

#### **Modelos Log-lineares**

Maria Teresa Leão Costa

UnB – IE Departamento de Estatística

Análise de Dados Categorizados

#### Modelos Log-lineares

- Com a abordagem log-linear, modela-se as frequências de uma tabela de contingência em termos da associação entre variáveis.
- Os modelos log-lineares especificam como a frequência de uma célula depende dos níveis das variáveis categóricas para aquela célula.

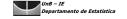




#### **Modelos Log-lineares**

- Metodologia apropriada quando n\u00e3o existe distin\u00e7\u00e3o clara entre vari\u00e1vel resposta e vari\u00e3vels explicativas.
- Todas as variáveis são tratadas como resposta e o foco está na independência estatística.

Os MODELOS LOG-LINEARES descrevem padrões de associação e interação entre variáveis categorizadas.



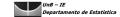


## **Modelos Log-lineares para Tabelas Bidimensionais**

#### Tabelas 2x2

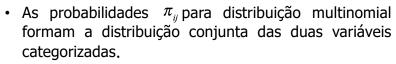
 Suponha uma amostra de tamanho n classsificada segundo duas variáveis binárias X e Y.

	Y		
Grupo	Sucesso	Insucesso	Total
1	$n_{II}$	$n_{12}$	$n_{I^+}$
2	$n_{2I}$	$n_{22}$	$n_{2+}$
Total	$n_{+1}$	$n_{+2}$	n



# 4

#### Modelos Log-lineares para Tabelas Bidimensionais



	Y		
X	1	2	Total
1	$\pi_{II}$	$\pi_{l2}$	$\pi_{I^+}$
2	$\pi_{2I}$	$\pi_{22}$	$\pi n_{2+}$
Total	$\pi_{+1}$	$\pi_{+2}$	1

- Os modelos log-lineares usam as frequências esperadas  $\mu_{ii} = n \cdot \pi_{ii}$  em vez das probabilidades  $\pi_{ii}$ .
  - $m{+}$   $n_{ij}$  frequências das células tem distribuição de Poisson com parâmetro  $\mu_{ij}$



Tomando o logaritmo, em ambos os lados, temos que:

$$\log(\theta) = \log(\mu_{11}) - \log(\mu_{12}) - \log(\mu_{21}) + \log(\mu_{22}) = 0$$

ou seja, a independência estatística pode ser expressa como uma combinação linear dos logaritmos das frequências esperadas.



Ou em termos da "odds ratio":

$$\theta = \frac{\pi_{11} \cdot \pi_{22}}{\pi_{12} \cdot \pi_{21}} = 1$$

A odds ratio pode ser expressa também em termos das frequências esperadas, onde  $\mu_{ij} = n \cdot \pi_{ij}$ 

$$\theta = \frac{\mu_{11} \cdot \mu_{22}}{\mu_{12} \cdot \mu_{21}} = 1$$





#### Modelo de Independência

■ Se Xe Ysão independentes então:

$$\pi_{ij} = \pi_{i+}.\pi_{+j}$$

 A expressão relacionada as frequências esperadas é dada por:

$$\mu_{ij} = n \, \pi_{i+} \pi_{+j}$$

#### Modelo de Independência

 Na escala logarítmica, a independência tem a forma aditiva

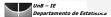
$$\log \mu_{ii} = \log n + \log \pi_{i+} + \log \pi_{+i}$$

 Denotando a variável linha por X e a variável coluna por Y, a expressão acima é equivalente a:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

Efeito linha- efeito da classificação na linha i

Efeito coluna- efeito da classificação na coluna j





■ H<sub>0</sub>) X e Y independentes

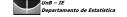
$$H_0$$
) $X$   $e$   $Y$  independentes  $\Leftrightarrow H_0$ ) $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$ 

 Os valores ajustados que satisfazem ao modelo log-linear de independência são:

$$\hat{\mu}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n}$$



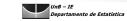
- O modelo especifica como as contagens esperadas das células variam de acordo com as categorias de Xe Y.
- O modelo considera as observações como contagens de células em vez de classificações de indivíduos.





Modelos loglineares para tabelas de contingência são *modelos lineares generalizados* que:

- tratam as contagens de células como observações independentes das distribuições de Poisson
- usam a *função de link -* log.



### Interpretação dos Parâmetros

- No modelo de independência a diferença entre dois parâmetros para uma dada variável indica o log da odds de uma resposta com relação a outra, nesta variável.
  - Exemplo:

Seja uma tabela Ix2 (X com I níveis e Y com dois níveis) Considere  $\pi = P(Y=1) = P(sucesso)$ .

Para a linha *i* temos que:

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) = \log(\mu_{i1}) - \log(\mu_{i2}) =$$

$$= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y$$

- $\Rightarrow$  Não depende de  $\emph{i}$ , istoé, logito para Y não depende do nível de X
- ⇒independe da parametrização usada



## Este modelo é similar a uma Análise de Variância com 2 fatores para uma variável resposta Y:

$$E(y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij}$$

onde:

μ – efeito médio geral

 $\begin{vmatrix} \alpha_i \\ \beta_i \end{vmatrix}$  efeitos principais dos fatores 1e2

 $|(\alpha\beta)_{ij}$  – efeito da interação

$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

onde:

 $\mu$  – termo baseado no tamanho da amostra

 $\lambda_i^X$  – efeito da classificação na linha i para variável X

 $\left|\lambda_{j}^{Y}
ight.$  – efeito da classificação na linha j para variável Y

 $\lambda_{ii}^{XY}$  – efeito da interação entre X e Y



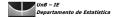
#### Modelo Log-linear Saturado

O Modelo Log-linear saturado para tabela 2x2 é:

$$\log(\mu_{ii}) = \mu + \lambda_i^X + \lambda_i^Y + \lambda_{ii}^{XY}$$
  $i = 1, 2$   $e$   $j = 1, 2$ 

onde  $\mu_{ij} = n\pi_{ij} \acute{e}$  a frequência esperada da célula (i,j).

- Os parâmetros λ<sup>XY</sup><sub>ij</sub> são termos de associação representam interações entre X e Y , em que o efeito de qualquer variável na frequência esperada das células depende da categoria da outra variável.
- O modelo descreve perfeitamente qualquer conjunto de frequências esperadas.
- É o modelo mais geral para tabelas de contingência bidimensional.





Observe que no modelo linear saturado para tabelas 2x2 :

$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$
  $i = 1,2$   $e$   $j = 1,2$ 

no. de parâmetros:

$$(\lambda_{1}^{X}, \lambda_{2}^{X}) (\lambda_{1}^{Y}, \lambda_{2}^{Y}) (\lambda_{11}^{XY}, \lambda_{12}^{XY}, \lambda_{21}^{XY}, \lambda_{22}^{XY})$$

⇒ Modelo Superparametrizado

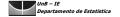
#### ■ Restrições:

$$\sum_{i=1}^{2} \lambda_{i}^{X} = 0 \;\; , \;\; \sum_{j=1}^{2} \lambda_{j}^{Y} = 0 \; , \;\;\; \sum_{i=1}^{2} \lambda_{ij}^{XY} = \sum_{j=1}^{2} \lambda_{ij}^{XY} = 0$$

produz 3 parâmetros não redundantes:

$$\lambda_1^X$$
,  $\lambda_1^Y$ ,  $\lambda_{11}^{XY}$ 

O quarto parâmetro,  $\mu$  , é fixado pelo tamanho da amostra n.



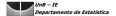
#### Interpretação dos Parâmetros

- Existe uma relação direta entre a razão de chances e os parâmetros de associação dos modelos log-lineares. A relação é mais simples para uma tabela 2x2.
- Para o modelo saturado

$$\log \theta = \log \left( \frac{\mu_{11} \mu_{22}}{\mu_{12} \mu_{21}} \right) = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21}$$

Assim:

$$\log(\theta) = (\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) - (\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\mu + \lambda_2^X + \lambda_1^Y + \lambda_{22}^{XY}) = = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$



## Relação direta entre log (odds ratio) e $\lambda_{ij}^{XY}$

Temos que "odds ratio":

$$\theta = \frac{\pi_{11} \cdot \pi_{22}}{\pi_{12} \cdot \pi_{21}} = 1$$

A odds ratio pode ser expressa também em termos das frequências esperadas, onde  $\mid \mu_{ii} = n \pi_{i+} \pi_{+i}$ 

$$\theta = \frac{\mu_{11} \cdot \mu_{22}}{\mu_{12} \cdot \mu_{21}} = 1$$

Tomando o logaritmo, em ambos os lados, temos que:

$$\log(\theta) = \log(\mu_{11}) - \log(\mu_{12}) - \log(\mu_{21}) + \log(\mu_{22}) = 0$$

ou seja, a independência estatística pode ser expressa como uma combinação linear dos logaritmos das fiequências esperadas.

#### Interpretação dos Parâmetros

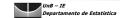
Das restrições :

Assim:

$$log(\theta) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} =$$

$$= \lambda_{11}^{XY} + (-\lambda_{21}^{XY}) + (+\lambda_{11}^{XY}) - \lambda_{21}^{XY} =$$

$$= \lambda_{11}^{XY} + (\lambda_{11}^{XY}) + (+\lambda_{11}^{XY}) + \lambda_{11}^{XY} = 4 \lambda_{11}^{XY}$$

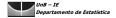


#### Interpretação dos Parâmetros

#### ■ Temos então que:

$$\log \theta = 4\lambda_{11}^{XY}$$

A razão de chances para uma tabela 2x2 é igual ao antilog de quatro vezes o parâmetros de associação no modelo loglinear saturado.



#### Modelo Log-linear de Independência

Se X e Y são independentes  $\theta$ =1 e  $log(\theta)$ =log(1)=0, mas :

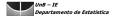
$$\log(\theta) = 4\lambda_{ij}^{XY} \quad i = 1, 2 \quad e \quad j = 1, 2$$

Assim testar independência equivale a testar:

$$H_0 \lambda_{ij}^{XY} = 0$$
  $i = 1,2$   $e$   $j = 1,2$ 

Neste caso temos o Modelo Log-linear de Independência

$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y \quad i = 1,2 \quad e \quad j = 1,2$$





#### Modelo Log-linear de Independência

Se X e Y são independentes  $\theta$ =1 e  $log(\theta)$ =log(1)=0, mas :

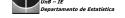
$$\log(\theta) = 4\lambda_{ij}^{XY} \quad i = 1, 2 \quad e \quad j = 1, 2$$

Assim testar independência equivale a testar:

$$H_0 \lambda_{ij}^{XY} = 0$$
  $i = 1,2$   $e$   $j = 1,2$ 

Neste caso temos o Modelo Log-linear de Independência

$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y \quad i = 1,2 \quad e \quad j = 1,2$$



#### Tabelas I x J

Modelo Log-linear saturado:

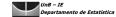
$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$
  $i = 1,..., I \ e \ j = 1,2,...,J$  com as restrições:

$$\sum_{i=1}^{I} \lambda_i^{X} = 0$$
 ,  $\sum_{j=1}^{J} \lambda_j^{Y} = 0$  ,  $\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0$ 

o que implica em IJ-1 parâmetros não redundantes.

Modelo Log-linear de Independência:

$$\log(\mu_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$$
  $i = 1, ..., I$   $e$   $j = 1, 2, ..., J$ 





No modelo saturado, como já foi visto, existe uma relação direta entre a log da odds ratio e os parâmetros de associação, \( \lambda\_{ij}^{XY} \), a saber:

$$\log(\theta) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

