Profissão: Cientista de Dados

Análise de Regressão

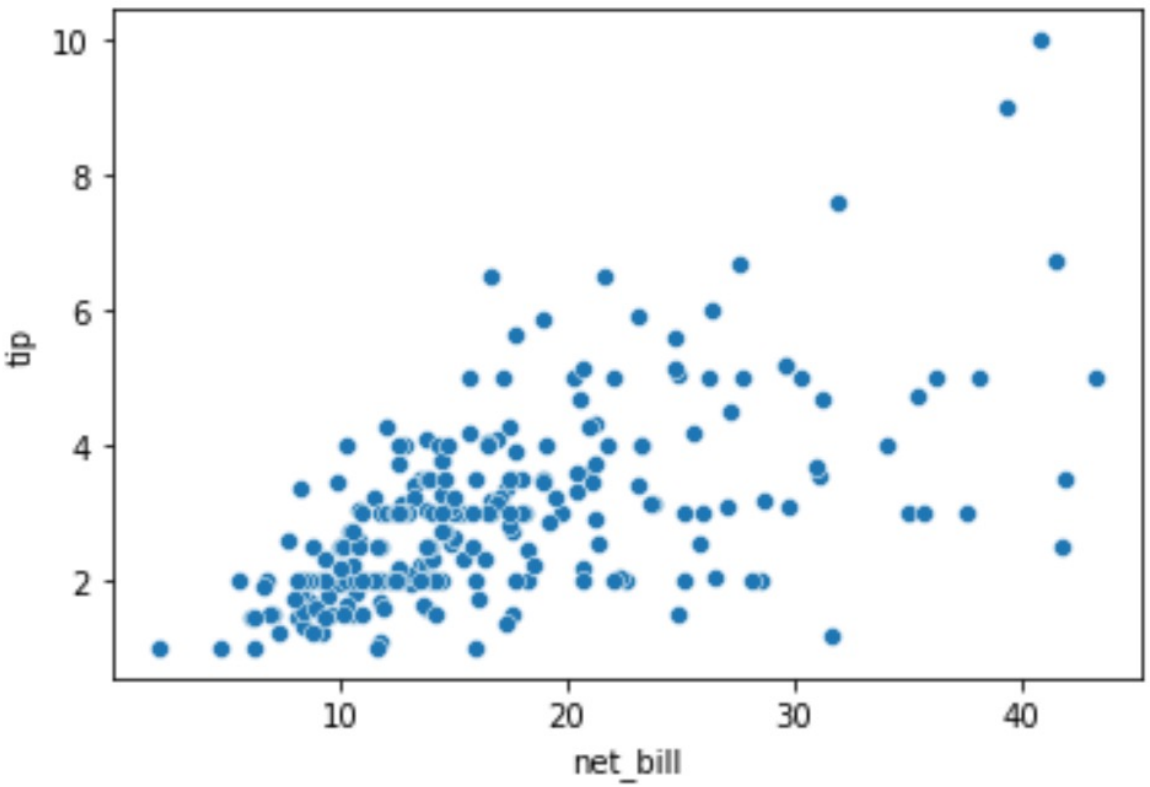
O que é um modelo?

O que é um modelo?

Base e premissas

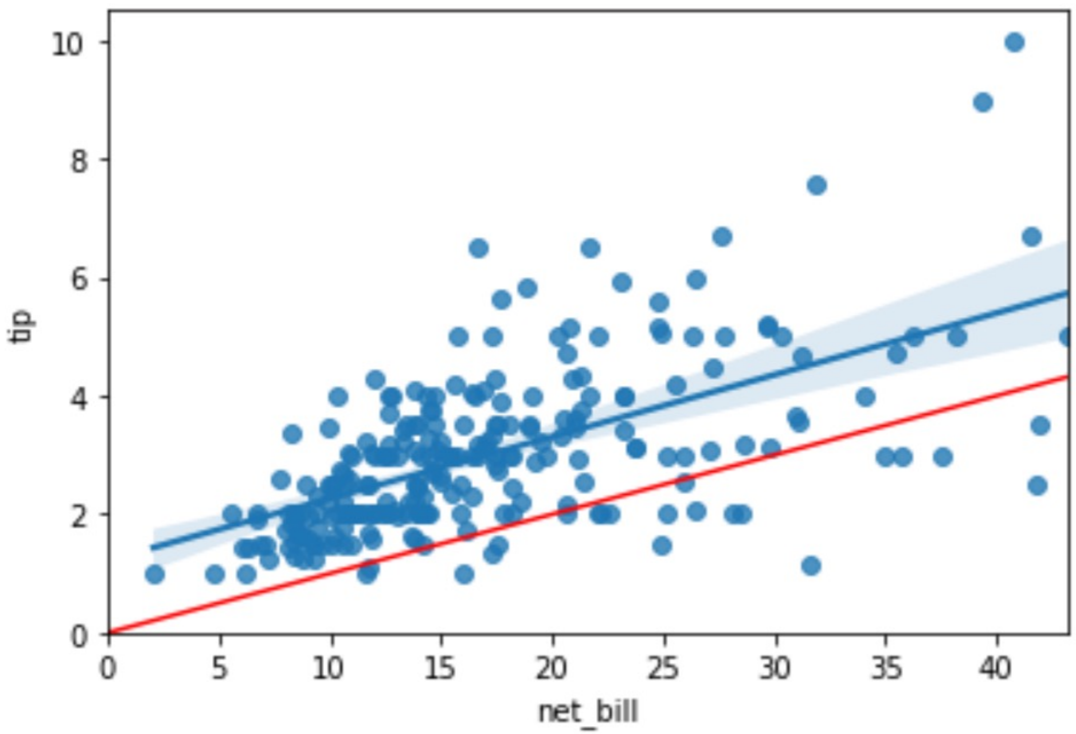
	total_bill	tip	sex	smoker	day	time	size	tip_pct
0	16.99	1.01	Female	No	Sun	Dinner	2	0.063204
1	10.34	1.66	Male	No	Sun	Dinner	3	0.191244
2	21.01	3.50	Male	No	Sun	Dinner	3	0.199886
3	23.68	3.31	Male	No	Sun	Dinner	2	0.162494
4	24.59	3.61	Female	No	Sun	Dinner	4	0.172069

O que é um modelo?



Introdução

Introdução



Y = tip_pct (% de gorjeta)
X = net_bill (valor da conta)

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

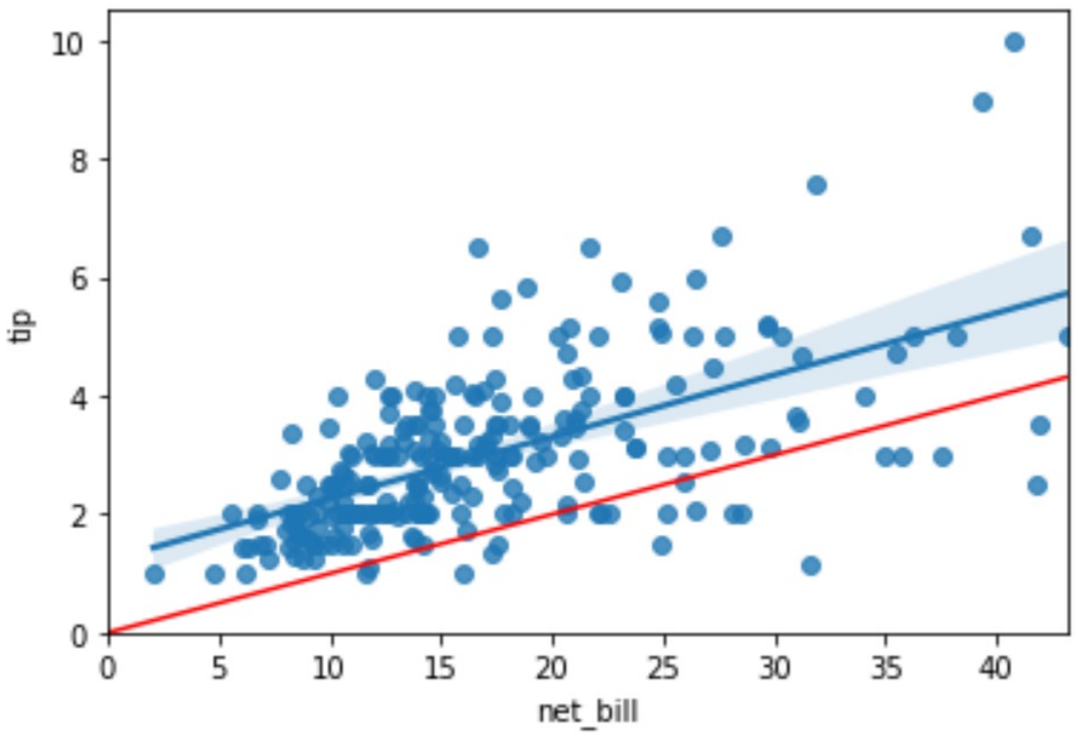
Com i = 1, ... N
$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

α e β são constantes desconhecidas.

Usando estimativas pontuais temos:

$$y = \hat{\alpha} + \hat{\beta} x + \hat{\varepsilon}$$

Entendendo a equação



$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$\hat{\alpha}$ = 1,33
$\hat{\beta}$ = 0,10

Se x = 10:

$$\hat{y} = 1{,}33 + 0{,}10 * 10 = 2{,}33$$

Se x = 20:

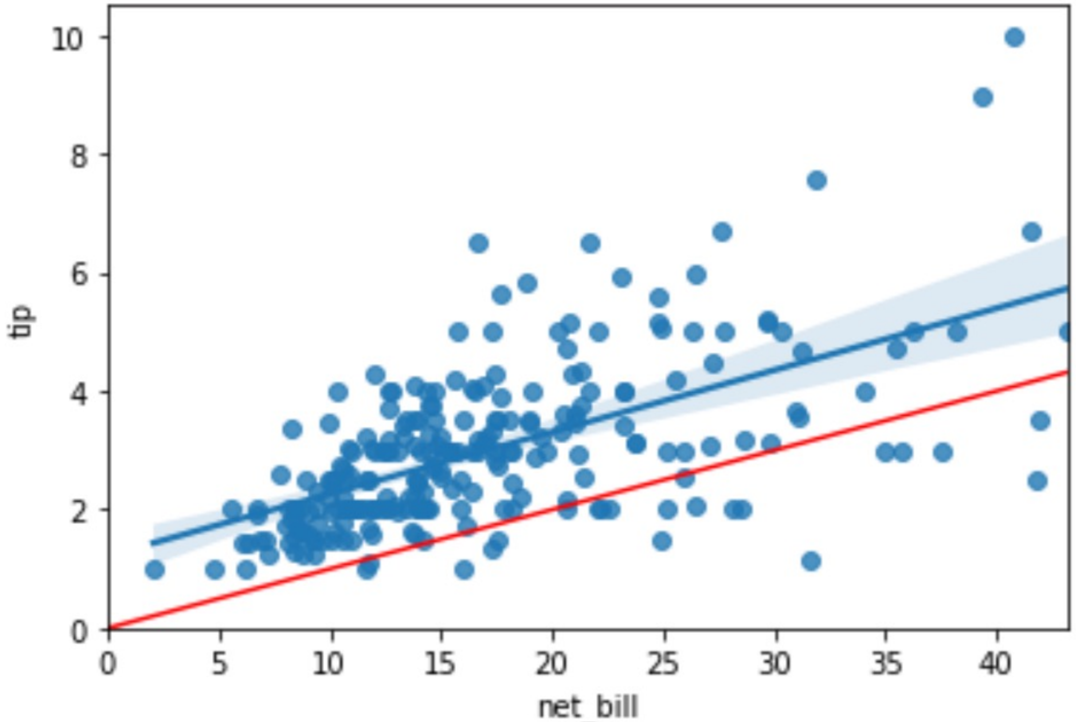
$$\hat{y} = 1{,}33 + 0{,}10 * 20 = 3{,}33$$

Se x = 30:

$$\hat{y} = 1{,}33 + 0{,}10 * 30 = 4{,}33$$

Interpretação dos parâmetros

Interpretação dos parâmetros



$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Com i = 1, ... N
$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

- α é o valor esperdo de y quando x é zero.
- β é o aumento esperado em y para cada unidade que se incrementa em x.
- ε é um erro aleatório do valor de y, quando predito pelo modelo.
- σ^2 é a variância dos erros ε_i

Isto é um modelo!

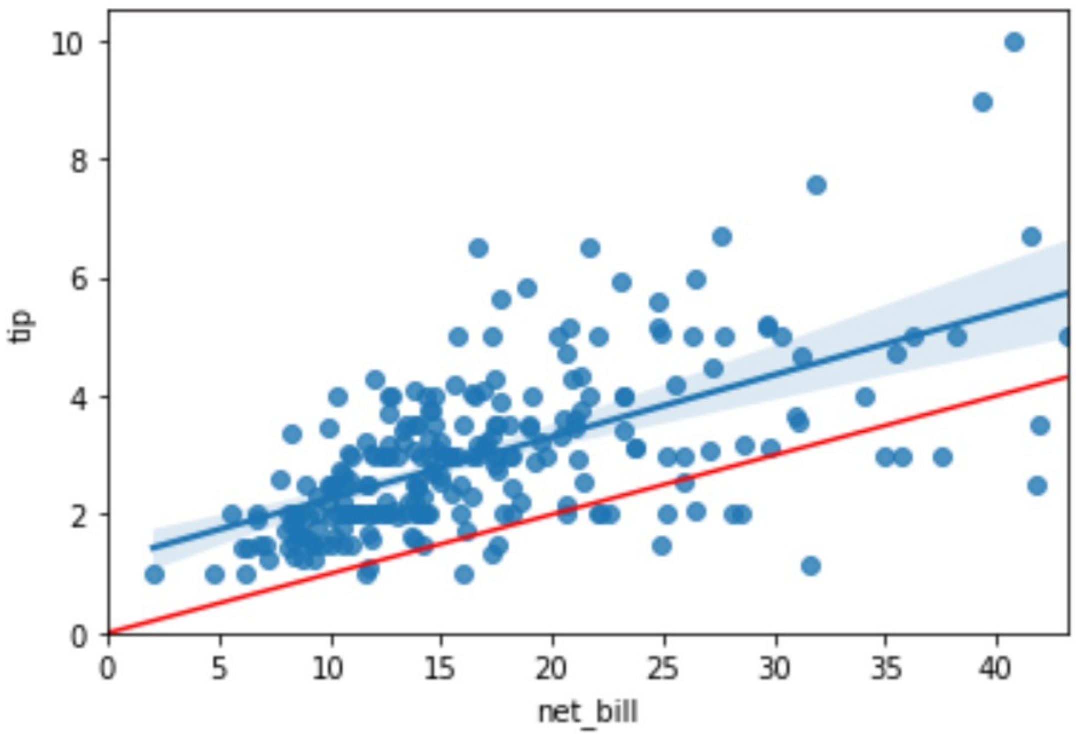
“All models are wrong,
but some are useful”

“Todos os modelos estão errados,
mas alguns são úteis”

George E. P. Box

Obtendo as estimativas

Os erros



Y = tip_pct (% de gorjeta)
X = net_bill (valor da conta)

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Com i = 1, ... N
$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

α e β são constantes desconhecidas.

Usando estimativas pontuais temos:

$$y = \hat{\alpha} + \hat{\beta} x + \hat{\varepsilon}$$

Os erros

$$\left.\begin{aligned} \hat{y}_i &= \hat{\alpha} + \hat{\beta} x_i \\ y_i &= \hat{\alpha} + \hat{\beta} x_i + \hat{\varepsilon}_i \end{aligned}\right\} \quad y_i = \hat{y}_i + \hat{\varepsilon}_i$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad \Rightarrow \qquad \hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

$$QME = \frac{1}{N} \sum_{i}^{N} \hat{\varepsilon}_i^{\,2} \qquad SQE = \sum_{i}^{N} \hat{\varepsilon}_i^{\,2}$$

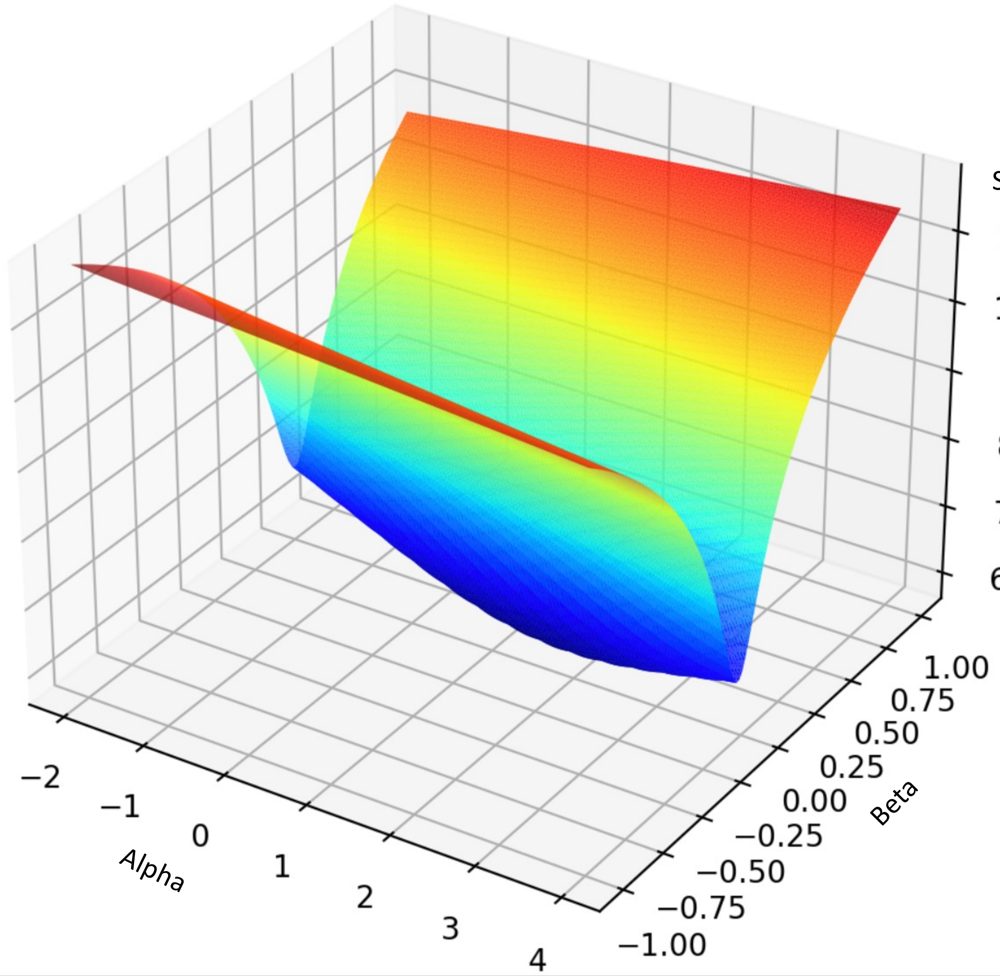
Minimizando os erros

$$SQE = \sum_{i}^{N} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i)\right)^2$$

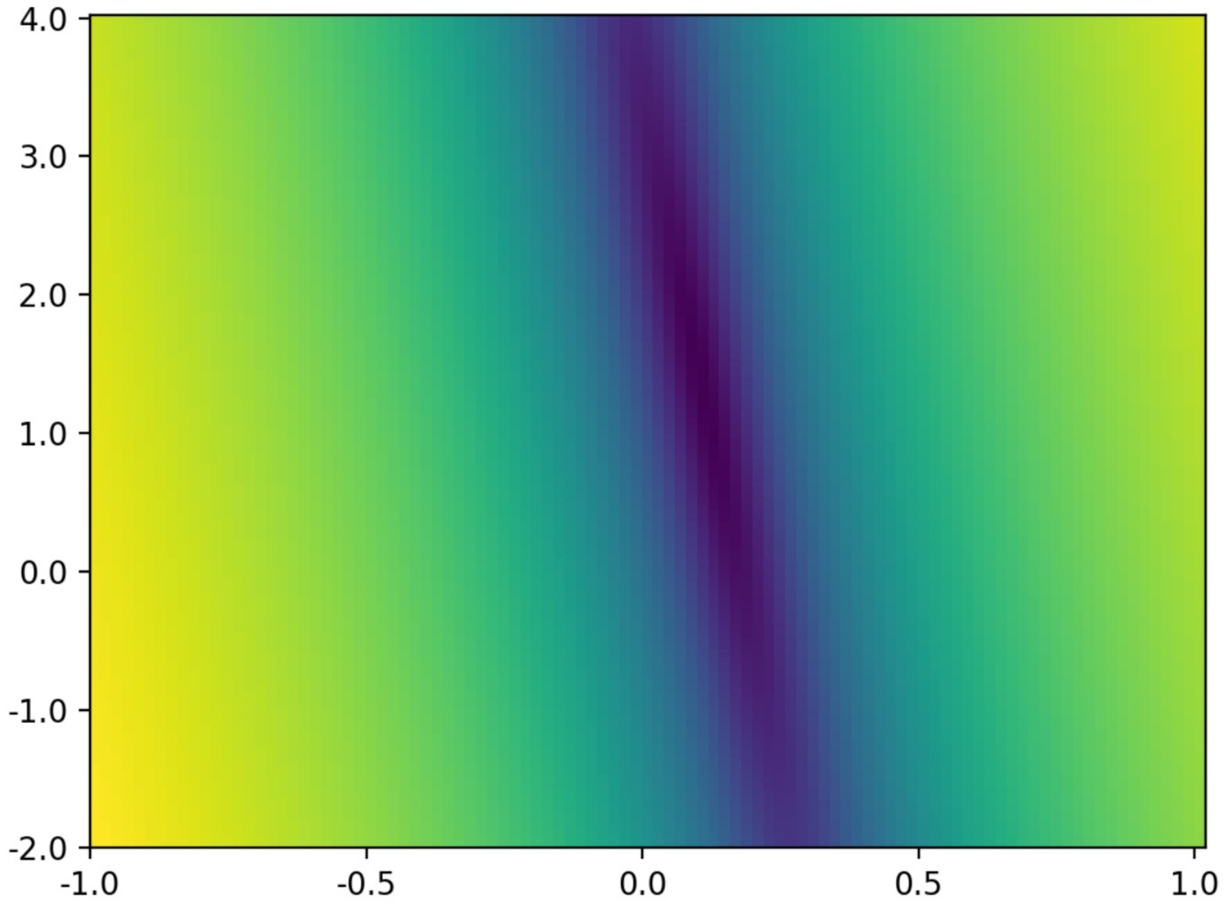
$\widehat{\alpha}$ e $\widehat{\beta}$ correspondem a a e b que minimizam:

$$SQE = \sum_{i}^{N} \left(y_i - (a + b x_i)\right)^2$$

Erro em função dos parâmetros



Erro em função dos parâmetros



Estimador de mínimos quadrados

$$\frac{\partial SQE}{\partial \alpha} = 0 \qquad \Rightarrow \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\frac{\partial SQE}{\partial \beta} = 0 \qquad \Rightarrow \qquad \hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})}$$

Propriedades dos estimadores de mínimos quadrados

- A soma dos resíduos é zero $\sum_{i=1}^{N} \hat{\varepsilon}_i = 0$
- Os resíduos não têm correlação linear com os preditores
- O ponto $(\bar{x}, \bar{y})$ está sempre na reta
- A distribuição dos estimadores é conhecida

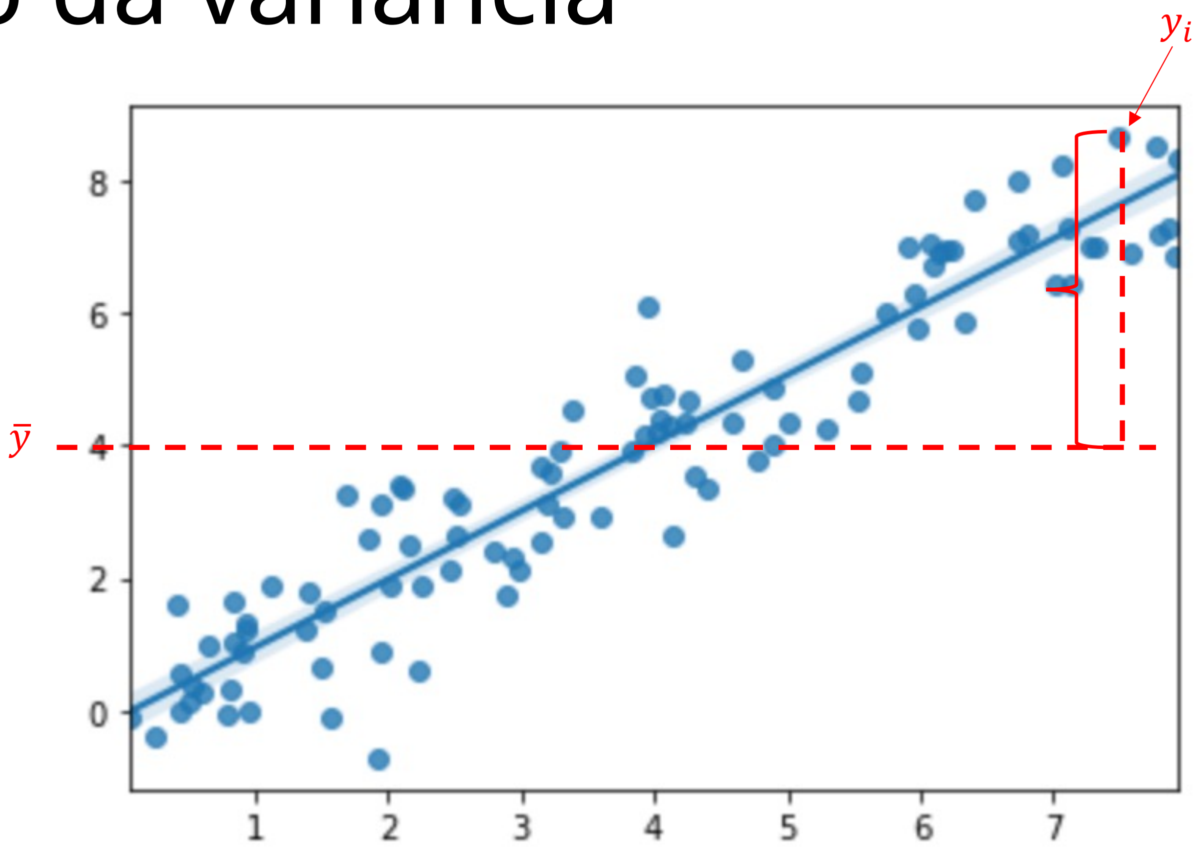
Qualidade do modelo

Decomposição da variância

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SQM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

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

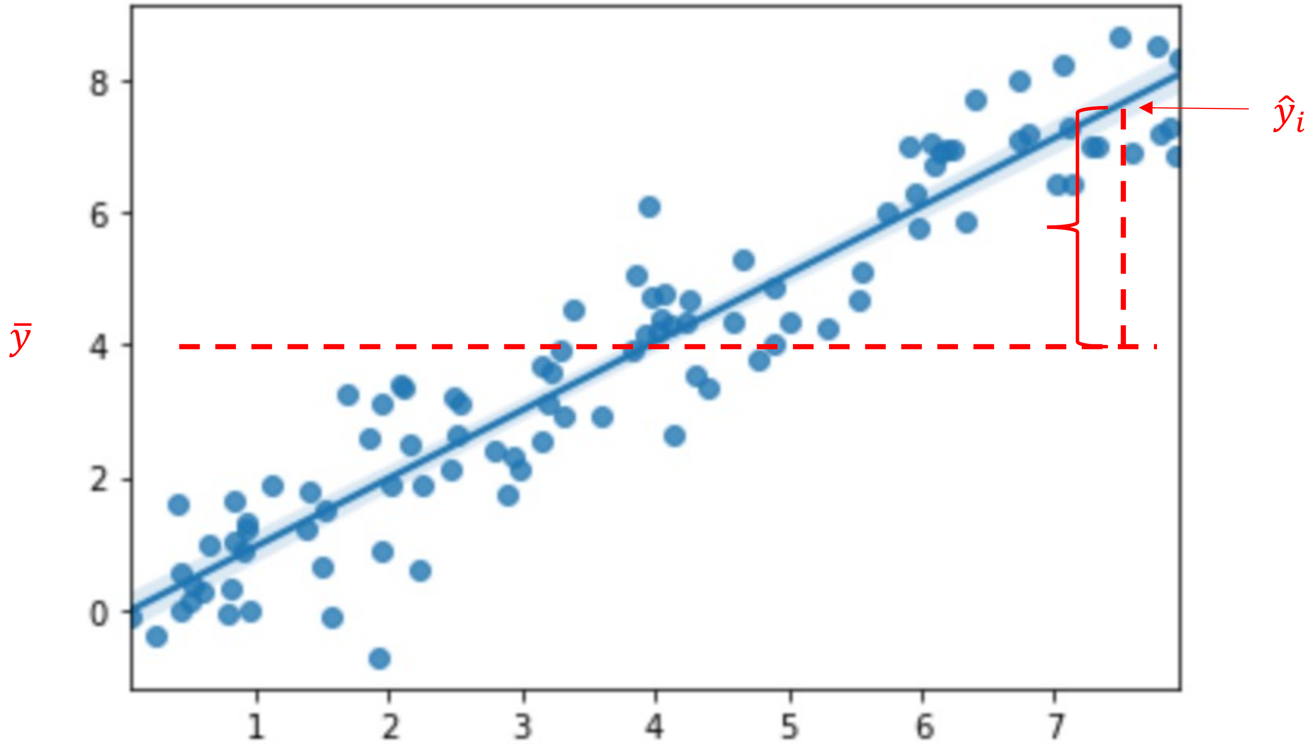


Decomposição da variância

$$\text{SQT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

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



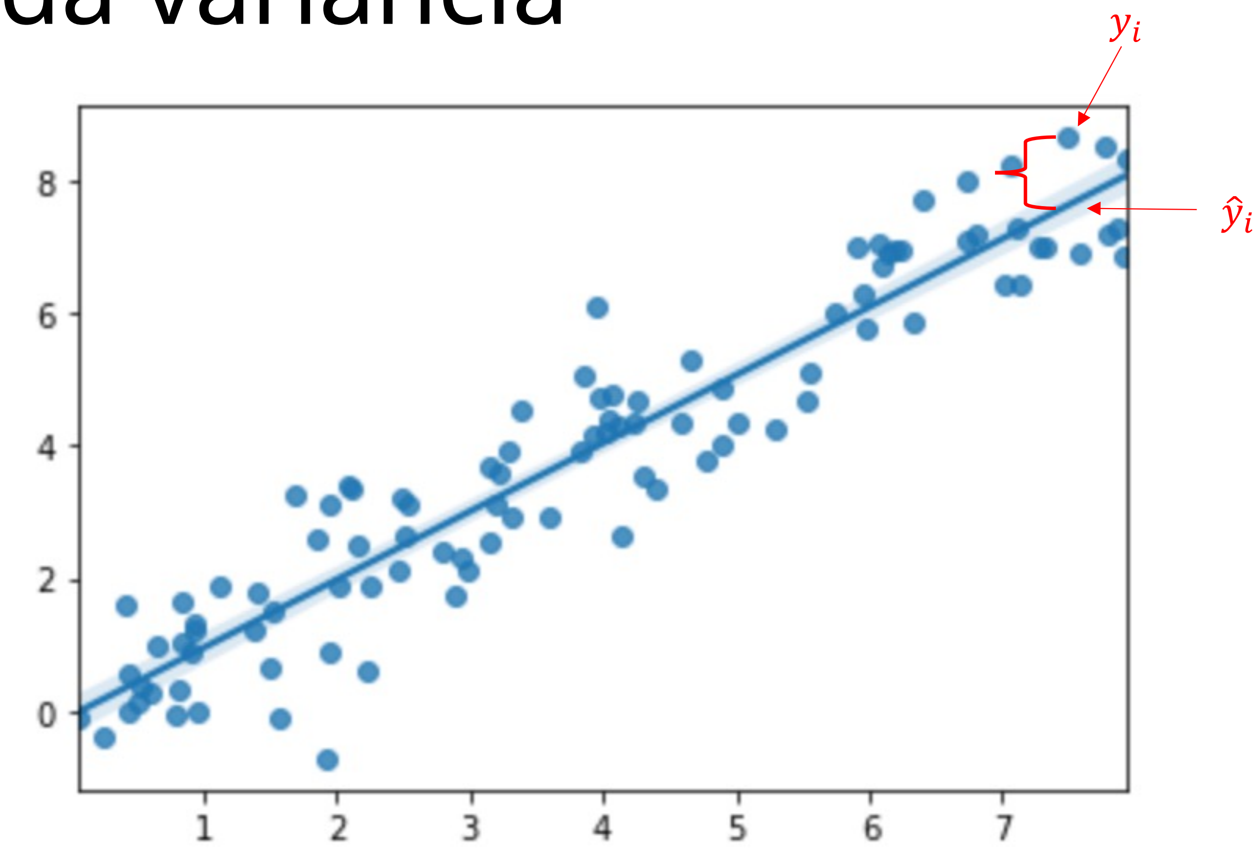
Decomposição da variância

$$\text{SQT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

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

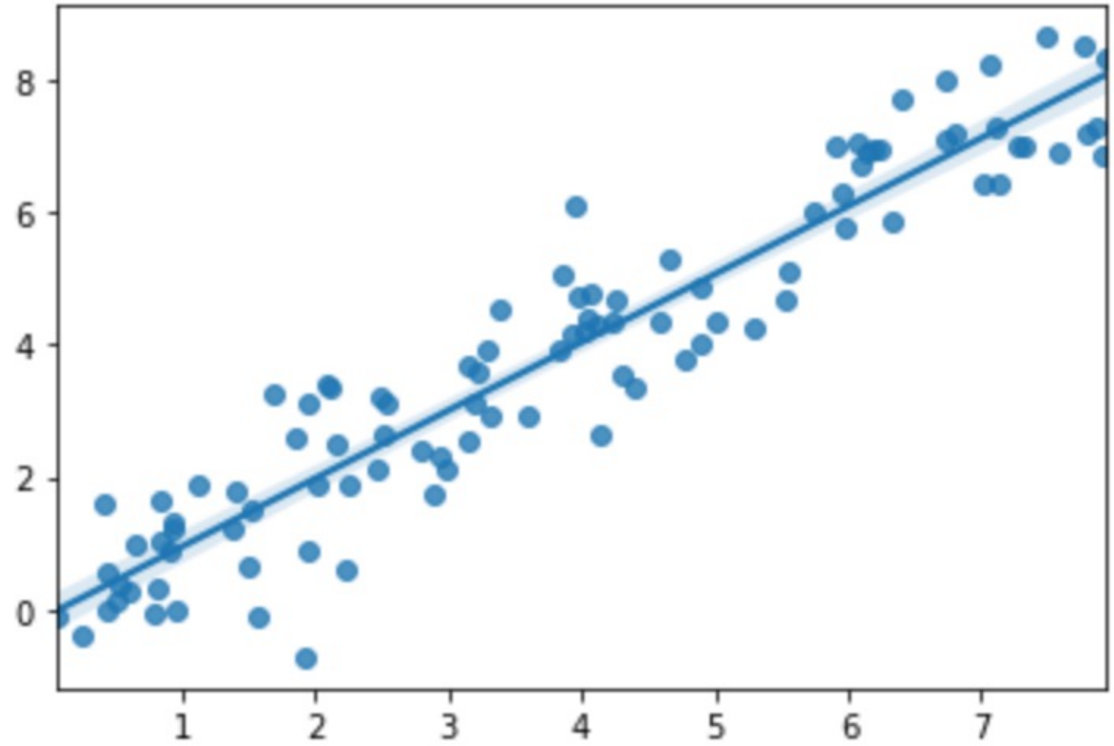
$$\text{SQT} = \text{SQM} + \text{SQE}$$



Coeficiente de determinação

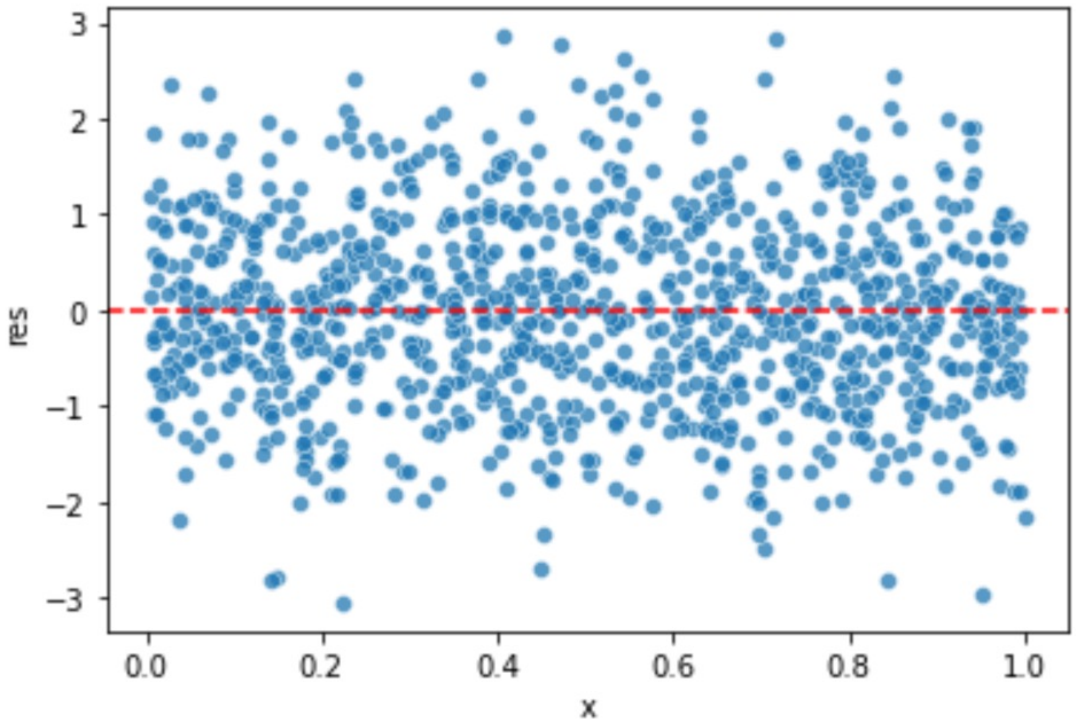
R^2 – coeficiente de determinação

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQE}{SQT}$$



Análise de Resíduos

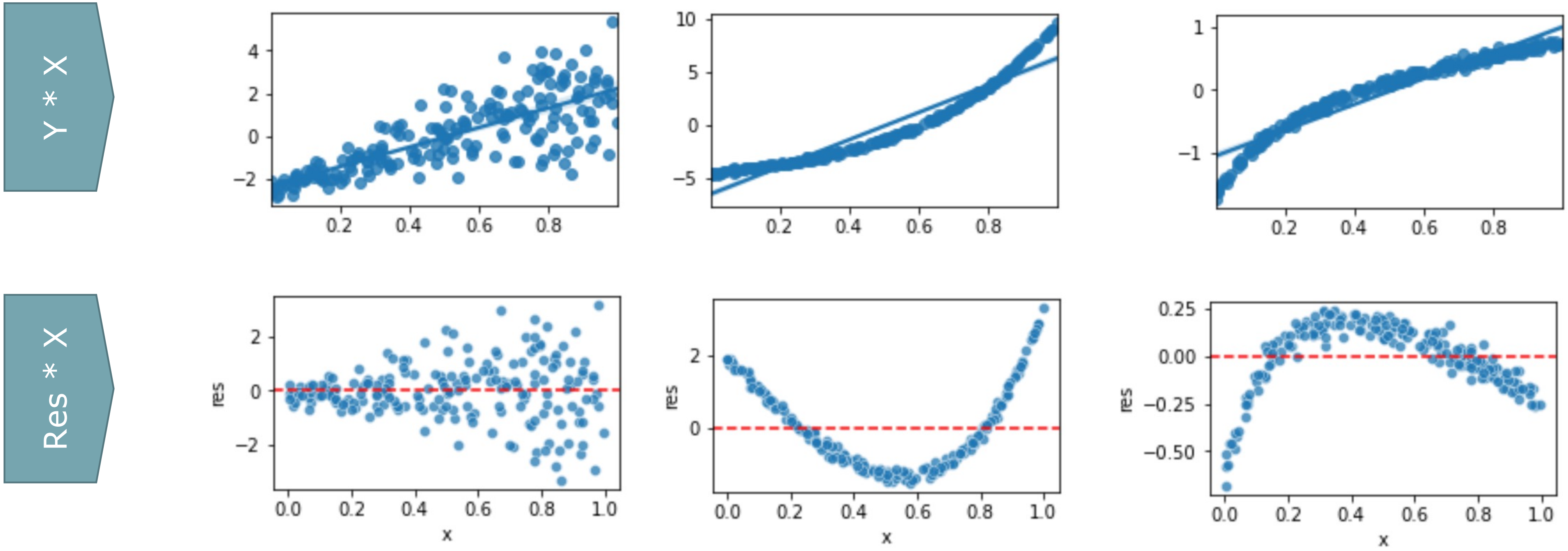
Análise de resíduos



Desejado:

- Nenhum padrão evidente
- Aspecto de independência
- Variância uniforme

Análise de resíduos



Resíduo aumenta | Relação convexa | Relação côncava

Multiplicação de matrizes

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 7 & 15 \\ 10 & 22 \end{pmatrix}$$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \underset{\color{blue}{3} \times 2}{\begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}} \underset{2 \times \color{red}{2}}{\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}} = \underset{\color{blue}{3} \times \color{red}{2}}{\begin{pmatrix} 5 & 11 \\ 7 & 15 \\ 10 & 22 \end{pmatrix}}$$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} \mathbf{1} & \mathbf{2} \\ 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} \mathbf{1} & 3 \\ \mathbf{2} & 4 \end{pmatrix} = \begin{pmatrix} \mathbf{5} & 11 \\ 7 & 15 \\ 10 & 22 \end{pmatrix}$$

$$1 \times 1 + 2 \times 2 = 5$$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} 1 & 2 \\ \boxed{\color{green}{\mathbf{1}} \quad \color{green}{\mathbf{3}}} \\ 2 & 4 \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix} \color{blue}{\mathbf{1}} \\ \color{blue}{\mathbf{2}} \end{matrix}} & \begin{matrix} 3 \\ 4 \end{matrix} \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ \color{red}{\mathbf{7}} & 15 \\ 10 & 22 \end{pmatrix}$$

$\color{green}{1} \times \color{blue}{1} + \color{green}{3} \times \color{blue}{2} = \color{red}{7}$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ \boxed{\mathbf{2} \quad \mathbf{4}} \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix}\mathbf{1} \\ \mathbf{2}\end{matrix}} & \begin{matrix}3 \\ 4\end{matrix} \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 7 & 15 \\ \mathbf{10} & 22 \end{pmatrix}$$

$$2 \times 1 + 4 \times 2 = 10$$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} \boxed{\mathbf{1} \quad \mathbf{2}} \\ 1 \quad 3 \\ 2 \quad 4 \end{pmatrix} \begin{pmatrix} 1 & \boxed{\begin{matrix}\mathbf{3} \\ \mathbf{4}\end{matrix}} \\ 2 & \end{pmatrix} = \begin{pmatrix} 5 & \mathbf{11} \\ 7 & 15 \\ 10 & 22 \end{pmatrix}$$

$1 \times 3 + 2 \times 4 = 11$

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} 1 & 2 \\ \boxed{\mathbf{1} \quad \mathbf{3}} \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & \boxed{\begin{matrix} \mathbf{3} \\ \mathbf{4} \end{matrix}} \\ 2 & \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 7 & \mathbf{15} \\ 10 & 22 \end{pmatrix}$$

1 x 3 + 3 x 4 = 15

Multiplicação de matrizes

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

$$A.B = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ \boxed{\mathbf{2} \quad \mathbf{4}} \end{pmatrix} \begin{pmatrix} 1 & \boxed{\mathbf{3}} \\ 2 & \boxed{\mathbf{4}} \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 7 & 15 \\ 10 & \mathbf{22} \end{pmatrix}$$

$2 \times 3 + 4 \times 4 = 22$

Multiplicação de matrizes e a regressão

$$\mathrm{X} = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 4 \end{pmatrix}$$

$$\mathrm{X}.\,\widehat{\mathrm{B}} = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \hat{\alpha} + 3\hat{\beta} \\ \hat{\alpha} + 2\hat{\beta} \\ \hat{\alpha} + 4\hat{\beta} \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{pmatrix}$$

$$\widehat{\mathrm{B}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$$

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Matrizes e regressão múltipla

$$X = \begin{pmatrix} 1 & 3 & 1 \\ 1 & 2 & 0 \\ 1 & 4 & 0 \end{pmatrix}$$

$$\widehat{B} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

$$X.\widehat{B} = \begin{pmatrix} 1 & 3 & 1 \\ 1 & 2 & 0 \\ 1 & 4 & 0 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\alpha} + 3\hat{\beta}_1 + 1\hat{\beta}_2 \\ \hat{\alpha} + 2\hat{\beta}_1 + 0\hat{\beta}_2 \\ \hat{\alpha} + 4\hat{\beta}_1 + 0\hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{pmatrix}$$