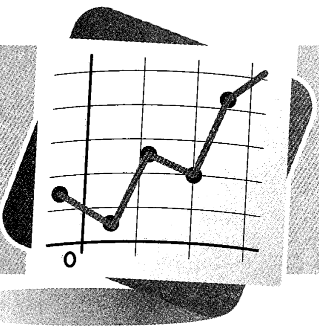


11 CORRELAÇÃO E REGRESSÃO

11.1 Introdução

Nos capítulos anteriores, nossa preocupação era descrever a distribuição de valores de uma única variável. Com esse objetivo, aprendemos a calcular medidas de tendência central e variabilidade.

Quando, porém, consideramos observações de duas ou mais variáveis, surge um novo problema: as **relações** que podem existir entre as variáveis estudadas. Nesse caso, as medidas estudadas não são eficientes.

Assim, quando consideramos variáveis como peso e altura de um grupo de pessoas, uso do cigarro e incidência do câncer, vocabulário e compreensão da leitura, dominância e submissão, procuramos verificar se existe alguma relação entre as variáveis de cada um dos pares e qual o grau dessa relação. Para isso, é necessário o conhecimento de novas medidas.

Sendo a relação entre as variáveis de natureza **quantitativa**, a **correlação** é o instrumento adequado para descobrir e medir essa relação.

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

NOTA:

- Ficaremos restritos às relações entre duas variáveis (correlação simples).

11.2 Correlação

11.2.1 Relação funcional e relação estatística

Como sabemos, o perímetro e o lado de um quadrado estão relacionados. A relação que os liga é perfeitamente definida e pode ser expressa por meio de uma sentença matemática:

$$2p = 4\ell,$$

onde 2p é o perímetro e ℓ é o lado.

Atribuindo-se, então, um valor qualquer a ℓ, é possível determinar **exatamente** o valor de 2p.

Consideremos, agora, a relação que existe entre o peso e a estatura de um grupo de pessoas. É evidente que essa relação não é do mesmo tipo da anterior; ela é bem menos precisa. Assim, pode acontecer que a estaturas diferentes correspondam pesos iguais ou que a estaturas iguais correspondam pesos diferentes. Contudo, em média, quanto maior a estatura, maior o peso.

As relações do tipo perímetro – lado são conhecidas como **relações funcionais** e as do tipo peso – estatura, como **relações estatísticas**.

Quando duas variáveis estão ligadas por uma **relação estatística**, dizemos que existe **correlação** entre elas.

NOTA:

- As relações funcionais são um caso limite das relações estatísticas.

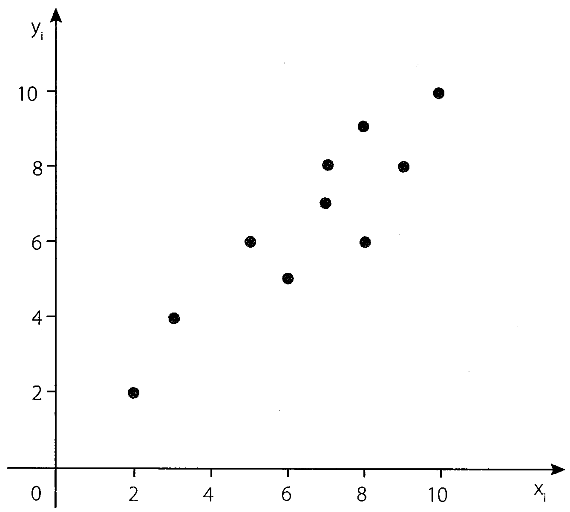
11.2.2 Diagrama de dispersão

Consideremos uma amostra aleatória, formada por dez dos 98 alunos de uma classe da faculdade **A** e pelas notas obtidas por eles em Matemática e Estatística:

Nºs	NOTAS	
	MATEMÁTICA (x_i)	ESTATÍSTICA (y_i)
01	5,0	6,0
08	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

TABELA 11.1

Representando, em um sistema coordenado cartesiano ortogonal, os pares ordenados (x_i, y_i), obtemos uma nuvem de pontos que denominamos **diagrama de dispersão.** Esse diagrama nos fornece uma ideia grosseira, porém útil, da correlação existente.

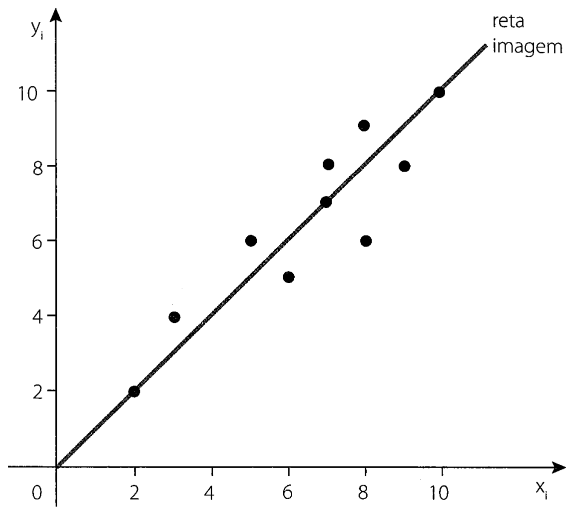


11.2.3 Correlação linear

Os pontos obtidos, vistos em conjunto, formam uma elipse em diagonal.

Podemos imaginar que, quanto mais fina for a elipse, mais ela se aproximará de uma reta. Dizemos, então, que a correlação de forma elíptica tem como "imagem" uma reta, sendo, por isso, denominada **correlação linear**.

É possível verificar que a cada correlação está associada como "imagem" uma relação funcional. Por esse motivo, as relações funcionais são chamadas **relações perfeitas**.



Como a correlação em estudo tem como "imagem" uma reta ascendente, ela é chamada **correlação linear positiva**.

Assim, uma correlação é:

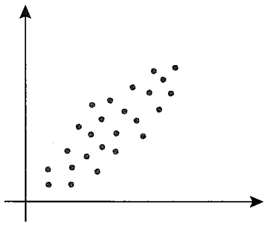
a. **linear positiva** se os pontos do diagrama têm como "imagem" uma reta ascendente;

b. **linear negativa** se os pontos têm como "imagem" uma reta descendente;

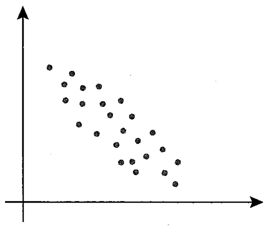
c. **não linear** se os pontos têm como "imagem" uma curva.

Se os pontos apresentam-se dispersos, não oferecendo uma "imagem" definida, concluímos que não há relação alguma entre as variáveis em estudo.

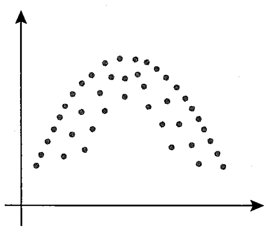
Temos, então:



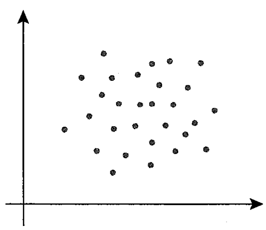
correlação linear positiva



correlação linear negativa



correlação não linear



não há correlação

11.2.4 Coeficiente de correlação linear

O instrumento empregado para a medida da correlação linear é o **coeficiente de correlação**. Esse coeficiente deve indicar o grau de intensidade da correlação entre duas variáveis e, ainda, o sentido dessa correlação (positivo ou negativo).

Faremos uso do **coeficiente de correlação de Pearson**, que é dado por:

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

onde **n** é o número de observações.

Os valores limites de **r** são −1 e +1, isto é, o valor de **r** pertence ao intervalo [−1, +1].

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, então **r = 0**.

Logicamente:

a. se $r = +1$, há uma correlação perfeita e positiva entre as variáveis;

b. se $r = -1$, há uma correlação perfeita e negativa entre as variáveis;

c. se $r = 0$, ou **não há correlação** entre as variáveis, ou a relação que porventura exista **não é linear**.

NOTAS:

- Para que uma relação possa ser descrita por meio do **coeficiente de correlação de Pearson** é imprescindível que ela se aproxime de uma função linear. Uma maneira prática de verificarmos a linearidade da relação é a inspeção do diagrama de dispersão: se a elipse apresenta saliências ou reentrâncias muito acentuadas, provavelmente trata-se de **correlação curvilínea**.
- Para podermos tirar algumas conclusões significativas sobre o comportamento simultâneo das variáveis analisadas, é necessário que:

  $$0{,}6 \leq |\, r \,| \leq 1.$$

  Se $0{,}3 \leq |\, r \,| < 0{,}6$, há uma correlação relativamente fraca entre as variáveis.

  Se $0 < |\, r \,| < 0{,}3$, a correlação é muito fraca e, praticamente, nada podemos concluir sobre a relação entre as variáveis em estudo.

Vamos, então, calcular o coeficiente de correlação relativo à Tabela 11.1. O modo mais prático para obtermos **r** é abrir, na tabela, colunas correspondentes aos valores de x_iy_i, x_i^2 e y_i^2. Assim:

	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
n = 10	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
	$\Sigma = 65$	$\Sigma = 65$	$\Sigma = 473$	$\Sigma = 481$	$\Sigma = 475$

TABELA 11.2

Logo:

$$r = \frac{10 \times 473 - 65 \times 65}{\sqrt{(10 \times 481 - 65^2)\ (10 \times 475 - 65^2)}} = \frac{4.730 - 4.225}{\sqrt{(4.810 - 4.225)\ (4.750 - 4.225)}}$$

$$= \frac{505}{\sqrt{585 \times 525}} = \frac{505}{554{,}18} = 0{,}911$$

Daí:

$$r = 0{,}91,$$

resultado que indica uma correlação linear positiva altamente significativa entre as duas variáveis.

Resolva

1. Complete o esquema de cálculo do coeficiente de correlação para os valores das variáveis x_i e y_i:

x_i	4	6	8	10	12
y_i	12	10	8	12	14

Temos:

	x_i	y_i	x_iy_i	x_i^2	y_i^2
n = 5	4	12	48	16	144
	....	....	....	....	....
	....	....	....	....	....
	....	....	....	....	....
	12	14	168	144	196
	$\Sigma = ...$	$\Sigma = ...$	$\Sigma = ...$	$\Sigma = ...$	$\Sigma = ...$

Logo:

$$r = \frac{.... \times .... - .... \times ....}{\sqrt{(.... \times .... - .... \times ....)\,(.... \times .... - .... \times ....)}} = \frac{.... - ....}{\sqrt{(.... - ....)\,(.... - ....)}} = ,$$

$$= \frac{.....}{\sqrt{....} \times ....} = \frac{.....}{\sqrt{....}} = \frac{.....}{.....} = ... ,$$

donde r = 0,42.

A correlação linear entre as variáveis X e Y é positiva, porém fraca.

11.3 Regressão

11.3.1 Ajustamento da reta

Sempre que desejamos estudar determinada variável em função de outra[1], fazemos uma **análise de regressão**.

Podemos dizer que a **análise de regressão** tem por objetivo descrever, através de um modelo matemático, a relação entre duas variáveis, partindo de **n** observações das mesmas.

A variável sobre a qual desejamos fazer uma estimativa recebe o nome de **variável dependente** e a outra recebe o nome de **variável independente**.

Assim, supondo **X** a variável independente e **Y** a dependente, vamos procurar determinar o ajustamento de uma reta à relação entre essas variáveis, ou seja, vamos obter uma função definida por:

$$Y = aX + b,$$

onde **a** e **b** são os parâmetros.

Sejam duas variáveis **X** e **Y**, entre as quais exista uma correlação acentuada, embora não perfeita, como, por exemplo, as que formam a Tabela 11.2.

---

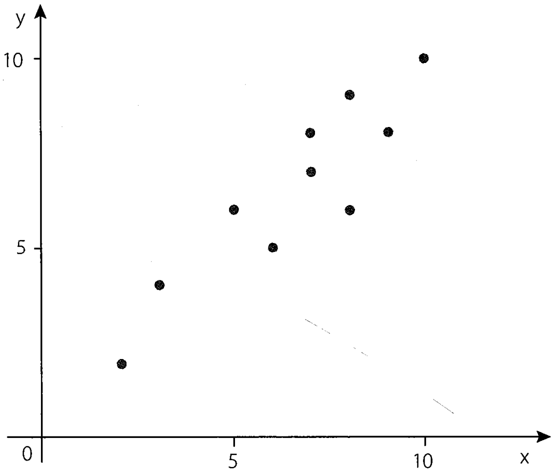
[1] Lembre-se de que estamos restritos à regressão linear simples.

Daí, temos:

x_i	5	8	7	10	6	7	9	3	8	2
y_i	6	9	8	10	5	7	8	4	6	2

TABELA 11.3

cujo diagrama de dispersão é dado por:



Podemos concluir, pela forma do diagrama, que se trata de uma correlação retilínea, de modo a permitir o ajustamento de uma reta, imagem da função definida por:

$$Y = aX + b$$

Vamos, então, calcular os valores dos parâmetros **a** e **b** com a ajuda das fórmulas:

$$a = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

e

$$b = \bar{y} - a\bar{x}$$

onde:

n é o número de observações;

$\bar{\mathbf{x}}$ é a média dos valores x_i $\left(\bar{x} = \frac{\sum x_i}{n}\right)$;

$\bar{\mathbf{y}}$ é a média dos valores y_i $\left(\bar{y} = \frac{\sum y_i}{n}\right)$.

NOTA:

- Como estamos fazendo uso de uma amostra para obtermos os valores dos parâmetros, o resultado, na realidade, é uma **estimativa** da verdadeira equação de regressão. Sendo assim, escrevemos:

$$\hat{Y} = aX + b,$$

onde $\hat{\mathbf{Y}}$ é o **Y estimado**.

Formemos, então, a tabela de valores:

	x_i	y_i	x_iy_i	x_i^2
n = 10	5	6	30	25
	8	9	72	64
	7	8	56	49
	10	10	100	100
	6	5	30	36
	7	7	49	49
	9	8	72	81
	3	4	12	9
	8	6	48	64
	2	2	4	4
	$\Sigma = 65$	$\Sigma = 65$	$\Sigma = 473$	$\Sigma = 481$

TABELA 11.4

Temos, assim:

$$a = \frac{10 \times 473 - 65 \times 65}{10 \times 481 - (65)^2} = \frac{4.730 - 4.225}{4.810 - 4.225} = \frac{505}{585} = 0,8632$$

Como:

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

vem:

$$b = 6,5 - 0,8632 \times 6,5 = 6,5 - 5,6108 = 0,8892,$$

donde:

$$a = 0,86 \text{ e } b = 0,89$$

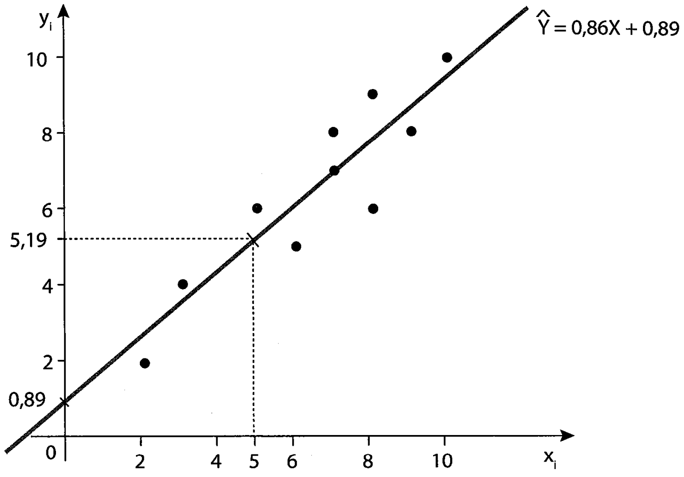
Logo:

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

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

$$X = 0 \Rightarrow \hat{Y} = 0,89$$
$$X = 5 \Rightarrow \hat{Y} = 0,86 \times 5 + 0,89 = 5,19$$

Assim, temos:



11.3.2 Interpolação e extrapolação

Voltando à Tabela 11.1, vemos que **4,0** não figura entre as notas de Matemática. Entretanto, podemos estimar a nota correspondente em Estatística fazendo **X = 4,0** na equação:

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

Assim:

$$X = 4,0 \Rightarrow \hat{Y} = 0,86 \times 4,0 + 0,89 = 4,33$$

O mesmo acontece com a nota **1,0**. Repetindo o procedimento, temos:

$$X = 1,0 \Rightarrow \hat{Y} = 0,86 \times 1,0 + 0,89 = 1,75$$

Como $4 \in [2, 10]$, dizemos que foi feita uma **interpolação**; e como $1 \notin [2,10]$, dizemos que foi feita uma **extrapolação**.

NOTA:

- Uma norma fundamental no uso de equações de regressão é a de nunca extrapolar, exceto quando considerações teóricas ou experimentais demonstrem a possibilidade de extrapolação.

Resolva

1. Complete o esquema para o ajustamento de uma reta aos dados:

x_i	2	4	6	8	10	12	14
y_i	30	25	22	18	15	11	10

Temos:

	x_i	y_i	x_iy_i	x_i^2
	2	30	60	4
	....	....	....	....
	....	....	....	....
	....	....	....	....
n = 7	....	....	....	....
	....	....	....	....
	14	10	140	196
	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$ ....

Logo:

$$a = \frac{.... \times .... - .... \times ....}{.... \times .... - (....)^2} = \frac{.... - ....}{.... - ....} = \frac{....}{....} = ...$$

$$b = ... - (....) = ... + ... = ...$$

donde:

$a = ...$ e $b = ...$,

isto é:

$$\hat{Y} = -1{,}7X + 32{,}3$$

Exercícios

1. Um grupo de pessoas fez uma avaliação do peso aparente de alguns objetos. Com o peso real e a média dos pesos aparentes, dados pelo grupo, obteve-se a tabela:

PESO REAL	18	30	42	62	73	97	120
PESO APARENTE	10	23	33	60	91	98	159

Calcule o índice de correlação.

2. Considere os resultados de dois testes, **X** e **Y**, obtidos por um grupo de alunos da escola **A**:

x_i	11	14	19	19	22	28	30	31	34	37
y_i	13	14	18	15	22	17	24	22	24	25

a. Verifique, pelo diagrama, se existe correlação retilínea.

b. Em caso afirmativo, calcule o coeficiente de correlação.

c. Escreva, em poucas linhas, as conclusões a que chegou sobre a relação entre essas variáveis.

3. A tabela abaixo apresenta a produção de uma indústria:

ANOS	1980	1981	1982	1983	1984	1985	1986	1987	1988
QUANTIDADES (t)	34	36	36	38	41	42	43	44	46

Calcule:

a. o coeficiente de correlação;

Sugestão: Para simplificar os cálculos, use para o tempo uma variável auxiliar, por exemplo: $x_i' = x_i - 1984$.

b. a reta ajustada;

c. a produção estimada para 1989.

NOTA:
- Lembre-se de que foi usada uma variável auxiliar.

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

TEMPERATURA (°C)	10	15	20	25	30
COMPRIMENTO (mm)	1.003	1.005	1.010	1.011	1.014

Determine:

a. o coeficiente de correlação;

b. a reta ajustada a essa correlação;

c. o valor estimado do comprimento da barra para a temperatura de 18°C;

d. o valor estimado do comprimento da barra para a temperatura de 35°C.

5. A variação do valor da UPC, relativamente a alguns meses de 1995, deu origem à tabela:

MESES	mai.	jun.	jul.	ago.	set.	out.	nov.
VALORES R$	10,32	10,32	11,34	11,34	11,34	12,22	12,22

a. Calcule o grau de correlação.

b. Estabeleça a equação de regressão de **Y** sobre **X**.

c. Estime o valor da UPC para o mês de dezembro.

Sugestão: Substitua os meses, respectivamente, por 1, 2, ..., 7.

6. A partir da tabela:

x_i	1	2	3	4	5	6
y_i	70	50	40	30	20	10

a. calcule o coeficiente de correlação;
b. determine a reta ajustada;
c. estime o valor de **Y** para X = 0.

7. Certa empresa, estudando a variação da demanda de seu produto em relação à variação de preço de venda, obteve a tabela:

PREÇO (x_i)	38	42	50	56	59	63	70	80	95	110
DEMANDA (y_i)	350	325	297	270	256	246	238	223	215	208

a. Determine o coeficiente de correlação.
b. Estabeleça a equação da reta ajustada.
c. Estime **Y** para X = 60 e X = 120.

8. Pretendendo-se estudar a relação entre as variáveis "consumo de energia elétrica" (x_i) e "volume de produção nas empresas industriais" (y_i), fez-se uma amostragem que inclui vinte empresas, computando-se os seguintes valores:

$$\sum x_i = 11{,}34,\ \sum y_i = 20{,}72,\ \sum x_i^2 = 12{,}16,\ \sum y_i^2 = 84{,}96 \text{ e } \sum x_i y_i = 22{,}13$$

Determine:

a. o cálculo do coeficiente de correlação;
b. a equação de regressão de **Y** para **X**;
c. a equação de regressão de **X** para **Y**.