Modelos log-lineares

Apresentamos exemplos de ajustes de modelos log-lineares com a função `glm` do pacote stats em R. Os resultados são destacados em cor azul.

Os dados do exemplo podem ser encontrados na Seção 8.2.4 do livro Agresti (2002), *Categorical Data Analysis* (2nd ed.). New York: Wiley. Referem-se a um levantamento amostral (*survey*) realizado em 1992 com 2276 estudantes do último ano do ensino médio na cidade de Dayton-OH. Os dados consistem de três variáveis indicadoras do uso de álcool (`A`), cigarro (`C`) e maconha (`M`). A tabela de contingências é armazenada na folha de dados (*data frame*) `dados`.

```
dados <- data.frame(expand.grid(
   M = factor(c("Sim", "Não"), C = factor(c("Sim", "Não"),
   A = factor(c("Sim", "Não"), contagem = c(911, 538, 44, 456, 3, 43, 2, 279))
```

O objeto `dados` consiste de quatro colunas. A função `expand.grid` cria uma folha de dados de três colunas com todas as combinações dos níveis (`Sim` e `Não`) das três variáveis (`A`, `C` e `M`). As frequências observadas (`contagem`) estão na quarta coluna.

```
dados
    M   C   A contagem
1 Sim Sim Sim      911
2 Não Sim Sim      538
3 Sim Não Sim       44
4 Não Não Sim      456
5 Sim Sim Não        3
6 Não Sim Não       43
7 Sim Não Não        2
8 Não Não Não      279
```

Neste exemplo adotamos as restrições de soma nula dos parâmetros λ.

```
options(contrasts = c("contr.sum", ""))
```

Observação 1. O comando `options(contrasts = c("contr.treatment", ""))` faz com que para cada fator os coeficientes dos primeiros níveis em ordem alfabética (neste exemplo, `Não`) sejam iguais a 0, ou seja, o primeiro nível de cada fator é o nível basal (*baseline*). O nível basal pode ser mudado com o argumento `levels` na criação da folha de dados. O comando abaixo muda o nível basal para `Sim`.

```
dados <- data.frame(expand.grid(
   M = factor(c("Sim", "Não"), levels = c("Sim", "Não")),
   C = factor(c("Sim", "Não"), levels = c("Sim", "Não")),
   A = factor(c("Sim", "Não"), levels = c("Sim", "Não"))),
   contagem = c(911, 538, 44, 456, 3, 43, 2, 279))
```

Ajustamos os modelos (A, C, M), (AC, M), (AM, CM), (AC, AM, CM) e (ACM). Para tanto, usamos a função `glm` aplicada ao modelo Poisson (`family = poisson`). As contagens fazem o papel da variável resposta.

```
## Modelo (A, C, M)
m1 <- glm(contagem ~ A + C + M, data = dados, family = poisson)
```

De outra forma,

```
m1 <- glm(contagem ~ ., data = dados, family = poisson)
```

notando que na fórmula "`contagem ~ .`" o lado direito representa todas variáveis em `dados`, exceto `contagem`.

```
## Modelo (M, AC)
m2 <- glm(contagem ~ A * C + M, data = dados, family = poisson)
```

Na fórmula acima, `A * C` significa que todos parâmetros envolvendo A e C (λ^A, λ^C e λ^{AC}) são incluídos no modelo, ao passo que `M` denota λ^M. Explicitando todos os componentes do modelo, escrevemos

```
m2 <- glm(contagem ~ A + C + M + A:C, data = dados, family = poisson)
```

sendo que `A:C` indica somente os parâmetros de associação λ^{AC} (interação de primeira ordem entre `A` e `C`).

```
## Modelo (AM, CM)
m3 <- glm(contagem ~ A * M + C * M, data = dados, family = poisson)

# Modelo (AC, AM, CM)
m4 <- glm(contagem ~ (A + C + M)^2, data = dados, family = poisson)
summary(m4)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.25153    0.11632  36.551  < 2e-16 *** μ
A1          -1.50400    0.11380 -13.217  < 2e-16 *** λ1^A
C1          -0.28227    0.05491  -5.140 2.74e-07 *** λ1^C
M1           1.19605    0.11850  10.093  < 2e-16 *** λ1^M
A1:C1        0.51363    0.04352  11.803  < 2e-16 *** λ11^AC
A1:M1        0.74650    0.11617   6.426 1.31e-10 *** λ11^AM
C1:M1        0.71197    0.04096  17.382  < 2e-16 *** λ11^CM
```

Na fórmula de `m4` acima a potência não significa o quadrado da soma, mas determina a inclusão de todas as parcelas com até duas variáveis. Os erros padrão foram estimados com base no modelo produto de distribuições Poisson independentes. Outras formas de ajustar o modelo `m4` são dadas abaixo.

```
m4 <- glm(contagem ~ .^2, data = dados, family = poisson)
m4 <- glm(contagem ~ A * C + A * M + C * M, data = dados, family = poisson)
m4 <- glm(contagem ~ A + C + M + A:C + A:M + C:M, data = dados, family = poisson)
m4 <- glm(contagem ~ A * C * M - A:C:M, data = dados, family = poisson)
```

Na última linha acima, do modelo saturado (`m5` logo abaixo) excluímos a interação de segunda ordem λ^{ACM} (`A:C:M`).

```
# Modelo (ACM)
m5 <- glm(contagem ~ .^3, data = dados, family = poisson)
```

O modelo saturado também pode ser ajustado com o comando `m5 <- glm(contagem ~ (A + C + M)^3, data = dados, family = poisson)`.

Observação 2. Com a função `model.matrix` verifique que as restrições de soma nula são válidas para as colunas da matriz modelo $\boldsymbol{X}$ exceto a primeira coluna.

Em seguida apresentamos as frequências esperadas estimadas obtidas de cada um dos cinco modelos ajustados. A função `fitted` fornece estas estimativas, listadas abaixo com pelo menos três dígitos significativos (`digits = 3`).

```
freqest <- cbind(dados, fitted(m1), fitted(m2), fitted(m3), fitted(m4), fitted(m5))
colnames(freqest) <- c("M", "C", "A", "Contagem", "(A,C,M)", "(M, AC)", "(AM, CM)",
                       "(AC, AM, CM)", "(ACM)")
print(freqest, digits = 3)
```

```
    M   C   A Contagem (A,C,M) (M, AC) (AM, CM) (AC, AM, CM) (ACM)
1 Yes Yes Yes      911   540.0   611.2   909.24       910.38   911
2  No Yes Yes      538   740.2   837.8   438.84       538.62   538
3 Yes  No Yes       44   282.1   210.9    45.76        44.62    44
4  No  No Yes      456   386.7   289.1   555.16       455.38   456
5 Yes Yes  No        3    90.6    19.4     4.76         3.62     3
6  No Yes  No       43   124.2    26.6   142.16        42.38    43
7 Yes  No  No        2    47.3   118.5     0.24         1.38     2
8  No  No  No      279    64.9   162.5   179.84       279.62   279
```

Conforme esperado, o modelo saturado reproduz exatamente as frequências observadas. Os demais modelos, exceto (AC, AM, CM), levam a estimativas das frequências esperadas distantes das frequências observadas.

Continuando o exemplo, estudamos as associações condicional e marginal em alguns modelos. Para M = *k,* a expressão do logaritmo da razão de chances é

$$\begin{aligned}\log \theta_{ij(k)} &= \log\left((\mu_{ijk}\,\mu_{i+1,j+1,k}) / (\mu_{i+1,jk}\,\mu_{i,j+1,k})\right)\\ &= \log\ \mu_{ijk} + \log\ \mu_{i+1,j+1,k} - \log\ \mu_{i+1,jk} - \log\ \mu_{i,j+1,k}. \qquad (1)\end{aligned}$$

No modelo (M, AC), $\log\ \mu_{ijk} = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC}$. Substituindo $\log\ \mu_{ijk}$ na expressão (1) e levando em conta as restrições de soma nula em uma tabela 2 x 2 x 2, obtemos $\log \theta_{11(1)} = 4\,\lambda_{11}^{AC}$. As estimativas dos parâmetros são

```
coefficients(m2)
```

```
(Intercept)          A1          C1          M1       A1:C1
5.0345262  -1.0065602   0.1864231   0.1577094   0.7184335
```

λ_{11}^{AC}

de maneira que a estimativa de λ_{11}^{AC} (razão de chances de uso de `A` e `C` dado que M = 1 = `Sim`) é

```
exp(4 * coefficients(m2)["A1:C1"])
```

```
 A1:C1
17.703
```

Na tabela marginal 2 x 2 de A e C a estimativa é a mesma (comprove) e a coincidência ocorre não apenas neste exemplo, pois A e C são conjuntamente independentes de M sob este modelo (a tabela tridimensional pode ser colapsada). Não estamos afirmando que o modelo `m2` = (M, AC) proporciona um bom ajuste. Para o modelo `m4` = (AC, AM, AC), temos log $\mu_{ijk} = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$. Substituindo log μ_{ijk} na expressão (1) obtemos log $\theta_{11(1)} = 4\,\lambda_{11}^{AC}$. As estimativas dos parâmetros são

```
coefficients(m4)
```

```
(Intercept)          A1          C1          M1
  4.2515335  -1.5039966  -0.2822714   1.1960453
λ11^AC    A1:C1       A1:M1       C1:M1
  0.5136335   0.7465036   0.7119723
```

e a estimativa da razão de chances $\theta_{11(1)}$ é dada por

```
exp(4 * coefficients(m4)["A1:C1"])
```

```
 7.803201
```

lembrando que na tabela marginal 2 x 2 de A e C a estimativa é 17,7, conforme visto acima, de modo que para estes dados, a associação marginal é mais forte do que associação condicional. Para o modelo (AC, AM, AC), a tabela tridimensional não pode ser colapsada e assim, para cada par de variáveis, padrões de associação parcial e marginal podem ser diferentes, como aconteceu neste exemplo com as variáveis A e C.

Agora ajustamos diversos modelos partindo do modelo saturado. Levando em conta que temos alguns modelos encaixados (*nested*), o ajuste de modelos mais simples é efetuado com uma atualização do modelo mais complexo usando a função `update`, notando que, por exemplo, `me1` = `m5`.

```
me1 <- glm(contagem ~ (A + C + M)^3, data = dados, family = poisson) # (ACM)
me2 <- update(me1, . ~ . - A:C:M) # (AC, AM, CM)
me3 <- update(me2, . ~ . - A:C)   # (AM, CM)
me4 <- update(me2, . ~ . - A:M)   # (AC, CM)
me5 <- update(me2, . ~ . - C:M)   # (AC, AM)
me6 <- update(me5, . ~ . - A:M)   # (M, AC)
me7 <- update(me5, . ~ . - A:C)   # (C, AM)
me8 <- update(me4, . ~ . - A:C)   # (A, CM)
me9 <- update(me8, . ~ . - C:M)   # (A, C, M)
```

Observação 3. Poderíamos iniciar com o modelo mais simples (A, C, M) e passar para modelos mais complexos adicionando elementos (“ . ~ . + “ nas fórmulas).

Organizamos os resultados dos testes de bondade de ajuste com as estatísticas G^2 e X^2 em uma tabela.

```
# Graus de liberdade
gl <- c(me9$df.residual, me8$df.residual, me7$df.residual, me6$df.residual,
        me5$df.residual, me4$df.residual, me3$df.residual, me2$df.residual)
```

O valor de G^2 encontra-se no componente `deviance` de cada objeto com o modelo ajustado (`me1`, `me2`, etc). A estatística X^2 é calculada a partir dos resíduos de Pearson (função `resid` com `type = "pearson"`).

```
# G2 e valor-p
G2 <- c(me9$deviance, me8$deviance, me7$deviance, me6$deviance,
        me5$deviance, me4$deviance, me3$deviance, me2$deviance)
pG2 <- pchisq(G2, gl, lower.tail = FALSE)

# X2 e valor-p
X2 <- c(sum(resid(me9, type = "pearson")^2), sum(resid(me8, type = "pearson")^2),
        sum(resid(me7, type = "pearson")^2), sum(resid(me6, type = "pearson")^2),
        sum(resid(me5, type = "pearson")^2), sum(resid(me4, type = "pearson")^2),
        sum(resid(me3, type = "pearson")^2), sum(resid(me2, type = "pearson")^2))

pX2 <- pchisq(X2, gl, lower.tail = FALSE)

# Bondade do ajuste
modelos <- c("(A, C, M)", "(A, CM)", "(C, AM)", "(M, AC)", "(AC, AM)",
             "(AC, CM)", "(AM, CM)", "(AC, AM, CM)")
bda <- data.frame(modelos, gl, G2, pG2, X2, pX2)
print(bda, digits = 4)
```

```
       modelos gl      G2        pG2        X2        pX2
1    (A, C, M)  4 843.827 2.464e-181 1411.3860 2.348e-304
2      (A, CM)  3  92.018  8.072e-20  505.5977 2.921e-109
3      (C, AM)  3 939.563 2.320e-203  824.1630 2.487e-178
4      (M, AC)  3 843.827 1.352e-182  704.9071 1.811e-152
5     (AC, AM)  2 497.369 9.946e-109  443.7611  4.350e-97
6     (AC, CM)  2  92.018  1.043e-20   80.8148  2.827e-18
7     (AM, CM)  2 187.754  1.697e-41  177.6149  2.700e-39
8 (AC, AM, CM)  1   0.374  5.408e-01    0.4011  5.265e-01
```

Adotando um nível de significância de 5%, apenas o modelo (AC, AM, CM) apresenta um ajuste satisfatório, sendo que $G^2 = 0{,}374$ ($p = 0{,}541$) e $X^2 = 0{,}401$ ($p = 0{,}526$) , ambas com 1 g.l. Este modelo inclui todos os coeficientes das associações entre pares de variáveis. Por sua vez, o modelo (AM, CM) de independência condicional entre o uso de álcool e cigarro dado o uso de maconha ($A \perp\!\!\!\perp C \mid M$) não ajusta bem os dados. Vimos que a hipótese de independência condicional pode ser testada com a estatística *CMH*.

Utilizando a estatística G^2, modelos encaixados podem ser comparados com a função `anova`, que realiza uma análise da desviância (ou do desvio, *deviance*). Exemplificamos comparando os modelos (ACM), (AC, AM, CM), (AM, CM)e (A, C, M).

```
(tanova = anova(me9, me7, me3, me2, me1))
```

```
Analysis of Deviance Table
Model 1: contagem ~ M + C + A
Model 2: contagem ~ M + C + A + M:A
Model 3: contagem ~ M + C + A + M:C + M:A
Model 4: contagem ~ M + C + A + M:C + M:A + C:A
Model 5: contagem ~ (M + C + A)^3

   Resid. Df Resid. Dev Df Deviance
1          4    1286.02
2          3     939.56  1   346.46
3          2     187.75  1   751.81
4          1       0.37  1   187.38
5          0       0.00  1     0.37
```

Na tabela acima as colunas `Resid. Df` e `Resid. Dev` significam os graus de liberdade e a estatística G^2 para cada modelo ajustado, respectivamente. A coluna `Df` representa os graus de liberdade da estatística de teste utilizada para comparar dois modelos encaixados, notando que `Df` é obtido por diferença entre duas linhas consecutivas da coluna `Resid. Df`. A coluna `Deviance` mostra a estatística de teste para a comparação entre os dois modelos. O nível descritivo (p) é calculado com o comando `pchisq(tanova$Deviance, tanova$Df, lower.tail = FALSE)`. Por exemplo, G^2(C, AM) – G^2(AM, CM) = 939,56 – 187,75 751,81, com g.l. = 1, de modo que há diferença significativa entre os modelos (AM, CM) e (C, AM) (p < 0,001). Não há diferença significativa entre o modelo saturado e o modelo (AC, AM, CM), pois G^2(AC, AM, CM) – G^2(ACM) = 0,37, com g.l. = 1 e p = 0,541.

O modelo (AC, AM, CM) é o modelo de associação homogênea para cada par de variáveis. Tomando o par (A,C), temos $\theta_{ij(1)} = \theta_{ij(2)}$, i, j = 1, 2. Tendo em vista que a variável C tem dois níveis, as igualdades se reduzem a $\theta_{11(1)} = \theta_{11(2)}$. A estimativa da razão de chances comum é

```
exp(4 * coefficients(me2)["A1:C1"])
```

```
 7.803201
```

A chance de uso de álcool para aqueles que já usaram cigarro é cerca de 7,8 vezes a chance de uso de álcool para aqueles que ainda não usaram cigarro, tanto para os participantes que já usaram maconha quanto para os que não usaram.

Ressalte-se que estimativas das razões de chances podem ser determinadas diretamente sem necessidade de reescrever a expressão de log $\theta_{ij(k)}$ dada pela equação (1). A partir de (1), em uma tabela 2 x 2 x 2,

$$\log \theta_{11(k)} = \log \mu_{11k} + \log \mu_{22k} - \log \mu_{12k} - \log \mu_{21k}, \quad k = 1, 2. \qquad (2)$$

Para k = 1, as parcelas do lado direito ocupam as posições 1, 7, 3 e 5 no vetor 8 x 1 do preditor linear $X\boldsymbol{\beta}$ (verifique as linhas de `dados`), enquanto que para k = 2 as posições são 2, 8, 4 e 6. Sendo assim, as estimativas de $\theta_{11(1)}$ e $\theta_{11(2)}$ são

```
xbetame2 <- me2$linear.predictors  X β
exp(xbetame2[1] + xbetame2[7] - xbetame2[3] - xbetame2[5])
```

```
7.803201  θ11(1)
```

```
exp(xbetame2[2] + xbetame2[8] - xbetame2[4] - xbetame2[6])
```

```
7.803201  θ11(2)
```

e as duas estimativas coincidem, conforme esperado, pois o modelo (AC, AM, CM) é o modelo de associação homogênea. De outra forma, utilizando as frequências esperadas estimadas:

```
fitted(me2)[2] * fitted(me2)[8] / (fitted(me2)[4] * fitted(me2)[6])
```

Um intervalo de confiança assintótico para $\theta_{11(1)}$ e $\theta_{11(2)}$ é dado por

```
conf <- 0.95      Coeficiente de confiança
sume2 <- summary(me2)
eplambda <- sume2$coefficients["A1:C1", "Std. Error"]  Erro padrão da estimativa de λ11^AC
emax <- qnorm((1 + conf) / 2) * eplambda    zconf x erro padrão da estimativa de λ11^AC
ICteta <- exp( 4 * (coefficients(me2)["A1:C1"] + c(-emax, emax)))
cat("\n IC de", conf * 100, "% para teta11(k):", ICteta, "\n")
```

```
IC de 95% para teta11(k): 5.547649 10.97581
```

No comando para o cálculo de `ICteta` primeiro calculamos o intervalo de confiança para λ_{11}^{AC}. Estes resultados indicam uma forte associação entre o uso de álcool e cigarro tanto para quem já usou maconha quanto para quem não usou.

Observação 4. Obtenha estimativas pontual e intervalar para as razões de chance condicionais envolvendo os pares (A, M) e também (C, M).

Finalmente, analisamos as associações condicionais entre A e C nos dois níveis da variável M utilizando o modelo saturado e a expressão (2).

```
xbetame1 <- me1$linear.predictors  Xβ
exp(xbetame1[1] + xbetame1[7] - xbetame1[3] - xbetame1[5])
```

```
13.80303  θ11(1)
```

```
exp(xbetame1[2] + xbetame1[8] - xbetame1[4] - xbetame1[6])
```

```
7.65514  θ11(2)
```

Os resultados ilustram que as estimativas das razões de chances podem depender bastante do modelo ajustado. Logo, inferências devem ser baseadas em um modelo que ajuste bem os dados.

Observação 5. Procure obter as estimativas dos erros padrão baseadas no modelo multinomial. Um estimador da matriz de covariâncias do estimador de máxima verossimilhança de $\boldsymbol{\beta}$ é dado por

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \left\{\mathbf{X}'\left[\mathbf{diag}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}'/n\right]\mathbf{X}\right\}^{-1}.$$

Na expressão acima, $\boldsymbol{X}$ denota a matriz modelo (obtida com a função `model.matrix`) sem a primeira coluna, que corresponde ao intercepto.

Observação 6. Modelos log-lineares também podem ser ajustados em R com as funções `loglin` (pacote stats) e `loglm` (pacote MASS).

Observação 7. Procure refazer estes ajustes em SPSS e com a PROC CATMOD em SAS.