

Department of Statistics

COURSE STATS 330/762

Model Answers for Assignment 5, 2011

1. *Read the data into R, and make a suitable data frame with factor levels "Yes", "No" for each classifying variable. Print out the data frame. [10 marks]*

The following R code does the trick:

```
> count = c(120,80,4,31,22,24,1,23 )
> picnic.df = data.frame(count,expand.grid(Ill = c("Yes","No"),
+ Potato = c("Yes","No"), Crab = c("Yes","No")))
> picnic.df
  count Ill Potato Crab
1   120 Yes   Yes  Yes
2    80 No    Yes  Yes
3     4 Yes   No   Yes
4    31 No    No   Yes
5    22 Yes   Yes  No
6    24 No    Yes  No
7     1 Yes   No   No
8    23 No    No   No
> xtabs(count~Ill + Potato + Crab, data=picnic.df)
```

```
, , Crab = Yes
```

```
      Potato
Ill   Yes  No
Yes  120   4
No   80  31
```

```
, , Crab = No
```

```
      Potato
Ill   Yes  No
Yes   22   1
No   24  23
```

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

The saturated model is fitted by

```
sat.glm = glm(count~Ill*Potato*Crab, family=poisson, data=picnic.df)
```

We can see which interactions are needed using **step** and **anova**:

```

> anova(sat.glm, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                7      285.974
Ill                                6      285.577  0.528743
Potato                             5      162.375 < 2.2e-16 ***
Crab                                4       68.151 < 2.2e-16 ***
Ill:Potato                          3       15.108 3.263e-13 ***
Ill:Crab                            2        6.387  0.003146 **
Potato:Crab                         1         0.269  0.013378 *
Ill:Potato:Crab                     0          0.000  0.604146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> step(sat.glm, formula(sat.glm), direction = "back", trace=0)

Call:  glm(formula = count ~ Ill + Potato + Crab + Ill:Potato +
Ill:Crab +
      Potato:Crab, family = poisson, data = picnic.df)

Coefficients:
(Intercept)              IllNo          PotatoNo          CrabNo
      4.7914             -0.4152           -3.5297           -1.7219
IllNo:PotatoNo  IllNo:CrabNo  PotatoNo:CrabNo
      2.6026              0.5431              0.8447

Degrees of Freedom: 7 Total (i.e. Null);  1 Residual
Null Deviance:      286
Residual Deviance: 0.2688      AIC: 52.6

```

From this output it looks like the homogeneous association model fits the data well, since all interactions except the 3-way seem to be required. This is the model where the conditional odds ratios between two variables are the same for each level of the third factor (e.g. in this example, the OR in the Ill x Potato tables table corresponding to Crab = Yes is the same as the OR in the Ill x Potato tables table corresponding to Crab = No.

3. Calculate 3 confidence intervals for the 3 conditional odds ratios. [10 marks]

We need to fit the homogeneous association model and examine the summary.

```
> ham.glm = glm(count~Ill*Potato*Crab-Ill:Potato:Crab, family=poisson,
data=picnic.df)
> summary(ham.glm)
```

Call:

```
glm(formula = count ~ Ill * Potato * Crab - Ill:Potato:Crab,
family = poisson, data = picnic.df)
```

Deviance Residuals:

	1	2	3	4	5	6	7
8							
	-0.04274	0.05251	0.24422	-0.08377	0.10066	-0.09507	-0.41071
	0.09842						

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	4.79139	0.09078	52.781	< 2e-16	***
IllNo	-0.41524	0.14316	-2.901	0.00373	**
PotatoNo	-3.52974	0.46637	-7.569	3.77e-14	***
CrabNo	-1.72189	0.22833	-7.541	4.66e-14	***
IllNo:PotatoNo	2.60259	0.48747	5.339	9.35e-08	***
IllNo:CrabNo	0.54313	0.31304	1.735	0.08274	.
PotatoNo:CrabNo	0.84466	0.33908	2.491	0.01274	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 285.97418 on 7 degrees of freedom
Residual deviance: 0.26879 on 1 degrees of freedom
AIC: 52.6

Number of Fisher Scoring iterations: 4

The confidence intervals are

```
> exp( 2.60259 + c(-1,1)* 0.48747 * 1.96) # OR between Ill and Potato
[1] 5.192155 35.094036
> exp( 0.54313 + c(-1,1)* 0.31304 * 1.96) # OR between Ill and Crab
[1] 0.9319945 3.1793870
> exp(0.84466 + c(-1,1)* 0.33908 * 1.96) # OR between Potato and Crab
[1] 1.197293 4.523368
```

From this it looks as though the conditional OR's between Illness and Potato and Crab and Potato are significantly greater than 1, but the same is not true of the conditional OR between Ill and Crab.

4. *What food do you think caused the food poisoning? Is it possible to tell?*
[10 marks]

There are four possible explanations: (a) neither food caused the illness, (b) potato salad caused the illness, (c) crabmeat caused the illness, and (d) both caused the illness.

Let's start by looking at the two-way tables of illness and potato and illness and crabmeat.

```
> xtabs(count~Ill + Potato , data=picnic.df)
```

	Potato	
Ill	Yes	No
Yes	142	5
No	104	54

```
>
```

```
> xtabs(count~Ill + Crab , data=picnic.df)
```

	Crab	
Ill	Yes	No
Yes	124	23
No	111	47

For the first table, the odds ratio (odds of being ill for those who ate potato salad/odds of being ill for those who didn't eat potato salad) is $142*54/(5*104) = 14.7$ so there is a very strong association between being ill and eating potato salad. The equivalent OR between Ill and Crab is $124*47/(23*111) = 2.3$ so this association is less strong. Of course, association is not the same as causation: it could be that persons who ate potato salad were more likely to have eaten something else that caused the illness.

Let's assume that the other foods have been ruled out, so it is one of potato and crab that is the culprit. If potato salad alone is to blame, then the association between illness and potato salad for those who ate crab should be the same as for those who did not eat crab. This is what the homogeneous association model is saying. The common OR for these two groups (ate crab, did not eat crab) is $\exp(2.60259) = 13.5$, which is highly significant. Thus, even for those who did not eat crab, the association between potato and illness is very strong, pointing to potato salad being the culprit.

The association between illness and crab is largely due to the fact that those who ate crab also tended to eat potato salad – the marginal table is

```
> xtabs(count~Potato + Crab , data=picnic.df)
```

	Crab	
Potato	Yes	No
Yes	200	46
No	35	24

and the OR is $200*24/(35*46) = 2.98$. Thus, the observed data are consistent with potato salad being the culprit.

Could crab alone have been to blame? If this were the case, then the OR's between Illness and crab conditional on potato salad would be the same, and significantly different from 1. But this is not the case, as the CI we calculated in Q3 included 1. Thus it seems unlikely that crab alone was to blame.

Could they both have been to blame? In this case, there should be a significant association between illness and crab in the group that did not eat potato salad . Let's examine this table:

```
> xtabs(count~Ill + Crab + Potato, data=picnic.df)
, , Potato = Yes
```

	Crab	
Ill	Yes	No
Yes	120	22
No	80	24

```
, , Potato = No
```

	Crab	
Ill	Yes	No
Yes	4	1
No	31	23

Of the 35 persons who ate crab, only 4 became ill. This compares with the 24 persons who ate neither crab nor potato of whom 1 became ill. Thus it was possible to become ill without eating either of the suspect foods. How does 4/35 compare to 1/24? The numbers are a bit small to run a Poisson regression. We can use Fisher's exact test which is OK for small numbers. Applying this to the "potato no" table, we get

```
> cond.table = xtabs(count~Ill + Crab + Potato, data=picnic.df)
> fisher.test(cond.table[, , 2])
# cond.table[, , 2] is the "potato no" 2x2 table
```

Fisher's Exact Test for Count Data

```
data: cond.table[, , 2]
p-value = 0.6392
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2653918 152.2799210
sample estimates:
odds ratio
 2.920307
```

Thus, it seems that there is no evidence that crab is the culprit, and that potato alone is responsible.