Unidade 1 - Correlação

1. Introdução

A correlação e a regressão são duas técnicas estreitamente relacionadas que envolvem uma forma de estimação. Mais especificamente, *a análise de correlação e regressão compreende a análise de dados amostrais para saber **se** e **como** duas ou mais variáveis estão relacionadas uma com a outra numa população.*
A análise de correlação dá um número que resume o grau de relacionamento entre duas variáveis. A análise de regressão tem como resultado uma equação matemática que descreve o relacionamento. A equação pode ser utilizada para estimar, ou predizer, valores futuros de uma variável quando se conhecem ou se supõem conhecidos valores da outra variável.
A análise de correlação é útil em trabalho exploratório, quando um pesquisador ou analista procura determinar quais variáveis são potencialmente importantes e o interesse está basicamente no grau ou força do relacionamento. Em outras situações, focaliza-se mais a natureza do relacionamento, isto é, a equação de predição, e a análise de regressão é o instrumento principal.
Os dados para a análise de regressão e correlação provêm de observações de variáveis emparelhadas. Para um problema de duas variáveis, isto significa que cada observação origina dois valores, um para cada variável. Por exemplo, um estudo que envolva características físicas pode focalizar a idade e a altura de cada indivíduo. As duas variáveis de interesse – idade e altura de cada pessoa – são então emparelhadas.

2.. Correlação

O objetivo da análise de correlação é a determinação da força do relacionamento entre duas observações emparelhadas. O termo *correlação* significa literalmente *co-relacionamento*, pois indica até que ponto os valores de uma variável estão relacionados com os de outra. Há muitos casos em que pode existir um relacionamento entre duas variáveis. Consideremos, por exemplo, questões como estas:

- ✓ A idade e a resistência física estão correlacionadas?
- ✓ Pessoas de maior renda tendem a apresentar melhor escolaridade?
- ✓ O sucesso num emprego pode ser predito com base no resultado de testes?
- ✓ A temperatura parece influenciar a taxa de criminalidade?
- ✓ Estudantes com maior capacidade de leitura tendem a obter melhores resultados em cursos de matemática?

Problemas como esses se prestam à análise de correlação. O resultado de tal análise é um coeficiente de correlação – um valor que quantifica o grau de correlação. Ao realizarmos uma análise de correlação é interessante construirmos um gráfico chamado *diagrama de dispersão* que nos auxilia a formular conclusões intuitivas sobre dados emparelhados. Observe os diagramas de dispersão a seguir:

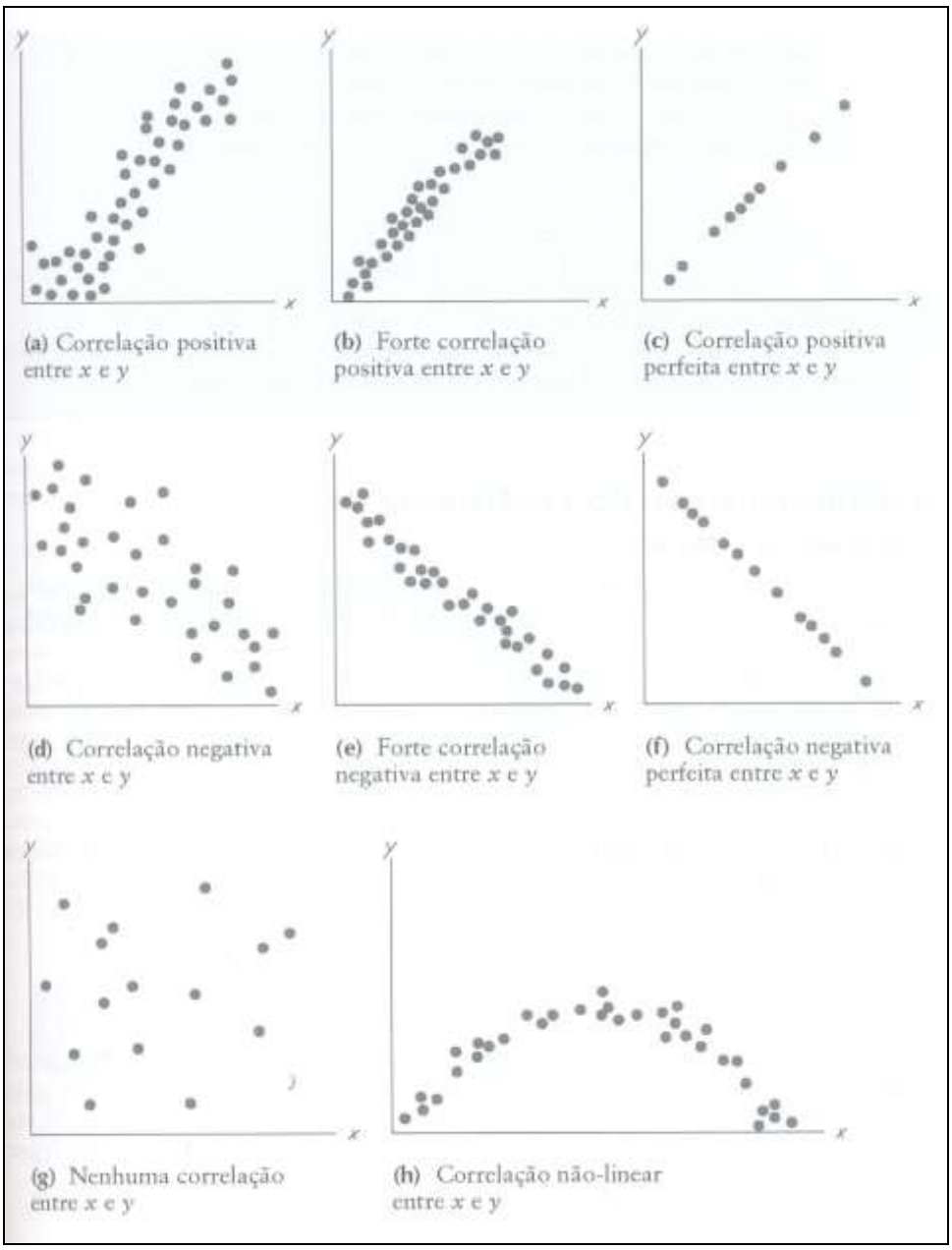

Figura 1 – Diagramas de dispersão
Fonte: Triola, 1998

Analisando a figura acima podemos observar que: Os gráficos **a**, **b** e **c** exibem um padrão de valores crescentes de y que correspondem a valores crescentes de x. Nos diagramas de dispersão **d**, **e** e **f** os valores de y decrescem quando os valores de x crescem. Nos gráficos **c** e **f** o padrão de pontos aproxima-se de uma linha reta, sugerindo uma relação mais forte entre x e y. Em contraste com os seis primeiros gráficos, o diagrama **g** não apresenta qualquer padrão definido e sugere que não há correlação (ou relacionamento) entre x e y. E, finalmente, o diagrama de dispersão **h** exibe um padrão que não é linear. Porém, as conclusões tiradas de diagramas de dispersão tendem a ser subjetivas necessitando, portanto, de métodos mais precisos e objetivos.

A forma mais comum de análise de correlação envolve dados contínuos. O grau de relacionamento entre duas variáveis contínuas é sintetizado por um coeficiente de correlação conhecido como *r de Pearson* ou *coeficiente de correlação momento-produto de Pearson*, em homenagem ao matemático Karl Pearson, que desenvolveu a técnica. Essa técnica só é válida se pudermos levantar certas suposições um tanto quando rígidas. As suposições são:

1. A amostra de dados emparelhados (x, y) é aleatória.
2. Os pares de dados (x, y) têm uma distribuição normal bivariada, ou seja, para qualquer valor fixo de x, os valores correspondentes de y tenham uma distribuição normal, e vice-versa.

O coeficiente de correlação tem duas propriedades que caracterizam a natureza de uma relação entre duas variáveis. Uma é o seu sinal (+ ou -) e a outra é a sua magnitude:

- ✓ Quando r tem sinal positivo indica que a valores altos **(baixos)** de uma das variáveis, correspondem a valores altos **(baixos)** da outra.
- ✓ Quando r tem sinal negativo indica que a valores altos **(baixos)** de uma das variáveis, correspondem a valores baixos **(altos)** da outra.
- ✓ Quando o valor de r está próximo de –1,00 ou +1,00 indica que os pares (x, y) estão muito próximos de uma reta, ou mesmo sobre uma reta.
- ✓ Quando o valor de r está próximo de 0 indica que os pares (x, y) estão mais dispersos.

Em geral, para calcularmos o coeficiente de correlação linear de Pearson utilizamos a seguinte expressão:

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

A notação empregada na fórmula para o cálculo do coeficiente de correlação pode ser definida da seguinte maneira:

n $\rightarrow$ representa o número de pares de dados.

$\sum x \rightarrow$ denota a soma de todos os valores de x.

$\sum x^2 \rightarrow$ indica que devemos elevar ao quadrado cada valor de x e somar os resultados.

$\left(\sum x\right)^2 \rightarrow$ indica que devemos somar os valores de x e elevar o total ao quadrado. É extremamente importante não confundir $\sum x^2$ com $\left(\sum x\right)^2$.

$\sum xy \rightarrow$ indica que devemos multiplicar cada valor de x pelo valor correspondente de y e somar então todos esses produtos.

r $\rightarrow$ representa o coeficiente de correlação linear para uma *amostra*.

$\rho \rightarrow$ representa o coeficiente de correlação linear para uma *população*.

Exemplo 1 – (Stevenson, 2001) Suponha-se que estejamos interessados em saber se o desempenho do estudante na universidade está relacionado com seu desempenho no curso secundário. Para avaliar isto, imaginemos 15 universitários escolhidos aleatoriamente numa grande universidade. Os dados encontram-se a seguir:

Aluno	Secundário (Classificação em %)	Universidade (Classificação)
1	80	1
2	82	1
3	84	2,1
4	85	1,4
5	87	2,1
6	88	1,7
7	88	2
8	89	3,5
9	90	3,1
10	91	2,4
11	91	2,7
12	92	3
13	94	3,9
14	96	3,6
15	98	4

Para iniciarmos o estudo da relação entre a classificação do estudante no secundário e na universidade vamos construir o diagrama de dispersão, pois esse gráfico proporciona uma visualização do relacionamento e se a relação linear é concebível:

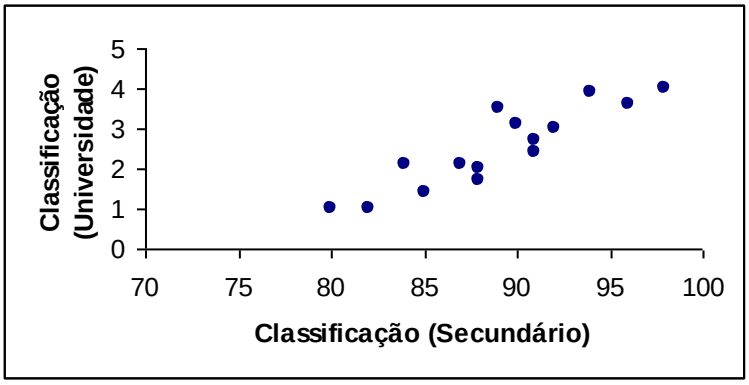

Figura 2 – Diagrama de dispersão das classificações dos 15 estudantes no secundário e na universidade.

Nosso gráfico parece indicar que existe uma relação positiva moderada, pois em geral médias baixas no secundário parecem estar associadas a médias baixas na universidade, enquanto que médias altas no secundário e na universidade parecem corresponder-se, embora haja algumas exceções.

Vamos agora calcular o coeficiente de correlação linear de Pearson para obtermos uma medida da magnitude desse relacionamento:

Aluno	Secundário (Classificação em %) x	Universidade (Classificação) y	xy	x^2	y^2
1	80	1	80	6400	1
2	82	1	82	6724	1
3	84	2,1	176,4	7056	4,41
4	85	1,4	119	7225	1,96
5	87	2,1	182,7	7569	4,41
6	88	1,7	149,6	7744	2,89
7	88	2	176	7744	4
8	89	3,5	311,5	7921	12,25
9	90	3,1	279	8100	9,61
10	91	2,4	218,4	8281	5,76
11	91	2,7	245,7	8281	7,29
12	92	3	276	8464	9
13	94	3,9	366,6	8836	15,21
14	96	3,6	345,6	9216	12,96
15	98	4	392	9604	16
Soma=	1335	37,5	3400,5	119165	107,75

$$r = \frac{15(3400{,}5) - (1335)(37{,}5)}{\sqrt{\left[15(119155) - (1335)^2\right]\left[15(107{,}75) - (37{,}5)^2\right]}} = +0{,}90$$

Exemplo 2 – Uma seguradora de automóveis estabeleceu a meta de expandir sua participação no mercado de seguros. Para isso, abriu novas sucursais e contratou novos corretores. A equipe responsável pelo treinamento dos novos corretores decidiu avaliar a relação entre o número de dias de treinamento e o desempenho alcançado em um teste simulado de vendas. O período de treinamento variou de meio dia a dois dias e meio. Após o treinamento, todos os corretores participaram de uma série de situações simuladas de vendas e, de acordo com seu desempenho, receberam uma nota que podia variar de 0 a 130. Os dados encontram-se a seguir:

Dias de treinamento	Desempenho
0,5	46
0,5	51
1	71
1	75
1,5	92
1,5	99
2	105
2	112
2,5	121
2,5	125

a) Construa o diagrama de dispersão para os dados.

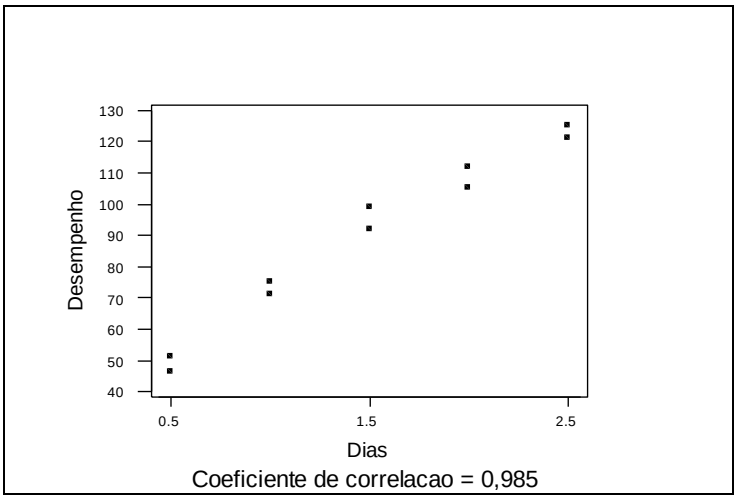

Figura 3 - Diagrama de dispersão para o desempenho no teste simulado de vendas em função do número de dias de treinamento.

b) Calcule o coeficiente de correlação linear de Pearson.

	X	Y	X*Y	X^2	Y^2
	Dias de treinamento	Desempenho			
	0,5	46	23	0,25	2116
	0,5	51	25,5	0,25	2601
	1	71	71	1	5041
	1	75	75	1	5625
	1,5	92	138	2,25	8464
	1,5	99	148,5	2,25	9801
	2	105	210	4	11025
	2	112	224	4	12544
	2,5	121	302,5	6,25	14641
	2,5	125	312,5	6,25	15625
Soma:	15	897	1530	27,5	87483
n:	10				

$$r = \frac{10(1530) - (15)(897)}{\sqrt{\left[10(27{,}5) - (15)^2\right]\left[10(87483) - (897)^2\right]}} = +0{,}985$$

Propriedades do coeficiente de correlação linear de Pearson

1. O valor de r está sempre entre –1,00 e +1,00. Isto é, -1,00 < r < +1,00.
2. O valor de r não varia se todos os valores de qualquer uma das variáveis são convertidos para uma escala diferente. Por exemplo, estamos estudando o relacionamento entre peso (em kg) e altura (em metros) e transformarmos todos os valores de peso para gramas, o valor de r não se modificará.
3. O valor de r não é afetado pela escolha de x ou y. Permutando todos os valores de x e y, r permanecerá inalterado.
4. r mede a intensidade, ou grau, de relacionamento linear. Não serve para medir a intensidade de um relacionamento não-linear.

Erros comuns que envolvem a correlação

Identificamos a seguir os erros mais comuns cometidos na interpretação de resultados que envolvem correlação:

1. *Devemos evitar a conclusão de que a correlação implica causalidade.* Um estudo mostrou uma correlação forte e positiva entre o número de doentes mentais no Reino Unido e o número de aparelhos de rádio, para o período de 1924 a 1937. Observe os resultados abaixo:

Ano	Número de doentes mentais por 10 000 habitantes	Número de aparelhos de rádio (em milhões)
1924	8	1350
1925	8	1960
1926	9	2270
1927	10	2483
1928	11	2730
1929	11	3091
1930	12	3647
1931	16	4620
1932	18	5497
1933	19	6260
1934	20	7012
1935	21	7618
1936	22	8131
1937	23	8593

Fonte: Montgomery & Peck (1992)

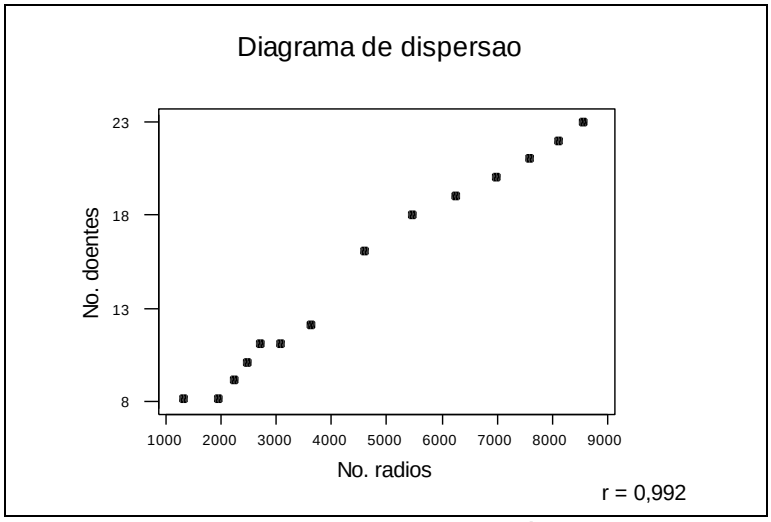

Figura 3 – Diagrama de dispersão do número de doentes mentais e do número de aparelhos de rádio.

✓ Mesmo apresentando um *r* tão alto é extremamente improvável que o número de doentes mentais esteja funcionalmente relacionado ao número de aparelhos de rádio existentes.

✓ A razão para esta forte correlação é o fato de que as duas variáveis aumentaram simultaneamente ao longo dos anos considerados.

✓ O número de doentes mentais aumentou porque os procedimentos para diagnóstico de doenças mentais foram se tornando cada vez mais sofisticados e o número de aparelhos de rádio aumentou devido à queda no preço, ou seja, havia a atuação de outros fatores também chamados de variáveis ocultas. Define-se formalmente uma variável oculta como uma variável que afeta as variáveis em estudo, mas não está incluída no estudo.

2. *Devemos prestar atenção à propriedade de linearidade.* A conclusão de que não há correlação linear significativa não quer dizer que x e y não estejam relacionados de alguma forma. Os dados da figura abaixo conduzem a um valor de r=0, que é uma indicação da ausência total de correlação linear entre as duas variáveis; mas pode-se ver facilmente pela figura que o padrão dos dados reflete forte relacionamento não-linear.

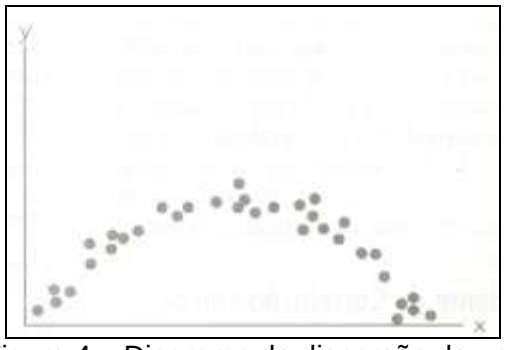

Figura 4 – Diagrama de dispersão de um relacionamento não-linear.

Exercícios

Para cada exercício a seguir:

a) Identifique as variáveis x e y.
b) Calcule o coeficiente de correlação.

1. Em uma amostra de 8 funcionários de uma empresa, observou-se duas variáveis: anos de empresa (A) e número de promoções recebidas (P). Os dados encontram-se a seguir:

A	5	6	6	7	7	8	8	8
P	2	2	1	2	0	3	1	0

r= -0,1944

2. Uma imobiliária realizou um estudo com o objetivo de verificar o sentido e a força de correlação entre a "idade" dos imóveis (em anos) e o preço do aluguel (em centenas de dólares):

"Idade"	3	12	5	7	8	19	10	22	15	8	25
Preço do aluguel	5	3,2	4	3,3	4,5	1,3	3	1,4	2,8	5,1	2,6

r = -0,8133

3. A Companhia dos Sonhos Gelados produz e comercializa sorvetes. A área comercial da empresa resolveu analisar alguns dados referentes aos últimos anos. Analisou a temperatura média no verão e o volume de vendas nesta mesma estação. Obteve os números apresentados na tabela seguinte:

Temperatura (em ºC)	32	28	33	27	26	36	34	30	31	29
Vendas (em mil unidades)	83	78	80	75	71	92	85	81	83	79

r = 0,9269

4. Uma empresa resolveu comparar o número de horas de treinamentos preventivos com o número de acidentes verificados nas suas instalações. Obteve os números apresentados na tabela seguinte:

Nº de treinamentos	14	12	18	25	32	44	17	28
Nº de acidentes	49	52	45	46	41	35	49	44

r = -0,9639

5. As vendas (em R$1000,00) da indústria Pirapora nos últimos 11 meses estão apresentadas na tabela seguinte:

Mês	1	2	3	4	5	6	7	8	9	10	11
Vendas	40	43	51	54	58	62	57	65	60	68	72

r = 0,9409

6. O Residencial Universitários é um complexo de 300 apartamentos localizados perto da Universidade do Bom Saber. A gerente Mariana Bagdeve suspeita que existe uma relação entre o número de apartamentos alugados em cada semestre e o número de estudantes matriculados na universidade. As matrículas na universidade (em 1000) e o número de apartamentos alugados durante os oito últimos semestres são:

Semestre	1	2	3	4	5	6	7	8
Nº de matrículas	7,2	6,3	6,7	7,0	6,9	6,4	7,1	6,7
Nº de apartamentos alugados	291	228	252	265	270	240	288	246

r = 0,9628