What is Multiple Linear Regression?

Recall that simple linear regression can be used to predict the value of a response based on the value of one continuous predictor variable. Depending on the context, the response and predictor variables might be referred to by other names. For simplicity, we'll generally stick with the terms response and predictor throughout this lesson. Let's return to an earlier example. Using the Cleaning data, we used simple linear regression to model the relationship between the response, Removal, and the predictor, OD. We found that a 1-unit increase in OD is associated with a 0.53-unit increase in Removal.

But, what do we do if we have more than one predictor variable?

For the cleaning example, we have three potential predictors, OD, ID, and Width.

One option is to fit separate regression models for the different predictors.

Continuing with this example, we learn that there is a significant relationship between Removal and ID, and that for a 1-unit increase in ID, Removal increases by 0.65 units on average.

We also learn that there is not a significant relationship between Removal and Width. In other words, there is no association between changes in Width and changes in Removal.

However, fitting simple linear regression models for each predictor ignores the information in the other variables.

Because we can make predictions based on only one model at a time, we lose potentially valuable information in the other predictors.

This means that we can't make predictions with a model based on one predictor using the information from the other predictors.

Also, because each simple model ignores the other variables when the coefficients are estimated, the coefficients might be misleading, especially if the predictors are correlated with one another.

Instead of fitting separate models for each predictor, we can include multiple predictors in the same model. When more than one predictor is used, the procedure is called multiple linear regression.

Recall the unknown, or true, linear regression model with one predictor. This equation describes how the mean of Y changes for given values of X. We can also write the equation in terms of the observed values of Y, rather than the mean. Because the individual data values for any given value of X vary randomly about the mean, we need to account for this random variation, or error, in the regression equation.

We add the Greek letter ε to the equation to represent the random error in the individual observations.

When we fit a multiple linear regression model, we add a slope coefficient for each predictor.

For the Cleaning example, with OD and ID as predictors, the model has slope coefficients for both predictors.

Each coefficient represents the average increase in Removal for every one-unit increase in that predictor, holding the other predictor constant.

What if we have more than two predictors?

As a generalization, let's say that we have p predictors. The multiple linear regression model can be extended to include all p predictors. Linear regression models can also include functions of the predictors, such as transformations, polynomial terms, and cross-products, or interactions. And later we'll see that linear models can also be fit with categorical predictors. A challenge when fitting multiple linear regression models is that we might need to estimate many coefficients. Although modern statistical software can easily fit these models, it is not always straightforward to identify important predictors and interpret the model coefficients. In the videos that follow, we talk about fitting and interpreting multiple linear regression models and some of the challenges involved.

Statistical Thinking for Industrial Problem Solving

Close