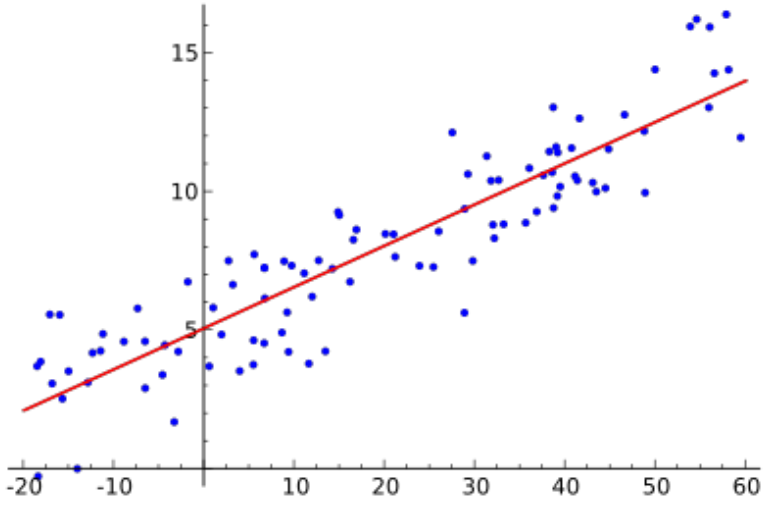
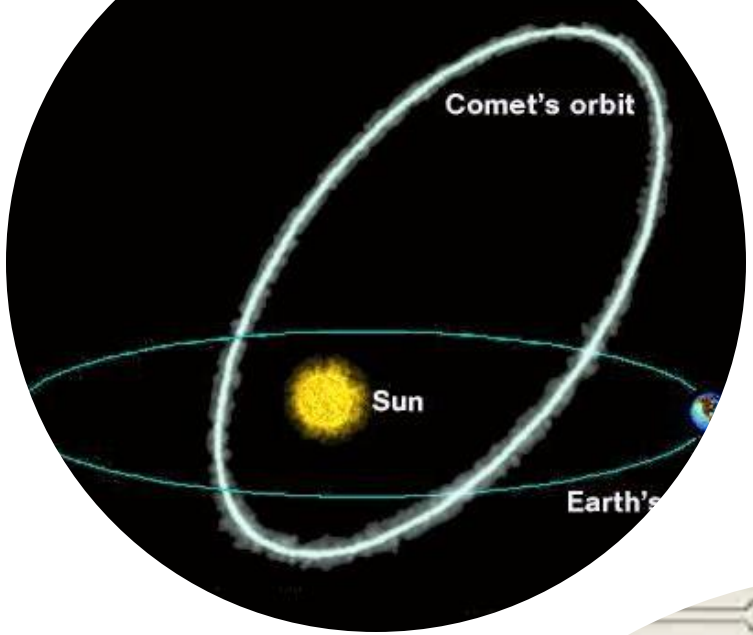
Regressão Linear



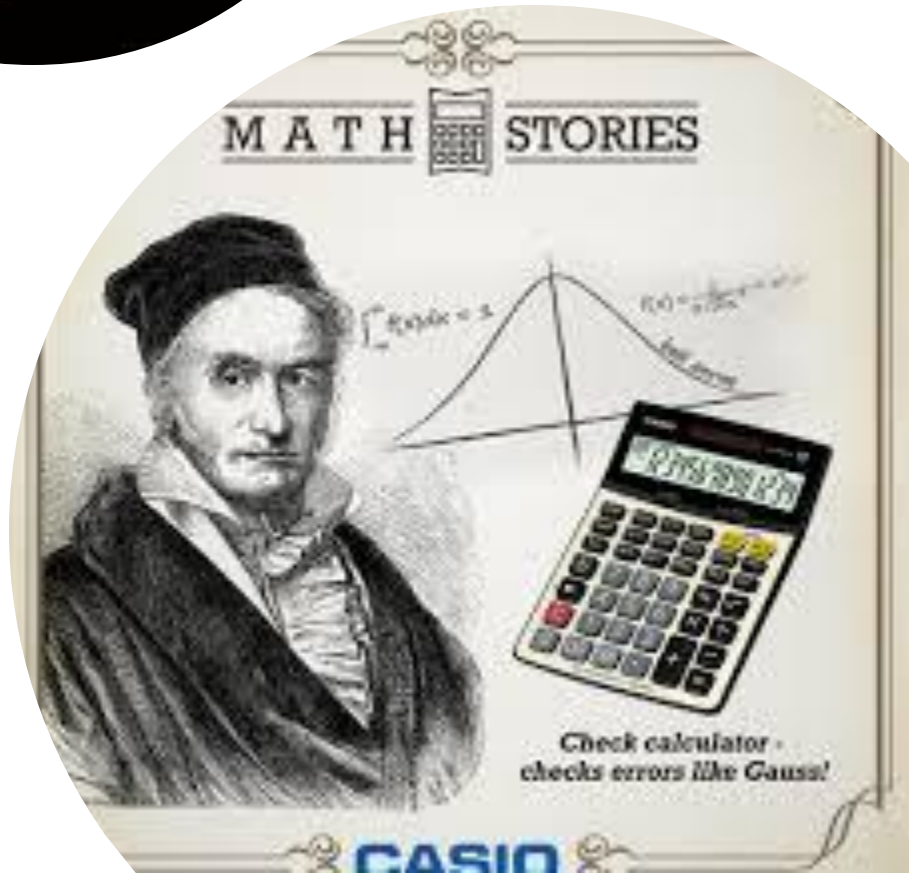
PROF. DR. VLADIMIR C ALENCAR
LANA - UEPB
WWW.VALENCAR.COM

A Inteligência Artificial moderna começou em 1800, com a invenção da Regressão Linear pelo Johann Carl Friedrich Gauss (Alemanha, 1777 - 1855).



Ele foi um matemático, astrônomo e físico alemão que contribuiu muito em diversas áreas da ciência, dentre elas a teoria dos números, estatística, análise matemática, geometria, geofísica, astronomia.

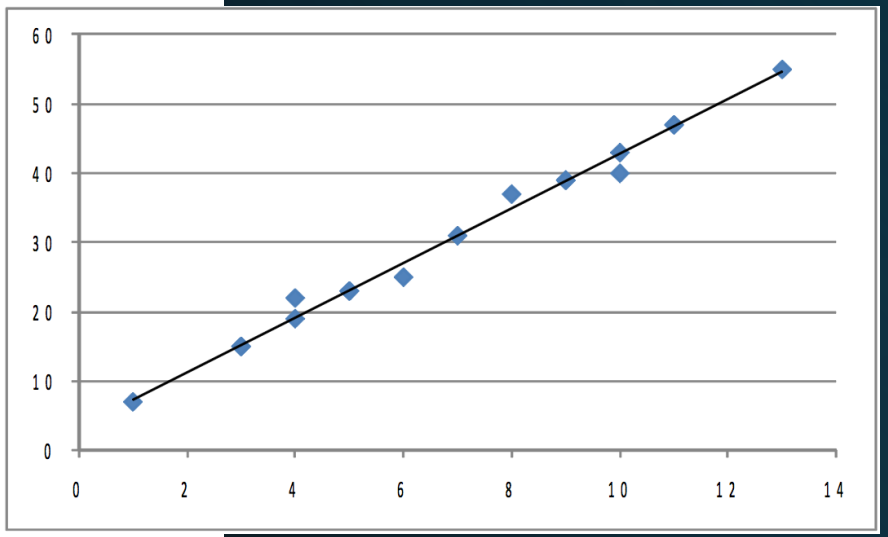
Ele usou a Regressão Linear para determinar as órbitas dos cometas ao redor do Sol.



- Alguns anos mais tarde, o estatístico inglês Sir Galton cunhou o termo “regressão” para prever o crescimento da ervilha e outros fenómenos biológicos.

- O nome veio do fato de que as coisas tendem a “regredir à média”.

- Seu trabalho foi utilizado pelo matemático inglês Karl Pearson para desenvolver os conceitos de correlação e distribuição em 1.900

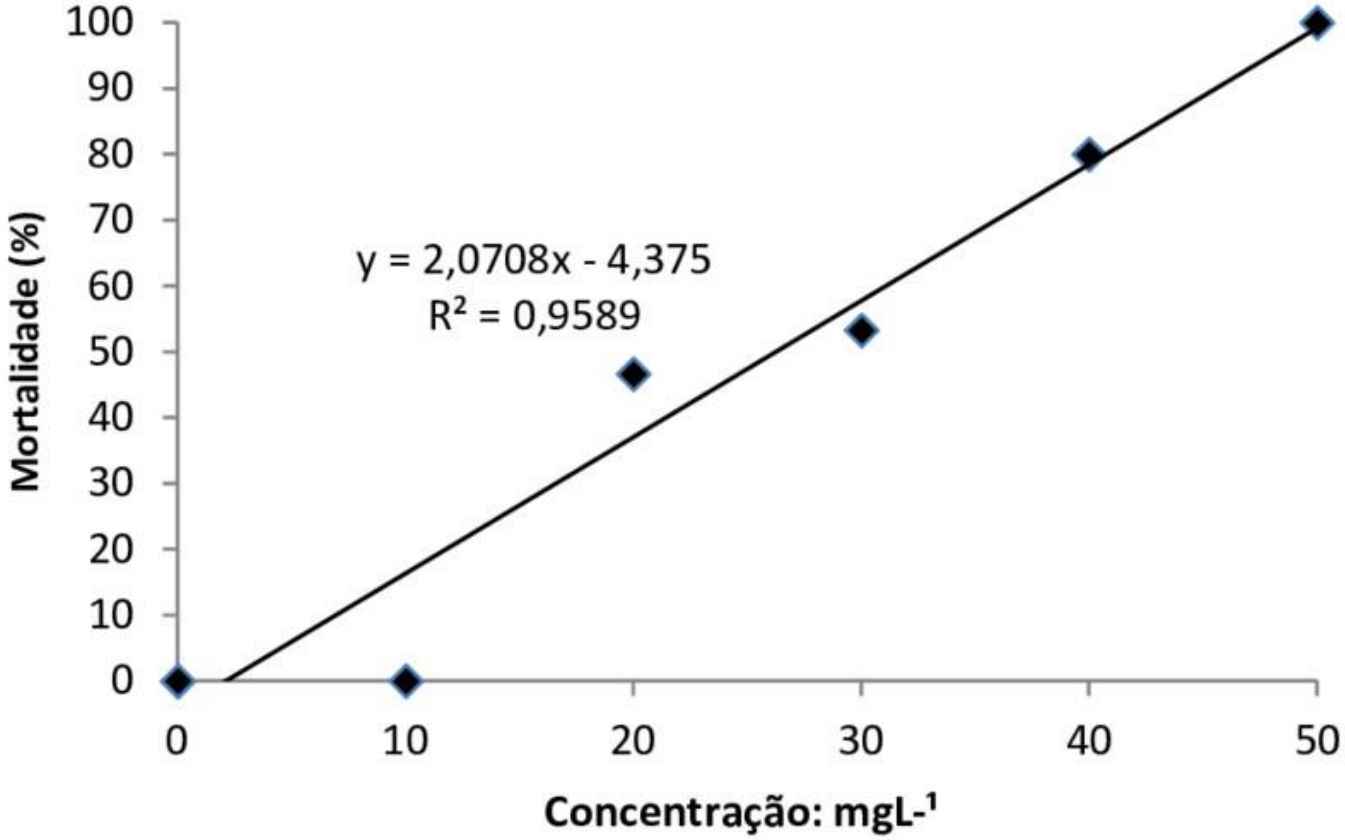


Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas de tal forma que uma variável possa ser predita a partir de outra.

Um estudo de regressão busca, essencialmente:

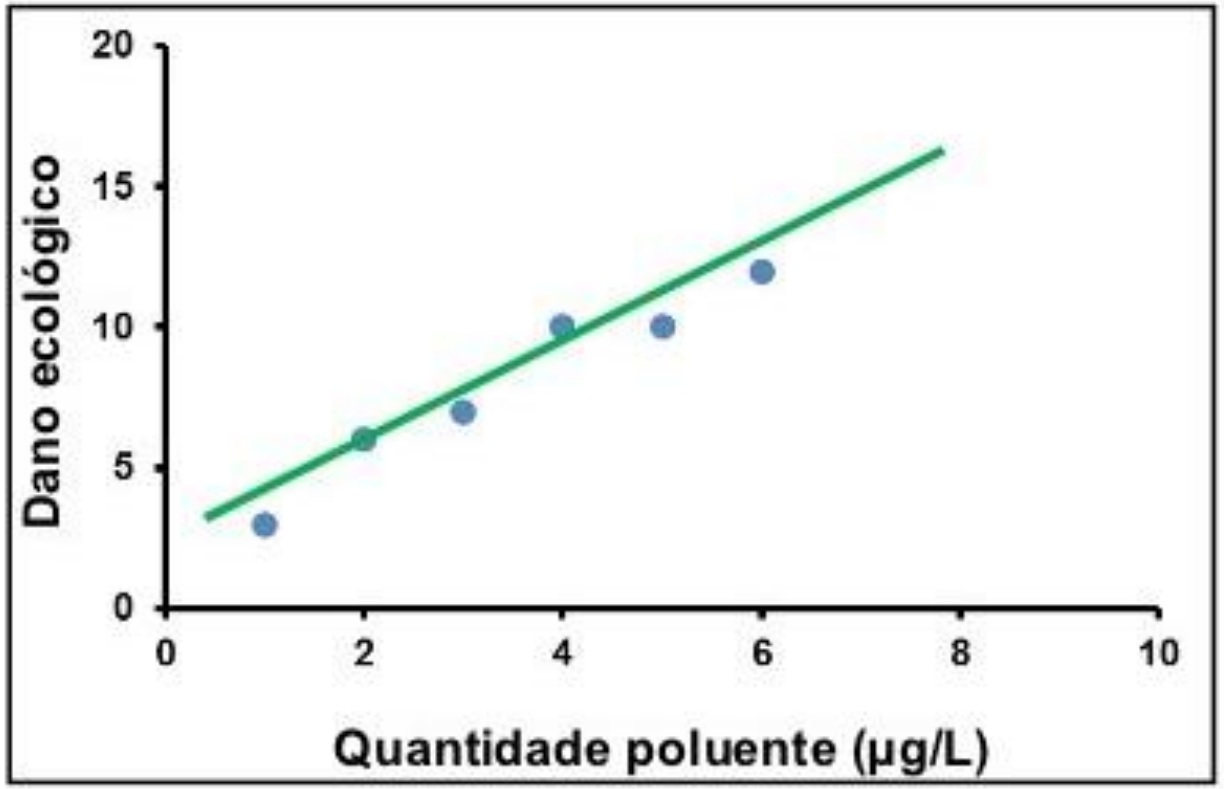
associar uma variável Y -> variável resposta ou dependente

a uma outra variável X -> variável explanatória ou independente



- Em uma determinada região um biólogo pretende estudar a relação entre um determinado poluente (P) despejado por uma fábrica em um riacho e o dano causado em curso d'água em um valor de dano qualquer.

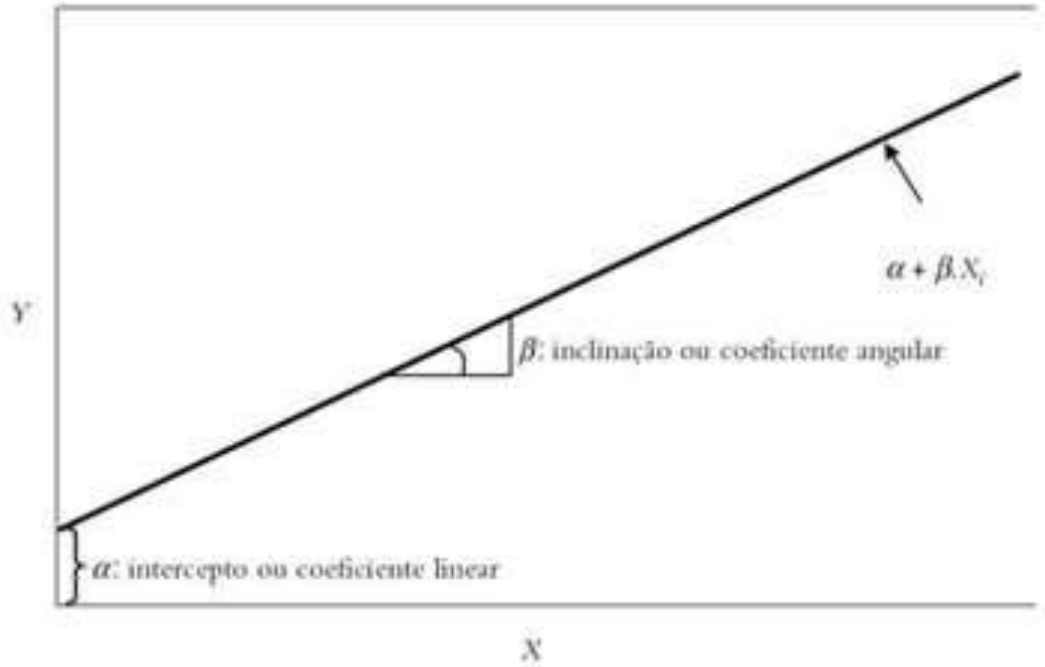
Quantidade Poluente (µg/L)	Dano Ecológico
1	3
2	6
3	7
4	10
5	10
6	12

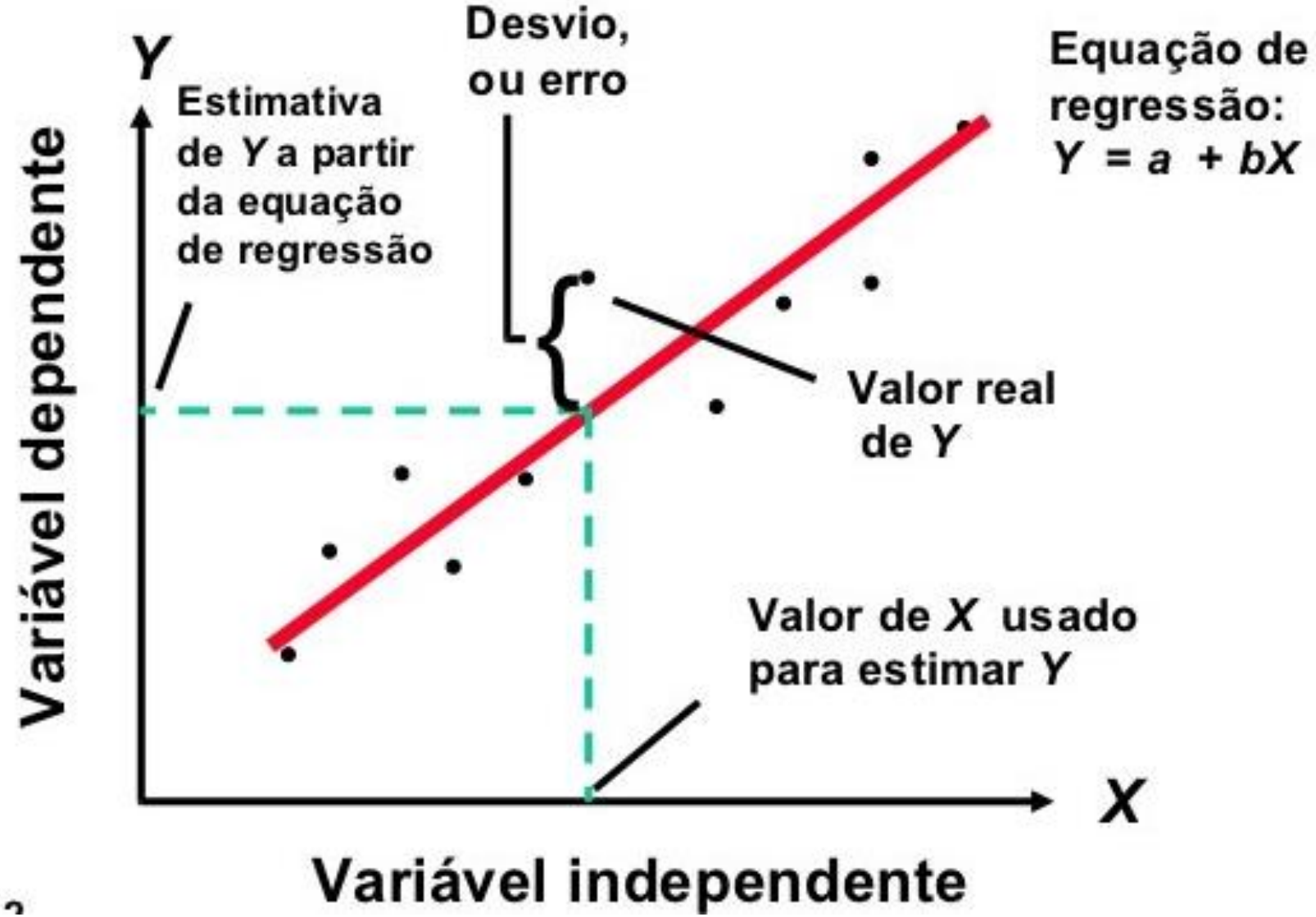


$$\hat{y} = a + bx$$

Onde:

- $\hat{y}$ = valor previsto de *y* dado um valor para *x*
- ***x*** = variável independente
- ***a*** = ponto onde a linha intercepta o eixo y
- ***b*** = inclinação da linha reta



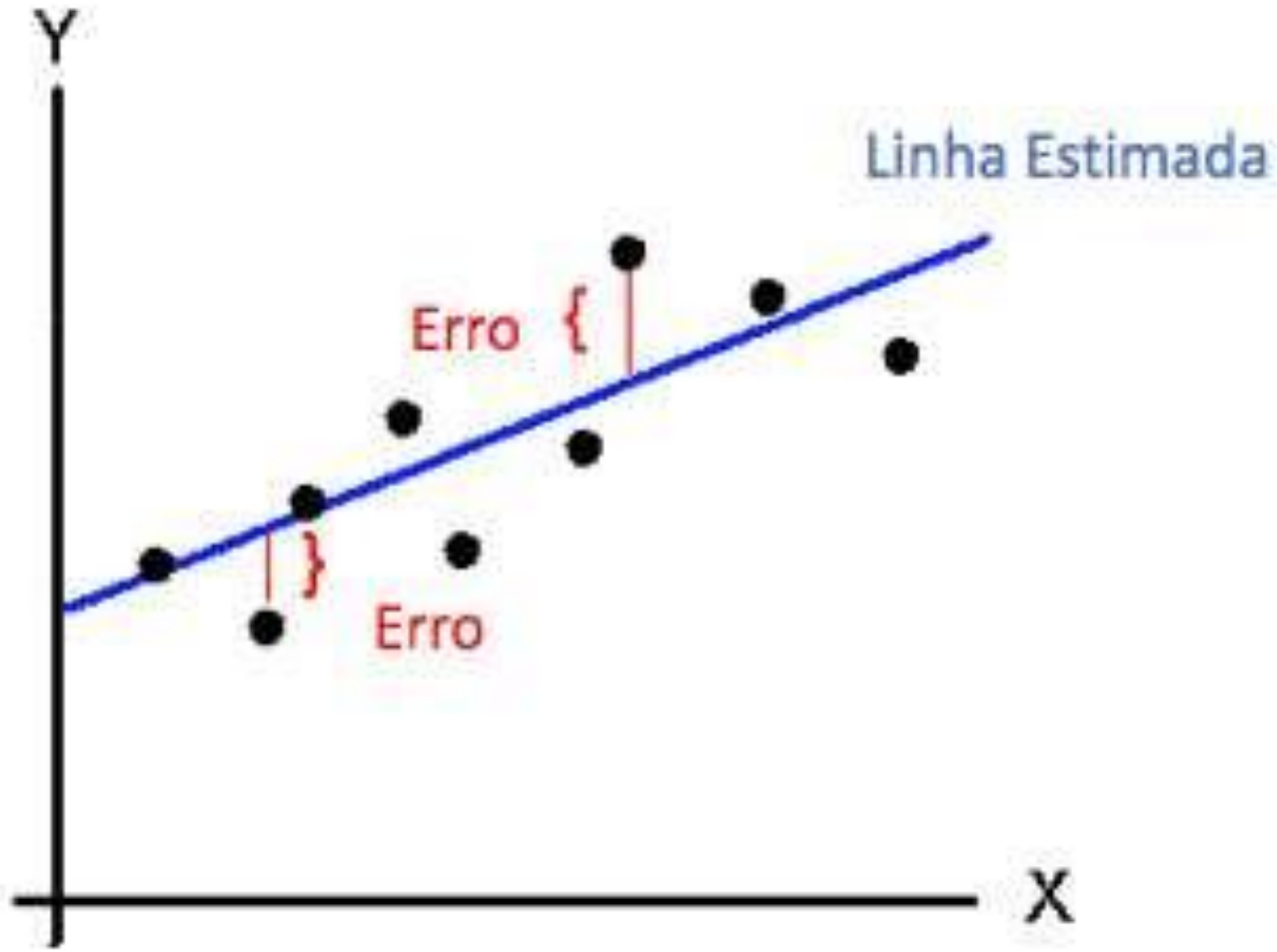

Y
Desvio,
ou erro
Equação de
regressão:
Y = a + bX
Estimativa
de Y a partir
da equação
de regressão
Valor real
de Y
Variável dependente
Valor de X usado
para estimar Y
X
Variável independente

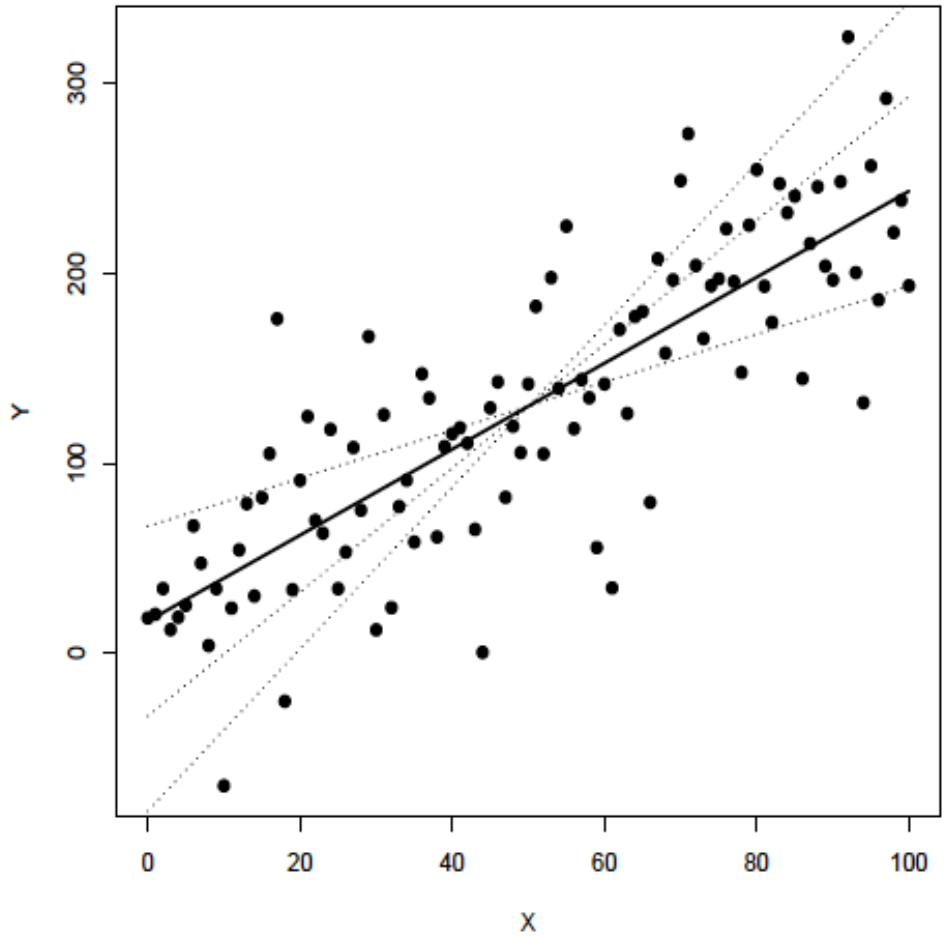
$Y = \alpha + \beta . X$

Deve-se determinar **α** e **β** de modo que a somatória dos quadrados dos **resíduos** seja a menor possível

Método de Mínimos Quadrados Ordinários – MQO

(Ordinary Least Squares - OLS)

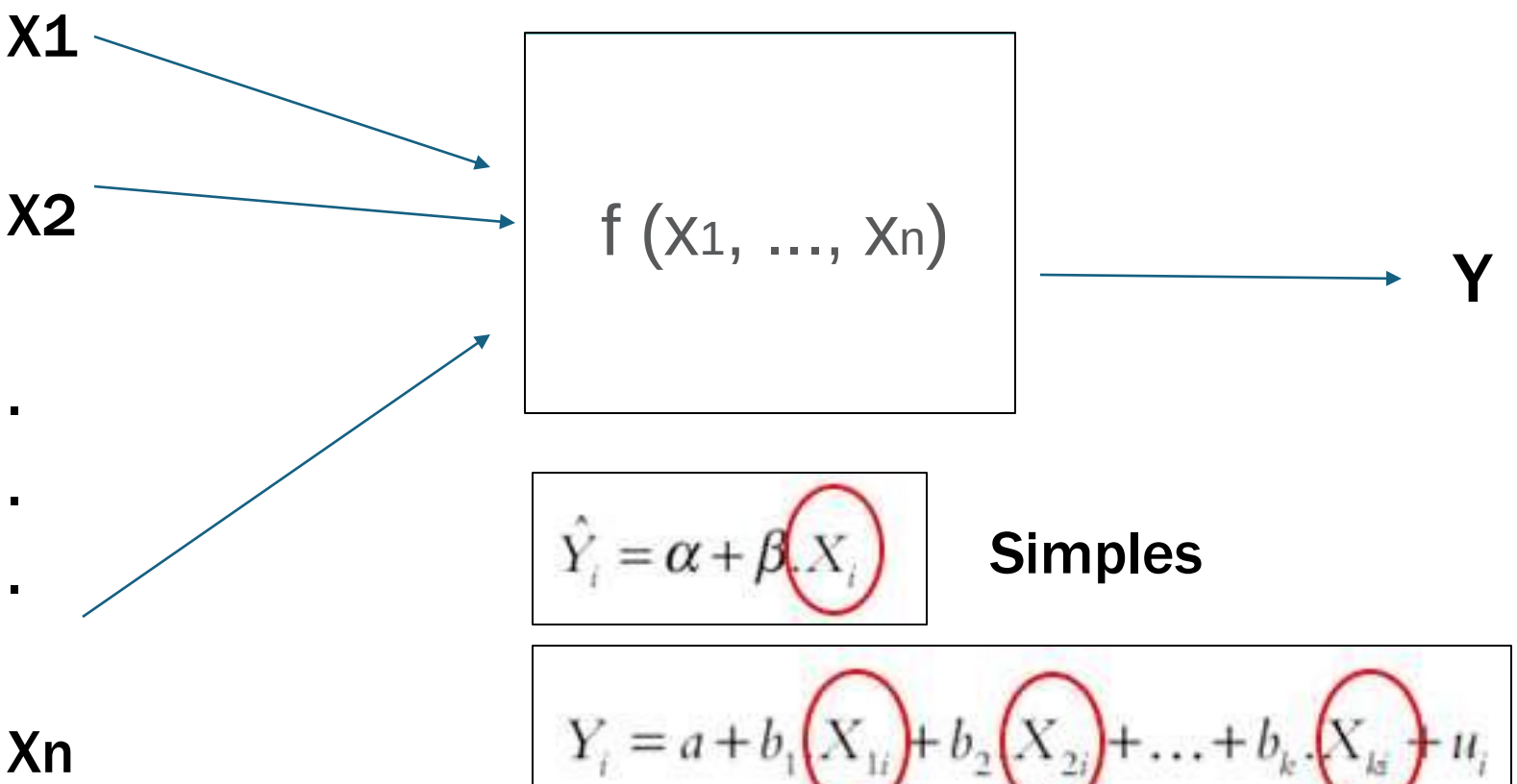




$$Y = \alpha + \beta \, . \, X$$

Os coeficientes dessa reta podem ser estimados pelo Método dos Mínimos Quadrados

Regressão Linear



$$\hat{Y}_i = \alpha + \beta . X_i$$

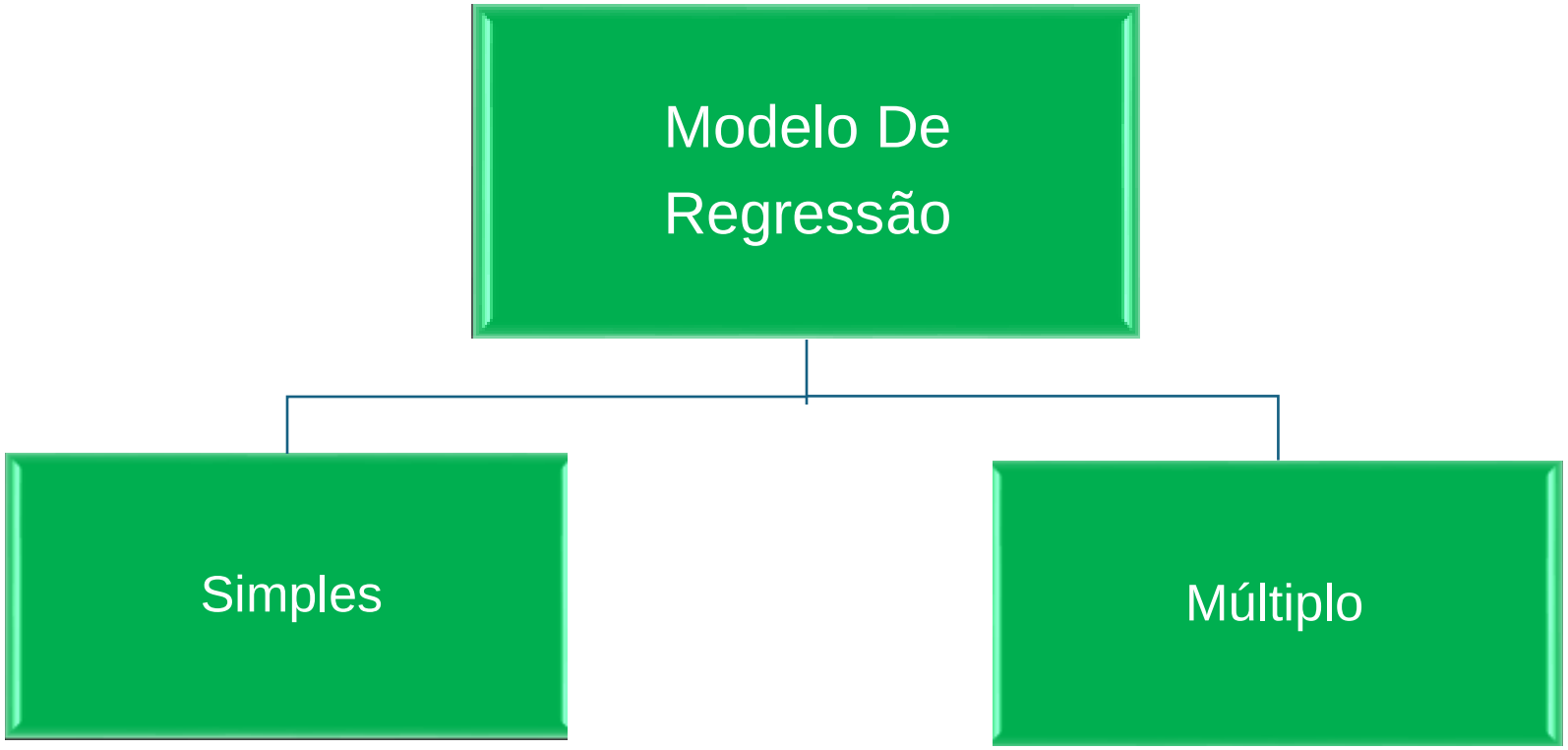
Simples

$$Y_i = a + b_1 . X_{1i} + b_2 . X_{2i} + \ldots + b_k . X_{ki} + u_i$$

Múltipla

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

Este relacionamento existe em apenas uma direção:
variável independente (x) → variável dependente (y)



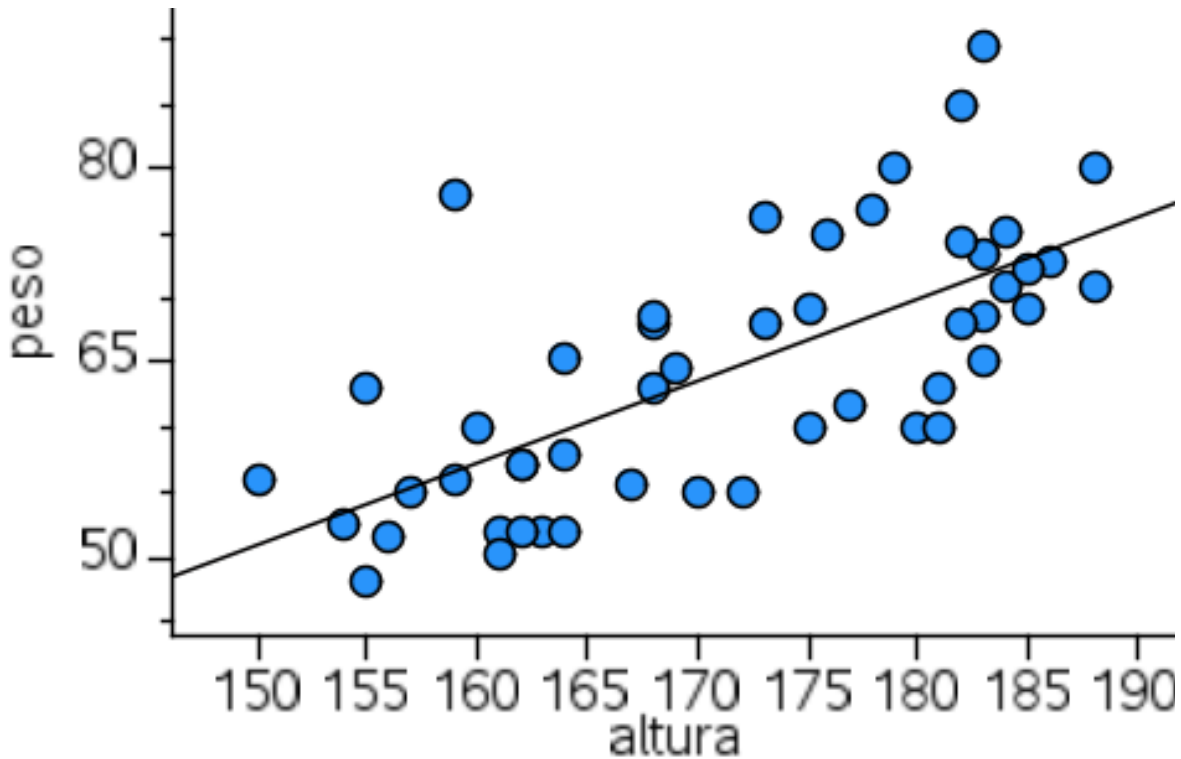
1 Variável Dependente Y

1 Variável Independente X

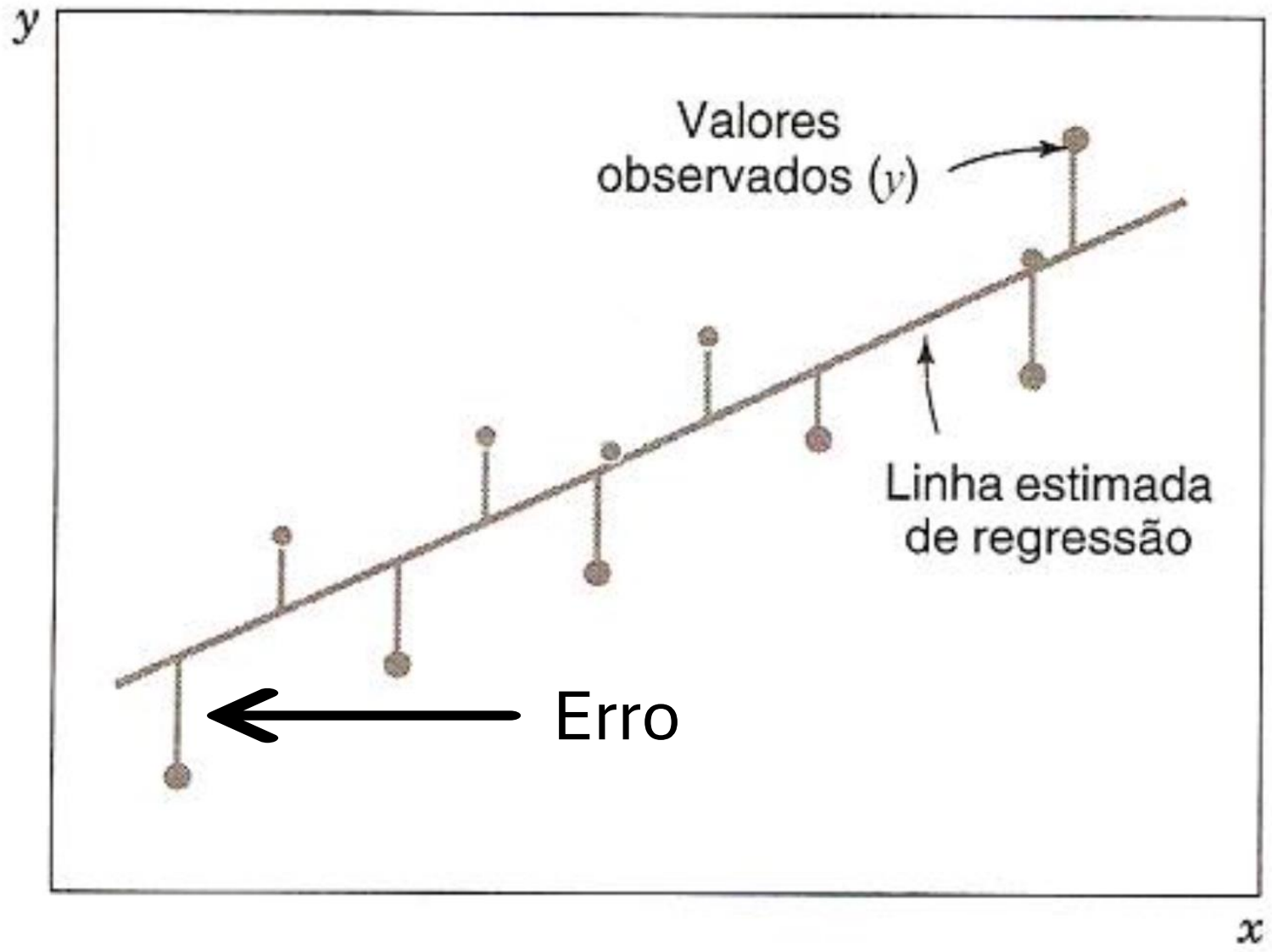
1 Variável Dependente Y

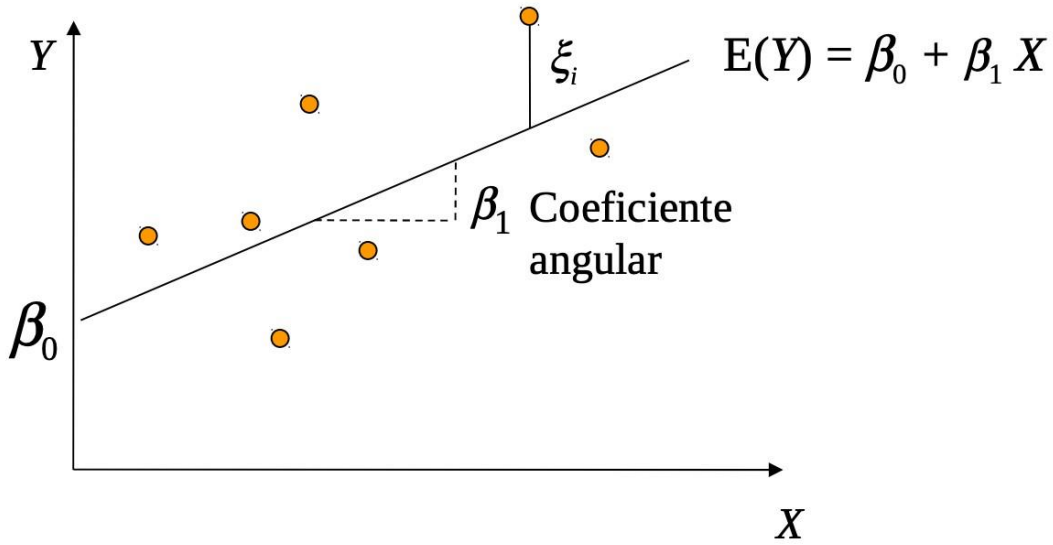
2 ou + Variáveis Independentes X, Xi

$$\hat{y} = a + bx$$



Reta de Regressão


y
Valores
observados (y)
Linha estimada
de regressão
Erro
x



$$E(Y) = \beta_0 + \beta_1 X$$

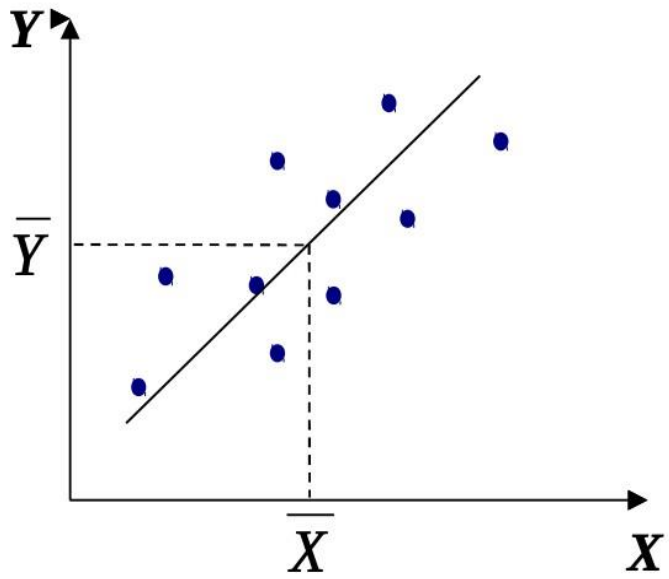
$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Erro Aleatório

Variável Dependente

Variável Independente

A reta de regressão passa sempre pelo ponto $(\overline{X}, \overline{Y})$



Em geral não se conhece os valores de β_0, β_1 e σ^2

Eles podem ser estimados através de dados obtidos por amostras.

O método utilizado na estimação dos parâmetros é o método dos mínimos quadrados, o qual considera os desvios dos Y_i de seu valor esperado:

$$\xi_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Em particular, o método dos mínimos quadrados requer que consideremos a soma dos n desvios quadrados, denotado por Q:

$$Q = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

$$Q = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

De acordo com o método dos mínimos quadrados, os estimadores de β_0 e β_1 são aqueles, denotados por b_0 e b_1, que tornam mínimo o valor de Q.

Derivando

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] X_i$$

Igualando-se essas equações a zero obtém-se os valores b_0 e b_1 que minimizam Q:

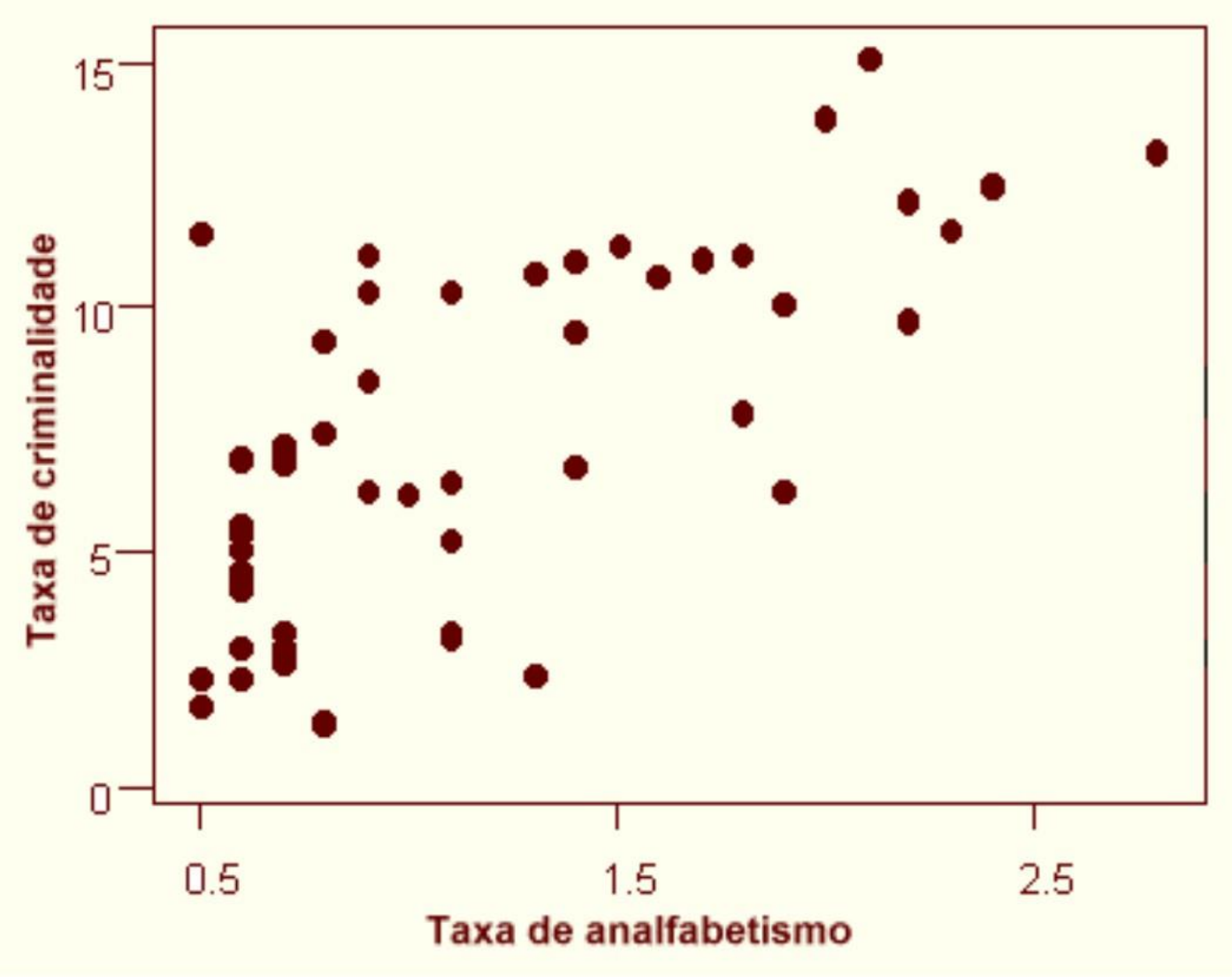
$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$E(Y) = \beta_0 + \beta_1 X$$

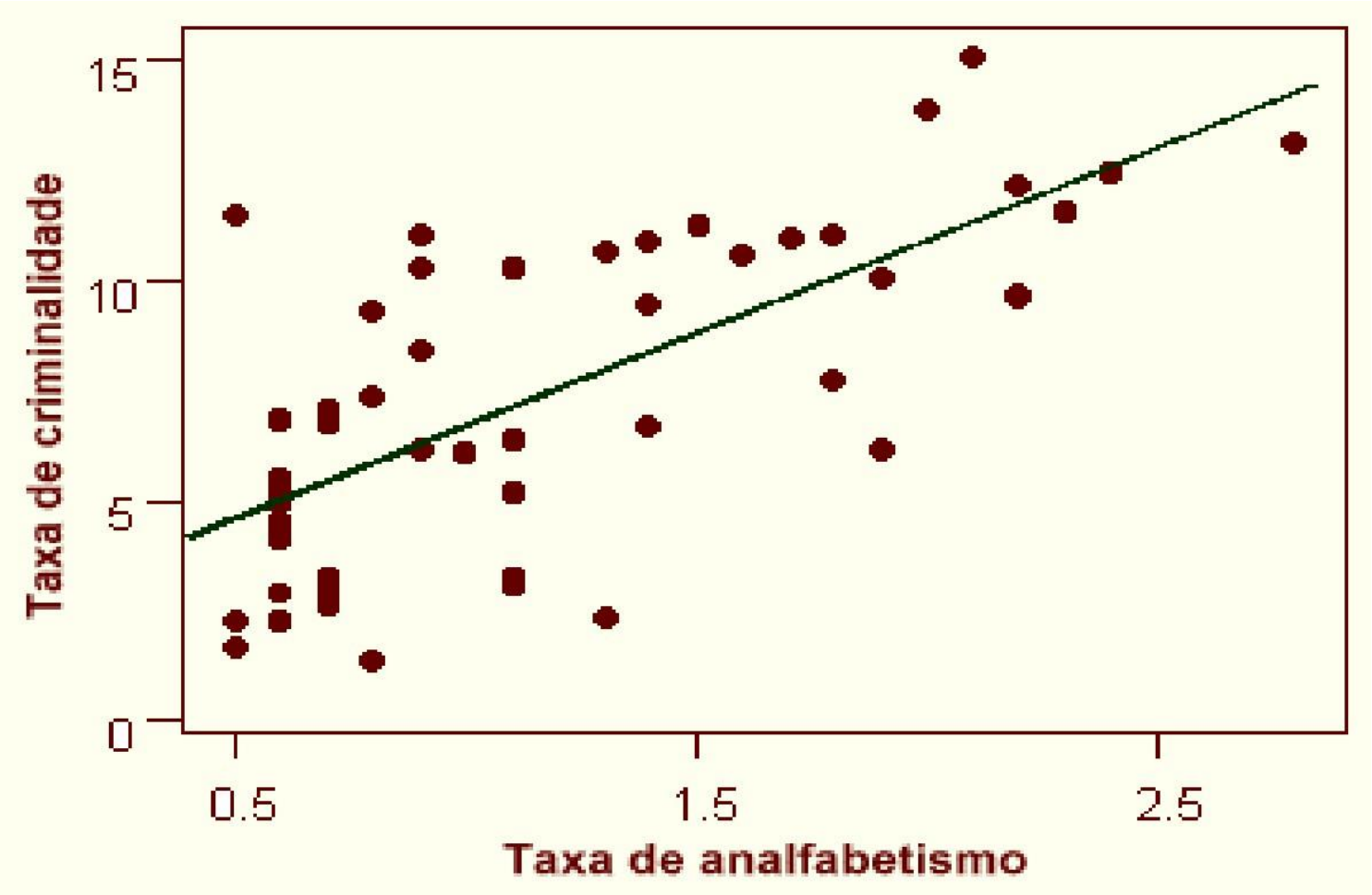
$$\hat{Y} = b_0 + b_1 X$$

$$e_i = Y_i - \hat{Y}_i \quad \text{(resíduo)}$$



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar.

Nota-se também uma tendência linear. Correlação entre X e Y: 0,702



$$Y\_prev = 2{,}397 + 4{,}257\ X$$

$$\hat{Y} = 2{,}397 + 4{,}257\ X$$

Y_prev - valor predito para a taxa de criminalidade

X : taxa de analfabetismo

INTERPRETAÇÃO:

Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,257 unidades.

Os **coeficientes de correlação** são métodos estatísticos para se medir as relações entre variáveis e o que elas representam.

O que a **correlação** procura entender é como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas.

Coeficiente de Correlação (R)

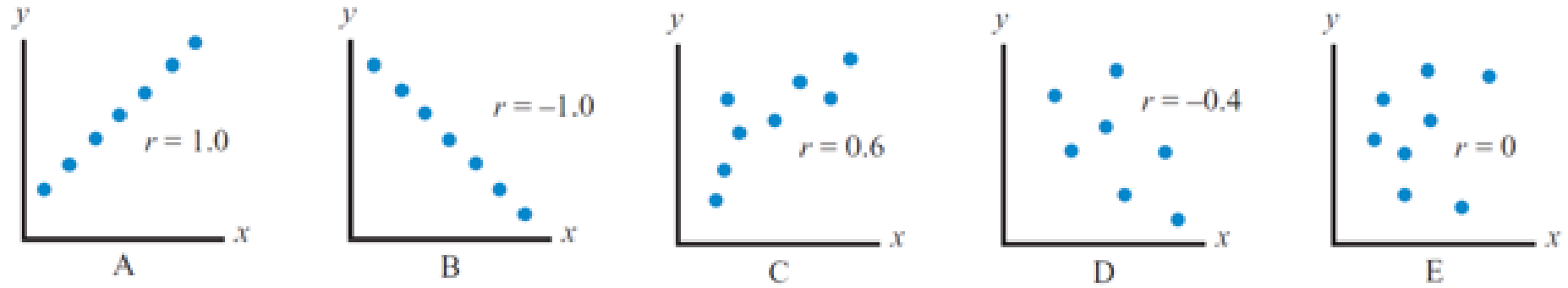


Gráfico A (*r* = 1.0): correlação positiva perfeita entre *x* e *y*

Gráfico B (*r* = -1.0): correlação negativa perfeita entre *x* e *y*

Gráfico C (*r* = 0.6): relação positiva moderada: y tende a aumentar se x aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

Gráfico D (*r* = -0.4): relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo: y tende a diminuir se x aumenta

Gráfico E (*r* = 0): Sem relação entre x e y

Os valores de **r** variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.

Correlação de Pearson (R)

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

Onde:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$$

Coeficiente de determinação (Ajuste) (R^2)

É uma medida de ajuste de um modelo estatístico linear generalizado, como a regressão linear simples ou múltipla, aos valores observados de uma variável aleatória.

O R^2 varia entre 0 e 1, por vezes sendo expresso em termos percentuais.

Nesse caso, expressa a quantidade da variância dos dados que é explicada pelo modelo linear.

Assim, quanto maior o R^2, mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra.

Por exemplo, um $R^2 = 0,8234$ significa que o modelo linear explica 82,34% da variância da variável dependente a partir do regressores (variáveis independentes) incluídas naquele modelo linear.

Coeficiente de determinação (Ajuste) (R^2)

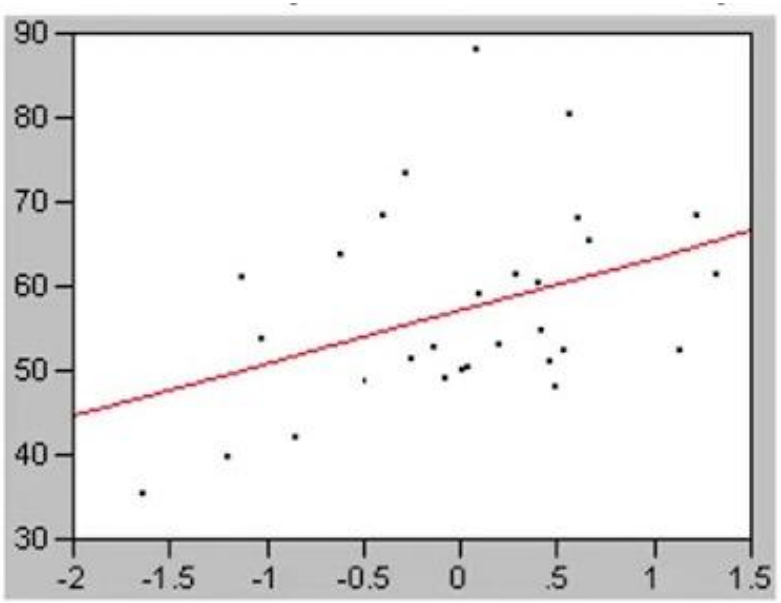
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Onde:

$\hat{y}$ – Valor Previsto (y_prev)

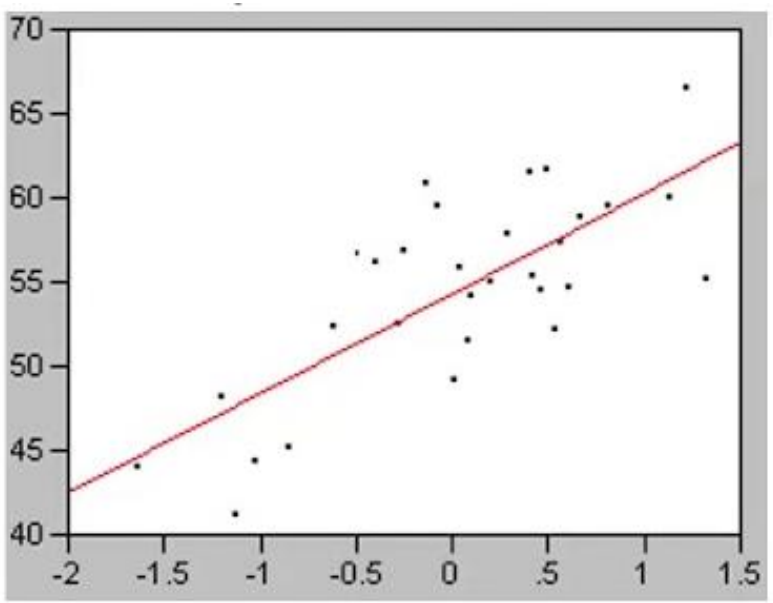
$\bar{y}$ – $Média\ de\ y$

Coeficiente de Determinação (Ajuste) (R^2)



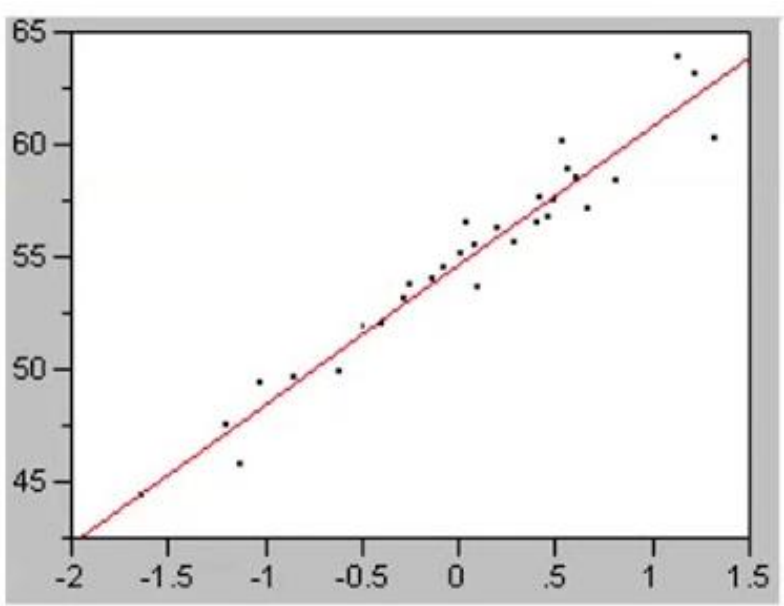
$R^2 = 0.16$

extremely poor
predictive ability



$R^2 = 0.56$

moderate
predictive ability



$R^2 = 0.94$

excellent
predictive ability

Métricas Regressão - Função de Custo

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

Mean Squared Error - MSE

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Absolute Error - MAE

Onde:

- y_i são os valores reais (observados),
- $\hat{y}_i$ são os valores previstos pelo modelo,
- n é o número de observações.

Métricas Regressão - Função de Custo

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

Root Mean Squared Error - RMSE

Onde:

- y_i são os valores reais (observados),
- $\hat{y}_i$ são os valores previstos pelo modelo,
- n é o número de observações.

Métricas Regressão - Função de Custo

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]$$

Cross-Entropy Loss

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\max(0, 1 - y^{(i)} \cdot h_\theta(x^{(i)}))$$ Hinge Loss

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(1-h_\theta(x^{(i)}))^{\gamma} \cdot y^{(i)}\log(h_\theta(x^{(i)}))$$ Focal Loss

Regressão Linear - Aplicações

1. Previsão de preços de imóveis no mercado imobiliário.
2. Previsão da demanda por energia elétrica.
3. Estimativa de salários com base em experiência e habilidades.
4. Previsão do crescimento econômico de um país (PIB).
5. Previsão de poluição do ar com base em fatores ambientais.
6. Estimativa de despesas médicas em seguros de saúde.
7. Previsão de consumo de combustível de veículos.
8. Estimativa da produção agrícola com base em dados climáticos.
9. Previsão de temperatura e clima.
10. Modelagem de crescimento populacional.

Regressão Linear

$$y_i = b + \underbrace{w_1 x_{i1} + \cdots + w_p x_{ip}}_{\mathbf{w} \cdot \mathbf{x}_i} .$$

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

Regressão Lasso

A principal diferença entre a **regressão linear** e a **regressão Lasso** está no termo de regularização adicionado à função de custo.

A regressão Lasso inclui um **termo de penalidade** baseado na soma dos valores absolutos dos coeficientes dos parâmetros, o que pode forçar alguns desses coeficientes a serem exatamente zero.

Isso é conhecido como **sparsity** (esparsidade), e significa que o Lasso realiza a **seleção de características** de forma automática.

Regressão Lasso

A função de custo da regressão Lasso é:

$$\text{Custo} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- y_i são os valores reais.
- $\hat{y}_i$ são os valores preditos pelo modelo.
- β_j são os coeficientes dos parâmetros.
- λ é o parâmetro de regularização, que controla a força da penalização (quanto maior o λ, maior a penalização).

Regressão Lasso

Vantagens da Regressão Lasso:

1. **Seleção automática de variáveis**: Ao penalizar os coeficientes, o Lasso tende a zerar os menos importantes.
2. **Prevenção de overfitting**: O termo de regularização evita que o modelo se ajuste excessivamente aos dados de treinamento.
3. **Melhoria da interpretabilidade**: Ao eliminar variáveis irrelevantes, o modelo se torna mais simples e mais fácil de interpretar.

Regressão Ridge

A **Regressão Ridge** é outra técnica de **regressão regularizada** usada em **machine learning**, projetada para lidar com problemas onde a **multicolinearidade** (correlações entre variáveis explicativas) pode ser um problema.

Assim como a **Regressão Lasso**, ela adiciona um termo de penalidade à função de custo da **regressão linear** para prevenir overfitting, mas, ao contrário do Lasso, a penalização na Regressão Ridge é baseada na soma dos quadrados dos coeficientes, o que faz com que esses coeficientes se aproximem de zero, mas não necessariamente sejam zerados.

Regressão Ridge

Função de Custo da Regressão Ridge:

A função de custo da Regressão Ridge é:

$$\text{Custo} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- y_i são os valores reais.
- $\hat{y}_i$ são os valores preditos pelo modelo.
- β_j são os coeficientes dos parâmetros.
- λ é o parâmetro de regularização, que controla a força da penalização (quanto maior o λ, maior a penalização).

Regressão Ridge

Vantagens da Regressão Ridge:

1. **Lida bem com multicolinearidade**: Quando há correlação entre variáveis explicativas, o Ridge estabiliza a estimativa dos coeficientes.
2. **Prevenção de overfitting**: Ao adicionar o termo de regularização, o Ridge reduz a complexidade do modelo e evita que ele se ajuste excessivamente aos dados de treinamento.
3. **Melhoria da generalização**: Ao encolher os coeficientes, o modelo se torna menos sensível a pequenas flutuações nos dados de treinamento, o que melhora a performance em novos dados.