Log-linear models and conditional independence

G. Marchetti

2024-05-13

Log-linear model and factorization

A joint probability vector for the model of conditional independence: X_2 independent of X_3 given X_1.

```
u1 <-  0.2
u2 <- -0.2
u3 <-  1.2
u12 <- 0.8
u13 <- 0.5
u23 <- 0
u123 <-0
lam <- c(0, u1, u2, u12, u3, u13, 0, 0)

L <- matrix(c(1,1,0,1), 2, 2)
M <- L  %x% L %x% L
p <- exp(M %*% lam)
p <- p/sum(p)
p
```

```
           [,1]
[1,] 0.03314280
[2,] 0.04048071
[3,] 0.02713503
[4,] 0.07376067
[5,] 0.11003799
[6,] 0.22158929
[7,] 0.09009148
[8,] 0.40376202
```

The contingency table

```
X <- expand.grid(X1 = c(0,1), X2 = c(0,1), X3 = c(0,1), stringsAsFactors = TRUE)
data3 <- data.frame(X, p)
data3
```

```
  X1 X2 X3          p
1  0  0  0 0.03314280
2  1  0  0 0.04048071
3  0  1  0 0.02713503
4  1  1  0 0.07376067
5  0  0  1 0.11003799
6  1  0  1 0.22158929
7  0  1  1 0.09009148
8  1  1  1 0.40376202
```

```
ftable(X1 + X2 ~ X3, xtabs(p ~. , data3))
```

```
    X1            0                       1
    X2            0            1           0           1
X3
0       0.03314280 0.02713503  0.04048071 0.07376067
1       0.11003799 0.09009148  0.22158929 0.40376202
```

The conditional odds-ratio are both 1

```
(0.05480844 * 0.14898478)/ (0.18197043 * 0.04487335)
```

```
[1] 1
```

```
(0.06694318 * 0.24563438)/(0.13480701 * 0.12197842)
```

```
[1] 1
```

An example

Some old data concerning breast cancer reported by Morrison n (1973). The three factors are

- X_1 diagnostic center
- X_2 nuclear grade
- X_3 survival after three years

Read the data

```
Freq <- c(35, 42, 59, 77, 47, 26, 112, 76)
df_bc <- data.frame(expand.grid(X1 = c("Boston", "Glamorgan"), X2 = c("malignant", "benign")
df_bc
```

```
         X1        X2      X3 Freq
1    Boston malignant    died   35
2 Glamorgan malignant    died   42
3    Boston    benign    died   59
4 Glamorgan    benign    died   77
5    Boston malignant survived   47
6 Glamorgan malignant survived   26
7    Boston    benign survived  112
8 Glamorgan    benign survived   76
```

Fit a saturated model

```
m_sat  <-glm(Freq ~ X1 * X2 * X3, family = poisson, data = df_bc)
m_sat
```

```
Call:  glm(formula = Freq ~ X1 * X2 * X3, family = poisson, data = df_bc)

Coefficients:
                 (Intercept)                    X1Glamorgan
                     3.55535                        0.18232
                    X2benign                     X3survived
                     0.52219                        0.29480
         X1Glamorgan:X2benign           X1Glamorgan:X3survived
                     0.08395                       -0.77437
           X2benign:X3survived  X1Glamorgan:X2benign:X3survived
```

```
                    0.34616                                    0.12034
```

```
Degrees of Freedom: 7 Total (i.e. Null);  0 Residual
Null Deviance:      89.97
Residual Deviance: -1.288e-14   AIC: 62.6
```

Fit a log-linear model

```r
m_ci <- glm(Freq ~ X1 * X2 + X1 * X3, family = poisson, data = df_bc)
m_ci
```

```
Call:  glm(formula = Freq ~ X1 * X2 + X1 * X3, family = poisson, data = df_bc)

Coefficients:
           (Intercept)             X1Glamorgan                  X2benign
               3.41662                 0.18384                   0.73494
          X3survived    X1Glamorgan:X2benign  X1Glamorgan:X3survived
               0.52561                 0.07599                  -0.67976

Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
Null Deviance:      89.97
Residual Deviance: 4.072     AIC: 62.67
```

The likelihood ratio test is $G_2^2 = 4.072$ that is not significant.

```r
anova(m_ci, m_sat, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Freq ~ X1 * X2 + X1 * X3
Model 2: Freq ~ X1 * X2 * X3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2     4.0724
2         0     0.0000  2   4.0724   0.1305
```

Fitted counts

```
m <- fitted(m_ci)
data_fit <- cbind(df_bc[, 1:3], m = m)
data_fit
```

```
          X1        X2       X3          m
1     Boston malignant     died   30.46640
2  Glamorgan malignant     died   36.61538
3     Boston    benign     died   63.53360
4  Glamorgan    benign     died   82.38462
5     Boston malignant survived   51.53360
6  Glamorgan malignant survived   31.38462
7     Boston    benign survived  107.46640
8  Glamorgan    benign survived   70.61538
```

```
m[1] * m[7] /(m[3]*m[5])
```

```
1
1
```

```
m[2] * m[8]/(m[4]*m[6])
```

```
2
1
```

Using a significance test for the same conditional independence

Use the package **bnlearn** to test conditional independence

```
library(readr)
data_bc<- read_rds("data_bc.rds")

ci.test("X2", "X3", "X1", data = data_bc)
```

```
    Mutual Information (disc.)

data:  X2 ~ X3 | X1
mi = 4.0724, df = 2, p-value = 0.1305
alternative hypothesis: true value is greater than 0
```