Print

Identifying Influential Observations

So what's the difference between an outlier and an influential observation? An outlier is an unusual data point, whereas an influential observation is an unusual data point that has a large effect on some part of the model, such as the model coefficients, the standard errors, or the predicted values. It can sometimes have a large residual compared to the rest of the points, but it's an observation so far away from the rest of the data that it singlehandedly exerts influence on the regression model.

For example, if deleting an observation results in a large change in parameter estimates, then that observation has a significant influence on the parameters. If deleting an observation results in a change in the standard errors, then the observation influences the precision of the parameters.

In this plot, an influential observation strongly affects the linear model's fit to the data. The points are scattered around the bottom left portion of the plot, and the influential observation is in the upper right corner of the plot. If the influential observation were removed, the best fitting line to the rest of the data would most likely be very different.

To identify outliers and influential observations in your data, you can use several diagnostic statistics. First, you can use STUDENT residuals as a way to detect outliers. To detect influential observations, you can use Cook's D statistics, RSTUDENT residuals, and DFFITS statistics. To determine which predictor variable is being influenced, you can use DFBETAS.

Close