RÉGRESSION À PLUSIEURS RÉGIMES (J1, N10)
(06 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

On appelle (modèle de) **régression à plusieurs régimes**, ou **régression à plusieurs phases**, ou **régression à régimes séparés**, ou **régression segmentée**, ou **régression par morceaux**, ou encore **régression à changements** (ou **déplacements**) **systématiques** (ou **structurels**) (en moyenne), un **modèle de régression** dont la fonction de **régression** admet plusieurs formes selon les parties de l'**espace d'observation** auxquelles appartiennent les observations des **variables exogènes** (cf aussi **modèle à structure variable**, **test de changement structurel**).

Un tel modèle est utilisé eg en contrôle de qualité (lorsque la **variable** à contrôler dépend d'autres variables déterminées : eg dérèglement d'un équipement) (cf **contrôle de réception**).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $(\mathcal{X}_0, \mathcal{B}_0)$ et $(\mathcal{Y}_0, \mathcal{C}_0)$ deux **espaces d'observation** probabilisables. On suppose que les opérations qui suivent ont un sens et que le **couple aléatoire** $(\xi, \eta) : \Omega \mapsto \mathcal{X}_0 \times \mathcal{Y}_0$ est tq l'**espérance conditionnelle** de η pr à ξ est de la forme :

(1) $E(\eta / \xi = x) = f_r(x, b_r)$ ssi $x \in B_r$, $\forall r \in N_R^*$,

où $B_r \in \mathcal{B}_0$, $\Pi_0 = (B_r)_{r=1,...,R}$ est une **partition** mesurable de $\mathcal{X}_0$ et $R \geq 2$ est un entier naturel. La forme (1) se synthétise selon :

(2) $E(\eta / \xi = x) = \Sigma_{r=1}^{R} \mathbf{1}(B_r(x)).\ f_r(x, b_r)$,

expression dans laquelle :

(a) $(f_r)_{r=1,...,R}$ est une **suite** donnée de fonctions $f_r : \mathcal{X}_0 \times \mathbf{R}^{Qr} \mapsto \mathbf{R}$ dont la forme analytique est connue, mais non le **paramètre** $b_r \in \mathbf{R}^{Qr}$ (qui est **inobservable**) ;

(b) l'application $x \mapsto E(\eta / \xi = x)$ est continue sur $\mathcal{X}_0$.

On appelle **régression à plusieurs régimes** une régression de la forme (1) (ou (2)).

Si le couple (ξ, η) est observé selon (X, y), où X est une (N,K)-**matrice** aléatoire, y un N-**vecteur aléatoire** à valeurs dans $\mathcal{Y}_0^N$ et N le nombre d'observations, on définit N équations :

(3) $y_n = E(y_n / \xi = X_n) + u_n = \Sigma_{r=1}^{R} \mathbf{1}(B_r(X_n)) .\ f_r(X_n, b_r) + u_n$

(où X_n désigne, $\forall n \in N_N^*$, la n-ième ligne de X), ie un **modèle de régression** dont on cherche à estimer les paramètres (inobservables) $b_r \in \mathbf{R}^{Qr}$ (où $Q_r \geq 1$) et les paramètres (inobservables) ∂B_r (**frontière** de B_r) ($\forall r \in N_R^*$).

(ii) Le nombre R, qui n'est pas nécessairement connu, est le **nombre de « régimes »** du modèle. La méthode habituellement utilisée pour l'**estimation** est la **méthode du maximum de vraisemblance** (dont les paramètres revêtent ainsi une forme complexe).

(iii) Souvent, le modèle est plus simple. En pratique, on a $\mathcal{X}_0 = \mathbf{R}^K$, $\mathcal{Y}_0 = \mathbf{R}$ et B_r est ($\forall\ r \in N_R^*$) un pavé de $\mathbf{R}^K$ (pavé dont les sommets sont, en général, inconnus).

Lorsque K = 1 (une seule variable exogène), on a $\mathcal{X}_0 = \mathbf{R}$ et $\Pi_0 = (B_r)_{r=1,...,R}$ est une subdivision de **R**, avec $B_r =]c_{r-1}\ ,\ c_r[$ ($\forall\ r \in N_R^*$). Lorsqu'on suppose que $x \mapsto E\ (\eta\ /\ \xi = x)$ est continue sur **R**, alors :

(4) $\quad f_{r-1}\ (c_{r-1}\ ,\ b_{r-1}) = f_r\ (c_{r-1}\ ,\ b_r), \qquad \forall\ r \in N_R^* \setminus \{1\}.$

Si R est donné, ainsi que les **points de « changement »** de régime, ou **points de « rupture »**, c_r ($\forall\ r \in N_R^*$), on se ramène à un modèle de régression non linéaire (cf **modèle non linéaire**). Si, de plus, les f_r sont toutes des fonctions linéaires (ou affines), on peut procéder à divers tests (eg le **test de CHOW**).

On admet en général que la **perturbation aléatoire** $u = (u_1\ ,...,\ u_N)'$ possède une **espérance conditionnelle** (pr à X) nulle dans la forme (3), ie $E\ u\ /\ X = 0$, ainsi qu'une **dispersion** conditionnelle $V\ u\ /\ X$ définie positive (cf **matrice définie positive**). Souvent, cette dernière est diagonale, ie $V\ u\ /\ X = \sigma^2\ D$ (cf **matrice diagonale**) et les termes $d_{\alpha\beta}$ de D sont tq $d_{\alpha\beta} = \delta_{\alpha\beta}\ /\ w_a$, où les w_a ($a \in N_N^*$) sont connus (et s'interprètent eg comme des « poids »). La méthode d'estimation est alors la **méthode des moindres carrés généralisés**, sous une forme élémentaire appelée **méthode des moindres carrés pondérés**. Cette dernière revient à minimiser pr aux b_r la quantité :

(5) $\quad \Sigma_{r=1}^{R}\ \Sigma_{D(n,r)}\ w_n\ .\ (y_n - f_r\ (X_n\ ,\ b_r))^2,$

où $D(n,r) = \{X_n : X_n \in B_r\}$, compte tenu des contraintes (4). On parle alors de **régression segmentée contrainte**.

(iv) D'autres versions du modèle sont basées sur des hypothèses différentes, eg :

(a) la fonction $x \mapsto E\ (\eta\ /\ \xi = x)$ n'est pas continue. Ceci autorise des « **ruptures** » (ie des « **catastrophes** ») correspondant aux **changements de régimes** (cf **théorie des catastrophes**) : dans ce type de modèle, les indices $n \in N_N^*$ sont remplacés par des indices $t \in N_T^*$ qui expriment explicitement que le « **temps** » intervient dans le déroulement du **phénomène** décrit par le modèle ;

(b) V u est diagonale par « **blocs** », les R blocs en question s'associant de façon naturelle aux R régimes du modèle : il y a donc **homogénéité** interne aux régimes, et **hétéroscédasticité** externe à ceux-ci (du point de vue de la variance conditionnelle V u / X, ie au second ordre).