

O coeficiente de correlação

Quanto mais próximo de 1 melhor?

Raquel M.C. Gonçalves^a

Ana M.N. Simões^a

1. Introdução

A simples pressão de uma tecla numa não menos simples calculadora de bolso é, hoje em dia, o quanto basta para obter o valor do coeficiente de correlação, medida do ajuste de uma determinada função aos resultados experimentais.

Ao acréscimo de frequência na determinação do coeficiente de correlação, contudo, tem vindo a corresponder um decréscimo na execução de gráficos, na sua maioria manuais e enfadonhos, o que, aliado à interpretação incorrecta do valor daquela grandeza — QUANTO MAIS PRÓXIMO DE 1 MELHOR — pode conduzir a conclusões erróneas relativamente à qualidade do ajuste.

A Química, como muitas outras ciências, já não dispensa o uso de análise estatística, e o cálculo do coeficiente de correlação é de utilidade inegável. Com este trabalho pretendemos contribuir para a sua adequada utilização.

Ilustra-se o exposto com alguns exemplos.

2. Definição

Sejam $(x_{i1}, x_{i2}, \dots, x_{in}, y_i)$ N observações que se relacionaram estatisticamente por uma função analítica linear nos coeficientes a_0 e a_j :

$$y = a_0 + \sum_{j=1}^n a_j x_j \quad (1)$$

utilizando o método dos mínimos quadrados.

O coeficiente de correlação entre y e x_j , r_{jy} , é definido por:

$$r_{jy} = \frac{s_{jy}^2}{s_j s_y} \quad (2)$$

s_j e s_y representam os desvios padrão estimados da amostra para x_j e y , respectivamente, e s_{jy}^2 a covariância estimada da amostra entre as variáveis citadas, isto é,

$$s_j = \left\{ \frac{1}{N-1} \sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 \right\}^{1/2} \quad (3)$$

$$s_y = \left\{ \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2 \right\}^{1/2} \quad (4)$$

$$s_{jy}^2 = \frac{1}{N-1} \sum_{i=1}^N [(x_{ij} - \bar{x}_j)(y_i - \bar{y})] \quad (5)$$

\bar{x}_j e \bar{y} são os valores médios:

$$\bar{x}_j = \frac{\sum_{i=1}^N (x_{ij})}{N} \quad \text{e} \quad \bar{y} = \frac{\sum_{i=1}^N (y_i)}{N}$$

No caso particular de uma recta, $n=1$ na equação (1), a equação (2) toma a seguinte forma:

$$r = r_{xy} = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{\{N \sum x_i^2 - (\sum x_i)^2\}^{1/2} \{N \sum y_i^2 - (\sum y_i)^2\}^{1/2}} \quad (6)$$

onde todos os somatórios se referem à variação do índice i de 1 até N . É esta, pois, a expressão que permite obter o vulgarizado coeficiente de correlação de uma recta, r .

O coeficiente de correlação, contudo, pode também ser definido entre quaisquer duas variáveis "independentes", x_j e x_k , por uma expressão idêntica à equação (2):

$$r_{jk} = \frac{s_{j k}^2}{s_j s_k} \quad (7)$$

em que s_j e s_k são obtidos por expressões do tipo da equação (3) e s_{jk}^2 da equação (5) onde o factor $(y_i - \bar{y})$ deve ser substituído por $(x_{ik} - \bar{x}_k)$.

Quer (2), quer (7), permite medir a correlação entre duas variáveis.

Por esse motivo, r , assim obtido, designa-se por *coeficiente de correlação simples*.

Este conceito, porém, pode ser extrapolado de modo a incluir correlações múltiplas entre grupos de variáveis tomadas simultaneamente. Tendo por base a equação (1), o *coeficiente de correlação múltipla*, R , é dado por:

$$R = \left\{ \sum_{j=1}^n (a_j s_{jy}^2 / s_y^2) \right\}^{1/2} = \left\{ \sum_{j=1}^n (a_j s_j r_{jy} / s_y) \right\}^{1/2} \quad (8)$$

Enquanto que o coeficiente de correlação simples entre x_j e y é útil para testar se a variável j deve ou não ser incluída na função de ajuste, o coeficiente de correlação múltipla indica sobre a adequação da função como um todo.

3. Significado

O coeficiente de correlação não é uma grandeza desligada da análise de regressão e do método dos mínimos

^a CECUL — Instituto Bento da Rocha Cabral, 14, 1200 Lisboa.

quadrados. Quer isto dizer que o tratamento estatístico das observações só deve ser efectuado após verificação do cumprimento dum certo número de condições. Transcrevem-se em seguida, de uma forma resumida, as hipóteses subjacentes ao método dos mínimos quadrados:

1 — A equação imposta deve ser uma função “correcta”; deve conter todas as variáveis com interesse.
2 — $\Delta y_i = y_i - y_{\text{calc}}$ são aditivos; a função de distribuição de y_{calc} tem a mesma forma da distribuição de Δy_i .
3 — As variáveis independentes não têm erro; toda a incerteza está concentrada em y_i .

4 — Δy_i são independentes entre si.

5 — Δy_i tem uma distribuição gaussiana; o critério de estimativa dos coeficientes a_0 e a_j é o dos mínimos quadrados.

6 — Δy_i tem distribuição em torno de zero; não há desvios sistemáticos.

7 — O desvio padrão dos valores experimentais é, frequentemente considerado constante.

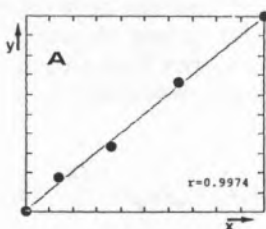
Uma fonte de erro possível na interpretação do coeficiente de correlação pode, pois, provir do não cumprimento de algumas destas hipóteses.

Outra fonte de erro possível advém da apresentação de valores do coeficiente de correlação sem o indicativo, indispensável, do número de graus de liberdade. De facto, valores daquela grandeza, só por si, não podem ser usados para indicar o grau de correlação entre variáveis. Um teste comum consiste em comparar os valores do coeficiente de correlação com a probabilidade de distribuição de uma população semelhante não correlacionada, P. Constroem-se, assim, tabelas de que é exemplo a Tabela 1 para o caso de uma relação linear. A variação do valor do coeficiente de correlação com o número de observações, para igual probabilidade de não correlação, é notável. Por exemplo, $r=0.991$ para uma amostra de 5 pares de pontos tem o mesmo significado estatístico que $r=0.324$ para 100 pares de pontos ($P=0.001$).

4. Exercícios

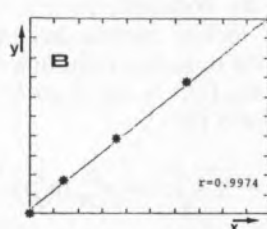
Problema 1

x	1	3	6	10	15
y	0.8	2.5	4.0	7.0	10.2



$$y = 0.2444 + 0.6651 x$$

x	1	3	6	10	15
y	0.8	2.4	4.4	7.0	10.0



$$y = 0.3533 + 0.6524 x$$

Qual a função de melhor ajuste?

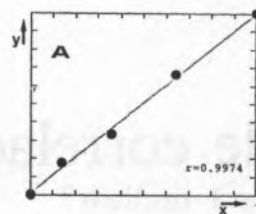
Resposta:



teste 6.
A função A tem melhor ajuste. A função B não cumpre a hipó-

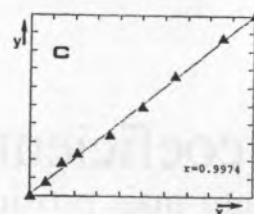
Problema 2

x	1	3	6	10	15
y	0.8	2.5	4.0	7.0	10.2



$$y = 0.2444 + 0.6651 x$$

x	1	2	3	4	6	8	10	13	15
y	0.8	1.5	2.5	3.0	4.0	5.4	7.0	9.0	10.2



$$y = 0.2204 + 0.6680 x$$

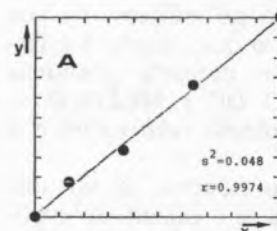
Qual a função de melhor ajuste?

Resposta:

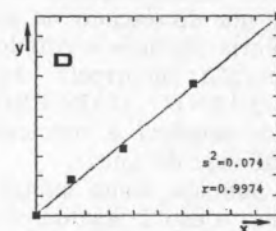
A função C tem melhor ajuste. Igual valor de r, para maior número de pontos, significa maior probabilidade de correlação entre as variáveis.

Problema 3

x	1	3	6	10	15
y	0.8	2.5	4.0	7.0	10.2



$$y = 0.2444 + 0.6651 x$$



$$y = 0.2371 + 0.6682x - 0.0002x^2$$

Qual a função de melhor ajuste?

Resposta:

A função A tem melhor ajuste. A equação que melhor se adapta aos valores experimentais é a que corresponde ao menor valor da variância estimada do ajuste $s^2 = \Sigma(\Delta y_i^2)/(N-n-1)$.

Tabela 1
r em função de N e P

N \ P	0.50	0.20	0.10	0.050	0.020	0.010	0.005	0.002	0.001
3	0.707	0.951	0.988	0.997	1.000	1.000	1.000	1.000	1.000
4	0.600	0.800	0.900	0.950	0.980	0.990	0.995	0.999	1.000
5	0.494	0.687	0.805	0.878	0.934	0.959	0.974	0.986	0.991
6	0.347	0.608	0.729	0.811	0.882	0.917	0.942	0.963	0.974
7	0.209	0.553	0.669	0.754	0.832	0.878	0.906	0.935	0.951
8	0.281	0.507	0.621	0.707	0.789	0.834	0.870	0.905	0.925
9	0.260	0.472	0.582	0.666	0.750	0.796	0.836	0.875	0.898
10	0.242	0.443	0.549	0.632	0.715	0.765	0.805	0.847	0.872
11	0.228	0.419	0.521	0.602	0.685	0.735	0.776	0.820	0.847
12	0.216	0.398	0.497	0.576	0.658	0.708	0.750	0.795	0.823
13	0.206	0.380	0.476	0.553	0.634	0.684	0.726	0.772	0.801
14	0.197	0.365	0.458	0.532	0.612	0.661	0.703	0.750	0.779
15	0.189	0.351	0.441	0.514	0.592	0.641	0.683	0.730	0.759
16	0.182	0.338	0.426	0.497	0.574	0.623	0.664	0.711	0.742
17	0.176	0.327	0.412	0.482	0.558	0.606	0.647	0.694	0.725
18	0.170	0.317	0.400	0.468	0.543	0.590	0.631	0.678	0.709
19	0.165	0.308	0.389	0.456	0.529	0.575	0.616	0.662	0.693
20	0.160	0.299	0.378	0.444	0.516	0.561	0.602	0.648	0.679
22	0.152	0.284	0.360	0.423	0.492	0.537	0.578	0.622	0.652
24	0.145	0.271	0.344	0.404	0.472	0.515	0.554	0.599	0.629
26	0.138	0.260	0.330	0.388	0.455	0.496	0.534	0.578	0.607
28	0.132	0.250	0.317	0.374	0.437	0.478	0.515	0.558	0.586
30	0.126	0.241	0.306	0.361	0.423	0.463	0.499	0.541	0.570
32	0.124	0.233	0.296	0.349	0.409	0.448	0.484	0.525	0.554
34	0.120	0.225	0.287	0.339	0.397	0.436	0.470	0.511	0.539
36	0.116	0.219	0.279	0.329	0.386	0.424	0.458	0.498	0.525
38	0.113	0.213	0.271	0.320	0.376	0.413	0.446	0.486	0.513
40	0.110	0.207	0.264	0.312	0.367	0.403	0.435	0.474	0.501
42	0.107	0.202	0.257	0.304	0.358	0.393	0.425	0.463	0.490
44	0.104	0.197	0.251	0.297	0.350	0.384	0.416	0.453	0.479
46	0.102	0.192	0.246	0.291	0.342	0.376	0.407	0.444	0.469
48	0.100	0.188	0.240	0.285	0.335	0.368	0.399	0.435	0.460
50	0.098	0.184	0.235	0.279	0.328	0.361	0.391	0.427	0.451
60	0.089	0.168	0.214	0.254	0.300	0.330	0.359	0.391	0.414
70	0.082	0.155	0.200	0.239	0.278	0.306	0.332	0.363	0.385
80	0.077	0.145	0.185	0.220	0.260	0.286	0.311	0.340	0.361
90	0.072	0.136	0.174	0.207	0.245	0.270	0.293	0.322	0.341
100	0.068	0.129	0.165	0.197	0.232	0.256	0.279	0.305	0.324

Bibliografia

- P.R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences", McGraw-Hill, New York, 1969.
- D.G. Watts, "Kinetic Data Analysis", Plenum Press, New York, 1981.
- R.J. Cvetanovic, D.L. Singleton e G. Paraskeopoulos, J. Phys. Chem., 1979, **83**, 50.
- M.D. Pattengill e D.E. Sands, J. Chem. Educ., 1979, **56**, 244.
- R.B. Huff e K.N. Carter, J. Chem. Educ., 1981, **58**, 49.
- P.F. Tiley, Chem. in Britain, 1985, 162.