Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Regressão Linear

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

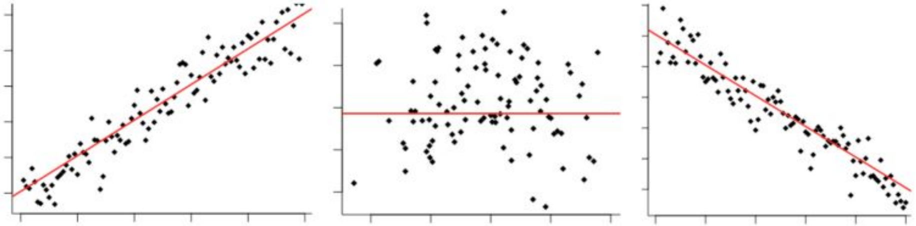
- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601

> git pull
```

Correlação

Conceito



$r > 0$ $r = 0$ $r < 0$

Medida de associação entre variéveis;

“Associação não implica na causa!” (ex.: Clima, Sorvete e Mosquito);

$$r = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x s_y}$$

Correlação

Conceito

Medidas não paramétricas de corelação:

- Spearman;
- Kendall

Intervalo de valores → -1 a 1

Função no R:

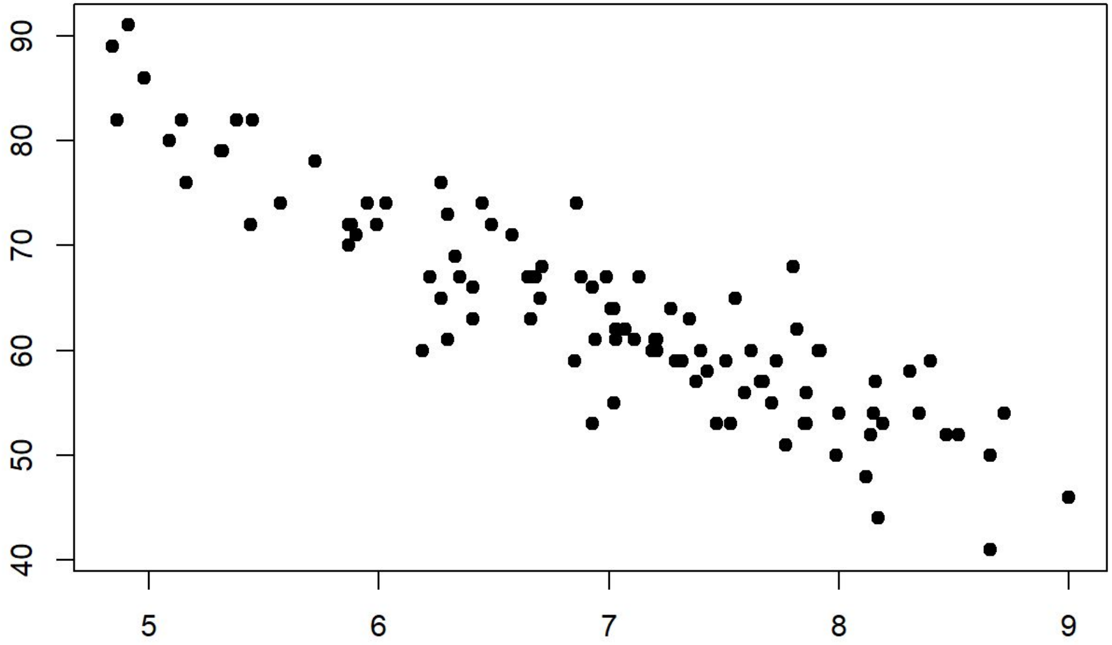
- cor();
- cor.test();

Regressão Linear

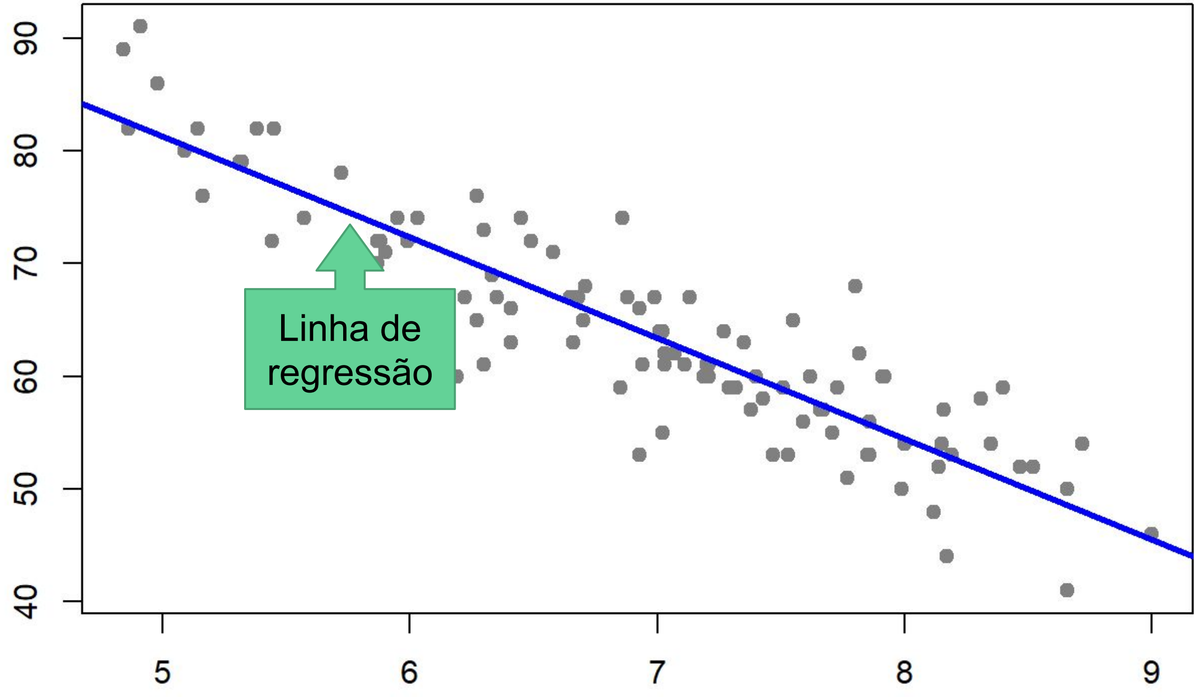
Conceito

"Um modelo estatístico que procura estabelecer uma relação linear entre a variável preditora (X) e a variável resposta (Y)."

Regressão Linear

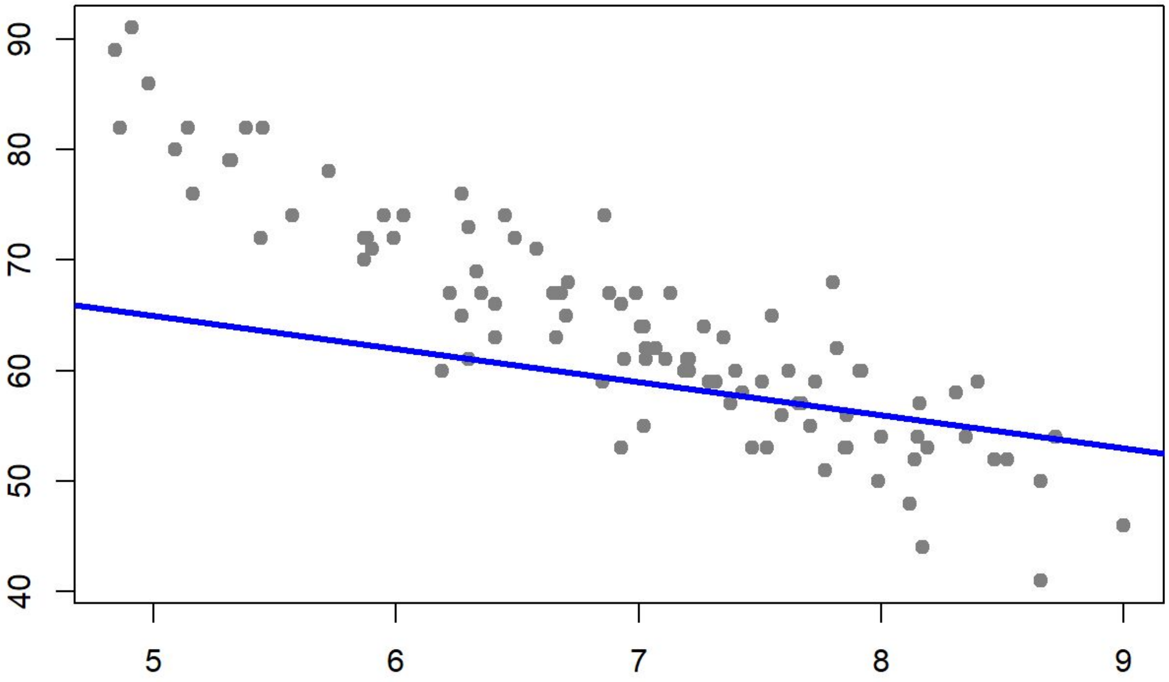


Regressão Linear



Modelo de regressão que melhor se ajusta aos dados

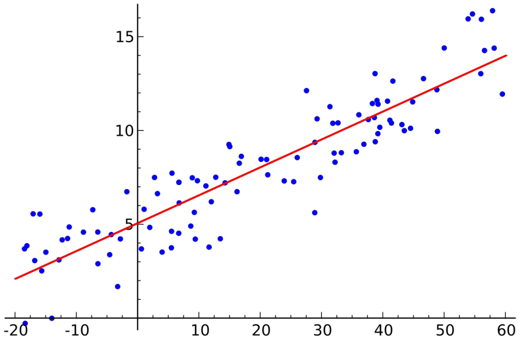
Regressão Linear



Modelo de regressão que não se ajusta bem com os dados

Regressão Linear

Uma linha



Média de Y para qualquer valor de X.

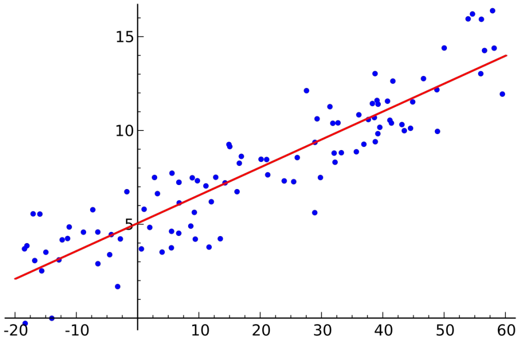
Fórmula para descrever uma função linear:

$$y = mx + c$$

- y e x ➜ duas variáveis;
- m ➜ slope, coeficiente linear;
- c ➜ onde a linha corta o y;

Regressão Linear

Uma linha



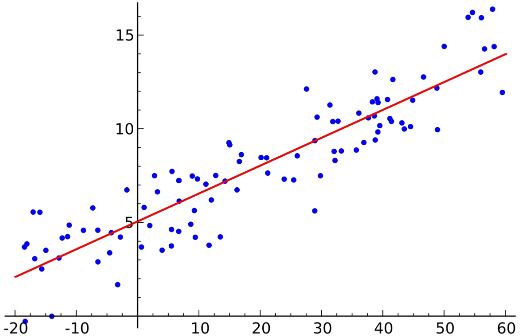
Se Y é a variável resposta e X é a variável preditora, então:

$$\hat{Y}_i = b_1 X_i + b_0$$

- X_i e Y_i → correspondem às coordenadas da i-ésima observação;
- $\hat{Y}_i$ → resultado estimado;
- b_1 → slope, coeficiente linear;
- b_0 → onde a linha corta o y;

Regressão Linear

Os dados não ajustam perfeitamente ao modelo



$$Y_i \neq \hat{Y}_i$$

Resíduo:

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

Se:

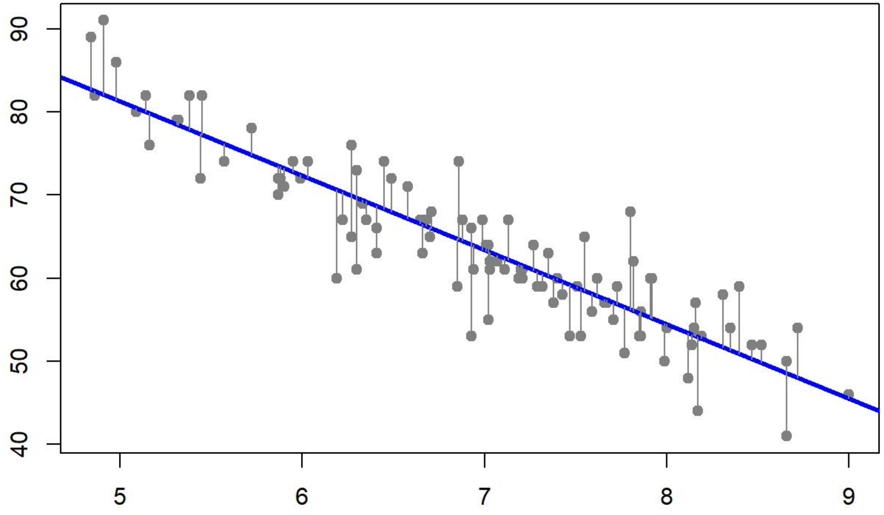
$$\hat{Y}_i = b_1 X_i + b_0$$

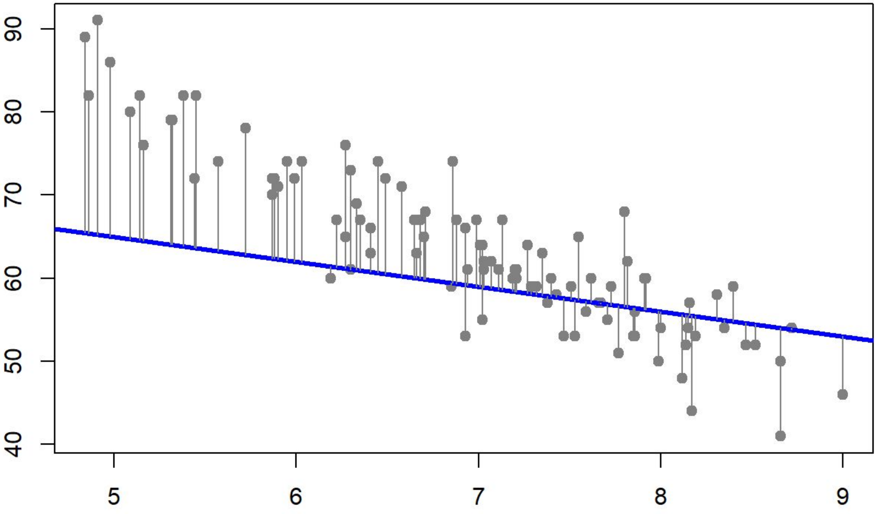
Modelo de regressão linear completo:

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

Regressão Linear

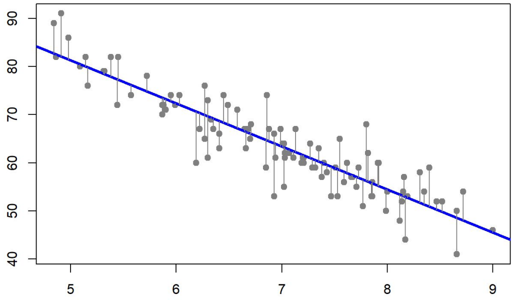
Comparação dos resíduos nos dois modelos lineares





Regressão Linear

Problema do modelo de regressão linear

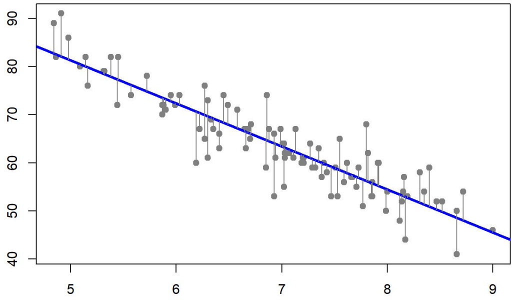


Encontrar uma estimativa de b_1 e b_0 que minimiza a soma do quadrado do resíduo:

$$\sum (Y_i - \hat{Y}_i)^2 = \sum \epsilon_i^2$$

Regressão Linear

Como estimar b_1 e b_0?



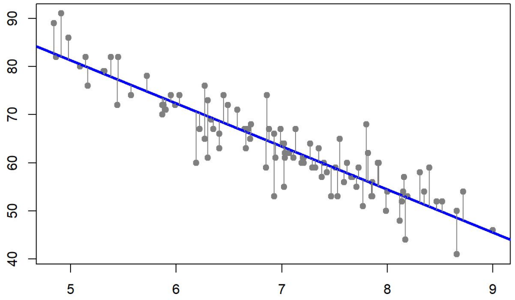
Método dos mínimos quadrados

- Encontra o melhor ajuste para um conjunto de dados minimizando a soma dos quadrados dos erros.

$$\sum (Y_i - \hat{Y}_i)^2 = \sum \epsilon_i^2$$

Regressão Linear

Mínimos quadrados



- Derivar a função:

$$\sum(Y_i - \hat{Y}_i)^2 = \sum \epsilon_i^2$$

- http://mathworld.wolfram.com/LeastSquaresFitting.html

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{cov(X,Y)}{var(X)}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$