Department of Statistics

COURSE STATS 330/762

Model answer for Assignment 5, 2012

1. *Type the data into R, and make a suitable data frame with factor levels "Yes","No" for smoking and leukoplakia, and levels "None", "0-40 gm", "41-80 gm", " >80 gm" for alcohol intake. Print out the data frame. Also print out a cross-tab like the table above.*
   [10 marks]

The following code constructs a suitable data frame:

```
counts = scan()
26 8 10 8
38 43 8 24
4 14 1 17
1 3 0 7

leuko.df = data.frame(Counts = counts, expand.grid(Smoker = c("Yes","No"),
    Leuko=c("Yes","No"), Alcohol = c("None", "0-40 gm", "41-80 gm",
     ">80 gm")))

    > leuko.df
       Counts Smoker Leuko  Alcohol
    1      26    Yes   Yes     None
    2       8     No   Yes     None
    3      10    Yes    No     None
    4       8     No    No     None
    5      38    Yes   Yes  0-40 gm
    6      43     No   Yes  0-40 gm
    7       8    Yes    No  0-40 gm
    8      24     No    No  0-40 gm
    9       4    Yes   Yes 41-80 gm
    10     14     No   Yes 41-80 gm
    11      1    Yes    No 41-80 gm
    12     17     No    No 41-80 gm
    13      1    Yes   Yes   >80 gm
    14      3     No   Yes   >80 gm
    15      0    Yes    No   >80 gm
    16      7     No    No   >80 gm

    The R function array can be used to turn the data frame into a table:

    leuko.array = array(leuko.df$Counts, c(2,2,4),
       dimnames=list(Smoker = c("Yes","No"),
       Leuko=c("Yes","No"),
       Alcohol = c("None", "0-40 gm", "41-80 gm",  ">80 gm")))
```

```
leuko.array
, , Alcohol = None

      Leuko
Smoker Yes No
   Yes  26 10
   No    8  8

, , Alcohol = 0-40 gm

      Leuko
Smoker Yes No
   Yes  38  8
   No   43 24

, , Alcohol = 41-80 gm

      Leuko
Smoker Yes No
   Yes   4  1
   No   14 17

, , Alcohol = >80 gm

      Leuko
Smoker Yes No
   Yes   1  0
   No    3  7
```

We can re-arrange the rows, columns and slices with the R function aperm:

```
> aperm(leuko.array, c(3,1,2))
, , Leuko = Yes

          Smoker
Alcohol    Yes No
  None      26  8
  0-40 gm   38 43
  41-80 gm   4 14
  >80 gm     1  3

, , Leuko = No

          Smoker
Alcohol    Yes No
  None      10  8
  0-40 gm    8 24
  41-80 gm   1 17
  >80 gm     0  7

A similar result can be obtained using xtabs:

> xtabs(Counts~Alcohol+Smoker+Leuko, leuko.df)
```

Mark allocation: 5 marks for making the data frame, 2 marks for printing it, and 3 marks for printing the table.

2. Fit a suitable Poisson regression model to the contingency table counts. State what your model means in terms of conditional independence, homogeneous association etc. [10 marks]

We will try fitting all possible combinations of interactions, storing the results in a list for easy processing:

```
model.list = vector(length=8, mode="list")

# independence model
model.list[[1]]=glm(Counts~Leuko+Smoker+Alcohol, family = poisson, data=leuko.df)

# Smoking independent of Leuko and Alcohol
model.list[[2]]=glm(Counts~Smoker+Leuko*Alcohol, family = poisson, data=leuko.df)

# Alchol independent of Smoking and Leuko
model.list[[3]]=glm(Counts~Smoker*Leuko+Alcohol, family = poisson, data=leuko.df)

# Leuko independent of Smoking and Alcohol
model.list[[4]]=glm(Counts~Smoker*Alcohol+Leuko, family = poisson, data=leuko.df)

# Leuko and Alcohol conditionally  independent given smoking
model.list[[5]]=glm(Counts~Smoker*Leuko+Smoker*Alcohol , family = poisson,
data=leuko.df)

# Smoking and Alcohol conditionally  independent given Leuko
model.list[[6]]=glm(Counts~Smoker*Leuko+Leuko*Alcohol , family = poisson,
data=leuko.df)

# Smoking and Leuko conditionally  independent given Alcohol
model.list[[7]]=glm(Counts~Smoker*Alcohol+Leuko*Alcohol , family = poisson,
data=leuko.df)

# Homogeneous association
model.list[[8]]=glm(Counts~(Smoker+Leuko+Alcohol)^2 , family = poisson,
data=leuko.df)

# Saturated model
model.list[[9]]=glm(Counts~Leuko*Smoker*Alcohol , family = poisson, data=leuko.df)
```

We can print out a table of AIC values:

```
     formula                                    AIC
 [1,] Counts ~ Leuko + Smoker + Alcohol          128.474
 [2,] Counts ~ Smoker + Leuko * Alcohol          125.072
 [3,] Counts ~ Smoker * Leuko + Alcohol          117.533
 [4,] Counts ~ Smoker * Alcohol + Leuko           99.356
 [5,] Counts ~ Smoker * Leuko + Smoker * Alcohol  88.414
 [6,] Counts ~ Smoker * Leuko + Leuko * Alcohol   114.131
 [7,] Counts ~ Smoker * Alcohol + Leuko * Alcohol 95.954
 [8,] Counts ~ (Smoker + Leuko + Alcohol)^2       87.786
 [9,] Counts ~ Leuko * Smoker * Alcohol           92.415
```

From this it appears that the homogeneous association model is a good fit, although there is some evidence that Leuko and Alcohol are conditionally independent given Smoker. The homogeneous association model is also the one identified by backward elimination.

Mark allocation: 5 marks for identifying the minimum AIC model, 4 marks for mentioning the homogeneous association, 1 for mentioning the conditional independence model.

3. *Fit a logistic regression model to the data using Leukoplakia as the response. Do both smoking and Alcohol have a relationship with the response? (Note that this is modeling the conditional distribution of Leukoplakia, given smoking and alcohol, whereas the model fitted in part 2 is modelling the joint distribution of all three variables. [10 marks]*

The following code will fit the model:

```
> n  = as.vector(apply(leuko.array, c(1,3), sum))
> r = as.vector(leuko.array[,1,])
> logistic.df = data.frame(r=r, n=n, expand.grid(Smoker = c("Yes","No"),
+    Alcohol = c("None", "0-40 gm", "41-80 gm",  ">80 gm")))
>
> model10 = glm(cbind(r,n-r)~Smoker*Alcohol, family=binomial, data=logistic.df)
```

The summary is

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               9.555e-01  3.721e-01   2.568   0.0102 *
SmokerNo                 -9.555e-01  6.233e-01  -1.533   0.1253
Alcohol0-40 gm            6.026e-01  5.383e-01   1.119   0.2629
Alcohol41-80 gm           4.308e-01  1.178e+00   0.366   0.7147
Alcohol>80 gm             2.161e+01  4.820e+04   0.000   0.9996
SmokerNo:Alcohol0-40 gm  -1.949e-02  7.776e-01  -0.025   0.9800
SmokerNo:Alcohol41-80 gm -6.249e-01  1.330e+00  -0.470   0.6384
SmokerNo:Alcohol>80 gm   -2.246e+01  4.820e+04   0.000   0.9996
---
Null deviance: 2.0941e+01  on 7  degrees of freedom
Residual deviance: 3.1675e-10  on 0  degrees of freedom
We can run an anova:

> anova(model10, test="Chi")
Analysis of Deviance Table
              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             7    20.9412
Smoker         1  12.9413         6     7.9999 0.0003214 ***
Alcohol        3   6.6281         3     1.3718 0.0847457 .
Smoker:Alcohol 3   1.3718         0     0.0000 0.7121620


> model11 = glm(cbind(r,n-r)~Smoker, family=binomial, data=logistic.df)
> anova(model11,model10, test="Chi")
Analysis of Deviance Table

Model 1: cbind(r, n - r) ~ Smoker
Model 2: cbind(r, n - r) ~ Smoker * Alcohol
         Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1               6     7.9999
2               0     0.0000  6   7.9999   0.2381
```

Seems like there is no strong evidence of an alcohol effect but that smoking is definitely related.

Mark allocation: 5 marks for fitting the model, 5 marks for a correct conclusion.

4. Do you notice any similarities between the two analyses? [5 marks]

The anova for the saturated model fit for the Poisson regression is

```
Analysis of Deviance Table

                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                    15    187.285
Leuko                 1   18.400        14    168.885 1.791e-05 ***
Smoker                1    6.143        13    162.742 0.0131938 *
Alcohol               3  106.682        10     56.060 < 2.2e-16 ***
Leuko:Smoker          1   12.941         9     43.119 0.0003214 ***
Leuko:Alcohol         3    9.402         6     33.716 0.0243953 *
Smoker:Alcohol        3   32.345         3      1.372 4.427e-07 ***
Leuko:Smoker:Alcohol  3    1.372         0      0.000 0.7121620  ---
```

We see that the three lines in the logistic anova table are the same as the Leuko:Smoker, Leuko:alcohol and Leuko:Smoker:Alcohol lines in the Poisson anova table. Same conclusion for the summary table, except the signs have changed.

Mark allocation: 5 marks for noting the similarity.

5. Can you explain these (Hard!!!) [5 marks]
   See separate sheet for solution. Mark allocation: Not many students will get this. Give 2 marks for an attempt, and full marks if they have tried to connect the joint and conditional probabilities. Be generous!

R hint: Use the functions **`expand.grid`** and **`xtabs`** for Q1.