

Ficha — Dados bivariados: correlação e reta de regressão

10.º Ano MACS 4 páginas 12 exercícios

Objetivo: Representar a associação entre duas variáveis num diagrama de dispersão, interpretar o coeficiente de correlação linear r , usar a reta de regressão para inferir valores e distinguir correlação de causalidade, reconhecendo a influência dos outliers.

Revisão rápida

Dois passos, duas variáveis — para estudar *associação* entre variáveis, o gráfico é o **diagrama de dispersão** (pares (x, y) como pontos):

- x = variável **independente** (explicativa, eixo horizontal) · y = variável **dependente** (resposta, vertical) — a escolha vem da pergunta.
- Leitura: pontos próximos de uma **reta crescente** → associação positiva · **decrecente** → negativa · «nuvem» sem padrão → sem associação linear.

Coeficiente de correlação linear r :

Propriedade	Significado
$-1 \leq r \leq 1$	é normalizado — compara-se entre amostras
sinal de r	direção: + crescente · – decrescente
$ r $	força: perto de 1 = forte (pontos colados à reta) · perto de 0 = fraca (nuvem)

Reta de regressão — a reta que **melhor ajusta** os pontos: $\hat{y} = a + b \cdot x$, onde $b = r \cdot (s_y/s_x)$ (pendente) e $a = \bar{y} - b \cdot \bar{x}$ (interceção). Serve para **inferir**: dado um valor de x , estima-se $\hat{y}$.

Três advertências: (1) **correlação $\neq$ causalidade** — as duas podem ser causadas por um terceiro fator (confundidor: o calor explica gelados e afogamentos); (2) **outliers** puxam r e a reta — analisa-se com e sem eles; (3) a reta **não extrapola** com segurança para longe dos dados observados.

Estratégia: definir x e y → desenhar a dispersão → medir r (sinal, força) → ajustar a reta → inferir com cautela → perguntar sempre: «e quem provoca o quê?».

Exercícios

1. Completar: para estudar a associação entre duas variáveis usa-se o diagrama de _____; o eixo x representa a variável _____ e o eixo y a variável _____.
2. Liga cada descrição ao padrão: (i) pontos colados a uma reta de baixo para cima; (ii) nuvem sem qualquer padrão; (iii) pontos próximos de uma reta decrescente; (iv) pontos em forma de parábola. Opções: correlação positiva forte · sem correlação linear · correlação negativa · associação não linear.
3. Completar as propriedades de r : $-1 \leq r \leq 1$ · o sinal de r indica a _____ da associação · o valor absoluto indica a _____ · $r = 0,92$ representa uma associação _____ e $r = 0,15$ uma _____.

4. Interpreta cada r e inventa o contexto que o justifica: (a) $r = 0,92$ · (b) $r = -0,75$ · (c) $r = 0,15$ · (d) $r = -0,99$. (Pistas: horas de estudo × nota, horas de ecrã × nota, ...)

5. Reta dada: $\hat{y} = 2,5x + 10$. (a) Interpreta o 2,5 (pendente) e o 10 (interceção). (b) Calcula $\hat{y}$ para $x = 4$ e para $x = 10$. (c) Porque se recomenda cautela ao usar a reta fora do intervalo dos dados?

6. Determinar a reta a partir das estatísticas: $\bar{x} = 5$, $\bar{y} = 20$, $s_x = 2$, $s_y = 4$, $r = 0,8$. Calcula $b = r(s_y/s_x)$, depois $a = \bar{y} - b\bar{x}$, escreve $\hat{y}$ e infere y para $x = 6$.

7. Correlação e causalidade: «Verifica-se $r \approx 0,9$ entre vendas de gelados e afogamentos nas praias.» Explica porque isto não prova que gelados causam afogamentos, identifica o fator comum e dá um segundo exemplo do mesmo tipo.

8. Outliers: um ponto distante da nuvem principal. (a) Para que valores de r o puxa — mais próximos de ± 1 ou de 0? Depende do alinhamento. (b) Porque é que se recomenda analisar os dados com e sem o outlier antes de concluir?

9. Verdadeiro ou falso. Corrige as falsas.

- a) $r = -0,9$ indica uma forte associação negativa. (V/F)
- b) $r = 0$ significa que não existe qualquer associação entre as variáveis. (V/F)
- c) A reta de regressão serve para estimar y dado um valor de x . (V/F)
- d) Se a correlação é elevada, a causalidade está provada. (V/F)
- e) A pendente b da reta mede a variação de $\hat{y}$ por unidade de x . (V/F)

10. Aplicação integrada: num estudo, horas estudadas por semana (x) e nota no teste (y) dão $r = 0,85$ e $\hat{y} = 1,2x + 8$. (a) Prever a nota de quem estuda 7 h. (b) O que significam os números 1,2 e 8. (c) Porque é que a previsão não é uma garantia para um aluno individual?

11. Tecnologia: descreve os passos, numa folha de cálculo, para obter o diagrama de dispersão, o coeficiente r e a reta de regressão a partir de duas colunas de dados — incluindo que opção se marca para mostrar a reta e a equação, e um cuidado na leitura.

12. Síntese: um anúncio afirma «quem usa este plano de estudos tem notas mais altas — prova: $r = 0,9$ ». Analisa criticamente a afirmação (correlação, causalidade, variáveis em falta) e escreve o que seria necessário para sustentar a conclusão.

Soluções — Dados bivariados: correlação e regressão

10.º Ano

MACS

Resolução completa

1. dispersão · independente (explicativa) · dependente (resposta).

2. (i) **correlação positiva forte** · (ii) **sem correlação linear** · (iii) **correlação negativa** · (iv) **associação não linear** (a reta não a descreve — por isso $r \approx 0$ não significa «sem relação»).

3. direção · força · positiva **forte** · positiva **fraca** (quase sem padrão linear).

4. (a) $r = 0,92 \rightarrow$ positiva e muito forte: ex.: **horas de estudo** $\times$ **nota** (quem estuda mais, tende a tirar notas mais altas, pontos colados à reta). (b) $r = -0,75 \rightarrow$ negativa forte: **horas de ecrã** $\times$ **nota** (mais ecrã, notas mais baixas). (c) $r = 0,15 \rightarrow$ positiva mas muito fraca: ex.: **n.º de sapatos** $\times$ **nota** — praticamente sem relação linear. (d) $r = -0,99 \rightarrow$ negativa quase perfeita: ex.: minutos restantes de bateria $\times$ tempo de utilização — quase determinística.

5. (a) Pendente **2,5**: cada unidade que x cresce, $\hat{y}$ cresce em média **2,5**. Interceção **10**: valor estimado de y quando $x = 0$ (muitas vezes fora do sentido prático — atenção à interpretação). (b) $x = 4 \rightarrow \hat{y} = 2,5 \times 4 + 10 = 20$; $x = 10 \rightarrow \hat{y} = 2,5 \times 10 + 10 = 35$. (c) A reta foi ajustada aos pontos observados — extrapolar é assumir que o padrão continua linear onde não há dados, o que quase nunca se confirma (e pode dar valores absurdos).

6. $b = 0,8 \times (4 \div 2) = 0,8 \times 2 = 1,6$; $a = 20 - 1,6 \times 5 = 20 - 8 = 12 \rightarrow \hat{y} = 1,6x + 12$. Para $x = 6$: $\hat{y} = 1,6 \times 6 + 12 = 9,6 + 12 = 21,6$.

7. A correlação alta não prova causalidade: as duas variáveis são causadas por um **terceiro fator** — o **calor** faz comprar mais gelados e faz mais gente entrar na água (mais afogamentos). Retirar os gelados não mudaria os afogamentos. Segundo exemplo: número de fogos e número de bombeiros no local correlacionam-se porque **ambos** são causados pelos fogos — os bombeiros não causam fogos. Para provar causalidade exige-se mecanismo plausível, controlar confundidores ou **experimentação** (grupo de controlo).

8. (a) Depende: se o outlier se alinha com a nuvem, **fortalece** artificialmente $|r|$ (puxa para ± 1); se está fora de alinhamento, **arruína** a correlação (puxa para 0). (b) Porque um único ponto pode **criar** ou **destruir** uma conclusão: analisar com e sem ele mostra se o padrão pertence aos dados ou ao acidente — se a conclusão muda, o outlier tem de ser justificado (erro de registo? caso excepcional?) antes de qualquer decisão.

9. a) **V** — o sinal dá a direção, 0,9 a força. b) **F** — só significa **sem associação linear** (pode haver não linear, ver ex. 2-iv). c) **V** — é a sua função ($\hat{y}$ estimado). d) **F** — «correlação não implica causalidade». e) **V** — é a definição de pendente: $\Delta\hat{y}$ por unidade de x .

10. (a) $\hat{y} = 1,2 \times 7 + 8 = 8,4 + 8 = 16,4$ (≈ 16 valores). (b) **1,2** = cada hora extra de estudo acrescenta em média 1,2 valores à nota; **8** = nota estimada para quem estuda 0 h (interceção — descreve o padrão, não um indivíduo real). (c) A reta descreve a **tendência da amostra**: $r = 0,85 < 1$ deixa dispersão — alunos iguais em x têm y diferentes (talento, método, sono); por isso a previsão é uma **estimativa com erro**, não uma garantia — e há limites naturais ($\text{nota} \leq 20 \rightarrow x \leq 10$ h).

11. Passos: (1) dispor os pares em duas colunas (x , y); (2) selecionar as duas colunas e **inserir gráfico de dispersão**; (3) no gráfico, ativar **tendência linear** e mostrar a **equação** e o **R^2** (o R^2 é r^2 — serve para ver a força); (4) ler r (raiz de R^2 com o sinal da pendente) e os coeficientes a , b ; (5) usar a equação para inferir (substituir x). Cuidado: a equação vale para o intervalo dos dados; e R^2 alto não prova causalidade.

12. Resposta possível: o anúncio comete três erros. (1) **Correlação $\neq$ causalidade**: $r = 0,9$ só mostra associação — talvez quem usa o plano já estude mais (a variável em falta: motivação/horas). (2) **Variáveis em falta**: o estudo não controlou tempo de estudo, escola ou apoio — confundidores. (3) **Generalização**: não se diz de quem é a amostra. Para sustentar a conclusão seria preciso **experimentação** (grupo com o plano e grupo de controlo, sorteados, mesmas condições) ou, no mínimo, um estudo que controle as variáveis em falta — e ainda assim a conclusão seria «causa», com intervalo de dúvida.

Nota: o percurso analítico é sempre o mesmo — **dispersão $\rightarrow r \rightarrow$ reta $\rightarrow$ previsão $\rightarrow$ advertência causal**. E a frase que resolve metade das questões: *correlação forte é suspeita, não é prova*.