**RÉGRESSION CONTRAINTE (J9)**
(22 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

Dans l'étude d'un **modèle de régression**, certaines **variables**, certaines **observations** effectuées sur ces variables, ou certains **paramètres** (inobservables) peuvent ne pas parcourir l'**ensemble** des valeurs possibles : en effet, il arrive que ces grandeurs soient astreintes à demeurer dans des parties strictes de leurs ensembles théoriques.

Ces parties sont le plus souvent définies par des contraintes simples (égalités ou inégalités), mais elles peuvent aussi résulter de contraintes plus complexes (eg **variétés différentielles**).

On peut distinguer deux situations courantes.

(i) **Contrainte sur les variables** ou **contrainte sur les observations**. Etant donné une **structure statistique** $(\Omega, \mathcal{T}, \mathcal{P})$, un **espace d'observation** des **variables exogènes** $(\mathcal{X}_0, \mathcal{B}_0)$ et un espace d'observation des **variables endogènes** $(\mathcal{Y}_0, \mathcal{C}_0)$, on suppose que le **couple aléatoire** $(\xi, \eta) : \Omega \mapsto \mathcal{X}_0 \times \mathcal{Y}_0$ définit un **modèle image** $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathcal{B}_0 \otimes \mathcal{C}_0)$ auquel on peut associer une fonction de régression (é, eg sous forme explicite :

(1) $\quad E\, \eta / \xi = f(\xi),$

ou sous forme implicite :

(2) $\quad E\, g(\eta, \xi) / \xi = 0,$

Dans ces formes, les opérations sont supposées licites (notamment, calculabilité de l'**espérance conditionnelle** et existence d'éléments « nuls »).

Par ailleurs, le critère de centralité E (espérance conditionnelle) peut être remplacé par toute autre notion de centralité (cf **relation fonctionnelle**, **régression**).

On dit alors que (1) (ou (2)) est un **modèle avec contrainte sur les variables**, ou **modèle sous contrainte relative aux variables**, ssi il existe une partie stricte (non vide) $S \subset \mathcal{X}_0 \times \mathcal{Y}_0$ tq (1) (ou (2)) soit vérifié par les valeurs (x, y) prises par le couple $(\xi, \eta)$ dans S.

Le problème immédiat consiste à estimer f (ou g) sous la contrainte ainsi définie.

Si la contrainte est de la forme d'équation (resp d'inéquation) :

(3)
$$S = \{(x, y) \in \mathcal{X}_0 \times \mathcal{Y}_0 : h(x, y) = 0\},$$
$$(\text{resp } S = \{(x, y) \in \mathcal{X}_0 \times \mathcal{Y}_0 : h(x, y) \leq 0\}),$$

on parle de **modèle contraint à l'égalité** (resp de **modèle contraint à l'inégalité**) **sur les variables**.

Si $\mathcal{X}_0$ x $\mathcal{Y}_0$ est un **espace métrique** et si S en est une **partie bornée**, on parle de **régression à variables bornées**.

Des exemples usuels de contraintes sur les variables sont les suivants :

(a) **identité** d'un **modèle d'interdépendance** (contrainte déterministe) ;

(b) **troncature** des **variables aléatoires** incluses dans la régression, donc des **échantillons** issus de ces va (contrainte déterministe) ;

(c) **censure** des échantillons de ces mêmes variables (**contrainte stochastique**).

On distingue parfois entre :

(a) contrainte sur les variables, exprimable sous la forme d'une liaison mathématique tq $\varphi(\xi, \eta) = 0$ (resp $\varphi(\xi, \eta) \geq 0$), et qui affecte donc a priori toutes les observations de ces variables ;

(b) contrainte sur certaines observations de ces variables, exprimable sous la forme d'une liaison entre certaines coordonnées de celles-ci (**situation statistique** moins fréquente).

(ii) **Contrainte sur les paramètres**. Comme précédemment, on admet l'existence d'un modèle de **régression non linéaire** de forme explicite :

(4) $\quad E\ \eta / \xi = f(\xi, \theta)$,

ou de forme implicite :

(5) $\quad E\ g(\xi, \eta, \theta) = 0$,

dans lequel le **paramètre d'intérêt** $\theta \in \Theta$ intervient explicitement.

On dit que (4) (ou (5)) est un **modèle contraint sur les paramètres**, ou un **modèle sous contraintes relative aux paramètres**, ssi il existe une partie stricte (non vide) (souvent une **partie bornée**) $\Lambda \subset \Theta$ tq (4) (ou (5)) soit vérifiée par toute valeur $\theta \in \Lambda$.

Il s'agit alors notamment d'estimer le paramètre $\theta$ sous la contrainte ainsi définie, ou de tester ses valeurs effectives.

Si la contrainte sur le paramètre est de la forme :

$$L = \{\theta \in \Theta : l(\theta) = 0\},$$
(6)
$$(\text{resp } L = \{\theta \in \Theta : l(\theta) \leq 0)\}),$$

on parle de **modèle contraint à l'égalité** (resp **modèle contraint à l'inégalité**) **sur les paramètres**.

Si $\Theta$ est un espace métrique et si $\Lambda$ en est une partie bornée, on parle de **régression à paramètres bornés**.

Le problème peut aussi se poser pour une régression sous forme non paramétrée, avec $E\ \eta / \xi = f(\xi)$ ou $E\ g(\xi, \eta) = 0$, dans laquelle f ou g est à estimer au sein d'une classe restreinte de **fonctions de régression** ou de **fonction d'interdépendance**.

(iii) Les deux difficultés précédentes peuvent se présenter simultanément (modèle contraint sur les variables et sur les paramètres). De façon plus générale, la notion de **loi scientifique** conduit à distinguer deux situations :

(a) d'une part, l'étude de la famille $\mathscr{P}^{\zeta}$ des **lois multivariées** $P^{\zeta}$ qui peuvent régir le **phénomène** considéré (lois « candidates »). Si l'on note $P^{\zeta}$ l'une de ces lois :

($a_1$) l'existence de contraintes sur les variables signifie que $\zeta$ ne parcourt qu'une partie stricte $D \in \mathscr{D}$ de l'ensemble $\mathscr{Z}$ de ses valeurs, à partir duquel est définit l'**espace probabilisable** $(\mathscr{Z}, \mathscr{D})$ ;

($a_2$) l'existence de contraintes sur les paramètres signifie que $P^{\zeta}$ ne parcourt qu'une partie stricte $P^0$ de la famille $\mathscr{P}^{\zeta}$. Dans le cas d'une **spécification** paramétrique, cette famille $\mathscr{P}^{\zeta}$ s'écrit sous la forme $(P_{\theta}^{\zeta})_{\theta \in \Theta}$, et c'est $\theta$ qui est contraint de parcourir une partie stricte $\Theta^0$ de $\Theta$ ;

(b) d'autre part, l'étude des **relations fonctionnelles** $\rho$ qui peuvent se déduire des lois précédentes. Les contraintes (variables ou paramètres) imposées à ces lois se répercutent donc sur ces relations.

(iv) En pratique, les problèmes d'estimation sont résolus à l'aide d'**observations** (X, Y) de $(\xi, \eta)$, notamment lorsque $\mathscr{X}_0 = \mathbf{R}^K$ et $\mathscr{Y}_0 = \mathbf{R}^G$ : par suite, les N observations $(X_n , Y_n)$ de $(\xi, \eta)$ (ie les lignes de (X,Y)) sont tq X est une **matrice** aléatoire à valeurs dans $M_{NK}$ (**R**) et Y une matrice aléatoire à valeurs dans $M_{NG}$ (**R**).

Dans ces cas, les contraintes portent sur les **matrices** (aléatoires) X et Y : eg H X = a et L Y = b.

(v) Les situations précédentes se rencontrent notamment dans le cas du **modèle de régression**. Les méthodes d'estimation mises en oeuvre nécessitent la résolution d'un **problème d'optimisation** sous contrainte (cf **programmation mathématique**).

Ainsi, le **modèle de régression multiple** non linéaire $y = F(b) + u$ avec contrainte sur le paramètre $b \in B$ (où B est une partie bornée de $\mathbf{R}^Q$) conduit au programme suivant :

(7) $\min ||y - F(b)||^2$ sous $b \in B$.

(vi) Les modèles contraints usuels sont les suivants :

(a) le modèle linéaire, estimé par la **méthode des mco**, avec contrainte linéaire sur le vecteur y des observations de la variable endogène $\eta$ (ou sur la matrice X des observations des variables exogènes $\xi$), lorsque des **identités** (eg équations de définition) existent.

C'est le cas de l'**équation de partage**, dans laquelle on estime des proportions $y_n \geq 0$ (avec $e_N' \, y = 1$), ou du **modèle à variable dépendante qualitative** (variable qualitative codée dans une partie bornée de **R** tq {0, 1}, ou [-1, +1}, ou {-1, 0, +1}).

Les contraintes (en particulier linéaires) sur $\xi$ engendrent souvent un phénomène de **colinéarité** (stricte) ;

(b) le **modèle linéaire**, estimé par la **méthode des moindres carrés ordinaires**, avec **contrainte sur le paramètre** b à l'égalité (resp à l'inégalité).

(vii) Un modèle bien spécifié, ou un modèle appliqué à des observations « régulières », peuvent rendre superflue ou inutile la prise en compte de certaines contraintes, les solutions obtenues étant identiques.

Ainsi (économie), une fonction de consommation standard s'écrit sous la forme affine $C = c\,R + b + \varepsilon$, dans laquelle le paramètre c (propension marginale à consommer le revenu disponible R) est borné selon $0 < c < 1$ et R désigne ce même revenu. L'estimation par la **méthode des mco** conduit au problème de minimisation de $\Sigma_{n=1}^{N} (C_n - c\,R_n - b)^2$ sous la contrainte $c \in ]0, 1[$, dont la solution est identique à celle des mco sans contrainte. Ceci suppose que les consommations observées $C_n$ soient (au moins pour la plupart) inférieures aux revenus observés $R_n$ .