Designation: 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, health, and 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

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

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**

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

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

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

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

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

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

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, 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)

[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, 2017. Published January 2018. Originally approved in 2019. Last previous edition approved in 2016 as E3080 – 16. DOI: 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.

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 statistical relationships between two or more variables Ref. (**1, 2**).[3] 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 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. Simple Linear Regression Analysis

5.1 *Simple Linear Regression Model:*

5.1.1 Select the response variable Y and the predictor variable X. The predictor X is assumed to have known values with little or no measurement error. The response Y has a distribution of values for a given X value, and this distribution is defined for all X values in a given range.

5.1.2 The regression function for the straight line relationship is $Y = \beta_0 + \beta_1 X$. 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.

5.1.3 The statistical distribution for Y is 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 $Y = \beta_0 + \beta_1 X + \varepsilon$, where ε is a random error that is normally distributed with mean zero and standard deviation σ (variance σ^2).

5.1.4 An example of a linear regression model is depicted in Fig. 1 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.

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

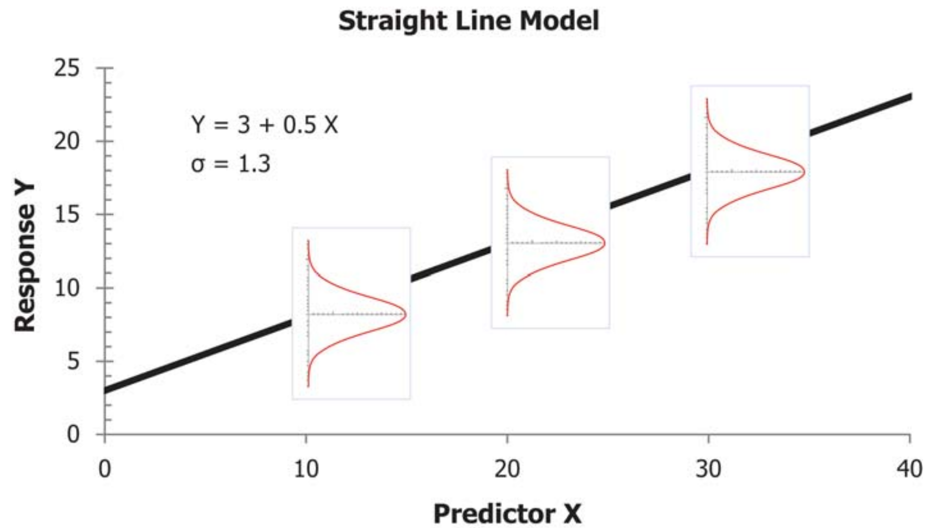

FIG. 1 Graphical Depiction of a Straight Line Regression Model

5.2 *Estimating Regression Model Parameters:*

5.2.1 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.

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

5.2.3.1 Calculate the averages of X and Y:

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

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

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.

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:

(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 *Evaluation of the Model:*

5.3.1 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.

5.3.2 *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

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			

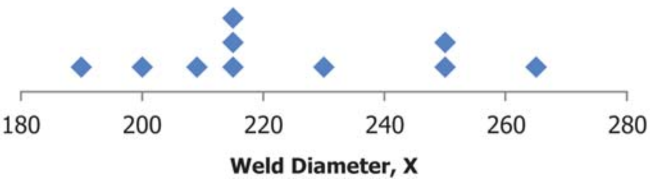

FIG. 2 Dot Plot of the Predictor Value *X*

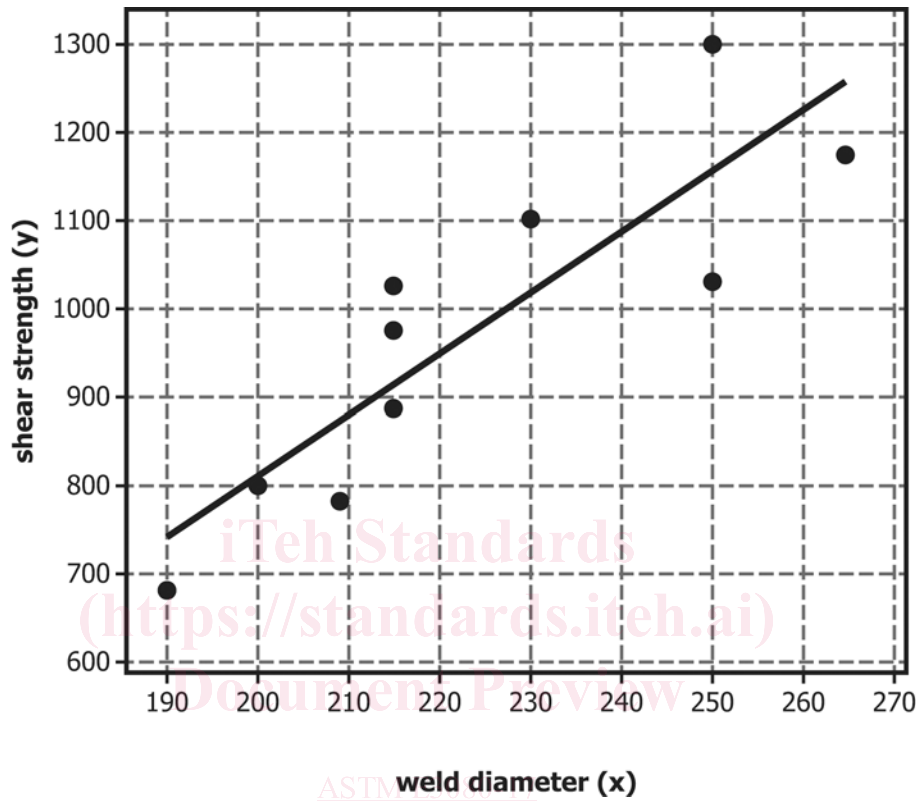

FIG. 3 Scatter Plot of Data with Fitted Linear Model

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:

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

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. 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.

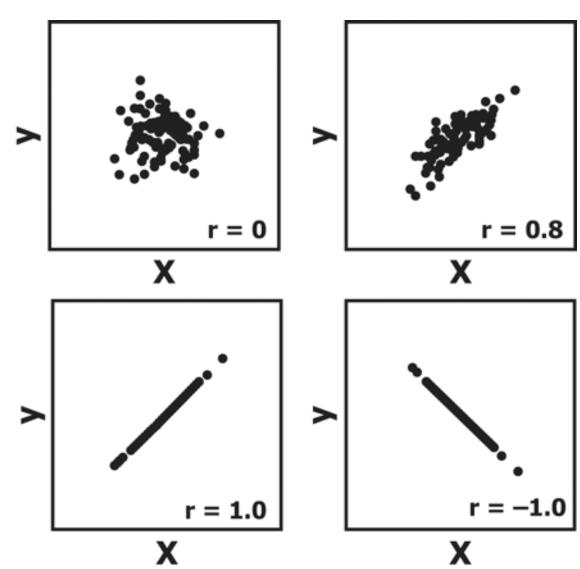

FIG. 4 Typical Scatter Plots for Selected Values of the Correlation Coefficient, *r*

5.3.3 *Residual Plots:*

5.3.3.1 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.

5.3.3.2 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. 5. 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.)

5.3.3.3 Plot of residuals against X (or equivalently against $\hat{Y}_i$) will detect certain departures from the assumptions. Residuals may also be plotted against time of testing (if available) or against another variable. Fig. 6 shows some of these patterns, and Appendix X1 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 best pattern – indicates no model deficiencies.

(2) Plot B – increasing variance with X, consider weighted regression or data transformations (see X1.4.4).

(3) Plot C – curvature in the relationship, consider adding a quadratic term or using a nonlinear model (see X1.4.1).

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

The residuals plot for the example in Fig. 7 indicates no obvious curvature, and a slight tendency for an increase in vertical scatter with increasing X, but more data points would be necessary to confirm this.

5.3.3.4 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.

5.4 *Use of the Model for Interval Estimates of Regression Parameters and Predicted Values:*

5.4.1 The estimates b_0 and b_1 of their respective model parameters β_0 and β_1 are point estimates. These estimates have minimum variance among unbiased estimates without specifying the distribution of Y. To give a sense of the precision for these estimates, interval estimates, or confidence intervals, can be provided. For these interval estimates, the form of the statistical distribution is required, and the normal distribution is usually specified, as in 5.1.3. 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 n and the values of the X_i. The confidence level is chosen. 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 in the following examples in this section.

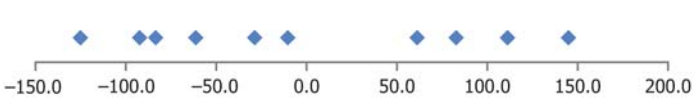

FIG. 5 Dot Plot of Residuals

5.4.2 A general form for the *confidence interval* for a point estimate E is:

$$E \pm t s_E \qquad (13)$$

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 confidence level. Practice E2586 provides description of confidence intervals, standard error, and degrees of freedom. 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 is the upper $(1 - \alpha/2)^{th}$ quantile of the Student's t distribution with $n - 2$ degrees of freedom, for a confidence level of $100(1 - \alpha)\%$. The value of t for a 95 % two-sided confidence interval with 8 degrees of freedom is 2.306.

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.

5.4.3 The standard error for the slope estimate is:

$$s_{b1} = s / \sqrt{S_{XX}} \qquad (14)$$

From the example:

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

The confidence interval for the slope β_1 is calculated as:

$$b_1 \pm t s_{b1} \qquad (15)$$

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

5.4.4 The standard error for the intercept estimate is:

$$s_{b0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}} \qquad (16)$$

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: