STAT 401A - Statistical Methods for Research Workers

Case statistics

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 17, 2014

Case statistics

Definition

Leverage (h_i) is a measure of the distance between an observation's explanatory variable values and the average of the explanatory variable values in the entire data set.

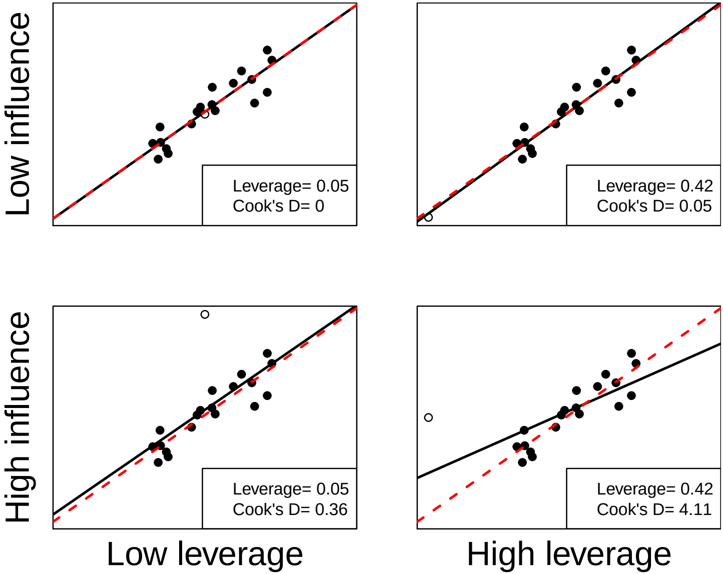
Rule-of-thumb: Possible concern when leverage $> 2p/n$ where p is the number of regression coefficients and n is the number of observations.

Definition

Cook's distance (D) is a measure of the overall effect on estimated regression coefficients when removing an observation.

Rule-of-thumb: Concerned when Cook's D ≈ 1.

Consider simple linear regression (point of interest is the open circle):



Residuals

- Residual (observed minus predicted):

$$r_i = \hat{e}_i = Y_i - \hat{\mu}_i$$

- (Internally) studentized residual

$$\frac{r_i}{\widehat{SD}(r_i)} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$$

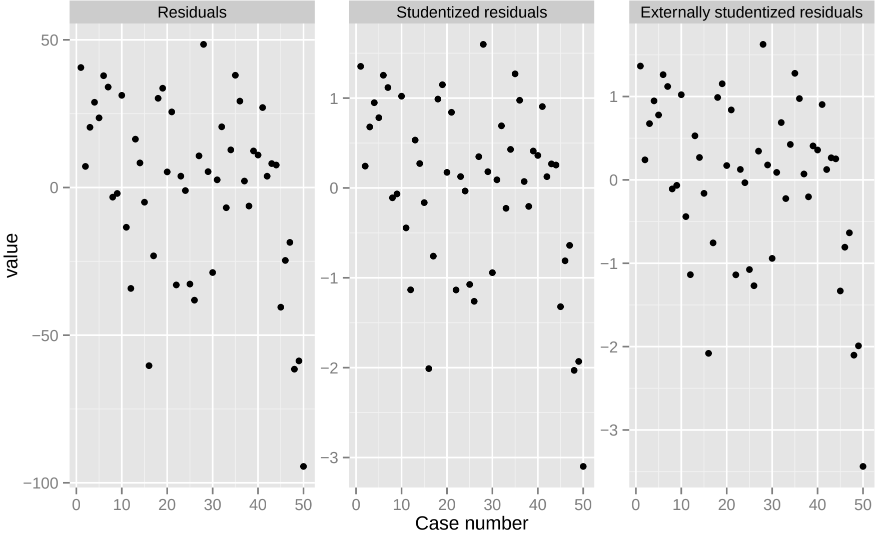
- Externally studentized residuals

$$\frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

  where $\hat{\sigma}_{(i)}$ is the estimate of the standard deviation about the regression line from the fit that excludes observation i.

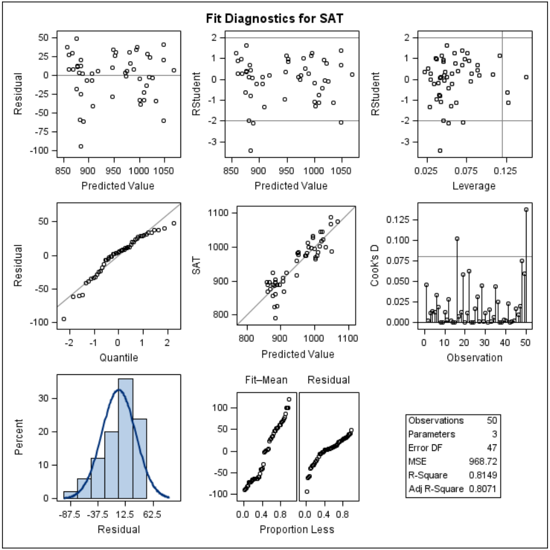
95% of studentized residuals should be within -2 and 2.

SAT residuals after adjusting for % taking and median class rank:



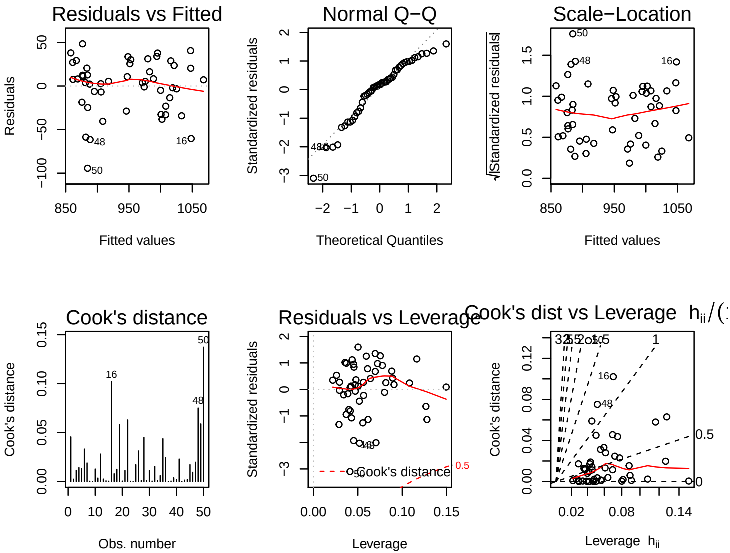
```
DATA case1201;
  INFILE 'case1201.csv' DSD FIRSTOBS=2;
  INPUT state $ sat takers income years public expend rank;
  ltakers = log(takers);
  IF state='Alaska' THEN DELETE;
  RUN;

PROC GLM DATA=case1201;
  MODEL sat = ltakers rank;
  RUN;
```

SAS diagnostics:



Observations	50
Parameters	3
Error DF	47
MSE	968.72
R-Square	0.8149
Adj R-Square	0.8071

```
mod = lm(SAT~log(Takers)+Rank, case1201)
opar = par(mfrow=c(2,3)); plot(mod, 1:6, ask=FALSE); par(opar)
```



Summary of case statistics

- Leverage: observations that might be influential
- Cook's distance: observations had large overall influence on their own
  - If influential, fit with and without to determine impact on questions of interest
- Residuals: observations are not being fit accurately by the model

Check out this app (on campus or VPN):
`http://shiny1.stat.iastate.edu/_Statistics/14-outlier/`