Print

Checking for Influential Observations

The RSTUDENT residuals are one of the statistics that you can use to identify influential observations. RSTUDENT residuals are similar to STUDENT residuals, except they're calculated after deleting the i^{th} observation. In other words, the RSTUDENT residual is the difference between the observed Y and the predicted value of Y that's calculated after excluding the observation from the regression. This difference is then divided by the standard error.

Like the studentized residuals, the RSTUDENT residuals are put on a standard deviation scale, so we can expect most observations to have RSTUDENT residuals less than 2 or less than 3 in magnitude. When you evaluate RSTUDENT residuals, the observation is probably influential if the RSTUDENT residual is different from the STUDENT residual. Also, if the absolute value of the RSTUDENT residuals is greater than 3, you've probably detected an influential observation.

The Cook's D statistic is most useful for identifying influential observations for explanatory models, when the purpose of your model is parameter estimation. You can calculate a Cook's D statistic for each observation in a data set as if that observation weren't in the data set.

This 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. If any observation has a Cook's D statistic greater than 4 divided by n, where n is the sample size, that observation is influential. Diagnostic plots that SAS creates for Cook's D show the suggested cutoff as a horizontal line, making it easy to determine each observation's influence on the parameter estimates.

DFFITS is most useful for predictive models. 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 in which that particular observation is 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.

$$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{(i)}}{s(\widehat{Y}_i)}$$

If the standardized difference between these predicted values is large, that particular observation has a large effect on the model fit. The suggested cutoff is $2\sqrt{\frac{p}{n}}$, 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.

Close