Loglinear models

3-way

Demetris Athienitis

Introduction

There are many different types of associations that can exist with three variables (in a 3-way contingency table)

- Two-way and three-way interactions.
- Conditional and marginal independencies.

Loglinear models can be used to describe associations among all three variables.

Definition (Associations)

- X, Y, Z are *mutual independent*, (X, Y, Z) if $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

- Y is *jointly independent* of X and Z, (XZ, Y) if $\pi_{ijk} = \pi_{+j+}\pi_{i+k}$

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

- X and Y are *conditionally independent* given Z, (XZ, YZ) if $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

Definition (Associations - continued)

- *Homogeneous association*, (XZ, XY, YZ) if two variables have the same association for all levels of the third, e.g. $\pi_{ij|k} = \pi_{ij|k'}$ same $\forall k, k'$

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

- *Non restricted association*, (saturated model) (XYZ)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$$

Example

Consider a $2 \times 2 \times 2$ with X, Y conditional independence (XZ, YZ)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- X and Y are conditionally independent given Z:

$$\log\left(\theta_{XY(k)}\right) = \log\left(\frac{\mu_{ijk}\mu_{i'j'k}}{\mu_{i'jk}\mu_{ij'k}}\right) = \cdots = 0 \Longrightarrow \theta_{XY(k)} = 1$$

- The $X - Z$ odds ratio is the same at all levels of Y:

$$\log\left(\theta_{X(j)Z}\right) = \log\left(\frac{\mu_{ijk}\mu_{i'jk'}}{\mu_{i'jk}\mu_{ijk'}}\right) = \cdots = \lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ}$$

- Similarly, $Y - Z$ odds ratio same at all levels of X. Model has no three-factor interaction.

Example

Consider the loglinear homogeneous association model denoted (XY, XZ, YZ)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

Each pair of variables is conditionally dependent, but association (as measured by odds ratios) is the same at all levels of third variable.

Example (Teen substance usage)

A survey of 2276 high school seniors

```
> ftable(teens, row.vars=c("alc","cigs"))
         mj yes  no
alc cigs
yes yes     911 538
    no       44 456
no  yes       3  43
    no        2 279

> teens.AC.AM.CM =  glm(Freq ~ alc*cigs + alc*mj + cigs*mj,
+       family=poisson, data=teens.df)
> summary(teens.AC.AM.CM)
---
    Null deviance: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
```

Example (continued)

```
> X2=sum(residuals(teens.AC.AM.CM,type="pearson")^2);X2
[1] 0.4011005
> 1-pchisq(X2,1)
[1] 0.5265215
```

(AC, AM, CM) model fits well with $G^2 = 0.37$ (and $X^2 = 0.4$) on 1 df.

Equivalently G^2 done via,

```
> teens.ACM <- update(teens.AC.AM.CM, . ~ alc*cigs*mj)
> anova(teens.AC.AM.CM, teens.ACM, test="Chisq")
Model 1: Freq~alc * cigs + alc * mj + cigs * mj
Model 2: Freq~alc * cigs * mj
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1    0.37399
2         0    0.00000  1  0.37399   0.5408
```

Example (continued)

Next we check if any 2-way interactions can be removed

```
> drop1(teens.AC.AM.CM, test="Chisq")
Single term deletions

Model:
Freq ~ alc * cigs + alc * mj + cigs * mj
         Df Deviance    AIC    L.R.T.  Pr(>Chi)
<none>           0.37  63.42
alc:cigs  1   187.75 248.80 187.38 < 2.2e-16 ***
alc:mj    1    92.02 153.06  91.64 < 2.2e-16 ***
cigs:mj   1   497.37 558.41 497.00 < 2.2e-16 ***
```

Example (continued)

To test for conditional independence of A and C given M

```
> teens.AM.CM <- update(teens.AC.AM.CM, . ~ alc*mj + cigs*mj)
> anova(teens.AM.CM, teens.AC.AM.CM, test="Chisq")
Analysis of Deviance Table

Model 1: Freq ~ alc + mj + cigs + alc:mj + mj:cigs
Model 2: Freq ~ alc * cigs + alc * mj + cigs * mj
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         2    187.754
2         1      0.374  1   187.38 < 2.2e-16 ***
```

Example (continued)

We can also get predicted counts under a variety of models and compare them to the actual data/saturated model

```
> table.7.4
  alc cigs  mj (A,C,M) (AC,M) (AM,CM) (AC,AM,CM) (ACM)
1 yes  yes yes   540.0  611.0  909.00     910.00   911
2 yes  yes  no   740.0  838.0  439.00     539.00   538
3 yes   no yes   282.0  211.0   45.80      44.60    44
4 yes   no  no   387.0  289.0  555.00     455.00   456
5  no  yes yes    90.6   19.4    4.76       3.62     3
6  no  yes  no   124.0   26.6  142.00      42.40    43
7  no   no yes    47.3  119.0    0.24       1.38     2
8  no   no  no    64.9  162.0  180.00     280.00   279
```

Example (continued)

```
> summary(teens.AC.AM.CM)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.63342    0.05970  94.361  < 2e-16 ***
alcyes           0.48772    0.07577   6.437 1.22e-10 ***
cigsyes         -1.88667    0.16270 -11.596  < 2e-16 ***
mjyes           -5.30904    0.47520 -11.172  < 2e-16 ***
alcyes:cigsyes   2.05453    0.17406  11.803  < 2e-16 ***
alcyes:mjyes     2.98601    0.46468   6.426 1.31e-10 ***
cigsyes:mjyes    2.84789    0.16384  17.382  < 2e-16 ***
---
    Null deviance: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417
```

Example (continued)

(AC, AM, CM) model, AC odds-ratio is the same at each level of M. $1 =$ yes and $2 =$ no for each variable, the estimated conditional AC odds ratio is

$$\frac{\hat{\mu}_{11k}\hat{\mu}_{22k}}{\hat{\mu}_{12k}\hat{\mu}_{21k}} = \exp(\hat{\lambda}_{11}^{AC} + \cancelto{0}{\hat{\lambda}_{22}^{AC}} - \cancelto{0}{\hat{\lambda}_{12}^{AC}} - \cancelto{0}{\hat{\lambda}_{21}^{AC}}) = e^{2.0545} = 7.8$$

A 95% CI is

$$e^{2.05 \mp (1.96)(0.174)} \longrightarrow (5.5, 11.0)$$

Common odds-ratio is reflected in the fitted values for the model:

$$\frac{(910)(1.38)}{(44.6)(3.62)} = 7.8 \qquad \frac{(539)(280)}{(455)(42.4)} = 7.8$$

Similar results hold for AM and CM conditional odds-ratios in this model.

Example (continued)

(AM, CM) model, $\lambda_{ij}^{AC} = 0$, and conditional AC odds-ratio (given M) is $e^0 = 1$ at each level of M, i.e., A and C are conditionally independent given M. Again, this is reflected in the fitted values for this model.

$$\frac{(909)(0.24)}{(45.8)(4.76)} = 1 \qquad \frac{(439)(180)}{(555)(142)} = 1$$

The AM odds-ratio is not 1, but it is the same at each level of C:

$$\frac{(909)(142)}{(439)(4.76)} = 61.87 \qquad \frac{(45.8)(180)}{(555)(0.24)} = 61.87$$

Similarly, the CM odds-ratio is the same at each level of A:

$$\frac{(909)(555)}{(439)(45.8)} = 25.14 \qquad \frac{(4.76)(180)}{(142)(0.24)} = 25.14$$

Remarks

- Loglinear models extend to any number of dimensions
- Loglinear models treat all variables symmetrically. Logistic regression models treat Y as response and other variables as explanatory. More natural approach when there is a single response.
- For modeling ordinal associations consider a 2-way table with assigned
    - row scores $u_1 \leq u_2 \leq \cdots \leq u_I$
    - column scores $v_1 \leq v_2 \leq \cdots \leq v_J$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$$

  where $\beta u_i v_j$ takes the role of λ_{ij}^{XY} but only 1 parameter is used, i.e. only 1 degree of freedom taken up, instead of $(I-1)(J-1)$

Residuals

Checking residuals is always important and done in the usual way as with any GLM, however a new graphical visualization may also be useful

R

In `mosaic{vcdExtra}`

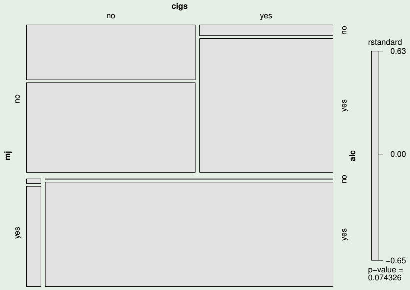
```
mosaic(glm object,...)
```

is capable of a mosaic plot of the residuals, where the area of each tile is proportional to the corresponding cell entry, given the dimensions of previous splits.

Example (Teen substance usage continued)

Getting and visualizing the standardized deviance residuals

```
rstandard(teens.AC.AM.CM)
      1       2       3       4       5       6       7       8
 0.6332 -0.6334 -0.6347  0.6331 -0.6527  0.6317  0.5933 -0.6335
> mosaic(teens.AC.AM.CM,~mj+cigs+alc,residuals_type = "rstandard")
```



We learned

- Loglinear models extended to 3-way tables and how 3 variables can be associated
- Looked at a $2 \times 2 \times 2$ example (where odds ratios calculated since each dimension was 2)
- ALWAYS look at residuals in any model