3 장에서는 종속변수(반응변수)가 이진형(binary)인 경우이고 설명 변수가 하나이고 측정형인 경우 분석 방법인 Logit Regression model, Probit model 을 살펴보았다. 물론 마지막 절에서는 설명변수가 분류형일 경우 Logit regression model 이 어떻게 이용될 수 있는지 살펴보았지만, 이것이 이 장에서 다룰 Log-linear 분석 몫이다.

Log-linear 분석은 종속변수와 독립 변수가 모두 범주형(분류형)인 경우 인과 관계를 분석하는 방법이다. 이 분석에서는 분할표의 셀 빈도를 변수들(설명 변수들과 반응 변수)의 관계로 표현한다. 반응 변수가 이진형이면 Log-linear 분석은 Logit 분석과 동일하다.

종속변수가 이진형이고 설명변수가 2 개 이상이고 Mixed(측정형, 분류형)인 경우는 Logistic Regression model 을 사용하면 된다. [페이지 44~ 회귀분석 측면 참고: PROC LOGISTIC 절차에서도 /SELECTION=MODEL, 즉 변수 선택이 가능, LOGISTIC 분석은 개체 분류하는 판별 분석에도 사용된다. 페이지 51 참고].

Recall Homework #8: 페이지 57 의 Homework6-3 문제 수정

TAX.txt 자료는 다음 변수에 대한 자료이다. 다음 절차에 의해 Logistic 분석을 실시하시오.

1) 적절한 변수를 선택하고 (유의수준=0.1) 분석 결과를 해석하시오. (회귀분석)
2) 판별 분석을 실시하고 Classification Table 을 보고 적절한 Phat 기준을 선택하시오. (분류에 참고)
3) 새로운 사람의 정보는 다음과 같다. 세금 보고 전문 기업은 이 사람에게 DM 발송을 할 필요가 있겠는가? 결혼, 자기 사업, 부양 가족=10 명, 세금 효율=3, 소득=12.3

종속변수: PREP(세금 보고 전문가 이용=1, 자신이 직접=0)

독립변수:
1)MA (결혼 여부, 1=결혼, 0=미혼) Indicator 변수
2)SE (자기 사업=1, 취업=0) Indicator 변수
3)DEP (부양 가족 수) : 측정형 변수(연속형)
4)TR (세금 효율:rate) : 측정형 변수(연속형)
5)INCOME (소득) : 측정형 변수(연속형)

4.1. Log-linear Model for 2 dimension

IxJ 분할표의 총 N(=i*j)개의 셀에서 n 개의 표본을 추출하는 다항 분포를 고려하자. 다항 분포에서 확률 π_{ij} 가 2 차원 분할표의 (2 dimension contingency table) 결합 밀도 함수를 형성한다. 만약 반응이 서로 독립이면 $\pi_{ij} = \pi_{i+}\pi_{+j}$ for $i = 1,2,...,I$ and $j = 1,2,...,J$ 이다. 그러므로 가 셀의 기대 도수 $E_{ij} = m_{ij} = n\pi_{i+}\pi_{+j}$ 이다. Log-linear 모형에서는 확률 π_{ij} 대신 m_{ij} 를 사용하여 모형을 설정한다.

2X2 분할표에 대해 예제(성별에 따른 사후 세계 믿음 여부 차이)를 통해 Log-linear model 을 설명해보기로 하자.

	믿는다	안 믿는다
남자	435	147
여자	375	134

```
data one;
    input gender $ postdeath $ f @@;
    cards;
m y 435 m n 147 f y 375 f n 134
run;
proc freq data=one order=data;
    weight f;
    table gender*postdeath/chisq nopercent nocol expect;
run;
```

```
          빈도|
        기대도수|
      행 백분율|y        |n        |   총합
---------------+---------+---------+
m              |     435 |     147 |    582
               |   432.1 |   149.9 |
               |   74.74 |   25.26 |
---------------+---------+---------+
f              |     375 |     134 |    509
               |   377.9 |   131.1 |
               |   73.67 |   26.33 |
---------------+---------+---------+
총합               810       281      1091
```

통계량	자유도	값	확률값
카이제곱	1	0.1621	0.6872
우도비 카이제곱	1	0.1620	0.6873

$\chi^2 = 0.16\,(df = 1)$ 이므로 성별의 차이는 없다.

4.1.1. Independence model

만약 두 변수간에 독립을 가정하면 (i, ,j) 셀의 기대빈도의 Log 는 다음과 같다.

$$\ln m_{ij} = \ln n + \ln \pi_{i+} + \ln \pi_{+j}$$

행 변수(일반적으로 독립 변수)를 X, 열 변수를 (종속 변수) Y 라 하면 위의 식은

Log-linear model of independence $\ln m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$ --- (1)

where $\lambda_i^X = \ln \pi_{i+} - (\sum_h \ln \pi_{h+})/I$, $\lambda_i^Y = \ln \pi_{+j} - (\sum_h \ln \pi_{+h})/J$,

$\mu = \ln n + (\sum_h \ln \pi_{h+})/I + (\sum_h \ln \pi_{+h})/J$.

제약 조건 $\sum \lambda_i^X = \sum \lambda_j^Y = 0$

모수 λ_i^X, λ_i^Y 는 평균에 대한 편차(deviation)이다.

$\ln \hat{m}_{ij}$	믿는다	안 믿는다
남자	6.069	5.010
여자	5.935	4.876

2X2 분할표의 경우 Independence model 의 모수 해석은

$$\ln\theta = \ln(\frac{m_{11}m_{22}}{m_{12}m_{21}}) = \ln m_{11} + \ln m_{22} - \ln m_{12} - \ln m_{21}$$
$$= (\mu + \lambda_1^X + \lambda_1^Y) + (\mu + \lambda_2^X + \lambda_2^Y) - (\mu + \lambda_1^X + \lambda_2^Y) - (\mu + \lambda_2^X + \lambda_1^Y) = 0$$

제약 조건 $\sum \lambda_i^X = \sum \lambda_j^Y = 0$ 과 $\ln m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$ 을 이용하여 식(1)의 모수에 대한 추정치를 구하면 다음과 같다. 유일 근(독립 모형에서는 각 요인에서 모수가 중복적으로 정의되어)이 아니므로 요인의 마지막 수준을 0 으로 하거나(방법 1: SAS GENMOD) 첫 수준을 0 으로 하거나(방법 2: SAS GENMOD) 모수의 합을 0 으로 한 방법(방법 3: SAS CATMOD)으로 모수를 추정할 수 있다.

	μ	λ_1^X	λ_2^X	λ_1^Y	λ_2^Y
방법 1	4.876	0.134	0	1.059	0
방법 2	6.069	0	-0.134	0	-1.059
방법 3	5.472	0.067	-0.067	0.529	-0.529

1 행 2 열을 보면 $\ln m_{12} = \mu + \lambda_1^X + \lambda_2^Y = 4.876 + 0.134 + 0 = 5.01 = \ln(149.9)$
그리고 어떤 방법을 사용하더라도 요인의 주효과(main effect)를 나타내는 모수간 차이는 항상 동일하다. 예를 들어 $\lambda_1^Y - \lambda_2^Y = 1.059$ 이다. 그러므로 $\ln\hat{\theta} = \ln(\frac{\hat{\pi}}{1-\pi}) = 1.059$ 이고 odds ratio 의 추정치 $\hat{\theta}$ 는 $e^{1.059} = 2.88$ 이다. (2x2 분할표 방법과 동일= $\exp[\frac{(435\times134)}{(375\times147)}]$)

4.1.2. Saturated model

만약 변수들간에 독립이 성립하지 않는다고 가정하자.

그리고 $n_{ij} = \ln m_{ij}$, $n_{i+} = \frac{\sum_j n_{ij}}{J}$, $n_{+j} = \frac{\sum_i n_{ij}}{I}$, $\mu = n_{++} = \frac{\sum_i \sum_j n_{ij}}{I \times J}$ 라 놓고

$\lambda_i^X = n_{i+} - n_{..}$, $\lambda_i^Y = n_{+j} - n_{..}$, $\lambda_{ij}^{XY} = n_{ij} - n_{i+} - n_{+j} + n_{..}$ 라 하면 다음과 같이 놓을 수 있다.

$$\lg m_{ij} = \mu + \lambda_i^X + \lambda_j^X + \lambda_{ij}^{XY}$$, 제약 조건 $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$ --- (2)

모수의 개수 μ 1 개, λ_i^X 형태의 비중복 모수 수 (I-1), λ_j^Y 는 (J-1), λ_{ij}^{XY} 는 (I-1)X(J-1) 이므로 총 모수 수는 IJ 개이다. 이 경우 모수의 수가 가자 많으므로 "꽉 찼다"는 의미의 saturated model 이라 한다.

(cf) Independence model = reduced model (귀무가설이 성립할 경우) 식 (1)
Saturated model = full model 식 (2)

식 (2)와 같은 모형을 hierarchical model (층화 모형) 이라 한다. 층화 모형이란 차수 항이 높은 요인이 있으면 저차 항은 반드시 포함되어 있는 것이다. λ_{ij}^{XY} 이 있으면 λ_i^X, λ_j^Y 이 들어 있는 경우이다. 층화 모형이 선호되는 이유는 낮은 차수 항이 포함되지 않으면 고차 항에(교차 효과와 비슷) 대한 해석이 어렵기 때문이다.

그리고 $n_{ij} = \ln m_{ij}$, $n_{i+} = \frac{\sum_j n_{ij}}{J}$, $n_{+j} = \frac{\sum_i n_{ij}}{I}$, $\mu = n_{++} = \frac{\sum_i \sum_j n_{ij}}{I \times J}$ 라 놓고

$\lambda_i^X = n_{i+} - n_{..}$, $\lambda_i^Y = n_{+j} - n_{..}$, $\lambda_{ij}^{XY} = n_{ij} - n_{i+} - n_{+j} + n_{..}$ 라 하면 다음과 같이 놓을 수 있다.

$$\ln m_{ij} = \mu + \lambda_i^X + \lambda_j^X + \lambda_{ij}^{XY}$$, 제약 조건 $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$ --- (2)

위의 모형은 교차(interaction) 항이 있는 two-way ANOVA 모형과 동일하다. λ_i^X 는 평균에 대한 편차이므로 만약 $\lambda_i^X > 0$ 이면 i 행 셀들의 기대치(물론 log 기대빈도의 평균)는 전체 분할표의 기대치보다 높다.

Saturated model 의 모수의 수는 $1+(I-1)+(J-1)+(I-1)(J-1)=IJ$ 이고 independent model 의 모수 수는 $1+(I-1)+(J-1)=I+J-1$ 이고 만약 모든 $\lambda_{ij}^{XY}=0$ 이면 두 변수는 서로 독립이다.

2X2 분할표의 경우 Saturated model 의 모수 해석은

$$\ln\theta=\ln(\frac{m_{11}m_{22}}{m_{12}m_{21}})=\ln m_{11}+\ln m_{22}-\ln m_{12}-\ln m_{21}$$
$$=(\mu+\lambda_1^X+\lambda_1^X+\lambda_{11}^{XY})+(\mu+\lambda_2^X+\lambda_2^X+\lambda_{22}^{XY})-(\mu+\lambda_1^X+\lambda_2^X+\lambda_{12}^{XY})-(\mu+\lambda_2^X+\lambda_1^X+\lambda_{21}^{XY})$$
$$=\lambda_{11}^{XY}+\lambda_{22}^{XY}-\lambda_{12}^{XY}-\lambda_{21}^{XY}$$

조건 $\sum_i \lambda_{ij}^{XY}=\sum_j \lambda_{ij}^{XY}=0$ 에 의해 $\lambda_{11}^{XY}=\lambda_{22}^{XY}=-\lambda_{12}^{XY}=-\lambda_{21}^{XY}$ ➔ $\log\theta=4\lambda_{11}^{XY}$

그러므로 $\lambda_{11}^{XY}=0$ (독립)이면 Odds ratio 는 1 이 된다. (Recall: 독립)

식 (2)는 $m_{ij}=\exp^{(\mu+\lambda_i^X+\lambda_j^X+\lambda_{ij}^{XY})}$ 이고 셀 확률 $\pi_{ij}=\frac{m_{ij}}{\sum\sum m_{ab}}$ 는

$$\pi_{ij}=\frac{\exp^{(\mu+\lambda_i^X+\lambda_j^X+\lambda_{ij}^{XY})}}{\sum\sum\exp^{(\mu+\lambda_i^X+\lambda_j^X+\lambda_{ij}^{XY})}}$$

IxJ 분할표에서는 (I-1)x(J-1)개의 연관성 모수만을 중복되지 않게 정의할 수 있고 독립성 검정은 (I-1)x(J-1)개의 모수들이 0 인지를 검정한다. 그러므로 2X2 에서는 1 개의 모수가 odds ratio 를 결정한다.
다음은 예제 자료(2x2 분할표)의 연관성 관련 모수를 추정한 예이다.

	λ_{11}^{XY}	λ_{12}^{XY}	λ_{21}^{XY}	λ_{22}^{XY}
방법 1	0.056	0	0	0
방법 2	0.014	−0.014	−0.014	0.014
방법 3	0	0	0	0.056

$\ln\theta=\lambda_{11}^{XY}+\lambda_{22}^{XY}-\lambda_{12}^{XY}-\lambda_{21}^{XY}=0.056$ ➔ $\hat{\theta}=e^{0.056}=1.057$

4.1.3 SAS 사용 예제

```
data one;
    input x y wt @@;
    cards;
    1 1 435 1 2 147 2 1 375 2 2 134
run;
title 'independent model';
proc catmod data=one;
    weight wt;
    model x*y=_response_/ml pred=freq noprofile noresponse nodesign;
    loglin x y;
run;

title 'saturated model';
proc catmod data=one;
    weight wt;
    model x*y=_response_/ml pred=freq noprofile noresponse nodesign;
    loglin x y x*y;
run;
```

[Independence model]

Analysis of Maximum Likelihood Estimates

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
x	1	0.0670	0.0303	4.88	0.0272
y	2	0.5293	0.0346	233.83	<.0001

75page

Maximum Likelihood Predicted Values for Response Functions

Function Number	Observed Function	Observed Standard Error	Predicted Function	Predicted Standard Error	Residual
1	1.177506	0.0988	1.192702	0.092066	-0.0152
2	0.092593	0.119438	0.134022	0.060686	-0.04143
3	1.029086	0.100645	1.05868	0.069234	-0.02959

76page

Maximum Likelihood Predicted Values for Frequencies

x	y	Observed Frequency	Observed Standard Error	Predicted Frequency	Predicted Standard Error	Residual
1	1	435	16.17276	432.099	14.45822	2.901008
1	2	147	11.27801	149.901	8.796712	-2.90101
2	1	375	15.68772	377.901	13.96713	-2.90101
2	2	134	10.84167	131.099	7.963844	2.901008

74page

Analysis of Maximum Likelihood Estimates

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
x	1	0.0603	0.0347	3.02	0.0822
y	2	0.5285	0.0347	232.39	<.0001
x*y	3	0.0140	0.0347	0.16	0.6873

74page[χ^2]

Maximum Likelihood Predicted Values for Response Functions

Function Number	Observed Function	Observed Standard Error	Predicted Function	Predicted Standard Error	Residual
1	1.177506	0.0988	1.177506	0.0988	0
2	0.092593	0.119438	0.092593	0.119438	0
3	1.029086	0.100645	1.029086	0.100645	0

Maximum Likelihood Predicted Values for Frequencies

x	y	Observed Frequency	Observed Standard Error	Predicted Frequency	Predicted Standard Error	Residual
1	1	435	16.17276	435	16.17276	1.16E-10
1	2	147	11.27801	147	11.27801	0
2	1	375	15.68772	375	15.68772	0
2	2	134	10.84167	134	10.84166	-196E-12

기대 도수와 관측 도수는 같다. (saturated model)

4.2. Log-linear Model for 3 dimension

인과 관계 연구에서 중요한 것은 예측 변수(predictor)와 통제 변수(control variable)를 어떻게 잘 선택하냐 하는 것이다. 하나의 반응변수와 하나의 설명 변수 간의 관계를 연구할 때 그 관계에 영향을 미치는 변량(covariate)을 조정해야 한다. 예를 들어 간접 흡연의 효과를 알아보기 위하여 남편이 흡연하는 아내들의 폐암 발생율과 남편이 비흡연자인 아내들의 폐암 발생율을 비교할 수 있을 것이다. 종속변수는 폐암 발생 여부, 설명변수는 남편 흡연 여부이다. 이 경우 인과 관계를 제대로 분석하려면 여자의 나이, 사회학적 수준, 근무 환경 등을 조정해야 한다.

4.2.1. Partial Association

변수가 3 개(X, Y, Z)이고 모두 범주형(분류형)이라면 다항 분할표를 얻을 수 있다. 이 경우 Z 의 값에 따라 X-Y 분할표를 얻을 수 있다. 이 분할표를 partial table 이라 하고 z 는 controlled 되었다고 한다. partial table 을 결합하여 얻어진 분할표를 X-Y marginal table 이라 하는데 이 경우 z 는 무시되었다고 본다.

4.2.2. Death Penalty Example

다음 Table 5.1 은 Radelet(1981)의 2x2x2 분할표로 살인 사건의 피고(defendant) 인종에 따른 사형 판결(death penalty)의 차이는 있는지 알아보고자 조사한 자료이다. 총 관측치 수는 326 명. 종속변수는 Death penalty, 설명변수는 Defendant race, 그리고 control 변수가 victim race 이다.

Table 5.1 Death Penalty Verdict by Defendant's Race and Victim's Race

Defendant's Race	Victim's Race	Death Penalty: Yes	Death Penalty: No	Percentage Yes
White	White	19	132	12.6
	Black	0	9	0.0
Black	White	11	52	17.5
	Black	6	97	5.8

Source: Reprinted with permission from Radelet (1981).

빨간 박스 안은 Victim(피해자)의 인종을 무시하고 구한 사형 언도 받은 사람 비율이다. 이것만 보면 흑인의 사형 언도 비율은 약 10%, 백인의 사형 언도 비율은 12%로 흑인이 낮다. [Table 5.2]

Table 5.2 Frequencies for Death Penalty Verdict and Defendant's Race

Defendant's Race	Death Penalty: Yes	Death Penalty: No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

그러나 victim 인종을 고려하여 보자. Victim 이 백인일 때 피의자 흑인의 사형 언도 비율은 4.9%(=17.5-12.6) 높고 victim 이 흑인일 때 피의자 백인의 사형 언도 비율은 흑인에 비해 5.8% 낮다. 즉 victim 인종을 control 하면 흑인의 사형 언도 비율이 높다. control 변수를 고려하면 왜 두 변수 간의 관계의 방향이 변하는가? Table 5.3.을 보자.

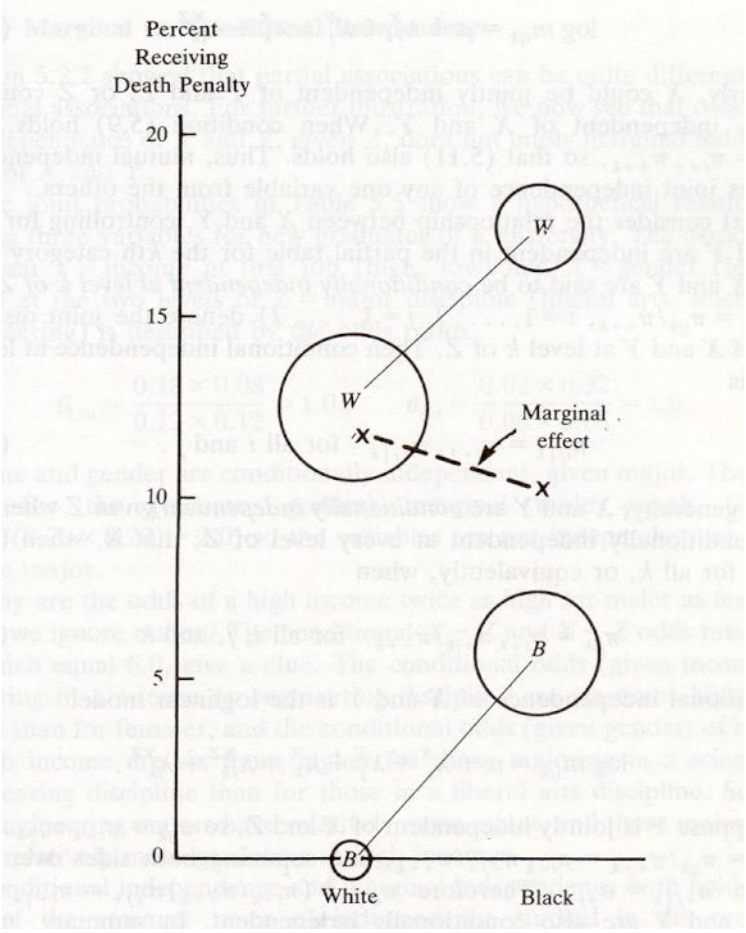
Table 5.3 의 Odds ratio 를 계산할 때는 셀이 0 인 셀이 있어 각 셀에 0.5 을 더하여 계산하였다. Marginal (victim 인종이 무시) 값을 살펴보면 Defendant 가 백인인 경우 흑인보다 사형 언도 받을 가능성은 1.18 배이다. Partial (victim 의 인종이 control)을 보면 victim 인종이 백인인 경우(Level1) 백인 defendant 사형 언도 가능성은 흑인의 0.67 배, victim 이 흑인일 경우 0.79 배로 marginal 의 결과와 반대가 된다. 이는 왜 그럴까? victim 인종과 defendant 인종 간의 odds ratio 의 값이 매우 높다. 즉 victim 백인인 경우 defendant 백인이 흑인에 비해 25.99 배이다.

Table 5.3 Odds Ratios for Death Penalty (P), Victim's Race (V), and Defendant's Race (D)[a]

		Variables		
Association		P–D	P–V	D–V
Marginal		1.18	2.71	25.99
Partial	Level 1	0.67	2.80	22.04
	Level 2	0.79	3.29	25.90

[a] The value 0.5 was added to each cell frequency before calculation of odds ratios.

victim 인종이 백인이 많고 백인이 백인을 많이 살해하므로 victim 인종만을 고려하지 않으면 백인이 흑인에 비해 사형 언도 받을 가능성이 높다고 결론지을 수 있으나 victim 인종을 고려하면 해석은 달라진다. 다음은 marginal 과 partial 의 효과의 차이를 보여준 것이다. 동그라미는 defendant 인종과 victim 인종의 결합에서 관측치 크기이다. 이렇게 marginal 과 partial 효과가 달라지는 경우를 Simpson Paradox 라 한다.



4.2.3. Independence 종류

변수 X, Y, Z 3 개 있다고 가정하자.

Table 5.4 Summary of Independence Models

Model	Probabilistic Form for π_{ijk}	Association Terms in Loglinear Model	Interpretation
(5.10)	$\pi_{i++}\pi_{+j+}\pi_{++k}$	None	Variables mutually independent
(5.12)	$\pi_{i+k}\pi_{+j+}$	λ_{ik}^{XZ}	Y independent of X and Z
(5.15)	$\pi_{i+k}\pi_{+jk}/\pi_{++k}$	$\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}$	X and Y independent, given Z

식 (5.10)의 mutual independence 는 log-linear model $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$

식 (5.12)은 log-linear model $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$

식 (5.15)은 log-linear model $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

4.2.4. Marginal vs. conditional Independence

다음 자료는 변수 X(성별), Y(소득수준), Z(전공)의 연관성 분석을 위한 자료라 하자.

Table 5.5 Conditional Independence Does Not Imply Marginal Independence

		Income	
Major	Gender	Low	High
Liberal Arts	Female	0.18	0.12
	Male	0.12	0.08
Science or Engineering	Female	0.02	0.08
	Male	0.08	0.32
Total	Female	0.20	0.20
	Male	0.20	0.40

전공이 주어진 경우 성별과 소득수준과의 연관성은 odds ratio 에 의해 계산되는데…

Liberal art: $\theta = \frac{0.18\times 0.18}{0.12\times 0.12} = 1$, Science: $\theta = \frac{0.02\times 0.32}{0.08\times 0.08} = 1$ ➔ 서로 독립 (conditional)

전공을 무시한 성별과 소득수준과의 연관성은 odds ratio 에 의해 계산되는데…

$\theta = \frac{0.2\times 0.4}{0.2\times 0.2} = 2$ ➔ 독립이 아님 (marginal)

전공을 무시할 때 소득 수준의 high 의 odds ratio 의 경우 여학생보다 남학생이 2 배 높다. 왜 이런 경우가? 해답은 성별과 전공, 소득 수준과 전공의 conditional adds ratio 는 6 이다. 소득이 주어진 경우 전공 과학의 전공 선택은 남자가 6 배 높고, 성별이 주어진 경우 소득 수준이 높은 사람은 과학 전공자가 인문과학 전공자보다 6 배 높다.

만약 Y 가 (X, Z)와 joint 독립이라면 $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$ 이다(conditionally independence). 만약 양변을 k 에 대해 합하면 $\pi_{ij+} = \pi_{i++}\pi_{+j+}$ 이므로 X, Y 는 marginal 독립이다. 그러므로 Y 가 (X, Z)와 독립이라면 X, Y 는 conditionally, marginally 독립이다.

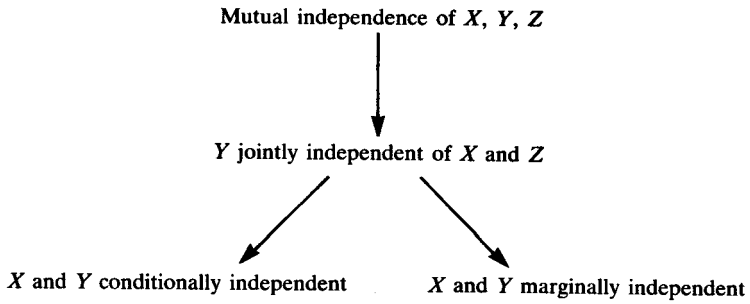


Figure 5.3 Relationships among types of X–Y independence.

Three-factor interaction model

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

4.3. Log-linear models for 3 dimension

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

$$\mu = \eta_{...}$$

$$\lambda_i^X = \eta_{i..} - \eta_{...}, \qquad \lambda_j^Y = \eta_{.j.} - \eta_{...}, \qquad \lambda_k^Z = \eta_{..k} - \eta_{...}$$

$$\lambda_{ij}^{XY} = \eta_{ij.} - \eta_{i..} - \eta_{.j.} + \eta_{...}$$

$$\lambda_{ik}^{XZ} = \eta_{i.k} - \eta_{i..} - \eta_{..k} + \eta_{...}$$

$$\lambda_{jk}^{YZ} = \eta_{.jk} - \eta_{.j.} - \eta_{..k} + \eta_{...}$$

$$\lambda_{ijk}^{XYZ} = \eta_{ijk} - \eta_{ij.} - \eta_{i.k} - \eta_{.jk} + \eta_{i..} + \eta_{.j.} + \eta_{..k} - \eta_{...}.$$

of the parameters for any index equals zero. That is,

$$\lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \cdots = \sum_k \lambda_{ijk}^{XYZ} = 0.$$

eneral loglinear model for a three-way table is

Table 5.6 Some Loglinear Models for Three-Dimensional Tables

Loglinear Model	Symbol
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	(X, Y, Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	(XY, Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	(XY, YZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	(XY, YZ, XZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$	(XYZ)

(X, Y, Z) ➔ 모두 독립

(XY, Z) ➔ Z 는 (X, Y)와 독립

(XY, YZ) ➔ Y 가 주어진 경우 X 와 Z 가 독립

(XY, YZ, XZ) ➔ X, Y, Z 의 어떤 쌍도 서로 조건적 독립이 아니고 3 차 교차 항이 없다.

(XYZ) ➔ X, Y, Z 의 어떤 쌍도 서로 조건적 독립이 아니고 각 쌍의 odds ratio 가 다른 변수에 따라 변한다.

Example

This analysis reproduces the predicted cell frequencies for Bartlett's data using a log-linear model of no three-variable interaction (Bishop, Fienberg, and Holland 1975, p. 89). Cuttings of two different lengths (Length=short or long) are planted at one of two time points (Time=now or spring), and their survival status (Status=dead or alive) is recorded.

```
title "Bartlett's Data";
data bartlett;
    input Length Time Status wt @@;
    datalines;
   1 1 1 156     1 1 2  84     1 2 1 84     1 2 2 156
   2 1 1 107     2 1 2 133     2 2 1 31     2 2 2 209
   ;

proc catmod data=bartlett;
    weight wt;
    model Length*Time*Status=_response_
    / noparm noresponse pred=freq;
    loglin Length|Time|Status;
quit;
```

Response Profiles

Response	Length	Time	Status
1	1	1	1
2	1	1	2
3	1	2	1
4	1	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	2	2	2

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Length	1	4.27	0.0388
Time	1	6.84	0.0089
Length*Time	1	6.84	0.0089
Status	1	49.92	<.0001
Length*Status	1	49.92	<.0001
Time*Status	1	94.75	<.0001
Length*Time*Status	1	2.26	0.1324
Likelihood Ratio	0	.	.

3 차 교차항만 유의하지 않으므로 모형은 다음과 같다.

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

```
proc catmod data=bartlett;
    weight wt;
    model Length*Time*Status=_response_
    / noparm noresponse pred=freq;
    loglin Length Time Status Length*Time Time*Status Length*Status;
run;
```

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Length	1	2.64	0.1041
Time	1	5.25	0.0220
Status	1	48.94	<.0001
Length*Time	1	5.25	0.0220
Time*Status	1	95.01	<.0001

Maximum Likelihood Predicted Values for Frequencies

			Observed		Predicted		
Length	Time	Status	Frequency	Standard Error	Frequency	Standard Error	Residual
1	1	1	156	11.43022	161.0961	11.07379	-5.09614
1	1	2	84	8.754999	78.90386	7.808613	5.096139
1	2	1	84	8.754999	78.90386	7.808613	5.096139
1	2	2	156	11.43022	161.0961	11.07379	-5.09614
2	1	1	107	9.750588	101.9039	8.924304	5.096139
2	1	2	133	10.70392	138.0961	10.33434	-5.09614
2	2	1	31	5.47713	36.09614	4.826315	-5.09614
2	2	2	209	12.78667	203.9039	12.21285	5.09614

Time 과 Status 의 관계를 해석하기 위하여 잔차(O-E)를 정리하면…

[Length=1]

Time \ Status	1	2
1	-5.096	5.096
2	5.096	-5.096

시간이 1 일 경우 Status 가 2, 시간이 2 일 경우 Status 가 1 일 가능성이 높다.

[Length=2]

Time \ Status	1	2
1	5.096	-5.096
2	-5.096	5.096

시간이 1 일 경우 Status 가 1, 시간이 2 일 경우 Status 가 2 일 가능성이 높다.