

# Reconhecimento de Padrões Discriminantes Lineares Regressão Logística

Prof. George von Borries  
Departamento de Estatística  
Universidade de Brasília

1 - 2024



# Modelos lineares

## ● Regressão:

$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$  com  $Y_i \stackrel{iid}{\sim} N(E(Y_i), \sigma^2)$ ,  
sendo  $X_1, \dots, X_p$  variáveis explicativas quantitativas conhecidas.

$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^2$  com  $Y_i \stackrel{iid}{\sim} N(E(Y_i), \sigma^2)$ ,  
também é um modelo de regressão linear (linear nos parâmetros).

## ● Análise de Variância:

$E(Y_i) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$  com  $Y_i \stackrel{iid}{\sim} N(E(Y_i), \sigma^2)$ ,  
sendo  $\alpha_i, \beta_j, (\alpha\beta)_{ij}$  variáveis qualitativas indicando níveis de efeitos de fatores  
(fixos ou aleatórios) e interações.

## ● Análise de Covariância:

$E(Y_i) = \mu + \alpha_i + \beta X_i$  com  $Y_i \stackrel{iid}{\sim} N(E(Y_i), \sigma^2)$ ,  
sendo  $\alpha_i$  qualitativa e  $X_i$  quantitativa.

## ● Modelos Lineares Generalizados:

$g[E(Y_i)] = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$  com  $Y_i \stackrel{iid}{\sim} FE(g[E(Y_i)], \sigma^2)$ ,  
uma distribuição da Família Exponencial.



## Modelo Linear Generalizado (GLM)

GLMs<sup>1</sup> são extensões de modelos de regressão por permitirem que a média de uma população dependa de uma **função linear de covariáveis** através de uma função de **ligação** não linear.

GLMs incluem, entre outros, os seguintes modelos:

- ① Regressão ordinária e modelos ANOVA.
- ② Regressão logística para respostas categóricas binárias ou com múltiplas categorias (distribuição binomial ou distribuição multinomial).
- ③ Modelos log-lineares para dados de contagem com distribuição de Poisson.

---

<sup>1</sup>A PROC GLM do SAS se refere a modelos lineares gerais, i.e., modelos ANOVA com distribuição normal.



## Componentes de um Modelo Linear Generalizado (GLM)

- ① O **componente aleatório** identifica a variável resposta  $\mathbf{Y}$  e sua distribuição de probabilidade. As observações  $\mathbf{Y} = (y_1, \dots, y_n)^T$  são tratadas como independentes.
- ② O **preditor linear ou componente sistemático** é especificado como  $\mathbf{X}\beta$  para um vetor de parâmetros  $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$  e uma matriz de delineamento  $\mathbf{X}$  que contem valores de  $p$  variáveis explicativas de  $n$  observações.
- ③ A **função de ligação**  $g$  é aplicada a cada componente  $E(\mathbf{Y})$  relacionando esta resposta média ao preditor linear,  $g[E(\mathbf{Y})] = \mathbf{X}\beta$ .



## ① Componente Aleatório

- Para uma amostra de tamanho  $n$ , as observações da variável resposta  $Y$  são representadas por  $\mathbf{Y} = (y_1, \dots, y_n)^\top$ .
- O componente aleatório de um GLM consiste em identificar a variável resposta  $Y$  e a distribuição de probabilidade para  $Y_1, \dots, Y_n$ .
- GLMs tratam  $Y_1, \dots, Y_n$  como sendo variáveis aleatórias independentes com distribuição de probabilidade da família exponencial,

$$f(y_i; \theta_i) = \exp [a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)].$$

Se  $a(y_i) = y_i$  a distribuição está na forma padrão (canônica) e  $b(\theta_i)$  é chamado de parâmetro natural.

- Na regressão ordinária,  $Y$  tem distribuição Normal com variância constante  $\sigma^2$ .
- Obs. A distribuição Normal faz parte da família exponencial.
- Observação:**

Outra forma de representar a família exponencial é,

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}.$$



## ② Componente Sistemático (Preditor Linear)

- Para uma amostra de tamanho  $n$ , as variáveis explicativas  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$  são relacionadas a parâmetros  $\{\eta\}$  pertencentes a  $\{E(Y_i)\}$  através de uma combinação linear, i.e,

$$\eta_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

ou na forma matricial,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

com  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ ,  $\boldsymbol{\beta}$  um vetor coluna  $p \times 1$  de parâmetros do modelo, e  $\mathbf{X}$  uma matriz  $n \times p$  de variáveis explicativas com valores  $\{x_{ij}\}$ .

- Note que o modelo é **linear** nos parâmetros. As variáveis explicativas podem ser funções não lineares, como interações ( $x_{i3} = x_{i1}x_{i2}$ ), termos quadráticos ( $x_{ij}^2$ ), etc.
- GLMs tratam  $y_i$  como variáveis aleatórias e  $X_i$  como fixo e por isso o preditor linear é chamado de **componente sistemático**.



## 3 Função de Ligação

- A função de ligação conecta a componente aleatória com o preditor linear.
- Seja  $\mu_i = E(Y_i)$ ,  $i = 1, \dots, n$ . GLM conecta  $\eta_i$  a  $\mu_i$  através de uma função monótona e diferenciável  $g(\cdot)$ , i.e,

$$\eta_i = g(\mu_i) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

- A função de ligação  $g$  que transforma o  $\mu_i$  no parâmetro natural da família exponencial é chamada de ligação canônica. Exemplos são:
  - A média para distribuição normal (ligação identidade)

$$\mu_i = E(Y_i) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

- O log da chance para a distribuição Binomial (ligação logito)

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

- O log da média para a distribuição Poisson (ligação log)

$$\log \mu_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$



## Regressão Logística binária

- $Y_i = 1$  se  $\mathbf{x} \in \omega_1$  e  $Y_i = 0$  se  $\mathbf{x} \in \omega_2$ .
- $\mathbf{x}$  é o espaço de características com valores discretos e/ou contínuos.
- Considere  $\pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1 | \mathbf{X} = \mathbf{x}_i)$ ,  $i = 1, \dots, n$ .
- $Y_i \sim \text{Bernoulli}(\pi)$  com  $E(Y_i) = \pi$ ;  $\text{Var}(Y_i) = \pi(1 - \pi)$

Neste caso,

$$\text{logito } [\pi_i] = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \log \left( \frac{P(Y_i = 1 | \mathbf{X} = \mathbf{x}_i)}{P(Y_i = 0 | \mathbf{X} = \mathbf{x}_i)} \right) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}$$

(A razão  $\frac{\pi_i}{1 - \pi_i}$  é chamada de chance ou *odds*)

A solução será,

$$\pi_i = \frac{\exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}\}}{1 + \exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}\}} \quad \text{e} \quad 1 - \pi_i = \frac{1}{1 + \exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}\}}.$$

A decisão será alocar  $\mathbf{x}$  a  $\omega_1$  se  $\pi_i > k$ , sendo  $k$  um ponto de corte (geralmente,  $k = 0.5$ ).



Este é um modelo linear generalizado com

- Ligação: Logito (log-odds)
- Preditor: linear ( $\beta_0 + \beta^T x$ )
- Distribuição da resposta: Binomial

Note que,

- $E(Y)$  é não linear.
- $\pi_i$  é monótono, i.e,
  - $\pi_i \geq \pi_j$  para  $x_i > x_j$       ou
  - $\pi_i \leq \pi_j$  para  $x_i > x_j$
- $0 \leq \pi_i \leq 1$ .
- a curva resposta tem formato em S.



## Problemas com o Modelo de Regressão Linear

Um modelo simples relacionando  $\pi$  a  $x$  é um modelo linear:

$$\mu_i = E(Y_i) = \pi_i = \beta_0 + \beta^T x$$

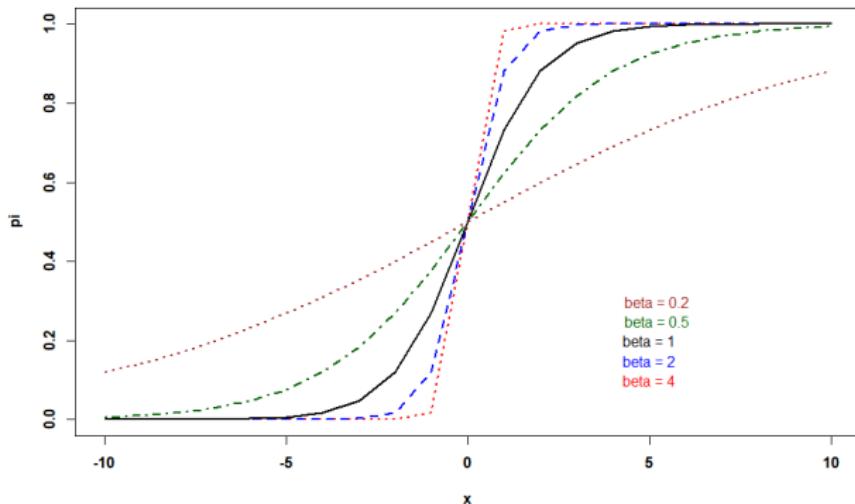
Mas,

- Para certos valores de  $x$ ,  $\pi_i > 1$  ou  $\pi_i < 0$ .  
Impossível, pois  $\pi_i$  é uma probabilidade!
- Solução de Mínimos Quadrados não é ótima porque  $Var(Y) = \pi_i(1 - \pi_i)$ .
- Estimadores de Máxima Verossimilhança não tem forma fechada  
(curso de estatística computacional).



## Exemplos de Curvas Logísticas

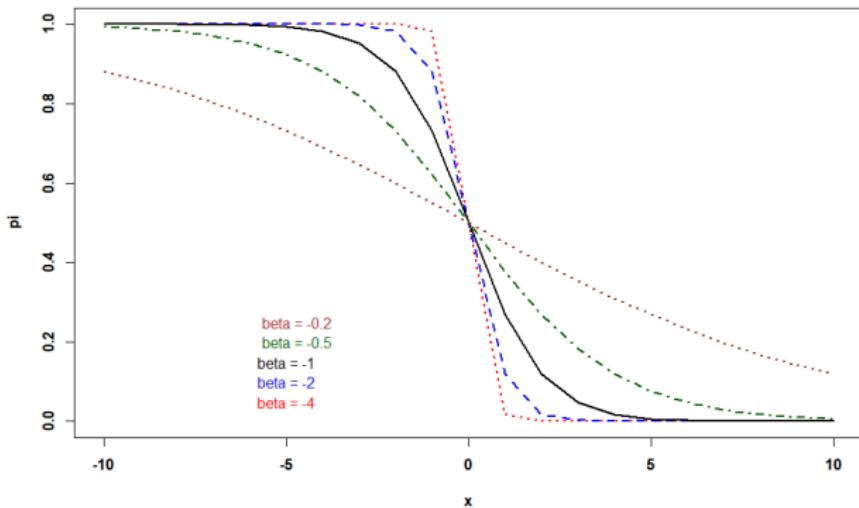
$\beta_0 = 0$  e  $\beta$  positivo,  $x = x$  (regressão logística simples) e  $\pi_i = \pi_i$ .



- $\beta_0$  indica o valor de  $x$  para o qual a  $\pi_i > 0.5$
- O valor de  $|\beta|$  indica o grau de separação entre  $\omega_1$  e  $\omega_2$ .

## Exemplos de Curvas Logísticas

$\beta_0 = 0$  e  $\beta$  negativo,  $x = x$  (regressão logística simples) e  $\pi_i = \pi_i$ .



- A maior inclinação ocorre em  $\pi_i = 0.5$  ou  $x = -\beta_0/\beta$ .
- Em biologia, este ponto é denominado de nível efetivo mediano e se caracteriza pela indecisão quando o ponto de corte é 0.5.



## Inferência - Estimação Pontual

- **Dados:**

$n$  observações independentes,  $(Y_i, x_i)$ ,  $i = 1, \dots, n$ .

$x$  corresponde a uma única variável explicativa.

- **Suposições:**

$Y_i \sim \text{Binomial}(1, \pi_i) = \text{Bernoulli}(\pi_i)$ , independentes.

- **Resposta binária:**

$Y_i = 1$  se  $x \in \omega_1$  e  $Y_i = 0$  se  $x \in \omega_2$ .

- **Modelo:**

$$\text{logito } [\pi_i] = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta x_i$$

ou

$$\pi_i = P(Y_i = 1 | X = x_i) = \frac{\exp\{\beta_0 + \beta x_i\}}{1 + \exp\{\beta_0 + \beta x_i\}}$$



- Função de verossimilhança para  $(\alpha, \beta)$ :

$$\begin{aligned}
 \ell(\beta_0, \beta) &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\
 &= \prod_{i=1}^n \left[ \frac{\exp\{\beta_0 + \beta x_i\}}{1 + \exp\{\beta_0 + \beta x_i\}} \right]^{y_i} \left[ \frac{1}{1 + \exp\{\beta_0 + \beta x_i\}} \right]^{1-y_i} \\
 &\stackrel{*}{=} \prod_{i=1}^n \frac{\exp\{[\beta_0 + \beta x_i] Y_i\}}{1 + \exp\{\beta_0 + \beta x_i\}}
 \end{aligned}$$

(\*) Uma vez que  $y_i = 0$  ou  $1$ .

- Log-verossilhança para  $(\beta_0, \beta)$ :

$$\begin{aligned}
 \mathcal{L}(\beta_0, \beta) &= \sum_{i=1}^n [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] \\
 &= \beta_0 \sum_{i=1}^n Y_i + \beta \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n \log[1 + \exp(\beta_0 + \beta x_i)]
 \end{aligned}$$



- Funções escores (primeiras derivadas):

$$U_1(\beta_0, \beta) = \frac{\partial \mathcal{L}(\beta_0, \beta)}{\partial \beta} = \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \pi_i = \sum_{i=1}^n x_i (Y_i - \pi_i)$$

$$U_2(\beta_0, \beta) = \frac{\partial \mathcal{L}(\beta_0, \beta)}{\partial \beta_0} = \sum_{i=1}^n Y_i - \sum_{i=1}^n \pi_i = \sum_{i=1}^n (Y_i - \pi_i)$$

- Os estimadores de Máxima Verossimilhança (MLE) são obtidos pela solução das equações

$$U_1(\hat{\beta}_0, \hat{\beta}) = 0$$

$$U_2(\hat{\beta}_0, \hat{\beta}) = 0$$

- Que implicam em

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}x_i\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}x_i\}},$$

i.e., # de sucessos observados = # de sucessos preditos.



- A matriz Hessiana (segundas derivadas) será formada por:

$$\frac{\partial^2 \mathcal{L}(\beta_0, \beta)}{\partial \beta^2} = - \sum_{i=1}^n x_i^2 \pi_i (1 - \pi_i)$$

$$\frac{\partial^2 \mathcal{L}(\beta_0, \beta)}{\partial \beta_0^2} = - \sum_{i=1}^n \pi_i (1 - \pi_i)$$

$$\frac{\partial^2 \mathcal{L}(\beta_0, \beta)}{\partial \beta_0 \partial \beta} = - \sum_{i=1}^n x_i \pi_i (1 - \pi_i)$$

- Resultando na Matriz de Informação de Fisher,

$$\mathcal{I} = \begin{pmatrix} \sum_{i=1}^n \pi_i (1 - \pi_i) & \sum_{i=1}^n x_i \pi_i (1 - \pi_i) \\ \sum_{i=1}^n x_i \pi_i (1 - \pi_i) & \sum_{i=1}^n x_i^2 \pi_i (1 - \pi_i) \end{pmatrix}$$

**Observação:** Expressões semelhantes são obtidas quando existem  $p$  variáveis explicativas. Neste caso, os escores e a diagonal da matriz de informação para um dado  $\beta_j$  ( $j = 0, \dots, p$ ) irão envolver o respectivo  $x_j$ , com  $x_{j=0} = 1$ . Os elementos fora da diagonal irão envolver produtos de  $x_i$  e  $x_j$ .



## Exemplo: Challenger Shuttle O-Ring Failures

- Data prevista de lançamento: 22 de Janeiro de 1986.
- Data do lançamento/acidente: 28 de Janeiro de 1986  
Cabo Canaveral, Flórida.
- Filme sobre o acidente:  
<https://www.youtube.com/watch?v=2FehGJQl0f0>
- Cobertura CBS news:  
<https://www.youtube.com/watch?v=N9kWG-1AOCA>
- Informações Gerais:  
<http://pt.wikipedia.org/wiki/STS-51-L>
- Livros:
  - McDonald, A.J. (2012) "Truth, Lies, and O-Rings: Inside the Space Shuttle Challenger Disaster". University Press of Florida.
  - Vaughan, D. (2016) "The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA". University of Chicago Press.



## Regressão Logística Múltipla

Modelo de regressão logística com duas ou mais de variáveis explicativas.

Sejam,

- Resposta binária:  $Y$
- $k$  preditores:  $\mathbf{x} = (x_1, \dots, x_k)$
- Quantidade a estimar:  $\pi(\mathbf{x}) = P(Y = 1 | x_1, \dots, x_k)$

O modelo será

$$\text{logito}(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

- O parâmetro  $\beta_j$  reflete o efeito do acréscimo de uma unidade em  $x_j$  no *log-odds* de  $Y = 1$ , mantendo os outros preditores constantes.
- $e^{\beta_j}$  é o efeito multiplicativo na chance (odds) de  $Y = 1$  para o acréscimo de uma unidade em  $x_j$ , mantendo os outros preditores constantes.



## Inferência

Suponha que temos  $n$  observações independentes,  $(x_{i1}, \dots, x_{ik}, Y_i)$ ,  $i = 1, \dots, n$  e desejamos estimar  $\beta_0, \beta_1, \dots, \beta_k$ .

Fazendo os escores iguais a zero temos as equações de estimação,

$$U_0(\beta_0, \beta) = \frac{\partial \ell(\beta_0, \beta)}{\partial \beta_0} = \sum_{i=1}^n Y_i - \sum_{i=1}^n \pi_i = 0$$

$$U_j(\beta_0, \beta) = \frac{\partial \ell(\beta_0, \beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} Y_i - \sum_{i=1}^n x_{ij} \pi_i = 0$$

para  $j = 1, \dots, k$ .

Estas equações são resolvidas numericamente.



A matriz de informação de Fisher resultante é

$$\begin{pmatrix} \sum_{i=1}^n \pi_i(1 - \pi_i) & \sum_{i=1}^n x_{i1}\pi_i(1 - \pi_i) & \dots & \sum_{i=1}^n x_{ik}\pi_i(1 - \pi_i) \\ \sum_{i=1}^n x_{i1}\pi_i(1 - \pi_i) & \sum_{i=1}^n x_{i1}^2\pi_i(1 - \pi_i) & \dots & \sum_{i=1}^n x_{i1}x_{ik}\pi_i(1 - \pi_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik}\pi_i(1 - \pi_i) & \sum_{i=1}^n x_{i1}x_{ik}\pi_i(1 - \pi_i) & \dots & \sum_{i=1}^n x_{ik}^2\pi_i(1 - \pi_i) \end{pmatrix}$$

A matriz de variância-covariância de  $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$  é o inverso da matriz de informação. Os valores dos estimadores  $\widehat{\text{Var}}(\hat{\beta}_j)$  ( $j = 1, \dots, k$ ) são os elementos da diagonal desta matriz. O erro padrão assintótico de  $\hat{\beta}_j$  é obtido por

$$\text{ASE}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}.$$



## Teste do Modelo

Hipóteses:

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

$$H_1 : \text{Pelo menos um } \beta_j \neq 0$$

Utilizamos a estatística da razão de verossimilhança,

$$G^2 = Q_L = -2 \log \left[ \frac{\ell(\hat{\beta}_0)}{\ell(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)} \right]$$

Sob  $H_0$ ,

$$G^2 \xrightarrow{d} \chi_k^2 \quad \text{para} \quad n \longrightarrow \infty$$

Rejeitamos  $H_0$  para valores elevados de  $Q_L$ .



## Teste para componentes do Modelo

Considere que o vetor de  $k$  coeficientes é dividido em dois grupos de  $q$  e  $r$  coeficientes. Seja  $\beta_{r|q}$  the partial contribution of  $\beta_r$  ao modelo reduzido que contém somente o subgrupo  $\beta_q$ .

Hipóteses:

$$H_0 : \beta_{r|q} = \mathbf{0} \Rightarrow \mathcal{L} = \mathcal{L}(\beta_0, \beta_q)$$

$$H_1 : \beta_{r|q} \neq \mathbf{0} \Rightarrow \mathcal{L} = \mathcal{L}(\beta_0, \beta_{q|r}, \beta_{r|q}) = \mathcal{L}(\beta_0, \beta)$$

Utilizamos a estatística da razão de verossimilhança,

$$G^2 = Q_L = -2 \log \left[ \frac{\ell(\hat{\beta}_0, \hat{\beta}_q)}{\ell(\hat{\beta}_0, \hat{\beta})} \right]$$

Sob  $H_0$ ,

$$G^2 \xrightarrow{d} \chi_r^2 \quad \text{para} \quad n \longrightarrow \infty$$

Rejeitamos  $H_0$  para valores elevados de  $Q_L$ .



## Teste para Coeficientes Individuais

Hipóteses:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

A estatística do teste (de Wald) é

$$Z = \frac{\hat{\beta}_j}{ASE(\hat{\beta}_j)} \sim N(0, 1)$$

Rejeitamos  $H_0$  para valores elevados de  $|Z|$ .

## Intervalos de Confiança para os parâmetros

Intervalos de confiança são formados da mesma maneira que no modelo com uma variável explicativa. O intervalo de  $100(1 - \alpha)\%$  de confiança para  $\beta_j$  é dado por

$$\hat{\beta}_j \pm Z_{\alpha/2} ASE(\hat{\beta}_j)$$



## Intervalos de Confiança para o logito

O logito estimado é dado por

$$\hat{g}(\mathbf{x}) = \text{logito}(\hat{\pi}(\mathbf{x})) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k.$$

com variância assintótica igual a

$$\widehat{\text{Var}}(\hat{g}(\mathbf{x})) = \sum_{j=0}^k x_j^2 \widehat{\text{Var}}(\hat{\beta}_j) + \sum_{j=0}^k \sum_{\ell=j+1}^k 2x_j x_\ell \widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_\ell)$$

sendo  $x_0 = 1$ .

O intervalo de  $100(1 - \alpha)\%$  de confiança para  $\text{logito}(\pi(\mathbf{x}))$  será

$$\hat{g}(\mathbf{x}) \pm Z_{\alpha/2} \text{ASE}(\hat{g}(\mathbf{x})) = (\hat{g}_L(\mathbf{x}); \hat{g}_U(\mathbf{x}))$$

onde  $\text{ASE}(\hat{g}(\mathbf{x})) = \sqrt{\widehat{\text{Var}}(\hat{g}(\mathbf{x}))}$ .



Uma vez que

$$\pi(x) = \frac{\exp[g(x)]}{1 + \exp[g(x)]}$$

podemos encontrar o intervalo de  $100(1 - \alpha)\%$  de confiança para  $\pi(x)$  substituindo os pontos extremos do intervalo para o logito, i.e,

$$\left( \frac{\exp[\hat{g}_L(x)]}{1 + \exp[\hat{g}_L(x)]}; \frac{\exp[\hat{g}_S(x)]}{1 + \exp[\hat{g}_S(x)]} \right)$$



## Seleção de Variáveis



**Modelo Parcimonioso** é aquele que explica bem o comportamento da variável resposta envolvendo o mínimo de parâmetros possíveis na estimação.

Um modelo com muitas variáveis não é necessariamente um modelo que gera um bom discriminante. Algumas sugestões que podem ajudar neste processo:

- ① Inicie o processo com uma análise cuidadosa de cada potencial preditor.
  - Para variáveis nominais e ordinais examine a tabela bidimensional da variável resposta e os k níveis da variável explicativa. Preste atenção em células com zeros (sem observações). Neste caso é preciso unir as categorias ou eliminá-las completamente.
  - Para variáveis contínuas examine o modelo de regressão logística.
  - Um modelo não deve conter muitas variáveis se o número de respostas de uma categoria for pequena. Peduzzi et al.<sup>2</sup> sugerem  $p$  (variáveis explicativas) não superior ao número de obs/10 na categoria de menor ocorrência, i.e., 40 sucessos e 100 falhas  $\Rightarrow$  modelo com no máximo 4 variáveis explicativas.
- ② Selecione as variáveis para a análise múltipla.
  - Hosmer e Lemeshow sugerem a inclusão de qualquer variável com  $p$ -valor  $< 0.25$  no modelo univariado. Esta abordagem pode resultar na inclusão de algumas variáveis questionáveis ou ignorar a possibilidade de variáveis em grupo serem mais importantes que individualmente.

<sup>2</sup>Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. (1996) *A simulation study of the number of events per variable in logistic regression analysis*. Journal of Clinical Epidemiology.



- ③ Utilize procedimentos baseados numa regra de decisão (estatística) para seleção de variáveis preditoras do modelo (*Stepwise/Forward/Backward*).

- Pode incluir uma variável por vez (*forward selection*) no modelo mais simples ou eliminar uma variável por vez (*backward selection*) do modelo mais complexo. Também é possível fazer adição ou remoção atendendo a regra de decisão em cada etapa (*stepwise selection*).
- No modelo logístico utilizamos o teste da razão de verossimilhança como critério de inclusão ou exclusão.
- A variável mais importante é aquela que produz a maior alteração na estatística da razão de verossimilhança.
- Os  $p$ -valores calculados no procedimento *Stepwise* não correspondem aos  $p$ -valores dos testes de hipóteses usuais. Estes resultados devem ser utilizados como indicadores da importância relativa entre variáveis.



## Comentários

- É melhor “errar” incluindo muitas variáveis e depois utilizar outra metodologia para obter um modelo mais parcimonioso.
- Uma possibilidade é começar com um modelo contendo variáveis consideradas importantes e depois utilizar o método stepwise para incluir novas variáveis.
- Quando existem muitos preditores pode ocorrer que algumas variáveis não importantes (“ruído”) sejam estatisticamente significantes.
- Uma ação útil é examinar se cada variável do modelo final tem efeito linear.
- Podemos utilizar o modelo final como um modelo inicial para procurar por interações importantes.

### Exemplo: Acceptance Analysis

Teixeira, R.M.; de Oliveira, L. de L.; Loureiro, L.M.R.; von Borries, G.; da Costa, T.H.M. *Sensory Analysis of Post-Exercise Coffee or Cocoa Milk Beverages for Endurance Athletes*. Beverages, 2020, 6(4), 61.

Nota: Este não é um estudo de classificação.



## Algoritmos para Seleção de Variáveis

**Seleção Backward:** inicia com um modelo complexo e remove as variáveis explicativas com maior p-valor, no teste  $H_0 : \beta_i = 0$ , sequencialmente. O processo para quando a retirada de uma variável produz um ajuste inadequado, de acordo com algum critério indicado pelo analista.

**Seleção Forward:** Segue o processo inverso. Inicia com o modelo sem variáveis explicativas (modelo nulo) e adiciona variáveis ao modelo até que a adição de uma nova variável não melhore o modelo (critério indicado pelo analista).

**Seleção Stepwise:** adiciona ou elimina variáveis em cada etapa, utilizando os critérios anteriores que resultam no melhor ajuste do modelo.

O R tem a função `stepAIC` no pacote MASS que permite aplicar os algoritmos seleção através do Critério de Informação de Akaike (AIC). O melhor modelo será aquele que minimiza AIC (ver adiante).



(Segundo parágrafo da página 126 de Agresti (2019), sobre cuidados no uso dos algoritmos de seleção.)

*Variable selection methods need not yield a meaningful model. You should regard its results with skepticism. When you evaluate many terms, one or two that are not truly important may look impressive merely due to chance. In any case, true statistical significance is not simple to judge for effects highlighted as being the most or the least significant, and it should not be the sole criterion for whether to include a term in a model. It is sensible to include a variable that is important for the purposes of the study and report its estimated effect even if it is not statistically significant. If the variable is a potential confounder, including it in the model may help to reduce bias in estimating relevant effects of key explanatory variables and may make it possible to compare results with other studies where the effect is significant, perhaps because of a larger sample size. On the other hand, with a very large  $n$ , sometimes a term might be statistically significant but not practically significant. You might then exclude it from the model because the simpler model is easier to interpret - for example, when the term is a complex interaction.*



## Verossimilhança e Medidas Relacionadas:

Em modelos com variáveis explicativas contidas num modelo menor, podemos utilizar o teste da razão de verossimilhança para escolher entre os modelos.

### 1 - 2 Loglikelihood

$$-2 \log(\mathcal{L}(\hat{\beta}_0, \hat{\beta})) = -2 \sum_{i=1}^n [y_i \log(\hat{\pi}_i) + (1 - y_i) \log(1 - \hat{\pi}_i)]$$

Modelos com mais variáveis tendem a ter maiores valores para este critério. Devemos considerar medidas que penalizem o acréscimo de parâmetros no modelo.

### 2 Critério de Informação de Akaike

$$\text{AIC} = -2 \log \mathcal{L} + 2(k + s)$$

sendo  $k$  é o número de níveis de resposta menos 1 e  $s$  é o número de variáveis explicativas. No modelo com resposta binária  $k + s$  é o número de parâmetros no modelo.

### 3 Critério de Schwarz (ou Bayesiano de Informação)

$$\text{BIC} = -2 \log \mathcal{L} + (k + s) \log(n)$$

sendo  $n$  é número de observações.



## Comentários:

- AIC e BIC penalizam o log likelihood pelo número de parâmetros no modelo.
- Menores valores de AIC e BIC indicam os melhores modelos.
- BIC também ajusta pelo tamanho do amostra.
- BIC tende a escolher modelos menores que AIC quando a amostra cresce.



## Passos para selecionar um modelo:<sup>3</sup>

- ① Modelo inicial com variáveis importantes e que mostram qualquer indício de relevância ( $p < 0.2$ ) quando consideradas isoladamente.
- ② Fazer seleção *backward* e manter as variáveis significantes para um nível mais rigoroso e as que influenciam estas variáveis se forem removidas do modelo.
- ③ Incluir novas variáveis que mostrarem contribuição importante na presença das variáveis retidas após o passo 2.
- ④ Verificar interações pertinentes entre as variáveis selecionadas utilizando  $\alpha = 0.05$ .
- ⑤ Realizar testes de diagnóstico.

---

<sup>3</sup>Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. (2013) *Applied Logistic Regression, 3<sup>rd</sup> Edition*, Wiley.



## Exemplo: Female Horseshoe Crabs and Their Satellites (Agresti, 2019)

*“Caranguejos-ferradura” tem um corpo dividido em três partes: cabeça, abdômen e cauda. O nome “horseshoe crab” se deve ao formato arredondado da cabeça que é semelhante a ferradura de um cavalo. (Fonte: nwf.org)*



Fotos:

The National Wildlife Federation (NWF) e

Danielle Brigida, USFWS at Cape May National Wildlife Refuge.



## Exemplo: Female Horseshoe Crabs and Their Satellites (Agresti, 2019)

O problema avalia se espécies do sexo feminino tem caranguejos macho no ninho (satélites). Os dados envolvem as seguintes variáveis:

Arquivo Crabs.dat:

- crab: caranguejo analisado.
- sat: número de satélites no ninho.
- $Y = 1$  se o ninho tem satélites e 0 caso contrário.
- weight: peso em gramas.
- width: comprimento (cm) da carapaça.
- color: cor  
1: clara, 2: pouco clara, 3: pouco escura, 4: escura.
- spine: condição da espinha  
1: ambas ok, 2: para uma desgastada ou quebrada,  
3: para ambas desgastadas ou quebradas.



## Exemplo: Female Horseshoe Crabs and Their Satellites (Agresti, 2019)

Um estudo procura verificar se as variáveis comprimento da carapaça e peso da espécie do sexo feminino são bons discriminantes para presença satélites no ninho.

O estudo utiliza as seguintes variáveis:

- crab: caranguejo analisado.
- y: 1 se o ninho tem satélites e 0 caso contrário.
- weight: peso em quilogramas.
- width: comprimento (cm) da carapaça.

São ajustados os seguintes modelos:

- **M0**: modelo  $y \sim 1$ , somente com intercepto.
- **M1**: modelo  $y \sim \text{width}$ .
- **M2**: modelo  $y \sim \text{weight}$ .
- **M3**: modelo  $y \sim \text{width} + \text{weight}$ .



## Exemplo: Female Horseshoe Crabs and Their Satellites - Análise 1

**Observação:** Ver programa R no site do curso.

- **Significância:** apesar de M1 e M2 apresentarem parâmetros significativamente diferentes de zero, M3 indica que os mesmos parâmetros são não significativos quando juntos no modelo.
- **Verossimilhança:** não produz resultados claros, uma vez que o modelos M1, M2 e M3 produzem resultados semelhantes. Existe uma pequena indicação de que os modelos M1 e M2 são melhores que o modelo M3.
- **Seleção Stepwise:** indica que o M1 é o melhor modelo, isto é, o que utiliza somente a variável comprimento como variável explicativa.



## Diagnóstico e Comparação de Modelos



## Teste de Qualidade do Ajuste para Preditores Contínuos (*Goodness-of-Fit*)

- Um modelo de regressão logística fornece as probabilidades estimadas para cada classe, considerando um conjunto de especificações das variáveis exploratórias.
- O número estimado de observações em cada classe é obtido pela multiplicação da probabilidade estimada e o número de sujeitos naquele cenário.
- Quando as variáveis são contínuas existem poucas observações com valores comuns, invalidando o uso de estatísticas como  $X^2$  de Pearson ou razão de verossimilhanças  $G^2$  que são baseadas na comparação dos valores observados e esperados para cada nível das variáveis explicativas.
- Uma solução é agrupar as observações de certo intervalo e calcular  $X^2$  e  $G^2$ .
- Uma abordagem mais simples é ajustar a regressão logística diretamente as contagens observadas.



- O teste de **Hosmer and Lemeshow** ajusta as observações em decimais baseados nas probabilidades preditas pelo modelo (SAS por exemplo).
  - As  $n/10$  observações com maiores probabilidades estimadas são colocadas na primeira categoria, e assim por diante.
  - Para cada grupo, os valores ajustados para uma resposta são a soma das probabilidades estimadas naquele grupo.
  - A estatística  $X^2$  de Pearson<sup>4</sup> é calculada e rejeitamos o modelo ajustado se  $X^2 > \chi^2_{gl,\alpha}$ , onde  $gl = g - 2$  e  $g$  é o número de grupos.  
 $g$  deve ser especificado no R.
- Um valor elevado da estatística para qualidade do ajuste indica que existe algum *lack of fit*, mas o teste não indica a natureza do problema.
- Uma maneira mais informativa para teste de *lack of fit* é ajustar um modelo mais complexo (contendo por exemplo um termo quadrático) e utilizar o teste da razão de verossimilhança para verificar se as covariáveis adicionais são úteis. Podemos utilizar também o teste de Wald ou Escore e rejeitar a hipótese nula para valores elevados da estatística utilizando valores críticos de  $\chi^2$  com  $gl$  igual a diferença entre o número de parâmetros nos dois modelos.

---

<sup>4</sup>Ver slide seguinte.



## Estatística de Pearson

Suponha que existam  $s$  conjuntos (cenários) de variáveis explicativas. Defina:

- $n_i$  = número de eventos no  $i$ -ésimo conjunto.
- $y_i$  = número de sucessos no  $i$ -ésimo conjunto.
- $\hat{\pi}_i$  = probabilidade estimada de sucessos no  $i$ -ésimo conjunto.
- $\hat{y}_i = n_i \hat{\pi}_i$  = número estimado de sucessos no  $i$ -ésimo conjunto.
- Resíduo de Pearson definido como

$$e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

O teste de Pearson para qualidade do ajuste (*goodness of fit*) é

$$X^2 = \sum_{i=1}^s e_i^2 = \sum_{i=1}^s \frac{(y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}$$

Valores elevados de  $X^2$  irão levar a conclusão que o modelo proposto não ajusta bem aos dados.



## Teste de Qualidade do Ajuste e Teste LR para comparação de modelos

- O **Modelo Nulo** é aquele que representa a média das observações, i.e., possui somente o intercepto (ou viés)  $\beta_0$ .
- O **Modelo Saturado** é o modelo mais completo e fornece um ajuste perfeito para a amostra de logitos.
- Lembre que a estatística Deviance  $G^2$  é utilizada para testar a qualidade do ajuste para o modelo  $M$  de regressão logística.



- Sejam  $\mathcal{L}_M$  e  $\mathcal{L}_S$  as verossimilhanças do modelo M e do modelo saturado S, respectivamente. Então,

$$G^2 = \text{Deviance} = -2[\log(\mathcal{L}_M) - \log(\mathcal{L}_S)]$$

Suponha que o modelo  $M_k$  é um caso especial do modelo  $M_1$ . Estes modelos são sub-modelos ou modelos aninhados (*nested*). Dado que  $M_1$  é adequado, a estatística LR (razão de verossimilhança) para testar se o modelo mais simples é adequado será

$$\begin{aligned} Q_L &= -2[\log(\mathcal{L}_{M_k}) - \log(\mathcal{L}_{M_1})] \\ &= -2[\log(\mathcal{L}_{M_k}) - \log(\mathcal{L}_S)] - 2[\log(\mathcal{L}_{M_1}) - \log(\mathcal{L}_S)] \\ &= \text{Deviance}_k - \text{Deviance}_1 \end{aligned}$$

Assim, podemos comparar modelos através da comparação das deviances.

Para grandes amostras esta estatística é aproximadamente  $\chi^2$  com  $gl$  igual a diferença de  $gl$  dos resíduos nos dois modelos.



## Exemplo: Female Horseshoe Crabs and Their Satellites - Análise 2

- O modelo M1, somente com preditor comprimento (width), não indica falta de adequabilidade do modelo. Não rejeita a hipótese de adequabilidade do modelo.

```
ResourceSelection :: hoslem.test(M1$y, M1$fitted.values, g=10)
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test
```

```
data: M1$y, M1$fitted.values
X-squared = 4.3855, df = 8, p-value = 0.8208
```

- O modelo M2, somente com peso, também não indica falta de adequabilidade do modelo. A hipótese de adequabilidade do modelo não é rejeitada.

```
> hoslem.test(M2$y, M2$fitted.values, g=10)
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test
```

```
data: M2$y, M2$fitted.values
X-squared = 7.8334, df = 8, p-value = 0.4499
```



## Exemplo: Female Horseshoe Crabs and Their Satellites - Análise 2

- O modelo M3, como comprimento e peso, também não indica falta de adequabilidade do modelo. A hipótese de adequabilidade do modelo não é rejeitada.

```
> hoslem.test(M3$y, M3$fitted.values, g=10)
Hosmer and Lemeshow goodness of fit (GOF) test

data: M3$y, M3$fitted.values
X-squared = 0.91631, df = 8, p-value = 0.9987
```

- O modelo M3 é melhor que o modelo M0 (somente intercepto)?

```
> lmtest::lrtest(M0,M3)
Likelihood ratio test

Model 1: y ~ 1
Model 2: y ~ width + weight
#Df  LogLik Df  Chisq Pr(>Chisq)
1    1 -112.879
2    3  -96.446  2 32.867  7.296e-08 ***
```

O teste razão de verossimilhança rejeita a hipótese de  $\beta = 0$ .  
Logo, existe evidência (ao nível de significância  $\alpha = 0.10$ ) de que o modelo M3 é melhor que o modelo M0.



## Exemplo: Female Horseshoe Crabs and Their Satellites - Análise 2

- O modelo M3 é melhor que o modelo M1 ( $\alpha = 0.05$ )?

H0: Modelo M3 não é mais relevante que M1.

H1: Modelo M3 é mais relevante que M1.

Neste caso podemos utilizar Deviance<sub>1</sub> – Deviance<sub>3</sub>, resultando em  $P(\chi^2_1 > 1.560777) = 0.2116$  e que não existe evidência da relevância da inclusão do efeito peso.

```
> -2*(logLik(M1) - logLik(M3))
'log Lik.' 1.560777

> 1-pchisq(1.560777,1)
[1] 0.2115515                      # ou

> lrtest(M1,M3)
Likelihood ratio test

Model 1: y ~ width
Model 2: y ~ width + weight
#Df LogLik Df Chisq Pr(>Chisq)
1    2 -97.226
2    3 -96.446  1  1.5608      0.2116
```

É interessante verificar que os modelos com uma variável são significativos, enquanto que o modelo com as duas variáveis se torna desnecessário.



## Medidas (Diagnóstico) de Influência

Analisam a influência que as observações tem na estimativa dos parâmetros e/ou da resposta média. Observações influentes são aquelas que alteram as estimativas quando retiradas do estudo.

- Seja a matriz  $\mathbf{H}$  (*hat*) aquela que produz os valores logito preditos quando aplicada aos logitos amostrais.
- Seja a influência  $h_i$  o  $i$ -ésimo elemento da diagonal da matriz  $\mathbf{H}$ .
  - Valores elevados de  $h_i$  indicam grande potencial de influência da  $i$ -ésima observação.
- Medidas comumente usadas para verificar influência (*leverage*)
  1. **Dfbeta** é a alteração (padronizada) na estimativa do parâmetro quando a  $i$ -ésima observação é eliminada.
  2. A distância de Cook indica o quanto os valores da regressão se alteram quando a  $i$ -ésima observação é eliminada.
  3. Alteração em  $X^2$  e  $G^2$  quando uma observação é eliminada.

Para cada medida em 1, 2 e 3, valores elevados indicam grande influência.



## Exemplo: Female Horseshoe Crabs and Their Satellites - Análise 3

```
> stats::influence.measures(M1)

Influence measures of
  glm(formula = y ~ width, family = binomial(link = "logit"),
       data = crabs) :

      dfb.1_   dfb.wdth    dffit cov.r   cook.d   hat.inf
1 -0.04761  0.049966  0.0604  1.021  1.13e-03  0.01234
2 -0.10405  0.100800 -0.1137  1.033  4.23e-03  0.02572
3 -0.00454  0.009480  0.0750  1.010  2.02e-03  0.00708
4 -0.06162  0.055360 -0.1112  1.008  5.03e-03  0.01006
5 -0.00454  0.009480  0.0750  1.010  2.02e-03  0.00708
6 -0.09487  0.089934 -0.1188  1.019  5.12e-03  0.01660
7  0.04128 -0.049136 -0.1269  0.995  8.74e-03  0.00754
8 -0.06603  0.059893 -0.1120  1.009  5.04e-03  0.01058
9 -0.09686  0.092061 -0.1191  1.020  5.10e-03  0.01736
10 -0.01871  0.011553 -0.1091  1.000  5.47e-03  0.00739
11 -0.08128  0.075671 -0.1155  1.013  5.11e-03  0.01302
12 -0.00618 -0.001162 -0.1109  0.999  5.85e-03  0.00715
13 -0.04748  0.049931  0.0612  1.021  1.17e-03  0.01209
14 -0.08452  0.082739 -0.0879  1.040  2.36e-03  0.03001  *

...
...
```



## Problemas Numéricos na Estimação



## 1 Tabelas de Contingência com Zeros

### 2 Separação Completa

- Quando um grupo de variáveis separa completamente os grupos de respostas.
- Neste caso todas as observações tem probabilidade um de serem alocadas a resposta correta.
- Neste caso  $\hat{\beta} = \infty$ , apesar do R reportar um valor finito.

### 3 Separação Quase Completa

- Situação semelhante a anterior.
- Separação completa e quase completa geralmente ocorrem em dados com poucas observações ou muitas categorias.

### 4 Colinearidade nas variáveis explicativas.

- Forte correlações entre as variáveis explicativas podem causar problemas com as estimativas dos parâmetros e dos erros padrão na regressão logística.
- A informação de um preditor é a mesma informação de outro preditor.
- Geralmente as estimativas dos erros padrão serão elevadas e os coeficientes associados não serão significativos.
- O problema é eliminado através da retirada de uma das variáveis correlacionadas.



## Modelos para Várias Classes (Respostas Politômicas)



**Regressão Logística (Ordinária)** - utilizada para classificar em duas classes (resposta binárias).

**Regressão Logística Politômica (Multicategórica)** - utilizada para classificação em  $c$  classes (respostas são nominais ou ordinais).

- ✓ Em cada combinação dos níveis das variáveis explicativas, o modelo assume que as contagens para as categorias de  $Y$  tem distribuição multinomial (modelo logito multinomial).
- ✓ Conhecido também como modelo de escolha discreta (*discrete choice model*) na literatura de negócios e econometria.



## Exemplo: Prediction models of rice cooking quality.

von Borries et al. (2017)<sup>a</sup> avaliam textura de arroz cozido através da relação com medidas de textura instrumental. A variável resposta consta de 4 categorias de viscosidade: muito pegajoso, pegajoso, levemente solto e solto. Dois problemas são tratados no artigo:

1. Apresentação gráfica de modelo com grande número de variáveis explicativas.
2. Resposta com quatro categorias em resposta ordinal - Regressão logística politômica.

---

<sup>a</sup> von Borries, G.; Bassinello, P.Z.; Rios, E.S.; Koakuzu, S.N. e Carvalho, R.N. (2017) Prediction models of rice cooking quality. *Cereal Chemistry*.



## Modelo com Categorias Não Ordenadas

- $Y$  é uma variável nominal com  $c$  classes (ou categorias).
- A ordem das classes é irrelevante.
- Sejam  $\{\pi_1, \dots, \pi_c\}$  as probabilidades das respostas para as  $c$  classes. Então  $\sum_{j=1}^c \pi_j = 1$ , em que  $\pi_j = P(Y_{ij} = 1 | \mathbf{X} = \mathbf{x}_i)$  e  $Y_{ij} = 1$  se  $\mathbf{x} \in \omega_j$ .
- Se temos  $n$  observações independentes baseadas em  $\{\pi_1, \dots, \pi_c\}$ , então as contagens nas categorias tem distribuição multinomial.
- Precisamos generalizar o problema logístico ordinário para tratar com  $c$  categorias.
- Uma vez que o modelo formou logitos para  $c - 1$  pares de categorias, qualquer outro logito é redundante.



- Para dados dicotômicos ( $c = 2$ ), o modelo logit é:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta X_i$$

- Estendendo para  $c = 3$  categorias,

$$\log\left(\frac{\pi_{i,1}}{1 - \pi_{i,1}}\right) = \beta_{01} + \beta_1 X_i, \quad \log\left(\frac{\pi_{i,2}}{1 - \pi_{i,2}}\right) = \beta_{02} + \beta_2 X_i$$

e

$$\log\left(\frac{\pi_{i,3}}{1 - \pi_{i,3}}\right) = \beta_{03} + \beta_3 X_i$$

Mas o modelo não é tratável. As probabilidades devem somar 1.



## Categorias Logito de Referência

- Vamos considerar então modelos para cada par de categorias:

$$\log \left( \frac{\pi_{i,1}}{\pi_{i,3}} \right) = \beta_{01} + \beta_1 x_i, \quad \log \left( \frac{\pi_{i,2}}{\pi_{i,3}} \right) = \beta_{02} + \beta_2 x_i$$

e

$$\log \left( \frac{\pi_{i,1}}{\pi_{i,2}} \right) = \beta_{03} + \beta_3 x_i$$

Estas equações são mutuamente consistentes e uma é redundante.  
Podemos obter a terceira equação das outras duas:

$$\begin{aligned}\log \left( \frac{\pi_{i,1}}{\pi_{i,2}} \right) &= \log \left( \frac{\pi_{i,1}/\pi_{i,3}}{\pi_{i,2}/\pi_{i,3}} \right) = \log \left( \frac{\pi_{i,1}}{\pi_{i,3}} \right) - \log \left( \frac{\pi_{i,2}}{\pi_{i,3}} \right) \\ &= (\beta_{01} + \beta_1 x_i) - (\beta_{02} + \beta_2 x_i) \\ &= (\beta_{01} - \beta_{02}) + (\beta_1 - \beta_2) x_i\end{aligned}$$

implicando em  $\beta_{03} = \beta_{01} - \beta_{02}$  e  $\beta_3 = \beta_1 - \beta_2$ .



- Para  $c$  categorias existem  $\binom{c}{2} = \frac{c(c-1)}{2}$  pares de categorias. Precisamos especificar somente  $c - 1$  pares e os demais serão redundantes.
- Considere  $c = 3$ . Escolhemos uma categoria de referência e obtemos

$$\log(\pi_{i,j}/\pi_{i,c}) = \beta_{0j} + \beta_j x_i \quad j = 1, 2 \text{ e } c = 3.$$

- As probabilidades para as três categorias, correspondendo as equações logito são<sup>5</sup>,

$$\pi_{i,1} = \frac{e^{\beta_{01} + \beta_1 x_i}}{1 + e^{\beta_{01} + \beta_1 x_i} + e^{\beta_{02} + \beta_2 x_i}} \quad \pi_{i,2} = \frac{e^{\beta_{02} + \beta_2 x_i}}{1 + e^{\beta_{01} + \beta_1 x_i} + e^{\beta_{02} + \beta_2 x_i}}$$

$$\pi_{i,3} = \frac{1}{1 + e^{\beta_{01} + \beta_1 x_i} + e^{\beta_{02} + \beta_2 x_i}}$$

$$\Rightarrow \pi_{i,1} + \pi_{i,2} + \pi_{i,3} = 1.$$

---

<sup>5</sup>O termo 1 em cada denominador e no numerador de  $\pi_{i,3}$  representa  $e^{\beta_{03} + \beta_3 x_i} = e^0$ , pois  $\beta_{03} = \beta_3 = 0$  para a categoria de referência.



- A análise **não** requer que a resposta seja ordenada de maneira específica. A escolha da categoria de referência é arbitrária, mas geralmente é feita de maneira a facilitar a interpretação dos dados.
- Programas estatísticos estimam  $c - 1$  equações logit simultaneamente utilizando o algoritmo iterativo *Fisher scoring*.
- No R podemos utilizar a função `vglm` da biblioteca VGAM.
- Na estimação simultânea, as mesmas estimativas são obtidas para cada par de categorias, não importando a categoria utilizada como referência.
- O tamanho da amostra precisa ser suficiente para modelar uma função resposta por subpopulação. Caso contrário, o pesquisador irá encontrar problemas de estimação e receberá um aviso de estimativas infinitas para os parâmetros. Redução do número de categorias na resposta pode ajudar a resolver o problema.

**Exemplo:** Idade de um carro em anos, para avaliação de IPVA.

$$\omega_1 = +30, \quad \omega_2 = (30, 20], \quad \omega_3 = (20, 10], \quad \omega_4 = (10, 1], \quad \omega_5 = 0 \text{ ou}$$

$$\omega_1 = +20, \quad \omega_2 = (20, 10], \quad \omega_3 = (10, 1], \quad \omega_4 = 0.$$

No primeiro caso,  $\omega_1$  e  $\omega_2$  indicam classes com isenção de IPVA.

No segundo caso, apenas  $\omega_1$  indica isenção de IPVA.

- Para  $c > 2$ , os coeficientes devem ser interpretados como efeitos nos contrastes entre pares de categorias, nunca como a probabilidade de pertencer a certa categoria.
- Na análise binária, se uma covariável  $X$  tem coeficiente positivo, interpretamos que  $X$  resulta num aumento da chance da resposta considerada. Isto **não** é obrigatoriamente verdade no modelo multinomial. Veja o exemplo seguinte.

### Exemplo: Alligator Food Choice (Agresti, 2019)

Agresti descreve um estudo sobre a alimentação de jacarés na Flórida. Os dados apresentam o tipo principal de alimento encontrado no estômago dos jacarés e o comprimento do animal, em metros. Três tipos de alimento são descritos: peixe ( $F = Fish$ ), invertebrados ( $I = Invertebrates$ ) e outros ( $O = Other$ ).

Os invertebrados eram, principalmente, bichos de maça, insetos aquáticos e pequenos crustáceos. A categoria outros inclui anfíbios, mamíferos, répteis, plantas, pedras e outros detritos. O comprimento dos jacarés variou entre 1.24 e 3.89 metros.



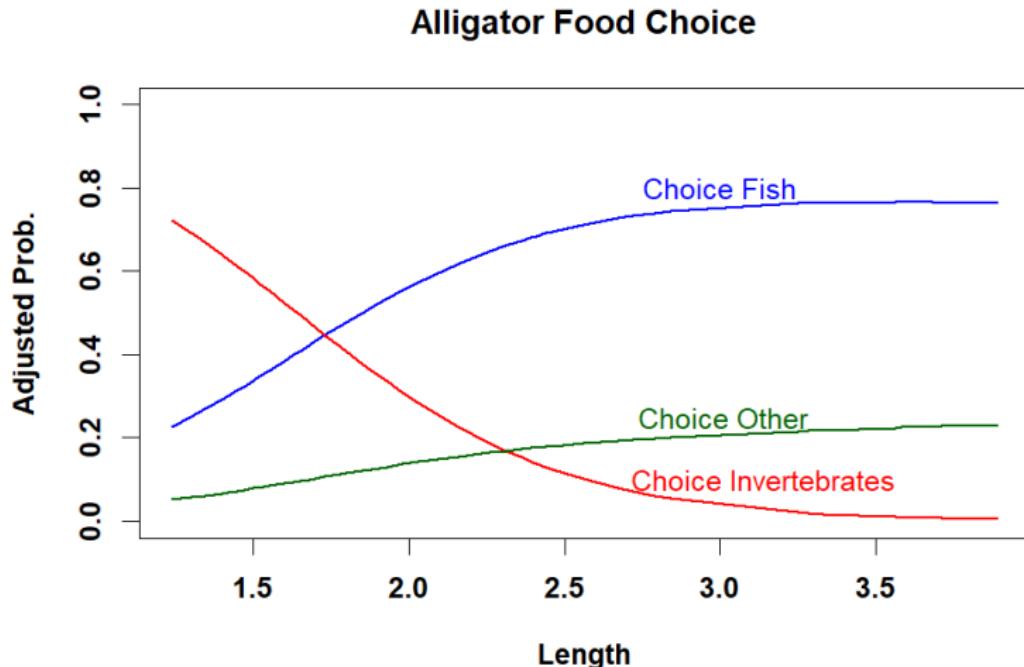
## Exemplo: Alligator Food Choice (Agresti, 2019)

- ✓ Para cada logito, a interpretação é feita como em modelos de regressão logística ordinários, condicional ao evento de que a resposta foi uma das duas categorias no logito. Exemplo: A estimativa do  $\log(odds)$  de que a resposta é peixe em vez de invertebrado é:  $\log(\hat{\pi}_1/\hat{\pi}_2) = -4.079713 + 2.355337$   
⇒ “Jacarés grandes tem maior chance de comer peixe em vez de invertebrados”.
- ✓ As estimativas simultânea e individual são similares, mas não são iguais. Estimativas individuais e simultâneas são assintoticamente não viesadas, mas as individuais tem maiores erros padrão.
- ✓ Nas estimativas individuais, os três conjuntos de estimativas tem de ser obtidos separadamente.
- ✓ O teste da razão de verossimilhança comparando o modelo nulo com o modelo ajustado permite testar  $H_0 : \beta_1 = \beta_2 = 0$ .



- Note no gráfico das probabilidades que

1.  $\pi_I(x) > \pi_F(x) > \pi_O(x)$  para jacarés menores.
2.  $\pi_F(x) > \pi_I(x) > \pi_O(x)$  para jacarés de médio porte.
3.  $\pi_F(x) > \pi_O(x) > \pi_I(x)$  para jacarés maiores.



Nota de Agresti, 2019, pág. 162.

*"As estimativas para uma equação são interpretadas como regressão logística binária, condicional ao evento em que o resultado pertence a uma das duas categorias. Por exemplo, dado que o tipo de alimento é peixe ou invertebrado, a equação de predição para  $\log(\hat{\pi}_1/\hat{\pi}_2)$  com coeficiente 2.355 para tamanho, indica que grandes jacarés são relativamente mais propensos a comer peixe em vez de invertebrados. A probabilidade condicional estimada para a comida de escolha peixe cresce no comprimento  $x$  de acordo com uma curva em formato S. Para jacarés de comprimento  $x+1$  metros, a chance estimada de que o alimento principal é peixe em vez de invertebrado é igual a  $\exp(2.355) = 10.5$  vezes a chance estimada em  $x$  metros.*

*A hipótese de que o alimento principal de escolha é independente do comprimento do jacaré é  $H_0 : \beta_1 = \beta_2 = 0$ .*

*A estatística do teste da razão de verossimilhança compara o modelo nulo com o modelo em análise através da diferença nas deviances. O p-valor de 0.0002 fornece forte evidência de que o comprimento afeta pelo menos um dos logitos."*



Nota de Agresti, 2019, pág. 163.

*"As probabilidades estimadas dos resultados (peixe, invertebrados, outros) são obtidas por,*

$$\begin{aligned}\hat{\pi}_1 &= \frac{\exp(1.6177 - 0.1101x)}{1 + \exp(1.6177 - 0.1101x) + \exp(5.6974 - 2.4654x)} \\ \hat{\pi}_2 &= \frac{\exp(5.6974 - 2.4654x)}{1 + \exp(1.6177 - 0.1101x) + \exp(5.6974 - 2.4654x)} \\ \hat{\pi}_3 &= \frac{1}{1 + \exp(1.6177 - 0.1101x) + \exp(5.6974 - 2.4654x)}\end{aligned}$$

O termo 1 em cada denominador e no numerador de  $\hat{\pi}_3$  representa  $\exp(\hat{\beta}_{03} + \hat{\beta}_3x)$  para  $\hat{\beta}_{03} = \hat{\beta}_3 = 0$ , quando a categoria 3 (outros) é a categoria de referência."

```
> fitted <- fitted(fit)
> fitted[ ,c(1,2,3)]
   F           I           O
1 0.2265307 0.721964000 0.05150528
2 0.2502564 0.692466814 0.05727683
3 0.2502564 0.692466814 0.05727683
4 0.2584591 0.682256244 0.05928463
...
57 0.7647495 0.007780500 0.22746996
58 0.7645799 0.007248095 0.22817198
59 0.7630060 0.004733748 0.23226027
```



# Modelo com Categorias Ordenadas

## Abordagens:

- ① Modelo de Logitos Cumulativos: mais utilizado e de mais fácil implementação computacional.
- ② Modelo Logito de Categorias Adjacentes: modelo atrativo, mas alguns aplicativos, como o SAS, só podem estimar o modelo quando os dados são agrupados.
- ③ Modelo Logito de Razão Sequencial: é mais especializado. Este modelo é destinado a situações em que as categorias ordinais representam uma avanço através de estágios.



- Quando a resposta tem apenas duas categorias, todos os modelos se reduzem ao modelo logístico usual para respostas binárias.
- Em alguns casos é necessário juntar categorias.
  - resultando em perda de informação e
  - dificultando a análise do objeto de estudo.
- Não seria incorreto ignorar a ordem e estudar o modelo logit generalizado. Entretanto, ao incorporar a ordem
  - temos interpretações mais simples e
  - testes, geralmente, com maior poder.
- Modelos com respostas ordinais impõem restrições que podem ser inapropriadas. Devemos checar se as restrições são válidas.



## 1 Modelo de Logitos Cumulativos

- Probabilidades cumulativas (pertencer as qualquer das classes  $\omega_1$  a  $\omega_j$ ):

$$P(Y \leq j) = \pi_1 + \dots + \pi_j, \quad j = 1, \dots, c.$$

- Estas probabilidades refletem a ordenação

$$P(Y \leq 1) \leq P(Y \leq 2) \leq \dots \leq P(Y \leq c) = 1$$

- Os logitos das primeiras  $c - 1$  probabilidades cumulativas são

$$\text{logito}[P(Y \leq j)] = \log \left( \frac{P(Y \leq j)}{1 - P(Y \leq j)} \right) = \log \left( \frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_c} \right)$$

- Cada logito cumulativo utiliza todas as  $c$  classes.
- Um modelo para o  $j$ -ésimo logito cumulativo pode ser reduzido em duas categorias. Uma categoria formada por  $\{1, \dots, j\}$  e outra categoria formada por  $\{j + 1, \dots, c\}$ . Assim,  $Y$  pode ser transformado em binário para cada  $j < c$ .



**Exemplo:** para  $c = 3$  e  $\{(X_i, Y_i), i = 1, \dots, n\}$ , é possível ajustar,

$$\log \left( \frac{\pi_{i,1}}{\pi_{i,2} + \pi_{i,3}} \right) = \beta_{01} + \beta_1 x_i$$

$$\log \left( \frac{\pi_{i,1} + \pi_{i,2}}{\pi_{i,3}} \right) = \beta_{02} + \beta_2 x_i$$

E as probabilidades baseadas nestes modelo serão

$$\pi_{i,1} = \frac{e^{\beta_{01} + \beta_1 x_i}}{1 + e^{\beta_{01} + \beta_1 x_i}}, \quad \pi_{i,2} = \frac{e^{\beta_{02} + \beta_2 x_i} - e^{\beta_{02} + \beta_1 x_i}}{(1 + e^{\beta_{01} + \beta_1 x_i})(1 + e^{\beta_{02} + \beta_2 x_i})}$$

$$\pi_{i,3} = \frac{1}{1 + e^{\beta_{02} + \beta_2 x_i}}$$

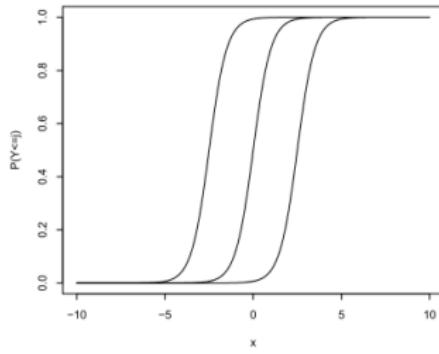
$$\Rightarrow \pi_{i,1} + \pi_{i,2} + \pi_{i,3} = 1$$



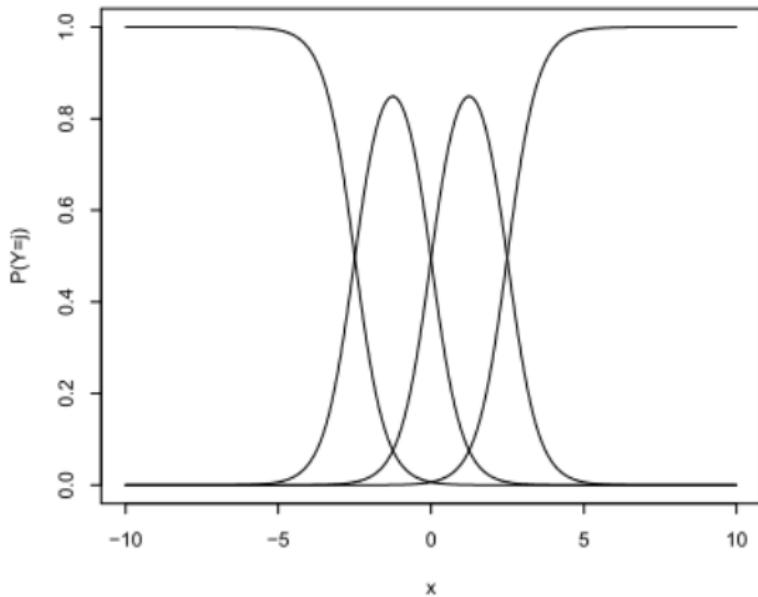
- O modelo logit cumulativo assume que o efeito de  $X$  é idêntico para todos os  $c - 1$  logitos cumulativos. Na expressão anterior,  $\beta_1 = \beta_2$ . Assim, a expressão será

$$\log \left( \frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_c} \right) = \beta_{0j} + \beta x_i, \quad j = 1, \dots, c - 1.$$

- Para todo  $j < (c - 1)$ , o modelo implica que a chance (odds) de pertencer a  $j$ -ésima categoria ou uma categoria inferior é um múltiplo de  $e^\beta$  para cada unidade acrescida em  $x$ . Este resultado é chamado de **modelo de chances proporcionais**.
- O modelo resulta em gráficos do tipo:



- As probabilidades para cada categoria resultam em gráficos do tipo:



Note que para cada valor de  $X$  as probabilidades somam 1.



O estudo analisa o nível de ferimento de motoristas numa amostra de acidentes de automóvel. Além do grau de ferimento, os seguintes dados são fornecidos: sexo do motorista, se o acidente foi em área urbana ou rural e se o motorista utilizava cinto de segurança.

- Resposta é *Injury* (nível do ferimento), sendo:
  - 1 = *not injured*;
  - 2 = *injured but not transported by emergency*;
  - 3 = *injured and transported by emergency medical services but not hospitalized*;
  - 4 = *injured and hospitalized but not die*;
  - 5 = *injured and died*.
- $X_1$  = *Gender* (sexo): *Female, Male* (Feminino, Masculino).
- $X_2$  = *Location* (localização): *Urban, Rural* (Urbano, Rural).
- $X_3$  = *Seat-Belt* (cinto de segurança): *No, Yes* (Não, Sim).
- A escala da resposta é arbitrária. As estimativas dos parâmetros serão as mesmas se utilizarmos a mesma ordem.
- As estimativas dos parâmetros serão invertidos se a categoria de referência for invertida (variável explicativa binária).



- Uma característica interessante do modelo de chances proporcionais é a invariância a escolha das classes de resposta. Isto significa que se o modelo é adequado para um escala resposta do tipo  $\{1, 2, 3, 4, 5\}$ , então será adequado também para categorias agrupadas do tipo  $\{< 3, = 3, > 3\}$ .
- No R podemos utilizar a função `polr` do pacote MASS ou `clm` do pacote ordinal com uma coluna que representa as categorias (formato longo).

```
> head(accidents.1)
  gender location seatbelt injury count
1     F    urban      no     i1   7287
2     F    urban      no     i2    175
3     F    urban      no     i3    720
4     F    urban      no     i4     91
5     F    urban      no     i5     10
6     F    urban     yes     i1 11587
```

- A função `vglm` do pacote VGAM também pode ser utilizada com categorias representadas em diferentes colunas (formato curto).

```
> accidents.s
  gender location seatbelt     i1     i2     i3     i4     i5
1     F    urban      no  7287  175  720   91   10
2     F    rural     yes 11587 126 577   48     8
3     M    urban      no 3246  73 710 159   31
4     M    rural     yes 6134  94 564   82   17
5     F    urban      no 10381 136 566   96   14
6     F    rural     yes 10969  83 259   37     1
7     M    urban      no  6123 141 710 188   45
8     M    rural     yes  6693  74 353   74    12
```

**Nota:** `polr` e `vglm` estimam funções um pouco diferentes e `vglm` só ajusta modelos de rank completo.



**Cuidado!** Johnson<sup>6</sup> (2011) destaca a importância de ler o manual para saber o que está sendo estimado em cada programa. O mesmo conjunto de dados pode apresentar diferentes estimativas em diferentes programas (ou funções do mesmo programa), e em todos os casos as estimativas estarão corretas.

- As funções `polr` e `clm` estimam,

$$\text{logito}[P(Y_i \leq j)] = \beta_{0j} - \beta_1 x_{1i} - \beta_2 x_{2i}, \quad j = 1, \dots, c - 1.$$

- A função `vglm` (e a Proc Logistic do SAS) calculam

$$\text{logito}[P(Y_i \leq j)] = \beta_{0j} + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad j = 1, \dots, c - 1.$$

- Devido a literatura médica, SAS e `vglm` tendem a indicar que um valor pequeno para os coeficientes é preferível para níveis mais baixos da variável resposta.  
Exemplo: resposta com diferentes níveis de evolução de um tumor. Valores pequenos para os coeficientes indicam ausência ou pequena evolução do tumor, i.e., coeficiente maior  $\Rightarrow$  maior chance de resposta com menor valor.
- `polr` segue uma ideia mais geral de que valores pequenos indicam menor probabilidade de observar níveis mais altos da variável.  
Exemplo: No caso do estudo de acidentes de automóvel, os pequenos valores para o coeficiente relacionado a uso de cinto de segurança, indicam uma menor probabilidade de acidentes com ferimentos graves,  
coeficiente maior  $\Rightarrow$  menor chance de resposta com menor valor.

<sup>6</sup>Johnson, P.E. (2011) Ordinal Outcomes Regression. Technical Report. Disponível no site do curso.



## Exemplo: Auto Accidents - Agresti (2019) - Exercício 6.19, pg. 192.

```
> summary(fitpa)
Call:
polr(formula = injury ~ seatbelt + location ,
      data = accidents.l,
      weights = count, Hess = T, method = "logistic")

Coefficients:
            Value Std. Error t value
seatbeltno  0.7594    0.02732   27.80
locationrural 0.7413    0.02678   27.68

> exp(cbind(OR = coef(fitpa), confint(fitpa)))
Waiting for profiling to be done...
          OR      2.5 %     97.5 %
seatbeltno 2.137084 2.025871 2.254846
locationrural 2.098701 1.991398 2.211859
```

- Os valores positivos dos coeficientes indicam uma maior chance de ferimentos graves quando não usa cinto de segurança em relação a chance de ferimentos graves quando usa cinto de segurança.
- Da mesma forma, a chance de ferimento grave quando o acidente é na área rural é maior do que a chance quando o acidente é na área urbana.



## Exemplo: Auto Accidents - Comentários

- AS razões de chances (OR) indicam

**Cinto de segurança:** que a chance de ferimento grave quando não usa cinto é cerca de duas vezes a chance de ferimento grave para os que usam cinto e a

**Localização:** a chance de ferimento grave em acidentes na área rural é cerca de duas vezes a chance de ferimento grave em acidentes na área urbana.

- O teste da razão de verossimilhança para comparação de diversos modelos, indica que o melhor modelo é aquele com variáveis explicativas referentes ao uso de cinto de segurança e localização, sem interação ( $\alpha = 0.05$ ).
- A função `polr` permite obter probabilidades para as diversas categorias da variável resposta ferimento, especificando as condições de uso de cinto de segurança e localização. Exemplo:

```
> pnr <- predict(fitpa, data.frame(seatbelt="no",
+                                     location="rural"), type = "probs")
> pyr <- predict(fitpa, data.frame(seatbelt="yes",
+                                     location="rural"), type = "probs")
> pyr - pnr
      i1          i2          i3          i4          i5
  0.086521858 -0.010500666 -0.061436966 -0.012338365 -0.002245862
```

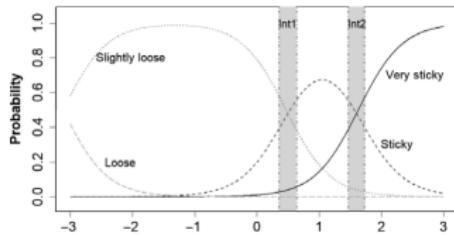
Área rural: morte mais provável se não utilizar cinto de segurança.

## Exemplo: Prediction models of rice cooking quality

von Borries et al. (2017)<sup>a</sup> avaliam textura de arroz cozido através da relação com medidas de textura instrumental. A variável resposta consta de 4 categorias de viscosidade: muito pegajoso, pegajoso, levemente solto e solto. Dois problemas são tratados no artigo:

1. Apresentação gráfica de modelo com grande número de variáveis explicativas.
2. Resposta com quatro categorias em resposta ordinal - Regressão logística politômica.

A redução do número de variáveis explicativas para duas Componentes Principais permitiu um bom ajuste do modelo e obtenção das probabilidades de previsão de pegajosidade do arroz, apresentadas na Figura 4 do artigo.



<sup>a</sup> von Borries, G.; Bassinello, P.Z.; Rios, E.S.; Koakuzu, S.N. e Carvalho, R.N. (2017) Prediction models of rice cooking quality. *Cereal Chemistry*.



## Observações Gerais

- Este tipo de modelo exige amostras relativamente grandes. Precisamos pelo menos 5 observações em cada categoria resposta, em cada nível do efeito principal.
- O teste de adequabilidade do modelo tende a rejeitar a suposição de modelo adequado para chances proporcionais nas seguintes situações:
  1. pequenas amostras;
  2. grande número de categorias;
  3. muitas variáveis explicativas.
- Estudo exploratório: quando existe preocupação em relação ao p-valor do teste de adequabilidade do modelo, podemos ajustar  $c - 1$  modelos binários e examinar as estimativas para similaridades e diferenças.
- Se não temos proporcionalidade, então a melhor abordagem pode ser tratar a resposta como nominal e ajustar o modelo logit com categoria de referência.
- O modelo de chances proporcionais NÃO é um bom modelo se as distribuições da resposta diferem em sua dispersão em vez de sua média.
- Quando o modelo proporcional não é adequado, é possível considerar um modelo cumulativo geral com efeitos separados para diferentes probabilidades cumulativas. **Problema:** as curvas podem se cruzar em alguns valores das variáveis preditoras, violando a ordem que probabilidades cumulativas devem ter. Ver Seção 6.3.2 de Agresti (2019).



## Modelo Logito de Categorias Adjacentes

- Logitos de categorias adjacentes são

$$\log \left( \frac{\pi_j}{\pi_{j+1}} \right), \quad j = 1, \dots, c - 1.$$

- Um modelo com um preditor  $x$  tem forma

$$\log \left( \frac{\pi_j}{\pi_{j+1}} \right) = \beta_{0j} + \beta x, \quad j = 1, \dots, c - 1. \quad (1)$$

- Para  $c = 3$  os logitos são  $\frac{\pi_1}{\pi_2}$  e  $\frac{\pi_2}{\pi_3}$ .
- O modelo determina os logitos para todos os pares de categorias adjacentes. Para o Modelo (1), o coeficiente de  $x$  para  $\log \left( \frac{\pi_j}{\pi_k} \right)$  é igual a  $\beta(k - j)$ . Logo, o **efeito depende da distância entre as categorias**, utilizando a escala da variável resposta.
- ✓ No R devemos utilizar `family = acat` na função `vglm`.



## Modelo Logito de Razão Sequencial

- Este modelo une as duas ideias anteriores, i.e., ajustamos  $c - 1$  logitos

$$\log\left(\frac{\pi_{i,1}}{\pi_{i,2} + \dots + \pi_{i,c}}\right), \log\left(\frac{\pi_{i,2}}{\pi_{i,3} + \dots + \pi_{i,c}}\right), \dots, \log\left(\frac{\pi_{i,c-1}}{\pi_{i,c}}\right),$$

- Neste caso, o modelo compara uma categoria com todas as categorias superiores.
- O modelo pode ser ajustado também por  $c - 1$  logitos da forma

$$\log\left(\frac{\pi_{i,1}}{\pi_{i,2}}\right), \log\left(\frac{\pi_{i,1} + \pi_{i,2}}{\pi_{i,3}}\right), + \dots + \log\left(\frac{\pi_{i,1} + \dots + \pi_{i,c-1}}{\pi_{i,c}}\right)$$

- Este modelo contrasta a probabilidade de cada categoria com a soma das probabilidades de categorias inferiores.
- Modelos logito cumulativos e de categorias adjacentes são candidatos para quase todo modelo com resposta ordinal.
- O modelo logito de razão sequencial é mais utilizado em situações em que o mecanismo sequencial define as categorias ordinais. Este tipo de situação é encontrado, por exemplo, em estudos de sobrevivência através de várias fases etárias e em progressão de estágios.
- No R devemos utilizar `family = sratio` na função `vglm`.



## Aplicações:

- ① Rios, Érica dos Santos (2015) [Modelos estatísticos para avaliação da qualidade culinária de arroz : textura e propriedades viscoamilográficas](#). Monografia de graduação em Estatística. Universidade de Brasília, Brasil.
- ② de Oliveira, Geiziane Silva (2015). [Modelos de regressão com resposta ordinal para avaliação de textura do arroz](#). Monografia de graduação em Estatística. Universidade de Brasília, Brasil.
- ③ de Oliveira, Geiziane Silva e von Borries, George (2017). [Comparing priors in Bayesian Logistic Regression for sensorial classification of rice](#). SAS Global Forum, Orlando, USA.
- ④ da Rocha, Laura Teixeira (2017) [Estudo de regiões de incerteza na avaliação e ajuste de escalas de classificação sensorial de arroz](#). Monografia de graduação em Estatística. Universidade de Brasília, Brasil.



## Referências:

- ① Agresti, A. (2013) Categorical Data Analysis, 3<sup>a</sup> edição. John Wiley & Sons.
- ② Agresti, A. (2019) An Introduction to Categorical Data Analysis, 3<sup>a</sup> edição. John Wiley & Sons.
- ③ Bilder, C.R. e Loughin, T.M. (2015) Analysis of Categorical Data with R. CRC Press.
- ④ Hosmer, D.W.; Lemeshow, S. e Sturdivant, R. (2013) Applied Logistic Regression. 3<sup>a</sup> edição. John Wiley & Sons.

