

### 0.0.1 Cook's Distance

- For variance components  $\gamma$ :  $CD(\gamma)_i$ ,
- For fixed effect parameters  $\beta$ :  $CD(\beta)_i$ ,
- For random effect parameters  $\mathbf{u}$ :  $CD(u)_i$ ,
- For linear functions of  $\hat{\beta}$ :  $CD(\psi)_i$

## Random Effects

A large value for  $CD(u)_i$  indicates that the  $i$ -th observation is influential in predicting random effects.

## linear functions

$CD(\psi)_i$  does not have to be calculated unless  $CD(\beta)_i$  is large.

*Cook (1977)* greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

Cook (1986) gave a completely general method for assessing influence of local departures from assumptions in statistical models.

### 0.0.2 Cook's Distance

In classical linear regression, a commonly used measure of influence is Cook's distance. It is used as a measure of influence on the regression coefficients.

For linear mixed effects models, Cook's distance can be extended to model influence diagnostics by defining.

$$C_{\beta i} = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}) (\hat{\beta} - \hat{\beta}_{[i]})}{p}$$

It is also desirable to measure the influence of the case deletions on the covariance matrix of  $\hat{\beta}$ .

### Cook's Distance

Cook's Distance ( $D_i$ ) is an overall measure of the combined impact of the  $i$ th case on all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the  $i$ -th case is deleted.

Importantly,  $D_{(i)}$  can be calculated without fitting a new regression coefficient each time an observation is deleted.

### 0.0.3 Cook's Distance

Cook's  $D$  statistics (i.e. colloquially Cook's Distance) is a measure of the influence of observations in subset  $U$  on a vector of parameter estimates.

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)}$$

If  $V$  is known, Cook's  $D$  can be calibrated according to a chi-square distribution with degrees of freedom equal to the rank of  $\mathbf{X}$ .

For LME models, Cook's distance can be extended to model influence diagnostics by defining.

It is also desirable to measure the influence of the case deletions on the covariance matrix of  $\hat{\beta}$ .

## 0.1 Cook's Distance for LMEs

Diagnostic methods for fixed effects are generally analogues of methods used in classical linear models. Diagnostic methods for variance components are based on 'one-step' methods. Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models.

For fixed effects parameter estimates in LME models, the Cook's distance can be extended to measure influence on these fixed effects.

$$CD_i(\beta) = \frac{(c_{ii} - r_{ii}) \times t_i^2}{r_{ii} \times p}$$

For random effect estimates, the Cook's distance is

$$CD_i(b) = g'_{(i)}(I_r + \text{var}(\hat{b})D)^{-2}\text{var}(\hat{b})g_{(i)}.$$

Large values for Cook's distance indicate observations for special attention.

### 0.1.1 Cook's Distance

Cooks Distance ( $D_i$ ) is an overall measure of the combined impact of the  $i$ th case of all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the  $k$ th case is deleted.  $D_{(k)}$  can be calculated without fitting a new regression coefficient each time an observation is deleted.

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

Cook's Distance is a well known diagnostic technique used in classical linear models, extended to LME models. For LME models, two formulations exist; a Cook's distance that examines the change in fixed fixed parameter estimates, and another that examines the change in random effects parameter estimates. The outcome of either Cook's distance is a scaled change in either  $\beta$  or  $\theta$ .

Cook's  $D$  statistics (i.e. colloquially Cook's Distance) is a measure of the influence of observations in subset  $U$  on a vector of parameter estimates (Cook, 1977).

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)}$$

If  $V$  is known, Cook's  $D$  can be calibrated according to a chi-square distribution with degrees of freedom equal to the rank of  $\mathbf{X}$  (?).

For fixed effects parameter estimates in LME models, the Cook's distance can be extended to measure influence on these fixed effects.

$$CD_i(\beta) = \frac{(c_{ii} - r_{ii}) \times t_i^2}{r_{ii} \times p}$$

For random effect estimates, the Cook's distance is

$$CD_i(b) = g'_{(i)}(I_r + \text{var}(\hat{b})D)^{-2}\text{var}(\hat{b})g_{(i)}.$$

Large values for Cook's distance indicate observations for special attention.

### 0.1.2 Change in the precision of estimates

The effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's distance, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large.

### 0.1.3 Cook's Distance

Cooks Distance ( $D_i$ ) is an overall measure of the combined impact of the  $i$ th case of all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the  $k$ th case is deleted.  $D_{(k)}$  can be calculated without fitting a new regression coefficient each time an observation is deleted.

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

Cook's Distance is a well known diagnostic technique used in classical linear models, extended to LME models. For LME models, two formulations exist; a Cook's distance that examines the change in fixed fixed parameter estimates, and another that examines the change in random effects parameter estimates. The outcome of either Cook's distance is a scaled change in either  $\beta$  or  $\theta$ .

### 0.1.4 Cook's Distance

Cook's  $D$  statistics (i.e. colloquially Cook's Distance) is a measure of the influence of observations in subset  $U$  on a vector of parameter estimates (Cook, 1977).

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)}$$

If  $V$  is known, Cook's  $D$  can be calibrated according to a chi-square distribution with degrees of freedom equal to the rank of  $\mathbf{X}$  (?).

### 0.1.5 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.[1] 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.

#### Interpretation

Specifically  $D_i$  can be interpreted as the distance one's estimates move within the confidence ellipsoid that represents a region of plausible values for the parameters.[clarification needed] This is shown by an alternative but equivalent representation of Cook's distance in terms of changes to the estimates of the regression parameters between the cases where the particular observation is either included or excluded from the regression analysis.

### 0.1.6 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.

# Bibliography

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics* 19, 15–18.