RÉGRESSIONS MULTIPLES (J9)
(09 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'étude simultanée de plusieurs **modèles de régression** **apparemment sans relation** (A. ZELLNER) est parfois préférable à leur étude séparée : on peut alors prendre en compte des « **simultanéités** », ie des **corrélations** éventuelles entre les **perturbations aléatoires** de ces modèles.

(i) Soit, $\forall\, g \in \mathbf{N}_G^*$:

(1) $\quad y\,(g) \;=\; X\,(g)\, b\,(g) + u\,(g), \qquad$ avec $E\, u\,(g) = 0,\;\; V\, u\,(g) = s^2_g\, I_N$,

un **modèle de régression linéaire** multiple, exprimé dans l'**espace des observations**, et dans lequel u (g) et y (g) sont des **va** à valeurs dans $\mathbf{R}^N$ (même ensemble de valeurs), $X\,(g) \in M_{NK(g)}\,(\mathbf{R})$ (avec $K_g \geq 1$) et $b\,(g) \in \mathbf{R}^{K(g)}$ est un paramètre d'intérêt (où K(g) désigne K_g).

La n-ième équation (observation) de l'équation vectorielle (1) s'écrit donc sous forme bi-indicée (modèle à deux **indices**, n et g), ie :

(2) $\quad y_n\,(g) \;=\; X_n\,(g)\, b\,(g) + u_n\,(g),$

où $X_n\,(g) = [x_{n1}\,(g)\, ,\ldots,\, x_{nK(g)}\;\, (g)]$ est la n-ième ligne de X (g).

La prise en compte de corrélations éventuelles entre les u(g) (donc entre les y(g)) peut s'effectuer en complétant la **spécification** (1) ou (2) à l'aide d'une hypothèse tq :

(3) $\quad C\,(u_\alpha\,(g),\, u_\beta\,(h)) \;=\; \delta_{\alpha\beta}\, \sigma_{gh}\;, \qquad \forall\,(g, h)$ et $\forall\,(\alpha, \beta)$.

où $\delta_{\alpha\beta}$ désigne le **symbole de KRONECKER**. Autrement dit, les perturbations relatives à une même obserbation $\beta = \alpha = n$ sont seules corrélées.

(ii) L'écriture d'ensemble du modèle précédent s'obtient par « empilement » des équations (1), d'où la forme suivante, appelée **modèle de régressions multiples** (termes exprimés au pluriel), ou **modèle à équations sans relation apparente** :

(4) $\quad y \;=\; X\, b + u, \qquad$ avec $E\, u = 0,\;\; V\, u = \Sigma \otimes I_N$,

où $\Sigma = (\sigma_{gh})_{(g,h)} \in S_G\,(\mathbf{R})$ (**matrice symétrique**, supposée être une **matrice définie positive**), $A \otimes B$ désigne le produit tensoriel des matrices A et B (cf **produit tensoriel algébrique**) et :

(5)
$\quad y \;=\; (y\,(1)\; /\!/\!/\; y\,(G))', \qquad X = (X\,(1)\; /\!/\!/\; X\,(G))' \qquad$ (**observables**),
$\quad b = (b\,(1)\; /\!/\!/\; b\,(G))', \qquad (u\,(1)\; /\!/\!/\; u\,(G))' \qquad$ (**inobservables**),

où /// représente des sauts de lignes ou de colonnes, ie :

$$y = \begin{pmatrix} y(1) \\ \vdots \\ y(g) \\ \vdots \\ y(G) \end{pmatrix}, \quad X = \begin{pmatrix} X(1) & O & \cdots & O \\ \vdots & \ddots & & \vdots \\ & & \ddots & \\ O & O & \cdots & X(G) \end{pmatrix}, \quad b = \begin{pmatrix} b(1) \\ \vdots \\ b(g) \\ \vdots \\ b(G) \end{pmatrix}, \quad u = \begin{pmatrix} u(1) \\ \vdots \\ u(g) \\ \vdots \\ u(G) \end{pmatrix}$$

(iii) La forme (4) n'est ainsi autre que celle d'un modèle de régression linéaire multiple, assorti d'une hypothèse de **dispersion** non scalaire.

Un tel modèle peut donc s'analyser (estimation, tests) à l'aide de méthodes usuelles : **méthode de moindres carrés**, **méthode du maximum de vraisemblance**, etc. Le premier type de méthodes permet de calculer :

(a) soit l'**estimateur des moindres carrés ordinaires** $b^\wedge = (X' X)^{-1} X' y$ de b, avec $V b^\wedge = (X' X)^{-1} X' (\Sigma \otimes I_N) X (X' X)^{-1}$;

(b) soit l'**estimateur des moindres carrés généralisés** $b_g^\wedge = \{X' (\Sigma^{-1} \otimes I_N) X\}^{-1} X' (\Sigma^{-1} \otimes I_N) y$ de b, avec $V b_g^\wedge = \{X' (\Sigma^{-1} \otimes I_N) X\}^{-1}$.

On montre que :

(a) si Σ est une **matrice diagonale**, alors $b_g^\wedge = b^\wedge$;

(b) si $X (g) = X (0)$, $\forall g \in N_G^*$ (matrices identiques entre elles), on a encore $b_g^\wedge = b^\wedge$.

En général, cependant, Σ n'est pas connue et il est nécessaire de l'estimer (cf **méthode des moindres carrés généralisés**).

(iv) Le modèle de régressions multiples précédent {(4),(5)} contient un cas particulier important, celui du **modèle de régression multidimensionnel**, dans lequel $X (g) = X (0)$, $\forall g \in N_G^*$ (cf (iii) (b) ci-dessus).

(v) Certains résultats peuvent s'étendre :

(a) lorsque le nombre d'observations N peut différer d'une équation g décrite en (1) à une autre h, soit N_g , $\forall g \in N_G^*$ (même formalisme) ;

(b) en supposant les équations (1) non linéaires (pr aux paramètres b(g)), ie :

(6) $\quad y (g) = F_g (b (g)) + u (g), \qquad$ avec $E\, u (g) = 0$, $V u (g) = \sigma_g^2 . I_N$.

(même démarche que pour une **régression non linéaire**).