바이오통계 과제

-제 4장 Logit and loglinear models-

2019314199

통계학과

김동환

목차

1. Logit model – binary response

Ⅰ. 데이터 소개

이 연구의 목표는 현재 피임 사용 여부가 현재 자녀를 원하고 있는 정도와 교육수준에 따라 달라지는지 알아보는데 있다. 해당 연구에서는 교육수준이 Lower, Upper인 여성을 대상으로 현재 아이를 원하는지 여부를 바탕으로 그들의 피임사용 여부를 조사한 자료이다.

교육수준과 아이를 현재 원하는가를 기준으로 피임여부 조사 표

Education	Desires More Children?	Contraceptive Use		Total
		Yes	No	
Lower	Yes	53	7	60
	No	38	9	47
Upper	Yes	212	52	264
	No	50	38	88
Total		353	106	459

출처: Current Use of Contraception Among Married Women by Age, Education and Desire for More Children Fiji Fertility Survey, 1975

Ⅱ. 분석

1)Goodness of fit

H_0: *자료가 모형에 적합하다.*

$H_1 : not\ H_0$

Deviance value = 0.7974, P-value = 0.3719

Pearson value = 0.8167, P-value = 0.3661

$Cannot\ reject\ H_0$

Testing Global Null Hypothesis : BETA = 0

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : not\ H_0$

P-value < 0.001 from LRT, Score, Wald test

$Cannot\ reject\ H_0$

LRT 검정 통계량 LR=-2[lnL($\hat{\beta}_0, \beta_1 = 0$) - lnL($\hat{\beta}_0, \hat{\beta}_1$)] ~$\chi_1^2$

Walds 검정 통계량 W = $\frac{(\hat{\beta}_1 - (\beta_1 = 0))^2}{\mathrm{var}(\hat{\beta}_1)} \sim \chi_1^2$

score 검정 통계량 $s(\theta) = \frac{d\ log\ L(\theta)}{d\theta}$ ~χ^2

Type 3 Analysis of Effects 결과

education

Chi-sqaure : 8.6292, P-value : 0.0033

desires more children

Chi-square : 18.3822, P-value < 0.0001

educatoin과 desires는 모두 피임율에 유의한 영향을 미친다.

2) Fitted logistic regression model

$$ln\frac{\pi(x)}{1-\pi(x)} = 1.3120 - 0.4556\ education + 0.5147 desires$$

Standard Error :	0.1515	0.1551	0.1201
		L : 1	N : 1
		U : -1	Y : -1

2-1) 교육 수준 Lower vs Upper

desires more children 이 고정되어있을 때 :

Education = L, Desires =N, $\hat{\beta}_1(L) = -0.4556$

Education = U, Desires =N, $\hat{\beta}_1(U) = 0.4556$

Odds ratio:

Education L vs U

log odds ratio $\hat{\beta}_1 - (-\hat{\beta}_1) = 2\hat{\beta}_1$=2 x (-0.4556) = -0.9112

odds ratio = $e^{2\hat{\beta}_1}$=0.402

교육수준이 Lower 일 때, 피임을 할 odds는 Upper일 때의 0.402배이다.

2-2) Desires more children? No vs Yes

Education 이 고정되어있을 때 :

Education = L, Desires = N, $\hat{\beta}_2(N) = 0.5147$

Education = L, Desires = Y, $\hat{\beta}_2(Y)$ = -0.5147

Odds ratio:

Desires N vs Y

log odds ratio $\hat{\beta}_2 - (-\hat{\beta}_2) = 2\hat{\beta}_2$=2 x0.5147 = 1.0294

odds ration = = $e^{2\hat{\beta}_2}$=2.800

아이를 원하지 않는 여성이 피임을 할 odds는 아이를 원하는 여성에 2.8배이다.

3) 결과

교육수준이 높은 경우 피임을 하는 비율이 더 높고,

아이를 원하지 않는 경우 피임을 하는 비율이 더 높다.

Ⅲ. 코드

SAS

```
data con;
input edu $ des $ outcome $ count @@;
cards;
l y yes 53 l y no 7
l n yes 38 l n no 9
u y yes 212 u y no 52
u n yes 50 u n no 38
;
run;

proc logistic data = con;
freq count;
class edu des;
model outcome = edu des / scale = none aggregate;
run;
```

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23.8286	2	<.0001
Score	23.8167	2	<.0001
Wald	22.5753	2	<.0001

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.7974	1	0.7974	0.3719
Pearson	0.8167	1	0.8167	0.3661

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
edu	1	8.6292	0.0033
des	1	18.3822	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3120	0.1514	75.0612	<.0001
edu	l	1	-0.4556	0.1551	8.6292	0.0033
des	n	1	0.5147	0.1201	18.3822	<.0001

R

```
sink('con.txt')
cat('edu des outcome count
l y yes 53
l y no 7
l n yes 38
l n no 9
u y yes 212
u y no 52
u n yes 50
u n no 38
')
sink()
con<-read.table('con.txt',sep=' ',header=T)
unlink('con.txt')

con1 <- con[rep(row.names(con),con$count),-4]
xtabs(~edu + des + outcome, data=con1)

con1$outcomey<-ifelse(con1$outcome=='yes',1,0)
con1$edul<-ifelse(con1$edu=='l',1,-1)
con1$desy<-ifelse(con1$des=='n',1,-1)
fit1<-glm(outcomey~edul+desy,data=con1,family='binomial')
summary(fit1)
```

```
> summary(fit1)

Call:
glm(formula = outcomey ~ edul + desy, family = "binomial", data = con1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1815   0.4409   0.6726   0.6726   1.0362

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3120     0.1514   8.664  < 2e-16 ***
edul          0.4556     0.1551   2.938  0.00331 **
desy         -0.5147     0.1201  -4.287 1.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 496.09  on 458  degrees of freedom
Residual deviance: 472.26  on 456  degrees of freedom
AIC: 478.26

Number of Fisher Scoring iterations: 4
```

2. Multinomial response (ordinal)

Ⅰ. 데이터 소개

이 연구의 목적은 어린이들의 성별과 나이가 티비를 자주 보는 정도에 영향을 미치는지 알아보는 것이다. 다음 자료는 성별과 나이그룹 x(7세-9세) , y(10세-12세)에 따라 티비를 얼마나 자주보는지 s (seldom or never), e (every week), a (almost daily)로 구분하여 조사하였다.

성별, 나이 그룹에 따른 티비를 자주 보는 정도

성 별	나이그룹 x(7-9), y(10-12)	티비를 얼마나 보는가			Total
		s	e	a	
남	x	88	28	5	121
남	y	78	20	4	102
여	x	87	14	5	106
여	y	93	17	3	113
Total					442

출처: Andersen, D. (1995): School children's leisure hours. (In Danish). Report no. 95:2. Copenhagen: Danish National Institute of Social Research.

Ⅱ. 분석

1) Score Test for the Proportional Odds Assumption

H_0: $Can\ assume\ Proportional\ Odds\ model.$

H_1: $Not\ H_0$.

검정 통계량 : $s(\theta) = \frac{d\ log\ L(\theta)}{d\theta}\ \sim \chi^2$

$\chi^2 = 0.1153, P - value = 0.9440$

$Cannot\ reject\ H_0$

2)Goodness of fit

Deviance and Pearson Goodness-of-Fit Statistics

H_0: *자료가 모형에 적합하다.*

H_1 : $not\ H_0$

Deviance value = 3.2933, P-value = 0.5100

Pearson value = 3.3130, P-value = 0.5069

$Cannot\ reject\ H_0$

3) Fitted model

$$\hat{L}_{j|i}(x) = ln\frac{F_j(x)}{1-F_j(x)} = \hat{\alpha}_j - \hat{\beta}_1 gender + \hat{\beta}_2 age$$

$$= \hat{\alpha}_j - (0.5212)gender + (-1.1273)age$$

standard error:	0.2522	0.2521
	f: 1	x: 1
	m: -1	y: -1

$estimated\ cutpoint\ parameters:\ \hat{\alpha}_{1=0.000162}, \hat{\alpha}_{2=0.7236}$

LRT 검정 통계량 LR=-2[lnL($\hat{\beta}_0, \beta_1 = 0$) - lnL($\hat{\beta}_0, \hat{\beta}_1$)] ~$\chi_1^2$

Walds 검정 통계량 W = $\frac{(\hat{\beta}_1 - (\beta_1 = 0))^2}{\mathrm{var}(\hat{\beta}_1)} \sim \chi_1^2$

score 검정 통계량 $s(\theta) = \frac{d\ log\ L(\theta)}{d\theta} \sim \chi^2$

Testing Global Null Hypothesis : BETA = 0

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : not\ H_0$

P-value < 0.001 from LRT, Score, Wald test

$reject\ H_0$

Type 3 Analysis of Effects 결과

gender

Chi-sqaure : 4.2711, P-value : 0.0388

desires more children

Chi-square : 19.9903, P-value : 0.0001

성별과 나이그룹은 모두 티비시청률에 유의한 영향을 미친다.

3-1) 성별 female vs male

age_group이 일정할 때:

Gender = f, age_group = A, $\hat{\beta}_1(f) = 0.5212$

Gender = m, age_group = A, $\hat{\beta}_1(m) = -0.5212$

$\hat{\beta}_1 - (-\hat{\beta}_1) = 2\hat{\beta}_1$=2 x 0.5212 = 1.0424

age_group 이 고정되어있을 때, 여성이 티비를 적게보는 범주에 속할 오즈가 남성의 $e^{2\hat{\beta}_1}$=2.836배이다.

3-2) 나이그룹 x(7-9),y(10-12)

성별이 일정할 때:

age_group = x, Gender = m, $\hat{\beta}_2(x) = -1.1273$

age_group = y, Gender = m, $\hat{\beta}_2(y) = 1.1273$

$\hat{\beta}_2 - (-\hat{\beta}_2) = 2\hat{\beta}_2$=2 x (-1.1273) = -2.2546

gender 가 고정되어있을 때, 그룹 x가 티비를 적게보는 범주에 속할 오즈가 그룹 y의 $e^{2\hat{\beta}_1}$=0.105배이다.

즉, 나이가 고정되어있으면 남성이 티비를 더 많이 보는 경향이 있고

성별이 고정되어있으면 나이그룹x가 티비를 더 많이 보는 경향이 있다.

4) Estimated cumulative probability

gender	age	pro	count	__LEVLEL__	p	$F_j(x)$
m	x	s	2	s	0.16134	P(Y=s) = 0.16134 P(Y=e) = 0.12262 p(Y=a) = 0.71604
m	x	s	2	e	0.28396	
m	y	s	19	s	0.64708	P(Y=s) = 0.64708 P(Y=e) = 0.14370 P(Y=a) = 0.20922
m	y	s	19	e	0.79078	
f	x	s	8	s	0.35299	P(Y=s) = 0.35299 P(Y=e) = 0.17635 P(Y=a) = 0.47066
f	x	s	8	e	0.52934	
f	y	s	11	s	0.83871	P(Y=s) = 0.83871 P(Y=e) = 0.07596 P(Y=a) = 0.08533
f	y	s	11	e	0.91467	

거의 안봄(s), 주마다 봄 (e) , 매일 봄 (a)

gender가 고정되어있으면, x가 y보다 P(Y=a) 가 높고

age가 고정되어있으면, 남성이 여성 보다 P(Y=a) 가 높다.

Ⅲ. 코드

SAS

```
data television;
input gender $ age $ pro $ count @@;
cards;
m x s 2 m x e 4 m x a 16
m y s 19 m y e 3 m y a 6
f x s 8 f x e 2 f x a 9
f y s 11 f y e 2 f y a 1
;
proc logistic order = data data=television;
freq count;
class gender age;
model pro = gender age /scale=none aggregate;
output out = prob pred=p;
run;

proc print data=prob(obs=20);
run;
```

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
0.1153	2	0.9440

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	3.2933	4	0.8233	0.5100
Pearson	3.3130	4	0.8283	0.5069

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.8587	2	<.0001
Score	24.3235	2	<.0001
Wald	20.3663	2	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	4.2711	0.0388
age	1	19.9903	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	s	1	0.000162	0.2556	0.0000	0.9995
Intercept	e	1	0.7236	0.2680	7.2876	0.0069
gender	f	1	0.5212	0.2522	4.2711	0.0388
age	x	1	-1.1273	0.2521	19.9903	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender f vs m	2.836	1.055	7.622
age x vs y	0.105	0.039	0.282

OBS	gender	age	pro	count	_LEVEL_	p
1	m	x	s	2	s	0.16134
2	m	x	s	2	e	0.28396
3	m	x	e	4	s	0.16134
4	m	x	e	4	e	0.28396
5	m	x	a	16	s	0.16134
6	m	x	a	16	e	0.28396
7	m	y	s	19	s	0.64708
8	m	y	s	19	e	0.79078
9	m	y	e	3	s	0.64708
10	m	y	e	3	e	0.79078
11	m	y	a	6	s	0.64708
12	m	y	a	6	e	0.79078
13	f	x	s	8	s	0.35299
14	f	x	s	8	e	0.52934
15	f	x	e	2	s	0.35299
16	f	x	e	2	e	0.52934
17	f	x	a	9	s	0.35299
18	f	x	a	9	e	0.52934
19	f	y	s	11	s	0.83871
20	f	y	s	11	e	0.91467

R

```
sink('pro.txt')
cat('gender age pro count
1 -1 s 2
1 -1 e 4
1 -1 a 16
1 1 s 19
1 1 e 3
1 1 a 6
-1 -1 s 8
-1 -1 e 2
-1 -1 a 9
-1 1 s 11
-1 1 e 2
-1 1 a 1
')
sink()
pro <-read.table('pro.txt',sep=' ',header=T)
unlink('pro.txt')
pro$ord_pro<-ordered(pro$pro,levels=c('s','e','a'))
fit1<-polr(ord_pro~gender+age,data=pro,weight=count)
summary(fit1)
```

```
> summary(fit1)

Re-fitting to get Hessian

Call:
polr(formula = ord_pro ~ gender + age, data = pro, weights = count)

Coefficients:
         Value Std. Error t value
gender  0.5212     0.2510   2.077
age    -1.1273     0.2499  -4.510

Intercepts:
    Value   Std. Error t value
s|e  0.0002  0.2573     0.0006
e|a  0.7236  0.2708     2.6719

Residual Deviance: 137.9976
AIC: 145.9976
```

3. Multinomial response (norminal)

Ⅰ. 데이터 소개

다음은 군인의 계급과 보직에 따라 사망자 수를 조사한 자료이며, 군인의 계급과 보직이 사망원인과 어떤 관련이 있는지 알아보고자 한다. 사망원인은 질병으로 인한 사망(d), 부상으로 인한 사망(j), 그 외 (s)로 분류하였고, 군인의 계급은 이등병인지의 여부 (이등병 : p , 이등병보다 높음 : h), 보직은 보병(i), 보병아님(ni)로 구분하였다.

군인 사망원인 자료

이등병인가? Private?	보병인가? Infantry?	consolidation 후			
		Disease	Injury	Other	Total
p	i	285	60	8	353
p	ni	15	50	2	67
h	i	13	18	1	32
h	ni	2	4	0	6
Total					458

출처: C. Lee (1999), "Selective Assignment of Military Positions in the Union Army: Implications for the Impact of the Civil War", Social Science History, Vol. 23, #1, pp 67-97.

Ⅱ. 분석

1) Goodness of fit

Maximum Likelihood Analysis of Variance

H_0: *자료가 모형에 적합하다.*

$H_1 : not\ H_0$

Likelihood Ratio value = 4.60, P-value=0.1005

$Cannot\ reject\ H_0$

Test for Beta = 0 by Maximum Likelihood Analysis of Variance

H_0: 베타의 계수가 0이다.

$H_1 : not\ H_0$

$\beta_{private}, Chi - square = 17.38, P - value = 0.0002$

$\beta_{infantry}, Chi - square = 69.63, P - value < 0.0001$

$reject\ H_0$

private과 infantry의 계수가 모두 0이 아니고 유의하다.

2) Fitted model

(disease)

$$ln\frac{\pi_d}{\pi_s} = 2.4746 + 0.3863\ private + 0.6946 infantry$$

Standard Error :	0.6199	0.5424	0.4113
		p : 1	i : 1
		h : -1	ni : -1

(injury)

$$ln\frac{\pi_j}{\pi_s} = 3.0391 - 0.4064\ private - 0.6047\ infantry$$

Standard Error :	0.6081	0.5412	0.4024
		p : 1	i : 1
		h : -1	ni : -1

범주 d와 j의 비교는 위 두 식의 차로 구할 수 있다.

$$ln\frac{\pi_d}{\pi_j} = -0.82931 + 0.7927\ private + 1.2993\ infantry$$

2-1) 첫 번째 식에서 그 외 원인으로 사망 vs 질병으로 사망의 추정오즈

계급 Private vs Higher, 보직이 동일할 때:

$\hat{\beta}_1 - (-\hat{\beta}_1) = 2\hat{\beta}_1$=2 x 0.3863 = 0.7726

이등병은 다른 계급보다 $e^{2\hat{\beta}_1}$=2.165배가 됨으로, 이등병은 다른 계급보다 질병으로 사망비율이 높다.

보직 Infantry vs Non-Infantry, 계급이 동일할 때:

$\hat{\beta}_2 - (-\hat{\beta}_2) = 2\hat{\beta}_2$=2 x 0.6946 = 1.3892

보병은 다른 보직보다 $e^{2\hat{\beta}_2}$=4.01배가 됨으로, 보병은 다른 보직보다 질병으로 사망비율이 높다.

2-2) 두 번째 식에서 그 외 원인으로 사망 vs 부상으로 사망의 추정오즈

계급 Private vs Higher, 보직이 동일할 때:

$\hat{\beta}_1 - (-\hat{\beta}_1) = 2\hat{\beta}_1$=2 x(- 0.4064) = -0.8128

이등병은 다른 계급보다 $e^{2\hat{\beta}_1}$=0.443배가 됨으로, 이등병은 다른 계급보다 부상으로 사망비율이 낮다.

보직 Infantry vs Non-Infantry, 계급이 동일할 때:

$\hat{\beta}_2 - (-\hat{\beta}_2) = 2\hat{\beta}_2$=2 x(- 0.6047) = -1.2094

보병은 다른 보직보다 $e^{2\hat{\beta}_2}$=0.298배가 됨으로, 보병은 다른 보직보다 부상으로 사망비율이 낮다.

3) 각 수준별 반응 확률에 대한 추정 값

$$ln\frac{\pi_d}{\pi_s} = 2.4746 + 0.3863\,x_{1l} + 0.6946x_{2l}$$

$$ln\frac{\pi_j}{\pi_s} = 3.30391 - 0.4064\,x_{1l} - 0.6047\,x_{2l}$$

3-1) Private = p , Infantry = i 일 때, $x_{1l} = 1, x_{2l} = 1$

$$disease:\ \pi_{11} = \frac{\exp(2.4746+0.3863+0.6946)}{1+\exp(2.4746+0.3863+0.6946)+\exp(3.0391-0.4064-0.6047)}$$=0.8074

$$injury:\pi_{12} = \frac{\exp(3.0391-0.4064-0.6047)}{1+\exp(2.4746+0.3863+0.6946)+\exp(3.0391-0.4064-0.6047)}$$=0.1742

$$other:\ \pi_{13} = \frac{1}{1+\exp(2.4746+0.3863+0.6946)+\exp(3.0391-0.4064-0.6047)}$$=0.02293

3-2) Private = p , Infantry = ni 일 때, $x_{1l} = 1, x_{2l} = -1$

$$disease:\ \pi_{21} = \frac{\exp(2.4746+0.3863-0.6946)}{1+\exp(2.4746+0.3863-0.6946)+\exp(3.0391-0.4064+0.6047)}$$=0.2239

$$injury:\pi_{22} = \frac{\exp(3.0391-0.4064+0.6047)}{1+\exp(2.4746+0.3863-0.6946)+\exp(3.0391-0.4064+0.6047)}$$=0.7463

$$other:\ \pi_{23} = \frac{1}{1+\exp(2.4746+0.3863-0.6946)+\exp(3.0391-0.4064+0.6047)}$$=0.0299

3-3) Private = h , Infantry = i 일 때, $x_{1l} = -1, x_{2l} = 1$

$$disease: \pi_{21} = \frac{\exp(2.4746-0.3863+0.6946)}{1+\exp(2.4746-0.3863+0.6946)+\exp(3.0391+0.4064-0.6047)} = 0.4194$$

$$injury: \pi_{22} = \frac{\exp(3.0391+0.4064-0.6047)}{1+\exp(2.4746-0.3863+0.6946)+\exp(3.0391+0.4064-0.6047)} = 0.5484$$

$$other: \pi_{23} = \frac{1}{1+\exp(2.4746-0.3863+0.6946)+\exp(3.0391+0.4064-0.6047)} = 0.0323$$

3-4) Private = h , Infantry = ni 일 때, $x_{1l} = -1, x_{2l} = -1$

$$disease: \pi_{21} = \frac{\exp(2.4746-0.3863-0.6946)}{1+\exp(2.4746-0.3863-0.6946)+\exp(3.0391+0.4064+0.6047)} = 0.2239$$

$$injury: \pi_{22} = \frac{\exp(3.0391+0.4064+0.6047)}{1+\exp(2.4746-0.3863-0.6946)+\exp(3.0391+0.4064+0.6047)} = 0.7463$$

$$other: \pi_{23} = \frac{1}{1+\exp(2.4746-0.3863-0.6946)+\exp(3.0391+0.4064+0.6047)} = 0.0299$$

Ⅲ. 코드

SAS

```
data soldier;
input private $ type $ cause $ count @@;
cards;
p i d 285 p i j 60 p i s 8
p ni d 15 p ni j 50 p ni s 2
h i d 13 h i j 17 h i s 1
h ni d 2 h ni j 4 h ni s 0
;
run;
```

```
proc catmod order = data;
weight count;
model cause = private type / pred=prob;
run;
```

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	28.60	<.0001
private	2	17.38	0.0002
type	2	69.63	<.0001
Likelihood Ratio	2	4.60	0.1005

Analysis of Maximum Likelihood Estimates

Parameter		Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	2.4746	0.6199	15.94	<.0001
		2	3.0391	0.6081	24.98	<.0001
private	p	1	0.3863	0.5424	0.51	0.4764
	p	2	-0.4064	0.5412	0.56	0.4527
type	i	1	0.6946	0.4113	2.85	0.0912
	i	2	-0.6047	0.4024	2.26	0.1329

Maximum Likelihood Predicted Values for Probabilities

private	type	cause	Observed Probability	Observed Standard Error	Predicted Probability	Predicted Standard Error	Residual
p	i	d	0.8074	0.021	0.8028	0.0211	0.0046
		j	0.17	0.02	0.1743	0.0201	-0.004
		s	0.0227	0.0079	0.0229	0.0079	-27E-5
p	ni	d	0.2239	0.0509	0.248	0.052	-0.024
		j	0.7463	0.0532	0.7236	0.0538	0.0226
		s	0.0299	0.0208	0.0284	0.0199	0.0014
h	i	d	0.4194	0.0886	0.4714	0.0877	-0.052
		j	0.5484	0.0894	0.4995	0.0875	0.0489
		s	0.0323	0.0317	0.0292	0.0288	0.0031
h	ni	d	0.3333	0.1925	0.0645	0.0278	0.2688
		j	0.6667	0.1925	0.9194	0.0352	-0.253
		s	0	0	0.016	0.0197	-0.016

R

```
sink('soldier.txt')
cat('private type cause count
1 1 d 285
1 1 j 60
1 1 s 8
1 -1 d 15
1 -1 j 50
1 -1 s 2
-1 1 d 13
-1 1 j 17
-1 1 s 1
-1 -1 d 2
-1 -1 j 4
-1 -1 s 0
')
sink()
sol <- read.table('soldier.txt',sep=' ',header=T)
unlink('soldier.txt')

sol$cause1<-factor(sol$cause,levels=c('d','j','s'))
fit1<-multinom(cause1~private+type,data=sol,weights=count)
summary(fit1)
```

```
> summary(fit1)
Call:
multinom(formula = cause1 ~ private + type, data = sol, weights = count)

Coefficients:
  (Intercept)    private       type
j   0.5935008 -0.8221805 -1.2984035
s  -2.4718707 -0.3882289 -0.6959062

Std. Errors:
  (Intercept)   private      type
j   0.2210817 0.1880727 0.1559379
s   0.6205780 0.5425276 0.4112875
```

R과 SAS의 계수는 다르지만 해석은 같다.

4. Logilinear model

Ⅰ. 데이터 소개

이 자료는 빈곤 여부, 성별, 인종에 따른 청소년 수를 조사한 것이다. 연구자는 이 세 범주형 변수 사이에 어떤 관계가 있는지 알아보고자 한다.

인종, 성별, 빈곤 여부에 따른 청소년 수

race	gender	poverty?		Total
		yes	no	
black	male	27	79	106
	female	18	78	96
white	male	301	233	534
	female	120	200	320
Total				1056

출처: Wenk and Hardesty (1993). "The Effects of Rural-to-Urban Migration on the Poverty Status of Youth in the 1980s", Rural Sociology, 58(1) pp. 76-92.

Ⅱ. 분석

1) 로그-선형 모형 선택 가설

H_0: *모형이 적합하다.*

$H_1: not\ H_0$

2) 우도비 검정통계량

$$G^2 = -2\ log\ \lambda = \Sigma_i^2 \Sigma_j^2 \Sigma_k^2 n_{ijk} log(\frac{n_{ijk}}{\hat{m}_{ijk}})$$

포화 log-linear모델 적합 결과 Maximum Likelihood Analysis of Variance에서
세 변수에 대한 교호작용 (XYZ)는 Chi-sqaure = 1.01, P-value는 0.3156로 유의수준 0.05에서 유의하지않다. 따라서 3요인 교호작용 (XYZ)를 제거하고 변수들 간의 2요인 교호작용을 고려한 더 간단한 모형을 적합한다.

2요인 교호작용 모델 (XY,XZ,YZ)를 적합한 결과 Likelihood Ratio가
Chi-sqaure = 0.99, P-value는 0.3186으로 유의수준 0.05에서 귀무가설 : 모형이 적합하다. 를 기각할 수 없으나, 교호작용 gender*type(YZ)가 Chi-square = 2.03, P-value는 0.1542로 유의수준 0.05에서 유의하지 않음으로, 이를 고려하여 교호작용 중 (YZ)를 제거하고 (XY,XZ)을 고려한 모델을 적합한다.

(XZ,XY) 적합 결과 Likelihood Ratio가 Chi-square = 3.02, P-value는 0.2214로 유의수준 0.05에서 귀무가설 : 모형이 적합하다.를 기각할 수 없고, Maximum Likelihood Analysis of Variance에서 모든 주효과와 교호작용이 유의수준 0.05에서 P-value <.0001로 유의함으로 최종 로그-선형 모형은 (XY,XZ)로 결정한다.

3) 모형적합 결과 표

로그-선형 모형	독립 유형	$G^2(M)(df)$	$G^2(M_2\|M_1)(df)$	${\chi^2}_{0.05}(df)$
(XY,YZ,XZ)	3요소 교호작용 없음	1.01(1)		
(XZ,XY)	조건부 독립 (conditional indep)	3.02(2)	2.01(1)	3.841(1)
(XY,YZ)		47.52(2)	46.51(1)	
(XZ,YZ)		30.05(2)	29.04(1)	
(XY,Z)	결합 독립 (jointly indep)	54.34(3)	53.33(2)	5.991(2)
(YZ,X)		81.38(3)	80.37(2)	
(XZ,Y)		36.87(3)	35.86(2)	
(X,Y,Z)	상호독립(mutual indep)	88.20(4)	87.19(3)	7.815(3)

최적 모형 선택 :

위에서의 설명과 더불어

$H_0 : {\lambda_{ij}}^{YZ} = 0$ 에서 $G^2(XY, XZ \mid XY, XZ, YZ) = G^2(XY, XZ) - G^2(XY, XZ, YZ)$ =3.02-1.01 = 2.01이고 $df(XY, XZ \mid XY, XZ, YZ) = df(XY, XZ) - df(XY, XZ, YZ)$=2-1 =1 이므로, P-value = P(${\chi^2}_{0.05} \geq 2.01$)이므로 $H_0 : {\lambda_{ij}}^{YZ} = 0$ 가 기각되지않는다. 따라서 최종 모형으로 (XY,XZ)를 선택한다.

4)결과

모든 주효과와 교호작용효과가 유의하다. 교호작용을 보면 poverty*gender로부터 청소년 중 남성이 더 빈곤을 겪고 있고, poverty*race로부터 흑인보다는 백인이 더 빈곤을 겪고있다.

Ⅲ. 코드

SAS

```
data pov;
input poverty $ gender $ type $ count;
cards;
poverty male black  27
poverty male white 301
poverty female black  18
poverty female white 120
not-poverty male black  79
not-poverty male white 233
not-poverty female black 78
not-poverty female white 200
;
run;

proc catmod order = data;
weight count;
model poverty*gender*type = _response_/NOITER PRED=FREQ;
loglin poverty|gender|type;
run;

proc catmod order = data data=pov;
weight count;
model poverty*gender*type = _response_/NOITER PRED=FREQ;
loglin poverty|gender poverty|type gender|type;
run;

proc catmod order = data data=pov;
weight count;
model poverty*gender*type = _response_/NOITER PRED=FREQ;
loglin poverty|gender poverty|type;
run;
```

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
poverty	1	56.19	<.0001
gender	1	15.98	<.0001
poverty*gender	1	9.67	0.0019
type	1	288.42	<.0001
poverty*type	1	37.57	<.0001
gender*type	1	3.08	0.0793
poverty*gender*type	1	1.01	0.3156
Likelihood Ratio	0	.	.

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
poverty	1	57.45	<.0001
gender	1	25.26	<.0001
poverty*gender	1	28.41	<.0001
type	1	301.54	<.0001
poverty*type	1	41.10	<.0001
gender*type	1	2.03	0.1542
Likelihood Ratio	1	0.99	0.3186

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
poverty	1	60.58	<.0001
gender	1	56.30	<.0001
poverty*gender	1	32.93	<.0001
type	1	317.56	<.0001
poverty*type	1	44.84	<.0001
Likelihood Ratio	2	3.02	0.2214

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
poverty	poverty	-0.3640	0.0468	60.58	<.0001
gender	male	0.2453	0.0327	56.30	<.0001
poverty*gender	poverty male	0.1876	0.0327	32.93	<.0001
type	black	-0.8126	0.0456	317.56	<.0001
poverty*type	poverty black	-0.3054	0.0456	44.84	<.0001

R

```
sink('poverty.txt')
cat('race gender poverty count
1 1 p 27
1 1 np 79
1 -1 p 18
1 -1 np 78
-1 1 p 301
-1 1 np 233
-1 -1 p 120
-1 -1 np 200
')
sink()
sol <- read.table('poverty.txt',sep=' ',header=T)
unlink('pverty.txt')

sol$poverty<-factor(sol$poverty,levels=c('np','p'))
sol$race<-factor(sol$race,levels=c('-1','1'))
sol$gender<-factor(sol$gender,levels=c('-1','1'))


fit0 <- glm(count ~ poverty + gender + race,
            data = sol, family = poisson)

fit1 <- glm(count ~ (poverty + gender + race)^2,
            data = sol, family = poisson)

fit2 <- glm(count ~ poverty * gender * race,
            data = sol, family = poisson)

fit3 <- glm(count ~ race * poverty + race * gender,
            data = sol, family = poisson)

fit4 <- glm(count ~ race * poverty + poverty * gender,
            data = sol, family = poisson)

fit5 <- glm(count ~ race * gender + poverty * gender,
            data = sol, family = poisson)

fit6 <- glm(count ~ race + gender * poverty,
            data = sol, family = poisson)

fit7 <- glm(count ~ race * poverty + gender,
            data = sol, family = poisson)

fit8 <- glm(count ~ gender * poverty + race,
            data = sol, family = poisson)

fit9 <- glm(count ~ race * gender + poverty,
            data = sol, family = poisson)
```

```
> AIC(fit0,fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9)
     df       AIC
fit0  4 147.07844
fit1  7  65.87710
fit2  8  66.88251
fit3  6  92.93365
fit4  6  65.89849
fit5  6 110.40668
fit6  5 115.22498
fit7  5  97.75195
fit8  5 115.22498
fit9  5 142.26014
```

AIC가 작은 fit1과 fit4를 고려해보자

```
> summary(fit1)

Call:
glm(formula = count ~ (poverty + gender + race)^2, family = poisson,
    data = sol)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-0.4922  0.3003  0.6554 -0.2955  0.1526 -0.1722 -0.2392  0.1874

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.28503    0.06993  75.578  < 2e-16 ***
povertyp         -0.47579    0.10962  -4.340 1.42e-05 ***
gender1           0.17727    0.09335   1.899   0.0576 .
race1            -0.89505    0.12396  -7.221 5.18e-13 ***
povertyp:gender1  0.71179    0.13353   5.330 9.80e-08 ***
povertyp:race1   -1.18238    0.18444  -6.410 1.45e-10 ***
gender1:race1    -0.23178    0.16267  -1.425   0.1542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 583.75951  on 7  degrees of freedom
Residual deviance:   0.99459  on 1  degrees of freedom
AIC: 65.877
```

```
> summary(fit4)

Call:
glm(formula = count ~ race * poverty + poverty * gender, family = poisson,
    data = sol)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-0.8523 -0.4452  1.2147  0.4637  0.2708  0.2651 -0.4212 -0.2826

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.31824    0.06490  81.948  < 2e-16 ***
race1            -1.01449    0.09316 -10.890  < 2e-16 ***
povertyp         -0.49254    0.10811  -4.556 5.22e-06 ***
gender1           0.11538    0.08248   1.399    0.162
race1:povertyp   -1.22148    0.18242  -6.696 2.14e-11 ***
povertyp:gender1  0.75038    0.13076   5.739 9.54e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 583.760  on 7  degrees of freedom
Residual deviance:   3.016  on 2  degrees of freedom
AIC: 65.898
```

fit1 은 (XY,XZ,YZ) 를 적합시킨 결과인데, SAS와 마찬가지로 gender*race 의 interaction term이 유의하지 않음을 알 수 있다. 따라서 SAS와 마찬가지로 gender*race를 제거한 (XY,XZ) 모형을 채택한다.