

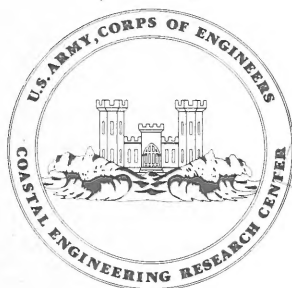
# Forcing Regression Through a Given Point Using Any Familiar Computational Routine

by

Edward B. Hands

TECHNICAL PAPER NO. 83-1

MARCH 1983



Approved for public release;  
distribution unlimited.

U.S. ARMY, CORPS OF ENGINEERS  
COASTAL ENGINEERING  
RESEARCH CENTER

Kingman Building  
Fort Belvoir, Va. 22060

6B  
458  
TL  
no. 83-1

Reprint or republication of any of this material shall give appropriate credit to the U.S. Army Coastal Engineering Research Center.

Limited free distribution within the United States of single copies of this publication has been made by this Center. Additional copies are available from:

*National Technical Information Service  
ATTN: Operations Division  
5285 Port Royal Road  
Springfield, Virginia 22161*

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



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE   |                       | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM                                  |
|---|-----------------------|--|
| 1. REPORT NUMBER<br>TP 83-1   | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER  |
| 4. TITLE (and Subtitle)<br><br>FORCING REGRESSION THROUGH A GIVEN POINT<br>USING ANY FAMILIAR COMPUTATIONAL ROUTINE   |                       | 5. TYPE OF REPORT & PERIOD COVERED<br><br>Technical Paper                    |
|   |                       | 6. PERFORMING ORG. REPORT NUMBER   |
| 7. AUTHOR(s)<br><br>Edward B. Hands   |                       | 8. CONTRACT OR GRANT NUMBER(s)   |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of the Army<br>Coastal Engineering Research Center (CEREN-GE)<br>Kingman Building, Fort Belvoir, VA 22060   |                       | 10. PROGRAM ELEMENT, PROJECT, TASK<br>AREA & WORK UNIT NUMBERS<br><br>D31677 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Department of the Army<br>Coastal Engineering Research Center<br>Kingman Building, Fort Belvoir, VA 22060  |                       | 12. REPORT DATE<br>March 1983  |
|   |                       | 13. NUMBER OF PAGES<br>20  |
| 14. MONITORING AGENCY NAME & ADDRESS (If different from Controlling Office)   |                       | 15. SECURITY CLASS. (of this report)<br><br>UNCLASSIFIED                     |
|   |                       | 15a. DECLASSIFICATION/DOWNGRADING<br>SCHEDULE                                |
| 16. DISTRIBUTION STATEMENT (of this Report)<br><br>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)  |                       |  |
| <div style="display: flex; justify-content: space-between;"> <div>Coastal engineering<br/>Data analysis</div> <div>Prediction equations<br/>Regression</div> </div>   |                       |  |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number)<br>This report describes a simple method for obtaining the prediction equation best fit to all data points (in the least squares sense) while forcing an exact fit at any known point. The decision to constrain the solution at a point should be justified on theoretical grounds without appeal to data. Examples are given. When required any familiar regression program can be forced to select the best line through a given point by simply adjusting and extending the data entry. All necessary changes to the program results (test statistics and estimates of regression parameters) can be accomplished without modifying the computer program. |                       |  |



## PREFACE


This report draws attention to the frequent, but often neglected, need to force a regression line through a known point while obtaining the best possible fit to all experimental data points. A simple method is described for solving this problem without modifying customary computational routines. This method can be applied to many problems, but is especially useful when calibrating empirical prediction formulas to fit site-specific coastal conditions or when choosing from among several theoretical prediction models. The work was carried out under the U.S. Army Coastal Engineering Research Center's (CERC) Shore Response to Offshore Dredging work unit, Shore Protection and Restoration Program, Coastal Engineering Area of Civil Works Research and Development.

The report was prepared by Edward B. Hands, Geologist, under the general supervision of Dr. C.H. Everts, Chief, Engineering Geology Branch, and Mr. N. Parker, Chief, Engineering Development Division. The author acknowledges the helpful suggestions received from C.B. Allen, C.H. Everts, R.J. Hallermeier, R.D. Hobson, and P. Vitale.

Technical Director of CERC was Dr. Robert W. Whalin, P.E.

Comments on this publication are invited.

Approved for publication in accordance with Public Law 166, 79th Congress, approved 31 July 1945, as supplemented by Public Law 172, 88th Congress, approved 7 November 1963.

  
TED E. BISHOP  
Colonel, Corps of Engineers  
Commander and Director

## CONTENTS

|   | Page |
|---|------|
| CONVERSION FACTORS, U.S. CUSTOMARY TO METRIC (SI) . . . . . | 5    |
| SYMBOLS AND DEFINITIONS. . . . .                            | 6    |
| I INTRODUCTION TO REGRESSION . . . . .                      | 9    |
| II A PROBLEM WITH THE CUSTOMARY APPROACH. . . . .           | 11   |
| III SOLUTION TO THE PROBLEM. . . . .                        | 12   |
| 1. Regression Through the Origin . . . . .                  | 12   |
| 2. Regression Through Any Arbitrary Point (a, b) . . . . .  | 12   |
| IV SELECTING BETWEEN MODELS I AND II. . . . .               | 13   |
| V EXAMPLES . . . . .  | 15   |
| LITERATURE CITED . . . . .                                  | 20   |

## TABLES

|   |    |
|---|----|
| 1 Adjustment of standard elements produced by programs using<br>extended data . . . . . | 14 |
| 2 Field calibration data . . . . .  | 16 |
| 3 Extended data set No. 1. . . . .  | 19 |
| 4 Extended data set No. 2. . . . .  | 19 |

## FIGURES

|   |    |
|---|----|
| 1 Application of Model I produces an intercept ( $\hat{\alpha}$ ), which may be a<br>useful estimate of a component of longshore flow which is independ-<br>ent of wave conditions and presumably pervades the entire data set. . | 10 |
| 2 Application of Model I identified a threshold value below which<br>waves cause no damage . . . . .  | 10 |
| 3 Application of Model II forces a zero-intercept solution . . . . .  | 11 |
| 4 Model II estimates an increase in Y per unit increase in X that<br>is nearly twice that predicted using Model I. . . . .  | 11 |
| 5 Real test data for example problem 1 . . . . .  | 17 |
| 6 Real test data for example problem 2 and fitted equations. . . . .  | 18 |

# CONVERSION FACTORS, U.S. CUSTOMARY TO METRIC (SI) UNITS OF MEASUREMENT

U.S. customary units of measurement used in this report can be converted to metric (SI) units as follows:

| Multiply           | by                      | To obtain                               |
|--------------------|-------------------------|---|
| inches             | 25.4                    | millimeters                             |
|                    | 2.54                    | centimeters                             |
| square inches      | 6.452                   | square centimeters                      |
| cubic inches       | 16.39                   | cubic centimeters                       |
| feet               | 30.48                   | centimeters                             |
|                    | 0.3048                  | meters                                  |
| square feet        | 0.0929                  | square meters                           |
| cubic feet         | 0.0283                  | cubic meters                            |
| yards              | 0.9144                  | meters                                  |
| square yards       | 0.836                   | square meters                           |
| cubic yards        | 0.7646                  | cubic meters                            |
| miles              | 1.6093                  | kilometers                              |
| square miles       | 259.0                   | hectares                                |
| knots              | 1.852                   | kilometers per hour                     |
| acres              | 0.4047                  | hectares                                |
| foot-pounds        | 1.3558                  | newton meters                           |
| millibars          | $1.0197 \times 10^{-3}$ | kilograms per square centimeter         |
| ounces             | 28.35                   | grams                                   |
| pounds             | 453.6                   | grams                                   |
|                    | 0.4536                  | kilograms                               |
| ton, long          | 1.0160                  | metric tons                             |
| ton, short         | 0.9072                  | metric tons                             |
| degrees (angle)    | 0.01745                 | radians                                 |
| Fahrenheit degrees | 5/9                     | Celsius degrees or Kelvins <sup>1</sup> |

<sup>1</sup>To obtain Celsius (C) temperature readings from Fahrenheit (F) readings, use formula:  $C = (5/9) (F - 32)$ .

To obtain Kelvin (K) readings, use formula:  $K = (5/9) (F - 32) + 273.15$ .

# SYMBOLS AND DEFINITIONS

F The F-value may be produced by a multiple regression program and is analogous to the t-value in simple regression (one independent variable). The F-value indicates the "significance" of  $r^2$  and is useful in selecting the most important independent variables.

$$F = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \hat{y})^2} \left( \frac{n - p - 1}{p} \right) = \frac{r^2}{1 - r^2} \left( \frac{n - p - 1}{p} \right)$$

$H_b$  height of breaking waves

n size of the sample

p total number of independent variables. Caution, several observed carriers may end up combined into a single independent variable; e.g.,  $X = (gH_b)^{1/2} \sin 2\alpha_b$  has two distinct carriers ( $H_b$  and  $\alpha_b$ ) but is one independent variable (see example problem 1). The value of p will be one less than the number of constants to be estimated in Model I, and is equal to the number of constants in Model II.

r sample correlation coefficient. The r-value produced by regression partially measures the closeness of fit between the linear predictor and data. Its square is called the coefficient of determination.

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = \frac{\Sigma(y - \bar{y})(x - \bar{x})}{\sqrt{\Sigma(y - \bar{y})^2 \Sigma(x - \bar{x})^2}} \quad (\text{Model I})$$

$$r^2 = \frac{\Sigma yx}{(\Sigma y)^2 (\Sigma x)^2} \quad (\text{Model II})$$

$SS_x$  sum of squares of x may be produced by the regression program and is useful for computing other values, e.g.,  $S_{\hat{\beta}}$ .

$$SS_x = \Sigma(x - \bar{x})^2$$

$S_{\hat{\beta}}$  standard error of the estimated slope,

$$S_{\hat{\beta}} = \sqrt{\frac{S_{\hat{y} \cdot x}^2}{SS_x}}$$

The larger  $S_{\hat{\beta}}$ , the less reliable is the estimate of slope.

$S_{y \cdot x}^2$  unbiased estimator of the variance of the random component  $\epsilon$ , e.g.,

$$S_{y \cdot x}^2 = \frac{\Sigma(y - \hat{y})^2}{n - p - 1} \quad \text{in Model I}$$

The number of independent variables, p, is 1 in simple regression with Model I. The mean square deviation from regression corresponds to



the simple variance used to measure the spread of values in a single data set. It is also sometimes called the *standard error of the estimate*. The value produced by regression to indicate uncertainty of the estimated  $y$ ; the value  $S_{y \cdot x}^2$  depends on the variances of all the estimated coefficients.

t The t-value produced in simple regression to test whether the estimated regression coefficient is "significantly" different from zero.

$$t = \frac{\hat{\beta} - 0}{S_{\hat{\beta}}}$$

v longshore current velocity

X independent variable in regression

x observed values of X. A string of n-values in simple regression; a n by p matrix in multiple regression

Y dependent variable to be estimated

y n observed values of Y

$\hat{y}$  estimated value of Y for given values of X

$\hat{\alpha}$  Y-intercept in a regression model

$\alpha_b$  angle between the crest of the breaking wave and the shoreline

$\hat{\beta}$  estimated regression coefficients in multiple regression or the slope of the line in simple regression

$$\hat{\beta} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \text{ (Model I)}$$

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} \text{ (Model II)}$$

$\epsilon$  zero-mean random component of Y assumed by both regression models



# FORCING REGRESSION THROUGH A GIVEN POINT USING ANY FAMILIAR COMPUTATIONAL ROUTINE

by

*Edward B. Hands*

## I. INTRODUCTION TO REGRESSION

The engineer frequently needs to estimate some response or dependent variable  $Y$  (e.g., sand transport rate, change in shoreline position, or structural damage), when given the magnitude of other factors, or independent variables  $X$  (e.g., longshore wave energy flux, storm frequency, elevation of storm surges, etc.). A common approach is to assume a linear model,

$$Y = a + \beta X + \epsilon \text{ (Model I)}$$

then adopt the principle of least squares; and use sample data to estimate the unknown parameters,  $\alpha$  and  $\beta$ . Both  $\beta$  and  $X$  can be considered as strings of numbers in the case of multiple regression with several independent variables;  $\epsilon$  indicates that the response is not being thought of as an exact linear function of  $X$ . The  $\epsilon$  represents random and unpredictable elements in  $Y$ ; therefore,  $\epsilon$  does not appear in the prediction equation:  $\hat{y} = \hat{a} + \beta x$ , where  $\hat{y}$ ,  $\hat{a}$ , and  $\hat{\beta}$  are estimates of the corresponding components in the conceptual Model I. The assumption that  $\epsilon$  has an expected value of zero indicates that the "average" response is considered linear. If  $\epsilon$  varies widely, Model I, though conceptually correct, may have only limited predictive value. In such a case the estimated mean value of  $Y$  would frequently be thrown off by noise in the data. If  $\epsilon$  varies only slightly, good predictions will be possible provided good estimates of  $\alpha$  and  $\beta$  are available. Adopting the principle of least squares means one is willing to define the best estimates of  $\alpha$  and  $\beta$  as those that minimize the sum of the squares of the deviations between the observed and predicted values (i.e.,  $y$  and  $\hat{y}$ ).

Customarily, no constraints are placed on the contenders for the best fit line. Of all possible lines in the  $XY$  plane, the prediction equation is chosen because it has the least sums of squares of deviations in  $\hat{y}$ 's from the data points. The  $y$ -intercept,  $\hat{a}$ , is the point where the best fit line intersects the  $Y$ -axis. The  $\hat{a}$  may be of special interest, e.g., in the regression of current speed against longshore wave energy flux measured in a field test (Fig. 1). An intercept substantially above zero would suggest that during the test a component of the longshore current was driven by mechanisms other than waves (e.g., tides or winds). In this case, the nonzero intercept would not only be meaningful, but would also provide a good estimate of the velocity of any steady, nonwave-generated coastal current during the test.

An additional example of unconstrained regression would be where greater and greater structural damage occurs as the wave forces exceed an undetermined threshold value. Again Model I applies and produces the correct regression coefficient ( $\hat{\beta}$ ). In the process it produces a meaningless response intercept well below zero (Fig. 2). In contrast with the previous example, the interest here is strictly in the prediction of future damage for given wave forces, not in the value of the intercept itself. The resulting linear relationship applies only to values of the independent variable above the threshold of wave effect.

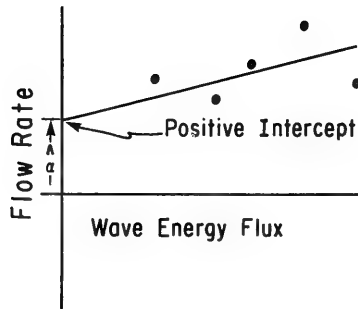


Figure 1. Application of Model I produces an intercept ( $\hat{\alpha}$ ), which may be a useful estimate of a component of longshore flow which is independent of wave conditions and presumably pervades the entire data set.

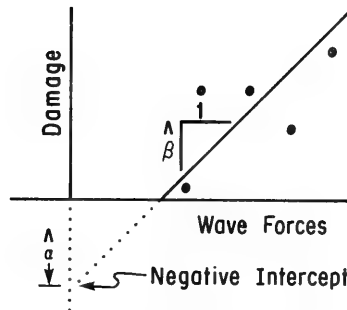


Figure 2. Application of Model I identified a threshold value below which waves cause no damage. A negative intercept is produced, but is of no interest in this particular problem.

Although the negative intercept ( $\hat{\alpha}$ ) is in itself meaningless, Model I is correct because there is no basis for constraining  $\hat{\alpha}$ .

## II. A PROBLEM WITH THE CUSTOMARY APPROACH

There are many cases where the logic of the application dictates the response at a particular value of  $X$ . For example, if the response is some change that is regressed against time then the response must be 0 when  $X = 0$  (Fig. 3). If there is no elapse time, there can be no change. If the linear assumption is valid, the appropriate conceptual mode is

$$Y = \beta X + \epsilon \text{ (Model II)}$$

and the customary predictive equation (based on Model I) is inappropriate and may give poor estimates of  $\beta$  (see Fig. 4). Yet the vast majority of regression programs (e.g., SPSS, IMSL, IBM's 5110 package, and TI-59) do not allow specification of a zero intercept or any constraint through a known point. Statistical texts usually do not cover this topic either. However, formulas for the zero-intercept case are given by Brownlee (1965) and Krumbein (1965).

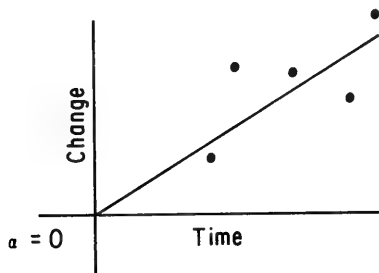


Figure 3. Application of Model II forces a zero-intercept solution.

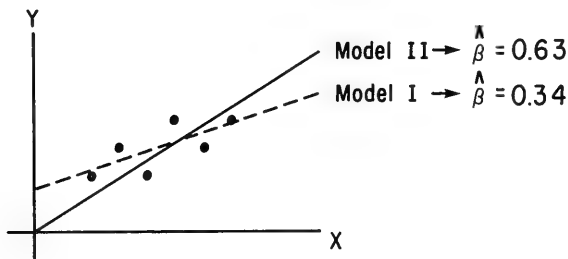


Figure 4. Model II estimates an increase in  $Y$  per unit increase in  $X$  that is nearly twice that predicted using Model I. The physical relationship between  $X$  and  $Y$  dictates which model should be adopted. If Model II is appropriate the solution can be obtained using a simple artifice described in this report to modify results of standard computer programs intended for Model I.

The value of  $Y$  may be known for a single value of  $X$  (not necessarily 0). The best prediction should then be sought from among the limited subset of lines through this point. All these lines will have a larger sum of squares  $(\Sigma[y - \hat{y}]^2)$  than the line that would have been selected by Model I. A simple procedure is described herein for picking from among these restricted candidates the one with the smallest  $\Sigma[y - \hat{y}]^2$ . Thus, regressing through the origin is but one specific case that can be solved by a general model forcing regression at an arbitrary point.

### III. SOLUTION TO THE PROBLEM

This report describes a method for getting the best fit to all data points (in the sense of least squares) while forcing an exact fit at any known point. A simple procedure for forcing regression through the origin was described by Hawkins (1980), who indicated the procedure was not well known. The author of this report knows of no references to the general case of an exact fit to an arbitrary point. However, if a fit can be constrained through the origin, then a simple transform of variables can force the line through any given point. The details of the through-the-origin solution will be explained first.

#### 1. Regression Through the Origin.

For each set of measured dependent and independent variables observed  $(y_i, x_i)$ , also enter, or program, a mirror-image set  $(-y_i, -x_i)$ . Thus, the computer is given an extended data set consisting of  $2n$  data points, only  $n$  of which were observed. By definition of this extended data set, the dependent and all the independent variables each individually sum to zero, forcing a zero intercept:

$$\hat{\alpha} = \bar{y} - \beta \bar{x} \quad \text{by the principle of least squares}$$

$$\hat{\alpha} = 0 \quad \text{because } \Sigma x \text{ and } \Sigma y = 0 \text{ and thus } \bar{x} = \bar{y} = 0 \text{ on the extended data set}$$

Thus a zero-intercept solution is obtained. Is it still the least squares solution for the observed data set? The principle of least squares by definition minimizes the sum of the squares of the deviations of the observed from the predicted values. Because each squared deviation from the observed data set generates an identical squared deviation in the extended data set, the sum of these two positive sequences is minimized over the extended data set only if it is also minimized over both the observed and the mirror-image sets. Thus, the regression coefficient produced in this manner is not only the least squares solution for the artificially extended data set, but for the observed data set as well. By this artifice the proper estimate is obtained for the regression coefficient ( $\beta$ ) with the prediction forced through the origin.

#### 2. Regression Through Any Arbitrary Point (a, b).

If the predicted response  $(\hat{Y})$  must be  $a$  when the independent variables ( $X$ ) are  $b$ , then regress an extended data set  $u$  on  $v$ , where  $u = x - a$  and  $v = y - b$ . If  $(a, b) = (0, 0)$ , then this collapses to the exact situation described above. If  $(a, b) \neq (0, 0)$ , the direct results,  $\hat{u} = \beta v$ , should be unraveled to produce the  $y$  prediction:

$$\hat{u} = \hat{y} - b = \hat{\beta}(x - a)$$

$$\hat{y} = (b - a\hat{\beta}) + \hat{\beta}x$$

NOTE: The proper estimate of the regression coefficient ( $\hat{\beta}$ ) now forces the prediction through the point (a, b) as desired. By using this procedure the correct regression coefficient is obtained by using any familiar computational routines. The second most frequently reported output from regression programs, the correlation coefficient (r), is also the correct, unbiased estimator for Model II.

If additional information is provided by the regression program, then corrections may be necessary before adopting them for the real data set. The estimate of the residual variance will be correct for simple regression (one independent variable) and can be easily adjusted for multiple regression (see Table 1). Any sums of squares, cross products, and F-values produced by the program will be exactly twice the correct values. The standard error of the estimated slope will be too small by a factor of  $\sqrt{2}$ . Therefore, the t-value, for testing the zero slope hypothesis, will be too large by the same factor.

Table 1 indicates the corrections for most of the elements produced by various regression programs. However, employing the described extended data procedures does *not* require consideration of any part of the output beyond that used in the standard unconstrained approach.

#### IV. SELECTING BETWEEN MODELS I AND II

If either the true or mean value (whichever interpretation fits the situation) of the dependent variable (Y) is unknown for all values of the independent variable in the range of concern, then the customary model (I) may be appropriate. However, if the postulated physical relationship between X and Y dictates constraint through any point (a, b) and the relationship is linear from the maximum observed x to x = a, then Model II should be used. To proceed with the customary evaluation of Model I would be equivalent to ignoring what is already known about the relationship between X and Y and, instead, relying totally on the limited information available in the sample data. The objective should be to obtain the best interpretation of the data, which does not override any more firmly established understanding of the situation.

Assuming Model II applies, it may still be useful to evaluate Model I to test in the conventional way (Draper and Smith, 1966) the significance of the estimated nonzero intercept. If this test fails to provide enough evidence to reject the strawman hypothesis ( $H_0: \alpha = 0$ ) then this failure may be cited as additional evidence strictly from the data, substantiating the choice of Model II to estimate  $\beta$ . The results of this formal test of hypothesis should not, however, be relied on as the criterion for selecting Model II. It should serve only as a source of auxiliary information clarifying the extent to which the sample data will support the model choice. The choice should be made on the basis of functional insight and understanding of the relationship between X and Y.

Comparing the correlation coefficients or r-values, produced using the real data and the extended data, is likewise not a valid method for choosing

Table 1. Adjustment of standard elements produced by programs using extended data.

| Parameters and test statistics    | Estimates obtained by constrained regression routine using extended data | Correct estimates for constrained model  |
|-----------------------------------|--|--|
| Regression coefficient: $\beta$   | $\hat{\beta}^1$  | $\hat{\beta}^1$  |
| Correlation coefficient: $\rho$   | $r$  | $r$  |
| Standard error of $\hat{\beta}^1$ | $S_{\hat{\beta}}^1$  | $S_{\hat{\beta}}^1 \left[ \frac{(2n - p - 1)}{(n - p)} \right]^{1/2}$<br>which is exactly $S_{\hat{\beta}}^1 \sqrt{2}$ if $p = 1$<br>and approximately $S_{\hat{\beta}}^1 \sqrt{2}$ if $\frac{n}{p}$ is moderately large |
| Sum of squares                    | $SS_x, SS_y, SS_{xy}$  | $\frac{SS_x}{2}, \frac{SS_y}{2}, \frac{SS_{xy}}{2}$  |
| t-test statistic                  | $t$  | $t/\sqrt{2}$   |
| F-value                           | $F$  | $\frac{F}{2}$  |
| Estimated residual variance       | $S_{y \cdot x}^2$  | $S_{y \cdot x}^2 \left[ \frac{(n - p/2 - 1/2)}{(n - p)} \right]^{1/2}$<br>which is exactly $S_{y \cdot x}^2$ of $p = 1$ and<br>approximately $S_{y \cdot x}^2$ if $\frac{n}{p}$ is moderately large                      |

<sup>1</sup>Hatted values are estimates of the true unhatted parameters (e.g.,  $\beta$  is an unknown parameter in a conceptual model and  $\hat{\beta}$  is its estimate based on evaluation of some measurements),  $p$  is the number of independent variables that were measured. Additional information on interpreting the standard elements of regression is available in Draper and Smith (1966).



between Models I and II. The value of  $r^2$  using Model I (observed data only) is often referred to as the reduction in variance of the estimator made possible by using the apparent association between X and Y. A value of  $r^2 = 0$  indicates that knowledge of the X-values makes no improvement in the prediction of Y and using the mean value of the y's as the estimator would not increase the sum of the squares of the deviations. At the other extreme if  $r^2 = 1$ , all sample points lie on a sloping straight line implying a strong predictive value. Similarly with Model II, higher  $r^2$  values indicate improved fit of the data; but comparing  $r^2$  values between Models I and II does not reveal which is correct or even preferable. There is a slight conceptual and a substantial computational difference between the  $r^2$  values for the two models. The two values should not be compared; both indicate the relative fit of various data to their own particular model. Either value can be used to measure "goodness of fit" in particular applications; or even to indicate the usefulness of several versions of the particular model chosen. For example comparison of r-values would indicate whether taking logs of the measurements, or raising them to a given power prior to regression, improved the fit. But comparison of the r-value would not be a valid basis for choosing between Models I and II.

## V. EXAMPLES

The following problems illustrate a frequent need to constrain the regression line in coastal engineering applications. The problems also illustrate the usefulness of  $r^2$  to rank different predictors in terms of how well they fit data. Before initially applying the described method to an actual problem, it may be helpful to reanalyze one of the small data sets used in these examples and compare the results with those published in this report.

### \*\*\*\*\* EXAMPLE PROBLEM 1 \*\*\*\*\*

Consider the requirement to simulate a long-term history of wave-induced longshore currents for a particular coastal site. Assume hindcasted wave data are available, but that current measurements were not made over the period of interest. According to the Shore Protection Manual (U.S. Army, Corps of Engineers, Coastal Engineering Research Center, 1977), the longshore current ( $v$ ) can be calculated as a function of the beach slope ( $m$ ), the gravitational acceleration ( $g$ ), and the angle and height of breaking waves ( $\alpha_b$ ,  $H_b$ , respectively).

$$v = 20.7 \, m \, (gH_b)^{1/2} \sin 2\alpha_b \quad (1)$$

The coefficient of proportionality (20.7) is based on typical mixing and frictional factors for the surf zone. Empirical formulas, like equation (1) can be adjusted by regression analysis of test data from the specific site of intended application. This will customize the formula to fit site-sensitive conditions. The longshore velocity also varies laterally within the surf zone. The problem of estimating the spatial structure of flow across the surf zone may be avoided by obtaining current measurements at the exact point where the long-term flow must be reconstructed, then regressing the test measurements against simultaneously determined breaker conditions. Steps in such an analysis are given below. Only a few data points are used in the example to encourage the reader to go through the computations and check the results. The data are taken from a frequently referenced field study done at Nags Head, North Carolina (Galvin and Savage, 1966).

GIVEN: Longshore current velocities ( $v$ ), breaker heights ( $H_b$ ), breaker angles ( $\alpha_b$ ), and the beach slope ( $m$ ) determined onsite during a short field evaluation (see Table 2).

Table 2. Field calibration data (from Galvin and Savage, 1966).

| Obsn. | $H_b$<br>(ft) | $m$   | $v$<br>(ft/s) |
|-------|---------------|-------|---------------|
| 1     | 2             | 0.03  | 2.42          |
| 2     | 3.2           | 0.026 | 4.33          |
| 3     | 1.8           | 0.029 | 1.96          |
| 4     | 8             | 0.026 | 1.26          |

REQUIRED: An equation that will predict wave-induced longshore currents for the test site.

ANALYSIS: Because the linearity expressed in equation (1) has a firm theoretical basis in the concept of radiation stress (Longuet-Higgins, 1970), and because according to this concept,  $v = 0$  whenever  $H_b = 0$  or  $\alpha_b = 0$ , the prediction line must pass through the origin (0, 0). So Model II must be used.

Let

$$Y = v$$

and

$$X = m(gH_b)^{1/2} \sin 2\alpha_b$$

Regress  $Y$  on  $X$  to determine the best estimate of the coefficient of proportionality between  $X$  and  $Y$ .

CORRECT RESULTS:

|                                 |                   |        |
|---------------------------------|-------------------|--------|
| Regression coefficient          | $\hat{\beta}$     | = 17   |
| Correlation coefficient         | $r$               | = 0.91 |
| Standard error of $\hat{\beta}$ | $S\hat{\beta}$    | = 4.6  |
| Test statistic for $\beta$      | $t$               | = 3.7  |
| Estimated residual variance     | $S^2_{y \cdot x}$ | = 1.8  |

CONCLUSION: The version of the Longuet-Higgins type equation that best fits this problem site (based on available current data) is:

$$v = 17 m (gH_b)^{1/2} \sin 2\alpha_b$$

NOTE: Fitting the equation to the data in this example produces results closer to those obtained with larger data sets (eq. 1) if the line is forced through the origin rather than being fit strictly to the data without this constraint (see Fig. 5).

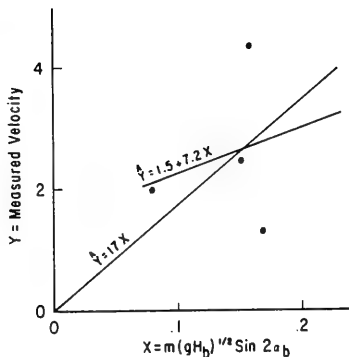


Figure 5. Real test data for example problem 1. Compare the correct fit through the origin with the customary fit.

\*\*\*\*\* EXAMPLE PROBLEM 2 \*\*\*\*\*

At least 10 equations relating the velocity of longshore currents to wave characteristics have appeared in the literature. Presumably more will appear as knowledge increases or theory is adapted to specific wave or bathymetric conditions (i.e., specialized for breaker type or bar dimensions). A recent article (Komar, 1979) questions the value of including a measure of beach slope in the general prediction equation and claims better results for

$$v = 0.585(gH_b)^{1/2} \sin 2\alpha_b$$

GIVEN: The same situation and data as in example problem 1.

REQUIRED: Determine the best fit version of the type

$$v = (gH_b)^{1/2} \sin 2\alpha_b$$

and compare the results with those obtained in example problem 1 to see if the beach slope is indeed of any value at this particular site.

ANALYSIS: For the same reasons stated in example problem 1, regression should require the prediction line to pass through the point (0, 0).

Let

$$Y = v$$

$$X = (gH_b)^{1/2} \sin 2\alpha_b$$

and regress Y on X using Model II with its extended data set (Fig. 6).

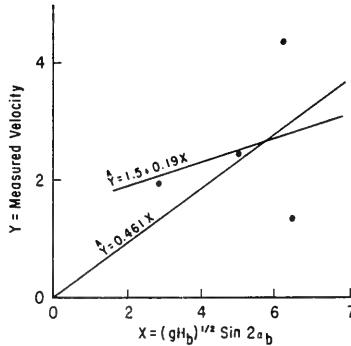


Figure 6. Real test data for example problem 2 and fitted equations. Compare the correct fit through the origin with the customary fit.

CORRECT RESULTS:

|                                  |                   |        |
|----------------------------------|-------------------|--------|
| Regression coefficient           | $\hat{\beta}$     | = 0.46 |
| Correlation coefficient          | $r$               | = 0.90 |
| Standard error of $\hat{\beta}$  | $S_{\hat{\beta}}$ | = 0.13 |
| Test statistic for $\hat{\beta}$ | $t$               | = 3.6  |
| Estimated residual variance      | $S_{y \cdot x}^2$ | = 1.8  |

CONCLUSION: The best predictor of the Komar type is:

$$v = 0.46(gH_b)^{1/2} \sin 2\alpha_b$$

It would be surprising to find a clear indication of whether beach slope should be included in the predictor for longshore currents by evaluating such a limited data set as chosen here to encourage reader computation. Indeed a comparison of Tables 3 and 4 reveals no significant differences between the correlation coefficients or any other test statistics. However, significant differences would be expected if a large reliable data set covering a wider range of conditions were compared by the methods illustrated in this report.

Table 3. Extended data set No. 1.

| Obsn. | X<br>(ft/s) | Y<br>(ft/s) |
|-------|-------------|-------------|
| 1     | 0.152       | 2.42        |
|       | -0.152      | -2.42       |
| 2     | 0.162       | 4.33        |
|       | -0.162      | -4.33       |
| 3     | 0.0827      | 1.96        |
|       | -0.0827     | -1.96       |
| 4     | 0.170       | 1.27        |
|       | -0.170      | -1.27       |

Table 4. Extended data set No. 2.

| Obsn. | X<br>(ft/s) | Y<br>(ft/s) |
|-------|-------------|-------------|
| 1     | 5.05        | 2.42        |
|       | -5.05       | -2.42       |
| 2     | 6.25        | 4.33        |
|       | -6.25       | -4.33       |
| 3     | 2.85        | 1.96        |
|       | -2.85       | -1.96       |
| 4     | 6.53        | 1.27        |
|       | -6.53       | -1.27       |

#### LITERATURE CITED

- BROWNLEE, K.A., *Statistical Theory and Methodology in Science and Engineering*, 2d ed., John Wiley & Sons, Inc., New York, 1965.
- DRAPER, N.R., and SMITH, H., *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, 1966.
- GALVIN, C.J., Jr., and SAVAGE, R.P., "Longshore Currents at Nags Head, North Carolina," Bulletin 11, U.S. Army, Corps of Engineers, Coastal Engineering Research Center, Washington, D.C., 1966, pp. 11-29.
- HAWKINS, D.M., "A Note on Fitting a Regression Without an Intercept Term," *The American Statistician*, Vol. 34, Nov. 1980, p. 233.
- KOMAR, P.D., "Beach-Slope Dependence of Longshore Currents," *Journal of Waterways, Port, Coastal and Ocean Divisions*, Vol. 105, Nov. 1979.
- KRUMBEIN, W.C., and GRAYBILL, F.A., *An Introduction to Statistical Models in Geology*, McGraw-Hill Book Co., New York, 1965.
- LONGUET-HIGGINS, M.S., "Longshore Currents Generated by Obliquely Incident Seawaves," Parts I and II, *Journal of Geophysical Research*, Vol. 75, No. 33, Nov. 1970.
- U.S. ARMY, CORPS OF ENGINEERS, COASTAL ENGINEERING RESEARCH CENTER, *Shore Protection Manual*, 3d ed., Vols. I, II, and III, Stock No. 008-022-00113-1, U.S. Government Printing Office, Washington, D.C., 1977, 1,262 pp.

|   |   |
|---|---|
| <p>Hands, Edward B.<br/>Forcing regression through a given point using any familiar computational routine / by Edward B. Hands.--Fort Belvoir, Va. : U.S. Army, Corps of Engineers, Coastal Engineering Research Center ; Springfield, Va. : available from NTIS, 1983.<br/>[20] p. : ill. ; 28 cm.--(Technical paper / Coastal Engineering Research Center ; no. 83-1)<br/>Cover title.<br/>"March 1983."<br/>Report describes a simple method for obtaining the prediction equation best fit to all data points (in the least squares sense) while forcing an exact fit at any known point. Examples are given. All necessary changes to the program results are accomplished without modifying customary computational routines.<br/>1. Coastal engineering. 2. Data analysis. 3. Prediction equations. 4. Regression. I. Title. II. Coastal Engineering Research Center (U.S.). III. Series: Technical paper (Coastal Engineering Research Center (U.S.)); no. 83-1.<br/>TC203 .U581tp no. 83-1 627</p> | <p>Hands, Edward B.<br/>Forcing regression through a given point using any familiar computational routine / by Edward B. Hands.--Fort Belvoir, Va. : U.S. Army, Corps of Engineers, Coastal Engineering Research Center ; Springfield, Va. : available from NTIS, 1983.<br/>[20] p. : ill. ; 28 cm.--(Technical paper / Coastal Engineering Research Center ; no. 83-1)<br/>Cover title.<br/>"March 1983."<br/>Report describes a simple method for obtaining the prediction equation best fit to all data points (in the least squares sense) while forcing an exact fit at any known point. Examples are given. All necessary changes to the program results are accomplished without modifying customary computational routines.<br/>1. Coastal engineering. 2. Data analysis. 3. Prediction equations. 4. Regression. I. Title. II. Coastal Engineering Research Center (U.S.). III. Series: Technical paper (Coastal Engineering Research Center (U.S.)); no. 83-1.<br/>TC203 .U581tp no. 83-1 627</p> |
| <p>Hands, Edward B.<br/>Forcing regression through a given point using any familiar computational routine / by Edward B. Hands.--Fort Belvoir, Va. : U.S. Army, Corps of Engineers, Coastal Engineering Research Center ; Springfield, Va. : available from NTIS, 1983.<br/>[20] p. : ill. ; 28 cm.--(Technical paper / Coastal Engineering Research Center ; no. 83-1)<br/>Cover title.<br/>"March 1983."<br/>Report describes a simple method for obtaining the prediction equation best fit to all data points (in the least squares sense) while forcing an exact fit at any known point. Examples are given. All necessary changes to the program results are accomplished without modifying customary computational routines.<br/>1. Coastal engineering. 2. Data analysis. 3. Prediction equations. 4. Regression. I. Title. II. Coastal Engineering Research Center (U.S.). III. Series: Technical paper (Coastal Engineering Research Center (U.S.)); no. 83-1.<br/>TC203 .U581tp no. 83-1 627</p> | <p>Hands, Edward B.<br/>Forcing regression through a given point using any familiar computational routine / by Edward B. Hands.--Fort Belvoir, Va. : U.S. Army, Corps of Engineers, Coastal Engineering Research Center ; Springfield, Va. : available from NTIS, 1983.<br/>[20] p. : ill. ; 28 cm.--(Technical paper / Coastal Engineering Research Center ; no. 83-1)<br/>Cover title.<br/>"March 1983."<br/>Report describes a simple method for obtaining the prediction equation best fit to all data points (in the least squares sense) while forcing an exact fit at any known point. Examples are given. All necessary changes to the program results are accomplished without modifying customary computational routines.<br/>1. Coastal engineering. 2. Data analysis. 3. Prediction equations. 4. Regression. I. Title. II. Coastal Engineering Research Center (U.S.). III. Series: Technical paper (Coastal Engineering Research Center (U.S.)); no. 83-1.<br/>TC203 .U581tp no. 83-1 627</p> |







