Évaluation de la régression bornée

Thierry Foucart

UMR 6086, Université de Poitiers, S P 2 M I, bd 3 téléport 2
BP 179, 86960 Futuroscope, Cedex FRANCE

Résumé. Le modèle linéaire est très fréquemment utilisé en statistique et particulièrement dans les secteurs de l'assurance, de la banque et du marketing. Il permet de déterminer les variables explicatives qui interviennent dans le risque mesuré chez les assurés et dans les choix effectués par la clientèle. Le problème considéré dans cet article apparaît lorsque ces variables sont liées statistiquement, par exemple le revenu et la catégorie socioprofessionnelle. Les estimations données par le critère des moindres carrés ordinaires deviennent alors instables et peuvent prendre des valeurs en contradiction avec les valeurs réelles. Il existe de nombreuses méthodes adaptées à ce type de données. Nous proposons ici d'évaluer l'efficacité de la régression bornée en procédant par simulation. Les résultats sont clairs : le gain en précision et en stabilité des coefficients de régression est impressionnant.

1 Introduction

Le modèle linéaire est une des méthodes statistiques les plus employées dans les sciences de l'homme et de la société. Il donne en effet une réponse à la question récurrente de l'effet propre d'une variable sur une autre. En assurance automobile par exemple, la question pourrait être : l'âge du conducteur joue-t-il un rôle dans le risque d'accident indépendamment des autres facteurs ? Ce risque dépend-il de son sexe toutes choses égales par ailleurs ? Pour répondre à ces questions, on effectue la régression du risque par les facteurs d'accident, et on étudie chacun des coefficients de régression de l'âge et du sexe : la réponse est considérée comme positive si ce coefficient est significativement non nul.
L'hypothèse « toutes choses égales par ailleurs », formalisée par le choix des variables explicatives du modèle linéaire, est toutefois très discutée depuis fort longtemps parce qu'elle ouvre la porte à des abus flagrants (Simiand, 1932). Sa formalisation demande beaucoup de précautions pour éviter des contradictions internes. Ces dernières se manifestent au plan mathématique par une relation linéaire exacte entre les variables explicatives. Dans ce dernier cas, l'analyse statistique est impossible, la matrice de corrélation n'étant pas inversible.
Ces contradictions ne sont pas toujours totales. Il existe des situations dans lesquelles les variables explicatives ne sont pas liées au sens linéaire du terme (il n'existe pas de combinaison linéaire strictement égale à 0), mais le sont au sens statistique (il existe une combinaison linéaire « presque »égale à 0). L'estimateur des moindres carrés ordinaires devient alors peu précis, et on est amené à utiliser d'autres estimateurs dont les plus classiques sont ceux de la

régression orthogonale (les variables explicatives sont choisies parmi les composantes principales de variance suffisante) et de la régression bornée que nous étudions ci-dessous.
Après avoir défini l'estimateur de la régression bornée (ou ridge regression) introduite par Hoerl et Kennard en 1970, nous donnons quelques exemples des conséquences de la colinéarité statistique entre les variables explicatives d'un modèle linéaire. Pour pouvoir comparer les estimations aux valeurs réelles des coefficients de régression, nous avons procédé par simulation. La généralisation de cette procédure donne un échantillon de l'estimateur borné et par suite une estimation de l'erreur quadratique moyenne. La comparaison de cette estimation avec l'erreur quadratique de l'estimateur des moindres carrés ordinaires montre clairement que la régression bornée améliore considérablement les estimations lorsqu'il existe une colinéarité statistique entre les variables explicatives.
Les données analysées dans cet article et les logiciels utilisés sont disponibles à l'adresse suivante : `http ://foucart.thierry.free.fr` (rubrique colinéarité).

2 Colinéarité et modèle linéaire.

2.1 Estimateurs dans le modèle linéaire

Le modèle linéaire consiste à représenter une relation entre une variable expliquée notée Y et p variables explicatives $X_1, \ldots, X_p$ par l'équation ci-dessous

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon,$$

dans laquelle

1. les coefficients $\beta_j, j = 1, \ldots, p$ sont des paramètres théoriques appelés coefficients de régression ;
2. la variable ε est une variable aléatoire appelée variable résiduelle, centrée et de variance σ^2 appelée variance résiduelle, indépendante des variables explicatives ;
3. on suppose fréquemment que la variable ε suit la loi normale $N(0, \sigma)$.

Dans toute la suite du texte, les variables explicatives sont centrées et réduites et la variable résiduelle suit la loi normale.
On considère un échantillon de taille n du vecteur $(Y, X_1, \ldots, X_p)$. On note $\mathbf{X}$ la matrice de n lignes et p colonnes contenant les observations $x_{i,j} (i = 1, \ldots, n$ et $j = 1, \ldots, p)$ des variables X_j et Y la matrice colonne contenant les observations $y_i (i = 1, \ldots, n)$ de la variable Y. La matrice $\mathbf{R}$ définie ci-dessous est la matrice des corrélations observées :

$$\mathbf{R} = \frac{1}{n} \mathbf{X}^t \mathbf{X}.$$

L'estimateur $\mathbf{B}$ des moindres carrés ordinaires du vecteur $\beta = (\beta_1, \ldots, b_p)^t$ est égal à :

$$\mathbf{B} = \frac{1}{n} \mathbf{R}^{-1} \mathbf{X}^t \mathbf{X}$$

C'est un estimateur efficace (de variance minimale dans la classe des estimateurs sans biais). On note $\mathbf{b} = (b_1, \ldots, b_p)^t$ l'observation du vecteur $\mathbf{B}$ et b_0 l'estimation de β_0 déduite des

moyennes observées des variables Y et $X_j, j = 1, \ldots, p$. La matrice variance $\mathbf{V_B}$ de l'estimateur $\mathbf{B}$ est égale à :

$$\mathbf{V_B} = \frac{\sigma^2}{n}\mathbf{R}^{-1}.$$

Le coefficient de détermination noté R^2 est le carré du coefficient de corrélation entre les valeurs observées $y_i, i = 1, \ldots, n$ et les valeurs y_i' estimées par le modèle :

$$y_i' = b_0 + b_1 x_{i,1} + \ldots + b_j x_{i,j} + \ldots + b_p x_{i,p}.$$

Les résidus e_i $(i = 1, \ldots, n)$ sont les différences entre les valeurs observées de Y et les valeurs estimées :

$$\forall i = 1, \ldots, n \quad \epsilon_i = y_i - y_i'$$

On sait que les résidus sont centrés et non corrélés aux variables explicatives. L'estimation sans biais de la variance résiduelle σ^2 est donnée par :

$$s'^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\epsilon_i^2.$$

2.2 Colinéarité et estimateur borné

Les effets de la colinéarité entre les variables explicatives résultent de l'inversion de la matrice $\mathbf{R}$ dans le calcul de l'estimateur $\mathbf{B}$ et de sa matrice variance $\mathbf{V_B}$. La colinéarité crée tout d'abord une grande instabilité des estimations des coefficients de régression : les variances des estimateurs, proportionnelles aux termes diagonaux de $\mathbf{R^{-1}}$, sont particulièrement élevées. Les signes des coefficients estimés peuvent même être contraires à ceux des vraies valeurs. Le coefficient de détermination R^2 peut aussi devenir très instable (Foucart, 2000). L'interprétation des résultats est finalement sujette à caution. Il n'est pas toujours facile de détecter cette colinéarité par une simple lecture de la matrice $\mathbf{R}$. En effet, cette colinéarité apparaît lorsqu'un coefficient de corrélation est proche d'une des bornes de l'intervalle dans lequel il peut varier conditionnellement aux autres (Foucart, 1997). On la recherche en examinant les valeurs propres de la matrice $\mathbf{R}$ et des indices comme les facteurs d'inflation (termes diagonaux de la matrice $\mathbf{R^{-1}}$) , l'indice de conditionnement (inverse de la plus petite valeur propre de la matrice $\mathbf{R}$) et l'indice de multicolinéarité (moyenne des facteurs d'inflation) : une petite valeur propre et des indices élevés indiquent une colinéarité statistique. On pourra sur ces points consulter l'ouvrage de Tomassone «La régression »(Masson, 1992). Les procédures de simulation utilisée ci-dessous visualisent dans un premier temps ces propriétés connues au plan mathématique. Elles montrent aussi que la régression bornée (ou ridge regression) proposée par Hoerl et Kennard (1970) donne de bien meilleurs résultats que la régression des moindres carrés ordinaires dans le cas de données statistiquement colinéaires.

L'idée générale est la suivante : un estimateur efficace X d'un paramètre réel m est un estimateur de variance minimale dans la classe des estimateurs sans biais $(E(X) = \mu)$, mais il ne minimise pas l'erreur quadratique définie par $E(\|X' - \mu\|^2)$, dans laquelle X' est un estimateur quelconque de μ. On peut donc chercher un estimateur biaisé X' dont l'erreur quadratique

est plus petite.

Cette situation se présente dans le cas du modèle linéaire lorsque les variables explicatives sont statistiquement colinéaires. Nous allons vérifier par simulation que l'estimateur $\mathbf{B_r} = (B_{r1}, \ldots, B_{rp})^t$ défini dans la régression bornée donne alors de meilleures estimations que l'estimateur efficace $\mathbf{B} = (B_1, \ldots, B_p)^t$ des moindres carrés ordinaires. L'estimateur $\mathbf{B_r}$ est obtenu suivant le critère des moindres carrés sous contrainte de norme :

$$\|\mathbf{B_r}\|^2 = b_{r1}^2 + b_{r2}^2 + \ldots + b_{rp}^2 \leq M$$

En fait, ce n'est pas le majorant M que l'on fixe : on montre en effet qu'il suffit de remplacer dans la formule de l'estimateur des moindres carrés ordinaires la matrice $\mathbf{R}$ par la matrice $\mathbf{R}+k\mathbf{I}$, où $\mathbf{I}$ est la matrice identique et k une constante réelle positive, pour limiter la norme de l'estimateur $\mathbf{B_r}$. En pratique, on recherche la meilleure constante k à l'aide de la représentation graphique des coefficients de régression en fonction de k, appelée ridge trace. Les formules concernant l'estimateur borné sont les suivantes

$$\mathbf{B}_r = \frac{1}{n}[\mathbf{R} + kI]^{-1}\mathbf{X}^t\mathbf{Y} \quad \mathbf{V}_{\mathbf{B_r}} = \frac{\sigma^2}{n}[\mathbf{R} + k\mathbf{I}]^{-1}\mathbf{R}[\mathbf{R} + k\mathbf{I}]^{-1}$$

3 Effets de la colinéarité. Exemples.

3.1 Colinéarité et matrice de corrélation

On considère quatre variables X_1, X_2, X_4 et X_4 dont les coefficients de corrélation sont donnés ci-dessous La colinéarité entre les variables ne peut guère être décelée par un simple

	X_1	X_2	X_3	X_4
X_1	1.000			
X_2	0.500	1.000		
X_3	0.500	0.500	1.000	
X_4	-0.500	0.400	0.300	1.000

TAB. 1 – *Matrice de corrélation des variables explicatives.*

examen de la matrice. Elle peut être mise en évidence de plusieurs façons :

1. Les facteurs d'inflation f_j associés à chaque coefficient de régression sont élevés ($f_1 = 62,\ f_2 = 26,\ f_3 = 14,\ f_4 = 50$) ;
2. l'indice de multicolinéarité I, égal à 1 en l'absence de toute colinéarité, est élevé : $I = 38$;
3. la matrice $\mathbf{R}$ possède une valeur propre très faible ($\lambda_4 = 0.007$). La combinaison linéaire des variables X_1, X_2, X_3 et X_4 définie par la quatrième composante principale est donc presque constante et égale à 0 : on ne peut pas choisir une valeur de X_4 en toute liberté lorsque les trois autres sont fixées ;
4. On utilise souvent l'indice de conditionnement, dont on trouvera une analyse dans Belsley (1980) : $\kappa = 1/\lambda_4 = 148.83$.

3.2 Effet de la colinéarité sur les coefficients de régression estimés

Les données ridge1 contiennent les observations de cinq variables sur cent individus statistiques obtenues par simulation. On veut expliquer la cinquième variable, Y, par les quatre premières X_1, X_2, X_3 et X_4, dont la matrice de corrélation est égale à la précédente. Les variables explicatives sont centrées et réduites. Les coefficients de corrélation observés entre les variables explicatives et la variable expliquée sont donnés dans le tableau ci-dessous

	X_1	X_2	X_3	X_4
y	0.540	0.216	-0.107	-0.491

TAB. 2 – *Coefficients de corrélation observés entre Y et X_1, X_2, X_3, X_4, (données ridge1).*

La régression des moindres carrés ordinaires donne les résultats suivants :

degré de liberté	Somme des carrés	Variance estimée	Pourcentage de variance totale
Tot 99	229.7305	2.320510	1
Exp 4	112.7185	1.088805	0.490655
Res 95	117.0120	1.231705	0.509345

TAB. 3 – *Analyse de variance (données ridge1, n = 100).*

	Estimation	écart-type	t de Student	facteur d'inflation
b_1	1.6339	0.8739	1.870	62.00
b_2	-0.1482	0.5659	-0.262	26.00
b_3	-1.0375	0.4153	-2.498	14.00
b_4	0.4439	0.7848	0.566	50.00
b_0	-0.1650	0.1110	-1.486	

TAB. 4 – *Estimation des coefficients de régression (données ridge1).*

Les variables explicatives étant réduites et l'écart-type de la variable expliquée égal à 1.516, on peut apprécier intuitivement la taille des coefficients de régression. Les coefficients de régression b_1, b_3 prennent des valeurs élevées en valeur absolue. Le coefficient de régression b_2 est négatif, malgré un coefficient de corrélation positif entre X_2 et $Y (0.216)$, et b_4 est positif malgré un coefficient de corrélation entre X_4 et Y fortement négatif (-0.491). Seul b_3 est significativement non nul pour un risque de première espèce $\alpha = 5\%$ $(t = -2.498)$. Le coefficient de détermination $(R^2 = 0.49)$ est hautement significatif. Ces résultats peuvent s'expliquer par la forte colinéarité statistique entre X_1, X_2, X_3, et X_4 détectée en paragraphe 3.1.

3.3 Effet de la colinéarité sur les variances

On étudie maintenant les données ridge2, obtenues par simulation suivant le même modèle que précédemment. La matrice de corrélation entre les variables explicatives reste égale à $\mathbf{R}$, les coefficients de régression théoriques sont les mêmes, mais les corrélations observées entre la variable expliquée et les variables explicatives sont les suivantes :

	X_1	X_2	X_3	X_4
y	0.486	0.084	-0.199	-0.584

TAB. 5 – *Coefficients de corrélation observés entre Y et X_1, X_2, X_3, X_4, (données ridge 2).*

La régression des moindres carrés ordinaires donne les résultats suivants

degré de liberté	Somme des carrés	Variance estimée	Pourcentage de variance totale
Tot 99	188.1299	1.900302	1
Exp 4	94.14017	0.9109364	0.500400
Res 95	93.98971	0.9893653	0.499600

TAB. 6 – *Analyse de variance (données ridge 2, n = 100).*

	Estimation	écart-type	t de Student	facteur d'inflation
b_1	0.4638	0.7832	0.592	62.00
b_2	0.3674	0.5072	0.724	26.00
b_3	-0.5204	0.3722	-1.398	14.00
b_4	-0.5594	0.7033	-0.795	50.00
Cst	-0.0985	0.0995	-0.990	

TAB. 7 – *Estimation des coefficients de régression (données ridge 2).*

La situation est paradoxale : le coefficient de détermination R^2 est hautement significatif ($R^2 = 0.50, n = 100$), mais aucun des coefficients de régression n'est significativement non nul. On peut apporter comme explication une surestimation des écarts-types des estimateurs B_j. Les estimations b_1, b_2, b_3 et b_4 ne paraissent pas en effet spécialement grandes (les variables explicatives sont réduites, et l'écart-type de la variable expliquée est égal à 1.372), et l'augmentation des variances des estimateurs due à la colinéarité a pour effet de diminuer les t de Student, les rendant ainsi non significatifs.

4 Simulation d'un échantillon de l'estimateur borné

4.1 Démarche

On peut, par simulation, visualiser de façon plus complète l'effet de la colinéarité sur les coefficients de régression. La démarche est la suivante

1. on choisit le nombre de variables explicatives p, le vecteur de régression théorique $\beta = (\beta_1, \ldots, \beta_p)^t$, le coefficient constant β_0, le coefficient de détermination R^2 et le nombre d'observations n. On fixe la matrice de corrélation $\mathbf{R}$ entre les variables explicatives. On en déduit la variance résiduelle théorique σ^2 ;
2. on simule un échantillon des variables explicatives $x_{i,j}$ de matrice de corrélation égale à $\mathbf{R}$. Il suffit pour cela de simuler un échantillon de taille n d'un vecteur aléatoire quelconque de dimension p, d'en effectuer l'analyse en composantes principales pour obtenir un vecteur Z dont la matrice de covariance est strictement égale à la matrice identique

$\mathbf{I}$, et d'effectuer une transformation linéaire de ce dernier de façon à obtenir un tableau de données X dont la matrice de corrélation est strictement égale à $\mathbf{R}$. Cette transformation est définie par la matrice triangulaire inférieure $\mathbf{T}$ obtenue par l'algorithme de Cholesky(Ciarlet, 1989)

$$\mathbf{R} = \mathbf{T}\mathbf{T}^t \qquad \mathbf{X} = \mathbf{Z}\mathbf{T}^t;$$

3. on simule un échantillon indépendant (ε_i) de taille n de la v.a. ε suivant la loi normale $N(0, \sigma)$, et on en déduit les valeurs simulées $y_i, i = 1, \ldots, n$ de la variable Y pour les valeurs $x_{i,j}$ du tableau $\mathbf{X}$

$$\forall i = 1, \ldots, n \quad y_i = b_0 + b_1 x_{i,1} + ... + b_p x_{i,p} + \varepsilon_i;$$

4. on calcule le vecteur de régression estimé $\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_p})^\mathbf{t}$;
5. on recommence la simulation effectuée en 3) pour obtenir une autre simulation du vecteur de régression avec le même tableau $\mathbf{X}$, etc.

Chaque échantillon de la variable expliquée donne une estimation $\mathbf{b}$ du vecteur de régression β pour les mêmes valeurs des variables explicatives. On en déduit l'erreur quadratique $\|\mathbf{b} - \beta\|^2$. En répétant m fois cette opération, on dispose donc d'un échantillon de m vecteurs $\mathbf{b}_l, l = 1, \ldots, m$ conditionnellement à $\mathbf{X}$. On peut calculer les erreurs quadratiques $d_l = \|\mathbf{b}_l - \beta\|$ pour $l = 1, \ldots, m$ en choisissant comme estimateur l'estimateur de la régression bornée $\mathbf{B_r}$ pour différentes valeurs de k (pour $k = 0$, l'estimateur de la régression bornée est confondu avec l'estimateur des moindres carrés ordinaires). Certains auteurs donnent des indications sur le choix de cette constante k (Nordberg, 1982).

4.2 Exemple

Dans l'exemple ci-dessous, la matrice de corrélation $\mathbf{R}$ entre les variables explicatives est donnée dans le tableau 1. Les coefficients de régression et la constante choisis sont les suivants :

$\beta_0 = 0$	$\beta_1 = 0.5$	$\beta_2 = 0.5$	$\beta_3 = -0.5$	$\beta_4 = -0.5$

Le modèle théorique est donc égal à

$$Y = 0.5X_1 + 0.5X_2 - 0.5X_3 - 0.5X_4 + \varepsilon\ .$$

Le coefficient R^2 étant fixé à 0.5, la variance résiduelle théorique est égale à 0.95. On effectue ensuite la régression linéaire bornée des données simulées en effectuant la démarche précédente. Le coefficient constant estimé b_0 n'est pas nécessairement nul, mais n'est pas pris en compte dans le calcul des distances entre les vecteurs de régression.
La taille de l'échantillon étant fixée à $n = 100$, nous donnons ci-dessous les résultats de la régression sur un échantillon obtenus par l'estimateur des moindres carrés ordinaires $(k = 0)$ et par l'estimateur borné pour différentes valeurs de la constante k.

Coefficients théoriques	0.5	0.5	-0.5	-0.5	Carré de la distance
Coefficients estimés	br_1	br_2	br_3	br_4	d^2
k= 0 (MCO)	0.123	0.850	-0.445	-0.826	0.374
k = 0.01	0.356	0.688	-0.539	-0.611	0.070
k = 0.05	0.452	0.575	-0.543	-0.498	0.010

TAB. 8 – *coefficients de régression estimés pour différentes valeurs de k. En dernière colonne : carré de la distance au vecteur de régression théorique.*

4.3 Généralisation

Pour généraliser ces résultats, nous avons généré, pour les mêmes valeurs $x_{i,j}$ des variables explicatives, cinquante échantillons de la v.a. Y, puis, pour chaque valeur de k, effectué les cinquante régressions et calculé les carrés d_l^2 des distances, leur moyenne et leur variance :

k	moyenne	variance
0	1.383	4.005
0.01	0.242	0.107
0.05	0.045	0.002

TAB. 9 – *moyennes et variances des carrés des distances $d_l^2 \quad (l = 1, \ldots, 50)$ pour $k = 0, 0.01$ et 0.05.*

D'après le tableau précédent, la régression classique donne des estimations beaucoup plus éloignées en moyenne des coefficients de régression théoriques que la régression bornée. La variance des carrés des distances est très élevée par rapport aux autres variances, et le meilleur estimateur des trois précédents est celui de la régression bornée pour $k = 0.05$.
La fonction de répartition observée des carrés des distances du vecteur de régression estimé par le critère des moindres carrés ordinaires au vecteur de régression théorique est donnée en figure 1 ci-dessous. Elle met en évidence la fréquence de vecteurs de régression estimés par les moindres carrés ordinaires très différents du vecteur théorique. La valeur maximale des carrés des distances obtenues par l'estimateur borné et calculées sur les cinquante échantillons en posant $k = 0.05$ est égale à 0.208 : dans plus de 60% des cas, la régression des moindres carrés ordinaires donne un estimation moins bonne du vecteur théorique que la pire donnée par la régression bornée.
Le tableau 10 contient les dix vecteurs de régression obtenus par les moindres carrés ordinaires les plus éloignés du vecteur théorique. Le coefficient de détermination R^2 et le coefficient constant b_0 sont à peu près correctement estimés. Ce n'est pas le cas des coefficients de régression, très mal reconstruits. Dans tous les vecteurs de régression estimés, un au moins des coefficients est de signe contraire au coefficient théorique et certains autres sont très élevés en valeur absolue. Dans la pratique, le risque d'obtenir ce genre de résultats est loin d'être négligeable, puisque ces échantillons représentent 20% du nombre total d'échantillons. L'idée qui vient naturellement est de déterminer sur ces cinquante échantillons la valeur de k qui donne les meilleurs résultats. En faisant varier k dans l'intervalle $[0, 1]$ avec un incrément de 0.001, on obtient $k = 0.078$ (tableau 11) :
La moyenne des carrés des distances est très faible par rapport à celle que l'on obtient par les estimateurs des moindres carrés ordinaires (0.039 au lieu de 1.383), et ces distances varient

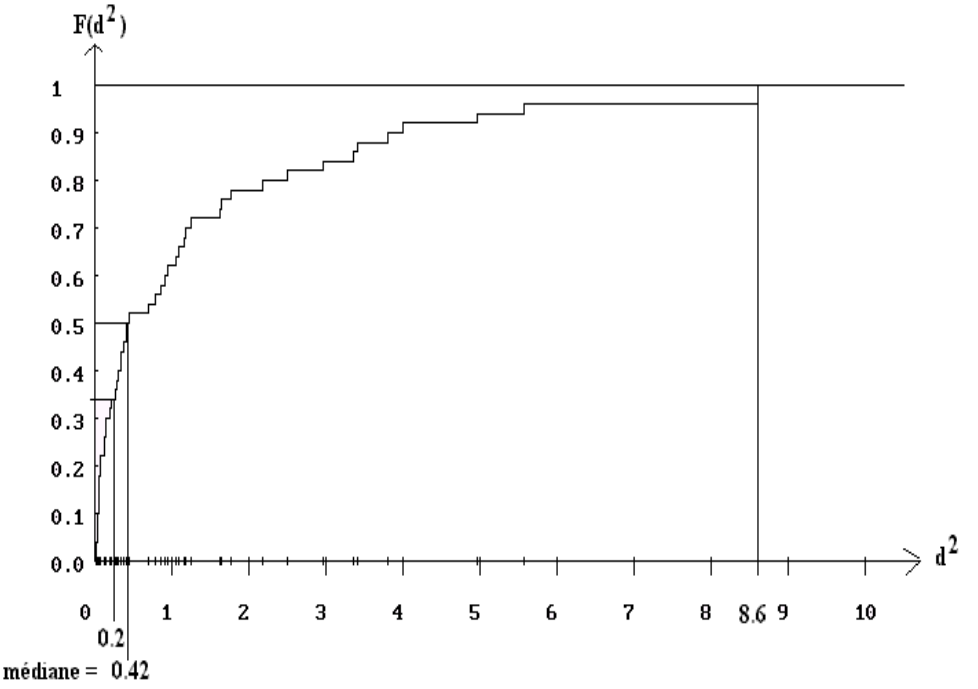

FIG. 1 – *fonction de répartition des carrés des distances (régression des moindres carrés ordinaires, échantillon simulé de* 50 *termes).*

	R^2	b_0	b_1	b_2	b_3	b_4	d^2
valeurs théoriques	0.500	0.000	0.500	0.500	-0.500	-0.500	
n° 50	0.597	-0.105	-0.444	1.135	-0.041	-1.495	2.494
n° 5	0.416	-0.231	-0.643	1.089	0.190	-1.413	2.963
n° 44	0.534	-0.020	-0.611	1.166	0.010	-1.690	3.353
n° 32	0.402	0.095	-0.841	1.248	-0.042	-1.413	3.400
n° 46	0.594	-0.081	-0.667	1.348	-0.045	-1.728	3.795
n° 47	0.424	0.117	1.693	-0.341	-1.118	0.714	3.986
n° 25	0.560	0.189	1.898	-0.336	-1.233	0.828	4.956
n° 42	0.589	0.071	-1.028	1.488	0.008	-1.910	5.556
n° 40	0.626	0.110	2.474	-0.642	-1.421	1.097	8.600
n° 23	0.407	0.119	-1.523	1.488	0.448	-2.124	8.605

TAB. 10 – *les dix vecteurs de régression les plus éloignés du vecteur théorique (régression des moindres carrés ordinaires)*

beaucoup moins : l'intérêt de la régression bornée est évident et considérable. La simulation montre aussi la robustesse de la méthode : la moyenne des carrés des distances diminue très rapidement lorsque k varie de 0 à 0.05, reste à peu près constante lorsque k varie de 0.05 à 0.1 environ, et augmente ensuite lentement à partir de 0.1 (cf. figure 2 ci-dessous). La variance des

k	moyenne	variance
0.078	0.039	0.001

TAB. 11 – *moyenne et variance des carrés des distances d_k^2 pour la valeur optimale $k = 0.078$ $(m = 50)$.*

distances, minimale aussi pour $k = 0.078$, suit la même évolution. La recherche précise de la meilleure valeur de la constante k ne présente visiblement guère d'intérêt. Nous avons procédé à plusieurs simulations identiques : les résultats ont toujours été analogues aux précédents.

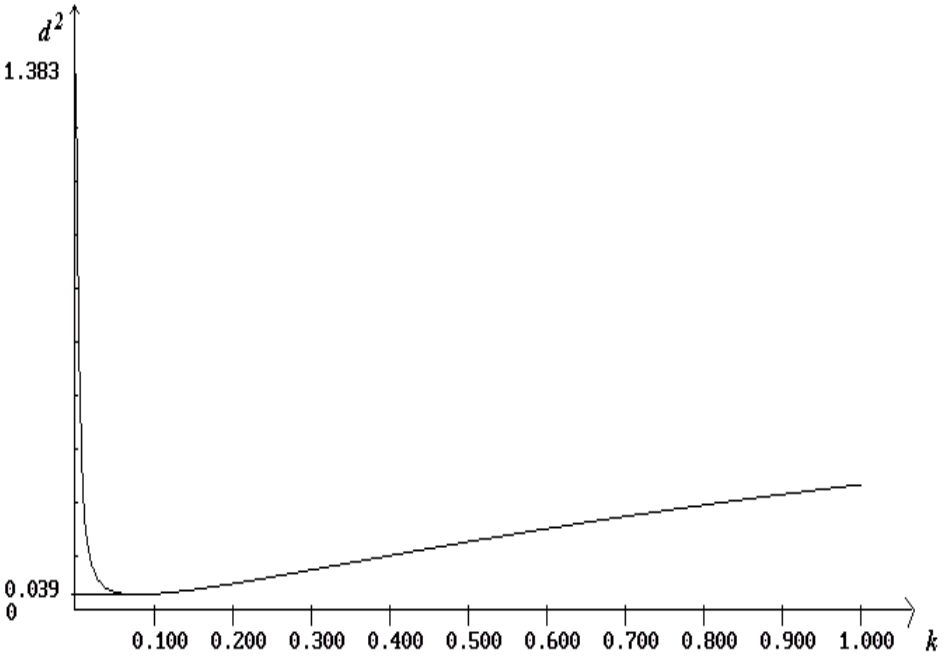

FIG. 2 – *moyenne des carrés des distances entre le vecteur théorique et le vecteur estimé en fonction de k.*

5 Applications

Dans le cas de données réelles quelconques, on ne connaît ni le vecteur de régression théorique, ni la valeur optimale de la constante k. Pour choisir k, on utilise les ridge traces : l'absence de colinéarité se traduisant par une ridge trace très régulière, on va choisir comme valeur celle pour laquelle les coefficients de régression sont stabilisés.

5.1 Régression bornée en l'absence de colinéarité (données ridge0)

Effectuons d'abord la régression bornée dans le cas où les variables explicatives ne sont pas colinéaires. Les données analysées ci-dessous (fichier ridge 0) ont été obtenues par simulation en supposant que les variables explicatives sont non corrélées. Les coefficients de régression et le coefficient de détermination théoriques sont les mêmes que ceux choisis précédemment pour créer les données ridge1 et ridge 2.

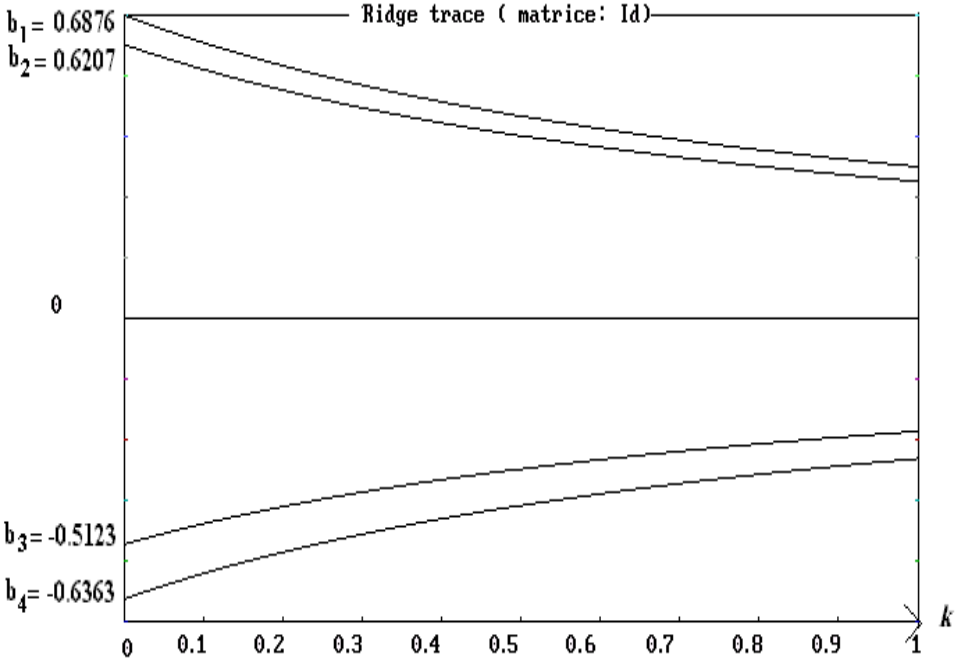

FIG. 3 – *ridge trace (données ridge 0).*

La figure 3 ci-dessus est la ridge trace obtenue en effectuant les régressions bornées pour des valeurs de k variant de 0 à 1. On observe une très grande stabilité des coefficients de régression par rapport à la constante k. Le choix de cette dernière n'intervient guère dans les estimations.

5.2 Régression bornée des données ridge 1

Revenons au données traitées dans le paragraphe 2.1. La figure 4 ci-dessous donne la représentation graphique des estimations b_1, b_2, b_3 et b_4 des coefficients de régression suivant les valeurs de k. Pour $k = 0$, ces valeurs sont celles que l'on obtient par la régression des moindres carrés ordinaires.
On observe l'instabilité de ces coefficients pour les faibles valeurs de k. Les coefficients de régression b_1 et b_3 diminuent très rapidement en valeur absolue, au contraire de b_2 et b_4. On

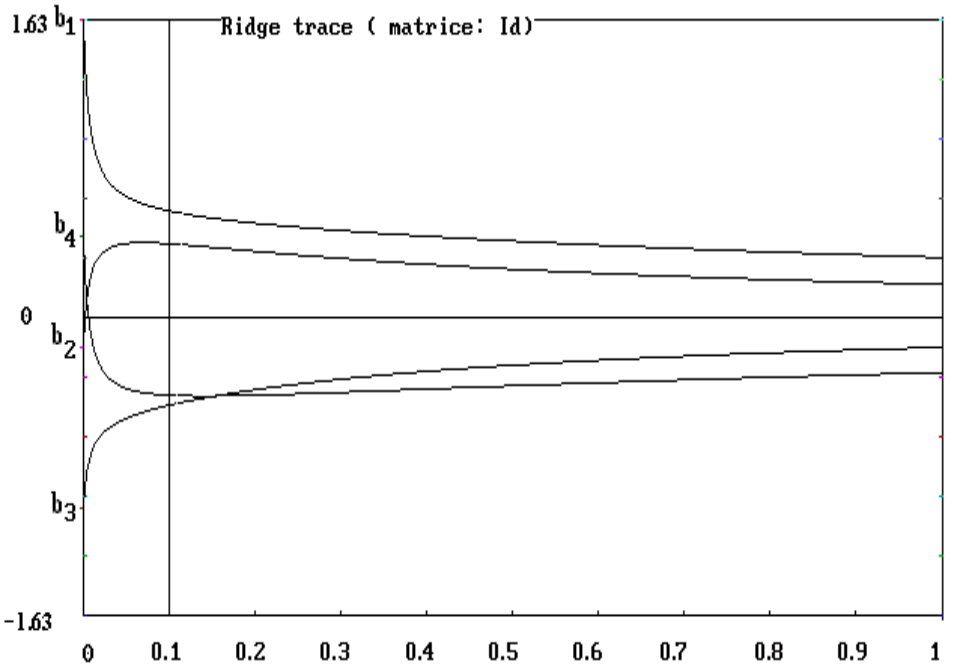

FIG. 4 – *ridge trace (données ridge 1).*

recherche sur ce graphique une valeur de k pour laquelle les coefficients de régression sont stabilisés : on peut prendre ici $k = 0.1$. La régression bornée donne alors les résultats suivants :

	br_1	br_2	br_3	br_4	br_0
Estimation	0.585141	0.405312	-0.480752	-0.426282	-0.164952
écart-type	0.05550	0.06988	0.07351	0.05690	
t de Student	6.955	3.826	-4.315	-4.943	

TAB. 12 – *Résultats de la régression bornée (données ridge1, k = 0.1).*

Le biais de l'estimateur $\mathbf{B}_r$ apparaît dans le coefficient de corrélation non nul entre les résidus et la variable expliquée estimée par le modèle : $r = 0.088$. Les écarts-types indiquent une très grande stabilité de $\mathbf{B}_r$ autour de son espérance $E(\mathbf{B}_r)$, et les t de Student montrent que tous les coefficients sont significatifs. Les simulations précédentes montrent que les coefficients obtenus sont largement plus proches des vraies valeurs que les estimations données par le critère des moindres carrés ordinaires.

5.3 Régression bornée des données ridge2

La ridge trace (figure 5) montre une bonne stabilité des coefficients de régression, et les effets de la colinéarité concernent donc surtout les variances des estimateurs.

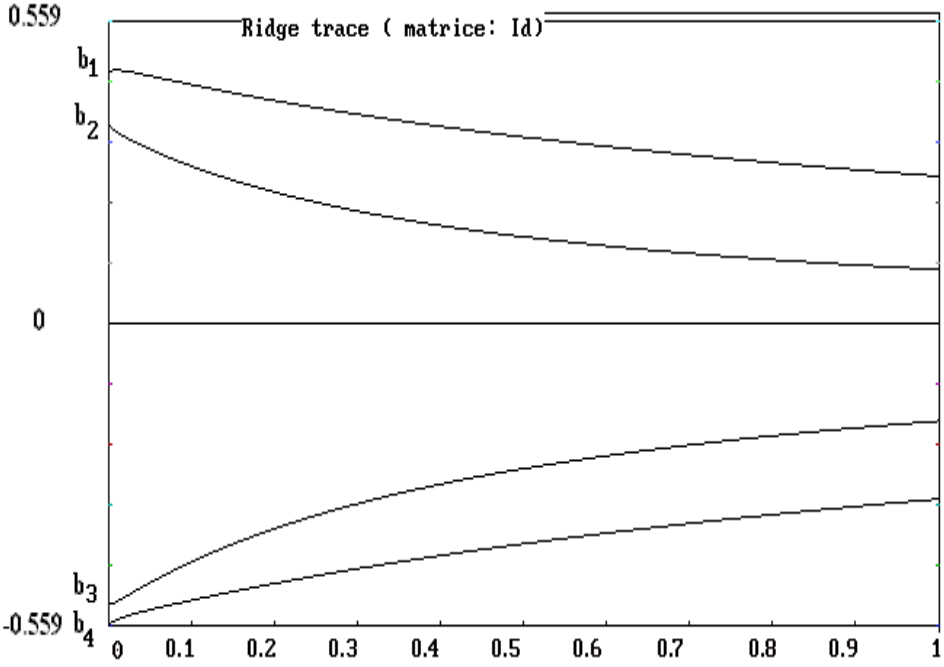

FIG. 5 – *ridge trace (données ridge 2).*

On peut le vérifier en choisissant une petite valeur de k, par exemple, $k = 0.01$ ou $k = 0.02$. Les résultats pour chacune de ces deux valeurs sont données dans les tableaux ci-dessous

	br_1	br_2	br_3	br_4	br_0
Estimation	0.468614	0.353795	-0.514348	-0.547913	-0.098518
écart-type	0.23287	0.16324	0.13169	0.21075	
t de Student	1.467	1.580	-2.848	-1.895	

TAB. 13 – *Résultats de la régression bornée pour* $k = 0.01$ *(données ridge 2).*

Il y a trés peu de différences entre les estimations suivant les valeurs de k, mais les valeurs sont bien plus stables pour $k = 0.02$. La colinéarité entre les variables explicatives exerce ici un effet sur les seules variances, et les estimations obtenues suivant le critère des moindres

	br_1	br_2	br_3	br_4	br_0
Estimation	0.466987	0.344683	-0.505688	-0.542312	-0.098518
écart-type	0.14950	0.11658	0.10273	0.13704	
t de Student	2.277	2.156	-3.589	-2.885	

TAB. 14 – *Résultats de la régression bornée pour $k = 0.02$ (données ridge 2)*

carrés ordinaires sont beaucoup plus proches des valeurs théoriques que les écarts-types des estimateurs ne l'indiquent.

6 Conclusion

Les coefficients de régression théoriques de ces applications sont en réalité connus : le modèle utilisé pour créer les données ridge1 et ridge 2 est celui qui a été précisé dans le paragraphe 3.2. Le coefficient de détermination est fixé à 0.5, et les coefficients de régression théoriques sont :

β_0 = 0	β_1 = 0.5	β_2 = 0.5	β_3 = -0.5	β_4 = -0.5

Les estimations obtenues dans le premier cas par la régression bornée (données ridge 1) sont beaucoup plus proches des valeurs théoriques que celles qui sont déduites du critère des moindres carrés ordinaires. Dans le second (données ridge 2), elles sont beaucoup plus stables. Compte tenu des résultats des simulations donnés dans le paragraphe 3, on pouvait s'y attendre. L'intérêt de la régression bornée apparaît finalement sur trois points :

1. Lorsque les coefficients estimés par le critère des moindres carrés ordinaires sont très différents des vraies valeurs, elle donne des estimations bien meilleures ;
2. elle permet de contrôler la stabilité des estimations ;
3. lorsque les variables explicatives sont non corrélées, elle ne modifie quasiment pas les estimations ;
4. La stabilité des résultats par rapport à la constante k limite l'importance d'en rechercher la meilleure valeur : une approximation même grossière, déduite simplement de la ridge trace, donnera des résultats en moyenne bien meilleurs que la régression des moindres carrés ordinaires. Par suite, en effectuant systématiquement une régression bornée pour une faible valeur de la constante k (par exemple $k = 0.01$) , les estimations des coefficients de régression ne peuvent être que meilleures, même lorsque la colinéarité entre les variables explicatives n'est pas très forte.

Toutefois, lorsque les valeurs théoriques des coefficients de régression sont elles-mêmes élevées en valeur absolue, la régression bornée est à éviter. C'est le cas par exemple lorsque les coefficients de régression théoriques sont égaux aux valeurs estimées sur les données ridge 1 : la ridge trace est la même, et par suite la régression bornée donne de très mauvais résultats. Il est donc indispensable d'analyser a priori la taille des coefficients de régression en suivant une démarche critique.
Ces simulations montrent le danger d'interpréter le signe des coefficients de régression sans

précaution. Même lorsque toutes les hypothèses mathématiques sont satisfaites (distribution gaussienne de la variable résiduelle, linéarité des liaisons) ce qui est le cas dans les exemples donnés puisqu'ils sont construits à partir de ces hypothèses, il est très possible d'obtenir des estimations très différentes des valeurs théoriques. Lorsque ces hypothèses ne sont qu'approximativement vérifiées, ce qui est le cas général des données réelles, la statistique produit des résultats qu'il est indispensable d'examiner avec prudence et de ne pas prendre pour certains même s'ils sont largement significatifs.

Références

Belsley D.A., Kuh E., Welsh R.E. (1980). Regression diagnostics : identifying influential data and sources of collinearity. *Wiley, New York.*

Ciarlet P.G. 1989). Introduction to Numerical Linear Algebra and Optimisation, *London, Cambridge University Press.*

Foucart T. (1997). Numerical Analysis of a Correlation Matrix. *Statistics,* 29/4, p. 347-361.

Foucart T. (2000). Colinéarité et Instabilité Numérique dans le Modèle Linéaire, *RAIRO Operations research*, Vol.34, 2, p. 199-212.

Hoerl A.E., R.W. Kennard (19701). Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

Hoerl A.E., R.W. Kennard (19702). Ridge regression : Applications to nonorthogonal problems. *Technometrics*, 12, 69-82.

Nordberg L., 1982. A procedure of determination of a good ridge parameter in linear regression. *Commun, Statist. Simula. Computa.* 11(3), p. 285-289.

Simiand F.,1932. Le salaire, l'évolution sociale et la monnaie. *Lien internet*
`http ://www.uqac.uquebec.ca/zone30/Classiques_des_sciences_sociales/index.html`.

Tomassone R., Lesquoy E. et Millier C. (1992) : La régression. Nouveaux regards sur une ancienne méthode statistique, *Masson, Paris*, 2e ed. .

Summary

The linear model is very frequently used in statistics and particularly in insurance, bank and marketing. It makes it possible to determine the explanatory variables which play part in the risk measured in the policy-holders and in the choices carried out by the customers. The problem considered in this article appears when these variables are dependent statistically, for example the income and the socio-professional group. Then, the estimates given by the

criterion of ordinary least squares become not very reliable and can take values in contradiction with the real values. There are many methods adapted to this type of data. We propose here to evaluate the effectiveness of the ridge regression while proceeding by simulations. The results are clear: the gain in precision and in stability of the regression coefficients is impressive.