How to Handle Nonconstant Variance in a Linear Regression Analysis

By Katie Daisey and Jennifer Brown

Q What is nonconstant variance and why is it a concern when model fitting?

A One of the underlying assumptions in linear regression is that of constant variance. But what exactly does this mean, when might it be violated, why is it important to check for constant variance, and can it be accounted for in a linear regression analysis if present?

A linear regression model includes two important sub-models: (1) the model connecting the expected value of the response variable, Y, with one or more independent predictor variables, and (2) the model describing the variance in Y for a given value of the predictor variable. Consider the theoretical simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$. Sub-model (1) is $\beta_0 + \beta_1 X$, which connects Y with the single predictor variable, X, through a linear function. Sub-model (2) is ε, often referred to as the error term, which models the variance of Y for a given value of X and is assumed to be normally distributed with mean 0 and constant variance σ^2 in a linear regression analysis.

Sub-model (2) relates to the data and fitted model in the following way. Let the fitted model be represented by $\hat{Y}_i = b_0 + b_1 X_i$, where Y_i is the i^{th} observed value with a predictor variable value of X_i, b_0 is the estimated value of the model intercept, b_1 is the estimated value of the model slope, and $\hat{Y}_i$ is the predicted value for Y_i based on the model. As illustrated in Figure 1 [E3080], the variance in Y for a given value of X is assumed follow a normal distribution with spread that remains constant at different values of X but with a center near the predicted value $\hat{Y} = b_0 + b_1 X$. In other words, the center of the normally distributed error around the model is changing as the value of X changes but the spread of Y stays the same at different values of X.

Though an error term does not appear in the fitted model, an error value, often referred to as a residual value, can be calculated for each observed value as the difference between the observed value and predicted value based on the model, or $e_i = Y_i - \hat{Y}_i$. The collection of residual values is used to estimate the variance of Y at a given X value. That is, $s^2 = \sum_{i=1}^{n} e_i^2 / (n-1)$. The square root of the value s^2 is then used to estimate the constant standard deviation value, σ.

Figure 2 shows examples of constant and non-constant variance in observed Y values for given values of a predictor variable, X. Plots A and B show observations taken at discrete values of X, which could illustrate data collected from a planned experiment, for instance. The Y values in plot A exhibit constant variance as X increases, while the Y values in plot B exhibit increasing variation as X increases. Plot C shows data for Y across a range of values for X, which could illustrate observational data, for example. The observed values for Y in plot C exhibit increasing variance as X increases. (Note that in some cases the observed values in Y may exhibit decreasing variance as X increases.)

There are many situations where the variance in Y is expected to be constant across the range of X values. For example, the Y values in Plot A could represent experimental data collected on a measuring device that is expected to have constant error (i.e., constant variance) or near-constant error within the range of the measurement instrument's capability (X values), often referred to as "working range." There

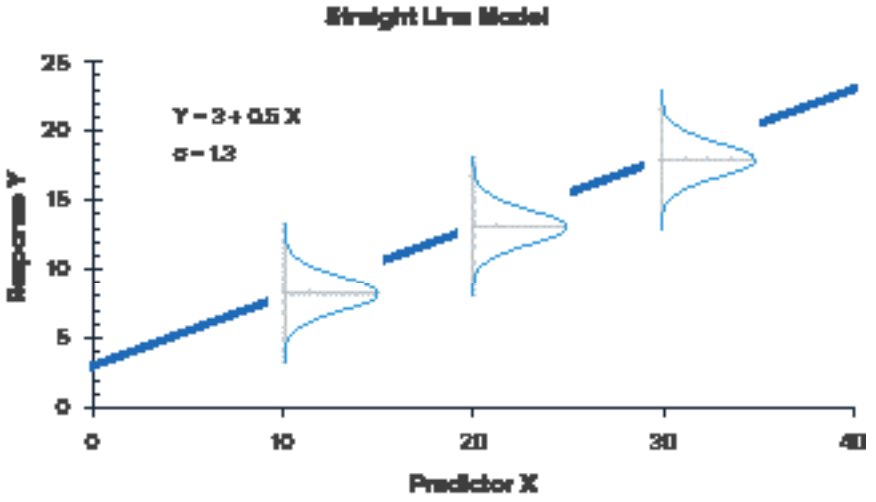

Figure 1 – Illustration of the normality and constant variance assumption in linear regression analysis, which appears as Figure 2 in E3080.

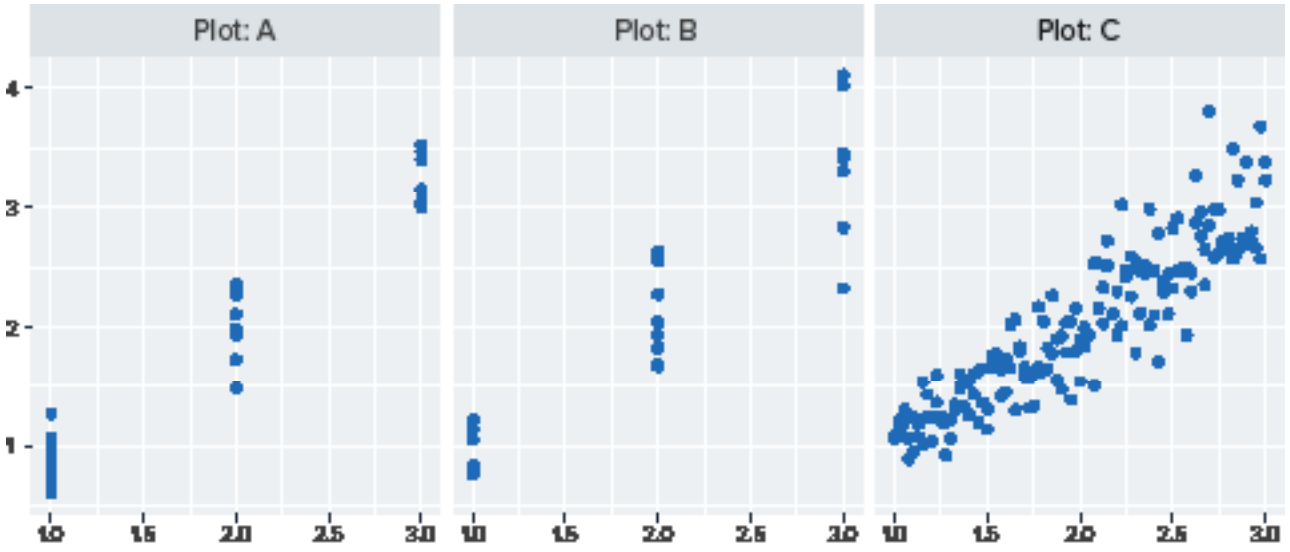

Figure 2 – Plot A illustrates constant variance. Plots B and C illustrate non-constant variance.

are also situations in which the variance is not expected to be constant. The Y values in Plot B, for example, could represent measured lengths of bolts manufactured to various nominal lengths (X), where bolts with longer nominal length may be expected to have higher variance in their final manufactured length (Y) compared to bolts with shorter nominal length.

Sometimes non-constant variance is obvious in the data and sometimes it is subtle. Formal statistical tests for non-constant variance are available and are generally recommended when performing a linear regression analysis. If variation is expected to be constant but the data or formal statistical tests suggest otherwise, this may be indicative of another variable acting on the process that needs to be accounted for.

It is important not to overlook the assumption of constant variance since the presence of non-constant variance can have a significant impact on variance estimates and, consequently, any inference made using confidence intervals or prediction intervals. The assumption of constant variance results in hyperbolic-shaped confidence bounds and prediction bounds, which roughly parallel the model but widen a bit as the distance from the data center $(\bar{X}, \bar{Y})$ increases [E3080]. If variation in Y increases (or decreases) as the value of X increases, the bounds will not reflect this, and erroneous inferences will be made. Figure 3 shows prediction bounds around two models of the same data, one resulting from a linear regression analysis that accounts for non-constant variance (solid lines exhibiting the megaphone shape) and the other resulting from a linear regression analysis that ignores the non-constant variance observed in the data (dashed lines roughly parallel to model). Any inference made using the constant variance prediction bounds for larger or smaller values of X in this case will be highly inaccurate. Hence, if non-constant variance is present, it must be accounted for.

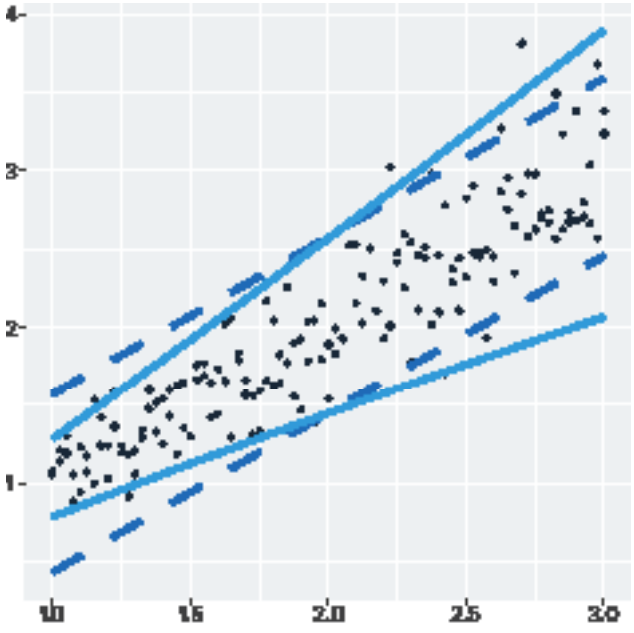

Figure 3 – Illustration of a prediction interval resulting from a linear regression analysis that accounts for non-constant variance (solid lines exhibiting the megaphone shape) and one resulting from a linear regression analysis that ignores non-constant variance (dashed lines roughly parallel to model).

E3080 discusses some of the common methods used in linear regression analysis to account for non-constant variance such as transformation of variables. Alternate methods, such as weighted least squares regression, may also serve as a remedial measure. In addition to constant variance, there are other important assumptions in linear regression analysis that must be verified. The reader is encouraged to see E3080 for information on all underlying assumptions, practical methods for verifying the assumptions hold, and remedial measures to try if one or more assumptions are violated. ■

Katie Daisey, Ph.D., is a scientist at Arkema Inc., supporting R&D and manufacturing in the areas of statistics, chemometrics, and digital transformation. Daisey currently serves as chair of the committee on quality and statistics (E11).

Jennifer Brown is a statistician with 17 years of experience in the aerospace industry. She is chair of the subcommittee on terminology (E11.70) and a member of the subcommittee on specialized NDT methods (E07.10).

John Carson, Ph.D., is senior statistician for Neptune and Co. and coordinator of Data Points. He is a member of the committees on quality and statistics (E11), petroleum products, liquid fuels, and lubricants (D02), air quality (D22), and more.