범주형자료분석방법론 HW4

이름: 김연주

학과: 통계학과

학번: 2021250461

1. Table10.1

```
> #table10.1
> t10.1=data.frame(expand.grid(cig=c("Yes","No"),alc=c("Yes","No"),mar=c("Yes","N
o"),sex=c("female","male"),race=c("white","other")), count=c(405,13,1,1,268,218,1
7,117,453,28,1,1,228,201,17,133,23,2,0, 0,23, 19,1,12,30,1,1,0,19,18,8,17))
> t10.1
   cig alc mar    sex  race count
1  Yes Yes Yes female white   405
2   No Yes Yes female white    13
3  Yes  No Yes female white     1
4   No  No Yes female white     1
5  Yes Yes  No female white   268
6   No Yes  No female white   218
7  Yes  No  No female white    17
8   No  No  No female white   117
9  Yes Yes Yes   male white   453
10  No Yes Yes   male white    28
11 Yes  No Yes   male white     1
12  No  No Yes   male white     1
13 Yes Yes  No   male white   228
14  No Yes  No   male white   201
15 Yes  No  No   male white    17
16  No  No  No   male white   133
17 Yes Yes Yes female other    23
18  No Yes Yes female other     2
19 Yes  No Yes female other     0
20  No  No Yes female other     0
21 Yes Yes  No female other    23
22  No Yes  No female other    19
23 Yes  No  No female other     1
24  No  No  No female other    12
25 Yes Yes Yes   male other    30
26  No Yes Yes   male other     1
27 Yes  No Yes   male other     1
28  No  No Yes   male other     0
29 Yes Yes  No   male other    19
30  No Yes  No   male other    18
31 Yes  No  No   male other     8
32  No  No  No   male other    17
```

```
> fit1=glm(count ~ cig + alc + mar + sex + race +sex*race, family=poisson,data=t1
0.1)
> fit2=glm(count ~ .^2, family=poisson, data=t10.1)
> fit3=glm(count~.^3,data=t10.1,family=poisson)
> fit4=glm(count~cig*mar+cig*sex+cig*race+alc*mar+alc*sex+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit5=glm(count~cig*alc+cig*mar+cig*sex+cig*race+alc*sex+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit6=glm(count~cig*alc+cig*sex+cig*race+alc*mar+alc*sex+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit7=glm(count~cig*alc+cig*mar+cig*sex+cig*race+alc*mar+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit8=glm(count~cig*alc+cig*mar+cig*sex+cig*race+alc*mar+alc*sex+mar*sex+mar*race
+sex*race,data=t10.1,family=poisson)
> fit9=glm(count~cig*alc+cig*mar+cig*race+alc*mar+alc*sex+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit10=glm(count~cig*alc+cig*mar+cig*sex+alc*mar+alc*sex+alc*race+mar*sex+mar*rac
e+sex*race,data=t10.1,family=poisson)
> fit11=glm(count~cig*alc+cig*mar+cig*sex+cig*race+alc*mar+alc*sex+alc*race+mar*ra
ce+sex*race,data=t10.1,family=poisson)
> fit12=glm(count~cig*alc+cig*mar+cig*sex+cig*race+alc*mar+alc*sex+alc*race+mar*se
x+sex*race,data=t10.1,family=poisson)
> fit13=glm(count~alc*cig+alc*mar+cig*mar+alc*sex+alc*race+sex*mar+sex*race+mar*ra
ce,data=t10.1,family=poisson)
> fit14=glm(count~alc*cig+alc*mar+cig*mar+alc*sex+alc*race+sex*mar+sex*race,data=t
10.1,family=poisson)
> fit15=glm(count~alc*cig+alc*mar+cig*mar+alc*sex+alc*race+sex*race,data=t10.1,fam
ily=poisson)

> model=c(1:15)
> Gsq=c(fit1$deviance,fit2$deviance,fit3$deviance,fit4$deviance,fit5$deviance,fit6
$deviance,fit7$deviance,fit8$deviance,fit9$deviance,fit10$deviance,fit11$deviance,
fit12$deviance,fit13$deviance,fit14$deviance,fit15$deviance)
> Xsq=c(sum(resid(fit1, type = "pearson")^2),
+       sum(resid(fit2, type = "pearson")^2),
+       sum(resid(fit3, type = "pearson")^2),
+       sum(resid(fit4, type = "pearson")^2),
+       sum(resid(fit5, type = "pearson")^2),
+       sum(resid(fit6, type = "pearson")^2),
+       sum(resid(fit7, type = "pearson")^2),
+       sum(resid(fit8, type = "pearson")^2),
+       sum(resid(fit9, type = "pearson")^2),
+       sum(resid(fit10, type = "pearson")^2),
+       sum(resid(fit11, type = "pearson")^2),
+       sum(resid(fit12, type = "pearson")^2),
+       sum(resid(fit13, type = "pearson")^2),
+       sum(resid(fit14, type = "pearson")^2),
+       sum(resid(fit15, type = "pearson")^2))
> DF=c(25,16,6,17,17,17,17,17,17,17,17,17,18,19,20)
> comp=data.frame(model,Gsq, Xsq, DF)
> comp
   model         Gsq         Xsq DF
1      1 1325.140761 1454.137896 25
2      2   15.340343   18.675326 16
3      3    5.272001    4.802283  6
4      4  201.199311  190.597369 17
5      5  106.957996  108.106684 17
6      6  513.472179  474.263793 17
7      7   18.716951   23.141176 17
8      8   20.320861   30.319719 17
9      9   16.317184   19.161474 17
10    10   15.783467   20.117758 17
11    11   25.161008   27.967052 17
12    12   18.928935   22.830848 17
13    13   16.735040   20.505053 18
14    14   19.908587   23.018092 19
15    15   28.805080   32.125171 20
```

```
> #fit1,2,3
> qchisq(0.95,df=25)
[1] 37.65248
> qchisq(0.95,df=16)
[1] 26.29623
> qchisq(0.95,df=6)
[1] 12.59159
```

fit1을 살펴보면, fit1이 적합하다는 귀무가설에 대해 G^2, X^2값 모두 임계값인 37.65248보다 크기에 귀무가설이 기각된다. 즉, fit1은 데이터를 잘 적합하지 못하고 saturated model을 지지하는 대립가설이 채택된다. 같은 방법으로 분석하면 fit2, fit3는 데이터를 잘 적합한다.

```
> #fit4-12
> qchisq(0.95,df=1)
[1] 3.841459
> Gsq[4]-Gsq[2]
[1] 185.859
> Xsq[4]-Xsq[2]
[1] 171.922
> Gsq[5]-Gsq[2]
[1] 91.61765
> Xsq[5]-Xsq[2]
[1] 89.43136
> Gsq[6]-Gsq[2]
[1] 498.1318
> Xsq[6]-Xsq[2]
[1] 455.5885
> Gsq[7]-Gsq[2]
[1] 3.376608
> Xsq[7]-Xsq[2]
[1] 4.46585
> Gsq[8]-Gsq[2]
[1] 4.980518
> Xsq[8]-Xsq[2]
[1] 11.64439
> Gsq[9]-Gsq[2]
[1] 0.9768414
> Xsq[9]-Xsq[2]
[1] 0.4861476
> Gsq[10]-Gsq[2]
[1] 0.4431241
> Xsq[10]-Xsq[2]
[1] 1.442432
> Gsq[11]-Gsq[2]
[1] 9.820665
> Xsq[11]-Xsq[2]
[1] 9.291726
> Gsq[12]-Gsq[2]
[1] 3.588592
> Xsq[12]-Xsq[2]
[1] 4.155522
```

Fit4와 fit2를 비교하면, 검정통계량 (G^2값과 X^2값의 차이)은 임계값보다 크다. 즉, fit4

를 지지하는 귀무가설이 기각되고 fit2를 지지하는 대립가설이 채택된다. 그에 따라 alc*cig항은 유의함을 알 수 있다. 같은 방법으로 fit5부터 fit12를 fit2와 비교하였다. 그 결과 alc*mar, cig*mar, alc*race, sex*mar항들은 모두 유의함을 알 수 있었다. 또한, alc*sex, mar*race항은 G^2 검정통계량은 유의하지 않았지만 X^2 검정통계량은 유의하였기에 해당 항이 유의하다고 판단하고 분석을 진행했다. 그리고 cig*sex, cig*race 항은 유의하지 않았다.

```
> #fit13,14,15
> Gsq[13]-Gsq[10]
[1] 0.9515726
> Xsq[13]-Xsq[10]
[1] 0.3872951
> Gsq[14]-Gsq[13]
[1] 3.173548
> Xsq[14]-Xsq[13]
[1] 2.513039
> Gsq[15]-Gsq[14]
[1] 8.896493
> Xsq[15]-Xsq[14]
[1] 9.107079
```

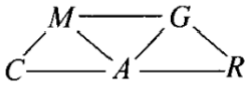
위와 같은 방법으로 fit13,14,15를 비교하였다. 그 결과 cig*race, mar*race항은 유의하지 않았고, sex*mar항은 유의하였다. 그에 따라 fit14를 채택하였다.

```
> #fit14 선택
> qchisq(0.95,df=19)
[1] 30.14353
> res14=resid(fit14, type = "pearson")/sqrt(1 - lm.influence(fit14)$hat)
> res14
          1           2           3           4           5           6
 1.99220082 -1.94454238 -0.26635745  0.79640308  0.06928227 -1.00055716
          7           8           9          10          11          12
-0.06123309  0.94353997  0.13921475  1.76675885 -0.86188249  0.33654953
         13          14          15          16          17          18
-0.73877576  0.39042443 -1.01855518  0.02721663 -1.13850914  0.57436494
         19          20          21          22          23          24
-0.40596493 -0.24820644  1.20029275  0.94327297 -0.84716892 -0.88548065
         25          26          27          28          29          30
-1.35010490 -0.58706376  1.42496183 -0.32932555  0.13216412  0.67777703
         31          32
 3.26385449 -0.74538103
```

Fit14와 saturated model을 비교한 결과 G^2, X^2값 모두 임계값보다 작으므로 fit14를 지지하는 귀무가설을 기각할 수 없다. 그에 따라 해당 모델이 데이터를 잘 적합함을 확인할 수 있다.

Fit14는 cig, alc, mar, sex, race, alc*cig, alc*mar, cig*mar, alc*sex, alc*race, sex*mar, sex*race항으로 구성되어 있다. 이는 아래 그림처럼 표현될 수 있다. 즉, cig,alc conditional association과 cig, mar conditional association은 (alc:cig,alc:mar,cig:mar)모델과 fit14에서 동일하다. 이를 통해 cig에서 mar로, cig에서 alc로 가는 path가 sex와 race

를 거치지 않음을 알 수 있으며 collapsibility condition이 성립한다.



마지막으로 fit14의 standardized residual를 살펴보면 2또는 3을 초과하는 값이 31번째 값 뿐이므로 나머지는 모두 잘 적합 되었음을 확인 가능하다.

2. Table 10.3

```
> #table10.3
> u1=c(rep(1,4),rep(2,4),rep(3,4),rep(4,4))
> v1=c(rep(c(1,2,3,4),4))
> count=c(81,68,60,38,24,26,29,14,18,41,74,42,36,57,161,157)
> pre=factor(u1,levels=4:1)
> teen=factor(v1,levels=4:1)
>
> t10.3=data.frame(pre=pre,teen=teen,u1=u1,v1=v1,count=count)
> t10.3
   pre teen u1 v1 count
1    1    1  1  1    81
2    1    2  1  2    68
3    1    3  1  3    60
4    1    4  1  4    38
5    2    1  2  1    24
6    2    2  2  2    26
7    2    3  2  3    29
8    2    4  2  4    14
9    3    1  3  1    18
10   3    2  3  2    41
11   3    3  3  3    74
12   3    4  3  4    42
13   4    1  4  1    36
14   4    2  4  2    57
15   4    3  4  3   161
16   4    4  4  4   157
>
```

```
> model1=glm(count ~ pre + teen + u1:v1, data=t10.3,family=poisson)
> model2=glm(count ~ pre + teen, data=t10.3,family=poisson)
> summary(model1)

Call:
glm(formula = count ~ pre + teen + u1:v1, family = poisson, data = t10.3)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.47349    0.43388   1.091    0.275
pre3        -0.01634    0.12641  -0.129    0.897
pre2         0.10772    0.19883   0.542    0.588
pre1         1.75369    0.23432   7.484 7.20e-14 ***
teen3        1.15514    0.12909   8.948  < 2e-16 ***
teen2        1.41556    0.19962   7.091 1.33e-12 ***
teen1        1.87966    0.24910   7.546 4.50e-14 ***
u1:v1        0.28584    0.02824  10.122  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 431.078  on 15  degrees of freedom
Residual deviance:  11.534  on  8  degrees of freedom
AIC: 118.21

Number of Fisher Scoring iterations: 4
```

```
> qchisq(0.95,df=9)
[1] 16.91898
```

```
> summary(model2)

Call:
glm(formula = count ~ pre + teen, family = poisson, data = t10.3)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.71317    0.07306  64.514  < 2e-16 ***
pre3        -0.85381    0.09026  -9.459  < 2e-16 ***
pre2        -1.48599    0.11483 -12.941  < 2e-16 ***
pre1        -0.50920    0.08051  -6.325 2.54e-10 ***
teen3        0.25529    0.08409   3.036  0.00240 **
teen2       -0.26796    0.09588  -2.795  0.00519 **
teen1       -0.45655    0.10136  -4.504 6.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 431.08  on 15  degrees of freedom
Residual deviance: 127.65  on  9  degrees of freedom
AIC: 232.33

Number of Fisher Scoring iterations: 5

> qchisq(0.95,df=8)
[1] 15.50731
```

Model1, model2모두 해당 모델을 지지하는 귀무가설과 saturated model를 지지하는 대립가설을 검정함을 통해 모델 적합성을 판단할 수 있다. Model1의 경우 G^2값이 임계값인 15.50731보다 작으므로 데이터를 잘 적합하고, model2는 반대로 데이터를 잘 적합하지 못한다.

```
> model2$deviance-model1$deviance
[1] 116.1192
> qchisq(0.95,model2$df.residual-model1$df.residual)
[1] 3.841459
```

두 모델의 deviance를 비교한 lrt 결과 beta 계수가 유의하다. 이를 통해 ordinal association의 존재를 확인할 수 있다. Model1의 summary 에서 나타난 wald test statistic을 통해서도 ordinal association을 나타내는 beta 계수가 유의함을 알 수 있었다. 즉, teenage birth control과 premarital sex의 양의 상관관계가 존재한다.

또한, 해당 beta값이 증가할수록 연관성은 강해지고, u와 v의 score 차이가 커질수록 local OR은 증가한다. 그리고 해당 모델에서는 row score와 column score이 1씩 차이나기에 row a, row a+1, column b, column b+1에서의 local OR은 exp(beta)와 같다. 즉, 모든 인접한 행과 열에서의 local OR이 동일한 uniform association이 확인된다.

```
> #local OR, exp(beta) 비교
> model1$fitted.values
        1         2         3         4         5         6         7
 80.85658  67.65406  69.39574  29.09363  20.75004  23.10650  31.54350
        8         9        10        11        12        13        14
 17.59996  24.39370  36.15178  65.68137  48.77315  32.99969  65.08766
       15        16
157.37940 155.53326
> (80.9*23.1)/(67.6*20.8)
[1] 1.329078
> exp(0.28584)
[1] 1.330879
>
> (65.7*155.5)/(48.8*157.4)
[1] 1.33006
> exp(0.28584)
[1] 1.330879
```