Department of Statistics

COURSE STATS 330

Model answers for Assignment 5, 2007

Question 1

1. *Type the data into R and construct a data frame suitable for fitting log-linear models. Print out the data frame. [5 marks]*

```
> counts=c(1105, 411111, 4624,157342, 14, 483, 497, 1008)
> accident.df = data.frame(Counts=counts, expand.grid(Ejected = c("Yes",
     "No"), Seatbelt = c("Yes", "No"), Injury=c("Nonfatal", "Fatal")))
> accident.df
  Counts Ejected Seatbelt   Injury
1   1105     Yes      Yes Nonfatal
2 411111      No      Yes Nonfatal
3   4624     Yes       No Nonfatal
4 157342      No       No Nonfatal
5     14     Yes      Yes    Fatal
6    483      No      Yes    Fatal
7    497     Yes       No    Fatal
8   1008      No       No    Fatal
```

2. *Fit a suitable log-linear model to the data. Are there any cells that are not fitted well by the model? [5 marks]*

If we fit the saturated model and do an anova, we obtain

```
> accident.glm = glm(Counts~Ejected*Seatbelt*Injury, data=accident.df,
+ family=poisson)
> anova(accident.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: Counts
Terms added sequentially (first to last)
                        Df Deviance Resid. Df  Resid. Dev   P(>|Chi|)
NULL                                        7     1624865
Ejected                  1   729871         6      894994           0
Seatbelt                 1   111458         5      783536           0
Injury                   1   772092         4       11444           0
Ejected:Seatbelt         1     7877         3        3568           0
Ejected:Injury           1     2423         2        1145           0
Seatbelt:Injury          1     1142         1           3  2.741e-250
Ejected:Seatbelt:Injury  1        3         0 -5.010e-11   9.115e-02
```

This suggests that the three-way interaction term is not significant, and that the homogeneous association model is appropriate for these data. Fitting this model, and using the summary function, we get

```
> ham.glm = glm( counts~(Ejected+Seatbelt+Injury)^2, data=accident.df,
+ family=poisson)
> summary(ham.glm)

Call:
glm(formula = counts ~ (Ejected + Seatbelt + Injury)^2, family = poisson,
    data = accident.df)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.20704  -0.01071  -0.10095   0.01731  -1.59987   0.31400   0.30951  -0.21583

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             7.00137    0.02992  233.99   <2e-16 ***
EjectedNo               5.92527    0.02996  197.76   <2e-16 ***
SeatbeltNo              1.43913    0.03321   43.33   <2e-16 ***
InjuryFatal            -3.96315    0.06944  -57.07   <2e-16 ***
EjectedNo:SeatbeltNo   -2.39964    0.03334  -71.97   <2e-16 ***
EjectedNo:InjuryFatal  -2.79779    0.05526  -50.63   <2e-16 ***
SeatbeltNo:InjuryFatal  1.71732    0.05402   31.79   <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: 1.6249e+06  on 7  degrees of freedom
Residual deviance: 2.8540e+00  on 1  degrees of freedom
AIC: 93.853

Number of Fisher Scoring iterations: 3
```

None of the Deviance residuals are large, indicating that all cells are reasonably well fitted.

3. *Compute the conditional odds ratio between seatbelt use and injury, conditional on Ejected. Do these odds ratios depend on being ejected from the car? What effect does seatbelt use have on the type of injury? Also compute a 95% confidence interval for the conditional odds ratio. [10 marks]*

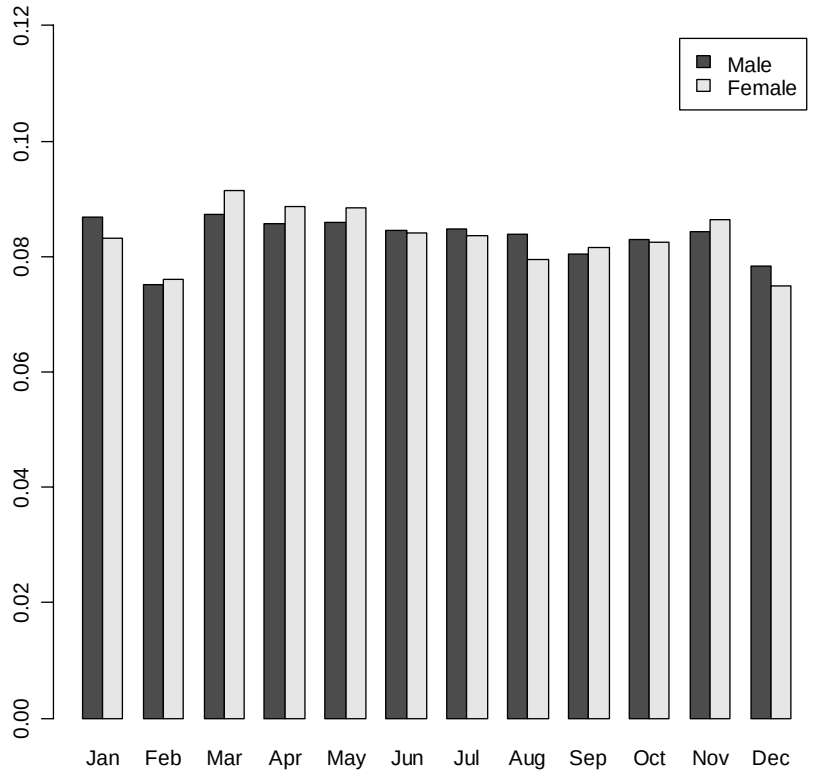
The homogeneous association model implies that the odds ratios between Injury and Seatbelt in the separate tables for Eject = “yes” and Eject=”No” are the same, ant that the log-odds ratio is the coefficient for the Seatbelt-Injury interaction. The estimated coefficient is 1.71732, so the odds ratio in both tables is exp(1.71732) = 5.569582. Thus, the odds of an accident being fatal are about 5 and a half times more if a seat belt is not worn. A 95% confidence interval is obtained by

```
> exp(confint(ham.glm))
Waiting for profiling to be done...
                             2.5 %       97.5 %
(Intercept)            1.034985e+03 1.163800e+03
EjectedNo              3.532271e+02 3.972509e+02
SeatbeltNo             3.952975e+00 4.502698e+00
InjuryFatal            1.656784e-02 2.175183e-02
EjectedNo:SeatbeltNo   8.497287e-02 9.683767e-02
EjectedNo:InjuryFatal  5.472342e-02 6.796101e-02
SeatbeltNo:InjuryFatal 5.013605e+00 6.196228e+00

i.e. (5.013,6.196) to 3 dp.
```

Q2. The following code makes the data frame and plots the relative frequencies of the various months separately for males and females:

```
counts = scan()
3755 3251 3777 3706 3717 3660 3669 3626 3481 3590 3650 3392
1362 1244 1496 1452 1448 1376 1370 1301 1337 1351 1416 1226
heights = matrix(counts, 2,12, byrow=T)
relheights = heights/apply(heights, 1, sum)
barplot(relheights, beside=T, legend.text = T, ylim=c(0,0.12))
```



We see from the graph that the distribution over the months is similar for males and females. However, the months seem to differ: there is a dip in December and Februrary, and a peak in March.

Are these observed differences due to chance? We can see if the distributions are the same for males and females by testing if gender and month are independent, or, alternatively if the interactions are zero in the Poisson model. We get

```
> suicide.df = data.frame(counts=counts, expand.grid(month= c("Jan",
"Feb", "Mar", "Apr", "May", "Jun", "Jul",
+ "Aug", "Sep", "Oct", "Nov", "Dec"), sex=c("Male","Female")))
>
> anova(glm(counts~month*sex, data=suicide.df, family=poisson),
test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: counts
Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         23    12704.5
month     11    118.4        12    12586.2 3.886e-20
sex        1  12574.2        11       12.0       0.0
month:sex 11     12.0         0  4.441e-14       0.4
```

The p-value corresponding to a deviance of 12.0 in 11 df is 0.3636 indicating that the distributions are the same for both sexes. However, the model with gender alone has a deviance of 130.35 on 22 df, corresponding to a p-value of 0.000, so that there is a significant difference between the months.