Homework 11 with solutions

The assignment should be typed, with properly labeled computer output. You are encouraged to work together to a reasonable degree, but the write up should be your own. Please make sure you show your work! You should drop-off a copy of your homework on ANGEL in the Homework drop-off box.

1. (20 pts) Fit the model of no second-order interaction (e.g., homogenous associations model) to the following hypothetical $3 \times 3 \times 2$ table of counts

125	7	11	5	106	18
124	6	22	3	109	9
146	6	0	2	111	0

under following conditions and comment on the fit of the model:

(a) Treat the observed zeros as sampling zeros. Make any adjustments necessary to data or the model, fit the model and comment on the model fit.

Solution: If the observed zeros as sampling zeros, it will cause convergence problems sometimes. Therefore we should add a small amount to the zeros before we fit a model. This is done automatically by PROC GENMOD in SAS or glm() in R. We can find that the best model we can fit includes all two-way interactions, i.e. the homogenous association model, but that we do have an issue with the large standard error for the. Also all the other models that include XY term show the similar issue with the estimate and its standard error. The models that do not include this term, show no issues with the fit. The homogenous model fits the best based on the overall fit statistics, e.g., $G^2 = 4.897$, $df = 4$, with p-value 0.29.

Alternatively, we can also add 0.5 to each cell, and refit the models. Notice the change in G^2; but the homogenous model still fits the best. Keep in mind however, that this can oversmooth the data. For example, if we fit the homogenous association model we do not observe any huge standard errors. The original problematic factor XY gets finite estimates as -3.3149, and the expected count for the zero cell (X=2, Y=2, Z=1) will be 0.41 which approximately equal to our assigned value 0.5.

Model	df	G^2	G^2 with 0.5
(x,y,z)	12	849.93	831.7
(xy,z)	8	799.4	787.63
(xy,xz)	6	798.12	786.36
(xy,xz,yz)	4	4.897	4.778

It would be acceptable if you directly replaced zero counts by some small number, such as 10^{-6}, and ran your analysis.

(b) Treat the zeros as structural zeros. Make any adjustments to the data or the model, fit the model and comment on its fit.

Solution:

If the zeros as structural zeros, we should create a dummy variable for every zero count and then fit the model. We include two numerical dummies for them, delta1 and delta2, where each is a vector of 0 counts and a count with 1 for one of the cell with the structural zero, e.g.,

```
> delta1
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
> delta2
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
```

Now the homogenous association model has the following fit statistics, $G^2 = 4.897$ with $df = 3$ with $pvalue = 0.1795$, and the model fits well. The estimated XY term for the zero count is not relevant, but we get the correct degrees of freedom.

Alternatively, We can use NA in R or SAS code to denote a structural zero, and this should also compute the correct degrees of freedom.

If you did not make any changes as above, but kept zero values and ran the model like you did in part (a), then you should have at least adjusted the degrees of freedom. For example, there are total of $T_e = 18$ cells in the table with $Z_e = 2$ two cells with zero. The number of parameters fitted in the homogenous association model is $T_p = 14$ and a number that cannot be estimated due to structural zeros is $Z_p = 1$. The usual $df = 18 - 14 = 4$, but with adjustment $df = (18 - 2) - (14 - 1) = 3$.

SAS Code:

```
/*------------
Problem 2
------------*/
/* prob-2-a */

data prob2;
input x y z count;
datalines;
1 1 1 125
1 2 1 7
1 3 1 11
2 1 1 124
```

```
2 2 1 6
2 3 1 22
3 1 1 146
3 2 1 6
3 3 1 0
1 1 2 5
1 2 2 106
1 3 2 18
2 1 2 3
2 2 2 109
2 3 2 9
3 1 2 2
3 2 2 111
3 3 2 0
;
run;

/* add 0.5 to each cell */
data prob2new;
set prob2;
count=count+0.5;
run;

/*(x,y,z)*/
proc genmod data=prob2new order=data;
class x y z;
model count= x y z / dist=poisson link=log;
run;

/* (xy,xz)*/
proc genmod data=prob2new order=data;
class  x y z;
model count =  x y z x*y x*z/ dist=poisson link=log;
run;

/* (xy,z)*/
proc genmod data=prob2new order=data;
class  x y z;
model count =  x y z x*y / dist=poisson link=log;
run;

/* (xy,xz,yz)*/
```

```
proc genmod data=prob2new order=data;
class  x y z;
model count =  x y z x*y x*z y*z / dist=poisson link=log;
run;

/* prob-2-b */

data prob2b;
set prob2;
delta1 = 0;
delta2= 0;
if count=0 and z=1 then delta1=1;
if count=0 and z=2 then delta2=1;
run;

proc genmod data=prob2b order=data;
class x y z;
model count = x y z x*y x*z y*z delta1 delta2/link=log
dist=poisson lrci type3 obstats;
run;
proc genmod data=prob2b order=data;
class x y z;
model count = x y z x*y x*z y*z delta1/link=log
dist=poisson lrci type3 obstats;
run;
```

R code

```
##Problem 2--
#part a#
X=factor(c(rep("x1",6),rep("x2",6),rep("x3",6)))
Y=factor(c(rep(c("y1","y1","y2","y2","y3","y3"),3)))
Z=factor(c(rep(c("z1","z2"),9)))
count=c(125,5,7,106,11,18,124,3,6,109,22,9,146,2,6,111,0,0)
count=count+0.5
table=xtabs(count~X+Y+Z)
model=glm(count~X+Y+Z,family=poisson(link=log))
summary(model)
model=glm(count~X+Y+Z+X*Y+X*Z,family=poisson(link=log))
summary(model)
model=glm(count~X+Y+Z+X*Y,family=poisson(link=log))
summary(model)
```

```
model=glm(count~X+Y+Z+X*Y+X*Z+Y*Z,family=poisson(link=log))
summary(model)

count=c(125,5,7,106,11,18,124,3,6,109,22,
9,146,2,6,111,0.0000001,0.0000001)

count=count+0.5
model=glm(count~X+Y+Z,family=poisson(link=log))
summary(model)
model=glm(count~X+Y+Z+X*Y+X*Z,family=poisson(link=log))
summary(model)
model=glm(count~X+Y+Z+X*Y,family=poisson(link=log))
summary(model)
model=glm(count~X+Y+Z+X*Y+X*Z+Y*Z,family=poisson(link=log))
summary(model)

#part b#
X=factor(c(rep("x1",6),rep("x2",6),rep("x3",6)))
Y=factor(c(rep(c("y1","y1","y2","y2","y3","y3"),3)))
Z=factor(c(rep(c("z1","z2"),9)))
count=c(125,5,7,106,11,18,124,3,6,109,22,9,146,2,6,111,0,0)
model=glm(count~X+Y+Z+X*Y+X*Z+Y*Z,family=poisson(link=log))
summary(model)

delta1=c(rep(0,16),1,0)
delta2=c(rep(0,16),0,1)
model=glm(count~X+Y+Z+X*Y+X*Z+Y*Z+delta1
+delta2,family=poisson(link=log))
summary(model)

count=c(125,5,7,106,11,18,124,3,6,109,22,9,146,2,6,111,NA,NA)
model=glm(count~X+Y+Z+X*Y+X*Z+Y*Z,family=poisson(link=log))
summary(model)
1-pchisq(4.897,3)
```

2. (25 pts) Children of ages 0-15 were classified according to whether they carried Streptococcus pyogenes and according to the size of their tonsils (you worked on this problem in homework 3!).

   Fit the *linear by linear association model* (see Lesson 6, Part D) and interpret the fit of the model. How does this compare to your results from HW3.

	Tonsil size		
	Normal	Slightly enlarged	Very enlarged
Carrier	19	29	24
Non-carrier	497	560	269

Solution: The linear by linear association model is:

$$\log(\mu_{ij}) = \lambda + \lambda_i^C + \lambda_j^T + \beta u_i v_j,$$

where $i = 1, 2$ and $j = 1, 2, 3$. The fit statistic is $G^2 = .2374$ with df=1 and p-value .6261. This indicates the model fits well at an alpha of .05 level. To compare the linear by linear association model with the independence model, the ΔG^2 is $7.3209 - .2374 = 7.0835$ with df=1 and p-value .0078. At an alpha of .05, the linear by linear association model fits better.

From the results, we can see that $\hat{\beta} = -.4286 < 0$, which indicates a negative association between the two variables. In addition, $\exp(-.4286) = .6514$, which describes that the estimated odds ratio for a unit change in row and column scores of "Carrier-NonCarrier" and "Normal-SlightlyEnlarged" equals $.6514$.

The conclusion of this model is the same as the results from HW3.

SAS Code

```
/*------------
Problem 3
------------*/
data prob3;
length size $17;
input carrier $ size $ count xcarrier ysize;
datalines;
yes Normal 19 1 1
yes Slightly-enlarged 29 1 2
yes Very-enlarged 24 1 3
no Normal 497 2 1
no Slightly-enlarged 560 2 2
no Very-enlarged 269 2 3
;

proc genmod data=prob3 order=data;
class carrier size;
```

```
model count=carrier size xcarrier*ysize / link=log
dist=possion lrci type3 obstats;
run;
```

R code

```
##Problem 3  -- by Minchen Wei##
Carrier=rep(c("yes","no"),c(3,3))
Tonsile=rep(c("Normal","Slightly Enlarged","Very Enlarged"),2)
Carrier=factor(Carrier)
Tonsile=factor(Tonsile)
count=c(19,29,24,497,560,269)
xCarrier=rep(c(1,2),c(3,3))
yTonsile=rep(1:3,2)
table=xtabs(count~Carrier+Tonsile)
model=glm(count~Carrier+Tonsile+
xCarrier*yTonsile,family=poisson(link=log))
summary(model)
```

3. (30pts) The table below contains data from the 2004 General Social Survey on reports of region of residence in 2004 and at age 16.

	Residence in 2004			
Residence at age 16	Northeast	Midwest	South	West
Northeast	425	17	80	36
Midwest	10	555	74	47
South	7	34	771	33
West	5	14	29	452

(a) Fit the symmetry and quasi-symmetry models. Interpret results.

Solution:
First we fit symmetry model and got

```
Null deviance: 4481.79  on 15  degrees of freedom
Residual deviance:  115.61  on  5  degrees of freedom
```

The result indicates that the symmetry model does not fit the data well.
Then we fit quasi-symmetry model and got

```
Null deviance: 4481.7873  on 15  degrees of freedom
Residual deviance:    3.7156  on  2  degrees of freedom
```

The result above indicates that the quasi-symmetry model fits the data well.

(b) Test marginal homogeneity by comparing the fits of these models.

Solution:

To test marginal homogeneity, we have $G^2(\text{marginal homogeneity}) = G^2(\text{symmetry}) - G^2(\text{quasi-symmetry}) = 115.61 - 3.7156 = 111.8944$ with $df = 5 - 3 = 2$. The p-value is almost 0. Therefore, the marginal homogeneity model does not fit.

SAS Code

```
/*------------
Problem 4
------------*/
options ls=90 nodate center nodate;
data residence;
input atage16 $ in2004 $ count;
symm=0;
if atage16=in2004='Neast' THEN symm=1;
if atage16=in2004='Midwest' THEN symm=2;
if atage16=in2004='South' THEN symm=3;
if atage16=in2004='West' THEN symm=4;
if atage16='Neast' AND in2004='Midwest' THEN symm=5;
if atage16='Midwest' AND in2004='Neast' THEN symm=5;
if atage16='Neast' AND in2004='South' THEN symm=6;
if atage16='South' AND in2004='Neast' THEN symm=6;
if atage16='Neast' AND in2004='West' THEN symm=7;
if atage16='West' AND in2004='Neast' THEN symm=7;
if atage16='Midwest' AND in2004='South' THEN symm=8;
if atage16='South' AND in2004='Midwest' THEN symm=8;
if atage16='Midwest' AND in2004='West' THEN symm=9;
if atage16='West' AND in2004='Midwest' THEN symm=9;
if atage16='South' AND in2004='West' THEN symm=10;
if atage16='West' AND in2004='South' THEN symm=10;
datalines;
Neast Neast 425
Neast Midwest 17
Neast South 80
Neast West 36
Midwest Neast 10
```

```
Midwest Midwest 555
Midwest South 74
Midwest West 47
South Neast 7
South Midwest 34
South  South 771
South West 33
West Neast 5
West Midwest 14
West South 29
West West 452
;

/* symmetry */
proc genmod data=residence order=data;
class symm;
model count=symm / link=log dist=poisson lrci type3 obstats;
run;

/* quasi-symmetry */
proc genmod data=residence order=data;
class atage16 in2004 symm;
model count=atage16 in2004 symm / link=log
dist=poisson lrci type3 obstats;
run;
```

R code

```
##Problem 4##
R04=rep(c("N","M","S","W"),c(4,4,4,4))
R16=rep(c("N","M","S","W"),4)
count=c(425,17,80,36,10,555,74,47,7,34,771,33,5,14,29,452)
R04=factor(R04)
R16=factor(R16)

# Create indicator variables for each diagonal cell
iN=(R04=="N")*(R16=="N")
iM=(R04=="M")*(R16=="M")
iS=(R04=="S")*(R16=="S")
iW=(R04=="W")*(R16=="W")
R04_N=(R04=="N")
```

```
R04_M=(R04=="M")
R04_S=(R04=="S")
R04_W=(R04=="W")
R16_N=(R16=="N")
R16_M=(R16=="M")
R16_S=(R16=="S")
R16_W=(R16=="W")
symm1=1*(R04=="N")*(R16=="N")
symm2=2*(R04=="M")*(R16=="M")
symm3=3*(R04=="S")*(R16=="S")
symm4=4*(R04=="W")*(R16=="W")
symm5=5*(R04=="N")*(R16=="M")+5*(R16=="N")*(R04=="M")
symm6=6*(R04=="N")*(R16=="S")+6*(R16=="N")*(R04=="S")
symm7=7*(R04=="N")*(R16=="W")+7*(R16=="N")*(R04=="W")
symm8=8*(R04=="M")*(R16=="S")+8*(R16=="M")*(R04=="S")
symm9=9*(R04=="M")*(R16=="W")+9*(R16=="W")*(R04=="M")
symm10=10*(R04=="S")*(R16=="W")+10*(R16=="S")*(R04=="W")

# Contigency Table
table=xtabs(count~R04+R16)
Contingency_Table=list(Frequency=table,Percent=prop.table(table),
RowPct=prop.table(table,1),ColPct=prop.table(table,2))
Contingency_Table
result=chisq.test(table,correct=FALSE)

# Simple Kappa Coefficient
# Using a function Kappa() in package vcd.
# Please first load packages VR, colorspace and grid and finally
load vcd.
library(vcd)
kappa=Kappa(table)
kappa

options(contrast=c("contr.treatment","contr.poly"))
# Independence Model
model=glm(count~R04+R16,family=poisson(link=log))
summary(model)

# Quasi-Independence Model
model=glm(count~R04+R16+iN+iM+iS+iW,family=poisson(link=log))
summary(model)
```

```
# Symmetry Model
model=glm(count~symm1+symm4+symm6+
symm2+symm5+symm3+symm7+symm8+
symm9+symm10,family=poisson(link=log))
summary(model)

# Quasi-Symmetry Model
model=glm(count~R04_N+R04_M+R04_S+R04_W+
R16_N+R16_S+R16_W+R16_M+
symm1+symm4+symm6+symm2+symm5+symm3+
symm7+symm8+symm9+symm10,family=poisson(link=log))
summary(model)
```