TEST DE COOK (I)
(04 / 11 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le **test de COOK** est un **test** de **détection** d'une **aberration** dans un « jeu » de **données** observées (cf **observation**).

(i) On considère un **modèle de régression linéaire** multiple exprimé dans un **espace d'observation** (X, y) selon :

(1) $y = X\,b + u$, avec $b \in \mathbf{R}^K$, $E\,u = 0$, $V\,u = \sigma^2\,I_N$.

On veut estimer b par la **méthode des moindres carrés ordinaires**, sachant que l'on suspecte l'existence de valeurs atypiques parmi les données (X, y).

La détection de ces aberrations peut s'effectuer comme suit. On note :

(a) [X, y] la matrice augmentée d'ensemble, à valeurs dans $M_{N,K+1}$ (**R**) ;

(b) $[X_{(n)}\ ,\ y_{(n)}]$ la matrice, à valeurs dans $M_{N-1,K+1}$ (**R**), qui se déduit de la précédente lorsqu'on enlève la n-ième observation (ie la ligne d'indice n) $(X_n\ ,\ y_n)$ (supposée « atypique »).

On considère le modèle « restreint » suivant :

(2) $y_{(n)} = X_{(n)}\,b_{(n)} + u_{(n)}$, avec $E\,u_{(n)} = 0$ et $V\,u_{(n)} = \sigma^2\,I_N$.

On note $b_{(n)}\hat{}$ l'**estimateur des mco** de b calculé à partir de (2) et $y_{(n)}\hat{} = X\,b_{(n)}\hat{}$. L'**influence** de cette observation d'indice n peut se mesurer à l'aide de la **distance** (cf aussi **courbe d'influence**) :

(3) $\Delta_n^2 = ||y\hat{} - y_{(n)}\hat{}||^2$.

On définit alors le **test de R.D. COOK** à partir de la **statistique de test** :

(4) $D_n^2 = A / B$,

avec :

(5)
$$A = \Delta_n^2,$$
$$B = (K + 1)\,(\sigma^2)\hat{},$$

où $(\sigma^2)\hat{}$ désigne l'estimateur des mco de σ^2, calculé sur le modèle (1).

Ce test nécessite, en outre, l'admission d'une hypothèse tq :

(6) $D_n^2 \sim \mathcal{F}_{K+1,N-(K+1)}$

(**loi de FISHER-SNEDECOR** à K+1 et N-(K+1) **degré de liberté**).

(ii) Le test de COOK n'est que « partiel », ou « marginal » (une seule observation aberrante). Il ne traite donc pas l'existence d'une hétérogénéité d'échantillonnage plus importante (cf **mélange légal**).