Identifying influential points 10.2 ~ 10.4.

(I) $\hat{\underset{\sim}{Y}} = H \underset{\sim}{Y}$ w/ $H = (h_{ij})_{n\times n}$. $\sum_{i=1}^{n} h_{ii} = p$.

$0 \le h_{ii} \le 1$, $h_{ii} \approx \frac{p}{n}$ $\forall i$.

Note. $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{i \ne j} h_{ij} Y_j$.

$\Rightarrow$ h_{ii} : the effect of Y_i on $\hat{Y}_i$.

$\hookrightarrow$ Flag i-th case if $h_{ii} > \frac{2p}{n}$, say.

(II) Deleted residuals.

- remove the i-th case from the data $(\underset{\sim}{Y}, \mathbb{X})$
  $\hookrightarrow$ resulting data $(\underset{\sim}{Y}_{(i)}\ \mathbb{X}_{(i)})$

$$\underset{\sim}{Y}_{(i)} = \mathbb{X}_{(i)} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad \text{w/} \quad \underset{\sim}{\varepsilon} \sim N_{n-1}(0, \sigma^2 I)$$

$\Rightarrow$ $\underset{\sim}{b}_{(i)}$: LSE of β using $(\underset{\sim}{Y}_{(i)}, \mathbb{X}_{(i)})$,

$\Rightarrow$ $\hat{Y}_{i(i)} = (\mathbb{X}\, \underset{\sim}{b}_{(i)})_i$: fitted value for the ith case

$= \underset{\sim}{x}_i^t\, b_{(i)}$ $\quad$ $\underset{p\times 1}{\underset{\sim}{x}_i}^t$: i-th row of $\mathbb{X}$.

Def. $d_i = Y_i - \hat{Y}_{i(i)}$: deleted residual for the ith case

$\Rightarrow$ $E(d_i) = E(Y_i) - E(\hat{Y}_{i(i)}) = \underset{\sim}{x}_i^t \underset{\sim}{\beta} - \underset{\sim}{x}_i^t \underset{\sim}{\beta} = 0$

$$\begin{aligned}
\sigma^2\{d_i\} &= \sigma^2 + \sigma^2\{\hat{Y}_{i(i)}\} \qquad \because Y_i \perp \hat{Y}_{i(i)} \\
&= \sigma^2 + \underset{\sim}{x}_i^t\, \sigma^2\{\underset{\sim}{b}_{(i)}\}\, \underset{\sim}{x}_i \\
&= \sigma^2 + \sigma^2\, \underbrace{\underset{\sim}{x}_i^t (\mathbb{X}_{(i)}^t \mathbb{X}_{(i)})^{-1} \underset{\sim}{x}_i}_{\overset{\Delta}{=}\, A}
\end{aligned}$$

unbiased est. of σ^2 using $(\underset{\sim}{Y}_{(i)}, \mathbb{X}_{(i)})$

is $MSE_{(i)} = \underset{\sim}{e}_{(i)}^t \underset{\sim}{e}_{(i)} / n-1-p$

w/ $\underset{\sim}{e}_{(i)} = \underset{\sim}{Y}_{(i)} - \mathbb{X}_{(i)} \underset{\sim}{b}_{(i)}$

$$\sigma^2\{d_i\}\Big|_{\sigma^2 = MSE_{(i)}} = s^2\{d_i\}$$

$\Rightarrow$

as usual $\dfrac{d_i}{s\{d_i\}} \sim t(n-1-p)$

i.e. $t_i^* \triangleq \dfrac{d_i}{s\{d_i\}} \sim t(n-1-p)$, $i = 1 \cdots n$

$\Rightarrow$ i-th case is an outlier / influential

if $|t_i^*| > t(1 - \frac{\alpha}{2n} ; n-1-p)$

↑ Bonferroni correction ---

(III) Cook's distance

$$D_i \triangleq \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ MSE}$$

$$= \frac{(\hat{\underset{\sim}{Y}} - \hat{\underset{\sim}{Y}}_{(i)})^t (\hat{\underset{\sim}{Y}} - \hat{\underset{\sim}{Y}}_{(i)})}{p\ MSE}$$

$\hat{\underset{\sim}{Y}} = H\underset{\sim}{Y} = \mathbb{X} \cdot b$

$\hat{\underset{\sim}{Y}}_{(i)} = \mathbb{X} \cdot b_{(i)}$

$MSE = \dfrac{(\underset{\sim}{Y} - \hat{\underset{\sim}{Y}})^t (\underset{\sim}{Y} - \hat{\underset{\sim}{Y}})}{n-p}$

↳ i-th case : major influence if $D_i > F(0.5 ; p, n-p$

-- little apparent influence if $D_i > F(0.2 ; p, n-p$

Rmk : t_i^* , D_i ↳ we need to fit another model --!
and $i = 1 \cdots n$!

↳ t_i^*, D_i's can be obtained using the fit w/ all data pts.

Note $\underset{\sim}{x}_{i\,1\times p}^{t}$: i-th row of $\mathbb{X}$.

$\therefore \quad \mathbb{X} = \begin{bmatrix} \underset{\sim}{x}_1^t \\ \underset{\sim}{x}_2^t \\ \vdots \\ \underset{\sim}{x}_n^t \end{bmatrix}$ and $\mathbb{X}_{(i)} = \begin{bmatrix} \underset{\sim}{x}_1^t \\ \vdots \\ \underset{\sim}{x}_{i-1}^t \\ \underset{\sim}{x}_{i+1}^t \\ \vdots \\ \underset{\sim}{x}_n^t \end{bmatrix}$ ← no i-th

$$\Rightarrow \mathbb{X}^t\mathbb{X} = \sum_{j=1}^n \underset{\sim}{x}_j \underset{\sim}{x}_j^t , \qquad \mathbb{X}_{(i)}^t \underset{\sim}{Y}_{(i)} = \mathbb{X}^t\underset{\sim}{Y} - \underset{\sim}{x}_i Y_i$$

$$\Rightarrow (\mathbb{X}_{(i)}^t \mathbb{X}_{(i)})^{-1} = \Big(\sum_{\substack{j=1\\ j\neq i}}^n \underset{\sim}{x}_j \underset{\sim}{x}_j^t\Big)^{-1} = \left(\mathbb{X}^t\mathbb{X} - \underset{\sim}{x}_i \underset{\sim}{x}_i^t\right)^{-1}$$

$$\overset{\text{Lemma 10}}{=\!=} (\mathbb{X}^t\mathbb{X})^{-1} + \frac{(\mathbb{X}^t\mathbb{X})^{-1}\underset{\sim}{x}_i \underset{\sim}{x}_i^t (\mathbb{X}^t\mathbb{X})^{-1}}{1 - \underbrace{\underset{\sim}{x}_i^t (\mathbb{X}^t\mathbb{X})^{-1}\underset{\sim}{x}_i}_{= h_{ii}}} \quad \cdots\cdots ①$$

$$\Rightarrow A = \underset{\sim}{x}_i^t (\mathbb{X}_{(i)}^t \mathbb{X}_{(i)})^{-1}\underset{\sim}{x}_i = \underset{\sim}{x}_i^t \; ① \; \underset{\sim}{x}_i$$

$$= h_{ii} + \frac{h_{ii}^2}{1-h_{ii}} \quad \cdots\cdots (*1)$$

and $\underset{\sim}{b}_{(i)} = (\mathbb{X}_{(i)}^t \mathbb{X}_{(i)})^{-1} \mathbb{X}_{(i)} \underset{\sim}{Y}_{(i)} = \cdots\cdots$

$$= \underset{\sim}{b} + \frac{(\mathbb{X}^t\mathbb{X})^{-1}\underset{\sim}{x}_i (\underset{\sim}{x}_i^t \underset{\sim}{b} - Y_i)}{1-h_{ii}} = \underset{\sim}{b} - \frac{(\mathbb{X}^t\mathbb{X})^{-1}\underset{\sim}{x}_i e_i}{1-h_{ii}} \quad \cdots (*2)$$

$$\Rightarrow d_i - e_i = \hat{Y}_i - \hat{Y}_{i(i)} = \underset{\sim}{x}_i^t(\underset{\sim}{b} - \underset{\sim}{b}_{(i)}) = \frac{h_{ii}e_i}{1-h_{ii}} \Rightarrow \boxed{d_i = \frac{e_i}{1-h_{ii}}}$$

$$MSE_{(i)}(n-1-p) = \sum_{\substack{j=1\\ j\neq i}}^n (Y_j - \underset{\sim}{x}_j^t \underset{\sim}{b}_{(i)})^2$$

$$= \sum_{\substack{j=1\\ j\neq i}}^n (\underline{Y_j - \underset{\sim}{x}_j^t \underset{\sim}{b}} + \underline{\underset{\sim}{x}_j^t \underset{\sim}{b} - \underset{\sim}{x}_j^t \underset{\sim}{b}_{(i)}})^2$$

$$= \sum_{\substack{j=1\\ j\neq i}}^n (e_j + \underset{\sim}{x}_j^t(\underset{\sim}{b} - \underset{\sim}{b}_{(i)}))^2 = \cdots\cdots \quad \text{by } (*2)$$

$$= \sum_{j\neq i} \left(e_j + \frac{h_{ij}e_i}{1-h_{ii}}\right)^2$$

$$= \cdots\cdots \qquad \sum h_{ij}^2 = h_{ii}, \ \sum h_{ij} e_j = 0$$

$$= \sum e_j^2 - \frac{e_i^2}{1-h_{ii}} = MSE(n-p) - e_i^2 \frac{1}{1-h_{ii}}$$

$$= SSE - e_i^2 \frac{1}{1-h_{ii}}$$

I, e.

$$MSE_{(i)} = \left(SSE - \frac{e_i^2}{1-h_{ii}}\right) \Big/ (n-p-1) \quad \cdots\cdots \quad (*3).$$

$(*1), (*3) \Rightarrow \quad s^2\{d_i\} = MSE_{(i)} \left(1 + h_{ii} + \frac{h_{ii}^2}{1-h_{ii}}\right)$

and $\quad d_i = \dfrac{e_i}{1-h_{ii}}$

$$\Rightarrow \ t_i^* = \frac{d_i}{s\{d_i\}} = \frac{e_i\sqrt{n-p-1}}{\sqrt{(1-h_{ii})SSE - e_i^2}}, \quad \forall i = 1, 2 \cdots n$$

calculated from the $\overset{\text{reg.}}{\wedge}$ fit w/ all data points.

$(*2)$

$$\Rightarrow \ D_i = (\hat{Y} - \hat{Y}_{(i)})^t (\hat{Y} - \hat{Y}_{(i)}) / pMSE$$

$$= (b - b_{(i)})^t X^t X (b - b_{(i)}) / p \cdot MSE$$

$$\overset{(*2)}{=} \left(\frac{(X^tX)^{-1} x_i e_i}{1-h_{ii}}\right)^t X^t X \left(\frac{(X^tX)^{-1} x_i e_i}{1-h_{ii}}\right) \Big/ p \cdot MSE$$

$$= \frac{e_i^2 h_{ii}}{(1-h_{ii})^2 \cdot p \cdot MSE}, \quad \forall i = 1, 2 \cdots n.$$

IV. $(DFFITS)_i = \dfrac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = \cdots = t_i^* \left(\dfrac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$

Flag i if $|\ | > 1$, small ~ moderate; if $|\ | > 2\sqrt{\frac{p}{n}}$ large data set.

$(DFBETAS)_{k(i)} = \dfrac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} C_{kk}}}$ $\quad C_{kk} = k\text{-th diag. of } (X^tX)^{-1}$.

Flag i if $|\ | > 1$. small to moderate data set

if $|\ | > \frac{2}{\sqrt{n}}$ large data set.

etc. $\cdots$ ~

Lemma 10

$$Q_{n\times n} = P_{n\times n} + U_{n\times q} V_{q\times n}$$

$$\Rightarrow Q^{-1} = P^{-1} - P^{-1} U (I_{q\times q} + V P^{-1} U)^{-1} V P^{-1}$$

<pf>

$$Q^{-1} = [P + UV]^{-1}$$

$$\Rightarrow Q^{-1}[P + UV] = I$$

$$\parallel$$

$$Q^{-1}P + Q^{-1}UV$$

$$\parallel$$

$$[Q^{-1} + Q^{-1}UVP^{-1}]P$$

$$\Rightarrow P^{-1} = Q^{-1} + Q^{-1}UVP^{-1} \Rightarrow Q^{-1} = P^{-1} - Q^{-1}UVP^{-1} \quad ①$$

and $$P^{-1}U = Q^{-1}U + Q^{-1}UVP^{-1}U \quad ②$$

$$= Q^{-1}U[I_q + VP^{-1}U]$$

$$② \Rightarrow \underline{Q^{-1}U} = P^{-1}U[I_q + VP^{-1}U]^{-1}$$

$$① \Rightarrow Q^{-1} = P^{-1} - \underline{Q^{-1}U}VP^{-1}$$

$$= P^{-1} - \underline{P^{-1}U[I_q + VP^{-1}U]^{-1}}VP^{-1}$$

#.

Take. $Q = X_{(i)}^{t} X_{(i)}$ $P = X^{t}X$ $U = -x_i$ $V = x_i^{t}$

$\Rightarrow \cancel{Q^{-1} = P^{-1}}$.