More on Unusual and Influential Data

Measuring Influence

One way to examine the influence of an observation is through its effect on estimated regression coefficients, such as

$$\text{DFBETA}_{ij} = D_{ij} = B_j - B_{j(-i)},$$

or with a scaled version

$$\text{DFBETAS}_{ij} = D^*_{ij} = \frac{D_{ij}}{SE_{(-i)}(B_j)},$$

both of which focus on the change in the estimated coefficient by leaving out the current observation. These measures are useful, but there are many of them to examine. The change in the entire vector of estimated coefficients can be calculated as:

$$\mathbf{d}_i = \mathbf{b} - \mathbf{b}_{(-i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{E_i}{1-h_i}.$$

Another influence measure, Cook's D, summarizes the influence of an observation on the entire vector of estimated coefficients:

$$\begin{aligned} D_i &= \frac{(\mathbf{b} - \mathbf{b}_{(-i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})}{(k+1)S_E^2} = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(-i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(-i)})}{(k+1)S_E^2} \\ &= \left(\frac{E_i'^2}{k+1}\right)\left(\frac{h_i}{1-h_i}\right). \end{aligned}$$

A related measure is DFFITS_i, which is defined as:

$$\text{DFFITS}_i = E_i^* \sqrt{\frac{h_i}{1-h_i}},$$

which will typically satisfy $D_i \approx \text{DFFITS}_i^2/(k+1)$. Since these measures all depend on residuals and hat values, a graphical alternative is to plot E_i^* against h_i values using reference lines like 0 and ± 2 for the studentized residuals and $2\overline{h}$ and $3\overline{h}$ for the hat values.

Influence on Standard Errors and Collinearity

The effect of an individual data point on the standard errors of the vector of estimated coefficients can be summarized in terms of the squared size of the

joint confidence region for the parameter vector estimate. The COVRATIO measure proposed by Belsley *et al* (1980) approximates the squared ratio of volumes of the deleted and full-data confidence regions for the vector of regression coefficients:

$$\text{COVRATIO}_i = \frac{1}{(1-h_i)\left(\frac{n-k-2+E_i^{*2}}{n-k-1}\right)^{k+1}}.$$

Observations that increase precision of the estimated vector have COVRATIO values > 1, while observations that decrease precision have values < 1.

The topic of collinearity will be addressed in depth in Chapter 13, but individual data points that influence collinearity may be detected by the COVRATIO statistic or the hat value. Another way to detect data that influence collinearity is via scatterplot matrices of the covariates.

Guidelines for Numerical Values of Diagnostic Statistics

I agree with the author's stress on the graphical examination of diagnostic statistics rather than extensive use of numerical guidelines. However, guidelines can be especially helpful for creating reference lines in plots for identifying outliers and influential points.

Hat values: The mean hat value is $\overline{h} = (k+1)/n$, and multiples such as $2\overline{h}$ or $3\overline{h}$ have been suggested for identifying high leverage points.

Studentized Residuals: Under model assumptions we would expect about 5% of the residuals to be outside the interval $|E_i^*| \leq 2$.

Other measures: For D_{ij}^*, $\left|D_{ij}^*\right| > 1$ or 2 can be used, or Belsley *et al* (1980) propose $2/\sqrt{n}$ as a size-adjusted cutoff. For DFFITS_i, Chatterjee and Hadi (1988) suggest the size-adjusted cutoff of

$$|\text{DFFITS}_i| > 2\sqrt{\frac{k+1}{n-k-1}},$$

which leads to the suggestion for Cook's D_i of

$$D_i > \frac{4}{n-k-1}.$$

For COVRATIO, Belsley *et al* (1980) suggest

$$|\text{COVRATIO}_i - 1| > \frac{3(k+1)}{n}.$$