RÉGRESSION AVEC ABERRATION(S) (J5)
(16 / 08 / 2021, © Monfort, Dicostat2005, 2005-2021)

Il arrive qu'une **loi multivariée** P^{ζ} soit un **mélange légal** de diverses autres **lois de probabilité**, dont l'une est celle que le **statisticien** considère comme une loi « principale », celle escomptée par lui, tandis que les autres jouent un rôle secondaire, parfois aussi un rôle perturbateur. Par suite, toute **relation fonctionnelle** directement issue de P^{ζ} ne reflète pas la véritable relation escomptée, celle reliant à titre principal les variables du modèle.

Cette **situation statistique** peut concerner, en particulier, un **modèle d'interdépendance** ou un **modèle de régression**.

(i) Soit $P^{(\xi, \eta)}$ la loi de probabilité d'un **couple aléatoire** (ξ, η), dans lequel ξ est une **variable exogène** et η une **variable endogène**. On suppose $P^{(\xi, \eta)}$ que parcourt une **famille de lois** donnée. On note $f = dP^{(\xi, \eta)} / d\mu$ la **densité** de $P^{(\xi, \eta)}$ pr à une **mesure** μ donnée, $c (. / \xi)$ la **densité conditionnelle** de η relativement à ξ, et :

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

la **fonction de régression** non linéaire associée à c, exprimée dans l'**espace des variables** (ξ, η) et dépendant d'un **paramètre** b. On suppose que toutes les opérations ont un sens.

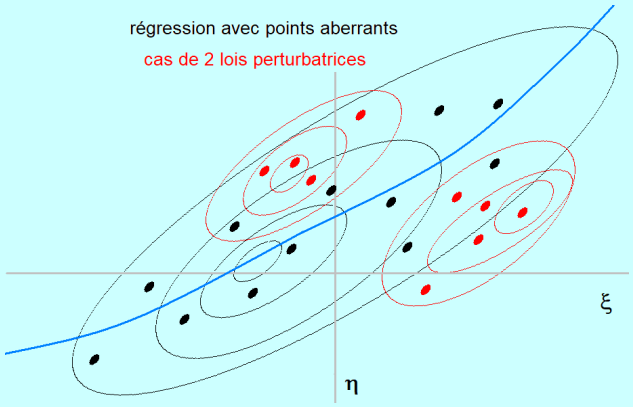
(a) on suppose que c est une densité c* contaminée par une **suite** constituée de p densités $(f_j)_{j=1,...,p}$, ie par définition (cf **contamination des lois**, **mélange légal**) :

(2) $\quad c = \alpha^* . c^* + \sum_{i=1}^{p} \alpha_i . c_i$,

avec $\alpha^* + \sum_{i=1}^{p} \alpha_i = 1$ et $(\alpha^*, (\alpha_i)_{i=1,...,p}) \in S_{p+1}$ (**simplexe** correspondant). Toutes ces densités ont donc pour variables le couple (ξ, η) considéré.

Le graphique suivant, exprimé dans l'espace $\mathbf{R}^{1+K}$ de ces variables, concerne le cas de deux lois perturbatrices, dont les densités c_i (réduites par leurs coefficients α_i) sont représentées sous forme de courbes de niveau en rouge, et les observations correspondantes (supposées en nombre fini) sont représentées par des « points » rouges. La loi principale est représentée par sa densité (réduite par son coefficient α^*) dont les lignes de niveau sont reproduites en noir, de même que les observations correspondantes.

La fonction de régression contaminée de η en ξ est représentée par une ligne bleue, correspondant à la densité contaminée c. On observe les inflexions apportées par les points aberrants rouges, alors que la vraie fonction de régression c* devrait être approximativement un hyperplan dans $\mathbf{R}^{1+K}$.



(b) la description précédente s'adapte directement au cas d'une loi $P^{(\xi, \eta)}$ (resp densité c) à queue épaisse.

(ii) On suppose que l'équation (1) est « observée » selon le modèle de **régression non linéaire** à **perturbation** additive (exprimé dans un **espace d'observation** (X, y)) :

(3) $y = F(b) + u$, avec $E\, u / X = 0$,

où (X, y) désigne les N **observations** de (ξ, η) et F une fonction associée à c, et dépendant de f et de X.

On dit que (3) est l'équation d'un **(modèle de) régression avec aberrations** ssi il existe une suite (X_i , y_i), avec $i \in I$ et $I \subset N_N^*$, constituée d'**aberrations** dans l'**ensemble** des **données**. Autrement dit, un point (X_i , y_i) de cette suite est un **point aberrant** (ou **point atypique**, ou encore **point excentré**) pour la lp $P^{(X, y)}$ (ie pour la **loi jointe** du couple (X, y)) : en général, un tel point figure dans une **queue** de cette loi (**loi à queue épaisse**) ou encore $P^{(X, y)}$ est une **loi contaminée** (eg mélange de lois dont les paramètres de centralité sont « très » différents, ou dont certaines possèdent une dispersion importante) (cf **contamination des lois**, **loi contagieuse**).

(iii) L'analyse d'un tel modèle implique donc d'identifier soit les queues épaisses (qui correspondent à des y_i ou à des x_{ik}), soit les composantes du mélange légal.

Ainsi, lorsque I est connu, que (3) est linéaire (ie Q = K et $F = X \in M_{NK}$ (**R**)) et qu'il s'agit d'estimer b, on peut associer à tout point aberrant (X_i , y_i) une **variable indicatrice** :

$$(4) \quad d_n = \begin{cases} 1 \text{ si } n = i, \\ 0 \text{ sinon.} \end{cases}$$

Le modèle devient alors un **modèle d'analyse de la covariance** auquel on peut appliquer les procédures d'estimation ou de test courantes (**méthode de moindres carrés**, **maximum de vraisemblance**, etc).

(iv) La même technique est encore applicable seulement pour certaines coordonnées de (X_i , y_i) : ainsi, y_i seul peut être aberrant, ou x_{ik} (k donné) seul peut être aberrant, etc.

(v) Une autre méthode consiste à remplacer (X_i , y_i) par un « point moyen » ou par une « prévision fictive » (exo-prévision) déduits des autres points (points non aberrants). Ceci suppose qu'un test préalable de détection de points aberrants (cf **test d'aberration**) ait été mis en oeuvre (cf eg **estimateur de pré-test**).

(v) Dans certains cas, il est possible de traiter les points aberrants :

(a) soit comme des observations manquantes (cf **lacune**) ;

(b) soit comme des **paramètres importuns** (ou **paramètres incidents**) supplémentaires.

Moyennant certaines hypothèses et certaines adaptations, les méthodes d'estimation ou de test usuelles sont applicables. Ainsi, on suppose souvent eg que la proportion du nombre d'aberrations est relativement « faible » pr au nombre des autres observations (notamment dans les procédures asymptotiques).

Enfin, la théorie de la **robustesse** permet de traiter un modèle de régression avec aberrations (cf **régression robuste**) en sorte que ces dernières n'aient qu'une influrence minimale sur l'estimation des paramètres ou les tests effectués sur ceux-ci.