

A. HATHOUT

Régression avec nombre variable de voisins et régression avec stratégie variable

Les cahiers de l'analyse des données, tome 10, n° 4 (1985),
p. 470-476

http://www.numdam.org/item?id=CAD_1985__10_4_470_0

© Les cahiers de l'analyse des données, Dunod, 1985, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

RÉGRESSION AVEC NOMBRE VARIABLE DE VOISINS ET RÉGRESSION AVEC STRATÉGIE VARIABLE

[REGR. STRA. VAR.]

par A. Hathout *

0 Stratégies de régression

Soit I un ensemble d'individus. D'une part chaque $i \in I$ est décrit par un ensemble J de variables explicatives dont les valeurs $P(i, j)$ contiennent un tableau de correspondance $I \times J$. D'autre part, pour chaque $i \in I$ on a un réel $v(i)$ dont l'intervalle de variation est divisé en un ensemble C de tranches, ou classes ; la tranche où tombe $v(i)$ étant notée $t(i)$.

Soit K un ensemble de CARDK algorithmes ou stratégies de régression. Chaque $k \in K$ estime pour tout i décrit par J une valeur $\hat{v}(i, k)$ en fonction des variables j .

En 1979, J.P. Benzécri nous a suggéré d'appliquer lors de l'estimation de $\hat{v}(i, k)$ la stratégie $k \in K$ censée commettre la moindre erreur, eu égard à un certain nombre de considérations relatives au voisinage de i . C'est pourquoi nous présentons ici une méthode d'estimation avec stratégie variable ainsi qu'une application montrant son efficacité.

1 Choix d'une stratégie par segmentation

Supposons définie une distance d entre deux profils d'individus décrits par J . Soit c une tranche : $c \in C$, on notera $g(c)$ le centre de gravité du sous nuage des profils sur J des individus principaux i ($i \in I$) pour lesquels la variable à expliquer rentre dans cette tranche ($t(i) = c$).

A tout individu i on associera une classe c pour laquelle la distance $d(i, g(c))$ prend la valeur minima (éventuellement il y a plusieurs c pour i ; le programme choisira alors le premier de ceux-ci suivant l'ordre séquentiel) cette classe sera notée $cl(i) : d(i, g(cl(i))) = \inf\{d(i, g(i)) \mid i \in C\}$. Il importe de souligner que $cl(is)$ peut être calculé pour un individu supplémentaire is , parce que les calculs de distance ne font intervenir que l'ensemble J des variables explicatives. Quant aux individus principaux, on n'a pas nécessairement $cl(i) = t(i)$, un individu i pouvant être plus proche du centre $g(c')$ que du centre $g(t(i))$ de la classe c où rentre $v(i)$. On notera donc :

$$I(c) = \{i \mid i \in I ; cl(i) = c\}$$

L'ensemble des individus principaux associés à la classe c . Considérons l'erreur $e(i, k)$ commise dans l'estimation de l'algorithme $v(i)$ par l'algorithme k :

(*) Statisticien.

$$e(i,k) = |v(i) - \hat{v}(i,k)|$$

On construira le tableau :

$$I(c,k) = \Sigma\{e(i,k) | i \in I, cl(i) = c\}$$

ou somme des erreurs commises par la stratégie k dans l'évaluation des valeurs v(i) pour les individus i ∈ I associés à la classe c.

A partir du tableau I on définira la stratégie S_I(c) à adopter associée à la classe c relativement à un ensemble I comme étant celle qui commet la plus petite erreur :

$$S_I(c) = k \Leftrightarrow \forall k' \neq k, I(c,k) \leq I(c,k').$$

De façon précise : S_I(c) est la stratégie qui a commis la plus petite erreur dans l'évaluation de v pour les individus de I associés à la classe c.

Plus brièvement S_I(c) = k se lira : la stratégie dans I pour c est k ou "dans I, la stratégie pour c est k". Remarquons que si nous notons :

$$I(c,*) = I(c, S_I(c))$$

nous avons par définition la propriété suivante :

$$\forall k \in K, I(c,*) \leq I(c,k)$$

I(c,*) est la plus petite erreur commise par les algorithmes de K dans c.

Maintenant, pour un individu supplémentaire is la méthode consistera à estimer v(is) par la stratégie qui a commis la plus faible erreur dans l'ensemble I(cl(is)) des individus principaux associés à la même classe cl(is) que is, soit :

$$\hat{v}(is) = \hat{v}(is, S_I(cl(is)))$$

2 Avantages et inconvénients des diverses stratégies

Pour la bonne intelligence de ce qui suit, rappelons :

- que nous abordons le problème de la régression en partant d'un point is projeté dans un ensemble I de points où on détermine un sous ensemble J₁ ⊂ I représentatif de is ; la valeur estimée v(is) étant la moyenne des valeurs de v(i) sur I₁

- que nos stratégies sont d'une part la régression d'après un nombre variable de voisins appelée dans la suite N. V. (1) et d'autre part la régression dite par boules (5) avec un nombre de voisins allant de 2 à 10, mais fixé *a priori* (nous écrirons N. F.).

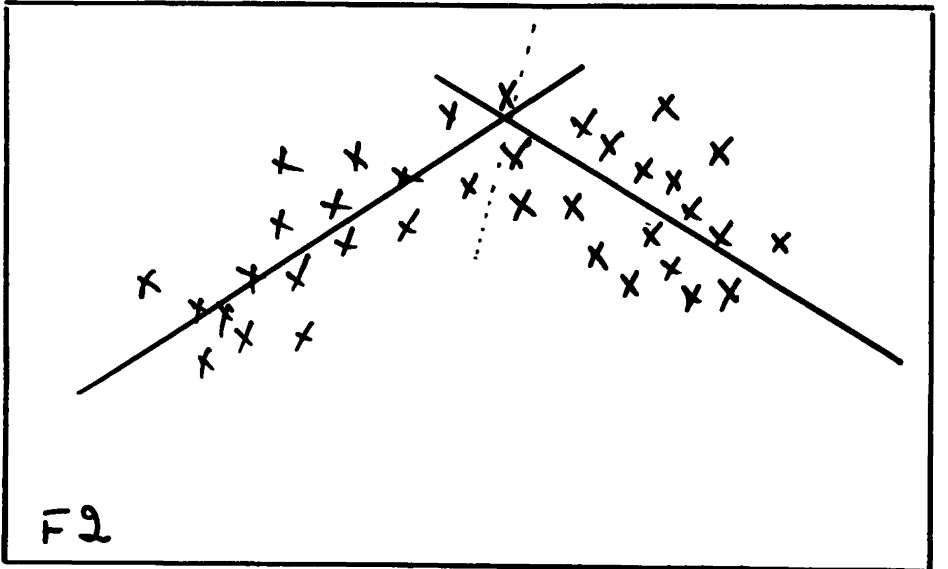
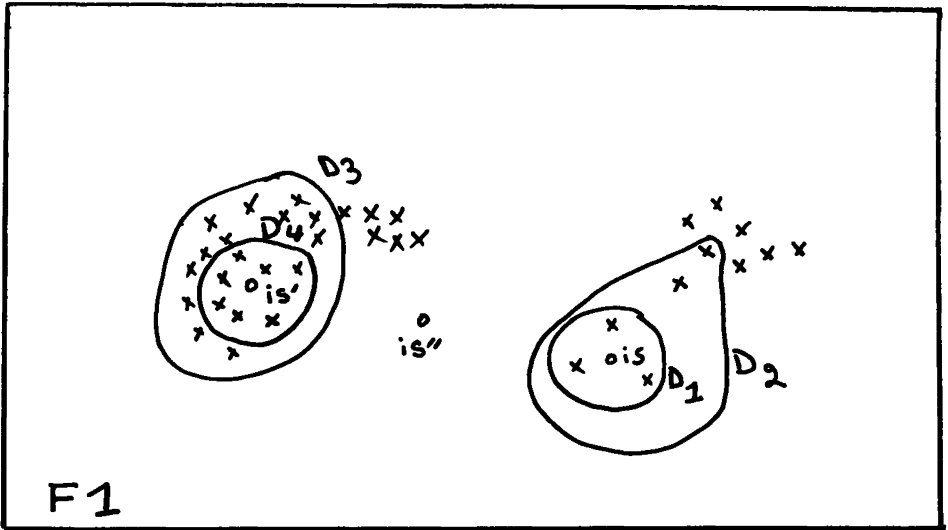
Considérons la figure F1.

Le point is sera représenté par les trois points de D₁ avec la régression N. Variable ce qui est judicieux. Par contre la régression N. Fixe avec 5 voisins assimilera is à D₂, qui semble trop dispersé.

(1) CAD, Vol VIII, n° 1, pp 19-26, 1983, [REGR. NVAR. VOIS.]

Le point is sera correctement assimilé aux points de D_4 par la régression N. Fixe avec 7 voisins alors que l'estimation ne sera pas forcément aussi satisfaisante avec la régression N. Variable ; l'expérience a montré que très souvent la régression N. Variable s'arrête à un seul voisin.

Le point is'' sera aussi mal estimé par la régression N. Fixe que par la régression N. Variable car il est éloigné de tout élément I ce qui pose le problème de régressibilité déjà abordé dans (1).



Soulignons enfin que même dans le cas d'une stratégie unique, l'avantage de la segmentation (Régression typologique) a également été observé expérimentalement (2) par d'autres chercheurs et la figure F2 nous autorise à penser que l'estimation faite à partir d'hyperplans de régression établis sur des classes prédéfinies donnent de meilleurs résultats que l'estimation d'après un seul hyperplan de régression.

3 Choix et validation empirique

Dans le choix, l'évaluation et l'utilisation d'un ensemble de stratégies, il faut distinguer plusieurs étapes, à chacune desquelles peuvent intervenir différents ensembles I d'individus décrits par le même ensemble J de variables explicatives et pour lesquels la variable à expliquer v est ou non connue.

1°) Choix des centres de tranche d'après lesquels on effectuera une segmentation en fonction des variables J : ce choix est fait pour un ensemble I, nous dirons IP, d'individus principaux pour lesquels v est connu.

2°) Segmentation d'une population sur chaque segment de laquelle on déterminera par essai une meilleure stratégie $S_I(c)$; sur cette population, notée IS, on devra également connaître v .

3°) Essai du système de stratégie variable déterminé en 2°) : on devra pour cela prendre une population IE sur laquelle v est connue.

4°) Application de la méthode : cette fois l'opération n'a d'intérêt que s'il s'agit d'un ensemble IA pour lequel la variable à estimer v est inconnue !

En principe, il semble préférable que IP, IS, IA soient trois ensembles distincts. Mais pratiquement si l'on dispose d'un nombre modeste d'individus pour lesquels v est connu, il pourra être préférable d'accepter que deux de ces ensembles ou les trois se confondent afin d'avoir pour chacun des effectifs suffisants.

Dans le cas présent on a fait deux essais

1° IP = IS = IA = I

2° IP = IS = I et IA = un échantillon aléatoire représentant le quart de I.

Nous présentons ci-dessous une brève description du jeu de données utilisées ainsi que des résultats obtenus.

4 L'exemple traité

La population est composée de 2135 individus interrogés en 1976 dans le cadre de l'Observatoire National d'entrée dans la vie active. Ils ont terminé en 1975 une formation de techniciens supérieurs et ont trouvé un emploi.

Cette population est décrite par 21 variables transformées en 264 modalités de présence-absence dont nous donnons ci-après la liste (3).

La variable à expliquer est le salaire découpé en 18 tranches croissantes.

(2) Thèse de 3° cycle présentée à Paris-Dauphine le 12/1/84 par M. TRANSO.

(3) cf. J. Lamoure et X. Viney Fév. 1982. "L'entrée dans la vie active des Jeunes sortant d'IUT et de STS. Complémentarités et Concurrences" Cahier n° 8 de l'Observatoire National d'Entrée dans la Vie Active.

```

(1)***SPECIALITE D'ETUDES
(2)***STATUT ETABLISSEMENT
(3)***AGE
(4)***REGION D'ETUDES
(5)***REGION DE RESIDENCE
(6)***ORIGINE SCOLAIRE
(7)***ORIENTATION DIPLOME
(8)***TEMPS MIS A LA RECHERCHE D'UN EMPLOI
(9)***COMMENT L'EMPLOI A ETÉ TROUVE
(10)***SERVICE DE TRAVAIL
(11)***STATUT DE L'EMPLOISSEMENT
(12)***TAILLE DE L'ETABLISSEMENT
(13)***SECTEUR D'ACTIVITE DE L'ETABLISSEMENT
(14)***NOMBRE HEURES DE TRAVAIL PAR SEMAINE
(15)***NOM DE L'EMPLOI
(16)***SITUATION DE FAMILLE
(17)***SITUATION MILITAIRE
(18)***REGION DE RESIDENCE
(19)***TYPE DE MIGRATION INTER-REGIONS
(20)***CONTEXTE EMPLOI
(21)***CONTEXTE CHOMAGE
(22)***CONTEXTE NON EMPLOI
(23)***POSSIBILITE CHOIX BTS DANS REGION ETUDES
(24)***POSSIBILITE CHOIX BTS DANS REGION RESIDENCE

```

L'ensemble K des stratégies essayées est le suivant :

- Pour $k = 1$, la stratégie adoptée est la régression d'après un nombre variable de voisins.

- Pour $k = 2$ à 10, la stratégie est la régression dite par boules avec un nombre fixe de voisins égal à 2 pour la stratégie $k = 2$, à 3 pour la stratégie $k = 3, \dots$ et à 10 pour la stratégie $k = 10$. (Ne pas confondre le nombre de voisins et la stratégie : nous utilisons toujours k pour indiquer la stratégie même si $k = 1$, ce qui implique un nombre de voisins variables). Nous commentons ci après les résultats obtenus dans les tableaux des erreurs commises par les diverses stratégies (10) dans chaque classe (18 tranches de salaires donc 18 classes). Le premier tableau, I porte sur tous les individus. Le deuxième tableau IA porte sur un échantillon IACI composé de 25% des individus de I .

Voici ce que nous observons dans I .

1° le total des erreurs commises par la régression N. Variable est inférieur à celui de n'importe laquelle des 9 autres stratégies i . e. que s'il fallait ne choisir qu'une seule stratégie, c'est la stratégie N. Variable que nous prendrions.

2° sur 7 des 18 classes, la régression N. Variable commet la plus petite erreur (7 est à comparer avec $1,8 =$ espérance du nombre total des classes où une stratégie prise parmi 10 commet la moindre erreur).

Voici ce que nous observons dans IA

1° sur 12 des 18 classes, l'égalité $S_I(c) = S_{IA}(c)$ est vérifiée autrement dit le choix de la stratégie afférente à la classe c est le même que dans l'échantillon de base.

2° sur 7 des 18 classes, la régression N. Variable commet, comme précédemment, la plus petite erreur; (les 7 classes ne sont pas toutes les mêmes que celles référencées pour I)

3° le total des erreurs commises par la régression N. Variable est toujours plus petit en comparaison avec les autres stratégies

ERREURS SUR L'ENSEMBLE I

A	NV.	2	3	4	5	6	7	8	9	10
1	464	357	362	359	351	360	350	356	350	364
2	1068	891	884	833	827	795	780	779	766	788
3	651	657	568	549	529	529	529	566	606	646
4	232	277	275	282	284	272	267	279	279	285
5	478	467	498	488	472	468	463	479	439	480
6	493	478	474	471	448	448	446	466	438	424
7	338	366	370	359	360	360	361	361	363	360
8	735	746	744	733	713	713	710	706	685	723
9	735	746	744	733	713	713	710	706	685	723
10	833	844	853	854	840	860	860	834	834	843
11	651	657	568	549	529	529	529	566	606	646
12	478	467	498	488	472	468	463	479	439	480
13	493	478	474	471	448	448	446	466	438	424
14	338	366	370	359	360	360	361	361	363	360
15	735	746	744	733	713	713	710	706	685	723
16	735	746	744	733	713	713	710	706	685	723
17	833	844	853	854	840	860	860	834	834	843
18	1058	891	884	833	827	795	780	779	766	788
19	274	618	991	991	991	991	991	991	991	991
20	274	618	991	991	991	991	991	991	991	991
21	274	618	991	991	991	991	991	991	991	991
22	274	618	991	991	991	991	991	991	991	991
23	274	618	991	991	991	991	991	991	991	991
24	274	618	991	991	991	991	991	991	991	991
25	274	618	991	991	991	991	991	991	991	991
26	274	618	991	991	991	991	991	991	991	991
27	274	618	991	991	991	991	991	991	991	991
28	274	618	991	991	991	991	991	991	991	991
29	274	618	991	991	991	991	991	991	991	991
30	274	618	991	991	991	991	991	991	991	991
31	274	618	991	991	991	991	991	991	991	991
32	274	618	991	991	991	991	991	991	991	991
33	274	618	991	991	991	991	991	991	991	991
34	274	618	991	991	991	991	991	991	991	991
35	274	618	991	991	991	991	991	991	991	991
36	274	618	991	991	991	991	991	991	991	991
37	274	618	991	991	991	991	991	991	991	991
38	274	618	991	991	991	991	991	991	991	991
39	274	618	991	991	991	991	991	991	991	991
40	274	618	991	991	991	991	991	991	991	991
41	274	618	991	991	991	991	991	991	991	991
42	274	618	991	991	991	991	991	991	991	991
43	274	618	991	991	991	991	991	991	991	991
44	274	618	991	991	991	991	991	991	991	991
45	274	618	991	991	991	991	991	991	991	991
46	274	618	991	991	991	991	991	991	991	991
47	274	618	991	991	991	991	991	991	991	991
48	274	618	991	991	991	991	991	991	991	991
49	274	618	991	991	991	991	991	991	991	991
50	274	618	991	991	991	991	991	991	991	991
51	274	618	991	991	991	991	991	991	991	991
52	274	618	991	991	991	991	991	991	991	991
53	274	618	991	991	991	991	991	991	991	991
54	274	618	991	991	991	991	991	991	991	991
55	274	618	991	991	991	991	991	991	991	991
56	274	618	991	991	991	991	991	991	991	991
57	274	618	991	991	991	991	991	991	991	991
58	274	618	991	991	991	991	991	991	991	991
59	274	618	991	991	991	991	991	991	991	991
60	274	618	991	991	991	991	991	991	991	991
61	274	618	991	991	991	991	991	991	991	991
62	274	618	991	991	991	991	991	991	991	991
63	274	618	991	991	991	991	991	991	991	991
64	274	618	991	991	991	991	991	991	991	991
65	274	618	991	991	991	991	991	991	991	991
66	274	618	991	991	991	991	991	991	991	991
67	274	618	991	991	991	991	991	991	991	991
68	274	618	991	991	991	991	991	991	991	991
69	274	618	991	991	991	991	991	991	991	991
70	274	618	991	991	991	991	991	991	991	991
71	274	618	991	991	991	991	991	991	991	991
72	274	618	991	991	991	991	991	991	991	991
73	274	618	991	991	991	991	991	991	991	991
74	274	618	991	991	991	991	991	991	991	991
75	274	618	991	991	991	991	991	991	991	991
76	274	618	991	991	991	991	991	991	991	991
77	274	618	991	991	991	991	991	991	991	991
78	274	618	991	991	991	991	991	991	991	991
79	274	618	991	991	991	991	991	991	991	991
80	274	618	991	991	991	991	991	991	991	991
81	274	618	991	991	991	991	991	991	991	991
82	274	618	991	991	991	991	991	991	991	991
83	274	618	991	991	991	991	991	991	991	991
84	274	618	991	991	991	991	991	991	991	991
85	274	618	991	991	991	991	991	991	991	991
86	274	618	991	991	991	991	991	991	991	991
87	274	618	991	991	991	991	991	991	991	991
88	274	618	991	991	991	991	991	991	991	991
89	274	618	991	991	991	991	991	991	991	991
90	274	618	991	991	991	991	991	991	991	991
91	274	618	991	991	991	991	991	991	991	991
92	274	618	991	991	991	991	991	991	991	991
93	274	618	991	991	991	991	991	991	991	991
94	274	618	991	991	991	991	991	991	991	991
95	274	618	991	991	991	991	991	991	991	991
96	274	618	991	991	991	991	991	991	991	991
97	274	618	991	991	991	991	991	991	991	991
98	274	618	991	991	991	991	991	991	991	991
99	274	618	991	991	991	991	991	991	991	991
100	274	618	991	991	991	991	991	991	991	991
NB	7	0	0	0	2	0	1	1	5	2

96 I

ERREURS SUR L'ENSEMBLE IA (25% DE I)

A	NV.	2	3	4	5	6	7	8	9	10
1	101	70	72	61	59	62	69	64	67	68
2	200	133	144	139	140	136	131	133	133	135
3	150	103	112	109	110	109	105	108	108	108
4	334	236	252	244	240	237	232	236	236	236
5	107	110	111	110	110	101	97	94	95	94
6	116	103	102	105	105	101	97	94	95	94
7	202	144	150	148	144	144	136	134	133	133
8	202	144	150	148	144	144	136	134	133	133
9	202	144	150	148	144	144	136	134	133	133
10	202	144	150	148	144	144	136	134	133	133
11	202	144	150	148	144	144	136	134	133	133
12	202	144	150	148	144	144	136	134	133	133
13	202	144	150	148	144	144	136	134	133	133
14	202	144	150	148	144	144	136	134	133	133
15	202	144	150	148	144	144	136	134	133	133
16	202	144	150	148	144	144	136	134	133	133
17	202	144	150	148	144	144	136	134	133	133
18	202	144	150	148	144	144	136	134	133	133
19	202	144	150	148	144	144	136	134	133	133
20	202	144	150	148	144	144	136	134	133	133
21	202	144	150	148	144	144	136	134	133	133
22	202	144	150	148	144	144	136	134	133	133
23	202	144	150	148	144	144	136	134	133	133
24	202	144	150	148	144	144	136	134	133	133
25	202	144	150	148	144	144	136	134	133	133
26	202	144	150	148	144	144	136	134	133	133
27	202	144	150	148	144	144	136	134	133	133
28	202	144	150	148	144	144	136	134	133	133
29	202	144	150	148	144	144	136	134		

5 Conclusion

1° Comme nous l'avons vu en 3, le programme nécessite plusieurs étapes ce qui le rend très coûteux. Sa mise en oeuvre "paramétrée" en un seul programme "vaut la peine", nous venons de le voir, mais elle nécessite une infrastructure de "Recherche".

2° Une critique de ce travail qui est aussi un espoir est que le mode conversationnel permettra à l'utilisateur spécialiste d'un domaine (médecin, financier,...) de régler la marche du programme en se basant sur le contenu de ses données autant ou plus que sur des critères numériques (*).

3° La méthode ne va pas au bout de ses propres pensées (vu le coût, elle y aurait du mal) : de même qu'il est licite de penser qu'à des parties différentes de l'espace correspondent des stratégies différentes, on peut se conjecturer qu'à des individus "excentrés" correspondront de meilleures estimations dans l'espace des axes issus d'une analyse des correspondances et cumulant seulement 60% de l'inertie alors que relativement à des individus "moyens" les données brutes donnerons de meilleures estimations.

(*) Recherche du (des) cas clinique (S) connu (S) les (S) plus proche (S) d'un cas nouveau ; et, en homéopathie, recherche du (des) type (S) homéopathique (S) dont les symptômes correspondent au mieux à ceux du patient.