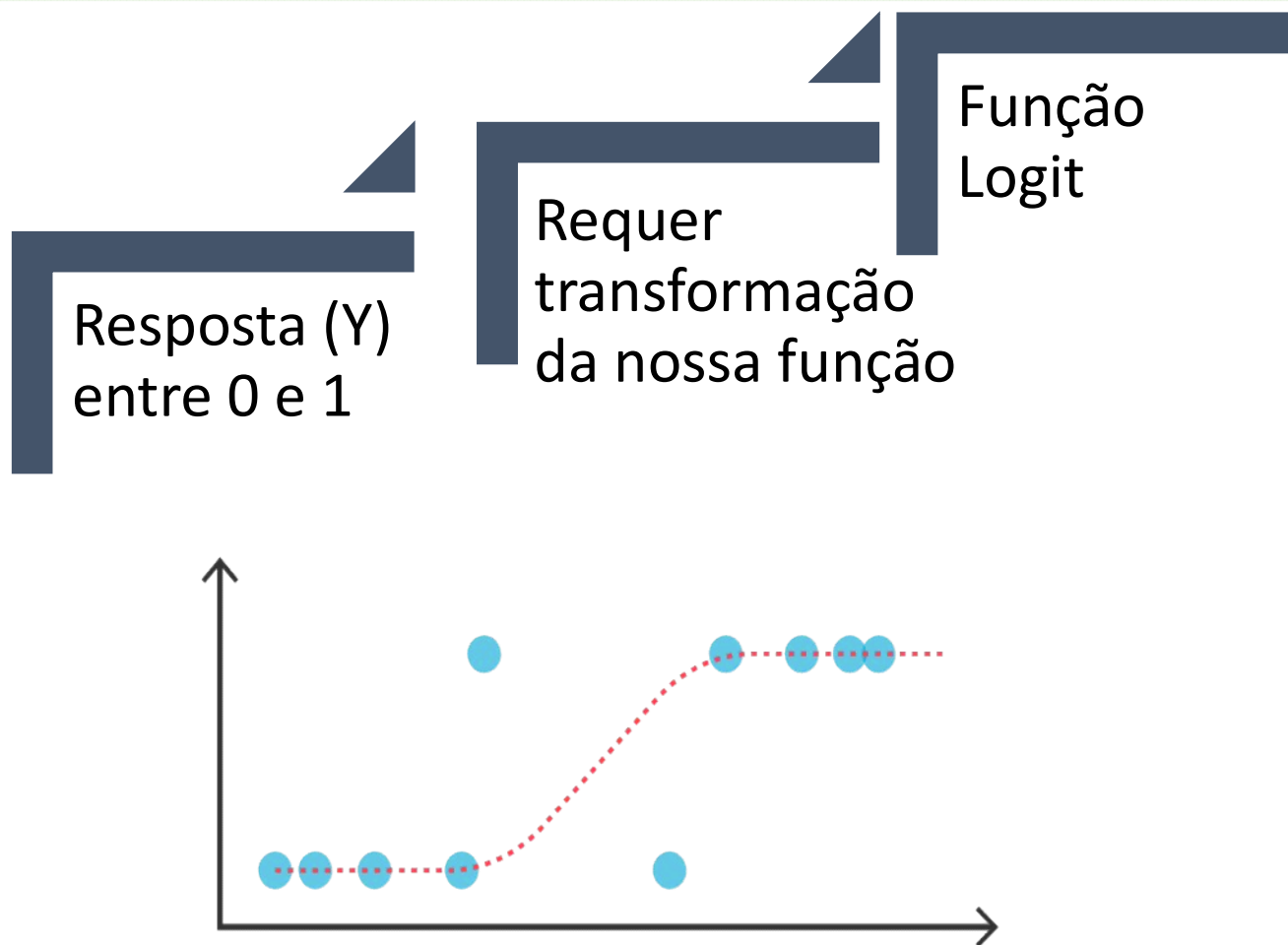




PUC Minas
Virtual

Regressão Logística

Regressão logística



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

Escolha da função de ligação

A função de ligação deve **ser contínua, diferenciável e monótona**

Essa função deve ser **capaz de produzir** os valores de η no intervalo 0 e 1.

Capaz de Linearizar a relação entre os componentes

Essa função deve proporcionar interpretações simples

Linearização

Escolha da Função de ligação

Componente Aleatório

Função de ligação

Componente Sistemático

Escolha da função de ligação

Obter como função de ligação, através da inversa de F , função distribuição acumulada (fda) de alguma variável aleatória contínua,

$$g(\pi_i) = F^{-1}(\pi_i) = X'B$$

Ou de forma equivalente,

$$\pi_i = F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})$$

Temos três funções de ligação usuais para dados binários:

**Logito , Probit e
complemento log-log (Distribuição Gumbel – valores extremos)**

Função de Ligação - Logito

A função de ligação logito baseia-se na fda da distribuição logística em sua forma padrão ($\mu = 0$ e $\sigma = 1$)

$$F(Z) = \frac{e^z}{1 + e^z}, z \in \mathbb{R}.$$

Aplicando essa função em :

$$\pi_i = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in})$$

Temos então o GLM baseado na função de ligação logito fica definido por :

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

Função de Ligação - Logito

GLM baseado na função de ligação logito fica definido por :

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

Ou, na escala do preditor linear ,

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

Função de Ligação - Logito

Temos que o modelo de regressão logística é definido pela utilização da função de ligação logito.

Logo, considerando $Y_i | x_i \sim \text{bin}(m_i, \pi_i)$, $i = 1, \dots, n$ independentes, então o modelo de regressão é dado por:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

E na escala da odds temos:

$$\frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$$

Regressão Logística

A regressão logística pode ser linearizada:

$$\eta = x' \beta$$

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

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

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

Regressão Logística – Analisando o erro

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

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

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

Logo os **erros não podem ser normais.**

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

$$\sigma^2_{y_i} = E(y_i)[1 - E(y_i)]$$

Estimação de Parâmetros

A forma geral de um modelo de regressão logística é

$$y_i = E(y_i) + \varepsilon_i$$

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

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

Estimação de Parâmetros

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

$$f_i(y_i) = p_i^{y_i} [1 - E(p_i)]^{1 - y_i}, \quad i = 1, 2, 3, \dots, n$$

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

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



PUC Minas
Virtual