AD 659 977

USNRDL-TR-67-99
11 July 1967

REGRESSION PARAMETERS FOR PAIRS OF EQUALLY UNCERTAIN VARIABLES

by

E. C. Freiling
G. R. Crocker

U.S. NAVAL RADIOLOGICAL DEFENSE LABORATORY

SAN FRANCISCO · CALIFORNIA · 94135

This document has been approved for public release; distribution is unlimited.

PHYSICAL CHEMISTRY BRANCH
E. C. Freiling, Head

NUCLEAR TECHNOLOGY DIVISION
R. Cole, Head

ADMINISTRATIVE INFORMATION

The work reported is part of a project sponsored by the Division of Biology and Medicine of the Atomic Energy Commission under Contract AT (49-7)-1963.

DDC AVAILABILITY NOTICE

This document has been approved for public release and sale; its distribution is unlimited.

ACCESSION for
CFSTI WHITE SECTION
DDC BUFF SECTION
UNANNOUNCED
JUSTIFICATION .....
.....
BY .....
DISTRIBUTION/AVAILABILITY
DIST. | AVAIL.
1

Eugene P. Cooper
Eugene P. Cooper
Technical Director

D. C. Campbell, CAPT USN
Commanding Officer and Director

ABSTRACT

The correlation of radiochemical data from samples of fractionated nuclear debris involves the treatment of two variables whose uncertainties are comparable. We considered three new criteria for the establishment of regression parameters for such correlations (least square perpendicular distances between points and the line, bisection of the angle formed by the certain-x and certain-y regression lines, and adoption of the geometric mean of the certain-x and certain-y regression slopes). We concluded that the geometric-mean slope $\hat{b}$ was most satisfactory. It is related to the usual certain-x regression slope $b_{x,x}$ and the coefficient of correlation r by the simple expression

$$\hat{b} = b_{x,x}/|r|$$

SUMMARY

Problem

The correlation of radiochemical data from fractionated debris does not meet the usual requirements for the application of least-squares analysis; namely, that one variable be known with much greater certainty than the other. Occasionally the mechanical application of the usual least-squares treatment produces results which appear to be specious.

Findings

Alternative treatments were developed and investigated. These are based upon criteria which are more appropriate to the situation at hand and also give results which are more reasonable: least square perpendicular distances between points and the line, bisection of the angle formed by the certain-x and certain-y regression lines, and adoption of the geometric mean of the certain-x and certain-y regression slopes. Of these, the geometric-mean slope was found to be the most satisfactory.

INTRODUCTION

In applying standard least-squares methods to the statistical analysis of relations between two variables one assumes the independent variable to be much better known than the dependent variable and then proceeds to determine a regression line by minimizing the squared deviations of the latter variable. There arise situations where this assumption is not at all fulfilled. A good example is the correlation of radiochemical data from fractionated nuclear debris, where dependent and independent variables are of nearly equal uncertainty. Here the regression slopes can be heavily influenced by uncertain data lying near the population extremities. Situations frequently arise where the calculated line differs significantly from what the eye would select, leaving the viewer with an uncomfortable feeling about the reliability of the correlation parameters. An example will be presented in a later section.

Several obvious solutions occur to this state of affairs. One is to minimize the squares of the perpendicular distances from the regression line instead of those of the vertical distances. Another is to use the geometric mean of the slopes of the lines for y on x and for x on y. Still another is to bisect the angle formed by these lines.

The purpose of this report is to develop, test, and evaluate these methods with a view to applying the results to the correlation of radiochemical data from fractionated nuclear debris.

NOTATION

The notation below refers to quantities taken from standard statistical development. Additional notation will be introduced in the text.

y_i = dependent variable

x_i = independent variable

n = number of data points

$\langle u_i \rangle$ = mean value of $u = \frac{1}{n} \Sigma u_i$

$S(u,v) = n(\langle u_i v_i \rangle - \langle u_i \rangle \langle v_i \rangle) = S(v,u)$

r = coefficient of correlation $= S(x,y)/\sqrt{S(x,x)S(y,y)}$

$\Theta_{u,v}$ = angle made with v axis by regression line obtained by assuming certainty in the u values (cf Figs. 1 and 2)

$b_{u,v} = \tan \Theta_{u,v}$

$a_{u,v}$ = v intercept made by regression line obtained by assuming certainty in the u values.

Figures 1 and 2 illustrate the notation and some relations between the quantities listed. A more complete set of relations is given in Table 1. Since these relations are either standard (see, for example, Ref. 1) or immediately derivable (from Figs. 1 and 2), their derivations are not belabored.

DERIVATION OF EQUATIONS

In this section equations will be derived in their most concise form. An investigator desiring to apply the equations to work completed or in progress will find these forms inconvenient. Therefore a later section will summarize the equations in practical form, i.e., in terms of the parameters most likely to be available. Specifically, these are $a_{x,y}$, $b_{x,x}$, r, and $\langle x_i \rangle$.

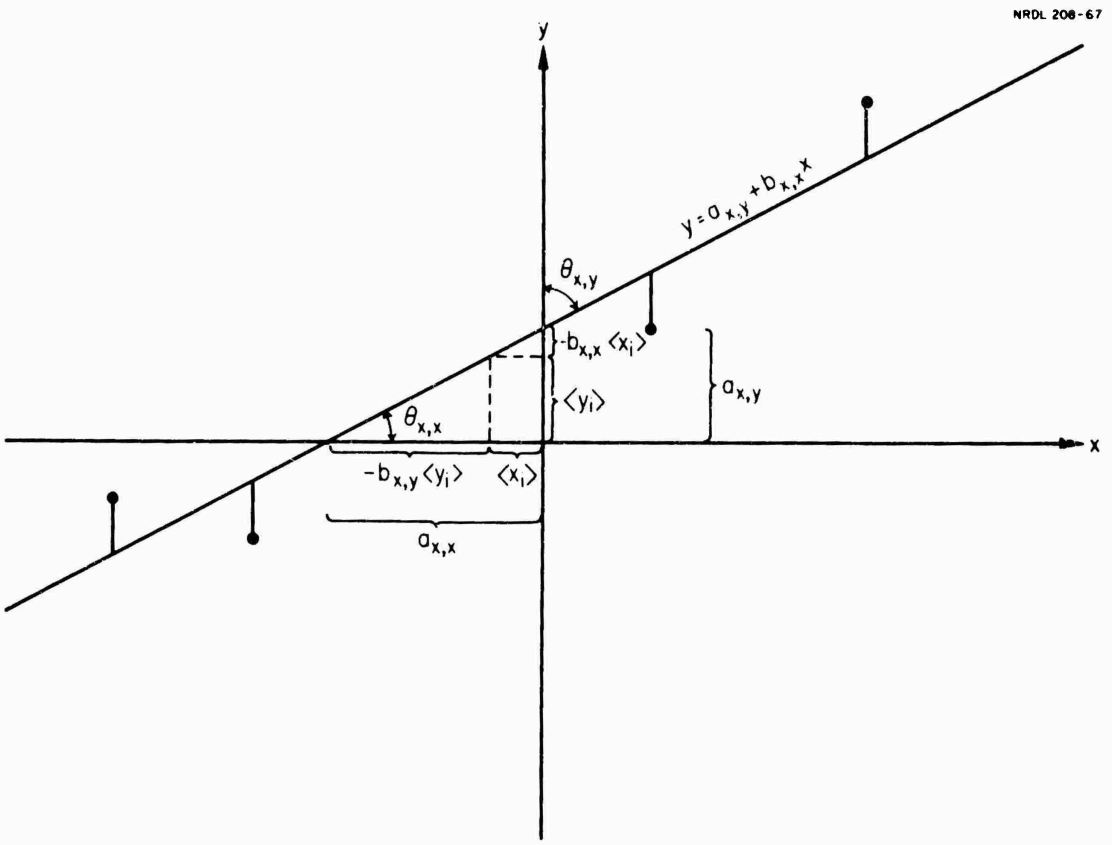

Fig. 1 Illustrated Quantities and Relations for a Regression Line Obtained by Minimizing Deviations in the y-direction.

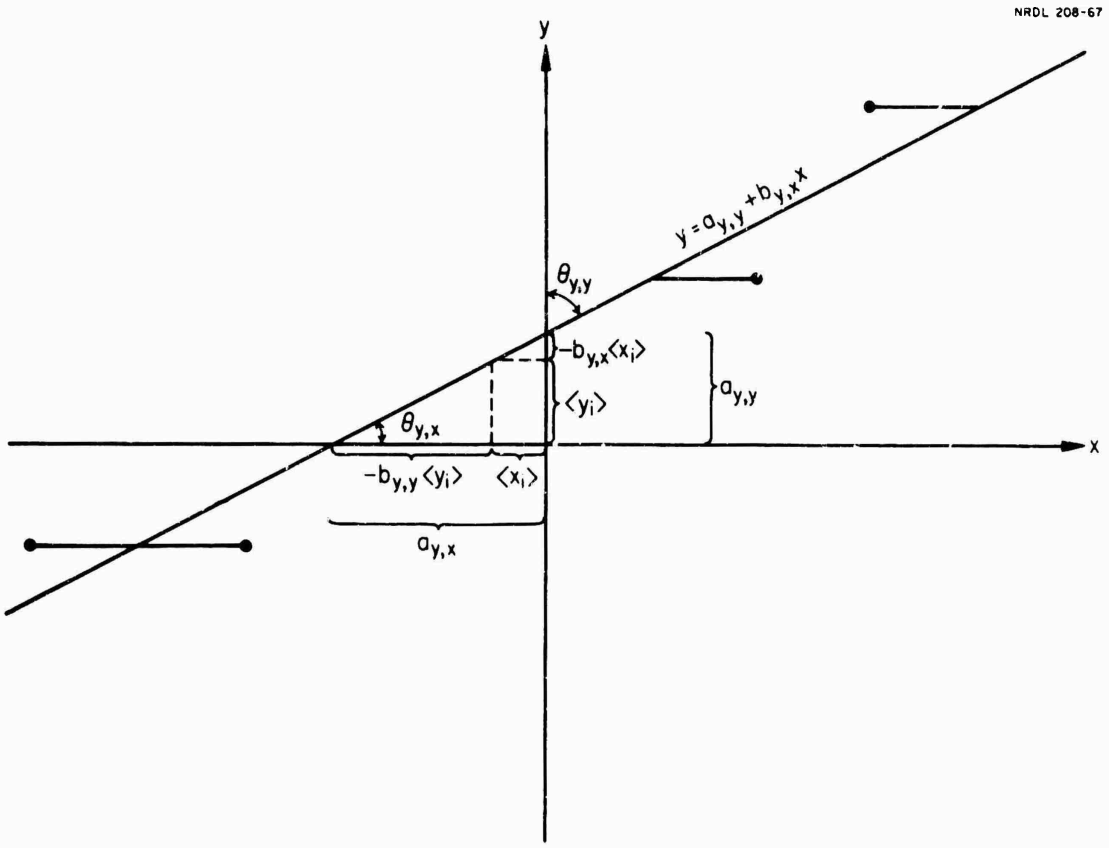

Fig. 2 Illustrated Quantities and Relations for a Regression Line Obtained by Minimizing Deviations in the x-direction.

TABLE 1

Summary of Basic Relations

Quantity	Expressed in Terms of a	Expressed in Terms of b	Expressed in Terms of S
$b_{x,x}$	$-\frac{a_{x,y}}{a_{x,x}}$	$\frac{1}{b_{x,y}}$	$\frac{S(x,y)}{S(x,x)}$
$b_{y,y}$	$-\frac{a_{y,x}}{a_{y,y}}$	$\frac{1}{b_{y,x}}$	$\frac{S(x,y)}{S(y,y)}$
r^2	$\frac{a_{x,y}\, a_{y,x}}{a_{x,x}\, a_{y,y}}$	$b_{x,x}\, b_{y,y}$	$\frac{[S(x,y)]^2}{S(x,x)S(y,y)}$
$a_{x,x}$	$-a_{x,y}/b_{x,x}$	$\langle x_i \rangle - b_{x,y} \langle y_i \rangle$	$\langle x_i \rangle - \frac{S(x,x)}{S(x,y)} \langle y_i \rangle$
$a_{y,y}$	$-a_{y,x}/b_{y,y}$	$\langle y_i \rangle - b_{y,x} \langle x_i \rangle$	$\langle y_i \rangle - \frac{S(y,y)}{S(x,y)} \langle x_i \rangle$

Geometric Mean Slope

The geometric mean slope $\hat{b}$ is given simply by

$$\hat{b}^2 = b_{x,x}\, b_{y,x}$$

$$= S(y,y)/S(x,x).$$

To complete the definition of this line, it is reasonable to impose the condition that it pass through the point $P_o(x_o,y_o)$ formed by the intersection of the two regression lines

$$y = a_{x,y} + b_{x,x}\, x$$

and

$$y = a_{y,y} + b_{y,x}\, x.$$

The coordinates of this point are

$$x_o = \frac{a_{y,y} - a_{x,y}}{b_{x,x} - b_{y,x}}$$

and

$$y_o = \frac{b_{x,x}\, a_{y,y} - b_{y,x}\, a_{x,y}}{b_{x,x} - b_{y,x}}$$

Hence, the intercept with the y-axis of the line with slope $\hat{b}$ and passing through P_o is

$$\hat{a}_y = y_o - \hat{b}x_o$$

$$= \frac{a_{y,y}\,(b_{x,x} - \hat{b}) - a_{x,y}\,(b_{y,x} - \hat{b})}{b_{x,x} - b_{y,x}}$$

and that with the x-axis is

$$\hat{a}_x = x_o - \frac{1}{\hat{b}}\, y_o$$

$$= -\hat{a}_y/\hat{b}$$

The Bisector of the Angle $\theta_{x,x} - \theta_{y,x}$

The line bisecting the angle between $\theta_{x,x}$ and $\theta_{y,x}$ will form the angle

$$\tilde{\theta} = \frac{1}{2}(\theta_{x,x} + \theta_{y,x})$$

with the x-axis. Hence we can write

$$\tan 2\tilde{\theta} = \tan (\theta_{x,x} + \theta_{y,x})$$

and by familiar trigonometric relations obtain an expression for the slope $\tilde{b}$ of the bisector:

$$\frac{2\tilde{b}}{1-\tilde{b}^2} = \frac{b_{x,x} + b_{y,x}}{1 - b_{x,x}\, b_{y,x}}$$

from which

$$\tilde{b} = \frac{b_{x,x}\, b_{y,x} - 1}{b_{x,x} + b_{y,x}} + \sqrt{\left(\frac{b_{x,x}\, b_{y,x} - 1}{b_{x,x} + b_{y,x}}\right)^2 + 1}$$

where the plus sign is chosen to make $\tilde{b}$ approach $b_{x,x}$ and $b_{y,x}$ when these latter two quantities approach each other.

As in the case of the geometric mean slope, the line will be made to pass through P_o. Proceeding as in that case:

$$\tilde{a}_y = y_o - \tilde{b}x_o$$

and

$$\tilde{a}_x = x_o - \frac{1}{\tilde{b}}\, y_o = -\tilde{a}_y/\tilde{b}$$

For the former intercept

$$\tilde{a}_y = \frac{a_{y,y}(b_{x,x}^2+1) - a_{x,y}(b_{y,x}^2+1) - (a_{y,y}-a_{x,y})\sqrt{(b_{x,x}^2+1)(b_{y,x}^2+1)}}{b_{x,x}^2 - b_{y,x}^2}$$

Least Squared Perpendicular Deviations

The square of the distance d_i of point $P_i(x_i,y_i)$ from a line $y = \overset{\upsilon}{a}_y + \overset{\upsilon}{b}_x$ is known from analytic geometry to be

$$d_i^2 = (y_i - \overset{\upsilon}{a}_y - \overset{\upsilon}{b}x_i)^2/(\overset{\upsilon}{b}^2 + 1).$$

Summing over i, differentiating partially with respect to $\overset{\upsilon}{a}_y$, and setting the result equal to zero give

$$\langle y_i \rangle - \overset{\upsilon}{a}_y - \overset{\upsilon}{b} \langle x_i \rangle = 0.$$

Carrying out a similar treatment with respect to $\overset{\upsilon}{b}$ gives

$$\overset{\upsilon}{b}(\langle y_i^2 \rangle - \langle x_i^2 \rangle - 2\overset{\upsilon}{a}_y \langle y_i \rangle + \overset{\upsilon}{a}_y^2) - (\overset{\upsilon}{b}^2-1)(\langle x_i y_i \rangle - \overset{\upsilon}{a}_y \langle x_i \rangle) = 0$$

Substitution for $\overset{\upsilon}{a}_y$ and solution for $\overset{\upsilon}{b}$ gives

$$\overset{\upsilon}{b} = \frac{S(y,y) - S(x,x)}{2S(x,y)} \pm \sqrt{\left(\frac{S(y,y) - S(x,x)}{2S(x,y)}\right)^2 + 1}$$

or

$$\overset{\upsilon}{b} = \frac{1}{2}(b_{y,x} - b_{x,y}) + \sqrt{\frac{1}{4}(b_{y,x} - b_{x,y})^2 + 1}$$

where the plus sign is chosen to make $\overset{\upsilon}{b}$ approach $b_{y,x}$ and $b_{x,y}$ when these latter two quantities approach each other.

The intercept with the y-axis is obtained by eliminating $\langle y_i \rangle$ from the equation

$$\overset{\upsilon}{a}_y = \langle y_i \rangle - \overset{\upsilon}{b} \langle x_i \rangle$$

and the equation

$$a_{x,y} = \langle y_i \rangle - b_{x,x} \langle x_i \rangle$$

to get

$$\overset{U}{a}_y = a_{x,y} + (b_{x,x} - \overset{U}{b}) \langle x_i \rangle$$

Analogously to the previous cases,

$$\overset{U}{a}_x = \langle x_i \rangle - \frac{1}{\overset{U}{b}} \langle y_i \rangle$$

Summary

The equations in this section have been derived in manners chosen for directness and have not always appeared in the most desirable form. To remedy this, Table 2 summarizes the results of this section in a way which illustrates the similarity among the chosen methods and the circumstances under which the parameters will converge. It is convenient at this point to introduce the quantity B, defined by either of the equations

$$b^2 - 2Bb - 1 = 0$$

or

$$b = B + \sqrt{B^2 + 1}$$

although we will not have use for it until we discuss the application of the equations.

For conversion of the equations to practical form, it is helpful to first convert the ingredients to practical form, and this is done in Table 3 by manipulation of relationships in Table 1. Application of Table 3 to previously developed equations gives Table 4.

APPLICATION

Consideration of some of the properties of the quantities we have discussed will provide helpful orientation.

TABLE 2

Summary of Equations in Symmetrical Form

Quantity	Treatment		
	Geometric Mean Slope	Least-Square Perpendicular Deviations	Slope of Bisecting Line
a_y	$y_o - \hat{b}x_o$	$\langle y_i \rangle - \overset{\upsilon}{b} \langle x_i \rangle$	$y_o - \tilde{b}x_o$
a_x	$x_o - \frac{1}{\hat{b}} y_o$	$\langle x_i \rangle - \frac{1}{\overset{\upsilon}{b}} \langle y_i \rangle$	$x_o - \frac{1}{\tilde{b}} y_o$
B^a	$\frac{b_{y,x} - b_{x,y}}{2\sqrt{b_{y,x}\, b_{x,y}}}$	$\frac{b_{y,x} - b_{x,y}}{2}$	$\frac{b_{y,x} - b_{x,y}}{1 + b_{y,x}\, b_{x,y}}$

a. Defined by either of the equations $b^2 - 2Bb - 1 = 0$ or $b = B + \sqrt{B^2 + 1}$.

TABLE 3

Summary of Equations for Conversion to Practical Form

Quantity	Equation	
$a_{x,x}$	$-a_{x,y}/b_{x,x}$	
$a_{y,x}$	$(1-r^2)\ \langle x_i \rangle - r^2 \frac{a_{x,y}}{b_{x,x}}$	(a)
$a_{y,y}$	$a_{x,y} - \frac{1-r^2}{r^2} b_{x,x}\ \langle x_i \rangle$	(a)
$b_{x,y}$	$\frac{1}{b_{x,x}}$	
$b_{y,x}$	$\frac{b_{x,x}}{r^2}$	
$b_{y,y}$	$\frac{r^2}{b_{x,x}}$	
$\langle y_i \rangle$	$a_{x,y} + b_{x,x}\ \langle x_i \rangle$	

(a) Ezekiel[2] calls r^2 the coefficient of determination and $1-r^2$ the coefficient of non-determination. He designates the latter by k^2. He calls k the coefficient of alienation.

TABLE 4

Summary of Equations in Practical Form

Treatment	$\frac{a_y - a_{x,y}}{\langle x_i \rangle}$	B^a	b
Geometric Mean ($\wedge$)	$\frac{r-1}{r} b_{x,x}$	$\frac{b_{x,x}^2 - r^2}{2\lvert r \rvert b_{x,x}}$	$\frac{b_{x,x}}{\lvert r \rvert}$
Least-Square Perpendicular Deviations ($\cup$)	$\frac{2r^2 b_{x,x}^2 - b_{x,x}^2 + r^2 - \sqrt{(b_{x,x}^2 - r^2)^2 + 4r^4 b_{x,x}^2}}{2r^2 b_{x,x}}$	$\frac{b_{x,x}^2 - r^2}{2r^2 b_{x,x}}$	-
Bisector ($\sim$)	$\frac{r^2(b_{x,x}^2 + 1) - \sqrt{(b_{x,x}^2 + 1)(b_{x,x}^2 + r^4)}}{(r^2 + 1) b_{x,x}}$	$\frac{b_{x,x}^2 - r^2}{(r^2 + 1) b_{x,x}}$	-

a. Defined by either of the equations $b^2 - 2Bb - 1 = 0$ or $b = B + \sqrt{B^2 + 1}$.

We first note that while S(u,v) may be either positive or negative, S(u,u) must always be positive. Reference to Table 1 shows that S(x,y), $b_{x,x}$, $b_{y,y}$, $b_{x,y}$, $b_{y,x}$ and r will therefore all have the same sign. It is also obvious from Table 1 that the value of r will lie between those of $b_{x,x}$ and $b_{y,y}$, while the value of 1/r will lie between those of $b_{x,y}$ and $b_{y,x}$.

Table 2 shows that the sign of the quantity B is the same in all three treatments and governed by $b_{y,x} - b_{x,y}$. From its definition, each value of B is seen to change sign as the value of b^2 goes through 1. However, most correlations of fractionation data give values in the range of $0 \leq b \leq 1$, so that B will lie primarily in the range $B \leq 0$.

Reference to Table 4 shows that, for $r^2 = 1$ (perfect correlation), all B's are equal. Since for positive correlation $\partial B/\partial r$ at $r^2 = 1$ is the same for both the geometric-mean and angle-bisection treatments, these will have similar values for good positive correlations.

In general

$$\hat{B} : \overset{\cup}{B} : \tilde{B} = |r| : 1 : 2r^2/(r^2 + 1)$$

so that $|\overset{\cup}{B}| \geq |\hat{B}| \geq |\tilde{B}|$. Now, the relation between b and B is complicated, but for the range of interest it can be visualized geometrically as shown by Fig. 3. From this figure it is apparent that as -B increases, b decreases, and therefore:

$$\tilde{b} \geq \hat{b} \geq \overset{\cup}{b} \qquad (0 \leq \overset{\cup}{b}, \hat{b}, \tilde{b} \leq 1)$$

Figure 4 shows the application of these methods to the correlation of Te^{132} data from Shot Sedan.[3] Two outliers are evident among the data. These were included in all the calculations except one, and that one is indicated on the graph. The slopes for certain-x ($b_{x,x}$) and certain-y ($b_{y,x}$) are seen to be extreme. The slope for angle bisection ($\tilde{b} = 0.654$)

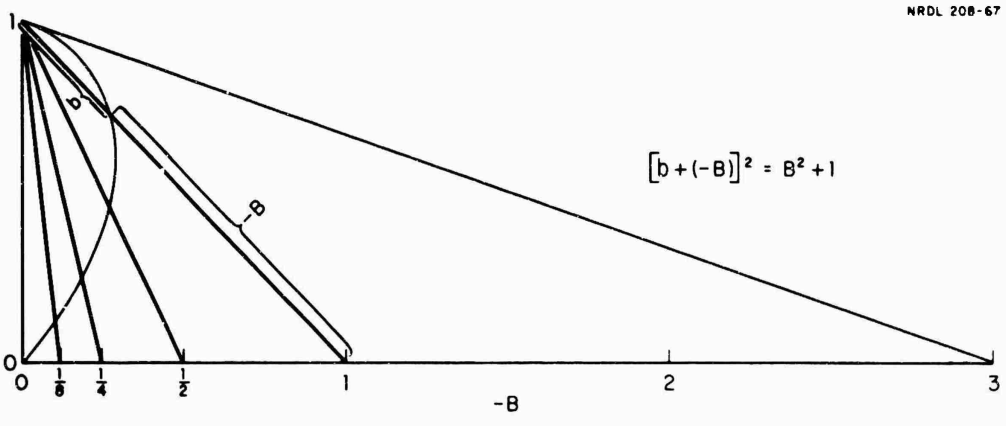

Fig. 3 Geometrical Relation Between b and -B. The members of the family of right triangles all have unit height and a base equal to -B. If a distance -B is laid off along a hypotenuse, the length of the remainder of the hypotenuse corresponds to b. The locus of these points is shown by the curved line.

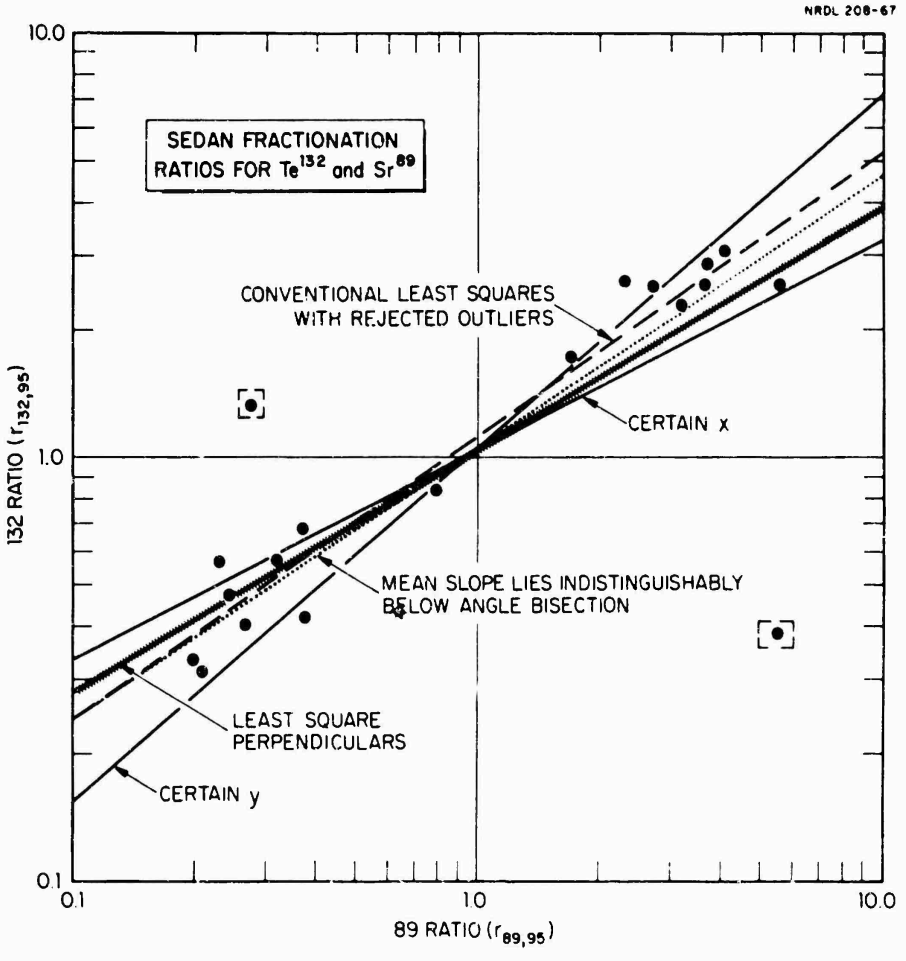

Fig. 4 Comparison of Regression Parameters. Outliers are in dashed squares.

is slightly greater than that for the mean slope ($\hat{b} = 0.645$), but the lines are indistinguishable on the graph. The slopes of these lines are somewhat larger than that for the least square perpendiculars.

RECOMMENDATION

The choice of a method from the alternatives presented must be made in light of the realizations that: (1) The choice is not critical; (2) Cases with low values of r are of little practical significance; (3) Cases of $r \approx 1$, $b \approx 1$ and $b \approx 0$ do not usually present a problem. The considerations we have presented argue in favor of the geometric-mean regression line ($\hat{b}$) for the following reasons: (1) Its parameters are very simple to calculate from quantities usually obtained by the conventional practice of regarding x as certainly known; (2) Since it gives results which are nearly equal to those obtained from the angle-bisection treatment, it has all the advantages of that method; and (3) It gives results for the slope which are intermediate to those obtained by rejecting outliers and those obtained by least square perpendiculars.

Although little experience has been obtained to date on the application of this method, no circumstances which would dictate another choice are foreseeable at this time.

The similarity between the lines obtained by rejection of outliers and the geometric-mean line indicates that the geometric mean should receive further attention as a means of handling the general problem of outliers.

REFERENCES

1. C. A. Bennett and N. L. Franklin, Statistical Analysis in Chemistry and the Chemical Industry, Wiley, New York (1954).

2. M. Ezekiel and K. A. Fox, Methods of Correlation and Regression Analysis, Wiley, New York (1959).

3. W. B. Lane, "Some Radiochemical and Physical Measurements of Debris From an Underground Nuclear Detonation," Project Sedan, Atomic Energy Commission Report PNE 229F, U. S. Naval Radiological Defense Laboratory, June 1963.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

U. S. Naval Radiological Defense Laboratory
San Francisco, California 94135

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

REGRESSION PARAMETERS FOR PAIRS OF EQUALLY UNCERTAIN VARIABLES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*

Edward C. Freiling
Glenn R. Crocker

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
16 October 1967	29	3

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AEC Contract AT(49-7)-1963	USNRDL-TR-67-99
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Atomic Energy Commission Washington, D. C. 20545

13. ABSTRACT

The correlation of radiochemical data from samples of fractionated nuclear debris involves the treatment of two variables whose uncertainties are comparable. We considered three new criteria for the establishment of regression parameters for such correlations (least square perpendicular distances between points and the line, bisection of the angle formed by the certain-x and certain-y regression lines, and adoption of the geometric mean of the certain-x and certain-y regression slopes). We concluded that the geometric-mean slope $\hat{b}$ was most satisfactory. It is related to the usual certain-x regression slope $b_{x,x}$ and the coefficient of correlation r by the simple expression

$$\hat{b} = b_{x,x} / |r|$$

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Regression						
Fractionation						
Statistics						
Nuclear debris						
Least squares						