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:

Formas de detectar pontos influentes.

1- Matriz "Hat"

$Y = X\beta + \varepsilon$

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

$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$

$$H = X(X'X)^{-1}X'Var(\hat{Y}) = H\sigma^2IH = H\sigma^2$$

$\hat{y}_1 = h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n$

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

h_{ii}→influência exercida por y_i em $\hat{y}_i$.

H – simétrica e idempotente ($H^2 = H$) e $\sum_{i=1}^{n} h_{ii} = k + 1$ (var. indep intercepto, ou k – modelo sem intercepto).

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

"Valor médio" de $h_{ii} = \frac{k+1}{n}$.

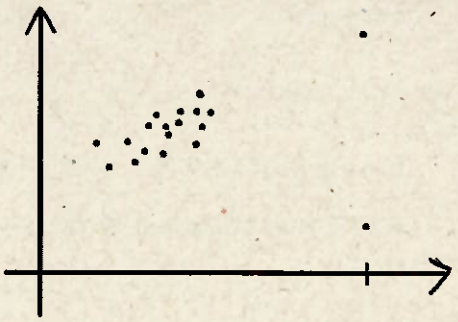
Analise os elementos da diagonal de H (h_{ii}) e dê especial atenção a pontos X_i tais que $h_{ii} > \frac{2(k+1)}{n}$.

$h_{ii} = 0$ ou $h_{ii} = 1 \rightarrow h_{ij} = 0, \forall i \neq j$

$a)\ h_{ii} = 0 \qquad h_{ij} = 0 \rightarrow \hat{y}_i = 0 \forall Y$ ($\hat{y}_i$não é afetado por Y)

$b)\ h_{ii} = 1 \qquad h_{ij} = 0 \rightarrow \hat{y}_i = y_i$

H independe de Y. Uma checada em H pode revelar pontos sensíveis, nos quais um valor discrepante em y terá um grande impacto no ajuste.



Exemplo:

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

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

$h_{i_j} = \frac{1}{n} + \frac{[(x_i - \bar{x})(x_j - \bar{x})]}{\sum(x_i - \bar{x})^2}$ Procure valores tais que $h_{i_j} > \frac{4}{n}$

$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \rightarrow$ Ponto influente: abcissa 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 infeormação.

-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 Y é alto.

Ex

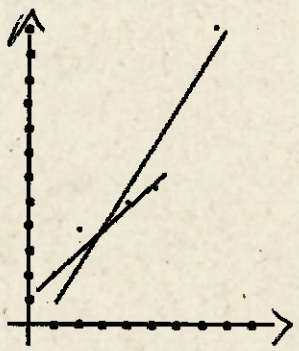
X	1	2	3	4	5	8
y	2	4	4	5	5.5	12

r = 0,9428

$\widehat{\beta_0} = 1,7$

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

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



Acrescentando o ponto (8,12)

R= 0.9576 SSR=54.55

$\widehat{\beta_0} = 0.33$ SSE = 4.61

$\hat{\beta}_1 = 1.33$ $\sigma^2 = 1.15$

$\hat{y} = 0.33 + 1.33x$

x_i	$\hat{y}_i$	e_i	h_{ii}	r_i	t_i
1	1.66	0.34	0.425	0.42	0.36
2	3.00	1.00	0.274	1.09	1.10
3	4.32	-0.32	0.188	-0.33	-0.27
4	5.65	-0.65	0.166	-0.67	-0.57
5	6.98	-1.48	0.211	-1.55	-1.92
8	11.00	1.00	0.730	1.89	3.92

Soma dos $h_{ii} = 1.994 \sim 2 = k + 1$

$$h_{ii} = \frac{1}{6} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{23}{6} = 3.83$$

$$\sum(x_i - \bar{x})^2 = 30.84$$

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

$$r_1 = \frac{0.34}{1{,}07\sqrt{1 - 0{,}425}}$$

2- Outros tipos de resíduos

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

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

→ ir para verso de 3

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

Válidas as suposições do modelo

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

$$t_i = \frac{y_i - \tilde{y}_i}{\widehat{\sigma_{(i)}}\,[1+\underline{X_i'}\,(X_{(i)}'X_{(i)})^{-1}\underline{X_{(i)}}]^{1/2}},\ com\ \tilde{y}_i = r_i\sqrt{\frac{n-k-2}{n-k-r_i^2}}$$

h_i pequeno - t_i pode ser alto porque e_i é alto, mas o "impacto" no ajuste deve ser menor.

h_i grande - t_i pode ser pequeno porque y_i é consistente com o modelo.

3- Medidas de influência

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

$\hat{\beta}_{(i)}$ – estimados de β computado sem o caso i.

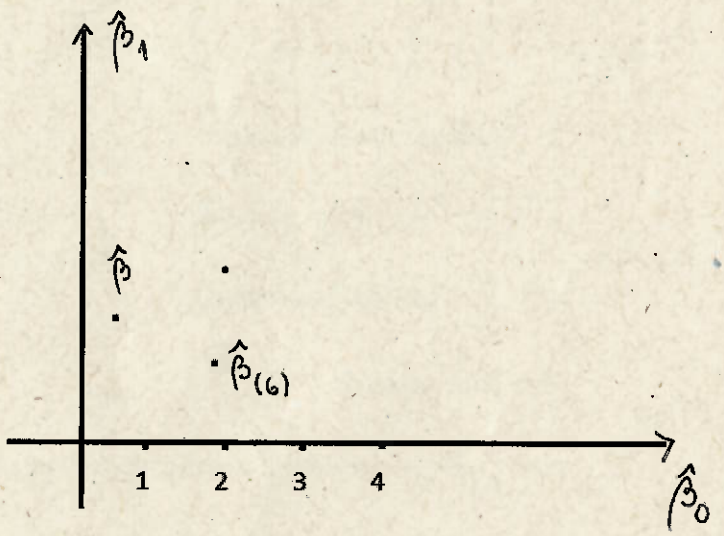
$X_{(i)}$ – matriz (n-1) x (K+1), obtida tirando-se a i-ésima linha de X.

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

No exemplo

$$\hat{\beta} = \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \begin{bmatrix}0.33\\ 1.33\end{bmatrix} \qquad \hat{\beta}_{(6)} = \begin{bmatrix}1.7\\ 0.8\end{bmatrix}$$

Análise gráfica:



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

Distância de Cook (D de Cook)

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

$$\begin{cases} X\widehat{\beta_{(i)}} = \widehat{Y_{(i)}} \\ X\hat{\beta} = \hat{Y} \end{cases} \rightarrow D_i = \frac{(\widehat{Y_{(i)}} - \hat{Y})'(\widehat{Y_{(i)}} - \hat{Y})}{(K+1)\widehat{\sigma^2}}$$

Sugestões:

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

Verifica-se que

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

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}– reflete a posição de $X_i$com relação a $\bar{X}$.

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

Não existe um ponto de corte para D. Alguns autores sugerem o quantil de ordem 0,5 da distribuição F com k e n – k – 1 gl. No entanto, isto não é completamente satisfatório porque sabe-se que D_i não tem distribuição F. Um ponto de corte comum é $D_i \geq 1$.

$$DFit = \frac{\widehat{Y_i} - \widehat{Y_{(i)}}}{\sqrt{\widehat{\sigma_{(i)}^2} h_{ii}}} = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} \frac{e_i}{\widehat{\sigma_i}(1 - h_{ii})^{1/2}} = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} t_i$$

Belsley, Kuh e Welsch [1980] sugerem como ponto de corte $|DFit| > 2\sqrt{\frac{p}{n}}$, p = k + 1 ou k.

Obs:

$\underset{\sim}{X}_i^{'}$ - i-ésima linha de X

$$X_{(i)}^{'} X_{(i)} = X'X - \underset{\sim}{X}_i \underset{\sim}{X}_i^{'}$$

$$\left(X'X - \underset{\sim}{X}_i \underset{\sim}{X}_i^{'}\right)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} \underset{\sim}{X}_i \underset{\sim}{X}_i^{'} (X'X)^{-1}}{1 - \underset{\sim}{X}_i^{'} (X'X)^{-1} \underset{\sim}{X}_i}$$

$$\Rightarrow \left[X_{(i)}^{'} X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} \underset{\sim}{X}_i \underset{\sim}{X}_i^{'} (X'X)^{-1}}{1 - h_{ii}}$$

pq $h_{ii} = X_i^{'} (X'X)^{-1} X_i$

$$\Rightarrow \underset{\sim}{\hat{\beta}} - \underset{\sim}{\hat{\beta}}_{(i)} = \frac{(X'X)^{-1} \underset{\sim}{X}_i e_i}{1 - h_{ii}}$$

Referências

Hoaglin, D.C. and Welsch, R.E. (1978). The Hat matrix in Regression and ANOVA. The American Statistician, 32, 17-32.

Belsley, D.A., E. Kuh and R.E. Welsch (1980). Regression Diagnostic: Identifying Influential Data and Sources of Collinearity, Wiley, New York

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression, New York: Chapman and Hall