MODÈLE LINÉAIRE (D2, J1)
(09 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) De façon générale, on appelle **modèle linéaire** tout **modèle statistique** dans lequel les **variables** (observées ou non) et les **paramètres** (inconnus) interviennent de façon linéaire ou affine (cf **espace affine**, **application affine**) (cf aussi **problème linéaire**).

Un tel modèle peut aussi se déduire d'un **modèle non linéaire** après **linéarisation**.

(ii) On adopte parfois une notion plus large (et plus souple pour les applications) : un **modèle linéaire** est alors un modèle statistique dans lequel seuls les paramètres (inconnus) interviennent de façon linéaire, les variables (observées ou non) pouvant intervenir de façon non linéaire.

Dans certains cas, le modèle peut être linéaire pr à certaines fonctions ψ des paramètres d'intérêt. Ainsi, le modèle représenté par l'équation :

(0) $\eta = \exp(c_1 \xi_1) + c_2^2 \xi_2 + \varepsilon = \psi_1(c_1, c_2) . \exp(\xi_1) + \psi_2(c_1, c_2) . \xi_2 + \varepsilon$

est linéaire pr aux fonctions $\psi_1(c_1, c_2) = \exp(c_1)$ et $\psi_2(c_1, c_2) = c_2^2$.

(iii) Le modèle linéaire est donc un **modèle « versatile »** car il peut s'adapter à de nombreux contextes, qui en constituent des cas particuliers ou des extensions (eg **variables binaires** ou **variables indicatrices**, **variables qualitatives** après divers **codages**).

Ainsi, dans le cas d'une seule variable endogène, le modèle linéaire peut prendre l'une des **formes unidimensionnelles** suivantes :

(a) le **modèle de régression multiple** linéaire (standard), dans lequel l'unique **variable endogène** η s'écrit sous la forme :

(1) $\eta = \xi' b + \varepsilon, \qquad \forall\, b \in \mathbf{R}^K,$

où $(\xi, b) \mapsto \xi' b$ est une forme bilinéaire (ie linéaire pr à chaque argument) (cf **forme multilinéaire**). On obtient un modèle du même type avec une équation tq :

(2) $\eta = \varphi(\xi)' c + \varepsilon, \qquad \forall\, c \in \mathbf{R}^L,$

où un **changement de variable** transforme les K variables exogènes $(\xi_1, ..., \xi_K)$ en L nouvelles « variables exogènes » $\varphi_\lambda = \varphi_\lambda(\xi)$, $\forall\, \lambda \in N_L^*$, à l'aide d'une fonction (non nécessairement linéaire) $\varphi : \mathbf{R}^K \mapsto \mathbf{R}^L$;

De même, le modèle :

(3) $\eta = \varphi(\xi)' \psi(c) + \varepsilon, \qquad \forall\, c \in C,$

peut parfois être considéré comme linéaire si le **paramètre d'intérêt** est $d = \psi(c) \in \mathbf{R}^M$, avec $d_m = \psi_m(c)$, $\forall\, m \in N_M^*$. Ici, l'ensemble C est a priori quelconque ;

(b) le **modèle d'analyse de la variance** (linéaire) ;

(c) le **modèle d'analyse de la covariance** (linéaire) ;

(d) le **modèle à variance composée** (linéaire) ;

(e) le **modèle linéaire généralisé** ;

(f) le **modèle d'interdépendance** (linéaire).

(iv) Les **versions muldimensionnelles** des modèles précédents sont aussi des modèles linéaires. Ainsi en est-il :

(a) du modèle des **régressions multiples** (linéaire), dont les équations peuvent s'écrire :

(4) $\eta_g = \xi' b_g + \varepsilon_g , \quad \forall\, g \in N_G^*,$

ou encore :

(5) $\eta_g = \varphi(\xi)' c_g + \varepsilon_g , \quad \forall\, g \in N_G^* ;$

(b) du **modèle d'interdépendance linéaire**, qui s'écrit :

(6) $B\,\eta + C\,\xi = \varepsilon, \quad \forall\, (B, C) \in M_G(\mathbf{R}) \times M_{NK}(\mathbf{R}),$

ou encore :

(7) $D\,\kappa(\eta) + E\,\varphi(\xi) = \varepsilon, \quad \forall\, (D, E) \in M_H(\mathbf{R}) \times M_{NL}(\mathbf{R}),$

où $\kappa : \mathbf{R}^G \mapsto \mathbf{R}^H$ est une fonction a priori quelconque.

(iii) Les propriétés de base des modèles linéaire sont fondées sur les opérations linéaires qu'ils permettent de mettre en œuvre (algèbre linéaire) (cf **opérateur linéaire**).

Ainsi, si le modèle (1) est « observé » sous la forme :

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

alors (8) équivaut à :

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

ou à :

(10) $\beta\, y = X\,(\beta\, b) + \beta\, u,$ avec $E\,(\beta\, u) = 0,\ V\,(\beta\, u) = \beta^2\, \sigma^2\, I_N , \quad \forall\, b \in \mathbf{R}_+^*,$

ou encore à :

(11) $Q\,y = X\,b + Q\,u$, avec $E(Q\,u) = 0$, $V(Q\,u) = \sigma^2 . I_N$,

quelle que soit la **matrice orthonormale** Q (cf **matrice orthogonale**) tq $Q\,X = X$ et $Q\,Q' = I_N$.

Les **estimateurs** usuels (**estimateur des mco**, **estimateur du mv**) $T_N = t_N(X, y)$ de b et $S_N = s_N(X, y)$ de σ^2 vérifient resp, $\forall$ y (cf **équivariance**) :

(12)
$$t_N(X, y + X\,\gamma) = \psi_N(y + X\,\gamma) = \psi_N(y) + \gamma, \qquad \forall\ \gamma \in \mathbf{R}^K,$$
$$t_N(X, \beta\,y) = \psi_N(\beta\,y) = \beta\,\psi_N(y), \qquad \forall\ \beta \in \mathbf{R}_+^*,$$

et :

(13) $s_N(X, Q\,y) = \varphi_N(Q\,y) = \varphi_N(y)$,

$\forall$ Q du type précédent, où ψ_N désigne la deuxième application partielle de t_N et φ_N celle de s_N .

(iv) Si, dans un **modèle de régression linéaire** standard :

(14) $y = X\,b + u$, avec $X \in M_{NK}(\mathbf{R})$ et $N \geq K$,

la matrice X vérifie $X'\,X = I_K$, elle définit une **variété de RIEMANN** particulière, appelée **variété de E. L. STIEFFEL**, notée V_{KN} .

On peut alors définir des **lp** (matricielles) sur cet ensemble : **loi de FISHER-LANGEVIN-MISES**, **loi de BINGHAM-FISHER**, etc.

Si X est de la forme $X = Z\,(Z'\,Z)^{1/2}$, alors $X \in V_{KN}$: on l'appelle parfois « **orientation** » de Z.