09 강

데이터분석방법론2

로지스틱회귀모형 (4)

통계·데이터과학과 이기재 교수

학습목차

1 제 9강. 로지스틱회귀모형(4)

1 명목형 반응변수들에 대한 기준범주 로짓 모형

2 순서형 반응변수들에 대한 누적 로짓 모형

3 대응쌍 자료의 주변동질성

학습개요 및 목표

이번 강의는 반응변수가 다범주인 로지스틱회귀모형에 대해 공부합니다. 명목형 변수와 순서형 변수인 경우로 구분하여 모형 적용 방법을 살펴봅니다.

1. 명목형 반응변수에 대한 기준범주 로짓모형을 설명할 수 있다.
2. 순서형 반응변수에 대한 누적 로짓모형을 설명할 수 있다.
3. 대응쌍 자료에 대해서 주변동질성을 검정할 수 있다.

제 9강. 로지스틱회귀모형(4)

1. 명목형 반응변수들에 대한 기준범주 로짓 모형
2. 순서형 반응변수들에 대한 누적 로짓 모형
3. 대응쌍 자료의 주변동질성

01

제 9강. 로지스틱회귀모형(4)

명목형 반응변수에 대한 로짓모형

명목형 반응의 로짓모형

다범주 로짓모형

- 명목형 반응변수 Y가 범주 $1, 2, 3, \cdots, c$를 갖는 경우($c>2$)
- 각 범주에 대응하는 반응확률 : $\{\pi_1, \pi_2, \cdots \pi_c\}, \sum_{j}^{c} \pi_j = 1$

- n명의 관측치를 c개 범주에 할당시키는 표본모형
  ➔ 다항분포(Multinomial Distribution)를 따름

1. 기준범주를 이용한 로짓모형

기준범주 로짓

- $\pi_j = (Y = j), \;\; j = 1, 2, \cdots, c$
- 임의로 하나의 기준범주(Baseline-category)를 선택한 후 이 범주와 나머지 각 반응범주와 짝을 지어 로짓을 정의함
- 기준범주 로짓(마지막 범주 c가 기준일 때)
  - $\log\left(\frac{\pi_j}{\pi_c}\right), \; j = 1, 2, \; \cdots, \; c-1$
  - 예측변수 x를 가진 기준범주 로짓모형

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x, \;\; j = 1, 2, \;\; \cdots, \; c-1$$

"각 로짓에 대해서 서로 다른 모수 (α_j, β_j) 가정"

1. 기준범주를 이용한 로짓모형

Note

① $\exp(\hat{\beta}_j)$는 반응범주 c에 대한 반응범주 j의 오즈에 예측변수 x가 1단위 증가함으로써 나타나는 승법효과임

② 임의의 범주 a와 b에 대하여

$$\log\left(\frac{\pi_a}{\pi_b}\right) = \log\left(\frac{\pi_a/\pi_c}{\pi_b/\pi_c}\right) = \log\left(\frac{\pi_a}{\pi_c}\right) - \log\left(\frac{\pi_a}{\pi_c}\right)$$
$$= (\alpha_a + \beta_a x) - (\alpha_b + \beta_b x)$$
$$= (\alpha_a - \alpha_b) + (\beta_b - \beta_a)x$$

③ 순서형 반응변수의 경우에도
"기준범주 로짓모형"을 적용할 수 있음
: 이 경우는 순서에 대한 정보를
무시하고 있기 때문에 정보의 손실이 있을 수 있음

2. 예제: 악어의 먹이 선택

미국 플로리다 주의 59마리 악어의 길이(미터)와 주요 먹이

1.24 I	1.30 I	1.30 I	1.32 F	1.32 F	1.40 F	1.42 I	1.42 F
1.45 I	1.45 O	1.47 I	1.47 F	1.50 I	1.52 I	1.55 I	1.60 I
1.63 I	1.65 O	1.65 I	1.65 F	1.65F	1.68 F	1.70 I	1.73 O
1.78 I	1.78 I	1.78 O	1.80 I	1.80 F	1.85 F	1.88 I	1.93 I
1.98 I	2.03 F	2.03 F	2.16 F	2.26 F	2.31 F	2.31 F	2.36 F
2.36 F	2.39 F	2.41 F	2.44 F	2.46 F	2.56 O	2.67 F	2.72 I
2.79 F	2.84 F	3.25 O	3.28 O	3.33 F	3.56 F	3.58 F	3.66 F
3.68 O	3.71 F	3.89 F					

- F=어류(Fish)
- O=기타(Other)
- I=연체류(Invertebrates)

[출처] M.F. Delany and Clint T. Moore

2. 예제: 악어의 먹이 선택

R 프로그램

```
-------------------------------------------------------------------------------
> Gators <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Alligators.dat",
+                       header=TRUE)
> Gators
     x y
1  1.24 I
2  1.30 I
...
59 3.89 F
> library(VGAM) # package for multivariate GLMs, such as multinomial models
> fit <- vglm(y ~ x, family=multinomial, data=Gators) # vglm = vector GLM
> coef(fit, matrix = TRUE)
            log(mu[,1]/mu[,3])  log(mu[,2]/mu[,3])
(Intercept)            1.6177              5.6974
x                     -0.1101             -2.4654
> summary(fit)
              Estimate Std. Error  z value  Pr(>|z|)
(Intercept):1   1.6177     1.3073    1.237   0.21591
(Intercept):2   5.6974     1.7937    3.176   0.00149
x:1            -0.1101     0.5171   -0.213   0.83137 # for log[P(Y=1)/P(Y=3)]
x:2            -2.4654     0.8996   -2.741   0.00613 # for log[P(Y=2)/P(Y=3)]
---
Residual deviance: 98.3412 on 114 degrees of freedom
Reference group is level 3 of the response # reference = baseline category
-------------------------------------------------------------------------------
```

2. 예제: 악어의 먹이 선택

기준범주 로짓모형에 대한 추정 결과

모수	로짓에 대한 먹이 선택 범주	
	(어류/기타)	(연체류/기타)
절편	1.618	5.697
길이	-0.110(0.517)	-2.465(0.900)

2. 예제: 악어의 먹이 선택

- $Y=$ "주요 먹이", $x=$ "악어의 길이", c=3인 경우
  - $\log(\widehat{\pi}_1/\widehat{\pi}_3)=1.618-0.110x$
  - $\log(\widehat{\pi}_2/\widehat{\pi}_3)=5.697-2.465x$
  - $\log(\widehat{\pi}_1/\widehat{\pi}_2)=(1.618-5.697)+[-0.110-(-2.465)]x$
    $=-4.08+2.355x$
  - 큰 악어일수록 "연체류(2)"보다 "어류(1)"를 선호하는 경향이 있음

- 길이가 x 미터인 악어에 비해 길이가 x+1미터인 악어의 주요 먹이는 "연체류" 가 아님
- "어류" 일 오즈의 추정값은 $\exp(2.355)=10.5$ 배임

2. 예제: 악어의 먹이 선택

```
-----------------------------------------------------------------------------------
> fit2 <- vglm(y ~ x, family=multinomial(refLevel="I"), data=Gators)
> summary(fit2) # now using y=2 (I = Invertebrates) as baseline category
              Estimate Std. Error  z value  Pr(>|z|)
(Intercept):1  -4.0797    1.4686   -2.778   0.00547
(Intercept):2  -5.6974    1.7937   -3.176   0.00149
x:1             2.3553    0.8032    2.932   0.00336 # for log[P(Y=1)/P(Y=2)]
x:2             2.4654    0.8996    2.741   0.00613 # for log[P(Y=3)/P(Y=2)]
---
Residual deviance: 98.3412 on 114 degrees of freedom # same for any baseline
Reference group is level 2 of the response
> confint(fit2, method="profile") # profile likelihood confidence intervals
                 2.5 %     97.5 %
x:1            1.01118    4.19907  # beta for log[P(Y=1)/P(Y=2)]
x:2            0.87752    4.46361  # beta for log[P(Y=3)/P(Y=2)]
-----------------------------------------------------------------------------------
```

모수	로짓에 대한 먹이 선택 범주	
	(어류/연체류)	(기타/연체류)
절편	-4.080	-5.697
길이	2.355(0.803)	2.465(0.900)

2. 예제: 악어의 먹이 선택

SAS 프로그램

```
DATA gator;
INPUT length choice $ @@;
CARDS;
1.24 I  1.30 I  1.30 I  1.32 F  1.32 F  1.40 F  1.42 I  1.42 F
1.45 I  1.45 O  1.47 I  1.47 F  1.50 I  1.52 I  1.55 I  1.60 I
1.63 I  1.65 O  1.65 I  1.65 F  1.65 F  1.68 F  1.70 I  1.73 O
1.78 I  1.78 I  1.78 O  1.80 I  1.80 F  1.85 F  1.88 I  1.93 I
1.98 I  2.03 F  2.03 F  2.16 F  2.26 F  2.31 F  2.31 F  2.36 F
2.36 F  2.39 F  2.41 F  2.44 F  2.46 F  2.56 O  2.67 F  2.72 I
2.79 F  2.84 F  3.25 O  3.28 O  3.33 F  3.56 F  3.58 F  3.66 F
3.68 O  3.71 F  3.89 F
;
PROC LOGISTIC;
   MODEL choice = length  / link=glogit aggregate  scale=none;
RUN;
```

2. 예제: 악어의 먹이 선택

분석 결과

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.8006	2	0.0002
Score	12.5702	2	0.0019
Wald	8.9360	2	0.0115

Analysis of Maximum Likelihood Estimates

Parameter	choice	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	F	1	1.6177	1.3073	1.5314	0.2159
Intercept	I	1	5.6974	1.7938	10.0881	0.0015
length	F	1	-0.1101	0.5171	0.0453	0.8314
length	I	1	-2.4654	0.8997	7.5101	0.0061

Odds Ratio Estimates

Effect	choice	Point Estimate	95% Wald Confidence Limits	
length	F	0.896	0.325	2.468
length	I	0.085	0.015	0.496

3. 반응확률의 추정

- $\log\left(\dfrac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x,\ j = 1,\ 2,\ \cdots,\ c-1$
  - $\dfrac{\pi_j}{\pi_c} = e^{\alpha_j + \beta_j x},\ \left(\sum_{j=1}^{c} \pi_j = 1\right)$
  - $\pi_j = \dfrac{e^{\alpha_j + \beta_j x}}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{c-1} + \beta_{c-1} x}},\ j = 1, 2, \cdots, c-1$

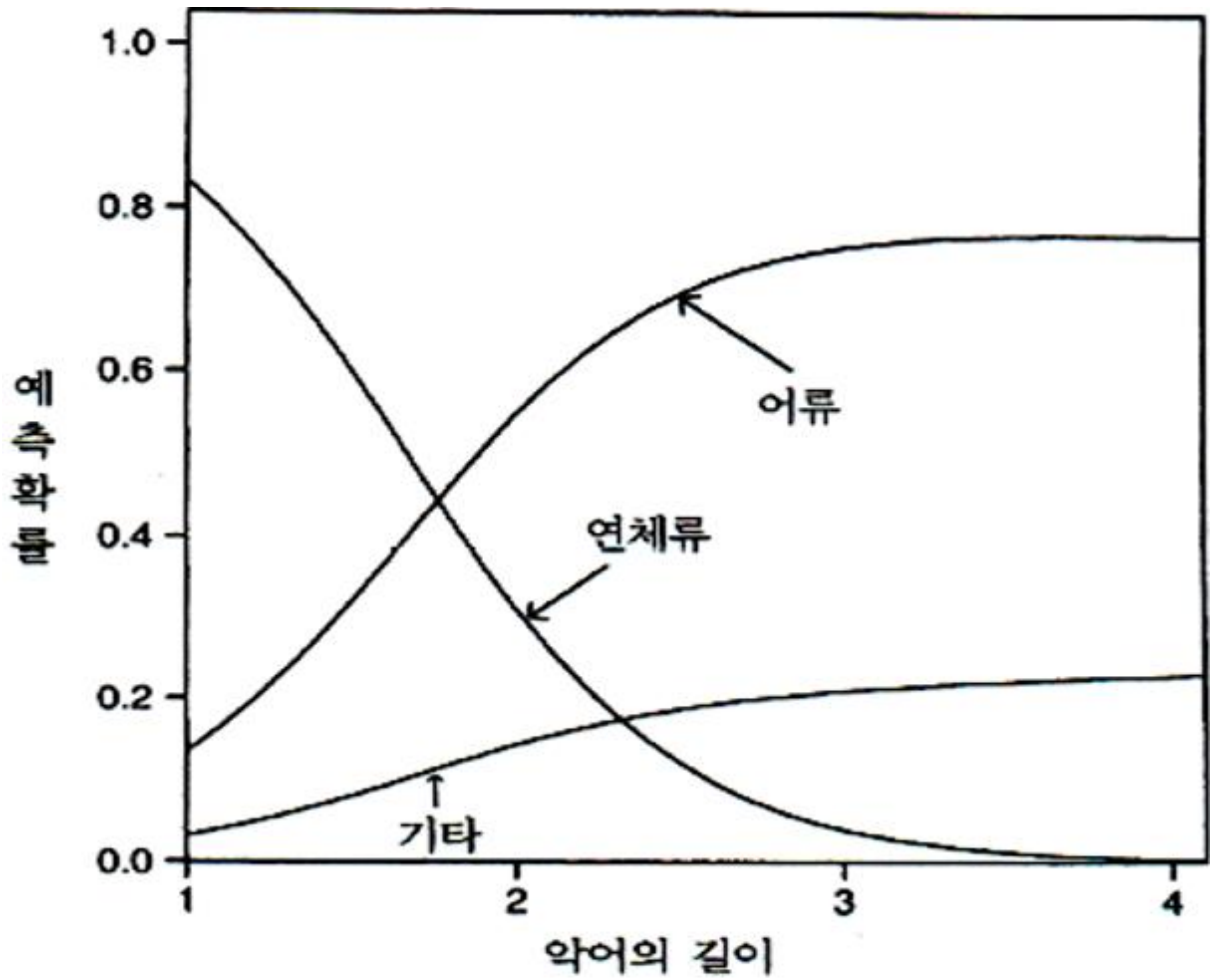
    $\pi_c = \dfrac{1}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{c-1} + \beta_{c-1} x}}$

3. 반응확률의 추정

- $\hat{\pi}_1 = \dfrac{e^{1.62-0.11x}}{1+e^{1.62-0.11x}+e^{5.70-2.47x}}$
- $\hat{\pi}_2 = \dfrac{e^{5.70-2.47x}}{1+e^{1.62-0.11x}+e^{5.70-2.47x}}$
- $\hat{\pi}_3 = \dfrac{1}{1+e^{1.62-0.11x}+e^{5.70-2.47x}}$

3. 반응확률의 추정

악어의 주요 먹이 선택에 대한 예측확률



4. 예제 : 사후 세계에 관한 연구

▪ 사후 세계에 관한 연구

인종	성별	사후 세계에 대한 믿음		
		믿는다	잘 모르겠다	믿지 않는다
백인	여성	371	49	74
	남성	250	45	71
흑인	여성	64	9	15
	남성	25	5	13

▪ $$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2$$

"아니오" 범주를 기준으로 설정한 경우
"사후세계에 대한 믿음에 대해 성별과 인종 간에 교호작용 효과는 없다"고 가정함

4. 예제 : 사후 세계에 관한 연구

- $$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2$$

```
-------------------------------------------------------------------------
> Afterlife <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Afterlife.dat", header=TRUE)
> Afterlife
   race gender yes undecided no
1 white female 371        49 74
2 white   male 250        45 71
3 black female  64         9 15
4 black   male  25         5 13
> library(VGAM)
> fit <- vglm(cbind(yes,undecided,no) ~ gender + race, family=multinomial,
+             data=Afterlife)
> summary(fit)
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1      1.3016      0.2265    5.747   9.1e-09
(Intercept):2     -0.6529      0.3405   -1.918    0.0551 .
gendermale:1      -0.4186      0.1713   -2.444    0.0145
gendermale:2      -0.1051      0.2465   -0.426    0.6700
racewhite:1        0.3418      0.2370    1.442    0.1493
racewhite:2        0.2710      0.3541    0.765    0.4442
---
Residual deviance: 0.8539 on 2 degrees of freedom
-------------------------------------------------------------------------
```

"이탈도 통계량=0.854, df=2
➔ 이 모형은 자료를 잘 적합함

4. 예제 : 사후 세계에 관한 연구

- $$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2$$

```
--------------------------------------------------------------------------------
> fit.race <- vglm(cbind(yes,undecided,no) ~ race, family=multinomial,
+                      data=Afterlife) # removing gender from model
> deviance(fit.race)
[1] 8.04650
> lrtest(fit, fit.race) # lrtest function available in VGAM package
Likelihood ratio test
Model 1: cbind(yes, undecided, no) ~ gender + race
Model 2: cbind(yes, undecided, no) ~ race
  #Df   LogLik  Df  Chisq   Pr(>Chisq)
1   2  -19.732
2   4  -23.329   2 7.1926     0.02742 # deviance diff. = 2(log-lik. diff.)
--------------------------------------------------------------------------------
```

H_0: $\beta_1^G = \beta_2^G = 0$ "사후세계 믿음에 대해 성별 주효과는 없다"는 가설에 대한 검정

5. 이산형 선택모형

Note

- 다범주 로짓모형은 시장조사에서 여러 선택의 가능성 중 특정 상품이 얼마나 선택되는지에 대한 분석을 할 때 중요한 도구로 사용되고 있음

[자동차 브랜드 선호도]

- 연간 수입
- 가족 수
- 교육 정도
- 거주지 (도심 또는 교외)

- 일반화된 모형으로 설명변수가 Y 범주에 따라 다른 값을 가질 수 있는 경우(자동차 브랜드별 옵션 가격 등)에 적용할 때 이산형 선택모형(discrete choice model)이라고 함

02

제 9장. 로지스틱회귀모형(4)

순서형 반응변수에 대한 로짓모형

순서형 반응변수에 대한 로짓 모형

- 반응범주들이 순서형인 경우는 순서를 고려한 로짓을 정의할 수 있음

➔ 순서를 고려한 로짓 모형은 해석이 간단하고 보통의 다범주 로짓 모형보다 더 좋은 검정력을 갖게 됨

- 누적확률(cumulative probability)

$$P(Y \le j) = \pi_1 + \cdots + \pi_j, \quad j = 1, 2, \ldots, c$$

순서형 반응변수에 대한 로짓 모형

- 누적 로짓(cumulative logit)

$$logit[P(Y \le j)] = \log\left[\frac{P(Y \le j)}{1 - P(Y \le j)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right]$$

$$j = 1, 2, \ ..., c-1$$

- $c = 3$ 인 경우의 누적 로짓

$$logit[P(Y \le 1)] = \log[\pi_1 / (\pi_2 + \pi_3)]$$

$$logit[P(Y \le 2)] = \log[(\pi_1 + \pi_2) / \pi_3)]$$

1. 비례오즈 누적 로짓 모형

- 예측변수 X 에 대하여

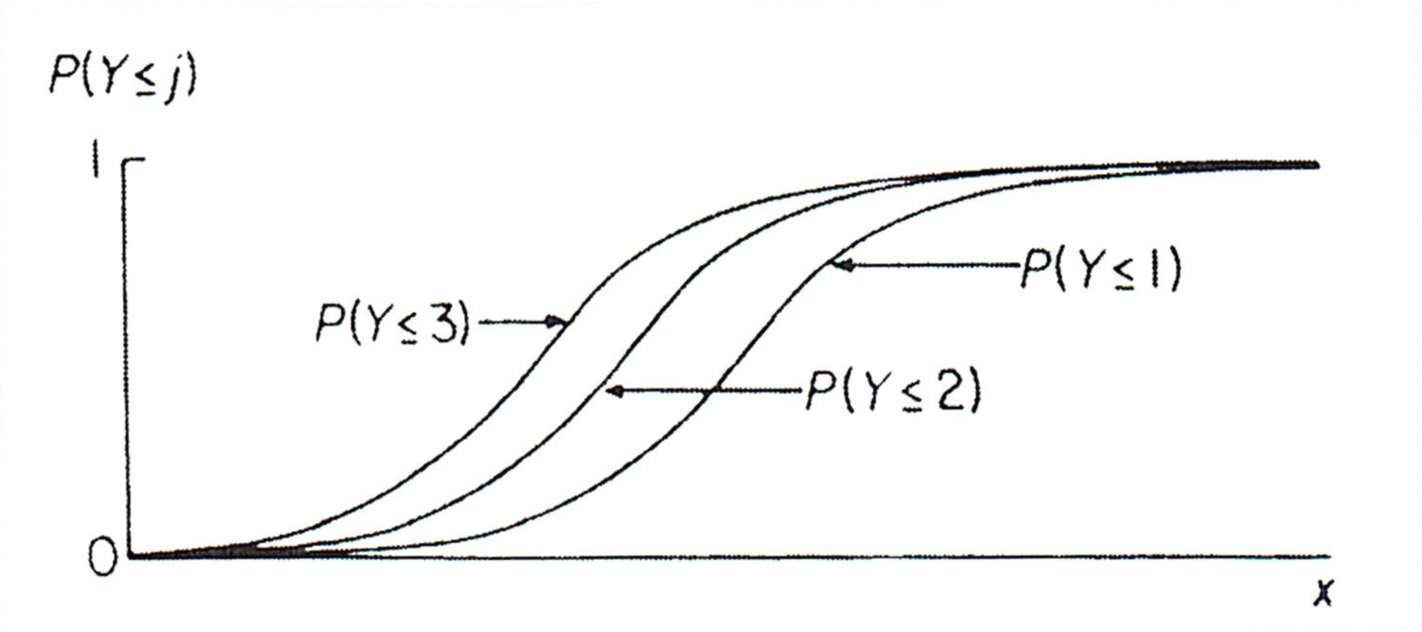
$$\text{logit}[P(Y \le j)] = \alpha_j + \beta x, \quad j = 1, 2, \cdots, c-1$$

"$1, \cdots, j$ 의 범주들을 하나의 범주로 합하고, $j+1$ 부터 c 까지의 범주를 다른 하나 범주로 보는 로지스틱 회귀모형과 비슷함"

- 각각의 누적 로짓에 대해서 서로 다른 절편 α_j 와 같은 기울기 β 를 가정함

1. 비례오즈 누적 로짓 모형

- 4개의 반응범주와 하나의 연속형 예측변수 x 에 대한 비례오즈 모형



"각각의 j 에 대해 β 의 효과가 같다는 것은 세 개의 곡선이 같은 모양을 갖는다는 의미임"

1. 비례오즈 누적 로짓 모형

- X 가 a과 b일 때 오즈비

$$\frac{P(Y \le j|X=a)/P(Y>j|X=a)}{P(Y \le j|X=b)/P(Y>j|X=b)}$$

➔ 오즈비의 로그

$$\log\left(\frac{P(Y \le j|X=a)}{P(Y>j|X=a)}\right) - \log\left(\frac{P(Y \le j|X=b)}{P(Y>j|X=b)}\right)$$

$$= \mathrm{log}it\,[P(Y \le j|X=a)] - \mathrm{log}it\,[P(Y \le j|X=b)]$$

$$= \beta(a-b), \quad j = 1, \cdots, c-1$$

1. 비례오즈 누적 로짓 모형

- $$\frac{odds\ of\ (Y \le j)\ at\ a}{odds\ of\ (Y \le j)\ at\ b} = e^{\beta(a-b)}$$

➔ 어떤 주어진 범주 이하의 반응에 대한 오즈는 x가 한 단위 증가하면 e^{β}배 만큼 증가함

➔ 비례오즈 모형(proportional odds model)

- $\beta = 0$ 만족 ⇔ X 와 Y 는 통계적으로 독립

2. 예제 : 정치성향과 가입정당의 관련성

- 개인의 정치성향과 가입정당의 관련성 자료

성별	정당	정치성향				
		매우 진보적	약간 진보적	중간	약간 보수적	매우 보수적
여성	민주당	25	105	86	28	4
	공화당	0	5	15	83	32
남성	민주당	20	73	43	20	3
	공화당	0	1	14	72	32

- 정치성향 5점 척도
  (1. 매우 진보적, 2. 약간 진보적, 3. 중간, 4. 약간 보수적, 5. 매우 보수적)
- 가입정당 ($x=1$ (공화당), $x=0$ (민주당)), 성별(남성=1, 여성=0)

2. 예제 : 정치성향과 가입정당의 관련성

R 프로그램

```
--------------------------------------------------------------------------------
> Polviews <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Polviews.dat",
+                          header=TRUE)
> Polviews # grouped data; ungrouped data file at website is Polviews2.dat
  gender party y1  y2 y3 y4 y5
1 female   dem 25 105 86 28  4
2 female repub  0   5 15 83 32
3   male   dem 20  73 43 20  3
4   male repub  0   1 14 72 32
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ party + gender,
+             family=cumulative(parallel=TRUE), data=Polviews)
> summary(fit) # "parallel=TRUE" imposes proportional odds structure
               Estimate  Std. Error   z value   Pr(>|z|)
(Intercept):1  -2.12233    0.16875   -12.577    <2e-16 # 4 intercepts for
(Intercept):2   0.16892    0.11481     1.471     0.141 # 5 y categories
(Intercept):3   1.85716    0.15103    12.297    <2e-16
(Intercept):4   4.65005    0.23496    19.791    <2e-16
partyrepub     -3.63366    0.21785   -16.680    <2e-16 # same effects
gendermale      0.04731    0.14955     0.316     0.752 # for all 4 logits
---
Residual deviance: 9.8072 on 10 degrees of freedom
--------------------------------------------------------------------------------
```

$$\text{logit}[P(Y \le j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2$$

비례오즈 누적로짓 모형에서 절편은 관심모수가 아님.

2. 예제 : 정치성향과 가입정당의 관련성

분석 결과 해석

① 비례오즈 로짓 모형에서 β_1에 대한 ML 추정

$\hat{\beta}_1 = -3.634\,(SE = 0.218)$

➔ 고정된 j에 대하여 진보적 방향으로 응답할 오즈는 민주당원에 비해 공화당원일 때 $\exp(-3.634) = 0.0026$배임

➔ 개인의 가입정당과 정치성향은 강한 관련성이 있으며 민주당원이 공화당원에 비해 더 진보적인 성향을 띠고 있음

2. 예제 : 정치성향과 가입정당의 관련성

분석 결과 해석

```
-----------------------------------------------------------------------
> attach(Polviews)
> data.frame(gender, party, fitted(fit)) # y1 = very lib., y5 = very conserv.
  gender  party     y1     y2     y3     y4     y5
1 female    dem 0.1069 0.4352 0.3228 0.1256 0.0095
2 female  repub 0.0031 0.0272 0.1144 0.5895 0.2657
3   male    dem 0.1115 0.4423 0.3165 0.1206 0.0090
4   male  repub 0.0033 0.0284 0.1189 0.5927 0.2566
-----------------------------------------------------------------------
```

2. 예제 : 정치성향과 가입정당의 관련성

모형의 모수에 대한 추론

```
---------------------------------------------------------------------------
> fit2 <- vglm(cbind(y1,y2,y3,y4,y5) ~ gender, # removing party effect
+               family=cumulative(parallel=TRUE), data=Polviews)
> lrtest(fit, fit2)
Likelihood ratio test

 Model 1: cbind(y1, y2, y3, y4, y5) ~ party + gender
 Model 2: cbind(y1, y2, y3, y4, y5) ~ gender
   #Df    LogLik  Df   Chisq  Pr(>Chisq)
 1  10   -35.203
 2  11  -236.827   1  403.25   < 2.2e-16
 > confint(fit, method="profile")
                     2.5 %     97.5 %
 partyrepub     -4.07164  -3.21786  # profile likelihood CI's for
 gendermale     -0.24639   0.34140  # beta_1 and beta_2 in full model
---------------------------------------------------------------------------
```

- 가입정당과 정치성향은 강한 연관성이 존재한다는 강한 근거

3. 순서형 분석의 검정력 증가

- 순서형 변수들에 대한 분할표에서 독립성 검정을 할 때 순서형 검정이 모든 변수들을 명목형 변수로 처리하는 카이제곱검정에 비해 더 적절하고 큰 검정력을 갖음
- Y의 순서정보를 이용하는 누적 로짓모형이 Y를 명목형 변수로 간주하여 분석하는 기준범주 로짓모형보다 더 큰 검정력을 갖음
- 적합도가 다소 떨어지더라도 좀더 간단한 모형이 효과의 대부분을 설명할 수 있다면 간단한 모형을 사용하는 것이 바람직함

4. 예제 : 총 가구 수입과 행복도

흑인에 대한 행복도와 총 가구 수입 자료(괄호 안 도수는 백인에 대한 자료)

총 가구 수입	행복도		
	행복하지 않음	좀 행복함	아주 행복함
평균 이하	37 (128)	90 (324)	45 (107)
평균	25 (66)	93 (479)	56 (295)
평균 이상	6 (35)	18 (247)	13 (184)

4. 예제 : 총 가구 수입과 행복도

- Y = 행복도
  (1 = 행복하지 않음, 2 = 좀 행복함, 3 = 아주 행복함)

  x = 총 가구수입(양적 변수로 처리)

  (1 = 평균 이하, 2 = 평균, 3 = 평균 이상)

- 비례오즈 모형

$$\text{logit}[P(Y \le j)] = \alpha_j + \beta x, \quad j = 1, 2$$

4. 예제 : 총 가구 수입과 행복도

```
------------------------------------------------------------------------------
> Happy <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Happy.dat",
+                      header=TRUE)
> Happy # data for sampled black Americans
  income y1 y2 y3
1      1 37 90 45
2      2 25 93 56
3      3  6 18 13
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3)~ income, family=cumulative(parallel=TRUE),
+              data=Happy)
       Estimate  Std. Error  z value  Pr(>|z|)  # not showing the two
income  -0.2668     0.1510   -1.768    0.0771  # intercept estimates
---
> fit0 <- vglm(cbind(y1,y2,y3)~ 1, family=cumulative, data=Happy) # null model
> lrtest(fit, fit0)
Model 1: cbind(y1, y2, y3) ~ income # treating happiness and income as ordinal
Model 2: cbind(y1, y2, y3) ~ 1
  #Df   LogLik  Df  Chisq  Pr(>Chisq)
1   3  -14.566
2   4  -16.121   1  3.109     0.07786 .
------------------------------------------------------------------------------
```

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta x, \;\; j = 1, 2$$

$\hat{\beta} = -0.267$이므로 총 수입이 증가할 수록 "행복하지 않음" 범주가 나올 경향은 점점 감소하는 것을 알 수 있음.

4. 예제 : 총 가구 수입과 행복도

```
--------------------------------------------------------------------------
> fit2 <- vglm(cbind(y1,y2,y3) ~ factor(income), family=multinomial,data=Happy)
> fit0 <- vglm(cbind(y1,y2,y3)~ 1, family=multinomial, data=Happy)
> # baseline cat. logit null model equivalent to cumulative logit null model
> lrtest(fit2, fit0)
Model 1: cbind(y1, y2, y3) ~ factor(income) # treats variables as nominal-scale
Model 2: cbind(y1, y2, y3) ~ 1
  #Df   LogLik  Df   Chisq  Pr(>Chisq)
1   0  -14.058                          # fit2 model is saturated
2   4  -16.121   4  4.1258      0.3892
--------------------------------------------------------------------------
```

$$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2, \quad j = 1, 2.$$

$$H_0 : \beta_{j1} = \beta_{j2} = 0, \quad j = 1, 2$$

- 가구수입에 대한 선형 효과와 같은 가정을 하지 않은 모형임.
- 귀무가설 하에서 더 많은 모수를 사용하기 때문에 검정력이 높지 않음.

5. 잠재변수 선형모형과 누적연결함수 관계

- 누적 로짓 모형에서 비례 오즈 형태를 가정하면 하나의 예측변수 효과는 $c-1$개의 누적 로짓 모형 식에서 모두 동일하게 됨
- 비례 오즈 구조는 단순 잠재변수모형에 의해서 자동적으로 만들어짐
- Y^*를 연속형 잠재변수라고 하고, Y^*의 절단점을 $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ 라고 함.

$$Y = j \quad if \ \ \alpha_{j-1} < Y^* \leq \alpha_j$$

- 잠재변수 Y^*의 평균이 설명변수와 관련된 회귀모형을 만족

$$Y^* = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

5. 잠재변수 선형모형과 누적연결함수 관계

- Y^*를 연속형 잠재변수라고 할 때

$$Y = j \quad if \;\; \alpha_{j-1} < Y^* \leq \alpha_j$$

- $P(Y \leq j) = P(Y^* \leq \alpha_j) = P(\epsilon \leq \alpha_j - \alpha - \beta_1 x_1 - \cdots - \beta_p x_p)$
- $\text{link}[P(Y \leq j)] = \alpha_j - \beta_1 x_1 - \cdots - \beta_p x_p$

- 오차항 ϵ 이 정규분포이면 이에 대응되는 연결함수는 프로빗(probit) 함수
- 오차항 ϵ 이 로지스틱 분포(정규분포처럼 종모양, 대칭이나 꼬리부분이 더 두터움)를 따르면 이에 대응되는 연결함수는 로짓(logit) 함수가 되고 잠재변수 모형은 누적 로짓 모형이 됨

6. 반응범주 선택에 대한 불변성

- 잠재변수 선형모형과 누적 로짓 모형에서 비례 오즈 형태를 가정한다면 Y의 범주를 어떻게 선택하든지 상관없이 효과에 대한 모수는 변하지 않음
- 정치성향을 측정한 연속형 변수에 대해서 이산형 변수로 바꿀 때 (진보적, 중간, 보수적) 범주로 나누든지 또는 (매우 개방적, 약간 개방적, 약간 보수적, 보수적, 매우 보수적) 범주로 나누든지 상관없이 모수의 효과는 동일하게 됨

03

제 9장. 로지스틱회귀모형(4)

대응쌍 자료의 주변동질성

대응쌍 자료의 주변동질성

- 동일한 대상에 대해서 두 번의 조사를 한 경우나 한 표본의 개체와 다른 표본의 개체간에 자연스러운 짝 관계(pairing, 쌍)가 있는 경우에 만들어진 대응쌍 자료의 분석방법
- 대응쌍이 흔히 발생하는 경우는 각 개체에 대해서 반복적으로 관측하는 경우로, 예를 들어 경시적(longitudinal) 연구에서 동일한 대상을 시간의 흐름에 따라 반복적으로 관측하는 경우임
- 주변동질성 검정의 문제를 다룸
- 교재 8장 1절 내용

1. 사례 ①: 환경개선과 관련한 일반사회조사 사례

- 1144명을 대상으로 환경개선을 위해서 (1) 더 높은 세금을 지불할 의향이 있는지 (2) 생활수준 긴축을 받아들일 의향이 있는지 응답하도록 함
- 조사 결과

더 많은 세금 지불	생활수준의 긴축		합계
	찬성	반대	
찬성	227	132	359
반대	107	678	785
합계	334	810	1144

  - 특성상 두 설문항목에 대한 응답결과는 서로 종속되어 있음(표본 오즈비=10.9)
  - 더 많은 세금 지불에 "찬성"한 비율 = 359/1144 = 0.314
  - 생활수준의 긴축에 "찬성"한 비율 = 334/1144 = 0.292

1. 사례 ①: 환경개선과 관련한 일반사회조사 사례

- 주요 관심사

$\pi_{1+} = \pi_{+1}$ ⇔ 주변동질성(marginal homogeneity)

$$\left[\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}\right]$$

$$\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21}$$

2. 사례 ②: 신약효과 실험

- 86명의 실험 대상자에 대해서 각각 랜덤하게 (신약 → 가짜약) 또는 (가짜약→신약) 순으로 복용하게 한 후 그 효과를 조사함

- 실험결과

		가짜약		계
		S	F	
신약	S	12	49	61
	F	10	15	25
계		22	64	86

2. 사례 ②: 신약효과 실험

- 확률분포

	S	F	
S	π_{11}	π_{21}	π_{+1}
F	π_{12}	π_{22}	π_{+2}
	π_{1+}	π_{2+}	1.0

- 주요 관심사

$\pi_{1+} - \pi_{+1}$ 에 대한 추론 ⇔ 주변동질성 만족 여부

3. 맥니마 검정 (McNemar test)

- H_0 : 주변동질성 만족 $\left(\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21}\right)$

$$\Leftrightarrow \quad \frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{1}{2}$$

➔ "대응쌍 자료에서 귀무가설을 만족하는 경우는 n_{12} 와 n_{21} 은 같은 기대도수를 갖게 됨"

3. 맥니마 검정 (McNemar test)

- $n^* = n_{12} + n_{21}$ 으로 정의하면 귀무가설이 성립할 때

$$n_{12} \sim B\left(n^*, \frac{1}{2}\right),$$

$$E(n_{12}) = n^*/2,$$

$$\sqrt{Var(n_{12})} = \sqrt{n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}$$

3. 맥니마 검정 (McNemar test)

- 검정통계량

$$Z = \frac{n_{12} - n^*/2}{\sqrt{n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} \sim N(0,1)$$

$$= \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

또는

$$Z^2 = \frac{(n_{12} - n_{21})}{n_{12} + n_{21}} \sim \chi_1^2$$

"맥니마 검정"

3. 맥니마 검정 (McNemar test)

- 맥니마 검정 결과

① 환경개선 관련 일반사회조사

$$z = \frac{132-107}{\sqrt{132+107}} = 1.617$$

➔ $Z^2 = 2.6151$, df=1, p-값 = 0.1059

3. 맥니마 검정 (McNemar test)

- 맥니마 검정 결과

② 신약효과 사례

	S	F	
S	12	49	61(71%)
F	10	15	
	22(26%)		86

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{49 - 10}{\sqrt{49 + 10}} = 5.1$$

$P-$ 값 < 0.0001

3. 맥니마 검정 (McNemar test)

- $\pi_{1+} - \pi_{+1}$ 에 대한 신뢰구간 작성
  - $\pi_{1+} - \pi_{+1}$ 의 추정량 : $p_{1+} - p_{+1}$
  - $SE = \sqrt{\widehat{Var}(p_{1+} - p_{+1})}$

$$= \frac{1}{n}\sqrt{(n_{12} + n_{21}) - \frac{(n_{12} - n_{21})^2}{n}}$$

  - 95% 신뢰구간 : $(p_{1+} - p_{+1}) \pm 1.96 \times SE$

10 강

다음시간안내

분할표에 대한 로그선형 모형

수고하셨습니다.