제 5 장
분할표 데이터 분석

목차

❖ 범주형자료 분석의 개념

❖ 분할표의 통계분석

- 분할표의 확률구조
- 비교 측도
- 카이제곱 근사 검정
- Fisher의 정확 검정
- Simpson의 역설
- McNemar의 짝지어진 표본 검정

❖ 로지스틱 회귀분석

❖ 로그 선형 모형

범주형 자료 (Categorical Data) 분석의 개념

- 범주형 변수(categorical variable) – 범주(category)를 값으로 갖는 변수
  - 명목형 변수(nominal variable) – 순서 없는 범주를 가지는 변수 (예: 종교 – 불교, 기독교, 천주교 등)
  - 순서형 변수(ordinal variable) – 순서가 있는 범주를 가지는 변수 (예: 자동차의 크기– 소형, 중형, 대형)

- 연속형 변수를 몇 개의 그룹으로 만들어 범주형 변수로 만들기도 한다 (예: 연속형인 시험 점수(0~100점)를 몇 개의 등급으로 나눈 학점(A, B, C, D, F)으로 변환하여 범주형 변수로 만든 경우)

- 범주형 자료 분석이라 함은 흔히 반응(종속) 변수가 범주형이고 설명(독립) 변수는 범주형일 수도 있고 아닐 수도 있는 데이터를 분석하는 것을 의미 한다.

분할표의 기본 분석

분할표(Contingency Table)

- 분할표(contingency table) – 아래표와 같이 범주형 데이터가 각 변수의 값에 따라 통계표 형태로 정리되어 있을 때 그러한 통계표를 흔히 분할표라고 한다.
- 차원(dimensionality) – 분할표의 구성에 관계된 변수의 수
- 수준(level) – 범주형 변수가 가지는 범주의 수
- 분할표의 예 – 2x2분할표

		Y		
		1	2	
X	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

		Y		
		1	2	
X	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
		π_{+1}	π_{+2}	1

분할표(Contingency Table)

❖ IxJ 분할표에서 i번째 행, j번째 열의 칸(cell)에 빈도와 확률을 각각 n_{ij}와 π_{ij}라고 하자. 그리고 각 행의 합, 열의 합, 그리고 전체 합과 확률은 다음과 같다.

$$n_{i+} = \Sigma_j n_{ij}, \quad n_{+j} = \Sigma_i n_{ij}$$
$$\Sigma_i n_{i+} = \Sigma_j n_{+j} = \Sigma_i \Sigma_j n_{ij} = n$$

$$\pi_{i+} = \Sigma_j \pi_{ij}, \quad \pi_{+j} = \Sigma_i \pi_{ij},$$
$$\Sigma_i \pi_{i+} = \Sigma_j \pi_{+j} = \Sigma_i \Sigma_j \pi_{ij} = 1$$

	1	2	…	J	
1	n_{11}	n_{12}	…	n_{1J}	n_{1+}
2	n_{21}	n_{22}	…	n_{2J}	n_{2+}
I	n_{I1}	n_{I2}	…	n_{IJ}	n_{I+}
	n_{+1}	n_{+2}	…	n_{+J}	n

	1	2	…	J	
1	π_{11}	π_{12}	…	π_{1J}	π_{1+}
2	π_{21}	π_{22}	…	π_{2J}	π_{2+}
I	π_{I1}	π_{I2}	…	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	…	π_{+J}	1

분할표의 확률구조

연구대상이 되는 모집단으로부터 임의로 추출된 어떤 개체가 두 범주형 변수인 X와 Y값에 따라 분류된다고 하자.

- 결합확률(Joint probability)와 결합확률분포(Joint probability distribution)
  - 확률변수를 여러 개 함께 고려하는 확률 분포. 두 범주형 변수인 X와 Y에 대해 결합 확률은 $\pi_{ij} = \Pr(X = i, y = j)$. 표본에 의한 추정은 $p_{ij} = n_{ij}/n$.
- 주변확률(Marginal probability)와 주변확률분포(Marginal probability distribution)
  - 특정한 하나의 확률변수를 고려하는 확률 분포. 두 범주형 변수인 X와 Y에 대해 주변확률 확률은 $\pi_{i+} = \Pr(X = i)$ 와 $\pi_{+j} = \Pr(Y = j)$. 표본에 의한 추정은 $p_{i+} = n_{i+}/n$와 $p_{+j} = n_{+j}/n$.

분할표의 확률구조

- 조건부확률(Conditional probability)와 조건부확률분포(Conditional probability distribution)
  - 한 변수의 각 수준에서 다른 변수에 관해 개별적인 확률분포. 범주형 변수인 X가 주어졌을 때 Y의 조건부 확률은 $\pi_{j|i} = \Pr(y = j|X = i)$. 표본에 의한 추정은 $p_{j|i} = n_{ij}/n_{i+}$.

- 독립성 (Independence)
  - 한 변수의 각 수준에서 다른 변수의 조건부분포가 동일할 때 두 변수는 통계적으로 독립이라 한다. 두 변수 X와 Y가 독립이면 임의의 어떤 열(행)의 어떤 수준에 대한 확률은 각 행(열)에서 동일하다.

2차원 분할표의 독립성 또는 동질성 검정

❖ 독립성 (independence) 가설
- H_0: $\pi_{ij} = \pi_{i+}\,\pi_{+j}$ for all *i* and *j*

❖ **동질성 (homogeneity) 가설**
- H_0: $\pi_{1|j} = \pi_{2|j} = \cdots = \pi_{I|j}$ for all *j*

	1	2	…	J	
1	π_{11}	π_{12}	…	π_{1J}	π_{1+}
2	π_{21}	π_{22}	…	π_{2J}	π_{2+}
I	π_{I1}	π_{I2}	…	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	…	π_{+J}	1

❖ 독립성가설과 동질성가설의 카이제곱(Chi-square)에 의한 검정 방법은 동일

2차원 분할표의 독립성 또는 동질성 검정

❖ 피어슨 카이제곱(Pearson's Chi-square) 검정

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2_{((I-1)(J-1))}$$

- 귀무 가설 하에서 추정기대빈도는

$$\hat{\mu}_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n$$

- 유의수준 α가 주어졌을 때, $X^2 > \chi^2_{((I-1)(J-1))}(\alpha)$라면 귀무 가설을 기각한다. 즉, p-값이 α보다 작으면 귀무 가설을 기각한다.
- 귀무 가설을 기각하는 것은 행 변수와 열 변수가 독립이 아니다 또는 행(열) 변수의 각 수준의 확률이 동일하지 않다라는 것을 의미한다.

2차원 분할표의 독립성 또는 동질성 검정

❖ 우도비 카이제곱(Likelihood Ratio Chi-square) 검정

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(n_{ij}/\widehat{\mu_{ij}}\right) \sim \chi^2_{((I-1)(J-1))}$$

.

- 귀무가설 하에서 추정기대빈도는

$$\widehat{\mu_{ij}} = n\widehat{\pi_{ij}} = n\widehat{\pi_{i+}}\widehat{\pi_{+j}} = n_{i+}n_{+j}/n$$

- 유의수준 α가 주어졌을 때, $X^2 > \chi^2_{((I-1)(J-1))}(\alpha)$라면 귀무 가설을 기각한다. 즉, *p*-값이 α보다 작으면 귀무 가설을 기각한다.
- 귀무 가설을 기각하는 것은 행 변수와 열 변수가 독립이 아니다 또는 행(열) 변수의 각 수준의 확률이 동일하지 않다라는 것을 의미한다.

그룹 비교를 위한 측도

- 비율의 차이(Difference of Proportions):
  - $D = p_{1|1} - p_{1|2} = n_{11}/n_{1+} - n_{21}/n_{2+}$
  - 범위: $-1 \sim 1$
  - 동질 또는 독립: D = 0
- 상대적 위험도(Relative Risk):
  - $RR = p_{1|1}/p_{1|2} = (n_{11}/n_{1+})/(n_{21}/n_{2+})$
  - 범위: $0 \sim \infty$
  - 동질 또는 독립: RR = 1
- 오즈비(Odds Ratio)
  - $OR = (p_{11}/p_{12})/(p_{21}/p_{22})$
    $= (n_{11}n_{22})/(n_{12}n_{21})$
  - 범위: $0 \sim \infty$
  - 동질 또는 독립: OR = 1

		Y		
		1	2	
X	1	**n_{11}**	n_{12}	n_{1+}
	2	**n_{21}**	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

예제 1: 내세의 믿음과 성별의 연관성 (확률구조)

성별	내세의 믿음		
	예	아니오	합
여성	435	147	582
남성	375	134	509
합	810	281	1091

결합확률

성별	내세의 믿음		
	예	아니오	합
여성	40%	13%	53%
남성	34%	12%	47%
합	74%	26%	100%

주변확률

- ✓ 성별을 설명변수로 내세의 믿음에 대한 조건부분포
- ✓ 여성 중 “예”라고 반응한 비율: 435 / 582 = 0.747
- ✓ 여성의 경우 조건부분포 (0.747, 0.253)
- ✓ 남성의 경우 조건부분포 (0.737, 0.263)

예제 2: 아스피린과 심근경색의 연관성 (비교측도)

❖ 아스피린 복용이 심근경색 예방에 효과가 있는지 연구

	심근경색		합
	Yes	No	
위약(placebo)	189	10,845	11,034
아스피린(aspirin)	104	10,933	11,037
합	293	21,778	22,071

예제 2: 아스피린과 심근경색의 연관성 (계속)

OR=(10933x189)/(104x10845) =1.832
Col1: RR = 0.990577/0.982871 = 1.008
Col2: RR = 0.009423/0.017129 = 0.550

drug * heart 교차표

			heart		
			no	yes	전체
drug	aspirin	빈도	10933	104	11037
		기대빈도	10890.5	146.5	11037.0
		drug 중 %	99.1%	0.9%	100.0%
	placebo	빈도	10845	189	11034
		기대빈도	10887.5	146.5	11034.0
		drug 중 %	98.3%	1.7%	100.0%
전체		빈도	21778	293	22071
		기대빈도	21778.0	293.0	22071.0
		drug 중 %	98.7%	1.3%	100.0%

위험도 추정값

	값	95% 신뢰구간	
		하한	상한
drug (aspirin / placebo)에 대한 승산비	1.832	1.440	2.331
코호트 heart = no	1.008	1.005	1.011
코호트 heart = yes	.550	.434	.698
유효 케이스 수	22071		

카이제곱 검정

	값	자유도	점근 유의확률 (양측검정)	정확한 유의확률 (양측검정)	정확한 유의확률 (단측검정)
Pearson 카이제곱	25.014[b]	1	.000		
연속수정[a]	24.429	1	.000		
우도비	25.372	1	.000		
Fisher의 정확한 검정				.000	.000
유효 케이스 수	22071				

a. 2x2 표에 대해서만 계산됨

b. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 146.48입니다.

예제 2: 아스피린과 심근경색의 연관성 (in R)

```
> aspirin<-matrix(c(189,104,10845,10933),ncol=2,nrow=2)
> chisq.test(aspirin,correct=FALSE)#연속수정 안 한 경우
        Pearson's Chi-squared test

data:  aspirin
X-squared = 25.014, df = 1, p-value = 5.692e-07
```

```
> chisq.test(aspirin) # 연속수정 한 경우

        Pearson's Chi-squared test with Yates'
continuity correction

data:  aspirin
X-squared = 24.429, df = 1, p-value = 7.71e-07
```

예제 2: 아스피린과 심근경색의 연관성 (in R)

```
aspirin.test<- chisq.test(aspirin,correct=FALSE)


> aspirin.test$observed  #관측빈도
     [,1]  [,2]
[1,]  189 10845
[2,]  104 10933

> aspirin.test$expected # 기대빈도
         [,1]     [,2]
[1,] 146.4801 10887.52
[2,] 146.5199 10890.48

> aspirin.test$p.value # p-value
[1] 5.691897e-07
```

소표본일 때 Fisher 정확 검정

- 표본의 수가 작을 때 근사적 방법인 피어슨 카이제곱 검정과 우도비 카이제곱 검정은 타당하지 않을 수 있다. 이 때 정확한 분포를 이용한 검정방법이 선호된다.
- 표본 수 판정기준: 칸의 기대 빈도가 5이하인 셀이 20% 이상일때

- Fisher의 정확(Fisher's Exact) 검정

  - 초기하분포(Hypergeometric Distribution):

$$\Pr(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$$

		Y		
		1	2	
X	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

  - n_{11}의 범위는 $m_- \le n_{11} \le m_+$
  - $m_- = \max(0, n_{1+}+n_{+1}-n)$, $m_+ = \min(n_{1+}, n_{+1})$

소표본일 때 Fisher 정확 검정

- 독립성검정에 대한 P-값은 관측도수 이상으로 대립가설을 선호하게 되는 초기하확률의 합이다.
- 주변합계가 주어질 때 n_{11}이 클 수록 표본오즈비 $\hat{\theta} = n_{11}n_{22} / n_{12}n_{21}$가 커지게 되고 대립가설을 더욱 선호하게 되는 뚜렷한 증거를 나타낸다.
- 따라서, P-값은 초기하분포에서 n_{11}이 관측값보다 크거나 같게 우측 꼬리확률이 된다.
- 이 검정은 2x2표의 경우 영국 통계학자 피셔(Fisher, Ronald A., 1890-1962)에 의해 1934년에 제안되었으며 Fisher의 정확검정(exact test)이라 부른다.
- 독립성의 귀무가설에 대해 대립가설

예제 3: 간염 치료의 효과 (소표본)

❖ 간염 치료제가 효과가 있는지 연구

	효과		합
	호전됨	호전 안됨	
치료군(Test)	10	2	12
대조군(Control)	2	4	6
합	12	6	18

❖ 표본의 수가 작으므로 피어슨 카이제곱 검정(P-값=0.0339)이나 우도비 카이제곱 검정 (P-값=0.0346)은 타당하지 않을 수 있다.
(75%이상의 칸의 기대 빈도가 5이하)

❖ 기대빈도 – (1,1): 12x12/18=8, (1,2): 6x12/18=4
(2,1): 12x6/18=4, (2,2): 6x6/18=2

예제 3: 간염 치료의 효과 (계속)

❖ 초기하분포에 의한 확률

(1,1)	(1,2)	(2,1)	(2,2)	확률
12	0	0	6	0.0001
11	1	1	5	0.0039
10	**2**	**2**	**4**	**0.0533**
9	3	3	3	0.2370
8	4	4	2	0.4000
7	5	5	1	0.2560
6	6	6	0	0.0498

❖ Ex) 위의 0.0533의 값을 초기하분포를 이용하여 R에서 계산하시오.

```
> choose(12,10)*choose(6,2)/choose(18,12)
[1] 0.05332902
```

예제 3: 간염 치료의 효과 (계속)

- 단측검정 – H_0: 독립, H_1: 양의 연관성 있음

  P-값 = $Pr(n_{11} \geq 10) = Pr(n_{11}=10) + Pr(n_{11}=11) + Pr(n_{11}=12)$

  $= 0.0533 + 0.0039 + 0.0001 = 0.0573$
- 단측검정 – H_0: 독립, H_1: 음의 연관성 있음

  P-값 = $Pr(n_{11} \leq 10) = Pr(n_{11}=6) + \cdots + Pr(n_{11}=10)$

  $= 0.0498 + \cdots + 0.0533 = 0.9961$
- 양측검정 – H_0: 독립, H_1: 연관성 있음

  P-값 = $Pr(n_{11} \geq 10) + Pr(n_{11}=6)$

  $= 0.0533 + 0.0039 + 0.0001 + 0.0498 = 0.1071$

❖ 피어슨 또는 우도비 카이제곱 (양측) 검정과 다른 결론

- 피어슨 카이제곱 검정: P-값=0.0339
- 우도비 카이제곱 검정: P-값=0.0346

예제 3: 간염 치료의 효과 (in R)

단측검정 – H_0: 독립, H_1: 양의 연관성 있음

P-값 = $Pr(n_{11} \geq 10) = Pr(n_{11}=10) + Pr(n_{11}=11) + Pr(n_{11}=12)$
$= 0.0533 + 0.0039 + 0.0001 = 0.0573$

```
> liver.data<- matrix(c(10,2,2,4),nrow=2)   # 데이터 입력
> fisher.test(liver.data,alternative="greater")

        Fisher's Exact Test for Count Data

data:  liver.data
p-value = 0.05726
alternative hypothesis: true odds ratio is greater than
1
95 percent confidence interval:
 0.9374086       Inf
sample estimates:
odds ratio
  8.457238
```

예제 3: 간염 치료의 효과 (in R)

단측검정 – H0: 독립, H1: 음의 연관성 있음

P-값 = Pr(n11≤10) = Pr(n11=6) + … + Pr(n11=10)

= 0.0498 + … + 0.0533 = 0.9961

```
> fisher.test(liver.data,alternative="less")

        Fisher's Exact Test for Count Data

data:  liver.data
p-value = 0.9961
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
   0.0000 108.2447
sample estimates:
odds ratio
  8.457238
```

예제 3: 간염 치료의 효과 (in R)

양측검정 – H0: 독립, H1: 연관성 있음

P-값 = $Pr(n_{11} \geq 10) + Pr(n_{11}=6)$

= 0.0533 + 0.0039 + 0.0001 + 0.0498 = 0.1071

```
> fisher.test(liver.data)

        Fisher's Exact Test for Count Data

data:  liver.data
p-value = 0.107
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.6896384 166.4344614
sample estimates:
odds ratio
  8.457238
```

예제 4: 차 (소표본)

❖ 런던의 로담스테드 실험연구소에서 근무했던 Fisher의 한 동료는 차를 마실 때 연유와 차 중에서 어느 것을 먼저 컵에 넣었는지를 구분할 수 있다고 주장했다. 그러한 주장을 검정하기 위해 여덟 컵의 차를 맛보도록 하는 시험을 고안했다. 컵은 랜덤 한 순서로 제시되었다.

1차 첨가	추측		합
	연유	차	
연유	3	1	4
차	1	3	4
합	4	4	8

$H_0 : \theta = 1$ 맛 감정가의 추측과 컵에 무엇을 먼저 넣는가의 순서는 무관하다.

vs

$H_1 : \theta > 1$ 순서와 추측간에 양적 연관성이 있다는 감정가의 주장을 반영.

Simpson의 역설(Paradox)

- 주변연관성(marginal association)이 조건부연관성(conditional association)과 다른 방향을 가질 수 있다는 결과를 Simpson의 역설이라고 한다.

- 두 변수가 사실 음의 상관이 있는데 lurking confounder 때문에 양의 상관이 있는 것처럼 나타날 때 발생한다.

- 1951년 Edward H. Simpson에 의해 처음 Journal of the Royal Statistical Society (Series B) 학술지 논문에서 이 현상이 설명되었다. 이후에 여러 통계학자에 의해 언급되었고, 1972년 Colin R. Blyth에 의해 Simpson's paradox로 이름 불리게 되었다.

예제 5: UC-Berkeley의 입학허가 (Simpson의 역설)

- 1973년 미국의 University of California – Berkeley의 대학원 입학 허가에 관한 데이터

- 입학허가에 남녀 차별이 있었는가를 조사

성별	불합격자	합격자	지원자
남자	1291(48.0%)	1400(52.0%)	2691
여자	1063(58.9%)	772(42.1%)	1835

예제 5: UC-Berkeley의 입학허가 (Simpson의 역설)

분야 * 합격여부 * 성별 교차표

성별			합격여부				전체	
			합격		불합격			
			빈도	분야의 %	빈도	분야의 %	빈도	분야의 %
남학생	분야	A	512	62.1%	313	37.9%	825	100.0%
		B	353	63.0%	207	37.0%	560	100.0%
		C	120	36.9%	205	63.1%	325	100.0%
		D	138	33.1%	279	66.9%	417	100.0%
		E	53	27.7%	138	72.3%	191	100.0%
		F	224	60.1%	149	39.9%	373	100.0%
	전체		1400	52.0%	1291	48.0%	2691	100.0%
여학생	분야	A	89	82.4%	19	17.6%	108	100.0%
		B	17	68.0%	8	32.0%	25	100.0%
		C	202	34.1%	391	65.9%	593	100.0%
		D	131	34.9%	244	65.1%	375	100.0%
		E	94	23.9%	299	76.1%	393	100.0%
		F	239	70.1%	102	29.9%	341	100.0%
	전체		772	42.1%	1063	57.9%	1835	100.0%

예제 5: UC-Berkeley의 입학허가 (Simpson의 역설)

- 대학원 입학에서 남학생의 합격률(52%)가 여학생의 합격률(42%)보다 높음
- 분야별로 세분화 하였을 때 여학생의 합격률이 더 높음 (A,B,D,F)
- 남학생들은 주로 A,B 분야에 많은 수를 지원함 (여학생은 적게 지원함)
- A, B, F 부분에서 합격률이 비교적 높음
- 2원 분할표의 경우 전공선택에서의 남녀별 성향차이가 반영되지 않아 이를 반영한 3원 분할표와 상반된 결과가 제시됨
- 이러한 현상을 Simpson의 역설이라고 함

짝지어진 표본(Matched sample)일 때 분할표 분석
(Paired T-test 와 비슷한 테스트. 자료가 binary인 경우)

Test 1	Test 2		합
	Positive	Negative	
Positive	a	b	a+b
Negative	c	d	c+d
합	a+c	b+d	n

$$H_0: \pi_b = \pi_c \text{ VS } H_a: \pi_b \neq \pi_c$$

(또는, $H_0: \pi_a + \pi_b = \pi_a + \pi_c$ VS $H_a: Not\ H_0$)

(또는, $H_0: \pi_c + \pi_d = \pi_b + \pi_d$ VS $H_a: Not\ H_0$)

짝지어진 표본(Matched sample)일 때 분할표 분석

❖맥니머 검정 통계량 (McNemar's Test Statistics)

$$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi^2_{(df=1)}$$

Test 1	Test 2		합
	Positive	Negative	
Positive	a	b	a+b
Negative	c	d	c+d
합	a+c	b+d	n

❖맥니머 검정 통계량 (McNemar's Test Statistics)-연속수정

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \sim \chi^2_{(df=1)}$$

P-value 가 달라짐.

짝지어진 표본(Matched sample)일 때 분할표 분석

❖ 세금인상과 복지확대에 관련한 의견에 대해 “예”라고 응답할 확률비교

예제 6: 세금인상과 복지확대에 관련 의견 (McNemar 검정)

세금 인상	복지확대		합
	찬성	반대	
찬성	227	132	359
반대	107	678	785
합	334	810	1144

- 귀무가설은 H_0: $\pi_{1+} = \pi_{+1}$ 또는 H_0: $\pi_{12} = \pi_{21}$

- 동일한 사람에게 두 가지 이슈에 대한 의견을 물어 보았기 때문에 표본은 독립적이 아님.
- 대응쌍(matched pair)을 이루는 이항반응에 대한 주변동질성(marginal homogeneity) 검정

예제 6: 세금인상과 복지확대에 관련 의견 (in R)

```
> McNemar.data<-
matrix(c(227,107,132,678),nrow=2,
dimnames=list("Tax"=c("Yes","No"),"Welfare"=c
("Yes","No")))
```

```
> McNemar.data
     Welfare
Tax   Yes  No
  Yes 227 132
  No  107 678
```

예제 6: 세금인상과 복지확대에 관련 의견 (in R)

```
> mcnemar.test(McNemar.data,correct=FALSE) #연속수정 안함

        McNemar's Chi-squared test

data:  McNemar.data
McNemar's chi-squared = 2.6151, df = 1, p-value = 0.1059
```

```
> MC.test<-(132-107)^2/(132+107) # 검정통계량 계산

> MC.test  # McNemar's chi-squared 위의 결과.
[1] 2.615063

> 1-pchisq(MC.test,df=1)  # P-value 구하기
[1] 0.1058533
```

예제 6: 세금인상과 복지확대에 관련 의견 (in R)

```
> mcnemar.test(McNemar.data,correct=TRUE) #연속수정 함

        McNemar's Chi-squared test

data:  McNemar.data
McNemar's chi-squared = 2.41, df = 1, p-value = 0.1206
```

```
> MC.test.c<-(abs(132-107)-1)^2/(132+107)  # 검정통계량 계산

> MC.test.c  # McNemar's chi-squared 위의 결과.
[1] 2.410042

> 1-pchisq(MC.test.c,df=1)  # P-value 구하기
[1] 0.1205591
```

일반화 선형모형
(Generalized Linear Model; GLM)

일반화선형모형(Generalized Linear Model; GLM)

- 선형모형(linear model)에서 반응은 설명변량들의 선형결합에 정규분포를 따르는 오차가 붙여지는 것으로 가정된다.
- 따라서 반응변수(종속변수)는 연속형이어야 한다.
- 그런데 실제로는 드물지 않게 종속변수가 이산형인 경우가 있고 연속형이지만 비음(非陰, nonnegative)인 경우가 있다.
- 이런 경우들에서는 선형모형이 적절하지 않으므로 대안 모형이 필요한데, 일차적으로 고려할 수 있는 것이 일반화선형모형(generalized linear model)이다.

GLM의 성분

❖ 모든 GLM(Generalizes Linear Model)에 공통된 세 가지 요소

- 랜덤성분(Random component)
  : 반응변수 Y의 확률분포를 규정함.
- 체계적성분(Systematic component)
  : 모형의 예측변수로 사용되는 설명변수들을 규정함.
- 연결함수(Link function)
  : 체계적성분과 랜덤성분의 기대값과의 함수관계를 나타냄.

❖ GLM은 선형 예측방정식을 통해 반응 평균의 어떤 함수와 설명변수간의 관계를 나타냄.

GLM의 성분

❖ **랜덤성분**

GLM의 랜덤성분은 표본크기가 N일 때 독립적인 반응변수 Y에 대해 $Y_1, \cdots, Y_N$ 의 확률분포를 가정

관측값이 이항반응인 경우 → 이항분포
음이 아닌 빈도 수를 나타내는 경우 → 포아송분포
연속형 반응인 경우 → 정규분포 가정함.

❖ **체계적성분**

Y의 평균을 $\mu = E(Y)$라 할 때, μ의 값은 설명변수의 수준에 따라 변함
GLM의 체계적 성분은 설명변수 x_j 의 선형식

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

을 써서 모형의 우변에 표현하고, 이와 같은 일차결합을 선형예측이라 함

GLM의 성분

❖ 연결

선형예측식과 $\mu = E(Y)$ 의 관계를 규정

Y 가 정규분포 → 항등연결 $g(\mu) = \mu$

포아송분포 → log항등연결 $g(\mu) = \log(\mu)$

이항분포 → logit연결 $g(\mu) = \log[\mu/(1-\mu)]$

❖ 랜덤성분의 확률분포는 평균의 특별한 함수인 자연모수를 갖게 되는데,
정규분포의 경우 자연모수는 평균 자신이 되고
포아송분포의 경우에는 평균의 로그가 된다
또한, 이항분포의 경우에는 성공 확률의 로짓이 자연모수가 된다
연결함수 $g(\mu)$ 로 자연모수를 사용할 때 정준연결이라 부른다.

로지스틱 회귀모형

- 2개의 범주를 취하는 반응변수 Y를 공변량(covariate) X로 설명하기 위한 대표적인 모형
- 반응변수는 2개의 범주를 취하는 범주형 변수이고, 설명변수는 범주형 변수 또는 연속형 변수.
- 반응변수가 2개의 범주를 취하는데 일반적인 회귀모형을 적용할 경우 발생하는 문제점:
  - 일반적인 회귀모형: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$
  - 관측치 Y_i는 0 또는 1이지만, 예측치 $\hat{Y}_i$는 $-\infty$에서 $+\infty$까지의 연속형 값.
- 로짓변환을 취한 후 회귀모형 적용해 문제 해결. 그래서 로짓(logit) 회귀 모형이라고도 함.

로지스틱 회귀모형

- 반응변수의 2개의 범주를 0과 1로 표시
- Pr(Y=1|X=x): X가 x로 주어졌을 때 Y가 1일 확률
- 로지스틱 회귀 모형의 정의:

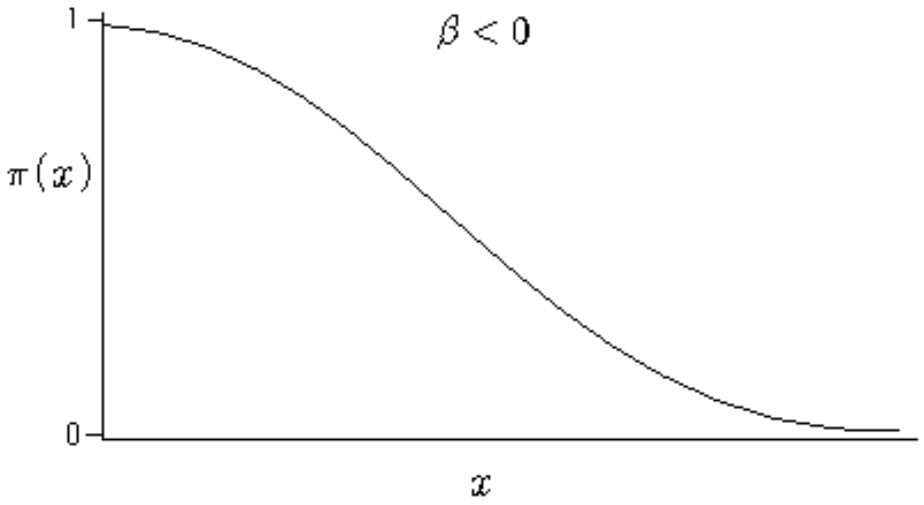
$$\log \frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = \beta_0 + \beta_1 x$$

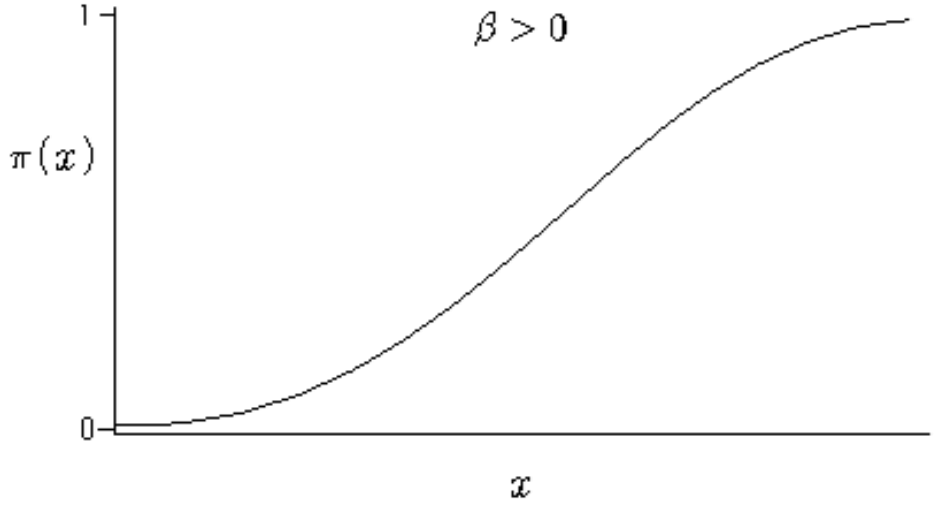
또는

$$\Pr(Y=1|X=x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}$$

- 해석:
  - β_1이 양수이면 X가 증가함에 따라 성공확률도 증가
  - β_1이 음수이면 X가 증가함에 따라 성공확률은 감소

로지스틱 곡선





- β의 크기는 곡선이 얼마나 빨리 증가 또는 감소하는지를 결정하며 $|\beta|$가 증가함에 따라 곡선은 더욱 가파른 변화를 보이며, 특히, $\beta = 0$ 일 때 곡선은 x 축에 평행인 직선이 된다.

로지스틱 회귀모형의 해석

- 이항반응 Y와 양적 설명변수 X에 대해 $\pi(x)$를 X=x 일 때 "성공"확률이라 할 경우, 확률 $\pi(x)$는 이항 분포의 모수.

- 로지스틱회귀모형은 $\pi(x)$의 로짓(logit)에 대해 선형식

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

- 로지스틱회귀 공식을 변형하여 다른 형태의 공식에 의해 성공 확률을 직접 나타낼 수도 있음.

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

로지스틱 회귀모형의 해석

- 곡선의 기울기가 가장 가파른 점은 $\pi(x)=0.5$가 되는 $x=-\alpha/\beta$일 때임.

$$\pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

- x값에 $-\alpha/\beta$ 를 대입하여 $\pi(x)$=0.5를 확인하거나 아래의 식에 $\pi(x)$=0.5를 대입한 후 이 식을 x 에 관해 풀면 됨.

$$logit[\pi(x)]=\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x$$

이 값을 중위수효과수준(Median Effective Level)이라 함.

즉, $x=EL_{50}$은 가능한 두 가지 반응의 확률이 각각 50%가 되는 수준을 말함.

로지스틱 회귀모형의 해석

❖ 오즈비의 해석

- 반응 1(즉, '성공')의 오즈는 $\mathrm{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$

$$\Rightarrow \quad \frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}$$

- x가 1단위 증가함에 따라 오즈는 e^{β} 배만큼씩 곱해져서 증가함을 알 수 있음.
- 즉, x 에서의 오즈에 e^{β} 를 곱하여 $x+1$의 오즈를 구함.
- $\beta = 0$ 일 때는 $e^{\beta} = 1$ 로 일정하므로 x의 변화와 관계가 없음.

적합도와 유의성

❖ 모형 적합도(goodness-of-fit):우도비검정(likelihood ratio test)

- H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs H_1: not H_0
- $-2\log(\text{우도비}) = -2[\log L_0 - \log L_1] \sim \chi^2_{(df)}$
- L_0와 L_1는 각각 H_0와 H_1의 가정 하에서 계산한 최대 우도
- 자유도 df=k
- P-값이 유의 수준 α보다 작으면 H_0기각. 즉, 현재 고려한 모형이 적절함.

❖ 모수의 유의성: Wald 검정

- H_0: $\beta_j = 0$ vs H_1: $\beta_j \neq 0$
- $(\hat{\beta}_j/SE(\hat{\beta}_j))^2 \sim \chi^2_{(1)}$
- 또는 $\hat{\beta}_j/SE(\hat{\beta}_j) \sim N(0,1)$
- P-값이 유의 수준 α보다 작으면 H_0기각. 즉, j번째 모수의 추정치는 유의함.

예제 1: 광고 효과 데이터

❖ 광고 노출 수가 구매 여부에 영향을 주는지 연구
(데이터 AD.txt)

광고노출 수	구매자 수	비구매자 수	합
0	3	7	10
1	5	5	10
2	4	6	10
3	7	3	10
4	8	2	10

- 반응 변수 Y는 2개의 범주(구매 또는 비구매)만 가지는 범주형
- 구매를 1, 비구매를 0로 표시
- 설명 변수 X는 광고 노출 수로 연속형

예제 1: 광고 효과 데이터

```
> AD.data<- read.table("AD.txt",header=T)
> attach(AD.data)  # 데이터 변수 사용가능.
> AD.data[1:10,]
   x y
1  0 1
2  0 1
3  0 1
4  0 0
5  0 0
6  0 0
7  0 0
8  0 0
9  0 0
10 0 0
> dim(AD.data)
[1] 50  2
```

1. 설명 변수 x는 광고 노출 수.(0~4) 연속형
2. 종속 변수 y는 구매는 1, 비구매는 0.

예제 1: 광고 효과 데이터 (계속)


```
> model.glm<-glm(y~x,family="binomial")# GLM 모형 적합하기.
> summary(model.glm)  # 모형의 전체적인 요약.
Call: glm(formula = y ~ x, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7196  -1.0378   0.7194   0.8977   1.5560
Coefficients:
            Estimate  Std.Error  z value  Pr(>|z|)
(Intercept)  -0.8568     0.5249  -1.632    0.1026
x             0.5192     0.2237   2.321    0.0203 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 62.989  on 48  degrees of freedom
AIC: 66.989

Number of Fisher Scoring iterations: 4
```

예제 1: 광고 효과 데이터 (계속)

```
Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 62.989  on 48  degrees of freedom
AIC: 66.989
```

위의 결과 값을 이용해서 모델의 적합도를 테스트 한다.

테스트하기 위한 통계량은

$-2\log(\text{우도비}) = -2[\log L_0 - \log L_1] \sim \chi^2_{(df)}$

```
> chisq.value<- (model.glm$null.deviance-model.glm$deviance)
# 위의 값이 −2log(우도비) = −2[log L0−log L1] 을 계산

> 1-pchisq(chisq.value,df=49-48) # p-value #모델이 유용함.
[1] 0.01426615
```

예제 1: 광고 효과 데이터 (계속)

```
> model.glm$coefficients  # 회귀계수
(Intercept)           x
 -0.8568081   0.5191543

> model.glm$fitted[1:5] # Y의 적합값. (Fitted probability)
1         2         3         4         5
0.2980067 0.2980067 0.2980067 0.2980067 0.2980067

> model.glm$fitted[45:50]
       45        46        47        48        49        50
0.7720299 0.7720299 0.7720299 0.7720299 0.7720299 0.7720299

> length(model.glm$fitted)
[1] 50
```

Y의 적합값: E(Y) =P(Y|X=x)

$$= \hat{\pi}(x) = \frac{exp(\hat{\alpha}+\hat{\beta}x)}{1+exp(\hat{\alpha}+\hat{\beta}x)}$$

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

예제 1: 광고 효과 데이터 (계속)

```
> length(which(round(model.glm$fitted)-y==0))/length(y)
[1] 0.62
```

Y의 적합값과 실제값의 비교: $\hat{\pi}(x)$을 0과 1로 변환(0.5 기준)
62% 정도의 일치율을 보임.

```
> table(y)
y
 0  1
23 27
> y.hat<-round(model.glm$fitted)
> table(y.hat)
y.hat
 0  1
20 30
> table(y.hat,y)
      y
y.hat  0  1
    0 12  8
    1 11 19
```

Confusion Matrix 가 자주 이용됨.

로지스틱 회귀모형에 관한 추론

❖ 효과에 대한 신뢰구간

- 로지스틱회귀모형 $\mathrm{logit}[\pi(x)] = \alpha + \beta x$ 에서 모수 에 대한 β 표본 신뢰구간은 $\hat{\beta} \pm z_{\alpha/2}(ASE)$

- 구간 양 끝점을 지수 변환하면 오즈의 증가배율인 e^{β} 의 신뢰구간을 얻게 됨.

```
> summary(model.glm)$coef  # 모형요약 중 회귀계수 부분만 추출.
               Estimate Std. Error   z value   Pr(>|z|)
(Intercept) -0.8568081  0.5249445 -1.632188 0.10263987
x            0.5191543  0.2236932  2.320832 0.02029593
```

```
> summary(model.glm)$coef[2,1]-1.96*summary(model.glm)$coef[2,2]
[1] 0.08071559
> summary(model.glm)$coef[2,1]+1.96*summary(model.glm)$coef[2,2]
[1] 0.957593
```

β 의 근사적인 95% 신뢰구간 = **(0.0807, 0.9756)**

로지스틱 회귀모형에 관한 추론

x가 1단위 증가함에 따라 오즈는 e^{β} 배만큼씩 곱해져서 증가함을 알 수 있음.

$$\text{Odds} = \frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}$$

- 또한, 광고노출이 1회 증가할 때 구매하게 될 오즈에 대한 95% 신뢰구간은 $(e^{0.0807}, e^{0.9756}) = (1.084, 2.653)$ 이 됨.
- 광고노출이 1회 증가할 때 구매하게 될 오즈는 최소 8.1%에서 최대 약 2.65배까지 증가한다고 추론할 수 있음.

다중 로지스틱 회귀모형

❖ 설명변수가 2개 이상일 때 다중 로지스틱 회귀 모형:

$$\Pr(Y=1|X_1=x_1,\ldots,X_k=x_k)=\frac{\exp(\beta_0+\beta_1x_1+\ldots+\beta_kx_k)}{1+\exp(\beta_0+\beta_1x_1+\ldots+\beta_kx_k)}$$

또는

$$\log\frac{\Pr(Y=1|X_1=x_1,\ldots,X_k=x_k)}{\Pr(Y=0|X_1=x_1,\ldots,X_k=x_k)}=\beta_0+\beta_1x_1+\ldots+\beta_kx_k$$

❖ 변수선택 방법은 일반적인 회귀분석 때와 동일

- 전진 선택법 (Forward Selection)
- 후진 소거법 (Backward Elimination)
- 단계 선택법 (Stepwise Selection)

예제 2: 잡지 구독 데이터

❖ 새 잡지의 구독 성향 조사(파일 Readers.txt 참조)

- 40명의 응답자로부터 성별(Gender: 0=Male, 1=Female), 나이(Age), 사회적 친화력(Socio), 정치적 성향(Polit), 새 잡지의 구독(Subs: 1=구독함, 0=구독하지 않음)를 측정
- 변수 Socio와 Polit는 10점 만점 척도로 측정 – 연속형으로 고려
- 반응변수가 구독의사(subs)일 때 나머지 변수들을 설명변수로 하는 로지스틱 회귀모형을 고려하여 여러 설명변수들이 구독의사에 미치는 영향력을 비교 분석

예제 2: 잡지 구독 데이터


```
> readers.data<- read.table("Readers.txt",header=T)
> attach(readers.data)
> model.glm2<- glm(subs~gender+age+socio+polit,family="binomial")
> summary(model.glm2)

Call:
glm(formula = subs ~ gender + age + socio + polit, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.02063    2.70336  -0.747   0.4548
gender여성  -1.46206    0.84778  -1.725   0.0846 .
age          0.13820    0.05922   2.334   0.0196 *
socio       -0.66273    0.31065  -2.133   0.0329 *
polit        0.07509    0.15926   0.471   0.6373
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.452  on 39  degrees of freedom
Residual deviance: 39.880  on 35  degrees of freedom
AIC: 49.88

Number of Fisher Scoring iterations: 5
```

예측력(Predictive Power)의 요약

❖ 유의한 k개의 설명변수로 적합된 로지스틱 회귀 모형을 이용하여 i 번째 객체가 1을 취할 예측 확률:

$$\hat{\pi}_i = \widehat{\Pr}(Y_i = 1 | X_{1i} = x_{1i}, \ldots, X_{ki} = x_{ki}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki})}$$

- $\hat{\pi}_i > \pi_0$ 라면 $\hat{Y}_i = 1$
- $\hat{\pi}_i \leq \pi_0$ 라면 $\hat{Y}_i = 0$
- π_0 는 흔히 0.5
- 민감도(Sensitivity) = $\Pr(\hat{Y}_i = 1 | Y_i = 1)$
- 특이도(Specificity) = $\Pr(\hat{Y}_i = 0 | Y_i = 0)$

❖ 분류표(Classification Table):

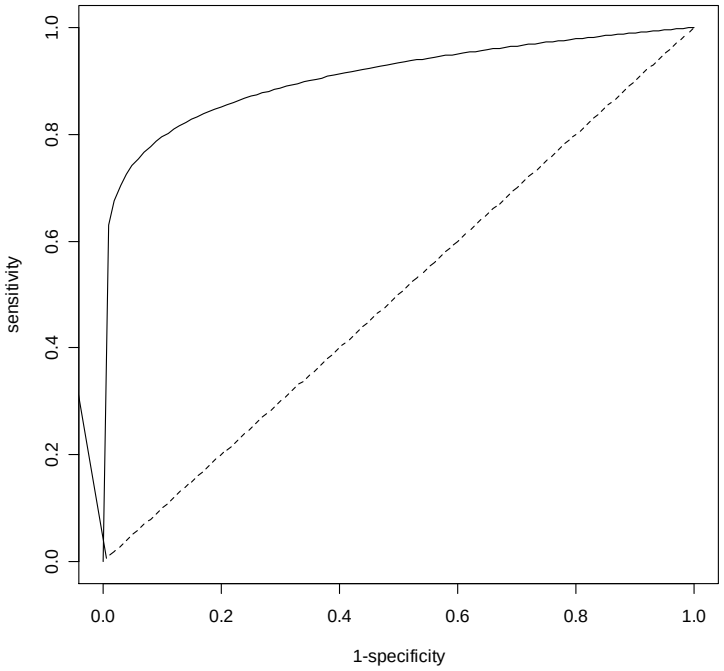
		$\hat{Y}$	
		1	0
Y	1		
	0		

예측력(Predictive Power)의 요약

❖ ROC(Receiver Operating Characteristic) 곡선:

- 모든 가능한 π_0에 대해 민감도와 (1-특이도)의 그림
- 곡선 아래 면적(AUC; Area Under the Curve)이 클수록 더 좋은 예측력 가짐
- 랜덤하게 예측한다면 45도 각도의 대각선

민감도



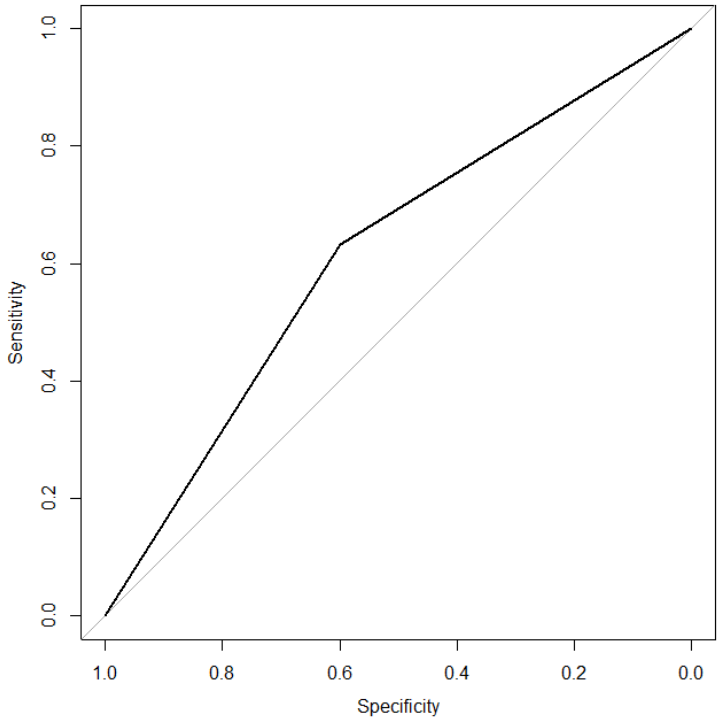
1-특이도

예제 1: 광고 효과 데이터

```
> library(pROC)
> roc(y.hat,y,plot=TRUE)  # 적합값과 관측값으로 ROC 그리기

Call:
roc.default(response = y.hat, predictor = y, plot = TRUE)

Data: y in 20 controls (y.hat 0) < 30 cases (y.hat 1).
Area under the curve: 0.6167   #AUC
```



로그-선형 모형
(Log-linear Model)

로그-선형 모형

❖ **다차원 분할표**(multi-dimensional contingency table)는 여러 복잡한 구조를 가질 수 있기 때문에 구조적 특성의 탐색을 위해서는 카이제곱 이상의 체계적인 통계적 방법이 필요하다.

❖ 로그-선형 모형이 바로 그런 통계적 분석을 위한 모형이다.

❖ 범주형 변수들 사이의 연관성이나 교호작용을 기술하고 분석하기 위한 모형이다.

❖ 로그-선형 모형에서는 모든 변수가 반응변수로 취급한다.

2차원 분할표 분석을 위한 로그-선형 모형

❖ 두 범주형 변수 X와 Y간에 서로 독립일 때 각 칸의 기대빈도 μ_{ij}는 다음과 같이 표현될 수 있다.

$$\mu_{ij} = n\,\pi_{i+}\,\pi_{+j}$$

❖ 독립모형 (Independence Model):

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

- λ_i^X : 변수 X의 i번째 수준 효과
- λ_j^Y : 변수 Y의 j번째 수준 효과

❖ 독립모형의 해석 (Ix2분할표의 경우):

$$\text{logit}[Pr(Y=1|X=i)]$$
$$= \log[Pr(Y=1|X=i)/Pr(Y=2|X=i)]$$
$$= \log(\mu_{i1}/\mu_{i2}) = \log(\mu_{i1}) - \log(\mu_{i2})$$
$$= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y)$$
$$= \lambda_1^Y - \lambda_2^Y = \beta_0$$

로그-선형 모형의 모수 추정 (2 x 2 분할표 경우)

❖ 모형의 가정 (다항분포)

$$P(Y) = \frac{n!}{y_{11}!\, y_{12}!\, y_{21}!\, y_{22}!} \pi_{11}^{y_{11}} \pi_{12}^{y_{12}} \pi_{21}^{y_{21}} \pi_{22}^{y_{22}}$$

모수의 추정: 우도 (또는 가능도)를 최대화 하는 추정법 (최대우도법, Maximum Likelihood Estimation)

-로그우도 함수를 최대화하는 모수의 값을 찾는다.

(라그랑지 승수법을 이용하여 우도를 각각의 모수에 대해 편미분하여 0으로 놓고 식을 푼다.)

로그-선형 모형의 모수 추정 (2 x 2 분할표 경우)

❖ 로그 우도 함수: (우도함수는 모수에 대한 함수)

$$\log L = \log \prod_{i=1}^{2} \prod_{j=1}^{2} P(y_{ij})$$

$$= \log \frac{n!}{y_{11}!\, y_{12}!\, y_{21}!\, y_{22}!} \pi_{11}^{y_{11}} \pi_{12}^{y_{12}} \pi_{21}^{y_{21}} \pi_{22}^{y_{22}}$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} y_{ij} \log \pi_{ij} + c + \lambda(1 - \pi_{11} - \pi_{12} - \pi_{21} - \pi_{22})$$

❖ 로그 우도 함수를 최대로 하는 값을 찾으면

$$\hat{\pi}_{ij} = \frac{y_{ij}}{n}$$

3차원 분할표 분석을 위한 로그-선형 모형

❖ 범주형 변수 X,Y,Z에 대한 로그-선형 모형은 변수들간의 연관성과 독립성에 기초하여 여러 모형을 정의할 수 있다.

- 독립모형: (X,Y,Z)

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

- 주변독립모형: (XY,Z)

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

- 조건부독립모형: (XY,XZ)

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

- 상호결합적 모형: (XY,XZ,YZ)

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

- 포화모형: (XYZ)

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

❖ 여러 모형을 비교하여 가장 적합한 모형을 선택 후, 연관성 해석

모형 선택

❖ 모형의 적합도 (Goodness-of-fit) 검정
- H_0: 현재적합모형 vs H_1: 포화모형
- Pearson카이제곱 $X^2 \sim \chi^2(df)$ 또는 우도비 카이제곱 $G^2 \sim \chi^2(df)$
- 자유도 df는 관측치의 수-현재적합모형의 모수 수
- p-값이 유의 수준 α 보다 작으면 현재 적합 모형이 좋지 않음

❖ 모형 M_0와 M_1비교 (M_1모형이 M_0포함할 경우)
- H_0: M_0 모형 vs H_1: M_1 모형
- $G^2(M_0|M_1) = G^2(M_0) - G^2(M_1) \sim \chi^2(df)$
- 자유도 df 는 두 모형의 모수 수 차이
- p-값이 유의 수준 α 보다 작으면 M_1모형이 더 좋음

❖ Akaike Information Criterion (AIC)을 최소화하는 모형 선택
- AIC = -2(최대 로그 우도- 모수의 수)
- 유사하게, G^2-2(자유도)를 최소화하는 모형 선택

예제 1: 술,담배,마리화나 사용간의 연관성

❖ 미국 고교생 (2276명)의 술(Alcohol), 담배(Cigarette), 마리화나(Marijuana) 사용간의 연관성 연구

(파일 Cigar.txt 참조)

술 (Alcohol)	담배 (Cigarette)	마리화나(Marijuana) Yes	No
Yes	Yes	911	538
Yes	No	44	456
No	Yes	3	43
No	No	2	279

```
> cigar.data<- read.table("Cigar.txt",header=TRUE)
> attach(cigar.data)
```

예제 1: 술,담배,마리화나 사용간의 연관성

아래의 명령어로 원래 데이터를 분석에 적합한 형태로 변환시켜준다.

```
> cigar.data2 <- array(data = c(911,3,538,43,44,2,456,279),
+                      dim = c(2,2,2),
+                      dimnames = list("Alco" = c("yes","no"),
+                                      "Mari" = c("yes","no"),
+                                      "Ciga" = c("yes","no")))
```

```
> cigar.data #변환 전
  Alco Ciga Mari Count
1    1    1    1   911
2    2    1    1     3
3    1    1    2   538
4    2    1    2    43
5    1    2    1    44
6    2    2    1     2
7    1    2    2   456
8    2    2    2   279
```

```
> cigar.data2 #변환 후
, , Ciga = yes

     Mari
Alco  yes  no
  yes 911 538
  no    3  43

, , Ciga = no

     Mari
Alco  yes  no
  yes  44 456
  no    2 279
```

예제 1: 술,담배,마리화나 사용간의 연관성

데이터의 탐색적 분석 1. (Exploratory Data Analysis)

```
> addmargins(cigar.data2)
, , Ciga = yes
      Mari
Alco  yes  no  Sum
  yes 911 538 1449
  no    3  43   46
  Sum 914 581 1495

, , Ciga = no
      Mari
Alco  yes  no Sum
  yes  44 456 500
  no    2 279 281
  Sum  46 735 781

, , Ciga = Sum
      Mari
Alco  yes   no  Sum
  yes 955  994 1949
  no    5  322  327
  Sum 960 1316 2276
```

예제 1: 술,담배,마리화나 사용간의 연관성

데이터의 탐색적 분석 **2. (Exploratory Data Analysis)**

```
> prop.table(cigar.data2)
, , Ciga = yes

     Mari
Alco          yes         no
  yes 0.400263620 0.23637961
  no  0.001318102 0.01889279

, , Ciga = no

     Mari
Alco          yes         no
  yes 0.0193321617 0.2003515
  no  0.0008787346 0.1225835
```

로그 선형 모형을 적합하기 위한 **R function**

1. loglin
2. glm

loglin 을 이용하여 여러 모형들의 적합을 시도

```
> loglin(cigar.data2,margin=list(1,2,3)) # 1. 독립모형
2 iterations: deviation 0
$lrt   # 우도비 카이제곱 G^2
[1] 1286.02

$pearson # Pearson 카이제곱 X^2
[1] 1411.386

$df  # 적합한 모형의 자유도 (관측치 수준의 개수 - 모수의 수 = 8-4)
[1] 4

$margin  # 적합한 모형의 구체적인 부분.여기서는 독립모형임.
$margin[[1]]
[1] "Alco"

$margin[[2]]
[1] "Mari"

$margin[[3]]
[1] "Ciga"
```

loglin 을 이용하여 여러 모형들의 적합을 시도

```
# 독립모형
> model1<-loglin(cigar.data2,margin=list(1,2,3))

# 주변독립모형, 3개
> model2<-loglin(cigar.data2,margin=list(1,c(2,3)))
> model3<-loglin(cigar.data2,margin=list(2,c(1,3)))
> model4<-loglin(cigar.data2,margin=list(3,c(1,2)))

# 조건부독립모형, 3개
> model5<-loglin(cigar.data2,margin=list(c(1,2),c(1,3)))
> model6<-loglin(cigar.data2,margin=list(c(2,3),c(1,2)))
> model7<-loglin(cigar.data2,margin=list(c(1,3),c(2,3)))

# 상호결합적모형
>model8<-loglin(cigar.data2,margin=list(c(1,2),c(1,3),c(2,3)))

# 포화모형
> model9<-loglin(cigar.data2,margin=list(c(1,2,3)))
```

위의 9가지 모형을 비교하여 가장 적절한 모형을 찾아냄.

시도된 모형들이 적합한가 검정(Testing)

```
# 조건부독립모형이 적절한가에 대한 검정의 예.
# 1. Pearson 카이제곱 검정
> model5$pearson # Pearson 카이제곱 값
[1] 497.3693
> model5$df  # 조건부 독립모형의 자유도.
[1] 2
> 1-pchisq(model5$pearson,df=model5$df) # 귀무가설 하에서 p-value
[1] 0
# 2. 우도비 카이제곱 검정
> 1-pchisq(model5$lrt,df=model5$df) # 귀무가설 하에서 p-value
[1] 0
```

모형의 적합도 (Goodness-of-fit) 검정

H_0: 현재적합모형 vs H_1: 포화모형

Pearson카이제곱 $X^2 \sim \chi^2(df)$ 또는 우도비 카이제곱 $G^2 \sim \chi^2(df)$

자유도 df는 관측치의 수-현재적합모형의 모수 수

p-값이 유의 수준 α 보다 작으면 현재 적합 모형이 좋지 않음

모형	Pearson X^2	우도비 G^2	자유도	G^2의 P-value	G^2-2(자유도)
M1 (A, C, M)	1411.4	1286.0	4	<0.001	1278.0
M2 (A, CM)	505.6	534.2	3	<0.001	528.2
M3 (M, AC)	704.9	843.8	3	<0.001	837.8
M4 (C, AM)	824.2	939.6	3	<0.001	933.6
M5 (AC, AM)	443.8	497.4	2	<0.001	493.4
M6 (AM, CM)	177.6	187.8	2	<0.001	183.8
M7 (AC, CM)	80.8	92.0	2	<0.001	88.0
M8(AC, AM, CM)	0.374	0.401	1	0.541	-1.6
M9 (ACM)	0.0	0.0	0	-	-

- (AC, AM, CM) 모형이 p-값이 0.05보다 크므로 가장 적합하다고 결론

```
> 1-pchisq(model8$lrt,df=model8$df) # 귀무가설 하에서 p-value
[1] 0.5408395
> 1-pchisq(model8$pearson,df=model8$df)
[1] 0.5265517
```

두 모형의 비교.(한 모형이 다른 모형을 포함할때)

모형	Pearson X^2	우도비 G^2	자유도	G^2의 P-값	G^2-2(자유도)
M6 (AM, CM)	177.6	187.8	2	<0.001	183.8
M8(AC, AM, CM)	0.4	0.4	1	0.54	-1.6

❖ 모형 M_0와 M_1비교 (M_1모형이 M_0포함할 경우)

- H_0: M_0 모형 vs H_1: M_1 모형
- $G^2(M_0|M_1) = G^2(M_0) - G^2(M_1) \sim \chi^2(df)$
- 자유도 df 는 두 모형의 모수 수 차이
- p-값이 유의 수준 α 보다 작으면 M_1모형이 더 좋음

```
> 1-pchisq(model6$lrt-model8$lrt, df=model6$df-model8$df)
[1] 0
```

- **(AM,CM)**와 (AC,AM,CM)비교:

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

= 187.8-0.4 = 187.4, 자유도 df = 2-1, p-값 < 0.001

그러므로 (AC,AM,CM)모형이 **(AM,CM)모형보다 더 좋음**

예제 1: 술,담배,마리화나 사용간의 연관성(계속)

❖ 로그-선형 모형 적합 후의 오즈비

모형	조건부 연관성		
	AC	AM	CM
(A,C,M)	1.0	1.0	1.0
(AM,CM)	1.0	61.9	25.1
(AC,AM,CM)	7.8	19.8	17.3
(ACM) 수준1	13.8	24.3	17.5
(ACM) 수준2	7.7	13.5	9.7

- 오즈비 계산 예: 7.8 = (910.4x1.4)/(44.6x3.6) : M=Yes
  = (538.6x279.6)/(455.4x42.4) : M=No
- 모형 선택에 따라 연관성의 해석이 달라짐. 예를 들어, 현재 데이터에 가장 적합한 모형인 (AC, AM, CM)으로는 AC의 오즈비는 7.8로 연관성이 높지만, (AM, CM)모형으로는 1.0으로 독립.
- 그래서 모형 선택이 중요함.