DÉTECTION DE LA MULTICOLINÉARITÉ DANS UN MODÈLE LINÉAIRE ORDINAIRE : quelques éléments pour un usage averti des indicateurs de Belsley, Kuh et Welsch

*Hélène Erkel-Rousse**

Introduction

Dans un modèle linéaire ordinaire à N observations et K variables explicatives, une situation de multicolinéarité "approchée" apparaît lorsque certains vecteurs colonnes de X, la matrice (NxK) des variables exogènes, forment un système de vecteurs "presque" liés, et qu'alors la matrice carrée X'X a un certain nombre de valeurs propres très petites, qui pèsent numériquement sur son inversion. On dit dans ce cas que la matrice X est mal conditionnée. Cette multicolinéarité peut engendrer de graves instabilités sur certaines estimations des paramètres du premier ordre obtenues par les moindres carrés ordinaires (MCO). Ces instabilités résultent d'une part de difficultés purement numériques liées à l'inversion d'une matrice presque singulière (on peut se heurter aux limites de l'algorithme de calcul). D'autre part, la matrice de variances-covariances de l'estimateur des MCO étant proportionnelle à l'inverse de la matrice X'X, une partie de ses éléments peuvent être très grands et certains estimateurs se caractériser ainsi par une forte variabilité. Ces manifestations reflètent la difficulté que l'économètre rencontre à séparer les effets respectifs des différentes variables explicatives en présence de multicolinéarité. Via son influence sur la matrice de variances-covariances de l'estimateur des MCO, la multicolinéarité s'apparente donc à un problème numérique doublé d'un problème statistique. Dans un modèle linéaire ordinaire affecté de multicolinéarité, l'estimateur des MCO demeure sans biais linéaire optimal ; mais il a l'inconvénient d'être peu robuste au voisinage de la multicolinéarité stricte.

Dès lors qu'on ne peut ignorer les conséquences fâcheuses de la multicolinéarité, il convient de disposer d'indicateurs de sa détection fiables et précis, et de maîtriser leur interprétation. On se propose ici de clarifier l'utilisation d'indicateurs de ce type proposés par D.A. Belsley, E. Kuh et R.E. Welsch (BKW), dans un ouvrage de référence

* À l'époque de la rédaction de cet article, Hélène Erkel-Rousse travaillait au CREST, INSEE.
Nous remercions la Société de Statistique de France d'avoir autorisé la reproduction de cet article, publié dans la *Revue de Statistique Appliquée*, 1995, XLIII(4), pp. 19-42.
L'auteur remercie Pierre Cazes et un referee anonyme pour leurs commentaires sur une version antérieure de l'article.

publié en 1980 : les indices de conditionnement et le tableau de décomposition des variances. Après les avoir définis et en avoir donné une interprétation géométrique (chapitre 1), on montrera que ces indicateurs sont souvent incorrectement interprétés, et on précisera une règle de lecture générale permettant une détection satisfaisante de la multicolinéarité (chapitre 2). On précisera alors comment se manifeste une multicolinéarité croissante entre une partie des vecteurs colonnes des variables explicatives : détérioration progressive de la précision des estimateurs des moindres carrés relatifs aux coefficients de ces vecteurs, absence de "contamination" des estimations des autres coefficients. On soulignera une difficulté d'interprétation des indicateurs de BKW liée à leur non invariance par changement d'origine. On proposera à cette occasion un point de vue sur les débats relatifs à l'opportunité de raisonner sur les variables explicatives véritables centrées (chapitre 3).

I- Les indicateurs de détection de la multicolinéarité de BKW : définition et interprétation géométrique

BKW ne sont pas les premiers auteurs à avoir mené une réflexion sur la détection de la multicolinéarité (Cf. par exemple Farrar et Glauber [1967], Gunst et Mason [1977], Mason, Gunst et Webster [1975], Silvey [1969]). Cependant, leurs travaux ont permis un progrès important dans ce domaine. En effet, les indicateurs qu'ils ont proposés en adaptant les résultats théoriques de S.D. Silvey (1969) au cadre de l'économétrie empirique permettent d'établir une règle de détection de la multicolinéarité avec identification de ses sources et évaluation de la gravité de ses conséquences pour l'estimation. En outre, les indicateurs de BKW ne s'apparentent pas à des tests de multicolinéarité, qui sont souvent critiqués dans la mesure où la multicolinéarité ne peut être appréhendée comme un problème d'inférence statistique si elle relève des données observées et non des variables aléatoires à la base de leur constitution (Cf. par exemple Maddala [1977] ou Erkel-Rousse [1994 et 1994/95]). Si certains auteurs leur ont depuis 1980 opposé des concurrents, comme les "VIF" (variance inflation factors) (Cf. Stewart [1987]), les indicateurs de BKW demeurent cependant parmi les plus riches, les plus célèbres et les plus accessibles à l'économètre praticien. Ils sont programmés dans la procédure REG (option COLLIN) du logiciel SAS, de même que les "VIF" et leur inverse, les "TOL" (tolérances).

I-1 Le cadre de référence

On se place dans le modèle linéaire ordinaire à K variables explicatives et N observations :

$$\underset{(N,1)}{y} = \underset{(N,K)}{X} \quad \underset{(K,1)}{b} + \underset{(N,1)}{u} \quad (1)$$

où y est le vecteur de la variable dépendante, X la matrice des variables explicatives, b le vecteur des paramètres du premier ordre, et u le vecteur des perturbations, de matrice de variances-covariances : $V(u) = \sigma^2 I_N$.

On suppose que la matrice X est formée de vecteurs colonnes $X_1, X_2, ..., X_k, ..., X_K$ non nuls et non linéairement dépendants, de sorte que le paramètre multidimensionnel b est identifiable et peut être estimé par les moindres carrés ordinaires (MCO). On note $\hat{b} = (X'X)^{-1}X'y$ l'estimateur des MCO de b et $\mathcal{L}(X)$ le sous-espace vectoriel engendré par les colonnes de X. On définit en outre la matrice X^* des vecteurs colonnes de X normés à 1, et on note $X_1^*, X_2^*, ..., X_k^*, ..., X_K^*$ les vecteurs colonnes de cette matrice. Bien entendu, les sous-espaces vectoriels $\mathcal{L}(X^*)$ et $\mathcal{L}(X)$ sont identiques. Les matrices $X'X$ et $X^{*'}X^*$ sont symétriques définies positives, donc diagonalisables, avec des valeurs propres toutes strictement positives. Une situation de multicolinéarité approchée se traduit par l'existence de valeurs propres proches de zéro dans les deux matrices diagonalisées.

Les indicateurs de BKW reposent sur des éléments de la diagonalisation de la matrice $X^{*'}X^*$ et non de celle de la matrice X'X. Ce point n'a pas toujours été perçu du fait qu'en amont de leurs calculs, BKW (1980) se bornent à indiquer que la matrice X est supposée avoir été normalisée, sans adopter une notation spécifique pour la matrice X^*. Il en résulte que les présentations des indicateurs de BKW omettent souvent de mentionner cette normalisation (Cf. Judge et *alii* [1980]), ou bien y font référence tout en appliquant les calculs de BKW à des matrices X non normalisées, ce qui est source de confusion. La normalisation de BKW revient à travailler sur le modèle (1) reparamétré de manière à y faire apparaître la matrice X^*, i.e. sur le modèle (2) :

$$y = X^* b^* + u \quad (2)$$

où b^* est le vecteur (K,1) des paramètres b_k^* liés aux b_k par la relation :

$$b_k^* = b_k \left|\left| X_k \right|\right|, \quad \left|\left| X_k \right|\right| \text{ étant la norme de } X_k.$$

Les indicateurs de BKW sont constitués des indices de conditionnement et du tableau de décomposition des variances des estimateurs.

I-2 Les indices de conditionnement

La diagonalisation de la matrice X^* s'écrit :

$$X^{*'}X^* = Q^* \Lambda^* Q^{*'}$$

où $\underset{(K,K)}{Q^*}$ est la matrice orthonormée des vecteurs propres de $X^{*'}X^*$, associée à la matrice $\Lambda^* = \text{diag}\,(\lambda_j^*)_{j=1,\ldots,K}$ de ses valeurs propres, supposées (sans perte de généralité car cela revient à réordonner les vecteurs X_k) classées par ordre décroissant :

$$\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_K^* \geq 0$$

Les indices de conditionnement notés $\left(\eta_j^*\right)_{j=1,2,\ldots,K}$ sont définis comme les racines carrées des rapports entre la plus grande valeur propre de la matrice $X^{*'}X^*$ (*resp.* la plus grande valeur singulière de la matrice X^*) et chacune de ses autres valeurs propres (*resp.* chacune de ses autres valeurs singulières), c'est-à-dire :

$$\eta_j^* = \sqrt{\frac{\lambda_1^*}{\lambda_j^*}} \qquad \forall\, j \in \{1, 2,\ldots, K\}$$

D'où : $1 = \eta_1^* \leq \ldots \leq \eta_j^* \leq \ldots \leq \eta_K^* \qquad \forall\, j \in \{2,\ldots, K\}$

η_K^* est appelé **l'indice de conditionnement maximal**. Lorsque les vecteurs colonnes de X sont orthogonaux, η_K^* est égal à 1. Lorsque plusieurs vecteurs colonnes de X sont liés par une relation linéaire exacte, c'est également le cas pour certains vecteurs colonnes de X^*. η_K^* est alors égal à $+\infty$, puisque λ_K^* est nulle et λ_1^* strictement positive. Des situations intermédiaires entre l'orthogonalité et la liaison linéaire exacte se traduisent par des indices de conditionnement vérifiant : $1 \leq \eta_j^* < +\infty \quad \forall\, j \in \{1, 2,\ldots, K\}$.

Des sauts importants entre indices de conditionnement successifs reflètent de fortes différences d'ordre de grandeur entre les valeurs propres de la matrice $X^{*'}X^*$ et traduisent donc le risque d'une situation de multicolinéarité plus ou moins sérieuse. Sur la base d'études de simulation effectuées sur des données de types différents, BKW [1980] et Belsley [1991] estiment que des liaisons faibles entre les vecteurs colonnes de X^* sont associées à des indices de l'ordre de 5 à 10, alors que des liaisons relativement fortes sont associées à des indices de l'ordre de 30 ou plus. Des indices supérieurs à la centaine correspondent notamment à des situations nettement pathologiques, mais on considère qu'il y a risque sérieux d'un problème de multicolinéarité dès que certains indices atteignent des ordres de grandeur proches de la trentaine. Des indices légèrement inférieurs à 30 (à partir de 20 à 25 environ) sont ambigus (Cf. Belsley [1991], Rousse [1990], et voir annexe). Ces seuils ne sont qu'indicatifs. Par exemple, un indice de conditionnement de l'ordre de la trentaine devient peu significatif lorsqu'il côtoie un autre indice de conditionnement de l'ordre de 3000 (Cf. Belsley [1991]).

I-3 Le tableau de décomposition des variances

Si les indices de conditionnement mesurent un degré de multicolinéarité entre les variables explicatives normées à 1, le tableau de décomposition des variances permet de détecter l'existence et l'origine de problèmes induits par cette situation de multicolinéarité, et de voir sur quels coefficients ces problèmes peuvent porter. La décomposition des variances complète donc utilement le diagnostic en identifiant les directions et les estimateurs susceptibles d'être entachés de multicolinéarité, quand il en existe. Elle s'obtient en raisonnant non pas sur les variables initiales, mais sur les composantes principales engendrant le sous-espace vectoriel $\mathcal{L}(X) = \mathcal{L}(X^*)$, qui ont l'avantage d'être non corrélées.

Q^* (définie *supra*) est la matrice des vecteurs axiaux factoriels ou des facteurs dans l'analyse en composantes principales non centrée de X^* avec comme métrique l'identité (Cf. Saporta [1990]).
Soit Z^* la matrice N x K des composantes principales. On a :

$$X^{*'}X^*Q^* = Q^*\Lambda^* \text{ et } Z^* = X^*Q^* \Rightarrow Z^{*'}Z^* = \Lambda^*$$

Le modèle (2) s'écrit alors :

$$y = Z^*\beta^* + u = X^*Q^*\beta^* + u \quad (3)$$

et donc : $b^* = Q^*\beta^*$,

d'où : $\hat{b}^* = Q^*\hat{\beta}^*$,

$\hat{b}^*$ et $\hat{\beta}^*$ étant respectivement les estimateurs des MCO de b^* et de β^* dans les modèles (2) et (3). D'où l'on déduit la relation entre les composantes de $\hat{\beta}$ et $\hat{b}$ (puisque $\hat{b}_k^* = \hat{b}_k \left|\left|X_k\right|\right|$) :

$$\hat{b}_k = \sum_{j=1}^{K} \frac{Q_{k,j}^* \hat{\beta}_j^*}{\left|\left|X_k\right|\right|}$$

$Q_{k,j}^*$ étant la $k^{ième}$ composante du $j^{ième}$ facteur (i.e. le terme (k,j) de Q^*). Comme :

$$V(\hat{\beta}^*) = \sigma^2(Z^{*'}Z^*)^{-1} = \sigma^2 \text{ diag}\left(\frac{1}{\lambda_j^*}\right)_{j=1,2,\ldots,K}$$

les estimateurs des β_j^* sont non corrélés deux à deux, et on peut décomposer la variance de tout $\hat{b}_k$ en fonction de celles des $\hat{\beta}_j^*$:

$$\forall k \in \{1, 2,..., K\}, \quad V(\hat{6}_k) = \frac{\sigma^2}{\left|\left|X_k\right|\right|^2} \sum_{j=1}^{K} \frac{(Q^*_{k,j})^2}{\lambda^*_j}$$

$$\text{On pose : } \forall (k,j) \in \{1, 2,..., K\}^2, \quad \pi^*_{j,k} = \frac{\sigma^2}{\left|\left|X_k\right|\right|^2} \frac{(Q^*_{k,j})^2}{\lambda^*_j V(\hat{6}_k)} = \frac{V\left(\frac{Q^*_{k,j}\hat{\beta}^*_j}{\left|\left|X_k\right|\right|}\right)}{V(\hat{6}_k)}$$

$\pi^*_{j,k}$, qui représente la part de variabilité de $\hat{6}_k$ "portée" par la $j^{ième}$ composante principale de X^*, est appelée $j^{ième}$ proportion de décomposition de la variance de $\hat{6}_k$. Le tableau de décomposition des variances de BKW est constitué des éléments $(\pi^*_{j,k})_{(j,k) \in \{1, 2,..., K\}^2}$. On note Π^* la matrice de ces éléments.

Remarque : Supposons que λ^*_j soit proche de zéro. Alors $V(\hat{\beta}^*_j)$ est élevée et peut peser fortement sur $V(\hat{6}_k)$ si la valeur de $(Q^*_{k,j})^2$ est assez grande. Supposons que ce soit le cas pour deux estimateurs $\hat{6}_k$ et $\hat{6}_\ell$, $k \neq \ell$. Ceci se traduit par des valeurs élevées de $\pi^*_{j,k}$ et $\pi^*_{j,\ell}$. $\hat{6}_k$ et $\hat{6}_\ell$ sont affectés d'une forte part de variabilité générée par une direction commune, celle de la $j^{ième}$ composante principale de X^*, ce qui laisse supposer un fort degré de multicolinéarité entre les vecteurs X_k et X_ℓ. Cette remarque guide la règle d'interprétation la plus célèbre des indicateurs de BKW, qui va être énoncée et relativisée dans le chapitre suivant.

II- La détection de la multicolinéarité à l'aide des indicateurs de BKW

II-1 La règle d'interprétation initiale de BKW

Dans le chapitre III-2 de leur ouvrage de référence (Cf. BKW [1980] pp. 108-109 et 112), BKW énoncent la règle d'interprétation suivante de leurs indicateurs, notée RI (pour règle initiale) :

Règle RI :

1) Supposons qu'au moins un indice de conditionnement η^*_j , $j \in \{1, 2,..., K\}$ soit élevé (supérieur à 30 environ).

2) Si, sur une ligne j correspondant à η_j^* élevé, *au moins deux* indices k et ℓ, $k \neq \ell$, sont tels que $\pi_{j,k}^*$ et $\pi_{j,\ell}^*$ soient élevés (en pratique supérieurs à 0,5 environ), alors $\hat{6}_k$ et $\hat{6}_\ell$, les estimateurs des MCO des coefficients des vecteurs X_k et X_ℓ, peuvent être entachés d'un problème de multicolinéarité reflétant une forte liaison entre X_k et X_ℓ.

La configuration décrite dans la règle RI correspond effectivement à une situation de multicolinéarité. Des calculs effectués sur des modèles à deux variables explicatives (constante incluse, quand il y en a une) la justifient pleinement dans ce cadre particulier (Cf. Rousse [1990] ou Erkel-Rousse [1994]).

Cependant, on se propose de montrer que **la règle RI appliquée au cadre de modèles à plus de deux variables explicatives omet un grand nombre de situations pourtant associées à des problèmes de multicolinéarité aigus.**

La première partie de la règle RI n'est pas remise en cause. Rappelons seulement que la prudence est de mise pour des indices compris entre 20 et 30, qui sont ambigus (Cf. annexe et aussi BKW [1980] et Belsley [1991]). C'est la seconde partie de la règle RI qui manque de généralité :

- la configuration qui y est décrite ne rend compte de manière exhaustive que des situations de multicolinéarité d'ordre 1, c'est-à-dire des cas où seul l'indice de conditionnement maximal est élevé. C'est pourquoi elle ne pose pas de problème lorsqu'elle est appliquée à des modèles à deux variables explicatives seulement ;

- en revanche, dans un modèle à plus de deux variables explicatives, lorsque p indices de conditionnement exactement $(\eta_j^*)_{j \geq K-p+1}$, $p \in \{1, 2,..., K-1\}$ sont élevés (multicolinéarité d'ordre p), la règle RI peut ne désigner qu'une partie des vecteurs colonnes de X impliqués dans la multicolinéarité. Supposons qu'**au moins deux éléments différents** $(\pi_{j,k}^*)_{k \in K_I}$, $\text{card}(K_I) > 2$ d'une **même ligne** $j \geq K-p+1$ du tableau Π^* soient élevés. Il est en général inexact de considérer que seuls les vecteurs (X_k) $k \in K_I$ sont fortement liés, et que par conséquent seuls les estimateurs des coefficients b_k, $k \in K_I$ peuvent être estimés de manière peu robuste.

II-2 Définition d'une règle de lecture rendant compte des situations de multicolinéarité d'ordre supérieur à 1

Percevant les insuffisances de la règle RI, BKW (1980) en ont proposé des amendements dans le sens d'un examen de sommes partielles sur les lignes d'éléments du tableau de proportions des variances, dans deux cas présentés comme "particuliers" :

- les situations dites "à dépendances concurrentes" (*competing dependencies*), où deux indices de conditionnement au moins $(\eta_j^*)_{j \in I}$ ont des valeurs à la fois élevées et très proches. BKW exhibent alors des exemples illustrant le risque que les proportions des variances des estimateurs $(\pi_{j,k}^*)_{j \in I}$, $\forall\, k \in \{1, 2,..., K\}$ soient mal réparties entre les axes $j \in I$;

- les situations dites "à dépendances dominantes" (*dominating dependencies*), où deux indices de conditionnement au moins $(\eta_j^*)_{j \in I}$ ont des valeurs élevées, mais d'ordres de grandeur très différents. Il peut être alors délicat de discerner les problèmes de multicolinéarité non dominants.

BKW [1980] avaient donc tous les éléments pour énoncer une règle générale englobant *a priori* les configurations décrites dans la règle RI ainsi que celles qu'ils présentent comme des cas particuliers. Malheureusement, ils ont adopté une présentation éclatée de leurs résultats, à notre avis lourde de conséquences pour la transmission de leurs travaux. La règle RI est d'abord introduite comme une règle générale, sans aucune référence aux cas nécessitant qu'y soient apportés des amendements (Cf. BKW pp. 108-109 et 112). Ceux-ci ne sont présentés qu'une quarantaine de pages plus loin dans le cadre d'exemples numériques (p 143). Un bilan ultérieur regroupe la règle RI et ses amendements dans les cas particuliers évoqués, mais un peu tard (pp. 152-155) et sous une forme très peu synthétique.

En dehors de ces aspects formels, BKW ont, nous semble-t-il, envisagé à tort les situations nécessitant des amendements comme des cas particuliers, alors même que ceux-ci ne sont que deux cas extrêmes d'un continuum de situations de multicolinéarité d'ordre strictement supérieur à 1, fréquemment observables dans des modèles à plus de deux variables explicatives. Dans des ouvrages ou articles ultérieurs (publiés de 1982 à 1991), D.A. Belsley reprend cette présentation particulièrement trompeuse pour des lecteurs pratiquant un examen rapide et non exhaustif de ses études. Plutôt que de raisonner sur la règle RI, qui semble incomplète et source d'erreurs, il nous semble préférable d'énoncer directement une règle générale, valable quel que soit le nombre de variables explicatives et l'ordre de multicolinéarité du modèle.

Règle générale RG :

1) En pratique, le seuil à partir duquel un indice de conditionnement peut être considéré comme élevé se situe à 30 environ, mais des valeurs un peu inférieures (à partir de la vingtaine) sont ambiguës, et doivent donc également attirer l'attention.

2) Supposons que p indices de conditionnement $(\eta_j^*)_{j \geq K-p+1}$,
$p \in \{1, 2,...,K-1\}$ exactement soient "élevés". Alors :

- les indices $k \in \{1, 2,..., K\}$ tels que $\sum_{j=K-p+1}^{K} \pi^*_{j,k}$ est "petite" (inférieure à 0,5 ou 0,6 environ) correspondent à des vecteurs X_k non impliqués dans des relations de quasi-colinéarité. Les estimateurs des MCO des coefficients b_k associés à ces vecteurs X_k ne sont pas affectés par la multicolinéarité ;

- les indices $k \in \{1, 2,...,K\}$ tels que $\sum_{j=K-p+1}^{K} \pi^*_{j,k}$ est proche de 1 (supérieure à 0,5 ou 0,6 environ) correspondent aux vecteurs X_k impliqués dans des relations de quasi-colinéarité. Les coefficients b_k associés à ces vecteurs X_k peuvent être estimés de manière très imprécise par leur estimateur des MCO $\hat{b}_k$;

- plus la somme $\sum_{j=K-p+1}^{K} \pi^*_{j,k}$ est proche de 1, plus le diagnostic est préoccupant pour la précision de $\hat{b}_k$. En outre, cette somme étant fixée, la précision de $\hat{b}_k$ est d'autant plus faible que les $\pi^*_{j,k}$ les plus élevés sont situés dans les lignes les plus basses du tableau de décomposition des variances (indices j proches de K).

La règle RG prévoit bien toutes les configurations possibles de multicolinéarité. Elle permet de conclure dans les deux cas "particuliers" décrits par BKW (1980) de "dépendances concurrentes" ou "dominantes". Elle décrit en outre tout le continuum possible dans les cas de multicolinéarité d'ordre multiple entre les situations de "dépendances concurrentes" et de "dépendances dominantes", et permet ainsi une interprétation correcte et plus aisée des indicateurs de BKW.

II-3 Exemple

On a simulé le modèle suivant à trois variables explicatives et 40 observations :

$$y = X_1 + 3X_2 + 2X_3 + u \quad (4)$$

où :

$$X_1 = 2X_2 + v$$
$$X_3 = X_1 - 3X_2 + w$$

Le vecteur u des perturbations est un vecteur normal centré de matrice de variances-covariances l'idendité I_{40} :

$$u \rightarrow N(0, I_{40})$$

$v \perp X_2$, $w \perp X_1$, $w \perp X_2$, $0 < ||v|| << ||X_k||$ et $0 < ||w|| << ||X_k||$, $\forall k \in \{1, 2, 3\}$, tels que :

- angle entre les vecteurs X_1 et X_2 : 0,2°,

- angle entre les vecteurs X_1 et X_3 : 117°,

- angle entre X_3 et sa projection orthogonale sur le sous-espace $\mathcal{L}(X_1,X_2)$: 2,9°

- v est le résultat de la projection orthogonale d'un vecteur v_1 (issu de 40 tirages indépendants dans une même loi uniforme centrée) sur le sous-espace orthogonal à $\mathcal{L}(X_2)$, X_2 étant issu de 40 tirages indépendants dans une loi uniforme de variance supérieure.

- w est le résultat de la projection orthogonale d'un vecteur w_1 (issu de 40 tirages indépendants dans une même loi uniforme centrée) sur le sous-espace orthogonal à $\mathcal{L}(X_1,X_2)$, X_1 étant calculé à partir de X_2 et de v.

- Les normes de v et w ont été calculées en fonction des angles entre les trois vecteurs colonnes de X fixés *a priori* de sorte que le modèle soit affecté de deux types de multicolinéarité, dont les conséquences sur la qualité des estimations seront un peu différentes. Plus précisément, le tableau en annexe donne une évaluation de l'angle entre les directions d'un vecteur X_k et de sa projection orthogonale sur l'espace vectoriel engendré par les autres vecteurs colonnes de X en deçà duquel on peut parler de situation de multicolinéarité. Cet angle semble être de l'ordre de 4 ou 5°. Dans le présent exemple, les trois vecteurs des variables explicatives sont donc fortement liés entre eux par deux "quasi-relations" linéaires indépendantes. On dit que le modèle est affecté d'une multicolinéarité d'ordre 2. Ce sont les vecteurs X_1 et X_2 qui sont les plus fortement liés. Ils seront responsables d'une multicolinéarité dite "primaire", première source des difficultés d'estimation. Le vecteur X_3 est davantage séparé des deux autres vecteurs, mais il est cependant peu éloigné de sa projection orthogonale sur le sous-espace $\mathcal{L}(X_1,X_2)$. Ceci engendrera une multicolinéarité dite "secondaire", dont les conséquences, moins fortes que celles de la multicolinéarité primaire mais néanmoins néfastes, ne seront pas identifiées par la règle RI.

Cette multicolinéarité d'ordre 2, qui implique les trois vecteurs X_1, X_2 et X_3, se traduit par des estimations fort médiocres des trois paramètres associés. En témoigne le tableau des résultats de l'estimation par les MCO dans le modèle (4) :

Variable explicative	Vraie valeur du paramètre	Valeur estimée du paramètre	Écart-type estimé	T de Student
X_1	$b_1 = 1$	$\hat{b}_1 = 3,22$	$\hat{\sigma}_{\hat{b}_1} = 7,0$	$t_{\hat{b}_1} = 0,46$
X_2	$b_2 = 3$	$\hat{b}_2 = -0,76$	$\hat{\sigma}_{\hat{b}_2} = 15,2$	$t_{\hat{b}_2} = -0,05$
X_3	$b_3 = 2$	$\hat{b}_3 = 2,70$	$\hat{\sigma}_{\hat{b}_3} = 0,8$	$t_{\hat{b}_3} = 3,38$

Les indicateurs de BKW prennent les valeurs suivantes :

Valeur propre	Indice de conditionnement	Tableau de décomposition des variances		
		$V(\hat{b}_1)$	$V(\hat{b}_2)$	$V(\hat{b}_3)$
$\lambda_1^* = 3$	$\eta_1^* = 1$	$\pi_{1,1}^* = 0^+$	$\pi_{1,2}^* = 0^+$	$\pi_{1,3}^* = 0^+$
$\lambda_2^* = 2.10^{-3}$	$\eta_2^* = 42$	$\pi_{2,1}^* = 10^{-3}$	$\pi_{2,2}^* = 10^{-3}$	$\pi_{2,3}^* = 0{,}980$
$\lambda_3^* = 4.10^{-6}$	$\eta_3^* = 857$	$\pi_{3,1}^* = 0{,}999$	$\pi_{3,2}^* = 0{,}999$	$\pi_{3,3}^* = 2.10^{-2}$

Ce sont les paramètres associés aux deux premières variables explicatives qui sont estimés avec la plus faible précision. C'est normal puisque l'examen des angles entre vecteurs colonnes de X montre que X_1 et X_2 sont plus colinéaires que X_3 et sa projection sur le sous-espace $\mathcal{L}(X_1,X_2)$. Cependant X_3 est touché par une multicolinéarité secondaire, et son coefficient n'est pas estimé avec une grande précision (l'estimation 2,7 de $b_3 = 2$ est entachée d'une erreur relative de 35%). Or, si on se fonde sur la règle RI, seul le problème de multicolinéarité primaire portant sur X_1 et X_2 est détecté. La statistique de Student associée à $\hat{b}_3$ ne permet pas de corriger le diagnostic.

Cependant, les indicateurs de BKW apportent toute l'information nécessaire quand ils sont interprétés selon la règle RG, ou en se souvenant de l'aménagement de la règle RI en cas de "dépendance dominante" :

- d'une part deux indices de conditionnement sont élevés, ce qui indique une situation de multicolinéarité d'ordre 2 entre les trois vecteurs. La forte part de variabilité des estimateurs des deux premiers paramètres portée par la dernière composante principale, celle qui est associée à la direction la moins structurante du sous-espace $\mathcal{L}(X)$, montre que la multicolinéarité la plus préjudiciable est celle qui touche X_1 et X_2 (diagnostic de la multicolinéarité primaire) ;

- la part de variabilité de $\hat{b}_3$ portée par les deux directions peu structurantes atteint la valeur $\pi_{2,3}^* + \pi_{3,3}^*$ proche de 100%, ce qui montre que X_3 est également impliqué dans la multicolinéarité. Mais il s'agit d'une multicolinéarité moins pathologique, comme l'indique le fait que $\pi_{3,3}^*$ est faible (diagnostic de la multicolinéarité secondaire). L'estimation de b_3 évite ainsi le pire, qui serait d'être fortement déterminée par la dernière composante principale, la moins structurante. Ceci se retrouve dans l'estimation certes médiocre et inutilisable de b_3 mais moins catastrophique que celles des deux autres paramètres. On doit y voir la manifestation de la multicolinéarité secondaire, très préjudiciable, mais moins prononcée que celle de la multicolinéarité primaire.

Remarque très importante en pratique :

Dans les brochures "SAS User's Guide Statistics" version 5 (page 672) et "SAS/STAT User's Guide" version 6 (pp. 1416-1417), seule la règle RI est énoncée pour guider

l'utilisateur de la procédure REG dans son interprétation des diagnostics de multicolinéarité (option COLLIN). Ce faisant, la plupart des praticiens utilisateurs de SAS risquent d'omettre un nombre important de situations de multicolinéarité.

II-4 L'absence de "contamination" d'une multicolinéarité localisée

La règle RG rappelle explicitement **une propriété importante concernant la portée de la multicolinéarité**. Supposons que certains vecteurs $(X_k)_{k \in K_M}$ des variables explicatives soient fortement liés, et que ceci se traduise par une situation de multicolinéarité préjudiciable pour la qualité des estimations.
Dans ce cas, les coefficients $(b_k)_{k \notin K_M}$ des variables explicatives non impliquées dans ces relations ne sont pas affectés par la multicolinéarité. Autrement dit, **il n'y a pas "contamination" des effets néfastes de la multicolinéarité à l'ensemble des estimations**. Plus généralement, les estimations des fonctions des paramètres qui seraient identifiables si les relations entre les vecteurs $(X_k)_{k \in K_M}$ étaient exactes (et non plus approchées) ne sont pas affectées par la multicolinéarité touchant les $(X_k)_{k \in K_M}$.
C'est ce que la règle RG souligne, dans l'énoncé du 2), premier tiret.

Cette propriété est logique, mais elle ne va pas obligatoirement de soi, si l'on se borne à un examen rapide de la formule :

$$\forall\, k \in \{1, 2, \ldots, K\}, \quad V(\hat{b_k}) = \frac{\sigma^2}{\left|\left|X_k\right|\right|^2} \sum_{j=1}^{K} \frac{(Q^*_{k,j})^2}{\lambda^*_j}$$

Certains auteurs en déduisent en effet qu'une multicolinéarité localisée peut affecter les estimations de la totalité des paramètres du premier ordre, une faible valeur de λ^*_j pesant sur la variance de *tous* les estimateurs $(\hat{b_k})_{k=1,\ldots,K}$. Cet argument n'est pas exact, car il ne prend pas en compte les variations des termes multiplicatifs $(Q^*_{k,j})^2$ en fonction de λ^*_j. Pour s'en convaincre, il suffit de reformuler chaque $V(\hat{b_k})$ en fonction de l'angle θ_k entre le vecteur X_k et sa projection orthogonale sur le sous-espace $\mathcal{L}(X^c_k)$ engendré par la matrice X^c_k des vecteurs colonnes de X dont on a retiré X_k (résultat classique, cf. par exemple Erkel-Rousse [1994] pour une démonstration) :

$$\forall\, k \in \{1, 2, \ldots, K\}, \quad V(\hat{b_k}) = \frac{\sigma^2}{\left|\left|X_k\right|\right|^2} \sum_{j=1}^{K} \frac{(Q^*_{k,j})^2}{\lambda^*_j} = \frac{\sigma^2}{\left|\left|X_k\right|\right|^2 (1 - \cos^2(\theta_k))}$$

En particulier, si X_k est orthogonal à X_k^c, on a :

$$V(\hat{b}_k) = \frac{\sigma^2}{||X_k||^2} \qquad \text{où : } \hat{b}_k = \frac{X_k'y}{||X_k||^2}$$

La somme $\sum_{j=1}^{K} \frac{(Q^*_{k,j})^2}{\lambda^*_j}$ est ici égale à 1, et ce, quel que soit l'ordre de multicolinéarité associé à la matrice X_k^c.

Dans le cas général cette somme, égale à l'inverse de $1-\cos^2(\theta_k)$, n'est rien d'autre que le $k^{ième}$ facteur d'inflation de la variance dû à la non-orthogonalité de X_k avec X_k^c, $1-\cos^2(\theta_k)$ correspondant à la tolérance.

La somme $\sum_{j=1}^{K} \frac{(Q^*_{k,j})^2}{\lambda^*_j}$ est donc une fonction continûment décroissante de l'angle que la direction du vecteur X_k forme avec le sous-espace $\mathcal{L}(X_k^c)$. La variance $V(\hat{b}_k)$ ne dépend de X_k^c qu'à travers l'angle θ_k entre X_k et sa projection orthogonale sur $\mathcal{L}(X_k^c)$. Si cet angle est suffisamment grand, autrement dit si X_k est extérieur à toute multicolinéarité éventuelle, alors peu importe que certains vecteurs colonnes de X_k^c soient ou non fortement liés à l'intérieur du sous-espace $\mathcal{L}(X_k^c)$. À norme du vecteur X_k fixée et précision du modèle σ^2 donnée, la variance $V(\hat{b}_k)$ ne devient élevée que si l'angle θ_k devient petit, c'est-à-dire si X_k lui-même s'approche de $\mathcal{L}(X_k^c)$ et entre ainsi dans une relation de multicolinéarité avec X_k^c.

Remarque :

Si les variables sont centrées, $\cos^2(\theta_k)$ est identique au coefficient de détermination (ou encore au carré du coefficient de corrélation multiple) de X_k par rapport à X_k^c.

Exemple :

On simule le modèle suivant à quatre variables explicatives et 40 observations :

$$y = X_1 + 3X_2 + 2X_3 - 5X_4 + u$$

$$\text{où : } u \rightarrow N(0,I_{40}) \quad \text{et} \quad \begin{cases} X_1 = 2X_2 + Z_1 & Z_1 \perp X_2 \\ X_3 = X_1 - 3X_2 + v & v \perp \mathcal{L}(X_1,X_2) \\ X_4 = X_1 - 2X_2 + 0{,}5X_3 + \beta Z_2 & Z_2 \perp \mathcal{L}(X_1,X_2,X_3) \end{cases}$$

$0 < ||v|| << ||X_1 - 3X_2||$ (première source de multicolinéarité, constante), tel que :

L'angle entre X_3 et sa projection orthogonale sur le sous-espace $\mathcal{L}(X_1,X_2)$ est de : 1,9°.
L'angle entre les vecteurs X_1 et X_2 est de 30° (Z_1 n'est pas de norme très petite devant X_2).
L'angle entre les vecteurs X_1 et X_3 est de 101°.
L'angle entre les vecteurs X_2 et X_3 est de 131°.

X_1, X_2, X_3, Z_1 et Z_2 sont des vecteurs fixés une fois pour toutes. X_2 a été obtenu par 40 tirages indépendants dans une loi uniforme. Z_1 résulte de la projection sur l'orthogonal de $\mathcal{L}(X_2)$ d'un vecteur issu de 40 tirages indépendants dans une autre loi uniforme. La norme de Z_1 n'est pas suffisamment petite pour que X_1 et X_2 soient fortement colinéaires. X_1 est calculé à partir de X_2 et de Z_1 et v résulte de la projection sur l'orthogonal de $\mathcal{L}(X_1,X_2)$ d'un vecteur issu de 40 tirages indépendants dans une même loi uniforme centrée de faible variance. La norme de v a été ajustée en fonction de l'angle entre X_3 et sa projection orthogonale sur le sous-espace $\mathcal{L}(X_1,X_2)$, fixé *a priori* de sorte que le modèle soit affecté systématiquement d'au moins une source de multicolinéarité. Z_2 résulte de la projection sur l'orthogonal de $\mathcal{L}(X_1,X_2,X_3)$ d'un vecteur issu de 40 tirages indépendants dans une loi uniforme.

D'une simulation à une autre, on fait varier l'angle de X_4 avec le sous-espace vectoriel $\mathcal{L}(X_1,X_2,X_3)$ via la valeur de β. Lorsque β est grand, le modèle n'est affecté que de la multicolinéarité d'ordre 1 touchant systématiquement X_1, X_2 et X_3. Lorsque β est petit, X_4 peut entrer dans la relation de quasi-colinéarité liant les autres vecteurs d'exogènes, augmentant ainsi l'ordre de la multicolinéarité, qui passe à 2.

Le tableau de la page suivante présente les résultats d'une partie des simulations effectuées sur ce modèle par Rousse [1990]. Ces résultats appellent les commentaires suivants :

- Quelle que soit la valeur de β, la qualité des estimations de b_1, b_2 et b_3 est gravement affectée par la multicolinéarité touchant X_1, X_2 et X_3. Ce problème est détecté par les valeurs systématiquement élevées de l'indice de conditionnement maximal η_4^* et des proportions π_{41}^*, π_{42}^* et π_{43}^* de décomposition des variances de $\hat{b}_1$, $\hat{b}_2$ et $\hat{b}_3$ sur la quatrième composante principale.

- Il est plus intéressant d'observer la dégradation régulière de la qualité de l'estimation de b_4 au fur et à mesure que le vecteur X_4 se rapproche du sous-espace vectoriel $\mathcal{L}(X_1,X_2,X_3)$.

  Pour $\beta = 100$, X_4 est quasiment orthogonal à ce sous-espace, et l'estimation de b_4 n'est absolument pas affectée par la forte liaison entre X_1, X_2 et X_3. La variance de $\hat{b}_4$ est entièrement portée par la seconde composante principale, orthogonale à la quatrième, sur laquelle repose la totalité des variances des autres estimateurs.

Pour des valeurs de β comprises entre 0,30 et 0,08, l'angle entre X_4 et $\mathcal{L}(X_1,X_2,X_3)$ diminue progressivement tout en restant supérieur à la valeur de 5° à partir de laquelle X_4 entrerait dans une relation de quasi-colinéarité avec les autres vecteurs des exogènes. L'erreur relative affectant l'estimation de b_4 augmente progressivement tout en restant acceptable. L'indice de conditionnement η_3^* rend compte

Résultats d'une partie des simulations sur le modèle : $y = X_1 + 3X_2 + 2X_3 - 5X_4 + u$

	Degré de multicolinéarité							
Valeur du coefficient β	100,0	0,30	0,15	0,08	0,05	0,04	0,03	0,01
Angle de X_4 avec $\mathcal{L}(X_1,X_2,X_3)$	89,6°	22,9°	11,9°	6,4°	4,8°	3,2°	2,4°	0,8°
	Estimation des paramètres du premier ordre							
$\hat{b}_1$ ($b_1 = 1$) ($t_{\hat{b}_1} \leq 0{,}6\ \forall\ \beta$)	0,3	0,4	0,4	0,5	0,7	0,8	0,9	2,2
$\hat{b}_2$ ($b_2 = 3$) ($t_{\hat{b}_2}$)	5,0 (2,0)	4,9 (2,0)	4,8 (1,9)	4,6 (1,8)	4,3 (1,5)	4,1 (1,3)	3,8 (1,1)	1,4 (0,2)
$\hat{b}_3$ ($b_3 = 2$) ($t_{\hat{b}_3}$)	2,7 (3,4)	2,7 (3,3)	2,6 (3,2)	2,6 (3,1)	2,5 (2,8)	2,5 (2,6)	2,4 (2,3)	1,8 (0,9)
$\hat{b}_4$ ($b_4 = -5$) valeur absolue de $t_{\hat{b}_4}$	− 5,0 (10^4)	− 5,1 (40,0)	− 5,1 (20,3)	− 5,2 (11,0)	− 5,4 (7,1)	− 5,5 (5,8)	− 5,6 (4,4)	− 6,9 (1,8)
	Indices de conditionnement non triviaux							
η_2^*	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5
η_3^*	1,6	5,7	10,8	18,8	26,6	30,4	34,7	42,5
η_4^*	119	133	137	147	165	181	211	516
	Tableau de décomposition des variances							
de $\hat{b}_k$, $\forall\ k < 4$:								
$\sum_{j=1}^{3} \pi_{jk}^* \leq$	0,001	0,003	0,007	0,020	0,038	0,047	0,055	0,023
$\pi_{4k}^* \geq$	0,999	0,997	0,993	0,980	0,962	0,953	0,945	0,977
de $\hat{b}_4$:								
π_{14}^*	0^+	0,014	0,004	0,001	0^+	0^+	0^+	0^+
π_{24}^*	0,999	0,029	0,008	0,002	0,001	0,001	0^+	0^+
π_{34}^*	0^+	0,943	0,935	0,829	0,649	0,536	0,386	0,061
π_{44}^*	0^+	0,014	0,053	0,168	0,350	0,463	0,613	0,939

Nota : $R_c^2 = \dfrac{||\hat{y}||^2}{||y||^2} > 0{,}9.$

du fait que X_4 se rapproche du sous-espace $\mathcal{L}(X_1,X_2,X_3)$ en s'élevant régulièrement, tout en restant inférieur à 20. Ces situations, où X_4 n'est pas impliqué dans la multicolinéarité, vont ainsi de pair avec des estimations de b_4 satisfaisantes.

Pour des valeurs de β inférieures à 0,08, l'angle de X_4 avec le sous-espace $\mathcal{L}(X_1,X_2,X_3)$ passe en deçà de 5°. L'erreur relative sur l'estimation de b_4 augmente, en franchissant le cap de 10% pour $\beta \leq 0,04$. Le cas $\beta = 0,05$ correspond à une zone limite vis-à-vis de l'implication de X_4 dans la relation de quasi-colinéarité liant les vecteurs des exogènes, traduite par une valeur de η_3^* comprise entre 20 et 30. Lorsque β devient inférieur à 0,05, X_4 entre sans ambiguïté dans une relation de quasi-colinéarité avec les autres vecteurs. Cette relation devient pathologique pour $\beta = 0,01$, comme le montrent les valeurs de η_3^* et de l'angle de X_4 avec $\mathcal{L}(X_1,X_2,X_3)$. L'erreur relative affectant l'estimation de b_4 est alors très élevée (près de 40%). L'examen de la décomposition de la variance de $\hat{b}_4$ montre que la part de sa variance portée par la (puis les) directions entachée(s) de multicolinéarité augmente progressivement ainsi que la proportion π_{44}^*. Lorsque β devient inférieur à 0,04, X_4 passe ainsi progressivement d'une implication dans une multicolinéarité "secondaire" à une implication dans une multicolinéarité "primaire". Ceci se traduit dans la chute continue de la précision de $\hat{b}_4$. Conformément à la théorie, la précision de $\hat{b}_4$ ne se dégrade donc profondément qu'à partir du moment où X_4 entre lui-même dans la relation de quasi-colinéarité. On observe en outre un continuum dans la gravité des conséquences de la multicolinéarité entre X_1, X_2 et X_3 sur la qualité de $\hat{b}_4$ au fur et à mesure que X_4 se rapproche de $\mathcal{L}(X_1,X_2,X_3)$.

Dans tous les exemples *supra*, les indicateurs de BKW se sont avérés très éclairants, ce qui est logique car ils sont spécifiquement adaptés à la détection de multicolinéarité. Soulignons que le seul examen des statistiques de Student usuelles ne constituerait en aucun cas un palliatif valable à l'absence d'une détection spécifique de la multicolinéarité (Erkel-Rousse [1994]). Pour autant, du fait de certaines propriétés de non invariance, en particulier par centrage, il arrive que les indicateurs de BKW posent des problèmes d'interprétation.

III- Conséquences de la non-invariance des indicateurs de BKW par centrage : faut-il centrer ou non les variables explicatives véritables ?

Les indicateurs de BKW sont modifiés par toute opération impliquant une variation des directions relatives des vecteurs . Si cette propriété semble souhaitable *a priori*, une de ses conséquences - la non-invariance par centrage - pose des problèmes en matière d'inter-

prétation des indicateurs de BKW. Cette difficulté a donné lieu à un débat passionné - loin d'être clos - sur l'opportunité de raisonner ou non sur des modèles centrés.

III-1 Le problème : Invariance des estimateurs des MCO et non-invariance des indicateurs de BKW par centrage

Soit le modèle linéaire ordinaire avec constante :

$$\underset{(N,1)}{y} = \underset{(N,1)}{e}\,\underset{(1,1)}{c} + \underset{(N,K-1)}{Z}\,\underset{(K-1,1)}{a} + \underset{(N,1)}{u} \qquad (5)$$

où e est le vecteur unité, Z la matrice des variables explicatives véritables, c la constante et a le vecteur des paramètres des variables explicatives véritables. X = (e|Z) est de plein rang colonnes. L'estimateur â des MCO de a dans le modèle "non centré" (5) s'obtient aussi en effectuant les MCO dans le modèle "centré" :

$$\underset{(N,1)}{y} = \underset{(N,K-1)}{\tilde{Z}}\,\underset{(K-1,1)}{a} + \underset{(N,1)}{\tilde{u}} \qquad (6)$$

où $\tilde{Z}$ est la matrice des vecteurs colonnes de Z centrés (théorème de Frisch-Waugh).

Si les estimateurs des MCO de a sont identiques dans les deux modèles, les indicateurs de BKW ne le sont pas : on montre que les indices de conditionnement sont plus élevés dans le modèle non centré que dans le modèle centré (Cf. Belsley [1991] pp. 201-204). Supposons que les indicateurs de BKW calculés sur le modèle non centré suggèrent que â est affecté d'un problème de multicolinéarité alors que les indicateurs calculés sur le modèle centré tendraient à indiquer le contraire. Que peut-on conclure ?

III-2 Le débat : faut-il raisonner sur le modèle initial ou sur le modèle centré ?

Depuis la publication de l'ouvrage de BKW (1980), cette question fait l'objet d'un débat théorique contradictoire entre les partisans du centrage et ses opposants.

Pour de nombreux auteurs (Cf. Marquardt [1980], Weisberg [1980], Montgomery et Peck [1982], Gunst [1983] ou Stewart [1987], par exemple ; voir aussi Askin [1982], Snee et Marquardt [1984], Thisted [1987]), la constante est introduite dans un modèle pour un motif purement technique - le centrage des résidus d'estimation, le paramètre d'intérêt étant le vecteur a des coefficients des variables explicatives véritables. Le vecteur e joue donc un rôle très spécifique par rapport aux autres vecteurs colonnes

de X. En conséquence, raisonner sur le modèle centré, qui isole le paramètre a et dissymétrise les statuts de la constante et des autres variables explicatives, semblerait pertinent. La baisse du degré de multicolinéarité lorsqu'on passe du modèle non centré au modèle centré correspondrait à la suppression des liaisons du vecteur constant et des niveaux des autres vecteurs, liaisons considérées comme "non essentielles" par opposition aux corrélations entre vecteurs des variables explicatives véritables. Cette vision assimile la multicolinéarité à des *corrélations* fortes entre variables explicatives.

Belsley, Kuh et Welsch s'opposent avec vigueur à cette vision des choses (Cf. BKW [1980], Belsley [1984, 1986, 1991]), ainsi que Lesage et Simon [1985] et Simon et Lesage [1988]. Selon Belsley [1991] par exemple, il serait faux de penser que, parce que la constante ne peut être corrélée à quoi que ce soit, elle ne peut être colinéaire avec quoi que ce soit. Il faudrait raisonner non pas en termes de "variables" (dont la constante ne fait pas véritablement partie) mais en termes de "vecteurs" (dont le vecteur e fait partie). Ainsi, la constante devrait être conservée lors d'analyses de la multicolinéarité. En outre, le centrage des vecteurs colonnes de Z serait à proscrire, car il affecte l'ampleur des effets sur les estimations des paramètres d'une modification relative des données (contrairement par exemple à la normalisation de la matrice X en X^*, qui a la faveur de BKW). Or, ce sont ces effets que l'on chercherait à appréhender. L'indice de conditionnement maximal issu du modèle centré ne répondrait donc pas au bon problème. Belsley [1991] présente à cet égard un exemple numérique où l'indice de conditionnement maximal issu du modèle non centré est très élevé, tandis que l'indice de conditionnement maximal issu du modèle centré est peu élevé (pp. 178-183). Or, dans cet exemple, une faible modification des données engendre une très forte variation de l'estimation de a (dans les deux modèles). La matrice X initiale était donc mal conditionnée, et le diagnostic issu du modèle centré trompeur.

III-3 Une réponse nuancée sur la question de l'opportunité du centrage : détection des "effets taille" par un double diagnostic de multicolinéarité

Si l'exemple présenté par Belsley est certes troublant, sa force démonstrative n'a pas valeur de généralité. Il en ressort surtout que les indicateurs de BKW ne s'avèrent pas aussi aisés d'utilisation qu'ils pouvaient le sembler. Certains des arguments de Belsley (1991) sont toutefois suffisamment convaincants pour qu'on n'ignore pas le diagnostic obtenu à partir du modèle non centré. Cependant, le refus par Belsley de reconnaître le statut particulier de la constante relève sans doute d'une vision trop "mathématicienne" et pas assez "statisticienne" du modèle linéaire. Surtout, à la lumière de l'expérience pratique de l'économétrie sur données réelles, il nous semble que les indicateurs issus des modèles non centrés tendent à conclure trop souvent à des problèmes de multicolinéarité.

C'est le cas lorsque le modèle est affecté d'un "effet taille". Un tel effet se manifeste lorsque les variables explicatives prennent des valeurs numériques très grandes. C'est une configuration fréquente par exemple en macroéconométrie quantitative, en présence d'agrégats macroéconomiques exprimés en valeur ou à prix constants (volume) parmi les variables explicatives. L'effet taille peut être responsable de diagnostics de multicolinéarité positifs suggérant de fortes liaisons entre la constante et ces agrégats, même si ceux-ci ont une tendance nette sur la période d'estimation. On obtient le même genre de résultat en présence d'un "trend" (i.e. une fonction linéaire du temps) exprimé en années (i.e. 19xx).

Un effet taille suggère que l'économètre a exprimé certaines variables explicatives dans une unité inadéquate, de sorte que l'évolution de ces variables est "écrasée" par leur forte valeur initiale. Il suffit de changer d'unité (exemple : exprimer le temps comme la différence entre l'année courante et l'année centrale de la période d'estimation) pour que le diagnostic de multicolinéarité disparaisse sans que les estimations n'en soient aucunement modifiées, hormis celle de la constante. Dans cette configuration, l'effet taille crée dans le modèle initial un diagnostic de multicolinéarité purement artificiel. Le diagnostic issu du modèle centré est alors moins trompeur, car l'effet taille est corrigé par le centrage. Au total, la comparaison des diagnostics obtenus à partir des modèles initial et centré est intéressante, car elle permet de discriminer les "fausses" situations de multicolinéarité des "vraies".

Exemple (sur données réelles) :

On modélise le volume d'importations françaises de produits manufacturés M en fonction de la demande intérieure DI exprimée aux prix de 1980, d'un terme de compétitivité-prix COMP, d'une tendance croissante OUV représentative de l'ouverture progressive des économies aux échanges[1] sur la période d'estimation 1970-1990 et du taux d'investissement lissé de la France relativement à ses principaux partenaires commerciaux[2,3]. Le modèle s'écrit en logarithmes :

$$\text{Log}(M_t) = \varepsilon_d.\text{Log}(DI_t) + \varepsilon_c.\text{Log}(COMP_t) + \tau.OUV_t + \varepsilon_i.\text{Log}(INV_t) + c + u_t$$

avec : $E(u_t) = 0$, $V(u_t) = \sigma^2$, $\text{Cov}(u_t,u_{t'}) = 0$
$\forall\ t,t' \in \{1970,..., 1990\}$, tels que $t' \neq t$

(1) Diminution institutionnelle des barrières à l'échange (baisse des tarifs douaniers, suppression des quotas à l'importation, etc.).

(2) Lissage du taux d'investissement relatif sur 6 ans. Terme représentatif de la compétitivité structurelle, qui est un phénomène de moyen terme. Cf. Erkel-Rousse (1992).

(3) Estimations effectuées sur données annuelles dans une approche d'économétrie traditionnelle.

L'estimation de ce modèle par les MCO donne les résultats suivants :

Variable explicative	Valeur estimée du paramètre	Écart-type estimé	T de Student
Constante	$\hat{c} = -5,9$	$\hat{\sigma}_{\hat{c}} = 1,16$	$t_{\hat{c}} = -5,1$
$Log(DI_t)$	$\hat{\varepsilon}_d = 1,3$	$\hat{\sigma}_{\hat{\varepsilon}_d} = 0,08$	$t_{\hat{\varepsilon}_d} = 16,1$
$Log(COMP_t)$	$\hat{\varepsilon}_c = -1,0$	$\hat{\sigma}_{\hat{\varepsilon}_c} = 0,16$	$t_{\hat{\varepsilon}_c} = -6,1$
OUV_t	$\hat{\tau} = 0,02$	$\hat{\sigma}_{\hat{\tau}} = 2.10^{-3}$	$t_{\hat{\tau}} = 14,7$
$Log(INV_t)$	$\hat{\varepsilon}_i = -0,7$	$\hat{\sigma}_{\hat{\varepsilon}_i} = 0,13$	$t_{\hat{\varepsilon}_i} = -5,3$

Le diagnostic de multicolinéarité est le suivant, sur le modèle initial (non centré) :

Valeur propre	Indice de conditionnement	Tableau de décomposition des variances				
		$V(\hat{c})$	$V(\hat{\varepsilon}_d)$	$V(\hat{\varepsilon}_c)$	$V(\hat{\tau})$	$V(\hat{\varepsilon}_i)$
$\lambda_1^* = 2,3$	$\eta_1^* = 1,0$	$\pi_{1,1}^* = 0$	$\pi_{1,2}^* = 0$	$\pi_{1,3}^* = 0^+$	$\pi_{1,4}^* = 0^+$	$\pi_{1,5}^* = 0^+$
$\lambda_2^* = 1,5$	$\eta_2^* = 1,2$	$\pi_{2,1}^* = 0$	$\pi_{2,2}^* = 0$	$\pi_{2,3}^* = 0,1^-$	$\pi_{2,4}^* = 0^+$	$\pi_{2,5}^* = 0^+$
$\lambda_3^* = 1,0$	$\eta_3^* = 1,5$	$\pi_{3,1}^* = 0$	$\pi_{3,2}^* = 0$	$\pi_{3,3}^* = 0^+$	$\pi_{3,4}^* = 0$	$\pi_{3,5}^* = 0,7$
$\lambda_4^* = 0,2$	$\eta_4^* = 3,2$	$\pi_{4,1}^* = 0$	$\pi_{4,2}^* = 0$	$\pi_{4,3}^* = 0,6$	$\pi_{4,4}^* = 0,1$	$\pi_{4,5}^* = 0,2$
$\lambda_5^* = 4.10^{-6}$	$\eta_5^* = 792,5$	$\pi_{5,1}^* = 1$	$\pi_{5,2}^* = 1$	$\pi_{5,3}^* = 0,3$	$\pi_{5,4}^* = 0,9$	$\pi_{5,5}^* = 0^+$

et sur le modèle centré :

Valeur propre	Indice de conditionnement	Tableau de décomposition des variances			
		$V(\hat{\varepsilon}_d)$	$V(\hat{\varepsilon}_c)$	$V(\hat{\tau})$	$V(\hat{\varepsilon}_i)$
$\lambda_1^* = 2,7$	$\eta_1^* = 1,0$	$\pi_{1,1}^* = 0^+$	$\pi_{1,2}^* = 0^+$	$\pi_{1,3}^* = 0^+$	$\pi_{1,4}^* = 0^+$
$\lambda_2^* = 1,0$	$\eta_2^* = 1,6$	$\pi_{2,1}^* = 0$	$\pi_{2,2}^* = 0^+$	$\pi_{2,3}^* = 0^+$	$\pi_{2,4}^* = 0,7$
$\lambda_3^* = 0,3$	$\eta_3^* = 3,2$	$\pi_{3,1}^* = 0^+$	$\pi_{3,2}^* = 0,8$	$\pi_{3,3}^* = 0,1$	$\pi_{3,4}^* = 0,3$
$\lambda_4^* = 3.10^{-2}$	$\eta_4^* = 9,6$	$\pi_{4,1}^* = 1^-$	$\pi_{4,2}^* = 0,2$	$\pi_{4,3}^* = 0,9$	$\pi_{4,4}^* = 0^+$

Les indicateurs de BKW calculés sur le modèle initial (non centré) indiquent que le logarithme de la demande intérieure, la variable d'ouverture des frontières et la constante sont fortement colinéaires, et que les estimateurs de leur coefficient peuvent être affectés par la multicolinéarité (c'est-à-dire être très sensibles à une petite modification des données et estimés avec une faible précision). Au contraire, aucune multicolinéarité n'est détectée sur le modèle centré. Que penser alors de la qualité des estimations données par $\hat{\varepsilon}_d$ et $\hat{\tau}$?

Cette contradiction doit alerter et suggérer immédiatement la possibilité d'un effet taille. Dans une application de macroéconomie, comme c'est le cas ici, il suffit que les volumes soient transformés en indices pour supprimer l'effet taille. On considère le modèle initial reparamétré de manière à faire apparaître des indices :

$$\text{Log}\left(\frac{M_t}{M_{1980}}\right) = \varepsilon_d.\text{Log}\left(\frac{DI_t}{DI_{1980}}\right) + \varepsilon_c.\text{Log}(COMP_t) + \tau.OUV_t + \varepsilon_i.\text{Log}(INV_t) + \gamma + u_t$$

où : $\gamma = c - \text{Log}(M_{1980}) + \varepsilon_d.\text{Log}(DI_{1980})$

Les estimateurs des MCO des coefficients associés aux variables explicatives véritables sont inchangés. De même, les éléments de diagnostic calculés sur ce nouveau modèle centré ne sont pas modifiés par le passage des volumes en indices.

En revanche, les indicateurs de BKW sur le nouveau modèle non centré sont modifiés:

Valeur propre	Indice de condition-nement	Tableau de décomposition des variances				
		$V(\hat{c})$	$V(\hat{\varepsilon}_d)$	$V(\hat{\varepsilon}_c)$	$V(\hat{\tau})$	$V(\hat{\varepsilon}_i)$
$\lambda_1^* = 2,8$	$\eta_1^* = 1,0$	$\pi_{1,1}^* = 0^+$	$\pi_{1,2}^* = 0^+$	$\pi_{1,3}^* = 0^+$	$\pi_{1,4}^* = 0^+$	$\pi_{1,5}^* = 0^+$
$\lambda_2^* = 1,0^+$	$\eta_2^* = 1,7^-$	$\pi_{2,1}^* = 0,2$	$\pi_{2,2}^* = 0^+$	$\pi_{2,3}^* = 0^+$	$\pi_{2,4}^* = 0^+$	$\pi_{2,5}^* = 0,6$
$\lambda_3^* = 1,0^-$	$\eta_3^* = 1,7^+$	$\pi_{3,1}^* = 0,6$	$\pi_{3,2}^* = 0^+$	$\pi_{3,3}^* = 0^+$	$\pi_{3,4}^* = 0^+$	$\pi_{3,5}^* = 0,2$
$\lambda_4^* = 0,2$	$\eta_4^* = 3,4$	$\pi_{4,1}^* = 0,1$	$\pi_{4,2}^* = 0^+$	$\pi_{4,3}^* = 0,8$	$\pi_{4,4}^* = 0,1$	$\pi_{4,5}^* = 0,2$
$\lambda_5^* = 3.10^{-2}$	$\eta_5^* = 9,8$	$\pi_{5,1}^* = 0,1$	$\pi_{5,2}^* = 1^-$	$\pi_{5,3}^* = 0,2$	$\pi_{5,4}^* = 0,9$	$\pi_{5,5}^* = 0^+$

Une fois l'effet taille corrigé, les indicateurs de BKW sur le modèle non centré ne détectent plus de multicolinéarité : les diagnostics sur modèles centré et non centré coïncident. Le diagnostic initial de multicolinéarité découlait donc artificiellement d'un effet taille, et s'avérait trompeur.

Bilan :

Contrairement aux avis très tranchés qui s'expriment dans le débat sur l'opportunité du centrage, notre avis est d'établir le diagnostic de la multicolinéarité à partir du double examen des indicateurs de BKW calculés sur les modèles initial et centré.

1) Si les diagnostics sont les mêmes, il n'y a aucun problème d'interprétation.

2) Si, au contraire, les indicateurs obtenus sur le modèle initial indiquent qu'il y a multicolinéarité alors que les indicateurs calculés sur le modèle centré indiquent l'inverse, il faut s'interroger sur la possibilité d'un effet taille :

- si un tel effet est envisageable, on le corrige en exprimant les variables explicatives tenues pour responsables (par le tableau de décomposition des variances) dans des unités plus appropriées. Puis on recalcule les indicateurs de BKW sur le modèle corrigé de l'effet taille (équivalent à un reparamétrage du modèle initial). Si les diagnostics sur les données corrigées non centrées et centrées deviennent similaires (absence de multicolinéarité), cela signifie que l'effet taille était effectivement à l'origine de la difficulté d'interprétation. On conclut à l'absence de multicolinéarité. Si, au contraire, les diagnostics demeurent incompatibles et qu'on est certain d'avoir éliminé tout effet taille du modèle, alors on se situe désormais dans le cas de figure suivant ;

- si aucun effet taille n'est envisageable, on se trouve face à l'obligation de prendre parti pour l'une ou l'autre des positions exprimées dans le débat sur le centrage. Pour ma part, je suis sensible au premier argument de Belsley, **dès lors que la question de l'unité appropriée pour exprimer les variables explicatives ne se pose pas.** Dans une configuration ambiguë de cette nature, on privilégiera le diagnostic établi sur le modèle non centré.

Conclusion

Les indicateurs de BKW font l'objet de fréquentes erreurs d'interprétation, qu'on s'est efforcé de dénoncer. Interprétés correctement, ces indicateurs constituent des outils précieux de détection de la multicolinéarité et de compréhension de ses sources. Cependant, leur propriété de non invariance par changement d'origine leur vaut une interprétation parfois délicate et suscite un inévitable débat sur l'opportunité de raisonner sur des modèles transformés par centrage. On a donné ici un point de vue à cet égard, sans avoir le moins du monde l'ambition de clore ce débat difficile.

ANNEXE

Dans Rousse [1990], on donne des indications sur les seuils à partir desquels apparaissent des problèmes de multicolinéarité engendrant des erreurs relatives inacceptables sur les estimations des paramètres du premier ordre. Les seuils d'indices de conditionnement η_j^* sont cohérents avec ceux que préconisent Belsley, Kuh et Welsch. On présente en outre des seuils relatifs à la borne inférieure θ des angles θ_k formés entre un vecteur colonne X_k et le sous-espace engendré par les autres colonnes de X (N.B. : θ varie dans]0,90°]).

Intensité des liaisons entre vecteurs colonnes de X	**Multicolinéarité**	**Angle θ (degrés)**	**Indices de conditionnement η_j^***
faibles	non	$\theta > 10$	$\eta_j^* \leq 10$
assez faibles	non	$4 \text{ ou } 5 \leq \theta \leq 10$	$10 \leq \eta_j^* \leq 20$
modérées ?	zone d'ambiguïté	$4 \text{ ou } 5 \leq \theta \leq 10$	$20 \leq \eta_j^* \leq 30$
modérées ou fortes	oui	$2 \leq \theta \leq 4 \text{ ou } 5$	$30 \leq \eta_j^* \leq 100$
très fortes	oui (pathologique)	$\theta < 2$	$\eta_j^* > 100$

Source : Rousse [1990]

BIBLIOGRAPHIE

R.G. ASKIN (1982) : Multicollinearity in regression: review and examples, *Journal of Forecasting*, vol. 1, n° 3, pp. 281-292.

D.A. BELSLEY (1984) : Demeaning conditioning diagnostics through centering, *The American Statistician*, vol. 38, n° 2, suivi de commentaires et de la réponse de l'auteur, pp. 73-93.

D.A. BELSLEY (1986) : Centering, the Constant, First-differencing, and Assessing conditioning, in *Model Reliability*, D.A. Belsley and E. Kuh Ed., M.I.T. Press, pp. 117-153.

D.A. BELSLEY (1991) : *Conditioning diagnostics: Collinearity and weak data in regression*, Wiley interscience.

D.A. BELSLEY, E. KUH, R.E. WELSCH (1980) : *Regression Diagnostics: Identifying influential data and sources of Collinearity*, Wiley Ed.

H. ERKEL-ROUSSE (1992) : Les performances extérieures de la France et de l'Allemagne, l'effet de l'investissement sur la compétitivité, *Économie et Statistique* n° 253, avril, pp. 35-47.

H. ERKEL-ROUSSE (1994) : Détection de la multicolinéarité dans un modèle linéaire ordinaire : Quelques éléments pour un usage averti des indicateurs de BELSLEY, KUH et WELSCH, projet de document de travail de la *Collection méthodologique de l'Insee*, 21 avril 1994.

H. ERKEL-ROUSSE (1994/95) : Multicolinéarité dans le modèle linéaire ordinaire : définition, détection, propositions de solutions, in *Introduction à l'économétrie du modèle linéaire*, polycopié Ensae, pp. 177-252.

H. ROUSSE (1990) : Détection et effets de la multicolinéarité dans les modèles linéaires ordinaires, un prolongement de la réflexion de BELSLEY, KUH et WELSCH, *Document de travail du Département des Études Économiques d'Ensemble de l'Insee*, n° 9002, juin.

D.E. FARRAR, R.R. GLAUBER (1967) : Multicollinearity in regression analysis: the problem revisited, *Review of Economics and Statistics*, vol. 49, pp. 92-107.

R.F. GUNST (1983) : Regression analysis with multicollinear predictor variables: definition, detection and effects, *Communications in statistics, theory and methods*, vol. 12, n° 19, pp. 2217-2260.

R.F. GUNST, R.L. MASON (1977) : Advantages of examing multicollinearities in regression analysis, *Biometrics*, vol. 33, pp. 249-260.

G.G. JUDGE, W.E. GRIFFITHS, R.C. HILL, T.C. LEE (1980) : *The Theory and Practice of Econometrics*, Wiley Ed.

J.P. LESAGE, S.D. SIMON (1985) : Numerical accuracy of statistical algorithms for microcomputers, *Computational Statistics and Data Analysis*, n° 3, pp. 47-57.

G.S. MADDALA (1977) : *Econometrics*, Mc Graw-Hill Ed.

D.W. MARQUARDT (1980) : You should standardize the predictor variables in your regression models, *Journal of the American Statistical Association*, vol. 75, pp. 74-103.

R.L. MASON, R.F. GUNST, J.T. WEBSTER (1975) : Regression analysis and problems of multicollinearity, *Communications in Statistics*, 4(3), pp. 277-292.

D.C. MONTGOMERY, E.A. PECK (1982) : *Introduction to linear regression analysis*, Academic: New York.

G. SAPORTA (1990) : *Probabilités, Analyse des données et Statistique*, Technip.

SAS Institute Inc. (1989) : *SAS/STAT User's Guide*, Version 6 Ed., Vol. 2, Cary.

P. SEVESTRE (1988) : *Econométrie II*, Polycopié Ensae.

S.D. SILVEY (1969) : Multicollinearity and Imprecise Estimations, *Journal of Royal Statistical Society*, Series B, vol. 31, pp. 539-552.

S.D. SIMON, J.P. LESAGE (1988) : The impact of collinearity involving the intercept term on the numerical accuracy of regression, *Computer Science in Economics and Management*, n° 1, pp. 137-152.

G.W. STEWART (1987) : Collinearity and least squares regression, *Statistical Science* vol. 2, n° 1, pp. 68-84, suivi des commentaires de D.A. BELSLEY, pp. 86-91 et de R.A.THISTED, pp. 91-93.

R.D. SNEE, D.W. MARQUARDT (1984) : Collinearity diagnostics depend on the domain of prediction, the model and the data, *The American Statistician*, vol. 38, n° 2, pp. 83-87.

S. WEISBERG (1980) : *Applied linear regression*, Wiley, New York.