

Fundamentos de Data Science, Data Mining e Análise Preditiva

Especialização em Ciência de Dados com Big Data, BI e Data
Analytics



Prof. Dr. Carlos Barros

Introdução a Correlação e Regressão

VARIÁVEIS

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

- EXISTE UMA RELAÇÃO MATEMÁTICA ENTRE ESTAS DUAS VARIÁVEIS?
- SE EXISTE, COMO POSSO MEDIR SUA FORÇA?
- PODERIA USAR ESSA RELAÇÃO PARA FAZER PREVISÕES?

GRÁFICO DE DISPERSÃO

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

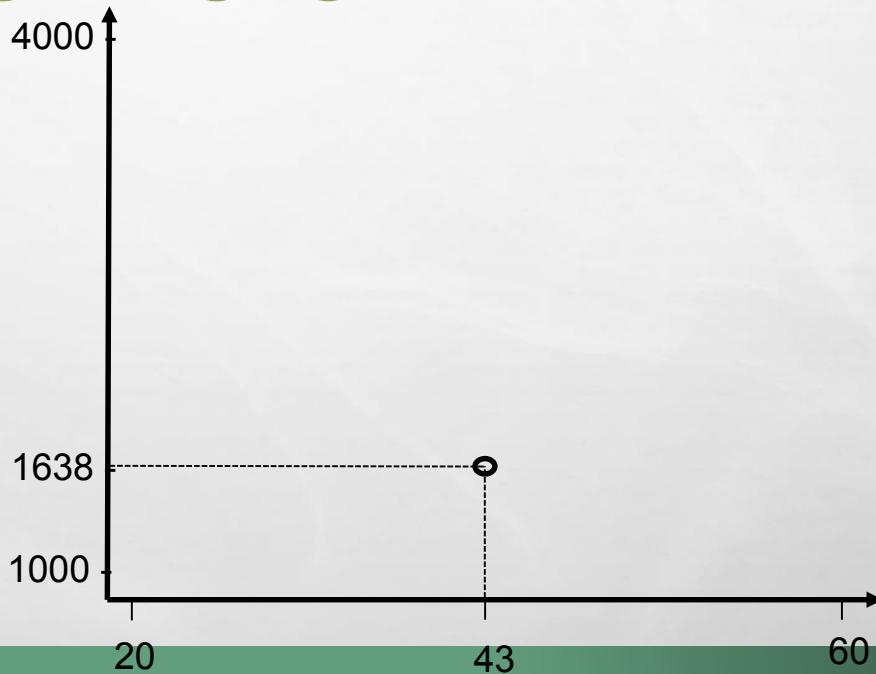
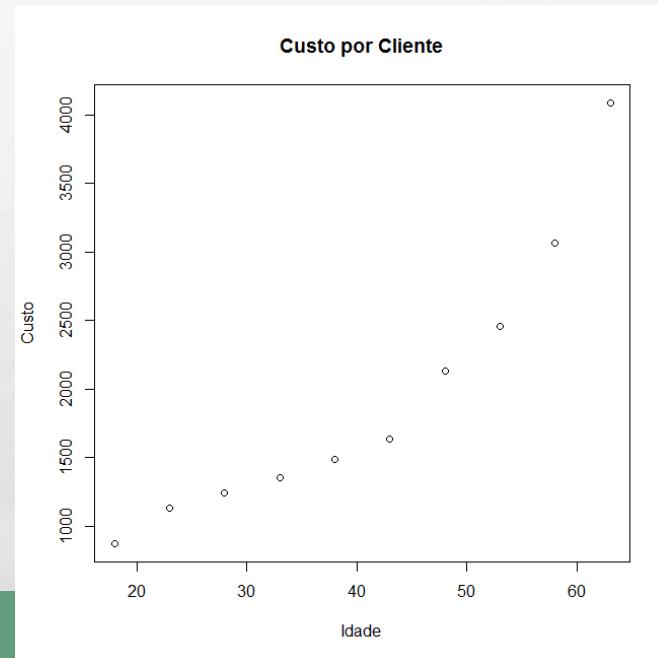


GRÁFICO DE DISPERSÃO

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090



PLANO CARTESIANO

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

Eixo Y (Vertical)
Variável de Resposta
Ou Dependente

Na regressão é o que queremos Prever

Qual vai ser o custo para o plano de saúde de um paciente com 45 anos de idade?

Custo

Y

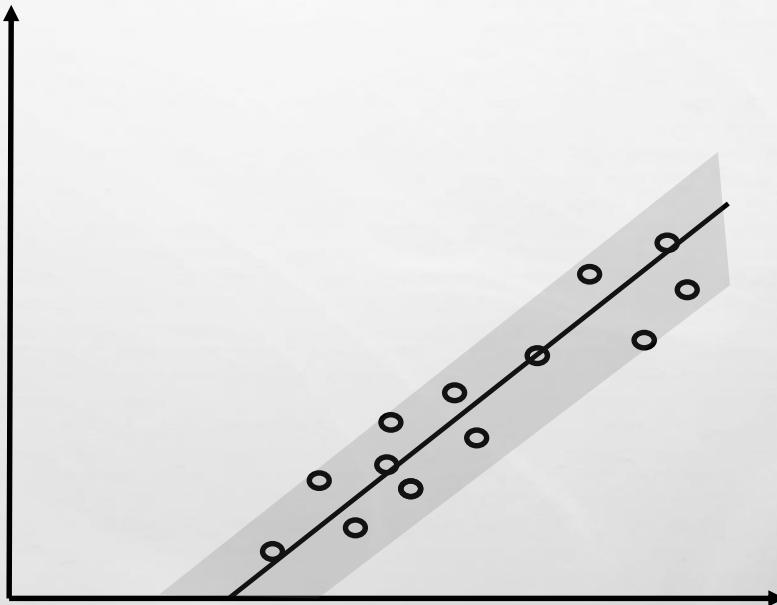
Eixo X (Horizontal)
Variável Explanatória
Ou Independente

Idade

X

Na regressão é o que explica, ou usamos para prever

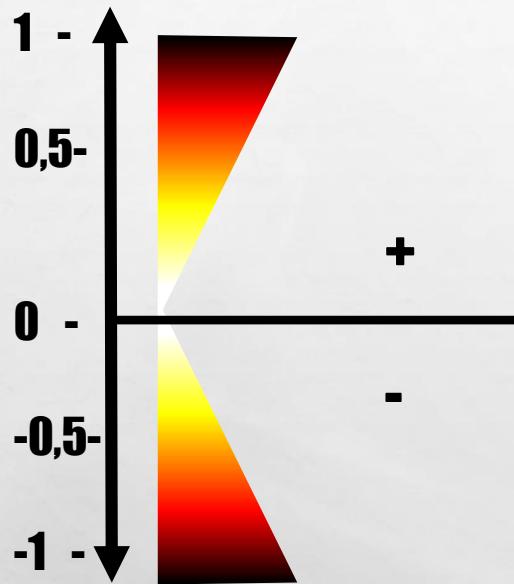
REGRESSÃO LINEAR



CORRELAÇÃO (R)

- MOSTRA A FORÇA E A DIREÇÃO DA RELAÇÃO ENTRE VARIÁVEIS
- PODE SER UM VALOR ENTRE -1 E 1
- A CORRELAÇÃO DE $A \sim B$ É A MESMA QUE $B \sim A$

FORÇA E DIREÇÃO

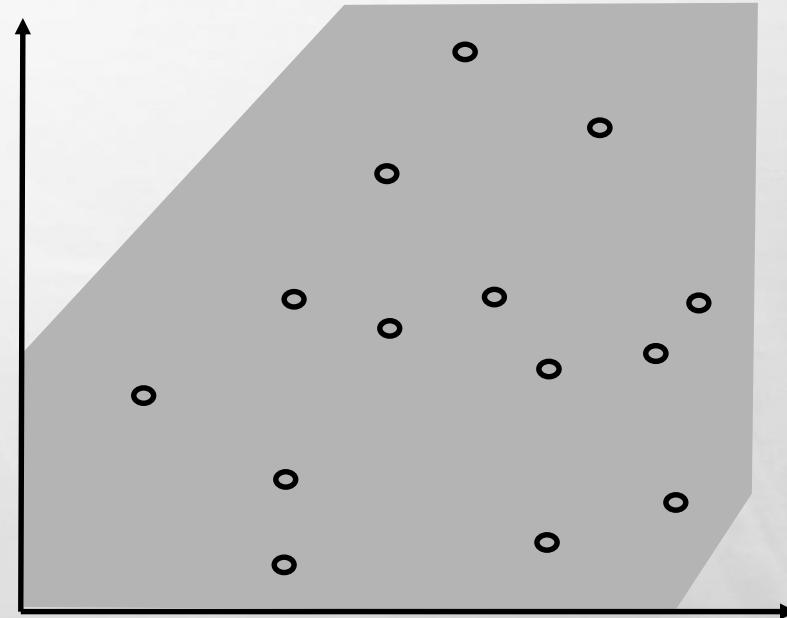
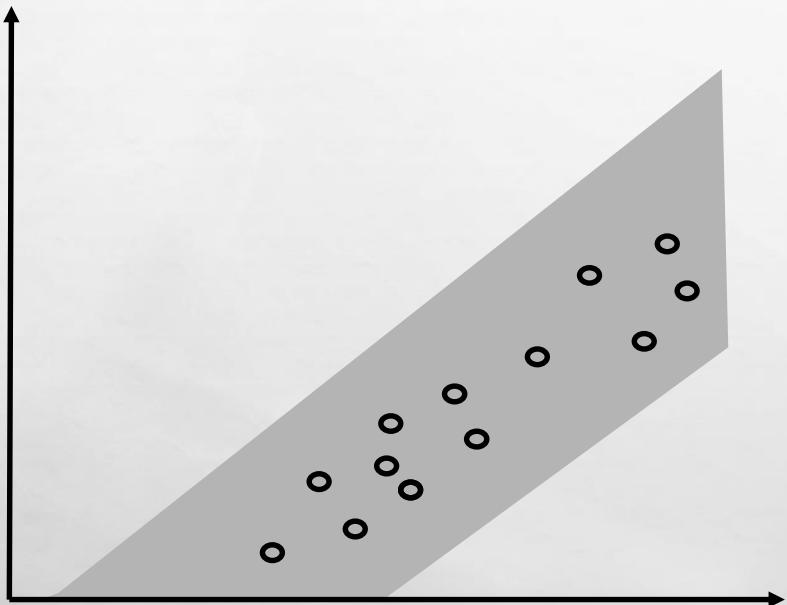


1 →	Perfeita
0,7 →	Forte
0,5 →	Moderada
0,25 →	Fraca
0 →	Inexistente
-0,25 →	Fraca
-0,5 →	Moderada
-0,7 →	Forte
-1 →	Perfeita

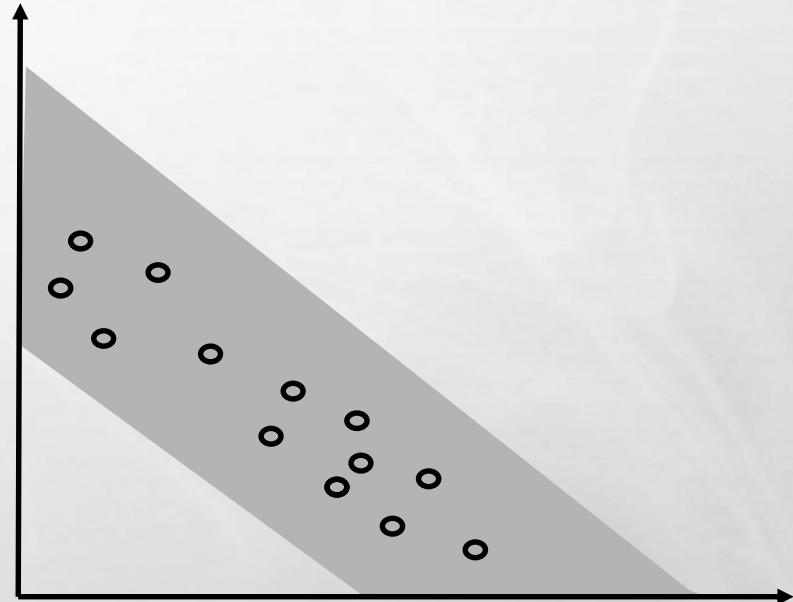
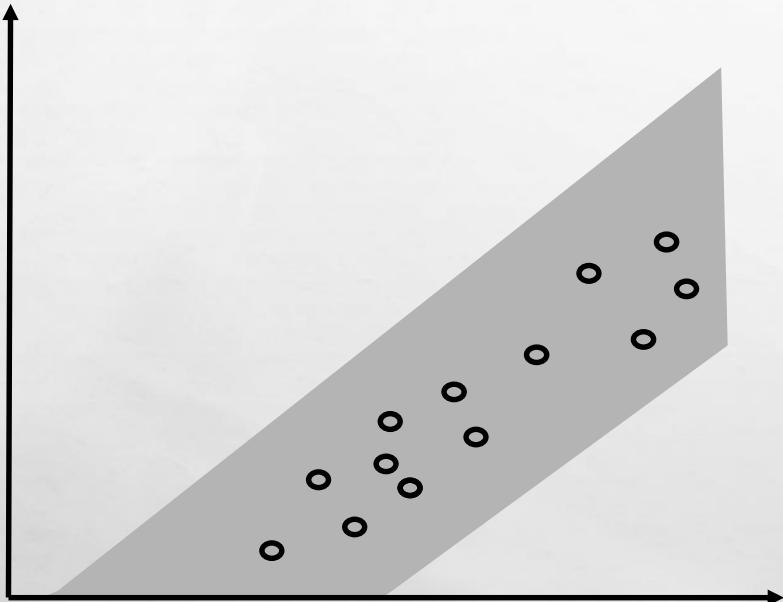
EXEMPLOS

- **1: POSITIVA PERFEITA**
- **-0,8: NEGATIVA FORTE**
- **0,23: POSITIVA FRACA**
- **0,09: POSITIVA FRACA**
- **-0,334 NEGATIVA FRACA**
- **0: INEXISTENTE**
- **0,6: POSITIVA MODERADA**
- **1,2: ERRO**

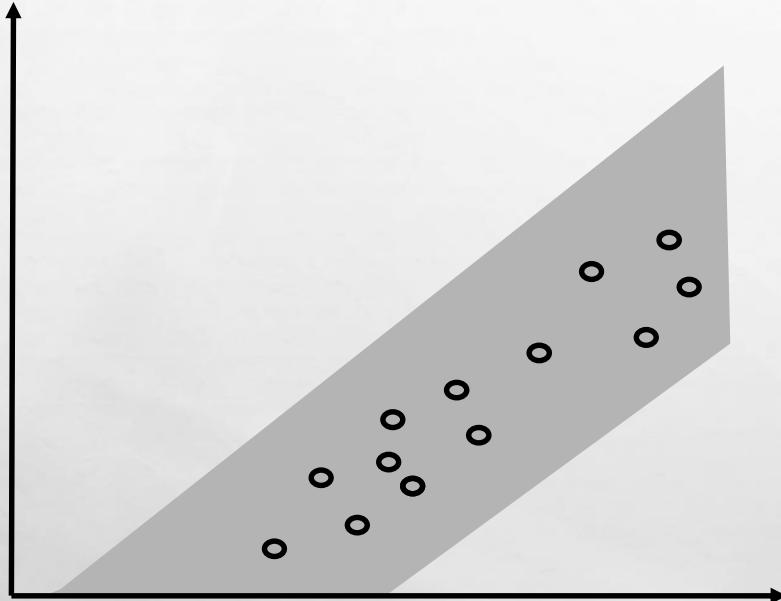
FORTE - FRACA



POSITIVA - NEGATIVA

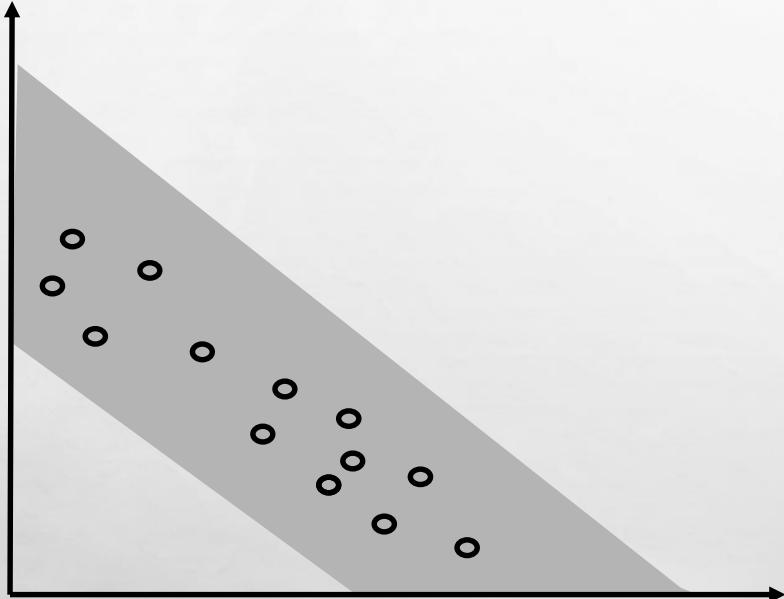


POSITIVA



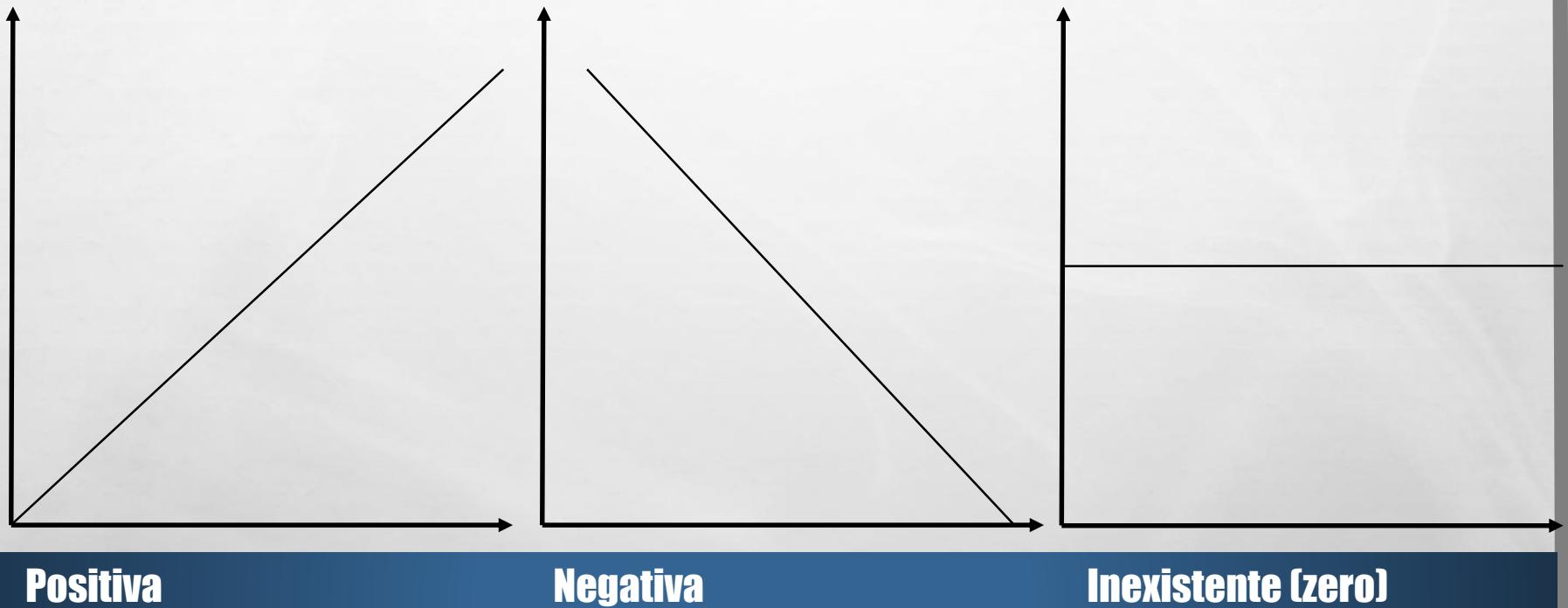
Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

NEGATIVA



Preco	Vendas
100	524
105	521
109	523
111	518
114	505
115	506
117	503
120	499

POSSIBILIDADES



COEFICIENTE DE DETERMINAÇÃO (R^2)

- MOSTRA O QUANTO O MODELO CONSEGUE EXPLICAR OS VALORES
- QUANTO MAIOR, MAIS EXPLICATIVO ELE É
- O RESTANTE DA VARIABILIDADE ESTÁ EM VARIÁVEIS NÃO INCLUÍDAS NO MODELO
- VARIA ENTRE ZERO ATÉ 1 (SEMPRE POSITIVO)
- CALCULA-SE COM O QUADRADO DO COEFICIENTE DE CORRELAÇÃO (R)

COEFICIENTE DE DETERMINAÇÃO (R^2)

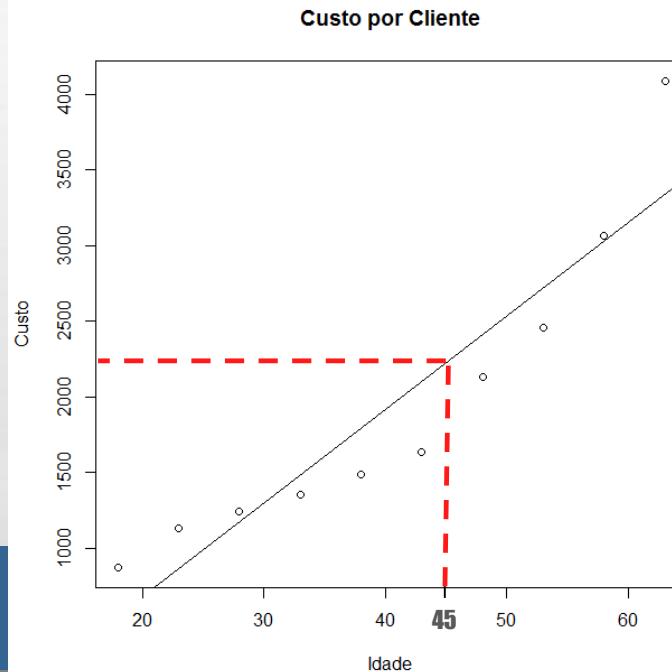
Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

- CORRELAÇÃO: 0,93
- R^2 : 0,86
- 86% DA VARIÁVEL DEPENDENTE CONSEGUE SER EXPLICADA PELAS VARIÁVEIS EXPLANATÓRIAS PRESENTES NO MODELO

PREVISÃO

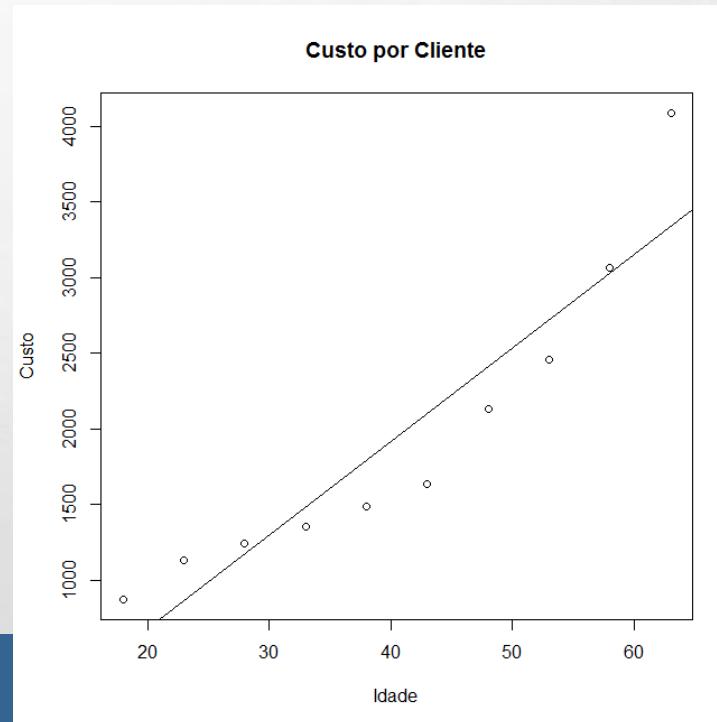
Qual vai ser o custo de um cliente com 45 anos de idade?

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090



COMO A LINHA É CONSTRUÍDA?

- Ponto de Encontro da Linha no Eixo Y (interseção) : $X=0$
- Inclinação: a cada unidade que aumenta a variável Independente (x), a variável de resposta (y) sobe o valor da inclinação
- Planilhas e Ferramentas Estatísticas calculam estes valores automaticamente



DADOS DE EXEMPLO

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

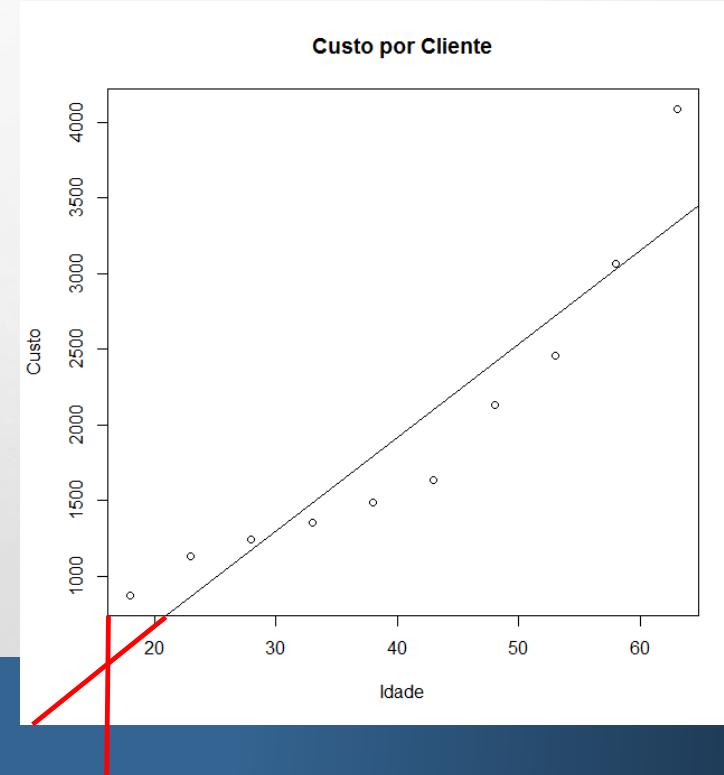
- Intersecção: -558,94
- Inclinação: 61,86

Previsão:

33 anos: 1356

34 anos: $1356 + 61,86$

= 1417,86

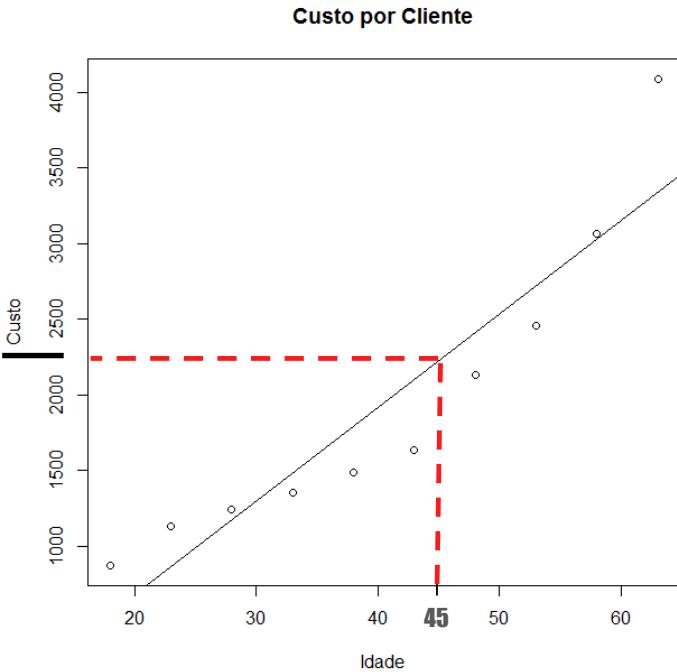


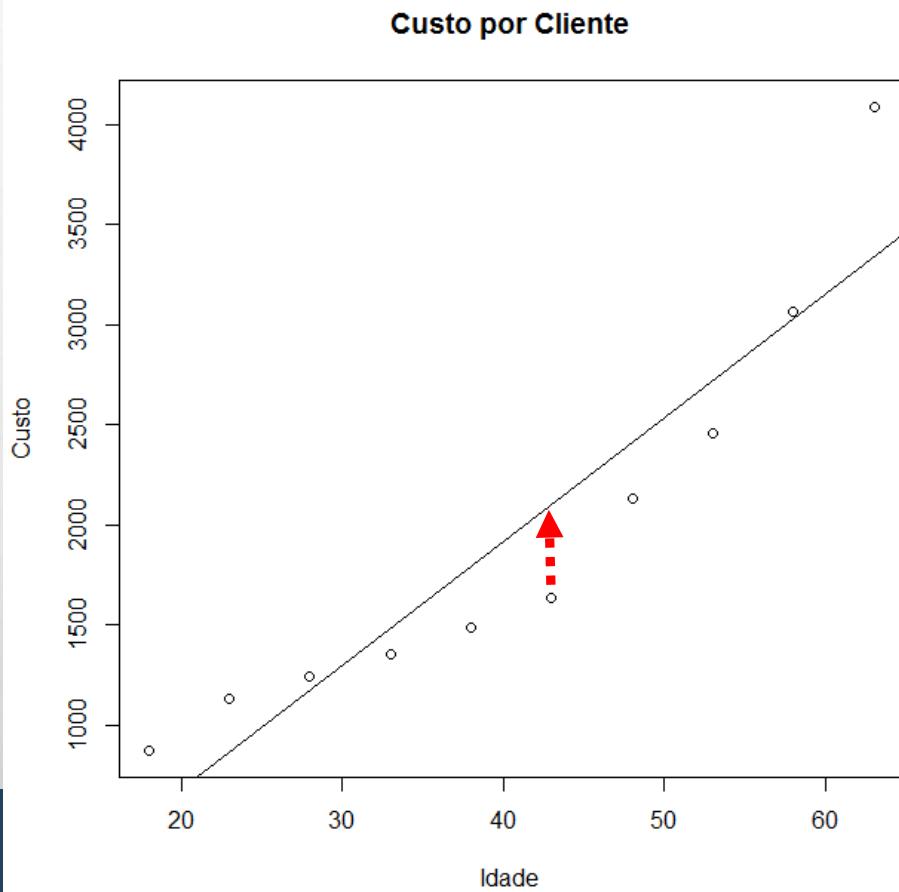
COMO PREVER?

- PREVISÃO = INTERSECÇÃO + (INCLINAÇÃO * VALOR A PREVER)
- QUANTO VAI CUSTAR UM CLIENTE COM 56 ANOS DE IDADE?
- $X = -558,94 + (61,86 * 56)$
- $X = \underline{2905,22}$
- QUALQUER SOFTWARE EXECUTA O CÁLCULO AUTOMATICAMENTE

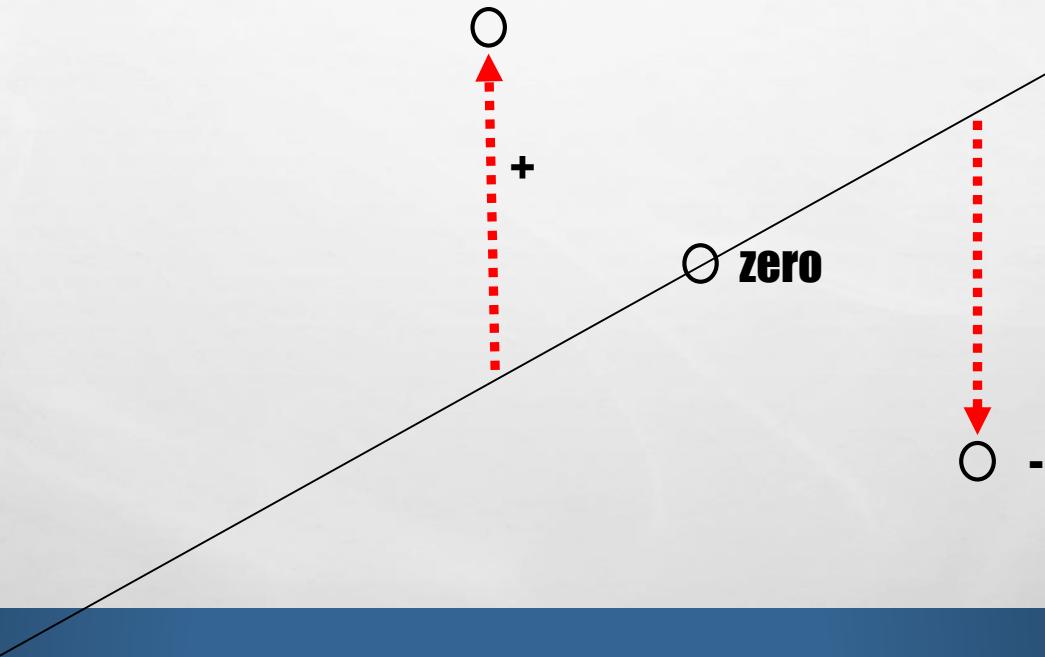
Régressão Linear - Residuais

2225

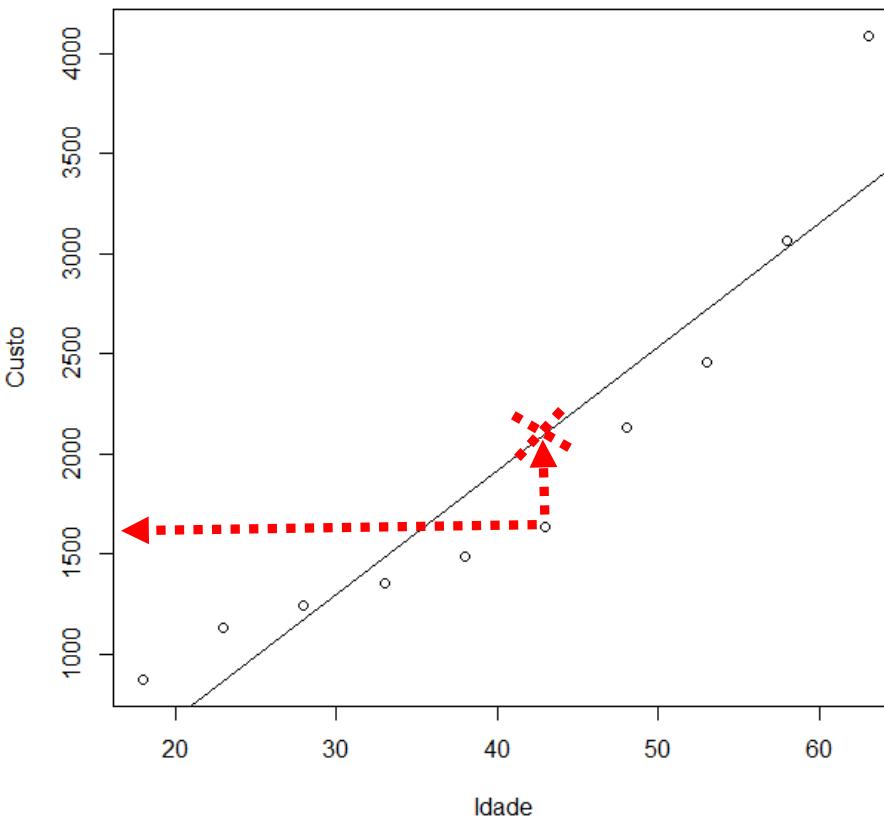




Valor ajustado = 1500
Residual = -500

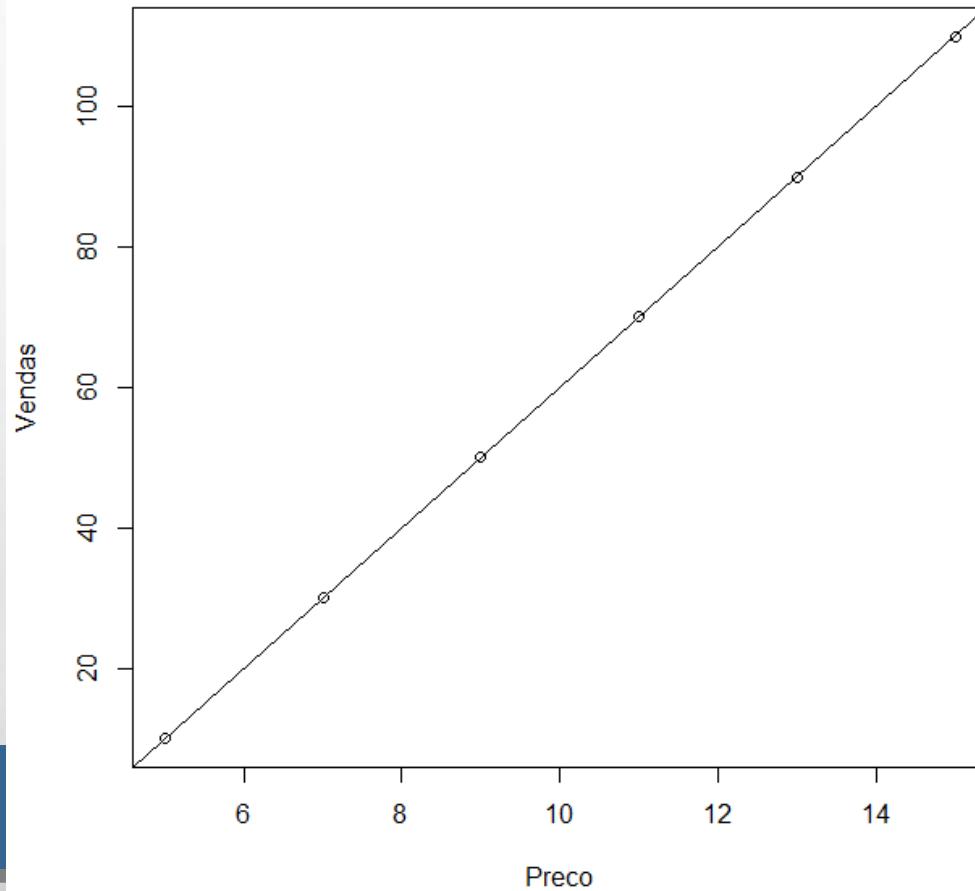


Custo por Cliente



Preco	Vendas
5	10
7	30
9	50
11	70
13	90
15	110

Soma de Residuais= 0
R (Correlação)= 1

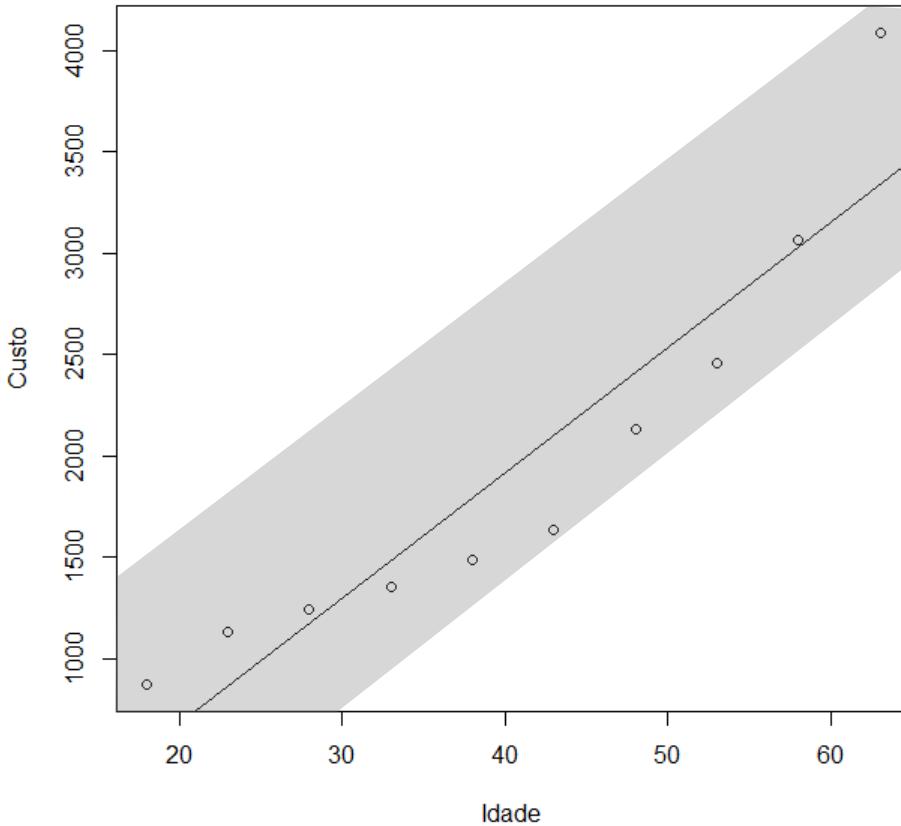


Régressão Linear - Outliers, Extrapolação, Correlação não é causa

OUTLIERS

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

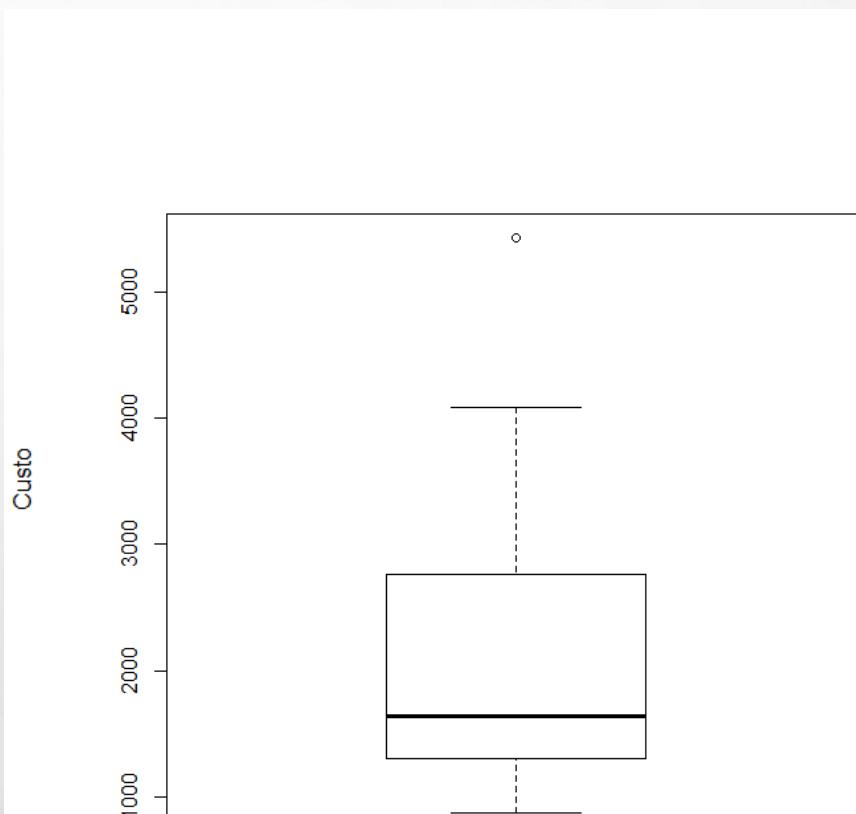
CORRELAÇÃO= 0,93



OUTLIERS

Idade	Custo
18	871
23	1132
24	5435
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

- CORRELACÃO= 0,34

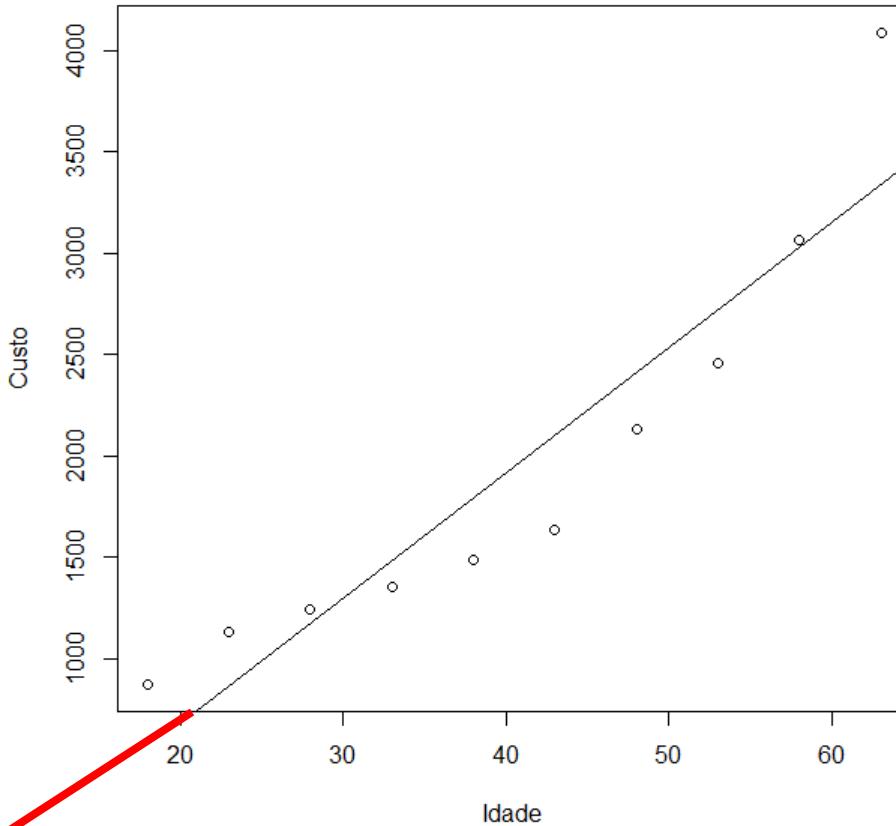


EXTRAPOLAÇÃO

16 = ?

Idade	Custo
18	871
23	1132
28	1242
33	1356
38	1488
43	1638
48	2130
53	2454
58	3066
63	4090

80 = ?



CORRELAÇÃO NÃO É CAUSA

Pessoas com mais treinamento tem melhor performance

Ou será que:

Elas receberam treinamento porque performaram melhor?

Candidatos vistos como carismáticos obtém mais votos

Ou será que:

Candidatos mais votados são vistos como carismáticos?

SIMPLES E MÚLTIPLA

- **SIMPLES**
 - UMA VARIÁVEL EXPLANATÓRIA PARA PREVER UMA VARIÁVEL DEPENDENTE
 - $Y \sim X$
- **MÚLTIPLA**
 - DUAS OU MAIS VARIÁVEIS EXPLANATÓRIAS PARA PREVER UMA VARIÁVEL DEPENDENTE
 - $Y \sim X_1 + X_2 + X_N$

ANALISAR CADA X COM Y

- **ANALISAR CADA VARIÁVEL INDEPENDENTE COM Y INDIVIDUALMENTE**
- **GERAR GRÁFICOS DE DISPERSÃO INDIVIDUAIS**
- **BUSCAR REDUNDÂNCIAS (MESMOS EFEITOS DE X SOBRE Y): EXPLICAÇÃO POSTERIOR**

COEFICIENTE DE DETERMINAÇÃO (R^2)

- LEMBRANDO QUE R^2 É O PERCENTUAL DE VARIAÇÃO DA VARIÁVEL DE RESPOSTA QUE É EXPLICADA PELO MODELO
- QUANDO SE COLOCAM MAIS VARIÁVEIS NO MODELO, A TENDÊNCIA É QUE R^2 AUMENTE, MESMO QUE A ADIÇÃO DA VARIÁVEL NÃO AUMENTE A PRECISÃO DO MODELO
- PARA ISSO, UTILIZA-SE R^2 AJUSTADO, QUE AJUSTA A VARIAÇÃO DO MODELO DE ACORDO COM O NUMERO DE VARIÁVEIS INDEPENDENTES QUE É INCLUÍDA NO MODELO
- R^2 AJUSTADO VAI SER SEMPRE MENOR QUE R^2

COLINEARIDADE E PARCIMÔNIA

- **COLINEARIDADE: DUAS VARIÁVEIS INDEPENDENTES QUE SÃO CORRELACIONADAS**
- **INCLUIR VARIÁVEIS INDEPENDENTES COLINEARES PODE PREJUDICAR O MODELO, CRIANDO PREVISÕES NÃO CONFIÁVEIS**
- **PARCIMÔNIA: NÃO COLOCAR VARIÁVEIS QUE NÃO MELHOREM O MODELO EM NADA: CRIAR MODELOS PARCIMONIOSOS**

REQUISITOS BÁSICOS

- 1. LINEARIDADE ENTRE A VARIÁVEL DEPENDENTE E AS VARIÁVEIS INDEPENDENTES**
- 2. QUE AS VARIÁVEIS SEJAM NORMALMENTE DISTRIBUÍDAS**
- 3. POUCA OU NENHUMA COLINEARIDADE**

RESIDUAIS

- PRÓXIMOS A DISTRIBUIÇÃO NORMAL
- VARIÂNCIA CONSTANTE EM RELAÇÃO A LINHA DE MELHOR AJUSTE
- INDEPENDENTES (SEM PADRÃO)

FÓRMULAS

Correlação

Inclinação

Interceptação

Previsão

CORRELAÇÃO DE PEARSON

Idade	Custo
18	871
23	1100
25	1393
33	1654
34	1915
43	2100
48	2356
51	2698
58	2959
63	3000
67	3100

$$r = \frac{cov(X, Y)}{\sqrt{var(x).var(y)}}$$

cov: covariância

var: variância

$$r = \frac{11869,71}{\sqrt{(255,5371).(564932,6942)}}$$

$$r = \frac{11869,71}{\sqrt{144361313,3}}$$

$$r = \frac{11869,71}{12015,04529}$$

$$\mathbf{r = 0,9879}$$

INCLINAÇÃO

Idade	Custo
18	871
23	1100
25	1393
33	1654
34	1915
43	2100
48	2356
51	2698
58	2959
63	3000
67	3100

$$m = r \left(\frac{S_y}{S_x} \right)$$

r: correlação (0,9879)

s: desvio padrão

$$m = 0,9879 \left(\frac{751,6200}{15,9855} \right)$$

m = 46,45

INTERCEPTAÇÃO

Idade	Custo
18	871
23	1100
25	1393
33	1654
34	1915
43	2100
48	2356
51	2698
58	2959
63	3000
67	3100

$$b = \bar{y} - m\bar{x}$$

$$b = 2104,182 - 46,45 * 42,09$$

\bar{y} : média de y

\bar{x} : média de x

$$\mathbf{b = 149,0577}$$

m : Inclinação (46,45)

PREVISÃO

Idade	Custo
18	871
23	1100
25	1393
33	1654
34	1915
43	2100
48	2356
51	2698
58	2959
63	3000
67	3100

$$P = b + (m * v)$$

b: interceptação (149,0577)

m: inclinação (46,45)

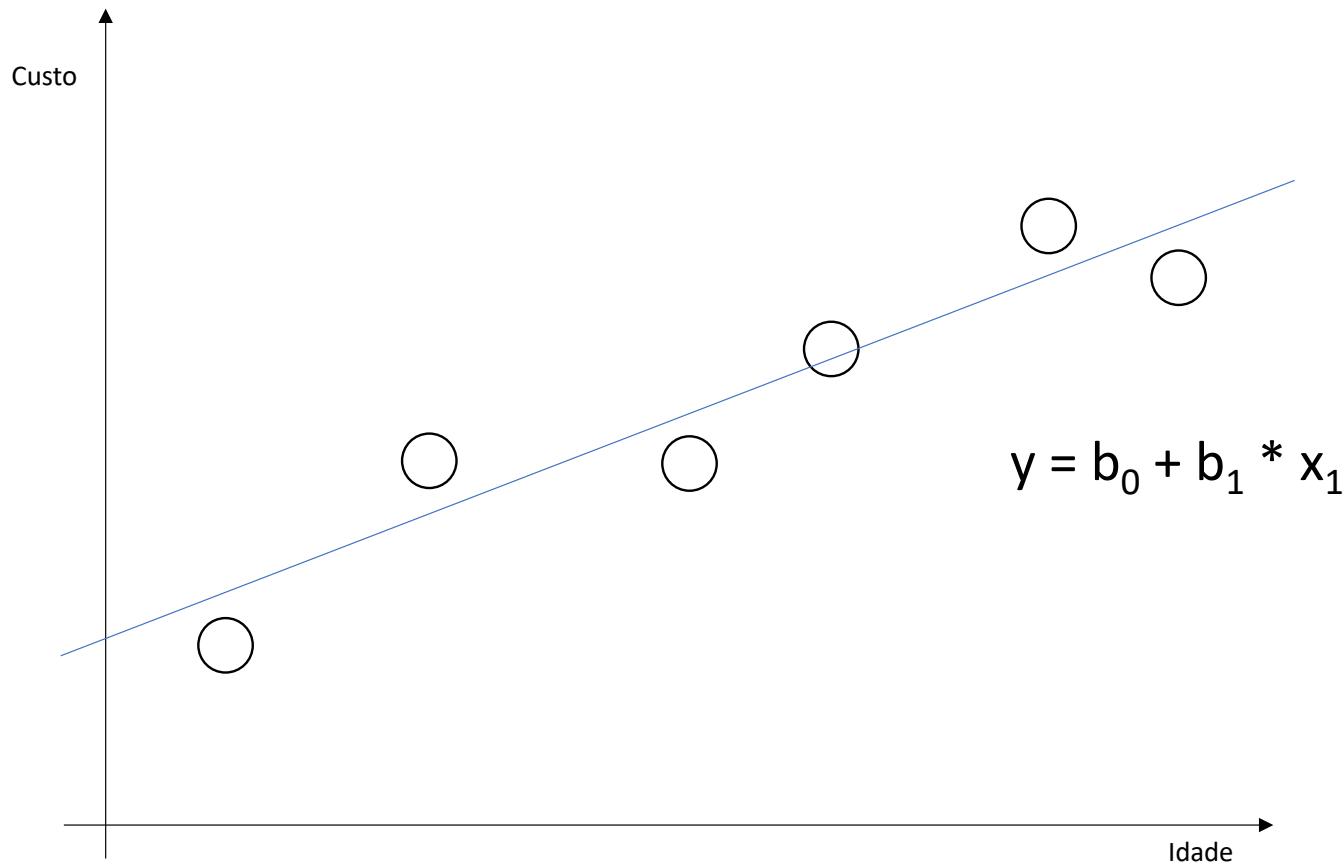
v: variável dependente

$$v = 54 \text{ anos}$$

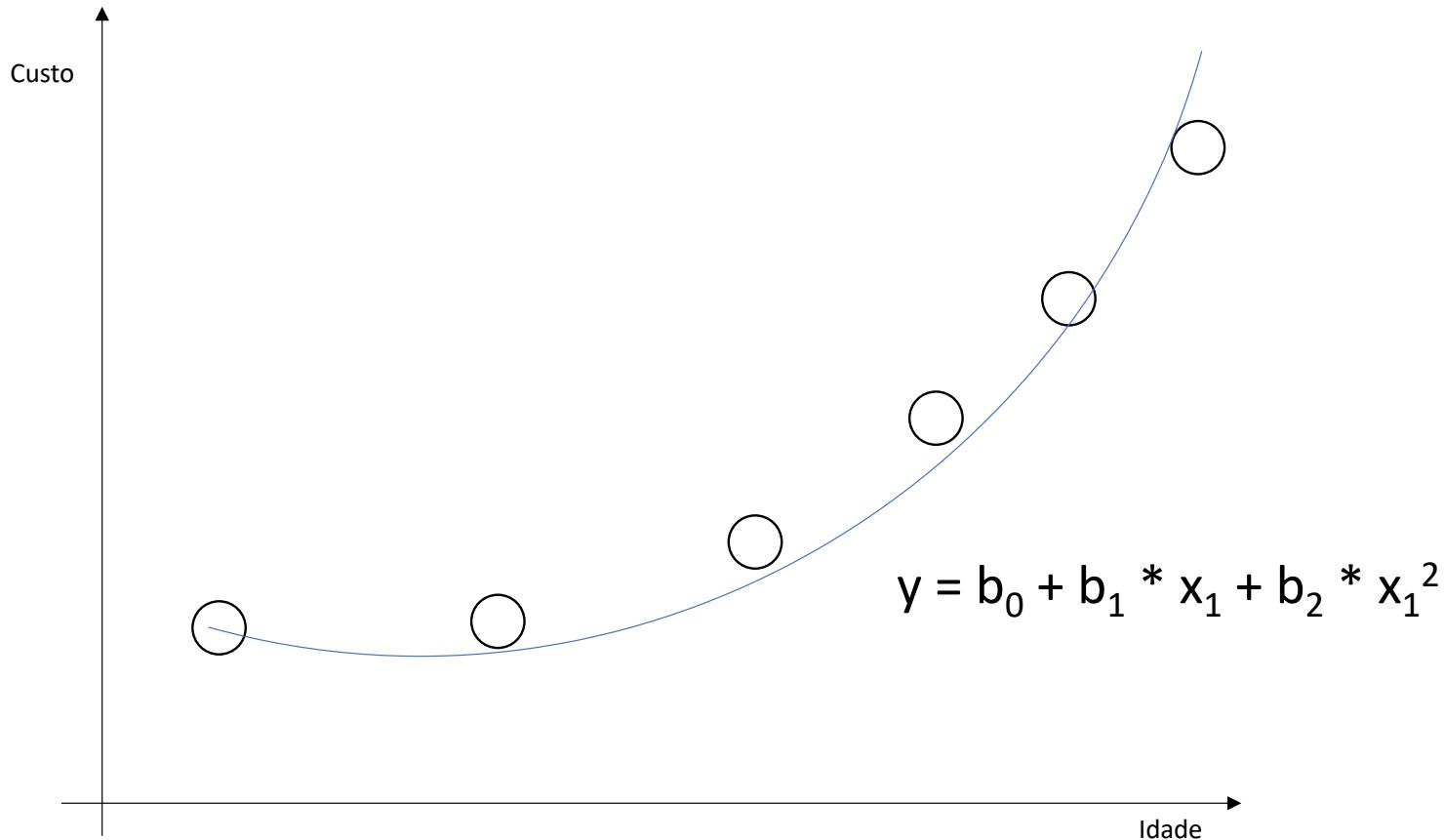
$$p = 149,0577 + (46,45 * 54)$$

$$p = 2657,355$$

Regressão linear simples



Régressão linear polinomial



Regressão Logística

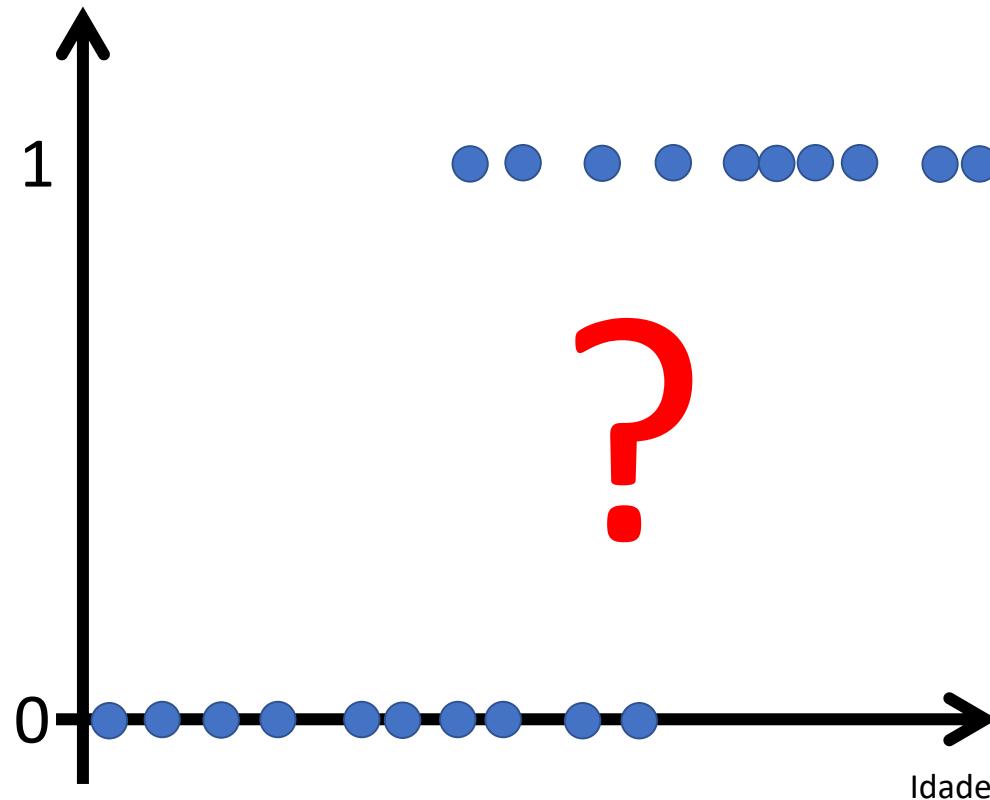
- Semelhante a regressão Linear, porém a variável de resposta é binária: sucesso ou fracasso
 - 1: sucesso
 - 0: fracasso
- O sucesso ou fracasso é representado através de probabilidade
- Também pode ser simples ou múltipla

Regressão Logística

	A	B	C
1	CANDIDATO	SITUACAO	DESPESAS
2	George Turner	0	10
3	Victor Johnson	0	100
4	Jerry Perry	1	1600
5	Shirley Cook	1	1500
6	Carolyn Bailey	1	3300
7	Susan Sanders	0	200
8	Anthony Harris	1	1800
9	Philip Richardson	1	1700
10	Eugene Phillips	0	300
11	Mildred Morris	1	1800
12	Richard Jones	0	100
13	Joan Hernandez	0	500
14	Lawrence Mitchell	1	3000
15	Annie Brooks	0	20
16	Stephen Simmons	0	200
17	Samuel Russell	1	700
18	Jason Brown	1	1600
19	Bobby Gonzalez	1	1900
20	Steven Coleman	0	100
21	Benjamin Ramirez	0	400

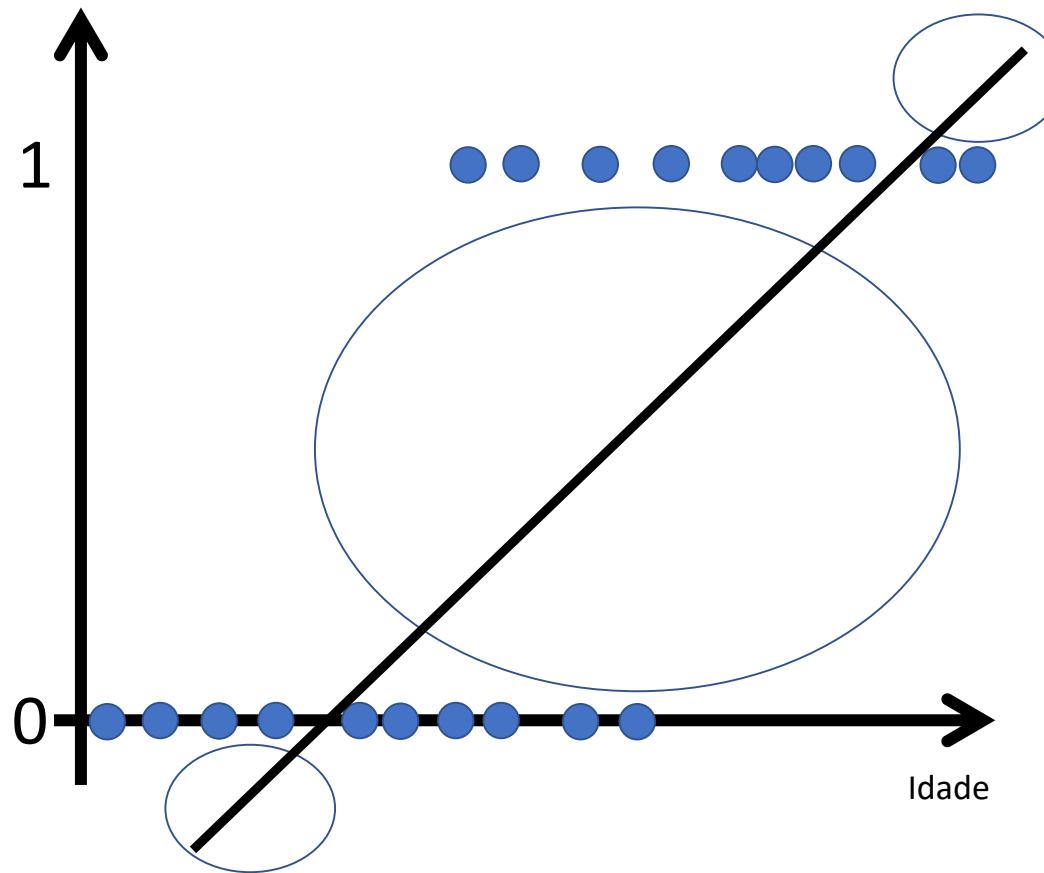
	A	B	C
1	CANDIDATO	DESPESAS	
2	A	0	
3	B	10	
4	C	200	???
5	D	500	
6	E	900	
7	F	1500	
8	G	3000	

Pagar (S/N)



Pagar (S/N)

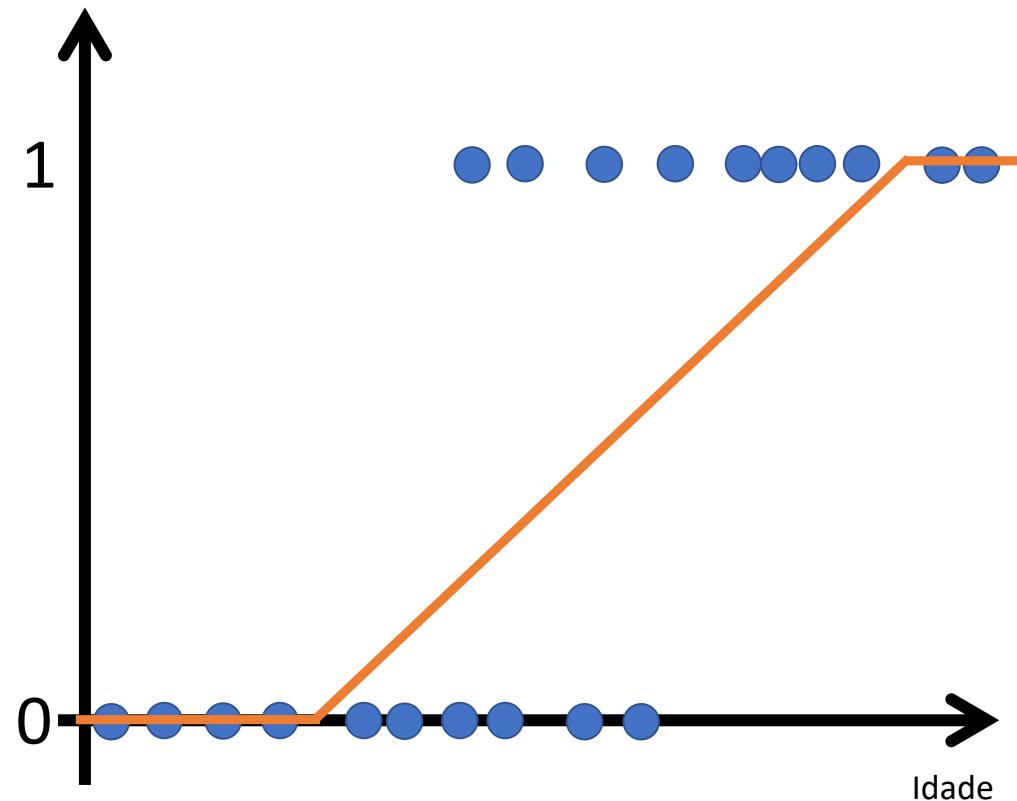
Existe algum tipo de correlação entre a idade e o pagamento?
Predizer uma probabilidade
Quanto maior a idade, maior a probabilidade de pagar



Pagar (S/N)

Função sigmoide

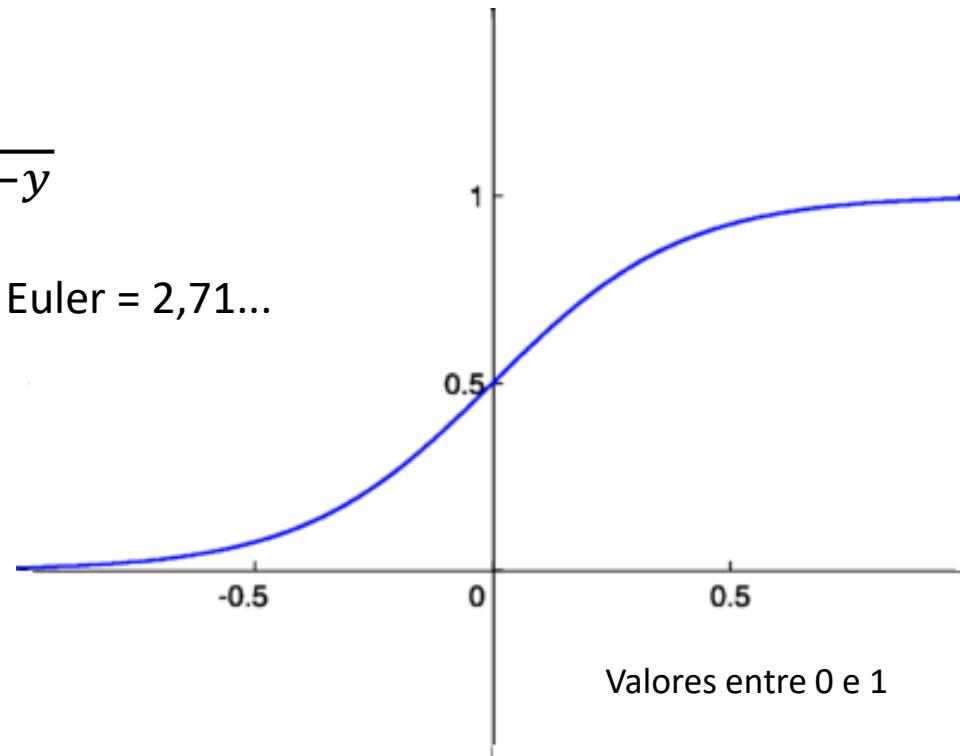
Encontrar a melhor linha para encaixar nos dados



Sigmoid (função sigmoide)

$$p = \frac{1}{1 + e^{-y}}$$

e = número de Euler = 2,71...



Valores entre 0 e 1

Se p for alto o valor será aproximadamente 1

Se p for pequeno o valor será aproximadamente 0

Não retorna valores negativos

Métricas de Erros

Previsão de valores numéricos (reais, inteiros)

Métricas diferentes da previsão de categorias

Uso:

- Regressão clássica
- Regressão ML
- Series Temporais
- Etc.

Mean Erro (ME)

Dependente de Escala

A média da diferença entre realizado e previsto

Previsto	Realizado	Dif.
3,34	3,00	-0,34
4,18	4,00	-0,18
3,00	3,00	0
2,99	3,00	0,01
4,51	4,50	-0,01
5,18	4,00	-1,18
8,18	4,50	-3,68

$$MAE = \sum_{i=1}^N \frac{p_i - t_i}{n}$$

$$ME = \frac{-5,38}{7} = -0,76$$

Mean Absolute Erros (MAE)

Dependente de Escala

A média da diferença absoluta entre o realizado e o previsto

Previsto	Realizado	Dif. Absoluta
3,34	3,00	0,34
4,18	4,00	0,18
3,00	3,00	0
2,99	3,00	0,01
4,51	4,50	0,01
5,18	4,00	1,18
8,18	4,50	3,68
		5,4

$$\text{MAE} = \sum_{i=1}^N \frac{|p_i - t_i|}{n}$$

$$\text{MAE} = \frac{5,4}{7} = 0,77$$

Root Mean Squared Error (RMSE)

Independente de Escala

O desvio padrão da amostra da diferença entre o previsto e o teste

Previsto	Realizado	Dif. ao Quad.
3,34	3,00	0,1156
4,18	4,00	0,0324
3,00	3,00	0
2,99	3,00	1E-04
4,51	4,50	1E-04
5,18	4,00	1,3924
8,18	4,50	13,5424

$$RMSE = \sqrt{\frac{\sum_{I=1}^N (p_i - t_i)^2}{N}}$$

$$RMSE = \sqrt{\frac{15,083}{7}}$$

$$RMSE = 1,46$$

Mean Percentage Error (MPE)

Independente de Escala (%)

Diferença percentual de erro

Previsto	Realizado	Erro %
3,34	3,00	-11,3333
4,18	4,00	-4,5
3,00	3,00	0
2,99	3,00	0,333333
4,51	4,50	-0,22222
5,18	4,00	-29,5
8,18	4,50	-81,7778

$$MPE = \frac{\sum_{I=1}^N \frac{(t_i - p_i)}{t_i - 100}}{N}$$

$$MPE = \frac{-127}{7}$$

$$MPE = -18,14$$

Mean Absolute Percentage Error (MAPE)

Independente de Escala (%)

Diferença absoluta percentual de erro

Previsto	Realizado	Erro abs.	Erro % abs.
3,34	3,00	0,1156	0,1133333
4,18	4,00	0,0324	0,045
3,00	3,00	0	0
2,99	3,00	1E-04	0,0033333
4,51	4,50	1E-04	0,0022222
5,18	4,00	1,3924	0,295
8,18	4,50	13,5424	0,8177778

%

$$\text{MAPE} = \frac{\sum_{i=1}^N \frac{|p_i - t_i|}{|t_i|}}{N}$$

$$\text{MAPE} = \frac{1,2766667}{7}$$

$$\text{MAPE} = 0,18$$

Previsto	Realizado	Diferença	Dif. Abs.	Dif. Quad.	Erro %	Erro % abs
3,34	3	-0,34	0,34	0,1156	-11,3333	11,33333
4,18	4	-0,18	0,18	0,0324	-4,5	4,5
3	3	0	0	0	0	0
2,99	3	0,01	0,01	1E-04	0,33333	0,333333
4,51	4,5	-0,01	0,01	1E-04	-0,22222	0,222222
5,18	4	-1,18	1,18	1,3924	-29,5	29,5
8,18	4,5	-3,68	3,68	13,5424	-81,7778	81,77778

ME	-0,76857
MAE	0,77143
RMSE	1,46789
MPE	-18,1429
MAPE	18,2381