MÓDULO III

Regressão linear simples

Alexandre Loureiros Rodrigues
Professor do departamento de Estatística - UFES

ESPECIALIZAÇÃO
INTELIGÊNCIA ARTIFICIAL
& CIÊNCIA DE DADOS

SEAD UFES | Superintendência de Educação a Distância

ÍNDICE

1. Modelos de regressão

Introdução

- A regressão é uma técnica usada para modelar a relação entre uma variável dependente (resposta) e uma ou mais variáveis independentes (característica ou features).

- Predizer valores contínuos com base em dados históricos

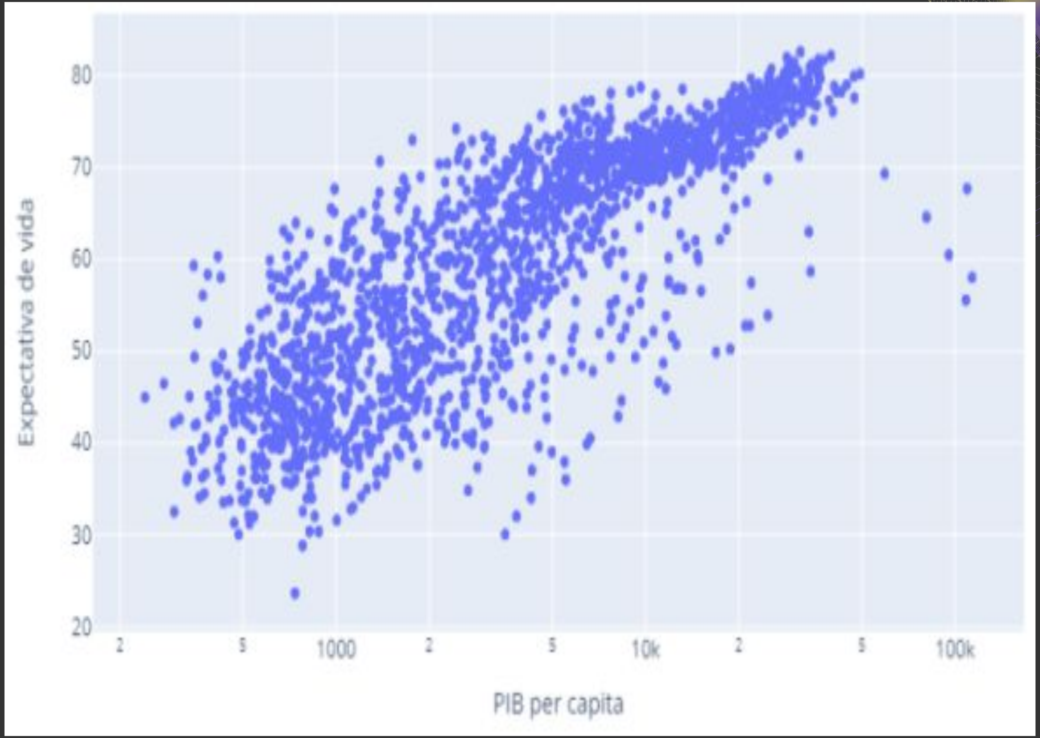
- Possivelmente o problema de aprendizado de máquinas mais frequente em nas áreas aplicadas.
    - Exemplo: Número ligações clandestinas por trafos em Cariacica

Introdução

Exemplo : Número ligações clandestinas em Cariacica

Introdução

Exemplo: Expectativa de vida VS PIB
Já vimos que a correlação linear mede a relação entre duas variáveis quantitativas. Como podemos ir além ?



2. Formulação Matemática

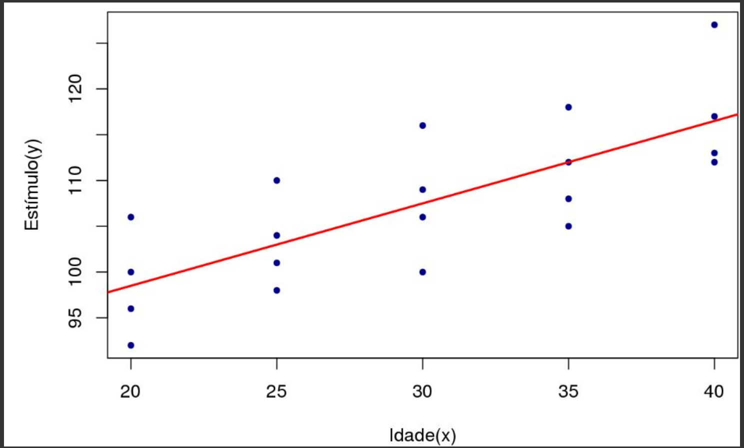
Formulação matemática

- Conjunto de treinamento $(X_1, Y_1),\ldots,(X_n, Y_n)$ - amostras independentes
- Desejamos criar uma função f que associa cada valor da característica X a um valor variável resposta Y.
  - $\hat{Y} = f(X)$
- A função f pode ser bem geral e muitas vezes nem é possível escrevê-la analiticamente.

Vamos começar de forma simples !
Usar função linear

Formulação matemática

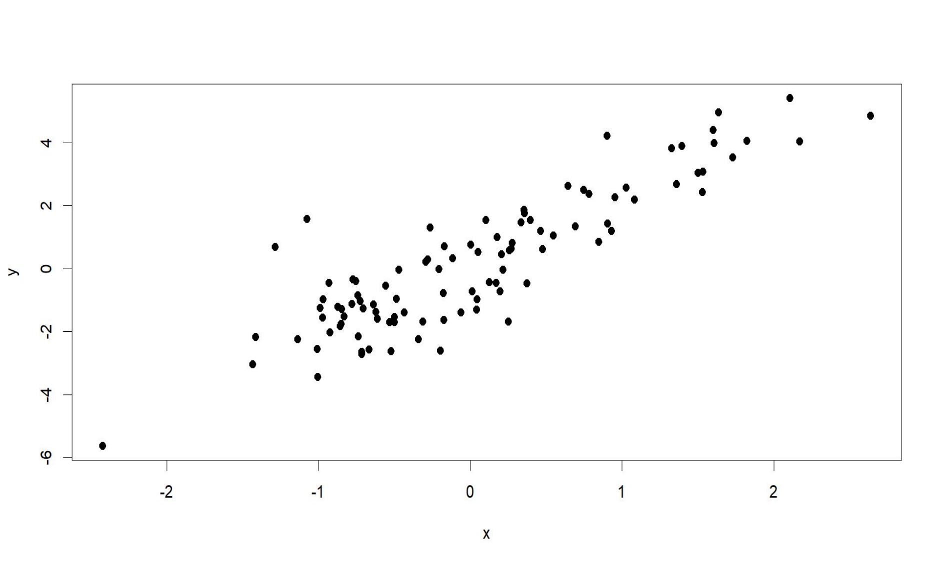
- Regressão linear simples
    - $\hat{Y} = \theta_1 + \theta_2 X$
- Equação de uma reta com intercepto θ_1 e coeficiente angular θ_2

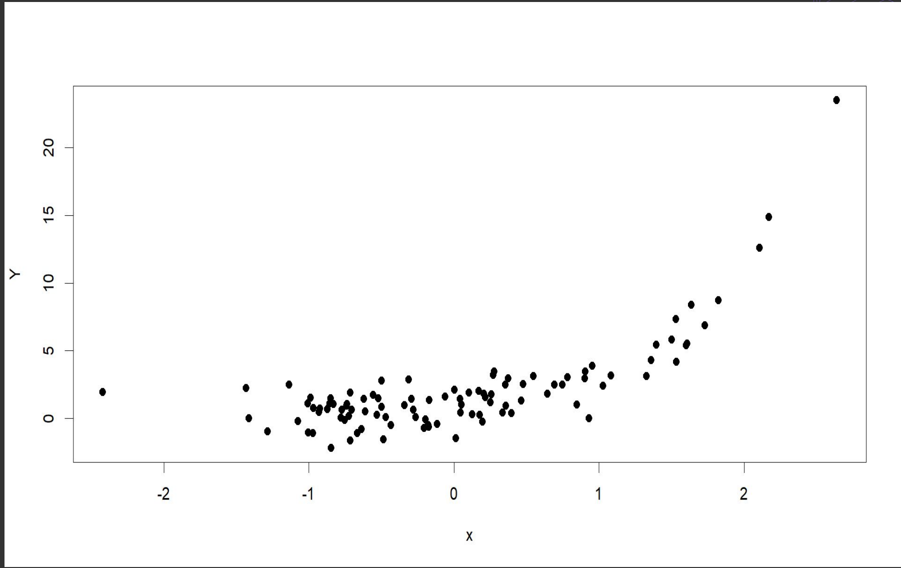


- Como saber quem são θ_1 e θ_2?
- Como saber se o modelo linear é uma boa abordagem para o problema de regressão?

Formulação matemática

- Regressão linear simples é adequada ?
  - Diagrama de dispersão





3. Ajuste do modelo

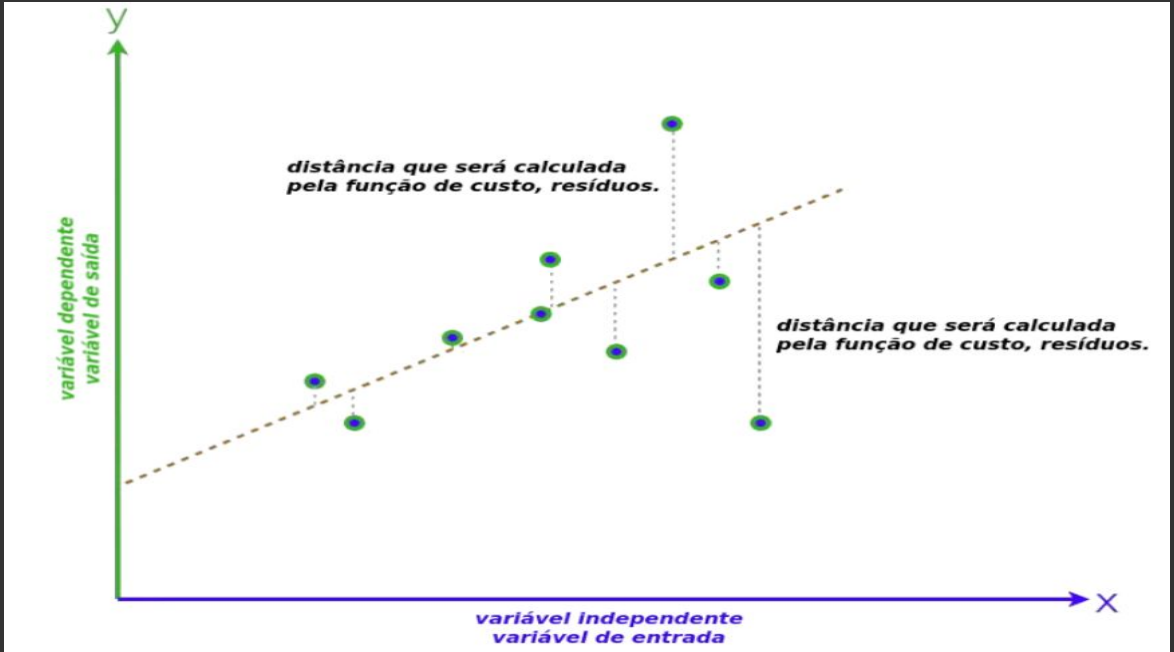
Ajuste do modelo - Função perda

- Queremos um modelo que se ajuste bem aos dados de treinamento
  - $\hat{Y}_i \approx Y_i,\ i = 1, \ldots, n$
  - Medimos a proximidade de $\hat{Y}_i$ e Y_i usando uma função perda:
    - Perda quadrática: $L\left(\hat{Y}_i, Y_i\right) = \left(\hat{Y}_i - Y_i\right)^2$
    - Outras funções perda: minmax, absoluta, etc.

- **A função perda para todo dado de treinamento é dada por:**

$$\sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2$$

Ajuste do modelo - Visão geométrica



Ajuste do modelo - Estimação

- Objetivo é encontrar θ_1 e θ_2 que minimizem a função perda geral

$$\sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2$$

$$= \sum_{i=1}^{n} [(\theta_1 + \theta_2 X_i) - Y_i]^2$$

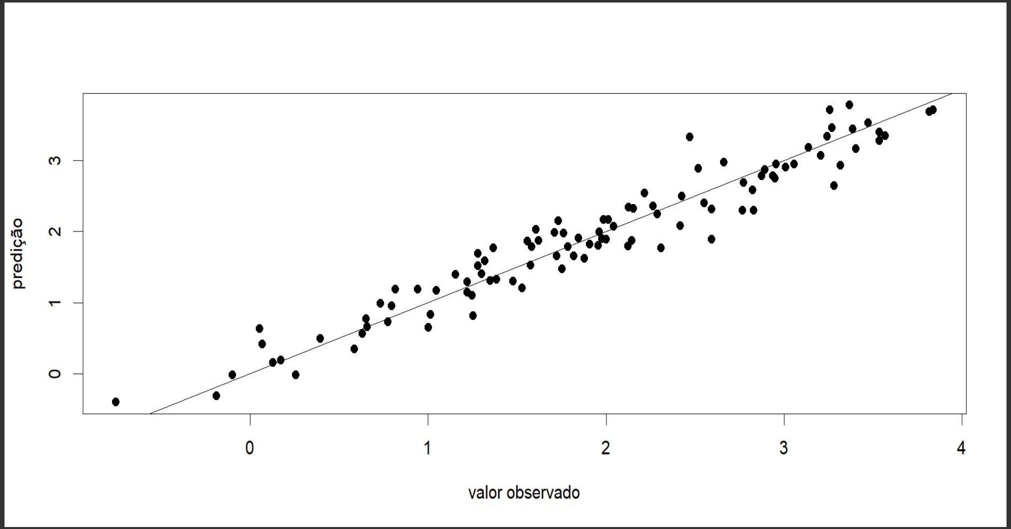
$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\theta}_1 = \bar{Y} - \hat{\theta}_2 \bar{X}$$

- No caso da perda quadrática, chamamos esta procedimento de estimação de método dos mínimos quadrados (MMQ).

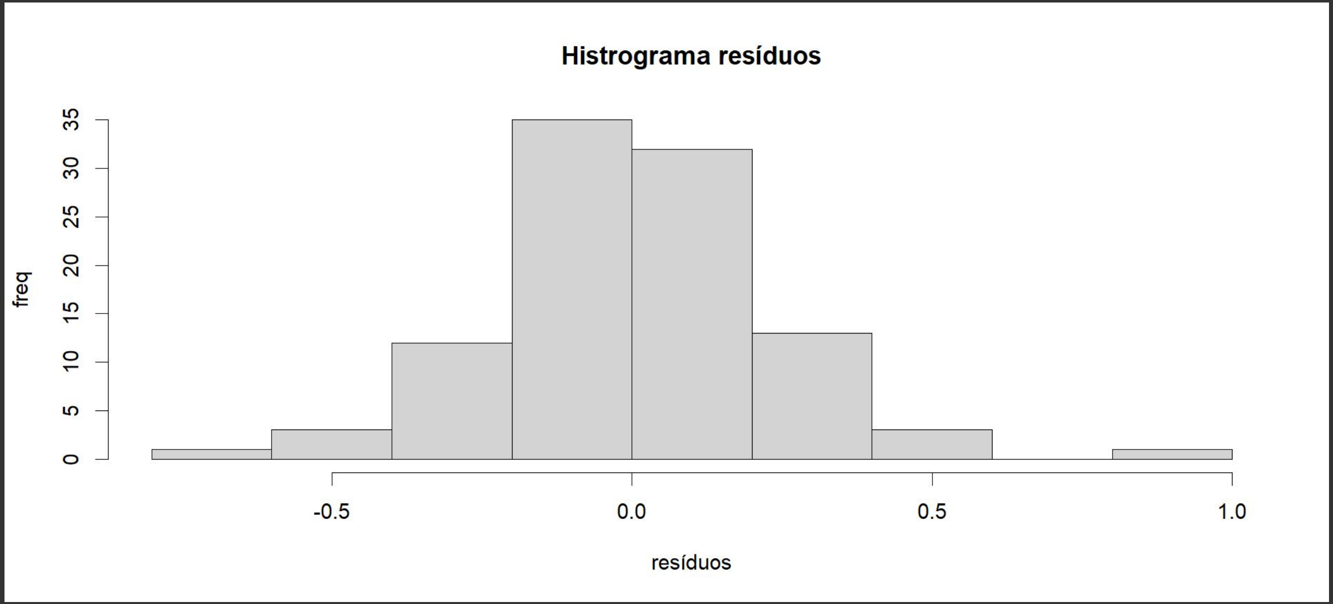
Avaliação de performance

- Diagrama de dispersão entre as predições e valores observados
  - $\hat{Y}_i$ vs Y_i



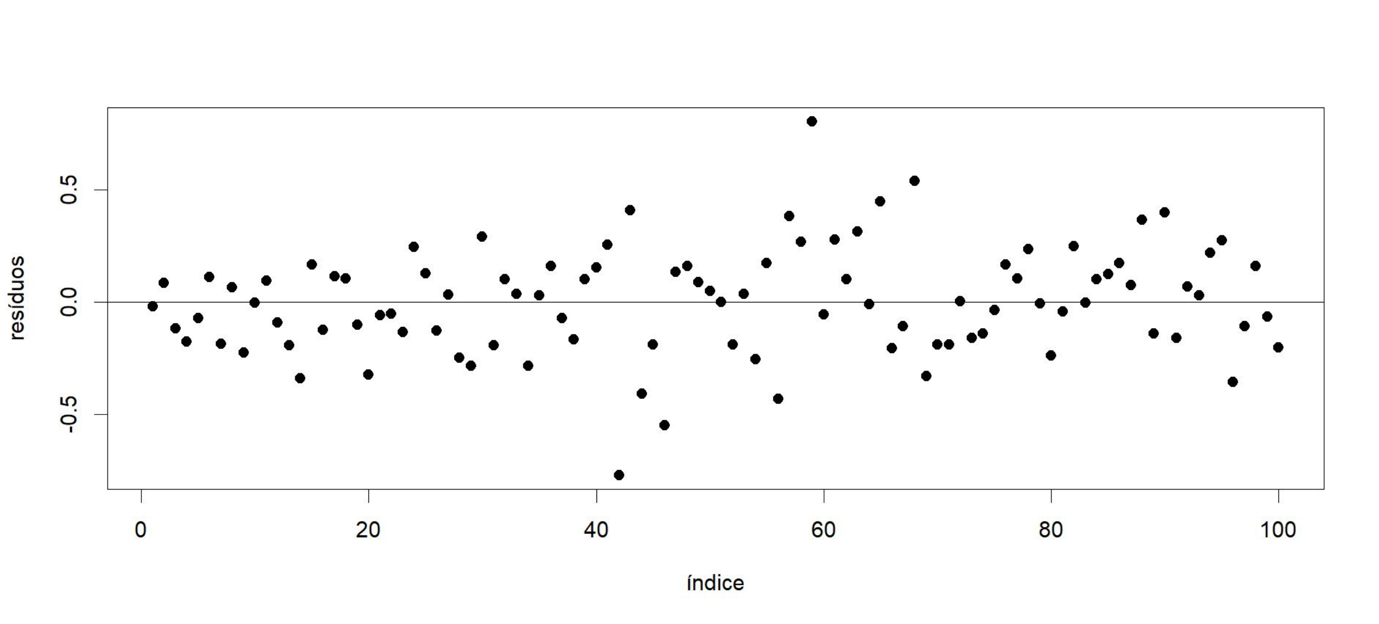
Desvios da reta podem indicar problemas no modelo

Avaliação de performance



$$e_i = \hat{Y}_i - Y_i$$

Avaliação de performance



Métricas de avaliação

- Objetivo de quantificar a avaliação do modelo

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100$$

Exemplo

Y = ALTURA

X = PESO

Z = PÉ (CM)

$\hat{Y}$ = a + bX

$\hat{Y}$ = u + vZ