Categorical Data Analysis

[2주차 클린업 교안]

👏 열화와 같은 반응에 힘입어 범주 클린업 2주차가 돌아왔습니다! 👏

이번 2주차는 일반화 선형 모형 GLM에 대해 자세히 알아보도록 해요.

어렵거나 모르는 부분이 있을 때 망설임 없이 저를 찾아와 주세요

그럼 2주차 클린업 시작해볼까요!

(2주차도 우리 모두 아자아자 파이팅~!)

임지훈 010-3499-4655

목차

임지훈 010-3499-4655

Ⅰ. GLM (Generalized Linear Model)

회귀분석에서 배운 선형회귀모델 (Linear Model)에 대해 떠올려보자. 선형회귀모델은 최소제곱법(LSE)를 통해 연속형 반응변수와 설명변수들 간의 최적의 선형관계식을 추정한다. 하지만 우리가 다루는 데이터의 반응변수가 항상 연속형 변수인 것은 아니다. 반응변수가 범주형이거나 도수 자료인 경우, 일반 선형회귀모델을 쓰는 것이 부적합하다. 그렇다면 어떻게 선형회귀모델의 한계점을 극복할 수 있을까?

1. GLM이란?

1) 정의

일반화 선형 모형(GLM)이란 연속형 반응변수들에 대한 모형(선형회귀모형)을 확장시켜 다양한 형태의 반응변수에 대한 모델들도 모두 포함한 더 넓은 범위의 모형들의 집합이다.

2) 필요성

(1) 정규분포를 포함한 다양한 확률분포를 사용할 수 있다.

일반 선형회귀모델은 4가지의 기본 가정 (독립변수와 종속변수 간의 선형성, 오차항의 정규성, 독립성 그리고 등분산성)을 만족해야 한다. (자세한 내용은 회귀팀 2주차 클린업 참고!) 하지만 우리가 다루는 범주형 변수는 오차항이 정규분포를 따르지 않기 때문에 최소제곱법(LSE)를 적용하는 것이 어렵다. 하지만 GLM은 최소제곱법 (LSE) 대신 **최대가능도법(MLE)**를 이용하여 모형을 적합하기 때문에 오차항의 정규분포 가정이 더 이상 필요 없어진다. 이처럼 정규분포 외의 다른 확률분포를 따르는 반응변수를 분석하려면 일반화 선형 모형(GLM)을 이용해야 한다.

(2) 변수 간의 연관성을 파악하고 반응변수를 예측할 수 있다.

GLM은 분할표 분석과 비교했을 때도 큰 이점을 갖는다. 분할표는 범주형 변수들 간의 연관성만을 파악할 수 있었던 반면, GLM은 범주형 변수들 간의 연관성 뿐 아니라 범주형 변수와 연속형 변수 간의 연관성 역시 분석 가능하다. 또한, 분할표는 주어진 자료에 대해서만 분석할 수 있었다면, GLM은 새로운 설명변수 값들이 주어졌을 때 그 설명변수에 따른 반응변수 값을 예측할 수 있다.

2. GLM의 구성 성분

GLM 구성 성분		
① 랜덤 성분	② 연결 함수	③ 체계적 성분
$\boldsymbol{\mu}\ (= E(Y))$	$\boldsymbol{g()}$	$\boldsymbol{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}$

GLM은 위의 표와 같이 랜덤 성분, 연결 함수 그리고 체계적 성분 총 3가지의 구성요소로 이루어진다. 아래의 일반적 형태를 확인하고, 각 구성요소에 대해 자세히 알아보도록 하자.

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

1) 랜덤 성분 (Random Component)

랜덤 성분은 반응변수 Y의 확률분포를 정해줌으로써 Y를 정의한다. 랜덤성분은 가정한 Y의 확률분포의 기댓값인 μ로 표기한다. 결국 우리가 알고 있는 Y를 확률분포가 무엇인지에 따라 표기만 다르게 하는 셈이다. 이 때 반응변수 Y의 관측값들은 서로 독립이라고 가정한다.

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

하지만 모든 확률분포를 GLM의 반응변수의 분포로 활용할 수 있는 것이 아니라, 지수족에 해당하는 확률분포만을 사용할 수 있음은 주의해야 한다. (이 내용은 [부록]에서 간단히!)

2) 체계적 성분 (Systematic Component)

체계적 성분이란 설명변수 X들을 명시하는 성분으로, X들의 선형결합의 형태로 표현한다. 주로 일반화 선형모형 (GLM)의 오른쪽 항에 위치한다. ($\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$)

체계적 성분은 **교호작용을 설명하는 항 ($x_i = x_a x_b$)**이나 **곡선효과를 나타내는 항 ($x_i = x_a^2$)**을 포함할 수 있다.

3) 연결 함수 (Link Function)

연결함수 g()는 랜덤 성분과 체계적 성분을 연결하여 두 성분의 범위를 조정하는 역할을 한다.

만약 랜덤 성분이 이항분포를 따르는 이진변수이고, 설명변수들이 연속형 변수라고 가정해 보자. 이 경우 체계적 성분 (우측 항)의 범위는 $-\infty$~∞를 따르기 때문에 랜덤 성분 (좌측 항)의 범위와 일치하지 않는다. 따라서 연결 함수를 통해 양변의 범위를 일치시키는 과정이 필요한 것이다.

GLM에서 사용되는 대표적인 연결 함수의 종류는 아래와 같다.

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

선형회귀모형은 랜덤 성분과 체계적 성분의 범위가 $-\infty$~∞ 로 일치하기에 연결함수가 필요 없다고 생각할 수 있지만, 이 경우에는 위의 '항등 연결 함수'에 의해 연결되어 있다고 생각하면 된다. 따라서 선형회귀모형도 GLM의 일종으로 속한다고 말할 수 있는 것이다.

3. GLM의 특징

1) 오차항의 다양한 분포 가정 가능

GLM은 정규분포 외에도 반응변수의 오차항이 가진 성질에 따라 어느 분포든 정의할 수 있다. 오차항의 확률분포가 무엇인지에 따라 어떤 연결함수를 사용하는지는 일반적으로 정해져 있다. (GLM의 종류에서 확인!)

2) 선형 관계식 유지

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

GLM의 체계적 성분은 선형성을 유지하고 있어 해석이 용이하다는 장점이 있다. 여기서 '선형'의 관계는 설명변수 X와 반응변수 Y간의 선형 관계를 말하는 것이 아닌 **회귀 계수 β**의 선형성을 가리킨다는 점을 주의해야 한

다. 따라서 $x_i = x_a x_b$나 $x_i = x_a^2$와 같이 교호작용과 곡선효과를 나타내는 항이 체계적 성분에 포함되어 있더라도 선형성이 유지되는 것이다.

3) 독립성 가정만 필요

선형회귀모형과 달리 GLM은 4가지 기본 가정 (정규성, 선형성, 독립성, 등분산성) 중 오차항에 대한 독립성만 만족하면 된다. 이를 위해서는 반응변수 간의 자기상관성을 검정해보아야 한다. 자기상관성(Autocorrelation)이란, 반응변수의 각 관측치가 상호 연관성을 띄는 것을 의미하며 시간이나 공간에서 연속된 일련의 관측치들 간 존재하는 상관관계를 의미한다. 일반적으로 회귀분석 후 더빈-왓슨 검정(DW Test)을 통해 자기상관성(혹은 오차의 독립성)을 검정한다. (자세한 내용은 회귀분석팀 클린업 참고!)

4) 제한적인 범위를 지닌 반응변수도 사용 가능

GLM은 랜덤 성분 (좌변)과 체계적 성분 (우변)의 범위가 상이하더라도 연결함수를 통해 일치시킬 수 있기 때문에 제한된 범위를 가진 반응변수(범주형 자료, 도수자료)도 반응변수로 삼을 수 있다.

결국, GLM은 **랜덤성분의 분포**와 **랜덤성분의 함수(연결함수)**를 일반화함으로써 기존의 선형회귀모형의 개념을 확장시켰다고 할 수 있다. 기존에는 설명변수들과 반응변수 간의 선형성만을 가정했다면, 연결함수를 통해 다양한 형태의 함수들이 등장한다. 또한 오차항의 정규성이 다양한 형태의 분포로 확장된 것이다.

4. GLM의 모형 적합

모형 적합(Model fitting)이란 주어진 데이터를 근거로 모형의 모수를 추정하는 과정이다. 모수란 모집단 분포의 특성을 규정짓는 척도로서 일반적으로 알려져 있지 않은 미지의 상수이다. 이 모수는 통계적 추론을 통해 표본의 특성을 근거로 **추정**하는데, 선형회귀모형에서 활용한 추정법이 바로 LSE(최소제곱추정법 : Least Square Estimation)다. 하지만 LSE는 잔차의 제곱합을 최소화하는 추정법으로 회귀의 4가지 기본 가정을 충족해야 하나, GLM은 해당 가정들을 모두 만족시키지 못하기에 사용할 수 없는 것이다.

따라서 GLM은 LSE 대신 **최대가능도추정법(MLE : Maximum Likelihood Estimation)**을 활용하여 모형을 적합시킨다. 가능도(Likelihood)란, 고정된 관측값이 어떤 확률분포를 따를 가능성을 의미한다. 이 가능도는 주어진 데이터를 후보 분포에 대입하여 얻은 값(분포 곡선의 높이)으로 수치화할 수 있는데, 각 데이터들의 가능도들을 모두 곱한 식을 아래와 같은 가능도 함수(Likelihood Function)이라고 부른다.

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

위 가능도함수가 최대가 되도록 하는 추정량 $\hat{\theta}$을 찾는 것이 **최대 가능도 추정법**이다. 그리고 **최대 가능도 추정량** (MLE : Maximum Likelihood Estimator)은 최대 가능도 추정법을 통해 얻은 추정량이다. 왼쪽의 식은 주어진 데이터들을 후보 확률분포에 대입하여 얻은 높이를 모두 곱했다는 뜻이고, 이 식을 최대화하는 최대 가능도 추정량 $\hat{\theta}$를 더 쉽게 찾을 수 있도록 Log 변환 시킨 형태가 오른쪽 식이다. Log 변환된 로그가능도함수를 편미분하여 0이 되도록 하는 방정식의 해가 곧 우리가 원하는 $\hat{\theta}$가 될 것이다. 이러한 MLE의 과정을 통해 GLM 모형 적합을 진행한다.

MLE를 이해하기 위해 간략한 예시 하나만 살펴보자. 표본크기 n의 확률표본 $X_1, \dots, X_n$이 모수가 λ인 지수분포 ($f_x(x) = \lambda e^{-\lambda x}$)로부터 추출되었다고 가정하자. 여기에 n개의 표본들을 대입하여 얻은 높이를 곱한 가능도함수에 로그변환을 진행하면 아래와 같이 변환된다.

$$L(\lambda\,;\, x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \quad \xrightarrow{\text{Log}} \quad lnL(\lambda) = nln\lambda - \lambda \sum_{i=1}^{n} x_i$$

이 때 최대 가능도 추정량 λ를 구하기 위해 편미분하여 값이 0이 되도록 하는 λ를 추정해본다.

$\frac{dlnL(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$을 성립시키는 최대가능도추정량 $\hat{\lambda}$는 $\frac{1}{\bar{X}}$이 된다. (위 MLE 과정에 대한 배경과 설명은 매우 간추린 것이니 이 점 참고해주세요!)

II. 유의성 검정

유의성 검정이란 모형의 모수에 대한 추정값이 유의한지 혹은 축소 모형의 적합도가 좋은지를 판단하는 검정이다.

회귀분석에서는 F-검정을 통해 회귀모델의 유의성 검정을 진행 하고, T-검정을 통해 각 회귀계수 β의 유의성을 검정하였다. 그렇다면 GLM에서는 어떻게 유의성을 검정하는지 아래에서 확인해보자.

1. 유의성 검정의 가설

GLM모형, $g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$의 유의성 검정 가설은 다음과 같다.

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

H_1 : 적어도 하나의 β는 0이 아니다.

귀무가설에 따르면 모든 모수 값이 0이므로 해당 GLM모형이 분석 가치가 없다는 의미이다. 반면 대립가설은 적어도 하나의 모수 값이 0이 아니므로 모형이 유의하다는 것을 의미한다. (모형 자체의 유의성을 검정하는 회귀분석의 F-검정과 같은 맥락!)

2. 유의성 검정의 종류

1) 왈드 검정 (Wald test)

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

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

왈드 검정은 회귀 계수에 대한 추정값과 표준오차만 사용하여 통계량을 구한다는 점에서 장점을 갖지만, 범주형 자료이거나 소표본인 경우 검정력이 감소된다는 단점이 있다. 따라서 GLM의 유의성 검정은 다음의 가능도비 검정을 주로 사용한다.

2) 가능도비 검정 (Likelihood-ratio test)

① 검정통계량 : $G^2 = -2\log\left(\frac{l_0}{l_1}\right) = -2(L_0 - L_1) \sim \chi^2{}_{df}$

② 기각역 : $G^2 \geq \chi_{\alpha,df}^2$

검정통계량은 귀무가설 하에서의 가능도함수(l_0)와 전체공간 하에서의 가능도함수(l_1)의 차이를 이용한다. 여기서 전체공간이란 모수에 대한 아무런 제약이 없는 상태로, 귀무가설+대립가설 상태를 의미한다.

LR test는 두 가능도 함수의 최대값을 비교하는 방식으로 검정통계량을 구하는데, 이 때 전체공간 하에서의 가능도함수(l_1)의 최대값은 **MLE를 통해 계산한 값**이다. 즉, 검정통계량 식을 풀이하면 아래와 같다.

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

이 때 가능도 함수 l_0이 최대값을 갖도록 하는 모수 추정량을 θ_0라고 하자. 만약 두 가능도 함수 간의 차이가 작다면 $\frac{l_0}{l_1}$은 1에 가까워진다. 이는 θ_0가 MLE(최대 가능도 추정량)와 근사한 값을 갖는다는 뜻으로, 귀무가설 하의 θ_0가 충분히 모수를 설명할 수 있다는 의미다. 반대로 두 가능도 함수 간의 차이가 크면 $\frac{l_0}{l_1}$가 1보다 작아져 $-2\log\left(\frac{l_0}{l_1}\right)$ 값은 기각역에 해당하는 큰 값을 가져 귀무가설을 기각하게 된다. 이 때 자유도(df)는 두 가설 (H_0, H_1)의 모수 개수의 차이를 의미한다.

즉, 검정 flow를 정리해보면,

① l_0와 l_1의 차이가 크다 → ② 검정통계량이 큰 값을 가진다 → ③ 작은 p-value값을 지닌다. (기각역에 속함) → ④ 귀무가설 기각, 적어도 하나의 β는 0이 아니다. → ⑤ 모형의 모수 추정값은 유의하다.

(1주차에서 가능도비 검정에 대해 알아보았던 것이 기억나나요? 1주차에서는 독립성 검정을 위해 가능도비를 사용하였다면 이번 2주차에서는 유의성 검정에 이용합니다. 따라서 가설과 검정통계량에서 차이를 보이지만, 검정 flow는 유사하답니다!)

가능도비 검정은 귀무가설 하의 가능도 함수(l_0)와 전체공간 하의 가능도 함수(l_1)에 대한 정보를 모두 사용한다. 따라서 타 유의성 검정에 비해 가장 많은 정보를 이용한다. 그래서 왈드 검정에 비해 검정력과 신뢰도에서 높은 강점을 보인다.

3. 이탈도 (Deviance)

이탈도를 이용해 우리가 관심있는 모형(M)을 포화모형(S)와 비교하여 모형의 적합성을 판단한다.

1) 관심모형과 포화모형

① 관심모형 (M)

관심모형이란 우리가 관심을 갖는 모형, 즉 유의성을 검정하고자 하는 모형을 뜻한다.

② 포화모형 (S)

포화모형이란 관측값들에 대하여 완벽하게 자료에 적합하는 모형, 즉 모든 관측값에 대하여 모수를 갖는 가장 복잡한 모형을 뜻한다.

예를 들어 관측값으로 범주팀의 행복(Y), 클린업 시간(x_1), 패키지 난이도(x_2), 교안 페이지 수(x_3), 마지막으로 범주 팀원들의 귀여움(x_4)이 주어졌다고 가정해보자.

이 중 우리는 클린업 시간(x_1)과 패키지 난이도(x_2)가 반응변수에 미치는 영향을 알고 싶다면, 관심모형과 포화모형은 아래와 같을 것이다.

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

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

이에 따른 유의성 검정을 위한 이탈도의 귀무가설(H_0)과 대립가설(H_1)은 다음과 같다.

H_0 : 관심모형(M)에 속하지 않는 모수는 모두 0이다. (=관심모형을 사용)

H_1 : 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다. (=관심모형 사용 불가)

귀무가설이 맞다면 관심모형만이 데이터를 잘 설명하고 있으므로 (다른 변수들은 모수가 0이므로 반응변수에 영향력이 없기 때문에) 관심모형을 사용한다. 반면 대립가설이 맞다면 관심모형 외에도 데이터를 잘 설명할 수 있는 모형이 있다는 뜻이므로 관심모형을 사용할 수 없다.

2) 이탈도(Deviance)

이탈도란 포화모형(S)와 관심모형(M)을 비교하기 위한 가능도비 통계량으로 식은 아래와 같다.

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

이탈도 역시 관심모형(M)에서 얻은 로그 가능도 함수의 최댓값(L_M)과 포화모형(S)에서 얻은 로그 가능도 함수의 최댓값(L_S)의 차이를 이용하여 계산한다.

이탈도를 사용하기 위해선 한 가지 조건이 있다. 이탈도는 포화모형에는 있지만 관심모형에는 없는 계수들이 0인지의 여부를 확인하는 것이기에 관심모형은 포화모형에 **내포된(nested)** 관계여야 한다.
(=(M의 모수 ⊂ S의 모수)

즉, 이탈도의 검정 flow를 정리해보면,

① 두 가능도 함수의 최댓값 간의 차이가 크다 → ② 이탈도가 크다→ ③ 작은 p-value값을 지닌다.
→ ④ 귀무가설 기각→ ⑤ 관심모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다
→ ⑥관심모형 M이 적합하지 않다, 관심모형 M 사용 불가능

이탈도 역시 앞선 가능도비 검정과 검정통계량의 형태와 검정 flow가 매우 유사하다.

4) 이탈도와 가능도비 검정의 관계

가능도비 검정 통계량은 두 모형 간의 이탈도 값의 차이와 같다. M_0는 단순한 형태의 관심모형, M_1은 복잡한 형태의 관심모형 그리고 S를 두 모형을 모두 포함한 포화모형이라고 가정하자. 이 때 두 모형(M_0, M_1)의 이탈도의 차이를 구해보면 아래와 같이 정리된다.

M_0의 이탈도 - M_1의 이탈도 (=모형 간의 이탈도의 차이)

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

$$= -2(L_0 - L_1)$$ (=가능도비 검정 통계량)

위의 식에 따르면 두 모형 간 이탈도의 차가 가능도비 검정 통계량과 동일하다. 이 성질을 통해 두 모형의 이탈도 차이를 구하면 어느 모형이 더 좋은 지 판단할 수 있다. (쉽게 말해 관심모형 vs 관심모형 간의 비교를 하는 것이다)

하지만, 이 과정에서도 이탈도를 활용하기 때문에 **모형 M_0은 모형 M_1에 내포된(nested)된 모형**이어야 한다. 만약 여러 모형을 비교하고 싶은데 내포된 경우가 아니라면, AIC, BIC와 같은 모형 선택(variable selection)을 위한 측도들을 활용하여 모형을 비교한다. (자세한 내용은 회귀팀 클린업 참고!)

즉, 이탈도의 차이를 이용한 두 관심모형을 비교하는 검정 flow를 정리해보면,

① 관심모형(M_0, M_1) 간 이탈도의 차이가 작다. → ② 가능도비 검정 통계량이 작다. → ③ 큰 p-value값을 지닌다.
→ ④ 귀무가설 기각하지 못한다. → ⑤ M_0에 포함되지 않는 모수들을 모두 0이다.
→ ⑥ 간단한 관심모형인 M_0이 더 적합하다.

III. GLM 모형의 종류

1) GLM 모형의 종류

모형들을 자세히 알아보기에 앞서 GLM 모형 전체를 (물론 이 역시 일부다…) 확인해보자. 앞서 랜덤성분이 어떤 확률분포를 따르는 지에 따라 어떤 연결함수를 활용하는지 일반적으로 정해져 있다고 언급했다. GLM의 구성요소에 따라 어떤 밑의 표와 같이 종류가 매우 다양한데, 이 중 중요한 모형인 색칠된 모형들에 선택과 집중 해보도록 하자!

<table>
<tr><th>GLM</th><th>랜덤성분</th><th>연결함수</th><th>체계적 성분</th></tr>
<tr><td>일반 회귀 분석</td><td rowspan="3">정규 분포</td><td rowspan="3">항등</td><td>연속형</td></tr>
<tr><td>분산 분석</td><td>범주형</td></tr>
<tr><td>공분산 분석</td><td>혼합형</td></tr>
<tr><td>선형 확률 모형</td><td rowspan="3">이항 자료</td><td>항등</td><td rowspan="7">혼합형</td></tr>
<tr><td>로지스틱 회귀 모형</td><td>로짓</td></tr>
<tr><td>프로빗 회귀 모형</td><td>프로빗</td></tr>
<tr><td>기준범주 로짓 모형</td><td rowspan="4">다항 자료</td><td rowspan="4">로짓</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>

2) GLM 모형의 식

위의 모형들의 랜덤 성분이 어떤 확률분포를 따르는 지에 따라 분류하여 나열한 모형별 식이다. 이 중 볼드체 되어 있는 모형은 이번 클린업을 통해 자세히 알아볼 모형들이다. (식만 보면 어려워 보이겠지만, 이번 클린업을 듣고 나면… 쉬울..지도?)

① 반응변수 : 이항 자료

- 선형 확률 모형

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

구성 : 이항 랜덤 성분과 항등 연결 함수

- **로지스틱 회귀 모형**

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

구성 : 이항 랜덤 성분과 로짓 연결 함수.

- 프로빗 회귀 모형

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

  구성 : 이항 랜덤 성분과 프로빗 연결

② 반응변수 : 다항 자료

- **기준 범주 로짓 모형**

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

  구성 : 다항 랜덤 성분(명목형)과 로짓 연결 함수

- **누적 로짓 모형**

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

  구성 : 다항 랜덤 성분(순서형)과 로짓 연결 함수

- 이웃범주 로짓 모형

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

  구성 : 다항 랜덤 성분(순서형)과 로짓 연결 함수

- 연속비 로짓 모형

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

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

  구성 : 다항 랜덤 성분(순서형)과 로짓 연결 함수

③ 반응변수 : 도수 자료

- **포아송 회귀 모형**

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

구성 : 포아송 랜덤 성분과 로그 연결 함수

- **음이항 회귀 모형**

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

구성 : 음이항 랜덤 성분과 로그 연결 함수

- **율자료 포아송 회귀 모형**

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

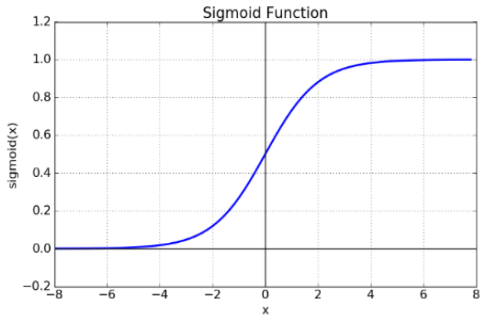
구성 : 포아송 랜덤 성분과 로그 연결 함수

IV. 로지스틱 회귀 모형

1. 로지스틱 회귀 모형이란?

로지스틱 회귀(Logistic Regression)이란 반응변수 Y가 이항자료일 때의 회귀이다. 즉, 반응변수는 이항분포를 따르게 되고, 성공일 확률로 표기될 것이다.

1) 형태



로지스틱 회귀 모형은 왼쪽 그림과 같이 S자 곡선 형태를 띈다. 해당 곡선은 $\pi(x)$와 x의 비선형 관계를 나타내며, 이와 같은 그래프의 형태를 시그모이드(Sigmoid) 형태의 함수라고 일컫는다.

즉, 로지스틱 회귀 모형은 일반선형모델로 설명할 수 없는 이항변수와 연속형 변수들 간의 관계를 GLM의 형태로 표현한 것이다.

2) 로지스틱 회귀모형의 장점

① 이항변수와 연속형 변수 간의 범위 일치

이항분포를 따르는 반응변수 $\pi(x)$ $(= P(Y = 1|X = x))$는 0~1 사이의 확률값 (Y가 성공할 확률)을 가질 것이고, 설명변수의 선형식은 $-\infty$~∞ 사이의 연속적인 값을 가져 양 변의 범위가 맞지 않는 문제가 생긴다. 따라서 연결함수를 통해 양 변의 범위를 맞춰주는 과정이 필요한데, 과정은 아래와 같다.

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

좌변 범위 : 0~1 ≠ 우변 범위 : $-\infty$~∞

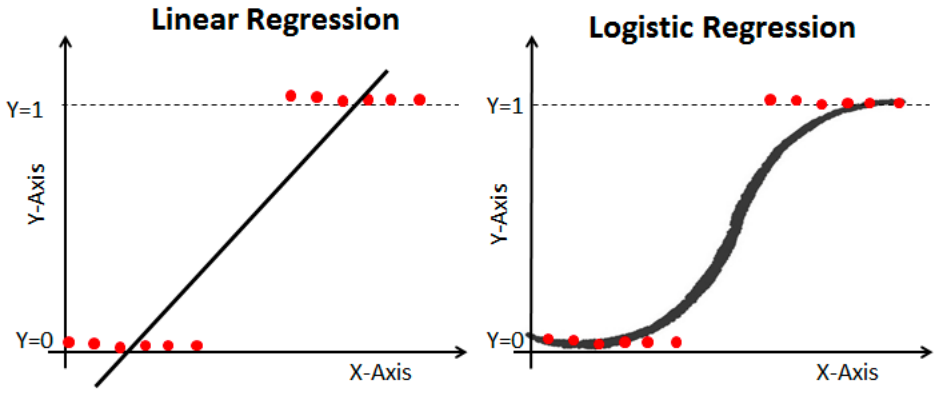
↓

좌변을 오즈 형태로 만든 뒤 로그 취하기! (로짓 연결함수)

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

좌변 범위 : $-\infty$ ~∞ = 우변 범위 : $-\infty$~∞

선형식을 이항변수 Y에 맞게 조정하는 위 과정을 그림을 통해 확인하면 다음과 같다.



로지스틱 회귀모형(오른쪽)은 일반선형모델(왼쪽)과 달리 0~1 사이의 범위를 벗어나는 값이 없다. 결국 로지스틱 회귀모형은 이항분포를 따르는 랜덤성분을 로짓 연결함수를 통해 체계적 성분과 연결한 형태의 GLM인 것이다! 이로써 연속형 변수를 통해 이항분포를 따르는 반응변수를 해석할 수 있게 되었다.

② 기본가정의 완화

범위 일치 이외에도 로지스틱 회귀모형은 가정으로부터 자유롭다는 장점을 갖는다. 일반회귀모형은 4가지 기본 가정(정규성, 등분산성, 선형성, 독립성)을 모두 만족해야 했다면 GLM의 일종인 로지스틱 회귀모형은 오직 독립성 가정만 만족하면 된다. 애초에 이항분포를 따르는 반응변수 자체가 정규성과 등분산성을 따르는 것은 불가하다. (이항분포의 분산 Var(x) = nπ(x)(1 − π(x))이 x에 대한 식으로 표현되기 때문에 등분산성 조건 성립이 불가한 것은 당연하다)

③ 후향적 연구 활용

새로운 데이터가 주어졌을 때 이를 통해 새로운 반응변수를 예측하는 것만 가능했던 일반선형모델과 달리 로지스틱 회귀 모형은 오즈비를 활용(바로 다음 챕터!)하므로 후향적 연구 분석에도 활용 가능하다.

3) 기울기

로지스틱 회귀 모형을 x에 대해 미분하면 아래와 같이 x에서의 접선의 기울기를 구할 수 있다.

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

(계산식은 부록 참고!)

로지스틱 회귀 모형의 기울기는 위의 식에서 알 수 있듯이 모수 β의 영향을 받는다. β가 양수 값을 가지면 상향곡선, β가 음수 값을 가지면 하향곡선의 형태를 띌 것이고, β의 절댓값이 클수록 기울기 변화율이 크기 때문에 더 가파른 형태의 로지스틱 회귀 모형의 형태를 띈다.

2. 로지스틱 회귀 모형의 해석

로지스틱 회귀 모형은 확률을 통해 해석하거나, 오즈를 이용하여 해석할 수 있다.

1) 확률을 통한 해석

로지스틱 회귀 모형을 변형하면 아래와 같이 확률에 관한 식으로 달리 표현할 수 있다.

$$\pi(x)=\frac{\exp\left(\beta_0+\beta_1x_1+\cdots+\beta_px_p\right)}{1+\exp\left(\beta_0+\beta_1x_1+\cdots+\beta_px_p\right)}$$

각 x_p를 대입하면 Y(반응변수)=1일 확률, 즉 $\pi(x)$ 값을 알 수 있다. 해당 $\pi(x)$ 값이 분석자가 지정한 cut-off point보다 크면 Y=1로, 작으면 Y=0로 예측한다.일반적으로 cut-off point는 0과 1의 가운데 값인 0.5로 설정하나, 이는 상황에 따라 다르다. (자세한 내용은 3주차 클린업에서!)

예를 들어, 키(cm)에 따른 연애 여부(Y)에 관한 로지스틱 회귀 모형이 $\log\left[\frac{\pi(x)}{1-\pi(x)}\right]=4+0.08x$라고 가정하자. 이를 확률에 관한 식으로 변형하면 아래와 같다.

$$\pi(x)=\frac{\exp\left(4+0.08x\right)}{1+\exp\left(4+0.08x\right)}$$

만약 관측치 중 하나가 174cm라면, 해당 관측치의 연애 확률 $\pi(x)$은 x에 174를 대입하여 계산한다.

$\pi(174) = \frac{\exp(4+0.08*174)}{1+\exp(4+0.08*174)}$ = 0.99 > 0.5 (cut-off point는 0.5로 가정) 이므로, 해당 관측치는 연애를 하고 있다($\hat{Y}$=1)고 판단한다.

2) **오즈비를 통한 해석**

로지스틱 회귀 모형은 로짓(로그오즈) 연결함수를 이용하였기 때문에 오즈를 이용하여 해석하는 것도 가능하다. 이 때 모형에 각각 x와 x+1을 대입한 두 값을 서로 빼면 오즈비 형태로 나온다. 아래 수식을 살펴보자.

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

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

$$\frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta} \cdots ③$$

마지막 ③번 식에서 분자는 설명변수가 x+1일 때 Y=1일 오즈이고 분모는 설명변수가 x일 때 Y=1일 오즈이다. 즉, 좌변은 오즈비이다. 이 오즈비가 e^{β} 값을 가지므로, 'x+1일 때 Y=1일 오즈가 x일 때 Y=1일 오즈보다 e^{β}배 높다'고 해석할 수 있다. 즉, 다른 설명변수가 고정된 상황에서 x가 1 단위 증가할 때마다 Y=1일 오즈가 e^{β}배 증가한다고 해석할 수 있다.

1)에서 살펴본 예시를 오즈비를 활용하여 해석해보자. 해당 로지스틱 회귀 모형 $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 4 + 0.08x$에서 키(cm)가 1 증가할 때마다 연애 중일 오즈가 $e^{0.08}$=약 1.08배 증가한다고 해석할 수 있다.

V. 다범주 로짓 모형 (multicategory Logit Model)

앞서 배운 로지스틱 회귀 모형이 이진 분류 문제를 다루는 모형이었다면, 다범주 로짓 모형은 3개 이상의 범주를 가진 반응변수로 확장시킨 모형이다. 다범주 로짓 모형은 로지스틱 회귀 모형과 동일하게 **로짓 연결 함수**를 사용하지만, **랜덤성분은 다항분포**를 따르는 차이점을 갖는다.

이 때, 반응변수의 범주가 3개 이상으로 늘어났기 때문에, 범주형 자료가 명목형 자료인지, 순서형 자료인지 구분할 필요가 있다. 자료의 성격에 따라 서로 다른 모델을 적용하는데, 우선 명목형 자료가 반응 변수로 주어졌을 때 어떤 모형을 사용하는지 알아보자.

1. 기준 범주 로짓 모형 (Baseline-Category Logit Model)

1) 기준 범주 로짓 모형의 정의

기준 범주 로짓 모형은 반응변수가 명목형 자료일 때 사용하는 다범주 로짓 모형의 일종이다. 기준 범주 로짓 모형은 반응 변수의 여러 개의 범주 중 마지막 범주를 기준 범주(Baseline-Category)로 선택한 뒤, 기준 범주와 나머지 범주들을 짝지어 로짓을 정의한다. 반응 변수가 총 J개의 범주로 구성되었다고 가정했을 때, 기준 범주 로짓은 아래와 같다.

$$\log\left(\frac{\pi_j}{\pi_J}\right), j = 1, \dots, J-1$$

분자(π_j)는 j번째 범주에 속할 확률이고 분모(π_J)는 기준 범주인 J번째 범주에 속할 확률을 의미한다. 위 식이 기준 범주 로짓 모형의 좌변을 구성하게 되고 완전한 GLM 모형은 다음과 같다.

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

j : 범주에 대한 첨자 / J : 기준 범주에 대한 첨자 / 1~K : 설명변수에 대한 첨자

위 식에서 β는 총 2개의 첨자를 가진다. 윗 첨자는 총 K개의 설명변수가 있을 때 각 설명변수 별로 회귀계수 β를 구분한다. 아래 첨자는 β가 기준 범주 J와 그 외 어떤 범주 간의 관계식인지를 표시한다. 즉, 기준 범주 J와 그 외의 범주 J-1개의 범주를 각각 비교하기 위해 **총 J-1개의 로짓 방정식**을 갖는다. 따라서 회귀 계수 β의 아래 첨자를 확인하면, 우리는 기준 범주 J와 어떤 범주 간의 로짓 방정식인지 유추할 수 있는 것이다. (만약 J=2라면 총 2개의 범주를 가진 로지스틱 회귀 모형을 의미하겠죠?)

로짓으로 표현한 위의 모형을 확률에 대한 식으로 재정의하면 아래와 같다.

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

예시를 통해 이해해보자. 반응변수 Y가 3개의 범주 (하니, 해린, 민지)를 갖는다고 가정하자. 이 때 반응변수 Y를 가장 좋아하는 멤버라 정의하고, 기준 범주를 민지로 삼아보자. 그렇다면 기준 범주 로짓모형은 3-1=2개가 만들어진다.

$$\log\left(\frac{\pi_{하니}}{\pi_{민지}}\right) = 5 + 0.27x_1 + \cdots + 0.59x_K$$

$$\log\left(\frac{\pi_{해린}}{\pi_{민지}}\right) = 2 + 0.22x_1 + \cdots + 0.46x_K$$

두 식의 체계적 성분을 확인해보면, 같은 설명변수 이더라도 회귀계수 β가 다른 값을 가지는 것을 알 수 있다.(물론 우.연.히 같을 수도 있음!) 각 기준 범주 로짓 모형의 좌변을 확률에 대한 식으로 재정의하여 해린이 가장 좋아하는 멤버일 확률을 구해본다면, 식은 아래와 같이 정리될 수 있다.

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

2) 해석

기준 범주 로짓 모형은 오즈와 기준 범주를 이용하여 모형을 해석하는데, 어떤 범주들 간의 의미를 해석하는가에 따라 2가지 방법이 있다.

① j범주와 J범주(기준 범주) 비교

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

다른 설명변수들이 고정되어 있을 때, x_i가 한 단위 증가하면 J범주 대신 j범주일 **오즈가 e^{β_j}배 증가**한다고 해석한다. 위 예시를 이용하면, 하니와 민지의 기준 범주 로짓 모형에서 다른 설명변수들이 고정되어 있을 때 x_1이 한 단위 증가하면 민지보다 하니를 좋아할 오즈가 $e^{0.27}$ = 약 1.31배 증가한다고 해석 할 수 있다.

② a범주와 b범주 비교

기준 범주 J가 아닌 서로 다른 두 범주 간의 관계를 해석할 수 있다. 각 범주의 기준 범주 로짓 모형을 빼주면 아래의 식을 통해 확인해보자.

$$\log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) = (\alpha_a + \beta_a^1 x_1 + \cdots + \beta_a^K x_K) - (\alpha_b + \beta_b^1 x_1 + \cdots + \beta_b^K x_K)$$

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

이 경우 다른 설명 변수들이 고정되어 있을 때 x_i가 한 단위 증가하면 b범주 대신 a범주일 오즈가 $e^{\beta_a - \beta_b}$ **배 증가한다고 해석한다.**

2. 누적 로짓 모형 (Cumulative Logit Model)

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

반응변수가 순서형 자료일 때 사용하는 순서형 다범주 로짓 모형 역시 기준 범주를 정하고 다른 범주들과 짝지어 비교하는 방식이다. 하지만 순서형 다범주 로짓 모형은 범주들 간의 순서 정보를 고려해야 하므로 범주를 순서대로 정렬시키고 두 덩어리로 나누는 Collapse 과정이 필요하다. 이 때 어떤 기준(cut point)로 collapse을 진행하는지에 따라 사용되는 모형이 결정된다. 아래 그림을 보자.

소형	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
소형	중형	대형	초대형
소형	중형	대형	초대형

왼쪽부터 각각 이웃 범주 로짓 모형, 연속비 로짓 모형, 누적 로짓 모형이다. 이 중에서 우리는 세 번째, **누적 로짓 모형**에 대해 집중적으로 다뤄보겠다. 누적 로짓 모형은 그림과 같이 전체 범주를 모두 사용한다는 특징을 갖는다.

1) 모형

누적 로짓 모형은 누적 확률에 로짓 연결함수를 사용하여 만들어진다. 반응변수가 총 J개의 범주를 가질 때 우선 첫 번째 범주부터 j번째 범주까지의 누적 확률은 나타내 보면 다음의 식과 같다.

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

위의 누적 확률을 로그 오즈의 형태로 조작하게 되면,

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

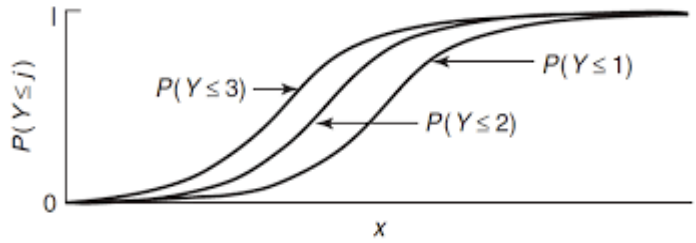
위의 식이 누적 로짓 함수의 좌변을 구성한다. 최종적으로 누적 로짓 모형의 형태는 아래와 같다.

$$logit[P(Y \le j|X=x)] = \ \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \ , j = 1, \dots, (J-1)$$

누적 로짓 모형은 기준 범주 로짓 모형과 유사하게 기준점을 설정하여 비교하는 원리이므로 총 J-1개의 로짓 방정식이 만들어진다.

반면, 체계적 성분에서 두 모형은 약간의 차이를 보인다. 회귀계수 β가 2개의 첨자를 가지고 있었던 기준 범주 로짓 모형과 달리 누적 로짓 모형의 β는 한 개의 첨자만을 가진다. 이 첨자는 각 설명변수별 회귀계수 β를 구분하는 역할을 한다. 반면 기준점과 어떤 범주 간의 로짓 방정식인지 구분해 주는 j 첨자는 α_j에는 있지만 β에는 붙어있지 않다. 그 이유가 무엇일까?

이는 누적 로짓 모형에서 모든 J-1개 로짓 방정식의 β가 동일하다고 가정하기 때문이다. 이것을 **비례오즈 가정(proportional odds)**라고 부른다. 즉 누적 로짓 모형에서 β는 어떤 범주를 j로 설정하는지에 영향을 받지 않고 오직 α값만 방정식에 따라 변화한다는 의미이다. 이를 그림으로 나타내면 아래와 같다.



위와 같이 서로 다른 로짓 방정식이 같은 기울기의 곡선 형태이지만 수평 이동을 한 것처럼 나타난다. (β는 같지만 α는 다르다는 뜻!)

예시를 통해 정확히 알아보자. 만약 범주팀 팀장에 대한 불만 지표가 적음/보통/많음/매우 많음, 총 4개의 범주로 구성되어 있을 때, 로짓 방정식들은 아래와 같이 나타날 것이다.

$$logit[P(Y \le \textit{적음})] = \ 8 + 0.07x_1 + \cdots + 0.6x_p$$
$$logit[P(Y \le \textit{보통})] = \ -5 + 0.07x_1 + \cdots + 0.6x_p$$
$$logit[P(Y \le \textit{많음})] = \ 12 + 0.07x_1 + \cdots + 0.6x_p$$

누적 로짓 모형은 총 3(4-1)개의 로짓 방정식을 갖고, 모두 같은 β값을 갖지만 α값은 로짓 방정식마다 상이하다.

2) 해석

누적 로짓 모형은 기준 범주 로짓 모형과 마찬가지로 오즈를 이용하여 해석한다.

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

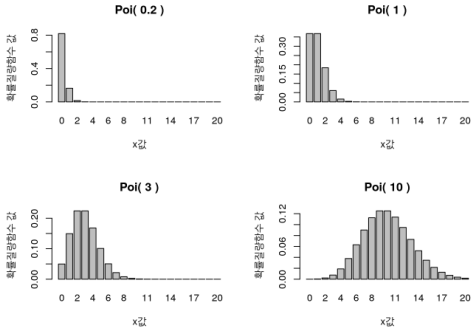
다른 설명변수가 고정되어 있다는 가정 하에, x_i가 한 단위 증가할 때 **$Y > j$에 비해 $Y \le j$일 오즈가 e^{β_i}만큼 증가**한다고 해석한다.

위의 범주팀장에 대한 불만도 예시에 따르면 어떤 범주인지와 무관하게 다른 설명변수가 고정되어 있을 때 x_1이 한 단위 증가하면 $Y > j$에 비해 $Y \le j$일 오즈가 $e^{0.07} =$ 약 1.07배 증가한다.

VI. 포아송 회귀 모형 (Poisson Regression Model)

지금까지 알아본 모델들을 간단히 복습해보자. 일반선형모형은 랜덤 성분이 정규분포를 따르고, 로지스틱 회귀 모형의 랜덤 성분은 이항분포를 따르고 마지막으로 다범주 로짓 모형의 랜덤 성분은 다항 분포를 따를 때 사용하는 GLM 모형이었다. 당연히 반응 변수가 도수 자료일 때, 즉 포아송 분포를 따를 때는 앞선 모형을 사용하는 것이 부적합하다.

포아송 회귀 모형은 반응변수가 포아송 분포를 따를 때 사용하는 GLM 모형이다. 이 때 포아송 분포란 단위 시간 안에 사건이 일어난 건수, 횟수를 표현하는 이산 확률 분포이다. 식과 그림은 아래와 같다.



$$f(x;\ \lambda) = \Pr\,(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}\ (x = 0,1, \ldots)$$

왼쪽 그림과 같이 포아송 분포는 모수 λ (평균) 값이 작을 때 편향된 분포를 갖지만 평균이 커질수록 정규분포와 유사한 모습을 띈다. (1주차 카이제곱 검정을 증명할 때 설명했던 것이 기억 나시나요?)

따라서 평균이 작아 편향된 포아송 분포의 경우 일반선형모델을 적용하는 것이 부적합하다. 정규성과 등분산성 가정에 어긋나기 때문이다. 이 경우 일반선형모델을 적용하면 표준오차나 유의수준이 편향되는 문제점을 초래할 수 있다. 이러한 한계점을 극복하기 위해 반응 변수가 도수 자료일 때 포아송 회귀 모형을 활용한다.

1. 포아송 회귀 모형

1) 형태

포아송 회귀 모형(Poisson Regression Model)은 반응변수가 도수 자료이며 랜덤 성분이 포아송 분포를 따르고 로그 연결함수를 사용하는 모형이다. 형태는 아래와 같다.

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

도수 자료를 나타내는 포아송 분포의 랜덤 성분은 $0 \sim \infty$ 사이의 값을 갖는 반면 체계적 성분은 $-\infty \sim \infty$ 사이의 값을 갖기 때문에 양 변의 범위를 일치시키기 위해 랜덤 성분을 로그 연결함수를 통해 체계적 성분과 동일한 범위를 갖도록 조정한다.

2) 해석

① 도수를 통한 해석

포아송 회귀 모형을 변형하여 μ(도수)에 관한 식으로 나타내 해석할 수 있다.

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

해당 식에 주어진 데이터의 설명변수들을 대입하면 μ(도수)에 대한 예측값, 즉 기대도수가 도출된다.

예를 들어 반응변수(Y)가 살면서 내가 로또 1등에 당첨될 횟수, 설명변수(X)는 내가 1년에 복권을 사용하기 위해 사용하는 금액이고 이에 대한 포아송 회귀 모형의 식이 $\log(\mu) = -2 + 0.0001x$라고 가정해보자. 만약 내가 1년에 복권을 사기 위해 만 원을 투자한다면, $e^{-2+0.0001*10000} = e^{-1} =$ 약 0.3679회가 나온다. 기대 도수가 1회 미만이므로 1년에 만원을 투자해서는 평생 로또 1등에 당첨될 수 없는 것이다.

② 오즈비를 통한 해석

포아송 회귀 모형을 오즈비를 이용하여 해석할 수 도 있다. 포아송 회귀 모형에 x+1과 x를 대입한 두 모형의 차이를 이용해 β에 관한 식을 얻을 수 있다.

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

다른 설명변수들이 고정되어 있다는 가정 하에, x가 한 단위 증가하면 **기대도수 μ가 e^{β}배만큼 증가한다**고 해석한다.

앞선 예시를 이용해보면, 1년에 복권을 사기위해 투자하는 돈이 1원이 늘어난다면, 로또 1등에 당첨될 기대도수는 $e^{0.0001}$배만큼 증가한다. ($e^{0.0001}$ = 거의 1에 가까운 수.. 전혀 차이를 만들지 못할 것이다)

2. 율자료 포아송 회귀 모형

포아송 회귀 모형은 설명변수들이 도수 자료(μ)에 미치는 영향을 표현하지만, 이는 시간, 공간 등의 요소의 차이를 반영하지 못한 채 μ 값의 예측값, 즉 기대도수 값 만을 산출한다. 이런 경우에는 비율자료를 사용하여 이러한 차이를 반영해 주어야 한다. 예를 들어 지역의 범죄 발생 건수는 그 지역의 인구 숫자에 큰 영향을 받기 때문에 우리는 비율자료를 통해 대소를 비교해야 한다.

1) 형태

율자료 포아송 회귀 모형은 앞선 포아송 회귀 모형과 같이 로그 연결함수를 사용하지만 기존의 μ값 대신 비율자료를 반응변수로 사용한다. 식으로 표현하면 아래와 같다.

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

여기서 t는 지표값을 의미하는데 비율자료를 구하기 위해 기준이 되는 값이다. 범죄 발생 비율의 경우 그 지역 인구의 모집단이 지표값이 된다. 위의 모형 역시 도수(μ)에 대한 식으로 변형해 표현할 수 있다.

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

2) 해석

율자료 포아송 회귀 모형도 오즈비를 이용하여 x+1과 x를 대입한 식 간의 차이를 이용해 해석할 수 있다.

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

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

즉, 다른 설명변수들이 고정되어 있다는 가정 하에 **기대비율이 e^{β}배 만큼 증가한다고 해석**한다. (이 때, 기대비율을 비교하고 있다는 점에 주의!)

3. 포아송 회귀 모형의 문제점

1) 과대산포 문제 (Overdispersion)

포아송 분포를 따르는 도수 자료는 등산포 가정, 즉 평균과 분산이 같다는 성질을 지닌다. 하지만 현실에서 평균과 분산이 정확히 일치하는 데이터는 흔치 않다. 일반적으로 분산이 평균보다 큰 값을 가져 등산포 가정이 어긋나는 경우가 많은데, 이러한 문제점을 **과대산포(Overdispersion)**이라고 한다. 과대산포 문제를 무시하고 포아송 회귀 모형 분석을 진행하면 분산을 과소평가하여 검정 결과가 왜곡되는 문제점이 발생한다. 따라서 과산포 검정을 통해 과대산포 여부를 확인하고 과대산포가 발견된 경우 '음이항 회귀 모형'을 이용하여 해결할 수 있다. (물론 그 외에도 Quasi-Poisson과 같은 방법들도 있다)

+) R에서 AER패키지 내에 있는 함수 dispersiontest() 통해 과산포 검정을 진행할 수 있습니다.

① 음이항 회귀 모형 (Negative Binomial Regression)

음이항 회귀 모형은 음이항 랜덤성분과 로그 연결함수로 구성된 GLM이다. 음이항 분포는 평균보다 큰 분산값을 갖고 있기 때문에 포아송 분포의 등산포 가정을 완화하는 효과가 있다. 평균과 분산은 다음과 같이 표현한다.

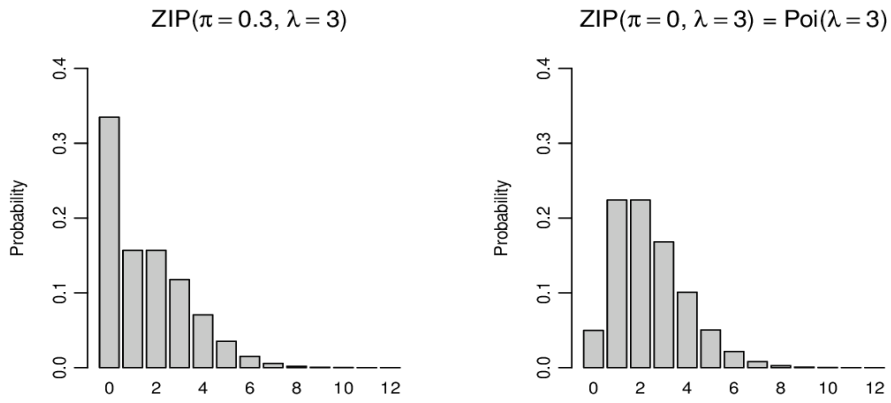
$$E(Y) = \mu, \qquad Var(Y) = \mu + D\mu^2$$

(평균과 분산 간 비선형성을 가정한 2차함수 형태)

위와 같이 평균은 포아송 분포와 같지만 분산은 기존 분산에 $D\mu^2$가 더해져 있다. D는 산포모수로 불리며 평균과 분산 간의 차이를 발생시키는 역할을 한다. 위와 음이항 분포의 성질을 이용하여 포아송 분포가 가지는 등산포 가정을 완화해 과대산포 문제를 해결한다.

2) 과대영 문제 (Excess Zeros)

포아송 분포에서 예상된 0 발생 횟수보다 실제로 더 많은 0이 발생하였을 때 과대영 문제(Excess Zeros)가 발생하였다고 말한다. 이는 실생활에서 매우 흔히 볼 수 있는 문제다. 예를 들어 로또 1등 당첨 횟수를 조사했을 때 대부분의 값이 0이 나올 것이다. 아래 그림을 비교해보자.



(왼쪽 : 과대영 문제 O / 오른쪽 : 과대영 문제 X)

왼쪽 그래프는 과대영 문제가 발생하지 않은 오른쪽 그래프에 비해 더 많은 0 값이 관측되었다. 이러한 과대영 문제는 영과잉 포아송 회귀모형(ZIP)이나 영과잉 음이항 회귀모형(ZINB)을 이용해 해결할 수 있다. 이 중 영과잉 포아송 회귀모형을 집중적으로 살펴보자.

① 영과잉 포아송 분포 (Zero Inflated Possion, ZIP)

영과잉 포아송 분포는 0만 발생하는 점확률분포와 0 이상의 정수 값이 발생하는 포아송 분포의 혼합분포구조를 가진다. 모형의 형식은 아래와 같다.

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

위의 형식에서 반응변수 Y는 베르누이 분포를 따르는데 0이 발생할 확률이 ϕi, 0 이상의 정수 값이 발생할 확률을 1- ϕi로 이분화해 표현한다. 이 때 $g(y_i)$는 0 이상의 정수 값들이 따르는 포아송 분포를 의미한다.
(이 $g(y_i)$가 음이항 분포를 따르고 이를 통해 GLM을 형성하면 영과잉 음이항 회귀모형(ZINB)가 됩니다!)
즉 영과잉 포아송 분포(ZIP)는 영과잉 부분(0)과 0이 아닌 부분(포아송 분포)으로 이분화시킨 모형이다.

이 ZIP을 GLM 모형으로 표현하면, 이를 영과잉 포아송 회귀모형(ZIPR)이라고 부른다. 해당 모형은 아래와 같이 두 개의 식으로 이루어진다.

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

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

위의 식은 ϕi에 관하여, 아래의 식은 λ에 관하여 만들어진 식이다. 전자는 0 값이 발생할 확률(ϕi)을 로짓 연결함수를 이용하여 표현한 식이고, 후자는 포아송 분포의 평균(λ)을 로그 연결함수를 이용하여 표현한 식이다.

[2주차 흐름 정리]

[실습코드]

이번 주차는 따로 과제로 제출하지 않고 직접 코드를 돌려보고 결과를 확인해 보면서 각 GLM 모형이 어떤 방식으로 작동되는지를 알아보도록 하겠습니다! 그 전에 오늘 배운 내용을 복습하시면서 코드를 돌려보고 궁금한 점은 언제든 질문해주세요! (그 전에도 언제든 질문 대환영~!!)

[3주차 예고]

이제 클린업도 1주밖에 남지 않았군요! 부족한 설명이었겠지만 범주형자료분석의 이론에 대해 깊이 공부해보는 기회가 되셨을 것이라고 생각합니다. 다음주 마지막 주차에서는 다양한 분류 평가지표와 샘플링 기법 그리고 인코딩 기법들에 대해 알아볼 것입니다. 특히 여러분들이 데이터를 다룰 때 실용적으로 활용할 수 있는 내용들이니 마지막까지 화이팅해 보아요!

[부록]

여기서 설명되는 부분은 정말x100 Only For 참고… 클린업 내용보다 심화되는 내용이지만 궁금하실 멋진 범주러들을 위해!

I. 지수족 (Exponential Family)

앞선 I. GLM(일반화 선형 모형) 단원에서 GLM은 랜덤성분의 확률분포를 정규분포에서 지수족에 해당하는 분포로 확장하였다고 설명했다. 그렇다면 지수족이란 무엇을 뜻하는 것인지 알아보자

지수족은 계산식이나 흐름이 매우매우 복잡하고 다루기 매우 어려운 부분입니다(저는 그랬어요…) 그래서 부록으로 둔 만큼 정말 압축된 내용만 설명하는 점 이해 부탁드려요. 최대한 우리의 목표인 범주형 자료분석, 특히 GLM과 연관성 있는 부분에 집중해서 설명을 하겠습니다. 통계적 추론 입문 수업에서 자세히 배우지만 여기선 간단히만 알아두는 걸로 해요!

1) 지수족의 정의

확률분포가 지수족(exponential family)에 속한다는 것은 다음의 조건 ①와 ②를 만족하는 경우이다.

① $f(x\,;\,\theta) = a(\theta)b(x)\exp\left(\sum \eta_i(\theta)t_i(x)\right)$

② A = {x | $f(x\,;\,\theta) > 0$} 가 θ에 의존하지 않는다.

①의 식의 우변을 4가지 구성성분으로 나누어 확인해보면, $a(\theta)$은 모수 θ만의 식(partition), b(x)는 x만의 식(support), $\eta_i(\theta)$는 자연 모수(natural parameter), $t_i(x)$는 충분통계량(sufficient statistic)이라고 불린다. 여기서 특히 우리는 자연 모수 $\eta_i(\theta)$와 충분통계량 $t_i(x)$이 갖는 중요성에 대해 기억할 필요가 있다.

2) 지수족의 필요성

① 충분통계량

그렇다면 지수족에 속하는 분포들이 왜 유용한지가 궁금하다. 그 이유는 지수족을 만족하는 표본들이 있을 때, 표본들의 결합확률분포(joint density)를 계산하면 완비충분통계량(C.S.S.)를 쉽게 구할 수 있기 때문이다.

$$f_{X_1,\dots,X_n}(X_1,\dots,X_n \mid \theta) = \{\prod_{i=1}^{n} h(x)\{c(\theta)\}^n \exp\,(\sum_{j=1}^{k}\{\eta_j(\theta)\sum_{i=1}^{n} \boldsymbol{t_j(x_i)}\})$$

위 식에서 충분통계량이라고 부르는 $t_i(x)$의 합이 완비충분통계량(C.S.S.)가된다.

$$T(X) = (\sum_{i=1}^{n} t_1(X_i), \dots, \sum_{i=1}^{n} t_k(X_i))$$

완비충분통계량은 모수 값을 완전히 설명할 수 있는 최소한의 함수식이다. 따라서 우리는 모든 데이터를 계산 시에 다 저장해서 사용할 필요가 없이 이 식에 나온 결과만을 저장하여 모수 추정에 활용하면 된다. 그리고 이 완비충분통계량의 함수인 T(X)의 기댓값이 적절히 변형되었을 때 이 값이 결국 θ의 최소분산 불편추정량(UMVUE)가 된다. (레만-쉐페 정리에 의하여!) UMVUE는 불편추정량 중에서 가장 작은 분산을 가진 값이기 때문에, 가장 모수를 잘 추정하고 있는 추정량이라고 생각하면 된다. 지수족에 해당하지 않는 확률분포의 최소분산 불편추정량을 구하는 것은 매우 복잡한 과정이지만, 지수족에 해당하면 완비충분통계량을 쉽게 구할 수 있고, 이를 이용하여 UMVUE까지 이르는 과정이 비교적 매우 편리해 진다는 장점이 있다.

② 자연 모수 (natural parameter)

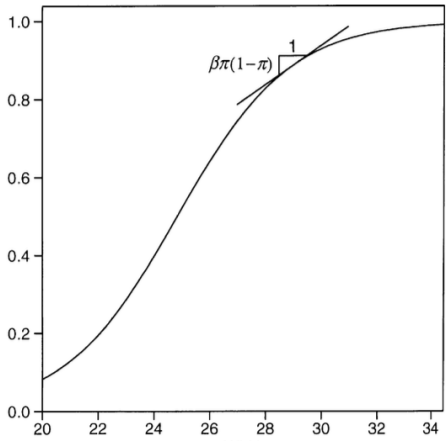
확률분포를 지수족의 표현으로 바꾸었을 때, $\eta_i(\theta)$에 해당하는 식을 자연모수라고 부른다. 이 자연모수는 우리가 배운 GLM의 연결함수를 추론하는 데에 큰 도움을 준다. 이항분포를 예시로 그 이유에 대해 알아보자.

$$f(x\,|n,p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \exp\big(nlog(1-p)\big) \exp\,(xlog\frac{p}{1-p})$$

위의 식에서 자연 모수 $\eta_i(\theta)$는 $log\frac{p}{1-p}$이다. 자연모수가 모수 p의 로짓(logit) 형태인 것을 알 수 있다. 이는 GLM에서 이항분포를 따르는 랜덤성분과 체계적 성분을 연결하는 로짓 연결함수와 같은 형태다. 이렇게 자연모수를 연결함수의 형태로 사용하는 것을 **정준연결함수**(canonical link function)이라고 부른다. 예를 들어, 포아송 분포의 자연 모수는 $log\lambda$이고 연결함수 역시 로그 연결함수를 사용한다. 반드시 정준연결함수가 GLM에 사용되는 것은 아니지만 대부분의 GLM이 비슷한 방식을 이용한다는 점에서 지수족의 자연 모수는 GLM 적용에 밀접한 연관성을 갖는다.

II. 로지스틱 회귀 기울기

앞선 IV. 로지스틱 회귀 모형 단원에서 로지스틱 회귀 모형의 시그모이드 함수의 기울기를 $\beta\pi(x)[1-\pi(x)]$ 이라고 설명했다. 이는 그림을 통해 표현하면 아래와 같다.



즉, 기울기는 β의 영향을 받으며, β의 부호에 따라 함수의 모양이 결정되며, |β|에 따라 기울기의 폭이 정해진다. 그렇다면, 어떻게 $\beta\pi(x)[1-\pi(x)]$ 식이 도출되었는지 궁금하다면 아래 증명식 참고! (x에 대한 곱하기 미분, 나누기 미분의 복잡한 형태일 뿐이랍니다!)

$$\frac{\partial\pi(x)}{\partial x} = \frac{\partial}{\partial x}\frac{e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})} = \frac{\partial}{\partial x}(e^{\alpha+\beta x}) * \left(\frac{1}{1+e^{\alpha+\beta x}}\right) + e^{\alpha+\beta x} * \frac{\partial}{\partial x}\left(\frac{1}{1+e^{\alpha+\beta x}}\right)$$

$$= \beta\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} + e^{\alpha+\beta x} * \left(-\frac{\beta e^{\alpha+\beta x}}{[1+e^{\alpha+\beta x}]^2}\right)$$

$$= \frac{\beta e^{\alpha+\beta x}(1+e^{\alpha+\beta x})}{[1+e^{\alpha+\beta x}]^2} - \frac{\beta(e^{\alpha+\beta x})^2}{[1+e^{\alpha+\beta x}]^2} = \frac{\beta e^{\alpha+\beta x}}{[1+e^{\alpha+\beta x}]^2}$$

$$= \beta\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)\left(\frac{1}{1+e^{\alpha+\beta x}}\right) = \beta\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)\left(1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)$$

$$= \boldsymbol{\beta\pi(x)(1-\pi(x))}$$

III. 순서형 다범주 로짓 모형

V. 다범주 로짓 모형 단원에서 순서형 자료를 다룰 때 collapse 방법에 따라 총 3가지 방법이 있다고 설명했다. 그 중 누적 로짓 모형에 대해서 살펴보았는데, 나머지 두 방법에 대해 간단히 살펴보자!

1. 이웃범주 로짓 모형 (Adjacent-Categories Logit)

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

이웃범주 로짓 모형은 cut-point를 기준으로 바로 옆에 있는 범주만을 사용한다. 이웃범주 로짓 모형은 이전 순서의 범주 대신 다음 순서의 범주에 속할 오즈에 대한 설명변수의 효과, 즉 β값이 동일하다고 가정한다.

즉, $J-1$개의 이웃범주 로짓 모형에 대한 회귀계수의 효과가 모두 동일하다고 여기는 것이다! (누적 로짓 모형하고 동일하죠?!) 이웃범주 로짓 모형의 형태는 다음과 같다!

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

→ 오즈를 이용하여, 다른 변수가 고정되어 있을 때 x가 한 단위 증가하면 **$Y = j$ 대신에 $Y = j+1$ 일 오즈가 e^{β}배**라고 해석한다.

누적 로짓 모형은 복수의 범주와 복수의 범주를 비교한 반면, 이웃범주 로짓 모형은 이웃한 두 범주 간의 오즈를 비교한다.

2. 연속비 로짓 모형 (Continuation-ratio Logit)

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

연속비 로짓 모형의 형태는 기준 범주(표에서 파란색 강조)보다 낮은 변수들 혹은 높은 범주들(표에서 주황색으로 강조)을 묶어서 로짓을 정의하기 때문에 2가지 형태를 가진다. 더불어 각 모형에서의 해석은 오즈를 이용한다.

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

→ 다른 변수가 고정되어 있을 때, x가 한 단위 증가하면 $\boldsymbol{Y \geq j+1}$ **대신에** $\boldsymbol{Y = j}$**일 오즈가** $\boldsymbol{e^{\beta}}$배이다.

② 기준범주보다 **낮은** 범주 : $\log\left(\frac{\pi_1+\cdots+\pi_j}{\pi_{j+1}}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k,\ j = 1, \ldots, (J-1)$

→ 다른 변수가 고정되어 있을 때, x가 한 단위 증가하면 $\boldsymbol{Y = j+1}$ **대신에** $\boldsymbol{Y \leq j}$**일 오즈가** $\boldsymbol{e^{\beta}}$배이다.

끝!!!