0.1 Cook's Distance

Some texts tell you that points for which Cook's distance is higher than 1 are to be considered as influential. Other texts give you a threshold of $4/N$ or $4/(Nk1)$, where N is the number of observations and k the number of explanatory variables. In your case the latter formula should yield a threshold around 0.1 .

John Fox (1), in his booklet on regression diagnostics is rather cautious when it comes to giving numerical thresholds. He advises the use of graphics and to examine in closer details the points with "values of D that are substantially larger than the rest". According to Fox, thresholds should just be used to enhance graphical displays.

In your case the observations 7 and 16 could be considered as influential. Well, I would at least have a closer look at them. The observation 29 is not substantially different from a couple of other observations.

(1) Fox, John. (1991). Regression Diagnostics: An Introduction. Sage Publications.

0.2 Interpreting Cook's Distance

A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

Purpose Cook's distance is useful for identifying outliers in the X values (observations for predictor variables). It also shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier.

Definition Cook's distance is the scaled change in fitted values. Each element in CooksDistance is the normalized change in the vector of coefficients due to the deletion of an observation.

Interpreting Cook's Distance

Some texts tell you that points for which Cook's distance is higher than 1 are to be considered as influential. Other texts give you a threshold of 4/N or 4/(Nk1), where N is the number of observations and k the number of explanatory variables. In your case the latter formula should yield a threshold around 0.1 .

John Fox (1), in his booklet on regression diagnostics is rather cautious when it comes to giving numerical thresholds. He advises the use of graphics and to examine in closer details the points with "values of D that are substantially larger than the rest". According to Fox, thresholds should just be used to enhance graphical displays.

In your case the observations 7 and 16 could be considered as influential. Well, I would at least have a closer look at them. The observation 29 is not substantially different from a couple of other observations.

I would add that influential cases are not usually a problem when their removal from the dataset would leave the parameter estimates essentially unchanged: the ones we worry about are those whose presence really does change the results.

Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. dfbeta refers to how much a parameter estimate changes if the observation in question is dropped from the data set. Note that with k covariates, there will be k+1 dfbetas (the intercept, 0, and 1 for each covariate). Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

There is one other point worth making here. In observational research, it is often difficult to sample uniformly across the predictor space, and you might have just a few points in a given area. Such points can diverge from the rest. Having a few, distinct cases can be discomfiting, but merit considerable thought before being relegated outliers. There may legitimately be an interaction amongst the predictors, or the system may shift to behave differently when predictor values become extreme. In addition, they may be able to help you untangle the effects of colinear predictors. Influential points could be a blessing in disguise.

Cook's Distance Formula

It is calculated as:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j \ - \hat{Y}_{j(i)})^2}{p \ \text{MSE}},$$

where:

- $\hat{Y}_j$ is the prediction from the full regression model for observation j;
- $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted;
- p is the number of fitted parameters in the model;
- MSE is the mean square error of the regression model.

For the case of simple linear regression, the following are the algebraically equivalent expressions

$$D_i = \frac{e_i^2}{p \ \text{MSE}} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^T (X^T X)(\hat{\beta} - \hat{\beta}^{(-i)})}{(1 + p)s^2},$$

where: h_{ii} is the leverage, i.e., the i-th diagonal element of the hat matrix $\mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T$; e_i is the residual (i.e., the difference between the observed value and the value fitted by the proposed model).

The Cook's distance, Di, of observation i is

is the jth fitted response value, where the fit does not include observation i. MSE is the mean squared error. p is the number of coefficients in the regression model. Cook's distance is algebraically equivalent to the following expression:

where ri is the ith residual, and hii is the ith leverage value.

CooksDistance is an n-by-1 column vector in the Diagnostics table of the LinearModel object.

How To After obtaining a fitted model, say, mdl, using fitlm or stepwiselm, you can:

Display the Cook's distance values by indexing into the property using dot notation, mdl.Diagnostics.CooksDistance Plot the Cook's distance values using plotDiagnostics(mdl,'cookd') For details, see the plotDiagnostics method of the LinearModel class. Determine Outliers Using Cook's Distance This example shows how to use Cook's Distance to determine the outliers in the data.