Print

Assumptions for Regression

When you perform a regression analysis, several assumptions must be met to provide valid tests of hypotheses and confidence intervals. The first assumption is that the linear model fits the data adequately. For example, in simple linear regression, the mean of the response variable is linearly related to the value of the predictor variable. To make sure that a linear relationship exists, you can create a scatter plot of the response, Y, versus the predictor, X.

The other three regression assumptions correspond to the error. The linear regression model assumes that the errors are normally distributed with mean 0, that the errors have an equal variance at each value of the predictor variable, and that the errors are independent. Let's focus on the first assumption, that the linear model fits the data adequately.

Four examples were developed by Anscombe in 1973 to demonstrate the importance of plotting your data to check this assumption. In each example, the scatter plot of the data values is different, but the regression equation is Y = 3.0 + 0.5X, and the R-square statistic is 0.67.

In the first plot, the data hovers around the regression line, so a regression line does in fact adequately describe the data. In the second plot, a simple linear regression model isn't adequate. The straight line doesn't fit the curvilinear relationship. To fit the data better, the model requires a quadratic term. However, the model and R-square values are still the same.

In the third plot, there seems to be an outlier (toward the top center of the plot) that's affecting the fit of the regression line. This outlier is an influential data value in that it's substantially changing the fit of the regression line. The points closely follow the regression line. Again, the model and R-square values are the same.

In the final plot, the outlier dramatically changes the fit of the regression line. The points line up vertically on the x-axis at the value 8, and on the y-axis from the value 5 to around 9. In fact, without the outlier, the slope would be undefined, yet the model and R-square values are the same.

The four plots illustrate that relying on the regression output to describe the relationship between our variables can be misleading. The regression equation and the R-square statistics are the same even though the relationships between the two variables are different. You should always produce a scatter plot before you conduct a regression analysis.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close