Print

The Simple Linear Regression Model

In the cleaning parts example, we collected data on 50 parts. We fit a regression model to predict Removal as a function of the OD of the parts. But what if we had sampled a different set of 50 parts and fit a regression line using these data? Would this produce the same regression equation?

By fitting a regression line to observed data, we are trying to estimate the true, unknown relationship between the variables. This fitted regression equation is just one estimate of the true linear model. In reality, the true linear model is unknown.

In simple linear regression, we assume that, for a fixed value of a predictor X, the mean of the response Y is a linear function of X. We denote this unknown linear function by the equation shown here where β0 is the intercept and β1 is the slope. The regression line we fit to data is an estimate of this unknown function.

The equation of the fitted line is denoted by this equation. Here, β0 and β1 are estimates of β0 and β1, respectively. The notation Y-hat indicates that the response is estimated from the data and that it is not an actual observation.

In the cleaning example, the intercept, β0, is 4.099 and the slope, β1, is 0.528. If we select a different sample of parts, our fitted line will be different. To illustrate, we use the Demonstrate Regression teaching module in the sample scripts directory. We'll simulate samples based on the cleaning parts example.

Let's say for the purpose of this simulation, that we know the true line and the true population characteristics. The true intercept is 4.00, the slope is 0.50, and the correlation is 0.90.The Y variable is Removal, and the X variable is OD. The mean of OD is 14, and the standard deviation is 5.0. The Sample Size is 50 parts. The Sample Data graph shows one sample, of size 50, drawn from this population. When I open the Summary of All Samples outline, we see the true line. In reality, this true line is not known, but we'll use it to see how much variability we observe in the lines from different samples.

The characteristics for this fitted line are shown under Summary of Fits. When I click Draw Additional Samples, a new sample is simulated. This is displayed as a red line in the Summary of All Samples graph. Each time I click Draw Additional Samples, a new sample is simulated. Notice that the slopes and intercepts for each of these samples are slightly different. I'll select Reset Samples and this time draw 1000 samples of size 50. The lines for the simulated data appear to form bands around the true line. The bands are widest at the low and high values of OD, and are narrowest at the mean of OD. These bands are essentially confidence bands, showing the uncertainty in our individual lines around the true line. We can see this uncertainty in the Summary of Fits table. The average intercept and slope for our 1000 samples are close to our population values, but there is a wide range in values for our individual lines.

In reality, we don't know the true line, and we fit a line for our one sample. This fitted line is just one point estimate for the true model. But, because of the variability we might observe if we were to draw different samples, statistical software reports confidence intervals and provides hypothesis tests to help us draw inferences about the true linear model.

We discuss this standard regression output later in this lesson.

Statistical Thinking for Industrial Problem Solving

Close