Identifying Influential Observations

Influential observations might be difficult to detect from simple scatter plots in a multiple regression setting. Several statistics are designed to assist in identifying influential observations. Diagnostic plots of these statistics provide a visual method to identify influential observations. Let's look at each of these in detail.

Recall that you use the STUDENT residual to check for outliers and the RSTUDENT residual to check for influential observations. The RSTUDENT residual is calculated as the residual divided by the standard error estimated with the current observation deleted. If the absolute value of the RSTUDENT residuals is greater than 2 or 3 (depending on sample size), you've probably detected an influential observation.

The leverage statistic measures how far an observation is from the cloud of observed data points. Observations far from the middle tend to be influential points. If their leverage values are greater than 2p/n, where *p* is the number of model parameters and *n* is the sample size, they might be influential. For more information on the Leverage statistic, click the Information button.

Cook's distance, or the Cook's D statistic, is the most common measure of the influence of an observation. It is calculated for each observation in the data set. For each observation, the Cook's D statistic is calculated as if that observation weren't in the data set. The Cook's D statistic measures the distance between the set of parameter estimates with that observation deleted from your regression analysis and the set of parameter estimates with all the observations in your regression analysis. An observation might have an adverse effect on the analysis if the Cook's D statistic is greater than 4/n, where *n* is the sample size.

DFFITS measures the impact that each observation has on its own predicted value. For each observation, DFFITS is calculated using two predicted values. The first predicted value is calculated from a model using the entire data set to estimate model parameters. The second predicted value is calculated from a model using the data set with that particular observation removed to estimate model parameters. The difference between the two predicted values is divided by the standard error of the predicted value, without the observation. If the standardized difference between these predicted values is large, that particular observation has a large effect on the model fit.

There are two versions of the rule of thumb for DFFITS. The general cutoff value is 2. The more precise cutoff is 2 times the square root of p divided by n, where *p* is the number of terms in the model, including the intercept, and *n* is the sample size. If the absolute value of DFFITS for any observation is greater than this cutoff value, you've detected an influential observation.

The DFBETAS statistic not only helps you to identify an influential observation, it also tells you which predictor variable is being influenced. DFBETAS, which stands for difference in betas, measure the change in each parameter estimate when an observation is deleted from the analysis. One DFBETA is calculated per predictor variable per observation. Each DFBETA is calculated by taking the estimated coefficient for that particular predictor variable, using all the data, and subtracting the estimated coefficient for that particular predictor variable with the current observation removed. This difference in the betas is divided by its standard error calculated with all the observations included in the analysis. This calculation is repeated for all predictor variables and all observations.

Large DFBETAS indicate observations that are influential in estimating a given parameter. When the absolute value of the DFBETA is greater than 2 times the square root of 1 divided by n, where *n* is the sample size, you have identified and influential observation. For more details on the DFBETAS statistic, click the Information button.

The covariance ratio is yet another statistic that you can use to identify influential observations. The Covariance Ratio measures the change in the precision of the parameter estimates when an observation is deleted from the model. It is calculated as the ratio of the determinant of the covariance matrix with the i^{th} observation deleted, to the determinant of the covariance matrix with all the observations included. Values of the covariance ratio greater than 1 indicate that the presence of the i^{th} observation increases the precision of the estimates. Values less than 1 indicate that the presence of the i^{th} observation decreases the precision of the estimates. Values near 1 indicate that the i^{th} observation has little effect on the precision of the estimates.

You can use ODS Graphics in PROC REG to create plots of values of RSTUDENT, LEVERAGE, Cook's D, DFFITS, DFBETAS, and COVRATIO, as well as plots of the studentized residuals plotted against the predicted values and against the leverage statistics. Here's a table summarizing the suggested cutoffs to be used for identifying influential observations with each of these statistics.

Close