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.

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.

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

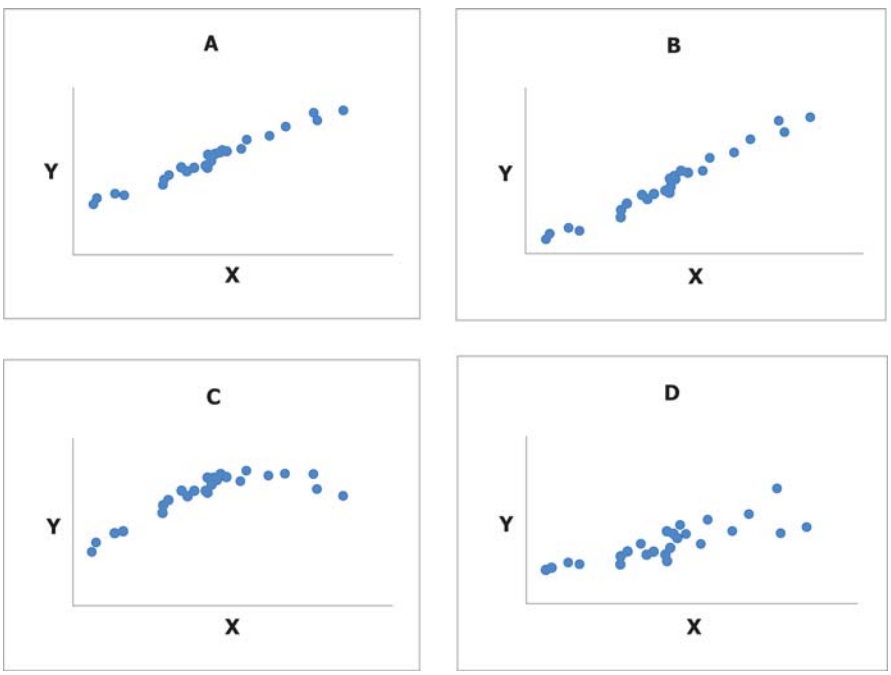

FIG. 1 Scatter Plots

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 $\sigma = 1.3$ 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

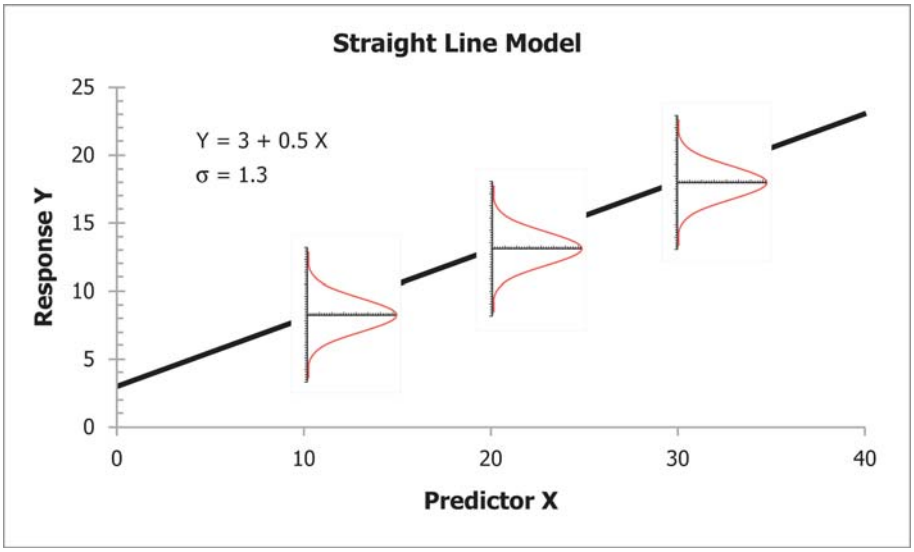

FIG. 2 Graphical Depiction of a Straight Line Regression Model