Designation: ~~E3080 – 16~~ E3080 – 17

An American National Standard

Standard Practice for Regression Analysis[1]

This standard is issued under the fixed designation E3080; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This practice covers regression analysis methodology for estimating, evaluating, and using the simple linear regression model to define the statistical relationship between two numerical variables.

1.2 The system of units for this practice is not specified. Dimensional quantities in the practice are presented only as illustrations of calculation methods. The examples are not binding on products or test methods treated.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate ~~safety~~ safety, health, and ~~health~~environmental practices and determine the applicability of regulatory limitations prior to use.*

1.4 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E178 Practice for Dealing With Outlying Observations

E456 Terminology Relating to Quality and Statistics

~~E2282 Guide for Defining the Test Result of a Test Method~~

E2586 Practice for Calculating and Using Basic Statistics

3. Terminology

3.1 *Definitions*—Unless otherwise noted, terms relating to quality and statistics are as defined in Terminology E456.

~~3.1.1 *characteristic, n*—a property of items in a sample or population which, when measured, counted, or otherwise observed, helps to distinguish among the items. **E2282**~~

3.1.1 *coefficient of determination, r^2, n*—square of the correlation coefficient.

~~3.1.3 *confidence interval, n*—an interval estimate [L, U] with the statistics L and U as limits for the parameter θ and with confidence level 1 – α, where Pr(L ≤ θ ≤ U) ≥ 1 – α. **E2586**~~

~~3.1.3.1 *Discussion*—~~

~~The confidence level, 1 – α, reflects the proportion of cases that the confidence interval [L, U] would contain or cover the true parameter value in a series of repeated random samples under identical conditions. Once L and U are given values, the resulting confidence interval either does or does not contain it. In this sense "confidence" applies not to the particular interval but only to the long run proportion of cases when repeating the procedure many times.~~

~~3.1.4 *confidence level, n*—the value, 1 – α, of the probability associated with a confidence interval, often expressed as a percentage. **E2586**~~

~~3.1.4.1 *Discussion*—~~

[1] This practice is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.10 on Sampling / Statistics.
Current edition approved ~~Nov. 1, 2016~~Nov. 1, 2017. Published ~~November 2016~~January 2018. Originally approved in 2019. Last previous edition approved in 2016 as E3080 – 16. DOI: ~~10.1520/E3080-16.~~10.1520/E3080-17.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

~~α is generally a small number. Confidence level is often 95 % or 99 %.~~

~~3.1.5 *correlation coefficient, n—for a population,ρ,* a dimensionless measure of association between two variables X and Y, equal to the covariance divided by the product of~~ ~~σ_X times σ_Y.~~

~~3.1.6 *correlation coefficient, n—for a sample, r*, the estimate of the parameter *ρ* from the data.~~

~~3.1.7 *covariance, n—of a population*, cov(X, Y), for two variables, X and Y, the expected value of~~ ~~$(X - \mu_X)(Y - \mu_Y)$.~~

~~3.1.8 *covariance, n—of a sample*; the estimate of the parameter cov(X,Y) from the data.~~

~~3.1.9 *dependent variable, n*—a variable to be predicted using an equation.~~

3.1.2 *degrees of freedom, n*—the number of independent data points minus the number of parameters that have to be estimated before calculating the variance. **E2586**

~~3.1.11 *deviation, d, n*—the difference of an observed value from its mean.~~

~~3.1.12 *estimate, n*—sample statistic used to approximate a population parameter. **E2586**~~

~~3.1.13 *independent variable, n*—a variable used to predict another using an equation.~~

~~3.1.14 *mean, n—of a population*, μ, average or expected value of a characteristic in a population – *of a sample*,~~ ~~$\bar{X}$~~~~, sum of the observed values in the sample divided by the sample size. **E2586**~~

~~3.1.15 *parameter, n*—see *population parameter*. **E2586**~~

~~3.1.16 *population, n*—the totality of items or units of material under consideration. **E2586**~~

~~3.1.17 *population parameter, n*—summary measure of the values of some characteristic of a population. **E2586**~~

~~3.1.18 *prediction interval, n*—an interval for a future value or set of values, constructed from a current set of data, in a way that has a specified probability for the inclusion of the future value. **E2586**~~

~~3.1.19 *regression, n*—the process of estimating parameter(s) of an equation using a set of data.~~

3.1.3 *residual, n*—observed value minus fitted value, when a model is used.

~~3.1.21 *statistic, n*—see *sample statistic*. **E2586**~~

3.1.4 ~~*quantile,*~~ *predictor variable, X, n*—~~value such that a fraction~~a variable ~~f~~used ~~of the sample or population is less than or equal to that value.~~to predict a response variable using a regression model.~~ **E2586**~~

3.1.4.1 *Discussion—*

Also called an *independent* or *explanatory* variable.

3.1.5 ~~*sample,*~~ *regression analysis, n*—a ~~group of observations or test results, taken from a larger collection of observations or test results, which serves to provide information that may be used as a basis for making a decision concerning the larger collection.~~statistical procedure used to characterize the association between two numerical variables for prediction of the response variable from the predictor variable.~~ **E2586**~~

~~3.1.24 *sample size, n, n*—number of observed values in the sample. **E2586**~~

3.1.6 ~~*sample statistic,*~~ *response variable, Y, n*—~~summary measure of the observed values of a sample.~~a variable predicted from a regression model.~~ **E2586**~~

3.1.6.1 *Discussion—*

Also called a *dependent* variable.

~~3.1.26 *standard error*—standard deviation of the population of values of a sample statistic in repeated sampling, or an estimate of it. **E2586**~~

~~3.1.26.1 *Discussion*—~~

~~If the standard error of a statistic is estimated, it will itself be a statistic with some variance that depends on the sample size.~~

3.1.7 ~~*standard deviation*—~~*sample correlation coefficient, r, n*—~~of~~ a ~~population,~~dimensionless ~~σ, the square root of the average or expected value of the squared deviation of a variable from its mean; —~~measure of association between two variables estimated from ~~of a sample, s,~~the ~~square root of the sum of the squared deviations of the observed values in the sample from their mean divided by the sample size minus 1.~~data.~~ **E2586**~~

3.1.8 ~~*variance, σ*~~*sample covariance,*~~^{2},~~ s_{xy}~~^{2}~~, *n*—~~square~~ an estimate of the ~~standard deviation of the population or sample.~~association of the response variable and predictor variable calculated from the data.~~ **E2586**~~

~~3.1.28.1 *Discussion—*~~

~~For a finite population, σ^2 is calculated as the sum of squared deviations of values from the mean, divided by *n*. For a continuous population, σ^2 is calculated by integrating $(x - \mu)^2$ with respect to the density function. For a sample, s^2 is calculated as the sum of the squared deviations of observed values from their average divided by one less than the sample size.~~

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *intercept, n—of a regression model*, β_0, the value of the response variable when the predictor variable is zero.

3.2.2 *regression model parameter, n*—a descriptive constant defining a regression model that is to be estimated.

3.2.3 *residual standard deviation, n—of a regression model*, σ, the square root of the residual variance.

3.2.4 *residual variance, n—of a regression model*, σ^2, the variance of the residuals (see *residual*).

3.2.5 *slope, n—of a regression model*, β_1, the incremental change in the response variable due to a unit change in the predictor variable.

3.3 *Symbols:*

b_0	=	intercept estimate (5.2.2)
b_1	=	slope estimate (5.2.2)
β_0	=	intercept parameter in model (5.1.2)
β_1	=	slope parameter in model (5.1.2)
E	=	general point estimate of a parameter (5.4.2)
e_i	=	residual for data point *i* (5.2.5)
ε	=	residual parameter in model (5.1.3)
F	=	*F* statistic (X1.3.2)
h	=	index for any value in data range (5.4.5)
i	=	index for a data point (5.2.1)
n	=	number of data points (5.2.1)
r	=	sample correlation coefficient (5.3.2.1)
r^2	=	coefficient of determination (5.3.2.2)
$S(b_0,b_1)$	=	sum of squared deviations of Y_i to the regression line (X1.1.2)
s_{b1}	=	standard error of slope estimate (5.4.3)
s_{b0}	=	standard error of intercept estimate (5.4.4)
s_E	=	general standard error of a point estimate (5.4.2)
σ	=	residual standard deviation (5.1.3)
s	=	estimate of σ (5.2.6)
σ^2	=	residual variance (5.1.3)
s^2	=	estimate of σ^2 (5.2.6)
s_X^2	=	variance of *X* data (X1.2.1)
s_Y^2	=	variance of *Y* data (X1.2.1)
S_{XX}	=	sum of squares of deviations of *X* data from average (5.2.3)
S_{XY}	=	sum of cross products of *X* and *Y* from their averages (5.2.3)
s_{XY}	=	sample covariance of *X* and *Y* (X1.2.1)
$s_{\hat{Y}h}$	=	standard error of $\hat{Y}_h$ (5.4.5)
$s_{\hat{Y}h(ind)}$	=	standard error of future individual *Y* value (5.4.6)
S_{YY}	=	sum of squares of deviations of *Y* data from average (5.2.3)
t	=	Student's *t* distribution (5.4.2)
X	=	predictor variable (5.1.1)
$\bar{X}$	=	average of *X* data (5.2.3)
X_h	=	general value of *X* in its range (5.4.5)
X_i	=	value of *X* for data point *i* (5.2.1)
Y	=	response variable (5.1.1)
$\bar{Y}$	=	average of *Y* data (5.2.3)
$\hat{Y}_{h(ind)}$	=	predicted future individual *Y* for a value X_h (5.4.6)
Y_i	=	value of *Y* for data point *i* (5.2.1)
$\hat{Y}_h$	=	predicted value of *Y* for any value X_h (5.4.5)
$\hat{Y}_i$	=	predicted value of *Y* for data point *i* (5.2.4)

3.4 *Acronyms:*

3.4.1 *ANOVA, n*—Analysis of Variance

3.4.2 *df, n*—Degrees of Freedom

3.4.3 *LOF, n*—Lack of Fit

3.4.4 *MS, n*—Mean Square

3.4.5 *MSE, n*—Mean Square Error

3.4.6 *MSR, n*—Mean Square Regression

3.4.7 *MST, n*—Mean Square Total

3.4.8 *PE, n*—Pure Error

3.4.9 *SS, n*—Sum of Squares

3.4.10 *SSE, n*—Sum of Squares Error

3.4.11 *SSR, n*—Sum of Squares Regression

3.4.12 *SST, n*—Sum of Squares Total

4. Significance and Use

4.1 Regression analysis is a statistical procedure that studies the ~~relations~~ statistical relationships between two or more ~~numerical~~ variables ~~and~~Ref. ~~**utilizes**~~**(1, 2**~~**existing**~~**).**[3] ~~data to determine a model equation for prediction of one variable from another. In this standard, a simple linear regression model, that is, a straight line relationship between two variables, is considered~~In general, one of these variables is designated as a response variable and the rest of the variables are designated as predictor variables. Then the objective of the model is to predict ~~(~~the~~1, 2).~~response from the predictor variables.

4.1.1 This standard considers a numerical response variable and only a single numerical predictor variable.

4.1.2 The regression model consists of: (*1*) a mathematical function that relates the mean values of the response variable distribution to fixed values of the predictor variable, and (*2*) a description of statistical distribution that describes the variability in the response variable at fixed levels of the predictor variable.

4.1.3 The regression procedure utilizes experimental or observational data to estimate the parameters defining a regression model and their precision. Diagnostic procedures are utilized to assess the resulting model fit and can suggest other models for improved prediction performance.

4.1.4 The regression model can be useful for developing process knowledge through description of the variable relationship, in making predictions of future values, and in developing control methods for the process generating values of the variables.

4.2 Section 5 in this standard deals with the simple linear regression model using a straight line mathematical relationship between the two variables where variability of the response variable over the range of values of the predictor variable is described by a normal distribution with constant variance. Appendix X1 provides supplemental information.

5. ~~Straight Line Regression and Correlation~~Simple Linear Regression Analysis

5.1 ~~*Two Variables*—~~*Simple Linear Regression Model:* ~~The data set includes two variables, *X* and *Y*, measured over a collection of sampling units, experimental units or other type of observational units. Each variable occurs the same number of times and the two variables are paired one to one. Data of this type constitute a set of *n* ordered pairs of the form (x_i, y_i), where the index variable (*i*) runs from 1 through *n*.~~

5.1.1 Select the response variable *Y* ~~is always to be~~and the predictor variable ~~*treated*~~*X*. ~~as a random variable.~~ The predictor *X* ~~may be either a random variable sampled from a population with an error that is negligible compared to the error of~~ is assumed to have known values with little or no measurement error. The response ~~*Y*,~~*Y* ~~or values chosen as in the design of an experiment where the values represent levels that are fixed and without error. We refer to~~ has a distribution of values for a given *X* ~~as the independent variable and~~ value, and this distribution is defined for all ~~*Y*~~*X* ~~as the dependent variable.~~values in a given range.

5.1.2 The ~~practitioner typically wants to see if a relationship exists between~~ regression function for ~~X and Y. In theory, many different types of relationships can occur between X and Y. The most common is a simple linear relationship of the form~~the straight ~~Y~~line ~~= α~~relationship is $Y=\beta_0+\beta_1X$~~+ β~~. The ~~X~~two ~~+ ε, where α~~ parameters for the function are the intercept β_0 and ~~β are model~~the slope β_1 ~~coefficients and ε is a random error term representing variation in the observed~~ . The intercept is the value of *Y* ~~at given~~ when ~~*X*,~~*X* ~~and is assumed to have a mean of 0 and some unknown standard deviation σ. A statistical analysis that seeks to determine a linear relationship between a dependent variable,~~ = 0, but ~~Y, and a single independent variable, X, is called simple linear regression. In this type of analysis it is assumed that the error structure is normally distributed with mean 0 and some unknown variance σ~~this parameter may not be of practical interest when the ~~^{2} throughout the~~ range of *X* ~~and~~is ~~Y.~~far ~~Further, the errors are uncorrelated with each other. This will be assumed~~removed from zero. The slope is the amount of incremental change in ~~*throughout*~~*Y* ~~the remainder of this section.~~units for a unit change in *X*.

5.1.3 The ~~regression problem is~~statistical distribution for ~~*to*~~*Y* ~~determine estimates of the coefficients α and β that "best" fit the data~~is assumed to be a normal (Gaussian) distribution having a mean of $\beta_0+\beta_1X$ ~~and allow estimation of σ. An additional measure of association, the correlation coefficient, ρ, can also be estimated from this type of data which indicates the strength of the linear relationship between~~ with a standard deviation σ. ~~X and Y.~~The ~~sample correlation coefficient,~~simple linear regression ~~r,~~model is

[3] The boldface numbers in parentheses refer to a list of references at the end of this standard.

~~the estimate of~~then stated as $Y=\beta_0+\beta_1X+\varepsilon$~~ρ. The square of the correlation coefficient,~~, where ε is a random error ~~r~~that^2~~, is called the coefficient of determination and has additional meaning for the~~ is normally distributed with mean zero and standard deviation σ (variance σ^2~~linear relationship between~~).~~X and Y.~~

5.1.4 ~~When a suitable model is found, it may~~An example of a linear regression model is depicted in Fig. 1 ~~be used to estimate the mean response at a given value~~ over a range of *X* ~~or~~ from 0 to ~~predict~~40 ~~*the*~~*X* ~~range of future~~units. Normal distributions of response *Y* ~~values from a given~~with σ = 1.3 ~~*X*.~~*Y* units are depicted at *X* = 10, 20, and 30 *X* units.

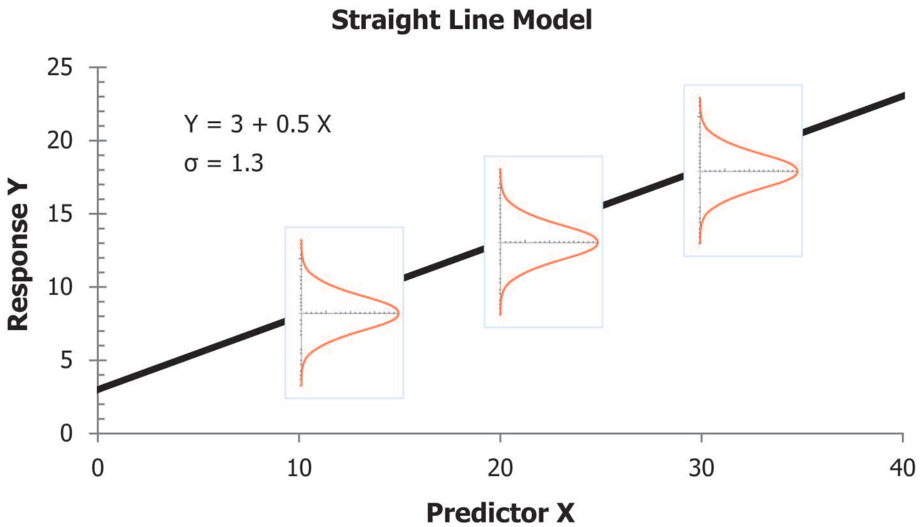

FIG. 1 Graphical Depiction of a Straight Line Regression Model

5.2 ~~*Method of Least Squares—*~~*Estimating Regression Model Parameters:* ~~The methodology considered in this standard and used to estimate the model parameters α and β is called the method of least squares. The form of the best fitting line will be denoted as $Y = a + bX$, where a and b are the estimates of α and β respectively. The ith observed values of X and Y are denoted as x_i and y_i. The estimate of Y at $X = x_i$ is written $\hat{y}_i = a + bx_i$. The "hat" notation over the y_i variable denotes that this is the estimated mean or predicted value of Y for a given x.~~

~~5.2.1 The least squares best fitting line is one that minimizes the sum of the squared deviations from the line to the observed y_i values. Note that these are vertical distances. Analytically, this sum of squared deviations is of the form:~~

$$S(\mathrm{a},\ b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{1}$$

5.2.1 The ~~sum~~ model parameters β_0, and β_1~~of squares,~~ , are estimated from a sample of data consisting of n pairs of values designated as (~~S,~~X_i, ~~is~~Y_i ~~written as a function~~), with the sample number i ranging from 1 through n. The data can arise in two different ways. Observational data consists of ~~a~~X and ~~b.~~Y ~~Minimizing this function involves taking partial derivatives~~ values measured on a set of ~~S~~n ~~with respect to~~ random samples. Experimental data consists of ~~a~~Y ~~and~~ values measured on n experimental units with ~~b.~~X ~~This will result in two linear equations that are then solved simultaneously for~~ values set at fixed values. In both cases the ~~a~~Y ~~and~~values ~~b.~~may ~~The resulting solutions are functions of the (~~have measurement error, but the ~~x~~X~~$_i$;~~ ~~y~~values~~$_i$) paired data.~~ are assumed known with negligible measurement error.

5.2.2 The regression line parameters β_0, and β_1 are estimated by the method of least squares, which finds their corresponding estimates b_0 and b_1 that minimize the sum of the squares of the vertical distances between the Y_i values and their respective line values at X_i. (For a further discussion of the least squares method, see X1.1.2.)

5.2.3 ~~Several algebraically equivalent formulas for the least squares solutions are found in the literature. The following describes one convenient form of the solution. First define sums of squares~~ Calculate the following statistics from the ~~S~~X~~$_{XX}$~~ and ~~S~~Y~~$_{YY}$ and the sum of cross products~~ values in the data set.~~S_{XY} as follows:~~

$$S_{XX} = (n - 1)s_x^2 = \sum_{i=1}^{n}(x_1 - \bar{x})^2 \tag{2}$$

$$S_{YY} = (n - 1)s_y^2 = \sum_{i=1}^{n}(y_1 - \bar{y})^2 \tag{3}$$

$$S_{XY} = \sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y}) = \sum_{i=1}^{n}(x_1 - \bar{x})y_1 \tag{4}$$

~~Note that in Eq 2 and Eq 3, s_x and s_y are the ordinary sample standard deviations of the X and Y data respectively. The last expression in Eq 4 follows from the middle expression because $\sum_{i=1}^{n}(x_1 - \bar{x})\bar{y} = 0$.~~

5.2.3.1 Calculate the averages of X and Y:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{2}$$

~~From the least squares solution, the slope estimate is calculated as:~~

$$b = \frac{\sum_{i-1}^{n}(x_i - \bar{x})y_i}{\sum_{i-1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}} \tag{5}$$

5.2.3.2 Calculate the sums of squared deviations S_{XX} and S_{YY} of X and Y from their respective averages and the sum of cross products S_{XY} of the X and Y deviations from their averages:

$$S_{XX} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{3}$$

$$S_{YY} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \tag{4}$$

$$S_{XY} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{5}$$

S_{XX} is a known fixed constant. S_{YY} and S_{XY} are random variables.

~~Once b is determined, the intercept term is calculated from:~~

~~$a = \bar{y} - b\bar{x}$ (6)~~

5.2.3.3 The least squares solution gives the parameter estimates:

$$b_1 = S_{XY}/S_{XX} \tag{6}$$

$$b_0 = \bar{Y} - b_1\bar{X} \tag{7}$$

[S_{YY} is not used here but will be used in subsequent sections.]

5.2.4 The *fitted values*$\hat{Y}_i$ for each data point Y_i are calculated from the estimated regression function as:

$$\hat{Y}_i = b_0 + b_1 X_i \tag{8}$$

5.2.5 The *residual*e_i is the difference between the response data point Y_i and its fitted value $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i \tag{9}$$

Residuals are graphically the vertical distances on the scatter plot between the response data points Y_i and the estimated regression line.

5.2.6 The estimates s^2 of the variance σ^2 and s of the standard deviation σ of the Y distribution are calculated as the sum of the squared residuals divided by their degrees of freedom:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-2)} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 / (n-2) \tag{10}$$

$$s = \sqrt{s^2} \tag{11}$$

These estimates have $n - 2$ degrees of freedom because of prior estimation of two parameters, the slope and intercept of the line, which removed two degrees of freedom from the data set of n data points prior to calculation of the residuals.

5.2.7 *Regression Analysis Procedure with Example*—The steps in the regression analysis procedure for the simple linear model, that are illustrated in the example below, are as follows:

(1) Choose the predictor variable X and response variable Y.

(2) Obtain data pairs of X and Y from available data or by conducting an experiment.

(3) Evaluate the distribution of the predictor variable and the XY relationship using plots.

(4) If the model is supported by the data plots, estimate the model parameters from the data.

(5) Evaluate the fitted model against the model assumptions.

(6) Use the regression model for future prediction of Y from X.

5.2.7.1 A data set from Duncan, Ref. **(3)** lists measurements of shear strength (inch-pounds) and weld diameter (mils) measured on 10 random test specimens, so this is an observational data set with $n = 10$ pairs. Regression analysis will be used to investigate the relationship between weld diameter and shear strength, with the objective of predicting shear strength Y from weld diameter X. The weld diameters are considered to be measured with small error. The data are listed in Table 1.

5.2.7.2 A dot plot of the X data is shown as Fig. 2, and the plot indicated that the data was spread out fairly evenly across the range of 190–270 mils and some of the parts had the same diameters.

5.2.7.3 A scatter plot of the data is recommended as a first or concurrent step for a visual look at the relationship, and most computer packages have this as an option. This is a plot of Y (on the vertical axis) versus X (on the horizontal axis) for each data pair. If a straight line relationship exists, the cluster of points will appear to be elongated in a particular direction along a straight line, and the plot will visually reveal any curvature or any other deviations from a straight line relationship, as well as any outlying data points. The estimated regression line can also be included on the plot to give a visual impression of the fit of the model to the data.

The scatter plot for this example is shown in Fig. 3. The shear strength appears to be increasing in a linear fashion with weld diameter. There is some scatter but no apparent outlying data points.

5.2.7.4 The calculations, with equation numbers for each calculation, are shown in Table 1. The averages of X and Y are respectively 233.9 mils and 975.0 inch-pounds. The deviations of X and Y from their averages are listed for each observation, and these are used to calculate values of the statistics S_{XX}, S_{YY}, and S_{XY}. The least squares estimates of the slope and intercept are calculated, resulting in the estimated model equation giving fitted values $\hat{Y}_i$=-569.47+6.898 X_i, and these values are listed for each observation. The residuals e_i=Y_i=$\hat{Y}_i$ are also listed for each observation. Estimates of the variance and standard deviation of the Y distribution are calculated from squares of the residuals. The estimated standard deviation is 99.90 inch-pounds.

5.2.7.5 The least squares straight line is depicted with the scatter plot in Fig. 3, and indicates that a straight line model appears to give a reasonable fit to this data set. Some additional comments from Table 1 are:

TABLE 1 ~~Weld Diameter (x) and Shear Strength (y)~~

i	~~x_i~~	~~y_i~~	~~$d_i = x_i - y_i$~~	~~$x_i - \bar{x}$~~	~~$(x_i - \bar{x})y_i$~~
~~1~~	~~190~~	~~680~~	~~−490.0~~	~~−33.9~~	~~−23,052.0~~
~~2~~	~~200~~	~~800~~	~~−600.0~~	~~−23.9~~	~~−19,120.0~~
~~3~~	~~209~~	~~780~~	~~−571.0~~	~~−14.9~~	~~−11,622.0~~
~~4~~	~~215~~	~~885~~	~~−670.0~~	~~−8.9~~	~~−7,876.5~~
~~5~~	~~215~~	~~975~~	~~−760.0~~	~~−8.9~~	~~−8,677.5~~
~~6~~	~~215~~	~~1025~~	~~−810.0~~	~~−8.9~~	~~−9,122.5~~
~~7~~	~~230~~	~~1100~~	~~−870.0~~	~~6.1~~	~~6,710.0~~
~~8~~	~~250~~	~~1030~~	~~−780.0~~	~~26.1~~	~~26,883.0~~
~~9~~	~~265~~	~~1175~~	~~−910.0~~	~~41.1~~	~~48,292.5~~
~~10~~	~~250~~	~~1300~~	~~−1050.0~~	~~26.1~~	~~33,930.0~~
~~average~~	~~223.9~~	~~975.0~~			
~~**stdev (*S*)**~~	~~24.196~~	~~191.645~~	~~170.987~~		
~~S^2~~	~~585.433~~	~~36,727.778~~	~~29,236.544~~		

TABLE 1 Data and Calculations for Straight Line Regression Model Example

Sample, i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$\hat{Y}_i$	e_i	Statistics	Results	EQ
1	190	680	-33.9	-295.0	741.2	-61.2	S_{XX}	5268.90	Eq 3
2	200	800	-23.9	-175.0	810.1	-10.1	S_{YY}	330550.00	Eq 4
3	209	780	-14.9	-195.0	872.2	-92.2	S_{XY}	36345.00	Eq 5
4	215	885	-8.9	-90.0	913.6	-28.6	Slope, b_1	6.8980	Eq 6
5	215	975	-8.9	0.0	913.6	61.4	Intercept, b_0	-569.47	Eq 7
6	215	1025	-8.9	50.0	913.6	111.4	Variance, s^2	9980.16	Eq 10
7	230	1100	6.1	125.0	1017.1	82.9	St. Dev., s	99.90	
8	250	1030	26.1	55.0	1155.0	-125.0			
9	250	1300	26.1	325.0	1155.0	145.0			
10	265	1175	14.1	200.0	1258.5	-83.5			
	$\bar{X}$	$\bar{Y}$							
Average	223.9	975.0	0.0	0.0	975.0	0.0			
Equation	Eq 1	Eq 2			Eq 8	Eq 9			

~~**parameter estimates**~~

~~***b***~~	~~6.898~~
~~***a***~~	~~−569.468~~
~~S_{XX}~~	~~5,268.900~~
~~S_{YY}~~	~~330,550.000~~
~~S_{XY}~~	~~36,345.000~~

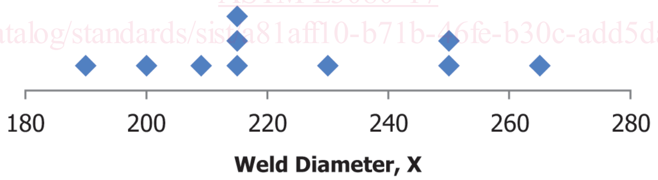

FIG. 2 Dot Plot of the Predictor Value *X*

(1) The least squares estimated model equation is $Y = -569.47 + 6.898\ X$. Clearly the negative intercept is not a plausible value for shear strength. This is apparently due to the fact that the data are far removed from the origin (0, 0). It is possible that there is some nonlinear behavior in the relationship approaching the origin.

(2) The averages of the deviations of X and Y from their averages are zero, and the average of the residuals are zero. These results follow from the property that sums of deviations from averages are zero.

(3) The average of the fitted values of Y is the same as the average of the Y data.

~~5.3 *Example*—An example for this kind of data and the associated basic calculations is shown in Table 1. This data is taken from Duncan (**3**), and shows the relationship between the measurement of shear strength, *Y*, and weld diameter, *X*, for 10 random specimens. Values for the estimated slope and intercept are $b = 6.898$ and $a = -569.468$. Fig. 2 shows the scatter plot and associated least squares linear fit.~~

~~In Eq 5, the slope estimate b is seen as a weighted average of the y_i where the weights, w_i, are defined as:~~

$$w_i = \frac{(x_i - \bar{x})}{S_{XX}} \qquad (7)$$

~~Values of x_i furthest from the average will have the greatest impact on the associated weight applied to observation y_i and on the numerical determination of the slope b.~~

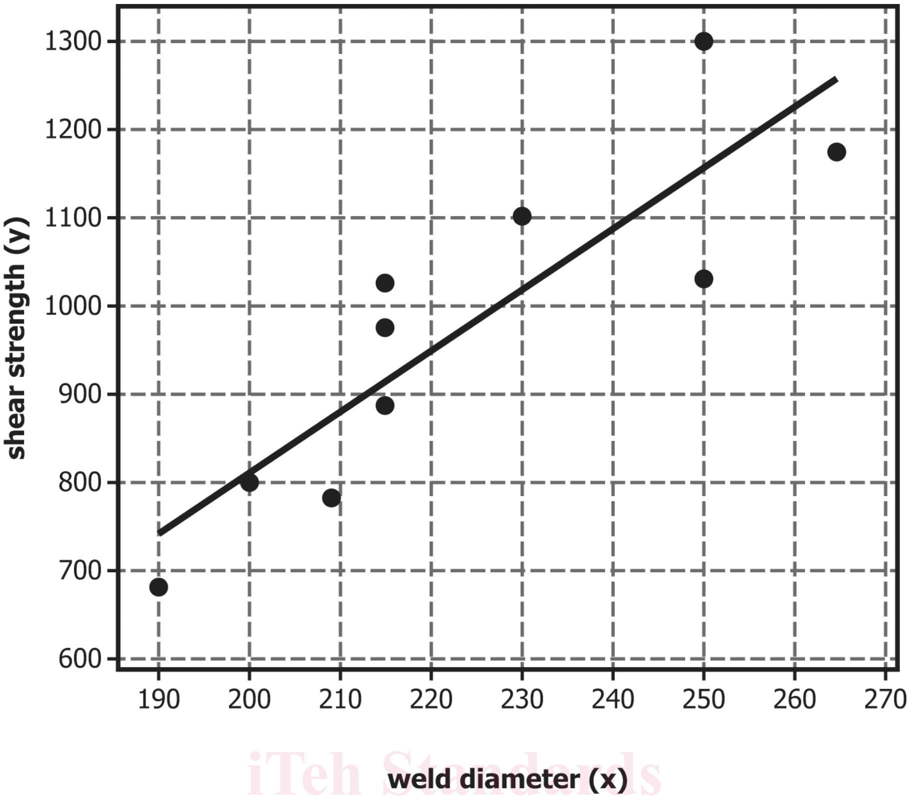

FIG. ~~1~~3 Scatter Plot of ~~Table 1~~Data with Fitted Linear Model

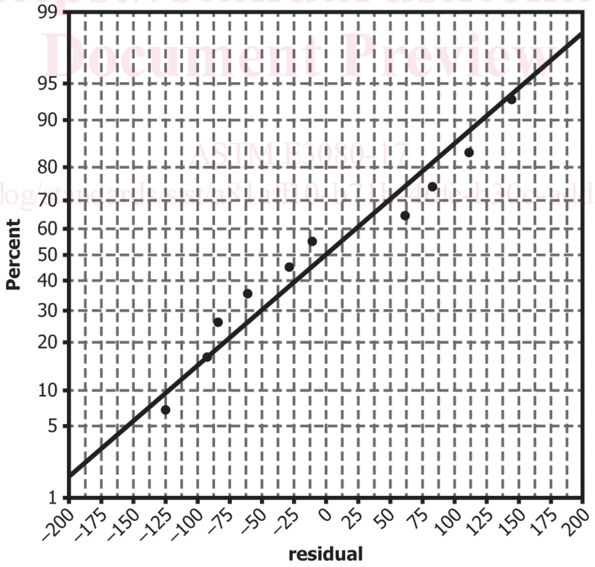

FIG. 8 Normal Probability Plot of Residuals

5.3 ~~*Correlation Coefficient*—~~*Evaluation of the Model:* ~~The population correlation coefficient, or Pearson Product Moment Correlation Coefficient, ρ, is a dimensionless parameter intended to measure the strength of a linear relationship between two variables. The estimated sample correlation coefficient, *r*, for a set of paired data (x_i, y_i) is calculated as:~~

$$r = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})y_i}{(n - 1)s_x s_y} \tag{8}$$

~~In Eq 8, the quantity $\frac{\sum_{i-1}^{n}(x - \bar{x})(y - \bar{y})}{(n - 1)}$ is referred to as the sample co-variance. Here again, the mean of *y* disappears from the~~

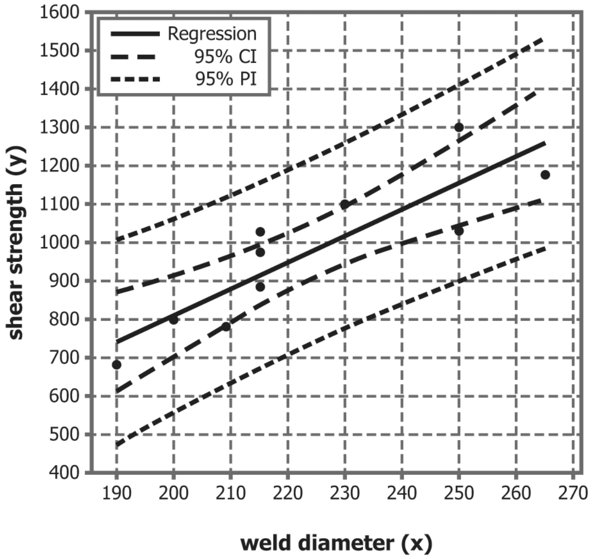

FIG. ~~5~~9 Regression Plot with 95 % Confidence and Prediction Intervals

~~right side of Eq 8, because~~ ~~$\sum_{i-1}^{n}(x-\bar{x})\bar{y}=0$.~~

5.3.1 ~~An alternative formula for r uses the standard deviation of the paired differences (~~This section discusses model evaluation through measures of association and plots of the residuals to check~~d~~$_i$ ~~=~~for ~~y~~departures$_i$ ~~=~~from ~~x~~the$_i$~~). Note that it does not matter in what order we calculate these differences. Either~~ model assumptions and the presence of data outliers.~~d$_i$ = y$_i$ – x$_i$ or d$_i$ = x$_i$ – y$_i$ will give the same result:~~

$$r=\frac{s_x^2+s_y^2-s_d^2}{2s_xs_y} \tag{9}$$

~~The correlation coefficient for the data in Table 1 using Eq 8 and Eq 9 are:~~

$$r=\frac{36,345}{(10-1)(24.196)(191.645)}=0.871$$

$$r=\frac{24.196^2+191.645^2-170.897^2}{2(24.196)(191.645)}=0.871$$

5.3.2 ~~The value of the correlation coefficient is always between –1 and +1. If *r* is negative (*y* decreases as *x* increases) then a line fit to the data will have a negative slope; similarly, positive values of *r* (*y* increased as *x* increases) are associated with a positive slope. Values of *r* near 0 indicate no linear relationship so that a line fit to the data will have a slope near 0. In cases where the (*x*,*y*) data have an *r* = –1 or *r* = +1, the relationship between *x* and *y* is perfectly linear. An *r* value near to +1 or –1 indicate that a line may provide an adequate fit to the data but does not "prove" that the relationship is linear since other models may provide a better fit (for example, a quadratic model). As values of *r* become closer to the extremes (–1 and +1) a line provides a stronger explanation of the relationship. Fig. 2 shows examples of what correlated data look like for several values of *r*.~~*Measures of Association Between X and Y:*

5.3.2.1 The sample correlation coefficient is a dimensionless statistic intended to measure the strength of a linear relationship between two variables. The estimated correlation coefficient, *r*, from a set of paired data (X_i, Y_i) is calculated from three statistics, S_{XX}, S_{YY}, and S_{XY}:

$$r=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \tag{12}$$

The value of the correlation coefficient ranges between –1 and +1. The sign of *r* is the same as the sign of slope estimate b_1. Values of *r* near 0 indicate a weak or nonexistent straight line relationship. An *r* value closer to either +1 or –1 indicates that a straight line provides an ever stronger explanation of the relationship. Fig. 4 shows examples of scatter plots that appear for selected values of *r*.

5.3.2.2 The coefficient of determination is the squared value of the correlation coefficient with symbol r^2. It measures the proportion of variation in the *Y* data explained by the predictor variable *X*.

5.3.2.3 For the example the sample correlation coefficient is:

$$r=\frac{36345}{\sqrt{(330550)(5268.9)}}=0.8709$$