RAIRO. RECHERCHE OPÉRATIONNELLE

THIERRY FOUCART

Colinéarité et instabilité numérique dans le modèle linéaire

RAIRO. Recherche opérationnelle, tome 34, n° 2 (2000), p. 199-212

<http://www.numdam.org/item?id=RO_2000__34_2_199_0>

RAIRO Operations Research
RAIRO Oper. Res. **34** (2000) 199-212

COLINÉARITÉ ET INSTABILITÉ NUMÉRIQUE DANS LE MODÈLE LINÉAIRE (*)

par Thierry FOUCART ([1])

Communiqué par I.C. LERMAN

Résumé. – *Nous donnons dans cet article l'expression du coefficient de corrélation multiple dans un modèle linéaire en fonction des coefficients de corrélation. Cette expression permet d'analyser d'un point de vue numérique l'instabilité des estimations dans le cas de variables explicatives quasi colinéaires que l'on rencontre fréquemment dans le modèle linéaire et le modèle autorégressif. L'approche numérique, que nous montrons sur deux exemples, complète ainsi l'approche habituelle de la quasi colinéarité, fondée sur les propriétés statistiques des estimateurs.*

Mots clés : Modèle linéaire multiple, quasi colinéarité, coefficients de corrélation multiple, matrice symétrique définie positive, factorisation de Cholesky.

Abstract. – *In this paper we give the expression of the multiple correlation coefficient in a linear model according to the coefficients of correlation. This expression makes it possible to analyze from a numerical point of view the instability of estimates in the case of collinear explanatory variables in the linear model or in the autoregressive model. This numerical approach, that we show on two examples, thus supplements the usual approach of the quasi colinearity, founded on the statistical properties of the estimators.*

Keywords: Linear multiple model, collinearity, multiple correlation coefficients, symmetric definite positive matrix, Cholesky's factorization.

INTRODUCTION

Nous établissons dans cet article la relation entre le coefficient de corrélation multiple et les coefficients de corrélation, généralisant par là les travaux concernant la relation de transitivité entre deux coefficients de corrélation d'une même matrice déjà publiés (Foucart, 1991).

Cette relation permet de préciser les effets d'une faible variation de ces coefficients dans deux situations statistiques fréquemment rencontrées qui

(*) Reçu en octobre 1997.

([1]) Département Mathématiques, S P 2 M I, boulevard 3, Téléport 2, BP. 179, 86960 Futuroscope Cedex, France.

Recherche opérationnelle/Operations Research, 0399-0559/00/02/$ 7.00

sont l'estimation des paramètres d'un modèle de régression linéaire multiple et celle des paramètres d'un modèle autorégressif.

En effet, considérons le modèle:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

dans lequel:

- $\mathbf{Y}$ est le vecteur $(y_1, y_2, \ldots, y_n)^t$ des n valeurs observées de la variable expliquée Y;
- $\mathbf{X}$ est la matrice des données à n lignes et p colonnes, la colonne j (de 1 à p) étant définie par le vecteur $(x_1^j, x_2^j, \ldots, x_n^j)^t$;
- les variables sont centrées et réduites:

$$\sum_{i=1}^{n} y_i = 0 \quad \sum_{i=1}^{n} y_1^2 = n \qquad \forall\, j = 1, \ldots, p \quad \sum_{i=1}^{n} x_i^j = 0 \quad \sum_{i=1}^{n} x_i^{j2} = n$$

- $\beta = (\beta_1, \beta_1, \ldots, \beta_p)^t$ est le vecteur des coefficients de régression;
- $\mathbf{r}$ est le vecteur des coefficients de corrélation observés $(r_1, r_2, \ldots, r_p)^t$ entre les variables explicatives X_j et la variable expliquée Y.
- ε est le vecteur résiduel $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$.

On note $\mathbf{R}$ la matrice de corrélation entre les p variables. On la suppose de rang p. La méthode classique d'estimation des paramètres est fondée sur le critère des moindres carrés. L'estimateur $\mathbf{B} = (b_1, b_2, \ldots, b_p)^t$ de β est alors donné par la formule ci-dessous:

$$\mathbf{B} = (1/n)\mathbf{R}^{-1}\mathbf{X}^t\mathbf{Y} = \mathbf{R}^{-1}\mathbf{r}.$$

Le coefficient de détermination est par définition le carré du coefficient de corrélation linéaire de la variable expliquée $\mathbf{Y}$ et de $\mathbf{B}^t\mathbf{X}$. Il est égal à:

$$R^2 = \mathbf{r}^t\mathbf{R}^{-1}\mathbf{r}.$$

L'instabilité numérique dans l'estimation de β va dépendre de la colinéarité éventuelle entre les variables explicatives. Les relations que nous établirons entre les coefficients de corrélation et les coefficients de détermination vont nous permettre d'expliquer de façon plus approfondie les instabilités numériques très souvent rencontrées.

De façon similaire considérons un modèle autorégressif stationnaire $AR(p)$:

$$X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \ldots + \theta_{pt} X_{t-p} + \varepsilon_t$$

où $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^t$ est le vecteur d'autorégression et ε_t un processus indépendant identiquement distribué (i.i.d.). On note ρ_i le coefficient de corrélation entre X_t et X_{t-i} et $\gamma_(i)$ leur covariance.

Il est bien connu que le processus $\{X_t, t \in \mathbf{Z}\}$ est stationnaire si le polynôme

$$P(z) = 1 - \theta_1 z - \theta_2 z^2 - \ldots - \theta_p z^p$$

a toutes ses racines à l'extérieur du cercle unité. On sait alors que l'estimateur du vecteur θ est fondé sur les observations $X_1, X_2, \ldots, X_n$ et déterminé par l'équation de Yule Walker (Brockwell et Davis, 1998):

$$\theta = \mathbf{R}^{-1}\rho$$

où ρ est le vecteur des coefficients de corrélation $(\rho_1, \rho_2, \ldots, \rho_p)^t$ et $\mathbf{R}$ la matrice des covariances $\gamma(ij)$, $1 \leq i \leq p$, $1 \leq j \leq p$.

Comme dans le modèle précédent, nous sommes amenés à étudier les coefficients d'autocorrélation et les coefficients d'autocorrélation partielle. Avec les résultats établis ci-dessous, nous pourrons également interpréter les instabilités numériques.

1. INTERPRÉTATION STATISTIQUE DES TERMES D'UNE MATRICE SYMÉTRIQUE DÉFINIE POSITIVE

Étant donnée une matrice de corrélation $\mathbf{R}$ entre p variables $X_1, \ldots, X_p$, on peut déterminer une matrice triangulaire inférieure $\mathbf{A} = (a_{i,j})$ telle que (Ciarlet, 1989) :

$$\mathbf{R} = \mathbf{A}\mathbf{A}^{\mathbf{t}}.$$

Lorsque la matrice $\mathbf{R}$ est symétrique définie positive, la matrice $\mathbf{A}$ est définie de la façon suivante:

$$\begin{aligned} &\forall\, i = 1, \ldots, p \qquad a_{i,1} = r_{1,i} \\ &\forall\, i = 2, \ldots, p \qquad a_{i,j} = \left[1 - \sum_{k=1}^{i-1} a_{i,k}^2\right]^{1/2} > 0 \end{aligned} \tag{1}$$

$$\forall\, i = 2, \ldots, p \qquad \forall\, j = i+1, \ldots, p \quad a_{j,i} = \frac{r_{i,j} - \sum_{k=1}^{i-1} a_{i,k} a_{j,k}}{a_{i,i}}. \tag{2}$$

1.1. Propriétés classiques

On note:

- R_i^2 : le coefficient de détermination de la variable X_i dans la régression par les autres variables $X_j, (j = 1, \ldots, p$ et $j \neq i)$.
- $R(X_i, X_j/X_k, k \neq i$ et $k \neq j)$: le coefficient de corrélation partielle entre les variables X_i et X_j conditionnellement à l'ensemble des autres variables.
- $R_{<i}^2$: le coefficient de détermination de la variable X_i dans la régression par les variables $X_j, j = 1, \ldots, i-1$.

Les relations classiques sont les suivantes (Hawkins et Eplett, 1982; Whittaker, 1990 p. 141):

- le terme $a_{i,i}^2$ est égal à $1 - R_{<i}^2$:

$$a_{i,i}^2 = 1 - R_{<i}^2 \tag{3}$$

- le coefficient de détermination dans la régression de X_p par les variables $X_i, i = 1, \ldots, p-1$ est égal à:

$$R_p^2 = \sum_{k=1}^{p-1} a_{p,k}^2 \tag{4}$$

- le coefficient de détermination $R_{p,p-l}^2$ dans la régression de X_p par les variables $X_i, i = 1, \ldots, p-l$ est égal à:

$$R_{p,p-l}^2 = \sum_{k=1}^{p-l} a_{p,k}^2. \tag{5}$$

Les $a_{p,k}$ pour $k \leq p-l$, sont en effet les termes obtenus par la factorisation de la matrice des corrélations entre les variables $X_i, i = 1, \ldots, p-l$ et X_p. On retrouve alors la formule (4).

- le terme diagonal $s_{i,i}$ de la matrice $\mathbf{S} = \mathbf{R}^{-1}$ est égal à $1/(1 - R_i^2)$, où R_i^2 est le coefficient de détermination de la variable X_i dans la régression par les autres variables X_j, $(j = 1, \ldots, p$ et $j \neq i)$;
- le coefficient de corrélation partielle entre les variables X_i et X_j conditionnellement à l'ensemble des autres s'exprime en fonction des termes de la matrice inverse:

$$R(X_i, X_j/X_k, k \neq i \text{ et } k \neq j) = -s_{i,j}/[s_{i,i}s_{j,j}]1/2. \tag{6}$$

1.2. Propriétés supplémentaires

Les matrices de corrélation entre p variables possèdent des propriétés supplémentaires récemment démontrées dans le cadre des matrices symétriques définies positives (Foucart, 1997):

- la matrice $\mathbf{R}$ est symétrique définie positive pour toute valeur du coefficient de corrélation $r_{i,j}$ à l'intérieur d'un intervalle $I_{i,j} =]\alpha_{i,j}, \gamma_{i,j}[$, où $\alpha_{i,j}$ et $\gamma_{i,j}$ sont des fonctions des autres termes de la matrice. On trouvera les expressions de $\alpha_{i,j}$ et $\gamma_{i,j}$ en annexe ;
- le coefficient de corrélation partielle $R(X_i, X_j/X_k, k \neq i \text{ et } k \neq j)$ est la distance du coefficient de corrélation $r_{i,j}$ au centre de l'intervalle $I_{i,j}$ rapportée à la demie longueur de cet intervalle:

$$R(X_i, X_j/X_k,\ k \neq i \text{ et } k \neq j) = [r_{i,j} - (\alpha_{i,j} + \gamma_{i,j})/2]/[(\gamma_{i,j} - \alpha_{i,j})/2].$$

Ce coefficient de corrélation partielle tend vers ± 1 lorsque $r_{i,j}$ tend vers l'une des bornes de son intervalle de variation.

L'intervalle $I_{i,j}$ est l'ensemble des valeurs de $r_{i,j}$ tel que la matrice $\mathbf{R}$ soit symétrique définie positive: on l'appelle *intervalle de variation* du coefficient de corrélation $r_{i,j}$.

L'ensemble $I_{i,j,k,l}$ des valeurs du couple de coefficients de corrélation $(r_{i,j}, r_{k,l})$ telles que la matrice $\mathbf{R}$ soit définie positive définit *l'ensemble de variation* de ce couple. On peut le construire en calculant pour chaque valeur du coefficient de corrélation $r_{i,j}$ variant dans l'intervalle $]-1, 1[$, l'intervalle de variation $]\alpha_{k,l}, \gamma_{k,l}[$ – éventuellement vide – du coefficient $r_{k,l}$.

2. EXPRESSION DU COEFFICIENT DE DÉTERMINATION EN FONCTION DES COEFFICIENTS DE CORRÉLATION

2.1. Expression du coefficient de détermination R^2 en fonction d'un coefficient de corrélation de la forme $r_{k,n}$

L'expression (2) appliquée au couple $(p, p-1)$ donne:

$$a_{p,p-1} = \frac{r_{p-1,p} - \sum_{k=1}^{p-2} a_{p-1,k} a_{p,k}}{a_{p-1,p-1}}.$$

En remplaçant $a_{p,p-1}$ par cette expression dans la formule (4), on aboutit à l'expression de R^2 en fonction de $r_{p-1,p}$ ci-dessous:

$$R^2 = \sum_{k=1}^{p-2} a_{p,k}^2 + \left[\frac{r_{p-1,p} - \sum_{k=1}^{p-2} a_{p-1,k} a_{p,k}}{a_{p-1,p-1}} \right]^2 . \tag{7}$$

Dans l'expression (7), tous les termes de la forme $a_{i,j}$ sont constants par rapport à $r_{p-1,p}$. Nous obtenons ainsi le coefficient de détermination R^2 comme fonction polynomiale de degré 2 de $r_{p-1,p}$, définie sur l'intervalle de variation $]\alpha_{p-1,p}, \gamma_{p-1,p}[$ du coefficient de corrélation $r_{p-1,p}$.

Les propriétés des coefficients de détermination et des coefficients de corrélation partielle nous donnent une interprétation statistique précise des termes du second membre de la formule (7):

- le premier terme est le coefficient de détermination $R^2_{p,p-2}$ obtenu dans la régression de X_p par les p-2 premières variables (*cf.* Sect. 1.1) ;
- le second terme caractérise l'augmentation du coefficient de détermination précédent lorsque l'on ajoute X_{p-1} dans l'ensemble des variables explicatives. Par définition du coefficient de corrélation partielle, il est égal à $(1 - R^2_{p,p-2}) R^2(X_{p-1}, X_p / X_i, i = 1, \ldots, p-2)$. On retrouve ici la relation entre le terme $r_{p-1,p}$ de la matrice de corrélation et le coefficient de corrélation partielle correspondant (*cf.* Sect. 1.2).

Pour exprimer R^2 en fonction d'un coefficient de corrélation de la forme $r_{p,k}$, il suffit de permuter les variables X_k et X_{p-1}.

2.2. Expression du coefficient de détermination R^2 en fonction d'un coefficient de corrélation de la forme $r_{k,l} (k \neq p, l \neq n)$

Nous donnons maintenant l'expression du coefficient de détermination en fonction du coefficient de corrélation $r_{p-2,p-1}$ entre les deux variables X_{p-2} et X_{p-1}.

Ce coefficient de corrélation intervient dans le calcul des termes suivants: $a_{p-1,p-2}$, $a_{p-1,p-1}$, $a_{p,p-1}$, $a_{p,n}$.

Les formules (1) et (2) donnent:

$$a_{p-1,p-2} = \frac{r_{p-2,p-1} - \sum_{k=1}^{p-3} a_{p-2,k} a_{p-1,k}}{a_{p-2,p-2}} \qquad (8)$$

$$a_{p-1,p-1} = \left[1 - \sum_{k=1}^{p-2} a_{p-1,k}^2\right]^{1/2}.$$

Les deux autres termes: $a_{p,p-1}$ et $a_{p,p}$, n'interviennent pas dans la formule (7). On peut donc exprimer R^2 en fonction de $r_{p-1,p}$ et de $a_{p-1,p-2}$:

$$R^2 = \sum_{k=1}^{p-2} a_{p,k}^2 + \left[\frac{r_{p-1,p} - \sum_{k=1}^{p-3} a_{p-1,k} a_{p,k} - a_{p-1,p-2} a_{p,p-2}}{1 - \sum_{k=1}^{p-3} a_{p-1,k}^2 - a_{p-1,p-2}^2}\right]^2. \qquad (9)$$

Le coefficient de détermination R^2 est la fonction des coefficients de corrélation $r_{p-1,p}$ et $r_{p-2,p-1}$ obtenue en composant les formules (8) et (9). Cette fonction est définie sur l'ensemble de variation du couple $(r_{p-1,p}, r_{p-2,p-1})$ défini dans la section 1.2. En fixant $r_{p-1,p}$, on obtient l'expression de R^2 en fonction de $r_{p-2,p-1}$.

Comme précédemment, de simples permutations permettront d'exprimer le coefficient de détermination R_i^2 de la variable X_i dans la régression par les autres variables $X_j (j = 1, p$ et $j \neq i)$ en fonction de tout coefficient de corrélation $r_{k,l}$ entre les variables X_k et $X_l (k \neq l)$.

Le premier terme du second membre est le même que dans la formule (7).

Le second terme apparaît ici comme la contribution de la variable X_{p-1} au coefficient de détermination R^2 en fonction de $r_{p-2,p-1}$ et $r_{p,p-1}$: dans le paragraphe précédent, le coefficient $r_{p-2,p-1}$ était constant. On peut en déduire le coefficient de corrélation partielle $R(X_{p-1}, X_p / X_i, i = 1, \ldots, p-2)$ en fonction de $r_{p-2,p-1}$ et $r_{p,p-1}$, et en déduire, par la formule (6), la relation entre $r_{p-1,p}, r_{p-2,p-1}$ et les termes $s_{p-1,p}, s_{p-1,p-1}$ et $s_{p,p}$ de la matrice inverse, relation plus générale que celle donnée dans la section 1.2 qui ne concerne que $r_{p,p-1}$.

En posant le deuxième terme du second membre égal à 0, on établit facilement que l'ensemble des couples $(r_{p-1,p}, r_{p-2,p-1})$ pour lesquelles le

coefficient de corrélation partielle – ou le terme $s_{p-1,p}$ de la matrice inverse – est nul, est une droite de $\mathbf{R}^2$.

2.3. Dérivées

Pour évaluer l'effet sur les coefficients de détermination d'une faible variation d'un coefficient de corrélation, il suffit de calculer les dérivées des fonctions précédentes.

Un calcul simple donne la dérivée de R^2 par rapport à $r_{p-1,p}$:

$$dR^2/dr_{p-1,p} = 2\frac{r_{p-1,p} - \sum_{k=1}^{p-2} a_{p-1,k}a_{p,k}}{a_{p-1,p-1}^2}. \tag{10}$$

Si le numérateur est non nul, cette dérivée tend vers $\pm\infty$ lorsque $a_{p-1,p-1}$ tend vers 0, c'est-à-dire, compte tenu des propriétés classiques de la section 1.1, lorsque $R^2_{<p-1}$ tend vers 1. On constate donc que R^2 est d'autant plus instable par rapport à $r_{p-1,p}$ que X_{p-1} est bien reconstruite par les autres variables explicatives. On voit ainsi comment la quasi colinéarité des $p-1$ variables explicatives entraîne l'instabilité du coefficient de détermination.

Si le numérateur est nul, cela signifie que R^2 est égal à $R^2_{p,p-2}$, et donc que le coefficient de corrélation partielle $R(X_p, X_{p-1}/X_k, k < p-1)$ est nul. On est ramené au cas précédent, en ne considérant que les $p-2$ premières variables explicatives.

Pour calculer la dérivée de R^2 par rapport à $r_{p-2,p-1}$, on utilise la dérivée d'une fonction composée :

$$da_{p-1,p-2}/dr_{p-2,p-1} = \frac{1}{a_{p-2,p-2}}$$

$$dR^2/da_{p-1,p-2} = \frac{U}{\left[1 - \sum_{k=1}^{p-2} a_{p-1,k}^2\right]^2}$$

où :

$$U = -2\left[r_{p-1,p} - \sum_{k=1}^{p-2} a_{p-1,k}a_{p,k}\right] a_{p,p-2}\left[1 - \sum_{k=1}^{p-2} a_{p-1,k}^2\right]$$

$$+ 2\left[r_{p-1,p} - \sum_{k=1}^{p-2} a_{p-1,k}a_{p,k}\right]^2 a_{p-1,p-2}$$

$$dR^2/dr_{p-2,p-1} = \frac{1}{a_{p-2,p-2}} \frac{U}{\left[1 - \sum_{k=1}^{p-2} a_{p-1,k}^2\right]^2}.$$

Soit, compte tenue de (1):

$$dR^2/dr_{p-2,p-1} = \frac{1}{a_{p-2,p-2}} \frac{U}{a_{p-1,p-1}^4}. \tag{11}$$

Le dénominateur de l'expression (11) dépend de $a_{p-2,p-2}$ et de $a_{p-1,p-1}^4$. R^2 est donc beaucoup plus instable par rapport à $r_{p-2,p-1}$, c'est-à-dire au coefficient de corrélation entre corrélations entre les deux variables explicatives X_{p-2} et X_{p-1}, que par rapport au coefficient de corrélation $r_{p-1,p}$ entre X_{p-1} et X_p, sous réserve bien sûr que U soit différent de 0.

Pour calculer la dérivée du coefficient de détermination R^2 par rapport à un coefficient de corrélation $r_{k,l}$, il suffit d'effectuer les permutations nécessaires des variables explicatives.

3. EXEMPLES NUMÉRIQUES : ÉTUDE DE L'INSTABILITÉ DES ESTIMATIONS DANS LE MODÈLE LINÉAIRE

3.1. Colinéarité statistique et numérique

La quasi colinéarité des variables explicatives d'un modèle linéaire se manifeste par la présence d'une ou plusieurs valeurs propres faibles de la matrice de corrélation, et a de nombreuses conséquences.

La plus fréquente est une forte instabilité des coefficients de régression obtenus par l'estimateur des moindres carrés. Cette conséquence est de nature statistique, et ne met pas en doute la validité des calculs numériques. On peut la caractériser par les variances des estimateurs ou de préférence, par les facteurs d'inflation de la variance définis par les termes diagonaux de la matrice $\mathbf{R}^{-1}$, qui prennent de grandes valeurs en cas de quasi colinéarité.

Les méthodes utilisées habituellement dans ce cas ont pour objectif d'atténuer les effets statistiques de la colinéarité, et de donner des estimations des coefficients de régression stables et interprétables. Elles sont fondées:

- sur l'indice de multicolinéarité, défini par la moyenne des inverses des valeurs propres (régression bornée: Hoerl et Kennard, 1970a et b; Cazes, 1975; Foucart, 1998);
- sur la sélection de composantes principales (Jolliffe, 1982; Naes et Helland, 1993);

- sur l'indice de conditionnement défini par l'inverse de la racine carrée de la plus petite valeur propre (Belsley *et al.*, 1980);
- sur la recherche de composantes particulières dans l'ensemble des variables explicatives (Helland, 1992, régression PLS: Wold *et al.*, 1984; Tenenhaus, 1998).

Il peut arriver aussi que les calculs manquent totalement de précision à cause de l'inversion de la matrice de corrélation. Lorsque la matrice $\mathbf{R}$ est à la limite de l'inversibilité – sa dernière valeur propre est quasi nulle –, le calcul en ordinateur reste parfois possible tout en donnant des résultats absurdes: il est recommandé dans ce cas-là d'effectuer le produit matriciel $\mathbf{R}\mathbf{R}^{-1}$ pour vérifier que l'on retrouve la matrice identité $\mathbf{I}$. Dans le cas où $\mathbf{R}\mathbf{R}^{-1}$ est différente de $\mathbf{I}$, on dit alors que la matrice $\mathbf{R}$ est « mal conditionnée ».

Notre démarche est ici différente: nous supposons exacts les calculs numériques et examinons les liaisons entre le coefficient de détermination et les coefficients de corrélation.

3.2. Exemples numériques

Nous étudions à titre d'exemples numériques deux matrices de corrélation construites terme à terme à l'aide d'un algorithme déduit de la définition de l'intervalle de variation d'un coefficient de corrélation d'une matrice $\mathbf{R}$.

La première matrice donne les corrélations entre quatre variables X_1, X_2, X_3 et Y:

TABLEAU 1
Matrice des corrélations entre les 4 variables.

	X_1	X_2	X_3	Y
X_1	1			
X_2	0,6000	1		
X_3	–0,2790	0,6	1	
Y	0,0446	0	0	1

Étudions tout d'abord la matrice de corrélation entre les trois variables explicatives. Les valeurs propres de cette matrice sont égales à:

$$\lambda_1 = 1,720419 \qquad \lambda_2 = 1,279000 \qquad \lambda_3 = 0,00058.$$

La troisième valeur propre, par sa petite taille, indique une quasi colinéarité entre les variables explicatives. Le coefficient de détermination noté comme précédemment R^2 est égal à 0,99536. Le calcul des dérivées du coefficient

de détermination R^2 par rapport aux coefficients de corrélation montre que la régression est très instable :

TABLEAU 2
Dérivées du coefficient de détermination R^2 de la variable Y.

	X_1	X_2	X_3
X_2	1194,429		
X_3	–994,579	1192,563	
Y	44,635	–53,520	44,565

Le coefficient de détermination R^2 est donc très sensible aux coefficients de corrélation entre les variables explicatives X_1, X_2, X_3, tandis qu'il l'est beaucoup moins aux coefficients de corrélation entre les variables explicatives et la variable expliquée $r_{1,4}$, $r_{2,4}$ et $r_{3,4}$: nous retrouvons le résultat expliqué dans le paragraphe précédent.

Considérons le couple (1,2) dont l'intervalle de variation est $]-0,9348038, 0,6000039[$. La figure 1 ci-dessous représente R^2 en fonction de $r_{1,2}$: R^2 reste très proche de 0 pour toute valeur de $r_{1,2}$ sauf près des bornes de son intervalle de variation où la convergence vers 1 est très rapide.

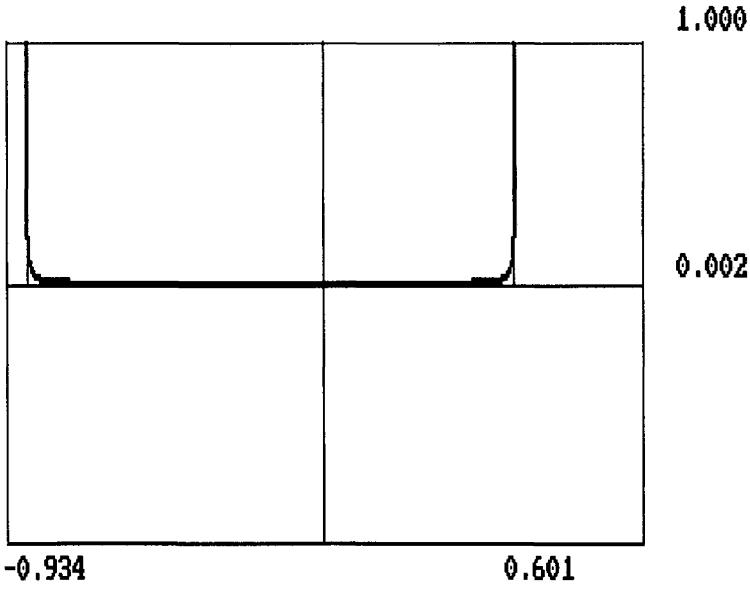

Figure 1.

En diminuant le coefficient $r_{1,2}$ de 0,001 ($r_{1,2} = 0,599$ au lieu de 0,6), on obtient un coefficient de détermination égal à 0,45260, beaucoup plus faible que le précédent (0,99536).

Étudions maintenant la matrice de corrélation (Tab. 3):

TABLEAU 3
Matrice des corrélations entre les 6 variables.

	X_1	X_2	X_3	X_4	X_5	Y
X_1	1					
X_2	0,300	1				
X_3	-0,300	0,500	1			
X_4	-0,500	0,600	0,400	1		
X_5	0,300	-0,200	0,300	-0,460	1	
Y	0,704	0,752	0,263	-0,022	-0,121	1

Les valeurs propres de la matrice des corrélations entre les variables explicatives sont les suivantes:

$$\lambda_1 = 2,166143, \qquad \lambda_2 = 1,385893, \qquad \lambda_3 = 1,225122,$$
$$\lambda_4 = 0,2222648, \qquad \lambda_5 = 0,00058.$$

La dernière valeur propre est $\lambda_5 = 0,00058$, égale à la dernière valeur propre de la matrice précédente. Elle caractérise ici aussi l'existence de quasi colinéarité entre les variables explicatives. Le coefficient de détermination est égal à 0,99599, mais la régression est beaucoup plus stable que la précédente, comme en témoignent les dérivées du coefficient de détermination R^2 par rapport aux coefficients de corrélation:

TABLEAU 4
Dérivées du coefficient de détermination R^2 de la variable Y.

	X_1	X_2	X_3	X_4	X_5
X_2	-0,46				
X_3	-0,19	-0,60			
X_4	0,41	1,28	0,53		
X_5	0,25	0,78	0,32	-0,68	
Y	0,54	1,71	0,71	-1,50	-0,91

En posant $r_{2,4} = 0,599$, on obtient comme coefficient de détermination 0,99535, presqu'égal au précédent.

En conclusion, on pourra difficilement trouver une valeur influente sur le coefficient de détermination R^2 dans le second cas alors que dans le premier, l'existence d'une telle valeur n'est pas exclue.

Tous les calculs précédents ont été effectués en double précision, mais cette précaution ne fait que repousser les limites de capacité de calcul de l'ordinateur. Pour vérifier nos calculs, nous les avons recommencés en simple précision : les résultats obtenus sont les mêmes que les premiers, au moins jusqu'à la quatrième décimale.

On peut donc considérer que la précision des calculs est suffisante ici pour mettre en évidence les effets de la quasi colinéarité dans les exemples donnés.

CONCLUSION

L'analyse des dérivées du coefficient de détermination par rapport aux coefficients de corrélation entre les variables explicatives apporte un éclairage nouveau sur l'instabilité numérique des estimations dans le cas de quasi colinéarité entre ces variables. Comme l'étude des dérivées sur les exemples numériques l'a montré, les résultats ne sont pas toujours instables, du moins au niveau du coefficient de détermination.

On notera aussi que l'instabilité numérique concerne toutes les méthodes statistiques faisant appel à l'inversion de matrice, comme l'analyse factorielle discriminante, l'analyse canonique etc.

ANNEXE

L'intervalle de variation du coefficient de corrélation $r_{p-1,p}$ est l'intervalle $]\alpha_{p-1,p}, \gamma_{p-1,p}[$ où :

$$\alpha_{p-1,p} = -a_{p-1,p-1}\left[1 - \sum_{k=1}^{p-2} a_{p,k}^2\right]^{1/2} + \sum_{k=1}^{p-2} a_{p-1,k}a_{p,k}$$

$$\gamma_{p-1,p} = a_{p-1,p-1}\left[1 - \sum_{k=1}^{p-2} a_{p,k}^2\right]^{1/2} + \sum_{k=1}^{p-2} a_{p-1,k}a_{p,k}.$$

Cette propriété s'étend à un terme quelconque non diagonal $r_{i,j}$ de la matrice par une simple permutation des lignes et des colonnes de la matrice i et $p-1$ d'une part, j et p d'autre part : la matrice $\mathbf{R}$ est en effet symétrique définie positive quel que soit l'ordre des variables X_j, et l'intervalle $I_{i,j}$, appelé *intervalle de variation* du coefficient de corrélation $r_{i,j}$, est invariant dans toute permutation de ces variables.

REMERCIEMENTS

L'auteur remercie les referes et l'un d'entre eux tout particulièrement pour leurs observations et leurs conseils qui ont permis une notable amélioration de l'article.

RÉFÉRENCES

1. D.A. Belsley, E. Kuh et R.E. Welsh, *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley, New York (1980).
2. P.J. Brockwell et R.A. Davis, *Time series: Theory and methods*. Springer Series in Statistics (1998).
3. P. Cazes, Protection de la régression par utilisation de contraintes linéaires et non linéaires. *Rev. Statist. Appl.* **XXIII** (1975) 37-57.
4. P.G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimisation*. Cambridge University Press, London (1989).
5. T. Foucart, Transitivité du produit scalaire. *Rev. Statist. Appl.* **XXXIX** (1991) 57-68
6. T. Foucart, Numerical Analysis of a Correlation Matrix. *Statistics* **29** (1997) 347-361.
7. T. Foucart, Stabiliby of the inverse correlation matrix. Partial ridge regression. *J. Statist. Plann. Inference* **77** (1999) 141–154.
8. D.M. Hawkins et W.J.R. Eplett, The Cholesky Factorization of the Inverse Correlation or Covariance Matrix in Multiple Regression. *Technometrics* **24** (1982) 191-198.
9. I.S. Helland, Maximum Likelihood Regression on Relevant Components. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **54** (1992) 637-647.
10. A.E. Hoerl et R.W. Kennard, Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics* **12** (1970a) 55-67.
11. A.E. Hoerl et R.W. Kennard, Ridge Regression: Applications to nonorthogonal problems. *Technometrics* **12** (1970b) 69-82.
12. I.T. Jolliffe, A note on the use of principal components in regression. *Appl. Statist.* **31** (1982) 300-303.
13. T. Naes et I.S. Helland, Relevant Components in Regression. *Scand. J. Statis.* **20** (1993) 239-250.
14. M. Tenenhaus, La régression PLS, théorie et pratique. Technip, Paris (1998).
15. J. Whittaker, *Graphical models in applied multivariate statistics*. Wiley, New York (1990).
16. S. Wold, A. Ruhe, H. Wold et W.J. Dunn III, The collinearity problem in linear regression, The partial least squares (PLS) approached to generalized inverses. *SIAM Sci. Stat. Comp.* **5** (1984) 735-743.