Modelos de regressão para dados politômicos

Prof. Caio Azevedo

Exemplo 17: 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 17 (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 17 (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)

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

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

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

- 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 i-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 i-ésimo jacaré; β_{1j} : parâmetro associado ao incremento (positivo/negativo) 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

- Além disso, temos 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)

$$\boldsymbol{Y}_i \overset{ind.}{\sim} \text{multinomial}_k(m_i, \boldsymbol{p}_i), \boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{i(k-1)})',$$

$$\ln(p_{ij}/p_{iJ}) = \sum_{r=1}^{p} x_{ri}\beta_{rj} = \eta_{ij}, i = 1, 2, ..., n$$

- $\boldsymbol{Y}_i$ é um vetor de tamanho $k-1$, retirando-se a categoria J, a qual é a categoria de referência. Ou seja, $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{ik})'$ retirando-se a variável Y_{iJ}, analiogamente para o vetor $\boldsymbol{p}_i = (p_{i1}, ..., p_{ik})'$. 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

- $p_{ij} = \dfrac{e^{\eta_{ij}}}{\sum_{j=1}^{k} e^{\eta_{ij}}}, \forall i, j$. 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)

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

- Verossimilhança (produto de multinomiais)

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

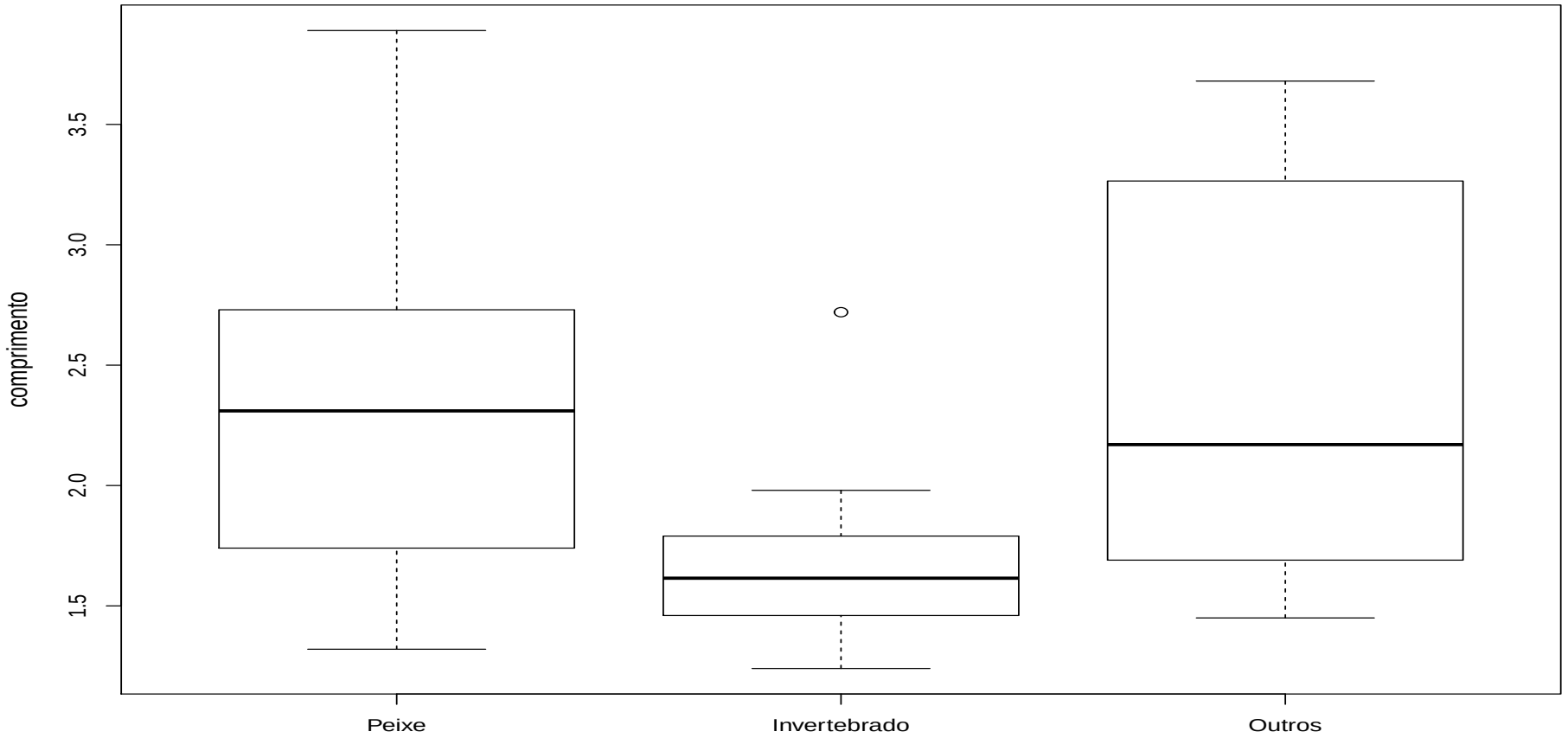
- O processo de maximização é conduzido de forma numérica (algoritmo de Newton-Raphson ou Escore de Fisher, por exemplo).

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

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 $\boldsymbol{C}\boldsymbol{\beta}$, podem ser conduzidos de maneira análoga.
- A análise e diagnóstco pode ser feita para cada uma as variáveis do vetor $\boldsymbol{Y}$ 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

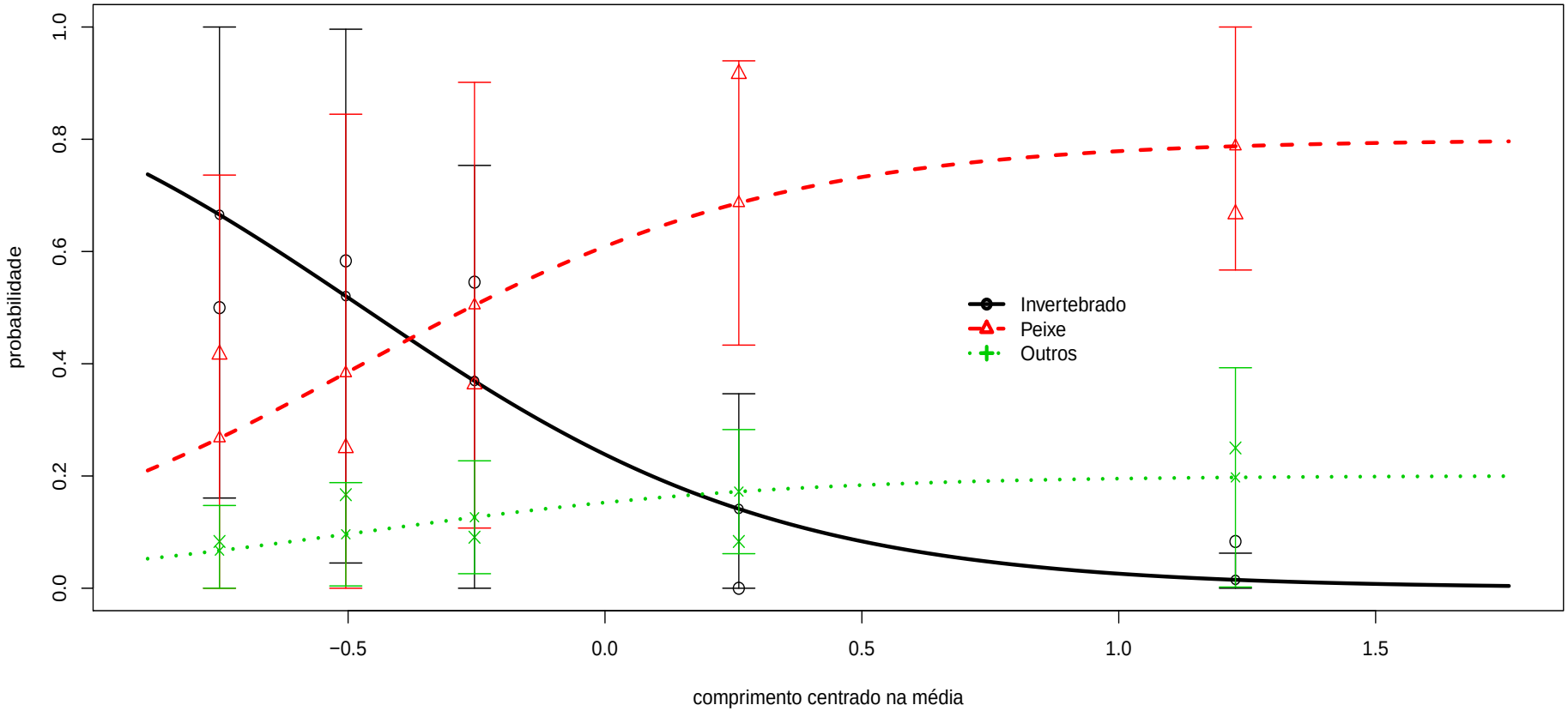
Parâmetro	Estimativa	EP	Estat. Z	p-valor
β_{01} (peixe)	1,383	0,422	3,277	0,0011
β_{02} (invertebrado)	0,445	0,544	0,819	0,4131
β_{11} (peixe)	-0,110	0,517	-0,213	0,8314
β_{12} (invertebrado)	-2,467	0,900	-2,740	0,0061

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.

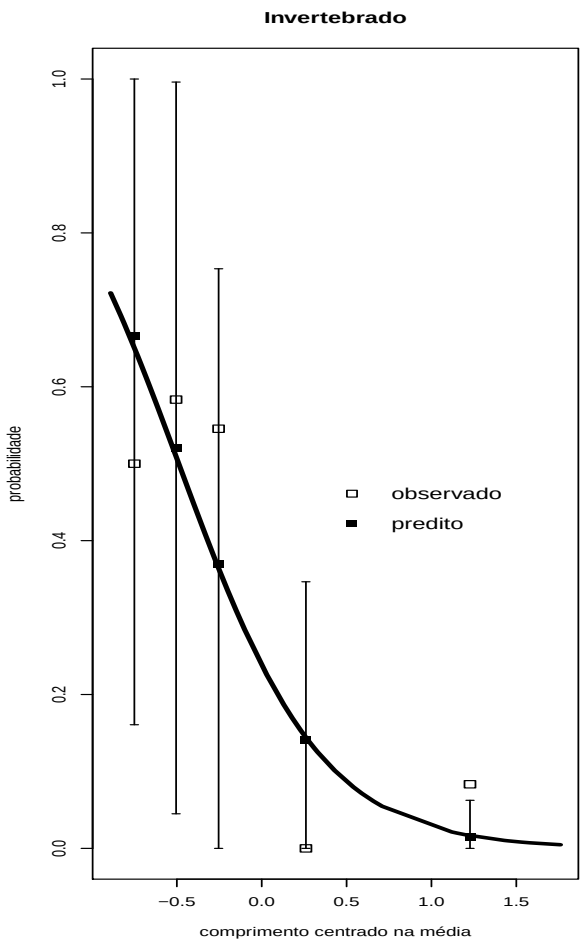
Proporções (probabilidades) estimadas

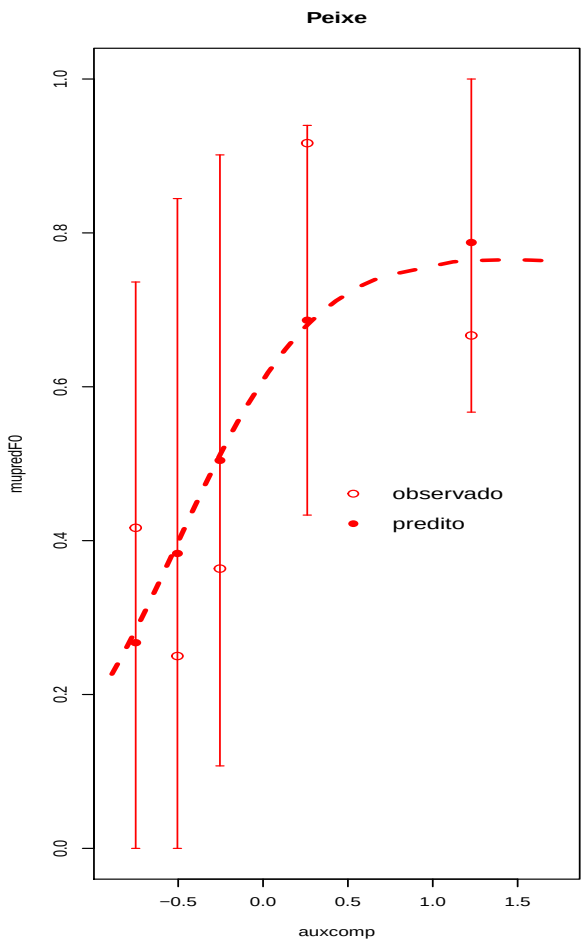
- Peixe: $p_{P_i} = \dfrac{e^{1,383}}{1 + e^{1,383} + e^{0,445-2,467x_i}}$
- Invertebrado : $p_{I_i} = \dfrac{e^{0,445-2,467x_i}}{1 + e^{1,383} + e^{0,445-2,467x_i}}$
- Outros : $p_{O_i} = \dfrac{1}{1 + e^{1,383} + e^{0,445-2,467x_i}}$

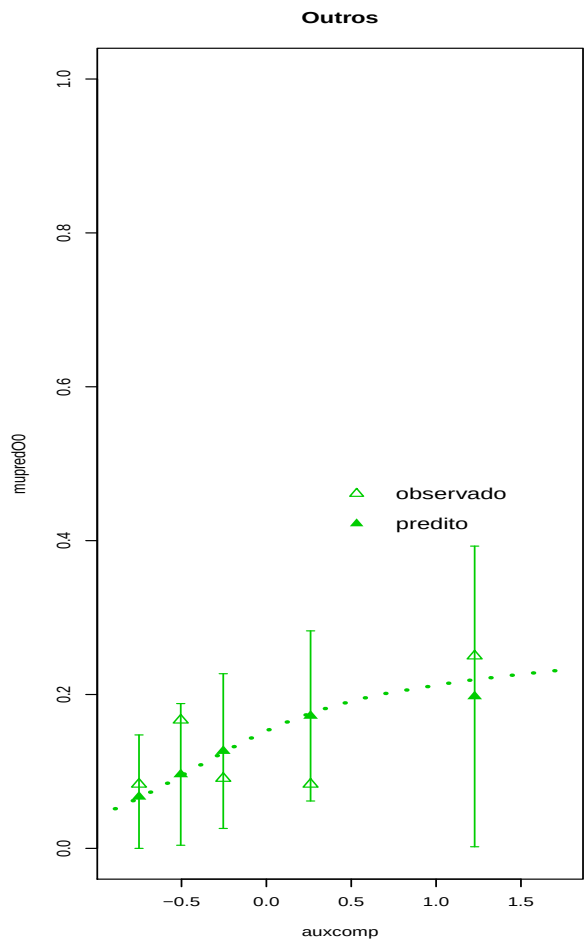
Proporções observadas e preditas pelo modelo



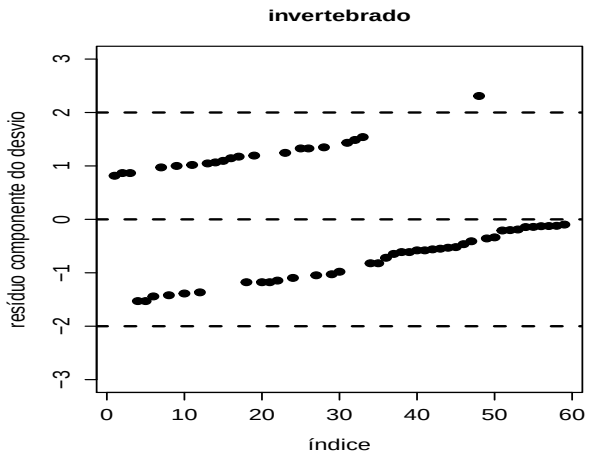
Proporções observadas e preditas pelo modelo

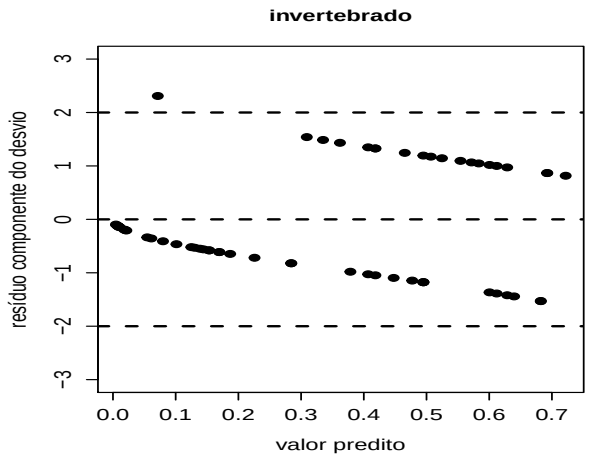


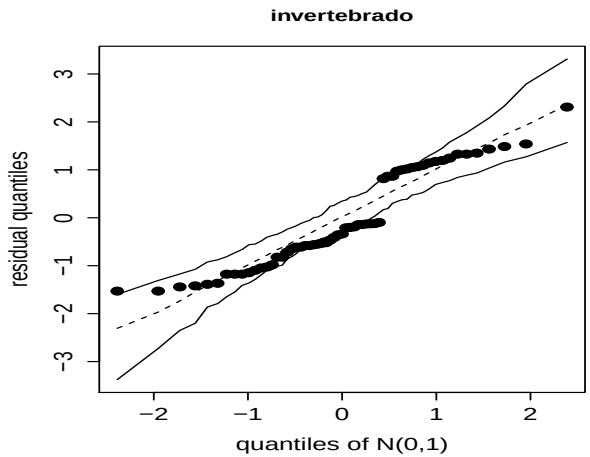


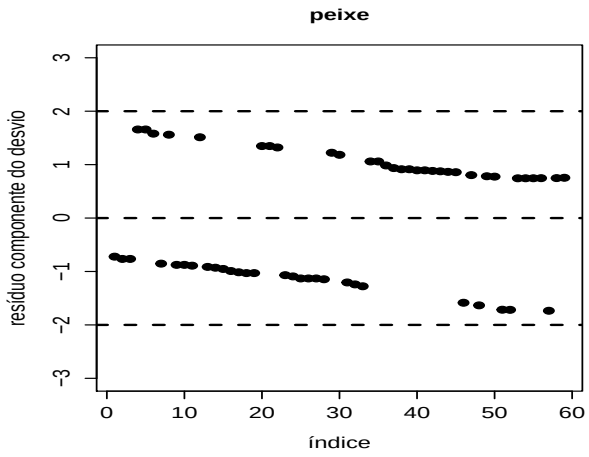


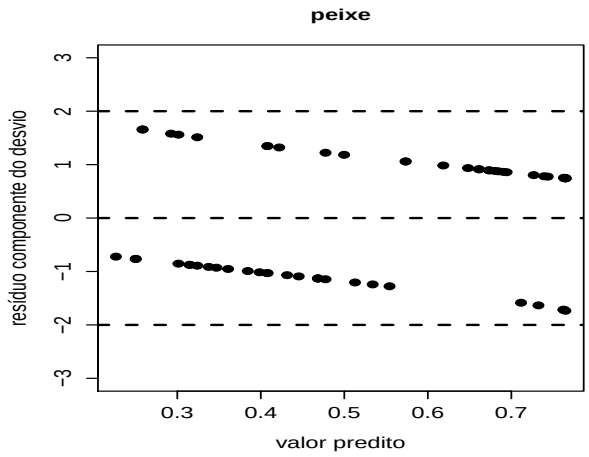
Análise do resíduo componente do desvio

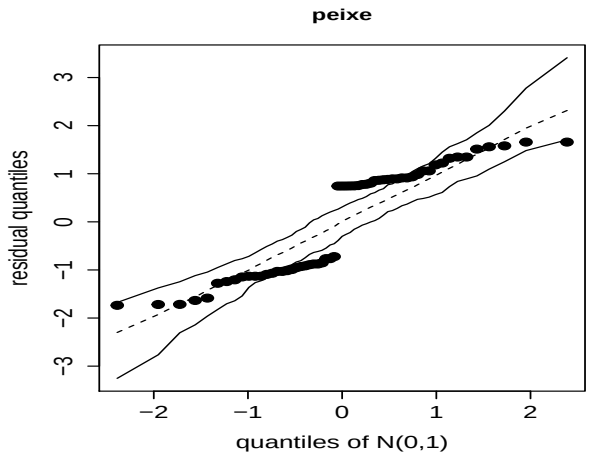












Desempenho de classificação

Classificação realizada a partir de valores simulados de 59 tetranomiais com parâmetros $\widetilde{p}_i$.

	Verdadeiro		
Predito	F	I	O
F	18	7	2
I	9	11	4
O	4	2	2

Desempenho de classificação

Estatísticas gerais

Acurácia (Ac) : 0,5254

IC(95%) : (0,3912, 0,657)

Taxa de não infomação (TNI) : 0,5254

P-Value [Acur $>$ TNI] : 0,5527

Kappa : 0,2133

Desempenho de classificação

Estatísticas por classe:

	Classe: F	Classe: I	Classe: O
Sensibilidade	0,5806	0,5500	0,2500
Especifidade	0,6786	0,6667	0,8824
Valor preditivo positivo	0,6667	0,4583	0,2500
Valor preditivo negativo	0,5938	0,7429	0,8824
Prevalência	0,5254	0,3390	0,1356
Taxa de detecção	0,3051	0,1864	0,0339
Prevalência de detecção	0,4576	0,4068	0,1356
Acurácia balanceada	0,6296	0,6083	0,5662

	Observado	
Predito	Evento	Não-Evento
Evento	A	B
Não Evento	C	D

Fórmulas:

- Sensibilidade = A/(A+C)
- Especifidade = D/(B+D)
- Prevalência = (A+C)/(A+B+C+D)
- Valor preditivo positivo = (Sensibilidade * Prevalência)/((Sensibilidade*Prevalência) + ((1-Especifidade)*(1-Prevalência)))

Fórmulas:

- Valor preditivo negativo = (Especifidade * (1-Prevalência))/(((1-Sensibilidade)*Prevalência) + ((Especifidade)*(1-Prevalência)))
- Taxa de detecção = A/(A+B+C+D)
- Prevalência de detecção= (A+B)/(A+B+C+D)
- Acurácia balanceada = (sensitivity+specificity)/2

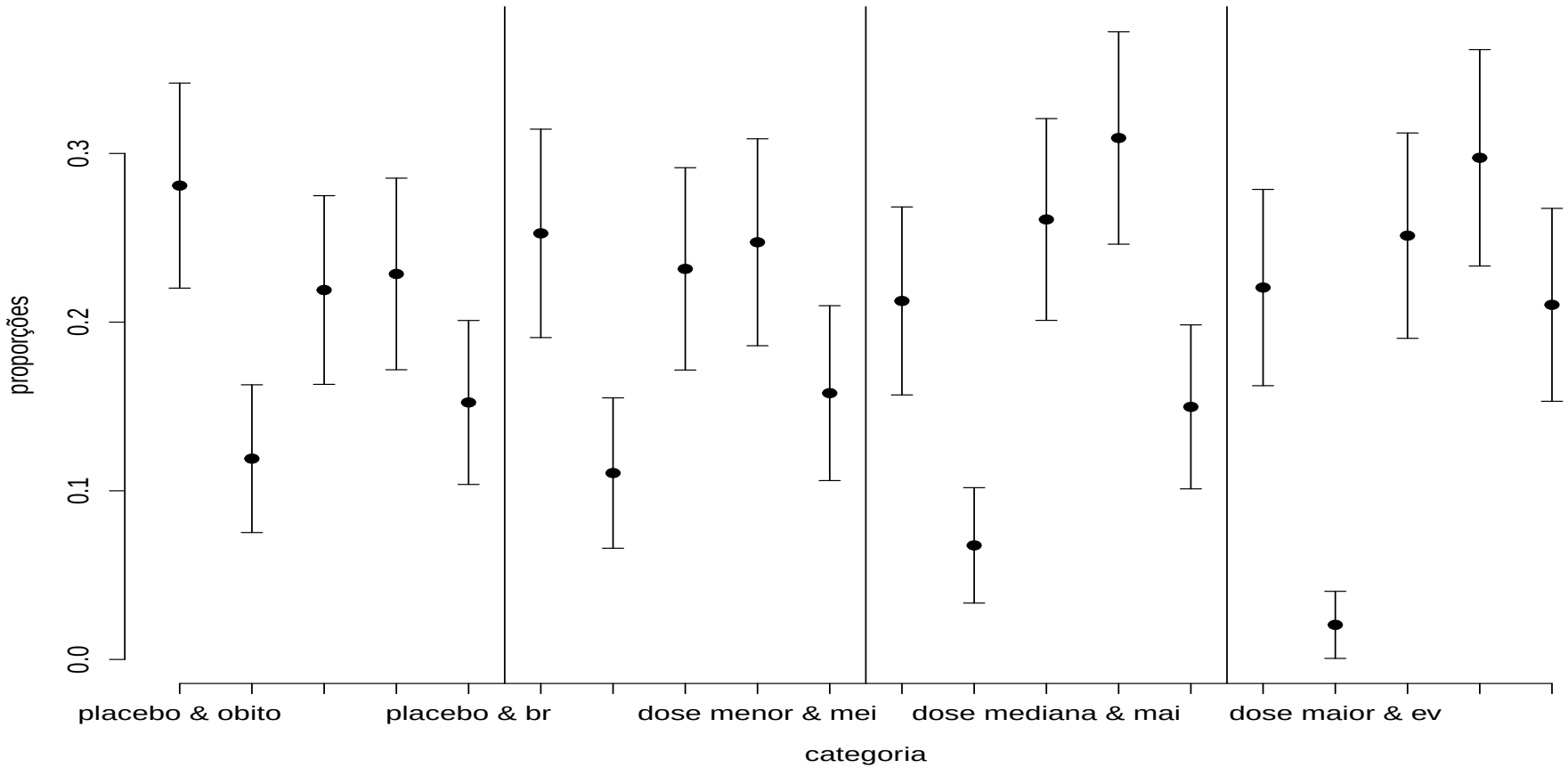
Exemplo 18: Efeito das doses de medicamentos intravenosos em pacientes com trauma hemorrágico subaracnóide

- O conjunto de dados foi extraído de Agresti (2010).
- Corresponde a distribuição de pacientes, que receberam algum tipo de tratamento (dose de medicamento intravenoso), de acordo com o desfecho clínico (óbito, estado vegetativo, incapacidade maior, incapacidade menor, boa recuperação), de acordo com a escala de resultado de Glasgow.
- Os totais de paciente, por cada dose, foram fixados (produto de multinomiais).

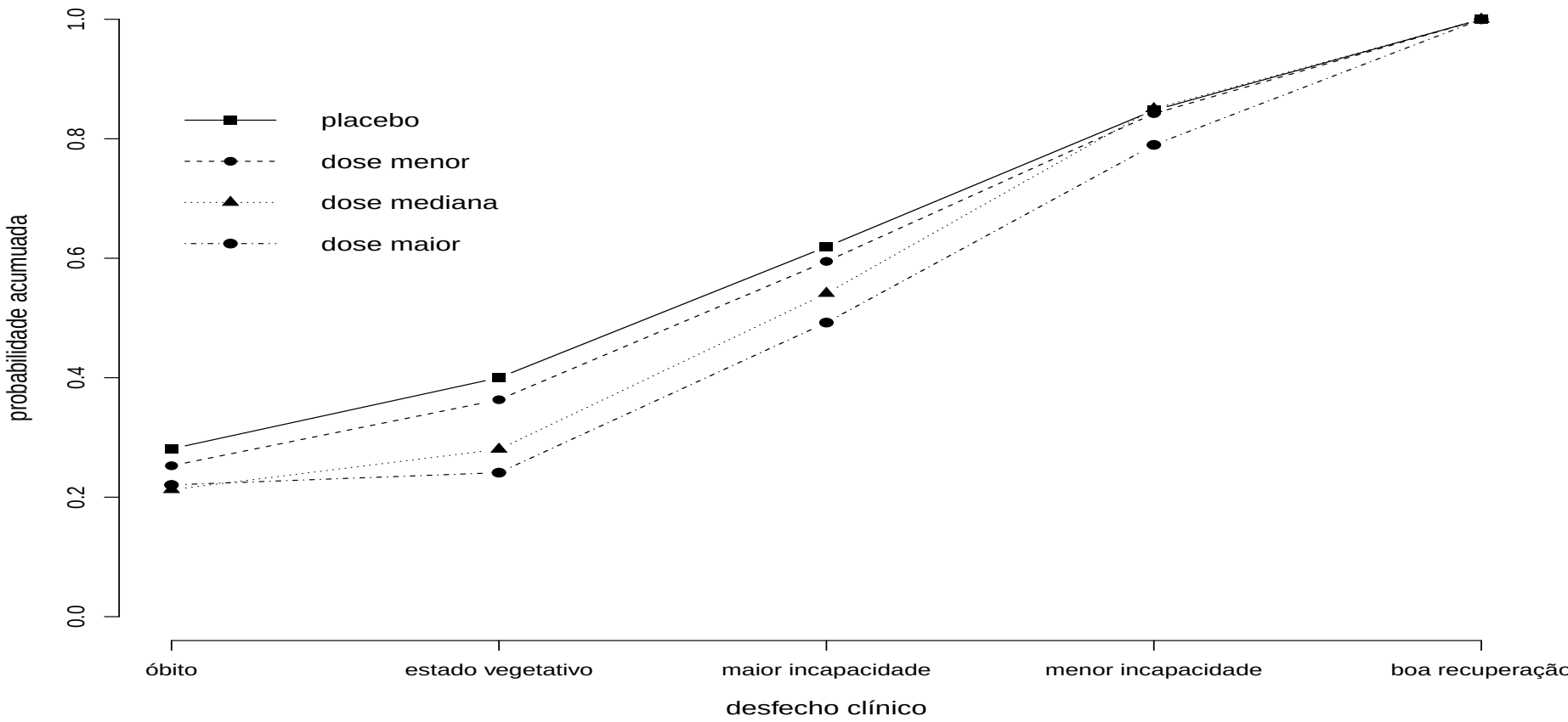
Dados

	Escala de resultado de Glasgow					
tratamento	óbito	est. veg.	mai. inc.	men. inc.	boa rec.	Total
placebo	59	25	46	48	32	210
dose menor	48	21	44	47	30	190
dose mediana	44	14	54	64	31	207
dose maior	43	4	49	58	41	195

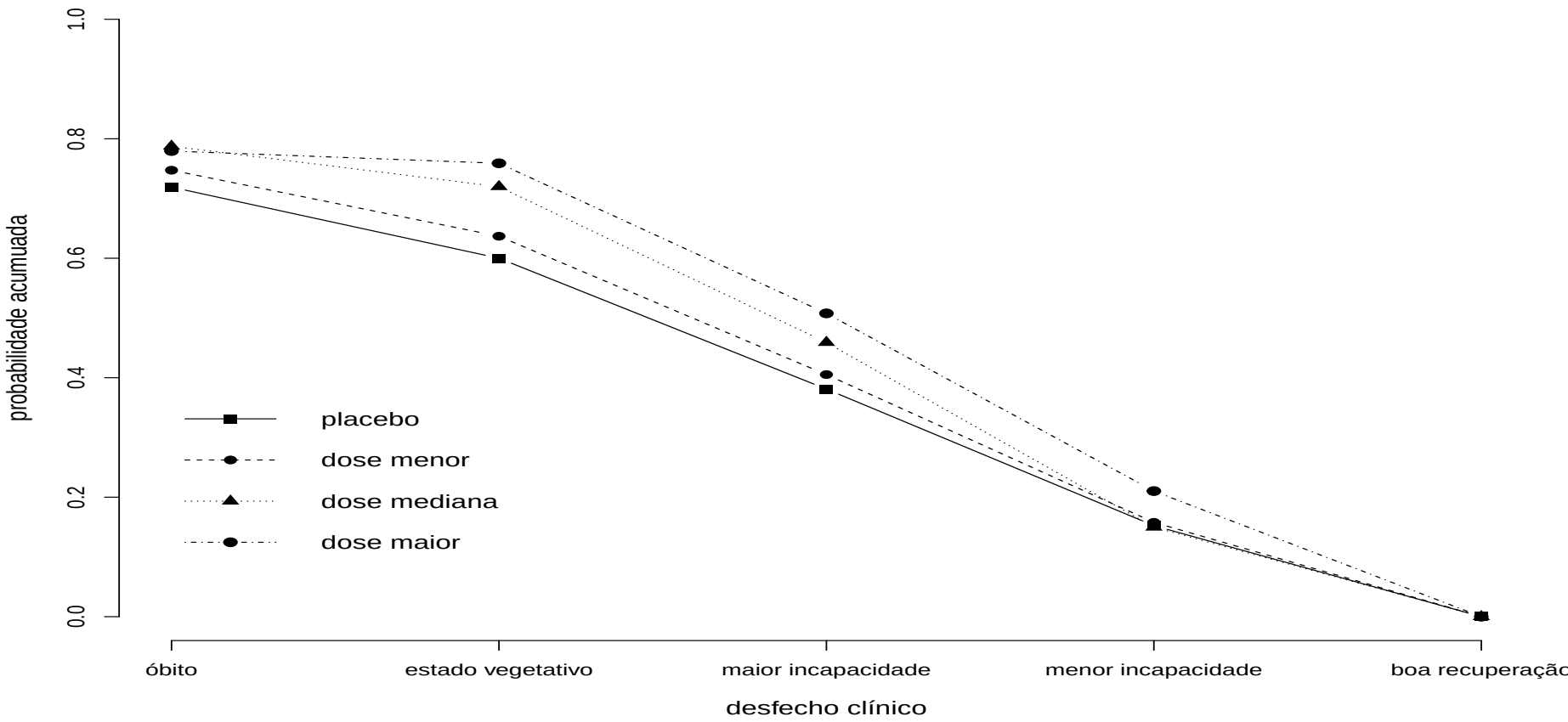
Análise descritiva



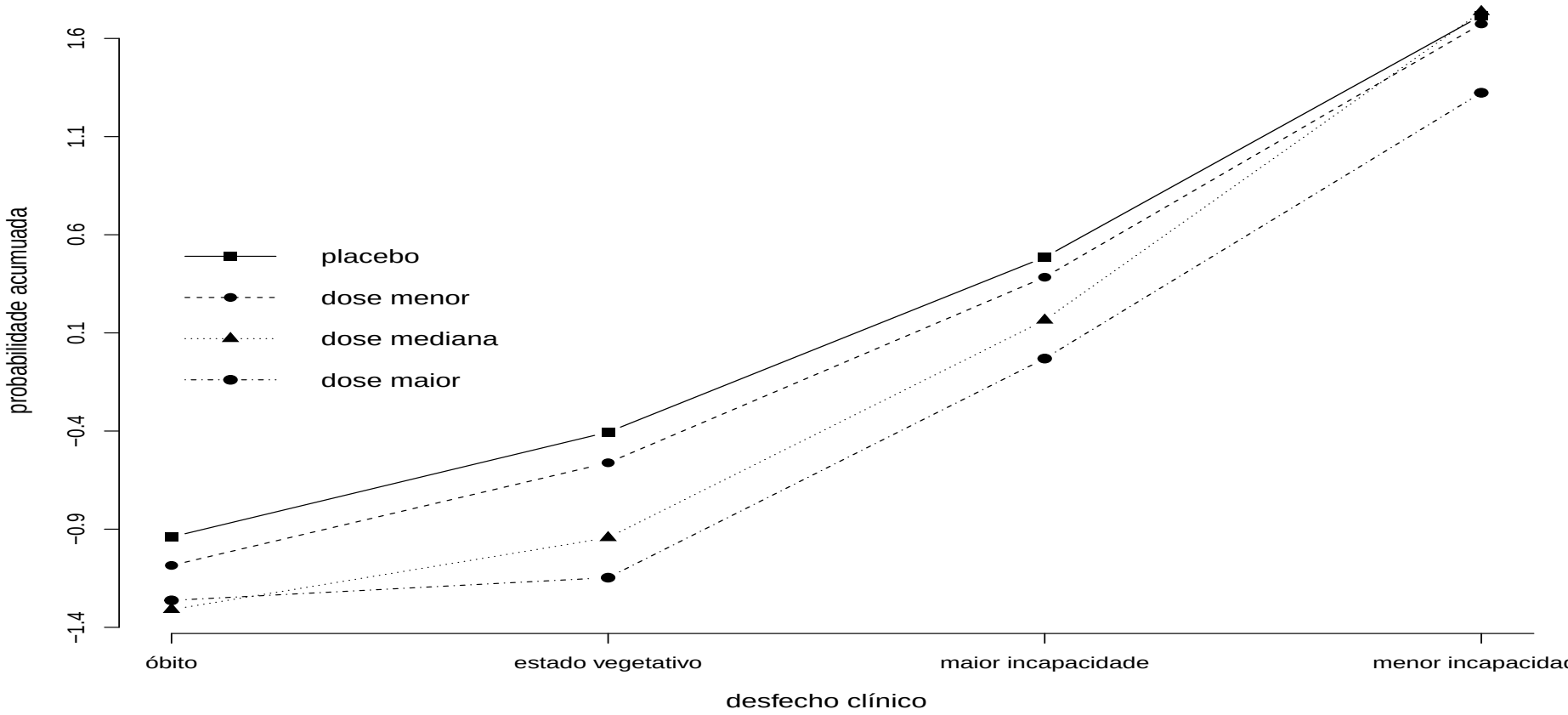
Função de distribuição acumulada



Função de sobrevivência



Logito(FDA/FDS)



Modelo de regressão (logitos cumulativos)

$$\begin{aligned}
\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})' &\overset{ind.}{\sim} \text{pentanomial}(n_i, \boldsymbol{p}_i) \\
\boldsymbol{p}_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5})', &\quad \sum_{j=1}^{5} Y_{ij} = n_i, \sum_{j=1}^{5} p_{ij} = 1, \\
\text{logito}(P(Y_{ij} \leq j)) = \text{logito}(P_{ij}) &= \alpha_j + \beta x_{ij}, x_{ij} = i, \\
&\quad i = 1, 2, 3, 4; j = 1, 2, 3, 4, 5,
\end{aligned}$$

$n_1 = 210, n_2 = 190, n_3 = 207, n_4 = 195$

- Y_{ij} : quantidade de indivíduos submetidos à dose i que apresentaram o desfecho clínico j.
- x_{ij} : valor da dose i, relativa ao desfecho clínico j.

Modelo de regressão (logitos cumulativos)

- $P(Y_{ij} \leq j) = \dfrac{e^{\alpha_j + \beta x_{ij}}}{1 + e^{\alpha_j + \beta x_{ij}}}$.
- $P(Y_{ij} = j) = P(Y_{ij} \leq j) - P(Y_{ij} \leq j-1) =$
$\dfrac{e^{\alpha_j + \beta x_{ij}}}{1 + e^{\alpha_j \beta x_{ij}}} - \dfrac{e^{\alpha_{j-1} + \beta x_{i(j-1)}}}{1 + e^{\alpha_{j-1} + \beta x_{i(j-1)}}}, P(Y_{ij} \leq 0) = 0$
- Para o mesmo nível j da resposta e diferentes níveis i da covariável:

$$\begin{aligned} \text{logito}(P_{i'j}) - \text{logito}(P_{ij}) &= \beta(x_{i'j} - x_{ij}) \rightarrow \\ \frac{P_{i'j}/(1-P_{i'j})}{P_{ij}/(1-P_{ij})} &= e^{\beta(x_{i'j} - x_{ij})} \end{aligned}$$

- Para o mesmo nível i da covariável e diferentes níveis j da resposta:

$$\begin{aligned} \text{logito}(P_{ij'}) - \text{logito}(P_{ij}) &= \alpha_{j'} - \alpha_j \rightarrow \\ \frac{P_{ij'}/(1-P_{ij'})}{P_{ij}/(1-P_{ij})} &= e^{\alpha_{j'} - \alpha_j} \end{aligned}$$

Estimativas dos parâmetros modelo

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z	p-valor
α_1	-0,72	0,16	[-1,03 ; -0,41]	-4,53	$<$ 0,0001
α_2	-0,32	0,16	[-0,63 ; -0,01]	-2,04	0,0417
α_3	0,69	0,16	[0,38 ; 1,00]	4,38	$<$ 0,0001
α_4	2,06	0,17	[1,72 ; 2,40]	11,84	$<$ 0,0001
β	-0,18	0,06	[-0,29 ; -0,07]	-3,12	0,0018

Análise preditiva