Técnicas Estatísticas de Predição

Otaviano Francisco Neves

Correlação

Correlação

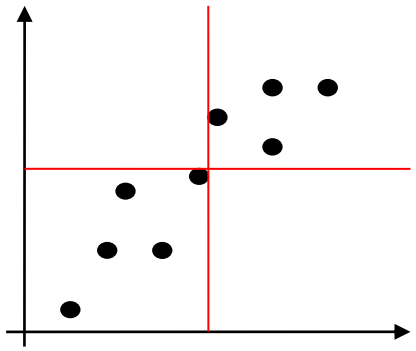
- É uma medida adimensional que está entre **-1 e 1** e mede a relação entre duas variáveis;
- **Correlação Negativa** indica relacionamento inversamente proporcional;
- **Correlação Positiva** indica relacionamento diretamente proporcional.

Correlação Amostral (*Pearson*)

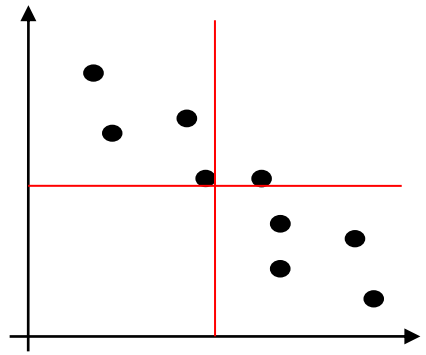
$$Corr(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{\sum (xy) - n\bar{x}\bar{y})}{\sqrt{\sum x^2 - n\bar{x}^2 \sum y^2 - n\bar{y}^2}}$$

Gráficos de dispersão

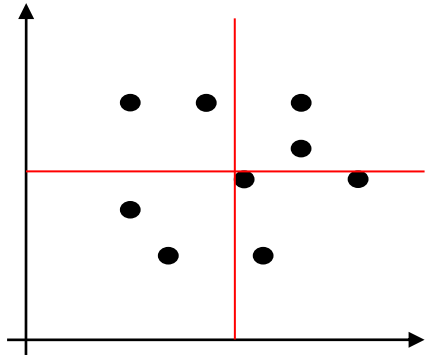
$$-1 \leq Corr(x, y) \leq 1$$



$Corr(x, y) \cong 1$

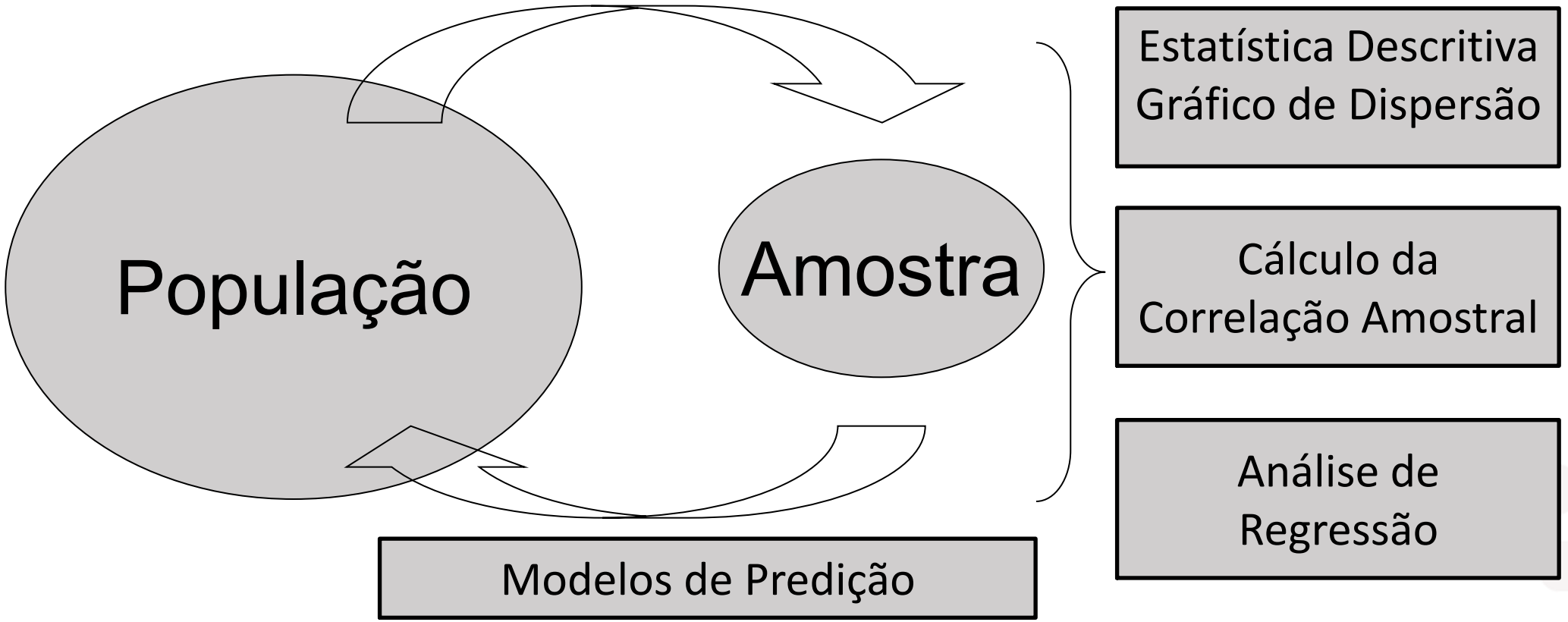


$Corr(x, y) \cong -1$



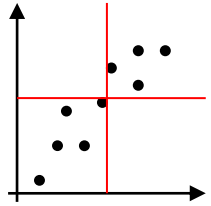
$Corr(x, y) \cong 0$

Esquema da Análise de Correlação



Esquema da Análise de Correlação

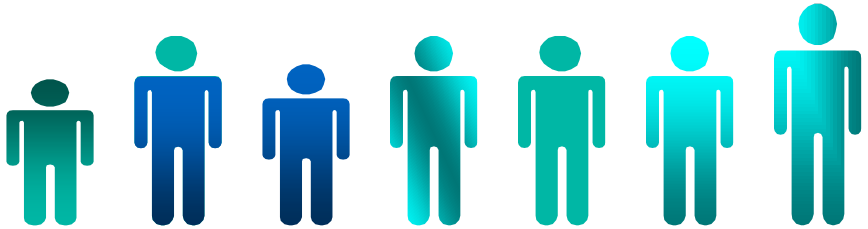
Altura vs. Peso

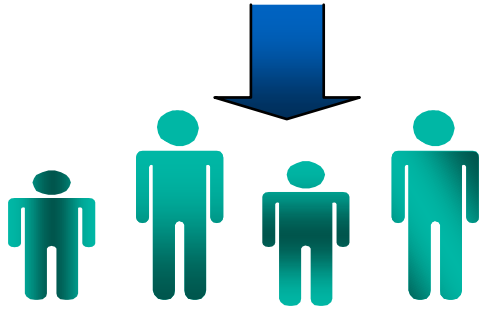


Questão:

Existe relação entre as duas variáveis?

População: N





Amostra: n

Corr. (x,y)

Qual modelo me descreve esta relação?

Exemplo

Os dados a seguir representam o tempo de serviço em anos de 10 funcionários de uma seguradora (X) e a quantidade de clientes que cada um possui (Y), verifique se existe uma associação entre as variáveis.

Dados

ID	A	B	C	D	E	F	G	H	I	J
X	2	3	4	5	4	6	7	8	8	10
Y	48	50	56	52	43	60	62	58	64	72

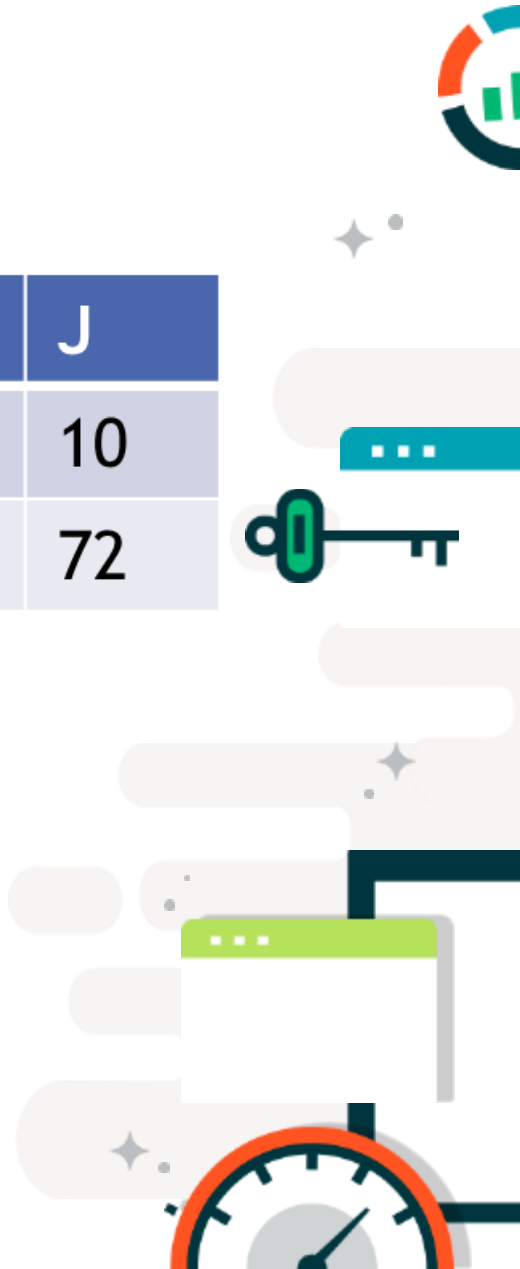
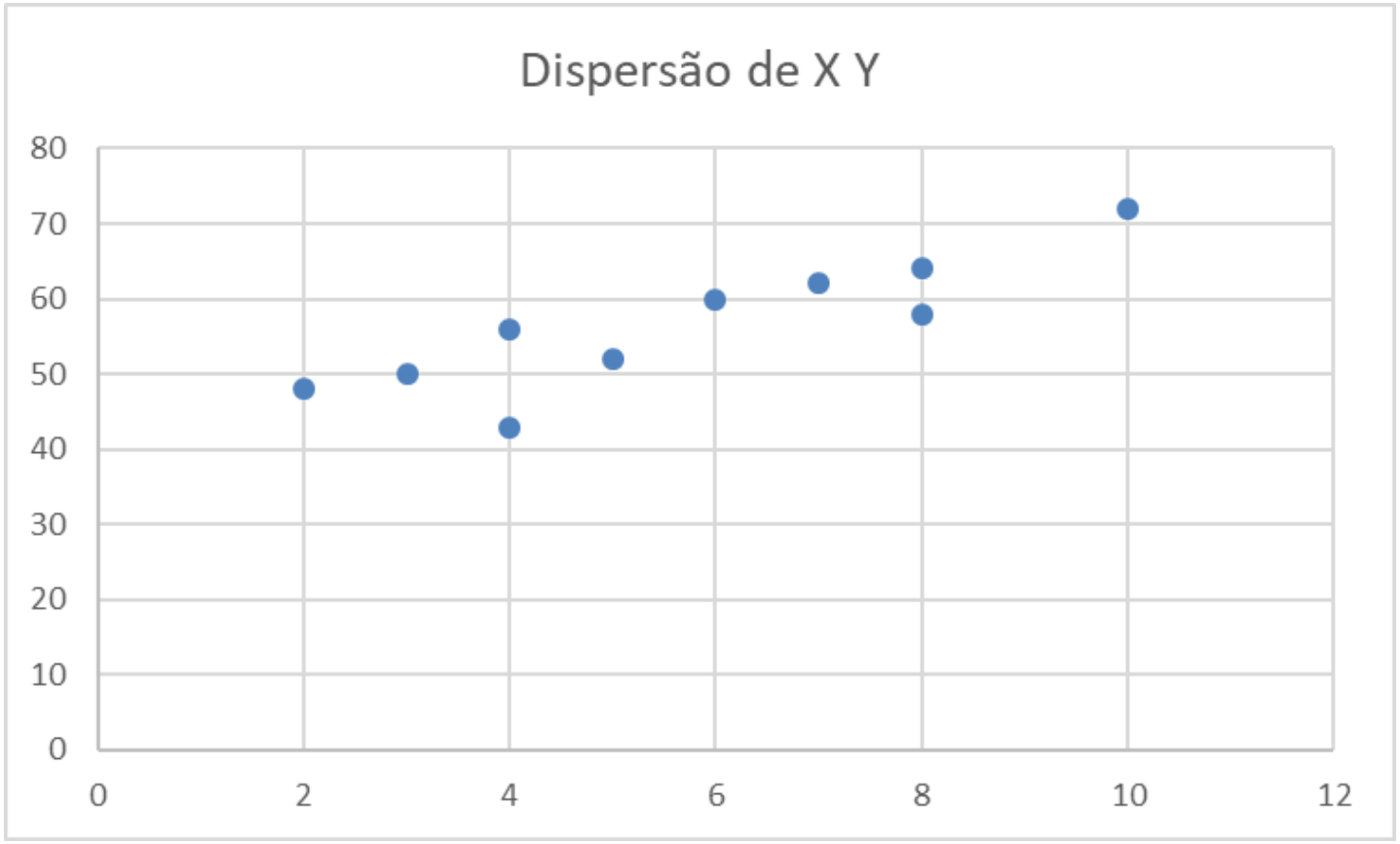


Gráfico de dispersão



Cálculo

x	y	x2	y2	x.y
2	48	4	2304	96
3	50	9	2500	150
4	56	16	3136	224
5	52	25	2704	260
4	43	16	1849	172
6	60	36	3600	360
7	62	49	3844	434
8	58	64	3364	464
8	64	64	4096	512
10	72	100	5184	720
57	**565**	**383**	**32581**	**3392**

Correlação Amostral

$$Corr(x,y) = \frac{\sum(xy) - n\bar{x}\bar{y})}{\sqrt{\sum x^2 - n\bar{x}^2 \sum y^2 - n\bar{y}^2}} =$$

$$\frac{3392 - 10.5{,}9.56.5}{\sqrt{386 - 10.(5{,}9)^2 \times 32581 - 10.(56{,}5)^2}} =$$

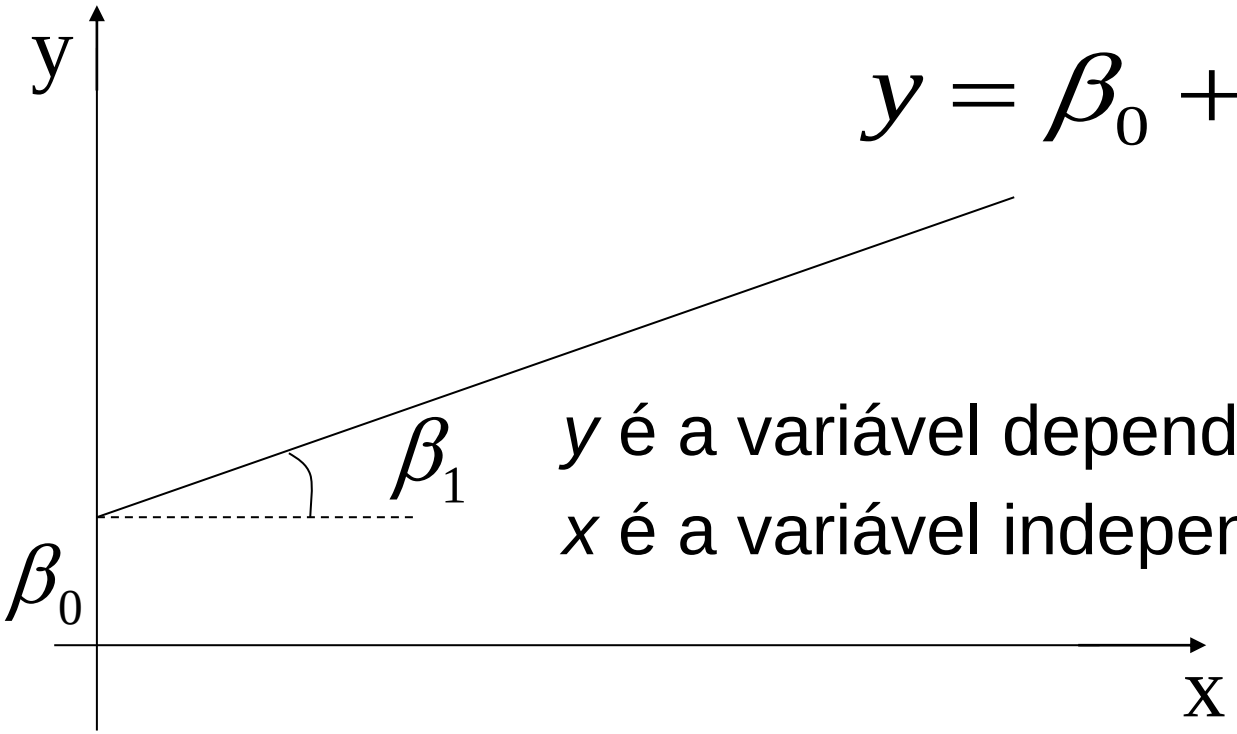
$$= \frac{171{,}5}{\sqrt{58{,}1 \times 658{,}5}} = \mathbf{0,8768}$$

Técnicas Estatísticas de Predição

Otaviano Francisco Neves

Regressão Linear Simples

Modelo Teórico

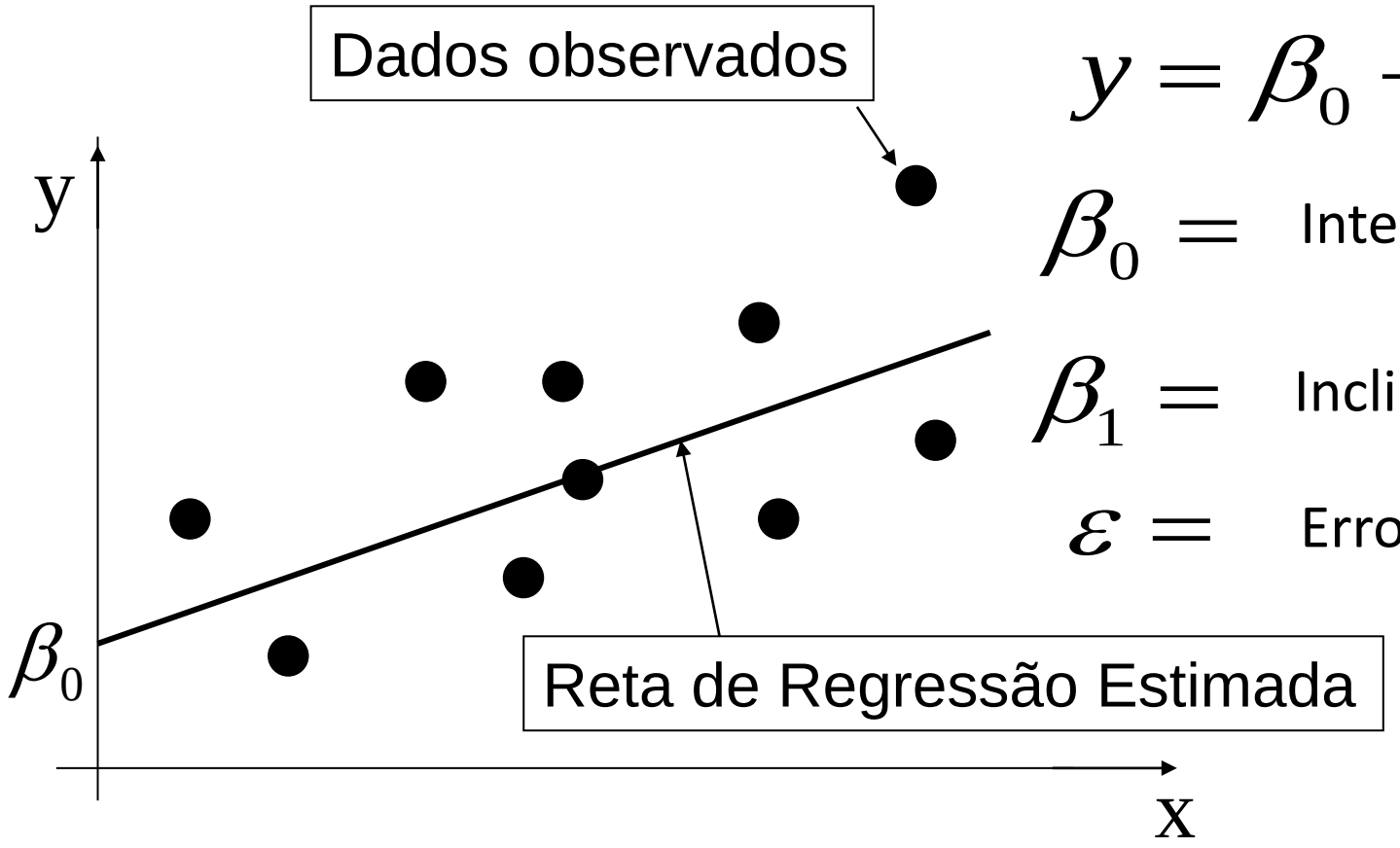


$$y = \beta_0 + \beta_1 x + \varepsilon$$

y é a variável dependente ou resposta.

x é a variável independente ou explicativa.

Modelo de Regressão Linear Simples - Ajuste



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

$\beta_0 =$ Intercepto

$\beta_1 =$ Inclinação

$\varepsilon =$ Erro Aleatório

Estimação : Regressão Linear Simples

$$S(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

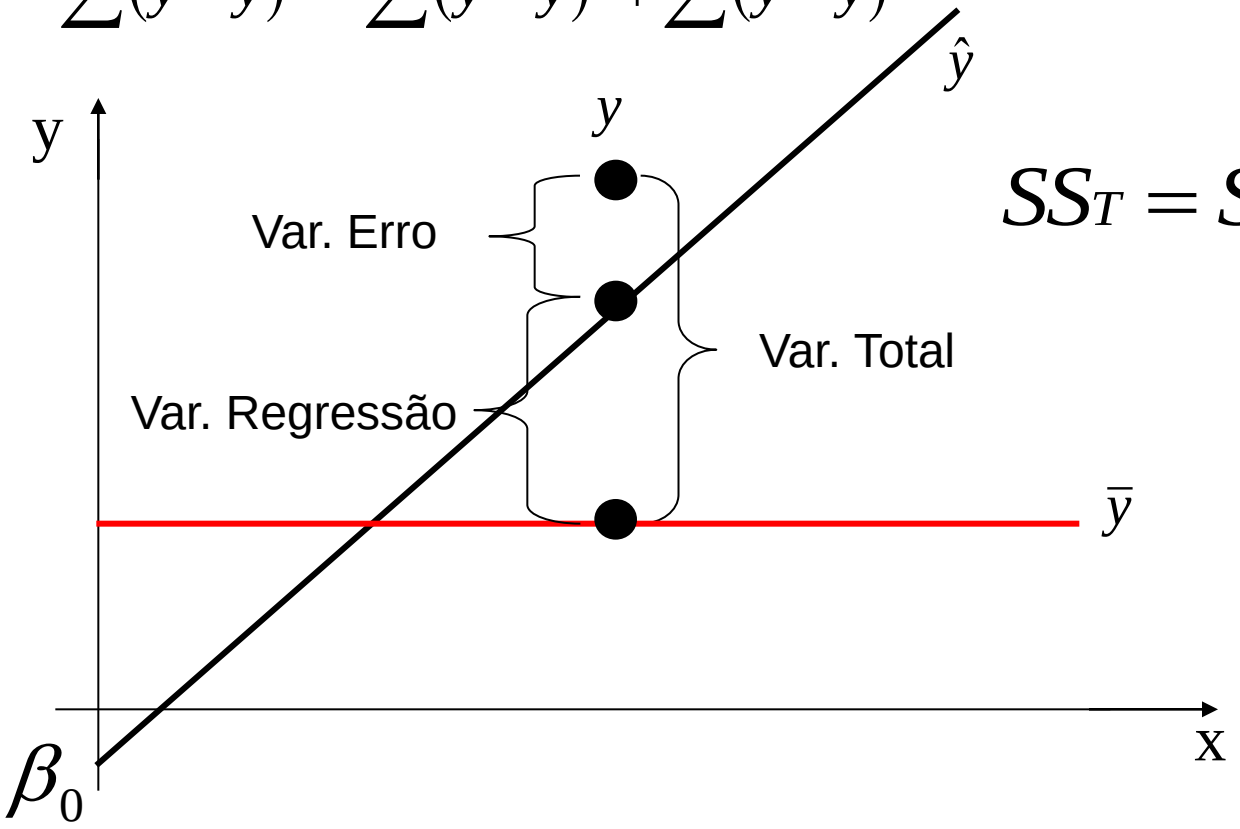
$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Partições da variabilidade

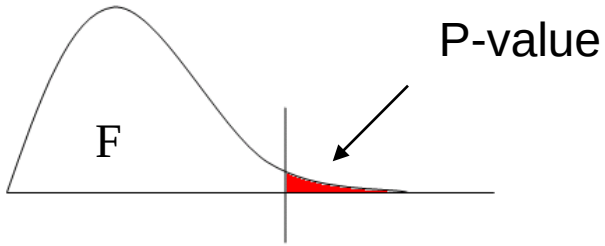
$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$



$$SS_T = SS_R + SS_E$$

Tabela Anova

Variação	Soma de quadrado	Graus de Liberdade	Erro Médio	F
Regressão	SS_R	1	$MS_R=SS_R$	MS_R/MS_E
Residual (error)	SS_E	n-2	$MS_E=SS_E/(n-2)$	
Total	SS_T	n-1		



Exemplo

Dados o tempo de serviço em anos de 10 funcionários de uma seguradora (X) e a quantidade de clientes que cada um possui (Y), verifique se existe uma associação entre as variáveis.

ID	A	B	C	D	E	F	G	H	I	J
X	2	3	4	5	4	6	7	8	8	10
Y	48	50	56	52	43	60	62	58	64	72

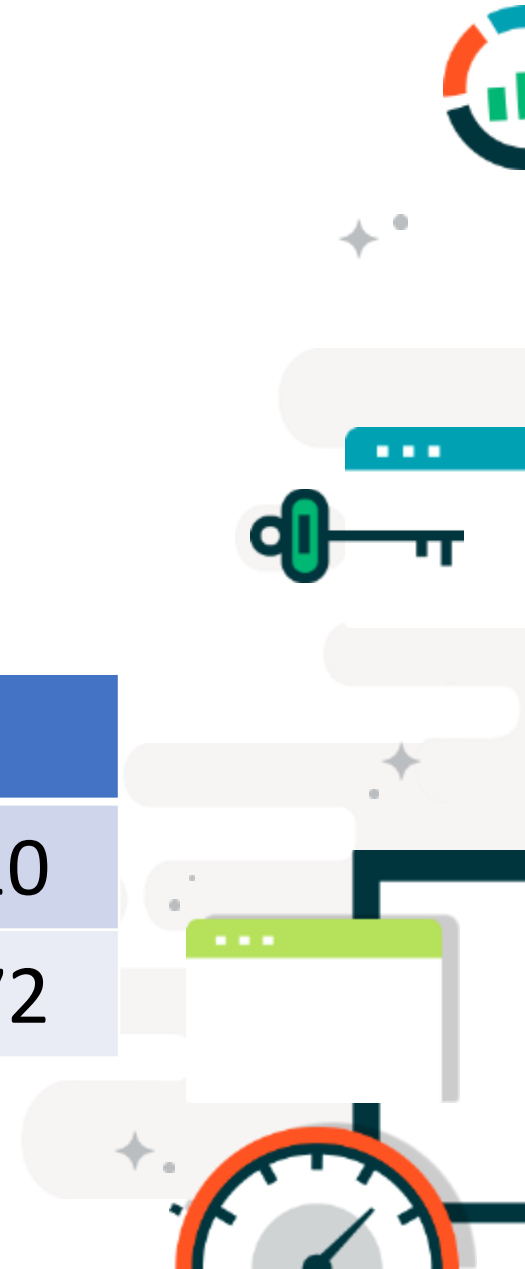
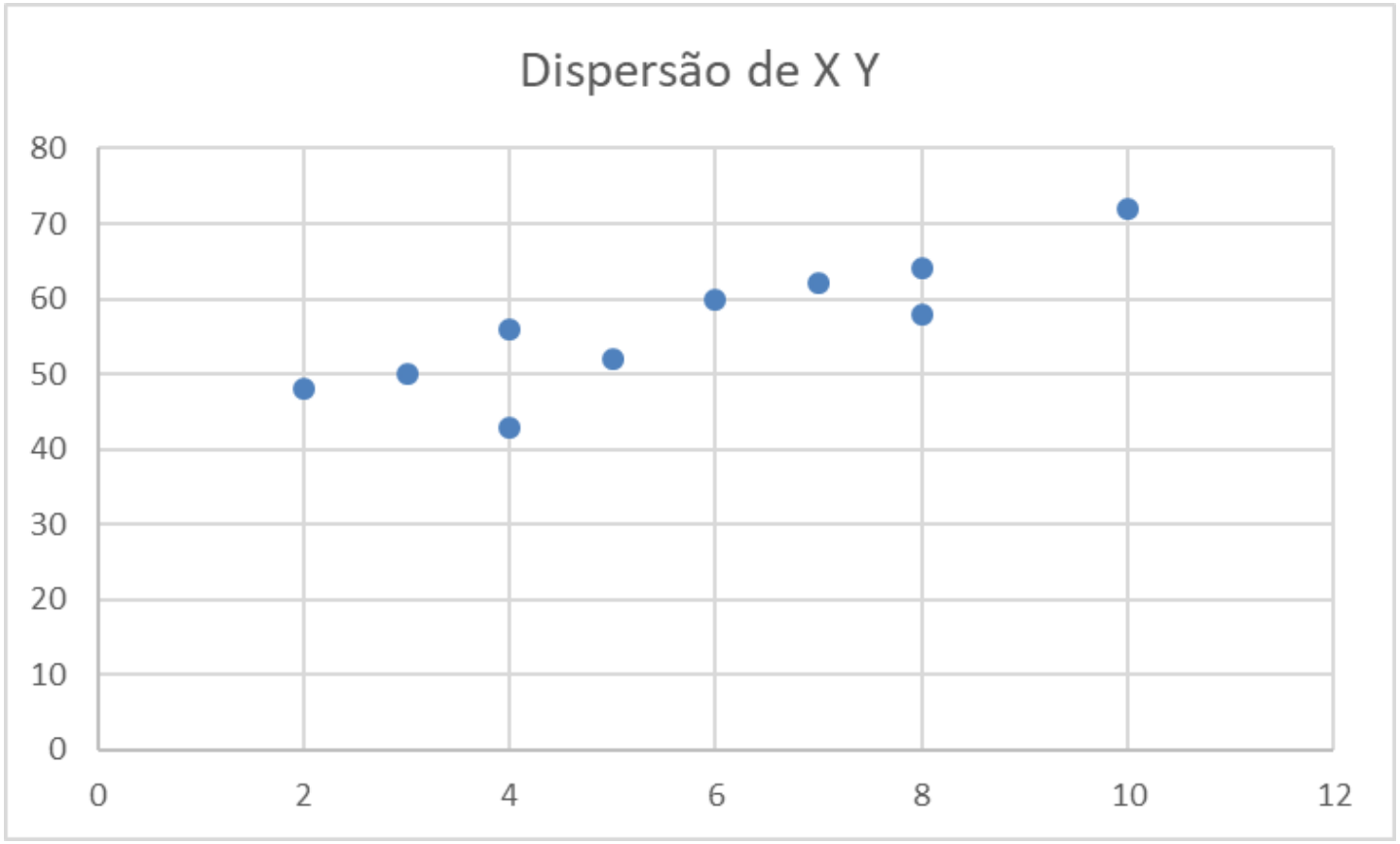


Gráfico de dispersão



Equação da Reta de Regressão

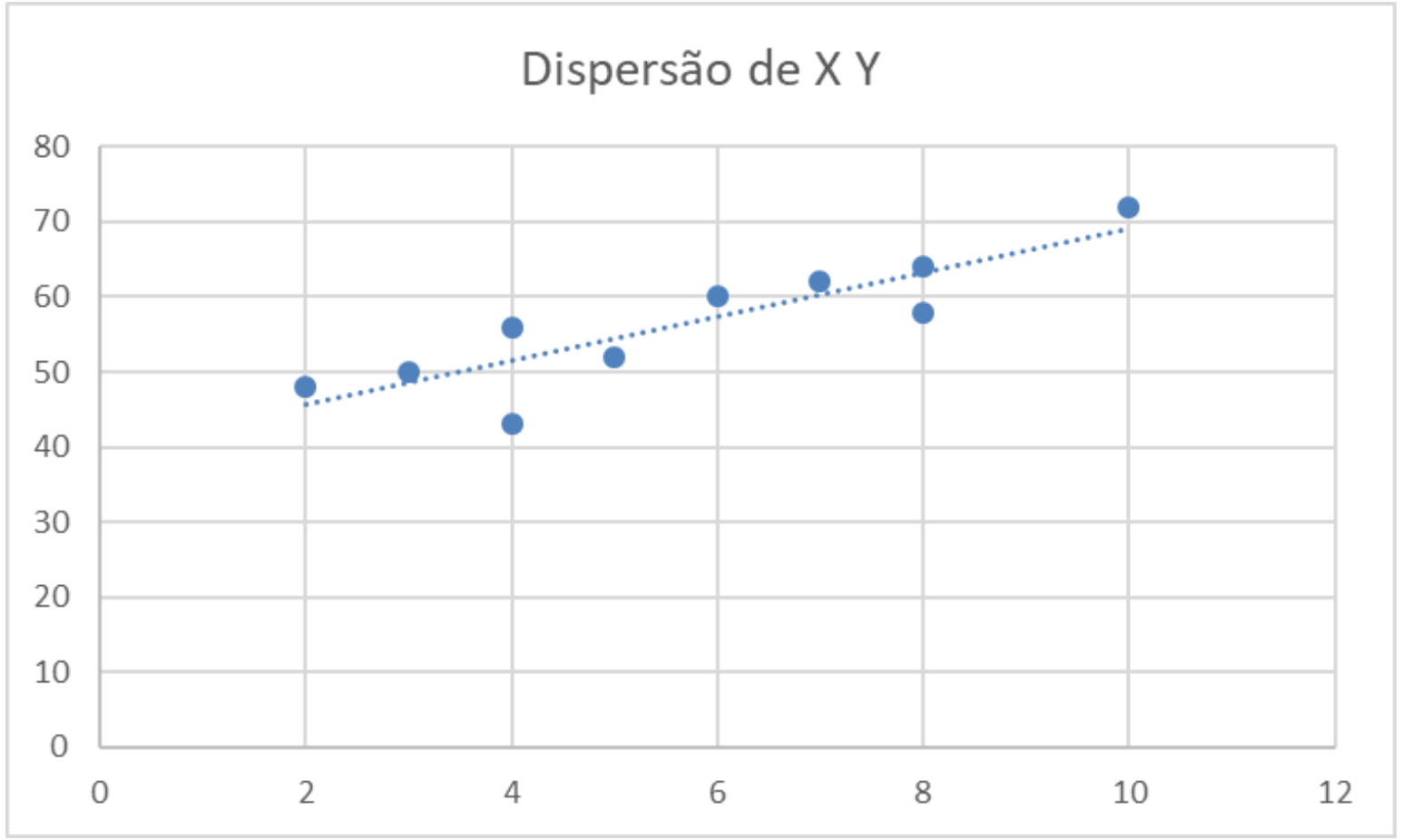


Tabela Anova

Variação	Soma de quadrado	Graus de Liberdade	Erro Médio	F
Regressão	SS_R	1	$MS_R=SS_R$	MS_R/MS_E
Residual (error)	SS_E	n-2	$MS_E=SS_E/(n-2)$	
Total	SS_T	n-1		

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	506,23	506,235	26,60	0,001
X	1	506,23	506,235	26,60	0,001
Error	8	152,27	19,033		
Lack-of-Fit	6	49,77	8,294	0,16	0,965
Pure Error	2	102,50	51,250		
Total	9	658,50			

Estimação dos Parâmetros

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	39,67	3,54	11,20	0,000
X	2,952	0,572	5,16	0,001

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

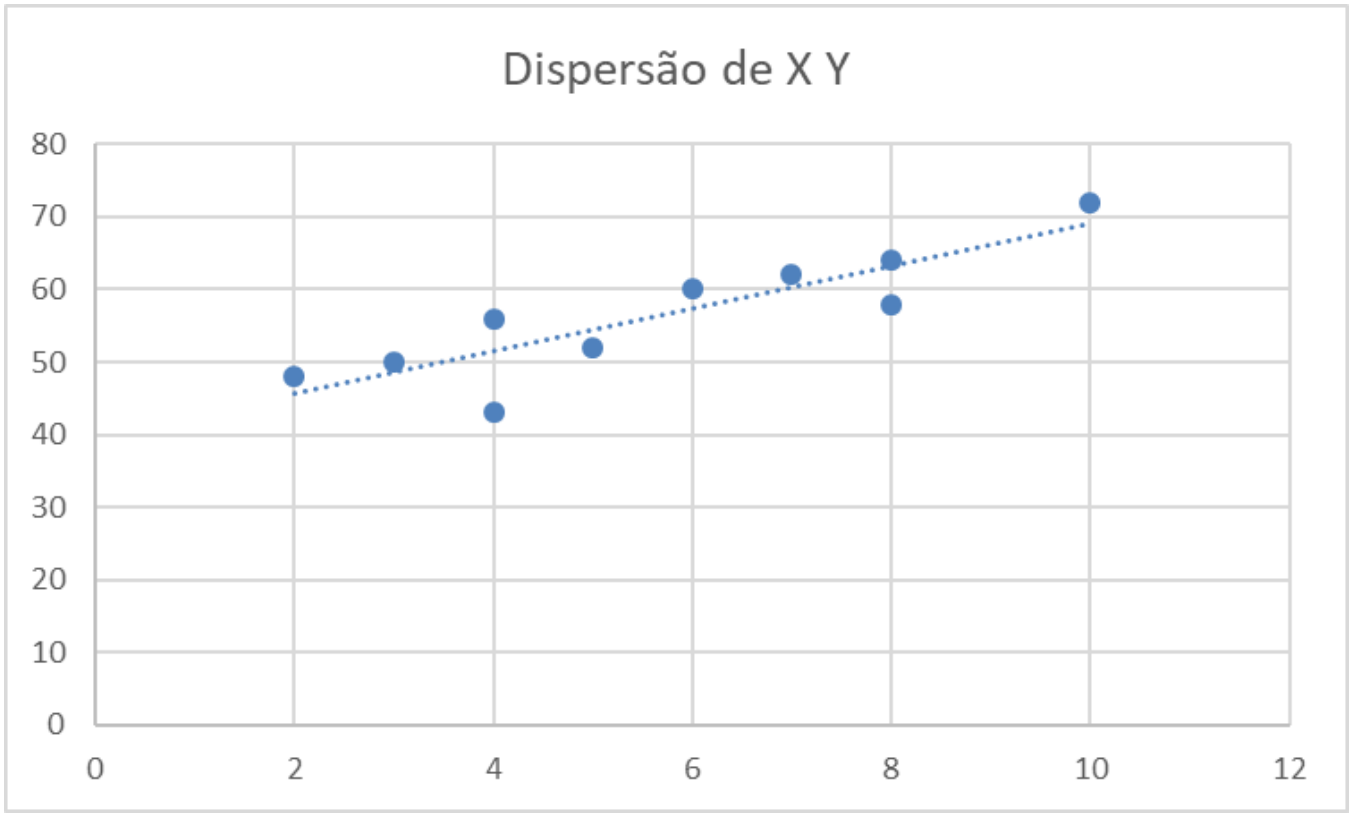
$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$\hat{\beta}_1 = \frac{10 * 3392 - 57 * 565}{10 * 383 - 57^2} = 2{,}9518$$

$$\hat{\beta}_0 = 56{,}5 - 2{,}9518 * 5{,}7 = 39{,}6747$$

Modelo de Regressão - Ajuste

$$\hat{y} = 39{,}67 + 2{,}95x + \varepsilon$$



Modelo de Predição

$$\hat{y} = 39{,}67 + 2{,}95 * 8 = 63{,}286 \cong 63\, clientes$$

Técnicas Estatísticas de Predição

Otaviano Francisco Neves

Regressão Linear Múltipla

Modelo de Regressão Linear Múltipla

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

Matriz de Regressão

$$y = X\beta + \varepsilon$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Estimadores de Mínimos quadrados

$$\hat{\beta} = \left(X' X\right)^{-1} X' y$$

$$\hat{y} = X\hat{\beta}$$

$$e = y - \hat{y}$$

Exemplo

Esses dados representam a resistência à tração (y) de uma ligação de fio em um processo de fabricação de semicondutores, comprimento de fio (x1) e altura da matriz (x2) para ilustrar a construção de um modelo empírico.

Dados

ID	y	x1	x2
1	9,95	2	50
2	24,45	8	110
3	31,75	11	120
.	.	.	.
.	.	.	.
.	.	.	.
24	22,13	6	100
25	21,15	5	400

Tabela ANOVA

Source of variation	Sum of Square	Degrees of Freedom	Mean Square	F
Regression	SS_R	k	$MS_R=SS_R/k$	MS_R/MS_E
Residual (error)	SS_E	n-(k+1)	$MS_E=SS_E/(n-(k+1))$	
Totals	SS_T	n-1		

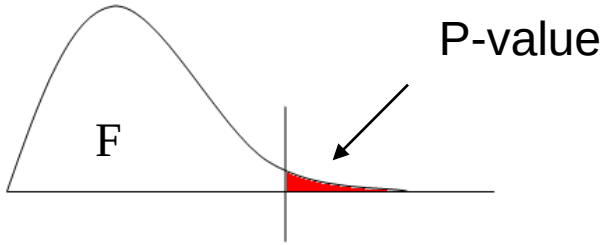
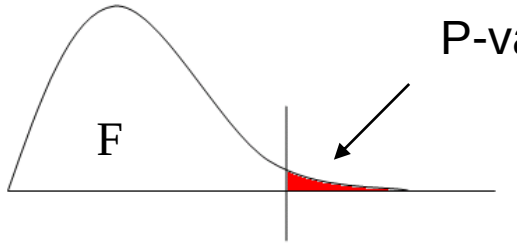


Tabela ANOVA

Source of variation	Sum of Square	Degrees of Freedom	Mean Square	F
Regression	2	5990,772	2995,39	572,1672
Residual (error)	22	115,1735	5,24	
Totals	24	6105,9447		

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5990,8	2995,39	572,17	0,000
x1	1	4507,5	4507,53	861,01	0,000
x2	1	104,9	104,92	20,04	0,000
Error	22	115,2	5,24		
Total	24	6105,9			



P-value<0,0001

Estimação dos Parâmetros

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	2,26	1,06	2,14	0,044
x1	2,7443	0,0935	29,34	0,000
x2	0,01253	0,00280	4,48	0,000

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t_0 = \frac{\hat{\beta}_1}{Se(\hat{\beta}_1)} = \frac{2,74}{0,0935} = 29,34$$

P-value < 0,0001

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

$$t_0 = \frac{\hat{\beta}_2}{Se(\hat{\beta}_2)} = \frac{0,012528}{0,002798} = 4,48$$

P-value < 0,0001

Modelo de Regressão Linear Múltipla

$$\hat{y} = 2{,}26 + 2{,}744x_1 + 0{,}013x_2 + \varepsilon$$

Técnicas Estatísticas de Predição

Otaviano Francisco Neves

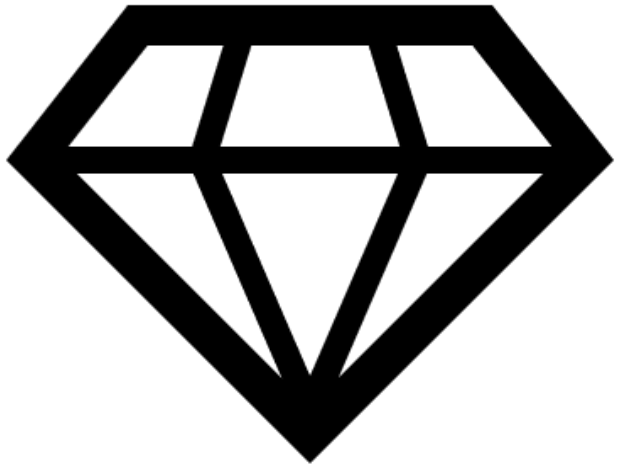
Regressão Linear Múltipla - Exemplo

Dados airbnb - IA

Data Sources

https://www.kaggle.com

- ✓ID da hospedagem
- ✓Nome da hospedagem
- ✓ID do Hóspede
- ✓Nome do Hóspede
- ✓Grupo de vizinhança
- ✓Latitude
- ✓Longitude
- ✓Tipo de hospedagem
- ✓Preço
- ✓Locação mínima
- ✓Número de avaliações
- ✓Taxa mensal de avaliações
- ✓Número máximo de hóspedes
- ✓Disponibilidade anual

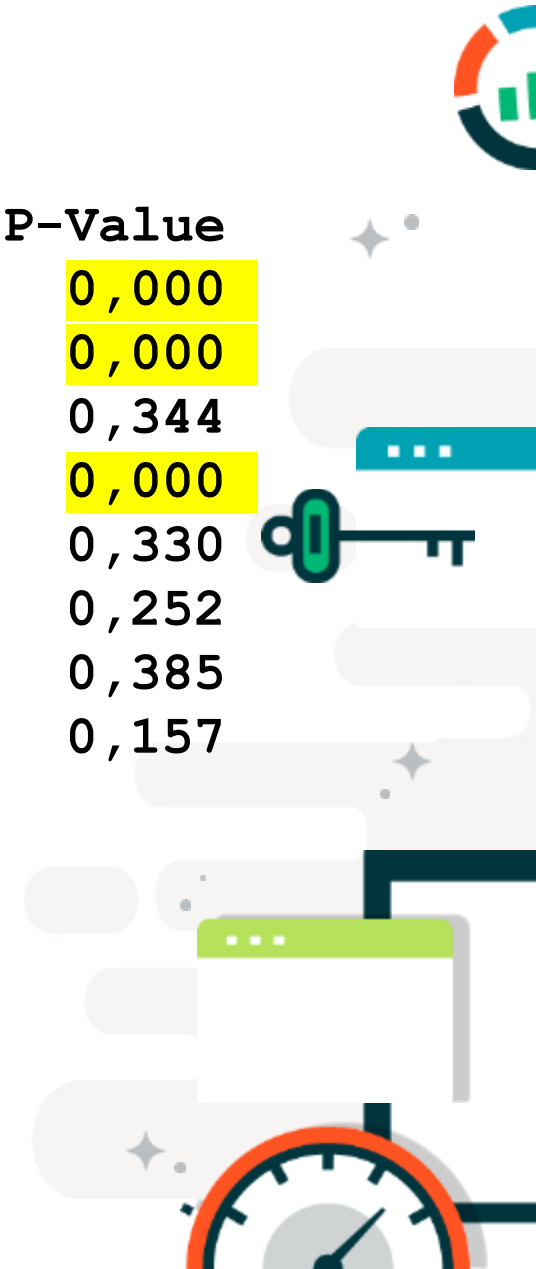


Objetivo

Modelar a taxa mensal de avaliações (ocupações)

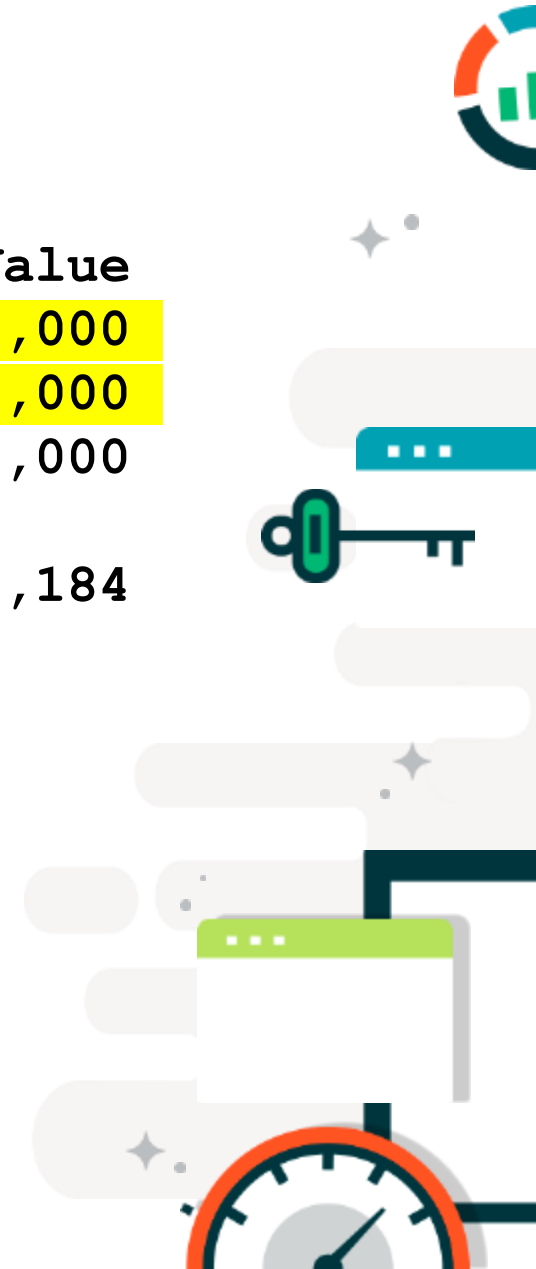
Análise de Variância

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	10,3820	1,48315	494,95	0,000
Preço	1	0,0425	0,04253	14,19	0,000
locação mínima	1	0,0027	0,00270	0,90	0,344
Número de avaliações	1	0,3653	0,36528	121,90	0,000
Número máximo de hóspedes	1	0,0029	0,00286	0,95	0,330
Disponibilidade anual	1	0,0040	0,00395	1,32	0,252
Grupo de vizinhança	1	0,0023	0,00227	0,76	0,385
Tipo de hospedagem	1	0,0060	0,00603	2,01	0,157
Error	203	0,6083	0,00300		
Total	210	10,9903			



Novo Modelo – ANOVA

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	10,3696	5,18481	1737,44	0,000
Preço	1	0,0480	0,04796	16,07	0,000
Número de avaliações	1	0,3563	0,35626	119,38	0,000
Error	208	0,6207	0,00298		
Lack-of-Fit	100	0,3258	0,00326	1,19	0,184
Pure Error	108	0,2949	0,00273		
Total	210	10,9903			

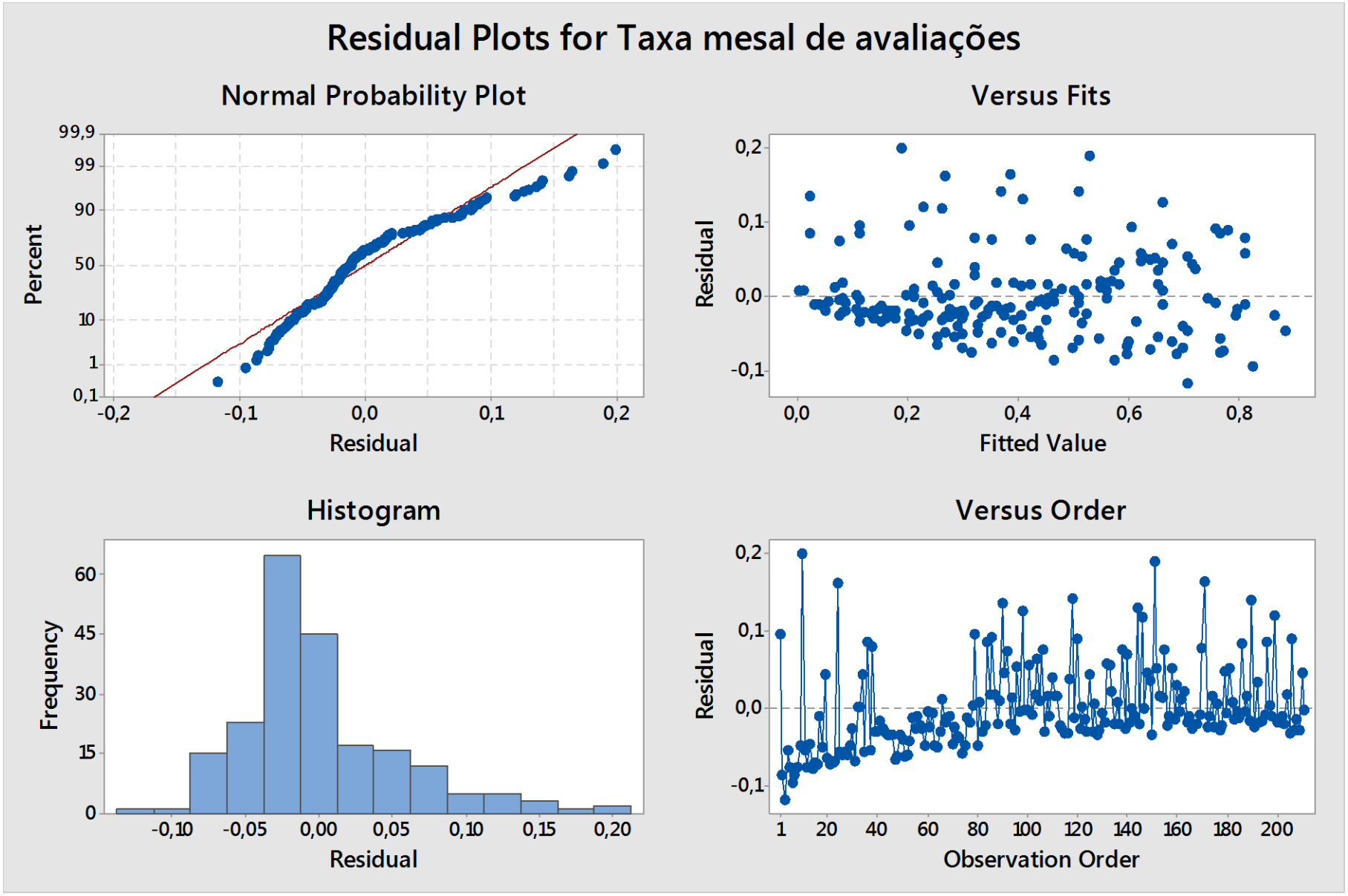


Qualidade de Ajuste

S = 0,0546275 - Desvio padrão do Erro

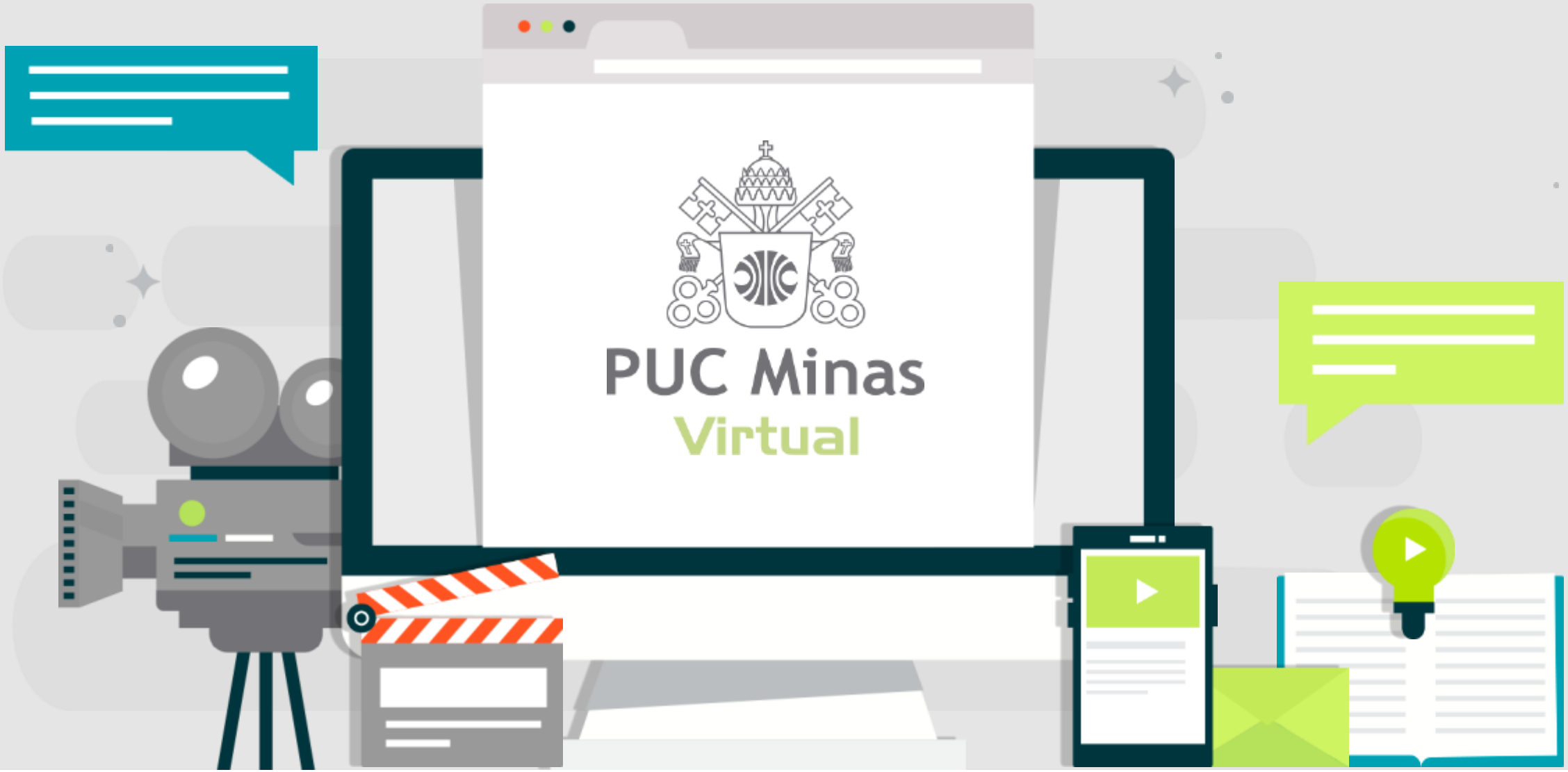
R^2 = SSR/SST = 94,35% - Coeficiente de determinação

Análise de Resíduo



Modelo

Taxa mesal de avaliações = 0,3543 - 0,002405 Preço + 0,006712 Número de avaliações


PUC Minas
Virtual