NOTICE: This standard has either been superseded and replaced by a new version or withdrawn.
Contact ASTM International (www.astm.org) for the latest information

Designation: E3080 – 19

An American National Standard

Standard Practice for Regression Analysis with a Single Predictor Variable[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 of a set of data to define the statistical relationship between two numerical variables for use in predicting one variable from the other.

1.2 The regression analysis provides graphical and calculational procedures for selecting the best statistical model that describes the relationship and for evaluation of the fit of the data to the selected model.

1.3 The resulting regression model can be useful for developing process knowledge through description of the variable relationship, in making predictions of future values, in relating the precision of a test method to the value of the characteristic being measured, and in developing control methods for the process generating values of the variables.

1.4 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.

1.5 *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, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.6 *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

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 *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**

3.1.2 *predictor variable, X, n*—a variable used to predict a response variable using a regression model.

3.1.2.1 *Discussion*—Also called an *independent* or *explanatory* variable.

3.1.3 *regression analysis, n*—a statistical procedure used to characterize the association between two or more numerical variables for prediction of the response variable from the predictor variable.

3.1.3.1 *Discussion*—In this practice, only a single predictor variable is considered.

3.1.4 *residual, n*—the observed value minus fitted value, when a regression model is used.

3.1.5 *response variable, Y, n*—a variable predicted from a regression model.

3.1.5.1 *Discussion*—Also called a *dependent* variable.

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

3.1.7 *sample correlation coefficient, r, n*—a dimensionless measure of association between two variables estimated from the data.

3.1.8 *sample covariance, s_{xy}, n*—an estimate of the association of the response variable and predictor variable calculated from the data.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *intercept, β_0, n—of a regression model*, the value of the response variable when the value of the predictor variable is equal to 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, σ, n—of a regression model*, the square root of the residual variance.

[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 Sept. 1, 2019. Published January 2020. Originally approved in 2016. Last previous edition approved in 2017 as E3080 – 17. DOI: 10.1520/E3080-19.

[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.

Copyright © ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959. United States

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

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

3.3 *Symbols:*

b_0 = intercept parameter estimate (5.5.1)
b_1 = slope parameter estimate (5.5)
b_{11} = curvature parameter estimate (8.1.1.1)
β_0 = intercept parameter in model (5.3.1)
β_1 = slope parameter in model (5.3.1)
β_{11} = curvature parameter in model (5.3.3)
E = general point estimate of a parameter (5.7)
e_i = residual for data point i (5.5.2)
ε = error term in model (5.4)
F = F statistic (6.5.2)
h = index for predicting any value in data range (6.4.3)
i = index for a data point (5.2)
L = lower confidence limit (5.7.2)
λ = Box-Cox parameter (A1.5.4)
n = number of data points (5.2)
p = number of parameters in regression model (5.7)
r = correlation coefficient (6.3.2.1)
r^2 = coefficient of determination (6.3.2.2)
$S(b_0,b_1)$ = sum of squared deviations of Y_i to the regression line (A1.1.2)
s_{b1} = standard error of slope estimate (6.4.1)
s_{b0} = standard error of intercept estimate (6.4.2)
s_E = general standard error of a point estimate (5.7)
σ = residual standard deviation (5.4.1)
s = estimate of σ (6.2.6)
σ^2 = residual variance (5.4.1)
s^2 = estimate of σ^2 (6.2.6)
s_X^2 = variance of X data (A1.2.1)
s_Y^2 = variance of Y data (A1.2.1)
S_{XX} = sum of squares of deviations of X data from average (6.2.3)
S_{XY} = sum of cross products of X and Y from their averages (6.2.3)
s_{XY} = sample covariance of X and Y (A1.2.1)
$s_{\hat{Y}_h}$ = standard error of $\hat{Y}_h$ (6.4.3)
$s_{\hat{Y}_{h(ind)}}$ = standard error of future individual Y value (6.4.4)
S_{YY} = sum of squares of deviations of Y data from average (6.2.3)
t = Student's t distribution (5.7)
U = upper confidence limit (5.7.2)
X = predictor variable (5.1)
$\bar{X}$ = average of X data (6.2.3)
X_h = general value of X in its range (6.4.3)
X_i = value of X for data point i (5.2)
Y = response variable (5.1)
$\bar{Y}$ = average of Y data (6.2.3)
$\dot{Y}$ = geometric mean of Y data (A1.5.4)
Y' = transformed Y (A1.5.2)
$\hat{Y}_{h(ind)}$ = predicted future individual Y for a value X_h (6.4.4)
Y_i = value of Y for data point i (5.2)
$\hat{Y}_h$ = predicted value of Y for any value X_h (6.4.3)
$\hat{Y}_i$ = predicted value of Y for data point i (5.5.1)

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 procedure that uses data to study the statistical relationships between two or more variables (**1, 2**).[3] This practice is restricted in scope to consider only a single numerical response variable and a single numerical predictor variable. The objective is to obtain a regression model for use in predicting the value of the response variable Y for given values of the predictor variable X.

4.2 A regression model consists of: (*1*) a *regression function* that relates the mean values of the response variable distribution to fixed values of the predictor variable, and (*2*) a *statistical distribution* that describes the variability in the response variable values at a fixed value of the predictor variable.

4.2.1 The regression analysis utilizes either *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.

4.3 The information in this practice is arranged as follows.

4.3.1 Section 5 gives a general outline of the steps in the regression analysis procedure. The subsequent sections cover procedures for estimation of specific regression models.

4.3.2 Section 6 assumes a straight line relationship between the two variables. This is also known as the simple linear regression model or a first order model. This model should be used as a starting point for understanding the XY relationship and ultimately defining the best fitting model to the data.

4.3.3 Section 7 considers a proportional relationship between the variables, where the ratio of one variable to the other is constant. The intercept is constrained to be zero. This model is useful for single point calibration, where a reference material is run periodically as a standard during routine testing to correct for drift in instrument performance over a given range of test results.

4.3.4 Section 8 discusses a regression function that considers curvature in the XY relationship, the second order polynomial model.

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

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/8a82b0ab-...

4.3.5 Annex A1 provides supplemental information of a more mathematical nature in regression.

4.3.6 Appendix X1 lists calculations for the curvature model estimates and exhibits a worksheet for these calculations.

5. Regression Analysis Procedure for a Single Predictor Variable

5.1 *Choose the response variable Y and the predictor variable X.* The predictor variable X is assumed to have known values with little or no measurement error. For given values of X, the response variable Y has a distribution of values representing the random effect of measurement errors, and these distributions are defined within a given range of the X values.

5.2 *Obtain a data set* consisting of n pairs of values designated as (X_i, Y_i), with the sample index i ranging from 1 through n. The data can arise in two different ways. Observational data consists of X and Y values measured on a set of n random test units. Experimental data consists of Y values measured on n test units with X values set at controlled values in an experimental study.

5.2.1 When designing an experiment for defining the XY association some considerations are:

(1) Range of X values.

(2) Number of distinct X values.

(3) Spacing of X values.

(4) Number of Y observations for each X value.

The answers depend on the objectives of the investigation, whether determining the nature of the regression function, estimating the slope or intercept of the simple linear model, or estimating the measurement error of Y, as well as other objectives.

5.2.1.1 The X values should cover the entire range of interest. Extrapolation beyond the range of observed X values may fail due to expanding estimation error outside the range and the uncertainty of whether the model gives an adequate description of the XY relationship outside the range. When inference is required for the Y intercept (the value of Y when X is zero) the range of X should extend down to zero or near zero.

5.2.1.2 Two X levels are necessary when the objective is to determine if there is an effect of X on Y, and to give an estimate of the effect (slope). Three X levels are necessary to evaluate any curvature in the relationship. Four or more X levels give better definition of the model shape, particularly if there is a possible asymptote or a threshold in the relationship. The X levels should be equally spaced. If X is transformed, such as to logarithms, the equal spacing should be with respect to the transformed X.

5.2.1.3 Usually the number of Y observations should be equal at each X level. When the objective is to estimate Y variance or evaluate variance constancy, then at least four observations are recommended at each X level.

5.3 *Choose a regression function that fits the data.* A *scatter plot* of the data is recommended for a visual look at the XY relationship, and most computer packages have this as an option. This is a plot of points on the XY plane having a value of Y (on the vertical axis) and a value of X (on the horizontal axis) for each data pair, where it is useful for evaluating the quality of the data and suggesting an appropriate regression function to define the XY relationship. Fig. 1 gives examples of four scatter plots that illustrate different situations.

5.3.1 Fig. 1A shows a cluster of points that appear to be elongated in a particular direction along a straight line that does not pass through the *origin* (X=0, Y=0). This pattern suggests

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/8a82b0ab-5726-470c-9570-ed1f632d33c2/astm-e3080-19

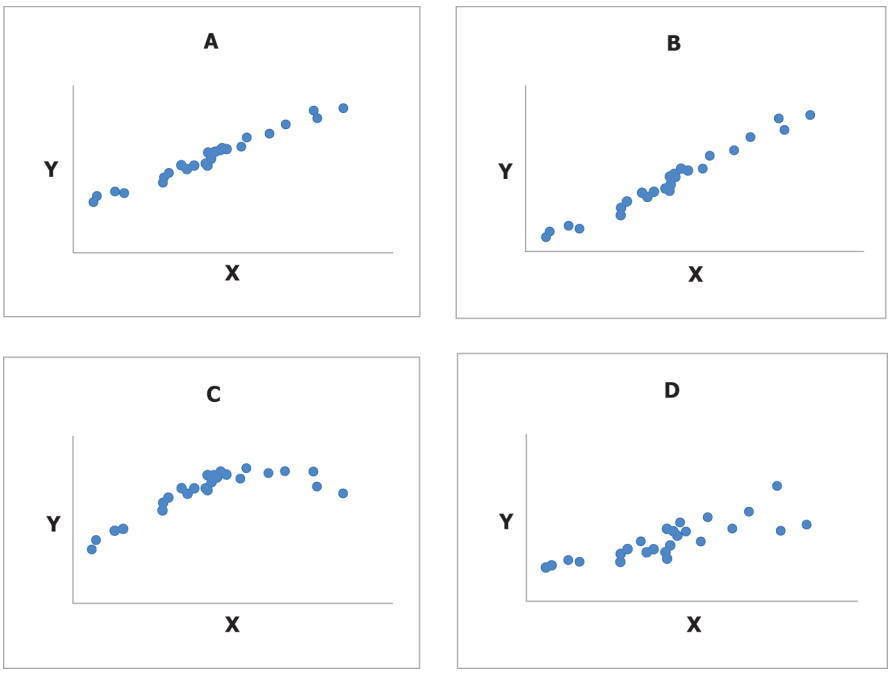

FIG. 1 Scatter Plots

the straight line regression function $Y=\beta_0+\beta_1X$. The two *parameters* for this function are the *intercept* β_0 and the *slope* β_1. The slope is the amount of incremental change in *Y* units for a unit change in *X*. The intercept is the value of *Y* when $X = 0$. Both parameters are necessary to define this regression function.

5.3.2 Fig. 1B suggests a straight line that appears to go through the origin, thus *Y* is proportional to *X*, and the regression function is $Y=\beta_1X$. An intercept term is not required because the *Y* intercept is constrained to equal zero, that is, the line goes through the origin.

5.3.3 Fig. 1C indicates curvature in the relationship, and there are several regression functions that can be used. For slight curvature, a simple model is to add a second order (X^2) term to the straight line function as $Y=\beta_0+\beta_1X+\beta_{11}X^2$.

5.3.4 Fig. 1D shows data with increasing variability with larger mean values. This suggests the need for a weighted regression procedure discussed in A1.4.2.

5.3.5 Data points appearing outside the swarm of data (*outliers*) can have an adverse effect on estimation of regression function parameters. For the straight-line function, outliers at the extremes of the *X* range can greatly affect the estimate of the slope and intercept parameters, and outliers in the middle of the range tend to affect the intercept estimate more than the slope. Outliers can be formally identified by statistical procedures (see Practice E178).

5.3.6 A special situation occurs when there are two data swarms separated by a gap. This may indicate that there were two sources of data with different values of a second lurking predictor variable. Such a data set consists essentially of two data points in cases of a large gap.

5.4 *Define the regression model* by adding an error term to the regression function that describes the variation in *Y* through a statistical distribution. For example, the *simple linear regression model* using the regression function in 5.3.1 is then stated as $Y=\beta_0+\beta_1X+\varepsilon$, where ε is a random error having a distribution with mean zero and standard deviation σ (variance σ^2).

5.4.1 The distribution for ε can often be assumed to have a normal (Gaussian) distribution with a constant standard deviation over the range of *X*. Thus, the distribution of *Y* at a given *X* is a normal distribution with a mean of $\beta_0+\beta_1X$ and a standard deviation of σ. An example of such a linear regression model is shown in Fig. 2 over a range of *X* from 0 to 40 *X* units. Normal distributions of response *Y* with σ = 1.3 *Y* units are depicted at *X* = 10, 20, and 30 *X* units.

5.4.2 Distributions other than the normal distribution may also be considered, depending on knowledge of the application. For example, low microbial counts may use a Poisson error distribution.

5.5 *Parameter estimation* uses the data set to provide the parameter estimates. For the simple regression functions described above, the procedures used are given in the following sections. In this practice, the parameters are lower-case Greek letters and the estimates are the corresponding lower-case Roman letters. For example, the estimate of the slope parameter β_1 is b_1.

5.5.1 The *fitted values of Y*, denoted $\hat{Y}_i$ (read *Y*-hat), for each data point (X_i, Y_i) are calculated from the estimated regression function. For the straight-line model, the fitted values of Y_i are $\hat{Y}_i=b_0+b_1X_1$. The right-hand function defines the regression line, which may be shown on the scatter plot of the data to evaluate model fit.

5.5.2 The estimates of the error term values ε are the *residuals* ε_i, calculated as $e_i=Y_i-\hat{Y}_i$, and these are used to estimate the standard deviation parameter σ. Note that the residual values are the vertical distances of the points from the regression line.

5.6 *Evaluation of the regression model* is performed to diagnose departure from model assumptions, such as model fit to the data, constancy of variance over the range of *X*, and conformance to the assumed error distribution. Residual plots are useful for these diagnostics.

5.6.1 A plot of the residuals against their *X* values (or

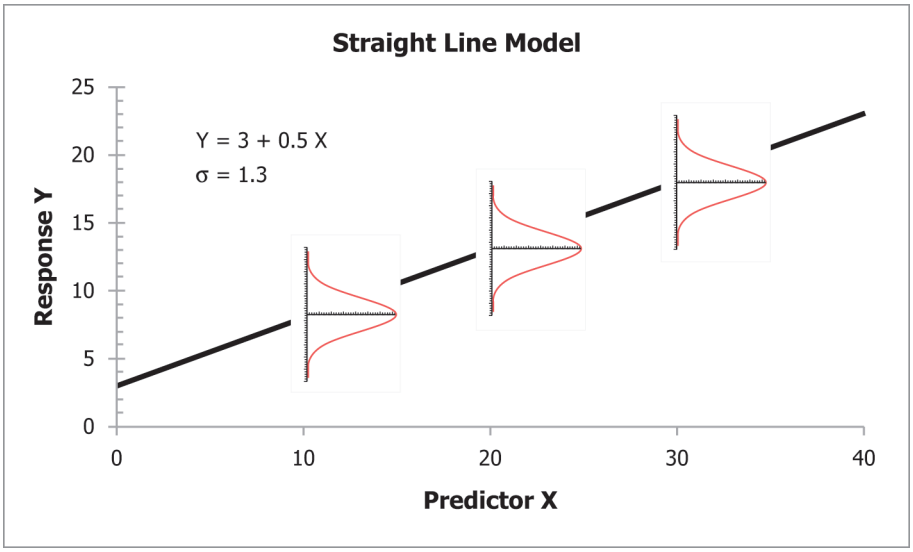

FIG. 2 Graphical Depiction of a Straight Line Regression Model

equivalently, against their $\hat{Y}_i$ values) will detect certain departures from the assumptions. Residuals may also be plotted against time of testing (if available) or against another known variable. Fig. 3 shows some of these patterns and discusses remedies for these departures. (The horizontal line on the plots indicates a value of zero for the average of the residuals.)

(1) Plot A – the desired horizontal pattern – indicates no model deficiencies

(2) Plot B – increasing variance with *X*, consider weighted regression (see A1.4.2) or data transformations (see A1.5)

(3) Plot C – curvature in the relationship, consider adding a quadratic term or using a nonlinear model (see Section 8)

(4) Plot D – possible effect of time order of testing or the effect of another variable denoted as *T*

5.6.2 Plotting the residuals against a vertical scale of the cumulative percentage of the normal distribution checks the assumption of normality in the model. The fitted cumulative normal distribution from the data is shown as a straight line on the plot if the residuals fit a normal distribution. Computer packages provide these plots and can also perform a more rigorous statistical test for normality. If the plot indicates a curve, a data transformation may be required to achieve a normal distribution.

5.6.3 Outlier testing in regression analysis takes two forms. Outlier testing can be performed upon any sets of multiple *Y*'s collected at each unique value of *X* studied. Additionally, outlier analysis can be performed on the entire set of residuals. In the latter case, finding an outlier could indicate an issue in either the *X* or *Y* value of the point in question or it may indicate other issues with the regression analysis.

5.7 *Use of the model for interval estimates of regression parameters and predicted Y values.*

5.7.1 The estimates of model parameters and fitted *Y* values are *point estimates*. For example, the estimate of the slope parameter β_1 is the estimate b_1 that has been calculated from the data. To give a sense of the precision for these estimates, *interval estimates*, or confidence intervals, can be provided. A general form for the *confidence interval* for a general point estimate *E* is:

$$E \pm t s_E \tag{1}$$

where s_E is the *standard error* of the estimate and *t* is a tabulated multiplier that is dependent upon *the degrees of freedom* of the standard error and the desired *confidence level*, stated as a percentage. Thus, we may state that the true value of the parameter being estimated lies within the confidence interval at a given confidence level. The degrees of freedom for the standard error are generally $n - p$, where *p* is the number of parameters in the regression model.

5.7.2 To calculate these interval estimates, the form of the statistical distribution for *Y* is required, and the normal distribution is often assumed. The widths of the interval estimates, given here as two-sided confidence intervals, are dependent on (*1*) the standard errors of the estimates, and (*2*) the level of confidence. The standard errors depend on the number of data pairs *n* and the values of the X_i.

The confidence level is defined as $100(1 - \alpha)$ %, where α is the probability that the confidence interval does not contain the parameter value. For example, $\alpha = 0.05$ (or a risk of 5 % non-coverage) corresponds to a confidence level of 95 %, which shall be used for the examples in this practice. The value of *t* is the upper $(1 - \alpha/2)$th quantile of the Student's *t* distribution with $n - p$ degrees of freedom, for a confidence level of $100(1 - \alpha)$ %. Values of *t* are found in statistical texts and in commercial statistical software packages.

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/8a82b0ab-5726-470c-9570-ed1f632d33c2/astm-e3080-19

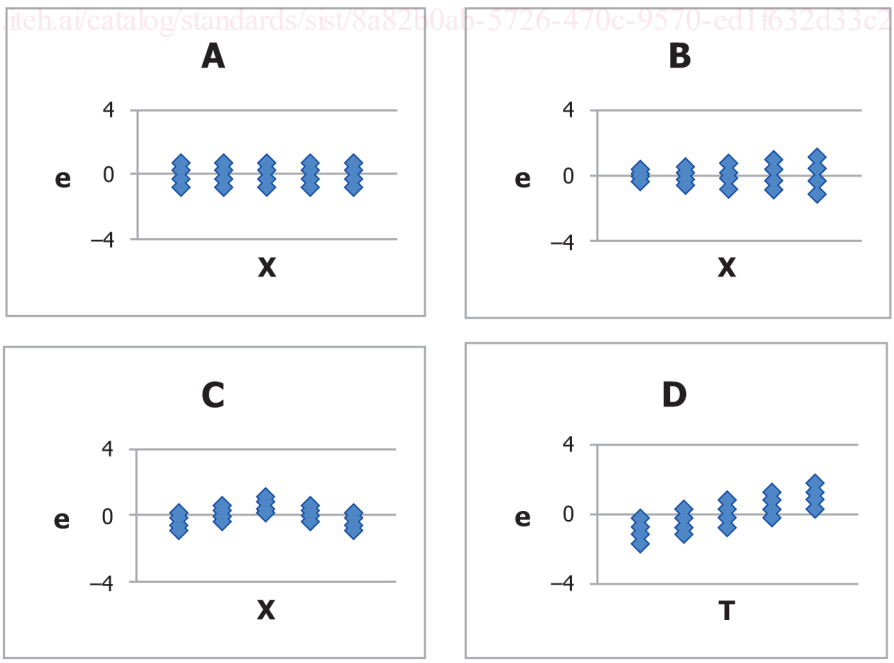

FIG. 3 Residual Plots – Some Patterns

5.7.3 The confidence interval can also be stated as the interval (L, U) between lower (L) and upper (U) confidence limits for the parameter being estimated. Practice E2586 provides discussion of confidence intervals, standard error, and degrees of freedom.

6. Simple Linear Regression Analysis

6.1 *Simple Linear Regression Model:*

6.1.1 This model defines the functional relationship between X and Y as a straight line in the XY plane.

6.1.2 The *regression function* for the straight line relationship is:

$$Y = \beta_0 + \beta_1 X \tag{2}$$

where the two *parameters* for the function are the *intercept* β_0 and the *slope* β_1.

The intercept is the value of Y when $X = 0$, but this parameter may not be of practical interest when the range of X is far removed from zero. The slope is the amount of incremental change in Y units for a unit change in X.

6.1.3 The statistical distribution for Y is usually assumed to be a normal (Gaussian) distribution having a mean of $\beta_0 + \beta_1 X$ with a standard deviation σ. The *simple linear regression model* is then stated as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3}$$

where ε is a random error that is normally distributed with mean zero and standard deviation σ (variance σ^2).

6.2 *Estimating Regression Model Parameters:*

6.2.1 The model parameters β_0, and β_1, are estimated from a sample of data consisting of n pairs of values designated as (X_i, Y_i), with the sample number i ranging from 1 through n. The data can arise in two different ways. Observational data consists of X and Y values measured on a set of n random samples. Experimental data consists of Y values measured on n experimental units with X values set at fixed values. In both cases the Y values may have measurement error, but the X values are assumed known with negligible measurement error.

6.2.2 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 A1.1.2.)

6.2.3 Calculate the following statistics from the X and Y values in the data set.

6.2.3.1 Calculate the averages of X and Y:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{5}$$

6.2.3.2 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:

$$S_{XX} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{6}$$

$$S_{YY} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \tag{7}$$

$$S_{XY} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{8}$$

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

6.2.3.3 The least squares solution gives the parameter estimates:

$$b_1 = S_{XY} / S_{XX} \tag{9}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{10}$$

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

$$\hat{Y}_i = b_0 + b_1 X_i \tag{11}$$

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

$$e_i = Y_i - \hat{Y}_i \tag{12}$$

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

6.2.6 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:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-2)} = \frac{\sum_{i-1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{(n-2)} \tag{13}$$

$$s = \sqrt{s^2} \tag{14}$$

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.

6.2.7 *Example*—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.

6.2.7.1 The scatter plot for this example is shown in Fig. 4. The shear strength appears to be increasing in a linear fashion with weld diameter. There is some scatter but no apparent outlying data points.

6.2.7.2 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},

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/0a82b0ab-5726-470c-9570-ed1f632d34c2/astm-e3080-19

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 6
2	200	800	–23.9	–175.0	810.1	–10.1	S_{YY}	330550.00	Eq 7
3	209	780	–14.9	–195.0	872.2	–92.2	S_{XY}	36345.00	Eq 8
4	215	885	–8.9	–90.0	913.6	–28.6	Slope, b_1	6.8980	Eq 9
5	215	975	–8.9	0.0	913.6	61.4	Intercept, b_0	–569.47	Eq 10
6	215	1025	–8.9	50.0	913.6	111.4	Variance, s^2	9980.16	Eq 13
7	230	1100	6.1	125.0	1017.1	82.9	St. Dev., s	99.90	Eq 14
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 4	Eq 5			Eq 8	Eq 9			

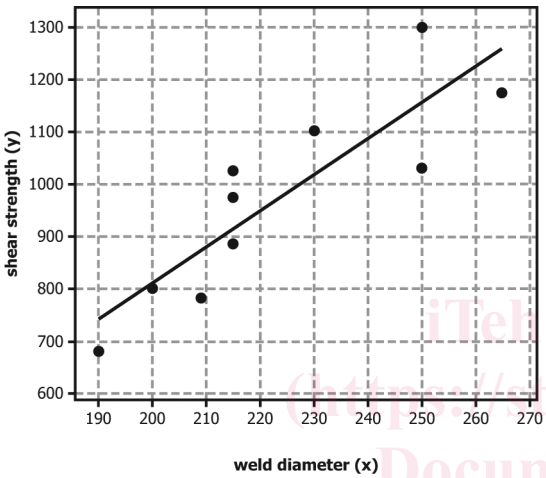

FIG. 4 Scatter Plot of Data with Fitted Linear Model

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.

6.2.7.3 The least squares straight line is depicted with the scatter plot in Fig. 4, and indicates that a straight line model appears to give a reasonable fit to this data set. Some additional comments from Table 1 are:

(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 so far removed from the origin (0, 0) that the estimate is poorly defined. It is also 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, $\hat{Y}_i$, is the same as the average of the Y data.

6.3 *Evaluation of the Model:*

6.3.1 This section discusses model evaluation through measures of association and plots of the residuals to check for departures from the model assumptions and the presence of data outliers.

6.3.2 *Measures of Association Between X and Y:*

6.3.2.1 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}:

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} \tag{15}$$

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. 5 shows examples of scatter plots that appear for selected values of r.

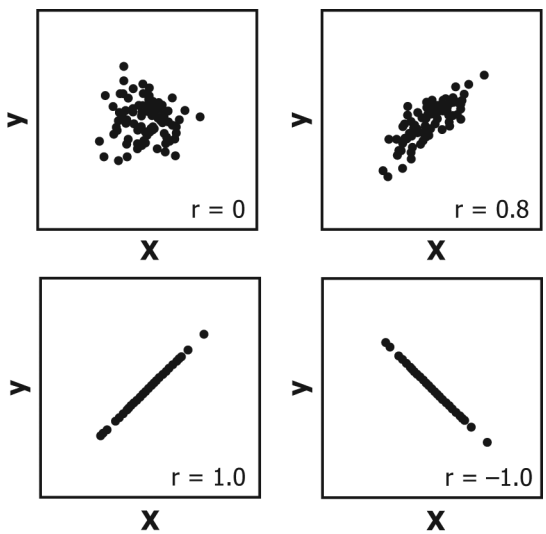

FIG. 5 Typical Scatter Plots for Selected Values of the Correlation Coefficient, *r*

iTeh Standards (https://standards.iteh.ai) Document Preview ASTM E3080-19

6.3.2.2 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.

6.3.2.3 For the example the sample correlation coefficient is:

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

The sample coefficient of determination for the example is $r^2 = 0.8709^2 = 0.7585$. This means that approximately 76 % of the variance in Y is explained by the straight line model (see 6.5.2). These measures are often used as acceptance criteria for linearity; but this usage should be discouraged, because these statistics are not absolute measures of linearity and should be used for comparative purposes only.

6.3.3 *Residual Plots:*

6.3.3.1 Plots of residuals e_i are used for evaluating outliers in the data and various model assumptions over the range of X, including normality, constant error variance, linearity of the regression function, and independence of the error terms. These check for outliers in the data, constancy of Y distribution variance, curvature of the regression function, lack of independence of errors, and normality of the Y distribution.

6.3.3.2 The residuals dot plot is a useful diagnostic for finding outliers, which may be harder to detect from the data set itself. Large outliers can distort the estimate of the regression line because the least squares procedure will tend to move the line towards the outlier, thus masking it. Formal outlier testing procedures can be found in Practice E178.

A residuals dot plot for the example is shown in Fig. 6. There are no apparent outliers at each end of the plot.

Additional graphics for this purpose are histograms, "stem and leaf" plots, and "box and whiskers" plots. (See Practice E2586.)

The plot of residuals against X in Fig. 7 indicates no discernable pattern, such as curvature or increasing scatter versus X, but this is a relatively small data set.

6.3.3.3 Plotting the residuals against a vertical scale of the cumulative percentage of the normal distribution checks the assumption of normality in the model. The fitted cumulative normal distribution from the data is shown as a straight line on the plot if the residuals fit a normal distribution. Computer packages provide these plots and can also perform a more rigorous statistical test for normality.

For the example, the residual plot against X in Fig. 8 indicates an approximate straight line pattern for the example, supporting a normal distribution for the residuals.

6.4 *Interval Estimates of Regression Parameters and Predicted Y Values*—This section shows the calculations for the interval estimates for b_0 and b_1 of their respective model parameters β_0 and β_1 for the simple linear model (see 5.7 for an introduction to this concept). Also given are calculations for

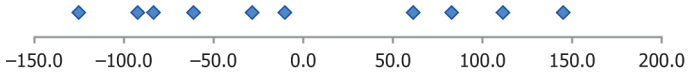

FIG. 6 Dot Plot of Residuals

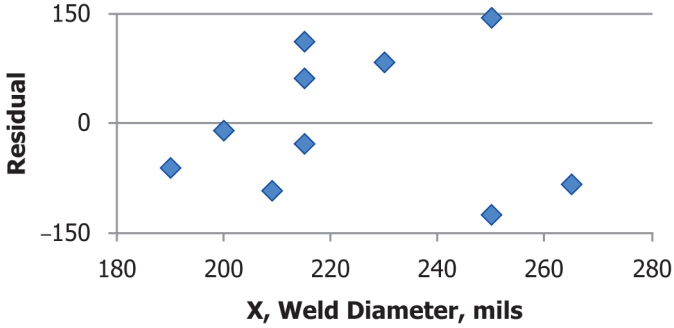

FIG. 7 Plot of Residuals versus *X* — Duncan Example

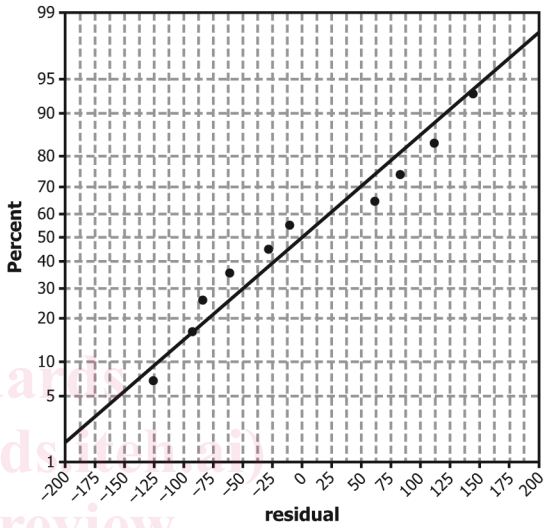

FIG. 8 Normal Probability Plot of Residuals

certain predicted values of Y at given values of X. For these calculations the estimate s of the standard deviation σ of the Y distribution is required with its degrees of freedom $n - 2$. Also required is the choice of the confidence level, and for these calculations a 95 % confidence interval will be used. In the example, the standard deviation estimate is s = 99.9 inch-pounds with $n - 2 = 10 - 2 = 8$ degrees of freedom. The value of t for a 95 % two-sided confidence interval with 8 degrees of freedom is 2.306.

6.4.1 *Confidence Interval for the Slope*—The standard error for the slope estimate is:

$$s_{b1} = s / \sqrt{S_{XX}} \tag{16}$$

From the example:

$$s_{b1} = 99.9 / \sqrt{5268.9} = 1.376$$

The confidence interval for the slope β_1 is calculated as:

$$b_1 \pm t s_{b1} \tag{17}$$

From the example, the 95 % confidence interval is:

$$6.898 \pm (2.306)(1.376) = 6.898 \pm 3.173$$

$$\text{or } (3.725, \quad 10.071)$$

If the slope confidence interval includes zero, this supports the assertion that there is no relationship between X and Y at the given level of confidence. In this example, the slope confidence

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/8a82b0ab-...

interval does not include zero, thus supporting the existence of a statistical relationship between Y and X.

6.4.2 *Confidence Interval for the Intercept*—The standard error for the intercept estimate is:

$$s_{b0} = s\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}} \tag{18}$$

From the example:

$$s_{b0} = 99.9\sqrt{\frac{1}{10}+\frac{223.9^2}{5268.9}} = 309.76$$

The confidence interval for the intercept β_0 is calculated as:

$$b_0 \pm ts_{b0} \tag{19}$$

In this example, the 95 % confidence interval is:

$$-597.5 \pm (2.306)(309.76) = -569.5 \pm 714.3$$

$$\text{or } (-1283.8,\quad 144.8)$$

If the confidence interval includes zero, this technically supports the assertion that the line may go through the origin (0,0) at the given level of confidence. However, this use of the confidence interval amounts to a rather large extrapolation outside the range of the data, which explains the implausible negative estimate mentioned in 6.2.7.3.

6.4.3 *Confidence Interval for the Predicted Value of the Mean Y at a Given X*—The *predicted value* $\hat{Y}_h$ for a mean response of Y at X_h is:

$$\hat{Y}_h = b_0 + b_1 X_h \tag{20}$$

The index h is used instead of the index i because the prediction is not necessarily from a value of X in the data set.

Predictions outside the range of X (extrapolation) should be performed with caution, as the regression function may not be valid outside this range.

The standard error for a mean $\hat{Y}_h$ response at a value $X = X_h$ is:

$$s_{\hat{Y}_h} = s\sqrt{\frac{1}{n}+\frac{\left(X_h - \bar{X}\right)^2}{S_{XX}}} \tag{21}$$

From the example, at X_h = 215 mils, the standard error is:

$$s_{\hat{Y}_h} = 99.9\sqrt{\frac{1}{10}+\frac{(215 - 223.9)^2}{5268.9}} = 33.88$$

The confidence interval for the mean Y response at a value $X = X_h$ is calculated as:

$$\hat{Y}_h \pm ts_{\hat{Y}_h} \tag{22}$$

From the example, the 95 % confidence interval for the average predicted value of 913.6 inch-pounds is:

$$913.6 \pm (2.306)(33.88) = 913.6 \pm 78.13$$

$$\text{or } (835.47,\quad 991.73)$$

Thus the expected mean response of Y at X = 215 falls between 835.47 and 991.73 with 95 % confidence.

6.4.4 *Confidence Interval for the Predicted Value of a Future Value Y at a Given X*—The standard error for an individual response $\hat{Y}_{h(ind)}$ at $X = X_h$ is calculated as:

$$s_{\hat{Y}_{h(ind)}} = s\sqrt{1+\frac{1}{n}+\frac{\left(X_h - \bar{X}\right)^2}{S_{XX}}} \tag{23}$$

From the example, at X_h = 215 mils, the standard error is:

$$s_{\hat{Y}_{h(ind)}} = 99.9\sqrt{1+\frac{1}{10}+\frac{(215 - 223.9)^2}{5268.9}} = 105.49 \text{ inch-pounds}$$

The confidence interval for a future new Y response at a value $X = X_h$ is calculated as:

$$\hat{Y}_{h(ind)} \pm ts_{\hat{Y}_{h(ind)}} \tag{24}$$

This is known as a *prediction interval*, an interval estimate in which would contain a future observation with a given probability based on the data set. Prediction intervals are wider than confidence intervals because a prediction interval applies to an individual value whereas the confidence interval applies to a mean response. In the example, the prediction interval at 95 % confidence for the predicted value of the response at a weld diameter of 215 mils is:

$$913.6 \pm (2.306)(105.49) = 913.6 \pm 243.26$$

$$\text{or } (670.34,\quad 1156.86)$$

6.4.5 An array of confidence intervals and prediction intervals shown as bands around the regression line is depicted in Fig. 9 for the example. The vertical intervals are narrowest at the *centroid*, $(\bar{X}, \bar{Y})$ of the data and become wider as the distance from the center increases. These bands are valid for a single predictions only. Multiple predictions using the same data set are discussed in A1.1.8.1. These bands can be useful in setting manufacturing requirements; for example, the confidence interval indicates that a minimum weld diameter of 200 mils would be required to obtain an average shear strength of 700 inch-pounds at 95 % confidence. The prediction interval suggests that a minimum shear strength of 220 mils would be necessary to guarantee that a single future item would have meet that shear strength with 95 % confidence.

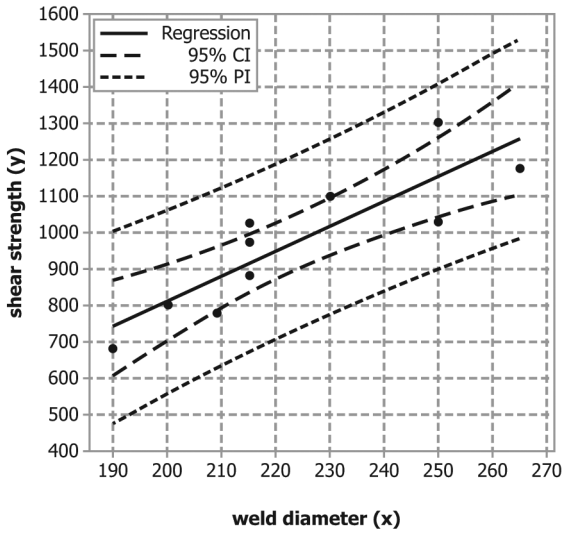

FIG. 9 Regression Plot with 95 % Confidence and Prediction Intervals

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/8a82b0ab-...

6.5 *Analysis of Variance (ANOVA) Calculations:*

6.5.1 Statistical analysis packages are often used for regression analysis. The output consists of the estimates of the regression parameters, various plots, and an ANOVA table. The calculations for the ANOVA table are shown in Table 2. This section discusses the ANOVA procedure and its relation to earlier calculations.

6.5.2 ANOVA partitions the total sum of squares in the Y data, *SST*, into the residual sum of squares, *SSE*, and the regression sum of squares, *SSR*. The degrees of freedom (*df*) for these sums of squares are respectively $n - 1$, $n - 2$ and 1. *SST* has been previously calculated as S_{YY} in Eq 7, and *SSE* has been previously calculated as the sum of the squared residuals, the numerator of s^2 in Eq 13. *SSR* is the sum of squares of deviations of the fitted values from their average $\bar{Y}$, which represents the variation removed from the Y data due to its estimated relationship with X. *SSR* may also be equivalently calculated as:

$$SSR = \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2 = b_1^2\Sigma\left(X_i - \bar{X}\right)^2 \tag{25}$$

This expression enables calculation of the sums of squares for regression and for error without first requiring calculation of fitted values and residuals.

The mean squares are variances, each calculated as a sum of squares divided by its degrees of freedom. The F statistic is the ratio of the regression mean square to the residual mean square, and is used to test the fit of the regression model, thus $F = MSR/MSE$. *MST* is the variance of the Y data, see Eq A1.14.

The p-value is the probability of obtaining a slope estimate as large as that obtained from the data, assuming that the true slope is zero. Low values of p, such as $p < 0.05$, are used to reject the condition that the true slope is zero, thus confirming that a relationship that is either linear, or that has a statistically-meaningful trend component, exists between X and Y.

6.5.3 The ANOVA table for the example is shown in Table 3. The F test indicated a high level of statistical significance for the validity of the model with a low p value of 0.001. The coefficient of determination $r^2 = SSR / SST = 250709 / 330550 = 0.7585$, which agrees with the value in 6.3.2.3.

7. Zero Intercept Linear Model

7.1 An associated model often considered along with the simple linear model is the model that constrains the intercept to be zero. Thus Y is *proportional* to X throughout the range. This model is useful in test methods where single-point calibration is conducted periodically, due to minor instabilities in the testing process. The regression model is:

$$Y = \beta_1 X + \varepsilon \tag{26}$$

where the *slope* β_1 is the single regression function parameter and ε is a random *error* term that is assumed to be normally distributed with mean zero and variance σ^2.

TABLE 3 ANOVA Table for Example

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	1	250709	250709	25.12	0.001
Residual	8	79841	9980		
Total	9	330550			

7.1.1 The slope estimate b_1 is calculated as:

$$b_1 = \sum_{i=1}^{n} X_i Y_i / \sum_{i=1}^{n} X_i^2 \tag{27}$$

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

$$\hat{Y}_i = b_1 X_i \tag{28}$$

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

$$e_i = Y_i - \hat{Y}_i \tag{29}$$

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

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-1)} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{(n-1)} \tag{30}$$

$$s = \sqrt{s^2} \tag{31}$$

These estimates have $n - 1$ degrees of freedom because of prior estimation of the slope of the line, which removed one degree of freedom from the data set of n data points prior to calculation of the residuals.

7.1.5 The standard error for the slope estimate is:

$$S_{b1} = s / \sqrt{\sum_{i=1}^{n} x_i^2} \tag{32}$$

The 100(1 – α)th two-sided confidence interval for the slope β_1 is calculated as:

$$b_1 \pm t s_{b1} \tag{33}$$

where t is the $(1 - \alpha/2)$th quantile of the t distribution with $n - 1$ degrees of freedom. The confidence bands for the line are also straight lines with zero intercepts having slopes defined by the confidence limits on the slope (see Fig. 11).

7.2 *Example*—An experiment was conducted to determine an instrument response over a range of 0 to 10 mg/L of a substance dissolved in a solvent. Five solution standards at 2, 4, 6, 8, and 10 mg/L concentrations were run in duplicate and the results are shown in Fig. 10. A zero-intercept model was considered because the data points appeared to lie in a straight line that approached the origin.

TABLE 2 ANOVA Table Calculations

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F* statistic	*p*-value
Regression	1	$SSR=\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$	$MSR = SSR / 1$	$F = MSR / MSE$	p
Residual	$n - 2$	$SSE=\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$	$MSE = SSE / n - 2$		
Total	$n - 1$	$SST=\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$	$MST = SST / n - 1$		

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E3080-19
https://standards.iteh.ai/catalog/standards/sist/0ab-5726-470c-9570-ed1f632d3/astm-e3080-19