Interpreting Regression Results

Earlier, we saw that the method of least squares is used to fit the best regression line. The total variation in our response values can be broken down into two components: the variation explained by our model and the unexplained variation or noise.

The total sum of squares, or SST, is a measure of the variation of each response value around the mean of the response. For each observation, this is the difference between the response value and the overall mean response.

The model sum of squares, or SSM, is a measure of the variation explained by our model. For each observation, this is the difference between the predicted value and the overall mean response. This is the variation that we attribute to the relationship between X and Y. Note that sometimes this is reported as SSR, or regression sum of squares.

The error sum of squares, or SSE, is a measure of the random error, or the unexplained variation. For each observation, this is the difference between the response value and the predicted value. This is the variation that is not explained by our regression model. This is also referred to as sum of squared errors. All of the variation in our response can be broken down into either model sum of squares or error sum of squares. Another way to think about sums of squares is to consider a right triangle. The total sum of squares can be broken down into error and model sums of squares.

Compare the sums of squares for Model 1 and Model 2. In Model 1, more of the total variation in the response is unexplained than in Model 2. In other words, Model 2 explains more of the total variation in the response than Model 1.

The sums of squares are reported in the ANOVA table, which was described in a previous module. In the context of regression, the p-value reported in this table gives us an overall test for the significance of our model. The p-value is used to test the hypothesis that there is no relationship between the predictor and the response. Or, stated differently, the p-value is used to test the hypothesis that the true slope coefficient is zero.

For the cleaning example, we fit a model for Removal versus OD. Because our p-value is very small, we can conclude that there is a significant linear relationship between Removal and OD. In a simple linear regression situation, the ANOVA test is equivalent to the t test reported in the Parameter Estimates table for the predictor. The estimates in the Parameter Estimates table are the coefficients in our fitted model. As we have discussed, we can use this model directly to make predictions. More specifically, we can use the model to predict average Removal within the range of values we observed for OD. This is an important point. The OD values in our sample range from 4 to 24.7.

It's important to keep in mind that extrapolating beyond this range can lead to unrealistic or unreliable predictions. We can also construct two types of intervals using our model: confidence intervals and prediction intervals.

Confidence intervals, which are displayed as confidence curves, provide a range of values for the predicted mean for a given value of the predictor. Note that these bands are essentially what we observed in the Demonstrate Regression simulation when we fit 1000 lines. The bands represent the uncertainty in the estimates of the true line.

Prediction intervals provide a range of values where we can expect future observations to fall for a given value of the predictor. Prediction intervals are useful when we are interested in using the model to predict individual values of the response.

The Demonstrate Regression simulation illustrated that estimates of the true slope can vary from sample to sample. There can be a large difference in the slope from one sample to another. Our slope estimate, 0.5283, is a point estimate for the true, unknown slope. So we use a confidence interval to provide a range of values for the true slope. For our example, the average increase in Removal for every 1-unit increase in OD is between 0.462 and 0.595. The confidence interval for the slope provides an additional test for size of the slope coefficient. This might be easier to interpret and explain than a p-value. Because our confidence interval does not contain zero, we can conclude that the true slope is nonzero.

Here are additional measures of the fit of the model. One popular statistic is RSquare, the coefficient of determination. RSquare provides a measure of the strength of the linear relationship between the response and the predictor. In simple linear regression, RSquare is the square of the correlation coefficient, r. This statistic, which falls between 0 and 1, measures the proportion of the total variation explained by the model. The closer RSquare is to 1, the more variation that is explained by the model. In our example, 84% of the variation in our response, Removal, is explained by the variable OD.

Note that the value of RSquare can be influenced by a number of factors, so here are a few cautions. If there are repeated measurements for the predictor, the maximum possible value of RSquare will be less than 1. So having repeated measurements, which is generally desirable, results in lower values of RSquare. Also, as we saw with the correlation coefficient, severe outliers can artificially inflate RSquare.

So, although RSquare is a useful measure, and in general a higher RSquare value is better, there is no cutoff value to use for RSquare that indicates that we have a good model. RSquare, and the similar measure RSquare Adjusted, are best used to compare different models on the same data. We describe RSquare Adjusted in the Multiple Linear Regression lesson.

Close