Disciplina:

Técnicas de Amostragem e Modelos de Regressão

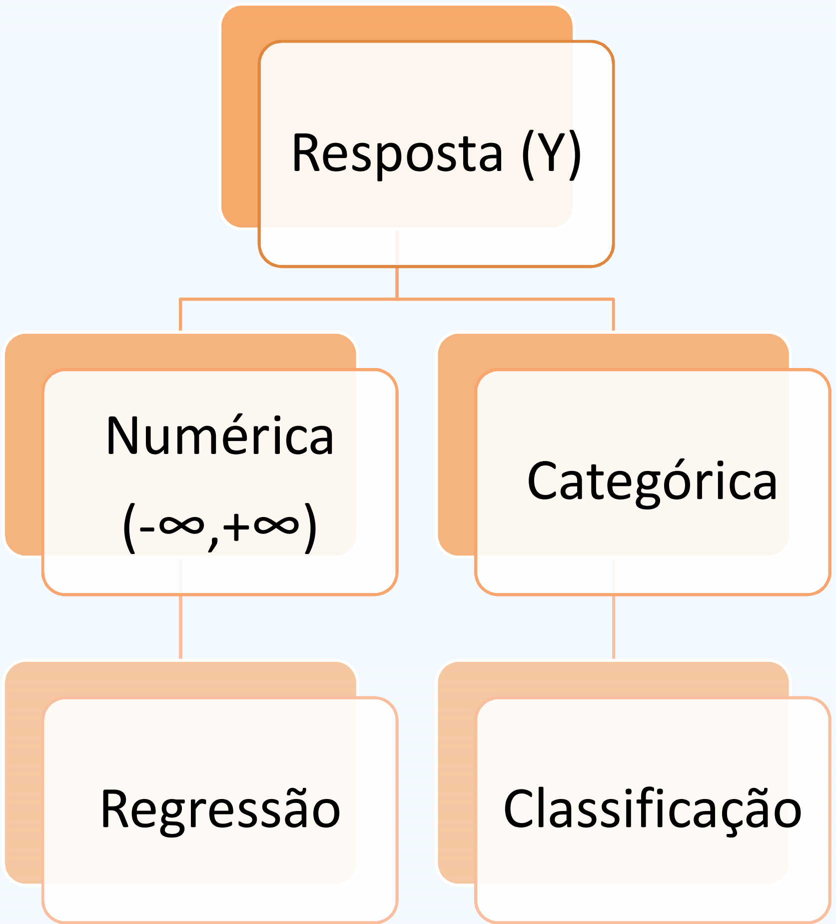
Professora: Anaíle Mendes Rabelo

PUC Minas

PUC Minas

Regressão Logística

Modelos de Machine Learning



Regressão Logística

Variáveis Categóricas

Binárias

Estudar a probabilidade de ocorrência

Obtemos uma resposta numérica (probabilidade)

Probabilidade Valores estão entre 0 e 1

Quando utilizar a Regressão logística

VARIÁVEL DEPENDENTE COM DISTRIBUIÇÃO BINOMIAL

REPROVAÇÃO NO TESTE DE HOMOCEDASTICIDADE

RESÍIDUOS NÃO TEM DISTRIBUIÇÃO NORMAL

Regressão logística

- Técnica de classificação usada para prever uma resposta qualitativa

Exemplos

- Fraudes – Probabilidade da transação ser fraudulenta ou não. Classificação usada para prever uma resposta qualitativa (binária)
- Marketplace - probabilidade do cliente comprar ou não a mercadoria

PUC Minas

Regressão Logística

Pressupostos da Regressão Logistica

A variável reposta precisa ser qualitativa, dicotômica ou binaria(modelo tradicional)

As preditoras podem ser quantitativas ou categóricas (transformadas em binarias ou Dummy)

Assume que as observações são independentes, que uma não afeta a outra

Por que não uma regressão linear?

Problema da ordenação:

Não podemos utilizar uma regressão linear para prever eventos categóricos.

Ex: Condição médica dos pacientes:
- AVC
- Parada Cardíaca
- Overdose

Podemos simplesmente ordenar e realizar a regressão linear?

1 = AVC
2 = Parada Cardíaca
3 = Overdose

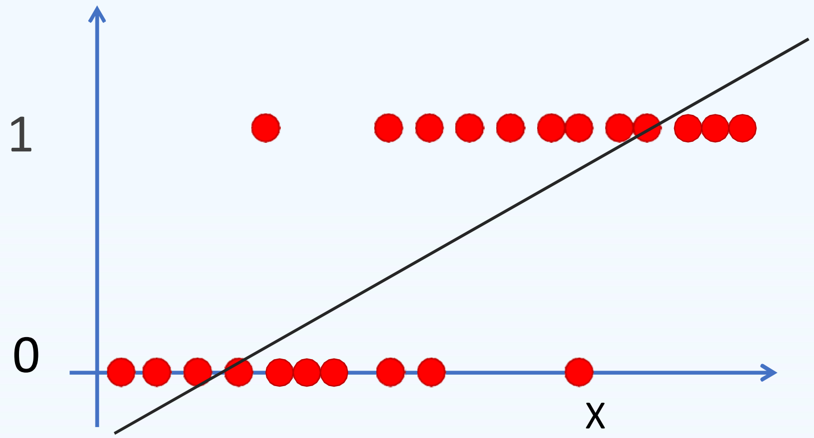
Devemos utilizar as variáveis dummy e transformar as variáveis.

PUC Minas

Regressão Logística - Definição Teórica

Permite **estimar a probabilidade** associada à **ocorrência de determinado evento** em vista de um conjunto de variáveis preditoras.

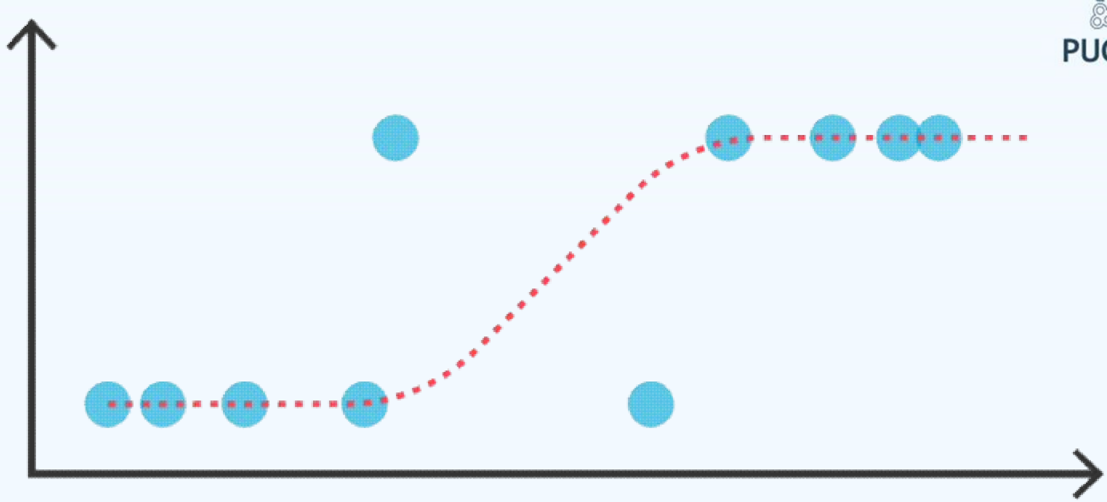
- Probabilidade de sucesso (1)
- Probabilidade de fracasso (0)



Ao interpretar Y como probabilidade, temos que realizar transformações para que a resposta de nossa regressão esteja entre 0 e 1

PUC Minas

Regressão Logística



Função Logística

- Retorna os valores entre 0 e 1
- Tem formato de “S”

Suponha que o modelo tenha a seguinte forma:

$$\mathrm{Y} = X'B + e$$

Em que $X' = [1, X_{i1}, X_{i2}, \dots, X_{in}]$, $\mathrm{B} = [\beta_0, \beta_1, \beta_2, \dots, \beta_n]$, e a variável resposta entre 0 e 1.

Assumimos que a variável reposta é uma variável aleatória de Bernoulli, com função de Probabilidade:

y_i	**Probabilidade**
1	P($\boldsymbol{y}_i$ = 1) = $\boldsymbol{p}_i$
0	P($\boldsymbol{y}_i$ = 0) = 1 - $\boldsymbol{p}_i$

Recapitulando - Distribuição de Bernoulli

A variável aleatória Y tem distribuição de Bernoulli se apresenta apenas **dois resultados possíveis**, representados por 0 (fracasso ou negativo) e 1 (sucesso ou positivo). O parâmetro $0 < p < 1$ é a **probabilidade de sucesso**. Dessa forma, a função de probabilidade é

$$p(y) = \begin{cases} 1 - p & \text{se "fracasso" ou } y = 0 \\ p & \text{se "sucesso" ou } y = 1, \end{cases}$$
$$= p^y \cdot (1-p)^{1-y}, \quad y \in \{0, 1\}.$$

Denotamos por $Y \sim \mathrm{Ber}(p)$.

Com isso, Y apresenta:

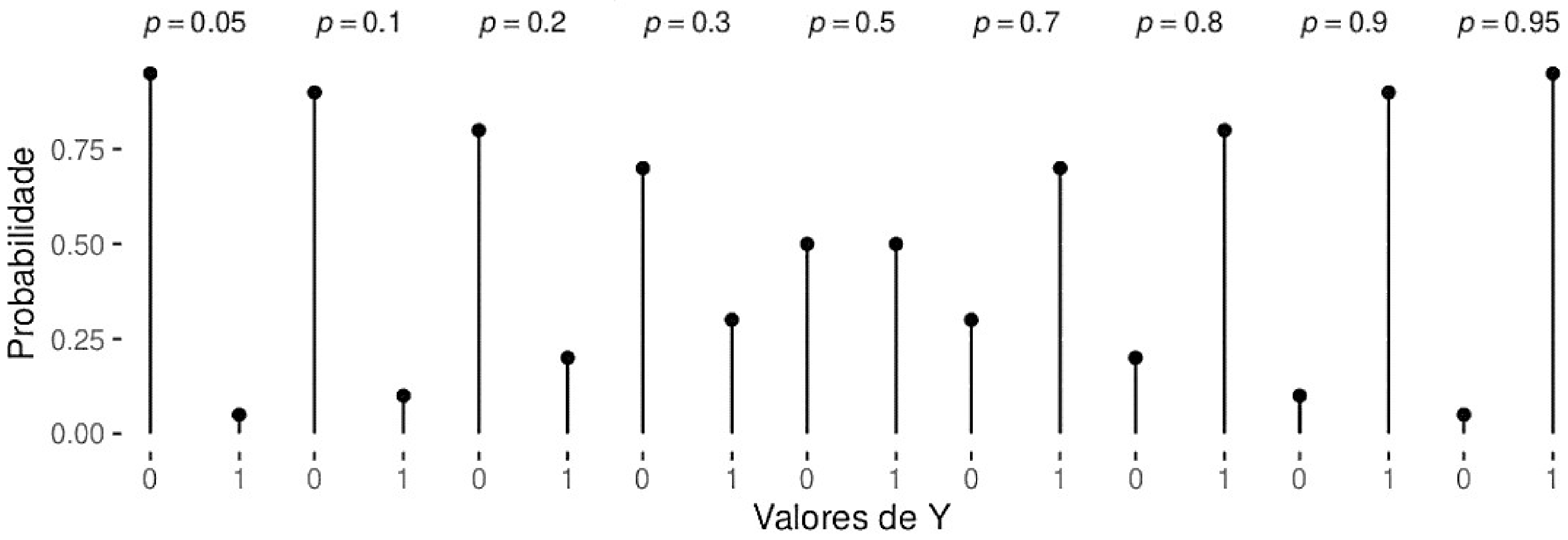
- $\mu = \mathrm{E}(Y) = p$.
- $\sigma^2 = \mathrm{V}(Y) = p \cdot (1-p)$.

Distribuição de Bernoulli

Onde,

P -> Probabilidade de sucesso

1 – p -> Probabilidade de fracasso



Regressão Logistica

- Uma vez que E(y_i) = 1(p_i) + 0(1- p_i) = p_i

- Temos que:

$$E(y_i) = x'_i\beta = p_i$$

Logo, a **resposta encontrada na regressão logística** sempre será a **probabilidade de sucesso (1)**

PUC Minas

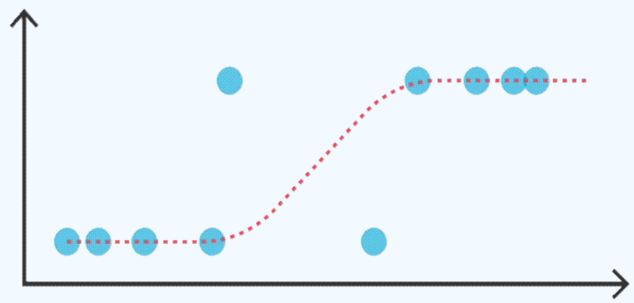
Regressão logística

Resposta (Y) entre 0 e 1

Requer transformação da nossa função

Função Logit

$$E(y) = \frac{e^{x'\beta}}{1 + e^{x'}} = \frac{1}{1 + e^{-x'\beta}}$$



Conseguimos transformar essa linha em forma de s, em uma linha reta (linearização), para que possa ser possível o cálculo dos coeficientes?

Regressão Logística

A regressão logística pode ser linearizada:

$$\eta = x'\beta$$

Ser o preditor linear, onde η é definido pela transformação.

$$\eta = ln\,\frac{p}{p-1}$$

Essa transformação é frequentemente chamada de **transformação logit** da probabilidade p e a razão $\frac{p}{p-1}$ é chamada de chance odds .

Função ligação que associa os valores esperados da função resposta aos preditores lineares do modelo.

Regressão Linear – Analisando o erro

- Como temos uma resposta binária (0 e 1), temos que os termos de erro só podem ter dois valores:

$$\varepsilon_i = 1 - x'_i\beta , \qquad y_i = 1$$
$$\varepsilon_i = - x'_i\beta , \qquad y_i = 0$$

Logo os **erros não podem ser normais**, e a **variância não é constante**.

$$E(\sigma^2{}_{yi}) = E\{y_i - E(y_i)\} = (1- p_i)^2\, p_i + (0- p_i)^2(1- p_i) = p_i(1- p_i)$$
$$\sigma^2{}_{yi} = E(y_i)[1 - E(y_i)]$$

Estimação de Parâmetros

- A estimação dos parâmetros de $x'_i\beta$ é realizada a partir do método de máxima verossimilhança;
- Como nosso dados seguem a distribuição de Bernoulli, então a distribuição de probabilidade é dada por:

$$f_i\,(y_i) = p_i^{y_i}[1\ - E(p_i)]^{1\,-\,y_i}\ , \qquad \text{i = 1, 2, 3, …., n}$$

- E cada observação assume o valor de 0 e 1.
- Logo a função de verossimilhança para v.a. independentes pode ser dada por:

$$\text{L}(y_1, y_2, y_3, \dots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = = \prod_{i=1}^{n} p_i^{y_i}[1\ - E(p_i)]^{1-\,y_i}$$

Estimação de Parâmetros

PUC Minas

A forma geral de um modelo de regressão logística é

$$y_i = \mathrm{E}(y_i) + \varepsilon_i$$

Em que as observações são variáveis aleatórias independentes de Bernoulli com valores esperados

$$E(y) = p_i = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

Interpretação dos Parâmetros

PUC Minas

- Considere o caso em que o preditor linear tem apenas uma variável preditora, de forma que o valor ajustado do preditor linear em um valor particular de x, xi é:

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1\, x_i$$

O valor ajustado em x_i +1 é:

$$\hat{\eta}(x_i + 1\,) = \hat{\beta}_0 + \hat{\beta}_1\, (x_i + 1)$$
$$\hat{\eta}(x_i + 1\,) = \hat{\beta}_0 + \hat{\beta}_1\, (x_i) + \hat{\beta}_1$$

- E a diferença dos valores previstos é:

$$\hat{\eta}(x_i + 1\,) - \hat{\eta}(x_i) = \hat{\beta}_1$$

$$\hat{\eta}(x_i + 1\,) - \hat{\eta}(x_i) = \ln\left(odds_{x_i+1}\right) - \ln\left(odds_{x_i}\right) = ln\,\frac{odds_{x_i+1}}{odds_{x_i}} = \hat{\beta}_1$$

$$\widehat{O_R} = \frac{odds_{x_i+1}}{odds_{x_i}} = \ e^{\hat{\beta}_1}$$

- Para as variáveis preditora binárias, podemos realizar a seguinte análise:

$$ln\frac{p}{1-p} = \hat{\beta}_0 + \hat{\beta}_1\, x_i$$

Para x_i = 0, temos :

$$ln\frac{p_0}{1-p_0} = \hat{\beta}_0$$

Para x_i = 1, temos :

$$ln\frac{p_1}{1-p_1} = \hat{\beta}_0 + \hat{\beta}_1$$

Logo:

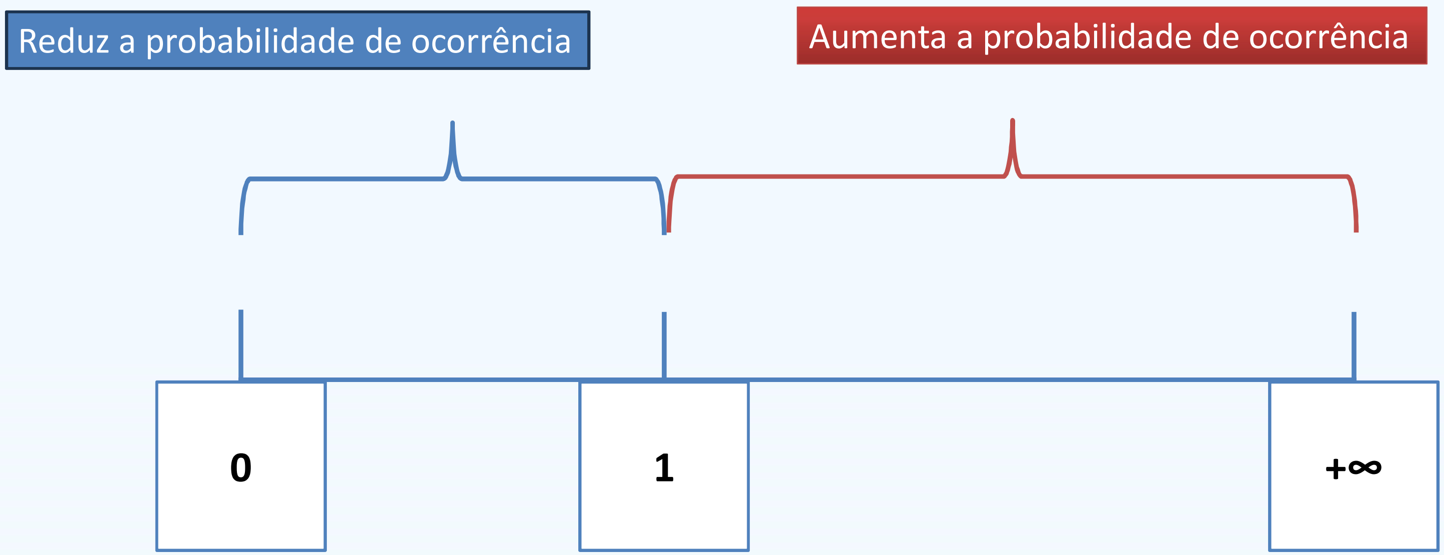
$$ln\frac{p_1}{1-p_1} - ln\frac{p_0}{1-p_0} = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0$$

$$ln\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \hat{\beta}_1 \rightarrow \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = e^{\hat{\beta}_1}$$

PUC Minas

ODDS

$$\widehat{O_R} = \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$



Exemplo de Interpretação:

Suponha que você esteja estudando a probabilidade de uma pessoa comprar um produto online com base em duas variáveis: idade e gênero. Após ajustar um modelo de regressão logística, você obtém os seguintes resultados:

- Para a variável idade, o odds ratio é 1.05.
- Para a variável gênero (sendo 1 para masculino e 0 para feminino), o odds ratio é 0.8.

Interpretação:

- Para a idade: A cada aumento de uma unidade na idade, as chances de comprar o produto aumentam em 5%.
- Para o gênero: As chances de comprar o produto são 20% menores para homens em comparação com mulheres.

PARÂMETROS DOS MODELOS

- Verificar a significância das variáveis do modelo

- Teste de hipótese para determinar se a variável preditora do modelo é significativamente relacionada com variável resposta do modelo

  – Teste de Wald