**Designation: E3080 – 23**

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, $\beta_0$, 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 Nov. 1, 2023. Published November 2023. Originally approved in 2016. Last previous edition approved in 2019 as E3080 – 19. DOI: 10.1520/E3080-23.

[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, $\sigma^2$, n—of a regression model*, the variance of the residuals (see *residual*).

3.2.5 *slope, $\beta_1$, 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) |
| $\beta_0$ | = | intercept parameter in model (5.3.1) |
| $\beta_1$ | = | slope parameter in model (5.3.1) |
| $\beta_{11}$ | = | 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) |
| $\varepsilon$ | = | 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) |
| $\lambda$ | = | Box-Cox parameter (A1.5.2) |
| $n$ | = | 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) |
| $\sigma$ | = | residual standard deviation (5.4.1) |
| $s$ | = | estimate of $\sigma$ (6.2.6) |
| $\sigma^2$ | = | residual variance (5.4.1) |
| $s^2$ | = | estimate of $\sigma^2$ (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.