A SHORT CUT METHOD FOR LINEAR REGRESSION

by

Shian-Koong Perng

Thesis submitted to the Graduate Faculty of the Virginia Polytechnic Institute in candidacy for the degree of MASTER OF SCIENCE

in

Statistics

September, 1961

Blacksburg, Virginia
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>4</td>
</tr>
<tr>
<td>II. LINEAR REGRESSION</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Assumptions and Notation</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1 In General</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2 Equally Spaced Data</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Method of Least Squares</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Method of Group Averages</td>
<td>8</td>
</tr>
<tr>
<td>2.3.1 Optimum Division for The Residual Equation</td>
<td>11</td>
</tr>
<tr>
<td>2.3.2 Efficiency of b' As a Function of The Number of Omitted Residual Equations</td>
<td>13</td>
</tr>
<tr>
<td>2.4 A Numerical Example</td>
<td>15</td>
</tr>
<tr>
<td>III. QUADRATIC POLYNOMIAL REGRESSION</td>
<td>15</td>
</tr>
<tr>
<td>3.1 Method of Least Squares</td>
<td>18</td>
</tr>
<tr>
<td>3.2 Method of Group Averages</td>
<td>20</td>
</tr>
<tr>
<td>3.3 Optimum Division of The Residual Equations</td>
<td>23</td>
</tr>
<tr>
<td>IV. FUNCTIONAL RELATIONSHIP SITUATION</td>
<td>27</td>
</tr>
<tr>
<td>4.1 Assumptions and Notation</td>
<td>27</td>
</tr>
<tr>
<td>4.2 Maximum Likelihood Method</td>
<td>28</td>
</tr>
<tr>
<td>Chapter</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.3 Method of Group Averages</td>
<td>28</td>
</tr>
<tr>
<td>4.4 Comparison of The Relative Efficiency of Two Estimators of $\beta$</td>
<td>33</td>
</tr>
<tr>
<td>4.4.1 Variance of $b_B$</td>
<td>33</td>
</tr>
<tr>
<td>4.4.2 Variance of $b_M$</td>
<td>34</td>
</tr>
<tr>
<td>4.4.3 The Relative Efficiency of $b_B$ and $b_M$</td>
<td>38</td>
</tr>
<tr>
<td>4.5 A Numerical Example</td>
<td>40</td>
</tr>
<tr>
<td>V. CONCLUSION</td>
<td>42</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>43</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>44</td>
</tr>
<tr>
<td>VITA</td>
<td>45</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

The method of Least Squares is usually used to estimate the parameters in a linear regression situation. It is however sometimes desirable to use some short-cut methods of estimation, which require less computational effort.

Bose [2] developed three alternative methods for evaluating the regression coefficient for relationship between variables $x$ and $y$, where the independent variable $x$ is observed at equally spaced intervals. The three methods are (1) method of Successive Differences, (2) method of Difference at Half Range, and (3) method of Range. Nair and Shrivastava [6] developed the method of Group Averages for the same case, and pointed out that Bose's methods could be seen as special cases of the method of Group Averages.

It is the purpose of this paper to review several versions of the method of Group Averages, and to compare their efficiencies relative to the method of Least Squares. The use of this method is also discussed for the case when the independent variable is measured with error, in which case the Least Squares method is known to produce biased results.
II. LINEAR REGRESSION

2.1 Assumptions and Notation

2.1.1 In General

Consider two sets of variables

\[ x_1 \ x_2 \ \ldots \ \ldots \ x_n \]
\[ y_1 \ y_2 \ \ldots \ \ldots \ y_n \]

where \( x_i \) are observed sample values of the independent variable, assumed to be accurately known, and the \( y_i \) are sample values of the dependent variates. In practice this means that \( x \) is really known with an error which is so small that it can be neglected. The expected value of \( y_i \), denoted by \( Y_i \), is defined

\[ \mathcal{E}(y_i) = Y_i \]

and a single linear relationship is assumed to hold between the value \( x_i \) and \( Y_i \). That is to say

\[ Y_i = \alpha + \beta x_i \]

or

\[ y_i = \alpha + \beta x_i + \epsilon_i \]

where \( \mathcal{E}(\epsilon_i \epsilon_j) = 0 \) for all \( i \neq j \)

and \( \epsilon_i \sim N(0, \sigma^2) \quad i = 1, 2, \ldots, n \).

The normality of \( \epsilon_i \) is only necessary for tests of hypotheses.
The problem to be solved is: given a set of paired observations

\[
\begin{cases}
x_1, x_2 \ldots x_n \\
y_1, y_2 \ldots y_n
\end{cases}
\]

to estimate, and to make tests of hypotheses concerning the parameters \( \alpha \), \( \beta \), and \( \sigma^2 \).

2.1.2 Equally Spaced Data

Suppose \( n \) values of \( x_1 \) be equally spaced, which, for the sake of simplicity, are assumed to be \( 1, 2, \ldots n \). Therefore the problem now is to fit a straight line of the form:

\[
y_1 = \alpha + \beta x_1,
\]

where \( x_1 = 1, 2, \ldots n \).

2.2 Method of Least Squares

Given a set of paired observations

\[
\begin{cases}
x_1, x_2 \ldots x_n \\
y_1, y_2 \ldots y_n
\end{cases}
\]

If the following linear relationship holds, i.e.,

\[
y_1 = \alpha + \beta x_1 + e_1,
\]

it is desired to minimize, with respect to \( \alpha \) and \( \beta \)
\[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2, \]

resulting in the Least Squares estimators of \( \alpha \) and \( \beta \):

\[ a = \bar{y} - bx \]
\[ b = -\frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = -\frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \]

and the variances

\[ \text{Var} (b) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
\[ \text{Var} (a) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} . \]

We can use Student's t as a test criterion to test the null hypothesis \( \beta = \beta_0, \quad \alpha = \alpha_0 \):

\[ t = \frac{b - \beta_0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}, \quad \text{d.f.} \ (n - 2), \]

and
\[
t = \frac{a - a_0}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}}}, \quad \text{d.f. (n - 2)},
\]

where \( s^2 = \frac{n}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-2) \),

and \( \hat{y}_i = a + bx_i \).

Now we assume \( x_i \) to be equally spaced. Then the Least Squares estimators of \( a, b \), and their variances will be:

\[
a = \bar{y} - b \left( \frac{n+1}{2} \right), \quad (2.2.1)
\]

\[
\frac{6}{n-1} 5 + \frac{12}{n(n^2-1)} \sum_{i=1}^{n} iy_i = b = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} (1 - \frac{n+1}{2}) y_i, \quad (2.2.2)
\]

\[
\text{Var} (a) = \frac{2(2n+1)\sigma^2}{n(n-1)} , \quad (2.2.3)
\]

and

\[
\text{Var} (b) = \frac{12 \sigma^2}{n(n^2-1)} . \quad (2.2.4)
\]

2.3 **Method of Group Averages**

Nair and Shrivastava [6] gave the Group Averages Method as follows. Under the same conditions as in previous sections, the following residual equations are obtained:
\[
\begin{align*}
Y_1 - a - \beta 1 &= e_1 \\
\vdots \\
Y_i - a - \beta 2 &= e_i \\
\vdots \\
Y_n - a - \beta n &= e_n
\end{align*}
\]

We divide the \( n \) residual equations into three groups, say, where the first \( p \) equations form the first group, the last \( q \) equations form the third group, and the middle \( n - p - q - r \) equations form the second group. By equating to zero the sum of the residuals in the first group we get a single equation, and by equating to zero the last group we get another equation. Ignoring the middle group, we have two simultaneous equations in the two unknowns \( a \) and \( \beta \):

\[
p a + (1 + 2 + \ldots + p) \beta = \frac{p}{1} \sum_{i=1}^{p} y_i
\]

\[
q a + [(n - q + 1) + \ldots + n] \beta = \frac{n}{n-q+1} \sum_{i=n-q+1}^{n} y_i
\]

Solving for \( a \) and \( \beta \), we obtain estimators \( a' \) and \( b' \):

\[
b' = \frac{2[\frac{1}{p} \sum_{i=p+1}^{n} y_i - \frac{1}{p} \sum_{i=1}^{p} y_i]}{2n - p - q}
\]

\[
= \frac{2[\bar{y}_3 - \bar{y}_1]}{2n - p - q}
\]

where \( \bar{y}_i = \text{mean of } i\text{th group} \),
and

\[ a' = \frac{2n+1-q}{p} \sum_{1}^{p} y_i - \frac{p+1}{q} \sum_{n-q+1}^{n} y_i \]
\[ = \frac{(2n+1-q) \bar{y}_1 - (p+1) \bar{y}_3}{2n - p - q} \]  \hspace{1cm} (2.3.2)

Using standard formulae and the conditions of the model,

\[ \text{Var} (b') = \frac{4}{(2n-p-q)^2} \text{Var} (\bar{y}_3 - \bar{y}_1) \]
\[ = \frac{4}{(2n-p-q)^2} \left[ \frac{\sigma^2}{q} + \frac{\sigma^2}{p} \right] \]
\[ = \frac{4(p+q)\sigma^2}{pq(2n-p-q)^2} \]  \hspace{1cm} (2.3.3)

and

\[ \text{Var} (a') = \frac{(2n+1-q)^2 \text{Var} (\bar{y}_1) + (p+1)^2 \text{Var} (\bar{y}_3)}{(2n-p-q)^2} \]
\[ = \sigma^2 \left[ \frac{(2n+1-q)^2}{p(2n-p-q)^2} + \frac{(p+1)^2}{q(2n-p-q)^2} \right] \]
\[ = \sigma^2 \left[ \frac{1}{p} \left(1+\frac{p+1}{2n-p-q}\right)^2 + \frac{1}{q} \left(1+\frac{p+1}{2n-p-q}\right)^2 \right] \]  \hspace{1cm} (2.3.4)

We can use Student's t as a test criterion to test null hypotheses \( \beta = \beta_0 \) and \( \alpha = \alpha_0 \).
where

\[ s^2 = \frac{\sum (y_i - \hat{Y}_i)^2}{n - 2} \]

and

\[ \hat{Y}_i = a' + b' x_i \]

2.3.1 Optimum Division for the Residual Equation

It is of interest to determine best values for \( p \) and \( q \), so as to maximize the efficiency of \( b' \), with respect to \( b \) where \( b \) is the estimator by Least Squares method, which is known to be the most efficient estimator of \( \beta \).

Defining the efficiencies as the inverse ratio of variances, we have:

\[
E(b') = \frac{\text{Var}(b)}{\text{Var}(b')} = \frac{3pq(2n-p-q)^2}{n(n^2-1)(p+q)} \quad (2.3.1.1)
\]

\[
E(a') = \frac{\text{Var}(a)}{\text{Var}(a')} = \frac{2(2n+1)/n(n+1)}{1/p(1+\frac{p+1}{2n-p-q})^2 + 1/\frac{p+1}{q2n-p-q}^2} \quad (2.3.1.2)
\]
Differentiating the logarithmic function of $E(b')$ with respect to $p$ and $q$ we obtain

\[
\begin{align*}
\frac{1}{p} - \frac{2}{2n-p-q} - \frac{1}{p+q} &= 0 \\
\frac{1}{q} - \frac{2}{2n-p-q} - \frac{1}{p+q} &= 0
\end{align*}
\]

Solving for $p$ and $q$ we have

\[ p = q = \frac{n}{3} \]

This result was also given by Nair and Shrivastava [6].

Thus the most efficient method of estimating $\beta$ by the methods of Group Averages is to divide the whole set of residual equations into three equal parts thus ignoring the middle set. Substituting $p = q = \frac{n}{3}$ into the efficiency equations (2.3.1.1), and (2.3.1.2), we have

\[
E(b') = \frac{\text{Var} (b)}{\text{Var} (b')} = \frac{8}{9} \cdot \frac{n^2}{n^2-1} > \frac{8}{9} \quad , \quad (2.3.1.3)
\]

\[
E(a') = \frac{\text{Var} (a)}{\text{Var} (a')} = \frac{16n^2(2n+1)}{3(n-1)(13n^2+18n+9)}
\]

\[
= \frac{32}{39} \cdot \frac{n^2(n+1)}{(n^2+13n^2+9)(n-1)}
\]

\[
= \frac{32}{39} \cdot \frac{n^3+\frac{1}{2}n^2}{n^3+\frac{5}{13}n^2-\frac{9}{13}n-\frac{9}{13}} > \frac{32}{39} \quad . \quad (2.3.1.4)
\]
If \( n \) is not a multiple of \( 3 \), two cases are possible, namely \( n = 3m \pm 1 \); in either case the optimum value of \( p \) and \( q \) is \( m \).

### 2.3.2 Efficiency of \( b' \) as A Function of The Number of Omitted Residual Equations

The efficiency of \( b' \) has been derived as:

\[
E(b') = \frac{\text{Var}(b')}{\text{Var}(b)} = \frac{3pq(2n-p-q)^2}{n(n^2-1)(p+q)}.
\]

In general, if we take an equal, but arbitrary, number of residual equations, i.e. \( p = q \), we have

\[
E(b') = \frac{6p(n-p)^2}{n(n^2-1)}.
\]

In this case the number of omitted residual equations, \( r = n - p - q = n - 2p \), and we have

\[
E(b') = \frac{3(n^3+n^2r-nr^2-r^3)}{4n(n^2-1)}.
\]  \hspace{1cm} (2.3.2.1)

Setting \( \frac{dE(b')}{dr} \) equal to zero, we have the previously obtained result:

\[
\frac{dE(b')}{dr} = n^2 - 2nr - 3r^2 = 0
\]

\[\therefore \quad r = \frac{n}{3}.\]

The efficiencies of the Group Averages estimator as a function of \( r \) for selected values of \( n \) are in Chart 1.
Chart 1

Percentage Efficiencies of $b'$ As A Function of $Y$ For $n=9, 21, 30$
2.4 A Numerical Example

Snedecor [8] gave the following numerical example. The heights of four soybean plants were measured a week after emergence, showing a mean height of 5 centimeters. Four other plants were similarly measured two weeks after emergence, and so on to the seventh week; see Table 1.

Table 1
Data for Example

<table>
<thead>
<tr>
<th>Age in Weeks</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height in centimeters</td>
<td>5</td>
<td>13</td>
<td>16</td>
<td>23</td>
<td>33</td>
<td>38</td>
<td>40</td>
</tr>
</tbody>
</table>

Using the Least Squares method for estimating parameters of a regression of height on age, we have:

\[
\bar{x} = 4 ,
\]

\[
\bar{y} = 24 ,
\]

\[
\sum_{1}^{7} (x_i - \bar{x})^2 = 28 ,
\]

\[
\sum_{1}^{7} (y_i - \bar{y}) = 1,080 ,
\]

\[
\sum_{1}^{7} (y_i - \bar{y})(x_i - \bar{x}) = 172 ,
\]
The estimated regression line equations:

\[ y_1 = 6.143x_1 - 0.572, \]

and

\[ s^2 = \frac{\sum (y_i - \bar{y}_1)^2}{5} \]

\[ = \frac{1}{3} [ \sum (y_i - \bar{y})^2 - \frac{7}{1} \frac{(\Sigma (x_i - \bar{x})(y_i - \bar{y}))^2}{\Sigma (x_i - \bar{x})^2} ] \]

\[ = \frac{23.92}{5} = 4.784. \]

The test for \( H_0: \beta = 0 \), \( H_A: \beta > 0 \),

\[ t_{0.05}(5) = \frac{6.143}{\sqrt{4.784}} \quad \text{d.f.} = 5 \]

\[ = 14.84* , \]

and we reject the null hypothesis.

Using the method of Group Averages, we obtain the estimates:

\[ b' = \frac{2(\bar{y}_3 - \bar{y}_1)}{\frac{n}{3}} = 6.42. \]
\[ a' = \frac{(5n + 3)\bar{y}_1 - (n + 3)\bar{y}_3}{4n} = -1.71 \]

and the estimated regression line equation is

\[ Y_1 = 6.42x_1 - 1.71 \]

The test for \( H_0: \beta = 0 \), \( H_A: \beta > 0 \),

\[ t_{0.05(5)} = \frac{6.42}{\sqrt{\frac{18}{n^2} s^2}} \]

\[ = \frac{6.42}{1.05} = 6.11^* \]

and we reject the null hypothesis. Note that the estimates and tests are quite comparable for both methods.
III. QUADRATIC POLYNOMIAL REGRESSION

As before we assume $x_1$ to be equally spaced, and for the sake of simplicity, the $x_1$'s will be assumed to be 1, 2, ... n. We now desire to estimate a quadratic regression of the form:

$$y_i = a + \beta x_i + \gamma x_i^2.$$ 

The estimators of $a$, $\beta$, and $\gamma$ will be obtained by the method of Least Squares and the method of Group Averages. The optimum division of residual equations for the Group Averages method will also be obtained.

3.1 Method of Least Squares

Estimates of $a$, $\beta$, and $\gamma$ are obtained by minimizing the following function:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - \beta i - \gamma i^2)^2,$$

which results in the normal equations:

$$\begin{align*}
\sum_{i=1}^{n} \sum_{i=1}^{n} a + \beta \sum_{i=1}^{n} i + \gamma \sum_{i=1}^{n} i^2 &= \sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} a \sum_{i=1}^{n} i + \beta \sum_{i=1}^{n} i^2 + \gamma \sum_{i=1}^{n} i^3 &= \sum_{i=1}^{n} i y_i \\
\sum_{i=1}^{n} a \sum_{i=1}^{n} i^2 + \beta \sum_{i=1}^{n} i^3 + \gamma \sum_{i=1}^{n} i^4 &= \sum_{i=1}^{n} i^2 y_i
\end{align*}$$
Using the well known relationship

\[
\sum_{l=1}^{n} i = \frac{1}{2} n(n + 1) \\
\sum_{l=1}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1) \\
\sum_{l=1}^{n} i^3 = \left(\frac{1}{2} n(n+1)\right)^2 \\
\sum_{l=1}^{n} i^4 = \frac{1}{30} n(n+1)(6n^3+9n^2+n-1)
\]

we solve the simultaneous equations for \(a\), \(b\), and \(c\), and obtain the Least Squares estimators \(\hat{a}\), \(\hat{b}\), \(\hat{c}\):

\[
a = \frac{3[H_1(3n^2+3n+2)-5H_2(2n+1)+10H_3]}{n(n-1)(n-2)},
\]

\[
b = \frac{6[H_1(n+1)(n+2)(-2n-1)+2H_2(16n^2+30n+11)-30(n+1)H_3]}{n(n^2-1)(n^2-4)},
\]

and

\[
c = \frac{30[H_1(n+1)(n+2)-6H_2(n+1)+6H_3]}{n(n^2-1)(n^2-4)},
\]

where

\[
H_1 = \frac{n}{1} y_1, \quad H_2 = \frac{n}{1} i y_1, \quad H_3 = \frac{n}{1} i^2 y_1.
\]

Further, we have

\[
\text{Var}(a) = \frac{3(3n^2+3n+2)\sigma^2}{n(n-1)(n-2)} \quad (3.1.1)
\]
\[ \text{Var} \ (b) = \frac{12(16n^2 + 30n + 11)}{n(n^2 - 1)(n^2 - 4)} \sigma^2 \]  
(3.1.2)

and \[ \text{Var} \ (c) = \frac{180 \sigma^2}{n(n^2 - 1)(n^2 - 4)} \]  
(3.1.3)

3.2 Method of Group Averages

Nair and Shrivastava [6] give the method of Group Averages for a quadratic polynomial as follows. Using the model

\[ y_1 = \alpha + \beta i + \gamma i^2 + e_1 \]

we have the following residual equations:

\[ \begin{cases} 
  y_1 - \alpha - \beta - \gamma = e_1 \\
  y_2 - \alpha - 2\beta - 2^2\gamma = e_2 \\
  \vdots \\
  y_n - \alpha - n\beta - n^2\gamma = e_n 
\end{cases} \]

Here we require three simultaneous equations to solve for \( a', b', c' \), the estimators of \( \alpha, \beta, \gamma \). From our experience of fitting a straight line, we will assume that these three equations are based on equal numbers of residual equations. We take the first \( p \) equations, middle \( p \) equations, last \( p \) equations to form the first, second and last group. Again using our experience from the linear case, we will assume that some equations between these groups will not be used; thus \( w \) and \( u \) equations, say, will be omitted between the
first and second, and second and third sets of \( p \), respectively.
Thus

\[
p + w + p + u + p = n.
\]

From the first, second and last groups, we get three simultaneous equations:

\[
\begin{align*}
\ell_1 a + k_1 \beta + m_1 \gamma &= S_1 \\
\ell_2 a + k_2 \beta + m_2 \gamma &= S_2 \\
\ell_3 a + k_3 \beta + m_3 \gamma &= S_3
\end{align*}
\]

where, under assumption of equally spaced \( x \)'s,

\[
\ell_1 - \ell_2 - \ell_3 = p
\]

\[
k_1 = \sum_{i=1}^{p} i = \frac{1}{2} p(p+1)
\]

\[
k_2 = \sum_{i=p+w+1}^{2p+w} \frac{i}{p+w+1} = \frac{1}{2} p(3p+2w+1)
\]

\[
k_3 = \sum_{i=n-p+1}^{n} \frac{i}{n-p+1} = \frac{1}{2} p(2n-p+1)
\]

\[
m_1 = \sum_{i=1}^{p} i^2 = \frac{1}{6} p(p+1)(2p+1)
\]

\[
m_2 = \sum_{i=p+w+1}^{2p+w} \frac{i^2}{p+w+1} = \frac{1}{6} p \left[ \frac{(3p+2w+1)(4p+2w+1)}{+2(p+w)(p+w+1)} \right]
\]
\[ m_3 = \sum_{n-p+1}^{n} i^2 = \frac{1}{6} p[(2n-p+1)(2n+1)+2(n-p)(n-p+1)] \]

\[ S_1 = \sum_{1}^{p} y_i \]

\[ S_2 = \sum_{p+w+1}^{2p+w} y_i \]

\[ S_3 = \sum_{n-p+1}^{n} y_i \]

Note that \( u \) does not appear in these equations, since, given \( n, p, \) and \( w \) uniquely determine \( u. \)

Solving the simultaneous equations for \( a', b', c' \) the estimators of \( a, b, \gamma. \) We obtain:

\[ a' = \frac{S_1 V_1 + S_2 V_2 + S_3 V_3}{12p(n-2p-w)(n-p)(p+w)} \]

\[ b' = \frac{S_1 Z_1 + S_2 Z_2 + S_3 Z_3}{p(n-2p-w)(n-p)(p+w)} \]

\[ c' = \frac{S_1(n-2p-w)+S_2(n-p)+S_3(p+w)}{p(n-2p-w)(n-p)(p+w)} \]

where

\[ V_1 = 18n^2p - 46np^2 + 12n^2w - 48pwn + 22p^2w \]

\[ + 6n^2 - 6np - 12p^2 + 4n - 4w - 18wp \]

\[ + 20p^3 - 12n^2w - 3p + 6pw^2 - 6w^2 \]
\[ v_2 = 10np^2 + 6np - 4n - 4p^3 + 4p - 6n^2 - 6n^2p \]

\[ v_3 = 8p^3 + 12p^2 + 4p + 18pw + 4w + 6w^2 + 14p^2w + 6pw^2 \]

\[ z_1 = 2p^2 + 2p + 3nw + w + w^2 - n^2 + np - n \]

\[ z_2 = n^2 - np - p + n = (n+1)(n-p) \]

\[ z_3 = -2p^2 - p - 3pw - w - w^2 = -(p+w)(2p+w+1) \]

3.3 **Optimum Division of the Residual Equations**

It is of interest to find the values for \( p, w, u \), such that the efficiencies of \( c' \) and \( b' \) will be maximized relative to the Least Squares estimators. First we obtain:

\[
\text{Var } (c') = \frac{\sigma^2[(n-2p-w)^2+(n-p)^2+(p+w)^2]}{p(n-2p-w)^2(n-p)^2(p+w)^2},
\]

and

\[
\text{Var } (b') = \frac{\sigma^2[z_1^2+z_2^2+z_3^2]}{p(n-2p-w)^2(n-p)^2(p+w)^2}.
\]

Hence the efficiencies of \( c' \), \( b' \) as compared to Least Squares estimators \( c \) and \( b \) are:

\[
E(c') = \frac{\text{Var } (c)}{\text{Var } (c')} = \frac{180}{n(n^2-1)(n^2-4)} \cdot \frac{p(n-2p-w)^2(n-p)^2(p+w)^2}{(n-2p-w)^2+(n-p)^2+(p+w)^2},
\]

(3.3.1)
and

\[ E(b^*) = \frac{\text{Var}(b)}{\text{Var}(b^*)} \]  \hspace{1cm} (3.3.2)

\[ = \frac{12(16n^2+30n+11)}{n(n^2-1)(n^2-4)} \frac{p(n-2p-w)^2(n-p)^2(p+w)^2}{(z_1^2+z_2^2+z_3^2)} , \]

where

\[ z_1^2+z_2^2+z_3^2 = (2p^2+2p+3nw+w-n^2+n-3p)^2 + (n+1)^2(n-p)^2 \]
\[ + (p+w)^2(2p+w+1)^2 . \]

Since we are fitting a quadratic regression curve, we are more interested in \( E(c^*) \) than in \( E(b^*) \). Since it is not likely that both efficiencies will be maximized by one division pattern, we will obtain those values for \( p, w, u \), such that \( E(c^*) \) will be maximized. If we maximize the numerator of (3.3.1) with respect to \( w \) we have:

\[ \frac{1}{p+w} - \frac{1}{n-2p-w} = 0 \]
\[ w = \frac{1}{2} (n-3p) . \]  \hspace{1cm} (3.3.3)

Fortunately (3.3.3) simultaneously minimizes the denominator with respect to \( w \). Therefore (3.3.3) is the condition to maximize \( E(c^*) \). Thus

\[ w = u = \frac{1}{2} (n-3p) , \]

since \( n = 3p + w + u \). With this value of \( w \), (3.3.1) reduces to
\[ E(c') = \frac{15}{2} \cdot \frac{p(n-p)^4}{n(n^2-1)(n^2-4)} \]

which attains maximum value when \( p = n/5 \). Thus for maximum efficiency of \( c' \) we should choose our groups in such a way that

\[ p = w = u = \frac{n}{5} \]

In other words, we should divide the set of residual equations into five equal parts and then use the first, third, and fifth parts for the residual equations, ignoring the second and fourth sets. This result is also given by Nair and Shrivastava. [6]

This division of the residual equations does not maximize the efficiency of \( b' \). The efficiencies of \( a' \) and \( b' \), using this division are, however, quite satisfactory, as will be shown. Let \( w = p = \frac{n}{5} \), then \( a' \), \( b' \), \( c' \), and their efficiencies are:

\[ a' = \frac{5}{48n^3} [S_1(67n^2+105n+50) - S_2(26n^2+150n+100) + S_3(7n^2+45n+50)] , \]

\[ b' = \frac{25[(10n+10)S_2 - (3n+5)S_3 - (7n+5)S_1]}{8n^3} , \]

and

\[ c' = \frac{125[S_3 + S_1 + 2S_2]}{8n^3} ; \]
E(a') = \frac{\text{Var}(a)}{\text{Var}(a')} = \frac{6912n^4(3n^2+3n+2)}{5(n-1)(n-2)} \phi_1 ,

\text{where } \phi_1 = 5214n^4 + 22500n^3 + 48150n^2 + 45000n + 15000 ,

E(b') = \frac{\text{Var}(b)}{\text{Var}(b')} = \frac{384 n^4(16n^3+30n+11)}{125 (n^2-1)(n^2-4)} \phi_2 ,

\text{where } \phi_2 = 79n^2 + 150n + 75 ,

and

E(c') = \frac{\text{Var}(c)}{\text{Var}(c')} = \frac{384 n^4}{625 (n^2-1)(n^2-4)} .

The efficiencies of a', b', c', are presented in Table 2 for a selected set of values of n.

Table 2

<table>
<thead>
<tr>
<th>n</th>
<th>E(a')</th>
<th>E(b')</th>
<th>E(c')</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>88.14</td>
<td>76.33</td>
<td>76.19</td>
</tr>
<tr>
<td>10</td>
<td>79.76</td>
<td>65.19</td>
<td>64.65</td>
</tr>
<tr>
<td>15</td>
<td>79.00</td>
<td>63.47</td>
<td>62.83</td>
</tr>
<tr>
<td>20</td>
<td>78.90</td>
<td>62.90</td>
<td>62.22</td>
</tr>
<tr>
<td>25</td>
<td>78.92</td>
<td>62.64</td>
<td>61.93</td>
</tr>
</tbody>
</table>
IV. FUNCTIONAL RELATIONSHIP SITUATION

4.1 Assumptions and Notation

We will now consider the problem of estimating a linear relationship between two variables, $X$ and $Y$, when both are subject to errors of observation. The problem is no longer to estimate a regression coefficient in the usual sense; it is rather to estimate the constants in a functional relation:

$$ Y_i = \alpha + \beta X_i; $$

however, $Y_i$ and $X_i$ cannot be directly observed because of errors of observation. Instead of $Y_i$ and $X_i$, values denoted $y_i$ and $x_i$ are observed where

$$ y_i = Y_i + e_i, $$
$$ x_i = X_i + d_i, $$

$$ e_i \sim \text{NID} (0, \sigma_e^2), \quad i = 1, \ldots, n, $$
$$ d_j \sim \text{NID} (0, \sigma_d^2), \quad j = 1, \ldots, n, $$

and $$ \mathcal{C}(e_i, d_j) = 0 \quad \text{for all } i, j. $$

We shall use the Maximum Likelihood method and the Group Averages method to estimate $\alpha$, $\beta$, $\sigma_e^2$, $\sigma_d^2$, and compare, under some restrictions, the estimators obtained by these two methods.
4.2 **Maximum Likelihood Method, \( \lambda = \sigma_d^2/\sigma_e^2 \) Known**

Lindley [5] gave the maximum likelihood estimators for the functional relationship situation as follows under the assumption \( \lambda = \frac{\sigma_d^2}{\sigma_e^2} \) is known:

\[
\hat{\beta} = b_M = \frac{1}{\lambda} \left[ n \sum (y_i - \bar{y})^2 - \frac{1}{\lambda} n \sum (x_i - \bar{x})^2 \right]
\]

where

\[
U = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) - \frac{1}{\lambda} \sum_{i=1}^{n} (x_i - \bar{x})}{2 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}
\]

\[
\hat{a} = \bar{y} - \hat{\beta} \bar{x}
\]

and

\[
\sigma_e^2 = \frac{1}{n-2} \left[ \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right]
\]

where the sign used for the first term in \( \hat{\beta} \) is that which will maximize the likelihood function.

Lindley pointed out that it is not feasible to obtain reasonable maximum likelihood estimators without additional information; a knowledge of the ratio \( \lambda = \sigma_d^2/\sigma_e^2 \) seemed to be the most convenient form of additional information to use.

4.3 **Method of Group Averages**

Wald [9] suggested that when the number of observations, \( n \), is even one can order the \( x_i \)'s such that \( x_i < x_{i+1} \),
divide the observations into two equal groups, and estimate the parameters of the functional relation

\[ Y_1 = a + \beta X_1 \]

as follows:

\[ \hat{\beta} = b_w = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} \quad (4.3.1) \]

\[ \hat{a} = a_w = \bar{y} - b_w \bar{x} \quad (4.3.2) \]

\[ \hat{\sigma}_d^2 = (s_x^2 - \frac{s_{xy}}{b_w}) \frac{n}{n-2} \quad (4.3.3) \]

\[ \hat{\sigma}_e^2 = (s_y^2 - b_w s_{xy}) \frac{n}{n-2} \quad (4.3.4) \]

where

\[ \bar{y} = \frac{1}{n} \sum_{1}^{n} y_1 / n, \quad \bar{y}_1 = \frac{\bar{x}}{\bar{x}_2} \frac{n}{n} / \bar{x}_2, \quad \bar{y}_2 = \frac{n}{n} / \bar{x}_2 + 1 \]

\[ \bar{x} = \frac{1}{n} \sum_{1}^{n} x_1 / n, \quad \bar{x}_1 = \frac{\bar{y}}{\bar{y}_1} \frac{n}{n} / \bar{y}_1, \quad \bar{x}_2 = \frac{n}{n} / \bar{x}_2 + 1 \]

\[ s_x^2 = \frac{1}{n} \left[ \sum_{1}^{n} (x_1 - \bar{x}_1)^2 + \sum_{\frac{n}{2}+1}^{n} (x_1 - \bar{x}_2)^2 \right] \]

\[ s_y^2 = \frac{1}{n} \left[ \sum_{1}^{n} (y_1 - \bar{y}_1)^2 + \sum_{\frac{n}{2}+1}^{n} (y_1 - \bar{y}_2)^2 \right] \]
and
\[
S_{xy} = \frac{n}{\frac{n}{2} + 1} \frac{\sum (x_1 - \bar{x}_1)(y_1 - \bar{y}_1) + \sum (x_1 - \bar{x}_2)(y_1 - \bar{y}_2)}{n} .
\]

All these estimators are consistent provided
\[
\Pr \{ |d_1| \geq \frac{1}{2} c \} \text{ is negligible}
\]
where \( c = \min |x_1 - x_{i+1}| \).

That is to say the probabilities that the \( d_1 \) (errors of observations on \( X \)) have values exceeding \( \pm \frac{1}{2} c \) may be neglected.

This assumption simply insures that the ordering of the \( x_1 \) is identical with the ordering according to the \( I_1 \).

Bartlett [1] provided a modification of Wald's method by using three groups instead of two groups to estimate \( \alpha, \beta, \sigma_e^2, \) and \( \sigma_d^2 \). The resulting estimators are:

\[
\hat{\beta} = b_B = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} \quad (4.3.5)
\]

\[
\hat{\alpha} = a_B = \bar{y} - b_B \bar{x} \quad (4.3.6)
\]

\[
\hat{\sigma}_e^2 = (S_y^2 - b_B S_{xy}) \frac{n}{n-3} \quad (4.3.7)
\]

\[
\hat{\sigma}_d^2 = (S_x^2 - S_{xy}/b_B) \frac{n}{n-3} \quad (4.3.8)
\]
where

\[
\bar{y}_3 = \frac{n}{3n+1} y_1/n \frac{2n}{3} ,
\]

\[
\bar{y}_2 = \frac{n}{3n+1} y_1/n \frac{2n}{3} ,
\]

\[
\bar{y}_1 = \frac{n}{3} y_1/n \frac{2n}{3} ,
\]

\[
\bar{x}_3 = \frac{n}{3n+1} x_1/n \frac{2n}{3} ,
\]

\[\bar{x}_2 \text{ and } \bar{x}_1 \text{ are defined as they are for } y ,\]

\[
S_x^2 = \frac{1}{n} \left[ \frac{n}{3} \sum (x_1 - \bar{x}_1)^2 + \frac{2n}{3} \sum (x_1 - \bar{x}_2)^2 + \frac{n}{3} \sum (x_1 - \bar{x}_3)^2 \right] ,
\]

\[S_y^2 \text{ and } S_{xy} \text{ are defined analogously.}\]

Nair and Banerjee [7] used a sampling experiment to show that Bartlett's method gave a more efficient estimator of \( \beta \). They assumed the functional relationship to be

\[Y_1 = 1 + X_1 ,\]

i.e. \( \alpha = \beta = 1 \).

The \( X_1 \) were given integral values from 1 to 30, thus the
corresponding values of \( Y_i \) were 2 to 31. However the observed variables \( x_i \) and \( y_i \) were of the form

\[
x_i = X_i + d_i, \\
y_i = X_i + e_i, \text{ where}
\]

\[
d_i \sim \text{NID}(0, 0.01) \\
e_i \sim \text{NID}(0, 0.01).
\]

Using random normal deviates, they obtained the required samples of size 30, with \( x_i, y_i \) both subject to uncorrelated errors. Since the variances were quite small the ordering according to \( x_i \) was, for all practical purposes, identical to the ordering according to the \( X_i \).

One hundred samples of 30 paired observations were obtained, and estimates of \( \alpha \) and \( \beta \) were obtained by using both Wald's and Bartlett's methods. The results of the samples are presented in Table 3.

**Table 3**

Results of Sampling Experiment

<table>
<thead>
<tr>
<th>Method of Estimating</th>
<th>( a )</th>
<th>S. D.</th>
<th>( b )</th>
<th>S. D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald's</td>
<td>0.998583</td>
<td>0.0104</td>
<td>1.010063</td>
<td>0.0991</td>
</tr>
<tr>
<td>Bartlett's</td>
<td>0.999793</td>
<td>0.0035</td>
<td>1.003634</td>
<td>0.0767</td>
</tr>
</tbody>
</table>
The above table indicates that while both the methods give consistent estimates for $\alpha$ and $\beta$, Bartlett's method gives the more efficient estimates. Therefore Bartlett's method will be used in succeeding sections.

4.4 Comparison of the Relative Efficiency of Two Estimators of $\beta$

For the comparison of the efficiency of the two estimators, $b_B$ (group average) and $b_M$ (maximum likelihood), we need to find their variances.

4.4.1 Variance of $b_B$

Dorff and Gurland [3] gave the asymptotic variance of $b_B$ as follows, using the notation of Section 4.1:

$$
\text{Var} \ (b_B) = \frac{2n}{3} \left( \sigma_e^2 + \beta^2 \sigma_d^2 \right) \frac{n}{3} \left( \frac{\sum X_i}{3} - \frac{\sum X_i}{1} \right)^2
$$

Since we shall compare the $\text{Var} \ (b_B)$ with the $\text{Var} \ (b_M)$ under the assumption $\lambda \sigma_e^2 = \sigma_d^2$, the variance formula to be used is:

$$
\text{Var} \ (b_B) = \frac{2n}{3} \sigma_e^2 (1 + \lambda \beta^2) \frac{n}{3} \left( \frac{\sum X_i}{3} - \frac{\sum X_i}{1} \right)^2
$$

(4.4.1.1)
4.4.2 Variance of $b_M$

The general procedure for obtaining an asymptotic variance of a maximum likelihood estimator is as follows:

1. Take the partial derivatives of the logarithms of the likelihood function with respect to all parameters, i.e.,
   \[ \frac{\partial \log L}{\partial \theta_i}, \quad i = 1 \ldots k \]
   where $k$ is the number of parameters.

2. Take the second derivatives for all equations in (1.) with respect to all parameters again, i.e.,
   \[ \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \quad \text{for } i, j = 1, 2, \ldots k. \]

3. Take the expected value of each second partial derivative, i.e.
   \[ \mathcal{E}(\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}). \]

4. Form the "information" matrix whose $i$, $j$th element is
   \[ -\mathcal{E}(\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}) \quad i, j = 1 \ldots k. \]

5. The diagonal elements of the inverse of this matrix will be the asymptotic variances, and off-diagonal elements will be the asymptotic covariances of maximum likelihood estimators of $\theta_1 \theta_2 \ldots \theta_k$.

We shall follow this procedure to obtain the asymptotic variance of $b_M$. The likelihood function is:
\[ L(e_1, e_2, \ldots, e_n, d_1, d_2, \ldots, d_n) \]
\[ = \frac{1}{(\sigma_e \sigma_d 2\pi)^n} \exp\left\{ \frac{1}{2} \sum_{1}^{n} \left( \frac{e_1^2}{\sigma_e^2} + \frac{d_1^2}{\sigma_d^2} \right) \right\}, \]

and since \[ e_i = y_i - y_1, \]
and \[ d_i = x_i - x_1, \]

\[ L = \frac{1}{(\sigma_e \sigma_d 2\pi)^n} \exp\left\{ \frac{1}{2} \sum_{1}^{n} \left[ \frac{(y_1 - y_1)^2}{\sigma_e^2} + \frac{(x_1 - x_1)^2}{\sigma_d^2} \right] \right\} . \]

Using the relationship:
\[ \lambda = \sigma_d^2/\sigma_e^2, \]
\[ Y_1 = a + \beta X_1, \]

and taking the logarithm, we have

\[ \log L = -\frac{n}{2} \log \sigma_e^2 - \frac{n}{2} \log \lambda \sigma_e^2 - n \log 2\pi - \frac{1}{2} \sum_{1}^{n} \frac{(y_1 - a - \beta X_1)^2}{\sigma_e^2} - \frac{1}{2} \left[ \frac{\sum_{1}^{n} (x_1 - x_1)^2}{\sigma_e^2} + \frac{\sum_{1}^{n} (x_1 - x_1)^2}{\lambda \sigma_e^2} \right] . \]

The first derivatives have been derived in Graybill [4], and are as follows:

\[ \frac{\partial \log L}{\partial x_i} = \frac{(y_t - a - \beta X_1)\beta}{\sigma_e^2} + \frac{x_i - x_1}{\lambda \sigma_e^2}, \quad i = 1, 2, \ldots, n, \]
\[
\frac{\partial \log L}{\partial \sigma_e^2} = -\frac{n}{\sigma_e^2} + \frac{1}{2\sigma_e^4} \left[ \frac{n}{1} \Sigma(y_i - \alpha - \beta X_i)^2 + \frac{1}{\lambda} \right],
\]

\[
\frac{\partial \log L}{\partial \alpha} = \frac{n}{\sigma_e^2} \Sigma(y_i - \alpha - \beta X_i)X_i.
\]

Using these, we take second partials and their expected values. Reversing the sign and inserting into the information matrix denoted by \( A \):

\[
A = \begin{bmatrix}
    \frac{1+\lambda\beta^2}{\lambda} & 0 & \ldots & 0 & 0 \\
    0 & \frac{1+\lambda\beta^2}{\lambda} & \ddots & \vdots & \vdots \\
    \vdots & \ddots & \ddots & \ddots & \vdots \\
    0 & \ldots & \frac{1+\lambda\beta^2}{\lambda} & 0 & 0 \\
    0 & \ldots & 0 & \frac{n}{\sigma_e^2} & 0 & 0 & \sigma_e^2
\end{bmatrix}
\]

\[
= \begin{bmatrix}
    \frac{\beta X_1}{\sigma_e^2} & \frac{\beta X_2}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & 0 \\
    \frac{\beta X_2}{\sigma_e^2} & \frac{\beta X_3}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & 0 \\
    \ldots & \ldots & \ddots & \ddots & \ldots \\
    \frac{\beta X_n}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & \frac{n \Sigma X_i^2 / \sigma_e^2}{\lambda} & \frac{n X_i^2 / \sigma_e^2}{\lambda} \\
    0 & 0 & \ldots & \frac{n X_i^2 / \sigma_e^2}{\lambda} & \frac{n \Sigma X_i / \sigma_e^2}{\lambda}
\end{bmatrix}
\]

\[
\begin{bmatrix}
    \frac{\beta X_1}{\sigma_e^2} & \frac{\beta X_2}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & 0 \\
    \frac{\beta X_2}{\sigma_e^2} & \frac{\beta X_3}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & 0 \\
    \ldots & \ldots & \ddots & \ddots & \ldots \\
    \frac{\beta X_n}{\sigma_e^2} & \ldots & \frac{\beta X_n}{\sigma_e^2} & \frac{n \Sigma X_i^2 / \sigma_e^2}{\lambda} & \frac{n X_i^2 / \sigma_e^2}{\lambda} \\
    0 & 0 & \ldots & \frac{n X_i^2 / \sigma_e^2}{\lambda} & \frac{n \Sigma X_i / \sigma_e^2}{\lambda}
\end{bmatrix}
\]
We partition the matrix $A$:

\[
A = \begin{pmatrix}
A_{11} & A_{12} \\
A_{12} & A_{22}
\end{pmatrix},
\]

and denote the inverse of $A$ as

\[
A^{-1} = \begin{pmatrix}
A_{11} & A_{12} \\
(A_{12})' & A_{22}
\end{pmatrix}.
\]

Since we are only interested in the variance of $b_M$, which is the first diagonal element of $A^{22}$, we need to obtain only $A^{22}$. Using the relationship

\[
A^{22} = [A_{22} - A_{12} A_{11}^{-1} A_{12}]^{-1}
\]

we have

\[
[A^{22}]^{-1} = \frac{1}{\sigma^2} \begin{pmatrix}
\Sigma X_i^2 & \Sigma X_i \\
\Sigma X_i & n
\end{pmatrix} - \frac{\beta}{\sigma^2} \begin{pmatrix}
X_1 & \cdots & X_n & 0 \\
1 & \cdots & 1 & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
\frac{\lambda \sigma^2}{1 + \lambda \beta^2} & 0 & \cdots & 0 & 0 \\
0 & \ddots & \ddots & \ddots & \ddots \\
\ddots & \ddots & \frac{\lambda \sigma^2}{1 + \lambda \beta^2} & 0 & \cdots \\
0 & \cdots & 0 & \frac{\sigma^4}{n} & 0 \\
0 & \cdots & 0 & 0 & \frac{1}{n}
\end{pmatrix} \begin{pmatrix}
X_1 \\
\cdots \\
X_n
\end{pmatrix} = \begin{pmatrix}
\frac{\beta}{\sigma^2} \\
\cdots \\
\cdots
\end{pmatrix}
\]
\[
[A^{22}]^{-1} = \frac{1}{\sigma^2 \left( 1 + \lambda \beta^2 \right)} \begin{bmatrix}
\sum_{i=1}^{n} x_i^2 & \frac{n}{1} \\
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n}
\end{bmatrix}
\]

\[
A^{22} = \frac{\sigma^2 \left( 1 + \lambda \beta^2 \right)}{n \left( \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 / n \right)} \begin{bmatrix}
n & -\sum_{i=1}^{n} x_i \\
-\sum_{i=1}^{n} x_i & n \sum_{i=1}^{n} x_i^2
\end{bmatrix}
\]

Hence

\[
\text{Var} \left( b_M \right) = \frac{n \sigma^2 \left( 1 + \lambda \beta^2 \right)}{n \left( \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 / n \right)}
\]

\[
= \frac{\sigma^2 \left( 1 + \lambda \beta^2 \right)}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]  \hspace{1cm} (4.4.2.1)

4.4.3 TheRelative Efficiency of \(b_B\) and \(b_M\)

The efficiency of Group Averages estimator \(b_B\) compared to the maximum likelihood estimator \(b_M\) is:

\[
E\left( b_B \right) = \frac{\text{Var} \left( b_M \right)}{\text{Var} \left( b_B \right)}
\]

\[
= \frac{\sum_{i=1}^{n} \frac{\frac{n}{3} x_i - \frac{n}{3} \sum_{i=1}^{n} x_i}{2n+1}^2}{\frac{2n}{3} \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]}
\]  \hspace{1cm} (4.4.3.1)
If we assume that $X_i$ are equally spaced, and for the sake of simplicity, let $X_i$ be 1, 2, ..., $n$,

$$\frac{2n}{3} \sum_{1}^{n} (X_i - \bar{X})^2 = \frac{n^2}{18} (n^2 - 1)$$

$$\left(\frac{n}{2} \sum_{1}^{n+1} X_i - \frac{n}{2} \sum_{1}^{n} X_i\right)^2 = \left(\frac{n}{2} \sum_{1}^{n+1} i - \frac{n}{2} \sum_{1}^{n} i\right)^2$$

$$= \left[\frac{n}{18} (5n + 3) - \frac{n}{18} (n + 3)\right]^2$$

$$= \frac{4n^4}{81}.$$ 

Hence

$$E(b_B) = \frac{\text{Var} (b_M)}{\text{Var} (b_B)}$$

$$= \frac{\frac{4n^4}{81}}{\frac{n^2}{18} (n^2 - 1)}$$

$$= \frac{8n^2}{9(n^2 - 1)} \geq \frac{8}{9}.$$

Thus we conclude that the relative efficiency of Group Averages estimator compared with the maximum likelihood estimator is always greater than $\frac{8}{9}$. Efficiencies of $b_B$ when the $X_i$ are not equally spaced are not as straightforward, but estimates can be made based on formula (4.4.3.1).
4.5 A Numerical Example

Graybill [4] gave the following example in which he assumed that all values of $x_i$ and $y_i$ below are recorded under the assumptions of Section (4.1) and $\lambda = 1$.

Table 4
Data for Numerical Example

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>6.3</th>
<th>7.9</th>
<th>8.2</th>
<th>9.4</th>
<th>10.8</th>
<th>11.5</th>
<th>12.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>11.1</td>
<td>9.9</td>
<td>8.2</td>
<td>7.3</td>
<td>6.1</td>
<td>6.0</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Using the maximum likelihood estimators, we have

$$\sum_{1}^{7} (x_i - \overline{x})^2 = 34.337 , \quad \sum_{1}^{7} (y_i - \overline{y})^2 = 31.714 ,$$

$$\sum_{1}^{7} (y_i - \overline{y})(x_i - \overline{x}) = -32.311 , \quad n = 7 ,$$

$$\overline{x} = 7.543 , \quad \overline{y} = 9.571 , \quad u = 0.040589 ,$$

and hence

$$b_M = -0.960 , \quad a_M = 16.81 , \quad \hat{\sigma}^2 = 0.139 .$$

Using the method of Group Averages we get the estimates...
\[ b_B = \frac{\overline{y}_3 - \overline{y}_1}{\overline{x}_3 - \overline{x}_1} = \frac{7.1 - 12.2}{10.5 - 5.1} = -0.944 \]

\[ a_B = \overline{y} - b_B \overline{x} = 16.684 \]

\[ \hat{\sigma}^2 = (S_{y}^2 - b_B S_{xy}) \frac{n}{n-3} \]

\[ = [0.813 - (-0.944)(-0.711)] \frac{7}{4} \]

\[ = 0.252 \]

The estimates are quite comparable. However, the Group Averages method is much easier to compute.
V. CONCLUSION

The use of the Group Averages method for the case of $X_1$ equally spaced in linear regression and quadratic polynomial regression can readily be recommended, since it involves less computational work and has quite satisfactory efficiencies when compared to the Least Squares method. The method will be particularly useful to those who have no automatic calculators, and for those uses where answers are needed quickly.

More caution is needed to use the Group Averages method in the functional relationship situation, since the following condition must be satisfied:

$$\Pr \{ |d_1| > \frac{1}{2} c \} \text{ is negligible}$$

where $c = \text{Min.} |X_{i+1} - X_i|$, as this insures $b_B$ to be a consistent estimator of $\beta$. However, the efficiency of the Group Averages estimator of $\beta$ is quite satisfactory compared with the maximum likelihood estimator, at least when the $X_i$ are equally spaced. Another advantage to Group Averages method is that no prior knowledge is required about the relative magnitudes of $\sigma_e^2$ and $\sigma_d^2$, while such information is essential for the maximum likelihood method.
BIBLIOGRAPHY

1. Bartlett, M. S. "Fitting a straight line when both variables are subject to error" Biometrics, 5 (1949) 207.


7. Nair, K. R. and Banerjee, K. S. "Note on fitting of straight line if both variables are subject to error" Sankhya, 6 (1942) 331.


ACKNOWLEDGEMENTS

The author wishes to express his thanks to

; without his help this thesis would not have been
written.

The author also thanks for his valuable
suggestions which led to the improvement of this thesis.

The author is also grateful to

who prepared the final typewritten copies.
The vita has been removed from the scanned document.
ABSTRACT

This thesis reviews and discusses the so-called "Group Averages method" in the linear regression, the quadratic regression, and the functional relation situations.

In the linear and quadratic regression situations, under the assumption of $X_1$ equally spaced, the efficiency of the Group Averages estimator is quite satisfactory as compared with Least Squares estimators.

In the functional relation situation we used the Group Averages method and the Maximum Likelihood method for estimation of parameters. To compare their efficiencies we used the variance of the Group Averages estimator which was given by Dorff and Gurland [3], and developed the variance of Maximum Likelihood estimators. Under the assumption of $X_1$ equally spaced, we found the efficiency of the Group Averages estimator to be quite satisfactory. However, caution is needed for using the Group Averages method in functional relationships, since it requires the following condition to be satisfied:

$$\Pr \{ |d_1| \geq \frac{1}{2} c \} \text{ negligible}$$

where $c = \min. |X_{i+1} - X_i|$. 