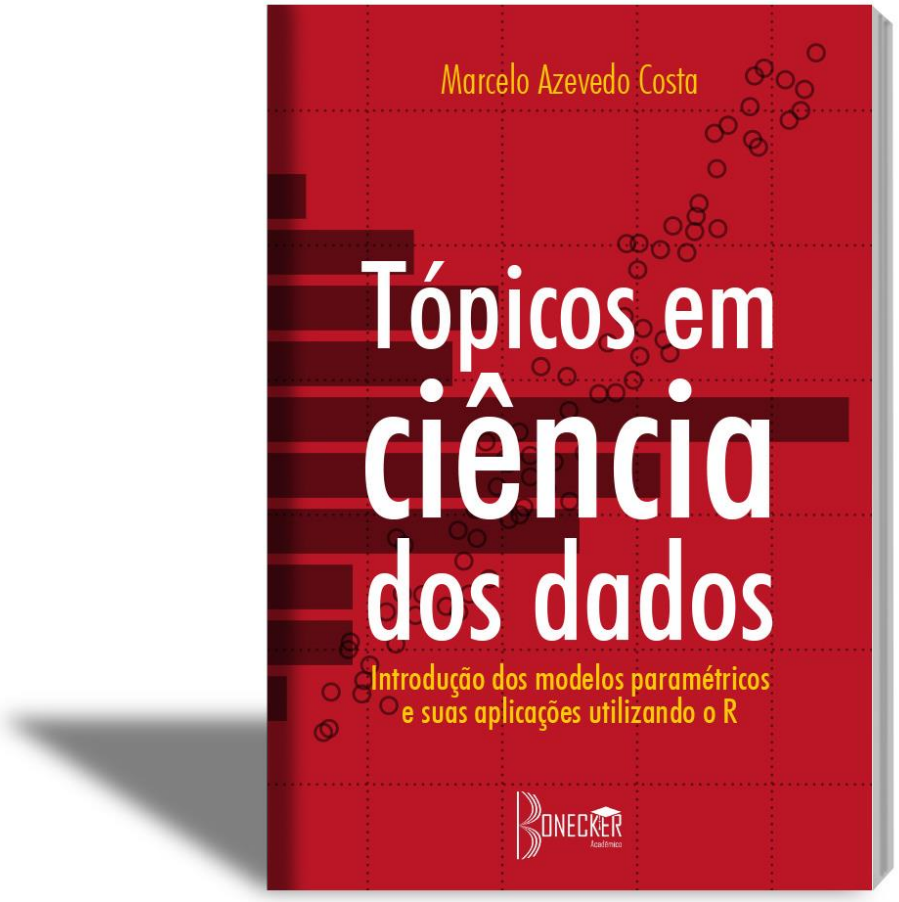
O Modelo de Regressão Linear Múltipla

Introdução à Ciência dos Dados

Marcelo Azevedo Costa
Departamento de Engenharia de Produção
Universidade Federal de Minas Gerais

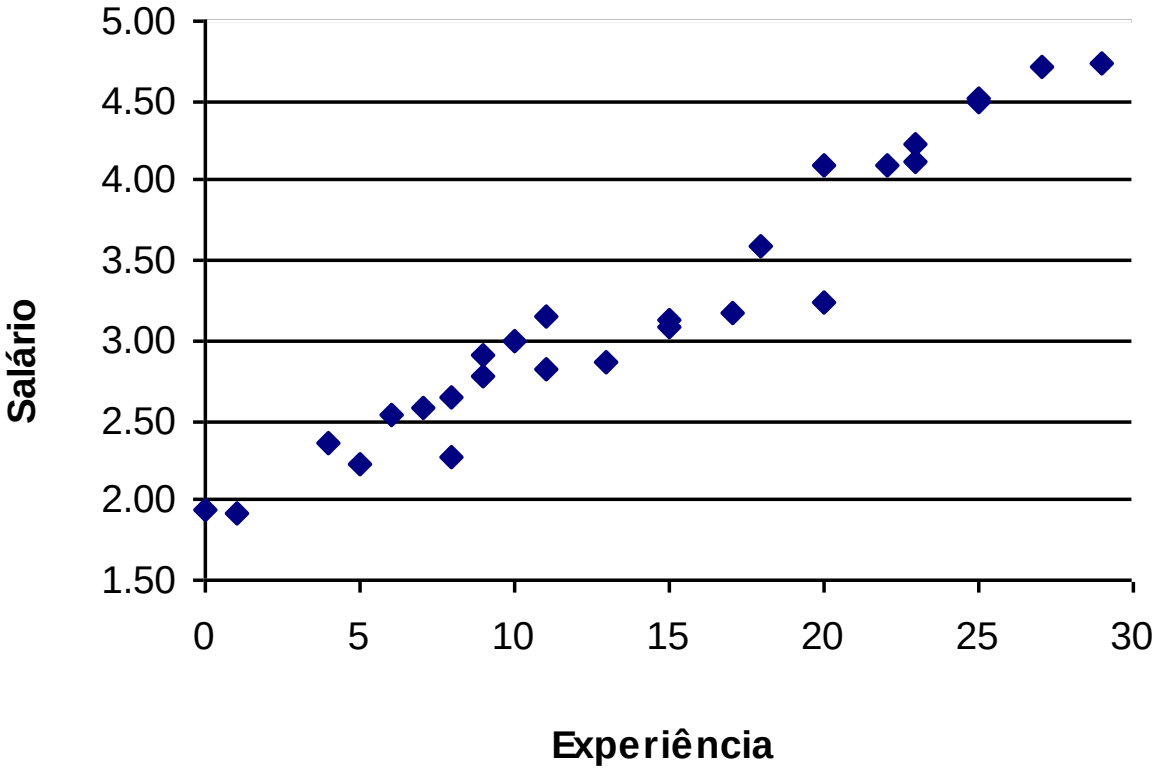
Capítulo 3: O Modelo de Regressão Linear Múltipla



Exemplo

Salario	Experiencia
1.9307	0
3.1769	17
2.2769	8
3.1307	15
2.7769	9
3.0923	15
2.6538	8
2.2230	5
2.8538	13
3.2307	20
2.8230	11
1.9076	1
2.5384	6
2.5692	7
4.2230	23
4.0923	20
3.6000	18
4.7076	27
3.1461	11
2.9923	10
4.7461	29
4.1153	23
2.3615	4
4.0923	22
4.5076	25
2.9076	9
4.4846	25

- Um investigador deseja estudar a possível relação entre o salário (em anos) e o tempo de experiência (em mil reais) no cargo de gerente de agências bancárias de uma grande empresa.



Regressão Linear Simples

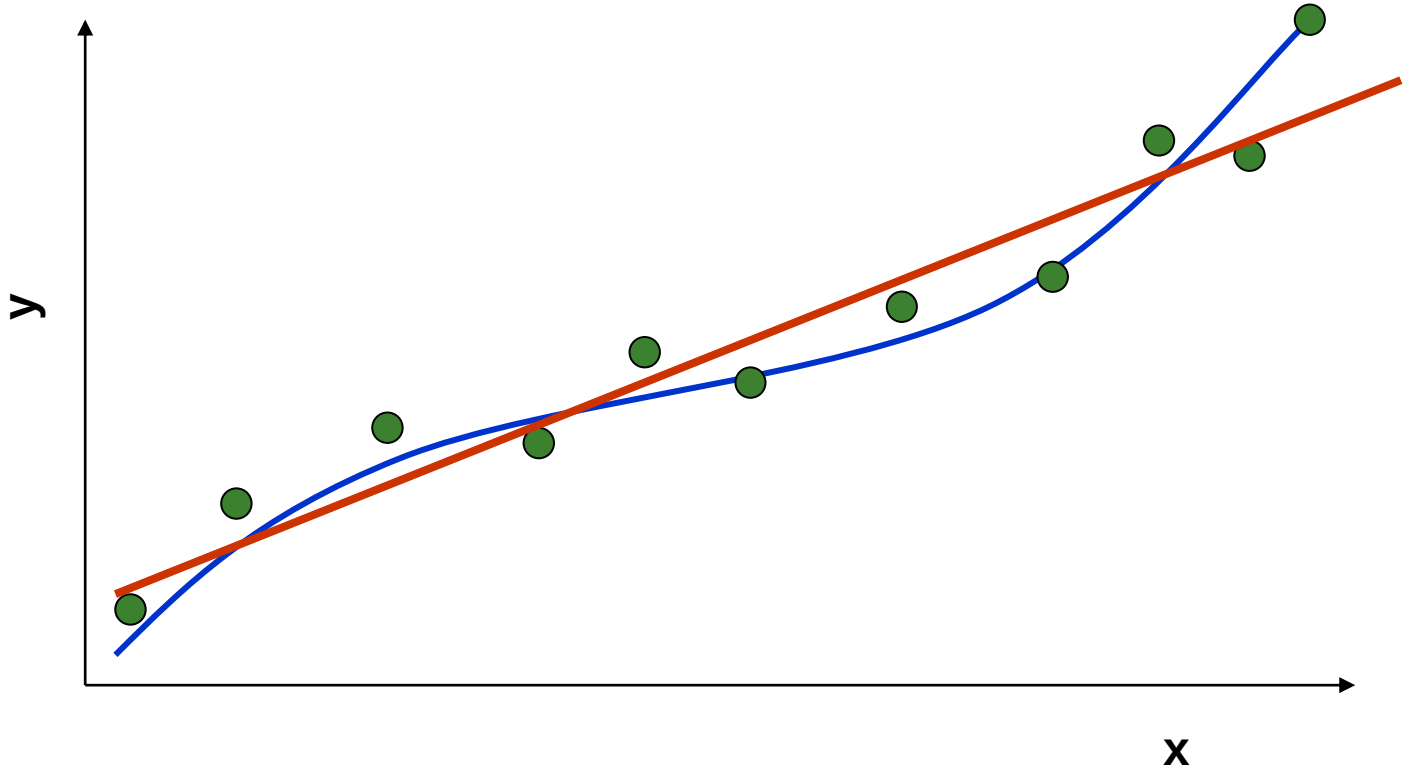
- Suposição: f(.) pode ser aproximada por uma reta

$$f(x_i) = \beta_0 + \beta_1 . x_i + \varepsilon_i$$

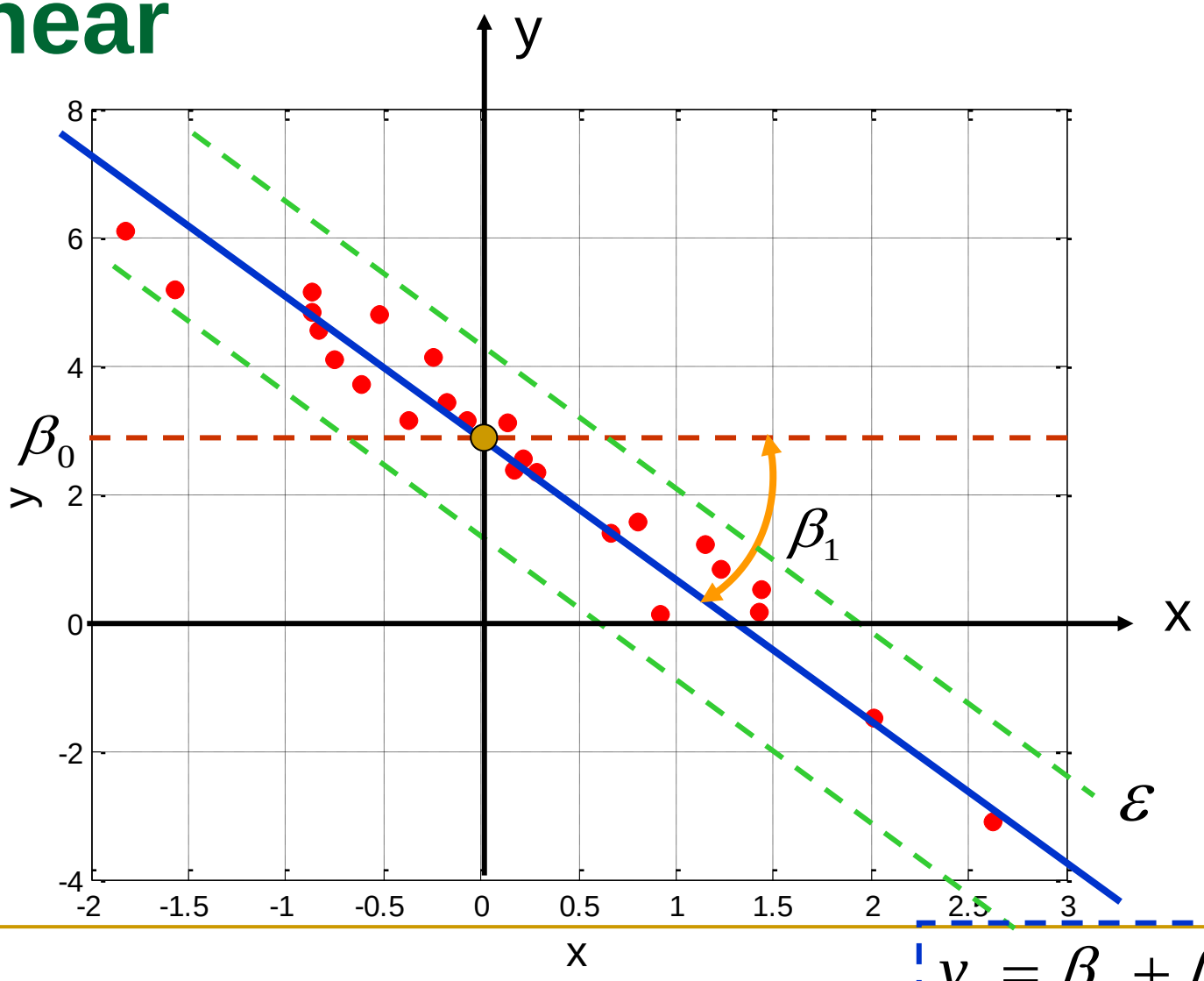
ε_i : é uma variável aleatória que expressa a não-adequação do modelo e componentes não explicadas pela reta de regressão

A escolha do modelo

- O modelo é uma aproximação simplificada da relação real entre as variáveis de interesse.

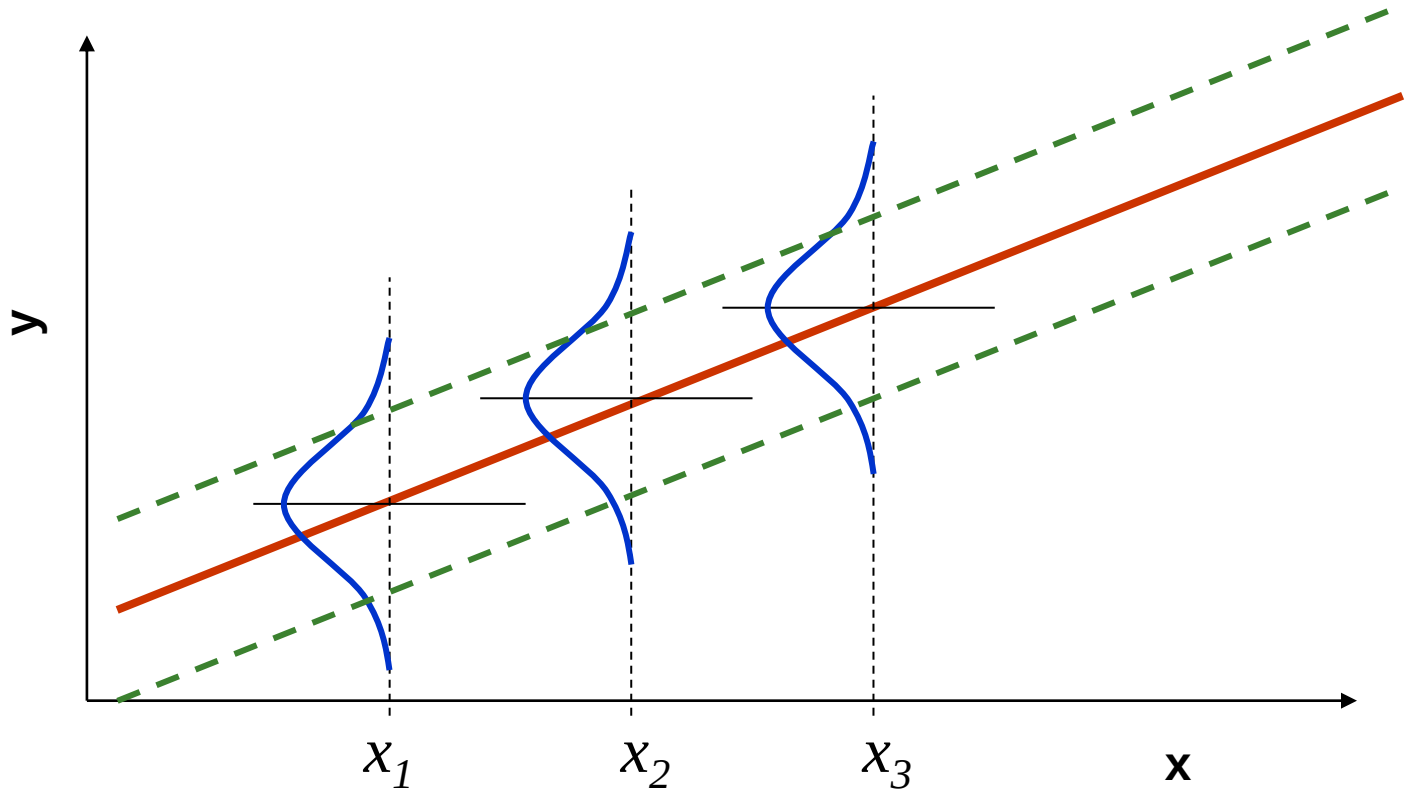


Interpretação Gráfica do Modelo Linear

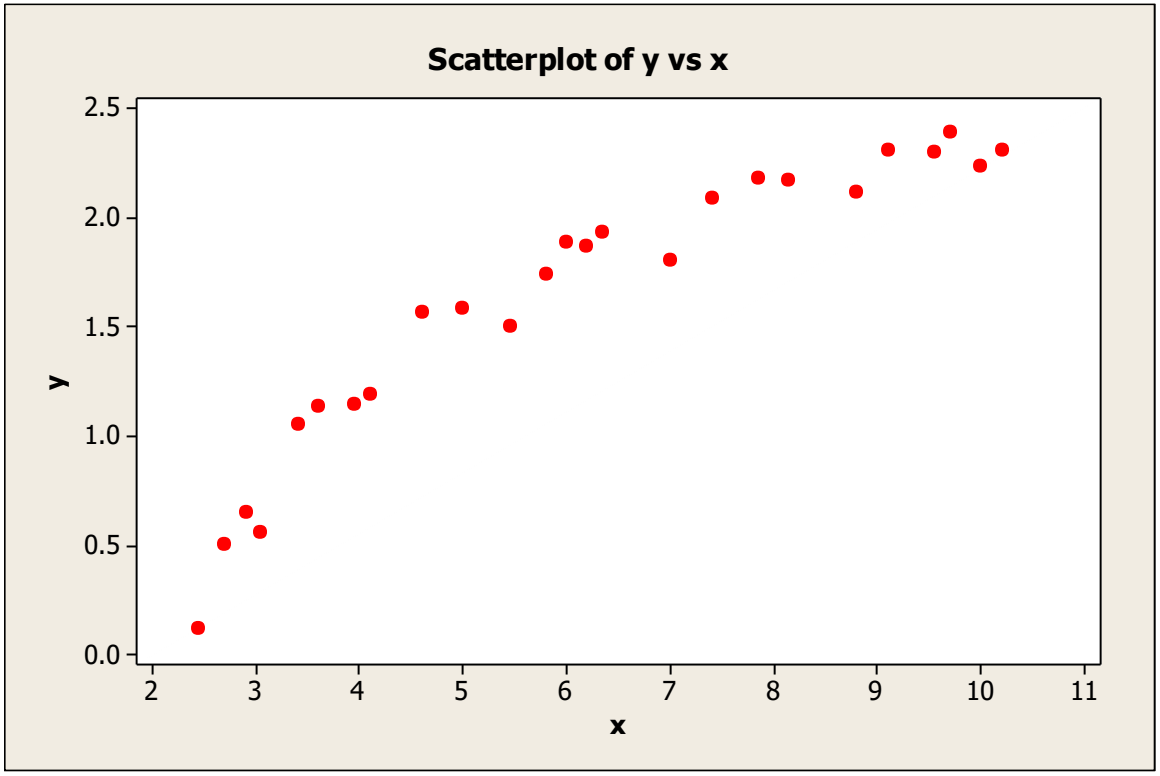


$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i$$

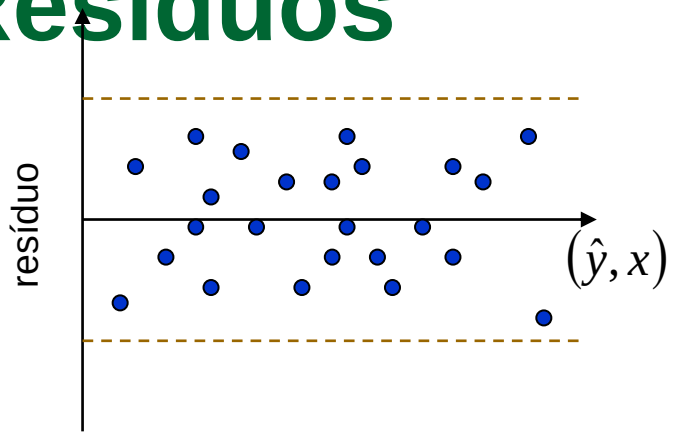
Modelo Homocedástico (variância constante)

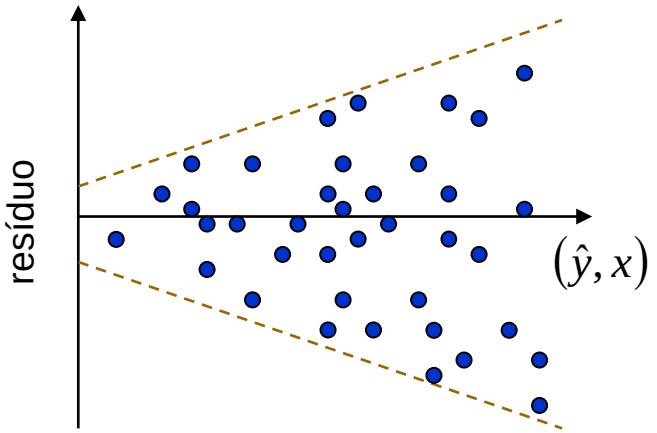


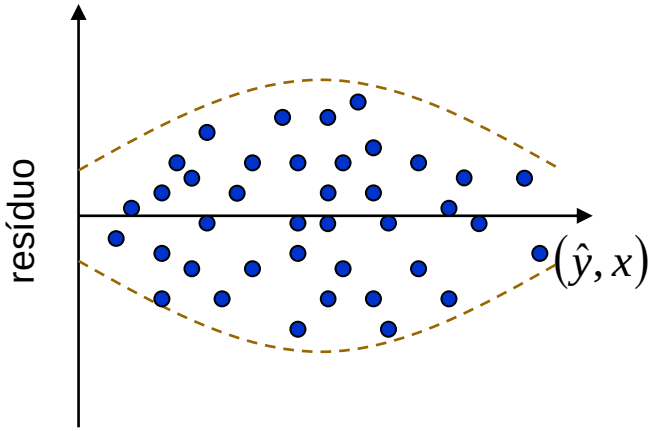
Quais as suposições necessárias para a seguinte base de dados?

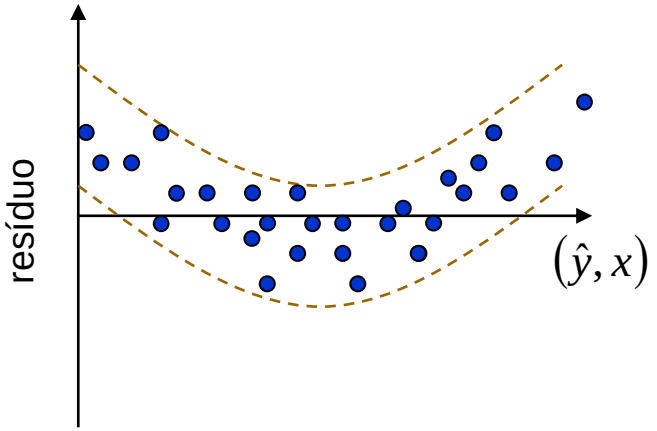


Caracterização Visual dos Resíduos

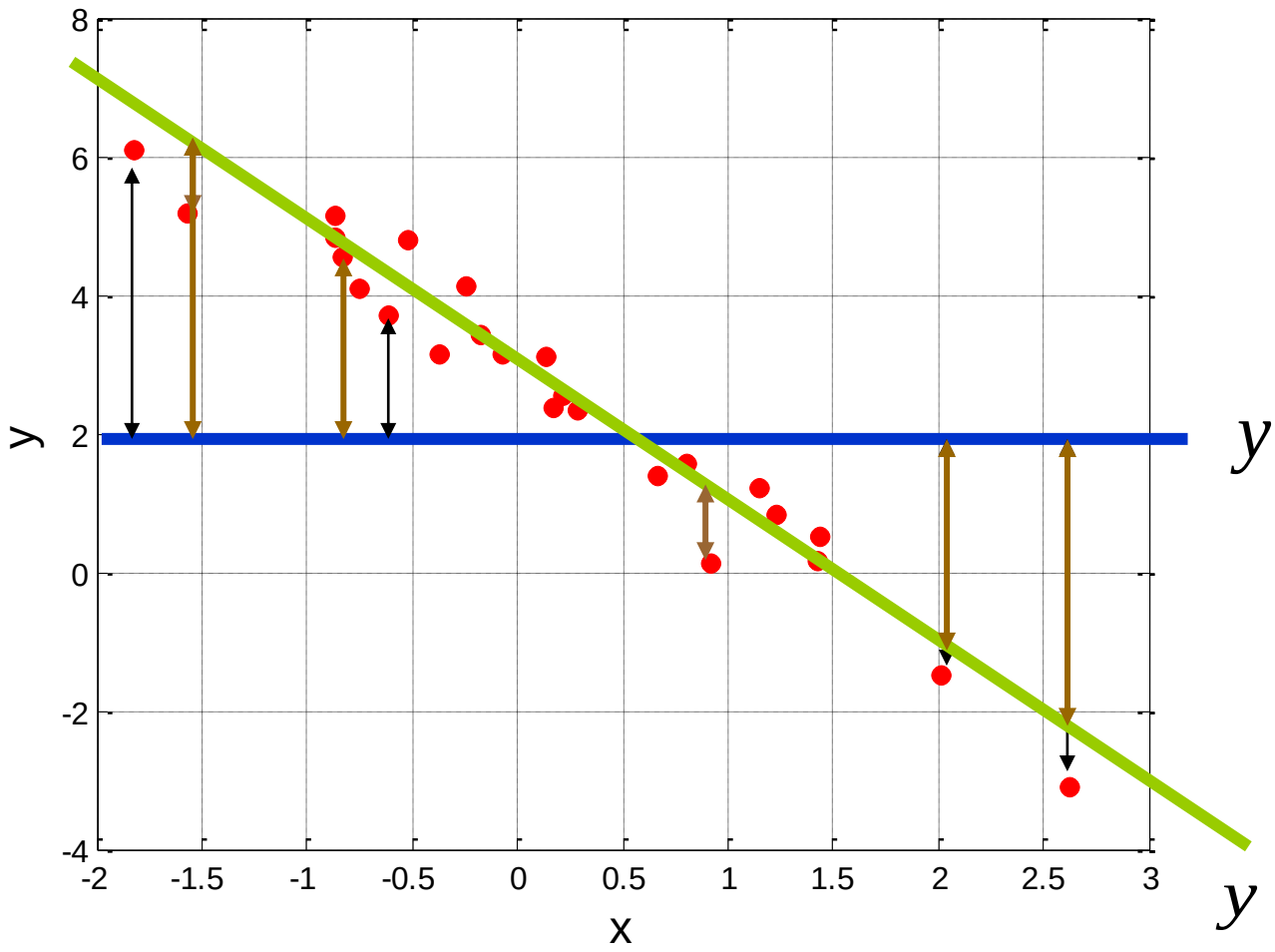








Coeficiente de Determinação (R^2)



$$y = \beta_0 = \bar{y}$$

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

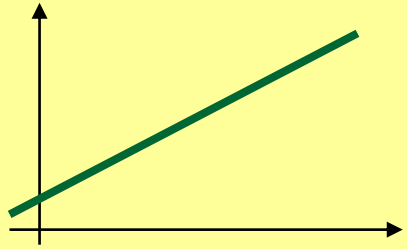
Decomposição da Soma dos Quadrados Totais

$$SQ_{Total} = SQ_{\text{Reg}} + SQ_{\text{Re}s}$$

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

SQ_{Reg}
Características assimiladas pelo Modelo

Modelo de regressão



Características não explicadas pela regressão

SQ_{Total}

Características padronizadas

SQ_{Res}

Coeficiente de Determinação (R^2)

$$R^2 = \frac{SQ_{\mathrm{Re}g}}{SQ_{Total}} = 1 - \frac{SQ_{\mathrm{Re}s}}{SQ_{Total}}$$

Pode ser interpretado como a proporção da variação explicada pelo regressor x

Propriedade:

$$0 \le SQ_{\mathrm{Re}g} \le SQ_{Total} \quad \rightarrow \quad 0 \le R^2 \le 1$$

Regressão Linear Múltipla

- Caso particular de duas variáveis

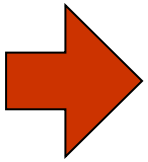
$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \varepsilon$$

- Caso geral de *p* variáveis

$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \cdots + \beta_k . x_k + \varepsilon$$

Forma Matricial

i	y	x_1	x_2
1	16,68	7	560
2	11,5	3	220
3	12,03	3	340
4	14,88	4	80
5	13,75	6	150
6	18,11	7	330
7	8	2	110
8	17,83	7	210
9	79,24	30	1460
10	21,5	5	605
11	40,33	16	688
12	21	10	215
13	13,5	4	255
14	19,75	6	462
15	24	9	448
16	29	10	776
17	15,35	6	200
18	19	7	132
19	9,5	3	36
20	35,1	17	770
21	17,9	10	140
22	52,32	26	810
23	18,75	9	450
24	19,83	8	635
25	10,75	4	150



i	y	x_0	x_1	x_2
1	16,68	1	7	560
2	11,5	1	3	220
3	12,03	1	3	340
4	14,88	1	4	80
5	13,75	1	6	150
6	18,11	1	7	330
7	8	1	2	110
8	17,83	1	7	210
9	79,24	1	30	1460
10	21,5	1	5	605
11	40,33	1	16	688
12	21	1	10	215
13	13,5	1	4	255
14	19,75	1	6	462
15	24	1	9	448
16	29	1	10	776
17	15,35	1	6	200
18	19	1	7	132
19	9,5	1	3	36
20	35,1	1	17	770
21	17,9	1	10	140
22	52,32	1	26	810
23	18,75	1	9	450
24	19,83	1	8	635
25	10,75	1	4	150

Matriz Resposta e Matriz de Regressores

y
16,68
11,5
12,03
14,88
13,75
18,11
8
17,83
79,24
21,5
40,33
21
13,5
19,75
24
29
15,35
19
9,5
35,1
17,9
52,32
18,75
19,83
10,75

$\mathbf{y}_{n\times 1}$

x_0	x_1	x_2
1	7	560
1	3	220
1	3	340
1	4	80
1	6	150
1	7	330
1	2	110
1	7	210
1	30	1460
1	5	605
1	16	688
1	10	215
1	4	255
1	6	462
1	9	448
1	10	776
1	6	200
1	7	132
1	3	36
1	17	770
1	10	140
1	26	810
1	9	450
1	8	635
1	4	150

$\mathbf{X}_{n\times p}$

Equação do modelo de regressão linear múltipla

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

Estimação dos Parâmetros

Soma dos Quadrados dos Erros

$$\mathrm{SQE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Ponto de mínimo

$$\left.\frac{\partial \mathrm{SQE}}{\partial \boldsymbol{\beta}}\right|_{\hat{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

Estimador de mínimos quadrados

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Propriedades dos Estimadores

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

$$\mathrm{var}(\hat{\beta}_j) = \sigma^2 \mathrm{C}_{jj}$$

onde C_{jj} é o elemento j da diagonal da matriz $C=(X^TX)^{-1}$.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \sim N\left(\boldsymbol{\beta}, \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\sigma^2\right)$$

Colinearidade/Multicolinearidade

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

Suponha as variáveis resposta e preditora previamente padronizadas

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

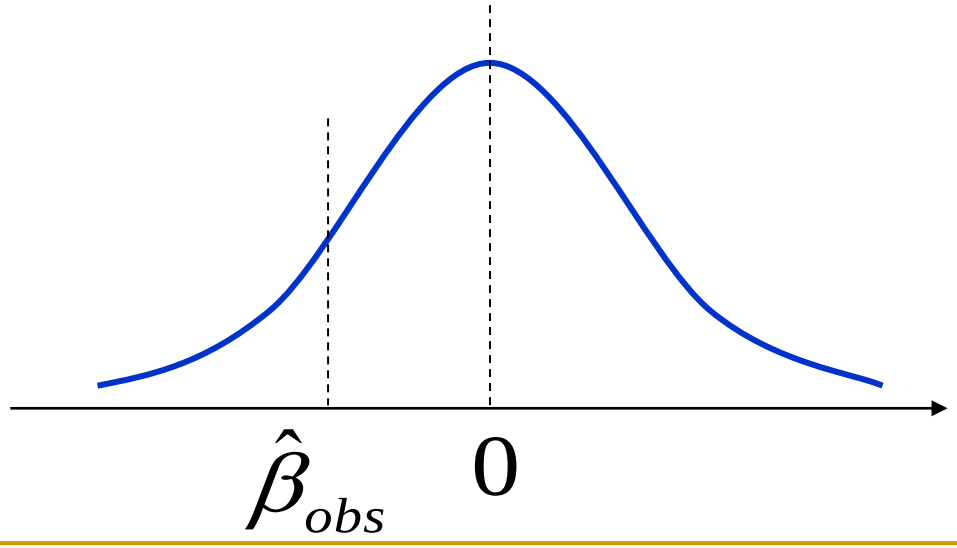
$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix}$$

No caso geral (p variáveis):

$$\mathbf{C}_{jj} = \frac{1}{1-R_j^2},\ j = 1,2,\ldots,p$$

Características da Multicolinearidade

- A multicolinearidade torna a variância de um estimador β_j muito elevada, neste caso é possível que, para amostras diferentes o mesmo estimador possa ser negativo para uma amostra e positivo para a outra amostra.



  - Na prática, a multicolinearidade causa a não rejeição da hipótese nula de um estimador (H_0: β_j=0), quando o mesmo é significativo.