

범주형자료분석팀

3팀

이정민
이경미
이현진
윤예빈
김준영

CONTENTS

1. GLM

2. 유의성 검정

3. 로지스틱 회귀 모형

4. 다범주 로짓 모형

5. 포아송 회귀 모형

1

GLM

GLM

잘 적합된 모형(*model*)의 장점

(1) 모형의 구조로부터 변수들 간의 연관성과 교호작용의 형태 파악

(2) 모수에 대한 추론을 통해 설명변수가 반응변수에 미치는 영향 파악

(3) 모수 추정값들로부터 효과들의 중요도와 크기 결정 가능

GLM

췌 익



연속형 반응변수가 주어진 상황

선형회귀모델(Linear Model)에서 **최소제곱법(LSM)**을 통해
연속형 반응변수와 설명변수들 간의 상관관계를 추정할 수 있음



우리가 다루는 데이터의 반응변수가 언제나 연속형은 아님!
반응변수가 범주형이거나 도수자료인 경우, 일반 선형회귀모델 사용 불가!

GLM

일반화 선형 모형 (*Generalized Linear Models*)

연속형 뿐만 아니라 다양한 형태의 반응변수에 대한
모형들을 포함하는 **광범위한 모형들의 집합**



보통의 선형회귀모형과 분산분석(ANOVA) 모형에서 나아가
범주형 변수에 대한 모형도 포함되어 있음

GLM의 필요성



정규분포를 포함한 다양한 확률분포를 사용할 수 있음



변수간 연관성을 파악하고 반응변수를 예측할 수 있음



GLM의 필요성



정규분포를 포함한 다양한 확률분포를 사용할 수 있음



반응변수가 **범주형 or 도수자료인 경우**, 오차항의 확률 분포가 정규분포를 따르지 않기 때문에 **LSE**를 사용하는 것은 바람직하지 못함

하지만 GLM은 MLE를 사용함으로써 분석을 가능하게 함!

GLM의 필요성



정규분포를 포함한 다양한 확률분포를 사용할 수 있음



반응변수가 **범주형 or 도수자료인 경우**, 오차항의 확률 분포가 정규분포를 따르지 않기 때문에 **LSE**를 사용하는 것은 바람직하지 못함

하지만 GLM은 MLE를 사용함으로써 분석을 가능하게 함!

GLM의 필요성



변수간 연관성을 파악하고 반응변수를 예측할 수 있음



범주형 변수들 간의 연관성만 파악할 수 있는 분할표와 달리
GLM은 범주형 변수와 연속형 변수 간 **연관성을 파악**할 수 있으며,
새로운 설명변수에 대한 **반응변수를 예측**할 수 있음



GLM의 구성 성분

GLM의 일반적인 형태

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

GLM 구성 성분

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



평균의 함수식과 설명변수 간의 관계를 선형예측식(체계적 성분)을 통해 정의

GLM의 구성 성분

랜덤성분 (*Random Component*)

반응변수 Y 를 정의하고, Y 의 확률분포를 가정함
이 때, 가정한 확률분포 하에서 Y 의 기댓값 μ 을 랜덤성분으로서 표기

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

지수족 (*Exponential Family*)에 해당하는 확률분포만 사용 가능

GLM의 구성 성분

체계적 성분 (*Systematic Component*)

설명변수 X 를 명시하는 성분으로, X 들의 선형 결합 형태

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

✓ 체계적 성분에서 포함할 수 있는 요소

- ① 교호작용을 설명하는 항 ($x_i = x_a x_b$)
- ② 곡선효과를 나타내는 항 ($x_i = x_a^2$)

GLM의 구성 성분

연결함수 (*Link Function*)

랜덤성분과 체계적 성분을 연결하는 함수

두 성분 간 범위를 맞춰주는 역할

종류	반응변수	표기
항등 연결 함수 (<i>Identity Link</i>)	연속형 자료	$g(\mu) = \mu$
로그 연결 함수 (<i>Log Link</i>)	포아송, 음이항 분포를 따르는 도수자료	$g(\mu) = \log(\mu)$
로짓 연결 함수 (<i>Logit Link</i>)	이항 분포를 따르는 0~1 사이의 값 (확률 등)	$g(\mu) = \log[\mu/(1 - \mu)]$

GLM의 특징

특징

- (1) 오차항의 다양한 분포 가정 가능
- (2) 선형 관계식 유지
- (3) 독립성 가정만 필요
- (4) 제한적인 범위를 지닌 반응변수도 사용 가능



GLM의 특징



오차항의 다양한 분포 가정 가능



GLM은 정규분포 외에도
반응변수의 오차항이 가진 성질에 따라 어떤 분포든 정의 가능 !

일반적으로 확률분포에 따른 연결함수는 정해져 있음

GLM의 특징



선형 관계식 유지



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



GLM의 체계적 성분은 선형 관계식의 형태를 띄고 있기 때문에 해석 용이



GLM의 특징

'선형'의 관계의 의미



선형 관계식 유지

설명변수 X 와 반응변수 Y 간의 관계를 말하는 것이 아닌,

회귀 계수 β 의 선형성을 가리킴!



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

교호작용이나 곡선효과를 나타내는 항이

우변에 위치하고 있더라도 선형성이 유지됨

GLM의 체계적 성분은 선형 관계식의 형태를 띄고 있기 때문에 해석 용이

GLM의 특징



독립성 가정만 필요

선형회귀모델과 달리, GLM은 회귀분석의 4가지 가정 중
오차항에 대한 독립성만 만족하면 됨



반응변수 간의 자기상관성을 검정해 보아야 함



오차들의 독립성 가정이 위배되었을 때 오차들 간 자기상관이 있다고 함

GLM의 특징



독립성 가정만 필요

선형회귀모델과 달리, GLM은 회귀분석의 4가지 가정 중
오차항에 대한 독립성만 만족하면 됨



반응변수 간의 자기상관성을 검정해 보아야 함



오차들의 독립성 가정이 위배되었을 때 오차들 간 자기상관이 있다고 함

GLM의 특징



제한적인 범위를 지닌 반응변수도 사용 가능

GLM은 연결함수를 통해 좌변의 랜덤성분과
우변의 체계적성분 간의 범위를 맞춰줄 수 있음



제한범위를 가진 반응변수(범주형 자료, 도수 자료 등)도 사용할 수 있음

GLM의 특징



제한적인 범위를 지닌 반응변수도 사용 가능

GLM은 연결함수를 통해 좌변의 랜덤성분과
우변의 체계적성분 간의 범위를 맞춰줄 수 있음



제한범위를 가진 반응변수(범주형 자료, 도수 자료 등)도 사용할 수 있음

GLM의 특징



(4) 제한적인 범위를 지닌 반응변수도 사용 가능

GLM은 보통의 선형모형을 일반화한 것

랜덤성분의 분포와 **랜덤성분의 함수(연결함수)**를

일반화한 것이 GLM

GLM은 **연결함수**를 통해 좌변의 랜덤성분과

우변의 체계적성분 간의 범위를 맞춰줄 수 있음

랜덤성분이 **정규분포가 아닌 다른 분포**를 가질 수 있으므로

범주형 변수 등을 다룰 수 있게 됨

제한범위를 가진 반응변수(범주형 자료, 도수 자료 등)도 사용할 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	

GLM의 종류

GLM	랜덤성분	연결함수	체계적 성분
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

(i) 반응변수 : 이항 자료

선형 확률 모형

$$\pi(x) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

이항 랜덤 성분과 항등 연결 함수

로지스틱 회귀 모형

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

이항 랜덤 성분과 로짓 연결 함수

프로빗 회귀 모형

$$\Phi^{-1}(\mu) = \text{probit}(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

이항 랜덤 성분과 프로빗 연결

GLM의 종류

(ii) 반응변수 : 다항 자료

누적 로짓 모형

$$P(Y \leq j) = \log \left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} \right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k, j = 1, \dots, J-1$$

이항 랜덤 성분과 로짓 연결 함수

연속비 로짓 모형

$$\log \left(\frac{\pi_j}{\pi_{j+1} + \cdots + \pi_J} \right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k, j = 1, \dots, J-1$$

$$\log \left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1}} \right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k, j = 1, \dots, J-1$$

다항 랜덤 성분(순서형)과 로짓 연결 함수

GLM의 종류

(ii) 반응변수 : 도수 자료

포아송 회귀 모형

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

포아송 랜덤 성분과 로그 연결 함수

음이항 회귀 모형

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

음이항 랜덤 성분과 로그 연결 함수

율자료 포아송 회귀 모형

$$\log(\mu/t) = \log(\mu) - \log(t) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

포아송 랜덤 성분과 로그 연결 함수

GLM의 모형 적합

모형 적합 (*Model Fitting*)

주어진 데이터를 근거로 모형의 모수를 추정하는 과정



GLM은 회귀의 기본 4가지 가정을 만족하지 못하므로
LSE를 사용하는 것이 바람직하지 않음



최대가능도추정법 (*Maximum Likelihood Estimation, MLE*)을 활용

GLM의 모형 적합

모형 적합 (*Model Fitting*)

주어진 데이터를 근거로 모형의 모수를 추정하는 과정



GLM은 회귀의 기본 4가지 가정을 만족하지 못하므로
LSE를 사용하는 것이 바람직하지 않음



최대가능도추정법 (*Maximum Likelihood Estimation, MLE*) 을 활용

GLM의 모형 적합

가능도 (*Likelihood*)

고정된 관측값이 어떤 확률분포를 따를 가능성

가능도들을 다 곱한 것이

아래의 가능도 함수 (*Likelihood Function*)



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

GLM의 모형 적합

최대가능도추정량 (*Maximum Likelihood Estimator*)

가능도함수 $P(x|\theta)$ 가 최대가 되도록하는 추정량 $\hat{\theta}$



확률표본 X_1, \dots, X_n 이 모수가 λ 인 지수분포 ($f_x(x) = \lambda e^{-\lambda x}$)를 따를 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 값은

$$\frac{d \ln L(\lambda)}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \text{ 임에 따라 } \hat{\lambda} = \frac{1}{\bar{x}}$$

2

유의성 검정

유의성 검정

유의성 검정

모형의 모수에 대한 추정값이 유의한지
혹은 축소 모형의 적합도가 좋은지 판단하는 검정

GLM 모형에서의 유의성 검정 가설

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

H_1 : 적어도 하나의 β 는 0이 아니다

귀무가설 기각 시 모형이 유의하다고 해석 !

유의성 검정

유의성 검정

모형의 모수에 대한 추정값이 유의한지
혹은 축소 모형의 적합도가 좋은지 판단하는 검정



유의성 검정

wald 검정

가능도비 검정

스코어 검정

왈드 검정 (*Wald Test*)

왈드 검정

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

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



회귀계수에 대한 추정값과 표준오차만 사용하여 통계량을 구함

왈드 검정 (Wald Test)

왈드 검정

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

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



하지만 범주형 자료이거나 소표본인 경우 검정력 감소

회귀계수에 대한 추정값과 표준오차만 사용하여 통계량을 구함



가능도비 검정으로 해결!

가능도비 검정 (*Likelihood-ratio Test*)

가능도비 검정

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

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



귀무가설 하에서의 가능도함수(l_0)와
전체공간 하에서의 가능도함수(l_1)의 비(ratio) 이용



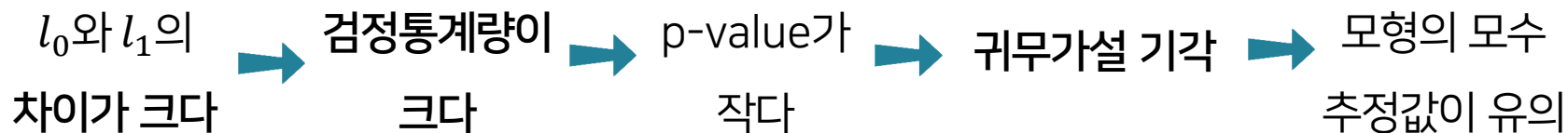
모수에 대한 아무런 제약이 없는 상태로 귀무가설 + 대립가설 상태

가능도비 검정 (*Likelihood-ratio Test*)

가능도비 검정통계량의 의미

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

검정 과정



가능도비 검정 (*Likelihood-ratio Test*)

가능도비 검정통계량의 의미

$$G^2 = -2 \log \left(\frac{\text{모수가 귀무가설 H}_0 \text{를 만족할 때 } (\beta = 0) \text{ 가능도 함수의 최대값}}{\text{모수가 아무런 제약 없이 가능도 함수의 최대값}} \right)$$


MLE로 계산

가능도비 검정은 귀무가설 하에서의 가능도함수와
전체공간 하에서의 가능도함수를 모두 사용

검정과정

왈드 검정에 비해 **검정력이 좋고 신뢰도도 높음!**

l_0 와 l_1 의 차이가 크다 → 검정통계량이 크다 → p-value가 작다 → 귀무가설 기각 → 모형의 모수 추정값이 유의

이탈도 (*Deviance*)

이탈도

관심모형(M)과 포화모형(S)을 비교하기 위한 가능도비 통계량

관심모형 (M)

유의성을 검정하고자 하는 모형

Ex) 범주팀의 행복 (Y) = $\beta_0 + \beta_1 \times$ 세미나 시간(x_1) + $\beta_2 \times$ 패키지 난이도(x_2)

포화모형 (S)

모든 관측값에 대해 모수를 갖는 가장 복잡한 모형

Ex) 범주팀의 행복 (Y) = $\beta_0 + \beta_1 \times$ 세미나 시간(x_1) + $\beta_2 \times$ 패키지 난이도(x_2)
+ $\beta_3 \times$ 교안 페이지 수(x_3) + $\beta_4 \times$ 클린업 발표 질문 개수(x_4)

이탈도 (*Deviance*)

이탈도의 귀무가설과 대립가설

H_0 : 관심모형에 속하지 않는 모수는 모두 0이다.

H_1 : 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다.



귀무가설 채택 ➡ 관심모형 사용

대립가설 채택 ➡ 관심모형에 **모수가 추가된 모형 필요**

이탈도 (Deviance)

L_M : 관심모형 하에서의 로그 가능도 함수의 최댓값

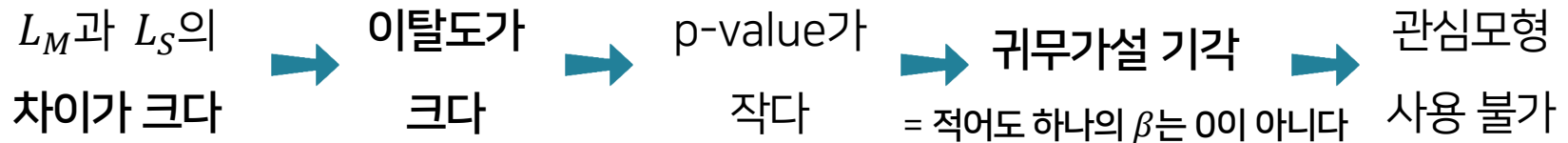
L_S : 포화모형 하에서의 로그 가능도 함수의 최댓값

이탈도

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



검정 과정



이탈도 (*Deviance*)

L_M : 관심모형 하에서의 로그 가능도 함수의 최댓값

L_S : 포화모형 하에서의 로그 가능도 함수의 최댓값

이탈도

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



즉 이탈도는 두 모형의 가능도 함수의 최댓값 차이($L_M - L_S$)를 이용
GLM 모형에서의 이탈도는 근사적으로 카이제곱분포를 따름

이탈도 (Deviance)



이탈도 사용 조건

L_M : 관심모형 하에서의 로그 가능도 함수의 최댓값
 L_S : 포화모형 하에서의 로그 가능도 함수의 최댓값

이탈도

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

이탈도는 포화모형에는 있지만 관심모형에는 없는

계수들이 0인지 여부를 확인하는 것이므로

관심모형은 포화모형에 **내포된(nested) 관계**여야 함!



관심모형(M)의 모수 \subset 포화모형(S)의 모수

즉 이 둘은 포화모형의 로그 가능도 함수의 최댓값 차이($L_M - L_S$)를 이용 .. 뭐?

GLM 모형에서의 이탈도는 근사적으로 카이제곱분포를 따름



이탈도와 가능도비 검정의 관계

$$\begin{aligned} & M_0 \text{의 이탈도} - M_1 \text{의 이탈도 (모형 간 이탈도의 차이)} \\ &= -2(L_0 - L_S) - (-2(L_1 - L_S)) \\ &= -2(L_0 - L_1) \text{ (=가능도비 검정 통계량)} \end{aligned}$$



두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음

이탈도와 가능도비 검정의 관계

M_0 의 이탈도 - M_1 의 이탈도 (모형 간 이탈도의 차이)

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

$$= -2(L_0 - L_1) \text{ (=가능도비 검정 통계량)}$$

히 히



이 성질을 활용해 관심모형 2개 중 어느 것이 적합한지 비교해볼 수 있음

이탈도와 가능도비 검정의 관계

검정 과정



관심모형(M_0, M_1)간 이탈도의 차이가 작음 (\rightarrow 가능도비 검정통계량이 작음)



p-value가 큼



귀무가설 기각 불가 ($\rightarrow M_0$ 에 포함되지 않는 모수들은 모두 0)



간단한 관심모형 M_0 이 더 적합

이탈도와 가능도비 검정의 관계



관심모형(M_0, M_1)은 이탈도의 차이가 적음 (가능도비 검정 통계량이 작음)

마찬가지로 이 과정에서도 이탈도를 활용하므로
모형 M_0 은 모형 M_1 에 내포된(nested)된 모형이어야 함

p-value가 큼

귀무가설 기각 그러나 내포된 경우가 아니라면 (수들은 모두 0)

AIC, BIC 등의 모형 선택 측도를 활용하여 모형 비교

간단한 관심모형 M_0 이 더 적합

3

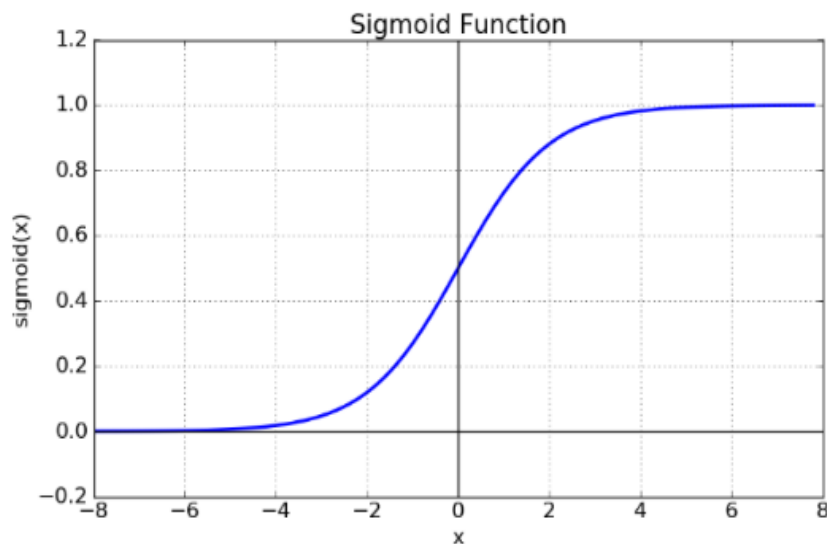
로지스틱 회귀 모형

로지스틱 회귀 모형

로지스틱 회귀 (*Logistic Regression*)

반응변수 Y가 이항자료일 때의 회귀

즉 반응변수는 이항분포를 따르고 성공일 확률로 표기됨



로지스틱 회귀 모형은
성공 확률과 x 의 비선형 관계를 나타냄
즉 이항변수와 연속형 변수들 간의 관계를
GLM의 형태로 표현한 것

시그모이드 함수, 딥러닝에서 활성화 함수로 자주 쓰임!

로지스틱 회귀 모형의 장점

이항변수와 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 : $0 \sim 1 \neq$ 우변 범위 : $-\infty \sim \infty$

⋮



이항분포를 따르는 반응변수 $\pi(x)(= P(Y = 1|X = x))$ 는 $0 \sim 1$ 의 값을 가지고
설명변수의 **선형식**은 $-\infty \sim \infty$ 의 값을 가져 양변의 범위가 맞지 않음

로지스틱 회귀 모형의 장점

이항변수와 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 : $0 \sim 1 \neq$ 우변 범위 : $-\infty \sim \infty$

|| 좌변을 오즈 형태로 만듭!

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

좌변 범위 : $0 \sim \infty \neq$ 우변 범위 : $-\infty \sim \infty$

|| 좌변에 로그 (로짓 연결함수)

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

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

로지스틱 회귀 모형의 장점

이항변수와 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 : $0 \sim 1 \neq$ 우변 범위 : $-\infty \sim \infty$

|| 좌변을 오즈 형태로 만듦!

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

좌변 범위 : $0 \sim \infty \neq$ 우변 범위 : $-\infty \sim \infty$

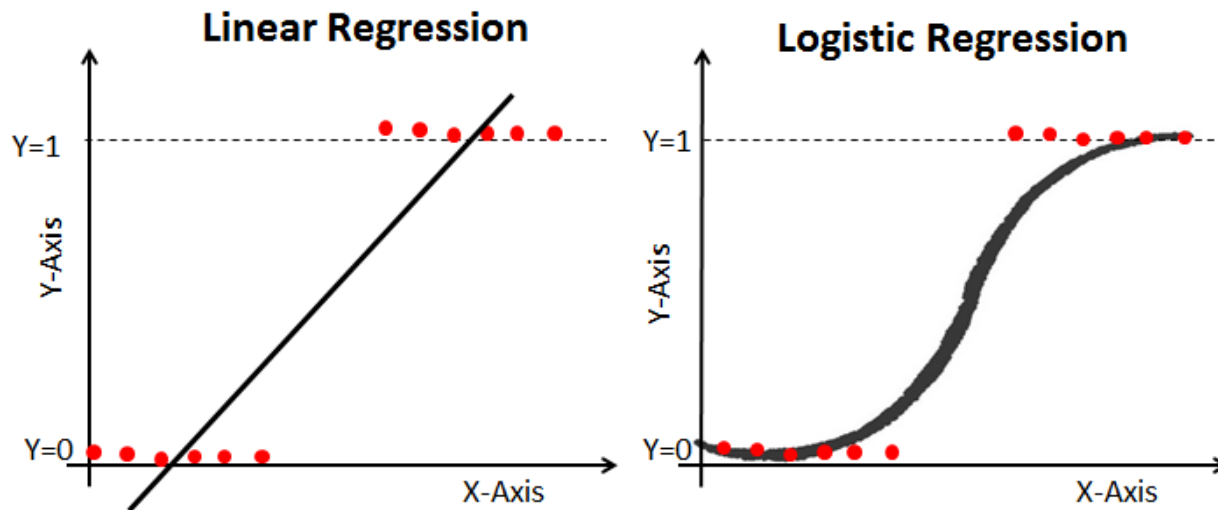
|| 좌변에 로그 (로짓 연결함수)

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

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

로지스틱 회귀 모형의 장점

이항변수와 연속형 변수 간의 범위 일치



일반 선형 모형

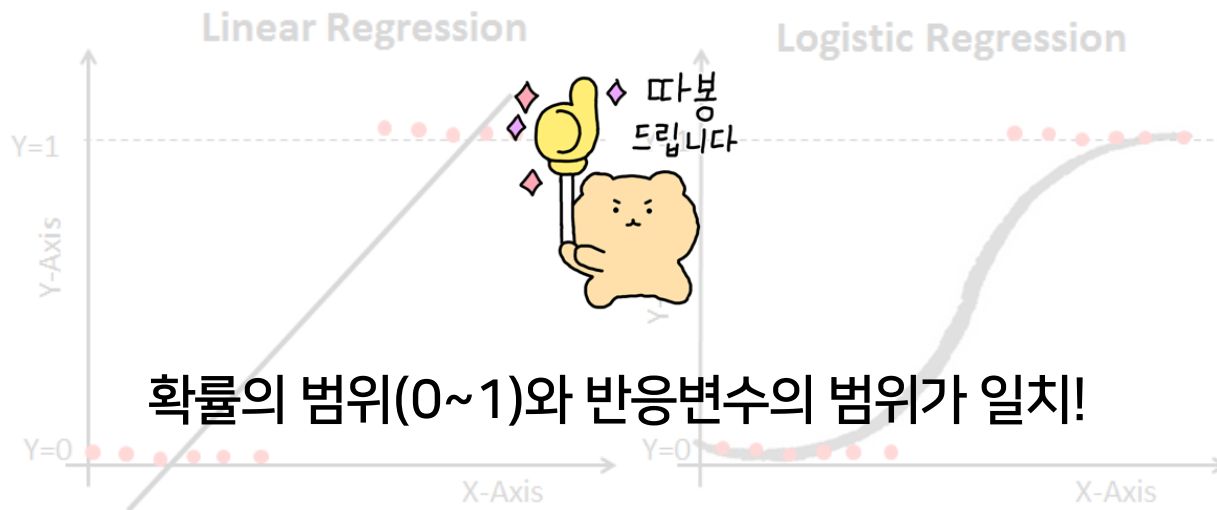
Y의 범위가 0과 1을 초과

로지스틱 회귀 모형

Y의 범위가 0과 1 사이

로지스틱 회귀 모형의 장점

이항변수와 연속형 변수 간의 범위 일치



일반 선형 모형

Y의 범위가 0과 1을 초과

로지스틱 회귀 모형

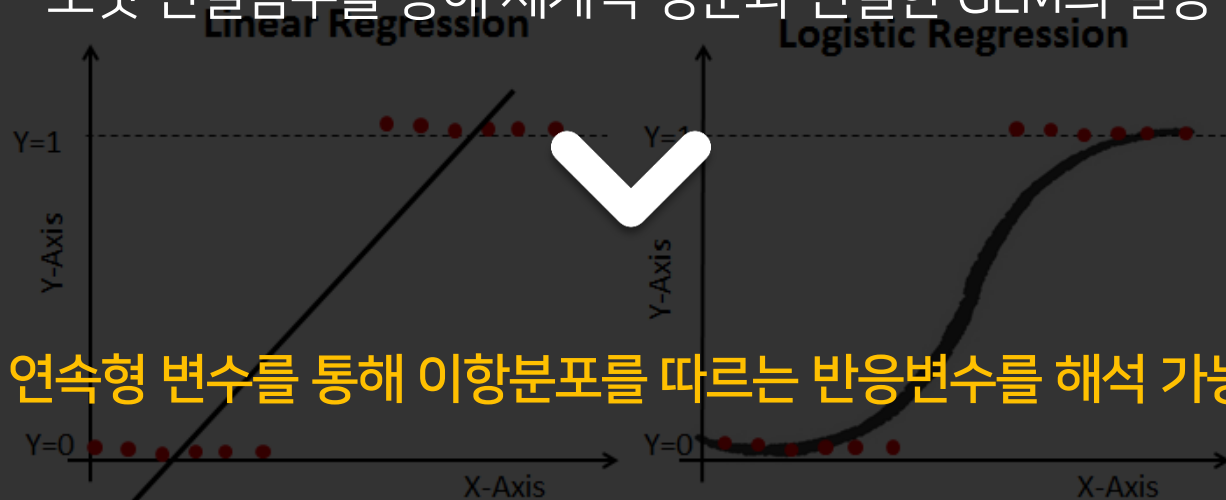
Y의 범위가 0과 1 사이

로지스틱 회귀 모형의 장점



로지스틱 회귀 모형은 이항 랜덤성분을

로짓 연결함수를 통해 체계적 성분과 연결한 GLM의 일종



연속형 변수를 통해 이항분포를 따르는 반응변수를 해석 가능

일반 선형 모형

Y의 범위가 0과 1을 초과

로지스틱 회귀 모형

Y의 범위가 0과 1 사이

로지스틱 회귀 모형의 장점



모수들이 오즈비와 관련되어 있기 때문에
후향적 연구에서도 사용 가능



일반 회귀 모형이 충족시켜야 하는 4가지 가정 중
오직 독립성 가정만 만족하면 됨

종 아

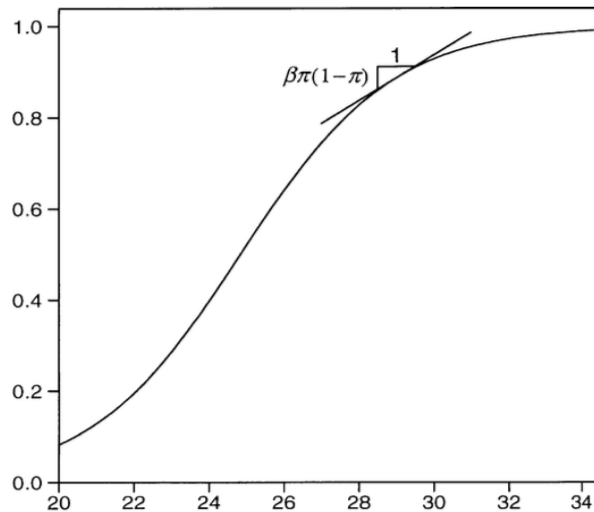


로지스틱 회귀 모형

로지스틱 회귀 모형의 기울기

로지스틱 회귀 모형을 x 에 대해 미분

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

회귀계수(β)에 따라 증가/감소 비율이 결정 β 가 양수면 상향곡선 β 가 음수면 하향곡선 β 의 절댓값이 클수록 가파른 형태

로지스틱 회귀 모형의 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 $\pi(x)$ 는 특정 x 를 대입해 구할 수 있음
 $\pi(x)$ 값이 cut-off point 보다 크면 $Y=1$, 작으면 $Y=0$ 으로 예측

일반적으로 0.5 사용! (그러나 상황에 따라 조절 가능)

로지스틱 회귀 모형의 해석

확률을 통한 해석

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

축구 직관 횡수(X)에 따른 싸인 유니폼 당첨 여부(Y)

$\log \left[\frac{\pi(x)}{1-\pi(x)} \right] = 2 + 0.02x$ 라고 가정할 경우 관측치 X가 20이라면,

$$\pi(x) = \frac{\exp(2+0.02x)}{1+\exp(2+0.02x)} = \text{약 } 0.92 > 0.5$$



싸인 유니폼 당첨 (Y=1) 이라고 판단!


로지스틱 회귀 모형의 해석

오즈비를 통한 해석

다른 설명변수가 모두 고정되어 있을 때
x가 한 단위 증가하면 오즈가 e^β 배 증가한다고 해석

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

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

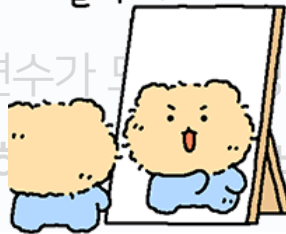

$$\frac{\pi(x+1)/[1 - \pi(x+1)]}{\pi(x)/[1 - \pi(x)]} = e^\beta$$

로지스틱 회귀 모형의 해석

오즈비를 통한 해석

할수 있다..!

다른 설명변수가 동일하게 되어 있을 때
 x 가 한 단위 증가하면 Y 가 e^β 배 증가한다고 해석



좌변의 분자는 '설명변수가 $x+1$ 일 때 Y 가 1일 오즈'이고

분모는 '설명변수가 x 일 때 Y 가 1일 오즈'이므로
 $x+1$ 일 때 Y 가 1일 오즈가 x 일 때 Y 가 1일 오즈보다 e^β 배 높다고 해석

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



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

로지스틱 회귀 모형의 해석

오즈비를 통한 해석

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

축구 직관 횡수(X)에 따른 싸인 유니폼 당첨 여부(Y)

$\log \left[\frac{\pi(x)}{1-\pi(x)} \right] = 2 + 0.02x$ 라고 가정할 경우 직관에 1번 더 간다면

싸인 유니폼에 당첨될 오즈가 $e^{0.02} = \text{약 } 1.02\text{배}$ 증가

4

다범주 로짓 모형

다범주 로짓 모형

다범주 로짓 모형 (*Multicategory Logit Model*)

3개 이상의 범주를 가진 반응변수로 확장시킨 모형
연결함수는 로짓 연결함수 사용, 랜덤성분은 다항분포 따름

반응변수의 범주가 3개 이상으로 늘어났기 때문에
명목형 자료와 순서형 자료를 구분해야 함!

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 연결함수 사용, 랜덤성분은 다항분포 따름



반응변수의 범주가 3개 이상으로 늘어났기 때문에
명목형 자료와 순서형 자료를 구분해야 함!

다범주 로짓 모형

기준 범주 로짓 모형

명목형 자료

누적 로짓 모형

순서형 자료

기준 범주 로짓 모형

기준 범주 로짓 모형 (*Baseline-Category Logit Model*)

반응변수가 명목형 자료일 때 사용하는 다범주 로짓 모형

기준 범주와 나머지 범주를 짝지어 로짓을 정의

일반적으로 반응변수의 여러 범주 중 마지막 범주!

기준 범주 J 와 그 외 범주들을 각각 짝지어 로짓을 정의

따라서 총 $(J-1)$ 개의 로짓 방정식이 생성

■ 기준 범주 로짓 모형

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



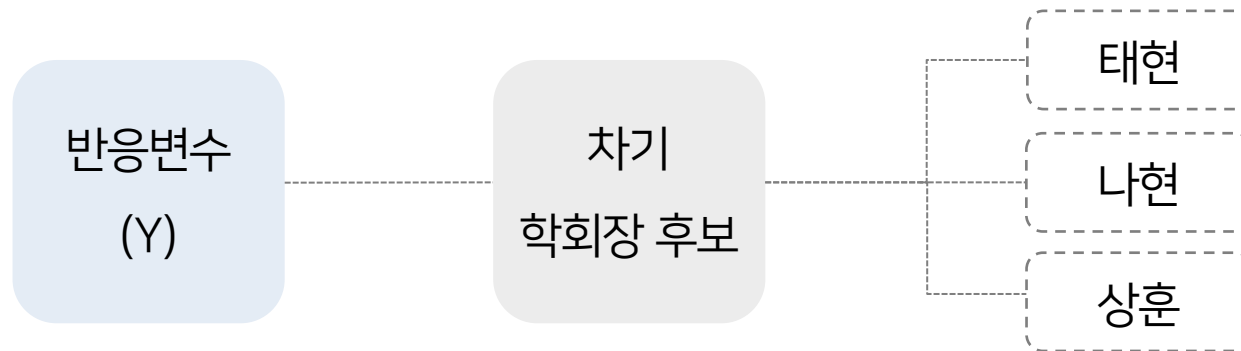
기준 범주 로짓 모형을 확률에 대한 식으로 재정의!

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

기존 범주 로짓 모형



차기 학회장 후보들의 당선 확률을 기존 범주 로짓 모형으로 알아보면



기존 범주를 '상훈'으로 정의

$$\log \left(\frac{\pi_{\text{나현}}}{\pi_{\text{상훈}}} \right) = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log \left(\frac{\pi_{\text{태현}}}{\pi_{\text{상훈}}} \right) = 2 + 0.22x_1 + \cdots + 0.46x_K$$

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

기준 범주 로짓 모형

차기 학회장 후보들의 당선 확률을 기준 범주 로짓 모형으로 알아보면

같은 설명변수여도 회귀계수 β 가 다른 값을 가지는 것을 알 수 있음



기준 범주를 '상훈'으로 정의

$$\log \left(\frac{\pi_{\text{나현}}}{\pi_{\text{상훈}}} \right) = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log \left(\frac{\pi_{\text{태현}}}{\pi_{\text{상훈}}} \right) = 2 + 0.22x_1 + \cdots + 0.46x_K$$

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

기준 범주 로짓 모형

나현이가 학회장이 될 확률

$$\pi_{\text{나현}} = \frac{e^{2+0.22x_1+\dots+0.46x_K}}{1 + e^{2+0.22x_1+\dots+0.46x_K} + e^{5+0.27x_1+\dots+0.59x_K}}$$



기준 범주 로짓 모형을 확률에 대한 식으로 재정의!

기준 범주를 '상훈'으로 정의

$$\log \left(\frac{\pi_{\text{나현}}}{\pi_{\text{상훈}}} \right) = 5 + 0.27x_1 + \dots + 0.59x_K$$

$$\log \left(\frac{\pi_{\text{태현}}}{\pi_{\text{상훈}}} \right) = 2 + 0.22x_1 + \dots + 0.46x_K$$

기준 범주 로짓 모형

j범주와 J범주(기준 범주) 간 비교

오즈를 이용한 해석

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



다른 설명변수들이 고정되어 있을 때

x_i 가 한 단위 증가하면 기준범주 대신 j범주일 오즈가 $e^{\beta_j^i}$ 배 증가

기준 범주 로짓 모형

기준 범주가 아닌 두 범주 간 비교

오즈를 이용한 해석

$$\begin{aligned}\log\left(\frac{\pi_a}{\pi_b}\right) &= \log\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) = \log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) \\ &= (\alpha_a + \beta_a^1 x_1 + \cdots + \beta_a^K x_K) - (\alpha_b + \beta_b^1 x_1 + \cdots + \beta_b^K x_K) \\ &= [\alpha_a - \alpha_b] + [(\beta_a^1 - \beta_b^1)x_1 + \cdots + (\beta_a^K - \beta_b^K)x_K]\end{aligned}$$



다른 설명변수들이 고정되어 있을 때

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

누적 로짓 모형

명목형 자료와는 달리 순서형 자료에서는 순서를 고려하므로
반응변수를 절단점에 따라 두 그룹으로 나누는 Collapse 과정이 필요

절단점
(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

누적 로짓 모형

순서형 자료는 절단점을 통해
Collapse를 진행하는 방법에 따라 다범주 로짓 모형들이 구분됨

절단점



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

이웃 범주 로짓 모형

절단점



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

연속비 로짓 모형

절단점



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

누적 로짓 모형

누적 로짓 모형

순서형 자료는 절단점을 통해
Collapse를 진행하는 방법에 따라 다범주 로짓 모형들이 구분됨

절단점



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

이웃 범주 로짓 모형

절단점



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

연속비 로짓 모형

절단점



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

누적 로짓 모형

누적 로짓 모형

누적 로짓 모형 (*Cumulative Logit Model*)

반응변수가 순서형 자료일 때 사용하는 다범주 로짓 모형

누적 확률에 로짓 연결함수를 사용



누적 로짓 모형

누적 확률 (Cumulative Probability)

반응변수가 총 J개의 범주를 가질 때
첫 번째 범주부터 j번째 범주까지의 누적확률



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

$j = 1, \dots, J$ 총 J개의 반응변수

누적 로짓 모형

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



누적확률을 로그 오즈의 형태로 조작

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



누적 로짓 모형의 최종 형태

$$\text{logit}[P(Y \leq j|X = x)]$$

누적 로짓 모형

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



누적확률을 로그 오즈의 형태로 조작

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



누적 로짓 모형의 최종 형태

$$\text{logit}[P(Y \leq j|X = x)]$$

누적 로짓 모형

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



누적확률을 로그 오즈의 형태로 조작

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



누적 로짓 모형의 최종 형태

$$\text{logit}[P(Y \leq j|X = x)]$$

누적 로짓 모형

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



누적확률을 로그 오즈의 형태로 조작

누적 로짓 모형의 최종 형태

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

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

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

누적 로짓 모형의 최종 형태

$$\text{logit}[P(Y \leq j | X = x)]$$



누적 로짓 모형과 기준 범주 로짓 모형 비교

누적 로짓 모형

$$\begin{aligned} & \text{logit}[P(Y \leq j | X = x)] \\ &= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned}$$

기준 범주 로짓 모형

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

공통점

기준점을 설정하여 비교하는 원리



총 J-1개의 로짓 방정식 생성

누적 로짓 모형과 기준 범주 로짓 모형 비교

누적 로짓 모형

$$\begin{aligned} & \text{logit}[P(Y \leq j | X = x)] \\ &= \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned}$$

기준 범주 로짓 모형

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

차이점

기준 범주에 상관없이 β 값 동일기준 범주에 따라 β 값 상이

누적 로짓 모형과 기준 범주 로짓 모형 비교

누적 로짓 모형

$$\begin{aligned} \text{logit}[P(Y \leq j | X = x)] \\ = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned}$$

기준 범주 로짓 모형

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

차이점

기준 범주에 상관없이 β 값 동일

비례 오즈 가정 때문!

4

다범주 로짓 모형



누적 로짓 모형과 기준 범주 로짓 모형 비교

비례 오즈 가정

누적 로짓 모형에서 (J-1)개의 로짓 방정식에 대한

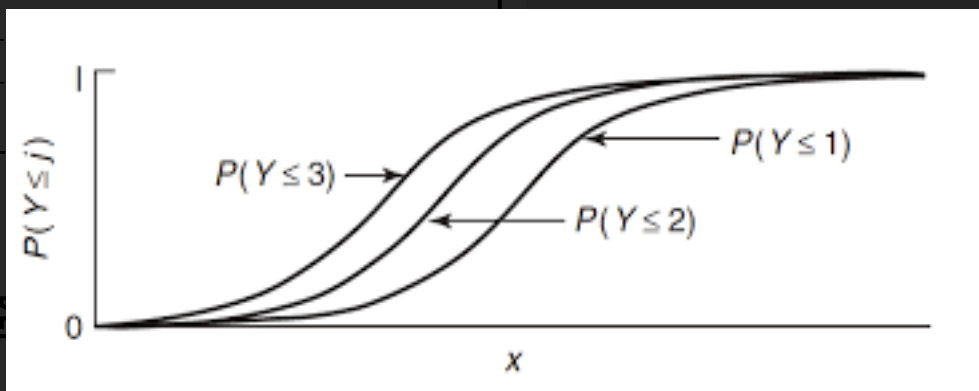
β 의 효과가 모두 동일하다는 가정

누적 로짓 모형

누적 로짓 모형

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \dots + \beta_K x_K$$

기준



β 값 동일

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \alpha_j + \beta_1 x_1 + \dots + \beta_K x_K$$

누적 로짓 모형의 β 값이 동일한 것은

서로 다른 로짓 방정식의 그래프가 수평 이동을 한 것처럼 나타남

→ α 는 다르지만 β 는 같음

누적 로짓 모형 예시

범주팀 클린업에 대한 불만도를
낮음 / 보통 / 높음 / 아주 높음으로 분류한 경우

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

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

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



기준 범주에 상관없이 **β 값이 동일**하다는 것을 확인!

누적 로짓 모형

오즈를 이용한 해석

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



다른 설명변수가 모두 고정되어 있을 때
 x_i 가 한 단위 증가하면 $Y > j$ 에 비해 $Y \leq j$ 일 오즈가 e^{β_i} 배 증가한다고 해석

5

포아송 회귀 모형

포아송 회귀 모형

랜덤성분의 분포

모형

정규 분포

일반 선형 회귀 모형

이항 분포

로지스틱 회귀 모형

다항 분포

다범주 로짓 모형

포아송 분포

포아송 회귀 모형

포아송 회귀 모형



포아송 분포 평균이 작은 경우

정규성과 등분산성 가정 충족하지 않아 일반 선형모형 적용 **불가능**



표준오차나 유의성 수준 편향 문제 발생



포아송 회귀 모형으로 해결!

포아송 회귀 모형



포아송 분포 평균이 작은 경우

정규성과 등분산성 가정 충족하지 않아 일반 선형모형 적용 **불가능**



표준오차나 유의성 수준 편향 문제 발생



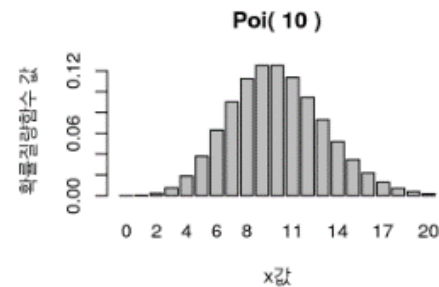
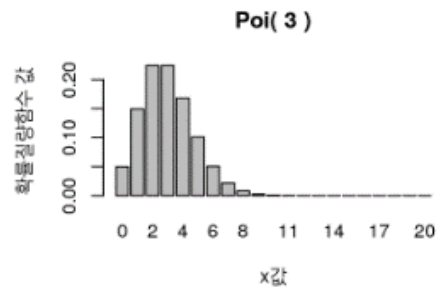
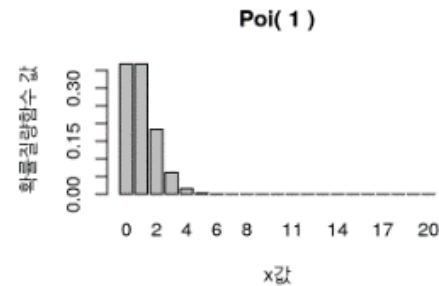
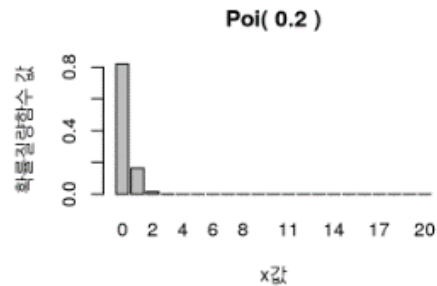
포아송 회귀 모형으로 해결!

포아송 회귀 모형

포아송 회귀 모형 (*Poisson Regression Model*)

반응변수가 도수 자료처럼 포아송 분포를 따를 때 사용

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





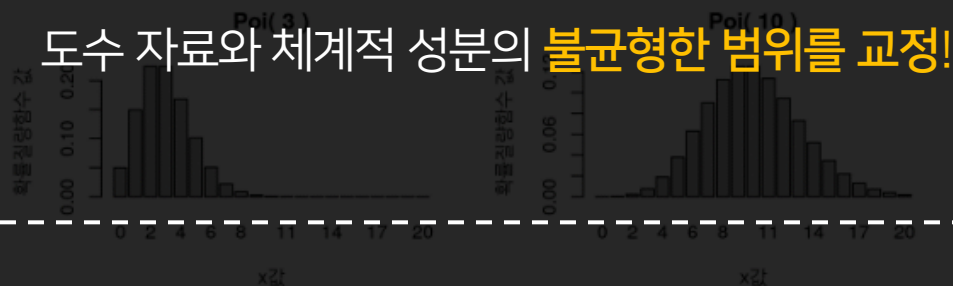
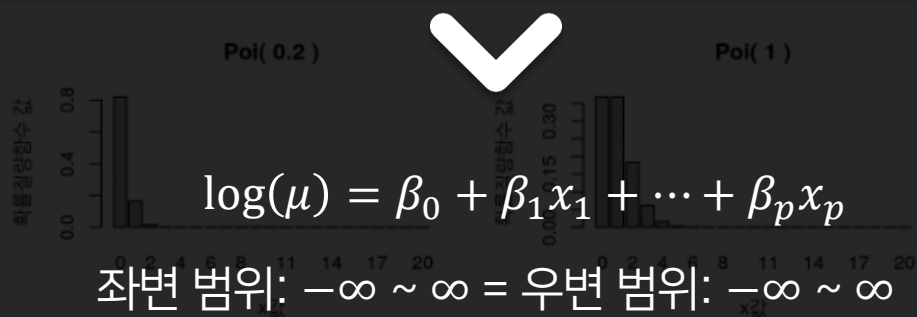
포아송 회귀 모형

로그 연결함수를 사용하는 이유

포아송 회귀 모형 (Poisson Regression Model)

반응변수가 $\mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ 를 따를 때 사용

좌변 범위: 음이 아닌 정수 \neq 우변 범위: $-\infty \sim \infty$



포아송 회귀 모형 해석

도수를 이용한 해석

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



포아송 회귀 모형을 도수($= \mu$)에 관한 식으로 변형




다른 설명변수들이 고정되어 있을 때
 x_i 가 한 단위 증가하면 μ 가 e^{β_i} 배 증가

포아송 회귀 모형 해석

오즈비를 이용한 해석

포아송 회귀 모형에 $(x+1)$ 과 x 를 대입한 후 빼기


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

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

다른 설명변수들이 고정되어 있을 때
 x 가 한 단위 증가하면 μ 가 e^{β} 배 증가

포아송 회귀 모형 예시

반응 변수 (Y)

평생 로또 1등에 당첨될 횟수

설명 변수 (X)

1년에 복권을 구매하는 금액

포아송 회귀모형

$$\log(\mu) = -2 + 0.0001x$$

포아송 회귀 모형 예시

도수를 이용한 해석

1년에 복권을 구매하는 금액 = 10,000원이라면

$$\mu = \exp(-2 + 0.0001 * 10,000) \approx 0.3679$$



기대도수가 1회 미만으로,

1년에 10,000원을 투자해서는 평생 한 번도 로또 1등에 당첨될 수 없음

포아송 회귀 모형 예시

오즈비를 이용한 해석

1년에 복권을 구매하는 금액이 1원 늘어난다면,

$$\frac{\mu(x+1)}{\mu(x)} = e^{0.0001} \approx 1.0001$$



평생 로또 1등에 당첨될 기대도수가 1.0001배만큼 증가함

포아송 회귀 모형의 한계



설명변수들이 시간, 공간 등의 요소의 차이를 반영하지 못함
 μ 에 대한 예측값인 기대도수만 산출 가능



비율 자료를 이용한 율자료 포아송 회귀 모형 사용

포아송 회귀 모형의 한계



설명변수들이 시간, 공간 등의 요소의 차이를 반영하지 못함
 μ 에 대한 예측값인 기대도수만 산출 가능



비율 자료를 이용한 율자료 포아송 회귀 모형 사용

율자료 포아송 회귀 모형

율자료 포아송 회귀 모형 (*Poisson Regression Model of Rate Data*)

기존의 기대도수 (μ) 대신 비율자료 ($\frac{\mu}{t}$)를 반응변수로 사용

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



기존 포아송 회귀 모형과 같이 로그 연결함수 사용

t 는 기준이 되는 지표 값을 나타냄

율자료 포아송 회귀 모형

율자료 포아송 회귀 모형 (*Poisson Regression Model of Rate Data*)

기존의 기대도수 (μ) 대신 비율자료 ($\frac{\mu}{t}$)를 반응변수로 사용

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



기존 포아송 회귀 모형과 같이 로그 연결함수 사용

t 는 기준이 되는 지표 값을 나타냄

율자료 포아송 회귀 모형

율자료 포아송 회귀 모형 (*Poisson Regression Model of Rate Data*)

기존의 기대도수 (μ) 대신 비율자료 ($\frac{\mu}{t}$)를 반응변수로 사용

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




기존 포아송 회귀 모형과 같이 로그 연결함수 사용

t 는 기준이 되는 지표 값을 나타냄

| 울자료 포아송 회귀 모형

오즈비를 이용한 해석

포아송 회귀 모형에 $(x+1)$ 과 x 를 대입한 후 빼기

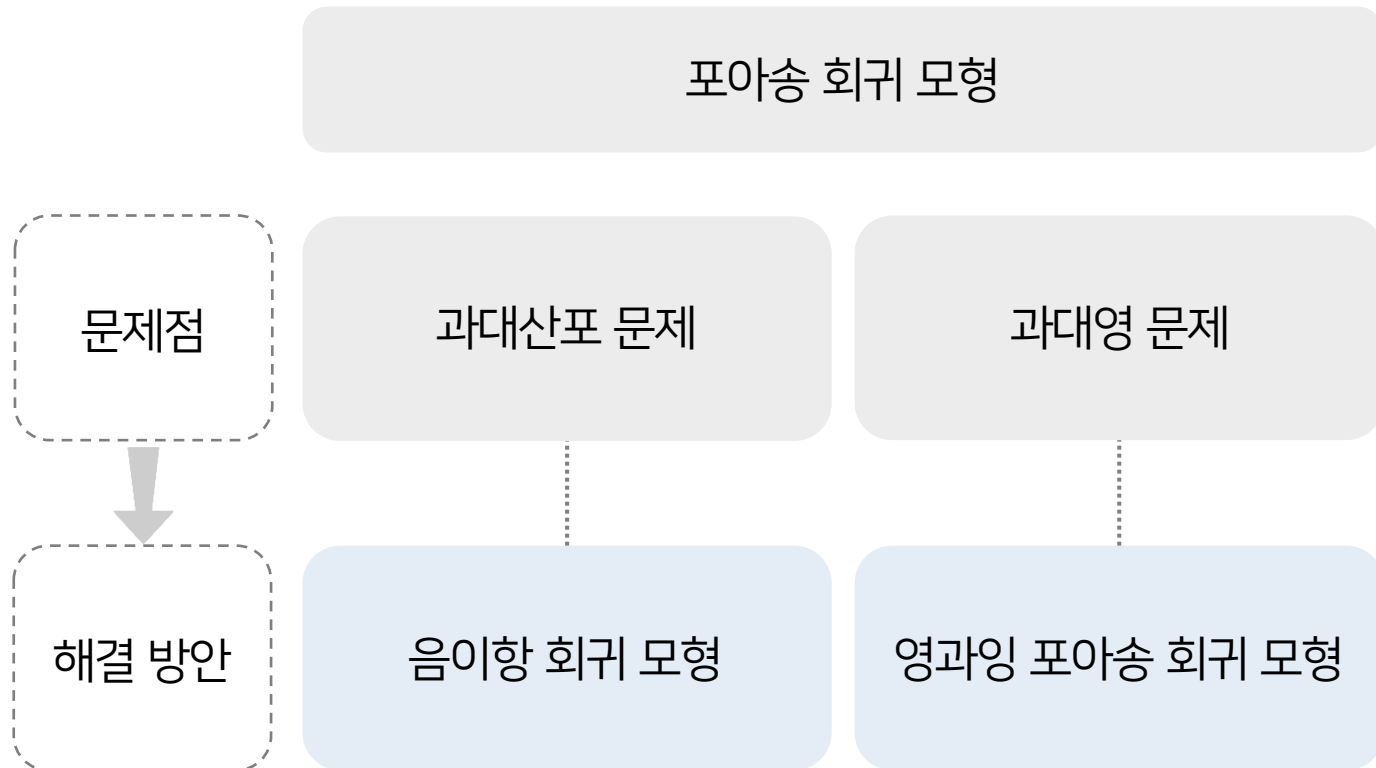

$$\log\left(\frac{\mu(x+1)}{t}\right) - \log\left(\frac{\mu(x)}{t}\right) = \log(\mu(x+1)) - \log(\mu(x))$$

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

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

다른 설명변수들이 고정되어 있을 때
 x 가 한 단위 증가하면 기대비율이 e^{β} 배 증가

포아송 회귀 모형의 문제점



포아송 회귀 모형의 문제점

과대산포 문제 (*Overdispersion*)

평균에 비해 분산이 크게 나타나 등산포 가정을 만족하지 못하는 경우



포아송분포에서 도수 자료의 평균과 분산이 같다는 성질



R에서 함수 `dispersiontest()` 을 통해 과산포 검정 진행 가능

포아송 회귀 모형의 문제점



현실에서 과대산포 문제 발생 시 분산이 과소평가되어 검정 결과가 왜곡됨



음이항 회귀 모형을 이용해 해결!

포아송 회귀 모형의 문제점



현실에서 과대산포 문제 발생 시 분산이 과소평가되어 검정 결과가 왜곡됨



음이항 회귀 모형을 이용해 해결!

포아송 회귀 모형의 문제점

음이항 회귀 모형 (*Negative Binomial Regression*)

랜덤성분이 음이항 분포를 따르며 로그 연결함수로 구성된 GLM

평균과 분산 간의 비선형성을 가정한 2차 함수 형태

$$\begin{aligned} E(Y) &= \mu \\ \text{Var}(Y) &= \mu + \mathbf{D}\mu^2 \end{aligned}$$

평균과 분산의 차이를
발생시키는 산포모수



음이항 분포는 평균보다 큰 분산 값을 가지기 때문에
포아송 분포의 등산포 가정을 완화해 과대산포 문제 해결

포아송 회귀 모형의 문제점

음이항 회귀 모형 (*Negative Binomial Regression*)

랜덤성분이 음이항 분포를 따르며 로그 연결함수로 구성된 GLM

평균과 분산 간의 비선형성을 가정한 2차 함수 형태

$$\begin{aligned} E(Y) &= \mu \\ \text{Var}(Y) &= \mu + \mathbf{D}\mu^2 \end{aligned}$$

평균과 분산의 차이를 발생시키는 산포모수

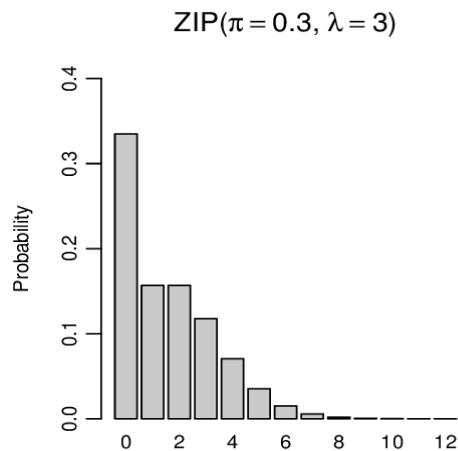


음이항 분포는 평균보다 큰 분산 값을 가지기 때문에
포아송 분포의 등산포 가정을 완화해 과대산포 문제 해결

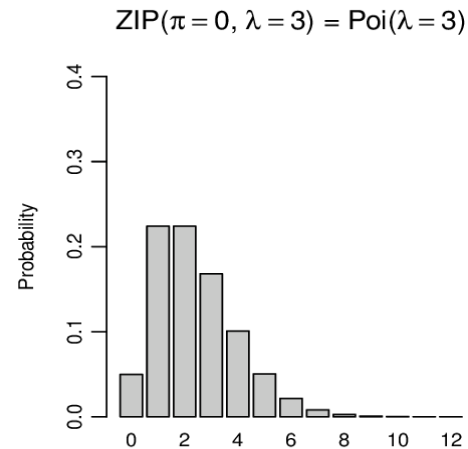
포아송 회귀 모형의 문제점

과대영 문제 (*Excess Zeros*)

포아송 분포에서 예상된 0 발생 횟수보다 실제로 더 많은 0이 발생한 경우



과대영 문제가 발생한 경우

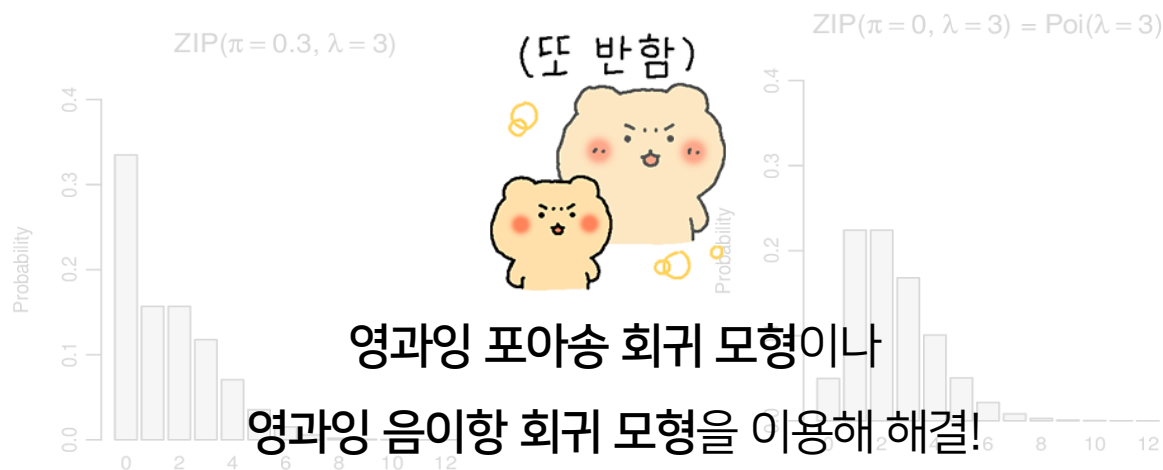


과대영 문제가 발생하지 않은 경우

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

이번 클린업에서는 영과잉 포아송 회귀 모형만을 다룸!

포아송 회귀 모형의 문제점

영과잉 포아송 회귀모형 (*Zero Inflated Poisson Regression, ZIP*)

0의 값만 갖는 점 확률분포와

0 이외의 값을 갖는 포아송 분포가 결합된 혼합분포 구조

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

베르누이 분포 0 이상의 정수 값 0이 발생할 확률 0 이상의 정수 값이 발생할 확률

포아송 회귀 모형의 문제점

영과잉 포아송 분포를 GLM으로 표현

⋮

0이 발생할 확률(ϕ_i)을 로짓 연결함수를 이용하여 표현

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

포아송 분포의 평균(λ)을 로그 연결함수를 이용하여 표현

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



THANK YOU

