MORE ON LEVERAGE AND VARIANCES OF RESIDUALS

(Reference: Section 7.6.3)

Recall: In simple linear regression, to establish that $\mathrm{Var}(y - \hat{y}|x) = \sigma^2(1 + \text{leverage})$ for a new observation from Y|x (chosen independently of the y_i's), we reasoned that since y and $\hat{y}$ are independent,

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

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

Instead, we need to go through a procedure much like that in finding $\mathrm{Var}(\hat{E}(Y|x))$, taking covariances into account. The result, generalized to multiple regression:

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

Notation: $h_i = h_{ii} = h(\underline{u}_i)$ (= $h(\underline{x}_i)$ by abuse of notation) is called the i^{th} *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 (which is large for a leverage), then $\mathrm{Var}(\hat{e}_i)$ 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 *it is advisable to check leverages to identify possible influential observations.*