1 Cook's Distance

- In statistics, Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing least squares regression analysis.
- 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.
- Cook's distance is the scaled change in fitted values. Each resulting element in a diagnostic calculation is the normalized change in the vector of coefficients due to the deletion of an observation.
- In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate data points that are particularly worth checking for validity; to indicate regions of the design space where it would be good to be able to obtain more data points.
- It is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.
- Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression.
- Points with a large Cook's distance are considered to merit closer examination in the analysis.
- 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 D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made

with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

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).

R code for computing Cook's Distance

```
attach(mtcars)
fit = lm(mpg ~ cyl + wt )
cooks.distance(fit)
plot(cooks.distance(fit),type="b",pch=18,col="red")

N = 32
k = 2
cutoff = 4/ (N-k-1)
abline(h=cutoff,lty=2)
```

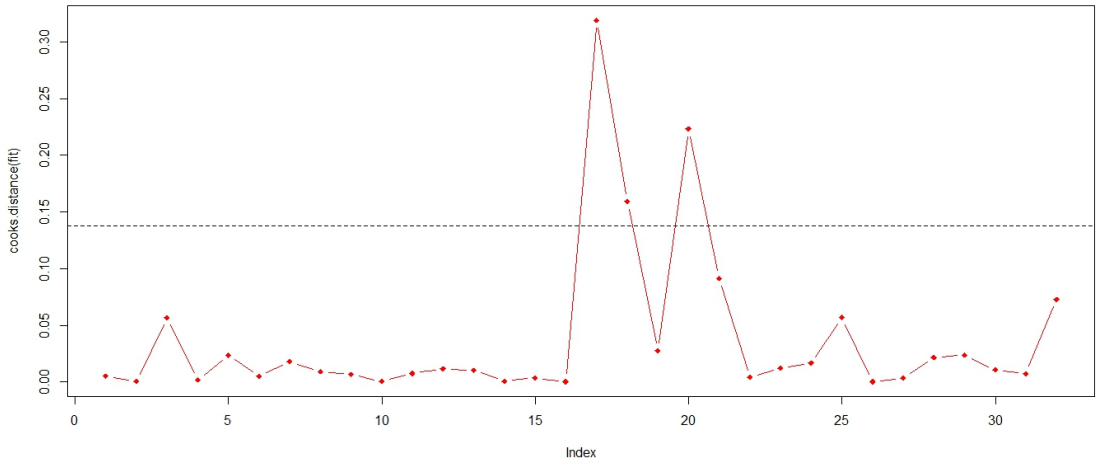

Figure 1:

1.1 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. (John Fox) *(1991). Regression Diagnostics: An Introduction. Sage Publications.*

- 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

$$\frac{4}{(N-k-1)},$$

where N is the number of observations and k the number of explanatory variables.

- The `R` help file advises that an observation with Cook's distance larger than three times the mean Cook's distance might be an outlier. .
- John Fox (mentioned above), 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.

Remark

- 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** : *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, β_0, and β_1 for each covariate).*
- Cook's distance is arguably more important if you are doing predictive modeling, whereas *DFBETA* is more important in explanatory modeling.
- **DFFITS**: Although the raw values resulting from the equations are different, Cook's distance and *DFFITS* are conceptually identical and there is a closed-form formula to convert one value to the other.