범주형자료분석팀

3팀

권가민
김수인
김준령
박윤아
이정민

INDEX

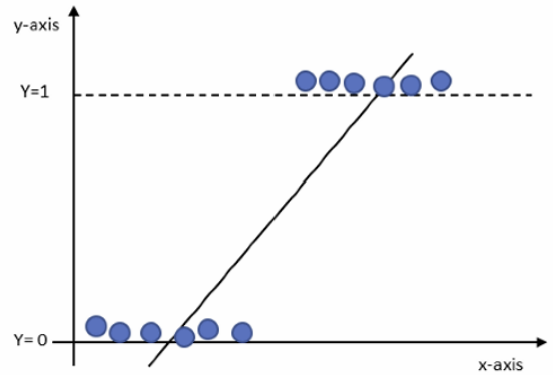
1

GLM

GLM(Generalized Linear Models)

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

우리가 다루는 데이터의 반응변수가
항상 연속형은 아님!
반응변수가 **범주형**이거나 **도수자료**인 경우,
일반 **선형회귀모델을 사용할 수 없음**



GLM(Generalized Linear Models)

일반화 선형 모형 (Generalized Linear Models)

연속형 반응변수들에 대한 모형뿐만 아니라

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

보통의 선형회귀모형과 분산분석(ANOVA) 모형에서 나아가

범주형 변수 등에 대한 모형도 GLM 안에 포함되어 있음

GLM

GLM의 필요성

정규분포를 포함한 **다양한 확률분포**를 사용할 수 있음

변수간 **연관성**을 파악하고 반응변수를 **예측**할 수 있음

GLM의 필요성

정규분포를 포함한 **다양한 확률분포**를 사용할 수 있음

독립변수와 종속변수 간의 선형성, 오차항의 정규성, 독립성, 등분산성

일반 선형회귀모델은 **4가지의 기본 가정**을 만족해야 함

하지만 반응변수가 범주형이거나 도수자료인 경우,

오차항의 확률분포가 **정규분포를 따르지 않음**

선형회귀모델에서의 모형적합방법인 **최소제곱법을 사용할 수 없음**!

GLM의 필요성

정규분포를 포함한 **다양한 확률분포**를 사용할 수 있음

독립변수와 종속변수 간의 선형성, 오차항의 정규성, 독립성, 등분산성

일반 선형회귀모델은 **4가지의 기본 가정**을 만족해야 함

하지만 반응변수가 범주형이거나 도수자료인 경우,

오차항의 확률분포가 **정규분포를 따르지 않음**

선형회귀모델에서의 모형적합방법인 **최소제곱법을 사용할 수 없음**!

GLM

GLM의 필요성

정규분포를 포함한 **다양한 확률분포**를 사용할 수 있음

GLM은 LSM 대신 최대가능도법을 이용해 모형을 적합하기 때문에,
오차항의 정규분포 가정이 **충족되지 않아도** 사용할 수 있음

GLM은 정규분포 뿐만 아니라 다른 분포를 따르는 반응변수도 분석할 수 있게 함

GLM

GLM의 필요성

정규분포를 포함한 **다양한 확률분포**를 사용할 수 있음

GLM은 LSM 대신 최대가능도법을 이용해 모형을 적합하기 때문에,
오차항의 정규분포 가정이 **충족되지 않아도** 사용할 수 있음

GLM은 정규분포 뿐만 아니라 다른 분포를 따르는 반응변수도 분석할 수 있게 함

GLM

GLM의 필요성

변수간 **연관성**을 파악하고 반응변수를 **예측**할 수 있음

GLM은 **범주형 변수와 연속형 변수 간 연관성**을 파악할 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
$\mu(= E(Y))$	$g()$	$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

GLM은 평균의 함수식과 설명변수 간의 관계를

선형예측식(**체계적 성분**)을 통해 정의한 것

GLM의 구성 성분

랜덤성분 (Random Component)

반응변수 Y를 정의하고, Y의 확률분포를 가정함

이 때, 가정한 확률분포 하에서 **Y의 기댓값** μ을 랜덤성분으로서 표기

반응변수 Y의 관측값들은 서로 독립이라고 가정

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

GLM의 구성 성분

랜덤성분 (Random Component)

반응변수 Y를 정의하고, Y의 확률분포를 가정함

이 때, 가정한 확률분포 하에서 **Y의 기댓값** μ을 랜덤성분으로서 표기

모든 확률분포를 GLM의 반응변수의 분포로 활용할 수는 없음

지수족(Exponential Family)에 해당하는 확률분포만 사용 가능

GLM의 구성 성분

체계적 성분 (Systematic Component)

설명변수 X를 명시하는 성분

X들의 **선형 결합 형태**

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

체계적 성분에는 ① **교호작용**을 설명하는 항 $(x_i = x_a x_b)$

② **곡선효과**를 나타내는 항 $(x_i = {x_a}^2)$

을 포함할 수 있음

GLM의 구성 성분

연결함수 (Link Function)

랜덤성분과 체계적 성분을 **연결**하는 함수

두 성분 간 **범위**를 맞춰주는 역할

좌변
이항분포를 따르는 이진변수

≠

우변
연속형 변수

우변의 범위는 $-\infty \sim \infty$를 따르므로 좌변의 범위와 일치하지 않음

연결함수를 통해 **양변의 범위**를 맞춰주는 과정이 필요함!

GLM의 구성 성분

연결함수 (Link Function)

랜덤성분과 체계적 성분을 **연결**하는 함수

두 성분 간 **범위**를 맞춰주는 역할

좌변
이항분포를 따르는 이진변수

≠

우변
연속형 변수

우변의 범위는 $-\infty \sim \infty$를 따르므로 좌변의 범위와 일치하지 않음

연결함수를 통해 **양변의 범위**를 맞춰주는 과정이 필요함!

GLM의 구성 성분

연결함수 (Link Function)

랜덤성분과 체계적 성분을 **연결**하는 함수

두 성분 간 **범위**를 맞춰주는 역할

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

GLM의 구성 성분

연결함수 (Link Function)

랜덤성분과 체계적 성분을 **연결**하는 함수

두 성분 간 **범위**를 맞춰주는 역할

랜덤성분에 대해 **정규분포**를 가정하고,

연결함수로는 '**항등 연결 함수**'를 사용한 것이 선형회귀모형

즉, 회귀모형과 같은 선형모형 역시 GLM의 일종

GLM의 특징

오차항의 다양한 분포 가정 가능

선형 관계식 유지

독립성 가정만 필요

제한적인 범위를 지닌 반응변수도 사용 가능

GLM의 특징

오차항의 다양한 분포 가정 가능

GLM은 정규분포 외에도 반응변수의 오차항이 가진 성질에 따라
어떤 분포든 정의할 수 있음

일반적으로 확률분포에 따른 연결함수는 정해져 있음

(잠시 후 GLM의 종류에서 확인!)

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 해석이 용이함

GLM의 특징

선형 관계식 유지

여기서 '선형'의 관계는

설명변수 X와 반응변수 Y 간의 관계를 말하는 것이 아닌,

회귀 계수 β의 선형성을 가리킴!

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

$x_i = x_a x_b$나 $x_i = {x_a}^2$와 같이

교호작용이나 곡선효과를 나타내는 항이

우변에 위치하고 있더라도 선형성이 유지됨

GLM의 체계적 성분은 선형 관계식의 형태를 띄고 있기 때문에 해석에 용이함

GLM의 특징

독립성 가정만 필요

선형회귀모형과 달리, GLM은 회귀분석의 4가지 가정 중

오차항에 대한 독립성만 만족하면 됨

반응변수 간의 **자기상관성**을 검정해 보아야 함

GLM의 특징

독립성 가정만 필요

선형회귀모형과 달리, GLM은 회귀분석의 4가지 가정 중

오차항에 대한 독립성만 만족하면 됨

반응변수 간의 **자기상관성**을 검정해 보아야 함

GLM의 특징

독립성 가정만 필요

자기상관성 (Autocorrelation): 반응변수의 각 관측치가 **상호 연관성**을 띄는 것

오차들이 서로 독립이 아니어서 **독립성 가정**이 위배되었을 때

오차들 간 **자기상관**이 있다고 함

자기상관성(혹은 오차의 독립성)은 회귀분석 후 더빈-왓슨(DW Test)을 통해 검정함

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

GLM

GLM의 특징

GLM은 보통의 **선형모형을 일반화**한 것

제한적인 범위를 지닌 반응변수도 사용 가능

랜덤성분의 분포와 랜덤성분의 함수(연결함수)를

일반화한 것이 GLM

GLM은 연결함수를 통해 좌변의 랜덤성분과

우변의 체계적성분 간의 **범위를 맞춰줄 수 있음**

랜덤성분이 정규분포가 아닌 다른 분포를 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

<table>
<tr><th>GLM</th><th>랜덤성분</th><th>연결함수</th><th>체계적 성분</th></tr>
<tr><td>기준범주 로짓 모형</td><td rowspan="4">다항 자료</td><td rowspan="4">로짓</td><td rowspan="4">혼합형</td></tr>
<tr><td>누적 로짓 모형</td></tr>
<tr><td>이웃범주 로짓 모형</td></tr>
<tr><td>연속비 로짓 모형</td></tr>
<tr><td>로그 선형 모형</td><td rowspan="4">도수 자료</td><td rowspan="5">로그</td><td>범주형</td></tr>
<tr><td>포아송 회귀 모형</td><td rowspan="4">혼합형</td></tr>
<tr><td>음이항 회귀 모형</td></tr>
<tr><td>카우시 모형</td></tr>
<tr><td>율자료 포아송 회귀 모형</td><td>비율 자료</td></tr>
</table>

GLM의 종류

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

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

이항 랜덤 성분과 로짓 연결 함수

프로빗 회귀 모형

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

이항 랜덤 성분과 프로빗 연결

GLM의 종류

반응변수 : 다항 자료

기준 범주 로짓 모형

$$logit\left[\frac{\pi_j}{\pi_i}\right] = \alpha_j + \beta_j^A x_1 + \cdots + \beta_j^K x_k, \qquad j = 1, \cdots, J-1$$

다항 랜덤 성분(명목형)과 로짓 연결 함수

이웃 범주 로짓 모형

$$log\left[\frac{\pi_{j+1}}{\pi_i}\right] = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k, \qquad j = 1, \cdots, J-1$$

다항 랜덤 성분(순서형)과 로짓 연결 함수

GLM의 종류

반응변수 : 다항 자료

누적 로짓 모형

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

다항 랜덤 성분(순서형)과 로짓 연결 함수

연속비 로짓 모형

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

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

다항 랜덤 성분(순서형)과 로짓 연결 함수

GLM의 종류

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

$$log\left(\frac{\mu}{t}\right) = log(\mu) - log(t) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

포아송 랜덤 성분과 로그 연결 함수

GLM의 모형 적합

모형 적합 (Model Fitting)

주어진 데이터를 근거로 모형의 모수를 추정하는 과정

모집단 분포의 특성을 규정짓는 척도, 모집단의 특성치

GLM은 회귀의 기본 4가지 가정을 만족하지 못하므로
LSE를 사용할 수 없음

최대가능도추정법(Maximum Likelihood Estimation, MLE)을 활용

GLM의 모형 적합

모형 적합 (Model Fitting)

주어진 데이터를 근거로 모형의 모수를 추정하는 과정

모집단 분포의 특성을 규정짓는 척도, 모집단의 특성치

GLM은 회귀의 기본 4가지 가정을 만족하지 못하므로
LSE를 사용할 수 없음

최대가능도추정법(Maximum Likelihood Estimation, MLE)을 활용

GLM의 모형 적합

가능도 (Likelihood)

고정된 관측값이 어떤 확률분포를 따를 **가능성**

각 데이터를 후보 분포에 대입해 얻은 값으로 수치화할 수 있음

가능도들을 다 곱한 것이

아래의 **가능도 함수**(Likelihood Function) $P(x|\theta)$

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

GLM의 모형 적합

로그가능도함수를 편미분하여 0이 되도록 하는 방정식의 해

최대가능도추정량 (Maximum Likelihood Estimator)

가능도함수 $P(x|\theta)$가 **최대**가 되도록 하는 추정량 $\hat{\theta}$

확률표본 $X_1, \cdots X_n$이 모수가 λ인 지수분포 $(f_x(x) = \lambda e^{-\lambda x})$를 따를 때,

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

$l(\lambda)$을 편미분해 값이 0이 되도록 하는 값은

$\frac{dlnL(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$임에 따라 $\hat{\lambda} = \frac{1}{\bar{x}}$이 됨

2

유의성 검정

유의성 검정

유의성 검정

모형의 **모수에 대한 추정값이 유의한지**

혹은 축소 모형의 적합도가 좋은지 판단하는 검정

GLM 모형에서의 유의성 검정 가설

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

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

H_1: 적어도 하나의 β는 0이 아니다

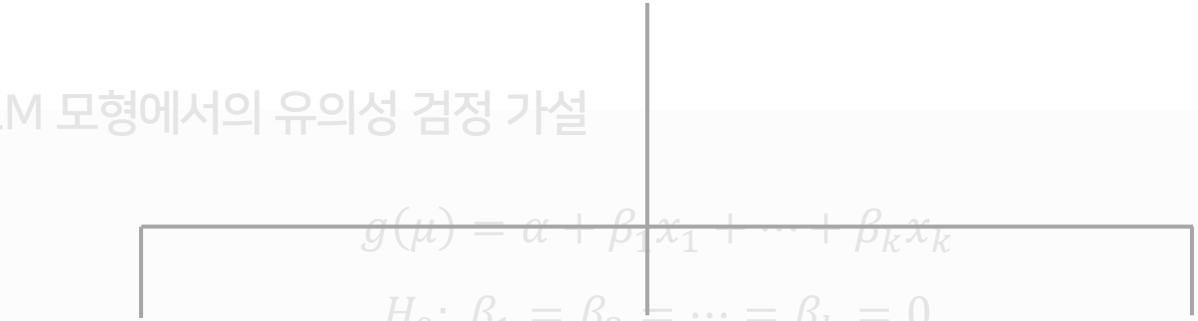
귀무가설을 기각하지 못할 경우 해당 GLM 모형은 유의하지 않음

유의성 검정

유의성 검정

모형의 **모수에 대한 추정값이 유의한지**

혹은 축소 모형의 적합도가 좋은지 판단하는 검정



GLM 모형에서의 유의성 검정 가설

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

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

H_1: 적어도 하나의 β는 0이 아니다

귀무가설이 참일 경우 해당 GLM 모형은 분석 가치가 없음

유의성 검정

왈드 검정 (Wald Test)

왈드 검정

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

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

왈드 통계량은 ML 추정량들의 **대표본 정규성**을 이용하기 때문에
회귀 계수에 대한 **추정값과 표준오차**만 사용함

유의성 검정

왈드 검정 (Wald Test)

왈드 검정

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

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

범주형 자료이거나 소표본인 경우 검정력이 감소

가능도비 검정 사용

가능도비 검정 (Likelihood-ratio Test)

가능도비 검정

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

기각역 : $G^2 \geq \chi^2_{\alpha,df}$

귀무가설을 만족하는 **단순모형에 대한 가능도 함수(l_0)**와
전체공간 하에서의 **완전모형에 대한 가능도 함수(l_1)**의 비를 이용
이때 분모의 가능도 함수(l_1) 최댓값은 MLE를 통해 계산

가능도비 검정 (Likelihood-ratio Test)

가능도비 검정

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

기각역 : $G^2 \geq \chi^2_{\alpha,df}$

즉, LRT는 **두 가능도 함수의 최댓값을 비교**하는 방식으로 진행되며
자유도는 H_0, H_1 간 모수 개수의 차이

가능도비 검정 (Likelihood-ratio Test)

가능도비 검정통계량의 의미

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

$\frac{l_0}{l_1}$ 가 **1에 가까워지면** 가능도 함수의 차이가 작다는 것을,

$\frac{l_0}{l_1}$ 가 **1보다 작아진다면** 가능도 함수의 차이가 크다는 것을 의미하고

이때 **귀무가설을 기각**함

l_1은 항상 l_0 보다 큰 값을 가짐

유의성 검정

가능도비 검정 (Likelihood-ratio Test)

가능도비 검정통계량의 의미

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

검정 과정

l_0 과 l_1 의 차이가 큼 → **검정통계량**의 값이 큼 → 작은 p-value →

귀무가설 기각, 적어도 하나의 β는 0이 아님 → 모형의 모수 추정값 유의함

유의성 검정

가능도비 검정 (Likelihood-ratio Test)

가능도비 검정통계량의 의미

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

검정 과정

귀무가설 하에서와 전체공간 하에서의 가능도 함수를 모두 사용해 왈드 검정에 비해 검정력이 좋고 신뢰도도 높음!

유의성 검정

이탈도 (Deviance)

관심모형과 포화모형

관심모형 (M)

검정 과정에서 우리가 관심 있는 모형

즉, **유의성을 검정하고자 하는 모형**

포화모형 (S)

주어진 관측값들에 대해 완벽하게 자료에 적합하는 모형

즉, **모든 관측값에 대해 모수를 갖는** 가장 복잡한 모형

유의성 검정

이탈도 (Deviance)

관심모형과 포화모형

귀무가설 (H_0)

H_0: 관심모형에 속하지 않는 모수는 모두 0이다

귀무가설 채택 시 **관심모형** 사용

대립가설 (H_1)

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

대립가설 채택 시 관심모형 사용불가, **모수가 추가된 모형** 필요

유의성 검정

이탈도 (Deviance)

관심모형과 포화모형

관심모형 (M)

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

포화모형 (S)

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$(x_3) + \beta_4 \times$ 범주 팀원들 간의 사랑(x_4)

4개의 관측값 중 반응 변수에 대해

세미나 시간과 패키지 난이도가 미치는 영향에만 관심이 있을 경우의 예시

유의성 검정

이탈도 (Deviance)

이탈도

포화모형과 관심모형을 비교하기 위한 **가능도비 통계량**

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

검정 과정

$L_M - L_S$가 큼 → **이탈도가** 큼 → 작은 p-value → **귀무가설 기각**

→ 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아님

→ **관심모형이 적합하지 않음**

유의성 검정

이탈도 (Deviance)

이탈도

포화모형과 관심모형을 비교하기 위한 **가능도비 통계량**

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

관심모형과 포화모형 하의 로그 가능도 함수의 최댓값 차이$(L_M - L_S)$ 이용

GLM 모형에서의 이탈도는 **근사적으로 카이제곱분포를 따름**

유의성 검정

이탈도 (Deviance)

이탈도 사용 조건

이탈도

이탈도는 포화모형에는 있고

관심모형에는 없는 계수들이 0인지 확인하는 과정

즉, 관심모형은 포화모형에 내포된(nested) 관계를 만족해야 함

M의 모수 ⊂ S의 모수

포화모형이 모수를 더 많이 포함하고 있어 $L_M \leq L_S$ 만족

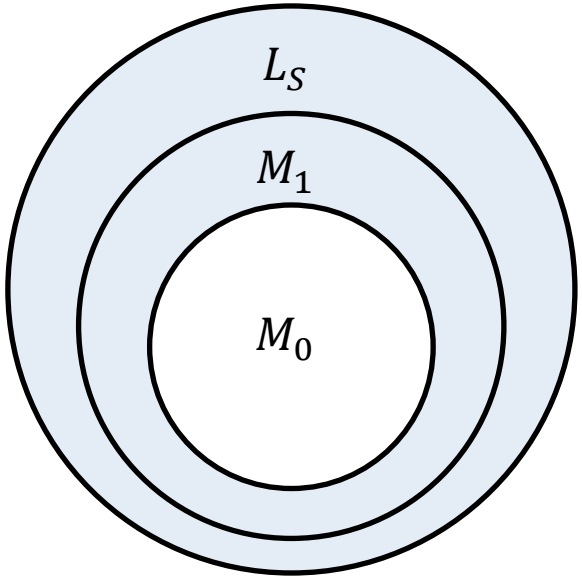
GLM 모형에서의 이탈도는 근사적으로 카이제곱분포를 따름

유의성 검정

이탈도 (Deviance)

이탈도와 가능도비 검정의 관계

두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음



M_0의 이탈도 $-$ M_1의 이탈도

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

$= -2(L_0 - L_1)$

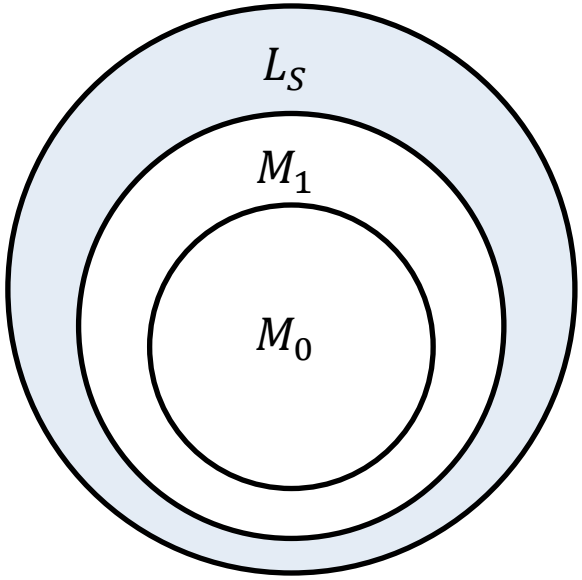
(=가능도비 검정 통계량)

유의성 검정

이탈도 (Deviance)

이탈도와 가능도비 검정의 관계

두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음



M_0의 이탈도 $-$ M_1의 이탈도

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

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

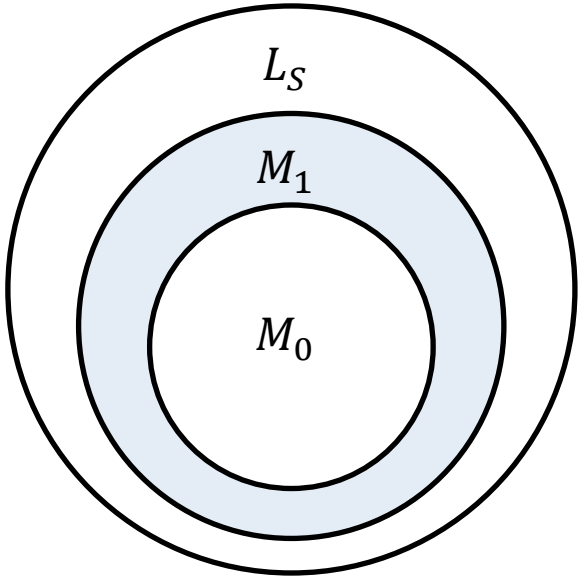
(=가능도비 검정 통계량)

유의성 검정

이탈도 (Deviance)

이탈도와 가능도비 검정의 관계

두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음



M_0의 이탈도 $-$ M_1의 이탈도

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

$= -2(L_0 - L_1)$

(=가능도비 검정 통계량)

유의성 검정

이탈도 (Deviance)

이탈도와 가능도비 검정의 관계

두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음

검정 과정

관심모형(M_0, M_1)간 이탈도의 차이가 작음 →

가능도비 검정통계량 작음 → 큰 p-value → **귀무가설 기각 불가** →

M_0에 포함되지 않는 모수들은 모두 0 → **간단한 관심모형 M_0 이 더 적합**

유의성 검정

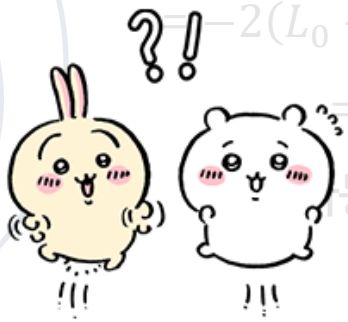
이탈도 (Deviance)

이탈도와 가능도비 검정의 관계

두 모형 간 이탈도 값의 차이는 가능도비 검정 통계량과 같음

모형 M_0이 모형 M_1에 내포된(nested) 모형이 아니라면

AIC, BIC와 같은 모형 선택을 위한 측도들을 활용!



3

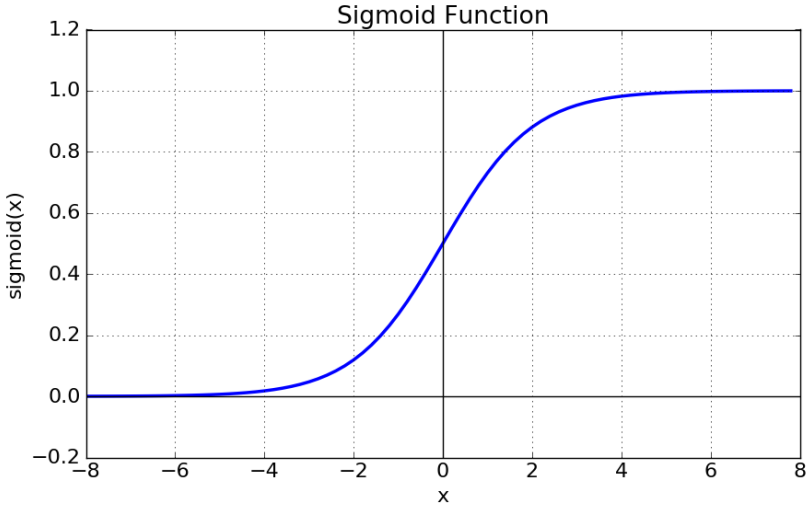
로지스틱 회귀 모형

로지스틱 회귀 모형

로지스틱 회귀 (Logistic Regression)

반응변수가 이항자료일 때의 회귀

즉, 반응변수는 이항분포를 따르게 되고 성공일 확률로 표기됨



로지스틱 회귀 모형은 **성공확률과 x의 비선형 관계를 표현**함

이항변수와 연속형 변수 간의 관계를 GLM의 형태로 표현한 것

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: 0~1 ≠ 우변 범위: $-\infty \sim \infty$

이항분포를 따르는 반응변수 $\pi(x) = P(Y = 1|X = x)$는 확률로 나타나

0~1의 값을 가지고 설명변수의 선형식은 $-\infty \sim \infty$의 값을 가짐!

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: 0~1 ≠ 우변 범위: $-\infty \sim \infty$

좌변을 오즈 형태로 만들기

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

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

좌변에 로그 취하기 (로짓 연결함수)

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

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: 0~1 ≠ 우변 범위: $-\infty \sim \infty$

좌변을 오즈 형태로 만들기

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

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

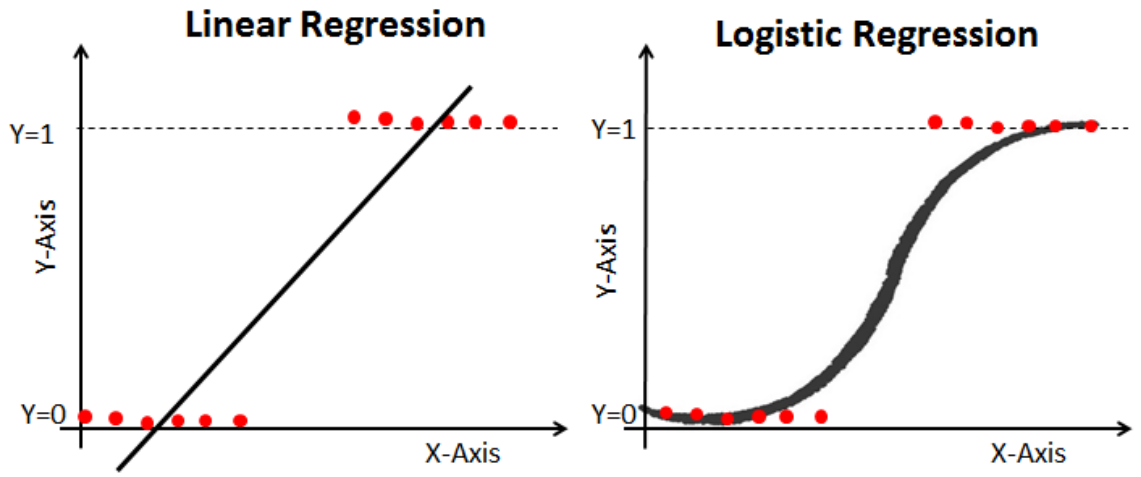
좌변에 로그 취하기 (로짓 연결함수)

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

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

로지스틱 회귀 모형의 장점

이항변수와 연속형 변수 간의 범위 일치



이항 랜덤 성분을 로짓 연결함수를 통해 체계적 성분과 연결한 GLM의 일종

일반선형모델과 달리 0~1 사이의 범위를 벗어나지 않으므로

확률의 범위와 반응 변수의 범위가 일치됨을 확인 가능!

로지스틱 회귀 모형

로지스틱 회귀 모형의 장점

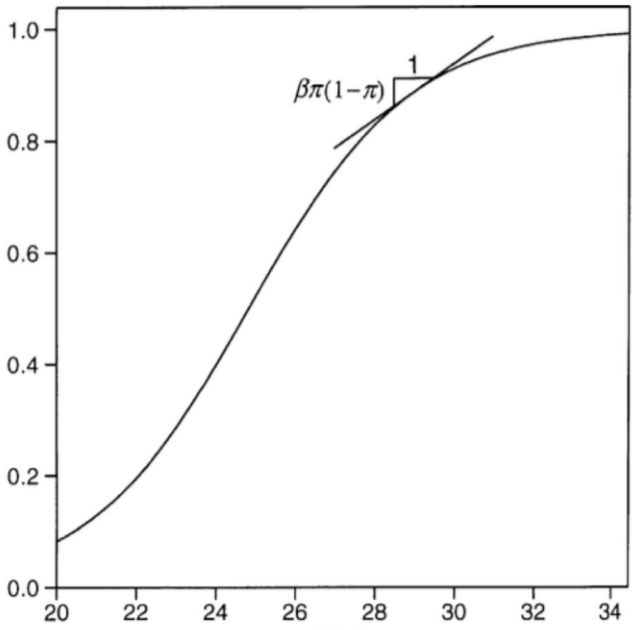
후향적 연구에도 사용 가능

일반 회귀모형의 기본 가정 중 **독립성**만 만족하면 됨

로지스틱 회귀 모형

로지스틱 회귀의 기울기

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



회귀계수 β에 따라 증가/감소비율 결정

$\beta > 0$ 일 때 상향곡선

$\beta < 0$ 일 때 하향곡선

β의 절댓값이 클수록 기울기 변화율 증가

로지스틱 회귀 모형의 해석

확률을 통한 해석

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

특정 x에서 $\pi(x)$가 분석자가 사전에 지정한

cut-off point보다 크면 Y=1, 작으면 Y=0으로 예측

보통 0.5지만 상황에 따라 조정 가능 (3주차 클린업에서!)

축구 직관 횟수(X)에 따른 싸인 유니폼 당첨 여부(Y)

$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 2 + 0.02x$ 로 가정할 경우 관측치 X=20이라면

$\pi(x) = \frac{\exp(2+0.02\times20)}{1+\exp(2+0.02\times20)} =$ 약 0.92 > 0.5

로지스틱 회귀 모형의 해석

확률을 통한 해석

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

특정 x에서 $\pi(x)$가 분석자가 사전에 지정한

cut-off point보다 크면 Y=1, 작으면 Y=0으로 예측

보통 0.5지만 상황에 따라 조정 가능 (3주차 클린업에서!)

축구 직관 횟수(X)에 따른 싸인 유니폼 당첨 여부(Y)

$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 2 + 0.02x$ 로 가정할 경우 관측치 X=20이라면

$\pi(x) = \frac{\exp(2+0.02\times20)}{1+\exp(2+0.02\times20)} =$ 약 $0.92 > 0.5$

싸인 유니폼 획득가능이라고 판단 (Y=1)

로지스틱 회귀 모형의 해석

오즈비를 통한 해석

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

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

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

다른 설명변수가 모두 고정되어 있을 때

x가 한 단위 증가하면 오즈가 e^{β}배 증가한다고 해석함

로지스틱 회귀 모형의 해석

오즈비를 통한 해석

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

축구 직관 횟수(X)에 따른 싸인 유니폼 당첨 여부(Y)

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

싸인 유니폼에 당첨될 오즈가 $e^{0.02}$ = 약 1.02배 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 구분해야함

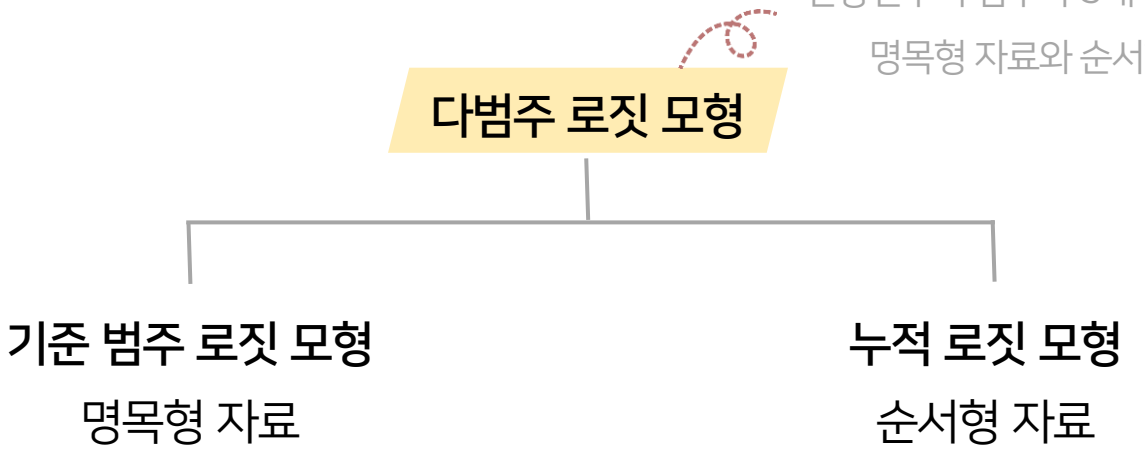
다범주 로짓 모형

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 구분해야함



다범주 로짓 모형

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 범주 중 마지막 범주

범주에 대한 첨자

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

기준 범주에 대한 첨자

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

다범주 로짓 모형

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 범주 중 마지막 범주

기준 범주 J와 그 외 범주들을 각각 짝지어 **로짓**을 정의

그 결과 $(J-1)$개의 로짓 방정식이 생성됨.

다범주 로짓 모형

기준 범주 로짓 모형

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

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

좌변을 확률에 대한 식으로 재정의

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

다범주 로짓 모형

기준 범주 로짓 모형

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

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

좌변을 확률에 대한 식으로 재정의

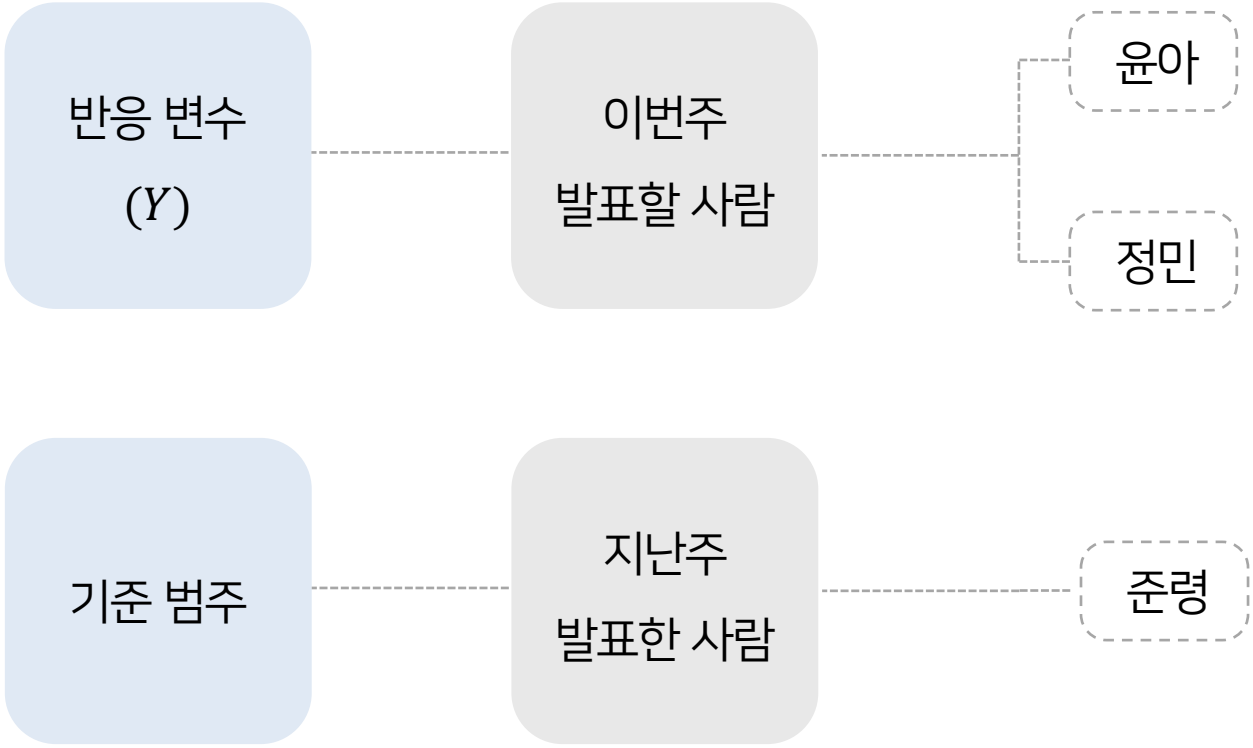
$$\pi_{ij} = \frac{e^{\alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^K x_K}}{\sum_{i=1}^{J} e^{\alpha_i + \beta_i^1 x_1 + \cdots + \beta_i^K x_K}}, j = 1, \cdots (J-1)$$

다범주 로짓 모형

기준 범주 로짓 모형

예시

범주 팀 신입 부원의 발표확률을 기준 범주 로짓 모형으로 알아보자!



다범주 로짓 모형

기준 범주 로짓 모형

예시

기준 범주를 '준령'으로 정의

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

$$\log\frac{\pi_{윤아}}{\pi_{준령}} = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log\frac{\pi_{정민}}{\pi_{준령}} = 2 + 0.22x_1 + \cdots + 0.46x_K$$

같은 설명변수여도 회귀계수 β가 다른 값을 가지는 것을 알 수 있음

다범주 로짓 모형

기준 범주 로짓 모형

예시

$$\log\frac{\pi_{\text{윤아}}}{\pi_{\text{준령}}} = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log\frac{\pi_{\text{정민}}}{\pi_{\text{준령}}} = 2 + 0.22x_1 + \cdots + 0.46x_K$$

좌변을 확률에 대한 식으로 재정의

윤아가 이번주에 발표할 확률

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

다범주 로짓 모형

기준 범주 로짓 모형

예시

$$\log\frac{\pi_{\text{윤아}}}{\pi_{\text{준령}}} = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log\frac{\pi_{\text{정민}}}{\pi_{\text{준령}}} = 2 + 0.22x_1 + \cdots + 0.46x_K$$

좌변을 확률에 대한 식으로 재정의

윤아가 이번주에 발표할 확률

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

다범주 로짓 모형

기준 범주 로짓 모형

① j범주와 기준 범주 간 비교

오즈를 이용한 해석

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

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

다른 설명변수들이 고정되어 있을 때

x_i가 한 단위 증가하면 기준범주 대신 **j 범주일 오즈**가 **$e^{\beta_j^i}$배 증가**

다범주 로짓 모형

기준 범주 로짓 모형

② 기준범주가 아닌 두 범주 간 비교

오즈를 이용한 해석

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

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

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

다른 설명변수들이 고정되어 있을 때

x_i가 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 고려**하므로
반응변수를 **절단점**에 따라 두 그룹으로 나누는 **Collapse 과정**이 필요함

절단점
(Cut-Point)

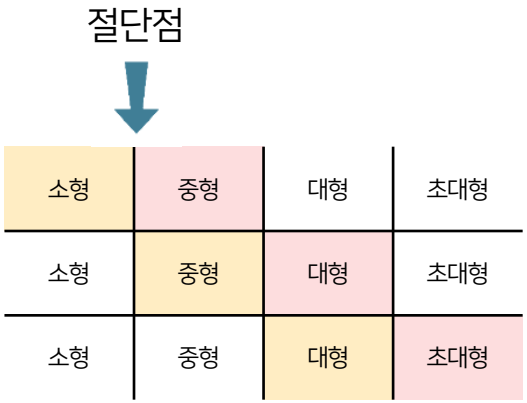
순서형 반응변수의 범주들을 나누는 기준으로
각 행마다 색깔이 바뀌는 경계점이 Cut-Point가 됨

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

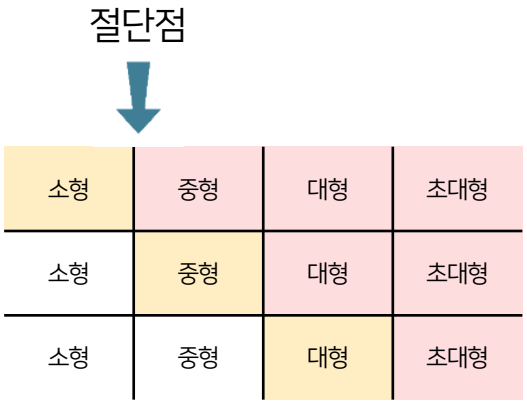
다범주 로짓 모형

누적 로짓 모형

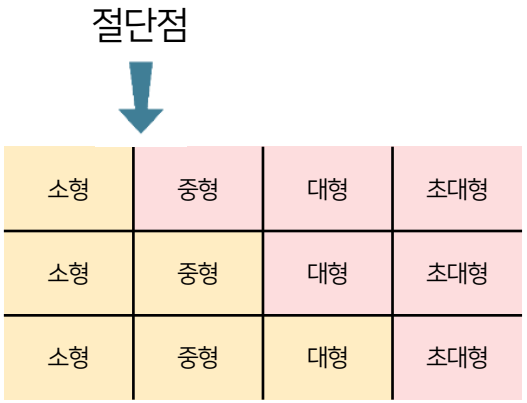
순서형 자료는 절단점을 정해 **Collapse를 진행하는 방법**에 따라
다범주 로짓 모형들이 구분됨



<이웃 범주 로짓 모형>



<연속비 로짓 모형>

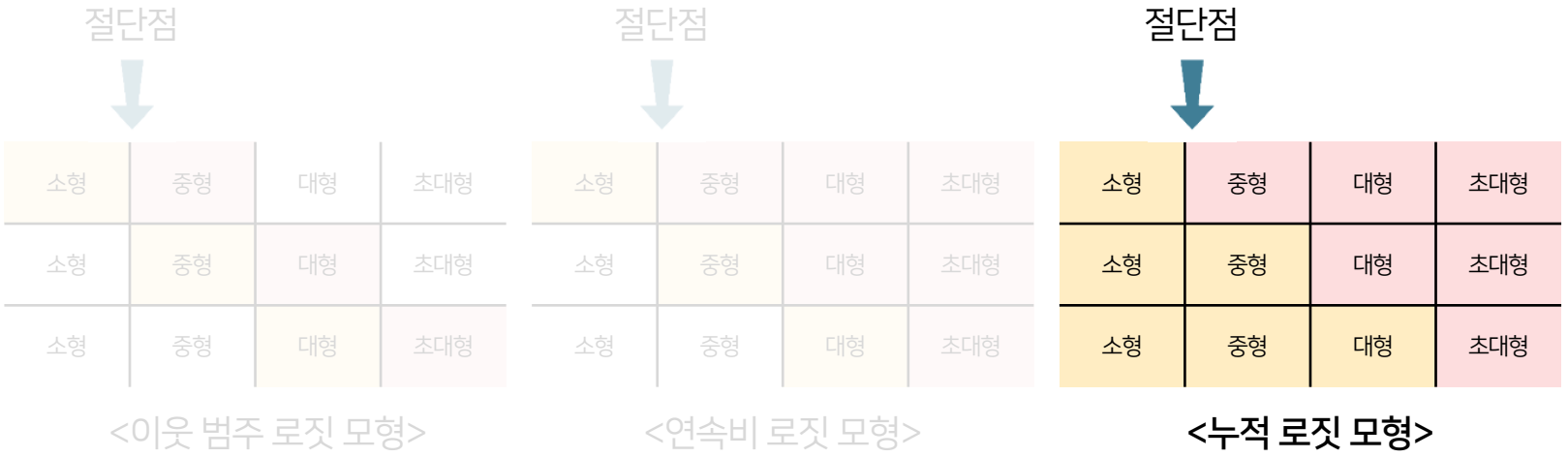


<누적 로짓 모형>

다범주 로짓 모형

누적 로짓 모형

순서형 자료는 절단점을 정해 **Collapse를 진행하는 방법**에 따라
다범주 로짓 모형들이 구분됨



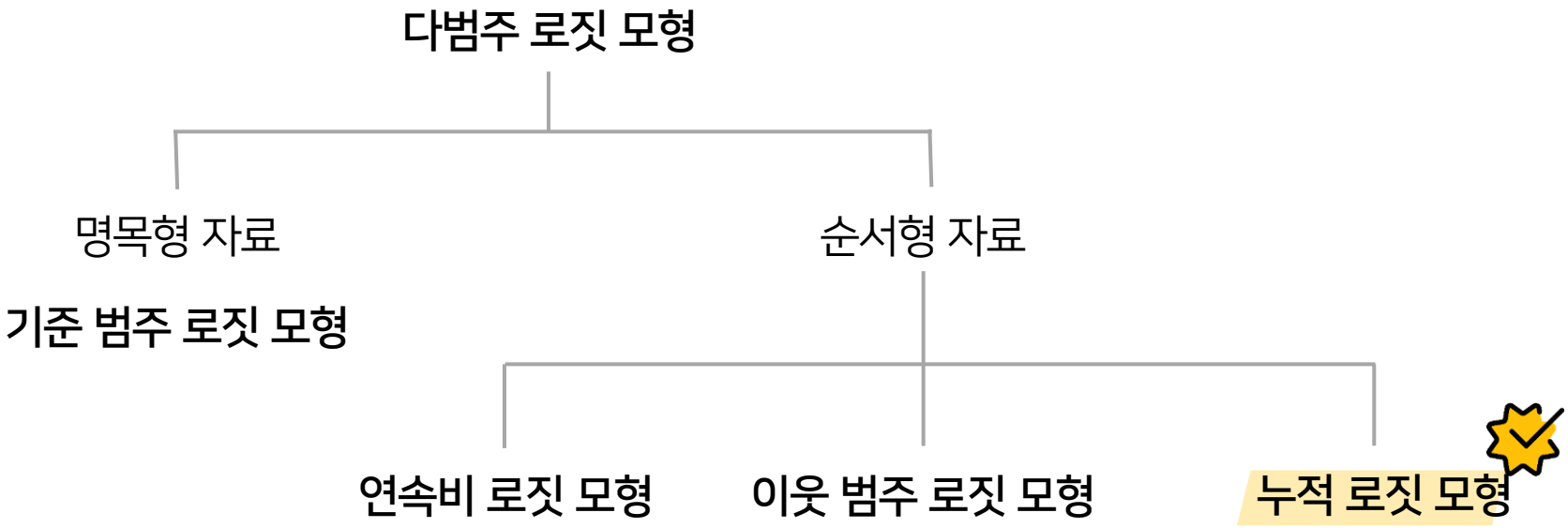
다범주 로짓 모형

누적 로짓 모형

누적 로짓 모형 (Cumulative Logit Model)

반응변수가 순서형 자료일 때 사용하는 다범주 로짓 모형

누적 확률에 **로짓 연결함수**를 사용



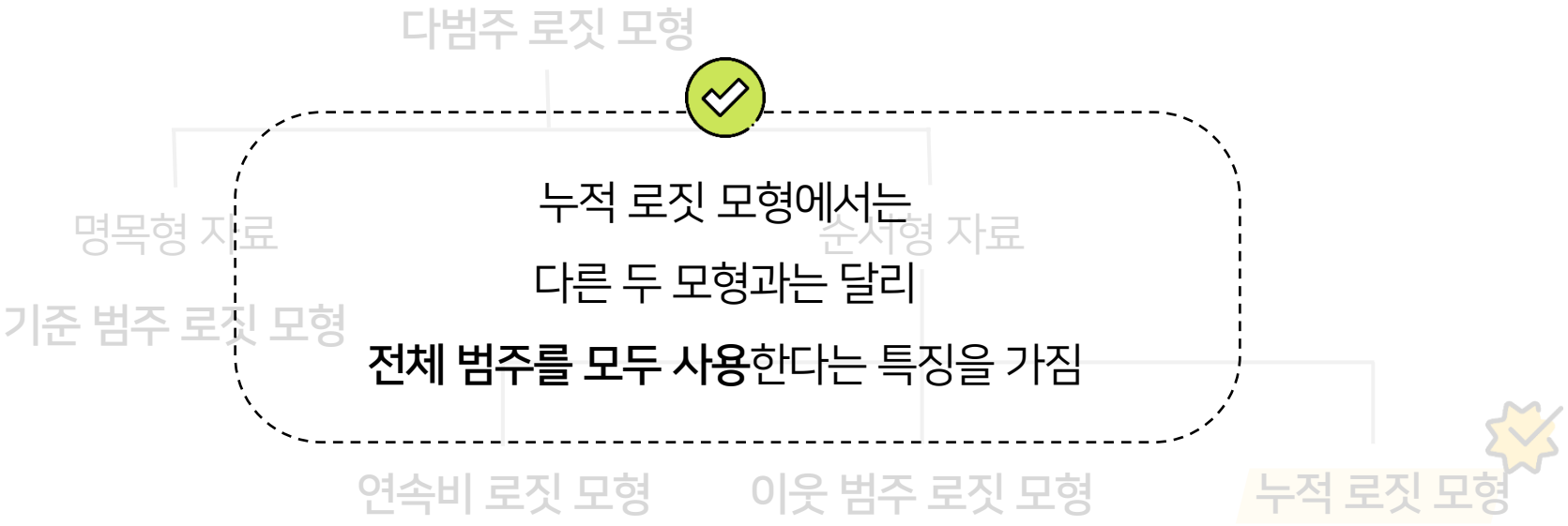
다범주 로짓 모형

누적 로짓 모형

누적 로짓 모형 (Cumulative Logit Model)

반응변수가 순서형 자료일 때 사용하는 다범주 로짓 모형

누적 확률에 **로짓 연결함수**를 사용



누적 로짓 모형에서는

다른 두 모형과는 달리

전체 범주를 모두 사용한다는 특징을 가짐

누적 로짓 모형

누적 확률 (Cumulative Probability)

반응변수가 총 J개의 범주를 가질 때

첫 번째 범주부터 j 번째 범주까지의 누적확률

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

$j = 1, \cdots J$ 반응변수가 총 J개로 구성되어 있음

다범주 로짓 모형

누적 로짓 모형

모형 유도

누적 확률

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

누적확률을 로그 오즈의 형태로 조작

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

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

누적 로짓 모형

모형 유도

누적 확률

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

누적확률을 로그 오즈의 형태로 조작

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

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

다범주 로짓 모형

누적 로짓 모형

누적 로짓 모형의 최종 형태

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

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

첫 번째 범주부터 j번째 범주까지의 누적 로짓

다범주 로짓 모형

누적 로짓 모형

누적 로짓 모형

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

기준 범주에 따라 $\boldsymbol{\beta}$**값 상이함**

기준 범주 로짓

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

기준 범주에 상관없이 $\boldsymbol{\beta}$**값이 동일함**

누적 로짓 모형의 β값이 동일한 것은

비례 오즈 가정 때문임!

다범주 로짓 모형

누적 로짓 모형

누적 로짓 모형

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

기준 범주에 따라 $\boldsymbol{\beta}$**값 상이함**

기준 범주 로짓

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

기준 범주에 상관없이 $\boldsymbol{\beta}$**값이 동일함**

누적 로짓 모형의 β값이 동일한 것은

비례 오즈 가정 때문임!

누적 로짓 모형

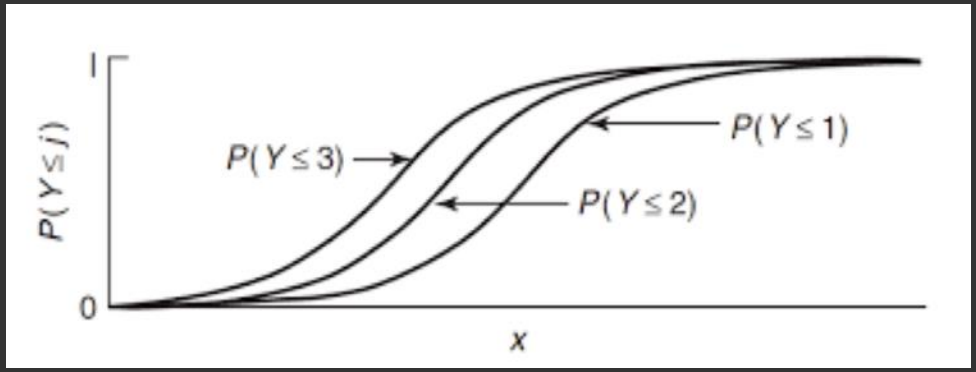
비례 오즈 가정 (Proportional Odds)

누적 로짓 모형에서 $(J-1)$개의 로짓 방정식에 대한 β의 효과가

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

모두 동일하다는 가정

기준 범주에 따라 $\boldsymbol{\beta}$**값 상이함**



서로 다른 로짓 방정식의 그래프가 **수평 이동**을 한 것처럼 나타남

누적 로짓 모형의 β값이 동일한 것은

α는 다르지만 β**는 같음**

비례 오즈 가정 때문임!

다범주 로짓 모형

누적 로짓 모형

예시

범주팀 클린업에 대한 만족도 지표

낮음 / 보통 / 높음 / 아주 높음

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

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

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

로짓 모형은 총 $3, (4-1)$개의 로짓 방정식을 가짐

다범주 로짓 모형

누적 로짓 모형

예시

범주팀 클린업에 대한 만족도 지표

낮음 / 보통 / 높음 / 아주 높음

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

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

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

기준 범주와 상관없이 **일관된 β 값**을 가지는 것을 확인!

누적 로짓 모형

오즈를 이용한 해석

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

다른 설명변수가 모두 고정되어 있을 때

x_i가 한 단위 증가하면 $Y > j$에 비해 **$Y \le j$일 오즈가 e^{β_i}배 증가**한다고 해석함

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

평균이 작은 **포아송 분포**의 경우
정규성과 등분산성 가정 충족하지 않아
일반 선형모형 적용 **불가능**

표준오차나 유의성 수준이 편향되는 문제를
포아송 회귀 모형으로 해결 가능

포아송 회귀 모형

평균이 작은 **포아송 분포**의 경우
정규성과 등분산성 가정 충족하지 않아
일반 선형모형 적용 **불가능**

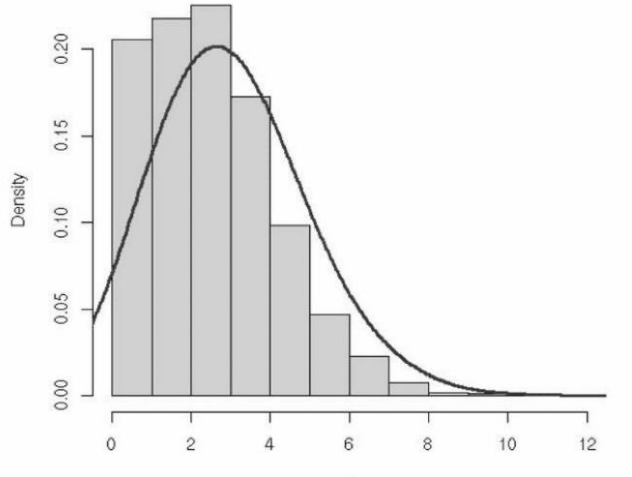
표준오차나 유의성 수준이 편향되는 문제를
포아송 회귀 모형으로 해결 가능

포아송 회귀 모형

포아송 회귀 모형 (Poisson Regression Model)

반응변수가 도수 자료처럼 포아송 분포를 따를 때 사용하는 모형

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

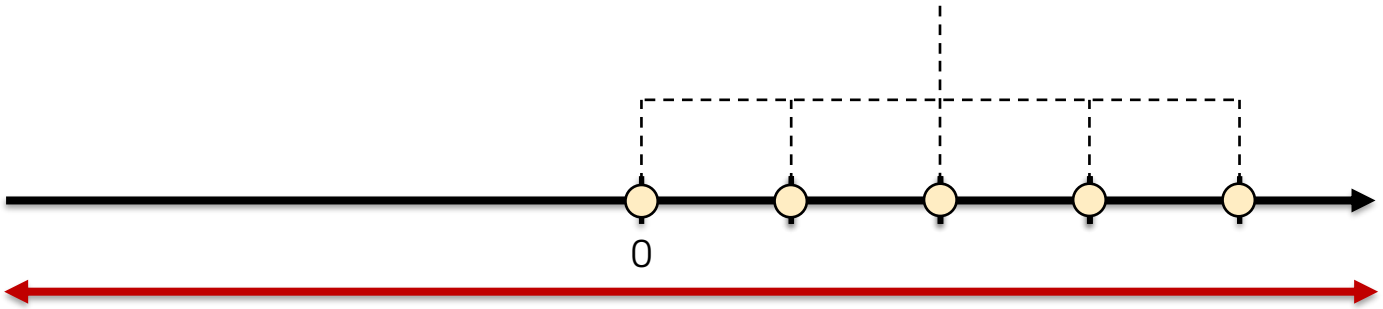


포아송 회귀 모형의 **랜덤성분**은 포아송 분포를 따름

포아송 회귀 모형

도수 자료와 체계적 성분의 **불균형한 범위를 교정**하기 위해
로그 연결함수를 사용

도수 자료는 음이 아닌 임의의 정수값



체계적 성분 $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$는 $(-\infty, \infty)$의 범위를 지님

포아송 회귀 모형

도수 자료와 체계적 성분의 **불균형한 범위를 교정**하기 위해
로그 연결함수를 사용

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

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

좌변에 로그 취하기 (로그 연결함수)

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

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

포아송 회귀 모형

① 도수를 이용한 해석

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

기대도수 μ를 도출할 수 있게 됨

다른 설명변수들이 고정되어 있을 때 x_i**가 한 단위 증가**하면 μ가 e^{β_i}**배 증가**

포아송 회귀 모형

포아송 회귀 모형에
$(x+1)$과 x를 대입한 후 빼기

② 오즈비를 이용한 해석

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

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

다른 설명변수들이 고정되어 있을 때 $\boldsymbol{x}$**가 한 단위 증가**하면 $\boldsymbol{\mu}$가 $\boldsymbol{e^{\beta}}$**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

1년에 복권을 사기 위해 10,000원을 투자한다면

도수를 이용한 해석

$$\mu = \exp(-2 + 0.0001 * 10000)$$
$$\approx 0.3679$$

기대도수가 1회 미만으로,

1년에 만원을 투자해서는 평생 한 번도 로또 1등에 당첨될 수 없는 것임!

포아송 회귀 모형

예시

1년에 복권을 사기 위해 쓰는 돈이 **1원** 늘어난다면

오즈비를 이용한 해석

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

평생 로또 1등에 당첨될 **기대도수**가 **1.0001배만큼 증가**함

포아송 회귀 모형

한계

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

μ에 대한 예측값인 **기대도수만 산출 가능**

비율 자료를 이용한 율자료 포아송 회귀 모형 사용!

포아송 회귀 모형

한계

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

μ에 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** $\left(\frac{\mu}{t}\right)$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 연결함수를 사용
- t는 기준이 되는 지표 값을 나타냄

율자료 포아송 회귀 모형

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

기존의 기대도수 (μ) 대신 **비율자료** $\left(\frac{\mu}{t}\right)$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 연결함수를 사용

t는 기준이 되는 지표 값을 나타냄

율자료 포아송 회귀 모형

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

기존의 기대도수 (μ) 대신 **비율자료** $\left(\frac{\mu}{t}\right)$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 연결함수를 사용

t는 기준이 되는 지표 값을 나타냄

포아송 회귀 모형

율자료 포아송 회귀 모형

오즈비를 이용한 해석

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

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

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

다른 설명변수들이 고정되어 있을 때

x가 한 단위 증가하면 **기대비율**이 **e^{β}배 증가**

포아송 회귀 모형

포아송 회귀 모형의 문제점

포아송 회귀 모형

포아송 회귀 모형의 문제점

과대산포 문제 (Overdispersion Problem)

예측되는 분산보다 더 큰 분산을 가질 때 나타나는 문제

등산포 가정을 만족하지 못할 경우

주어진 도수 자료의 평균과 분산이 같다는 가정

R에서 AER 패키지 내의 함수 **dispersiontest()** 통해

과대산포 검정을 진행할 수 있음

포아송 회귀 모형

포아송 회귀 모형의 문제점

현실에서는 포아송 분포를 따르는 도수 자료가

등산포 가정을 만족하지 않는 경우 존재

과대산포 문제로 인해 **분산이 과소평가**되어 **검정 결과가 왜곡됨**

음이항 회귀 모형을 이용하여 해결!

포아송 회귀 모형

포아송 회귀 모형의 문제점

현실에서는 포아송 분포를 따르는 도수 자료가

등산포 가정을 만족하지 않는 경우 존재

과대산포 문제로 인해 **분산이 과소평가**되어 **검정 결과가 왜곡됨**

음이항 회귀 모형을 이용하여 해결!

포아송 회귀 모형

포아송 회귀 모형의 문제점

음이항 회귀 모형 (Negative Binomial Regression)

랜덤성분이 음이항 분포를 따르고 로그 연결함수로 구성된 GLM

평균과 분산 간의 비선형성을 가정

$$E(Y) = \mu$$

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

평균과 분산의 차이를 발생시키는 산포모수

포아송 회귀 모형

포아송 회귀 모형의 문제점

음이항 회귀 모형 (Negative Binomial Regression)

랜덤성분이 음이항 분포를 따르고 로그 연결함수로 구성된 GLM

음이항 분포는
평균보다 큰 분산 값을 가짐

포아송 분포의
등산포 가정을 완화하여
과대산포 문제 해결

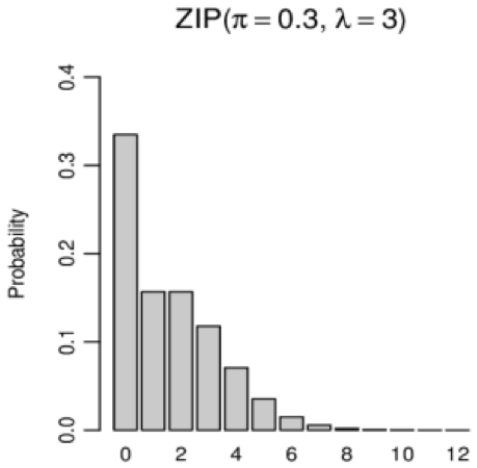
포아송 회귀 모형

포아송 회귀 모형의 문제점

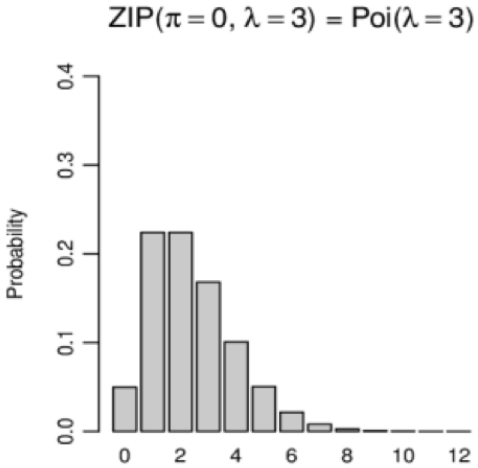
과대영 문제 (Excess Zeros)

포아송 분포에서 예상된 0 발생 횟수보다

실제로 더 많은 0이 발생한 경우



과대영 문제가 발생한 경우



과대영 문제가 발생하지 않은 경우

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 **영 과잉 포아송 회귀 모형**이나
영 과잉 음이항 회귀 모형을 통해 해결 가능

이번 클린업에선 영 과잉 포아송 회귀 모형만을 다룸!

포아송 회귀 모형

포아송 회귀 모형의 문제점

영 과잉 포아송 회귀모형 (ZIP, Zero Inflated Poisson Regression)

0의 값만 갖는 **점 확률분포**와

0 이외의 값을 갖는 **포아송 분포**가 결합된 **혼합 분포구조**

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

- Y: 베르누이 분포
- ϕ_i: 0이 발생할 확률
- $g(y_i)$: 0 이상의 정수 값
- $1-\phi_i$: 0 이상의 정수 값이 발생할 확률

포아송 회귀 모형

포아송 회귀 모형의 문제점

영 과잉 포아송 분포를 GLM으로 표현

0값이 발생할 확률(ϕ_i)을 로짓 연결함수를 이용하여 표현

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

포아송 분포의 평균(λ)을 로그 연결함수를 이용하여 표현

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

THANK YOU