

Introdução à regressão linear



Plataforma completa de aprendizado
contínuo em programação.

#BoostingPeople

rocketseat.com.br

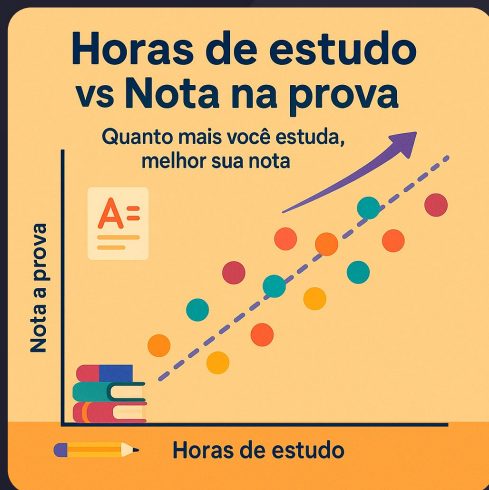
Todos os direitos reservados © Rocketseat S.A.

Será que dá pra prever vendas com base em propaganda?

O QUE É REGRESSÃO?

- Uma forma de modelar a relação entre duas variáveis.
- Prever o valor de Y com base em X.
- Muito usada em negócios, economia, marketing, operações.

Exemplo do cotidiano



A regressão tenta colocar uma "reta" que explica essa tendência.

Por que aprender regressão?

- Ajuda a tomar decisões baseadas em dados.
- Previsão de custos, demanda, vendas, tempo.
- É base para modelos mais avançados (ML, AI).

A equação da reta

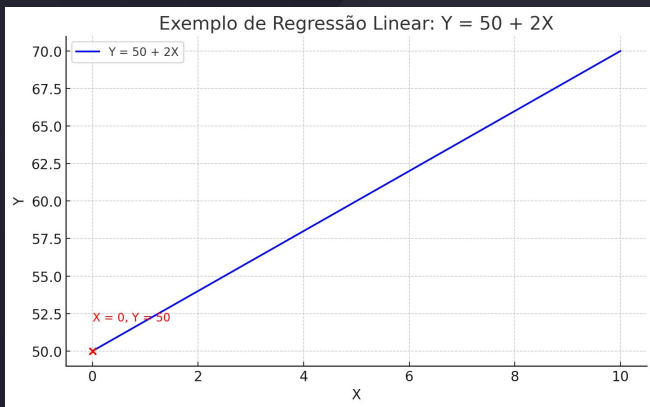
$$Y = a + bX$$

- **a**: intercepto (onde a reta cruza o eixo Y)
- **b**: coeficiente angular (quanto Y varia quando X aumenta 1 unidade)

Interpretando a equação

Ex:

$$Y = 50 + 2X$$



Cada aumento de 1 unidade em X aumenta Y em 2 unidades

Como a linha é calculada?

Método dos mínimos quadrados

- Para cada ponto, ele olha o quanto ele está distante da linha (isso é o erro).
- Em vez de só somar esses erros (o que pode dar zero, se uns forem positivos e outros negativos), ele eleva ao quadrado (por isso chama “mínimos quadrados”).
- Depois, ele escolhe a linha que tiver a menor soma desses quadrados de erro.



Como a linha é calculada?

Método dos mínimos quadrados

- O método dos mínimos quadrados encontra a linha que erra menos, no geral, tentando passar o mais perto possível de todos os pontos.



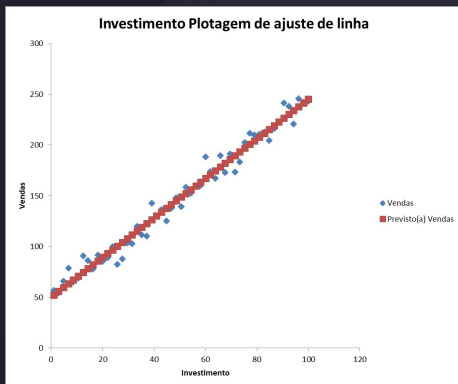
Aplicando na prática (excel)



Prever vendas com base em investimentos em mídia

Aplicando na prática (excel)

Ele explica 97,6% da variação na variável dependente (ex: Vendas) com base em uma variável independente (ex: Investimento).



A correlação é muito forte. O erro padrão é baixo o suficiente considerando a escala dos dados.

Aplicando na prática (excel)

Métrica	O que significa	O que esse valor mostra
R múltiplo	É o valor absoluto da correlação entre os valores reais e os previstos . Vai de 0 a 1.	0,9879 → muito alta correlação linear
R-quadrado (R^2)	Representa a proporção da variação de Y explicada por X .	97,6% da variação nas vendas é explicada pelo investimento
R-quadrado ajustado	R^2 corrigido para o número de variáveis no modelo. Muito útil se você tiver múltiplos X.	Praticamente igual ao R^2 → sinal de modelo consistente
Erro padrão (standard error)	Mede o desvio médio entre os valores reais e os previstos. Quanto menor, melhor.	9,07 → os valores previstos estão em média a ± 9 unidades dos reais
Observações	Quantidade de linhas da amostra.	53 registros usados no modelo

Estatística de regressão	
R múltiplo	0,987905
R-Quadrado	0,975956
R-quadrado ajustado	0,975484
Erro padrão	9,071746
Observações	53

Aplicando na prática (excel)

ANOVA					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	1	170361,6	170361,6	2070,093	5,84621E-43
Resíduo	51	4197,125	82,29657		
Total	52	174558,7			

Essa é a tabela de ANOVA (Análise de Variância). Ela serve para testar se o modelo é estatisticamente significativo, ou seja:

“Essa regressão realmente explica alguma coisa, ou o resultado poderia ter sido só sorte?”

H_0 : “Os coeficientes do modelo são iguais a zero.”

Ou seja:

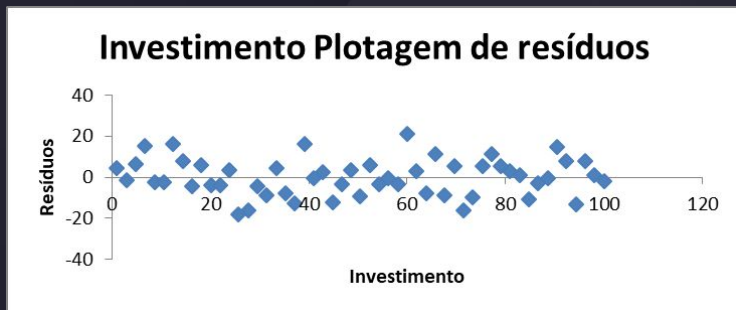
H_0 : “Não há relação linear entre a variável independente (X) e a dependente (Y).”

Aplicando na prática (excel)

ANOVA					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	1	170361,6	170361,6	2070,093	5,84621E-43
Resíduo	51	4197,125	82,29657		
Total	52	174558,7			

Coluna	Significado
gl (graus de liberdade)	N-1 para total, 1 para regressão (uma variável X), N-2 para resíduo
SQ (Soma dos Quadrados)	Medida da variância (quanto os dados se espalham)
MQ (Média dos Quadrados)	SQ dividida pelo gl (serve para comparar variância explicada e erro)
F (Estatística F)	Razão entre variância explicada e não explicada. Quanto maior, melhor o modelo.
F de significância	O valor-p da regressão. Mede a chance de a relação ser aleatória.

O que são os erros (resíduos)?



Diferença entre o valor real e o valor previsto

Interpretação dos coeficientes

	<i>Coeficientes</i>
Interseção	50,42325206
Investimento	1,946740701

Coeficiente positivo = relação direta

Coeficiente negativo = relação inversa

Correlação e regressão

CORRELAÇÃO

Mede relação

X

REGRESSÃO

Prediz valores

Coeficiente de determinação R^2

Mede o quanto da variação de Y é explicada por X

"O quão boa é essa reta?" -> Coeficiente de determinação

- Vai de 0 a 1
- $R^2 = 0,85 \rightarrow 85\%$ da variação em Y é explicada por X
- R^2 baixo pode indicar que faltam variáveis

$$R^2 = 1 - \frac{SQ_{resíduo}}{SQ_{total}}$$

Regressão simples vs múltipla

Situação	Regressão simples	Regressão múltipla
Você olha uma causa	Temperatura → Sorvete	Temperatura + Promoção + Feriado → Sorvete
Você quer uma explicação direta	Y vs X	Y vs $(X_1 + X_2 + X_3 + \dots)$

Quando não usar regressão?

- ❑ Se a relação não for linear
- ❑ Se houver outliers extremos
- ❑ Se a variável dependente for categórica (use logística)

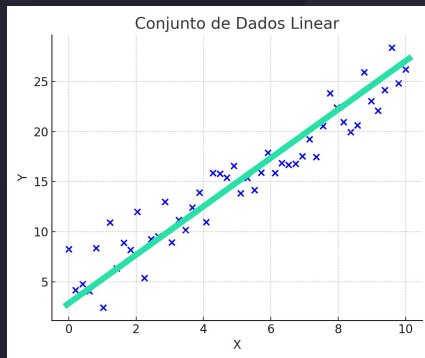
Premissas da regressão linear

Premissa	Significa que...	Se violada, causa...
Linearidade	Relação $X \rightarrow Y$ é reta	Modelo não representa a realidade
Independência dos erros	Erros não influenciam uns aos outros	Previsões instáveis, especialmente no tempo
Homocedasticidade	Erros têm variância constante	Inferência distorcida
Normalidade dos resíduos	Erros seguem distribuição normal	Testes estatísticos pouco confiáveis
Sem multicolinearidade	Xs são independentes entre si	Coefficientes instáveis, interpretação difícil

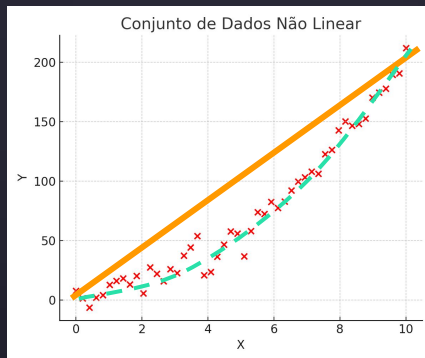
Premissas da regressão linear

LINEARIDADE

✓ PODE



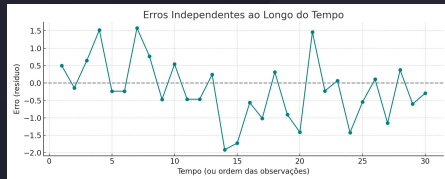
✗ NÃO PODE



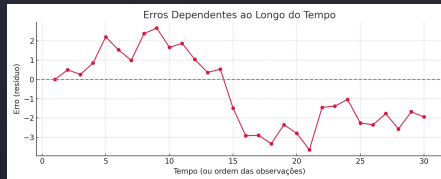
Premissas da regressão linear

INDEPENDÊNCIA DOS ERROS

✓ PODE



✗ NÃO PODE

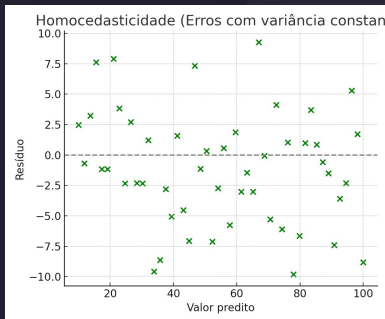


TESTE DE DURBIN-WATSON

Premissas da regressão linear

HOMOCEDASTICIDADE

✓ PODE



✗ NÃO PODE

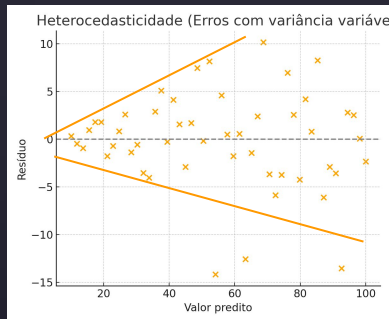
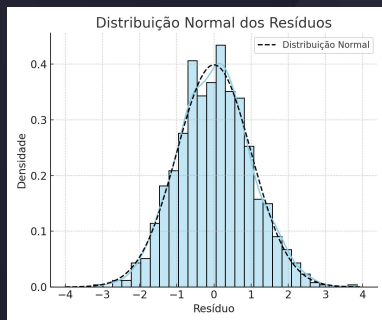


GRÁFICO DE RESÍDUOS

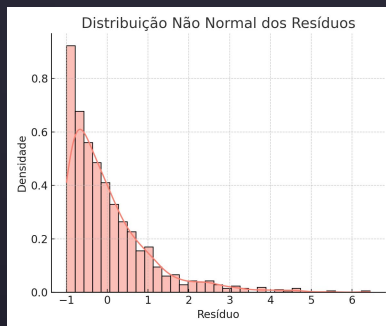
Premissas da regressão linear

NORMALIDADE DOS RESÍDUOS

✓ PODE



✗ NÃO PODE

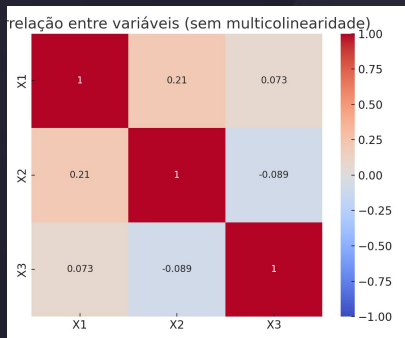


Teste de Shapiro-Wilk

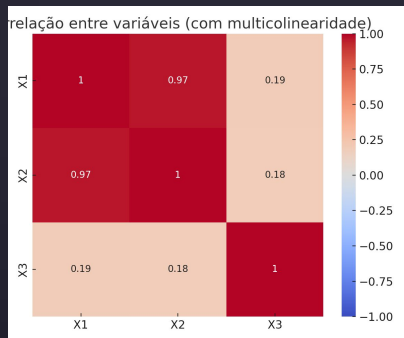
Premissas da regressão linear

SEM MULTICOLINEARIDADE

✓ PODE



✗ NÃO PODE



VIF (Variance Inflation Factor)

Conclusão

- Regressão linear não é bola de cristal
- Simples, mas exige cuidado
- Abre portas para modelagem avançada