Multiple and Curvilinear Regression Analysis of Light Scattering Data

W. A. J. BRYCE

Light scattering data from a Sofica instrument are subjected to a regression analysis. The method of analysis is given in detail. The standard errors of M_w, $\langle S^2\rangle^{\frac{1}{2}}$, A_2 and $A_3Q(\theta)$ are estimated. Manual and analytical results are compared.

POLYMERS produced by an anionic mechanism may have a narrow molecular weight distribution. The width of the distribution curve is normally estimated by comparing M_w and M_n values. If the ratio is to have any significance then the standard errors of the molecular weights must be calculated.

A test of a dilute solution theory usually involves the second virial coefficient and the radius of gyration. Hence it is important that the ranges of these parameters should be known.

Three of the above parameters (M_w, A_2 and $\langle S^2\rangle^{\frac{1}{2}}$) may be obtained from light scattering experiments. Zimm[1] plots usually show a small degree of curvature, hence the accuracy of the extrapolation is very dependent on the personal judgement of the experimenter. Any estimation of the errors involved may be classified similarly.

In this paper it is intended to give in detail the method employed in fitting a multiple and curvilinear regression by the method of least squares to light scattering data.

An Elliott Algol computer programme has been written to give M_w, $\langle S^2\rangle^{\frac{1}{2}}$, A_2 and $A_3Q(\theta)$ values and their standard errors.

The method has been applied to three sets of data: (1) data from the *Sofica Handbook;* (2) data from an experiment in this Department; (3) data from a recent publication by Huglin *et al.*[2].

In the latter publication a similar type of analysis was reported. Huglin *et al.* did not calculate standard errors and it will be shown that on a theoretical basis two more terms should be added to Huglin's equation.

THEORY

A Zimm plot usually has a certain degree of curvature. This would suggest that one should consider adding higher terms of $\sin^2 \frac{1}{2}\theta$ and C and probably cross terms. The corrected form (see Appendix) of the equation originally published by Zimm gives

$$\frac{KC}{I} = \frac{1}{M_w P_\theta} + 2A_2C + [3A_3Q_\theta + 4A_2^2 M P_\theta(1 - P_\theta)]C^2 \tag{1}$$

which may be expanded to

$$\frac{KC}{I} = \frac{1}{M_w} + \frac{D\sin^2 \frac{1}{2}\theta}{3M_w} + \frac{D^2 \sin^4 \frac{1}{2}\theta}{36M_w} + 2A_2C + 3A_3Q_\theta C^2 + \frac{4A_2^2 DM \sin^2(\frac{1}{2}\theta)C^2}{3} \tag{2}$$

where $D = 16\pi^2 \langle S^2 \rangle / \lambda'^2$, λ' is the wavelength in the media, $\langle S^2 \rangle^{\frac{1}{2}}$ is the radius of gyration and $\sin^4(\frac{1}{2}\theta) C^2$ and higher terms have been omitted.

The above equation may be written as

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 \tag{3}$$

where $Y = KC/I$, $b_0 = 1/M_w$, $b_1 = \langle S^2 \rangle / M_w$, $b_2 = \langle S^2 \rangle^2 / M_w$, $b_3 = A_2$, $b_4 = A_3 Q_\theta$, $b_5 = A_2^2 \langle S^2 \rangle M$, $X_1 = 52{\cdot}63[n \sin \frac{1}{2}\theta / \lambda]^2$, $X_2 = 692{\cdot}68[n \sin \frac{1}{2}\theta / \lambda]^4$, $X_3 = 2C$. $X_4 = 3C^2$, and $X_5 = 210{\cdot}54[n \sin \frac{1}{2}\theta / \lambda]^2 C^2$. Here $n \sin \frac{1}{2}\theta / \lambda$ is employed as a variable as suggested by Carpenter[3].

The problem is to derive the values of the constants (b_0, b_1, ... b_5) and their standard errors by means of a multiple and curvilinear regression analysis. The terms omitted by Huglin *et al.* were b_2X_2 and b_5X_5.

The above equation may be transformed into

$$Y - \overline{Y} = b_1(X_1 - \overline{X}_1) + b_2(X_2 - \overline{X}_2) + b_3(X_3 - \overline{X}_3) + b_4(X_4 - \overline{X}_4) + b_5(X_5 - \overline{X}_5) \tag{4}$$

where the bar above a variable denotes the mean value of that variable and

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - b_3\overline{X}_3 - b_4\overline{X}_4 - b_5\overline{X}_5 \tag{5}$$

This may be written as

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 \tag{6}$$

where $x_1 = X_1 - \overline{X}_1$. The deviation between a predicted value and an observed experimental value is given by

$$y - Y_{(P)} = y - b_1x_1 - b_2x_2 - b_3x_3 - b_4x_4 - b_5x_5 \tag{7}$$

The best fit is that which makes the sum of the squares of these deviations a minimum. The approach is that of the least squares method and it can be shown that:

$$\begin{array}{l} b_1C_{11} + b_2C_{12} + b_3C_{13} + b_4C_{14} + b_5C_{15} = A_{1y} \\ b_1C_{21} + \quad . \quad . \quad . \quad . \quad = A_{2y} \\ \vdots \\ b_5C_{51} + \quad . \quad . \quad . \quad . \quad = A_{5y} \end{array}$$

where $C_{11} = \sum x_1^2$, $C_{12} = \sum x_1x_2$, $A_{1y} = \sum x_1y$. The above may be written in matrix notation as

$$\begin{vmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} \\ C_{21} & . & . & . & . \\ . & & & & . \\ C_{51} & . & . & . & C_{55} \end{vmatrix} \begin{vmatrix} b_1 \\ . \\ . \\ b_5 \end{vmatrix} = \begin{vmatrix} A_{1y} \\ . \\ . \\ A_{5y} \end{vmatrix}$$

$$CB = A$$

for which the solution is $B = C^{-1}A$.

$$\begin{vmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_5 \end{vmatrix} = \begin{vmatrix} C^{11} & C^{12} & C^{13} & C^{14} & C^{15} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ C^{51} & & & & C^{55} \end{vmatrix} \begin{vmatrix} A_{1y} \\ \cdot \\ \cdot \\ \cdot \\ A_{5y} \end{vmatrix}$$

where C^{11} represents the inverse element of C_{11}.

The above has been derived with reference to the mean values. A similar set of equations may be obtained using the actual values.

$$\begin{vmatrix} \sum X_1X_1 & \sum X_1X_2 & \sum X_1X_3 & \sum X_1X_4 & \sum X_1X_5 & \sum X_1 \\ \sum X_2X_1 & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \sum X_5X_1 & & & & \sum X_5X_5 & \sum X_5 \\ \sum X_1 & \cdot & \cdot & \cdot & \sum X_5 & F \end{vmatrix} \begin{vmatrix} b_1 \\ \\ \\ \\ b_5 \\ b_0 \end{vmatrix} = \begin{vmatrix} \sum X_1Y \\ \cdot \\ \cdot \\ \cdot \\ \sum X_5Y \\ \sum YY \end{vmatrix}$$

where F is the total number of observations. This may be written $Db = R$. Hence $b = D^{-1}R$.

The sum of squares due to regression is

$$b_1A_{1Y} + b_2A_{2Y} + b_3A_{3Y} + b_4A_{4Y} + b_5A_{5Y}$$

The residual sum of squares due to regression is

$$A_{YY} - [b_1A_{1Y} + b_2A_{2Y} + b_3A_{3Y} + b_4A_{4Y} + b_5A_{5Y}] \tag{9}$$

hence the residual variance is

$$S^2 = [A_{YY} - b_1A_{1Y} - b_2A_{2Y} - b_3A_{3Y} - b_4A_{4Y} - b_5A_{5Y}]/[F - (5+1)] \tag{10}$$

where $F - (5+1)$ represents the number of degrees of freedom.

The solutions to the above equations are:

$$\begin{aligned} b_1 &= A_{1Y}C^{11} + A_{2Y}C^{12} + A_{3Y}C^{13} + A_{4Y}C^{14} + A_{5Y}C^{15} \\ &\cdot \quad \cdot \qquad\qquad\qquad\qquad \cdot \\ &\cdot \quad \cdot \qquad\qquad\qquad\qquad \cdot \\ &\cdot \quad \cdot \qquad\qquad\qquad\qquad \cdot \\ b_5 &= A_{1Y}C^{51} + A_{2Y}C^{52} \quad \cdot \quad \cdot \quad A_{5Y}C^{55} \end{aligned}$$

$C^{12} = C^{21}$, $C^{34} = C^{43}$ etc.

The standard errors are given by:

$$\begin{aligned} \text{S.E.}(b_1) &= S\sqrt{C^{11}} \\ \text{S.E.}(b_2) &= S\sqrt{C^{22}} \\ \cdot \quad & \quad \cdot \\ \cdot \quad & \quad \cdot \\ \text{S.E.}(b_5) &= S\sqrt{C^{55}} \end{aligned} \tag{11}$$

The standard error in b_0 is given by either

$$\text{(i) S.E.}(b_0)=S\sqrt{D^{66}}$$

or

$$\text{(ii) S.E.}(b_0)=S[(1/F)+C^{11}\overline{X}_1^2+C^{22}\overline{X}_2^2\ldots\ldots+2C^{12}\overline{X}_1\overline{X}_2+2C^{13}\overline{X}_1\overline{X}_3\ldots\ldots]^{\frac{1}{2}}$$

$$=S[(1/F)+\sum_{\substack{j=1\\i=1}}^{\substack{j=5\\i=5}} C^{ji}\overline{X}_i\overline{X}_j]^{\frac{1}{2}}$$

The confidence limits for the coefficients are given by

$$b_T\pm t_\alpha S\sqrt{C^{TT}}=b_T\pm t_\alpha S.E.(b_T)$$

t_α is obtained from tables for $F-6$ degrees of freedom.

From equation (3) $\langle S^2\rangle=b_1/b_0$ and $\langle S^2\rangle^2=b_2/b_0$. The variance of a ratio is given by Fieller's[4] theorem.

Let $b_T/b_0=R$. If $t_\alpha^2S^2\,\text{Var}\,(b_0)/b_0^2$ is small then Fieller's theorem gives

$$\text{Var}\,(R)=S^2[\text{Var}\,(b_T)-2R\,\text{Covar}\,(b_Tb_0)+R^2\,\text{Var}\,(b_0)]/b_0^2$$

To solve the above, the Covar (b_Tb_0) is required. This is given by the respective terms of the inverse matrix (D^{-1}).

Covar (b_1b_0) is given by D^{16} and Covar (b_2b_0) by D^{26}.

S.E.$(R)=\sqrt{\text{Var.}(R)}$

COMPUTATION

The Algol programme solves the equations in the manner described above. If the absolute values of the variables were used in the matrices then they would contain elements with a large numerical range (10^{-10} to 10^{42}). To avoid any errors due to this large range each variable was multipled by a scaling factor (some power of ten) such that the mean value was in the range 10^{-1} to 10^{+1}. *Table 1* gives the results for the three sets of data in scaled units.

Table 1. Results

	HUGLIN	SOFICA	14L
b_0	3·875/0*	1·572/0	4·123/0
S.E.b_0	6·504/−2	1·935/−1	8·700/−2
b_1	3·328/−1	4·138/−1	1·805/−1
S.E.b_1	4·033/−2	1·103/−1	3·765/−2
b_2	1·644/−2	1·228/−1	4·043/−2
S.E.b_2	2·466/−2	1·069/−1	3·585/−2
b_3	3·858/−1	1·914/−1	4·576/−1
S.E.b_3	7·599/−3	1·568/−1	6·530/−2
b_4	1·410/−2	5·433/−2	1·627/−2
S.E.b_4	4·794/−4	4·232/−2	1·634/−2
b_5	−2·289/−3	1·364/−2	6·834/−3
S.E.b_5	4·653/−4	4·059/−2	1·499/−2
S.S.E.R.†	99·98%	90·03%	99·14%

*/-2 represents $\times 10^{-2}$.

†S.S.E.R. Sum of squares explained by regression.

Table 2

	HUGLIN	SOFICA	14L[5]
t_a	2·05	2·00	2·04
M_w (reported)	2·54/5†	7·55/5	2·55/5
M_w (regression)	2·58/5	6·79/5	2·43/5
ΔM_w‡	±0·09/5	±1·71/5	±0·10/5
ΔM_w as %	±3·5	25·0	±4·1
$\langle S^2\rangle^{\frac{1}{2}}$ (reported)	288/−8	490/−8	181/−8
$\langle S^2\rangle^{\frac{1}{2}}$ (regression)	290/−8	477/−8	203/−8
$\Delta\langle S^2\rangle^{\frac{1}{2}}$	±41/−8	±287/−8	±48/−8
$\Delta\langle S^2\rangle^{\frac{1}{2}}$ as %	±14·0	±60·0	±24·0
A_2 (reported)	3·86/−4	3·3/−4	4·59/−4
A_2 (regression)	3·86/−4	1·914/−4 N.S.	4·56/−4
ΔA_2	±0·16/−4	±3·199/−4	±1·35/−4
ΔA_2 as %	±4·2	±167·0	±30·0
$A_3Q(\theta)$ (reported)	1·31/−2		
$A_3Q(\theta)$ (regression)	1·410/−2	N.S.*	N.S.
$\Delta A_3Q(\theta)$	±0·097/−2		
$\Delta A_4Q(\theta)$ as %	±6·9		

*N.S. not significant.
†/5 represents ×10⁵.
‡The Δ values quoted are the 95% confidence limit values.

DISCUSSION OF RESULTS

The signifiance of a coefficient is measured by its ratio to its standard error. For b_1, $t=b_1/S\sqrt{C^{11}}$. The value of t obtained is compared with that in the tables for $F-6$ degrees of freedom and the chosen significance level (5·0 per cent). If the value is less than that in the table then the coefficient is not statistically significant at that level. For the data considered, $t \triangleq 2{\cdot}0$, hence any coefficient that is less than twice its standard error is not statistically significant.

A coefficient may be non-significant under two circumstances. (1) The independent variable has no effect on the dependent variable. (2) The errors in the experimental are such that the test is inconclusive.

For a functional relationship, such as we have here, then case 1 is not admissible and any non-significance must be attributable to experimental error. This does not mean that the variable should be omitted from the regression but rather that the experimental accuracy should be improved.

Coefficient b_2 is not significant for all three samples but has the correct order of magnitude. The results for b_5 (Huglin) are anomalous. Why this should be is not apparent. The experimental errors in I vary over the range of measurement and hence may be the reason for this anomaly.

The data for 14L and Huglin's work suggest that the molecular weight can be determined to within five per cent (95 per cent confidence level). The ranges for the other parameters ($\langle S^2\rangle^{\frac{1}{2}}$, A_2) differ widely and only the Huglin data give small ranges.

The manual results (14L, Sofica) compare favourably with the regression values. The results for A_2 (Sofica) should be noted.

APPENDIX

In the papers on light scattering by B. H. Zimm[1,6] there is a series of misprints. In particular, equations 25 and 26 of ref. 1 and equations 1 and 4 of ref. 6 contain misprints.

Equation 25[1] and equation 1[6] should read

$$I(\theta,C)=K\{MP(\theta)C+2A_2M^2P^2(\theta)C^2+[4A_2^2M^3P^4(\theta)-3A_3M^2P^2(\theta)Q(\theta)]C^3\}$$

Equation 26[1] and equation 4[6] should read

$$\frac{KC}{I(\theta,C)}=\frac{1}{MP(\theta)}+2A_2C+[3A_3Q(\theta)+4A_2^2MP(\theta)\{1-P(\theta)\}]C^2$$

These misprints have been discussed in private correspondence with Professor Zimm.

Department of Chemistry,
University of Aberdeen,
Old Aberdeen, AB9 2UE, Scotland

(*Received February 1969*)

REFERENCES

[1] ZIMM, B. H. *J. chem. Phys.* 1948, **16**, 1093
[2] EVANS, J. M., HUGLIN, M. B. and LINDLEY, J. *J. appl. Polym. Sci.* 1967, **11**, 2159
[3] CARPENTER, D. K. *J. Polym. Sci. A2*, 1966, **4**, 923
[4] FINNEY, D. J. *Statistical Method in Biological Assay.* Charles Griffin: London, 1952
[5] McGIBBON, G. 'Synthesis and solution properties of branched polystyrenes'. *Ph.D. Thesis,* Aberdeen
[6] ZIMM, B. H. *J. phys. Colloid Chem.* 1948, **52**, 260