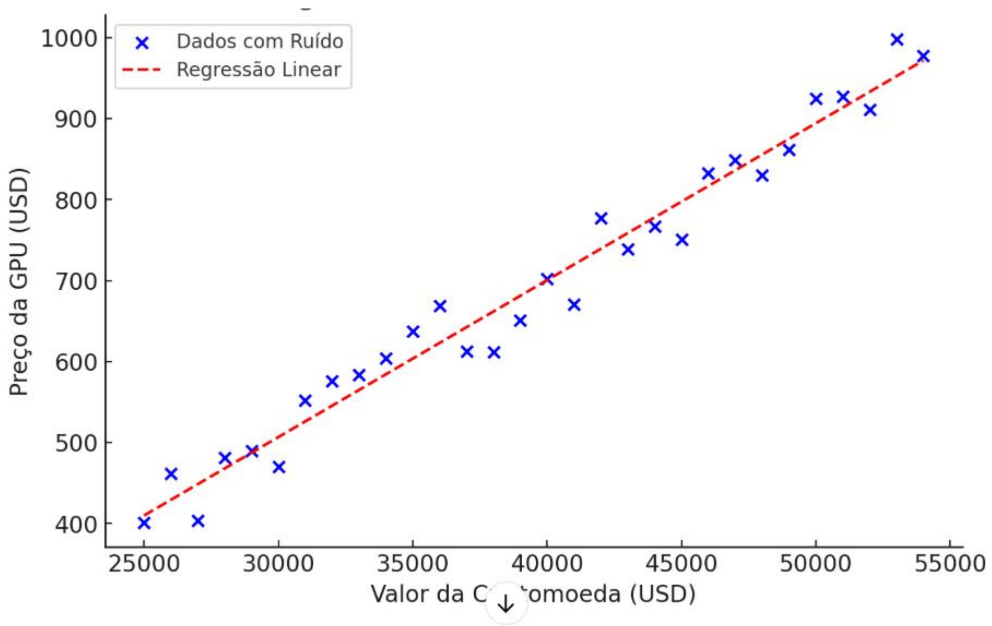
Motivação

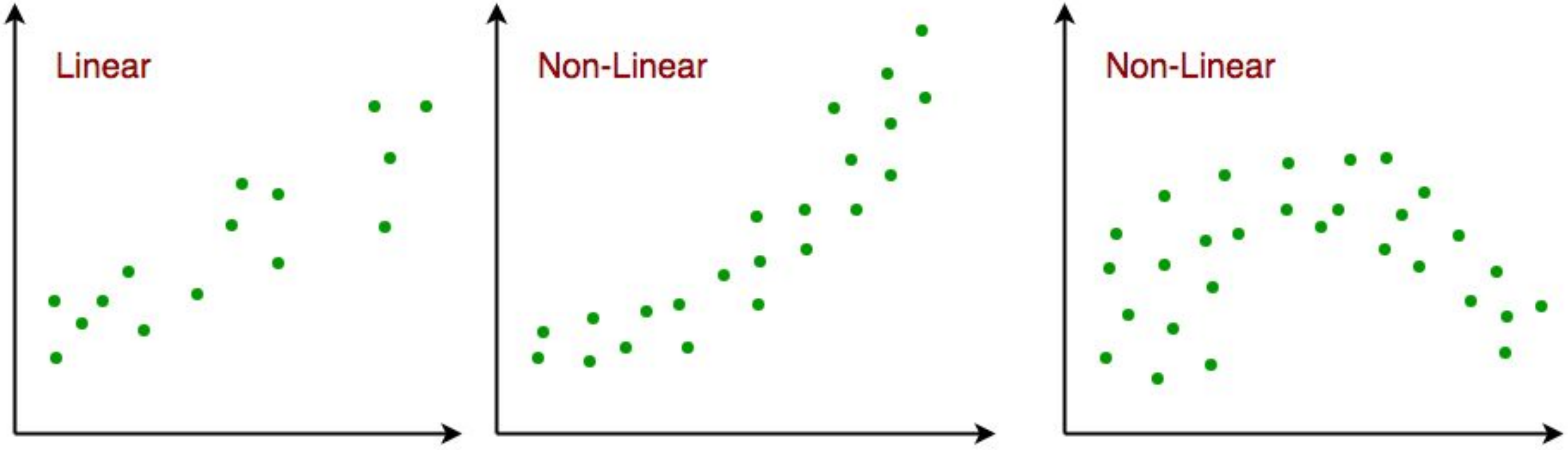
- Você precisa prever a popularidade de suas postagens em uma rede social, você quer utilizar como parâmetro da postagem a duração do vídeo.

- Você precisa prever o preço de uma casa com base na distância do centro da cidade.

- Prever a nota na prova com base na quantidade de horas de estudo.

Valor da Criptomoeda (USD)	Preço da GPU (USD)
25000	401
26000	462
27000	404
28000	481
29000	490
30000	470
31000	552
32000	576
33000	584
34000	604
35000	637
36000	669
37000	613
38000	612
39000	651
40000	702
41000	671
42000	777
43000	739




Linear
Non-Linear
Non-Linear

Regressão Linear - Simples

A regressão linear modela a relação entre uma variável dependente Y e uma ou mais variáveis independentes X através de uma equação linear.

Coeficiente

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Variável Dependente

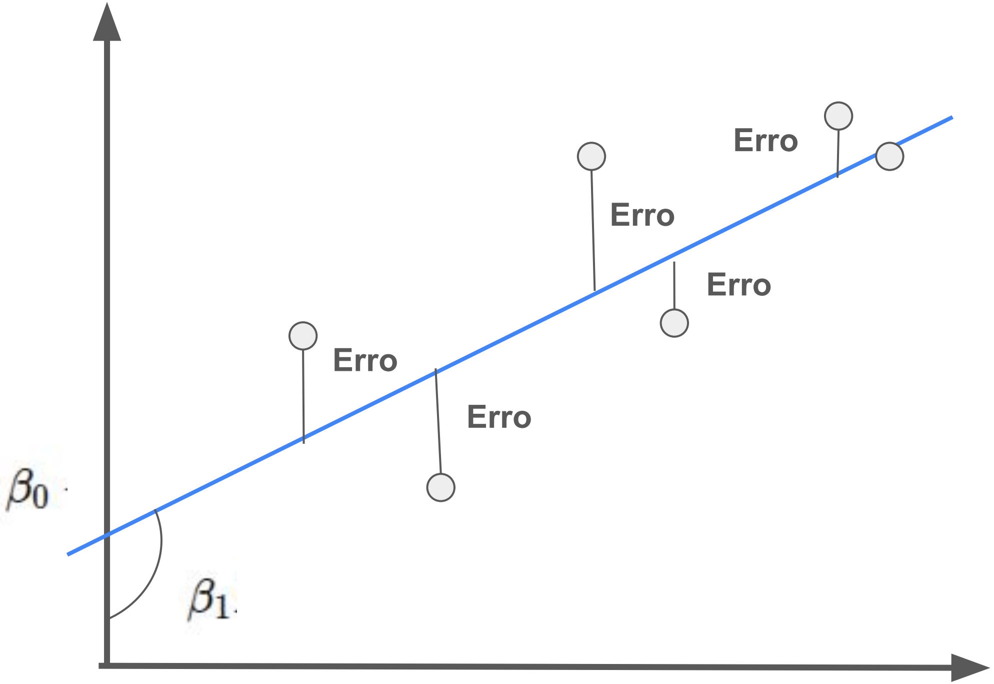
$$y = \beta_0 + \beta_1 X + \epsilon$$

Erro residual

Variável Independente

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

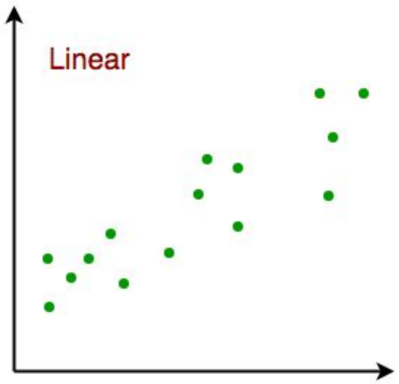
Constante/Intercepto



β_0
β_1
Erro
Erro
Erro
Erro
Erro
Erro

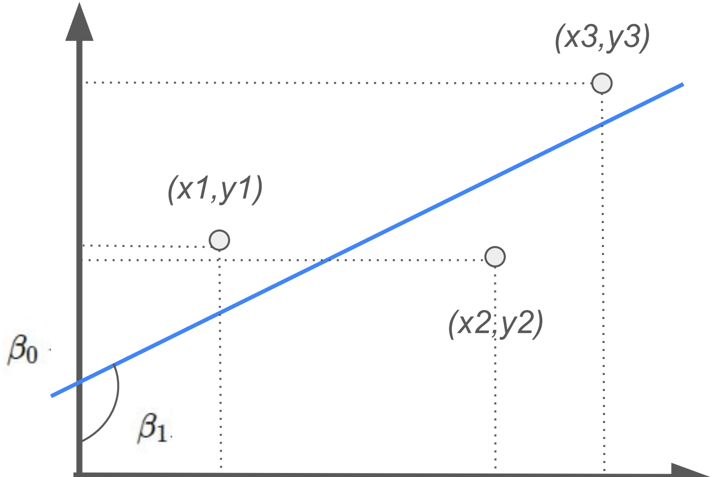
Condições para utilização

Linearidade: As variáveis independentes e dependentes têm uma relação linear entre si. Isso implica que as mudanças na variável dependente seguem aquelas na variável independente de forma linear. Isso significa que deve haver uma linha reta que pode ser desenhada através dos pontos de dados. Se a relação não for linear, a regressão linear não será um modelo preciso.



Condições para utilização

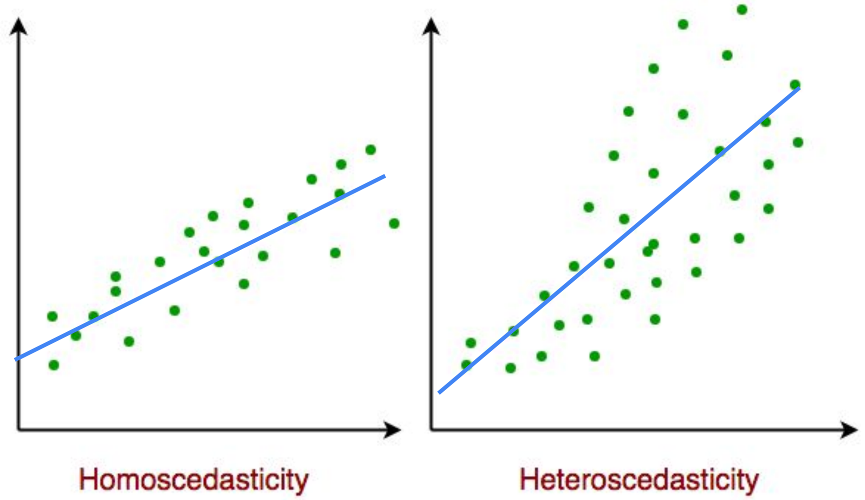
Independência: As observações no conjunto de dados são independentes umas das outras. Isso significa que o valor da variável dependente para uma observação não depende do valor da variável dependente para outra observação. Se as observações não forem independentes, a regressão linear não será um modelo preciso (Teste de Durbin-Watson).



Os valores de *x1,x2 e x2* são independentes entre si.

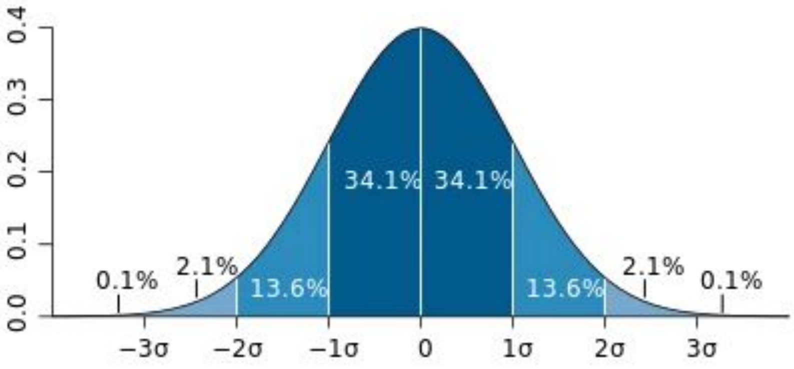
Condições para utilização

Homocedasticidade: Em todos os níveis da(s) variável(is) independente(s), a variância dos erros é constante. Isso indica que a quantidade da(s) variável(is) independente(s) não tem impacto na variância dos erros. Se a variância dos resíduos não for constante, a regressão linear não será um modelo preciso (teste de Breusch-Pagan)



Condições para utilização

Normalidade: Os resíduos devem ser distribuídos normalmente. Isso significa que os resíduos devem seguir uma curva em forma de sino. Se os resíduos não forem distribuídos normalmente, a regressão linear não será um modelo preciso (Teste de Shapiro-Wilk).



Treinamento

1 - Calcular as médias de x e y.

2 - Calcular o coeficiente angular.

3 - Calcular o intercepto.

4 - Calcular o erro residual.

 4.1 - Calcular as previsões.

 4.2 - Erro Quadrático Médio (MSE)

 4.3 - Erro Médio Absoluto (MAE)

 4.4 - Calcular o R²

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Treinamento - Dados

X	Y
1	2
2	2.5
3	3
4	4
5	4.5

Erro / Resíduo

O erro médio absoluto (MAE - *Mean Absolute Error*) mede a média da diferença entre o valor real com o predito. Como a diferença pode ser positiva ou negativa, é adicionado um módulo entre a diferença dos valores.

$$MAE\left(y, \hat{y}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

A métrica é pouco afetada por valores discrepantes (*outliers*).

Erro / Resíduo

O erro quadrático médio (MSE - *Mean Squared Error*) é uma métrica que calcula a média de diferença entre o valor predito com o real. Entretanto, ao invés de usar o módulo do resultado entre o valor de y e ŷ, nesta métrica a diferença é elevada ao quadrado.

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

Esta métrica penaliza os valores de erro muito altos.

R^2

A métrica R^2, R-dois ou coeficiente de determinação; representa o percentual da variância dos dados que é explicado pelo modelo.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Os resultados variam de 0 a 1, geralmente também são expressos em termos percentuais, ou seja, variando entre 0% e 100%. Quanto maior é o valor de R^2, mais explicativo é o modelo em relação aos dados previstos.

Regressão Linear - Múltipla

Intercepto ou
constante do modelo

Incremento
na variável 02

variável 02
(Ex: Horas)

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Predição da
Variável Dependente

Incremento
na variável 01

variável 01
(Ex: Idade)

Exercício

Calcule a equação resultante da regressão linear da tabela abaixo, calcule o MSE, MAE e R^2. Ao final, verifique se o modelo explica os dados utilizando o valor do coeficiente de determinação.

X	Y
2	4.5
4	7.0
6	9.5
8	12.0
10	14.5