REVUE DE STATISTIQUE APPLIQUÉE

N. H. FONTON

R. PALM

Comparaison empirique de méthodes de prédiction en régression linéaire multiple

Revue de statistique appliquée, tome 46, n° 3 (1998), p. 53-64.

<http://www.numdam.org/item?id=RSA_1998__46_3_53_0>

Rev. Statistique Appliquée, 1998, **XLVI** (3), 53-64

COMPARAISON EMPIRIQUE DE MÉTHODES DE PRÉDICTION EN RÉGRESSION LINÉAIRE MULTIPLE

N.H. Fonton (1), R. Palm (2)

(1) Faculté des Sciences Agronomiques, Université Nationale du Bénin, Campus d'Abomey-Calavi, BP 526, Cotonou (Bénin)

(2) Faculté Universitaire des Sciences Agronomiques, Avenue de la Faculté d'Agronomie, 8, 5030 Gembloux (Belgique)

RÉSUMÉ

Dans cet article, on étudie les erreurs des prédictions réalisées à partir d'équations de régression linéaires multiples établies en combinant différentes méthodes de sélection des variables à diverses méthodes d'estimation des coefficients de régression. L'étude repose sur des données simulées.

La régression pseudo-orthogonale avec sélection des variables et calcul des coefficients par la méthode suggérée par Lawless et Wang (1976) a donné, dans l'ensemble, les meilleurs résultats. Toutefois, le gain de précision n'est que de l'ordre de 3 à 4 % par rapport à la méthode classique des moindres carrés avec sélection des variables pas à pas (*stepwise*).

Les simulations ont montré également que le pouvoir prédictif des équations estimées est sensiblement plus faible que le pouvoir prédictif des modèles théoriques. En particulier, aucune méthode n'a permis d'établir des équations permettant de réaliser des prédictions valables lorsque le nombre de variables explicatives potentielles est supérieur au nombre d'observations dans l'échantillon et que, simultanément, le coefficient de détermination théorique est faible.

Mots-clés : *régression linéaire, prédiction, sélection des variables, estimation des coefficients de régression.*

ABSTRACT

A Monte Carlo simulation study of the predictions given by multiple linear regression equations has been performed. 78 models of regression have been considered by combining 13 methods of variables selection with 6 methods of estimation of regression coefficients.

Ridge regression with variables selection and estimation of the regression coefficients as suggested by Lawless and Wang (1976) gives the overall best results. Nevertheless, the increase in precision is only about 3 or 4 percent, with regard to the widely used ordinary least squares estimation combined with the stepwise selection procedure.

It has also been observed that the square root of the average squared error of prediction is much larger than the residual variance of the true model. None of the compared methods leads up to regression equations which are usefull for making predictions when the number of

predictors is greater than the number of observations and when, in the meantime, the value of R^2 for the true model is rather small.

Keywords : *linear regression, prediction, variables selection, estimation of regression coefficients.*

1. Introduction

L'établissement des équations de régression multiple est un problème auquel se trouvent fréquemment confrontés de nombreux utilisateurs de l'outil statistique. Un exemple concret typique est la construction de modèles statistiques-empiriques de prévision des rendements des cultures. Il s'agit, dans ce cas, d'exprimer les rendements observés au cours des 20 ou 30 années antérieures en fonction des conditions météorologiques observées durant ces années, ces conditions pouvant être décrites par des centaines de variables.

Parmi les études relatives à l'établissement des équations de régression, un grand nombre d'articles ont été consacrés à la sélection des variables. Des synthèses relatives à ce sujet sont données par Hocking (1976), Miller (1990) et Thompson (1978a, 1978b).

Une autre voie abondamment explorée concerne la recherche de méthodes alternatives à la régression classique pour réduire l'effet de la colinéarité. Ces méthodes ont été passées en revue par Palm et Iemma (1995), notamment. Cette dernière approche du problème n'est cependant pas indépendante de la précédente, car plusieurs de ces méthodes alternatives proposent également la suppression de variables redondantes, selon des critères qui peuvent être différents de ceux traditionnellement utilisés en régression multiple classique.

Le but de la présente étude est d'analyser la qualité des prédictions réalisées à partir d'équations de régression établies par diverses méthodes de sélection des variables, combinées à diverses méthodes d'estimation des coefficients de régression. L'étude repose sur des données simulées, afin de permettre un meilleur contrôle des différents facteurs pris en considération.

Nous présenterons d'abord les différentes méthodes d'établissement d'équations qui ont été comparées (paragraphe 2). Ensuite, nous décrirons le plan de simulation des données qui a été adopté (paragraphe 3). Nous examinerons et discuterons alors les résultats obtenus (paragraphe 4) avant de tirer quelques conclusions (paragraphe 5).

2. Méthodes comparées

Nous avons combiné treize procédures de sélection de variables à six méthodes d'estimation des coefficients de régression des variables sélectionnées. Nous présentons d'abord les méthodes d'estimation et nous examinons ensuite les procédures de sélection.

Soit $\boldsymbol{X}_A$ la matrice des p variables explicatives qui ont été sélectionnées parmi les k variables potentielles et $\boldsymbol{y}$ la variable à expliquer, ces variables étant observées sur n individus. Nous considérons, en outre, que les variables ont été préalablement standardisées, de manière à ce que les moyennes soient nulles et que les sommes des carrés soient égales à l'unité.

La première méthode d'estimation envisagée est la méthode classique des moindres carrés, désignée par le sigle *MC* :

$$\widehat{\boldsymbol{\beta}}_A = \left(\boldsymbol{X}'_A \boldsymbol{X}_A\right)^{-1} \boldsymbol{X}'_A \boldsymbol{y}.$$

Les méthodes suivantes font appel à la régression pseudo-orthogonale :

$$\widehat{\boldsymbol{\beta}}_A(d) = \left(\boldsymbol{X}'_A \boldsymbol{X}_A + d\boldsymbol{I}\right)^{-1} \boldsymbol{X}'_A \boldsymbol{y} \quad (d \geqslant 0).$$

Elles se distinguent par la valeur adoptée pour la constante d. Nous avons en effet retenu, d'une part, la valeur d_1 proposée par Hoerl *et al.* (1975) et, d'autre part, la valeur d_2 proposée par Lawless et Wang (1976) :

$$d_1 = p\,\widehat{\sigma}^2 \Big/ \left(\widehat{\boldsymbol{\beta}}'_A \widehat{\boldsymbol{\beta}}_A\right) \quad \text{et} \quad d_2 = p\,\widehat{\sigma}^2 \Big/ \left(\widehat{\boldsymbol{\beta}}'_A \boldsymbol{X}'_A \boldsymbol{X}_A \widehat{\boldsymbol{\beta}}_A\right),$$

$\widehat{\sigma}^2$ étant la variance résiduelle estimée à partir de la régression au sens des moindres carrés :

$$\widehat{\sigma}^2 = \left(\boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'_A \boldsymbol{X}'_A \boldsymbol{y}\right) \Big/ (n - p - 1).$$

Ces méthodes seront respectivement désignées, par la suite, par les sigles *HKB* et *LW*.

La quatrième méthode, notée *JS*, repose sur l'estimateur de James et Stein (1961) :

$$\widehat{\boldsymbol{\beta}}_A(JS) = c\,\widehat{\boldsymbol{\beta}}_A,$$

avec :

$$c = 1 - \frac{(p-2)(n-p-1)}{(n-p+1)\,p\,F_{obs}},$$

si l'expression est positive et $c = 0$ si l'expression est négative (Sclove, 1968).Dans cette expression, F_{obs} est le rapport entre le carré moyen lié au modèle et le carré moyen résiduel, traditionnellement donné dans le tableau d'analyse de la variance associé à la régression ordinaire au sens des moindres carrés.

En ce qui concerne les deux dernières méthodes d'estimation testées, elles prennent en compte le biais de sélection, qui résulte du fait que les mêmes données sont utilisées pour la sélection des variables et pour l'estimation des coefficients des variables sélectionnées.

Ce biais de sélection est la différence entre l'espérance mathématique du vecteur des coefficients de régression estimés correspondant à tous les vecteurs $\boldsymbol{y}$ qui satisfont la condition de sélection de l'ensemble A de variables explicatives et l'espérance mathématique des coefficients pour tous les vecteurs possibles de $\boldsymbol{y}$, indépendamment du fait qu'ils conduisent ou non à la sélection de l'ensemble A de variables explicatives (Miller, 1990). Le principe des méthodes basées sur le biais de sélection est d'estimer ce biais et de le soustraire de l'estimation obtenue par la méthode des moindres carrés. Pour estimer ce biais, deux procédures décrites par Miller (1990) et illustrées à l'aide d'un exemple numérique par Fonton (1995), ont été testées. Il s'agit de la méthode

de Monte Carlo d'estimation progressive du biais et de la méthode de vraisemblance conditionnelle. Ces méthodes seront désignées, respectivement, par les sigles *CA* et *ML*.

Pour sélectionner les variables à faire figurer dans les équations nous avons retenu la procédure classique de choix des variables pas à pas (*stepwise selection*), la procédure de sélection progressive (*forward selection*), la procédure de sélection basée sur les tests de signification des coefficients de régression calculés par régression pseudo-orthogonale et la procédure de sélection des variables basée sur l'analyse en composantes principales proposée par Baskerville et Toogood (1982).

La méthode pas à pas se distingue de la méthode progressive par la possibilité d'exclusion du modèle d'une variable préalablement introduite, alors que dans la méthode progressive les variables introduites à une étape donnée sont définitivement incluses dans le modèle. Pour ces méthodes, l'introduction d'une variable (et la sortie d'une variable, dans le cas de la régression pas à pas) se fait en fixant un seuil de signification, qui a été exprimé en fonction de la valeur observée de la variable t de Student.

Trois valeurs ont été retenues : $t = 1,5$, $t = 2$ et $t = 2,5$. Elles correspondent à des niveaux de probabilité de l'ordre de 15 %, 5 % et 1,5 %. Ces méthodes de sélection seront désignées, par la suite, par les sigles *sw1,5*, *sw2* et *sw2,5* pour les procédures pas à pas et par les sigles *prog1,5*, *prog2* et *prog2,5* pour les méthodes progressives.

Pour les sélections basées sur les tests de signification des coefficients de régression estimés par régression pseudo-orthogonale, une régression sur toutes les variables a été réalisée en utilisant, d'une part, la valeur d_1, proposée par Hoerl *et al.* (1975) et, d'autre part, la valeur d_2 proposée par Lawless et Wang (1976). Ces constantes ont été définies ci-dessus, mais dans le cas présent, l'ensemble A correspond à toutes les variables explicatives potentielles. Les tests de signification ont été réalisés pour les trois niveaux de signification ($t = 1,5$, t =2 et $t = 2,5$) et les variables non significatives ont été éliminées. Les méthodes seront désignées, par la suite, par *hkb1,5*, *hkb2* et *hkb2,5* dans le cas de l'utilisation de d_1, et par *lw1,5*, *lw2* et *lw2,5*, dans le cas de l'utilisation de d_2.

Enfin, la technique inspirée de Baskerville et Toogood (1982), notée par la suite *bt*, est basée sur l'analyse en composantes principales (ACP) de la matrice formée par la variable à expliquer et les k variables explicatives. Soit :

$$\boldsymbol{Z} = [\boldsymbol{y}\,\boldsymbol{X}],$$

la matrice des $k+1$ variables, l_j ($j = 1,\ k+1$) les valeurs propres et $\boldsymbol{c}_j$ les vecteurs propres normés à l'unité issus de l'ACP de $\boldsymbol{Z}$.

Si on divise chacun des vecteurs propres par $\sqrt{l_j}$, soit :

$$\boldsymbol{d}_j = \boldsymbol{c}_j / \sqrt{l_j}\,,$$

et si on juxtapose ces résultats, on obtient la matrice $\boldsymbol{D}$, dont l'élément général est d_{ij} ($i = 0, \ldots, k; j = 1, \ldots, k+1$). Sur la base de cette matrice $\boldsymbol{D}$, la sélection des variables a été opérée de la manière suivante. Parmi les colonnes de $\boldsymbol{D}$, correspondant à des valeurs propres inférieures ou égales à 0,10, on repère celle qui présente la valeur

la plus élevée de d_{0j} et, dans cette colonne, on retient les variables i pour lesquelles la valeur d_{ij} $(i = 1, \ldots, k)$ est supérieure, en valeur absolue, à l'unité.

Les treize procédures de sélection, combinées aux six méthodes d'estimation, donnent donc lieu à 78 techniques d'établissement d'équations de régression. Elles seront représentées par la juxtaposition des sigles. Ainsi, la technique résultant de l'estimation des paramètres par la régression pseudo-orthogonale en utilisant la constante d_2, les variables ayant été préalablement sélectionnées par la procédure pas à pas avec pour valeur théorique $t = 2$, sera notée *LW*sw2*.

La combinaison des méthodes de sélection des variables et des méthodes d'estimation des coefficients des variables sélectionnées peut conduire à des techniques de régression qui présentent sans doute peu de justifications théoriques. Elle a cependant l'avantage de permettre, lors de l'analyse des résultats, la séparation des deux facteurs et l'examen de leur éventuelle interaction.

D'autre part, et plus particulièrement en ce qui concerne le choix des variables, de nombreuses autres méthodes proposées dans la littérature n'ont pas été retenues, de manière à maintenir le volume des calculs dans des limites raisonnables. C'est ainsi, par exemple, que nous n'avons pas retenu les différents critères de sélection tels que AIC, BIC ou DEC, etc., étudiés par Kundu et Murali (1996).

Enfin, comme nous le préciserons au paragraphe 3, nous avons procédé à une présélection des variables lorsque le nombre de variables potentielles, k, est supérieur au nombre d'observations, n. En effet, il est impossible, dans ces cas, d'estimer les coefficients de régression avec les techniques basées sur le principe des moindres carrés. Cette situation, couramment rencontrée dans la pratique, constitue un problème supplémentaire pour l'établissement des équations de régression.

3. Génération des données

L'objectif est d'analyser le comportement des techniques d'établissement d'équations de régression, en relation notamment avec la structure des variables explicatives potentielles. Pour définir cette structure, deux facteurs ont été pris en considération. Il s'agit du nombre de variables explicatives disponibles et du degré de colinéarité des variables.

Trois valeurs ont été retenues pour le nombre de variables potentielles : $k = 10$, $k = 20$ et $k = 40$. Ces valeurs nous paraissent couvrir une large gamme de situations couramment rencontrées en pratique.

La colinéarité a été quantifiée par la moyenne des inverses des valeurs propres. Cet indice, noté *IC*, correspond aussi à la trace de l'inverse de la matrice de corrélation des variables explicatives, divisée par le nombre de variables. Trois valeurs de cet indice ont été retenues : $IC = 1$, $IC = 20$ et $IC = 40$. La première valeur correspond à l'absence de corrélation entre les variables et les deux autres valeurs représentent des degrés croissants de la colinéarité.

La procédure retenue pour la génération des variables explicatives est celle utilisée par Hoerl *et al.* (1986), qui permet de modifier le degré de colinéarité d'une matrice. Les valeurs propres utilisées ont été obtenues par la procédure de Bendel et Afifi (1977).

Les variables à expliquer ont été générées à partir du modèle suivant :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

où $\boldsymbol{X}$ est la matrice des k variables explicatives potentielles, $\boldsymbol{\beta}$ est le vecteur des coefficients de régression et $\boldsymbol{\varepsilon}$ est le vecteur des résidus. Ces résidus sont des réalisations indépendantes de variables normales réduites. Le vecteur $\boldsymbol{\beta}$ contient cinq valeurs positives et égales et $k - 5$ valeurs nulles. Les positions dans le vecteur $\boldsymbol{\beta}$ des valeurs non nulles ont été choisies au hasard et la valeur des cinq coefficients identiques et non nuls a été déterminée de manière à obtenir une valeur fixée du coefficient de détermination multiple. Deux valeurs de ce coefficient ont été retenues : $R^2 = 0,4$ et $R^2 = 0,8$.

Pour chaque matrice $\boldsymbol{X}$, caractérisée par un nombre k de variables ($k = 10$, $k = 20$ et $k = 40$) et par un indice de colinéarité *IC* (*IC* = 1, *IC* = 20 et *IC* = 40) et pour chaque valeur de R^2 ($R^2 = 0,4$ et $R^2 = 0,8$), deux vecteurs $\boldsymbol{y}$ ont été générés, par simple permutation aléatoire des positions occupées, dans le vecteur $\boldsymbol{\beta}$, par les coefficients non nuls.

Pour chacune des neuf structures des données explicatives (trois valeurs de k combinées à trois valeurs de *IC*), 3.100 individus ont été générés. Vingt échantillons de 15 individus, vingt échantillons de 30 individus et vingt échantillons de 60 individus ont été sélectionnés sans remise. Les 1.000 individus non sélectionnés sont réservés à l'évaluation de la qualité des prédictions (paragraphe 4). Sur chacun des 60 échantillons, les 78 techniques de régression sont appliquées quatre fois : sur les deux variables $\boldsymbol{y}$ caractérisées par un coefficient de détermination multiple théorique de 0,4 et sur les deux variables $\boldsymbol{y}$ caractérisées par un coefficient de détermination multiple théorique de 0,8. Au total 168.480 équations ont été établies.

Comme nous l'avons signalé au paragraphe 2, nous avons procédé à une présélection des variables, lorsque le nombre de variables potentielles, k, est supérieur au nombre d'observations n. Ainsi, douze variables ont été présélectionnées lorsque $n = 15$ et $k = 20$ ou $k = 40$ et 24 variables ont été présélectionnées lorsque $n = 30$ et $k = 40$. Pour opérer cette présélection, nous avons procédé de la façon suivante. Dans le cas $n = 15$ et $k = 20$, les variables ont été réparties aléatoirement en deux groupes de 10 variables. Dans chacun de ces groupes, nous avons recherché la combinaison de 6 variables qui a conduit à la variance résiduelle minimum. Les deux ensembles de 6 variables ainsi obtenus représentent les 12 variables présélectionnées. Pour $n = 30$ et $k = 40$ les variables ont été réparties en deux groupes de 20 variables et 12 variables ont été sélectionnées dans chaque groupe. Pour $n = 15$ et $k = 40$, les variables ont été réparties en 4 groupes de 10 variables et 3 variables ont été sélectionnées dans chaque groupe.

4. Analyse des résultats

Le critère retenu pour comparer les différentes techniques d'établissement des équations de régression présentées au paragraphe 2 est l'erreur quadratique moyenne de prédiction. Elle est basée sur les écarts entre les valeurs observées, y_i, et les valeurs estimées par une équation donnée, $\widehat{y}_i$, les écarts étant calculés pour les 1.000 individus

réservés à la validation dont il a été question au paragraphe 3 :

$$\overline{e}_q = \sqrt{\frac{1}{1.000} \sum_{i=1}^{1.000} (y_i - \widehat{y}_i)^2}.$$

Rappelons que l'écart-type résiduel théorique est égal à l'unité et que l'écart-type total théorique de la variable à expliquer est égal à 1,29 si $R^2 = 0,4$ et à 2,24 si $R^2 = 0,8$, compte tenu de la manière dont les vecteurs $\boldsymbol{y}$ ont été simulés (paragraphe 3).

L'analyse de la variance de l'ensemble des résultats a mis en évidence l'existence de nombreuses interactions, ce qui nous a conduit à réaliser quatre analyses séparées. Les quatre cas considérés résultent de la combinaison des deux valeurs de R^2 (0,4 et 0,8) et de deux situations concernant le rapport entre le nombre de variables explicatives potentielles, k, et le nombre d'observations dans l'échantillon, n ($k < n$ et $k > n$).

Ces quatre analyses de la variance ont montré que les différences entre méthodes de sélection, méthodes d'estimation et degré de colinéarité sont toutes significatives et que l'interaction entre les méthodes d'estimation et les méthodes de sélection est non significative lorsque $k < n$ mais qu'elle reste significative lorsque $k > n$.

Le tableau 1 donne, pour les quatre situations envisagées, les moyennes et, entre parenthèses, les écarts-types des $\overline{e}_q$ pour les treize méthodes de sélection et pour les six méthodes d'estimation. Ces écarts-types sont les écarts-types empiriques des $\overline{e}_q$ moyens, calculés pour les 20 répétitions. Ils mesurent donc essentiellement la variabilité liée aux facteurs contrôlés (structure des variables explicatives, effectifs des échantillons et méthode d'estimation ou de sélection).

Nous allons tout d'abord examiner les résultats lorsque k est inférieur à n et nous analyserons ensuite le cas où k est supérieur à n.

Pour $k < n$, le tableau montre que la meilleure méthode de sélection est la méthode *lw1,5* et la meilleure méthode d'estimation est la méthode *LW*. La combinaison *LW*lw1,5* est aussi la meilleure combinaison parmi les 78 combinaisons testées. La valeur moyenne correspondante des $\overline{e}_q$ vaut 1,18 pour $R^2 = 0,4$ et 1,24 pour $R^2 = 0,8$.

L'examen du tableau 1 montre aussi, pour $k < n$, que les différences sont plus marquées pour les méthodes de sélection que pour les méthodes d'estimation. Ce point est confirmé lorsqu'on examine le classement des 78 moyennes. Ainsi, pour $R^2 = 0,4$ et quelle que soit la méthode d'estimation, les méthodes de sélection *lw1,5* et *lw2* conduisent à des valeurs de $\overline{e}_q$ inférieures ou égales à 1,22 tandis que les méthodes *bt*, *hbk2* et *hbk2,5* donnent des valeurs moyennes de $\overline{e}_q$ comprises entre 1,26 et 1,42; pour $R^2 = 0,8$ les valeurs moyennes de $\overline{e}_q$ sont comprises entre 1,24 et 1,46 pour *lw1,5* et *lw2* et entre 1,43 et 1,74 pour *bt*, *hbk2* et *hbk2,5*.

Lorsque $k > n$, la meilleure méthode de sélection est la méthode *stepwise*, avec une valeur de t fixée à 2,5 pour $R^2 = 0,4$ et à 2 ou 2,5 pour $R^2 = 0,8$. La meilleure méthode d'estimation est la méthode *CA*. La combinaison *CA*sw2,5* conduit aux valeurs moyennes de $\overline{e}_q$ de 1,35, pour $R^2 = 0,4$, et de 1,95, pour $R^2 = 0,8$. Du fait de l'interaction entre les méthodes de sélection et les méthodes d'estimation lorsque $k > n$, la combinaison *CA*sw2,5* n'est cependant pas la meilleure combinaison. En

TABLEAU 1

Valeurs moyennes et écarts-types de $\overline{e}_q$, en fonction de la méthode d'estimation des paramètres et de la méthode de sélection des variables : valeurs pour les quatre cas envisagés.

Méthodes	$R^2 = 0,4$ et $k < n$		$R^2 = 0,4$ et $k > n$		$R^2 = 0,8$ et $k < n$		$R^2 = 0,8$ et $k > n$	
Sélection								
sw1,5	1,21	(0,11)	1,51	(0,15)	1,31	(0,20)	1,96	(0,30)
sw2	1,21	(0,09)	1,41	(0,12)	1,33	(0,24)	1,95	(0,31)
sw2,5	1,22	(0,08)	1,36	(0,08)	1,41	(0,29)	1,95	(0,29)
prog1,5	1,22	(0,11)	1,60	(0,18)	1,31	(0,19)	2,05	(0,35)
prog2	1,21	(0,09)	1,55	(0,16)	1,31	(0,21)	2,01	(0,33)
prog2,5	1,22	(0,09)	1,51	(0,15)	1,33	(0,23)	1,98	(0,31)
hkb1,5	1,24	(0,09)	1,60	(0,19)	1,37	(0,22)	2,14	(0,37)
hkb2	1,26	(0,07)	1,54	(0,17)	1,47	(0,27)	2,12	(0,33)
hkb2,5	1,28	(0,05)	1,47	(0,12)	1,63	(0,33)	2,14	(0,30)
lw1,5	1,20	(0,09)	1,54	(0,16)	1,29	(0,18)	2,01	(0,32)
lw2	1,21	(0,09)	1,50	(0,14)	1,31	(0,19)	1,97	(0,31)
lw2,5	1,23	(0,07)	1,46	(0,12)	1,40	(0,24)	1,97	(0,30)
bt	1,33	(0,15)	2,02	(0,43)	1,68	(0,19)	2,55	(0,54)
Estimation								
MC	1,26	(0,12)	1,67	(0,31)	1,38	(0,27)	2,17	(0,43)
HKB	1,23	(0,12)	1,55	(0,19)	1,36	(0,25)	2,05	(0,35)
LW	1,21	(0,09)	1,46	(0,14)	1,35	(0,25)	1,97	(0,30)
JS	1,24	(0,11)	1,63	(0,26)	1,37	(0,27)	2,14	(0,42)
CA	1,23	(0,07)	1,38	(0,10)	1,51	(0,24)	1,96	(0,28)
ML	1,24	(0,09)	1,57	(0,25)	1,40	(0,26)	2,09	(0,38)

effet, pour $R^2 = 0,4$, la meilleure combinaison est *ML*sw2,5* (moyenne des $\overline{e}_q$ égale à 1,33) alors que pour $R^2 = 0,8$, la meilleure combinaison est *CA*lw2* (moyenne des $\overline{e}_q$ égale à 1,89).

La méthode *LW*lw1,5*, qui, rappelons-le, est la meilleure combinaison pour $k < n$, a conduit, pour $k > n$, aux moyennes des $\overline{e}_q$ de 1,47 pour $R^2 = 0,4$ et de 1,93 pour $R^2 = 0,8$. Par rapport aux meilleures combinaisons mentionnées ci-dessus, on voit que la différence est faible, du moins pour $R^2 = 0,8$.

Dans la mesure où la méthode *LW*lw1,5* est globalement la meilleure méthode (moyenne générale des $\overline{e}_q$ égale à 1,37), nous avons comparé cette méthode à la méthode plus classique et plus couramment utilisée, à savoir la sélection des variables par *stepwise* avec une valeur de t fixée à 2 et l'estimation des paramètres par les moindres carrés (*MC*sw2*). Cette méthode, qui constitue donc notre référence, a donné lieu à une valeur moyenne des $\overline{e}_q$ égale à 1,42 et se situe en 18ème position lorsque les 78 méthodes sont classées par ordre croissant des moyennes des $\overline{e}_q$. En moyenne, le gain de précision est de l'ordre de 3 à 4 %. Une analyse plus détaillée

des rapports entre les $\overline{e}_q$ obtenus pour le modèle *LW*lw1,5* et les $\overline{e}_q$ obtenus pour le modèle *MC*sw2* a montré que ce rapport n'est lié ni à la taille des échantillons, ni au degré de colinéarité des matrices des variables explicatives potentielles, ni à la valeur de R^2. Par contre, il augmente avec le nombre de variables explicatives prises en considération. Ainsi, ce rapport vaut, en moyenne, 0,96 pour $k = 10$, 0,95 pour $k = 20$ et 1,00 pour $k = 40$. Dans ce dernier cas, un tiers des rapports sont supérieurs à l'unité. En définitive, on peut donc constater que le gain de précision lié à l'utilisation de la méthode *LW*lw1,5* par rapport à la méthode de référence, reste assez limité.

Indépendamment des informations relatives au choix des méthodes, le tableau 1 fait également apparaître que les valeurs moyennes des $\overline{e}_q$ sont toutes supérieures, et souvent même nettement supérieures, à l'écart-type résiduel théorique, qui a été fixé à l'unité. Une analyse plus approfondie des discordances entre les erreurs quadratiques moyennes de prédiction et l'écart-type résiduel théorique a été réalisée dans le cas du modèle *LW*lw1,5*. Nous avons, dans ce but, calculé les quantités :

$$\text{«}R^{2^*}\text{»} = \frac{\sigma_y^2 - \overline{e}_q^2}{\sigma_y^2},$$

qui représentent la part de la variance de la variable y qui est expliquée par la régression. Dans cette relation, σ_y^2 est la variance théorique de y, directement liée à la valeur de R^2. Les coefficients «R^{2^*}» sont définis et s'interprètent comme des coefficients de détermination multiple. Ils peuvent toutefois prendre des valeurs négatives, lorsque $\overline{e}_q$ est plus grand que σ_y, c'est-à-dire lorsque le modèle de régression introduit de la variabilité supplémentaire lors de la prédiction. Ces coefficients «R^{2^*}» ont ensuite été divisés par les valeurs théoriques R^2. Ce rapport indique donc quelle fraction de R^2 est effectivement expliquée par la régression : une valeur proche de l'unité signifie une bonne concordance entre l'écart-type résiduel théorique et l'écart quadratique moyen. Une valeur proche de zéro ou négative signifie que la capacité prédictive du modèle est nulle et que, par conséquent, la régression établie est sans intérêt pour la prédiction.

On a constaté que ces rapports sont essentiellement fonction de la valeur théorique de R^2 et du rapport k/n. Les distributions de ces rapports étant dissymétriques et de variances non constantes, nous avons repris, dans le tableau 2, les premier et troisième quartiles observés pour les différentes situations considérées. Dans l'interprétation de ces résultats, on ne perdra pas de vue qu'il s'agit de la distribution de valeurs moyennes. La variabilité des rapports provient principalement de la structure des données (effectif des échantillons et caractéristiques des matrices $\boldsymbol{X}$) et non du caractère aléatoire de l'échantillonnage, les valeurs $\overline{e}_q$ étant des moyennes de 20 répétitions.

L'examen du tableau 2 montre clairement l'inefficacité des équations de régression pour la prédiction lorsque le rapport k/n est supérieur à l'unité et $R^2 = 0,4$ et lorsque k/n est égal à 2,6 et $R^2 = 0,8$. Pour $R^2 = 0,4$, il faut que la taille de l'échantillon soit trois fois plus grande que le nombre de variables potentielles, k, pour que les équations de régression calculées expliquent 50 % de la variabilité expliquée par les modèles théoriques. La situation est, par contre, plus favorable lorsque $R^2 = 0,8$, du moins si $k < n$.

TABLEAU 2

Premier et troisième quartiles, q_1 et q_3, de la distribution des rapports «R^{2^*}»/R^2, en fonction de R^2 et de k/n.

k/n	$R^2 = 0,4$		$R^2 = 0,8$	
	q_1	q_3	q_1	q_3
0,17	0,61	0,83	0,94	0,96
0,33	0,49	0,66	0,89	0,93
0,67	0,03	0,47	0,77	0,86
1,33	–1,09	0,00	0,28	0,67
2,67	–1,57	–0,84	–0,22	0,29

5. Conclusions

L'objectif de cette étude est de comparer la qualité des prédictions réalisées à partir d'équations de régression établies de diverses manières : treize méthodes de sélection des variables ont été combinées à six méthodes d'estimation des paramètres des variables sélectionnées. Les comparaisons ont été réalisées sur base de données simulées.

On s'est volontairement placé dans un contexte défavorable, caractérisé par un nombre élevé de variables explicatives potentielles ($k = 10$, $k = 20$ et $k = 40$) par rapport à la taille des échantillons ($n = 15$, $n = 30$ et $n = 60$). De plus, on a considéré qu'on ne dispose d'aucune connaissance a priori concernant les variables explicatives supposées utiles ou concernant les ordres de grandeur des coefficients de régression ou des éventuelles contraintes liant ces coefficients. Nous pensons, en effet, que de telles situations défavorables sont courantes dans la pratique.

Plusieurs conclusions peuvent être tirées de cette étude. Tout d'abord on peut constater que, pour $k < n$, la méthode *LW*lw1,5* a donné les meilleurs résultats. Cette méthode correspond à la régression pseudo-orthogonale avec sélection des variables et calcul des coefficients par la méthode suggérée par Lawless et Wang (1976), la valeur de la variable t de Student étant fixée à 1,5. Pour $k > n$ et pour un coefficient de détermination théorique élevé ($R^2 = 0,8$), la méthode ci-dessus n'est que très légèrement moins bonne que la méthode *CA*lw2*, qui est, dans ce cas, la meilleure. Comparée à la méthode plus classique des moindres carrés avec choix des variables par sélection pas à pas, la régression pseudo-orthogonale n'offre cependant qu'un gain de précision de l'ordre de 3 à 4 %.

Cette constatation confirme les conclusions de Draper et Van Nostrand (1979). Selon ces auteurs, les améliorations par rapport aux moindres carrés ordinaires sont très faibles lorsque le vecteur β est bien estimé, c'est-à-dire si la colinéarité n'est pas un problème sérieux et si β n'est pas trop proche de zéro; d'autre part, si β est mal estimé du fait de la colinéarité ou parce qu'il est proche de zéro, l'importance de l'amélioration est loin d'être manifeste. De nombreuses autres études destinées à comparer les performances de la régression pseudo-orthogonale n'ont d'ailleurs pu mettre en évidence la supériorité systématique de la régression pseudo-orthogonale. Des informations à ce sujet sont données notamment par Van Nostrand, en commentaire à l'article de Smith et Campbell (1980).

On a constaté également qu'aucune méthode ne permet d'établir une équation conduisant à des prédictions valables si le nombre de variables explicatives potentielles est supérieur au nombre d'observations dans l'échantillon et que, simultanément, le coefficient de détermination théorique est faible ($R^2 = 0,4$). Dans ce cas, en effet, l'écart quadratique moyen des erreurs de prédiction est supérieur à l'écart-type marginal de la variable à expliquer.

De façon plus générale, on a noté que le pouvoir prédictif des équations estimées est sensiblement plus faible que le pouvoir prédictif du modèle théorique, sauf si le nombre d'individus est au moins trois fois plus grand que le nombre de variables et que le coefficient de détermination théorique est élevé ($R^2 = 0,8$).

Cette conclusion généralise une constatation que nous avons pu faire à plusieurs reprises lors de la construction de modèles agrométéorologiques de type statistique-empirique pour la prévision des rendements des cultures : la sélection des variables et l'estimation des paramètres des variables sélectionnées conduit à des modèles qui s'ajustent bien aux données, mais dont les capacités prédictives sont nulles (Palm et Dagnelie, 1993). La présente étude montre clairement que le problème de l'établissement des équations utiles pour la prédiction ne peut pas être résolu par le simple remplacement d'un algorithme de sélection des variables et d'estimation des paramètres, comme la méthode des moindres carrés pas à pas, par un autre algorithme, comme, par exemple, la régression pseudo-orthogonale, ou par la prise en compte du biais de sélection. La solution, si elle existe, se trouve sans doute dans l'augmentation de l'information disponible. Cette augmentation peut se faire soit par l'addition de nouvelles observations ou par la prise en compte d'information externe, concernant la nature de la relation recherchée. Une telle information externe peut permettre une réduction a priori du nombre de variables ou l'imposition de contraintes sur les paramètres. Des informations à ce sujet et des exemples d'applications sont donnés par Cazes (1975, 1977, 1978) et par Cazes et Turpin (1971).

Bibliographie

BASKERVILLE J.C., TOOGOOD J.H. (1982). Guided regression modelling for prediction and exploration of structure with many explanatory variables. *Technometrics* 24, p. 9-17.

BENDEL R.B., AFIFI A.A. (1977). Comparison of stopping rules in forward stepwise regression. *J. Amer. Stat. Assoc.* 72, p. 46-53.

CAZES P. (1975). Protection de la régression par utilisation de contraintes linéaires et non linéaires. *Rev. Stat. Appl.* 23, p. 37-57.

CAZES P. (1977). Estimation biaisée et estimation sous contraintes dans le modèle linéaire. *In : Premières journées internationales «Analyse des données et informatique».* Versailles, I.R.I.A., p. 223-232.

CAZES P. (1978). Méthodes de régression : la régression sous contraintes. *Cah. Anal. Données* 3, p. 147-165.

CAZES P., TURPIN P.Y. (1971). Régression sous contraintes : application à l'estimation de la courbe granulométrique d'un aérosol. *Rev. Stat. Appl.* 19, p. 23-44.

DRAPER N.R., VAN NOSTRAND R.C. (1979). Ridge regression and James-Stein estimation : review and comments. *Technometrics* 21, p. 451-466.

FONTON H.N. (1995). *Comparaison des méthodes de prédiction en régression linéaire multiple.* Gembloux, Faculté universitaire des Sciences agronomiques, 230 p.

HOCKING R.R (1976). The analysis and selection of variables in linear regression. *Biometrics* 32, p. 1-49.

HOERL A.E., KENNARD R.W., BALDWIN K.F. (1975). Ridge regression : some simulations. *Comm. Stat.* A4, p. 105-123.

HOERL R.W., SCHUENEMEYER J.H., HOERL A.E (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics* 28, p. 369-380.

JAMES W., STEIN C. (1961). Estimation with quadratic loss. *In* : Neyman J. (ed.). *Proceedings of the fourth Berkeley Symposium* (vol. 1). Berkeley, University of California Press, p. 361-379.

KUNDU D., MURALI G. (1996). Model selection in linear regression. *Comput. Stat. Data Anal.* 22, p. 461-469.

LAWLESS J.F., WANG P. (1976). A simulation study of ridge and other regression estimators. *Comm. Stat.* A5, p. 307-323.

MILLER A.J. (1990). *Subset selection in regression.* Chapman and Hall, London, 229 p.

PALM R., DAGNELIE P. (1993). *Tendance générale et effets du climat dans la prévision des rendements agricoles des différents pays des Communautés européennes.* Luxembourg, Office des publications officielles des Communautés Européennes, 128 p.

PALM R., IEMMA A.F. (1995). Quelques alternatives à la régression classique dans le cas de la colinéarité. *Rev. Stat. Appl.* 43, p. 5-33.

SCLOVE S.L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Stat. Assoc.* 63, p. 596-606.

SMITH G., CAMPBELL F. (1980). A critique of some ridge regression methods. *J. Amer. Stat. Assoc.* 75, p. 74-91.

THOMPSON M. (1978a). Selection of variables in multiple regression. Part I : a review and evaluation. *Int. Stat. Rev.* 46, p. 1-19.

THOMPSON M. (1978b). Selection of variables in multiple regression. Part II : chosen procedures, computation and examples. *Int. Stat. Rev.* 46, p. 129-146.