PUC Minas

Introdução a Modelagem Estatística

MODELOS ESTATÍSTICOS

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

- **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.

OS 5 PRINCIPAIS MODELOS ESTATÍSTICOS

- Regressão Linear
- Análise de Variância (ANOVA)
- Regressão Logística
- Análise de Sobrevivência
- Séries Temporais

PUC Minas

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

- **Compreensão do fenômeno**
  - Entender como diferentes variáveis se interagem e se influenciam. Retirar insights sobre o fenômeno estudado, permitindo uma compreensão mais profunda.
- **Previsão e predição**
  - Prever resultados futuros com base em dados históricos. Essas previsões podem ser valiosas para tomada de decisões e planejamento estratégico.
- **Teste de hipóteses**
  - Testar hipóteses e avaliar a significância estatística de diferentes fatores. Eles ajudam a determinar se uma relação observada entre variáveis é estatisticamente significativa ou simplesmente resultado do acaso.

PUC Minas

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

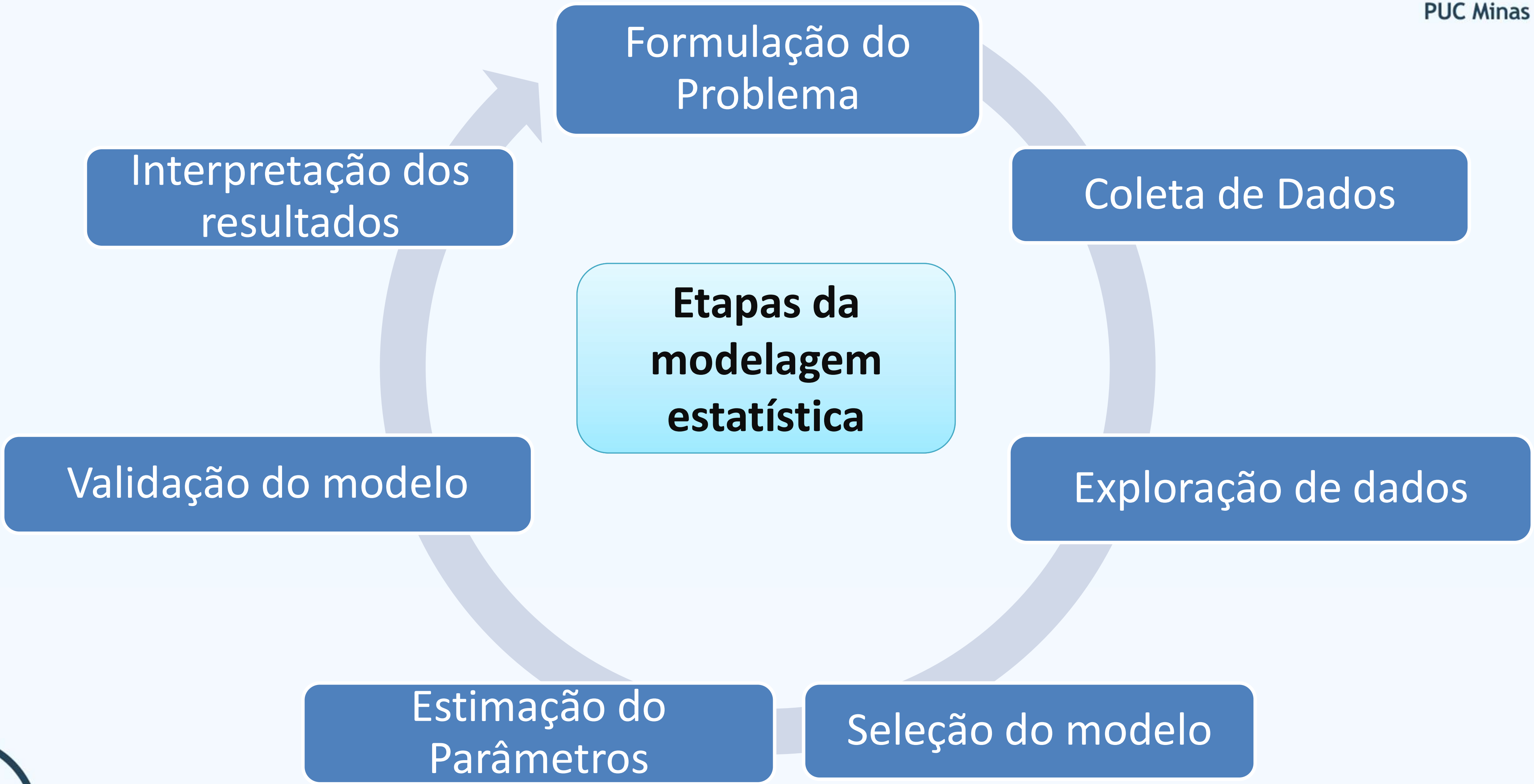
- Permitem fazer inferências sobre a população com base em uma amostra. Eles ajudam a generalizar conclusões e insights para além dos dados observados.

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.


Formulação do Problema
Coleta de Dados
Exploração de dados
Seleção do modelo
Estimação do Parâmetros
Validação do modelo
Interpretação dos resultados
Etapas da modelagem estatística

PUC Minas

FUNDAMENTOS PARA MODELOS DE REGRESSAO LINEAR

FUNDAMENTOS PARA MODELOS DE REGRESSÃO

MEDIDAS RESUMO

MEDIDAS DE ASSOCIAÇÃO

Foto de Aleks Magnusson no Pexels.

PUC Minas

Classificação de Variáveis Aleatórias

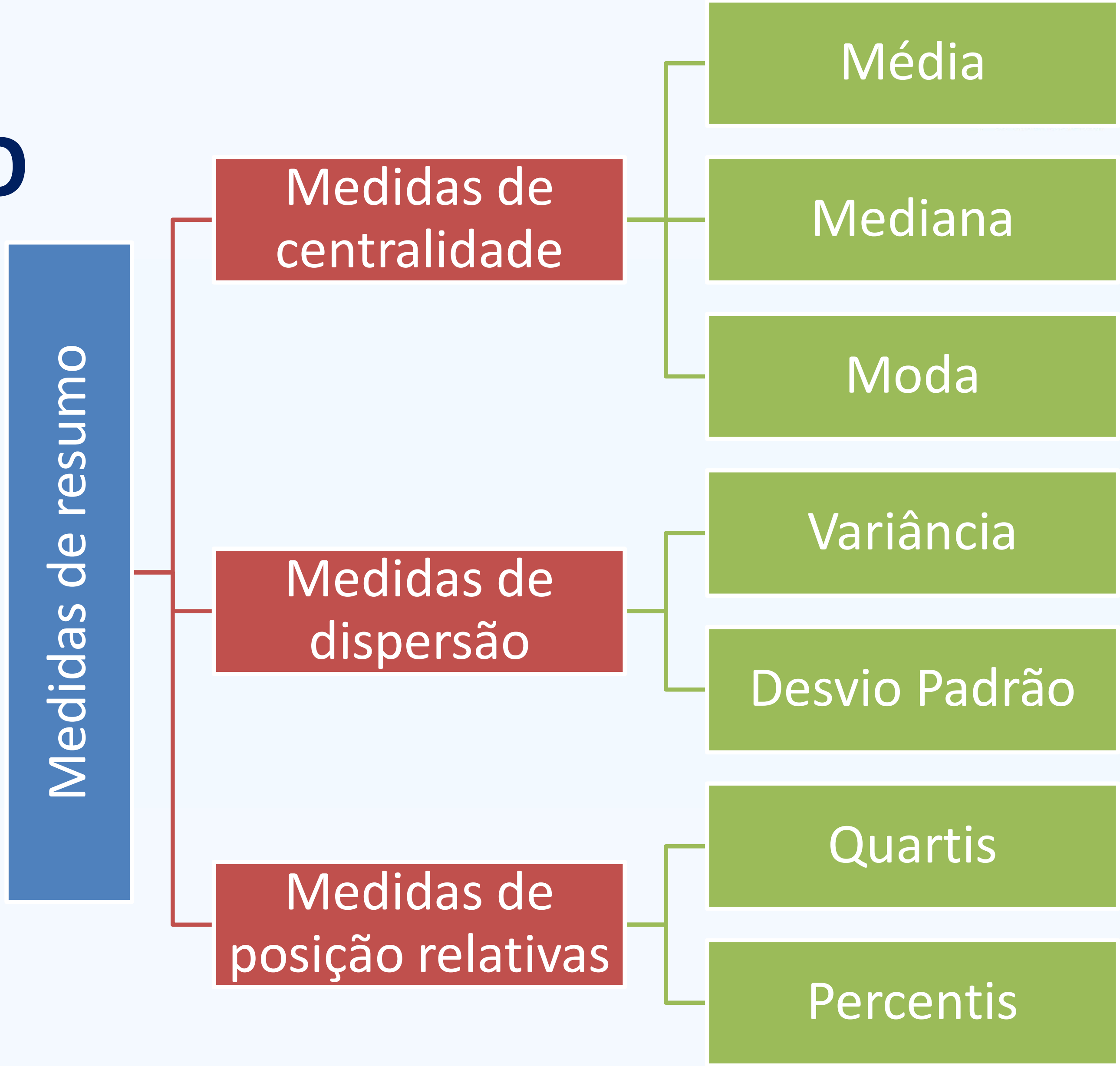
- Variáveis Aleatórias
  - Discreta
    - Número de filhos, número de sucessos em n tentativas
  - Contínua
    - massa (kg) do objeto , altura (m) do indivíduo, idade, temperatura, diâmetro da peça,

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?

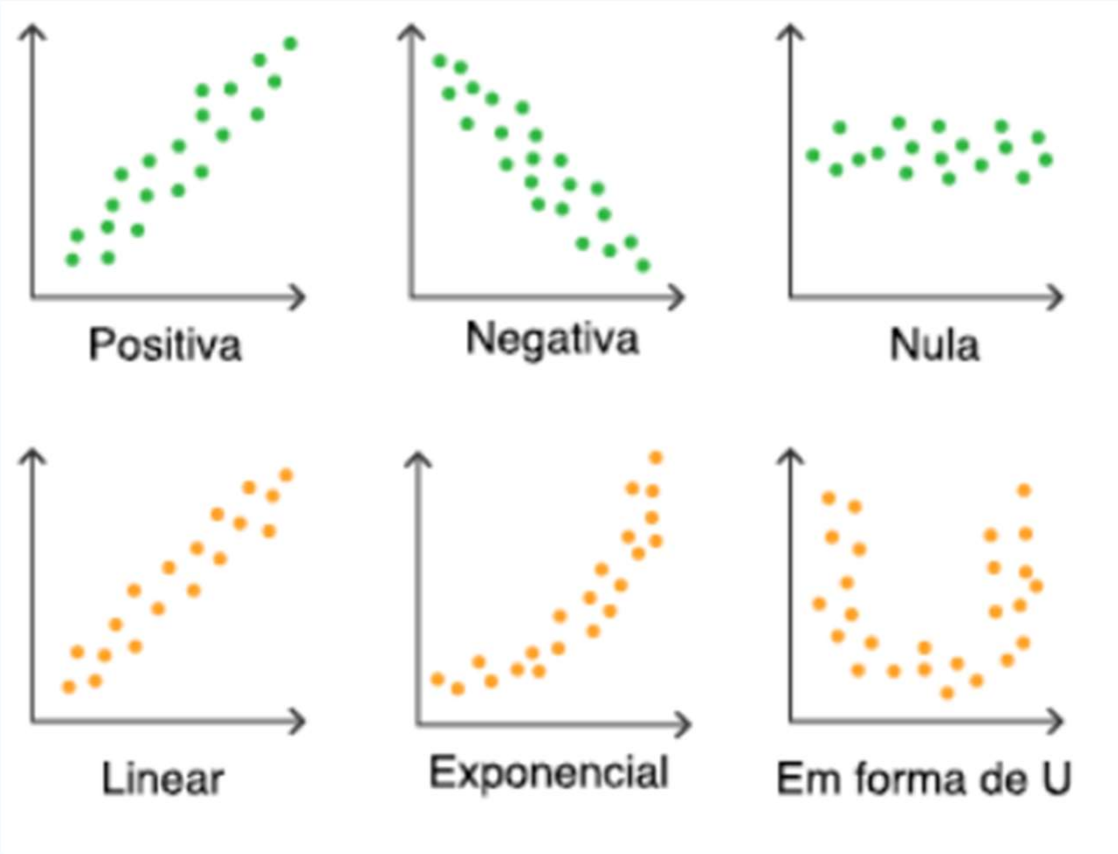


PUC Minas

MEDIDAS DE ASSOCIAÇÃO

Correlação linear

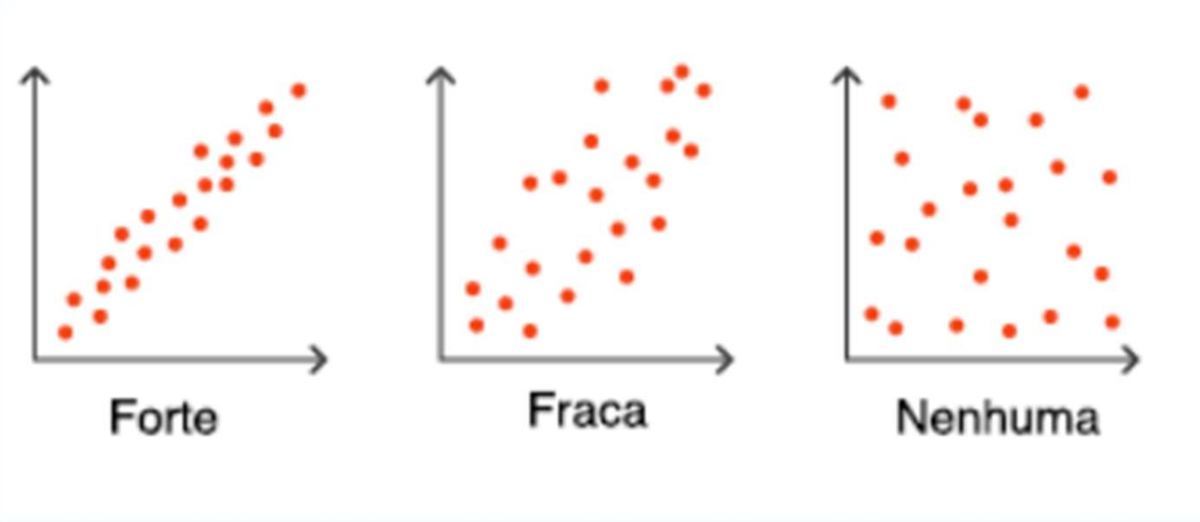
Determinado através de gráficos de dispersão e do coeficiente de variação



MEDIDAS DE ASSOCIAÇÃO

Correlação linear

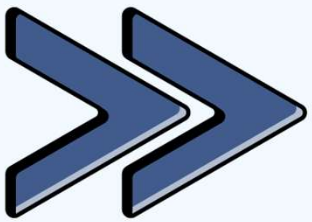
Força da Correlação



CARACTERÍSTICAS DA CORRELAÇÃO

Pode ser um valor entre -1 e 1

-1 ↔ **1**



Mostra a força e a direção entre as variáveis

A correlação de A ~ B é a mesma que B ~ A

COVARIÂNCIA E CORRELAÇÃO

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

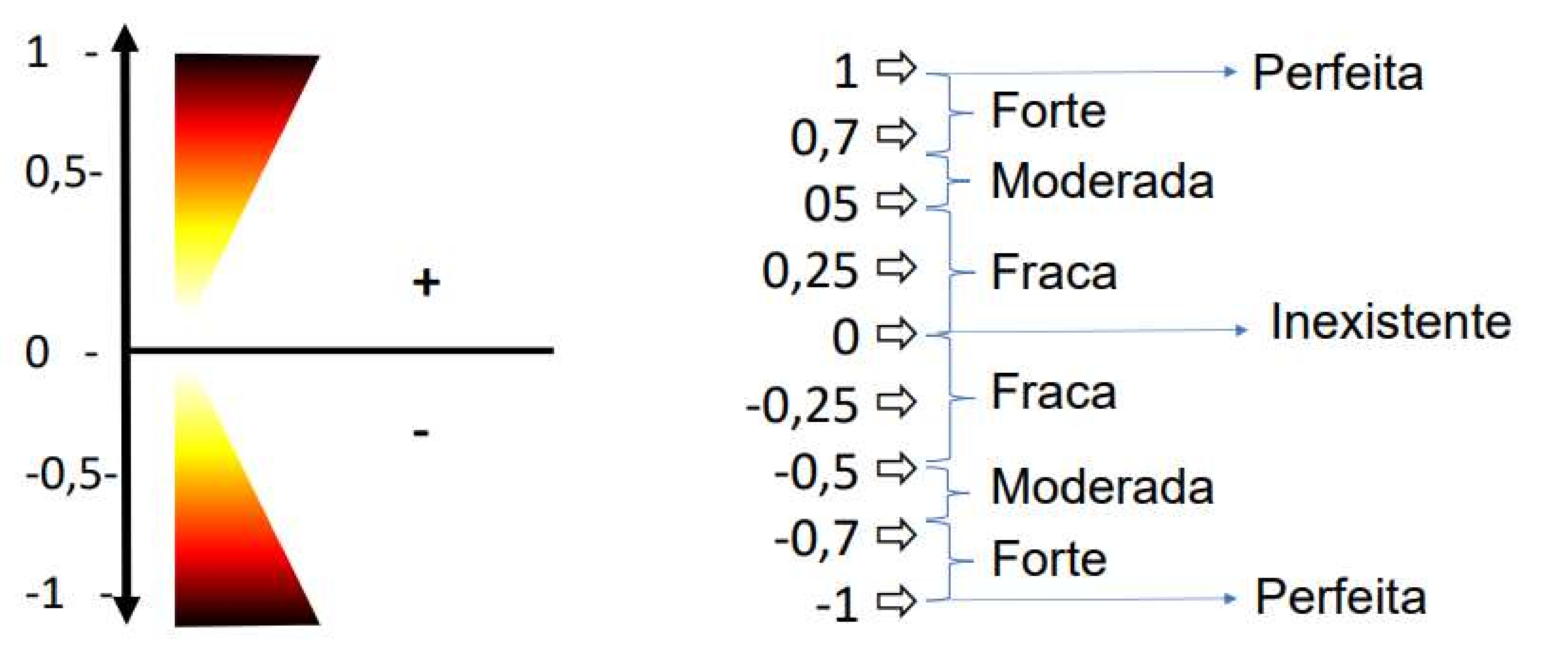
$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^{n} (y_{1i} - \overline{y_1}) \cdot (y_{2i} - \overline{y_2})$$

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

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

PUC Minas

FORÇA DA CORRELAÇÃO

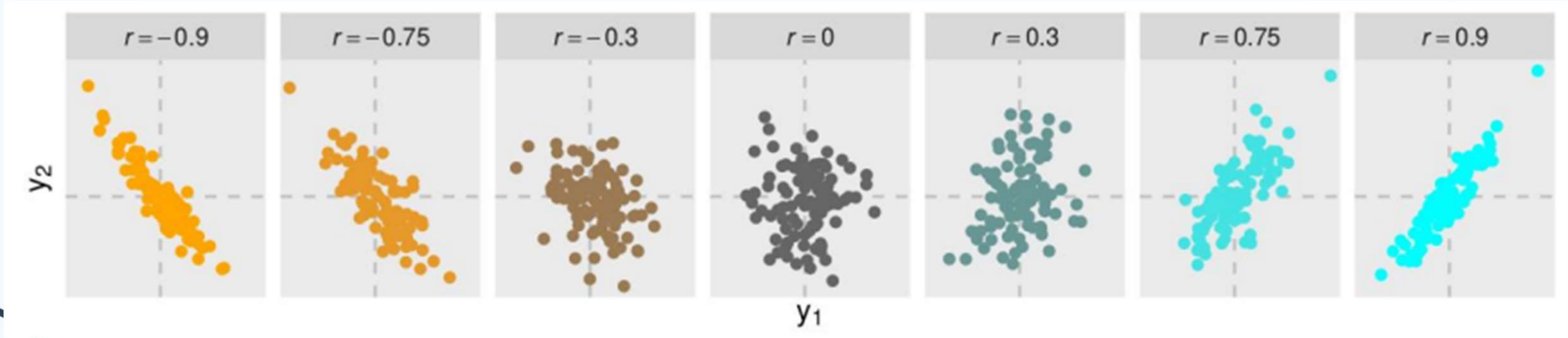


PUC Minas

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).



PUC Minas

TESTE DE HIPÓTESE PARA A CORRELAÇÃO

Sejam as hipóteses nula e alternativa:

H0 : r = 0.

Ha : r ≠ 0

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

Teste t

Graus de Liberdade

gl = n -2

TIPOS DE CORRELAÇÃO

- Pearson → PARAMÉTRICO
- Postos → NÃO PARAMÉTRICO
- Kendall → NÃO PARAMÉTRICO

PUC Minas

COEFICIENTE DE PEARSON

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

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

PUC Minas

COEFICIENTE DE SPEARMAN – CORRELAÇÃO DE POSTOS

- *Teste **não paramétrico** -> pode ser aplicado em **variáveis não normais***
- *Mede a força da relaçao 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

Postos

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

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

d^2

d^2
$(12-9)^2$
$(10{,}5 - 12)^2$
$(10{,}5-1)^2$
$(9-6)^2$
$(8-11)^2$
$(7-4)^2$
$(6-8)^2$
$(5-2{,}5)^2$
$(4-2{,}5)^2$
$(3-5)^2$
$(2-10)^2$
$(1-7)^2$

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*

PUC Minas

CORRELAÇÃO DE KENDALL

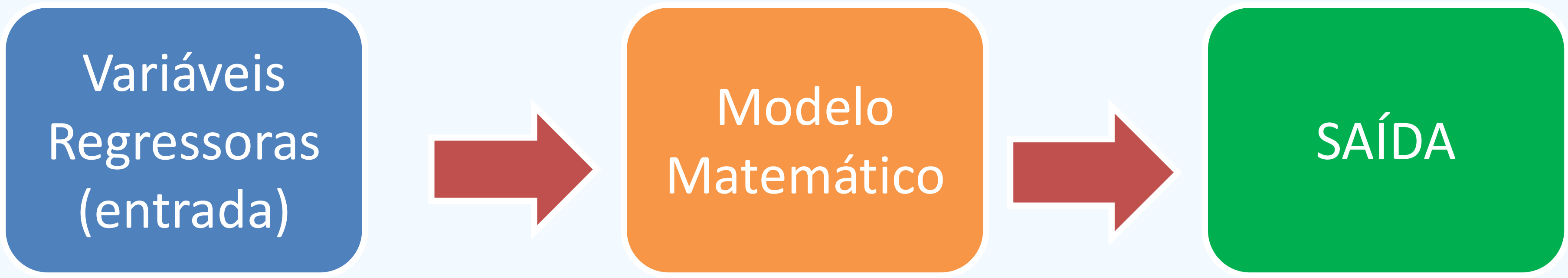
$\boldsymbol{x_i > x_j}$ **e** $\boldsymbol{y_i > y_j}$ **ou se** $\boldsymbol{x_i < x_j}$ **e** $\boldsymbol{y_i < y_j}$

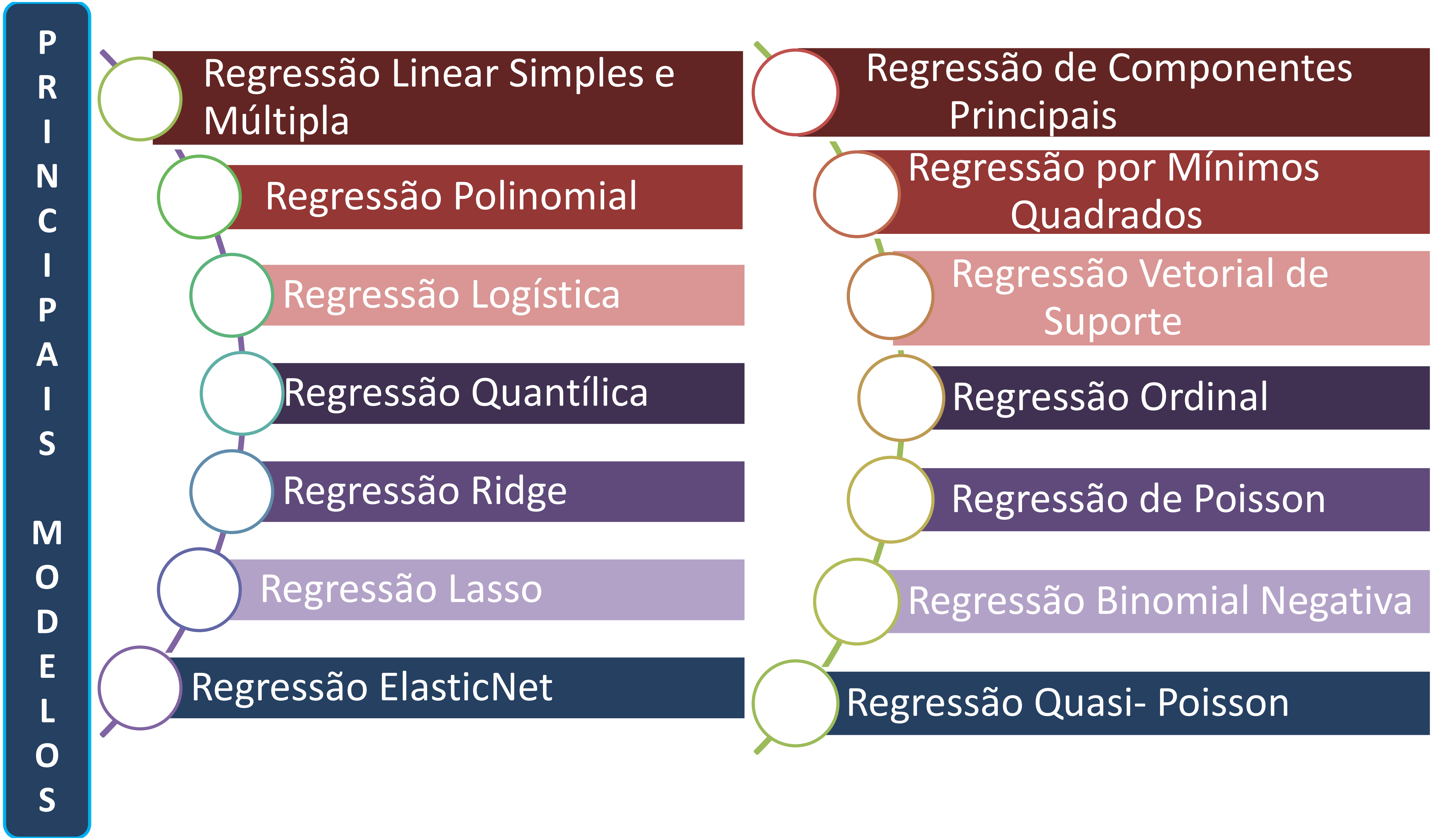
$$\tau = \frac{(Qtd\ de\ pares\ concordantes) - (Qtd\ de\ pares\ discordantes)}{n(n-1)/2}$$

$\boldsymbol{x_i > x_j}$ **e** $\boldsymbol{y_i < y_j}$ **ou se** $\boldsymbol{x_i < x_j}$ **e** $\boldsymbol{y_i > y_j}$

REGRESSAO LINEAR

MODELOS DE REGRESSÃO




PRINCIPAIS MODELOS
Regressão Linear Simples e Múltipla
Regressão Polinomial
Regressão Logística
Regressão Quantílica
Regressão Ridge
Regressão Lasso
Regressão ElasticNet
Regressão de Componentes Principais
Regressão por Mínimos Quadrados
Regressão Vetorial de Suporte
Regressão Ordinal
Regressão de Poisson
Regressão Binomial Negativa
Regressão Quasi- Poisson

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

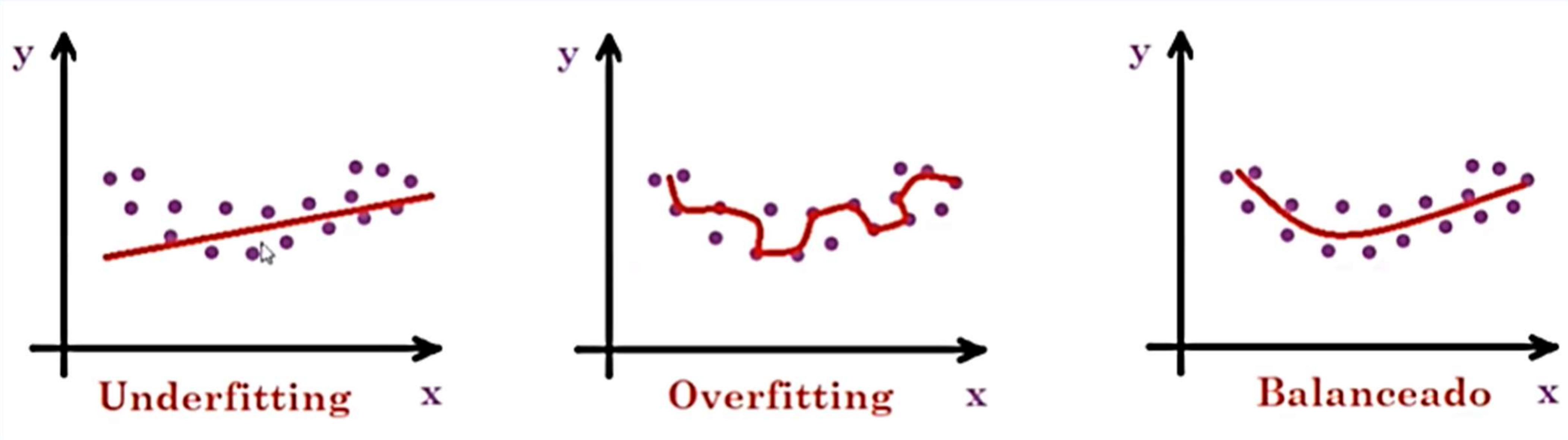
- Algoritmo não se encaixa com os dados de entrada

Overfitting

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

PUC Minas

INTRODUÇÃO SOBRE AJUSTE DE MODELOS



PUC Minas

REGRESSÃO LINEAR

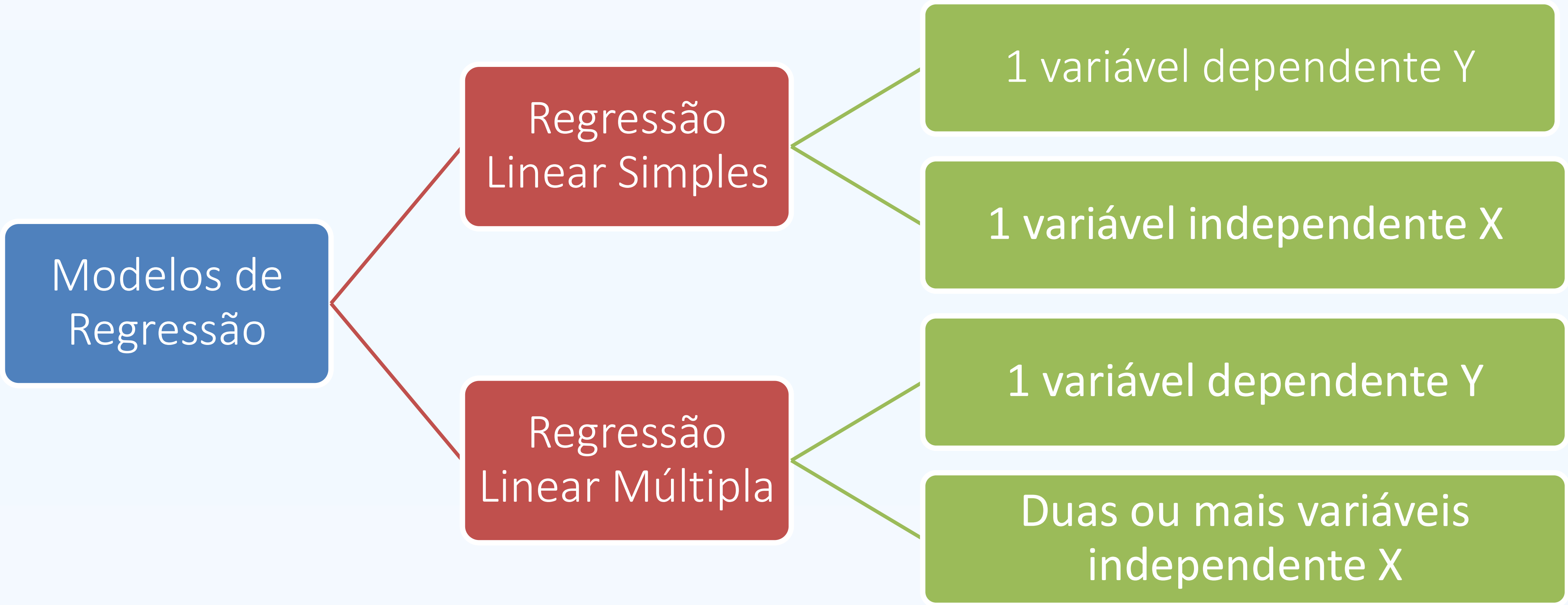
- *Técnica estatística que visa modelar a relação entre* ***uma variável dependente (ou resposta)*** *e* ***uma ou mais variáveis independentes (ou preditoras).***
- *Utilizada para prever valores contínuos e entender a relação linear entre as variáveis.*
- *O termo "linear" se refere ao fato de que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

PUC Minas

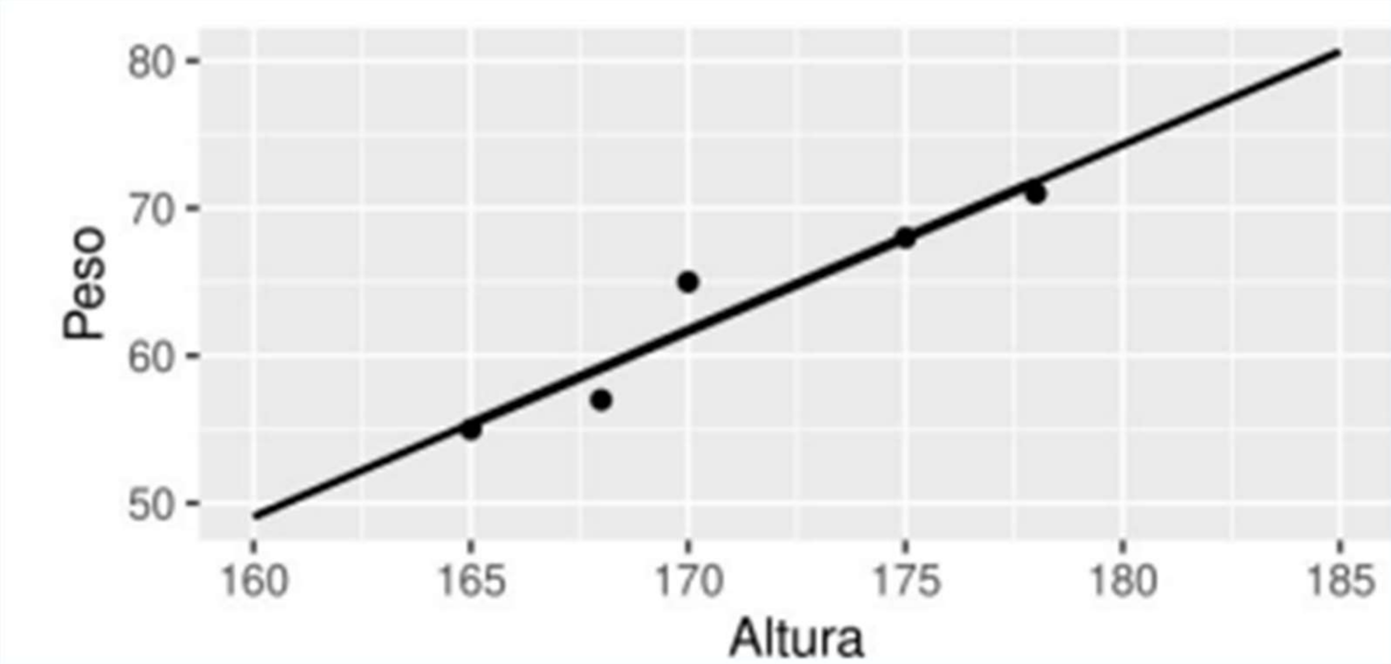
REGRESSÃO LINEAR



PUC Minas

REGRESSÃO LINEAR SIMPLES

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



$$\boxed{Y = a + bx}$$

PUC Minas

REGRESSÃO LINEAR SIMPLES

Variável independente

Intercepto Y

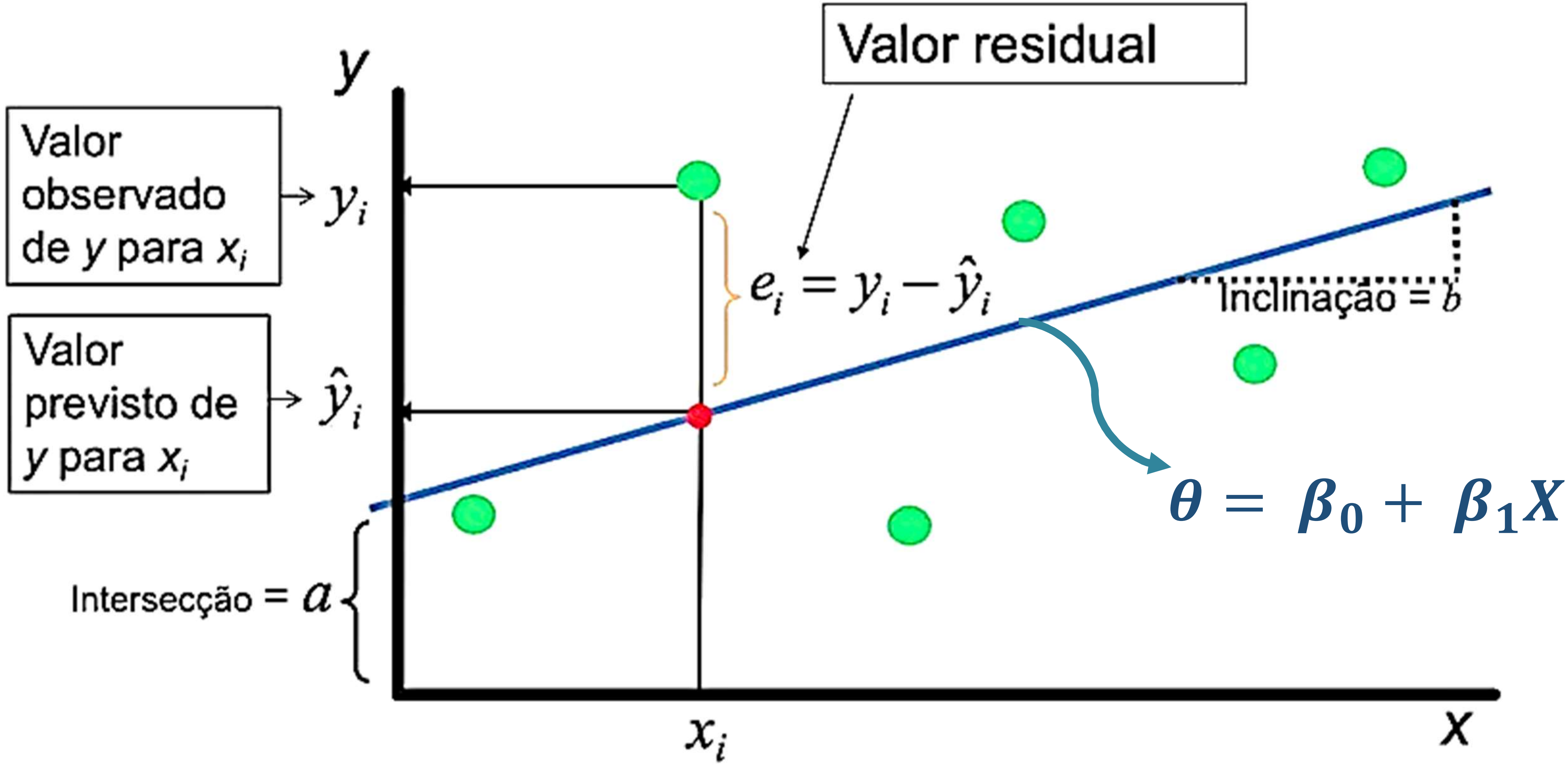
Coeficiente Angular

Erro Aleatório

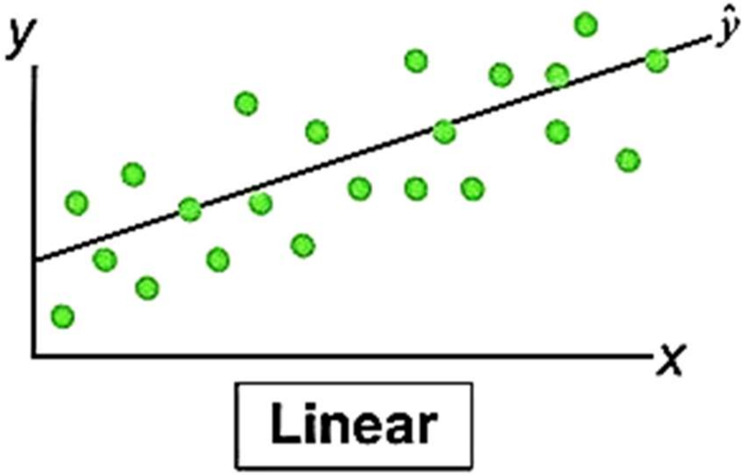
Variável Dependente

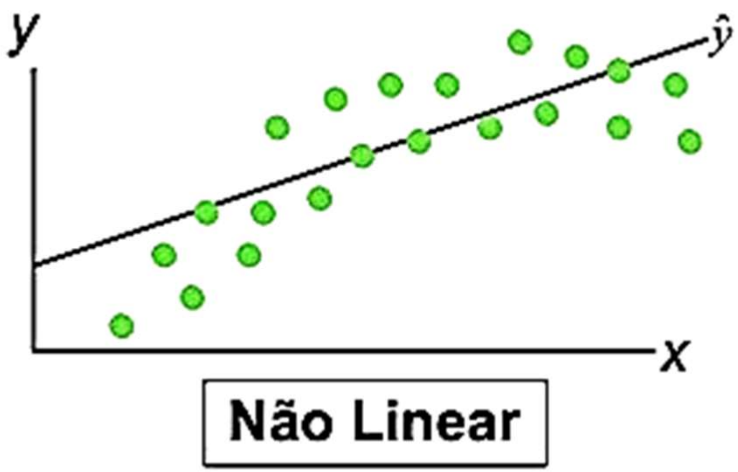
$$\underbrace{\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i}_{\text{Componente Linear}} + \underbrace{e_i}_{\text{Componente do Erro Aleatório}}$$

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

- Forma algébrica;
- Forma matricial.

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) $[f(x) + g(x)]' = f'(x) + g'(x)$.

$$f'(x) = x^3 \quad \Rightarrow \quad f'(x) = 3x^{3-1} = 3x^2$$

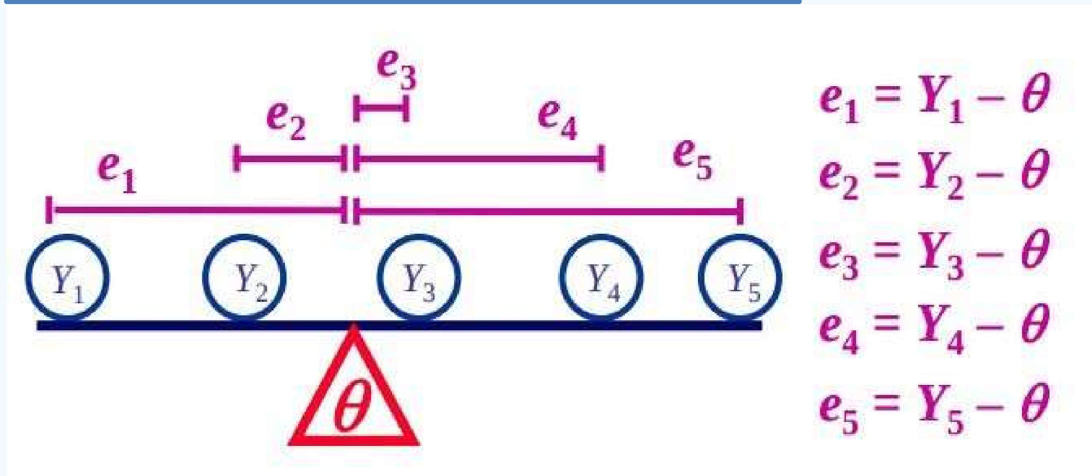
$$f'(x) = 3x^4 \quad \Rightarrow \quad f'(x) = 4 .3x^{4-1} = 12x^3$$

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Equação da reta $\boldsymbol{\theta = \beta_0 + \beta_1 X}$

Modelo de Regressão Linear $\mathbf{Y} = \boldsymbol{\beta_0 + \beta_1 X} + \mathbf{e}$



Queremos encontrar uma função que minimize os erros.

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

1º Passo : Definir o Erro Quadrático Total

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

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

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

Para uma amostra de tamanho n teremos:

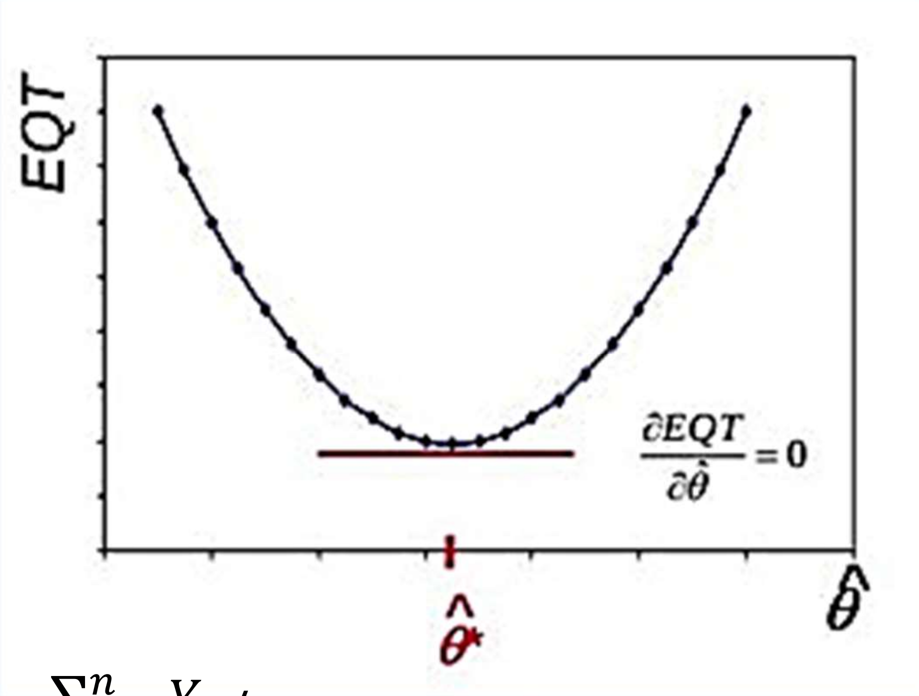
$$Y_i = \hat{\theta} + \hat{e_i} \Rightarrow \text{EQT} = \sum_{i=1}^{n} \widehat{e_i^2} = \sum_{i=1}^{n} (Y_i - \hat{\theta})^2$$

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

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

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

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



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

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

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Para encontrar os valores que minimizam o EQT:

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

$$\hat{\alpha} = \bar{Y} - \widehat{\beta_1}\, \bar{X}$$

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Para encontrar os valores que minimizam o EQT:

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

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA ALGÉBRICA

Definindo $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ que minimizam o EQT:

$$Y = \beta_0 + \beta_1\, X + e \quad \left\{ \begin{array}{l} \widehat{\beta_1} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \dfrac{\widehat{Cov}\,(X,Y)}{\widehat{Var}\,(X)} \\[2em] \widehat{\beta_0} = \bar{Y} \;-\; \widehat{\beta_1}\,\bar{X} \end{array} \right.$$

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

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

$$Y = XB + e$$

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$\mathrm{Y} = XB + e$$

$$\mathrm{e} = Y - XB$$

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

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

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

=Y'Y – 2BX'Y + B'X'XB

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

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

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

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

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

PUC Minas

PREMISSAS DA REGRESSAO LINEAR SIMPLES

PREMISSAS DA REGRESSÃO LINEAR SIMPLES

- Análise de Outiliers de resíduos
- Homocedasticidade
- Normalmente distribuído

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

$$\left\{ \begin{array}{l} \textbf{Média = 0} \\ \textbf{Variância constante} \\ \textbf{Covariância = 0} \end{array} \right.$$

ANÁLISE DE OUTLIERS

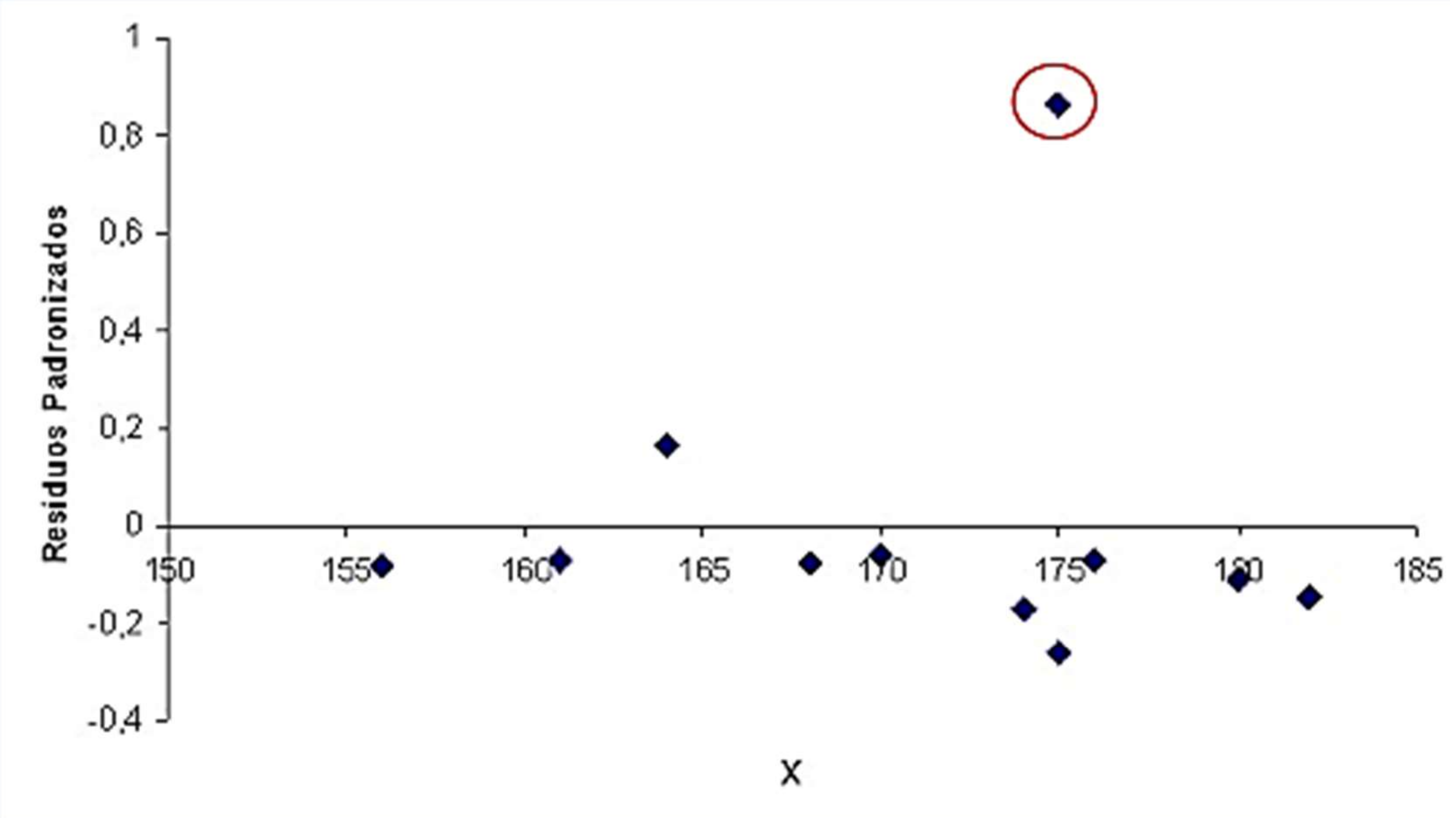


Gráfico de Resíduos padronizados vs Valores ajustados

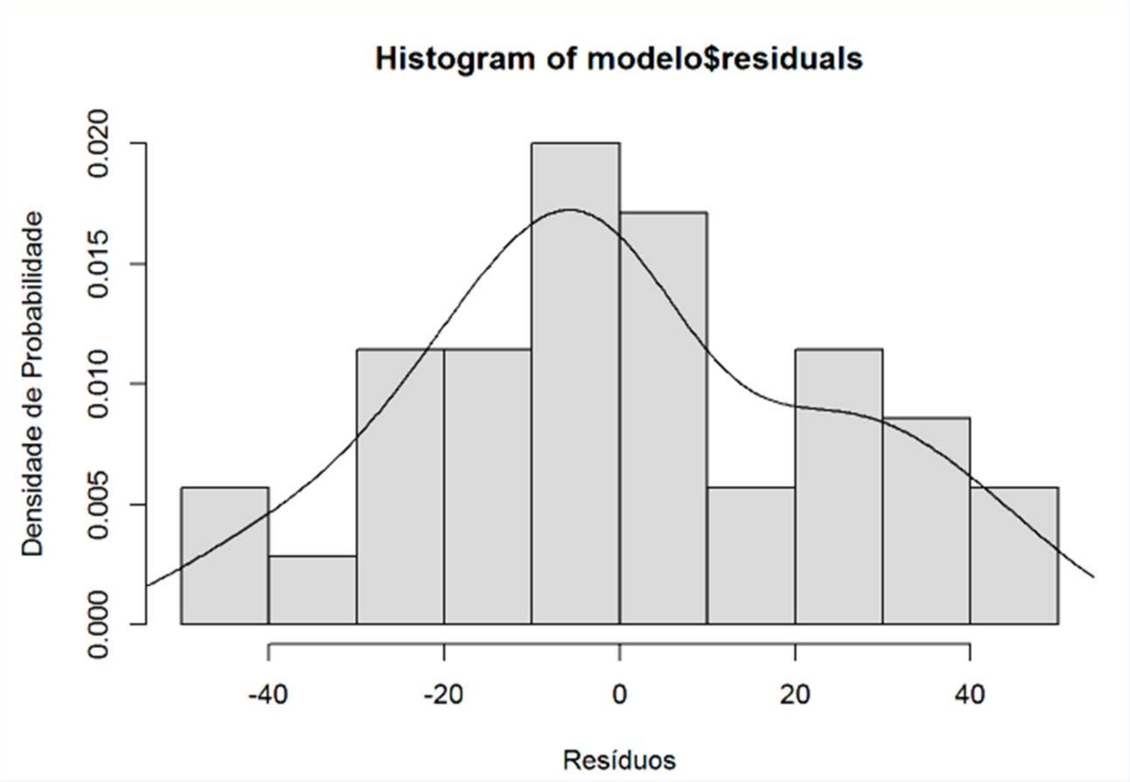
PUC Minas

NORMALIDADE DOS RESÍDUOS

Teste de Shapiro Wilk

H_0 = distribuição normal : $p > 0.05$

H_1 = distribuição não normal : $p <= 0.05$



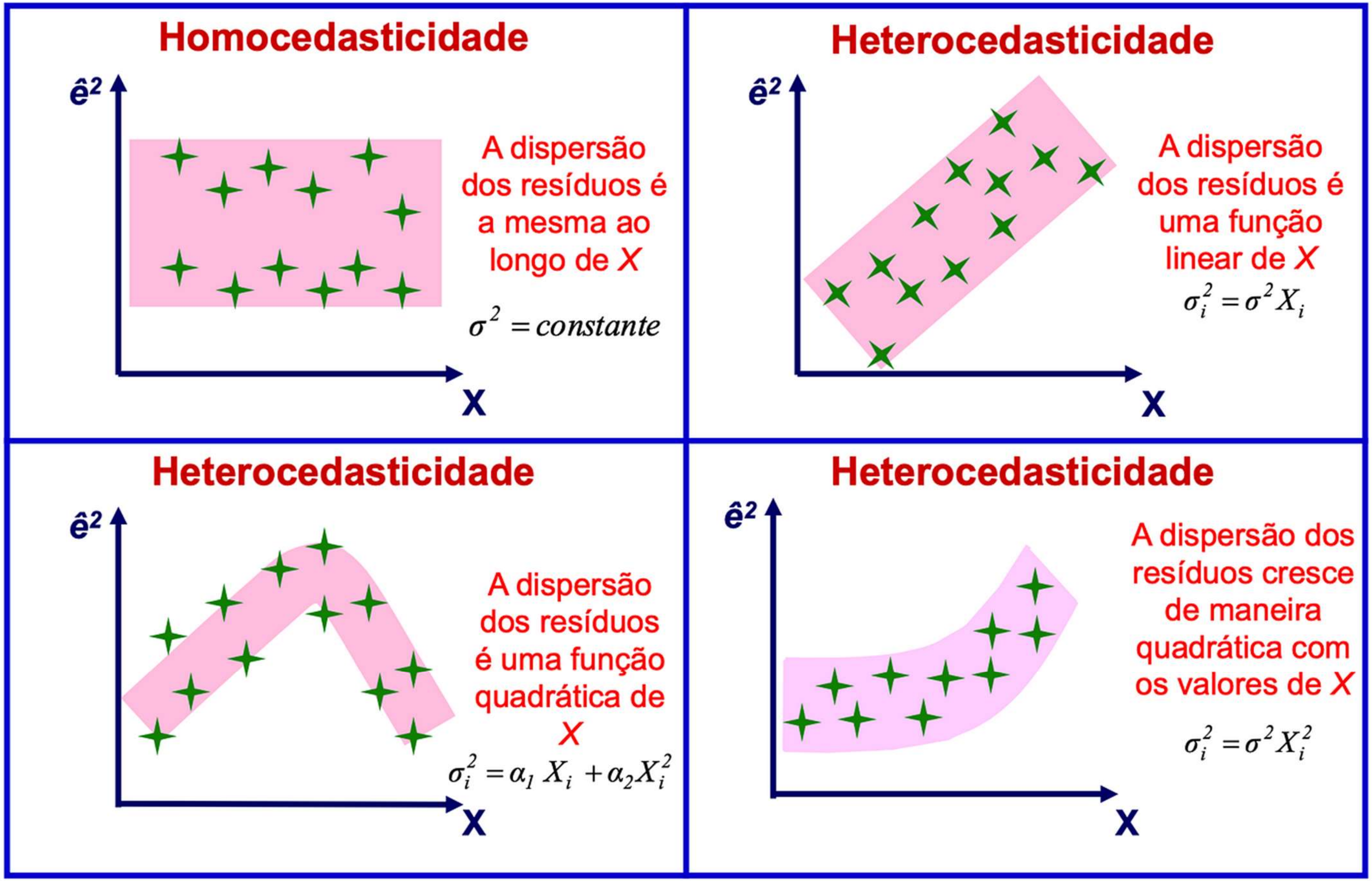
ANÁLISE DA HOMOCEDASTICIDADE DOS RESÍDUOS

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

Teste Breusch-Pagan (Homocedasticidade)

Ho = existe homocedasticidade : $p > 0.05$

Ha = não existe homocedasticidade : $p <= 0.05$


Homocedasticidade
ê²
A dispersão dos resíduos é a mesma ao longo de X
$\sigma^2 = constante$
X
Heterocedasticidade
ê²
A dispersão dos resíduos é uma função linear de X
$\sigma_i^2 = \sigma^2 X_i$
X
Heterocedasticidade
ê²
A dispersão dos resíduos é uma função quadrática de X
$\sigma_i^2 = \alpha_1 X_i + \alpha_2 X_i^2$
X
Heterocedasticidade
ê²
A dispersão dos resíduos cresce de maneira quadrática com os valores de X
$\sigma_i^2 = \sigma^2 X_i^2$
X

PUC Minas

TESTE - T

Avaliando a significância de cada parâmetro β do modelo

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

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

$$H_0 : p - valor \geq 0{,}05$$
$$H_1 : p - valor < 0{,}05$$

PUC Minas

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

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

PUC Minas

TESTE – T – ESTATÍSTICA DE TESTE

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

A estatística do teste segue uma distribuição t de Student.

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

PUC Minas

TESTE – T – ESTATÍSTICA DE TESTE

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

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

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.

PUC Minas

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.

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

H_1: Pelo menos um β_j é diferente de zero

$$H_0 = F_{Calc} \leq F_{Crítico} \; ou \; p - valor \geq 0{,}05$$
$$H_1 = F_{Calc} > F_{Crítico} \; ou \; p - valor < 0{,}05$$

PUC Minas

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.

PUC Minas

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

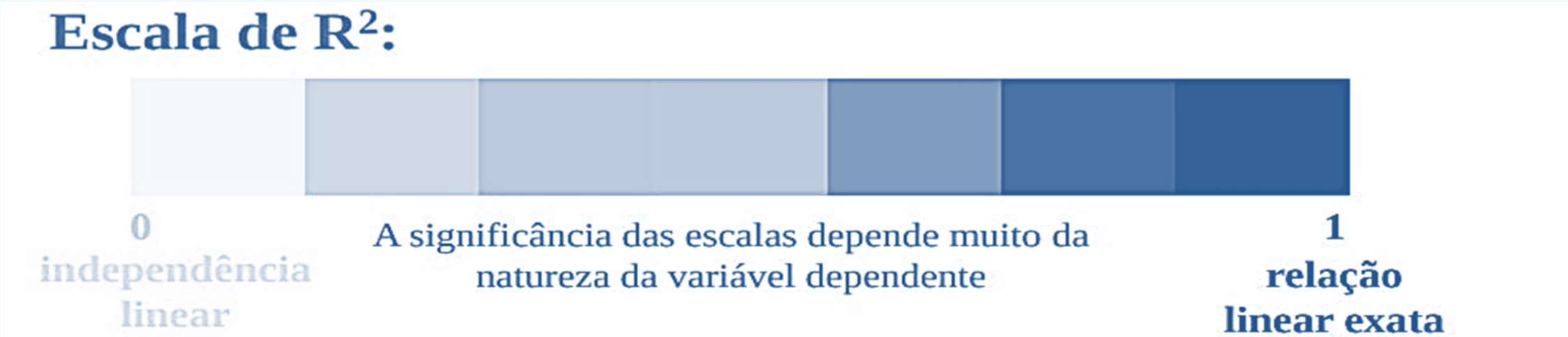
PUC Minas

COEFICIENTE DE DETERMINAÇÃO (R²)

Como avaliar o modelo?

O **coeficiente de determinação (R²)** estima a proporção da variabilidade da variável dependente (Y) que é explicada pelas(s) variáveis independente do modelo de regressão.

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} \widehat{e_i}^2}{\sum_{i=1}^{n} y_i^2} = \widehat{\beta_1}^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$



```
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

Coeficientes estimados

Teste -t

```
Coefficients:
            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 ***
---
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,	Adjusted R-squared:  0.8499 
F-statistic: 51.98 on 1 and 8 DF,  p-value: 9.161e-05
```

R^2

Teste F

OLS Regression Results

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

R^2

Teste F

Coeficientes

	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