

10 강

데이터분석방법론2

분할표 및 도수자료에 대한 로그 선형모형

통계·데이터과학과 이기재 교수



학습목차

1 제 10강. 분할표 및 도수자료에 대한 로그 선형모형

- 1 분할표 자료에 대한 로그 선형모형
- 2 로그 선형모형의 추론
- 3 로그 선형모형과 로지스틱 회귀모형의 관련성
- 4 독립성 그래프와 붕괴 가능성



학습개요 및 목표

로지스틱회귀 모형은 반응변수와 설명변수로 구분할 수 있을 때 적용했습니다. 로그선형모형은 GLM의 일종이지만 모든 변수들을 반응변수로 간주하여 분석합니다. 로그선형모형을 통해 분할표 분석에서 단순히 독립성 여부만을 검정했던 것과 비교하여 변수들간의 연관성 관계를 구체적으로 살펴볼 수 있게 됩니다.

- 1 삼차원 분할표에 대한 로그 선형모형을 설명할 수 있다.
- 2 로그선형모형 추론법과 로지스틱회귀 모형과의 관련성을 설명할 수 있다.
- 3 독립성 그래프의 붕괴 가능성의 의미를 설명할 수 있다.



제 10강. 분할표 및 도수자료에 대한 로그 선형모형

- 1 분할표 자료에 대한 로그 선형모형
- 2 로그 선형모형의 추론
- 3 로그 선형모형과 로지스틱 회귀모형의 관련성
- 4 독립성 그래프와 붕괴 가능성

01

제 10강. 분할표 및 도수자료에 대한 로그 선형모형

분할표 자료에 대한 로그 선형모형

I. 분할표 자료에 대한 로그 선형모형

- 로지스틱회귀모형은 반응변수 Y 와 설명변수 x_1, x_2, \dots 을 구분할 수 있는 경우에 적용
- 로그 선형모형은 모든 변수들을 반응변수로 간주하여 분석함
- n 개의 관측값을 두 개의 범주형 변수에 따라 교차 분류하여 $I \times J$ 분할표를 작성한 경우
 - 행변수 X 와 열변수 Y 가 서로 독립
 $\Leftrightarrow \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J$
 - 칸 확률 $\{\pi_{ij}\}$ 는 다항분포(multinomial distribution)의 모수
 - $\{\pi_{ij}\}$ 대신 기대도수 $\{\mu_{ij} = n\pi_{ij}\}$ 를 사용 (포아송 추출모형)
 $\left[\rightarrow X \text{ 와 } Y \text{ 의 독립성 가정 } \Leftrightarrow \text{ 모든 } i, j \text{ 에 대하여 } \mu_{ij} = n\pi_{i+}\pi_{+j} \right]$

1. 0차원 분할표에 대한 독립성 로그 선형모형

- 행 변수를 X , 열 변수를 Y 로 표시할 때

- X 와 Y 가 독립

$$\Leftrightarrow \mu_{ij} = n\pi_{i+}\pi_{+j} \quad (\mu_{ij} = n\pi_{ij})$$

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

λ_i^X : 변수 X 의 i 번째 수준의 효과

λ_j^Y : 변수 Y 의 j 번째 수준의 효과

(기대도수의 로그 함수값은 행 효과 λ_i^X 와 열 효과 λ_j^Y 의 가법함수)

1. 이차원 분할표에 대한 독립성 로그 선형모형

- 두 범주형 변수의 독립성 \Leftrightarrow 로그 선형모형의 적합성

- X^2 와 G^2 (독립성검정)를 이용하여 로그 선형모형의 적합도 판단

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \quad G^2 = 2 \sum n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

$\hat{\mu}_{ij} = n_{i+}n_{+j}/n$, X^2 와 G^2 는 자유도 $(I-1)(J-1)$ 인

카이제곱 분포로 근사

- 로그 선형모형에서는 반응변수와 설명변수를 구별하지 않으며,
 X 와 Y 를 모두 반응변수로 간주함

2. 독립성 모형에서의 모수 해석

- 로그선형모형은 GLM의 한 가지 유형임
 - 각각의 칸 도수를 포아송 분포에서 추출된 독립적인 관측값으로 간주함
 - SAS에서는 PROC GENMOD로 분석 가능

2. 독립성 모형에서의 모수 해석

- 반응변수 Y 의 수준이 두 개인 $I \times 2$ 분할표에 대한 독립성 로그 선형모형 적용

- i 번째 행에서 $Y=1$ 인 확률 π 에 대한 로짓

$$\begin{aligned}
 \log[P(Y=1)/1-P(Y=1)] &= \log(\mu_{i1}/\mu_{i2}) \\
 &= \log(\mu_{i1}) - \log(\mu_{i2}) \\
 &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) \\
 &= \lambda_1^Y - \lambda_2^Y
 \end{aligned}$$

〔 $\rightarrow Y$ 에 대한 로짓은 변수 X 의 수준에 영향을 받지 않음 〕

2. 독립성 모형에서의 모수 해석

예

일반사회조사의 사후세계에 대한 믿음 조사 결과

	Yes	No	계
행복하지 않음	190	32	222
보통 행복	611	113	724
매우 행복	326	51	377
계	1127	196	1323

2. 독립성 모형에서의 모수 해석

독립성 로그 선형모형을 적합하기 위한 R프로그램

```

> HappyHeaven <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                             HappyHeaven.dat", header=TRUE)
> HappyHeaven      # Data file HappyHeaven at text website
  happy  heaven count
1   not     no   32
2   not    yes  190
3 pretty    no  113
4 pretty    yes  611
5  very     no   51
6  very     yes  326
> fit <- glm(count ~ happy + heaven, family=poisson, data=HappyHeaven)
> # canonical link for Poisson is log, so "(link=log)" is not necessary
> # loglm function in MASS library also fits loglinear models
> summary(fit) # independence loglinear model
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.49313    0.09408   37.13  < 2e-16
happypretty   1.18211    0.07672   15.41  < 2e-16
happyvery     0.52957    0.08460    6.26 3.86e-10
heavenyes     1.74920    0.07739   22.60  < 2e-16
---
Residual deviance:  0.89111  on 2  degrees of freedom

```

3. 이차원 분할표에 대한 포화 모형

- 독립성을 만족하지 않는 경우의 로그 선형모형 적합
 - $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$
여기서 $\{\lambda_{ij}^{XY}\}$ 는 독립성에서 벗어난 정도를 반영하는 연관성을 나타냄
 - 이차원 분할표에 대한
포화로그선형모형(saturated loglinear model)이라고 함
 - 독립성 모형은 $\lambda_{ij}^{XY} = 0$ 인 특별한 경우임

3. 이차원 분할표에 대한 포화 모형

- 로그오즈비와 $\{\lambda_{ij}^{XY}\}$ 는 직접적으로 연관됨

예

2X2 분할표


$$\begin{aligned}
 \log\theta &= \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21}) \\
 &= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
 &\quad - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\
 &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}
 \end{aligned}$$

3. 0차원 분할표에 대한 포화 모형

- $\{\lambda_{ij}^{XY}\}$ 는 로그오즈비를 결정함 (모든 i, j 에 대하여 $\lambda_{ij}^{XY} = 0$)
 $\rightarrow \log\theta = 0$ (X 와 Y 는 서로 독립)
- 모형에 $\log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$ 에 대한 모수의 수

모수	중복되지 않은 모수	비고
λ	1	
λ_i^X	$I-1$	($\lambda_I^X = 0$ 으로 놓을 수 있음)
λ_j^Y	$J-1$	($\lambda_J^Y = 0$ 으로 놓을 수 있음)
λ_{ij}^{XY}	$(I-1)(J-1)$	
계	IJ	

3. 이차원 분할표에 대한 포화 모형

- 모수의 총 수는 분할표의 칸 개수와 같음
- 포화 로그 선형모형은 가장 일반적인 모형임
- （ 이 모형은 기대도수를 완전하게 설명하고
관측도수를 완전하게 적합함 ($\hat{\mu}_{ij} = n_{ij}$)
- 실질적으로 로그 선형모형을 적합하여 분석할 때는 포화모형보다 간단하게 해석할 수 있는 비포화모형을 사용하여 분석함

3. 이차원 분할표에 대한 포화 모형

- 사례 : $I \times J$ 분할표에서 독립성 로그 선형모형

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

$$df = IJ - [1 + (I-1) + (J-1)] = (I-1)(J-1)$$

- X^2 또는 G^2 을 이용한 독립성 검정

⇔ 독립성 로그 선형모형에 대한 적합성 검정

- 포화모형 $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$ 을 적용하면

$$df = 0 \text{ 이고 완전 적합됨 } (\hat{\mu}_{ij} = n_{ij})$$

4. 삼차원 분할표에 대한 포화 모형

- $\{\mu_{ijk}\}$: 삼차원 분할표에서 칸 기대도수(cell expected frequency)

1 : 가장 일반적인 로그 선형모형(포화모형)

- $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$
- (XYZ) 로 표시함
- 삼차교호작용항을 포함하며 자료를 완전하게 적합함

4. 삼차원 분할표에 대한 포화 모형

- $\{\mu_{ijk}\}$: 삼차원 분할표에서 칸 기대도수(cell expected frequency)

2 : 동질연관성모형

- $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
- 어떤 두 변수 사이의 조건부오즈비는 나머지 변수의 모든 수준에서 동일함
- (XY, XZ, YZ) 로 표시함

4. 삼차원 분할표에 대한 포화 모형

- $\{\mu_{ijk}\}$: 삼차원 분할표에서 칸 기대도수(cell expected frequency)

3 : X 와 Y 사이의 조건부 독립성 모형

- $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
- (XZ, YZ) 로 표시함
- Z 를 통제했을 때 X 와 Y 사이에 연관성이 없음
 $\log[\theta_{XY(Z)}] = 0$ (Z 가 주어졌을 때 X 와 Y 는 서로 독립)

4. 삼차원 분할표에 대한 포화 모형

- $\{\mu_{ijk}\}$: 삼차원 분할표에서 칸 기대도수(cell expected frequency)

4 : 상호독립모형

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

5. 조건부 연관성을 나타내는 이차 교호 작용항 모수

- 로그선형모형에서 최고차 항은 모형에 대한 해석과 직접 관련됨
- 예 : $2 \times 2 \times k$ 분할표에서 (XZ, YZ) 모형의 경우

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

$$\begin{aligned} \log \theta_{XY(k)} &= \log \left(\frac{\mu_{11k} \mu_{22k}}{\mu_{12k} \mu_{21k}} \right) \\ &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} \end{aligned}$$

 위 식에서 오른쪽 부분은 k 에 의존하지 않음
 조건부 오즈비는 Z 의 수준에 상관없이 동일함

- (XY, XZ, YZ) 는 Y 의 모든 수준에서 동일한 XZ 오즈비를 가짐
- 삼차 교호작용항 λ_{ijk}^{XYZ} 을 포함하지 않는 모형은 동질연관성 만족

6. 음주, 흡연, 마리화나 경험 예제

■ 조사 data ($2 \times 2 \times 2$ 분할표)

알코올(A)	담배(C)	마리화나(M)	
		예	아니오
예	예	911	538
	아니오	44	456
아니오	예	3	43
	아니오	2	279

6. 음주, 흡연, 마리아나 경험 예제

■ 로그 선형모형 적합

- $(A, C, M) : \log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M$
- $(AC, M) : \log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC}$
- $(AM, CM) : \log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$
- $(AC, AM, CM) : \log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$
- $(ACM) : \log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM} + \lambda_{ijk}^{ACM}$

6. 음주, 흡연, 마리아나 경험 예제

■ 로그 선형모형 적합을 위한 R 프로그램

```

> Drugs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Substance.dat",
+                      header=TRUE)
> Drugs
  alcohol cigarettes marijuana count # data file has 8 rows, for 8 cell counts
1      yes          yes          yes   911
...
8       no           no           no   279
> A <- Drugs$alcohol; C <- Drugs$cigarettes; M <- Drugs$marijuana
> fit <- glm(count ~ A + C + M + A:C + A:M + C:M, family=poisson, data=Drugs)
> summary(fit) # homogeneous association model

              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.6334      0.0597   94.36  < 2e-16
Ayes         0.4877      0.0758    6.44 1.22e-10
Cyes        -1.8867      0.1627  -11.60  < 2e-16
Myes        -5.3090      0.4752  -11.17  < 2e-16
Ayes:Cyes     2.0545      0.1741   11.80  < 2e-16 # AC log odds ratio = 2.0545
Ayes:Myes     2.9860      0.4647    6.43 1.31e-10
Cyes:Myes     2.8479      0.1638   17.38  < 2e-16
---
Residual deviance:  0.37399  on 1  degrees of freedom

```

6. 음주, 흡연, 마리아나 경험 예제

■ 로그 선형모형 적합을 위한 SAS 프로그램

DATA drug;

INPUT a c m count @@;

CARDS;

1 1 1 911 1 1 2 538 1 2 1 44 1 2 2 456

2 1 1 3 2 1 2 43 2 2 1 2 2 2 2 279

RUN;

PROC GENMOD DATA=drug ;

Class a c m;

MODEL count = a c m a*c a*m c*m/ dist=poisson link=log lrci type3 obstats;

RUN;

6. 음주, 흡연, 마리화나 경험 예제

■ 각 로그 선형모형에 대한 적합 결과

음주	흡연	마리화나	로그 선형모형				
			(A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
예	예	예	540.0	611.2	909.24	910.4	911
		아니오	740.2	837.8	438.84	538.6	538
	아니오	예	282.1	210.9	45.76	44.6	44
		아니오	386.7	289.1	555.16	455.4	456
아니오	예	예	90.6	19.4	4.76	3.6	3
		아니오	124.2	26.6	142.16	42.4	43
	아니오	예	47.3	118.5	0.24	1.4	2
		아니오	64.9	162.5	179.84	279.6	279

6. 음주, 흡연, 마리아나 경험 예제

■ (AM, CM) 모형에서 조건부연관성과 주변부연관성 계산

1 : AC오즈비 계산(M의 두 수준에서 각각 구함)

$$1.0 = \frac{909.24 \times 0.24}{45.76 \times 4.76} = \frac{438.84 \times 179.84}{555.16 \times 142.16}$$

2 : AC의 주변부 연관성

		C	
		예	아니오
A	예	1348.08	600.92
	아니오	146.82	180.06

$$\frac{1348.08 \times 180.06}{600.92 \times 146.82} = 2.7$$

“(AM, CM)모형에서 AC간에는 조건부독립성은 성립하지만
주변부 독립성은 만족하지 않음”

02

제 10강. 분할표 및 도수자료에 대한 로그 선형모형

로그 선형모형의 통계적 추론

1. 카이제곱 적합성 검정

- 칸의 적합값과 관측값을 비교함으로써 모형의 적합도를 검정함

- 가능도비 통계량 (모형의 이탈도, Deviance) :

$$G^2 = 2 \sum n_{ijk} \log\left(\frac{n_{ijk}}{\widehat{\mu_{ijk}}}\right)$$

- 피어슨 통계량 :

$$X^2 = \sum \frac{(n_{ijk} - \widehat{\mu_{ijk}})^2}{\widehat{\mu_{ijk}}}$$

- 자유도 = (총 칸의 개수) - (추정할 모형의 모수 개수)
- 포화모형의 자유도는 0이고, 자유도는 모형이 복잡할수록 줄어듦

1. 카이제곱 적합성 검정

■ 음주(*A*), 흡연(*C*), 마리화나(*H*) 경험 여부에 대한 로그 선형모형 적합결과

모형	G^2	X^2	자유도	P - 값*
(<i>A, C, M</i>)	1286.0	1411.4	4	<0.001
(<i>A, CM</i>)	534.2	505.6	3	<0.001
(<i>C, AM</i>)	939.6	824.2	3	<0.001
(<i>M, AC</i>)	843.8	704.9	3	<0.001
(<i>AC, AM</i>)	497.4	443.8	2	<0.001
(<i>AC, CM</i>)	92.0	80.8	2	<0.001
(<i>AM, CM</i>)	187.8	177.6	2	<0.001
(<i>AC, AM, CM</i>)	0.4	0.4	1	0.54
(<i>ACM</i>)	0.0	0.0	0	-
* G^2 의 P - 값				

“모형 (*AC, AM, CM*)이 잘 적합함”

2. 로그 선형모형의 칸 표준화잔차

- 각 칸의 잔차(residual)을 통해서 모형의 적합 정도를 살필 수 있음

잔차로 피어슨 잔차(Pearson residual)와 수정 잔차(adjusted residual)을 고려함

- 로그선형모형의 표준화 잔차값

경험유무			관찰값	모형 AM, CM		모형 AC, AM, CM	
A	C	M		적합값	표준화잔차	적합값	표준화잔차
예	예	예	911	909.2	3.70	910.4	0.63
		아니오	538	438.8	12.80	538.6	-0.63
	아니오	예	44	45.8	-3.70	44.6	-0.63
		아니오	456	55.2	-12.80	455.4	0.63
아니오	아니오	예	3	4.8	-3.70	3.6	-0.63
		아니오	43	142.2	-12.80	42.4	0.63
	아니오	예	2	0.2	3.70	1.4	0.63
		아니오	279	179.8	12.80	279.6	-0.63

2. 로그 선형모형의 칸 표준화잔차

Note

- ① 표준화 잔차는 근사적으로 표준정규분포를 따름
- ② 표준화 잔차 값이 2~3보다 큰 경우에 적합결여를 나타냄

3. 조건부연관성에 대한 유의성검정

- 서로 다른 로그 선형모형을 비교하여 조건부연관성에 대하여 검정 가능
- 예 : (AC, AM, CM) 모형 고려
 - 음주(A)와 흡연(C) 간에 조건부연관성이 없다는 귀무가설의 검정
→ $H_0 : \lambda^{AC} = 0$ 의미
 - 모형(AM, CM)이 적합되는 가를 분석하는 것과 같은 의미

3. 조건부연관성에 대한 유의성검정

- 모형(AC, AM, CM)에서 $\lambda^{AC} = 0$ 을 검정하는 검정 통계량

$$G^2[(AM, CM)|(AC, AM, CM)] = G^2(AM, CM) - G^2(AC, AM, CM)$$

- 모형(AM, CM)의 $G^2 = 187.8 (df = 2)$
- 모형(AC, AM, CM)의 $G^2 = 0.4 (df = 1)$

➔ $G^2[(AM, CM)|(AC, AM, CM)] = 187.4, df = 1, P < 0.0001$

AC 간에 강한 조건부연관성이 있음

모형(AC, AM, CM)을 적합해야 함

4. 로그 선형모형의 베이지안 적합

- 동질연관성모형에서 로그 선형모형의 모수에 대해 사전분포로 평균 0, 표준편차 10인 정규분포 가정

```
-----
> library(MCMCpack) # b0 = prior mean, B0 = prior precision = 1/variance
> fitBayes <- MCMCpoisson(count ~ A + C + M + A:C + A:M + C:M, b0=0, B0=0.01,
+                          mcmc=10000000, data=Drugs)
> summary(fitBayes)
1. Empirical mean and standard deviation # showing only association parameters
      Mean      SD
Ayes:Cyes  2.0644  0.17481
Ayes:Myes  3.0639  0.48161
Cyes:Myes  2.8569  0.16442
2. Quantiles for each variable:
      2.5%      25%      50%      75%      97.5%
Ayes:Cyes  1.7296  1.9451  2.0616  2.1806  2.4152
Ayes:Myes  2.2111  2.7274  3.0321  3.3655  4.0983
Cyes:Myes  2.5437  2.7444  2.8540  2.9660  3.1879

> mean(fitBayes[,6] < 0) # posterior prob. that AM log odds ratio < 0
[1] 0                      # (parameter 6 in model is AM log odds ratio)
-----
```

5. 고차원 로그 선형모형

- 삼차원 분할표에 대한 로그 선형모형은 다차원 분할표로 확장 가능
- 변수 W, X, Y, Z 를 갖는 사차원 분할표에 대한 로그 선형모형을 고려함

1 : (WX, WY, WZ, XY, XZ, YZ) 모형

- 동질연관성 구조를 가짐
(임의의 두 변수는 조건부 종속이며 나머지 다른 두 변수의
결합 수준에 대하여 동일한 오즈비를 가짐)

5. 고차원 로그 선형모형

2 : (WX, WY, WZ, XZ, YZ) 모형 : (XY 항이 없음)

- X 와 Y 는 W 와 Z 의 수준들의 각 조합에서 조건부 독립 만족
- 이 모형은 삼차 교호작용항 WXY, WXZ, WYZ, XYZ 등을 포함할 수 있음
- 교재 P263의 자동차 사고와 안전벨트의 사례

5. 고차원 로그 선형모형

■ 교재 P263의 자동차 사고와 안전벨트의 사례

성별(G)	사고장소(L)	안전벨트(S)	부상(I)	
			아니오	예
여성	도시	미착용	7,287	996
		착용	11,587	759
	농촌	미착용	3,246	973
		착용	6,134	757
남성	도시	미착용	10,381	812
		착용	10,969	380
	농촌	미착용	6,123	1084
		착용	6,693	513

5. 고차원 로그 선형모형

분할표에 대한 로그 선형모형

: 반응변수와 설명변수를 구별하지 않고 범주형 변수들간의 연관성을 다룸

로지스틱 회귀모형

: 하나의 범주형 반응변수가 여러 설명변수에 의해서 어떻게 영향을 받는지 설명

6. 통계적 유의성과 실제적 유의성: 비유사성지수

- 표본의 크기는 추론과정의 결과에 큰 영향을 끼칠 수 있음
- 표본의 크기가 작은 경우는 적합성검정을 통과한 가장 단순한 모형보다 실제 상황은 훨씬 복잡할 수 있음
- 반면, 표본이 큰 경우는 통계적으로는 유의하다고 나온 효과가 실제로는 아주 약하거나 크게 중요하지 않을 수 있음
- 표본이 클 때는 표본자료와 모형 적합값의 유사성을 비교하는 것이 바람직함(이 방법은 표본크기에 영향을 크게 받지 않음)
- 비유사성지수(dissimilarity index)

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

cell counts $\{n_i = np_i\}$ and fitted values $\{\hat{\mu}_i = n\hat{\pi}_i\}$

03

제 10강. 분할표 및 도수자료에 대한 로그 선형모형

로그 선형모형과 로지스틱회귀모형의 관련성

1. 로그 선형모형 해석을 위한 로지스틱회귀 모형 사용

- 로그 선형모형에서 한 변수가 이항형 반응변수이고, 나머지 변수는 설명변수(범주형)인 경우를 고려함
- 삼차원 분할표에서 동질 연관성 모형

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

1. 로그 선형모형 해석을 위한 로지스틱회귀 모형 사용

- Y 를 이항형 반응변수로 가정하고, X 와 Z 를 설명변수로 가정
- X 의 수준이 i 이고 Z 의 수준이 k 라고 할 때

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

$$\rightarrow \text{logit}[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z$$

1. 로그 선형모형 해석을 위한 로지스틱회귀 모형 사용

- 이항반응변수 Y 를 갖는 삼차원 분할표에서 로그 선형모형과 로지스틱회귀모형의 관계

로그선형모형	로지스틱회귀모형
(Y, XZ)	α
(XY, XZ)	$\alpha + \beta_i^X$
(YZ, XZ)	$\alpha + \beta_k^Z$
(XY, YZ, XZ)	$\alpha + \beta_i^X + \beta_k^Z$
(XYZ)	$\alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$

- 로그 선형모형은 적어도 두 변수들이 반응변수일 때 사용하는 것이 바람직하고, 반응변수가 하나인 경우는 로지스틱회귀 모형을 적합하는 것이 좋음

04

제 10강. 분할표 및 도수자료에 대한 로그 선형모형

독립성 그래프와 붕괴 가능성

1. 독립성 그래프 (Independence Graph)

■ 로그 선형모형에서 그래프를 이용하여 조건부 독립성을 표현하는 방법

- 로그 선형모형의 독립성 그래프는 꼭지점의 집합이며 각 꼭지점은 변수를 나타냄
 - 분할표의 차원의 수 만큼 꼭지점들이 있고, 두 꼭지점은 선분으로 연결되거나 연결되지 않음
 - 선분으로 연결되지 않는 경우는 꼭지점에 대응하는 두 변수간의 조건부 독립성을 나타냄

1. 독립성 그래프 (Independence Graph)

- 예 : 사차원 분할표 (W, X, Y, Z 로 구성)에 대해서 로그 선형모형 (WX, WY, WZ, YZ) 고려

- XY 와 XZ 의 연관성이 빠져 있음

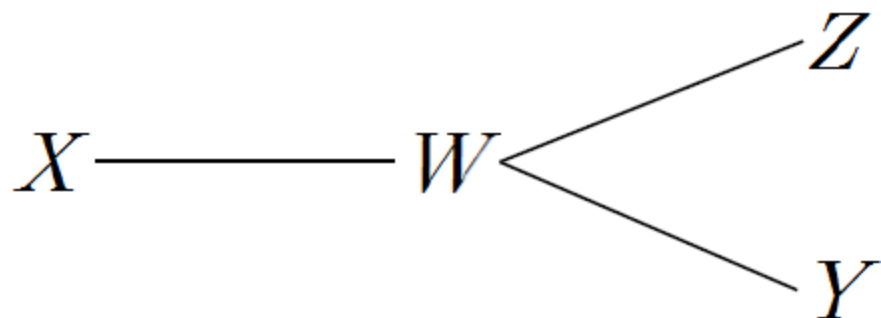
X 와 Y , X 와 Z 는 각각 나머지 두 변수가 주어졌을 때,

조건부 독립임

W 와 X , $\{ W, Y, Z \}$ 간의 모든 가능한 쌍들은 연관성을 가짐

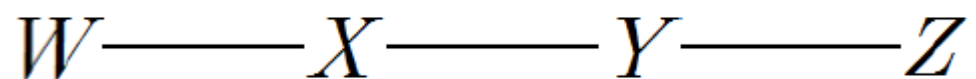
1. 독립성 그래프 (Independence Graph)

■ 독립성 그래프



$$(WX, WY, WZ, YZ) \equiv (WX, WYZ)$$

■ 로그선형모형 (WX, XY, YZ)의 독립성 그래프



2. 삼차원 분할표에서 붕괴가능성 조건

- 부분분할표의 연관성은 주변 연관성과 다를 수 있음

(Z 가 주어졌을 때 X 와 Y 가 조건부 독립이라도 X 와 Y 가 반드시 주변 독립인 것은 아님)

- 붕괴가능성 조건(Collapsibility Condition)

삼차원 분할표에서 Z 와 X 가 조건부 독립이거나 Z 와 Y 가 조건부 독립이면 XY 의 주변 오즈비와 조건부 오즈비는 서로 같음

- 로그선형모형(XY, YZ) 또는 (XY, XZ)에 해당하면 XY 의 연관성이 부분 분할표나 주변 분할표에서 동일함

($X-Y-Z$ 또는 $Y-X-Z$)

11

강

다음시간안내

선형혼합모형

수고하셨습니다.