

Disciplina:

**TÉCNICAS DE AMOSTRAGEM E
REGRESSÃO LINEAR**

Professora: Anaíle Mendes Rabelo



Amostragem Aplicada à IA e Ciência de Dados



Por que o tamanho da base importa?

O tamanho ideal de uma base de dados **não depende apenas** de fórmulas estatísticas tradicionais como margem de erro ou intervalo de confiança.

Em Ciência de Dados moderna, precisamos considerar múltiplos fatores que influenciam diretamente a qualidade e a eficácia dos nossos modelos.

01

Variabilidade dos dados

Quanto maior a dispersão, mais observações necessárias

02

Complexidade do modelo

Modelos sofisticados exigem mais dados para treinar

03

Número de variáveis

Cada feature adicional demanda mais registros

04

Representatividade

A amostra deve refletir fielmente a população

05

Qualidade dos dados

Dados limpos valem mais que volume bruto





Características de um Tamanho Ideal



Representatividade

Captura adequadamente as características e a diversidade da população-alvo



Estabilidade Estatística

Garante estimativas confiáveis com intervalos de confiança aceitáveis



Eficiência de Recursos

Otimiza tempo de coleta, armazenamento e processamento computacional



Minimização de Vieses

Evita redundâncias e garante distribuição equilibrada entre classes

Critérios Estatísticos Clássicos

As fórmulas tradicionais são o **ponto de partida** para calcular o tamanho amostral, mas não bastam para bases reais de Data Science.

Tipo de Variável	Medida Principal	Fórmula	Parâmetros
Quantitativa	Média (μ)	$n = (Z\sigma/E)^2$	σ (desvio padrão), E (erro)
Qualitativa	Proporção (p)	$n = Z^2 p(1-p)/E^2$	p (proporção), E (erro)

Exemplo Prático

Para estimar a renda média ($\sigma = R\$ 1.500$) com erro de R\$ 300 e 95% de confiança:

$$n = \left(\frac{1.96 \times 1500}{300} \right)^2 \approx 96 \text{ observações}$$

⚠️ Atenção: Essas fórmulas assumem amostragem aleatória simples e não consideram fatores como número de variáveis, complexidade do modelo ou desbalanceamento de classes.



Além das Fórmulas Clássicas

Em bases de dados reais para Ciência de Dados, **outros critérios são fundamentais** e muitas vezes mais importantes que as fórmulas estatísticas tradicionais.



Dimensionalidade

Número de variáveis/features no dataset



Complexidade

Tipo e sofisticação do modelo utilizado



Desbalanceamento

Distribuição desigual entre classes



Variabilidade

Dispersão e heterogeneidade dos dados

Dimensionalidade: O Desafio das Muitas Variáveis

A Maldição da Dimensionalidade

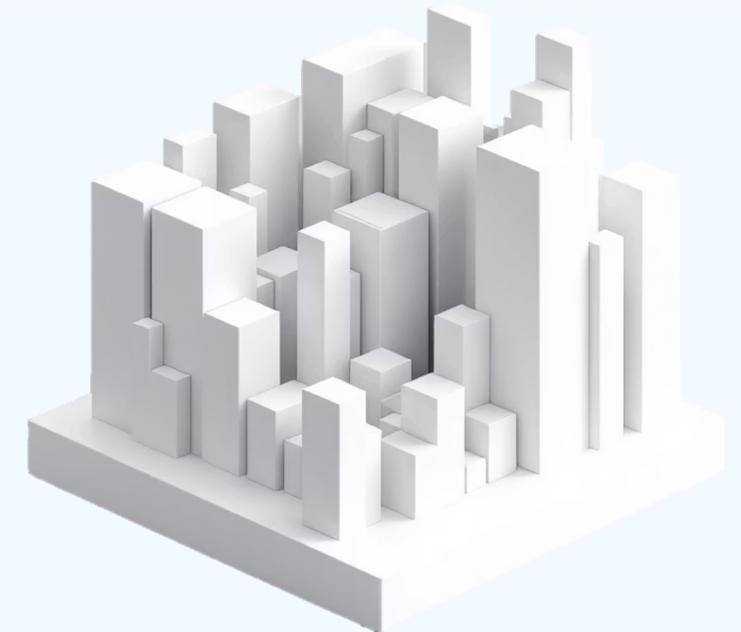
Quanto maior o número de variáveis (features), mais observações são necessárias para evitar overfitting e garantir que cada variável tenha informação suficiente para ser estimada com precisão.

Regra Prática Geral

$$n \geq 10 \text{ a } 30 \times \text{número de variáveis}$$

Exemplo Aplicado

- Dataset com **20 variáveis** → mínimo de 200 a 600 registros
- Dataset com **50 variáveis** → mínimo de 500 a 1.500 registros
- Dataset com **100 variáveis** → mínimo de 1.000 a 3.000 registros



💡 Dica: Em modelos de regressão múltipla, uma regra conservadora é ter pelo menos 15-20 observações por preditor para evitar instabilidade nas estimativas.

Complexidade do Modelo

Modelos mais sofisticados possuem mais parâmetros para estimar e, portanto, **exigem mais dados** para treinar adequadamente.



Modelos Lineares Simples

Regressão linear/logística: Requerem menos dados, geralmente suficiente com algumas centenas de observações.



Modelos Baseados em Árvores

Random Forest, XGBoost: Necessitam de milhares de observações para capturar padrões complexos sem overfitting.



Deep Learning

Redes neurais profundas: Exigem dezenas ou centenas de milhares de exemplos, especialmente para problemas de visão computacional.

1

Regressão Logística

≥ 50 eventos por variável preditora

2

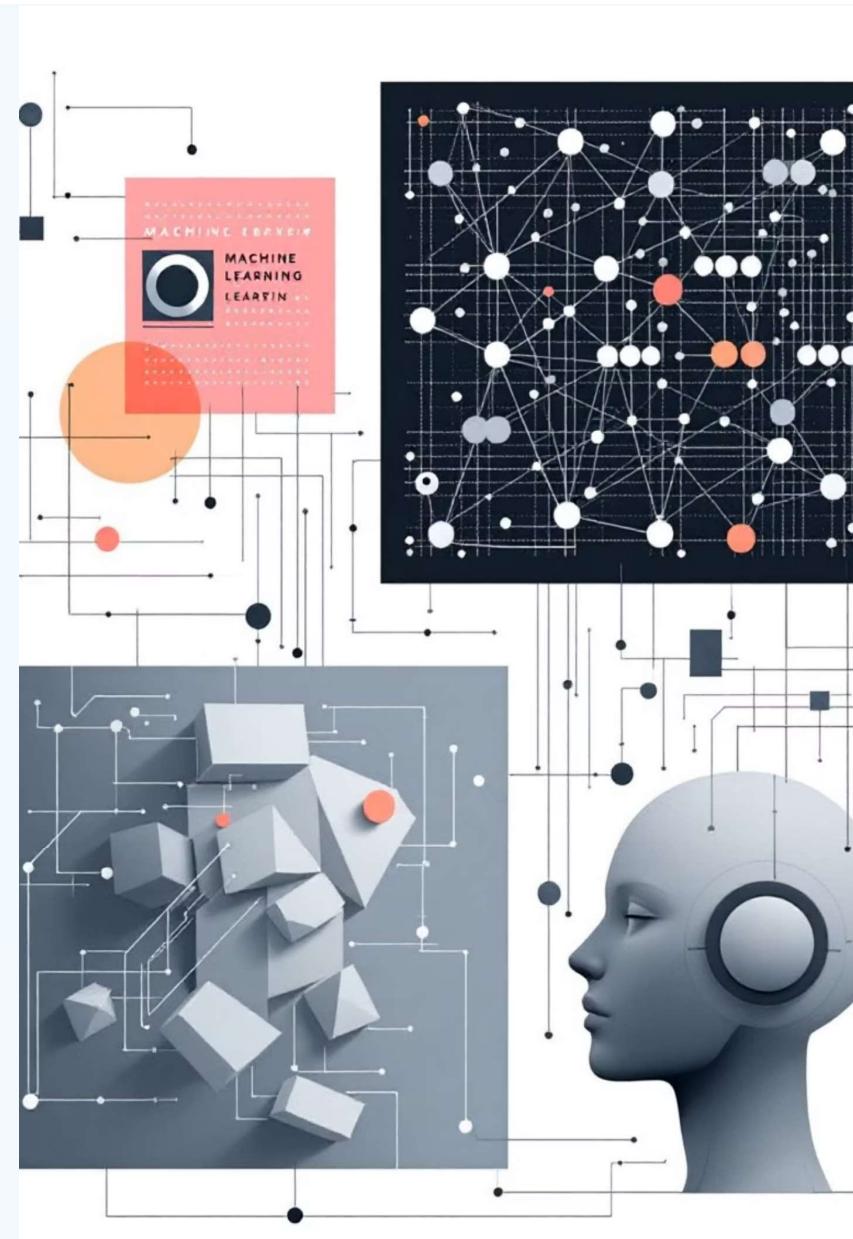
Random Forest

≥ 1.000 observações mínimo

3

Deep Learning

≥ 1.000 observações por classe (ideal: 10.000+)



Desbalanceamento de Classes

O Problema

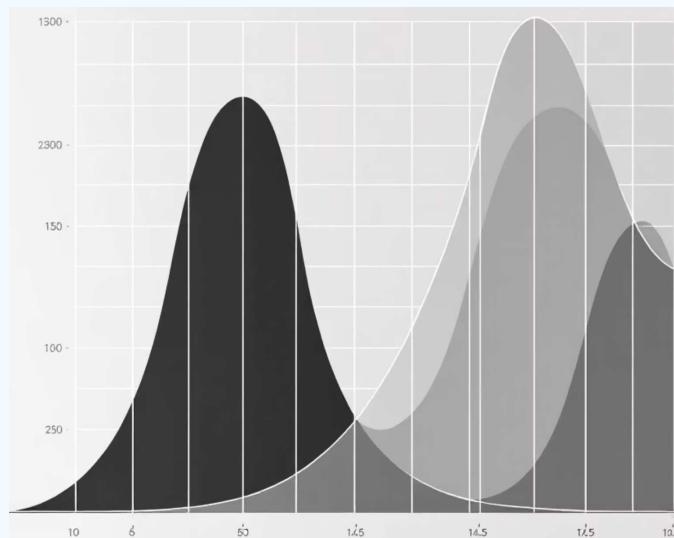
Quando temos distribuição desigual entre classes (ex: 90% de uma classe vs 10% de outra), o modelo pode ignorar completamente a classe minoritária.

Exemplo Real

Detecção de fraudes em cartões de crédito:

- 99% transações legítimas
- 1% transações fraudulentas

O modelo pode simplesmente classificar tudo como "não-fraude" e ter 99% de acurácia!



Variabilidade e Poder Estatístico

Alta Variabilidade = Maior Amostra

Quando os dados apresentam grande dispersão (alto desvio padrão σ), precisamos de **mais observações** para capturar adequadamente os padrões.

$$n \propto \sigma^2$$

Se o desvio padrão dobra, o tamanho amostral necessário **quadruplica**!

Guia de Decisão

Critério	Tipo	O que Observar	Regra Prática
Variabilidade	Quant/Qual	Alta variabilidade exige mais observações	Dobrar n se σ dobra
Erro tolerável	Ambos	Menor erro desejado requer mais dados	n cresce com $1/E^2$
Nº variáveis	Ambos	Cada variável consome informação	$n \geq 30 \times n_{\text{var}}$
Complexidade	Ambos	Modelos complexos precisam de mais exemplos	+1.000 por classe
Desbalanceamento	Qualit	Classe rara precisa de representação adequada	≥ 100 por classe
Qualidade	Ambos	Base limpa vale mais que base grande	Priorize qualidade

População Finita

Se a população é limitada, aplique o **fator de correção**: $n_{\text{ajustado}} = n/(1 + n/N)$

Dados Limpos > Dados Abundantes

Evite: valores faltantes, viés de seleção, duplicatas e outliers não tratados

Teste com Amostra Piloto

Antes de escalar, valide com uma amostra pequena para estimar parâmetros reais

"Em Ciência de Dados, a boa amostragem é o primeiro modelo bem treinado."

Introdução a Modelagem Estatística

MODELOS ESTATÍSTICOS



- **Objetivo:** Explorar o papel fundamental dos modelos estatísticos na análise de dados e na tomada de decisões;
- Estudar a relação entre as variáveis
- Os modelos podem fornecer insights valiosos e prever tendências, auxiliando em diversos contextos.

<https://www.linkedin.com/pulse/dados-intelig%C3%A1ncia-transforma%C3%A7%C3%A3o-digital-o-que-mais-castro-cip-/?originalSubdomain=pt>

OS 5 PRINCIPAIS MODELOS ESTATÍSTICOS

Regressão Linear

Análise de Variância (ANOVA)

Regressão Logística

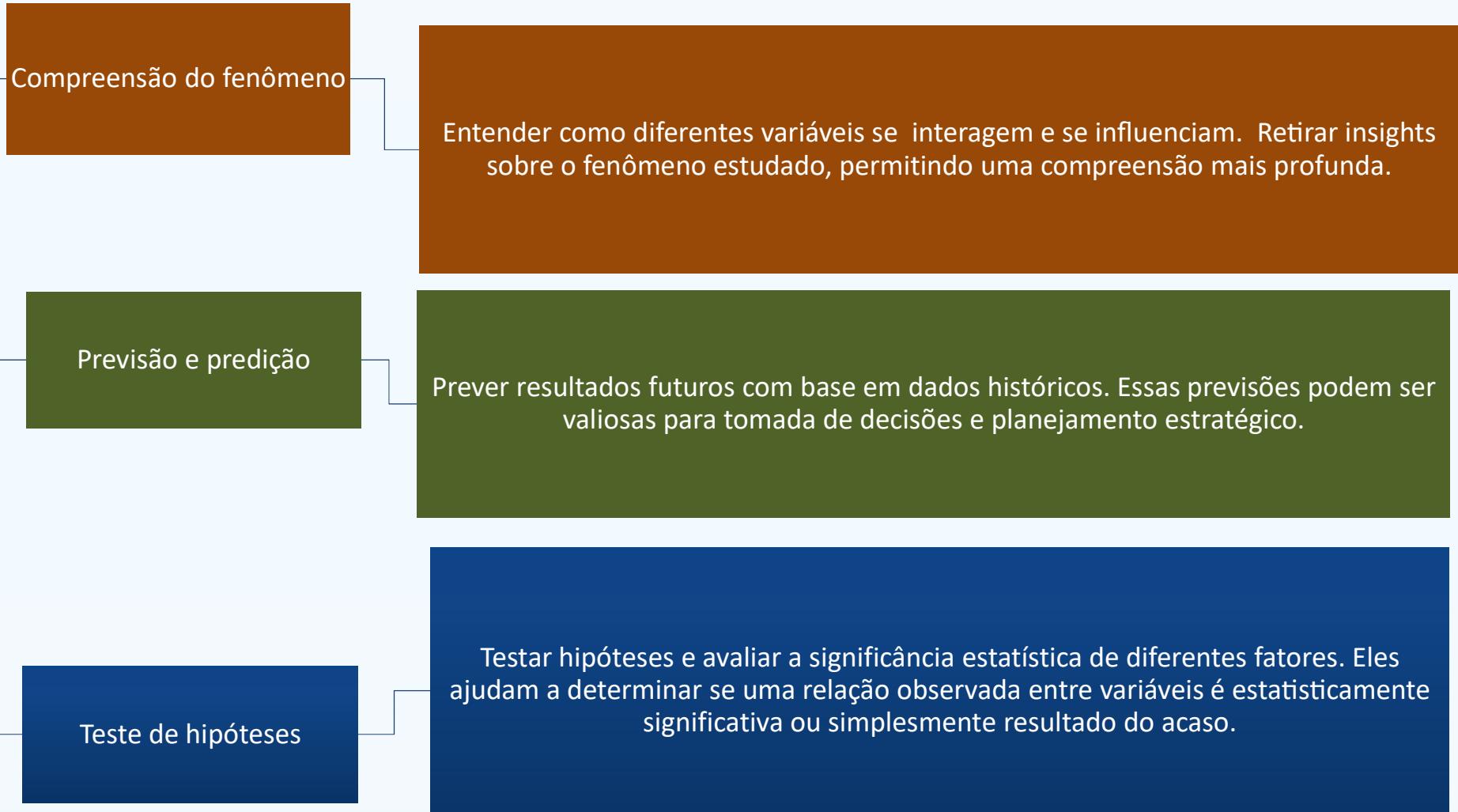
Análise de Sobrevida

Séries Temporais

O QUE É UM MODELO ESTATÍSTICO

- Representação simplificada e abstrata de um fenômeno ou sistema real que se baseia em princípios estatísticos e matemáticos.
- Descreve a relação entre variáveis e fornece uma estrutura para entender, analisar e prever dados.
- São amplamente utilizados em ciências de dados para compreender padrões, explorar relações e tomar decisões informadas com base em evidências quantitativas.

Porque modelar



OBJETIVOS DA MODELAGEM ESTATÍSTICA

Descrição

- Descrever e resumir os dados, identificando padrões, tendências e características importantes. Ajudam a comunicar informações complexas de forma concisa.

Inferência

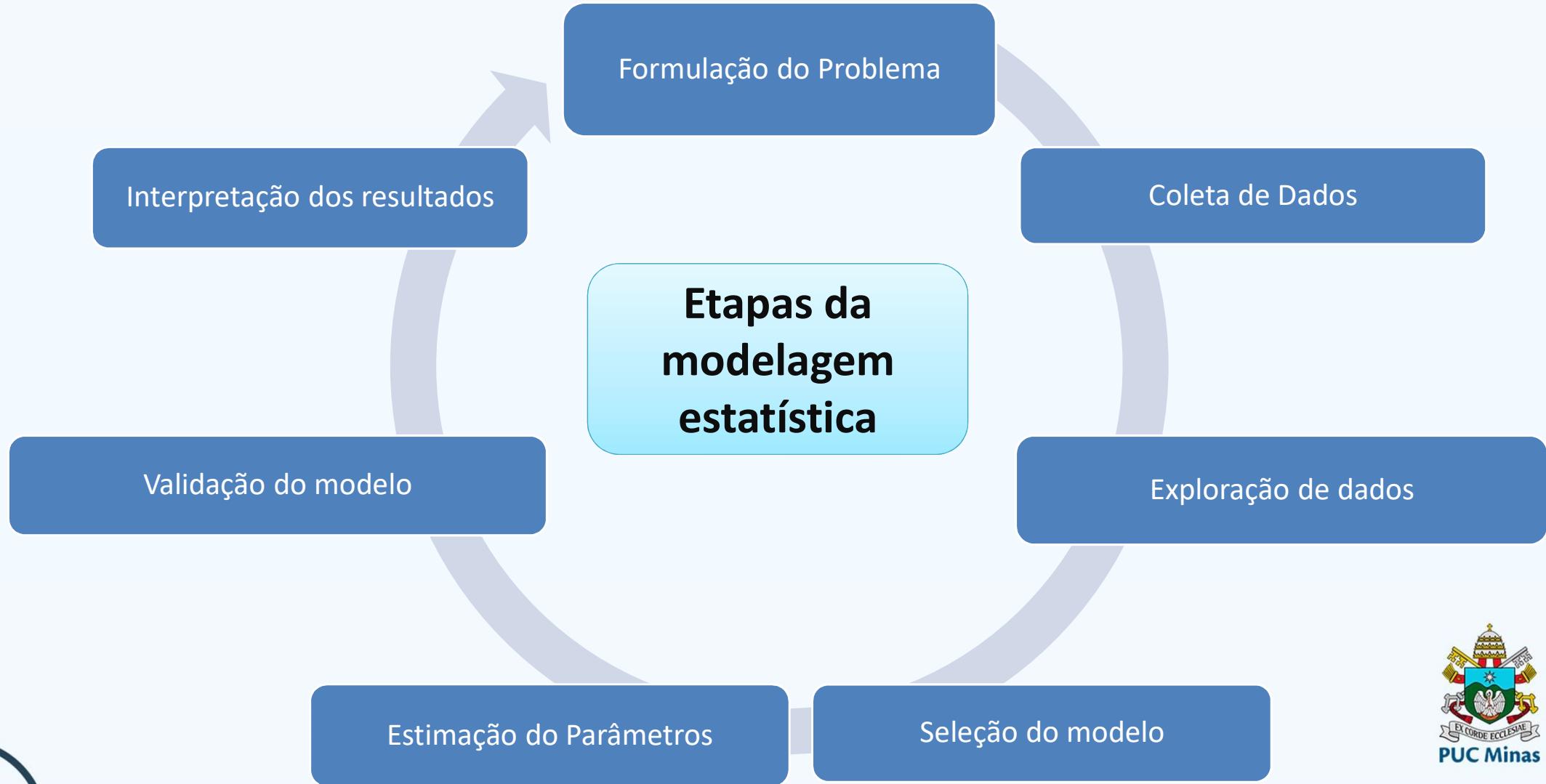
- Permitem fazer inferências sobre a população com base em uma amostra. Eles ajudam a generalizar conclusões e insights para além dos dados observados.

Previsão

- Prever eventos futuros com base em dados históricos. Fornecem uma base para tomar decisões informadas e antecipar resultados.

Controle e otimização

- Otimizar processos e controlar variáveis importantes. Eles ajudam a identificar os fatores-chave que influenciam um resultado específico e fornecem uma base para a melhoria contínua.



FUNDAMENTOS PARA MODELOS DE REGRESSAO LINEAR

FUNDAMENTOS PARA MODELOS DE REGRESSÃO



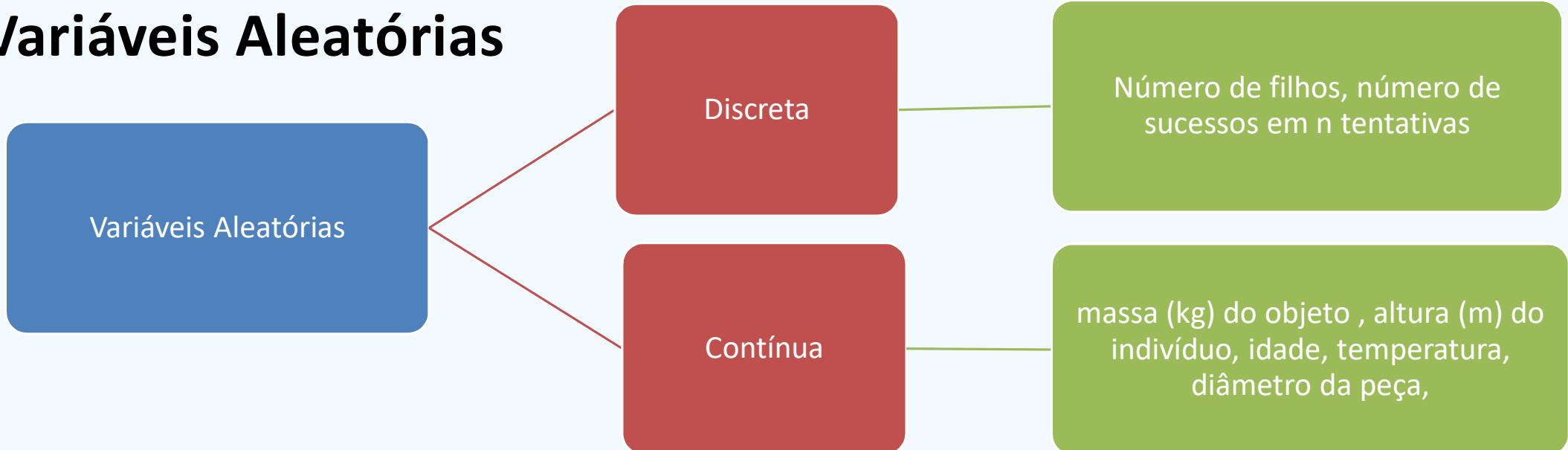
MEDIDAS RESUMO



MEDIDAS DE
ASSOCIAÇÃO

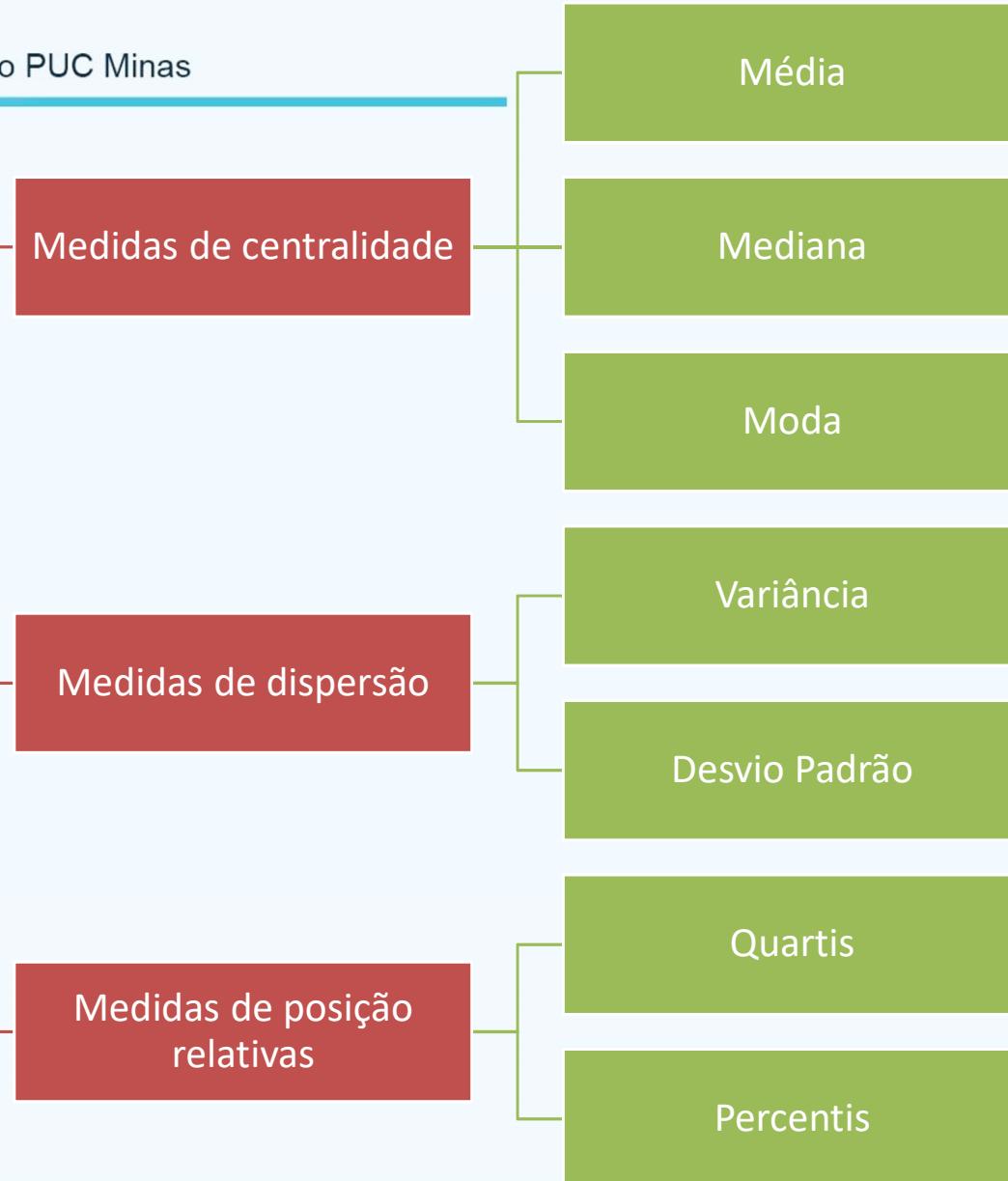


Classificação de Variáveis Aleatórias



Variáveis discretas → suporte em um conjunto de valores enumeráveis (finitos ou infinitos)

Variáveis contínuas → suporte em um conjunto não enumerável de valores



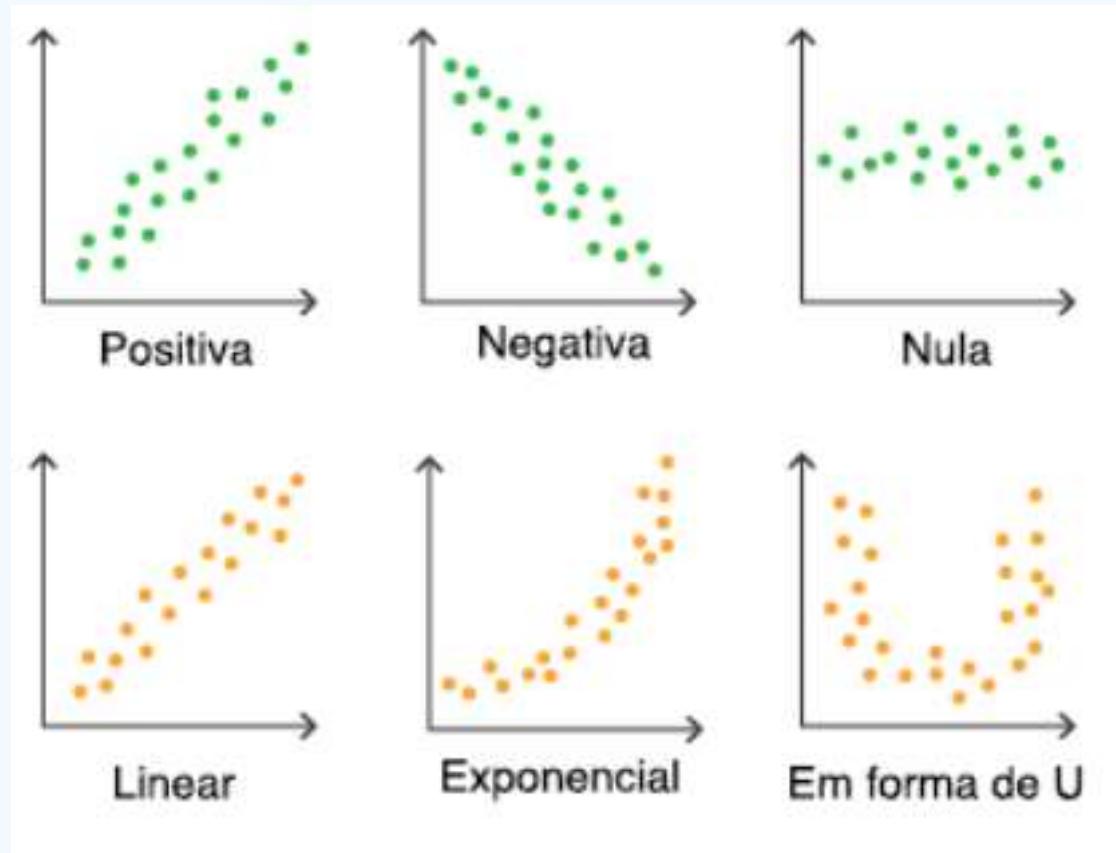
MEDIDAS RESUMO

Qual aplicabilidade
de cada medida?

MEDIDAS DE ASSOCIAÇÃO

Correlação linear

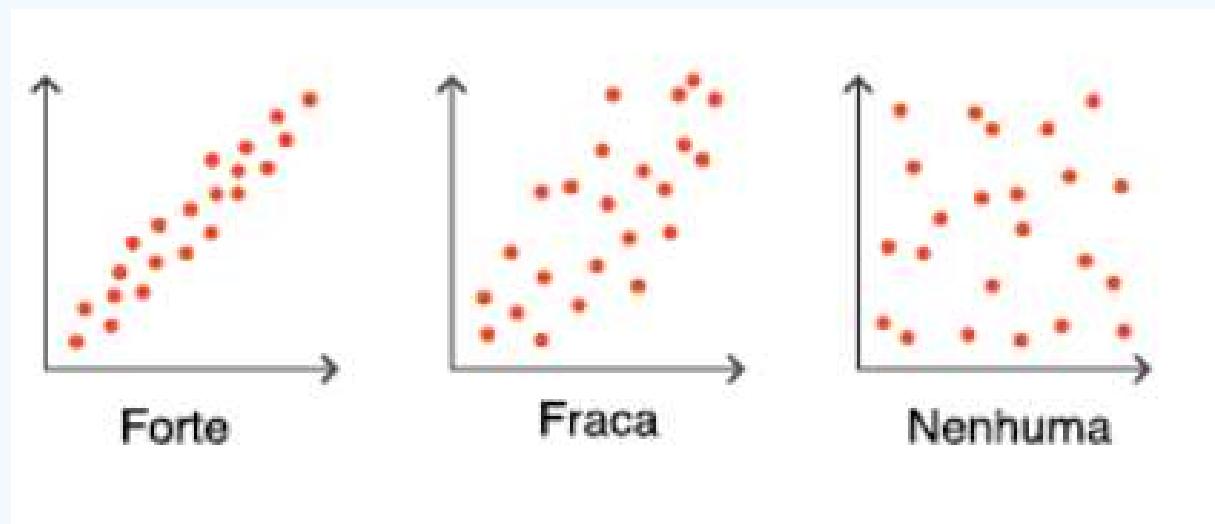
Determinado através de gráficos de dispersão e do coeficiente de variação



MEDIDAS DE ASSOCIAÇÃO

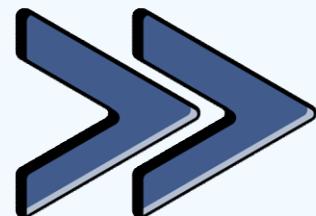
Correlação linear

Força da Correlação



CARACTERÍSTICAS DA CORRELAÇÃO

Pode ser um valor entre -1 e 1



Mostra a força e a direção
entre as variáveis



A correlação de $A \sim B$
é a mesma que $B \sim A$

COVARIÂNCIA E CORRELAÇÃO

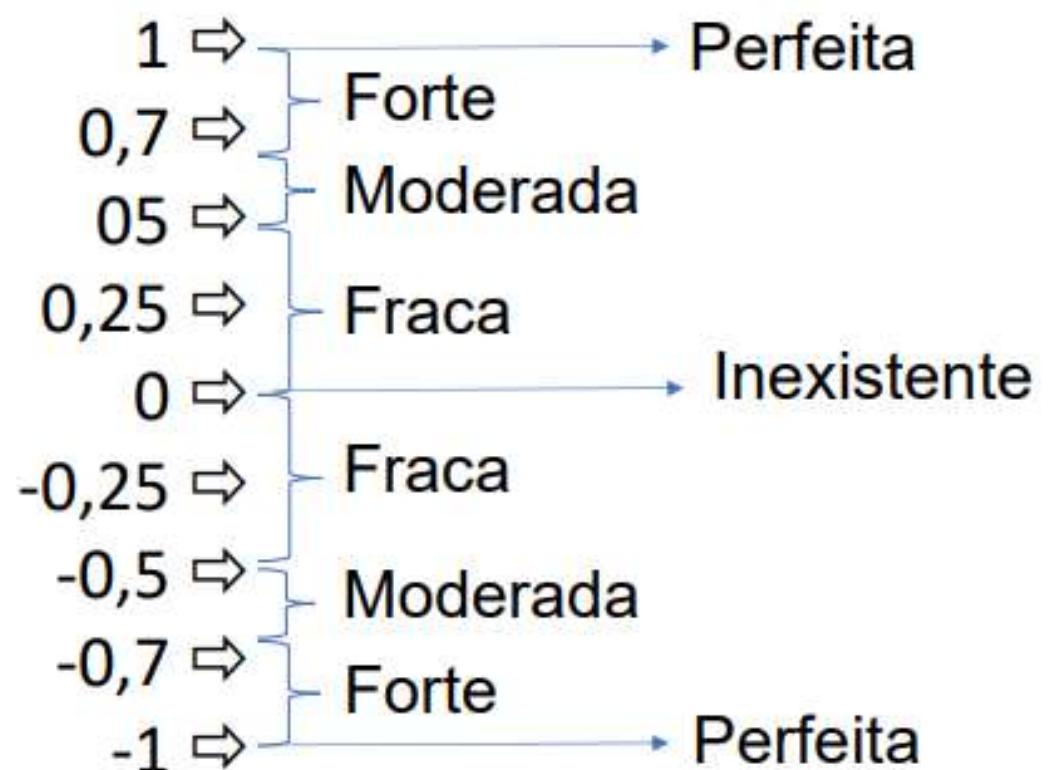
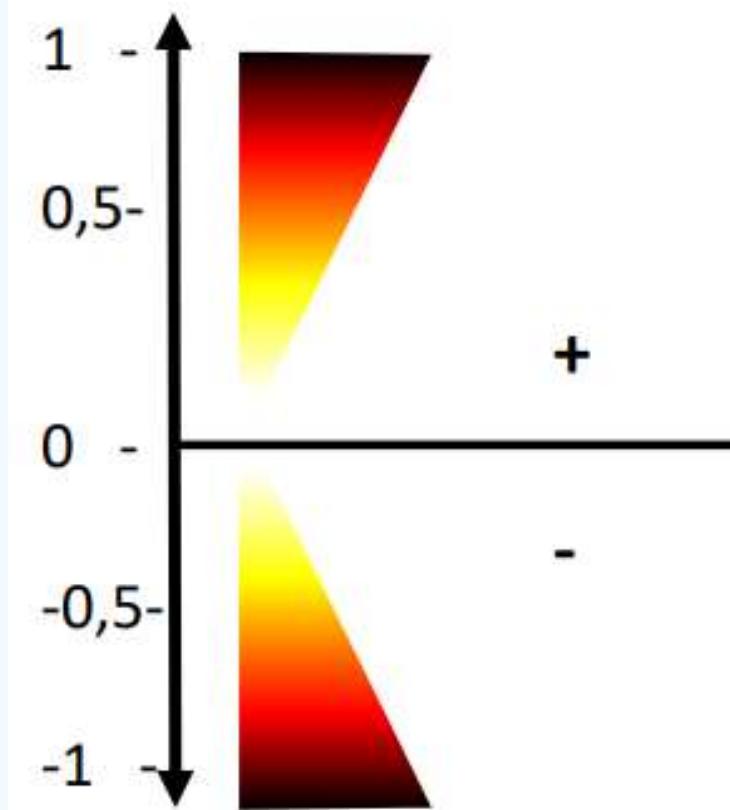
- A covariância amostral entre duas variáveis Y1 e Y2 é:

$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^n (y_{1i} - \bar{y}_1) \cdot (y_{2i} - \bar{y}_2)$$

- A correlação amostral entre duas variáveis Y1 e Y2 é (Coeficiente de Pearson):

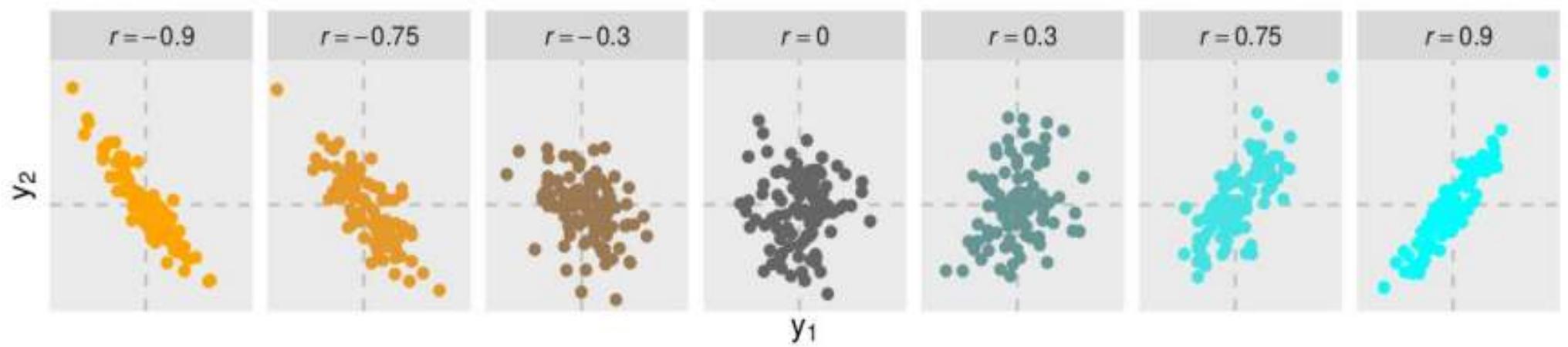
$$r = \frac{\sum_{i=1}^n (y_{1i} - \bar{y}_1) \cdot (y_{2i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^n (y_{1i} - \bar{y}_1)^2} \cdot \sqrt{\sum_{i=1}^n (y_{2i} - \bar{y}_2)^2}} = \frac{\text{Cov}(y_1, y_2)}{\sqrt{V(y_1) \cdot V(y_2)}}$$

FORÇA DA CORRELAÇÃO



É usado para determinar se existe relação linear entre variáveis aleatórias quantitativas. A correlação r assume valores entre -1 e 1.

- Quando $r > 0$, então existe uma associação (linear) positiva.
- Quando $r < 0$, então existe uma associação (linear) negativa.
- Quando $r = 0$, então não existe uma associação (linear).



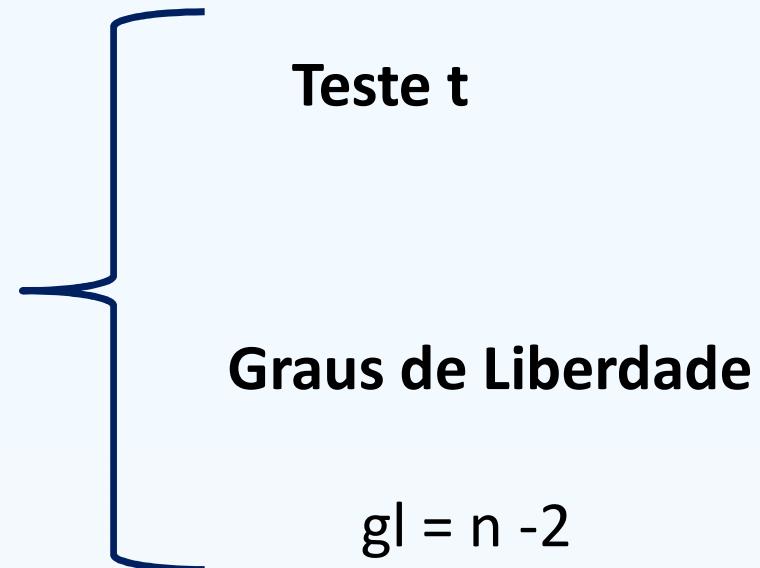
TESTE DE HIPÓTESE PARA A CORRELAÇÃO

Sejam as hipóteses nula e alternativa:

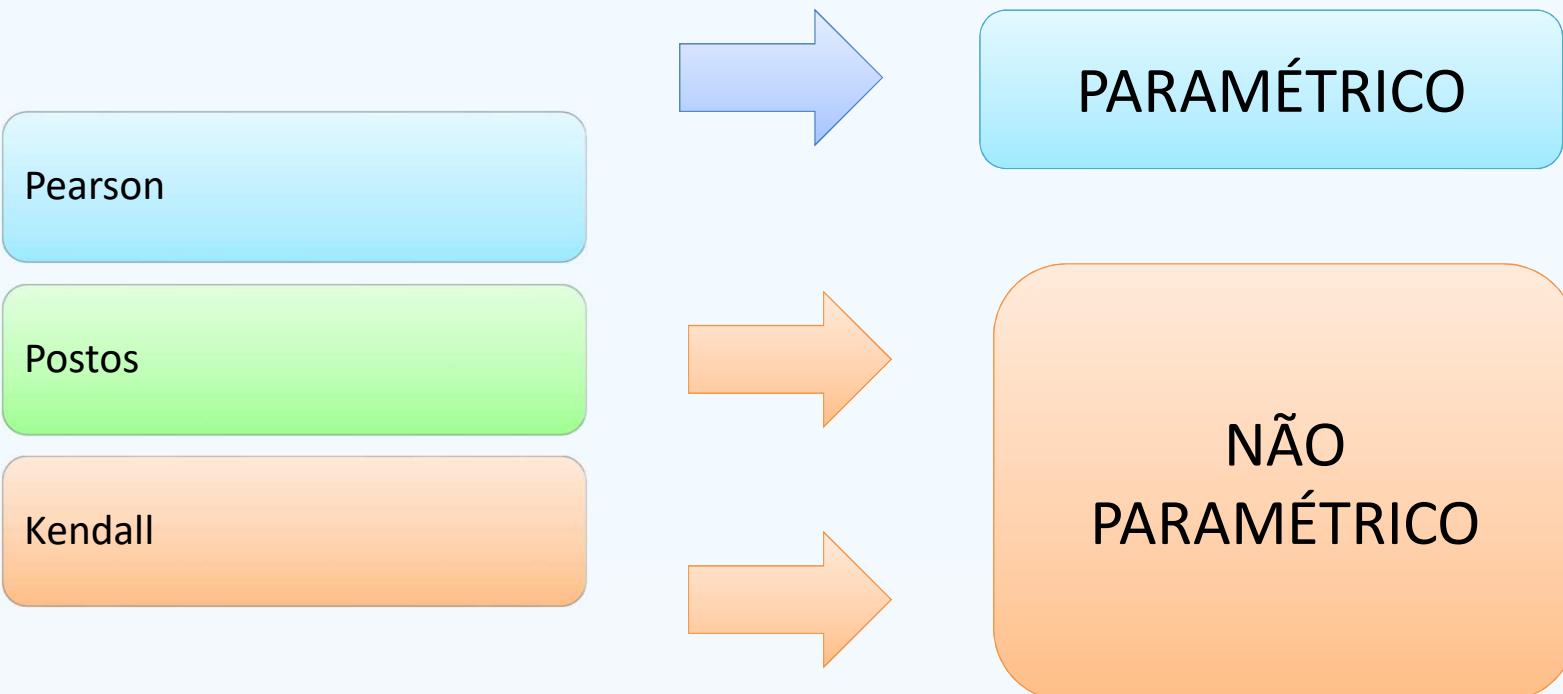
$$H_0 : r = 0.$$

$$H_a : r \neq 0$$

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$



TIPOS DE CORRELAÇÃO



COEFICIENTE DE PEARSON

- Forma mais precisa de medir a **correlação linear** entre duas grandezas
- Teste **paramétrico** -> temos que ter **variáveis normais**

$$r = \frac{\sum_{i=1}^n (y_{1i} - \bar{y}_1) \cdot (y_{2i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^n (y_{1i} - \bar{y}_1)^2} \cdot \sqrt{\sum_{i=1}^n (y_{2i} - \bar{y}_2)^2}} = \frac{Cov(y_1, y_2)}{\sqrt{V(y_1) \cdot V(y_2)}}$$

COEFICIENTE DE SPEARMAN – CORRELAÇÃO DE POSTOS

- *Teste **não paramétrico** -> pode ser aplicado em variáveis não normais*
- *Mede a força da relação entre duas variáveis (lineares ou não lineares)*
- *Utiliza os postos de entradas de amostras de dados pareados*
- *Pode ser utilizado para dados contínuos e ordinais*

$$r_R = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

n = nº de amostras
 d_i = diferença de alcance de cada elemento

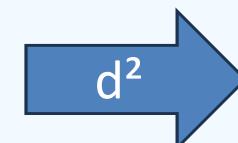
X	Y
83	82
75	92
75	54
73	70
72	88
62	64
60	80
58	62
54	62
52	69
51	83
48	79

as



$$r_R = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

X	Y
12	9
10,5	12
10,5	1
9	6
8	11
7	4
6	8
5	2,5
4	2,5
3	5
2	10
1	7



d²
(12-9) ²
(10,5 – 12) ²
(10,5-1) ²
(9-6) ²
(8-11) ²
(7-4) ²
(6-8) ²
(5-2,5) ²
(4-2,5) ²
(3-5) ²
(2-10) ²
(1-7) ²

CORRELAÇÃO DE KENDALL

- Teste **não paramétrico** -> quando temos amostras pequenas (<30)*
- Populações com grandes quantidade de empates (valores repetidos)*
- Pode ser utilizado juntamente com o spearman para comparação*
- Pode ser utilizado para dados contínuos e ordinais*

CORRELAÇÃO DE KENDALL

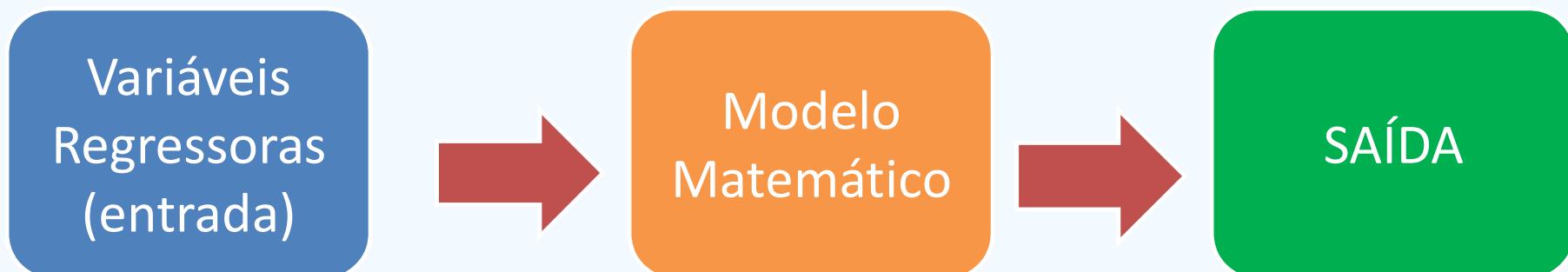
$x_i > x_j$ e $y_i > y_j$ ou se $x_i < x_j$ e $y_i < y_j$

$$\tau = \frac{(Qtd\ de\ pares\ concordantes) - (Qtd\ de\ pares\ discordantes)}{n(n - 1)/2}$$

$x_i > x_j$ e $y_i < y_j$ ou se $x_i < x_j$ e $y_i > y_j$

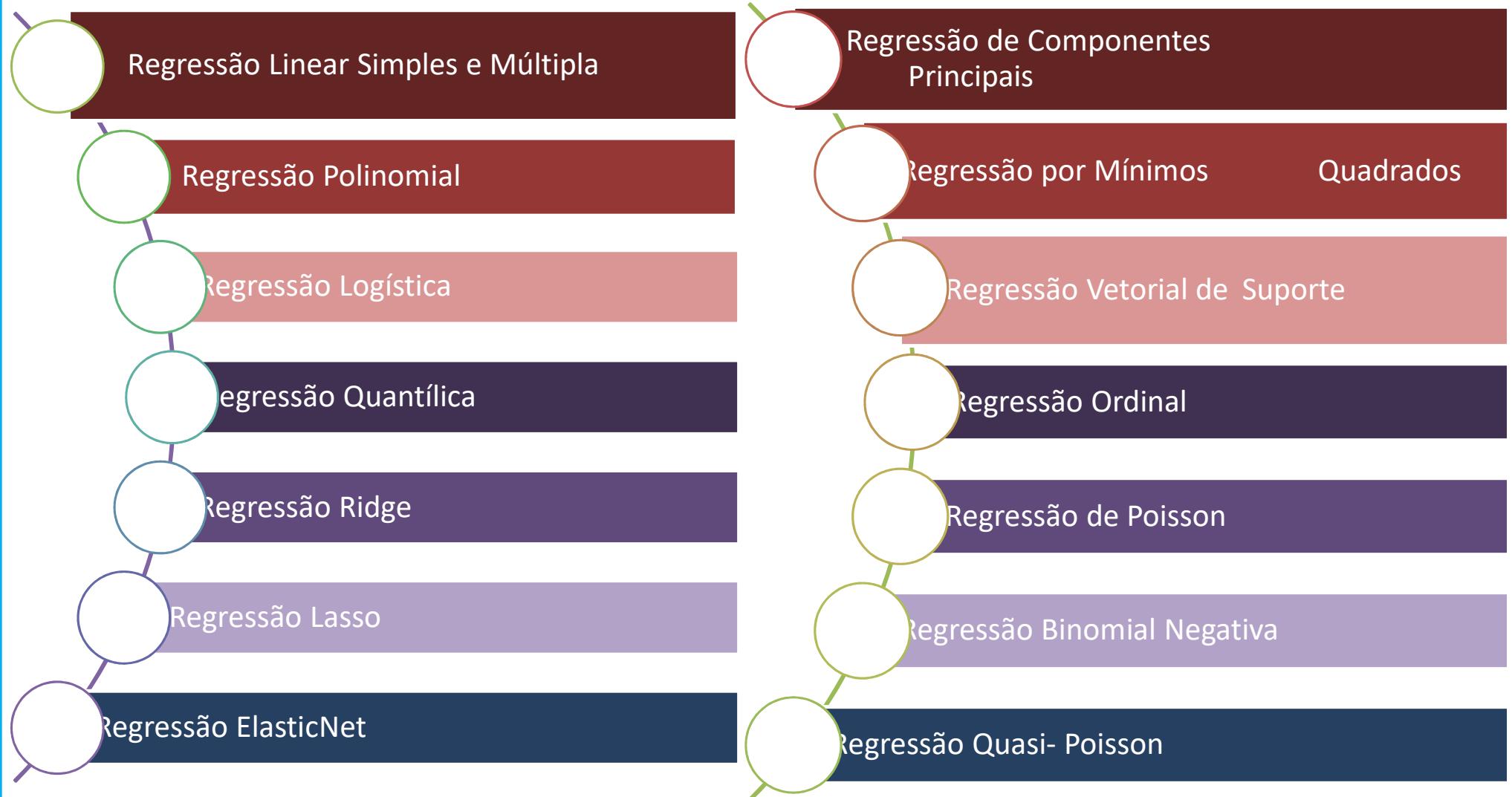
REGRESSAO LINEAR

MODELOS DE REGRESSÃO



P
R
I
N
C
I
P
A
I
S

M
O
D
E
L
O
S



ESCOLHENDO O TIPO DE REGRESSÃO

Outliers

- Valores discrepantes

Normalidade dos resíduos

- Distribuição simétrica com as medidas de centralidade tendendo a igualdade

Multicolinearidade

- Variáveis independentes altamente correlacionadas

Homocesticidade

- Homogeneidade de variância

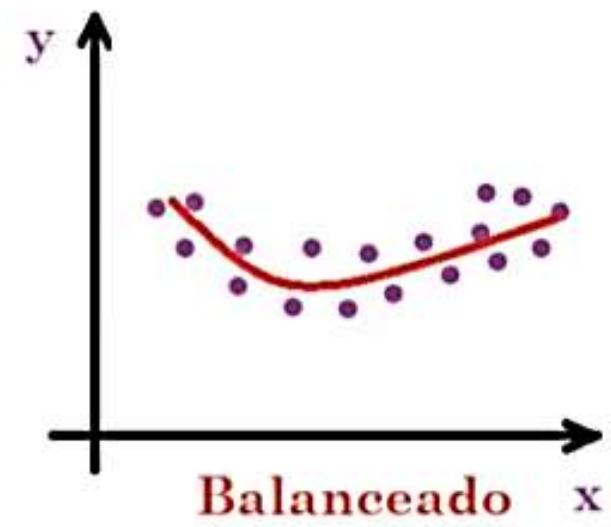
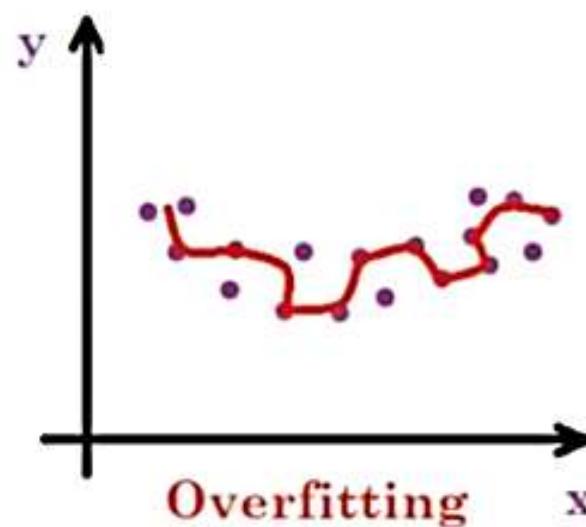
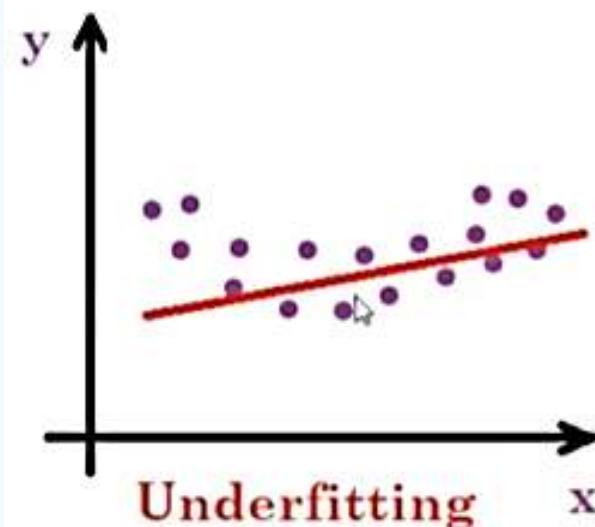
Underfitting

- Algoritmo não se encaixa com os dados de entrada

Overfitting

- Algoritmos ótimo para os dados de entrada mas ruim para teste

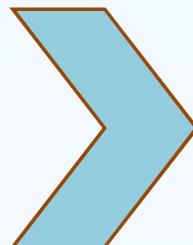
INTRODUÇÃO SOBRE AJUSTE DE MODELOS



REGRESSÃO LINEAR

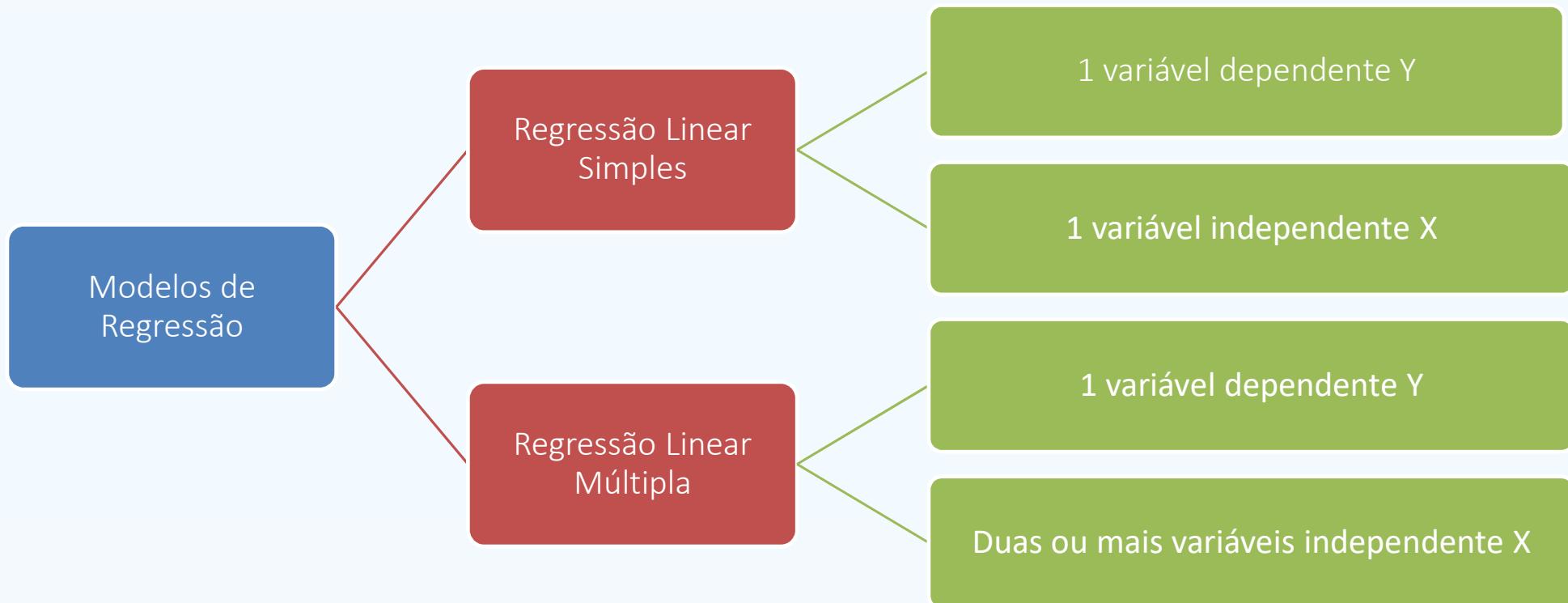
- Técnica estatística que visa modelar a relação entre **uma variável dependente (ou resposta)** e **uma ou mais variáveis independentes (ou preditoras)**.
- Utilizada para prever valores contínuos e entender a relação linear entre as variáveis.
- O termo "linear" se refere ao fato de que a **relação entre as variáveis é modelada através de uma linha reta**.

Uma variável independente x explica a variação em outra variável, que é chamada de variável dependente y.



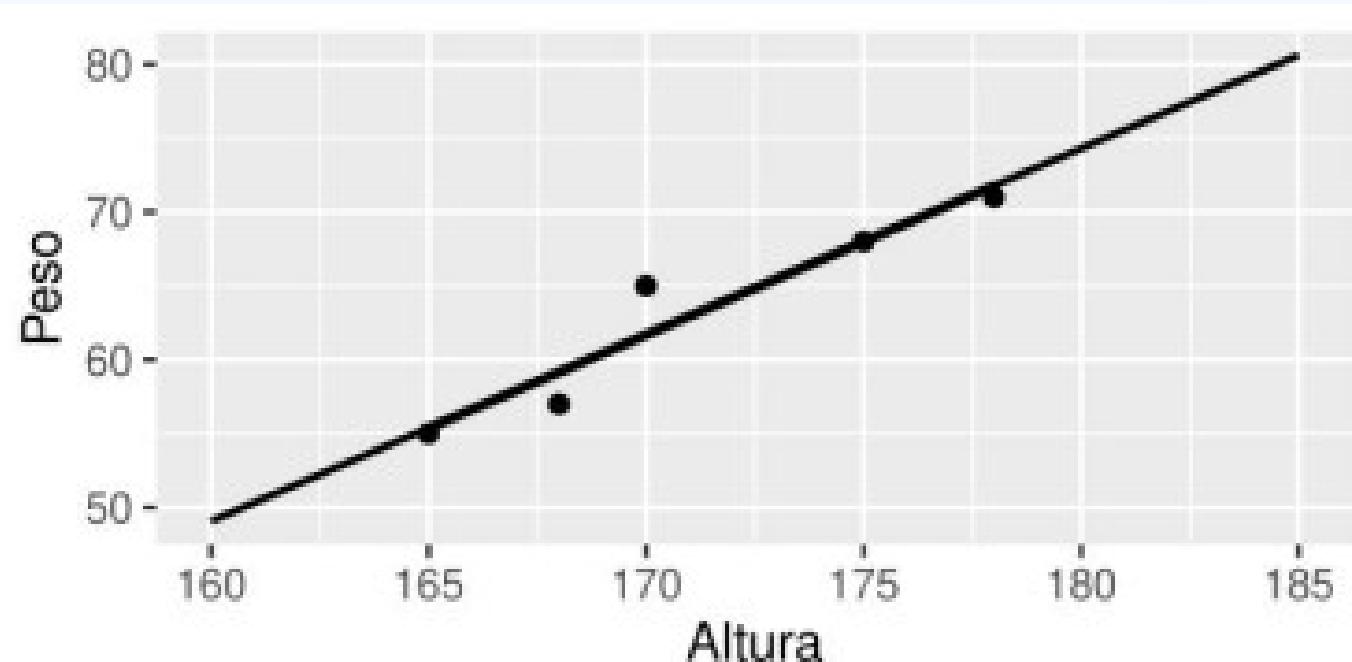
Este relacionamento existe em apenas uma direção: variável independente (x) -> variável dependente (y).

REGRESSÃO LINEAR



REGRESSÃO LINEAR SIMPLES

Altura	Peso
165	55
168	57
170	65
175	68
178	71



$$Y = a + bx$$

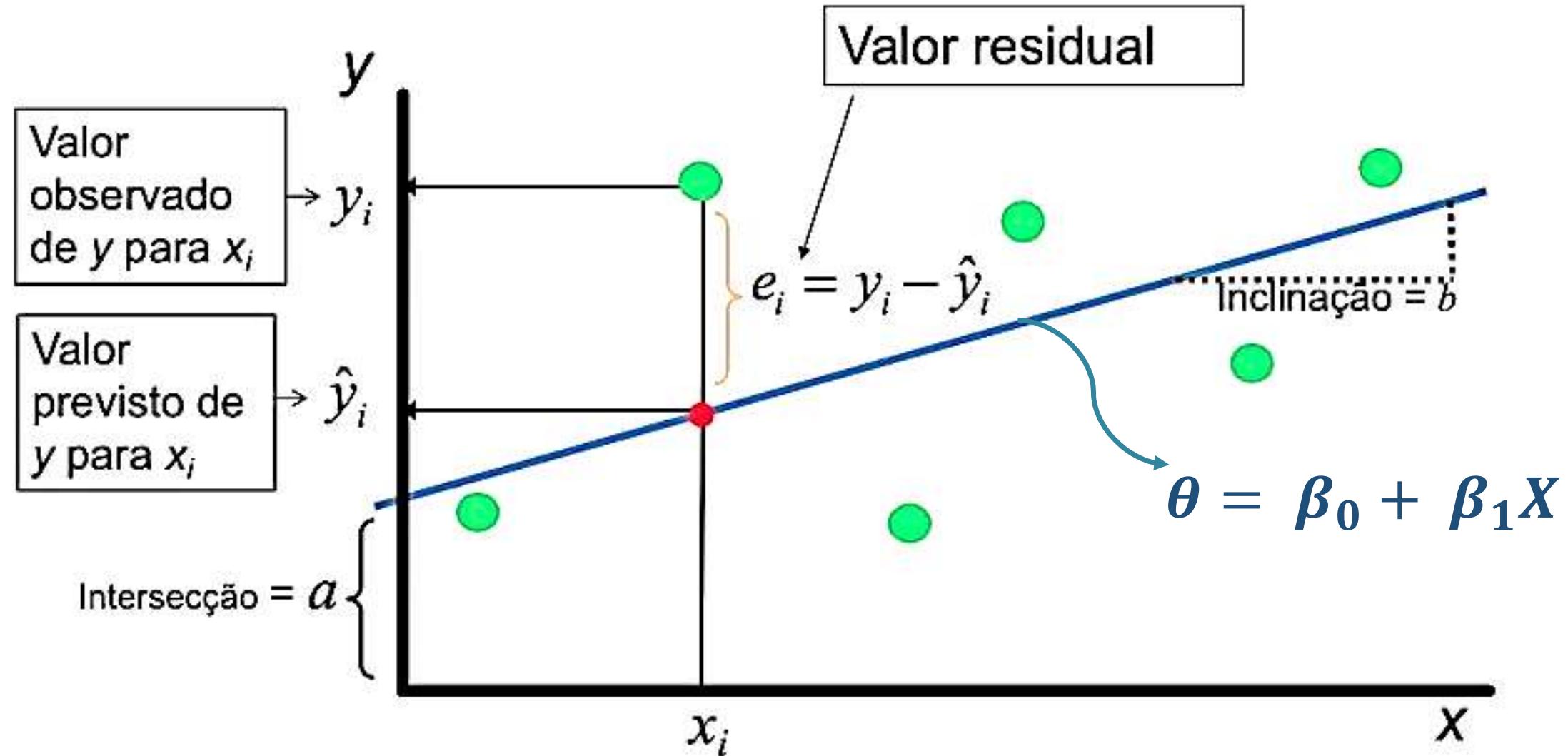
REGRESSÃO LINEAR SIMPLES

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

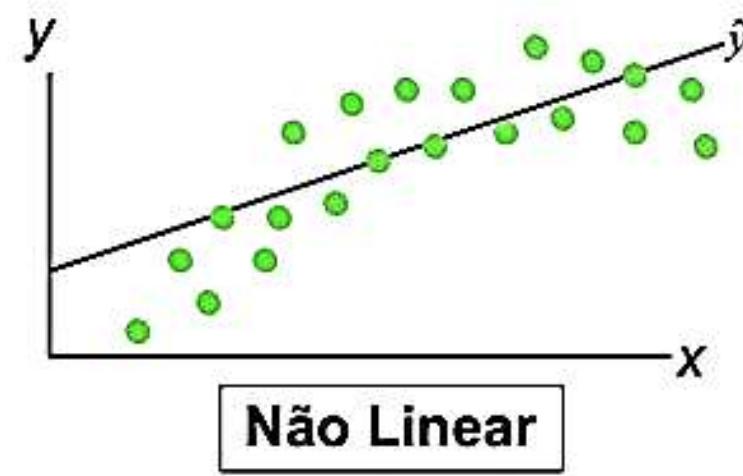
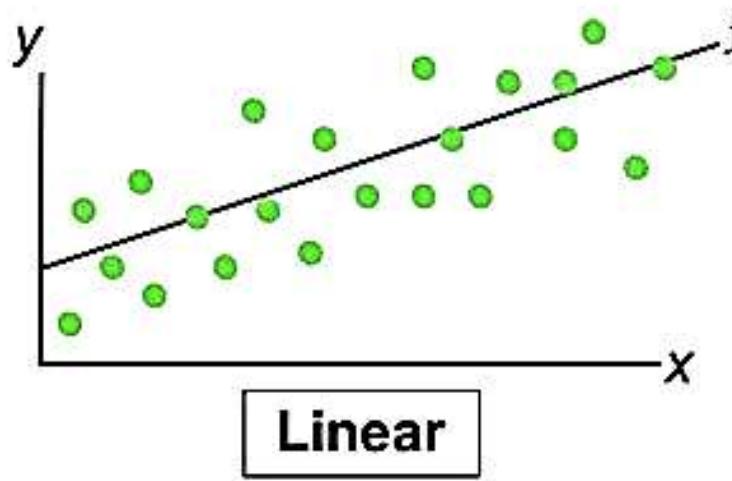
Diagram illustrating the components of the Simple Linear Regression equation:

- Variável Dependente** (Dependent Variable) points to \hat{Y}_i .
- Intercepto Y** (Y-intercept) points to $\hat{\beta}_0$.
- Coeficiente Angular** (Angular Coefficient) points to $\hat{\beta}_1$.
- Variável independente** (Independent Variable) points to X_i .
- Componente Linear** (Linear Component) is indicated by a bracket under the terms $\hat{\beta}_0 + \hat{\beta}_1 X_i$.
- Componente do Erro Aleatório** (Error Component) is indicated by a bracket under the term e_i .
- Erro Aleatório** (Error) points to e_i .

REGRESSÃO LINEAR SIMPLES



PREMISSAS DA REGRESSÃO LINEAR SIMPLES

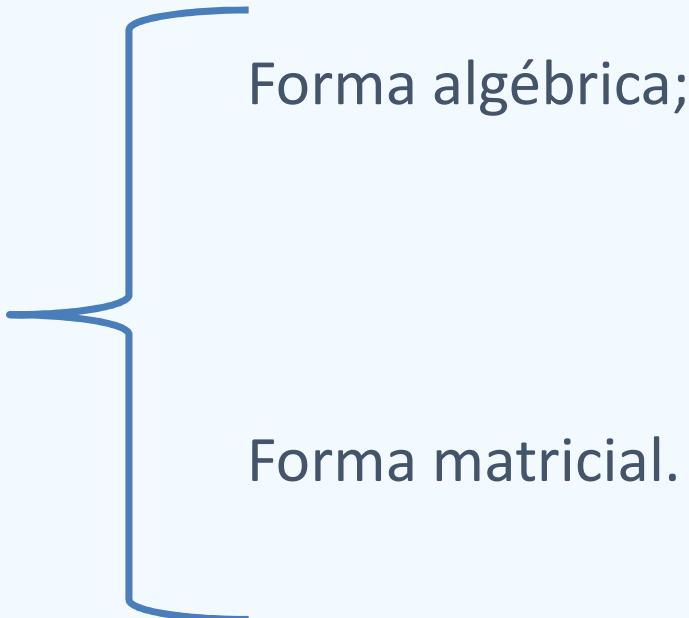


- O **relacionamento** entre as variáveis independentes e a variável dependente devem ser **linear**.
- Correlação de **moderada a forte**

ESTIMAÇÃO DOS PARÂMETROS

ESTIMADORES DOS COEFICIENTES DA REGRESSÃO

Estimadores dos Coeficientes de Regressão



FUNDAMENTOS TEÓRICOS

Regras de Derivação

- i) Se $f(x) = a$, então $f'(x) = 0$.
- ii) Se $f(x) = ax$, então $f'(x) = a$.
- iii) (Regra do tombo) Se $f(x) = x^a$, então $f'(x) = a \cdot x^{a-1}$.
- iv) (Derivada da soma) $[f(x) + g(x)]' = f'(x) + g'(x)$.

$$\begin{array}{ccc} f'(x) = x^3 & \xrightarrow{\hspace{1cm}} & f'(x) = 3x^{3-1} = 3x^2 \\ \\ f'(x) = 3x^4 & \xrightarrow{\hspace{1cm}} & f'(x) = 4 \cdot 3x^{4-1} = 12x^3 \end{array}$$

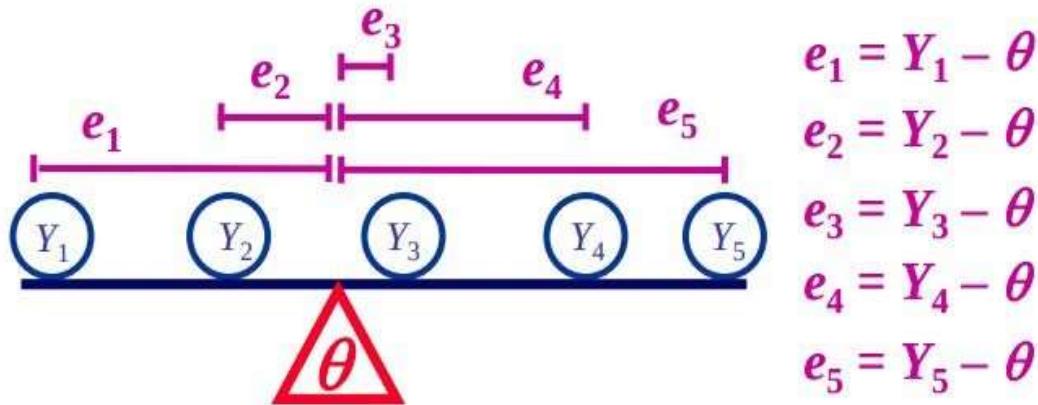
ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Equação da reta

$$\theta = \beta_0 + \beta_1 X$$

Modelo de Regressão Linear

$$Y = \beta_0 + \beta_1 X + e$$



Queremos encontrar
uma função que
minimize os erros.

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

1º Passo : Definir o Erro Quadrático Total

$$EQT = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$$

$$EQT = (Y_1 - \theta)^2 + (Y_2 - \theta)^2 + (Y_3 - \theta)^2 + (Y_4 - \theta)^2 + (Y_5 - \theta)^2$$

$$EQT = \sum_{i=1}^5 (Y_i - \theta)^2$$

Para uma amostra de tamanho n teremos:

$$Y_i = \hat{\theta} + \hat{e}_i \quad \Rightarrow \quad EQT = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (Y_i - \hat{\theta})^2$$

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

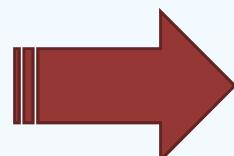
2º Passo : Encontrar $\hat{\theta}$ que Minimize o EQT

$$\frac{dEQT}{d\hat{\theta}} = 0$$

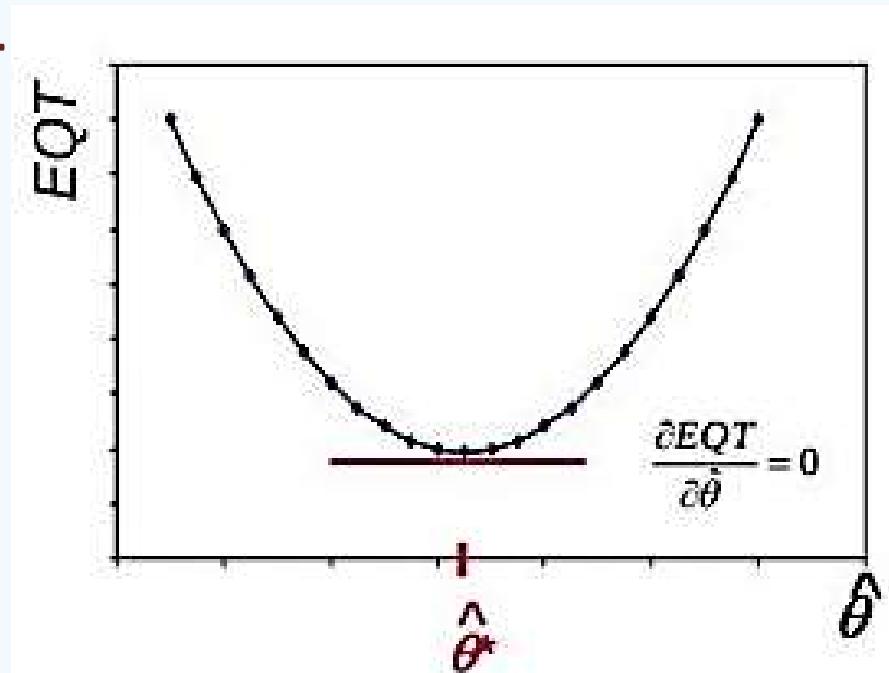
$$EQT = \sum_{i=1}^n (Y_i - \hat{\theta})^2$$

$$\frac{dEQT}{d\hat{\theta}} = \sum_{i=1}^n 2(Y_i - \hat{\theta})(-1) = -2 \sum_{i=1}^n Y_i + 2 \sum_{i=1}^n \hat{\theta}$$

$$\frac{dEQT}{d\hat{\theta}} = -2 \sum_{i=1}^n Y_i + 2n\hat{\theta} = 0$$



$$\hat{\theta} = \frac{\sum_{i=1}^n Y_i}{n}$$



ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Para encontrar os valores que minimizam o EQT:

$$\frac{dEQT}{d\hat{\alpha}} = -2 \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta} X_i)](-1) = 0$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Para encontrar os valores que minimizam o EQT:

$$\frac{dEQT}{d\hat{\beta}} = -2 \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta} X_i)](-1) = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Definindo β_0 e β_1 que minimizam o EQT:

$$Y = \beta_0 + \beta_1 X + e$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$$
$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

$$Y = XB + e$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \\ n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & \cdot \\ \vdots & \vdots \\ n & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ \vdots \\ e_n \end{bmatrix}$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$Y = XB + e$$

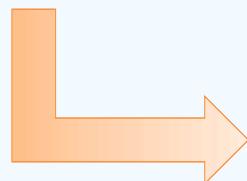
$$e = Y - XB$$

$$SQR = e'e = (Y - XB)'(Y - XB)$$

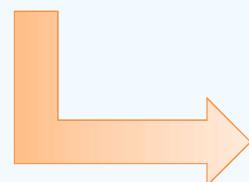
$$e'e = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 & e_5 & \dots & e_n \end{bmatrix} X \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ \vdots \\ e_n \end{bmatrix} = e^2_1 + e^2_2 + e^2_3 + e^2_4 + e^2_5 + \dots + e^2_n$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$SQR = e'e = (Y - XB)'(Y - XB)$$



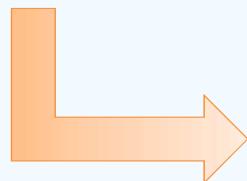
$$= Y'Y - 2BX'Y + B'X'XB$$



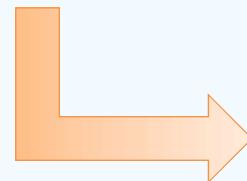
$$\frac{\partial SQR}{\partial B} = -2X'Y + 2B'X'X \equiv 0$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

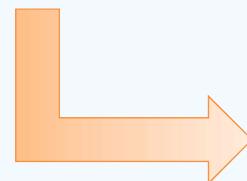
$$-2X'Y + 2B'X'X = 0$$



$$2B'X'X = 2X'Y$$



$$B'X'X = X'Y$$



$$\hat{B} = (X'X)^{-1}X'Y$$

PREMISSAS DA REGRESSAO LINEAR SIMPLES

PREMISSAS DA REGRESSÃO LINEAR SIMPLES

- Análise de Outliers de resíduos
- Homocedasticidade
- Normalmente distribuído

$$e_i = y_i - \hat{y}_i$$

Média = 0
Variância constante
Covariância = 0

ANÁLISE DE OUTLIERS

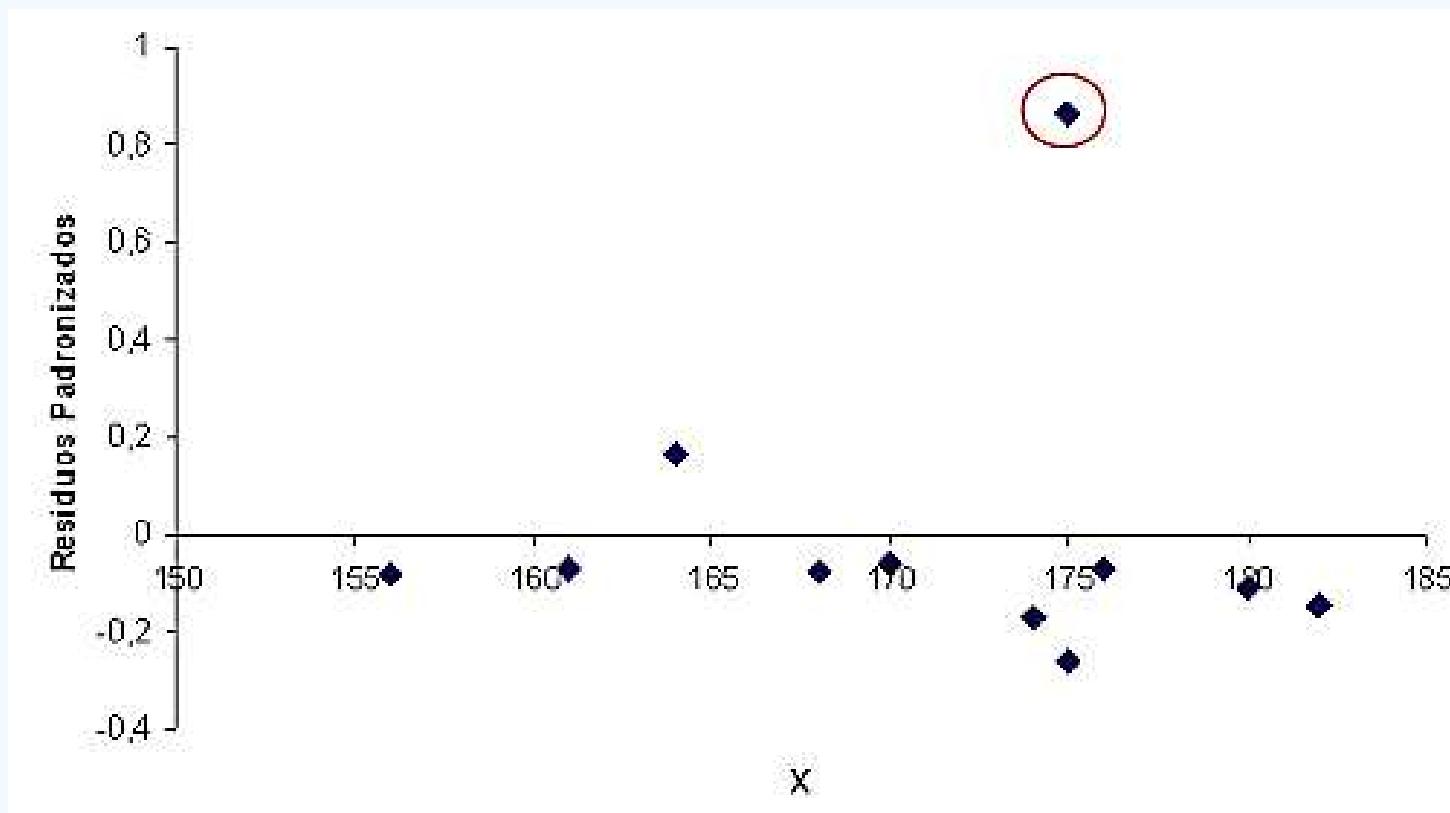


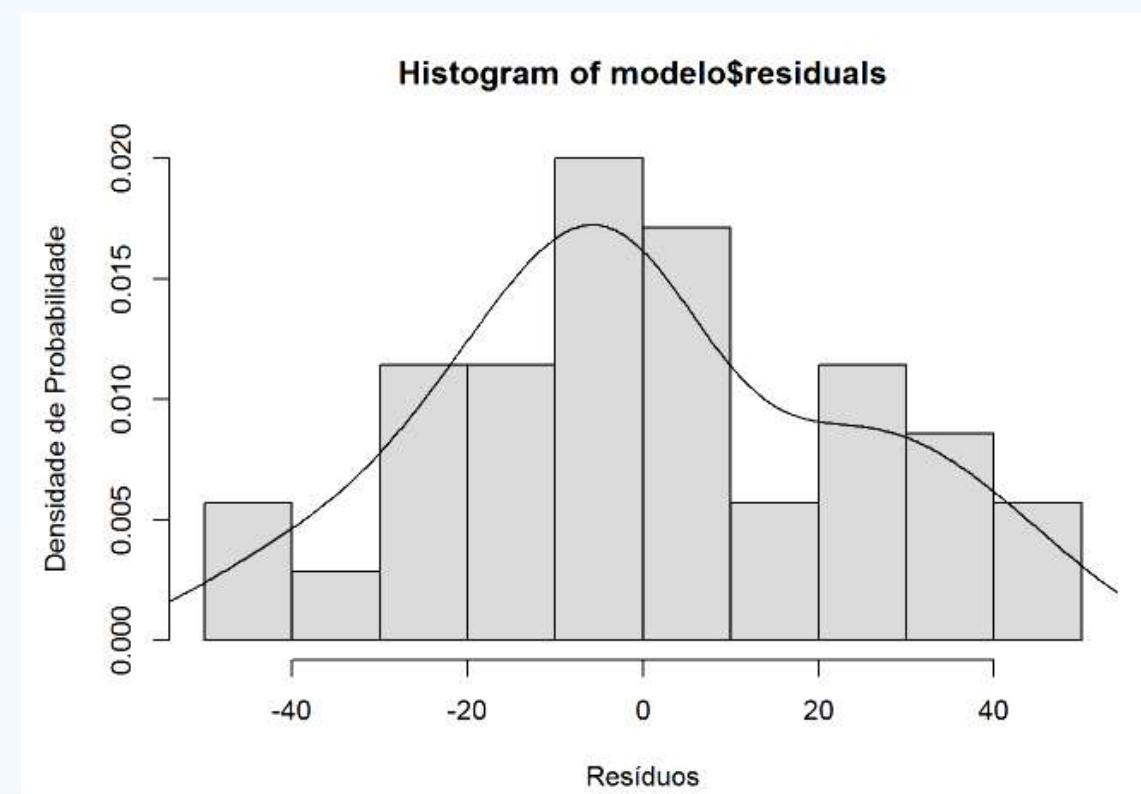
Gráfico de Resíduos padronizados vs Valores ajustados

NORMALIDADE DOS RESÍDUOS

Teste de Shapiro Wilk

H_0 = distribuição normal : $p > 0.05$

H_1 = distribuição não normal : $p \leq 0.05$



ANÁLISE DA HOMOCEDASTICIDADE DOS RESÍDUOS

Homocedasticidade: A variância dos erros é, condicionada aos valores das variáveis explanatórias, será constante.

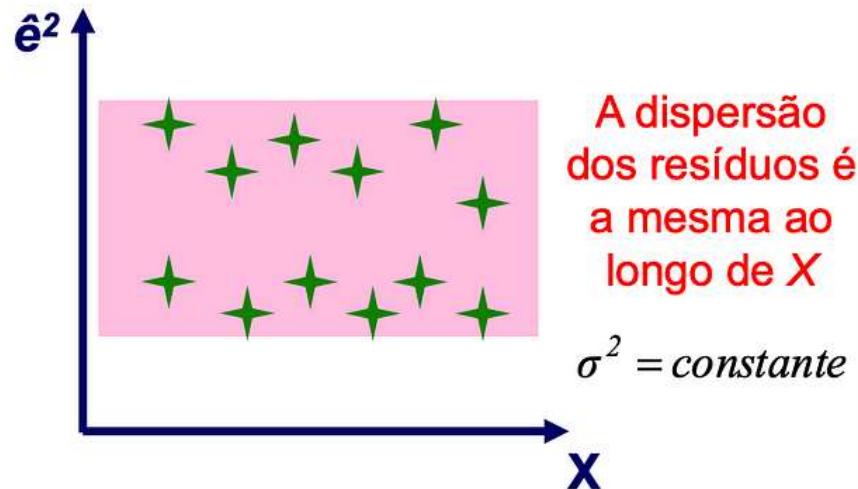
Teste Breusch-Pagan (Homocedasticidade)

H_0 = existe homocedasticidade : $p > 0.05$

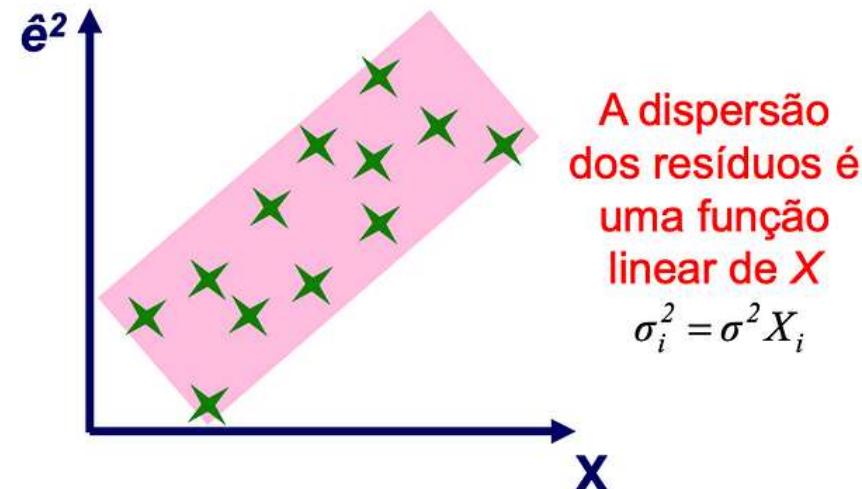
H_a = não existe homocedasticidade : $p \leq 0.05$



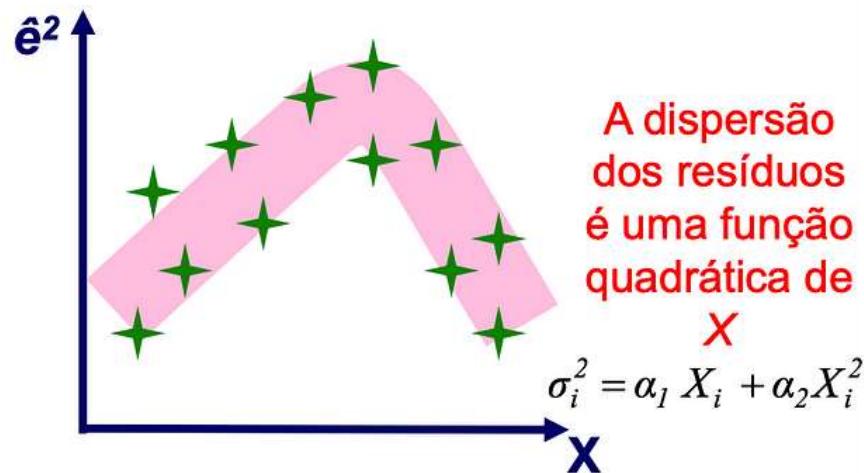
Homocedasticidade



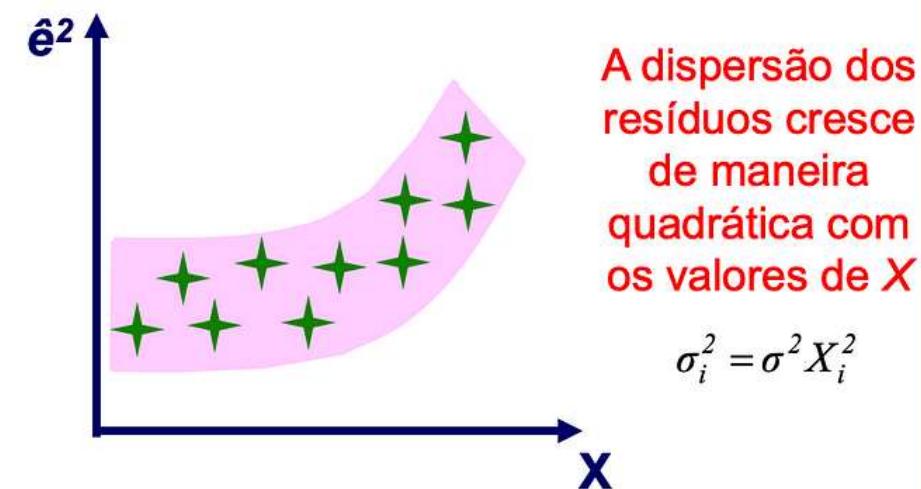
Heterocedasticidade



Heterocedasticidade



Heterocedasticidade



TESTE - T

Avaliando a significância de cada parâmetro β do modelo

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

$$H_0: p - valor \geq 0,05$$

$$H_1: p - valor < 0,05$$

TESTE – T – FORMULAÇÃO DAS HIPÓTESES

- Para cada coeficiente de regressão β , a hipótese nula (H_0) geralmente afirma que não há efeito significativo.
- A hipótese alternativa (H_1), por outro lado, afirma que há um efeito significativo.
- Os testes são bilaterais

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

TESTE – T – ESTATÍSTICA DE TESTE

A estatística do teste é calculada usando a estimativa do coeficiente ($\hat{\beta}_1$) e seu erro padrão ($SE(\hat{\beta}_1)$).

A estatística do teste segue uma distribuição t de Student.

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

TESTE – T – ESTATÍSTICA DE TESTE

O desvio padrão do estimador de um coeficiente de regressão ($SE(\hat{\beta}_1)$) pode ser calculado usando a seguinte fórmula:

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

onde:

- s é o desvio padrão dos resíduos do modelo (erro padrão residual),
- n é o número de observações,
- x_i são os valores da variável independente,
- \bar{x}_i é a média dos valores da variável independente.

TESTE – T – ESTATÍSTICA DE TESTE

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

onde:

y_i são os valores observados da variável dependente,
 \hat{y}_i são os valores previstos pela regressão.

TESTE - F

Avalia a significância global de um modelo de regressão linear, ou seja, para testar se pelo menos uma das variáveis independentes tem um efeito significativo sobre a variável dependente.

O teste F é comumente usado em modelos de regressão múltipla.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

H_1 : Pelo menos um β_j é diferente de zero

$$H_0 = F_{Calc} \leq F_{Crítico} \text{ ou } p\text{-valor} \geq 0,05$$

$$H_1 = F_{Calc} > F_{Crítico} \text{ ou } p\text{-valor} < 0,05$$

TESTE – F – ESTATÍSTICA DE TESTE

A estatística do teste F é calculada como:

$$F = \frac{(SQR/q)}{(SQE/(n-k-1))}$$

Onde:

- **SQR** é a soma dos quadrados da regressão
- q é o número de coeficientes a serem testados (neste caso, q=k),
- **SQE** é a soma dos quadrados dos resíduos
- n é o número de observações,
- k é o número de variáveis independentes no modelo.

TESTE – F – ESTATÍSTICA DE TESTE

Soma dos Quadrados da Regressão (SQR):

$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

\hat{y}_i são os valores previstos pela regressão para a observação
 \bar{y} é a média dos valores observados da variável dependente.
SQR, mais o modelo está explicando a variabilidade nos dados.

TESTE – F – ESTATÍSTICA DE TESTE

Soma dos Quadrados dos Erros (SQE):

$$SQE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

onde:

y_i são os valores observados da variável dependente,

\hat{y}_i são os valores previstos pela regressão para a observação

COEFICIENTE DE DETERMINAÇÃO (R^2)

Como avaliar o modelo?

O **coeficiente de determinação (R^2)** estima a proporção da variabilidade da variável dependente (Y) que é explicada pelas(s) variáveis independentes do modelo de regressão.

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n \widehat{e}_i^2}{\sum_{i=1}^n y_i^2} = \widehat{\beta}_1^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n y_i^2}$$

Escala de R^2 :



0
independência
linear

A significância das escalas depende muito da
natureza da variável dependente

1
relação
linear exata

call:

lm(formula = custo ~ idade, data = dados)

Modelo a ser criado

Residuals:

Min	1Q	Median	3Q	Max
-463.37	-277.04	-45.04	218.15	751.27

Análise de Resíduos

Coefficients: Coeficientes estimados

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-558.949	368.759	-1.516	0.168
idade	61.868	8.582	7.209	9.16e-05 ***

Teste -t

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 389.7 on 8 degrees of freedom

Multiple R-squared: 0.8666, R^2 Adjusted R-squared: 0.8499

F-statistic: 51.98 on 1 and 8 DF, p-value: 9.161e-05

Teste F

OLS Regression Results

Dep. Variable:	y	R-squared:	<u>0.942</u>	R ²
Model:	OLS	Adj. R-squared:	0.939	
Method:	Least Squares	F-statistic:	291.1	→ Teste F
Date:	Wed, 23 Jan 2019	Prob (F-statistic):	1.47e-12	
Time:	16:51:04	Log-Likelihood:	-35.596	

Coeficientes

nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	364.1800	13.765	26.457	0.000	335.261	393.099
x1	-1.0320	0.060	-17.062	0.000	-1.159	-0.905

→ Teste T

Omnibus:	0.433	Durbin-Watson:	2.235
Prob(Omnibus):	0.805	Jarque-Bera (JB):	0.115
Skew:	-0.182	Prob(JB):	0.944
Kurtosis:	2.926	Cond. No.	9.26e+03

→ Resíduos