Common Issues

In the previous videos and exercise, we used the Metal Coatings scenario to fit models with main effects and two-way interactions. However, we can run into problems estimating model coefficients if we don't have enough data.

For this discussion, we return to the Impurity Logistic example. Let's say we believe there might be an interaction between the Reactor and the Shift. When we fit a logistic model for Outcome with the five main effects and the interaction between Reactor and the Shift, we see some unusual results.

The Parameter Estimates table reports that all of the coefficients for Reactor, Shift, and their interactions are Unstable. Notice that the standard errors for these estimates are all extremely large, and the chi-square test statistics for these coefficients are zero. Why does this happen? We only have 100 observations, which isn't a lot of data for a logistic model with categorical predictors. More importantly, we don't have data for all combinations of Outcome, Reactor, and Shift, so we can't properly estimate the interaction.

We can see the problem in this graph. When the Shift is 2 and the Reactor is 1, the Outcome is always Pass. And, when Shift is 1 and the Reactor is 1, only one observation has an Outcome of Fail. In order to estimate all of the model coefficients, we need to have a sufficient amount of data. But just how much data we need can get a bit complicated - it depends on the structure of the data set and the proportion of the observations in the target category. If you encounter issues like this, you might need to collect more data, you might need to eliminate a predictor, or you might be able to combine categories of a nominal predictor with many levels.

Fortunately, for this example, we have the underlying continuous response, Impurity. Earlier we fit a multiple linear regression model with all main effects and two-way interactions using Impurity, and we saw that we can easily estimate all of the model coefficients.

Remember that continuous variables have much more information content than categorical variables. In general, regression models with continuous responses require much less data than models with categorical responses.

Another common issue in logistic regression is called separation. This occurs if a variable is a perfect, or a near perfect, predictor of the response. In this simple example, when the Predictor is below zero the Response is always zero, and when the Predictor is above zero the Response is always one. The logistic curve is nearly a vertical line, separating the zeros from the ones. When this occurs, the slope coefficient approaches infinity, and it can't properly be estimated.

The Parameter Estimates table reports both the slope and the intercept as unstable.

There is a bright side to this situation for a problem-solving team. If a continuous predictor completely explains the categorical response, we might have found the root cause of the problem!

Close