Université de Montréal

Sur la détection de données aberrantes en régression linéaire multivariée

par

Nathalie Malo

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

Décembre 2000

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Sur la détection de données aberrantes en régression linéaire multivariée

présenté par

Nathalie Malo

a été évalué par un jury composé des personnes suivantes :

Yves Lepage
(président-rapporteur)

Robert Cléroux
(directeur de recherche)

Aziz Lazraq
(co-directeur)

Jean-François Angers
(membre du jury)

Mémoire accepté le:
Mars 2001

SOMMAIRE

Dans plusieurs domaines, souvent l'intérêt porte spécifiquement sur ce qu'on nomme une donnée aberrante. Par contre, en statistique, on doit tenir compte de la présence d'une telle observation. Dans un cas comme dans l'autre, le problème demeure le même : réussir à déceler ces données à l'écart.

Le but de ce mémoire est donc de définir et de comparer divers algorithmes robustes et non robustes de détection d'aberrances une à la fois ou en groupes. Étant donné l'utilisation très répandue de la régression en recherche et la complexité des aberrances multivariées, les algorithmes ici étudiés reposent sur le modèle de la régression linéaire multivariée.

Les premières méthodes sont basées sur l'indice de redondance de Stewart et Love (1968). Elles utilisent soit la fonction d'influence soit l'inclusion successive (FORWARD) de variables en régression multivariée. Quant aux autres méthodes provenant de la littérature, elles se basent sur une forme quadratique des résidus pour détecter une translation dans la moyenne.

Finalement, puisqu'une aberrance multivariée peut être identifiée par une méthode et non par une autre ainsi que selon un certain modèle et non selon un autre, on recommande d'utiliser plus d'une de ces procédures. En effet, l'emploi des différentes fonctions d'influence de l'indice de redondance de Stewart et Love (1968) jumelé à un algorithme utilisant une forme quadratique robuste des résidus donne des résultats très satisfaisants pour l'identification de données aberrantes en régression linéaire multivariée.

REMERCIEMENTS

Tout a débuté il y a environ cinq ans, le jour des inscriptions à l'université. Toujours aussi indécise, cette fois j'ai pourtant rempli une seule demande comprenant deux choix : mathématiques ou musique! Déjà, j'étais très reconnaissante envers mes professeurs du cégep, spécialement Serge Robert et Yves Camerlain, qui m'ont permis de découvrir mon intérêt pour les mathématiques et qui furent les premiers à croire en moi. Malgré leur recommandation d'attendre après la première année du bacc, j'étais alors convaincue de détester la statistique...

Depuis, j'ai parcouru un grand bout de chemin. Qui aurait cru que cette jeune fille qui détestait jadis également l'informatique deviendrait un jour coadministratrice du laboratoire de statistique! Là, j'ai eu droit au meilleur "boss" qu'on peut avoir, Miguel Chagnon. Merci beaucoup pour ce travail qui s'est avéré une formation tant amusante qu'enrichissante et pour tes précieux conseils que j'apprécie encore aujourd'hui. Merci également à Christian Léger qui m'a dirigée vers la statistique après un premier cours et qui a souvent pensé à moi pour divers emplois.

Plus particulièrement, j'aimerais remercier mon directeur, Robert Cléroux, qui fut le premier à me donner la chance de travailler en statistique. Merci pour votre grande compréhension et pour votre support financier qui m'ont permis de réaliser plusieurs projets dont le plus important est incontestablement ce mémoire. Merci à Aziz Lazraq pour ses explications, pour m'avoir appris à programmer en *S-Plus* et avec qui il fut intéressant de travailler tout un été. Merci aussi à la

Faculté des Études Supérieures pour la bourse de fin d'études qu'elle m'a accordée sous la recommandation du Département de Mathématiques et de Statistique.

Enfin, je n'aurais jamais pu me rendre jusqu'ici sans le soutien moral et financier de mes parents. Merci papa et maman d'avoir toujours été là pour m'aider. Merci également à Phil pour tous les petits à côtés et pour ton calme lors des moments plus stressants.

Finalement, merci à Alexis et Alexandre pour m'avoir appris qu'il est possible de terminer un devoir à l'avance... et non toujours à la dernière minute! Merci aussi a tous les autres amis, professeurs, coads et ex-coads sans qui la vie étudiante n'aurait jamais été aussi agréable. Bref, merci à tous pour m'avoir fait sourire à chaque jour depuis mon entrée à l'Université de Montréal.

Aujourd'hui j'adore la statistique et ses maintes applications, j'apprécie l'informatique et surtout, j'ai grâce à vous acquis des connaissances qui me permettront de travailler... et de continuer à apprendre!

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

En recherche médicale, les individus d'intérêts sont ceux qui répondent étrangement à divers traitements. En exploitation minière, on cherche une minéralisation spécifique cachée sous la surface. Pour un gestionnaire, une caisse reflète un profil financier bien différent de celui des autres banques. Quant à l'éleveur de plantes ou d'animaux, il recherche les individus supérieurs aux autres pour ses utilisations futures. Des situations comme celles-ci on peut en nommer plusieurs. Elles se retrouvent dans la vie de tous les jours, dans pratiquement tous les domaines. Mais qu'ont elles en commun? Dans tous ces exemples, l'étude porte spécifiquement sur ce qu'on nomme, en statistique, une observation aberrante. Dans la littérature, on parle aussi de données douteuses, de données erronées, d'observations à l'écart, d'observations influentes, de valeurs surprises, de valeurs extrêmes, d'aberrances, etc.

D'un autre côté, le statisticien se doit de travailler sur des données expérimentales fiables. Il doit donc tenir compte de la présence de données aberrantes. En effet, il pourrait s'avérer que ces données soient la source de contamination dans les analyses futures. Pour toutes ces raisons, il est primordial de disposer de techniques statistiques permettant l'identification de ces données influentes. Ainsi, le statisticien sera en mesure d'évaluer la qualité de l'échantillon, ce qui est essentiel dès le tout début d'une nouvelle étude.

Or, le processus d'identification de données aberrantes possède une longue histoire. Durant les dernières décennies, plusieurs méthodes ont été proposées pour détecter une donnée à l'écart. La plupart de ces méthodes se rapportent

au cas univarié où il suffit d'examiner les points dans les ailes de la distribution échantillonnale. Cependant, il n'est pas évident de définir exactement la notion de donnée aberrante multivariée. En effet, une observation aberrante dans $\mathbb{R}^p$ ne l'est pas nécessairement dans un sous-espace donné. D'où l'application de techniques univariées à la projection sur chaque axe ne mène pas obligatoirement à de bons résultats. Dans la littérature, on retrouve quelques méthodes permettant de déceler des données aberrantes multivariées. Les premières s'appliquant seulement à des populations normales.

Selon Gnanadesikan (1977), il existe trois raisons pour lesquelles les conséquences d'avoir une observation influente sont intrinsèquement plus complexes dans un échantillon multivarié que dans le cas univarié. La première est que les aberrances multivariées peuvent fausser non seulement les mesures de position et d'échelle, mais également les mesures d'orientation telle la corrélation. Comme on vient de le voir, il est aussi plus difficile de caractériser une aberrance multivariée. Enfin, on retrouve plusieurs variétés d'aberrances multivariées : une grande erreur dans l'une des composantes du vecteur ou une erreur moyenne systématiquement dans chacune des composantes.

À cause de l'utilisation très répandue de la régression, pour des chercheurs dans plusieurs domaines, l'importance porte sur la détection d'observations influentes sous un modèle de régression linéaire. En effet, une telle observation qui s'éloigne de la majorité peut fausser l'ajustement du modèle tentant d'exprimer la relation présente entre les autres données. Sous le critère des moindres carrés, ces observations à l'écart peuvent également influencer excessivement l'estimation des paramètres, altérer les intervalles de confiance et réduire la puissance des tests statistiques. Encore une fois, en régression multivariée le problème devient plus complexe. En effet, une observation douteuse peut être déclarée aberrante par une méthode et non par une autre, de même que selon un certain modèle et

non selon un autre.

Aujourd'hui, on s'intéresse donc à l'identification de données aberrantes multivariées, mais également à un problème tout aussi complexe : la détection de groupes de données aberrantes. Ce dernier a pour but de contrer l'effet de masquage qui survient lorsqu'une donnée influente cache la présence d'une autre ou lorsque des observations sont influentes ensemble, mais pas individuellement. Encore peu de méthodes portent sur ce sujet.

Le but de notre étude est donc de faire le point sur différentes méthodes de détection d'aberrances. Pour ce faire, on travaillera sous le contexte de la régression linéaire multivariée. Le modèle alors utilisé est décrit au premier chapitre. Notons qu'il existe des méthodes statistiques robustes qui sont valides même lorsque l'échantillon contient quelques données aberrantes. Toutefois, on s'intéresse ici à leur identification.

Dans le cadre de ce mémoire, toujours sous le modèle de la régression linéaire multivariée, on adaptera différents algorithmes utilisant une mesure d'association afin d'en construire de nouveaux basés sur l'indice de redondance de Stewart et Love (1968). L'emploi de ce coefficient nous permettra entre autres de définir, grâce à la connaissance de sa distribution exacte, des méthodes de sélection de variables avec inférence au premier chapitre.

Quant au deuxième chapitre, on le débutera en généralisant au cas multivarié une méthode de détection d'aberrances existante en régression linéaire multiple basée sur l'inclusion successive (FORWARD) de variables. Puis, on la comparera à deux algorithmes pour l'identification de données aberrantes en régression linéaire multivariée provenant de la littérature. Enfin, on développera une méthode basée sur la fonction d'influence de l'indice de redondance de Stewart et Love.

L'intérêt particulier de cet algorithme est qu'il pourra également être généralisé à l'influence d'un groupe de points. Combiné à une classification hiérarchique, on verra qu'il en découle une méthode de détection d'ensembles aberrants en régression multivariée.

Malheureusement, tous ces algorithmes sont basés sur des estimateurs très influençables par la présence de valeurs à l'écart. C'est pourquoi, au dernier chapitre, on comparera non seulement les méthodes précédentes, mais également une nouvelle version de chacune d'elles obtenue en remplaçant leur estimateur habituel par un estimateur robuste. Ces différentes méthodes seront d'abord appliquées à un exemple en régression linéaire multiple, puis à quatre autres jeux de données en régression multivariée.

Bref, dans ce mémoire, il sera question de l'identification de données aberrantes et d'ensembles d'observations influentes en régression linéaire multivariée. Notons que nous n'allons pas porter de jugement quant au traitement spécial qui doit leur être accordé. En présence d'aberrances, c'est avec l'aide de l'expérimentateur que le statisticien pourra décider si ces données doivent être corrigées ou supprimées.

Chapitre 1

RAPPEL SUR LA RÉGRESSION MULTIVARIÉE

Dans ce premier chapitre, il est question des aspects théoriques de la régression linéaire multivariée qui seront utiles à la compréhension des différentes méthodes de détection de données aberrantes décrites au chapitre suivant. On rappellera entre autres le modèle de régression linéaire multivariée ainsi que l'obtention des estimateurs des moindres carrés et des estimateurs à vraisemblance maximale. Puis, on citera les propriétés ainsi que le rôle dans la régression multivariée de deux mesures d'association.

1.1. Modèle de régression linéaire multivariée

Dans cette section, on s'intéresse à modéliser la relation entre un ensemble de p variables aléatoires à expliquer, $Y = (\mathcal{Y}_1,\mathcal{Y}_2,\cdots,\mathcal{Y}_p)'$, et un ensemble de q variables contrôlables, $X = (\mathcal{X}_1,\mathcal{X}_2,\cdots,\mathcal{X}_q)'$, où $\mathcal{X}_1$ peut être égal à 1. Plus précisément, en régression linéaire multivariée, on veut exprimer simultanément toutes les composantes $\mathcal{Y}_j$, $j = 1,2,\cdots,p$, du vecteur aléatoire Y comme des combinaisons linéaires des q variables contrôlables à des erreurs aléatoires près. Pour ce faire, considérons un ensemble de p équations de régression linéaire de la façon suivante :

$$\mathcal{Y}_1 = \beta_{11}\mathcal{X}_1 + \beta_{12}\mathcal{X}_2 + \cdots + \beta_{1q}\mathcal{X}_q + \epsilon_1 = \underset{\sim}{\beta_1}'X + \epsilon_1 \qquad (1.1.1)$$

$$\begin{aligned}
\mathcal{Y}_2 &= \beta_{21}\mathcal{X}_1 + \beta_{22}\mathcal{X}_2 + \cdots + \beta_{2q}\mathcal{X}_q + \epsilon_2 = \underset{\sim}{\beta_2}'X + \epsilon_2 \\
&\vdots \\
\mathcal{Y}_p &= \beta_{p1}\mathcal{X}_1 + \beta_{p2}\mathcal{X}_2 + \cdots + \beta_{pq}\mathcal{X}_q + \epsilon_p = \underset{\sim}{\beta_p}'X + \epsilon_p
\end{aligned}$$

où les $\underset{\sim}{\beta_j} = (\beta_{j1},\beta_{j2},\cdots,\beta_{jq})'$, $j = 1,2,\cdots,p$, sont des vecteurs de taille q.

Le système d'équations linéaires précédent peut donc s'écrire sous la forme

$$\begin{array}{ccccccc}
Y & = & \mathcal{B}' & & X & + & \underset{\sim}{\epsilon} \\
(p \times 1) & & (p \times q) & & (q \times 1) & & (p \times 1)
\end{array}$$

où l'espérance du vecteur aléatoire d'erreur $\underset{\sim}{\epsilon} = (\epsilon_1,\epsilon_2,\cdots,\epsilon_p)'$ est nulle et où les termes d'erreur associés à différentes variables dépendantes peuvent être corrélés.

Si pour un vecteur X_α de taille q fixé on observe le vecteur correspondant Y_α de taille p et si on recommence indépendamment n fois, alors on obtient le modèle empirique suivant :

$$Y_\alpha = \mathcal{B}'X_\alpha + \underset{\sim}{\epsilon_\alpha},\ \alpha = 1,...,n \tag{1.1.2}$$

où $E(\underset{\sim}{\epsilon_\alpha}) = 0$ et où les $\underset{\sim}{\epsilon_\alpha}$ sont indépendants avec matrice de variance-covariance $\Sigma = E(\underset{\sim}{\epsilon_\alpha}\underset{\sim}{\epsilon_\alpha}')$ inconnue.

Puis, à partir des observations $Y_1,Y_2,\cdots,Y_n$ de Y et $X_1,X_2,\cdots,X_n$ de X du modèle en (1.1.2), on peut écrire le modèle de régression linéaire multivariée sous sa forme matricielle

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E} \tag{1.1.3}$$

où $\mathbf{Y} = [Y_1\ Y_2 \cdots Y_n]' = (Y_{ij}) : n \times p$ est une matrice dont chaque ligne Y_i', $i = 1,2,\cdots,n$ représente une observation du vecteur aléatoire Y;

$\mathbf{X} = [X_1\ X_2 \cdots X_n]' = (X_{ik})$: $n \times q$ est une matrice dont chaque ligne X_i', $i = 1,2,\cdots,n$ représente une observation du vecteur des variables contrôlables X;

$\mathcal{B} = [\underset{\sim}{\beta_1}\underset{\sim}{\beta_2} \cdots \underset{\sim}{\beta_p}] = (\beta_{kj})$: $q \times p$ est la matrice des paramètres inconnus;

$\mathcal{E} = [\underset{\sim}{\epsilon_1}\ \underset{\sim}{\epsilon_2} \cdots \underset{\sim}{\epsilon_n}]' = (\epsilon_{ij})$: $n \times p$ est une matrice dont les lignes $\underset{\sim}{\epsilon_i}'$ sont les vecteurs d'erreur de taille p associés à chaque observation.

Soit $X_i' = (x_{i1},x_{i2},\cdots,x_{iq})$ les valeurs prises par les variables contrôlables pour la i^e observation. Alors la i^e donnée ou le i^e cas se note (Y_i',X_i') où Y_i, $i = 1,2,\cdots,n$, est défini selon le modèle empirique en (1.1.2). Plus tard, on s'intéressera aux cas aberrants.

Pour chacun des modèles précédents, on suppose que la matrice $\mathbf{X}$ est de plein rang, c'est-à-dire que $rang(\mathbf{X}) = q$. Également, on suppose que $n \geq p + q$ et que les vecteurs Y_α, $\alpha = 1,2,\cdots,p$, sont multinormaux.

D'un autre côté, considérons les vecteurs $\underset{\sim}{Y_j}$: $n \times 1$, $j = 1,2,\cdots,p$, formés à partir des colonnes de la matrice $\mathbf{Y} = [\underset{\sim}{Y_1}\underset{\sim}{Y_2} \cdots \underset{\sim}{Y_p}]$. Ainsi, chaque $\underset{\sim}{Y_j}$ représente un vecteur aléatoire formé des n observations de l'une des p composantes du vecteur conceptuel Y. Notons que si ces composantes étaient indépendantes, alors la régression linéaire multivariée serait équivalente à effectuer p régressions linéaires multiples, soit une pour chacun des vecteurs $\underset{\sim}{Y_j}$ sur la matrice $\mathbf{X}$ comme suit :

$$\begin{array}{ccccccc} \underset{\sim}{Y_j} & = & \mathbf{X} & & \underset{\sim}{\beta_j} & + & \underset{\sim}{\varepsilon_j} \\ (n \times 1) & & (n \times q) & & (q \times 1) & & (n \times 1) \end{array} \tag{1.1.4}$$

où chaque vecteur $\underset{\sim}{\beta_j}$ tel que défini en (1.1.1) contient les coefficients de la régression univariée du vecteur $\underset{\sim}{Y_j}$ sur la matrice $\mathbf{X}$, $j = 1,2,\cdots p$.

1.1.1. Estimateurs des moindres carrés

On s'intéresse maintenant à l'estimation des paramètres inconnus. On généralisera d'abord l'estimation des moindres carrés au cas multivarié pour estimer les coefficients de régression. Puis, on trouvera les estimateurs à vraisemblance maximale des matrices $\mathcal{B}$ et Σ.

Revenons donc au modèle précédent (1.1.4) où l'on considère l'ensemble des p régressions linéaires univariées.

À partir des valeurs des variables dépendantes et indépendantes, on cherche l'estimation par la méthode des moindres carrés des paramètres β. Puisque la matrice $\mathbf{X}$ est de plein rang alors, en particulier, $\mathbf{X}'\mathbf{X}$ est non singulière et donc inversible.

Les estimateurs des moindres carrés $\underset{\sim}{\beta_j}$, $j = 1,2,\cdots,p$, peuvent être déterminés exclusivement à partir des observations du vecteur $\underset{\sim}{Y_j}$. Conformément à la solution en régression linéaire univariée, on trouve

$$\underset{\sim}{\beta_j} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{Y_j}.$$

En regroupant ces p estimateurs, on obtient la matrice des estimateurs des moindres carrés pour le cas multivarié

$$\hat{\mathcal{B}} = [\underset{\sim}{\hat{\beta}_1}\underset{\sim}{\hat{\beta}_2}\cdots\underset{\sim}{\hat{\beta}_p}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\underset{\sim}{Y_1}\underset{\sim}{Y_2}\cdots\underset{\sim}{Y_p}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.1.5}$$

Notons qu'on peut également obtenir l'estimation des coefficients de régression directement sous le contexte de la régression multivariée. Par la méthode des moindres carrés, il s'agit donc de minimiser la somme des carrés des erreurs donnée par la trace de $\mathcal{E}'\mathcal{E}$.

Or, par le modèle (1.1.3) sous sa forme matricielle, on a que la matrice des erreurs est donnée par $\mathcal{E} = \mathbf{Y} - \mathbf{X}\mathcal{B}$ et on obtient les carrés des erreurs

$$\begin{aligned}\mathcal{E}'\mathcal{E} &= (\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathcal{B} - \mathcal{B}'\mathbf{X}'\mathbf{Y} + \mathcal{B}'\mathbf{X}'\mathbf{X}\mathcal{B}.\end{aligned}$$

On veut donc minimiser la somme des carrés des erreurs exprimée par

$$\begin{aligned}tr(\mathcal{E}'\mathcal{E}) &= tr(\mathbf{Y}'\mathbf{Y}) - tr(\mathbf{Y}'\mathbf{X}\mathcal{B}) - tr(\mathbf{Y}\mathcal{B}'\mathbf{X}') + tr(\mathcal{B}'\mathbf{X}'\mathbf{X}\mathcal{B}) \\ &= tr(\mathbf{Y}'\mathbf{Y}) - 2tr(\mathbf{Y}'\mathbf{X}\mathcal{B}) + tr(\mathcal{B}'\mathbf{X}'\mathbf{X}\mathcal{B}).\end{aligned}$$

Pour ce faire on calcule la dérivée partielle suivante

$$\frac{\partial tr(\mathcal{E}'\mathcal{E})}{\partial \mathcal{B}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathcal{B} = 0$$

de laquelle on obtient l'estimateur des moindres carrées $\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

À partir de cet estimateur, on calcule la matrice des valeurs prédites

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathcal{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.1.6}$$

ainsi que la matrice des résidus

$$\hat{\mathcal{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = [I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = H\mathbf{Y} \tag{1.1.7}$$

où I_n est la matrice identité de dimension $n \times n$. Il est facile de démontrer que la matrice H est symétrique ($H' = H$) et idempotente ($H^2 = H$).

Notons que les conditions d'orthogonalité entre les résidus, les valeurs prédites et les colonnes de $\mathbf{X}$ de la régression linéaire univariée tiennent toujours pour le cas multivarié. Elles proviennent principalement de l'équation $\mathbf{X}'[I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{X}' - \mathbf{X}' = 0$. Ainsi, les résidus $\hat{\varepsilon}_j$ $(j = 1,2,\cdots,p)$ sont perpendiculaires aux colonnes de $\mathbf{X}$, car

$$\mathbf{X}'\hat{\mathcal{E}} = \mathbf{X}'[I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = 0$$

et les valeurs prédites $\hat{Y}_j$ sont perpendiculaires aux vecteurs résiduels $\hat{\varepsilon}_k$ puisque

$$\hat{\mathbf{Y}}'\hat{\mathcal{E}} = \hat{\mathcal{B}}'\mathbf{X}'[I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = 0.$$

Étant donné que $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathcal{E}}$, on obtient finalement

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\mathcal{E}})'(\hat{\mathbf{Y}} + \hat{\mathcal{E}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathcal{E}}'\hat{\mathcal{E}} + 0 + 0' = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathcal{E}}'\hat{\mathcal{E}} \qquad (1.1.8)$$

où $\mathbf{Y}'\mathbf{Y}$, $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ et $\hat{\mathcal{E}}'\hat{\mathcal{E}}$ sont les matrices formées des sommes de carrés et des produits croisés respectivement dus au total, à la régression et à l'erreur.

1.1.2. Estimateurs à vraisemblance maximale

D'autre part, en considérant le modèle empirique donné en (1.1.2) et en supposant la normalité des ϵ_α, on peut également obtenir les estimateurs à vraisemblance maximale des matrices $\mathcal{B}$ et Σ.

Soit $Y_\alpha \sim \mathcal{N}(\mathcal{B}'X_\alpha,\Sigma)$, $\alpha = 1,2,\cdots,n$. La vraisemblance est donnée par

$$L(\mathcal{B},\Sigma) = \frac{|\Sigma^{-1}|^{\frac{n}{2}}}{(2\pi)^{\frac{pn}{2}}} \cdot \exp\left[\frac{-1}{2}\sum_{\alpha=1}^{n}(Y_\alpha - \mathcal{B}'X_\alpha)'\Sigma^{-1}(Y_\alpha - \mathcal{B}'X_\alpha)\right]$$

et son logarithme naturel par

$$\begin{aligned}\ln L(\mathcal{B},\Sigma) &= \frac{-np}{2}\ln(2\pi) + \frac{n}{2}\ln(|\Sigma^{-1}|) \\ &- \frac{1}{2}\sum_{\alpha=1}^{n}(Y_\alpha - \mathcal{B}'X_\alpha)'\Sigma^{-1}(Y_\alpha - \mathcal{B}'X_\alpha).\end{aligned}$$

Par la dérivée partielle suivante

$$\begin{aligned}\frac{\partial \ln L(\mathcal{B},\Sigma)}{\partial \mathcal{B}} &= \frac{-1}{2}\sum_{\alpha=1}^{n} -2\Sigma^{-1}(Y_\alpha - \mathcal{B}'X_\alpha)X'_\alpha \\ &= \Sigma^{-1}\left[\sum_{\alpha=1}^{n} Y_\alpha X'_\alpha - \mathcal{B}\sum_{\alpha=1}^{n} X_\alpha X'_\alpha\right]\end{aligned}$$

on a que la vraisemblance est maximale en

$$\hat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \qquad (1.1.9)$$

D'où l'estimateur à vraisemblance maximale précédent est dans ce cas-ci le même que l'estimateur des moindres carrés trouvé en (1.1.5).

En remplaçant $\mathcal{B}$ par son estimateur $\hat{\mathcal{B}}$ on obtient

$$\begin{aligned} \ln L(\hat{\mathcal{B}},\Sigma) &= \frac{-np}{2}\ln(2\pi) + \frac{n}{2}\ln(|\Sigma^{-1}) \\ &- \frac{1}{2}tr\left[\Sigma^{-1}\cdot\sum_{\alpha=1}^{n}(Y_\alpha - \hat{\mathcal{B}}'X_\alpha)(Y_\alpha - \hat{\mathcal{B}}'X_\alpha)'\right]. \end{aligned}$$

On veut donc maximiser cette équation par rapport à Σ^{-1}. Pour ce faire, on aura besoin du théorème suivant dont la preuve est énoncée à la page 104 de Mardia *et al.* (1982).

Théorème 1.1.1. *Soit* $f(C) = \frac{n}{2}\ln(|C|) - \frac{1}{2}tr(CD)$ *où* $C = (c_{ij})$ *est une matrice semi-définie positive et où* $D = (d_{ij})$ *est une matrice définie positive. Alors le maximum de* $f(C)$ *est atteint par* $C = nD^{-1}$ *et est donné par* $f(nD^{-1})$ *où* $f(nD^{-1}) = \frac{pn}{2}\ln(n) - \frac{n}{2}\ln(|D|) - \frac{pn}{2}$.

Par le théorème précédent avec les matrices $C = \Sigma^{-1}$ et $D = \sum_{\alpha=1}^{n}(Y_\alpha - \mathcal{B}'X_\alpha)(Y_\alpha - \mathcal{B}'X_\alpha)'$ on trouve l'estimateur à vraisemblance maximale de Σ

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n}\sum_{\alpha=1}^{n}(Y_\alpha - \hat{\mathcal{B}}'X_\alpha)(Y_\alpha - \hat{\mathcal{B}}'X_\alpha)' \qquad (1.1.10) \\ &= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}). \end{aligned}$$

Par la définition des deux estimateurs obtenus dans cette section, on peut voir que $\hat{\mathcal{B}}$ est sans biais pour $\mathcal{B}$, mais que ce n'est pas le cas pour $\hat{\Sigma}$. En effet, c'est l'estimateur $\frac{n}{n-q}\hat{\Sigma}$ qui est sans biais pour Σ.

1.2. L'INDICE DE REDONDANCE DE STEWART ET LOVE

Le modèle de régression linéaire étant maintenant établi, on y reviendra un peu plus loin au cours de cette section. Pour l'instant, définissons une mesure de corrélation qui sera utile dans les algorithmes de sélection de variables de la section suivante, ainsi que dans ceux de détection d'aberrances du second chapitre.

Pour ce faire, considérons deux vecteurs aléatoires $X^{(1)} : p \times 1$ et $X^{(2)} : q \times 1$ de moyenne $E(X^{(i)}) = \mu^{(i)}$ et de covariance $\Sigma_{ij} = E[(X^{(i)} - \mu^{(i)})(X^{(j)} - \mu^{(j)})']$, $i = 1,2$ et $j = 1,2$.

Soit $\mu = E(X) = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$ le vecteur de moyennes du vecteur aléatoire $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} : (p+q) \times 1$. Sa matrice de variance-covariance, définie positive, est

$$\Sigma = E[(X - \mu)(X - \mu)'] = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Supposons que μ et Σ sont inconnus.

Au niveau de l'échantillon aléatoire de taille n obtenu à partir de X, on a $X_\alpha = \begin{pmatrix} X_\alpha^{(1)} \\ X_\alpha^{(2)} \end{pmatrix}$, $\alpha = 1,2,\ldots,n$. La moyenne échantillonnale se définit par $\overline{X}^{(i)} = \frac{1}{n} \sum_{\alpha=1}^{n} X_\alpha^{(i)}$, $i = 1,2$. Quant à la matrice de variance-covariance empirique, elle peut s'écrire sous la forme suivante :

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}. \tag{1.2.1}$$

Lingoes et Schonemann (1974) ont introduit le coefficient de corrélation échantillonnal entre les deux vecteurs $X^{(1)}$ et $X^{(2)}$ suivant :

$$RLS = RLS(X^{(1)},X^{(2)}) = \frac{tr(S_{12}S_{21})^{\frac{1}{2}}}{\sqrt{tr(S_{11})tr(S_{22})}}. \tag{1.2.2}$$

Cette mesure de corrélation possède les propriétés algébriques suivantes :

1. $0 \leq RLS \leq 1$.
2. Si $p = q = 1$, alors RLS est équivalent à la valeur absolue du coefficient de corrélation simple |r|.
3. RLS est une mesure d'association. Elle est symétrique, mais comme elle n'est pas fonction des corrélations canoniques, elle n'est donc pas invariante sous les transformations linéaires en général.
4. Si A est une matrice telle que $A'A = \text{kI}$ où I est la matrice identité et $k > 0$ est un scalaire, alors $RLS(AX^{(1)}, X^{(2)}) = RLS(X^{(1)}, X^{(2)})$.
5. Soient les vecteurs $X^{(1)} : p \times 1$, $X^{(i)} : q \times 1$ et les scalaires a_i, $i = 2,3,\cdots,m$. Soit S_{ij} la matrice de variance-covariance échantillonnale entre $X^{(i)}$ et $X^{(j)}$ pour $i,j = 1,2,\cdots,m$. Lorsque $p = q = 1$ et que $RLS(X^{(1)}, \sum_{i=2}^{m} a_i X^{(i)})$ est maximisé selon les a_i, alors il est équivalent à R, le coefficient de corrélation multiple.

De plus, Lazraq et Cléroux (1989) décrivent une distance entre deux nuages de points en termes de RLS, de façon analogue à Robert et Escoufier (1976). Considérons les deux nuages de points

$$Y_1 = (X_1^{(1)} - \overline{X}^{(1)}, X_2^{(1)} - \overline{X}^{(1)}, \cdots, (X_n^{(1)} - \overline{X}^{(1)}) : p \times n$$
$$Y_2 = (X_1^{(2)} - \overline{X}^{(2)}, X_2^{(2)} - \overline{X}^{(2)}, \cdots, (X_n^{(2)} - \overline{X}^{(2)}) : q \times n$$

l'un dans $\mathcal{R}^p$ et l'autre dans $\mathcal{R}^q$. Supposons que $q < p$, alors on peut ajouter une colonne de $p - q$ zéros à chaque vecteur $X_i^{(2)} - \overline{X}^{(2)}$, $i = 1,2,\cdots,n$.

Pour toute matrice E, on définit la norme euclidienne correspondante comme étant $||E|| = \sqrt{tr(E'E)}$ et on définit la distance entre Y_1 et Y_2 par

$$dist(Y_1,Y_2) = \left|\left| \frac{Y_1}{\sqrt{tr(Y_1'Y_1)}} - \frac{Y_2}{\sqrt{tr(Y_2'Y_2)}} \right|\right|.$$

Théorème 1.2.1. *Soit $T : p \times p$ une matrice orthogonale, il est possible de montrer que*

$$\min_T dist(Y_1,TY_2) = \sqrt{2}\sqrt{1 - RLS(X^{(1)},X^{(2)})}. \tag{1.2.3}$$

Alors $RLS = 1$ si et seulement si il existe une rotation T tel que les deux nuages de points s'associent complètement. D'où RLS est un coefficient de corrélation et une donnée qui influence RLS est à la fois une donnée qui influence la distance minimale entre Y_1 et TY_2.

La démonstration du théorème précédent se trouve aux pages 253 à 256 de Seber (1984).

Revenons maintenant au contexte de la régression linéaire multivariée. Soient les deux vecteurs aléatoires $X^{(1)} \equiv Y_\alpha : p \times 1$ et $X^{(2)} \equiv X_\alpha : q \times 1$ du modèle (1.1.2). On veut donc expliquer linéairement $X^{(1)}$ à partir de $X^{(2)}$.

Or, le problème consiste à trouver une matrice M de dimensions $q \times p$ telle que le coefficient RLS($X^{(1)}$, $M'X^{(2)}$) soit maximal sous la contrainte $S_{12}M - M'S_{22}M = 0$ (les équations de régression). La solution est $M' = S_{12}S_{22}^{-1}$. Les nouvelles variables $M'X^{(2)}$ sont les p fonctions linéaires obtenues par la régression linéaire multivariée de $X^{(1)}$ sur $X^{(2)}$.

La valeur maximale du carré de RLS est donc donnée par

$$RI = RI(X^{(1)},X^{(2)}) = \max\ RLS^2(X^{(1)},M'X^{(2)}) = \frac{tr(S_{12}S_{22}^{-1}S_{21})}{tr(S_{11})}. \tag{1.2.4}$$

Il s'agit de l'indice de redondance de Stewart et Love (1968) noté RI qui représente la fraction de la variabilité totale de $X^{(1)}$ expliquée par la régression sur $X^{(2)}$. Ce dernier peut également s'écrire sous la forme d'une moyenne pondérée par les variances des carrés des coefficients de corrélation multiple entre les composantes du vecteur à prédire $X^{(1)}$ et le vecteur de prédiction $X^{(2)}$:

$$RI = \frac{\sum_{i=1}^{p} S^2_{x_i^{(1)}} R_{x_i^{(1)} \cdot x_1^{(2)}, x_2^{(2)}, \ldots x_q^{(2)}}}{\sum_{i=1}^{p} S^2_{x_i^{(1)}}} \tag{1.2.5}$$

où $S^2_{x_i^{(1)}}$ est la variance empirique de la composante $x_i^{(1)}$ du vecteur $X^{(1)}$ et $R^2_{x_i^{(1)} \cdot x_1^{(2)}, x_2^{(2)}, \ldots, x_q^{(2)}} = 1 - \dfrac{SCE_i}{SCT_i}$ est le carré du coefficient empirique de corrélation multiple pour la régression linéaire multiple de $x_i^{(1)}$ sur $x_1^{(2)}, x_2^{(2)}, \ldots, x_q^{(2)}$ avec SCE_i et SCT_i qui sont respectivement les sommes de carrés dues à l'erreur et totale.

Les propriétés de cette mesure de redondance sont les suivantes :

1. $0 \leq RI \leq 1$.
2. Si $p = q = 1$, alors RI est équivalent à r^2, le carré du coefficient de corrélation simple.
3. Lorsque $p = 1$, RI devient le carré du coefficient de corrélation multiple R^2 entre une variable et un vecteur.
4. RI est une mesure d'association non symétrique et non invariante en général.
5. Soit A une matrice de dimensions $p \times p$ telle que $A'A = kI$ où I est la matrice identité et k est un scalaire et soit une autre matrice $B : q \times q$ non singulière. RI est invariant sous la transformation linéaire $X^{(1)} \longrightarrow AX^{(1)}$ et $X^{(2)} \longrightarrow BX^{(2)}$.

Notons qu'au niveau de la population on définit l'équivalent de RI donné à l'équation (1.2.4) par

$$\rho I = \frac{tr(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})}{tr(\Sigma_{11})}. \tag{1.2.6}$$

1.2.1. Fonction d'influence

Au chapitre suivant, lors du contexte de la détection de valeurs aberrantes en régression multivariée, nous aurons a considérer la fonction d'influence de RI telle que définie ci-bas.

Définition 1.2.1. *Soit F la fonction de répartition d'une variable aléatoire X avec paramètre θ. Alors θ peut s'écrire comme une fonctionnelle de F, $\theta = T(F)$. Si δ_x est une fonction de répartition attribuant la probabilité 1 au point x, alors $\widetilde{F} = (1-\varepsilon)F + \varepsilon\delta_x$ est une perturbation de F par δ_x. Or, on a $\widetilde{\theta} = T(\widetilde{F})$ et la fonction d'influence théorique d'un point x sur le paramètre θ se définit par :*

$$I(x;\theta) = \lim_{\varepsilon\to 0}(\frac{\widetilde{\theta}-\theta}{\varepsilon}). \tag{1.2.7}$$

En remplaçant F par la fonction de répartition empirique, notée F_n, et ε par $\frac{-1}{n-1}$ dans l'équation (1.2.7) on obtient la fonction d'influence empirique. Une approximation de cette dernière consiste en la fonction d'influence expérimentale. Celle-ci détermine l'influence de x_i sur l'estimateur $\hat{\theta}$ de θ basé sur les observations $x_1, x_2, \ldots, x_n$:

$$I\_(x_i;\hat{\theta}) = (n-1)(\hat{\theta} - \hat{\theta}_{-i}), i = 1, \ldots, n. \tag{1.2.8}$$

où $\hat{\theta}_{-i}$ est un estimateur basé sur les $n-1$ observations $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$.

Lazraq et Cléroux (1989) ont obtenu la fonction d'influence théorique de RI suivante :

$$I(x;\rho I) = \rho I\left[\frac{2z^{(1)'}Bz^{(2)}}{tr(\Sigma_{11}^*)} - \frac{z^{(1)'}z^{(1)}}{tr(\Sigma_{11})} - \frac{z^{(2)'}B'Bz^{(2)}}{tr(\Sigma_{11}^*)}\right] \tag{1.2.9}$$

où $\Sigma_{11}^{*} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $B = \Sigma_{12}\Sigma_{22}^{-1}$ est la matrice de régression de $X^{(1)}$ sur $X^{(2)}$ et $z^{(i)} = x^{(i)} - \mu^{(i)}$, $i = 1,2$.

Notons que l'équation (1.2.9) peut également s'écrire sous une forme quadratique

$$I(x;\rho I) = z'Qz \tag{1.2.10}$$

où $z = \begin{pmatrix} z^{(1)} \\ z^{(2)} \end{pmatrix}$, $Q = \rho I \begin{pmatrix} \frac{-I_p}{tr(\Sigma_{11})} & \frac{B}{tr(\Sigma_{11}^{*})} \\ \frac{B'}{tr(\Sigma_{11}^{*})} & \frac{-B'B}{tr(\Sigma_{11}^{*})} \end{pmatrix}$ et I_p est une matrice identité de dimension $p \times p$.

Puis, Lazraq et Cléroux (1989) ont obtenu l'équation suivante pour la variance de la fonction d'influence :

$$\begin{aligned} \sigma^2 &= Var[I(x;\rho I)] \qquad (1.2.11) \\ &= 2\rho I^2 \left[\frac{tr\Sigma_{11}^2}{(tr\Sigma_{11})^2} - \frac{(4tr(\Sigma_{11}\Sigma_{11}^{*}) - 2tr\Sigma_{11}^{*2})}{tr\Sigma_{11}^{*}tr\Sigma_{11}} + \frac{(2tr\Sigma_{11}\Sigma_{11}^{*} - tr\Sigma_{11}^{*2})}{(tr\Sigma_{11}^{*})^2} \right]. \end{aligned}$$

Considérons maintenant les cas particuliers de la réduction de la fonction d'influence de RI en régression linéaire multiple et en régression linéaire simple.

À partir des résultats du tableau 1.2.1, en régression linéaire multiple, on obtient la fonction d'influence du carré du coefficient de corrélation multiple et sa variance associée.

$$\begin{aligned} I(x;R^2) &= R^2 \left[\frac{2z_1\underset{\sim}{\beta}'z^{(2)}}{R^2\sigma_1^2} - \frac{z_1^2}{\sigma_1^2} - \frac{z^{(2)\prime}\underset{\sim}{\beta}\underset{\sim}{\beta}'z_{(2)}}{R^2\sigma_1^2} \right] \\ &= \frac{1}{\sigma_1^2}[2z_1\underset{\sim}{\beta}'z^{(2)} - z_1^2R^2 - (\underset{\sim}{\beta}'z^{(2)})^2] \end{aligned}$$

et

$$Var[I(x;R^2)] = 2R^2 \left[1 - \frac{4\sigma_1^2R^2\sigma_1^2 - 2(R^2\sigma_1^2)^2}{R^2\sigma_1^2\sigma_1^2} + \frac{2\sigma_1^2R^2\sigma 1^2 - (R^2\sigma_1^2)^2}{(R^2\sigma_1^2)^2} \right]$$

TABLEAU 1.2.1. *Notation et réduction des expressions*

Régression multivariée $p \neq 1$ et $q \neq 1$	Régression multiple $p = 1$ et $q \neq 1$	Régression linéaire simple $p = q = 1$
$z^{(1)} : p \times 1$	$z_1 : 1 \times 1$	$z_1 = x_1 - \mu_1 : 1 \times 1$
$z^{(2)} : q \times 1$	$z^{(2)} : q \times 1$	$z_2 = x_2 - \mu_2 : 1 \times 1$
$\Sigma_{11} : p \times p$	$\sigma_1^2 : 1 \times 1$	$\sigma_1^2 : 1 \times 1$
$\Sigma_{12} : p \times q$	$\sigma_{(1)} : 1 \times q$	$\sigma_{12} : 1 \times 1$
$\Sigma_{22} : q \times q$	$\Sigma_{22} : q \times q$	$\sigma_2^2 : 1 \times 1$
RI	$R^2 = \frac{\sigma_{(1)}\Sigma_{22}^{-1}\sigma_{(1)}}{\sigma_1^2}$	$r^2 = \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}$
$\Sigma_{11}^* = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} : p \times p$	$\sigma_{11}^* = R^2\sigma_1^2 : 1 \times 1$	$\sigma_{11}^* = r^2\sigma_1^2 : 1 \times 1$
$B = \Sigma_{12}\Sigma_{22}^{-1} : p \times q$	$\underset{\sim}{\beta}' = \sigma_{(1)}\Sigma_{22}^{-1} : 1 \times q$	$b = \frac{\sigma_{12}}{\sigma_2^2} = \frac{\sigma_1}{\sigma_2}r : 1 \times 1$

$$
\begin{aligned}
&= 2R^2[1 - 4 + 2R^2 + \frac{2}{R^2} - 1] \\
&= 4(R^2)^2.
\end{aligned}
$$

On obtient également des résultats similaires pour le carré du coefficient de régression simple lorsque $p = q = 1$:

$$
\begin{aligned}
I(x; r^2) &= \frac{1}{\sigma_1^2}\left[2z_1z_2\frac{\sigma_1}{\sigma_2}r - z_1^2r^2 - \frac{\sigma_1^2}{\sigma_2^2}r^2z_2^2\right] \\
&= 2\frac{z_1}{\sigma_1}\frac{z_2}{\sigma_2}r - \frac{z_1^2}{\sigma_1^2}r^2 - \frac{z_2^2}{\sigma_2^2}r^2 \\
&= 2r\left[-\frac{r}{2}(\widetilde{y_1}^2 + \widetilde{y_2}^2) - \widetilde{y_1}\widetilde{y_2}\right]
\end{aligned}
$$

où $\widetilde{y_i} = \frac{z_i}{\sigma_i} = \frac{x_i - \mu_i}{\sigma_i}$, $i = 1,2$ et

$$Var[I(x; r^2)] = 4(r^2)^2.$$

D'autre part, à partir de l'équation (1.2.8) on obtient la fonction d'influence expérimentale de RI

$$I\_(x_i; RI) = (n-1)(RI - RI_{-i}),\ i = 1, \ldots, n. \tag{1.2.12}$$

C'est cette dernière qui nous permettra au chapitre suivant d'identifier des données aberrantes. En effet, une observation x_j ayant une grande influence en valeur absolue sera considérée comme un candidat à l'aberrance.

Or, on recherche deux types de données aberrantes. Le premier, survient lorsque l'influence de cette observation, mesurée par $I\_(x_j; RI)$, est négative. Cela signifie qu'enlever x_j augmente la valeur de RI tel qu'illustré par le graphique 1.2.1.

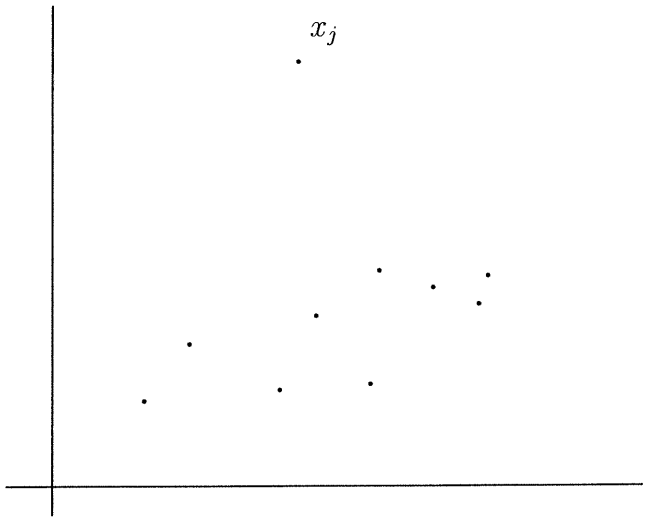

FIGURE 1.2.1. *Illustration du premier type d'aberrance*

Donc, dans ce cas, enlever x_j augmente la qualité du modèle. Par contre, lorsque la valeur de l'influence est positive, cela signifie qu'enlever x_j diminue RI ainsi que la qualité du modèle comme dans la figure 1.2.2.

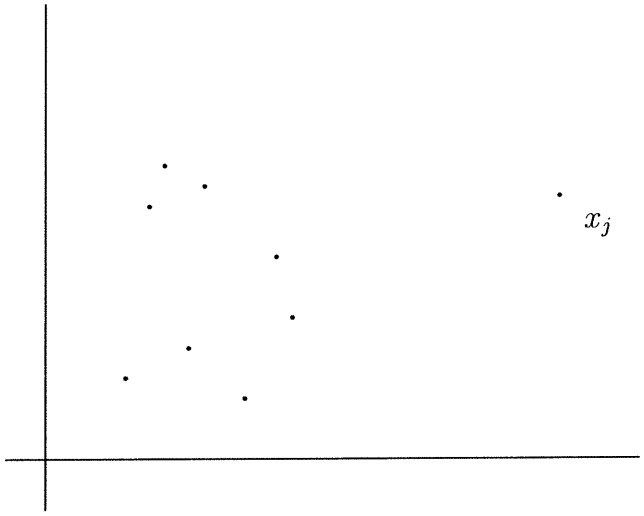

FIGURE 1.2.2. *Illustration du second type d'aberrance*

Dans ce deuxième cas, il s'agit d'un autre type d'aberrance qui pourrait indiquer que le modèle incluant l'observation x_j n'est peut être pas approprié.

1.2.2. Distribution exacte

L'avantage de l'indice de redondance de Stewart et Love par rapport à d'autres mesures de corrélation vectorielle, tel le coefficient RV de Escoufier (1973), est la connaissance de sa loi exacte. En effet, celle-ci nous permettra éventuellement d'utiliser RI dans un algorithme de sélection de variables avec inférence.

Lazraq et Cléroux (1988) ont obtenu la distribution exacte de $\frac{RI}{1-RI}$ sous l'hypothèse nulle $H_0 : \Sigma_{12} = 0$. Soit lorsque les variables explicatives sont indépendantes du vecteur à expliquer. Elle est donnée par le théorème suivant.

Théorème 1.2.2. *Si le vecteur* $\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ *est multinormal et, sous* H_0, $P\left[\frac{RI}{1-RI} \leq r\right] = P\left[W \leq 0\right]$ *où* W *est distribué comme* $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$ *où les* W_i *sont indépendantes et identiquement distribuées selon une* $\mathcal{N}(0,1)$, *et où* $\lambda_1, \lambda_2, \ldots, \lambda_{p(n-1)}$ *sont les* p *valeurs propres de* Σ_{11} *chacune répétées* q *fois et les* p *valeurs propres de* $-r\Sigma_{11}$ *chacune répétées* $n-1-q$ *fois.*

1.3. Régression multivariée et sélection de variables

En se basant sur la maximisation de l'indice de Stewart et Love tel que défini en (1.2.4), on veut déterminer un algorithme de sélection de variables.

Tel que vu à la section précédente, en régression linéaire multivariée, on cherche une matrice M de dimensions $q \times p$. On veut substituer à $X^{(1)}$ un ensemble $M'X^{(2)}$ de combinaisons linéaires qui peuvent expliquer $X^{(1)}$ de telle manière que la qualité de la prédiction de $X^{(1)}$ par $X^{(2)}$, mesurée par $\rho I(X^{(1)}, M'X^{(2)})$, soit maximale sous la contrainte $S_{12}M - M'S_{22}M = 0$. Les nouvelles variables $M'X^{(2)}$ sont celles obtenues par la régression linéaire multivariée de $X^{(1)}$ sur $X^{(2)}$. On trouve donc $M = S_{22}^{-1}S_{21}$.

Retournons à l'équation (1.1.8), où on a défini $\hat{\mathcal{E}}'\hat{\mathcal{E}}$ comme étant la matrice des produits scalaires des vecteurs d'erreur. Si on normalise les matrices pour que les moyennes soient nulles, alors on a $\mathbf{Y'Y} = (n-1)S_{11}$, $\mathbf{Y'X} = (n-1)S_{12}$, $\mathbf{X'X} = (n-1)S_{22}$ et d'après la définition de la matrice des résidus donnée en (1.1.7) on peut écrire

$$\begin{aligned}\hat{\mathcal{E}}'\hat{\mathcal{E}} &= (H\mathbf{Y})'H\mathbf{Y} = \mathbf{Y}'H'H\mathbf{Y} = \mathbf{Y}'H^2\mathbf{Y} = \mathbf{Y}'H\mathbf{Y} \\ &= \mathbf{Y}'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (n-1)S_{11} - (n-1)S_{12}S_{22}^{-1}S_{21} \\ &= (n-1)S_{11.2}.\end{aligned}$$

À partir de cette matrice résiduelle notée $S_{11.2}$ on parvient à la relation suivante :

$$tr(S_{11.2}) = tr(S_{11})\left[1 - \frac{tr(S_{12}S_{22}^{-1}S_{21})}{tr(S_{11})}\right] = tr(S_{11})[1 - RI]. \qquad (1.3.1)$$

D'où maximiser RI est équivalent à minimiser la trace de la matrice résiduelle $S_{11.2}$. Par (1.3.1) on peut définir l'indice de Stewart et Love comme suit :

$$RI = 1 - \frac{tr(S_{11.2})}{tr(S_{11})} = 1 - \frac{tr(\hat{\mathcal{E}}'\hat{\mathcal{E}})}{tr(\mathbf{Y}'\mathbf{Y})}.$$

1.3.1. Indice de redondance partiel

Pour pouvoir passer à un algorithme de sélection de variables, un dernier élément doit être défini. En effet, on aura besoin d'un indice de redondance partiel ainsi que de sa distribution. Alors soit $x_j^{(2)}$ une composante du vecteur $X^{(2)}$ et soit $X_{-j}^{(2)}$ un sous-vecteur de $X^{(2)}$ de taille t ne contenant pas $x_j^{(2)}$. Après avoir éliminé de $X^{(1)}$ et de $x_j^{(2)}$ l'effet linéaire de $X_{-j}^{(2)}$, on obtient $S_{\cdot X_{-j}^{(2)}}$, la matrice de covariance partielle de $\begin{pmatrix} X^{(1)} \\ x_j^{(2)} \end{pmatrix}$: $(p+1) \times 1$, donnée par :

$$S_{\cdot X_{-j}^{(2)}} = S - \begin{pmatrix} S_{X^{(1)}X_{-j}^{(2)}} \\ S_{x_j^{(2)}X_{-j}^{(2)}} \end{pmatrix} S^{-1}_{X_{-j}^{(2)}X_{-j}^{(2)}} \left(S_{X_{-j}^{(2)}X^{(1)}} S_{X_{-j}^{(2)}x_j^{(2)}}\right).$$

À partir de cette dernière, Lazraq et Cléroux (1988) ont défini

$$RI_{.X_{-j}^{(2)}}(X^{(1)},x_j^{(2)}) = \frac{tr(S_{X^{(1)}x_j^{(2)}.X_{-j}^{(2)}} s^{-1}_{x_j^{(2)}x_j^{(2)}.X_{-j}^{(2)}} S_{x_j^{(2)}X^{(1)}.X_{-j}^{(2)}})}{tr(S_{X^{(1)}X^{(1)}.X_{-j}^{(2)}})}. \tag{1.3.2}$$

Il s'agit de l'indice de redondance partiel entre le vecteur $X^{(1)}$ et la variable $x_j^{(2)}$ après avoir éliminé de $X^{(1)}$ et de $x_j^{(2)}$ l'effet linéaire de $X_{-j}^{(2)}$.

De façon analogue au théorème 1.2.2, la distribution de $\frac{RI_{.X^{(2)}}}{1-RI_{.X^{(2)}}}$ est donnée par le théorème 1.3.1.

Théorème 1.3.1. *Si le vecteur* $\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ *est multinormal, pour tout sous-vecteur* $X_{-j}^{(2)} : t \times 1$ *de* $X^{(2)}$ *ne contenant pas* $x_j^{(2)}$, *et sous* $H_0 : \Sigma_{X^{(1)}x_j^{(2)}.X_{-j}^{(2)}} = 0$, $P[\frac{RI_{.X^{(2)}}}{1-RI_{.X^{(2)}}} \leq s] = P[U \leq 0]$ *où* U *est distribué comme* $\sum_{i=1}^{p(n-1-t)} \delta_i U_i^2$ *où les* U_i *sont indépendantes et identiquement distribuées* $\mathcal{N}(0,1)$ *et où* $\delta_1,\delta_2,\cdots,\delta_{p(n-t-1)}$ *sont les* p *valeurs propres de* $\Sigma_{X^{(1)}X^{(1)}.X_{-j}^{(2)}}$ *de multiplicité 1 et les* p *valeurs propres de* $-s\Sigma_{X^{(1)}X^{(1)}.X_{-j}^{(2)}}$ *de multiplicité* $n-2-t$.

Lazraq et Cléroux (1988) ont aussi obtenu la relation de récurrence suivante entre RI et $RI_{.X^{(2)}}$:

$$\begin{aligned} RI\left(X^{(1)}, \begin{pmatrix} X_{-j}^{(2)} \\ x_j^{(2)} \end{pmatrix}\right) &= RI(X^{(1)},X_{-j}^{(2)}) \\ &\quad + RI_{.X_{-j}^{(2)}}(X^{(1)},x_j^{(2)})[1 - RI(X^{(1)},X_{-j}^{(2)})]. \end{aligned} \tag{1.3.3}$$

Puisqu'une formule analogue existe au niveau de la population, la régression linéaire multivariée avec inférence devient possible et fait l'objet de la prochaine section.

1.3.2. Algorithme de sélection de variables

Les résultats de la section précédente nous permettent maintenant d'écrire des algorithmes de sélection de variables avec inférence de type inclusion successive (FORWARD), élimination successive (BACKWARD) et pas à pas (STEPWISE). On s'intéresse ici à la sélection successive de variables lors de la régression du vecteur $X^{(1)}$ sur $X^{(2)}$ tel que défini par Lazraq et Cléroux (1988).

1. Pour chacune des q composantes $x_j^{(2)}$ du vecteur $X^{(2)}$, on calcule $RI(X^{(1)},x_j^{(2)})$ selon l'équation (1.2.4). On choisit la variable $x_j^{(2)}$ qui maximise le RI calculé précédemment et on l'ajoute au modèle.
2. De façon similaire, on cherche parmi les $q-t$ composantes restantes la variable $x_j^{(2)}$ qui, conjointement avec les t variables déjà entrées dans le modèle (représentées par le vecteur $X_{-j}^{(2)}$), maximisera $RI\left(X^{(1)}, \begin{pmatrix} X_{-j}^{(2)} \\ x_j^{(2)} \end{pmatrix}\right)$.
3. À chaque étape, on calcule le coefficient de redondance partiel $RI_{.X^{(2)}}$ d'après la relation de récurrence (1.3.3) ainsi que la statistique $\dfrac{RI_{.X^{(2)}}}{1-RI_{.X^{(2)}}}$.
4. Puis, on teste l'hypothèse $H_0 : \rho I\left(X^{(1)}, \begin{pmatrix} X_{-j}^{(2)} \\ x_j^{(2)} \end{pmatrix}\right) = \rho I(X^{(1)},X_{-j}^{(2)})$, reflétant le fait que l'apport de $x_j^{(2)}$ après $X_{-j}^{(2)}$ est négligeable? Cette hypothèse est équivalente à $H_0 : \rho I_{.X_{-j}^{(2)}} = 0$ qui est elle-même équivalente à $H_0 : \Sigma_{X^{(1)}x_j^{(2)}.X_{-j}^{(2)}} = 0$. On rejette H_0 au niveau α si $\dfrac{RI_{.X^{(2)}}}{1-RI_{.X^{(2)}}} >$ u où u est le $100(1-\alpha)^e$ centile de U distribué selon le théorème 1.3.1. Si H_0 est rejetée, alors la composante $x_j^{(2)}$ entre dans le modèle.
5. Ainsi de suite jusqu'à ce qu'on ne rejette pas l'hypothèse H_0 au niveau α. Le vecteur $X_{-j}^{(2)} : t \times 1$ de la dernière étape représente le sous-ensemble des variables de $X^{(2)}$ expliquant linéairement le mieux le vecteur $X^{(1)}$.

Notons que dans le cas où $p=1$, le calcul de $\dfrac{RI_{.X^{(2)}}}{1-RI_{.X^{(2)}}}$ en 3 revient à calculer la statistique F habituelle. On a déjà vu que RI est équivalent à R^2 dans ce cas-ci.

Par la relation de récurrence en (1.3.3) on obtient :

$$\begin{aligned} RI_{.X^{(2)}} &= \frac{R_{t+1}^2 - R_t^2}{1 - R_t^2} \\ &= \frac{1 - \frac{SCE_{t+1}}{SCT} - (1 - \frac{SCE_t}{SCT})}{1 - (1 - \frac{SCE_t}{SCT})} \\ &= \frac{SCE_t - SCE_{t+1}}{SCE_t} \end{aligned} \tag{1.3.4}$$

où l'indice t représente l'étape lorsqu'il y a déjà t variables incluses dans le modèle. SCE et SCT sont des abréviations pour la somme des carrés due à l'erreur et la somme des carrés totale de la régression de $X^{(1)}$ sur $X_{-j}^{(2)}$.

Finalement, on a bien

$$\begin{aligned} \frac{RI_{.X^{(2)}}}{1 - RI_{.X^{(2)}}} &= \frac{\frac{SCE_t - SCE_{t+1}}{SCE_t}}{\frac{1-(SCE_t - SCE_{t+1})}{SCE_t}} \\ &= \frac{SCE_t - SCE_{t+1}}{SCE_{t+1}} = F. \end{aligned} \tag{1.3.5}$$

Le test utilisé à l'étape 4 est alors remplacé par rejeter H_0 si $(n-2-t) \cdot F$ est plus grand qu'une loi F de Fisher avec 1 et $n-2-t$ degrés de liberté au niveau α.

En conclusion, c'est à l'aide de tous les éléments théoriques énoncés précédemment qu'on parviendra, au chapitre suivant, à définir différentes méthodes de détection de valeurs aberrantes en régression linéaire multivariée.

Chapitre 2

ALGORITHMES DE DÉTECTION DE DONNÉES ABERRANTES EN RÉGRESSION MULTIVARIÉE

Lors de ce second chapitre, on va parcourir quelques méthodes existantes dans la littérature permettant l'identification de données douteuses. Trois de ces algorithmes seront généralisés au cas multivarié avec utilisation de l'indice de Stewart et Love. Au départ, il sera question de la détection de valeurs aberrantes une à la fois. Puis, on s'intéressera plus particulièrement à des groupes d'observations pouvant être aberrants. Enfin, comme ces valeurs influencent beaucoup les méthodes statistiques, on voudra améliorer le critère de robustesse.

2.1. Utilisation de la régression successive dans l'identification de données aberrantes

Une première approche de l'identification de données aberrantes parvient de Mickey *et al.* (1967). Leur but consiste à pointer des observations qui sont suffisamment non usuelle pour justifier la reconsidération des données et du modèle. Ils ont donc mis au point une méthode permettant de détecter des valeurs influentes sous l'hypothèse de linéarité. Cette dernière met en pratique une nouvelle utilisation de l'inclusion successive (FORWARD) de variables en régression linéaire multiple.

Comme pour la plupart des méthodes, le critère de détection d'aberrances se base sur les résidus. En effet, Mickey *et al.* (1967) considèrent qu'une donnée peut être jugée aberrante si sa suppression entraîne une réduction importante de la somme des carrés due à l'erreur. Dans un contexte de sélection progressive, l'idée est de trouver l'observation qui une fois supprimée cause la plus grande réduction de la somme des carrés due à l'erreur. Puis, une fois cette observation identifiée, en trouver une autre dont la suppression réduit le plus la somme des carrés due à l'erreur et ainsi de suite. Les observations sont donc ordonnées selon leur apport à la matrice résiduelle.

2.1.1. Le cas de la régression linéaire multiple

Sous le contexte de la régression linéaire où $p = 1$, on a le modèle

$$\underset{(n \times 1)}{\underset{\sim}{Y}} = \underset{(n \times q)}{\mathbf{X}} \underset{(q \times 1)}{\underset{\sim}{\beta}} + \underset{(n \times 1)}{\underset{\sim}{\varepsilon}}.$$

Pour chaque observation à être enlevée on introduit une nouvelle variable qui vaut 1 pour le cas à être supprimé et 0 sinon. On augmente donc la matrice de données $\mathbf{X}$ par un matrice identité I_n de dimension $n \times n$:

$$\mathbf{X} \longrightarrow \mathbf{X}^* = (\mathbf{X}|I_n) : n \times (q+n)$$

et on augmente également le vecteur $\underset{\sim}{\beta}$ en un nouveau vecteur $\underset{\sim}{\beta}^* : (q+n) \times 1$

$$\underset{\sim}{\beta}' \longrightarrow \underset{\sim}{\beta}^{*\prime} = (\beta_1, \beta_2, \cdots, \beta_q, \nu_1, \nu_2, \cdots, \nu_n) = (\underset{\sim}{\beta}'|\underset{\sim}{\nu}').$$

Alors le modèle devient $\underset{\sim}{Y} = \mathbf{X}^* \underset{\sim}{\beta}^* + \underset{\sim}{\varepsilon} = \mathbf{X}\underset{\sim}{\beta} + I_n \underset{\sim}{\nu} + \underset{\sim}{\varepsilon}$.

L'algorithme consiste à forcer toutes les colonnes de la matrice $\mathbf{X}$ à entrer dans le modèle de régression. Puis, les autres colonnes de $\mathbf{X}^*$ (celles de la matrice identité) sont ajoutées successivement à l'aide de la procédure FORWARD. La première colonne à entrer dans le modèle selon cette méthode est celle qui réduit

le plus la somme des carrés due à l'erreur puisque l'on prendra $\hat{\nu}_k$ tel que ε_k devient égal à 0. Ce vecteur indique donc le cas qui influence le plus la matrice résiduelle et qui doit être considéré comme candidat à l'aberrance.

La procédure se termine lorsque la valeur de F (voir l'équation 1.3.5) calculée à chaque étape est plus petite qu'un nombre spécifié, c'est-à-dire lorsque la réduction maximale est trop petite. On a donc retenu un ensemble de variables qui est un sous-ensemble des colonnes de la matrice I_n. Si la j^e colonne de la matrice identité est choisie, alors le j^e cas est une valeur aberrante possible.

Finalement, cette méthode basée sur l'inclusion successive de variables en régression linéaire multiple ordonne les colonnes de la matrice identité par ordre décroissant d'influence sur la somme des carrés due à l'erreur. Elle ordonne donc les différentes observations par ordre décroissant d'influence sur la matrice résiduelle, ce qui est équivalent à l'influence sur R, le coefficient de corrélation multiple. En effet, si la j^e colonne de I_n entre en premier dans le modèle selon la procédure FORWARD, c'est que cette variable contribue le plus à réduire la somme des carrés due à l'erreur ou à augmenter le R^2. Les calculs en ajoutant les colonnes de la matrice identité une à une sont équivalents aux calculs en éliminant les cas correspondants un à un.

2.1.2. La généralisation au cas multivarié

On s'intéresse maintenant au modèle de régression linéaire multivariée décrit au premier chapitre à l'équation (1.1.3). Sous ce contexte, on veut généraliser la méthode de Mickey *et al.* (1967) vue précédemment. On procède donc de façon similaire.

On transforme d'abord la matrice $\mathbf{X}$ en une matrice augmentée $\mathbf{X}^*$

$$\mathbf{X} \longrightarrow \mathbf{X}^* = (\mathbf{X}|I_n) : n \times (q+n)$$

où I_n est la matrice identité de dimension $n \times n$.

Puis, on augmente la matrice $\mathcal{B}$ en une nouvelle matrice $\mathcal{B}^*$

$$\mathcal{B} \longrightarrow \mathcal{B}^* = (\mathcal{B}|\mathcal{V}) : (q+n) \times p$$

où $\mathcal{V} = (\underset{\sim}{\nu_1}\underset{\sim}{\nu_2}\cdots\underset{\sim}{\nu_p}) = (\nu_{ij}) : n \times p$.

Le modèle de l'équation (1.1.3) sous sa forme matricielle devient

$$\mathbf{Y} = \mathbf{X}^*\mathcal{B}^* + \mathcal{E} = \mathbf{X}\mathcal{B} + I_n\mathcal{V} + \mathcal{E}$$

tandis que le modèle empirique donné en (1.1.2) devient

$$Y_\alpha = \mathcal{B}' X_\alpha + \mathcal{V}_\alpha + \underset{\sim}{\epsilon_\alpha},\ \alpha = 1,2,\cdots,n.$$

On procède donc à une sélection de variables (FORWARD), tel que vu à la fin du premier chapitre, en forçant les X_α à entrer dans le modèle. Pour chaque colonne de la matrice identité qui entrera dans le modèle, quel que soit $\mathcal{B}$, on prendra $\hat{\mathcal{V}}_\alpha$ de sorte que $\underset{\sim}{\epsilon_\alpha} = 0$.

On a également vu que le critère de sélection de l'inclusion successive est basé sur la valeur de l'indice de redondance de Stewart et Love tel que défini en (1.2.5). Donc, la colonne de I_n qui entre dans le modèle est celle qui influence le plus à la hausse RI, ou le plus à la hausse l'ensemble des $R^2_{x_i^{(1)}\cdot x_1^{(2)},x_2^{(2)},\ldots,x_q^{(2)}}$ (coefficients de corrélation multiple), ou le plus à la baisse l'ensemble des sommes des carrés due à l'erreur des régressions linéaires multiples notées SCE_i.

Encore une fois, cette procédure permet d'ordonner en ordre décroissant les cas les plus influents, c'est-à-dire les candidats à l'aberrance. Notons que toutes les fonctions, programmées avec la version 3.4 du logiciel *S-Plus*, nécessaires à la mise en pratique de cet algorithme sont données à l'annexe A.

2.2. UTILISATION DE L'INFLUENCE D'UN POINT DANS L'IDENTIFICATION DE DONNÉES ABERRANTES

Depuis son introduction par Hampel (1974), la fonction d'influence d'un point X sur un paramètre θ est devenue un outil classique de la statistique. Plusieurs auteurs l'utilisent dans la détection de données aberrantes. La seconde méthode présentée au cours de cette section en est un exemple. Il s'agit d'une généralisation de la procédure de Cléroux *et al.* (1990) au modèle de régression linéaire multivariée. Leur méthode est basée sur la fonction d'influence d'une mesure d'association. Ici, on utilisera l'indice de redondance de Stewart et Love.

À partir de l'équation (1.2.9), on veut pour chaque observation x déterminer son influence sur RI. Ainsi, on pourra classer les observations en ordre décroissant d'influence. L'idée étant que l'observation qui a la plus grande influence est celle qui risque le plus d'être aberrante.

En pratique, on utilisera la fonction d'influence empirique $I\_(x_i; RI) = (n-1)(RI - RI_{-i})$ pour identifier les données douteuses. Puisque les paramètres $\mu^{(i)}$ et Σ_{ij},i,j = 1,2, sont inconnus, on les remplace par leurs estimateurs respectifs $\overline{X}^{(i)}$ et S_{ij}. On obtient ainsi les estimateurs des influences de tout point x_i sur RI notés $\hat{I}\_(x_i; RI)$.

De plus, pour chaque point on calcule également son influence relative à l'aide de la formule suivante :

$$100\frac{(\hat{I}\_(x_i,RI))}{(n-1)RI} = 100\frac{(RI - RI_{-i})}{RI}$$

afin d'identifier les candidats à l'aberrance.

Enfin, Cléroux *et al.* (1990) soupçonnent comme données douteuses tous les points x_i tels que

$$|\hat{I}\_(x_i; RI)| \geq 3\hat{\sigma}$$

où $\hat{\sigma}$ est obtenu à partir de l'écart type σ de la fonction d'influence (voir 1.2.11) en remplaçant les paramètres inconnus par leurs estimateurs habituels.

Notons qu'il s'agit ici d'une procédure heuristique qui s'avère toutefois moins compliquée que la procédure de test pour au plus deux aberrances proposée par Lazraq et Cléroux (1989). Dans ce dernier cas, on se base sur la distribution de la fonction d'influence théorique.

Posons $T = I(x; \rho I) = z'Qz$ tel que défini en (1.2.10) où z a comme moyenne le vecteur 0 et comme matrice de variance-covariance Σ. Si z est multinormale, alors T a la même distribution que $\sum_{i=1}^{p+q} \lambda_i W_i^2$ où les variables aléatoires W_i sont indépendantes et identiquement distribuées $\mathcal{N}(0,1)$ et où les λ_i sont les valeurs propres de la matrice ΣQ.

En pratique, les percentiles de cette distribution peuvent être calculés par une méthode exacte en utilisant l'algorithme de Imhof (1961). Cet algorithme est basé sur la propriété qu'une forme quadratique de variables aléatoires normales est distribuée comme une combinaison linéaire de variables aléatoires khi-deux dont la fonction caractéristique peut s'écrire facilement. La distribution exacte de la forme quadratique est alors obtenue en inversant numériquement la fonction caractéristique de la combinaison linéaire de variables aléatoires khi-deux correspondante. Une description détaillée de cet algorithme ainsi que la sous-routine "fquad" programmée en Fortran sont présentés par Koerts et Abrahamse (1969).

Le critère pour déterminer si une observation à l'écart est significativement aberrante est basé sur les valeurs extrêmes de la fonction d'influence. Pour chacune des observations, on a son influence notée $T_i = z_i' Q z_i$, $i = 1,2,\cdots,n$, puis on considère $T_{(n)} = \max_{1\leq i\leq n} T_i$ et $T_{(1)} = \min_{1\leq i\leq n} T_i$. La valeur de $T_{(n)}$ sera habituellement positive tandis que celle de $T_{(1)}$ sera habituellement négative. Notons la distribution de T par $F(t) = P[T \leq t]$. Puisque les T_i sont indépendants, la distribution de $T_{(n)}$ est $G(t) = [F(t)]^n$ et celle de $T_{(1)}$ est $H(t) = 1 - [1 - F(t)]^n$.

La procédure du test approximatif pour deux aberrances est donc la suivante :

1. Pour chacune des observations, calculer son influence estimée $\hat{I}(x_i; RI)$, $i = 1,2,\cdots,n$, obtenue en remplaçant Σ par S et ρI par RI dans l'équation (1.2.9) de l'influence théorique.
2. Considérer les points x_k et x_l tels que leur influence est donnée par les valeurs extrêmes $\max_{1\leq i\leq n} \hat{I}(x_i; RI) = \hat{I}(x_k; RI) = \gamma$ et $\min_{1\leq i\leq n} \hat{I}(x_i; RI) = \hat{I}(x_l; RI) = \delta$.
3. Calculer les valeurs propres de la matrice $S\hat{Q}$ où $\hat{Q}$ est obtenue en remplaçant Σ par S dans Q.
4. À l'aide de la sous-routine Fortran "fquad" (algorithme d'Imhof), calculer les probabilités $p_\gamma = F(\gamma)$ que T soit plus petit que γ et $p_\delta = F(\delta)$ que T soit plus petit que δ.
5. Calculer la probabilité d'excéder la plus grande valeur positive $p_{max} = P[T_{(n)} \geq \gamma] = 1 - P[T_{(n)} \leq \gamma]$ en utilisant $P[T_{(n)} \leq \gamma] = G(\gamma) = [F(\gamma)]^n = [p_\gamma]^n$.
6. Si cette probabilité est plus petite que α, alors considérer x_k comme une donnée aberrante au niveau α.
7. Semblablement pour x_l, calculer la probabilité de ne pas excéder la plus petite valeur négative $p_{min} = P[T_{(1)} \leq \delta] = H(\delta) = 1 - [1 - F(\delta)]^n = 1 - [1 - p_\delta]^n$.

8. Si cette probabilité est plus petite que α, alors considérer x_l comme une donnée aberrante au niveau α.

Enfin, pour détecter des données aberrantes en régression multivariée à l'aide de la fonction d'influence de RI, on pourra à la fois utiliser la procédure heuristique ainsi que le test approximatif pour au plus deux aberrances, tous deux décrits ci-haut.

2.3. AUTRES MÉTHODES DE DÉTECTION DE DONNÉES ABERRANTES MULTIVARIÉES

Dans cette section, on présente deux autres méthodes, provenant de la littérature, pour l'identification d'observations douteuses en régression linéaire multivariée. Encore une fois notons que les versions *S-Plus* de ces algorithmes ainsi que de ceux de la section précédente et de la section suivante se trouvent à l'annexe B.

2.3.1. La méthode de Srivastava et von Rosen

La méthode de Srivastava et von Rosen (1998) pour la détection de données douteuses dans des modèles linéaires multivariés se base sur un test du rapport de vraisemblance. On verra un peu plus loin que la statistique de test utilisée est en fait le maximum de n statistiques dépendantes, où n représente le nombre d'observations.

Srivastava et von Rosen procèdent avec deux modèles de régression linéaire multivariée. Le premier modèle noté H_0 et donné par

$$E(\mathbf{Y}) = \mathbf{X}\mathcal{B} \tag{2.3.1}$$

survient lorsqu'aucune des n observations n'est aberrante. Ce modèle est équivalent au modèle déjà vu en (1.1.3). Par contre, si le i^e cas est aberrant, à partir

du modèle empirique défini en (1.1.2), une translation dans la moyenne de cette observation peut s'exprimer de la façon suivante :

$$E(Y_i) = \mathcal{B}' X_i + \delta$$

où δ : $p \times 1$ est un vecteur de constantes inconnues. Dans ce dernier cas, on considère plutôt le modèle H_i donné par

$$E(\mathbf{Y}) = \mathbf{X}\mathcal{B} + \underset{\sim}{a_i}\delta' = (\mathbf{X}, \underset{\sim}{a_i}) \begin{pmatrix} \mathcal{B} \\ \delta' \end{pmatrix} = \mathbf{X}_i^* \mathcal{B}^*$$

où $\underset{\sim}{a_i}$ représente la i^e colonne de la matrice identité de dimension $n \times n$, c'est-à-dire que $I_n = (\underset{\sim}{a_1}, \underset{\sim}{a_2}, \cdots, \underset{\sim}{a_p})$.

Sous le modèle H_0, on a déjà trouvé les estimateurs à vraisemblance maximale de $\mathcal{B}$ et de Σ en (1.1.9) et en (1.1.10) respectivement. Posons $A = (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})$: $p \times p$ et $R = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$: $n \times n$. On a donc $\hat{\Sigma} = n^{-1}A$. La matrice A définie positive peut aussi s'écrire sous la forme suivante :

$$\begin{aligned} A &= (\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathcal{B}}) \\ &= (\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'(Y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= \mathbf{Y}'(I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'(I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= \mathbf{Y}'H'H\mathbf{Y} = \mathbf{Y}'H^2\mathbf{Y} = \mathbf{Y}'H\mathbf{Y} \\ &= \mathbf{Y}'(I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= \mathbf{Y}'(I_n - R)\mathbf{Y}. \end{aligned} \tag{2.3.2}$$

De la même façon, pour le modèle H_i, on obtient $\hat{\mathcal{B}}^* = (\mathbf{X}_i^{*\prime}\mathbf{X}_i^*)^{-1}\mathbf{X}_i^{*\prime}\mathbf{Y}$ l'estimateur à vraisemblance maximale de $\mathcal{B}^*$ ainsi que celui de Σ, $\hat{\Sigma}^* = n^{-1}A_i$ avec la matrice définie positive

$$A_i = \mathbf{Y}'(I_n - R_i)\mathbf{Y}$$

où $R_i = \mathbf{X}_i^*(\mathbf{X}_i^{*\prime}\mathbf{X}_i^*)^{-1}\mathbf{X}_i^{*\prime}$.

Soient $X_i : q \times 1$ et $Y_i : p \times 1$ les vecteurs du modèle empirique défini en (1.1.2), $i = 1,2,\cdots,n$. Supposons que $q \leq n - p - 2$. Alors pour le i^e cas on a les valeurs prédites $\hat{Y}_i = \hat{\mathcal{B}}' X_i$ et les résidus $\underset{\sim}{\hat{\epsilon}_i} = Y_i - \hat{Y}_i$ obtenus respectivement à partir des lignes des matrices $\hat{Y}$ et $\hat{\mathcal{E}}$ des équations (1.1.6) et (1.1.7). On note $r_{ii} = X_i'(\mathbf{X}'\mathbf{X})^{-1}X_i$. Il est possible de montrer que

$$A_i = A - (1 - r_{ii})^{-1}\underset{\sim}{\hat{\epsilon}_i}\underset{\sim}{\hat{\epsilon}_i}'.$$

Le test du rapport de vraisemblance pour l'hypothèse H_0 contre l'hypothèse H_i, sous la normalité, est donné par le rapport des deux estimateurs de Σ sous H_0 et H_i respectivement. Les grandes valeurs de ce rapport $\dfrac{|A|}{|A_i|}$ indiquent le rejet de l'hypothèse nulle. Pour chacune des n observations, soit la statistique

$$T_i = \frac{\underset{\sim}{\hat{\epsilon}_i}' A^{-1} \underset{\sim}{\hat{\epsilon}_i}}{1 - r_{ii}} \tag{2.3.3}$$

alors le rapport devient

$$\lambda_i = \frac{|A|}{|A_i|} = (1 - T_i)^{-1} = 1 + T_i(1 - T_i)^{-1}$$

et l'hypothèse H_0 est rejetée pour de grandes valeurs de T_i. D'où le test du rapport de vraisemblance maximale se base sur la statistique

$$T = \max_{1 \leq i \leq n} T_i$$

et est donné par rejeter l'hypothèse nulle si la valeur de T est grande. Pour obtenir le niveau de signification de la statistique T, on a besoin de sa distribution. Par contre, lorsque les T_i sont dépendants, celle-ci n'est pas disponible.

En pratique, dans une situation où on a $P(T \geq t) \leq \sum_{i=1}^{n} P(T_i \geq t)$, on utilisera un test conservateur basé sur la borne supérieure de Bonferroni.

Sous H_0, Srivastava et von Rosen (1998) montrent que les distributions des T_i sont identiques et que

$$\frac{f-p+1}{p}(\lambda_i - 1) = \frac{f-p+1}{p}\frac{T_i}{1-T_i} \sim F_{p,f-p+1}$$

où $f = n-q-1$ et $F_{r,s}$ dénote la distribution de Fisher avec r et s degrés de liberté.

Enfin, la valeur-p pour le test du rapport de vraisemblance de H_0 contre H_i appliquée à la statistique T, sera plus petite ou égale à la borne théorique suivante :

$$\begin{aligned} P(T \geq t) &\leq nP(T_1 \geq t) \\ &= nP\left(F_{p,f-p+1} \geq \frac{f-p+1}{p}\frac{t}{1-t}\right) \end{aligned}$$

et on considère le i^e cas ayant la valeur de T_i maximale comme étant aberrant si sa valeur-p est plus petite que le niveau de signification désiré.

2.3.2. La méthode de Naik

Également sous le modèle de la régression linéaire multivariée, l'idée proposée par Naik (1989) consiste à utiliser le coefficient d'aplatissement empirique multivarié des résidus défini par Mardia (1970).

Soit $\hat{\mathcal{E}}$: $n \times p$ la matrice des résidus telle que définie en (1.1.7) et soient $\hat{\epsilon_1}', \hat{\epsilon_2}', \cdots, \hat{\epsilon_n}'$ les n lignes de la matrice résiduelle $\hat{\mathcal{E}}$. La procédure de Naik consiste à calculer pour chacune des n observations la valeur de

$$\hat{\epsilon_i}' A^{-1} \hat{\epsilon_i}, i = 1,2,\cdots,n$$

où A est définie de la même façon que pour la méthode précédente en (2.3.2). Notons que Srivastava et von Rosen (1998) utilisent cette même forme quadratique qu'ils standardisent par $1 - r_{ii}$ tel que vu à l'équation (2.3.3).

Si une ou plusieurs de ces formes quadratiques sont inhabituellement grandes, alors les observations correspondantes seront considérées comme candidats à l'aberrance. Naik développe un test pour données aberrantes basé sur un sous-ensemble de résidus standardisés indépendants. Tout comme Srivastava et von Rosen (1998), le test de Naik considère des données aberrantes au sens de la translation dans la moyenne.

2.4. Détection d'ensembles de données aberrantes multivariées

Les différentes méthodes décrites précédemment permettent de détecter les données aberrantes isolées, mais pas nécessairement celles qui se retrouvent en groupes. En effet, ces méthodes souffrent de l'effet de "masquage" tel que nommé par Murphy (1951). Ce phénomène, d'abord discuté par Pearson et Chandra Sekar (1936), survient lorsque la présence d'observations extrêmes non déclarées comme données aberrantes a tendance à masquer l'influence d'autres observations encore plus extrêmes lors de la détection d'aberrances une à la fois. Par conséquent, on fait face au problème de masquage lorsqu'une donnée aberrante cache la présence d'une autre ou lorsque des observations sont influentes ensemble, mais non individuellement. Par exemple, considérons l'échantillon contenant les points à l'écart A et B tel qu'illustré par la figure 2.4.1.

Ici, A seul ne sera pas considéré comme une aberrance puisqu'il est masqué par B. Semblablement, A masque l'influence de B qui ne sera pas non plus considéré aberrant. Par contre, l'ensemble regroupant les deux points A et B sera détecté comme un ensemble aberrant. Donc, afin de déjouer l'effet de masquage dont souffrent les méthodes de détection d'une aberrance à la fois, on aura recours à un algorithme permettant de tester l'influence d'un groupe de plusieurs observations. C'est pourquoi, dans cette section, on s'intéresse principalement à la notion d'ensembles de points douteux. Plus particulièrement, on procédera à

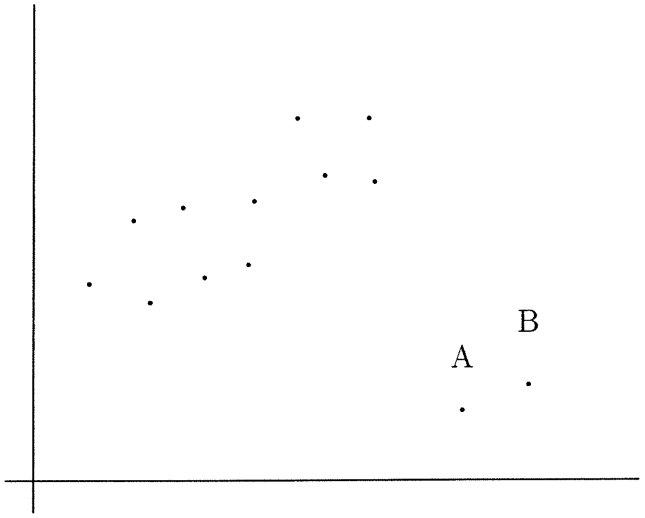

FIGURE 2.4.1. *Illustration de l'effet de masquage*

une généralisation des résultats obtenus à la section 2.2 sur la fonction d'influence du coefficient de corrélation vectorielle RI à la détection de groupes de données aberrantes.

Malheureusement, un danger alternatif à l'effet de masquage a été décrit par Fieller (1976). Ce dernier survient lorsque de fausses conclusions sont dues à l'effet qu'il nomme débordement (*swamping*). Plus précisément, il s'agit d'une observation non influente qui serait incorrectement incluse dans un groupe de données jugées aberrantes. Pour mieux comprendre cet autre problème, considérons l'exemple de la page 71 de Barnett et Lewis (1978). Soit l'échantillon contenant les valeurs 3, 4, 7, 8, 10, 13 et 951. Une procédure de détection d'ensembles aberrants appliquée au groupe contenant les valeurs 13 et 951 considérera cette paire comme étant aberrante. Or, la valeur 13 qui n'est pas exceptionnellement grande est déclarée aberrante à cause de son regroupement avec la valeur extrême 951. Par conséquent, le choix des ensembles à tester devient primordial.

Ainsi, afin d'éviter l'énumération systématique de tous les regroupements possibles et de contrer l'effet de débordement (*swamping*), on devra procéder à la formation de groupes pertinents. Pour ce faire, on aura recours à une méthode de classification pour former les groupes susceptibles d'être des ensembles douteux. Enfin, on sera en mesure de définir l'influence d'un groupe de points sur l'indice de redondance de Stewart et Love et donc de proposer une méthode heuristique ainsi qu'un test approximatif de détection d'ensembles de données aberrantes.

2.4.1. La classification hiérarchique

Tel que mentionné ci-haut, une étape primordiale à la détection d'ensembles de données aberrantes est la formation de groupes pertinents. Pour ce faire, il est nécessaire que l'ensemble des points du groupe constitue un amas compact de points. C'est pour cette raison qu'il semble naturel d'utiliser une procédure de classification hiérarchique. Or, par groupe pertinent, on entend un ensemble dont les éléments s'unissent à un bas niveau de distance dans la classification et restent ensembles longtemps avant de s'associer à d'autres points.

Avant d'aller plus loin, discutons un peu de la classification (référence : chapitre 12 de Johnson et Wichern, 1992). Il s'agit d'une technique de regroupement ne faisant aucune hypothèse quant au nombre de groupes ou à la structure des groupes. À cause de contraintes tel le temps disponible, on se voit dans l'impossibilité de déterminer le meilleur regroupement à partir de la liste de toutes les meilleures structures possibles. C'est pourquoi les algorithmes de classification recherchent un bon mais pas nécessairement le meilleur regroupement. Puisque le critère auquel on s'intéresse est la proximité entre les éléments regroupés, les méthodes de classification sont basées sur une distance.

Soient $U = (u_1,u_2,\cdots,u_p)$ et $V = (v_1,v_2,\cdots,v_p)$ deux vecteurs de taille p. En classification, on préfère souvent la distance Euclidienne définie par

$$\begin{aligned} d(U,V) &= \sqrt{(u_1-v_1)^2+(u_2-v_2)^2+\cdots+(u_p-v_p)^2} \\ &= \sqrt{(U-V)'(U-V)} \end{aligned}$$

à la distance statistique qui elle se définit par

$$d(U,V) = \sqrt{(U-V)'S^{-1}(U-V)}.$$

La raison est que la matrice de variance-covariance échantillonnale S ne peut pas être calculée sans une connaissance préalable des différents groupes.

La procédure générale des méthodes de classification hiérarchique est la suivante :

1. Initialement, chacun des éléments est considéré individuellement de façon à ce qu'il y ait autant de groupes que d'éléments. Puisque dans notre cas il s'agit des observations, on débute avec n groupes.
2. Sous le critère de la distance euclidienne, à chaque étape, on regroupe les deux cas qui se ressemblent le plus. Ces sous-groupes sont donc combinés selon leur similarité.
3. Éventuellement, à mesure que la similarité décroit, tous les sous-groupes seront fusionnés en un seul groupe.

Notons que le résultat des méthodes hiérarchiques peut être représenté par un diagramme en deux dimensions nommé dendogramme. Un tel dendogramme illustre alors les regroupements effectués à des niveaux successifs. C'est donc en étudiant ce dernier qu'on peut former les groupes pertinents.

Puis, parmi les procédures hiérarchiques on préférera pour la classification d'observations les méthodes de liaison. Cependant, il existe trois différentes méthodes de liaison dont voici les principales étapes du regroupement de n cas :

1. Au départ on a n groupes, contenant chacun une seule observation, ainsi qu'une matrice symétrique de distances notée $D = (d_{ij}) : n \times n$.
2. On identifie à travers la matrice D la paire de groupes les plus près. Soit la distance entre les groupes U et V notée d_{UV}.
3. On combine les groupes U et V en un nouveau groupe noté (UV) puis on met à jour la matrice des distances. Pour ce faire, on supprime d'abord les lignes et les colonnes correspondants aux groupes U et V. On ajoute ensuite une ligne et une colonne donnant les distances entre le groupe (UV) et les autres groupes restants notées $d_{(UV)W}$.
4. On calcule les distances précédentes selon la formule associée à la méthode de liaison employée.
   - Pour la méthode de liaison simple : $d_{(UV)W} = \min[d_{UW}, d_{VW}]$.
   - Pour la méthode de liaison complète : $d_{(UV)W} = \max[d_{UW}, d_{VW}]$.
   - Pour la méthode de liaison moyenne : $d_{(UV)W} = \frac{\sum_i \sum_j d_{ij}}{n_{(UV)} n_W}$ où $i \in (UV), j \in W$ et $n_{(UV)}$ et n_W représentent le nombre d'observations des groupes (UV) et W respectivement.
5. On répète $n-1$ fois les étapes 2, 3 et 4 jusqu'à ce que toutes les observations ne forment qu'un seul groupe. À chacun des niveaux, on note les groupes qui sont formés afin de produire le dendogramme associé.

Enfin, il ne reste plus qu'à choisir parmi les trois méthodes de liaison. Voyons leurs avantages. Tout d'abord, la méthode de liaison simple fusionne les groupes selon la distance entre leurs membres les plus près (voir figure 2.4.2). C'est-à-dire que les groupes sont joints par leur lien le plus court entre eux. Cette technique ne permet toutefois pas de discerner des groupes différents mais rapprochés ainsi qu'une configuration non elliptique tel qu'illustré à la figure 2.4.3 en a) et en b)

respectivement.

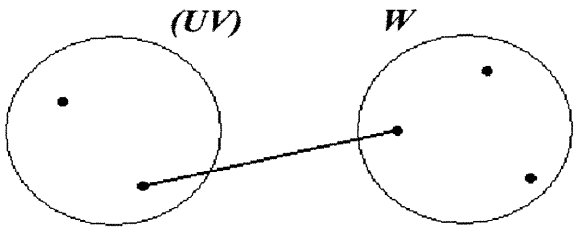

FIGURE 2.4.2. *Distance $d_{(UV)W}$ pour la méthode de liaison simple*

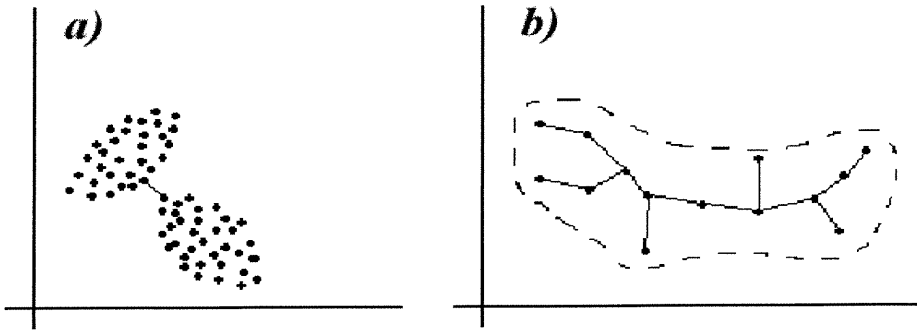

FIGURE 2.4.3. *Désavantage de la méthode de liaison simple*

À l'opposé, la méthode de liaison complète fusionne les groupes selon la distance entre leurs membres les plus éloignés (voir figure 2.4.4). Ceci a l'avantage d'assurer que tous les éléments d'un groupe sont à l'intérieur d'une distance maximale les uns des autres.

Puis, la méthode de liaison moyenne fusionne les groupes selon la distance moyenne entre les paires de membres des ensembles respectifs. Cette méthode se rapproche de la configuration de la liaison complète. Par contre, puisque les

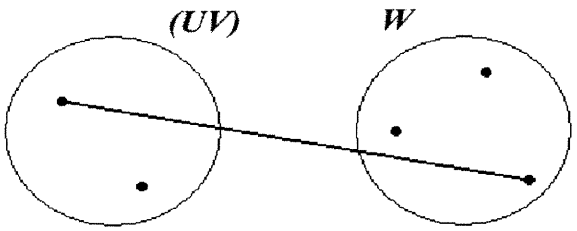

FIGURE 2.4.4. *Distance $d_{(UV)W}$ pour la méthode de liaison complète*

distances sont définies différemment pour chacune de ces deux méthodes, il n'est pas surprenant que les regroupements s'effectuent à des niveaux différents.

Or, le but étant de former des ensembles d'observations douteuses afin de détecter des groupes de données aberrantes, on demeure conservateur en choisissant la méthode de liaison complète. En effet, pour mieux identifier les données influentes, on ne voudrait pas qu'un candidat à l'aberrance soit facilement regroupé aux autres observations provoquant ainsi un effet de débordement (*swamping*).

2.4.2. Influence d'un groupe de points

Étant en mesure de former des groupes pertinents de candidats à l'aberrance, on veut maintenant tester l'influence de ces ensembles. On procède de façon similaire à Cléroux *et al.* (1990) en généralisant les fonctions d'influence. On veut pouvoir calculer l'influence d'un groupe de points sur l'indice de redondance RI.

Soit $\overline{X}_G = \dfrac{\sum_{i=1}^{m} X_i}{m}$ la moyenne des m points X_i d'un groupe G. Si les X_i sont des variables aléatoires indépendantes de moyenne μ et de matrice de variance-covariance Σ, alors on sait que $\overline{X}_G$ a pour moyenne μ et pour matrice de variance-covariance $\dfrac{\Sigma}{m}$.

En appliquant la formule (1.2.9) à $\overline{X}_G$, soit en remplaçant Σ_{11} par $\frac{\Sigma_{11}}{m}$ et Σ_{11}^* par

$$\Sigma_{11}^* = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \longrightarrow \frac{\Sigma_{12}}{m}\left(\frac{\Sigma_{22}}{m}\right)^{-1}\frac{\Sigma_{21}}{m} = \frac{1}{m}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \frac{\Sigma_{11}^*}{m}$$

et en conservant $B = \Sigma_{12}\Sigma_{22}^{-1}$, on obtient la fonction d'influence de $\overline{X}_G$ sur RI

$$I(\overline{X}_G; RI) = mRI\left[\frac{2U^{(1)'}BU^{(2)}}{tr(\Sigma_{11}^*)} - \frac{U^{(1)'}U^{(1)}}{tr(\Sigma_{11})} - \frac{U^{(2)'}B'BU^{(2)}}{tr(\Sigma_{11}^*)}\right] \qquad (2.4.1)$$

où $U = \begin{pmatrix} U^{(1)} \\ U^{(2)} \end{pmatrix} = \begin{pmatrix} \overline{X}_G^{(1)} - \mu^{(1)} \\ \overline{X}_G^{(2)} - \mu^{(2)} \end{pmatrix}$. Le résultat est très simple puisqu'il suffit de remplacer x par $\overline{X}_G$ dans l'équation (1.2.9) et de multiplier le résultat par m.

Par (2.4.1), et de façon similaire à la forme quadratique obtenue en (1.2.10), la fonction d'influence de $\overline{X}_G$ peut s'écrire

$$I(\overline{X}_G; RI) = U'QU.$$

Étant donné que la variance de U est $\frac{\Sigma}{m}$, on trouve que

$$Var[I(\overline{X}_G; RI)] = Var[mU'QU] = m^2 Var[U'QU] = m^2\left(\frac{\sigma}{m}\right)^2 = \sigma^2$$

où σ^2 est donné par l'équation (1.2.11).

Enfin, l'algorithme pour la détection d'ensembles de points douteux est le suivant :

1. Par une classification hiérarchique, avec la méthode de liaison complète basée sur la distance euclidienne, on forme les groupes pertinents.
2. Pour chacun de ces groupes, on calcule son influence échantillonnale selon l'équation (2.4.1).
3. On compare les influences calculées précédemment à la variance échantillonnale $\hat{\sigma}^2$ où $\hat{\sigma}^2$ est obtenu à partir de l'équation (1.2.11).

4. On considère aberrantes les observations d'un groupe dont l'influence en valeur absolue est plus grande que $3\hat{\sigma}$.

Encore une fois, il s'agit d'une procédure de test heuristique. Le test approximatif pour au plus deux aberrances basé sur l'algorithme d'Imhof (1961) effectué à la section 2.2 ne s'applique plus. En effet, puisqu'on définit arbitrairement, à l'aide de la classification hiérarchique, les groupes à tester alors on perd la notion d'influence minimale et maximale, d'où le test approximatif n'a plus de sens.

2.5. Robustesse dans les méthodes de détection d'aberrances

Dans la section précédente, lors de la définition de la procédure de détection de groupes d'observations aberrantes, en présence d'un échantillon on a dû estimer les paramètres inconnus. En effet, sous le modèle de la régression linéaire multivariée, on a entre autres estimé la matrice de variance-covariance Σ par $S = \frac{A}{n-1}$ où A est donnée en (2.3.2). Cette estimation est nécessaire pour le calcul de l'indice de redondance de Stewart et Love, des fonctions d'influence et de la variance échantillonnale $\hat{\sigma}^2$.

Or, l'estimateur S est grandement influencé par la présence d'une ou de plusieurs données aberrantes. Pour contrer ce problème, on a recours à des estimateurs robustes. On utilise les mêmes que ceux employés par Cléroux *et al.* (1990). Il s'agit des estimateurs robustes pour la moyenne et pour la matrice de variance-covariance proposés par Huber (1977). Ils sont obtenus à l'aide de la fonction "robuste" de la librairie "DomouSe" de la version 3.4 du logiciel *S-Plus*. Cette dernière fait appel aux sous-routines "robsts" programmées en Fortran par Marazzi (1985).

On reprendra les différentes méthodes de ce chapitre en remplaçant les estimateurs habituels par ces estimateurs robustes. Plus précisément, on remplacera Σ par son estimateur robuste plutôt que par S pour calculer les fonctions d'influence de RI estimées ainsi que l'estimé de sa variance $\hat{\sigma}^2$. D'autre part, dans les méthodes de Srivastava et von Rosen (1998) et de Naik (1989), on remplacera la matrice A par l'estimateur robuste de Σ multiplié par $(n-1)$. Seulement dans le cas de la méthode de Mickey *et al.* (1967), on ne pourra pas appliquer cet estimateur robuste puisque la matrice $\mathbf{X}$ est augmentée de la matrice identité I_n. Quant à l'inférence, ni la borne théorique de Srivastava et von Rosen ni le test approximatif basé sur l'algorithme d'Imhof (1961) ne pourront être appliqués dans le cas robuste puiqu'ils se basent sur la loi de l'estimateur S.

En conclusion, qu'elles se retrouvent seules ou en groupes, on connaît maintenant divers algorithmes de détection de données aberrantes en régression linéaire multivariée. Au chapitre suivant, on comparera ces différentes méthodes afin de déterminer les plus performantes dans différentes situations. Pour ce faire, on les appliquera à quelques jeux de données pertinents. On s'intéressera surtout aux avantages et aux désavantages des méthodes de détection d'une aberrance à la fois versus la méthode d'identification de groupes de données aberrantes. Aussi, on voudra savoir s'il y a une différence significative entre l'utilisation d'un algorithme et sa version robuste.

Chapitre 3

COMPARAISON DES DIVERS ALGORITHMES DE DÉTECTION D'ABERRANCES EN RÉGRESSION MULTIVARIÉE

Au cours de ce dernier chapitre, on va appliquer chacun des algorithmes du chapitre précédent à quelques exemples pertinents. Ensuite, on comparera tous les résultats ainsi obtenus. On tentera d'identifier les avantages et les désavantages des différentes méthodes de détection de données aberrantes en régression linéaire multivariée. Est-ce que l'un de ces algorithmes est généralement plus performant que les autres? L'utilisation d'estimateurs robustes permet-elle d'améliorer significativement les méthodes? L'identification de groupes d'aberrances est-elle plus efficace que la détection de données aberrantes une à la fois? Voilà le type de questions auxquelles on tentera de répondre.

3.1. Cas particulier de la régression linéaire multiple

Tout d'abord, on veut tester chacune des méthodes sur un exemple où $p = 1$. Pour ce faire, on choisit les données présentées au chapitre 5 de Daniel et Wood (1971). Ces dernières se rapportent à l'opération d'un mécanisme provoquant l'oxydation de l'ammoniac en acide nitrique. La variable à expliquer est 10 fois le pourcentage d'ammoniac entrant dans le mécanisme qui s'échappe sous forme d'oxyde nitrique. Cette variable est une mesure inverse de l'efficacité globale du mécanisme. Le vecteur $X^{(2)}$ contient trois variables contrôlables. La première est l'entrée d'air dans le mécanisme, soit le taux d'opération du mécanisme. La

deuxième est la température de l'eau froide qui entre dans la tour d'absorption d'oxyde nitrique et la troisième est la concentration de l'acide nitrique produit qui se trouve dans le liquide d'absorption. Ces données représentent 21 jours consécutifs d'opération du mécanisme, soit 21 observations pour chacune des variables.

Puisque ce jeu de données a auparavant été étudié par différents auteurs, on pourra comparer nos résultats à ceux déjà obtenus. En effet, Daniel et Wood (1971), après une analyse approfondie, concluent que les observations 1, 3, 4, et 21 sont aberrantes. Andrews et Pregibon (1978) ont également identifié les mêmes aberrances, mais ils ont ajouté que l'observation 2 peut aussi être considérée comme extrêmement étrange. D'autre part, Cook (1979) a trouvé que les observations 2, 4, et 21 sont extrêmement influentes sur une autre mesure de distances. Voyons maintenant ce qu'on obtient à l'aide des divers algorithmes du chapitre 2.

Premièrement, regardons le tableau 3.1.1 qui présente les étapes de la méthode de Mickey *et al.* (1967). On remarque que les observations 21 et 4 sont les deux premières à entrer dans le modèle de régression de façon très significative avec une valeur-p de 0,004. Celles-ci peuvent donc être considérées comme aberrantes. Puis, au niveau 5%, on pourrait porter un intérêt particulier aux observations douteuses 3, 1, et 13 avec les valeurs-p respectives de 0,0381, 0,0021 et de 0,0185. Notons que les observations 14, 8 et 7 ont également une valeur-p inférieure à 5%, mais elles entrent trop tard dans le modèle pour être considérées aberrantes selon cette procédure.

En second lieu, jetons un coup d'oeil aux estimés des fonctions d'influence théorique et empirique de RI du tableau 3.1.2. Ces derniers sont illustrés à la figure 3.1.1. On remarque que l'ordre des variables les plus influentes est le même pour les deux fonctions. La valeur calculée de RI est très élevée, soit de 0,9136,

TABLEAU 3.1.1. *La méthode de Mickey et al. (1967) appliquée aux données de Daniel et Wood (1971)*

Observation	*RI* partiel	*RI*	Valeur-p	Fisher
21	0,4094	0,9490	0,0042	11,0922
4	0,4339	0,9711	0,0040	11,4990
3	0,2724	0,9790	0,0381	5,2403
1	0,5310	0,9901	0,0021	14,7198
13	0,3821	0,9939	0,0185	7,4218
20	0,2499	0,9954	0,0819	3,6648
2	0,3275	0,9969	0,0519	4,8695
14	0,5388	0,9986	0,0101	10,5155
8	0,4417	0,9992	0,0360	6,3294
16	0,4215	0,9995	0,0585	5,0997
12	0,3231	0,9997	0,1415	2,8643
19	0,5445	0,9999	0,0583	5,9771
7	0,8973	1	0,0041	34,9648
5	0,6891	1	0,0819	6,6481
15	0,6116	1	0,2180	3,1491
6	1	1	1	3,1491

et celle de $\hat{\sigma}$ est de 0,1652. Ce qui nous donne la valeur de 0,4956 pour trois écarts types. D'où la procédure heuristique déclare que seule l'observation 21 est aberrante selon la fonction d'influence théorique. Si on appliquait cette même heuristique à l'influence empirique, alors l'observation 1 serait également aberrante. Par le test approximatif pour au plus deux aberrances, on rejette l'hypothèse que les observations les plus influentes, soient 21 et 1, sont aberrantes avec une valeur-p respective de 0,1811 et de 0,3324.

TABLEAU 3.1.2. *Estimés des fonctions d'influence théorique et empirique des données de Daniel et Wood (1971)*

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}\_(x_i; RI)$	Influence relative (%)
21	**-0,5010**	-2,7419	21	**-0,7044**	-3,8550
1	0,3993	2,1853	1	**0,5469**	2,9932
2	0,2813	1,5396	2	0,3475	1,9018
4	-0,2221	-1,2156	4	-0,2793	-1,5289
3	0,1163	0,6363	3	0,1111	0,6078
9	-0,0902	-0,4939	9	-0,1059	-0,5796
6	-0,0872	-0,4773	6	-0,0945	-0,5174
16	0,0846	0,4630	16	0,0933	0,5105
18	0,0738	0,4037	18	0,0809	0,4427
12	-0,0576	-0,3151	12	-0,0782	-0,4282
19	0,0572	0,3132	11	-0,0690	-0,3779
11	-0,0568	-0,3109	7	-0,0688	-0,3768
17	0,0534	0,2925	19	0,0618	0,3383
7	-0,0534	-0,2921	17	0,0436	0,2385
5	-0,0281	-0,1539	5	-0,0297	-0,1624
14	0,0255	0,1394	14	0,0272	0,1486
15	0,0219	0,1197	8	-0,0186	-0,1017
13	0,0158	0,0866	20	-0,0154	-0,0844
20	-0,0140	-0,0764	13	0,0142	0,0778
8	-0,0135	-0,0741	15	0,0136	0,0744
10	-0,0051	-0,0282	10	-0,0086	-0,0469

Pour ce qui est des deux méthodes basées sur une forme quadratique des résidus, les résultats apparaissent au tableau 3.1.3. Pour la méthode de Naik (1989),

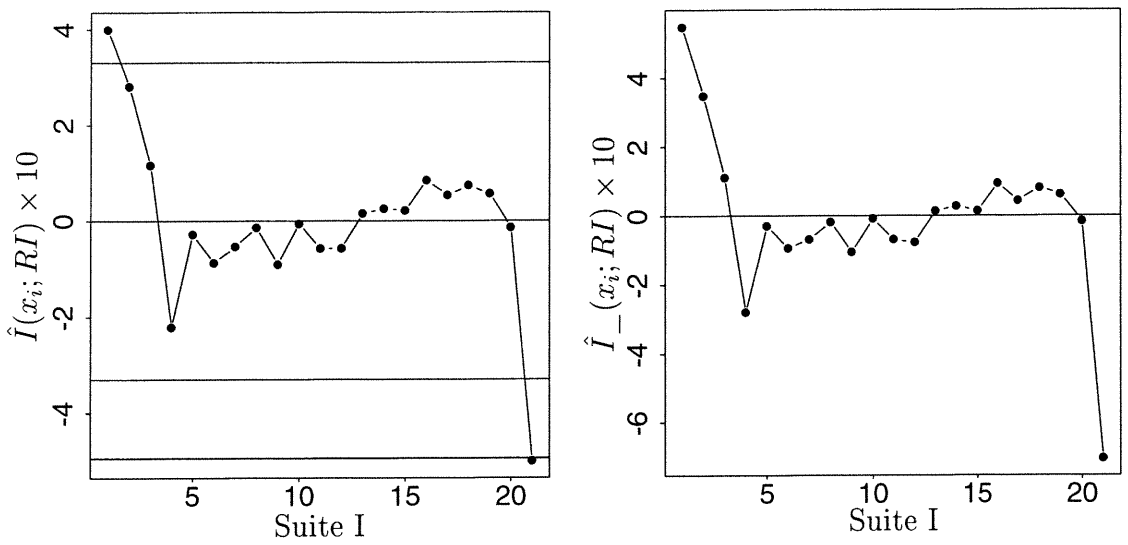

FIGURE 3.1.1. *Graphiques des estimés des fonctions d'influence théorique et empirique de RI pour les données de Daniel et Wood (1971)*

les trois plus grandes formes quadratiques qui se distinguent des autres sont celles des observations 21, 4 et 3. D'où ces dernières sont considérées comme candidats à l'aberrance. Semblablement, par la méthode de Srivastava et von Rosen (1998), on soupçonne les mêmes observations mais également l'observation 1. Pour la valeur maximale de T_i correspondant à l'observation 21, on calcule la borne théorique. Cette dernière étant de 0,6740, on ne rejette pas l'hypothèse que l'observation 21 n'est pas une donnée aberrante.

Voyons maintenant si les résultats précédents tiennent dans le cas d'estimateurs robustes. Les estimations robustes des deux fonctions d'influence se trouvent au tableau 3.1.4 et sont illustrées à la figure 3.1.2. Dans ce cas-ci, on obtient un *RI* robuste encore plus élevé, soit de 0,9288, ainsi qu'un plus petit écart type robuste estimé de 0,1373. La valeur de trois écarts types est donc de 0,4118 et la même observation (21) est considérée aberrante par la procédure heuristique

TABLEAU 3.1.3. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données de Daniel et Wood (1971)*

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
21	0,2428	21	0,2929	21	0,0719	21	0,0522
4	0,1249	4	0,1815	4	0,0370	4	0,0323
3	0,0834	3	0,1160	3	0,0247	3	0,0207
1	0,0501	1	0,0585	1	0,0148	1	0,0104
9	0,0385	9	0,0553	9	0,0114	9	0,0098
12	0,0332	6	0,0506	12	0,0098	6	0,0090
6	0,0329	12	0,0432	6	0,0097	12	0,0077
11	0,0275	11	0,0389	11	0,0081	11	0,0069
15	0,0227	7	0,0319	15	0,0067	7	0,0057
7	0,0218	15	0,0312	7	0,0065	15	0,0056
2	0,0181	2	0,0206	2	0,0054	2	0,0037
5	0,0104	5	0,0164	5	0,0031	5	0,0029
17	0,0083	17	0,0129	17	0,0025	17	0,0023
13	0,0077	13	0,0114	13	0,0023	20	0,0020
8	0,0074	20	0,0112	8	0,0022	13	0,0020
20	0,0070	8	0,0108	20	0,0021	8	0,0019
10	0,0060	10	0,0090	10	0,0018	10	0,0016
16	0,0032	16	0,0046	16	0,0009	16	0,0008
19	0,0014	19	0,0020	19	0,0004	19	0,0004
18	0,0008	18	0,0012	18	0,0002	18	0,0002
14	0,0000	14	0,0000	14	0,0000	14	0,0000

pour la fonction d'influence théorique robuste estimée. Par contre, dans le cas

TABLEAU 3.1.4. *Estimés robustes des fonctions d'influence théorique et empirique des données de Daniel et Wood (1971)*

Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)	Observation	$\hat{I}\_(x_i; RI)$ robuste	Influence relative (%)
21	**-0,6348**	-3,4172	1	**0,8511**	4,5817
1	0,3153	1,6975	2	**0,6517**	3,5081
4	-0,2826	-1,5213	3	**0,4152**	2,2353
2	0,2051	1,1041	21	-0,4002	-2,1544
3	0,0872	0,4695	16	0,3974	2,1396
6	-0,0688	-0,3703	18	0,3851	2,0730
16	0,0626	0,3373	19	0,3660	1,9702
18	0,0598	0,3218	17	0,3478	1,8721
9	-0,0561	-0,3020	14	0,3313	1,7837
17	0,0521	0,2806	13	0,3184	1,7141
19	0,0478	0,2574	15	0,3178	1,7107
7	-0,0355	-0,1912	10	0,2956	1,5914
11	-0,0330	-0,1775	20	0,2888	1,5545
12	-0,0255	-0,1372	8	0,2856	1,5375
5	-0,0251	-0,1350	5	0,2745	1,4777
20	-0,0194	-0,1043	7	0,2353	1,2669
14	0,0184	0,0991	11	0,2351	1,2658
8	-0,0052	-0,0279	12	0,2259	1,2163
15	0,0037	0,0197	6	0,2096	1,1285
13	-0,0031	-0,0169	9	0,1983	1,0674
10	0,0008	0,0042	4	0,0248	0,1336

de l'influence empirique robuste estimée, les observations 1, 2, et 3 dépassent le seuil de 0,4118. Notons toutefois qu'il est moins juste d'appliquer cet intervalle ici

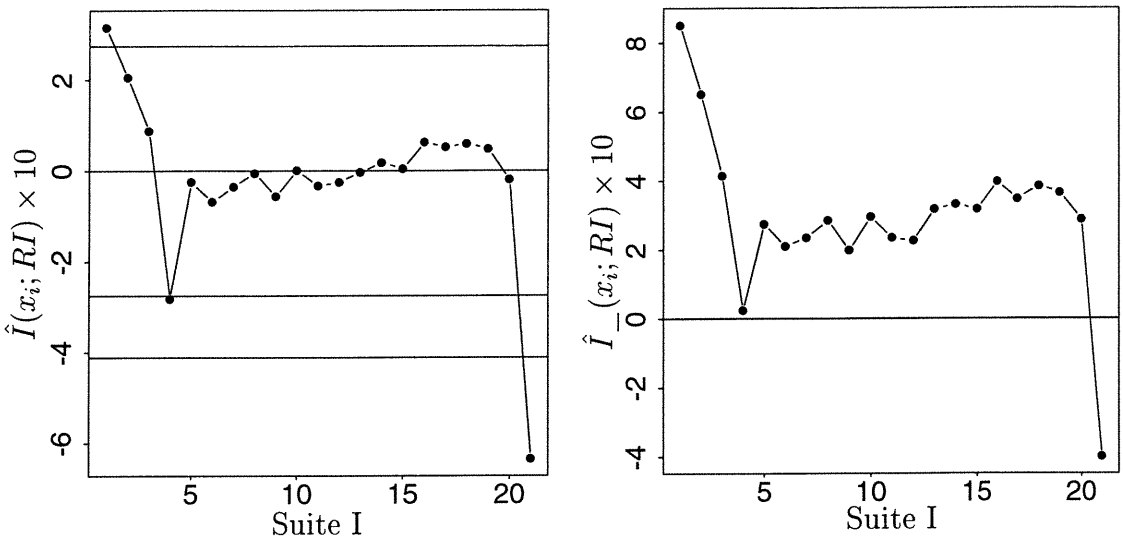

FIGURE 3.1.2. *Graphiques des estimés robustes des fonctions d'influence théorique et empirique de RI pour les données de Daniel et Wood (1971)*

puisque la variance (σ^2) est calculée à partir de l'équation de l'influence théorique.

Finalement, pour ce qui est des formes quadratiques robustes du tableau 3.1.3, les résultats sont sensiblement les mêmes qu'auparavant.

À l'aide des résultats de la classification hiérarchique présentés à la figure 3.1.3, on classe les observations en neuf groupes pertinents. Pour ce faire, on trace de façon *ad hoc* une coupure, indiquée par un trait d'axe pointillé, de manière à retenir que les ensembles dont les éléments s'unissent à un bas niveau de distance dans la classification et restent ensembles longtemps avant de s'associer à d'autres points. Notons que seuls les groupes de cardinalité inférieure ou égale à 15% sont considérés. L'influence théorique ainsi que sa version robuste pour chacun de ces groupes sont calculées au tableau 3.1.5. Rappelons qu'on a 0,4956 et 0,4118 comme valeur de $3\hat{\sigma}$ pour le cas habituel et pour le cas robuste

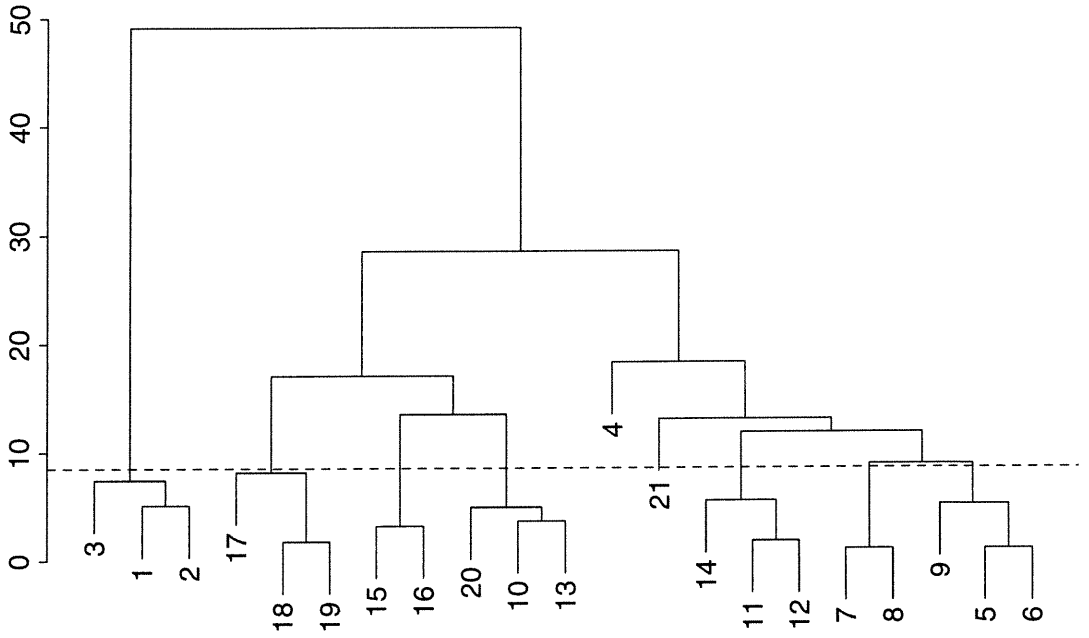

FIGURE 3.1.3. *Dendogramme pour les données de Daniel et Wood (1971)*

respectivement. On déclare donc aberrants deux groupes, soit le premier comportant les observations 1, 2 et 3 ainsi que le sixième correspondant à l'observation 21.

En conclusion, pour cet exemple des données de Daniel et Wood (1971) en régression linéaire multiple, de façon globale, par tous les algorithmes, on considère l'observation 21 comme donnée aberrante. Les résultats obtenus par les deux méthodes basées sur les formes quadratiques coïncident avec ceux de Daniel et Wood (1971). Tout comme Andrews et Pregibon (1978), seule la méthode pour la détection d'ensembles aberrants considérera également influente l'observation 2.

TABLEAU 3.1.5. *Influence (robuste) d'un groupe de points pour les données de Daniel et Wood (1971)*

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{1, 2, 3\}$	**1,0091**	**0,8092**
$G_2 = \{4\}$	-0,2221	-0,2826
$G_3 = \{17, 18, 19\}$	0,1903	0,1651
$G_4 = \{15, 16\}$	0,1163	0,0751
$G_5 = \{10, 13, 20\}$	0,0390	0,0334
$G_6 = \{21\}$	**-0,5010**	**-0,6348**
$G_7 = \{11, 12, 14\}$	-0,0415	-0,0025
$G_8 = \{7, 8\}$	-0,0625	-0,0364
$G_9 = \{5, 6, 9\}$	-0,1985	-0,1477

Dorénavant pour les exemples de la section suivante, afin d'alléger la lecture, on ne présentera dans les tableaux que les dix observations les plus douteuses pour chacune des méthodes. Aussi, on calculera uniquement la fonction d'influence théorique de RI et non celle empirique. Ce qui permettra de conserver les tests basés sur de l'inférence s'appliquant seulement à la fonction d'influence théorique.

3.2. QUELQUES EXEMPLES EN RÉGRESSION LINÉAIRE MULTIVARIÉE

Ayant déjà étudié le cas particulier de la régression linéaire lorsque $p = 1$ à la section précédente, on s'intéresse maintenant au cas multivarié. Pour ce faire, on analysera les résultats de cinq différents jeux de données en régression multivariée.

3.2.1. Les données sur le tabac

Débutons avec les données de Woltz *et al.* (1948). Il s'agit de données relatives aux constituants organiques et inorganiques d'un échantillon de 25 feuilles

TABLEAU 3.2.1. *Généralisation de la méthode de Mickey et al. (1967) appliquée aux données TABAC*

Observation	*RI* partiel	*RI*	Valeur-p
22	0,2659	0,8055	0,0202
6	0,2301	0,8503	0,0373
14	0,2964	0,8946	0,0189
1	0,2166	0,9175	0,0579
3	0,2944	0,9418	0,0273
2	0,2400	0,9557	0,0620
9	0,2311	0,9660	0,0812
16	0,2875	0,9758	0,0576
15	0,2640	0,9822	0,0845
4	0,2055	0,9858	0,1637

de tabac. Sur chacune des feuilles, ils ont relevé neuf variables. La première mesure le taux de consommation en pouces par 1000 secondes tandis que les huit autres représentent les pourcentages de sucre, de nicotine, d'azote, de chlore, de potassium, de phosphore, de calcium et de magnésium. Par la régression linéaire multivariée, on veut expliquer le vecteur composé des trois premières variables par le vecteur contenant les six derniers pourcentages.

La généralisation de l'algorithme de Mickey *et al.* (1967) nous donne les résultats du tableau 3.2.1. Au niveau 5%, on considère comme candidats à l'aberrance les observations 22, 6, et 14.

La régression multivariée de ces données donne un *RI* élevé de 0,7351 et un écart type de 0,4361. Dans le cas robuste, ces valeurs passent à 0,7270 et à 0,4471 respectivement. Or, dans les deux cas, si on regarde le tableau 3.2.2 et la figure 3.2.1, on voit qu'aucune observation n'a une influence qui dépasse trois écarts

TABLEAU 3.2.2. *Estimés (habituels et robustes) de la fonction d'influence théorique des données TABAC*

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)
22	-1,2575	-7,1280	22	-1,2993	-7,4463
10	0,5830	3,3044	10	0,6565	3,7624
14	-0,5601	-3,1748	14	-0,5773	-3,3085
2	0,5537	3,1386	2	0,5588	3,2025
6	-0,4363	-2,4730	6	-0,4583	-2,6263
5	0,4003	2,2691	5	0,4256	2,4390
7	0,2825	1,6010	7	0,2955	1,6938
12	-0,2686	-1,5224	15	0,2805	1,6075
15	0,2607	1,4778	3	-0,2791	-1,5995
23	0,2550	1,4456	12	-0,2778	-1,5920

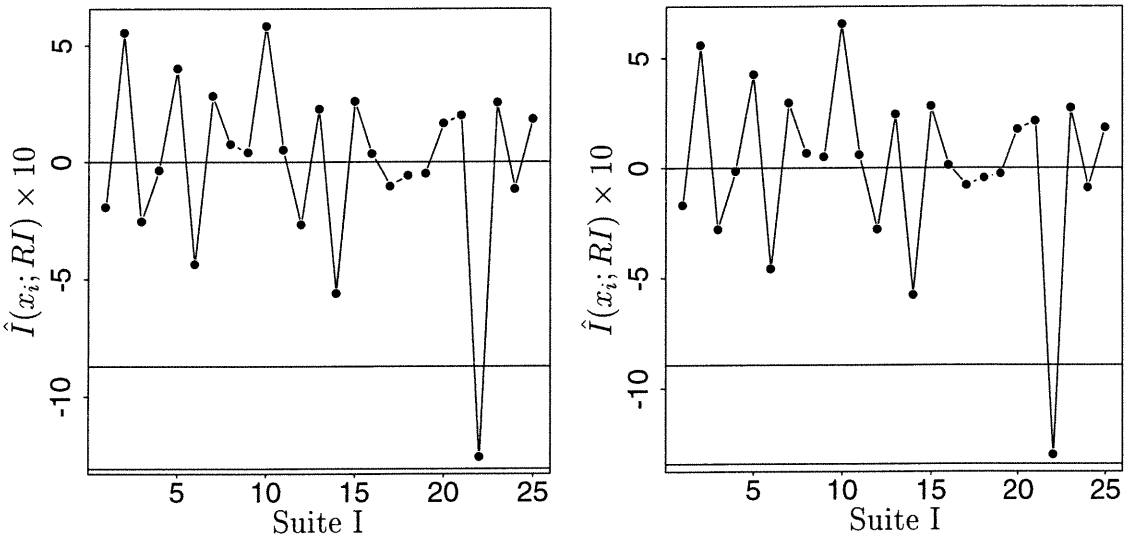

FIGURE 3.2.1. *Graphiques des estimés (habituels et robustes) des fonctions d'influence théorique de RI pour les données TABAC*

types en valeur absolue. Par contre, l'observation 22 se démarque des autres avec

TABLEAU 3.2.3. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données TABAC*

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
2	0,4164	22	0,4012	2	0,2929	6	0,1983
22	0,3082	2	0,2877	6	0,2696	22	0,1968
6	0,2672	1	0,2844	22	0,2273	2	0,1930
1	0,2648	6	0,1989	14	0,1068	14	0,0841
24	0,2297	3	0,1776	4	0,1049	4	0,0817
3	0,2234	14	0,1496	23	0,0909	23	0,0707
10	0,2154	24	0,1462	17	0,0798	21	0,0647
14	0,1901	23	0,1428	21	0,0779	13	0,0597
23	0,1827	10	0,1420	10	0,0778	11	0,0586
20	0,1339	20	0,1362	9	0,0763	9	0,0576

une influence relative de -7,13% et une influence relative robuste de -7,45%. Puis, les observations 10, 14 et 2 suivent avec une influence un peu moins grande.

À la vue du tableau 3.2.3, l'observation 2 correspondant à la valeur de T_i maximale selon la méthode de Srivastava et von Rosen (1998) n'est toutefois pas aberrante si on la compare à la borne théorique de 0,7778. Les autres plus grandes valeurs de T_i correspondent à l'observation 22 puis aux observations 6 et 1 auxquelles on devrait porter une attention particulière. D'un autre coté, selon l'algorithme de Naik (1989), les observations se distinguant par leur grande forme quadratique sont dans l'ordre 22, 2 et 1. Les versions robustes de ces deux méthodes concordent en déclarant aberrantes les observations 2, 6 et 22.

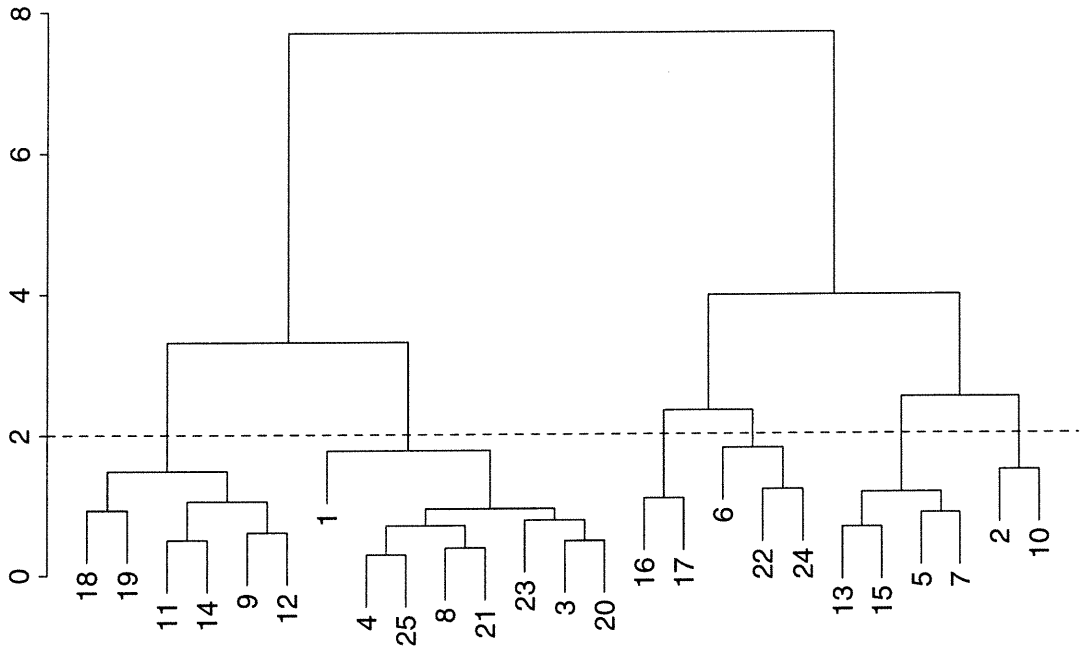

FIGURE 3.2.2. *Dendogramme pour les données TABAC*

À l'aide du dendogramme de la figure 3.2.2, on forme quatre groupes pertinents. Pour chacun de ces groupes d'observations, regardons son influence donnée au tableau 3.2.4. Pour trois écarts types on a calculé une valeur de 1,3083 et de 1,3413 dans le cas robuste. Or, selon la procédure heuristique, le premier groupe comprenant les observations 2 et 10 ainsi que le quatrième comprenant les observations 6, 22 et 24 sont tous deux déclarés aberrants.

De façon globale, tous les algorithmes considèrent comme aberrance la feuille de tabac numéro 22. Puis dans la majorité des cas, les observations 2 et 6 sont également considérées aberrantes. Les observations 1, 10, 14 et 24 sont parfois soupçonnées. Pour l'ensemble des données sur le tabac, les versions robustes performent mieux que les deux méthodes correspondantes basées sur des formes

TABLEAU 3.2.4. *Influence (robuste) d'un groupe de points pour les données TABAC*

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{2, 10\}$	**1,5616**	**1,6203**
$G_2 = \{5, 7, 13, 15\}$	1,2548	1,3305
$G_3 = \{16, 17\}$	0,2045	0,2211
$G_4 = \{6, 22, 24\}$	**-1,5203**	**-1,528**

quadratiques. Bien qu'au niveau de la robustesse il ne semble pas y avoir une différence importante dans le cas des fonctions d'influence de RI, celle appliquée à des groupes de données détecte plus de candidats à l'aberrance que lorsqu'on procède pour une observation à la fois.

3.2.2. Les données ventes

Le jeu de données analysé au cours de cette section est celui présenté à la page 456 de Johnson et Wichern (1992). Il provient d'une étude sur le rendement de vendeurs par rapport à leurs capacités intellectuelles. Les trois variables à expliquer sont la croissance des ventes, le profit provenant des ventes et les ventes relatives à de nouveaux clients. Tandis que les quatre variables contrôlables représentent respectivement une note d'esprit de créativité, d'habileté mécanique, d'esprit d'abstraction et une note d'esprit mathématique. Ce jeu de données nommé "ventes" comporte 50 observations pour chacune des sept variables.

Au tableau 3.2.5, on trouve les résultats provenant de la généralisation de la méthode de Mickey *et al.* (1967). On voit que les deux premières variables à entrer dans le modèle de régression multivariée correspondent aux observations 8 et 10. Ces dernières peuvent être considérées aberrantes avec une valeur-p significative

TABLEAU 3.2.5. *Généralisation de la méthode de Mickey et al. (1967) appliquée aux données VENTES*

Observation	*RI* partiel	*RI*	Valeur-p
8	0,1678	0,9656	0,0002
10	0,1009	0,9691	0,0067
48	0,0767	0,9714	0,0286
28	0,0842	0,9739	0,0214
32	0,0907	0,9762	0,0181
19	0,0898	0,9784	0,0218
44	0,0967	0,9805	0,0174
22	0,1145	0,9827	0,0092
16	0,1327	0,9850	0,0043
36	0,1489	0,9872	0,0026

de 0,0002 et de 0,0067.

Encore une fois, lors de la régression linéaire multivariée, on obtient un *RI* très élevé, soit de 0,9587, et un *RI* robuste de 0,9644. L'écart type estimé étant de 0,0663 et de 0,0575 dans le cas robuste, ceci nous donne les valeurs respectives de 0,1989 et de 0,1725 pour $3\hat{\sigma}$. Comparons à ces deux bornes la valeur absolue des influences du tableau 3.2.6 illustrées à la figure 3.2.3. Seule l'observation 8 est à l'extérieur de l'intervalle ainsi formé dans le cas robuste avec une influence de -0,2738 et peut donc être considérée aberrante selon cette heuristique. Par contre, si on s'en tient aux influences relatives, l'observation 10 se distingue également des autres et devient ainsi un deuxième candidat à l'aberrance. Quant au test approximatif basé sur l'algorithme d'Imhof, on ne rejette pas l'hypothèse que les deux observations 8 et 44 ayant l'influence minimale et maximale ne sont pas aberrantes avec une valeur-p de 0,0933 et de 0,8362.

TABLEAU 3.2.6. *Estimés (habituels et robustes) de la fonction d'influence théorique des données VENTES*

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)
8	-0,1634	-0,3479	8	**-0,2738**	-0,5793
10	-0,1207	-0,2570	10	-0,1565	-0,3311
44	0,0887	0,1889	19	-0,1093	-0,2312
23	0,0838	0,1784	22	-0,0896	-0,1895
19	-0,0803	-0,1710	23	0,0863	0,1826
2	0,0791	0,1683	2	0,0751	0,1590
29	0,0685	0,1458	29	0,0687	0,1454
21	0,0619	0,1319	38	-0,0617	-0,1306
22	-0,0555	-0,1182	21	0,0574	0,1214
30	0,0501	0,1066	35	0,0558	0,1181

TABLEAU 3.2.7. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données VENTES*

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
10	0,2331	10	0,3924	10	0,1477	10	0,1405
8	0,2039	8	0,3133	19	0,1021	8	0,0926
48	0,1288	4	0,1821	8	0,0995	19	0,0902
22	0,1135	28	0,1574	22	0,0818	16	0,0728
19	0,1132	44	0,1326	44	0,0780	44	0,0695
16	0,0928	16	0,1312	16	0,0738	22	0,0694
32	0,0752	48	0,1230	38	0,0689	38	0,0591
37	0,0711	38	0,1106	37	0,0507	37	0,0422
44	0,0640	19	0,1010	46	0,0442	46	0,0370
38	0,0623	22	0,0990	15	0,0345	48	0,0312

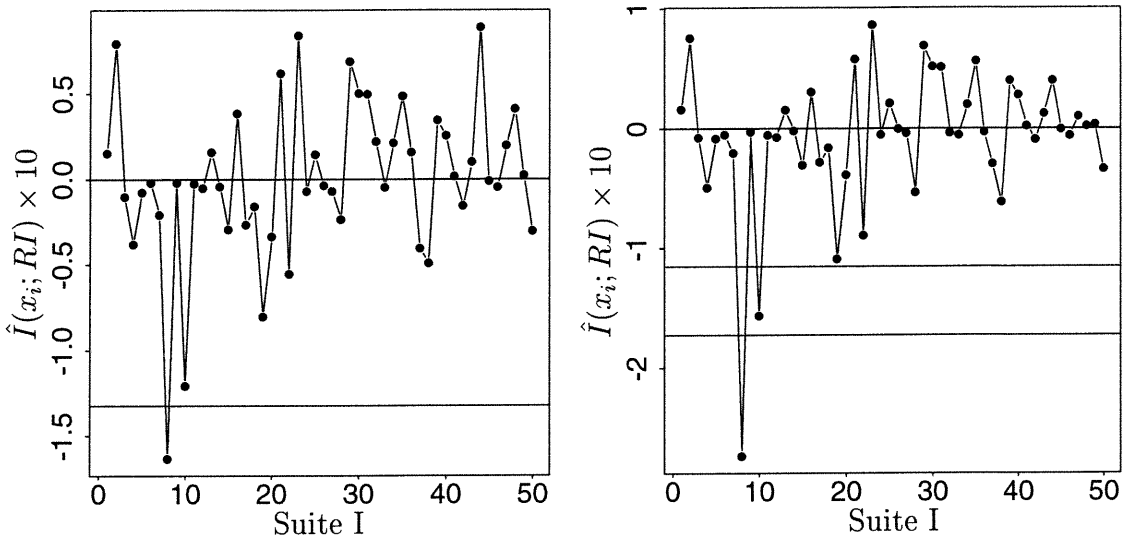

FIGURE 3.2.3. *Graphiques des estimés (habituels et robustes) des fonctions d'influence théorique de RI pour les données VENTES*

Pour les méthodes de Srivastava et von Rosen (1998) et de Naik (1989), les résultats sont présentés au tableau 3.2.7. Dans le cas utilisant l'estimateur S, les observations 8 et 10 ont une forme quadratique particulièrement grande et représentent donc de bons candidats à l'aberrance. Par contre, si on considère la version robuste de ces algorithmes, alors seule l'observation 10 se démarque des autres. Selon la borne théorique de 0,4571, on ne rejette pas l'hypothèse que l'observation 10 n'est pas aberrante.

Par le dendogramme de la figure 3.2.4, on teste l'influence de quatre groupes d'observations. Les résultats se trouvent au tableau 3.2.8. Comme auparavant, seul le deuxième ensemble contenant uniquement l'observation 8 a une influence robuste dépassant le seuil de trois écarts types.

Pour cet exemple, on peut dire que tous les algorithmes donnent des résultats très proches. De façon générale, la huitième et la dixième observations doivent

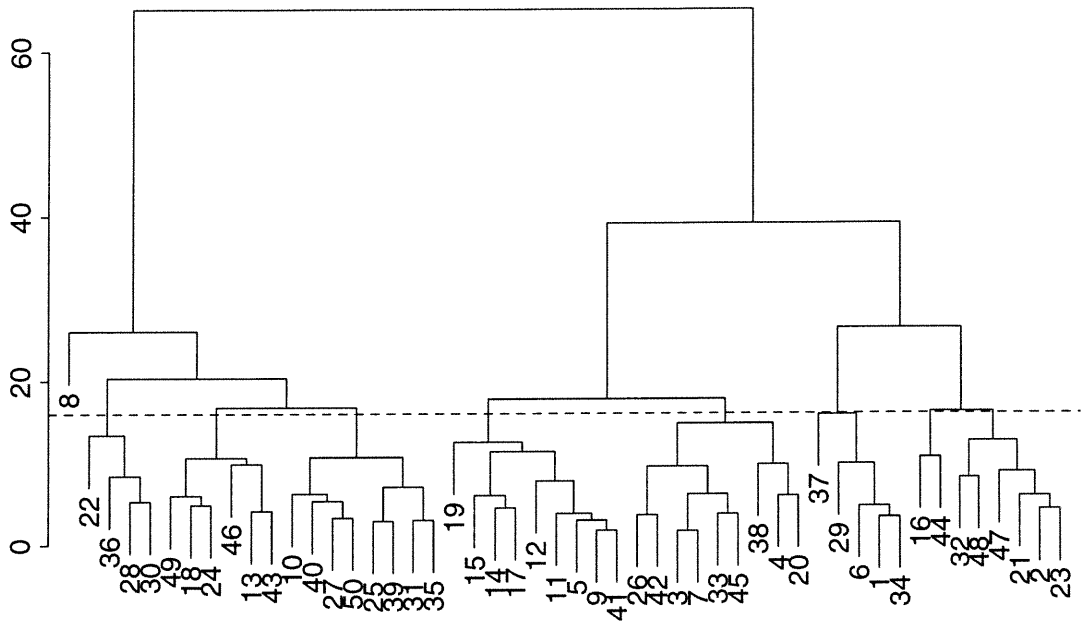

FIGURE 3.2.4. *Dendogramme pour les données VENTES*

TABLEAU 3.2.8. *Influence (robuste) d'un groupe de points pour les données VENTES*

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{1, 6, 29, 34, 37\}$	0,1340	0,1361
$G_2 = \{8\}$	-0,1634	**-0,2738**
$G_3 = \{22, 28, 30, 36\}$	0,1334	0,1035
$G_4 = \{16, 44\}$	0,1528	0,1026

être fortement considérées au niveau du critère de l'aberrance.

TABLEAU 3.2.9. *Généralisation de la méthode de Mickey et al. (1967) appliquée aux données de Rohwer*

Observation	*RI* partiel	*RI*	Valeur-p
37	0,1415	0,3823	0,0219
32	0,1183	0,4554	0,0399
5	0,1166	0,5189	0,0432
26	0,1047	0,5693	0,0615
27	0,1033	0,6138	0,0673
36	0,1112	0,6567	0,0591
9	0,1134	0,6957	0,0607
35	0,1574	0,7436	0,0230
30	0,1246	0,7755	0,0535
12	0,0919	0,7962	0,1297

3.2.3. Les données de Rohwer

On s'intéresse maintenant aux données de Rohwer présentées à la section 4.3 de Timm (1975). Il s'agit de données relatives à des élèves de maternelle provenant d'un milieu où le statut socio-économique est élevé. Des données similaires ont également été recueillies pour des sujets d'une région où le statut socio-économique est plus faible. Les variables à expliquer sont les résultats de trois tests différents : un test de réussite (SAT), un test de vocabulaire Peabody (PPVT) et un test de Ravin (RPMT). Quant aux cinq variables contrôlables, elles sont notées N, S, NS, NA et SS. L'échantillon comporte 37 observations.

Par le tableau 3.2.9, les observations 37, 32 et 5 sont les candidats à l'aberrance au niveau 5% selon la généralisation de la méthode de Mickey *et al.* (1967). Ces observations n'entrent toutefois pas dans le modèle avec une valeur-p très

TABLEAU 3.2.10. *Estimés (habituels et robustes) de la fonction d'influence théorique des données de Rohwer*

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)
13	-0,9546	-9,4526	3	0,9107	9,2494
3	0,9059	8,9700	13	-0,9102	-9,2438
17	-0,7726	-7,6498	37	-0,7959	-8,0832
37	-0,7481	-7,4075	17	-0,7957	-8,0813
31	-0,6923	-6,8552	30	0,6876	6,9832
32	-0,6758	-6,6919	31	-0,6372	-6,4719
30	0,6429	6,3658	32	-0,6100	-6,1953
29	0,5074	5,0242	29	0,5214	5,2957
9	-0,4685	-4,6391	10	-0,4870	-4,9456
33	-0,4609	-4,5639	9	-0,4869	-4,9454

significative.

Cette fois-ci on obtient un RI moins élevé que dans les exemples précédents, soit de 0,2805 et de 0,2735 dans le cas robuste. Pour l'écart type, on a un estimé de 0,6011 et un estimé robuste de 0,6004. À la figure 3.2.5, on remarque qu'aucune observation n'est à l'extérieur de l'intervalle $\pm 3\hat{\sigma}$. Mais si on regarde le tableau 3.2.10, les observations 3 et 13 ont une influence relative plus grande que les autres et sont donc susceptibles d'être aberrantes. Par contre le test approximatif ne confirme pas l'aberrance de ces deux observations avec une valeur-p respective de 0,8517 et de 0,8782.

Pour ce qui est des formes quadratiques du tablau 3.2.11, les observations 7 et 37 ont les plus grandes valeurs de T_i tandis que seule l'observation 7 se démarque par la méthode de Naik (1989). Pourtant, avec une borne théorique de 3,8670

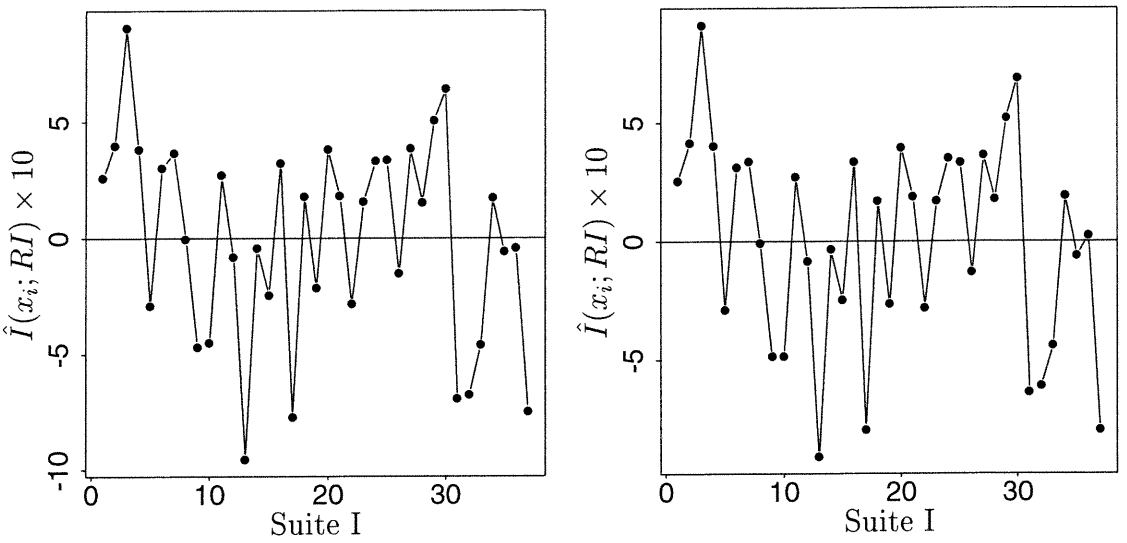

FIGURE 3.2.5. *Graphiques des estimés (habituels et robustes) des fonctions d'influence théorique de RI pour les données de Rohwer*

TABLEAU 3.2.11. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données de Rohwer*

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
7	0,1883	7	0,2575	7	0,2670	7	0,2143
37	0,1834	37	0,1704	37	0,1615	37	0,1501
26	0,1631	26	0,1578	5	0,1281	5	0,1168
5	0,1430	30	0,1465	26	0,1176	26	0,1113
32	0,1303	5	0,1313	8	0,1048	23	0,0961
16	0,1236	23	0,1292	23	0,1045	8	0,0927
8	0,1188	16	0,1178	28	0,1024	28	0,0885
9	0,1106	32	0,1157	32	0,0989	32	0,0871
30	0,1033	28	0,1150	30	0,0968	15	0,0844
36	0,0919	15	0,1134	34	0,0958	34	0,0823

l'observation 7 n'est pas considérée aberrante selon le critère basé sur la borne

supérieure de Bonferroni. Quant aux versions robustes de ces deux méthodes, encore une fois elles concordent en déclarant l'observation 7 comme premier candidat à l'aberrance.

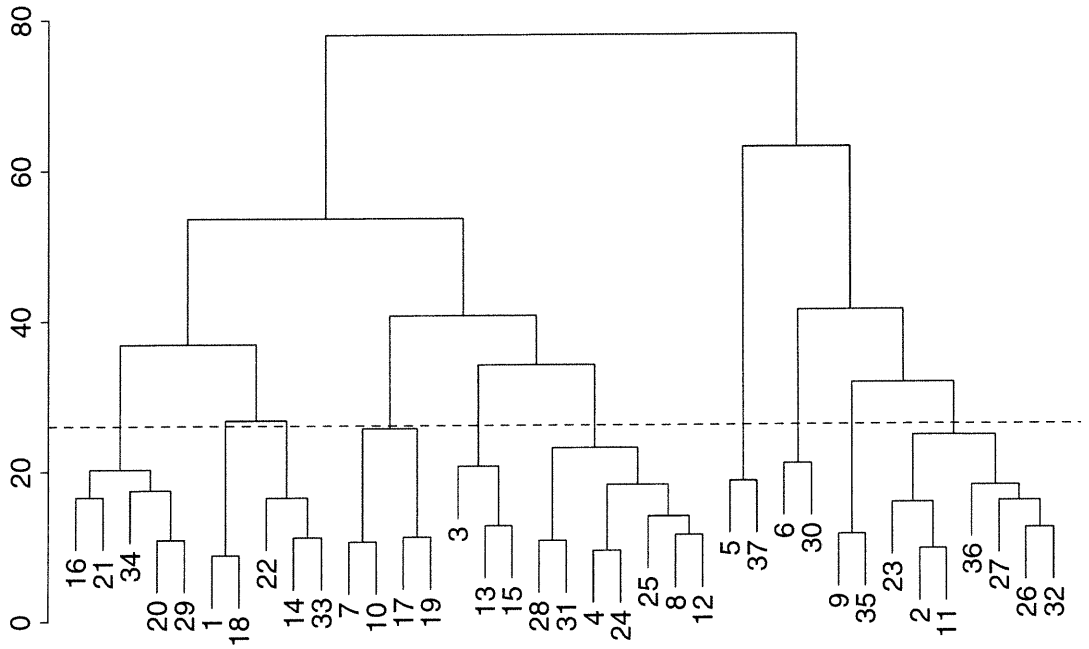

FIGURE 3.2.6. *Dendogramme pour les données Rohwer*

Sept différents groupes sont formés à partir du dendogramme de la figure 3.2.6. Avec les valeurs de 1,8033 et de 1,8012 pour trois écarts types (estimés habituel et robuste), aucune influence du tableau 3.2.12 n'est plus grande en valeur absolue. Or, selon cette procédure heuristique, aucun de ces ensembles n'est déclaré aberrant.

Pour les données de Rohwer, par tous les algorithmes, aucune observation n'est significativement déclarée aberrante. Selon les deux méthodes basées sur

TABLEAU 3.2.12. *Influence (robuste) d'un groupe de points pour les données de Rohwer*

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{5, 37\}$	-0,9993	-1,0439
$G_2 = \{6, 30\}$	0,9971	1,0269
$G_3 = \{7, 10, 17, 19\}$	-0,8655	-0,9819
$G_4 = \{3, 13, 15\}$	0,2215	0,2457
$G_5 = \{9, 35\}$	-0,5566	-0,5802
$G_6 = \{1, 18\}$	0,4607	0,4406
$G_7 = \{14, 22, 33\}$	-0,5452	-0,5268

des formes quadratiques et d'après la procédure FORWARD multivariée, on devrait s'interroger à propos des observations douteuses 7 et 37. Contrairement, selon les fonctions d'inluence de RI, ce sont les observations 3 et 13 qui ont le plus attiré l'attention. Puisque le RI calculé est faible, alors la qualité du modèle de régression utilisée laisse à désirer et c'est pourquoi les divers algorithmes ne concordent pas. En effet, il n'est pas pertinent de chercher des aberrances sous un mauvais modèle de régression.

3.2.4. Les données de Gerrild et Lantz

Gerrild et Lantz (1969) ont recueilli des données relatives au pétrole brut à partir de grès de la réserve pétrolière de Elk Hills en Californie. Basé sur sa composition chimique, chaque pétrole brut peut être attribué à l'une des trois zones de grès notées π_1, π_2 et π_3. Les cinq variables suivantes sont considérées:

X_1 = *vanadium (en pourcentage de cendres)*;

$$
\begin{aligned}
X_2 &= \sqrt{fer\ (en\ pourcentage\ de\ cendres)};\\
X_3 &= \sqrt{berylium\ (en\ pourcentage\ de\ cendres)};\\
X_4 &= 1/[hydrocarbones\ saturés\ (en\ pourcentage\ de\ la\ surface)];\\
X_5 &= hydrocarbones\ aromatiques\ (en\ pourcentage\ de\ la\ surface).
\end{aligned}
$$

Le vecteur à expliquer contient les trois premières variables qui représentent des éléments de traces. Quant aux variables contrôlables, soient les deux dernières, elles sont déterminées à partir de l'aire sous la courbe produite par l'analyse chimique d'une chromatographie en phase gazeuse. Dans le livre de Johnson et Wichern (1992), on observe les valeurs de ces cinq variables pour 56 cas dont leur appartenance à la population π_i, $i = 1,2,3$, est connue.

3.2.4.1. *Toutes les polulations* ($\pi_1 \cup \pi_2 \cup \pi_3$)

On considère d'abord l'ensemble des trois populations, soit le total des 56 observations.

Au tableau 3.2.13 on voit que trois observations (41, 20 et 10) entrent en premier dans le modèle avec une valeur-p d'environ 1%. Elles sont donc les plus suspectes à l'aberrance d'après la généralisation de la méthode de Mickey *et al.* (1967).

Par la régression linéaire multivariée des trois éléments de traces sur les deux mesures d'hydrocarbones on trouve un RI encore une fois peu élevé, soit de 0,3259 et de 0,3839 dans le cas robuste. Quant à la variance, son estimé est de 0,6796 et son estimé robuste est de 0,6829. Au tableau 3.2.14 et à la figure 3.2.7, on remarque que l'influence de l'observation 41 dépasse en valeur absolue $3\hat{\sigma} =$ 2,0388 et il en est de même pour l'influence robuste avec $3\hat{\sigma}$ robuste = 2,0487.

TABLEAU 3.2.13. *Généralisation de la méthode de Mickey et al. (1967) appliquée aux données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

Observation	*RI* partiel	*RI*	Valeur-p
41	0,0995	0,3930	0,0100
20	0,0924	0,4490	0,0142
10	0,1055	0,5071	0,0101
11	0,0755	0,5443	0,0357
47	0,0791	0,5804	0,0324
16	0,0819	0,6148	0,0299
27	0,0931	0,6506	0,0201
40	0,0924	0,6829	0,0209
7	0,0842	0,7096	0,0289
18	0,0795	0,7327	0,0354

TABLEAU 3.2.14. *Estimés (habituels et robustes) de la fonction d'influence théorique des données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)
41	**-2,8946**	-16,1505	41	**-3,4768**	-16,4733
11	-1,6431	-9,1678	11	-1,7901	-8,4815
30	1,4042	7,8349	40	-1,3915	-6,5930
40	-1,2740	-7,1085	30	1,2937	6,1297
25	1,0161	5,6694	16	-1,2618	-5,9784
16	-0,9474	-5,2862	47	-1,2329	-5,8415
47	-0,9219	-5,1436	25	0,9891	4,6866
46	0,8942	4,9893	46	0,8468	4,0121
55	-0,7365	-4,1093	55	-0,8358	-3,9601
18	-0,7101	-3,9621	20	-0,7282	-3,4505

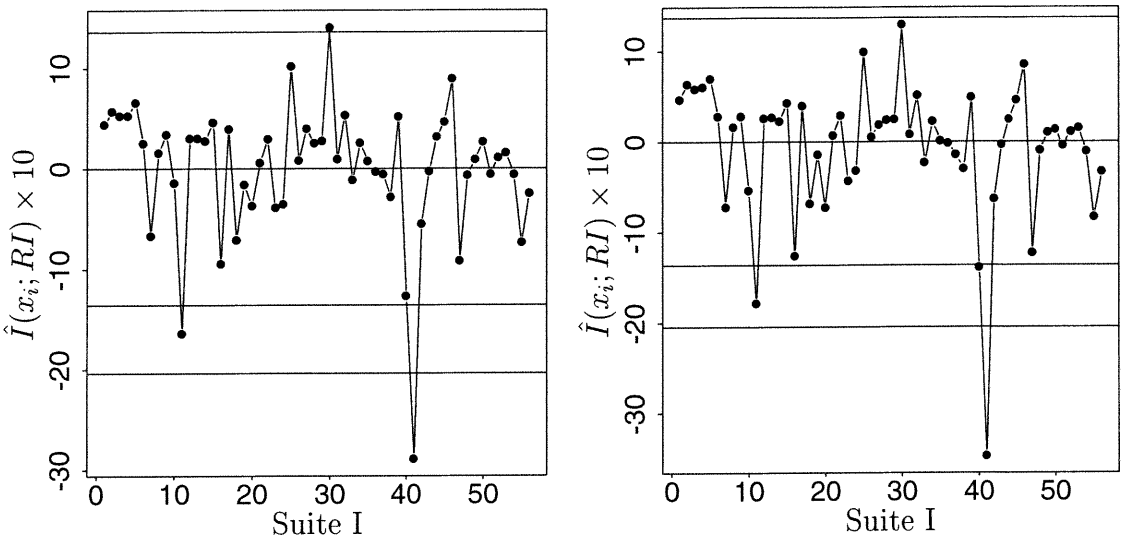

FIGURE 3.2.7. *Graphiques des estimés (habituels et robustes) des fonctions d'influence théorique de RI pour les données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

De plus, par le test approximatif basé sur l'algorithme d'Imhof, l'observation 41 est déclarée aberrante au niveau 5% avec une borne théorique de 0,0364. Ce n'est toutefois pas le cas pour l'observation 30 qui a la plus grande influence positive, car la borne théorique associée est de 0,8890.

Avec les deux méthodes basées sur les formes quadratiques, on obtient les mêmes résultats. Par le tableau 3.2.15 on voit que seule l'observation 20 peut être considérée comme aberrante tandis que dans le cas robuste l'observation 10 est également douteuse. Cependant, avec une borne supérieure de 6,9551 on ne rejette pas l'hypothèse que l'observation 20 n'est pas un candidat à l'aberrance.

On forme jusqu'à huit groupes pertinents à partir du dendogramme de la figure 3.2.8. Ces derniers ainsi que leur influence associée sont présentés au tableau 3.2.16. Ici, deux ensembles d'observations sont considérés aberrants puisque leur

TABLEAU 3.2.15. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
20	0,2727	20	0,3164	20	0,4930	20	0,4865
41	0,1407	10	0,2277	10	0,3186	10	0,3126
10	0,1371	41	0,2162	38	0,1493	38	0,1452
40	0,1133	11	0,1529	37	0,1427	37	0,1406
11	0,1083	40	0,1245	40	0,1322	40	0,1307
7	0,0917	38	0,1108	41	0,1274	41	0,1164
18	0,0823	55	0,1028	50	0,1142	50	0,1131
21	0,0747	37	0,0895	11	0,0872	11	0,0863
52	0,0675	7	0,0892	18	0,0854	18	0,0814
55	0,0666	21	0,0817	7	0,0822	7	0,0798

influence en valeur absolue est plus grande que trois écarts types tant dans le cas habituel que dans le cas robuste. Or, les observations 16, 33 et 47 du deuxième groupe ainsi que les observations 27, 30, 44 et 46 du sixième groupe représentent des aberrances selon cette procédure heuristique.

Enfin, pour les données de Gerrild et Lantz (1969) d'après les méthodes de détection d'aberrance une à la fois, il faut surveiller les observation 41, 20 et 10. Par contre, l'algorithme de détection d'ensemble aberrant donne des résultats tout à fait différents en considérant douteux les deux groupes contenant les observations 16, 33, 47 et 27, 30, 44, 46. Comme dans l'exemple précédent, le modèle de régression est de faible qualité et les divers algorithmes donnent des résultats

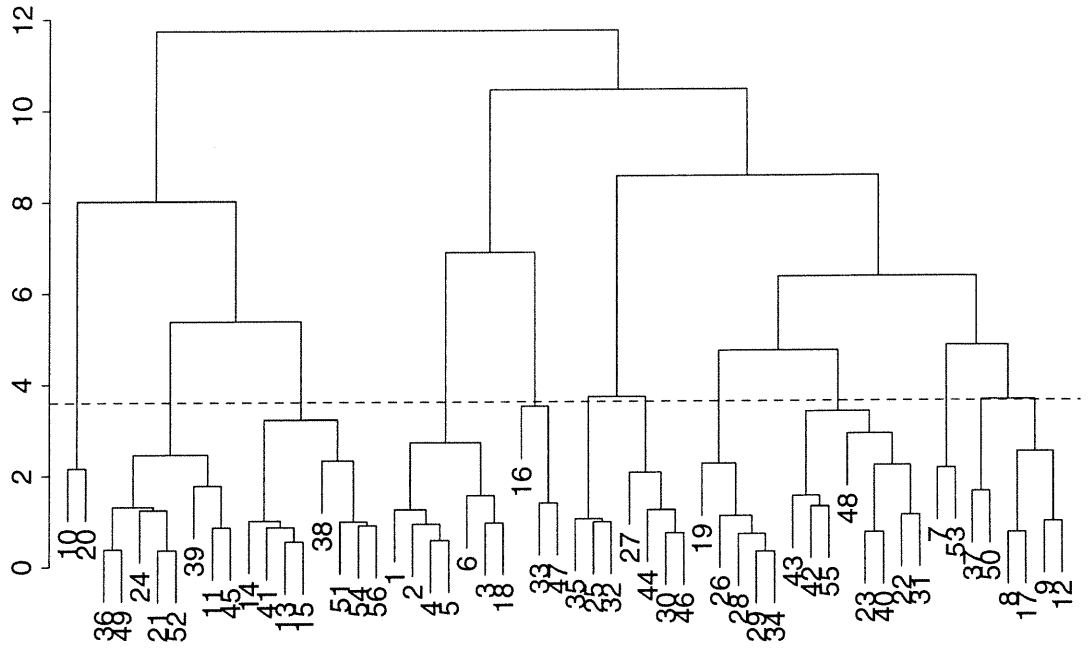

FIGURE 3.2.8. *Dendogramme pour les données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

TABLEAU 3.2.16. *Influence (robuste) d'un groupe de points pour les données de Gerrild et Lantz (1969)* ($\pi_1 \cup \pi_2 \cup \pi_3$)

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{10, 20\}$	-0,4459	-1,1904
$G_2 = \{16, 33, 47\}$	**-2,0420**	**-2,7569**
$G_3 = \{7, 53\}$	-0,2268	-0,2551
$G_4 = \{19, 26, 28, 29, 34\}$	0,9808	0,934
$G_5 = \{25, 32, 35\}$	1,8178	1,7301
$G_6 = \{27, 30, 44, 46\}$	**3,9159**	**3,6440**
$G_7 = \{37, 50\}$	0,2324	0,0081
$G_8 = \{8, 9, 12, 17\}$	1,7374	1,7351

discordants.

TABLEAU 3.2.17. *Généralisation de la méthode de Mickey et al. (1967) appliquée aux données de Gerrild et Lantz (1969)* (π_3)

Observation	*RI* partiel	*RI*	Valeur-p
2	0,1996	0,4535	0,0024
23	0,1551	0,5382	0,0099
37	0,1343	0,6003	0,0194
35	0,1206	0,6485	0,0321
19	0,1500	0,7012	0,0178
32	0,1538	0,7472	0,0169
1	0,1373	0,7819	0,0272
9	0,1331	0,8109	0,0312
25	0,1287	0,8352	0,0355
20	0,1330	0,8571	0,0332

3.2.4.2. *Population π_3 seulement*

Finalement, on examinera uniquement les 38 observations provenant de la troisième population.

En ordre d'importance, les observations à considérer comme étant les plus douteuses sont 2, 23, et 37 si on se fie aux valeurs-p du tableau 3.2.17 pour la généralisation de la méthode de Mickey *et al.* (1967).

Pour la troisième population de grès, la régression des trois éléments traces sur les deux mesures d'hydrocarbone nous donne un *RI* encore faible, soit de 0,3172, ainsi qu'un écart type de 0,6501. Dans le cas robuste, ces valeurs augmentent à 0,3598 et à 0,6519. Seule l'influence robuste de l'observation 23 se trouve à l'extérieur de l'intervalle de $\pm 3\hat{\sigma} = \pm 1{,}9557$ tel qu'en font foi le tableau 3.2.18 et la

TABLEAU 3.2.18. *Estimés (habituels et robustes) de la fonction d'influence théorique des données de Gerrild et Lantz (1969) (π_3)*

Observation	$\hat{I}(x_i; RI)$	Influence relative (%)	Observation	$\hat{I}(x_i; RI)$ robuste	Influence relative (%)
32	1,2924	11,0123	23	**-1,9823**	-14,8891
23	-1,1969	-10,1991	37	-1,3569	-10,1919
37	-1,1032	-9,4003	32	1,2845	9,6478
12	1,0942	9,3240	22	-1,2289	-9,2305
22	-1,0339	-8,8097	12	1,1946	8,9729
6	-0,9521	-8,1127	6	-0,9957	-7,4784
1	-0,6833	-5,8221	1	-0,7176	-5,3897
28	0,6314	5,3800	28	0,6701	5,0331
7	0,5826	4,9644	7	0,6699	5,0320
9	0,5611	4,7811	36	0,4632	3,4792

figure 3.2.9. Cependant, l'observation 23 ayant l'influence minimale et l'observation 32 ayant l'influence maximale ne sont pas déclarée aberrantes d'après le test approximatif avec une valeur-p respective de 0,7330 et de 0,7276.

Selon le tableau 3.2.19, dans tous les cas c'est l'observation 2 qui a une forme quadratique beaucoup plus grande que celle des autres observations et qui par conséquent est le premier candidat à l'aberrance. Par le test de Srivastava et von Rosen (1998), on associe à l'observation 2 une borne théorique de 0,0660 et on est près de la déclarer aberrante au niveau 5%. De plus, l'observation 23 pourrait aussi être considérée douteuse, mais si on ajoute le critère de robustesse alors c'est plutôt l'observation 20 qui devient le deuxième candidat à l'aberrance.

À l'aide de la figure 3.2.10, on forme les cinq groupes d'observations du tableau 3.2.20. On compare la valeur absolue de ces influences à trois écarts-types,

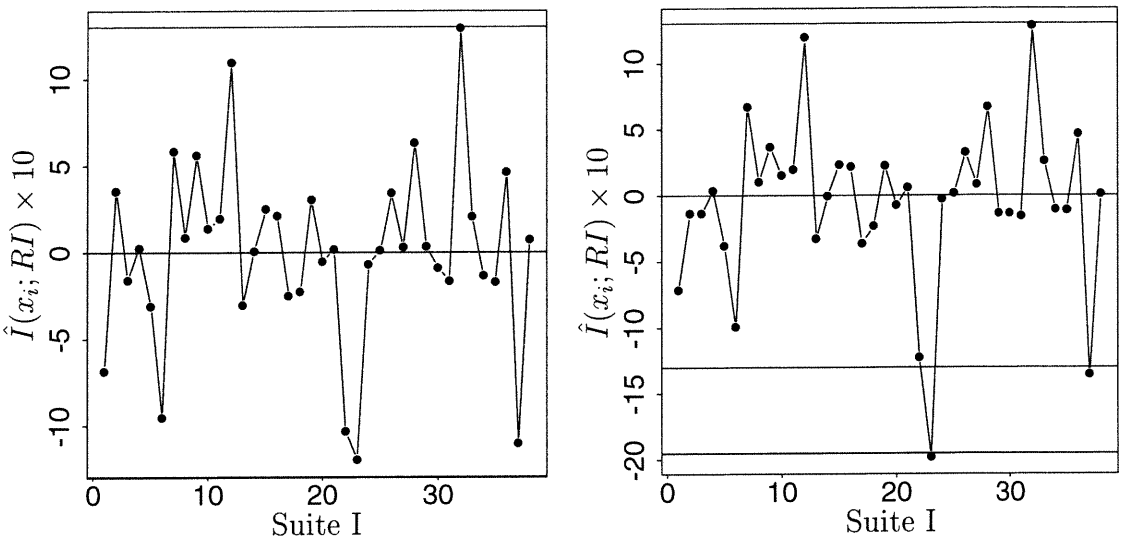

FIGURE 3.2.9. *Graphiques des estimés (habituels et robustes) des fonctions d'influence théorique de RI pour les données de Gerrild et Lantz (1969)* (π_3)

soit à 1,9503 et à 1,9557 dans le cas robuste. Ainsi le dernier groupe incluant les observations 9, 12, 26 et 28 est aberrant et le premier groupe contenant les observations 2, 19 et 32 est plutôt douteux.

D'après le dernier algorithme, les observations 9, 12, 26 et 28 semblent être aberrantes ensemble bien qu'elles ne l'étaient pas individuellement. L'observation 2 est un candidat important à l'aberrance. Les méthodes d'identification d'une aberrance à la fois déclarent également douteuses les observations 20, 23 et 37. Encore une fois, puisque le *RI* calculé est petit, le modèle de régression ainsi que la pertinence des différents résultats obtenus laissent à désirer.

TABLEAU 3.2.19. *Les méthodes (robustes) de Srivastava et von Rosen (1998) et de Naik (1989) appliquées aux données de Gerrild et Lantz (1969) (π_3)*

Observation	T_i	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$	Observation	T_i robuste	Observation	$\underset{\sim}{\epsilon_i}' A \underset{\sim}{\epsilon_i}$ robuste
2	0,3630	2	0,4018	2	0,5385	2	0,5298
23	0,1984	23	0,2729	20	0,2263	20	0,2183
37	0,1368	37	0,2173	23	0,1761	23	0,1514
6	0,1321	20	0,1901	22	0,1349	22	0,1313
3	0,1280	3	0,1627	19	0,1267	19	0,1229
35	0,1242	35	0,1489	37	0,1266	37	0,1228
20	0,1117	34	0,1404	3	0,1213	3	0,1148
22	0,1100	22	0,1352	34	0,1056	34	0,0984
33	0,1057	6	0,1299	6	0,1017	6	0,0975
18	0,1053	18	0,1048	24	0,0940	24	0,0886

TABLEAU 3.2.20. *Influence (robuste) d'un groupe de points pour les données de Gerrild et Lantz (1969) (π_3)*

Groupe d'observations	Influence $\hat{I}(\overline{X}_G; RI)$	Influence robuste
$G_1 = \{2, 19, 32\}$	**2,3694**	1,9172
$G_2 = \{24, 25, 35, 37\}$	0,0373	0,2118
$G_3 = \{15, 29\}$	0,3857	0,2228
$G_4 = \{7, 14, 17\}$	0,5323	0,5239
$G_5 = \{9, 12, 26, 28\}$	**3,245**	**3,3704**

3.3. COMPARAISON GÉNÉRALE DES DIVERS ALGORITHMES

Pour terminer, évaluons la performance globale de chacune des méthodes pour l'ensemble des jeux de données étudiés précédemment. De façon générale, les différents algorithmes pour la détection de données aberrantes une à la fois donnent

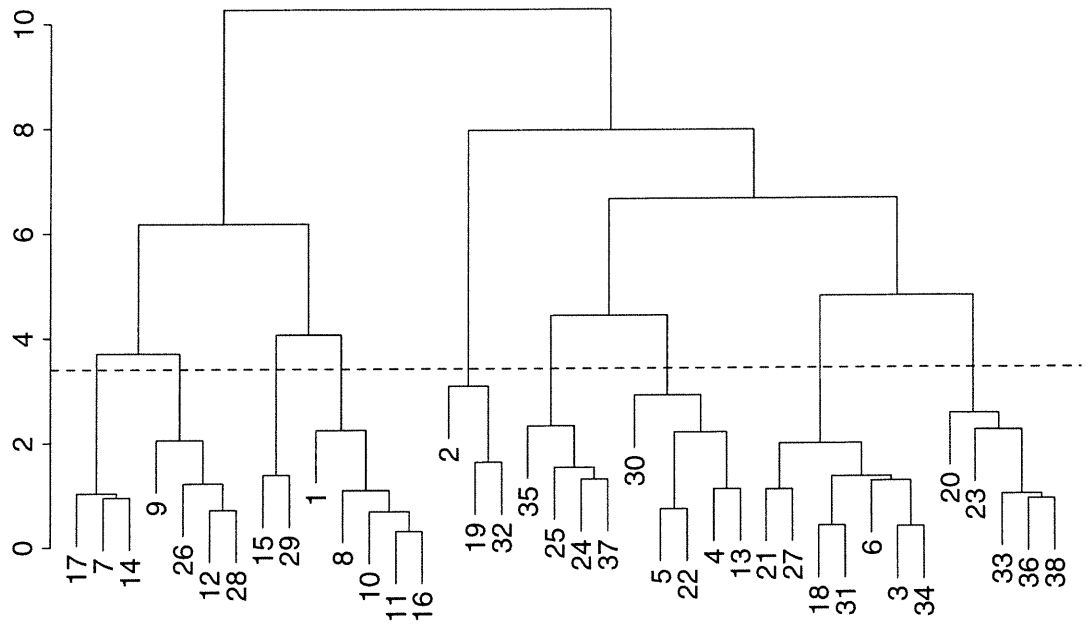

FIGURE 3.2.10. *Dendogramme pour les données de Gerrild et Lantz (1969)* (π_3)

des résultats très similaires lorsque le modèle de régression est de bonne qualité, c'est-à-dire lorsque RI est assez grand. La généralisation de la méthode de Mickey *et al.* (1967) à cependant le désavantage d'être plus complexe et plus longue à calculer que les autres. Comme il fut question à la fin de la première section de ce chapitre, la fonction d'influence théorique de RI est préférée à la fonction d'influence empirique puisqu'on peut calculer sa variance ainsi qu'un test approximatif.

Quant à la robustesse, pour la fonction d'influence de RI, elle a permis dans quelques cas la détection de données aberrantes qui n'avaient pas été décelées selon la procédure heuristique non robuste. Ceci est vrai autant pour l'influence

d'une donnée ou d'un groupe de points. Pour ce qui est de la méthode de Srivastava et von Rosen (1998) et de celle de Naik (1989), on préférera leur version robuste qui permettent d'obtenir les mêmes résultats par les deux algorithmes.

Aussi, on a remarqué que tous les tests basés sur l'inférence ne sont que rarement significatif. Or, ceci nous importe peu puisque le but est de pointer les observations douteuses et non de juger de leur aberrance. Ce jugement pourra être posé qu'après une évaluation de ces données dans leur contexte.

Dans les derniers exemples, la méthode de détection d'ensembles aberrants a donné des résultats différents à ceux des autres méthodes. Ceci est dû à la mauvaise qualité du modèle. Or, la méthode de Mickey *et al.* (1967) et les méthodes basées sur la fonction d'influence de RI, que ce soit pour la détection d'une aberrance à la fois ou d'un groupe d'observations, ont l'avantage de mesurer la qualité du modèle en calculant RI.

En pratique, si le modèle de régression est pertinent, on recommande d'utiliser à la fois les versions robustes et non robustes de la fonction d'influence de RI et les versions robustes de Srivastava et von Rosen (1998) ou de Naik (1989). Ici, l'inférence importe peu. Par la suite, on devra retourner à l'expérimentateur afin de préciser si les données détectées comme douteuses sont vraiment des aberrances. Si tel est le cas, ce dernier nous aidera également à déterminer le traitement spécial qui doit leur être accordé : les corriger si possible, les enlever et les imputer à partir des autres données, etc.

CONCLUSION

Le but de ce mémoire est de généraliser certains algorithmes de détection de données aberrantes sous le modèle de la régression linéaire multivariée, tel que défini au premier chapitre, et de les comparer à d'autres méthodes existantes dans la littérature.

La première procédure d'intérêt est celle de Mickey *et al.* (1967) qui se base sur l'inclusion successive de variables en régression linéaire multiple. On a vu au premier chapitre que l'indice de redondance de Stewart et Love (1968) se réduit à R^2 lorsque $p = 1$ et que la connaissance de sa distribution exacte permet de définir des algorithmes de sélection de variables en régression linéaire multivariée. D'où on a pu généraliser cette méthode pour $p > 1$. Ainsi cette dernière permet d'évaluer la qualité du modèle de régression par RI et de détecter une donnée aberrante à la fois. Cependant, comme la procédure d'inclusion successive (FORWARD) multivariée est assez complexe, le temps d'exécution de cet algorithme est plus considérable que celui des autres méhodes.

En second lieu, on a reproduit la procédure basée sur la fonction d'influence de RI ainsi que le test approximatif correspondant définis par Lazraq et Cléroux (1989). Cette méthode a l'avantage d'identifier des observations influentes selon la pertinence du modèle de régression linéaire multivariée. Ensuite, dans le but de contrer l'effet de masquage, on a généralisé cet algorithme pour la détection d'ensembles aberrants de façon similaire à Cléroux *et al.* (1990). Pour ce faire, afin de définir des groupes d'observations pertinents on a eu recours à la classification hiérarchique avec la distance euclidienne et sous le critère de la liaison complète.

De plus, dans le cas d'un échantillon, pour calculer les fonctions d'influence on a d'abord utilisé l'estimateur S. Malheureusement, ce dernier est très influençable par la présence de valeurs à l'écart. On a donc contruit de nouvelles versions pour ces deux méthodes en utilisant l'estimateur robuste de Huber (1977).

D'autre part, on a voulu comparer ces différents algorithmes à d'autres provenant de la littérature. On a donc repris les méthodes de Naik (1989) et de Srivastava et von Rosen (1998). Notons ici que la différence majeure est que ces deux procédures détectent des données aberrantes au sens d'une translation dans la moyenne. Encore une fois, on a créé les versions robustes afin d'améliorer ces deux méthodes. Ceci nous a permis d'obtenir exactement les mêmes résultats pour les versions robustes de ces deux méthodes.

Puis, on a appliqué chacune des méthodes à différents jeux de données en régression linéaire multivariée. À partir des résultats ainsi obtenus au troisième chapitre, on en déduit que tous les algorithmes sont pertinents que sous un modèle de régression de bonne qualité, c'est-à-dire dont la valeur de RI est grande. Les procédures basées sur l'indice de redondance RI ont l'avantage de permettre d'évaluer cette qualité du modèle.

En somme, afin de maximiser nos chances de détecter tous les candidats à l'aberrance on utilisera d'abord les méthodes habituelles et robustes basées sur la fonction d'influence de RI pour la détection d'une donnée aberrante à la fois et pour l'identification d'ensembles aberrants. À ces dernières, on jumelera l'emploi d'un des deux algorithmes robustes de détection de valeurs influentes au sens de la translation dans la moyenne.

Une fois les données douteuses ainsi décelées, l'inférence importe peu puisque l'on devra retourner auprès de l'expérimentateur afin de juger de leur aberrance

et s'il y a lieu de décider du traitement spécial qui doit leur être accordé.

Annexe A

FONCTIONS SPLUS POUR LA SÉLECTION DE VARIABLES FORWARD EN RÉGRESSION MULTIVARIÉE

```
#GENERALISATION DE LA METHODE DE MICKEY ET AL. AU CAS MULTIVARIE
nforward_function(x, n, p, q, f = T)
{
#fonction FORWARD
#x:la matrice des données de dimension n*(p+q)
#p:la dimension du vecteur dépendant (y)
#q:la dimension du vecteur indépendant (x)
#n:le nombre d'observations
#f=true pour la généralisation de la méthode de Mickey, Dunn et Clark
                a   <- 1
                b   <- q
                d   <- p
                tab <- c(rep(1, q))
                ino <- 0
                h   <- 0
        if(f) {
                x <- cbind(x, diag(n))
                a <- q + 1
                b <- q + n
                d <- p + q
                tab <- c(rep(1, b))
```

```
        tab[1:q] <- p + (1:q)
        ino <- q
        h   <- nric(tab, ino, x, p, q)
}
s   <- var(x)
s11 <- s[1:p, 1:p]
tr  <- trace.mat(s11)
cat("var", "ino", "RIP", "RI", "pvalue", "f", sep = "   ", "\n")
while(ino < b) {
        rm  <- 0
        ino <- ino + 1
        for(j in (a:b)) {
                k  <- j + p
                lv <- match(k, tab)
                if(is.na(lv)) {
                        tab[ino] <- k
                        ri <- nric(tab, ino, x, p, q)
                        if(rm < ri) {
                          rm <- ri
                          kk <- p + j
                        }
                }
        }
        rip <- (rm - h)/(1 - h)
        h   <- rm
        tab[ino] <- kk
        if(rip < 1) {
                if(ino == 1){
                        s113   <- s11
                        fisher <- ntest2(rip, n, ino)
                        prob   <- test(rip, s113, n, p, ino)}
```

```
                        else if (p == 1) {
                                fisher <- ntest2(rip, n, ino)
                                prob   <- ntest(fisher, n, ino)}
                        else {  s113   <- ssp(s11, s, tab, p, ino)
                                fisher <- NULL
                                prob   <- test(rip, s113, n, p, ino)}
                }
                else prob <- 1
                im <- tab[ino] - d
                im <- round(im, 0)
                rm <- round(rm, 4)
                prob   <- round(prob, 4)
                rip    <- round(rip, 4)
                fisher <- round(fisher, 4)
                cat(im, ino, rip, rm, prob, fisher, sep = "    ", "\n")
        }
}
##############################################################################

#CALCUL DE LA TRACE D'UNE MATRICE
trace.mat_function(mat)
{
        sum(diag(as.matrix(mat)))
}
##############################################################################

#CALCUL DU RI PARTIEL
nric_function(tab, ino, x, p, q)
{
        num <- 0
        den <- 0
```

```
        for(i in (1:p)) {
                t1  <- tab[1:ino]
                Y   <- x[, i]
                reg <- lsfit(x[, t1], Y)
                e   <- resid(reg)
                sce <- t(e) %*% e
                y   <- Y - mean(Y)
                sct <- t(y) %*% y
                r2  <- 1 - (sce/sct)
                s2  <- var(Y)
                num <- num + (r2 * s2)
                den <- den + s2
        }
        ri <- num/den
        ri
}
##############################################################################
#CALCUL DE LA MATRICE SIGMA ETOILE

ssp_function(s11, s, tab, p, ino)
{
        nim  <- ino - 1
        dimp <- 1:p
        t1   <- tab[1:nim]
        s33  <- s[t1, t1]
        s13  <- s[dimp, t1]
        s113 <- s13 %*% solve(s33) %*% t(s13)
        s11 - s113
}
##############################################################################
```

```
#CALCUL DE LA VALEUR-P POUR LE TEST MULTIVARIE
test_function(ri, s113, n, p, ino)
{
           r   <- ri/(1 - ri)
        lambda <- eigen(s113)$values
        coef   <- c(lambda,  - r * lambda)
        mult   <- c(rep(1, p), rep(n - 1 - ino, p))
        1 - fquad(coef, 0, mult)
}
##############################################################################

#CALCUL DE LA VALEUR-P LORSQUE P=1
ntest_function(f, n, ino)
{
        1 - pf(f, 1, n - 1 - ino)
}
##############################################################################

#CALCUL DE LA FISHER LORSQUE P=1
ntest2_function(ri, n, ino)
{
        r   <- ri/(1 - ri)
        df2 <- n - 1 - ino
        nb  <- r * df2
        nb
}
```

Annexe B

FONCTIONS SPLUS DES AUTRES ALGORITHMES DE DÉTECTIONS D'ABERRANCES EN RÉGRESSION MULTIVARIÉE

```
#FONCTION D'INFLUENCE THEORIQUE DE RI ET INFERENCE
inf.ri_function(x, n, p, q, plot=F, robustesse=F)
{
p1 <- p + 1
pq <- p + q

#CALCUL DE L'ESTIMATEUR (ROBUSTE) DE LA MATRICE DE VARIANCE-COVARIANCE
#ET CALCUL DE RI (ROBUSTE):
if (robustesse){ s <- robuste(x)$covariance
                ri <- rv("rvi", x, p, robust=T)}
else {s  <- var(x)
      ri <- rv("rvi", x, p)}
s11 <- s[1:p, 1:p]
s12 <- s[1:p, p1:pq]
s21 <- t(s12)
s22 <- s[p1:pq, p1:pq]
if (p==1) {
                print(c("R^2 = ", ri))
                s11_as.matrix(s11)
```

```
                s12_t(as.matrix(s12))
                s21_t(as.matrix(s21))
}
else{ print(c("RI = ", ri))}

#CALCUL DE L'ECART TYPE ESTIME POUR LE TEST HEURISTIQUE:
        tr <- trace.mat(s11)
s11.etoile <- s12 %*% solve(s22) %*% s21
 tr.etoile <- trace.mat(s11.etoile)
B  <- s12%*%solve(s22)
d2 <- 2*(ri^2)*(trace.mat(s11%*%s11)/(tr^2) - (4*trace.mat(s11%*%s11.etoile)
        - 2*trace.mat(s11.etoile%*%s11.etoile))/(tr.etoile*tr)
        + (2*trace.mat(s11%*%s11.etoile)
        - trace.mat(s11.etoile%*%s11.etoile))/(tr.etoile^2))
d1 <- sqrt(d2)
print(c("sigma = ", d1))

#CALCUL DES ESTIMES DE LA FONCTION D'INFLUENCE THEORIQUE DE RI:
x1 <- x[,1:p]
x2 <- x[,p1:pq]
xbar  <- apply(x, 2, mean)
x1bar <- xbar[1:p]
x2bar <- xbar[p1:pq]
z1 <- t(x1) - x1bar
z2 <- t(x2) - x2bar
 I <- ri*diag((2*t(z1)%*%B%*%z2/tr.etoile) - (t(z1)%*%z1/tr)
                - (t(z2)%*%t(B)%*%B%*%z2/tr.etoile))
I.rel <- 100*I/((n-1)*ri)
I     <- round(I, 4)
I.rel <- round(I.rel, 4)
mat <- cbind(1:n, I, I.rel)
```

```
mat1 <- mat[rev(order(abs(mat[,2]))),]
dimnames(mat1)[[2]] <- c("obs.i", "influence", "inf.relative (%)")

#TEST APPROXIMATIF POUR AU PLUS 2 ABERRANCES (IMHOF):
Ik  <- max(I)
Il  <- min(I)
ip  <- diag(1, p)
A11 <- -ip/trace.mat(s11)
A12 <- B/tr.etoile
A21 <- t(A12)
A22 <- -t(B)%*%B/tr.etoile
A1  <- cbind(A11, A12)
A2  <- cbind(A21, A22)
A   <- rbind(A1, A2)
A   <- ri*A
sa  <- s%*%A
l1  <- eigen(sa)$values
Fk  <- fquad(l1, Ik)
p3  <- 1 - (Fk)^n
Fl  <- fquad(l1, Il)
p4  <- 1 - ((1 - Fl)^n)
print(c("test approximatif : Pmax = P[Tn > I(Xk; RI)] =", p3))
print(c("test approximatif : Pmin = P[T1 < I(Xl; RI)] =", p4))

#GRAPHIQUE DE LA FONCTION D'INFLUENCE THEORIQUE ESTIMEE DE RI:
if (plot){
        plot(1:n, 10*I, xlab="seq", ylab="ixri", type="b")
        abline(h=c(-30*d1, -20*d1, 0, 20*d1, 30*d1))
        }
return(mat1)
}
```

```
##############################################################################

#FONCTION D'INFLUENCE EMPIRIQUE DE RI
inf.jacknife_function(x, n, p, q, plot=F, robustesse=F)
{
#CALCUL DE L'ESTIMATEUR (ROBUSTE) DE LA MATRICE DE VARIANCE-COVARIANCE
#ET CALCUL DE RI (ROBUSTE):
if (robustesse){ s <- robuste(x)$covariance
                 ri <- rv("rvi", x, p, robust=T)}
else {s  <- var(x)
      ri <- rv("rvi", x, p)}
s11 <- s[1:p, 1:p]
s12 <- s[1:p, (p+1):(p+q)]
s22 <- s[(p+1):(p+q), (p+1):(p+q)]
if (p==1) {
                 print(c("R^2 = ", ri))
                 s11_as.matrix(s11)
                 s12_t(as.matrix(s12))
                 s21_t(as.matrix(s21))
}
else {print(c("RI=", ri))}

#CALCUL DES ESTIMES DE LA FONCTION D'INFLUENCE EMPIRIQUE DE RI:
        tr <- trace.mat(s11)
s11.etoile <- s12%*%solve(s22)%*%t(s12)
 tr.etoile <- trace.mat(s11.etoile)
ri.i <- rep(NA, n)
for (i in (1:n)) {
        x.i <- x[-i,]
        ri.i[i] <- rv("rvi", x.i, p)
        }
```

```
I     <- (n - 1)*(ri - ri.i)
I.rel <- 100*I/((n-1)*ri)
I     <- round(I, 4)
I.rel <- round(I.rel, 4)
mat <- cbind(1:n, I, I.rel)
mat1 <- mat[order(abs(mat[,2])),]
mat2 <- apply(mat1, 2, rev)
dimnames(mat2)[[2]] <- c("obs.i", "influence", "inf.relative (%)")

#GRAPHIQUE DE LA FONCTION D'INFLUENCE EMPIRIQUE ESTIMEE DE RI:
if (plot){
        plot(1:n, 10*I, xlab="seqj", ylab="ixrij", type="b")
        abline(h=0)
        }
return(mat2)
}
################################################################################

#METHODE DE NAIK
formquad_function(x, n, p, q, robustesse=F)
{
Y <- x[, 1:p]
X <- x[, (p+1):(p+q)]
E <- lsfit(X, Y)$residuals
if (robustesse){ S <- robuste(Y)$covariance*(n-1) }
else { S <- t(E)%*%E}
ESE  <- diag(E%*%solve(S)%*%t(E))
ESE  <- round(ESE, 4)
Obs  <- 1:n
mat1 <- cbind(Obs, ESE)
mat2 <- mat1[rev(order(mat1[,2])),]
```

```
return(mat2)
}
#############################################################################

#METHODE DE SRIVASTAVA ET VON ROSEN ET INFERENCE
tmax_function(x, n, p, q, robustesse=F)
{
Y  <- x[, 1:p]
X  <- as.matrix(x[, (p+1):(p+q)])
XX <- solve(t(X)%*%X)
R  <- X%*%XX%*%t(X)
Id <- diag(rep(1, n))
if (robustesse){ S <- robuste(Y)$covariance*(n-1) }
else { S <- t(Y)%*%(Id - R)%*%Y}
E  <- lsfit(X, Y)$residuals
Ri <- diag(R)
Ti <- diag(E%*%solve(S)%*%t(E))*(1/(1 - Ri))

#CALCUL DE LA BORNE THEORIQUE (BONFERRONI)
if (!robustesse){
        f    <- n - q - 1
        cte  <- (f - p + 1)/p
        Tmax <- max(Ti)
        fisher   <- cte*Tmax/(1 - Tmax)
        valeur.p <- (1 - pf(fisher, p, f - p + 1))
        n.valp   <- n*valeur.p
        print(c("Borne theorique (Bonferroni) pour T =", n.valp))
        }
Obs  <- 1:n
Ti   <- round(Ti, 4)
mat1 <- cbind(Obs, Ti)
```

```
mat2 <- mat1[rev(order(mat1[,2])),]
return(mat2)
}
##############################################################################

#FONCTION D'INFLUENCE THEORIQUE DE RI POUR UN GROUPE DE POINTS
inf.gr_function(x, n, p, q, groupe, plot=F, robustesse=F)
{
p1 <- p + 1
pq <- p + q
m  <- length(groupe)
xg <- x[groupe,]

# DENDOGRAMME:
if (plot) {
        x_as.matrix(x)
        d <- dist(x)
        h <- hclust(d)
        plclust(h)
        }

#CALCUL DE L'ESTIMATEUR (ROBUSTE) DE LA MATRICE DE VARIANCE-COVARIANCE
#ET CALCUL DE RI (ROBUSTE):
if (robustesse){ s <- (robuste(x)$covariance)
                ri <- rv("rvi", x, p, robust=T)}
else {s  <- (var(x))
      ri <- rv("rvi", x, p)}
s11 <- s[1:p, 1:p]
s12 <- s[1:p, p1:pq]
s21 <- t(s12)
s22 <- s[p1:pq, p1:pq]
```

```
if (p==1) {
        s11 <- as.matrix(s11)
        s12 <- t(as.matrix(s12))
        s21 <- t(as.matrix(s21))
        }

#CALCUL DES ESTIMES DE LA FONCTION D'INFLUENCE THEORIQUE DE RI:
if (m == 1) {xg.bar <- xg}
else {xg.bar <- apply(xg, 2, mean)}
xg1.bar <- xg.bar[1:p]
xg2.bar <- xg.bar[p1:pq]
mu  <- apply(x, 2, mean)
mu1 <- mu[1:p]
mu2 <- mu[p1:pq]
u1 <- xg1.bar - mu1
u2 <- xg2.bar - mu2
if (p==1 && m==1) {u2 <- t(u2)}
tr <- trace.mat(s11)
s11.etoile <- s12 %*% solve(s22) %*% s21
tr.etoile  <- trace.mat(s11.etoile)
B <- s12%*%solve(s22)
I <- m*ri*((2*t(u1)%*%B%*%u2/tr.etoile) - (t(u1)%*%u1/tr)
                - (t(u2)%*%t(B)%*%B%*%u2/tr.etoile))
I <- round(I, 4)
return(I)
}
```

BIBLIOGRAPHIE

Andrews, D. et D. Pregibon (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society, Series B* **40**(1), 85–93.

Barnett, V. et T. Lewis (1978). *Outliers in Statistical Data.* New York: John Wiley & Sons.

Cléroux, R., J.-M. Helbling, et N. Ranger (1990). Détection d'ensembles de données aberrantes en analyse des données multivariées. *Statistique Appliquée* **38**(1), 5–21.

Cook, R. (1979). Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174.

Daniel, C. et F. Wood (1971). *Fitting Equations to Data.* New York: John Wiley and Sons.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29**, 751–760.

Fieller, N. (1976). *Some Problems related to the Rejection of Outlying Observations.* Ph.D. Thesis, University of Sheffield.

Gerrild, P. et R. Lantz (1969). *Chemical Analysis of 75 Crude Oil Samples from Pliocene Sand Units, Elk Hills Oil Field, California.* U.S. Geological Survey Open-File Report.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York: John Wiley & Sons.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Huber, P. (1977). Robust covariances, in statistical decision theory and releated topics. *Academic Press* **2**, 165–191.

Imhof, P. (1961). Computing the distribution of quadratic forms in normal variates. *Biometrika* **48**, 419–426.

Johnson, R. et D. Wichern (1992). *Applied Multivariate Statistical Analysis* (Third ed.). Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Koerts, J. et A. Abrahamse (1969). *On the Theory and Application of the General Linear Model.* Rotterdam University Press.

Lazraq, A. et R. Cléroux (1988). Un algorithme pas à pas de sélection de variables en régression linéaire multivariée. *Statistique et analyse des données* **13**, 39–58.

Lazraq, A. et R. Cléroux (1989). On the detection of multivariate data outliers and regression outliers. *Data Analysis, Learning Symbolic and Numerical Knowledge, E. Diday ed IMRIA, Nova Sci. Publ. New York*, 133–140.

Lingoes, J. et P. Schonemann (1974). Alternative measures of fit for the Schonemann-Carroll matrix fitting algorithm. *Psychometrika* **39**, 423–427.

Marazzi, A. (1985). Robust affine invariant covariances, doc $n°6$, division de statistique et informatique. *Institut Universitaire de Médecine Sociale et Préventive, Lausanne.*

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

Mardia, K., J. Kent, et J. Bibby (1982). *Multivariate Analysis.* Academic Press.

Mickey, M., O. Dunn, et V. Clark (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research* **1**, 105–111.

Murphy, R. (1951). *On Tests for Outlying Observations.* Ph.D. Thesis, Princeton University, University Microfilms Inc., Ann Arbor, Mich.

Naik, D. (1989). Detection of outliers in the multivariate linear regression model. *Comm. Statist.A - Theory Methods* **18**, 2225–2232.

Pearson, E. et C. Chandra Sekar (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika* **28**, 308–320.

Robert, p. et Y. Escoufier (1976). A unifying tool for linear multivariate statistical methods: The rv-coefficient. *Applied Statistics* **25**, 257–265.

Seber, G. (1984). *Multivariate Observations.* New York: John Wilwy and Sons.

Srivastava, M. et D. von Rosen (1998). Outliers in multivariate regression models. *Journal of Multivariate Analysis* **65**, 195–208.

Stewart, D. et W. Love (1968). A general canonical correlation index. *Psychometric Bulletin* **70**, 160–163.

Timm, N. (1975). *Multivariate Analysis with Applications in Education and Psychology.* Brooks & Cole.

Woltz, W., W. Reid, et W. Colwell (1948). Sugar and nicotine in cured bright tabacco as related to mineral element composition. *Proc. Soil Sci. Am.* **13**, 385–387.