Regressão Logística

Marcelo Azevedo Costa
Departamento de Engenharia de Produção
Universidade Federal de Minas Gerais

Introdução

- Seja ξ um experimento e Y uma variável aleatória discreta associada

$$Y \in \{0,1\}$$ [{sucesso,fracasso},{sim,não},{verdadeiro, falso}]

Y é uma variável aleatória com distribuição de bernoulli, se sua pdf é definida como:

$$P(Y = y) = p^{y}(1-p)^{1-y}$$

$$Y \sim Bernoulli(p)$$

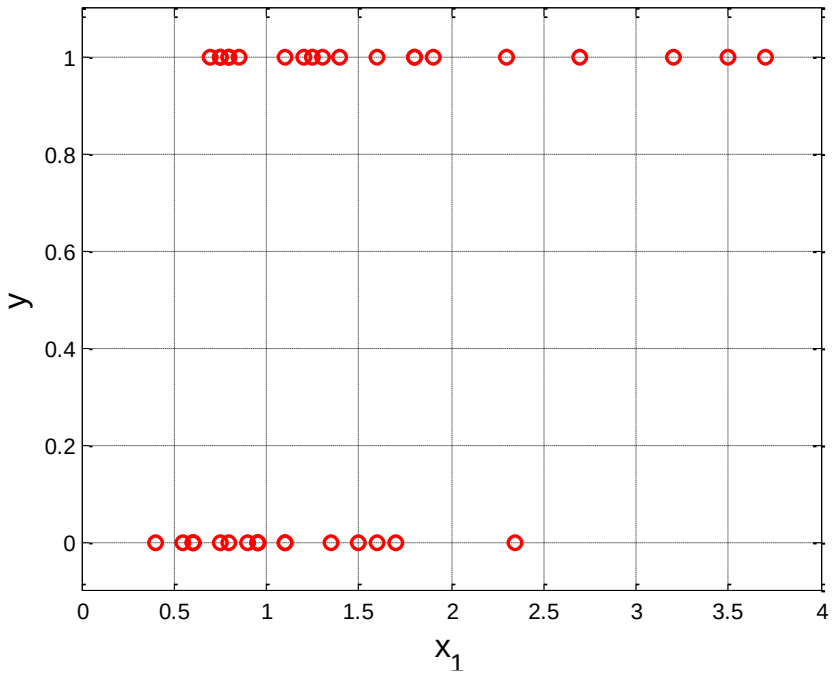
$$Y = \begin{cases} 0, & \text{com prob.}\,(1-p) \\ 1, & \text{com prob.}\,(p) \end{cases}$$

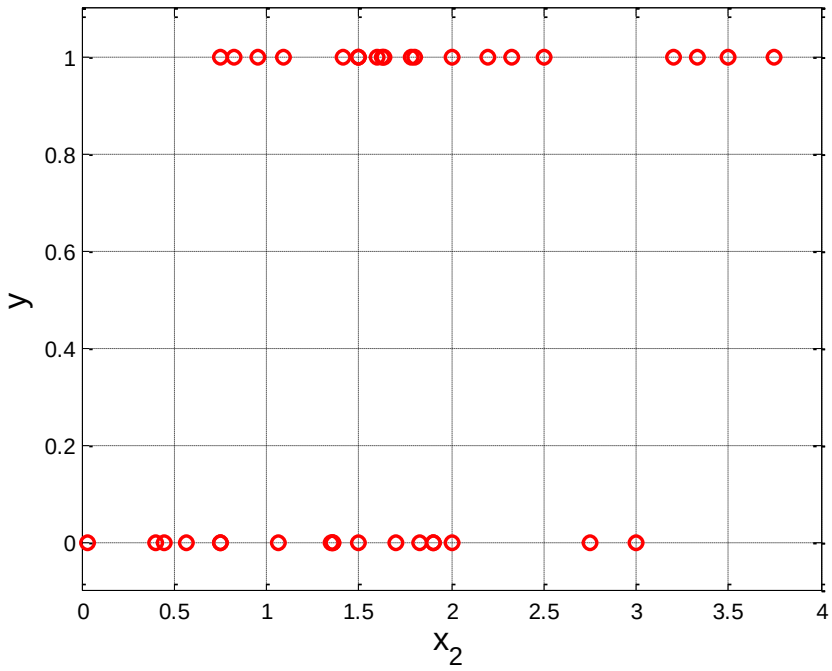
Análise Exploratória

- Dados do experimento sobre a influência da razão e do volume de ar, inspirado na ocorrência de vaso-constrição da pele dos dedos da mão

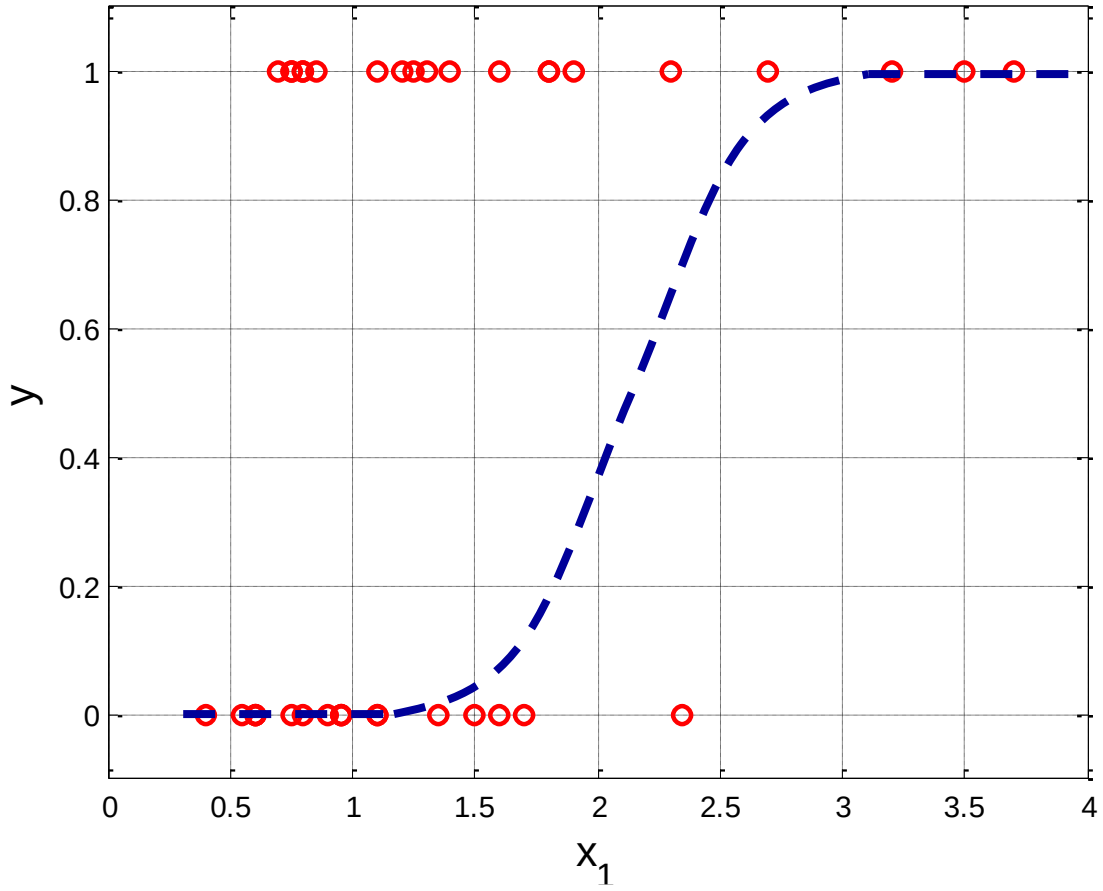
  - **y** : ocorrência (y=1) ou ausência (y=0) de compressão de vasos
  - $\mathbf{x_1}$: log do volume de ar inspirado
  - $\mathbf{x_2}$: logaritmo da razão de ar inspirado

Gráficos de Dispersão





Análise da Proporção



- Restrição:

$$0 \leq p \leq 1$$

- *p* pode ser interpretado como a probabilidade de um indivíduo ser um "*sucesso*"

Modelo de Regressão Logística

- O modelo logístico linear é utilizado para analisar respostas binárias através de um conjunto de variáveis explicativas

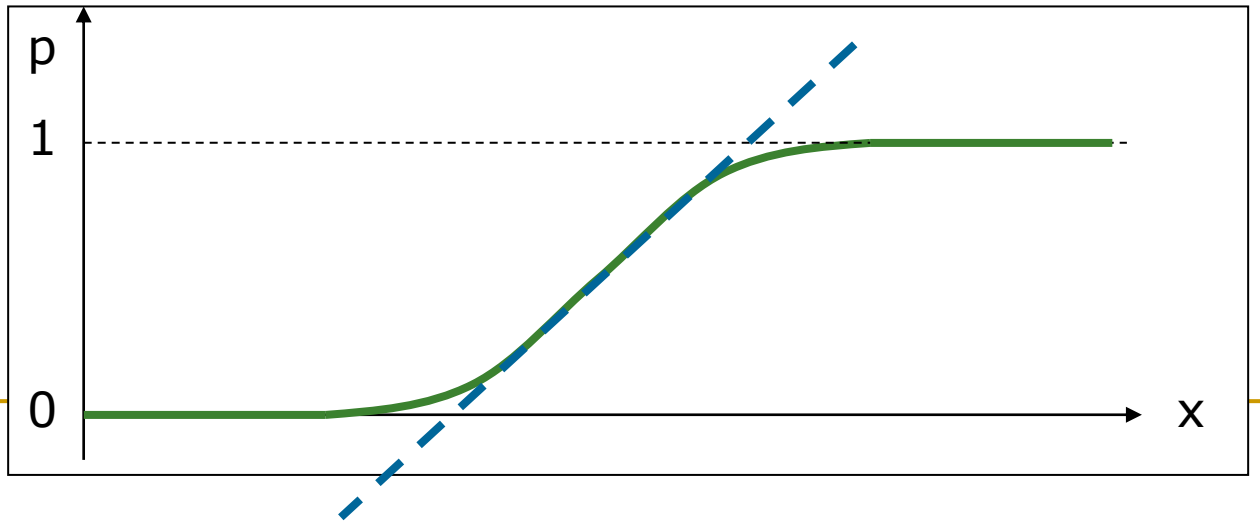
- A relação entre a probabilidade de sucesso p (π ou μ) e o conjunto de variáveis explicativas é dada através da função de ligação logística.

Especificação do modelo

- Suponha *n* observações bernoulli na forma y_i, i = 1,2, ..., n e p variáveis explicativas: $x_1, x_2, \ldots, x_p$.

- p_i ou $p(X_i)$ é a probabilidade de sucesso correspondente à i-ésima observação.

$$\text{logit}(p_i) = \log\left\{\frac{p_i}{1-p_i}\right\} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$



Equação da Proporção

Seja: $$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Então: $$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Desde que y_i seja uma observação proveniente de uma distribuição bernoulli com média p_i, o valor esperado de y_i é:

$$E(y_i) = \left(\frac{e^{\eta_i}}{1+e^{\eta_i}} \right)$$

Sensibilidade/Especificidade

- **Eventos**
  - T_+ : teste positivo
  - T_- : teste negativo
  - D_+ : indivíduo portador da doença
  - D_- : indivíduo não portador da doença

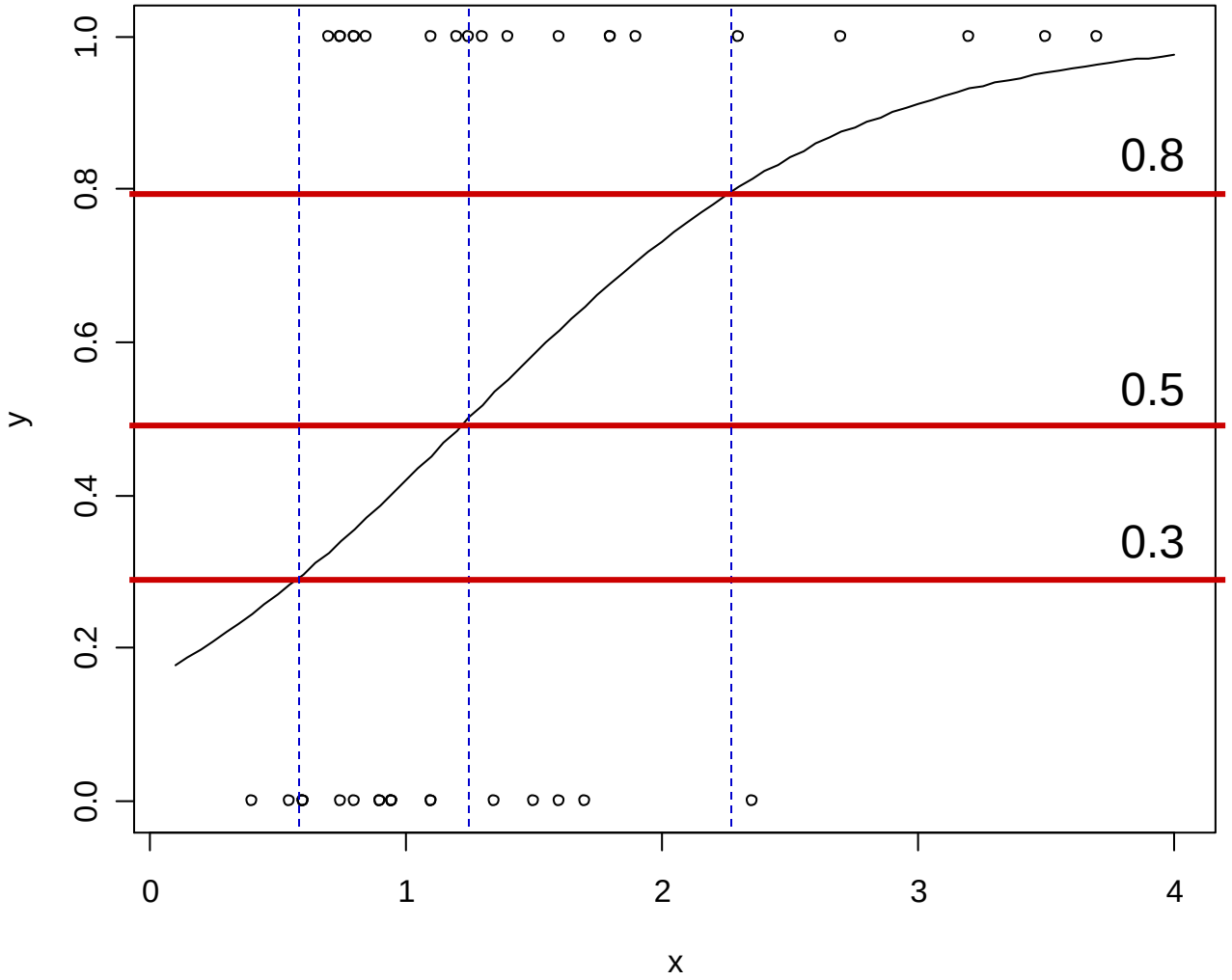
Sensibilidade do Teste	**Especificidade do Teste**
$s = P(T_+ \mid D_+)$	$e = P(T_- \mid D_-)$

Mais definições

- **Sensibilidade**: é a probabilidade que o resultado do teste (modelo) seja positivo quando a doença é presente.

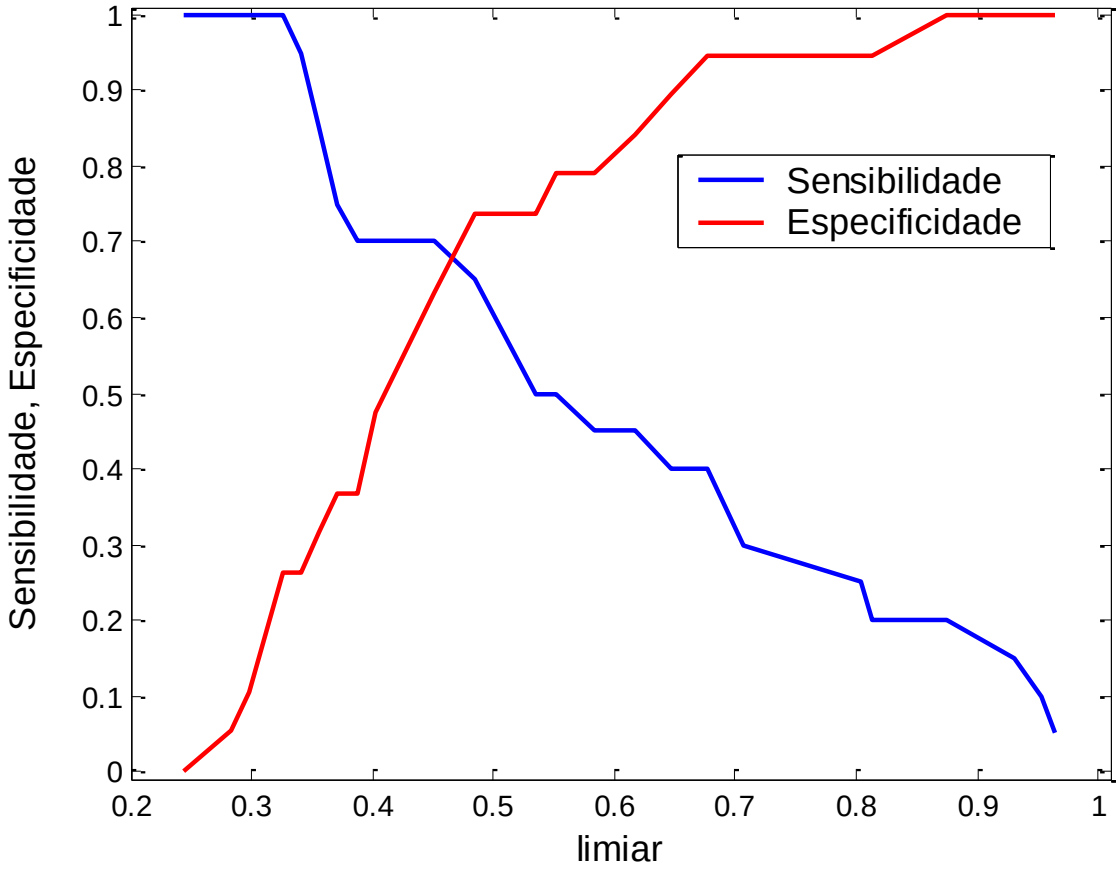
- **Especificidade**: é a probabilidade que o resultado do teste (modelo) seja negativo quando a doença não está presente.

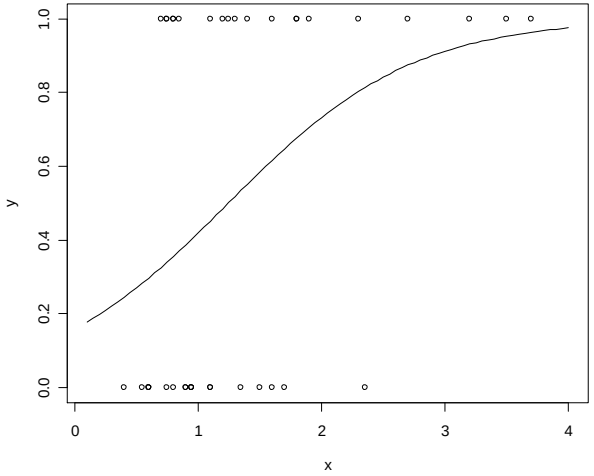
Curva ROC no Modelo Logístico



aula06.R

Sensibilidade e Especificidade versus limiar (logística)





Comparação de Modelos com a curva ROC

AUC: Area Under the Curve

- **O valor de AUC de um classificador é equivalente a probabilidade de que o classificador atribua um índice de detecção maior para uma elemento (verdadeiramente) positivo de uma amostra aleatório em relação a um elemento (verdadeiramente) negativo também escolhido arbitrariamente.**

- **Em termos práticos classificadores com maiores índices de AUC apresentam maior poder de discriminação.**

- O valor de AUC pode ser utilizado para selecionar classificadores

AUC : Comparação de Classificadores