MORE ON LEVERAGE AND VARIANCES OF RESIDUALS

Recall: In simple linear relation, to establish that

$$\mathrm{Var}(y - \hat{y}|x) = \sigma^2(1 + \text{leverage})$$

for a new observation y (chosen independently of the y_i's), we reasoned that since y and $\hat{y}$ are independent,

$$\mathrm{Var}(y - \hat{y}|x) = \sigma^2 + \mathrm{Var}(\hat{y}|x)$$

$$= \sigma^2 + \mathrm{Var}(\hat{E}(Y|x))$$

Similar reasoning is used to establish the result for multiple regression.

But we *cannot* apply this to find $\mathrm{Var}(y_i - \hat{y}|x)$. *Why not?*

Instead, we need to take covariances into account.

The result (generalized to multiple regression):

$$\mathrm{Var}(\hat{e}_i|\underline{x}) = \sigma^2(1 - h(\underline{u}_i))$$

Notation:

$$h_i = h_{ii}$$

$$= h(\underline{u}_i)$$

$$(= h(\underline{x}_i) \text{ by abuse of notation})$$

$= \textit{the } i^{th} \textit{ leverage}.$

So:

$$\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_i)$$

Consequence:

Since $\mathrm{Var}(\hat{e}_i) \geq 0$,

$$h_i \leq 1.$$

Note:

i) h could be > 1 for other values of $\underline{x}$.

ii) $h \geq 0$ since $\mathrm{Var}(\hat{E}(Y|\underline{x})) = h\sigma^2$

Practical consequence:

If h_i is close to 1, then Var($\hat{e}_i$) is small.

Recalling that E($\hat{e}_i$) = 0, this implies that $\hat{e}_i$ is small -- so the least squares fit is close to ($\underline{u}_i$, y_i).

In other words:

If h_i is close to 1, then $\underline{x}_i$ is influential.

Thus: *Check leverages to identify possible influential observations.*