Department of Statistics

COURSE STATS 330

Model answers for Assignment 5, 2005

Note: No data is supplied for this assignment. You must type your own and construct suitable data frames using the `rep` or `expand.grid` functions.

Question 1.

The data in table 1 come from a classic British study into the effect of smoking on the incidence of coronary heart disease and lung cancer.

1. *Make a data frame containing the data and fit a model to explain the coronary death rate (in deaths per 1000 person years) in terms of age and smoking. Make a table of the rates. [10 marks]*

To make a data frame **`smoke.df`** containing the data, we can type

```
count<-c(2,12,28,28,31,32,104,206,186,102)
exposure<-c(18793,10673,5710,2585,1462,52407,43248,28612,12663, 5317)
smoke<-factor(rep(0:1,c(5,5)))
age<-factor(rep(c("35-44","45-54","55-64","65-74","75-84"),2))
smoke.df<-data.frame(count,exposure,smoke,age)
```

giving

```
> smoke.df
    age smoke exposure count
1  35-44     0    18793     2
2  45-54     0    10673    12
3  55-64     0     5710    28
4  65-74     0     2585    28
5  75-84     0     1462    31
6  35-44     1    52407    32
7  45-54     1    43248   104
8  55-64     1    28612   206
9  65-74     1    12663   186
10 75-84     1     5317   102
```

To fit the model to calculate rates per 1000, we use log(exposure/1000) as an offset. The code is

```
smoke.glm<-glm(count~smoke*age,  family=poisson,  offset=log(exposure/1000),
data=smoke.df)
summary(smoke.glm)
anova(smoke.glm)
```

This results in

```
Call:
glm(formula = count ~ smoke * age, family = poisson, data = smoke.df,
    offset = log(exposure/1000))

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.2403     0.7071  -3.168  0.00153 **
smoke1           1.7470     0.7289   2.397  0.01653 *
age45-54         2.3575     0.7638   3.087  0.00202 **
age55-64         3.8303     0.7319   5.233 1.67e-07 ***
age65-74         4.6228     0.7319   6.316 2.68e-10 ***
age75-84         5.2945     0.7296   7.257 3.95e-13 ***
smoke1:age45-54 -0.9868     0.7901  -1.249  0.21167
smoke1:age55-64 -1.3630     0.7562  -1.802  0.07148 .
smoke1:age65-74 -1.4424     0.7565  -1.907  0.05656 .
smoke1:age75-84 -1.8472     0.7572  -2.440  0.01471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9.3509e+02  on 9  degrees of freedom
Residual deviance: 3.5083e-14  on 0  degrees of freedom
AIC: 75.068

Number of Fisher Scoring iterations: 3
>
> anova(smoke.glm, test="Chisq")
Analysis of Deviance Table


          Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                         9     935.09
smoke      1    29.10         8     905.99  6.874e-08
age        4   893.86         4      12.13 3.570e-192
smoke:age  4    12.13         0  3.508e-14       0.02
```

The interaction terms are clearly required.

To get the predicted rates, we need to subtract the offsets from the predicted log-mean, then exponentiate:

```
Rate1<-exp(predict(smoke.glm) - log(smoke.df$exposure/1000))
```

Since this is a saturated model, we can also use

```
Rate2<-smoke.df$count/(smoke.df$exposure/1000)
```

Arrange in a table:

```
Result<-matrix(rate1,5,2)
Dimnames(Result)<-list(levels(smoke.df$age), c("Nonsmokers","Smokers"))
 > Result
```

```
     Nonsmokers    Smokers
35-44  0.1064226  0.6106055
45-54  1.1243324  2.4047355
55-64  4.9036778  7.1997763
65-74 10.8317215 14.6884624
75-84 21.2038304 19.1837502
```

2. *Does smoking have an effect on coronary death rates? If so, does the effect depend on age? [5 marks]*

Since the anova table indicates that both smoking (p=0.01653) and the interaction with age (p=0.02) are significant, smoking has an effect and the effect depends on age.

3. *If you find that it does, can you think of an intuitive reason why this might be so? [5 marks]*

Smokers have a higher death rate than non-smokers but the difference gets less with age (and actually seems to reverse at the highest age group. This could be that all but the most coronary resistant smokers have died off by the time they get to 80 years of age.

Question 2

A survey to assess attitudes to AIDS education posed the following questions; (1) Do you think the Government should pay all health-related costs for AIDs victims (2) Do you think the government should allocate funds to promote safe sex practices. Table 2 gives the answers of 621 respondents to this questionnaire, classified by gender.

1. *Make a data frame and fit a suitable log-linear model to these data. Do we require a saturated model or is a smaller model adequate? [10 marks]*

```
> counts<-c(76,6,114,11,160,25,181,48)
> gender<-rep(c("M","F"),c(2,2))
> fund<-rep(c("Yes","No"),4)
> pay<-rep(c("Yes","No"),c(4,4))
>
> AIDS.df<-data.frame(counts, gender, fund, pay)
>
> AIDS.glm<-glm(counts~gender*fund*pay, family=poisson, data=AIDS.df)
> summary(AIDS.glm)

Call:
glm(formula = counts ~ gender * fund * pay, family = poisson,
    data = AIDS.df)
```

```
Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.87120    0.14434  26.820  < 2e-16 ***
genderM                 -0.65233    0.24664  -2.645  0.00817 **
fundYes                  1.32730    0.16235   8.175 2.95e-16 ***
payYes                  -1.47331    0.33428  -4.407 1.05e-05 ***
genderM:fundYes          0.52900    0.26946   1.963  0.04962 *
genderM:payYes           0.04619    0.56428   0.082  0.93476
fundYes:payYes           1.01101    0.35502   2.848  0.00440 **
genderM:fundYes:payYes -0.32833    0.59339  -0.553  0.58005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  4.4582e+02  on 7  degrees of freedom
Residual deviance: -7.5495e-15  on 0  degrees of freedom
AIC: 61.382

Number of Fisher Scoring iterations: 3

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

                Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                7     445.82
gender           1    12.23         6     433.59  4.706e-04
fund             1   346.94         5      86.66  1.967e-77
pay              1    70.34         4      16.32  4.996e-17
gender:fund      1     3.20         3      13.12       0.07
gender:pay       1     1.45         2      11.67       0.23
fund:pay         1    11.37         1       0.30  7.484e-04
gender:fund:pay  1     0.30         0 -7.550e-15       0.58
```

Seems like the 3 way interactions are insignificant. If we fit the homogeneous association model (ie no 3-way interaction), then the residual deviance p-value is 0.5834314. The model

```
counts ~ gender + fund + pay + fund:pay + gender:fund
```

also fits well with a residual deviance p-value of 0.3037501. Finally, fitting the model

```
counts ~ gender  +  inf * health
```

gives a residual deviance of 5.581 on 3 df, with pvalue 0.1338734, so this model also seems OK. This model assumes that the conditional OR between gender and pay, given fund is 1.

2. *Describe your model(s) in terms of conditional odds ratios and independence. [5 marks]*

The homogeneous association model i.e. the model `counts~gender*fund*pay-gender:fund:pay` implies that the conditional odds ratios between `gender` and `pay` given `fund` are the same. The model `counts~gender + fund*pay` is the model where gender is independent of the other two factors, and the model `counts ~ gender + fund + pay + fund:pay + gender:fund` is the model of conditional independence of gender and fund, given pay.

3. *Give a confidence interval for the conditional odds ratio between `gender` and `pay`, conditional on `fund`. [5 marks]*

The homogeneous association model gives

```
>   AIDS3.glm<-glm(counts~gender*fund*pay-gender:fund:pay,   family=poisson,
data=AIDS.df)
> summary(AIDS3.glm)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.8521     0.1415  27.219  < 2e-16 ***
genderM         -0.5976     0.2242  -2.666  0.00768 **
fundYes          1.3514     0.1575   8.578  < 2e-16 ***
payYes          -1.3750     0.2750  -5.001 5.71e-07 ***
genderM:fundYes  0.4636     0.2406   1.927  0.05401 .
genderM:payYes  -0.2516     0.1749  -1.438  0.15035
fundYes:payYes   0.8997     0.2852   3.155  0.00160 **

Null deviance: 445.82335  on 7  degrees of freedom
Residual deviance:   0.30072  on 1  degrees of freedom
AIC: 59.683
```

The estimate of the gender-pay interaction is -0.2516, with a se of 0.1749, so a 95% CI for the log-OR is -0.2516 +/- 1.96*0.1749. We get

```
> -0.2516 +c(-1,1)*1.96*0.1749
[1] -0.594404  0.091204
> exp(-0.2516 +c(-1,1)*1.96*0.1749)
[1] 0.5518914 1.0954925
```

so that the CI for the conditional OR is (0.5518914, 1.0954925).

Basing the CI on the model without the gender-fund interaction gives an interval of (0.5783792, 1.1400426), a very similar answer.