P. CAZES

Méthodes de régression

Les cahiers de l'analyse des données, tome 3, nº 2 (1978), p. 147-165
<http://www.numdam.org/item?id=CAD_1978__3_2_147_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 2 - p. 147-165

MÉTHODES DE RÉGRESSION

I. — La régression sous contraintes

[RÉGR. CONTR.]

par P. Cazes (1)

Introduction

On sait que la régression usuelle fournit parfois des résultats dénués d'intérêt, soit parce que les coefficients de régression sont très sensibles aux fluctuations d'échantillonnage, soit parce que ces coefficients ne sont pas interprétables (certains coefficients étant par exemple négatifs, alors que pour avoir un sens, ils doivent être positifs). On se trouve en particulier dans les cas précédents si certaines variables explicatives sont très corrélées, ou si le nombre de ces variables est grand (vis-à-vis du nombre d'observations).

Nous considérons ici pour les comparer un certain nombre de méthodes utilisées pour protéger la régression contre ces résultats fortuits et mal interprétables ; après quelques rappels sur la régression usuelle,et ses limitations, nous détaillons les techniques de régression sous contraintes et de régression biaisée (estimateur sous contraintes de positivité des coefficients de régression, estimateur borné (ridge) où l'on impose à la norme du vecteur de régression de ne pas être trop élevée etc). Nous montrons en particulier que la plupart des estimateurs biaisés rencontrés dans la littérature peuvent être considérés d'une part comme des estimateurs sous contraintes, et d'autre part comme des cas particuliers de l'estimateur borné (ridge) généralisé, ce dernier pouvant lui-même être considéré, (sous des hypothèses de normalité) comme un estimateur bayésien.

Enfin, nous faisons des rappels sur quelques techniques de régression par l'analyse des données : régression par boule et régression par l'analyse des correspondances.

Nous ne détaillerons pas ici les diverses applications que nous avons traitées, nous contentant d'en énumérer ci-dessous un certain nombre, en renvoyant le lecteur aux publications correspondantes.

Nous avons commencé à appliquer la régression sous contrainte de positivité des coefficients de régression dans [5], pour tenter d'expliquer le taux d'un élément trace dans une roche en fonction des taux des différents éléments majeurs constitutifs de la roche, des coefficients de régression négatifs n'ayant pas de sens d'un point de vue géologique.Nous avons ensuite (cf [6], [7]) appliqué les programmes conçus pour la régression sous contrainte de positivité, conjointement avec ceux conçus pour la régression bornée et la régression sur variables entachées d'erreurs à un problème physique qui sans être à strictement parler une régression est justiciable des mêmes calculs : l'estimation (d'après un modèle) de la courbe granulométrique d'un aérosol ; les trois types de contraintes ont donné des résultats voisins et interprétables, alors que la régression usuelle ne fournissait que des résultats aberrants. Notons que dans toutes ces régressions , l'on avait imposé à la somme des coefficients de régression d'être égale à 1, l'ensemble de ces coefficients pouvant être considéré comme une distribution de probabilités.

(1) Maître assistant, laboratoire de statistique ; Université P.& M.Curie

Nous avons également (cf [1]) appliqué les trois types précédents de programmes sous contraintes à un autre problème d'estimation posé par l'efficacité de photomultiplicateurs à dynodes. Il s'agissait encore d'estimer des proportions (de somme inférieure ou égale à 1). Seules la régression bornée et la régression sur variables entachées d'erreurs ont fourni des résultats physiquement acceptables, la régression usuelle fournissant des coefficients de régression négatifs, ou plus grands que 1, et la régression sous contrainte de positivité des coefficients de régression nuls, encadrant des coefficients positifs (zéros en milieu de spectre), ce qui était incompatible avec le modèle physique, qui supposait une certaine continuité des coefficients de régression. En fait, le modèle précédent, établi par le physicien faisant cette étude revient à estimer une courbe continue g(x) par un histogramme à 7 ou 8 classes. Nous avons dans un second temps repris cette étude en cherchant une estimation plus fine de g(x) à l'aide d'un histogramme de 30 à 50 classes. Pour réaliser cette estimation, le nombre de variables explicatives étant élevé (de 30 à 50), nous avons effectué l'analyse des correspondances du tableau des variables explicatives (qui était en fait un tableau de lois de probabilités) ; l'approximation obtenue en projetant la variable à expliquer en supplémentaire sur les 4 ou 5 premiers axes de cette analyse permet d'estimer la courbe g(x) (cf [14]).

Nous avons encore (cf [15]) utilisé la régression sous contrainte de positivité afin d'expliquer le rendement y de la culture de la betterave en fonction d'un nombre assez important de variables. Outre cette régression sous contrainte de positivité des coefficients de régression (influence positive des variables sur y), nous avons utilisé la régression pas à pas, et la régression sous contrainte de positivité sur la variable z = max y - y, max y désignant la valeur maximale de y sur l'échantillon étudié, cette dernière analyse permettant de voir l'influence négative des variables explicatives dans la reconstitution de y. L'ensemble de ces méthodes a permis d'affiner les résultats obtenus avec la régression usuelle qui a bien sûr été faite.

Nous avons enfin (cf [12] et [13]) appliqué la régression par boule et la régression par l'analyse des correspondances (après découpage en classes des variables explicatives et à expliquer) pour essayer d'expliquer le taux de matière organique (ou kérogène) dans une roche en fonction d'un grand nombre de caractéristiques hétérogènes telles que le taux de carbonate, la granulométrie, la couleur, etc.

Signalons pour terminer que l'on peut trouver d'autres exemples d'application dans [19] , [22], [27], [34] en ce qui concerne la régression biaisée et la régression sous contraintes, et dans [3] en ce qui concerne la régression par l'analyse des données.

Dans le présent cahier, après quelques rappels (§§ 1 , 2), nous considérons en détail les méthodes de régression sous contraintes. Des articles ultérieurs seront consacrés aux critères bayésiens (§ 4) et aux méthodes fondées sur l'analyse des correspondances (§ 5, accompagné d'un article d'application). Une bibliographie détaillée (≃35 titres) paraîtra à la fin de la série.

1 *Rappels sur la régression*

Etant donné un espace euclidien R^J dont on désignera par N le produit scalaire définissant la norme, et p + 1 vecteurs $x^J_1, \ldots, x^J_p, x^J_{p+1}$, on cherche en régression à expliquer l'un de ces vecteurs en fonction linéaire des autres, de telle sorte que la norme de la différence entre ce vecteur et son approximation soit minimale. Supposons pour fixer les idées que l'on veuille expliquer x^J_{p+1} que l'on notera aussi y^J ou $x^J_{I_2}$, I_2 désignant l'ensemble à un élément : p + 1. Posant $I_1 =]p]$ *,

* *la notation]n] désigne l'ensemble des entiers 1, 2,..., n.*

$I = I_1 \cup I_2 =]p+1]$, $X^J_I = \{X^J_1, \ldots, X^J_{p+1}\}$, $X^J_{I_1} = \{X^J_1, \ldots, X^J_p\}$, on recherchera donc la combinaison linéaire

$yy^J = b^{I_1} \circ X^J_{I_1} = \Sigma\{b^i\, X^J_i \mid i \in I_1\}$ * telle que $y^J - yy^J$ soit de norme minimale. La solution $yy^J_\circ$ est donc la projection de y^J sur le sous espace W_1 engendré par les $\{X^J_i \mid i \in I_1\}$. Nous supposerons que les $\{X^J_i \mid i \in I_1\}$ sont linéairement indépendants, cas auquel on peut toujours se ramener, quitte à éliminer certains d'entre eux ; ces vecteurs forment donc une base de W_1 qui est isomorphe à R^{I_1}, base que nous utiliserons. Nous désignerons aussi par W le sous-espace de R^J engendré par W_1 et y^J (sous-espace qui est isomorphe à R^I, si y^J n'est pas dans W_1) et par V_{II} ou plus simplement V la matrice des produits scalaires $\{\langle X^J_i, X^J_{i'}\rangle_N \mid i, i' \in I \times I\}$, qui est la matrice associée à la métrique (ou pseudo métrique** si y^J est dans W_1, auquel cas W et W_1 sont confondus) induite par N sur W quand on repère W par les $\{X^J_i \mid i \in I\}$.

Posons $b^I = (b^{I_1}, 0)$, $\varphi^I = (-b^{I_1}, 1) = \delta^I_{p+1} - b^I$ (en bref, b^I (resp. φ^I) est la fonction sur I dont la restriction à I_1 vaut b^{I_1} (resp. $-b^{I_1}$) tandis que sa restriction à l'ensemble à un élément I_2 vaut 0 (resp. 1), δ^I_{p+1} étant la fonction nulle sur I_1 et égale à 1 sur $I_2 = \{p+1\}$) ; on a alors : $yy^J = b^I \circ X^J_I = \Sigma\{b^i X^J_i \mid i \in I\}$, $y^J - yy^J = \varphi^I \circ X^J_I = \Sigma\{\varphi^i X^J_i \mid i \in I\}$; désignant par H_1 le sous espace de R^I engendré par les p premiers vecteurs $\delta^I_1, \ldots, \delta^I_p$ de la base canonique de R^I (i.e. vecteurs dont toutes les composantes sont nulles, sauf une d'entre elles qui vaut 1) et par H'_1 le sous espace affin de H_1 et passant par δ^I_{p+1}, on est alors ramené à chercher dans R^I muni de la métrique (ou pseudo métrique) V, la combinaison linéaire φ^I de H'_1 de norme minimale. La solution $\varphi^I_\circ = \delta^I_{p+1} - b^I_\circ$ est donc la projection de l'origine sur H'_1, tandis que $b^I_\circ$ est la projection de δ^I_{p+1} sur H_1.

En découpant V en blocs suivant I_1 et I_2 :

$$V = \begin{array}{cc} \begin{array}{cc} I_1 & I_2 \end{array} & \\ \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} & \begin{array}{c} I_1 \\ I_2 \end{array} \end{array} \qquad (1)$$

où V_{11} désigne la métrique induite par N et $X^J_{I_1}$ sur R^{I_1} et V_{22} le carré de la norme de y^J, et en désignant par A_{H_1}, $A_{H'_1}$, A_{W_1} les projecteurs respectifs sur H_1, H'_1 et W_1, R^I étant muni de la métrique V et R^J de la métrique N, on a, puisque d'après les hypothèses faites V_{11} est régulière:

* *Nous utilisons ici les notations tensorielles, avec des conventions analogues à celles adoptées dans [Note Lim.] TII B n° 1. Ainsi l'écriture* $yy^J = b^{I_1} \circ X^J_{I_1}$ *signifie que :* $\forall\, j \in J : yy^j = \Sigma\{b^i\, X^j_i \mid i \in I_1\}$, *formule que l'on peut aussi écrire :* $yy^j = b^{I_1} \circ X^j_{I_1}$, yy^j *(resp.* X^j_i*) désignant la* $j^{ème}$ *composante de* yy^J *(resp.* X^J_i*) tandis que* $X^j_{I_1}$ *désigne l'ensemble des* $\{X^j_i \mid i \in I_1\}$.

** *i.e. métrique pour laquelle la nullité de la norme d'un vecteur n'implique pas que ce vecteur est nul.*

$$\left.\begin{aligned} b^I_\circ &= A_{H_1}\ \delta^I_{p+1} \\ yy^J_\circ &= A_{W_1}\ y^J_\circ = b^I_\circ \circ x^J_I = b^{I_1}_\circ \circ x^J_{I_1} \\ \varphi^I_\circ &= A_{H'_1}\ O^I \end{aligned}\right\} \quad (2)$$

$$\left.\begin{aligned} b^{I_1}_\circ &= V^{-1}_{11}\ V_{12} \\ \| y^J - yy^J_\circ \|^2_N &= \| \varphi^I_\circ \|^2_V = V_{22} - V_{21}\ V^{-1}_{11}\ V_{12} \end{aligned}\right\} \quad (3)$$

où l'indice $\circ$ signifie que l'on a affaire à la solution de la régression usuelle par opposition aux solutions que l'on obtiendra si l'on impose des contraintes (cf § 3).

Suivant le cas, on raisonnera soit dans R^J, soit dans R^I, ou soit dans R^{I_1}.

On peut caractériser la qualité de l'ajustement obtenu en remplaçant y^J par yy^J par l'indice

$$S^2 = \| y^J - yy^J \|^2_N / \| y^J \|^2_N = \| \varphi^I \|^2_V / V_{22} \quad (4)$$

qui est l'indice que l'on minimise dans la régression, et dont la valeur minimale $S^2_\circ$, obtenue pour $yy^J = yy^J_\circ$, n'est autre que le carré du sinus entre y^J et $yy^J_\circ$.

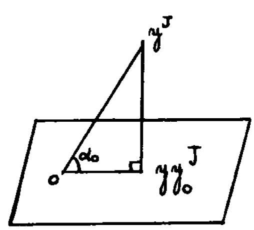
Au lieu de considérer S^2, on utilise aussi la quantité

$$R^2 = \| yy^J \|^2_N / \| y^J \|^2_N = \| b^I \|^2_V / V_{22} \quad (5)$$

dont la valeur maximale $R^2_\circ$ est obtenue pour $yy^J = yy^J_\circ$.

Notons que $R^2_\circ$ n'est autre que le carré du cosinus entre y^J et $yy^J_\circ$, et l'on a :

$$R^2_\circ = 1 - S^2_\circ = V^{-1}_{22}\ V_{21}\ V^{-1}_{11}\ V_{12}$$



$\cos \alpha_0 = R_0 ; \sin \alpha_0 = S_0$

Il revient donc au même de minimiser S^2 que de maximiser R^2. L'intérêt d'utiliser S^2 plutôt que R^2 réside dans le fait que si l'on impose des contraintes la relation

$$R^2 + S^2 = 1 \quad (6)$$

n'est plus en général vérifiée, et que dans ce cas, on ne cherche pas à maximiser R^2, mais à minimiser S^2 sous les contraintes imposées.

Remarques

1) Dans le cas où l'on a un modèle, si $E(y^J)$ et $Var(y^J)$ désignent respectivement l'espérance mathématique et la variance de y^J, on pose :

$$\left.\begin{aligned} E(y^J) &= \beta^{I_1} \circ X^J_{I_1} \\ \mathrm{Var}(y^J) &= \sigma^2\ \Gamma_\circ \end{aligned}\right\} \quad (7)$$

$\Gamma_\circ$ étant une matrice définie positive connue, σ^2 et β^{I_1} étant inconnus. Dans ce cas, on adopte pour métrique la métrique $N = \Gamma_\circ^{-1}$; $b_\circ^{I_1}$ ainsi que $yy^J_\circ$ sont alors des estimateurs sans biais, de dispersion minimale parmi les estimateurs fonctions linéaires de y^J, (théorème de Gauss-Markov) de β^{I_1} et $E(y^J)$, tandis que $s^2 = \| y^J - yy^J_\circ \|^2/(\mathrm{Card}\, J - p)$ est un estimateur sans biais de σ^2. Dans ce cas, la matrice variance de b^{I_1} s'écrit :

$$\mathrm{Var}\ b^{I_1} = V_{11}^{-1}\ \sigma^2 \quad (8)$$

2) Dans le cas de la régression usuelle, J est un échantillon d'individus affectés de poids $p_j = 1/\mathrm{Card}\, J$; on prend alors pour métrique N la métrique des poids qui n'est autre ici au facteur $1/\mathrm{Card}\, J$ près que la métrique usuelle de R^J. On supposera de plus que le tableau X^J_I est centré $(\Sigma\{p_j\ X^j_I \mid j \in J\} = 0)$. Dans ce cas $R^2_\circ$ n'est autre que le carré du coefficient de corrélation multiple de y^J par rapport aux $\{X^J_i \mid i \in I_1\}$.

Si les p_j ne sont pas tous égaux, on a affaire à la régression pondérée où N est toujours la métrique des poids de matrice la matrice diagonale de $j^{\text{ème}}$ terme diagonal p_j.

Si y^J est gaussien, il suffit de raisonner conditionnellement aux $\{X^J_i \mid i \in I_1\}$ pour retrouver le modèle (7).

3) Si l'on désire expliquer q variables $X^J_{p+1}, \ldots, X^J_{p+q}$ en fonction des $\{X^J_i \mid i \in I_1\}$ en effectuant q régressions, la première formule (3) est encore valable ; il suffit de poser $I_2 = \{p+1, \ldots, p+q\}$, $I = I_1 \cup I_2$, $X^J_{I_2} = \{X^J_i \mid i \in I_2\}$, $X^J_I = (X^J_{I_1}, X^J_{I_2})$, et d'effectuer la partition (1) de V suivant I_1 et I_2. Si $b^{I_1}_{I_2} \circ X^J_{I_1}$ désigne l'approximation de $X^J_{I_2}$, on a donc :

$$b^{I_1}_{I_2} = V_{11}^{-1}\ V_{12}$$

4) Au lieu de rechercher dans R^I la combinaison linéaire φ^I de norme minimale pour V, sous la condition que la dernière composante φ^{p+1} de φ^I soit égale à 1, on peut rechercher la combinaison linéaire φ^I minimisant toujours $\|\varphi^I\|^2_V$ mais sous la condition de normalisation $\|\varphi^I\|^2_M = 1$, M désignant une métrique donnée de R^I. La solution φ^I est fournie par le vecteur propre normé (pour M) associé à la plus petite valeur propre de $M^{-1}V$. Dans ce cas où toutes les variables jouent un rôle symétrique, on parle de moindres carrés orthogonaux.

2 *Limitations de la régression*

Les résultats fournis par la régression peuvent dans certains cas être illusoires, car très sensibles aux fluctuations d'échantillonnage, ou même non interprétables. C'est en particulier le cas si le nombre p des variables explicatives est relativement élevé vis-à-vis de la dimension Card J de l'échantillon sur lequel on travaille, ou si les variables explicatives sont très corrélées, en sorte que la régression excellente sur l'échantillon particulier traité peut ne rien valoir pour des cas nouveaux ; ou encore si le programme fournit des coefficients quelconques, alors que pour avoir un sens, les coefficients de régression

b^i, ou les quantités yy^j approximations des y^j doivent être compris entre deux bornes, par exemple être positifs, ou compris entre 0 et 1, cas fréquents en pratique.

Reprenons chacun des points précédents :

Supposons donc qu'on ait un grand nombre de variables explicatives ; prenons pour fixer les idées, le cas limite où $p = \text{Card}\,J$. Dans ce cas, les $\{x^J_i \mid i = 1, p\}$ forment, sauf cas particulier une base de R^J, et les coefficients de régression b^i ne sont autres que les coordonnées de y^J sur cette base. On a donc une reconstitution parfaite de y^J ($yy_o^J = y^J$; $R^2_o = 1$), et ceci quel que soit la nature de la dépendance de la variable à expliquer y avec les variables explicatives $\{x_i \mid i = 1, p\}$; la reconstitution de y est donc illusoire, et l'on conçoit qu'il en soit encore de même, si le nombre de variables explicatives p, tout en étant plus petit que Card J, est voisin de ce nombre.

Supposons maintenant que les variables explicatives sont très corrélées, auquel cas la matrice V_{11} est pseudo-singulière, et donc les coefficients de régression donnés par (3) vont avoir des valeurs élevées ; de plus si on est dans le cas d'un modèle, la variance de certains de ces coefficients va également d'après (8) être élevée. Dans ces conditions pour expliquer une valeur y^j de y on a une reconstitution yy^j somme d'effets $b^i\, x_i^j$ grands en valeur absolue et se retranchant ; on obtient ainsi une reconstitution illusoire de y^j. On peut noter que l'augmentation du nombre de variables explicatives peut conduire à l'obtention de variables très corrélées (sur l'échantillon J sur lequel on travaille).

Pour protéger la régression contre ces résultats sans grande signification, on peut utiliser des techniques de sélection de variables, des techniques de régression orthogonale, ou un compromis entre ces deux types de techniques.

Du point de vue sélection de variables, la technique la plus intéressante nous semble être l'introduction ascendante des variables, qui est une technique de pas à pas : au pas 1, on introduit la variable explicative la plus corrélée à y ; soit x_{i_1} cette variable ; au pas 2, on choisit parmi les variables restantes, la variable x_i telle que le carré R^2 du coefficient de corrélation multiple de y par rapport à (x_{i_1}, x_i) soit maximum ; de façon générale, ayant sélectionné $x_{i_1}, \ldots, x_{i_{h-1}}$, on introduit au pas h la variable x_i telle que le carré du coefficient de corrélation multiple de y par rapport à $x_{i_1}, \ldots, x_{i_{h-1}}, x_i$ soit maximum. On arrête la procédure quand l'augmentation du coefficient R^2 quand on passe d'un pas au pas suivant n'est pas significative, ce qui revient à tester à l'aide du t de Student usuel que le coefficient de régression associé à la variable introduite n'est pas significativement différent de zéro.

Cette technique a été améliorée par Effroymson (cf [17]), qui à chaque pas remet à l'épreuve les variables déjà introduites ; l'on peut ainsi dans cette méthode de pas à pas (*step-wise*) éliminer une variable déjà introduite à un pas précédent. On doit noter toutefois que même avec cette technique améliorée on n'obtient pas en général en gardant r variables, la meilleure régression (au sens du critère R^2), correspondant à ce nombre r.

De plus, au lieu de baser le critère d'arrêt sur le coefficient R^2, on peut songer à utiliser la méthode de l'échantillon de base et échantillon d'épreuve, méthode très générale et qui vaut, quel que soit le type de régression effectuée : régression usuelle, pas à pas, sous contrainte, etc... . Rappelons-en le principe, avant de voir son application comme critère d'arrêt d'une régression pas à pas. On divise l'échantillon initial J en deux parties J_1 et J_2, (Card J_1 + Card J_2 = Card J, Card J_2 étant de l'ordre de 10 à 20% de Card J, les éléments de J_2 étant tirés au hasard dans J) ; on effectue la régression sur J_1 dit échantillon de base, d'où un coefficient R_1^2, puis on applique la formule de régression ainsi trouvée sur J_2 dit échantillon d'épreuve, d'où un coefficient R_2^2, carré de la corrélation (mesurée sur J_2) entre y et son approximation. Si R_2 est de l'ordre de R_1, on a une formule de régression qui peut permettre de faire de la prévision (i.e. prévoir la valeur de y pour un cas nouveau connaissant pour ce cas la valeur des variables explicatives, la précision de la prévision étant fournie par R_1 ou R_2 qui sont du même ordre de grandeur). Si R_2 est très inférieur à R_1, la formule de régression trouvée n'est valable que sur l'échantillon J_1 (i.e. sur l'échantillon sur lequel elle a été établie), et non pour un cas nouveau.

En ce qui concerne la régression pas à pas, on effectuera la régression sur J_1 et l'on portera sur un graphique en fonction du pas (ou du nombre de variables conservées) R_1 et R_2 ; et l'on arrêtera la procédure quand R_2 n'augmentera plus ; notons que cette procédure est très utilisée en analyse discriminante pas à pas, le coefficient R étant remplacé par le critère des bien classés.

Du point de vue de la régression orthogonale, qui revient à faire de la régression sur des variables explicatives non corrélées, on peut faire l'analyse factorielle du tableau $X_{I_1}^J$ des variables explicatives (analyse en composantes principales ou analyse factorielle des correspondances) en mettant y^J en supplémentaire et prendre comme nouvelles variables les facteurs issus de cette analyse (cf §5.2.1). On sélectionne alors les facteurs associés aux plus grandes valeurs propres, et parmi ces facteurs on ne retient que les plus corrélés à y. Notons qu'ainsi on risque d'éliminer un facteur très corrélé à y, mais associé à une valeur propre faible ; un tel facteur n'a souvent qu'un pouvoir explicatif fortuit dans la mesure où, étant associé à une faible valeur propre, il est très sensible aux fluctuations d'échantillonnage ; il est donc normal de ne pas en tenir compte. Mais si l'on veut que la détermination même des facteurs soit en rapport avec le pouvoir explicatif de ceux-ci, on pourra analyser un tableau de Burt croisant les modalités de la variable à expliquer avec celles des variables explicatives ; ce qui est une sorte d'analyse canonique particulière cf fin de ce §.

Si l'on veut interpréter les coefficients de régression, et si les facteurs représentent des entités complexes, ce qui peut être le cas en particulier des facteurs de rang supérieur à deux ou trois, facteurs qui peuvent représenter un pourcentage d'inertie non négligeable et qui peuvent donc être sélectionnés, s'ils sont assez corrélés à y, il semble préférable au lieu de faire la régression sur facteurs, de faire la régression sur des variables sélectionnées d'une part en fonction de leur corrélation avec y, et d'autre part de leur position et de la qualité de leur représentation dans le sous espace des premiers axes factoriels. C'est cette dernière méthode, synthèse en quelque sorte des techniques précédentes ; analyse factorielle plus sélection de variables à partir de cette analyse, méthode qui a l'avantage de faire voir la structure des variables explicatives, qui nous semble du point de vue pratique la plus intéressante.

Si l'on a un modèle, et que l'on veuille garder toutes les variables du modèle, les méthodes de protection précédentes ne peuvent pas s'appliquer.

Pour protéger la régression, on peut faire de la régression biaisée (cf § 4) ou imposer des contraintes (cf § 3).

Le fait d'imposer des contraintes est également requis dans certains cas, si l'on veut pouvoir interpréter les résultats de la régression : si par exemple comme on l'a dit au début de ce paragraphe, pour avoir un sens les coefficients de régression doivent être compris entre deux bornes, ou si y^j varie entre deux bornes, auquel cas son approximation yy^j devra être entre ces mêmes bornes, pour avoir un résultat sensé, on rajoutera ces contraintes, i.e. on minimisera la norme de $y^J - yy^J$ sous les contraintes précédentes, ce qui revient à tenir compte des informations *a priori* dont on dispose pour effectuer la régression.

Signalons pour terminer ce paragraphe que quand on a un mélange hétérogène de variables explicatives : quantitatives, semi-quantitatives, qualitatives, on ne peut, sauf cas particulier comme l'analyse de covariance, appliquer les techniques de régression usuelle. Dans ce cas on homogénéïsera le problème en rendant toutes les variables, y compris la (ou les) variable à expliquer, qualitatives, par découpage en classes des variables quantitatives, et en faisant l'analyse des correspondances du tableau de contingence croisant l'ensemble des modalités (ou des classes) de la (ou des) variable à expliquer avec l'ensemble des modalités de toutes les variables explicatives, analyse où l'on ajoutera en supplémentaire les individus caractérisés par les variables explicatives (cf § 5.2.2). On pourra alors sur les facteurs ainsi obtenus sur les individus, effectuer soit une régression usuelle, soit une régression par boule (i.e. une régression où la valeur de y en un point M de l'espace des variables explicatives, ici l'espace des premiers facteurs, est estimée par la moyenne des valeurs de y pour les points situés dans un voisinage de M (cf § 5.1)). On peut noter que la méthode précédente qui de par sa conception, protège la régression contre des résultats fortuits, s'applique également si y est qualitatif ; dans ce cas on pourra effectuer sur les facteurs calculés sur les individus soit une analyse discriminante usuelle, soit une analyse discriminante par voisinage.

3 *Régressions sous contraintes*

3.1 *Généralités*

On recherche ici, avec les notations du § 1 que l'on conserve, le point $yy^J = b^{I_1} \circ X^J_{I_1} = b^I \circ X^J_I$ de l'espace W_1 engendré par les $\{X^J_i \mid i \in I_1\}$, le plus proche de y^J, R^J étant toujours muni de la métrique N, mais l'on impose des contraintes aux coefficients de régression $\{b^i \mid i \in I_1\}$, ce qui revient à restreindre le domaine où b^{I_1} varie à une partie C de l'espace R^{I_1} :

$$b^{I_1} \in C \subset R^{I_1}$$

et donc le domaine où yy^J varie à une partie D de W_1 :

$$yy^J \in D = \{z^J = b^{I_1} \circ X^J_{I_1} = \Sigma\{b^i X^J_i \mid i \in I_1\} \mid b^{I_1} \in C\} \subset W_1$$

(C désignant le sous ensemble de R^{I_1} où doit se trouver b^{I_1} pour que les contraintes soient vérifiées *)

La solution que l'on notera $yy^J_+ = b^{I_1}_+ \circ X^J_{I_1} = b^I_+ \circ X^J_I$ est donc la projection de y^J sur D.

Nous supposerons que cette projection est unique, ce qui est en particulier réalisé dans le cas très fréquent en pratique où C (et donc D) est un convexe fermé.

Compte tenu de la relation :

$$\|y^J - b^{I_1} \circ X^J_{I_1}\|^2_N = \|y^J - b^{I_1}_\circ \circ X^J_{I_1}\|^2_N + \|(b^{I_1}_\circ - b^{I_1}) \circ X^J_{I_1}\|^2_N = \|\varphi^{I}_\circ\|^2_V + \|b^{I_1}_\circ - b^{I_1}\|^2_{V_{11}} \quad (9)$$

* On suppose que les contraintes ne sont pas incompatibles, i.e. que C est non vide.

l'on déduit que yy_+^J (resp. $b_+^{I_1}$) est la projection de $yy_\circ^J$ (resp. $b_\circ^{I_1}$) sur D (resp. C), R^J (resp. R^{I_1}) étant muni de la métrique N (resp. V_{11}).

Si la solution $b_\circ^{I_1}$ de la régression classique vérifie les contraintes, i.e. si $b_\circ^{I_1} \in C$, $b_+^{I_1} = b_\circ^{I_1}$, sinon $b_+^{I_1}$ (resp. yy_+^J) est la projection de $b_\circ^{I_1}$ (resp. y^J ou $yy_\circ^J$) sur la frontière de C (resp. D).

On caractérisera la qualité de l'ajustement obtenu en approchant y^J par yy_+^J à l'aide du critère S^2 (cf (4)) dont la valeur S_+^2 pour $yy^J = yy_+^J$ est bien sûr supérieure ou égale à la valeur $S_\circ^2$ obtenue dans la régression sans contrainte. Si l'origine appartient à C, S_+^2 est plus petit que 1. Si l'on considère la valeur R_+^2 du critère R^2 (cf(5)) pour $yy^J = yy_+^J$, on a en général $R_+^2 + S_+^2 \neq 1$ (i.e. la relation(6) n'est pas vérifiée) sauf dans le cas où C est délimité par des sous espaces vectoriels, ce qui correspond à des contraintes linéaires d'égalité ou d'inégalité.

Remarque : Posant $b^I = (b^{I_1}, 0)$, $\varphi^I = (-b^{I_1}, 1) = \delta_{p+1}^I - b^I$, $\varphi_+^I = \delta_{p+1}^I - b_+^I$, la relation (9) s'écrit encore :

$$\left.\begin{aligned} \|\varphi^I\|_V^2 &= \|\varphi_\circ^I\|_V + \|\varphi^I - \varphi_\circ^I\|_V^2 \\ &= \|\varphi_\circ^I\|_V^2 + \|b^I - b_\circ^I\|_V^2 \end{aligned}\right\} \quad (10)$$

Posant : $C_1 = \{b^I = (b^{I_1}, 0) \mid b^{I_1} \in C\} \subset H_1 \subset R^I$

$C_1' = \{z^I = \delta_{p+1}^I - b^I \mid b^I \in C_1\} \subset H_1' \subset R^I$

l'on déduit de (10) que b_+^I qui est la projection de δ_{p+1}^I sur C_1, est aussi la projection de $b_\circ^I$ sur C_1, tandis que φ_+^I qui est la projection de l'origine sur C'_1 est aussi la projection de $\varphi_\circ^I$ sur C'_1, R^I étant toujours muni de la métrique V.

3.2 *Etude du cas particulier où l'on impose des contraintes linéaires d'inégalité.*

3.2.1 *Cas général*

Nous n'étudierons pas le cas où l'on impose des contraintes linéaires d'égalité, puisque cela revient à faire une projection sur un sous espace vectoriel (ou affin) et qu'en incorporant ce type de contrainte directement dans le modèle étudié, on revient à la régression usuelle.

Supposons donc qu'on ait des contraintes linéaires d'inégalité, contraintes qui peuvent toujours se mettre sous la forme :

$$\forall \alpha \in A \quad : b^{I_1} \circ F_{I_1}^\alpha \leq G^\alpha \qquad (11)$$

ou de façon condensée :

$$b^{I_1} \circ F_{I_1}^A \leq G^A \qquad (12)$$

A, $F_{I_1}^A$ et G^A étant connus.

L'ensemble C de R^{I_1} où doit se trouver b^{I_1} pour que les contraintes soient vérifiées est donc un polyèdre convexe fermé, délimité par les sous espaces affins $b^{I_1} \circ F_{I_1}^\alpha = G^\alpha$. La forme linéaire $b_+^{I_1}$ minimisant $\|y^J - b^{I_1} \circ X_{I_1}^J\|_N^2$ est alors telle que

$$b^{I_1} \circ F_{I_1}^{A_1} = G^{A_1} \qquad (13)$$

$$b^{I_1} \circ F_{I_1}^{A_2} \lneq G^{A_2} \qquad (14)$$

où (A_1, A_2) constitue une partition de A fonction de y^J, partition où A_1 ou A_2 peuvent être l'ensemble vide.

En d'autres termes, si la solution $b_o^{I_1}$ de la régression classique vérifie les contraintes (12), i.e. si $b_o^{I_1}$ appartient à C, $b_+^{I_1}$ est égal à $b_o^{I_1}$ et A_1 est l'ensemble vide ; sinon $b_+^{I_1}$ étant la projection (pour V_{11}) de $b_o^{I_1}$ sur la frontière de C appartient à l'une des faces de C, face caractérisée par le sous ensemble A_1 de A pour lequel les inégalités sont remplacées par des égalités. Notons qu'une fois que A_1 est connu, pour déterminer $b_+^{I_1}$, il suffit de projeter y^J sur le sous espace W'_1 de W_1, engendré par les combinaisons linéaires $b^{I_1} \circ X_{I_1}^J$ satisfaisant (13), ce sous espace étant un sous espace affin si $G^{A_1 I_1}$ est différent de zéro. Dans le cas où G^A est nul, la solution de la régression sous contrainte étant obtenue par projection sur un sous espace vectoriel ($\forall A_1 \subset A$, $G^{A_1} = 0$), la relation (6) reste valable.

3.2.2 *Etude du cas où l'on impose aux coefficients de régression d'être positifs.*

Ce cas très fréquent en pratique revient à particulariser les contraintes (12) en prenant $A = I_1$, $G^A = 0$, $F_{I_1}^A = -\delta_{I_1}^{I_1}$, $\delta_{I_1}^{I_1}$ étant l'application identité de R^{I_1} dans lui-même. Il présente un grand intérêt quand on a affaire à une régression entre variables positives ; en effet, imposer aux coefficients de régression d'être positifs implique une reconstitution yy^j de chaque valeur y^j de y par une somme de termes positifs, et évite donc une reconstitution illusoire de y^j par une suite de termes de signes différents et se compensant.

Pour minimiser $Q = \|y^J - b^{I_1} \circ X_{I_1}^J\|_N^2$ qui est une fonction du second degré des coefficients de régression b^{i}, en imposant à ces coefficients d'être positifs, on peut appliquer des méthodes de programmation quadratique, comme la méthode du gradient ou des méthodes dérivées comme la méthode de D'Esopo ; dans cette dernière méthode, au lieu de choisir la direction de déplacement suivant le gradient, on se déplace successivement, pour minimiser Q en respectant les contraintes, suivant les axes de coordonnées de R^{I_1} ; c'est cette dernière méthode que nous avons programmée, car elle a l'intérêt pratique d'être relativement rapide.

Comme l'on sait que $b_+^{I_1}$ est soit égal à $b_o^{I_1}$, si tous les coefficients de la régression classique sont positifs, soit appartient à une face de C qui est le cône positif de R^{I_1}, il suffit pour calculer $b_+^{I_1}$ de déterminer la face à laquelle il appartient, i.e. le sous ensemble A_1 de I_1 pour lequel les coefficients de régression sont nuls, ou ce qui revient au même le sous ensemble $A_2 = I_1 - A_1$ sur lequel porte effectivement la régression. En ce sens, on peut dire que la régression sous contraintes de positivité des coefficients de régression est une technique de sélection de variables ; d'où l'idée d'employer un algorithme d'introduction progressive des variables, avec le cas échéant élimination de variables déjà rentrées pour trouver la solution $b_+^{I_1}$, i.e. le sous ensemble A_2, qui peut être l'ensemble I_1 tout entier, ou l'ensemble vide. L'avantage de cette façon d'opérer réside dans le fait qu'en un nombre fini de pas, on a la solution exacte obtenue par projection de y^J sur la face engendrée par les $\{X_i^J \mid i \in A_2\}$, alors que dans les méthodes itératives de programmation quadratique évoquée plus haut, on sait qu'elles convergent, quand le nombre d'itérations tend vers l'infini, mais la convergence peut être lente, et de plus la solution obtenue est une solution approchée.

Nous décrivons au paragraphe suivant l'algorithme dont nous venons de parler, et que nous empruntons à Lawson et Hanson (cf [24]).

3.2.3 *Algorithme*

1) Poser $A_1 = I_1$; $A_2 = \emptyset$; $b^{I_1} = 0^{I_1}$

2) Calculer $GR_{I_1} = \langle y^J - b^{I_1} \circ X^J_{I_1} , X^J_{I_1} \rangle_N$

3) Si $A_1 = \emptyset$ ou si $GR_i \le 0$, $\forall\, i \in A_1$, aller en 12) ; sinon aller en 4)

4) Sélectionner un indice $t \in A_1$ tel que : $GR_t = \max\{GR_i \mid i \in A_1\}$

5) Faire passer t de A_1 à A_2

6) Faire la régression classique sur $X^J_{A_2}$; soit c^{A_2} le vecteur de régression obtenu ; poser $c^{I_1} = (0^{A_1}, c^{A_2})$

7) Si $c^i \ge 0$, $\forall\, i \in A_2$, poser $b^{I_1} = c^{I_1}$ et aller en 2) ; sinon aller en 8)

8) Sélectionner l'indice $q \in A_2$ tel que :
$$b^q/(b^q - c^q) = \min\{b^j/(b^j - c^j) \mid c^j \le 0,\ j \in A_2\}$$

9) Poser $\alpha = b^q/(b^q - c^q)$

10) Poser $b^{I_1} = b^{I_1} + \alpha(c^{I_1} - b^{I_1})$

11) Faire passer de A_2 à A_1 tous les indice $i \in A_2$ pour lesquels $b^i = 0$* ; aller en 6)

12) Fin ; b^{I_1} est la solution cherchée.

3.2.4 *Commentaire de l'algorithme*

Dans tout le déroulement du programme, les coefficients de b^{I_1} sont toujours positifs ou nuls, et en fin de programme b^{I_1} est la solution cherchée. L'ensemble A_2 correspond aux variables qui sont dans la régression, et l'ensemble A_1 aux variables qui n'y sont pas.

On part du cas où l'on n'a aucune variable dans la régression et donc où $b^{I_1} = 0$.

GR_{I_1} n'est rien d'autre au coefficient 1/2 près que le gradient de la quantité $-Q = - \| y^J - b^{I_1} \circ X^J_{I_1} \|^2_N$ que l'on cherche à maximiser, puisqu'on veut minimiser Q.

Aux étapes 1, 2, 3, $b^{I_1} \circ X^J_{I_1} = b^{A_2} \circ X^J_{A_2}$ est la projection de y^J sur le sous espace engendré par les $\{X^J_i \mid i \in A_2\}$; il en résulte puisque tous les coefficients de b^{I_1} sont positifs ou nuls (positifs sur A_2, nuls sur A_1) que si $A_2 = I_1$, i.e. $A_1 = \emptyset$, on a obtenu la solution qui est dans ce cas identique à la solution sans contrainte. De même si GR_i est négatif pour tout i appartenant à A_1 , on a obtenu un point stationnaire, qui est donc la solution. Notons que GR_i est nul (au pas 2), par construction pour tout i appartenant à A_2 , puisque la variable associée X^J_i étant dans la régression, elle est orthogonale à $y^J - b^{I_1} \circ X^J_{I_1}$.

* *En particulier, ayant $b^q = 0$ par construction, et $q \in A_2$, on fait passer q de A_2^1 à A_1.*

L'indice t que l'on introduit dans la régression étant tel que $<y^J - b^{I_1} \circ X^J_{I_1}, X^J_t> = GR_t$ est positif, le coefficient c^t obtenu dans la régression à l'étape 6 quand on vient de l'étape 5 est positif (cf supra).

Dans la boucle partant de l'instruction 6 et allant jusqu'à l'instruction 11, on élimine de la régression les variables précédemment rentrées, et qui du fait de l'introduction de X^J_t ont des coefficients de régression négatifs. Notons que lorsqu'on sort de cette boucle pour revenir en 2), i.e. quand les variables rentrées dans la régression ont tous leurs coefficients positifs, la variable X^J_t rentrée à l'étape 5 n'a pas été éliminée, et se trouve donc dans la régression alors qu'elle n'y était pas ; de plus lorsque l'on revient à l'étape 2, la quantité $\| y^J - b^{I_1} \circ X^J_{I_1} \|^2$ que l'on minimise est plus petite que la valeur qu'elle avait prise lors du précédent passage en 2) ce qui assure la convergence de l'algorithme.

Pour démontrer les résultats précédents concernant la boucle 6 - 11, désignons par $b_1^{I_1}$ la valeur de b^{I_1} à l'entrée de cette boucle (i.e. la valeur de b^{I_1} à l'étape 2), par $b_{r+1}^{I_1}$ et $c_r^{I_1}$ la valeur de b^{I_1} et de c^{I_1} à l'issue du $r^{\text{ème}}$ tour de cette boucle. Nous noterons également par $Q(b^{I_1})$ la quantité critère $\| y^J - b^{I_1} \circ X^J_{I_1} \|^2$ que l'on minimise.

Si l'on sort de la boucle pour r = 1, on a $b_2^{I_1} = c_1^{I_1}$ et donc $Q(b_2^{I_1}) = Q(c_1^{I_1}) < Q(b_1^{I_1})$, puisque pour $b_2^{I_1}$ on a une variable de plus (la variable t) dans la régression que pour $b_1^{I_1}$.

Si l'on sort pour $r \geq 2$, on a chaque fois que l'on passe en 10) :

$$
\begin{aligned}
Q(b_{r+1}^{I_1}) &= \| y^J - b_{r+1}^{I_1} \circ X^J_{I_1} \|^2 = \\
&= \| \alpha (y^J - c_r^{I_1} \circ X^J_{I_1}) + (1-\alpha)(y^J - b_r^{I_1} \circ X^J_{I_1}) \|^2 \\
&\leq \alpha \| y^J - c_r^{I_1} \circ X^J_{I_1} \|^2 + (1-\alpha) \| y^J - b_r^{I_1} \circ X^J_{I_1} \|^2 \\
&= \alpha\, Q(c_r^{I_1}) + (1-\alpha)\, Q(b_r^{I_1}) \\
&\leq Q(b_r^{I_1})
\end{aligned}
$$

La première inégalité est l'inégalité de convexité, tandis que la seconde résulte du calcul même de $c_r^{I_1}$ qui minimise Q, pour le sous ensemble de variables ayant un coefficient non nul pour $b_r^{I_1}$.

Enfin, si l'on sort de la boucle au $s^{\text{ème}}$ tour on a, d'après 7)

$$Q(b_{s+1}^{I_1}) = Q(c_s^{I_1}) \leq Q(b_s^{I_1})$$

l'inégalité ayant lieu pour la même raison que précédemment.

On a donc dans tous les cas

$$Q(b_{s+1}^{I_1}) \leq Q(b_1^{I_1})$$

ce qui implique que la variable t a un coefficient de régression b_{s+1}^t strictement positif, i.e. se trouve pour $b_{s+1}^{I_1}$ dans la régression, sinon on aurait l'inégalité inverse puisque l'ensemble A_2 des variables rentrant dans la régression et associé à $b_1^{I_1}$ contiendrait l'ensemble A_2 associé à $b_{s+1}^{I_1}$.

Il nous reste maintenant à montrer que pour $r = 1$, le coefficient c_1^t est positif, i.e. que l'on ne risque pas de ressortir de la boucle avec le même sous ensemble A_2 qu'en 2), avant l'adjonction de t. On a donc le résultat suivant à montrer :

Si $GR_t = \langle y^J - b_1^{A_2} \circ X_{A_2}^J , X_t^J \rangle \gtrless 0$, avec $\langle y^J - b_1^{A_2} \circ X_{A_2}^J , X_i^J \rangle = 0$ $\forall\, i \in A_2$, ceci traduisant que $b_1^{A_2} \circ X_{A_2}^J$ est la régression de y^J par rapport à $X_{A_2}^J$, le coefficient c^t de X_t^J dans la régression de y^J par rapport aux $\{X_i^J \mid i \in A_2 \cup t = A'_2\}$ est strictement positif.

Le résultat découle immédiatement du lemme suivant où l'on a posé $z^J = y^J - b_1^{A_2} \circ X_{A_2}^J$, ce qui ne change pas dans la régression sur A'_2 la valeur du coefficient relatif à la variable t.

Lemme : Si z^J est orthogonal aux variables $\{X_i^J \mid i \in A_2\}$ et si $\langle z^J, X_t^J \rangle = GR^t \gtrless 0$, alors le coefficient c^t de régression de X_t^J dans la régression de z^J sur les variables $\{X_i^J \mid i \in A_2 \cup t\}$ est strictement positif.

Démonstration :

Soit $b^{A_2} \circ X_{A_2}^J + c^t X_t^J$ la régression de z^J sur les variables de $A_2 \cup t$. On a :

$$\| b^{A_2} \circ X_{A_2}^J + c^t X_t^J \|^2 = \langle b^{A_2} \circ X_{A_2}^J + c^t X_t^J , z^J \rangle$$
$$= c^t\, GR_t$$

d'où l'on déduit puisque GR_t est positif que c^t est positif ou nul, c^t ne pouvant pas en fait être nul, car s'il l'était, la rég on de z^J sur l'ensemble $A_2 \cup t$ se réduirait à 0, et donc z^J serait orthogonal à X_t^J, auquel cas GR_t serait nul contrairement à l'hypothèse.

Remarques :

1) On peut débuter l'algorithme avec un sous ensemble A_2 non vide, à condition que tous les coefficients de la régression de y^J sur $X_{A_2}^J$ soient positifs ou nuls.

2) On peut avant d'introduire la variable t à l'étape n° 4 tester que X_t^J n'est pas combinaison linéaire des $\{X_i^J \mid i \in A_2\}$. Si X_t^J est trop proche du sous espace vectoriel engendré par les $\{X_i^J \mid i \in A_2\}$, on sélectionnera la variable $t' \in A_1-\{t\}$ telle que $GR_{t'}$ soit maximum; si ce maximum est négatif, on arrêtera l'algorithme, sinon on introduira t' dans A_2 si elle n'est pas trop corrélée à $X_{A_2}^J$. Si t' est trop corrélée à $X_{A_2}^J$, on recherchera $t'' \in A_1-\{t,t'\}$, telle que $GR_{t''}$ soit maximum, et on continuera le processus comme pour t'.

3.2.5 *Comparaison de l'algorithme précédent avec POSIC.*

Pour juger de la rapidité de l'algorithme précédent, surnommé NNLS (*Non Negative Least Square*) par Lawson et Hanson, dont nous avons repris le paquet de cartes, nous l'avons comparé à l'algorithme que nous avons programmé, et dont une première version se trouve dans [5]. Dans notre algorithme appelé POSIC (ou POSI2) nous adoptons la démarche suivante :

1) Calcul de la matrice V des produits scalaires entre les $\{X_i^J | i \in I\}$ (cf § 1).

2) Régression usuelle sur I_1 et sélection du sous ensemble I'_1 de I_1 associé aux variables ayant un coefficient de régression strictement positif. Si $I'_1 = I_1$, on a la solution de régression sous contraintes, sinon on va en 3).

3) Régression usuelle sur I'_1 et sélection du sous ensemble I''_1 de I'_1 associé aux variables ayant un coefficient de régression strictement positif.

4) Attaque du sous programme ESOPN de régression sous contraintes de positivité par la méthode de D'ESOPO (cf § 3.2.2) avec comme point de départ le point b^{I_1} défini de la façon suivante :

$\forall\, i \in I_1 - I''_1 \;:\; b^i = 0$

$\forall\, i \in I''_1 \;:\;$ b^i est la valeur (qui par construction est positive) du coefficient de régression calculé en 3).

Nous avons comparé les performances de POSIC et NNLS, en ce qui concerne la rapidité, à l'aide de deux jeux de données, le premier relatif à des données physiques, le second à des données agronomiques.

Dans le premier cas (cf tableau 1), on avait huit variables explicatives, et l'on a fait dix essais ; d'un essai à l'autre, seul le vecteur y^J à expliquer variait, la matrice $X_{I_1}^J$ des variables explicatives (en nombre égal à 8) restant constante. Le temps d'exécution de NNLS * qui variait assez peu d'un essai à l'autre, et qui restait compris entre 111 et 130 ms, s'est révélé toujours supérieur, quand il n'y avait pas d'itération dans ESOPN , au temps d'exécution de POSIC qui variait entre 46 et 56 ms ; par contre, quand il y avait des itérations dans ESOPN, (itérations qui étaient de l'ordre de la trentaine), les deux programmes avaient des durées d'exécution comparables, POSIC ayant l'avantage de fournir d'un point de vue statistique des résultats plus complets que ceux de NNLS. On peut noter que sur ce jeu de données, il y avait au plus un coefficient de régression sous contraintes nul, ce qui exigeait pour NNLS la rentrée de pratiquement toutes les variables, et ce qui explique que NNLS soit relativement plus long pour ce jeu de données que POSIC. On peut encore remarquer, en ce qui concerne POSIC, que le calcul de la matrice V demande 34 ms, alors que la régression usuelle (triangularisation de V par la méthode de Doolittle-Choleski, et inversion) ne demande que 6,4 ms, ce qui explique que quand il n'y a pas d'itération dans ESOPN (ou quand il y en a peu) POSIC soit relativement rapide.

Dans le second cas, on avait des données relatives à l'étude du rendement de la culture de la betterave (cf introduction et [15]) que l'on désirait expliquer en fonction de 23 variables divisées en deux groupes de cardinaux respectifs 14 et 9. On a fait la régression totale et les régressions sur chacun des deux groupes, en centrant et sans centrer les variables. Les résultats comparatifs entre POSIC et NNLS de ces six essais sont donnés dans le tableau 2. On voit que lorsque ESOPN exige un nombre important d'itérations, POSIC est beaucoup plus lent que NNLS ; alors que si le nombre d'itérations dans ESOPN est faible (inférieur à 20), POSIC est beaucoup plus rapide que NNLS, sauf si le nombre de coefficients de régression non nuls à l'issue de la régression sous contraintes est faible, auquel cas NNLS rentrant peu de variables est très rapide.

En conclusion, si le nombre d'itérations dans ESOPN est faible (inférieur à 20 ; il est souvent de l'ordre de 1 à 5 itérations), ce qui est

* *Tous les essais ont été effectués sue l'UNIVAC 1110 de la faculté des sciences d'Orsay.*

fréquemment le cas, et ce qui se produit quand le point de départ b^{I_1} de ESOPN se trouve sur la face où est située la solution $b_+^{I_1}$, ou au voisinage de cette face, POSIC est plus rapide que NNLS, sinon NNLS est plus rapide.

essai	1	2	3	4	5	6	7	8	9	10
NI	0	0	0	0	0	34	0	0	25	28
t_{ESOPN}	2,2	2,2	2,2	0	0	62	0	2,2	46,1	51,3
t_{POSIC}	55,5	55,5	55,5	46,1	46,7	115,3	46,2	55,5	98,9	104,0
t_{NNLS}	111,2	129,4	129,9	117,8	115,9	113,8	116	113,9	107,6	107,6
N	7	7	7	8	8	7	8	7	7	7

Temps de calcul comparés entre NNLS et POSIC dans le cas de données physiques : Card J = 100 ; card I_1 = 8.

D'un essai à l'autre, seul change le vecteur y^J à expliquer, la matrice $X_{I_1}^J$ des variables explicatives restant la même. NI désigne le nombre d'itérations effectuées dans le sous programme ESOPN, et N le nombre de coefficients de régression non nuls à l'issue de la régression sous contraintes.

On a indiqué dans le temps de calcul de POSIC la part due au sous programme ESOPN. Notons que le temps de calcul dans POSIC de la matrice V est égal à 34,4 ms, alors que la régression usuelle (triangularisation de V par la méthode de Doolittle-Choleski et inversion) demande 6,4 ms.

Les temps donnés dans le tableau sont en millisecondes.

Tableau n° 1

essai	1	2	3	4	5	6
Card I_1	23	23	14	14	9	9
NCEN	0	1	0	1	0	1
$t_{\text{calcul de V}}$	150	164	69	78	37,4	43,3
$t_{\text{régression usuelle}}$	68	68	21	21	8,3	8,3
t_{ESOPN}	1131	59	75	34	128	7,0
t_{POSIC}	1365	315	180	148	184	66
t_{NNLS}	519	506	247	243	100	46
NI	118	5	17	7	60	2
N	7	11	7	9	2	2

Temps de calcul comparés entre POSIC et NNLS dans le cas de données agronomiques (card J = 93). D'un essai à l'autre seule change la matrice $X_{I_1}^J$ des variables explicatives, le vecteur à expliquer y^J restant le même. Si NCEN = 1, on centre les variables, sinon NCEN = 0. NI désigne toujours le nombre d'itérations dans ESOPN, et N le nombre de coefficients de régression non nuls à l'issue de la régression sous contraintes.

On a donné certains temps de calcul internes à POSIC, comme celui relatif au calcul de V, à la régression usuelle (inversion de V après triangularisation), ou au sous programme ESOPN. Les temps donnés dans le tableau sont en millisecondes.

Tableau n° 2

Signalons que NNLS a le désavantage par rapport à POSIC, de ne pas donner la solution de la régression usuelle sauf bien sûr si tous les coefficients de la régression usuelle sont positifs ; de plus NNLS travaille directement sur la matrice $X_I^J = (X_{I_1}^J, y^J)$ alors que POSIC après avoir calculé la matrice V des produits scalaires entre les X_i^J (cf § 1) ne travaille plus que sur V. Il en résulte que par suite des erreurs d'arrondi dans le calcul de V, NNLS est plus précis que POSIC. Par contre, si Card J est élevé (> 300) NNLS risque de devenir beaucoup moins rapide que POSIC. En outre pour Card J très élevé (Card J > 10.000), NNLS semble impraticable, soit par dépassement de capacité de la mémoire (puisque X_I^J est en mémoire centrale), problème soluble par accroissement du temps à l'aide d'un fichier auxiliaire, soit par un temps de calcul prohibitif ; alors que par un aménagement facile de POSIC, on peut au lieu d'introduire X_I^J en mémoire centrale, lire X_I^J ligne par ligne et calculer V au fur et à mesure.

Remarques :

1) Pour que l'algorithme NNLS devienne praticable, pour de grandes valeurs de Card J , il faudrait, comme pour POSIC, raisonner directement sur la matrice variance V (calculée ligne par ligne),ce qui implique un remaniement complet du programme de Lawson et Hanson.

2) Un programme de régression fournissant les résultats de la régression usuelle, de la régression sous contraintes de positivité des coefficients, et de la régression *step-wise* (cf § 2) est en cours de mise au point par A. Bohy. Ce programme qui travaille sur le tableau X_I^J en mémoire centrale, et qui emploie l'algorithme NNLS, utilise le fait que cet algorithme et l'algorithme *step-wise* de Effroyymson sont très semblables ; en effet dans les deux cas ce sont des techniques d'introduction (ascendante) des variables, avec le cas échéant élimination de variables déjà introduites.

3.2.6 Remarques sur la régression sous contraintes de positivité des coefficients

1) La régression sous contraintes de positivité permet d'étudier les liaisons positives de la variable à expliquer y^J avec les$\{X_i^J \mid i \in I_1\}$. Si l'on veut étudier les liaisons négatives de y^J avec ces variables, il suffit de remplacer y^J par $-y^J$(ou $X_{I_1}^J$ par $-X_{I_1}^J$) et d'appliquer la régression sous contraintes de positivité. Remarquons que si l'on a affaire à une régression entre variables positives, il vaut mieux pour garder le caractère additif de la reconstitution remplacer y^j par $\max\{y^j \mid j \in J\} - y^j$. On a d'ailleurs intérêt, pour garder cette reconstitution additive, à ne pas centrer les variables, mais à ajouter, le cas échéant, aux variables,la variable certaine (constante 1).

Régression usuelle, régression sous contraintes de positivité et régression sous contraintes de négativité, permettent de mieux appréhender la liaison de y^J avec les$\{X_i^J \mid i \in I_1\}$; nous avons appliqué cette façon de procéder, conjointement avec la régression pas à pas, dans l'étude dont nous avons déjà parlé, sur le rendement de la betterave, en fonction d'un grand nombre de variables explicatives (cf[15])

2) Si l'on a un modèle (ce qui est en particulier le cas si on fait l'hypothèse de normalité cf § 1, remarques 1 et 2), on peut, une fois déterminé le sous ensemble A_2 correspondant aux coefficients de régression

strictement positifs, calculer, en raisonnant conditionnellement à A_2 la matrice variance du vecteur $b_+^{A_2}$ des coefficients de régression non nuls, ce qui d'un point de vue pratique est fort intéressant. Nous avons également appliqué cette façon de faire, (qui vaut aussi dans le cas de contraintes linéaires d'inégalité quelconques), à l'étude déjà citée sur le rendement de la betterave.

3.3 *Régression sous contrainte quadratique d'inégalité*

3.3.1 *Cas général*

On suppose ici R^I muni d'une forme bilinéaire symétrique R, dont la restriction à R^{I_1} est R_{11} , et l'on désire minimiser la somme des moindres carrés $Q = \|y^J - b^{I_1} \circ X^J_{I_1}\|^2_N = \|\varphi^I\|^2_V = \|\varphi^I_o\|^2_V + \|b^{I_1} - b_o^{I_1}\|^2_{V_{11}}$ (φ_o^I et $b_o^{I_1}$ désignant toujours la solution de la régression usuelle (cf § 1)) sous la contrainte :

$$T = R(\varphi^I, \varphi^I) \leq 0 \qquad (15)$$

Partitionnant R suivant I_1 et $I_2 = \{p+1\}$:

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \qquad (16)$$

la contraite précédente peut se mettre sous la forme, puisque $\varphi^I = (-b^{I_1}, 1)$: $T = R_{11}(b^{I_1}, b^{I_1}) - 2 R_{21} b^{I_1} + R_{22} \leq 0$ (17)

Nous désignerons par U la surface du second degré T = 0, qui délimite le domaine C de R^{I_1} défini par (17).

La solution $b_+^{I_1}$ qui est dans R^{I_1} la projection (pour la métrique V_{11}) de $b_o^{I_1}$, solution de la régression usuelle sur C est de la forme :

$$b_+^{I_1} = (V_{11} + k R_{11})^{-1} (V_{12} + k R_{12}) \qquad (18)$$

où k est un multiplicateur de Lagrange défini de façon à ce que (17) soit vérifié : si $b_o^{I_1}$ vérifie (17), k = 0 ; sinon $b_+^{I_1}$ étant la projection de $b_o^{I_1}$ sur la frontière de C, i.e. sur U, k est déterminé de telle sorte que T = 0, la forme de l'expression (18) ayant été obtenue par dérivation de l'expression Q + k T par rapport à b^{I_1}.

On peut noter que si $b_o^{I_1}$ n'appartient pas à C, la projection de $b_o^{I_1}$ sur U, qui existe toujours, peut ne pas être unique, e.g. si C n'est pas un convexe fermé (si C est l'extérieur d'un ellipsoïde par exemple), et si $b_o^{I_1}$ appartient à un élément de symétrie de U. Dans les cas particuliers que nous étudions ci-après, et qui sont les cas rencontrés en pratique, la projection de $b_o^{I_1}$ sur U (quand $b_o^{I_1}$ n'appartient pas à C) est toujours unique.

Remarque :

Au lieu de rechercher φ^I minimisant $\|\varphi^I\|^2_V$ sous la contrainte (15) , avec la condition $\varphi^{p+1} = 1$, on peut rechercher φ^I minimisant $\|\varphi^I\|^2_V$ sous la contrainte plus générale $R(\varphi^I, \varphi^I) \leq -a^2$, avec $\|\varphi^I\|^2_M = 1$, a étant fixé et M une métrique donnée de R^I.

On obtient ainsi une généralisation des moindres carrés orthogonaux (cf § 1, remarque 4 *in fine*). Si $R(\varphi^I, \varphi^I) = -(\varphi^{p+1})^2$ on recherchera donc φ^I, minimisant $\|\varphi^I\|_V^2$ sous les conditions $\|\varphi^I\|_M^2 = 1$ et $(\varphi^{p+1})^2 \geq a^2$. Ce dernier problème a été étudié, dans le cas où la matrice associée à la métrique M, (dans la base canonique de R^I) est diagonale, par Legendre (cf[27]) qui parle de *ridge*-régression orthogonale.

3.3.2 *Régression bornée*

On suppose ici R^I muni de la métrique M dont la restriction à R^{I_1} est M_{11} et l'on impose la contrainte

$$\|b^{I_1} - c^{I_1}\|_{M_{11}}^2 \leq a^2 \tag{19}$$

où a et c^{I_1} sont connus.

L'intérêt de cette contrainte est de limiter la norme du vecteur de régression et donc de limiter dans la reconstitution de la variable à expliquer y la valeur absolue des termes qui rentrent dans cette reconstitution, ce qui évite une reconstitution illusoire de y par une somme de termes contraires, donc se compensant, grands en valeur absolue par rapport à y, cette reconstitution illusoire survenant en particulier quand on a des variables explicatives très corrélées (cf § 2).

La contrainte (19) qui est un cas particulier de la contrainte (17) s'obtient à partir de cette dernière en posant :

$$R_{11} = M_{11} \;;\; R_{12} = M_{11}c^{I_1} \;;\; R_{22} = \|c^{I_1}\|_{M_{11}}^2 - a^2 \tag{20}$$

La solution $b_+^{I_1}$, qui est la projection de $b_o^{I_1}$ sur l'intérieur de l'ellipsoïde défini par (19), qui est un convexe fermé, est unique, et est donnée par (18) où l'on a remplacé R_{11} par M_{11} et R_{12} par $M_{11}c^{I_1}$:

$$b_+^{I_1} = (V_{11} + kM_{11})^{-1}\,(V_{12} + k\,M_{11}\,c^{I_1}) \tag{21}$$

et l'on peut montrer (cf [8]) que dans le cas où $c^{I_1} = 0$, l'unique valeur de k minimisant Q sous la contrainte (19) est positive ou nulle.

Si $c^{I_1} = 0$, on obtient l'estimateur borné (*ridge*) donné par Hoerl et Kennard (cf [21] et [22]) dans le cas où M_{11} est la métrique usuelle de R^{I_1}. Si de plus $M_{11} = V_{11}$ on obtient l'estimateur raccourci $b_o^{I_1}/(1+k)$.

Remarques :

1) Si l'on pose $b^I = (b^{I_1}, 0)$, $c^I = (c^{I_1}, 0)$, $\varphi^I = \delta^I_{p+1} - b^I$, $\psi^I = \delta^I_{p+1} - c^I$, (19) s'écrit aussi

$$\|b^I - c^I\|_M^2 = \|\varphi^I - \psi^I\|_M^2 \leq a^2$$

2) Si l'on impose les contraintes :

$$\left.\begin{array}{l}\|b^{I_1}\|_{M_{11}}^2 \leq a^2\\ \Sigma\{b^i \mid i \in I_1\} = 1\end{array}\right\}\tag{22}$$

et si l'on pose

$$I'_1 = \{1, 2, \ldots\ p-1\}$$
$$I''_1 = \{p\}$$

le vecteur $b_+^{I_1}$ minimisant Q sous les contraintes (22) est de la forme :

$$\left.\begin{array}{l}b_+^{I'_1} = (A_{I'_1I'_1} + kE_{I'_1I'_1})^{-1}\,(A_{I'_1I_2} + k\,F_{I'_1I_2})\\ b_+^p = 1 - \Sigma\{b^i \mid i \in I'_1\}\end{array}\right\}\tag{23}$$

avec :

$$\left.\begin{aligned} \forall\, i,i' \in I'_1 \ : A_{ii'} &= v_{ii'} + v_{pp} - v_{ip} - v_{i'p} \\ A_{ip+1} &= v_{ip+1} + v_{pp} - v_{ip} - v_{p+1,p} \\ E_{ii'} &= m_{ii'} + m_{pp} - m_{ip} - m_{i'p} \\ F_{ip+1} &= m_{pp} - m_{ip} \end{aligned}\right\} (24)$$

m_{ij} (resp. v_{ij}) désignant le terme général de M_{11} (resp. V).

3.3.3 *Régression sur variables entachées d'erreurs.*

On suppose ici que J est un échantillon muni d'une mesure p_J de masse totale 1, et que le tableau X_I^J est centré, R^J étant muni de la métrique des poids de matrice, la matrice diagonale des $\{p_j \mid j \in J\}$ (cf § 1 remarque 2) ; V est alors la matrice variance associée à X_I^J et $Q = \|y^J - b^{I_1} \circ X_{I_1}^J\|_N^2 = \|\varphi^I\|_V^2$ n'est autre que la variance résiduelle de la régression. Si les variables sont entachées d'erreurs, la régression n'aura de sens que si Q est supérieur à la variance d'erreur, ce qui s'écrit en désignant par M la métrique d'erreurs :

$$Q = \|\varphi^I\|_V^2 \geq \|\varphi^I\|_M^2 \qquad (25)$$

Posant R = M - V, on retrouve la contrainte (15). Le vecteur $b_+^{I_1}$ minimisant Q sous (25) s'écrit donc d'après (18) où l'on a posé $\lambda = k/(1-k)$, et en partitionnant M de façon analogue à R et V :

$$b_+^{I_1} = (V_{11} + \lambda M_{11})^{-1} (V_{11} + \lambda M_{12}) \qquad (26)$$

Si l'erreur sur y est non corrélée avec les erreurs relatives aux variables explicatives, ce qui est en particulier le cas si l'on suppose les erreurs non corrélées, hypothèse que l'on fait généralement, $M_{12} = 0$, et l'on retrouve l'estimateur borné. Dans ce cas l'on peut montrer (cf [8] *) que la valeur de λ dans (26) minimisant Q sous la contrainte (25) est unique, cette valeur étant positive ou nulle.

Remarques :

1) Si $\varphi_\circ^I$ ne vérifie pas (25), la solution $\varphi_+^I = (-b_+^{I_1}, 1)$ se trouve sur la surface :

$$\|\varphi^I\|_V^2 = \|\varphi^I\|_M^2 \qquad (27)$$

ce qui traduit que la variance résiduelle de la régression est égale à la variance d'erreur.

Minimiser $Q = \|\varphi^I\|_V^2$ sous la contrainte (27) revient donc encore à minimiser $\|\varphi^I\|_M^2$ sous cette contrainte, i.e. à projeter (pour la métrique M) l'origine sur la surface définie par (27).

2) Si l'on a un modèle non centré, Q est le moment résiduel d'ordre 2 non centré de la régression, et il faudra remplacer la contrainte (25) par la contrainte :

$$Q - (\varphi^I \circ \bar{X}_I)^2 \geq \|\varphi^I\|_M^2 \qquad (28)$$

où $\bar{X}_I = (\bar{X}_{I_1}, \bar{y})$ désigne la moyenne de $X_I^J = (X_{I_1}^J, y^J)$.

Dans ce cas, si E désigne la matrice de terme général $\bar{X}_i \bar{X}_{i'}$ $(i,i' \in I)$, on a :

$$R = M + E - V \qquad (29)$$

et la solution de la régression est fournie par (26) où M est remplacé par M+E.

* *La démonstration est faite dans le cas où M_{11} est diagonale, cas auquel on peut toujours se ramener par une transformation linéaire sur $X_{I_1}^J$.*