CAHIERS DU BURO

C. PARDOUX

Sur la sélection de variables en régression multiple : une mise au point

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 39-40 (1982), p. 101-133

<http://www.numdam.org/item?id=BURO_1982__39-40__101_0>

SUR LA SÉLECTION DE VARIABLES EN RÉGRESSION MULTIPLE : UNE MISE AU POINT

C. PARDOUX
Université Paris IX - Dauphine

TABLE DES MATIÈRES

SOMMAIRE

Les procédures et critères utilisés pour la sélection de variables explicatives en régression multiple sont présentés. Les méthodes par étapes, les moins coûteuses en calcul, donnent dans certains cas des résultats voisins de ceux obtenus par la recherche de tous les "meilleurs" sous-ensembles (i.e. les sous-ensembles qui pour chaque nombre p de variables explicatives maximisent le coefficient de corrélation multiple R_p) ; les résultats peuvent être très différents quand il y a notamment de fortes corrélations entre les variables explicatives. Le statisticien se doit de ne pas accepter les résultats donnés par une méthode comme la solution définitive ; il incorporera éventuellement aux statistiques l'information disponible. Un exemple illustre les principales méthodes.

I. INTRODUCTION

L'objet de cet exposé est de montrer l'intérêt de la sélection de variables dans les modèles de régression multiple, et d'étudier les principales méthodes actuellement utilisées.

Dans ces mêmes Cahiers du B.U.R.O. , le problème de la sélection de variables a déjà fait l'objet d'une partie d'un article de BRENOT, CAZES et LACOURLY [7]. Depuis la parution de cet article, de nombreux travaux ont été réalisés sur ce sujet, et une nouvelle mise au point s'imposait.

Ce problème de sélection de variables ne se pose, bien sûr, que si l'analyste dispose d'une bonne connaissance sur les variables qui lui permet de construire une classe de modèles a priori pertinents. Donc, il ne doit être abordé qu'après une première analyse sérieuse des données. La détection des liaisons entre les variables explicatives peut être faite par l'examen de la matrice des corrélations et une analyse en composantes principales, ou par l'examen de l'inverse de la matrice de Cholesky. Cette dernière, notée V , est la matrice non singulière des variances-covariances de la variable expliquée et des variables explicatives, V peut se décomposer en un produit de deux matrices triangulaires, transposées l'une de l'autre : $V = AA'$. Si A désigne la matrice triangulaire inférieure, son inverse B est aussi triangulaire inférieure ; HAWKINS et EPLETT [21] ont montré que cette matrice triangulaire B pouvait être considérée comme un résumé particulièrement efficace des relations en régression multiple entre les prédicteurs et la variable dépendante.

Les principales méthodes de sélection seront classées en deux catégories :

- Les méthodes par étapes, les moins coûteuses en calcul, ne garantissent pas l'obtention de la "meilleure" régression théorique à chaque stade ; mais pour le praticien, le sous-ensemble optimal quant à la signification, au coût d'observation ou de mesure, n'est pas toujours celui qui correspond à la meilleure régression (il y a d'ailleurs souvent plusieurs sous-ensembles presque équivalents pour le théoricien) ; les logiciels BMDP [12], SAS [4] et SPSS [31] ont des programmes de sélection par étapes avec des variantes propres à chacun d'eux.

- La recherche des "meilleurs" sous-ensembles de variables amène à choisir selon divers critères de qualité de prédiction dont nous étudierons les propriétés et les liens ; le critère C_p de MALLOWS [28] sera particulièrement retenu ; nous verrons qu'à l'aide d'une décomposition en une somme de termes correspondant chacun à une observation, il permet d'étudier plus spécifiquement l'adéquation d'un sous-modèle.

Le programme 9 R de BMDP [12] recherche le "meilleur" sous-ensemble selon trois critères au choix : le carré du coefficient de corrélation multiple R^2 , le carré ajusté du coefficient de corrélation multiple et le critère C_p de MALLOWS . SAS [4] a un programme (RSQUARE) qui donne les valeurs de R^2 pour toutes les combinaisons possibles de variables. Ces programmes sont coûteux en calcul. Néanmoins certains algorithmes, comme celui de FURNIVAL et WILSON [17] utilisé par BMDP, sont très performants et permettent de traiter des modèles contenant jusqu'à 27 variables explicatives.

PERK [6] montre que les résultats obtenus par les méthodes par étapes et la recherche des "meilleurs" sous-ensembles selon le critère R^2 peuvent être très différents dans les cas où un ensemble de prédicteurs est collectivement significatif sans qu'aucun d'eux considéré isolément ne le soit.

Enfin, il est naturel de tenir compte de toute l'information disponible pour traiter un problème. Plusieurs auteurs (AITKIN [1] ;

DEMPSTER, SCHATZOFF et WERMUTH [9] ; LAMOTTE [24]) ont attiré l'attention sur la prise en compte d'une information sur les coefficients (positivité, par exemple) ou d'une information sur les prédicteurs donnée par l'expérience (certains d'entre eux doivent être inclus dans le modèle, par exemple).

Nous commenterons à partir d'un exemple les résultats obtenus par ces différentes méthodes.

II. DE L'INTERET DE LA SELECTION DE VARIABLES

II.1 Notations et hypothèses

Le modèle général s'écrit : $\underset{(nx1)}{Y} = \underset{(nxr)}{X}\ \underset{(rx1)}{\beta} + \underset{(nx1)}{\epsilon}$ (1)

où : n , le nombre d'observations est plus grand ou égal à r , le nombre de variables explicatives ; Y est un vecteur aléatoire à n composantes (variable dépendante) ; X est une matrice connue supposée de rang r de variables non aléatoires (la matrice X peut contenir un vecteur colonne constant), β est le vecteur des paramètres inconnus ; et ϵ , le vecteur à n dimensions des erreurs aléatoires.

Les hypothèses sur les distributions de Y et ϵ sont les suivantes : les y_i sont des v.a. de moyenne $x_i \beta$ (x_i désignant le $i^{ème}$ vecteur ligne de X), de même variance σ^2 et ne sont pas corrélées entre elles :

$$E(Y) = X\beta \quad \Leftrightarrow \quad E(\epsilon) = C$$

$$\text{Var } Y = \sigma^2 I = \text{Var } \epsilon \quad .$$

Le modèle linéaire "complet" est donc supposé non biaisé.

L'estimateur des moindres carrés de β s'écrit :

$$\hat{b} = (X'X)^{-1} X'Y \quad .$$

A un vecteur ligne x de variables explicatives, correspond l'estimateur suivant de E(y) :

$$\hat{y} = x \hat{b} \ .$$

Dans la suite, on notera X_p , la matrice des p colonnes de X correspondant aux variables retenues pour le sous-modèle et par X_q , la matrice correspondant aux autres variables ($p + q = r$).

Le modèle (1) peut s'écrire : $Y = X_p \beta_p + X_q \beta_q + \epsilon$.

L'estimateur $\hat{b}_p$ des moindres carrés de β_p pour le sous-modèle : $Y = X_p \beta_p + \epsilon'$, s'écrit : $\hat{b}_p = (X_p' X_p)^{-1} X_p' Y$.

Si on désigne par b_p, le vecteur des composantes de $\hat{b}$ correspondant aux variables retenues et par b_q , le vecteur des composantes de $\hat{b}$ correspondant aux autres variables, on a la relation suivante entre $\hat{b}_p$, b_p et b_q (cf. ULMO [38]) :

$$\hat{b}_p = b_p + (X_p' X_p)^{-1} X_p' X_q b_q \quad .$$

On aura donc : $\hat{b}_p = b_p$ si et seulement si : $b_q = 0$ ou $X_p' X_q = 0$, i.e. si les variables composant X_p sont orthogonales aux variables composant X_q .

II.2 *Le recours aux estimateurs biaisés*

Si les variables explicatives sont quasiment colinéaires, l'estimateur sans biais des moindres carrés de β a des coefficients de variances élevées. Cet estimateur est alors instable et le modèle permet de reconstruire uniquement les données considérées.

A la suite d'une analyse en composantes principales, on peut être amené à mesurer la "vraie" dimension du sous-espace engendré par l'ensemble des variables explicatives et à procéder alors à une première élimination de certaines variables. On peut aussi remplacer l'ensemble des variables par un certain nombre de composantes principales : c'est la régression orthogonale (cf. MANSFIELD, WEBSTER et GUNST [29]). Mais, il vaut mieux en général garder les variables d'origine pour leur signification.

La minimisation de la somme des carrés des résidus (ou erreurs):

$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, est le critère le plus utilisé pour estimer les coefficients β . D'autres critères : minimisation de la somme des valeurs absolues des résidus relatifs, sont jugés mieux adaptés aux modèles économiques (cf. NARULA et WELLINGTON [30]). Si on cherche à minimiser un critère de ce type, on est évidemment amené à inclure dans le modèle les r variables (à moins que certaines d'entre elles n'aient rien à voir avec la variable dépendante). Mais pour des raisons déjà évoquées, il vaut souvent mieux ne pas utiliser toutes les variables : il faut réaliser le "meilleur" compromis entre un nombre minimum p de prédicteurs et un "bon" modèle ; c'est le principe de parcimonie.

ALLEN [2] a introduit la notion d'"Erreur quadratique moyenne de prédiction". Dans le cas où on retient p variables explicatives, l'"Erreur quadratique moyenne de prédiction" pour une nouvelle observation (y,x) est égale à la valeur moyenne du carré de la diffé-

rence entre y et son approximation $\hat{y}_p$: $MSEP(\hat{y}_p) = E(y - \hat{y}_p)^2$ avec : $\hat{y}_p = x_p \hat{b}_p$, $\hat{b}_p$ étant l'estimateur de β_p associé au n-échantillon de la régression, et x_p (resp. x), le vecteur ligne des valeurs des p (resp. r) variables relatives à la nouvelle observation.

Si on suppose y de moyenne $x\beta$ et de variance σ^2 , on montre que :

$$MSEP(\hat{y}_p) = E(y - \hat{y}_p)^2 = \sigma^2 + \text{var } \hat{y}_p + (E(\hat{y}_p) - x\beta)^2 \quad ;$$

les deux derniers termes représentent donc l'erreur quadratique moyenne de $\hat{y}_p$ considéré comme estimateur de $x\beta$:

$$MSE(\hat{y}_p) = E[\,(\hat{y}_p - x\beta)^2\,] \quad .$$

Si on cherche le sous-ensemble de variables tel que $MSEP(\hat{y}_p)$ soit minimum, on peut être amené à ne pas inclure toutes les variables dans le modèle.

Ainsi, des estimateurs biaisés sont parfois préférables à des estimateurs sans biais. Ceci a été confirmé par de multiples travaux, en particulier par DEMPSTER, SHATZOFF et WERMUTH [9] qui par une étude utilisant la simulation, ont comparé 56 alternatives différentes à la méthode des moindres carrés usuelle.

III. LES METHODES PAR ETAPES

Si on a r variables explicatives, les meilleurs sous-ensembles de variables pour un critère donné sont parmi les (2^r-1) combinaisons possibles de p = 1,2,...,r variables, ce nombre de combinaisons devenant impressionnant dès que r dépasse 7 ou 8 . C'est la raison pour laquelle les méthodes de sélection par étapes qui évitent d'explorer tous les modèles possibles, ont eu beaucoup de succès. Elles consistent à rechercher un sous-ensemble de variables en procédant par incorporation ou retrait de variables, chaque opération donnant un résultat qui influence les opérations suivantes.

III.1 Introduction "ascendante" des variables

La première variable à entrer dans la régression est celle qui a le coefficient de corrélation le plus élevé avec la variable dépendante. On ajoute ensuite une variable à la fois dans l'équation de régression. La variable introduite à chaque étape est celle qui maximise le rapport suivant :

$$F_{p+(j)} = (n-p-1)\frac{RSS_p - RSS_{p+(j)}}{RSS_{p+(j)}} = (n-p-1)\frac{R^2_{p+(j)} - R^2_p}{1-R^2_{p+(j)}}$$

où : RSS_p est la somme des carrés des résidus correspondant au modèle à p variables ; $RSS_{p+(j)}$, la somme des carrés des résidus du modèle contenant ces mêmes p variables et une nouvelle variable j ; R^2_p , le carré du coefficient de corrélation multiple de Y en fonction des p variables et $R^2_{p+(j)}$, le carré du coefficient de corrélation multiple de Y en fonction de ces mêmes p variables et de la nouvelle variable j .

On arrête la procédure en choisissant la régression qui précède l'introduction d'une variable "non significative" au sens de la statistique F de FISHER-SNEDECOR. Deux points importants sont à soulever au sujet de la validité de ce test :

- A chaque étape, le modèle contenant seulement une ou plusieurs variables sélectionnées, est supposé sans biais ; or, c'est le modèle complet qui est supposé non biaisé.

- Même si on fait les hypothèses permettant de supposer que $F_{p+(j)}$ suit une loi de FISHER-SNEDECOR, la v.a. $\max_j F_{p+(j)}$ ne suit pas cette loi.

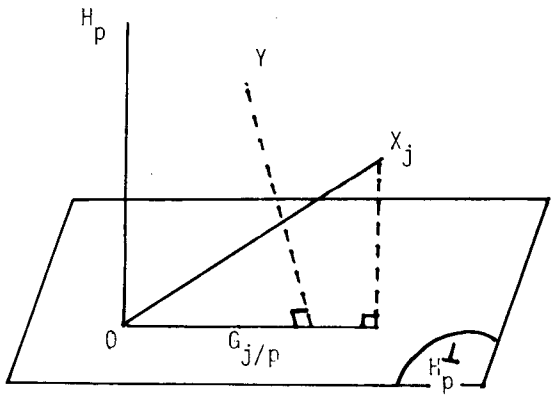
L'utilisation du test de FISHER-SNEDECOR a pour ces raisons été critiquée en particulier par DERFLINGER et STAPPLER [10] ; DRAPER, GUTTMAN et KANEMASU [13] ; POPE et WEBSTER [33].

DERFLINGER et STAPPLER ont construit un autre test en considérant la statistique $W = \max_j |T_j|$ où :

$$T_j = \frac{G_{j/p}}{\sqrt{\sum_{k=1}^{n-r} D^2_{k/n-r}}} \quad ; \; G_{j/p} \text{ est proportionnel à la projection}$$

de Y sur la variable obtenue en projetant la $j^{ième}$ variable sur le complémentaire dans R^n du sous-espace H_p engendré par les p variables déjà introduites ; on a : $E(G_{j/p}) = 0$ et $Var(G_{j/p}) = \sigma^2$ sous l'hypothèse H_0 : $\begin{cases} \beta_j = \hat{b}_{p,j} & j = 1,\ldots,p \\ \beta_j = 0 & j = p+1,\ldots,r \end{cases}$

D_k est la projection de Y sur le $k^{ième}$ vecteur d'une base orthonormée du complémentaire dans R^n du sous-espace engendré par les r variables ; $\sum_{k=1}^{n-r} D^2_{k/n-r}$ est donc un estimateur sans biais de σ^2 .



Plutôt que de recourir à des tests difficiles à mettre en oeuvre, le praticien a toujours la possibilité d'utiliser des techniques descriptives : représenter graphiquement les valeurs de F_p pour voir l'étape à laquelle se produit un saut.

Notons que rechercher à chaque étape la variable j telle que $F_{p+(j)}$ soit maximum, est équivalent à rechercher la variable j telle que le coefficient de corrélation partielle entre la variable dépendante et la variable j sachant les p variables déjà sélectionnées, soit maximum en valeur absolue ; ce qui est encore équivalent à rechercher la variable j telle que le coefficient de corrélation multiple $R_{p+(j)}$ soit maximum.

La sélection se fait ici par l'intermédiaire de la statistique F , mais le coefficient de corrélation multiple entre la variable dépendante et le sous-ensemble retenu en fin de procédure est-il significatif ? Ce problème a été abordé par DIEHR et HOFLIN [11], RENCHER et PUN [34], WILKINSON et DALLAL [40]. Ces derniers ont construit par simulation des tableaux permettant de tester la nullité du carré du coefficient de corrélation multiple dans le cas d'utilisation de la procédure de sélection ascendante.

Avec cette méthode d'introduction ascendante des variables, on n'étudie pas comment une nouvelle variable introduite modifie le rôle des précédentes : une variable peut très bien perdre de son importance lorsqu'elle est combinée à d'autres variables. Dans certains cas, ce peut être la combinaison de certaines variables plutôt que leur présence individuelle qui importe le plus. C'est pourquoi, EFROYMSON [15] a jugé important de modifier cette technique et a mis au point la régression pas à pas (stepwise regression) qui est une méthode d'incorporations successives avec possibilité de retrait à chaque étape.

III.2 Elimination "descendante" des variables

Cette méthode consiste à :

- Effectuer la régression avec les r variables ;

- Eliminer la variable qui provoque la plus faible diminution du rapport $F_{r-(j)}$ donné ci-dessous ;

- Eliminer ensuite parmi les (r-1) variables restantes celle qui provoque la plus faible diminution du $F_{r-1-(j)}$; et ainsi de suite ...

$$F_{p-(j)} = (n-p) \cdot \frac{RSS_{p-(j)} - RSS_p}{RSS_p} = (n-p) \cdot \frac{R_p^2 - R_{p-(j)}^2}{1 - R_p^2} \quad .$$

La notation p-(j) signifie qu'on a enlevé la variable j de la régression à p variables.

Notons qu'utiliser la quantité $F_{p-(j)}$ revient à utiliser le carré du t de Student associé à la variable j dans la régression à p variables.

On arrête le processus d'élimination en choisissant le modèle qui précède l'élimination d'une variable "significative" au sens de F . Comme précédemment, l'application du test de FISHER-SNEDECOR à la statistique $\min_j F_{p-(j)}$ est à contester.

Cette méthode est particulièrement recommandée si on désire avoir la formule de régression sur l'ensemble des variables explicatives. Elle donne aussi la meilleure régression à r-1 termes.

III.3 Les méthodes par étapes utilisées dans les logiciels

- BMDP [12] propose dans son programme 2R "Stepwise regression" quatre méthodes qui peuvent être utilisées pour introduire ou éliminer des variables à chaque étape ; supposons que le modèle contienne p variables :

i/ F : La variable j avec le plus petit $F_{p-(j)}$ est éliminée si $F_{p-(j)}$ est plus petit qu'une valeur limite. Si aucune variable ne satisfait ce critère, la variable j avec le plus grand $F_{p+(j)}$ est introduite si $F_{p+(j)}$ dépasse une valeur limite.

ii/ FSWAP : la variable j ayant le plus petit $F_{p-(j)}$ est éliminée si $F_{p-(j)}$ ne dépasse pas une valeur limite. Si aucune variable ne satisfait ce critère, on échange une variable de l'équation avec une variable qui n'est pas encore dans l'équation si cet échange accroît le coefficient de corrélation multiple R . Si aucune variable ne peut être échangée, on introduit une variable comme dans le critère précédent.

iii/ R : la variable j avec le plus petit $F_{p-(j)}$ est éliminée si sa suppression mène à un coefficient de corrélation multiple R plus grand en valeur absolue que celui auparavant obtenu avec le même nombre de variables. Si aucune variable ne satisfait ce critère, on introduit une variable selon F .

iv/ RSWAP : la variable j avec le plus petit $F_{p-(j)}$ est éliminée par le critère R . Si aucune variable ne satisfait ce critère, une variable de l'équation est échangée avec une variable pas encore dans l'équation si l'échange accroît le coefficient de corrélation multiple R en valeur absolue. Si aucune variable ne peut être échangée, une variable est introduite selon F .

Ces méthodes demandent le choix délicat d'une valeur limite.

La méthode F est la moins coûteuse en calcul, la méthode R est un peu plus coûteuse. FSWAP et RSWAP demandent beaucoup plus de calculs que F et R , et leurs coûts sont comparables au coût de calcul du programme 9R de recherche des "meilleurs" sous-ensembles lorsque le nombre de variables explicatives ne dépasse pas 27 .

- SAS [4] propose cinq méthodes de sélection par étapes dans un programme appelé "Stepwise procedure" :

i/ FORWARD : cette option utilise la méthode d'introduction ascendante ; si on ne spécifie pas le seuil du F , il est égal à SLENTRY = 0,5 .

ii/ BACKWARD : programme de la méthode d'élimination descendante ; si on ne spécifie pas le seuil du F , il est égal à SLSTAY = 0,1 .

iii/ STEPWISE : c'est la procédure pas à pas avec incorporation et retrait en se référant aux seuils SLENTRY et SLSTAY .

iv/ MAXR (Maximum R^2 improvement) : procédure originale, présentée comme supérieure à la méthode pas à pas et comme presque aussi bonne que la méthode de sélection à partir du calcul de toutes les régressions possibles ; cette méthode améliore la procédure d'introduction ascendante des variables en remplaçant s'il y a lieu, à chaque étape chaque variable du modèle par celle qui donne le plus grand accroissement de R^2 .

v/ MINR (Minimum R^2 improvement) : procédure qui donne usuellement le même "meilleur" modèle que MAXR, mais qui demande plus de calculs que MAXR .

- SPSS [31] propose seulement un programme d'introduction ascendante des variables : FORWARD (STEPWISE) INCLUSION.

Remarquons que les méthodes par étapes qui ont été améliorées pour donner des résultats voisins de ceux qui sont donnés par le calcul de toutes les régressions, sont aussi coûteuses en calcul (ou presque) que la recherche de toutes les meilleures régressions.

III.4 Conclusion

Rappelons que ces méthodes par étapes ne donnent pas toujours les mêmes résultats lorsqu'on les pratique sur les mêmes données, et aucune d'elles ne garantit l'obtention du "meilleur" sous-ensemble à p variables. On le reconstatera dans l'exemple donné au paragraphe V .

BERK [6] donne des résultats théoriques intéressants :
La sélection ascendante donne le meilleur sous-ensemble (i.e. celui dont R^2 est le plus élevé) pour chaque taille de sous-ensemble si et seulement si la sélection descendante donne le meilleur sous-ensemble de chaque taille.

Il s'ensuit que si les sélections ascendante et descendante ne concordent pas, alors aucune ne donne les "meilleurs" sous-ensembles.

Un autre énoncé du théorème est que si tous les "meilleurs" sous-ensembles sont emboîtés, alors les sélections ascendante et descendante donnent les meilleurs sous-ensembles.

Il est fréquent que la sélection ascendante donne le "meilleur" sous-ensemble pour des petits nombres de variables, mais pas pour des grands nombres de variables ; et que la sélection descendante donne des résultats contraires.

GUNST et MASON [19] exposent un exemple de sélection de variables qu'ils ont traité avec plusieurs méthodes.
Ils analysent les différences des résultats donnés qui se trouvent être dûes pour leur problème aux effets de colinéarité entre variables explicatives. Ils arrivent à la même conclusion que BERK : on a de meilleurs résultats par la procédure de recherche du "meilleur" sous-ensemble de chaque taille lorsqu'un ensemble de variables est collectivement significatif sans qu'aucune d'elles considérée isolément ne le soit au sens du test F .

IV. LA RECHERCHE DES MEILLEURS SOUS-ENSEMBLES

Nous allons étudier et comparer les différents critères de sélection utilisés pour la recherche du "meilleur" sous-ensemble de prédicteurs. Les critères proposés sont des fonctions de RSS_p , somme des carrés des résidus de la régression obtenue avec les p variables sélectionnées. Ces critères ont l'avantage par rapport

aux méthodes par étapes de nous permettre de choisir entre des modèles non emboîtés.

IV.1 Le carré moyen résiduel

On cherche le sous-ensemble qui minimise le critère :

$$RMS_p = \frac{RSS_p}{n-p} \quad .$$

Ce critère "pénalise" la diminution de RSS_p dûe à l'adjonction d'une nouvelle variable par la diminution du dénominateur.
On montre que :

$$E(RSS_p) = (n-p)\,\sigma^2 + \sum_{i=1}^{n} (E(\hat{y}_{p_i}) - x_i\,\beta)^2$$

$$\Rightarrow \quad E(RMS_p) = \sigma^2 + \frac{1}{n-p} \sum_{i=1}^{n} (E(\hat{y}_{p_i}) - x_i\,\beta)^2 \quad .$$

IV.2 Le carré du coefficient de corrélation multiple

$$R_p^2 = 1 - \frac{RSS_p}{TSS} \qquad \text{où :} \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R_p^2 croît avec p et est maximum lorsque le modèle contient les r variables (R_p^2 est maximum lorsque RSS_p est minimum).
On peut tracer la courbe $\max(R_p^2)$ en fonction de p et retenir la plus petite valeur de p pour laquelle $\max(R_p^2)$ se stabilise.

IV.3 Le carré ajusté du coefficient de corrélation multiple

$$\bar{R}_p^2 = 1 - \frac{n-1}{n-p}(1-R_p^2) = 1 - \frac{n-1}{n-p}\frac{RSS_p}{TSS} = 1 - (n-1)\frac{RMS_p}{TSS} \quad .$$

On recherche le sous-ensemble qui maximise $\bar{R}^2_p$; ce sous-ensemble est le même que celui qui minimise RMS_p .

On montre qu'éliminer une variable j d'un ensemble de p variables au sens de ce critère ($\bar{R}^2_p \leqslant \bar{R}^2_{p-j}$) revient à l'éliminer par la méthode d'élimination descendante si :

$$F_{p-(j)} = (n-p) \cdot \frac{RSS_{p-(j)} - RSS_p}{RSS_p} \leqslant 1 .$$

De même, ajouter une variable j à p variables au sens de ce critère ($\bar{R}^2_p \leqslant \bar{R}^2_{p+j}$) revient à l'ajouter par la méthode d'introduction ascendante si :

$$F_{p+(j)} = (n-p-1) \frac{RSS_p - RSS_{p+j}}{RSS_{p+(j)}} \geqslant 1 .$$

IV.4 La statistique C_p de MALLOWS

GORMAN et TOMAN [18] se référant aux travaux de MALLOWS considèrent le critère Γ_p suivant ($\sigma^2 \Gamma_p$ représentant le total de l'erreur quadratique moyenne) :

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2} \sum_{i=1}^{n} MSE(\hat{y}_{p_i}) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} (var(\hat{y}_{p_i}) + (E(\hat{y}_{p_i}) - x_i \beta)^2) \\
&= \frac{1}{\sigma^2} (p\sigma^2 + E(RSS_p) - (n-p)\sigma^2) \\
&= \frac{E(RSS_p)}{\sigma^2} + 2p - n .
\end{aligned}$$

MALLOWS a proposé l'utilisation de la statistique suivante :

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

où $\hat{\sigma}^2$ est un estimateur de σ^2 généralement pris égal au carré moyen

résiduel RMS_r obtenu en utilisant toutes les variables (RMS_r est un estimateur sans biais de σ^2 si on suppose le modèle complet non biaisé) ; dans ce cas, on a : $C_r = r$.

Sous l'hypothèse de normalité des erreurs, la statistique F pour tester l'hypothèse : $\beta_q = 0$, est :

$$F_p = \frac{RSS_p - RSS_r}{(r-p).RMS_r} \quad .$$

On démontre que : $$F_p = 1 + \frac{C_p - p}{r - p}$$

ce qui est encore équivalent à : $C_p = (r - p)(F_p - 1) + p$.

Ainsi, des valeurs de C_p beaucoup plus grandes que p - ou des valeurs de F_p beaucoup plus grandes que 1 - , indiquent que nous n'avons pas un sous-modèle adéquat.

Pour comparer les différents sous-modèles, on peut représenter graphiquement les observations des C_p , ou celles des F_p . SPJØVOLL [36] recommande de représenter les $\frac{C_p}{p}$ au lieu des C_p si on préfère comparer les points avec une ligne horizontale.

SPJØTVOLL donne les valeurs de l'espérance et de la variance de la statistique C_p lorsque les y_i sont supposés indépendants, de distribution normale et de même variance σ^2 et lorsque le sous-modèle considéré est adéquat.

Les distributions de C_p et de F_p pouvant être très dissymétriques, SPJØTVOLL recommande d'évaluer les probabilités

$$P_p = Pr\ (C_p \geqslant \text{la valeur observée de } C_p)$$

$$= Pr\ (F_p \geqslant \text{la valeur observée de } F_p) \ .$$

Il suggère d'examiner les P_p à l'aide d'une représentation graphique.

Mentionnons que dans le cas où $\hat{\sigma}^2 = RMS_r$, on a les relations suivantes entre le critère C_p et les critères R_p^2 et $\bar{R}_p^2$:

$$C_p = \begin{cases} (n-r) \cdot \dfrac{1 - R_p^2}{1 - R_r^2} + 2p - n \\ (n-p) \cdot \dfrac{1 - \bar{R}_p^2}{1 - \bar{R}_r^2} + 2p - n \end{cases} .$$

Ces égalités impliquent que pour chaque valeur de p , on a le minimum de C_p et le maximum de R_p^2 ou de $\bar{R}_p^2$ pour le même ensemble de variables ; mais ces critères n'amènent pas en général à sélectionner le même sous-ensemble de variables (on le constatera sur l'exemple du paragraphe VI).

On montre de plus qu'éliminer une variable j d'un ensemble de p variables au sens du critère C_p ($C_{p-j} \leq C_p$) revient à l'éliminer par la méthode d'élimination "descendante" des variables si :

$$F_{p-(j)} = (n-p) \frac{RSS_{p-(j)} - RSS_p}{RSS_p} \leq 2 \frac{n-p}{n-r} \frac{1 - R_r^2}{1 - R_p^2} .$$

De même, ajouter une variable j à p variables au sens du critère C_p ($C_{p+j} \leq C_p$) revient à l'ajouter par l'introduction "ascendante" si :

$$F_{p+(j)} = (n-p-1) \frac{RSS_p - RSS_{p+(j)}}{RSS_{p+(j)}} \geq 2 \frac{1 - R_r^2}{1 - R_{p+(j)}^2} \frac{n-p-1}{n-r} ,$$

IV.5 "Prédiction Sum of Squares" (noté PRESS) de ALLEN

ALLEN [3] cherche à minimiser la quantité :

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{p_i})^2$$

où : $\hat{y}_{p_i} = x_{p_i} \tilde{b}_p$, $\tilde{b}_p$ étant obtenu de la même façon que $\hat{b}_p$, mais seulement à partir de (n - 1) observations, la $i^{ème}$ étant exclue. En d'autres termes, chaque observation est "prédite" en utilisant les (n - 1) autres observations. ALLEN a montré que :

$$PRESS_p = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{p_i})^2}{(1 - Q_i)^2} \qquad \text{ou : } Q_i = x_{p_i} (X_p' X_p)^{-1} x'_{p_i} \quad .$$

Ainsi $PRESS_p$ peut être interprété comme une somme pondérée des carrés des résidus où le poids relatif au $i^{ème}$ résidu est fonction de la variance de $\hat{y}_{p_i}$ ($Var(\hat{y}_{p_i}) = Q_i \sigma^2$). On verra au paragraphe VI que ce critère ne donne pas pour chaque valeur de p le sous-ensemble qui maximise le coefficient de corrélation multiple.

Peut-on ne pas calculer toutes les régressions possibles pour utiliser ces critères ? Plusieurs auteurs dont FURNIVAL et WILSON [17] ont développé des algorithmes de calculs qui pour identifier les meilleurs sous-ensembles de chaque taille p ($1 \leq p \leq r$), amènent à évaluer seulement une faible partie des ($2^r - 1$) sous-ensembles. Le choix du "meilleur sous-ensemble" parmi ceux qui sont sélectionnés peut être guidé par la valeur de C_p, ou de R_p^2 , ou de $\bar{R}_p^2$. Ces trois critères sont ceux qui ont été retenus par le programme 9R de BMDP . Ce programme qui utilise l'algorithme de FURNIVAL et WILSON est très performant et permet d'analyser des problèmes contenant jusqu'à 27 variables.

V. LES CRITERES NON AGREGES

Tous les critères qui viennent d'être étudiés sont des statistiques agrégées. Celles-ci mesurent l'adéquation moyenne du modèle aux données, mais ne reflètent pas l'adéquation propre à chacune des régions de l'espace des observations. On peut envisager des procédures tenant compte des données individuellement. ALLEN [3] propose de rechercher un sous-ensemble de variables indépendantes pour chaque ensemble d'observations. WEISBERG [39] décompose le critère C_p en n composantes, ce qui permet d'étudier un sous-modèle particulier en déterminant le rôle de chacune des observations dans la structure de C_p .

V.1 Notations

Soit $U = X(X'X)^{-1} X'$, la matrice associée à la projection orthogonale sur le sous-espace H engendré par les colonnes de X .

Soit $V = X_p(X_p' X_p)^{-1} X_p'$, la matrice associée à la projection orthogonale sur le sous-espace H_p engendré par les colonnes de X_p .

Soit $Z = (I - V)X_q$; les vecteurs colonnes de Z sont les projections des vecteurs colonnes de X_q sur le sous-espace orthogonal à H_p.

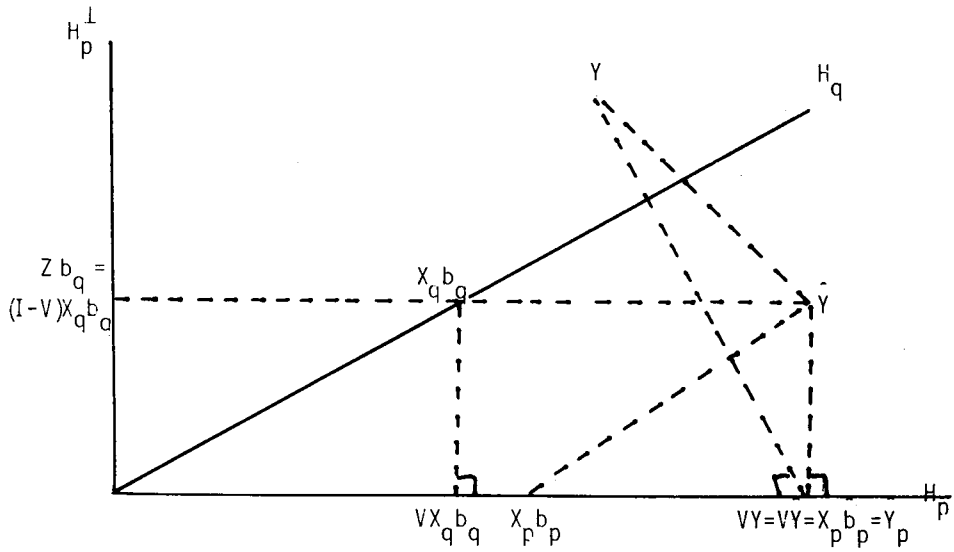
Posons : $W = Z(Z'Z)^{-1} Z'$; W est l'opérateur de projection sur le sous-espace H' de H orthogonal à H_p . On a donc : $U = V + W$.

Le modèle : $Y = X_p \beta_p + X_q \beta_q + \epsilon$ peut alors se réécrire ainsi : $Y = X_p \gamma_p + Z \beta_q + \epsilon$.

Notons que d'après II.1, l'estimateur des moindres carrés de γ_p pour ce dernier modèle sera le même que celui de β_p pour le modèle : $Y = X_p \beta_p + \epsilon'$. L'estimateur des moindres carrés du vecteur (γ_p, β_q) de ce modèle sera donc : $(\hat{b}_p, b_q)$, car comme le suggère la

figure ci-dessous, on peut montrer que l'estimateur b_q de β_q est le même dans les deux versions :

$$Y = X_p \beta_p + X_q \beta_q + \epsilon \qquad \text{et} \qquad Y = X_p \gamma_p + Z \beta_q + \epsilon \quad .$$



V.2 La procédure d'ALLEN

ALLEN [2] donne une méthode pour rechercher le sous-ensemble qui minimise $MSEP(\hat{y}_p)$ pour chaque y . Nous allons en exposer le principe :

Si on utilise les r variables, l'"Erreur carrée moyenne de prédiction" est : $MSEP(\hat{y}_r) = \sigma^2(1 + x(X'X)^{-1} x')$.

Si on utilise un sous-ensemble de p variables :

$$MSEP(\hat{y}_p) = \sigma^2(1 + x_p(X_p'X_p)^{-1} x_p') + (E(\hat{y}_p) - x\beta)^2 \quad .$$

D'après un résultat énoncé au paragraphe II.1 :

$$E(\hat{y}_p) = E(x_p\ \hat{b}_p) = E[\,x_p(b_p + (X_p'\ X_p)^{-1}\ X_p'\ X_q\ b_q)]$$

$$= x_p\ \beta_p + x_p\ (X_p'\ X_p)^{-1}\ X_p'\ X_q\ \beta_q \quad .$$

En posant : $z = x_q - x_p\ (X_p'\ X_p)^{-1}\ X_p'\ X_q$, on obtient :

$$E(\hat{y}_p) - x\beta = E(\hat{y}_p) - x_p\ \beta_p - x_q\ \beta_q = -z\ \beta_q$$

on en déduit :

$$MSEP(\hat{y}_p) = \sigma^2(1 + x_p(X_p'\ X_p)^{-1}\ x_p') + (z\ \beta_q)^2 \quad .$$

D'autre part, on peut montrer que :

$$x\ (X'\ X)^{-1}\ x' = x_p(X_p'\ X_p)^{-1}\ x_p' + z(Z'\ Z)^{-1}\ z' \quad .$$

Cette dernière égalité entraîne :

$$MSEP(\hat{y}_p) - MSEP(\hat{y}_r) = (z\ \beta_q)^2 - z(Z'\ Z)^{-1}\ z'\ \sigma^2 \ .$$

Si cette quantité est négative, on considère que $\hat{y}_p$ est meilleur que $\hat{y}_r$. Puisqu'elle dépend des paramètres inconnus β_q et σ^2 , la technique d'ALLEN requiert des estimateurs de ces quantités.

Cette procédure nécessite un effort important de calcul pour trouver un sous-ensemble optimal de variables (différent au besoin) pour chaque y . Elle peut permettre d'identifier les données exceptionnelles.

V.3 Décomposition de la statistique C_p

WEISBERG [39] décompose le critère C_p en n composantes, chacune d'elles correspondant à une observation.

Nous avons vu que la statistique C_p de MALLOWS est un estimateur de la quantité :

$$\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} MSE(\hat{y}_{pi}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} [\, var(\hat{y}_{pi}) + (E(\hat{y}_{pi}) - E(y_i))^2 \,] \quad .$$

Puisque : $E(y_i) = E(\hat{y}_i)$ si on suppose le modèle complet non biaisé, on obtient si on désigne par v_{ii} , le $i^{ème}$ terme diagonal de V :

$$\Gamma_p = \sum_{i=1}^{n} \left(v_{ii} + \frac{(E(\hat{y}_{pi}) - E(\hat{y}_i))^2}{\sigma^2} \right) .$$

On a : $(E(\hat{y}_{pi}) - E(\hat{y}_i))^2 = E(\hat{y}_{pi} - \hat{y}_i)^2 - var(\hat{y}_{pi} - \hat{y}_i)$.

D'après ce qui précède :

$$var(\hat{y}_{pi} - \hat{y}_i) = var(z_i\, b_q) = w_{ii}\, \sigma^2 = (u_{ii} - v_{ii})\, \sigma^2$$

si on désigne par u_{ii} , le $i^{ème}$ terme diagonal de U , et par z_i , la $i^{ème}$ ligne de la matrice Z .

Si on estime $E(\hat{y}_{pi} - \hat{y}_i)^2$ par sa valeur observée, on obtient comme estimateur de Γ_p , la quantité :

$$\sum_{i=1}^{n} \left[\frac{(\hat{y}_{pi} - \hat{y}_i)^2}{\hat{\sigma}^2} + v_{ii} - (u_{ii} - v_{ii}) \right] ,$$

quantité qui est égale à C_p , en effet :

$$\sum_{i=1}^{n} v_{ii} = \text{trace } V = p \ , \quad \sum_{i=1}^{n} u_{ii} = \text{trace } U = r \ , \text{ et}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_{pi})^2 = \sum_{i=1}^{n} (\hat{y}_{pi} - \hat{y}_i)^2 + (n-r)\, \sigma^2$$

si : $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-r}$.

Posons : $$C_{pi} = \frac{(y_{pi} - \hat{y}_i)^2}{\hat{\sigma}^2} + v_{ii} - (u_{ii} - v_{ii}) \quad .$$

L'examen des C_{pi} donne une information sur le rôle de chaque observation dans la détermination de C_p .

On démontre que sous les hypothèses de normalité :

$$E(C_{pi}) \simeq v_{ii} + (u_{ii} - v_{ii})(z_i \beta_q)^2$$

$$(n-r)\text{var}(C_{pi}) \simeq 2(u_{ii} - v_{ii})^2 (1 + 2(z_i \beta_q))^2 .$$

Ainsi, on peut comparer chaque C_{pi} avec v_{ii} , et un C_{pi} très différent de v_{ii} indique que ni le biais, ni la quantité $(u_{ii} - v_{ii})$ ne sont très petits. Si la quantité $(u_{ii} - v_{ii})$ est grande, $\text{var}(C_{pi})$ sera grande et C_{pi} aura en moyenne une contribution plus importante dans la détermination de C_p .

WEISBERG illustre sa méthode par un exemple. Il sélectionne d'abord les sous-ensembles de variables qui ont les plus petites valeurs de C_p . Il représente ensuite sur des graphiques séparés les ensembles de points (v_{ii}, C_{pi}). Il montre comment ces graphiques permettent d'interpréter C_p et de choisir le sous-modèle le mieux adapté à l'ensemble des données. Cette procédure peut être particulièrement intéressante quand on analyse des données à des fins de prédiction, car il est alors utile de tenir compte de l'évolution des $(C_{pi} - v_{ii})$.

VI. UN EXEMPLE DE SELECTION DE VARIABLES

HAGA et OKUNO [20] ont illustré l'utilisation de critères de sélection avec 30 ensembles d'observations générées au hasard à partir du modèle suivant :

$$x_1 = e_1$$
$$x_2 = 0,3\ e_1 + g\ e_2$$
$$x_3 = 0,3\ e_1 + 0,6\ g\ e_2 + 0,8\ g\ e_3$$
$$x_4 = x_1 + 0,5\ x_2 + 0,3\ x_3 + 0,5\ e_4$$
$$x_5 = 0,5\ x_1 + x_2 + 0,5\ e_5$$
$$y = x_1 + 0,5\ x_2 + 0,3\ x_3 + 0,5\ e \quad ,$$

où : $g = \sqrt{1-(0,3)^2} = \sqrt{0,91}$ et , les e_i (i=1,2,...5) et e sont indépendants et distribués normalement avec une moyenne nulle et un écart-type égal à 1 .

Moyennes, écarts-type et coefficients de corrélation sont donnés dans le tableau suivant, ainsi que les valeurs des critères de sélection correspondant aux régressions multiples relatives à tous les sous-ensembles possibles de 5 variables.

Pour ce modèle, le meilleur ensemble de variables est l'ensemble des 3 variables (x_1,x_2,x_3) , car x_4 et x_5 ne contribuent en rien à y malgré la corrélation relativement élevée entre x_4 et y -qui est dûe à la combinaison linéaire : $x_1 + 0,5\ x_2 + 0,3\ x_3$ commune à x_4 et y .

Variable	Moyenne	Ecart-type	Coefficient de corrélation					
			x_1	x_2	x_3	x_4	x_5	y
x_1	-0.405	1.087	1	.171	.204	.769	.560	.787
x_2	-0.386	0.922		1	.751	.649	.790	.615
x_3	-0.116	1.136			1	.660	.719	.619
x_4	-0.622	1.433				1	.886	.869
x_5	-0.661	1.176					1	.784
y	-0.667	1.692						1

p	n° des variables sélectionnées					$PRESS_p$	RSS_p	$100R_p^2$	$100\bar{R}_p^2$	$100R_p^{*2}$	C_p
1	1					35.583	31.624	61.95	60.59	59.32	56.71
		2				59.366	51.695	37.80	35.58	33.50	110.47
			3			58.647	51.271	38.31	36.11	34.04	109.33
				4		23.967	20.333	75.53	74.66	73.84	26.47
					5	37.597	32.034	61.45	60.08	58.79	57.82
2	1	2				14.748	11.858	85.73	84.67	83.68	5.77
	1		3			15.979	13.398	83.87	82.68	81.56	9.91
		2	3			21.717	17.455	78.99	77.44	75.98	20.77
	1			4		21.608	17.352	79.12	77.57	76.12	20.48
		2		4		58.114	46.990	43.46	39.27	35.35	99.87
			3	4		25.199	19.963	75.98	74.20	72.53	27.47
	1				5	38.733	32.030	61.46	58.60	55.93	59.80
		2			5	24.747	20.032	75.89	74.11	72.44	27.67
			3		5	38.516	31.503	62.09	59.29	56.66	58.39
				4	5	26.422	20.255	75.63	73.82	72.13	28.25
3	1	2	3			12.317	9.774	88.24	86.88	85.61	2.18
	1	2		4		15.859	11.820	85.77	84.13	82.60	7.68
	1		3	4		15.447	11.858	85.73	84.08	82.54	7.77
		2	3	4		17.133	13.086	84.25	82.43	80.73	11.06
	1	2			5	15.650	12.598	84.84	83.09	81.45	9.75
	1		3		5	22.877	16.325	80.35	78.09	75.97	19.74
		2	3		5	26.883	19.901	76.05	73.29	70.70	29.32
	1			4	5	40.451	31.327	62.30	57.96	53.89	59.93
		2		4	5	26.913	19.950	75.99	73.22	70.63	29.45
			3	4	5	27.232	20.027	75.90	73.12	70.52	29.65
4	1	2	3	4		12.909	9.386	88.70	86.89	85.20	3.16
	1	2	3		5	12.690	9.528	88.53	86.70	84.98	3.53
	1	2		4	5	16.811	11.803	85.79	83.52	81.40	9.63
	1		3	4	5	18.499	12.596	84.84	82.41	80.15	11.75
		2	3	4	5	28.693	19.884	76.07	72.24	68.66	31.27
5	1	2	3	4	5	14.166	9.329	88.77	86.43	84.24	5

HAGA et OKUNO ont calculé les valeurs des critères $PRESS_p$, RSS_p, R_p^2, $\bar{R}_p^2$, R_p^{*2} pour toutes les régressions possibles.

R_p^{*2} est un estimateur d'une fonction de Γ_p (Cf. HAGA et OKUNO) et

on a : $R_p^{*2} = 1 - \frac{(n+p)(n-1)}{(n+1)(n-p)} (1 - R_p^2)$.

Ils n'ont pas utilisé le critère C_p dont nous avons pu calculer les valeurs en utilisant la relation entre C_p et R_p^2 (Cf. paragraphe IV.4).

On peut, d'autre part, montrer que C_p et R_p^{*2} sont liés par la relation suivante :

$$C_p = \frac{(n-p)(n+r)}{n+p} \frac{1-R_p^{*2}}{1-R_r^{*2}} + 2p - n \qquad \text{si } \hat{\sigma}^2 = RMS_r .$$

Les critères $PRESS_p$, R_p^{*2} et C_p amènent à sélectionner le même ensemble de variables (x_1, x_2, x_3) , tandis que $\bar{R}_p^2$ amène à l'ensemble (x_1, x_2, x_3, x_4).

Lorsque $p = 1$, 2 ou 3 , tous les critères donnent le même sous-ensemble de variables. Par contre, lorsque $p = 4$, $PRESS_p$ retient $(x_1, x_2, x_3$ et $x_5)$ tandis que les autres critères retiennent $(x_1, x_2, x_3$ et $x_4)$. Il est important de remarquer que $PRESS_p$ ne maximise pas toujours le coefficient de corrélation multiple pour un nombre donné de variables.

HACA et OKUNO ont aussi simulé 9 ensembles d'observations toujours à partir du modèle précédent. Ils ont appliqué à ces 9 ensembles de données des méthodes de sélection par étapes et les méthodes de sélection selon les critères $\bar{R}^2$, R^{*2} et PRESS . Le tableau ci - après nous permet d'analyser les différences dans les résultats. Les méthodes par étapes ne concordent pas toujours entre elles et ne sélectionnent pas toujours l'ensemble (x_1, x_2, x_3). C'est l'introduction ascendante qui donne les moins bons résultats. Les sélections qui amènent le plus souvent à retenir les variables (x_1, x_2, x_3) sont faites à partir des critères R^{*2} et PRESS . Cette analyse est en accord avec la conclusion du paragraphe III.

	Introduction "ascendante"	Régression pas à pas	Elimination "descendante"	Variables sélectionnées d'après les critères suivants		
				$\bar{R}^2$	R^{*2}	PRESS
1	(1,2,4,5)	(1,2,3)	(1,2,3)	(1,2,3,5)	(1,2,3)	(1,2,3)
2	(1,3,4,5)	(1,3,5)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
3	(1,2,4)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
4	(1,2,3,4)	(1,2,3)	(1,2,3)	(1,2,3,4)	(1,2,3)	(1,2,3)
5	(1,2,3,4)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
6	(1,2,3,4)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
7	(1,2,3,4)	(1,2,3,5)	(1,2,3,5)	(1,2,3,5)	(1,2,3,5)	(1,2,3,5)
8	(1,2,3,4)	(1,2,3,4)	(1,2,3,4)	(1,2,3,4)	(1,2,3,4)	(1,2,3,4)
9	(1,2,4)	(1,2,4)	(1,2,4)	(1,2,3)	(1,2,3)	(1,2,3)

VI. CONCLUSION

Ce problème de sélection de variables se pose dans tous les domaines d'application de la statistique. Cet article avait pour but de faire le point sur les fondements des diverses méthodes utilisées en régression multiple et d'analyser les différences dans les résultats obtenus. Nous avons aussi montré qu'on pouvait affiner le choix des variables en décomposant le critère C_p , ce qui semble notamment intéressant pour les modèles de prévision.

Le problème de la sélection de variables dans des modèles de régression non linéaire n'a pas été envisagé ici, mais a été en particulier étudié par PEDUZZI, HARDY et HOLFORD [32], et la sélection par étapes pour les modèles linéaires logistiques a fait l'objet du programme LR de BMDP .

BIBLIOGRAPHIE

[1] M.A. AITKIN (1974) : Simultaneous inference and the choice of variable subsets in multiple regression. Technometrics 16, 221-228.

[2] D.M. ALLEN (1971) : Mean Square Error of Prediction as a criterion for selecting variables. Technometrics 13, 469-475.

[3] D.M. ALLEN (1974) : The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125-7.

[4] A.J. BARR, J.H. GOODNIGHT, J.P. SALL et J.T. HELLWIG (1979) : A user's guide to SAS . Raleigh, North Carolina : SAS institute.

[5] E.M.L. BEALE (1970) : A note on procedures for variable selection in multiple regression. Technometrics 12, 909-14.

[6] K.N. BERK (1978) : Comparing subset regression procedures. Technometrics 20, 1-6.

[7] J. BRENOT, P. CAZES, N. LACOURLY (1975) : Pratique de la régression : qualité et protection. Cahiers du B.U.R.O. n° 23.

[8] F. CAILLIEZ, J.P. PAGES (1976) : Introduction à l'analyse des données. SMASH.

[9] A.P. DEMPSTER, M. SCHAFZOFF et N. WERMUTH (1977) : A simulation study of alternatives to ordinary least squares. JASA 72, 77-9.

[10] G. DERFLINGER, H. STAPPLER (1976) : A correct test for the stepwise regression analysis. 76 COMPSTAT 2, 131-138.

[11] G. DIEHR et D.R. HOFLIN (1974) : Approximating the distribution of the sample R^2 in best subset regressions. Technometrics 16, 317-20.

[12] W.J. DIXON, M.B. BROWN (1981) : BMDP Biomedical Computer Programs. Statistical software. Berkeley : University of California Press.

[13] N.R. DRAPER, I. GUTTMAN, H. KANEMASU (1971) : The distribution of certain regression statistics. Biometrika 58, 295-8.

[14] N.R. DRAPER, H. SMITH (1966) : Applied regression analysis. Wiley, New-York.

[15] M.A. EFROYMSON (1960) : Multiple regression analysis ; dans RALSTON-WILF : Mathematical methods for digital computers, Wiley, New-York.

[16] A.B. FORSYTHE, L. ENGELMAN, R. JENNRICH (1973) : A stopping rule for variable selection in multiple regression. JASA 68, 75-7.

[17] G.M. FURNIVAL, R.W. Jr. WILSON (1974) : Regressions by leaps and bounds. Technometrics 16, 499-512.

[18] J.W. GORMAN, R.J. TOMAN (1966) : Selection of variables for fitting equations to data. Technometrics 8, 27-51.

[19] R.F. GUNST et R.L. MASON (1977) : Advantages of examining multicollinearities in regression analysis. Biometrics 33, 249-60.

[20] T. HAGA, T. OKUNO (1976) : Selection of variables in multiple regression analysis. Lecture note in mathematics 550, 713-722.

[21] D.M. HAWKINS, W.J.R. EPELTT (1982) : The Cholesky factorization of the inverse correlation or covariance matrix in multiple regression. Technometrics 24, 191-8.

[22] R.R. HOCKING (1972) : Criteria for selection of a subset regression : which one should be used ? Technometrics 14, 967-970.

[23] R.R. HOCKING (1976) : The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

[24] L.R. LAMOTTE (1978) : Bayes linear estimators. Technometrics 29, 281-90.

[25] G.P. McCABE (1978) : Evaluation of regression coefficient estimates using α-acceptability. Technometrics 20, 131-9.

[26] C.C. McDONALD et R.C. SCHWING (1973) : Instabilities of regression estimates relating air pollution to mortality. Technometrics 15, 463-481.

[27] R.J. McKAY (1979) : The adequacy of variable subsets in multivariate regression. Technometrics 21, 475-479.

[28] C.L. MALLOWS (1973) : Some comments on C_p . Technometrics 15, 661-75.

[29] E.R. MANSFIELD, J.T. WEBSTER et R.F. GUNST (1977) : An analytic variable selection technique for principal component regression. Applied statistics 26, 34-40.

[30] S.C. NARULA et J.F. WELLINGTON (1977) : Prediction, linear regression and the minimum sum of relatives errors. Technometrics 19, 185-190.

[31] H.H. NIE, C.H. HULL, J.G. JENKINS, J. STEINBRENNER ET D.H. BENT (1978): SPSS, Statistical Package for the Social Science, Second Edition, New-York : McGRAW-HILL.

[32] P.N. PEDUZZI, R.J. HARDY et T.R. HOLFORD (1980) : A stepwise variable selection procedure for nonlinear regression models. Biometrics 36, 511-16.

[33] P.T. POPE, J.T. WEBSTER (1972) : The use of an F-statistic in stepwise regression procedures. Technometrics 14, 327-40.

[34] A.C. RENCHER et F.C. PUN (1980) : Inflation of R^2 in best subset regression. Technometrics 22, 49-53.

[35] T. SAWA (1978) : Information criteria for discriminating among alternative regression models. Econometrica 46, 1273-91.

[36] E. SPJØTVOLL (1977) : Alternatives to plotting C_p in multiple regression. Biometrika 64, 1-8.

[37] D. SPREVAK (1976) : Statistical properties of estimates of linear models. Technometrics 18, 283-9.

[38] J. ULMO (1971) : Problèmes et programmes de régression. R.S.A. 19, 27-39.

[39] S. WEISBERG (1981) : A statistic for allocating C_p to individual cases. Technometrics 23, 27-31.

[40] L. WILKINSON, G.E. DALLAL (1981) : Tests of significance in forward selection regression with an F-to-enter stopping rule. Technometrics 23, 377-380.