Interpreting Results in Explanatory Modeling

As we discussed in the Simple Linear Regression lesson, we can use regression for different reasons. Two common goals of regression are explanatory modeling and predictive modeling. In explanatory modeling, we use regression to determine which variables have an effect on the response or help explain the response.

In this context, we are generally interested in identifying the predictors that tell us the most about the response, and in understanding the magnitude and direction of the model coefficients. That is, we want to know how the response values change as we change the values of a given predictor. In predictive modeling, we use regression to develop a model that accurately predicts values of the response variable based on the values of the predictors.

In this context, we are not as interested in understanding which predictors are important, or in estimating the model coefficients. We are interested in developing a model that accurately predicts future response values.

In this lesson, we focus on regression for explanatory modeling, but we'll also see how to use regression models for predictive purposes. We'll formally discuss using regression for predictive modeling in another module.

For this discussion, we switch to the Impurity example. In this scenario, a polymer is being produced. A catalyst is required for the chemical reactions to occur to produce the polymer. The catalyst contains a chemical that can create an impurity in the polymer. Impurity is directly related to Yield. That is, Percent Impurity plus Percent Yield is 100.

We use regression to understand the relationship between Impurity and three predictors, Temp, Catalyst Conc, and Reaction Time. The estimated, or fitted, regression equation is shown here.

Recall that in simple linear regression, we can conduct a hypothesis test to determine whether there is a relationship between the response and the predictor. We test the null hypothesis that the true slope coefficient, β1, is zero. In multiple regression, we test the null hypothesis that all the regression coefficients are zero, versus the alternative that at least one slope coefficient is nonzero.

To perform this hypothesis test, we compute a test statistic, the F Ratio. Recall that the F Ratio is a statistical signal-to-noise ratio. It's a ratio of the variation explained by our model (Mean Square Model) and the unexplained variation (Mean Square Error).

When there is no relationship between the response and any of the predictors, the model will not explain much of the variation in the response. Mean Square Model and Mean Square Error will be approximately the same, and the F Ratio will be close to 1. On the other hand, if the alternative hypothesis is true, at least one coefficient is nonzero. The model will explain at least some of the variation in the response. Mean Square Model will be greater than Mean Square Error, and the F Ratio will be greater than 1.

Because our decision-making about the magnitude of the F Ratio can be influenced by both the number of parameters in the model and the number of observations in our data set, we can't rely on the F Ratio alone to make decisions about our null hypothesis.

Fortunately, statistical software packages report p-values for all test statistics. As we have seen, p-values measure the strength of the evidence against our null hypothesis. The F Ratio, and the corresponding p-value, are reported in the ANOVA table. In our example, the F Ratio is 122.8, and the p-value is very small, less than 0.0001. We can safely conclude that at least one term in our model is significant. This is what is known as a whole model test. The ANOVA table enables us to make decisions about the significance of our model on the whole, but it doesn't tell us which predictors are significant.

For this, we use the information reported in the Effects Test table. This information is also reported in the Effect Summary table. The F ratios and p-values provide information about whether each individual predictor is related to the response.

These tests are known as partial tests, because each test is adjusted for the other predictors in the model. As we saw earlier, if the predictors are correlated, the p-values can change a great deal as other variables are added to or removed from the model.

Note that there are other types of tests for individual predictors that are available, but this discussion is beyond the scope of this course and we limit our discussion to partial t tests.

Returning to our example, both Temp and Catalyst Conc are highly significant. But Reaction Time is not significant, given the other terms in the model. The coefficients for Temp and Catalyst Conc are both positive, indicating that as the values of each of these variables increase, holding everything else constant, Impurity also increases.

To better understand how the predicted response changes as we change values of the individual predictors, we use the Prediction Profiler. This is illustrated in the next video.

But, before we get ahead of ourselves, there are some potential issues we need to explore. As in simple linear regression, we need to check our residuals to make sure we don't see any issues.

A residual analysis is used to investigate nonlinearity in the relationship between the response and predictors, non-constant error variance, and autocorrelation, or non-independence, of the errors. Residual analysis can also help us identify outliers or unusual observations that might be influencing our model.

Because we're dealing with multiple predictors, there are a few other issues we need to consider. First, there is the somewhat complicated issue of identifying important variables. We might want to add additional predictors to the model to explain more of the variation in the response. Or we might want to simplify the model by removing nonsignificant terms.

We also need to identify whether there is correlation between the predictors, or collinearity. This can make it difficult to determine which variables are most useful in explaining the response and can cause many issues in estimating our coefficients.

We address these topics in the videos that follow.

Close