JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LINDA VITTORIA DE CAROLIS

Le rapport de corrélation multiple et ses applications

Journal de la société statistique de Paris, tome 133, n° 1-2 (1992), p. 98-105

<http://www.numdam.org/item?id=JSFS_1992__133_1-2_98_0>

LE RAPPORT DE CORRÉLATION MULTIPLE ET SES APPLICATIONS[1]

par Linda Vittoria DE CAROLIS
Professeur à l'Université de Rome

Résumé

Dans cet article on introduit le rapport de corrélation multiple qui généralise à k ($k > 2$) caractères le rapport de corrélation de Pearson. Particulièrement on met en évidence que en utilisant le rapport de corrélation multiple dans le procédé stepwise on obtient une amélioration des résultats.

Abstract

In this research the multiple correlation ratio, which extends to k variables ($k > 2$) the Pearson's correlation ratio, is introduced. Particularly it is made evident that the multiple correlation ratio utilization in the stepwise procedure allows for better results.

1. Introduction

Considérons k caractères : le rapport de corrélation multiple que nous proposons permet premièrement d'étudier la dépendance en moyenne d'un caractère quantitatif, que nous notons avec le symbole ${}_kX$, de l'ensemble des $k-1$ caractères résiduels $({}_1X, \ldots, {}_{k-1}X)$, qui peuvent être qualitatifs aussi. En outre, en supposant que les $k-1$ caractères résiduels soient quantitatifs, ils peuvent être employé dans le procédé stepwise. A cet égard si nous désirons exprimer ${}_kX$ en fonction de $r \leqslant k-1$ variables explicatives, il s'agit de spécifier, parmi les $k-1$ caractères résiduels, ceux qui effectivement influent sur ${}_kX$. Au moyen du procédé stepwise on obtient ce résultat en supposant une liaison linéaire entre ${}_kX$ et les variables résiduelles [4]. Pour cela cette méthode présente le désavantage d'omettre l'examen d'une dépendance mathématique fonctionnelle non linéaire. Par conséquent, il faut conduire une étude plus générale qui peut être effectuée, comme nous le proposons, en employant la fonction de régression et le rapport de corrélation multiple. De cette façon, on a une amélioration des résultats.

1. Certains résultats de cette communication ont été présentés par L.V. de Carolis lors de la 47e session de l'Institut International de Statistique, Paris, 29 août – 6 septembre 1989.

2. Le rapport de corrélation multiple

Considérons tout d'abord le cas de trois caractères quantitatifs (ou variables) ${}_1X$, ${}_2X$, ${}_3X$ dont on a effectué n observations. Si l'on désire mettre en évidence une éventuelle interprétation de ${}_3X$ au moyen de l'un ou de tous les deux caractères résiduels, on commence par ranger les n données dans le tableau numéro 1. On spécifie que ${}_1X$ et ${}_2X$ présentent trois modalités chacun et ${}_3X$ deux. On considère ce cas soit pour fixer l'attention soit en vue de l'application du rapport de corrélation multiple au procédé stepwise comme nous verrons dans la suite. Dans ce tableau, on a noté le nombre d'ordre des modalités des caractères en bas et à droite de la lettre qui les contressigne et les fréquences avec le symbole n_{ijh} $(i=1,2;\ j,\ h=1,2,3)$ où les indices i,j,h se refèrent respectivement aux modalités de ${}_3X$, ${}_1X$ et ${}_2X$. Du tableau numéro 1, on obtient la distribution marginale de ${}_3X$ (tableau numéro 2). On obtient en outre les neuf distributions conditionnelles de ${}_3X$ liées aux couples des modalités $({}_1X_j,\ {}_2X_h)$ (tableau numéro 3).

On a les liaisons suivantes concernant les fréquences :

$$n_{ij.}=\sum_{h=1}^{3} n_{ijh}\ ;\quad n_{i..}=\sum_{j=1}^{3} n_{ij.}\ ;\quad n=\sum_{i=1}^{2} n_{i..}=\sum_{i,j,h} n_{ijh}\ ;\quad n_{.jh}=\sum_{i=1}^{2} n_{ijh}$$

Remarquons que les distributions conditionnelles sont autant que les produits des modalités de ${}_1X$ et ${}_2X$. En nous référant aux distributions 1, 2, 3 on obtient la liaison d'indépendance de ${}_3X$ du couple $({}_1X,{}_2X)$ ci-après :

$$n_{ijh}=\frac{n_{i..}n_{.jh}}{n}\quad (i=1,2;\ j,\ h=1,2,3) \tag{1}$$

Considérons donc les moyennes ${}_3\overline{X}_{jh}$ des distributions conditionnelles du tableau numéro 3 :

$${}_3\overline{X}_{jh}=\frac{\sum_{i=1}^{2} {}_3X_i n_{ijh}}{n_{.jh}};\quad (j,\ h=1,2,3) \tag{2}$$

Si nous notons ${}_3\overline{X}$ la moyenne arithmétique de la distribution marginale de ${}_3X$ on peut démontrer [3] qu'elle est aussi la moyenne arithmétique des moyennes conditionnelles pondérée avec le nombre total des observations des distributions conditionnelles respectives, on a donc :

$$\frac{\sum_{j,h} {}_3\overline{X}_{jh}n_{.jh}}{\sum_{j,h} n_{.jh}}=\frac{\sum_{i=1}^{2} {}_3X_i n_{i..}}{n}={}_3\overline{X} \tag{3}$$

Par conséquent si les moyennes conditionnelles sont égales, concluons que ${}_3X$ est indépendant en moyenne du couple $({}_1X,\ {}_2X)$ si ceci ne se vérifie pas, il y a

Caractères	*Modalités*																	
${}_3X$	${}_3X_1$									${}_3X_2$								
${}_1X$	${}_1X_1$			${}_1X_2$			${}_1X_3$			${}_1X_1$			${}_1X_2$			${}_1X_3$		
${}_2X$	${}_2X_1$	$2X_2$	$2X_3$	$2X_1$	$2X_2$	$2X_3$	$2X_1$	$2X_2$	$2X_3$	$2X_1$	$2X_2$	$2X_3$	$2X_1$	$2X_2$	$2X_3$	$2X_1$	$2X_2$	$2X_3$
Fréquences n_{ijh}	n_{111}	n_{112}	n_{113}	n_{121}	n_{122}	n_{123}	n_{131}	n_{132}	n_{133}	n_{211}	n_{212}	n_{213}	n_{221}	n_{222}	n_{223}	n_{231}	n_{232}	n_{233}
Totaux $n_{ij.}$	$n_{11.}$			$n_{12.}$			$n_{13.}$			$n_{21.}$			$n_{22.}$			$n_{23.}$		
Totaux $n_{i..}$	$n_{1..}$									$n_{2..}$								
Total	n																	

Tableau 1. Distribution des caractères ${}_1X$, ${}_2X$, ${}_3X$

${}_3X$	${}_3X_1$	${}_3X_2$	Total
$n_{i..}$	$n_{1..}$	$n_{2..}$	n

Tableau 2. Distribution marginale de ${}_3X$

${}_3X$	${}_3X_1$	${}_3X_2$	Total
n_{ijh}	n_{1jh}	n_{2jh}	$n_{.jh}$

Tableau 3. Distribution conditionnelle de ${}_3X$ liée au couple $({}_1X_j, {}_2X_h)$

dépendance en moyenne. Observons que si ${}_3X$ est indépendant du couple $({}_1X,\ {}_2X)$, il est aussi indépendant en moyenne mais le vice-versa n'est pas vrai. Si donc on désire mesurer le degré de dépendance en moyenne, on peut considérer l'indice relatif donné par la racine carrée positive du rapport entre $\sigma_3\overline{X}^2$ (variance des moyennes conditionnelles) donnée par l'expression :

$$\sigma^2{}_3\overline{X} = \frac{\sum_{j,h} ({}_3\overline{X}_{jh} - {}_3\overline{X})^2 n_{.jh}}{n} \tag{4}$$

et son maximum σ_3X^2 (variance marginale) donnée par l'expression :

$$\sigma^2{}_3X = \frac{\sum_{i=1}^{2} ({}_3X_i - {}_3\overline{X})^2 n_{i..}}{n} \tag{5}$$

Ce rapport est le rapport de corrélation à trois variables de ${}_3X$ en fonction du couple $({}_1X,\ {}_2X)$ que nous notons $\mathcal{M}_3X\,({}_1X,\ {}_2X)$, on a donc :

$$0 \leqslant \mathcal{M}_3X\,({}_1X, {}_2X) = \sqrt{\frac{\sigma^2{}_3\overline{X}}{\sigma^2{}_3X}} = \sqrt{\frac{\sum_{j,h} ({}_3\overline{X}_{jh} - {}_3\overline{X})^2 n_{.jh}}{\sum_{i} ({}_3X_i - {}_3\overline{X})^2 n_{i..}}} \leqslant 1 \tag{6}$$

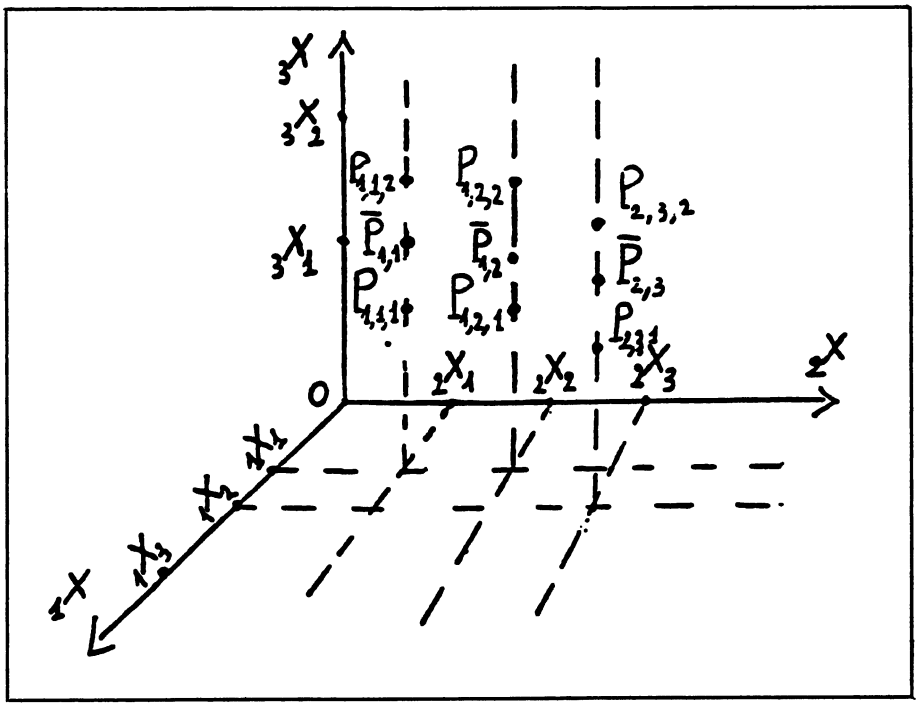
Le rapport de corrélation à trois variables est égal à zéro si toutes les moyennes conditionnelles sont égales et par conséquent ${}_3X$ est indépendant en moyenne du couple $({}_1X,\ {}_2X)$. Il est égal à un quand les moyennes conditionnelles coïncident avec les modalités de ${}_3X$; en ce cas la distribution à trois variables présente le maximum de la dépendance de ${}_3X$ de $({}_1X,\ {}_2X)$; par conséquent le maximum de la dépendance en moyenne coïncide avec le maximum de la dépendance et vice-versa. En généralisant au cas de k caractères, on a que le rapport de corrélation multiple de ${}_kX$ en fonction de l'ensemble des $k-1$ variables résiduelles est donné par l'expression :

$$0 \leqslant \mathcal{M}_kX\,({}_1X,\ \ldots,\ {}_{k-1}X) = \sqrt{\frac{\sigma^2{}_k\overline{X}}{\sigma^2{}_kX}} = \sqrt{\frac{\sum_{h_1\ldots h_{k-1}} ({}_k\overline{X}_{h_1\ldots h_{k-1}} - {}_k\overline{X})^2 n_{h_1\ldots h_{k-1}}}{\sum_{i=1}^{2} ({}_kX_i - {}_k\overline{X})^2 n_{i..}}} \leqslant 1 \tag{7}$$

Où ${}_k\overline{X}_{h_1\ldots k_{-1}k-1}$ est la moyenne conditionnelle de ${}_kX$ liée à l'ensemble des $k-1$ modalités $({}_1X_{h_1},\ \ldots,\ {}_{k-1}X_{h_{k-1}})$, $h_1\ldots h_{k-1}$ est l'une des 3^{k-1} dispositions avec répétition des numéros 1, 2, 3 (qui sont les numéros d'ordre des modalités de chaque variable résiduelle) pris $k-1$ à $k-1$ et la sommation $\sum_{h_1\ldots h_{k-1}}$ est étendue à toutes ces dispositions avec répétitions.

La formule (7) évidemment peut être généralisée en considérant que les k caractères aient respectivement s_1, ..., s_k modalités. En ce cas la sommation $\sum_{h_1 \ldots h_{k-1}}$ à numérateur est étendue à tous les s_1, s_2, ..., s_{k-1} ensembles de $k-1$ numéros h_1, h_2, ..., h_{k-1} obtenus en associant les numéros d'ordre (1, 2, ..., s_1) des modalités de ${}_1X$ avec ceux des $k-2$ caractères ${}_2X$, ..., ${}_{k-1}X$ et la sommation au dénominateur est par rapport à l'indice i ($i = 1, 2, \ldots, s_k$). Naturellement la (7) est valable si ${}_kX$ est quantitatif et les $k-1$ caractères résiduels sont de nature quelconque.

Pour utiliser le rapport de corrélation multiple dans le procédé stepwise, il convient de considérer l'interprétation géométrique de ce que nous venons d'exposer. En nous référant tout d'abord au cas de trois variables, si nous représentons la distribution du tableau numéro 1 dans l'espace euclidien à trois dimensions, on a la figure suivante :



Remarquons que l'on a 18 points distincts $p_{j,h,i}$ ayant coordonnées (${}_1X_j$, ${}_2X_h$, ${}_3X_i$), chacun répété n_{ijh} fois, que nous appelons points effectifs. Il y a en plus neuf points de régression $\overline{P}_{j,h}$ de ${}_3X$ en (${}_1X$, ${}_2X$) ayant pour coordonnées (${}_1X_j$, ${}_2X_h$, ${}_3\overline{X}_{jh}$).

On appelle surface de régression quelconque la surface qui contient les points de régression et son équation, où ${}_3X$ dépend du couple (${}_1X$, ${}_2X$), est la fonction de régression de ${}_3X$ en (${}_1X$, ${}_2X$).

Dans la suite nous prendrons comme fonction de régression une fonction

rationnelle entière. On démontre [3] que le carré du rapport de corrélation de ${}_3X$ en $({}_1X,\ {}_2X)$ est donné aussi par l'expression :

$$0 \leqslant \mathcal{M}^2\,{}_3X\,({}_1X,\ {}_2X) = 1 - \frac{\overline{\sigma}^2\,{}_3X}{\sigma^2\,{}_3X} \leqslant 1 \tag{8}$$

où

$$\sigma^2\,{}_3X = \frac{\sum_{i,j,h} ({}_3X_i - {}_3\overline{X}_{jh})^2 n_{ijh}}{n} \tag{9}$$

est la variance conditionnelle moyenne.

Par conséquent $\mathcal{M}^2\,{}_3X\,({}_1X,\ {}_2X)$ est un indice relatif qui mesure le degré de convergence des points effectifs vers les points de régression et précisément il vaut zéro si les points de régression appartiennent au plan d'équation ${}_3X = {}_3\overline{X}$;en ce cas il y a le maximum de divergence entre les points effectifs et ceux de régression ; il vaut un si les points effectifs et ceux de régression coïncident. Si en outre nous ajustons à la distribution une fonction rationnelle entière avec la méthode des moindres carrés on démontre [3] que la fonction de régression est la fonction la meilleure qu'on puisse ajuster à la distribution avec la méthode des moindres carrés. On peut évidemment généraliser ces résultats au cas de k variables.

3. Applications

Parlons maintenant de l'utilisation du rapport de corrélation à k variables dans le procéde stepwise. On commence donc par considérer les $k-1$ distributions à deux variables des caractères ${}_kX\,{}_iX$ (où $i = 1,\ ...,\ k-1$) et les moyennes conditionnelles ${}_k\overline{X}_t$ (où $t = 1, 2, 3$) des distributions de ${}_kX$ liées par les modalités ${}_iX_t$ de ${}_iX$. Observons qu'en ce cas, en nous référant à chaque variable résiduelle, on a trois points de régression $\overline{P}_t$ ayant pour coordonnées $({}_iX_t,\ {}_k\overline{X}_t)$. Le carré du rapport de corrélation de ${}_kX$ en ${}_iX$, $\mathcal{M}^2\,{}_kX,\ {}_iX$, mesure l'intensité de l'explication de ${}_kX$ par ${}_iX$ au moyen de la fonction de régression donnée par l'équation de la parabole suivante :

$${}_kX = c_0 + c_1\,{}_iX + c_2\,{}_iX^2 \tag{10}$$

dont les coefficients peuvent être obtenus ou en imposant le passage par les trois points de régression ou en employant la méthode des moindres carrés. On peut justifier à ce point l'hypothèse que chaque caractère résiduel ait trois modalités au minimum parce qu'en considérant deux modalités on aurait deux points de régression et comme fonction de régression l'équation d'une droite au lieu de celle de la parabole (10) et par conséquent dans cette première phase on retomberait dans le procédé stepwise traditionnel. Observons que, en général, il est convenable de conserver le nombre de modalités ici proposé pour les k caractères pour avoir des distributions conditionnelles ayant assez d'observations. En reprenant le procédé stepwise, après avoir déterminé les $k-1$ rapports de corrélation de ${}_kX$ en ${}_iX$, $\mathcal{M}_kX,\ {}_iX$, on considère comme premier caractère explicatif ${}_{j_1}X$ si $\mathcal{M}^2\,{}_kX,\ {}_{j_1}X$ a la

valeur plus grande. Pour améliorer l'explication on passe ensuite à choisir un deuxième caractère ${}_{j2}X$ si $\mathcal{M}^2{}_kX,\ {}_{j_2}X$ est immédiatement inférieur ou égal à $\mathcal{M}^2{}_kX,\ {}_{j_1}X$. On considère donc les neuf distributions conditionnelles de ${}_kX$ liées aux couples de modalités de ${}_{j_1}X$ et ${}_{j_2}X$ et le rapport de corrélation à trois variables $\mathcal{M}_kX({}_{j_1}X,\ {}_{j_2}X)$ dont le carré est une mesure de l'intensité de l'explication de ${}_kX$ par le couple $({}_{j_1}X,\ {}_{j_2}X)$ au moyen de la fonction de régression, en ce cas l'équation d'un paraboloïde de l'espace à trois dimensions. J'ai démontré [3] que, en augmentant le nombre des variables explicatives, le rapport de corrélation est non décroissant. Par conséquent pour évaluer l'amélioration obtenue en ajoutant ${}_{j_2}X$ on peut employer l'indice d'amélioration suivant (11), que nous notons ${}_{j_1}X^{\varepsilon}{}_{j_2}X$, qui utilise les rapports de corrélation à deux variables et à trois variables

$$0 \leqslant {}_{j_1}X^{\varepsilon}{}_{j_2}X = \frac{\mathcal{M}^2{}_kX\,({}_{j_1}X,\ {}_{j_2}X) - \mathcal{M}^2{}_kX,\ {}_{j_1}X}{1 - \mathcal{M}^2{}_kX,\ {}_{j_1}X} \leqslant 1 \tag{11}$$

Cet indice vaut zéro si le numérateur est nul, en ce cas il n'y a aucune amélioration dans l'explication de ${}_kX$ en ajoutant à ${}_{j_1}X$ le caractère ${}_{j_2}X$. L'indice vaut un si $\mathcal{M}^2{}_xX\,({}_{j_1}X,\ {}_{j_2}X)$ est égal à un, en ce cas les points effectifs appartiennent au paraboloïde de régression et par conséquent l'entrée en jeu de ${}_{j_2}X$ permet d'expliquer parfaitement ${}_kX$ au moyen de la fonction de régression. Dans tous les autres cas si l'on a $0{,}5 < {}_{j_1}X^{\varepsilon}{}_{j_2}X < 1$ on inclut ${}_{j_2}X$ parmi les variables explicatives et on passe à considérer une troisième variable ${}_{j_3}X$ telle que $\mathcal{M}^2{}_kX\,({}_{j_1}X,\ {}_{j_3}X)$ soit immédiatement inférieur ou égal à $\mathcal{M}^2{}_kX\,({}_{j_1}X,\ {}_{j_2}X)$. En continuant de cette façon, en ajoutant une à la fois une nouvelle variable et en supposant d'obtenir successivement des valeurs de l'indice d'amélioration supérieures à 0,5, le procédé s'achève quand, après $r-1$ variables explicatives, l'introduction d'une $r^{\text{-ième}}$ variable ne produit aucune amélioration appréciable en considérant l'indice ${}_{j_{r-1}}X^{\varepsilon}{}_{j_r}X$ suivant qu'on utilise les rapports de corrélation à r variables et à $r+1$ variables :

$${}_{j_{r-1}}X^{\varepsilon}{}_{j_r}X = \frac{\mathcal{M}^2{}_kX\,({}_{j_1}X, {}_{j_2}X, \ldots, {}_{j_r}X) - \mathcal{M}^2{}_kX\,({}_{j_1}X, {}_{j_2}X, \ldots, {}_{j_{r-1}}X)}{1 - \mathcal{M}^2{}_kX\,({}_{j_1}X, {}_{j_2}X, \ldots, {}_{j_{r-1}}X)} \tag{12}$$

Observons que le rapport de corrélation à $r+1$ variables ($r \leqslant k-1$) utilise 3^r moyennes conditionnelles et il est donné par la formule (13) issue de la (7) :

$$\mathcal{M}_kX\,({}_{j_1}X,\ {}_{j_2}X,\ \ldots,\ {}_{j_r}X) = \sqrt{\frac{\sum\limits_{h_1 \ldots h_r} ({}_k\overline{X}_{h_1 \ldots h_r} - {}_k\overline{X})^2 n_{h_1 \ldots h_r}}{\sum\limits_{i=1}^{2} ({}_kX_i - {}_k\overline{X})^2 n_{i\ldots}}} \tag{13}$$

où la $\sum\limits_{h_1 \ldots h_r}$ est étendue à toutes les 3^r dispositions avec répétitions des numéros (1, 2, 3). Remarquons en outre que l'on a 3^r points de régression dans l'espace à $r+1$ dimensions. Par conséquent ${}_kX$ sera expliqué par les r variables au moyen de la fonction de régression, qui est en ce cas l'équation d'un hyperparaboloïde [2] de degré h de l'espace à $r+1$ dimensions, qui contient les 3^r points.

Outre le procédé stepwise, le rapport de corrélation multiple, en mesurant la dépendance en moyenne d'une variable d'un ensemble de caractères de nature

quelconque, peut trouver des nombreuses applications particulièrement dans la statistique appliquée à la médecine. A cet égard je suis en train de faire une application de ce qui vient d'être exposé dans une recherche sur les causes possibles des allergies.

On peut ajouter qu'il est possible de tester la signification du rapport de corrélation à k caractères, formule (7) (cas d'un nombre quelconque de modalités) en généralisant le procédé employé pour deux [5]. En effet le rapport suivant :

$$F = \frac{g_2}{g_1} \frac{\sigma^2{}_k\overline{X}}{\overline{\sigma}^2{}_kX} = \frac{g_2}{g_1} \frac{\mathcal{M}^2{}_kX({}_1X, \ldots, {}_{k-1}X)}{1-\mathcal{M}^2{}_kX({}_1X, \ldots, {}_{k-1}X)}, \tag{14}$$

où $g_1 = s_1 \cdot s_2 \ldots s_{k-1} - 1$ et $g_2 = n - s_1 \cdot s_2 \ldots s_{k-1}$ sont les degrés de liberté respectivement de $\sigma^2{}_k\overline{X}$, variance des moyennes conditionnelles et de $\overline{\sigma}^2{}_kX$, variance conditionnelle moyenne, est une détermination de la variable F de Fisher-Snédécor.

Particulièrement, dans le cas du procédé stepwise, en nous référant au rapport de corrélation (13), on a :

$$F = \frac{n-3^r}{3^r-1} \frac{\mathcal{M}^2{}_kX({}_{j_1}X, \ldots, {}_{j_r}X)}{1-\mathcal{M}^2{}_kX({}_{j_1}X, \ldots, {}_{j_r}X)} \tag{15}$$

Le test (15) peut être employé au lieu de l'indice d'amélioration ${}_{j_{r-1}}X\,\varepsilon\,{}_{j_r}X$ (12). En effet, après avoir déterminé le rapport de corrélation (13), qui se réfère à l'inclusion d'un nouveau caractère ${}_{j_r}X$ si le test F (15) est significatif, nous ajouterons ${}_{j_r}X$ aux variables explicatives, sinon nous le repousserons.

Je conclus en faisant remarquer que, en vue des applications du rapport de corrélation multiple, on a régigé un software [1], pour le moment limité au rapport de corrélation à trois caractères et à son utilisation dans le procédé stepwise.

BIBLIOGRAPHIE

(1) ANZALONE P., MANCINI P. (1991), *Il rapporto di correlazione tripla nella procedura stepwise : un metodo di calcolo computerizzato*. Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma, Serie A-Ricerche.

(2) DE CAROLIS L.V. (1987), *Geometry of the Random Variables. Italian Contributions to the Methodology of Statistics*, CLEUP, Padova.

(3) DE CAROLIS L.V. (1989), *Il rapporto di correlazione multipla nella procedura stepwise*. Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma, Serie A-Ricerche.

(4) EFROYMSON M.A. (1960), Multiple regression analysis *in* Ralston A. and Wilf H.S. (eds), *Mathematical Methods for Digital Computers*, Wiley & Sons, New York, 191-203.

(5) MORICE E., CHARTIER F. (1954), *Méthode statistique*. Imprimerie Nationale, Paris, 347-348.