ESTIMATEUR DE LA RÉGRESSION (C5, H)
(22 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

Une **fonction de régression** peut être estimée de plusieurs façons : eg **méthode des moindres carrés**, **méthode du maximum de vraisemblance**, **méthode des fonctions splines**, **méthode du noyau**.

(i) La mise en oeuvre d'une **méthode non paramétrique** est adaptée au cas où la **régression** ne dépend d'aucun **paramètre** explicite (modèle de régression non paramétrique), ie n'est pas indexable par un ensemble fini d'**indices** (ou de « paramètres ») .

(ii) Le modèle associé peut s'écrire (dans l'**espace des variables**) eg sous la forme :

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

où ξ est une liste de K **variables exogènes** et η une **variable endogène** (ou une liste de G variable endogènes, dans le cas multidimensionnel).

Le problème consiste à estimer la fonction de régression f, au vu des observations $(X_n , y_n)_{n=1,\dots,N}$ de (ξ, η). Autrement dit, on pose, avec des hypothèses usuelles :

(2) $y_n = f(X_n) + u_n$, avec $E\,u_n = 0$, $C(u_\alpha , u_\beta) = \delta_{\alpha\beta} . \sigma^2$, $\forall (\alpha, \beta) \in (N_N^*)^2$.

(iii) La **méthode du noyau** (appliquée à cette situation) consiste à estimer f à l'aide de l'estimateur par le **noyau** suivant :

(3) $f_N^{\#}(x) = \Sigma_{n=1}^N p((x - X_n) / h) . y_n / \Sigma_{n=1}^N p((x - X_n) / h)$, $\forall x \in \mathbf{R}^K$,

où :

(a) p est une **fonction de poids** ($p \geq 0$ et $\int p\, d\lambda_K = 1$) (eg une **densité** symétrique pr à 0 et de **support** Supp p = [-1, +1]) ;

(b) $h > 0$ est une **largeur de fenêtre** à déterminer (ie à estimer) pour permettre le meilleur ajustement. On l'appelle **paramètre de lissage** ou **paramètre de complexité**.

Pour cet estimateur très simple (de type NADARAYA), on montre (lorsque K = 1) que, si f est de classe C^2, alors, lorsque $N \to +\infty$ et $h \to 0+$ en sorte que $N . h \to +\infty$ (cf aussi **théorème de NADARAYA**) :

(4)
$E\,f_N^{\#}(x) \sim f(x) + (1 / 2) f''(x) h^2 + 1 -1\, 2 \int p(z)\, dz$, $\forall x \in \mathbf{R}$,
$V\,f_N^{\#}(x) \sim \sigma^2 N h + -1 \int p^2(z)\, dz$, $\forall x \in \mathbf{R}$.

(iv) Une autre forme d'estimateur par le noyau s'écrit :

(5) $f_N^{\#\#}(x) = \{(N\,h)\}^{-1} \Sigma_{n=1}^{N} J\{(x - X_n) / h\} . y_n , \qquad \forall\, x \in \mathbf{R}^K,$

forme dans laquelle $J : \mathbf{R}^K \mapsto \mathbf{R}$ est un noyau donné a priori et h la largeur de fenêtre (à déterminer à partir des observations (X_n , y_n), et généralement notée h_N). On peut, par exemple, minimiser (pr à h) l'approximation suivante :

(6) $Q^2{}_N(h) = N^{-1} E \Sigma_{n=1}^{N} (f_N^{\#\#}(X_n) - f(X_n))^2$

de l'**écart quadratique moyen intégré** de $f_N^{\#\#}$:

(7) $Q^2(f_N^{\#\#}, f) = E \int (f_N^{\#\#}(x) - f(x))^2 dx.$