REVUE DE STATISTIQUE APPLIQUÉE

R. PALM
A. F. IEMMA

Quelques alternatives à la régression classique dans le cas de la colinéarité

Revue de statistique appliquée, tome 43, n° 2 (1995), p. 5-33
<http://www.numdam.org/item?id=RSA_1995__43_2_5_0>

Rev. Statistique Appliquée, 1995, **XLIII** (2), 5-33

QUELQUES ALTERNATIVES À LA RÉGRESSION CLASSIQUE DANS LE CAS DE LA COLINÉARITÉ

R. Palm[1], A.F. Iemma[2]

RÉSUMÉ

Cette note décrit, de façon succincte, les principes de quelques méthodes de régression biaisée : régression en fonction des composantes principales, régression proposée par Webster, Gunst et Mason, régression par les moindres carrés partiels, régression pseudo-orthogonale et utilisation des estimateurs de James et Stein. Toutes ces méthodes sont illustrées par un exemple numérique simple.

SUMMARY

In this note, we briefly describe some biased regression methods : principal component regression, latent root regression analysis, partial least square regression, ridge regression, and use of James-Stein's estimators. All these methods are illustrated by an example.

1. Introduction

La minimisation de la somme des carrés des écarts résiduelle est certainement le critère le plus souvent retenu quand il s'agit d'ajuster une équation de régression linéaire multiple à un ensemble de données.

Des raisons à la fois pratiques et théoriques justifient le recours systématique à ce critère d'ajustement. En effet, l'estimation des coefficients de régression par les moindres carrés ordinaires peut être réalisée sans difficulté par n'importe quel logiciel statistique et, d'autre part, les estimateurs des moindres carrés jouissent, sous certaines conditions d'application, d'un ensemble de propriétés intéressantes.

[1] Chef de travaux et Maître de conférences à la Faculté des Sciences agronomiques de Gembloux.

[2] Professeur titulaire à l'*Escola Superior de Agricultura* «Luiz de Queiroz», Piracicaba, São Paulo (Brésil), et Maître de conférences à la Faculté des Sciences agronomiques de Gembloux durant l'année académique 1990-1991.

Toutefois, lorsque les variables explicatives présentent des phénomènes de multicolinéarité, l'estimation par les moindres carrés ordinaires a notamment l'inconvénient de conduire à des estimateurs dont les variances sont très grandes.

Le remède généralement utilisé dans ce cas est la réduction de l'intensité de la colinéarité par l'élimination d'une ou de plusieurs variables explicatives. Se pose alors le problème, que nous n'aborderons pas dans cette note, du choix des variables à faire figurer dans l'équation.

Une autre approche du problème consiste à faire appel à des techniques de régression spécialement mises au point pour atténuer les effets de la multicolinéarité. Nous nous proposons d'examiner les principes des méthodes les plus courantes dans ce domaine et de les illustrer par un exemple numérique simple.

Nous rappelons d'abord quelques notions de régression classique, en insistant particulièrement sur les problèmes liés à la colinéarité (paragraphe 2). Ensuite, nous décrivons successivement trois techniques de régression basées sur le calcul de nouvelles variables explicatives : il s'agit de la régression en fonction des composantes principales (paragraphe 3), de la régression proposée par Webster, Gunst et Mason (paragraphe 4) et de la régression par les moindres carrés partiels (paragraphe 5). Nous envisageons alors deux méthodes basées sur des estimateurs «rétrécis» : la régression pseudo-orthogonale (paragraphe 6) et la régression utilisant les estimateurs de James et Stein (paragraphe 7). Enfin, nous terminons par une discussion (paragraphe 8).

2. Régression classique au sens des moindres carrés

Soit le modèle théorique suivant :

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

où $\boldsymbol{x}$ est le vecteur, de dimensions $1 \times p$, relatif aux variables explicatives, $\boldsymbol{\beta}$ est le vecteur des p coefficients de régression théoriques et ε est le résidu. On considère en outre que les résidus relatifs à des individus différents sont des réalisations indépendantes d'une même variable aléatoire normale de moyenne nulle et d'écart-type σ. En pratique, l'objectif poursuivi est d'estimer $\boldsymbol{\beta}$ à partir de données observées sur n individus choisis de manière aléatoire et simple et pour lesquels le modèle ci-dessus est applicable. Soit $\boldsymbol{y}$ le vecteur, de dimensions $n \times 1$, des valeurs observées de la variable à expliquer et $\boldsymbol{X}$ la matrice, de dimensions $n \times p$, des variables explicatives. L'estimation au sens des moindres carrés du vecteur $\boldsymbol{\beta}$ consiste à minimiser la quantité $\boldsymbol{e}'\,\boldsymbol{e}$, $\boldsymbol{e}$ étant le vecteur, de dimensions $n \times 1$, des n résidus observés, qui intervient dans la relation matricielle suivante :

$$\boldsymbol{y} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}} + \boldsymbol{e}\,,$$

avec :

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

De manière générale, les équations de régression comportent un terme indépendant. Pour en tenir compte, il suffit de considérer qu'une variable explicative est constante et égale à l'unité : le vecteur $\boldsymbol{x}$ du modèle théorique comporte alors un élément supplémentaire et la matrice $\boldsymbol{X}$ une colonne supplémentaire, par rapport au nombre de variables explicatives. Une autre solution consiste à centrer les variables par rapport à leurs moyennes. Cette seconde solution sera adoptée dans cette note. De plus, pour uniformiser et simplifier la présentation de diverses méthodes de régression, nous considérons non seulement que les observations sont centrées mais encore standardisées, de manière telle que la somme des carrés de chacune des variables soit égale à l'unité. En désignant par y_i^0 et par x_{ij}^0 les observations initiales des $p+1$ variables, on a :

$$y_i = \frac{y_i^0 - \overline{y}^0}{\sqrt{\text{SCE}_{y^0}}} \quad \text{et} \quad x_{ij} = \frac{x_{ij}^0 - \overline{x}_j^0}{\sqrt{\text{SCE}_{x_j^0}}} \quad (i = 1, \ldots, n\,;\ j = 1, \ldots, p)\,,$$

SCE_{y^0} et $\text{SCE}_{x_j^0}$ étant les sommes des carrés des écarts des variables $\boldsymbol{y}^0$ et $\boldsymbol{x}_j^0$. De cette manière, la matrice $\boldsymbol{X}'\boldsymbol{X}$ est la matrice de corrélation des p variables explicatives et le vecteur $\boldsymbol{X}'\boldsymbol{y}$ est le vecteur des corrélations simples de $\boldsymbol{y}$ avec chacune des variables explicatives.

A titre d'illustration, nous considérons l'exemple proposé par Dagnelie (1982) concernant l'étude de la relation entre le rendement du froment d'hiver (y_i^0, en quintaux par hectare) et quatre variables météorologiques (x_{i1}^0 = précipitations des mois de novembre et décembre, en mm; x_{i2}^0 = température moyenne du mois de juillet, en degrés centigrades; x_{i3}^0 = précipitations du mois de juillet, en mm; x_{i4}^0 = radiation du mois de juillet, en ml d'alcool mesurés à l'actinomètre de Bellani). Les données concernent onze années successives ($i = 1, \ldots, 11$), de l'année culturale 1920-1921 à l'année culturale 1930-1931. Le tableau 1 reprend les données standardisées.

TABLEAU 1

Variable à expliquer ($\mathbf{y}$) et variables explicatives ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ et $\mathbf{x}_4$) : données standardisées.

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	0,4043	−0,3568	0,4562	−0,6329	0,5755
2	−0,2817	−0,3426	−0,6314	0,1349	−0,3772
3	0,0568	0,1066	0,4809	−0,1538	0,1521
4	0,0121	−0,0422	−0,1865	0,1426	0,0696
5	0,1523	−0,3504	0,1348	0,1659	−0,1228
6	−0,2042	0,5899	0,0112	0,2796	0,0600
7	0,3491	−0,1447	0,0112	−0,0771	−0,1902
8	0,4043	−0,2066	0,1348	−0,2813	0,4559
9	−0,1043	0,1287	0,0112	−0,1469	−0,0280
10	−0,5964	0,3095	−0,2359	0,5700	−0,4679
11	−0,1923	0,3088	−0,1865	−0,0013	−0,1270

L'estimation au sens des moindres carrés de $\boldsymbol{\beta}$, si on prend en considération les quatre variables, est :

$$\widehat{\boldsymbol{\beta}}' = [-0,4234 \;\; 0,2742 \;\; -0,1871 \;\; 0,2871] \; .$$

A partir des moyennes, $\overline{y}^0$ et $\overline{\mathbf{x}}^0$, et des sommes des carrés des écarts des variables, SCE_{y^0} et $\mathrm{SCE}_{x_j^0}$, données par Dagnelie (1982) :

$$\overline{y}^0 = 25,659 \quad \text{et} \quad \overline{x}^0 = [138,03 \;\; 17,755 \;\; 74,45 \;\; 1.242,4 \;\; 25,659] \; ,$$
$$\mathrm{SCE}_{y^0} = 44,9581 \quad \text{et} \quad \mathbf{SCE}_{x^0} = [19.732,98 \;\; 16,3673 \;\; 13.468,59 \;\; 529.147] \; ,$$

on peut déterminer l'équation de régression pour les variables exprimées dans les unités originales :

$$\widehat{\beta}_j^0 = \widehat{\beta}_j \sqrt{\frac{\mathrm{SCE}_{y^0}}{\mathrm{SCE}_{x_j^0}}} \; ,$$

soit : $$\widehat{\boldsymbol{\beta}}^{0\prime} = [-0,02021 \;\; 0,4545 \;\; -0,01081 \;\; 0,002646] \; ,$$

et : $$\beta_0 = \overline{y}^0 - \overline{\mathbf{x}}^0 \, \widehat{\boldsymbol{\beta}}^0 = 17,89 \, .$$

L'équation s'écrit finalement :

$$y^0 = 17,89 - 0,0202\, x_1^0 + 0,455\, x_2^0 - 0,0108\, x_3^0 + 0,00265\, x_4^0 \, .$$

La caractéristique essentielle de l'estimateur des moindres carrés est d'être l'estimateur linéaire non biaisé de variance minimum, la matrice de variances et covariances des paramètres étant égale à :

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}' \boldsymbol{X})^{-1} \, .$$

D'autre part, le carré moyen de l'erreur de $\widehat{\boldsymbol{\beta}}$ est égal à :

$$E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sigma^2 \, \mathrm{tr} \, (\boldsymbol{X}' \boldsymbol{X})^{-1} \, ,$$

$\mathrm{tr}(\boldsymbol{X}' \boldsymbol{X})^{-1}$ représentant la trace de la matrice $(\boldsymbol{X}' \boldsymbol{X})^{-1}$. Ce carré moyen est en fait la somme des variances des p coefficients de régression. La quantité $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ est une mesure de la distance qui sépare, dans l'espace des paramètres, le vecteur estimé $\widehat{\boldsymbol{\beta}}$ du vecteur inconnu $\boldsymbol{\beta}$. Notons que cette distance ne présente un intérêt que parce que les variables $\boldsymbol{x}_j$ ont été préalablement standardisées.

Parmi les estimateurs linéaires non biaisés, l'estimateur des moindres carrés, $\widehat{\boldsymbol{\beta}}$, est celui qui minimise ce carré moyen, puisque, dans ce cas, il est de variance minimum. Toutefois, comme nous le verrons par la suite, il existe des estimateurs biaisés qui peuvent donner lieu à un carré moyen de l'erreur plus faible.

Pour l'exemple présenté ci-dessus, la matrice des variances et covariances des paramètres est égale à :

$$\widehat{\sigma}^2(\boldsymbol{X}'\,\boldsymbol{X})^{-1} = 0,03420 \begin{bmatrix} 1,5159 & -0,3776 & -1,3298 & -0,4342 \\ -0,3776 & 2,3007 & 0,5038 & -1,3810 \\ -1,3298 & 0,5038 & 4,0633 & 2,5946 \\ -0,4342 & -1,3810 & 2,5946 & 4,0125 \end{bmatrix}.$$

Il s'agit en réalité de la matrice estimée, puisqu'on a remplacé la variance résiduelle σ^2 par son estimation :

$$\widehat{\sigma}^2 = \boldsymbol{e}'\,\boldsymbol{e}/(n-p-1)\,.$$

L'estimation du carré moyen de l'erreur de $\widehat{\boldsymbol{\beta}}$ est, par conséquent, égale à :

$$\widehat{\sigma}^2\,\mathrm{tr}\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1} = (0,03420)(11,8924) = 0,40672\,.$$

On peut constater que les formules relatives au calcul de $\widehat{\boldsymbol{\beta}}$, $\mathbf{V}(\widehat{\boldsymbol{\beta}})$ et $\mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$ font intervenir l'inverse de la matrice $\boldsymbol{X}'\,\boldsymbol{X}$ et ne peuvent donc être utilisées que si la matrice $\boldsymbol{X}'\,\boldsymbol{X}$ est non singulière, c'est-à-dire s'il n'existe pas de relations linéaires entre les variables explicatives. L'existence de relations linéaires conduit à la situation connue sous le nom de colinéarité exacte entre les variables explicatives. Dans ce cas, le vecteur $\widehat{\boldsymbol{\beta}}$ est indéterminé, une infinité de vecteurs différents conduisant à la même valeur minimum de $\boldsymbol{e}'\,\boldsymbol{e}$. Si la matrice $\boldsymbol{X}'\,\boldsymbol{X}$ est quasi singulière, on se trouve en présence d'un phénomène de colinéarité approximative, qui, d'une part, cause des problèmes de précision numérique lors du calcul des coefficients et, d'autre part, conduit à des variances des coefficients et donc aussi à un carré moyen de l'erreur de $\widehat{\boldsymbol{\beta}}$ importants, comme nous le montrerons au paragraphe suivant.

Pour réduire les inconvénients liés à la colinéarité sans remettre en cause le principe des moindres carrés, il y a lieu de supprimer une ou plusieurs variables explicatives. On se trouve alors confronté au problème du choix des variables en régression multiple, que nous ne discuterons pas dans cette note. Une synthèse bibliographique a été faite par Hocking (1976) et Thompson (1978a, 1978b) et les résultats obtenus par différentes procédures de sélection dans le cas de l'exemple présenté ci-dessus sont donnés par Dagnelie (1982).

Des informations complémentaires relatives à la régression multiple classique peuvent être trouvées, par exemple, dans les livres de Draper et Smith (1981), de Theil (1971) ou de Weisberg (1985).

3. Régression en fonction des composantes principales

La régression en fonction des composantes principales, appelée aussi régression orthogonalisée[3], repose sur l'utilisation, comme variables explicatives, des valeurs des composantes principales calculées à partir de la matrice $\boldsymbol{X}$.

Soit $\lambda_1, \ldots, \lambda_r$ les r valeurs propres non nulles de $\boldsymbol{X}' \boldsymbol{X}$, qui, compte tenu de la standardisation des variables adoptée (paragraphe 2), est la matrice de corrélation des variables initiales. En l'absence de colinéarité, r est égal à p; sinon r est plus petit que p. Soit aussi :

$$\boldsymbol{U} = [\boldsymbol{u_1} \quad \boldsymbol{u_2} \quad \ldots \quad \boldsymbol{u_r}] ,$$

la matrice, de dimensions $p \times r$, constituée des r vecteurs propres normés à l'unité correspondant aux r valeurs propres de $\boldsymbol{X}' \boldsymbol{X}$. La matrice des composantes principales, de dimensions $n \times r$, est alors égale à :

$$\boldsymbol{Z} = [\boldsymbol{z_1} \quad \boldsymbol{z_2} \quad \ldots \quad \boldsymbol{z_r}] = \boldsymbol{X}\, \boldsymbol{U} .$$

Les n éléments du vecteur $\boldsymbol{z_k}$ $(k = 1, \ldots, r)$ sont les valeurs de la $k^{\text{ième}}$ composante pour les n individus et, compte tenu de la standardisation utilisée, la somme des carrés des éléments de ce vecteur vaut λ_k.

Nous allons d'abord examiner le cas où il n'y a pas colinéarité exacte et où on introduit dans l'équation de régression l'ensemble des p variables $\boldsymbol{z_k}$, ce qui permet de bien déterminer l'incidence de la colinéarité exacte ou approximative sur la qualité du vecteur $\widehat{\boldsymbol{\beta}}$. Nous envisagerons ensuite le cas, plus pratique, où la variable à expliquer est mise en relation avec un nombre plus réduit de composantes principales.

En l'absence de colinéarité exacte et si on prend en considération les p composantes principales, on a :

$$\boldsymbol{y} = \boldsymbol{Z}\, \widehat{\boldsymbol{\alpha}} + \boldsymbol{e} ,$$

et le vecteur $\widehat{\boldsymbol{\alpha}}$ est obtenu par la méthode des moindres carrés :

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Z}' \boldsymbol{Z})^{-1} \boldsymbol{Z}' \boldsymbol{y} = \boldsymbol{\Lambda}^{-1} \boldsymbol{Z}' \boldsymbol{y} ,$$

$\boldsymbol{\Lambda}^{-1}$ étant la matrice diagonale dont les éléments sont les inverses des valeurs propres de $\boldsymbol{X}' \boldsymbol{X}$. Cette dernière égalité résulte du fait que les variables $\boldsymbol{z_k}$ sont non corrélées, de moyennes nulles et de sommes de carrés égales à λ_k. On a donc aussi :

$$\widehat{\alpha}_k = \boldsymbol{z_k'}\, \boldsymbol{y} / \lambda_k = r_{yz_k} / \sqrt{\lambda_k} \quad (k = 1, \ldots, p) ,$$

r_{yz_k} étant le coefficient de corrélation de $\boldsymbol{y}$ et $\boldsymbol{z_k}$.

---

[3] En anglais : *principal component regression.*

Disposant du vecteur $\widehat{\boldsymbol{\alpha}}$, on obtient facilement le vecteur $\widehat{\boldsymbol{\beta}}$ par la relation :

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{U}\,\widehat{\boldsymbol{\alpha}}\,,$$

car :

$$\widehat{\boldsymbol{y}} = \boldsymbol{Z}\,\widehat{\boldsymbol{\alpha}} = \boldsymbol{X}\,\boldsymbol{U}\,\widehat{\boldsymbol{\alpha}}\,.$$

Le vecteur $\widehat{\boldsymbol{\beta}}$ ainsi obtenu par l'intermédiaire des composantes principales est théoriquement identique au vecteur $\widehat{\boldsymbol{\beta}}$ obtenu par les moindres carrés. L'intérêt du passage par les composantes principales peut être numérique, car on évite l'inversion d'une matrice qui, dans certains cas, est quasi singulière. Il est également théorique, car il permet, notamment, d'expliciter la variance des coefficients de régression $\widehat{\beta}_j$ en fonction des valeurs propres de $\boldsymbol{X}'\,\boldsymbol{X}$ afin de mettre en évidence l'incidence de la colinéarité.

En effet, suite à la non-corrélation des composantes principales, les variances des coefficients de régression $\widehat{\alpha}_k$ sont égales à :

$$\mathrm{V}\,(\widehat{\alpha}_k) = \sigma^2/\lambda_k \quad (k = 1, \ldots, p)\,.$$

Il en résulte que :

$$\mathrm{V}\,(\widehat{\beta}_j) = \mathrm{V}\left(\sum_{k=1}^{p} u_{jk}\,\widehat{\alpha}_k\right) = \sigma^2 \sum_{k=1}^{p} \frac{u_{jk}^2}{\lambda_k} \quad (j = 1, \ldots, p)\,,$$

u_{jk} étant le $j^{\text{ième}}$ élément du $k^{\text{ième}}$ vecteur propre. Ces variances sont donc d'autant plus grandes que les éléments j des vecteurs propres $\boldsymbol{u}_k$ sont importants alors que la valeur propre correspondante λ_k est faible.

D'autre part, grâce aux propriétés des inverses (Graybill, 1969), on sait que les valeurs propres de l'inverse d'une matrice non singulière sont égales aux inverses des valeurs propres de cette même matrice. On peut donc affirmer que les valeurs propres de $(\boldsymbol{X}'\,\boldsymbol{X})^{-1}$ sont égales aux inverses des λ_k $(k = 1, \ldots, p)$. Il en résulte que le carré moyen de l'erreur de $\widehat{\boldsymbol{\beta}}$ est égale à :

$$\mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sigma^2\,\mathrm{tr}\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1} = \sigma^2 \sum_{k=1}^{p} \frac{1}{\lambda_k}\,.$$

Ce carré moyen est donc d'autant plus grand que les dernières valeurs propres de $\boldsymbol{X}'\,\boldsymbol{X}$ sont faibles et la présence d'une ou de plusieurs valeurs propres de $\boldsymbol{X}'\,\boldsymbol{X}$ tendant vers zéro fait tendre le carré moyen vers l'infini.

Si la matrice $\boldsymbol{X}'\,\boldsymbol{X}$ est singulière, on ne dispose que de r valeurs propres non nulles et de r composantes principales. La formule relative à la variance de $\widehat{\beta}_j$ reste

d'application, en considérant que l'indice k varie de 1 à r et non plus de 1 à p. Le passage par les composantes principales permet d'obtenir une solution particulière parmi l'infinité de solutions qui conduisent au même minimum de la somme des carrés des écarts résiduelle, alors que la méthode des moindres carrés ordinaires ne le permet pas, sauf, bien sûr, si on élimine $p - r$ variables choisies parmi celles qui sont impliquées dans le phénomène de colinéarité ou en ajustant l'équation globale avec $p - r$ contraintes linéaires sur les coefficients. La solution obtenue par l'intermédiaire des composantes principales traduit alors ce phénomène de colinéarité. Ainsi, s'il existe, par exemple, entre x_1, x_2, x_3 et x_4 une relation du type :

$$x_4 = x_1 + x_2 - x_3 ,$$

on obtiendra une solution telle que :

$$\widehat{\beta}_4 = \widehat{\beta}_1 + \widehat{\beta}_2 - \widehat{\beta}_3 .$$

Les relations donnant $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ et $\mathrm{V}(\widehat{\beta}_j)$ restent également valables lorsqu'on ne prend en considération que q composantes parmi les r composantes disponibles, que ces q composantes soient les composantes associées aux plus grandes valeurs propres de $\boldsymbol{X}'\boldsymbol{X}$ ou non. Il suffit de considérer que $\boldsymbol{U}$ est constitué des q vecteurs propres correspondant aux composantes retenues et de considérer que l'indice k prend les valeurs correspondant aux q composantes retenues. Le vecteur $\widehat{\boldsymbol{\beta}}$ qu'on obtient alors est cependant différent du vecteur $\widehat{\boldsymbol{\beta}}$ obtenu par les moindres carrés. Nous le noterons $\widehat{\boldsymbol{\beta}}_{cp}$. En particulier, il ne s'agit plus d'un estimateur non biaisé. Mais le biais introduit peut être largement compensé par la réduction de la variance des $\widehat{\beta}_j$, si on néglige des composantes correspondant à des valeurs propres faibles et que ces composantes sont peu corrélées à la variable explicative.

Pour l'exemple traité, on a :

$$\lambda_1 = 2,5932 , \quad \lambda_2 = 0,9783 , \quad \lambda_3 = 0,2852 , \quad \lambda_4 = 0,1433 ,$$

et

$$\boldsymbol{U} = \begin{bmatrix} -0,2908 & 0,8713 & 0,3322 & 0,2142 \\ 0,5062 & 0,4248 & -0,7423 & 0,1107 \\ -0,5773 & 0,1360 & -0,4184 & -0,6879 \\ 0,5709 & 0,2046 & 0,4043 & -0,6846 \end{bmatrix} .$$

Les composantes principales sont données dans le tableau 2.

Ce tableau donne aussi les coefficients de corrélation de $\boldsymbol{y}$ avec chacune des composantes ainsi que le coefficient de détermination multiple des équations obtenues pour un nombre croissant de composantes. Comme les composantes sont, par construction, non corrélées, le coefficient R^2 s'obtient très facilement en cumulant les $r^2_{yz_k}$. Enfin, ce tableau donne encore les coefficients de régression, $\widehat{\alpha}_k$, de $\boldsymbol{y}$ en fonction de chacune des composantes principales.

TABLEAU 2

Valeurs des composantes principales, coefficients de corrélation simple de $\mathbf{y}$ *et des composantes,* r_{yz_k}*, coefficients de détermination multiple en fonction du nombre de composantes principales retenues,* R^2*, et coefficients de régression,* $\widehat{\alpha}_k$.

	Composantes			
i	1	2	3	4
1	1,0286	−0,0855	0,0403	0,0154
2	−0,5132	−0,6256	0,1459	0,0221
3	0,3880	0,3073	−0,1957	0,0777
4	−0,1247	−0,0823	0,0929	−0,1755
5	0,0042	−0,2506	−0,3356	−0,0902
6	−0,2930	0,5691	0,0949	−0,1058
7	−0,0163	−0,1707	−0,1011	0,1535
8	0,5510	−0,0677	0,1333	−0,1480
9	0,0371	0,0912	0,0845	0,1490
10	−0,8056	0,1512	−0,1497	−0,0316
11	−0,2560	0,1636	0,1902	0,1333
r_{yz_k}	0,860	−0,217	−0,080	−0,049
R^2	0,739	0,786	0,792	0,795
$\widehat{\alpha}_k$	0,5339	−0,2191	−0,1499	−0,1282

Le tableau 3 donne les coefficients de régression $\widehat{\beta}_{cp}$ obtenus par les régressions sur les composantes principales lorsqu'on retient la première, les deux premières, les trois premières ou les quatre composantes principales et le tableau 4 donne, à la constante σ^2 près, les variances de ces coefficients.

TABLEAU 3

Vecteurs $\widehat{\beta}_{cp}$*, en fonction du nombre de composantes principales retenues*

	Nombres de composantes			
Variables	1	2	3	4
x_1	−0,1552	−0,3461	−0,3959	−0,4234
x_2	0,2703	0,1772	0,2884	0,2742
x_3	−0,3082	−0,3380	−0,2753	−0,1871
x_4	0,3048	0,2599	0,1993	0,2871

TABLEAU 4

Variances des coefficients de régression en fonction du nombre de composantes principales retenues (toutes les valeurs doivent être multipliées par σ^2).

Variables	Nombres de composantes 1	2	3	4
x_1	0,0326	0,8086	1,1956	1,5159
x_2	0,0988	0,2833	2,2151	2,3007
x_3	0,1285	0,1474	0,7613	4,0633
x_4	0,1257	0,1685	0,7416	4,0125
Totaux	0,3856	1,4078	4,9137	11,8924

L'examen de ces tableaux montre, d'une part, que les vecteurs $\widehat{\beta}_{cp}$ sont modifiés de façon non négligeable lorsque le nombre de composantes retenues augmente et, d'autre part, que les variances augmentent très nettement avec le nombre de composantes retenues. On peut vérifier également que, si on tient compte des quatre composantes, on retrouve les résultats donnés pour la régression ordinaire au paragraphe 2.

On constate que la première composante explique près de 74 % de la variabilité de y, que les deux premières composantes expliquent 79 %, alors que les quatre composantes n'expliquent que 80 %. Il y a donc intérêt à négliger les deux dernières, et éventuellement même les trois dernières composantes, qui n'apportent guère d'information pour y, mais qui sont responsables de la forte augmentation de la variance des coefficients de régression.

Pour cet exemple, la valeur absolue du coefficient de corrélation de y avec les composantes principales diminue lorsque le numéro d'ordre de la composante augmente. Il est donc tout à fait naturel de retenir les q premières composantes et de négliger les $p - q$ dernières composantes. D'une façon générale cependant, les corrélations, prises en valeur absolue, ne sont pas nécessairement décroissantes. Des exemples sont donnés par Jolliffe (1982), notamment. Néanmoins, selon Jackson (1991), la suppression automatique des composantes associées à de faibles valeurs propres est une pratique fort répandue. Jolliffe (1982) rappelle que l'idée originale était de traiter les composantes principales de la même manière que des variables explicatives ordinaires et d'évaluer l'opportunité de leur introduction dans l'équation. Greenberg (1975) signale que la prise en considération de composantes correspondant à de faibles valeurs propres augmente la variance des coefficients de régression mais diminue le biais, si ces composantes sont corrélées avec y.

Dans certaines situations cependant, la prise en considération des premières composantes principales uniquement peut se justifier si on considère, *a priori*, que les dernières composantes prennent en compte des fluctuations aléatoires dans les variables explicatives. C'est le cas par exemple, dans les problèmes de calibrage en spectrométrie (Martens et Naes, 1989). Il s'agit, dans ce cas, de trouver une relation permettant d'exprimer la teneur en un élément donné, les protéines par exemple,

en fonction des réponses aux différentes longueurs d'onde. Les teneurs en protéines, déterminées par voie chimique classique sur un ensemble d'individus, des échantillons de blé par exemple, constituent le vecteur $\boldsymbol{y}$ et les hauteurs des spectres aux différentes longueurs d'onde (de 1.100 à 2.500 nanomètres, par pas de 2 nanomètres, par exemple) constituent la matrice $\boldsymbol{X}$. Pour l'exemple considéré, on disposerait de 700 variables explicatives. Par ailleurs, on peut considérer que les spectres sont, d'une part, fonction de la constitution chimique de l'objet soumis au spectromètre et, d'autre part, fonction d'une série de facteurs extérieurs tels que la température, l'humidité, le mode de préparation de l'objet, etc.

L'idée sous-jacente à l'utilisation de la régression en fonction des composantes principales est la suivante. On s'efforce d'extraire de $\boldsymbol{X}$ l'information utile, qui est vraisemblablement contenue dans les premières composantes, et on néglige les fluctuations «aléatoires» ou le «bruit» qui est contenu dans les dernières composantes principales. Comme on sait que (Palm, 1993) :

$$\boldsymbol{X} = \sum_{i=1}^{r} \boldsymbol{X}_i ,$$

on peut encore écrire : $$\boldsymbol{X} = \boldsymbol{X}_u + \boldsymbol{R} ,$$

avec :

$$\boldsymbol{X}_u = \sum_{i=1}^{q} \boldsymbol{X}_i \quad \text{et} \quad \boldsymbol{R} = \sum_{i=q+1}^{r} \boldsymbol{X}_i ,$$

$\boldsymbol{X}_u$ représentant donc l'information utile et $\boldsymbol{R}$ représentant le bruit. Les matrices $\boldsymbol{X}_i$ étant égales à :

$$\boldsymbol{X}_i = \boldsymbol{z}_i \, \boldsymbol{u}_i' ,$$

on a aussi :

$$\boldsymbol{X} = \boldsymbol{Z} \, \boldsymbol{U}' + \boldsymbol{R} ,$$

$\boldsymbol{R}$ étant une matrice de résidus.

Implicitement donc, la régression en fonction des premières composantes principales revient en fait à accepter un modèle linéaire pour $\boldsymbol{X}$, et ensuite un autre modèle linéaire pour $\boldsymbol{y}$.

Il faut noter aussi qu'il existe des algorithmes permettant d'extraire les valeurs propres et les vecteurs propres en séquence, en commençant par les valeurs propres les plus grandes (Jackson, 1991). De tels algorithmes sont particulièrement utiles lorsque le nombre de variables explicatives est élevé et qu'on ne retient que les premières composantes principales. Dans ce cas, il n'est, en effet, pas nécessaire de déterminer les dernières composantes.

L'analyse en composantes principales peut aussi fournir des informations sur le choix des variables explicatives de départ, en relation avec le problème de colinéarité, et ce quelle que soit la technique d'ajustement retenue. Un exemple pratique est

donné par Palm (1988). Le problème a été étudié en relation avec la régression en fonction des composantes principales par Mansfield *et al.* (1977), notamment, en vue d'éliminer un des inconvénients majeurs de la régression en fonction des composantes principales, qui est de fournir une équation de régression contenant toutes les variables explicatives initiales, même si des composantes principales ont été négligées.

Enfin, notons encore que différents indices sont proposés dans la littérature afin de quantifier le degré de colinéarité existant entre les variables explicatives (Jackson, 1991; Stewart, 1987). Un des plus simples est la racine carrée du rapport entre la plus grande et la plus petite valeur propre de $\boldsymbol{X}'\boldsymbol{X}$. Pour l'exemple traité, ce rapport vaut 4,25, ce qui correspond à un degré de colinéarité relativement faible.

4. Méthode de Webster, Gunst et Mason

Une variante de la régression en fonction des composantes principales a été proposée par Webster *et al.* (1974) sous l'appellation de régression par l'analyse des valeurs latentes[4]. Ces auteurs suggèrent de calculer les composantes non pas à partir de la matrice $\boldsymbol{X}'\boldsymbol{X}$, mais à partir de $\boldsymbol{A}'\boldsymbol{A}$ avec :

$$\boldsymbol{A} = [\boldsymbol{y}\ \boldsymbol{X}] \ .$$

Ils incluent donc la variable à expliquer dans le calcul des composantes principales.

En l'absence de colinéarité, on dispose alors de $p+1$ valeurs propres et de $p+1$ vecteurs propres, constituant la matrice $\boldsymbol{U}$:

$$\boldsymbol{U} = [\,\boldsymbol{u_1} \quad \boldsymbol{u_2} \quad \ldots \quad \boldsymbol{u_{p+1}}\,] \ ,$$

avec : $\boldsymbol{u}'_k = [\,u_{0k} \quad u_{1k} \quad \ldots \quad u_{pk}\,] \quad (k = 1, \ldots, p+1)$,

et les $p+1$ vecteurs des valeurs des composantes principales sont donnés par :

$$\boldsymbol{Z} = [\,\boldsymbol{z_1} \quad \boldsymbol{z_2} \quad \ldots \quad \boldsymbol{z_{p+1}}\,] = \boldsymbol{A}\,\boldsymbol{U} \ .$$

Avant de calculer l'équation de régression sur la base de ces composantes principales, on élimine les composantes qui sont sans grand intérêt, c'est-à-dire celles qui traduisent des phénomènes de colinéarité, exacte ou approximative, et qui ne possèdent aucun pouvoir explicatif. Plus concrètement, on élimine les composantes correspondant à une faible valeur de λ_k et pour lesquelles u_{0k} est également faible. Rappelons que λ_k mesure la variabilité de la composante k et que $u_{0k}\sqrt{\lambda_k}$ est égal au coefficient de corrélation de la $k^{\text{ième}}$ composante et $\boldsymbol{y}$. En pratique, Webster *et al.*, proposent d'éliminer la composante k lorsque, simultanément, $\lambda_k < 0,05$ et $u_{0k} < 0,10$.

Les composantes principales jugées sans intérêt étant éliminées, une technique particulière de calcul du vecteur des coefficients de régression est mise en œuvre. La

[4] En anglais : *latent root regression analysis.*

régression ordinaire de $\boldsymbol{y}$ sur les composantes n'aurait, en effet, aucun sens, puisque celles-ci sont déjà fonction linéaires de $\boldsymbol{y}$; il suffirait d'ailleurs d'inclure toutes les composantes pour que la régression conduise à une variation résiduelle nulle. Les éléments du vecteur des coefficients de régression estimé, $\widehat{\beta}_W$, s'obtiennent par les relations suivantes :

$$\widehat{\beta}_{W_j} = c \sum_{k=1}^{q} u_{jk}\, u_{0k}/\lambda_k \quad (j = 1, \ldots, p)\,,$$

avec :

$$c = -1 \Big/ \left(\sum_{k=1}^{q} u_{0k}^2/\lambda_k \right)\,,$$

les signes de sommation étant étendus aux q composantes retenues. Une justification de la procédure est donnée par Webster *et al.* D'autre part, compte tenu de la standardisation utilisée (paragraphe 2), la constante c, changée de signe, correspond à la somme des carrés des écarts résiduelle.

Pour l'exemple considéré, les cinq valeurs propres de la matrice $\boldsymbol{A}'\,\boldsymbol{A}$ sont :

$$\lambda_1 = 3,3988\,, \quad \lambda_2 = 1,0157\,, \quad \lambda_3 = 0,2947\,,$$
$$\lambda_4 = 0,16068 \quad \text{et} \quad \lambda_5 = 0,13018\,,$$

et la matrice des vecteurs propres correspondants est :

$$\boldsymbol{U} = \begin{bmatrix} 0,5024 & 0,1700 & 0,2134 & 0,6169 & 0,5409 \\ -0,2927 & -0,8093 & -0,2308 & 0,1736 & 0,4194 \\ 0,4230 & -0,4823 & 0,6774 & -0,3384 & -0,1226 \\ -0,4997 & -0,0397 & 0,5053 & 0,5846 & -0,3895 \\ 0,4830 & -0,2861 & -0,4325 & 0,3647 & -0,6040 \end{bmatrix}.$$

Les composantes principales et les coefficients de corrélation des composantes avec $\boldsymbol{y}$ sont donnés dans le tableau 5.

D'après la règle empirique de Webster *et al.*, aucune composante ne devrait être éliminée. D'autre part, on remarque aussi que les coefficients de corrélation ne diminuent pas nécessairement pour des valeurs croissantes de k. La plus forte corrélation, 0,926, s'observe pour la première composante et la plus faible corrélation, 0,116, s'observe pour la troisième composante.

Le tableau 6 donne les coefficients de régression, $\widehat{\beta}_W$, pour un nombre croissant de composantes. Il contient aussi les sommes des carrés des écarts résiduelles correspondantes. On constate que, si on prend en considération les cinq composantes, on retrouve les résultats de la régression classique au sens des moindres carrés. On constate également que, si on ne prend en considération que la première, les deux premières ou les trois premières composantes, la régression est très mauvaise, puisque la somme des carrés des écarts résiduelle est plus grande que l'unité, c'est-à-dire plus grande que la somme des carrés des écarts de $\boldsymbol{y}$. Il en sera d'ailleurs toujours ainsi en

TABLEAU 5

Valeurs des composantes principales et des coefficients de corrélation des composantes avec $\mathbf{y}$.

i	Composantes 1	2	3	4	5
1	1,0947	−0,0020	−0,0910	−0,1270	−0,0880
2	−0,5579	0,6365	−0,1775	−0,0783	−0,0433
3	0,3510	−0,3459	0,1698	−0,1436	−0,0155
4	−0,0981	0,1006	−0,0721	0,1720	−0,0859
5	0,0939	0,2730	0,3416	0,0396	−0,0716
6	−0,3813	−0,5459	−0,0567	0,1580	−0,0096
7	0,1692	0,2285	0,1588	0,0720	0,2717
8	0,6814	0,0517	−0,1140	0,1697	−0,0503
9	−0,0254	−0,1134	−0,1064	−0,1418	0,0703
10	−1,0009	−0,1269	0,1318	−0,0717	−0,1033
11	−0,3266	−0,1563	−0,1843	−0,0489	0,1255
r_{yz_k}	0,926	0,171	0,116	0,247	0,195

TABLEAU 6

Vecteurs β_W, *en fonction d'un nombre croissant de composantes, et sommes des carrés des écarts résiduelles,* SCE_r, *correspondantes.*

Variables	Nombres de composantes 1	2	3	4	5
x_1	0,5826	1,7400	1,3443	−0,1222	−0,4234
x_2	−0,8419	0,1772	−1,8359	0,3149	0,2742
x_3	0,9945	0,7836	−1,1092	−0,9635	−0,1871
x_4	−0,9613	−0,2287	1,1259	−0,4229	0,2871
SCE_r	13,4636	9,7330	3,8866	0,3809	0,2052

pratique si on considère uniquement la première composante. Dans ce cas, la somme des carrés des écarts résiduelle est, en effet, égale à λ_1/u_{01}^2 et ce rapport est plus grand que l'unité du fait de la corrélation entre les variables constituant la matrice $\boldsymbol{A}$.

Enfin, il faut encore signaler que Webster *et al.*, proposent aussi l'élimination éventuelle de variables explicatives de départ sur la base des valeurs et des vecteurs propres de $\boldsymbol{A}'\,\boldsymbol{A}$.

5. Régression par les moindres carrés partiels

L'utilisation de la régression par les moindres carrés partiels[5] s'est essentiellement développée dans le domaine de la chimie, en relation avec les problèmes de calibrage, comme celui que nous avons évoqué au paragraphe 3 (Martens et Naes, 1989).

La méthode présente à la fois des analogies avec la régression en fonction des composantes principales et avec la régression par l'analyse des valeurs latentes, décrite au paragraphe précédent.

Comme dans le cas de la régression en fonction des composantes principales, la méthode consiste à remplacer les p variables explicatives initiales par $q \leqslant r$ combinaisons linéaires $\boldsymbol{t_k}$ $(k = 1, \ldots, q)$ de ces variables :

$$\boldsymbol{T} = [\boldsymbol{t_1} \quad \boldsymbol{t_2} \quad \ldots \quad \boldsymbol{t_q}] = \boldsymbol{X}\,\boldsymbol{V}\,,$$

et à utiliser ces combinaisons linéaires comme variables explicatives :

$$\boldsymbol{y} = \boldsymbol{T}\,\widehat{\boldsymbol{\alpha}} + \boldsymbol{e}\,.$$

La détermination des vecteurs $\boldsymbol{t_k}$, que nous appellerons facteurs, se fait en tenant compte de $\boldsymbol{y}$, rejoignant pour cet aspect la méthode de Webster *et al.* (1974).

Les matrices $\boldsymbol{V}$ et $\boldsymbol{T}$ ne sont donc plus les matrices constituées des vecteurs propres de $\boldsymbol{X}'\,\boldsymbol{X}$ et des scores correspondants. Elles n'ont pas les propriétés des matrices $\boldsymbol{U}$ et $\boldsymbol{Z}$ définies au paragraphe 3.

Des algorithmes de calcul de la régression par les moindres carrés partiels sont donnés notamment par Martens et Naes (1989). Par ces algorithmes, les facteurs sont extraits un à un et on exprime successivement $\boldsymbol{y}$ en fonction d'un facteur, de deux facteurs, etc., jusqu'à q facteurs.

Le choix du nombre optimal, q, de facteurs à prendre en considération est important car l'utilisation d'un nombre de facteurs supérieur à l'optimum conduit à une rapide détérioration du modèle. Le choix de q se fait en étudiant l'évolution d'une mesure de la qualité du modèle en fonction du nombre de facteurs et on retient la valeur de q qui donne le meilleur modèle. Idéalement, cette étude doit se faire sur des individus n'appartenant pas à l'échantillon utilisé pour le calcul du modèle et on peut déterminer, par exemple, le carré moyen de l'erreur de prédiction pour un nombre croissant de facteurs. Des techniques de validation interne, c'est-à-dire basées sur les n individus de l'échantillon sont également proposées dans la littérature pour les cas où la validation externe mentionnée ci-dessus est inapplicable, par manque de données (Martens et Naes, 1989). Il faut noter que le problème de la validation des modèles de régression n'est pas propre à la régression par les moindres carrés partiels, mais dans ce cas spécifique, il est particulièrement important, du fait de la sensibilité de la méthode à la présence de facteurs excédentaires.

---

5 En anglais : *partial least squares regression.*

Pour l'exemple considéré, la matrice $\boldsymbol{V}$ est :

$$\boldsymbol{V} = \begin{bmatrix} -0,4334 & -0,9352 & -0,3304 & -0,1247 \\ 0,4562 & -0,2461 & 0,6455 & -0,3145 \\ -0,5692 & 0,1030 & 0,7244 & 0,5666 \\ 0,5292 & -0,2536 & -0,0055 & 0,7882 \end{bmatrix},$$

et les quatre facteurs sont donnés dans le tableau 7. On peut remarquer que $\boldsymbol{V}$ n'est pas une matrice orthogonale mais que les facteurs sont non corrélés.

TABLEAU 7

Valeurs des facteurs, coefficients de corrélation simple de $\boldsymbol{y}$ *et des facteurs,* r_{yt_k}, *coefficients de détermination multiple en fonction du nombre de facteurs retenus,* R^2, *et coefficients de régression,* $\widehat{\alpha}_k$.

	Facteurs			
i	1	2	3	4
1	1,0275	0,0103	−0,0493	−0,0039
2	−0,4160	0,5854	−0,1946	0,0205
3	0,3412	−0,2725	0,1630	−0,1318
4	−0,1112	0,0824	−0,0035	0,1996
5	0,0539	0,3428	0,3237	−0,0015
6	−0,3780	−0,5409	0,0146	0,1286
7	0,0110	0,1728	0,0003	−0,1791
8	0,5524	0,0155	−0,0510	0,1833
9	0,0181	−0,1311	−0,1415	−0,1249
10	−0,8138	−0,0540	0,1610	−0,0102
11	−0,2854	−0,2108	−0,2226	−0,0806
r_{yt_k}	0,877	0,146	0,060	0,018
R^2	0,770	0,791	0,794	0,795
$\widehat{\alpha}_k$	0,5491	0,1512	0,1164	0,0451

Le tableau 7 donne également les coefficients de corrélation de $\boldsymbol{y}$ avec chacun des facteurs, ainsi que le coefficient de détermination multiple des équations correspondant à un nombre croissant de facteurs, qui s'obtient en cumulant les $r^2_{yt_k}$.

Par rapport à la régression en fonction des composantes principales, on constate que, pour un même nombre de facteurs, les valeurs de R^2 sont plus élevées dans le cas des moindres carrés partiels. On remarque aussi que les valeurs absolues des coefficients de corrélation r_{yt_k} sont décroissantes pour cet exemple, mais il ne s'agit pas là d'une règle générale.

Les éléments du vecteur $\widehat{\boldsymbol{\alpha}}$ sont également donnés dans le tableau 7 et on peut sans difficulté calculer les vecteurs $\widehat{\boldsymbol{\beta}}_{PLS}$ pour un nombre quelconque, q, de facteurs.

Il suffit d'effectuer le produit :

$$\widehat{\boldsymbol{\beta}}_{PLS} = \boldsymbol{V}\,\widehat{\boldsymbol{\alpha}}\,,$$

en éliminant de $\boldsymbol{V}$ les $p-q$ dernières colonnes et en éliminant de $\widehat{\boldsymbol{\alpha}}$ les $p-q$ dernières lignes. Ces coefficients sont donnés dans le tableau 8 pour un nombre croissant de facteurs. On peut constater que, pour les quatre facteurs, on retrouve l'équation de régression multiple au sens des moindres carrés ordinaires.

TABLEAU 8

Vecteurs $\widehat{\boldsymbol{\beta}}_{PLS}$, en fonction du nombre de facteurs retenus.

Variables	Nombres de facteurs 1	2	3	4
x_1	−0,2379	−0,3793	−0,4178	−0,4234
x_2	0,2505	0,2133	0,2884	0,2742
x_3	−0,3125	−0,2969	−0,2126	−0,1871
x_4	0,2906	0,2522	0,2516	0,2871

Stone et Brooks (1990) ont montré que la régression linéaire classique, la régression en fonction des composantes principales et la régression par les moindres carrés partiels pouvaient être considérées comme des cas particuliers d'une procédure beaucoup plus générale de régression. On peut considérer en effet que, pour les trois méthodes, on définit une ou plusieurs combinaisons linéaires des variables explicatives :

$$\boldsymbol{X}\,\boldsymbol{c_1}\,,\ \boldsymbol{X}\,\boldsymbol{c_2}\,,\ \ldots\,,\ \boldsymbol{X}\,\boldsymbol{c_q}\quad (q \leqslant p)\,,$$

les vecteurs $\boldsymbol{c_k}$ $(k = 1, \ldots, q)$ étant normés à l'unité et les différentes combinaisons linéaires étant non corrélées :

$$(\boldsymbol{X}\,\boldsymbol{c}_i)'\,(\boldsymbol{X}\,\boldsymbol{c}_j) = 0\quad (i \neq j)\,.$$

On calcule ensuite une régression multiple en considérant que ces diverses combinaisons linéaires sont les variables explicatives.

Dans le cas de la régression classique au sens des moindres carrés, la combinaison linéaire est déterminée de manière à rendre maximum le carré du coefficient de corrélation de $\boldsymbol{y}$ et $\boldsymbol{X}\,\boldsymbol{c}_1$:

$$r^2 = \frac{\left((\boldsymbol{X}\,\boldsymbol{c}_1)'\,\boldsymbol{y}\right)^2}{\boldsymbol{y}'\,\boldsymbol{y}\,(\boldsymbol{X}\,\boldsymbol{c}_1)'\,(\boldsymbol{X}\,\boldsymbol{c}_1)} = \frac{(\boldsymbol{c}_1'\,\boldsymbol{X}'\,\boldsymbol{y})^2}{\boldsymbol{y}'\,\boldsymbol{y}\,\boldsymbol{c}_1'\,(\boldsymbol{X}'\,\boldsymbol{X})\,\boldsymbol{c}_1}\,,$$

sous la contrainte $\boldsymbol{c}_1'\,\boldsymbol{c}_1 = 1$. La solution est :

$$\boldsymbol{c}_1 = \widehat{\boldsymbol{\beta}}/(\widehat{\boldsymbol{\beta}}'\,\widehat{\boldsymbol{\beta}})^{1/2}\,,$$

$\widehat{\beta}$ étant le vecteur des coefficients de la régression classique. On peut par ailleurs montrer que, pour tout autre combinaison linéaire $X c_2$, telle que :

$$(X c_2)' (X c_1) = c_2 (X' X) c_1 = 0 ,$$

le coefficient de corrélation avec y est nul, ce qui justifie le fait qu'on se limite à une seule combinaison linéaire. Quant au coefficient de régression, $\widehat{\alpha}_1$, de y en fonction de $X c_1$, il vaut $(\widehat{\beta}' \widehat{\beta})^{1/2}$.

Dans le cas de la régression en fonction des composantes principales, on a, selon les notations du paragraphe 3 :

$$X c_k = z_k \quad \text{et} \quad c_k = u_k \quad (k = 1, \ldots, q) .$$

Le vecteur c_k est calculé de manière à assurer le maximum de la somme des carrés des éléments du vecteur $X c_k$, c'est-à-dire aussi le maximum de :

$$(X c_k)' (X c_k) = c_k' (X' X) c_k ,$$

sous les contraintes que c_k soit de norme unitaire et que la combinaison linéaire $X c_k$ soit non corrélée aux combinaisons linéaires précédentes.

Enfin, dans le cas de la régression par les moindres carrés partiels, on a :

$$X c_k = t_k \quad et \quad c_k = v_k \quad (k = 1, \ldots, q) .$$

Le vecteur c_k est alors calculé de manière à maximiser le carré de la covariance de $X c_k$ et de y :

$$\text{cov} (X c_k , y) = c_k' X' y ,$$

sous les mêmes contraintes que celles énoncées ci-dessus.

6. Régression pseudo-orthogonale

Nous avons signalé, au paragraphe 3, que le carré moyen de l'erreur de $\widehat{\beta}$ est, dans le cas des moindres carrés ordinaires, proportionnel à la somme des inverses des valeurs propres de $X' X$.

Il en résulte qu'en présence du phénomène de colinéarité approximative, les estimateurs des moindres carrés sont très instables. Ainsi, dans certains cas, de faibles modifications dans les données peuvent notamment conduire au changement du signe de l'un ou l'autre coefficient. De plus, la norme de $\widehat{\beta}$ est trop grande, c'est-à-dire que la somme des carrés des éléments de $\widehat{\beta}$ peut être nettement supérieure à la somme des carrés des éléments de β.

La régression pseudo-orthogonale[6] a été proposée par Hoerl et Kennard (1970a, 1970b), afin d'atténuer ces inconvénients. Par cette méthode, l'estimation du vecteur β est obtenue en ajoutant aux éléments de la diagonale de $\boldsymbol{X}'\boldsymbol{X}$ une quantité non négative, $\boldsymbol{k}$. On a alors :

$$\widehat{\boldsymbol{\beta}}_{r(k)} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{k}\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (k \geqslant 0)\,.$$

La solution ainsi obtenue est d'autant plus proche de la solution des moindres carrés que k est petit; en particulier, si $k = 0$, on retrouve la solution des moindres carrés.

Les propriétés de l'estimateur $\widehat{\beta}_{r(k)}$ ont été étudiées par Hoerl et Kennard (1970a) et sont reprises par de nombreux auteurs, notamment par Marquart (1970). Nous nous limiterons à l'énoncé des propriétés les plus importantes pour la compréhension de la technique. On peut tout d'abord montrer que $\widehat{\beta}_{r(k)}$ est une transformation linéaire de $\widehat{\beta}$, qui ne dépend que de $\boldsymbol{X}$ et de k :

$$\widehat{\boldsymbol{\beta}}_{r(k)} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}_{(k)}\,\widehat{\boldsymbol{\beta}}\,.$$

Il en résulte automatiquement que $\widehat{\beta}_{r(k)}$ est un estimateur biaisé de β, puisque $\widehat{\beta}$ est un estimateur non biaisé.

Le carré moyen de l'erreur de $\widehat{\beta}_{r(k)}$ est donné par la relation :

$$\mathrm{E}[(\widehat{\boldsymbol{\beta}}_{r(k)} - \boldsymbol{\beta})'\,(\widehat{\boldsymbol{\beta}}_{r(k)} - \boldsymbol{\beta})] = \mathrm{tr}\,[\mathbf{V}(\widehat{\boldsymbol{\beta}}_{r(k)}) + \boldsymbol{\beta}'(\boldsymbol{Q}_{(k)} - \boldsymbol{I})'(\boldsymbol{Q}_{(k)} - \boldsymbol{I})\,\boldsymbol{\beta}]\,,$$

$\mathbf{V}(\widehat{\beta}_{r(k)})$ étant la matrice de variances et covariances de $\widehat{\beta}_{r(k)}$:

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}_{r(k)}) = \sigma^2\,\boldsymbol{Q}_{(k)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{Q}'_{(k)}\,.$$

Cette espérance mathématique est constituée de deux termes. Le premier terme est lié à la variance de $\widehat{\beta}_{r(k)}$ et est une fonction décroissante de k : la variance diminue lorsque k augmente. Le second terme correspond au carré du biais; il est nul si $k = 0$ et est une fonction croissante de k. La réduction de la variance se fait donc au détriment du biais.

On peut, par ailleurs, affirmer qu'il existe toujours une valeur de k telle que le carré moyen de l'erreur de $\widehat{\beta}_{r(k)}$ est inférieur au carré moyen de l'erreur de $\widehat{\beta}$. Cette valeur de k est cependant fonction de σ^2 et de β, qui sont inconnus. Différentes méthodes sont proposées pour déterminer la valeur de k qui rend minimum l'espérance mathématique du carré moyen de l'erreur de $\widehat{\beta}_{r(k)}$ (Hoerl et Kennard, 1988). A titre

6 En anglais : *ridge regression.*

d'exemple, parmi les solutions simples, on peut citer les deux méthodes suivantes (Hoerl *et al.*, 1975; Lawless et Wang, 1976) :

$$k_1 = p\sigma^2/(\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}) \quad \text{et} \quad k_2 = p\widehat{\sigma}^2/(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 1/F_{obs}\,,$$

F_{obs} étant 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.

On peut aussi montrer que la norme du vecteur $\widehat{\boldsymbol{\beta}}_{r(k)}$ est inférieure à la norme du vecteur $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}}'_{r(k)}\widehat{\boldsymbol{\beta}}_{r(k)} < \widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} \quad (k > 0)\,.$$

En pratique, cela signifie que les coefficients de l'équation de régression pseudo-orthogonale sont, dans l'ensemble, plus petits, en valeur absolue, que les coefficients obtenus par la méthode des moindres carrés ordinaires. En fait, la norme est une fonction décroissante de k, qui tend vers zéro lorsque k tend vers l'infini. Cette réduction de la norme du vecteur rattache la régression pseudo-orthogonale aux méthodes «rétrécissantes[7]», dont il sera également question au paragraphe 7.

En relation avec cette norme, on peut considérer aussi que la régression pseudo-orthogonale est une forme de régression sous contrainte (Cazes, 1975, 1978). En effet, le vecteur $\widehat{\boldsymbol{\beta}}_{r(k)}$ est le vecteur qui assure le minimum de la somme des carrés des écarts entre les valeurs observées et les valeurs estimées de $\boldsymbol{y}$:

$$\mathrm{SCE}_{r(k)} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{r(k)})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{r(k)})\,,$$

sous la contrainte que la norme de $\widehat{\boldsymbol{\beta}}_{r(k)}$ soit inférieure ou égale à une valeur fixée, d, qui est liée à k par la relation (Tomassone *et al.*, 1983) :

$$\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{X}'\boldsymbol{y} = d^2\,,$$

si $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}$ est supérieur à d^2, sinon $k = 0$.

Une autre façon d'exprimer la même réalité consiste à se fixer une valeur pour SCE_r. Si cette valeur n'est pas égale au minimum, on sait qu'une infinité de vecteurs $\widehat{\boldsymbol{\beta}}$ peuvent conduire à cette somme de carrés. Et parmi cette infinité de solutions, $\widehat{\boldsymbol{\beta}}_{r(k)}$ est la solution de norme minimum.

Le tableau 9 donne l'évolution en fonction de k des coefficients de régression, de la norme du vecteur $\widehat{\boldsymbol{\beta}}_{r(k)}$ et de la somme des carrés des écarts résiduelle.

[7] En anglais : *shrinkage methods.*

TABLEAU 9

Evolution en fonction de k des coefficients de régression, de la norme du vecteur $\widehat{\beta}_{r(k)}$ et de la somme des carrés des écarts résiduelle.

k	0	0,2	0,4	0,6	0,8
x_1	−0,4234	−0,3434	−0,2980	−0,2657	−0,2409
x_2	0,2742	0,2331	0,2106	0,1949	0,1824
x_3	−0,1871	−0,2372	−0,2388	−0,2315	−0,2220
x_4	0,2871	0,2468	0,2302	0,2171	0,2057
$\widehat{\beta}'_{r(k)}\widehat{\beta}_{r(k)}$	0,3719	0,2894	0,2431	0,2094	0,1829
$\mathrm{SCE}_{r(k)}$	0,2052	0,2122	0,2258	0,2426	0,2610

L'application des formules données ci-dessus pour le choix de k conduit aux valeurs suivantes :

$$k_1 = (4)(0,03420)/0,3719 = 0,368 \quad \text{et} \quad k_2 = 1/5,81 = 0,172\,.$$

Par rapport aux moindres carrés ordinaires, l'augmentation de la somme de carrés des écarts résiduelle serait de l'ordre de 3 % pour k_2 et de l'ordre de 10 % pour k_1, mais la norme du vecteur serait réduite respectivement de l'ordre de 20 % et de 35 %.

Hoerl et Kennard (1970a, 1970b) ont proposé l'établissement de graphiques donnant l'évolution des différents coefficients de régression en fonction de k[8] Sur la base de ces graphiques, ils suggèrent, d'une part, de choisir la valeur de k et, d'autre part, d'éliminer éventuellement certaines variables explicatives. Ils suggèrent de retenir la valeur de k la plus faible à partir de laquelle l'évolution des coefficients en fonction de k est stabilisée et ils proposent de ramener à zéro les coefficients de régression qui, pour la valeur de k retenue, sont proches de zéro, ce qui revient à éliminer ces variables.

Pour l'exemple traité, la figure 1 montre que les variations des coefficients sont relativement faibles, car le phénomène de colinéarité, bien que présent, n'est pas particulièrement accentué. Des réductions bien plus importantes peuvent s'observer en pratique, surtout lorsqu'on dispose d'un plus grand nombre de variables explicatives comme le montrent, par exemple, les problèmes traités par Hoerl et Kennard (1970b) ou Tomassone *et al.* (1983). La figure suggère également qu'une valeur faible de k, par exemple 0,2, semble adéquate, et qu'aucune variable ne peut être éliminée *a priori*.

Enfin, signalons encore qu'une extension naturelle de la régression pseudo-orthogonale peut être envisagée en remplaçant la matrice scalaire $k\,\boldsymbol{I}$, qui est rajoutée à $\boldsymbol{X}'\,\boldsymbol{X}$, par la matrice $\boldsymbol{U}\,\boldsymbol{K}\,\boldsymbol{U}'$, où $\boldsymbol{K}$ est une matrice diagonale et $\boldsymbol{U}$ la matrice des vecteurs propres normés à l'unité de $\boldsymbol{X}'\,\boldsymbol{X}$ (Hoerl et Kennard, 1970a).

8 En anglais : *ridge trace.*

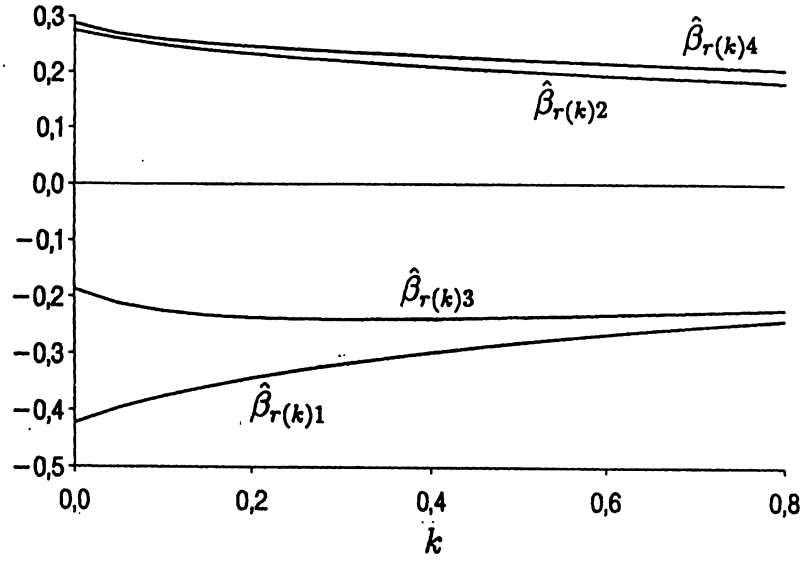

FIGURE 1

Evolution des coefficients de régression, $\widehat{\beta}_{r(k)i}$, en fonction de k.

7. Estimateurs de James et Stein

L'application à la régression des estimateurs proposés, dans un contexte plus général, par James et Stein (1961) conduit, comme la régression pseudo-orthogonale, à la notion d'estimateurs «rétrécis». L'objectif est de trouver un compromis entre le biais et la variance afin de minimiser l'espérance mathématique du carré moyen de l'erreur sur le vecteur des coefficients de régression.

Indépendamment de tout problème de régression, ces auteurs se sont intéressés à l'estimateur d'un vecteur γ de p paramètres, alors qu'on dispose de p estimateurs non biaisés indépendants, $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$, dont les distributions d'échantillonnage sont normales et de même variance $\sigma^2_{\widehat{\gamma}}$. Ils ont montré que la meilleure solution, pour $p \geqslant 3$, ne consiste pas à prendre le vecteur de ces p estimateurs :

$$\widehat{\gamma} = [\widehat{\gamma}_1 \quad \widehat{\gamma}_2 \quad \ldots \quad \widehat{\gamma}_p]' ,$$

mais qu'il y a lieu de multiplier ce vecteur par une constante c, inférieure à l'unité :

$$\widehat{\gamma}_{JS} = c\widehat{\gamma},$$

si, du moins, on souhaite minimiser le carré moyen de l'erreur de $\widehat{\gamma}$:

$$\mathrm{E}[(\widehat{\gamma}_{JS} - \gamma)'(\widehat{\gamma}_{JS} - \gamma)] = \mathrm{E}\left(\sum_{j=1}^{p}(\widehat{\gamma}_{JS_i} - \gamma_i)^2\right).$$

Cette espérance mathématique peut encore s'écrire :

$$\mathrm{E}\left(\sum_{j=1}^{p}(\widehat{\gamma}_{JS_i} - \gamma_i)^2\right) = \sum_{j=1}^{p}\mathrm{E}[\widehat{\gamma}_{JS_i} - \mathrm{E}(\widehat{\gamma}_{JS_i})]^2 + \sum_{j=1}^{p}[\mathrm{E}(\widehat{\gamma}_{JS_i}) - \gamma_i]^2$$

$$= p\,c^2\,\sigma_\gamma^2 + (1-c)^2 \sum_{j=1}^{p}\gamma_i^2$$

le premier terme représentant la somme des variances des estimateurs et le second terme correspondant à la somme des carrés des biais.

La valeur de c qui minimise l'espérance mathématique s'obtient en annulant la dérivée de l'expression. On trouve :

$$c = \frac{\sum_{j=1}^{p}\gamma_i^2}{p\,\sigma_\gamma^2 + \sum_{j=1}^{p}\gamma_i^2} = 1 - \frac{p\,\sigma_\gamma^2}{p\,\sigma_\gamma^2 + \sum_{j=1}^{p}\gamma_i^2}.$$

Par ailleurs, on sait que $\widehat{\gamma}_i^2$ est une estimation non biaisée de $\sigma_\gamma^2 + \gamma_i^2$ et on peut, dès lors, remplacer le dénominateur de l'expression donnée ci-dessus par $\Sigma\,\widehat{\gamma}_i^2$, ce qui permet d'éliminer les paramètres inconnus γ_i. Cette substitution fait cependant en sorte que c devient une variable aléatoire et la prise en compte de la variance de c conduit à l'estimateur de James et Stein (1961) :

$$\gamma_{\widehat{JS}_i} = \left(1 - (p-2)\,\sigma_\gamma^2 \Big/ \sum_{j=1}^{p}\widehat{\gamma}_i^2\right)\widehat{\gamma}_i\,,$$

ou encore :

$$\widehat{\gamma}_{JS} = \left(1 - c_0\,\frac{\sigma_\gamma^2}{\widehat{\gamma}'\widehat{\gamma}}\right)\widehat{\gamma},$$

avec :

$$c_0 = p - 2\,.$$

Pour $p \geqslant 3$, la multiplication des estimateurs non biaisés par la constante c introduit donc un biais qui est plus que compensé par la réduction de la variance. James et Stein (1961) ont également montré qu'une amélioration pouvait encore être apportée

à l'estimateur en donnant à c la valeur zéro lorsque le calcul de c par la formule donnée ci-dessus conduit à une valeur négative.

De façon plus générale, l'espérance mathématique du carré de l'erreur relative à l'estimateur de James et Stein est inférieure à l'espérance mathématique du carré de l'erreur de l'estimateur non biaisé lorsque c_0 est compris entre 0 et $2\,(p-2)$.

Les résultats de James et Stein [1961] peuvent être appliqués à l'estimation d'un vecteur de coefficients de régression. Du point de vue théorique, on aborde le problème dans le cas de la régression en fonction des composantes principales, $\boldsymbol{Z}_0$, standardisées de manière telle que $\boldsymbol{Z}'_0\,\boldsymbol{Z}_0 = \boldsymbol{I}$, afin d'assurer la non-corrélation et l'égalité des variances des coefficients de régression.

Dans ces conditions, l'estimateur de James et Stein s'écrit :

$$\widehat{\boldsymbol{\alpha}}_{JS} = \left(1 - \frac{(p-2)\,\sigma^2}{\widehat{\boldsymbol{\alpha}}'_0\,\widehat{\boldsymbol{\alpha}}_0}\right)\widehat{\boldsymbol{\alpha}}_0\,,$$

$\widehat{\boldsymbol{\alpha}}_0$ étant le vecteur de régression en fonction des composantes principales standardisées et σ^2 étant la variance résiduelle. Cette dernière peut être remplacée par l'estimation habituelle $\widehat{\sigma}^2$, à k degrés de liberté, ce qui entraîne encore une légère modification de la relation :

$$\widehat{\boldsymbol{\alpha}}_{JS} = \left(1 - \frac{k\,(p-2)\,\widehat{\sigma}^2}{(k+2)\,\widehat{\boldsymbol{\alpha}}'_0\,\widehat{\boldsymbol{\alpha}}_0}\right)\widehat{\boldsymbol{\alpha}}_0\,.$$

Si la quantité figurant entre parenthèses est négative, on la remplace cependant par zéro.

Le retour aux coefficients des variables initiales se fait alors sans difficultés, en utilisant une transformation analogue à celle donnée au paragraphe 3 :

$$\widehat{\boldsymbol{\beta}}_{JS} = \boldsymbol{W}\,\widehat{\boldsymbol{\alpha}}_{JS}\,,$$

avec :
$$\boldsymbol{W} = \boldsymbol{U}\,\boldsymbol{\Lambda}^{-1/2}\,.$$

Dans ces relations, $\widehat{\boldsymbol{\alpha}}_0$ est le vecteur des coefficients de régression relatif aux composantes principales de variance unitaire, $\boldsymbol{U}$ est la matrice des vecteurs propres normés à l'unité de la matrice $\boldsymbol{X}'\,\boldsymbol{X}$ et $\boldsymbol{\Lambda}^{-1/2}$ est la matrice diagonale dont les éléments sont égaux à $1/\sqrt{\lambda_i}$.

En pratique, la correction des coefficients de régression peut être effectuée directement sur le vecteur $\widehat{\boldsymbol{\beta}}$, sans passer par les composantes principales, car :

$$\widehat{\boldsymbol{\beta}}_{JS} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\,\widehat{\boldsymbol{\alpha}}_{JS} = c\,\boldsymbol{U}\,\boldsymbol{\Lambda}^{-1/2}\,\widehat{\boldsymbol{\alpha}}_0 = c\,\widehat{\boldsymbol{\beta}}\,.$$

Du fait que, pour la régression au sens des moindres carrés ordinaires, la somme des carrés liée à la régression est égale à :

$$\widehat{\boldsymbol{y}}'\,\widehat{\boldsymbol{y}} = (\boldsymbol{Z}_0\,\widehat{\boldsymbol{\alpha}}_0)'\,(\boldsymbol{Z}_0\,\widehat{\boldsymbol{\alpha}}_0) = \widehat{\boldsymbol{\alpha}}'_0\,\boldsymbol{Z}'_0\,\boldsymbol{Z}_0\,\widehat{\boldsymbol{\alpha}}_0 = \widehat{\boldsymbol{\alpha}}'_0\,\widehat{\boldsymbol{\alpha}}_0\,,$$

la constante c peut encore s'écrire :

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

cette dernière approximation n'étant satisfaisante que si le nombre de degrés de liberté de la variance résiduelle, k, est suffisamment grand.

F_{obs} est la valeur observée de la variable de Snedecor relative au test de signification classique des coefficients de régression. On constate donc que la constante c est d'autant plus faible que F_{obs} est petit. D'autre part, dans la mesure où on fixe c égal à zéro si F_{obs} est plus petit que $(p-2)/p$, on peut dire que l'estimateur de James et Stein inclut implicitement une forme de test de signification de l'ensemble des coefficients de régression, ceux-ci étant automatiquement fixés à zéro si la valeur F_{obs} est trop faible. L'utilisation de l'estimateur de James et Stein en relation avec le test de signification d'un sous-ensemble de coefficients de régression est évoquée par Sclove (1968).

On notera que le passage par les composantes principales a comme conséquence qu'on ne minimise pas l'espérance mathématique de $(\widehat{\boldsymbol{\beta}}_{JS} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}}_{JS} - \boldsymbol{\beta})$ mais bien de :

$$(\widehat{\boldsymbol{\alpha}}_{JS} - \boldsymbol{\alpha}_0)'(\widehat{\boldsymbol{\alpha}}_{JS} - \boldsymbol{\alpha}_0) = (\widehat{\boldsymbol{\beta}}_{JS} - \boldsymbol{\beta})'\mathbf{X}'\,\mathbf{X}(\widehat{\boldsymbol{\beta}}_{JS} - \boldsymbol{\beta}),$$

ce qui, comme le montre Copas (1983), revient à minimiser le carré moyen de l'erreur de prédiction :

$$\mathrm{E}(y - \widehat{y}_{JS})^2.$$

On a donc implicitement modifié le critère initial. Il faut cependant remarquer que la minimisation de l'erreur de prédiction se justifie entièrement si le but du calcul de l'équation est d'établir un modèle permettant de réaliser des prédictions de y. L'utilisation des estimateurs de James et Stein ne se justifie par contre pas si l'objectif est l'interprétation d'un seul coefficient de régression, celui-ci pouvant être estimé bien plus mal par la méthode de James et Stein que par la méthode des moindres carrés ordinaires.

Une autre justification de l'utilisation des estimateurs de James et Stein dans le cadre de la régression multiple a également été donnée par Copas (1983). L'argument repose sur la relation existant entre les observations y_i et les valeurs estimées par les moindres carrés ordinaires $\widehat{y}_i$ pour des individus autres que ceux utilisés lors de l'estimation de l'équation de régression. Sous l'hypothèse que les individus supplémentaires proviennent de la même population que les individus de l'échantillon utilisé pour la construction du modèle, Copas (1983) a montré que la pente de la droite exprimant y en fonction $\widehat{y}$ est, en espérance mathématique, inférieure à l'unité, ce qui signifie que, en moyenne, il y a surestimation des valeurs élevées de y et sous-estimation des valeurs faibles de y. Les prédictions ont donc tendance à être trop extrêmes, et on doit s'attendre à une réduction de l'espérance mathématique du carré

des écarts entre y_i et $\widehat{y}_i$, en utilisant une équation de prédiction du type :

$$\tilde{y} = c\,\widehat{y}$$

où la constante c est inférieure à l'unité. Cette constante c est précisément le coefficient intervenant dans l'estimateur de James et Stein.

Pour l'exemple considéré, la valeur F_{obs} obtenue par le rapport du carré moyen associé à la régression en fonction des quatre variables au carré moyen résiduel est égale à 5,81. La valeur à attribuer à la constante c est donc égale à :

$$c = 1 - \frac{(4-2)(6)}{(6+2)(4)(5,81)} = 0,93546 \text{ ou } 0,94\,.$$

Les coefficients de régression obtenus par les moindres carrés ordinaires (paragraphe 2) doivent donc être réduits de 6 % :

$$\widehat{\beta}'_{JS} = c\,\widehat{\beta}' = [-0,3961 \;\; 0,2565 \;\; 0,1750 \;\; 0,2686]\,.$$

La variance résiduelle associée à cette dernière équation est égale à 0,03475, soit une augmentation de l'ordre de 1,5 % par rapport à la variance résiduelle de la régression au sens des moindres carrés ordinaires.

8. Discussion

Dans les paragraphes précédents, nous avons présenté quelques alternatives à l'estimation classique des coefficients de régression par la méthode des moindres carrés. Celles-ci visent à atténuer les inconvénients de la multicolinéarité importante qui se manifeste dans beaucoup de problèmes de régression multiple. On notera cependant que toutes ces méthodes n'éliminent pas entièrement le problème du choix des variables explicatives, puisque la méthode de Webster *et al.*, la régression pseudo-orthogonale et la méthode basée sur les estimateurs de James et Stein proposent également la suppression de variables redondantes, selon des critères qui peuvent être différents de ceux traditionnellement utilisés en régression multiple classique.

Nous avons déjà fait allusion à des situations pratiques justifiant le recours à la régression en fonction des composantes principales. Il s'agit des cas où on sait que les variables explicatives contiennent une composante aléatoire en plus de l'information spécifique. L'utilisation systématique des composantes principales à la place des variables explicatives initiales, sous prétexte que ces dernières sont corrélées, ne se justifie pas d'un point de vue pratique, dans la mesure où l'équation de régression contient toutes les variables explicatives initiales, même si certaines composantes ont été négligées. L'interprétation de l'équation est alors pratiquement impossible et l'utilisation de celle-ci pour la réalisation d'estimations peut s'avérer très lourde, notamment si les variables sont difficiles à obtenir.

La méthode des moindres carrés partiels, utilisée presque exclusivement en chimiométrie, serait, d'après ses promoteurs, une alternative intéressante et son

application pourrait être envisagée dans un contexte tout à fait analogue à la régression en fonction des composantes principales.

La méthode de Webster *et al.*, est considérée par Draper et Smith (1981) comme relativement arbitraire. Ces derniers pensent qu'elle est susceptible de rendre des services à un utilisateur expérimenté qui en ferait constamment l'usage, mais ils estiment que la méthode ne peut pas être recommandée pour la majorité des utilisateurs.

Dans une synthèse bibliographique consacrée à la régression par les estimateurs de James et Stein et à la régression pseudo-orthogonale, Draper et Van Nostrand (1979) concluent que 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 à attendre est loin d'être manifeste. Les auteurs, de même que Draper et Smith (1981), émettent également des réserves à l'égard de nombreuses études de simulations concluant à la supériorité de la régression pseudo-orthogonale. Ils constatent, en effet, que ces études ont souvent été réalisées en imposant des restrictions aux valeurs réelles des paramètres, ce qui correspond précisément à la situation pour laquelle la régression pseudo-orthogonale est, du point de vue théorique, la méthode adéquate. Par conséquent, ils suggèrent de limiter l'utilisation de la régression pseudo-orthogonale aux cas où l'utilisateur est convaincu que les valeurs absolues des coefficients de régression ne sont vraisemblablement pas grandes et au cas où il souhaite effectivement imposer une contrainte sur la longueur du vecteur β.

Il faut, par ailleurs, également rappeler que l'exemple traité ne constitue qu'un support pratique, choisi en raison de ses dimensions réduites et destiné à illustrer les principes exposés. Il ne présente pas de phénomène de colinéarité trop accentué et il est donc normal que les méthodes alternatives ne soient pas supérieures à la régression usuelle. Nous avons cependant fait allusion, au paragraphe 6, à des exemples traités dans la littérature où la colinéarité est plus marquée et où les auteurs montrent que les solutions obtenues par la régression pseudo-orthogonale sont préférables à celles fournies par la régression usuelle. Les résultats de plusieurs exemples concrets sont également discutés par Hocking (1976).

Enfin, on notera aussi que les méthodes qui ont été décrites ne sont pas les seules solutions alternatives possibles. Ainsi, nous avons signalé que la régression pseudo-orthogonale est une forme de régression sous contrainte, la contrainte portant sur la norme du vecteur $\widehat{\beta}$ (paragraphe 6). D'autres formes de contraintes peuvent aussi être envisagées : contraintes de posivité des coefficients de régression, ou, d'une manière plus générale, contraintes imposant aux coefficients d'être compris entre deux limites fixées, contraintes imposant aux valeurs estimées de la variable à expliquer d'être positives ou comprises entre deux limites, etc. Des informations relatives à ces méthodes et des exemples d'applications sont donnés par Cazes (1975, 1977, 1978) et par Cazes et Turpin (1971). Cette approche de la régression revient à tenir compte des informations *a priori* dont on dispose pour effectuer la régression et se justifie certainement lorsque ces informations sont disponibles.

9. Bibliographie

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.

COPAS J.B. (1983). Regression, prediction and shrinkage. *J. R. Stat. Soc., Ser. B*, 45, p. 311-354.

DAGNELIE P. (1982). *Analyse statistique à plusieurs variables*. Gembloux, Presses agronomiques, 362 p.

DRAPER N.R., SMITH H. (1981). *Applied regression analysis*. New York, Wiley, 709 p.

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

GRAYBILL F.A. (1969). *Introduction to matrices with applications in statistics*. Belmont, Wadsworth, 372 p.

GREENBERG E. (1975). Minimum variance properties of principal component regression. *J. Amer. Stat. Assoc.* 70, p. 194-197.

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

HOERL A.E., KENNARD R.W. (1970a). Ridge regression : biased estimation for nonorthogonal problems. *Technometrics* 12, p. 55-66.

HOERL A.E., KENNARD R.W. (1970b). Ridge regression : applications to non-orthogonal problems. *Technometrics* 12, p. 69-82.

HOERL A.E., KENNARD R.W. (1988). Ridge regression. *In :* Kotz S., Johnston N.L. (eds). *Encyclopedia of statistical sciences* (vol. 8). New York, Wiley, p. 129-136.

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

JACKSON J.E. (1991). *A user's guide to principal components*. New York, Wiley, 570 p.

JOLLIFFE I.T. (1982). A note on the use of principal components in regression. *Appl. Stat.* 31, p. 300-303.

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

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.

MANSFIELD E.R., WEBSTER J.T., GUNST R.F. (1977). An analytical variable selection technique for principal component regression. *Appl. Stat.* 26, p. 34-40.

MARQUARDT D.W. (1970). Generalized inverses, ridge regression biased linear estimation and non linear estimation. *Technometrics* 12, p. 591-612.

MARTENS H., NAES T. (1989). *Multivariate calibration.* New York, Wiley, 419 p.

PALM R. (1988). Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.

PALM R. (1994). Les méthodes d'analyse factorielle : principes et applications. *Biom. Praxim.* 34, p. 35-80.

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

STEWART G.W. (1987). Collinearity and least squares regression. *Stat. Sci.* 2, p. 68-100.

STONE M., BROOKS R.J. (1990). Continuum regression : cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Stat. Soc., Ser. B*, 52, p. 237-269.

THEIL H. (1971). *Principles of econometrics.* New York, Wiley, 736 p.

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.

TOMASSONE R., LESQUOY E., MILLIER C. (1983). *La régression : nouveaux regards sur une ancienne méthode statistique.* Paris, Masson, 180 p.

WEBSTER J.T., GUNST R.F., MASON R.L. (1974). Latent root regression analysis. *Technometrics* 16, p. 513-522.

WEISBERG S. (1985). *Applied linear regression.* New York, Wiley, 324 p.