LES CAHIERS DE L'ANALYSE DES DONNÉES

P. CAZES

Méthodes de régression. III. L'analyse des données

Les cahiers de l'analyse des données, tome 3, nº 4 (1978), p. 385-391

<http://www.numdam.org/item?id=CAD_1978__3_4_385_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 4 - p. 385-391

MÉTHODES DE RÉGRESSION

III. – L'Analyse des données

[RÉGR. ANAL.]

par P. Cazes [1]

5 *Régression par l'analyse des données*

Ayant déjà étudié la régression sur composantes principales (cf § 2 et 4.7) nous ne parlerons ici que de régression par boule et par l'analyse des correspondances.

5.1 *Rappels sur la régression par boule.* (cf [3], [25])

Pour expliquer une variable y^J en fonction des variables $\{X_i^J | i \in I_1\}$, on peut considérer que toute observation j de J est un point de l'espace R_{I_1}, point que l'on peut affecter de la valeur y^j de y.

Supposant R_{I_1} euclidien, i.e. muni d'une métrique, pour prévoir la valeur y^s de y pour un point supplémentaire s, dont on connait les valeurs des variables explicatives, on recherche dans R_{I_1} les r points de J les plus proches de s. La moyenne de l'écart type des valeurs de y associées à ces r points permet d'estimer y^s, ainsi que la précision de l'estimation y y^s ainsi obtenue. L'avantage de cette façon d'opérer fort utilisée en météorologie, et qui ne peut s'appliquer que si l'on a un échantillon d'effectif suffisant est de fournir pour chaque estimation y y^s de y^s une précision (i.e. un écart type) non uniforme, fonction de s, contrairement à la régression usuelle(**), où l'on a une précision de reconstitution uniforme.

(*) *Suite de l'article paru sous le même titre dans les Cahiers, Vol. III n° 2 et n° 3.*

(1) *Maître-Assistant, I.S.U.P., Laboratoire de Statistique. Université Pierre & Marie Curie ; Paris.*

(**) *On suppose ici que y, $\{X_i | i \in I_1\}$ sont des variables aléatoires et que l'on a un échantillon J de ces variables (cf § 1 remarque 2) ; et non pas que l'on a un modèle, car dans ce cas la précision de la reconstitution dépend des valeurs $\{X_i^s | i \in I_1\}$ de X_i pour s, valeurs qui sont supposées connues et non aléatoires.*

Pour caractériser la qualité de la régression on peut calculer la corrélation sur J muni de la mesure uniforme entre y^J et son approximation yy^J. Notons que contrairement au cas de la régression usuelle, cette corrélation peut décroître si on augmente le nombre de variables explicatives.

Comme métrique de R_{I_1}, on emploie en général la métrique usuelle, soit que l'on raisonne dans l'espace des variables explicatives initiales, ces dernières ayant été centrées et réduites, soit que l'on raisonne dans l'espace des premiers axes factoriels d'une analyse factorielle (analyse en composantes principales, ou analyse des correspondances) effectuée sur le tableau des variables explicatives (cf §§ 2, 4.7, 5.2.1) ou sur le tableau croisant (après division en classes de toutes les variables) les modalités de la (ou les) variable à expliquer avec l'ensemble des modalités de toutes les variables explicatives (cf § 5.2.2).

Signalons enfin pour terminer que la régression par boule s'applique aussi si y est qualitative auquel cas on parle aussi de discrimination par boule ou par voisinage ; il suffit de calculer parmi les voisins d'un individu supplémentaire s, le pourcentage des voisins appartenant à chaque classe de y et d'affecter s à la classe correspondant au pourcentage le plus important, l'ensemble des pourcentages calculés permettant de juger de la qualité de l'affectation de s. Pour se rendre compte de la qualité globale de la discrimination, il suffit de calculer le taux d'erreur obtenu sur J en affectant de cette façon chaque observation j de J.

Notons que dans ce dernier cas (y qualitatif), on peut opérer de manière itérative pour essayer de diminuer le taux d'erreurs (cf[31bis]): au pas 1 on fait la discrimination par boule sur y (i.e. la discrimination par boule usuelle), d'où pour chaque observation j de J une estimation $yy^j_{(1)}$ de y^j, $yy^j_{(1)}$ étant une des classes de y, égale à y^j s'il n'y a pas d'erreur d'affectation, différente sinon ; au pas 2, on fait la discrimination par boule non plus sur y mais sur les valeurs $yy_{(1)}$ de y estimées au pas 1 ; et de façon générale, au pas k, on fait la discrimination par boule en prenant comme variable à expliquer l'estimation $yy_{(k-1)}$ de y obtenue à l'étape k - 1.

Si le processus se stabilise (i.e. si le nombre de différences entre $yy^j_{(k)}$ et $yy^j_{(k-1)}$ $(1 \leq j \leq n)$ tend vers zéro) et si le nombre d'erreurs effectuées en affectant l'observation j à la classe $yy^j_{(k)}$ de y (si on arrête le processus à l'itération k) est plus faible qu'en affectant j à $yy^j_{(1)}$ résultat de la discrimination par boule usuelle, on aura intérêt pour prévoir la classe d'un individu supplémentaire s à remplacer dans l'échantillon initial y^J par $yy^J_{(k)}$, ce qui revient, dans une certaine mesure à effectuer un "lissage" de y^J.

5.2 *Régression par l'analyse des correspondances*

On étudiera ici deux méthodes de régression par l'analyse des correspondances ; dans la première on raisonne sur le tableau initial X^J_I, tandis que dans la seconde, on construit le tableau croisant (après un découpage préalable en classes de toutes les variables) l'ensemble des modalités de la variable à expliquer avec l'ensemble des modalités de toutes les variables explicatives.

5.2.1 *Régression après l'analyse des correspondances du tableau des variables explicatives*

Dans cette méthode déjà évoquée (cf § 2), l'on effectue l'analyse des correspondances du tableau $X_{I_1}^{J}$ des variables explicatives, et l'on projette la variable à expliquer en supplémentaire sur les premiers axes factoriels de cette analyse ; l'on peut ensuite le cas échéant exprimer l'approximation ainsi obtenue de y en fonction des variables initiales, ceci étant en particulier intéressant quand on a un modèle et que les coefficients de régression sont interprétables. On obtient ainsi l'équivalent de la régression sur variables orthogonales et de l'estimateur de Marquardt (cf § 4.7) mais en effectuant l'analyse des correspondances de $X_{I_1}^{J}$ et non l'analyse en composantes principales, l'analyse des correspondances s'imposant dans certains cas, par exemple si y^J est une loi de probabilité, mélange en proportions inconnues, d'un certain nombre de lois de probabilité connues. En particulier si $y^j = \int_a^b (x(u))^j g(u)du$ où y^J et $(x(u))^J$ sont des lois de probabilité sur J, et si l'on désire estimer la fonction g(u), on découpera l'intervalle (a,b) en 50 ou 100 intervalles et l'on exprimera y^j en fonction d'une combinaison linéaire $\sum_i b^i (x(u_i))^j$ des lois de probabilité associées aux centres u_i des intervalles précédents. L'on estimera les coefficients de la combinaison linéaire précédente, et donc la fonction g(u) à partir des 5 ou 6 premiers facteurs issus de l'analyse des correspondances du tableau des 50 ou 100 lois de probabilité considérées (cf [PHOTOMULTIPLICATEUR], ce cahier, pp 393-417).

D'un point de vue numérique si $F_\alpha(y)$ désigne l'abscisse de la projection de y sur l'axe factoriel α (y compris l'axe factoriel trivial, correspondant à α = 0 et pour lequel $F_0(y) = 1$), le $i^{\text{ème}}$ coefficient de régression que nous noterons b_*^i s'écrit, comme il est aisé de le voir, à l'aide de la formule de transition :

$$b_*^i = ((y\,tot)/(X\,tot))\Sigma\{F_\alpha(y)F_\alpha(i)/\lambda_\alpha \mid \alpha = 0,\ r\}$$

formule où r désigne le nombre de facteurs (non triviaux) conservés, Xtot la somme des éléments du tableau $X_{I_1}^{J}$ des variables explicatives, ytot la somme des y^j (dans le cas particulier des lois de probabilité envisagé ci-dessus, ytot vaut 1, tandis que Xtot est égal au nombre p de lois de probabilités retenues pour expliquer y), et $F_\alpha(i)$ l'abscisse de la projection du point i sur l'axe α.

5.2.2 *Régression par l'analyse des correspondances après découpage en classes*

Pour étudier les liaisons entre y et $\{X_i \mid i \in I_1\}$, on rend, comme on l'a déjà dit (cf §§ 2, 5.1), toutes les variables qualitatives par découpage en classes, et l'on fait l'analyse des correspondances du tableau $t_{K_1K_2}$ croisant l'ensemble K_2 des modalités de la variable à expliquer y avec l'ensemble K_1 des modalités de toutes les variables explicatives : $K_1 = \cup\{L_i \mid i \in I_1\}$, où L_i désigne l'ensemble des modalités de X_i. Rajoutant en supplémentaire de $t_{K_1K_2}$ le tableau disjonctif complet p_{JK_1} associé aux variables explicatives, on peut projeter chaque observation j sur les axes factoriels et effectuer sur les facteurs ainsi obtenus une régression usuelle ou une régression par boule.

Signalons quelques propriétés de la régression par l'analyse des correspondances.

1) Si y est indépendant d'une variable X_i, la restriction à L_i de tous les facteurs non triviaux sur K_1 de $t_{K_1K_2}$ est nulle, et l'analyse de $t_{K_1K_2}$ est équivalente à celle du tableau $t_{K_1-L_i,K_2}$ croisant y avec les variables explicatives autres que X_i.

2) Désignant par $K = K_1 \cup K_2$ l'ensemble des modalités de toutes les variables, l'analyse des correspondances du tableau $t_{K_1K_2}$ est équivalente à celle du tableau t_{KK_2} croisant K avec K_2.(*)

Les propriétés 1) et 2) découlent immédiatement de l'équation des facteurs des tableaux concernés.

3) Les facteurs non triviaux φ_α^J calculés sur les individus sont centrés. Ils sont en général corrélés, sauf si les variables explicatives sont indépendantes auquel cas l'analyse des correspondances de $t_{K_1K_2}$ est équivalente à l'analyse des correspondances du tableau disjonctif complet p_{KJ} croisant toutes les variables avec elles-mêmes.

Notons que si l'on a plusieurs variables à expliquer, et si K_2 désigne l'ensemble des modalités de ces variables, les propriétés 1) et 3) se généralisent aisément (cf [9] et [10]).

En particulier, si les variables explicatives (i.e. associées à K_1) sont indépendantes, les facteurs φ_α^J obtenus en rajoutant en supplémentaire de $t_{K_1K_2}$, le tableau p_{JK_1}, sont non corrélés et de variance 1/p, où p désigne, rappelons-le le nombre Card I_1 de variables explicatives.

BIBLIOGRAPHIE

La bibliographie sur les méthodes de régression donnée ci-dessous est loin d'être exhaustive. En particulier ne sont pas cités les articles relatifs à des méthodes comme la régression robuste qui n'ont pas été traitées ici.

En ce qui concerne la régression bornée (*ridge regression*) une très abondante littérature lui a été consacrée depuis une dizaine d'années. On n'a renvoyé le lecteur qu'aux principaux articles. En particulier un grand nombre d'articles (dont un certain nombre paru dans *Communication in Statistics*) ayant trait à des simulations n'ont pas été reportés ici de façon à ne pas trop alourdir cette bibliographie.

[1] BALLINI, J.P., CAZES, P., TURPIN, P.Y. (1976) :
Single electrons multiplication as a combination of poissonian pulse height distributions using constraint methods, Nucl. Inst. and Meth. 124, pp 319-330

[2] BHATTACHARYA, P.K. (1966) :
Estimating the mean of a multivariate normal population with general quadratic loss function, Ann. of Math. Stat. 37, pp 1819-1824

[3] BORDET, J.P. (1973) :
Etudes de données géophysiques. Modélisations statistiques par régression factorielle ; thèse 3° c. Paris VI

(*) *Cette propriété nous a été signalée par R. Haeflinger*

[4] BRENOT, J. (1977) :

Contributions à la pratique du modèle linéaire : qualité, protection et estimation biaisée ; thèse de 3° cycle, Paris VI

[5] CAZES, P. (1970) :

Application de l'analyse des données au traitement de problèmes géologiques ; thèse de 3° cycle, Paris VI

[6] CAZES, P., TURPIN, P.Y. (1971) :

Régression sous contrainte. Application à l'estimation de la courbe granulométrique d'un aérosol, R. S. A., Vol XIX, n° 4, pp 23-44

[7] CAZES, P. (1975) :

Protection de la régression par utilisation de contraintes linéaires et non linéaires, R. S. A., Vol XXIII, n° 3, pp 37-57

[8] CAZES, P. (1975) :

Techniques de prévision linéaire : Etude et protection. Cours : école d'été d'analyse numérique de l'E.D.F., l'I.R.I.A. et le C.E.A. .

[9] CAZES, P. (1976) :

Régression par l'analyse des données. Congrès Européen des Statisticiens, Grenoble.

[10] CAZES, P. (1977) :

Etude de quelques propriétés extrémales des facteurs issus d'un sous tableau d'un tableau de Burt [Extr. Fac.], Les Cahiers de l'Analyse des Données, Vol II n°2, pp 143-160

[11] CAZES, P. (1977) :

Estimation biaisée et estimation sous contraintes dans le modèle linéaire, Premières journées internationales analyse des données et informatique, I.R.I.A., pp 223-232

[12] CAZES, P., REYRE, Y. (1976) :

La fossilisation du kérogène en milieu argilo carbonaté. Etude statistique de ses liaisons avec les propriétés lithologiques et pétrologiques dans l'Oxfordien du bassin de Paris (partie orientale), BULL. B.R.G.M. (2° série) section IV, n°2, pp 85-102

[13] CAZES, P. (1976) :

Régression par boule et par l'analyse des correspondances, R.S.A., Vol XXIV, n° 4, pp 5, 22

[14] CAZES, P. (1978)

Estimation de la statistique de multiplication du premier étage d'un photomultiplicateur à dynodes, [PHOTOMULTIPLICATEUR], *Les Cahiers de l'Analyse des Données* , Vol. III n°4, pp 393-417.

[15] CONESA, A., CAZES, P. et collaborateurs (1976) :

Etude globale de la culture de la betterave à sucre sur le périmètre du haut Chéliff, II, Analyse en régression, Ann. agron., 27 (1), pp 61-84

[16] DEMSTER, A.P., SCHATZOFF, M., WERMUTH, N. (1977)

A simulation study of alternative to ordinary least squares (avec dicussion), JASA, Vol 72, n° 357, pp77-106

[16 bis] DRAPER, N., SMITH, H. (1966) :

Applied Regression Analysis, John Wiley *and* Sons, New-York

[17] EFROYMSON, A. (1962) :

Multiple Regression Analysis, in RALSTON, A., WILF, H.S. (1962) : *Mathematicals Methods for Digital Computers*, John Wiley *and* Sons, New-York.

[18] FAREBROTHER, R.W. (1975) :

The minimum mean square error linear estimator and ridge regression, Technometrics, Vol 17, n°1, pp 127-128

[18 bis] GUNST, R.F., MASON, R.L. (1977) :

Biased estimation in regression : an evaluation using mean squared error, JASA, Vol 72, n° 359, pp 616-628

[19] HOCKING, R.R. (1976) :

The analysis and selection of variables linear regression, Biometrics, Vol 32, n° 1, pp 1-49

[20] HOCKING, R.R., SPEED, F.M., LYNN, M.J. (1976) :

A class of biased estimators in linear regression, Technometrics, Vol 18, n° 4, pp 425-437

[21] HOERL, A.E., KENNARD, R.W. (1970) :

Ridge regression : biased estimation for non orthogonal problems, Technometrics, Vol 12, n° 1, pp 55-67

[22] HOERL, A.E., KENNARD, R.W. (1970) :

Ridge regression : application to non orthogonal problems, Technometrics, Vol 12, n° 2, pp 69-82

[23] JAMES, W., STEIN, C. (1961) :

Estimation with quadratics loss : Procedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley 1, pp 361-379

[24] LAWSON, C.L., HANSON, R.J. (1974) :

Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New-Jersey

[25] LEBEAUX, M.O. (1974) :

Programme de régression et de classification utilisant la notion de voisinage, thèse de 3° cycle, Paris VI

[26] LEBEAUX, M.O. (1977) :

Notice sur l'utilisation du programme Poubel [POUBEL], Les Cahiers de l'Analyse des Données, Vol II, n°4, pp 467-481

[27] LEGENDRE, P. (1977) :

Une extension des moindres carrés pour la régression sur variables entachées d'erreurs. Problème avec ou sans contraintes, thèse de 3° cycle; Toulouse

[28] LINDLEY , D.V., SMITH, A.F.M. (1972) :

Bayes Estimates for the linear model, J.R.S.S., série B, Vol 34, pp 1-18

[29] MALLOWS, C.L. (1973) :

Some Comments on C_p , *Technometrics*, Vol 15, n° 4 , pp 661-675

[30] MARQUARDT, D.W. (1970) :

Generalized inverses, ridge regression, biased linear estimation and non linear estimation, Technometrics, Vol 12, n° 3, pp 591-612

[31] SCLOVE, S.L. (1968) :

Improved estimators for coefficients in linear regression, J.A.S.A., Vol 63, n° 322, pp 596-606

[31 bis] TIRET, L. (1978) :

Apport de l'analyse des données à l'orientation diagnostique en psychiatrie infantile.Etude de certaines caractéristiques médicales, thèse de 3° cycle , Paris VI

[32] TORO- VIZCARONDO C., WALLACE, T.D. (1968) :

A test of the mean square error criterion for restrictions in linear regression, J.A.S.A., Vol 63,n° 322, pp 558-572

[33] VINOD, H.D. (1976) :

Simulation and extension of a minimum mean squared error estimator in comparison with Stein's, Technometrics, Vol 18 n° 4, pp 491-496

[34] VINOD, H.D. (1976) :

Application of new ridge regression methods to a study of Bell system scale economies, J.A.S.A., Vol 71, n° 356, pp 835-841

[35] WEBSTER, J.T., GUNST, R.F., MASON, R.L. (1974) :

Latent root regression analysis, Technometrics, Vol 16, n° 4, pp 513-522