

Example

Consider a dataset

y	x ₁	x ₂
17.68	7	560
12.50	3	220
9.00	2	110
80.24	30	1460
41.33	16	688

Fit a MLR model and

- Calculate residuals, standardized residuals, studentized residuals, R-studentized Residuals.
- Calculate Cook's distance and conclude
- Calculate DFBETAS, DFFITS
- Calculate AIC, BIC and Mallows's C_p values.

Solⁿ: Here $n=5$ (# data points)
 $p=3$ (# parameters β_0, β_1 and β_2)

$$y = \begin{bmatrix} 17.68 \\ 12.50 \\ 9.00 \\ 80.24 \\ 41.33 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 2 & 110 \\ 1 & 30 & 1460 \\ 1 & 16 & 688 \end{bmatrix}$$

↔ Fitting MLR Model.

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

From the data

$$X'X = \begin{bmatrix} 5 & 58 & 3038 \\ 58 & 1218 & 59608 \\ 3038 & 59608 & 2979044 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 160.75 \\ 3247.74 \\ 159226.24 \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}$$

$$\therefore \hat{\beta} = \begin{bmatrix} 3.247615 \\ 2.799459 \\ -0.005877 \end{bmatrix}$$

$\therefore$ Fitted Model eqnⁿ is

$$\hat{y}_i = \underset{\hat{\beta}_0}{3.247615} + \underset{\hat{\beta}_1}{2.799459} x_{1i} - \underset{\hat{\beta}_2}{0.005877} x_{2i}$$

(4) Calculation of various residuals. (Scaled residuals)

Observation no.	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	$\frac{e_i}{\sqrt{MS_{RES}}}$
1	19.5527	-1.8727	-0.6176
2	10.3530	2.147	0.7081
3	8.20006	0.8	0.2638
4	78.6509	1.5891	0.5241
5	43.9955	-2.6655	-0.8791

$$MS_{RES} = \hat{\sigma}^2 = \frac{SS_{RES}}{n-3} = \frac{18.3867}{2} = 9.1933$$

$\downarrow$
 since 3 parameter β_0, β_1 and β_2 are estimated

$$y'y - \sum y_i^2 = 8696.4589 \text{ and}$$

$$\beta'x'y = \begin{bmatrix} 3.247615 & 2.799459 & -0.005877 \end{bmatrix} \begin{bmatrix} 160.75 \\ 3247.74 \\ 159226.24 \end{bmatrix} = 8678.1964$$

$$\therefore SS_{RES} = y'y - \beta'x'y = 18.2625$$

$$\rightarrow SST = y'y - \frac{(\sum y_i)^2}{n} = 8696.4589 - \frac{(160.7521)^2}{5} = 3528.2113$$

$$\rightarrow SS_{RES} = \beta'x'y - \frac{(\sum y_i)^2}{n} = 8678.1964 - \frac{(160.7521)^2}{5} = 3508.0196$$

$$\left[\begin{array}{l} \text{adj } A = \\ 75361928 \quad 8304552 \quad -243020 \\ 8304552 \quad 5665776 \quad -121836 \\ -243020 \quad -121836 \quad 2726 \end{array} \right]$$

↔ Calculation of Hat Matrix ($H = X(X'X)^{-1}X'$)

$$H = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 2 & 110 \\ 1 & 30 & 1460 \\ 1 & 16 & 688 \end{bmatrix}_{5 \times 3} \begin{bmatrix} 75361928 & 8304552 & -243020 \\ 8304552 & 5665776 & -121836 \\ -243020 & -121836 & 2726 \end{bmatrix}_{3 \times 3} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 7 & 3 & 2 & 30 & 16 \\ 560 & 220 & 110 & 1460 & 688 \end{bmatrix}$$

$$H = \frac{1}{120178896} \begin{bmatrix} -2597408 & -20263176 & 430688 \\ 46811184 & -1502040 & -8808 \\ 65238832 & 6234144 & -186832 \\ -30310712 & 399272 & 81860 \\ 41037000 & 15133800 & -316908 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 7 & 3 & 2 & 30 & 16 \\ 560 & 220 & 110 & 1460 & 688 \end{bmatrix}$$

3x5

5x3

$$H = \frac{1}{120178896} \begin{bmatrix} 9674540 & & & & \\ & 40367304 & & & \\ & & 57155600 & & \\ & & & 101123048 & \\ & & & & 65145096 \end{bmatrix}$$

[It is not necessary to calculate the off-diagonal elements as it is not used in any calculations]

From the above $h_{11} = 0.8050$

$$h_{22} = 0.3358$$

$$h_{33} = 0.4755$$

$$h_{44} = 0.8414$$

$$h_{55} = 0.5420$$

Since

$$H = \begin{bmatrix} 0.8050 & & & & \\ & 0.3358 & & & \\ & & 0.4755 & & \\ & & & 0.8414 & \\ & & & & 0.5420 \end{bmatrix}$$

Using h_{ii} values we can compute studentized residuals. (r_i)

Obs. no	e_i	$r_i = \frac{e_i}{\sqrt{MS_{RES}(1-h_{ii})}}$
1	-1.8727	-1.3986
2	2.147	0.8688
3	0.8	0.3643
4	1.5891	1.3160
5	-2.6655	-1.2990

Any

Externally Studentized residuals. [R-Student residuals]

$$t_i = \frac{e_i}{\sqrt{S_{e_i}^2(1-h_{ii})}}, \quad i=1, 2, \dots, n.$$

where

$$S_{e_i}^2 = \frac{(n-p)MS_{RES} - e_i^2/(1-h_{ii})}{n-p-1}$$

here $n=5$, $p=3$, $MS_{RES} = 9.1933$

↓
considering all data in data points

e_i and h_{ii} values are same as computed previously.

For comparison

↓

Obs ⁿ no.	e_i	h_{ii}	$S_{e_i}^2$	t_i	g_i
1	-1.8727	0.8050	0.4019	-6.6894	-1.3986
2	2.147	0.3358	11.4465	0.7786	0.8688
3	0.8	0.4755	17.1663	0.2661	0.3643
4	1.5891	0.8414	2.4645	2.5417	1.3160
5	-2.6655	0.5420	2.8737	-2.3234	-1.2990

→ Identification of Leverage points ←

Any obsⁿ for which the hat diagonal exceeds twice the average [i.e., $2p/n$] is remote enough from the rest of the data is considered as leverage point.

$$\frac{2p}{n} = \frac{2(3)}{5} = \frac{6}{5} = \underline{\underline{1.2}}$$

(Since here $2p/n > 1$, the cut-off doesn't apply)

→ Cook's distance Calculation ←

Formula:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X'X (\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{RES.}}$$

or

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}, \quad i=1, 2, \dots, n.$$

Obs ⁿ no.	h_{ii}	r_i	D_i
1	0.8050	-1.3986	2.6917
2	0.3358	0.8688	0.1272
3	0.4755	0.3643	0.0401
4	0.8414	1.3160	3.0625
5	0.5420	-1.2990	0.6656

$D_i > 1$
 $\therefore$ Indicating
 pt to be an
 influential
 point.

→ DFBETAS_{j,i} Calculation ←

Formula:

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where C_{jj} is j th diagonal element of $(X'X)^{-1}$

$j = 0, 1, 2, \dots, p-1$ and $i = 1, 2, 3, \dots, n$

For $j=0$ and $i=1$

$$DFBETAS_{0,1} = \frac{\hat{\beta}_0 - \hat{\beta}_{0(1)}}{\sqrt{S_{(1)}^2 C_{00}}}$$

We know
 that

$$\hat{\beta}_0 = 3.2476$$

$$\therefore \hat{\beta}_{0(1)} = 3.04$$

$$C_{00} =$$

$$\hat{\beta}_{0(1)} = (X'X)^{-1} X'y$$

$$= \begin{bmatrix} 4 & 51 & 2478 \\ 51 & 1169 & 55688 \\ 2478 & 55688 & 2665444 \end{bmatrix}^{-1} \begin{bmatrix} 143.07 \\ 3123.98 \\ 14925.44 \end{bmatrix}$$

Deleting the 1st dataset

Obs ⁿ no.	y	x ₁	x ₂
2	12.50	3	220
3	9.00	2	110
4	80.24	30	1460
5	41.33	16	688

$$(X'X)^{-1} = \begin{bmatrix} \boxed{0.629476} & 0.087790 & 0.002419 \\ 0.087799 & \boxed{0.192943} & 0.004112 \\ -0.002419 & -0.0041127 & \boxed{0.0000885} \end{bmatrix}$$

C_{11} C_{11} C_{22}

$$DFBETAS_{0,1} = \frac{3.2476 - 3.04}{\sqrt{0.4019 (0.629476)}} = 0.4127$$

iii) we can calculate $DFBETAS_{j,i}$ & j, i

Conclusion :-

$\hookrightarrow DFBETAS_{j,i}$ is an $n \times p$ matrix

if $|DFBETAS_{j,i}| > 2/\sqrt{n}$ [here $2/\sqrt{5} = 0.8944$]
 then i^{th} point
 has large effect on j^{th} parameter.

$\rightarrow$ Calculation of DFFITS $\leftarrow$

Formula :

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

for $i=1$ $\hat{y}_1 = 19.5527$

Computing $\hat{y}_{(i)}$ after deleting the i^{th} element.

$$\hat{y}_{(1)} = 3.04 + 1.1804(3) + 0.0285(220) = 12.8512$$

$$\therefore DFFITS_1 = \frac{19.5527 - 12.8512}{\sqrt{0.4019 (0.8050)}}$$

$$DFFITS_1 = 11.7818$$

Conclusion :-

if $|DFFITS_i| > 2\sqrt{p/n}$ [here $2\sqrt{3/5} = 1.5491$]

$\therefore$ Suggests attention
 as it can be
 influential point