Deletion Diagnostics

Bahjat F. Qaqish
BIOS 667

Ordinary least-quares: The setup is the usual one with design matrix X of dimensions $n \times p$, parameters β, OLS estimates $\hat{\beta}$ and residual variance estimate s^2. Estimates obtained from the data with the ith observation deleted are denoted (i) as in $\hat{\beta}_{(i)}$. The diagonal elements of the hat matrix are $h_i = x_i^\top (X^\top X)^{-1} x_i$. Since $X^\top X = \sum_{i=1}^n x_i x_i^\top$ and $X^\top y = \sum_{i=1}^n y_i x_i$, after deleting the ith observation, $X^\top X$ becomes $X^\top X - x_i x_i^\top$ and $X^\top y$ becomes $X^\top y - y_i x_i$. The fitted values are $\hat{\mu}_i = x_i^\top \hat{\beta}$, and the residuals are $r_i = y_i - \hat{\mu}_i$. The standardized residuals are

$$\frac{r_i}{\sqrt{(1-h_i)}}$$

and the Studentized standardized residuals are the scaled version

$$\frac{r_i}{s\sqrt{(1-h_i)}}.$$

A matrix formula: For matrix A and (column) vectors b and c,

$$(A + bc^\top)^{-1} = A^{-1} - \frac{A^{-1} b c^\top A^{-1}}{1 + b^\top A^{-1} c},$$

assuming the dimensions are conformable and the inverses all exist.

In the OLS setup, take $A = X^\top X$, $b = x_i$ and $c = -x_i$. This gives an expression for $(X^\top X - x_i x_i^\top)^{-1}$, which is then used to obtain,

$$\hat{\beta} - \hat{\beta}_{(i)} = (X^\top X)^{-1} x_i \frac{r_i}{1-h_i}.$$

Cook's distance is a measure of influence of an observation on the parameter estimates $\hat{\beta}$. It is usually defined as (the quadratic form)

$$D_i = \frac{1}{ps^2} (\hat{\beta} - \hat{\beta}_{(i)})^\top (X^\top X) (\hat{\beta} - \hat{\beta}_{(i)}).$$

Note: There are various related but slightly different forms of Cook's distance in the literature. The above reduces to

$$D_i = \frac{r_i^2}{ps^2} \frac{h_i}{(1-h_i)^2},$$

which shows that influence is essentially the product of residual (r_i^2 or $(r_i/s)^2$) times leverage ($h_i/(1-h_i)^2 = h_{(i)}/(1-h_i)$).

Note that the above formulae give *exact* results, not approximations. There is no need to refit the model n times. Once the fitted values and the hat matrix are available, all the above diagnostics are easily computed.

Weighted least-quares (WLS): The problem is transformed to OLS by defining $t_i = y_i \sqrt{w_i}$ and $z_i = x_i \sqrt{w_i}$. Now, all the OLS formulae given above apply to (t, Z).

Generalized linear models for independent outcomes: These models require iteration to find estimates. Approximations (usually accurate) to the quantities that would be computed after full iteration are possible: The iterative weights are used as if they were the weights in WLS.

GEE: Here two deletions are possible; either one observation or one cluster. So there are two versions of each of the quantities defined above; one for observation deletion and one for cluster deletion.

References: The (minimum) recommended reading for OLS is chapter 13 in the BIOS 663 textbook. The classic books on the topic are Cook & Weisberg (1982) and Belsley, Kuh & Welsch (1980) - both highly recommended. For generalized linear models for independent outcomes, section 12.7 of the book by McCullagh & Nelder (1989, 2nd ed.) gives the basics. For dependent outcomes and GEE, see Preisser & Qaqish (1996, Biometrika).