P. CAZES

Méthode de régression. II. Critères bayésiens

Les cahiers de l'analyse des données, tome 3, nº 3 (1978), p. 257-268

<http://www.numdam.org/item?id=CAD_1978__3_3_257_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 3 - p. 257-268

MÉTHODE DE RÉGRESSION

II. — Critères bayésiens (à suivre)

[RÉGR. CONTR.]

par P. Cazes [1]

4 Régression biaisée et régression bayésienne : l'estimateur borné généralisé

4.1 Régression bayésienne

On suppose ici que y^J est gaussien et que conditionnellement à β^{I_1}, y^J a pour espérance mathématique $\beta^{I_1} \circ X^J_{I_1}$ et pour matrice variance $\sigma^2\ \Gamma_\circ$, où σ^2 est inconnu et où $\Gamma_\circ$ est une matrice définie positive connue à partir de laquelle est définie la métrique $N = \Gamma_\circ^{-1}$ de R^J (cf § 1, remarque 1). On suppose de plus que β^{I_1} suit une loi normale centrée de matrice variance $\Sigma\ \sigma^2$. Pour estimer β^{I_1}, nous allons chercher la loi *a posteriori* de β^{I_1}, i.e. la loi de β^{I_1} conditionnellement à y^J. Cette loi se met sous la forme

$$f(y^J \mid \beta^{I_1})\ f(\beta^{I_1})/g(y^J)$$

où $f(y^J \mid \beta^{I_1})$, $f(\beta^{I_1})$, $g(y^J)$ désignent respectivement les lois de y^J sachant β^{I_1}, de β^{I_1} et de y^J.

La loi précédente se mettant sous la forme :

$$(k/g(y^J))\ \exp\ (-\frac{A}{2})$$

où k est une constante, et A une fonction du second degré des composantes de β^{I_1}, on en déduit que conditionnellement à y^J, β^{I_1} suit une loi normale dont les caractéristiques, espérance mathématique b^{I_1} et matrice variance Σ_β, s'obtiennent en identifiant les termes du premier et du second degré en β^{I_1} de l'expression A avec les termes correspondants de l'expression

$$B = \Sigma_\beta^{-1}(\beta^{I_1} - b^{I_1},\ \beta^{I_1} - b^{I_1})\ .$$

L'expression A s'écrivant :

$$A = (1/\sigma^2)(\Gamma_\circ^{-1}(y^J - \beta^{I_1} \circ X^J_{I_1},\ y^J - \beta^{I_1} \circ X^J_{I_1}) + \Sigma^{-1}(\beta^{I_1},\ \beta^{I_1}))$$

soit encore puisque $N = \Gamma_\circ^{-1}$, et d'après la décomposition en blocs de V (cf § 1, formule (1)) :

$$A = (1/\sigma^2)(V_{11}(\beta^{I_1},\ \beta^{I_1}) - 2\ V_{21}\ \beta^{I_1} + V_{22} + \Sigma^{-1}(\beta^{I_1},\ \beta^{I_1}))\ ,$$

on en déduit que :

$$b^{I_1} = (V_{11} + \Sigma^{-1})^{-1}\ V_{12} \tag{30}$$

---

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

$$\Sigma_\beta = (V_{1i} + \Sigma^{-1})^{-1} \sigma^2 \qquad (31)$$

L'estimateur du maximum de vraisemblance de β^{I_1} connaissant y^J, i.e. l'estimateur *a posteriori* de β^{I_1} ou estimateur bayésien n'est rien d'autre que $E(\beta^{I_1}|y^J) = b^{I_1}$; il est donc donné par la formule (30). On retrouve bien, si on n'a aucune information *a priori* sur β^{I_1} ($V_{11} + \Sigma^{-1} = V_{11}$, β^{I_1} ayant une dispersion infinie) l'estimateur classique de la régression.

4.2 Orthogonalisation de la régression

On suppose R^{I_1} muni de la métrique M_{11}, dont la matrice associée (par rapport à la base canonique de R^{I_1}) notée également M_{11} peut se mettre sous la forme TT' du produit de deux matrices transposées l'une de l'autre *, l'application linéaire associée à T', et notée aussi T' étant une isométrie de R^{I_1} muni de la métrique M_{11} dans R^{I_1} muni de la métrique unité δ.

Si l'on effectue l'analyse en composantes principales de $X_{I_1}^J$ (au sens de la recherche du tenseur $XX_{I_1}^J$, de rang t donné, le plus proche de $X_{I_1}^J$), R^J étant toujours muni de la métrique N, on est ramené à diagonaliser $M_{11}^{-1} V_{11} = T'^{-1} T^{-1} V_{11}$, ou encore $H = T^{-1} V_{11} T'^{-1}$. On a alors la décomposition canonique classique :

$$H = T^{-1} V_{11} T'^{-1} = U D_\lambda U' \qquad (32)$$

où $U = (U_1^{I_1}, \ldots, U_p^{I_1})$ désigne la matrice orthogonale des vecteurs propres $U_i^{I_1}$ de H et D_λ la matrice diagonale des valeurs propres λ_i associées.

Posant :

$$\gamma^{I_1} = \beta^{I_1} \circ (T' U')_{I_1}^{I_1} \qquad (33)$$

$$Z_{I_1}^J = (U T'^{-1})_{I_1}^{I_1} \circ X_{I_1}^J \qquad (34)$$

on a :

$$\gamma^{I_1} \circ Z_{I_1}^J = \beta^{I_1} \circ X_{I_1}^J \quad ** \qquad (35)$$

L'intérêt de raisonner sur γ^{I_1} et $Z_{I_1}^J$ plutôt que sur β^{I_1} et $X_{I_1}^J$ réside dans le fait que les $\{Z_i^J \mid i \in I_1\}$ sont orthogonaux (pour N), le passage de γ^{I_1} à β^{I_1} se faisant très simplement par la formule (33).

* *De façon générale, nous désignerons par A' la transposée d'une matrice ou d'une application linéaire A.*

** *Matriciellement (ou dans le langage des applications), on a si X (resp. Z) désigne la matrice (p x card J) associée à* $X_{I_1}^J$ *(resp.* $Z_{I_1}^J$*) :* $\gamma^{I_1} = U' T' \beta^{I_1}$; $Z' = X' T'^{-1} U$; $Z' \gamma^{I_1} = X' \beta^{I_1}$, *le renversement de l'écriture dans les formules (33) à (35) étant dû au fait que l'on raisonne dans l'espace* R^{I_1}, *dual de* R_{I_1}, *i.e. dans l'espace des combinaisons linéaires des variables explicatives.*

Posons $y^J = z^J_{I_2}$, $z^J_I = (z^J_{I_1}, z^J_{I_2})$, et désignons par W la matrice des produits scalaires $\{<z^J_i , z^J_{i'}>_N \mid i, i' \in I\}$. Cette matrice se partitionne suivant I_1 et I_2 de la façon suivante :

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \begin{pmatrix} D_\lambda & W_{12} \\ W_{21} & V_{22} \end{pmatrix} \qquad (36)$$

W_{11} étant par construction la matrice diagonale D_λ , tandis que W_{22} est égal à V_{22} carré de la norme de y^J.

La solution (pour γ^{I_1}) $g_o^{I_1}$ de la régression usuelle s'écrit alors: $g_o^{I_1} = W_{11}^{-1} W_{12} = (D_\lambda)^{-1} W_{12}$ tandis que la matrice variance (dans le cas d'un modèle) de $g_o^{I_1}$ est égale à $\sigma^2 (D_\lambda)^{-1}$. Notons que $g_o^{I_1}$ est bien sûr liée à $b_o^{I_1}$ par la formule de transformation (33), et que l'on a, δ désignant toujours la métrique unité de R^{I_1} :

$$\|\gamma^{I_1} - g_o^{I_1}\|^2_\delta = \|\beta^{I_1} - b_o^{I_1}\|^2_{M_{11}} \qquad (37)$$

$$\|\gamma^{I_1} - g_o^{I_1}\|^2_{D\lambda} = \|\beta^{I_1} - b_o^{I_1}\|^2_{V_{11}} \qquad (38)$$

Sous l'hypothèse de normalité, les expressions figurant dans (38) suivent au facteur $(1/\sigma^2)$ près une loi de chi deux à card $I_1 = p$ degrés de liberté.

4.3 *Le critère de l'erreur quadratique moyenne*

Nous faisons ici l'hypothèse que l'on a un modèle, i.e. que les relations (7) du § 1 sont vérifiées. Compte tenu de (35) on a donc :

$$E(y^J) = \beta^{I_1} \circ X^J_{I_1} = \gamma^{I_1} \circ Z^J_{I_1} \qquad (39)$$

Soit b^{I_1} (resp. g^{I_1}) un estimateur de β^{I_1} (resp. γ^{I_1}), nous caractériserons la qualité de cet estimateur par l'erreur quadratique moyenne E (MSE en anglais : *Mean Square Error*) :

$$E = E(\|b^{I_1} - \beta^{I_1}\|^2_{M_{11}}) = E(\|g^{I_1} - \gamma^{I_1}\|^2_\delta) \qquad (40)$$

et l'on recherchera les estimateurs rendant minimum le critère précédent.

Si on se limite aux estimateurs sans biais de β^{I_1} (resp. γ^{I_1}), fonctions linéaires de y^J, $b_o^{I_1}$ (resp. $g_o^{I_1}$) minimise E qui vaut alors E_o :

$$E_o = \sigma^2 \, \Sigma\{1/\lambda_i \mid i \in I_1\} \qquad (41)$$

Si les variables explicatives sont très corrélées, certains des λ_i sont faibles, et E_o peut être très élevée ; c'est la raison pour laquelle on considère des estimateurs de β^{I_1} et γ^{I_1} biaisés, mais qui peuvent présenter une valeur du critère E plus faible que E_o.

4.4 *L'estimateur borné généralisé*

On suppose ici, comme au § 4.1, que conditionnellement à γ^{I_1} (ou β^{I_1}) y^J est gaussien, et que de plus γ^{I_1} suit une loi *a priori* gaussienne , centrée de matrice variable $(D_k)^{-1} \sigma^2$ où D_k est une matrice diagonale de $i^{\text{ème}}$ terme diagonal k_i.

L'estimateur bayésien $g_*^{I_1}$ de γ^{I_1}, qui se déduit de (30) où V_{11} est remplacé par $W_{11} = D_\lambda$, V_{12} par W_{12} et Σ^{-1} par D_k , s'écrit si D désigne la matrice diagonale de $i^{ème}$ terme diagonal $d_i = \lambda_i/(\lambda_i + k_i)$;

$$g_*^{I_1} = (D_\lambda + D_k)^{-1}\, W_{12} = (D_\lambda + D_k)^{-1}\, D_\lambda\, g_\circ^{I_1} = D\, g_\circ^{I_1} \qquad (42)$$

L'estimateur $b_*^{I_1} = g_*^{I_1} \circ (U\,T'^{-1})_{I_1}^{I_1}$ associé à $g_*^{I_1}$ s'écrit alors :

$$b_*^{I_1} = (V_{11} + T\,U\,D_k\,U'\,T')^{-1}\, V_{12} \qquad (43)$$

Les estimateurs $b_*^{I_1}$ et $g_*^{I_1}$ sont les estimateurs bornés généralisés qui ont été introduits dans la littérature dans le cas particulier où $M_{11} = \delta$.

En particularisant les coefficients k_i, on obtient un certain nombre d'estimateurs tels que l'estimateur borné usuel si $k_i = k$ (cf § 3.3.2); l'estimateur raccourci $b_*^{I_1} = b_\circ^{I_1}/(1+k)$ ou $g_*^{I_1} = g_\circ^{I_1}/(1+k)$, si $k_i = k\lambda_i$; l'estimateur de Marquardt si k_i est nul ou infini, ce dernier estimateur étant un estimateur sur composantes principales (cf § 2), où sont éliminés les facteurs associés à des k_i infinis, etc...

Nous ne ferons plus par la suite, sauf mention contraire, l'hypothèse de normalité. Nous supposerons simplement que l'on a un modèle, i. e. que les relations (7) et (39) sont vérifiées, et nous étudierons, en supposant les k_i connus, les propriétés de l'estimateur borné généralisé, et des estimateurs dérivés, en faisant abstraction de la manière dont nous avons introduit cet estimateur. Nous raisonnerons de préférence sur $g_*^{I_1}$ et γ^{I_1}, toutes les propriétés trouvées se transférant immédiatement à $b_*^{I_1}$ et β^{I_1} par la transformation (33).

$g_*^{I_1*}$ est un estimateur biaisé de γ^{I_1} (sauf si D est l'identité, i.e. si tous les k_i sont nuls, cas trivial puisqu'alors $g_*^{I_1} = g_\circ^{I_1}$) d'espérance mathématique $D\gamma^{I_1}$ et de matrice variance $D^2(D_\lambda)^{-1}\sigma^2 = (D_k+D_\lambda)^{-2} D_\lambda\, \sigma^2$. La valeur de l'erreur quadratique moyenne E pour $g_*^{I_1}$ s'écrit (cf [4])

$$E = E(\| g_*^{I_1} - \gamma^{I_1} \|_\delta^2) = \Sigma\{(\lambda_i\sigma^2 + k_i^2\,(\gamma^i)^2/(\lambda_i + k_i)^2 \mid i \in I_1\}. \quad (44)$$

E est minimum si $k_i = (k_i)_\circ = \sigma^2/(\gamma^i)^2$ la valeur minimale E_{min} de E s'écrivant :

$$E_{min} = \sigma^2\, \Sigma\{(\gamma^i)^2/(\sigma^2 + \lambda_i(\gamma^i)^2) \mid i \in I_1\}$$

Cette valeur est bien sûr plus petite que la valeur $E_\circ$ associée à $g_\circ^{I_1}$, puisque $E_\circ$ correspond au cas particulier où tous les k_i sont nuls.

D'un point de vue pratique, les $(k_i)_\circ$ ou ce qui est équivalent les $(d_i)_\circ = \lambda_i/(\lambda_i + (k_i)_\circ)$ ne sont pas connus puisqu'ils dépendent de γ^{I_1} et de σ^2 qui sont inconnus. On peut estimer ces coefficients en remplaçant γ^{I_1} et σ^2 par $g_\circ^{I_1}$ et $s^2 = \| y^J - g_\circ^{I_1} \circ z_{I_1}^J \|_N^2/(\mathrm{Card}\, J - p)$ (cette dernière quantité étant rappelons- le un estimateur sans biais de σ^2), puis le cas échéant réestimer ces coefficients en remplaçant γ^{I_1} par l'estimateur g^{I_1} ainsi obtenu, σ^2 étant toujours estimé par s^2 et réitérer le processus, le passage de l'itération t à l'itération t + 1, s'écrivant avec des notations évidentes, et si l'on opère sur les coefficients $(d_i)_\circ$:

$$d_i^{(t+1)} = \lambda_i/(\lambda_i + s^2/(g_{(t)}^i)^2) = \lambda_i/(\lambda_i + s^2/(d_i^{(t)}\, g_\circ^i)^2)$$

ce qui s'écrit encore :

$$d_i^{(t+1)} = (d_i^{(t)})^2/((d_i^{(t)})^2 + 1/F_i) \qquad (45)$$

où $F_i = \lambda_i (g_o^i)^2/s^2$ n'est rien d'autre que le carré t_i^2 du t de Student associé à g_o^i. Il est immédiat de voir que si F_i est plus petit que 4, $d_i^{(t)}$ tend vers zéro, ce qui revient à éliminer l'influence de la $i^{\text{ème}}$ composante principale z_i^J de $X_{I_1}^J$, puisque que le coefficient associé g_*^i est alors nul. Par contre, si F_i est plus grand ou égal à 4, comme la valeur initiale $d_i^{(o)}$ de d_i vaut 1 (on part de g_o^i), $d_i^{(t)}$ converge vers la valeur $1/2 + (1/4 - 1/F_i)^{1/2}$.

Les valeurs de d_i ainsi obtenues, et donc les valeurs associées de k_i étant aléatoires, l'erreur quadratique moyenne E de l'estimateur correspondant ne peut plus être calculée à partir de la formule (44), cette formule supposant les k_i non aléatoires, et l'on ne peut plus rien dire *a priori* des propriétés de cet estimateur. Mais l'on peut montrer (cf [2] et [31]) (en raisonnant sur les $g_o^i/\sqrt{\lambda_i}$ qui sont non corrélés et de même variance σ^2) qu'il existe des valeurs k_i (ou ce qui est équivalent des coefficients d_i) fonction de y^J, donc aléatoires, et telles que l'erreur quadratique moyenne E de l'estimateur associé, soit plus petite que la valeur E_o correspondant à $g_o^{I_1}$.

4.5 <u>*L'estimateur raccourci*</u> $g_R^{I_1} = c\, g_o^{I_1}$ $(0 \le c \le 1)$.

Cet estimateur est obtenu en posant dans (42) $k_i = k\lambda_i = (1-c)\lambda_i/c$. L'erreur quadratique moyenne s'écrit alors d'après (41) et (44) :

$$E_R = c^2\, E_o + (1-c)^2\, \alpha^2 \qquad (46)$$

où α^2 désigne le carré de la norme de γ^{I_1} pour δ (ou de β^{I_1} pour M_{11}).

Elle passe par un minimum pour $c = c_o = \alpha^2/(E_o + \alpha^2)$, valeur qui est comprise entre 0 et 1, la valeur minimale associée de E_R étant égale à $\alpha^2 E_o/(E_o + \alpha^2)$. Cette valeur est plus petite que la valeur E_o de E associé à $g_o^{I_1}$ (puisque $g_o^{I_1}$ correspond au cas où c = 1) et plus élevée que la valeur E_{min} associée à l'estimateur borné généralisé optimal puisque les valeurs $(k_i)_o/\lambda_i$ de k_i/λ_i associées à l'estimateur sont en général différentes. La valeur c_o de c minimisant E_R dépendant de γ^{I_1} (ou β^{I_1}) et σ^2 est inconnue. On peut, comme pour les $(k_i)_o$, l'estimer en remplaçant γ^{I_1} et σ^2 par leurs estimateurs des moindres carrés $g_o^{I_1}$ et s^2, puis le cas échéant réestimer c_o à partir de l'estimateur $g_R^{I_1}$ ainsi obtenu, et continuer le processus qui est analogue à (45) où d_i est remplacé par c, et F_i par $F = \hat{\alpha}^2/\hat{E}_o = \|g_o^{I_1}\|_\delta^2 / (s^2\, \Sigma\{1/\lambda_i \mid i \in I_1\})$, $\hat{\alpha}$ et $\hat{E}_o$ étant les valeurs de α et E_o quand on remplace σ^2 par s^2 et γ^{I_1} par $g_o^{I_1}$. Si donc F est plus petit que 4, la valeur limite de c est nulle, sinon, puisque la valeur initiale de c vaut 1 (on part de $g_o^{I_1}$), la valeur limite de c vaut $1/2 + (1/4 - 1/F)^{1/2}$. Notons que si $M_{11} = V_{11}$, auquel cas D_λ est la métrique unité δ, $F = (n-p)\, R^2/(p(1-R^2))$, R désignant le coefficient de corrélation multiple de y^J par rapport à $z_1^J, \dots, z_p^J$ (ou ce qui est équivalent $x_1^J, \dots, x_p^J$) ; dans ce cas, F est le F de Fisher-Snédécor associé à la régression.

La valeur de c_o ainsi obtenue étant aléatoire, on ne peut rien dire des propriétés de l'estimateur raccourci ainsi obtenu, et la formule (46) n'est plus valable.

Par contre, si D_λ est la matrice unité, i.e. si l'on prend comme métrique M_{11} dans R^{I_1} la métrique $M_{11} = V_{11}$, et si l'on pose $G = R^2/(1 - R^2)$ on peut montrer (cf [31]) que si $p \geq 3$ et si l'on choisit $c = (1 - a(G)/G)$ où $a(G)$ est une fonction monotone non décroissante de G telle que $0 \leq a(G) \leq 2(p-2)/(\mathrm{Card}\, J - p + 2)$ alors $c\, g_o^{I_1}$ a une erreur quadratique moyenne plus petite que $g_o^{I_1}$. Si $a(G)$ est la fonction constante, on obtient l'estimateur de James et Stein et on peut montrer (cf [23]) que E_R est minimum si $a = (p-2)/(\mathrm{Card}\, J - p + 2)$. Si l'on choisit $a(G) = a$ si $G > a$, et $a(G) = G$ si $G \leq a$ on obtient l'estimateur $(1 - a/G)^+\, g_o^{I_1}$ * dont l'erreur quadratique moyenne est plus petite que E_o pour $0 \leq a \leq 2(p-2)/(\mathrm{Card}\, J - p + 2)$, estimateur qui est également meilleur au sens de E_R que l'estimateur $(1 - \frac{a}{G})\, g_o^{I_1}$.

Remarques

1) Toutes les propriétés précédentes, se transposent immédiatement à l'estimateur raccourci $c\, b_o^{I_1}$, puisque F est invariant dans le passage de γ^{I_1}, g^{I_1} à β^{I_1}, $g_o^{I_1}$.

2) L'estimateur

$$g^{I_1} = [D_\lambda(\gamma^{I_1}, g_o^{I_1})/(\sigma^2 + D_\lambda(\gamma^{I_1}, \gamma^{I_1})]\gamma^{I_1} \tag{47}$$

réalise (cf [18]) dans la classe des fonctions linéaires en y^J (g^{I_1} est fonction linéaire de y^J par l'intermédiaire de $g_o^{I_1}$) le minimum de l'erreur quadratique moyenne tensorielle $E((g^{I_1} - \gamma^{I_1}) \otimes (g^{I_1} - \gamma^{I_1}))$**. γ^{I_1} et σ^2 étant inconnus, on peut rechercher un estimateur g^{I_1} minimisant $\| y^J - g^{I_1} \circ Z_{I_1}^J \|_N^2$ et vérifiant (47) où γ^{I_1} a été remplacé par g^{I_1}, σ^2 par s^2 ou par $\| y^J - g^{I_1} \circ Z_{I_1}^J \|^2/(\mathrm{Card}\, J - p)$. On obtient dans les deux cas un estimateur raccourci $c\, g_o^{I_1}$ dont le paramètre c est aléatoire, et dont les propriétés ont été étudiées par simulation sur plusieurs exemples par Vinod (cf [33]). Vinod a également étudié l'estimateur de James et Stein ainsi que l'estimateur raccourci proposé par Farebrother (cf [18]) et obtenu en remplaçant dans (47) γ^{I_1} par $g_o^{I_1}$ et σ^2 par s^2.

Notons que de la même façon, l'estimateur b^{I_1} obtenu en remplaçant dans (47) γ^{I_1} par β^{I_1}, $g_o^{I_1}$ par $b_o^{I_1}$ et D_λ par V_{11} minimise dans la classe des estimateurs de β^{I_1} fonction linéaire de y^J, $E((b^{I_1} - \beta^{I_1}) \otimes (b^{I_1} - \beta^{I_1}))$

3) Au lieu de raccourcir toutes les composantes de $g_o^{I_1}$, on peut simplement raccourcir $q\,(q < p)$ des composantes de $g_o^{I_1}$, les q dernières par exemple, ce qui revient si I_1'' désigne l'ensemble des composantes raccourcies et $I_1' = I_1 - I_1''$ à considérer l'estimateur

$$g^{I_1} = (g_o^{I'_1},\ c\, g_o^{I''_1})$$

* $(1 - \frac{a}{G})^+$ *désigne la partie positive de* $1 - \frac{a}{G}$; *égale à* $1 - \frac{a}{G}$ *si* $1 - \frac{a}{G}$ *est positif ou nul (i.e. si* $G \geq a$*), et à zéro sinon.*

** *En ce sens que* $\forall\, u_{I_1} \in R_{I_1}$, $\forall\, h^{I_1} = y^J \circ A_J^{I_1}$:

$$[(h^{I_1} - \gamma^{I_1}) \otimes (h^{I_1} - \gamma^{I_1}) - (g^{I_1} - \gamma^{I_1}) \otimes (g^{I_1} - \gamma^{I_1})] \circ (u_{I_1} \otimes u_{I_1}) \geq 0$$

L'erreur quadratique moyenne associée s'écrivant :

$$E = \sigma^2\ \Sigma\{1/\lambda_i \mid i \in I'_1\} + c^2\ \sigma^2\ \Sigma\{1/\lambda_i \mid i \in I''_1\} + (1-c)^2 \Sigma\{\gamma_i^2 \mid i \in I''_1\}$$

soit encore : $E = E'_o + c^2\ E''_o + (1-c)^2\ \alpha''^2$

où E'_o (resp. E''_o) désigne l'erreur quadratique moyenne associée à $g_o^{I'1}$ (resp. $g_o^{I''1}$) et α''^2 le carré de la norme de $g^{I''1}$ pour la métrique usuelle de $R^{I''1^o}$, on a des résultats analogues, en ce qui concerne la valeur optimale de c, et l'estimation de cette valeur, à ceux donnés dans le cas de l'estimateur raccourci : il suffit de raisonner sur les q composantes de I''_1. En particulier si $q \geq 3$, si D''_λ* est la métrique unité et si $G'' = R''^2/(1-R^2)$ où R" désigne le coefficient de corrélation multiple de y^J par rapport aux variables $\{z_i^J \mid i \in I''_1\}$, $g^{I_1} = (g_o^{I'1}, (1 - a(G'')/G'')\ g_o^{I''1})$ a une valeur de E plus petite que la valeur E_o associée à $g_o^{I_1}$, pourvu que a(.) soit une fonction monotone non décroissante, comprise entre 0 et $2(q-2)/(\mathrm{Card} J-p+2)$. Si a(G") est une constante a, la valeur minimale de E est obtenue pour $a = (q-2)/(\mathrm{Card}\ J - p + 2)$. On peut encore noter que l'estimateur associé à $c = (1-a/G'')^+$, où a est une constante comprise entre 0 et $2(q-2)/(\mathrm{Card}\ J-p+2)$, estimateur qui revient à annuler toutes les composantes de I''_1 si G" est plus petit que a, est meilleur au sens de E que l'estimateur associé à $c = 1 - a/G''$.

4.6 *L'estimateur borné* $b_B^{I_1} = (V_{11} + k\,M_{11})^{-1} V_{12}$

Cet estimateur déjà rencontré au § 3.3 dans le cas de la régression sous contraintes quadratiques, avec un paramètre k aléatoire, puisque dépendant de y^J, alors qu'ici k est supposé fixé, s'obtient en remplaçant dans (43) k_i par k (i.e. en remplaçant D_k par $k\delta$, δ désignant toujours la matrice unité). Remplaçant k_i par k dans (44), il est immédiat de voir qu'au voisinage de k = 0 (i.e. de $b_o^{I_1}$), E est une fonction décroissante de k et que pour k positif, il existe une valeur (et une seule), k_o de k rendant ce critère minimum, cette valeur étant la racine positive de l'équation $dE/dk = 2\Sigma\{\lambda_i (k(\gamma^i)^2 - \sigma^2)/(\lambda_i + k)^3 \mid i \in I_1\} = 0$. k_o étant assez difficile à calculer, on adopte souvent pour k la valeur $k_1 = p\sigma^2/\|\beta^{I_1}\|^2_{M_{11}}$, k_1 minimisant E si $M_{11} = V_{11}$, auquel cas on a l'estimateur raccourci.

D'un point de vue pratique, de nombreuses solutions ont été proposées dans la littérature pour estimer k_o ou k_1, auquel cas, (44) n'est plus valable, k étant aléatoire. On peut par exemple pour estimer k_1 utiliser la valeur $p\,s^2/\|b_o^{I_1}\|^2_{M_{11}}$ préconisée par Farebrother (cf [18]) et obtenue en remplaçant dans l'expression de k_1, β^{I_1} par $b_o^{I_1}$ et σ^2 par s^2, ou encore utiliser un processus itératif de la forme : $k_{t+1} = p\,s^2/\|b_t^{I_1}\|^2$ où $b_t^{I_1}$ désigne l'estimateur associé à la valeur k_t de k à l'itération t, et k_{t+1} la valeur de k à l'itération t + 1. On peut aussi utiliser la valeur $p\,s^2/(\|b_o^{I_1}\|^2_{M_{11}} - p\,s^2)$ préconisée par Mallows (cf [29]) les espérances mathématiques du numérateur et du dénominateur de l'expression précédente étant, dans le cas où $M_{11} = V_{11}$, respectivement égales à $p\,\sigma^2$ et $\|\beta^{I_1}\|^2_{M_{11}}$.

On peut pour estimer k_o rechercher un estimateur $\hat{E}$ de l'erreur quadratique moyenne E, et rechercher empiriquement pour quelle valeur de k, $\hat{E}$ passe par un minimum. Pour obtenir $\hat{E}$, il suffit dans (44) où l'on fait $k_i = k$ de

* *D''_λ étant la restriction de D_λ à $R^{I''1}$.*

remplacer σ^2 par s^2, et $(\gamma^i)^2$ soit par $(g_o^i)^2$ soit par $(g_o^i)^2 - s^2/\lambda_i$. Dans ce dernier cas, on a un estimateur sans biais, mais qui peut ne pas présenter de minimum pour k positif alors que dans le premier cas, il existe une seule valeur positive de k rendant minimum $\hat{E}$.

On peut aussi (cf [16]) choisir k de façon à ce que $\Sigma\{(g_o^i)^2/((k^{-1} + \lambda_i^{-1})s^2) \mid i \in I_1\}$ soit égal à p, $\Sigma\{(g_o^i)^2/((k^{-1} + \lambda_i^{-1})\sigma^2) \mid i \in I_1\}$ ayant une espérance mathématique égale à p, sous les hypothèses bayésiennes et de normalité faites au début du § 4.4, la loi marginale de $g_o^{I_1}$ étant une loi normale centrée de matrice variance $(D_k^{-1} + D_\lambda^{-1})\sigma^2 = (k^{-1}\delta + D_\lambda^{-1})\sigma^2$, δ étant la matrice unité d'ordre p. Notons que cette façon de faire vaut également pour l'estimateur raccourci (cf § 4.5). si l'on remplace $k_i = k$ par $k_i = k\lambda_i = ((1-c)/c)\lambda_i$, k et donc c étant alors déterminés de telle sorte que $k\Sigma\{\lambda_i(g_o^i)^2/((1+k)s^2) \mid i \in I_1\}$ soit égal à p.

Une autre procédure (la *ridge trace*) préconisée par Hoerl et Kennard (cf [22]) pour estimer k est de visualiser les coefficients b_B^i de $b_B^{I_1}$ en fonction de k, et de retenir la valeur minimale positive de k pour laquelle il y a stabilisation de tous les coefficients b_B^i. Cette dernière procédure semble particulièrement intéressante d'un point de vue pratique, puisqu'elle permet de voir l'évolution de $b_B^{I_1}$ avec k.

Signalons pour terminer une méthode suggérée par Vinod (cf [34]) pour déterminer une valeur de k , non aléatoire. Posant $r_i = \lambda_i/(\lambda_i + k)^2$, on recherche la valeur de k minimisant l'indice

$$\Sigma\{(r_i/\bar{r} - 1)^2 \mid i \in I_1\} = \Sigma\{(r_i/\bar{r})^2 \mid i \in I_1\} - p$$

où $\bar{r}$ désigne la moyenne empirique des r_i.

Cet indice est nul si tous les λ_i sont égaux auquel cas $V_{11} = M_{11}$. Si de plus M_{11} est la métrique unité, le système initial est orthogonal.

Notons qu'en posant $m = p - \Sigma\{d_i \mid i = 1, p\} = p - \Sigma\{\lambda_i/(\lambda_i + k_i) \mid i = 1,p\}$, on peut dans le cas de l'estimateur borné généralisé $b_*^{I_1}$, effectuer une visualisation des coefficients b_*^i en fonction de m, et rechercher le minimum de l'indice précédent en fonction de m, m pouvant s'interpréter comme la diminution du rang du tableau $X_{I_1}^J$.

4.7 *L'estimateur de Marquardt et les estimateurs dérivés*

On suppose ici que l'on a une partition de I_1 en deux sous ensembles I'_1 et I''_1, et que k_i est nul pour tout i de I'_1, tandis que k_i est infini pour tout i de I''_1, ce qui revient, si l'on désigne par $g_M^{I_1}$ l'estimateur associé de γ^{I_1}, à poser $g_M^i = g_o^i$ pour $i \in I'_1$ et $g_M^i = 0$ pour $i \in I''_1$. Cet estimateur n'est rien d'autre que l'estimateur sur composantes principales (cf § 2) obtenu en ne gardant que les composantes principales associées à I'_1 du tableau $X_{I_1}^J$. Si I'_1 correspond aux composantes principales associées aux plus fortes valeurs propres, et M_{11} est la métrique unité ($M_{11} = \delta$), on obtient l'estimateur proposé par Marquardt (cf [30]).

Si E_M désigne l'erreur quadratique moyenne associée à $g_M^{I_1}$ l'on déduit de (41) et (44) que :

$$E_o - E_M = \Sigma\{(\sigma^2/\lambda_i - (\gamma^i)^2) \,|\, i \in I''_1\} \qquad (48)$$

$g_M^{I_1}$ sera donc meilleur que $g_o^{I_1}$ au sens de E, si cette quantité est positive ; une condition suffisante pour qu'il en soit ainsi est d'éliminer toutes les composantes telles que $\tau_i^2 = \lambda_i(\gamma^i)^2/\sigma^2$ soit plus petit que 1. Notons que si l'on estime τ_i^2 par $t_i^2 = \lambda_i(g_o^i)^2/s^2$, t_i^2 n'est rien d'autre que le carré du t de Student associé à γ^i, quand on teste l'hypothèse $\gamma^i = 0, \tau_i^2$ étant le paramètre de décentrement associé à t_i^2.

D'un point de vue pratique, après avoir éliminé les composantes principales associées aux plus faibles valeurs propres (cf § 2), on ne gardera parmi les composantes restantes que les composantes liées à y^J i.e. ayant des t_i^2 assez élevées, ce qui revient à tester l'hypothèse que $\gamma^i = 0$, ou l'hypothèse moins forte que τ_i^2 n'est pas trop élevé, par exemple qu'il est plus petit que 1. Les coefficients k_i nuls ou infinis étant alors aléatoires, les formules (44) et (48) ne sont plus valables.

Remarques :

1) L'estimateur $b_M^{I_1}$ de β^{I_1} associé à $g_M^{I_1}$, qui est l'estimateur que l'on considère en pratique, si l'on veut raisonner sur les variables initiales s'écrit si $U = (U_1, U_2)$ désigne la partition de U associée à I'_1 et I''_1, U_1 correspondant aux vecteurs propres gardés, et U_2 aux vecteurs propres éliminés :

$$b_M^{I_1} = b_o^{I_1} \circ (T'U'_1\ U_1 T'^{-1})_{I_1}^{I_1} = b_o^{I_1} - b_o^{I_1} \circ (T'U'_2\ U_2 T'^{-1})_{I_1}^{I_1}$$

2) Supposons que $I'_1 =]r]$; Marquardt a également proposé l'estimateur obtenu en posant : $k_i = 0$ si $i \in I'_1$; $k_{r+1} = k\lambda_{r+1}$; $k_i = \infty$ si $i \in I''_1 - \{r+1\}$; cet estimateur peut être appelé l'estimateur fractionnaire de Marquardt. Le coefficient k, ou ce qui est équivalent le coefficient $d_{r+1} = 1/(1+k')$ minimisant E est tel que $d_{r+1} = [(1/\tau_{r+1})^2 + 1]^{-1}$ (il suffit pour obtenir cette valeur de considérer qu'on raccourcit un vecteur à une dimension g_o^{r+1}, et d'appliquer la formule donnée au § 4.5 pour le paramètre de raccourcissement c_o).

On peut estimer cette valeur de d_{r+1} par $[(1/t_{r+1})^2 + 1]^{-1}$ expression obtenue en remplaçant σ^2 par s^2 et γ^{r+1} par g_o^{r+1} dans $\tau_{r+1}^2 = \lambda_{r+1}(g_o^{r+1})^2/\sigma^2$, et le cas échéant réitérer la procédure d'estimation de d_{r+1}, procédure qui converge vers 0 (k_{r+1} infini) si t_{r+1}^2 est plus petit que 4, et vers $1/2 + (1/4 - (1/t_{r+1})^2)^{1/2}$ si t_{r+1}^2 est plus grand ou égal à 4.

3) Au lieu de considérer les composantes principales associées au tableau $X_{I_1}^J$ des variables explicatives, Webster, Gunst et Mason (cf [35]) considèrent les composantes principales du tableau X_I^J de toutes les variables, ce qui revient si R^I est muni de sa métrique canonique à diagonaliser V, dont la décomposition canonique s'écrit :

$$V = W D_\mu W'$$

où D_μ désigne la matrice diagonale des valeurs propres μ_i de V, et $W = \{W_i^I \,|\, i \in I\}$ la matrice des vecteurs propres correspondants.

Adoptant comme base de R^I la base formée par les W_i^I, et désignant par a^i la coordonnée de la combinaison linéaire φ^I sur $W^I_i (1 \le i \le p+1)$, on a :

$$y^J - b^{I_1} \circ X^J_{I_1} = \varphi^I \circ X^J_I = a^I \circ W^I_I \circ X^J_I$$
$$= a^I \circ W^{p+1}_I \, y^J + a^I \circ W^{I_1}_I \circ X^J_{I_1}$$

D'où l'on déduit :

$$\left.\begin{aligned} a^I \circ W^{p+1}_I &= 1 \\ b^{I_1} &= -a^I \circ W^{I_1}_I \end{aligned}\right\} \quad (49)$$

S'il existe une valeur propre nulle, μ_{p+1} par exemple, les $\{X^J_i | i \in I_1\}$ étant supposés linéairement indépendants , y^J est dans le sous espace engendré par ces vecteurs ; le coefficient W^{p+1}_{p+1} est alors différent de zéro, et la solution de la régression est : $\varphi_\circ{}^I = (1/W^{p+1}_{p+1}) \, W^I_{p+1}$, soit $b^{I_1}_\circ = -(1/W^{p+1}_{p+1}) \, W^{I_1}_{p+1}$.

Si toutes les valeurs propres μ_i sont différentes de 0, la somme des moindres carrés $\| y^J - b^{I_1} \circ X^J_{I_1} \|^2_N = \| \varphi^I \|^2_V = \Sigma\{\mu_i (a^i)^2 | i \in I\}$ sera minimum sous la contrainte $\Sigma\{a^i \, W^{p+1}_i \mid i \in I\} = 1$ si :

$$a^i = (W^i_{p+1} / \mu_i) / \Sigma\{(W^i_{p+1})^2 / \mu_i | i \in I\} \quad (50)$$

Si on élimine les composantes i associées simultanément à une valeur propre μ_i petite, et à un coefficient W^{p+1}_i petit, i.e. les composantes traduisant les relations entre les $\{X^J_i \, | i \in I_1\}$, mais pas entre y^J et les $\{X^J_i \, | i \in I_1\}$, on obtient en désignant par I' le sous ensemble de I associé aux composantes gardées un estimateur $b^{I_1}_W$ donné par les formules (49) et (50) où I est remplacé par I'.

4) L'estimateur de Marquardt et l'estimateur étudié dans la remarque précédente peuvent être considérés comme des estimateurs sous contrainte d'égalité puisqu'ils reviennent à projeter y^J sur un sous espace de l'espace engendré par les $\{X^J_i \, | i \in I_1\}$.

Pour le premier estimateur, on impose la contrainte $\beta^{I_1} \circ (T'U'_2)^{I''_1}_{I_1} = 0$, tandis que pour le second, on a, en posant $I'' = I - I'$, $W_2 = \{W^I_i \, | i \in I''\} = W^I_{I''}$, $\varphi^I = (-\beta^{I_1}, 1)$: $\varphi^I \circ (W'_2)^{I''}_I = 0$.

4.8 Régression sous contraintes et régression biaisée

Tous les estimateurs biaisés rencontrés dans ce § 4 : estimateur borné généralisé et estimateurs dérivés, peuvent être considérés comme des estimateurs sous contraintes, avec une différence importante néanmoins : quand on obtient ces estimateurs en imposant des contraintes, les paramètres k_i sont aléatoires, alors qu'*a priori*, pour l'estimateur borné généralisé, et les estimateurs dérivés, les k_i sont connus. Ainsi l'estimateur borné généralisé peut être obtenu en imposant à chaque composante $\gamma^i = b^{I_1} \circ (T'U')^i_{I_1}$ de γ^{I_1} d'être comprise entre deux valeurs l_i et L_i fixées. Si $g^{I_1}_+$ désigne l'estimateur sous contraintes associé, on a, puisque ces contraintes (vu l'orthogonalité du modèle associé à γ^{I_1}) sont indépendantes : $g^i_+ = l_i$ si $g^i_\circ < l_i$; $g^i_+ = g^i_\circ$ $(k_i = 0)$ si $l_i \le g^i_\circ \le L_i$; $g^i_+ = L_i$ si $g^i_\circ > L_i$; notons qu'ici la valeur de k_i qui se déduit de l'équation $(\lambda_i/(\lambda_i + k_i)) \, g^i_\circ = g^i_+$ peut être négative si l_i et L_i sont de même signe, contrairement à ce qui se passe pour l'estimateur borné généralisé.

On a également vu que l'estimateur borné était obtenu en imposant la contrainte que la norme du vecteur de régression (pour la métrique M_{11}), ne soit pas trop élevée, et que comme cas particulier, on obtenait pour $M_{11} = V_{11}$ l'estimateur raccourci. De même, si l'on ne raccourcit (d'une façon égale) qu'une partie I''_1 des composantes de $g_o^{I_1}$, (cf § 4.5, remarque 3), cela revient à borner la norme de $g^{I''_1}$ pour la métrique D''_λ induite par D_λ sur $R^{I''_1}$, ce qui revient à borner la norme de b^{I_1} pour une certaine pseudo-métrique (cf [11]). L'estimateur fractionnaire de Marquardt enfin (cf § 4.7, remarque 2) peut être obtenu en projetant y^J sur le sous espace vectoriel correspondant aux composantes principales gardées (y compris la composante raccourcie) et en imposant à la valeur de la composante raccourcie d'être bornée en valeur absolue.

Signalons que si l'on projette (R^{I_1} étant muni de la métrique V_{11}) $b_o^{I_1}$ sur un convexe fermé C contenant β^{I_1} (i.e. si on impose la contrainte, $b^{I_1} \in C$), il découle immédiatement des propriétés de la projection sur un convexe que l'estimateur $b_+^{I_1}$ correspondant a une erreur quadratique moyenne (mesurée avec la métrique $M_{11} = V_{11}$, auquel cas $D_\lambda = \delta$) plus petite que celle associée à $b_o^{I_1}$. De même, si l'on impose des contraintes d'égalité, comme c'est le cas pour l'estimateur de Marquardt, ou l'estimateur de Webster, Gunst et Mason, même si β^{I_1} ne satisfait pas ces contraintes, on peut obtenir un estimateur $b_+^{I_1}$ meilleur au sens de l'erreur quadratique moyenne que $b_o^{I_1}$. Une condition suffisante pour qu'il en soit ainsi, et ce quelle que soit la métrique M_{11} de R^{I_1}, est que le paramètre de décentrement associé au test de l'hypothèse que β^{I_1} vérifie les contraintes d'égalité, soit plus petit que 1 (cf [4] et [32]).

4.9 *Conclusion*

Que penser des différents estimateurs biaisés présentés dans ce § 4 :

D'un point de vue pratique, l'estimateur raccourci, bien qu'ayant donné lieu à de très nombreux travaux, nous semble d'un intérêt restreint, puisqu'en fait, il ne modifie pas les rapports entre les différentes composantes de l'estimateur classique $b_o^{I_1}$ de β^{I_1}, alors que ces rapports peuvent ne pas avoir de sens si les variables explicatives sont très corrélées, et que l'on a des coefficients de régression peu stables car très sensibles aux fluctuations d'échantillonnage. L'estimateur borné généralisé séduisant dans son principe a le désavantage d'introduire p valeurs k_i à estimer. Finalement ce sont l'estimateur borné et l'estimateur de Marquardt qui d'un point de vue pratique semblent les plus intéressants et les plus utilisés. Ces deux estimateurs protègent bien la régression, comme on l'a déjà dit plus haut (cf § 2) dans le cas ou l'on a des variables explicatives très corrélées. Notons que l'estimateur borné est d'autant plus intéressant à utiliser que l'on a une connaissance *a priori* de la régression. Par exemple, si pour avoir un sens, les coefficients de régression doivent être compris entre deux limites, il peut être plus intéressant d'utiliser l'estimateur borné, plutôt que d'imposer les contraintes précédentes ; c'est en particulier le cas (cf[1]) si l'estimateur des moindres carrés ne vérifie pas ces contraintes et s'il existe une valeur positive de k (on choisira *a priori* la plus petite) telle que l'estimateur borné vérifie ces contraintes. On évite ainsi d'avoir plusieurs coefficients de régression atteignant leur valeur limite, ce qui dans certains cas peut être difficile à interpréter.

En ce qui concerne l'estimateur de Marquardt, nous préférons l'utiliser, plutôt que l'estimateur de Webster, Gunst et Mason, ce dernier revenant à faire l'analyse factorielle du tableau de toutes les variables, alors que le premier revient à faire l'analyse factorielle du tableau des variables explicatives (i.e. étudier la structure des variables explicatives) et à projeter y^J en supplémentaire sur les axes factoriels, ce qui permet de visualiser la liaison de y^J avec ces variables.

Si l'on n'est pas dans le cadre d'un modèle, et si l'on a un échantillon d'effectif assez élevé (Card J > 100), on a souvent intérêt, plutôt que d'utiliser la régression classique et les techniques de régression développées dans ce § 4, à faire la régression par l'analyse des correspondances et la régression par boule, qui sont étudiées au § suivant.

Remarque :

De nombreuses simulations ont été effectuées pour étudier les estimateurs biaisés de β^{I_1}. Nous nous contenterons de renvoyer le lecteur à [16] où 57 estimateurs de β^{I_1} sont comparés, et à [18 bis] où l'on trouve dans la bibliographie un certain nombre d'articles relatifs à ces simulations.