SAS

Closing Thoughts on Multiple Linear Regression

Multiple linear regression is a broad and complex topic, and we have only scratched the surface in this lesson.

In the closing exercise, we'll provide an opportunity to practice fitting a multiple linear regression model using the body fat data. In this practice exercise, you'll be asked to explore the variables to understand potential relationships and data quality issues, fit a regression model using all the predictors, conduct a residual analysis, explore influential observations, assess multicollinearity, and select the important predictors.

The Body Fat data are relatively "clean," and the data set provides a nice opportunity to practice your regression modeling skills. In reality, regression modeling is rarely this straightforward, and you might face a number of complications.

To close, here are some things to consider when you use multiple linear regression modeling in the context of problem solving.

Do you have enough data to fit a particular model? One rule of thumb is that you need at least 10 observations per predictor. However, if you have categorical predictors with many levels, you might need much more data to estimate all the model coefficients.

Do you need to include polynomial terms or transformations as model effects, or do you need to transform the response? A residual analysis might suggest that you need to fit a more complex model than you saw in this lesson, or you might have subject-matter expertise that suggests the need for polynomials or transformations. As always, we strongly suggest that you consult an expert for guidance.

Do you have high-quality data that represents the current state of the process? Data quality issues can stem from a number of potential causes. For example, is the measurement system capable? Are large numbers of records missing from some of the variables? Are the data too old to be useful? Is the source of the data well documented? And, do you have data for all the variables that might be important?

If data are of questionable quality and validity, then new data should be obtained before you perform a statistical analysis.