CORRELAÇÃO LINEAR E REGRESSÃO

Quando consideramos variáveis como peso e altura de um grupo de pessoas, uso do cigarro e incidência de câncer, procuramos verificar se existe alguma relação entre as variáveis e qual o grau dessa relação.

Ex: "Existe uma forte correlação entre desemprego e evasão escolar", "Observa-se estreita correlação entre o aumento do policiamento e a diminuição da criminalidade".

Constata-se, freqüentemente, a existência de uma relação entre duas (ou mais) variáveis. Se tal relação é de natureza quantitativa, a correlação é o instrumento adequado para descobrir e medir essa relação.

Uma vez caracterizada a relação, procuramos descrevê-la por meio de uma função matemática. A regressão é o instrumento adequado para a determinação dos parâmetros dessa função.

Quando estão em jogo somente duas variáveis, fala-se em correlação e regressão simples. Quando se trata de mais de duas variáveis, fala-se de correlação e regressão múltipla.

Exemplos:

variável independente (explicativa)		**variável dependente (resposta)**
renda	→	consumo
gasto com o controle da qualidade (R$)	→	número de defeitos nos produtos
memória ram do computador (gb)	→	tempo de resposta do sistema (seg) área construída do
imóvel (m2)	→	preço do imóvel (R$)

Uma **correlação simples** é uma relação entre duas variáveis. Os dados podem ser representados por pares ordenados (x,y) onde x é a variável independente (explicativa) e y é a variável dependente (resposta).

Coleta-se dados exibindo os valores correspondentes das variáveis. Faz-se o gráfico dos pontos em sistema de coordenadas retangulares. O conjunto resultante é chamado Diagrama de Dispersão.

Exemplo: X e Y representam, respectivamente, as notas de uma amostra de 10 alunos de matemática e estatística de uma classe de 98 alunos de uma faculdade. Uma amostra de 10 indivíduos acusaria alturas X_1, X_2, ..., X_{10}, e os correspondentes pesos Y_1, Y_2, ..., Y_{10}, . Os pontos a serem marcados no gráfico seriam, então (X_1,Y_1), (X_2,Y_2), ..., (X_{10},Y_{10}).

nº	Notas	
	Matemática (x_i)	Estatística (y_i)
1	5,0	6,0
8	8,0	9,0
24	7,0	8,0
38	10,0	10,0
44	6,0	5,0
58	7,0	7,0
59	9,0	8,0
72	3,0	4,0
80	8,0	6,0
92	2,0	2,0

Figura 1

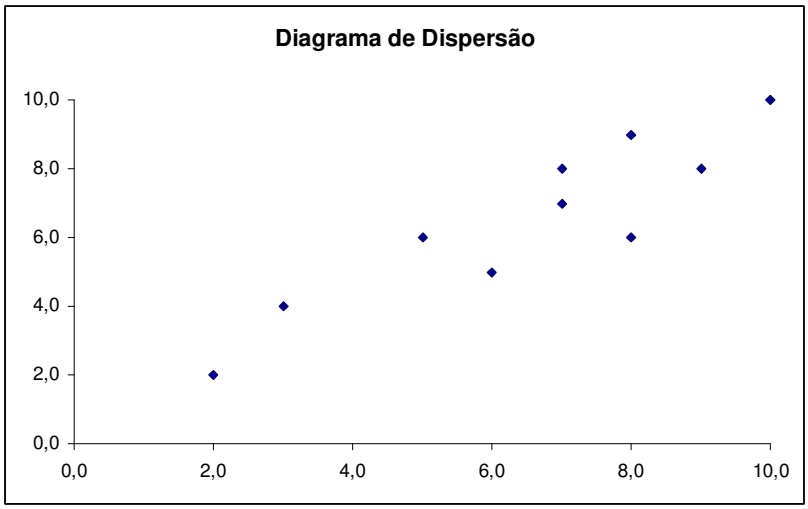


Figura 2

Pelo diagrama de dispersão (Figura 2), muitas vezes, se pode visualizar uma curva aproximativa dos dados. Os pontos obtidos, vistos em conjunto, formam uma elipse em diagonal. Quanto mais fina for a elipse, mais ela se aproximará de uma reta. Quando os dados parecem bem aproximados por uma reta, dizemos que há uma correlação linear entre as variáveis e que a correlação de forma elíptica tem como "imagem" uma reta (Figura 3). Quando existe um relacionamento entre as variáveis e tal relacionamento não é linear, diz-se, então, que há uma correlação não-linear entre as variáveis (Figura 5). Finalmente, há os casos em que o diagrama não sugere nenhum tipo de correlação entre as variáveis; neste caso diz-se que não há correlação linear (Figura 6).

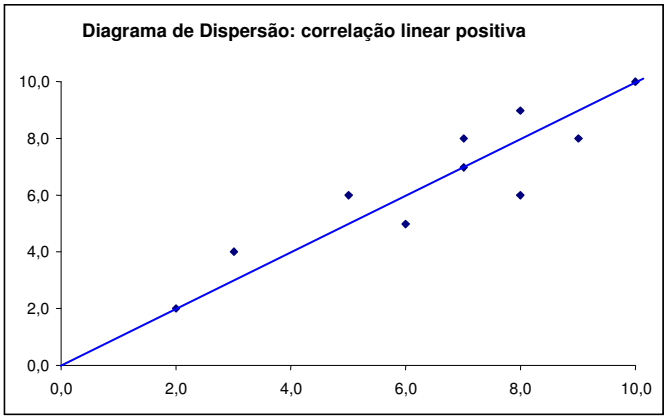


Figura 3

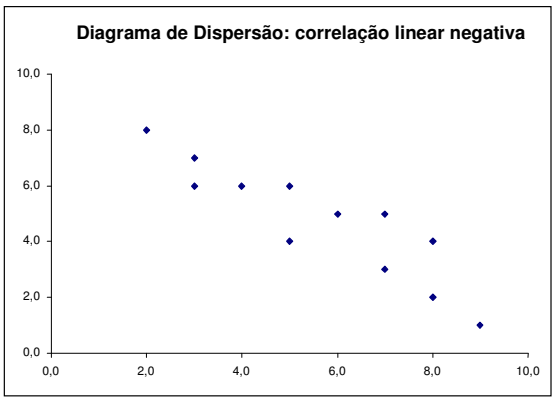


Figura 4

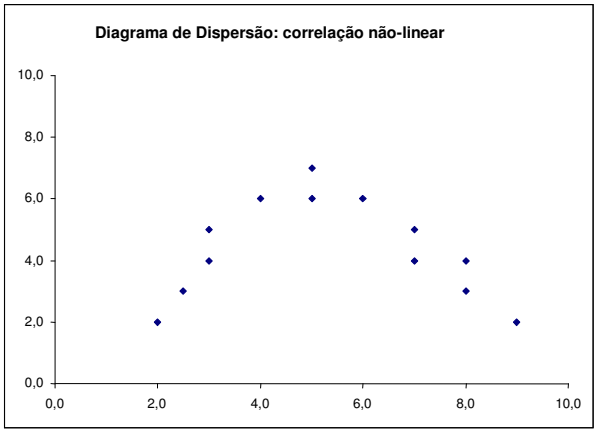


Figura 5

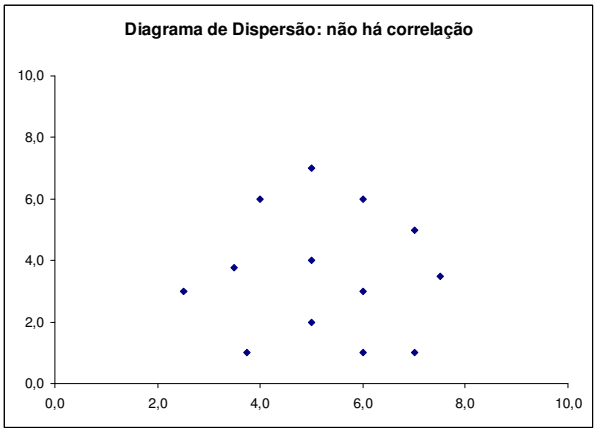


Figura 6

Então, caracterizada a relação, procuramos descrevê-la através de uma função matemática. A regressão é o instrumento adequado para a determinação dos parâmetros dessa função.

Coeficiente de correlação

Embora seja útil verificar a existência de correlação através do diagrama de dispersão, ele não nos fornece, com precisão, o grau de aderência entre as séries, ou seja, quão próximos estão os pontos em torno da reta.

O coeficiente de correlação linear é o instrumento empregado para a medida da correlação linear, indicando o grau de intensidade da correlação entre duas variáveis e, ainda, o sentido dessa correlação (positivo – Figura 1 ou negativo – Figura 2).

Pode ser utilizado o coeficiente de correlação de Pearson (em homenagem ao estatístico inglês Karl Pearson (1857-1936). O símbolo r representa o coeficiente de correlação amostral:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n.\sum x_i^{\,2} - \left(\sum x_i\right)^2\right]\left[n.\sum y_i^{\,2} - \left(\sum y_i\right)^2\right]}}$$

onde n é o número de observações.

Os valores limites de r (coeficiente de correlação) são –1 e +1, isto é, o valor de r pertence ao intervalo [-1,1].

Há correlação linear perfeita entre X e Y quando todos os pontos se encontram sobre uma mesma reta.

Se x e y tiverem forte correlação linear positiva, r estará próximo de 1. Se x e y tiverem forte correlação linear negativa, r estará próximo de -1. Se não existir correlação linear ou ainda se a correlação linear for fraca, r estará próximo de zero.

Quanto mais perto do zero, mais fraca é a correlação.

Assim:

a) se a correlação entre duas variáveis é perfeita e positiva, então r = 1;
b) se a correlação é perfeita e negativa, então r = -1;
c) se não há correlação entre as variáveis, r = 0.

Notas

Para a relação ser descrita por meio do **coeficiente de correlação de Pearson** é imprescindível que ela se aproxime de uma função linear.

Conclusões sobre o comportamento simultâneo das variáveis analisadas:

$0.6 \le |r| \le 1$: relação significativa entre as variáveis

$0.3 \le |r| < 0.6$: há uma relação relativamente fraca entre as variáveis

$0 < |r| < 0.3$: a relação é muito fraca e, praticamente nada podemos concluir sobre a relação entre as variáveis em estudo.

Exemplo: Calcular o coeficiente de correlação relativo a Figura 1.

Matemática (x_i)	Estatística (y_i)	x_iy_i	x_i^2	y_i^2
5	6	30	25	36
8	9	72	64	81
7	8	56	49	64
10	10	100	100	100
6	5	30	36	25
7	7	49	49	49
9	8	72	81	64
3	4	12	9	16
8	6	48	64	36
2	2	4	4	4
65	65	473	481	475

Logo:

$$r = \frac{10X\,473 - 65X\,65}{\sqrt{\left(10X\,481 - 65^2\right)\left(10X\,475 - 65^2\right)}} = \frac{4730 - 4225}{\sqrt{(4810 - 4225)(4750 - 4225)}} = \frac{505}{\sqrt{585X\,525}}\,\frac{505}{554{,}18} = 0{,}911$$

r = 0,911 indica uma correlação linear positiva altamente significativa entre as duas variáveis.

Exercício:

Calcule o coeficiente de correlação para os valores:

x_i	y_i	x_iy_i	x_i^2	y_i^2
4	12			
6	10			
8	8			
10	12			
12	14			

Regressão

O problema de se determinar equações de curvas que se ajustem a determinados conjuntos de dados observados é chamado ajustamento de curvas. Na prática, o próprio diagrama de dispersão geralmente sugere o tipo de curva a ser adotada. Assim, para as Figuras 3 e 4 poderíamos usar a reta ($Y = aX + b$) enquanto que, para a Figura 5, tentaríamos uma parábola ($Y = aX^2 + bX + c$). Às vezes é útil construir o diagrama em termos de variáveis transformadas. Assim, por exemplo, se log Y versus X conduz a uma reta, experimentaríamos $\log Y = a + bX$ como equação aproximadora.

Um dos principais objetivos do ajustamento é estimar uma das variáveis (a variável dependente) em função da outra (variável independente). Tal processo de estimação é denominado regressão.

Uma reta de regressão, também chamada de reta do ajuste ótimo, é aquela para o qual a soma dos quadrados dos resíduos é um mínimo. Sua equação pode ser usada para prever o valor de y para um dado valor de x.

Se Y for estimado em função de X por meio de uma equação, tal equação é denominada *equação de regressão de Y sobre X* e a curva ajustada é a *curva de regressão de Y sobre X*.

Para estimar a relação entre as variáveis:

$Y_i = \alpha X_i + \beta + \varepsilon_i$

Onde Y é a variável dependente, X e a variável dependente e ε é o erro aleatório.

Como trabalhamos com amostras, estaremos sempre calculando estimadores destes parâmetros, na forma:

$Y = aX + b$

Qual a função que melhor representa essa relação? Qual delas tem melhor aderência?

Portanto, deve ser escolhida a reta que apresenta as menores distâncias em relação aos pontos do diagrama; mais precisamente, aquela que minimiza a soma dos quadrados dos desvios (desvio é a distância com um sinal algébrico de + ou -). Por conseguinte, o método que utiliza esse critério para estimar os coeficientes a e b é conhecido como método dos mínimos quadrados.

Vamos considerar a regressão linear simples, utilizada quando uma reta representa de maneira satisfatória a relação entre as variáveis, ou seja, $Y = aX + b$ é a *equação de regressão de Y sobre X*.

O método mais simples utilizado para a determinação de *a* e *b* é o método dos mínimos quadrados. Após diversas simplificações é possível chegar a:

$$a = \frac{\sum\left[\left(x_i - \bar{x}\right)y_i\right]}{\sum\left(x_i - \bar{x}\right)^2} \qquad \text{e} \qquad \boxed{b = \bar{y} - a\bar{x}}$$

sendo $\bar{x}$ a média aritmética dos x; e $\bar{y}$ a média aritmética dos y.

Outra forma de se calcular o a é através da fórmula: $$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n.\sum x_i^{\,2} - \left(\sum x_i\right)^2}$$

Exemplo:

Matemática (x_i)	Estatística (y_i)	x_iy_i	x_i^2	y_i^2
5	6	30	25	36
8	9	72	64	81
7	8	56	49	64
10	10	100	100	100
6	5	30	36	25
7	7	49	49	49
9	8	72	81	64
3	4	12	9	16
8	6	48	64	36
2	2	4	4	4
65	65	473	481	475

$$a = \frac{10*473 - 65*65}{10*481 - 65^2} = \frac{4730 - 4225}{4810 - 4225} = \frac{505}{585} = 0{,}8632$$

$$\bar{x} = \frac{65}{10} = 6{,}5 \quad e \quad \bar{y} = \frac{65}{10} = 6{,}5$$

$$b = 6{,}5 - 0{,}8632 * 6{,}5 = 6{,}5 - 5{,}6108 = 0{,}8892$$

logo:

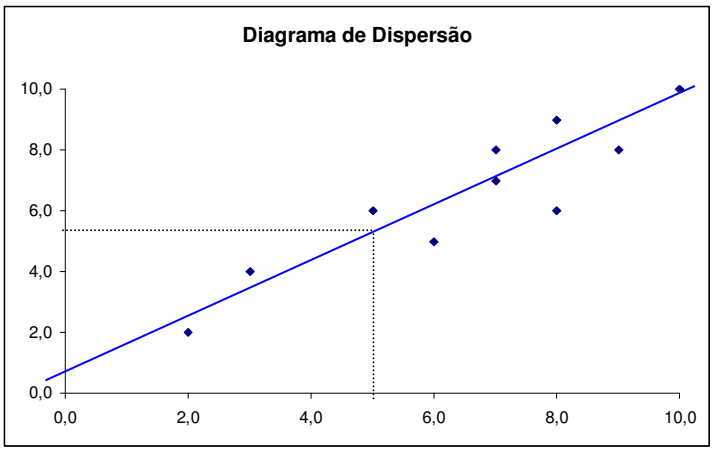
a = 0,86 e b = 0,89

$$\hat{Y} = 0{,}86X + 0{,}89$$

Para traçar a reta no gráfico, basta determinar dois de seus pontos:

X = 0 → $\hat{Y} = 0{,}89$

X = 5 → $\hat{Y} = 0{,}86 * 5 + 0{,}89 = 5{,}19$



Aplicação das retas de regressão

Após ter obtido a equação de uma reta de regressão, você pode usar essa equação para prever os valores y sob o intervalo dos dados (interpolação) se a correlação entre x e y for significante.

Exemplo:
Um executivo de publicidade pode prever as vendas da companhia baseado nos gastos de propaganda. Para prever valores y, substitua o valor dado de x na equação de regressão, calculando então $\hat{Y}$, o valor y previsto.

A equação de regressão para os gastos em propaganda (em milhares de dólares) e as vendas da companhia (em milhares de dólares) é

$\hat{Y} = 50{,}729x + 104{,}061$

Use esta equação para prever a expectativa de vendas da companhia com base nos gastos de propaganda a seguir:

a) US$ 1500 → $\hat{Y} = 50{,}729*1500 + 104{,}061$ → US$ 76.197,56
b) US$1800 → $\hat{Y} = 50{,}729*1800 + 104{,}061$ → US$ 91.416,26
c) US$2500 → $\hat{Y} = 50{,}729*2500 + 104{,}061$ → US$ 126.926,56

Obs: os valores previstos têm sentido somente para valores x no intervalo dos dados (ou próximos a ele).

Lista de Exercícios

Correlação e Regressão

1. A tabela abaixo apresenta os dados referentes à variação da demanda de um produto produzido (y_i) em relação à variação do preço da venda (x_i):

$\mathbf{x_i}$	40	45	52	58	65	70	85	90	100	120
$\mathbf{y_i}$	320	305	290	280	275	270	250	245	230	210

a. construa o diagrama de dispersão;
b. ajuste uma reta aos dados, ou seja, estabeleça a equação de regressão de y sobre x;
c. trace a reta no diagrama de dispersão;
d. determine y quando x = 80 e y quando x = 130.

2. Calcule o coeficiente de correlação relativo à tabela abaixo que apresenta as notas de Cálculo e Estatística de catorze alunos (n=14) e:

Cálculo (x_i)	8	7	4	9	6	4	7	6	5	8	2	7	3	6
Estatística (y_i)	7	9	4	7	5	6	9	6	8	9	4	6	2	7

a) construa o diagrama de dispersão;
b) estabeleça a equação de regressão de y sobre x;
c) trace a reta no diagrama de dispersão;

3. A tabela abaixo apresenta os dados referentes à variação do preço de venda do seu produto (y_i) em função do preço de custo (x_i):

$\mathbf{x_i}$	40	50	70	75	80	95	110	120
$\mathbf{y_i}$	130	140	145	160	160	170	180	200

a) construa o diagrama de dispersão;
b) estabeleça a equação de regressão de y sobre x;
c) trace a reta no diagrama de dispersão;
d) determine x quando y = 165 e x quando y = 190.

4. A tabela abaixo apresenta valores que mostram como o comprimento de uma barra de aço varia conforme a temperatura:

Temperat. (graus C)	10	15	20	25	30
Comprim/o (mm)	1003	1005	1010	1011	1014

a) determine o coeficiente de correlação;
b) estabeleça a equação de regressão de y sobre x;
c) o valor estimado do comprimento da barra para a temperatura de 18 graus C e para a temperatura de 35 graus C.

5. A tabela abaixo representa os pesos respectivos x e y de uma amostra de 12 pais e deus filhos mais velhos. Calcule o coeficiente de correlação e estime a linha de regressão de y para x .

x_i	65	63	67	64	68	62	70	66	68	67	69	71
y_i	68	66	68	65	69	66	68	65	71	67	68	70

6. Num determinado país, na última década, o aumento (x_i) percentual do nível de preços e a expansão percentual dos meios de pagamentos (y_i), de determinado produto de exportação, verificaram-se conforme a tabela abaixo:

ano	**1990**	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**
x_i	13	9	20	35	40	22	18	35	38	43
y_i	18	12	17	47	32	25	20	40	52	38

a) Estabeleça a equação de regressão de y sobre x;
b) Determine o coeficiente de correlação
c) Esboce o diagrama de dispersão.