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

2. Referenced Documents

2.1 *ASTM Standards:*[2]

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**

3.1.2 *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 $1 - \alpha$, where $\Pr(L \le \theta \le U) \ge 1 - \alpha$. **E2586**

3.1.3.1 *Discussion*—The confidence level, $1 - \alpha$, 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, $1 - \alpha$, of the probability associated with a confidence interval, often expressed as a percentage. **E2586**

3.1.4.1 *Discussion*—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.

3.1.10 *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**

---

[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. Published November 2016. DOI: 10.1520/E3080-16.

[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.

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.

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

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

3.1.22 *quantile, n*—value such that a fraction *f* of the sample or population is less than or equal to that value. **E2586**

3.1.23 *sample, 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. **E2586**

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

3.1.25 *sample statistic, n*—summary measure of the observed values of a sample. **E2586**

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.

3.1.27 *standard deviation—of a population*, σ, the square root of the average or expected value of the squared deviation of a variable from its mean; *—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. **E2586**

3.1.28 *variance, σ^2, s^2, n*—square of the standard deviation of the population or sample. **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.

4. Significance and Use

4.1 Regression analysis is a statistical procedure that studies the relations between two or more numerical variables and utilizes existing 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 (**1, 2**).[3]

5. Straight Line Regression and Correlation

5.1 *Two Variables*—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 *Y* is always to be treated as a random variable. *X* may be either a random variable sampled from a population with an error that is negligible compared to the error of *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 *X* as the independent variable and *Y* as the dependent variable.

5.1.2 The practitioner typically wants to see if a relationship exists between *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 $Y = \alpha + \beta X + \varepsilon$, where α and β are model coefficients and ε is a random error term representing variation in the observed value of *Y* at given *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, *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^2 throughout the range of *X* and *Y*. Further, the errors are uncorrelated with each other. This will be assumed throughout the remainder of this section.[4]

5.1.3 The regression problem is to determine estimates of the coefficients α and β that "best" fit the data 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 *X* and *Y*. The sample correlation coefficient, *r*, is the estimate of ρ. The square of the correlation coefficient, r^2, is called the coefficient of determination and has additional meaning for the linear relationship between *X* and *Y*.

5.1.4 When a suitable model is found, it may be used to estimate the mean response at a given value of *X* or to predict the range of future *Y* values from a given *X*.

5.2 *Method of Least Squares*—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 *i*th 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

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

[4] The normal distribution of the error structure is not required to fit the linear model to the data but is required for performing standard model analysis such as residual analysis, confidence and prediction intervals and statistical inference on the model parameters.

y_i values. Note that these are vertical distances. Analytically, this sum of squared deviations is of the form:

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

5.2.2 The sum of squares, S, is written as a function of a and b. Minimizing this function involves taking partial derivatives of S with respect to a and b. This will result in two linear equations that are then solved simultaneously for a and b. The resulting solutions are functions of the (x_i, y_i) paired data.

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 S_{XX} and S_{YY} and the sum of cross products 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 $\sum_{i=1}^{n} (x_1 - \bar{x})\bar{y} = 0$.

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

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

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

$$a = \bar{y} - b\bar{x} \tag{6}$$

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:

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

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.

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

5.4.1 An alternative formula for r uses the standard deviation of the paired differences ($d_i = y_i - x_i$). Note that it does not matter in what order we calculate these differences. Either $d_i = y_i - x_i$ or $d_i = x_i - y_i$ will give the same result:

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		

parameter estimates	
b	6.898
a	–569.468

S_{XX}	5,268.900
S_{YY}	330,550.000
S_{XY}	36,345.000

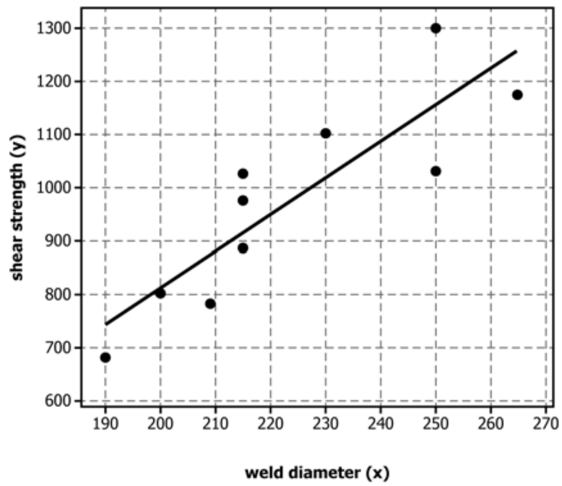

FIG. 1 Scatter Plot of Table 1 Data with Fitted Linear Model

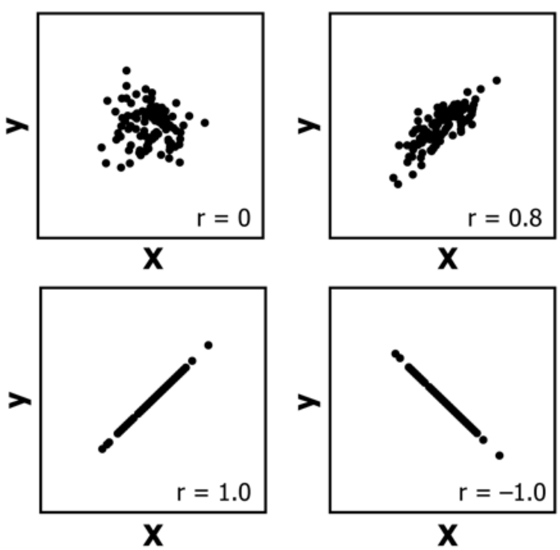

FIG. 2 Typical Scatter Plots for Selected Values of the Correlation Coefficient, *r*

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

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

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

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

5.4.2 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*.

5.4.3 An alternative formula for the estimated slope *b* as a function of the correlation coefficient, *r*, and standard deviations of the variables *X* and *Y* is:

$$b = \frac{r s_y}{s_x} \tag{10}$$

5.5 *Residuals*—For any specified x_i in the data set, the *residual* at x_i is the difference $e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$, the difference between the observed value of *Y* and the predicted value (the $\hat{y}_i$ value) at the observed value of *X*. The e_i term estimates the true random error term ε_i from the theoretical linear model (see 5.1.1). The predicted values of *Y* are computed using the estimated model equation $\hat{y}_i = a + bx_i$.

5.5.1 The residuals for a straight line regression with slope and intercept fit by least squares will always sum to zero. The sample correlation coefficient between the residuals e_i and the values of the independent variable, x_i, or the estimated values, $\hat{y}_i$ will also be zero.

5.5.2 The estimate of the *residual error variance*, σ^2, is calculated either using the squared residuals or from intermediate quantities S_{YY} and S_{XY} and *b*:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2} = \frac{S_{YY} - bS_{XY}}{n - 2} \tag{11}$$

5.5.3 The square root of Eq 11 is the estimate of the *unknown error standard deviation*, σ. The estimated values $\hat{y}_i$ require that we estimate two model parameters for their calculation, which results in a loss of two degrees of freedom in the denominator (*n* – 2).

5.5.4 *Example*—For the example, the estimate of σ is calculated using the information from Table 2:

Using Eq 11:

$$\hat{\sigma} = \sqrt{\frac{330{,}550 - 6.898\,(36{,}345)}{10 - 2}} = 99.9$$

Using the residual error estimated σ, the standard errors (*se*) for the estimates of slope and intercept may be calculated:

TABLE 2 Calculate the Estimate of σ

i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	680	741.16	–61.16	3,740.18
2	800	810.14	–10.14	102.76
3	780	872.22	–92.22	8,504.42
4	885	913.61	–28.61	818.39
5	975	913.61	61.39	3,769.03
6	1025	913.61	111.39	12,408.27
7	1100	1017.08	82.92	6,876.07
8	1030	1155.04	–125.04	15,634.61
9	1175	1258.51	–83.51	6,973.86
10	1300	1155.04	144.96	21,013.86
			SUM	79,841.31
			$\hat{\sigma}$	99.9

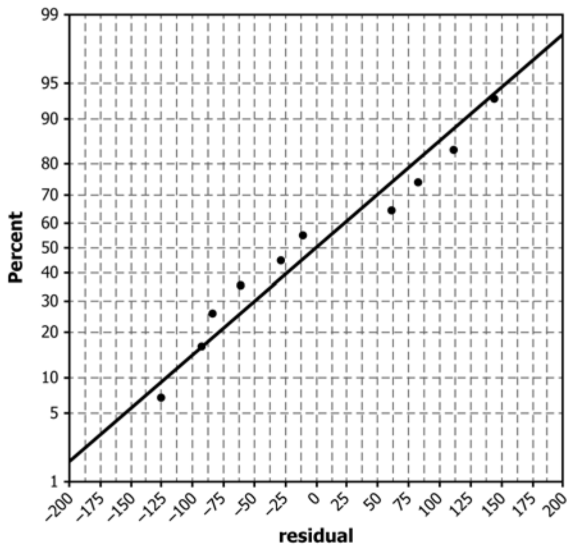

FIG. 3 Normal Probability Plot of Residuals

$$se(b) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \quad (12)$$

$$se(a) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \quad (13)$$

Standard errors for the slope and intercept for the example are:

$$se(b) = \frac{99.9}{\sqrt{5268.9}} = 1.376$$

and:

$$se(a) = 99.9\sqrt{\frac{1}{10} + \frac{223.9^2}{5268.9}} = 309.8$$

5.6 Recall that the assumed model is $Y = \alpha + \beta X + \varepsilon$. Generally, the ε terms are assumed independent and normally distributed with mean 0 and variance σ^2. If the data are true to these assumptions, the residuals will follow a normal distribution and be consistent with respect to time order (that is, be in a state of statistical control). A broad collection of diagnostic tools are available for performing residual analysis as well as other model diagnostic tasks. A few of the key tools are described below and illustrated in Figs. 3 and 4 using the data from the example.

5.6.1 *Probability Plots*—The residuals should be checked for an approximate normal distribution using a normal probability plotting technique. Various types of residuals may be calculated (for example, ordinary, standardize, and deleted *t*).

5.6.2 *Control Charts*—Residuals can be plotted on a control chart for individuals and moving ranges. This technique is checking for statistical control of the residuals.

5.6.3 In addition, residuals can be plotted against the independent variable. This is checking for homogeneity of variance across values of the independent variable.

5.7 The population coefficient of determination, ρ^2, is the square of the correlation coefficient ρ. The sample coefficient of determination is the square of *r*. The interpretation of r^2 is as the fraction reduction in the variance of *Y* from knowledge of *X* in advance. In other words, the fraction of the total variation in *Y* explained by the model and therefore removed by the linear trend. This interpretation is derived from a relation between the residual variance, σ^2, and overall variance of *Y*. This interpretation is mainly useful when values of *X* are sampled from a population and less useful when values of *X* are selected in a designed experiment.

5.7.1 *Example*—Using *r* from the above calculation, the sample coefficient of determination for the example is r^2 = 0.8712^2 = 0.759, to 3 significant digits. This means that approximately 76 % of the variation in *y* is explained by the model we are using. If the variance in the *Y* values is calculated and compared to the residual variance (see 5.4), then the ratio of the residual variance to the *Y* variance will be approximately (100 – 76) % = 24 %.

5.8 Uses of the linear regression model for calculating confidence intervals for the mean response and prediction intervals for a future response depend on the residuals being normally distributed and independent.

5.8.1 *Mean Value Estimates*—The estimated mean value, $\hat{y}$, at a specific value of the independent variable, say x_0, is determined using the fitted model $\hat{y} = a + bx_0$ directly. To construct a confidence interval for the mean response at a specific x_0, use the following form:

$$a + bx_0 \pm t_{1-\alpha/2,\, n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}} \quad (14)$$

The estimate $\hat{\sigma}$ is calculated first using Eq 11, then taking the square root; $t_{1-\alpha/2}$ is a positive Student's *t* quantile with *n* – 2 degrees of freedom that leaves a probability α/2 to the right. Quantities $\bar{x}$ and s_x are the sample mean and standard deviation of the *X* values in the data set.

5.8.2 *Example*—Suppose we are interested in the mean response at the specific value x_0 = 215. The estimated mean response at x_0 = 215 is:

$$\hat{y}_0 = -569.468 + 6.898(215) = 913.6$$

A 95 % confidence interval for the mean response at x_0 = 215 is:

$$\begin{aligned} &-569.468 + 6.898(215) \\ &\pm 2.306(99.9)\sqrt{\frac{1}{10} + \frac{(215 - 223.9)^2}{(10-1)(24.196)^2}} \\ &= 913.6 \pm 78.13 \end{aligned} \quad (15)$$

Thus the expected response at x_0 = 215 falls between 835.47 and 991.73 with 95 % confidence.

5.9 *Prediction Intervals*—For a specified value of the independent variable, x_0, we can also determine a prediction interval for a future response. The future response is unobserved and that is the point of the prediction. The standard formula for the prediction interval in this case is identical to Eq 14 with the addition of the number "1" inserted under the radical. This form is:

$$a + bx_0 \pm t_{1-\alpha/2,\, n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}} \quad (16)$$

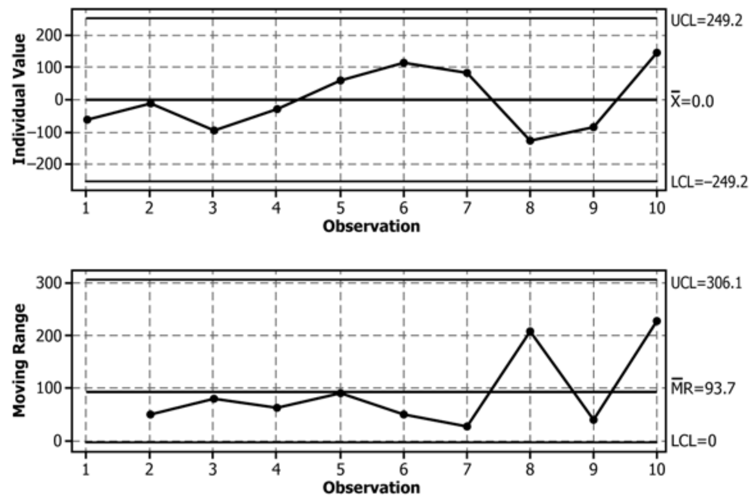

FIG. 4 Control Chart for Residuals

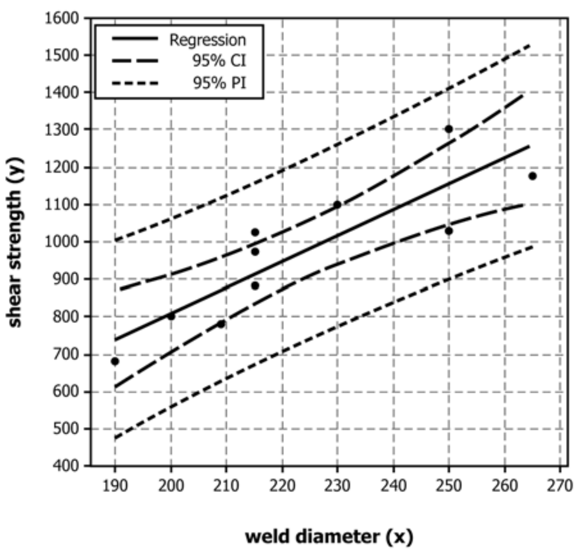

FIG. 5 Regression Plot with 95 % Confidence and Prediction Intervals

In simplest form, a prediction interval is an interval estimate in which one or more future observations would fall, with a certain probability, given what has already been observed. There are many variations on this theme. Prediction intervals are substantially wider than confidence intervals because a prediction interval applies to an individual value whereas a confidence interval applies to the mean response. Prediction intervals are often used in regression analysis.

5.9.1 *Example*—The prediction for a future value at the specific x_0 = 215 using 95 % confidence:

$$-569.468+6.898(215)$$
$$\pm 2.306(99.9)\sqrt{1+\frac{1}{10}+\frac{(215\ -\ 223.9)^2}{(10\ -\ 1)(24.196)^2}}$$
$$=913.6\pm 243.26$$

The 95 % prediction interval is between 670.34 and 1156.86.

5.9.2 Confidence and prediction interval limits are often plotted together on the scatter plot. This display is shown in Fig. 5 for the example.

6. Keywords

6.1 bivariate; correlation; least squares; regression

REFERENCES

(1) Draper, N., and Smith, H., *Applied Regression Analysis*, 2nd ed., John Wiley & Sons, 1981.

(2) Neter, J., et al, *Applied Linear Statistical Models*, 4th ed., Richard D. Irwin, Inc., 1989.

(3) Duncan, A. J., *Quality Control and Industrial Statistics*, 5th ed., Irwin, Homewood, IL, 1986.