Print

Fitting a Model with Curvature

In this example, a ball was dropped from rest at time 0 seconds from a height of 400 cm. The distance that the ball had fallen (in centimeters) was recorded by a sensor at various times. How would you describe the relationship between these two variables?

We fit a regression model, using Distance (cm) as a response and Time (sec) as a predictor. How well does a straight line describe the relationship between these two variables?

There appears to be some curvature in the relationship between the two variables that the straight line doesn't capture. Some points are systematically above the line, and others are below the line. But there is a tendency to ignore the graphical output and look first at the statistical output. Notice that both the model and the linear slope coefficient are highly significant, and that more than 95% of the variability in Distance (cm) is explained by Time (sec). But should we use this model to make predictions? A good practice, before interpreting statistical output, is to look at the graphical displays of the data and the residuals.

Let's take a look at the residual plots. Notice the curved pattern in the residual plot. This plot displays the variation left over after we've fit our linear model. In this example, the plot magnifies the subtle pattern we see in the bivariate plot. The residual plot also provides insights into how we might improve our model. In this case, we might need a more complex model -- one that addresses the curvature we see. To explain this curvature, we might fit a second-order polynomial model to the data. For this example, the polynomial model appears to do a better job of explaining the relationship between Time (sec) and Distance (cm).

The residual by predicted plot now looks much better. There is no obvious pattern, and the residuals appear to be scattered about zero.

Looking at RSquare, we see that nearly all of the variation in the response is explained by the model. The model is still highly significant, and there is a new term in the Parameter Estimates table. This is a quadratic effect. Both the linear term and the quadratic effect are highly significant. So, even though our initial linear model was significant, the model is improved with the addition of a quadratic effect. Note that this is still considered a linear model because the quadratic term was added in a linear fashion. In this model, note how the quadratic term is written. This means that the polynomial has been centered. The values of Time (sec) were centered by subtracting the mean. Centering polynomials is a standard technique used when fitting linear models with higher-order terms. It leads to the same model predictions, but does a better job of estimating the model coefficients.

In this example, the residual analysis pointed to a problem, and fitting a polynomial model made sense. In most real-life scenarios, fitting the best possible model when there are unusual patterns in data is not as straightforward. For example, you might need to apply a transformation to the response or the predictor. Or you might be missing other important effects that explain the relationship.

The decision on how to proceed with the analysis should be guided by subject matter knowledge and the context of the problem. As always, we recommend consulting with an expert to determine the best course of action.

Close