庫克距離 Cook's Distance

意義

庫克距離定義為刪除某一個樣本觀察值後，回歸模型變化量的總和。是常見用來衡量某一個觀察值，對回歸模型影響程度的方法。通常庫克距離越大，代表該觀察值是離群值，影響回歸線擬合的程度也越大。

公式

$$D_i = \frac{e_i^2}{(p+1) \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

e_i: 第 i 個觀察值的殘差。

p: 回歸模型中，自變數的數目。

MSE: 殘差的均方和。

h_{ii}: 第 i 個觀察值的槓桿值。

範例

下表為飲酒習慣與壽命的調查，回歸模型為 $y = 86.28 - 0.73x$ ，求各觀察值的庫克距離。

x (alcohol)	y (life)	predict y	residuals	leverage	Cook's D
23	65	69.54	-4.54	0.20664	0.030
4	92	83.37	8.63	0.68009	2.201
29	58	65.18	-7.18	0.30631	0.146
35	72	60.81	11.19	0.52559	0.299
14	68	76.09	-8.09	0.28138	0.159

上表已知e_i、P=2、h_{ii}，計算回歸模型的 MSE。

$$MSE = \frac{SSE\ (Sum\ of\ Square\ of\ Errors)}{n-k}$$

$$= \frac{(-4.54)^2 + (8.63)^2 + (-7.18)^2 + (11.19)^2 + (-8.09)^2}{5-2}$$

$$= 112.44$$

第 1 個樣本觀察值的庫克距離：

$$D_1 = \frac{(-4.54)^2}{2 \times 112.44}\left[\frac{0.20664}{(1-0.20664)^2}\right] = 0.030$$

第 2 個樣本觀察值的庫克距離：

$$D_2 = \frac{(8.63)^2}{2 \times 112.44}\left[\frac{0.68009}{(1-0.68009)^2}\right] = 2.201$$

第 3 個樣本觀察值的庫克距離：

$$D_3 = \frac{(-7.18)^2}{2 \times 112.44}\left[\frac{0.30631}{(1-0.30631)^2}\right] = 0.146$$

第 4 個樣本觀察值的庫克距離：

$$D_4 = \frac{(11.19)^2}{2 \times 112.44}\left[\frac{0.52559}{(1-0.52559)^2}\right] = 0.299$$

第 5 個樣本觀察值的庫克距離：

$$D_5 = \frac{(-8.09)^2}{2 \times 112.44}\left[\frac{0.28138}{(1-0.28138)^2}\right] = 0.159$$