Chapter 6. Diagnostics for Leverage and Influence

Detection of Influential Observations

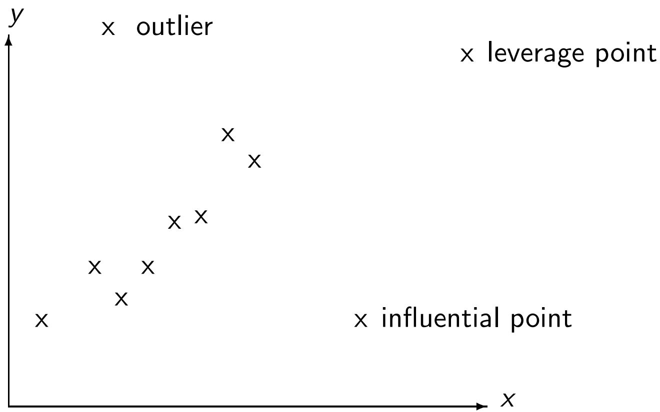
- Large residuals $\Longrightarrow$ outliers, *not good*.

  It does not necessarily mean that the observations are influential in fitting the chosen model.

- Unusual x-value (far away from others) $\Longrightarrow$

  (i) **leverage point**: may not affect the parameters estimation, but can have a dramatic effect on the model summary statistics, R^2, $\hat{\sigma^2}$, etc.

  (ii) **influential point**: y-value is also unusual. It has a noticeable impact on the model coefficients; it "pulls" the regression model in its direction.



y
x outlier
x leverage point
x influential point
x

Recall: 1. $H = X(X'X)^{-1}X'$ is called the "*hat matrix*".

2. $H = (h_{ij}), \quad H' = H$ and $H^2 = H$.

Note: 1. $Var(\mathbf{e}) = \sigma^2(I - H)$.

2. $Var(\hat{\mathbf{y}}) = \sigma^2 H$.

3. $\sum_{i=1}^{n} Var(\hat{y}_i)/n = \sigma^2 \sum_{i=1}^{n} h_{ii}/n = \sigma^2 tr(H)\ /n$

$= \sigma^2 rank(H)/n = p\sigma^2/n.$

Note: 4. $\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$.

5. h_{ij} indicates how heavily y_j contributes to $\hat{y}_i$,

   h_{ii} = leverage.

6. $tr(H) = \sum_{i=1}^{n} h_{ii} = p$. If $h_{ii} >> \frac{p}{n}$, **high** leverage.

   Traditionally, if $h_{ii} > \frac{2p}{n}$ (as $\frac{2p}{n} < 1$), then it is of high leverage.

7. Observations with large *leverages* **and** large *residuals* are likely to be **influential**.

Measures of Influence

1. The Cook's distance.

$$\begin{aligned} D_i &= (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})/(p\hat{\sigma^2}) \\ &= (\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})/(pMS_{Res}) \\ &= \text{The distance between } \hat{\beta} \text{ and } \hat{\beta}_{(i)} \text{ standardized by} \\ & \quad \widehat{Var}(\hat{\beta}) = MS_{Res}(X'X)^{-1}, \end{aligned}$$

where $\hat{\beta}_{(i)}$ = LSE when the i^{th} observation is deleted from the data.

Note: 1. If deletion of the i^{th} observation makes little difference to the fitted values, D_i will be *small* $\Longrightarrow$ **not** influential.

2. Usually, $D_i > 1$ is considered to be influential.

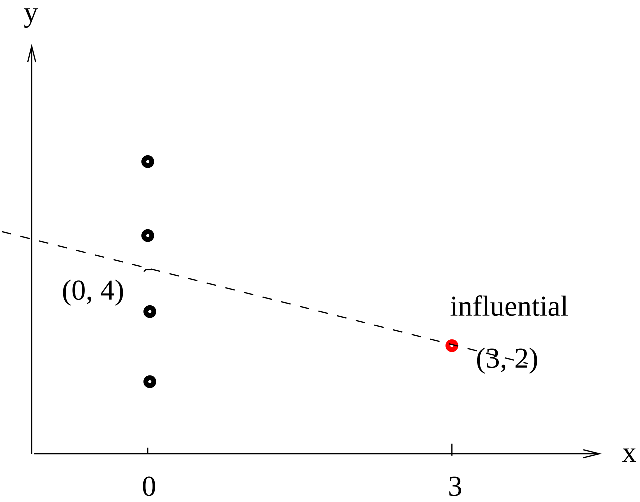
Note that $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (X'X)^{-1}\mathbf{x}_i e_i/(1-h_{ii})$ (*Exercise.*)

$$\begin{aligned}
\therefore \quad D_i &= \left(\frac{(X'X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right)'(X'X)\left(\frac{(X'X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right)/(\widehat{p\sigma^2}) \\
&= e_i\mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i e_i \,/\, [\widehat{p\sigma^2}(1-h_{ii})^2] \\
&= \left(\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2\left(\frac{h_{ii}}{1-h_{ii}}\right)\frac{1}{p} \\
&= (\text{internally studentized residual, } r_i)\cdot\left(\frac{Var\ \hat{y}_i}{Var\ e_i}\right)(\frac{1}{p})
\end{aligned}$$

$\therefore D_i$ large $\iff$ high *residual* (large response) or high *leverage* (large x-value).

Ex.1. $n = 5,\ p = 2,$

$x = 0,\ y_1 = 1,\ y_2 = 3,\ y_3 = 5,\ y_4 = 7;\ \ x = 5,\ y_5 = 2.$



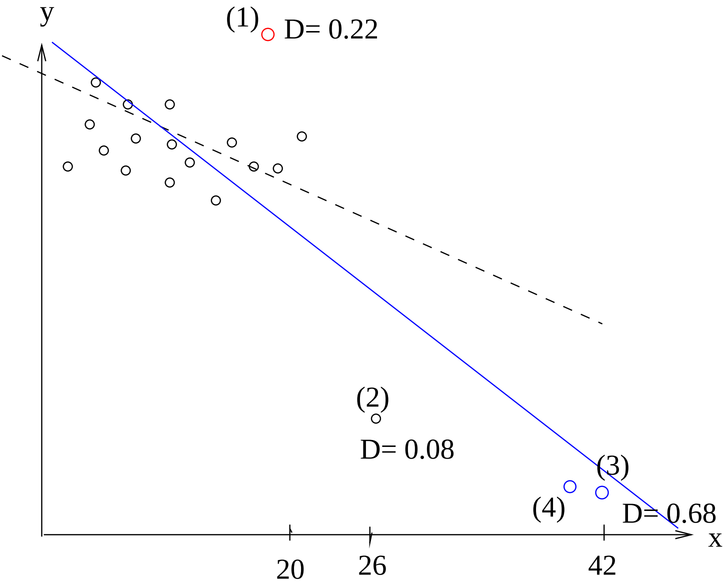
Ex.1. (Cont'd)

$$e_i = y_i - 4,\ i = 1,2,3,4;\quad e_5 = 0.$$

$$Var(e_i) = 0.75\sigma^2,\ i = 1,2,3,4;\quad Var(e_5) = 0.$$

$$D_i = \frac{e_i^2}{0.75\hat{\sigma^2}}\frac{1}{3}\frac{1}{2},\ i = 1,2,3,4;\quad D_5 = \text{indeterminate } (h_{55} = 1).\quad \square$$

Ex.2. x = age (in months) of a child at first word.



Ex. 2.(Cont'd)

(1) $D_i = 0.22$ outlier.

(2) $D_i = 0.08$ (3) $D_i = 0.68$ large x-values.

Not sensible to collect data in between.

Here $x = 26$ or $x = 42$, unusual; not reliable.

Note: If (4) were also observed, then (3)(4) would both be influential. However, delete (3) or (4) will not reduce much on D. $\square$

<u>Higher-order Cook's distance</u>: Omit pairs, triplets, etc., then compute the distance of $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_{(.)}$.

<u>Ex.</u>2.(Cont'd)

Omitting (2)(3) (together) gives large Cook's distance (0.67).

$\therefore$ (2)(3) highly influential. $\square$

Read pp. 191 – 195 in Text.

2. The DFFITS Statistics.

Define the number of standard deviations that the fitted value $\hat{y}_i$ changes if the i^{th} observation is removed to be

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}.$$

Note that

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \left[\frac{e_i}{S_{(i)}(1-h_{ii})^{1/2}}\right] = \sqrt{\frac{h_{ii}}{1-h_{ii}}} \cdot t_i$$

Note: 1. Recall that t_i = R-student residual.

2. $|DFFITS_i| > 2\sqrt{p/n}$ needs attention.

3. The DFBETAS Statistics.

Influence on regression coefficients can be measured by

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}, \quad (X'X)^{-1} = (C_{ij}).$$

It indicates the influence of observation i on $\hat{\beta}_j$.

Let $R = (X'X)^{-1}X' = (r_{ji}) = \begin{pmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \\ \vdots \\ \mathbf{r}_p' \end{pmatrix}$, so

$$DFBETAS_{j,i} = \frac{r_{ji}}{\sqrt{\mathbf{r}_j'\mathbf{r}_j}} \frac{e_i}{S_{(i)}(1 - h_{ii})} = \frac{r_{ji}}{\sqrt{\|\mathbf{r}_j\|^2}} \frac{t_i}{\sqrt{1 - h_{ii}}}.$$

Note: $|DFBETAS_{j,i}| > 2/\sqrt{n}$ calls for attention.

Measures of Model Performance

Consider the ratio of the determinants of the estimated covariance matrices

$$COVRATIO_i = \frac{\det(S_{(i)}^2 (X_{(i)}' X_{(i)})^{-1})}{\det(\hat{\sigma}^2 (X'X)^{-1})}.$$

Note that

$$\begin{aligned}
\det(X_{(i)}' X_{(i)}) &= \det(X'X - \mathbf{x}_i \mathbf{x}_i') = \det[(X'X)(I - (X'X)^{-1} \mathbf{x}_i \mathbf{x}_i')] \\
&= \det(X'X) \det(1 - \mathbf{x}_i' (X'X)^{-1} \mathbf{x}_i) \\
&= [\det(X'X)](1 - h_{ii})
\end{aligned}$$

$$\therefore \; COVRATIO_i = \left(\frac{S_{(i)}^2}{\hat{\sigma}^2} \right)^p (1 - h_{ii})^{-1}.$$

Note: 1. High leverage $\Longrightarrow$ $COVRATIO_i$ large $\uparrow$

$\Longrightarrow$ not influential. (Improve the precision.)

2. Outlier: $S_{(i)}^2/\hat{\sigma^2} << 1$.

3. $COVRATIO_i \rightarrow 1$, okay;

influential if it is in $[1 - \frac{3p}{n}, 1 + \frac{3p}{n}]^c$.

Homework 7: (Page 199) 6.1, 6.2, 6.10, 6.11, 6.15.

Due: Dec. 19, 2008.