Print

Introduction to Regression Analysis

We're often interested in understanding the relationship among several variables. Scatterplots and scatterplot matrices can be used to explore potential relationships between pairs of variables.

Correlation provides a measure of the linear association between pairs of variables, but it doesn't tell us about more complex relationships. For example, if the relationship is curvilinear, the correlation might be near zero. You can use regression to develop a more formal understanding of relationships between variables. In regression, and in statistical modeling in general, we want to model the relationship between an output variable, or a response, and one or more input variables, or factors.

Depending on the context, output variables might also be referred to as dependent variables, outcomes, or simply Y variables, and input variables might be referred to as explanatory variables, effects, predictors or X variables.

We can use regression, and the results of regression modeling, to determine which variables have an effect on the response or help explain the response. This is known as explanatory modeling.

We can also use regression to predict the values of a response variable based on the values of the important predictors. This is generally referred to as predictive modeling. Or we can use regression models for optimization, to determine settings of factors to optimize a response.

Our optimization goal might be to find settings that lead to a maximum response or to a minimum response. Or the goal might be to hit a target within an acceptable window. For example, let's say we're trying to improve process yield. We might use regression to determine which variables contribute to high yields, we might be interested in predicting process yield for future production, given values of our predictors, or

we might want to identify factor settings that lead to optimal yields.

We might also use the knowledge gained through regression modeling to design an experiment that will refine our process knowledge and drive further improvement. Consider an example where we are interested in the cleaning of metal parts. We have 50 parts with various inside diameters, outside diameters, and widths. Parts are cleaned using one of three container types. Cleanliness is a measure of the particulates on the parts. This is measured before and after running the parts through the cleaning process. The response of interest is removal. This is the difference between pre-cleaning and post-cleaning measures.

We're interested in whether the inside diameter, outside diameter, part width, and container type have an effect on the cleanliness, but we're also interested in the nature of these effects. The relationship we develop linking the predictors to the response is a statistical model or, more specifically, a regression model.

The term regression describes a general collection of techniques used in modeling a response as a function of predictors. The only regression models that we'll consider in the lesson are linear models. An example of a linear model for the cleaning data is shown. In this model, if the outside diameter increases by 1 unit, with the width remaining fixed, the removal increases by 1.2 units.

Likewise, if the part width increases by 1 unit, with the outside diameter remaining fixed, the removal increases by 0.2 units. This model enables us to predict removal for parts with given outside diameters and widths. For example, the predicted removal for parts with an outside diameter of 5 and a width of 3 is 16.6 units.

In this example, we have two continuous predictors. When more than one predictor is used, the procedure is called multiple linear regression. When only one continuous predictor is used, we refer to the modeling procedure as simple linear regression. For the remainder of this lesson, we focus on simple linear regression. In the next lesson, we introduce multiple linear regression.

A scatterplot indicates that there's a fairly strong positive relationship between Removal and OD (the outside diameter). To understand whether OD can be used to predict or estimate Removal, we fit a regression line.

The fitted line estimates the mean of Removal for a given fixed value of OD. The value 4.099 is the intercept and 0.528 is the slope coefficient. The intercept, which is used to anchor the line, estimates Removal when the outside diameter is zero.

Because diameter can't be zero, the intercept isn't of direct interest. The slope coefficient estimates the average increase in Removal for a 1-unit increase in outside diameter. That is, for every 1-unit increase in outside diameter, Removal increases by 0.528 units on average.

In a previous module, you were introduced to ANOVA. Let's quickly compare regression and ANOVA. In simple linear regression, both the response and the predictor are continuous. In ANOVA, the response is continuous, but the predictor, or factor, is nominal.

The results are statistically related. In both cases, we're building a general linear model. But the goals of the analysis are different.

Regression gives us a statistical model that enables us to predict a response at different values of the predictor, including values of the predictor not included in the original data.

ANOVA measures the mean shift in the response for the different categories of the factor. As such, it is generally used to compare means for the different levels of the factor.

You will learn more about the simple linear regression model, and will see how to fit regression models, in upcoming videos.

Close