Fitting the Multiple Linear Regression Model

Recall that the method of least squares is used to find the best-fitting line for the observed data. The estimated least squares regression equation has the minimum sum of squared errors, or deviations, between the fitted line and the observations. When we have more than one predictor, this same least squares approach is used to estimate the values of the model coefficients.

For example, when we have two predictors, the least squares regression line becomes a plane, with two estimated slope coefficients. The coefficients are estimated to find the minimum sum of squared deviations between the plane and the observations.

This extends to more than two predictors, but finding the least squares solution becomes much more complicated and requires matrix algebra. Fortunately, most statistical software packages can easily fit multiple linear regression models.

Let's revisit the Cleaning data one more time, focusing on only two predictors, OD and ID. We see that both OD and ID are positively correlated with Removal. And we also see that they are correlated with one another. This means that parts with larger outside diameters also tend to have larger inside diameters. In our individual models, OD and ID are both significant predictors of Removal, with very small p-values.

Here, we fit a multiple linear regression model for Removal, with both OD and ID as predictors. Notice that the coefficients for the two predictors have changed. The coefficient for OD (0.559) is pretty close to what we see in the simple linear regression model, but it's slightly higher. But look at the coefficient for ID! Now it's negative, and it's no longer significant. How do we interpret these results?

In multiple linear regression, the significance of each term in the model depends on the other terms in the model. OD and ID are strongly correlated. When OD increases, ID also tends to increase. So, when we fit a model with OD, ID doesn't contribute much additional information about Removal.

We see this more clearly when we look at the model fit statistics. Recall that RSquare is a measure of the variability in the response explained by the model. A similar measure, RSquare Adjusted, is used when fitting multiple regression models. We'll describe RSquare Adjusted in more detail later in this lesson.

A second important measure of model fit, the Root Mean Square Error, or RMSE, is a measure of the unexplained variation in the model. This is, essentially, a measure of how far the points are from the fitted line, on average. When the root mean square error is lower, the points are generally closer to the fitted line. For a predictive model, this corresponds to a model that predicts more precisely.

In our individual model for OD, RSquare is 0.84 and the root mean square error is 1.12. What is the RSquare Adjusted for the multiple regression model with both ID and OD? It's basically the same, 0.83. And, the root mean square error for the model with both predictors, 1.13, is very similar to the root mean square error for the model with just OD. So, we don't learn anything more about Removal when we add ID to the model than we already know with OD alone.

These somewhat contradictory results are actually fairly common, and later we'll see how to address the problem. For now, let's explore the issue further with a new example.

Take the relationship between drownings and ice cream consumption. We introduced this example in an exercise in the correlation lesson. When we fit a regression model for DrowningRate as a function of IceCreamRate, the model is highly significant. Higher drowning rates are associated

with higher ice cream consumption rates. But, can we interpret this to mean that ice cream consumption is directly associated with drownings?

When we take a closer look, we see that there is also a significant relationship between DrowningRate and Year. Over time, the drowning rate is decreasing.

When we fit a multiple regression model with both IceCream Rate and Year, only Year is significant. On average, the drowning rate decreases by 0.12 per year. Ice cream consumption is no longer a significant predictor of drownings, after adjusting for changes over time.

Remember the previous discussion of correlation versus causation. Just because we see significant results when we fit a regression model for two variables this does not necessarily mean that a change in the value of one variable causes a change in the value of the second variable, or that there is a direct relationship between the two variables.

Close