REVUE DE STATISTIQUE APPLIQUÉE

T. FOUCART

Collinéarité dans une matrice de produit scalaire

Revue de statistique appliquée, tome 40, n° 3 (1992), p. 5-17

<http://www.numdam.org/item?id=RSA_1992__40_3_5_0>

Rev. Statistique Appliquée, 1992, **XXXX** (3), 5-17

COLLINÉARITÉ DANS UNE MATRICE DE PRODUIT SCALAIRE

T. Foucart

Département de Mathématiques, Université d'Orléans

RÉSUMÉ

Nous complétons dans cet article des résultats déjà publiés sur l'intervalle de variation de chaque terme d'une matrice symétrique définie positive et les appliquons à la recherche des collinéarités des variables explicatives dans une régression.

Mots-clés : *Produit scalaire, Inégalité de Schwarz, Matrice symétrique définie positive, Corrélation, Collinéarité, Régression.*

ABSTRACT

In this paper we complete results that have been proved in a previous paper about interval of variation of each term of a symmetric definite positive matrix, and apply them to seek collinearities between dependent variates in a regression analysis.

Key-words : *Dot product, Schwarz's inequality, Symmetric definite positive matrix, Correlation, Collinearity, Regression.*

1. Intervalle de variation d'un terme d'une matrice symétrique définie positive

1.1 Rappel des résultats déjà établis (Foucart, 1991)

Soit R une matrice symétrique définie positive ; on considère le terme $r_{i,j}$ de la i^{e} ligne et de la j^{e} colonne.

On étudie tout d'abord les termes non diagonaux $(i \neq j)$.

Nous avons montré qu'il existe un intervalle $I_v =]a, b[$ dont les bornes sont des fonctions continues des autres termes de la matrice et tel que, quel que soit $r_{i,j} \in]a, b[$, la matrice R est symétrique définie positive.

Nous appelons cet intervalle «Intervalle de variation» du terme $r_{i,j}$ et le notons I_v.

L'intervalle I_v est inclus dans l'intervalle I_s déduit de l'inégalité de Schwarz, que nous appellons pour simplifier intervalle de Schwarz :

$$I_v =]a, b[\subset I_s = [-[r_{i,i} r_{j,j}]^{1/2}, [r_{i,i} r_{j,j}]^{1/2}]$$

La démonstration de cette propriété consiste à placer le terme $r_{i,j}$ à la $n-1^{\text{e}}$ ligne et à la n^{e} colonne de la matrice. On factorise alors la matrice ainsi obtenue, que nous noterons encore R, par la méthode de Choleski (Ciarlet P.G., 1989) :

$$R = BB^t$$

où B est la matrice triangulaire inférieure définie par :

$$\forall i = 1, n \qquad b_{i,1} = r_{1,i} / \sqrt{r_{1,1}} \tag{1}$$

$$\forall i = 2, n \qquad b_{i,i} = \left[r_{i,i} - \sum_{k=1}^{i-1} b_{i,k} \right]^{1/2} \tag{2}$$

$$\forall i = 2, n,$$

$$\forall j = i+1, n \qquad b_{j,i} = \frac{r_{i,j} - \sum_{k=1}^{i-1} b_{i,k} b_{j,k}}{b_{i,i}} \tag{3}$$

L'intervalle de variation $]a, b[$ est alors donné par les formules suivantes :

$$a = -b_{n-1,n-1} \left[r_{n,n} - \sum_{k=1}^{n-2} b_{n,k}^2 \right]^{1/2} + \sum_{k=1}^{n-2} b_{n-1,k} b_{n,k} \tag{4}$$

$$b = b_{n-1,n-1} \left[r_{n,n} - \sum_{k=1}^{n-2} b_{n,k}^2 \right]^{1/2} + \sum_{k=1}^{n-2} b_{n-1,k} b_{n,k} \tag{5}$$

Pour caractériser les intervalles de variation, nous avons proposé le coefficient suivant, que nous appelons coefficient de contrainte :

$$c_k = 1 - (b-a) / (2[r_{n,n} r_{n-1,n-1}]^{1/2})$$

Ce coefficient prend la valeur 1 pour $a = b$: l'intervalle de variation est vide et la contrainte est maximale, et la valeur 0 lorsque $I_v = I_s$: les autres termes de la matrice n'imposent aucune contrainte supplémentaire à l'inégalité de Schwarz.

Etudions maintenant les termes diagonaux $r_{i,i}$: par la transposition adéquate, on peut placer le terme $r_{i,i}$ à la n^{e} ligne et à la n^{e} colonne de la matrice, et la relation (2) montre alors que le terme $r_{n,n}$ est minoré par un nombre fonction continue des autres termes :

$$r_{n,n} > \sum_{k=1}^{n-1} b_{n,k}^2 \tag{6}$$

1.2 Positionnement d'un coefficient dans son intervalle

Nous supposons maintenant que la matrice R est la matrice des corrélations entre n variables.

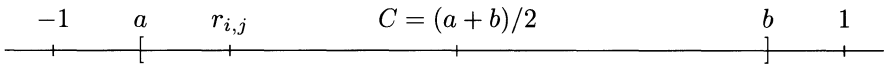
Rappelons avant de continuer la propriété fondamentale de la matrice S inverse de R : cette matrice est, au facteur σ^2/N près où σ^2 est la variance résiduelle et N le nombre d'observations, la matrice des covariances entre les estimateurs des coefficients de régression. En outre, les termes diagonaux de cette matrice sont égaux à $1/(1 - R_i^2)$, où R_i^2 est le coefficient de détermination de la variable x_i régressée par les autres variables explicatives x_j (Hawkins D.M., Eplett W.J.R., 1982).

Le terme $1/(1 - R_i^2)$ est appelé facteur d'inflation de la variable x_i.

Pour situer chaque terme de la matrice R dans son intervalle de variation, nous proposons un indice que nous avons appelé indice de distorsion :

$$d_{i,j} = \frac{2\left(\frac{a+b}{2} - r_{i,j}\right)}{b-a}$$

qui est proportionnel à la différence du terme $r_{i,j}$ et du centre de son intervalle de variation $]a, b[$, et qui varie entre -1 et $+1$. On suppose ici $i \neq j$



Une petite valeur de $d_{i,j}$ signifie que $r_{i,j}$ est proche du centre de l'intervalle, une valeur proche de ± 1 que $r_{i,j}$ est proche de l'une des bornes.

Une propriété supplémentaire donne une interprétation intéressante de ce coefficient : il est égal au coefficient de corrélation entre les estimateurs des coefficients de régression correspondants. Ainsi, un coefficient de corrélation proche de l'une des bornes de son intervalle de variation crée une corrélation élevée entre les estimateurs des coefficients de régression correspondants.

Théorème : on a la relation suivante :

$$d_{i,j} = \frac{s_{i,j}}{[s_{i,i}s_{j,j}]^{1/2}}$$

où les termes $s_{i,j}$ sont les termes de la matrice S inverse de la matrice des corrélations R.

Les formules (3), (4) et (5) permettent de calculer facilement $(a + b)/2 - r_{n-1,n}$ et $b - a$. On trouve :

$$r_{n-1,n} - (a+b)/2 = b_{n,n-1}b_{n-1,n-1}$$
$$b - a = 2b_{n-1,n-1}[b_{n,n}^2 + b_{n,n-1}^2]^{1/2}$$

On en déduit l'indice de distorsion :

$$d_{n-1,n} = -\frac{b_{n,n-1}}{[b_{n,n-1}^2 + b_{n,n}^2]^{1/2}}$$

Pour montrer l'égalité, il suffit de calculer les termes $s_{n-1,n-1}$, $s_{n-1,n}$ et $s_{n,n}$ de la matrice $S = R^{-1}$. Par définition de la matrice B, on a :

$$R = BB^t$$

L'inverse de la transposée de B étant la transposée de l'inverse de B, on en déduit :

$$S = R^{-1} = (B^{-1})^t B^{-1}$$

Posons $C = (B^{-1})^t$. On a, C étant comme B^t triangulaire supérieure :

$$\begin{aligned} c_{n,n} &= 1/b_{n,n} \\ c_{n-1,n-1} &= 1/b_{n-1,n-1} \\ c_{n-1,n} &= -\frac{b_{n,n-1}}{b_{n,n} b_{n-1,n-1}} \end{aligned}$$

En effectuant le produit matriciel CC^t, on obtient :

$$\begin{aligned} s_{n,n} &= \frac{1}{b_{n,n}^2} \\ s_{n-1,n} &= -\frac{b_{n,n-1}}{b_{n,n}^2 b_{n-1,n-1}} \\ s_{n-1,n-1} &= \frac{1}{b_{n-1,n-1}^2} + \frac{b_{n,n-1}^2}{b_{n,n}^2 b_{n-1,n-1}^2} \end{aligned} \tag{7}$$

On déduit de toutes ces formules :

$$\frac{s_{n-1,n}}{(s_{n,n} s_{n-1,n-1})^{1/2}} = -\frac{b_{n,n-1}}{(b_{n,n}^2 + b_{n,n-1}^2)^{1/2}}$$

Remarque : il est facile de voir que si tous les coefficients de corrélation des variables explicatives x_i sont au centre de leurs intervalles de variation, la matrice S est diagonale, la matrice R, qui en est l'inverse, également, ce qui signifie que les variables explicatives sont non-corrélées.

Etudions maintenant le cas des termes diagonaux. Les formules (2) et (7) nous donnent la relation entre $s_{n,n}$ et $r_{n,n}$:

$$s_{n,n} = \frac{1}{r_{n,n} - \sum_{k=1}^{n-1} b_{n,k}^2}$$

Le terme diagonal de la matrice inverse de R étant égal à $1/(1-R_n^2)$ et $r_{n,n}$ étant égal à 1, on a :

$$1 - \sum_{k=1}^{n-1} b_{n,k}^2 = 1 - R_n^2$$

Soit :

$$R_n^2 = \sum_{k=1}^{n-1} b_{n,k}^2$$

La borne inférieure du terme diagonal $r_{n,n}$ (cf (6)) est donc égale au coefficient de détermination R_n^2 de la variable x_n régressée par les variables $x_i, i = 1, n-1$: à un coefficient de détermination élevé correspond un terme diagonal proche de sa valeur minimale.

Cette propriété est vraie quel que soit le terme diagonal que l'on peut toujours placer à la dernière ligne et à la dernière colonne de la matrice. Il apparaît donc que plus la distance entre un terme diagonal $r_{j,j}$ et son minorant est faible, plus le coefficient de détermination entre la variable x_i et les variables x_j pour $j \neq i$ est proche de 1 : on retrouve ici le facteur d'inflation.

Les propriétés que nous venons de démontrer peuvent être appliquées de manière réciproque, c'est-à-dire que la matrice de corrélation R entre les variables explicatives x_j donne les coefficients de distorsion des corrélations entre les estimateurs des coefficients de régression : un coefficient de corrélation élevé entre deux variables explicatives signifie que les estimateurs des coefficients de régression correspondants ont un coefficient de corrélation proche de l'une de ses bornes. Il y a dualité entre les deux matrices.

1.3 Recherche de la collinéarité

On sait que, lorsque l'on effectue la régression d'une variable expliquée y par des variables explicatives x_i, les collinéarités entre les variables explicatives (réduites ou non) sont gênantes d'une part pour des raisons numériques (la matrice des corrélations R est mal conditionnée et les calculs imprécis) et d'autre part pour des raisons statistiques : elles peuvent augmenter sensiblement les variances des estimateurs des coefficients de régression, ou donner des estimations difficiles à interpréter (un coefficient positif est par exemple compensé par un coefficient

négatif). La recherche de la collinéarité entre des variables explicatives présente donc un double intérêt.

Plusieurs méthodes existent déjà pour rechercher et comprendre ces collinéarités.

Parmi elles, l'analyse en composantes principales nous paraît fondamentale. En effet, la collinéarité exacte de deux variables explicatives se traduit par la nullité d'une valeur propre de la matrice des corrélations : la taille des plus petites valeurs propres donne donc une indication que l'on peut résumer à l'aide de l'indice de multicollinéarité (Tomassone, 1983 ou Belsley, 1980) :

$$F = 1/n \sum_{i=1}^{n} 1/\lambda_i$$

où les λ_i sont les valeurs propres de la matrice de corrélation R.

Dans le cas de non-corrélation des variables explicatives en effet, ces valeurs propres sont égales à 1 et F prend la valeur 1, alors qu'une petite valeur propre augmente considérablement le critère F.

Ce critère général peut être complété par l'étude des coefficients de détermination R_i^2 de chaque variable x_i par rapport aux autres variables explicatives dont il peut être présenté comme un indice synthétique puisqu'il est égal à la moyenne des facteurs d'inflation :

$$F = 1/n \sum_{i=1}^{n} 1/(1 - R_i^2)$$

Un coefficient R_i^2 élevé signifie que la variable x_i peut être expliquée presque exactement par les autres, et donc que la matrice est proche de la collinéarité.

On peut minorer facilement le coefficient R_i^2 par le carré du coefficient de corrélation le plus élevé entre x_i et les autres variables explicatives, mais cette minoration peut être très insuffisante parce que le coefficient de détermination R_i^2 peut être très proche de 1 même lorsque les coefficients de corrélation restent tous faibles (Belsley, 1980).

Pour atténuer l'effet des collinéarités, il est naturel d'effectuer la régression sur les composantes principales en éliminant celles dont la variance est trop faible. En les conservant toutes, on retrouve évidemment la régression classique.

On peut aussi borner la norme du vecteur de régression (Hoerl A.E., Kennard R.W., 1970) : cela revient à ajouter une constante positive k aux termes diagonaux de la matrice et à estimer les coefficients de régression par la formule :

$$B = (R + kI)^{-1} XY$$

où X est le tableau des variables explicatives centrées réduites et Y la colonne définie par les observations de la variable expliquée y centrée et réduite.

On notera que cette méthode appelée ridge regression revient à ajouter une constante k aux valeurs propres λ de la matrice R et à en laisser invariants les

vecteurs principaux : on effectue une homothétie de rapport $\lambda_a/(\lambda_a + k)$ sur les coordonnées de la projection de la variable expliquée sur les composantes principales de rang α. La projection effectuée n'est plus orthogonale au sens de la covariance.

La ridge regression généralisée consiste à ajouter une constante k_α à chaque valeur propre λ_α, et à effectuer une homothétie de rapport $\lambda_\alpha/(\lambda_\alpha + k_\alpha)$ sur les coordonnées sur les composantes principales (Cazes, 1991).

Nous proposons une approche différente, qui modifie à la fois les valeurs propres et les vecteurs principaux de R : l'interprétation que nous avons donnée des corrélations entre les estimateurs des coefficients de régression nous permet de déterminer les corrélations entre les variables explicatives responsables des collinéarités ; il suffit ensuite de modifier ces corrélations en les écartant des bornes de leur intervalle de variation et d'estimer les coefficients de régression par l'estimateur :

$$B' = R'^{-1}XY$$

où R' est la matrice des corrélations corrigées. On obtient ainsi l'estimateur :

$$Y' = XB' \tag{8}$$

Dans la mesure où ces corrections sont minimes, on obtiendra des estimateurs satisfaisants des coefficients de régression.

Une autre procédure consiste à augmenter les termes diagonaux de R jugés trop proches de leur borne inférieure ; on obtient ainsi une matrice R' et un estimateur B' analogues aux précédents.

2. Application numérique

Nous reprenons les données publiées par Tomassone *et alt.* (1983) pour appliquer les méthodes précédentes et en comparer les résultats.

Les données de base sont constituées de 33 observations de 11 variables, la onzième étant la variable expliquée. L'objectif de la régression est de connaître l'influence de certaines caractéristiques de peuplements forestiers sur le développement d'un parasite du pin appelé la processionnaire.

On trouvera dans l'ouvrage de Tomassone les détails des variables explicatives, notre objectif se limitant ici à l'étude des propriétés numériques de la matrice de corrélation que nous donnons ci-dessous et dans laquelle on distingue immédiatement les coefficients de corrélation entre les variables explicatives élevés : entre x_4 et x_5 (0.905), entre x_3 et x_6 (0.980), entre x_6 et x_9 (0.909), entre x_3 et x_9 (0.877) et enfin entre x_8 et x_9 (0.854) :

Matrice étudiée

1	1									
2	0.121	1								
3	0.538	0.322	1							
4	0.321	0.137	0.414	1						
5	0.284	0.113	0.295	0.905	1					
6	0.515	0.301	0.980	0.439	0.306	1				
7	0.269	−0.152	0.128	0.058	−0.079	0.151				
8	0.360	0.262	0.759	0.772	0.596	0.810	0.060	1		
9	0.364	0.326	0.877	0.460	0.267	0.909	0.063	0.854	1	
10	−0.100	0.129	0.206	−0.045	−0.025	0.130	0.138	0.054	0.175	1
11	−0.534	−0.429	−0.518	−0.425	−0.201	−0.528	−0.230	−0.541	−0.594	−0.063	1

Les résultats de la régression effectuée à l'aide de l'estimateur habituel des moindres carrés sont les suivants :

Analyse de variance et régression
(Matrice initiale)

Variance résiduelle : 0.45759 Ecart-type résiduel : 0.67646

Coefficient de détermination : 0.6949 $F(10, 22) = 5.0117$

Coef.	estimation	écart-type	t de Student	Facteur d'inflation
1	−0.459	0.161	−2.846	1.878
2	−0.316	0.128	−2.458	1.190
3	0.521	0.762	0.683	41.874
4	−1.082	0.471	−2.295	16.023
5	0.801	0.361	2.219	9.385
6	−0.206	0.903	−0.228	58.760
7	−0.036	0.151	−0.236	1.649
8	0.342	0.447	0.765	14.423
9	−0.585	0.394	−1.486	11.169
10	−0.089	0.151	−0.590	1.651

L'indice de multicollinéarité est $F = 15.80$.

2.1 La question que nous nous posons est la suivante : quelles sont les corrélations responsables de la quasicollinéarité des variables ?

On retrouve dans la matrice des coefficients de distorsion (qui sont les corrélations entre les estimateurs des coefficients de régression) des valeurs élevées : $d_{4,5}(-0.85)$, $d_{3,6}(-0.92)$ mais les autres termes sont faibles : $d_{6,9}(-0.31)$, $d_{3,9}(0.078)$ et $d_{8,9}(-0.39)$: il est naturel de se demander pourquoi.

Nous proposons une interprétation numérique en examinant les intervalles de variation des coefficients de corrélation des variables explicatives.

Matrice des coefficients de distorsion

1	1									
2	−0.020	1								
3	−0.274	−0.092	1							
4	−0.004	0.013	−0.288	1						
5	−0.114	−0.015	0.194	−0.847	1					
6	0.095	0.069	−0.919	0.419	−0.309	1				
7	−0.286	0.150	0.332	−0.436	0.469	−0.397	1			
8	0.098	−0.002	0.337	−0.599	0.190	−0.422	0.177	1		
9	0.041	−0.091	0.078	−0.155	0.378	−0.314	0.247	−0.393	1	
10	0.318	−0.047	−0.508	0.289	−0.300	0.486	−0.395	−0.090	−0.248	1

Matrice des valeurs minimales

1	0.467									
2	−0.562	0.160								
3	0.382	0.166	0.976							
4	0.138	−0.089	0.360	0.938						
5	0.015	−0.190	0.253	0.373	0.893					
6	0.428	0.189	0.731	0.416	0.245	0.983				
7	−0.527	−0.773	0.038	−0.287	−0.252	−0.018	0.394			
8	0.185	0.020	0.729	0.608	0.524	0.751	−0.114	0.931		
9	0.154	0.024	0.834	0.371	0.197	0.852	−0.124	0.724	0.910	
10	−0.531	−0.619	−0.038	−0.196	−0.387	0.062	−0.864	−0.172	− 0.135	0.394

Matrice des valeurs maximales

1	1									
2	0.776	1								
3	0.626	0.452	1							
4	0.503	0.369	0.444	1						
5	0.498	0.408	0.358	0.949	1					
6	0.620	0.429	0.990	0.495	0.339	1				
7	0.710	0.687	0.309	0.194	0.400	0.223	1			
8	0.573	0.503	0.820	0.813	0.702	0.834	0.309	1		
9	0.591	0.577	0.927	0.524	0.425	0.938	0.373	0.910	1	
10	0.732	0.810	0.286	0.228	0.171	0.328	0.572	0.241	0.361	1

Les coefficients de corrélation $r_{4,5}$ et $r_{3,6}$ sont proches de leurs bornes supérieures alors que ce n'est pas le cas des coefficients $r_{3,9}$, $r_{6,9}$ et $r_{8,9}$.

Un fort coefficient de corrélation ne crée donc pas nécessairement de collinéarité entre l'ensemble des variables. Par contre, la propriété de dualité entre R et R^{-1} montre qu'il indique une collinéarité entre les estimateurs des coefficients de régression.

2.2 Posons-nous maintenant la question inverse : existe-t-il des corrélations peu élevées créant des collinéarités ?

On peut examiner directement la matrice des coefficients de distorsion ci-dessus : chaque terme est la distance relative du coefficient de corrélation au centre de son intervalle : il est clair qu'il n'y a pas d'autre liaison de la même nature que les précédentes.

Mais cela pourrait se produire : pour le montrer, nous changeons un coefficient de corrélation de la matrice et posons : $r_{1,5} = 0.02$ très proche de sa borne inférieure. Pour cette valeur presque nulle, l'indice de multicollinéarité est élevé ($F = 94.495$) et la plupart des facteurs d'inflation deviennent importants :

$$R_1^2 = 0.978 \quad R_2^2 = 0.191 \quad R_3^2 = 0.979 \quad R_4^2 = 0.997 \quad R_5^2 = 0.996$$
$$R_6^2 = 0.993 \quad R_7^2 = 0.675 \quad R_8^2 = 0.980 \quad R_9^2 = 0.986 \quad R_{10}^2 = 0.338$$

En posant $r_{1,5} = 0.02$, nous avons donc créé des collinéarités entre les variables explicatives.

On peut remarquer le paradoxe suivant : en posant $r_{1,5} = 0.02$, on a considérablement augmenté le coefficient de détermination R_1^2, qui passe de 0.467 à 0.978.

2.3 Notre dernière question est la suivante : peut-on modifier des coefficients de corrélation de façon à diminuer les collinéarités ?

Les coefficients responsables des collinéarités sont essentiellement $r_{4,5}$ et $r_{3,6}$. Nous les diminuons légèrement (de 0.02) et posons :

$$r_{3,6} = 0.95955 \qquad r_{4,5} = 0.88466$$

Nous donnons ci-dessous les résultats de la régression effectuée à l'aide de l'estimateur B' défini ci-dessus : $Y' = XB'$ est donc l'estimateur de Y.

Pour interpréter ces résultats et comparer Y' à l'estimateur $\widehat{Y}$ obtenu de manière habituelle, c'est-à-dire par l'estimateur efficace $B = R^{-1}XY$, nous calculons les paramètres suivants (*cf.* figure ci-dessous) :

(i) les paramètres optimaux qui sont obtenus par la régression multilinéaire habituelle (estimateur optimal car efficace) :

– $\widehat{Y}$ est la projection orthogonale de Y sur l'espace F engendré par les variables explicatives.

– le coefficient de détermination R^2 est le cosinus carré des vecteurs définis par la variable expliquée Y et son estimation $\widehat{Y}$: $\cos^2 \Theta$.

– la variance résiduelle estimée VR est la somme des carrés des écarts $y_i - \widehat{y}_i$ pondérés par le facteur $1/(N-n-1)$ où N est le nombre d'observations et n le nombre de variables explicatives. Cette variance résiduelle est minimale dans la classe des estimateurs sans biais.

– la corrélation entre les résidus et la variable expliquée estimée est nulle à cause de l'orthogonalité.

(ii) les paramètres de la seconde régression :

– Y' n'est pas la projection orthogonale de Y sur l'espace F.

– le coefficient de détermination R'^2 est comme le précédent le cosinus carré des vecteurs définis par la variable expliquée Y et son estimation Y' : $\cos^2 \Theta'$.

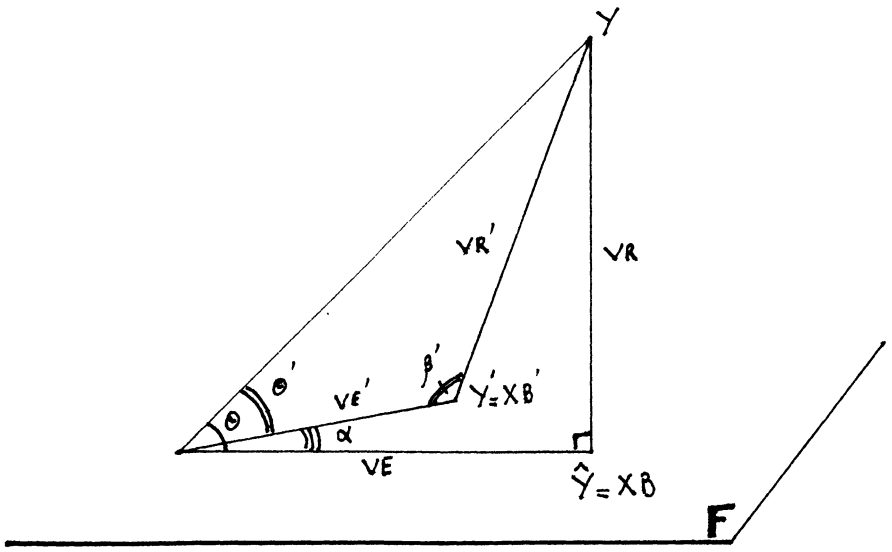
– la variance résiduelle estimée VR' est la somme des carrés des écarts $y_i - y'_i$ pondérés par le facteur $1/(N-n-1)$ où N est le nombre d'observations et n le nombre de variables explicatives.

– la corrélation entre les résidus et la variable expliquée estimée Y' est d'autant plus proche de 0 que le vecteur $Y - Y'$ et l'espace F sont proches de l'orthogonalité : cos β'.

(iii) des paramètres permettant de comparer les estimations Y et Y' :

– la corrélation entre les estimations Y et Y' de la variable expliquée mesure la proximité entre l'estimation obtenue par l'estimateur efficace et l'estimation obtenue par l'estimateur B' considéré : cos α.

– le rapport des variances expliquées VE et VE', qui sont les carrés des normes des vecteurs $\widehat{Y}$ et $\widehat{Y}'$, complète le paramètre précédent pour juger de l'écart entre $\widehat{Y}$ et Y'.



Projection de la variable expliquée sur l'espace F engendré par les variables explicatives X_j, $j = 1, n$.

Analyse de variance et régression (Matrice corrigée)

Variance résiduelle estimée :	0.4713	(optimale :	0.4576)
Ecart-type résiduel estimé :	0.6865	(optimal :	0.6765)
Coefficient de détermination :	0.6861	(optimal :	0.6949)
$F(10, 22) = 4.8075$		(optimal :	5.0117)

Corrélation entre résidus et variable expliquée estimée : 0.0285
Rapport des variances expliquées optimale et estimée : 1.0531
Corrélation entre les estimations de la variable expliquée : 0.9936

Coef.	estimation	écart-type	t de Student	Facteur d'inflation
1	−0.456	0.165	−2.774	1.895
2	−0.314	0.130	−2.412	1.184
3	0.287	0.312	0.921	6.821
4	−0.718	0.333	−2.154	7.783
5	0.529	0.263	2.012	4.846
6	0.195	0.357	0.544	8.947
7	−0.098	0.138	−0.709	1.326
8	0.176	0.420	0.418	12.369
9	−0.702	0.401	−1.751	11.248
10	−0.045	0.134	−0.340	1.251

On peut comparer ces résultats à ceux que donne la ridge regression (Tomassonne, 1983) : par rapport à cette dernière, les coefficients significativement non nuls sont les mêmes, les facteurs d'inflation sont légèrement supérieurs (le 8 et le 9 en particulier), mais l'estimation est plus proche de l'estimation efficace :

Analyse de variance et régression (Ridge regression, $k = 0.05$)

Variance résiduelle estimée :	0.4988	(optimale :	0.4576)
Ecart-type résiduel estimé :	0.7063	(optimal :	0.6765)
Coefficient de détermination :	0.6714	(optimal :	0.6949)
$F(10, 22) = 4.4956$		(optimal :	5.0117)

Corrélation entre résidus et variable expliquée estimée : 0.1095
Rapport des variances expliquées optimale et estimée : 1.2150
Corrélation entre les estimations de la variable expliquée : 0.9829

Coef.	estimation	écart-type	t de Student	Facteur d'inflation
1	−0.396	0.145	−2.736	1.384
2	−0.300	0.126	−2.393	1.043
3	0.170	0.196	0.868	2.550
4	−0.522	0.211	−2.474	2.950
5	0.427	0.194	2.201	2.492
6	0.127	0.184	0.690	2.243
7	−0.107	0.127	−0.845	1.064
8	−0.018	0.226	−0.082	3.387
9	−0.441	0.225	−1.960	3.343
10	−0.034	0.124	−0.274	1.024

3. Conclusion

L'application que nous avons proposée dans le deuxième paragraphe des propriétés établies dans le premier montre l'intérêt de ces dernières dans l'étude des collinéarités des variables explicatives d'une régression.

En premier lieu, elles montrent clairement que les collinéarités ne dépendent pas de la taille des coefficients de corrélation par rapport à ± 1, mais par rapport aux bornes de leur intervalles de variation, et que des valeurs très proches de 0 peuvent conduire à de telles situations.

En second lieu, elles permettent de déterminer les variables responsables des collinéarités et de diminuer les variances des estimateurs en modifiant légèrement les coefficients de corrélation concernés.

Bibliographie

[1] BELSLEY D.A., KUH E., WELSH R.E. (1980) Regression diagnostics : identifying influential data and sources of collinearity. Wiley, New York.

[2] CIARLET P.G. (1989) Introduction to Numerical Linear Algebra and Optimisation, Cambridge University Press, London.

[3] CAZES P. (1991) Méthodes de régression, cours de D.E.A., Université Paris IX Dauphine, Paris.

[4] FOUCART T. (1991) Transitivité du produit scalaire, Revue de Statistique Appliquée, vol XXXIX n°3, p.57-68.

[5] HAWKINS D.M., EPLETT W.J.R. (1982) The Cholesky Factorization of the Inverse Correlation or Covariance Matrix in Multiple Regression, Technometrics, vol 24 n°3, p.191-198.

[6] HOERL A.E., KENNARD R.W. (1970) Ridge Regression : application to non orthogonal problems, Technometrics, vol 12 n°2, p.69-82.

[7] TOMASSONE R., LESQUOY E., MILLIER R. (1983) La régression, nouveaux regards sur une ancienne méthode statistique, Masson, Paris.