

F. BENZÉCRI

J.-P. BENZÉCRI

Calculs de corrélation entre variables et juxtaposition de tableaux

Les cahiers de l'analyse des données, tome 14, n° 3 (1989),
p. 347-354

http://www.numdam.org/item?id=CAD_1989__14_3_347_0

© Les cahiers de l'analyse des données, Dunod, 1989, 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/>

CALCULS DE CORRÉLATION ENTRE VARIABLES ET JUXTAPOSITION DE TABLEAUX

[CORREL. JUXT.]

J.-P. & F. BENZÉCRI

1 Place des calculs de régression et corrélation dans l'analyse des données multidimensionnelles

Nous avons souvent manifesté de la réserve devant la pratique des calculs de corrélation et de régression. En effet, les liens existant entre deux variables sont ordinairement trop complexes pour qu'il soit judicieux de tenter de les exprimer par un seul nombre; et la possibilité de connaître une ou plusieurs variables à expliquer en fonction d'un groupe de variables explicatives apparaît au mieux après l'analyse du tableau de correspondance croisant les modalités de ces variables-ci et de celles-là.

Cependant il peut être satisfaisant de constater qu'un facteur est étroitement corrélé à une variable qui ne figure pas explicitement dans le tableau d'où est issu ce facteur; et, tandis que les formules de régressions polynomiales sont souvent illusoire, on peut s'intéresser à une simple formule de régression linéaire par morceaux construite en raccordant des régressions linéaires calculées sur des intervalles empiétant de la variable explicative.

Dans la présente note, nous présentons donc deux programmes, 'corel' et 'juxtab' qui ont été intégrés au logiciel MacSAIF, pour effectuer de tels calculs.

Partant d'un tableau dont le format est celui d'un tableau de correspondance usuel (ou encore d'un tableau de facteurs) lisible par 'zrang' (cf. [NOTE CRÉ. TAB.], §2, in *CAD*, vol. XIV, n°1, 1989), 'corel' calcule des coefficients de corrélation entre colonnes; en donnant pour poids aux lignes soit la constante 1, soit des poids variables fournis par une colonne du tableau. Avec le coefficient de corrélation sont données des formules de régression; éventuellement, l'intervalle de variation d'une des variables est restreint à des bornes.

Comme les calculs de corrélation sont le plus utiles entre variables ou facteurs appartenant à des tableaux différents, 'juxtab' permet de juxtaposer deux tableaux, pour en créer un seul sur lequel opérera 'corel'.

Au §2, on présente le dialogue d'utilisation de 'corel' sur le cas d'une fonction construite précisément pour rappeler les résultats illusoire aux quels peuvent conduire les calculs de régression, surtout dans le cas où l'intervalle de variation de la variable explicative n'est pas judicieusement délimité.

Au §3, on considère des données réelles, recueillies au cours d'un essai thérapeutique et dont l'étude approfondie fait l'objet d'un article (cf. [ÉCHELLES], in *CAD*, Vol XIV, n°3, 1989). Afin d'effectuer une régression linéaire par morceaux d'une donnée (variable supplémentaire à expliquer) en fonction du premier facteur (variable explicative) on juxtapose le tableau des facteurs au tableau des données; ce qui offre l'occasion de présenter les fonctions de 'juxtab'.

2 Calculs de corrélation et de régression pour une fonction linéaire par morceaux

Étant (comme l'analyse factorielle) fondés sur des propriétés d'inertie, les calculs de corrélation et d'ajustement linéaire (au sens des moindres carrés) tendent à être dominés par les éléments les plus écartés; lesquels sont éventuellement étrangers au problème étudié et ne peuvent, en tout cas, être considérés seuls, sans égard à la structure des éléments centraux.

En analyse des correspondances, l'influence des éléments excentriques sur un facteur peut et doit être contrôlée sur la colonne CTR. Au contraire, les calculs de corrélation et les analyses en composantes principales (qui sont des diagonalisations de matrices de corrélations) sont souvent effectués sans contrôle et comme machinalement.

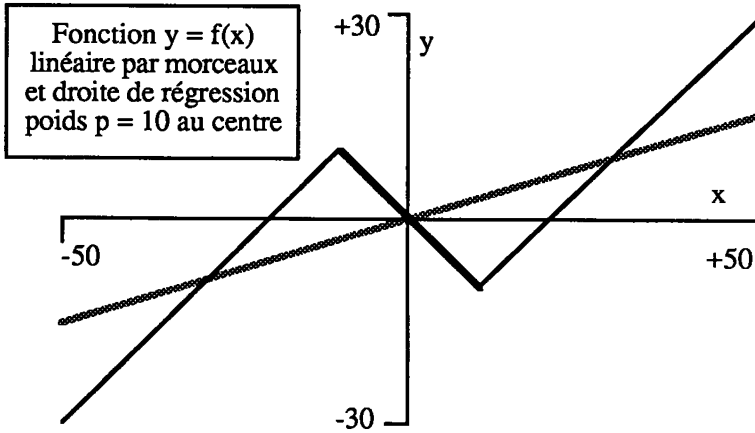
Dans le présent §, on considère le cas très simple d'un tableau de données qui ne comprend qu'une variable explicative x et une variable à expliquer y , celle-ci étant rigoureusement liée à celle-là par une relation fonctionnelle linéaire par morceaux; avec dans la partie centrale du champ de variation de x , une pente (-1); et sur les bords une pente (+1).

En pondérant de quatre façons différentes les données individuelles, on obtient, sur l'ensemble de l'intervalle de variation de x , les coefficients de corrélation les plus divers; tandis que, bien évidemment, en se bornant à des intervalles où y dépend linéairement de x , on a pour coefficient (-1) ou (+1).

De façon précise, le tableau des données comporte 101 lignes et 6 colonnes dont les sigles sont

$$\{x, y, p1, p10, p100, pmil\}.$$

Dans la colonne x on a 101 nombres entiers algébriques successifs, de -50 à +50; dans la partie centrale, (-10, 10), $y = -x$; dans les parties latérales on a, respectivement, $y = x+20$ sur (-50,-11) et $y = x-20$ sur (11,50).



On devine que, dans la colonne p1, se trouvent des nombres tous égaux à 1 et qui serviront de poids dans les calculs les plus simples; dans les colonnes p10, p100 et pml, on trouve respectivement les valeurs 10, 100 et 1000 dans l'intervalle central (-10, 10); tandis que dans les parties latérales (autrement dit, pour les 40 premières lignes et les 40 dernières) le poids est toujours 1.

Il faut enfin préciser que le tableau est présenté comme pour une analyse de correspondance ou pour 'zrang' (cf. *supra*); c'est-à-dire qu'il y a une ligne de titre, puis une ligne débutant par le nombre 6 et contenant les sigles des 6 colonnes; tandis que chacune des 101 lignes suivantes débute par un sigle (choisi, arbitrairement, comme 'b' pour l'intervalle central; et 'a' et 'c' pour les parties latérales $x < 0$ et $x > 0$) et contient ensuite les 6 valeurs de x , y et des poids.

```
ce programme calcule des correlations entre colonnes
le fichier des donnees est *:*:tc$
*:*:tc$w
ce nom est il confirme oui(0) ou non(N) 0
```

```
.....
```

c	48	29	1	1	1	1
c	49	29	1	1	1	1
c	50	30	1	1	1	1

```
tableau pour les calculs de correlation
ci dessous sont rappelés les num et sigles des colonnes
1 x; 2 y; 3 p1; 4 p10; 5 p100; 6 pml;
y a t il une colonne de poids 0 ou N 0
numero de la colonne de poids = 2
ERREUR la colonne a des valeurs < 0
numero de la colonne de poids = 1
ERREUR la colonne a des valeurs < 0
numero de la colonne de poids = 3
ce choix est il confirme 0 ou N 0
```

Le dialogue de 'corel' commence par le choix du tableau des données; lequel s'affiche; on choisit alors, éventuellement une colonne de poids; dont le programme vérifie qu'elle ne contient pas de nombre négatif.

```

tableau pour les calculs de corrélation
ci dessous sont rappelés les num et sigles des colonnes
  1  x; 2  y; 3  p1; 4  p10; 5  p100; 6  pml;
y a t il une colonne de poids 0 ou N 0
numero de la colonne de poids = 5
ce choix est il confirme 0 ou N 0
ptot = 2.1800000e+3
faut il creer un listage des correlations calculees 0 ou N 0
ci dessous sont rappelés les num et sigles des colonnes
  1  x; 2  y; 3  p1; 4  p10; 5  p100; 6  pml;
numero de la premiere col a corréler = 1
numero de la deuxieme col a corréler = 2
faut-il assujétir x a des bornes 0 ou N N
corr(x,y) = -3.2563000e-1 ; poids retenu = 1.0000000e+0
y - 0.0000000e+0 ≠ -2.5123396e-1 * (x - 0.0000000e+0)
x - 0.0000000e+0 ≠ -4.2205638e-1 * (y - 0.0000000e+0)
faut il calculer un autre corr 0 ou N 0
ci dessous sont rappelés les num et sigles des colonnes
  1  x; 2  y; 3  p1; 4  p10; 5  p100; 6  pml;
numero de la premiere col a corréler = 1
numero de la deuxieme col a corréler = 2
faut-il assujétir x a des bornes 0 ou N 0
borne inf de x = -100
borne sup de x = -10.5
ptot = 4.0000000e+1
corr(x,y) = 1.0000000e+0 ; poids retenu = 1.8348624e-2
y - -1.0500000e+1 ≠ 1.0000000e+0 * (x - -3.0500000e+1)
x - -3.0500000e+1 ≠ 1.0000000e+0 * (y - -1.0500000e+1)
faut il calculer un autre corr 0 ou N 0
ci dessous sont rappelés les num et sigles des colonnes
  1  x; 2  y; 3  p1; 4  p10; 5  p100; 6  pml;
numero de la premiere col a corréler = 1
numero de la deuxieme col a corréler = 2
faut-il assujétir x a des bornes 0 ou N 0
borne inf de x = -10.5
borne sup de x = 10.5
ptot = 2.1000000e+3
corr(x,y) = -1.0000000e+0 ; poids retenu = 9.6330275e-1
y - 0.0000000e+0 ≠ -1.0000000e+0 * (x - 0.0000000e+0)
x - 0.0000000e+0 ≠ -1.0000000e+0 * (y - 0.0000000e+0)
faut il calculer un autre corr 0 ou N 0
ci dessous sont rappelés les num et sigles des colonnes
  1  x; 2  y; 3  p1; 4  p10; 5  p100; 6  pml;
numero de la premiere col a corréler = 1
numero de la deuxieme col a corréler = 2
faut-il assujétir x a des bornes 0 ou N 0
borne inf de x = 10.5
borne sup de x = 100
ptot = 4.0000000e+1
corr(x,y) = 1.0000000e+0 ; poids retenu = 1.8348624e-2
y - 1.0500000e+1 ≠ 1.0000000e+0 * (x - 3.0500000e+1)
x - 3.0500000e+1 ≠ 1.0000000e+0 * (y - 1.0500000e+1)
faut il calculer un autre corr 0 ou N N
faut il calculer des corr sur un autre tableau 0 ou N N

```

Ne pas choisir de colonne de poids équivalent à choisir un poids constant (colonne 3 = p1, dans le cas présent).

Le programme demande ensuite s'il y a lieu de créer un listage des coefficients calculés (listage qui porte le nom du tableau de base suivi du suffixe 'relx'); quelle que soit la réponse, les résultats s'afficheront à l'écran.

```

tableau pour les calculs de corrélation
corr(x,y) = -3.2563000e-1
y - 0.0000000e+0 ≈ -2.5123396e-1 * (x - 0.0000000e+0)
x - 0.0000000e+0 ≈ -4.2205638e-1 * (y - 0.0000000e+0)
bornes de x : inf = -1.0000000e+2 ; sup = -1.0500000e+1
poids de l'intervalle retenu = 1.8348624e-2
corr(x,y) = 1.0000000e+0
y - -1.0500000e+1 ≈ 1.0000000e+0 * (x - -3.0500000e+1)
x - -3.0500000e+1 ≈ 1.0000000e+0 * (y - -1.0500000e+1)
bornes de x : inf = -1.0500000e+1 ; sup = 1.0500000e+1
poids de l'intervalle retenu = 9.6330275e-1
corr(x,y) = -1.0000000e+0
y - 0.0000000e+0 ≈ -1.0000000e+0 * (x - 0.0000000e+0)
x - 0.0000000e+0 ≈ -1.0000000e+0 * (y - 0.0000000e+0)
bornes de x : inf = 1.0500000e+1 ; sup = 1.0000000e+2
poids de l'intervalle retenu = 1.8348624e-2
corr(x,y) = 1.0000000e+0
y - 1.0500000e+1 ≈ 1.0000000e+0 * (x - 3.0500000e+1)
x - 3.0500000e+1 ≈ 1.0000000e+0 * (y - 1.0500000e+1)

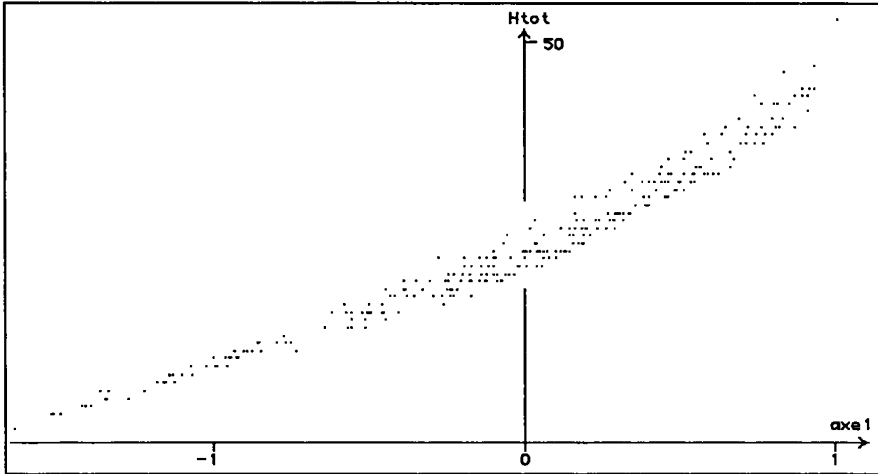
```

Nous publions face à face le déroulement du dialogue et le listage des résultats obtenus dans le cas où la colonne de poids est p100. Le lecteur remarquera qu'il y a une suite d'alinéas de même structure: nous soulignons que c'est la première des deux variables nommées qui seule peut être assujétie à varier entre des bornes. Il va sans dire que dialogue et listage concordent; à ceci près que sont affichés à l'écran les totaux bruts, ptot, des poids des éléments retenus (i.e. ceux pour lesquels la première variable satisfait aux conditions éventuellement imposées); tandis que sur le listage figure seulement le pourcentage de la masse totale retenue.

Quant aux résultats, on ne s'étonnera pas de trouver (quels que soient les poids choisis) des coefficients de corrélation égaux à (+1) ou (-1) et des formules de régression parfaites quand x est assujétie à varier dans un sous-intervalle où y en est une fonction linéaire exacte. Il vaut seulement la peine de considérer les résultats obtenus sur l'ensemble de l'intervalle.

poids	pcentre %	corr(x,y)	y ≈ a*x
p1	20.8 %	8.51 e-1	a = 4.14 e-1
p10	72.4 %	5.69 e-1	a = 3.08 e-1
p100	96.3 %	-3.26 e-1	a = -2.51 e-1
pmil	99.6 %	-8.93 e-1	a = -8.58 e-1

Il est naturel qu'avec le poids *pmil* l'intervalle central prédomine, la droite de régression globale (de pente a) lui étant à peu près ajustée; ce qui ne s'accorde aucunement avec les valeurs latérales. Celles-ci s'imposent au contraire quand le poids est constant (p1). Pour les poids p10 et p100, les résultats n'ont aucun intérêt du point de vue de la régression linéaire. Il faut particulièrement noter qu'avec le poids p10, alors que l'intervalle central a un poids relatif de 72.4%, l'ajustement de la droite de régression est dominé par les valeurs latérales. C'est pourquoi on a choisi de figurer cette droite avec le graphique de $y = f(x)$.



3 Régression linéaire du total *Htot* de 12 notes en fonction du facteur F1 issu de l'analyse de ces notes codées linéairement par morceaux

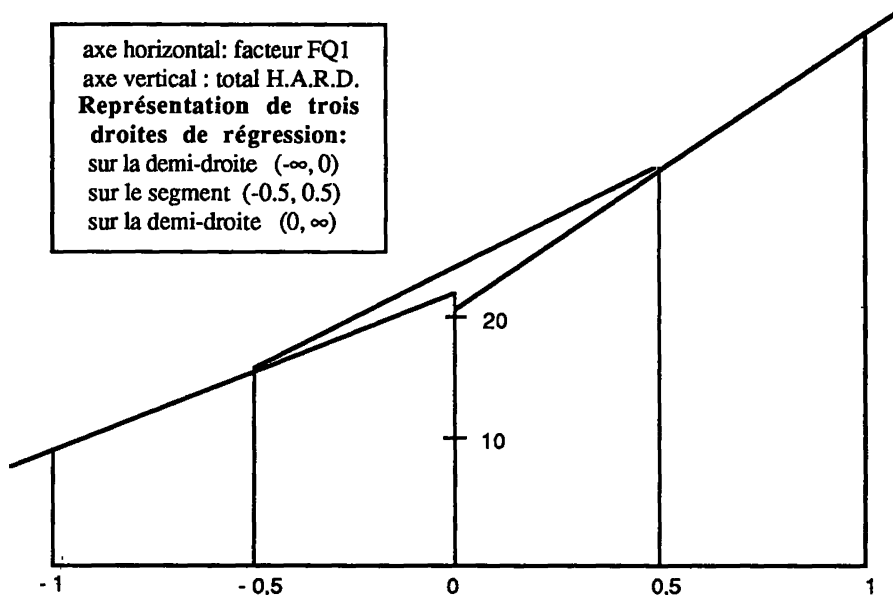
Les données de base constituent un tableau à 320 lignes et 13 colonnes: 320 "diagrammes HARD", comprenant chacun 12 notes mesurant divers symptômes dépressifs (Tristesse, Angoisse, Lassitude,...) et une colonne 'Htot', ou total. Les données ont été analysées après codage de chaque note suivant 3 modalités {-,=,+} (cf. [ÉCHELLES], op. laud. §2.1).

```

ce programme cree un tableau à partir de 2 autres
par juxtaposition(J) ou superposition(U)
fait il juxtaposer(J) ou superposer(U) 2 tableaux J
le nom du premier tableau a prendre est *:mxH4
*:mxH4ww
ce nom est il confirme oui(O) ou non(N) 0
le nom du deuxieme tableau a prendre est *:mxH4QiFac
*:mxH4QiFacww
ce nom est il confirme oui(O) ou non(N) 0
le suffixe du tableau a creer est +QF
+QF
ce sigle est il confirme oui(O) ou non(N) 0
*:mxH4ww
titre du premier tableau:
depression de l'adulte
*:mxH4QiFacww
titre du deuxieme tableau:
i:Qww:depression de l'adulte

```

En vue d'étudier, par 'corel' le lien entre la note totale *Htot* et le facteur F1 issu de l'analyse, on a, par 'jxtab', juxtaposé le tableau de base (à 13 colonnes) et le tableau 'iFac' qui contient les facteurs obtenus en analysant les données recodées barycentriquement (d'où la présence de la lettre 'Q' dans le nom du tableau des facteurs: cf. [CODAGE LIN.], in *CAD*, Vol XIV, n°2 pp. 203-210,



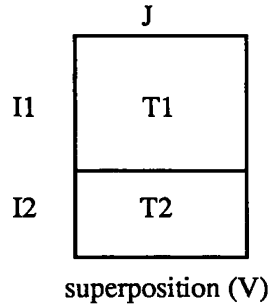
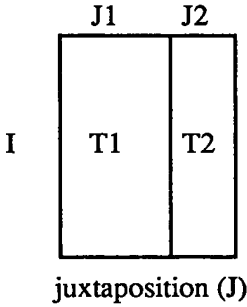
1989). Il va sans dire qu'on peut sur le tableau ainsi créé explorer toute relation entre variable de base et facteur, mais nous nous bornerons à Htot et axe1.

```

depression de l'adulte (1:mxH4+QF ; relx; 320indiv de poids 1)
corr(axe1,Htot) = 9.7648260e-1
Htot - 2.4093750e+1 = 1.6010971e+1 * (axe1 - 2.1900632e-10)
axe1 - 2.1900632e-10 = 5.9554056e-2 * (Htot - 2.4093750e+1)
bornes de axe1 : inf = -1.0000000e+1 ; sup = 0.0000000e+0
poids de l'intervalle retenu = 4.4375000e-1
corr(axe1,Htot) = 9.8126164e-1
Htot - 1.4654930e+1 = 1.2864991e+1 * (axe1 - -5.8468994e-1)
axe1 - -5.8468994e-1 = 7.4844548e-2 * (Htot - 1.4654930e+1)
bornes de axe1 : inf = 0.0000000e+0 ; sup = 1.0000000e+1
poids de l'intervalle retenu = 5.5625000e-1
corr(axe1,Htot) = 9.5478851e-1
Htot - 3.1623596e+1 = 2.1828388e+1 * (axe1 - 4.6643805e-1)
axe1 - 4.6643805e-1 = 4.1763098e-2 * (Htot - 3.1623596e+1)
bornes de axe1 : inf = -5.0000000e-1 ; sup = 5.0000000e-1
poids de l'intervalle retenu = 5.7187500e-1
corr(axe1,Htot) = 9.3949085e-1
Htot - 2.4005464e+1 = 1.6433024e+1 * (axe1 - 6.1265490e-2)
axe1 - 6.1265490e-2 = 5.3711542e-2 * (Htot - 2.4005464e+1)

```

Comme le précise le listage 'relx', on a effectué des régressions linéaires de Htot en fonction de axe1 sans restriction sur axe1, puis sur les intervalles $(-\infty, 0)$, $(-0.5, +0.5)$ et $(0, \infty)$. On a ainsi pu tracer 3 droites de régression partielle qu'on comparera utilement au nuage des points placé en regard. On a également fait des calculs de corrélation entre Htot et l'axe1 issu de l'analyse des diagrammes HARD codés avec 7 modalités par note: les corrélations sont alors plus élevées.



NB1. Comme l'indique le début du dialogue, le programme 'jxtab' permet aussi de superposer deux tableaux T1 et T2. Il faut encore signaler que 'jxtab' effectue diverses vérifications. Pour être superposables, T1 et T2 doivent avoir même nombre de colonnes et les sigles des colonnes doivent coïncider. Pour être juxtaposables, T1 et T2 doivent avoir même nombre de ligne; et les sigles des lignes doivent coïncider; comme les sigles des lignes ne sont pas énumérés en tête des tableaux, la vérification se fait au fur et à mesure que sont lues les lignes de T1 et T2, pour créer celles du tableau somme.

```
faut il juxtaposer(J) ou superposer(V) 2 tableaux V
le nom du premier tableau a prendre est *:mxH4
*:mxH4ww
ce nom est il confirme oui(O) ou non(N) 0
le nom du deuxieme tableau a prendre est *:mxH4QiFac
*:mxH4QiFacww
ce nom est il confirme oui(O) ou non(N) 0
le suffixe du tableau a creer est /Q
/Q
ce sigle est il confirme oui(O) ou non(N) 0
*:mxH4ww
titre du premier tableau:
depression de l'adulte
*:mxH4QiFacww
titre du deuxieme tableau:
i:Qww:depression de l'adulte
ERREUR les tableaux n ont pas le meme nombre de colonnes
```

NB2. Pour être acceptés par 'jxtab', deux tableaux doivent être tous deux en format 'ww' (ou, éventuellement, tous deux en format 'zz'; cf. [NOTE CRÉ. TAB.], in CAD, Vol XIV, n°1, 1989, §1.3). Pour combiner deux tableaux de formats différents, il suffit de recourir au programme 'zrang' qui copie tout tableau en lui donnant le format 'ww'. On voit sur les copies d'écran que pour 'corel' et 'jxtab', comme pour 'zrang', 'soustab', 'qori',... les noms de tableaux sont donnés sans suffixe. Le nom du tableau créé par 'jxtab' est celui du premier tableau donné, complété par un suffixe (sigle limité à 4 lettres) que choisit l'utilisateur.