INSTITUTO SUPERIOR DE TECNOLOGIAS AVANÇADAS

Ciência de Dados

Licenciatura Engenharia Informática

2º Semestre – 2021/2022

Ricardo Jesus Ferreira

ricardojesus.ferreira@my.istec.pt

Trabalhar com Dados

- Medidas de Dispersão, Frequência, Localização
- Analise de um dataset com Python
- Bibliotecas Python para analise de dados
- Ferramentas Python para visualização de dados

Dataset - Iris

- 150 linhas
- Download a partir da UCI Machine Learning Repository
- Atributos da informação
  - Caule
    - Comprimento e largura
  - Pétalas
    - Comprimento e largura
  - Espécie

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
4.3	3.0	1.1	0.1	Iris-setosa

Média

- Em estatística, média é definida como o valor que demonstra a concentração dos dados de uma distribuição

- Seja n o número total de valores e x_i cada valor, em que $i = 1, \dots, n$

- Média aritmética é a soma dos valores x_i dividido pelo número total de valores n

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana

- É o valor que separa a metade maior e a metade menor de uma amostra, uma população ou uma distribuição de probabilidade
  - Elementos Ímpar
  - Elementos Par

- Se a lista tiver um numero ímpar de elementos - calcula-se ordenando todos os elementos da lista e escolhe-se o que fica no meio

- Se a lista tiver um numero par de elementos – calcula-se a média dos dois valores que ficam no centro da lista ordenada

Moda

- Em estatística, moda como média e mediana é uma medida de dispersão, de localização ou de tendência central que mostra a frequência dos dados.

- Ordenando os elementos de um conjunto de dados e obtém-se a moda extraindo o(s) elemento(s) com maior repetição

- Uma amostra pode ser unimodal (uma moda), bimodal (duas modas), multimodal (várias modas) e amodal (nenhuma moda)

Amplitude

- Em estatística, a amplitude representa a diferença entre o maior e o menor valor de um conjunto de dados

- Mostra a dispersão dos valores de uma série
  - Se a amplitude for um valor elevado, então os valores na série estão afastados uns dos outros
  - Se a amplitude for um número baixo, então, os valores na série estão próximos uns dos outros.

Variância

- Em estatística, a variância de uma variável aleatória ou processo estocástico é uma medida da sua dispersão estatística, indicando "o quão longe" em geral os seus valores se encontram do valor esperado
  - Uma baixa variância indica que os valores do conjunto estão mais próximos
  - Uma alta variância, indica que os valores do conjunto estão mais espaçados

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2,$$

Percentagem

- A percentagem é uma medida de frequência que identifica a proporção de um determinado valor para uma variável

- Para calcular uma percentagem, é necessário o número total de observações e o número total de observações para um valor específico de uma variável

Percentagem

- A alteração percentual dá-nos uma compreensão de como uma medida muda com o tempo. Pode calcular a alteração relativa subtraindo o valor inicial do valor final e, em seguida, dividindo-se pelo valor absoluto do valor inicial

$$\text{Percentage change} = \frac{\Delta V}{V_1} = \frac{V_2 - V_1}{V_1} \times 100\%.$$

Percentagem

- A diferença percentual compara dois valores para uma variável

- Calcula-se a diferença percentual subtraindo o valor inicial do valor final e, em seguida, dividindo-se pela média dos dois valores

$$\text{Relative change}(x, x_{\text{reference}}) = \frac{\text{Actual change}}{|x_{\text{reference}}|} = \frac{\Delta}{|x_{\text{reference}}|} = \frac{x - x_{\text{reference}}}{|x_{\text{reference}}|}.$$

Intervalo de Confiança

- Um intervalo de confiança descreve a possibilidade de uma amostra conter o verdadeiro parâmetro populacional numa gama de valores em torno da média

$$\bar{X} \pm Z \frac{s}{\sqrt{n}}$$

Correlação

- É uma medida estatística que indica se duas variáveis são linearmente dependentes

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

Correlação

- É obtida através dos desvios padrão e da covariância

$$Cor(X,Y) = \frac{Cov(X,Y)}{s_x s_y}$$

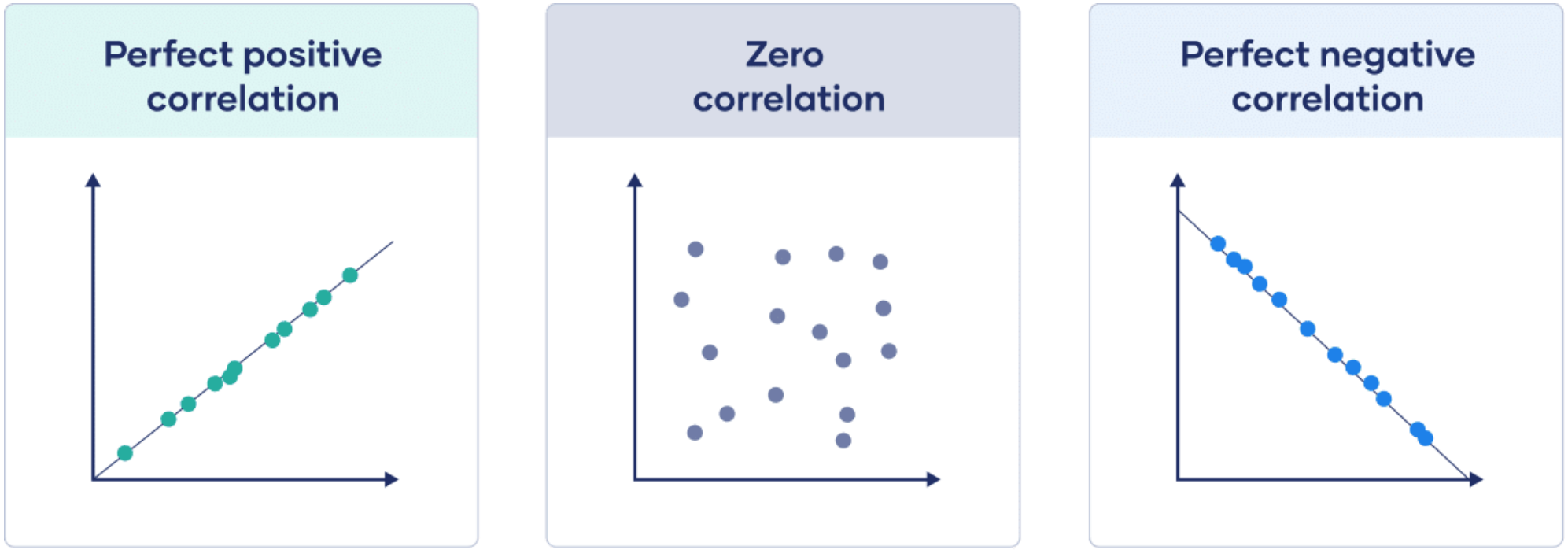
covariance

$$= \quad Cor(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

standard deviations

Correlação

- Esta pode ser positiva, negativa ou nula



Correlação

- Varia entre -1 e 1

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

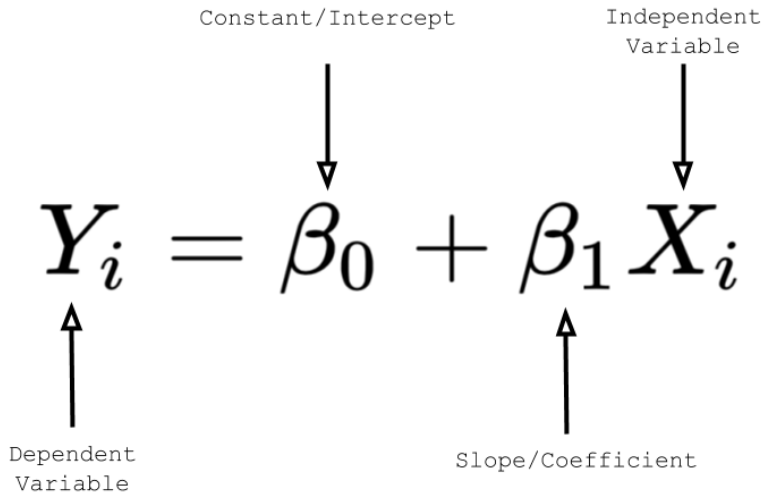
Regressão Linear

- A Regressão Linear é utilizada para estimar a variável dependente (y) baseado na variável independente (x)

- Serve para encontrar uma relação linear entre variáveis

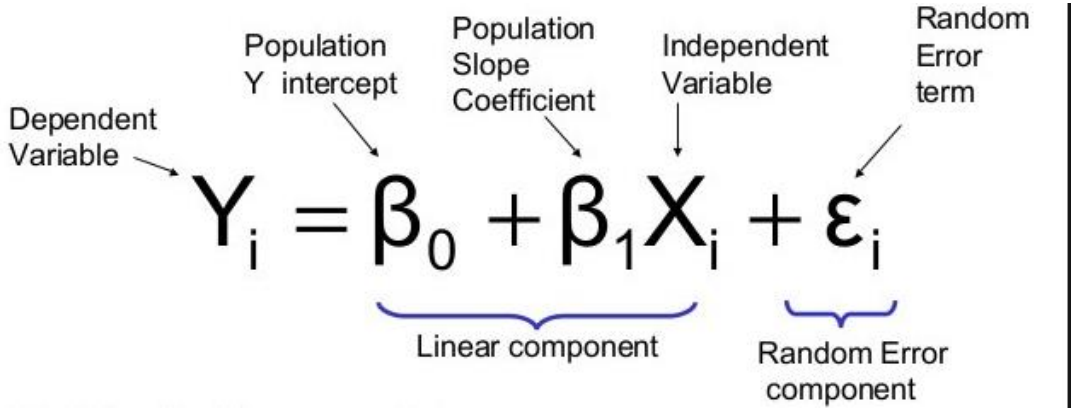
Regressão Linear

- Y - variável dependente
- X - variável independente
- β_0 - valor de y quando $x = 0$
- β_1 - Inclinação da reta



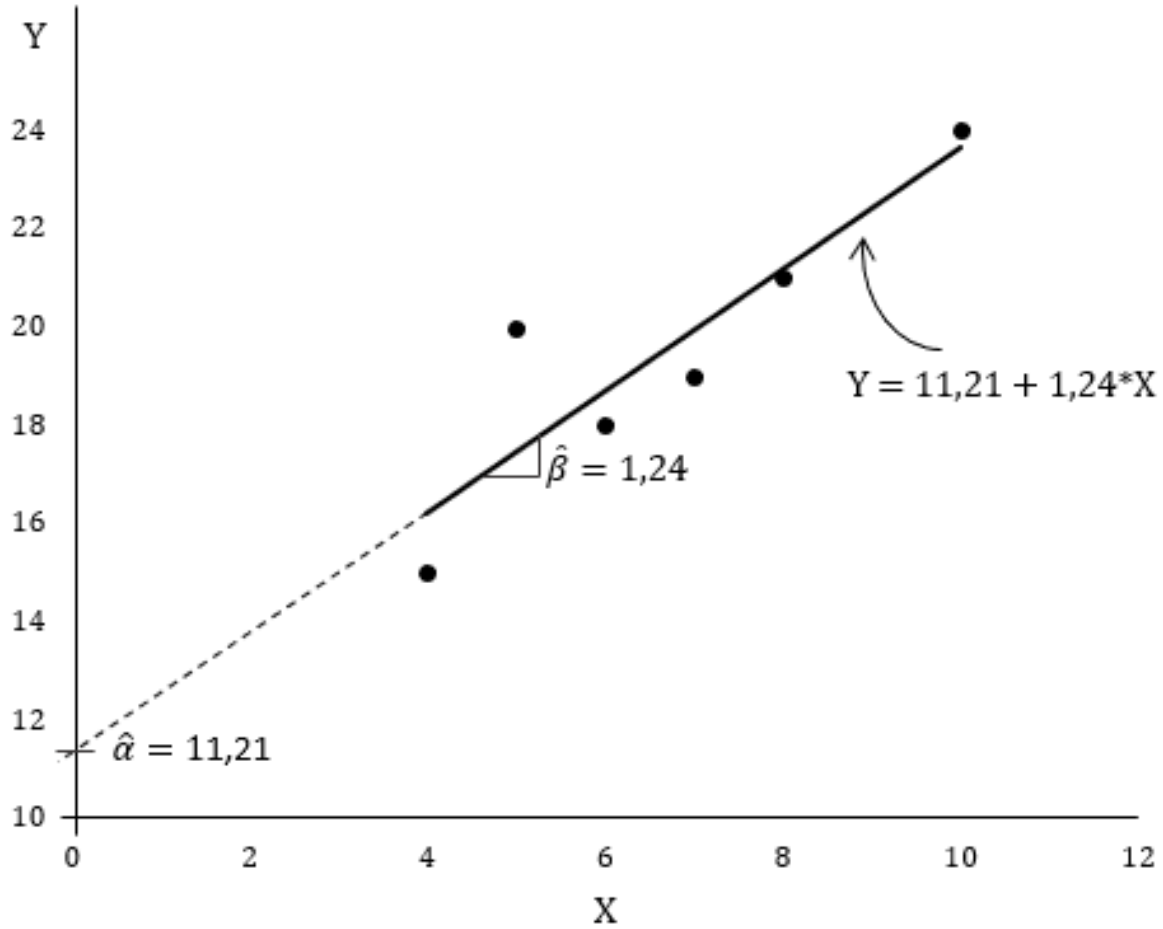
Regressão Linear

- Numa estimação não podemos ignorar a existência de erros
- Este erro representa a diferença entre a realidade dos dados e os valores estimados



Regressão Linear

X	4	6	7	5	8	10
Y	15	18	19	20	21	23



Regressão Linear

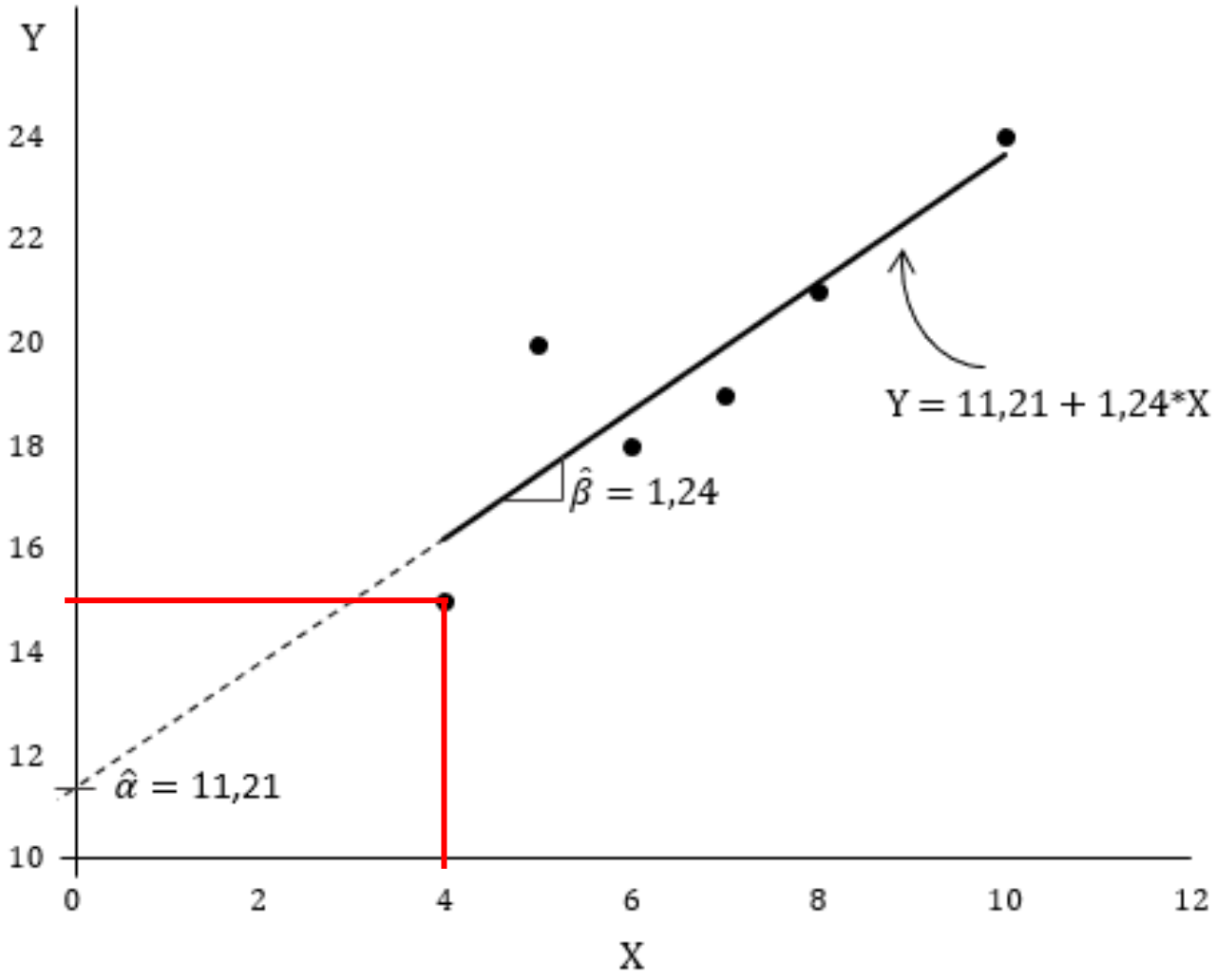
$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$y = 11{,}21 + 1{,}24x + \varepsilon$$

$$15 = 11{,}21 + 1{,}24x + \varepsilon$$

$$\varepsilon = 11{,}21 + 1{,}24 * 4 - 15$$

$$\varepsilon = 1{,}17$$



Regressão Linear

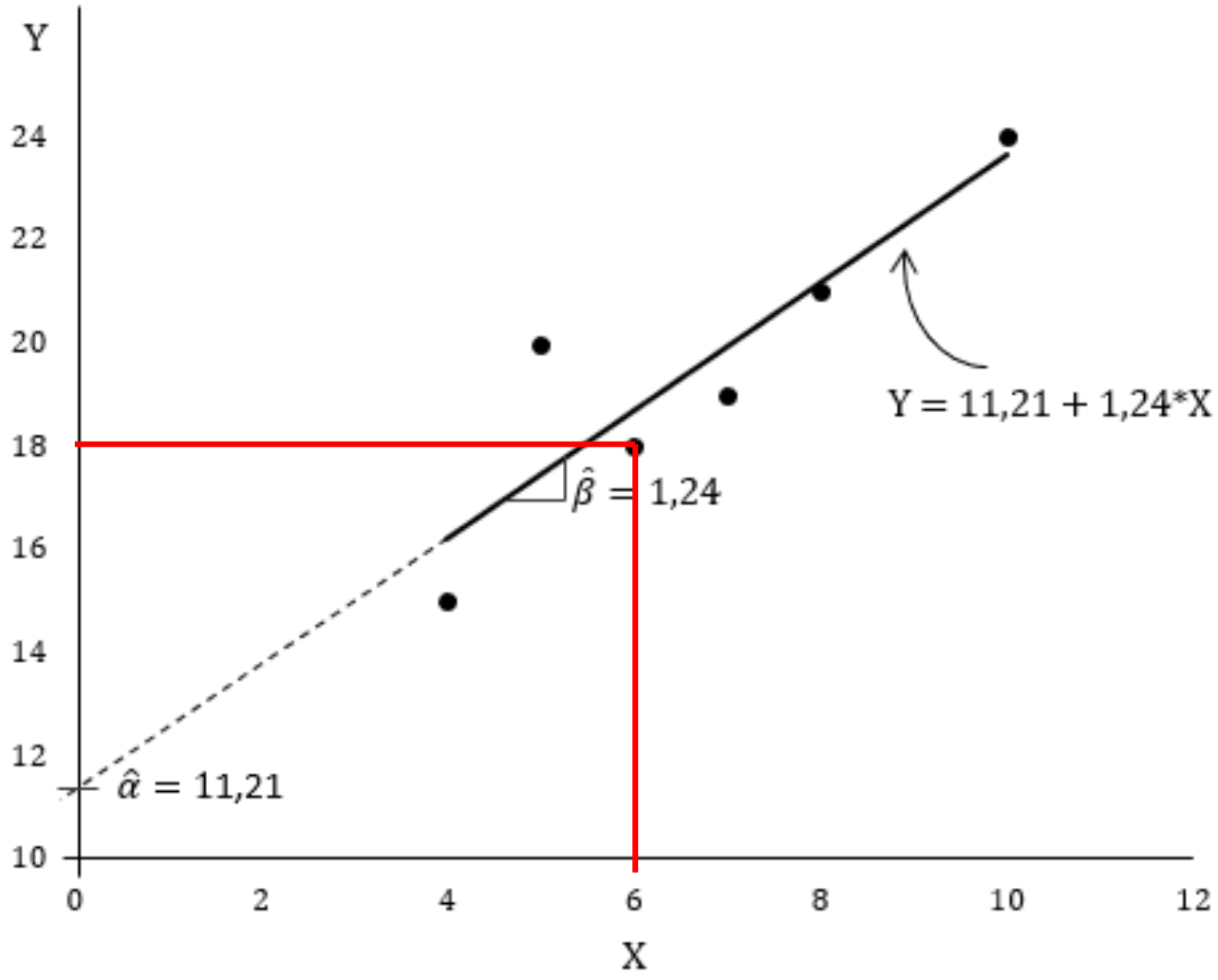
$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$y = 11{,}21 + 1{,}24x + \varepsilon$$

$$18 = 11{,}21 + 1{,}24x + \varepsilon$$

$$\varepsilon = 11{,}21 + 1{,}24 * 6 - 18$$

$$\varepsilon = 0{,}65$$



Bibliografia

- B. Gomez,(2020) “Resolviendo problemas de Big Data”, Alfaomega.
- D. Insua, (2019)“Big data: Conceptos, tecnologías y aplicaciones”, CSIC.
- H. Jones, (2019)“Analítica de datos”, HJ,.
- J. Somed, (2020)“Big Data Analytics”, JLC.
- D. Petković (2020)“Microsoft® SQL Server® 2019 A Beginner’s Guide - Seventh Edition”, McGraw Hill.