

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,] 0.05480844
[2,] 0.06694318
[3,] 0.04487335
[4,] 0.12197842
[5,] 0.18197043
[6,] 0.13480701
[7,] 0.14898478
[8,] 0.24563438
```

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	0.05480844
2	1	0	0	0.06694318
3	0	1	0	0.04487335
4	1	1	0	0.12197842
5	0	0	1	0.18197043
6	1	0	1	0.13480701
7	0	1	1	0.14898478
8	1	1	1	0.24563438

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

	X1	0	1	
X2	0	1	0	1
X3				
0	0.05480844	0.04487335	0.06694318	0.12197842
1	0.18197043	0.14898478	0.13480701	0.24563438

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 (1973). The three factors are

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

Read the data

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

ftable(X1 + X2 ~ X3, table(data_bc))
```

	X1	Boston	Glamorgan		
	X2	malignant	benign	malignant	benign
X3					
died		35	59	42	77
survived		47	112	26	76

Fit a saturated model

```
df_bc <- as.data.frame(table(data_bc))
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

```
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.

```
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

Using a significance test for the same conditional independence

Use the package `bnlearn`

```
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
```