

# 범주형자료분석팀

2팀  
정희철  
김민서  
이주형  
심수현

# INDEX

---

1. 일반화 선형 모형
2. 유의성 검정
3. 로지스틱 회귀 모형
4. 다범주 로짓 모형
5. 포아송 회귀 모형

# 1

## 일반화 선형 모형

## 일반화선형모형(GLM)

## 일반화 선형 모형

연속형/범주형 반응변수에 대한 모형들을 모두 포함한 모형의 집합

일반화의 대상

랜덤성분의 분포

랜덤성분의 함수

## 일반화선형모형(GLM)

## GLM의 필요성

반응변수가 범주형 자료이거나 도수자료일 때는 오차항이 정규분포를 따르지 않기 때문에 일반선형회귀를 사용할 수 없음



분할표는 독립성 검정으로 범주형 변수 간의 연관성만을 파악하지만,  
GLM은 변수 간의 연관성 뿐만 아니라 반응변수에 대해 예측 가능

## 일반화선형모형(GLM)

## GLM의 필요성

반응변수가 범주형 자료이거나 도수자료일 때는 오차항이 정규분포를 따르지 않기 때문에 일반선형회귀를 사용할 수 없음



분할표는 독립성 검정으로 범주형 변수 간의 연관성만을 파악하지만,  
GLM은 변수 간의 연관성 뿐만 아니라 반응변수에 대해 예측 가능

## GLM의 구성성분

## Generalized Linear Model

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

## GLM의 구성성분

랜덤성분  
 $\mu$

연결 함수  
 $g(\cdot)$

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

## GLM의 구성성분

## 랜덤성분 Random Component

반응변수 Y에 대해 가정한 분포의 기대값

반응변수	확률분포	표기
이진형	이항분포	$\pi(x)$
연속형	정규분포	$\mu$
도수자료	포아송분포	$\lambda$



## GLM의 구성성분

## 체계적 성분

설명변수  $X$ 들의 선형결합 :  $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$



회귀분석과 마찬가지로,

체계적 성분에는 **교호작용**을 설명하는 항이나

**곡선효과**를 나타내는 항을 넣을 수 있음

교호작용

$$x_i = x_a x_b$$

곡선효과

$$x_i = x_a^2$$

## GLM의 구성성분

## 체계적 성분

설명변수  $X$ 들의 선형결합 :  $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$



회귀분석과 마찬가지로,

체계적 성분에는 **교호작용**을 설명하는 항이나  
**곡선효과**를 나타내는 항을 넣을 수 있음

교호작용

$$x_i = x_a x_b$$

곡선효과

$$x_i = x_a^2$$

## GLM의 구성성분

## 연결함수 Link Function

랜덤성분과 체계적 성분의 범위를 맞춰주는 역할

종류	반응변수	표기
항등 연결 함수 (Identity Link)	연속형	$g(\mu) = \mu$
로그 연결 함수 (Log Link)	도수자료	$g(\mu) = \log(\mu)$
로짓 연결 함수 (Logit Link)	확률, 비율	$g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$

## GLM의 구성성분

# 연결함수 Link Function

★ 연결함수를 사용하는 이유

랜덤성분과 체계적 성분의 범위를 맞춰주는 역할

체계적 성분은 범위에 아무런 제약이 없어

종류	범위	표기
<p>한도 연결 함수 (Identity Link)</p>	$(-\infty, \infty)$ 사이의 값을 갖지만,	$g(\mu) = \mu$
<p>로그 연결 함수 (Log Link)</p>	<p>이름 맞춰줘야 함!</p> <p>도수자료</p>	$g(\mu) = \log(\mu)$
<p>로짓 연결 함수 (Logit Link)</p>	확률, 비율	$g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$

랜덤성분은 분포에 따라 범위에 제약이 있기 때문에

## GLM의 특징

## ① 선형관계식 유지

$g(\mu)$ 와  $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$ 가  
선형관계식을 갖기 때문에 해석 용이

## ② 독립성 가정만 필요

오차항의 독립성만 만족하면  
적용가능

## ③ 다양한 분포 가정 가능

오차항의 독립성만 만족하면 됨  
→ 다양한 분포 가정 가능

## ④ 다양한 반응변수 사용 가능

연결함수: 양변 범위 맞춰주는 역할  
→ 다양한 반응변수에 적용 가능

## GLM의 특징

## ① 선형관계식 유지

$g(\mu)$ 와  $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$ 가  
선형관계식을 갖기 때문에 해석 용이

## ② 독립성 가정만 필요

오차항의 독립성만 만족하면  
적용가능

## ③ 다양한 분포 가정 가능

오차항의 독립성만 만족하면 됨  
→ 다양한 분포 가정 가능

## ④ 다양한 반응변수 사용 가능

연결함수: 양변 범위 맞춰주는 역할  
→ 다양한 반응변수에 적용 가능

## GLM의 특징

## ① 선형관계식 유지

$g(\mu)$ 와  $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$ 가  
선형관계식을 갖기 때문에 해석 용이

## ② 독립성 가정만 필요

오차항의 독립성만 만족하면  
적용가능

## ③ 다양한 분포 가정 가능

오차항의 독립성만 만족하면 됨  
→ 다양한 분포 가정 가능

## ④ 다양한 반응변수 사용 가능

연결함수: 양변 범위 맞춰주는 역할  
→ 다양한 반응변수에 적용 가능

## GLM의 특징

## ① 선형관계식 유지

$g(\mu)$ 와  $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$ 가  
선형관계식을 갖기 때문에 해석 용이

## ② 독립성 가정만 필요

오차항의 독립성만 만족하면  
적용가능

## ③ 다양한 분포 가정 가능

오차항의 독립성만 만족하면 됨  
→ 다양한 분포 가정 가능

## ④ 다양한 반응변수 사용 가능

연결함수: 양변 범위 맞춰주는 역할  
→ 다양한 반응변수에 적용 가능



## GLM의 종류

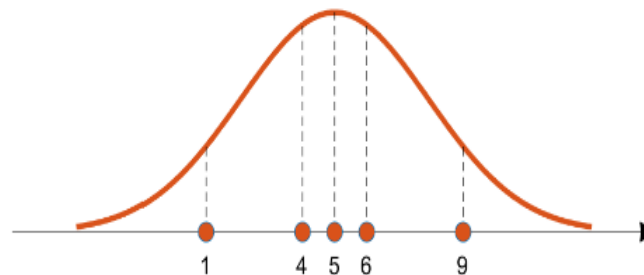
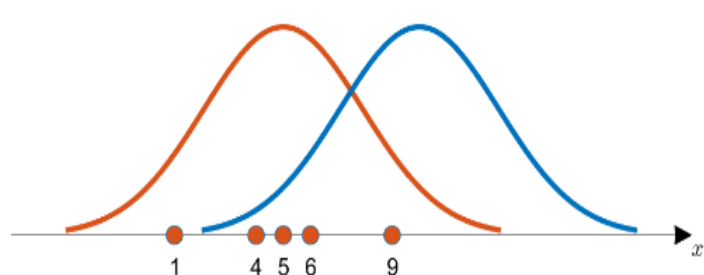
GLM	랜덤성분	연결함수	체계적 성분
로지스틱	이항 자료	로짓	혼합형
기준범주	다항 자료		
누적 로짓			
포아송	도수 자료	로그	
음이항			
율자료 포아송	비율 자료		

## GLM의 모형 적합



MLE

최대가능도 추정법



MLE, 즉 **최대가능도 추정법**을 사용하여 모형적합  
BUT, closed form으로 나오지 않는 경우가 대부분  
→ 알고리즘으로 추정!

MLE에 대한 설명은 회귀팀 1주차 클린업 참고!

# 2

유의성 검정

## 유의성 검정



## 유의성 검정

모형의 모수 추정값이 통계적으로 유의한지에 대한 검정  
축소 모형의 적합도가 좋은지에 대한 검정 또한 가능

## 유의성 검정의 종류

## 왈드 검정 (Wald Test)

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

$$Z \geq |Z_\alpha|, \quad Z^2 \geq \chi_\alpha^2$$

검정통계량 계산 시 계수값과 표준오차만 사용하여 간단함

## 유의성 검정의 종류

## 가능도비 검정

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

$$G^2 \geq x_{\alpha, df}^2$$

귀무가설 하에서 계산되는 로그가능도 함수  $l_0$ 과  
MLE에 의해 계산되는 로그가능도 함수  $l_1$ 의 차이

## 검정 Flow

$l_0$ 와 함수  $l_1$ 의 차이가 **크다**



검정통계량이 **크다**



P-value 값이 **작다**



귀무가설 **기각**,  
적어도 하나의  $\beta$ 는 **0이 아님**



모형의 모수 **추정값 유의**

## 유의성 검정의 종류

## 가능도비 검정

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

$$G^2 \geq x_{\alpha, df}^2$$

**자유도** = 두 모형의 차원의 차이  
 =  $H_0$ 과  $H_1$ 의 모수의 개수 차이

## 검정 Flow

$l_0$ 와 함수  $l_1$ 의 차이가 크다



검정통계량이 크다



P-value 값이 작다



귀무가설 기각,  
 적어도 하나의  $\beta$ 는 0이 아님



모형의 모수 추정값 유의

## 유의성 검정의 종류

## ★가능도비 검정

검정 Flow

가능도비 검정

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

귀무가설 하 ( $\beta = 0$ ) & 전체공간 하 ( $\hat{\beta}$ : MLE)  
 가능도 함수에 대한 정보 모두 이용  
 $G^2 \geq x_{\alpha, df}^2$

 $l_0$ 와 함수  $l_1$ 의 차이가 크다

검정통계량이 크다



P-value 값이 작다



: 왈드 검정에 비해 검정력과 신뢰도 ↑

귀무가설 기각,

적어도 하나의  $\beta$ 는 0이 아님

모형의 모수 추정값 유의

자유도

= 두 모형의 차원의 차이  
 MLE에 의해 계산 =  $H_0$ 과  $H_1$ 의 모수의 개수 차이



## 이탈도

## 관심모형 (M)

우리가 관심있는 모형, 유의성 검정을 진행할 모형

$$\text{시험점수}(Y) = \beta_0 + \beta_1 \times \text{공부시간}(x_1) + \beta_2 \times \text{시험난이도}(x_2)$$

## 포화모형 (S)

주어진 관측값들에 대해 완벽하게 자료에 적합하는 모형

$$\begin{aligned} \text{시험점수}(Y) = & \beta_0 + \beta_1 \times \text{공부시간}(x_1) + \beta_2 \times \text{시험난이도}(x_2) + \\ & \beta_3 \times \text{통학시간}(x_3) + \beta_4 \times \text{수면시간}(x_4) \end{aligned}$$

이탈도

이탈도

포화모형과 관심모형을 비교하기 위한 가능도비 통계량

$$-2 \log \left( \frac{L_M}{L_S} \right) = -2(l_M - l_S)$$

$L_M$  : 관심모형에서 얻은 로그 가능도 함수의 최댓값

$L_S$  : 포화모형에서 얻은 로그 가능도 함수의 최댓값

→ 두 모형의 **가능도 함수의 최댓값**을 이용!

## 이탈도



이탈도는 포화모형에는 있지만 **관심모형에는 없는**  
 포화모형과 관심모형을 비교하기 위한 **가능도비 통계량**  
**계수들이 0인지의 여부를 확인하는 것**

$$-2 \log \left( \frac{L_M}{L_S} \right) = -2(l_M - l_S)$$



$L_M$ : 관심모형에서 얻은 로그 가능도 함수의 최댓값

$L_S$ : 포화모형에서 얻은 로그 가능도 함수의 최댓값



**관심모형은 포화모형에 **내포된(nested)** 관계여야 함!**

→ 두 모형의 **가능도 함수의 최댓값**을 이용!

## 이탈도

## 이탈도

포화모형과 관심모형을  
비교하기 위한 가능도비 통계량

$$-2 \log \left( \frac{L_M}{L_S} \right) = -2(l_M - l_S)$$

## 검정 Flow

두 가능도 함수의 최댓값 차이가 **크다**



이탈도가 **크다**



P-value 값이 **작다**



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



관심모형 M **사용 불가능!**

## 이탈도와 가능도비 검정의 관계

$$M_0 \text{의 가능도비} - M_1 \text{의 가능도비} = -2(l_0 - l_S) - \{-2(l_1 - l_S)\} = -2(l_0 - l_1)$$

$M_0$  : 단순한 형태의 관심모형,  $M_1$  : 복잡한 형태의 관심모형

$S$  : 두 모형을 모두 포함하는 포화모형



두 모형 간 로그가능도비의 차 = 이탈도 검정 통계량

이탈도 활용 → 모형  $M_0$ 은 모형  $M_1$ 에 **내포된 모형**이어야 함

## 이탈도와 가능도비 검정의 관계

$$M_0 \text{의 가능도비} - M_1 \text{의 가능도비} = -2(l_0 - l_S) - \{-2(l_1 - l_S)\} = -2(l_0 - l_1)$$

$M_0$  : 단순한 형태의 관심모형,  $M_1$  : 복잡한 형태의 관심모형

$S$  : 두 모형을 모두 포함하는 포화모형



여러 모형을 비교하고 싶지만 내포되지 않은 경우

→ **AIC, BIC** 등을 활용해서 모형 비교

이탈도 활용 → 모형  $M_0$ 과 모형  $M_1$ 에 대한 모형이어야 함

## 이탈도와 가능도비 검정의 관계

$M_0$ 의 가능도비 -  $M_1$ 의 가능도비

$$\begin{aligned} & -2(l_0 - l_S) - \{-2(l_1 - l_S)\} \\ & = -2(l_0 - l_1) \end{aligned}$$

## 검정 Flow

관심모형 간 이탈도의 차이가 **작다**



가능도비 검정 통계량이 **작다**



P-value 값이 **크다**



귀무가설 **기각 못함**,  $M_0$ 에 포함되지  
않는 모수들이 **모두 0이다**



간단한 관심모형  $M_0$ 이 **더 적합**

# 3

로지스틱 회귀 모형



## 로지스틱 회귀 모형

## 로지스틱 회귀 모형

반응변수가 **이항자료**일 때 사용하는 회귀 모형

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



범위가 같지 않음!

## 로지스틱 회귀 모형

## 로지스틱 회귀 모형

반응변수가 **이항자료**일 때 사용하는 회귀 모형

① 좌변을 오즈로 설정

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

좌변 범위:  $(0, \infty)$

우변 범위:  $(-\infty, \infty)$

여전히  
범위가 같지 않음!

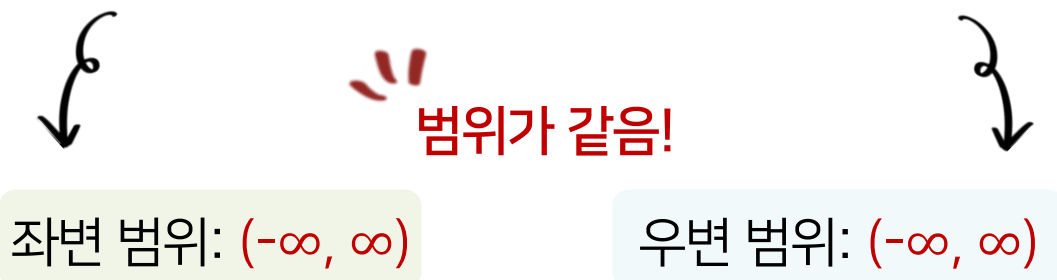
## 로지스틱 회귀 모형

## 로지스틱 회귀 모형

반응변수가 **이항자료**일 때 사용하는 회귀 모형

## ② 오즈에 로그 취하기

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



## 로지스틱 회귀 모형

## 로지스틱 회귀 모형

반응변수가 **이항자료**일 때 사용하는 회귀 모형

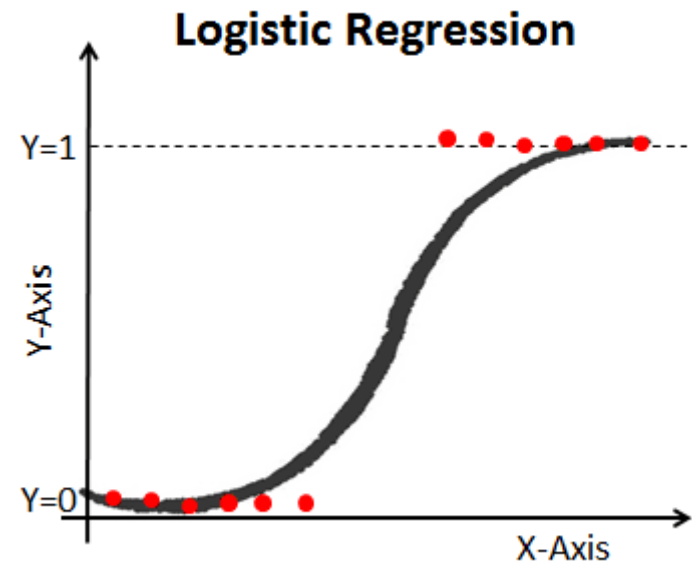
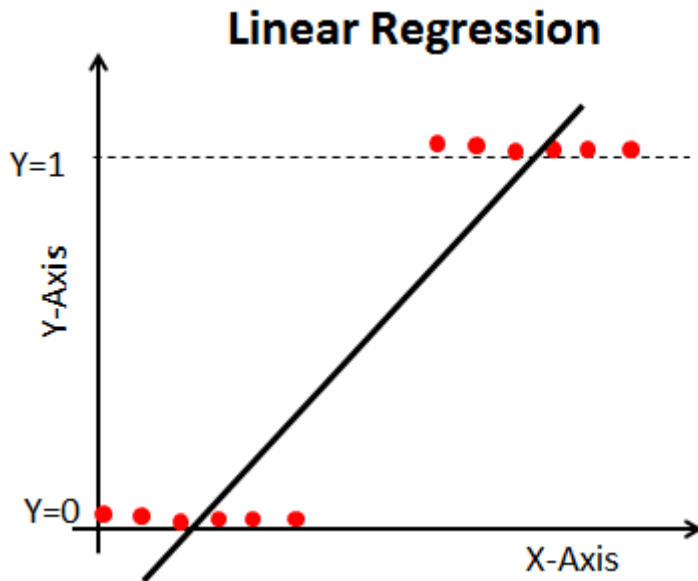
## ② 오즈에 로그 취하기

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

로지스틱 연결 함수로 좌변과 우변의 범위 맞춤

좌변 범위:  $(-\infty, \infty)$  → 로지스틱 회귀 모형 우변 범위:  $(-\infty, \infty)$

## 로지스틱 회귀 모형



시그모이드 형태

확률을 따르는 S자 곡선의 함수  
 $\pi(x)$ 와  $x$ 의 비선형 관계 나타냄

## 로지스틱 회귀 모형

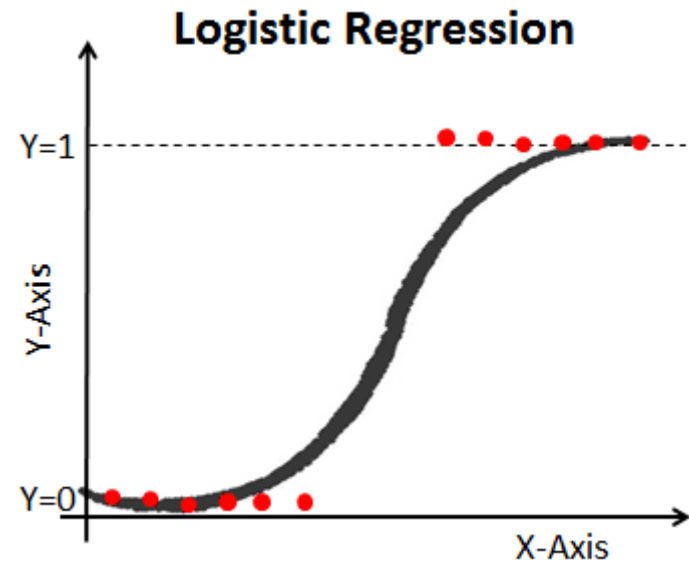
## 특징

## ① 양변의 범위 일치시킴

$$0 \leq \pi(x) \leq 1, 0 \leq 1 - \pi(x) \leq 1$$

$$0 \leq \frac{\pi(x)}{1 - \pi(x)} \leq \infty$$

$$-\infty \leq \log \frac{\pi(x)}{1 - \pi(x)} \leq \infty$$



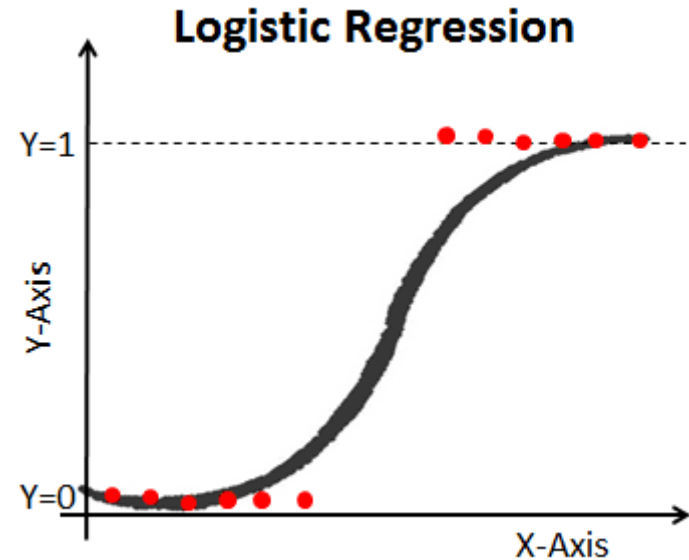
## 로지스틱 회귀 모형

## 특징

## ② 가정으로부터 자유로움

정규성, 등분산성, 선형성 가정 필요 X

독립성 가정만 만족하면 됨



## 로지스틱 회귀 모형의 해석

회귀계수  $\beta$ 의 해석

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

로지스틱 회귀 모형의 접선의 기울기

회귀계수  $\beta$ 가 양수이면 **상향 곡선**, 음수면 **하향 곡선**

$|\beta|$  증가  $\rightarrow$  변화율 증가



## 로지스틱 회귀 모형의 해석

## 확률에 기초한 해석

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

$x$ 값을 대입해 특정 범주에 속할 확률을 알 수 있음

확률값이 **cutoff point**보다 크면 **Y=1**, 작으면 **Y=0**으로 예측

## 로지스틱 회귀 모형의 해석

## 확률에 기초한 해석

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

$x$ 값을 대입해 특정 범주에 속할 확률을 알 수 있음

확률값이 **cutoff point**보다 크면  $Y=1$ , 작으면  $Y=0$ 으로 예측



일반적으로 0.5 사용

3주차 클린업에서 다룰 예정!

## 로지스틱 회귀 모형의 해석

## 오즈비를 이용한 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 모형에 각각  $x$ 와  $x + 1$ 을 대입한 후 빼주기

## 로지스틱 회귀 모형의 해석

## 오즈비를 이용한 해석

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



다른 설명변수가 모두 고정되어 있을 때,



$x$ 가 한 단위 증가하면  $Y = 1$ 일 오즈가  $e^{\beta}$ 배 만큼 증가

## 로지스틱 회귀 모형의 해석

Ex) 학점에 따른 합격유무에 관한 로지스틱 회귀 모형

## 오즈비를 이용한 해석

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

$Y = 1$  : 합격,  $Y = 0$  : 불합격,  $x$  = 학점



$x$ 가 한 단위 증가하면

$Y = 1$ (합격)일 오즈가  $e^3$ 배, 즉 20.086배 만큼 증가

## 로지스틱 회귀 모형의 해석

Ex) 학점에 따른 합격유무에 관한 로지스틱 회귀 모형

## 확률을 이용한 해석

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



학점이 2.5에서 4.5로 증가할 때

$Y = 1$ (합격)일 확률이 0.00001만큼 증가

# 4

다범주 로짓 모형

## 다범주 로짓 모형

## 다범주 로짓 모형

랜덤성분이 **다항분포**를 따르고 연결함수가 **로짓 연결함수**인 GLM

반응변수의 범주가

3개 이상



자료가 명목형인지 순서형인지 구분해야 함



자료의 종류에 따라

적용하는 모델이 달라지기 때문 !




## 4 다범주 로짓 모형



다범주 로짓 모형

# 로지스틱 회귀 모형 vs 다범주 로짓모형

다범주 로짓 모형

랜덤성분이	로지스틱 회귀모형	다범주 로짓모형
공통점	연결함수 = 로짓 연결함수 자료가 명목형인지 순서형인지 구분해야 함	
3개 이상 차이점	반응변수  = 성공/실패의 이항분포	↓ 자료의 종류에 따라 반응변수 = 다항분포 적합한 모델이 달라지기 때문!

## 다범주 로짓 모형

## 다범주 로짓 모형

랜덤성분이 **다항분포**를 따르고 연결함수가 **로짓 연결함수**인 GLM

반응변수의 범주가

3개 이상



자료가 명목형인지 순서형인지 구분해야 함



**자료의 종류에 따라**

**적용하는 모델이 달라지기 때문 !**

## 기준 범주 로짓 모형

## 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

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

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

마지막 범주 J가 기준이 될 때의 기준범주 로짓

-> 기준범주 J에 비해 **j범주일 확률의 오즈**

## 기준 범주 로짓 모형

### 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

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

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



범주 J가 기준이 될 때의 기준범주 로짓

-> 기준범주 J에 비해 **j범주일 확률의 오즈**

## 기준 범주 로짓 모형

### 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

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

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

$J$ : 기준 범주

$j$ : 범주에 대한 첨자

$A \sim K$ : 설명 변수  $x$ 에 대한 첨자

## 기준 범주 로짓 모형

### 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

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

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



**J-1**개의 로짓방정식으로 구성 -> 각 식마다 다른 모수들을 가짐

**J=2**인 경우 **이항 반응변수에 대한 로지스틱 회귀 !**

## 기준 범주 로짓 모형

### 기준 범주 로짓 모형

반응변수가  $J$ 개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
 기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$  에서  $j = 1, \dots, J - 1$ 이기 때문에

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

**J-1개의 로짓방정식**으로 구성됨



**J-1개의 로짓방정식**으로 구성 -> 각 식마다 다른 모수들을 가짐

**J=2인 경우 이항 반응변수에 대한 로지스틱 회귀 !**

## 기준 범주 로짓 모형

### 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

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

$$\hookrightarrow \pi_j = \frac{e^{\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K}}{\sum_{i=1}^J e^{\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K}} \quad j = 1, \dots, J-1$$

✓ 모형 공식을 변형하여 **j범주에 속할 확률**을 구할 수 있음!



## 기준 범주 로짓 모형

## 기준 범주 로짓 모형

반응변수가 J개 범주를 가지는 **명목형 변수**일 때 사용하는 모형  
기준 범주를 선택한 후 기준범주와 타 범주를 짝지어 로짓 정의

## 특징

- ① 명목형 범주를 다루기 때문에 순서 고려X
- ② **오즈와 기준 범주**를 사용해 해석 가능

## 기준 범주 로짓 모형 - 해석

① 기준 범주에 비해 j범주일 로그 오즈를 보고 해석할 경우

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

$$\log \left( \frac{\pi_j(x+1)}{\pi_J(x+1)} \right) - \log \left( \frac{\pi_j(x)}{\pi_J(x)} \right) = [\alpha_j + \beta(x+1)] - [\alpha_j + \beta(x)]$$

$$\log \left( \frac{\pi_j(x+1)/\pi_J(x+1)}{\pi_j(x)/\pi_J(x)} \right) = \beta$$

$$\frac{\pi_j(x+1)/\pi_J(x+1)}{\pi_j(x)/\pi_J(x)} = e^\beta$$

기준 범주 로짓 모형에 각각  $x$ 와  $x+1$ 을 대입한 후 빼주기

## 기준 범주 로짓 모형 - 해석

① 기준 범주에 비해 j범주일 로그 오즈를 보고 해석할 경우

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

$$\frac{\pi_j(x+1)/\pi_J(x+1)}{\pi_j(x)/\pi_J(x)} = e^\beta$$



X가  $x$ 일때보다  $x+1$ 일 때의 **오즈가  $e^\beta$ 배** 높음

## 기준 범주 로짓 모형 - 해석

① 기준 범주에 비해 j범주일 로그 오즈를 보고 해석할 경우

$$\frac{\pi_j(x+1)/\pi_J(x+1)}{\pi_j(x)/\pi_J(x)} = e^\beta \quad j = 1, \dots, J-1$$

다른 설명 변수가 고정되어 있을 때

$x$ 가 한 단위 증가하면 J범주 대신 j범주일 오즈가  $e^\beta$ 배 증가

## 기준 범주 로짓 모형 - 해석

② 기준 범주가 아닌 또 다른 범주끼리의 관계를 해석할 경우

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

다른 설명 변수가 고정되어 있을 때

$x$ 가 한 단위 증가하면 1범주 대신 2범주일 오즈가  $e^{\beta_2 - \beta_1}$ 배 증가

## 순서형 다범주 로짓 모형

## 순서형 다범주 로짓 모형

순서형 반응변수에 대한 로짓모형

순서 정보를 고려하여 기준범주를 정하고 범주끼리 비교

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

## 순서형 다범주 로짓 모형

## 순서형 다범주 로짓 모형

순서형 반응변수에 대한 로짓모형

순서 정보를 고려하여 기준범주를 정하고 범주끼리 비교



순서대로 정렬 후 두 덩어리로 나누는 **collapse 과정** 필요

-> collapse하는 기준인 **cut point**에 따라 모형 결정

## 순서형 다범주 로짓 모형

## 순서형 다범주 로짓 모형

순서형 반응변수에 대한 로짓모형

순서 정보를 고려하여 기준범주를 정하고 범주끼리 비교

이웃범주 로짓 모형

소형	중형	대형	초 대형
소형	중형	대형	초 대형
소형	중형	대형	초 대형

연속비 로짓 모형

소형	중형	대형	초 대형
소형	중형	대형	초 대형
소형	중형	대형	초 대형

누적 로짓 모형

소형	중형	대형	초 대형
소형	중형	대형	초 대형
소형	중형	대형	초 대형



순서형 다범주 로짓 모형

## 순서형 다범주 로짓 모형

순서형 반응변수에 대한 로짓모형

순서 정보를 고려하여 기준범주를 정하고 범주끼리 비교

이웃범주 로짓 모형

연속비 로짓 모형

누적 로짓 모형

소형	중형	대형	초대형
소형	중형	대형	초대형
중형	대형	초대형	

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

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

다른 두 모형과 달리 **전체 범주를 모두 사용!**

이번 클린업에서 집중적으로 다룰 예정



## 누적 로짓 모형-모형

## 누적 로짓 모형

누적확률에 로짓 연결함수를 씌운 형태

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

## 누적확률

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

-> 첫번째 범주부터 j번째 범주까지의 누적확률

-> j번째 범주 아래의 확률을 모두 더한 것

## 누적 로짓 모형-모형

## 누적 로짓 모형

누적확률에 로짓 연결함수를 씌운 형태

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



누적확률

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

→ 첫번째 범주부터 j번째 범주까지의 누적확률

→ j번째 범주까지의 확률을 모두 더한 것

## 누적 로짓 모형-모형

## 누적 로짓 모형

누적확률에 로짓 연결함수를 씌운 형태

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

## Step 1. 누적확률을 오즈의 형태로 변환

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

$$\rightarrow \log \left( \frac{P(Y \leq j|X=x)}{1-P(Y \leq j|X=x)} \right)$$

## 누적 로짓 모형 - 모형

## 누적 로짓 모형

누적확률에 로짓 연결함수를 씌운 형태

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

## Step 2. 변환한 식에 로그를 씌우기

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

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

## 누적 로짓 모형 - 모형

## 누적 로짓 모형

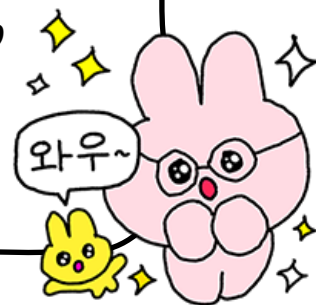
누적확률에 로짓 연결함수를 씌운 형태

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

Step 3. 누적 로짓 모형의 최종 형태

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

$j = 1, \dots, J$



## 기준범주 로짓 모형 vs 누적 로짓 모형

	기준범주 로짓 모형	누적 로짓 모형
공통점	기준점을 두고 이분화된 두 범위의 확률을 비교하는 방식 → J-1개의 로짓 방정식으로 구성	
차이점	$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$ → $\alpha$ 와 회귀계수에 모두 첨자 j	$\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$ → 회귀계수에선 첨자 j 사라짐

## 4

## 다범주 로짓 모형

## 기준범주 로짓 모형 vs 누적 로짓 모형

	기준범주 로짓 모형	누적 로짓 모형
공통점	기준점을 두고 이분화된 두 범위의 확률을 비교하는 방식 → J-1개의 로짓 방정식으로 구성	
차이점	$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$ → $\alpha$ 와 회귀계수에 모두 첨자 j	$\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$ → 회귀계수에선 첨자 j 사라짐



## 기준범주 로짓 모형 vs 누적 로짓 모형

	기준범주 로짓 모형	누적 로짓 모형
공통점	<p>J-1개의 로짓 방정식에 대한 <math>\beta</math>의 효과가 모두 동일하다고 가정하기 때문 (비례오즈 가정)</p> <p>→ J-1개의 로짓 방정식으로 구성</p>	
차이점	$\alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_K$ <p>→ <math>\alpha</math>와 회귀계수에 모두 첨자 j</p>	$\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$ <p>→ 회귀계수에선 첨자 j 사라짐</p>

## 4

## 다범주 로짓 모형

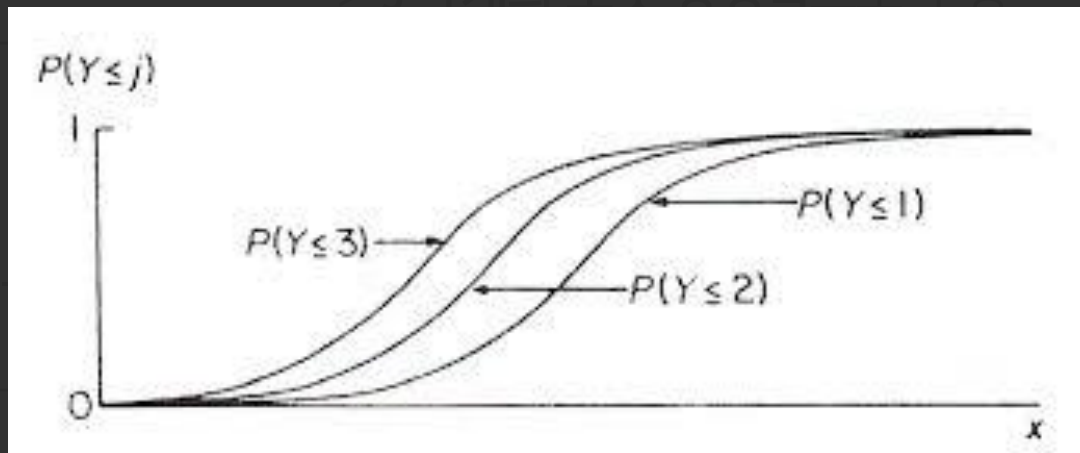
기준범주 로짓 모형 vs ~~비기준범주~~ 로짓 모형

# 비례오즈 가정

J-1개의 로짓 방정식에 대해  $\alpha$ 만 방정식에 따라 변할 뿐

모든 로짓 방정식은 동일한  $\beta$ 값을 지님

= 절편,  $\alpha$ 값만 변화, 기울기  $\beta$ 값은 변하지 않음!  
 J-1개의 로짓 방정식에 대한  $\beta$ 의 효과가 모두 동일하다고 가정하기 때문 (비례오즈 가정)



$+ \beta_p x_p$

자 j 사라짐

## 누적 로짓 모형 - 해석

기준 범주 로짓 모형처럼 **오즈를 이용**하여 해석



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

다른 설명 변수가 고정되어 있을 때  **$x$ 가 한 단위 증가**하면

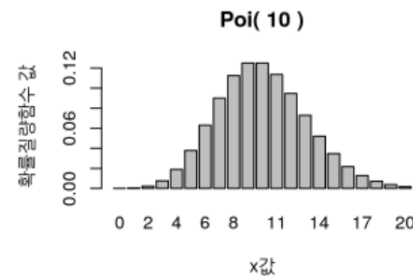
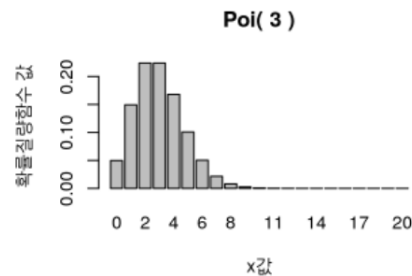
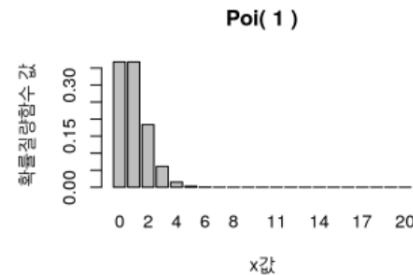
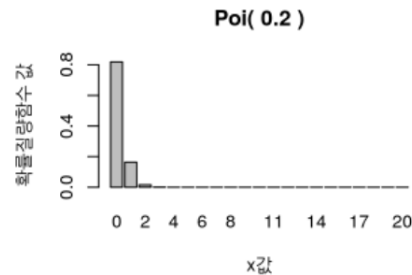
**$Y > j$ 에 비해  $Y \leq j$  일 오즈가  $e^\beta$ 만큼 증가**

# 5

포아송 회귀 모형

## 포아송 분포

모수  $\lambda$  값(=평균)이 **작을수록 오른쪽**으로 치우친 분포  
 작은 건수로 많은 관측치가 몰리는 현상 발생

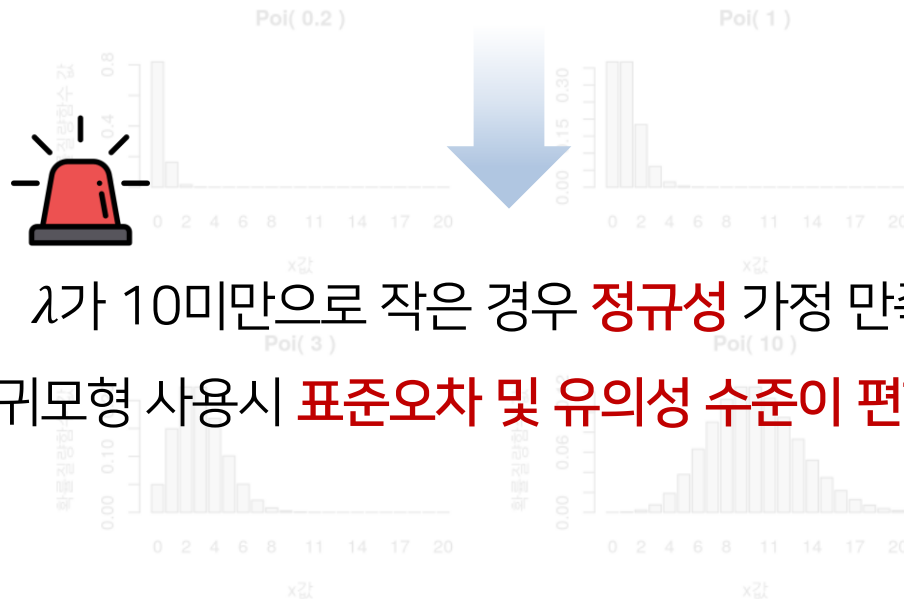


## 5

## 포아송 회귀 모형

## 포아송 분포

모수  $\lambda$  값(=평균)이 **작을수록 오른쪽**으로 치우친 분포  
작은 건수로 많은 관측치가 몰리는 현상 발생



$\lambda$ 가 10미만으로 작은 경우 **정규성** 가정 만족 X

→ 일반 선형회귀모형 사용시 **표준오차 및 유의성 수준이 편향**되는 문제 발생!

## 포아송 회귀 모형

반응변수 Y가 **도수자료**인 경우 사용하는 회귀모형  
랜덤성분이 **포아송 분포**를 따르고 연결함수가 **GLM**인 회귀모형

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



음수가 아닌 정수값을 가지는 도수자료  $\mu$ 를  
체계적 성분의 범위와 맞춰주기 위해 연결함수로 **로그** 사용

## 포아송 회귀 모형

반응변수 Y가 **도수자료**인 경우 사용하는 회귀모형  
랜덤성분이 **포아송 분포**를 따르고 연결함수가 **GLM**인 회귀모형

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



음수가 아닌 정수값을 가지는 도수자료  $\mu$ 를  
체계적 성분의 범위와 맞춰주기 위해 연결함수로 **로그** 사용



## 포아송 회귀 모형 - 해석

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

↓  $\mu$ 에 관한 식으로 정리

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

## ① 도수를 통한 해석

추정된 회귀 계수를 대입하면

$\mu$ 에 대한 예측값(=기대도수)을 얻을 수 있음 !

## 포아송 회귀 모형 - 해석

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



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

## ② 차이를 통한 해석

추정된 회귀 계수를 대입하면

$\mu$ 에 대한 예측값(=기대도수)을 얻을 수 있음!

## 포아송 회귀 모형 - 해석

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



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

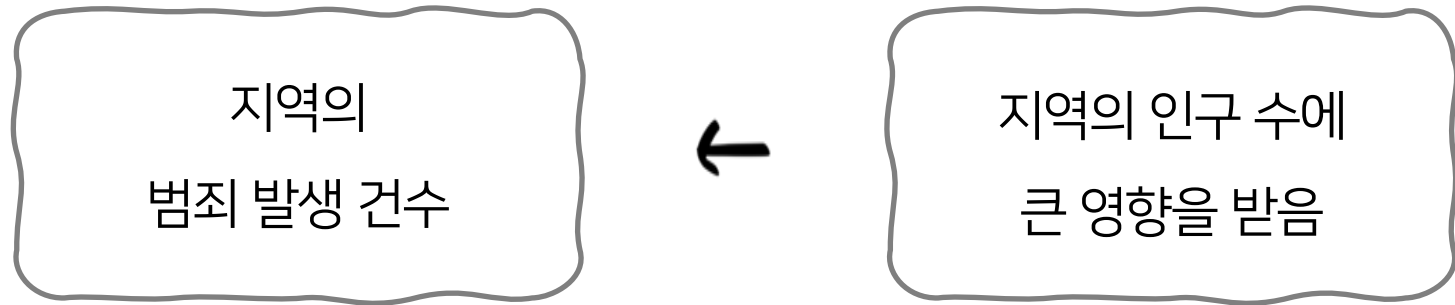
## ② 차이를 통한 해석



다른 설명 변수가 고정되어 있을 때,  
 $x$ 가 한 단위 증가하면 **기대도수  $\mu$ 가  $e^{\beta}$  배만큼 증가!**

추정된 회귀 계수를 대입하면  
 $\mu$ 에 대한 예측값 (= 기대도수)을 얻을 수 있음!

## 율자료 포아송 회귀 모형



특정 사건이 다른 크기의 지표(ex : 인구, 시간)에 걸쳐 발생

-> **비율 자료**를 사용해야 정확한 크기 판단 가능

## 율자료 포아송 회귀 모형

반응변수 Y가 **비율자료**인 경우 사용하는 회귀모형  
랜덤성분이 **포아송 분포**를 따르고 연결함수가 **GLM**인 회귀모형

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

$t$  : **지표값**. 수정항(offset)으로 지칭.

-> 비율을 구할 때 **분모에 들어가는 값!**

ex) 범죄 발생 비율 -> 지표값 = 그 지역 인구의 모집단

## 율자료 포아송 회귀 모형

반응변수 Y가 **비율자료**인 경우 사용하는 회귀모형  
랜덤성분이 **포아송 분포**를 따르고 연결함수가 **GLM**인 회귀모형

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



$t$  : **지표값**, 수정항(offset)으로 지칭.

→ 비율을 구할 때 **분모에 들어가는 값!**

ex) 범죄 발생 비율 -> 지표값 = 그 지역 인구의 모집단

## 율자료 포아송 회귀 모형

반응변수 Y가 **비율자료**인 경우 사용하는 회귀모형  
랜덤성분이 **포아송 분포**를 따르고 연결함수가 **GLM**인 회귀모형

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



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



포아송 회귀 모형처럼 **도수**( $\mu$ )에 대한 식으로 표현 가능 !

율자료 포아송 회귀 모형 : 해석

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



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



다른 설명 변수가 고정되어 있을 때,  
 $x$ 가 한 단위 증가하면 **기대도수  $\mu$ 가  $e^{\beta}$  배만큼 증가!**



포아송 회귀 모형의 문제점 : 과대산포 문제

포아송 분포

평균과 분산이 같다는 특징을 가짐



등산포 가정

랜덤 성분이 포아송 분포를 따른다고 가정할 때  
반응변수인 도수 자료의 평균과 분산이 같다는 가정

포아송 회귀 모형의 문제점 : 과대산포 문제



**등산포 가정을 만족하는 데이터가 매우 적다!**

포아송 분포

일반적인 데이터는 분산이 평균보다 크게 나타남

→ 과대산포 문제



등산포 가정

과대산포 문제를 무시하고 포아송 모형을 적합시

회귀 계수 추정량의 표준오차가 편향되어 작아짐

반응변수인 도수 자료의 평균과 분산이 같다는 가정

지대장나!



음이항 회귀 모형

## 음이항 회귀 모형

음이항 랜덤성분과 로그 연결함수로 구성된 GLM

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

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu + D\mu^2$$



$D$  : 산포모수. 음이항 분포에서 분산이 평균과 큰 값을 갖도록 만드는 요소



음이항 분포의 분산 : 포아송 분포와 달리 분산에  $D\mu^2$  가 더해진 형태 !

음이항 회귀 모형

## 음이항 회귀 모형

음이항 랜덤성분과 로그 연결함수로 구성된 GLM

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

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu + D\mu^2$$



$D$  : 산포모수. 음이항 분포에서 분산이 평균과 큰 값을 갖도록 만드는 요소



음이항 분포의 분산 : 포아송 분포와 달리 분산에  $D\mu^2$  가 더해진 형태 !

음이항 회귀 모형

## 음이항 회귀 모형

음이항 랜덤성분과 로그 연결함수로 구성된 GLM

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



포아송 분포가 가지는 **등산포 가정을 완화**하기 위해서

랜덤성분으로 음이항 분포를 사용해 **과대산포 문제 해결** 가능 !

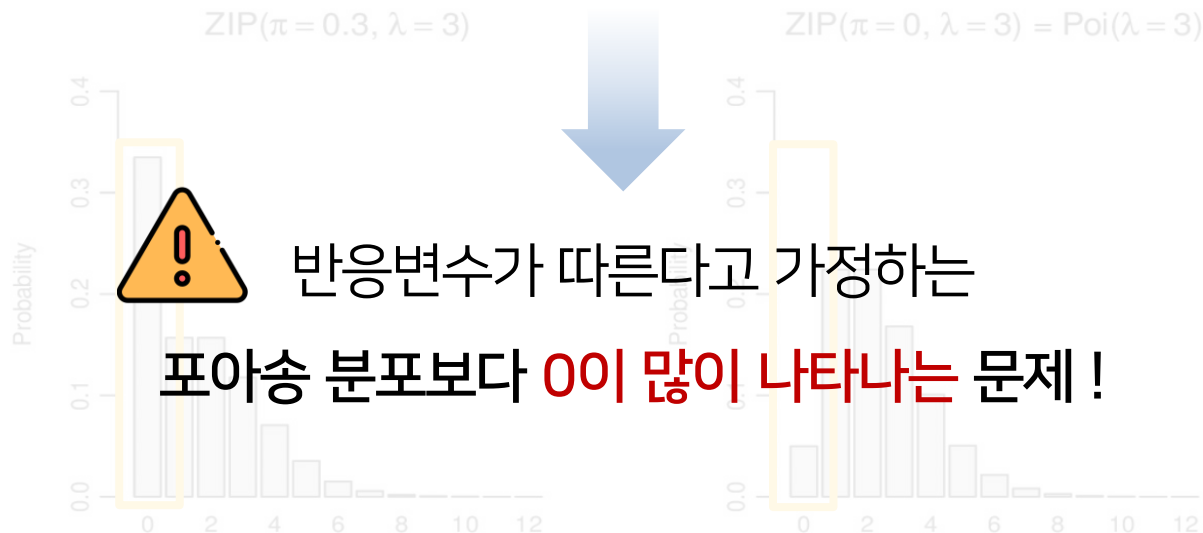
$D$  : 산포모수. 음이항 분포에서 분산이 평균과 큰 값을 갖도록 만드는 요소

음이항 분포의 분산 : 포아송 분포와 달리 분산에  $D\mu^2$  가 더해진 형태 !

포아송 회귀 모형의 문제점 : 과대영 문제

## 과대영 문제

특정 평균을 갖는 포아송 분포에서 나타나는 0보다  
**표본 도수 자료가 더 많은 0을 갖는 경우**

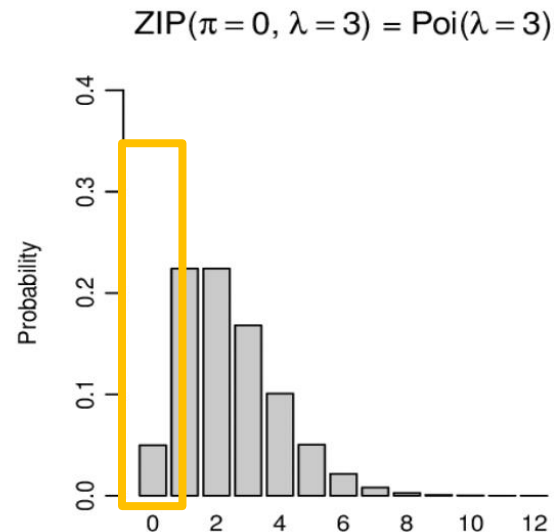
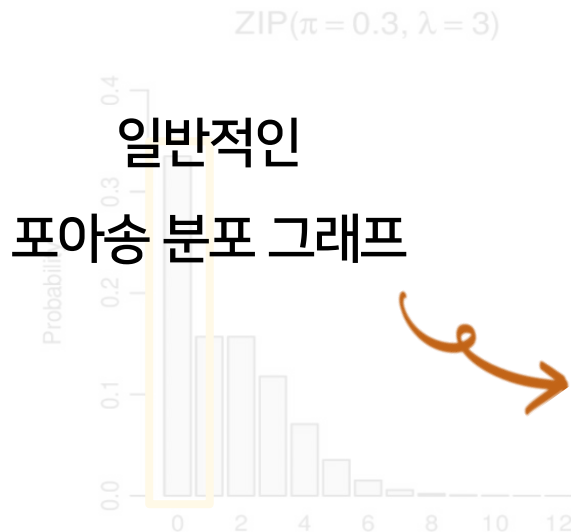


포아송 회귀 모형의 문제점 : 과대영 문제

## 과대영 문제

특정 평균을 갖는 포아송 분포에서 나타나는 0보다

**표본 도수 자료가 더 많은 0을 갖는 경우**



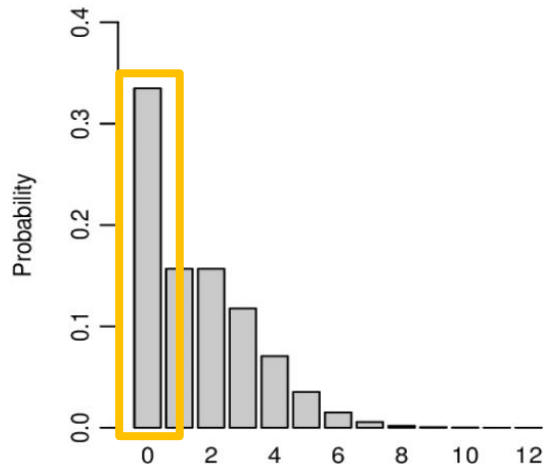
포아송 회귀 모형의 문제점 : 과대영 문제

## 과대영 문제

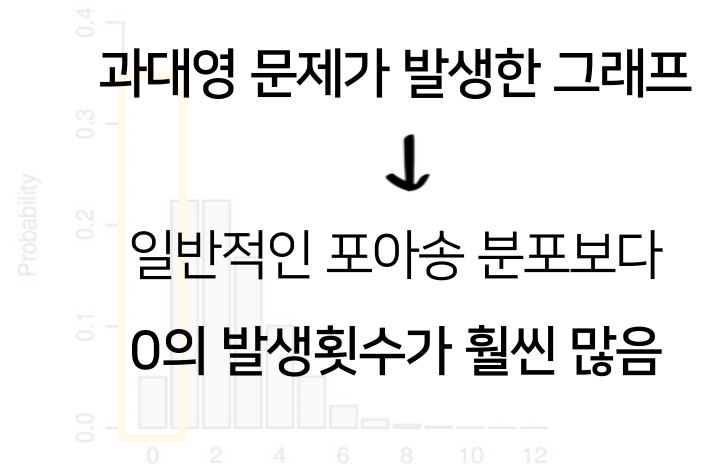
특정 평균을 갖는 포아송 분포에서 나타나는 0보다

**표본 도수 자료가 더 많은 0을 갖는 경우**

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



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





영과잉 포아송 분포

영과잉 포아송 분포

0의 값만을 갖는 점 확률 분포와 포아송 분포의 혼합분포 구조

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



$Y \sim \text{Bern}(\phi_i)$ , 베르누이 확률분포를 따름

영과잉 포아송 분포

영과잉 포아송 분포

0의 값만을 갖는 점 확률 분포와 포아송 분포의 혼합분포 구조

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

Y가 0의 값을 가질 확률

Y가 0이외의 값을 가질 확률



영과잉 포아송 분포

영과잉 포아송 분포

0의 값만을 갖는 **점 확률 분포**와 **포아송 분포**의 **혼합분포 구조**



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

영과잉 포아송 회귀 모형, 영과잉 음이항 회귀 모형

으로 **과대영 문제 해결** 가능 !

영과잉 포아송 회귀모형

영과잉 포아송 회귀모형

영과잉 포아송 분포를 사용하여 만든 GLM

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

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

영과잉 포아송 회귀모형

## 영과잉 포아송 회귀모형

영과잉 포아송 분포를 사용하여 만든 GLM



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



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

베르누이 분포의 성공확률( $\phi_i$ )에 대한 로짓연결 함수

영과잉 포아송 회귀모형

## 영과잉 포아송 회귀모형

영과잉 포아송 분포를 사용하여 만든 GLM

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

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



포아송 분포의 평균( $\lambda$ )에 대한 로그 연결함수

# 다음 주 예고

---

혼동행렬

ROC 곡선

샘플링

인코딩

대응작 검정 방법