JOURNAL OF PHYSICAL ORGANIC CHEMISTRY, VOL. 10, 797–813 (1997)

REVIEW COMMENTARY

HOW TO GET WRONG RESULTS FROM GOOD EXPERIMENTAL DATA: A SURVEY OF INCORRECT APPLICATIONS OF REGRESSION†

OTTO EXNER*

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 16610 Prague 6, Czech Republic

Examples are given from older and more recent literature (kinetics, ionization equilibria, complex formation in solution, dipole moment determination, thermochemistry, resonance energies, NMR shifts, photoelectron spectroscopy) where experimental data were processed in an incorrect way from the point of view of statistics. The results were more or less biased, sometimes completely wrong. Corrected procedures, based entirely on the least-squares method, are reported; in several cases methods are proposed. Some hints are given as to how these mistakes can be avoided, how they can be revealed in the literature and how the literature data can be recalculated: the last task is the most difficult. © 1997 John Wiley & Sons, Ltd.

J. Phys. Org. Chem. **10**, 797–813 (1997) No. of Figures: 13 No. of Tables: 0 No. of References: 96

Keywords: statistics; regression; correlation analysis; incorrect data processing; least-squares method

Received 20 January 1997; revised 12 April 1997; accepted 5 May 1997

INTRODUCTION

In the whole field of correlation analysis, there is no procedure more common than plotting two sets of data against each other. Very often a straight line is drawn through the points (less often a more complex curve) and its parameters are estimated. Outside correlation analysis, an apparently identical procedure is also common in which experimental results are compared with a theory. In any case we are concerned with a statistical procedure, even when it is meant only as a rough approximation, and the laws of mathematical statistics may be violated. There is then a possibility of obtaining results competely at variance with the original data. Some of these mistakes may be surprising, as can be seen from the following examples.

A SIMPLE EXAMPLE

The relevant problems can be illustrated by a familiar example, Brown's selectivity relationship:[1]

$$\log p_f = aS_f \quad (1)$$

Several electrophilic reactions of a benzene derivative (e.g. toluene) were studied kinetically and logarithms of partial rate factors in the *para* position, p_f, plotted against Brown's selectivity factor S_f, equation (1). The linear dependence [Figure 1(A)] looks fine and everything seems to be in order unless we inquire what exactly is the selectivity factor. It turns out that it has been defined also from the electrophilic reactions of toluene,[1] taking into account also the partial rate factor in the *meta* position:

$$S_f = \log p_f - \log m_f \quad (2)$$

When we introduce S_f from equation (2) into equation (1), we obtain

$$\log p_f = a(\log p_f - \log m_f) \quad (3)$$

The deficiency of this correlation is seen immediately since one variable is present on both sides of the equation. One can imagine[2] a limiting case when all values of m_f would be almost constant, or at least only slightly variable compared with the variability of p_f. Then equation (3) would simply express the dependence of p_f on itself. In the other limiting case, if p_f were less variable than m_f, no problems would occur. Real examples are between these two limits. Simple rearrangement of equation (3) gives [note that there is a misprint (wrong sign) in equation (4) in Ref. 2)]:

* *Correspondence to:* O. Exner.

† Presented in part at the VIIth International Conference on Correlation Analysis in Chemistry, Fukuoka, 2–6 September 1996.

Contract grant sponsor: Grant Agency of the Czech Republic; *Contract grant number:* 203/96/1658.

CCC 0894–3230/97/110797–17 $17.50

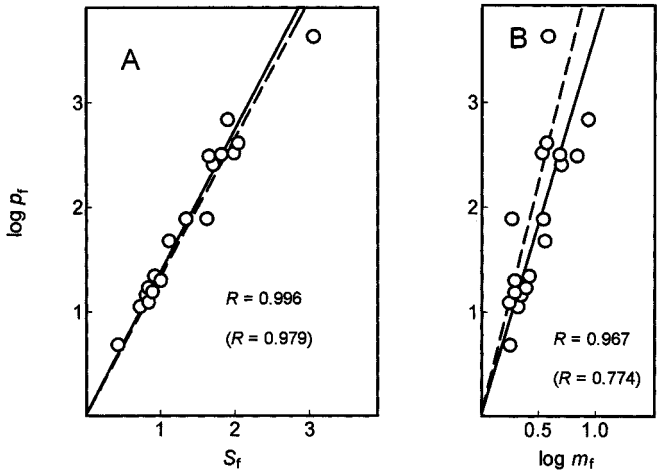

Figure 1. Brown's selectivity rule in electrophilic reactions of toluene: (A) the incorrect plot (Ref. 1) according to equation (1); (B) correct plot (Ref. 2) according to equation (3) with separation of the variables. Full lines, the correct regression line in (B) and its picture in (A); broken lines, the apparent regression line in (A) and its picture in (B). The regression lines were forced through the origin and errors in both coordinates were assumed equal (when non-forced, the correlation coefficients are given in parentheses)

$$\log p_f = \frac{a}{a-1} \log m_f \qquad (4)$$

This is a relationship between two independent experimental quantities, $\log p_f$ and $\log m_f$. The pertinent plot [figure 1(B)], on visual inspection, would not be considered as a valid linear regression. A relatively high correlation coefficient is obtained only when the line is forced through the origin: with a free intercept, R would be much lower [Figure 1(B) in parentheses] and the slope considerably different.

The incorrect plot [Figure 1(A)] thus differs from the correct plot [Figure 1(B)] in two respects:

(1) the value of the slope of the linear relationship is wrong, its estimate is biased; in Figure 1(A) and (B) the lines with correct slopes [determined from figure 1(B) and transferred into Figure 1(A)] are drawn fully, the wrong lines determined in Figure 1(A) and transferred into Figure 1(B) are broken;
(2) the fit is estimated wrongly; in our case it is overestimated, in less frequent cases it may be also underestimated; in Figure 1(A) and (B) the correlation coefficients are given on the plots; that in Figure 1(A) is an apparent value since it expresses partly also the dependence of p_f on itself.

The example also shows a simple and efficient method of detecting an erroneous treatment. It is sufficient to substitute for all derived and modified quantities their precursors, if necessary in several steps, so going back to the original experimental data [substitution for S_f in Figure 1(A) and in equation (1)]. These data are then recalculated from the reported correlations and compared with the actual experiments [transferring the broken line from Figure 1(A) and comparing with the points in Figure 1(B)]. Simple comparison may be sufficient in many cases. If not, statistical tests are possible.

Returning to the selectivity relationship, one should not conclude that it is not valid at all. It represents in fact the Hammett equation (with the constants σ^+) arranged for one substituent and various reactions. However, in the particular case of toluene derivatives the accuracy is low since the substituent effects are relatively small compared with more efficient substituents.

The subsequent examples represent a selection from those I have been collecting for many years. In all examples, the two mentioned defects will be found repeatedly but their relative importance may vary considerably from one case to another. The mentioned test served to detect wrong procedures in all cases but with differing results. The examples are arranged according to the underlying statistical model.

LINEAR REGRESSION

The model of linear regression assumes the following hard conditions:[3]

1. The explanatory (independent) variable x is an exact quantity, free of any error.
2. An (exact) linear relationship, equation (5), exists between x and the response function (dependent variable) η.
3. The exact values of η are not available, one knows only the values y differing by a random quantity ε ('error') $y = \eta + \varepsilon$. The distribution of ε need not necessarily be Gaussian but must meet certain conditions. In chemistry

ε is mostly the experimental error and these conditions are fulfilled. When ε represents an additional explanatory variable (i.e. an unknown factor, controlling also η), this condition may become questionable.

4. The errors ε are independent of x.

When these conditions are met, one can estimate the parameters α and β of equation (5) within the framework of the least-squares method[3] to obtain their estimates a and b in the regression equation (6):

$$\eta = \alpha + \beta x \tag{5}$$

$$y = a + bx \tag{6}$$

In many applications condition (1) is not met. Regression models are also available[4] for both variables loaded with error but in practice it is mostly sufficient to choose the more accurate quantity as x and the classical model can be used with a good approximation. Condition (2) may be rather hard: if it is not met, we could estimate something which does not exist. In correlation analysis, it is usually viewed as a preliminary hypothesis. When it cannot be *a posteriori* disproved, it is accepted as possibly valid. Most important for the following considerations may be condition (4). For instance, when y is not a direct experimental value but has been transformed, its error can vary substantially. Particular attention must been given to equations in which one variable is involved on both sides [e.g. equation (3)]. Then an error in this quantity appears in both coordinates mutually correlated and both conditions (1) and (4) are violated.

Let us reconsider the above example. In Figure 1(A), x is loaded with an error which is equal to or still greater than the error in y, both errors are mutually dependent. In Figure 1(B), the only problem is that x is also loaded with error. When the right regression model[4] is used, the estimate of a differs only slightly from that obtained in common regression; the correlation coefficient is the same in the two models.

The linear dependence between reaction enthalpies $\Delta H°$ and entropies $\Delta S°$ in a series of related reactions, the compensation effect, has been much discussed.[5–8] For the ionization of substituted anilines a good linear dependence was obtained[7] [Figure 2(A)] according to the equation

$$\Delta H° = \beta \Delta S° + \text{constant} \tag{7}$$

The proportionality constant β is called the isoequilibrium (or isokinetic) temperature.[8] Experimental values of $\Delta H°$ were determined calorimetrically and $\Delta S°$ values were obtained indirectly from $\Delta H°$ and from the measured pK (i.e. from $\Delta G°$):

$$\Delta S° = (\Delta H° - \Delta G°)/T \tag{8}$$

Substituting $\Delta S°$ from equation (8) into equation (7) gives a relationship between experimentally independent quantities $\Delta H°$ and $\Delta G°$:

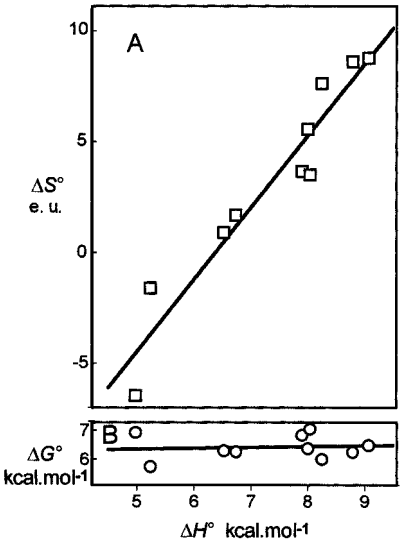

Figure 2. The enthalpy–entropy relationship for equilibria in the ionization of methyl-substituted anilines in water: (A) plot of $\Delta H°$ vs $\Delta S°$ (Ref. 7); (B) plot of original experimental quantities $\Delta G°$ vs $\Delta H°$

$$\Delta G° = \frac{\beta - T}{\beta} \Delta H° + \text{constant} \tag{9}$$

The corresponding plot [Figure 2(B)] shows no dependence: the values of $\Delta G°$ are approximately constant in comparison with the variations of $\Delta H°$. Mathematically this example is the same as the preceding one [Figure 1(A) and (B)] but there is a difference in the possible interpretation. In the selectivity relationship, the parameter S_f has been defined arbitrarily: it has no physical meaning and its definition is justified only by its (apparent) correlation with experimental quantities. In the enthalpy–entropy relationship, $\Delta S°$ has a clear physical meaning and could also be determined by a direct experiment: in that case there would be no objections against the plot, Figure 2(A). The correct interpretation should be that $\Delta G°$ is approximately constant within the reaction series: from it a mathematically necessary correlation between $\Delta H°$ and $\Delta S°$ comes into existence. It is thus true that plots of $\Delta H°$ vs $\Delta S°$ can have a meaning,[6] but it is also true that they can express only the trivial fact that $\Delta G°$ is approximately constant.[5] In the latter case, the slope β is near to the experimental temperature. However, the value of β itself is not sufficient for deciding whether a correlation is real or not.[6,8]

NONLINEAR REGRESSION

Linear equations are mostly only the first approximation. More generally, an observable quantity y depends on an explanatory variable x and several parameters α, β, γ

through a nonlinear equation:

$$y = f(x, \alpha, \beta, \gamma, \ldots) \quad (10)$$

With the assumptions, essentially the same as for the linear regression, one can estimate the parameters using the least-squares condition:

$$SSQ = SD^2(N-p) = \sum_i [(y_i - f(x_i, \alpha, \beta, \gamma)]^2 = \min \quad (11)$$

The condition of minimum sum of squares *SSQ* determines all the parameters in an unambiguous way and gives also the standard deviation *SD* of the fit with respect to the number of data *N* and number of parameters *p*. There is no problem in devising a correct computer program. Programmers tend to focus attention mostly on the shortest way to reach the minimum (the Newton–Raphson method, etc.). However, for smaller data sets and few parameters, as is common in experimental chemistry, a simpler program may be more useful, calculating *SSQ* successively for all possible values of a certain parameter: this may be particularly useful when only this parameter causes the nonlinearity.[8] A more difficult task may concern the confidence intervals of the parameters. The uncertainties of individual parameters may be very different and also strongly mutually dependent. Generally we must find all the possible sets of parameters which yield any acceptable standard deviation SD_a. The latter is most correctly determined by an *F*-test:[9]

$$SD_a^2 = SD^2 \left[1 + \frac{p}{N-p} F_{p, N-p}(\alpha) \right] \quad (12)$$

For two parameters, α and β, the result may be pictured by plotting them against each other. Equation (12) is then represented by a contour line: all acceptable combinations of α and β are situated inside this curve (see Figure 4, later, as an example).

In the past, the nonlinear equations have often been transformed into linear equations, sometimes with great mathematical ingenuity. This had some justification in the pre-computer era since otherwise a solution was difficult, sometimes impossible. Nevertheless, it still occurs at present, here and there. Evidently, the estimated parameters are biased and the literature thus contains many data that are completely at variance with the experimental facts. Particularly bad are transformations in which the original variables each appear on both sides of the equation. Still more mistakes occur with the uncertainty of the parameters. Very often it is assumed that one parameter (α) is already known and the uncertainty of the second (β) is searched for this given value of α. When the parameters are strongly dependent, their uncertainty is badly underestimated, sometimes by an order of magnitude or more.

Dipole moments in solution are commonly determined from the measured permittivity ε as a function of concentration, expressed as c_2 or as the weight fraction w_2, respectively, in different theories. The second, less important experimental quantity may be either the density or refractive index according to various methods.[10] Complications occur when the compound forms a dimer. Then we have to estimate three parameters: the dipole moments of the monomer and dimer, μ_M and μ_D, in addition to the equilibrium constant *K*. Even with the frequent assumption that $\mu_D = 0$, the equation is rather complex. Several ingenious methods for its linearization have been advanced[11,12] and used more or less extensively.[13,14] For instance, in equation (13) the terms in square brackets were plotted against each other and *K* and the polarizability α_M calculated from the slope and intercept.[12] From α_M, one obtains μ_M.

$$\frac{1}{\dfrac{A}{c_2}\dfrac{\varepsilon-\varepsilon_1}{\varepsilon+2} - B} = \frac{1}{\alpha_M - \alpha_D/2} +$$

$$\frac{2K}{(\alpha_M - \alpha_D/2)^2}\left[A\frac{\varepsilon-\varepsilon_1}{\varepsilon+2} - Bc_2 \right] \quad (13)$$

A and *B* are composite quantities considered as constants (*B* approximately); the proper variables are the permittivity ε and concentration c_2. (The subscript 1 always refers to the solvent and subscript 2 to the solute; quantities without a subscript belong to the solution.) The essential mistake with equation (13) is that both ε and c_2 are involved on both sides. Therefore, the finding[13,14] was not surprising that this procedure yields wrong results. The following equation can be derived in which the variables are separated[13] (c_2 has been replaced by w_2):

$$\frac{\varepsilon-1}{\varepsilon+2} = \frac{\varepsilon_1-1}{\varepsilon_1+2}\frac{1-w_2}{1+d_1\beta w_2} + \frac{2P^\circ_M - P^\circ_D}{8000K}\left[\left(1 + \frac{8000Kd_1w_2}{M+Md_1\beta w_2} \right)^{1/2} - 1 \right] + \frac{P^\circ_D d_1 w_2}{2M(1+d_1\beta w_2)} + \frac{aR_M d_1 w_2}{M+Md_1\beta w_2} \quad (14)$$

Equations (13) and (14) differ slightly in using either the refractive index *n* (involved in B) or density *d* (involved in β) as the auxiliary quantity; R_M is the molar refraction and *a* is a constant. Another detail is that in equation (14) there is not ε alone on the left-hand side but its function. However, $(\varepsilon-1)/(\varepsilon+2)$ is a fairly accurately linear function of ε in the range of measured values. An example in Figure 3 reveals that equation (13) may yield a completely wrong result which has in fact nothing in common with the original experimental data. According to equation (13), a rough linear dependence was obtained, Figure 3(A), and μ_M and *K*

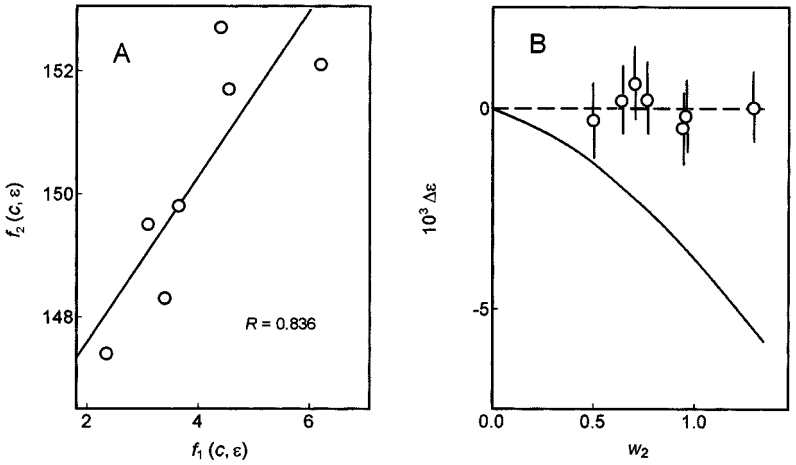

Figure 3. Dipole moment of a possibly dimerizing substance $C_5H_4N^+C^-(CN)_2$: (A) incorrect plot of composed variables (Ref. 12) according to equation (13) with the regression line; (B) correct plot with separated variables (Ref. 14) according to equation (14). Broken line calculated in (B) with the assumption that no dimerization occur and used as reference; solid curve derived in (A) as the apparent regression line and transferred into (B)

were calculated from it.[12] We recalculated[13] the required values of ε from these μ_M and K values and compared them with the acutal experimental ε in Figure 3(B). As reference, the values of ε were used as calculated with the assumption that the compound does not dimerize at all (broken line): both the experimental points and the wrong calculations are shown as deviations in relative values of ε. The hypothesis of no dimerization cannot be rejected. Figure 3 is a very good example of testing complex equations by recalculating back the original experimental data. In the more recent literature correct statistical programs have been already reported.[15] Nevertheless, even the suffer from the less important defect, underestimating the uncertainty of the parameters. In our papers,[13,14] a correct statistical procedure for calculating μ_M, μ_D and K was proposed. However, very good experimental data[16] are needed if these parameters are to be obtained with any reliability.

Similar problems are encountered with the dipole moments of complexes. Although there are more parameters, this case is more favorable. The reason is that dipole moments of the two components, μ_A and μ_B (in other terms the respective polarizations P_A and P_B), can be determined separately on pure compounds A and B. Subsequent measurements of ε and d of their mixtures, at variable weight fractions w_A and w_B, can serve to determine the dipole moment of the complex μ_{AB} and the equilibrium constant K. Nevertheless, a linearized equation was also advanced[17] and applied widely:[18,19] with an excess of B, an apparent polarization P_A^* (in the presence of B at a given concentration) is determined from the experimental ε and d. This is repeated for different w_B and introduced into the equation:

$$\left[\frac{1}{P_A^* - P'_A}\right] = \frac{1}{P_{AB} - P'_A - P_B} + \frac{M_B}{K(P_{AB} - P'_A - P_B)}\left[\frac{1}{w_B d}\right] \tag{15}$$

(here P'_A may equal or need not equal[18] P_A exactly). When the terms in square brackets are plotted against each other, one obtains from the intercept the polarization of the complex P_{AB} and when this is known, also K from the slope. The equation is not loaded with such great defects as equation (13). The proper explanatory variable is w_B but using $x = 1/w_B d$ is acceptable since the experimental density of the solution d is almost constant. More important is that the distribution of errors in $1/(P_A^* - P'_A)$ depends on x. Equation (15) was solved according to P_A^* and a program written[20] minimizing sum of squares in P_A^* instead in $1/(P_A^* - P'_A)$. In a few cases[19] the results from equation (15) were completely at variance with the facts and plots similar to Figure 3(A) were obtained. Mostly the results were not so sharp since the experimental points were scattered and the difference between correct and incorrect theoretical curves was not so great. In these cases one obtains the parameters K and μ_{AB} with great uncertainties, and sometimes no convergence was reached.[20] An example which is not as bad is shown in Figure 4. The strong dependence means that either the complex is present at high concentration and has a low dipole moment or vice versa. We obtained[20] $K = 1081$ with confidence limits 340–2290; the original literature[18] reported 1178 ± 25 (units $cm^3\,mol^{-1}$).

In some particular cases even biased estimates may appear sufficiently precise and a procedure which is wrong

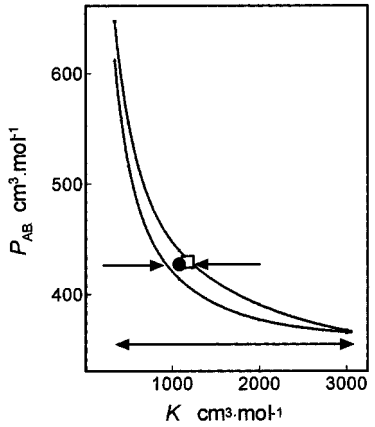

Figure 4. Dipole moment determination of a complex of 1-phenyl-1,3-butanedione with triethylamine: the contour map of the equilibrium constant K vs polarization of the complex P_{AB} (function of μ_{AB}). Values acceptable according to equation (12) are situated inside the curve. ●, Least-squares solution (Ref. 20); □, values from the original literature (Ref. 18); long double-headed arrow, the correct confidence interval of K; short arrows delimit the apparent confidence interval for a fixed value of P_{AB}

in principle gives practically the same result as a rigorous treatment. One such case was encountered in another example from the theory of dipole moments.[21] When μ is calculated according to the more exact Onsager theory, the resulting equation can be linearized into the form

$$\left[(1-\theta_2)\frac{\varepsilon-n_1^2}{2\varepsilon+n_1^2}+\theta_2\frac{\varepsilon-n_2^2}{2\varepsilon+n_2^2}\right]$$

$$=\frac{4\pi N\mu^2}{9000kT}\left[\left(\frac{2+n_2^2}{2\varepsilon+n_2^2}\right)^2\varepsilon c_2\right] \qquad (16)$$

This equation is essentially a function of measured ε in dependence on the concentration c_2. The refractive indices of the solvent (n_1) and solute (n_2) and the density of the solute d_2 are constants. In the auxiliary quanitity θ_2, the concentration is also involved:

$$\theta_2=M_2c_2/1000d_2 \qquad (17)$$

The task is to estimate the parameter μ. As in the foregoing examples, equation (16) was treated as linear and the terms in square brackets plotted against each other: from the slope one obtains μ. Separation of variables would be cumbersome. However, a rigorous solution is possible by successive approximations: For a trial value of μ, equation (16) is solved numerically for ε. This is repeated for all experimental values of c_2 and the sum of squares $\Sigma(\varepsilon_{calc}-\varepsilon_{exp})^2$ is calculated. By trying further values of μ, the best one is found giving the minimum sum of squares. Remarkably, this rigorous procedure gave practically the same results as the linearized equation (16).[21]

The next example is very important; it has been treated extensively in the literature[9,22–28] and a correct statistical solution was advanced many years ago.[25–28] Nevertheless, a popular incorrect method[22] is still being used.[29,30] When complex formation is followed spectroscopically, the absorbance A of the complex may be measured at a constant concentration c_Y of one component and at variable concentrations c_X of the second component, the latter being always in excess. The well known Benesi–Hildebrand equation[22] represents a transformation of the dependence into a formally linear form (the optical path l is a constant):

$$\left[\frac{lc_Y}{A}\right]=\frac{1}{\varepsilon}+\frac{1}{K\varepsilon}\left[\frac{1}{c_X}\right] \qquad (18)$$

When the terms in square brackets are used as explanatory variable and response function, respectively, one can calculate the equilibrium constant K and extinction coefficient ε of the complex from the slope and intercept. This equation is not so defective as those in the preceding examples: when c_X is exact, $1/c_X$ is also exact. When the errors in A are constant, those in $1/A$ are not but the mistake is not serious when A is not too variable. The main defect is that K is obtained as a ratio of two estimates and has an unsymmetrical distribution. Correct statistical solutions[25–28] were based on a nonlinear regression when equation (18) was solved for A:

$$A=\frac{K\varepsilon lc_Xc_Y}{Kc_X+1} \qquad (19)$$

Using different programs, both K and ε can be obtained by minimizing errors in A. Our least-squares recalculation of several examples[9] confirmed most of the previous work[25–27] and revealed that even the Benesi–Hildebrand[22] method gives acceptable results when the experimental pattern is well arranged. However, many older papers do not satisfy this condition and reliable results cannot be obtained by any method. Figure 5(A) is an example. The range of experimental concentrations is narrow and the experimental points are compatible both with the least-squares solution (line 1) and with the shifted values of K and ε (line 2 or 3). A simple remedy would be one experimental point more, say at $c_X=0{\cdot}015$. There is thus the very reasonable requirement that statistical treatment should proceed simultaneously with the experiments which can be completed stepwise.[28] However, the main problem is with the uncertainty of the parameters K and ε, which has been underestimated in all previous literature. Similarly as in the preceding example, the estimates of the two parameters are strongly dependent on each other, particularly when the experimental pattern is insuitable.[27] According to Figure 5(B), broadly varying values of K are acceptable provided that ϵ is also simultaneously shifted. With the aid of equation (12) we

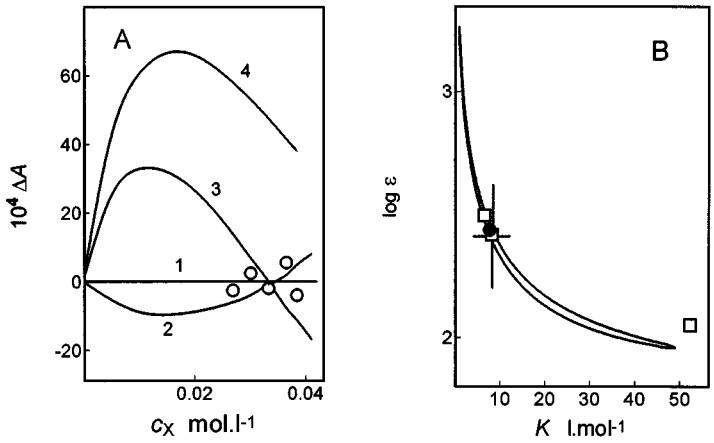

Figure 5. Spectrometric determination of the 1:1 complex of anthracene with iodine. (A) Experimental points and calculated theoretical curves as a function of concentration: 1, the least-squares solution (Ref. 9) as reference; 2 and 3, solutions with the limiting acceptable values of K and ε; 4, the theoretical curve with K and ε given in the original literature, Ref. 24. (B) Contour map of SD as function of possible K and log ε values (Ref. 9). Values acceptable according to equation (12) are situated inside the contour line. ●, Least-squares solution; □, values given in Refs 24 and 26 or calculated according to Ref. 22

obtained[9] for the confidence interval of K 0–49 (best value 7·8); previous recalculation[26,27] gave 8·3±4·6 and the original literature[24] 52·35. The Benesi–Hildebrand method[22] would yield 6·6 with no idea about the confidence interval (units mol^{-1}). Later examples from the literature are not as bad since more experimental points were measured with better accuracy. In the recent literature, the problem is that the data are not reported at all, sometimes not even the concentration range.[29,30] The example shows that the problems are not completely eliminated even with very common and broadly used methods.

To the same category belongs also the familiar Michaelis–Menten equation[31] for the kinetics of enzyme catalyzed reactions:

$$v=\frac{V_{\max}s}{K+s} \tag{20}$$

The equation is very important as the simplest form in this field and has been treated statistically in a number of papers.[32–37] One has to estimate the parameters V_{max} and K on the basis of experimental substrate concentrations s and corresponding observed reaction rates v. The problem was solved[32,33] on the basis of nonlinear regression, equation (11). The only possible improvement could be in better estimation of the uncertainties of V_{max} and K according to equation (12). In the past, several linearization procedures were suggested, which used reciprocal values of variables[34] or contained the same variable on both sides of the equation.[35] Of course, they gave biased results. In addition, several methods were advanced[36,37] not based on the least-squares principle. As far as I know, there has been no systematic reinvestigation of the previous results obtained by incorrect methods. Hence one cannot estimate how much they deviate from the least-squares values.

SYSTEM OF LINEAR REGRESSIONS

A well known example of this category is the isokinetic relationship (IKR) or compensation effect. Although the problem has been reviewed,[8,38] possible mistakes have been pointed out several times[39–42] and several correct statistical treatments have been advanced,[40,43-46] the fundamental error[47–51] is repeated again and again, even recently and in prominent journals.[52–55] Sometimes also correct and incorrect methods have been mixed together.[56,57] A linear dependence of activation enthalpy and activation entropy within a series of related reactions has the same form as equation (7):

$$\Delta H^{\neq}=\beta\Delta S^{\neq}+\text{constant} \tag{21}$$

However, there is a fundamental difference compared with equation (7) in that neither $\Delta H^{\neq}$ nor $\Delta S^{\neq}$ has been determined directly. The primary experimental quantities are the rate constants which are measured at different temperatures: $\Delta H^{\neq}$ and $\Delta S^{\neq}$ are obtained from the linear regression:

$$\log(k/T)=-\Delta H^{\neq}/2{\cdot}303\boldsymbol{R}T+\Delta S^{\neq}/2{\cdot}303\boldsymbol{R}+\log(\boldsymbol{R}/\boldsymbol{N}_{\mathrm{A}}\boldsymbol{h}) \tag{22}$$

When the Arrhenius theory is used instead, the treatment is simpler and the conclusions are the same. It follows that in equation (21) both variables are loaded with error and these errors are strongly correlated. Any experimental error, any misprint or misunderstanding which makes $\Delta H^{\neq}$ greater, also makes $\Delta S^{\neq}$ greater in accordance with equation (22). It follows that in a plot according to equation (21), the

erroneous point does not deviate from the line but moves along the line. Even when the experimental errors are negligible, the same effect is introduced by the imperfect interdependence of $\Delta H^{\neq}$ and $\Delta S^{\neq}$, particularly when a reaction has been included differing from the others in the series (e.g. another reaction mechanism). This does not make the supposed relationship, equation (21), worse: on the contrary, it is more accurate. The problem is not restricted to kinetics. It applies also to equilibrium processes, when ΔH° is not determined calorimetrically but from the temperature dependence of the equilibrium constant and the data are not still treated by equation (9) but by the same procedure as outlined here.[58] (another possibility may be that ΔH° is determined simultaneously with K from titration calorimetry, see the next section.)

A statistically correct solution can be achieved by returning to the original experimental quantities, viz. by substituting $\Delta S^{\neq}$ from equation (21) into equation (22). We obtain

$$\log k = (\Delta H^{\neq}/2{\cdot}303\boldsymbol{R})(\beta^{-1} - T^{-1}) + \text{constant} \qquad (23)$$

The value of T on the left-hand side of equation (22) can be included in the constant and the immaterial difference between the Arrhenius theory and theory of the activated complex is thus removed. In the coordinates T^{-1} and log k, equation (23) represents a family of straight lines with different slopes $-\Delta H^{\neq}/2{\cdot}303\boldsymbol{R}$: all lines intersect in one point at $T = \beta$. A linear dependence in the coordinates $\Delta H^{\neq}$ and $\Delta S^{\neq}$ is thus mathematically strictly equivalent to a common point of intersection in the coordinates log k and T^{-1}. The results of this test are often surprising. For a cyclization reaction of 1,3,5-trinitrobenzene with substituted 1-phenyl-1,3-butanediones a good linear relationship was obtained,[52] Figure 6(A). In the correct plot, Figure 6(B), no common point of intersection is apparent. Note that the lines for the 4-Cl and 4-Br substituents behave as outliers in Figure 6(B). However, in Figure 6(A) they do not deviate from the line but are shifted along the line: they make the apparent relationship only more trustworthy.

A least-squares solution of the problem is not trivial. Let us consider several regression lines with a certain number of points on each. They are supposed to intersect in a common point. The task is to determine the two coordinates of this point and the slopes of all the lines, in order to achieve the minimum sum of squares of the deviations of all points. For a simple regular pattern of points, the problem was solved algebraically,[40] and in a general case by successive approximations.[43] Several hypotheses can be also tested:[58] whether the lines are parallel ($\Delta H^{\neq}$ approximately constant) or whether they intersect on the y-axis ($\Delta S^{\neq}$ constant). The original programs were written[58] in the autocode of an HP 9820 desk calculator. At least two programs have been written[59,60] in FORTRAN for use on a PC but they have not been published. In the present author's opinion, there is no advantage in using the approximate solution[38] since its reliability in particular cases is unknown.

From the foregoing discussion, one must not conclude that a valid IKR does not exist. When a correct statistical treatment was used, many cases of its validity were revealed,[58] but the slope β was often different from that obtained from an incorrect $\Delta H^{\neq}/\Delta S^{\neq}$ plot. In one case at least,[61] the point of intersection lies within the interval of experimental temperatures so that the validity of IKR is seen on a graph such as Figure 6(B) without any statistics. In many more cases, however, the accuracy of kinetic measurements is not sufficient; then one cannot reject the IKR or cannot even reject a hypothesis that all $\Delta H^{\neq}$ or all $\Delta S^{\neq}$ are equal.

The mathematical problem of several regression lines with one point of intersection is more general. It was encountered outside solution kinetics and the fundamental

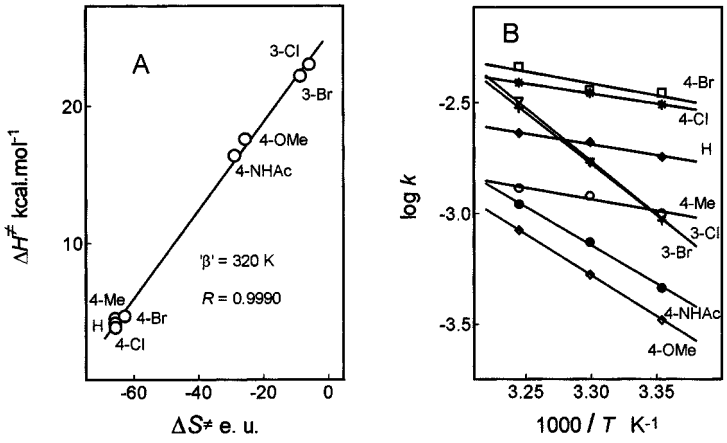

Figure 6. Isokinetic relationship in a series of related reactions (substituted 1-phenyl-1,3-butanediones with 1,3,5-trinitrophenol): (A) incorrect plot (Ref. 52) of $\Delta H^{\neq}$ vs $\Delta S^{\neq}$ with the apparent slope 'β'; (B) correct plot of the original experimental quantities log k vs T^{-1}; in correspondence with Figure 6(A), the lines should intersect at $1000/T = 3{\cdot}12$

mistake was repeated many times in different forms. Where the conductivity of semiconductors is concerned, a relationship such as equation (21) is called the Meyer–Neldel rule.[62] In contradistinction to kinetics, the experiments are accurate and the temperature interval is wide. It follows that the straight lines in the coordinates conductivity vs T^{-1} are determined exactly but they need not intersect exactly in one point. In heterogeneous catalysis IKR is called the Cremer law.[50,63] Further problems which need similar statistical treatment concern the viscosity of solutions,[64] diffusion in solids,[65] thermionic emission of electrons,[66] determination of molecular weight by centrifugation,[67] characteristic temperature of polymers[68] and chromatography of certain series of drugs.[69] In most cases, the proper statistical problem has not been recognized.

SYSTEM OF NONLINEAR REGRESSIONS

The following problem is similar to that in the preceding section but somewhat more complex. A linear dependence was claimed between $\Delta H°$ and $\Delta S°$ in complex formation within a series of amino acids,[6] dextrins[70] or various cryptands[71] according to equation (7). Compared with the examples treated in that section, however, there is a difference in the origin of the two variables. The values of $\Delta H°$ and $\log K$ ($\Delta G°$) were obtained together for each compound from titration calorimetry, then $\Delta S°$ was calculated according to equation (8) and a plot of $\Delta H°$ vs $\Delta S°$ constructed according to equation (7). In a titration calorimetric experiment,[72,73] the heat Q evolved in the volume V was measured as a function of the concentrations c_X and c_Y. The calculation is identical with that in the spectrometric determination: the difference from equation (19) is that c_Y cannot be neglected in comparison with c_X:

$$\frac{\Delta H}{K}=\frac{Vc_Xc_Y\Delta H^2}{Q}-(c_X+c_Y)\Delta H+\frac{Q}{V} \tag{24}$$

This statistical problem has been solved. From the dependence of the experimental Q on the concentrations c_X and c_Y, the parameters $\Delta H°$ and $\log K$ were estimated by nonlinear regression;[74] their strong mutual dependence in certain cases was shown in a graph similar to Figure 4 or 5(B). This analysis[74] is correct; it could only be improved by choosing an acceptable value of SD_a according to equation (12).

When a linear dependence between $\Delta H°$ and $\Delta S°$ is claimed,[6,70,71] a similar problem arises as in the isokinetic relationship. The two dependent parameters must not be introduced into equation (7). On the contrary, $\Delta S°$ calculated from this equation must be introduced into equation (24). When $\Delta G°$ is converted into $\log K$ we obtain

$$\Delta H\exp\left[\frac{\Delta H(\beta-T)+C}{\beta RT}\right]=\frac{Vc_Xc_Y\Delta H^2}{Q}-(c_X+c_Y)\Delta H+\frac{Q}{V} \tag{25}$$

The task is to estimate the two general parameters, β and C, and in addition $\Delta H°$ of all reactions of the series. This means processing of the immediate experimental results (c_X, c_Y and Q) of all titration calorimetric experiments at once. A least-squares solution is possible but would evidently be much more complex than with equation (23). For a recalculation, the original titration data would be needed but they are not given in the literature.[70,71] In any case, there is no doubt that the linear dependences reported are artifacts but one cannot decide at present whether they should possess a different slope or whether no relationship exists at all.

MULTIPLE LINEAR REGRESSION

The regression equation in a multiple linear regression has the form:

$$y=a+b_1x_1+b_2x_2+\cdots \tag{26}$$

The conditions are analogous to those for simple regression. In the literature much attention was given to the stochastic dependence existing in some cases between two or more explanatory variables (called intercorrelation or internal dependence, respectively).[75] When, for instance, x_1 and x_2 are strongly correlated, one obtains b_1 and b_2 with great uncertainty. However, the overall fit is not depreciated. The regression equation can be useful for predicting the y values but the individual terms and the regression coefficients b cannot be interpreted. More important is probably a proof that all terms in equation (26) are actually necessary. A fully convincing procedure is to compare with each simpler equation in which one term has been omitted. The significance of each term is then proved by an F-test.

Difficulties are encountered with graphical representation. A completely correct representation in two dimensions is, in fact, not possible, even with only two explanatory variables. An evidently erroneous attempt[76,77] was to divide equation (26) by x_2 to obtain:

$$y/x_2=b_1(x_1/x_2)+b_2 \tag{27}$$

In this procedure one parameter has been lost: a necessary assumption was that a equals zero. Equation (27) was treated as a simple regression with the variables x_1/x_2 and y/x_2; it was represented graphically, and b_1 and b_2 were either calculated from equations for simple regression or estimated from the plot. They were then used in equation (26) with $a=0$. The whole procedure is evidently a great mistake: the main defect is that equation (27) expresses partly the dependence of x_2 on itself. This defect is the greater the more important is x_2 in the multiple regression. In one example,[77] the following dependence on the two variables, denoted σ_N and σ_S, was obtained in this way:

$$\log k=(0{\cdot}76\pm0{\cdot}01)\sigma_N-(0{\cdot}47\pm0{\cdot}01)\sigma_S \quad (\text{'}R\text{'}=0{\cdot}993) \tag{28}$$

The apparent correlation coefficient 'R' is that of the simple regression, equation (26). Instead, the correct expression, obtained from multiple regression, reads (with R from the

multiple regression)

$$\log k = (0{\cdot}88 \pm 0{\cdot}09)\sigma_N + (0{\cdot}25 \pm 0{\cdot}15)\sigma_S \quad (R = 0{\cdot}985) \quad (29)$$

Note that the uncertainties in the two regression coefficients are very different; the term with σ_S is evidently not statistically significant. The two equations can be compared by another kind of graphical representation in which y is plotted vs $(x_1 + x_2 b_2/b_1)$ (Figure 7). The preference of equation (29) is evident [Figure 7(B)]. Since the term with σ_S is insignificant, simple regression with σ_N alone is also sufficient [Figure 7(C)].

A plot as shown in Figure 7(A) and (B) is acceptable provided that b_1 and b_2 have been predetermined by multiple regression: it can then give a true picture of the fit achieved (correlation coefficient $R_{1\cdot23}$) but gives no idea about the relative importance of the two variables (correlation coefficients $R_{12\cdot3}$ and $R_{13\cdot2}$) and about the uncertainty of the regression coefficients. Such plots were even suggested for searching the best value of b_2/b_1 by attempting successively various values.[78] There is evidently no reason to do this when programs for multiple regression are easily available. The same criticism applies to suggestions for determining the regression coefficients by graphical procedures.[79]

CORRELATIONS WITH A HIDDEN VARIABLE

In a classical textbook on quantum chemistry,[80] a linear dependence is presented between calculated delocalization energies *DE* (in β units) and experimental resonance energies *RE*. Calculations were carried out within the framework of simple Hückel method and the correlation was excellent for benzenoid hydrocarbons from benzene to perylene. We calculated $R = 0{\cdot}994$ and obtained Figure 8(A). There are almost no objections from the point of view of statistics. The explanatory variable is exact; the response function may be viewed as loaded with a random error. The purely experimental error certainly increases with x but it is not the main source of deviation. There is evidently a strong correlation between the two variables in Figure 8(A). Nevertheless, the conclusion is not fair that calculations of *DE* give a true picture of the strength of conjugation in the given molecules. The problem is that both *RE* and *DE* depend on the molecular weight *M*. When the enthalpy of combustion is determined experimentally, one obtains first the specific value related to 1 g. This must be multiplied by the (known or assumed) *M* to obtain the molar enthalpy of combustion and from that *RE*. On the other hand, calculated *DE* depends strongly on the number of π-electrons and hence, in the case of benzenoid hydrocarbons, directly on *M*. When the two variables in Figure 8(A) are divided by *M*, one obtains Figure 8(B), showing no correlation. (See also the difference between the two regression lines, one forced through the origin, the other not.) The correct interpretation of Figure 8(B) is that *DE* is unable to predict the experimental specific resonance energy. On the other hand, the interpretation of Figure 8(A) should be that *RE* depends strongly on the size of the molecule and so also does the calculated *DE*. In this sense we can call *M* a 'hidden variable.' Note that a very good correlation is obtained even by plotting *RE* vs the number of π-electrons (not shown).

A more recent example[81] is shown in Figure 9. Here some non-benzenoid compounds are included (butadiene, cyclopentadiene and azulene) and the difference between the two plots is still more striking. For the aliphatic compounds the theory thus yields results completely different from those for benzenoid hydrocarbons; azulene stays in the middle. Benzenoid hydrocarbons are situated practically in one point. (Note also the different values of R when the regression line is forced through the origin.) The result is that the significance of resonance energies, either experi-

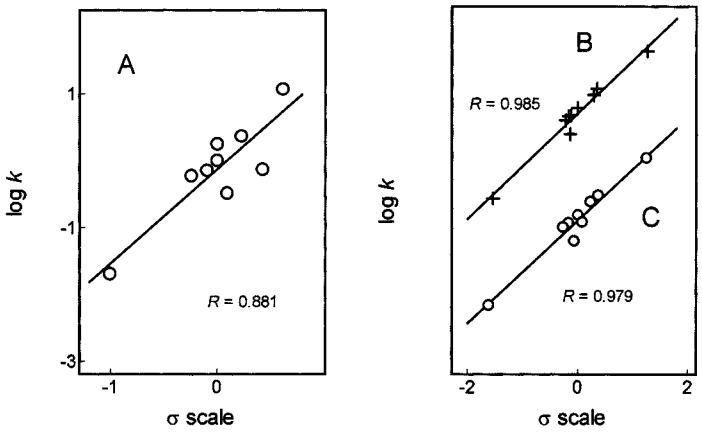

Figure 7. Attempted representation of multiple linear regression in two dimensions, hydrolysis of substituted phenylazobenzothiazolium dyes: (A) plot of log k vs a false blend of constants σ which was obtained incorrectly (Ref. 77), equation (28) based on equation (27); (B) plot vs the correct blend of σ obtained from equation (29) based on the multiple regression equation (26); (C) plot vs the constants σ_N as defined in Ref. 77. Correlation coefficients given in (A) and (B) are those from the multiple regression, $R_{1\cdot23}$

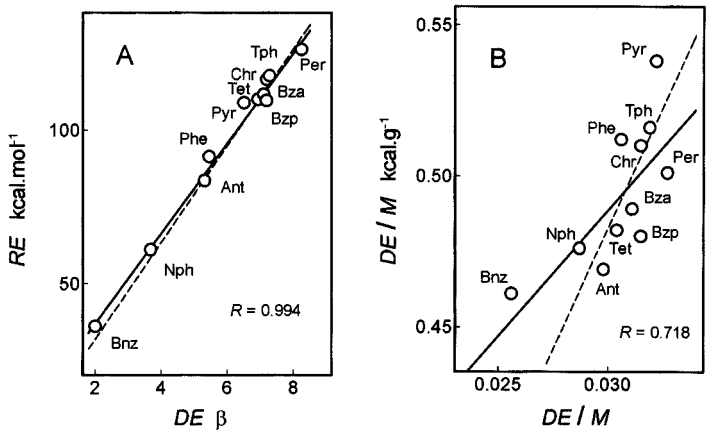

Figure 8. Quantities dependent on the molecular weight: (A) plot of the experimental molar resonance energy of aromatic hydrocarbons vs the calculated delocalization energy according to Ref. 80; (B) plot of specific resonance energies related to 1 g. Broken lines and regression coefficients *R* correspond to common regression; full lines were forced through the origin

mental or from quantum chemical calculations, must not be overestimated. (The determination of *RE* is not purely experimental; it assumes additivity of enthalpies of formation and its transferability from one compound to another.)

Several molar quantities dependent on *M* are well known[82] and have been extensively exploited, mainly several decades ago. The molecular weight is in these cases not much 'hidden' since it is involved explicitly in the defining equation. Examples are the molar refraction *MR* or the parachor *P*, which are functions of the refractive index *n* or surface tension γ, respectively:

$$MR = \frac{n^2 - 1}{n^2 + 2} d^{-1} M \tag{30}$$

$$P = \gamma^{1/4} d^{-1} M \tag{31}$$

The structural dependence of these quantities was always treated in terms of additivity for atoms or groups, and certain specific corrections. With respect to the preceding examples, one can have doubts as to whether the additive character is not only (or mainly) due to the exact additivity of the molecular weight.[82] The problem can be demonstrated

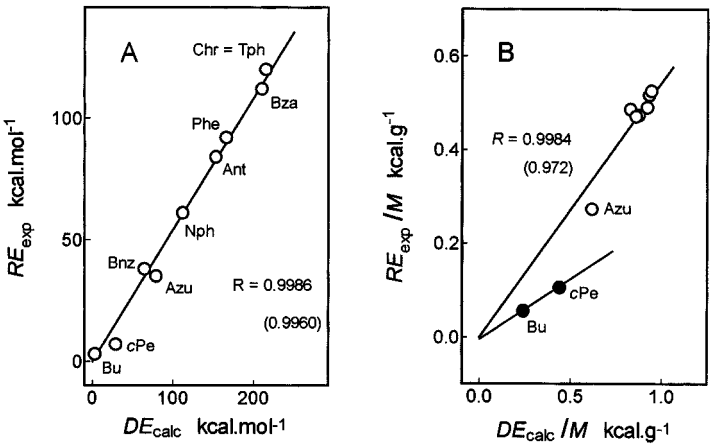

Figure 9. Resonance energies of aromatic and conjugated hydrocarbons: (A) plot of experimental vs calculated molar quantities according to Ref. 81; (B) plot of specific resonance energies related to 1 g. Regressions were forced through the origin (when not forced, the correlation coefficients are given in parentheses)

on homologous series of organic compounds which always served as a basis for determining the increments of individual groups. For instance, the parachor in a homologous series depends linearly on the number of carbon atoms, and thus also on M:[83]

$$\gamma^{1/4}d^{-1}M=\frac{p(CH_2)}{m(CH_2)}M+\frac{m(CH_2)p(X)-m(X)p(CH_2)}{m(CH_2)} \quad (32)$$

where $p(CH_2)$ and $p(X)$ are the increments of the CH_2 group and of the constant part of the molecule, respectively; $m(CH_2)$ and $m(X)$ are molecular weights of these groups. The dependence of P on M is represented by a family of parallel straight lines,[83] one for each homologous series, as in Figure 10(A). When the lines are not exactly parallel, the differences in slopes are not clearly seen. The correlation is evidently due also to the dependence of M on itself. By dividing equation (32) by M we obtain

$$\gamma^{1/4}d^{1-1}=\frac{p(CH_2)}{m(CH_2)}+\frac{m(CH_2)p(X)-m(X)p(CH_2)}{m(CH_2)}M^{-1} \quad (33)$$

This equation represents a family of straight lines intersecting in one point on the y-axis. Figure 10(B) reveals that this is not fulfilled and the parachor is not a truly additive quantity. An approximate additive character was simulated by the additivity of molecular weight and the whole parachor story was simply a statistical mistake.[83] When some correct conclusions were drawn from the parachor values, they could be obtained also from the molar volume: the measured surface tension γ has no positive effect on the additive character. Since the values of surface tension have always been discussed in terms of parachor, almost nothing is known about the structure dependence of this quantity.

In the same way, the additive relationships for other quantities were tested. The simple molar volume is additive with reasonable accuracy,[84] particularly for monofunctional compounds except the first members of each homologous series. Contrary to the literature claims, the molar volume at the boiling point is not additive.[84] The same conclusion concerns the so-called rheochor and further additive quantities based on viscosity,[85] also the so-called molar refractive index.[86] The last quantity was particularly simple, the refractive index n_D^{20} multiplied by the molecular weight. When calculated from tabulated increments for a given structure, it should give a rough estimate[87] of n_D^{20}. In contrast, molar refraction, equation (30), has remained as an example of an actually additive quantity with very good accuracy:[82] introducing the refractive index into the molar volume improves very significantly the additive character and extends the range of validity.

The above tests and their results will not be reproduced here, but a new example is shown in Figure 10. In a recent review,[88] the thermodynamic quantities $\Delta_fH°$(g), $\Delta_fH°$(l) and $\Delta_fH°$(s), $S°$(g), $S°$(l) and $S°$(s), $C_p°$(g), $C_p°$(l) and $C_p°$(s) were calculated as additive but with numerous specialized values of increments corrupting the additive character. A test of $\Delta_fH°$(l) on 16 homologous series (125 compounds) is shown in Figure 11(A) and (B): the additivity is very good. As expected, the fit in the correct regression, Figure 11(B), is worse (median value of R 0·9987) than in Figure 11(A) (0·99986) but it is still very good. The increment of the CH_2 group was derived within the framework of the correct regression, Figure 11(B), from the median value of all intercepts: we obtained $-25{\cdot}42$ kJ mol^{-1}. From the regressions in Figure 11(A), viz. from the median value of all slopes, one would obtain $-25{\cdot}35$ kJ mol^{-1}. These figures show that the results of correct and incorrect statistical procedures become very similar when the fit is very close. In fact, the correlation coefficients in Figure 11(B) are not

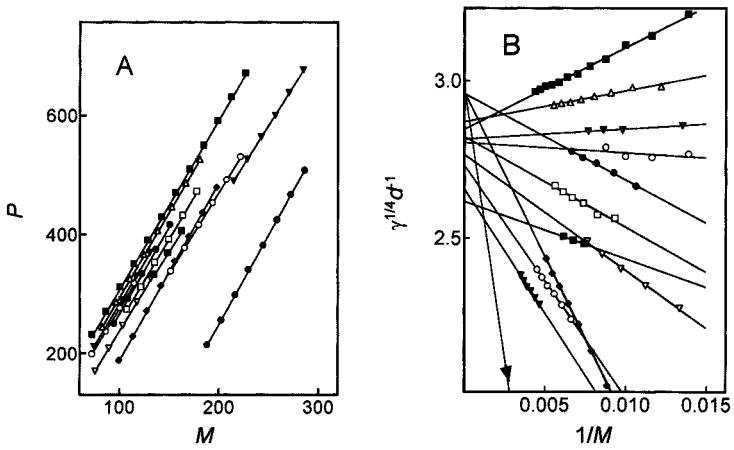

Figure 10. Molar additive properties: testing the parachor on 12 homologous series according to Ref. 83. (A) Plot of $P=\gamma^{1/4}d^{-1}M$ vs the molecular weight, equation (32); (B) plot of the 'specific parachor' $\gamma^{1/4}d^{-1}$ vs the reciprocal molecular weight, equation (33). Data mostly from Ref. 87 and earlier papers by the same author

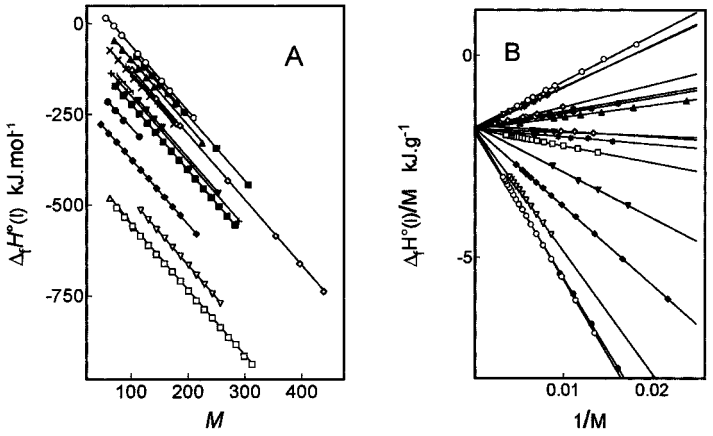

Figure 11. Molar additive properties: enthalpy of formation in the liquid state in 16 homologous series (data from Ref. 88). (A) Plot of molar quantities, $\Delta_f H°(l)$, vs the molecular weight; (B) plot of specific quantities, related to 1 g, vs the reciprocal molecular weight

informative. When the slope of a straight line is near to zero, R is also small. Nevertheless, this straight line has the same positive importance for the overall fit as the others. The proper statistics in this case are the standard deviation SD, and its ratio to s_0, the mean square deviation of all experimental values from their mean:

$$\psi = SD / s_0 \tag{34}$$

This dimensionless characteristic has been suggested[82] for comparing the accuracy of various relationships (particularly of relationships other than linear regression) for different quantities, i.e. differing in their magnitude and/or in the physical dimension. (One can, for instance, compare the accuracy of an empirical relationship for the refractive index with that for the boiling point.) Of course, one can use also the characteristics SD^2/s_0^2 or $1 - SD^2/s_0^2$ with the same result. For our example (Figure 11), we obtained $\psi = 0{\cdot}0068$ in Figure 11(B) and an apparent, seemingly slightly better, value $\psi = 0{\cdot}0052$ in Figure 11(A).

In the same way, we further tested $C_p^0(g)$ on 15 homologous series (116 compounds) and obtained $\psi = 0{\cdot}026$ in the correct plot; the apparent ψ would be $0{\cdot}0045$. The greater difference and worse value of ψ in this case is due to the fact that $C_p^0(g)$ values are fairly close for different compounds of the same size of molecule. The increment of the CH_2 group is $22{\cdot}95$ kJ mol^{-1} and the apparent value in the incorrect plot is $22{\cdot}86$ kJ mol^{-1}. Some molar additive quantities named here may seem of little importance today. On the other hand, new such quantities were suggested recently (in addition to the mentioned thermodynamic quantities[88]) particularly for molar enthalpies of vaporization,[89,90] or for partial molar volumes in solution.[91]

A different case of a quantity dependent on the size of molecule was claimed recently for ^{13}C NMR shifts in alkanes.[92] The sum of shifts of all carbon atoms, $\Sigma\delta$, was correlated with the parameter L defined as a function of three topological indices p_i:

$$\Sigma\delta = kL(+C) \tag{35}$$

$$L = 2p_1 + p_2 - p_3 - 2 \tag{36}$$

The 'molecular path counts' p_i denote the number of possible paths consisting of 1, 2 or 3 adjoining bonds in the given molecule.[92,93] The dependence obtained is shown in Figure 12(A). The fit is very good but still somewhat worse for branched alkanes; the intercept is negligible. Again, there are no objections against using the regression model. However, the size of the molecule may be the hidden independent variable. The value of $\Sigma\delta$ certainly depends more on the number of carbon atoms than on their character. The dependence of L is less clear. For straight-chain alkanes, L is a simple function of the number of carbon atoms $L = 2N - 3$. For branched alkanes, L increases with increasing degree of branching and depends on the number of atoms in a more complex way, which cannot be reproduced by an equation. When we divide both coordinates by the number of carbon atoms N, we obtain Figure 12(B). The variable $\Sigma\delta/N$ represents the mean value of the shift; L/N has no clear meaning. The fit may still seem good but is due to the straight-chain alkanes; for branched compounds alone it is much worse. Note also that the shifts δ are defined with respect to an arbitrary reference compound and any correlation should be valid irrespective of which reference was chosen. The reference compound used in this case[92] was tetramethylsilane. When the shifts are referenced to benzene, the relationship, equation (35), breaks down completely (not shown). Originally, the whole concept was applied only to isomeric hydrocarbons of the same molecular weight,[93] hence all the problems were avoided.

EFFECT OF ORDERING

This very interesting statistical problem[94] concerned photoelectron spectroscopy but can be encountered in any kind of

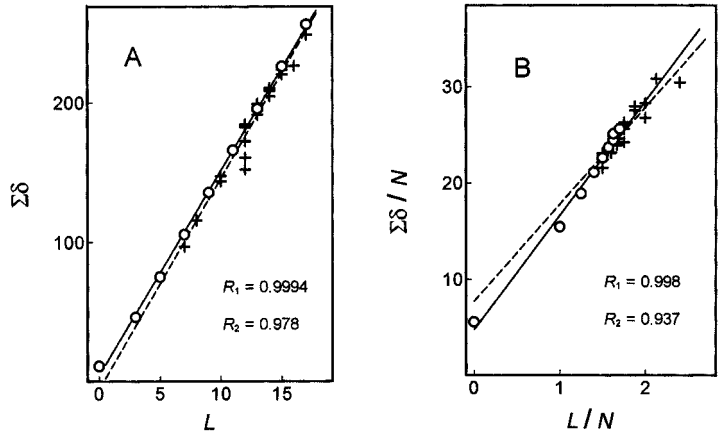

Figure 12. Quantities dependent on the molecular size. (A) Plot of the sum of ^{13}C NMR shifts of all C atoms vs a function of topological indices L according to Ref. 92: ○, C_2–C_{10} straight-chain hydrocarbons (correlation coefficient R_1); +, C_4–C_8 branched hydrocarbons (correlation coefficient R_2). (B) The same plot with quantities related to one carbon atom (divided by the number of carbon atoms N)

spectroscopy and even in other fields. A series of experimental numbers (e.g. ionization energies I in a photoelectron spectrum) are often compared with the results of a theoretical treatment in the form of a correlation diagram such as Figure 13(A). When the theory agrees with the experiment, the two given lines should always coincide. Since many theories are loaded with certain systematic errors, it is currently accepted as sufficient when the theoretical and experimental figures are merely linearly dependent. In this case the proper representation is a linear regression as in Figure 13(B). According to this regression, the theoretical values may be empirically linearly corrected (scaled) and then plotted in the correlation diagram, as has been done in Figure 13(A).

Heilbronner and Schmelzer[94] have discovered a shortcoming in the whole procedure. The problem is that the theoretical values belong each to a certain, exactly defined transition but the experimental values are not so assigned. One must assume that the first signal belongs to the first transition, the second signal to the second transition, etc. This is, however not certain. It is possible in principle that the calculations are completely wrong and the last observed transition belongs, say, to the first calculated. In other words, the experimental values have been ordered to form a sequence corresponding to the theoretical values. A mistake which may arise was demonstrated in a dramatic way: a series of experimental values were correlated with a series of random numbers assigned successively to the individual signals.[94] In Figure 13(D) we used our own series of random numbers and obtained $R = 0{\cdot}9816$; in the original paper[94]

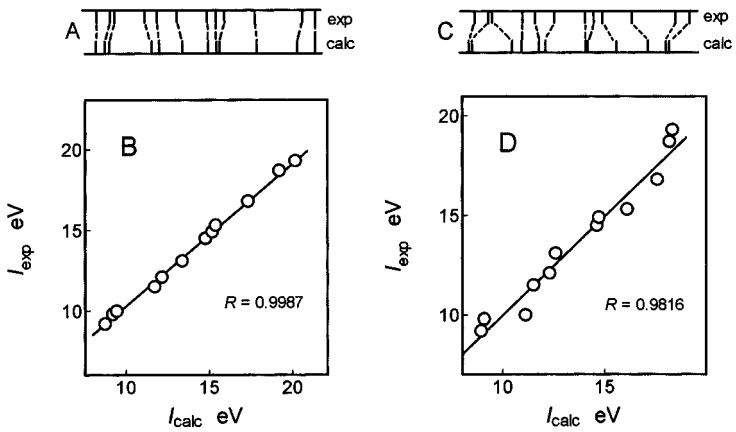

Figure 13. Comparison of calculated and experimental ionization energies of phosphabenzene (according to Ref. 94): (A) calculated in Ref. 95 and scaled; (B) calculated in Ref. 95 and not scaled; (C) scaled random numbers as 'calculated'; (D) not scaled random numbers as 'calculated'

$R = 0{\cdot}983$ was obtained with another set of random numbers. The 'theoretical' values were then linearly scaled and plotted in the correlation diagram in Figure 13(C). At first sight, the fit may be estimated as sufficient but compared with Figure 13(A) it is evidently much worse. It is also this fit which enables the real and apparent correlations to be distinguished. The authors calculated possible values of R by a Monte Carlo approach and gave the critical values for different significance levels.[94] A regression can be considered real if the critical value of R is exceeded. The heart of the problem is ordering of one variable before regression: this introduces a certain kind of dependence of the two variables. This problem may be encountered always when a set of data has been obtained simultaneously from one experiment: the whole set has a clear physical meaning but the individual items are not assigned.

DISCUSSION

Errors of incorrect statistics

The foregoing examples are rather different in character and represent only a fraction of all possible errors. Therefore, only few general conclusions can be drawn. While it is certain that a correct method must yield correct results, it is not evident what an incorrect method could yield. We have encountered wrong results, completely at variance with the original experiments, but also cases where the errors did not exceed the uncertainty of rounding-off. This depends not only on the kind of the statistical error but also strongly on the particular data.

Two limiting cases may be recognized. When the data are very accurate and the model fits very well, the statistics are changed into simple algebra and all transformations are allowed. Then correct and incorrect statistical methods may give the same result. In the opposite case, when the experimental error is great and/or the model inadequate, it may happen that no significant results are obtainable at all and no statistical method can help. Cases treated in this paper are in between. In them, no general rules can be formulated as to when the error will be great and when not: one must simply avoid the incorrect procedures as far as possible. The practical tasks can be divided into the following four categories.

Avoiding the mistakes

In this era of computers, there is no reason for a formal linearization of nonlinear equations. All such attempts [e.g. equations (13), (15), (16) and (18)] can be *a priori* rejected. More difficulties can arise with apparently simple quantities, denoted with a simple symbol, which are in fact composite [selectivity factor, equation (1)]. Even with well defined quantities of clear physical meaning, it may be of importance how they have been experimentally determined and how the experimental error is propagated [enthalpy and entropy in equation (7)]. Certain apparently simple procedures can be particularly dangereous when one does not recognize that any nontrivial statistical procedure has been applied [isokinetic relationship, equation (21)]. In this connection, one can ask the question whether even simple plotting one quantity against another can in some cases be 'forbidden.' A pure plot is certainly unobjectionable, whatever the way in which the variables have been obtained. However, representing the points by circles of a given diameter implies an idea about their uncertainty and particularly assumes independence of the errors in x and in y. This can sometimes be badly misleading. When a line is drawn through an array of points, it is clearly a statistical procedure. It implies that the points could lie on the line exactly under certain conditions, e.g. if there were no experimental errors or if a secondary factor influencing y were absent. This can sometimes be a serious mistake.

We believe that all the mistakes mentioned can be avoided with proper attention. More difficult is the proper interpretation for which no general rules can be recommended. Every case must be considered separately. In particular, one should stress that regression or correlation does not imply a causal dependence.

Revealing the mistakes in the literature

Using the criteria in the preceding section, one can recognize a treatment that is incorrect in principle. For a proof, an efficient method has also already been mentioned. It is necessary to calculate the original experimental values as they should be to satisfy exactly the required theory or correlation. Sometimes, one must go back to the experiment in several steps [see, for instance, equation (25)]. Comparison with actual experimental values may in some cases be shocking [Figure 3(B)]; in most cases it is so convincing that no special statistical tests are needed. If not, one can calculate the differences $\Delta = y(\text{experimental}) - y(\text{calculated})$ and test their distribution: zero mean value, approximately normal distribution, most important their independence of x. In our opinion, this processing may be sufficient in all cases.

Proposing correct statistical procedures

Within the framework of the least-squares method, this task is solved by equation (11) for linear or nonlinear regression, and even for systems of more equations. The mathematical solution can be more or less complex in individual cases but is always only a question of the computer program. Data in experimental chemistry are usually not too numerous (as they are, for instance, in crystallography), hence the problem of quick convergence is not so important. Very often, simple programs are sufficient, following *SD* as a function of a parameter which is given successively all possible values. In this way also the uncertainties of the parameters and their dependence are objectively represented [Figures 4 and 5(B)].

All methods recommended in this paper and all conclu-

sions have been based on the least-squares method. This approach is certainly suitable when the deviations are caused entirely by experimental errors or by any other random variable. However, in many chemical theories they are due merely to an imperfect model, i.e. to an additional factor which was not discovered. In statistical terms we may say that an additional explanatory variable was omitted. This variable does not possess a normal distribution and is not even random. Then robust methods,[96] which suppress the effect of outliers and concentrate on the known explanatory variable, may be more efficient. In other cases, attention is focused on predicting the response function: then methods may be preferred that avoid serious mistakes even at the cost of biased results. This all depends on the particular requirements in the given case. In our opinion, the least-squares method was sufficient in all cases treated in this paper. Note that several quoted procedures[36,37,79] were outside this method. They must not be considered as 'worse' for this reason, they can only be less suitable for a given problem. Most of the procedures criticized here were originally based on the least-squares method but yielded ultimately something which was not the least-squares estimate.

Recalculating or appraising literature data

Surprisingly, this task may be the most difficult. Literature data can be mostly recalculated with not too great effort when the original experiments are given. This is, however, seldom the case in the recent literature. Particularly with methods commonly used, even the general features (number of the measurements, range of concentrations, etc.) are often lacking. The parameters published may be viewed with suspicion but their actual reliability remains unknown (see, for instance, Ref. 9). Moreover, such parameters are transferred into further literature and may appear in reviews together with more reliable data. A remedy may be expected in publishing original data systematically on the Internet. In the reviews, one should at least specify the method by which any value has been obtained; methods recognized as incorrect could be particularly noted.

CONCLUSIONS

Statistical errors are common in the literature, both older and contemporary. Some examples given here may seem to be merely of historical interest but the results live on further in the secondary literature. For some broadly used methods, correct solutions were given some time ago and are being used more and more. On the other hand, the first statistical problems are just being encountered in some newly developing fields. In our opinion, one should give proper attention particularly to cases in which a statistical problem is not evident at first sight.

ACKNOWLEDGEMENT

The work was supported by the grant No. 203/96/1658 of the Grant Agency of the Czech Republic.

REFERENCES

1. H. C. Brown and C. R. Smoot, *J. Am. Chem. Soc.* **78**, 6255 (1956).
2. O. Exner, *Org. React. (Tartu)* **21**, 3 (1984).
3. For instance: H. Mager, *Moderne Regressionsanalyse*, Chapt. 1.1. Otto Salle Verlag, Frankfurt am Main (1982).
4. O. Exner and K. Zvára, *Chemom. Intell. Lab. Syst.*, submitted.
5. O. Exner, *Collect. Czech. Chem. Commun.* **38**, 799 (1973).
6. E. Grunwald and C. Steel, *J. Am. Chem. Soc.* **117**, 5687 (1995).
7. T. W. Zawidzki, H. M. Papée, W. J. Canady and K. J. Laidler, *Trans. Faraday Soc.* **55**, 1725 (1959).
8. O. Exner, *Prog. Phys. Org. Chem.* **10**, 411 (1973).
9. O. Exner, *Chemom. Intell. Lab. Syst.* (1997), in press.
10. O. Exner, *Dipole Moments in Organic Chemistry*, Chap. 7.1. Georg Thieme, Stuttgart (1975).
11. K. Bauge and J. W. Smith, *J. Chem. Soc.* 4244 (1964).
12. C. Trainer, J. F. Skinner and R. M. Fuoss, *J. Phys. Chem.* **68**, 3406 (1964).
13. O. Exner, *J. Mol. Struct.* **216**, 153 (1990).
14. O. Exner, *Collect. Czech. Chem. Commun.* **55**, 1435 (1990).
15. J. A. Walmsley, E. J. Jacob and H. B. Thompson, *J. Chem. Phys.* **80**, 2745 (1976).
16. A. Koll, M. Rospenk, L. Stefaniak and J. Wójcik, *J. Phys. Org. Chem.* **7**, 171 (1994).
17. A. V. Few and J. W. Smith, *J. Chem. Soc.* 2781 (1949).
18. G. R. Saad, M. M. Naoum and H. A. Rizk, *Can. J. Chem.* **67**, 284 (1989); **68**, 480 (1990).
19. V. Shanmugasundaram and M. Meyyapan, *Indian J. Chem.* **10**, 936 (1972).
20. O. Exner, *Can. J. Chem.* **70**, 1873 (1992).
21. P. J. Bauer, O. Exner, R. Ruzziconi, T. D. An, C. Tarchini and M. Schlosser, *Tetrahedron* **50**, 1707 (1994).
22. H. A. Benesi and J. H. Hildebrand, *J. Am. Chem. Soc.* **71**, 2703 (1949).
23. N. J. Rose and R. S. Drago, *J. Am. Chem. Soc.* **81**, 6138 (1959).
24. R. Bhattacharya and S. Basu, *Trans. Faraday Soc.* **54**, 1286 (1958).
25. K. Conrow, G. D. Johnson and R. E. Bowen, *J. Am. Chem. Soc.* **86**, 1025 (1964).
26. D. R. Rosseinsky and H. Kelawi, *J. Chem. Soc. A* 1207 (1969).
27 G. Carta and G. Crisponi, *J. Chem. Soc., Perkin Trans. 2* 53 (1982).
28. G. Carta, J. *Chem. Soc., Perkin Trans.* 2, 1219 (1992).
29. C. Laurence, G. Guihéneuf and B. Wojtkowiak, *J. Am. Chem. Soc.* **101**, 4793 (1979).
30. P. Maslak and W. H. Chapman, *J. Org. Chem.* **55**, 6334 (1990).
31. L. Michaelis and M. L. Menten, *Biochem. Z.* **49**, 333 (1913).
32. M. Hamala and E. Paulinyova, *Drev. Výsk. (Bratislava)* **28**, 29 (1983).
33. G. A. Sagnella, *Trends Biochem. Sci.* **10**, 100 (1985).
34. H. Lineweaver and D. Burk, *J. Am. Chem. Soc.* **56**, 658 (1934).
35. G. Scatchard, *Ann. N. Y. Acad. Sci.* **51**, 660 (1949).

36. R. Eisenthal and A. Cornish-Bowden, *Biochem. J.* **139**, 715 (1974).
37. J. Koštiř, *Chem. Listy* **79**, 989 (1988).
38. W. Linert, *Chem. Soc. Rev.* **23**, 429 (1994).
39. O. Exner, *Collect. Czech. Chem. Commun.* **29**, 1094 (1964).
40. O. Exner, *Collect. Czech. Chem. Commun.* **37**, 1425 (1972).
41. S. Wold, *Chem. Scr.* **2**, 145 (1972).
42. W. Linert, R. W. Soukup and R. Schmid, *Comput. Chem.* **6**, 47 (1982).
43. O. Exner and V. Beránek, *Collect. Czech. Chem. Commun.* **38**, 781 (1973).
44. R. R. Krug, W. G. Hunter and R. A. Grieger, *J. Phys. Chem.* **80**, 2335 and 2341 (1976).
45. V. I. Shimulis, *Kinet. Katal.* **24**, 715 (1983).
46. H. J. Chen and S. P. Yuang, *Commun. Statist.-Theor. Methods* **11**, 395 (1982).
47. J. E. Leffler, *J. Org. Chem.* **20**, 1202 (1955).
48. R. F. Brown, *J. Org. Chem.* **27**, 3015 (1962).
49. P. Beltrame and M. Simonetta, *Gazz. Chim. Ital.* **89**, 495 (1959).
50. G. C. Bond, *Catalysis by Metals*, p. 140. Academic Press, London (1962).
51. R. Lumry and S. Rajender, *Biopolymers* **9**, 1125 (1970).
52. L. M. Gnanados and D. Kalaivani, *J. Org. Chem.* **50**, 1178 (1985).
53. S. S. Kim. S. Y. Choi and C. H. Kang, *J. Am. Chem. Soc.* **107**, 4234 (1985).
54. T. Asano, T. Okada, S. Shinkai, K. Shigematsu, Y. Kusano and O. Manabe, *J. Am. Chem. Soc.* **103**, 5161 (1981).
55. V. Dorovska-Taran, R. Momtcheva, N. Gulubova and K. Martinek, *Biochim. Biophys. Acta* **702**, 37 (1982).
56. E. M. Y. Quinga and G. D. Mendenhall, *J. Org. Chem.* **50**, 2836 (1985).
57. V. Palm and S. Kaasik, *Org. React. (Tartu)* **30**, 73 (1996).
58. O. Exner, *Collect. Czech. Chem. Commun.* **40**, 2762 (1975).
59. P. Müller and J.-C. Perlberger, *Helv. Chim. Acta* **57**, 1943 (1974).
60. O. Pytela, M. Večeřa and P. Vetešnik, *Collect. Czech. Chem. Commun.* **46**, 898 (1981).
61. A. B. Dekelbaum and B. V. Passet, *Org. React. (Tartu)* **11**, 383 (1974).
62. A. Yelon, B. Movaghar and H. M. Branz, *Phys. Rev. B* **46**, 12244 (1992).
63. A. K. Galwey, *Adv. Catal.* **26**, 247 (1977).
64. W. Good and J. Stone, *Electrochim. Acta* **17**, 1813 (1972).
65. J. Shinar, D. Davidov and D. Shaltiel, *Phys. Rev. B* **30**, 6331 (1984).
66. R. Vanselow, *Surf. Sci.* **149**, 381 (1985).
67. R. M. Johnsen, *Chem. Scr.* **1**, 81 (1971).
68. D. D. Eley, *J. Polym. Sci.* **17**, 73 (1967).
69. T. Cserháti and K. Magyar, *J. Biochem. Biophys. Methods.* **24**, 249 (1992).
70. B. Zhang and R. Breslow, *J. Am. Chem. Soc.* **115**, 9353 (1993).
71. A. F. Danil de Namor, M. C. Ritt, M.-J. Schwing-Weil and F. Arnaud-Neu, *J. Chem. Soc., Faraday Trans.* **87**, 3231 (1991).
72. J. J. Christensen, J. Ruckman, D. J. Eatough and R. M. Izatt, *Thermochim. Acta* **3**, 203 (1972).
73. D. J. Eatough, J. J. Christensen and R. M. Izatt, *Thermochim. Acta* **3**, 219 (1972).
74. R. Karlsson and L. Kullberg, *Chem. Scr.* **9**, 54 (1976).
75. H. Mager, P. P. Mager and A. Barth, *Tetrahedron* **35**, 1953 (1979).
76. A. C. Farthing and B. Nam, in *Steric Effects in Conjugated Systems*, edited by G. W. Gray, p. 131. Butterworths, London (1958).
77. A. P. D'Rozario, A. Williams and B. Parton, *J. Chem. Soc., Perkin Trans. 2* 1781 (1987).
78. P. R. Wells, S. Ehrenson and R. W. Taft, *Prog. Phys. Org. Chem.* **6**, 147 (1968).
79. R. S. Drago, *Applications of Electrostatic–Covalent Models in Chemistry.* Surfside Scientific, Gainesville, FL (1994).
80. A. Streitwieser, *Molecular Orbital Theory for Organic Chemists*. Wiley, New York (1962).
81. S. Behrens, A. M. Köster and K. Jug, *J. Org. Chem.* **59**, 2546 (1994).
82. O. Exner, *Collect. Czech. Chem. Commun.* **31**, 3222 (1966).
83. O. Exner, Collect. *Czech. Chem. Commun.* **32**, 24 (1967).
84. O. Exner, *Collect. Czech. Chem. Commun.* **32**, 1 (1967).
85. O. Exner, *Collect. Czech. Chem. Commun.* **32**, 4327 (1967).
86. O. Exner, *Izv. Inst. Org. Khim. Bulg. Akad. Nauk* **3**, 87 (1967); *Chem. Abstr.* **69**, 2191m (1968).
87. A. I. Vogel, *J. Chem. Soc.* 1833 (1948).
88. E. S. Domalski and E. D. Hearing, *J. Phys. Chem. Ref. Data* **22**, 805 (1993).
89. J. S. Chickos, D. G. Hesse and J. F. Liebman, *J. Org. Chem.* **54**, 5250 (1989).
90. J. P. Guthrie and K. F. Taylor, *Can. J. Chem.* **61**, 602 (1983).
91. F. Shadidi, *Can. J. Chem.* **61**, 1414 (1983).
92. Y. Miyashita, T. Okuyama, H. Ohsako and S. Sasaki, *J. Am. Chem. Soc.* **111**, 3469 (1989).
93. M. Randić, *J. Magn. Reson.* **39**, 43 (1980).
94. E. Heilbronner and A. Schmelzer, *Nouv. J. Chim.* **4**, 23 (1980).
95. W. von Niessen, G. H. F. Diercksen and L. S. Cederbaum, *Chem. Phys.* **10**, 345 (1975).
96. O. Exner, I. Kramosil and I. Vajda, *J. Chem. Inf. Comput. Sci.* **33**, 407 (1993).

