Designation: E3080 – 16

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 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 $\Pr(L \leq \theta \leq U) \geq 1 - \alpha$. **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:

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

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