Print

Multiple Logistic Regression

For the Impurity Logistic example, we have three continuous predictors and two categorical predictors.

When we fit simple logistic models for Outcome, we see that Temp and Catalyst Conc are significant, but the model for Outcome and Reaction Time is not significant. What about the two categorical predictors, Reactor and Shift? The mosaic plot for Outcome and Reactor shows differences between the three Reactors, but there doesn't appear to be a difference between the two shifts.

As we saw with multiple linear regression, we can include many predictors in one model. The predictors can be continuous or categorical, and we can also include interaction terms. When we have multiple predictors and our response is categorical, the modeling procedure is called multiple logistic regression.

Here, we fit a model for Outcome, and include all of the main effects from Temp through Shift in the model. A test for the significance of the model as a whole is reported. Here, we are testing that the full model with the specified predictors is better than the reduced model with just the intercept and no predictors. The p-value indicates that model is highly significant. This tells us that at least one of the predictors is significant in predicting the probability that a batch will fail. Catalyst Conc, Temp, and Reactor are all highly significant. Shift is not significant, given the other terms in the model. When we remove Shift from the model, we see that the remaining predictors are all significant, with p-values less than 0.05. Recall that the simple logistic model for Outcome and Reaction Time was not significant. When we account for the information in the other predictors, Reaction Time is now a significant predictor of Outcome. How good is our model?

The misclassification rate for our model is 7%. This is our error rate. Only 7% of the observations are incorrectly classified by our model. Recall that our model with just Temp has a misclassification rate of 17%, so this model does much better. We can also state this in terms of correct classifications. Overall, the model correctly classified 93% of the observations. This is our accuracy rate. The model correctly classified 71 out of 74 passes, but is not quite as good at classifying the fails, catching only 22 out of 26 of them.

The coefficients in our logistic model are reported in the Parameter Estimates table. The Parameter Estimates table also reports test statistics and p-values for the model coefficients. Note that these p-values might be slightly different from the p-values reported in the Effect Summary table, because a different statistical test is conducted to test for model coefficients.

For more information on the statistical output reported in these tables, refer to the Read About It for this module.

To get a better understanding of the parameter estimates, we use the Prediction Profiler. The model predicts the probability of Fail at different values of the predictors. The starting values for the continuous predictors are the means for each predictor, and the starting value for Reactor is the first level, Reactor 1.

At these values, the probability of Fail is only 0.008. The slope of the line for Temp is flat, and then increases sharply. Holding the other factors constant, when Temp is between approximately 26 and 28 degrees Celsius, the probability of Fail is very low. But, when Temp increases beyond 28 degrees, the probability of Fail shows a sharp increase.

We see a similar pattern for Catalyst Conc. When we increase Catalyst Conc to 2.5, holding everything else constant, we can see that the probability of Fail increases from 0.008 to 0.831. Notice that the slopes of the lines for the other predictors have changed. For example, when Catalyst Conc is 2.5, increasing the Reaction Time will result in a lower probability of Fail.

In this example, we've fit a multiple logistic regression model with only main effects. We see how to fit this model in JMP in the next video.

Close