RÉGRESSION AVEC OBSERVATIONS MANQUANTES (J1, J5)

Il peut arriver que l'**inférence statistique** doive se réaliser alors qu'il existe des lacunes dans l'ensemble des observations utilisées (cf **censure**, **lacune**, **observation manquante**). Cette **situation statistique** (résultant d'un « *effet de masque* » ou « *effet d'occultation* ») nécessite une adaptation des **procédures statistiques** habituelles, notamment les procédures d'estimation des paramètres.

Dans le cas, important en pratique, de la **régression**, la loi de probabilité de l'ensemble des variables qui interviennent (loi multivariée) génère des « observations » de ces variables, dont certaines sont réellement observées, d'autres non.

On considère une **structure statistique** $(\Omega, \mathcal{T}, \mathcal{P})$, un **espace d'observation** $(\mathcal{Y}_0, \mathcal{C}_0)$, une **va** endogène $\eta : \Omega \mapsto \mathcal{Y}_0$ et le **modèle d'échantillonnage** $(\mathcal{Y}_0{}^N, \mathcal{C}_0{}^{\otimes N}, P^y)$ qui en résulte, où $y : \Omega \mapsto \mathcal{Y}_0{}^N$ est le N-uple constitué des observations de η.

On note le **modèle de régression multiple** standard usuel selon :

(1) $E\, y = X\, b$, avec $V\, y = \sigma^2 . I_N$,

où l'on suppose que $\mathcal{Y}_0 = \mathbf{R}$, $\mathcal{X} = M_{NK}(\mathbf{R})$ et $b \in \mathbf{R}^K$.

On dit que ce modèle représente une **régression avec observations manquantes**, ou **régression en présence d'observations manquantes**, ssi il existe une **partie** (en général connue) S_y (resp S_k) de $\mathbf{R}^N$ tq y_n (resp x_k) est observée ssi $y \in S_y$ (resp ssi $x_k \in S_k$), non observée sinon (où $k \in N_K{}^*$).

Dans le cas le plus simple, la matrice $X = [x_1 ,..., x_K]$ (vecteurs colonnes) est entièrement observée. On peut (à une permutation sur les observations (X_n , y_n) près) écrire :

(2) $y = X\, b + u$,

où l'on note $y = (y^1 /// y^2)'$ et $X = (X^1 /// X^2)'$, ie :

$$y = \begin{pmatrix} y^1 \\ y^2 \end{pmatrix}, \quad X = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix}$$

où y^1 est un vecteur aléatoire observé (observations présentes) et y^2 un vecteur aléatoire non observé (observations manquantes, ou lacunes).

La **méthode des moindres carrés ordinaires** (mco) appliquée au modèle (2) consiste à résoudre le programme mathématique suivant (cf **programmation mathématique**) :

(3) $\min \|y - X\, b\|$, sous $b \in \mathbf{R}^K$ et $y^2 \in \mathbf{R}^{N(2)}$,

où l'on suppose que $y_1 \in \mathbf{R}^{N(1)}$, $y_2 \in \mathbf{R}^{N(2)}$ et $N_1 + N_2 = N$ (en notant, par commodité, N(i) pour désigner N_i). Dans ce programme, y_2 est considéré comme un paramètre supplémentaire (**paramètre importun**) à estimer.

(i) la **méthode de K.D. TOCHER** comporte trois étapes :

(a) estimation de b par la **méthode des moindres carrés ordinaires** sur le modèle avec mise à zéro de y_2 :

(4) $(y^1 /// 0)' = X\, b + u$,

ie :

$$\begin{pmatrix} y^1 \\ 0 \end{pmatrix} = X\ b + u$$

d'où l'estimateur $b_0{}^\wedge = (X'\, X)^{-1} (X^1)'\, y^1$;

(b) calcul de l'« estimateur » (ou prévision) suivant(e) de y^2 :

(5) $(y^2{}_0)^\wedge = \{I - X^2 (X'\, X)^{-1} (X^2)'\}^{-1} X^2\, b_0{}^\wedge$;

(c) estimation de b par la **méthode des mco** sur le modèle (après estimation des lacunes y^2) :

(6) $(y^1 /// (y^2{}_0)^\wedge)' = X\, b + u$,

ie :

$$\begin{pmatrix} y^1 \\ \widehat{y_0^2} \end{pmatrix} = X\ b + u$$

ce qui fournit un pseudo-**estimateur des mco** $b_\sim{}^\wedge$ de b.

Le nombre de **degrés de liberté** de la **forme quadratique** $\|y - X\, b_\sim{}^\wedge\|^2$ ainsi obtenue est donc $(N - N^2) - K$. C'est ce nombre qui est à prendre en compte (eg pour effectuer des tests).

(ii) Dans le cas plus général où X est partiellement observé, on peut (à une **permutation** près sur les observations) écrire (1) sous la forme :

(7) $y = X\, b + u$, avec $y = (y^1 /// y^2 /// y^3 /// y^4)'$ et $X = (X^1 /// X^2 /// X^3 /// X^4)'$,

ie :

$$y = \begin{pmatrix} y^1 \\ y^2 \\ y^3 \\ y^4 \end{pmatrix}, \quad X = \begin{pmatrix} X^1 \\ X^2 \\ X^3 \\ X^4 \end{pmatrix}$$

Sous cette forme, y^1 et y^3 sont observées et y^2 et y^4 sont manquantes, tandis que (alternativement) X^1 et X^2 sont observées et X^3 et X^4 sont manquantes.

La **méthode de A.A. AFIFI - R.M. ELASHOFF** permet d'estimer (7) en trois étapes :

(a) estimation de b par la méthode des mco sur le modèle :

(8) $(y^1 /// 0 /// y^3 /// 0)' = (X^1 /// X^2 /// 0 /// 0)'\, b + u$,

ie :

$$\begin{pmatrix} y^1 \\ 0 \\ y^3 \\ 0 \end{pmatrix} = \begin{pmatrix} X^1 \\ X^2 \\ 0 \\ 0 \end{pmatrix} b + u$$

d'où un estimateur, encore noté $b_0^{\wedge}$, de b (après mise à zéro de y^2 et y^4) ;

(b) calcul de l' « estimateur » $(y_0^2)^{\wedge}$ de y^2 à l'aide de la formule (5) précédente, dans laquelle X est remplacée par la matrice du second membre de (8) (ie après mise à zéro de X^3 et X^4) ;

(c) estimation de b par la méthode des mco sur le modèle :

(9) $(y^1 /// (y_0^2)^{\wedge} /// 0 /// y^3 /// 0)' = (X^1 /// X^2 /// 0 /// 0)'\, b + u$,

ie :

$$\begin{pmatrix} y^1 \\ \widehat{y_0^2} \\ y^3 \\ 0 \end{pmatrix} = \begin{pmatrix} X^1 \\ X^2 \\ 0 \\ 0 \end{pmatrix} b + u$$

(iii) Dans les modèles précédents, d'autres méthodes d'estimation de b généralement utilisées sont fondées sur la maximisation de la **vraisemblance**, les observations manquantes figurant dans celle-ci jouant le rôle de **paramètres incidents** ou de **paramètres fantômes** (cf **paramètre importun**).

La **méthode du maximum de vraisemblance** ainsi adaptée doit, pour l'étude de ses propriétés asymptotiques, être assorties d'hypothèses tq la relation entre le nombre d'observations manquantes et le nombre total d'observations.