MODÈLE MULTI-INDICÉ (G2, G11, H7, K1, K15)
(23 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) De façon générale, un **modèle multi-indicé** est une **représentation statistique** dans laquelle les **variables** considérées sont observées selon plusieurs **critères** d'**observation**, eg :

(a) « **unité statistique** x **période** d'**observation** » : eg modèle de « panel » ;

(b) « **unité statistique** x **espace** » : eg modèle « spatial » ;

(c) « **espace** x **temps** » : eg processus spatio-temporel.

Ainsi, dans un **modèle d'analyse de la variance**, divers facteurs peuvent influer sur une grandeur numérique donnée : les données sont alors disposées selon une « dimension » **facteur** et une dimension **observation**.

(ii) Un modèle multi-indicé se distingue d'un modèle comportant des variables « multiples », eg :

(a) un **modèle multivarié** (cf **loi multivariée**), dans lequel la liste des **variables d'intérêt** est constituée de variables de différents types : **variables qualitatives**, **variables numériques** ;

(b) un **modèle multidimensionnel** (cf **analyse multidimensionnelle**, **loi multidimensionnelle**), moins général mais plus courant que le précédent, dans lequel les variables d'intérêt sont des **vecteurs aléatoires** (très souvent réels), ie prennent leurs valeurs dans des ensembles à plusieurs dimensions (au sens algébrique) (cf **espace vectoriel**).

(iii) Chaque « observation » d'un **modèle multi-indicé** est un **tableau statistique** multidimensionnel, de la forme générale à H « dimensions » :

(0) $T = (t_I)_{I \in \mathcal{T}}$, avec $t_I \in \mathbf{R}$, $\forall\, I = (i_1, ..., i_H) \in \mathcal{T}$,

où $\mathcal{T} = \Pi_{h=1}^{H} \mathcal{T}_h$ et $\mathcal{T}_h = \{1, ..., n_h\}$ (multi-**indice**).

On note T_H (**R**) l'ensemble des tableaux statistiques réels H-dimensionnels.

(iv) A titre d'exemple, soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** fondamental et :

(1) $\eta, \xi_1, ..., \xi_K : \Omega \mapsto \mathbf{R}$

une suite (ou « liste ») consituée de K **vars** ξ_k .

Un **modèle de régression multiple** linéaire tq (dans l'**espace des variables**) :

(2) $\eta = \zeta + \varepsilon = \Sigma_{k=1}^{K} b_k \xi_k + \varepsilon$, avec $E\, \varepsilon = 0$,

est alors « observé » (dans l'**espace des observations**) sous la forme :

(3) $y_I = z_I + u_I = \sum_{k=1}^{K} b_k x_{k,I} + u_I$,

avec E $u_I = 0$, $\forall I \in \mathcal{T}$, dans laquelle on peut noter $Y = (y_I)_{I \in \mathcal{T}} : \Omega \mapsto T_H$ (**R**), $X_k = (x_{k,I})_{I \in \mathcal{T}} : \Omega \mapsto T_H$ (**R**), $\forall k \in N_K^*$, et $u = (u_I)_{I \in \mathcal{T}} : \Omega \mapsto T_H$ (**R**) sont 1 + K + 1 tableaux aléatoires. Un tel modèle est ainsi multi-indicé par $I = (i_1, ..., i_H) \in \mathcal{T}$.

(v) De même, un **modèle de régression multiple** (non linéaire) tq :

(4) $\eta = f(\xi_1, ..., \xi_K) + \varepsilon$, avec E $\varepsilon = 0$,

est « observé » sous la forme :

(5) $y_I = f(x_{1,I}, ..., x_{K,I}, b) + u_I$, avec E $u_I = 0$, $\forall I \in \mathcal{T}$,

dans laquelle $b \in \mathbf{R}^Q$.

Sous l'hypothèse de non **corrélation** et d'**homoscédasticité** suivante :

(6) $C(u_I, u_J) = \delta_{IJ} . \sigma_u^2$, $\forall (I, J) \in \mathcal{T}^2$,

on peut estimer le **paramètre** b par la **méthode des moindres carrés ordinaires**, qui équivaut à la **méthode du maximum de vraisemblance** dans le cas gaussien, où $u_I \sim \mathcal{N}_1(0, \sigma_u^2)$ (**loi normale multidimensionnelle** centrée), $\forall I \in \mathcal{T}$.

Pour celà, on « empile » les observations de chaque variable $\eta, \xi_1, ..., \xi_k$, eg selon l'**ordre lexicographique**. Ainsi, dans le cas de η :

(7) $y = (y_{11...11} /// y_{11...n(H)} /// y_{11...21} /// y_{11...n(H-1)n(H)} /// y_{1n(2)...n(H-1)n(H)} /// y_{n(1)n(2)...n(H-1)n(H)})$,

où /// dénote un changement de ligne, ce qui se représente dans le schéma ci-après :

$$y = \begin{pmatrix} y_{1\,1\,\cdots\,1\,1} \\ \cdots \\ y_{1\,1\,\cdots\,n_H} \\ y_{1\,1\,\cdots\,2\,1} \\ \cdots \\ y_{1\,1\,\cdots\,n_{H-1}\,n_H} \\ \cdots \\ y_{1\,n_2\,\cdots\,n_{H-1}\,n_H} \\ \cdots \\ y_{n_1 n_2\,\cdots\,n_{H-1} n_H} \end{pmatrix}$$

Ce **vecteur aléatoire** est à valeurs dans $\mathbf{R}^N$, avec $N = \Pi_{h=1}^{H} n_h$, ce qui permet de ramener (3) à la forme standard :

(8) $y = X\,b + u, \qquad$ avec $E\,u = 0,\ V\,u = \sigma_u^2 . I_N$,

et (5) à la forme classique :

(9) $y = F(b) + u, \qquad$ avec $E\,u = 0,\ V\,u = \sigma_u^2 . I_N$,

où $X \in M_{NK}(\mathbf{R})$ et $F : \mathbf{R}^Q \mapsto \mathbf{R}^N$ est une fonction vectorielle (qui dépend, en général, de X).

(vi) La « **partie certaine** » z_l de certains modèles multi-indicés de forme additive :

(10) $y_l = z_l + u_i$, $\qquad$ avec $E\,u_l = 0, \qquad \forall\, l \in \mathcal{I}$,

est souvent décomposée de façon adaptée.

Ceci est le cas des modèles (8) et (9), ainsi que :

(a) du **modèle avec constantes spécifiques**, dans lequel :

(11) $z_l = b_{1,i(1)} + \ldots + b_{H,i(H)} = \Sigma_{h=1}^{H} b_{h,i(h)}$,

fréquemment utilisé en **analyse de la variance** ;

(b) du modèle plus général :

(12) $z_l = \Sigma_{k=1}^{K} b_k x_{k,l} + \Sigma_{h=1}^{H} b_{h,i(h)}$,

fréquemment utilisé en **analyse de la covariance**, etc.

(vii) Dans le cas d'un **modèle bi-indicé** (H = 2), un **modèle d'analyse de la variance** multidimensionnel peut s'écrire sous la forme :

(13) $Y_I\ (|I|, |T|) = X_I\ (|I|, |K|)\ b\ (|K|, |L|) + X_T\ (|L|, |T|) + U\ (|I|, |T|)$,

dans laquelle $Y = (y_{it})_{(i,t)}$ est tq l'observation y_{it} de la variable endogène η se décompose en fonction des observations $x_{I,i(k)}$ et $x_{T,It}$, où $X_I = (x_{I,ik})_{(i,k)}$ désigne la matrice des **effets inter-individuels**, $X_T = (x_{T,It})_{(I,t)}$ la matrice des **effets intra-individuels**, $i \in I$ un individu (**unité statistique**) et $t \in T$ le **temps** (en notant $|I|$ = Card I, $|T|$ = Card T, etc).

(vii) Certains modèles multi-indicés (10) comportent une partie aléatoire u_I décomposée de façon adaptée. Il est en effet souvent plus réaliste de supposer que le vecteur aléatoire empilé u (ou y) possède une **matrice de dispersion** non scalaire (ie $V\ u \neq \sigma_u^2\ .\ I_N$).

Ceci est notamment le cas du **modèle à erreurs composées**, pour lequel :

(14) $u_I = u_{i(1)} + \ldots u_{i(H)} + v_I\ ,\qquad \forall\ I \in \mathcal{I}$,

avec :

$$E\ u_{i(h)} = 0, \qquad \forall\ i_h \in \mathcal{I}_h\ ,\quad \forall\ h \in N_H^*,$$

(15) $$V\ u_{i(h)} = \sigma_h^2, \qquad \forall\ i_h \in \mathcal{I}_h\ ,\quad \forall\ h \in N_H^*,$$

$$C\ (u_{g,i(g)}\ ,\ u_{h,i(h)}) = 0, \qquad \forall\ (i_g\ ,\ i_h) \in \mathcal{I}_g \times \mathcal{I}_h\ ,\ \forall\ (g, h) \in N_H^* \times N_H^*\ (g \neq h).$$

(viii) Diverses méthodes (eg **méthode du mv** ou **méthode des mcg**) permettent de traiter ces modèles.

Les **paramètres d'intérêt** b des différentes formes précédentes (ou de leurs variantes) ne sont pas toujours estimables (donc les modèles ne sont pas toujours identifiables) (cf **estimabilité**, **identifiabilité**) et nécessitent l'introduction de **contraintes** de même nature que celles des modèles d'analyse de la variance ou de la covariance (cf aussi **singularité**).

(ix) Le « temps » est souvent l'un des « critères » selon lesquels on observe les variables η, ξ_1 ,..., ξ_K . Les hypothèses (tq (18)) sont alors modifiées pour tenir compte des caractéristiques propres au temps (**relation d'ordre** total, **autocorrélation** temporelle, etc).

De même, les modèles précédents sont souvent « observés » sur des **séries temporelles** de **coupes instantanées** : eg unités statistiques « suivies » au cours du temps (données de panels).

(x) Il existe enfin des modèles « mixtes », dans lesquels, à la fois, les parties certaine z_I et aléatoire u_I sont décomposées comme ci-dessus.