STAT 401A - Statistical Methods for Research Workers

Case statistics

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 16, 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.

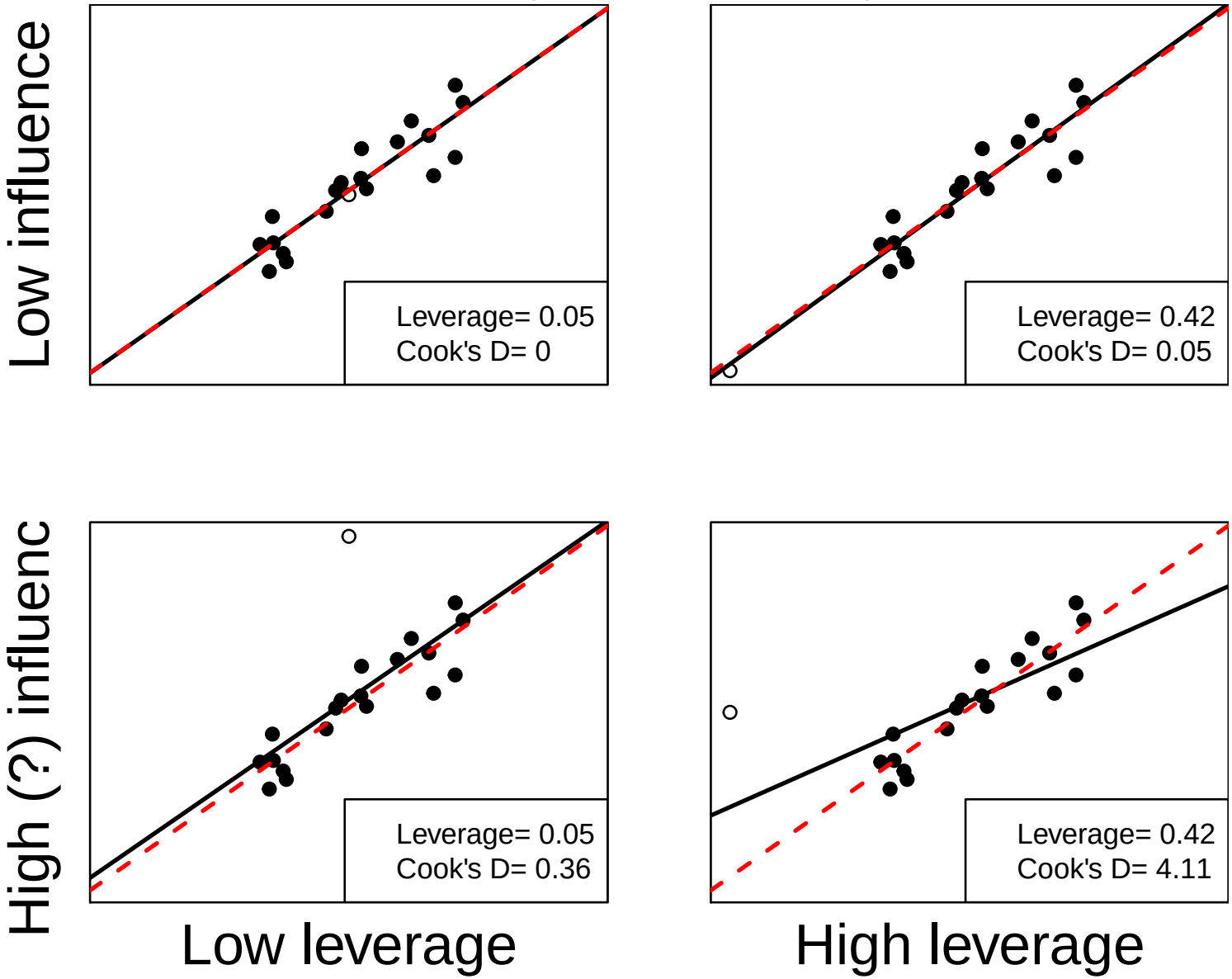
Rule-of-thumb: $> 2p/n$ where p is the number of regression coefficients and n is the number of observations.

Definition

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

Rule-of-thumb: ~ 1.

Consider simple linear regression (solid data point):



Residuals

- Residual (observed minus predicted):

$$e_i = Y_i - \hat{\mu}_i$$

- (Internally) studentized residual

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

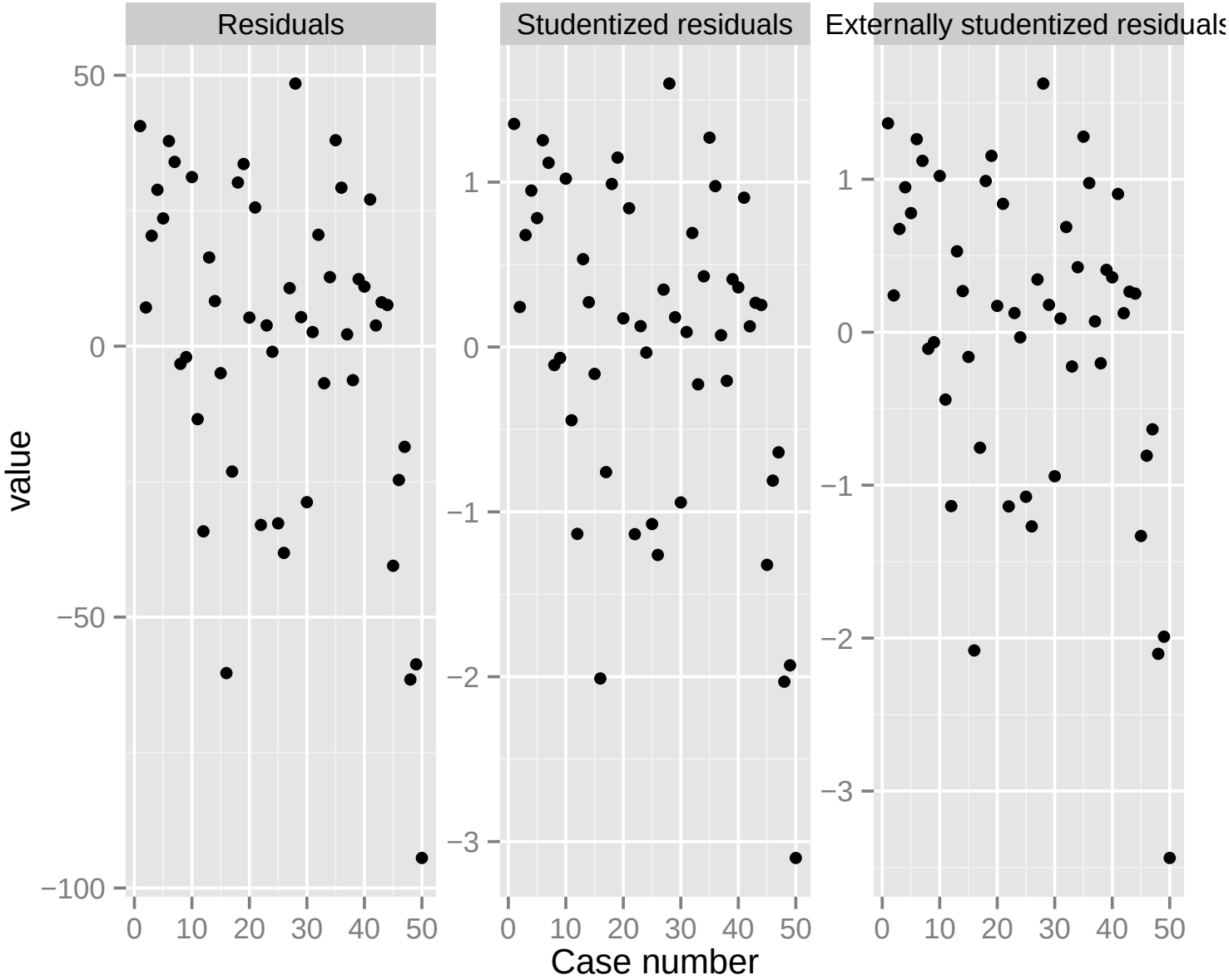
- Externally studentized residuals

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

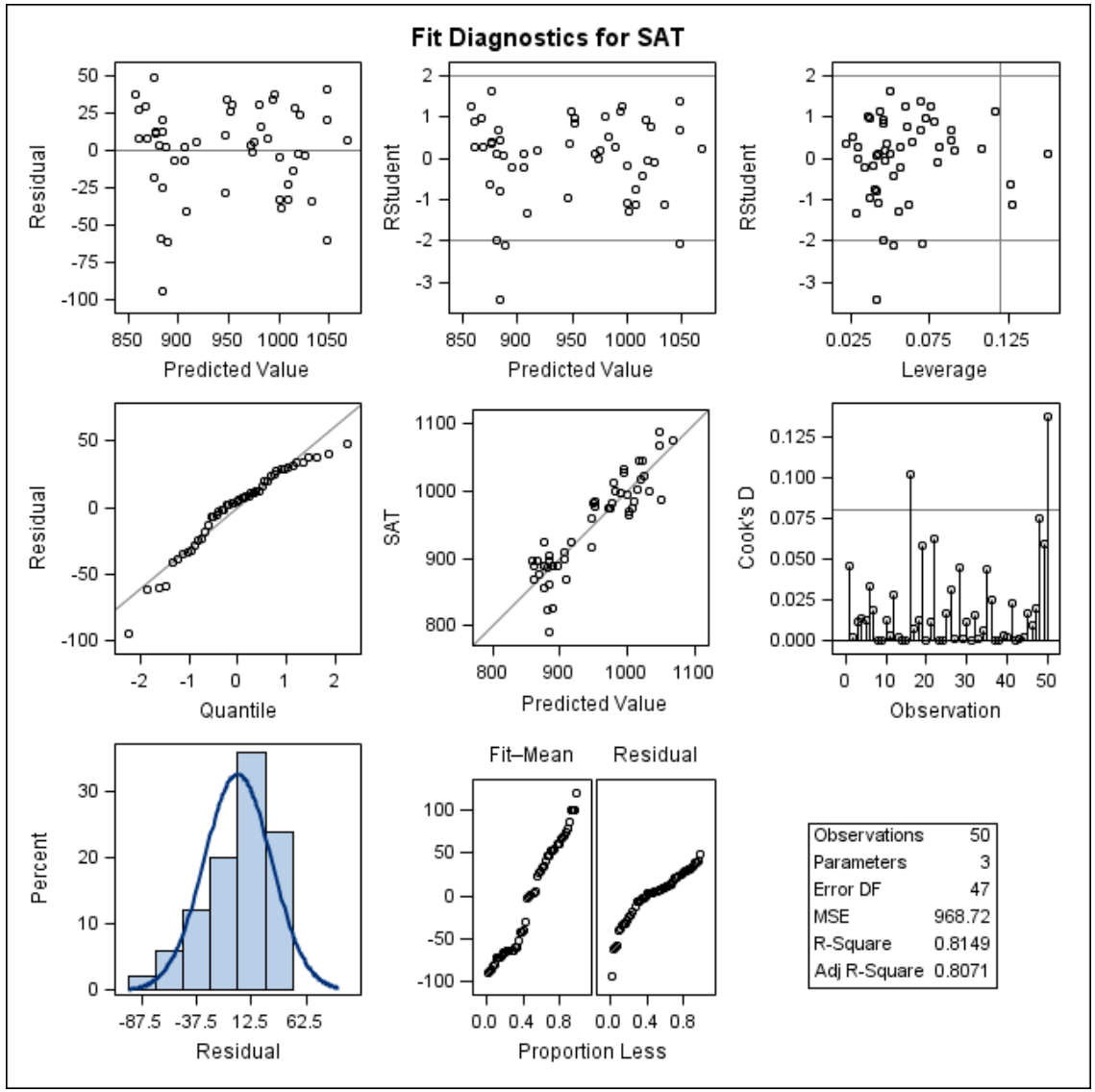
where $\hat{\sigma}_{(i)}$ is the estimate of the standard deviation about the regression line from the fit that excludes observation i.

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

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



SAS diagnostics:



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/`