Modelos de regressão para dados politômicos

Prof. Caio Azevedo

Exemplo 16: Estudo sobre hábitos alimentares de jacarés

- O conjunto de dados foi extraído de Agresti (2007).
- 59 jacarés foram aleatoriamente selecionados do Lago George, Flórida, EUA.
- De cada um deles foi medido o comprimento (em metros) e também o volume do principal tipo de alimento existente no estômago.
- Classificação do tipo de alimento: peixe (P), invertebrado (I) e outros (O).

Exemplo 16 (cont.)

- Invertebrados: cobras, insetos aquáticos, lagostim.
- Outros: anfíbios, mamíferos, material vegetal, pedras e outros detritos, tartarugas e inclusive restos de jacarés bebês.
- Objetivo: verificar se existe alguma influência do comprimento dos jacarés nos hábitos alimentares.
- **Observação: estamos considerando que as categorias (P,I,O) são nominais**.

Exemplo 16 (cont.)

- Variáveis resposta: vetor que indica o tipo de alimento encontrado no estômago do jacaré (P, I, O).
- Temos assim, um vetor aleatório trinomial (Bernoulli trivariada), que assume valor 1 para a categoria à qual pertence o tipo de alimento encontrado no estômago do jacaré e 0, para as outras.
- Categorias: 1- P; 2-I; 3-O
- Variável explicativa: comprimento do jacaré.

Banco de dados (versão 1)

jacaré	comprimento	tipo de alimento
1	1,24	I
2	1,30	I
3	1,30	I
4	1,32	P
5	1,32	P
⋮	⋮	⋮
57	3,68	O
58	3,71	P
59	3,89	P

Banco de dados (versão 2)

jacaré	comprimento	tipo de alimento		
		I	P	O
1	1,24	1	0	0
2	1,30	1	0	0
3	1,30	1	0	0
4	1,32	0	1	0
5	1,32	0	1	0
⋮	⋮	⋮		
57	3,68	0	0	1
58	3,71	0	1	0
59	3,89	0	1	0

Modelo de regressão (logitos de referência)

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2})' \overset{ind.}{\sim} \text{trinomial}(1, \mathbf{p}_i), \mathbf{p}_i = (p_{i1}, p_{i2})',$$

$$\ln(p_{i1}/p_{i3}) = \beta_{01} + \beta_{11}x_i \quad ; \quad \ln(p_{i2}/p_{i3}) = \beta_{02} + \beta_{12}x_i,$$

$$p_{i3} = 1 - p_{i1} - p_{i2} \qquad i = 1, 2, ..., 59$$

- Y_{ij} : 1 se o i-ésimo indivíduo pertence à j-ésima (j=1,2,3) categoria e 0 caso contrário.
- x_i : comprimento do j-ésimo jacaré centrado na média dos valores observados $\overline{x} = \frac{1}{59}\sum_{i=1}^{59} x_i^*$, em que x_i^* é o comprimento do j-ésimo jacaré; β_{1j} : parâmetro associado ao incremento (positivo) na probabilidade do indivíduo i pertencer à categoria j para o aumento em uma unidade no comprimento.

Intepretação dos parâmetros

- $\ln(p_{ij}/p_{i3}), j = 1,2$ representa o log da chance do jacaré i pertencer à categoria j em relação à pertencer a categoria 3 (referência).
- Note que quaisquer outras chances são obteníveis a partir das duas anteriores.

$$\begin{aligned} \ln(p_{i1}/p_{i2}) &= \ln\left(\frac{p_{i1}/p_{i3}}{p_{i2}/p_{i3}}\right) = \ln(p_{i1}/p_{i3}) - \ln(p_{i2}/p_{i3}) \\ &= (\beta_{01} - \beta_{02}) + (\beta_{11} - \beta_{12})x_i \end{aligned}$$

Intepretação dos parâmetros

- Note que $p_{ij}/p_{i3} = e^{\beta_{0j}+\beta_{1j}x_i} \rightarrow p_{ij} = e^{\beta_{0j}+\beta_{1j}x_i}p_{i3}, j = 1, 2$

  assim,

$$
\begin{aligned}
p_{i1} + p_{i2} + p_{i3} &= 1 \rightarrow p_{i1}/p_{i3} + p_{i2}/p_{i3} + 1 = 1/p_{i3} \\
\rightarrow p_{i3} &= \frac{1}{p_{i1}/p_{i3} + p_{i2}/p_{i3} + 1}
\end{aligned}
$$

- Logo, $p_{i1} = \dfrac{e^{\beta_{0j}+\beta_{1j}x_i}}{1 + e^{\beta_{01}+\beta_{11}x_i} + e^{\beta_{02}+\beta_{12}x_i}}, j = 1, 2$ e

  $p_{i3} = \dfrac{1}{1 + e^{\beta_{01}+\beta_{11}x_i} + e^{\beta_{02}+\beta_{12}x_i}}$.

Intepretação dos parâmetros

- Se, $x_i = 0$, então $p_{ij} = \frac{e^{\beta_{0j}}}{1+e^{\beta_{01}}+e^{\beta_{02}}}, j = 1, 2$
- Chances (entre pertencer e não pertencer a cada categoria) $p_{ij}/(1 - p_{ij}) = \frac{e^{\beta_{0j}+\beta_{1j}x_i}}{1+e^{\beta_{0j'}+\beta_{1j'}x_i}}, j \neq j' \in \{1, 2\}$ podem ser estimadas a partir dos valores de $p_{ij}, j = 1, 2, 3$.
- Se $\beta_{1j} > 0$ quanto maior seu valor, maior a probabilidade do indivíduo pertencer à categoria j, ocorrento o contrário se $\beta_{1j} < 0$.
- **Exercício: obter as razões de chances**.

Modelo de regressão geral (logitos de referência)

$$\begin{aligned}
\mathbf{Y}_i &\overset{ind.}{\sim} \text{multinomial}_k(m_i, \mathbf{p}_i), \mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{i(k-1)})', \\
\ln(p_{ij}/p_{iJ}) &= \sum_{r=1}^{p} x_{ri}\beta_{rj} = \eta_{ij}, i = 1, 2, ..., n
\end{aligned}$$

- $\mathbf{Y}_i$ é um vetor de tamanho $k-1$, retirando-se a categoria J, a qual é a categoria de referência. Ou seja, $\mathbf{Y}_i = (Y_{i1}, ..., Y_{ik})'$ retirando-se a variável Y_{iJ}, analiogamente para o vetor $\mathbf{p}_i = (p_{i1}, ..., p_{ik})'$. Seja ainda $A = \{1, 2, ..., k\} - \{J\}$.
- Y_{ij} : 1 se o i-ésimo indivíduo pertence à j-ésima ($j \in A$) categoria e 0 caso contrário.
- x_{ri} : valor da $r-$ésima categoria associada ao i-ésimo indivíduo.

Intepretação dos parâmetros

- $p_{ij} = \frac{e^{\eta_{ij}}}{\sum_{j=1}^{k} e^{\eta_{ij}}}, \forall i, j$. Se j =J, então $\beta_{rJ} = 0, \forall r$ e, assim $\eta_{rJ} = 0$.
- As outras quantidades têm interpretações análogas.

Estimação

- Verossimilhança (produto de multinomiais)

  $L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} p_{ij}^{y_{ij}}, \sum_{j=1}^{p} y_{ij} = m_i, \forall i, y_{ij} \in \{0, 1, ..., m_i\}, \forall i, j$

- Verossimilhança (produto de multinomiais)

  $l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_{ij}^{y_{ij}} + const.$

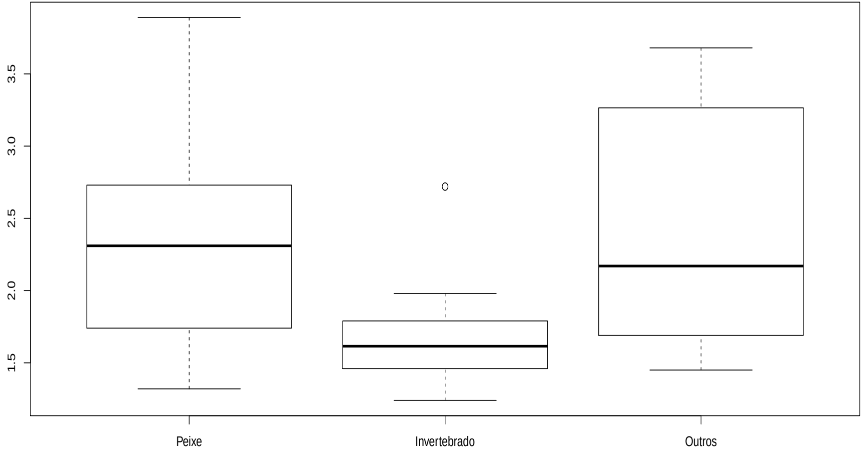
- O processo de maximização é conduzido de forma numérica (algoritmo de Newton-Raphson ou Escore de Fisher, por exemplo).

- Seja $\widehat{\boldsymbol{\beta}}$ o emv de $\boldsymbol{\beta}$. Em geral, para n suficientemente grande, $\widehat{\boldsymbol{\beta}} \approx N_p(\beta, I^{-1}(\boldsymbol{\beta}))$.

Validação dos modelos

- Os resultados vistos para os MLG's (testes de hipótese, desvio, análise de diagnóstico) podem ser adaptados para a classe de modelos em questão.
- Por exemplo, os testes $\mathbf{C}\boldsymbol{\beta}$, podem ser conduzidos de maneira análoga.
- A análise e diagnóstco pode ser feita para cada uma as variáveis do vetor $\mathbf{Y}$ as quais, nesse caso, tem distribuição binomial.

Boxplots



Medidas descritivas

Alimento	Média	Mediana	DP	Var.	CV(%)	Mín.	Máx.
P	2,36	2,31	0,76	0,58	32,23	1,32	3,89
I	1,66	1,61	0,33	0,11	19,70	1,24	2,72
O	2,42	2,17	0,88	0,78	36,44	1,45	3,68

Ajuste do modelo

Parâmetro	Estimativa	EP	Estat. Z	p-valor
β_{01} (peixe)	1,383	0,422	3,277	0,0011
β_{02} (invertebrado)	0,445	0,544	0,819	0,4131
β_{11} (peixe)	-0,110	0,517	-0,213	0,8314
β_{12} (invertebrado)	-2,467	0,900	-2,740	0,0061

Há uma equivalência entre as categorias "peixe" e "outras", em termos do coeficiente angular e entre as categorias "invertebrado" e "outras" em relação ao intercepto.

Proporções (probabilidades) estimadas

- Peixe: $p_{P_i} = \dfrac{e^{1,383}}{1 + e^{1,383} + e^{0,445 - 2,467x_i}}$
- Invertebrado : $p_{I_i} = \dfrac{e^{0,445 - 2,467x_i}}{1 + e^{1,383} + e^{0,445 - 2,467x_i}}$
- Outros : $p_{O_i} = \dfrac{1}{1 + e^{1,383} + e^{0,445 - 2,467x_i}}$

Proporções observadas e preditas pelo modelo