ON A SIMPLE METHOD OF CURVE FITTING

By K. R. NAIR and M. P. SHRIVASTAVA

Statistical Laboratory, Calcutta

1. The Problem

1. Suppose $y_1, y_2, \ldots y_n$ are the ordinates of an empirical curve corresponding to the values $x_1, x_2, \ldots x_n$ of the independent variable, x, where $x_1 \leqslant x_2 \leqslant x_3 \ldots \leqslant x_n$. Let it be required to find a mathematical curve

$$\dot{Y} = f(x, \alpha, \beta, \gamma, \ldots) \qquad \ldots (1{\cdot}10)$$

which shall represent the empirical curve as closely as possible.

After specifying the form of the mathematical equation (1·10) several methods for the estimation of the parameters $\alpha, \beta, \gamma, \ldots.$ are available, as follows[1]:

(1) The method of Selected Points, (2) The Graphic method, (3) The method of Averages, (4) The method of Least Squares, (5) The method of Moments.

2. The Method of Least Squares is the standard method for estimating the unknown parameters. We start by forming the residual equations:—

$$\left.\begin{aligned} y_1 - f(x_1, \alpha, \beta, \gamma, \ldots.) &= v_1 \\ y_2 - f(x_2, \alpha, \beta, \gamma, \ldots.) &= v_2 \\ &\vdots \\ y_i - f(x_i, \alpha, \beta, \gamma, \ldots.) &= v_i \\ &\vdots \\ y_n - f(x_n, \alpha, \beta, \gamma, \ldots.) &= v_n \end{aligned}\right\} \qquad \ldots (1{\cdot}20)$$

If we assume the v's to be distributed about zero according to the Normal Law, with same standard deviation, the Method of Maximal Likelihood of estimating $\alpha, \beta, \gamma, \ldots$ reduces to the minimising of Σv^2. If, however, no such assumption about the distribution of v's is made, the process of minimising Σv^2 is only a matter of convenience, in so far as we are making the curve pass as closely as possible (with respect to the ordinates) to the observed set of points.

We shall confine ourselves only to the parabolic form of (1.10), which is linear, with respect to the parameters $\alpha, \beta, \gamma, \ldots.$ Their least square estimates from the sample will be denoted by $a, b, c, \ldots$

3. Looking on the method (3), namely, the *Method of Averages*, we at once find that it has ease of computation to commend it. In this method also, we start with a specified form of the equation as (1·10), and get the same residual equations as (1·20). To determine the

parameters we should divide the residual equations in as many groups as there are parameters. By equating to zero the sum of the residuals in the first group we get a single equation in the unknown constants. Equating to zero the sum of the residuals in the second group we get a second equation in the constants, and so on. By solving simultaneously the equations obtained from the several groups, we obtain the estimates of the unknown parameters in the assumed equation.

Thus if there are three parameters α, β and γ, then denoting their estimates by a', b' and c' we can, in the general case, divide the n residual equations in three groups of p, q and r equations yielding the following simultaneous equations in a', b' and c'.

$$\left.\begin{aligned} \sum_{i=1}^{p} f(x_i, \; a', b', c') &= \sum_{i=1}^{p} y_i \\ \sum_{i=1}^{q} f(x_{p+i}, \; a', b', c') &= \sum_{i=1}^{q} y_{p+i} \\ \sum_{i=1}^{r} f(x_{p+q+i}, a', b', c') &= \sum_{i=1}^{r} y_{p+q+i} \end{aligned}\right\} \quad .. \quad (1{\cdot}30)$$

where

$$p+q+r=n \quad .. \quad (1{\cdot}31)$$

Since none of the residual equations is excluded and as the algebraic sum of the residuals in each group is zero, the algebraic sum of all the residuals will be zero. That is to say, the mean of all the observed values will be equal to the mean of all the estimated values.

The Method of Group Averages

4. We shall now make in method (3) a slight modification which consists in allowing some of the residual equations to be omitted. Then $p+q+r<n$. The algebraic sum of the residuals in each group will be zero. But the algebraic sum of all the residuals will not be zero. We shall call this method, the *Method of Group Averages*. The averages of the observed and the expected y's in each group will be equal, but those for the whole set will not. This new method that we are putting forward will be seen to run a middle course between Methods (1) and (3) and even include them as two extreme cases. Thus if p, q and r are made as small as unity, we get method (1); if p, q and r are made big enough to satisfy the condition (1·31) we have method (3).

Mr. S. S. Bose's Work on Linear Regression Coefficient [2]

5. Mr. S. S. Bose had considered elsewhere, three methods alternative to the Method of Least Squares, for evaluating the regression coefficient for observational data on two variables x and y, where the independent variable x varied by equal steps and where the number of observations of y at each value of x was same. All his three methods can be seen to be special cases of either the Method of Averages or of the Method of Group Averages.

Thus Mr. Bose's *Method of Successive Differences* consists in forming the first group of residuals with v_1, v_3, v_5, ... and the second group of residuals with v_2, v_4, v_6, ... The *Method of Differences at half-Range* takes the two groups as the first n 2 of the v's and the last $n/2$ of the v's, if n is even, and as the first $(n-1)/2$ of the v's and the last $(n-1)$ 2 of the v's, if n is odd. The third method, of using the Range, is the same as having in the first group only v_1 and in the second group only v_n. As Mr. Bose had considered in his paper only the problem of estimating β, namely, the linear regression coefficient of y on x and not of getting the complete regression line (which needs estimating both α and β), the connection between his Methods and those of the Method of Group Averages is not at once apparent.

Object of the Paper

8. The object of the present paper is to find the best values for p, q, r, etc. $(p+q+r+\ldots \leqslant n)$ each of which is the number of residuals (all contiguous when the data are arranged according to the magnitude of x) included in the 1st, 2nd, 3rd, etc. groups, so that the relative efficiencies of the estimates of $\alpha, \beta, \gamma, \ldots$ are the maximum possible, in comparison to those obtained by the Method of Least Squares. We shall work out separately the cases of fitting of a straight line and of a parabola of the second degree. The general case of a parabola of a higher degree will follow similarly.

It will be presently seen that there is no unique set of values for p, q and r which will *severally* maximise the relative efficiency of the estimates of α, β, and γ. At the same time we are not aware of any method having been as yet put forward to measure the joint efficiency of simultaneous estimates of a number of parameters. If such an expression was available one could have found a unique set of values for p, q and r which will maximise the efficiency of the Method of Group Averages. What we have done now is only to determine p, q, r, etc ... so as to maximise the efficiency of the coefficient of the *highest* degree.

The problem of getting an estimate of the residual variation of y after fitting by the method of group averages has been discussed in the special case of a straight line fit.

2. The Fitting of a Straight Line

1. Let the n values of x be equally spaced and let them, for the sake of simplicity, be assumed to be $1, 2, \ldots n$. We assume $y_1, y_2, \ldots y_n$ to be equally precise. We have to fit a straight line of the form

$$Y=\alpha+\beta x \qquad \ldots \quad (2 \cdot 10)$$

We shall get estimates of α and β by (1) the Method of Least Squares, and (2) the Method of Group Averages, and shall work out the sampling variances in either case and the efficiencies of estimates obtained by the latter method in comparison to the former.

(2·1) Method of Least Squares

1. Let a and b be the estimates of α and β by the method of Least Squares. Then a and b are such that

$$\sum_{i=1}^{n} (y_i-a-bi)^2 \qquad \ldots \quad (2 \cdot 1{,}10)$$

is a minimum. Hence we have

$$b=\frac{12}{n(n^2-1)} \Sigma\left(i-\frac{n+1}{2}\right) y_i \qquad \ldots \quad (2 \cdot 1{,}11)$$

and

$$a=\bar{y}-b\left(\frac{n+1}{2}\right) \qquad \ldots \quad (2 \cdot 1{,}12)$$

2. If σ^2 be the variance of a single observation on y the variance of b is given by

$$V(b) = \frac{12\sigma^2}{n(n^2-1)} \qquad .. \quad (2\cdot1{,}20)$$

and the variance of a is given by

$$V(a) = \frac{2(2n+1)\sigma^2}{n(n-1)} \qquad .. \quad (2\cdot1{,}21)$$

(2.2) THE METHOD OF GROUP AVERAGES

1. Let a' and b' be the estimates of α and β. The residual equations are

$$\left.\begin{aligned} y_1-\alpha-1\beta&=r_1 \\ y_2-\alpha-2\beta&=r_2 \\ y_i-\alpha-i\,\beta&=r_i \\ y_n-\alpha-n\beta&=r_n \end{aligned}\right\} \qquad .. \quad (2\cdot2{,}10)$$

We should form two groups, the members of each group being contiguous. In general we may take the first p residual equations, in the first group, and the last q equations, in the second group. Tho $n-p-q$ equations in the middle will be omitted. The two simultaneous equations for solving a' and b' will bo

$$\left.\begin{aligned} pa'+(1+2+\ldots+p)b' &= \sum_{i=1}^{p} y_i \\ qa'+\{(n-q+1)+(n-q+2)+\ldots+n\}\,b' &= \sum_{i=n-q+1}^{n} y_i \end{aligned}\right\} \qquad (2\cdot2{,}11)$$

Eliminating a' we get

$$b' = \frac{2\left[\dfrac{1}{q}\sum\limits_{n-q+1}^{n} y_i - \dfrac{1}{p}\sum\limits_{1}^{p} y_i\right]}{2n-p-q} \qquad .. \quad (2\cdot2{,}12)$$

And

$$a' = \frac{\dfrac{2n-q+1}{p}\sum\limits_{1}^{p} y_i - \dfrac{p+1}{q}\sum\limits_{n-q+1}^{n} y_i}{2n-p-q} \qquad .. \quad (2\cdot2{,}13)$$

2. The sampling variances of b' and a' are given by

$$V(b') = \frac{4(p+q)\sigma^2}{pq(2n-p-q)^2} \qquad .. \quad (2\cdot2{,}20)$$

and

$$V(a') = \sigma^2\left[\frac{1}{p}\left(1+\frac{p+1}{2n-p-q}\right)^2 + \frac{1}{q}\left(\frac{p+1}{2n-p-q}\right)^2\right] \qquad .. \quad (2\cdot2{,}21)$$

3. The efficiency of b' compared to that of the most efficient estimate, namely b, on the assumption of normal theory, is

$$E(b') = \frac{V(b)}{V(b')} = \frac{3pq(2n-p-q)^2}{n(n^2-1)(p+q)} \qquad .. \quad (2\cdot2{,}30)$$

and the efficiency of a' as compared to that of the efficient estimate a is given by

$$E(a')=\frac{V(a)}{V(a')}=\frac{2(2n+1)/n(n-1)}{\frac{1}{p}\left(1+\frac{p+1}{2n-p-q}\right)^2+\frac{1}{q}\left(\frac{p+1}{2n-p-q}\right)^2} \quad .. \quad (2{\cdot}2,31)$$

4. The values of p and q are at our disposal. Let us choose them in such a way that the efficiency of b' is maximum possible. Differentiating the logarithmic function of (2·2,30) partially with respect to p and q we get,

$$\left.\begin{aligned} \frac{1}{p}-\frac{2}{2n-p-q}-\frac{1}{p+q}&=0 \\ \frac{1}{q}-\frac{2}{2n-p-q}-\frac{1}{p+q}&=0 \end{aligned}\right\} \quad .. \quad (2{\cdot}2,40)$$

Both these equations are satisfied by the values

$$p=q=\frac{n}{3} \quad .. \quad (2{\cdot}2,41)$$

Thus the most efficient way for estimating β for a linear fitting by the method of Group Averages is to divide the whole set of residual equations into three equal parts and then reject the middle one. This does not, however, simultaneously maximise the efficiency of a'.

5. Substituting $p=q=n/3$ in (2.2,30) and (2.2,31) we have

$$E(b')=\frac{8}{9}\cdot\frac{n^2}{n^2-1}>\frac{8}{9} \quad .. \quad (2{\cdot}2,50)$$

(Putting $n=3$ in (2.2,50), $E(b')=1$, as it ought to be, for when $n=3$ the least square method and the method of maximal group-averages, namely, with $p=q=1$, give identical estimates of β.) and

$$E(a')=\frac{16}{3(n-1)}\frac{n^2(2n+1)}{(13n^2+18n+9)}>\frac{32}{39} \quad .. \quad (2{\cdot}2,51)$$

6. The results (2.2,50) and (2.2,51) hold when n is a multiple of 3. When n is not a multiple of 3, two cases are possible, namely, $n=3m\pm1$, in either of which cases the best value of p is m.

(2·3) Discussion

7. Mr. S. S. Bose had found that the 'method of differences at half range' was the best among the three methods he was considering. This evidently is a special case of our method, and is obtained by putting $p=q=n/2$, if n is even; and $p=q=\frac{1}{2}(n-1)$, if n is odd, the middle term being omitted. It is, however, obvious from the above that $p=q=n/3$ gives for the linear coefficient, an efficiency of about 90 per cent, whereas $p=q=n/2$ gives only an efficiency of about 75 per cent.

Mr. Bose's 'method of range' is also a special case of ours, namely, when $p=q=1$. By substituting these values in (2·2,30) and (2·2,31) we get the efficiencies of b' and a' as $\frac{6(n-1)}{n(n+1)}$ and $\frac{2(n-1)(2n+1)}{n(n^2+1)}$ which decrease rapidly as n increases. At $n=2$ both efficiencies are 100 as they ought to be; and at $n=3$, that of the first only is 100. For higher values of n the efficiencies rapidly deteriorate. His third method of 'successive differences' amounts to forming groups with non-contiguous elements and has brought about a great fall in efficiency even though the whole data are utilized. The reason for this fall will be at once clear by a graphical interpretation of the situation.

8. Suppose we have plotted all the (x, y) points. By the method of Group Averages our endeavour is to take two groups of points and to draw the mean ordinates for each group. The line joining the tops of the two ordinates will be our fitted line. Evidently the stability of this line depends on two opposing tendencies. The larger the number of points included in each group, the greater the accuracy of the corresponding average ordinate, and therefore, the closer will the fit of the estimated line be to the expected one. On the other hand, the farther apart these average ordinates are, the less will the line (or its gradient) be susceptible to sampling errors of estimation of their lengths. The first condition, if fulfilled, brings the average ordinates closer together, whereas the second diminishes the accuracy of their lengths, being based on less and less number of observations at the two extremes; hence only through a well balanced compromise between these two tendencies can we get the best possible line, that is to say, a line whose shift about the true line will be as small as possible.

9. If the two groups are formed by the odd and even ordinates, the mean ordinates of the two groups will lie very close together. The inclination of the line joining their tops will therefore be very sensitive to fluctuations in the lengths of the mean ordinates, and its samping variance will be high, thus lowering the efficiency. On the other hand, the method of range consists of joining the tops of the two extreme ordinates. Here the ordinates are farthest apart, but the fluctuations due to random sampling in single ordinates are likely to be much greater than those of the mean ordinates of a group. Thus the 'method of successive differences' suffers from lack of fulfilment of the second criterion of stability while the 'method of range' ignores the first criterion. The 'method of differences at half range' however, effects a compromise. Our method, namely, that of forming the two groups with the first one-third and the last one-third sets of ordinates, brings about, however, the best possible compromise.

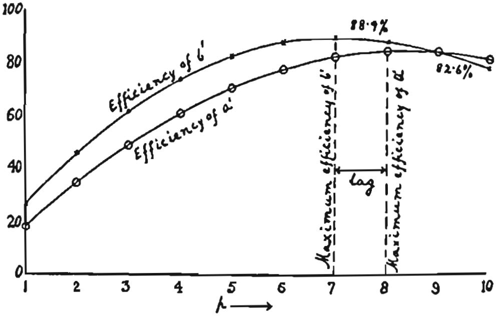

Chart (1) Efficiency of the Method of Group Averages ($n=21$).

10. Chart 1 shows, for sake of illustration, the efficiencies of a' and b' for various values of $p=q=1, 2, \ldots\ldots \frac{n-1}{2}$ for a particular value of n, namely, 21. It will be noticed that while the highest efficiency of b' is reached at $p=q=7$, the highest efficiency for a' is reached for a higher value of $p=q$, the nearest integer to this value being 8.

Table 1. Efficiency of the Estimates a' and b' for Different Sizes of Sample

m	N=3m			N=3m+1			N=3m−1		
	N	Percentage Efficiency of		N	Percentage Efficiency of		N	Percentage Efficiency of	
		a	b'		a'	b'		a'	b'
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	3	93·3	100·0	4	79·4	90·0	—	—	—
2	6	85·3	91·5	7	80·3	89·3	5	88·0	90·0
3	9	83·8	90·0	10	80·7	89·1	8	85·9	89·3
4	12	83·2	89·5	13	81·0	89·0	11	84·9	89·1
5	15	82·9	89·3	16	81·2	89·0	14	84·4	89·0
6	18	82·8	89·2	19	81·3	89·0	17	84·0	89·0
7	21	82·6	89·1	22	81·4	88·9	20	83·8	89·0
8	24	82·5	89·0	25	81·6	88·9	23	83·5	88·9
9	27	82·5	89·0	28	81·5	88·9	26	83·3	88·9
10	30	82·4	89·0	—	—	—	29	83·2	88·9

11. Table 1 shows the efficiencies of a' and b' for different values of n from 3 to 30, in separate columns for the types $3m$, $3m+1$ and $3m-1$, keeping $p=q=m$ which is the best value for estimating β. Charts (2.1), (2.2) and (2.3) represent those efficiencies graphically for the cases $n=3m$, $3m+1$ and $3m-1$ respectively, with $p=m$ in each case. It will be seen that when $n=3m$ or $3m-1$, the efficiencies of both a' and b' remain greater than their respective limiting values when $n \to \infty$. In the case of $n=3m+1$ the efficiency of b' for finite values of n is greater than its limiting value for $n \to \infty$, but that of a' is always less than (the lowest value being 79.4 per cent) the limiting value at $n \to \infty$.

3. Fitting of a Parabola of the Second Degree

1. As before let the n values of x, for the sake of simplicity, be assumed to be 1, 2,... n, and let $y_1, y_2, \ldots y_n$ be each equally precise. We have to fit a parabola of the form

$$Y=\alpha+\beta x+\gamma x^2 \tag{3·10}$$

We shall get estimates of α, β and γ, first by the method of Least Squares, and second, by the method of Group Averages. We shall examine for which size of grouping maximum efficiency could be attained.

(3·1). Method of Least Squares

1. Let a, b, c be the estimates of α, β and γ respectively by this method. Then

$$\sum_{i=1}^{n} (y_i-a-bi-ci^2)^2 \tag{3·1,10}$$

should be a minimum.

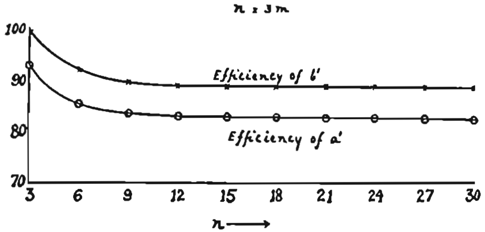

CHART (2·1)

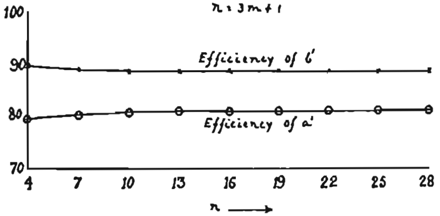

CHART (2·2)

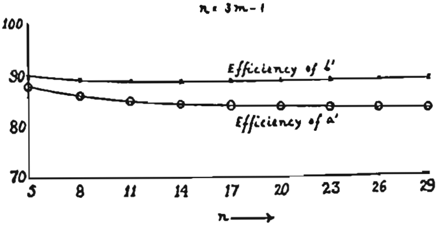

CHART (2·3)

The following normal equations will determine the values of a, b and c

$$\left.\begin{array}{l} na+b\,\Sigma i+c\,\Sigma i^2=\Sigma\;y_i \\ a\Sigma i+b\,\Sigma i^2+c\,\Sigma i^3=\Sigma i\,y_i \\ a\Sigma i^2+b\,\Sigma i^3+c\,\Sigma i^4=\Sigma i^2\,y_i \end{array}\right\} \quad .. \quad (3{\cdot}1{,}11)$$

where each Σ stands for summation over all values observed, and

$$\left.\begin{array}{l} \Sigma i=\tfrac{1}{2}n(n+1),\;\Sigma i^2=\tfrac{1}{6}\;n(n+1)\;(2n+1) \\ \Sigma i^3=\{\tfrac{1}{2}n(n+1)\}^2,\;\Sigma i^4=\tfrac{1}{30}\,n(n+1)\,(6n^3+9n^2+n-1) \end{array}\right\} \quad .. \quad (3{\cdot}1{,}12)$$

The variances of a, b and c can be easily worked out and are :—

$$V(a)=\frac{3(3n^2+3n+2)\,\sigma^2}{n(n-1)\,(n-2)} \quad .. \quad (3{\cdot}1{,}13)$$

$$V(b)=\frac{12(16n^2+30n+11)\,\sigma^2}{n(n^2-1)\,(n^2-4)} \quad .. \quad (3{\cdot}1{,}14)$$

$$V(c)=\frac{180\sigma^2}{n(n^2-1)\;(n^2-4)} \quad .. \quad (3{\cdot}1{,}15)$$

(3·2). Method of Group Averages

1. Let a', b', c' be the estimates of α, β, γ respectively. The residual equations are

$$\left.\begin{array}{l} y_1-\alpha-\beta-\gamma=v_1 \\ y_2-\alpha-2\beta-2^2\gamma=v_2 \\ \vdots \\ y_i-\alpha-i\,\beta-i^2\,\gamma=v_i \\ \vdots \\ y_n-\alpha-n\beta-n^2\gamma=v_n \end{array}\right\} \quad .. \quad (3{\cdot}2{,}10)$$

Here we require three simultaneous equations to solve for α, β and γ. From our experience of fitting the straight line it is easy to see that these three equations should all be based on an equal number, say p, of contiguous residual equations (3·2,10) so that $p=q=r$. Let us here omit q' residual equations between the first and the second group and r' residual equations between the second and the third groups so that we shall have

$$3p+q'+r'=n \quad (3{\cdot}2{,}11)$$

Let the three equations so formed be

$$\left.\begin{array}{l} A_1\,a'+B_1\,b'+C_1\,c'=S_1 \\ A_2\,a'+B_2\,b'+C_2\,c'=S_2 \\ A_3\,a'+B_3\,b'+C_3\,c'=S_3 \end{array}\right\} \quad .. \quad (3{\cdot}2{,}12)$$

where

$$A_1=A_2=A_3=p$$

$$B_1=\sum_{1}^{p}\,i=\tfrac{1}{2}\,p(p+1)$$

$$B_2=\sum_{p+q'+1}^{2p+q'}\,i=\tfrac{1}{2}\,p(3p+2q'+1)$$

$$B_3=\sum_{n-p+1}^{n}\,i=\tfrac{1}{2}\,p(2n-p+1)$$

$$C_1 = \sum_{1}^{p} i^2 = \tfrac{1}{6}\,p(p+1)\,(2p+1)$$

$$C_2 = \sum_{p+q'+1}^{2p+q'} i^2 = \tfrac{1}{6}\,p[(3p+2q'+1)\,(4p+2q'+1)+2(p+q')\,(p+q'+1)]$$

$$C_3 = \sum_{n-p+1}^{n} i^2 = \tfrac{1}{6}\,p[(2n-p+1)\,(2n+1)+2(n-p)\,(n-p+1)]$$

$$S_1 = \sum_{1}^{p} y_i,\; S_2 = \sum_{p+q'+1}^{2p+q'} y_i,\; S_3 = \sum_{n-p+1}^{n} y_i$$

After solving these equations the variances of a', b' and c' can be easily obtained. We shall start with the variance of c' which is :—

$$V(c') = \frac{[(n-2p-q')^2+(n-p)^2+(p+q')^2]}{p(n-2p-q')^2\;(n-p)^2\;(p+q')^2}\,\sigma^2 \qquad .. \quad (3{\cdot}2{,}13)$$

2. Hence the efficiency of c' as compared to that of the efficient estimate c is

$$E(c') = \frac{V(c)}{V(c')} = \frac{180}{n(n^2-1)(n^2-4)}\;.\;\frac{p(n-2p-q')^2\,(n-p)^2\,(p+q')^2}{(n-2p-q')^2+(n-p)^2+(p+q')^2} \qquad .. \quad (3{\cdot}2{,}20)$$

Now, the values of p and q' are at our disposal. Let us so express q' in terms of p that $E(c')$ shall be maximum possible. Maximising the numerator, and simultaneously minimising (3.2,20) with respect to q', we have

$$\left.\begin{aligned} \frac{1}{p+q'}-\frac{1}{n-2p-q'} &= 0 \\ \text{and}\qquad p+q'\;-(n-2p-q') &= 0 \end{aligned}\right\} \qquad .. \quad (3{\cdot}2{,}21)$$

both being identical; and giving

$$q' = \tfrac{1}{2}\,(n-3p) \qquad .. \quad (3{\cdot}2{,}22)$$

which is nothing but the condition that

$$q' = r' = \tfrac{1}{2}\,(n-3p) \qquad .. \quad (3{\cdot}2{,}23)$$

i.e. the two omitted groups should also consist of an equal number of residual equations. With this value of q' (3.2,20) reduces to

$$E(c') = \frac{15}{2}\;.\;\frac{p(n-p)^4}{n(n^2-1)\,(n^2-4)} \qquad .. \quad (3{\cdot}2{,}24)$$

which attains the maximum value when $p=\dfrac{n}{5}$.

Thus for the maximum possible efficiency of c' we should choose our groups in such a way that

$$p = q = r = q' = r' = \frac{n}{5} \qquad .. \quad (3{\cdot}2{,}25)$$

that is to say, *we should divide the whole set of residual equations into five equal parts and then reject the second and the fourth ones.*

3. With such a choice of groups, to form the simultaneous equations (3·2,12), we find that the efficiencies of the estimates a', b' and c' in relation to the efficient estimates a, b and c of the population parameters α, β and γ are given respectively by :

$$E(a') = \frac{6012\;n^4(3n^2+3n+2)}{5(n-1)\;(n-2)\;\phi_1(n)}\,, \qquad .. \quad (3{\cdot}2{,}30)$$

where $$\phi_1(n)=5214n^4+22500n^3+48150n^2+45000n+15000\,; \qquad .. \quad (3{\cdot}2{,}31)$$

$$E(b')=\frac{384}{125}\cdot\frac{n^4(16n^2+30n+11)}{(n^2-1)\,(n^2-4)\,\phi_2(n)}, \qquad .. \quad (3{\cdot}2{,}32)$$

where $$\phi_2(n)=79n^2+150n+75\,; \qquad .. \quad (3{\cdot}2{,}33)$$

$$\text{and } E(c')=\frac{384}{625}\cdot\frac{n^4}{(n^2-1)\,(n^2-4)}. \qquad .. \quad (3{\cdot}2{,}34)$$

4. The efficiencies of a', b' and c' have been worked out and presented in Table 2 for a few values of n namely, 5, 10, 15, 20 and 25. We notice that the efficiencies are greatest for $n=5$, and then decrease steadily to their limiting values when $n\rightarrow\infty$, except in the case of a' where the minimum efficiency, namely, about 78·9 per cent, is reached in the neighbourhood of $n=20$ and slightly increases for higher values of n until it rises to 79.5 per cent when $n\rightarrow\infty$.

Table 2. Percentage Efficiencies of a', b', c' for Different Sizes of Sample

n	E(a′)	E(b′)	E(c′)
(1)	(2)	(3)	(4)
5	88·14	70·33	76·19
10	79·76	65·19	64·65
15	79·00	63·47	62·83
20	78·90	62·90	62·22
25	78·92	62·61	61·93

It is interesting to note that although we have so selected our group size, namely, $n/5$ as to maximize the efficiency of c' even this maximum value is less than the efficiencies of a' and b' for the same group size.

5. The efficiency factors (3·2,30), (3·2,32) and (3·2,34) are, however, valid only when n is of the form $n=5m$ where m is an integer. If n is of the form $n=5m\pm t$ $(t=1, 2)$ it follows, from an analogy with the fitting of the straight line, that the best value for p is m, but this will give equal values for q' and r' only when $n=5m\pm2$.

6. Again, it may incidentally be noted here that had we used the method (3), namely, the Method of Averages, by forming the three groups each based on $n/3$ observations, the efficiencies of the estimates would have been greatly reduced. Thus the efficiency in this case of the estimate of γ can be obtained by putting $p=\frac{n}{3}$, $q'=0$ in (3·2,20) and is

$$E(c')=\frac{40}{81}\cdot\frac{n^4}{(n^2-1)\,(n^2-4)} \text{ or 49 per cent} \qquad .. \quad (3{\cdot}2{,}60)$$

By an analogy to the case of the straight line, this is at once seen to be an extension, to the fitting of a second degree parabola, of Mr. S. S. Bose's method of differences at half range for straight line fitting.

4. Fitting of a Parabola of the p^{th} Degree

1. From the two cases discussed above—namely, the fitting of a straight line and a parabola of the second degree — it is easily seen that the method of Group Averages is quite general. To fit a parabolic curve of the p^{th} degree, namely,

$$y=a_0+a_1\,x+a_2\,x^2+\;\ldots\;a_p\,x^p \qquad .. \quad (4{\cdot}10)$$

involving $p+1$ unknown parameters, we need $p+1$ simultaneous equations in the constants, a_0, a_1, a_2, ... a_p. We may then divide the n residuals in the order of values of x into

$2p+1$ equal parts, if n is a multiple of $2p+1$, equate severally to zero the sum of the residuals in all the $p+1$ odd groups, rejecting completely the even groups. This yields $p+1$ equations which suffice to determine the $p+1$ unknown parameters.

5. Estimate of Residual Variance

1. When a trend is present it will be necessary to get an unbiased estimate of the variance of the random variations in y. When the trend is determined by the method of least squares this is easily obtained from the corresponding analysis of variance table.

2. The problem before us is, how to estimate the residual variance when the method of group averages is used in determining the trend. Suppose we calculate the sum of squares S'_1 of deviations of observed values from the corresponding trend values obtained by the method of group averages. We shall find out the expectation of S'_1 when the trend is linear and $n=3p$ where p is an integer. This comes out to be

$$\text{Exp. } (S'_1)=\left[3p-1-\frac{3p^2+1}{8p^2}\right]\sigma^2_y, \qquad .. \quad (5{\cdot}20)$$

An unbiased estimate of σ^2_y is therefore obtained by dividing the observed S'_1, by

$$3p-1-\frac{3p^2+1}{8p^2} \qquad .. \quad (5{\cdot}21)$$

which is approximately equal to $n-\frac{11}{8}$. It is self-evident that the divisor should be greater than $n-2$, which is the divisor (known as 'degrees of freedom') for the residual sum of squares when the trend is based on normal theory, that is, when evaluated by method of least squares.

3. If we use the residuals of only the first p and the last p observed values which alone have been utilised in getting the equation for the linear trend, the corresponding residual sum of squares, S'_2 will have its expectation as

$$\text{Exp. } (S'_2)=(p-1)\left[2+\frac{p+1}{12p^2}\right]\sigma^2_y \qquad .. \quad (5{\cdot}30)$$

An unbiased estimate of σ^2_y is therefore obtained by dividing the observed S'_2 by

$$(p-1)\left[2+\frac{p+1}{12p^2}\right] \qquad .. \quad (5{\cdot}31)$$

which is very nearly equal to $2\left(\frac{n}{3}-1\right)+\frac{1}{12}$, that is, $\frac{1}{12}$ more than the degrees of freedom in the case of the normal theory.

4. For higher degree curves fitted by the Method of Group Averages the proper divisors of the residual sum of squares are not easy to calculate. But it appears that a sufficiently accurate estimate can be obtained, when n is not small, if we divide the residual sum of squares by the usual degrees of freedom of least square theory. At any rate such an estimate will be little higher than the correct estimate, so that we err on the safer side!

References

1. Bose, S. S. (1938). Relative Efficiencies of Regression Coefficients Estimated by the Method of Finite Differences; *Sankhyā, Vol.* **3**, pp. 339-346.

2. Reitz, H. L.: *Hand-book of Mathematical Statistics*

Paper received: 6 December, 1939