RÉGRESSION « SPLINE » (J1)
(26 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'expression (impropre) de **régression spline** ne correspond pas à un concept, mais à une méthode de représentation et d'estimation d'un **modèle de régression** (le plus souvent de forme non paramétrique) à l'aide de **fonctions splines** (cf aussi **estimateur spline de la régression**).

(i) Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** et $(\xi, \eta) : \Omega \mapsto \mathbf{R}^2$ un **couple aléatoire** réel. On considère le **modèle non linéaire** suivant (exprimé dans un **espace de variables**) :

(1) $\eta = f(\xi) + \varepsilon$, avec $E\,\varepsilon / \xi = 0$, $V\,\varepsilon / \xi = \sigma_\varepsilon^2$,

et l'on veut estimer la **fonction de régression** f.

On note :

(2) $y_n = f(X_n) + u_n$, avec $E\,u_n / X = 0$, $V\,u_n / X = \sigma_\varepsilon^2$, $\forall\, n \in N_N^*$,

la représentation « observée » de l'équation (1) dans l'**espace d'observation** $\mathbf{R}^N$, obtenue en y substituant les **observations** y_n de η et X_n de ξ. On suppose que $C(u_\alpha, u_\beta) = \delta_{\alpha\beta} \cdot \sigma_\varepsilon^2$, $\forall\, (\alpha, \beta)$ (**matrice diagonale**) (cf **corrélation**, **matrice de corrélation**), que la **variable exogène** ξ est bornée, ie que $\xi \in [a, b]$ et donc que $a \leq X_1 < ... < X_N \leq b$ (avec $-\infty < a$ et $b < +\infty$) (eg $a = 0$ et $b = 1$).

(ii) Si f est suffisamment régulière (eg continue), son **développement de TAYLOR** (avec reste intégral) s'écrit :

(3) $f(x) = \Sigma_{j=0}^{p-1} a_j x^j + \delta(x) = \alpha(x) + \delta(x)$,

où :

(4) $\delta(x) = \int \mathbf{1}_{[a\,,\,x]}(u) \cdot \{(p-1)!\}^{-1} \cdot f^{(p)}(u) \cdot (x-u)^{p-1}\, du$,

et où $f^{(p)}$ désigne la dérivée d'ordre p de f.

Par suite, étant donné q valeurs distinctes δ_i ($i \in N_q^*$) tq $a < \delta_1 < ... < \delta_q < b$, on peut développer δ selon :

(5) $\delta(x) = \Sigma_{i=1}^{q} b_i (x - \delta_i)_+^{p-1} + \gamma(x) = \beta(x) + \gamma(x)$,

où l'on note $z_+ = \max(z, 0)$ la **partie positive** z^+ de $z \in \mathbf{R}$.

On dit alors que la représentation {(3),(4)} de f à l'aide de la fonction « spline » d'ordre p définie par $s = \alpha + \beta$ est une **représentation spline** associée à la **suite** $(\delta_i)_{i=1,...,q}$.

La fonction s constitue une **approximation** de f qui possède des dérivées continues jusqu'à l'ordre p - 2 (fonction de classe C^{p-2}). Si la suite $(\delta_i)_{i=1,...,q}$ est donnée (eg connue), l'ensemble des fonctions splines d'ordre p est un **espace vectoriel** (de dimension p + q), dont une **base** naturelle est $\{1, x,..., x^{p-1}, (x - \delta_1)_+^{p-1}, ..., (x - \delta_q)_+^{p-1}\}$.

(iii) On appelle :

(a) **régression spline** à limites (ou segments) fixes le substitut du modèle (1) défini par :

(6) $\eta = \alpha(\xi) + \beta(\xi) + \varphi$, avec $E\,\varphi / \xi = 0$, $V\,\varphi / \xi = \sigma_\varphi^2$;

(b) alternativement, **modèle de régression spline** à limites (ou segments) fixes le substitut du modèle (2) défini par :

(7) $y_n = \alpha(X_n) + \beta(X_n) + v_n$, avec $V\,v_n / X = 0$, $C(v_\alpha, v_\beta) = \delta_{\alpha\beta}\,\sigma_\varphi^2$.

Cette substitution consiste à admettre que le reste γ est « négligeable ». Les limites δ_i peuvent être les valeurs X_n elles-mêmes (avec q = N et $\delta_i = X_n$, $\forall\, n \in N_N^*$).

Lorsque les coefficients de (5) sont nuls (ie $b_1 = ... = b_q = 0$) et que $\gamma = 0$, on obtient le **modèle de régression polynômiale** usuel. En faisant varier p et $(\delta_i)_{i=1,...,q}$, le modèle « spline » ainsi défini peut donc s'adapter de façon « souple » au cas où f est inconnue ou de forme analytique complexe (cf **complexité**).

Par suite :

(a) lorsque les limites δ_i sont fixées (connues), on peut estimer le modèle (2) en estimant seulement le modèle (7), ie en estimant $\alpha + \beta$ au lieu de f. En pratique, on estime les **paramètres** a_i et b_i (avec $p + q < N$) à l'aide d'une méthode ad hoc : **méthode des moindres carrés** ou **méthode du maximum de vraisemblance** ;

(b) lorsque les limites δ_i sont inconnues, on les considère comme des **paramètres** supplémentaires, ce qui implique a priori $p + 2q < N$ (modèle « spline » à limites, ou à segments, variables).

(iv) Si l'on admet pour **fonction de perte** une **fonction quadratique pénalisée**, ie une fonction de la forme :

(8) $L(f) = N^{-1} \sum_{n=1}^{N} \{y_n - f(X_n)\}^2 + \lambda^{-1} \int \{f^{(p)}(u)\}^2\, du$,

on appelle **estimateur « spline » de la régression** (1) la solution en f du programme mathématique :

(9) $\min L(f)$ sous $f \in W^{(p)\,2}(a, b)$,

où $W^{(p)\,2}(a, b)$ désigne l'**espace de S.L. SOBOLEV** des fonctions f : **R** $\mapsto$ **R** dont la dérivée d'ordre p est de carré intégrable sur [a, b]. Cette méthode d'estimation est parfois simplement appelée **régression « spline »**.

Le **paramètre** λ (ou λ-1), appelé **« paramètre » de pénalisation**, ou **« paramètre » de régularisation**, définit un équilibre (ou arbitrage) entre un ajustement des moindres carrés (premier terme du second membre de (8)) et une pénalisation pour non régularité de f (second terme, souvent appelé **terme pénalisant** ou **terme pénalisateur**).

On détermine généralement λ à partir de (X, y), notamment à l'aide d'une méthode de **validation croisée**.

On montre que la solution (unique) $f_N^{\sim}$ (simplement notée $f^{\sim}$) de (9) est une **fonction spline polynômiale** de degré 2 p - 1. Elle dépend donc, en général, de (f, X, λ), avec $X = (X_1, ..., X_N)$.

En particulier, lorsque p = 2, on définit une **régression spline cubique**.

On montre que $f^{\sim}$ est liée à l'**estimateur par le noyau** (cf **noyau**) de f, au sens où elle est de la forme (cf **méthode du noyau**) :

(10) $\quad f^{\sim} = N^{-1} \sum_{n=1}^{N} K_N(., X_n) . y_n$.

Sous certaines hypothèses, le choix de (X, λ) permet d'obtenir une solution $f_N^{\sim}$ convergeant vers f lorsque $N \rightarrow +\infty$.

(v) Les notions précédentes s'étendent directement au cas de plusieurs variables exogènes ξ_k (avec $k \in N_K^*$) ou au cas de plusieurs **variables endogènes** η_g (avec $g \in N_G^*$).

(vi) La régression spline s'applique aussi au modèle de **régression à plusieurs régimes**, ou régression à structure variable (régression « segmentée ») (cf **modèle à structure variable**).

Ce type de méthodes constitue aussi une aide à la **modélisation** (ie à la **spécification** de **modèles**) : eg étude des réponses d'**unités expérimentales** (matériel expérimental) à des **traitements** (cf **stimulus**), analyse des **surfaces de réponse** (cf aussi **dispositif expérimental**).