범주형자료분석팀

2팀

김찬영
이혜인
김서윤
심은주
진수정

INDEX

0. 지난 주 리뷰

1. GLM

2. 유의성 검정

3. 로지스틱 회귀 모형

4. 포아송 회귀 모형

5. 로그 선형 모형

6. 부록

GLM 일반화 선형모형

범주형 반응변수에 대한
비선형 관계

연속형 반응변수에 대한
선형 관계

ex) 회귀모형, 분산분석 모형

→ 범주형 반응변수에 대한 모형까지 포함하는
광범위한 모형의 집합

- 특징

1. 비선형 관계를 포함

2. 선형 관계식 유지

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_i$$ ➡ 해석 용이

3. 범위가 제한된 반응변수 사용가능

: 연결함수 $g(\mu)$로 범위 맞추기

4. 독립성 가정만 필요

~~선형성~~ ~~등분산성~~ ~~정규성~~ "독립성"

"유의성 검정"

- 모형의 모수 추정 값이 유의한지 검정
- 축소 모형의 적합도가 좋은지에 대한 검정

• 가설

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

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

H_1: 적어도 하나의 β는 0이 아니다

• 종류

왈드 검정	스코어 검정	가능도비 검정

"가능도비 검정"

$$-2\log\left(\frac{l_0}{l_1}\right) = -2(L_0 - L_1) \sim X_1^2$$

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

H_1: 적어도 하나의 β는 0이 아니다

l_0: 귀무가설 하에서 계산되는 가능도 함수의 최댓값

l_1: MLE에 의해 계산되는 가능도 함수의 최댓값

$$-2\log\left(\frac{\text{모수가}H_0\text{을 만족할 때의 가능도 함수의 최대값}}{\text{모수에 대한 아무런 제한 조건이 없는 완전모형의 가능도 함수의 최대값}}\right)$$

- 가설

H_0: 관심 모형에 포함되지 않는 모수는 모두 0 ➡ 관심 모형을 사용하자!

H_1: 관심 모형에 포함되지 않는 모수 중 적어도 하나는 0이 아님 ➡ 관심 모형은 안되겠군!

- 이탈도: $-2(L_M - L_S)$

  - 포화모형과 관심모형을 비교하기 위한 가능도비 통계량
  - S에는 있지만 M에는 없는 계수들이 0인지 확인 가능 → 모형이 Nested일 때만 사용 가능!
  - 모형 적합도 검정에도 사용
  - 근사적으로 카이제곱분포 따름

Logistic Model : logit을 link function으로 사용

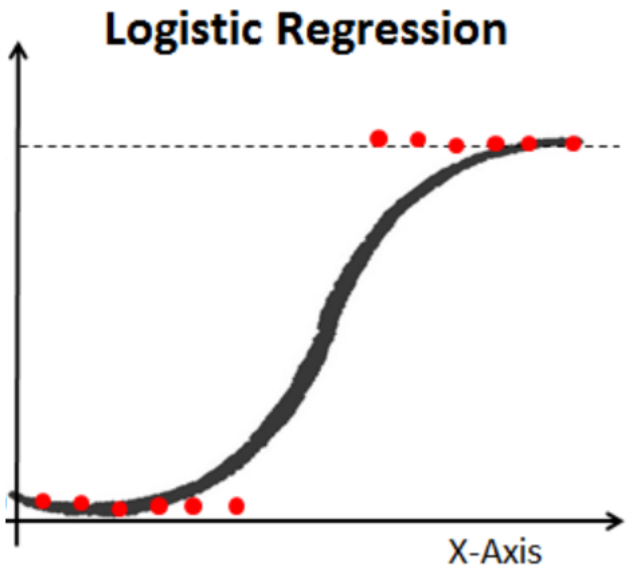
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta \quad \Leftrightarrow \quad \pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

- 범위 문제 해결

$$0 \le \pi_i \le 1, \qquad 0 \le 1-\pi_i \le 1$$

$$0 \le \frac{\pi_i}{1-\pi_i} < \infty$$

$$\Rightarrow \quad -\infty < \log\left(\frac{\pi_i}{1-\pi_i}\right) < \infty$$



Poisson Regression : log를 link function으로 사용

$$\log\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- 평균과 분산이 같아야 함

  E(Y)= μ, Var(Y)= μ

- 예측된 것보다 실제 분산이 더 큰 과대산포 문제가 발생하기도 함
- 반응변수(도수자료)를 이항자료로 범주화하여 로지스틱 회귀로도 분석 가능

  예시) [있다, 없다]로 범주화

율자료 포아송 회귀모형

: 반응변수 Y가 비율자료(rate)일 때 사용

$$\log\frac{\mu}{t} = \log\mu - logt = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

수정항

- 수정항(offset)이 존재
- 어떤 사건이 시간 혹은 공간별로 크기를 나타내는 다른 지표(t)에 걸쳐 나타낼 때 사건 발생률에 대한 모형을 설정

- 로지스틱 모형 vs 로그선형 모형

	로지스틱회귀 모형	로그선형 모형
사용 변수	Y: 범주형 X: 혼합형	모두 범주형
설명변수와 반응변수 구분	O	X
목적	결과 예측(분류)	변수 간 연계성 확인
	X, Z변수에 따른 Y변수의 확률은 얼마일까?	X, Y, Z변수들 간의 연관성이 있을까?

독립 로그 선형 모형

: XY가 독립이라는 조건 하의 모형

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

상수항 모수

행 효과 (주효과)

열 효과 (주효과)

- λ_{ij}^{XY} = 0 : 교호작용항이 없다!

→ XY는 연관이 없다!

→ 조건부오즈비=1로, 조건부 독립이다!

이 모형.. 처음보지만 왠지 익숙해.. 너 누구야..

포화 로그 선형 모형

: 변수들이 독립이 아닐 경우 사용

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

교호작용항

- 연관성 모수 λ_{ij}^{XY} 포함 : 독립이 아니다!

  → λ_{ij}^{XY}=0 이면 독립성 만족

- 모든 모수 고려 : 완벽한 적합, 해석이 어렵다

해석 NON-이지.. 사용 잘 하지 않는다!

로그선형 모형 해석 꿀팁!

λ_{ij}^{XY}와 같은 교호작용항을 중점으로 보자!

조건부 독립성 : 두 변수에 대한 교호작용항이 나타나지 않은 경우!

XY에 대한 교호작용항 λ_{ij}^{XY} 이 없다? 즉, 0이다?

→ Z 변수 통제 시, XY는 조건부 독립이다!

동질 연관성 : 두 변수에 대한 교호작용항이 나타나 있는 경우!

XY에 대한 교호작용항 λ_{ij}^{XY} 이 있다?

→ Z 변수 통제 시, XY 간의 동질연관성이 나타난다!

완전 독립 모형 : 3차원 (X,Y,Z)

$$\log(\mu_{ijk}) = \lambda + \underbrace{\lambda_i^X + \lambda_j^Y + \lambda_k^Z}_{\text{주효과}}$$

- 주효과만 고려 ($\lambda_i^X, \lambda_j^Y, \lambda_k^Z$), 교호작용항은 없다!
- 모든 변수 간 상호 독립

  조건부 독립 만족 → 모든 조건부 오즈비=1
- 대부분의 자료에 적합X

  현실에 거의 없기때문..

한 변수 독립 모형

(XY, Z) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$

(XZ, Y) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$

(YZ, X) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$

- 두 변수 간 교호작용 1개씩 존재: λ_{ij}^{XY}, λ_{ik}^{XZ}, λ_{jk}^{YZ}
- 교호작용 나타나지 않는 변수 → 조건부 독립
- (XY, Z)의 경우: Z변수 통제 시, XY의 관계는 동질연관성!

  → 교호작용항 λ_{ij}^{XY}만 나타나있다!

  → 나머지 YZ와 XZ의 관계는 조건부 독립!

조건부 독립 모형

(XY,XZ) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

(XY,XZ) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

(XY,XZ) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

- 두 변수 간 교호작용 2개씩 존재
- (XY,XZ)의 경우: X변수 통제 시, YZ의 관계는 조건부 독립

→ 교호작용항 λ_{jk}^{YZ}만 나타나지 않았다!

→ YZ에 대한 조건부 오즈비 = 1

→ XY, XZ의 관계는? 교호작용항 있으니까 동질연관성!

동질 연관성 모형 : 3차원 (XY,XZ,YZ)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- 두 변수 간 교호작용항 3개 모두 포함
- 모든 두 변수끼리 동질연관성 만족

→ 조건부 독립성은 나타나지 않는다!

→ 모든 교호작용항이 다 나타나 있으므로!

포화 모형 : 3차원 (XYZ)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$$

3차 교호작용항 추가

- 모든 모수 사용하므로, 데이터 완벽 적합
- 3차 교호작용까지 고려하여 해석이 너-무 복잡

  → 잘 사용하지 않는다!

복잡한 건 싫어..! 잘가고..

- 적합성 검정 및 모형 비교

"적합성 검정": 카이제곱 적합검정법 사용

- 검정통계량: X^2, G^2
- 칸잔차(표준화 잔차) 사용

  → 모형 적합성을 각 칸에 대하여 더 자세히 확인 가능!

"모형 비교": 유의성 검정 사용

- 이탈도 차를 사용하는 가능도비 검정
- 부분연관성 검정 가능!

"차이 지수" : 실제적 유의성 검증 위해 사용

$$D = \sum \frac{|n_i - \hat{\mu}_i|}{2n} = \sum \frac{|p_i - \hat{\pi}_i|}{2}$$

- 표본자료값(n_i, p_i)과 모형적합값($\hat{\mu}_i, \hat{\pi}_i$)이 서로 얼마나 가까운 지 요약

- 범위: 0~1
  → 작을수록 실제적으로 유의한 것!
  → 모형의 적합 결여에 대한 결과가 실제적으로 중요한 의미 갖는지 판단!

• 로지스틱 모형과의 관련성

로그 선형모형		로지스틱 모형	로지스틱 기호
(X, Y, Z)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	X	X
(Y, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	α	(-)
(X, YZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	X	X
(Z, XY)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	X	X
(XY, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	$\alpha + \beta_i^X$	(X)
(YZ, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	$\alpha + \beta_k^Z$	(Z)
(XY, YZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	X	X
(XY, YZ, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	$\alpha + \beta_i^X + \beta_k^Z$	(X+Z)
(XYZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$	$\alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$	(X*Z)

** 붉은색 음영: 로그선형모형 = 로지스틱모형일 경우*

로그선형 모형 중에서도 설명변수 간의 교호작용λ_{ik}^{XZ}이 포함되어 있는 경우만 로지스틱 모형과 동치관계

로지스틱은 아까도 보았듯이.. Y vs. X와 Z 간의 관계를 궁금해한다!

"독립성 그래프"

역할

로그선형 모형을 시각적으로 나타내 줌으로써,

모형에 내재되어있는 관계를 밝히는 데 도움을 준다!