J. Pharm. Pharmacol. 1981, 33: 460–461
Communicated January 23, 1981

0022–3573/81/070460–02 $02.50/0

Estimation of the built-in element of correlation in the least-squares fit of results to some algebraic expressions, including the Hansch equation

R. B. BARLOW, *Department of Pharmacology, Medical School, University of Bristol, Bristol BS8 1TD, U.K.*

In a previous note (Barlow 1981) it was pointed put that in the fitting of results to the Hansch equation (log 1/C $= a\pi^2 + b\pi + c\sigma + d$) there were several factors which might produce a bias towards correlation. An attempt was made to assess their total contribution by calculating the correlation coefficient, r, for some published results in which the values of π and σ were retained but values of log 1/C were replaced by random numbers lying in the same range as the original values. The bias appeared to be considerable; for instance, with the original results of Hansch et al (1962) 4·75% of the values of r obtained with random values of log 1/C were better than 0·6 and 1% were better than 0·7.

The apparent correlation could arise because of the limits arbitrarily set to the numbers fitted—values of log 1/C fall in a particular range and values of π and σ are usually unevenly distributed. It could also arise because of the fitting of relatively small numbers to an equation which contains 4 coefficients (a, b, c and d); with only 4 results the fit must appear perfect. The relation between the number of results, the number of variables and the values of r which must be exceeded for significance at a particular level can be derived from statistical theory but it can also be studied experimentally (and perhaps more easily) by using random-generated numbers in place of actual data. To do this, Topliss & Edwards (1979) used an IBM 360/158 computer and a Fortran multiple-regression analysis program but it is possible to make similar calculations with inexpensive microcomputers, such as a Commodore PET 2001. With small numbers the built-in element of correlation is considerable and this note attempts to show its extent in a form which may be easily appreciated.

Random-generated values of x, y, and, where appropriate, z, were fitted by least-squares to the expressions

$$y = mx + c \quad \text{(i)}$$

$$y = a + bx + cx^2 \quad \text{(ii)}$$

$$y = a + bx + cz \quad \text{(iii)}$$

$$y = ax^2 + bx + cz + d \quad \text{(iv)}$$

and the actual and fitted values of y (y and y_c) were used to calculate the proportion of the variance explained by regression, $r^2 \left(= \frac{S(y_c - \bar{y})^2}{S(y - \bar{y})^2}\right)$. The number of points fitted, N, was usually 6, 10, 16 and 25 and the number of trials made, nt, was usually 500 or 1000.

The results are summarized in Table 1 and the values for a straight line can be checked against what would be expected from the corresponding values of Student's $t \left(= \sqrt{\frac{r^2(N-2)}{(1-r^2)}}\right)$. In 1000 trials of the fit of 10 pairs of random-generated values of x and y to a straight line, for instance, 10% of the values of r^2 were above 0·29, 5% were above 0·38 and 1% were above 0·58 (Fig. 1); the corresponding limits calculated from Student's t with 8 degrees of freedom are 0·30 ($P = 0·1$), 0·40 ($P = 0·05$) and 0·58 ($P = 0·01$).

Table 1. Numbers show the limit above which 10%, 5% and 1% of the values of r^2 were found to lie in the calculations of the fit of random-generated values of x, y and z to the equation shown. The number of points fitted is indicated by N and the number of trials made is indicated by nt; $\bar{r}^2$ is the mean of the values of r^2 (but the distribution is not symmetrical).

N=	$\bar{r}^2$	10%	5%	1%	nt
(i) $y = mx + c$					
6	0·201	0·529	0·662	0·833	1000
10	0·109	0·291	0·384	0·577	1000
16	0·063	0·168	0·231	0·403	1000
25	0·038	0·103	0·144	0·247	1000
(ii) $y = a + bx + cx^2$					
6	0·398	0·779	0·886 (0·966)	0·964 (0·993)	500
10	0·220	0·485	0·592 (0·704)	0·662 (0·830)	500
16	0·133	0·294	0·392 (0·466)	0·501 (0·598)	500
25	0·091	0·206	0·248 (0·305)	0·355 (0·410)	500
(iii) $y = a + bx + cz$					
6	0·386	0·773	0·867	0·948	500
10	0·212	0·452	0·540	0·669	500
16	0·120	0·270	0·330	0·491	500
25	0·079	0·190	0·233	0·324	500
(iv) $y = ax^2 + bx + cz + d$					
10	0·340	0·636	0·716 (0·806)	0·840 (0·901)	1000
14	0·231	0·449	0·536 (0·617)	0·672 (0·740)	1000
14*	0·228	0·432	0·515	0·599	500
16	0·200	0·400	0·465 (0·550)	0·620 (0·673)	1000
20	0·155	0·304	0·397 (0·449)	0·503 (0·566)	1000
20*	0·163	0·331	0·393	0·508	500
25	0·125	0·255	0·295 (0·365)	0·412 (0·470)	1000
35	0·087	0·178	0·219 (0·264)	0·290 (0·349)	1000
35*	0·097	0·169	0·198	0·288	500

Note that there is most uncertainty about the values of the 1% limit because these involve results for only small numbers (5 out of 500 trials). The asterisk indicates that in these calculations the original experimental values of π and σ (Hansch et al 1962; Hansch & Fujita 1964) were retained and only the values of log 1/C were replaced by random-generated values. Values in parentheses are the limits calculated from F.

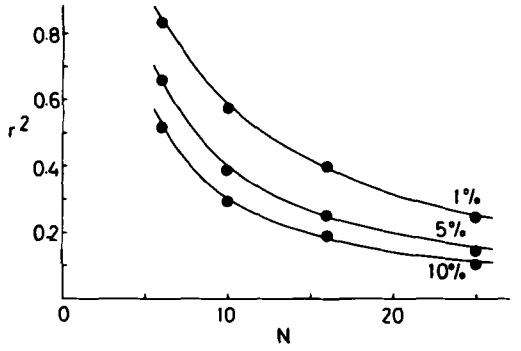

FIG. 1. The limits above which 10%, 5% and 1% of the values of r^2 lie are plotted against the number of points N, when random-generated values of x and y are fitted to $y = mx + c$. The continuous lines have been drawn (by eye) from the limiting values of r^2 calculated from Student's t; the points are the values obtained from 1000 trials with random-generated values of x and y.

With the fit to equations (ii) and (iii) the distribution of r^2 appears to be very similar, even though three variables (x, y and z) are involved in (iii) instead of two. The effect of the number of coefficients to be calculated on r^2 is illustrated in Fig. 2a, in which the value of r^2 above which 10% of the results lie is plotted against the number of points being fitted, N, for a straight line, parabola and for the Hansch equation; values for the Hansch equation for 10%, 5% and 1% of the results are shown in Fig. 2b. The limits of r can be calculated from statistical theory using the values of F in the variance ratio test (Martin 1978). With N points fitted to an equation containing p coefficients the ratio of the variance attributable to regression (with p degrees of freedom) to that attributable to error (with N-p-1 degrees of freedom) is

$$F = \frac{N\text{-}p\text{-}1}{p}\left(\frac{r^2}{1\text{-}r^2}\right)$$

so

$$r^2 = \frac{F}{\frac{N\text{-}p\text{-}1}{p} + F}$$

With $P = 0{\cdot}05$ and N = 10, F = 5·19 (Diem & Lentner 1970) and the limiting value of r^2 is 0·806. Other limiting values are shown in parentheses in Table 1 and in Fig. 2b. These are set slightly higher than the 'experimental' limits obtained in this work.

The calculations show that it is the fitting of small numbers to the Hansch equation, rather than the restriction of the values of π and σ which accounts for the apparent correlation with random biological data previously noted. The range of the values of π and σ merely limits the usefulness of the equation because they determine spanned substituent space (Hansch 1977).

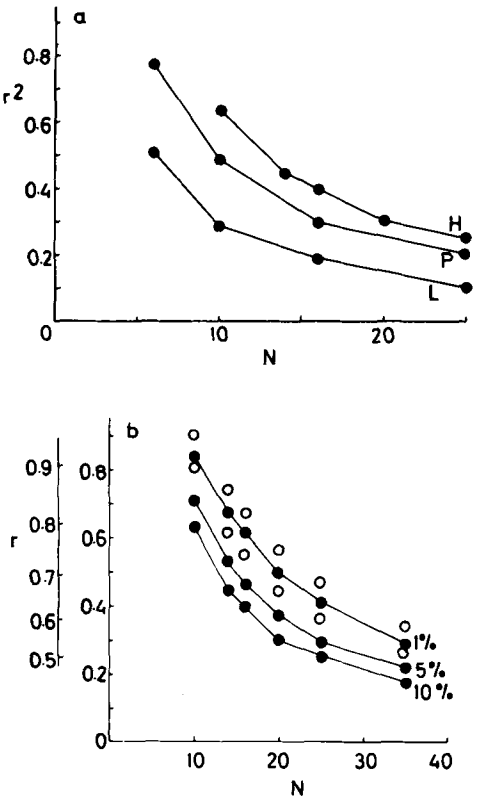

FIG. 2a. The effect of the complexity of the fitting equation. The limits above which 10% of the values of r^2 lie are plotted against the number of points fitted, N, for a straight line (L), parabola (P) and Hansch equation (H).

b. The fit of random-generated data to the Hansch equation. The limits above which 10%, 5% and 1% of the values of r^2 lie is plotted against the number of points fitted, N. Values of r are also indicated; note that with 20 points 5% of the random-generated values have $r > 0{\cdot}6$ and 1% have $r > 0{\cdot}7$. Open circles indicate limits calculated from F.

REFERENCES

Barlow, R. B. (1981) J. Pharm. Pharmacol. 33: 62–63

Diem, K., Lentner, C. (1970) Documenta Geigy; Scientific Tables, Geigy, Basle, pp. 40–41

Hansch, C., Maloney, P. P., Fujita, T., Muir, R. M. (1962) Nature (London) 194: 178–180

Hansch, C., Fujita, T. (1964) J. Am. Chem. Soc. 86: 1616–1626

Hansch, C. (1977) in: Keverling Buisman, J. A. (ed.) Biological Activity and Chemical Structure. Elsevier, Amsterdam, pp. 292–293

Martin, Y. C. (1978) Quantitative Drug Design. Marcel Dekker, New York, p. 185

Topliss, J. G., Edwards, R. P. (1979) J. Med. Chem. 22: 1238–1244