범주형자료분석팀

2팀
김보현
송승현
최용원
김동희
오주원

INDEX

1

GLM

GLM이란?

GLM(일반화 선형 모형)

연속형 반응변수에 대한 모형뿐만 아니라

다양한 형태의 반응변수에 대한 모형들을 포함하는 **광범위한 모형들의 집합**

선형회귀분석, ANOVA 등 모두 포함!

선형회귀모형의 한계를 극복해

범주형 · 도수형 반응변수가 주어진 경우에도

설명변수와 반응변수 간의 최적의 관계식 추정 가능

GLM의 필요성

선형 회귀의 가정 위배

범주형 · 도수형 반응변수는 자료의 오차항이 **정규분포**를 따르지 않아
선형회귀의 LSE 사용 불가

다양한 확률분포 사용

GLM은 LSE 대신 MLE를 사용해 모형을 적합
정규분포 이외의 반응변수에 대한 분석 가능

GLM의 필요성

VS 분할표 분석

분할표는 **범주형** 변수들 간의 연관성만을 파악

변수 간의 연관성 파악

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

새로운 설명변수에 대한 **반응변수 예측** 가능

GLM의 구성성분

GLM의 형태

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

구성성분

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

GLM의 구성성분

랜덤성분

가정한 확률분포 하에서 서로 **독립**인 반응변수 Y의 **기댓값**

반응변수로 **지수족**에 해당하는 분포만을 사용 가능

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

GLM의 구성성분

체계적 성분

설명변수 X를 명시하는 성분

X의 **선형결합**으로 표현 가능

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

교호작용을 설명하는 항이나 **곡선효과**를 나타내는 항 포함 가능

$x_i = x_a x_b$ $x_i = x_a^2$

교호작용 : 한 요인의 효과가 다른 요인의 수준에 의존하는 현상

GLM의 구성성분

연결 함수 (Link Function)

랜덤 성분과 **체계적 성분**을 연결하여 두 성분의 범위를 **일치**

랜덤성분이 **이항분포**를 따르고 **설명변수**가 **연속형**이라면,

체계적 성분의 범위는 제약이 없어 $-\infty \sim \infty$의 범위를 따르지만

랜덤 성분은 분포에 따라 범위에 제약이 발생해 둘의 범위가 **불일치**

GLM의 구성성분

연결 함수 (Link Function)

랜덤 성분과 체계적 성분을 연결하여 두 성분의 범위를 일치

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

GLM의 특징

오차항의 다양한 분포 가정 가능

반응변수의 오차항이 가진 성질에 따라
정규분포 외에도 **어떤 분포든 정의 가능**

선형 관계식 유지

회귀 계수 β의 선형 관계식 형태로 해석 용이
교호작용이나 곡선효과 포함하더라도 선형성 유지

GLM의 특징

독립성 가정만 필요

회귀팀 2주차 클린업 참고!

선형회귀모형과 달리 **오차항에 대한 독립성**만 만족하면 됨

반응변수 간의 자기상관성 검정

→ 더빈 왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 통해 랜덤성분과 체계적성분 간의 범위 교정

제한된 범위를 가진 반응변수 사용 가능

→ 범주형 자료, 도수자료

1 GLM

GLM의 특징

독립성 가정만 필요

회귀팀 2주차 클린업 참고!

선형회귀모형과 달리 **오차항에 대한 독립성**만 만족하면 됨

반응변수 간의 자기상관성 검정

더빈 왓슨 검정

제한적인 범위의 반응변수 사용 가능

연결함수를 통해 랜덤성분과 체계적성분 간의 범위 교정

제한된 범위를 가진 반응변수 사용 가능

범주형 자료, 도수자료

정리해보자!

GLM은 랜덤성분의 분포와 연결함수를 일반화한 것으로,
범주형 변수를 다룰 수 있음

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	혼합형
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		

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	혼합형
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		

GLM의 모형 적합

GLM은 회귀의 기본 4가지 가정 만족하지 못하므로
최소제곱법 (Least Square Estimation, LSE) 사용 불가능

최대가능도추정법 (Maximum Likelihood Estimation, MLE)

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

가능도(likelihood) : 고정된 관측값이 어떤 확률분포를 따를 가능성

GLM의 모형 적합

최대가능도추정량 (Maximum Likelihood Estimator)

최대가능도추정법으로 찾은 **가능도함수가 최대**가 되도록 하는 추정량 $\hat{\lambda}$

$X_1, \dots, X_n$이 모수가 λ인 지수분포를 따른다고 가정했을 때…

$$L(\lambda\,; x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \xrightarrow{\;log\;} \ln L(\lambda) = l(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$l(\lambda)$을 편미분해 값이 0이 되도록 하는 값

$$\hat{\lambda} = \frac{n}{\sum_{k=1}^{n} X_k}$$

2

유의성 검정

유의성 검정

분석 가치가 없는 GLM 모형

유의성 검정 가설

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

H_1 : 적어도 하나의 β는 0이 아니다.

유의성 검정

Significance Testing

- **왈드 검정** — Wald Test
- **스코어 검정** — Score Test
- **가능도비 검정** — Likelihood-ratio Test

2 유의성 검정

왈드 검정 (Wald Test)

검정통계량

$$Z = \frac{\hat{\beta}}{S.E.} \sim N(0,1) \text{ 또는 } Z^2 = \left(\frac{\hat{\beta}}{S.E.}\right)^2 \sim \chi_1^2$$

기각역

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

회귀계수에 대한 **추정값과 표준오차**만 사용하여 통계량을 구함

2 유의성 검정

왈드 검정 (Wald Test)

범주형 자료이거나 소표본인 경우

검정통계량 $Z = \frac{\hat{\beta}}{S.E.} \sim N(0,1)$ 또는 $Z^2 = \left(\frac{\hat{\beta}}{S.E.}\right)^2$ 검정력 감소

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

가능도비 검정 사용하여

회귀계수에 대한 추정값과 표준오차만 사용하여 통[illegible]함

유의성 검정

2 유의성 검정

가능도비 검정 (Likelihood-ratio Test)

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 하에서 가능도함수 (l_0)와

전체공간 하에서의 가능도함수 (l_1)의 **차이** 이용

귀무가설 + 대립가설

2 유의성 검정

가능도비 검정 (Likelihood-ratio Test)

$$G^2 = -2\log\left(\frac{\text{모수가 귀무가설 } H_0\text{를 만족할 때 } (\beta=0) \text{ 가능도 함수의 최댓값}}{\text{모수가 아무런 제약이 없을 때 가능도 함수의 최댓값}}\right)$$

MLE로 계산

L_0 , L_1 두 가능도 함수의 최댓값 비교

검정 과정

l_0 와 l_1 의 차이가 큼 ⟶ 검정통계량의 값이 큼 ⟶ p-value 값이 작음

⟶ **귀무가설 기각** (적어도 하나의 β는 0이 아님) ⟶ **모형의 모수 추정값 유의함**

가능도비 검정 (Likelihood-ratio Test)

귀무가설 하에서의 가능도 함수와 전체공간 하에서의 가능도 함수 모두 사용하여

타 검정에 비해 많은 정보 이용

왈드 검정에 비해 **좋은 검정력과 높은 신뢰도**

2 유의성 검정

이탈도

이탈도 (Deviance)

관심모형과 포화모형의 비교를 통해 **모형의 적합성** 판단하기 위한 가능도비 통계량

관심 모형 (M) 유의성을 검정하고자 하는 모형

$$\text{범주팀의 행복 } (Y) = \beta_0 + \beta_1 \times \text{클린업 시간 } (x_1) + \beta_2 \times \text{패키지 난이도 } (x_2)$$

포화 모형 (S) 모든 관측값에 대해 모수를 갖는 가장 복잡한 모형

$$\text{범주팀의 행복 } (Y) = \beta_0 + \beta_1 \times \text{클린업 시간 } (x_1) + \beta_2 \times \text{패키지 난이도 } (x_2) +$$
$$\beta_3 \times \text{교안 페이지 수 } (x_3) + \beta_4 \times \text{범주팀원들의 귀여움 } (x_4)$$

2 유의성 검정

이탈도

검정통계량

$$G^2 = -2\log\left(\frac{l_m}{l_s}\right) = -2(L_M - L_S) \sim \chi^2_{df}$$

기각역

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

H_0: 관심모형에 속하지 않는 모수는 모두 0이다. → 관심 모형 사용

H_1 : 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다. → 관심 모형 사용 불가

2 유의성 검정

이탈도

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

두 모형의 가능도 함수의 최댓값 차이 비교

검정 과정

두 가능도 함수의 최댓값 차이가 큼 ⟶ 이탈도가 큼 ⟶ p-value 값이 작음

⟶ **귀무가설 기각** ⟶ 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아님

⟶ **관심 모형 M이 적합하지 않으므로 사용 불가능**

2 유의성 검정

이탈도

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

이탈도는

포화모형에는 있지만 관심모형에는 없는 계수들이 0인지의 여부를 확인하므로

관심모형은 포화모형에 내포된 (nested) 관계여야함!

두 모형의 가능도 함수의 최댓값 차이 비교

검정 과정

관심모형(M)의 모수 ⊂ 포화모형(S)의 모수

두 가능도 함수의 최댓값 차이가 큼 ⟶ 이탈도가 큼 ⟶ p-value 값이 작음

⟶ **귀무가설 기각** ⟶ 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아님

⟶ **관심 모형 M이 적합하지 않으므로 사용 불가능**

이탈도와 가능도비 검정의 관계

단순한 형태의 관심모형 복잡한 형태의 관심모형

M_0의 이탈도 - M_1의 이탈도 = $-2(L_0 - L_s) - \left(-2(L_1 - L_s)\right)$

$= -2(L_0 - L_1)$

두 모형 간의 이탈도 값의 차이 = 가능도비 검정 통계량

모형의 적합도 비교 시 관심모형 간의 이탈도 차이 사용

이탈도와 가능도비 검정의 관계

단순한 형태의 관심모형 복잡한 형태의 관심모형

M_0의 이탈도 - M_1의 이탈도 = $-2(L_0 - L_s) - \left(-2(L_1 - L_s)\right)$

$= -2(L_0 - L_1)$

검정 과정

관심모형(M_0, M_1) 간 이탈도 차이가 작음 ⟶ 가능도비 검정 통계량이 작음

⟶ p-value 값이 큼 ⟶ **귀무가설 기각하지 못함**

⟶ M_0에 포함되지 않는 모수들은 모두 0 ⟶ **간단한 관심모형인 M_0이 더 적합**

2 유의성 검정

이탈도와 가능도비 검정의 관계

단순한 형태의 관심모형 M_0 복잡한 형태의 관심모형 M_1

이탈도 활용 시

모형 M_0은 모형 M_1에 내포된 (nested) 모형이어야 함

$= -2(L_0 - L_1)$

내포된 경우가 아니라면,

AIC, BIC와 같은 모형 선택 (Variable Selection) 측도 활용하여 모형 비교

관심모형(M_0, M_1) 간 이탈도 차이가 작음 ⟶ 가능도비 검정 통계량이 작음

⟶ p-value 값이 큼 ⟶ 귀무가설 기각하지 못함

회귀팀 3주차 클린업 참고!

⟶ M_0에 포함되지 않는 모수들은 모두 0 ⟶ 간단한 관심모형인 M_0이 더 적합

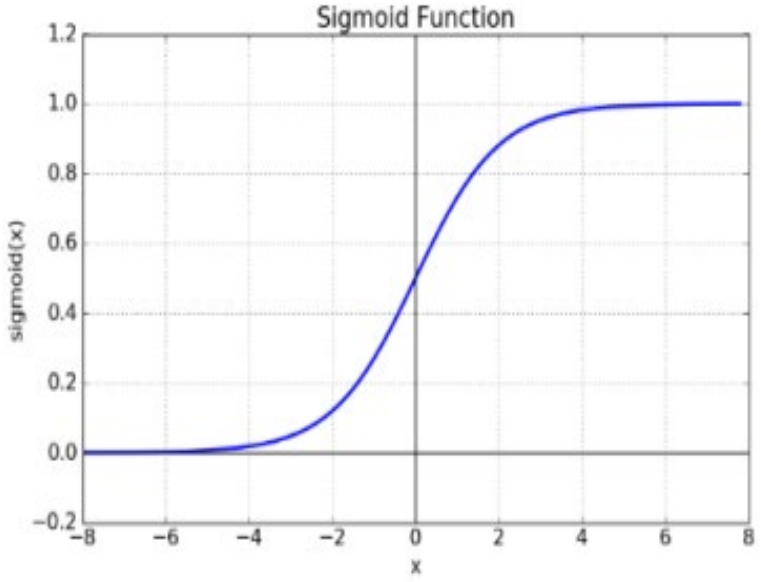
3

로지스틱 회귀모형

로지스틱 회귀모형

로지스틱 회귀모형(Logistic Regression)

반응변수 Y가 이항자료일 때의 회귀모형



확률을 따르는 S자 곡선 형태

$\pi(x)$와 x 의 비선형 관계를 나타냄

시그모이드(Sigmoid) 형태의 함수

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

이항 변수와 **연속형 변수** 간의 범위 일치

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

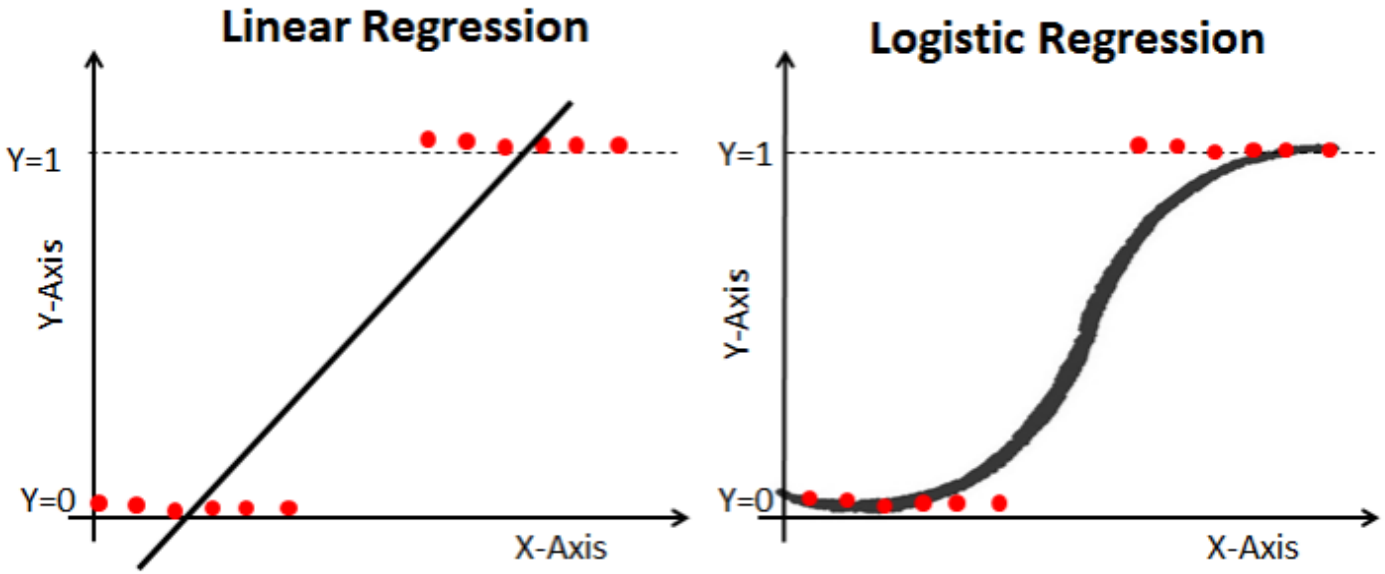
좌변 범위: 0~1 ≠ 우변 범위: $-\infty \sim \infty$

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



일반선형모형

Y의 범위가 0과 1을 초과

로지스틱 회귀모형

Y의 범위가 0과 1 사이

3 로지스틱 회귀모형

로지스틱 회귀모형의 장점

로지스틱 회귀모형은 **이항 랜덤성분**을

로짓 연결함수를 통해 체계적 성분과 연결한 **GLM**의 일종

연속형 변수를 통해 이항분포를 따르는 **반응변수**를 해석 가능

일반선형모형	로지스틱 회귀모형
Y의 범위가 0과 1을 초과	**Y의 범위가 0과 1 사이**

로지스틱 회귀모형의 장점

후향적 연구에서도 사용 가능

기본가정의 완화

모수들이 오즈비와 관련되어 있어

후향적 연구에도 사용 가능

일반 회귀 모형이 충족 시켜야 하는

4가지 가정(정규성,등분산성,선형,독립성)

중 오직 **독립성** 가정만 만족하면 됨

로지스틱 회귀모형의 장점

후향적 연구에서도 사용 가능

기본가정의 완화

모수들이 오즈비와 관련되어 있어

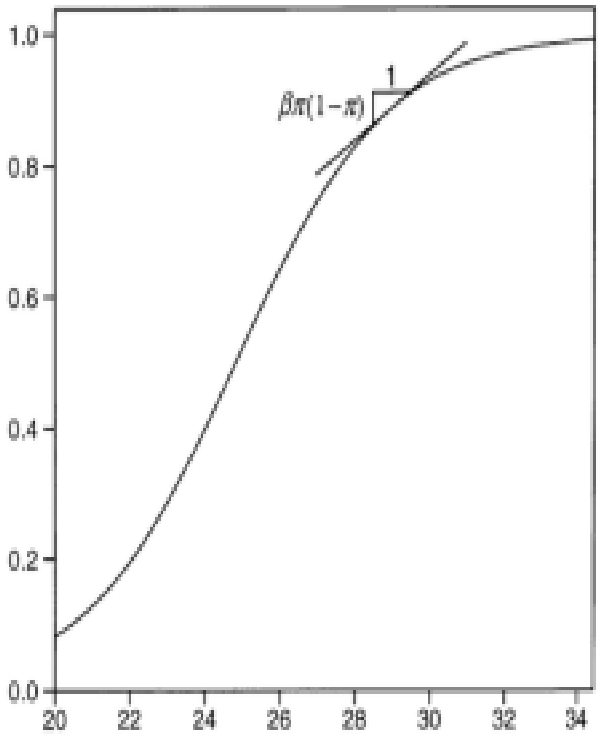
후향적 연구에도 사용 가능

일반 회귀 모형이 충족 시켜야 하는

4가지 가정(정규성,등분산성,선형,독립성)

중 오직 **독립성** 가정만 만족하면 됨

로지스틱 회귀모형의 기울기



로지스틱 회귀모형의 접선의 기울기

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

기울기는 모수 β의 영향을 받음

β가 **양수**라면 **상향곡선**의 형태

β가 **음수**라면 **하향곡선**의 형태

β의 절댓값이 클수록 가파른 형태

로지스틱 회귀모형의 해석

확률을 통한 해석

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

Y가 1일 성공의 확률인 $\boldsymbol{\pi(x)}$ 는 특정 x를 대입해 구할 수 있음

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

일반적으로 0.5 사용

로지스틱 회귀모형의 해석

오즈비를 통한 해석

로지스틱 회귀모형에서 $x+1$과 x대입 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 증가할 때마다 **Y=1일 오즈가 e^{β}배 증가**

4

다범주 로짓모형

다범주 로짓모형

다범주 로짓모형 (Multicategory Logit Model)

범주가 3개 이상이므로, 주어진 자료가 명목형 자료인지, 순서형 자료인지 구분

3개 이상의 범주를 가진 반응변수로 확장시킨 모형

연결함수는 **로짓 연결함수** 사용, 랜덤성분은 **다항분포** 따름

명목형 자료 → 기준 범주 로짓모형

순서형 자료 → 누적 로짓모형

4 다범주 로짓모형

기준 범주 로짓모형

기준 범주 로짓모형 (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

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

$$= \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^k x_k \,, j = 1, \ldots, \mathrm{J}-1$$

기준 범주 로짓모형

기준 범주 로짓모형 (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

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

$$= \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^k x_k \,, j = 1, \ldots, \mathrm{J}-1$$

j : 범주에 대한 첨자

J : 기준 범주에 대한 첨자

1 ~ k: 설명변수에 대한 첨자

기준 범주 로짓모형

기준 범주 로짓모형 (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

기준 범주 J와 그 외 범주들을 각각 짝 지어 로짓을 정의

그 결과 $J-1$ 개의 **로짓 방정식이 생성**

기준 범주 로짓모형

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

$$= \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^k x_k \,, j = 1, \ldots, \mathrm{J}-1 \;\text{ with } \sum_{j=1}^{J} P(Y=j|X=x) = 1$$

기준 범주 로짓모형의 **확률**에 대한 식

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

4 다범주 로짓모형

기준 범주 로짓모형 예시

반응변수를 해리포터 3인방 중 가장 좋아하는 마법사로!

기준 범주를 해리포터로 지정

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

$$\log\left(\frac{\pi_{론}}{\pi_{해리포터}}\right) = 5 + 0.27x_1 + \ldots + 0.59x_k$$

$$\log\left(\frac{\pi_{헤르미온느}}{\pi_{해리포터}}\right) = 2 + 0.22x_1 + \ldots + 0.46x_k$$

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

4 다범주 로짓모형

기준 범주 로짓모형 예시

반응변수를 해리포터 3인방 중 가장 좋아하는 마법사로!

각 기준 범주 로짓 모형의 좌변을 **확률**에 대한 식으로 재정의

가장 좋아하는 마법사가 헤르미온느일 확률

$$\pi_{\text{헤르미온느}} = \frac{e^{2+0.22x1+\cdots+0.46xk}}{e^{2+0.22x1+\cdots+0.46xk} + e^{5+0.27x1+\cdots+0.59xk}}$$

기준 범주 로짓모형의 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 증가하면 기준범주 대신 **j범주일 오즈가 e^{β_j}배 증가**

기준 범주 로짓모형의 해석

a범주와 b범주 비교

$$\log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) = \left(\alpha_a + \beta_a^1 x_1 + \ \ldots + \beta_a^k x_k\right) - \left(\alpha_b + \beta_b^1 x_1 + \ \ldots + \beta_b^k x_k\right)$$

$$= [\alpha_a - \alpha_b] + [(\beta_a^1 - \beta_b^1)x_1 + \cdots + (\beta_a^k - \beta_b^k)x_k]$$

다른 설명변수들이 고정되어 있을 때

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

주의 : a, b 범주는 기준범주가 아님!!

누적 로짓모형

순서형	이웃 범주 로짓모형 (Adjacent-Categories Model)
	연속비 로짓모형 (Continuation-ratio Logit Model)
	누적 로짓모형 (Cumulative Logit Model)

반응변수가 **순서형** 자료라면, 명목형 자료와는 달리 '순서'를 고려

4 다범주 로짓모형

누적 로짓모형

cut-point

순서형 반응변수의 범주들을 나누는 기준으로
각 행마다 색깔이 바뀌는 경계점이 cut-point

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

전체 범주를 모두 사용하여 검정력이 좋고 해석하기 용이

누적 로짓 모형

누적 로짓 모형

누적 확률에 **로짓 연결함수**를 사용한 형태

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

누적 확률

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

: 각각의 j에 대해 반응변수가 j 혹은 그 이하의 범주에 속할 확률

: 첫 번째 범주부터 j번째 범주까지의 누적 확률

누적 로짓 모형

누적 로짓 모형

누적 확률에 **로짓 연결함수**를 사용한 형태

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

누적 확률을 **로그 오즈**의 형태로 변환

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

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

누적 로짓 모형

누적 로짓 모형

누적 확률에 **로짓 연결함수**를 사용한 형태

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

누적 로짓 모형의 최종 형태

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

기준 범주 로짓 모형 vs 누적 로짓 모형

	기준 범주 로짓 모형	누적 로짓 모형
공통점	기준점을 두고 두 범위의 확률을 비교하는 방식 → J-1개의 로짓 방정식으로 구성	
차이점	$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$ → α와 회귀계수에 모두 첨자 j 존재	$\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$ → **회귀계수에서는 첨자 j 사라짐**

기준 범주 로짓 모형 vs 누적 로짓 모형

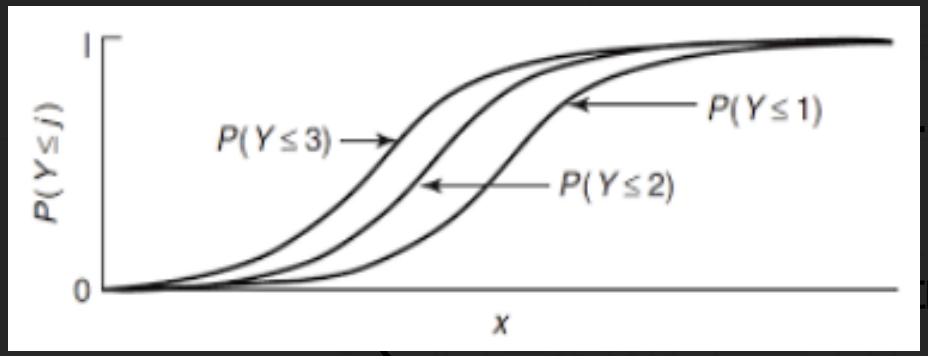
	기준 범주 로짓 모형	누적 로짓 모형
공통점	기준점을 두고 두 범위의 확률을 비교하는 방식 → J-1개의 로짓 방정식으로 구성	
차이점	$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$ → α와 회귀계수에 모두 첨자 j 존재	$\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$ → 회귀계수에서는 첨자 j 사라짐

J-1개의 로짓 방정식에 대한 β의 효과가
모두 동일하다고 가정하기 때문 (비례오즈 가정)

4 다범주 로짓 모형

기준 범주 로짓 모형 vs 누적 로짓 모형

비례오즈 가정



대한 β의 효과가

때문 (비례오즈 가정)

J-1개의 로짓 방정식은 절편인 α**값만 변할 뿐**

기울기에 해당하는 β**는 모두 같은 값**을 가짐 $x_1 + \cdots + \beta_p x_p$

→ 각 곡선은 **같은 모형**을 가져 **수평 이동**을 한 것처럼 나타남

점자 j 사라짐

4 다범주 로짓모형

누적 로짓 모형 예시

범주팀 클린업에 대한 불만도

낮음 / 보통 / 높음 / 아주 높음

$$logit[P(Y \leq \text{낮음})] = 5 + 0.05x_1 + \cdots + 0.6x_p$$

$$logit[P(Y \leq \text{보통})] = 8 + 0.05x_1 + \cdots + 0.6x_p$$

$$logit[P(Y \leq \text{높음})] = 12 + 0.05x_1 + \cdots + 0.6x_p$$

→ 모두 **β는 같지만** α값은 다르다는 것을 확인 가능!

누적 로짓 모형 해석

기준 범주 로짓 모형처럼 오즈를 이용하여 해석

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

다른 설명변수가 **고정**되어 있다는 가정 하에

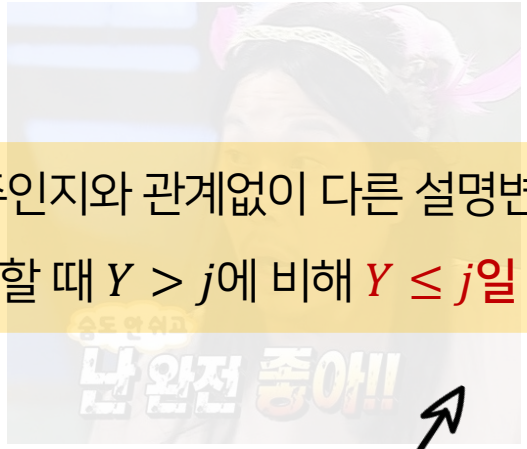
x_i**가 한 단위 증가**할 때 $Y > j$ 에 비해 $Y \le j$ **일 오즈가** e^{β_j}**만큼 증가**

4 다범주 로짓모형

누적 로짓 모형 해석

범주팀 클린업에 대한 불만도

낮음 / 보통 / 높음 / 아주 높음



어떤 범주인지와 관계없이 다른 설명변수가 고정되어 있을 때

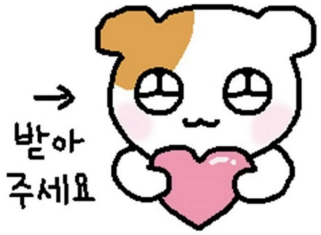
x_1이 한 단위 증가할 때 $Y > j$에 비해 $Y \leq j$**일 오즈가** $e^{0.05}$ = 약 1.05배 **증가**

$$logit[P(Y \leq \text{낮음})] = 5 + 0.05x_1 + \cdots + 0.6x_p$$

$$logit[P(Y \leq \text{보통})] = 8 + 0.05x_1 + \cdots + 0.6x_p$$

$$logit[P(Y \leq \text{높음})] = 12 + 0.05x_1 + \cdots + 0.6x_p$$

→ 모두 β**는 같지만** α값은 다르다는 것을 확인 가능!



5

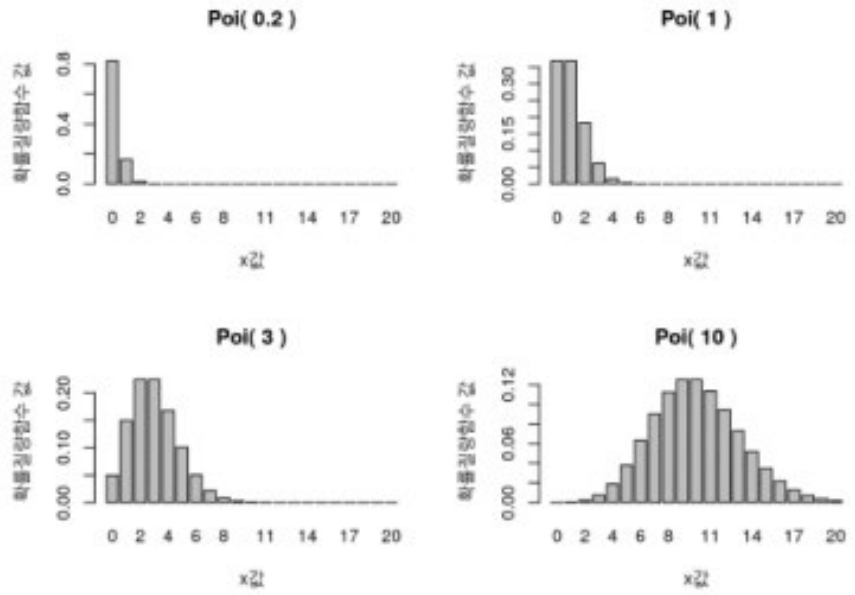
포아송 회귀 모형

포아송 회귀 모형

포아송 분포

단위 시간 안에 발생한 사건의 건수, 횟수를 표현하는 이산확률 분포

모수 λ, 평균이 작을수록 편향된 분포



포아송 회귀 모형

포아송 분포

단위 시간 안에 발생한 사건의 건수, 횟수를 표현하는 이산확률 분포

모수 λ, 평균이 작을수록 편향된 분포

정규성과 등분산성 가정 위배 & **표준오차나 유의성 수준 편향**되는 문제

포아송 회귀 모형으로 해결 가능!

포아송 회귀 모형

포아송 회귀 모형

반응변수가 **도수 자료**일 때 사용하는 모형

랜덤성분이 **포아송 분포**를 따름

양 변의 범위를 맞춰주기 위해 **로그 연결함수** 사용

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

도수 자료 (μ)	체계적 성분
음이 아닌 임의의 정수 값 0 ~ ∞ 사이의 값	-∞ ~ ∞ 사이의 값

포아송 회귀 모형 해석

① 도수를 통한 해석

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

↓ μ에 관한 식으로 정리

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

추정된 회귀 계수를 대입하면

μ(도수)에 대한 **예측값(=기대도수)** 도출 가능!

포아송 회귀 모형 해석

② 차이를 통한 해석

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

$$\downarrow$$

$$\frac{\mu(x+1)}{\mu(x)} = e^{\beta}$$

다른 설명변수들이 **고정**되어 있을 때

X가 한 단위 증가하면 **기대도수** μ가 e^{β}**배**만큼 **증가**

5 포아송 회귀 모형

포아송 회귀 [illegible] 해석

② 차이를 통 [illegible]

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

$$\frac{\mu(x+1)}{\mu(x)} = e^{\beta}$$

포아송 회귀 모형의 한계

포아송 회귀 모형은 도수 자료(μ)를 반응변수로 삼음

→ 시간, 공간 등의 요소의 차이 반영 X

→ μ값의 예측값(= 기대도수 값)만을 산출

비율자료를 활용한 율자료 포아송 회귀 모형 사용!

다른 설명변수들이 고정되어 있을 때

X가 한 단위 증가하면 기대도수 μ가 e^{β}배만큼 증가

5 포아송 회귀 모형

율자료 포아송 회귀 모형

도시별 범죄 발생 건수 ← 해당 도시의 인구수에 크게 영향을 받음

시공간 등 **크기를 나타내는 지표**에 걸쳐 사건이 발생한 확률

→ **비율자료**를 이용해야 정확한 결과 산출 가능!

율자료 포아송 회귀 모형

율자료 포아송 회귀 모형

기존의 도수(μ) 대신 **비율자료**를 반응변수로 사용

앞선 포아송 회귀 모형과 같이 **로그 연결함수** 사용

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

t : 지표값, 비율을 구할 때 분모에 들어가는 값

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

→ 기대도수 = 지표 t에 비례

율자료 포아송 회귀 모형 해석

차이를 통한 해석

$$\log(\mu(x+1)/t) - \log(\mu(x)/t) = \log(\mu(x+1)) - \log(\mu(x)) = \log\left(\frac{\mu(x+1)}{\mu(x)}\right) = \beta$$

$$\downarrow$$

$$\frac{\mu(x+1)}{\mu(x)} = e^{\beta}$$

다른 설명변수들이 **고정**되어 있을 때

X가 한 단위 증가하면 **기대비율**이 e^{β}**배**만큼 **증가**

기대도수 아님!

포아송 회귀 모형의 문제점

① 과대산포 문제 (Overdispersion)

등산포 가정

포아송 분포의 평균과 분산이 같다는 성질

현실에서 대부분의 데이터는

평균에 비해 **분산이 크게 나타나** 등산포 가정 만족 X

= 과대 산포

→ 이를 무시하고 포아송 회귀 모형을 적용한다면

분산이 과소평가되어 **검정 결과**가 **왜곡**

포아송 회귀 모형의 문제점

① 과대산포 문제 (Overdispersion)

등산포 가정

포아송 분포의 평균과 분산이 같다는 성질

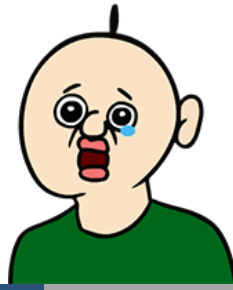
현실에서 대부분의 데이터는

평균에 비해 **분산이 크게 나타나** 등산포 가정 만족 X

= 과대 산포

→ 이를 무시하고 포아송 회귀 모형을 적용한다면

분산이 과소평가되어 **검정 결과**가 **왜곡**



포아송 회귀 모형의 문제점

① 과대산포 문제 (Overdispersion)

대표적 해결 방법인 **음이항 회귀 모형**을 이용해 해결 가능!

음이항 회귀 모형

음이항 회귀 모형

음이항 랜덤성분과 **로그 연결함수**로 구성된 GLM

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

D : **산포모수**, 분산이 평균보다 큰 값을 가지도록 해줌

포아송 분포와 달리 **분산에 $D\mu^2$가 더해진** 형태

음이항 회귀 모형

음이항 회귀 모형

음이항 랜덤성분과 **로그 연결함수**로 구성된 GLM

① D가 모든 예측값에 대해 동일하다고 가정

② 평균과 분산 간 비선형적인 관계가 있다고 가정

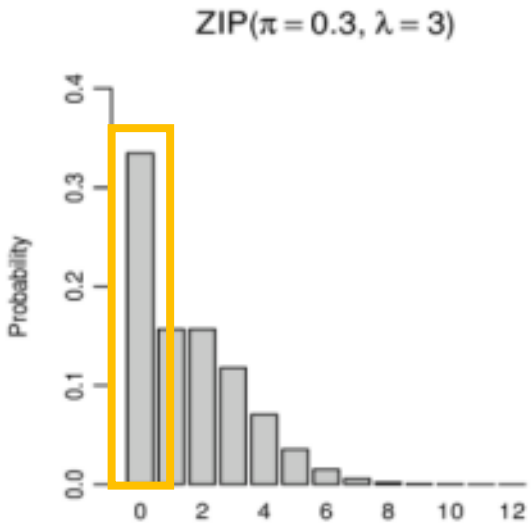
이러한 성질을 이용해 포아송 분포가 가지는

등산포 가정을 완화하고 **과대산포 문제 해결** 가능

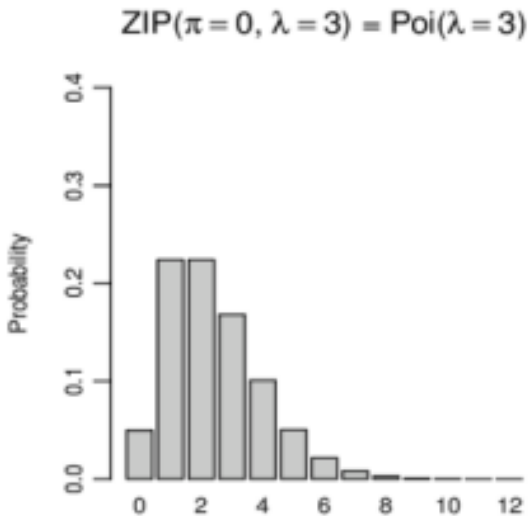
포아송 회귀 모형의 문제점

② 과대영 문제 (Excess Zeros)

포아송 분포를 통해 예상된 0 발생 횟수보다 **실제로 더 많은 0이 발생**한 경우



과대영 문제가 **발생**한 경우



과대영 문제가 발생하지 않은 경우

포아송 회귀 모형의 문제점

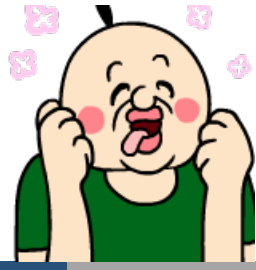
② 과대영 문제 (Excess Zeros)

포아송 분포를 통해 예상된 0 발생 횟수보다 **실제로 더 많은 0이 발생**한 경우

이러한 **과대영 문제**는 **영과잉 포아송 회귀 모형**이나

영과잉 음이항 회귀 모형을 이용해 **해결 가능**

이번 클린업에서는 영과잉 포아송 모형을 다룰 예정!



영과잉 포아송 회귀 모형

영과잉 포아송 분포

0의 값만 가지는 점 **확률분포**와 **포아송 분포**가 결합된 **혼합분포 구조**

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

- Y: 베르누이 분포를 따름
- $g(y_i)$: 0 초과 정수 값들이 따르는 포아송분포
- ϕ_i: 0이 발생할 확률
- $1-\phi_i$: 0 이외의 값을 가질 확률

영과잉 포아송 회귀 모형

영과잉 포아송 회귀 모형

영과잉 포아송 분포를 이용해 만든 **GLM**

총 2가지 식으로 이루어짐

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

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

영과잉 포아송 회귀 모형

영과잉 포아송 회귀 모형

영과잉 포아송 분포를 이용해 만든 **GLM**

총 2가지 식으로 이루어짐

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

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

0값이 발생할 확률(ϕ_i)을

로짓 연결함수를 사용해 표현

영과잉 포아송 회귀 모형

영과잉 포아송 회귀 모형

영과잉 포아송 분포를 이용해 만든 **GLM**

총 2가지 식으로 이루어짐

포아송분포의 평균(λ)을

로그 연결함수를 사용해 표현

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

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

다음주 예고

1. 혼동행렬
2. ROC 곡선
3. 샘플링
4. 인코딩

THANK YOU