

## Modelos log-lineares em uma tabela $2 \times 2$

Os dados foram coletados em um estudo transversal. As variáveis do exemplo são gênero (gender:  $X$ ) e crença na vida após a morte (belief:  $Y$ ).  $X$  tem níveis feminino (F) e masculino (M), enquanto  $Y$  tem níveis sim (Sim) e não ou indeciso (Não). Os dados estão na Tabela 2.1, p. 17 em Agresti (1996, *An Introduction to Categorical Data Analysis*, Wiley: New York).

A entrada de dados abaixo segue a mesma ordem de linhas e colunas apresentada no livro. O argumento `levels` altera a ordem dos níveis (*default* em R: ordem alfabética).

```
count <- c(435, 147, 375, 134)
gender <- factor(c("F", "F", "M", "M"))
belief <- factor(c("Sim", "Não", "Sim", "Não"), levels = c("Sim", "Não"))

cat("\n n =", sum(count))

n = 1091
```

Em seguida são ajustados três modelos em ordem crescente de complexidade. O primeiro (`m0`) é o modelo nulo, que ignora as duas variáveis. O segundo (`mind`) é o modelo de independência entre  $X$  e  $Y$ , enquanto `msat` corresponde ao modelo saturado.

```
m0 <- glm(count ~ 1, family = poisson)
mind <- glm(count ~ gender + belief, family = poisson)
msat <- glm(count ~ gender * belief, family = poisson)
```

Na fórmula `count ~ gender + belief` o sinal `+` não significa adição. Indica que as duas variáveis são incluídas no modelo, além do intercepto.

Na fórmula `count ~ gender * belief` o sinal `*` não significa produto. Indica que as duas variáveis são incluídas no modelo e também a interação `gender:belief` ( $X:Y$ ), além do intercepto. A fórmula `count ~ gender * belief` pode ser escrita de forma expandida como `count ~ gender + belief + gender:belief`.

Outra forma de ajustar o modelo saturado é

```
msat <- glm(count ~ (gender + belief)^2, family = poisson)
```

Na fórmula `count ~ (gender + belief)^2` a operação `^2` não significa o quadrado da soma. Indica que são incluídos todos os termos com até duas variáveis (ou seja, `gender`, `belief` e `gender:belief`), além do intercepto.

A função `update` pode ser usada para ajustar modelos encaixados.

```
m0 <- glm(count ~ 1, family = poisson)
mind <- update(m0, . ~ . + gender + belief)
msat <- update(mind, . ~ . + gender:belief)
```

De outra forma, iniciando com o modelo saturado e obtendo modelos mais simples,

```
msat <- glm(count ~ gender * belief, family = poisson)
mind <- update(msat, . ~ . - gender:belief)
m0 <- update(mind, . ~ . - gender - belief)
```

O ajuste de um modelo com a função `glm` produz diversos resultados, cujos nomes podem ser vistos com a função `names`.

```
names(mind)
```

```
"coefficients" "qr" "iter" "converged" "data"
"residuals" "family" "weights" "boundary" "offset"
"fitted.values" "linear.predictors" "prior.weights" "model" "control"
"effects" "deviance" "df.residual" "call" "method"
"R" "aic" "df.null" "formula" "contrasts"
"rank" "null.deviance" "y" "terms" "xlevels"
```

Nota 2. Tente explicar os resultados de `mind[1]`, `mind[[1]]`, `mind["coefficients"]`, `mind[["coefficients"]]` e `mind$coefficients`.

O ajuste do modelo de independência (`mind`) é verificado com as estatísticas  $G^2 = -2 \log(\text{razão de verossimilhanças})$  e  $X^2$  de Pearson.

```
ecount <- mind$fitted.values
cat("\n Contagens observadas e esperadas estimadas: \n",
    count, "\n", round(ecount, 1))
```

```
Contagens observadas e esperadas estimadas:
435    147    375    134
432.1 149.9  377.9  131.1
```

```
X2 <- sum((count - ecount)^2 / ecount)
G2 <- 2 * sum(count * (log(count) - log(ecount)))

cat("\n X2:", X2, ",", mind$df.residual, "g.l. (p =",
    pchisq(X2, mind$df.residual, lower.tail = FALSE), ")")
cat("\n G2:", G2, ",", mind$df.residual, "g.l. (p =",
    pchisq(G2, mind$df.residual, lower.tail = FALSE), ")")

X2: 0.162084 , 1 g.l. (p = 0.6872451 )
G2: 0.1619951 , 1 g.l. (p = 0.6873263 )
```

Portanto, a um nível de significância de 5%, a hipótese de independência entre gênero e crença na vida após a morte não pode ser rejeitada ( $p > 0,05$ ).

Observações. 1. A estatística  $X^2$  pode ser calculada como a soma dos quadrados dos resíduos de Pearson: `X2 <- sum(resid(mind, type = "pearson")^2)`.

2. O valor da estatística  $G^2$  também é dado por `mind$deviance` (desviância do modelo).

3. As frequências esperadas estimadas também podem obtidas com a função `fitted`, ou seja, `ecount <- fitted(mind)`.

As estimativas dos parâmetros do modelo de independência (`mind`) são apresentadas abaixo.

```
summary(mind)
  Coefficients:
                Estimate Std. Error z value Pr(>|z|)
λ (Intercept)  6.06865    0.04512 134.488 <2e-16
λ2x genderM   -0.13402    0.06069  -2.208  0.0272
λ2y beliefNão -1.05868    0.06923 -15.291 <2e-16
```

Por *default*, a parametrização em R (casela de referência) utiliza a primeira categoria (ou nível) como referência (nível basal). Neste exemplo temos estimativas  $\lambda 1^X = 0$  ( $X = 1$ : gênero = F) e  $\lambda 1^Y = 0$  ( $Y = 1$ : crença na vida após a morte = sim). Em SAS, os módulos (PROC) GENMOD e CATMOD utilizam outras parametrizações.

Para analisar os dados, consideramos  $Y$  como variável resposta. No exemplo obtemos que a estimativa de  $\lambda 1^Y - \lambda 2^Y$  é igual a  $0 - (-1,05868) = 1,05868$ . Esta diferença é constante, independente da parametrização. A estimativa da chance de crença na vida após a morte é  $\exp(1,05868) = 2,88$ , para cada gênero (F e M).

Os três modelos encaixados (`m0`, `mind` e `msat`) podem ser comparados com a estatística de teste  $G^2$  (LRT) implementada na função `anova`.

```
anova(m0, mind, msat, test = "LRT")
Analysis of Deviance Table

Model 1: count ~ 1
Model 2: count ~ gender + belief
Model 3: count ~ gender * belief

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3    272.685
2         1     0.162  2  272.523 <2e-16 ***
3         0     0.000  1   0.162  0.6873
```

A coluna `Deviance` mostra as diferenças entre as desviâncias de modelos em linhas seguidas e `Df` representa as respectivas diferenças nos graus de liberdade. O modelo de independência representa um ganho de ajuste comparado ao modelo nulo (`m0`); por quê?

Nota 2. Justifique as frequências esperadas estimadas obtidas do modelo nulo, dadas por `fitted(m0)`.

As estimativas dos parâmetros do modelo saturado (`msat`) são apresentadas abaixo.

```
summary(msat)
```

Coefficients:

		Estimate	Std. Error	z value	Pr(> z )
$\lambda$	(Intercept)	6.07535	0.04795	126.711	<2e-16
$\lambda 2^X$	genderM	-0.14842	0.07047	-2.106	0.0352
$\lambda 2^Y$	beliefNão	-1.08491	0.09540	-11.372	<2e-16
$\lambda 2^{XY}$	genderM:beliefNão	0.05583	0.13868	0.403	0.6873

O parâmetro de associação entre  $X$  e  $Y$  é  $\lambda 2^{XY}$ . A razão de chances tem estimativa  $\exp(\lambda 2^{XY}) = \exp(0,05583) = 1,057418$ , indicando uma associação fraca entre  $X$  e  $Y$ , que não é significativa ( $p = 0,6873$ ). De outra forma, a estimativa é dada por  $435 \times 134 / (375 \times 147)$ . Um intervalo de confiança assintótico de 95% para a razão de chances é apresentado abaixo. Foi utilizada a função `exp` porque os coeficientes ( $\lambda$ ,  $\lambda 2^X$ , etc) foram obtidos de um modelo em escala logarítmica.

```
exp(confint(msat, parm = "genderM:beliefNão", level = 0.95))
```

```
      2.5 %      97.5 %  
0.8054086 1.3875675
```

Nota 3. Interprete o resultado acima.

Nota 4. O gráfico abaixo mostra a distribuição condicional da variável crença na vida após a morte em relação ao gênero. O gráfico sugere que as duas variáveis são independentes?

```
pcond <- prop.table(matrix(count, ncol = 2, byrow = TRUE), margin = 1)  
rownames(pcond) <- c("Feminino", "Masculino")  
colnames(pcond) <- c("Crença: sim", "Crença: não")
```

```
library(lattice)  
barchart(pcond, ylab = "Proporção", xlab = "Gênero", stack = FALSE,  
         horizontal = FALSE, scale = list(cex = 1.5),  
         auto.key = list(space = "top", columns = 2))
```

