로지스틱 회귀모형의 기초

1 로지스틱 회귀 분석의 단순 모형

일반적으로 지금까지 배워온 회귀분석의 확률 모형에서는 반응변수 y는 연속형 확률변수이다. 따라서 예측변수 x의 값과 반응변수의 관계를 다음과 같은 회귀식으로 설명한다.

$$E(y|x) = \beta_0 + \beta_1 x \tag{1}$$

하지만 반응변수의 값이 연속형 변수가 아니라 두 개의 가능한 결과만을 가지는 이항변수라면 위에서 주어진 회귀식은 적절하지 못하다. 왜냐하면 반응변수의 기대값이 0과 1사이의 확률로 나타나기 때문이다.

$$E(y|x) = 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x) = P(y=1|x)$$

따라서 반응변수의 기대값의 범위와 예측변수가 있는 선형예측식(linear predictor) $\beta_0 + \beta_1 x$의 범위가 일치하지 않아서 선형회귀식 (1)을 그대로 사용할 수 없다.

위의 문제를 해결하기 위한 방법중의 하나는 다음과 같은 함수 m를 생각하여 변환된 선형예측식의 범위를 $[0,1]$로 만드는 것이다.

$$m : \Re \to [0,1] \quad \text{and } m(x) \text{ is monotone function.}$$

따라서 다음과 같은 이항변수를 반응변수로 하는 새로운 회귀식을 만들 수있다.

$$E(y|x) = m(\beta_0 + \beta_1 x) \tag{2}$$

주로 쓰이는 변환함수로 다음과 같은 로지스틱 함수(logistic function)가 있다.

$$m(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3}$$

반응변수가 베르누이분포를 따를 때 위의 로지스틱활수를 사용하는 회귀식을 로지스틱 회귀식이라고 한다.

$$P(y=1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1} \tag{4}$$

위의 로지스틱 회귀식을 다시 역으로 정리하면 다음과 같은 식을 얻을 수 있다.

$$\log\left[\frac{P(y=1|x)}{1 - P(y=1|x)}\right] = \log\frac{p(x)}{1-p(x)} = g(p(x)) = \beta_0 + \beta_1 x \tag{5}$$

식 (5)에서 나타난 함수 $g(p) = \log[p/(1-p)]$를 로짓함수(logit function)이라고 부르며 이는 로지스틱 함수의 역함수로서 0과 1 사이의 값을 가지는 확률을 실수 전체로 변환하는 함수로서 선형예측식의 범위와 일치하게 한다.

이렇게 관측값의 평균 (베르누이분포에서는 성공확률)과 선형예측식의 관계를 설정하는 함수를 결합함수(link function)라고 하며 g라고 표시한다.

$$g[E(y|x)] = g(p(x)) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x \tag{6}$$

따라서 로짓함수는 결합함수의 하나이며 다른 종류의 결합함수도 생각할 수 있다. 예를 들어 $\Phi(x) = P(Z \leq x)$를 표준정규분포의 분포함수라 한다면 다음과 같은 결합함수를 생각할 수 있고 이를 probit함수라고 부른다.

$$g(p(x)) = \Phi^{-1}(p(x)) = \beta_0 + \beta_1 x$$

2 로지스틱 회귀분석에서 계수의 의미: 오즈 비

일반적인 회귀분석의 모형 (1)에서 계수 β_1은 기울기로서 예측변수 x의 단위가 1 증가할 때 반응변수의 평균이 β_1만큼 증가하는 것으로 해석할 수 있다. 하지만 로지스틱 회귀모형 (4) 에서는 이러한 해석을 할 수 없다.

로지스틱 회귀모형에서 기울기 β_1의 의미를 알아보기 위하여 다음과 같은 주어진 확률 p에 대한 몇 가지 함수를 알아야 한다.

- 오드 (odd)

  $$\text{odd} = \frac{p}{1-p}$$

  예로 성공의 확율이 1/3일 때 오드는 1/2가 되며 이는 평균적으로 세 번의 시행할 때 한 번 성공하고 두 번 실패한다는 의미이다. 반대로 성공의 확율이 2/3일 때 오드는 $2 = 2/1$가 되며 이는 평균적으로 세 번의 시행할 때 두 번 성공하고 한 번 실패한다는 의미이다. 성공의 확률이 1/2일 때 오드는 1이 된다.

- 오즈 비 (odds ratio)

  $$\text{odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

  오즈 비는 두 개의 성공 확률 p_1과 p_2를 비교할 때 쓰는 양이다. 두 개의 오드를 비율로서 비교하는 양이며 오즈 비가 1일 경우에 두 확률은 같다.

이제 단순 로지스틱 회귀식 (5)을 생각하고 예측변수 x를 0과 1의 값을 가지는 이항변수로 가정한다. $x = 1$인 경우는

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_0 + \beta_1)$$

Table 1: 나이와 만성심장질환의 관계

	나이 ≥ 55 $(x=1)$	나이 < 55 $(x=0)$	합
CHD 있음 $(y=1)$	21	22	43
CHD 없음 $(y=0)$	6	51	57
합	27	73	100

이며 $x=0$인 경우는

$$\frac{P(y=1|x=0)}{1-P(y=1|x=0)} = \exp(\beta_0)$$

위에서 주어진 두 개의 오드, 즉 $x=1$인 경우와 $x=0$인 경우의 두 오드의 비를 구하면 다음과 같다.

$$\frac{\frac{P(y=1|x=1)}{1-P(y=1|x=1)}}{\frac{P(y=1|x=0)}{1-P(y=1|x=0)}} = \exp(\beta_1)$$

이는 다시 쓰면

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_1)\frac{P(y=1|x=0)}{1-P(y=1|x=0)}$$

위의 식에서 볼 때 예측변수 x가 1 의 값을 가질 때 반응 변수의 오드가 예측변수 x가 0일 경우의 오드의 $\exp(\beta_1)$배로 변하는 것을 알 수 있다. 따라서 $\exp(\beta_1)$는 반응변수의 오드의 증가량으로 볼 수 있다. 이는 두 성공확률의 오즈 비가 $\exp(\beta_1)$을 말한다. 위의 식에 로그를 취하면 다음과 같은 관계를 얻는다.

$$\log\left[\frac{P(y=1|x=1)}{1-P(y=1|x=1)} \Big/ \frac{P(y=1|x=0)}{1-P(y=1|x=0)}\right] = \beta_1$$

즉 오즈 비의 로그값이 단순 로지스틱 회귀식에서 기울기 β_1으로 나타난다.

간단한 예제를 통하여 오즈비와 로지스틱 회귀의 기울기의 관계를 명확히 해보자. 100명의 사람들을 55세 이상의 사람($x=1$)과 55세 미만의 사람($x=0$)의 그룹으로 나누었을 때 각 그룹에서 만성심장질환(CHD)가 있는 사람($y=1$)과 없는 사람($y=0$)의 수가 표 1 에 주어져있다. 여기서 나이에 대한 CHD유무의 오즈비는 다음과 같이 계산된다.

$$\text{Odds Ratio } = \frac{\frac{21/27}{6/27}}{\frac{22/73}{51/73}} = \frac{\frac{21}{6}}{\frac{22}{51}} = 8.11$$

위의 표 1의 자료를 이용하여 로지스틱회귀를 적합시키면 결과가 아래와 같고 회귀계수 β_1의 추정값은 오즈비의 로그값임을 알 수 있다.

$$\hat{\beta}_1 = \log(8.11) = 2.094$$