

---

# 1 Leverage and Influence

## 1.1 Summary of Influence Statistics

### 1.2 `rstudent`

The studentized residual RSTUDENT is estimated by  $s(i)^2$  without the  $i$ th observation, not by  $s^2$ . For example,

$$\text{RSTUDENT} = \frac{r_i}{s_{(i)} \sqrt{(1 - h_i)}}$$

Observations with RSTUDENT larger than 2 in absolute value may need some attention.

- **Studentized Residuals** Residuals divided by their estimated standard errors (like t-statistics). Observations with values larger than 3 in absolute value are considered outliers.
- **Leverage Values (Hat Diag)** Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than  $2(k + 1)/n$  are considered to be potentially highly influential, where  $k$  is the number of predictors and  $n$  is the sample size.
- **DFFITS** Measure of how much an observation has effected its fitted value from the regression model. Values larger than  $2\sqrt{(k + 1)/n}$  in absolute value are considered highly influential.
- **DFBETAS** Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than  $2/\sqrt{n}$  in absolute value are considered highly influential.

The measure that measures how much impact each observation has on a particular predictor is DFBETAs. The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

- **Cooks D** Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than  $4/n$  are considered highly influential.

## 2 Other Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

Influence arises at two stages of the LME model. Firstly when  $V$  is estimated by  $\hat{V}$ , and subsequent estimations of the fixed and random regression coefficients  $\beta$  and  $u$ , given  $\hat{V}$ .

### 2.0.1 DFBETA

A group of measures that measures how much impact each observation has on a particular predictor are the DFBETAs. The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \quad (1)$$

$$= B(Y - Y_{\bar{a}}) \quad (2)$$

For  $k$  predictors in the model, there are  $k + 1$  dfbetas.

### 2.0.2 DFFITS

DFFITS is a statistical measure designed to show how influential an observation is in a statistical model. It is closely related to the studentized residual.

$$DFFITS = \frac{\hat{y}_i - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}}$$

### 2.0.3 COVRATIO

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the  $i$ th observation:

$$COVRATIO = \frac{\det(s_{(i)}^2(X'_{(i)}X_{(i)})^{-1})}{\det(s^2(X'X)^{-1})}$$

Observations with

$$|COVRATIO - 1| \geq \frac{3k}{n}$$

where  $k$  is the number of parameters in the model and  $n$  is the number of observations used to fit the model, are worth further investigation.

## 2.1 Influential Observations : DFBeta and DFBetas