범주형자료분석팀

2팀

김정훈
김상태
김민정
박정현
이윤희

INDEX

1 GLM(Generalized Linear Model)

GLM이란?

[1] 종속변수가 범주형인 경우

예) binary변수 (합격 / 불합격)

다항변수 (공화당 / 민주당 / 무소속)

[2] 종속변수가 count인 경우

예) 하루에 마시는 물이 몇 잔인지

GLM

랜덤성분이 정규분포가 아닌

다른 분포를 갖게 일반화

평균의 함수를 연결함수로 모형화

GLM(Generalized Linear Model)

②

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

GLM 의 구성 성분

연결 함수

[1] 정의

랜덤성분의 기댓값과 체계적 성분을 연결

GLM 구성 성분

랜덤 성분	연결 함수	선형예측식 (체계적 성분 : x_k)
μ	$g()$	$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

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

[2] 필요성

종속변수의 평균을 선형개념으로 변환하는 데 필요

→ 좌변과 우변의 범위가 동일해짐

연결함수의 종류

로짓함수

반응변수 : Binomial한 경우에 사용

모양 : $g(\mu) = \log\left[\frac{\mu}{1-\mu}\right] = logit\ \mu$

설명 : 오즈에 로그를 씌운 것

로그함수

반응변수 : 포아송 분포 or 음이항 분포

거듭하는 횟수를 나타내는 도수 분포

모양 : $g(\mu) = \log[\mu]$

항등함수

반응변수 : 연속형 반응변수

모양 : $g(\mu) = \mu$

1 확률과 가능도

확률변수의 종류에 따른 확률과 가능도

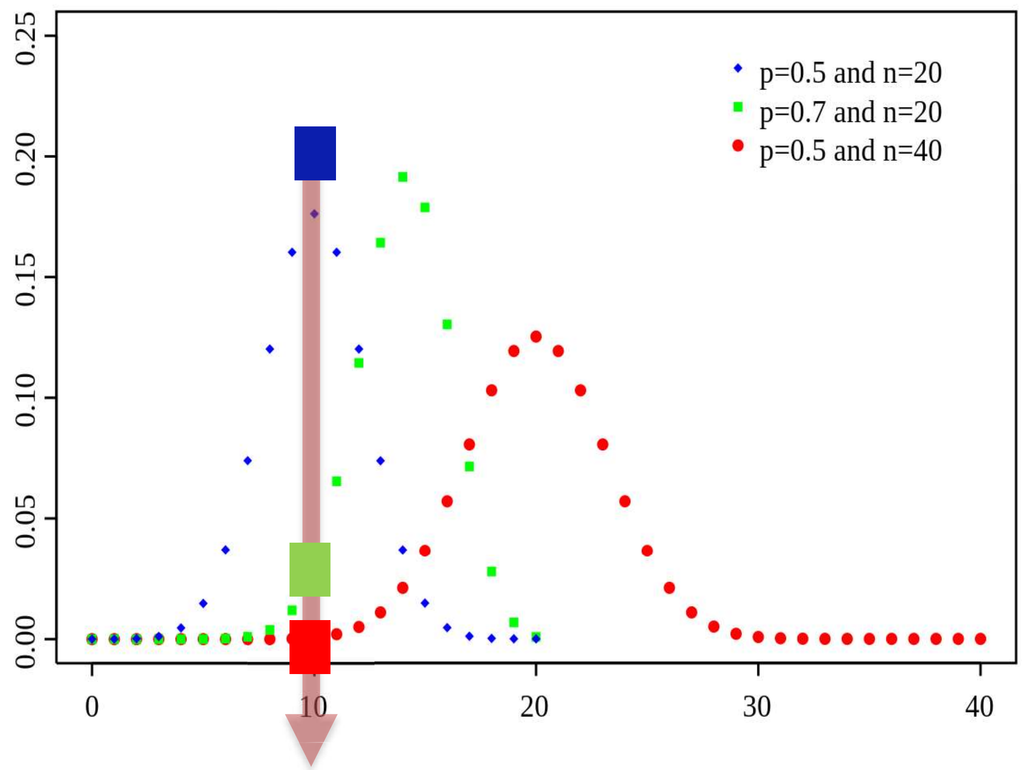
① 이산확률변수:

관측값 x가 고정되어 있을 때, 분포의 모수에 따라 변화하는 확률질량함수의 Y값

=

확률변수가 해당 분포를 따를 가능도

- X~b(20,0.5)일 가능도= 0.59
- X~b(20,0.7)일 가능도= 0.04
- X~b(40,0.5)일 가능도=0.001

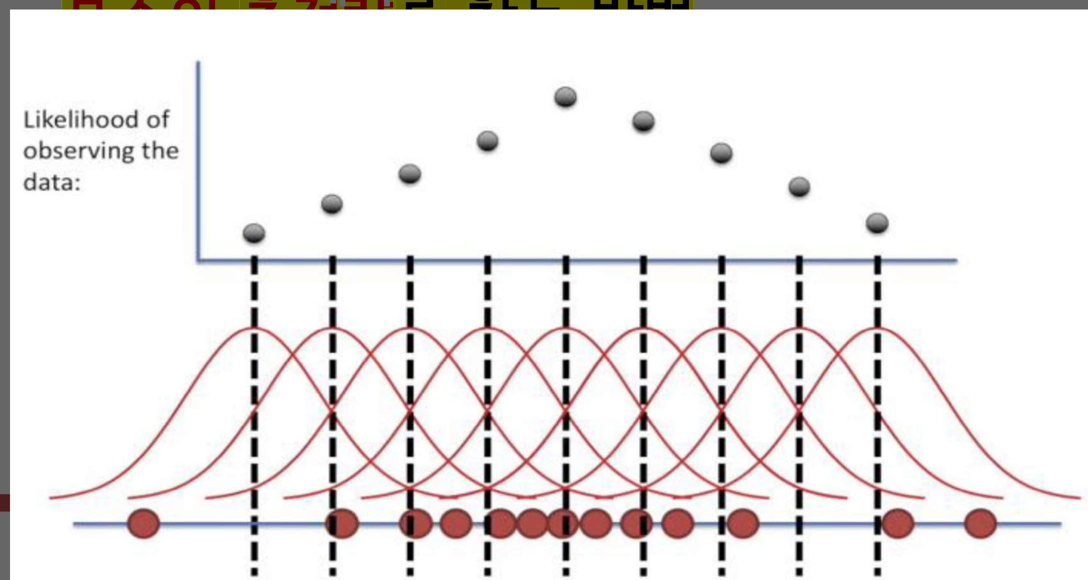


최대가능도추정법(MLE)

원리

주어진 관측값에 대해 해당 분포를 가질 **가능도**를 가장 크게 하는

③



모수 추정

최대가능도 모수 추정

관측값이 **가능도**가 제일 **큰 모수의** 정규분포를 따를 거라 추정

유의성 검정

정의

- 모형의 모수 추정값이 유의한지 검정
- 축소 모형의 적합도가 좋은지에 대한 검정

가설

$$\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 \boldsymbol{x}_1 + \cdots + \boldsymbol{\beta}_k \boldsymbol{x}_k$$

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

H_1 : 적어도 하나의 $\boldsymbol{\beta}$는 0이 아니다

종류

- 왈드 검정
- 스코어 검정
- 가능도비 검정

이탈도

1 정의 : $2(L_S - L_M)$

2 포화모형 S와 관심모형 M을 비교하기 위한 가능도비 통계량
S에는 있지만 M에는 없는 계수들이 0인지 확인 가능 → 모형이 Nested일 때만 사용 가능

예시

$$M : Y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

$$S : Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$L_M \uparrow$ → 이탈도 통계량 ↓ → P-value ↑ → $\beta_3 = 0$, M 모형 적합도 좋음

로지스틱 회귀 모형

반응변수

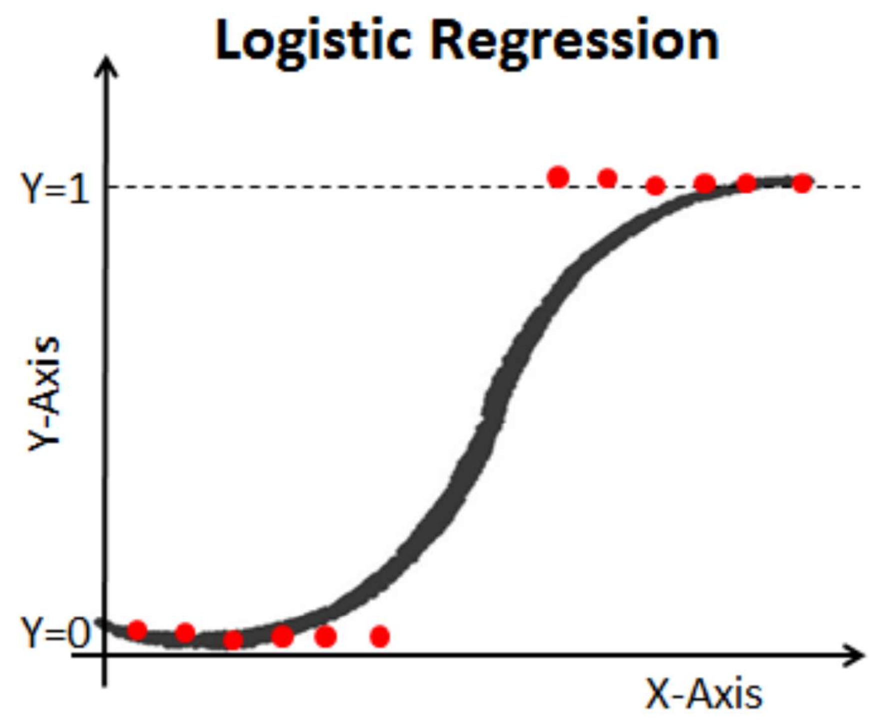
- 이항자료(이항분포)

: 성공이 나타날 확률

연결함수

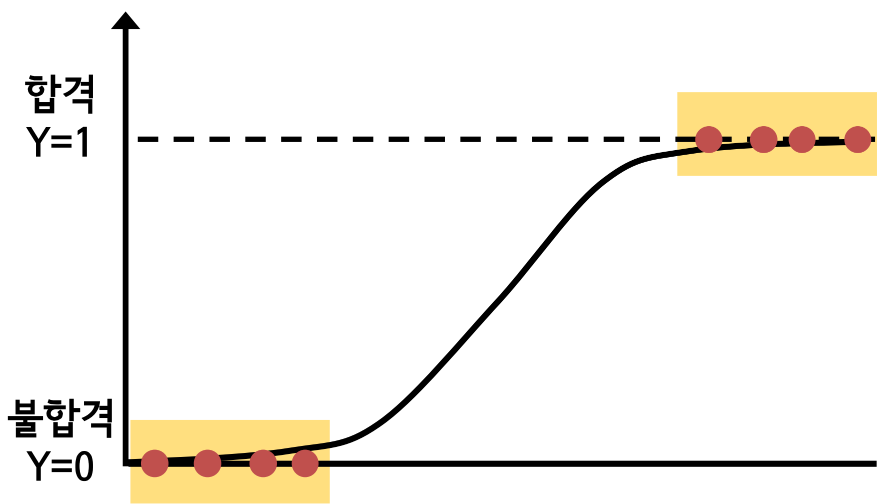
- 로짓연결함수

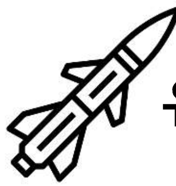
: Y가 이항변수인 경우 곡선으로 fitting



<로지스틱 회귀 모형>

① 좌우변의 범위 일치





유도 과정

1. $0 \leq \pi(x) \leq 1$
2. $0 \leq \pi(x) - 1 \leq 1$
3. $0 \leq \dfrac{\pi(x)}{1 - \pi(x)} \leq \infty$
4. $-\infty \leq \log \dfrac{\pi(x)}{1 - \pi(x)} \leq \infty$

오즈비를 이용한 해석

[예시: 학점에 따른 애인 유무]

X = 학점 / Y = 1 (커플) / Y = 0 (솔로)

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 3 + 2x$$

"x가 한 단위 증가할 때,
Y=1(커플)일 오즈가 $\mathbf{e}^2$배 증가함"

x(학점)가 한 단위 증가할 때마다
커플일 오즈가 $\mathbf{e}^2 = 0.7389$배 증가함..!

더미 변수의 개념

"You can use **dummy** to refer to things that are not real, but have been made to look or behave as if they are real"

- Source: Collins English Dictionary -

더미 변수

1 개념

범주형 변수를 연속형 변수'**스럽게**' 만든 것

2 필요성

회귀분석 등 연속형 변수로만 가능한 분석기법을 사용할 수 있게 만들어 줌

이탈도 차이를 이용

검정 통계량[1)]이 낮음

↓

P-value가 높음

↓

귀무가설 채택
'간단한 모형에 포함되지 않는 모수(β)는 모두 0'

↓

간단한 모형 채택!

1) 검정 통계량: $-2(L_o - L_1)$