Introdução à Regressão Logística

Parte 1

Unidade III

INTRODUÇÃO

- Os métodos apresentados para análise de tabelas de contingência nos ajudam a *investigar os efeitos de uma variável explicativa categórica sobre uma variável de resposta categorizada.*

- **Ideia:**
  - *Utilizar **modelos** para analisar dados categorizados*.

- Construção de modelo tem como foco estimação de parâmetros que descrevem a natureza e a intensidade da associação (efeitos) em termos de um pequeno número de parâmetros.
  - resultam na análise feita pelo métodos apresentados tabelas de contingência
  - é mais informativo que testar a significância meramente.
- Contudo, os modelos podem lidar com situações mais complicadas, como analisar simultaneamente os efeitos de diversas variáveis explicativas, que podem ser categóricas ou quantitativas ou ambas.

- **Vantagens:**
  - A forma estrutural do modelo descreve os padrões de associação e interação.
  - A magnitude dos parâmetros do modelo determinam a força e a importância dos efeitos.
  - As inferências sobre os parâmetros avaliam quais variáveis explicativas estão realmente associadas à variável resposta Y, enquanto são controlados os efeitos de outras variáveis, como possíveis variáveis confundidoras.
  - Os valores previstos do modelo suavizam os dados e fornecem estimativas melhoradas da média de Y em possíveis valores de variáveis explicativas.

INTRODUÇÃO

- Estudo da relação entre características dos elementos da população pesquisada.

- **Objetivo:** criar um modelo que descreva a relação entre um resultado (*variável resposta* ou *dependente*) e um conjunto de variáveis independentes (*preditoras* ou *explicativas* – covariáveis)

Modelos Lineares Generalizados

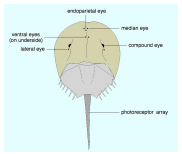
- Os Modelos de Regressão são amplamente utilizados neste caso:
  - quando a variável resposta é **quantitativa** - Modelo de Regressão Linear (Simples ou Múltiplo)
  - variável resposta é **categorizada**, tendo dois ou mais valores possíveis - Modelo de Regressão Logística.
  - variável de **resposta discreta** para as quais o resultado é uma contagem - Modelos Loglineares.

Exemplos

- Em uma análise sobre se as empresas têm ou não um departamento de *marketing*, de acordo com o tamanho da empresa.
- Em um estudo de participação na força de trabalho de mulheres casadas, em função da idade, número de filhos e da renda do cônjuge.
- Em um estudo longitudinal de doença coronariana em função da idade, gênero, histórico de tabagismo, nível de colesterol, porcentagem do peso corporal ideal e pressão arterial

PROBLEMA

- *Os dados analisados neste exemplo foram extraídos de um estudo de ninhadas de caranguejo do tipo ferradura realizado por J. Brockmann, a ser impresso em Ethology (1996).*

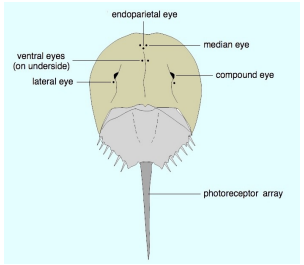


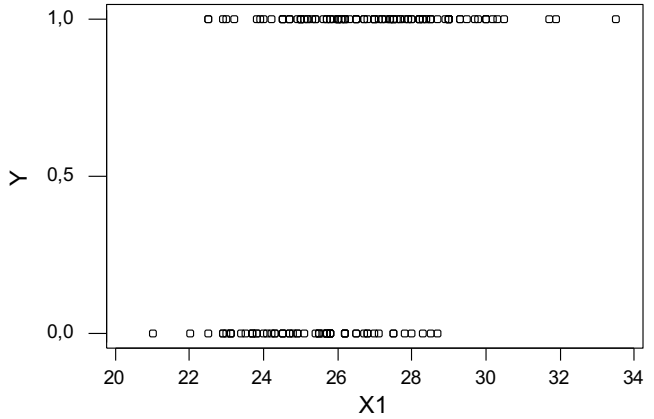
- Neste estudo cada fêmea de caranguejo ferradura tinha um caranguejo macho ligado a ela em seu ninho. O estudo investiga fatores que afetam a existência de qualquer outros machos, chamados *satélites*, morando perto da fêmea. Pensou-se que um dos possíveis fatores seria a *largura da carapaça da fêmea de caranguejo.*

PROBLEMA

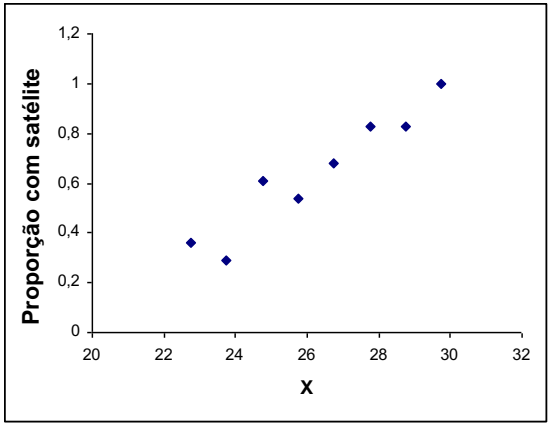
Para tal foi selecionada uma amostra de 173 fêmeas e as seguintes características foram investigadas para cada uma delas:

- X - largura da carapaça da fêmea *em cm;*
- Y - *se a fêmea tem pelo menos um satélite (1 – sim e 0- não)*





Largura	Nº de Casos	Satélite		Média (Proporção com satélite)
		Sim	Não	
< 23,25	14	5	9	0,36
23,25 |-- 24,25	14	4	10	0,29
24,25 |-- 25,25	28	17	11	0,61
25,25 |-- 26,25	39	21	17	0,54
26,25 |-- 27,25	22	15	7	0,68
27,25 |-- 28,25	24	20	4	0,83
28,25 |-- 29,25	18	15	3	0.83
≥ 29,25	14	14	0	1,00



PORQUE USAR REGRESSÃO LOGÍSTICA ?

Natureza da relação entre a variável resposta e a variável independente

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Considerando que:

$$E(\varepsilon_i) = 0$$

temos que:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

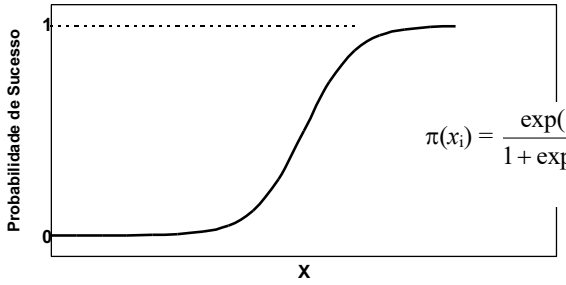
Esta expressão implica na possibilidade de $E(Y_i)$ assumir qualquer valor entre $-\infty$ e $+\infty$.

- Na ***Regressão Logística Binária*** a variável binária ***Y*** pode assumir apenas dois resultados possíveis:
  - 1 quando ocorre "sucesso"
  - 0 no caso de "insucesso".
- Sua distribuição de probabilidade é a ***distribuição de Bernoulli*** especificada pelas probabilidades:

  $P(Y_i = 1) = \pi_i$ (probabilidade de sucesso)

  $P(Y_i = 0) = 1 - \pi_i$ (probabilidade de insucesso)
- Esta distribuição tem média $E(Y_i) = \pi_i$ ***(ou seja, entre 0 e 1)*** e variância $V(Y_i) = \pi_i\ (1 - \pi_i\ )$ ***(ou seja, não é constante)***.

Modelo de Regressão Logística Simples



$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

Propriedades da função Logística:

1. A função logística é monótona (crescente ou decrescente, dependendo do sinal de β_1).
2. É quase linear no intervalo de crescimento e nas extremidades aproxima-se gradualmente de 0 e 1.
3. Pode ser linearizada.

Modelo de Regressão Logística Simples

$$E(Y_i) = \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- Transformação logito:

$$\pi^*(x_i) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right),$$

- resposta média logito:

$$\pi^*_i = \pi^*(x_i) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 X_i$$

Estimação dos Parâmetros do Modelo

Considere que para cada valor fixado x_i, $i = 1,...,c$ da variável explicativa X observamos variáveis aleatórias independentes $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, com

$$Y_{ij} \sim Bernoulli(\pi_i), \quad j = 1,2,...,n_i,$$

e

$$\pi_i = \pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

Isto significa que para cada valor fixado da variável explicativa x_i observou-se n_i replicações independentes da variável resposta. Vamos supor que os vetores aleatórios que definem cada grupo de replicações $(Y_{i1},...,Y_{in_i}) \quad i = 1,2,...,c$ são independentes. Assim temos que as variáveis aleatórias definidas por

$$Z_i = \sum_{j=1}^{n_i} Y_{ij}$$

representam o número de sucessos dentre as n_i repetições do grupo correspondente ao valor x_i, têm distribuição $Binomial(n_i, \pi_i), \quad i = 1,...,c$ e, além disso são independentes.

Assim, dada uma amostra observada $Z_1, Z_2, ..., Z_c$, a função de distribuição conjunta é dada por:

$$g(Z_1, Z_2, ..., Z_c) = \prod_{i=1}^{c} f(z_i) = \prod_{i=1}^{c} \binom{n_i}{Z_i} (\pi_i)^{Z_i} (1-\pi_i)^{n_i - Z_i} \qquad (5)$$

e denominada função de verossimilhança.

Pelo método de máxima verossimilhança deve-se maximizar o logaritmo da função de verossimilhança associada, dado por:

$$ln(g(Z_1, Z_2, ..., Z_c)) = ln\left(\prod_{i=1}^{c} \binom{n_i}{Z_i} (\pi_i)^{Z_i} (1-\pi_i)^{n_i - Z_i}\right) =$$

$$= \sum_{i=1}^{c} \left[ln\binom{n_i}{Z_i} + Z_i\, ln(\pi_i) + (n_i - Z_i)\, ln(1-\pi_i) \right]$$

$$= \sum_{i=1}^{c} \left[ln\binom{n_i}{Z_i} + Z_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\, ln(1-\pi_i) \right] =$$

$$= \sum_{i1=}^{c} \left[ln\binom{n_i}{Z_i} + Z_i(\beta_o + \beta_1 x_i) - n_i\, ln(1 + exp(\beta_0 + \beta_1 x_i)) \right] = l(\beta_0, \beta_1)$$

Pode-se provar que o máximo global da função l é o único ponto crítico da função, ou seja, o único ponto (b_0, b_1) que é a solução da equação

$$l'(\beta_0, \beta_1) = 0.$$

As derivadas parciais de l são dadas por

$$\frac{\partial l}{\partial \beta_0}(\beta_0, \beta_1) = \sum_{i=1}^{c}\left[Z_i - n_i \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}\right]$$

$$\frac{\partial l}{\partial \beta_1}(\beta_0, \beta_1) = \sum_{i=1}^{c}\left[Z_i x_i - n_i x_i \left(\frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}\right)\right]$$

Não existe uma forma algébrica fechada para (b_0, b_1) solução do sistema

$$\left(\frac{\partial l}{\partial \beta_0}(\beta_0, \beta_1), \frac{\partial l}{\partial \beta_1}(\beta_0, \beta_1)\right) = (0,0)$$

A solução estão deve ser encontrada utilizando métodos iterativos, como os procedimentos de *Newton-Raphson* ou do *Score*. Algoritmos estão implementados na maioria dos programas estatísticos.

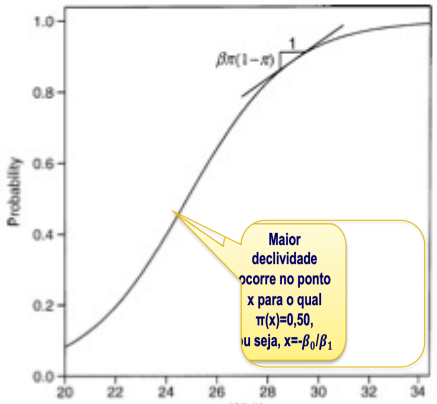
Interpretação de β_1

- O parêmetro β_1 determina a taxa de crescimento ou decrescimento para curva (em forma de ***S***) para $\pi(x)$.

$$\frac{d\pi(x)}{dx} = \beta_1 \pi(x)\left(1 - \pi(x)\right)$$

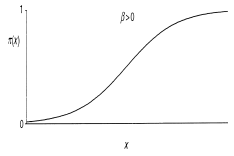
(depende do valor de x)

  - **sinal de** β_1 indica se a curva cresce $(\beta_1 > 0)$ ou decresce $(\beta_1 < 0)$
  - A medida que $|\beta_1|$ aumenta, a taxa de mudança cresce.

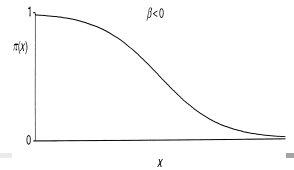


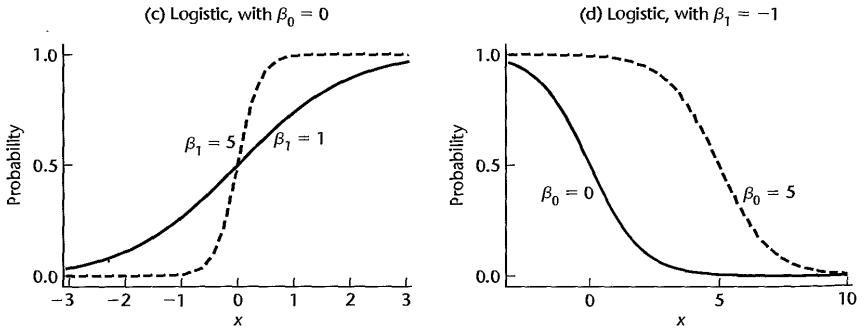
Interpretação de β_1

- **sinal de** β_1 indica se a curva cresce $(\beta_1 > 0)$



ou decresce $(\beta_1 < 0)$





- Considerando a fórmula:

$$logito(\pi(x)) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

  - O logito cresce de β_1 unidades para cada unidade que x aumenta.

  - Lembrando da odds (chance estimada se sucesso) está presente na expressão do modelo tem-se que:

$$\frac{\pi(x)}{1-\pi(x)} = exp(\beta_0 + \beta_1 x) = e^{\beta_0}\left(e^{\beta_1}\right)^x$$

- Para um valor $x+1$ a chance de sucesso é dada por:

$$\frac{\pi(x+1)}{1-\pi(x+1)} = exp(\beta_0 + \beta_1(x+1)) = e^{\beta_0}\left(e^{\beta_1}\right)^x e^{\beta_1}$$

  - De modo que a razão de chances entre $x+1$ e x é dada por e^{β_1}.

  - Para cada unidade que aumenta em x a chance estimada de sucesso (odds) é multiplicada por e^{β_1}. Isto é, a "odds" no nível $x+1$ é igual a odds no nível x multiplicada por e^{β_1}.

Exemplo – Problema Caranguejo Ferradura

Number of Observations Read	173
Number of Observations Used	173

Response Profile		
Ordered Value	y	Total Frequency
1	1	111
2	0	62

Probability modeled is y='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	198.453
SC	230.912	204.759
-2 Log L	225.759	194.453

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.3059	1	<.0001
Score	27.8752	1	<.0001
Wald	23.8872	1	<.0001

Estimated Covariance Matrix		
Parameter	Intercept	x
Intercept	6.910227	-0.26685
x	-0.26685	0.01035

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
x	1	0.4972	0.1017	23.8872	<.0001

$$\widehat{\beta}_1 = 0,4972 \Rightarrow \widehat{\theta} = e^{0,4972} = 1,6441$$