Análise de Dados Categóricos
Modelos log-lineares

Enrico A. Colosimo/UFMG

Depto. Estatística - ICEx - UFMG

Modelos log-lineares

- Modelo simétrico: todas as variáveis são respostas.
- O objetivo é modelar a distribuição conjunta das variáveis em uma tabela de contingência.
- Modelo log-linear: descreve padrões de associação entre variáveis categóricas. Particularmente útil para modelar a relação entre duas ou mais respostas
- Modelagem em termos de médias/frequências esperadas.

Tabela $r \times c$.

Frequências esperadas sob a hipótese de independência:

$$\widehat{\mu}_{ij} = \widehat{E}_{ij} = \frac{n_{i+} n_{+j}}{n}$$

- $\mu_{ij} = \mu \alpha_i \beta_j$.
- α_i é um efeito da variável linha;
- β_j é um efeito da variável coluna.
- $\log \mu_{ij}$ é um modelo linear;
- Modelo de Poisson com ligação logarítmica.
- Modelo multinomial $\mu_{ij} = n \pi_{i+} \pi_{+j}$, não tem intercepto pois n é fixo.

Tabela $r \times c$.

1. Modelo de Independência

$$log E_{ij} = \mu + \lambda_i^x + \lambda_j^y;\ i = 1, \ldots, r;\ j = 1, \ldots, c.$$

- λ_i^x é o efeito de linha e
- λ_j^y é o efeito de coluna.
- de forma a estabelecer a condição de identificabilidade do modelo, os parâmetros devem satisfazer a alguma restrição. Por exemplo,

$$\sum_{i=1}^{r} \lambda_i^x = \sum_{j=1}^{c} \lambda_j^y = 0 \qquad \text{ou} \qquad \lambda_r^x = \lambda_c^y = 0$$

Número de parâmetros: $r + c - 1$

Modelo sob Dependência

2 Modelo de Dependência (Modelo Saturado)

$$logE_{ij} = \mu + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy}$$

Número de Parâmetros:

- $\mu(1), \lambda_i^x(r-1), \lambda_j^y(c-1), \lambda_{ij}^{xy}((r-1)(c-1))$
- total = rc

pois: $\sum_{i=1}^{r} \lambda_{ij}^{xy} = \sum_{j=1}^{c} \lambda_{ij}^{xy} = 0$

- Hipótese de Independência

$H_0 : \lambda_{ij}^{xy} = 0$

$i = 1, ..., r \quad j = 1, ..., c$

Observações

- Os modelos são motivados por aqueles de ANOVA para respostas contínuas.
- Restrições são impostas para obter identificabilidade dos modelos.
- Da mesma forma que em ANOVA, as estimativas dos parâmetros dependem das restrições mas não as estimativas das frequências esperadas.
- E também, não temos interesse nas estimativas dos parâmetros mas somente nas das frequências esperadas.
- No modelo multinomial: $\pi_{ij} = \frac{E_{ij}}{\sum_i \sum_j E_{ij}}$.

Modelos log-lineares para 3 variáveis (X,Y,Z)

1. Modelo de Independência Mútua (X, Y, Z)

   $$logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z$$

   $$i = 1, ..., r \quad j = 1, ..., c \quad k = 1, ...l$$

2. Modelo de X independente de Y e Z (X, YZ) (Independência marginal)

   $$logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{jk}^{yz}$$

Modelos log-lineares para 3 variáveis(X,Y,Z)

3. Modelo de X e Y independente dado Z (XY, YZ) (Independência condicional)

$$logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^k + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$$

4. Modelo de associação 2 a 2 (XY, XZ, YZ)

$$logE_{ijk} = \mu + \lambda_i^x + \lambda_i^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$$

5. Modelo de terceira ordem - Saturado (XYZ)

$$logE_{ijk} = \mu + \lambda_i^x + \lambda_i^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}$$

Exemplo Ilustrativo - Interpretação dos modelos

Raça x Ensino x Cidade

- X Ensino médio: sim/não
- Y Raça: branca (b), não branca ($\bar{b}$)
- Z Cidade: A e B

Caso 1: Independência mútua (X, Y, Z)

Modelo: $logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z$

Tabela: Cidade A

		Raça	
		b	$\bar{b}$
Ens	S	2	2
Médio	N	2	2

Tabela: Cidade B

		Raça	
		b	$\bar{b}$
Ens	S	2	2
Médio	N	2	2

Exemplo Ilustrativo

Caso 2: Independência Marginal (XY, Z)

Modelo: $logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy}$

Tabela: Cidade A

		Raça	
		b	$\bar{b}$
Ens	S	3	1
Médio	N	1	3

Tabela: Cidade B

		Raça	
		b	$\bar{b}$
Ens	S	3	1
Médio	N	1	3

Exemplo Ilustrativo

Caso 3: Independência Condicional (XZ, YZ)

Modelo: $logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$

Tabela: Cidade A

		Raça	
		b	$\bar{b}$
Ens	S	3	3
Médio	N	1	1

Tabela: Cidade B

		Raça	
		b	$\bar{b}$
Ens	S	3	1
Médio	N	3	1

Exemplo Ilustrativo

Caso 4: Dependência dois a dois (XY, XZ, YZ)

Modelo: $logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$

Tabela: Cidade A

		Raça	
		b	$\bar{b}$
Ens	S	2	2
Médio	N	2	2

Tabela: Cidade B

		Raça	
		b	$\bar{b}$
Ens	S	4	0
Médio	N	0	4

Observações

1. Parâmetros não tem interpretação.
2. Sempre que efeitos de alta ordem aparecer no modelo, os de ordem inferior também devem aparecer.
3. Redução de tabelas: se X independente de Y e Z (independência marginal), podemos somar sobre X e obter uma tabela de duas entradas.
4. Teste de M-H é o teste de independência condicional.
5. O modelo saturado apresenta tantos parâmetros quanto forem as caselas da tabela.

Ajustando Modelos log-lineares

- Utilizar o método de MV para estimar os parâmetros dos modelos.
- O objetivo principal é encontrar o modelo mais simples (interpretar a forma de associação) que ajusta os dados.

Estatísticas:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{(n_{ijk}-\widehat{E}_{ijk})^2}{\widehat{E}_{ijk}}$$

$$RV = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} n_{ijk} log\frac{n_{ijk}}{\widehat{E}_{ijk}}$$

Objetivo: Estimar $\widehat{E}_{ijk}$ e os respectivos, graus de liberdade.

Estimando Modelos log-lineares

- Lembre que as frequências esperadas (médias) são as mesmas para todos os desenhos amostrais.
- Selecionar um modelo log-linear e estimar os E_{ijk} por máxima verossimilhança.
- Selecionar o modelo adequado para os dados e interpretar as associações entre as variáveis.

Estimação MV para o Modelo de Poisson

Os $n'_{ijk}s$ são V.A's independentes de Poisson com frêquencias esperadas (médias) E_{ijk}

$$L() = \prod_{i=1}^{r} \prod_{j=1}^{c} \prod_{k=1}^{l} \frac{e^{-E_{ijk}} (E_{ijk})^{n_{ijk}}}{n_{ijk}!}$$

e tomando o logarítmo de $L(E)$ temos

$$l() = logL() \propto \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} [-E_{ijk} + n_{ijk} log(E_{ijk})]$$

Estimação MV para o Modelo de Poisson

- $\log E_{ijk}$ define o modelo.
- As frequências estimadas para o modelo saturado são as frequências observadas.
- Por exemplo, estimar as frequências esperadas para:
  Modelo Marginal

$$\log(E_{ijk}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy}$$

$$l(E) = -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} exp[\mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{zy}] + n\mu +$$

$$\sum_{i=1}^{r} n_{i++}\lambda_i^x + \sum_{j=1}^{c} n_{+j+}\lambda_j^y + \sum_{k=1}^{l} n_{++k}\lambda_k^z + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij+}\lambda_{ij}^{xy}$$

Estimação MV para o Modelo de Poisson

- Encontrar o EMV dos parâmetros do modelo e a partir deles $\widehat{E}_{ijk}$.
- Para alguns modelos temos que usar métodos iterativos.
- Uma forma alternativa é usar um método indireto, pois $\widehat{E}_{ijk}$ é a mesma para todos os desenhos amostrais e restrições do modelo.

Estimando os E_{ijk} utilizando o desenho multinomial

1. Modelo de independência mútua (X, Y, Z)

   $H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ ou

   $H_0 : E_{ijk} = n\pi_{i++}\pi_{+j+}\pi_{++k} = \frac{E_{i++}E_{+j+}E_{++k}}{n^2}$

   então, $\hat{E}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2}$

2. Modelo de independência marginal (X, YZ)

   $H_0 : \pi_{ijk} = \pi_{i++}\pi_{+jk}$

   $H_0 : E_{ijk} = \frac{E_{i++}E_{+jk}}{n}$

   então, $\hat{E}_{ijk} = \frac{n_{i++}n_{+jk}}{n}$

Estimando modelos log-lineares

3. Modelo de independência condicional (XZ, YZ)

$$H_0 : \pi_{ij(k)} = \pi_{+j(k)}\pi_{i+(k)}$$

$$H_0 : \pi_{ij(k)} = \frac{\pi_{ijk}}{\pi_{++k}} = \frac{\pi_{+jk}}{\pi_{++k}}\frac{\pi_{i+k}}{\pi_{++k}}$$

$$E_{ijk} = \frac{E_{+jk}E_{i+k}}{E_{++k}} \text{ e } \hat{E}_{ijk} = \frac{n_{+jk}n_{i+k}}{n_{++k}}$$

4. Modelo condicionalmente dependente (XY, XZ, YZ)

$$H_0 : logE_{ijk} = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}$$

$$H_0 : \pi_{ijk} = \pi_{ij+}\pi_{i+k}\pi_{+jk}$$

$$E_{ijk} = \frac{E_{ij+}E_{i+k}E_{+jk}}{n^2} \text{ e } \hat{E}_{ijk} = \frac{n_{ij+}n_{i+k}n_{+jk}}{n^2}$$

Estimando Modelos log-lineares

- Os graus de liberdade são obtidos subtraindo os gls do modelo saturado com aquele de interesse.
- O modelo adequado é aquele que satisfaz os seguintes critérios:
  - X^2 e/ou G^2 não significativo (modelo adequado).
  - Modelo com menos parâmetros entre os que satisfazem à anterior (parcimônia).
- As estimativas foram obtidas a partir do modelo multinomial. No entanto, elas valem para o produto de multinomiais (marginais fixas) assim como para o produtos de Poissons.

Graus de liberade

- Comparar o modelo saturado ($r \times c \times l$) com o modelo reduzido(ajustado)
- Exemplo a seguir com o modelo de Poisson.
- Tabela: $2x2x2 = 8$ caselas

$(X, Y, Z) = 8 - 4 = 4$

$(XY, Z) = 8 - 5 = 3$

$(XY, XZ) = 8 - 6 = 2$

$(XY, XZ, YZ) = 8 - 7 = 1$

$(XYZ) = 8 - 8 = 0$

Graus de liberdade dos testes

Modelo	GL
(X,Y,Z)	(rcl) -r -c -l +2
(XY,Z)	(l-1)(rc-1)
(XZ,Y)	(c-1)(rl-1)
(YZ,X)	(r-1)(cl-1)
(XY,YZ)	c(r-1)(l-1)
(XY,XZ)	r(c-1)(l-1)
(XZ,YZ)	l(r-1)(c-1)
(XY,XZ,YZ)	(r-1)(c-1)(l-1)
(XYZ)	0

Exemplo: Consumo de drogas entre alunos de ensino médio

Exemplo: Pesquisa realizada com alunos do último ano do ensino médio no estado de Ohio, E.U., (Agresti, 2013, pag. 346)

X: Uso de álcool (sim, não)

Y: Uso de cigarro (sim, não)

Z: Uso de maconha (sim, não)

Uso de Álcool	Uso de cigarro	Uso de Sim	maconha não
Sim	Sim	911	538
	Não	44	456
Não	Sim	3	43
	Não	2	279

Análise Descritiva: Estratificando pelo Uso de Álcool

Álcool = Sim

		Maconha			
		Sim		Não	
Cigarro	Sim	911	(63%)	538	1449
	Não	44	(9%)	456	500

$$\hat{RC} = 18$$

Álcool = Não

		Maconha			
		Sim		Não	Total
Cigarro	Sim	3	(7%)	43	46
	Não	2	(1%)	279	281

$$\hat{RC} = 10$$

Análise Descritiva: Estratificando pelo Uso de Cigarro

Cigarro = Sim

		Maconha			
		Sim		Não	Total
Álcool	Sim	911	(63%)	538	1449
	Não	3	(7%)	43	49

$$\hat{RC} = 24$$

Cigarro = Não

		Maconha			
		Sim		Não	Total
Álcool	Sim	44	(9%)	456	500
	Não	2	(1%)	279	281

$$\hat{RC} = 13$$

Análise Descritiva: Estratificando pelo Uso de Maconha

Maconha: Sim

		Cigarro			
		Sim		Não	Total
Álcool	Sim	911	(95%)	44	955
	Não	3	(60%)	2	5

$$\widehat{RC} = 14$$

Maconha: Não

		Cigarro			
		Sim		Não	Total
Álcool	Sim	538	(54%)	456	994
	Não	43	(13%)	279	322

$$\widehat{RC} = 8$$

Exemplo:Valores Esperados (A: Álcool, C: Cigarro e M: Maconha)

Valores ajustados para os modelos log-lineares ($\widehat{E}_{ijk}$)

Uso de Álcool	Uso de Cigarro	Uso de Maconha	Modelos (A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
Sim	Sim	Sim	540	611	909	910	911
		Não	740	838	439	539	538
	Não	Sim	282	211	46	45	44
		Não	387	289	555	455	456
Não	Sim	Sim	91	19	4,8	3,6	3
		Não	124	27	142	42	43
	Não	Sim	47	119	0,24	1,4	2
		Não	65	163	180	280	279

Exemplo: A: Álcool, C: Cigarro e M: Maconha

Testes X^2 e G^2.

Modelo	G^2	X^2	GL	Valor-p
(A,C,M)	1286	1411,4	4	$< 0,001$
(A,CM)	534	505,6	3	$< 0,001$
(C,AM)	940	824,2	3	$< 0,001$
(M,AC)	844	704,9	3	$< 0,001$
(AC,AM)	497	443,8	2	$< 0,001$
(AC,CM)	92	80,8	2	$< 0,001$
(AM,CM)	188	177,6	2	$< 0,001$
(AC,AM,CM)	0,4	0,4	1	0,54
(ACM)	0	0	0	-

O modelo mais simples indica a associação entre as 3 variáveis (2 a 2)

Exemplo - Interpretações

- (AC,M)

  Condicional: $\widehat{RC}_{AC} = \frac{611x119}{211x19} = \frac{838x163}{289x27} = 17,7$

  Marginal: $\widehat{RC}_{AC} = \frac{(611+838)(119+163)}{(211+289)(19+27)} = 17,7$

- (AM,CM) Independência condicional de A e C controlando por M

  $logE_{ijk} = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$

  Condicional:

  $\widehat{RC}_{AC} = 1,0 = \frac{909x0,24}{46x4,8} = \frac{439x180}{555x142}$

  $\widehat{RC}_{AM} = \frac{909x142}{439x4,8} = 61,9$

  $\widehat{RC}_{CM} = 25,1$

Exemplo - Interpretações

Marginal: (Inapropriada)

$$\widehat{RC}_{AC} = \frac{(909+439)(0{,}24+180)}{(46+555{,}1)(4{,}8+142{,}2)} = 27$$

$$\widehat{RC}_{AM} = 61,9 \quad \widehat{RC}_{CM} = 25,1$$

- (AC,AM,CM) Permite todos os pares de associação mas mantém RC iguais (homogêneos) entre duas variáveis em cada nível da terceira

$$\widehat{RC}_{AC} = \frac{910x1{,}4}{45x3{,}6} = 7,9 = \frac{539x280}{455x42}$$

A chance de um estudante que fuma também beber é cerca de 8 vezes a chance daquele que não fuma. (Isso vale tanto para os que usam maconha quanto para os que não usam)

$$\widehat{RC}_{AM} = 19,8 \quad \widehat{RC}_{MC} = 17,3$$

Resumo

Estimando RC

	Cond			Marg		
Modelo	A-C	A-M	C-M	A-C	A-M	C-M
(A,C,M)	1,0	1,0	1,0	1,0	1,0	1,0
(AC,M)	17,7	1,0	1,0	17,7	1,0	1,0
(AM,CM)	1,0	61,9	25,1	27,0	61,9	25,1
(AM,CM,AC)	**7,9**	**19,8**	**17,3**	17,7	61,9	25,1

Observe a importância em se escolher ”o modelo correto/adequado”.

Exemplo 2

Exemplo (Pena de morte nos Estados Unidos (Agresti, 2013))

X:Raça do réu

Y:Raça da vítima

Z:Pena de morte

Modelo	G^2	χ^2	GL	Valor-p
(X,Y,Z)	402.84	(419.56)	4	0.000 (0.000)
(YZ,X)	385.96	(391.24)	5	0.000 (0.000)
(XZ,Y)	401.30	(410.15)	3	0.000 (0.000)
(XY,Z)	22.27	(19.70)	3	0.009 (0.000)
(XZ,YZ)	384.43	(386.58)	2	0.000 (0.000)
(XZ,XY)	20.73	(22.14)	2	0.000 (0.000)
(YZ,XY)	5.39	(5.81)	2	0.067 (0.055)
(XY,XZ,YZ)	0.38	(0.20)	1	0.540 (0.660)

Exemplo 2

A pena de morte é maior para o réu negro controlando por raça da vítima

Porque a associação entre pena de morte e raça do réu muda de direção quando controlamos por raça da vítima?

- Existe uma associação muito forte entre raça da vítima e do réu
- Pena de morte é mais provável quando a vítima é branca do que negra

Desta forma, brancos tendem a matar brancos e matando brancos é mais provável receber a pena de morte.

Paradoxo de Simpson: Condicional associação difere da marginal associação

Exemplo 2

Conclusão

- Estes resultados sugerem uma importante associação entre a raça do réu e a raça da vítima
- O modelo mais simples que se ajusta a estes dados é o (YZ,XY). A pena de morte é independente da raça do réu, dado a raça da vítima

Considerações Finais

- Os modelos log-lineares são úteis para investigar a estrutura de associação entre variáveis categóricas, em especial para três variáveis.
- É possível utilizá-los para situações envolvendo mais que três variáveis. No entanto, a interpretação final fica extremamente complexa.