Modelos log-lineares em tabelas tridimensionais

A Tabela 1 apresenta o número de respondentes em um estudo transversal classificados de acordo com a ideologia política, o partido político e o sexo. Os dados encontram-se em Agresti, A. (1996, *An Introduction to Categorical Data Analysis*, Wiley: New York, p. 203). Analise estes dados.

Tabela 1. Distribuição dos respondentes quanto à ideologia política, ao partido político e ao sexo (EUA, 1991).

		Ideologia política				
Sexo	Partido	Muito liberal	Levemente liberal	Moderada	Levemente conservadora	Muito conservadora
Feminino	Democrata	44	47	118	23	32
	Republicano	18	28	86	39	48
Masculino	Democrata	36	34	53	18	23
	Republicano	12	18	62	45	51

Uma questão de interesse é analisar a associação entre partido político e ideologia política tendo sexo como variável confundidora. Os dados consistem das variáveis partido político (X), ideologia política (Y) e sexo (Z). A tabela de contingências é armazenada na folha de dados (*data frame*) `dados`.

```
dados <- data.frame(expand.grid(
   ideologia = factor(c("ML", "LL", "Mod", "LC", "MC"),
                levels = c("ML", "LL", "Mod", "LC", "MC")),
   partido = factor(c("Democrata", "Republicano")),
   sexo = factor(c("Feminino", "Masculino"))),
   contagem = c(44, 47, 118, 23, 32,
                18, 28, 86, 39, 48,
                36, 34, 53, 18, 23,
                12, 18, 62, 45, 51))
```

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 das três variáveis na sequência `ideologia-partido-sexo`, de modo que a entrada de dados em `contagem` segue as linhas da Tabela 1.

Como ideologia é uma variável ordinal com níveis de "Muito liberal" (`ML`) a "Muito conservadora" (`MC`), a ordem é fixada com o argumento `levels` (por *defalut*, a ordem é alfabética) . As frequências observadas (`contagem`) estão na quarta coluna.

```
n <- sum(dados$contagem)
cat("\n Tamanho da amostra:", n)
```

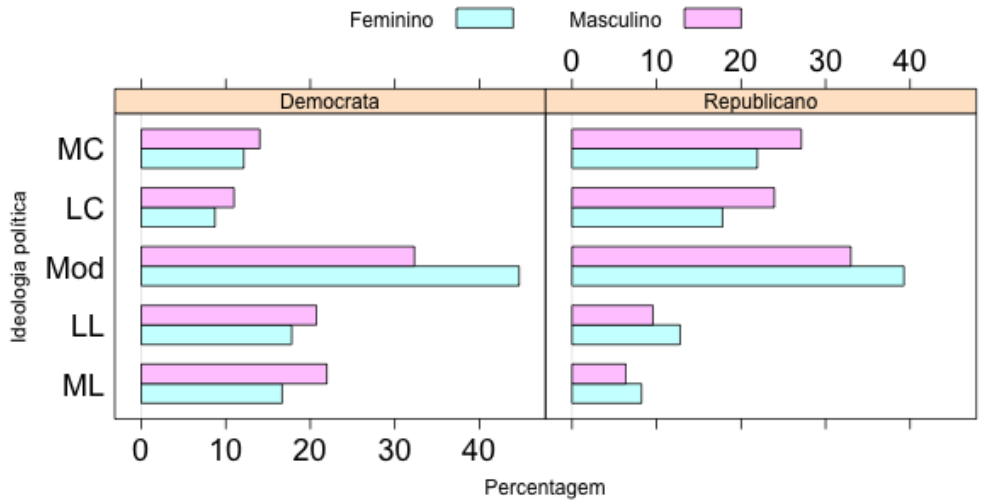
```
    Tamanho da amostra: 835
```

```
dados
```

	ideologia	partido	sexo	contagem
1	ML	Democrata	Feminino	44
2	LL	Democrata	Feminino	47
3	Mod	Democrata	Feminino	118
4	LC	Democrata	Feminino	23
5	MC	Democrata	Feminino	32
6	ML	Republicano	Feminino	18
7	LL	Republicano	Feminino	28
8	Mod	Republicano	Feminino	86
9	LC	Republicano	Feminino	39
10	MC	Republicano	Feminino	48
11	ML	Democrata	Masculino	36
12	LL	Democrata	Masculino	34
13	Mod	Democrata	Masculino	53
14	LC	Democrata	Masculino	18
15	MC	Democrata	Masculino	23
16	ML	Republicano	Masculino	12
17	LL	Republicano	Masculino	18
18	Mod	Republicano	Masculino	62
19	LC	Republicano	Masculino	45
20	MC	Republicano	Masculino	51

Os dados são representados em um gráfico de barras de frequências relativas. A função `barchart` é parte do pacote `lattice`.

```
library(lattice)
tabfreq <- xtabs(~ ideologia + partido + sexo, data = dados) *
   dados$contagem
tabfreqr <- prop.table(tabfreq, margin = 2:3) * 100 # Freq. relativas
barchart(tabfreqr, xlab = "Percentagem", ylab = "Ideologia política",
   stack = FALSE, scale = list(cex = 1.5),
   auto.key = list(space = "top", columns = 2))
```



Nota 1. Explique o gráfico da pag. 2 e a partir dele comente a associação entre as três variáveis.

Neste exemplo adotamos as restrições da casela de referência, *default* em R.

Ajustamos os modelos (X, Y, Z), (XY, Z), (XZ, YZ), (XY, XZ, YZ) e (XYZ). Para tanto, usamos a função `glm` com o modelo probabilístico produto de distribuições Poisson independentes (`family = poisson`).

Nota 2. Descreva cada um dos modelos acima em termos das variáveis da Tabela 1.

```
## Modelo (X, Y, Z)
m1 <- glm(contagem ~ partido + ideologia + sexo, data = dados, family =
      poisson)
```

De outra forma,

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

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

```
## Modelo (XY, Z)
m2 <- glm(contagem ~ partido * ideologia + sexo, data = dados, family =
      poisson)
```

Na fórmula acima, `partido * ideologia` significa que todos parâmetros envolvendo X e Y (λ^X, λ^Y e λ^{XY}) são incluídos no modelo, ao passo que `sexo` denota λ^Z. Explicitando todos os componentes do modelo (não é necessário), escrevemos

```
m2 <- glm(contagem ~ partido + ideologia + sexo + partido:ideologia, data =
      dados, family = poisson)
```

sendo que `partido:ideologia` indica somente os parâmetros de associação entre X e Y (λ^{XY}).

```
## Modelo (XZ, YZ)
m3 <- glm(contagem ~ partido * sexo + ideologia * sexo, data = dados,
      family = poisson)

# Modelo (XY, XZ, YZ)
m4 <- glm(contagem ~ (partido + ideologia + sexo)^2, data = dados, family =
      poisson)
```

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. Uma forma mais compacta para `m4` é dada abaixo.

```
m4 <- glm(contagem ~ .^2, data = dados, family = poisson)
```

Nota 3. Escreva o modelo `m4` de três outras formas distintas em R.

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

O modelo saturado também pode ser ajustado com o comando abaixo.

```
m5 <- glm(contagem ~ (partido + ideologia + sexo)^3, data = dados,
     family = poisson).
```

Nota 4. Com a função `model.matrix` verifique que as restrições da casela de referência são válidas para as colunas da matriz modelo $\boldsymbol{X}$, exceto a primeira coluna, nos modelos `m1` a `m5`.

Organizamos os resultados dos testes de ajuste com as estatísticas G^2 e X^2 em uma tabela. O modelo saturado não é incluído, pois com ele o ajuste é perfeito.

```
# Graus de liberdade
gl <- c(m1$df.residual, m2$df.residual, m3$df.residual, m4$df.residual)
```

O valor de G^2 encontra-se no componente `deviance` de cada objeto com o modelo ajustado (`m1`, `m2`, `m3` e `m4`). A estatística X^2 é calculada a partir dos resíduos de Pearson (função `resid` com `type = "pearson"`).

```
# G2 e valor-p
G2 <- c(m1$deviance, m2$deviance, m3$deviance, m4$deviance)
pG2 <- pchisq(G2, gl, lower.tail = FALSE)

# X2 e valor-p
X2 <- c(sum(resid(m1, type = "pearson")^2), sum(resid(m2, type = "pearson")^2),
    sum(resid(m3, type = "pearson")^2), sum(resid(m4, type = "pearson")^2))
pX2 <- pchisq(X2, gl, lower.tail = FALSE)
```

Os resultados das estatísticas de ajuste são apresentados abaixo com pelo menos quatro dígitos significativos.

```
# Bondade do ajuste
modelos <- c("(X, Y, Z)", "(XY, Z)", "(XZ, YZ)", "(XY, XZ, YZ)")
print(data.frame(modelos, gl, G2, pG2, X2, pX2), digits = 4)
```

```
       modelos gl         G2       pG2        X2       pX2
1    (X, Y, Z) 13  7.985e+01 1.177e-11 8.150e+01 5.761e-12
2      (XY, Z)  9  1.752e+01 4.119e-02 1.734e+01 4.370e-02
3     (XZ, YZ)  8  6.380e+01 8.329e-11 6.242e+01 1.561e-10
4 (XY, XZ, YZ)  4  3.245e+00 5.176e-01 3.235e+00 5.192e-01
```

Adotando um nível de significância de 5%, excluindo o modelo saturado, vemos que apenas o modelo `m4` (XY, XZ, YZ) apresenta um ajuste satisfatório, sendo que $G^2 = 3{,}245$ ($p = 0{,}5176$) e $X^2 = 3{,}235$ ($p = 0{,}5192$), com 4 g.l. Este modelo inclui todos os coeficientes das associações entre pares de variáveis.

Por sua vez, o modelo `m3` (XZ, YZ) de independência condicional entre partido (X) e ideologia política (Y) dado o sexo (Z) não ajusta bem os dados.

Nota 5.Vimos que a hipótese de independência condicional pode ser testada com a estatística *CMH*. Realize o teste com a estatística *CMH*.

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`).

```
# Frequências esperadas estimadas
freqest <- cbind(dados, fitted(m1), fitted(m2), fitted(m3), fitted(m4),
                 fitted(m5))
colnames(freqest) <- c("X", "Y", "Z", "Contagem", "(X,Y,Z)", "(XY, Z)",
   "(XZ, YZ)", "(XY, XZ, YZ)", "(XYZ)")
print(freqest, digits = 3)
```

```
     X           Y         Z Contagem (X,Y,Z) (XY, Z)  (XZ, YZ) (XY, XZ, YZ) (XYZ)
1   ML   Democrata  Feminino       44    32.6    46.3      33.9         46.6    44
2   LL   Democrata  Feminino       47    37.7    46.9      41.0         49.8    47
3  Mod   Democrata  Feminino      118    94.6    98.9     111.5        114.4   118
4   LC   Democrata  Feminino       23    37.1    23.7      33.9         22.2    23
5   MC   Democrata  Feminino       32    45.7    31.8      43.7         31.0    32
6   ML Republicano  Feminino       18    31.0    17.4      28.1         15.4    18
7   LL Republicano  Feminino       28    35.8    26.6      34.0         25.2    28
8  Mod Republicano  Feminino       86    89.9    85.6      92.5         89.6    86
9   LC Republicano  Feminino       39    35.2    48.6      28.1         39.8    39
10  MC Republicano  Feminino       48    43.4    57.3      36.3         49.0    48
11  ML   Democrata Masculino       36    23.8    33.7      22.4         33.4    36
12  LL   Democrata Masculino       34    27.4    34.1      24.2         31.2    34
13 Mod   Democrata Masculino       53    68.9    72.1      53.6         56.6    53
14  LC   Democrata Masculino       18    27.0    17.3      29.4         18.8    18
15  MC   Democrata Masculino       23    33.3    23.2      34.5         24.0    23
16  ML Republicano Masculino       12    22.6    12.6      25.6         14.6    12
17  LL Republicano Masculino       18    26.1    19.4      27.8         20.8    18
18 Mod Republicano Masculino       62    65.5    62.4      61.4         58.4    62
19  LC Republicano Masculino       45    25.7    35.4      33.6         44.2    45
20  MC Republicano Masculino       51    31.6    41.7      39.5         50.0    51
```

O modelo saturado reproduz exatamente as frequências observadas. Os demais modelos, exceto `m4` (XY, XZ, YZ), levam a estimativas das frequências esperadas distantes das frequências observadas.

Nota 6. Represente graficamente as frequências esperadas estimadas e as frequências observadas para os modelos acima, exceto `m5`.

Passamos a estudar as associações utilizando o modelo `m4`. O modelo `m4` (XY, XZ, YZ) é o modelo de associação homogênea para cada par de variáveis. A tabela tridimensional não pode ser colapsada e assim, para cada par de variáveis, padrões de associação parcial e marginal podem ser diferentes. Tomando o par X: partido (binária) e Y: ideologia. Usamos a notação

$$RC_{j(k)} = \frac{\dfrac{P(Y=1|X=1,Z=k)}{P(Y=j|X=1,Z=k)}}{\dfrac{P(Y=1|X=2,Z=k)}{P(Y=j|X=2,Z=k)}} = \frac{m_{11k} \times m_{2jk}}{m_{1jk} \times m_{21k}}.$$

Levando em conta que a tabela é 2 × 5 × 2, temos

$$RC_{j(1)} = RC_{j(2)}, \quad j = 2, 3, 4, 5.$$

De acordo com o modelo,

$$\log\left(RC_{j(1)}\right) = \lambda_{11}^{XY} + \lambda_{2j}^{XY} - \lambda_{1j}^{XY} - \lambda_{21}^{XY}.$$

As estimativas dos parâmetros são apresentadas abaixo, lembrando as restrições da casela de referência ($\lambda 1^X = \lambda 1^Y = \lambda 1^Z = 0$ e $\lambda i1^{XY} = \lambda 1j^{XY} = \lambda i1^{XZ} = \lambda 1k^{XZ} = \lambda j1^{YZ} = \lambda 1k^{YZ} = 0$, para i =1,2, j =1,…,5 e k=1,2).

```
summary(m4)
```

```
Coefficients:
                                  Estimate
(Intercept)                        3.84112
partidoRepublicano                -1.10530
ideologiaLL                        0.06687
ideologiaMod                       0.89857
ideologiaLC                       -0.73965
ideologiaMC                       -0.40730
sexoMasculino                     -0.33189
partidoRepublicano:ideologiaLL     0.42419
partidoRepublicano:ideologiaMod    0.86099
partidoRepublicano:ideologiaLC     1.68693
partidoRepublicano:ideologiaMC     1.56340
partidoRepublicano:sexoMasculino   0.27555
ideologiaLL:sexoMasculino         -0.13565
ideologiaMod:sexoMasculino        -0.37175
ideologiaLC:sexoMasculino          0.16266
ideologiaMC:sexoMasculino          0.07634
```

	λ
partidoRepublicano	$\lambda 2^X$
ideologiaLL	$\lambda 2^Y$
ideologiaMod	$\lambda 3^Y$
ideologiaLC	$\lambda 4^Y$
ideologiaMC	$\lambda 5^Y$
sexoMasculino	$\lambda 2^Z$
partidoRepublicano:ideologiaLL	$\lambda 22^{XY}$
partidoRepublicano:ideologiaMod	$\lambda 23^{XY}$
partidoRepublicano:ideologiaLC	$\lambda 24^{XY}$
partidoRepublicano:ideologiaMC	$\lambda 25^{XY}$
partidoRepublicano:sexoMasculino	$\lambda 22^{XZ}$
ideologiaLL:sexoMasculino	$\lambda 22^{YZ}$
ideologiaMod:sexoMasculino	$\lambda 32^{YZ}$
ideologiaLC:sexoMasculino	$\lambda 42^{YZ}$
ideologiaMC:sexoMasculino	$\lambda 52^{YZ}$

O nível de referência para a variável ideologia é muito liberal (Y = 1: `ML`). Para respondentes com ideologia levemente liberal (Y = 2: `LL`), a razão de chances comum (masculino e feminino) é

$$RC_{2(1)} = RC_{2(2)} = \exp\left(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}\right) = \exp\left(\lambda_{22}^{XY}\right),$$

com estimativa

```
exp(coefficients(m4)["partidoRepublicano:ideologiaLL"])
```

```
1.528353
```

Analogamente, para respondentes com ideologia levemente conservadora (Y = 4: `LC`), a razão de chances comum (respondentes masculinos e femininos) é $RC_{4(1)} = RC_{4(2)} = \exp(\lambda 24^{XY})$, com estimativa

```
exp(coefficients(m4)["partidoRepublicano:ideologiaLC"])
```

```
5.402858
```

A chance de um respondente ter ideologia muito liberal (`ML` é o nível de referência para Y) em relação a ter ideologia levemente conservadora (`LC`) é cerca de 5,4 vezes maior para os respondentes identificados com o partido Democrata (`Democrata` é o nível basal para X) do que para os respondentes identificados com o partido Republicano, não importando o sexo.

Este resultado pode ser apresentado como a seguinte estimativa:
Tanto para mulheres quanto para homens, temos que

$$\frac{\text{chance(muito liberal } \textit{versus} \text{ levemente conservador} \mid \text{é democrata)}}{\text{chance(muito liberal } \textit{versus} \text{ levemente conservador} \mid \text{é republicano)}} = 5,40.$$

O intervalo de confiança assintótico de 95% indica que esta associação é forte.

```
exp(confint(m4, parm = "partidoRepublicano:ideologiaLC", level = 0.95))
```

```
   2.5 %    97.5 %
3.108175 9.599845
```

Nota 6. Avalie a associação marginal entre partido político (X) e ideologia política (Y) e compare com os resultados do modelo de associação homogênea (`m4`).

Nota 7. O modelo (XY, Z), identificado como `m3`, apresenta $G^2 = 17,52$ ($p = 0,0412$) e $X^2 = 17,34$ ($p = 0,0437$), com 9 g.l. (vide pag. 4). Logo, a rejeição de `m3` não é tão forte. Descreva este modelo em palavras e discuta as associações entre as variáveis.

Nota 8. Comparando os modelos que são encaixados, pag. 4, quais são suas conclusões?

Nota 9. Para o modelo probabilístico produto de distribuições Poisson independentes, a matriz de covariâncias assintótica do estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$ (formado por todos os λ's) é dada pela inversa da matriz de informação de Fisher, cuja expressão é

$$\hat{\text{cov}}(\widehat{\boldsymbol{\beta}}) = \{\boldsymbol{X}^\top \mathbf{diag}(\widehat{\boldsymbol{m}})\boldsymbol{X}\}^{-1}. \qquad (1)$$

Na expressão (1), $\boldsymbol{X}$ denota a matriz modelo (função `model.matrix`). O vetor com as frequências esperadas estimadas $\widehat{m}_{111}, \widehat{m}_{112}, \ldots, \widehat{m}_{IJK}$ é denotado por $\widehat{\boldsymbol{m}}$ e $\mathbf{diag}(\widehat{\boldsymbol{m}})$ é a matriz diagonal formada pelos elementos de $\widehat{\boldsymbol{m}}$ na diagonal principal.

Os erros padrão das estimativas para o modelo `m4` podem ser obtidos observando a terceira coluna do resultado de `summary(m4)` ou diretamente, utilizando

```
summary(m4)$coefficients[, "Std. Error"].
```

Nota 10. Para o modelo probabilístico multinomial, partindo do modelo Poisson independente e calculando probabilidades condicionais em n (fixando n), o termo constante (intercepto) λ cancela e $\boldsymbol{\beta}_1$ representa o vetor $\boldsymbol{\beta}$ sem o intercepto. A matriz de covariâncias do estimador de máxima verossimilhança do vetor $\boldsymbol{\beta}_1$ é dada por

$$\hat{\text{cov}}(\widehat{\boldsymbol{\beta}}_1) = [\boldsymbol{X}_1^\top \{\mathbf{diag}(\hat{\boldsymbol{m}}) - \hat{\boldsymbol{m}}\hat{\boldsymbol{m}}^\top/n\}\boldsymbol{X}_1]^{-1}. \qquad (2)$$

Na expressão (2), $\boldsymbol{X}_1$ denota a matriz modelo $\boldsymbol{X}$ sem a primeira coluna, que corresponde ao intercepto. Pode ser provado que a matriz em (2) coincide com a matriz em (1) após eliminação da primeira linha e da primeira coluna (referentes ao intercepto λ). Portanto, resultados obtidos com base no modelo Poisson independente são válidos também para o modelo multinomial. As estimativas de máxima verossimilhança dos coeficientes $\boldsymbol{\beta}_1$ são as mesmas do modelo probabilístico Poisson independente.

Nota 11. Para o modelo probabilístico produto de multinomiais independentes, os resultados obtidos com base no modelo Poisson independente são válidos, desde que seja incluído um termo para a distribuição marginal fixada pelo desenho do estudo. Por exemplo, se para cada nível das variáveis X e Z tivermos distribuições multinomiais para Y (como variável resposta), o modelo deve incluir os coeficientes λ^{XZ}.

Nota 12. Modelos log-lineares também podem ser ajustados em R com as funções `loglin` (pacote `stats`) e `loglm` (pacote `MASS`).

Nota 13. Procure refazer estes ajustes em SPSS e com a `PROC CATMOD` em SAS.

Nota 14. A variável ideologia política na Tabela 1 é ordinal. Sendo assim, outros modelos (não estudados na disciplina) poderiam ser propostos.