United States
Department of the Interior
Geological Survey

[Reprints - Open file series]

Straight Line Fitting of an Observation Path by Least Normal Squares

by

W. Kirby

Open-File Report 74-187

Reston, Virginia

1974

Contents

Illustrations

Metric Conversion Table

English	x	factor	=	Metric
feet (ft)		0.3048		metres (m)
yards (yd)		0.9144		metres (m)
nautical miles		1.8532		kilometres (km)

Straight Line Fitting of an Observation Path by Least Normal Squares

W. Kirby

Abstract

In certain hydrographic problems, and perhaps in other geophysical problems, information must be collected as profiles along straight-line courses. When the inevitable deviations from perfect linearity occur, one must then find the straight line course that best approximates the actual observation path. Classical least squares (regression) does not solve this problem, because the line thus fitted depends upon which coordinate is taken as the dependent variable.

A coordinate-free solution is obtained by minimizing the sum of squares of normal distances between the line and the observation points. As in classical least squares, the line of best fit passes through the geometrical centroid of the observation points. The slope of this line, however, is closer to 1.0 than the slope of the classical regression line. The least normal squares and classical least squares solutions coincide when the observation path is nearly perfectly straight or when it runs generally parallel to one of the coordinate axes.

Introduction

Classical regression theory deals with the problem of finding the linear function $y = mx + b$ which best approximates a set of observed values y_i at fixed observation points x_i. The criterion is the sum of squared errors of approximation,

$$\Sigma_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{1}$$

where $y_i = mx_i + b$. The parameters m and b are selected so as to minimize this sum, whence the name least squares. This classical regression problem has wide applicability to the fitting of functional relationships to empirical data.

Not every curve-fitting problem, however, fits the classical regression mold. In estuarine modeling, for example, and no doubt in other hydrographic and topographic problems, one sometimes wishes to characterize the bottom topography by profiles measured (in the field) along straight lines. There exist various systems for collecting such profiles, but inevitably, because of the rigors of field operations, at least some of the observation points will deviate from the desired straight line course.

Figure 1 illustrates typical observation paths followed during bathymetric profiling in Tampa Bay, Florida. Observations were taken at 10-20-ft (3-6-m) intervals along these paths. These cross sections, measured at 500-ft (150-m) intervals along the navigation channel, have been used, for example, to estimate dredging volumes for a proposed channel improvement.

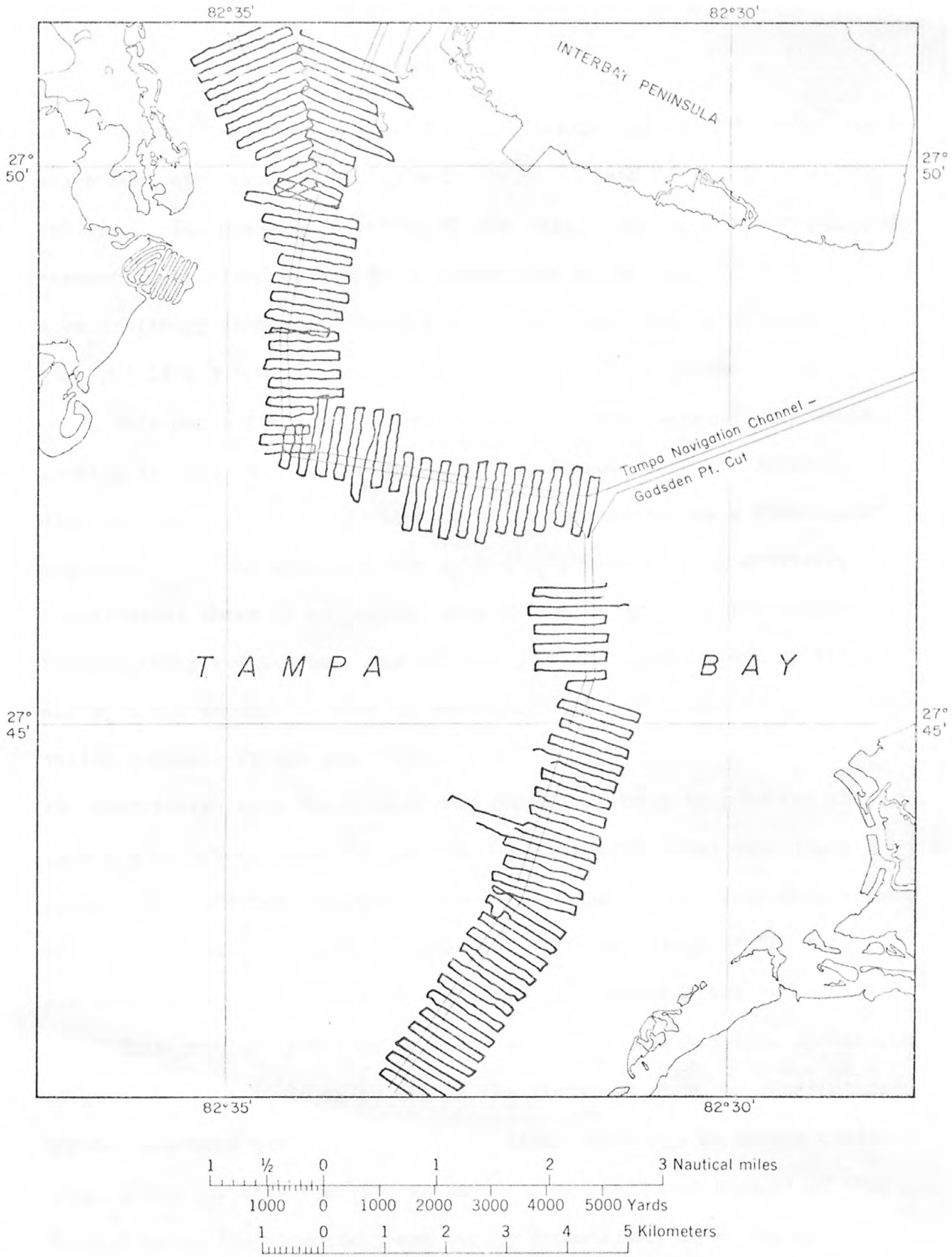

Figure 1. Typical bathymetric profiling paths in Tampa Bay, Florida. (Courtesy of R. A. Baltzer)

The curve-fitting problem here is to find the straight line course which best approximates the actual path of the observation vehicle. The measured position of the vehicle at each observation point is assumed to be given by a pair of coordinate values (x_i, y_i) relative to some arbitrary cartesian coordinate system. What has to be found is the straight line $y = mx + b$ which best fits these data points.

This curve-fitting problem differs from the classical regression problem in that the data values x_i and y_i do not represent logically distinct variables, one of which is to be represented as a function of the other. On the contrary, the x_i and y_i represent only arbitrary coordinates; there is no logical distinction or order of precedence between them; and, in fact, any orthogonal linear combination of the x_i and y_i would provide an equally meaningful representation of the observation points. By the same token, there is no desire to represent one of the coordinates as a function of the other, but only to pass the straight line $y = mx + b$ as close as possible to all the observed coordinate pairs. The classical regression approach fails to deal with this aspect of this problem: it is well known that different lines of "best fit" are obtained when the roles of the x and y coordinates are interchanged.

This problem is solved by postulating a coordinate-free definition of goodness of fit, in which errors are distances from the observation points, measured perpendicular to the line. Although we cannot claim originality for this concept, we have not seen any development of this concept along the elementary and purely geometrical lines reported here.

Nonetheless, Cramér (1946, p. 275) has used arguments from analytic geometry to derive what he calls the "orthogonal mean square regression line," which is precisely the line we derive below by more elementary methods. Similarly, in considering linear regression with both the ordinates and the abcissas subject to random error, Kenney and Keeping (1951, p. 213) use the method of maximum likelihood to arrive at the same equations we obtain below, in effect deriving least normal squares from maximum likelihood. In the same vein, Ware (1972) has compared methods for estimating the intended course of a particle that randomly strays away from a straight line. Ware's work, like Kenny's and Keeping's, rests on statistical hypotheses which may or may not be satisfied in the practical problem of fitting a straight line course. What we think is important, however, is that the immediate practical course-fitting problem can be resolved and compared and contrasted with the classical regression solution without recourse to statistical hypotheses, simply by invoking a principle of least normal squares.

The Normal Distance

An appropriate objective for this straight-line course fitting problem, therefore, is to minimize the sum of squared distances between the line and the data points, the distances being measured normal to the line. Referring to figure 1 and assuming temporarily that the slope m is finite--that the line is not vertical--the normal distance δ_i from the data point to the line is given by

$$\delta_i^2 = \frac{\epsilon_i^2}{m^2 + 1} \tag{2}$$

where

$$\epsilon_i = \hat{y} - y_i = mx_i + b - y_i \tag{3}$$

The objective function to be minimized by proper choice of m and b, therefore, is the sum of the normal squares, δ_i^2, as follows:

$$\Delta(m,b) = \Sigma_{i=1}^{n} (mx_i + b - y_i)^2/(m^2 + 1) \tag{4}$$

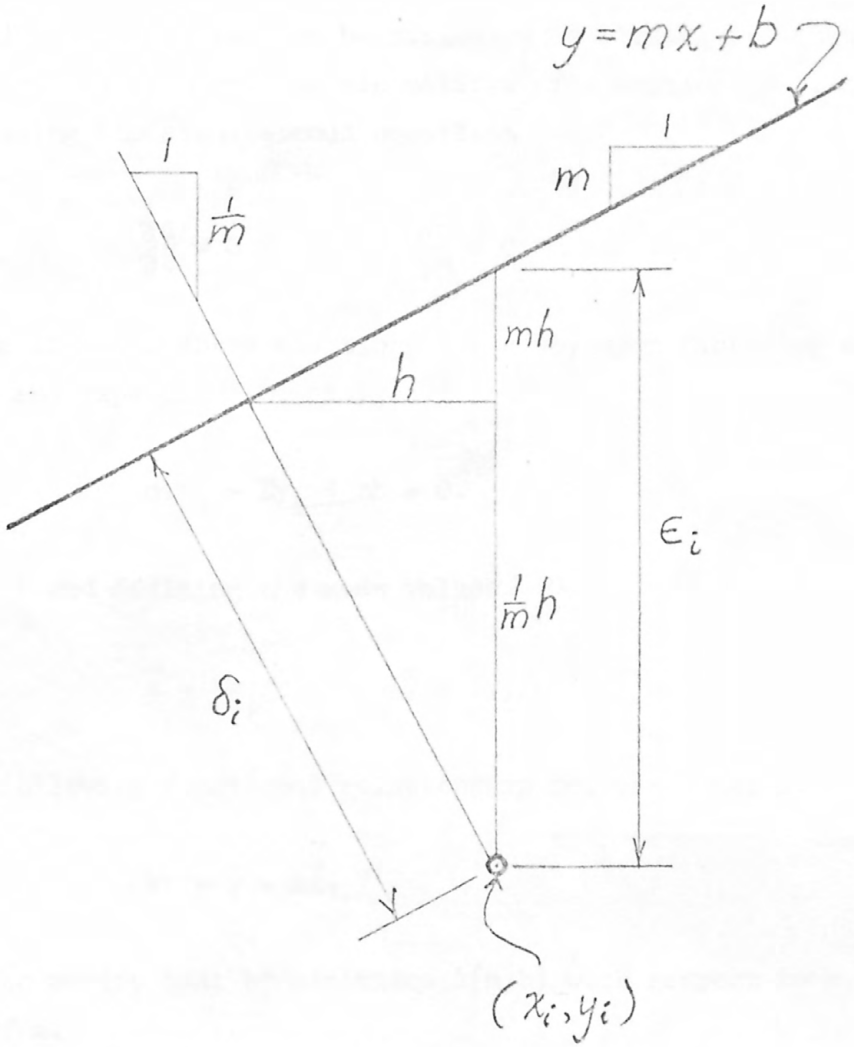

Figure 1. Definition sketch for calculating the normal distance.

This objective function can be minimized by finding its stationary points and testing each one for minimality. The stationary points are found by solving the simultaneous equations

$$\frac{\partial \Delta}{\partial b} = 0 \qquad \frac{\partial \Delta}{\partial m} = 0. \tag{5}$$

From the first of these equations one gets, upon factoring out $2/(m^2 + 1)$, and expanding the sum,

$$m\Sigma x_i - \Sigma y_i + nb = 0.$$

Solving for b and defining the mean values

$$\bar{x} = \Sigma x_i/n \qquad \bar{y} = \Sigma y_i/n \tag{6}$$

yields the following functional relationship between b and m:

$$b^* = \bar{y} - m\bar{x}. \tag{7}$$

It is easy to verify that b^* minimizes $\Delta(m,b)$ with respect to b, at any value of m.

In coordinate-free geometric terms, equation 7 makes the following assertion: the line of best fit passes through the centroid $(\bar{x},\bar{y})$ of the data points.

Because the optimal value of b is now known for any m, it is sufficient to consider the restriction of the objective function to the optimal line $b^* = \bar{y} - m\bar{x}$. Upon substituting equation 7 into equation 4, the restricted objective function becomes

$$\Delta^*(m) = \Delta(m,b^*) = \sum_{i=1}^{n} (mx_i + [\bar{y} - m\bar{x}] - y_i)^2/(m^2 + 1) \tag{8}$$

which is a function of m only, because at any m only the corresponding optimal value of b is considered.

Recognizing in equation 8 the occurrence of deviations from the means and denoting them by

$$\xi_i = x_i - \bar{x} \qquad \eta_i = y_i - \bar{y}$$

the sum becomes

$$\Sigma(m\xi_i - \eta_i)^2 = m^2\Sigma\xi_i^2 - 2m\Sigma\xi_i\eta_i + \Sigma\eta_i^2.$$

Recognizing the significance of sums of squares and cross products and defining the variances, covariance, and correlation coefficient of the data yields the following form of the restricted objective function:

$$\Delta^*(m) = n\,\frac{m^2 s_x^2 - 2mrs_x s_y + s_y^2}{m^2 + 1} \tag{9}$$

in which

$$s_x^2 = \Sigma\xi_i^2/n \qquad s_y^2 = \Sigma\eta_i^2/n$$

$$s_{xy} = \Sigma\xi_i\eta_i/n \qquad r = s_{xy}/s_x s_y \tag{10}$$

To minimize $\Delta^*(m)$ one first finds its stationary points by differentiation. The result, after factoring out $2n/(m^2 + 1)^2$ and combining terms, is the quadratic equation

$$(m^2 - 1)rs_x s_y + m(s_x^2 - s_y^2) = 0 \tag{11}$$

in which m is the unknown and r, s_x, and s_y are known geometrical properties of the set of data points $\{(x_i, y_i)\}$.

To solve this equation one first notes that if either s_x or s_y is zero, the data all lie on one of the (straight) coordinate lines, so the solution is trivial. In the contrary case, when both s_x and s_y exceed zero, one implements the finite-slope assumption underlying the error measure by requiring that the scatter of the data in the y direction not exceed that in the x direction, in the sense that

$$s_y \leq s_x \tag{12}$$

(This condition can be achieved, if necessary, by relabeling the coordinate axes.) Then, dividing by $s_x s_y$, one obtains

$$(m^2 - 1)r + m\left(\frac{s_x}{s_y} - \frac{s_y}{s_x}\right) = 0 \tag{13}$$

Now if the correlation r is zero, then the slope m must be either zero or indeterminate. (In the latter case the objective function is constant with respect to m.) If r is nonzero, however, division by r yields

$$m^2 + m\left(\frac{s_x}{rs_y} - \frac{s_y}{rs_x}\right) - 1 = 0$$

Those who are in the know will recognize the terms

$$m_x = rs_x/s_y \qquad\qquad m_y = rs_y/s_x$$

as the slopes of classical regressions of x on y and of y on x. A more convenient formulation of the equation, however, seems to be in terms of what we shall call the <u>scale ratio</u> of y with respect to x, defined by

$$\psi = s_y/s_x \tag{14}$$

By hypothesis $\psi \leq 1$. In these terms equation 13 becomes

$$m^2 + m(\frac{1}{\psi} - \psi)/r - 1 = 0. \tag{15}$$

The solution of this equation is

$$m = -(A/r) \pm \sqrt{1 + (A/r)^2} \tag{16}$$

where

$$A = (\frac{1}{\psi} - \psi)/2 \tag{17}$$

By laborious evaluation of $\partial^2 \Delta^*/\partial m^2$ at the stationary points, one finds that the condition for minimization is that rm exceed - A. Thus the minimizing value of m is the upper branch when r is positive and the lower one when r is negative. These results may be summarized in the following

Assertion:

$$m^* = \begin{cases} 0 & \text{if } r = 0 \\ \dfrac{-A + \sqrt{r^2 + A^2}}{r} & \text{if } r \neq 0 \end{cases} \tag{18}$$

This completes the solution of the minimum-normal-distance straight-line-fitting problem.

Comparison with Classical Least Squares

The value of m^* may be compared with the slope of the classical regression line in a special case which can be achieved by proper choice of coordinates. The classical regression line has slope $m_y = rs_y/s_x = r\psi$. When ψ is small, so that the data are nearly horizontal, the following approximation may be used for equation 18:

$$m^* \cong r\psi(\frac{1}{1 - \psi^2}) \qquad (\psi \cong 0)$$

In this case m^* is slightly larger than the classical regression slope m_y. This approximation holds also when ψ is fixed and r is very small. If ψ is close to 1, and r is made very small, the approximation takes the simpler form

$$m^* \cong \frac{r}{2(1 - \psi)} \qquad (\psi \cong 1,\ r \cong 0)$$

Finally, in the important special case of r close to 1, the slope is given by

$$m^*(1) = \psi \tag{19}$$

$$\frac{d}{dr} m^*(1) = \psi \frac{1 - \psi^2}{1 + \psi^2} \tag{20}$$

Thus, when r is close to 1, m^* agrees with the classical regression slope--as it should because all the data lie very close to the line.

Some numerical computations of m^* are illustrated in figure 2. A graph of the corresponding classical least squares slopes,

$$m_y = r\psi \tag{21}$$

would be a family of straight lines radiating from the origin to the terminal points (at r = 1) of the m^* curves.

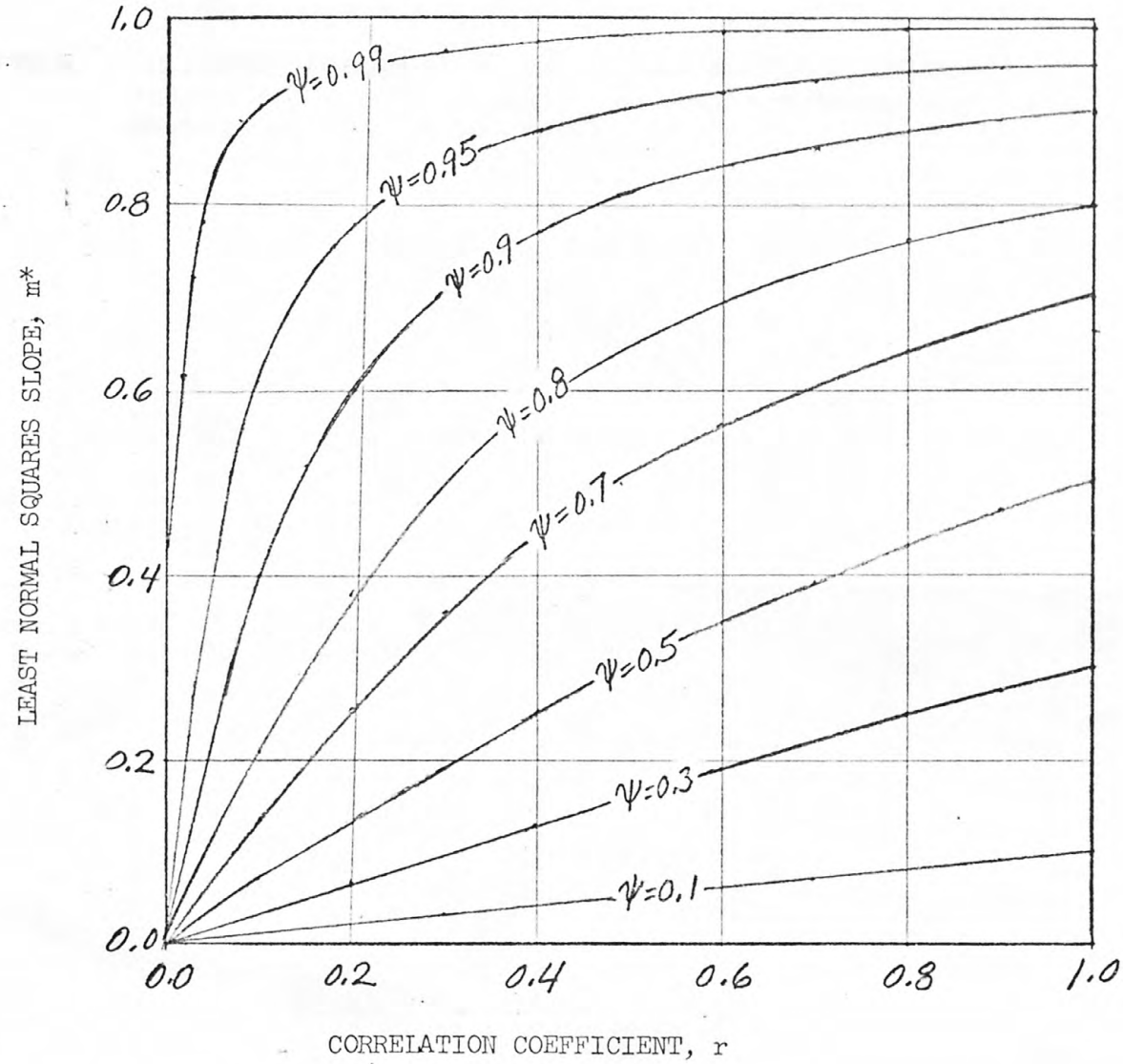

Figure 3. Least normal squares slope as a function of correlation coefficient r and scale ratio $\psi = s_y/s_x$.

References

Cramér, Harald, 1946, Mathematical Methods of Statistics: Princeton, N. J., Princeton Univ. Press, 575 pp.

Kenney, J. F., and Keeping, E. S., 1951, Mathematics of Statistics, Part 2, 2n edition: Princeton, N. J., D. Van Nostrand Co., 429 pp.

Ware, J. H., 1972, Fitting of straight lines when both variables are subject to error and ranks of the means are known: Jour. Am. Stat. Assoc., v. 67, pp. 891-897.