PARTE IV
CORRELAÇÃO E REGRESSÃO

Relação entre duas variáveis

Nos problemas de duas amostras discutidos anteriormente, concentrámo-nos na comparação de valores de parâmetros das distribuições de duas variáveis x e y.
Vamos considerar agora com mais cuidado a possível relação entre duas variáveis.
Por exemplo, é natural pensar que a altura e o peso de um indivíduo estão relacionados ou que o preço de um determinado produto (por exemplo, vinho) e o montante da colheita também estão relacionados.

Quando se fala de uma relação determinística, fala-se numa correspondência biunívoca entre duas variáveis.
Por exemplo, o perímetro P de uma circunferência e o raio r da mesma circunferência estão relacionados; a relação que liga essas duas variáveis é *definida e inalterável* e pode expressar-se pela seguinte fórmula: $P = 2\pi r$.

Relação estatística

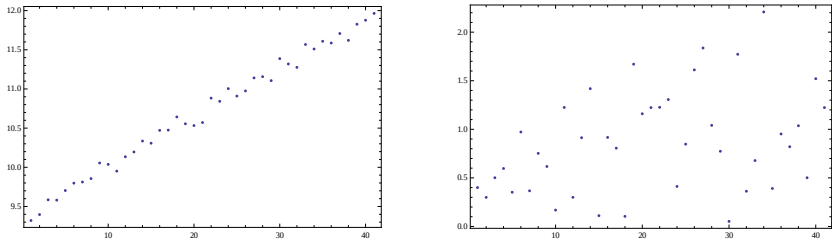
Quando falamos numa relação estatística, como por exemplo a relação entre o peso e altura de um indivíduo, pode suceder (e sucede) que indivíduos com a mesma altura tenham pesos diferentes, mas, em média, quanto maior é altura de um indivíduo, maior é o seu peso; no caso do preço do vinho, em média, quanto maior é a colheita, menor é o preço.

Assim, *uma relação estatística entre duas variáveis ocupa-se da variação em média*. Os fenómenos não estão ligados de forma determinística, mas a intensidade de um é acompanhada pela intensidade do outro no mesmo sentido (relação positiva) ou no sentido inverso (relação negativa).

Diagrama de dispersão

Vamos, então, dedicar-nos ao caso em que temos uma amostra de dimensão n constituída por pares de observações $(x_k, y_k); k = 1, 2, \ldots, n$, em que a primeira entrada do par é relativa à medição de uma variável x no indivíduo k e a segunda entrada é relativa à medição de uma variável y no mesmo indivíduo k.
Vamos supor que as variáveis em causa são quantitativas e expressas em escala (no mínimo) intervalar.

Para uma primeira indicação do tipo de possível associação entre as variáveis, é conveniente elaborar o chamado diagrama de dispersão, o qual é, simplesmente, a representação gráfica dos pontos (x_k, y_k) (como pontos de um plano).



No caso de os pontos do diagrama de dispersão tenderem a colocar-se aproximadamente sobre uma recta (como no gráfico da esquerda), dizemos que as variáveis estão linearmente correlacionadas.
Para medir, numericamente, o grau de correlação linear entre duas variáveis podemos usar uma estatística, conhecida por coeficiente de correlação amostral de Pearson, o qual é definido por

$$r = \frac{1}{n-1}\sum_{k=1}^{n}\frac{x_k-\bar{x}}{s_x}\frac{y_k-\bar{y}}{s_y},$$

onde s_x e s_y representam os desvios padrões amostrais das amostras provenientes das variáveis x e y, respetivamente.

Note-se que permutando as duas amostras, isto é, considerando a amostra emparelhada (y_k, x_k), o valor do coeficiente de correlação mantém-se inalterado. Também se mostra facilmente que este coeficiente é invariante para mudanças de localização e escala dos dados. Pode provar-se que o coeficiente de correlação r satisfaz as seguintes propriedades:

- $-1 \leq r \leq 1$;
- $r = \pm 1$ se e só se os n pontos (x_k, y_k) estiverem sobre uma recta;
- se as variáveis não estiverem relacionadas, então $r = 0$;
- se $r = 0$, então não existe relação linear entre as variáveis (podendo, no entanto, existir uma relação não linear entre as variáveis).

Em resumo, podemos dizer que r mede o grau de relação linear entre as duas variáveis. Quanto mais próximo de 1 estiver $|r|$, mais forte é a associação linear entre as variáveis. Se as variáveis não aparentam qualquer padrão ou, havendo padrão, este não for linear, então $r \approx 0$.

Quando pelo menos uma das variáveis está apenas em escala ordinal, utiliza-se o chamado coeficiente de correlação ordinal de Spearman, o qual coincide com o coeficiente de correlação de Pearson, mas aplicado às ordens dos dados.
Assim, cada par (x_k, y_k) é substituído pelo par $((x_k), (y_k))$, onde (x_k) representa a ordem da observação x_k na colecção dos dados (com significado análogo para (y_k)) e calcula-se o respetivo coeficiente de correlação de Pearson. Pode mostrar-se que tal equivale ao uso da seguinte fórmula

$$r_S = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)},$$

onde $d_k = (x_k) - (y_k)$.

$$r_S = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)}, \quad d_k = (x_k) - (y_k).$$

Pode provar-se que, tal como para o coeficiente r anterior, também $-1 \leq r_S \leq 1$.
Além disso, se a ordenação for totalmente concordante, teremos $d_k = 0$, para $k = 1, \ldots, n$ donde virá $r_S = 1$.
Se a ordenação for totalmente discordante (se uma ordenação for inversa da outra), pode mostrar-se que será $\sum_{k=1}^{n} d_k^2 = \frac{n(n^2-1)}{3}$, donde virá

$$r_S = 1 - \frac{6 \times \frac{n(n^2-1)}{3}}{n(n^2-1)} = 1 - 2 = -1.$$

Exemplo

Pediu-se a ambos os membros de um casal que ordenassem 10 determinados fatores, na educação dos filhos, do mais importante (10) para o menos importante (1). Os dados recolhidos estão apresentados na tabela seguinte:

Fator	1	2	3	4	5	6	7	8	9	10
Ord. Marido	6	3	1	7	2	8	5	9	5	10
Ord. Mulher	6	3	2	9	1	7	5	8	4	10

Calculemos o coeficiente de correlação de Spearman relativo a estes dados.
Como os dados estão já na forma de ordens, basta aplicar-lhes diretamente a fórmula $r_S = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)}$.
Neste caso, tem-se $d_1 = 0$, $d_2 = 0$, $d_3 = -1$, $d_4 = -2$, $d_5 = 1$, $d_6 = 1$, $d_7 = 0$, $d_8 = 1$, $d_9 = 1$ e $d_{10} = 0$, vindo, então

$$r_S = 1 - \frac{6 \times (1+4+1+1+1+1)}{10 \times 99} = 1 - \frac{54}{990} = 0.94545.$$

Como r_S está bastante próximo de 1, podemos concluir que existe uma grande concordância entre o casal sobre quais os fatores mais relevantes na educação dos seus filhos.

Vimos como um diagrama de dispersão é um processo gráfico para detetar, visualmente, relações entre dados bivariados.
Vimos também como o coeficiente de correlação pode ser usado para medir a associação linear entre duas variáveis.
A análise de regressão, que estudaremos agora, fornece-nos ferramentas para descrever, numericamente, relações entre variáveis, de modo a permitir fazer previsões.
Ao contrário da correlação, na regressão há que distinguir entre a variável resposta ou dependente - aleatória - e a variável independente ou variável preditora, em muitas situações, controlada pelo experimentador - (em princípio) não aleatória, que supomos medida sem erro.
Como exemplo, consideremos uma experiência laboratorial em que são administradas doses x_k (escolhidas, e portanto, controladas e não aleatórias) de um certo medicamento, em diversos animais, e se medem determinadas respostas y_k da administração dessas doses de medicamento.
Neste caso, não faz sentido trocar os papéis de x e y.

Regressão linear simples

O diagrama de dispersão pode evidenciar alguma relação funcional entre x e y (y como função de x). Quando tentamos descrever essa relação funcional, isto é, quando ajustamos um modelo $\hat{y} = f(x)$, falamos de regressão de y em x (ou de y sobre x).
No caso da regressão linear simples pressupõe-se haver uma relação de linearidade entre as variáveis x e y.Existe assim um modelo da forma

$$\hat{y} = a + bx,$$

considerando-se que os valores observados y_k são flutuações amostrais (com erro) em torno dos valores fornecidos pelo modelo – chamados valores previstos ou valores ajustados de y – isto é dos valores

$$\hat{y}_k = a + bx_k.$$

O erro de predição, desvio ou resíduo e_k correspondente à k-ésima observação é a diferença entre o valor observado y_k e o valor previsto $\hat{y}_k$, isto é, é dado por

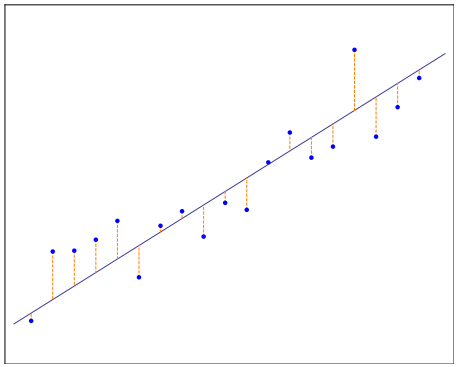
$$e_k = y_k - \hat{y}_k.$$

A soma dos quadrados dos desvios, vulgarmente designada por SSE (do inglês, "Sum of Squares due to Error") é, assim, dada por

$$SSE = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 = \sum_{k=1}^{n} (y_k - a - bx_k)^2.$$

Um dos critérios mais usados para encontrar a recta $\hat{y} = a + bx$ que "melhor"se ajusta aos dados é usar o chamado processo dos mínimos quadrados: neste caso, os valores dos parâmetros a e b que definem a recta $\hat{y} = a + bx$ são determinados de forma a que seja minimizada a soma dos quadrados dos desvios SSE.

Note-se que o valor absoluto do resíduo $|e_k| = |y_k - \hat{y}_k|$ é a distância entre o ponto (x_k, y_k) e o ponto, $(x_k, \hat{y}_k)$, isto é, o ponto alinhado com este, na vertical, mas situado sobre a recta.



Assim, ao minimizarmos a soma total dos quadrados dos desvios, estaremos a minimizar a soma total dos quadrados das distâncias, medidas na vertical, dos pontos (x_k, y_k) à recta $a + bx$.

Pode mostrar-se que os valores de a e b para os quais a expressão

$$SSE(a,b) = \sum_{k=1}^{n}(y_k - a - bx_k)^2$$

é mínima são dados por:

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

$$b = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2} = \frac{\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}}{(n-1)s_x^2}.$$

A recta $\hat{y} = a + bx$ assim obtida é chamada recta de regressão de y em x.

Facilmente se verificam as seguintes propriedades.

RR1 A recta de regressão passa pelo ponto $(\bar{x}, \bar{y})$, isto é, tem-se

$$\bar{y} = a + b\bar{x} \qquad (1)$$

É imediato, já que que $a = \bar{y} - b\bar{x}$

RR2 O valor de b pode ser obtido pela fórmula

$$b = r\frac{s_y}{s_x},$$

onde r é o coeficiente de correlação de Pearson para os dados (x_k, y_k): $r = \frac{1}{n-1}\sum_{k=1}^{n}\frac{x_k-\bar{x}}{s_x}\frac{y_k-\bar{y}}{s_y}$.

$$r\frac{s_y}{s_x} = \left(\frac{1}{n-1}\sum_{k=1}^{n}\frac{x_k-\bar{x}}{s_x}\frac{y_k-\bar{y}}{s_y}\right)\frac{s_y}{s_x}$$

$$= \sum_{k=1}^{n}\frac{(x_k-\bar{x})(y_k-\bar{y})}{(n-1)s_x^2}$$

$$= \frac{\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})}{\sum_{k=1}^{n}(x_k-\bar{x})^2} = b$$

RR3

$$\sum_{k=1}^{n} e_k = \sum_{k=1}^{n}(y_k - \hat{y}_k) = 0.$$

$$\sum_{k=1}^{n} e_k = \sum_{k=1}^{n}(y_k - a - bx_k) = \sum_{k=1}^{n} y_k - na - b\sum_{k=1}^{n} x_k$$
$$= n\bar{y} - na - nb\bar{x} = n(\bar{y} - a - \bar{x}) = 0$$

RR4

$$\sum_{k=1}^{n} y_k = \sum_{k=1}^{n} \hat{y}_k.$$

$$\sum_{k=1}^{n}(y_k - \hat{y}_k) = 0 \Rightarrow \sum_{k=1}^{n} y_k - \sum_{k=1}^{n}\hat{y}_k = 0 \Rightarrow \sum_{k=1}^{n} y_k = \sum_{k=1}^{n}\hat{y}_k.$$

Qualidade do ajustamento

Tendo encontrado a recta de regressão para os dados, a próxima questão que, naturalmente se levanta, é a de tentar aferir a qualidade do ajustamento obtido.
Uma vez que os resíduos medem a discrepância entre a recta e os dados observados, um gráfico de dispersão dos valores (x_k, e_k) pode ajudar a pôr em evidência as eventuais deficiências de considerarmos $a + bx$ como modelo para a relação entre y e x.

A inferência estatística baseada neste modelo (i.e. por exemplo, quando usamos a recta $\hat{y} = a + bx$ para estimar um valor da variável y para um certo valor de x diferente dos x_k) assenta no pressuposto de que os erros de ajustamento têm um comportamento aleatório normal, com valor médio nulo, não estão correlacionados e têm variância constante.

Por isso, na representação gráfica dos resíduos:

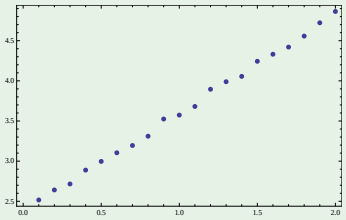
- não devem existir padrões ou tendências, devendo a distribuição dos pontos no plano ter um aspecto aleatório;
- os pontos devem estar dispostos (ao acaso) numa banda horizontal (uma vez que se espera variância constante para os desvios) centrada no eixo dos xx (uma vez que se espera uma média nula para os desvios).

Exemplo

Considere-se a seguinte amostra de pares (x_k, y_k):

$(0.1, 2.51649), (0.2, 2.64119), (0.3, 2.7158), (0.4, 2.8884), (0.5, 2.99668),$
$(0.6, 3.10415), (0.7, 3.19486), (0.8, 3.31053), (0.9, 3.52461), (1., 3.57375),$
$(1.1, 3.68104), (1.2, 3.89518), (1.3, 3.98911), (1.4, 4.05582), (1.5, 4.24346),$
$(1.6, 4.33153), (1.7, 4.42073), (1.8, 4.55742), (1.9, 4.72277), (2., 4.86406).$

Na figura seguinte apresenta-se o diagrama de dispersão dos dados, o qual evidencia uma relação linear entre x e y.

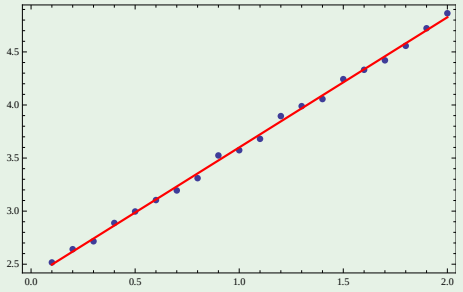


Exemplo (cont.)

Neste caso, calculando a recta de regressão $\hat{y} = a + bx$, obtém-se

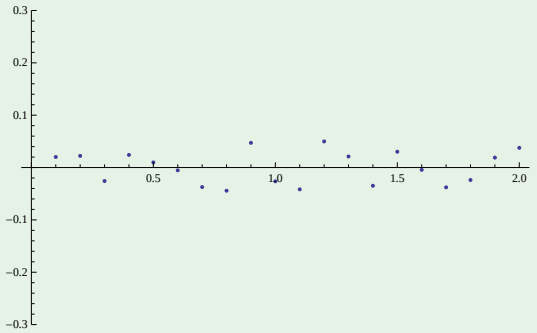
$$\hat{y} = 2.37366 + 1.2264x.$$

Na figura seguinte apresentam-se novamente os pontos (x_k, y_k), sobrepondo, no mesmo gráfico, a recta de regressão acima obtida, sendo patente o "bom" ajustamento da recta aos pontos considerados.



Exemplo (cont.)

Na figura seguinte estão representados os pontos (x_k, e_k) onde $e_k = y_k - \hat{y}_k$.



Pode provar-se facilmente que $\sum_{k=1}^{n}(y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) = 0$. Assim, temos

$$\begin{aligned}\sum_{k=1}^{n}(y_k - \bar{y})^2 &= \sum_{k=1}^{n}(y_k - \hat{y}_k + \hat{y}_k - \bar{y})^2 \\ &= \sum_{k=1}^{n}(y_k - \hat{y}_k)^2 + \sum_{k=1}^{n}(\hat{y}_k - \bar{y})^2 + 2\sum_{k=1}^{n}(y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) \\ &= \sum_{k=1}^{n}(\hat{y}_k - \bar{y})^2 + \sum_{k=1}^{n}(y_k - \hat{y}_k)^2\end{aligned}$$

A equação anterior costuma escrever-se como

$$SST = SSA + SSE$$

em que SST, SSA e SSE representam, respectivamente:

- a soma de quadrados total $\rightarrow \sum_{k=1}^{n}(y_k - \bar{y})^2$
- a soma dos quadrados devida ao ajustamento $\rightarrow \sum_{k=1}^{n}(\hat{y}_k - \bar{y})^2$
- e a soma de quadrados devida ao erro $\rightarrow \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$

Então, tem-se

- ajuste perfeito (relação linear perfeita) $\Rightarrow SSE = 0 \Rightarrow$

$$\frac{SSE}{SST} = 0 \quad \text{e} \quad 1 - \frac{SSE}{SST} = 1$$

- ajustamento totalmente desadequado (ausência total de relação linear) $\Rightarrow SSA = 0 \Rightarrow$

$$\frac{SSE}{SST} = 1 \quad \text{e} \quad 1 - \frac{SSE}{SST} = 0$$

- ajustamento intermédio (relação linear imperfeita) $\Rightarrow SSA \neq 0$ e $SSE \neq 0 \Rightarrow$

$$0 < \frac{SSE}{SST} < 1 \quad \text{e} \quad 0 < 1 - \frac{SSE}{SST} < 1$$

Mas,

$$\begin{aligned}1 - \frac{SSE}{SST} = \frac{SSA}{SST} &= \frac{\sum_{k=1}^{n}(\hat{y}_k - \bar{y})^2}{\sum_{k=1}^{n}(y_k - \bar{y})^2} \\ &= \frac{b^2\sum_{k=1}^{n}(x_k - \bar{x})^2}{\sum_{k=1}^{n}(y_k - \bar{y})^2} = b^2\frac{s_x^2}{s_y^2} = r^2\end{aligned}$$

onde r é o coeficiente amostral de Pearson para os dados (x_k, y_k).
(Na segunda igualdade usámos: $\hat{y}_k - \bar{y} = b(x_k - \bar{x})$)
Assim, r^2, que varia entre 0 e 1, mede o grau de linearidade dos dados.
Este número chama-se coeficiente de determinação e é tanto maior quanto mais o modelo linear se adequa aos dados.

Dividindo ambos os membros da equação

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2+\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

por $n-1$, vem

$$\frac{\sum_{k=1}^{n}(y_k-\bar{y})^2}{n-1}=\frac{\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2}{n-1}+\frac{\sum_{k=1}^{n}(y_k-\hat{y}_k)^2}{n-1}$$

Mas, pode provar-se facilmente que que a média dos $\hat{y}_k$ é $\bar{y}$; além disso, temos $\sum_{k=1}^{n}(y_k-\hat{y}_k)^2=\sum_{k=1}^{n}e_k^2=\sum_{k=1}^{n}(e_k-\bar{e})^2$, uma vez que $\bar{e}=0$. Assim a equação anterior pode ser escrita como

$$s_y^2=s_{\hat{y}}^2+s_e^2,$$

em que s_y^2 representa a variância total da amostra dos y_k, $s_{\hat{y}}^2$ representa a variância explicada pelo ajustamento (i.e. pela regressão linear de y em x) e s_e^2 representa a variância residual, devida a erro.

Temos também

$$r^2=\frac{\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2}{\sum_{k=1}^{n}(y_k-\bar{y})^2}=\frac{s_{\hat{y}}^2}{s_y^2},$$

o que mostra que r^2 representa a fracção (ou percentagem) da variância total que é devida ao ajustamento.

A regressão linear simples $\hat{y}=a+bx$ insere-se no caso mais geral de um modelo linear, isto é, de um modelo da forma

$$\hat{y}=a+b_1x_1+b_2x_2+\cdots$$

Trata-se de um modelo linear nos parâmetros $a,b_1,b_2,\ldots$.
Nesta fórmula, $x_1,x_2,\ldots$ podem ser diferentes variáveis (teremos um modelo de regressão linear múltipla), podem ser funções de uma mesma variável (dizemos então que temos regressão curvilínea), por exemplo

$$\hat{y}=a+b_1x+b_2x^2,$$

podendo ainda ter-se uma combinação dos dois casos, por exemplo,

$$\hat{y}=a+b_1x_1+b_2x_2+b_3x_1^2+b_4\cos x_2.$$

Todos estes casos são resolvidos de forma semelhante à regressão linear simples, pelo critério dos mínimos quadrados, sendo os parâmetros $a,b_1,b_2,\ldots$ determinados de forma a minimizar a soma dos quadrados dos desvios $\sum_{k=1}^{n}e_k^2=\sum_{i=1}^{n}(\hat{y}_k-y_k)^2$.

Um caso particularmente importante deste tipo de modelos é o da regressão polinomial, em que se procura ajustar aos dados um polinómio de um determinado grau m,

$$\hat{y}=a+b_1x+b_2x^2+\ldots+b_mx^m.$$

Também é possível (embora seja um problema de mais difícil resolução) ajustar um modelo não linear, ou seja, um modelo da forma

$$\hat{y}=f(x,a,b_1,b_2,\ldots)$$

em que $a,b_1,b_2,\ldots$ são parâmetros e f é uma função *não linear* desses parâmetros. Por exemplo, um modelo desse tipo será

$$\hat{y}=ae^{bx},$$

(modelo de crescimento/decrescimento exponencial).