

STATISTIQUE ET ANALYSE DES DONNÉES

GILBERT RITSCHARD

Régression robuste et problèmes de colinéarité

Statistique et analyse des données, tome 15, n° 2 (1990), p. 77-96

http://www.numdam.org/item?id=SAD_1990__15_2_77_0

© Association pour la statistique et ses utilisations, 1990, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

REGRESSION ROBUSTE ET PROBLEMES DE COLINEARITE

Gilbert RITSCHARD

Département d'économétrie, Université de Genève
2 rue Dancet, CH-1211 Genève 4

Résumé

Cet article montre comment des données atypiques peuvent masquer ou induire des problèmes de colinéarité approchée dans le cadre de la régression linéaire. Il met ainsi en évidence les implications que peut entraîner, du point de vue de la colinéarité, la réduction de l'importance des données atypiques par les estimateurs robustes. L'article traite également du diagnostic des problèmes de colinéarité, en proposant une extension de l'indice de conditionnement (condition number) et des indices de colinéarité de Stewart (1987) au cas de la régression robuste.

Mots clés : colinéarité, données atypiques, régression robuste.

Classification AMS : 62 J 05

Classification STMA : 07.150

Abstract

This paper shows how atypical data can mask or induce near collinearity problems in linear regression. It thus pinpoints what can happen, from the collinearity viewpoint, when robust estimators automatically download atypical data. The paper also focuses on near collinearity diagnostics. The condition number and Stewart's (1987) collinearity indexes are extended to the case of robust estimation.

Keywords : collinearity, influential data, robust regression.

Manuscrit reçu le 2 juin 1988, révisé le 24 août 1990

1. INTRODUCTION

La statistique robuste, qui a connu un essor important ces dix ou quinze dernières années, est née du souci de se prémunir des effets pervers qui peuvent résulter de la présence, dans les données, d'observations atypiques, c'est-à-dire d'observations très dissemblables des autres.

Dans le contexte de la régression en particulier, toute une panoplie d'estimateurs robustes ont été développés à la suite notamment de l'étude de Huber (1973). Il s'agit essentiellement d'estimateurs de classe M ayant une sensibilité aux données atypiques bornée (voir par exemple Hampel et al., 1986, ch.6), mais également d'estimateurs fondés sur la notion de rang (Jureckova, 1971) ou encore d'estimateurs dits à haut point de rupture, dont le plus célèbre est sans doute l'estimateur du moindre résidu médian de Rousseeuw (Rousseeuw, 1984).

De façon générale, la robustesse s'obtient par le biais de mécanismes qui réduisent ou annulent automatiquement l'importance des données atypiques dans l'estimation. Comme on le conçoit aisément, négliger ainsi l'information fournie par certaines observations va, outre le gain en robustesse, entraîner certaines conséquences secondaires.

La connaissance de ces effets secondaires est évidemment essentielle lorsqu'il s'agit de choisir un estimateur robuste pour une application donnée. En faisant abstraction du coût de mise en oeuvre, ce choix se ramène en général à un arbitrage entre degré de robustesse et efficacité.

Dans des domaines tels que l'économétrie, où l'on se heurte fréquemment aux difficultés de la colinéarité approchée, il nous paraît important de tenir compte également de l'impact que le recours à un estimateur robuste peut avoir sur cette colinéarité.

C'est à cet aspect, auquel la littérature sur la régression robuste n'a jusqu'ici accordé que peu d'attention, que nous consacrons cet article. Pour cette première approche du problème, nous nous référons essentiellement aux estimateurs robustes de classe M, même si les principes mis en évidence s'appliquent plus généralement à tout estimateur robuste.

Dans la section 2, nous montrons à l'aide de quelques cas de figure comment la présence d'une seule donnée atypique peut, selon les cas, engendrer ou masquer une situation de colinéarité approchée. Au travers d'une discussion des situations mises en évidence, nous montrons que le recours aux estimateurs robustes devrait, en règle générale, permettre d'éviter les problèmes de colinéarité accidentelle, (générée par une seule observation) et permettre de déceler les colinéarités structurelles masquées.

La section 3 traite des instruments de mesure de la colinéarité approchée. Dans une première sous-section, nous rappelons les indicateurs utilisés dans le cadre de la régression par les moindres carrés. Une seconde sous-section propose ensuite une extension de ces mesures au cas de l'estimation par des méthodes robustes. Ces extensions sont utilisées au paragraphe 4 pour illustrer, à l'aide de quelques exemples, le comportement des estimateurs robustes dans des situations de colinéarité accidentelle et/ou masquée.

Dans la section finale, nous discutons de la portée et des limites des indicateurs proposés, et concluons en évoquant quelques directions de recherches futures.

2. DONNEES ATYPIQUES ET COLINEARITE APPROCHEE

Dans le cadre de la régression linéaire, on distingue en général deux types non exclusifs de données atypiques : les observations aberrantes et les points levier (voir par exemple Ritschard, 1990). Cette distinction se justifie en particulier eu égard à la façon différente dont elles affectent les résultats de l'estimation : l'estimation des coefficients de régression, mais aussi de leurs écarts-types, et par suite les aides à l'évaluation que constituent les R^2 , F de Fisher-Snedecor, t de Student, etc.

Soit

$$y_i = x_i' \beta + \varepsilon \quad i = 1, 2, \dots, n$$

avec notamment $E(\varepsilon_i) = 0$, le modèle linéaire dont il s'agit d'estimer les coefficients $\beta \in R$ à partir des n observations $(x_i, y_i) \in X \times Y \subseteq R^k \times R$.

Les *données aberrantes* sont celles pour lesquelles la réponse y_i aux variables dépendantes x_i n'obéit manifestement pas au même modèle que la majorité des autres données. A défaut d'un traitement spécial, les données aberrantes vont exercer un effet d'attraction sur, par exemple, l'hyperplan des moindres carrés en raison d'un écart $r_i = y_i - x_i' \hat{\beta}$ important.

Les *points levier* correspondent aux observations qui sont très éloignées des autres dans l'espace X des variables explicatives. Ils exercent leurs effets d'attraction sur l'hyperplan des moindres carrés en raison de cette position atypique dans X .

Pour le problème qui nous occupe ici, à savoir la colinéarité approchée, nous définissons ci-après une troisième catégorie de données atypiques : les données à forte incidence sur la colinéarité (voir également Mason-Gunst, 1985, Hadi, 1987, et Hadi-Velleman, 1987).

Au préalable, il convient de rappeler que les problèmes de la colinéarité approchée apparaissent lorsqu'il existe une dépendance linéaire trop forte entre les variables explicatives, c'est-à-dire lorsque l'une au moins des variables x s'exprime presque comme une combinaison linéaire des autres.

Les difficultés d'estimation qui apparaissent dans ce cas peuvent être comparées à celles que l'on rencontre lorsqu'on cherche à placer un plateau de table (l'hyperplan de régression) sur des pieds presque alignés (les observations dans l'espace X . Un seul pied décalé par rapport aux autres suffit à assurer une certaine stabilité au plateau de table. Inversement, lorsqu'un pied se trouve relativement très éloigné des autres, le plateau sera instable, que ces autres pieds soient ou non alignés. Dans les deux cas, la position atypique d'un seul pied s'avère cruciale pour la stabilité, en l'assurant dans le premier cas et, au contraire, en générant l'instabilité dans le second.

Par analogie, on peut alors donner la définition suivante.

Définition 2.1 : Une observation ou un petit groupe d'observations est dit à *forte incidence sur la colinéarité*, si sa présence, ou sa suppression génère une situation de colinéarité approchée. On distingue :

- les données *source de colinéarité*, dont la présence est cause d'un problème de colinéarité;
- les données *masquant la colinéarité*, dont la suppression conduit à une situation de colinéarité approchée.

Les figures 2.1 et 2.2 qui représentent la position des données dans l'espace des variables explicatives, illustrent respectivement le cas d'un point source de colinéarité et d'un point masquant la colinéarité.

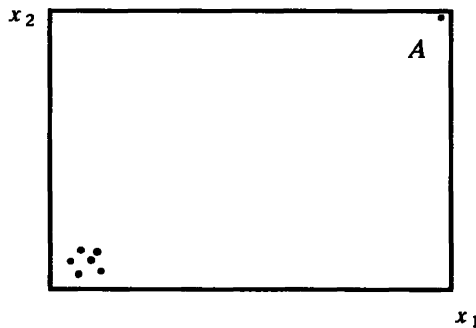


Figure 2.1 : Point A source de colinéarité

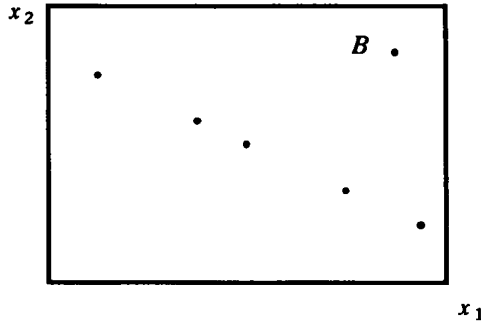


Figure 2.2 : Point B masquant la colinéarité

Remarquons que, comme l'illustre la figure 2.3, une même donnée peut être source d'une certaine dépendance linéaire et simultanément masquer une autre dépendance. Ainsi, paradoxalement, une donnée peut à la fois générer et masquer la colinéarité. En d'autres termes, une colinéarité peut en cacher une autre.

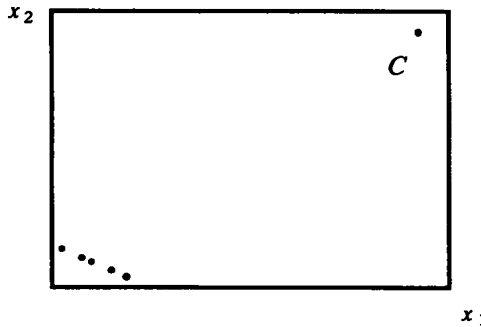


Figure 2.3 : Une colinéarité peut en cacher une autre

Les situations de colinéarité générées par une seule observation peuvent être considérées comme accidentelles. Celles qui, comme les colinéarités masquées, traduisent une dépendance plus profonde, seront, par opposition, qualifiées de structurelles.

Notons que seuls de forts points levier peuvent être source de colinéarité accidentelle, tandis qu'un point peut masquer une colinéarité structurelle sans être nécessairement très éloigné des autres observations.

Il découle notamment de la remarque précédente que les estimateurs robustes qui réduisent l'importance des points levier, permettront systématiquement d'éviter les problèmes dus à la colinéarité accidentelle.

Quant aux situations de colinéarité structurelle masquée, que l'utilisation des moindres carrés simples ne permet évidemment pas de détecter, elles pourront affecter la stabilité des estimations fournies par les estimateurs robustes. En effet, même s'il est en règle générale moins fort que celui des points source de colinéarité, l'effet levier des données masquant la colinéarité doit tout de même être suffisamment important pour que l'effet de masque joue. On peut donc raisonnablement s'attendre à ce que les estimateurs qui limitent les effets levier suppriment par là-même systématiquement ces effets de masque. En ce sens, on peut dire que les estimateurs robustes sont plus sensibles que les moindres carrés aux problèmes de colinéarité approchée structurelle.

Notons encore que si les estimateurs qui limitent les effets levier suppriment systématiquement la forte incidence sur la colinéarité que peuvent exercer quelques points particuliers, il n'en est pas de même des estimateurs qui, comme l'estimateur de Huber ou celui du moindre carré médian de Rousseeuw, limitent uniquement l'influence des observations aberrantes. Ceux-ci n'éliminent les fortes incidences sur la colinéarité que lorsque les données concernées sont également des points aberrants, ce qui ne saurait être systématique. Il n'en demeure pas moins que ces estimateurs peuvent également avoir des implications au niveau des problèmes de colinéarité.

3. INDICATEURS DE COLINEARITE APPROCHEE POUR ESTIMATEURS ROBUSTES

Après avoir mis en évidence à la section précédente le rôle que les données atypiques peuvent jouer du point de vue de la colinéarité, nous considérons dans cette section le problème de la détection des situations de colinéarité approchée. Plus précisément, il s'agit d'adapter les indicateurs qui permettent de juger de la stabilité des estimations des moindres carrés ordinaires au cas de la régression robuste.

3.1. Distance à la colinéarité dans le cas des moindres carrés

Notre propos n'est pas ici de faire une revue exhaustive des mesures de diagnostic existantes, on pourra pour cela se référer notamment à Belsley-Kuh-Welsch (1980, ch.3) et à Stewart (1987), mais de rappeler brièvement les mesures que nous nous proposons ensuite d'étendre au cas de la régression robuste.

La colinéarité correspond au cas où il existe une dépendance linéaire exacte entre les k colonnes de la matrice

$$X = (x_i')_{i=1,2,\dots,n} = (x_j)_{j=1,2,\dots,k}$$

du modèle de régression

$$(3.1) \quad y = X\beta + \varepsilon ,$$

c'est-à-dire lorsque le rang de X est inférieur à k . Il y a colinéarité approchée lorsque X est de rang complet, mais qu'une colonne, x_j , par exemple, s'exprime approximativement comme combinaison linéaire des autres; c'est-à-dire s'il existe une relation de la forme :

$$(3.2) \quad x_j = X_{(j)}\alpha + e_j$$

où $X_{(j)}$ de dimension $n \times (k - 1)$ désigne la matrice X sans sa j -ème colonne, pour laquelle le vecteur d'écarts $e_j \in \mathbb{R}^n$ est très petit.

Les coefficients de corrélation multiple $R_j, j = 1, 2, \dots, k$ mesurant l'association de chaque variable x_j avec l'ensemble des $(k - 1)$ variables de $X_{(j)}$, dont les carrés R_j^2 sont les coefficients de détermination des régressions (3.2), se présentent de ce point de vue comme des indicateurs tout désignés du degré de colinéarité approchée.

Pour des raisons d'interprétation, on considère de préférence certaines transformations de ces coefficients de corrélation multiple. En particulier, on se réfère souvent aux nombres

$$\kappa_j^2 = \frac{1}{1 - R_j^2}$$

qui se trouvent être les éléments diagonaux de l'inverse C^{-1} de la matrice des corrélations C . Lorsque les colonnes de X sont centrées et normées, on a $C = X'X$. Sous les hypothèses classiques d'indépendance et d'homoscédasticité des termes d'erreurs ε_j ($E(\varepsilon_i \varepsilon_j) = 0$ si $i \neq j$ et $E(\varepsilon_i^2) = \sigma^2$) on a alors $\text{Var}\hat{\beta} = \sigma^2 C^{-1}$, c'est-à-dire pour un coefficient générique :

$$\text{Var}\hat{\beta}_j = \sigma^2 \kappa_j^2 .$$

C'est pourquoi ces coefficients sont connus sous le nom de *facteur d'inflation de la variance* ou *VIF_j*, d'après la terminologie anglaise. On notera que cette interprétation en termes d'inflation de variance n'est plus aussi pertinente lorsqu'on relâche les hypothèses sur la distribution des termes d'erreur.

Nous préférons pour notre part la terminologie de Stewart (1987), qui appelle *indice de colinéarité* (collinearity index) la racine du *VIF*, soit

$$(3.3) \quad \kappa_j = \frac{1}{\sqrt{1 - R_j^2}}$$

Comme R_j^2 est le coefficient de détermination pour la régression (3.2), on a pour des variables centrées :

$$R_j^2 = 1 - \frac{\|e_j\|^2}{\|x_j\|^2}$$

d'où, en remplaçant dans l'expression (3.3) :

$$(3.4) \quad \kappa_j^{-1} = \frac{\|e_j\|}{\|x_j\|}$$

Le vecteur e_j est, d'après (3.2), le vecteur des perturbations qu'il faut exercer sur x_j pour rendre la dépendance $x_j = X_{(j)}\alpha$ exacte. Ce vecteur est minimal ($\|e_j\|$ minimal) lorsque α est le vecteur des coefficients de régression des moindres carrés. Ainsi, l'inégalité (3.4) permet d'interpréter l'indice de colinéarité κ_j comme l'inverse de la *distance relative minimale par rapport à une dépendance linéaire parfaite* pour x_j .

Dans la même optique, l'indice de colinéarité maximal

$$\kappa_{\max} = \max_j \{\kappa_j\}$$

peut s'interpréter comme l'inverse de la plus petite *distance relative à la colinéarité*.

La définition (3.3) de l'indice de colinéarité se réfère aux données centrées. Dans certains cas, il peut être plus utile de considérer la distance relative à la colinéarité, selon l'expression (3.4), par rapport aux données non centrées. On définit alors de façon plus générale l'indice de colinéarité par

$$\kappa_j = \frac{\|x_j\|}{\|e_j\|} = \|x_j\| \|x_j^+\|$$

où $\|x_j^+\|$ est la norme de la j -ème ligne de $X^+ = (X'X)^{-1}X'$.

L'*indice de conditionnement* (condition number), peut-être plus connu, est une autre manière de mesurer cette distance à la colinéarité. En notant $\|X\|$ la norme spectrale de la matrice X ,

$$\|X\| = \max_{|b|=1} \|Xb\| ,$$

cet indice de conditionnement, noté $\kappa(X)$, est défini comme le produit de la norme spectrale de X par la norme spectrale de son quasi-inverse $X^+ = (X'X)^{-1}X'$:

$$\kappa(X) = \|X\| \|X^+\| .$$

Il se trouve (c.f. Stewart, 1987) que $\|X^+\|$ est l'inverse de la norme spectrale de la plus petite matrice E qui rend $X + E$ exactement colinéaire. Ainsi,

$$\kappa^{-1}(X) = \frac{\|E\|}{\|X\|}$$

ce qui montre que l'indice de conditionnement s'interprète également en termes d'inverse de distance relative à la colinéarité.

L'indice maximal de colinéarité κ_{\max} présente notamment l'avantage d'être, contrairement à l'indice de conditionnement $\kappa(X)$, insensible aux variations indépen-

dantes d'échelle des variables explicatives. Par ailleurs, en exploitant mieux l'information de la matrice X , κ_{max} donne une idée plus précise du degré de colinéarité. Ainsi, pour le cas où les colonnes de X sont centrées et normées, on a (c.f. Stewart, 1987)

$$\kappa_{max} \leq \kappa(X) \leq k \kappa_{max}$$

d'où il ressort que l'indice de conditionnement ne fournit qu'une borne supérieure pour la mesure plus intuitive que représente κ_{max} .

Pratiquement, compte tenu de la sensibilité de l'indice de conditionnement aux variations d'échelle, on ne s'intéresse en général, du point de vue de la colinéarité, qu'aux indices calculés sur des données préalablement normées. Voir notamment à cet égard l'exemple 1 donné à la section 4.

L'indice de conditionnement, tout comme l'indice de colinéarité, peut être calculé sur les données centrées ou non centrées. Les valeurs obtenues seront évidemment toujours plus faibles dans le cas centré où la norme de x_j , de même que celle de X , est minimale. Pratiquement, s'agissant d'évaluer la distance à la colinéarité, on obtient des mesures intuitivement mieux interprétables, comme le souligne Belsley (1984), en la rapportant à des normes de variables non centrées. L'inverse des indices donne dans ce cas le pourcentage de variation des variables nécessaire pour atteindre la colinéarité. Dans le cas centré, on a le pourcentage de variation nécessaire des variables centrées, et l'on perd alors la référence à l'ordre de grandeur des variables. Le choix entre données centrées ou non centrées est évidemment important, car les valeurs des indices peuvent être très différentes, comme l'illustre notamment l'exemple de Belsley repris ci-après comme exemple 2 à la section 4.

Ce choix dépend en fait du but poursuivi. S'agissant d'évaluer la sensibilité des estimateurs à des perturbations des variables, les indices sur données non centrées paraissent préférables. Par contre, si l'on cherche à déceler la présence de problèmes de précision numérique, on se référera aux valeurs minimales fournies par les indices obtenus avec les données centrées.

Pour conclure cette sous-section, il convient encore de faire les remarques suivantes sur le calcul des indices discutés.

Pour l'indice de conditionnement $\kappa(X)$ on exploite la décomposition singulière de X , $\kappa(X)$ s'obtenant comme le rapport entre μ_{max} la plus grande valeur singulière de X et μ_{min} la plus petite (c.f. par exemple Belsley-Kuh-Welsch, 1980) :

$$\kappa(X) = \frac{\mu_{max}}{\mu_{min}} .$$

Quant aux indices de colinéarité, ils peuvent être calculés en exploitant la décomposition QR de la matrice X (c.f. Stewart, 1974) :

$$X = Q_X R_X$$

où Q_X est une matrice $n \times k$ dont les colonnes forment une base orthonormée de l'espace engendré par les colonnes de X , et R_X est une matrice $k \times k$ triangulaire supérieure. On peut en effet montrer (Stewart, 1987) que :

$$\kappa_k = \frac{\|x_k\|}{|r_{kk}|}$$

où r_{kk} est le k -ème, donc le dernier, terme diagonal de R_X .

3.2. Extension au cas de la régression robuste

Tels que définis précédemment, les indices de colinéarité et de conditionnement ne sont évidemment pas pertinents pour juger de la stabilité d'estimations robustes. En effet, les estimations robustes se présentent en fait comme des estimations fondées sur l'ensemble transformé d'observations résultant des mécanismes qui limitent automatiquement le rôle des données atypiques. La stabilité des estimations va alors dépendre non pas de la colinéarité entre données originales X , mais de la colinéarité au sein de l'ensemble transformé.

Dès lors, une démarche naturelle pour analyser les problèmes de colinéarité dans le cadre de la régression robuste consiste à mettre en évidence l'ensemble de données transformées et à calculer les indices classiques introduits au paragraphe 3.1 sur ces données transformées. C'est la voie que nous suivons ici en nous référant en particulier au cas des estimateurs robustes de classe M.

Rappelons qu'un estimateur T de classe M est un estimateur défini implicitement par l'équation vectorielle :

$$(3.5) \quad \sum_{i=1}^n \eta(x_i, y_i - x_i' T) x_i = 0$$

où $\eta : \mathcal{R}^k \times \mathcal{R} \rightarrow \mathcal{R}$ est une fonction du vecteur de variables explicatives x_i et de l'écart $r_i = y_i - x_i' T$ au plan de régression. L'équation ci-dessus généralise les conditions du premier ordre du maximum de vraisemblance sous les hypothèses classiques. En particulier pour $\eta(x_i, r_i) = r_i$, on retrouve l'estimateur des moindres carrés ordinaires.

Dans le cadre de la régression robuste, la fonction η permet d'introduire des pondérations en fonction de l'effet levier par le biais de x_i et de l'aberrance par le biais de r_i .

Les trois types d'estimateurs robustes les plus courants sont ceux de :

- *Huber* défini par $\eta(x, r) = \psi_c(r)$
- *Mallows* défini par $\eta(x, r) = \omega(x) \psi_c(r)$

• *Schweppe* défini par $\eta(x, r) = \omega(x) \psi_c(r/\omega(x))$,

où les $\omega(x)$ sont des fonctions de pondération qui permettent de limiter l'influence des points levier, et $\psi_c(r)$ la fonction de Huber permettant de borner l'influence des résidus trop grands :

$$\psi_c(r) = \begin{cases} r & \text{si } |r| < c \\ c \operatorname{signe}(r) & \text{sinon} \end{cases}$$

Remarquons qu'en posant :

$$p(x, r; \eta, T) = \frac{\eta(x, r)}{r}$$

le système (3.5) s'écrit :

$$\sum_{i=1}^n p_i(x_i, r_i; \eta, T) r_i x_i = 0$$

ce qui montre que les estimateurs de classe M se présentent comme des estimateurs des moindres carrés pondérés. On notera cependant qu'il s'agit de moindres carrés pondérés itérés dont les poids sont notamment fonction des observations.

Pour les trois types d'estimateurs définis ci-dessus, les poids p sont :

• Huber

$$p(x, r; \cdot) = \frac{\psi_c(r)}{r} = \begin{cases} 1 & \text{si } |r| < c \\ c/|r| & \text{sinon} \end{cases}$$

• Mallows

$$p(x, r; \cdot) = \omega(x) \frac{\psi_c(r)}{r} = \begin{cases} \omega(x) & \text{si } |r| < c \\ c \omega(x)/|r| & \text{sinon} \end{cases}$$

• Schweppe

$$p(x, r; \cdot) = \frac{\omega(x)}{r} \psi_c\left(\frac{r}{\omega(x)}\right) = \begin{cases} 1 & \text{si } |r|/\omega(x) < c \\ c \omega(x)/|r| & \text{sinon} \end{cases}$$

En désignant par P la matrice diagonale des pondérations, les données transformées finales sur lesquelles se fonde la régression robuste seront :

$$[X^* \ y^*] = P^{1/2} [X \ y] .$$

L'application des indices de colinéarité et de conditionnement à la matrice $X^* = P^{1/2} X$ conduit alors aux définitions suivantes :

Définition 3.1 : On appelle *indices de colinéarité pour régression robuste* les indices :

$$\kappa_j = \frac{\|x_j^{**}\|}{\|e_j^{**}\|} = \|x_j^{**}\| \|x_j^{**\dagger}\| \quad j = 1, 2, \dots, k$$

où x_j^{**} est la j -ème ligne de $X^{**} = (X^* X^*)^{-1} X^*$.

Définition 3.2 : On appelle *indice de conditionnement pour régression robuste* l'indice :

$$\kappa(X^*) = \frac{\|X^*\|}{\|X^{*+}\|} .$$

Comme dans le cas classique, les indices κ_j et $\kappa(X)$ représentent des inverses de distances relatives à la colinéarité. Les distances sont cependant évaluées ici dans la métrique P . Ainsi, par exemple, $\|x_j^*\|^2 = x_j' P x_j$. Pour ce qui est de l'indice de conditionnement, on rappelle qu'il est préférable de le calculer sur des données normées.

Nous avons vu que dans le cas classique, l'indice de colinéarité κ_j calculé sur les données centrées (pas nécessairement normées) s'interprète également, selon la relation (3.3), en termes de la corrélation multiple R_j^2 de x_j avec les autres variables $X_{(j)}$, ou de façon équivalente comme la racine de la part de la variance de x_j non reproduite par la régression de x_j sur $X_{(j)}$. Cet indice est également, dans le cas classique centré, égal à la racine du VIF_j .

Pour voir ce qu'il en est ici, il convient tout d'abord de préciser l'opération de centrage à considérer. Les estimateurs robustes conduisant à pondérer les observations selon leur position, il s'agit de centrer des données pondérées :

$$x_{ij}^{\mathcal{P}} = x_{ij} - \frac{1}{\sum_i p_i} \sum_i p_i x_{ij} .$$

En notant $\omega = (\omega_i)$ le vecteur des termes diagonaux $p_i^{1/2}$ de $P^{1/2}$, la matrice des données originales centrées s'écrit :

$$X^{\mathcal{P}} = X - \frac{1}{\omega' \omega} \omega \omega' P^{1/2} X ,$$

où $\mathbf{1}$ est un vecteur de 1. Les indices de colinéarité étant définis en termes de données transformées, ils doivent être calculés sur la matrice des données centrées transformées :

$$\begin{aligned} X^{*\mathcal{C}} &= P^{1/2} X^{\mathcal{P}} \\ &= M X^* \end{aligned}$$

$$\text{avec } M = I - \frac{1}{\omega' \omega} \omega \omega' .$$

Notons encore que dans le cas d'une régression avec constante, cette opération de centrage conduit à annuler la colonne associée à la constante, qu'il s'agira alors d'éliminer avant de calculer les indices de colinéarité.

Pour les données centrées, l'indice de colinéarité est :

$$\kappa_j^{\mathcal{C}} = \frac{\|M x_j^*\|}{\|M e_j^*\|} .$$

Le numérateur $\|Mx_j^*\|$ est, à un facteur multiplicatif près, égal à l'écart-type pondéré de x_j calculé à l'aide de la métrique P . De même, le dénominateur est l'écart-type pondéré des termes d'erreurs e_j de la régression pondérée de x_j sur $X_{(j)}$. L'interprétation en termes de racine de part non reproduite de variance ou de corrélation multiple reste donc valable.

Quant à l'interprétation en termes de facteur d'amplification de la variance, elle nécessite ici l'hypothèse d'homoscédasticité des erreurs transformées $\varepsilon^* = P\varepsilon$, c'est-à-dire

$$(3.6) \quad E(\varepsilon^*\varepsilon^{*\prime}) = \sigma^2 I \Leftrightarrow E(\varepsilon\varepsilon') = \sigma^2 P^{-1} .$$

Cette hypothèse est légitime lorsqu'on utilise par exemple un estimateur de Huber, qui ne réduit l'importance que des observations aberrantes. Lorsque, par contre, le poids d'observations est réduit en raison de leur effet levier, l'hypothèse ci-dessus n'a plus la même légitimité. Or, du point de vue de la colinéarité, c'est précisément la limitation de l'influence de ces données atypiques dans l'espace X des variables explicatives qui nous intéresse. De façon générale, le carré du coefficient de colinéarité des données centrées n'est plus rigoureusement un facteur d'amplification de la variance. Il s'agit de l'amplification de $\sigma^2(X^*X^*)^{-1}$, alors que la matrice de covariance est, par exemple sous l'hypothèse $E(\varepsilon\varepsilon') = \sigma^2 I$ et pour P indépendant de ε :

$$\text{Var} \hat{\beta}^* = \sigma^2 (X^*X^*)^{-1} X^{*\prime} P X^* (X^*X^*)^{-1} .$$

4. ILLUSTRATIONS

Exemple 1

Considérons à titre de premier exemple la matrice X de données initiales suivantes :

$$X = \begin{pmatrix} 1 & 0 & 4 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \\ 1 & 3 & 1 \\ 1 & 4 & 0 \\ 1 & 2 & 100 \end{pmatrix}$$

où les éléments de la troisième colonne à, l'exception du dernier, sont liés à ceux des deux premières colonnes par la relation :

$$x_{i3} = 4x_{i1} - x_{i2} \quad \text{pour } i = 1, 2, 3, 4, 5$$

La sixième donnée est un fort point levier, dont l'importance sera réduite par des estimateurs robustes du type Mallows ou Schweppe. Comme la première colonne de X est une constante, on a évidemment $\kappa_2^c = \kappa_3^c$ pour les données centrées. On vérifie que la corrélation entre x_2 et x_3 , est très faible. Elle vaut :

$$\text{corr}(x_2, x_3) = R_2 = R_3 = 0,0353 \text{ ,}$$

valeur dont on déduit cet indice de colinéarité centré :

$$\kappa_2^c = \kappa_3^c = \frac{1}{\sqrt{1 - R_2^2}} = 1,0006 \text{ .}$$

Le tableau 4.1 donne la valeur des indices de colinéarité κ_i et de l'indice de conditionnement $\kappa(X)$ pour les données X centrées et non-centrées, ainsi que pour les données obtenues en pondérant de quatre manières différentes la donnée atypique.

On notera que les indices de colinéarité donnent une image moins pessimiste de l'intensité de la colinéarité que l'indice de conditionnement. Les deux indices n'évoluent pas toujours dans le même sens.

Tableau 4.1
Indices de colinéarité κ_i et de conditionnement $\kappa(\cdot)$
pour différentes pondérations de la donnée atypique de l'exemple 1

racine du poids de x_6	données non centrées				données normées	données centrées		données centrées normées
	κ_1	κ_2	κ_3	$\kappa(X)$	$\kappa(X^n)$	$\kappa_2^c = \kappa_3^c$	$\kappa(X^c)$	$\kappa(X^{cn})$
ω_6								
1	1.93	1.85	1.12	83.1	3.72	1.0006	28.3	1.04
0.5	1.81	1.76	1.05	41.9	3.38	1.002	15.2	1.07
0.1	1.98	1.82	1.16	10.7	3.73	1.05	3.45	1.37
0.01	9.29	5.85	5.68	29.7	18.7	3.38	6.61	6.60
0.001	91.3	55.9	55.9	290.5	184.5	32.3	64.5	64.5

Pour les données initiales, les indices de colinéarité et l'indice de conditionnement des données normées indique un très léger problème de colinéarité. Une régression sur ces données donnera lieu à des estimations légèrement instables. L'indice de conditionnement des données non normées est donné à titre indicatif, mais est sans grand intérêt. Les indices de colinéarité étant insensibles aux changements individuels d'échelle des variables, ils sont évidemment identiques dans le cas normé et non normé. Selon ces indices κ_i , la colinéarité est à une variation de plus de 50% des variables. Notons que la légère baisse des indices lorsqu'on réduit un peu l'importance de la donnée atypique, montre que le soupçon de colinéarité détecté est en fait ici une colinéarité induite par

cette donnée. Les indices κ_i^c évalués sur les données centrées sont du même ordre de grandeur que l'indice de conditionnement $\kappa(X^{cn})$ des données centrées normées. Leur valeur proche de 1 indique l'absence de problèmes de précision numérique.

Comme prévu, les indices montrent comment la colinéarité masquée apparaît lorsqu'on diminue le poids de la donnée atypique. La colinéarité ne ressort cependant que pour des pondérations très petites. N'oublions pas que le poids de la donnée correspond en fait au carré du ω_i indiqué. Seuls les estimateurs qui tendent à annuler le poids de cet important point levier seront alors affectés dans leur précision numérique par cette colinéarité masquée.

Exemple 2

Pour ce deuxième exemple, nous reprenons les données utilisées par Belsley (1984) pour illustrer sa discussion sur l'opportunité de centrer ou non les données. Nous les reproduisons au tableau 4.2 avec deux données atypiques supplémentaires, soit les données 21 et 22.

Tableau 4.2
Données de l'exemple 2 (Belsley)

i	x_{i1}	x_{i2}	x_{i3}
1	1	.996926	1.000060
2	1	.997091	.998779
3	1	.997300	1.000680
4	1	.997813	1.002240
5	1	.997898	1.000650
6	1	.998140	1.000500
7	1	.998556	.999596
8	1	.998537	1.002620
9	1	.999414	1.003210
10	1	.999678	1.001300
11	1	.999926	.997579
12	1	.999995	.998597
13	1	1.000630	.995316
14	1	1.000950	.995966
15	1	1.001180	.997125
16	1	1.001770	.998951
17	1	1.002310	1.001020
18	1	1.003060	1.001860
19	1	1.003940	1.003530
20	1	1.004690	1.000210
21	1	1	100
22	1	100	-100

Belsley trouve, sans les deux points atypiques évidemment, un indice de conditionnement normé de 1342 dans le cas non centré et de 1 dans le cas centré. Manifestement, au vu du diagramme de dispersion des données 1 à 20 représenté à la figure 4.1, les points ne sont pas alignés, ce qu'indique clairement l'indice de conditionnement centré. La valeur astronomique de l'indice non centré reflète par contre le fait que la colinéarité peut être atteinte par des variations de l'ordre de 1/1342 ème seulement des variables. En ce sens, on est proche de la colinéarité, et une régression sur ces données donnera des estimations très instables.

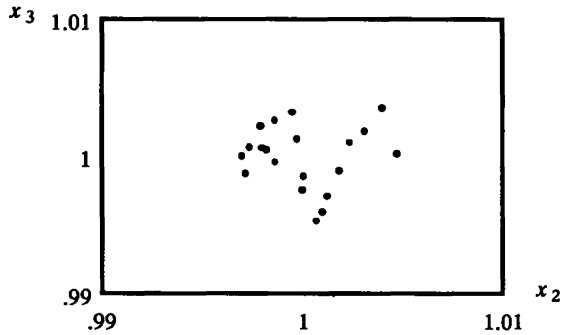


Figure 4.1 : Données originales de Belsley

Le tableau 4.3 illustre l'évolution des indices de colinéarité lorsqu'on limite le poids des deux données atypiques. On constate que les deux données supplémentaires masquent le problème de colinéarité des données non centrées, soit le problème de stabilité des estimateurs. En annulant progressivement le poids des données atypiques, les valeurs des indices tendent vers celles des données originales.

Tableau 4.3
Données de Belsley et deux points atypiques:
indices de colinéarité κ_i et de conditionnement $\kappa(\cdot)$

racine du poids de x_{21} et x_{22}	données non centrées			données normées	données centrées	données centrées normées
	κ_1	κ_2	κ_3	$\kappa(X^n)$	$\kappa_2^c = \kappa_3^c$	$\kappa(X^{cn})$
$\omega_{21} = \omega_{22}$						
1	1.08	1.52	1.46	2.65	1.46	2.53
0.1	1.45	1.58	1.50	2.87	1.43	2.45
0.01	10.2	6.61	4.73	21.6	1.43	2.45
0.001	100.5	63.7	44.7	214.7	1.42	2.42
0.0001	531.7	348.1	283.6	1132.4	1.10	1.54
0.00001	631.8	443.7	444.8	1340.3	1.00006	1.01

En ne retenant que la donnée atypique x_{21} , celle-ci induit une colinéarité accidentelle qui cache le problème structurel des données originales. Le tableau 4.4 montre l'évolution des indices suite à une réduction de l'importance de ce point levier.

Tableau 4.4
Données de Belsley et un seul point atypique:
indices de colinéarité κ_i et de conditionnement $\kappa(\cdot)$

racine du poids de x_{21}	données non centrées			données normées	données centrées	données centrées normées
	κ_1	κ_2	κ_3	$\kappa(X^n)$	$\kappa_2^c = \kappa_3^c$	$\kappa(X^{cn})$
ω_{21}						
1	456.9	456.9	1.04	941.3	1.00	1.0010
0.1	446.0	446.0	1.10	953.6	1.00	1.0001
0.01	445.9	445.9	4.63	1086.5	1.00	1.0000
0.001	448.2	445.9	45.0	1093.6	1.00	1.0001
0.0001	548.2	445.9	318.6	1191.0	1.00	1.0009
0.00001	631.8	445.9	447.1	1340.3	1.00	1.0012

Exemple 3

Ce dernier exemple montre l'évolution des indices dans le cas où la colinéarité masquée implique trois variables, mais pas la constante. La matrice considérée est la suivante :

$$X = \begin{pmatrix} 1 & 3 & 4 & 7 \\ 1 & -2 & 3 & 1 \\ 1 & 2 & -3 & -1 \\ 1 & 5 & 1 & 6 \\ 1 & 4 & -2 & 2 \\ 1 & -3 & -1 & -4 \\ 1 & 2 & 2 & 100 \end{pmatrix}$$

pour laquelle on a :

$$x_{i4} = x_{i2} + x_{i3} \quad \text{pour } i = 1, 2, \dots, 6.$$

Le tableau 4.5 montre comment cette colinéarité apparaît lorsqu'on réduit l'importance de la septième donnée. On peut noter ici que l'indice de colinéarité κ_1 associé à la constante, qui n'intervient pas dans la relation ci-dessus, reste proche de 1.

Tableau 4.5
Exemple 3, colinéarité masquée n'impliquant pas la constante:
indices de colinéarité κ , et de conditionnement $\kappa(\cdot)$

racine du poids de x_7	données non centrées				données normées	données centrées			données centrées normées
	κ_1	κ_2	κ_3	κ_4	$\kappa(X^n)$	κ_2^c	κ_3^c	κ_4^c	$\kappa(X^{cn})$
ω_7									
1	1.22	1.17	1.08	1.17	2.02	1.01	1.05	1.06	1.42
0.1	1.13	1.41	1.22	1.50	2.88	1.26	1.21	1.41	2.40
0.01	1.13	8.60	6.67	10.8	23.15	7.68	6.61	9.76	19.48
0.001	1.13	85.3	65.9	107.7	230.9	76.2	65.3	97.0	194.3
0.0001	1.13	852.6	658.8	1077.	2329.1	761.9	653.3	970.7	1942.5

5. CONCLUSION

Cet article ne constitue qu'une amorce de discussion des problèmes de colinéarité approchée en relation avec la régression robuste. Il s'agissait ici essentiellement de caractériser les situations où quelques données atypiques peuvent influencer fortement sur le degré de colinéarité. Nous avons également vu comment les mesures classiques de distance à la colinéarité pouvaient s'étendre au cas de la régression robuste.

Pour conclure, nous nous proposons de commenter encore brièvement la portée et les limites des indices considérés, ce qui nous amènera à mettre en évidence quelques perspectives prometteuses de recherche.

L'intérêt des indices de colinéarité dépend évidemment de ce que l'on veut détecter. Pour ce qui est de la colinéarité, les effets concernent principalement deux aspects :

- l'efficacité des estimateurs (inflation de la variance)
- les difficultés de calcul numérique des estimations.

En ce qui concerne le premier point, les extensions proposées peuvent fournir des indications précieuses. On rappelle néanmoins que l'interprétation des coefficients de colinéarité en termes de facteur d'inflation de la variance suppose quelques restrictions sur la distribution des termes d'erreur. Dans cette optique, des indices d'inflation de variance fondés directement sur la matrice des covariances asymptotiques des estimateurs robustes pourraient s'avérer plus pertinents.

Pour ce qui est des problèmes de calcul numérique, l'intérêt des notions proposées est moins évident. En effet, le calcul des indices suppose connu la matrice de pondération P . Or celle-ci est déterminée, dans le cas des estimateurs de classe M, itérativement avec les estimations de β , si bien que les indices ne peuvent pas, contrairement au cas classique, être obtenus avant le calcul des estimations. On ne saurait dès lors trop en attendre pour détecter a priori les difficultés de calcul qui peuvent se présenter. Notons qu'ici, les méthodes de détections des points levier à haut point de rupture qui se fondent sur des distances robustes, telles que par exemple la distance MVE de Rousseeuw :

$$d_{MVE}^2(x_i, c_{MVE}) = (x_i - c_{MVE})' V_{MVE}^{-1} (x_i - c_{MVE})$$

définie à partir du centre c_{MVE} de l'ellipse de volume minimal contenant la moitié des observations et la matrice des covariances V_{MVE} de cette moitié d'observations (c.f. Rousseeuw-Leroy, 1987), offrent néanmoins une possibilité d'obtenir un ensemble de données sans points levier indépendamment du calcul des estimations.

De manière générale, le recours aux estimateurs robustes et aux indices appropriés de colinéarité devrait, en permettant notamment de repérer les situations de colinéarité accidentelle et de colinéarité masquée, conduire à une meilleure maîtrise des problèmes de colinéarité propres à un ensemble de données.

Le diagnostic des situations particulières de colinéarité sera d'autant plus efficace que l'estimateur robuste choisi sera fortement sensible à la colinéarité, au sens où il aura tendance à affecter de poids très faibles les points levier à forte incidence sur la colinéarité. En particulier les estimateurs à haut point de rupture exploitant les méthodes d'ellipse minimale mentionnées ci-dessus, seront très performants sur ce plan.

Qu'il s'agisse d'obtenir des estimations robustes, ou de diagnostiquer la source de problèmes de colinéarité, il serait souhaitable, pour guider l'utilisateur dans le choix d'un estimateur, de disposer d'une mesure de la sensibilité théorique à la colinéarité. Pour les estimateurs de classe M, un tel indicateur pourrait par exemple être dérivé de la fonction *CVF* de changement de variance (c.f. Hampel et al., 1986, p. 324-325), en définissant un équivalent théorique de la colinéarité empirique.

Références bibliographiques

- Belsley, D.A., (1984) Demeaning Conditioning Diagnostics Through Centering. *The American Statistician*, 38(2), 73-77.
- Belsley, D.A., E. Kuh et R.E. Welsch (1980) *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

- Jureckova, J.** (1971) Non parametric Estimate of Regression Coefficient, *The Annals of Mathematical Statistics*, 42, 1328-1338.
- Hadi, A.S.** (1987) Diagnosing collinearity-influential observations, *Computational Statistics and Data Analysis*, 7, 143-159.
- Hadi, A.S. et P.F. Velleman** (1987) Comment : Diagnosing Near Collinearities in Least Squares Regression, *Statistical Science*, 2(1), 93-98.
- Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw et W.A. Stahel** (1986) *Robust Statistics, The Approach Based on Influence Functions*, Wiley, New York.
- Huber, P.J.** (1973) Robust Regression : Asymptotics, Conjectures, and Monte Carlo, *The Annals of Statistics*, 1, 799-821.
- Mason, R.L. et R.F. Gunst** (1985), Outlier-induced collinearities, *Technometrics*, 27, 401-407.
- Ritschard, G.**, (1990) Détection de données atypiques, in *La modélisation : confluent des sciences*, (Brissaud, Forsé, Grafmeyer et Zighed, eds), Editions du CNRS, Paris.
- Rousseeuw, P.J.** (1984) Least Median of Squares Regression, *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J. et A.M. Leroy** (1987) *Robust Regression and Outlier Detection*, Wiley, New York.
- Stewart, G.W.** (1981) Collinearity and Least Squares Regression, *Statistical Science*, 2(1), 68-100.