

CORRELAÇÃO E REGRESSÃO (PLT: Pág. 333)

Introdução Em muitos dos fenômenos, duas variáveis podem estar relacionadas. Por exemplo, o volume de vendas de uma empresa podem sofrer ações de fatores como investimento em propaganda, quantidade de concorrentes, atendimento, etc.

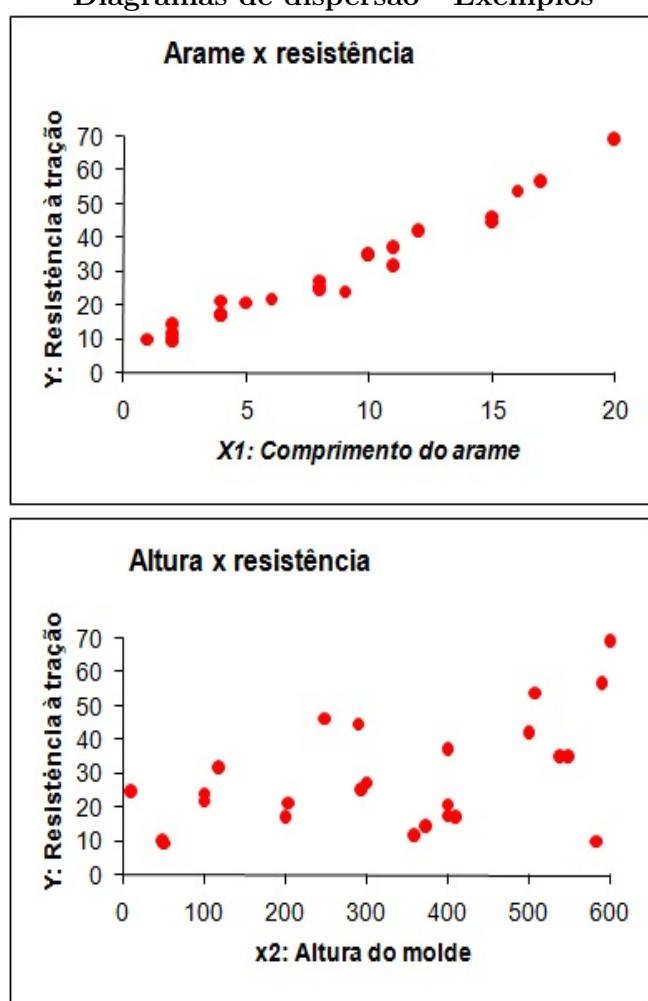
Tendo em mãos os valores observados de duas variáveis (X e Y) pode-se construir um modelo matemático da forma $y = f(x)$ que descreva o fenômeno em estudo.

Diagrama de Dispersão O gráfico construído tomando-se os valores de X com seus respectivos valores Y é chamado diagrama de dispersão. Em muitos casos, a nuvem de pontos do gráfico de dispersão delineia uma reta, sugerindo que o fenômeno pode ser descrito, ao menos aproximadamente, por uma função linear. Por meio da expressão desta função, o fenômeno pode ser analisado, auxiliando na tomada de decisões.

Exemplo 1 Essa tabela contém dados de três variáveis que foram coletados em uma planta de fabricação de semicondutores. Nessa planta, o semicondutor final é um arame colado a uma estrutura. As variáveis reportadas são a resistência (X_1), a tração (uma medida da quantidade de força requerida para romper a cola) (X_2), o comprimento do arame e a altura da matriz.

Diagramas de dispersão - Exemplos

	X_1	X_2	Y
1	2	50	9,95
2	8	11	24,45
3	11	120	31,75
4	10	550	35
5	8	295	25,02
6	4	200	16,86
7	2	375	14,38
8	2	52	9,6
9	9	100	24,35
10	8	300	27,5
11	4	412	17,08
12	11	400	37
13	12	500	41,95
14	2	360	11,66
15	4	205	21,65
16	4	400	17,89
17	20	600	69
18	1	585	10,3
19	10	540	34,93
20	15	250	46,59
21	15	290	44,88
22	16	510	54,12
23	17	590	56,63
24	6	100	22,13
25	5	400	21,15



(Gráficos elaborados utilizando Excel)

Os pontos destes diagramas de dispersão delineiam uma reta descendente; isto é, sugerem uma função linear crescente que descreve uma relação do tipo “quanto maior o x , maior o respectivo y ”. Dizemos, neste caso, que as variáveis são positivamente correlacionadas.

“quanto maior o x , menor o respectivo y ” → negativamente correlacionadas
 “quanto maior o x , maior o respectivo y ” → positivamente correlacionadas

OBS: Há casos em que as variáveis não se correlacionam linearmente: a correlação pode ser curvilínea.

Coeficiente de correlação linear de Pearson (r) mede a “intensidade” que relaciona as duas variáveis.

$$\frac{n \sum (x \cdot y) - \sum x \cdot \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2] \times [n \sum y^2 - (\sum y)^2]}} \quad (-1 \leq r \leq 1)$$

$\left\{ \begin{array}{l} r < 0: \text{negativamente correlacionadas} \\ r > 0: \text{positivamente correlacionadas} \end{array} \right.$

$\left\{ \begin{array}{l} r = \pm 1: \text{todos os pontos estão sobre uma reta} \\ r = 0: \text{não existe nenhuma correlação linear entre as variáveis} \end{array} \right.$

Teste de significância da correlação Este teste é utilizado para verificar se a correlação entre as variáveis é significativa ou não. O teste pode ser efetuado a vários graus de significância. Por exemplo, ao adotarmos $\alpha = 5\%$, que é padrão, a conclusão obtida tem 95% de chance de estar correta.

Critério

Se $|r| > (\text{valor crítico}) \Rightarrow \text{Correlação é significativa}$

A tabela apresentada no final deste material fornece os chamados *valores críticos* para o teste.

Observação: Mesmo que as variáveis estejam fortemente correlacionadas, não podemos dizer que uma variável é a causa de outra.

Exemplo 2 Voltando aos dados do exemplo anterior, vamos estudar a correlação entre as variáveis X_1 e Y .

X_1	Y	$(X_1)^2$	Y^2	$X_1 \cdot Y$
2	9,95	4	99,00	19,9
8	24,45	64	597,80	195,6
11	31,75	121	1008,06	349,25
10	35	100	1225,00	350
8	25,02	64	626,00	200,16
4	16,86	16	284,26	67,44
2	14,38	4	206,78	28,76
2	9,6	4	92,16	19,2
9	24,35	81	592,92	219,15
8	27,5	64	756,25	220
4	17,08	16	291,73	68,32
11	37	121	1369,00	407
12	41,95	144	1759,80	503,4
2	11,66	4	135,96	23,32
4	21,65	16	468,72	86,6
4	17,89	16	320,05	71,56
20	69	400	4761,00	1380
1	10,3	1	106,09	10,3
10	34,93	100	1220,10	349,3
15	46,59	225	2170,63	698,85
15	44,88	225	2014,21	673,2
16	54,12	256	2928,97	865,92
17	56,63	289	3206,96	962,71
6	22,13	36	489,74	132,78
5	21,15	25	447,32	105,75
206	725,82	2396	27178,53	8008,47

$$r = \frac{25 \times 8008,47 - 206 \times 725,82}{\sqrt{[25 \times 2396 - (206)^2] \times [25 \times 27178,53 - (725,82)^2]}} = 0,982$$

Para $n = 25$ e $\alpha = 5\%$, o valor crítico é 0,396 e $|r| = 0,982 > 0,396$. Logo, podemos concluir que a correlação é significativa.

Reta de regressão

Equação da reta de regressão (ou equação dos mínimos quadrados) é uma função matemática que tenta descrever melhor a nuvem de pontos do diagrama de dispersão como uma reta. Ela é uma equação da forma $y = ax + b$, onde a e b são calculadas pelas fórmulas:

$$a = \frac{n \sum (x \cdot y) - \sum x \cdot \sum y}{n \sum x^2 - (\sum x)^2} \quad \text{e} \quad b = \bar{y} - a\bar{x}, \quad \text{onde: } \begin{cases} n = \text{número de pontos observados} \\ \bar{x} = \frac{\sum x}{n} \text{ e } \bar{y} = \frac{\sum y}{n} \text{ (médias aritméticas)} \end{cases}$$

Este método de encontrar a equação da reta é denominado critério dos mínimos quadrados. Por meio desta equação podemos estimar o valor de uma variável que corresponde ao valor da outra variável.

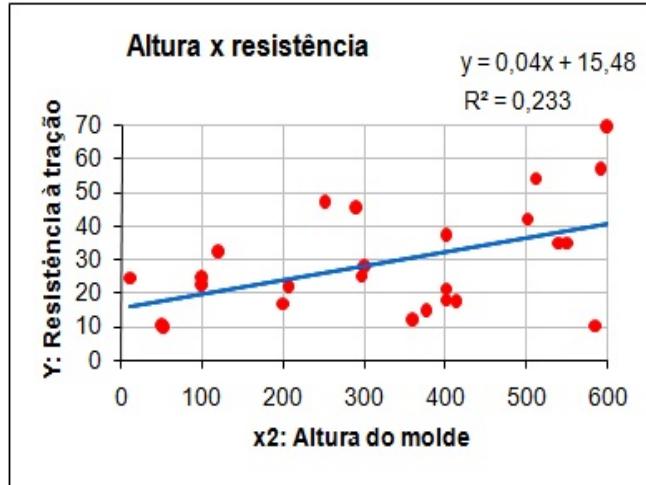
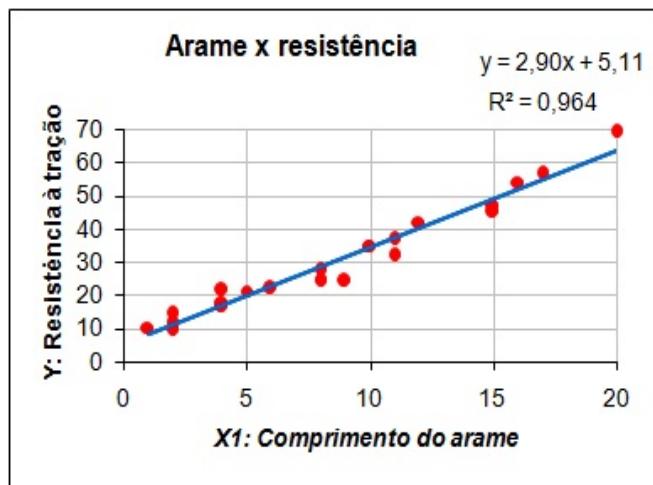
Exemplo 3 Voltando aos dados do exemplo anterior, vamos determinar a equação da reta ajustada.

$$a = \frac{n \sum (x \cdot y) - \sum x \cdot \sum y}{n \sum x^2 - (\sum x)^2} = \frac{25 \times 8008,47 - 206 \times 725,82}{25 \times 2396 - (206)^2} = 2,90$$

$$\bar{x} = \frac{\sum x}{n} = \frac{206}{25} = 8,24 \quad \text{e} \quad \bar{y} = \frac{\sum y}{n} = \frac{725,82}{25} = 29,03. \quad \text{Logo, } b = \bar{y} - a\bar{x} = 29,03 - 8,24 \times 2,90 = 5,13$$

Portanto, a reta de regressão é $y = 2,90x + 5,13$

OBS: Se repetirmos os mesmos procedimentos para a variável X_2 , teremos $r = 0,483$. Ao compararmos este coeficiente com o que determinamos para X_1 percebemos que a correlação com a variável Y é menor para X_2 .



(Gráficos elaborados utilizando Excel)

EXERCÍCIO 1 A tabela a seguir traz o custo para a produção de camisetas.

Quantidade (x)	Custo (y)	$x \times y$	x^2	y^2
0	100			
5	110			
10	120			
20	140			
50	200			
100	300			

- Construa um diagrama de dispersão.
- Calcule o coeficiente de Pearson e interprete.
- Obtenha a equação da reta que melhor descreve a relação entre as duas variáveis.
- Qual o custo estimado para a produção de 150 camisetas?
- Se o empresário dispõem de R\$ 350,00, quantas camisetas pode produzir?

EXERCÍCIO 2 A tabela abaixo são as notas de Estatística dos alunos de uma amostra e suas respectivas quantidades de faltas.

Aluno	Nota (x)	Faltas (y)	$x \times y$	x^2	y^2
1	0,0	10			
2	0,5	8			
3	1,5	6			
4	3,0	6			
5	4,5	4			
6	5,0	2			
7	6,0	4			
8	7,0	2			
9	8,5	0			
10	9,5	2			
<i>Total</i>					

- Construa um diagrama de dispersão.
- Calcule o coeficiente de Pearson e interprete.
- Obtenha a equação da reta que melhor descreve a relação entre as duas variáveis.
- Estime a quantidade de faltas do aluno com nota 8,0 ($x = 8,0$).
- Estime a quantidade de faltas do aluno com nota 2,0.
- Estime a nota do aluno com 2 faltas ($y = 2$).
- Estime a nota do aluno com 8 faltas.
- Estime a nota do aluno com 20 faltas.

EXERCÍCIO 3 A tabela a seguir relaciona os pesos (em centenas de libras) e as taxas de consumo de combustível em rodovias (em mi/gal) para uma amostra de carros de passeio novos.

X (Peso)	Y (Consumo)	$X \cdot Y$	X^2	Y^2
29	31			
35	27			
28	29			
44	25			
25	31			

34	29			
30	28			
33	28			
28	28			
24	33			

- a) Calcule o coeficiente de Pearson e interprete.
b) Obtenha a equação da reta que melhor descreve a relação entre as duas variáveis.

Valores críticos do Coeficiente de Correlação de Pearson

n	$\alpha = 5\%$	$\alpha = 1\%$
4	0,950	0,999
5	0,878	0,959
6	0,811	0,917
7	0,754	0,875
8	0,707	0,834
9	0,666	0,798
10	0,632	0,765
11	0,602	0,735
12	0,576	0,708
13	0,553	0,684
14	0,532	0,661
15	0,514	0,641
16	0,497	0,623
17	0,482	0,606
18	0,468	0,590
19	0,456	0,575
20	0,444	0,561
25	0,396	0,505
30	0,361	0,463
35	0,335	0,430
40	0,312	0,402
45	0,294	0,378
50	0,279	0,361
60	0,254	0,330
70	0,236	0,305
80	0,220	0,286
90	0,207	0,269
100	0,196	0,256