형	이항분포	$\pi(x)$
연속형	정규분포	μ
도수자료	포아송분포	μ 또는 λ

GLM의 반응변수로
지수족에 해당하는
확률분포만을
사용할 수 있음을 주의!

1 GLM이란?

GLM의 구성성분

체계적 성분 (Systematic Component)

설명변수 X들을 명시하는 성분

X의 **선형결합**의 형태로 표현

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

교호작용을 설명하는 항이나 **곡선효과**를 나타내는 항 포함 가능

$x_i = x_a x_b$

$x_i = x_a^2$

1 GLM이란?

GLM의 구성성분

연결 함수 (Link Function)

랜덤 성분과 체계적 성분을 연결하여 **두 성분의 범위를 조정**

체계적 성분의 범위는 제약이 없어 **−∞ ~ ∞의 범위**를 따르지만

랜덤 성분은 분포에 따라 **범위**에 **제약**이 발생해 **둘의 범위가 불일치**

연결 함수를 이용해 체계적 성분과 랜덤 성분의 **범위**를 **일치**시킴

1 GLM이란?

GLM의 구성성분

연결 함수 (Link Function)

랜덤 성분과 체계적 성분을 연결하여 **두 성분의 범위를 조정**

종류	반응변수	표기
항등 연결 함수 (Identity Link)	연속형 자료	$g(\mu) = \mu$
로그 연결 함수 (Log Link)	도수 자료 포아송 · 음이항 분포	$g(\mu) = \log(\mu)$
로짓 연결 함수 (Logit Link)	0~1 사이의 값 이항 분포	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

1 GLM이란?

GLM의 특징

오차항의 다양한 분포 가정 가능

반응변수의 오차항이 가진 성질에 따라
정규분포 외의 분포도 정의할 수 있음

선형 관계식 유지

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

회귀계수 β와 Y 간의
선형성을 유지하기 때문에 해석 용이

독립성 가정만 필요

오차항에 대한 독립성만 만족하면 됨
→ 자기상관성 검정 필요
ex) 더빈-왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 이용해 양변의 범위를 일치
→ 제한된 범위의 반응변수 사용 가능
ex) 범주형 자료, 도수 자료

1 GLM이란?

GLM의 특징

오차항의 다양한 분포 가정 가능

반응변수의 오차항이 가진 성질에 따라
정규분포 외의 분포도 정의할 수 있음

선형 관계식 유지

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

회귀계수 β와 Y 간의
선형성을 유지하기 때문에 해석 용이

독립성 가정만 필요

오차항에 대한 독립성만 만족하면 됨
→ 자기상관성 검정 필요
ex) 더빈-왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 이용해 양변의 범위를 일치
→ 제한된 범위의 반응변수 사용 가능
ex) 범주형 자료, 도수 자료

1 GLM이란?

GLM의 특징

오차항의 다양한 분포 가정 가능

반응변수의 오차항이 가진 성질에 따라
정규분포 외의 분포도 정의할 수 있음

선형 관계식 유지

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

회귀계수 β와 Y 간의
선형성을 유지하기 때문에 해석 용이

독립성 가정만 필요

오차항에 대한 독립성만 만족하면 됨

→ 자기상관성 검정 필요

ex) 더빈-왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 이용해 양변의 범위를 일치

→ 제한된 범위의 반응변수 사용 가능

ex) 범주형 자료, 도수 자료

1 GLM이란?

GLM의 특징

오차항의 다양한 분포 가정 가능

반응변수의 오차항이 가진 성질에 따라
정규분포 외의 분포도 정의할 수 있음

선형 관계식 유지

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

회귀계수 β와 Y 간의
선형성을 유지하기 때문에 해석 용이

독립성 가정만 필요

오차항에 대한 독립성만 만족하면 됨
→ 자기상관성 검정 필요
ex) 더빈-왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 이용해 양변의 범위를 일치
→ 제한된 범위의 반응변수 사용 가능
ex) 범주형 자료, 도수 자료

1 GLM이란?

GLM의 모형 적합

GLM은 회귀의 4가지 가정을 모두 충족하지 못하기 때문에 LSE로 모형 적합 불가능

정규성, 선형성, 독립성, 등분산성

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta) \xrightarrow{\log} L(\theta|x) = \log p(x|\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

MLE를 이용해 모형 적합

각 데이터들의 가능도를 모두 곱한 가능도 함수가 최대가 되는 추정량을 찾음

고정된 관측값이 어떤 확률분포를 따를 가능성

2

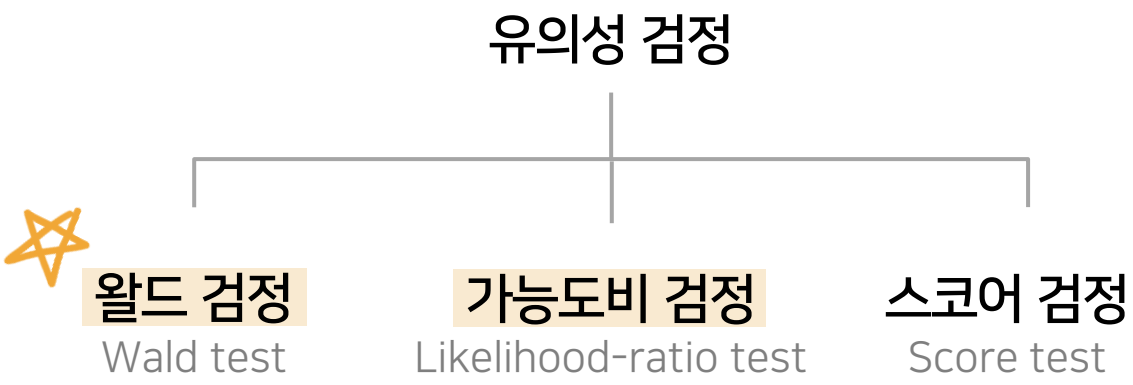
유의성 검정

유의성 검정

유의성 검정

모형의 모수에 대한 **추정값**이 **유의**한지

혹은 **축소 모형의 적합도**가 좋은지를 판단하는 검정



2 유의성 검정

왈드 검정

가설

귀무가설 $H_0 : \beta_j = 0$

대립가설 $H_1 : \beta_j \neq 0$

검정통계량

$$Z = \frac{\hat{\beta}}{SE} \sim N(0,1) \text{ 혹은 } Z^2 = \left(\frac{\hat{\beta}}{SE}\right)^2 \sim \chi_1^2$$

기각역

$$Z \geq |z_a| \text{ 혹은 } Z^2 \geq \chi_{\alpha,1}^2$$

왈드 검정의 장점

✓ 회귀 계수에 대한 추정값과 표준오차만 사용하여 통계량을 구해 간단

2 유의성 검정

왈드 검정

가설

귀무가설 $H_0 : \beta_j = 0$

대립가설 $H_1 : \beta_j \neq 0$

검정통계량

$$Z = \frac{\hat{\beta}}{SE} \sim N(0,1) \text{ 혹은 } Z^2 = \left(\frac{\hat{\beta}}{SE}\right)^2 \sim \chi_1^2$$

기각역

$$Z \geq |z_\alpha| \text{ 혹은 } Z^2 \geq \chi_{\alpha,1}^2$$

왈드 검정은 범주형 자료이거나 소표본인 경우 검정력 감소

가능도비 검정을 이용해 GLM 유의성 검정 시행

왈드 검정의 장점

✓ 회귀 계수에 대한 추정값과 표준오차만 사용하여 통계량을 구해 간단

2 유의성 검정

가능도비 검정

가설

귀무가설 H_0 :

$\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_k = 0}$

대립가설 H_1 :

적어도 하나의 β는 0이 아니다

검정통계량

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

기각역

$$G^2 \geq \chi^2_{\alpha,df}$$

귀무가설 하에서의 **가능도 함수**와 **전체공간** 하에서의 **가능도 함수**의 **차이**를 이용

모수에 대한 아무런 제약이 없는 상태

2 유의성 검정

가능도비 검정

가설

귀무가설 H_0 :

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

대립가설 H_1 :

적어도 하나의 β는 0이 아니다

검정통계량

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

기각역

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

MLE로 계산

l_0 모수가 귀무가설을 만족할 때 가능도 함수의 최대값

l_1 모수가 아무런 제약이 없을 때 가능도 함수의 최대값

df 두 가설의 모수의 개수의 차이

귀무가설 하에서의 가능도 함수와 전체공간 하에서의 가능도 함수의 차이를 이용

모수에 대한 아무런 제약이 없는 상태

2 유의성 검정

가능도비 검정

가설

귀무가설 H_0 :

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

대립가설 H_1 :

적어도 하나의 β는 0이 아니다

검정통계량

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

기각역

$$G^2 \geq \chi^2_{\alpha,df}$$

검정 과정

l_0와 l_1의 차이가 큼 → 검정통계량의 값이 큼 → p-value 값이 작음

→ 귀무가설 기각 → 적어도 하나의 β는 0이 아님 → 모형의 모수 추정값이 유의

2 유의성 검정

가능도비 검정

가설

귀무가설 H_0 :

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

대립가설 H_1 :

적어도 하나의 β는 0이 아니다

검정통계량

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

가능도비 검정은 귀무가설과 전체공간 하에서의

가능도 함수에 대한 정보를 모두 사용

$$G^2 \geq \chi^2_{\alpha,df}$$

다른 유의성 검정보다 가장 많은 정보 이용

검정력과 신뢰도가 높음

검정 과정

l_0와 l_1의 차이가 큼 → 검정통계량의 값이 큼 → p-value 값이 작음

→ 귀무가설 기각 → 적어도 하나의 β는 0이 아님 → 모형의 모수 추정값이 유의

2 유의성 검정

이탈도

이탈도 (Deviance)

관심있는 모형(M)과 **포화모형**(S)을 비교해 **모형의 적합성**을 판단

유의성을 검정하고자 하는 모형

모든 관측값에 대하여
모수를 갖는 가장 복잡한 모형

이탈도의 가설

귀무가설 H_0 : 관심모형에 속하지 않는 모수는 모두 0이다 (= 관심모형 사용)

대립가설 H_1 : 관심모형에 속하지 않은 모수 중 적어도 하나는 0이 아니다
(= 관심모형 사용 불가)

2 유의성 검정

이탈도

이탈도 (Deviance)

관심있는 모형(M)과 **포화모형**(S)을 비교해 **모형의 적합성**을 판단

유의성을 검정하고자 하는 모형

모든 관측값에 대하여
모수를 갖는 가장 복잡한 모형

이탈도의 가설

귀무가설 H_0 : 관심모형에 속하지 않는 모수는 모두 0이다 (= 관심모형 사용)

대립가설 H_1 : 관심모형에 속하지 않은 모수 중 적어도 하나는 0이 아니다
(= 관심모형 사용 불가)

2 유의성 검정

이탈도

가설

귀무가설 H_0 : 관심모형에 속하지 않는 모수는 모두 0이다

대립가설 H_1 : 관심모형에 속하지 않은 모수 중 적어도 하나는 0이 아니다

검정통계량

$$-2\log\left(\frac{l_m}{l_s}\right) = -2(L_M - L_s)$$

관심모형은 포화모형에 내포된 관계여야 함

이탈도는 포화모형에는 있지만 관심모형에는 없는 계수들이 0인지를 확인하는 것이기 때문!

2 유의성 검정

이탈도

가설

귀무가설 H_0 : 관심모형에 속하지 않는 모수는 모두 0이다

대립가설 H_1 : 관심모형에 속하지 않은 모수 중 적어도 하나는 0이 아니다

검정통계량

$$-2\log\left(\frac{l_m}{l_s}\right) = -2(L_M - L_s)$$

검정 과정

두 가능도 함수의 최댓값 간의 **차이가 큼** → 이탈도의 값이 큼 → p-value 값이 작음

→ **귀무가설 기각** → **관심모형이 적합하지 않음**

2 유의성 검정

이탈도와 가능도비 검정의 관계

두 관심모형 간의 이탈도 값의 차이 = 가능도비 검정 통계량

단순한 형태의 관심모형 ------ M_0의 이탈도 $-$ M_1 의 이탈도 ------ 복잡한 형태의 관심모형

$$= 2(L_0 - L_s) - \left(-2(L_1 - L_s)\right) = -2(L_0 - L_1)$$

L_s: 포화 모형(M_0, M_1을 포함)에서 얻은 로그 가능도 함수의 최댓값

이탈도 차이를 통해 관심모형과 관심모형 간의 비교가 가능해짐

어떤 관심모형이 더 좋은 모형인지를 판단 가능!

2 유의성 검정

이탈도와 가능도비 검정의 관계

두 관심모형 간의 이탈도 값의 차이 = 가능도비 검정 통계량

$\boldsymbol{M_0}$의 이탈도 $-$ $\boldsymbol{M_1}$ 의 이탈도

$$= 2(L_0 - L_s) - \left(-2(L_1 - L_s)\right) = \boldsymbol{-2(L_0 - L_1)}$$

이탈도 차이를 통해 **관심모형과 관심모형 간의 비교**가 **가능**해짐

어떤 관심모형이 더 좋은 모형인지를 판단 가능!

2 유의성 검정

이탈도와 가능도비 검정의 관계

두 관심모형 간의 이탈도 값의 차이 = 가능도비 검정 통계량

M_0의 이탈도 $- M_1$의 이탈도

$= 2(L_0 - L_S) - (-2(L_1 - L_S)) = -2(L_1 - L_0)$

이탈도를 활용하기 때문에

M_0은 M_1에 내포된 모형이어야 함

만약 내포된 관계가 아니라면

AIC, BIC와 같은 모형 선택을 위한 측도들을 활용해 모형 비교

자세한 내용은 회귀분석팀 클린업 참고!

이탈도 차이를 통해 관심모형과 관심모형 간의 비교가 가능해짐

어떤 관심모형이 더 좋은 모형인지를 판단 가능!

2 유의성 검정

이탈도와 가능도비 검정의 관계

두 관심모형 간의 이탈도 값의 차이 = 가능도비 검정 통계량

$\boldsymbol{M_0}$의 이탈도 − $\boldsymbol{M_1}$ 의 이탈도

$$= 2(L_0 - L_s) - \left(-2(L_1 - L_s)\right) = \boldsymbol{-2(L_0 - L_1)}$$

검정 과정

두 관심모형 간의 이탈도 차이가 큼 → 가능도비 검정통계량의 값이 작음

→ p-value 값이 큼 → **귀무가설 기각하지 못함**

→ M_0에 포함되지 않은 모수들이 모두 0임 → **간단한 관심모형 $\boldsymbol{M_0}$이 더 적합**

3

GLM 모형의 종류

GLM 모형의 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		

GLM 모형의 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		

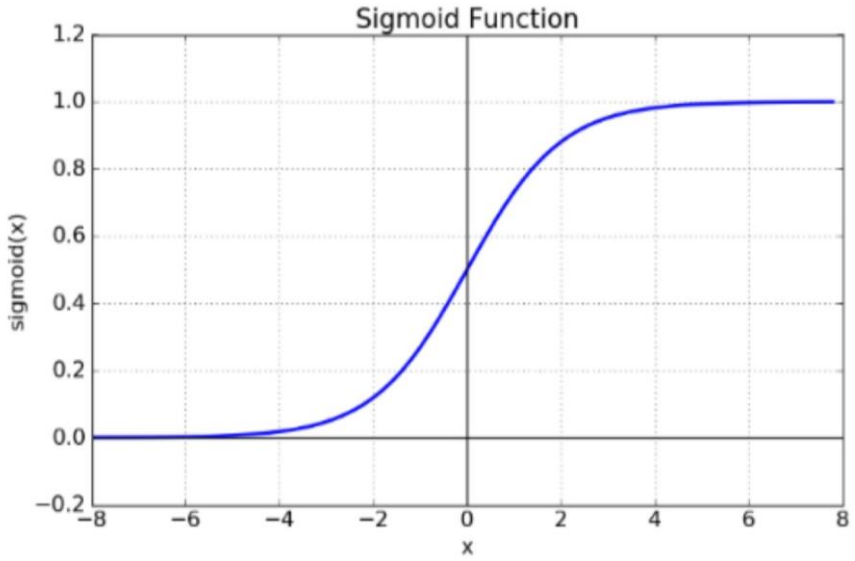
4

로지스틱 회귀모형

로지스틱 회귀모형

로지스틱 회귀모형 (Logistic Regression)

반응변수 Y가 **이항자료**일 때의 회귀모형



로지스틱 회귀모형
Logistic Regression

- ✓ $\pi(x)$와 x의 **비선형 관계**를 나타냄
- ✓ 일반선형모델로 설명할 수 없는 **이항변수**와 **연속형 변수들 간의 관계**를 **GLM** 형태로 표현

로지스틱 회귀모형의 장점

로지스틱 회귀모형의 장점

✓ 이항 변수와 연속형 변수 간의 범위 일치

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

좌변 범위 : 0~1 ≠ 우변 범위 : $-\infty \sim \infty$

좌변을 오즈 형태로 만든 후 로그 취하기

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

좌변 범위 : $-\infty \sim \infty$ = 우변 범위 : $-\infty \sim \infty$

로지스틱 회귀모형의 장점

로지스틱 회귀모형의 장점

✓ 이항 변수와 연속형 변수 간의 범위 일치

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

좌변 범위 : 0~1 ≠ 우변 범위 : $-\infty \sim \infty$

좌변을 오즈 형태로 만든 후 로그 취하기

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

좌변 범위 : $-\infty \sim \infty$ = 우변 범위 : $-\infty \sim \infty$

로지스틱 회귀모형의 장점

로지스틱 회귀모형의 장점

✓ 이항 변수와 연속형 변수 간의 범위 일치

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

좌변 범위 : 0~1 ≠ 우변 범위 : $-\infty \sim \infty$

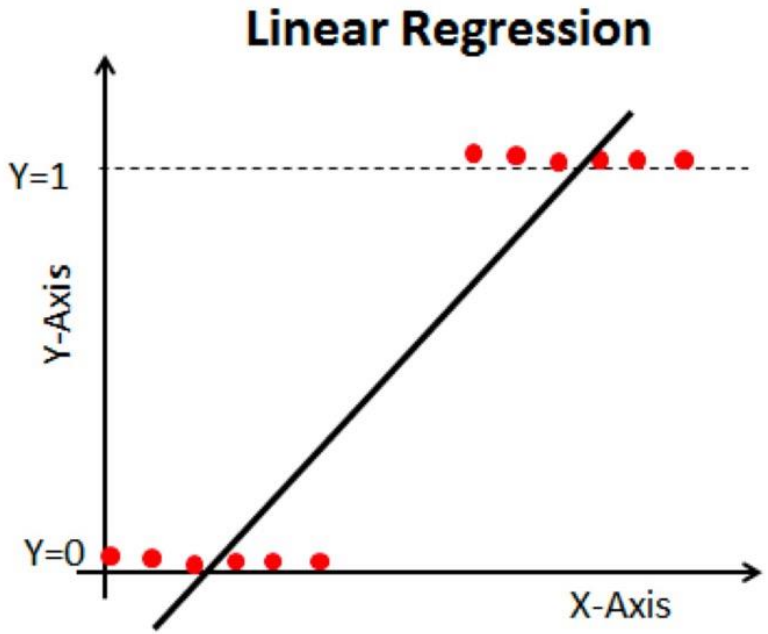
좌변을 오즈 형태로 만든 후 로그 취하기

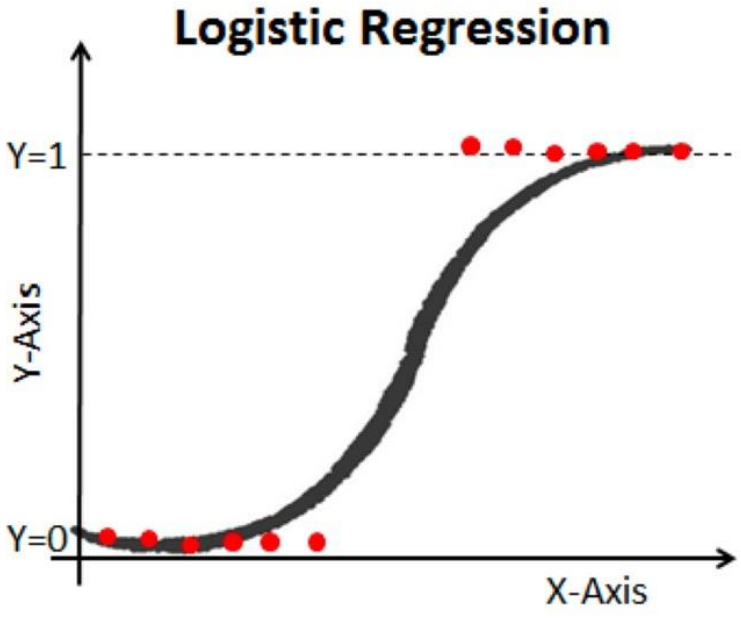
$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

좌변 범위 : $-\infty \sim \infty$ = 우변 범위 : $-\infty \sim \infty$

4 로지스틱 회귀모형

로지스틱 회귀모형의 장점





일반선형모델

Y의 범위가 0과 1을 초과

로지스틱 회귀모형

Y의 범위가 0과 1 사이

4 로지스틱 회귀모형

로지스틱 회귀모형의 장점

일반선형모델

Y의 범위가 0과 1을 초과

로지스틱 회귀모형

Y의 범위가 0과 1 사이

4 로지스틱 회귀모형

로지스틱 회귀모형의 장점

로지스틱 회귀모형의 장점

- ✓ 기본 가정의 완화
- ✓ 후향적 연구 분석에 활용 가능

독립성 가정만 **만족**하면 됨

일반선형회귀 모형의

정규성, 등분산성, 선형성 가정

만족할 필요 없음

오즈비를 사용하기 때문에

후향적 연구 분석에 활용 가능

로지스틱 회귀모형의 장점

로지스틱 회귀모형의 장점

✓ 기본 가정의 완화

✓ 후향적 연구 분석에 활용 가능

독립성 가정만 만족하면 됨

일반선형회귀 모형의
정규성, 등분산성, 선형성 가정
만족할 필요 없음

오즈비를 사용하기 때문에
후향적 연구 분석에 활용 가능

로지스틱 회귀모형의 기울기

로지스틱 회귀모형의 접선의 기울기

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

로지스틱 회귀 모형의 기울기는 모수 β의 영향을 받음

$\beta > 0$: **상향 곡선**

$\beta < 0$: **하향 곡선**

$|\beta|$가 **클수록** 기울기 변화율이 크므로 **가파른 형태**를 띰

로지스틱 회귀모형의 해석

확률을 통한 해석

로지스틱 회귀모형 변형해 확률에 관한 식으로 표현

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

$x_1, \ldots, x_p$를 대입하여 Y=1일 확률인 $\boldsymbol{\pi(x)}$를 알 수 있음

$\pi(x)$ 값이 cut-off point보다 크면 Y=1, 작으면 Y=0으로 예측

일반적으로 0.5 사용

로지스틱 회귀모형의 해석

오즈비를 통한 해석

로지스틱 회귀모형에 $x+1$과 x 대입 후 빼기

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

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

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

로지스틱 회귀모형의 해석

오즈비를 통한 해석

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

다른 설명변수가 **고정**되어 있을 때

x가 **한 단위 증가**할 때마다 **Y=1일 오즈가 e^{β}배 증가**

5

다범주 로짓모형

5 다범주 로짓모형

다범주 로짓 모형

다범주 로짓 모형 (Multicategory Logit Model)

3개 이상의 범주를 가진 반응변수로 확장시킨 모형

연결함수는 **로짓 연결 함수** 사용, 랜덤성분은 **다항분포** 따름

반응변수의 범주가 3개 이상으로 늘어났기 때문에
명목형 자료와 순서형 자료를 구분해야함

다범주 로짓 모형

기준 범주 로짓 모형
명목형 자료

누적 로짓 모형
순서형 자료

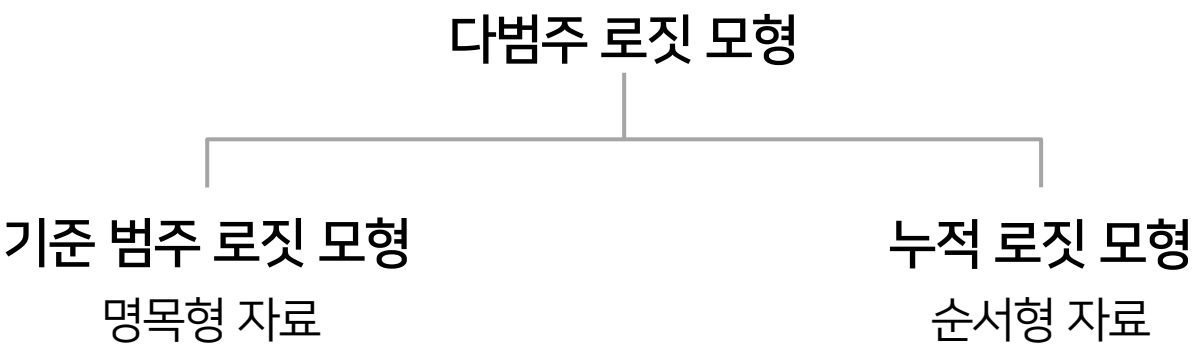
다범주 로짓 모형

다범주 로짓 모형 (Multicategory Logit Model)

3개 이상의 범주를 가진 반응변수로 확장시킨 모형

연결함수는 **로짓 연결 함수** 사용, 랜덤성분은 **다항분포** 따름

반응변수의 범주가 3개 이상으로 늘어났기 때문에
명목형 자료와 순서형 자료를 구분해야함



5 다범주 로짓모형

기준 범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

반응변수가 **명목형 자료**일 때 사용하는 **다범주 로짓 모형**

기준 범주와 나머지 범주를 짝지어 로짓을 정의

일반적으로 반응변수의 여러 개의 범주 중 **마지막 범주**

j : 범주에 대한 첨자

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

J : 기준 범주에 대한 첨자

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

5 다범주 로짓모형

기준 범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

반응변수가 **명목형 자료**일 때 사용하는 **다범주 로짓 모형**

기준 범주와 나머지 범주를 짝지어 로짓을 정의

일반적으로 반응변수의 여러 개의 범주 중 **마지막 범주**

j : 범주에 대한 첨자

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

J : 기준 범주에 대한 첨자

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

5 다범주 로짓모형

기준 범주 로짓 모형

기준 범주 로짓 모형 (Baseline-Category Logit Model)

반응변수가 **명목형 자료**일 때 사용하는 **다범주 로짓 모형**

기준 범주와 나머지 범주를 짝지어 로짓을 정의

일반적으로 반응변수의 여러 개의 범주 중 **마지막 범주**

기준범주와 그 외 J-1개의 범주를 각각 비교하기 때문에

총 **J-1개의 로짓 방정식 생성**

5 다범주 로짓모형

기준 범주 로짓 모형

j : 범주에 대한 첨자

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

J : 기준 범주에 대한 첨자

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

기준 범주 로짓 모형을 **확률**에 대한 식으로 재정의!

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

5 다범주 로짓모형

기준 범주 로짓 모형 예시

피셋 하입보이 지훈이의 뉴진스 최애 멤버를 기준 범주 로짓 모형으로 알아보자!

기준 범주를 민지로 정의

반응변수의 범주가 3개이기 때문에
총 2개의 기준 범주 로짓 모형 생성

$$\log\left(\frac{\pi_{\text{하니}}}{\pi_{\text{민지}}}\right) = 5 + 0.27x_1 + \cdots + 0.59x_k$$

$$\log\left(\frac{\pi_{\text{해린}}}{\pi_{\text{민지}}}\right) = 2 + 0.22x_1 + \cdots + 0.46x_k$$

같은 설명변수여도 회귀계수 β가 **다른 값**을 가지는 것을 알 수 있음

5 다범주 로짓모형

기준 범주 로짓 모형 예시

피셋 하입보이 지훈이의 뉴진스 최애 멤버를 기준 범주 로짓 모형으로 알아보자!

각 기준 범주 로짓 모형의 좌변을 **확률**에 대한 식으로 재정의

해린이 최애 멤버일 확률

$$\pi_{\text{해린}} = \frac{e^{\,2+0.22x_1+\cdots+0.46x_K}}{e^{\,2+0.22x_1+\cdots+0.46x_K} + e^{\,5+0.27x_1+\cdots+0.59x_K}}$$

5 다범주 로짓모형

기준 범주 로짓 모형의 해석

j범주와 J범주(기준범주) 비교

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

다른 설명변수들이 **고정**되어 있을 때

x_i가 **한 단위 증가**하면 기준범주 대신 **j 범주일 오즈가 e^{β_j}배 증가**

기준 범주 로짓 모형의 해석

a범주와 b범주 비교

$$\log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right)$$

$$= (\alpha_a + \beta_a^1 x_1 + \cdots + \beta_a^K x_K) - (\alpha_b + \beta_b^1 x_1 + \cdots + \beta_b^K x_K)$$

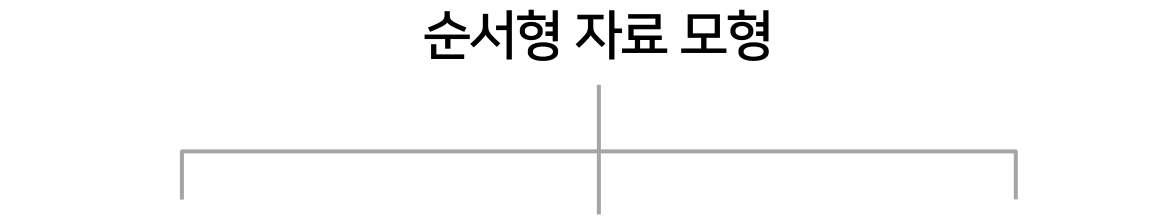
$$= [\alpha_a - \alpha_b] + \left[(\beta_a^1 - \beta_b^1)x_1 + \cdots + (\beta_a^K - \beta_b^K)x_K\right]$$

다른 설명변수들이 **고정**되어 있을 때

x_i가 **한 단위 증가**하면 b범주 대신 **a범주일 오즈가 $e^{\beta_a^i - \beta_b^i}$배 증가**

5 다범주 로짓모형

누적 로짓 모형



cut-point

순서형 반응변수의 범주들을 나누는 기준으로
각 행마다 색깔이 바뀌는 경계점이 cut point가 됨

소형	중형	대형	초대형
소형	중형	대형	초대형
소형	중형	대형	초대형

모든 범주를 기준범주로 사용해서 순서형 반응변수의 범주들을 나눔

5 다범주 로짓모형

누적 로짓 모형

순서형 자료 모형

연속비 로짓 모형 | 이웃 범주 로짓 모형 | 누적 로짓 모형

cut-point

순서형 반응변수의 범주들을 나누는 기준으로
각 행마다 색깔이 바뀌는 경계점이 cut point가 됨

소형	중형	대형	초대형
소형	중형	대형	초대형
소형	중형	대형	초대형

모든 범주를 **기준범주**로 사용해서 **순서형 반응변수**의 범주들을 나눔

5 다범주 로짓모형

누적 로짓 모형

반응변수가 총 J개의 범주를 가질 때

첫 번째 범주부터 j범주까지의 **누적확률**

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

누적 확률을 로그 오즈의 형태로 조작

$$\log\left(\frac{P(Y \leq j|X = x)}{1 - P(Y \leq j|X = x)}\right) = \log\left(\frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_J(x)}\right)$$

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

5 다범주 로짓모형

누적 로짓 모형

반응변수가 총 J개의 범주를 가질 때

첫 번째 범주부터 j범주까지의 **누적확률**

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

누적 확률을 **로그 오즈**의 형태로 조작

$$\log\left(\frac{P(Y \leq j|X = x)}{1 - P(Y \leq j|X = x)}\right) = \log\left(\frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_J(x)}\right)$$

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

5 다범주 로짓모형

누적 로짓 모형

반응변수가 총 J개의 범주를 가질 때

첫 번째 범주부터 j범주까지의 **누적확률**

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

누적 확률을 **로그 오즈**의 형태로 조작

최종적인 **누적 로짓 모형**의 형태

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

5 다범주 로짓모형

누적 로짓 모형과 기준 범주 로짓 모형 비교

누적 로짓 모형

$$logit[P(Y \leq j|X = x)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$$

기준 범주 로짓

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

공통점

기준점을 설정하여 비교하는 원리

→ 총 J-1개의 로짓 방정식 생성

차이점

기준 범주 로짓 모형

기준범주에 따라 β값이 상이

누적 로짓 모형

기준범주에 상관없이 β값이 동일

5 다범주 로짓모형

누적 로짓 모형과 기준 범주 로짓 모형 비교

누적 로짓 모형

$$logit[P(Y \leq j|X = x)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$$

기준 범주 로짓

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

누적 로짓 모형의 β값이 동일한 이유

비례 오즈 가정 때문!

차이점

기준 범주 로짓 모형

기준범주에 따라 β값이 상이

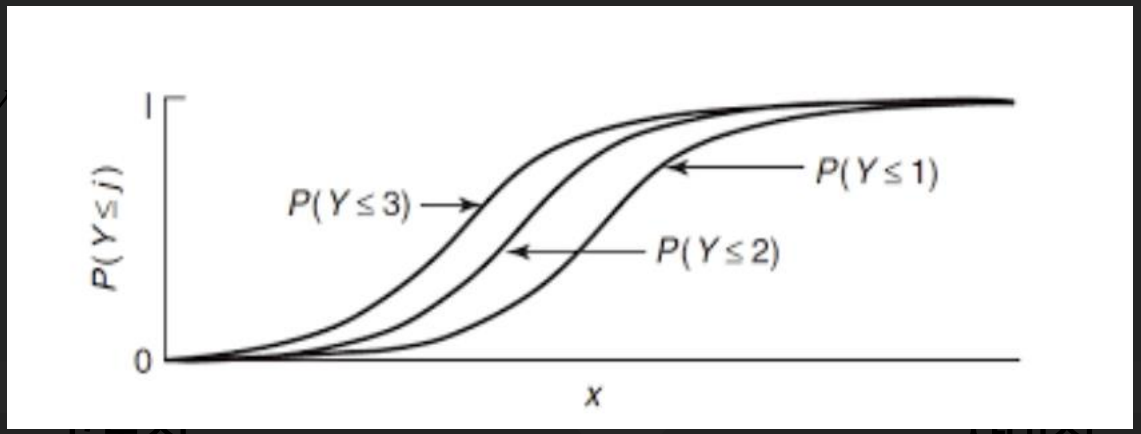
누적 로짓 모형

기준범주에 상관없이 β값이 동일

5 다범주 로짓모형

누적 로짓 모형과 기준 범주 로짓 모형 비교

비례 오즈 가정



서로 다른 로짓 방정식이 같은 기울기의 곡선 형태이지만
수평 이동을 한 것처럼 나타남
→ α는 다르지만 β는 같음

5 다범주 로짓모형

누적 로짓 모형 예시

범주팀 팀장에 대한 불만 지표

적음 / 보통 / 많음 / 매우 많음

$$logit[P(\leq \text{적음})] = 8 + 0.07x_1 + \cdots + 0.6x_p$$

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

$$logit[P(\leq \text{많음})] = 12 + 0.07x_1 + \cdots + 0.6x_p$$

기준 범주와 상관없이 **일관된 β값**을 가지는 것을 확인!

5 다범주 로짓모형

누적 로짓 모형 해석

오즈를 이용한 해석

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

다른 설명변수들이 **고정**되어 있을 때

x_i가 **한 단위 증가**하면 $Y > j$에 비해 $\boldsymbol{Y \leq j}$**일 오즈가** e^{β_i}**배 증가**

6

포아송 회귀 모형

포아송 회귀 모형

포아송 분포의 **평균이 작은** 경우 **분포**가 **편향**됨

정규성과 등분산성 가정 불충족

일반선형모형 적용 불가능

포아송 회귀 모형으로 해결 가능

포아송 회귀 모형

포아송 분포의 **평균이 작은** 경우 **분포**가 **편향**됨

정규성과 등분산성 가정 불충족

일반선형모형 적용 불가능

포아송 회귀 모형으로 해결 가능

포아송 회귀 모형

포아송 회귀 모형 (Poisson Regression Model)

반응변수가 **도수 자료**처럼 **포아송 분포**를 따를 때 사용

반응변수가 **도수 자료**이고

랜덤성분이 **포아송 분포**를 따르며 **로그 연결함수**를 사용

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

6 포아송 회귀 모형

포아송 회귀 모형

포아송 회귀 모형 (Poisson Regression Model)

반응변수가 도수 자료일 때 로그 연결함수를 사용하는 포아송 분포를 따를 때 사용

로그 연결함수를 사용하는 이유

포아송 분포의 랜덤 성분의 범위는 0과 ∞ 사이지만

체계적 성분의 범위는 - ∞과 ∞ 사이임

반응변수가 도수 자료이고

양 변의 범위를 동일하게 조정하기 위해 로그 연결함수 이용!

랜덤성분이 포아송 분포를 따르며 로그 연결함수를 사용

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

포아송 회귀 모형 해석

도수를 이용한 해석

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

포아송 회귀 모형을 도수(=μ)에 관한 식으로 변형

다른 설명변수들이 **고정**되어 있을 때

x_i가 **한 단위 증가**하면 μ가 e^{β_i}**배 증가**

포아송 회귀 모형 해석

오즈비를 이용한 해석

포아송 회귀 모형에 $x+1$과 x 대입 후 빼기

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

다른 설명변수들이 고정되어 있을 때

x가 한 단위 증가하면 μ가 e^{β}배 증가

6 포아송 회귀 모형

포아송 회귀 모형의 한계

설명변수들이 시간, 공간 등의 **요소의 차이**를 **반영하지 못함**

μ에 대한 예측값인 **기대도수만 산출** 가능

비율자료를 활용한 **율자료 포아송 회귀 모형** 사용!

6 포아송 회귀 모형

포아송 회귀 모형의 한계

설명변수들이 시간, 공간 등의 **요소의 차이**를 **반영하지 못함**

μ에 대한 예측값인 **기대도수만 산출** 가능

비율자료를 활용한 율자료 포아송 회귀 모형 사용!

6 포아송 회귀 모형

율자료 포아송 회귀 모형

율자료 포아송 회귀 모형

기존의 μ값 대신 **비율자료**를 반응변수로 사용

기존 포아송 회귀 모형과 같이 **로그 연결함수**를 사용

μ 대신 $\frac{\mu}{t}$ 사용

$$\log(\frac{\mu}{t}) = \log(\mu) - \log(t) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

t: 기준이 되는 지표값

6 포아송 회귀 모형

율자료 포아송 회귀 모형 해석

오즈비를 이용한 해석

포아송 회귀 모형에 $x+1$과 x 대입 후 빼기

$$\log(\mu(x+1)/t) - \log(\mu(x)/t)$$

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

다른 설명변수들이 고정되어 있을 때

x가 한 단위 증가하면 기대비율이 e^{β}배 증가

6 포아송 회귀 모형

포아송 회귀 모형의 문제점

현실에서는 포아송 분포를 따르는 **도수 자료**가

등산포 가정을 **만족하지 않는 경우** 발생

분산이 평균보다 큰 값을 가져 등산포 가정이 어긋나는 **과대산포** 문제 발생

과대산포 발생 시 **분산**이 **과소평가**되어 **검정 결과**가 **왜곡**

음이항 회귀 모형을 이용해 해결

6 포아송 회귀 모형

포아송 회귀 모형의 문제점

현실에서는 포아송 분포를 따르는 **도수 자료**가

등산포 가정을 **만족하지 않는 경우** 발생

분산이 평균보다 큰 값을 가져 등산포 가정이 어긋나는 **과대산포** 문제 발생

과대산포 발생 시 **분산**이 **과소평가**되어 **검정 결과**가 **왜곡**

음이항 회귀 모형을 이용해 해결

6 포아송 회귀 모형

음이항 회귀 모형

음이항 회귀 모형 (Negative Binomial Regression)

랜덤성분이 **음이항 분포**를 따르고 **로그 연결함수**를 사용

평균과 분산 간의 비선형성을 가정한 2차 함수 형태

$$E(Y) = \mu$$

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

평균과 분산의 차이를 발생시키는 산포모수

음이항 분포는 평균보다 큰 분산 값을 갖기 때문에

포아송 분포의 등산포 가정을 완화할 수 있는 성질을 이용해 과대산포 문제 해결

6 포아송 회귀 모형

음이항 회귀 모형

음이항 회귀 모형 (Negative Binomial Regression)

랜덤성분이 **음이항 분포**를 따르고 **로그 연결함수**를 사용

평균과 분산 간의 비선형성을 가정한 2차 함수 형태

$$E(Y) = \mu$$

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

평균과 분산의 차이를 발생시키는 산포모수

음이항 분포는 **평균보다 큰 분산** 값을 갖기 때문에

포아송 분포의 **등산포 가정**을 **완화**할 수 있는 성질을 이용해 **과대산포 문제 해결**

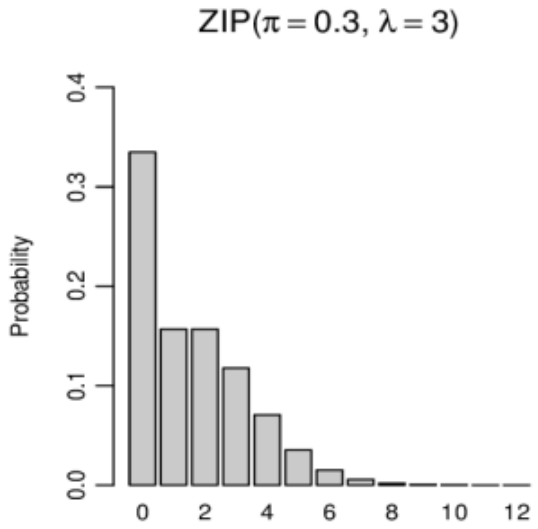
6 포아송 회귀 모형

포아송 회귀 모형의 문제점

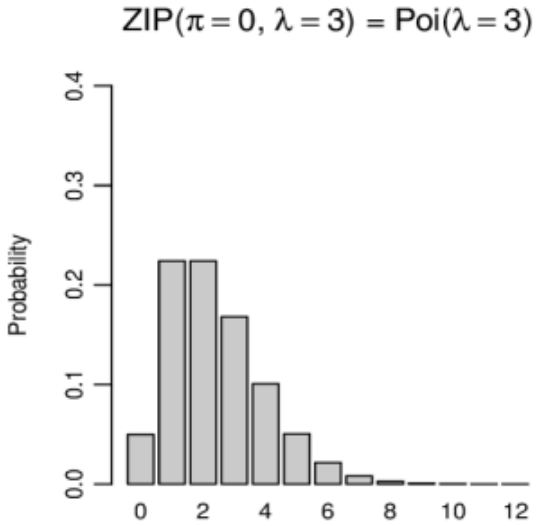
과대영 문제 (Excess Zeros)

포아송 분포에서 예상된 0 발생 횟수보다

실제로 더 많은 0이 발생한 경우



과대영 문제가 발생한 경우



과대영 문제가 발생하지 않은 경우

6 포아송 회귀 모형

포아송 회귀 모형의 문제점

과대영 문제 (Excess Zeros)

포아송 분포에서 예상된 0 발생 횟수보다

실제로 더 많은 0이 발생한 경우

과대영 문제는 영과잉 포아송 회귀 모형이나

영과잉 음이항 회귀 모형을 통해 해결 가능

6 포아송 회귀 모형

영과잉 포아송 회귀모형

영과잉 포아송 분포

$$Y = f(x) = \begin{cases} 0, & with\ probability\ \phi_i \\ g(y_i), & with\ probability\ 1-\phi_i \end{cases}$$

- Y: 베르누이분포를 따름
- $g(y_i)$: 0 이상의 정수 값들이 따르는 포아송 분포
- ϕ_i: 0이 발생할 확률
- $1-\phi_i$: 0 이상의 정수 값이 발생 확률

0만 발생하는 **점확률분포**와

0 이상의 정수 값이 발생하는 **포아송분포**의 **혼합분포 구조**

영과잉 부분과 0이 아닌 부분으로 이분화한 형태

영과잉 포아송 회귀모형

영과잉 포아송 회귀모형 (ZIP)

영과잉 포아송 분포를 GLM으로 표현

총 2가지 식이 생성됨

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

0값이 발생할 확률(ϕ_i)을 로짓 연결함수를 이용하여 표현

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

포아송분포의 평균(λ)을 로그 연결함수를 이용하여 표현

다음주 예고

1. 혼동행렬
2. ROC 곡선
3. 샘플링
4. 인코딩

감사합니다