RÉGRESSION SUR COMPOSANTES PRINCIPALES (J1, K5)
(12 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'estimation d'un **modèle de régression** dont les **variables exogènes** sont colinéaires (resp quasi-colinéaires) peut se réaliser à l'aide des **composantes principales** déduites de la **matrice d'observation** de ces variables (cf **colinéarité**, **quasi-colinéarité**, **estimateur de HOERL-KENNARD** d'une régression « raccourcie »).

Cette méthode (H. HOTELLING - K.G. KENDALL) peut notamment être utilisée préalablement à une régression usuelle utilisant ces variables, puisque cette dernière conduit à un **estimateur** soit non défini, soit à forte **dispersion**. Elle permet donc une **réduction de dimension** de l'espace des **paramètres**, sans exclure a priori de l'analyse certaines des variables exogènes considérées.

(i) On considère un **modèle de régression linéaire** multiple, noté (dans un **espace de variables**) $\eta = \xi' b + \varepsilon$ (avec $E\,\varepsilon = 0$ et $V\,\varepsilon = \sigma^2$), « observé » (dans un **espace d'observation**) selon $y = X b + u$ (avec $E\,u = 0$ et $V\,u = \sigma^2 . I_N$).

On suppose que X et y représentent des variables resp centrées pr à leurs **moyennes empiriques** : ie on observe η selon le vecteur aléatoire h et ξ selon la matrice $T \in M_{NK}$ (**R**), et l'on transforme ensuite ces observations selon $y = P h$ et $X = P T$ (où P désigne la **matrice de centrage par rapport à la moyenne**).

Le modèle précédent est appelé **régression sur composantes principales** lorsque la (N,K)-matrice X est remplacée par la (N,L)-matrice Z constituée des $L < K$ premiers vecteurs propres (ie des composantes principales centrées) z_l de X, avec :

(1) $\quad z_l = \lambda_l^{-1/2} X w_l , \qquad \forall\, l \in N_L^*,$

où w_l est le **vecteur propre** de la **matrice** $X' X \in M_K$ (**R**) associé à la l-ième plus grande **valeur propre** de X' X et où z_l désigne la l-ième colonne de Z (cf **décomposition spectrale d'une matrice**).

(ii) La méthode consiste donc à remplacer l'équation :

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

dans laquelle X est une **matrice** dont les vecteurs colonnes sont colinéaires (ou quasi-colinéaires), par l'équation :

(3) $\quad y = Z c + v,$

puis à estimer le nouveau paramètre $c \in \mathbf{R}^L$ par la **méthode des moindres carrés ordinaires** (mco).

(iii) On montre alors que :

(a) l'**estimateur des moindres carrés ordinaires** de c s'écrit $c_L\hat{} = (Z' Z)^{-1} Z' y = Z' y$, les composantes principales ayant été supposées normées ($Z' Z = I_L$). Comme

le « vrai » modèle est (2), l'**estimateur sur composantes principales**, ou **estimateur spectral**, $c_L\hat{}$ est biaisé (cf **biais**, **écart quadratique moyen**) ;

(b) si $V\ v = \sigma_v^2\ I_N$ (hypothèse faite pour estimer (3) par la méthode des mco), alors $V\ c_L\hat{} = V\ v$;

(c) la **variance** σ_v^2 est généralement estimée selon la formule usuelle des mco, soit :

(4) $(\sigma_v^2)\hat{} = (N - L)^{-1} . ||v\hat{}||^2 = (N - L)^{-1} . (y\hat{})' (I_L - Z\ Z') y\hat{}$.

Cette **statistique** n'est autre que l'estimateur de la variance de $c_l\hat{}$: ie $(V\ c_l\hat{})\hat{} = (\sigma_v^2)\hat{}$, $\forall\ l \in N_L^*$.

(iv) Utilisée en cas de **colinéarité** stricte ($rg\ X'\ X < K$) ou approchée ($rg\ X'\ X \# K$ ou $0 < Det\ (X'\ X) << +\infty$), la méthode précédente est aussi utilisée pour réduire le nombre de variables « explicatives » ξ_k , surtout si celles-ci sont initialement « trop » nombreuses (ie eg $K > N$) (cf aussi **principe de parcimonie**, **degré de liberté**), et que le nombre N des observations ne peut être augmenté.

Dans ces situations, la méthode revient implicitement à sélectionner des variables exogènes ξ_k à travers les **composantes principales** les plus importantes de X ainsi qu'à redéfinir (éventuellement) les concepts sous-jacents à la liste des variables exogènes (donc cette liste elle-même) (cf aussi **régression pas à pas**).

Si Q est une **matrice orthogonale** qui diagonalise X' X, ie si la **décomposition spectrale** de X s'écrit $Q'\ X'\ X\ Q = L$, avec $L = Diag\ \{l_1\ ,...,\ l_K\}$ et $l_1 \geq ... \geq l_K$, l'**estimateur sur composantes principales** $b_L\hat{}$ de b correspondant aux $L < K$ plus petites valeurs propres s'exprime en fonction de l'**estimateur des mco** $b\hat{}$ de b (déduit de (2)) selon :

(5) $b_L\hat{} = \{I_K - (X'\ X)^{-1}\ W\ (W'\ (X'\ X)^{-1}\ W)^{-1}\ W'\} . b\hat{}$,

où $W \in M_{KL}$ (**R**) est la matrice composée des L premières colonnes de Q. Autrement dit, $b_L\hat{}$ est l'estimateur des mco sous contrainte linéaire $W'\ b = 0$.

(v) Soit $\tau = (\tau_1\ ,..,\ \tau_K)$ une liste de K variables exogènes initiales, observé selon un (N,K)-**tableau statistique** initial T, et ζ une variable endogène initiale observée selon un N-vecteur z.

Ces données sont transformées selon $X = P\ T\ S^{-1}$ et $y = P\ z\ S^{-1}$ (cf **centrage**, **variable centrée**, **variable réduite**), où P désigne la **matrice de centrage**, S une **matrice de dispersion**. On note X' X la **matrice de corrélation** empirique.

On peut alors écrire la décomposition :

(6) $y = (X\ Q)(Q'\ b) + u = Z\ c + u$,

du modèle $y = X\,b + u$ initial, où $Q = [q_1, ..., q_K]$ désigne la (K,K)-matrice des vecteurs propres de X' X, $Z = X\,Q$ est la (N,K)-matrice des composantes principales associées à X et $c = Q'\,b \in \mathbf{R}^K$. Par suite, les « nouvelles » variables exogènes correspondant aux colonnes de Z sont des combinaisons linéaires orthogonales des variables initiales, définies selon :

(7) $z_k = X\,q_k\,, \quad \forall\, k \in N_K^*,$

et les « nouveaux » paramètres c_k sont des combinaisons linéaires orthogonales des paramètres initiaux, définies selon :

(8) $c_k = q_k'\,b, \quad \forall\, k \in N_K^*.$

Chaque coefficient c_k s'interprète eg comme contribution de la composante principale correspondante, soit z_k , dans l'équation (6). La **méthode de régression sur composantes principales** consiste :

(a) à considérer des parties $L \subset N_K^*$ tq Card L = L ;

(b) à définir le modèle, « restreint » à L, suivant :

(9) $y = Z_L\,c_L + u_L\,,$

où $Z_L = X\,Q_L$ (Q_L étant la sous-matrice de Q contenant les L vecteurs propres indicés par L), c_L est la **suite** des paramètres indicée par L et u_L la nouvelle « perturbation aléatoire » (qui dépend de L) ;

(c) à estimer b selon :

(10) $b^{\#} = Q_L\,c_L^{\#},$

où $c_L^{\#}$ est un estimateur de c_L (eg l'**estimateur des mco** ou l'**estimateur du mv**). On note $b_L^{\#}$ ou $b^{\#}$ (L) au lieu de $b^{\#}$.

Chacune des variables z_k dépend (en général) de l'ensemble des variables exogènes initiales (ie de X) : b peut ainsi être entièrement estimé par $b_L^{\#}$.

(vi) Ce type de méthode pose surtout un problème de **spécification** de modèle, puisque l'on est passé de la forme (2) à la forme (9). Il faut donc l'assortir de critères permettant de jauger l'importance de cette **modification** et son impact sur la qualité de l'**inférence statistique** (cf aussi **résistance**, **robustesse**).