

## MORE ON LEVERAGE AND VARIANCES OF RESIDUALS

(Reference: Section 7.6.3)

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

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

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

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

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

*Notation:*  $h_i = h_{ii} = h(\underline{u}_i)$  ( $= h(\underline{x}_i)$  by abuse of notation) is called the  $i^{\text{th}}$  leverage.

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

*Consequence:* Since  $\text{Var}(\hat{e}_i) \geq 0$ ,

$$h_i \leq 1.$$

*Note:*

- i)  $h$  could be  $> 1$  for other values of  $\underline{x}$ .
- ii)  $h \geq 0$  since  $\text{Var}(\hat{E}(Y|\underline{x})) = h\sigma^2$

*Practical consequence:* If  $h_i$  is close to 1 (which is large for a leverage), then  $\text{Var}(\hat{e}_i)$  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.*