STATISTICS 538, Lecture #11

More Log-Linear Models

November 29, 2010

The 'eyegrade' data, 7477 women cross-classified according to left-eye vision and right-eye vision

matched pairs - really eyes are the 'units', but they come in pairs!

```
> xtabs(y~left+right, data=eyegrade)
```

	y	right	left
1	1520	best	best
	266	best	second
	⋮		
16	492	worst	worst

left \ right	best	second	third	worst
best	1520	234	117	36
second	266	1512	362	82
third	124	432	1772	179
worst	66	78	205	492

Log-linear model with all main effect and interactions

saturated!

16 datapoints,
16 coefficients

```
> summary(ft1)
glm(formula = y ~ .^2, family = poisson, data = eyegrade)

Coefficients:
                          Estimate Std. Error
(Intercept)                7.32647    0.02565
rightsecond               -1.87114    0.07022
rightthird                -2.56429    0.09594
...
leftworst                 -3.13681    0.12574
rightsecond:leftsecond     3.60884    0.09671
rightthird:leftsecond      2.87244    0.12541
...
rightworst:leftworst       5.75177    0.21359
```

1

3+3

9

Notation for the saturated model? Interesting sub-models?

$L = \{0, 1, 2, 3\}$, $R = \{0, 1, 2, 3\}$

best worst

$$E\{Y_{\ell r}\} = \exp\{\beta_0 + \beta^L_\ell + \beta^R_r + \beta^{LR}_{\ell r}\}$$

1 3 3 9

'ref. level constraints'

$\beta^L_0 = 0$, $\beta^R_0 = 0$

$\beta^{LR}_{0r} = 0$,

$\beta^{LR}_{\ell 0} = 0$

i.e. $p_{\ell r} = \Pr\{L = \ell, R = r\}$

$$\propto e^{\beta^L_\ell + \beta^R_r + \beta^{LR}_{\ell r}}$$

Independence $p_{\ell r} = \Pr\{L = \ell\}\Pr\{R = r\} \Leftrightarrow \boxed{\beta^{LR} = 0}$

df: 16 ↓ 7

Symmetry $p_{ab} = p_{ba}$ for all a, b i.e. $\Pr\{L = a, R = b\} = \Pr\{L = b, R = a\}$

$\boxed{\beta^L = \beta^R, \ \beta^{LR} \text{ symmetric matrix}}$ ← df: 16 ↓ 10

Independence?

```
> anova(ft1)
Analysis of Deviance Table

Model: poisson, link: log

Response: y
Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev
NULL                          15     8692.3
right       3   1047.9        12     7644.4
left        3    972.9         9     6671.5
right:left  9   6671.5         0  -8.66e-15
```

reject H_0: left vision ind. of right vision

Symmetry? Approach differently than text

text: define a new predictor (with 10 levels) giving unordered pairs of levels i.e. L=2, R=1 and R=2, L=1 give the same value of this predictor

$\beta^L : \beta^R$ β_1^L β_2^L β_1^R β_2^R

```
> cntrsts
[1,] 0 1 0 0 -1  0  0 0 0 0  0 0 0  0 0 0
[2,] 0 0 1 0  0 -1  0 0 0 0  0 0 0  0 0 0
[3,] 0 0 0 1  0  0 -1 0 0 0  0 0 0  0 0 0
[4,] 0 0 0 0  0  0  0 0 1 0 -1 0 0  0 0 0
[5,] 0 0 0 0  0  0  0 0 0 1  0 0 0 -1 0 0
[6,] 0 0 0 0  0  0  0 0 0 0  0 0 1  0 -1 0
```

with respect to a bigger model with coefficients $\underset{\sim}{\beta}$

represent null as $H_0 : \underset{6\times 16}{C}\ \underset{16\times 1}{\underset{\sim}{\beta}} = \underset{6\times 1}{\underset{\sim}{0}}$

β^{LR}_{34} β^{LR}_{43}

β^{LR} symmetric

Symmetry, continued

```
> library(gregmisc)  ### for estimable()
> estimable(ft1, cm=cntrsts, beta0=rep(0,6),
            joint.test=T)

   X2.stat DF   Pr(>|X^2|)
1 18.81970  6 0.004479225
```

C

$\underset{\sim}{0}$

16 6 10

reject symmetry hypothesis

What about quasi-symmetry?

just the rows for $\beta^{LR}_{ab} = \beta^{LR}_{ba}$

$(0,0,0)$

```
> estimable(ft1, cm=cntrsts[4:6,], beta0=rep(0,3),
             joint.test=T)

   X2.stat DF Pr(>|X^2|)
1 7.224664  3 0.06507147
```

Data from a 20-year cohort study involving 1314 women

FREQ in 1992 in 1972

```
> femsmoke
    y smoker dead   age
1   2    yes  yes 18-24
2   1     no  yes 18-24
3   3    yes  yes 25-34
4   5     no  yes 25-34
5  14    yes  yes 35-44
...
26 28     no   no 65-74
27  0    yes   no   75+
28  0     no   no   75+
```

7 categories

2×2×7

'Poster-child' for Simpson's paradox

```
> xtabs(y~smoker+dead, data=femsmoke)
      dead
smoker yes  no
   yes 139 443
   no  230 502
```

20 year mortality

← 24% for smokers

← 31% for non-smokers

```
> xtabs(y~smoker+dead, subset=(age=="35-44"),
        data=femsmoke)
      dead
smoker yes  no
   yes  14  95
   no    7 114
```

← 13% mortality for smokers

← 6% for non-smokers

in 1972 smoking and age negatively correlated i.e. smoking more popular amongst younger women

Log-linear model with all main effect and interactions

not quite saturated

```
glm(formula = y ~ .^2, family = poisson, data = femsmoke)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.54284    0.58736   0.924 0.355384
smokerno         -0.29666    0.25324  -1.171 0.241401
deadno            3.43271    0.59014   5.817 6.00e-09 **
age25-34          0.92902    0.68381   1.359 0.174273
...
smokerno:deadno   0.42741    0.17703   2.414 0.015762 *
...
deadno:age65-74  -5.08798    0.61951  -8.213  < 2e-16 **
deadno:age75+   -27.31727 8839.01146  -0.003 0.997534
---
    Null deviance: 1193.9378  on 27  degrees of freedom
Residual deviance:    2.3809  on  6  degrees of freedom
AIC: 180.58
```

Drop a term??? Scientific plausibility?

recall AB interaction coefficients are main effects in (A|B,C) and (B|A,C) models

absence of AB interaction = conditional independence of A and B given C

```
> drop1(ft, test="Chi")
Single term deletions

Model:
y ~ (smoker + dead + age)^2
            Df Deviance    AIC    LRT Pr(Chi)
<none>             2.38 180.58
smoker:dead  1     8.33 184.52   5.95 0.01475 *
smoker:age   6    92.63 258.83  90.25 < 2e-16 ***
dead:age     6   632.30 798.49 629.92 < 2e-16 ***
```

What is the null when we test the smoker:dead interaction?

smoking and 20-year mortality unassociated given age

No evidence that we need to add three-way interactions - what is the corresponding interpretation?

no 3-way interaction implies:

odds ratio for smoking and mortality given age is the same across age groups

So estimate odds (20-yr mortality) to be $e^{.427} = 1.53$ times higher for smokers than non-smokers, within each age band