Print

Multiple Linear Regression with Interactions

Earlier, we fit a linear model for the Impurity data with only three continuous predictors. This is what we'd call an additive model.

According to this model, if we increase Temp by 1 degree C, then Impurity increases by an average of around 0.8%, regardless of the values of Catalyst Conc and Reaction Time. The presence of Catalyst Conc and Reaction Time in the model does not change this interpretation. Likewise, if we increase Catalyst Conc by 1 unit, Impurity increases by around 2.1% on average, regardless of the values of Temp or Reaction Time.

But, it's possible that the effect of one predictor on Impurity is dependent on the values of another predictor. For example, the effect of Temp on Impurity might be dependent on the value of Catalyst Conc or Reaction Time, or both. This dependency is known in statistics as an interaction effect.

For illustration, consider a simple example involving the breaking strength of a tool at different speeds using two different materials. Notice that the slopes of the lines for Speed versus the response, Strength, are different for the two values of Material.

This is easier to see if we overlay the data with the fitted lines for the two materials on the same plot. At low values of Speed, Material 1 has higher breaking strength. But at high values of Speed, Material 2 has higher breaking strength. This is a classic interaction.

The effect of Speed on Strength depends on Material. And, if we flip this around, the effect of Material on Strength depends on Speed.

We can extend our model to account for this dependency by including an interaction term in the model. In the two-predictor case, the two-way interaction term is constructed by computing the product of X1 and X2.

Let's return to the Impurity example. We fit a model with the three continuous predictors, or main effects, and their two-way interactions. Because we have three main effects, there are three possible two-way interactions. The interaction between Catalyst Conc and Reaction Time is significant, along with the interaction between Temp and Reaction Time. However, the interaction between Temp and Catalyst Conc is not significant.

We can visualize these interactions using interaction plots. Each interaction plot in this matrix shows the interaction of the row effect with the column effect. For each pair of variables there are two interaction plots, enabling us to visualize the interactions from different perspectives.

Take for example, the interaction plots between Temp and Catalyst Conc. The slopes of the lines for Temp and Catalyst Conc are parallel. This means that, on average, the effect of Temp on Impurity doesn't change as Reaction Time increases, and vice versa.

The slopes of the lines in the interaction plots between Reaction Time and Temp are nonparallel. And the slopes of the lines for the interaction between Catalyst Conc and Reaction Time are also nonparallel.

The Prediction Profiler makes it easier to understand these interactions. Take for example, the interaction between Temp and Reaction Time. Do you see how the slope for Reaction Time changes as we change the value of Temp from the low level to the high level? Understanding these interactions is important because it provides additional insights into our response. What about the two categorical predictors, Reactor and Shift?

Earlier we fit a model with all five predictors. If we have enough data, and if it makes sense to do so, we can fit a model with all possible two-way interactions. But this is where things can get complicated, particularly if we have categorical predictors.

In this case, our model with all two-way interactions includes five main effects and 10 interactions. Many of these terms are not significant. Because Reactor has three levels, the model includes the intercept plus 19 parameter estimates.

RSquare Adj and Root Mean Square Error have both improved from our earlier models. But this model is overly complex, and we need to consider the goal of our analysis.

In explanatory modeling, we're primarily interested in understanding the important predictors, and in making statements about the coefficients for these predictors. We're not as concerned with making predictions about future performance of the system we are studying. In predictive modeling, we are interested in developing a model that predicts the response as accurately as possible.

For both of these goals, it makes sense to simplify our model to include only the most important effects. We discuss this topic of variable selection, or model reduction, in the upcoming videos.

Close