범주형 자료분석 2주차 교안

신나는 행복 범주 클린업 2주차 시간이 돌아왔습니다~(짝짝짝)

지난 시간에는 분할표, 독립성 검정, 연관성 지표에 대해 공부했었죠?

이번주는 GLM과 유의성 검정을 배워볼 예정입니다!

오늘의 주인공이자 범주의 핵심, 로지스틱 회귀도 배울 예정이니 기대해주세요~

벌써 2주차가 됐네요!

이번 주차가 끝나면 클린업의 절반 이상 한 것이니까 오늘도 힘내서 열심히 해봅시다~!

목차

1. GLM (Generalized Linear Model)

■ GLM이란?

1) 정의

일반화 선형 모형(GLM)은 연속형 반응변수에 대한 모형뿐만 아니라 범주형 반응변수에 대한 모형들을 모두 포함하는 모형의 집합이다. 쉽게 말해, 선형회귀모형처럼 우리가 알고 있는 모형들을 일반화해서 나타낸 더 넓은 범위의 모형이라고 할 수 있다. 일반화를 할 때 두가지를 일반화하는데, ①랜덤성분의 분포를 일반화하고, ②랜덤성분의 함수를 일반화한다. 랜덤성분? 무슨 말이지? 당황하지 말고 GLM의 구성성분을 살펴보자.

2) 구성 성분

GLM 구성 성분		
랜덤 성분	연결 함수	체계적 성분
$\mu\ (= E(Y))$	$g()$	$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$
$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$		

GLM의 모양은 맨 아래와 같고 랜덤성분, 연결함수, 체계적 성분 이 세 가지로 구성된다. 하나하나 살펴보도록 하자.

- 랜덤성분

랜덤성분(random component)은 반응변수 Y를 정의하는 성분이다. Y의 확률분포를 정해줌으로써 Y를 정의하고, 가정한 확률분포 하에서 Y의 기댓값인 μ로 표기한다. 쉽게 말해 그냥 우리가 알고 있는 Y랑 같고 표기만 다르다고 생각하면 된다.

표본크기가 N인 반응변수 Y의 관측값을 $Y_i (i = 1,2, \dots, N)$라고 하자. (각 Y_i는 독립으로 가정한다.) 만약 관측값 Y_i가 성공, 실패와 같은 이진형이면 이항분포로 Y를 정의하고, 이항분포의 평균인 $\pi(x)$로 랜덤성분을 표기한다. 혹은 관측값 Y_i가 몸무게 같은 연속형 자료라면 정규분포를 가정하고 평균인 μ로 랜덤성분을 표기한다. 또는 관측값 Y_i가 시간당 발생하는 횟수를 나타낸다면 포아송 분포를 가정하고 평균인 λ로 랜덤성분을 표기하는 식이다.

위에서 GLM은 ①랜덤성분의 분포를 일반화한다고 했는데, 이는 GLM은 랜덤성분의 분포로 아무 분포나 가질 수 있다는 의미이다. 선형회귀모형을 생각해보면, 오차항이 정규분포를 따라야 한다는 가정이 있다. 따라서 선형회귀모형의 랜덤성분으로 정규분포 밖에 사용할 수 없는 반면, GLM은 모든 분포가 가능하다. 즉, 선형회귀모형은 GLM의 한 종류인 셈이다.

- 체계적 성분

체계적 성분(systematic component)은 설명변수 X를 명시하는 성분으로, $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$와 같이 X의 선형결합으로 구성된다. x_i에는 x_a^3과 같이 고차항이나 교호작용을 설명하는 $x_a x_b$을 넣을 수 있다.

연결함수

연결함수(link function) $g()$는 랜덤 성분과 체계적 성분을 연결하는 성분으로, 두 성분의 범위를 맞춰주는 역할을 한다. 말그대로 연결을 위한 함수인 셈이다.

GLM의 형태를 보면 좌변에 랜덤성분, 우변에 체계적 성분이 있는데, 체계적 성분에서 설명변수 X가 연속형이면 그 범위가 $-\infty$부터 ∞까지가 된다. 만약 반응변수 Y도 연속형이면 상관없지만, 우리의 관심사인 범주형 변수라면 그 범위가 체계적성분과 맞지 않게 된다. 이럴 경우에 연결함수를 사용하여 양변의 범위를 맞춰준다.

위에서 GLM이 ②랜덤성분의 함수를 일반화한다고 했는데, 일반선형회귀에서 랜덤성분의 함수로 μ(=Y)만 썼던 것과는 달리, 연결함수 $g(\mu)$를 사용해서 아무 함수나 쓸 수 있다는 뜻이다! 대신 양변의 범위를 맞춰주는 함수를 선택해야 한다.

GLM의 연결함수의 종류는 다양한데, 대표적인 세 가지만 살펴보도록 하자.

- **항등 연결함수(Identity Link)** : $g(\mu) = \mu$

항등 연결함수는 반응변수 Y가 연속형일 때 사용하며, 앞서 말했듯 Y가 범주형인 경우에는 보통 양변의 범위가 맞지 않게 되기 때문에 다른 연결함수를 사용한다. 항등 연결함수를 사용하는 대표적인 사례는 일반선형회귀모형이다. 랜덤성분의 정규분포를 가정하고 연결함수로 항등함수를 사용한 식이 우리가 아는 일반선형회귀모형이다.

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

로그 연결함수는 반응변수 Y가 개수나 횟수를 나타내는 도수자료(count data)일 때 많이 사용한다. 주로 반응변수가 포아송 분포나 음이항 분포를 따를 때 사용한다.

- **로짓 연결함수(logit link)** : $g(\mu) = \log[\frac{\mu}{(1-\mu)}]$

로짓(logit)은 오즈에 로그를 씌운 것으로, 반응변수 Y가 0과 1사이의 값을 가질 때 유용하다. 주로 반응변수가 이항분포를 따를 때 사용한다. 이따 배우겠지만 연결함수가 로짓 연결함수고 랜덤성분을 이항분포로 가정한 GLM이 바로 "로지스틱 회귀"이다.

3) GLM의 필요성

반응변수가 범주형이거나 도수자료일 때는 오차항이 정규분포를 따르지 않기 때문에 일반선형회귀 모형을 사용할 수 없게 된다. 하지만, GLM은 일반선형회귀와는 달리 이런 상황에도 굴하지 않고 모형을 만들 수 있다. 이는 최소제곱법을 사용해서 모형을 적합하는 일반선형회귀와는 달리, GLM이 최대가능도법(Maximum Likelihood Method)을 사용해 모형을 적합하기 때문이다. 따라서 정규성 조건을 맞출 필요도 없기에 보다 더 포괄적인 범위의 반응변수도 다룰 수 있다.

1주차 때 살펴봤던 분할표와 비교했을 때도 GLM은 장점을 갖는다. 분할표는 독립성 검정으로 범주형 변수 간의 연관성을 파악만 하지만, 모형을 쓰면 변수 간의 연관성을 파악할 뿐만 아니라 반응변수를 예측할 수도 있다.(사실 이건 모형 그 자체의 장점이다..ㅎㅎ)

4) GLM의 특징

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

① 선형 관계식 유지

GLM은 선형관계식을 유지하기 때문에 해석이 용이하다는 장점이 있다. 하지만 아까 GLM의 체계적 성분의 x_i에는 고차항도 올 수 있다고 했는데 그러면 비선형이 아닌가? 생각할 수도 있다. 여기서 선형이라고 함은 설명변수 X와 반응변수 Y의 선형관계가 아니라 회귀계수 β의 선형성을 의미한다.

② 범위가 제한된 반응변수도 사용이 가능

앞서 말했듯 GLM에서 양변의 범위가 다른 경우, 연결함수 $g(\mu)$를 통해 범위를 조정할 수 있기 때문에 범위가 제한된 반응변수도 사용 가능하다.

③ 독립성 가정만 필요

GLM은 기존의 선형회귀분석의 가정인 정규성, 등분산성, 독립성, 선형성 중에서 "독립성" 가정만 만족하면 된다. 즉 오차항이 독립이라는 가정을 만족해야 한다는 뜻인데, 이를 위해 자기상관성 검정이 필요하다. 자기상관성(autocorrelation)이란 오차항이 서로 의존적인 것을 의미하고, 자기상관성은 일반적으로 더빈-왓슨 검정(DW Test)을 통해 확인한다. (자세한 내용은 회귀분석팀의 1주차 클린업을 참고~)

④ 오차항의 다양한 분포를 가정

앞서 말했듯 GLM은 정규성에 대한 가정을 지키지 않아도 되기 때문에 랜덤성분에는 오차항의 성질에 따라 어느 분포든 정의할 수 있다. 랜덤성분의 분포를 무엇으로 정의하는지에 따라 연결함수가 정해지고, GLM의 모형이 결정된다. 따라서 GLM은 다양한 종류가 존재한다.

■ GLM의 종류

<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>Quasi-Poisson 모형</td></tr>
<tr><td>율자료 포아송 회귀 모형</td><td>비율 자료</td></tr>
</table>

상당히 많은 종류가 있다! 이중에서 우리는 색칠된 모형을 살펴볼 예정이다. (6개뿐이니 안심ㅎ)

1) 이항 자료

- 선형 확률 모형 : $\pi(x) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

  : 이항 랜덤 성분 & 항등 연결 함수

- **로지스틱 회귀 모형** : $\boldsymbol{logit}[\boldsymbol{\pi}(\boldsymbol{x})] = \mathbf{log}\,(\frac{\boldsymbol{\pi}(\boldsymbol{x})}{\mathbf{1}-\boldsymbol{\pi}(\boldsymbol{x})}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_k x_k}$

  : 이항 랜덤 성분 & 로짓 연결 함수.

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

  : 이항 랜덤 성분 & 프로빗 연결 함수(표준정규분포 함수 $\Phi(\mu)$의 역함수)

2) 다항 자료

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

  : 다항 랜덤 성분(명목형) & 로짓 연결 함수

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

  : 다항 랜덤 성분(순서형) & 로짓 연결 함수

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

  : 다항 랜덤 성분(순서형) & 로짓 연결 함수

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

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

  : 다항 랜덤 성분(순서형) & 로짓 연결 함수

3) 도수 자료

- **포아송 회귀 모형** : $\mathbf{log}(\boldsymbol{\mu}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_k x_k}$

  : 포아송 랜덤 성분 & 로그 연결 함수

- **음이항 회귀 모형**: $\mathbf{log}(\boldsymbol{\mu}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_k x_k}$

  : 음이항 랜덤 성분 & 로그 연결 함수

- **Quasi-Poisson 모형**: $\mathbf{log}(\boldsymbol{\mu}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_k x_k}$

  : 포아송 랜덤 성분 & 로그 연결 함수

- 율자료 포아송 회귀 모형 : $\log(\mu/t) = \log(\mu) - \log(t) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

  : 포아송 랜덤 성분 & 로그 연결 함수

■ GLM의 모형 적합

앞서 GLM은 최대가능도 방법을 사용해서 모형을 적합한다고 언급했다. 모형 적합이란 주어진 데이터를 근거로 모형의 모수를 추정하는 것을 의미한다. 즉, GLM은 **최대가능도 추정법**(method of maximum likelihood estimation)을 통해 모형의 모수를 추정한다. 또 언급하는 거지만, GLM은 랜덤성분이 정규분포가 아닌 경우도 있기 때문에 일반선형회귀처럼 LSE(최소제곱추정법)은 사용해 모형을 적합할 수 없다.

가능도함수(likelihood function) $f(x;\theta)$는 결합확률밀도(질량)함수를 모수에 대해 정의한 함수로, 결합확률밀도함수 $f(x;p)$와 같다고 생각하면 된다. **가능도**는 관측값이 고정된 상태에서 그 관측값이 관찰될 가능성으로, 쉽게 말해 가능도함수에서 x값이 주어졌을 때의 함수값이다. 최대가능도 추정법이란 이 가능도함수가 최대가 되는 추정량 $\hat{\theta}$를 찾는 방법이다. 이렇게 해서 구한 추정량 $\hat{\theta}$을 **최대가능도 추정량**(MLE, Maximum likelihood estimator)이라고 한다. (자세한 내용은 통계적추론입문 시간에..) 따라서 GLM은 이 MLE로 구성된 모델이다.

2. 유의성 검정

■ 유의성 검정이란?

유의성 검정이란 모형의 모수 추정값이 유의한지를 검정하는 것을 의미한다. 혹은 축소 모형의 적합도가 좋은지에 대한 검정이기도 하다. 예를 들면 회귀분석에서는 t검정으로 회귀 계수 β가 유의한지 검정하거나 F검정으로 모델이 유의한지를 검정했었다. 여기서는 GLM에서 유의성 검정하는 방법에 대해 알아볼 예정이다. GLM 모형 $g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ 에 대하여 가설은 다음과 같다.

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

대립가설 H_1 : 적어도 하나의 β는 0이 아니다.

귀무가설은 모든 모수값이 0이다 즉, 모형은 쓸모없다라는 것이고, 대립가설은 적어도 하나의 모수값은 0이 아니다 즉, 모형은 그래도 쓸만하다는 것이다. 회귀분석에서 F검정을 통해 모형의 유의성을 검정한 것과 같은 가설을 갖는다!

■ ML을 이용한 유의성 검정

GLM은 MLE로 구성된 모델이라고 했다. 따라서 이에 맞는 검정 방법으로 유의성을 검정해야한다. MLE에 대한 추정량을 검정하는 방법 중 왈드 검정과 가능도비 검정을 알아볼 예정이다.

1) **왈드 검정**(Wald test)

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

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

: 왈드 검정은 추정값과 표준오차만 사용하기 때문에 간단하다는 장점이 있지만, 범주형이나 소표본인 경우 가능도비 검정보다 검정력이 떨어진다는 단점이 있어서 GLM의 유의성 검정에는 주로 가능도비 검정을 많이 사용한다.

2) **가능도비 검정**(Likelihood-ratio test)

1주차 독립성 검정에서 봤던 그 가능도비 검정과 같은 애다! 다른 것을 검정하다보니 가설과 검정통계량은 다르지만 원리는 같다. 저번에는 관측도수와 기대도수의 차이를 비교했다면, 이번에는 가능도함수의 차이를 비교한다.

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

- 기각역 : $G^2 \geq \chi^2_{\alpha,df}$

: 가능도비 검정은 귀무가설 하에서의 가능도함수 l_0과 전체공간 하에서의 가능도함수 l_1의 차이를 이용한다. 여기서 전체공간은 아무런 제약이 없는 상태, 즉 귀무가설 + 대립가설 하를 의미한다.

가능도 함수의 차이는 각 가능도 함수의 최댓값을 비교해서 알아본다. 가능도함수 l_1은 MLE일 때 최대값을 갖는다. l_0이 최대값을 갖도록 하는 추정량을 θ_0이라 하면, 함수의 차이가 적다는 것은 MLE와 θ_0이 가까운 값을 가진다는 것을 의미하고, 모수를 동등하게 잘 설명한다고 할 수 있다. 따라서 함수의 차이가 적을수록 $\frac{l_0}{l_1}$ 값이 1에 가까워지고, 검정통계량 G^2는 최솟값 0을 갖는다. 함수의 차이가 클수록 $\frac{l_0}{l_1}$은 1보다 작아지고 음의 부호와 로그 값을 갖는 검정통계량의 값은 커져서 귀무가설을 기각하게 된다.

가능도비 검정의 flow를 정리하자면,

l_0과 l_1의 차이가 커지게 되면 -> 검정 통계량이 커지고 -> p-value가 작아져서 -> 귀무가설을 기각하며-> 적어도 하나의 β는 0이 아니게 되므로 -> 내가 쓴 모형의 모수 추정값은 유의하다 -> 모형은 쓸만하다!고 할 수 있다.

가능도비 검정은 귀무가설 하에서와 전체공간 하에서의 가능도 함수에 대한 정보를 사용하기 때문에 가장 많은 양의 정보를 사용한다고 할 수 있다. 따라서 왈드 통계량 보다 검정력도 좋고 신뢰도도 높아 많이 쓰인다.

■ 이탈도

1) 관심모형과 포화모형

이탈도에 대해서 알아보기 전에 관심모형과 포화모형이 무엇인지 알아보자.

관심모형

관심모형 M은 말그대로 우리가 관심이 있는 모형, 즉 유의성을 검정할 모형을 말한다.

포화모형 (Saturated model)

포화모형 S는 관측값에 대하여 완벽하게 자료를 적합하는 모형, 즉 모든 관측값에 대해 모수를 갖는 가장 복잡한 모형이다

예를 들면, $\textit{범주팀 행복정도}\,(Y) = \beta_0 + \beta_1 \times \textit{지현이의 귀여움}(x_1) + \beta_2 \times \textit{패키지 난이도}(x_2)$

$+ \beta_3 \times \textit{스터디 시간}(x_3) + \beta_4 \times \textit{차농남의 드립력}(x_4)$

귀무가설 H_0 : 관심모형 M에 포함되지 않는 모수는 모두 0이다.

대립가설 H_1 : 적어도 하나는 0이 아니다.

이 두 모형을 사용한 가설은 위와 같다. 귀무가설이 맞다면 관심모형만이 데이터를 잘 설명하고 있다는 뜻이니 관심모형을 사용하고, 대립가설이 맞다면 관심모형이 아니어도 데이터를 잘 설명하는 모형이 있다는 뜻이니 관심 모형을 사용할 수 없다.

2) 이탈도란?

이탈도(deviance)란 포화모형 S와 관심모형 M을 비교하기 위한 가능도비 통계량이다. 가능도비 검정처럼 관심모형 M과 포화모형 S의 가능도 함수의 최댓값의 차이를 통해 이탈도를 계산한다.

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

이탈도는 S에는 있지만 M에는 없는 계수들이 0인지 확인하는 통계량이기 때문에 모형이 내포(nested)될 때만 (M의 계수 ⊂ S의 계수) 사용 가능하다.

이탈도 공식을 살펴보면 위에서 봤던 가능도비 검정 통계량 G^2과 모양이 같다. 따라서 이탈도를 활용해 모형이 데이터에 적합한 모형인지 검정이 가능하다. 검정의 flow는 다음과 같다.

가능도 함수의 최댓값 차이가 작으면 -> 이탈도가 작아지고 -> p-value가 커져서 -> 귀무가설을 기각 못하고 -> 관심 모형에 포함되지 않는 계수들이 0이니까 -> 관심모형 M은 쓸만하다! -> 관심모형M 사용

3) 이탈도와 가능도비 검정의 관계

가능도비 검정 통계량은 모형 간의 이탈도 차와 같다. M_0을 간단한 관심모형, M_1을 복잡한 관심모형, S를 두 모형을 모두 포함하는 포화모형이라고 하면,

M_0의 이탈도 - M_1의 이탈도 (=모형 간의 이탈도의 차)

$= -2(L_0 - L_S) - \{-2(L_1 - L_S)\}$

$= -2(L_0 - L_1)$ (=가능도비 검정 통계량)

따라서 모형 간의 이탈도의 차는 가능도비 검정 통계량과 같다고 할 수 있다. 이 성질을 이용해 비교하고 싶은 모형끼리의 이탈도 차이를 보면 어느 모형이 더 좋은 지 알 수 있게 된다. 쉽게 말해 관심모형 VS 관심모형이 되는 것이다!

하지만 이탈도를 통해 모형을 비교하려면 M_0은 M_1의 내포모형이어야 한다. 내포된 경우가 아니라면 모형을 비교할 때 이탈도를 사용할 수 없고 AIC나 BIC 등 모형 선택의 기준이 되는 측도를 통해 비교해야 한다.(자세한 건 회귀분석팀 클린업 참고~)

모형 간의 이탈도 차이를 통해 모형 M_0과 $M_1(M_0 \subset M_1)$을 비교할 때의 검정 flow는 다음과 같다.

관심모형 간의 이탈도 차이(=가능도비 검정통계량)가 작으면 -> p-value가 커져서 -> 귀무가설을 기각 못하고 -> M_0에 포함되지 않는 계수들이 0이니까 -> M_0이 더 쓸만하다! -> M_0 사용

3. 로지스틱 회귀 모형

■ 로지스틱 회귀 모형이란?

범주의 핵심이자 2주차의 주인공 로지스틱 회귀다! **로지스틱 회귀**(Logistic Regression)는 반응변수 Y가 이항자료일 때의 회귀를 의미한다.

$$Y \sim Ber(\pi)\,,$$

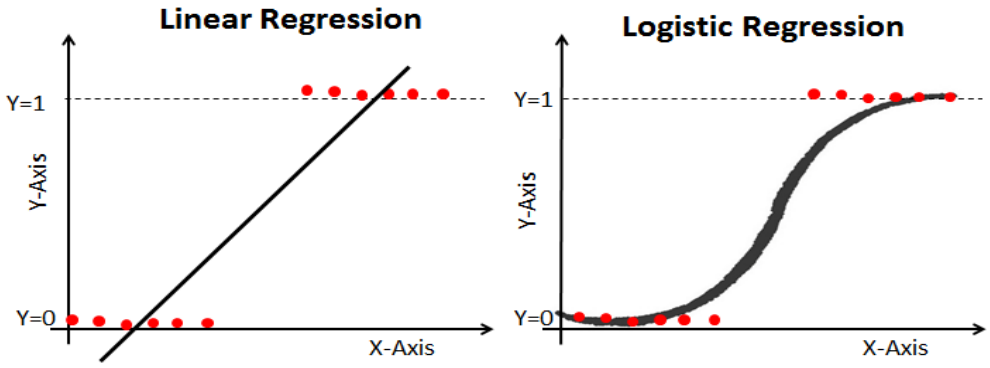
$$where\ \ \pi = P(Y=1)$$

반응변수가 성공(1) 또는 실패(0)의 이항분포를 따르기 때문에 기존의 일반 선형회귀는 사용할 수 없어서 GLM을 사용해 모형을 정의한다.

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

로지스틱 회귀의 모형은 이와 같다. 구성성분을 따져보자면, 랜덤성분은 이항 랜덤성분이고 연결함수는 로짓 연결함수이다.

여기서 체계적 성분의 범위는 $-\infty$에서 ∞인데, 원래 랜덤성분인 $\boldsymbol{\pi}(\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{Y}=\mathbf{1}|\boldsymbol{X}=\boldsymbol{x})$는 확률이니까 범위가 0~1이다. 따라서 범위를 맞추기 위해 오즈의 형태인 $\frac{\boldsymbol{\pi}(\boldsymbol{x})}{\mathbf{1}-\boldsymbol{\pi}(\boldsymbol{x})}$로 바꾸고(범위 : 0~∞), 로그를 취해서 범위를 $-\infty$에서 ∞로 맞추어 체계적 성분과 범위를 맞춘다.



로지스틱 회귀 모형의 함수 형태는 오른쪽과 같다. 빨간색 점처럼 Y가 0과 1뿐인 경우에, 선형회귀모형보다는 로지스틱 회귀 모형이 훨씬 데이터를 잘 설명하고 있는 것을 볼 수 있다. 로지스틱 회귀 모형의 함수를 **시그모이드 형태**라고 한다. 확률을 따르는 S자 곡선으로 $\pi(x)$와 x의 비선형 관계를 나타낸다.

로지스틱 회귀 모형은 GLM 답게 가정으로부터 자유롭다는 장점을 갖는다. 일반선형회귀와 달리 독립성 가정만 만족하면 된다. 또한 모형이 오즈비와 관련되어 있기 때문에 후향적 연구에도 사용할 수 있다.

■ 로지스틱 회귀 모형의 해석

로지스틱 회귀 모형을 해석하는 방법에 대해 알아보자. 로지스틱 회귀 모형 식을 변형하면, 다음과 같다.

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

만약 x값이 주어진다면 위 식에 대입해서 $Y = 1$일 확률을 알 수 있게 된다. 즉 $\pi(x) = P(Y = 1|X = x)$ 값을 알게 된다. 이 값이 0.5보다 크면 $Y = 1$, 작으면 $Y = 0$으로 예측하게 된다. 이 0.5를 cutoff point라고 하는데 항상 0.5인 것은 아니고, 자세한 내용은 3주차 클린업 때 배울 예정이다. 예시를 들어보자.

로지스틱 회귀모형 식이 $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 4 + 0.08x$이고 $Y = 1$(연애성공), $Y = 0$(실패), x는 키(cm)를 의미한다고 하자. 이를 정리하면 아래와 같은 식이 나온다.

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

이 식을 토대로 키 174cm인 장희가 연애를 성공할 수 있을지 살펴보자.

$\pi(174) = \frac{\exp(4+0.08\times174)}{1+\exp(4+0.08\times174)} = 0.99 > 0.5$이므로, 장희는 연애에 성공할 수 있다는 결론이 나온다!

방금은 확률로 모형을 해석하는 법을 알아봤다면, 이번에는 오즈를 이용하여 해석하는 법을 알아보자. 오즈를 이용한다면 회귀 계수 β의 의미도 파악할 수 있다. 로지스틱 회귀모형의 연결함수는 로짓(로그오즈) 연결함수이므로 오즈를 이용하여 해석할 수 있다. 이 때 모형에 각각 x와 $x + 1$을 대입한 뒤 빼 주면 오즈비의 형태가 나온다. 아래 수식을 살펴보자.

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

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

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

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

최종 식의 좌변의 분자는 $x + 1$일 때 $Y = 1$일 오즈이고, 분모는 x일 때 $Y = 1$일 오즈이다. 즉 좌변은 오즈비이다. 우리가 1주차에서 오즈비를 해석할 때 "분자가 될 오즈가 분모가 될 오즈보다 몇 배 높다"라고 해석했었다. 이를 적용해서 위 식을 해석하면, " $x + 1$일 때 $Y = 1$일 오즈가, x일 때 $Y = 1$일 오즈보다 e^{β}배 높다"고 해석할 수 있다. 즉, (다른 설명변수가 고정되어 있을 때) x가 한 단위 증가할 때 $Y = 1$일 오즈가 e^{β}배만큼 증가한다고 해석할 수 있다. 빼기로 오즈비를 만들어 주기 때문에 한 단위가 아니어도 해석이 가능하다. 만약 $x + 2$를 대입했다면 두 단위 증가할 때 얼마만큼 증가한다고 해석할 수 있다.

위의 예시를 다시 들어보자. 로지스틱 회귀모형 식이 $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 4 + 0.08x$이고 $Y = 1$(연애성공), $Y = 0$(실패), x는 키(cm)를 의미한다면, x가 한 단위(1cm) 증가할 때 $Y = 1$(연애성공)일 오즈가 $e^{0.08} = 1.08$배 증가한다고 해석하면 된다.

4. 다범주 로짓 모형

■ 다범주 로짓 모형이란?

다범주 로짓 모형(Multicategory Logit Model)은 랜덤성분이 다항분포를 따르고 연결함수가 로짓 연결함수인 GLM이다. 쉽게 말해 반응변수의 범주가 3개 이상인 로짓 모형을 의미한다. 앞에서 살펴본 로지스틱 회귀 모형은 연결함수가 로짓 연결함수로 동일하지만 반응변수가 성공/실패의 이항분포이고, 다범주 로짓 모형은 반응변수가 A/B/AB/O처럼 다항분포를 따른다는 점에서 차이가 있다.

반응변수의 범주가 3개 이상으로 늘어났기 때문에 명목형 자료인지 순서형 자료인지 구분할 필요가 있다. 자료의 종류에 따라 적용하는 모델이 달라지기 때문이다. 먼저 명목형 자료인 경우의 다범주 로짓모형을 살펴보자.

■ 명목형 다범주 로짓모형

명목형 다범주 로짓모형으로는 **기준범주 로짓모형**(Baseline-Category Logit Model)이 있다. 기준범주 로짓모형은 연결함수를 정할 때 반응변수 Y의 여러 범주 중에서 하나를 기준범주로 선택한 뒤, 기준범주와 나머지 범주를 짝지어 로짓을 정의한다. 만약 범주가 J개 있다면 기준범주 로짓은 다음과 같이 정의된다.

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

오즈의 분자는 j번째 범주일 확률, 분모는 기준범주 J인 확률로 로짓이 정의됐다. 연결함수를 만드는 법을 알았으니 완전한 GLM을 만들어보면 다음과 같다.

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^p x_p, j = 1, \dots, (J-1)$$

여기서 j는 j번째 범주를 뜻하고, J는 기준범주를 뜻하고, β에 붙어있는 지수는 제곱이 아니고 설명변수 x의 계수임을 뜻하는 첨자이다. 그런데 j는 왜 눈치 없이 모든 베타에 다 붙어있어서 우릴 헷갈리게 만들까?(년씨눈 ㅠ) 왜냐하면 기준범주 로짓모형은 1개의 식으로만 만들어지는 것이 아니기 때문이다. 반응변수 Y의 범주가 J개라고 하면, 기준범주 로짓모형은 J-1개의 로짓 방정식을 갖고, 각각의 방정식마다 서로 다른 β값을 갖는다. 만약 J=2라면, 반응변수 Y의 범주가 2개 즉 이항변수란 뜻이니 위에서 본 로지스틱 회귀모형이 된다!

여태까지 뭐가 지나간 건지 하나도 이해가 안 된다. 예시를 들어서 이해해보자! 반응변수 Y가 최애 블랙핑크 멤버라고 해보자. 그렇다면 Y의 범주는 제니,로제,지수,리사 4개가 될 것이다. 그 중에서 제니를 기준범주로 삼는다고 하면 로짓 모형은 다음과 같이 4-1=3개의 식으로 구성될 것이다.

$$\log\left(\frac{\pi_{로제}}{\pi_{제니}}\right) = 8 + 0.7x_1 + \cdots - 0.2x_p$$

$$\log\left(\frac{\pi_{지수}}{\pi_{제니}}\right) = 4 + 0.02x_1 + \cdots + 3x_p$$

$$\log\left(\frac{\pi_{리사}}{\pi_{제니}}\right) = -0.6 + 11x_1 + \cdots + 14x_p$$

각 식의 체계적 성분에서 계수의 값이 다 다름을 확인할 수 있다.(우연히 같을 수도 있겠지만!) 위에서 본 기본꼴에 j가 눈치 없이 붙어있는 이유가 바로 이 때문이다. 우리의 j는 모형이 몇 번째 모형인지 알려주는 알고 보니 친절한 첨자였던 것이다~(짜식 뭐라해서 미안ㅎ)

로지스틱 회귀 모형에서 했던 것처럼 모형 공식을 변형하여 확률에 대한 식으로 정리할 수도 있다.

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

분자에는 j범주일때의 식이 e의 지수로 들어가고 분모에는 전체 식의 합이 들어간다.블랙핑크 예시를 다시 들면 최애가 로제일 확률은 다음과 같다.

$$\pi_{로제} = \frac{e^{8+0.7x_1+\cdots-0.2x_p}}{e^{8+0.7x_1+\cdots-0.2x_p} + e^{4+0.02x_1+\cdots+3x_p} + e^{-0.6+11x_1+\cdots+14x_p}}$$

기존범주 로짓모형은 오즈와 기준범주를 통해 해석을 하는데, 기준범주에 비해 j범주일 로그 오즈를 보고 해석을 하는 식이다. 다범주 로짓모형의 식 하나를 보면,

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^p x_p,$$

이 경우에 (다른 설명변수가 고정되어 있을 때) x_i가 한 단위 증가할 때 J범주 대신 j범주일 오즈가 e^{β^i}만큼 증가한다고 해석한다. 위의 블핑 예시를 들면, (다른 설명변수가 고정되어 있을 때) x_1가 한 단위 증가할 때 최애가 제니 대신 로제일 오즈가 $e^{0.7} = 2.01$배만큼 증가한다고 해석할 수 있다.

기준범주와의 로짓끼리 빼서 해석하는 법도 가능하다. 무슨 말인지 아래 식을 살펴보자.

$$\log\left(\frac{\pi_2}{\pi_J}\right) - \log\left(\frac{\pi_1}{\pi_J}\right) = \left(\alpha_2 + \beta_2^1 x_1 + \cdots + \beta_2^p x_p\right) - \left(\alpha_1 + \beta_1^1 x_1 + \cdots + \beta_1^p x_p\right)$$

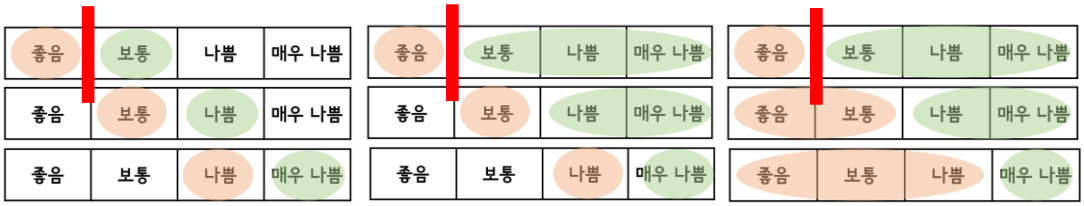
$$= [\alpha_2 - \alpha_1] + [(\beta_2^1 - \beta_1^1)x_1 + \cdots + (\beta_2^p - \beta_1^p)x_p]$$

이 경우에 (다른 설명변수가 고정되어 있을 때) x_i가 한 단위 증가할 때 1범주 대신 2범주일 오즈가 $e^{\beta_2^i - \beta_1^i}$만큼 증가한다고 해석하면 된다. 위에는 기준범주 VS 그냥 범주의 관계를 해석하는 방법이고, 지금 이 식은 그냥 범주 VS 그냥 범주의 관계 해석하는 방법이다.

■ 순서형 다범주 로짓모형

이제 순서형 반응변수에 대한 로짓모형인 순서형 다범주 로짓모형을 살펴보자. 순서형 다범주 로짓모형 역시 위에서처럼 기준범주를 정하고 범주끼리 비교하는 형식이다. 하지만 순서형 다범주 로짓모형은 순서 정보를 고려하기 때문에 범주를 순서대로 정렬시킨 후 두 덩어리로 나누는 collapse 과정이 필요한데, 이

때 collapse하는 기준인 cut point를 어떻게 정하는지에 따라 모형이 결정된다. 아래 그림을 보자.



첫번째처럼 나누면 이웃범주 로짓모형, 두번째는 연속비 로짓모형, 세번째는 누적 로짓모형이다. 이 중에서 우리는 모든 범주를 사용하는 누적 로짓모형만 살펴볼 예정이다.(궁금하면 따로 연락주세요~)

누적 로짓모형(Cumulative Logit Model)은 누적확률에 로짓 연결함수를 씌운 모형(GLM)이다. 일단 연결함수 부분인 누적 로짓을 살펴보자.

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

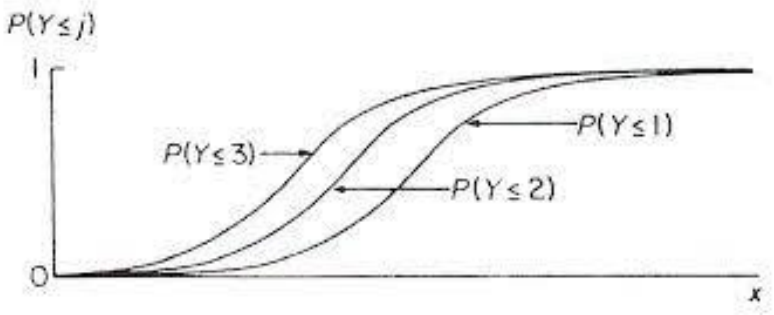
누적확률의 오즈 형태는 두번째 식과 같고, 1-누적확률은 결국 그 반대의 확률을 의미하니까 오즈는 j범주 이전과 j범주 이후로 나뉘게 된다. 따라서 위의 세번째 그림처럼 나눠지는 것이다.

최종적으로 누적 로짓모형의 형태는 다음과 같다.

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

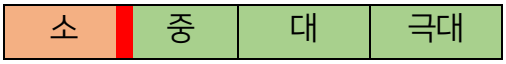
일단 모형의 체계적 성분에 첨자 j가 보인다. 누적 로짓모형 역시 기준범주를 정하고 비교하는 형식이기 때문에 때문에 위에서처럼 한 모형 안에 J-1개의 로짓 방정식이 생긴다. 하지만 이상하게도, 기준범주 로짓모형에서 모든 β에 껴있던 눈치 없는 j가 누적 로짓모형에서는 α에만 붙어있는 것을 확인할 수 있다.(깔끼빠빠 시전) 이는 한 모형 안에서 J-1개의 로짓 방정식의 회귀계수 β의 효과가 동일하다는 가정 때문이다. 이것을 **비례오즈 가정**(proportional odds)이라고 한다. 즉, 모형 안의 J-1개의 로짓 방정식은 α값만 다르고 동일한 β값을 갖는다.

비례오즈 가정 덕분에 누적 로짓모형은 다음과 같은 그래프를 갖는다.



방정식은 절편인 α값만 다르고 동일한 기울기 β값을 갖기 때문에, 위와 같이 똑같은 모양의 그래프가 그려진다. 이들은 기울기가 같기 때문에 교차하지도 않는다. 일종의 평행인 셈이다.

예시를 들어보자. 반응변수 Y가 회귀팀 심OO씨의 시비로 인한 장희의 빡침 정도로 소/중/대/극대의 순서형 범주를 갖는다고 할 때, 누적 로짓모형은 다음과 같다. 빨간 선은 cut point이다.



$$logit\left[P(Y \leq 소)\right] = 8 + 0.07x_1 + \cdots + 0.6x_p$$

$$logit\left[P(Y \leq 중)\right] = -5 + 0.07x_1 + \cdots + 0.6x_p$$

$$logit\left[P(Y \leq 대)\right] = 12 + 0.07x_1 + \cdots + 0.6x_p$$

누적 로짓모형은 다음과 같이 3개(4-1)의 로짓 방정식을 갖고, 비례오즈 가정으로 인해 회귀계수 β의 값 모두 같고 α값만 다름을 확인할 수 있다.

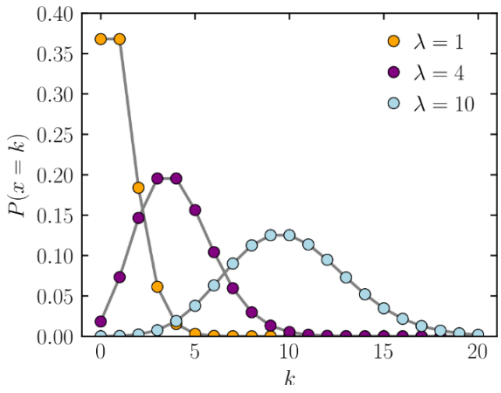
그렇다면 누적 로짓 모형은 어떻게 해석할 수 있을까? 기준범주 로짓모형처럼 오즈를 이용하면 된다. (다른 설명 변수가 고정되어 있을 때) x가 1단위 증가하면, $Y > j$ 에 비해 $Y \leq j$일 오즈가 e^{β}만큼 증가한다고 해석할 수 있다.

위의 예시를 보면 다른 설명 변수가 고정되어 있을 때, x_1이 1단위 증가하면, $Y > j$ 에 비해 $Y \leq j$일 오즈가 $e^{0.07} = 1.07$배만큼 증가한다고 해석할 수 있다.

5. 포아송 회귀 모형

지금까지 우리는 반응변수 Y가 이항분포와 다항분포를 따를 때 어떤 GLM모형을 쓰는지 살펴보았다. 이번엔 반응변수 Y가 포아송 분포를 따를 때의 GLM 모형을 살펴보고자 한다.

반응변수 Y가 포아송 분포를 따르게 되면 역시나 일반선형회귀를 사용할 수 없다. 정규성이나 등분산성을 만족하지 않기 때문이다. 따라서 포아송 분포를 따르는 도수 자료에 억지로 일반선형회귀를 사용하면 표준오차나 유의 수준이 편향되는 문제가 발생한다. 포아송 분포는 아래 그림처럼 그래프가 편향되어 있기 때문이다. (사실 평균 λ가 충분히 크면 아래 그림처럼 정규분포와 비슷한 모양을 갖기 때문에 일반선형회귀를 문제없이 사용 가능하다ㅎ) 어찌 됐건 반응변수 Y가 포아송 분포를 따를 때는 역시나 오늘의 주제 GLM을 써야한다.



■ 포아송 회귀 모형

포아송 회귀모형(Poisson Regression Model)은 반응변수 Y가 도수자료인 경우의 회귀모형으로, 랜덤 성분이 포아송 분포를 따르고, 연결함수가 로그 연결함수인 GLM이라고 할 수 있다. 모양은 다음과 같다.

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

연결 함수는 로그를 사용하는데, 이는 도수자료(μ)는 음수가 아닌 정수 값($0 \sim \infty$)을 갖는데 반해 체계적 성분은 ($-\infty \sim \infty$)가 범위이므로 이를 맞춰 주기 위함 때문이다.

포아송 회귀모형을 해석하는 방법은 로지스틱 회귀와 비슷하다. 하나는 식을 변형해 도수로 나타내서 해석하는 방법이다. 식을 변형하면 다음과 같이 도수를 표현할 수 있다.

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

다른 하나는 역시 차이를 이용해서 해석하는 방법이 있다. 포아송 회귀모형 식에 $x+1$과 x를 대입해서 빼 주면 다음과 같은 결과가 나온다.

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

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

이것은 (다른 설명변수가 고정되어 있을 때) x가 한 단위 증가할 때 기대도수 μ가 e^{β}배만큼 증가한다고 해석할 수 있다. 그렇다면 이해를 위해 예시를 들어보자.

반응변수 Y를 살면서 연애하는 횟수라고 하고, x는 키(cm)를 의미한다. 이때 포아송 회귀 모형의 식이 $\log(\mu) = -2 + 0.01x$라고 하자. 따라서, 키가 1cm 증가할 때 기대도수 μ가 $e^{0.01} = 1.01$배만큼 증가한다고 해석할 수 있다.

위 식을 정리하면, $\mu = \exp\left(-2 + 0.01x\right)$가 된다. 이 식을 토대로 키 174cm 장희가 살면서 연애를 몇 번 할 수 있을지 예측해보자.

$$\mu = e^{-2+0.01\times 174} = e^{-0.26} = 0.77$$

기대도수 값이 1이 안된다. 그렇다 장희는 살면서 단 한 번도 연애를 할 수 없는 것이다.

■ 과대산포 문제

포아송 분포는 평균과 분산이 같다는 특징이 있다. 이를 **등산포 가정**이라고 한다. 하지만 현실에서 이를 만족하는 데이터는 많지 않다. 일반적으로는 분산이 평균보다 더 크게 나타난다. 이런 현상을 **과(대)산포**(Overdispersion)문제라고 한다. 이 과대산포 문제를 무시하고 포아송 모형을 적합시키면 회귀계수의 표준오차를 더 작게 왜곡시켜서 검정의 결과가 부정확해지는 오류가 발생한다. 과대산포 유무는 과산포 검정을 통해 확인할 수 있고, 문제가 있다고 결과가 나오면 여러가지 방법으로 해결할 수 있다. 그 중에서 우리는 Quasi-Poisson 모형과 음이항 회귀모형을 통해 해결하는 법을 알아볼 예정이다.

1) Quasi-Poisson 모형

Quasi-는 사전적 의미로 유사,준-의 의미를 갖는다. 즉 **Quasi-Poisson 모형**은 유사 포아송 혹은 준포아송 모형으로 포아송 모형과는 비슷하면서도 조금은 다른 모형이라고 생각하면 된다. Quasi-Poisson 모형의 형태를 살펴보자.

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

Quasi-Poisson 모형은 포아송 모형과 동일하게 포아송 랜덤성분과 로그연결함수로 이루어진 GLM이다. 하지만 포아송 모형과 달리 분산이 평균보다 큰 값을 가질 수 있도록 하는 산포모수 θ를 추가한다. 이때 Quasi-Poisson 모형은 분산이 평균과 선형관계에 있음을 가정한다. 따라서 분산을 평균의 선형함수 형태로 표현한다.

$$E(Y) = \mu, \qquad Var(Y) = \theta\mu$$

만약 산포모수 θ가 1이면 포아송 회귀모형과 같다.

2) 음이항 회귀모형

음이항 회귀모형(Negative Binomial Regression Model) 역시 과대산포 문제가 발생했을 때 사용 가능하다. 음이항 회귀 모형은 음이항 랜덤성분과 로그연결함수로 이루어진 GLM이다. 애초에 음이항 분포는 분산이 평균보다 큰 분포이기 때문에, 포아송 분포의 등산포 가정을 완화하기 위해 랜덤성분으로 음이항 분포를 사용한 것이다. 이때, 음이항 회귀 모형은 분산이 평균과 비선형관계에 있음을 가정하고 산포모수 D를 사용하여 2차함수의 형태로 표현한다.

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

만약 산포모수 D가 0이면 포아송 회귀모형과 같다.

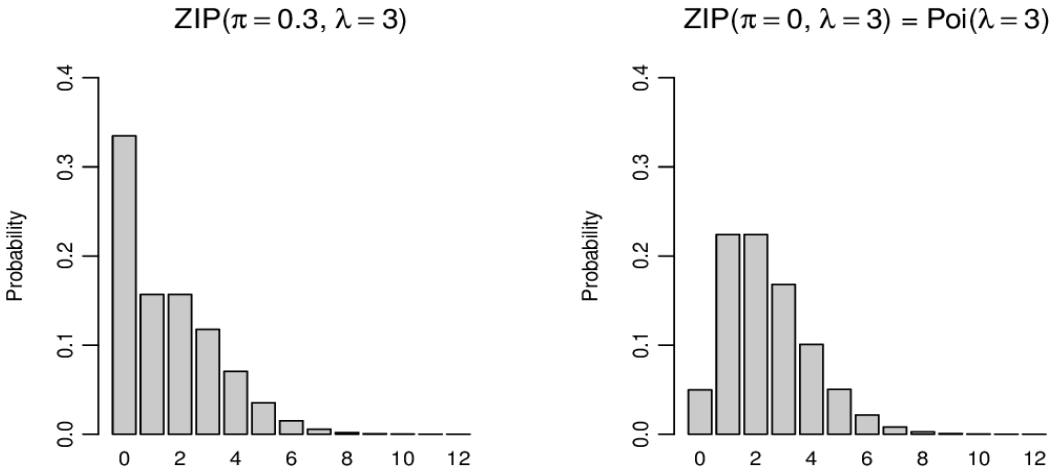
회귀분석에서 등분산성을 만족하지 못할 때 가중치를 두어 회귀계수를 구하는 WLS(Weighted Least Square) 방법이 있었다.(자세한 내용은 회귀분석팀 클린업에서~) Quasi-Poisson 모형과 음이항 회귀모형 역시 산포모수를 이용해 가중치를 둔다는 점에서 이와 같은 원리라고 생각하면 된다.

과대산포 문제가 있을 때 Quasi-Poisson 모형이나 음이항 회귀모형을 사용할 경우 포아송 모형과 회귀계수 β값은 동일하지만, 포아송 모형보다 회귀계수의 표준오차가 증가해서 검정의 결과가 조금 더 정확해진다.

Quasi-Poisson 모형과 음이항 회귀모형 중 어느 것이 더 좋은지는 판별할 수 없다. 따라서 데이터가 주어졌을 때 과대산포 문제가 있다고 판단되면 두 모형 모두 사용해본 후 비교해서 더 나은 모형을 선택하는 것이 바람직하다!

■ 과대영 문제

포아송 회귀모형에서 발생할 수 있는 또 다른 문제는 **과대영**(Excess Zeros) 문제이다. 포아송 분포에서 예상되는 것보다 더 많은 0이 나타날 때 과대영 문제가 발생했다고 한다. 아래 그림을 보면 이해가 바로 된다.



왼쪽 그림은 과대영 문제가 발생했을 때 그래프이고, 오른쪽 그림은 일반 포아송 분포의 그래프이다. 이처럼 원래 포아송 분포의 0보다 훨씬 많은 0이 나타날 때를 과대영이라고 한다. 생각보다 과대영 데이터는 흔하다. 예를 들어 로또 당첨자 수 데이터를 보면 대다수의 사람들은 당첨이 되지 않으니 당첨자 수는 0이 많이 있을 것이다.

이렇게 과대영 문제가 생겼을 때 영과잉 음이항 회귀모형(ZINB) 또는 영과잉 포아송 모형(ZIP)을 사용할 수 있다. 여기서는 ZIP만 살펴볼 예정이다. (궁금하면 저에게 따로 말해주세용~)

영과잉 포아송 모형(ZIP, Zero Inflated Poisson)에서 반응변수 Y는 0의 값이 발생하는 점확률분포와 0보다 큰 정수값을 갖는 포아송 분포의 혼합분포 구조를 가지고 있다. 이를 쉽게 적어보면 아래와 같다.

$$Y = \begin{cases} 0\ , & with\ probability\ p \\ 포아송\ 분포(평균\ \lambda)\ , & with\ probability\ 1-p \end{cases}$$

여기서 p는 베르누이 확률로, 즉 Y는 확률 p로 0이 되거나 확률 1-p로 평균 λ인 포아송 분포를 따르게 된다. 쉽게 말해 ZIP은 영과잉인 부분(0)과 0이 아닌 부분(포아송)을 나누는 모형이다. 만약에 0이 아닌 부분을 음이항 분포로 가정하면 ZINB가 된다.

이 ZIP을 이용해서 GLM을 만들 수 있다. 이는 영과잉 포아송 회귀모형(ZIPR)이라 불린다. 이 모형에는 아래처럼 p에 대한 식과 λ에 대한 식이 있다. 전자는 로짓 연결함수를 사용하고 후자는 로그 연결함수를 사용한다.

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

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

과제는 코드 실습한 거 반드시 이해해오기!!

다음주 예고!

혼동행렬, 평가지표, 샘플링, 인코딩을 배울겁니다!

그때까지 안뇽~