범주형자료분석팀

2팀
박지성
박지민
서희나
윤경선
이지윤

INDEX

GLM의 필요성

변수 간의 연관성을 파악 및 반응변수 예측 가능

분할표

✓ 범주형 변수들간의 연관성 파악
(독립성 검정)

GLM

✓ 범주형 변수와 연속형 변수 간의
연관성 파악 가능

✓ 새로운 설명변수가 주어진다면
반응변수 예측 가능

GLM의 구성성분

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

랜덤 성분과 체계적 성분을 연결하는 역할

두 성분의 범위를 맞춰주는 역할

만약 반응변수가 **이항 자료**, 설명변수가 **연속형 자료**이라면?

$[0,1]$ $\mathbf{g(\mu)} \neq \boldsymbol{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}$ $(-\infty,\ \infty)$

양변의 범위가 다름! → 연결 함수 사용

1 GLM

GLM의 특징

선형 관계 유지

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

'선형'의 관계 → 회귀 계수 β의 선형성

독립성 가정만 필요

오차항의 독립성만 만족하면 됨

자기상관성을 검정해야 됨

GLM의 모형 적합

모형 적합(Model Fitting)

주어진 데이터를 근거로 모형의 모수를 추정하는 과정

GLM은 **최대가능도 추정량(MLE)**로 **구성**된 모델!

최대가능도추정법

(Maximum Likelihood Estimation)

가능도 함수가 최대가 되는

추정량 $\hat{\theta}$ 찾기

독립성만 만족하면 됨

유의성 검정의 종류

또 보네, 가능도비 ☺ →

가능도비 검정(LR test)

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

✓ 기각역 : $G^2 \geq \chi^2_{\alpha,df}$

l_0 : 귀무가설 하에서의 가능도 함수 / l_1 : 전체 공간에서의 가능도 함수

df : H_0와 H_1 의 모수의 개수 차이

두 가능도 함수의 최댓값을 비교하는 방식

2 유의성 검정

이탈도

포화모형과 관심모형을 비교하기 위한 가능도비 통계량

$$\textit{이탈도(deviance)} = -2\log\left(\frac{l_m}{l_s}\right) = -2(L_m - L_s)$$

L_m : 관심도형에서 얻은 로그 가능도 함수의 최댓값

L_s : 포화모형에서 얻은 로그 가능도 함수의 최댓값

두 모형의 가능도 함수의 최댓값의 차이를 계산

2 유의성 검정

이탈도와 가능도비 검정의 관계

M_0 의 이탈도 - M_1의 이탈도 = -2$(L_0 - L_s) - \{-2(L_1 - L_s)\} = -2(L_0 - L_1)$

M_0 : 단순한 형태의 관심모형 / M_1 : 복잡한 형태의 관심모형 /

S : 두 모형을 포함하는 포화모형

두 모형 간 이탈도의 차 = **가능도비 검정 통계량**

이탈도를 활용하기 때문에 모형 M_0 은 모형 M_1에 내포된 모형이어야 함

양변의 범위를 맞추는 과정

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

좌변 범위 : 0~1 ≠ 우변 범위 : −∞~∞

좌변을 오즈 형태로

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

좌변 범위 : 0~ ∞ ≠ 우변 범위 : −∞~∞

좌변을 로그를 취하기 (로짓 연결함수)

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

좌변 범위 : −∞ ~ ∞ = 우변 범위 : −∞~∞

모형의 해석

오즈비를 이용한 해석

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

다른 설명변수가 모두 고정되어 있을 때

x가 한 단위 증가하면 $Y = 1$일 **오즈가 e^{β}배만큼 증가**

기준 범주 로짓 모형 | 형태

기준 범주 로짓 : j번째 범주에 속할 확률(분자), 기준범주에 속할 확률(분모)로 구성

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

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

- ✓ j : 범주에 대한 첨자, J : 기준 범주에 대한 첨자, $1\sim K$: 설명변수에 대한 첨자
- ✓ 총 $J - 1$개의 로짓 정의 → 그에 따라 $J - 1$개의 로짓 방정식 정의

4 다범주 로짓모형

누적 로짓 모형 | 형태

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

기준 범주 로짓 모형: $log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j^1 x_1 + \cdots + \beta_j^K x_K \quad , j = 1, \ldots, (J-1)$

비교 w. 기준 범주 로짓 모형

공통점

✓ 기준점을 두고 이분화된 두 범위의 확률을 비교하는 형식

→ J -1 개의 로짓 방정식이 만들어 진다.

차이점

✓ 누적 로짓 모형에선 $J-1$개의 로짓 방정식에 대한 β의 효과가 모두 동일하다고 가정

→ 비례 오즈 가정 (proportional odds)

포아송 회귀 모형 | 형태

반응변수가 도수자료인 경우에 사용되는 회귀 모형

랜덤성분이 포아송 분포를 따름

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

도수자료(μ) : 0~ ∞ 사이의 값을 지님

체계적 성분과 범위를 맞추기 위해 로그 연결함수 사용

율자료 포아송 회귀 모형 | 형태

랜덤성분은 포아송 분포, 연결함수는 로그 연결함수로 구성된 GLM모형

$$\log(\mu/t) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_p$$

지표값 : 수정항(offset)으로 지칭

→ 비율을 구할 때 분모에 들어가는 값

포아송 회귀모형의 문제점

과대산포 문제 – 해결 : 음이항 회귀 모형

음이항 랜덤성분과 로그 연결함수로 구성된 GLM

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

음이항 분포의 분산 → 포아송 분포와 다르게 평균에 $D\mu^2$ 가 더해짐

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

산포모수(D) : 음이항 분포에서 분산이 평균보다 큰 값을 갖도록 만드는 요소

포아송 회귀모형의 문제점

과대영 문제 – 해결 : 영과잉 포아송 회귀 모형

영과잉 포아송 분포

Y가 0의 값을 가질 확률

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

Y가 0이 아닌 경우에 따르는 포아송 분포

0의 값만을 갖는 점확률 분포와
포아송 분포의 혼합분포구조