범주형자료분석팀

2팀

김찬영
이혜인
김서윤
심은주
진수정

INDEX

0

지난 주 리뷰

분할표

: 범주형 변수의 결과의 도수들을 각 칸에 넣어 표로 정리한 것

- 2차원 분할표 (I*J)

  :두 개의 변수만을 분류한 분할표

	Y			합계
X	n_{11}	…	n_{1j}	n_{1+}
	…	…	…	…
	n_{i1}	…	n_{ij}	n_{i+}
합계	n_{+1}	…	n_{+j}	n

설명 변수 : X
반응 변수 : Y

0 지난 주 리뷰

- 2차원 분할표 독립성 검정

대표본
- 명목형
  - 피어슨 카이제곱 검정
  - 가능도비 검정
- 순서형
  - 멘탈 스코어 검정

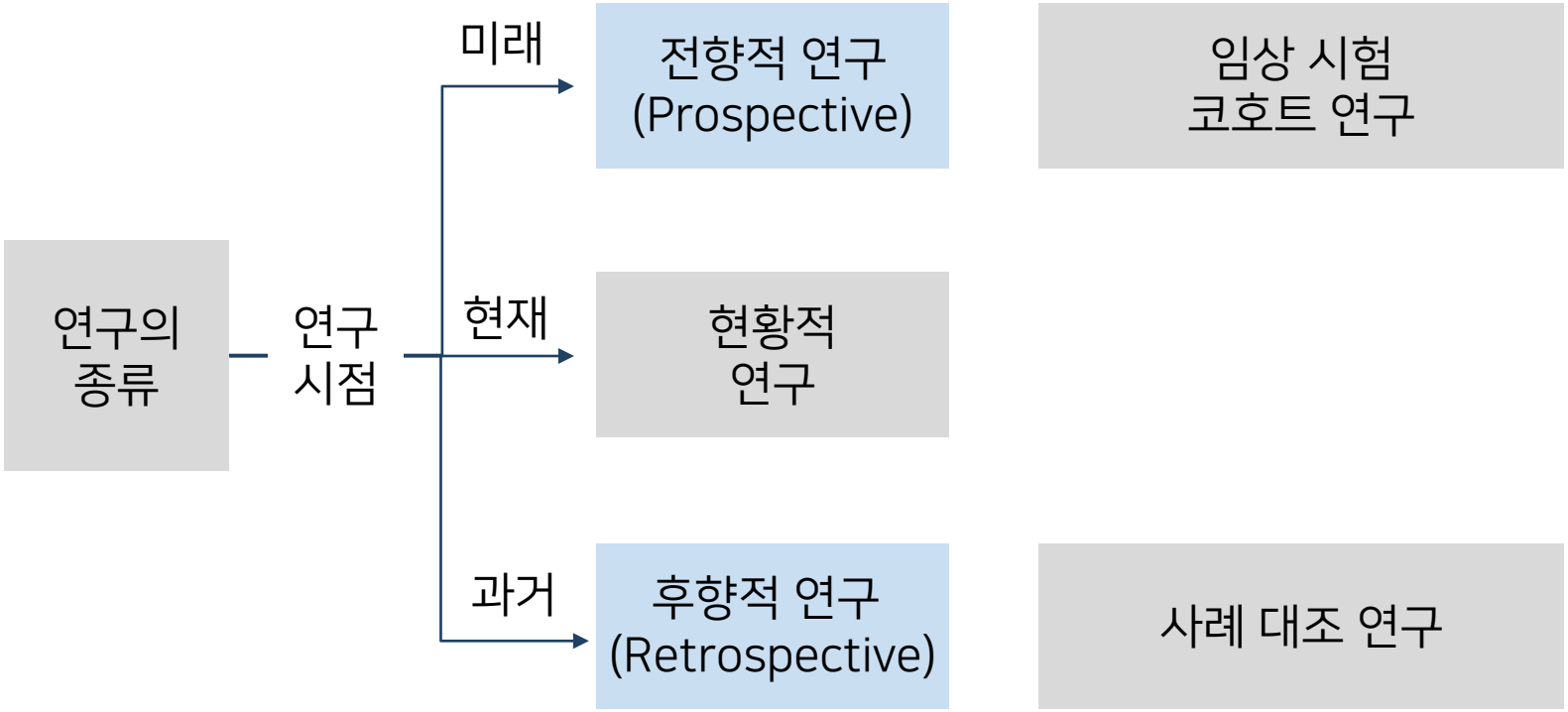
소표본
- 피셔의 정확 검정

- 3차원 분할표 독립성 검정

로그 선형 모형 비교

3차원 이상의 고차원 모형은
모형으로 다루는 것이 효과적!

0 지난 주 리뷰


연구의 종류
연구 시점
미래
현재
과거
전향적 연구 (Prospective)
현황적 연구
후향적 연구 (Retrospective)
임상 시험 코호트 연구
사례 대조 연구

"오즈비" : 여러 모형에서 기초가 되는 모수

각 행의 오즈끼리의 비 : Odds ratio(θ) = $\frac{_1\text{행의 오즈}}{_2\text{행의 오즈}} = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)}$

성별	연인 여부		Odds
	예	아니오	
여성	0.8144	0.1856	4.3879
남성	0.7928	0.2072	3.8262

→ $\frac{4.3879}{3.8262} = 1.1468$

"여성이 연인이 있을 오즈가 남성이 연인이 있을 오즈보다 약 1.15배 높다"

1

GLM

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

GLM 일반화 선형모형

익숙한 보통의 선형 회귀모형

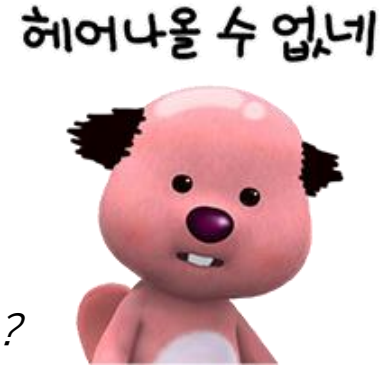
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 랜덤성분이 **정규분포가 아닌 다른 분포**를 갖도록 일반화
- 랜덤성분에 대한 함수인 **연결함수를 모형화**하여 일반화

- 필요성

1. 범주형 자료는 보통의 선형회귀모형의 사용 X
   - 오차항의 확률분포가 정규분포가 아니기 때문!

2. 정규분포에 근사하도록 만드는 작업 필요 X
   - GLM: ML 방법을 사용 → LSE와 같은 정규성 조건이 필요 없다!



GLM은 가정이 널-널하구나.. 아주 매력쟁이네..?

- 분할표 분석과 비교했을 때

분할표 분석

변수 간의 효과 파악

범주형 자료만 표현 가능

VS

GLM

반응 변수에 대한 예측 가능
&
변수 간의 연관성 파악 가능

설명변수에 연속형 변수
사용 가능

- 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"

- 구성 성분

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

① 선형예측식 (체계적 성분)	② 랜덤 성분	③ 연결 함수
$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$	$\mu(= E(Y))$	$g(\,)$

3개의 구성성분

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

- **체계적 성분 (선형예측식)**

  : 설명변수 X들의 선형결합

  - $x_i = x_a x_b$: x_a와 x_b의 교호작용을 설명할 수 있음
  - $x_i = x_a^2$: x_a의 곡선효과를 나타낼 수 있음

- **랜덤 성분**

  : 반응변수 Y에 대한 확률 분포 가정 → **Y 분포의 기대값인 평균 $\mu = E(Y)$ 사용**

  - Y 끼리 독립성은 만족해야 함

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

- 연결함수

  : 랜덤성분과 체계적 성분을 연결

반응변수가 범주형 변수일 경우,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i : 연속형이 아닌 값

$\beta_1 x_i$: $(-\infty, \infty)$

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

- 연결함수

  : 랜덤성분과 체계적 성분을 연결

반응변수가 범주형 변수일 경우, 연결함수$g(\,)$ 이용하면

$$g(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$(-\infty, \infty)$ $(-\infty, \infty)$

범위가 일-치!

- 연결함수의 종류

항등 연결 함수

: $g(\mu) = \mu$

- 연속형 반응변수
- 이항 반응변수

→ 구조적 결함

로그 연결 함수

: $g(\mu) = log(\mu)$

- 도수자료
- 포아송 반응변수
- 음이항 반응변수

로짓 연결 함수

: $g(\mu) = log[\frac{\mu}{1-\mu}]$

- 오즈의 로그 값을 모형화
- 이항 반응변수

- GLM의 종류

GLM	랜덤요소	연결함수	체계적 성분
선형 확률 분포	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 종류

범주팀장 모형 버려..? 아-니요!
다른 모형이 궁금하시면 교안과 친절한 안내로 돌려드립니다..

GLM	랜덤요소	연결함수	체계적 성분
선형 확률 분포	이항 자료	항등	혼합형
로지스틱 회귀 모형		로짓	
프로빗 회귀 모형			
기준범주 로짓 모형	다항 자료		
누적 로짓 모형			
이웃범주 로짓 모형			
연속비 로짓 모형			
로그 선형 모형	도수 자료	로그	범주형
포아송 회귀 모형			
음이항 회귀 모형			혼합형
카우시 모형			
율자료 포아송 회귀 모형	비율 자료		

T.M.G. (Too Much GLM)…
이므로 선택과 집중을 해보자..!

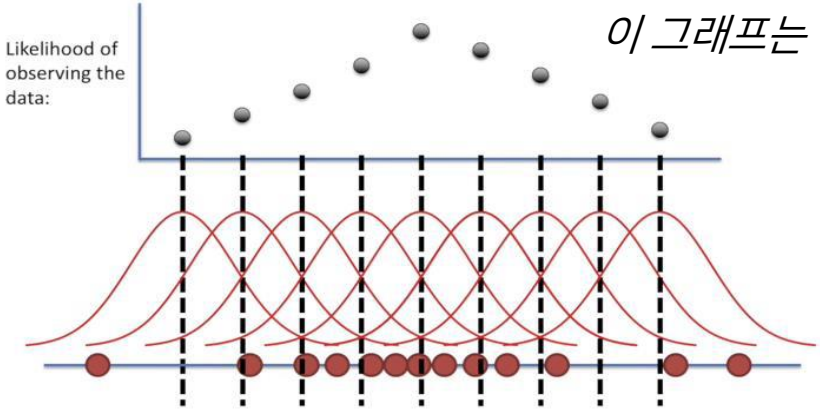
- 모형 적합

  : 샘플로 주어진 데이터를 근거로 모형의 모수를 추정하는 것

"MLE" 최대가능도추정법

↓

관측값이 고정됐을 때,그 관측값이 어떤 확률분포를 따를 가능성



이 그래프는 언제 봐도 ㄹㅈㄷ..

가능도에 대한 구체적인 설명은 지난 학기 근본범주팀 클린업에 쏘-클린하게 정리되어 있다는 소문이..

"MLE" 최대가능도추정법

: GLM은 수치적 근사로 계산

- 피셔 점수화 알고리즘
- 뉴턴-랩손 알고리즘 : 컴퓨터를 이용한 수치적 근사 *→ 손으로 풀고 싶다면*

소년들이잘못해서 가는곳은 소년원

대학생이잘못해서 가는곳은 대학원

*** LSE** (최소제곱 추정법) **는 왜 안될까?**

→ 오차에 대한 **정규분포 가정**이 없기 때문!

2

유의성 검정

"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

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

가능도비 검정의 FLOW

검정통계량 값이 크네? → **P-값이 작겠네?** → **귀무가설 기각이겠네?**

적어도 하나의 β는 0이 아니겠군! → **모형의 모수 추정값은 유의하구나!**

지난 주 FLOW와 똑같다! 다들 기억하시나요...? 네네네! ☆

"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

- 장점

두 가지 경우의 로그 가능도 함수에 대한 정보 사용

→ 가장 많은 양의 정보 사용!

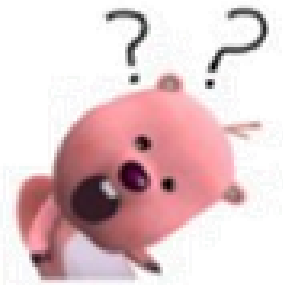
"가능도비 검정"

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

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

H_1: 적어도 하나의 β는 0이 아니다

*** 이탈도를 통해 구할 수 있음!**



이탈도가 뭔데요 어?!

- 관심모형 & 포화모형

관심 모형 (M)

유의성 검정을 진행할 모형

L_M : 모형 M에서 얻은
로그 가능도 함수의 최대값

포화 모형 (S)

관측값을 완벽하게 적합하는 모형

L_S : 모형 S에서 얻은
로그 가능도 함수의 최대값

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- 관심모형 & 포화모형

관심 모형 (M)

유의성 검정을 진행할 모형

L_M : 모형 M에서 얻은
로그 가능도 함수의 최대값

포화 모형 (S)

관측값을 완벽하게 적합하는 모형

L_S : 모형 S에서 얻은
로그 가능도 함수의 최대값

범주팀복지 (Y)

$= \beta_0 + \beta_1$ 스터디시간 (x_1) $+ \beta_2$ 교안 페이지 수 (x_2)

범주팀복지 (Y)

$= \beta_0 + \beta_1$ 스터디시간 (x_1) $+ \beta_2$ 교안 페이지 수 (x_2)

$+ \beta_3$ 스터디 시간& 교안페이지 수 $(x_1 x_2)$

- 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 때~~만 ~~사용 가능!~~
  - 모형 적합도 검정에도 사용
  - 근사적으로 카이제곱분포 따름

- 가능도비 검정과의 관계

M_0의 이탈도 $- M_1$의 이탈도

$= -2(L_0 - L_S) - [-2(L_1 - L_S)] = -2(L_0 - L_1)$

$=$ 가능도비 검정통계량

→ 모형간의 이탈도의 차

- 이탈도를 이용한 모형비교

  - M_0은 M_1의 nested 모델이어야 함(→ M_0의 계수 $\subset$ M_1의 계수)

    → 아니라면 AIC와 같은 값들을 이용해 모형 비교

  - M_0가 M_1에 비해 적합이 잘 되지 않는다면,

    → 두 이탈도의 차이가 커져 검정통계량이 커짐

(AIC는 권교수님의 전문 분야..)

FLOW 다시 정리해보자!

이탈도 차이가 작네?

그럼 검정통계량 값도 작네!

P-값은 크네?

관심모형에 없는 계수들은 모두 0이군!

이 관심모형 잘 적합 하는구나!

이 FLOW 하나면 어떤 검정이든 해석이 든-든하군요..! 검정계의 뜨끈한 국밥 FLOW..

3

로지스틱 회귀모형

로지스틱 회귀모형

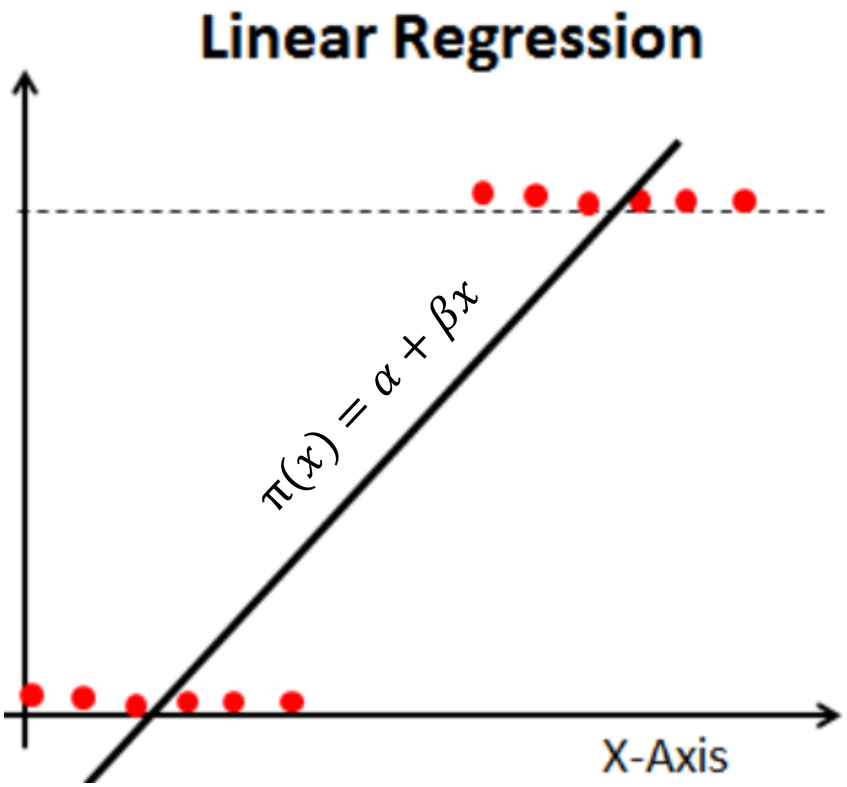
: 반응변수 Y가 이항반응변수일 때 사용

예시

학점에 따른 합격 여부를 확인하는 문제

- $Y \sim Bernoulli(\pi)$ 이항분포 사용

일반 선형 모델($\pi(x) = \alpha + \beta x$)로
적합한다면?



구조적 문제점

1. 범위가 일치하지 않음

$$\pi(x) < 0 \quad \text{or} \quad \pi(x) > 1$$

2. 분산이 일정하지 않음

$$Var(Y) = \pi(x)(1 - \pi(x))$$

➡ 등분산성 조건에 위배됨

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

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

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

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

Logistic Model

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

- 가정으로부터 자-유로움

~~정규성~~ ~~등분산성~~ ~~선형성~~ 독립성

Logistic Model

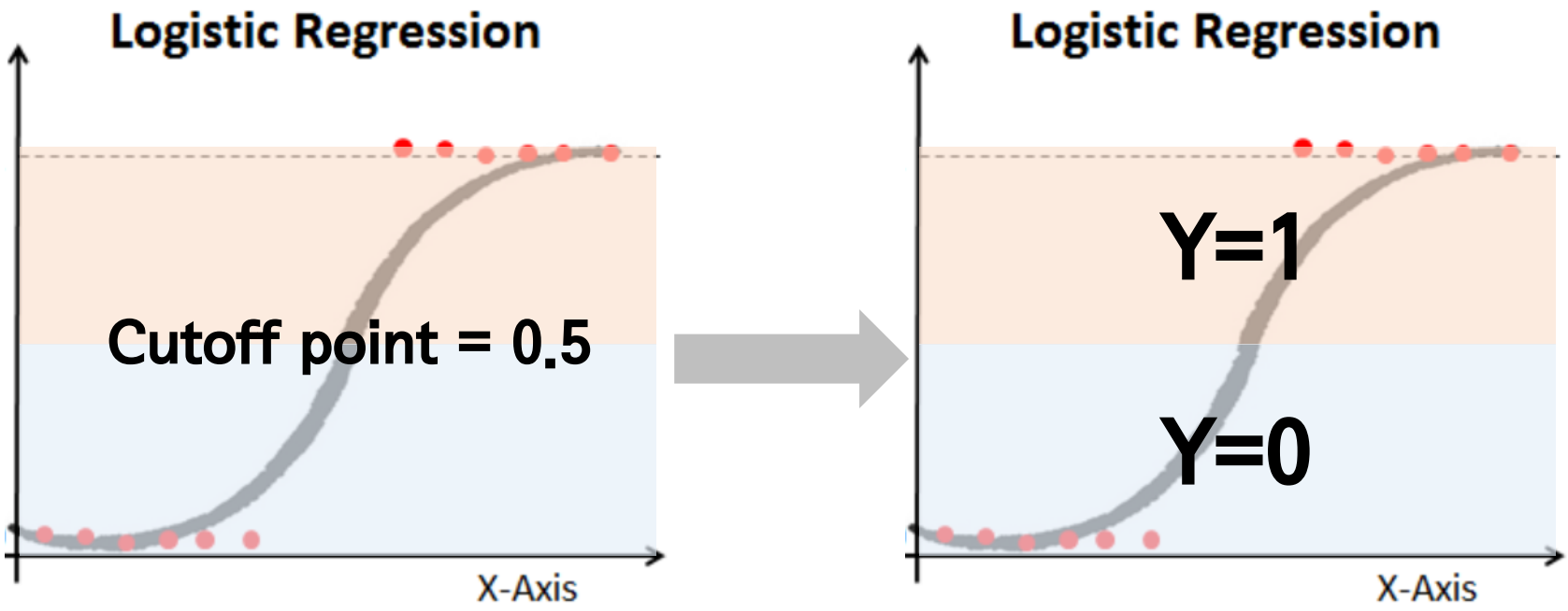
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta \quad \Leftrightarrow \quad \pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

- 후향적 연구에도 사용 가능

  모수들이 **오즈, 오즈비**와 관련되어 있기 때문에 **후향적 연구에 사용 가능**

  (프로빗 모형 등의 다른 모형은 사용 불-가)

위대한 오-즈비 하고 싶은 거 다 해..



Cut-off point를 중심으로 0 또는1로 예측

- 해석

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

- β의 해석

  곡선의 증가비율 혹은 감소비율을 결정

  $\beta > 0$: 곡선이 상향, $\beta < 0$: 곡선이 하향

  $|\beta|$가 증가함에 따라 변화율이 증가

- 해석

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \iff \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\log\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}\right) = \beta \Rightarrow \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))} = e^{\beta}$$

- 오즈비 해석

  X가 한 단위 증가함에 따라 오즈는 e^{β}배 증가

예시를 통해서 알아보자!

해석

학점에 따른 합격 여부 (x: 학점, π: 합격할 확률)

$$\log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \Leftrightarrow \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 6 + 2.5x$$

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

- 오즈비 해석

  X가 한 단위 증가함에 따라 오즈는 e^{β}배 증가

학점이 한 단위 증가할 때마다

합격할 오즈는 $e^{2.5} \approx 12.18$배 증가한다! 이-지!

- *좋은 모델이란…? 갑자기 분위기 데마감성…*

예측의 정확도

해석의 용이성

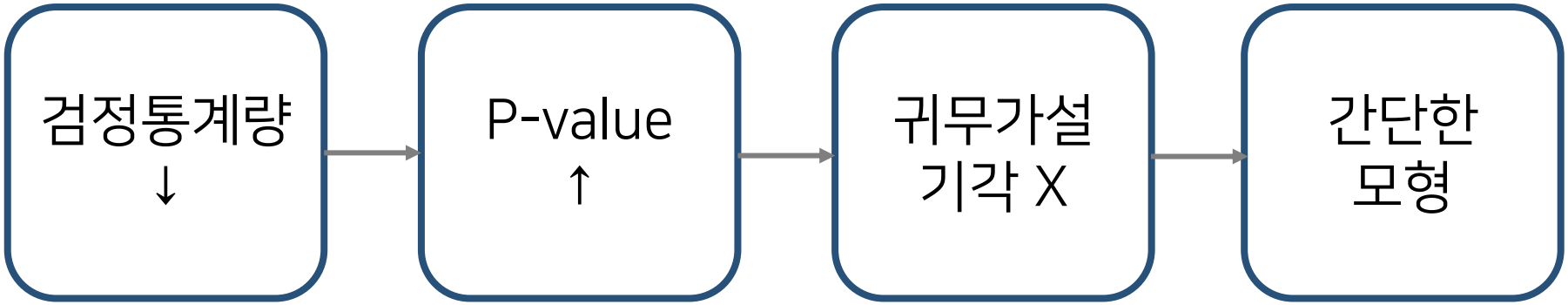
두 가지 조건은 일반적으로 상충되지만 둘 다 이루어져야 좋은 모델!

이탈도의 차이를 이용하여 모형들을 비교해서 최적의 모델을 찾아보자!

"모형 비교 Flow"

귀무가설: 간단한 모형 사용하자!

대립가설: 복잡한 모형 사용하자!



이탈도의 차를 활용한 모형 비교감성 그-대로 가져가면 되네! 범주.. 쉽네?ㅎ

Case1. 순서형 예측변수의 양적변수 취급

예시

격리 수준과 지역에 따른 특정 질병의 발병률은?

	incidence	area	isolation
1	1	3	3.317
2	0	1	7.554
3	1	1	5.883
4	0	3	5.932
5	0	1	5.308
6	1	3	4.934

양적변수취급 X

	area2	area3
Obs1	0	1
Obs2	0	0
Obs3	0	0

양적변수취급 O

	area
Obs1	3
Obs2	1
Obs3	1

Case1. 순서형 예측변수의 양적변수 취급

예시

격리 수준과 지역에 따른 특정 질병의 발병률은?

```
Likelihood ratio test

Model 1: incidence ~ isolation + area
Model 2: incidence ~ isolation + area
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -16.579
2   4 -15.690  1 1.7787     0.1823
```

양적변수취급 X

양적변수취급 O

P-value가 크네?

양적변수 취급하자!

Case 2. 교호작용항 포함 vs. 미포함

예시

격리 수준과 지역에 따른 특정 질병의 발병률은?

교호작용 X $\log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 \text{격리수준} + \beta_2 \text{지역}$

교호작용 O $\log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 \text{격리수준} + \beta_2 \text{지역} + \beta_3 x_1 x_2 \text{격리수준*지역}$

Case 2. 교호작용항 포함 vs. 미포함

예시

격리 수준과 지역에 따른 특정 질병의 발병률은?

```
Likelihood ratio test

Model 1: incidence ~ isolation + area
Model 2: incidence ~ isolation * area
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -15.690
2   6 -15.364  2 0.6522     0.7217
```

P-value가 크네?

단순한 게 최고다…

교호작용항 빼자!

교호작용 모형 비교 시 유-의할 점!

Case 2. 교호작용 포함

- 교호작용이 없는 모형은 해석이 더 용이

귀무가설 $\beta_3 = 0 \quad \Rightarrow \quad \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2$

- 주효과가 포함되지 않고 교호작용만을 가지고 적합하는 것은 No!

대립가설 $\beta_3 \neq 0 \quad \Rightarrow \quad \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

- 교호작용이 없는 경우, 조건부 독립성을 만족한다!

예시

학점과 성별에 따른 연인 여부를 확인하는 문제

아아… 이것은 '로그 선형 모형' 파트에서 다룰 것이다 범주 쉽..다 했지..?

4

포아송 회귀모형

포아송 회귀모형

: 반응변수 Y가 도수자료일 때 사용

예시

교통사건 건수를 예측하는 문제

- Y~ Poisson(μ) 포아송 분포 사용

Poisson Regression : log를 link function으로 사용

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

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

과대산포 문제(Over dispersion)란?

: 예측된 분산보다 실제 데이터가 더 큰 분산을 가질 때 나타나는 문제

"Poisson Regression": log를 link function으로 사용

- 발생 원인

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

1. 포아송 회귀 모형은 평균과 분산이 같다는 등산포를 가정

   그러나 가정과는 달리 실제 데이터에서는 보통 분산이 평균보다 크기 때문

   - 평균과 분산이 같아야 함

2. 등산포 가정을 만족하도록 하는 변수를 모두 사용하지 않고,
   하나의 변수만 통제하면 과대산포가 발생

   - 예측된 것보다 실제 분산이 더 큰 과대산포 문제가 발생하기도 함

이게 개체의 이질성인데..무슨 말인지는
다음 장에서 예시를 통해 알아보자!

   - 도수자료인 반응변수를 [있다, 없다]와 같이 이항자료로 범주화하여 나타내면 로지스틱 회귀로 분석할 수 있음

학점

체지방률 → 주량

근육량

세 가지 변수들의 고정된 수준 조합에서
Y(주량) ~ Poisson 라고 해보자!

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

체지방률만 고려한다면…?

체지방률	표본 수	주량 총합	주량 표본평균	주량 표본분산
<15	17	87	5.12	6.27
15-25	57	200	3.92	15.32
>25	26	27	1.04	3.08

체지방률 만으로는 변동을 다 잡아내지 못하기 때문에
고정된 수준조합이 와장창..되는 개체들 간의 이질성 나타난다!

이로 인해 과대산포 발-생!

과대산포 문제(Over dispersion) 해결법

"Poisson Regression": log를 link function으로 사용

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

- 확인 방법

  과산포 검정으로 과대산포 문제가 있는지 확인

- 해결 방법
  - 평균과 분산이 같아야 함

  유의미한 설명변수 & 파생변수 추가

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

  이상치 제거

  올바른 연결함수 사용

  다른 모형 사용 (음이항 회귀 모형 등)

  - 종속변수(Y)의 분산이 더 큰 과대산포 문제가 발생
  - 도수자료인 반응변수를 [있다, 없다]와 같이 이항자료로 범주
  - 지스틱 회귀로 분석할 수 있음

음이항..? 뭔지 말해주고 가..!

음이항 회귀 모형

"Poisson Regression": log를 link function으로 사용

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

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

음이항 회귀도 마찬가지로, 로그연결함수를 이용!

- 평균과 분산이 같아야 함

분산이 평균보다 더 큰 값을 갖도록 하는 모수 (산포모수 D) 를 하나 더 가짐

- 예측된 것보다 실제 분산이 더 큰 과대산포 문제가 발생하기도 함

$$\mathrm{E}(\mathrm{Y}) = \mu\,, Var(Y) = \mu + D\mu^2$$

- 도수자료인 반응변수를 [있다, 없다]와 같이 이항자료로 범주화하여 나타내면 로지스틱 회귀로 분석할 수 있음

산포모수 D=0 이면 포아송 회귀 모형과 동일

Poisson Regression : log를 link function으로 사용

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

해석

$$\log(\mu) = \alpha + \beta x$$

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

X가 한 단위 증가함에 따라 기대도수 μ는 e^{β}배 증가

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

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

예시

연령별 취업률을 알아보는 문제

여기서는 지표 t 가 연령이 되겠쥬?

R에서, 율자료는 formula인자에 Y ~ X + offset(log(t)) 형태로 나타내 주면 된답니다!

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

해석

X가 한 단위 증가하면 기대비율 $\frac{\mu}{t}$가 e^{β}배 증가

해석 방법도 다 비-슷비슷.. 여기선 기대비율이란 것만 알면 이-지이지

"Summary"

로지스틱 회귀모형

반응변수	이항분포
연결함수	로짓연결함수
형태	$\log\frac{\pi_i}{1-\pi_i} = x_i^T\beta$

포아송 회귀모형

반응변수	포아송분포
연결함수	로그연결함수
형태	$\log\mu = x_i^T\beta$

5

로그 선형 모형

로그 선형 모형

1. 범주형 자료들이 각 칸도수에 어떤 영향을 미치는 지 확인할 수 있음

범주형 변수만 고려 가능!

	Y			합계
X	n_{11}	…	n_{1j}	n_{1+}
	…	…	…	…
	n_{i1}	…	n_{ij}	n_{i+}
합계	n_{+1}	…	n_{+j}	n

범주형 변수들 사이 연관성 및 교호작용 분석 모형

로그 선형 모형

1. 범주형 자료들이 각 칸도수에 어떤 영향을 미치는 지 확인할 수 있음

2. 설명변수와 반응변수 구분이 없음

	Y			합계
X	n_{11}	…	n_{1j}	n_{1+}
	…	…	…	…
	n_{i1}	…	n_{ij}	n_{i+}
합계	n_{+1}	…	n_{+j}	n

X와 Y의 상관관계가 아니다

- 반응변수 Y가 여러 개인 다변량 분석에도 활용 가능

로그 선형 모형

1. 범주형 자료들이 각 칸도수에 어떤 영향을 미치는 지 확인할 수 있음
2. 설명변수와 반응변수 구분이 없음
3. 포아송 분포를 사용함

- 기대도수 " $\mu_{ij} = n\pi_{ij}$" 사용 → 분할표의 칸도수를 설명하는 것
- 분할표 내에서 포아송 회귀 모형 사용 가능
  → 칸도수는 도수자료이므로!

로그 선형 모형 의 목표

내포 모형 *VS* 포화모형

분할표를 가장 잘 설명하는 모형을 찾는 것

모형 1등 피뗏으로 따라와..

- 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?

예시를 통해 자세히 알아보자!

- 로지스틱 모형 vs 로그선형 모형

		로지스틱 회귀 모형	로그 선형 모형
사용 변수	X	밤샘 술파티 여부	밤샘 술파티 여부
	Y	숙취 여부	숙취 여부
	Z	나이 (연속형)	나이대 (초반, 중반, 후반) *범주형만 사용하기에 범주화
목적		X, Z 변수들에 따른 Y의 확률? (Y vs. X, Z)	분할표 내의 X,Y,Z 변수의 연관성은? (X vs. Y vs. Z)
결과		22살이고 밤새 술을 마신 내가 다음 날 숙취가 있을 확률은 “%”겠구나!	나이대, 술파티 여부, 숙취 여부는 연관성이 이렇게 되는구나!

- 모형 생성 과정(2차원 분할표)

XY가 서로 독립일 때, $\mu_{ij} = n\pi_{i+}\pi_{+j}$으로 표현 [승법적 관계식]

$$\log(\mu_{ij}) = \log(n)+\log(\pi_{i+})+\log(\pi_{+j})$$ 양변에 log()

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$ (X, Y는 단순 변수 형태) [가법적 관계식]

독립 로그 선형 모형

: XY가 독립이라는 조건 하의 모형

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

- λ : 상수항 모수
- λ_i^X : 행 효과 (주효과)
- λ_i^Y : 열 효과 (주효과)

- $\lambda_{ij}^{XY} = 0$: 교호작용항이 없다!

  → XY는 연관이 없다!

  → 조건부오즈비=1로, 조건부 독립이다!

이 모형.. 처음보지만 왠지 익숙해.. 너 누구야..

- 2차원 분할표 독립성 검정 vs 모형 적합성 검정

	2차원 분할표	로그선형모형
	독립성 검정	독립성 모형의 적합성 검정
H_0	두 변수가 독립이다	잘 적합(교호작용X, 독립)
H_1	두 변수가 독립이 아니다	잘 적합 안됨(교호작용 필요, 연관성)
의의	분할표 내의 독립성 검정과 모형의 적합성 검정의 귀무가설이 동일	

너였구나.. H_0: $\mu_{ij} = n\pi_{i+}\pi_{+j}$ 저번 주 독립성 검정 귀무가설 기억 새록새록이쥬?

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!

3차원 로그 선형 모형

: 기존 검정방법으로는 각 변수들 간 관계 파악 불가

Breslow-Day test,
Cochran-Mantel-Haenszel test..

		Y		합계
Z	X	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{++2}

모형 간의 비교 필요

↓

9가지 모형 존재

누가 그렇게 많으래…?

- 완전 독립 모형

1. $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$

- 한 변수 독립 모형

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

- 조건부 독립 모형

5. $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{YZ}$
6. $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{YZ}$
7. $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{YZ}$

- 동질 연관성 모형

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

- 포화 모형

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

우선 한 번 쫙 봐주시고..

- 3차원 분할표에서의 로그선형모형

<3차원 분할표>

Z	X	Y		조건부 오즈비
		Y_1	Y_2	
Z_1 제어	X_1	π_{111}	π_{121}	θ_1
	X_2	π_{211}	π_{221}	
Z_1	X_1	π_{112}	π_{122}	θ_2
	X_2	π_{212}	π_{222}	
Z_1	X_1	π_{113}	π_{123}	θ_3
	X_2	π_{213}	π_{223}	

동질연관성

$$\theta_{XY(1)} = \theta_{XY(2)} = \theta_{XY(3)}$$

조건부독립

$$\theta_{XY(1)} = \theta_{XY(2)} = \theta_{XY(3)} = 1$$

본격적으로 들어가기 전에, 지난 주 내용을 복습해보자… 기억나는가?

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)

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

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!

모형비교 이후, 이것까지 확인해보자!

아직 한 발 남았다..엉?

통계적 유의성 VS. 실제적 유의성

표본에 크기에 따라…

통계적으로 유의O → 실제로는 유의 X
통계적으로 유의X → 실제로는 유의 O

차이지수를 이용하여 검정

"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

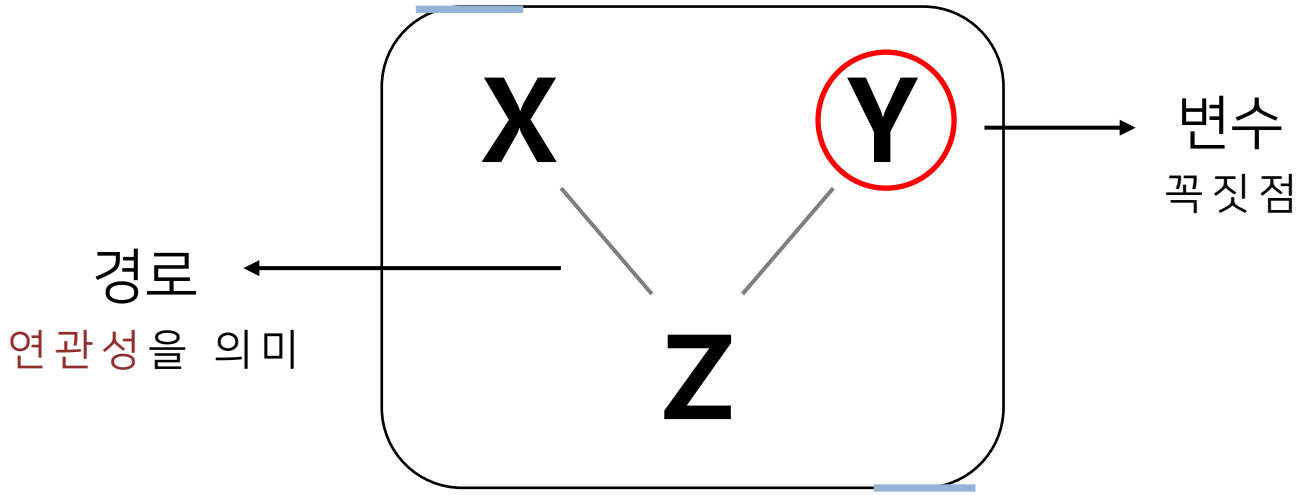
로지스틱은 아까도 보았듯이.. Y vs. X와 Z 간의 관계를 궁금해한다!

"독립성 그래프"

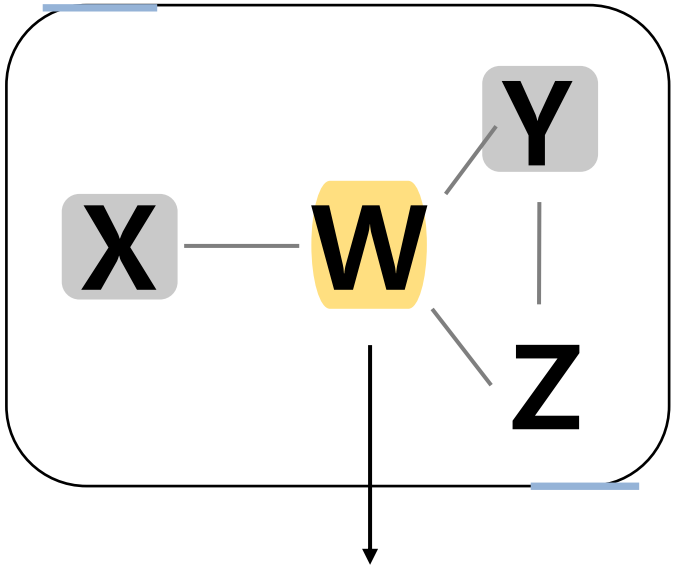
역할

로그선형 모형을 시각적으로 나타내 줌으로써,

모형에 내재되어있는 관계를 밝히는 데 도움을 준다!



- 분리

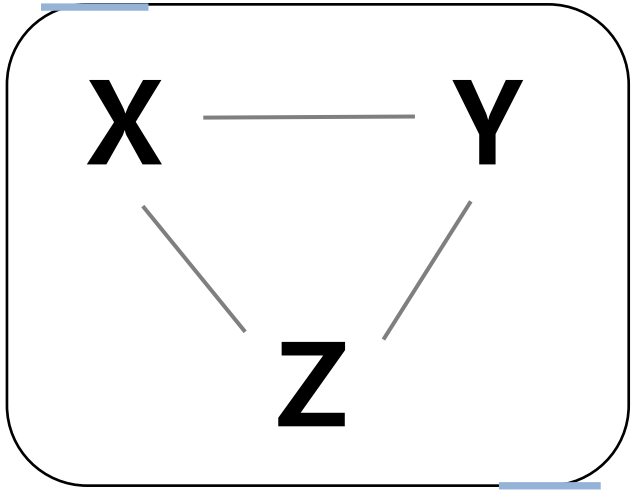


분리시키는 변수

W에 의해 X와 Y 분리

분리시키는 변수(W)가 통제될 때,
분리된 변수들(X, Y)은 서로 조건부 독립이다!

- 특징

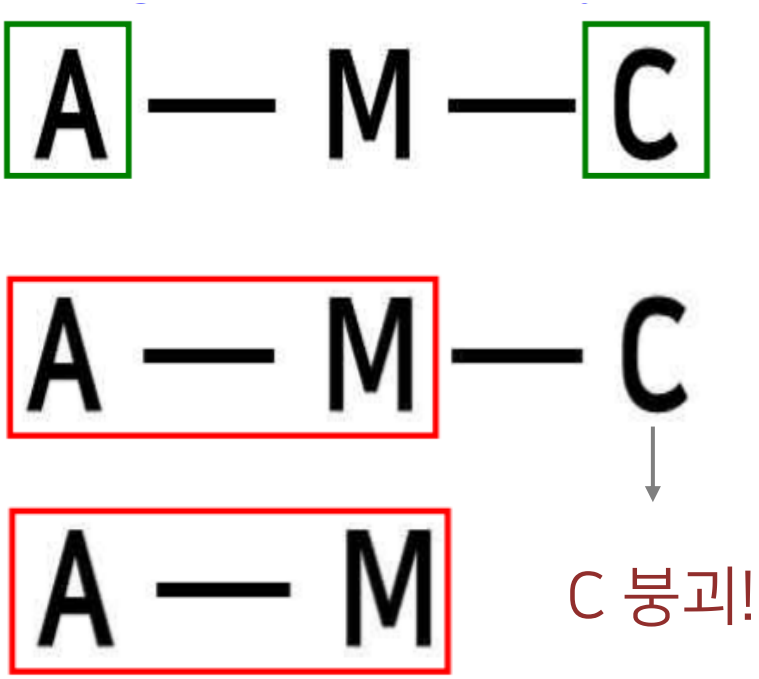


독립성 그래프로
모든 로그 선형모형을 나타낼 수는 없다!

(XY, YZ, XZ)? or (X,Y,Z)?

➡ 독립성 그래프와 로그 선형모형이 모두 1:1 대응되는 것은 아니다..

- 붕괴가능성 조건



A와 C가 조건부 독립
(M으로 분리)

C를 수준 별로 보나,
붕괴하여 합쳐 보나,
A-M의 연관성은 달라지지 X

즉, AM은 **붕괴가능성 조건 만족!**

그래서.. 붕괴가능성 조건이 성립되면 어쨌다는 건지..

- 붕괴가능성 조건

붕괴가능성 조건 성립

부분분할표 오즈비 값 = 주변분할표 오즈비 값

조건부 연관성 = 주변 연관성

6

부록

"가변수(Dummy Variables)"

: 설명변수가 범주형 변수일 때, 이를 연속형 변수 감성으로 바꿔준 것

- 연속형 변수만을 설명변수로 받는 분석기법도 사용 가능하게 됨
- J개의 수준을 갖는 인자를 표현하기 위해서는 J-1개의 더미변수 필요!

	신호등
Obs1	빨간색
Obs2	노란색
Obs3	초록색

	빨간색	노란색
Obs1	1	0
Obs2	0	1
obs3	0	0

"가변수(Dummy Variables)"

: 설명변수가 범주형 변수일 때, 이를 연속형 변수 감성으로 바꿔준 것

	신호등
Obs1	빨간색
Obs2	노란색
Obs3	초록색

	빨간색	노란색
Obs1	1	0
Obs2	0	1
obs3	0	0

더미변수는 전체를 하나로 취급해야 함

더미변수 중 일부만 유의미할 때에는 더미변수를 모두 없애거나 재범주화

더미변수를 만드는 원-핫 인코딩 방법은 다음주에 자세히…

- 로지스틱 모형과의 관련성

(Y,XZ)

$$\begin{aligned}
\text{logit}[P(Y=1)] &= \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \log\left[\frac{P(Y=1|X=i,Z=k)}{P(Y=2|X=i,Z=k)}\right] \\
&= \log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) = \log(\mu_{i1k}) - \log(\mu_{i2k}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\
&\quad - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}) \\
&\quad \| \qquad\qquad \| \qquad\qquad \| \\
&\quad \alpha \qquad\quad \beta_i^X \qquad\quad \beta_k^Z
\end{aligned}$$

"선형-대-선형 모형"

필요성

순서형 변수를 분석하기 위한 방법 필요!

→ 여태까지 배운 모형들은 모든 변수를 명목형으로 간주

설명

순서형 로그 선형 모형: 순서연관성을 모형화한 것

독립성 모형보다 복잡, 포화모형보다 단순

각 행에 크기 순으로 점수 할당

"선형-대-선형 연관성 모형 "

$$\log(\mu_{ij}) = \lambda_i^X + \lambda_j^Y + \boldsymbol{\beta\mu_i v_j}$$

$\boldsymbol{\beta\mu_i v_j}$: 독립성으로부터의 편차

$\boldsymbol{\beta}$: 연관성 방향과 강도

➡ $\boldsymbol{\beta = 0}$: 독립성 모형(연관성X)

독립성 모형과 선형-대-선형 연관성 모형의 G^2(이탈도)의 차이로 모형 선택!

3주차 예고

1. Confusion Matrix
2. ROC 곡선
3. Unbalanced Data
4. Encoding

THANK YOU