Statistique et analyse des données

Aziz Lazraq
Robert Cléroux

Un algorithme pas à pas de sélection de variables en régression linéaire multivariée

Statistique et analyse des données, tome 13, n° 1 (1988), p. 39-58
<http://www.numdam.org/item?id=SAD_1988__13_1_39_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 1 - p. 39-58

UN ALGORITHME PAS A PAS DE SELECTION DE VARIABLES EN REGRESSION LINEAIRE MULTIVARIEE

par

Aziz Lazraq et Robert Cléroux

Département d'informatique et de recherche opérationnelle, Université de Montréal.

Recherche subventionnée par le Conseil de recherches en sciences naturelles et en génie du Canada et par la fondation FCAR (Gouvernement du Québec)

Manuscrit reçu le 4.1.88, révisé le 24.7. 88

RESUME

Dans cet article on propose un algorithme pas à pas de sélection de variables en régression linéaire multivariée ou en analyse en composantes principales par rapport à des variables instrumentales. On établit les distributions nécessaires de l'indice de redondance de Stewart et Love et d'un indice partiel, ce qui permet d'aborder la sélection des variables. A chaque étape où une variable est retenue on effectue un test statistique afin de déterminer si elle est significative. Dans l'affirmative, toutes les variables préalablement retenues sont remises en cause en effectuant d'autres tests statistiques. La procédure se termine lorsqu'aucune variable n'est retenue et qu'aucune variable n'est éliminée.

ABSTRACT

In this paper, a stepwise algorithm is proposed for selecting variables in the multivariate linear regression model. We obtain the distributions of a function of Stewart and Love's redundancy index and of a function of a related partial index, and this allows for variable selection. At each step where a variable is selected to enter the model, a test of hypothesis is performed in order to determine wether its contribution is significant or not. If it is, all the variables previously entered in the model are in turn tested for significance. The procedure stops when no variable enters the model and no variable is removed from the model.

MOTS CLES

Sélection de variables, algorithme pas à pas, sélection progressive, élimination successive, régression linéaire multivariée, A.C.P.V.I.

INDICES DE CLASSIFICATION STMA

06:010, 06:900, 07:070

1. INTRODUCTION

Dans cet article nous nous intéressons au problème de sélection de variables dans un modèle de régression linéaire multivariée ou d'analyse en composantes principales par rapport à des variables instrumentales (A.C.P.V.I.). Le rapport entre ces deux méthodes est indiqué dans la prochaine section. Plusieurs auteurs ont récemment traité de ce problème en analyse des données; voir par exemple Bonifas, Escoufier, Gonzalez et Sabatier (1984), Gonzalez et Chami (1984), Escoufier (1986) et Dambroise, Escoufier et Massotte (1987) ainsi que les références citées par eux.

Nous proposons, dans le présent article, une méthode pas à pas de sélection de variables basée sur l'indice de redondance de Stewart et Love (1968). A chaque étape où une variable est choisie on effectue un test statistique afin de déterminer si elle est significativement non nulle. Dans l'affirmative, toutes les variables préalablement retenues sont remises en cause et on effectue un test statistique afin de déterminer si l'une d'elles est devenue redondante par suite de la dernière sélection. La procédure s'arrête lorsqu'aucune nouvelle variable n'est sélectionnée et qu'aucune variable n'est retranchée de la sélection.

Dans la section 2 on rappelle l'indice de redondance de Stewart et Love. Sa distribution est obtenue dans la section 3 dans le cas où la distribution des observations est multinormale. Dans la section 4 on introduit un indice de redondance partielle. Une relation de récurrence entre les indices de redondance et de redondance partielle est établie dans la section 5. La section 6 présente l'algorithme pas à pas de sélection de variables et finalement on traite un exemple dans la section 7.

2. L'INDICE DE REDONDANCE DE STEWART ET LOVE

Soit un vecteur aléatoire $\underline{Y} : p \times 1$ que nous chercherons à expliquer linéairement à partir du vecteur $\underline{X} : q \times 1$. Supposons que $\begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$ soit la matrice de covariance du vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ où $\Sigma_{yy} : p \times p$ et $\Sigma_{xx} : q \times q$. La matrice de covariance empirique de $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$, obtenue à partir d'un échantillon de taille n, est notée

$$\begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} A_{yy} & A_{yx} \\ A_{xy} & A_{xx} \end{bmatrix} = \frac{A}{n-1} \tag{2.1}$$

et l'indice de redondance de Stewart et Love (1968) est donnée par

$$RI = RI(\underline{Y}, \underline{X}) = \frac{tr(S_{yx} S_{xx}^{-1} S_{xy})}{tr(S_{yy})} \tag{2.2}$$

où tr désigne la trace d'une matrice. L'indice (2.2) peut également s'écrire sous la forme

$$RI = \frac{\sum_{i=1}^{p} s_{y_i}^2 R_{y_i x_1 x_2 \ldots x_q}^2}{\sum_{i=1}^{p} s_{y_i}^2} \tag{2.3}$$

où $s_{y_i}^2$ est la variance empirique de Y_i et $R_{y_i x_1 x_2 \ldots x_q}^2$ est le carré du coefficient empirique de corrélation multiple entre Y_i et $X_1, X_2, \ldots, X_q$. L'indice de redondance est donc une moyenne pondérée (par les variances) des carrés des coefficients de corrélation multiple entre les composantes du vecteur à prédire et le vecteur de prédiction. Par la suite on l'utilisera avantageusement dans un algorithme de sélection de variables. D'autres propriétés de RI se trouvent dans Lazraq et Cléroux (1987).

Le coefficient RV (Escoufier (1973), Robert et Escoufier (1976)) est une alternative à l'indice de Stewart et Love. Cependant, puisque sa distribution exacte est encore inconnue, on ne saurait l'utiliser dans un algorithme de sélection de variables avec inférence.

L'algorithme de sélection de variables proposé dans cet article est basé sur la maximisation de RI. C'est également un algorithme de sélection de variables en A.C.P.V.I. En effet selon Rao (1965), en A.C.P.V.I. on cherche à remplacer $\underline{Y}$ par $M'\underline{X}$, où $M : q \times p$ de telle manière que l'efficacité de la prédiction de $\underline{Y}$ par $\underline{X}$ soit maximale. Une des mesures proposées par l'auteur pour évaluer cette efficacité est la trace de la matrice résiduelle qui doit être la plus petite possible. Dans le contexte de la régression linéaire multivariée on doit avoir $M = S_{xx}^{-1} S_{xy}$ ($M'\underline{X}$ est le meilleur prédicteur de $\underline{Y}$ au sens des moindres carrés) et la matrice résiduelle est

$$S_{yy \cdot x} = S_{yy} - S_{yx} S_{xx}^{-1} S_{xy} \tag{2.4}$$

Egalement

$$tr(S_{yy \cdot x}) = tr(S_{yy}) \left[1 - \frac{tr(S_{yx} S_{xx}^{-1} S_{xy})}{tr(S_{yy})}\right] \tag{2.5}$$

$$= tr(S_{yy})(1 - RI). \tag{2.6}$$

et minimiser $tr(S_{yy \cdot x})$ est équivalent à maximiser RI.

3. DISTRIBUTION DE $RI / (1 - RI)$

Dans cette section nous obtenons la loi exacte de $RI / (1 - RI)$ dans le cas où les variables explicatives impliquées sont indépendantes du vecteur à expliquer, c'est-à-dire sous l'hypothèse nulle $H_0 : \sum_{yx} = 0$.

Il est connu que si le vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ est multinormal, alors la distribution de la matrice $A : (p + q) \times (p + q)$ définie par (2.1) est la distribution de Wishart avec paramètres $\sum$ et $n - 1$, notée $W_{p+q}(\sum, n - 1)$. Le Lemme suivant est démontré dans Mardia, Kent et Bibby (1979), p. 71. Il sera utile pour la suite.

Lemme (3.1) : Si $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ est multinormal, alors

a) $A_{yy\cdot x} = A_{yy} - A_{yx} A_{xx}^{-1} A_{xy}$ possède la distribution $W_p(\sum_{yy\cdot x}, n-1-q)$ où $\sum_{yy\cdot x} = \sum_{yy} - \sum_{yx} \sum_{xx}^{-1} \sum_{xy}$ et $A_{yy\cdot x}$ est indépendante de A_{xx} et de A_{yx} .

b) Sous $H_0 : \sum_{yx} = 0$, $A_{yy} - A_{yy\cdot x}$ possède la distribution $W_p(\sum_{yy}, q)$ et les matrices $A_{yx} A_{xx}^{-1} A_{xy}$, A_{xx} et $A_{yy\cdot x}$ sont conjointement indépendantes.

A partir de (2.2) on peut écrire

$$RI = \frac{tr(A_{yx} A_{xx}^{-1} A_{xy})}{tr(A_{yy\cdot x}) + tr(A_{yx} A_{xx}^{-1} A_{xy})}$$

d'où

$$\frac{RI}{1-RI} = \frac{tr(A_{yx} A_{xx}^{-1} A_{xy})}{tr(A_{yy\cdot x})} .$$

Par le Lemme précédent et sous H_0 on peut écrire

$$\frac{RI}{1-RI} = \frac{trE_1}{trE_2}$$

où E_1 possède la distribution $W_p(\sum_{yy}, q)$, E_2 possède la distribution $W_p(\sum_{yy}, n-1-q)$ et où E_1 et E_2 sont indépendantes.

On obtient maintenant la loi exacte de $RI / (1-RI)$. Par définition de la loi de Wishart, on peut écrire

$$E_1 = \sum_{i=1}^{q} Z_i Z_i' \quad \text{et} \quad E_2 = \sum_{i=q+1}^{n-1} Z_i Z_i'$$

où $Z_1, Z_2, \ldots, Z_{n-1}$ sont idd et $N(0, \sum_{yy})$. Egalement,

$$V_1 = trE_1 = \sum_{i=1}^{q} Z'_i Z_i \quad \text{et} \quad V_2 = trE_2 = \sum_{i=q+1}^{n-1} Z'_i Z_i .$$

Posons $Z' = (Z'_1, Z'_2, \ldots, Z'_{n-1}) : 1 \times p(n-1)$. Alors Z suit la loi $N(0, \Phi)$ où

$$\Phi = \begin{bmatrix} \sum_{yy} & 0 & \cdots & 0 \\ 0 & \sum_{yy} & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & \sum_{yy} \end{bmatrix} : p(n-1) \times p(n-1) .$$

Soient deux matrices $p(n-1) \times p(n-1)$ définies par

$$A = \begin{bmatrix} I_{pq} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{et} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & I_{p(n-1-q)} \end{bmatrix} .$$

Alors $V_1 = Z' AZ$, $V_2 = Z' BZ$ et la distribution de $RI / (1 - RI)$ est celle de $\dfrac{Z' AZ}{Z' BZ}$. Or,

$$p\left[\frac{Z' AZ}{Z' BZ} \leq r\right] = p[Z' AZ - rZ' BZ \leq 0] = p[Z' KZ \leq 0]$$

où $K = A - r B$. Mais $Z'KZ$ est distribué comme $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$ où les W_i sont iid et N(0, 1), et où les λ_i sont les valeurs propres de la matrice ΦK (voir Johnson et Kotz, (1970), p. 150).

Finalement, puisque

$$|\Phi K - \lambda I| = |\sum_{yy} - \lambda I|^q \, |-r\sum_{yy} - \lambda I|^{n-1-q} ,$$

les $p(n-1)$ valeurs propres de ΦK sont

(i) les p valeurs propres de $\sum_{yy}$, chacune ayant multiplicité q , et

(ii) les p valeurs propres de $-r \sum_{yy}$, chacune ayant multiplicité $n - 1 - q$.

Nous avons donc démontré le théorème suivant:

Théorème (3.1) :

Si le vecteur $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ est multinormal et, sous H_0 , $p[RI / (1 - RI) \leq r] = p[W \leq 0]$ où W est distribué comme $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$ où les W_i sont iid et $N(0, 1)$, et où $\lambda_1, \lambda_2, \ldots, \lambda_{p(n-1)}$ sont les p valeurs propres de $\sum_{yy}$ ayant chacune multiplicité q et les p valeurs propres de $-r\sum_{yy}$ ayant chacune multiplicité $n - 1 - q$.

En pratique, pour calculer $p[RI / (1 - RI) \leq r]$ on utilisera l'algorithme de Imhof (1961).

4. UN INDICE DE REDONDANCE PARTIELLE

Soit $Z = \begin{bmatrix} X_j \\ \underline{X}^{(2)} \end{bmatrix}$ où X_j est une composante quelconque du vecteur $\underline{X}$ et $\underline{X}^{(2)} : t \times 1$ est un sous-vecteur de $\underline{X}$ ne contenant pas X_j . Nous cherchons à définir l'indice de redondance partielle $RI_{\underline{X}^{(2)}}(\underline{Y}, X_j)$ entre le vecteur $\underline{Y}$ et la variable X_j après avoir éliminé de $\underline{Y}$ et de X_j l'effet linéaire de $\underline{X}^{(2)}$. En d'autres mots, $RI_{\cdot \underline{X}^{(2)}}(\underline{Y}, X_j)$ est l'indice de redondance entre les résidus des régressions linéaires de $\underline{Y}$ sur $\underline{X}^{(2)}$ et de X_j sur $\underline{X}^{(2)}$.

Afin de préciser davantage, soit $S = \begin{bmatrix} S_{yy} & S_{yx_j} \\ S_{x_j y} & s_{x_j x_j} \end{bmatrix}$ la matrice de covariance empirique du vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X}_j \end{bmatrix}$ et soit

$$S_{\cdot \underline{x}^{(2)}} = S - \begin{bmatrix} S_{y\underline{x}^{(2)}} \\ S_{x_j \underline{x}^{(2)}} \end{bmatrix} S^{-1}_{\underline{x}^{(2)}\underline{x}^{(2)}} (S_{\underline{x}^{(2)}y} \; S_{\underline{x}^{(2)}x_j}) \qquad (4.1)$$

la matrice de covariance partielle de $\begin{bmatrix} Y \\ X_j \end{bmatrix}$ après avoir éliminé de $\underline{Y}$ et de X_j l'effet linéaire de $\underline{X}^{(2)}$.

Nous obtenons donc de (4.1)

$$S_{yy\,x^{(2)}} = S_{yy} - S_{yx^{(2)}}\, S^{-1}_{x^{(2)}x^{(2)}}\, S_{x^{(2)}y}$$

$$s_{x_jx_j\cdot x^{(2)}} = s_{x_jx_j} - S_{x_jx^{(2)}}\, S^{-1}_{x^{(2)}x^{(2)}}\, S_{x^{(2)}x_j}$$

$$S_{yx_j\cdot x^{(2)}} = S_{yx_j} - S_{yx^{(2)}}\, S^{-1}_{x^{(2)}x^{(2)}}\, S_{x^{(2)}x_j}$$

et par conséquent l'indice de redondance partielle par

$$RI_{\cdot x^{(2)}} = RI_{\cdot x^{(2)}}\,(\underline{Y},\, X_j) = \frac{tr(S_{yx_j\cdot x^{(2)}}\, s^{-1}_{x_jx_j\cdot x^{(2)}}\, S_{x_jy\cdot x^{(2)}})}{tr(S_{yy\cdot x^{(2)}})} \tag{4.2}$$

Les propriétés de cet indice sont les suivantes:

a) si $\underline{Y}$ est une variable aléatoire Y , alors

$$RI_{\cdot x^{(2)}}(Y,\, X_j) = \frac{S_{yx_j\cdot x^{(2)}}\, S^{-1}_{x_jx_j\cdot x^{(2)}}\, S_{x_jy\cdot x^{(2)}}}{S_{yy\cdot x^{(2)}}} = \frac{S_{yx_j\cdot x^{(2)}}\, S_{x_jy\cdot x^{(2)}}}{S_{x_jx_j\cdot x^{(2)}}\, S_{yy\,x^{(2)}}} = \hat{\rho}^2_{yx_j\cdot x^{(2)}} \tag{4.3}$$

où $\hat{\rho}_{yx_j\cdot x^{(2)}}$ est le coefficient de corrélation partielle entre les variables Y et X_j après élimination de l'effet linéaire de $\underline{X}^{(2)}$.

b) à partir de (4.3) et par analogie à (2.3) on peut écrire

$$RI_{\cdot x^{(2)}}\,(\underline{Y},\, X_j) = \frac{\sum_{i=1}^{p} s^2_{y_i\cdot x^{(2)}}\, \hat{\rho}^2_{y_ix_j\,x^{(2)}}}{\sum_{i=1}^{p} s^2_{y_i\,x^{(2)}}} \tag{4.4}$$

c'est-à-dire que $RI_{\cdot x^{(2)}}(\underline{Y}, X_j)$ est une moyenne pondérée (par les variances partielles) des carrés des coefficients de corrélation partielle $\hat{\rho}_{y x_j \cdot x^{(2)}}$.

c) Dans le contexte du Lemme (3.1), puisque $(n-1)\, S_{\cdot x^{(2)}}$ possède la distribution $W_{p+1}(\sum_{\cdot x^{(2)}}, n-1-t)$ où

$$\sum_{\cdot x^{(2)}} = \begin{bmatrix} \sum_{yy \cdot x^{(2)}} \sum_{y x_j \cdot x^{(2)}} \\ \sum_{x_j y \cdot x^{(2)}} \sigma_{x_j x_j \cdot x^{(2)}} \end{bmatrix}$$

on voit que

(i) $S_{yy \cdot x^{(2)}, x_j} = S_{yy \cdot x^{(2)}} - S_{y x_j \cdot x^{(2)}}\, s^{-1}_{x_j x_j \cdot x^{(2)}}\, S_{x_j y \cdot x^{(2)}}$

(ii) $A_{yy \cdot x^{(2)} x_j} = (n-1)\, S_{yy \cdot x^{(2)} x_j}$ possède la distribution $W_p(\sum_{yy \cdot x^{(2)}, x_j}, n-2-t)$ où $\sum_{yy \cdot x^{(2)} x_j} = \sum_{yy \cdot x^{(2)}} - \sum_{y x_j \cdot x^{(2)}} \sigma^{-1}_{x_j x_j \cdot x^{(2)}} \sum_{x_j y \cdot x^{(2)}}$

(iii) sous $H_0 : \sum_{y x_j \cdot x^{(2)}} = 0$, la matrice $A_{yy \cdot x^{(2)}} - A_{yy \cdot x^{(2)}, x_j} = (n-1)\, S_{y x_j \cdot x^{(2)}}\, s^{-1}_{x_j x_j \cdot x^{(2)}}\, S_{x_j y \cdot x^{(2)}}$ possède la distribution $W_p(\sum_{yy \cdot x^{(2)}}, 1)$ et est indépendante de la matrice $A_{yy \cdot x^{(2)}, x_j}$.

Par analogie au théorème (3.1), la distribution de

$$\frac{RI_{\cdot x^{(2)}}}{1 - RI_{\cdot x^{(2)}}}$$

est donnée par le théorème suivant:

Théorème (4.1):

Si le vecteur $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ est multinormal, pour tout sous-vecteur $\underline{X}^{(2)} : t \times 1$ de $\underline{X}$, ne contenant pas X_j , et sous $H_0 : \sum_{y x_j \cdot x^{(2)}} = 0$, $p[RI_{\cdot x^{(2)}} / (1 - RI_{\cdot x^{(2)}}) \leq s] = p[U \leq 0]$ où U est distribué comme $\sum_{i=1}^{p(n-1-t)} \delta_i U_i^2$ où les U_i sont iid et $N(0, 1)$ et où

$\delta_1, \delta_2, \ldots, \delta_{p(n-t-1)}$ sont les p valeurs propres de $\sum_{yy \cdot x^{(2)}}$ ayant chacune multiplicité 1 et les p valeurs propres de $-s \sum_{yy \cdot x^{(2)}}$ ayant chacune multiplicité $n - 2 - t$.

Ici encore, pour calculer $p[RI_{\cdot x^{(2)}} / (1 - RI_{\cdot x^{(2)}}) \leq s]$ en pratique on utilisera l'algorithme de Imhof (1961).

5. RELATION DE RECURRENCE ENTRE *RI* ET $RI_{\cdot x^{(2)}}$

Dans cette section on établit une relation de récurrence entre *RI* et $RI_{\cdot x^{(2)}}$ sur laquelle seront basés les tests statistiques de l'algorithme pas à pas de la section suivante.

On connait la relation suivante entre le coefficient de corrélation multiple et le coefficient de corrélation partielle (voir, par exemple, Graybill (1976), p. 445).

$$R^2_{y_i x^{(2)}, x_j} = R^2_{y_i \cdot x^{(2)}} + (1 - R^2_{y_i \cdot x^{(2)}}) \, \hat{\rho}^2_{y_i, x_j \cdot x^{(2)}} \tag{5.1}$$

où $\underline{X}^{(2)} : t \times 1$ est un sous-vecteur de $\underline{X}$ ne contenant pas X_j et où Y_i est la i^{e} composante de $\underline{Y}$.

En multipliant les deux membres de (5.1) par $s^2_{y_i}$, la variance empirique de Y_i, puis en sommant sur i on obtient

$$\sum_{i=1}^{p} s^2_{y_i} R^2_{y_i x^{(2)}, x_j} = \sum_{i=1}^{p} s^2_{y_i} R^2_{y_i \cdot x^{(2)}} + \sum_{i=1}^{p} s^2_{y_i} \hat{\rho}^2_{y_i, x_j \cdot x^{(2)}} - \sum_{i=1}^{p} s^2_{y_i} R^2_{y_i x^{(2)}} \hat{\rho}^2_{y_i, x_j \cdot x^{(2)}} \tag{5.2}$$

Or, on a

$$\sum_{i=1}^{p} s^2_{y_i} \hat{\rho}^2_{y_i, x_j x^{(2)}} = \sum_{i=1}^{p} [s^2_{y_i x^{(2)}} + s_{y_i x^{(2)}} S^{-1}_{x^{(2)} x^{(2)}} s_{x^{(2)} y_i}] \, \hat{\rho}^2_{y_i, x_j \cdot x^{(2)}} \tag{5.3}$$

et

$$\sum_{i=1}^{p} s^2_{y_i} R^2_{y_i x^{(2)}} \hat{\rho}^2_{y_i, x_j x^{(2)}} = \sum_{i=1}^{p} s^2_{y_i} \frac{s_{y_i x^{(2)}} S^{-1}_{x^{(2)} x^{(2)}} s_{x^{(2)} y_i}}{s^2_{y_i}} \hat{\rho}^2_{y_i, x_j x^{(2)}} = \sum_{i=1}^{p} s_{y_i x^{(2)}} S^{-1}_{x^{(2)} x^{(2)}} s_{x^{(2)} y_i} \hat{\rho}^2_{y_i, x_j \cdot x^{(2)}} \tag{5.4}$$

de sorte que (5.2) devient

$$\sum_{i=1}^{p} s_{y_i}^2 \, R^2_{y_i \cdot x^{(2)},\, x_j} = \sum_{i=1}^{p} s_{y_i}^2 \, R^2_{y_i \, x^{(2)}} + \sum_{i=1}^{p} s^2_{y_i \, x^{(2)}} \, \hat{\rho}^2_{y_i,\, x_j \, x^{(2)}} \, . \tag{5.5}$$

En divisant (5.5) par $\sum_{i=1}^{p} s_{y_i}^2$ et en utilisant (2.3) et (4.4) il suit

$$RI\left[\underline{Y}, \begin{bmatrix} \underline{X}^{(2)} \\ X_j \end{bmatrix}\right] = RI(\underline{Y}, \underline{X}^{(2)}) + \frac{\sum_{i=1}^{p} s^2_{y_i \cdot x^{(2)}} \, \hat{\rho}^2_{y_i,\, x_j \cdot x^{(2)}}}{\sum_{i=1}^{p} s^2_{y_i \cdot x^{(2)}}} \; \frac{tr(S_{yy \, x^{(2)}})}{tr(S_{yy})}$$

$$= RI(\underline{Y}, \underline{X}^{(2)}) + RI_{\cdot x^{(2)}}(\underline{Y}, X_j) \, [1 - RI(\underline{Y}, \underline{X}^{(2)})] \, . \tag{5.6}$$

On a une formule analogue au niveau de la population

$$\rho I\left[\underline{Y}, \begin{bmatrix} \underline{X}^{(2)} \\ X_j \end{bmatrix}\right] = \rho I(\underline{Y}, \underline{X}^{(2)}) + \rho I_{\cdot x^{(2)}}(\underline{Y}, X_j) \, [1 - \rho I(\underline{Y}, \underline{X}^{(2)})] \tag{5.7}$$

dans laquelle les matrices de covariance empirique sont remplacées par les matrices de covariance théoriques.

6. ALGORITHME PAS A PAS DE SELECTION DE VARIABLES

La formule (5.7) nous permet de tester l'hypothèse $H_0 : \rho I\left[\underline{Y}, \begin{bmatrix} \underline{X}^{(2)} \\ X_j \end{bmatrix}\right] = \rho I(\underline{Y}, \underline{X}^{(2)})$ ou l'apport de X_j après $X^{(2)}$ est négligeable.

Par (5.7) cette hypothèse est équivalente à $H'_0 : \rho I_{\cdot x^{(2)}}(\underline{Y}, X_j) = 0$ qui est elle-même équivalente, par (4.2), à $H''_0 : \sum_{yx_j \cdot x^{(2)}} = 0$. Or le théorème (4.1) nous permet de tester H''_0 , pour tout X_j et pour tout sous-vecteur $\underline{X}^{(2)} : r \times 1$ ne contenant pas X_j : on rejette H''_0 au niveau α si $RI_{\cdot x^{(2)}} / (1 - RI_{\cdot x^{(2)}}) > u$, le $100(1 - \alpha)$–ième centile de la distribution de U.

(6.1) Algorithme de sélection progressive (angl. Forward Selection)

On cherche à expliquer linéairement le vecteur $\underline{Y} : p \times 1$ à partir du vecteur $\underline{X} : q \times 1$. Soit T un sous-ensemble de $\{1, 2, \ldots, q\}$ et désignons par $\underline{X}_T$ le sous-vecteur de $\underline{X}$ dont les composantes sont indicées par l'ensemble T . L'algorithme est le suivant:

(i) Initialisation: $Q \leftarrow \{1, 2, \ldots, q\}$, $T \leftarrow \varnothing$

(ii) Calculer $\max_{1 \leq j \leq q} RI(\underline{Y}, X_j) = RI(\underline{Y}, X_h)$

Tester $H_0 : \rho I(\underline{Y}, X_h) = 0$ au niveau α_1 .

si accepter: stop (pas de régression possible)

si rejeter: $T \leftarrow T \cup \{h\}$

(iii) Calculer $\max_{j \in C_Q^T} RI(\underline{Y}, \underline{X}_{T \cup \{j\}}) = RI(\underline{Y}, \underline{X}_{T \cup \{h\}})$

Tester $H'_0 : \rho I_{\underline{X}_T}(\underline{Y}, X_h) = 0$ au niveau α_1 .

si accepter: stop ($\underline{X}_T$ explique $\underline{Y}$)

si rejeter: $T \leftarrow T \cup \{h\}$

si $T \neq Q$, aller à (iii)

si $T = Q$, stop ($\underline{X}_T$ explique $\underline{Y}$)

où C_Q^T désigne le complément de T par rapport à Q .

(6.2) Algorithme d'élimination successive (angl. Backward Elimination)

(i) Initialisation: $Q \leftarrow \{1, 2, \ldots, q\}$, $T \leftarrow Q$

(ii) Calculer $\min_{j \in T} RI_{\underline{X}_{T-(j)}}(\underline{Y}, X_j) = RI_{\underline{X}_{T-(h)}}(\underline{Y}, X_h)$

Tester $H'_0 : \rho I_{\underline{X}_{T-(h)}}(\underline{Y}, X_h) = 0$ au niveau α_2 .

si rejeter: stop ($\underline{X}_T$ explique $\underline{Y}$)

si accepter: $T \leftarrow T - \{h\}$

si $T \neq \varnothing$, aller à (ii)

si $T = \varnothing$, stop (pas de régression possible)

(6.3) Algorithme pas à pas (angl. Stepwise)

(i) Commencer en utilisant la procédure de sélection progressive

(ii) A chaque fois qu'une variable est sélectionnée, remettre en cause les variables retenues précédemment, en utilisant la procédure d'élimination successive

(iii) Arrêter lorsqu'aucune variable n'est sélectionnée et qu'aucune variable n'est éliminée.

Dans cet algorithme, les tests d'hypothèses sont effectués au niveau α_1 pour la sélection progressive et au niveau α_2 pour l'élimination successive. Il est commun d'utiliser $\alpha_1 = .10$ et $\alpha_2 = .10$. Costanza et Afifi (1979) suggèrent d'utiliser $.10 \leq \alpha_1 \leq .25$ et $.10 \leq \alpha_2 \leq .25$.

7. EXEMPLE

Les données suivantes furent obtenues par Woltz, Reid et Colwell (1948) et se trouvent également dans Anderson et Bancroft (1952), p. 205. Elles sont présentées dans le tableau 7.1 où

Y_1 = taux de consumabilité de la cigarette (en pouces par seconde, 1 po ≈ 2.5 cm)

Y_2 = pourcentage de sucre dans la feuille de tabac

Y_3 = pourcentage de nicotine

X_1 = pourcentage d'azote

X_2 = pourcentage de chlore

X_3 = pourcentage de potassium

X_4 = pourcentage de phosphore

X_5 = pourcentage de calcium

X_6 = pourcentage de magnésium

L'objectif est d'expliquer le vecteur $\underline{Y} : 3 \times 1$ à partir d'un sous-vecteur du vecteur $\underline{X} : 6 \times 1$ des variables inorganiques de la feuille de tabac. Nous avons $p = 3$, $q = 6$ et $n = 25$.

Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6
1.55	20.05	1.38	2.02	2.90	2.17	.51	3.47	.91
1.63	12.58	2.64	2.62	2.78	1.72	.50	4.57	1.25
1.66	18.56	1.56	2.08	2.68	2.40	.43	3.52	.82
1.52	18.56	2.22	2.20	3.17	2.06	.52	3.69	.97
1.70	14.02	2.85	2.38	2.52	2.18	.42	4.01	1.12
1.68	15.64	1.24	2.03	2.56	2.57	.44	2.79	.82
1.78	14.52	2.86	2.87	2.67	2.64	.50	3.92	1.06
1.57	18.52	2.18	1.88	2.58	2.22	.49	3.58	1.01
1.60	17.84	1.65	1.93	2.26	2.15	.56	3.57	.92
1.52	13.38	3.28	2.57	1.74	1.64	.51	4.38	1.22
1.68	17.55	1.56	1.95	2.15	2.48	.48	3.28	.81
1.74	17.97	2.00	2.03	2.00	2.38	.50	3.31	.98
1.93	14.66	2.88	2.50	2.07	2.32	.48	3.72	1.04
1.77	17.31	1.36	1.72	2.24	2.25	.52	3.10	.78
1.94	14.32	2.66	2.53	1.74	2.64	.50	3.48	.93
1.83	15.05	2.43	1.90	1.46	1.97	.46	3.48	.90
2.09	15.47	2.42	2.18	.74	2.46	.48	3.16	.86
1.72	16.85	2.16	2.16	2.84	2.36	.49	3.68	.95
1.49	17.42	2.12	2.14	3.30	2.04	.48	3.28	1.06
1.52	18.55	1.87	1.98	2.90	2.16	.48	3.56	.84
1.64	18.74	2.10	1.89	2.82	2.04	.53	3.56	1.02
1.40	14.79	2.21	2.07	2.79	2.15	.52	3.49	1.04
1.78	18.86	2.00	2.08	3.14	2.60	.50	3.30	.80
1.93	15.62	2.26	2.21	2.81	2.18	.44	4.16	.92
1.53	18.56	2.14	2.00	3.16	2.22	.51	3.73	1.07

Tableau 7.1: Données sur le tabac

Les procédures (6.1), (6.2) ainsi que (6.3) furent utilisées et les résultats sont les suivants. Dans les deux premiers tableaux on indique la variable retenue ou éliminée aux différentes étapes de la procédure, on donne les valeurs de RI_{χ^2} ainsi que de RI également pour chaque étape et finalement le niveau empirique de signification de chaque test est indiqué. Dans le troisième tableau on indique la variable impliquée ainsi que la valeur de RI .

(7.1) Sélection progressive

Etape	Variable retenue	$RI_{.x^{(2)}}$	RI	Niveau empirique du test
1	1	.500	.500	.000
2	2	.263	.631	.008
3	3	.142	.684	.069
4	4	.108	.718	.128
5	6	.048	.731	.340
6	5	.014	.735	.666

Au niveau $\alpha_1 = .05$ on retiendrait les variables X_1 et X_2 puis on effectuerait une régression linéaire de $\underline{Y}$ sur X_1 et X_2 . Au niveau $\alpha_1 = .10$ on retiendrait les variables X_1, X_2 et X_3 .

(7.2) Elimination successive

Etape	Variable éliminée	$RI_{.x^{(2)}}$	RI	Niveau empirique du test
1	5	.014	.731	.666
2	3	.029	.723	.476
3	4	.128	.683	.096
4	6	.140	.631	.071
5	2	.263	.500	.008

Au niveau $\alpha_2 = .10$ on retiendrait les variables X_1, X_2, X_6 et X_4 et on effectuerait une régression de $\underline{Y}$ sur ces variables.

(7.3) Algorithme pas à pas

Etape	Variable retenue	Variable éliminée	RI
1	1	--	.500
2	2	--	.631
3	3	--	.684

Avec $\alpha_1 = .10$ et $\alpha_2 = .10$ la procédure pas à pas nous conduit à retenir les variables X_1 , X_2 et X_3 . Comme c'est le cas pour la régression linéaire multiple, les procédures de sélection progressive et d'élimination successive ne conduisent pas nécessairement aux mêmes sous-ensembles. La procédure pas à pas, qui est un mélange de ces deux dernières, peut également conduire à un sous-ensemble de variables différents des deux précédents.

L'algorithme de sélection pas à pas proposé dans cet article permet de forcer au départ la présence de certaines variables. Cependant, ces variables pourraient ne pas être retenues par la suite. Par exemple, en forçant la présence de X_3 , le sous-ensemble retenu demeure $\{X_1, X_2, X_3\}$. En forçant la présence de X_6 , le sous-ensemble retenu devient $\{X_1, X_2, X_4, X_6\}$.

Cet algorithme, programmé en FØRTRAN, peut être obtenu des auteurs.

REMERCIEMENTS

Les auteurs remercient l'éditeur ainsi que les rapporteurs pour leurs précieux conseils.

BIBLIOGRAPHIE

[1] Anderson, R.L. et Bancroft, T.A.: Statistical Theory in Research, 1952, McGraw-Hill, New York.

[2] Bonifas, L., Escoufier, Y., Gonzalez, P.L. et Sabatier R.: Choix de variables en analyse en composantes principales, Revue de Statistique Appliquée, 1984, 32, 5-15.

[3] Costanza, M.C. et Afifi, A.A.: Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis, Jour. Amer. Stat. Assoc., 1979, 74, 777-785.

[4] Dambroise, E., Escoufier, Y. et Massotte, P.: Application de l'analyse de données à l'élaboration de mini-sondages d'opinion, Revue de Statistique Appliquée, 1987, 35, 9-24.

[5] Escoufier, Y.: Le traitement des variables vectorielles, Biometrics, 1973, 29, 751-760.

[6] Escoufier, Y.: A propos du choix des variables en analyse des données, Metron, 1986, 44, 31-47.

[7] Gonzalez, P.L. et Chami, H.: Amélioration d'un réseau de surveillance de la pollution atmosphérique, R.A.I.R.O., 18, 1984, 369-384.

[8] Graybill, F.A.: Theory and Application of the Linear Model, 1976, Duxbury Press, Boston, Mass.

[9] Imhof, P.: Computing the Distribution of Quadratic Forms in Normal Variates, Biometrika, 1961, 48, 419-426.

[10] Johnson, N.L. et Kotz, S.: Continuous Univariate Distributions, vol. 2, 1970, Houghton Mifflin, New York.

[11] Lazraq, A. et Cléroux, R.: Etude comparative de différentes mesures de liaison entre deux vecteurs aléatoires et tests d'indépendance, Publ. 610, Dép. I.R.O., U. de Montréal, 1987, à paraître dans Statistique et Analyse des Données.

[12] Mardia, K.V., Kent, J.T. et Bibby, J.M.: Multivariate Analysis, 1979, Academic Press, London.

[13] Rao, C.R.: The Use and Interpretation of Principal Component Analysis in Applied Research, Sankhya, A, 1965, 26, 329-358.

[14] Robert, P. et Escoufier, Y.: A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient, Applied Statistics, 1976, 25, 257-265.

[15] Stewart, D. et Love, W.: A General Canonical Correlation Index, Psycho. Bull., 1968, 70, 160-163.

[16] Woltz, W.G., Reid, W.A. et Colwell, W.E.: Sugar and Nicotine in Cured Bright Tobacco as Related to Mineral Element Composition, Proc. Soil Sci. Soc. Am., 1948, 13, 385-387.