Model Quality

A Primer on Regression Model Diagnostic Methods

By Jennifer Brown

Q How good is my linear-regression model?

A The practice for regression analysis with a single predictor variable (E3080) was revised in 2019 to emphasize the practical application of simple linear regression. It now steps the user through the procedure for performing a linear-regression analysis between two numerical variables for the purpose of predicting one variable from the other, placing greater focus on the model, estimation of model parameters, model evaluation, and methods for predicting a mean value or future value.

A vital step in performing a regression analysis is evaluating the model, specifically, the significance of the predictor variable; how well the underlying analysis assumptions are met; and how well the predictor variable explains the variation in the response. If left unchecked, any predictions based on the model or insight into potential process knowledge or controls may be misleading.

Here are some simple graphical and computational model diagnostic methods described in E3080 that are typically used to evaluate a simple linear-regression model. Some of these methods can be extended to linear-regression models with more than one predictor variable.

First, let's look at the simple linear-regression model and underlying assumptions.

A simple linear-regression model takes the form of the straight-line regression function $Y = \beta_0 + \beta_1 X$ where X is the predictor variable; Y is the response variable; β_0 is the intercept, representing the value of Y when $X = 0$; and β_1 is the slope, representing the change in Y for a unit change in X. Given a set of paired values (X, Y), E3080 describes how to calculate the slope and intercept.

It is not enough, however, to merely calculate the slope and intercept. For the model to be useful, a relationship must exist. That is, the slope must be different enough from 0 to say that X is a statistically significant predictor of Y. Though this does not directly imply a cause-effect relationship between the two variables, it does indicate that X explains some portion of the variation observed in Y.

One method described in E3080 to assess how different the slope is from 0 is to construct a 95% confidence interval around the estimated value for the slope. If the confidence interval does not contain 0, then X is considered a statistically significant predictor of Y.

Assessing the significance of the predictor variable is one critical step in establishing a predictive model. Another is to verify that the model adequately represents the relationship between X and Y. When plotted against the data, the model should fit through the middle of the data, with the observed response values randomly falling either above or below the model. This variation represents random error. For each observed response value Y, an error value can be calculated as the observed value minus the predicted value based on the model. This value is often referred to as the residual.

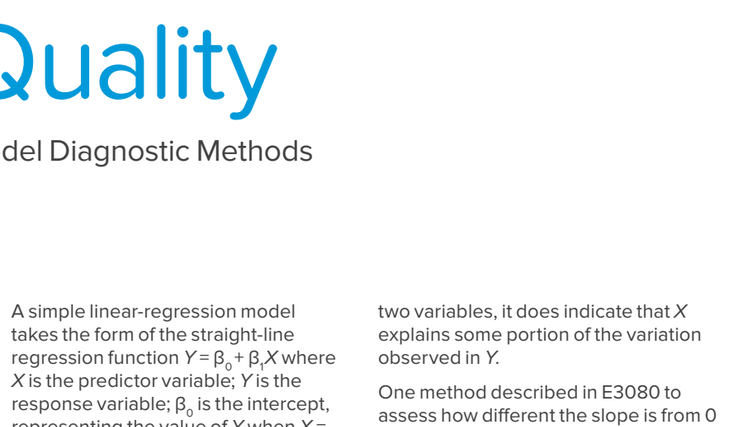

Figure 1 — This plot illustrates the underlying assumptions in simple linear-regression analysis. It appears as Figure 2 in E3080.

The collection of residual values is used to evaluate the underlying assumptions in simple linear regression: 1) the distribution of the residuals is normally distributed with mean 0 and constant variance σ^2; 2) the residuals are independent; and 3) the relationship is a straight-line regression function. A graphical depiction of these assumptions is shown in **Figure 1.**

A plot of the residuals against their X values, known as a residual plot, is used to evaluate the linearity, constant variance, and independence assumptions.

Figure 2 shows four example residual plots, where e denotes the residual.

Plot A shows the desired pattern of random variation around 0 over the range of X values that indicates the assumptions of linearity and constant variance hold. If the model runs straight through the middle of the relationship and the variation around the model is constant over the range of X values, then the residuals should vary randomly around 0 within a given range when plotted against their X values. In other words, the pattern in the residual plot will mimic how the model fits the data.

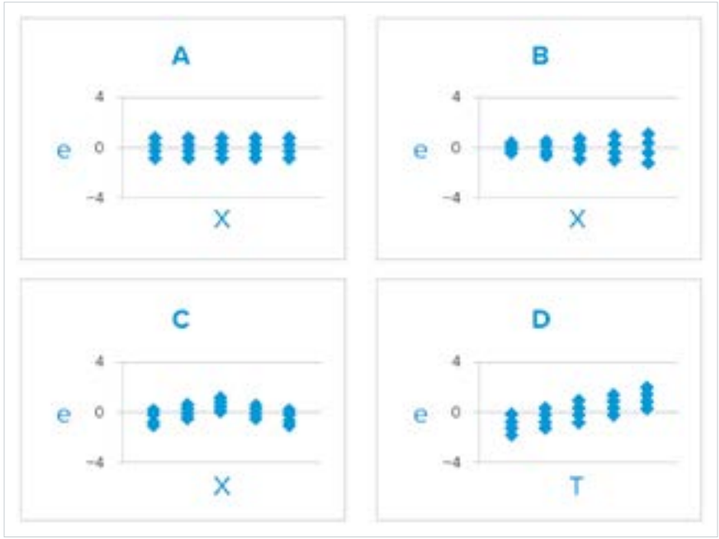

Figure 2 — This plot shows example patterns that may be observed in residual plots. It appears as Figure 3 in E3080.

If, for example, as X increases, the observed values fall below the model; then above the model; then below the model; the residual plot will exhibit a "rainbow" pattern as in Plot C, indicating the assumption of linearity does not hold. In some cases, the linearity assumption may hold, but the variation around the model is not constant. Plot B shows increasing variation as X increases. Though the presence of non-constant variance may or may not be practically significant, it should not be ignored, as it can result in misleading interval estimates for predicted values since the statistical interval formulas assume constant variance. If test order is available and a plot of the residuals against test order shows a pattern like in Plot D, then the assumption of independent residuals does not hold. A residual plot also serves as a visual tool for detecting outliers, which may also negatively impact the model.

The assumption of normally distributed residuals centered around 0 can be visually assessed with a histogram. If the normality assumption holds, the histogram should be roughly bell-shaped and symmetric about 0.

If any one of the underlying regression assumptions does not hold or influential outliers are present, the model may not be adequate. However, all is not lost. E3080 includes a discussion of potential remedial measures. Hence, regression analysis may be an iterative process to ensure all underlying model assumptions hold.

The final step in assessing model adequacy is to evaluate how well the significant predictor variable X explains the variation observed in Y. For simple linear regression, this is evaluated by looking at the coefficient of determination. Denoted r^2, this value is the square of the correlation coefficient, r, and is often expressed as a percent. A value close to 100% indicates that X explains almost all the variation in Y. A low value may indicate that there are other variables in addition to (or in lieu of) X that may help (or better) explain the variation in Y. If the predictor variable is insignificant, this will often be reflected by a value close to 0. If the underlying regression assumptions do not hold, the interpretation of r^2 is meaningless.

In summary, the model resulting from a simple linear-regression analysis is considered "good" and is of significant practical use only if the:

- Predictor variable is significant,
- Underlying model assumptions are met, and
- The predictor variable explains a large percentage of the variation observed in the response.

Because commercial software is readily available to perform the necessary calculations, the primary emphasis should be placed on model evaluation to ensure a quality model is produced. ■

Jennifer Brown worked as a statistician in the aerospace industry for over 12 years and is now an independent statistical consultant. She is a member of the committee on quality and statistics (E11) and chair of its subcommittee on terminology (E11.70). She is also a member of the subcommittee on specialized nondestructive testing methods (E07.10) and the technical contact for two of its data-related standards..

John Carson, Ph.D., senior statistician for Neptune and Co., is the Data Points column coordinator. He is a member of the committee on quality and statistics (E11) and a member of the committees on petroleum products, liquid fuels, and lubricants (D02); air quality (D22); and environmental assessment, risk management, and corrective action (E50).