NASA

Technical Memorandum 80733

Relationships Among the Slopes of Lines Derived from Various Data Analysis Techniques and the Associated Correlation Coefficient

Steven C. Cohen

```
(NASA-TM-80733)  RELATIONSHIPS AMONG THE            N80-31097
SLOPES OF LINES DERIVED FROM VARIOUS DATA
ANALYSIS TECHNIQUES AND THE ASSOCIATED
CORRELATION COEFFICIENT (NASA)   7 p                Unclas
HC A02/MF A01                         CSCL 12A G3/64 31322
```

JULY 1980

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

RELATIONSHIPS AMONG THE SLOPES OF LINES DERIVED FROM VARIOUS DATA ANALYSIS TECHNIQUES AND THE ASSOCIATED CORRELATION COEFFICIENT

Steven C. Cohen
Geodynamics Branch

Goddard Space Flight Center
Greenbelt, Maryland 20771

July 1980

GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland 20771

RELATIONSHIPS AMONG THE SLOPES OF LINES DERIVED FROM VARIOUS DATA ANALYSIS TECHNIQUES AND THE ASSOCIATED CORRELATION COEFFICIENT

Steven C. Cohen
Geodynamics Branch
Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

There are several techniques for fitting a straight line to a collection of data points. In the expression $Y = a + bX$ the parameters of interest are the intercept, a, and slope, b. Herein these parameters are subscripted by y if they are derived by minimizing the variance in Y, by x if the variance in X is minimized, and by xy if a reduced major axis analysis issued (see text). The correlation coefficient is designated by r. This paper notes that the slopes and correlation coefficients are related through $r^2 = b_y/b_x = (b_y/b_{xy})^2$. The corresponding standard deviations and correlation coefficient are related by $r^2 = S_{b_y}/S_{b_x} = S_{b_{xy}}/S_{b_x}$.

RELATIONSHIPS AMONG THE SLOPES OF LINES DERIVED FROM VARIOUS DATA ANALYSIS TECHNIQUES AND THE ASSOCIATED CORRELATION COEFFICIENT

This note points some simple and insightful features associated with fitting a straight line to data which may be useful to, but widely unknown by, many members of the scientific community. Specifically there exist relations connecting the correlation coefficient and slopes of the lines derived using alternative data analysis criteria. In the following discussion X_i and Y_i are to be regarded as observations, neither of which can be preferentially regarded as an independent or dependent variable. The analyses use the data along with error minimization criteria to derive the intercept, a, and slope, b, in the assumed relationship between X and Y, namely $Y = a + bX$. If the variance in Y is minimized, the derived constants are

$$b_y = \frac{N\Sigma X_i Y_i - \Sigma X_i\, \Sigma Y_i}{N\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{S_{xy}}{S_x^2} \qquad (1a)$$

$$a_y = \frac{\Sigma Y_i - b_y \Sigma X_i}{N} = \overline{Y} - b_y\, \overline{X} \qquad (1b)$$

where S_{xy} is the covariance of X and Y, S_x is the standard deviation in X, $\overline{Y}$ and $\overline{X}$ are mean values and the sums are taken over the N sets of observations. Alternatively if the variance in X is minimized

$$b_x = \frac{N\Sigma Y_i^2 - (\Sigma Y_i)^2}{N\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i} = \frac{S_y^2}{S_{xy}} \qquad (2a)$$

$$a_x = \overline{Y} - b_x \overline{X} \qquad (2b)$$

Taking the ratio of the slope b_y to the slope b_x results in an expression which can be recognized as the square of the correlation coefficient, r, between X and Y, i.e.

$$\frac{b_y}{b_x} = \frac{(N\Sigma X_i Y_i - X_i \Sigma Y_i)^2}{[N\Sigma X_i^2 (\Sigma X_i)^2]\ [N\Sigma Y_i^2 - \Sigma Y_i)^2]} = r^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2 \tag{3}$$

In general $b_x \geqslant b_y$ since b_y is derived from minimizing the variance in Y and b_x from minimizing the variance in X. The slopes are equal if the data is perfectly correlated, $r^2 = 1$; otherwise, the degree to which the slopes differ is a measure of the departure from perfect correlation.

An alternative procedure for fitting a straight line to data, one which treats X and Y in a more symmetric way than those so far considered, minimizes the sum of the triangular areas formed by the derived straight line and lines parallel to the coordinate axes through the data points. This reduced major axis formulation results in (Kermack and Haldane, 1950)

$$b_{xy} = \frac{N\Sigma Y_i^2 - (\Sigma Y_i)^2}{N\Sigma X_i^2 - (\Sigma X_i)^2} = \left(\frac{S_y}{S_x}\right)^2 \tag{4a}$$

$$a_{xy} = Y - b_{xy}\,\bar{X} \tag{4b}$$

It now follows that

$$b_{xy}^2 = b_y b_x = \left(\frac{b_y}{r}\right)^2 = (r\,b_x)^2 \tag{5}$$

Thus if two of the parameters r, b_x, b_y, and b_{xy} are known, the other two can be determined with little further effort. The standard deviation in the slopes are also related tc one another and the correlation coefficient by

$$S_{b_y} = S_{b_{xy}} = r^2\, S_{b_x} \tag{6}$$

The preceeding expressions are useful for one set of straight line parameters from another, a procedure which is aided by the expression for the intercept, $a_j = \bar{Y} - b_j \bar{X}$ where j = x, y, or xy. They are also useful for assessing the consequences of either using an inappropriate minimization criterion or inverting a linear regression expression to get X = A + BY where A = –a/b and B = 1/b without considering the implied changes in the selection of independent and dependent variables.

Reference

Kermack, K A. and J. B. S. Haldane, "Organic Correlation and Allometry," Biometrika, 37, 30-41, 1950.