# RELATIVE EFFICIENCIES OF REGRESSION COEFFICIENTS ESTIMATED BY THE METHOD OF FINITE DIFFERENCES

By S. S. BOSE

Statistical Laboratory, Calcutta.

## (1) Introduction.

The regression coefficient of $y$ on $x$ is defined by the expression :

$$b = \frac{S(x-\bar{x})(y-\bar{y})}{S(x-\bar{x})^2} \qquad \dots \quad (1{\cdot}1)$$

where $x$ and $y$ are the two variates, $\bar{x}$ and $\bar{y}$ their mean values, $b$ is the regression of $y$ on $x$ and the summation is taken for all values in the sample.

If the independent variable is taken at equal intervals (as is usually done in time series and in many physical and biological experiments), we have $S(x-\bar{x})^2 = n(n^2-1)/12$ so that

$$b = \frac{S(x-\bar{x})(y-\bar{y})}{n(n^2-1)/12} \qquad \dots \quad (1{\cdot}2)$$

The standard error of $b$ is easily calculated from the expression (1·1)

$$b = \frac{1}{S(x-\bar{x})^2}\left[(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \dots\dots + (x_n-\bar{x})(y_n-\bar{y})\right]$$

If the variance of $y$ is $\sigma^2$ the variance of the linear expression in the numerator = $\sigma^2 \,.\, S(x-\bar{x})^2$ and the variance of $b$ is therefore given by

$$\sigma_b^2 = \frac{\sigma^2 \,.\, S(x-\bar{x})^2}{\left[S(x-\bar{x})^2\right]^2} = \frac{\sigma^2}{S(x-\bar{x})^2} \qquad \dots \quad (1{\cdot}2)$$

In the present paper, it is proposed to estimate the regression coefficient by the method of differences of the values of $y$, and to calculate the relative efficiencies of these estimates in relation to the least square solution. The method of calculation is slightly different for even and odd pairs and the two cases have been treated separately.

### (2) METHOD OF SUCCESSIVE DIFFERENCES.

(a) *Even number of samples,* $n = 2m.$

Let the values of $y$ corresponding to given equally spaced values of $x$ be as follows

$$\begin{array}{ll} x_1 & x_2 \;\ldots\ldots\ldots\; x_m \;\ldots\ldots\ldots\; x_{2m} \\ y_1 & y_2 \;\ldots\ldots\ldots\; y_m \;\ldots\;\ldots\ldots\; y_{2m} \end{array}$$

If the relation between $x$ and $y$ be given by $Y = \alpha+\beta x$ where $Y$, $\alpha$ and $\beta$ are population values obtained from an infinite set of values of $x$ and $y$, we have

$$\begin{array}{lllllll} y_1 & = & Y_1 & + \; e_1 & = (\alpha+\beta x_1) & + \; e_1 \\ y_2 & = & Y_2 & + \; e_2 & = (\alpha+\beta x_2) & + \; e_2 \\ \bullet & \bullet & \bullet & \bullet \quad \circ & \bullet \quad \bullet & \bullet \quad \bullet \\ y_{2m} & = & Y_{2m} & + \; e_{2m} & = (\alpha+\beta x_{2m}) & + \; e_{2m} \end{array}$$

where $e_1$, $e_2$, ...... $e_{2m}$ are errors of $y_1$, $y_2$ ...... $y_{2m}$. It is assumed that these errors are distributed normally with mean $= 0$, and variance $= \sigma^2$.

$$\begin{array}{rl} \text{Now, } y_2 - y_1 & = (Y_2 - Y_1) + (e_2 - e_1) = \beta(x_2 - x_1) + (e_2 - e_1) \\ y_4 - y_3 & = (Y_4 - Y_3) + (e_4 - e_3) = \beta(x_4 - x_3) + (e_4 - e_3) \\ y_{2m} - y_{2m-1} & = (Y_{2m} - Y_{2m-1}) + (e_{2m} - e_{2m-1}) = \beta(x_{2m} - x_{2m-1}) + (e_{2m} - e_{2m-1}) \end{array}$$

Adding,

$$(y_2+y_4+\ldots\ldots y_{2m})-(y_1+y_3+\ldots\ldots y_{2m-1}) = \beta.d.m + (e_2+e_4+\ldots\ldots e_{2m})-(e_1+e_3+\ldots e_{2m-1})$$

where $\qquad d = x_2-x_1 = x_4-x_3 = \ldots\ldots = x_{2m}-x_{2m-1}$

Therefore, $\quad \beta \,.\, d \,.\, m = \left[S(y_{2t})-S(y_{2t-1})\right]_1^m - \left[S(e_{2t})-S(e_{2t-1})\right]_1^m \qquad \ldots \quad (2{\cdot}0)$

If $m$ is indefinitely large, we have assumed that the mathematical expectation of $S(e)$ would vanish.

We may therefore define

$$b_1 = \frac{\left[S(y_{2t})-S(y_{2t-1})\right]_1^m}{md} = \frac{2}{n} \,.\, \frac{1}{d} \left[S(y_{2t})-S(y_{2t-1})\right]_1^m \qquad \ldots \quad (2{\cdot}1)$$

where $n=2m$. Then the mathematical expectation of $E(b_1) = \beta$, so that the estimate $b_1$ is unbiassed.

We have

$$\begin{array}{lll} y_2 - y_1 & = \beta d + (e_2 - e_1) & = \beta d + e'_1 \\ y_4 - y_3 & = \beta d + (e_4 - e_3) & = \beta d + e'_2 \\ \bullet & \bullet \qquad \bullet & \bullet \qquad \bullet \\ y_{2m} - y_{2m-1} & = \beta d + (e_{2m} - e_{2m-1}) & = \beta d + e'_m \end{array}$$

Now, $\beta.l+e'$ is distributed normally with mean $=\beta d$, and variance $= 2\sigma^2$ (twice the variance of $e$). Since $db_1 = \frac{1}{m} . \mathrm{S}_1^m (\beta d + e')$, the variance of $db_1$ is $2\sigma^2/m$.

Thus the variance of $$b_1 = \frac{4\sigma^2}{nd^2} \qquad \text{... (2·2)}$$

and the standard error of $$b_1 = \frac{2\sigma}{d\sqrt{n}} \qquad \text{... (2·3)}$$

(*b*) *Odd number of samples,* $n = 2m+1$.

Using successive first differences, the regression coefficient is given by

$$b'_1 = \frac{(y_2-y_1) + (y_4-y_3) + \dots \dots \dots}{md} \qquad \text{... (2·4)}$$

The variance of $$b'_1 = \frac{2\sigma^2}{md^2} = \frac{4\sigma^2}{(n-1)d^2} \qquad \text{... (2·5)}$$

and the standard error of $$b'_1 = \frac{2\sigma}{d\sqrt{(n-1)}} \qquad \text{... (2·6)}$$

### (3) Method of Differences at Half Range.

(*a*) *Even number of samples,* $n = 2m$.

Another estimate of $b$ is obtained by first dividing the $2m$ set of values in two equal parts and then taking differences at intervals of $m$ as in the scheme shown below :

$$\begin{aligned} y_{m+1} - y_1 &= (Y_{m+1}-Y_1) + (e_{m+1}-e_1) = m\beta d + \epsilon_1 \\ \cdot \quad & \quad \cdot \qquad \quad \cdot \qquad \quad \cdot \qquad \quad \cdot \\ y_{2m}-y_m &= (Y_{2m}-Y_m) + (e_{2m}-e_m) = m\beta d + \epsilon_m \end{aligned}$$

Hence, $$\beta = \frac{[\mathrm{S}(y_{m+1}) - \mathrm{S}(y_1)]_1^m - [\mathrm{S}(\epsilon_1)]_1^m}{m^2 d} \qquad \text{... (3·0)}$$

The sample estimate of $\beta$ is given by

$$b_2 = \frac{[\mathrm{S}(y_{m+1}) - \mathrm{S}(y_1)]}{m^2 d} \qquad \text{... (3·1)}$$

The variance of $$b_2 = \frac{2\sigma^2}{m^3 \,.\, d^2} = \frac{16\sigma^2}{n^3 \,.\, d^2} \qquad \text{... (3·2)}$$

The standard error of $$b_2 = \frac{4\sigma}{n^{3/2} \,.\, d} \qquad \text{... (3·3)}$$

(*b*) *Odd number of samples*, $n = 2m+1$.

Omitting the middle term $y_{m+1}$, the estimate of $b'_2$ is obtained as follows:

$$b'_2 = \frac{(y_{m+2}-y_1)+(y_{m+3}-y_2)+\ldots\ldots\ldots+(y_{2m+1}-y_m)}{m(m+1)d} \quad \ldots \quad (3{\cdot}4)$$

The variance of $b'_2 = \dfrac{2\sigma^2}{m(m+1)d} = \dfrac{16\sigma^2}{(n-1)(n+1)^2d^2} \quad \ldots \quad (3{\cdot}5)$

The standard error of $b'_2 = \dfrac{4\sigma}{(n+1)d.\sqrt{(n-1)}} \quad \ldots \quad (3{\cdot}6)$

## (4) Method of Range.

The regression coefficient may also be calculated from the range. Thus if $y_1$ and $y_n$ are the two extreme values of $y$ corresponding to $x_1$ and $x_n$ respectively, the estimate of the regression coefficient is given by

$$b_3 = \frac{y_n-y_1}{(n-1)d} \quad \ldots \quad (4{\cdot}1)$$

The variance of $b_3 = \dfrac{2\sigma^2}{(n-1)^2d^2} \quad \ldots \quad (4{\cdot}2)$

and the standard error of $b_3 = \dfrac{\sigma\sqrt{2}}{(n-1)d} \quad \ldots \quad (4{\cdot}3)$

## (5) Relative Efficiency of the Regression Coefficients.

The variances of the four estimates of regression coefficients may be compared from the point of view of relative precision. Assuming the interval between $x$'s to be unity (which does not lead to any loss of generality) we have the following results.

Table 1. Variance of Regression Coefficients.

| | Even | Odd |
|---|---|---|
| (1) Least Square | $12\sigma^2/n(n^2-1)$ | $12\sigma^2/n(n^2-1)$ |
| (2) Successive differences | $4\sigma^2/n$ | $4\sigma^2/(n-1)$ |
| (3) Difference at half range | $16\sigma^2/n^3$ | $16\sigma^2/(n-1)(n+1)^2$ |
| (4) Method of Range | $2\sigma^2/(n-1)^2$ | $2\sigma^2/(n-1)^2$ |

The standard errors of the regression coefficients for different values of $n$ have been shown in Table 2 and Chart I.

Table 2. Standard Error of Estimates of Regression Coefficient in terms of Population Standard Deviation as Unit.

| $n$ | Least Square | Successive Difference | Difference at half Range | Range |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 2 | 1·414 | 1·414 | 1·414 | 1·414 |
| 3 | ·707 | 1·414 | ·707 | ·707 |
| 4 | ·447 | 1·000 | ·500 | ·471 |
| 5 | ·316 | 1·000 | ·333 | ·354 |
| 10 | ·101 | ·633 | ·126 | ·157 |
| 15 | ·060 | ·534 | ·067 | ·101 |
| 20 | ·039 | ·447 | ·045 | ·074 |
| 25 | ·028 | ·408 | ·031 | ·059 |
| 30 | ·021 | ·365 | ·024 | ·049 |
| 35 | ·017 | ·343 | ·019 | ·042 |
| 40 | ·014 | ·333 | ·016 | ·036 |
| 45 | ·011 | ·302 | ·013 | ·032 |
| 50 | ·010 | ·283 | ·011 | ·029 |

If $\sigma_0^2$ is the variance of the least square estimate and $\sigma^2$ that of another estimate, the relative efficiency has been defined by R. A. Fisher as $I = 100\sigma_0^2/\sigma^2$. The relative efficiencies of the four different estimates are shown in Table 3.

Table 3. Relative Efficiency of Regression Estimates.

| | Even | Odd |
|---|---|---|
| Successive Difference | $3/(n^2-1)$ | $3/n(n+1)$ |
| Difference at half Range | $3n^2/4(n^2-1)$ | $3(n+1)/4n$ |
| Method of Range | $6(n-1)/n(n+1)$ | $6(n-1)/n(n+1)$ |

It will be seen that the efficiency of the method of successive differences falls very rapidly with $n$. With 3 pairs of observations the efficiency is as small as 25 per cent only, and with 18 pairs, it is reduced to less than 1 per cent.

Next to the method of least squares, the method of differences at half range appears to give the best results. The limiting efficiency is 75 per cent, so that at any stage the loss of information is never more than 25 per cent as compared to the least square solution.

The range method gives as efficient estimates as the least square solution up to 3 pairs of observations but, the efficiency begins to fall rapidly as the number of observations increases. With 4 pairs of data, the efficiency is 90 per cent and is, in fact, better than the method of differences at half range. With 5 pairs or more, there is considerable loss of information. The relative efficiencies of regression coefficients for different values of $n$ are shown in Table 4 and Chart 2.

In practice however the efficiency of a statistic is not the only consideration. Davies and Pearson[1] have aptly noted: "An estimate ($E_2$) with a larger standard error than the theoretically most reliable ($E_1$) will sometimes be so much simpler to calculate that it will

CHART 1. STANDARD ERRORS OF THE REGRESSION COEFFICIENTS FOR DIFFERENT VALUES OF *n*.

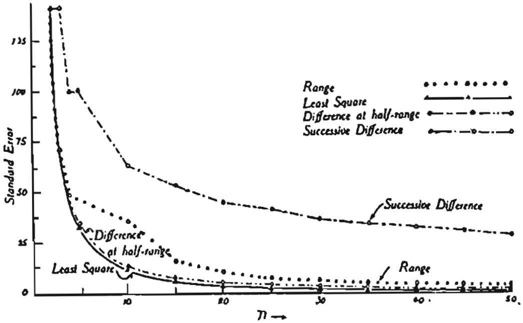

CHART 2. RELATIVE EFFICIENCIES OF REGRESSION COEFFICIENTS FOR DIFFERENT VALUES OF *n*.

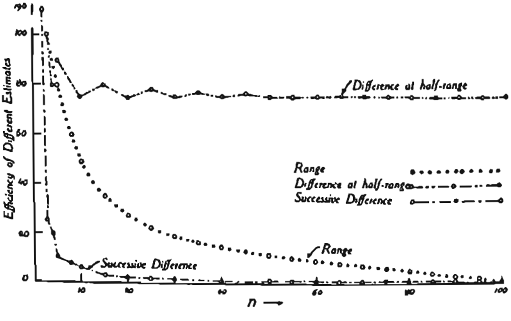

TABLE 4. EFFICIENCY OF THE DIFFERENT ESTIMATES OF REGRESSION COEFFICIENT IN RELATION TO THE LEAST SQUARE SOLUTION.

| n | Successive Difference | Difference at half range | Range |
|---|---|---|---|
| (1) | (2) | (3) | (4) |
| 2 | 100·00 | 100·00 | 100·00 |
| 3 | 25·00 | 100·00 | 100·00 |
| 4 | 20·00 | 80·00 | 90·00 |
| 5 | 10·00 | 90·00 | 80·00 |
| 10 | 3·03 | 75·76 | 49· |
| 15 | 1·25 | 80·00 | 35·00 |
| 20 | 0·75 | 75·10 | 27·14 |
| 25 | 0·46 | 78·00 | 22·15 |
| 30 | 0·33 | 75·08 | 18·71 |
| 35 | 0·24 | 77·14 | 16·19 |
| 40 | 0·10 | 75·05 | 14·27 |
| 45 | 0·14 | 76·67 | 12·75 |
| 50 | 0·12 | 75·03 | 11·53 |
| ∞ | 0·00 | 75·00 | 0·00 |

be preferred because for the purpose required the greater simplicity in calculation outbalances any loss of accuracy involved."

Hence, when regression coefficients have to be calculated on a large scale and a high degree of precision is not essential, the method of differences at half range may be recommended for actual computation. The method is simple and usually there is considerable saving in time. An actual procedure is shown in Appendix 1. Where greater precision is required, the method of least squares must of course be used.

## (6) SAMPLING EXPERIMENT.

A sampling experiment was conducted as follows: 144 samples of $x$ and $y$ are constructed with 4 pairs of values in each, with a population regression coefficient $(\beta) = 2$. Four estimates of regression coefficient were then made from each of the 144 sets The frequency distribution of each of these estimates is shown in Appendix 2. It should be sufficient here to show the close agreement between the observed and the expected values of the mean regression coefficients, standard deviations and the relative efficiencies of the four different estimates. This is given in Table 5.

TABLE 5. MEAN, STANDARD ERROR AND EFFICIENCY OF REGRESSION COEFFICIENT ($N = 144$, $n = 4$, $\sigma^2 = 1$).

| | MEAN | | STANDARD ERROR | | RELATIVE EFFICENCY | |
|---|---|---|---|---|---|---|
| | Observed | Expected | Observed | Expected | Observed | Expected |
| Least Square .. | 2·0034 ± ·0373 | 2·0000 | 0·4216 | 0·4172 | – | 100·00 |
| Successive Difference | 2·1722 ± ·0833 | 2·0000 | 0·9100 | 1·0000 | 21·30 | 20·00 |
| Difference at half range | 2·0372 ± ·0417 | 2·0000 | 0·4740 | 0·5000 | 79·10 | 80·00 |
| Range ... ... | 2·0417 ± ·0393 | 2·0000 | 0·4720 | 0·4713 | 79·78 | 90·00 |

It is a pleasure to acknowledge my indebtedness to Mr. S. N. Sen for drawing my attention to the present problem.

## Appendix 1.

### Methods of Calculating Regression Coefficient by the Method of Differences

| $x$ | $y$ | Successive Difference | Method of Difference at half range | Range |
|---|---|---|---|---|
| 1 | 3·84 | 1·74 | | |
| 2 | 5·58 | ... | | |
| 3 | 6·63 | 2·63 | | |
| 4 | 9·26 | .. | | |
| 5 | 11·42 | 3·25 | 7·58 | |
| 6 | 14·67 | ... | 9·09 | |
| 7 | 15·67 | 1·39 | 9·04 | |
| 8 | 17·06 | ... | 7·80 | 13·22 |

$$4\,\underline{)\,9·01} \qquad 16\,\underline{)\,33·51} \qquad 7\,\underline{)\,13·22}$$

$$b_1 = 2·25 \qquad b_2 = 2·09 \qquad b_3 = 1·89$$

The Least square estimate $b_0 = 2·015$

## Appendix. 2.

### Frequency Distribution of Regression Coefficient Calculated by Four different Methods.

($\beta = 2$, $n = 4$)

| | Least Square | Successive Difference | Difference at half range | Method of Range | | Least Square | Successive Difference | Difference at half range | Method of Range |
|---|---|---|---|---|---|---|---|---|---|
| ·3–·5 | ... | 1 | 1 | ... | 3·3 | ... | 5 | ... | 1 |
| ·7 | 1 | 3 | 0 | ... | 3·5 | .. | 5 | ... | 1 |
| ·9 | 0 | 4 | 1 | 1 | 3·7 | ... | 1 | ... | 0 |
| 1·1 | 2 | 5 | 6 | 8 | 3·9 | | 4 | ... | 1 |
| 1·3 | 6 | 13 | 4 | 6 | 4·1 | ... | 0 | ... | ... |
| 1·5 | 9 | 8 | 10 | 5 | 4·3 | ... | 1 | ... | .. |
| 1·7 | 18 | 10 | 18 | 14 | 4.5 | ... | 1 | .. | ... |
| 1·9 | 19 | 14 | 24 | 24 | 4·7 | ... | 1 | .. | ... |
| 2·1 | 25 | 17 | 19 | 20 | 4·9 | ... | 0 | ... | ... |
| 2·3 | 25 | 11 | 26 | 24 | 5·1 | ... | 1 | ... | ... |
| 2·5 | 28 | 11 | 16 | 17 | 5·3 | .. | 0 | . | ... |
| 2·7 | 7 | 16 | 12 | 13 | 5·5 | ... | 0 | ... | .. |
| 2·9 | 3 | 3 | 6 | 2 | 5·7 | ... | 0 | ... | .. |
| 3·1 | 1 | 8 | 1 | 3 | 5·9 | .. | 1 | ... | ... |
| | | | | | | 144 | 144 | 144 | 144 |

*Paper received 24 November, 1937*]