

Diagnóstico

Motivação:

Em algumas análises, as estatísticas básicas podem mudar muito quando um elemento amostral é retirado. Este ponto será denominado "influente".

Técnicas de Diagnóstico:

Famos de detectar pontos influentes

1- Matriz "Hat"

$$\underline{y} = X \underline{\beta} + \underline{\varepsilon}$$

$$\hat{\underline{\beta}} = (X'X)^{-1} X' \underline{y}$$

$$\hat{\underline{y}} = X \hat{\underline{\beta}} = X (X'X)^{-1} X' \underline{y} = H \underline{y}$$

$$H = X (X'X)^{-1} X'$$

$$\hat{y}_i = h_{i1} y_1 + h_{i2} y_2 + \dots + h_{in} y_n$$

$h_{ij} \rightarrow$ influência exercida por y_j em $\hat{y}_i$

$h_{ii} \rightarrow$ " " " y_i em $\hat{y}_i$

2

H - simétrica e idempotente ($H^2 = H$) e

$$\sum_{i=1}^n h_{ii} = (k+1) \quad h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \quad (0 \leq h_{ii} \leq 1)$$

k - var. indep intercepto (ou k - modelo sem intercepto)

"Valor médio de h_{ii} " : $\frac{k+1}{n} \Rightarrow$ ~~considerar~~

Análise os elementos da diagonal de $H(h_{ii})$ e ~~considere~~ de especial

atenção a pontos $\underline{x}_i$ tais que $h_{ii} > \frac{2(k+1)}{n}$

$$h_{ii} = 0$$

ou

$$h_{ii} = 1$$

$$\Rightarrow h_{ij} = 0 \quad \forall i \neq j$$

$$\hat{y}_i$$

$$\forall \underline{x}_i \text{ (não é afetado por } \gamma)$$

a) $h_{ii} = 0$

$$h_{ij} = 0 \Rightarrow \hat{y}_i = 0$$

b) $h_{ii} = 1$

$$h_{ij} = 0 \Rightarrow \hat{y}_i = y_i$$

x_i ponto influente

H independe de $\tilde{y}$. Uma checada em H pode revelar pontos sensíveis, nos quais um valor discrepante em y terá um grande impacto no ajuste

Exemplo

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$X = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & & x_n \end{bmatrix}'$$

$$h_{ij} = \frac{1}{n} + \frac{[(x_i - \bar{x})(x_j - \bar{x})]}{\sum (x_i - \bar{x})^2}$$

Procure valores tais que $h_{ii} > \frac{4}{n}$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

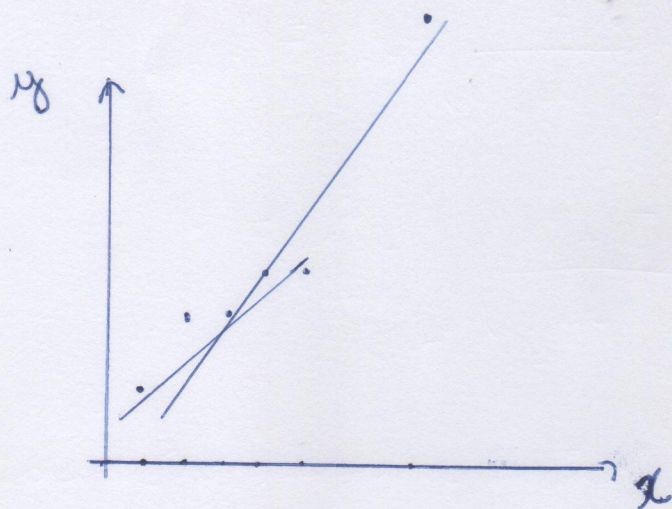
→ ponto influente: abrisse distante de $\bar{x}$

Obs:

- Para $k > 2$, é difícil visualizar graficamente pontos x_i distantes do grupo
- e a diagonal de H é uma importante fonte de informações.
- O efeito do i -ésimo caso na regressão é mais provável ser alto se h_{ii} é alto, mas sua importância é incerta dependendo dos y_j . Existem medidas que combinam h_{ii} e os y_j .

Obs: Nem sempre o resíduo associado a um ponto "discrepante" em $\hat{y}$ é alto.

x	1	2	3	4	5	8
y	2	4	4	5	5,5	12



$$r = 0,9428$$

$$\hat{\beta}_0 = 1,7$$

$$\hat{\beta}_1 = 0,8$$

$$\hat{y} = 1,7 + 0,8x$$

Aumentando o ponto (8,12)

$$r = 0,9576$$

$$SSR = 54,55$$

$$\hat{\beta}_0 = 0,33$$

$$SSE = 4,61$$

$$\hat{\beta}_1 = 1,33$$

$$\hat{\sigma}^2 = 1,15$$

$$\hat{y} = 0,33 + 1,33x$$

x_i	$\hat{y}_i$	e_i	h_{ii}	r_i
1	1,66	0,34	0,425	0,42
2	3,00	1,00	0,274	1,09
3	4,32	-0,32	0,188	-0,33
4	5,65	-0,65	0,166	-0,67
5	6,98	-1,48	0,211	-1,55
8	11,00	1,00	0,730	1,80
			1,994	
			$\approx K+1 = 2$	

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

$$h_{ii} \geq \frac{2(K+1)}{n} = \frac{4}{6} = 0,666$$

$$\bar{x} = \frac{23}{6} = 3,83$$

$$\sum (x_i - \bar{x})^2 = 30,84$$

$$r_3 = \frac{0,34}{1,07 \sqrt{1 - 0,425}}$$

2- Outros tipos de resíduos

$$e_i = y_i - \hat{y}_i \rightarrow \text{resíduo usual}$$

$$z_i = \frac{e_i}{\hat{\sigma}} \rightarrow \text{resíduo padronizado}$$

→ ir p/ verso de 3

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1-h_{ii}}} \rightarrow \text{resíduo "internamente studentizado"} \\ (\text{internally studentized Residual})$$

Validas as suposições do modelo

$$\frac{r_i}{n-k-1} \sim \text{Beta} \left(\frac{1}{2}, \frac{n-k-1}{2} \right)$$

ir p/ verso de 3, coluna r_i

$$t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{(i)} \left[1 + \frac{X_i' (X_{(i)}' X_{(i)})^{-1} X_i \right]^{1/2}} \\ = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{1-h_{ii}}} =$$

onde

$$\tilde{y}_i = X_i \hat{\beta}_{(i)} = r_i \sqrt{\frac{n-k-2}{n-k-r_i^2}}$$

$\hat{\beta}_{(i)}$ e $\hat{\sigma}_{(i)}^2$ obtidos com base na amostra de $n-1$ obs.,
excluídas a i -ésima obs.

Teste p/ outlier: Concluimos que y_i é outlier se

$$|t_i| > t_c \quad t_c \rightarrow \text{dist } t_{n-k-2}$$

h_{ii} pequeno - t_i pode ser alto pq e_i é alto, mas o "impacto" no ajuste deve ser menor

h_{ii} grande - t_i pode ser pequeno pq y_i é consistente com o modelo

→ resíduo "externamente studentizado"

pq o caso i não entra no cálculo de $\hat{\sigma}_{(i)}$

3- Medidas de Influência

Os métodos mais comuns de detectar influência eliminam um caso do conjunto de dados de cada vez.

$\hat{\beta}_{\sim(i)}$ - estimador de β computado sem o caso i

$X_{(i)}$ - matriz $(n-1) \times (k+1)$, obtida tirando-se a i -ésima linha de X

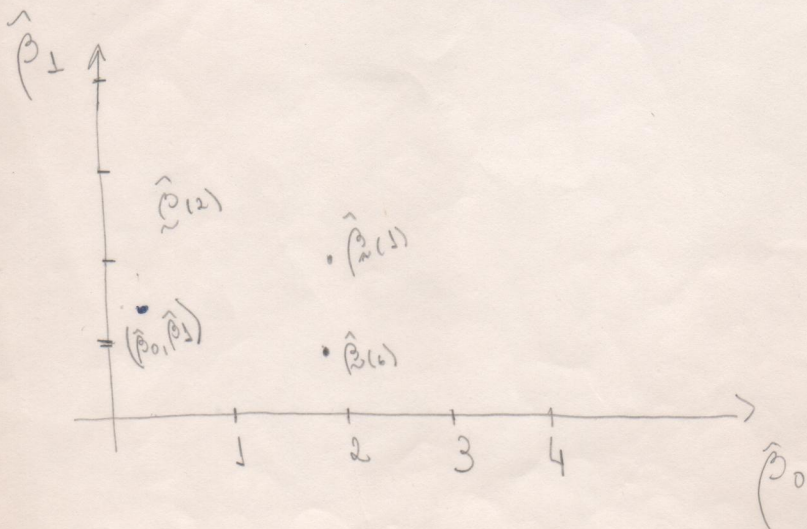
$$\hat{\beta}_{\sim(i)} = \left(X_{(i)}' X_{(i)} \right)^{-1} X_{(i)}' Y_{\sim(i)}$$

No exemplo

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 0,33 \\ 1,33 \end{bmatrix}$$

$$\hat{\beta}_{\sim(6)} = \begin{bmatrix} 1,7 \\ 0,8 \end{bmatrix}$$

Análise Gráfica



Estimativas de (β_0, β_1) eliminando um caso de cada vez

6
Como medir a distância entre $\hat{\beta}$ e $\hat{\beta}(i)$?

Distância de Cook (D de Cook)

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' (X'X) (\hat{\beta}_{(i)} - \hat{\beta})}{(k+1) \hat{\sigma}^2}$$

$$\begin{aligned} X \hat{\beta}_{(i)} &= \hat{y}_{(i)} \\ X \hat{\beta} &= \hat{y} \end{aligned} \Rightarrow D_i = \frac{(\hat{y}_{(i)} - \hat{y})' (\hat{y}_{(i)} - \hat{y})}{(k+1) \hat{\sigma}^2}$$

Sugestões:

- Analise o caso com o maior D_i .
- Analise valores de D_i próximos de 1 e maiores ou iguais a 1

Verifica-se que

$$D_i = \frac{1}{k+1} r_i^2 \left(\frac{h_{ii}}{1-h_{ii}} \right)$$

D_i é crescente em r_i^2 e em h_{ii}

r_i - reflete falta de ajuste do modelo no ponto i

h_{ii} - " a posição de x_i com relação a $\bar{x}$

D_i alto é devido à $|r_i|$ alto ou h_{ii} alto, ou ambos