Correlação e Regressão

Teoria

Introdução

Em muitas áreas de aplicação, o investigador pretende encontrar um modelo matemático que relacione as variáveis em estudo.

Os modelos matemáticos permitem, por exemplo, prever o tamanho de uma população num determinado instante de tempo.

Os modelos matemáticos são geralmente obtidos a partir de experiências ou observações.

Objectivo

Estudar a relação linear entre duas variáveis quantitativas.

Exemplos:

- Idade e altura das crianças
- Duração da prática de desporto e ritmo cardíaco
- Tempo de estudo e nota na prova
- Taxa de desemprego e taxa de criminalidade
- Expectativa de vida e taxa de analfabetismo

Objectivo

Investigar a presença ou ausência de relação linear sob dois pontos de vista:

a) Quantificando a força dessa relação: **correlação**

b) Explicitando a forma dessa relação: **regressão**

Representação gráfica de duas variáveis quantitativas: Diagrama de dispersão

Diagrama de dispersão

O **diagrama de dispersão** é o conjunto dos pontos do tipo (x, y) representados num referencial, onde x e y são os valores das variáveis X e Y, respectivamente, para cada uma das observações.

Diagrama de dispersão

Exemplo

Considere as variáveis altura (m) e peso (kg)

Altura (X)	*Peso (Y)*
1,75	76
1,69	71
1,85	90
1,8	81
1,7	70
1,74	73
1,87	76
1,68	65
1,76	70
1,95	92
1,96	90
1,62	55

Desenhe o diagrama de dispersão

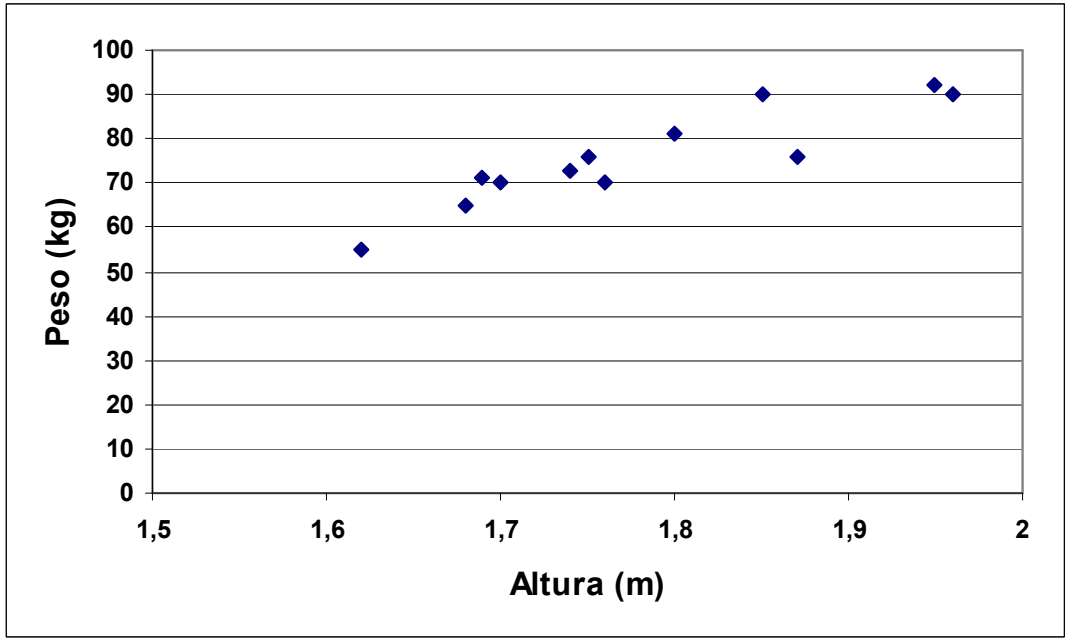


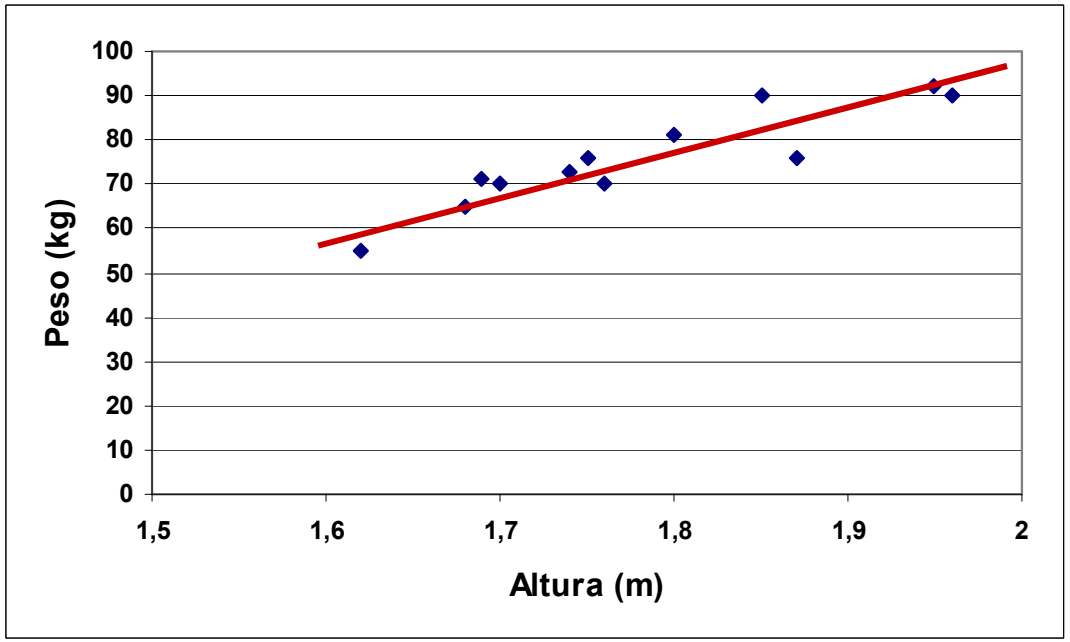
Diagrama de dispersão

Portanto este diagrama permite decidir empiricamente se um relacionamento linear entre X e Y deve ser assumido.

Por análise do diagrama de dispersão pode-se também concluir (empiricamente) se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através do enxame de pontos.

Diagrama de dispersão

Do exemplo anterior, podemos considerar uma reta que passa entre os pontos



Correlação

Definição:

O coeficiente de correlação é uma medida da intensidade ou do grau de associação linear entre as variáveis analisadas. Foi estabelecido por Karl Pearson, em 1896. Por essa razão é denominado **Coeficiente de Correlação de Pearson**. É dado por:

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

Correlação

Propriedade: $-1 \leq r \leq 1$

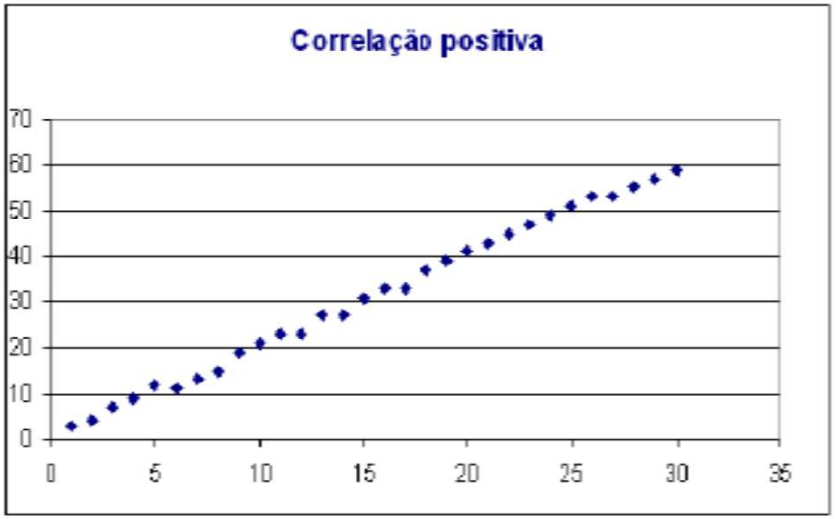
Casos particulares:

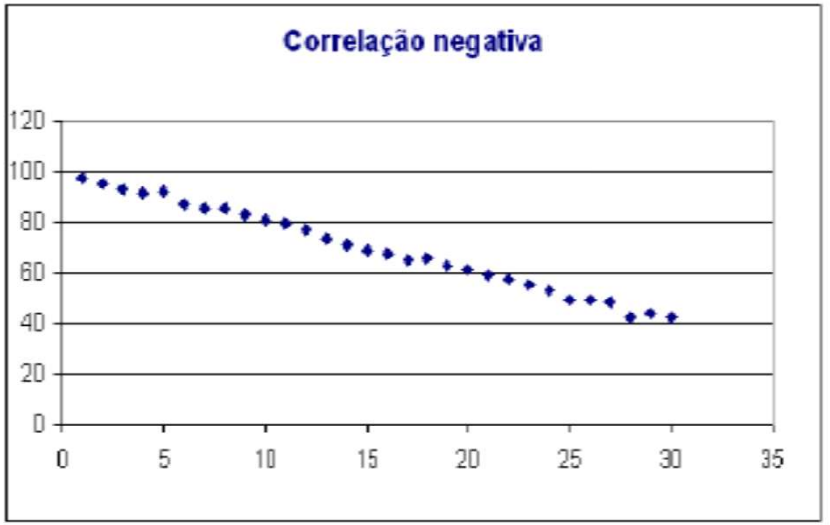
$r = 1 \Rightarrow$ correlação linear positiva e perfeita

$r = -1 \Rightarrow$ correlação linear negativa e perfeita

$r = 0 \Rightarrow$ inexistência de correlação linear

Correlação





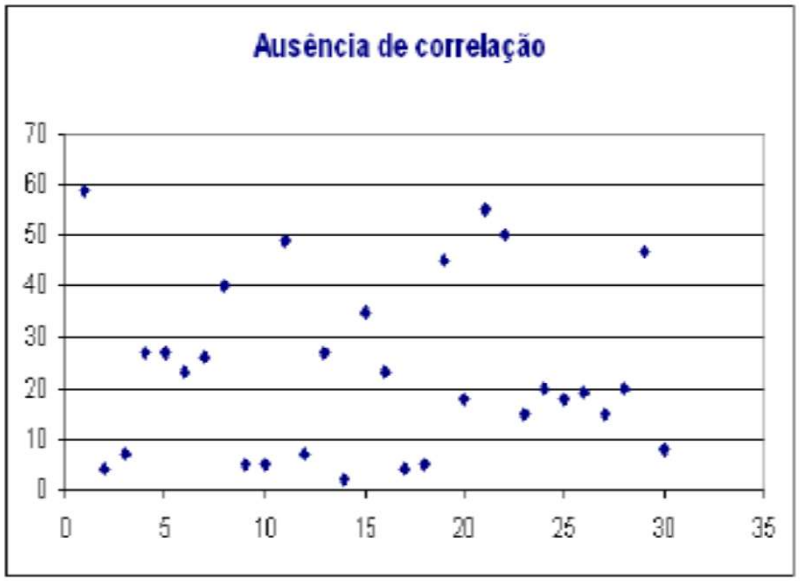
$r \cong 1$

$r \cong -1$

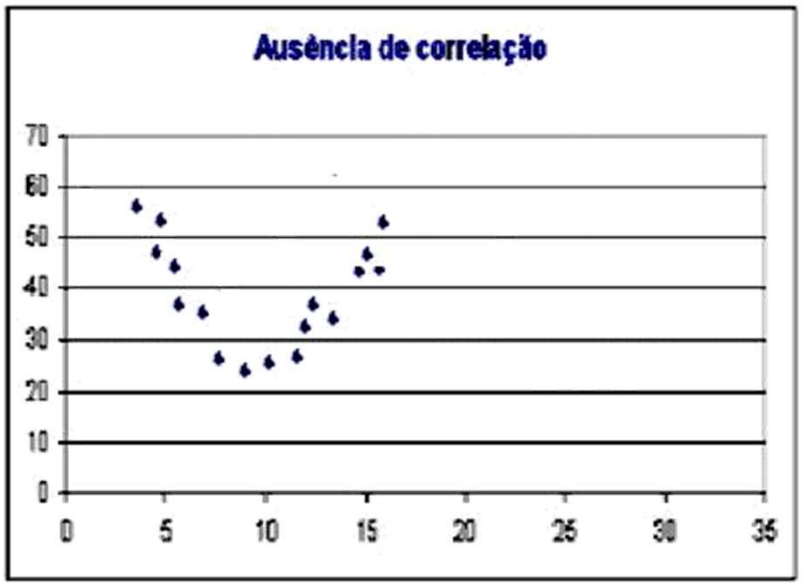
Quanto mais próximo **r** estiver de 1 ou -1 mais forte será a associação linear entre as variáveis

Correlação

Uma correlação igual a zero, ou próxima de zero, não significa que as duas variáveis não estão relacionadas, mas apenas que não existe uma relação linear entre elas, por exemplo



Não existe relação



Existe relação quadrática

Correlação

| r |	A correlação diz-se
0	nula
0 – 0,3	fraca
0,3 – 0,6	regular
0,6 – 0,9	forte
0,9 – 1	muito forte
1	perfeita

Correlação

Definição:

O **coeficiente de determinação** (r^2), é uma medida da proporção da variação de uma característica que é explicada estatisticamente pela outra variável.

Note que:

- $0 \leq r^2 \leq 1$;
- $r^2 \cong 1$ (próximo de 1) significa que grande parte da variação de uma variável é explicada linearmente pela variação da outra variável;
- $r^2 \cong 0$ (próximo de 0) significa que grande parte da variação de uma variável não é explicada linearmente pela variação da outra variável;

Correlação

Exemplo

Fez-se um estudo acerca da idade gestacional (semanas) e do peso ao nascer (kg) de 7 bebés. Os resultados obtidos estão representados na tabela:

Idade gestacional	28	32	35	38	39	41	42
Peso ao nascer	1,25	1,25	1,75	2,25	3,25	3,25	4,25

Verifique se existe uma relação linear entre estas variáveis. Em caso afirmativo, qual a proporção da variabilidade no peso do bebé que pode ser explicada pela idade gestacional?

Correlação

Exemplo (cont.)

Considerem-se as variáveis

X = idade gestacional em semanas

Y = peso ao nascer em kg

Pretende-se calcular ***r*** para verificar se existe uma relação linear,

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

Correlação

Exemplo (cont.)

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
28	1,25			
32	1,25			
35	1,75			
38	2,25			
39	3,25			
41	3,25			
42	4,25			
Σ				

Correlação

Exemplo (cont.)

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	28	1,25	35	784	1,5625
	32	1,25	40	1024	1,5625
	35	1,75	61,25	1225	3,0625
	38	2,25	85,5	1444	5,0625
	39	3,25	126,75	1521	10,5625
	41	3,25	133,25	1681	10,5625
	42	4,25	178,5	1764	18,0625
$\sum$	**255**	**17,25**	**660,25**	**9443**	**50,4375**

Correlação

Exemplo (cont.)

Substituindo em *r* vem

$$r = \frac{7 \times 660{,}25 - 255 \times 17{,}25}{\sqrt{\left[7 \times 9443 - 255^2\right]\left[7 \times 50{,}4375 - 17{,}25^2\right]}}$$

$$= \frac{4621{,}75 - 4398{,}75}{\sqrt{\left[66101 - 65025\right]\left[353{,}0625 - 297{,}5625\right]}} = \frac{223}{\sqrt{1076 \times 55{,}5}}$$

$$= \frac{223}{\sqrt{59718}} = \frac{223}{244{,}37} \cong 0{,}91$$

Como *r* está próximo de 1, pode-se concluir que existe uma relação linear muito forte entre as variáveis

Correlação

Exemplo (cont.)

Calculando o **Coeficiente de Determinação r^2** vem

$$r^2 = 0{,}91^2 \cong 0{,}83$$

então pode-se afirmar que aproximadamente 83% da variabilidade do peso ao nascer, de um bebé, pode ser explicada pelo número de semanas de gestação.

Os restantes 17% serão explicados por outros factores.

Regressão linear

A análise de regressão estuda o relacionamento entre uma variável chamada **variável dependente** e outras variáveis chamadas **variáveis independentes**.

Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.

Este modelo é designado por **modelo de regressão linear simples** se define uma relação linear entre a variável dependente e uma variável independente.

Regressão linear

Graficamente o modelo de regressão linear simples é apresentado como a reta que melhor aproxima a relação linear entre a variável dependente e a variável independente.

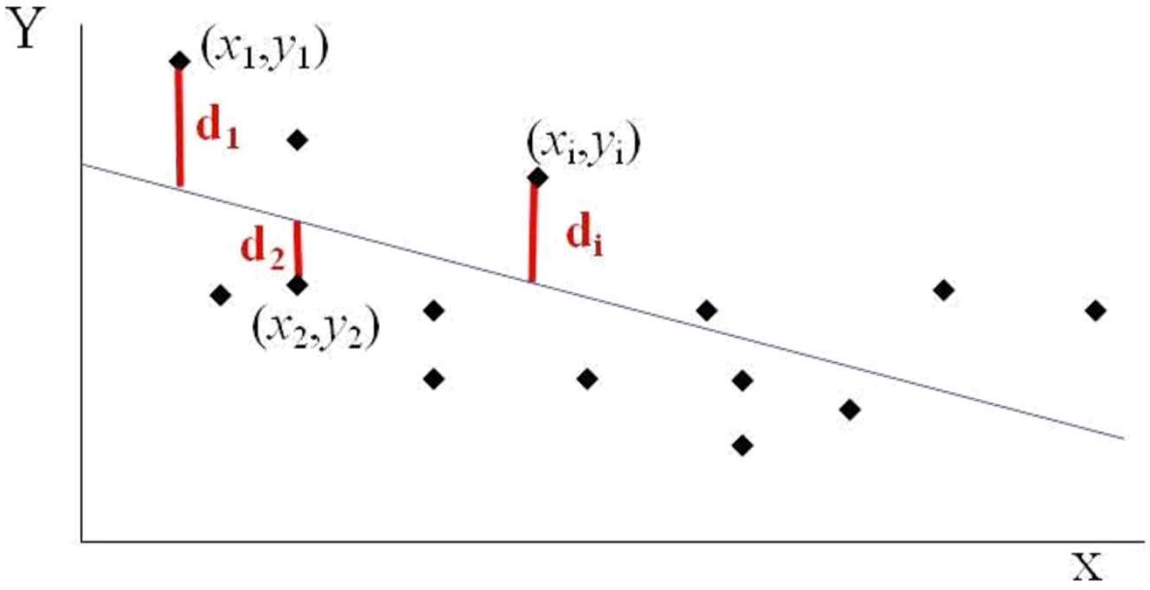
A representação matemática do modelo é então, a equação da reta

$$Y = a X + b$$

onde Y representa a variável dependente e X representa a variável independente.

Regressão linear

Os parâmetros **a** e **b** da reta são determinados através do Método dos Mínimos Quadrados, que minimiza o somatório dos quadrados dos resíduos (diferença entre o valor observado da variável dependente e o previsto pela recta de regressão)



Regressão linear

Obtêm-se assim as expressões para os parâmetros

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

e

$$b = \overline{Y} - a\,\overline{X} = \frac{\sum_{i=1}^{n} y_i}{n} - a\,\frac{\sum_{i=1}^{n} x_i}{n}$$

Regressão linear

Exemplo:

Considerando o exemplo que relaciona o peso ao nascer com a idade gestacional, definir a reta de regressão.

A equação da reta é

$$Y = a X + b$$

onde Y representa o peso e X a idade gestacional

Regressão linear

Exemplo (cont.)

Assim

$$a = \frac{7 \times 660{,}25 - 255 \times 17{,}25}{7 \times 9443 - 255^2} = \frac{223}{1076} \cong 0{,}21$$

e

$$b = \frac{17{,}25}{7} - 0{,}21 \times \frac{255}{7} = 2{,}46 - 0{,}21 \times 36{,}4 \cong -5{,}18$$

A equação da reta é

$$Y = 0{,}21X - 5{,}18$$

Regressão linear

Exemplo (cont.)

Podemos então estimar, por exemplo, o peso de um bebé com 40 semanas.

Substituindo na equação X por 40 vem

$$Y = 0{,}21 \times 40 - 5{,}18 = 3{,}22$$

ou seja, um bebé com 40 semanas de gestação terá um peso aproximado de 3,22 kg