RIA-76-U275
Cy No. 1

R-TR-76-011

THREE DEFINITIONS OF BEST LINEAR APPROXIMATION

by

JAMES J. HURT

1 APRIL 1976

PREPARED BY

RESEARCH DIRECTORATE

PREPARED FOR

RESEARCH DIRECTORATE
GENERAL THOMAS J. RODMAN LABORATORY
ROCK ISLAND ARSENAL
ROCK ISLAND, ILLINOIS 61201

DISCLAIMER

The findings of this report are not to be construed as an official department of the Army position, unless so designated by other authorized documents.

DISPOSITION INSTRUCTIONS

Destroy this report when no longer needed. Do not return to the originator.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: R-TR-76-011	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: Three Definitions of Best Linear Approximation	5. TYPE OF REPORT & PERIOD COVERED: Final
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): James J. Hurt	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Research Directorate, SARRI-LR-S GEN Thomas J. Rodman Laboratory Rock Island Arsenal, Rock Island, IL 61201	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 1T161101A91A
11. CONTROLLING OFFICE NAME AND ADDRESS: DAMA-ARZ-B Chief of Research and Development HQ, Department of the Army	12. REPORT DATE: 1 Apr 76
	13. NUMBER OF PAGES: 21
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Data Approximation
Linear Approximation
Least Squares Approximation
Statistics
Mean Values
Covariance Matrix
Correlation Coefficient

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Three definitions of best (in the least squares sense) linear approximation to given data points are presented. The relationships between these three are discussed along with their relationship to basic statistics such as mean values, the covariance matrix, and the (linear) correlation coefficient. For each of the three definitions, the best line is solved in closed form in terms of the data centroid and the covariance matrix.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

TABLE OF CONTENTS

 The following page is blank.

LIST OF FIGURES

The following page is blank.

LIST OF TABULAR DATA

The following page is blank.

1.0 INTRODUCTION

The problem frequently arises in practice where n data points, (x_1, y_1), (x_2,y_2),..., and (x_n,y_n), are given. These points can be plotted on an x-y graph giving a scattergram of the data. The problem is to find the line that gives a "best" approximation to the data, and to determine the "quality" of this approximation. In the following paragraphs, three definitions of a "best" approximation are given, the corresponding solutions for the best line are derived, and an expression for each "quality" is determined. In the process, the relationship between this data approximation problem and elementary statistics will also be revealed.

 The following page is blank.

2.0 NOTATION AND SOME PRELIMINARY RESULTS

Since the subject matter involves many sums from i = 1 to n, the following shorthand notation will be used:

$$\Sigma\, x_i = x_1 + x_2 + \ldots + x_n$$

No limits are indicated on the Σ summation sign since the limits i = 1 and i = n are understood.

The following elementary statistics about the data are needed:

$$\mu_x = \frac{1}{n}\,\Sigma\, x_i = \text{mean of } x_i \tag{2.1}$$

$$S_{xx} = \frac{1}{n}\,\Sigma\,(x_i - \mu_x)^2 = \text{variance of } x_i \tag{2.2}$$

$$\mu_y = \frac{1}{n}\,\Sigma\, y_i = \text{mean of } y_1 \tag{2.3}$$

$$S_{yy} = \frac{1}{n}\,\Sigma\,(y_i - \mu_y)^2 = \text{variance of } y_i \tag{2.4}$$

$$S_{xy} = \frac{1}{n}\,\Sigma\,(x_i - \mu_x)\,(y_i - \mu_y) = \text{covariance of } x_i \text{ and } y_i \tag{2.5}$$

Some elementary algebra results in the following useful identities.

$$\Sigma x_i = n\mu_x \tag{2.6}$$

$$\Sigma (x_i)^2 = nS_{xx} + n\mu_x^2 \tag{2.7}$$

$$\Sigma y_i = n\mu_y \tag{2.8}$$

$$\Sigma (y_i)^2 = nS_{yy} + n\mu_y^2 \tag{2.9}$$

$$\Sigma (x_i y_i) = nS_{xy} + n\mu_x\mu_y \tag{2.10}$$

These identities will be referred to repeatedly in the following sections. We shall assume that neither S_{xx} nor S_{yy} are zero (that both S_{xx} and S_{yy} are positive).

Finally, many quadratics will appear. Any quadratic Q(x) of the form (2.11).

$$Q(x) = A - 2Bx + Cx^2 \tag{2.11}$$

can be written as

$$Q(x) = (A - Cx_m^2) + C\,(x - x_m)^2$$

where

$$x_m = B/C \tag{2.12}$$

When $C > 0$, then the quadratic $Q(x)$ has a minimum at $x = x_m$. It's minimum value is

$$Q(x_m) = A - Cx_m^2 \tag{2.13}$$

Eqns (2.12) and (2.13) will be referred to for the minimizing argument and minimum value of any quadratic that appears in the following sections.

3.0 FIRST DEFINITION OF BEST LINEAR APPROXIMATION

When the x's are considered the independent variables, any line can be written in slope-intercept form as eqn (3.1)

$$y = mx + b \tag{3.1}$$

where m is the slope and b is the y-intercept. Each data point will, in general, not satisfy this equation but will leave residuals $r_1, r_2, \ldots, r_n$ where

$$r_i = y_i - mx_i - b, \quad i = 1,2,\ldots,n \tag{3.2}$$

The first definition of the best linear approximation is defined as that line where m and b minimize the sum of squared-residuals G:

$$G = \Sigma \ (r_i)^2 \tag{3.3}$$

The optimal line is determined in the following two steps. First, for any (fixed) slope m, the value of the y-intercept b which minimizes G is determined. To do this, substitute (3.2) into (3.3) then expand:

$$G = \Sigma \ (y_i - mx_i - b)^2$$

$$= \Sigma \ (y_i - mx_i)^2 - 2 \Sigma \ (y_i - mx_i) \ b + \Sigma \ b^2 \tag{3.4}$$

Now, equations (2.6) and (2.8) reduce eqn (3.4) to eqn (3.5).

$$G = \Sigma \ (y_i - mx_i)^2 - 2n(\mu_y - m\mu_x) \ b + nb^2 \tag{3.5}$$

This is a quadratic in b similar to eqn (2.11). It's minimum is given by eqn (3.6).

$$b_m = (n\mu_y - mn\mu_x) \ / \ n = \mu_y - m\mu_x \tag{3.6}$$

As an intermediate result, note that a horizontal line (constant function) is given by m = 0. In this event, the best value for b is μ_y and the resulting value for G is

$$G = \Sigma \ (y_i)^2 - n\mu_y{}^2 = nS_{yy}$$

These results are summarized by Theorem 1.

Theorem 1: The best constant approximation to the y_i data points is given by μ_y, the mean of the y_i. The "quality" of this approximation is given by the variance of y_i, $S_{yy} = G/n$.

The second step is to use eqn (3.6) for the y-intercept in the formula for G, eqn (3.4). Eqn. (2.13) gives the value for G as eqn (3.7).

$$G = \Sigma (y_i - mx_i)^2 - n (\mu_y - m\mu_x)^2 \tag{3.7}$$

Expanding this and using eqns (2.7), (2.9), and (2.10), eqn (3.7) reduces to eqn (3.8):

$$G = nS_{yy} - 2nS_{xy}m + nS_{xx}m^2 \tag{3.8}$$

This is a quadratic similar to eqn (2.11), hence its minimum occurs at

$$m_m = (nS_{xy}) / (nS_{xx}) = S_{xy} / S_{xx} \tag{3.9}$$

and its minimum value is:

$$G = nS_{yy} - nS_{xx}m_m^2$$

$$= nS_{yy} - nS_{xy}^2 / S_{xx} \tag{3.10}$$

Eqn (3.10) leads to the following definition of the "quality" of the approximation:

$$R = S_{xy}^2 / (S_{xx} S_{yy}) \tag{3.11}$$

With this value for R, eqn (3.10) becomes

$$G = nS_{yy} (1 - R) \tag{3.12}$$

Since both S_{xx} and S_{yy} are positive, R is positive. Since G is the sum of squared terms, it is positive and R is no greater than one. A value of zero for R says that a constant function is the best approximation to the y-data and a value of one for R, implying that $G^2 = 0$ and every residual $r_i = 0$, occurs when the line passes through every data point. All this is contained in:

Theorem 2: The best linear approximation to the y_i data points is given by the line (3.13)

$$y = \mu_y + S_{xy} (x - \mu_x) / S_{xx} \tag{3.13}$$

and the "quality" of this approximation is given by the quantity R in eqn. (3.11).

4.0 SECOND DEFINITION OF BEST LINEAR APPROXIMATION

The second definition of a "best" line is given when the y's are considered the independent variables and the x's are considered the dependent variables. This is simply an interchange of the roles of x and y from the previous section. The derivation of the equations follow in exactly the same order with x in place of y and y in place of x. The results are stated here:

Theorem 3: The best constant approximation to the x_i data points is given by μ_x, the mean of the x_i. The "quality" of this approximation is given by the variance of x_i, $S_{xx} = G^2/n$.

Theorem 4: The best linear approximation to the x_i data points is given by the line (4.1)

$$x = \mu_x + S_{xy} (y - \mu_y) / S_{yy} \tag{4.1}$$

and the "quality" of this approximation is given by the quantity R in eqn (4.2)

$$R = S_{xy}^2 / (S_{yy} S_{xx}) \tag{4.2}$$

4-1 The following page is blank.

5.0 GEOMETRIC BEST LINEAR APPROXIMATION

The third definition of a "best" line is given when neither the x's nor the y's are given preference. The x's were given preference in Section 3.0 and the y's were given preference in Section 4.0. In each case, the distance from a point to a line, the residual, was measured parallel to one of the axis. In this section, since neither axis can be so chosen, the geometric distance from a point to a line must be used.

The derivation of the best geometric line is much easier if vector notation is used. Let $P_1, P_2, \ldots, P_n$ be the data points

$$P_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad , \; i = 1,2,\ldots,n \tag{5.1}$$

For any vector V let V^T denote the transpose of V.

In vector notation, any line can be represented in parametric form as

$$L = A + Dt \tag{5.2}$$

Where A and D are vectors and t is the independent parameter. For any point P, the distance from this line to P, d(P,L), is the minimum distance from P to points on the line. For any point on the line, the square of this distance is d^2:

$$d^2 = (P - A - Dt)^T (P - A - Dt) \tag{5.3}$$

which expands to:

$$d^2 = (P - A)^T (P - A) - 2 D^T (P - A) t + D^T D t^2 \tag{5.4}$$

Eqn (5.4) is a quadratic similar to eqn (2.11), hence its minimum occurs at

$$t_m = D^T (P - A) / (D^T D) \tag{5.5}$$

and its minimum value is

$$\begin{aligned} d(P,L)^2 &= (P - A)^T (P - A) - D^T D t_m^2 \\ &= (P - A)^T (P - A) - \left[D^T (P - A) \right]^2 / (D^T D) \end{aligned} \tag{5.6}$$

Since all vectors are real, $D^T (P-A) = (P-A)^T D$ and equation (5.6) becomes

$$d(P,L)^2 = (P - A)^T (P - A) - ((P - A)^T D) (D^T (P - A)) / D^T D)$$

$$= (P - A)^T \left[I - DD^T / (D^T D) \right] (P - A) \qquad (5.7)$$

Where I is the identity matrix.

Given the data points $P_1, P_2, \ldots, P_n$, the "best" line L is that line which minimizes the sum of the distances - square, G:

$$G = \Sigma\, d(P_i, L)^2 \qquad (5.8)$$

Using eqn (5.7)

$$G = \Sigma\, (P_i - A)^T \left[I - DD^T / (D^T D) \right] (P_i - A) \qquad (5.9)$$

This "best" line is found in two steps. For the first step, fix the vector D and choose A so as to minimize eqn (5.9). For this purpose, let

$$M = I - DD^T / (D^T D) \qquad (5.10)$$

and

$$A = A_o + A_1 S \qquad (5.11)$$

for arbitrary vectors A_0 and A_1 with S a real number. With this form for A, eqn (5.9) expands to

$$G = \Sigma\, (P_i - A_o)^T M (P_i - A_o) - 2\, \Sigma A_1^T M (P_i - A_o)\, S + \Sigma A_1^T M A_1\, S^2 \qquad (5.12)$$

The vector A_0 minimizes G if the linear term in (5.12) is zero for every vector A_1, i.e., if

$$\Sigma M (P_i - A_o)$$

$$= M (\Sigma\, P_i - n A_o) = 0 \qquad (5.13)$$

While there are many vectors A_0 which satify (5.13), the obvious and easiest to use is the centroid:

$$A_o = \Sigma\, P_i / n = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad (5.14)$$

This result is summarized here.

Theorem 5: The best linear approximation to the data points passes through the centroid.

When the vector A is chosen as the centroid A_0, the value of G, eqn. (5.9), expands to become:

$$G = \Sigma\ (P_i - A_o)^T\ (P_i - A_o) - \Sigma\ (P_i - A_o)^T\ (D\ D^T)\ (P_i - A_o)\ /\ (D^T\ D) \quad (5.15)$$

The value for G is a minimum when the value for the second term in eqn (5.15) is a maximum, so attention will now focus on this second term, H.

$$H = \Sigma\ (P_i - A_o)^T\ (D\ D^T)\ (P_i - A_o)\ /(D^T\ D) \quad (5.16)$$

Since matrix multiplication is associative, H becomes

$$H = \Sigma\ \left[\ (P_i - A_o)^T\ D\right]\ \left[D^T\ (P_i - A_o)\right]\ /\ (D^T D) \quad (5.17)$$

Since $(P_i - A_o)^T\ D = D^T\ (P_i - A_o)$, H becomes

$$H = \Sigma \left[D^T\ (P_i - A_o)\right]\ \left[(P_i - A_o)^T\ D\right] / (D^T D) \quad (5.18)$$

Finally, using associativity once again

$$H = \Sigma\ D^T \left[(P_i - A_o)\ (P_i - A_o)^T\right] D\ /\ (D^T D)$$

$$= D^T \left[\ \Sigma\ (P_i - A_o)\ (P_i - A_o)^T\right]\ D\ /\ (D^T D) \quad (5.19)$$

A_0 is the centroid so

$$(P_i - A_o)\ (P_i - A_o)^T = \begin{bmatrix} x_i - \mu_x \\ y_i - \mu_y \end{bmatrix} \left[x_i - \mu_x,\ y_i - \mu_y\right]$$

$$= \begin{bmatrix} (x_i - \mu_x)^2 & (x_i - \mu_x)\ (y_i - \mu_y) \\ (y_i - \mu_y)\ (x_i - \mu_x) & (y_i - \mu_y)^2 \end{bmatrix} \quad (5.20)$$

and the matrix in eqn (5.19) is the covariance matrix:

$$\Sigma\ (P_i - A_o)\ (P_i - A_o)^T = \begin{bmatrix} nS_{xx} & nS_{xy} \\ nS_{xy} & nX_{yy} \end{bmatrix} \quad (5.21)$$

The quantity H in eqn (5.19) is the Rayleigh Quotient for the covariance matrix (5.21), which attains a maximum equal to the largest eigenvalue of (5.21) when the vector D is a corresponding eigenvector. The theory of eigenvalues/eigenvectors would have to be used if the vectors P_i had more

than two components. In the case of only two components, let

$$D = \begin{bmatrix} \cos u \\ \sin u \end{bmatrix}$$

for some angle u. Then $D^T D = \cos^2 u + \sin^2 u = 1$ and eqn (5.19) becomes

$$H = S_{xx} \cos^2 u + 2S_{xy} \cos u \sin u + S_{yy} \sin^2 u \qquad (5.23)$$

Eqn (5.23) can be simplified by using the trigonometric identities

$$\cos 2u = 2 \cos^2 u - 1 = 1 - 2 \sin^2 u \qquad (5.24)$$

and

$$\sin 2u = 2 \cos u \sin u$$

the simplified form for H becomes

$$H = S_{xx} \tfrac{1}{2} (1 + \cos 2u) + S_{sy} \sin 2u + S_{yy} \tfrac{1}{2} (1 - \cos 2u) \qquad (5.25)$$

$$= \tfrac{1}{2} (S_{xx} + S_{yy}) + \tfrac{1}{2} (S_{xx} - S_{yy}) \cos 2u + S_{xy} \sin 2u$$

Further simplification is achieved by defining r and u_o by eqns (5.26), (5.27), and (5.28).

$$r^2 = \tfrac{1}{4} (S_{xx} - S_{yy})^2 + S_{xy}^2 \qquad (5.26)$$

$$r \cos 2u_o = \tfrac{1}{2} (S_{xx} - S_{yy}) \qquad (5.27)$$

$$r \sin 2u_o = S_{xy} \qquad (5.28)$$

Figure 1 gives a geometric interpretation to these equations.

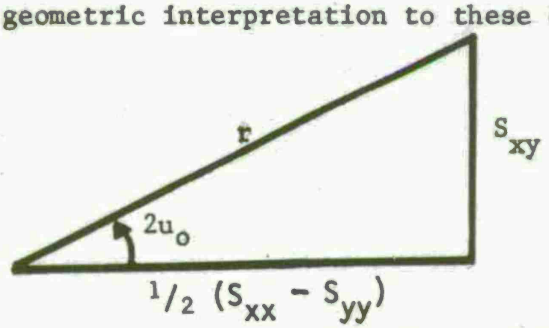

Figure 1 - Geometric definition of r and u_0

Using these values for R and u_o, H becomes

$$H = \tfrac{1}{2} (S_{xx} + S_{yy}) + r \cos (2u - 2u_o) \qquad (5.29)$$

This is maximized by choosing $u=u_o$. For this value of u, the value of H becomes

$$H = \tfrac{1}{2}(S_{xx} + S_{yy}) + r \tag{5.30}$$

Since $\Sigma (P_i - A_o)^T (P_i - A_o) = S_{xx} + S_{yy}$ the value of G, eqn. (5.15), becomes

$$G = \Sigma(P_i - A_o)^T (P_i - A_o) - \tfrac{1}{2}(S_{xx} + S_{yy}) - r$$

$$= \tfrac{1}{2}(S_{xx} + S_{yy}) - r \tag{5.31}$$

Since $r \geq 0$ and $G \geq 0$, equation (5.31) leads to an obvious definition for the "quality" of the approximation. These results are summarized in Theorem 6.

Theorem 6: The geometrically best linear approximation to the data points $P_1, P_2, \ldots P_n$ is given as

$$L = A + Dt$$

where the vector A is the centroid of the data points:

$$A = \Sigma P_i/n \tag{5.32}$$

and the vector D has components

$$D = \begin{bmatrix} \cos u_o \\ \sin u_o \end{bmatrix} \tag{5.33}$$

where u_o satisfies eqns (5.26), (5.27), and (5.28). The quality of this linear approximation is given by eqn (5.34) where r is defined by eqn (5.26).

$$R = \frac{2r}{(S_{xx} + S_{yy})} \tag{5.34}$$

The formula $L = A + Dt$ becomes

$$x = M_x + Ct \tag{5.35}$$

$$y = M_y + St$$

where $C = \cos(u_o)$ and $S = \sin(u_o)$ are defined by equations (5.26), (5.27), and (5.28). If the angle $2u_o$ is constrained between $-\pi$ and $+\pi$, then the formulas for C and S are

$$C = \sqrt{(1 + C_2)/2} \tag{5.36}$$

and

$$S = \text{sign}(S_{xy})\sqrt{(1 - C_2)/2} \tag{5.37}$$

where

$$C_2 = \cos(2u_o) = (S_{xx} - S_{yy})/(2r) \tag{5.38}$$

and

$$r = \sqrt{\tfrac{1}{4}(S_{xx} - S_{yy})^2 + S_{xy}^2} \tag{5.39}$$

6.0 REVIEW OF THE THREE DEFINITIONS

For the purposes of clarity, the main results are copied here. When the x-axis is treated as the independent variable, the best approximation is given by eqn (3.13)

$$y = \mu_y + S_{xy} (x - \mu_x) / S_{xx} \tag{3.13}$$

and the "quality" by eqn (3.11).

$$R = S_{xy}^2 / (S_{xx} S_{yy}) \tag{3.11}$$

When the y-axis is treated as the independent variable, the best approximation is given by eqn (4.1)

$$x = \mu_x + S_{xy} (y - \mu_y) / S_{yy} \tag{4.1}$$

and the "quality" by eqn (4.2)

$$R = S_{xy}^2 / (S_{xx} S_{yy}) \tag{4.2}$$

When neither axis is treated as independent, the best approximation is given by eqn (5.35)

$$\begin{aligned} x &= \mu_x + Ct \\ y &= \mu_y + St \end{aligned} \tag{5.35}$$

where C and S are defined by eqns (5.36), (5.37), (5.38), and (5.39). The "quality" of this approximation is given by eqn (5.34).

$$R = \frac{2r}{(S_{xx} + S_{yy})} \tag{5.34}$$

In all cases, the values for the "quality" R will be between zero and one. A value of R near one indicates that the linear approximation is very good in that the sum of squared residuals will be very nearly zero. A value of R near zero indicates that the linear approximation is very poor in that the sum of squared residuals will be large. The value for R is not dependent on the magnitudes of the data points but only on how well the data can be approximated by a line. Acceptable values for R are a function of the type of problem being solved and the required accuracy of the linear approximation.

The different approximations come about because of different meanings for the phrase "distance from a point to a line." Figure 2 shows the

three definitions used in this report. When the x-axis is treated as the independent variables, eqns (3.13) and (3.11), the error distances are measured parallel to the y-axis, e_1 in Figure 2. When the y-axis is treated as the independent variable, eqns (4.1) and (4.2), the error distances are measured parallel to the x-axis, e_2 in Figure 2. When neither axis is treated as the independent variable, eqns (5.35) and (5.34), the error distances are measured perpendicular to the line, e_3 in Figure 2. The choice between these three should be determined by the problem. If the data is such that one variable is clearly the independent variable and the other is dependent, then the appropriate formula, eqn (3.13) or eqn (4.1), should be used. The third method is useful when both variables are measured in the same units and there is no clear indication of which is the dependent variable.

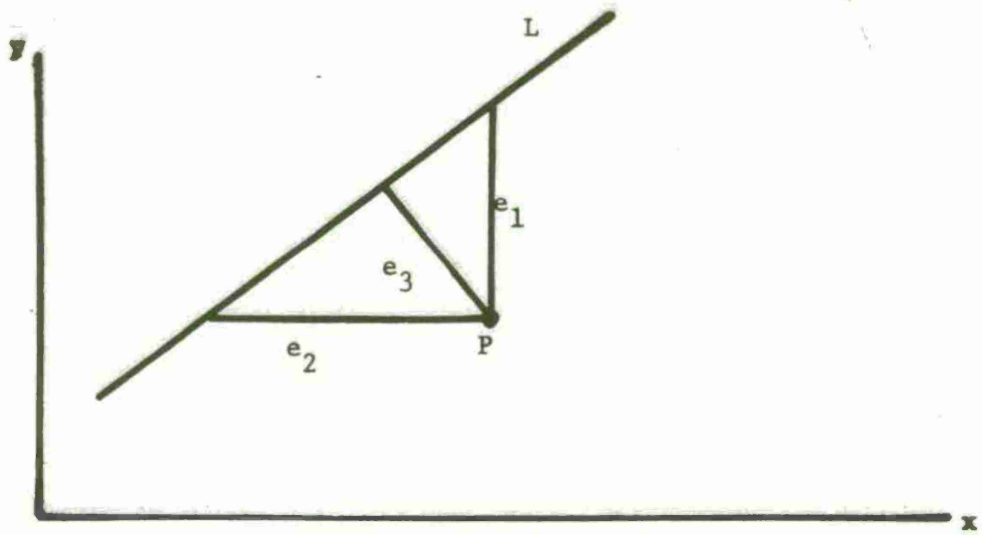

Figure 2 Three Distances From a Point to a Line

Comparing the three approximations is easier when the data has been translated and normalized such that $\mu_x = \mu_y = 0$ and $S_{xx} = S_{yy} = 1$ (subtract mean and divide by standard deviation). In this case, the co-variance S_{xy} is called the (linear) correlation coefficient. Eqns (3.13) and (3.11) reduce to $y = S_{xy}\, x$ and $R = S_{xy}^2$. The "quality" is the square of the correlation coefficient. Eqns (4.1) and (4.2) reduce to $x = S_{xy}\, y$ and $R = S_{xy}^2$. Eqns (5.35), (5.36), (5.37), (5.38), (5.39), and (5.34) quickly reduce to $y = \text{sign}(S_{xy})x$ and $R = |S_{xy}|$ if $S_{xy} \neq 0$. When $S_{xy} = 0$, there

is no best approximation and any line through the centroid will do as well as any other with a "quality" of R = 0. Note that, in this third case, the "quality" is the magnitude of the correlation coefficient, or the square root of the "quality" for the first two approximations. Indeed, whenever $S_{xx} = S_{yy}$, this third definition of "quality" will be the magnitude of the correlation coefficient. The first two definitions of "quality" are always the square of the correlation coefficient, so the three definitions of "quality" are best compared when the third definition has been squared. Note that the slopes for the three cases, S_{xy}, $1/S_{xy}$, and $\text{sign}(S_{xy})$, are equal only when $S_{xy} = 1$ or $S_{xy} = -1$. In all other cases, these three approximations lead to three different lines.

To illustrate the differences between these three linear approximations, two examples are discussed. The first example is a real application with the raw data shown in Table 1. The x-values are the lengths of subroutines (in binary bits) compiled on a CDC-6600 computer. The y-values are the lengths of these same subroutines (in binary bits) compiled on an IBM-370 computer. Both variables are measured in the same units (bits) and there is no clear indication which should be the dependent variable. This data, plotted in Figure 3, shows a strong linear relationship. The various statistics for this data are:

$$\mu_x = 7908.6 \qquad \mu_y = 8347.4$$

$$S_{xx} = 4.839 \times 10^7 \qquad S_{yy} = 3.816 \times 10^7$$

$$S_{xy} = 4.253 \times 10^7$$

When the x-axis is chosen as the independent variable, the best linear approximation is

$$y = 1395.3 + 0.879x$$

with a "quality" R = .980. When the y-axis is chosen as the independent variable, the best linear approximation is

$$x = -1395.0 + 1.115y$$

with a "quality" R = .980. Solving this approximation for y gives

$$y = 1251.7 + 0.897x$$

which is a different line than the first approximation.

When neither axis is chosen as the independent variable, the best

TABLE 1 RAW DATA FOR EXAMPLE 1

#	x	y
1	13260	13680
2	1920	3168
3	3240	4480
4	2760	3856
5	11820	11840
6	4140	4736
7	3960	4432
8	9900	7592
9	10320	8560
10	1440	2448
11	3420	4624
12	1800	3184
13	19140	18520
14	6540	7328
15	16680	17808
16	3780	4160
17	7080	7168
18	14280	13696
19	5520	3600
20	31920	27936
21	24120	23600
22	1800	2816
23	10080	10272
24	2160	3536
25	1800	3344
26	1920	4176
27	6300	6656
28	6720	6896
29	2160	3120
30	4020	4768
31	8880	9696
32	6240	8160
33	17640	16736
34	10440	10352
35	7080	8160
36	1380	2656
37	4200	6080
38	22380	21984
39	3480	5008
40	10980	12256
41	3840	4768
42	1620	2736

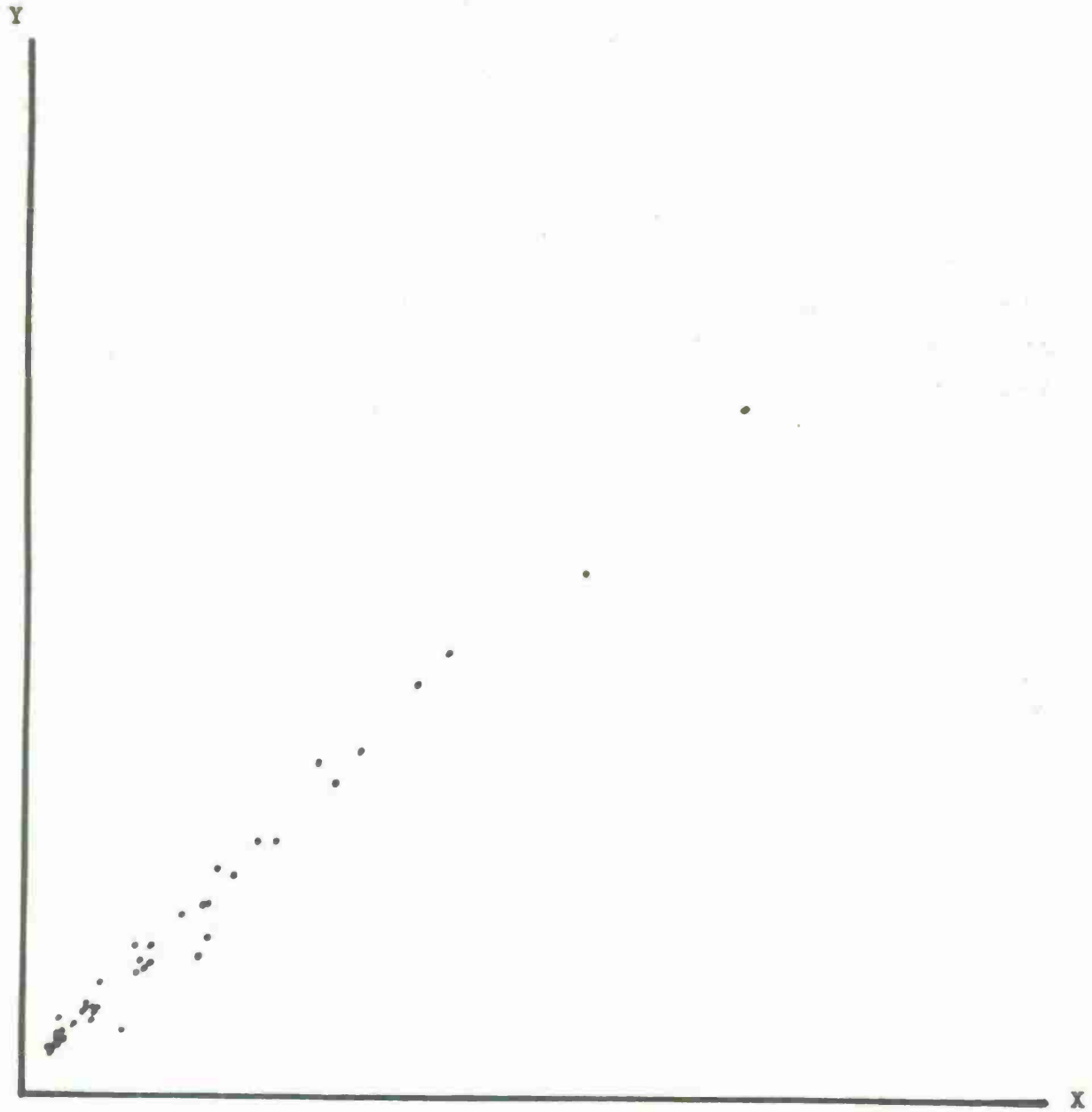

Figure 3 Plot of Raw Data for Example 1

linear approximation is

$$x = 7908.6 + 0.748t$$

$$y = 8347.4 + 0.664t$$

with a "quality" R=0.990 (R^2 = .980). This last approximation, when solved for y in terms of x, gives

$$y = 1332.4 + 0.887x$$

which is a third line. All three lines pass through the centroid (μ_x, μ_y) of the data and have three different slopes (0.879, 0.897, and 0.887). These slopes differ by very little and the three "qualities" are very high (.980, .980, and .980) indicating that the linear approximations are very good.

The second example is contrived to show how different the three approximations can be. The raw data is given in Table 2. The various statistics for this data are:

$$\mu_x = 0.5 \qquad \mu_y = 0.15$$

$$S_{xx} = 0.1 \qquad S_{yy} = 0.0078$$

and

$$S_{xy} = 0$$

When the x-axis is chosen as the independent variable, the best linear approximation is

$$y = 0.15$$

with a "quality" R = 0. This is a horizontal line!

When the y-axis is chosen as the independent variable, the best linear approximation is

$$x = 0.5$$

with a "quality" R = 0. This is a vertical line!

When neither axis is chosen as the independent variable, the best linear approximation is

$$x = 0.5 + t$$

$$y = 0.15$$

and the "quality" is R = .85 (R^2 = .72). This is the same horizontal line

as the first linear approximation!

TABLE 2 RAW DATA FOR EXAMPLE 2

#	x	y	#	x	y
1	.0	.00	7	.6	.24
2	.1	.09	8	.7	.21
3	.2	.16	9	.8	.16
4	.3	.21	10	.9	.09
5	.4	.24	11	1.0	.00
6	.5	.25			

 The following page is blank.

A.0 APPENDIX - PROGRAM LISTING

```
10 REM "COMPUTE BEST LINEAR APPROXIMATIONS"
20 REM "CODED IN HP-9830 BASIC BY JAMES HURT 31 MAR 76"
30 DIM S[6]
40 REM "ZERO SUMS"
50 MAT S=ZER
60 REM "ENTER DATA"
70 DISP S[1];
80 INPUT X,Y
90 REM "TEST FOR END OF DATA"
100 IF X<0 THEN 210
110 REM "UPDATE SUMS"
120 S[1]=S[1]+1
130 S[2]=S[2]+X
140 S[3]=S[3]+X*X
150 S[4]=S[4]+Y
160 S[5]=S[5]+Y*Y
170 S[6]=S[6]+X*Y
180 REM "PRINT DATA"
190 PRINT S[1];X;Y
200 GOTO 60
210 REM "COMPUTE STATISTICS"
220 S[2]=S[2]/S[1]
230 S[3]=S[3]/S[1]-S[2]*S[2]
240 S[4]=S[4]/S[1]
250 S[5]=S[5]/S[1]-S[4]*S[4]
260 S[6]=S[6]/S[1]-S[2]*S[4]
270 REM "PRINT STATISTICS"
280 PRINT
290 PRINT S[1];"POINTS"
300 PRINT "MEANS";S[2];S[4]
310 PRINT "VARIANCES";S[3];S[5];S[6]
320 REM "COMPUTE STANDARD DEVIATIONS"
330 X=SQR(S[3])
340 Y=SQR(S[5])
350 REM "CORRELATION COEFFICIENT"
360 R=S[6]/(X*Y)
370 PRINT "STD. DEV.";X;Y;R
380 REM "X IS INDEPENDENT"
390 X=S[6]/S[3]
400 Y=S[4]-X*S[2]
410 PRINT "Y=";Y;"+X*";X
420 REM "Y IS INDEPENDENT"
430 X=S[6]/S[5]
440 Y=S[2]-X*S[4]
450 PRINT "X=";Y;"+Y*";X
460 REM "QUALITY"
470 R=S[6]*S[6]/(S[3]*S[5])
480 PRINT "QUALITY=";R
```

Program Listing (Cont'd)

```
490 REM "C2=COS(2*UO)"
500 X=(S[3]-S[5])/2
510 R=SQR(X*X+S[6]*S[6])
520 C2=X/R
530 REM "X=COS(UO) AND Y=SIN(UO)"
540 X=SQR((1+C2)/2)
550 Y=SQR((1-C2)/2)
560 IF S[6] >= 0 THEN 580
570 Y=-Y
580 PRINT
590 PRINT "X=";S[2];"+T*";X
600 PRINT "Y=";S[4];"+T*";Y
610 REM "QUALITY"
620 R=2*R/(S[3]+S[5])
630 PRINT "QUALITY=";R
640 PRINT
650 PRINT
660 PRINT
670 END
```

DISTRIBUTION LIST

A. Dept. of the Defense

Defense Documentation Center
ATTN: TIPCR (12)
Cameron Station
Alexandria, VA 22314

B. Dept. of the Army

Commander
Rock Island Arsenal
ATTN: SARRI-L (1)
SARRI-LPC (2)
SARRI-LR (5)
SARRI-LR-S (Hurt) (10)
Rock Island, IL 61201

 The following page is blank.

DISTRIBUTION LIST UPDATE

- - - FOR YOUR CONVENIENCE - - -

Government regulations require the maintenance of up-to-date distribution lists for technical reports. This form is provided for your convenience to indicate necessary changes or corrections.

If a change in our mailing lists should be made, please check the appropriate boxes below. For changes or corrections, show old address *exactly* as it appeared on the mailing label. Fold on dotted lines, tape or staple the lower edge together, and mail.

☐ Remove Name From List

☐ Change or Correct Address

Old Address:	Corrected or New Address:

COMMENTS

_____

_____

_____

_____

_____

Date: _____ Signature: _____

Technical Report # R-TR-76-011

SARRI Form 900-643 (One-Time) (1 Feb 75)

FOLD HERE

Return Address:

_____

_____

_____

OFFICIAL BUSINESS
Penalty for Private Use $300

POSTAGE AND FEES PAID
DEPARTMENT OF THE ARMY
DOD 314

Commander
Rock Island Arsenal
Attn: SARRI-LR
Rock Island, Illinois 61201

FOLD HERE

AD_____ACCESSION NO._____

Research Directorate, General Thomas J. Rodman Laboratory
Rock Island Arsenal, Rock Island, Illinois 61201

Three Definitions of Best Linear Approximation

Prepared by: James J. Hurt
Security Class. (of this report) Unclassified
Technical Report R-TR-76-011

21 Pages, Incl Figures

Three definitions of best (in the least squares sense) linear approximation to given data points are presented. The relationships between these three are discussed along with their relationship to basic statistics such as mean values, the covariance matrix, and the (linear) correlation coefficient. For each of the three definitions, the best line is solved in closed form in terms of the data centroid and the covariance matrix.

UNCLASSIFIED

1. Data Approximation
2. Linear Approximation
3. Least Squares Approximation
4. Statistics
5. Mean Values
6. Covariance Matrix
7. Correlation Coefficient

I. James J. Hurt
II. Rock Island Arsenal
III. Research Directorate
General Thomas J. Rodman Laboratory
Rock Island Arsenal

DISTRIBUTION

Approved for public release; distribution unlimited.

AD_____ACCESSION NO._____

Research Directorate, General Thomas J. Rodman Laboratory
Rock Island Arsenal, Rock Island, Illinois 61201

Three Definitions of Best Linear Approximation

Prepared by: James J. Hurt
Security Class. (of this report) Unclassified
Technical Report R-TR-76-011

21 Pages, Incl Figures

Three definitions of best (in the least squares sense) linear approximation to given data points are presented. The relationships between these three are discussed along with their relationship to basic statistics such as mean values, the covariance matrix, and the (linear) correlation coefficient. For each of the three definitions, the best line is solved in closed form in terms of the data centroid and the covariance matrix.

UNCLASSIFIED

1. Data Approximation
2. Linear Approximation
3. Least Squares Approximation
4. Statistics
5. Mean Values
6. Covariance Matrix
7. Correlation Coefficient

I. James J. Hurt
II. Rock Island Arsenal
III. Research Directorate
General Thomas J. Rodman Laboratory
Rock Island Arsenal

DISTRIBUTION

Approved for public release; distribution unlimited.

DEPARTMENT OF THE ARMY
ROCK ISLAND ARSENAL
ROCK ISLAND, ILLINOIS 61201

OFFICIAL BUSINESS

PENALTY FOR PRIVATE USE, $300.

SARRI-LR

POSTAGE AND FEES PAID
DEPARTMENT OF THE ARMY

DoD 314

SPECIAL FOURTH-CLASS RATE
BOOK

POSTMASTER: If Undeliverable (Section 159.41 Postal Manual) Do Not Return