Print

Interpreting Logistic Regression Results

In the previous video, we used logistic regression to understand the relationship between Outcome and one predictor, Temp. Let's take a closer look at the logistic regression analysis results.

In linear regression, we used the method of least squares to find the best fitting model and estimate the model coefficients. Least squares is actually a special case of a more general approach for estimating statistical models, the method of maximum likelihood. The basic idea behind maximum likelihood is to find estimates for the model coefficients that are the most consistent with the observed data. When using logistic regression, the method of maximum likelihood finds unique values for the coefficients in the logistic model that are the most likely given the data that you have measured. For more information on maximum likelihood and the fitting procedure, see the Read About It for this module.

Let's take another look at the Parameter Estimates table. JMP conducts a chi-square test to test for the significance of the model coefficients. We omit the statistical details, and focus on the p-value reported in the table. With a p-value less than 0.0001, we can conclude that Temp is a highly significant predictor of Outcome.

The estimated coefficients in the table are for the log odds, or the logit, of Fail versus Pass. We can use this formula to compute the probabilities for a given value of Temp. For example, if the Temp is 28, the probability of Fail is 0.456. Because the probabilities sum to 1.0, the probability of Pass at this Temp can be computed from the probability of Fail. The probability of Pass is 0.544.

Let's say we want to predict whether a batch will pass or fail based on the temperature alone. We can predict the Outcome, or make a classification of the "most likely Outcome," using the predicted probabilities. Based on the predicted probability of membership in one of the two classes, the predicted class is the one with the highest probability.

In this example, the probability of Pass is higher than the probability of Fail, so we classify the batch as a Pass. Comparing the actual class to the predicted class gives us a measure of the performance of our model. We classify each row in our data table as a Pass or a Fail, and then see if these classifications are correct or incorrect. The results are often summarized in a table referred to as a confusion matrix. This confusion matrix shows the percent of correct and incorrect classifications.

Here, we can see that 12% of the observations were actual fails that were correctly classified as fails. And we can see that 71% of the observations were passes that were correctly classified as passes. So, 83% of the batches were correctly classified. Likewise, 14% of the observations were fails that were misclassified as passes, and 3% of the observations were passes that were incorrectly classified as fails. So, 17% of the batches were misclassified by our model.

We use this value, the misclassification rate, as a measure of the predictive performance of logistic regression models. Note that there are many other measures of performance and goodness of fit, but in this lesson, we limit our discussion to the misclassification rate.

In the next video, we see how to fit simple logistic regression models and interpret the statistical results, in JMP.

Close