범주형 자료분석 2주차 교안

여러분을 GLM의 세계에 초대할 2주차 클린업입니다! 저번주에는 분할표와 독립성 검정, 그리고 연관성 지표에 대해 공부했습니다. 이번주는 모델에 대해 배워볼 건데요. 특히 범주의 꽃, 로지스틱 회귀를 뿌셔볼 예정~!

팀장이 좋아하는 앙꼬 넣어봤어요… 2주차도 빠이팅

- 목차 -

Table of Contents

1. GLM (일반화선형모형, Generalized Linear Model)

■ GLM (일반화선형모형)이란?

1) GLM의 정의

일반화 선형 모형 (GLM)의 정의를 내리기 전에 먼저 우리가 잘 알고 있는 **선형 회귀 모형**에 대해 생각해보자. 일반 선형회귀모형은 최소제곱법으로 연속형 변수 사이의 회귀식을 추정한다. 선형회귀모형은 독립변수와 종속변수 사이의 선형성, 오차항의 정규성과 독립성, 등분산성의 4가지 기본 가정을 만족해야 한다. *(회귀모형에 대한 더 자세한 내용은 회귀분석팀 클린업 참고!)*

반면 반응변수가 범주형 자료이거나 count data인 경우도 비일비재하다.

- 반응변수가 범주형 변수인 경우 예시 : 이항 분포의 변수 (0 or 1, 성공 or 실패), 다항 분포의 변수 (지지 정당-민주당/무소속/공화당) 등
- 반응변수가 도수 자료(count data)인 경우 예시 : 포아송·음이항 분포의 변수 (물 마시는 횟수, 교통사고 건수) 등

이렇듯 일반선형모형의 가정들이 적용될 수 없는 경우를 위해 일반선형모형을 보다 확장한 것이 **일반화 선형모형(Generalized Linear Model)**이다. 연속형 반응변수에 대한 회귀모형이나 분산분석 모형 뿐만 아니라 범주형 반응변수에 대한 모형들을 포함하는 광범위한 모형의 집합이라고 할 수 있다.

GLM에서 '일반화(generalized)' 는 보통의 회귀모형을 두 가지로 일반화한 것을 의미한다.

> 첫째, 랜덤성분이 정규분포를 포함한 다른 분포를 갖도록 **일반화**
> 둘째, 랜덤성분의 함수인 연결함수(link function) "$g(\,)$"로 모형화하여 **일반화**

보통의 회귀모형은 랜덤성분에 대해 정규분포를 가정하고, 평균 그 자체를 항등연결함수 $g(\mu)=\mu$로 모형화한 것이다. 즉, 일반 선형회귀모형 역시 GLM의 한 종류인 것이다! *(지금까지 알아온 일반 선형회귀모형은 GLM의 빙산의 일각에 불과하다…)*

랜덤성분? 연결함수? 아직 낯선 용어들의 향연에 당황할 수 있지만, 앞으로 GLM의 구성성분과 특징들을 살펴보면 쉽게 이해하게 될 것이다. 우선 GLM이 기존의 회귀모형을 포함한 더욱 넓은 범위의 모형이라는 것을 알아 두자!

2) GLM의 필요성

위에서 살펴봤듯이, 반응변수가 범주형 자료이거나 count data인 경우 우리는 일반선형모형을 쓸 수 없다. 그렇기에 일반화된 선형모형인 GLM에 대한 필요성이 생기는 것이다. 더 구체적으로 알아보자.

우리가 다루는 범주형 자료분석은 오차항의 확률분포가 정규분포가 아니기 때문에 선형회귀모형의 사용이 불가하다. 그러나 GLM의 적합과정은 ML 방법(최대우도법)을 사용하기 때문에 회귀분석의 LSE(최소제곱법)와 같은 정규성 조건이 필요 없는 것이다! GLM은 이처럼 반응변수에 대한 가정이 상대적으로 널널하기 때문에 정규분포 외에 다른 분포들도 사용이 가능하다. 따라서 이항분포와 같이 정규분포 외에 다른 분포를 따르는 범주형 반응변수를 다루고 싶다면 GLM을 써야 한다.

더불어 분할표 분석과 비교했을 때도 GLM은 장점을 갖는데, 분할표로는 변수 간의 효과 파악(= 독립성

검정)만 가능하다면, GLM은 **변수 간 연관성 파악**과 더불어 **반응 변수에 대한 예측**도 가능하다. 또한 분할표 분석은 모든 변수가 범주형 자료일 때만 표현이 가능하지만, GLM은 설명변수에 연속형 변수가 있어도 사용이 가능하다. *(말그대로 모델링의 장점이다!)*

3) GLM의 특징

- 오차항의 다양한 분포를 가정

  오차항의 정규분포를 따른다는 정규성 가정을 만족해야 하는 일반선형회귀와 다르게 GLM은 다양한 분포를 가정할 수 있다. 오차항의 확률분포가 무엇이냐에 따라 일반적으로 사용하는 연결함수는 정해져 있다. *(이 부분은 GLM의 종류에서 알아보자!)*

- 선형 관계식 (Linear Predictor) 유지

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

  : GLM은 다음과 같이 선형 관계식을 유지하기 때문에 해석이 용이하다는 장점이 있다. 이 때 사용되는 선형의 의미는 회귀계수 β에 대한 선형성을 의미한다. 즉, 선형과 비선형을 구분할 때 설명변수 X와 반응변수 Y의 관계를 기준으로 생각해서는 안된다. 우리가 추정해야 하는 미지수는 설명변수나 반응변수가 아니라 회귀 계수 β이기 때문이다.

- 범위가 제한되는 반응변수도 사용이 가능

  : 연결함수 $g(\mu)$를 통해 범위를 조정할 수 있다.

- 독립성 가정만 필요

  : 기존의 회귀 가정인 정규성, 등분산성, 독립성, 선형성 중에서 독립성 가정만 만족하면 된다! 즉 자기상관성 검정이 필요하다! **자기상관성(autocorrelation)**이란 반응변수(Y변수)의 각 관측치가 상호 연관성을 띄는 것으로 시간 또는 공간적으로 연속된 일련의 관측치들 간에 존재하는 상관관계를 뜻한다.

  일반적으로 회귀분석 후 더빈-왓슨 검정(DW Test)을 통해 자기상관성(혹은 오차의 독립성)을 검정한다. 시간적 자기상관성이 존재하는 경우 시간을 고려하는 시계열 모델을 사용하고, 공간적 자기상관성의 경우 Moran's I 검정을 통해 확인하고 공간 회귀모형 등을 사용할 수 있다.

■ GLM의 구성성분

GLM 구성 성분		
랜덤 성분 **(random component)**	**연결 함수** **(link function)**	**체계적 성분** **(systematic component)**
$\boldsymbol{\mu}\ (= E(Y))$	$\boldsymbol{g()}$	$\boldsymbol{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}$
$\boldsymbol{g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k}$		

GLM의 일반적인 형태는 다음과 같고, 세 개의 구성성분으로 이루어진다.

1) 랜덤 성분

랜덤 성분(random component)은 반응변수 Y를 정의하고 반응변수에 대한 확률분포를 가정하는데, 이 때 가정한 확률분포의 기대값인 평균 μ를 랜덤성분으로 표기한다.

우선 반응변수 Y의 관측값을 서로 독립이라고 가정하는데, 관측값 Y_i의 결과가 성공 혹은 실패를 나타낸다면 이항분포를 사용해 이항 랜덤성분을 가정하고, 관측값 Y_i가 시간당 발생하는 횟수를 나타낸다면 포아송 랜덤성분을 가정하는 원리이다. Y가 이항분포를 따르는 경우 이항분포의 평균인 $\pi(x)$를 사용하고, 포아송분포를 따르는 경우는 포아송 분포의 평균인 μ (또는 λ)를 사용한다.

2) 체계적 성분

체계적 성분(systematic component)는 설명변수 X들의 선형결합으로, 선형예측식 $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ 을 말한다.

체계적 성분에는 교호작용항이나 곡선효과를 나타내는 항을 넣을 수도 있다.

- $\boldsymbol{x_i = x_a x_b}$: x_a와 x_b의 교호작용을 설명하는 항
- $\boldsymbol{x_i = x_a^2}$: x_a의 곡선효과를 나타내는 항

3) 연결함수

연결함수(link function) $g()$는 랜덤성분과 체계적 성분을 연결하며 둘의 범위를 맞춰주는 역할을 한다.

GLM은 앞서 설명한 것처럼 반응변수가 연속형 변수가 아닌 범주형 변수인 경우나 정규분포 외에 다른 분포를 따르는 경우를 표현할 수 있다. GLM의 반응변수가 범주형 변수인 경우, 체계적 성분인 우변은 $-\infty$부터 ∞까지의 범위를 가지는 반면, 랜덤성분(반응변수 Y의 평균)인 좌변은 이산형 등 연속형이 아닌 값을 가지게 되므로, 양변의 범위가 맞지 않는 결함이 나타나게 된다. 이러한 문제를 해결하기 위해 연결함수가 존재하는 것이다!

GLM에는 연결함수의 종류가 굉장히 많다. 보통은 연결함수로 항등함수를 사용한다. 랜덤성분의 정규분포를 가정하고, 항등함수를 사용한 식이 바로 우리가 아는 일반회귀모형 식이다! 이제 왜 일반회귀식이 GLM에 포함되는지 알겠죠?

그렇다면 이제 다양한 연결함수의 종류를 알아보자!

- **항등 연결 함수 (identity link)** : $g(\mu) = \mu$

  반응변수가 연속형일 때 사용하며, 이항 반응변수일 때 사용하게 되면 위에서 말한 것처럼 양변의 범위가 맞지 않는 구조적 결함이 나타난다. 예를 들어, 좌변은 [0,1] 의 범위를 우변은 $(-\infty, \infty)$의 범위를 갖게 되어 결함이 나타나는 경우를 말한다.

- **로그 연결 함수 (log link)** : $g(\mu) = \log(\mu)$

  반응변수가 도수자료(count data), 즉 음이 아닌 값인 경우에 많이 사용한다. 보통 포아송 분포나 음이항 분포를 따를 때 사용한다.

- **로짓 연결 함수 (logit link)** : $g(\mu) = \log[\mu/(1-\mu)]$

  로짓(logit)은 오즈에 로그를 씌운 값으로, 반응변수가 확률처럼 0과 1사이의 값을 가질 때 유용하다. 대표적으로 이항 반응변수인 경우 사용하며, 스포하자면 이게 바로 로지스틱 회귀이다!

- 프로빗 연결 함수 (Probit link) : $g(\mu) = \Phi^{-1}(\mu) = probit(\mu)$

  프로빗 연결 함수는 표준정규분포의 누적분포함수에 역함수를 취한 것이라고 생각하면 된다. 반응변수가 확률처럼 0과 1사이의 값일 때 유용하다.

이외에도 여러 연결함수의 종류가 있지만, 대표적으로 이 세 가지를 가장 많이 사용한다.

■ 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	
기준범주 로짓 모형	다항 자료	로짓	
누적 로짓 모형			
이웃범주 로짓 모형			
연속비 로짓 모형			
로그 선형 모형	도수 자료	로그	범주형
포아송 회귀 모형			혼합형
음이항 회귀 모형			
카우시 모형			
율자료 포아송 회귀 모형	비율 자료		

GLM의 종류가 이렇게나 많다…

우리는 이중에서 핑크색에 해당하는 모형만 메인으로 다루고자 한다. *(선택과 집중 감성… 만약 지식의 갈증을 느껴서 더 알고 싶은 부분이 있다면 구글링 혹은 저에게 질문해주세요! ㅎㅎ)*

1) 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

단순한 모형이지만 이항자료를 다루기에는 양변의 범위가 맞지 않는 구조적 결함이 있다.

- **로지스틱 회귀 모형**

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

구성 : 이항 랜덤 성분과 로짓 연결 함수.

- 프로빗 회귀 모형 (프로빗 모형)

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

구성 : 이항 랜덤 성분과 프로빗 연결

계획된 실험 등 특별한 경우에만 사용되고 자주 쓰이지는 않는다.

2) 다항 자료

- **기준범주 로짓 모형**

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

구성 : 다항 랜덤 성분(명목형)과 로짓 연결 함수

- **누적 로짓 모형**

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

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 : 다하 랜덤 성분(순서형)과 로짓 연결 함수

3) 도수 자료

- **포아송 회귀 모형**

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

구성 : 포아송 랜덤 성분과 로그 연결 함수

- **음이항 회귀 모형**

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

  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

- 로그 선형 모형

메인으로 다룰 모형들은 각 파트에서 세세히 알아보도록 하자.

■ GLM의 모형 적합

모형 적합, 즉 model fitting 이란 주어진 데이터를 근거로 모형의 모수를 추정하는 것을 의미한다.

GLM은 LSE(최소제곱추정법)을 사용할 수 없다! LSE는 오차의 제곱합을 최소화하는 방법으로 회귀의 기본 가정들을 만족해야 하지만 GLM은 그러한 가정이 없기 때문에 당연한 말이다. 따라서 GLM은 모형을 적합할 때 ML(Maximum Likelihood Estimation), 즉 최대가능도추정법을 사용한다. 이 때 **가능도(Likelihood)**란 관측값이 고정됐을 때, 그 관측값이 어떤 확률분포를 따를 가능성을 뜻한다. *(한편 **확률**(probability)이란 확률분포가 고정되어 있을 때 특정 값이 관측될 가능성을 뜻하니, 둘을 구분해서 잘 알아 두자!)*

MLE의 핵심 아이디어를 이해하기 위해 간단한 예시를 살펴보자.

왼쪽 그림에서 데이터 x는 주황색 곡선과 파란색 곡선 중 어떤 곡선으로부터 추출되었을 확률이 더 높을까? 눈으로 보기에도 파란색 곡선보다는 주황색 곡선에서 이 데이터들을 얻었을 가능성이 더 커 보인다. 데이터들의 분포가 주황색 곡선의 중심에 더 일치하는 것처럼 보이기 때문이다. 이 예시를 보면 우리가 데이터를 관찰함으로써 이 데이터가 추출되었을 것으로 생각되는 분포의 특성을 추정할 수 있음을 알 수 있다.

오른쪽 그림에서는 주황색 후보 분포에 대해 각 데이터들의 가능도를 점선의 높이로 나타냈다. 수치적으

로 이 가능도를 계산하기 위해서는 각 데이터 샘플에서 후보 분포에 대한 높이를 계산해서 다 곱한 것으로 사용할 수 있다. 우리가 생각할 수 있는 모든 후보군 분포에 대해 가능도를 계산해서 비교하면 우리는 데이터를 가장 잘 설명할 수 있는 확률분포를 얻어낼 수 있게 되는 것이다. 그 식을 바로 **가능도 함수(Likelihood function)**이라고 한다.

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

보통은 로그를 취해 아래와 같이 **로그 가능도 함수(log-likelihood function)**을 이용한다.

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

결국 최대가능도추정법이란 이 가능도 함수가 최대가 되는 모수 θ를 찾는 방법이라 할 수 있다. 모수 θ에 대해 편미분하고 0이 되는 θ를 찾는 과정을 통해 **최대가능도추정량(Maximum Likelihood Estimator)**를 찾을 수 있다. *(더 자세한 설명과 증명은 통계적추론입문 수업에서…)*

2. 유의성 검정

■ 유의성 검정이란?

유의성 검정이란 모형의 모수 추정값이 유의한지에 대한 검정이다. 더불어 축소 모형의 적합도가 좋은지에 대한 검정이기도 하다!

회귀분석에서 회귀 계수의 유의성 검정(t검정)이나 회귀 모델의 유의성 검정(F검정)에 대해 배웠으니 익숙할 것이다. 이 파트에서는 **LR test (likelihood ratio test, 가능도비 검정)**을 통해 GLM의 전체 베타 계수를 검정하는 방법에 대해 알아보자.

모형 $g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ 에 대하여 가설은 다음과 같다.

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

H_1 : 적어도 하나의 β는 0이 아니다.

■ 가능도비 검정

가능도비 검정은 귀무가설 하에서 계산되는 가능도 함수 $\boldsymbol{l_0}$와 MLE에 의해 계산되는 가능도 함수 $\boldsymbol{l_1}$의 차이를 이용한다. 검정 통계량은 다음과 같다.

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

말로 풀이하면 다음과 같다.

$$-2\log\left(\frac{\text{모수가 귀무가설 } \mathrm{H_0} \text{ 을 만족할 때 } (\beta = 0\text{일 때}) \text{ 가능도 함수의 최댓값}}{\text{모수에 대한 아무 제한조건이 없을 때의 가능도 함수의 최댓값}}\right)$$

가능도비 검정의 flow는 1주차의 독립성 검정과도 유사한데, l_0과 l_1의 차이가 커지게 되면 -> 검정 통계량이 커지고 > p-value가 작아져서 -> 귀무가설을 기각하며-> 적어도 하나의 β는 0이 아니게 되므로 -> 모형의 모수 추정값은 유의하다고 볼 수 있다!

이 때 자유도는 두 모형의 차원의 차이(Ω_0과 Ω_1의 차원의 차이)로, 귀무가설과 대립가설 간의 모수의 개수 차이와 같다.

가능도비 검정은 $\hat{\beta}$(MLE)와 $\beta = 0$(귀무가설 하)인 두 가지 경우의 로그 가능도 함수에 대한 정보를 사용한다. 가장 많은 양의 정보를 사용하는 검정이기 때문에 더 좋은 검정력을 갖는다.

왈드 검정(Wald Test)과 비교하자면, 검정 통계량 $[\hat{\beta}/SE \sim Z,\ (\hat{\beta}/SE) \sim X_1^2]$ 을 계산할 때 계수 값과 표준오차만 사용하면 되기 때문에 간단하지만 가능도비 검정보다 검정력이 낮다.

이러한 가능도비 검정은 이탈도 차이를 통한 모형 비교에서도 사용된다.

■ 이탈도 (deviance)

1) 관심모형과 포화모형

이탈도에 대해 알기 전에 먼저 관심모형과 포화모형의 의미를 알아보자.

- **관심모형** M은 유의성 검정을 진행하고자 하는 모형이다. $\boldsymbol{L_M}$은 모형 M에서 얻은 로그 가능도 함수의 최댓값이다.

  ex) $범주팀\ 복지(Y) = \beta_0 + \beta_1 \times 스터디\ 시간(x_1) + \beta_2 \times 교안\ 페이지\ 수(x_2)$

- **포화모형** S는 각 관측값에 대하여 완벽하게 자료를 적합하는 모형으로 가능한 모형 중 가장 많은 수의 모수를 갖는 (=가장 복잡한) 모형이다. *포화모형의 모수의 개수는 관측치의 개수와 같다.* $\boldsymbol{L_S}$는 모형 S에서 얻은 로그 가능도 함수의 최댓값이다.

  ex) $범주팀\ 복지(Y) = \beta_0 + \beta_1 \times 스터디\ 시간(x_1) + \beta_2 \times 교안\ 페이지\ 수(x_2)$

  $+\beta_3 \times 스터디\ 시간 \times 교안\ 페이지\ 수(x_1 x_2)$

가설은 다음과 같다.

H_0 : 관심 모형에 속하지 않는 모수는 모두 0이다. ➔ 관심 모형 사용

H_1 : 관심 모형에 속하지 않는 모수 중 적어도 하나는 0이 아니다. ➔ 관심 모형 사용 불가

2) 이탈도(deviance)란?

이탈도(deviance)란 포화 모형 S와 관심 모형 M을 비교하기 위한 가능도비 통계량이다.

$$이탈도(deviance) = -2(L_M - L_S)$$

이탈도는 S(포화 모형)에는 있지만 M(관심모형)에는 없는 계수들이 0인지 확인하는 통계량이기 때문에 모형이 nested일 때 (M의 계수 ⊂ S의 계수)만 사용이 가능하다.

이탈도를 통해 모형 적합도를 검정할 수 있다. flow를 살펴보면, 이탈도(검정 통계량)가 작다 -> p-value가 크다 -> 귀무가설 기각 안함 -> 관심 모형에 포함 안된 계수들이 0이다 (오 관심모형 적합도 좋은데?) -> 더 간단한 관심 모형을 적합하는 것이 좋겠다! 와 같은 방식으로 적합도 검정이 가능하다.

3) 이탈도와 가능도비 검정의 관계

가능도비 검정 통계량은 모형 간의 이탈도의 차와 같다.

M_0은 단순한 모형(Reduced Model), M_1은 복잡한 모형((Full Model)일 때,

$$M_0\text{의 이탈도} - M_1\text{의 이탈도 (모형 간의 이탈도의 차)}$$
$$= -2(L_0 - L_S) - \{-2(L_1 - L_S)\} = -2(L_0 - L_1) = \text{가능도비 검정 통계량}$$

위 식을 통해 간단한 모형의 이탈도와 복잡한 모형의 이탈도의 차가 가능도비 검정통계량과 같다는 것을 알 수 있다.

하지만 이탈도를 통해 모형을 비교하려면, M_0은 M_1의 내포(nested) 모형이어야 한다. 만약 내포된 경우가 아니라면 AIC와 같은 모형 선택의 기준이 되는 측도를 통해 모형 비교를 해야 한다. *(AIC는 다른 변수 조합으로 이루어진 모형을 비교할 때 쓰는 비교 지표로 회귀분석에서 배운다!)*

모형 M_0이 M_1에 비해 적합이 잘 되지 않는다면, 두 이탈도의 차이가 커지므로 이 검정통계량은 큰 값을 갖게 된다! 이러한 원리로 내가 관심 있는 모형이 더 적합한 모형인지 알 수 있다. 검정 flow를 살펴보면, 이탈도 차이가 작으면 -> 검정 통계량 값이 작다 -> p-value가 높다 -> 내가 관심있는 모형에 없는 계수들은 0이구나 -> 간단한 축소 모형인 관심 모형이 더 적합하다!

3. 로지스틱 회귀 모형

■ 로지스틱 회귀 모형(Logistic Regression Model)이란?

드디어 범주의 꽃, 로지-스틱 회귀 모형 파트이다! *오늘 스터디가 끝나면 여러분들은 glm(formula, family = binomial(link='logit'), data) 코드가 다시 보일 것이다!*

회귀분석 수업을 들었다면, 마지막 챕터였던 로지스틱 회귀를 기억할 것이다. 기존 회귀 모델과 비교되는 로지스틱 회귀 모형의 가장 큰 차이점은 바로 **반응변수가 이항자료**라는 것이다! 반응변수 Y가 성공 혹은 실패의 이항분포를 따르는 변수이기 때문에 기존 회귀 모델을 그대로 적용할 수가 없게 된다.

그러면 문제를 바꿔서, 종속변수 Y를 범주 대신 (범주1이 될) 확률로 두고 식을 세워보자. 우변의 선형예측식은 그대로 두고 좌변을 확률로 바꾸면 다음과 같아진다.

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

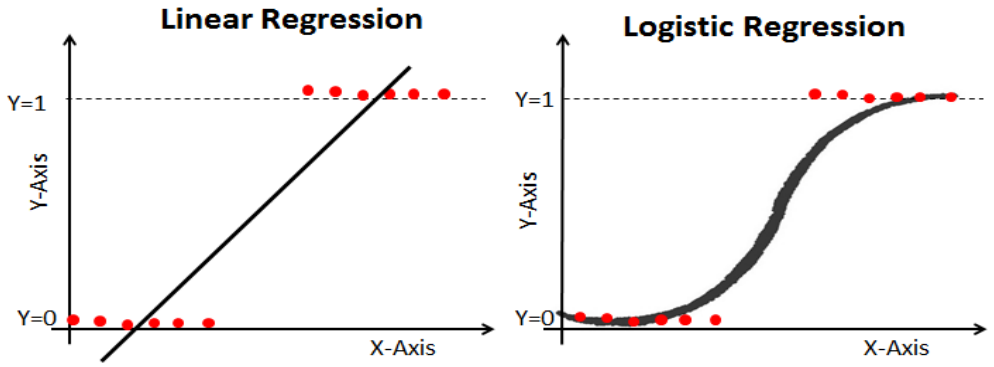
GLM 파트에서 배웠듯이, 이 선형확률모형에는 구조적 결함이 있다. 확률인 좌변의 범위는 (0,1)이지만 우변의 범위는 $-\infty$에서 ∞이기 때문이다. 여기서 식을 바꿔서 좌변을 오즈로 설정하면 다음과 같다.

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

하지만 이번에도 양변의 범위가 맞지 않는다. 오즈의 범위는 0에서 무한대이기 때문이다. 그래서 우리는 오즈에 로그를 취해주는 것이다.

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

이렇게 로짓 연결 함수를 사용함으로써, 우리는 좌변과 우변의 범위를 $-\infty$에서 ∞로 맞춰줄 수 있게 된 것이다. 이 식이 바로 로지스틱 회귀 모형이다.



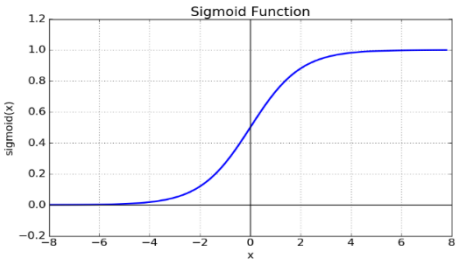
로지스틱 회귀를 왜 쓰는지 한번에 이해할 수 있는 그림 ~~(literally 선 넘는 문제를 해결한다)~~

로짓 연결 함수를 통해 양변의 범위가 일치되는 과정을 정리하면 다음과 같다.

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

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

$$-\infty \le \log\left(\frac{\pi(x)}{1-\pi(x)}\right) \le \infty$$



참고로 로지스틱 회귀와 같은 함수 모양을 **시그모이드 형태**라고 한다. 확률을 따르는 S자 곡선의 함수로 $\pi(x)$와 x의 비선형 관계를 나타낸다.

딥러닝에서 활성화 함수로 자주 등장하는 시그모이드 함수는 로지스틱 함수라고도 불린다. 1주차 딥러닝 클린업 참고!

범위 일치 이외에도 로지스틱 회귀모형은 장점을 가진다. 바로 가정으로부터 자유롭다는 것인데, 일반회귀모형과 달리 정규성, 등분산성, 선형성의 가정이 필요 없다. 오직 독립성 가정만 만족하면 된다. 애초에 이항분포를 따르는 반응변수 자체가 정규성과 등분산성을 만족할 수 없다. 이항분포의 분산 $\mathrm{Var(x)} = \pi(x)(1-\pi(x))$ 식만 살펴봐도 x에 대한 식이기 때문에 등분산성 조건이 위배되는 것은 당연!

■ 로지스틱 회귀 모형의 해석

이제 적합한 로지스틱 회귀 모형을 해석하는 방법에 대해 알아보자. 로지스틱 회귀 모형 식을 변형하면 다음

과 같이 확률에 대한 식으로 표현할 수 있다.

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

확률 식에 x값을 대입하면 특정 범주에 속할 확률 $\pi(x)$, 즉 $P(Y = 1|X = x)$을 알 수 있다. 만약 **확률 값이 cutoff point보다 크면 $Y = 1$, 작으면 $Y = 0$으로 예측할 수 있다!**

일반적으로 사용되는 cutoff point는 0.5이지만 상황에 따라 다르게 적용할 수 있다. 이와 관련된 내용은 3주차 ROC 곡선 파트에서 다룰 예정!

그렇다면 로지스틱 회귀 모형의 **회귀 계수 β**는 어떻게 해석하면 될까? x에 대한 로지스틱 회귀모형의 변화율, 즉 기울기를 구해보자. 미분을 하면 x에서의 접선의 기울기를 구할 수 있다.

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

이 변화율은 모수 β를 갖는 로지스틱 회귀모형의 접선의 기울기이다. 변화율 식에서 회귀계수 β는 곡선의 증가비율 혹은 감소비율을 결정한다. β가 양수이면 상향 곡선, 음수이면 하향 곡선이 그려지며, $|\beta|$이 증가함에 따라 변화율이 증가한다.

로지스틱 회귀모형의 연결함수는 로짓 함수, 즉 로그 오즈 함수이므로 오즈와 오즈비를 이용하여 해석할 수 있다는 장점이 있다. 회귀 모형에 각각 x와 $x + 1$을 대입한 뒤 빼 주면 오즈비의 형태가 나온다. 수식을 보면 바로 이해할 수 있다.

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

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

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

즉, **x가 한 단위 증가하면 $Y = 1$일 오즈가 e^{β}배 만큼 증가한다**고 해석하면 된다! 설명변수가 여러 개인 다중 로지스틱 모형의 경우, 일반 선형회귀와 마찬가지로 다른 설명변수가 모두 고정되어 있을 때라는 조건이 유지된다. 학점에 따른 합격유무 예시를 들어볼까?

로지스틱 회귀모형에 적합하여 나온 식이 $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 4 + 3x$이고 $Y = 1$(합격), $Y = 0$(불합격), x는 학점을 의미한다면, x가 한 단위 증가할 때 $Y = 1$(합격)일 오즈가 e^3 즉, 20.086배 증가한다고 해석하면 되는 것이다!

또한 한 단위가 아니더라도 추정된 $\hat{\pi}(x_1)$과 $\hat{\pi}(x_2)$를 통해 $\hat{\pi}(x_2) - \hat{\pi}(x_1)$를 구하여 x_1에서 x_2로 증가할 때의 확률의 변화도 알 수 있다. 이 경우 위의 예시를 들면, 학점이 2.5에서 4.5로 증가할 때, $\frac{\exp(4+3\times4.5)}{1+\exp(4+3\times4.5)} - \frac{\exp(4+3\times2.5)}{1+\exp(4+3\times2.5)} = 0.00001$만큼 $Y = 1$(합격)일 확률이 증가한다고 해석 가능하다.

4. 다범주 로짓 모형

■ 다범주 로짓 모형(Multicategory Logit Model)이란?

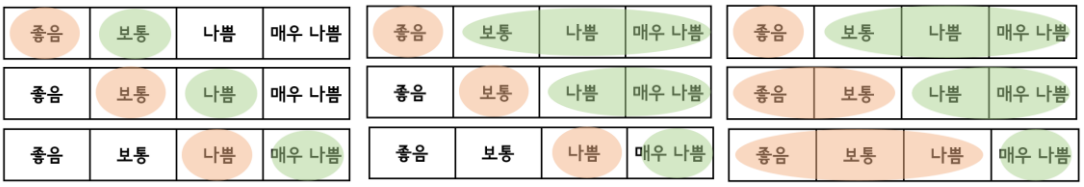
지금까지 배운 로지스틱이 Yes or no의 이항 분류 문제를 다룬다면, 다범주 로짓 모형은 반응변수의 범주(카테고리)가 3개 이상인 것으로 확장한 모형이다. 연결함수는 로지스틱과 마찬가지로 로짓연결함수를 쓰지만, **반응변수가 다항 분포를 따르는 다항자료**인 것이 특징이다. *(설명변수가 여러 개인 다중 로지스틱 회귀모형(Multiple Logistic Regression)과 이름이 비슷하지만 다른 모형이다! 헷갈리지 말자~)*

이 때 다항자료인 Y변수가 명목형 자료인지 순서형 자료인지 구분해서 접근할 필요가 있다. 자료의 종류에 따라 적용하는 모델이 달라지게 된다. 결론부터 말하자면, Y변수가 명목형이라면 기준 범주 로짓 모형을 사용하고, 순서형이라면 누적 로짓 모형, 이웃 범주 로짓 모형 또는 연속비 로짓 모형을 사용한다.

이 중 우리는 가장 많이 쓰이는 **기준 범주 로짓 모형**과 **누적 로짓 모형**을 집중적으로 살펴볼 것이다.

다범주 로짓 모형	명목형	기준 범주 로짓 모형 (Baseline-Category Logit Model)
	순서형	이웃 범주 로짓 모형 (Adjacent-Categories Model)
		연속비 로짓 모형 (Continuation-ratio Logit Model)
		누적 로짓 모형 (Cumulative Logit Model)

순서형 다범주 로짓 모형은 순서 정보를 고려하기 때문에 범주(카테고리)를 순서대로 정렬 시킨 후 두 개의 범주로 나눠주는 collapse 과정이 필요한데, 이 때 collapse 시키는 기준인 cut point를 어떻게 정하는지에 따라 사용되는 모형이 달라진다.



(1) 이웃범주 로짓 모형, (2) 연속비 로짓 모형, (3) 누적 로짓 모형의 cut point collapse 방법을 그림으로 표현한 것이다. 우리는 이 중 누적 로짓 모형에 대해서만 알아볼 거지만, 모형들의 차이점이 궁금할까봐 그림을 가져왔다! 참고로 누적 로짓 모형은 이웃 범주 로짓이나 연속비 로짓 모형과 달리 전체 범주를 모두 사용해서 해석한다는 특징이 있다.

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

반응변수 Y가 J개의 범주(카테고리)를 갖는 명목형 변수일 때, 명목형 반응변수에 대한 로짓 모형은 기준범주(Baseline-Category)를 선택한 뒤 이 범주와 나머지 범주의 짝을 지어 로짓을 정의한다. 마지막 범주 J가 기준이 될 때, 기준 범주 로짓은 다음과 같이 정의된다.

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

쉽게 생각하면 이 로짓은 반응이 J범주에서 일어났다는 조건 하에서, 반응이 j범주일 로그 오즈가 된다. 기준 범주 로짓 모형은 다음과 같은 모양을 갖는다.

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

이 때 J는 기준 범주, j는 범주에 대한 첨자, A~K는 설명변수 x에 대한 첨자이다. 기준 범주 로짓 모형은 J-1개의 로짓 방정식으로 이루어졌으며 각각의 식마다 다른 모수들을 갖는다. J=2인 경우는 이항 반응변수에 대한 보통의 로지스틱 회귀가 된다!

모형 공식을 변형하여 j범주에 속할 확률을 이렇게 나타낼 수 있다.

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

기준 범주 로짓 모형은 명목형 범주를 다루기 때문에 순서를 고려하지 않는다. 또한 오즈와 기준 범주를 사용해서 해석 할 수 있는데, 기준 범주에 비해 j범주일 로그 오즈를 보고 해석한다. 즉, **x가 1단위 증가하면** (다른 설명 변수가 고정되어 있을 때) **J범주 대신 j범주일 오즈가 e^{β}배 증가**한다고 해석하면 된다!

기준범주가 아닌 또 다른 범주끼리 로짓을 빼 주면 그 범주의 관계로도 해석이 가능하다.

$$\log\left(\frac{\pi_2}{\pi_1}\right) - \log\left(\frac{\pi_2/\pi_J}{\pi_1/\pi_J}\right) = \log\left(\frac{\pi_2}{\pi_J}\right) - \log\left(\frac{\pi_1}{\pi_J}\right)$$

$$= [\alpha_2 - \alpha_1] + [(\beta_2^A - \beta_1^A)x_1 + \cdots + (\beta_2^K - \beta_1^K)x_K]$$

이 경우 **x가 1단위 증가하면** (다른 설명 변수가 고정되어 있을 때) **1범주 대신 2범주일 오즈가 $e^{\beta_2 - \beta_1}$배 증가**한다고 해석하면 된다! (어렵지 않죠?)

■ 누적 로짓 모형 (Cumulative Logit Model)

이제 순서형 반응변수에 대한 로짓 모형인 누적 로짓 모형이다! 누적 확률(cumulative logit model)이 무엇이었는지 떠올려보자.

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

1범주부터 j범주까지의 확률을 모두 더한 확률이다. 이 누적확률에 로짓연결함수를 씌우면 누적 로짓 모형이 된다.

$$logit[P(Y \le j|X=x)] = \log\left(\frac{P(Y \le j|X=x)}{1 - P(Y \le j|X=x)}\right) = \log\left(\frac{P(Y \le j|X=x)}{P(Y > j|X=x)}\right)$$

$$= \log\left(\frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_J(x)}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p, j = 1, \ldots, (J-1)$$

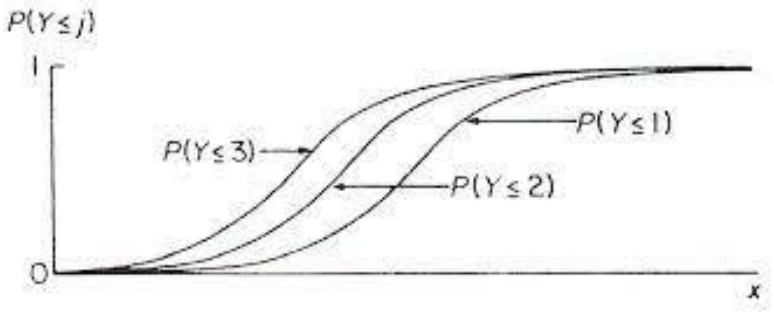
α_j가 다른 J-1개의 로짓 방정식이 만들어지는데, 이 때 절편 α_j를 제외하고 회귀계수 β_k에는 j첨자가 존재하지 않는다. 이는 **J-1개의 로짓 방정식에서의 회귀계수 β의 효과가 동일하다**는 것을 의미한다. 이것을 **비례 오즈(proportion odds) 가정**이라고 한다.

앞에서도 살짝 알아봤지만, 순서형 다항 로짓 모형은 범주(카테고리)의 순서를 고려하는데, cut point를 기준으로 범주들을 두 그룹으로 분류한다. ((ex) [매우 나쁨, 나쁨]과 [보통, 좋음, 매우 좋음] - 이 때

cut point는 나쁨과 보통 사이가 되겠죠?)

이렇게 cut point를 사용하여 두 그룹으로 나눌 때, 각각의 누적확률에 대하여 같은 비례상수(β)를 사용한다. 따라서 cut point가 달라도 β값은 항상 같다! 만약 비례 오즈 가정이 성립하지 않으면 Y변수가 순서형 변수라도 명목형 로짓 모형을 사용해야 한다. *(그러면 범주 별로 β_j를 추정하기 때문에 모수의 개수가 늘어나므로 더 복잡한 모형을 써야 한다는 단점이 있다.)*

비례 오즈 가정을 만족하면 β(기울기)가 고정되고 α(절편)만 달라지기 때문에 각 비례 오즈 모형의 곡선은 같은 모양을 가진다.



비례 오즈 모형에서의 누적확률

그렇다면 누적 로짓 모형은 어떻게 해석하면 될까? 마찬가지로 오즈를 이용하면 된다. **x가 1단위 증가하면** (다른 설명 변수가 고정되어 있을 때) **$[Y > j]$ 대신 $[Y \leq j]$일 오즈가 e^{β} 배 증가**한다고 해석할 수 있다.

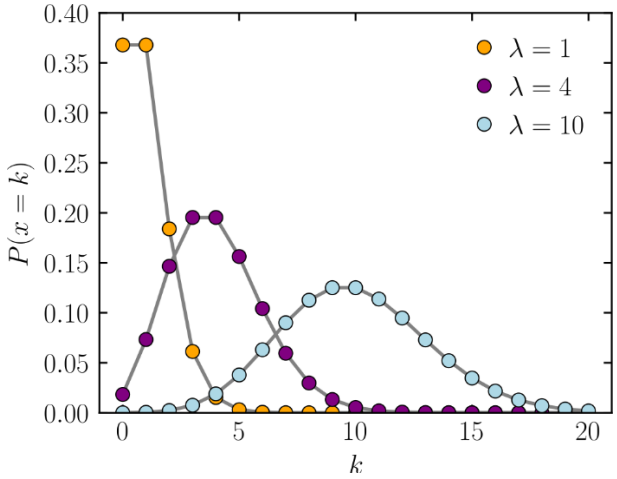
5. 포아송 회귀모형 (Poisson Regression Model)

■ 포아송 회귀모형이란?

포아송(Poisson) 분포가 무엇이었는지 기억해보자. *(통계학원론과 수리통계학에서 배웠다!)* 포아송 분포란 단위 시간 동안 어떤 사건이 일어난 건수 또는 횟수를 표현하는 이산 확률 분포이다. 따라서 구간에서 발생하는 사건의 횟수를 추정하는데 유용하다.

포아송 회귀모형은 이러한 포아송 분포를 따르는 **도수 자료(count data)**를 반응변수로 갖는 GLM이다. 로지스틱 회귀가 랜덤성분이 이항 분포를 따른다고 가정했다면, 포아송 회귀는 포아송 분포를 따른다고 가정하는 것이다.

포아송 분포를 따르는 반응변수의 평균이 작을 경우(10 미만), 일반 회귀 모형을 적합하면 표준 오차와 유의수준이 편향되는 문제가 발생한다. 왜 그럴까? 포아송 분포의 PDF(확률질량함수) 모양을 살펴보면 더 와 닿게 이해할 수 있다.



평균(위의 그림에서 λ)이 작을수록 작은 계급에 많은 관측치가 몰려 있으며, 오른쪽으로 길게 치우친 분포(skewed to the right)를 띠는 것을 볼 수 있다. 따라서 OLS 회귀 모형을 적합하면 정규성, 등분산성 등의 가정을 만족하지 못하게 된다.

또한 count data는 음수가 아닌 정수 값을 갖는데 OLS 회귀모형을 적합했을 때 음수값이 반환되는 문제도 생기게 된다. 로지스틱에서 로짓 연결 모형을 쓰는 것과 같은 원리로, 좌변과 우변의 범위를 맞춰주기 위해 포아송 회귀 모형은 로그연결함수를 사용하며, 모양은 다음과 같다.

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

■ 포아송 회귀 모형의 해석

포아송 회귀 모형 식 $\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ 를 변형하면 다음과 같이 도수를 표현할 수 있다.

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

추정된 회귀계수를 대입하면 기대도수, 즉 예측값 $\hat{\mu}$을 구할 수 있다.

또한 로지스틱 회귀와 마찬가지로 $x+1$과 x를 대입해서 빼주는 방식으로도 계수 β를 해석할 수 있다.

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

즉, **x가 1단위 증가하면** (다른 설명 변수가 고정되어 있을 때) **기대도수 μ가 e^{β} 배 증가한다!**

■ 과대산포(Overdispersion) 문제

포아송 분포의 가장 큰 특징 중 하나는 평균과 분산이 같다는 것이다. 즉 랜덤성분이 포아송 분포를 따른다고 가정한다면, 반응변수인 도수 자료의 평균과 분산이 같다는 가정을 만족해야 하는 것이다. 이를 **등산포 가정**이라고도 하는데, 문제는 이러한 가정을 만족하는 데이터가 실제로 많지 않다는 점이다. 일반적으로 분산이 평균보다 크게 나타나는 경우가 나타나는데, 이를 과산포 또는 **과대산포(overdispersion)** 문제라고 한다. 과대산포 문제를 무시하고 포아송 모형을 적합시키면 회귀 계수 추정량의 표준오차가 편향되어 작아지는 현상이 발생한다.

이러한 과대산포 문제가 발생할 때 적용할 수 있는 대안 모델로 음이항 회귀모형(Negative Binomial

Regression)이 사용된다. 음이항 회귀도 포아송 회귀와 마찬가지로 로그연결함수를 사용한다.

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

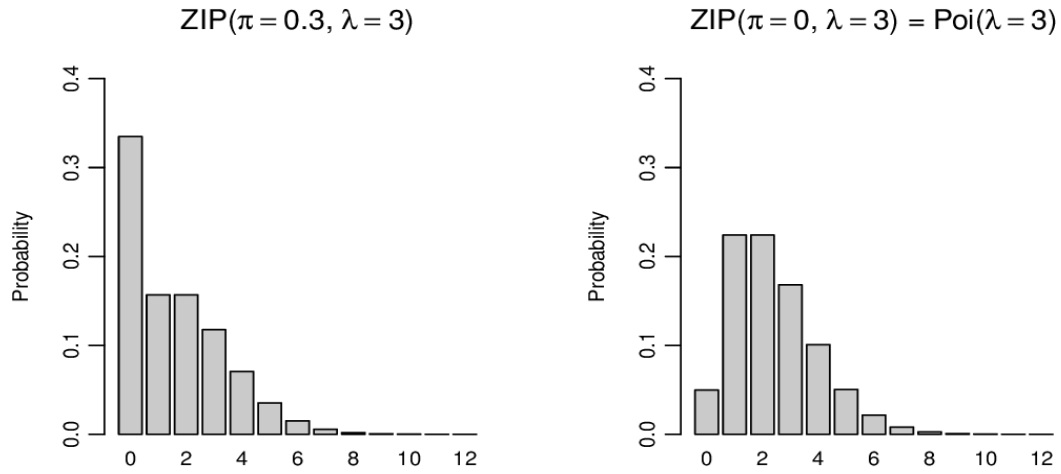
음이항 회귀 모형은 평균에는 영향이 없으면서 과대산포를 유발하는, 설명되지 않는 추가적인 가변성이 있다고 가정한다. 즉 분산이 평균보다 큰 값을 갖도록 하는 모수인 **산포모수 D**를 하나 더 갖는다.

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

이 때 산포모수 D가 0이면 포아송 회귀 모형과 동일하며, 모든 예측값에 대해 산포모수 D는 동일한 값을 갖는다고 가정한다. 참고로 R에서는 AER 패키지의 dispersiontest() 함수를 통해 쉽게 과대산포 검정이 가능하다.

■ 과대영(Excess Zeros) 문제

특정 평균 μ (또는 λ)를 갖는 포아송 분포에서 나타나는 '0'보다 표본 도수 자료가 더 많은 '0'을 갖는 경우 **과대영(excess zeros)**이 발생했다고 한다. 반응변수가 따른다고 가정하는 포아송 분포보다 0이 많이 나타나는 문제라고 생각하면 된다. 과대영 문제는 실제 데이터에서 생각보다 많이 발생한다. 예를 들면 학생들의 결석 일수를 표시하는 출석표를 생각해보자. 대부분 학생들은 결석을 하지 않으니까(^^) 대부분이 0인 데이터가 되는 것이다.



$\lambda = 3$일 때 제로팽창 포아송 분포와 일반 포아송 분포

이렇게 과대영 문제가 생겼을 때, ZIP 회귀모형을 대안으로 사용할 수 있다. **ZIP(Zero-Inflated Poisson, 제로팽창 포아송) 회귀 모형**은 크게 두 부분으로 나뉘는데, (1) 로지스틱 회귀모형을 사용하여 항상 '0'인 집단(structural zeros)과 항상 '0'은 아니지만 조사 시점에 '0'이라고 응답한 집단에 속할 확률을 구하고, (2) 포아송 회귀 또는 음이항 회귀모형으로 structural zeros를 포함하지 않는 나머지 관측치를 추정하는 모델을 적합하는 것이다. 즉 y_i에 대하여 두 가지 가능한 분포를 고려하는 것이다.

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

이 때 확률 ϕi는 제로팽창 확률로 0에서의 팽창확률, 즉 항상 0인 집단(structural zeros)이 될 확률을 뜻한다. 함수 $g(y_i)$는 항상 0은 아닌 집단이 따르는 포아송 또는 음이항 분포이다.

R에서는 pscl 패키지의 zeroinfl() 함수를 통해 ZIP모델 (또는 ZINB 모델)을 적합할 수 있다.

(ZIP과 ZINB 모델에 대해 더 알고 싶으면 링크를 참고하자..! 자세히 설명해주는 한글 자료가 많이 없다..
https://stats.idre.ucla.edu/r/dae/zip/ [ZIP], https://stats.idre.ucla.edu/r/dae/zinb/ [ZINB])