NOTICE

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE
Relationships Among the Slopes of Lines Derived from Various Data Analysis Techniques and the Associated Correlation Coefficient

Steven C. Cohen

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 cor-
relation coefficient is designated by \( r \). This paper notes that the slopes and correlation coeffi-
cients are related through \( r^2 = \frac{b_y}{b_x} = \left( \frac{b_y}{b_{xy}} \right)^2 \). The corresponding standard deviations and
correlation coefficient are related by \( r^2 = \frac{S_y}{S_x} = \frac{S_{xy}}{S_x} \).
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

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

$$a_y = \frac{\Sigma Y_i - b_y \Sigma X_i}{N} = \bar{Y} - b_y \bar{X}$$  \hspace{1cm} (1b)

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

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

$$a_x = \bar{Y} - b_x \bar{X}$$  \hspace{1cm} (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. 
In general \( b_x > b_y \) 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 (Kemack and Haldane, 1950)

\[
\begin{align*}
\frac{b_y}{b_x} &= \frac{(N\sum X_i Y_i - X_i \sum Y_i)^2}{N\sum X_i^2 (\sum Y_i)^2} \\
\frac{r^2}{\left(\frac{S_{xy}}{S_x S_y}\right)^2} &= r^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2
\end{align*}
\]

(3)

It now follows that

\[
b_{xy} = b_y b_x = \left(\frac{b_y}{r} + b_x\right)^2 = (r b_x)^2
\]

(4b)

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 to one another and the correlation coefficient by

\[
S_{b_y} = S_{b_{xy}} = r^2 S_{b_x}
\]

(6)

The preceding expressions are useful for one set of straight line parameters from another, a procedure which is aided by the expression for the intercept, \( a_j = Y - b_y X \) 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 \( X = \Lambda + B Y \) where \( \Lambda = -a/b \) and \( B = 1/b \) without considering the implied changes in the selection of independent and dependent variables.
Reference