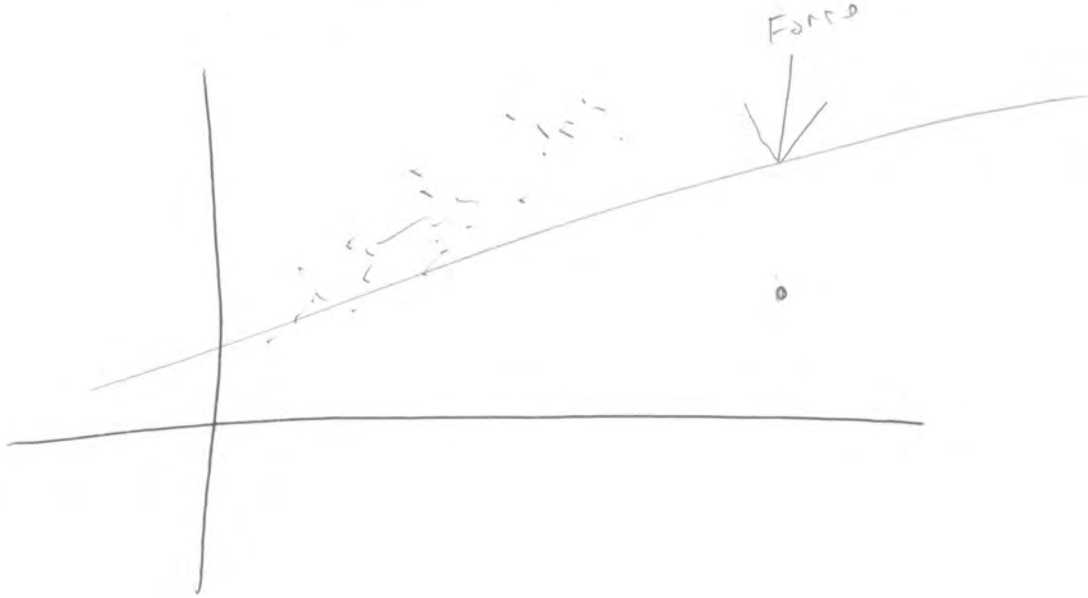
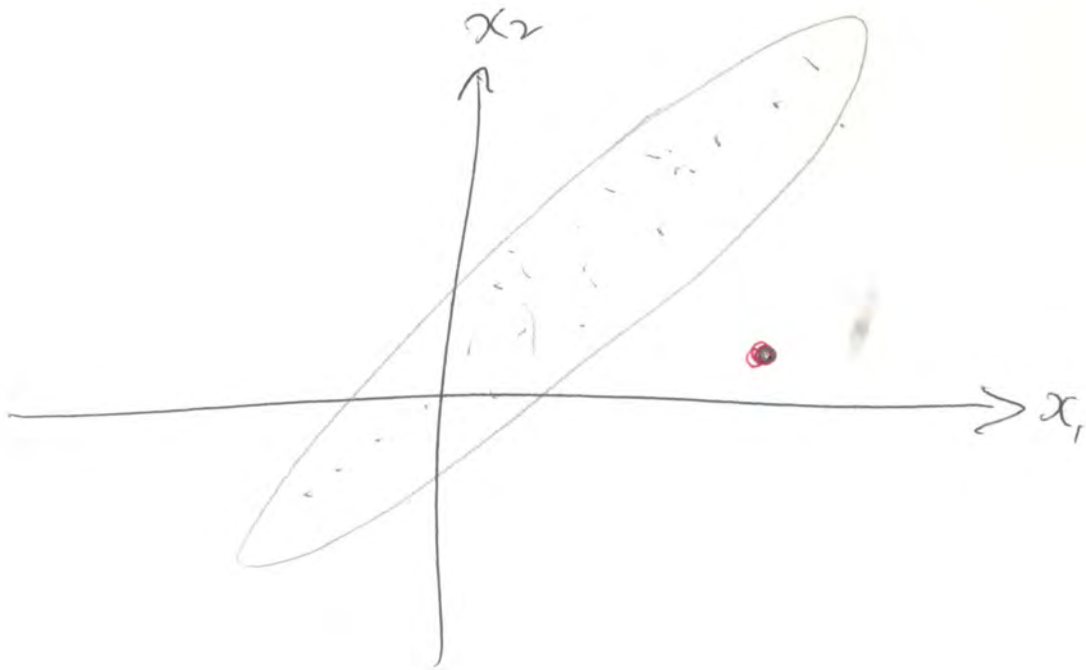
Influential Observations

1

An outlier in x can exert "leverage" on the least squares line



Outlier in multiple dimensions can be hard to spot



Starred section 8.2.1, p. 156

2

(a) h_{ii} <u>indirectly</u> reflects how far $\underset{\sim}{x}_i$ is from $\bar{\underset{\sim}{x}}$, vector of IV sample means ("centroid")

Transpose of row i of X

(b) $0 \le h_{ii} \le 1$

$$\mathrm{cov}(\hat{y}) = \sigma^2 H \text{ so } h_{ii} \ge 0$$

$$\mathrm{cov}(e) = \sigma^2 (I - H), \text{ so}$$

$$\mathrm{var}(e_i) = \sigma^2 (1 - h_{ii}) \ge 0 \Rightarrow h_{ii} \le 1$$

(c) Average h_{ii} is small

$$\mathrm{tr}(H) = \mathrm{tr}\left(X (X'X)^{-1} X' \right) = k+1, \text{ and}$$

$$\bar{h} = \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{k+1}{n} \rightarrow 0 \text{ as } n \rightarrow \infty$$

(d) Residuals e_i reflect ε_i better when h_{ii} are small

(next page)

In residual plots, we seem to be using e_i as a substitute for ε_i

It's more than
$$y = X\beta + \varepsilon$$
$$= Xb + e$$

Recall
$$e = (I-H)y = (I-H)(X\beta + \varepsilon)$$
$$= (I-H)X\beta + (I-H)\varepsilon$$
$$= X\beta - HX\beta + (I-H)\varepsilon$$
$$= X\beta - X\underbrace{(X'X)^{-1}X'X}_{I}\beta + (I-H)\varepsilon$$
$$= (I-H)\varepsilon, \quad \text{so}$$

$$e = \varepsilon - H\varepsilon$$

And we see that $e = \varepsilon$, except for $H\varepsilon$

$e_i = \varepsilon_i - \underset{\sim}{h}_i'\varepsilon$ where $\underset{\sim}{h}_i'$ is a row of H

($\&$ $\underset{\sim}{h}_i$ is a col of H)

Good way to represent this is with a <u>selection vector</u>

$$\underset{v_3'}{(0\ 0\ 1\ 0)}\ \underset{e}{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}} = e_3 \qquad \text{Picks out } e_3$$

$$\underset{v_3'}{(0\ 0\ 1\ 0)}\ \underset{A}{\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}}\ \underset{v_3}{\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}}$$

$$= (a_{31}\ a_{32}\ a_{33}\ a_{34}) \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = a_{33}$$

And in general

$$v_j' A v_j = a_{jj}$$

Now back to $e = \varepsilon - H\varepsilon$

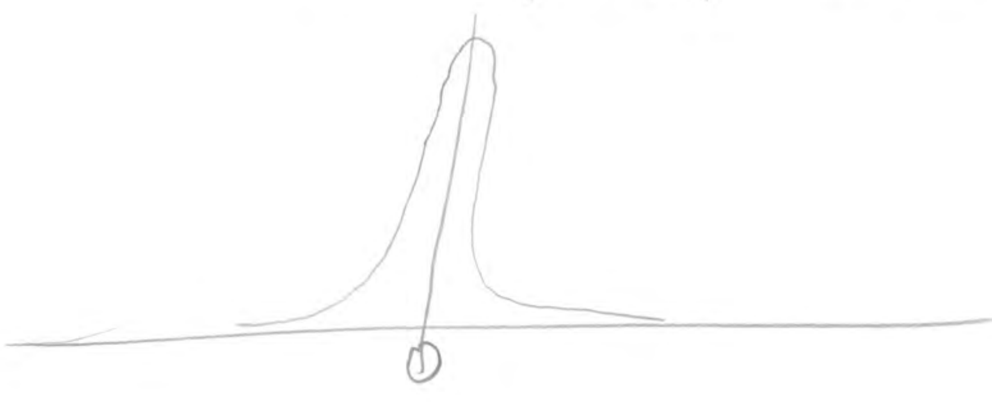
$$\Rightarrow v_i' e = v_i'\varepsilon - v_i' H \varepsilon$$

$$\Rightarrow e_i = \varepsilon_i - \underbrace{v_i' H}_{1 \times n} \varepsilon$$

A normal scalar RV

$$E(v_i' H \varepsilon) = 0$$

$$\begin{aligned} Var(v_i' H \varepsilon) &= cov(v' H \varepsilon) \\ &= v_i' H \, cov(\varepsilon) \, (v_i' H)' \\ &= v_i' H \sigma^2 I H v_i \\ &= \sigma^2 v_i' H v_i = \sigma^2 h_{ii} \end{aligned}$$



Know average $h_{ii} \longrightarrow 0$ as $n \to \infty$

As individual $h_{ii} \longrightarrow 0$, e_i gets closer to ε_i

And the diagnostic plots make more sense

SMALL h_{ii} are good, Large n is good

(e) Thm 5.1, p. 106

If $\max(h_{ii}) \to 0$ as $n \to \infty$, the distribution of b approaches $N(\beta, \sigma^2(X'X)^{-1})$ even if ε is not normal.

"Rule of thumb" $\max(h_{ii}) < 0.2$

The influence of the most influential obs $\to 0$

(2) $DFBETA = b - b_{(i)}$

Transpose of row i

$$= \frac{(X'X)^{-1} \underset{\sim}{x}_i e_i}{1 - h_{ii}}$$

(8.8), p. 158

DFBETAS: e_i^* instead of e_i

(3) $DFFIT = \hat{y}_i - \hat{y}_{(i)} = \dfrac{h_{ii} e_i}{1 - h_{ii}}$

p. 157

DFFITS

use e_i^*

(4) Thm 5.1, p. 106

~~[illegible]~~

Normality does not matter for tests & C.I.s provided $\max(h_{ii}) \to 0$ as $n \to \infty$

Rule of thumb

$\max(h_{ii}) < 0.2$

(4) Cook's Distance $D_i = \dfrac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{} = \dfrac{e_i^2}{s^2(k+1)} \left(\dfrac{h_{ii}}{1-h_{ii}}\right)^2$