Université de Neuchâtel
Faculté de droit et des sciences économiques

Etude comparative de différents estimateurs dans les modèles linéaires en présence de données aberrantes et du problème de la multicolinéarité en économie

Thèse

présentée à la Faculté de droit et des sciences économiques
pour obtenir le grade de docteur ès sciences économiques

par

Jean-Pierre Renfer

Neuchâtel
1997

Imprimerie de l'Evole SA, Neuchâtel

Monsieur Jean-Pierre RENFER est autorisé à imprimer sa thèse de doctorat ès sciences économiques intitulée :

"Etude comparative de différents estimateurs dans les modèles linéaires en présence de données aberrantes et du problème de la multicolinéarité en économie".

Il assume seul la responsabilité des opinions énoncées.

Neuchâtel, le 29 avril 1997

Le Doyen
de la Faculté de droit
et des sciences économiques

Pierre-Henri Bolle

A Isabelle qui elle n'a jamais douté

Préface

Comment caractériser l'aboutissement d'un travail conduisant à une thèse de Doctorat ? Les mots se bousculent sans qu'il n'y ait de vainqueur. Importance ? Souffrance ? Persévérance ? Délivrance ? A moins que ce ne soit Conviction ? Satisfaction ? Concrétisation ? Ou encore Commencement? Enrichissement ? Achèvement ? Aucun de ces mots ne sauraient traduire exactement les moments traversés au cours de cette recherche. Tous ont tenté de s'imposer sans jamais y parvenir. Pour moi, cette thèse m'aura permis de franchir une étape, de me dépasser pour aller plus loin et entreprendre de nouvelles activités dans le domaine des mathématiques et de la statistique.

Je tiens à remercier aujourd'hui les professeurs qui, chacun à leur manière, m'ont permis d'acquérir de nouvelles connaissances dans ce domaine. Je remercie le Professeur Y. Dodge, directeur de thèse. Grâce à lui j'ai pu bénéficier durant l'été 1991 d'un séjour dans le centre d'analyse multivariée du Professeur C.R. Rao à l'Université de Pennsylvanie, Etats-unis. Je le remercie également de m'avoir confié son remplacement pour le cours de mathématiques appliquées durant les années passées au sein du groupe de statistique de l'Université de Neuchâtel. Finalement, je le remercie de sa confiance pour m'avoir proposé un poste de chef de travaux ainsi que de l'opportunité de donner le cours de logiciels statistiques dans le cadre du diplôme postgrade en statistique. Je tiens également à remercier vivement madame le professeur J. Jurečkovà de l'Université de Prague pour m'avoir invité à faire un séjour scientifique durant l'été 1994. Ma profonde gratitude va également au Professeur S. Portnoy, ainsi qu'à son collègue R. Koenker, de l'Université d'Illinois à Champaign-Urbana qui a aimablement accédé à ma demande de séjourner une année dans le département de statistique avec une bourse du fonds national. J'ai ainsi eu la chance de pouvoir continuer mes recherches durant l'année académique 1995-1996 pour déposer ma thèse quelques mois après mon retour. Je remercie enfin le Professeur M. Dubois, co-rapporteur ainsi que le professeur P. Rousseeuw pour les nombreuses et intéressantes discussions que j'ai pu avoir avec lui.

Neuchâtel, Avril 1997 Jean-Pierre Renfer

Résumé

Le chapitre 1 est une introduction dans laquelle la problématique de cette étude a été expliquée, le but défini, les notations introduites et les principaux résultats de la littérature rappelés. Le chapitre 2 replace brièvement dans le contexte historique la découverte des estimateurs L_1 et L_2. Les méthodes de calcul des estimateurs L_1 ont été présentées dans le chapitre 3 en développant et en illustrant les deux principaux types d'algorithmes.

Plus de la moitié de la thèse a été consacrée au chapitre 4 dans lequel apparaissent les résultats nouveaux. Une contribution importante est celle de l'élaboration d'un estimateur hybride combinant les estimateurs L_1 et ridge, reconnus pour leur bon comportement respectif en présence de données aberrantes et du problème de la multicolinéarité. Cependant lorsque les deux problèmes se posent simultanément, aucune de ces deux méthodes n'est la plus appropriée. Une large place a été consacrée à ce nouvel estimateur, appelé estimateur L_1-ridge. Les résultats d'une étude approfondie, basée sur plusieurs critères de comparaison, ont montré que l'estimateur L_1-ridge se comporte mieux que les autres estimateurs lorsque les deux problèmes se posent simultanément. Dans cette étude comparative, une mesure de la stabilité des estimateurs a également été proposée. La stabilité de l'estimateur L_1-ridge apparaît nettement supérieure à celle de ses concurrents lorsque le degré de multicolinéarité devient important. De plus, cet estimateur peut être rapidement calculé puisque tout algorithme permettant de résoudre le problème de l'estimation L_1 peut s'appliquer aux données augmentées pour l'obtenir.

Sur le plan théorique, cette thèse fournit également des résultats nouveaux sur les propriétés qui caractérisent l'estimateur L_1-ridge. Dans le cadre de la régression simple, des formules exactes ont pu être obtenues pour calculer cet estimateur. Un important résultat a également pu être démontré dans le cas de la régression multiple qui fournit une condition suffisante pour laquelle les paramètres s'annulent. La découverte de ce résultat trouve son application dans le chapitre 5 qui présente une application originale de l'estimation L_1-ridge à un problème très important en régression, celui de la sélection d'un modèle. Cette nouvelle méthode présente l'avantage d'être simple et rapide. Elle présente en outre une fiabilité accrue en présence du problème des données aberrantes du fait que l'estimation L_1-ridge y est peu sensible.

Table des matières

Liste des Figures

Liste des Tableaux

CHAPITRE 1

INTRODUCTION

1.1 Introduction

Le but poursuivi dans cette étude est de comparer différentes méthodes d'estimation des paramètres dans les modèles linéaires lorsque l'on est confronté à deux problèmes majeurs en analyse de régression, à savoir le problème des données aberrantes et celui de la multicolinéarité.

Le modèle de régression linéaire multiple s'écrit sous la forme

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_p + \boldsymbol{\varepsilon}$$

où $\mathbf{y}$ est le $(n \times 1)$ vecteur d'observations de la variable dépendante, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ sont les $(n \times 1)$ vecteurs d'observations des variables explicatives, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, les paramètres à estimer et $\boldsymbol{\varepsilon}$ un $(n \times 1)$ vecteur d'erreurs. Sous forme matricielle, ce modèle peut s'écrire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

où $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

La notation en gras est utilisée pour désigner vecteurs et matrices.

La méthode la plus connue et sans doute la plus utilisée pour estimer les paramètres dans le modèle (1.1) est la méthode des moindres carrés. Malgré certaines propriétés optimales lorsque les erreurs sont indépendantes, identiquement distribuées selon une loi normale, le recours à des méthodes alternatives peut s'avérer judicieux. C'est le cas notamment lorsque la variable dépendante contient des données aberrantes, c'est-à-dire des observations qui se trouvent très éloignées de l'ensemble des données. Dans la littérature, ce genre d'observations porte parfois le nom de "valeurs surprises", "valeurs extrêmes", "valeurs contaminantes", "valeurs atypiques",... parmi d'autres. Pour traiter ce problème, une approche consiste à utiliser des méthodes dites robustes, puisque peu ou pas sensibles à ce type de données, comme par exemple en minimisant la somme des valeurs absolues des erreurs. Cette méthode, appelée estimation L_1 est décrite dans le prochain paragraphe. Une autre approche est basée sur les diagnostics obtenus à partir de l'estimation des moindres carrés. Comme il n'est pas rare dans la pratique de rencontrer le problème des données aberrantes, notamment en économie, leur détection est très importante. Elle permet de remettre en cause plusieurs aspects d'un problème comme par exemple de changer de modèle, d'envisager des transformations sur les variables ou encore de vérifier si le processus de mesure est correct. Dans ce cadre, un autre problème tout aussi important que l'on peut rencontrer en analyse de régression est celui de la multicolinéarité. La multicolinéarité est un terme utilisé pour caractériser une situation dans laquelle les variables permettant d'expliquer un phénomène sont fortement corrélées entre elles. Dans ce cas, les paramètres estimés par la méthode des moindres carrés peuvent avoir une variance exagérément grande et rendre ainsi les tests d'hypothèses sur les paramètres peu fiables. De plus, en présence du problème de la multicolinéarité, de faibles changements dans les valeurs des données peuvent modifier considérablement les paramètres estimés. C'est d'ailleurs pour examiner ces changements que la technique de régression ridge a été initialement introduite. Elle permet d'obtenir des paramètres de régression plus précis et plus stables que ceux estimés par la méthode des moindres carrés lorsque le problème de multicolinéarité est important. Une autre approche en cas de multicolinéarité consiste à utiliser les techniques de sélection de variables pour ne retenir que celles qui expliquent le mieux le phénomène étudié

sans être pour autant corrélées entre elles. Une nouvelle technique de sélection de variables est proposée dans le Chapitre 5 consacré aux applications.

1.2 Méthodes d'estimation

Il existe plusieurs méthodes d'estimation des paramètres; les méthodes considérées par la suite, notamment dans le Chapitre 4 consacré à la comparaison des différents estimateurs, seront principalement basées sur l'*estimation* L_1, l'*estimation* L_2 (méthode des moindres carrés) et l'*estimation ridge*.

1.2.1 Estimation L_1

Le problème de l'estimation L_1 des paramètres dans le modèle (1.1) est défini de la manière suivante.
Trouver le vecteur de paramètres $\boldsymbol{\beta}'_{L_1} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)$ qui minimise

$$\sum_{i=1}^{n} \mid y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}) \mid = \sum_{i=1}^{n} \mid y_i - \mathbf{x}_i' \boldsymbol{\beta}_{L_1} \mid = \sum_{i=1}^{n} \mid \widehat{e}_i \mid \quad (1.2)$$

où $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \cdots, x_{ip})$ est la i-ème ligne de la matrice $\mathbf{X}$, $\widehat{\boldsymbol{\beta}}_{L_1}$ l'estimateur L_1 et $\mathbf{e}'_{L_1} = (\widehat{e}_1, \ldots, \widehat{e}_n)$ le vecteur des résidus.

A noter que contrairement à l'estimation classique des moindres carrés, $\boldsymbol{\beta}_{L_1}$ ne peut être exprimé comme une fonction explicite de $\mathbf{X}$ et $\mathbf{y}$. L'estimateur L_1 possède la propriété remarquable d'être résistant aux valeurs aberrantes de la variable dépendante. Il s'agit d'une situation analogue à celle que l'on peut rencontrer dans le cas univarié, où l'estimateur L_1 est la médiane (peu ou pas sensible aux valeurs aberrantes) tandis que l'estimateur L_2 est la moyenne arithmétique (très sensible aux valeurs aberrantes). Remarquons que cet estimateur, bien qu'historiquement découvert avant celui des moindres carrés, est resté méconnu en raison des difficultés liées à son calcul et à l'absence de théorie asymptotique. Grâce aux progrès relativement récents de la programmation linéaire et aux performances des ordinateurs, cet estimateur se calcule aujourd'hui beaucoup plus facilement. Un bref historique de l'estimation L_1 et de la découverte de la méthode des moindres carrés

est donné dans le Chapitre 2. Les différentes méthodes pour calculer les estimateurs L_1 seront discutées au Chapitre 3.

1.2.2 Estimation L_2 (Méthode des moindres carrés)

Plus connue sous le nom de méthode des moindres carrés, l'estimation L_2 des paramètres dans le modèle (1.1) est définie de la façon suivante. Trouver le vecteur de paramètres $\boldsymbol{\beta}'_{L_2} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)$ qui minimise

$$\sum_{i=1}^{n}(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}))^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_{L_2})^2 = \sum_{i=1}^{n}\widehat{e}_i^2 \quad (1.3)$$

L'estimateur L_2 peut être exprimé en fonction de $\mathbf{X}$ et $\mathbf{y}$ par la formule

$$\boldsymbol{\beta}_{L_2} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (1.4)$$

L'estimateur des moindres carrés est sans doute le plus connu et le plus utilisé. Il doit sa popularité en partie au fait qu'il est facile à calculer, que sa théorie est simple, bien développée et largement documentée. De plus, il possède la variance minimale parmi tous les estimateurs linéaires non biaisés lorsque les erreurs sont non-corrélées, ont une moyenne nulle et une variance inconnue σ^2, c'est-à-dire

$$\mathbf{E}(\boldsymbol{\varepsilon}) = 0 \text{ et } \mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

où $\mathbf{V}$ est la matrice de *variance- covariance.*

Pour obtenir l'estimateur des moindres carrés de l'équation (1.4), la matrice $(\mathbf{X}'\mathbf{X})$ est supposée inversible. Or, dans le cas où les colonnes (variables) de la matrice $\mathbf{X}$ ne sont pas linéairement indépendantes, il n'est plus possible de calculer l'inverse de la matrice $(\mathbf{X}'\mathbf{X})$. On peut cependant en trouver la solution en calculant une matrice $\mathbf{G}$ telle que

$$\boldsymbol{\beta}_{L_2} = \mathbf{G}\mathbf{X}'\mathbf{y}$$

Cette matrice $\mathbf{G}$, notée ici $(\mathbf{X}'\mathbf{X})^-$ est appelée *inverse généralisée* de la matrice $(\mathbf{X}'\mathbf{X})$. Sa définition et ses propriétés sont décrites dans Arthanari and Dodge (1993).

Un autre problème survient lorsque la matrice $(\mathbf{X}'\mathbf{X})$ est mal conditionnée, c'est-à-dire lorsque les variables explicatives sont fortement corrélées entre elles. Ce problème est lié à celui de la multicolinéarité et peut notamment être traité en faisant appel aux estimateurs ridge.

1.2.3 Estimation ridge

Dans la pratique, il se peut que les variables indépendantes représentées par la matrice $\mathbf{X}$ soient linéairement dépendantes. Lorsque cette dépendance linéaire est exacte, on parle de multicolinéarité exacte (on dit aussi que les variables sont mathématiquement colinéaires). Cependant, le cas le plus fréquent est celui dans lequel l'une des variables n'est qu'approximativement une combinaison linéaire des autres. On parle dans ce cas de multicolinéarité proche (on dit aussi que les variables sont statistiquement colinéaires). Dans cette situation, la matrice $(\mathbf{X}'\mathbf{X})$ possède au moins une valeur propre proche de zéro (toutes les valeurs propres de cette matrice sont non-négatives) indiquant que son déterminant est proche de zéro. Par la suite, nous nous intéresserons essentiellement à la multicolinéarité proche et parlerons alors simplement de multicolinéarité. L'un des inconvénients majeurs en cas de multicolinéarité est l'imprécision de l'estimation L_2 des paramètres. En effet, la matrice de variance-covariance de $\boldsymbol{\beta}_{L_2}$ est donnée par

$$\mathbf{V}(\boldsymbol{\beta}_{L_2}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Ainsi, la variance totale des paramètres obtenus par la méthode des moindres carrés est

$$Var(\boldsymbol{\beta}_{L_2}) = \sigma^2 Tr(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \sum_{i=1}^{s} \frac{1}{\lambda_i} \tag{1.5}$$

où $Tr(\mathbf{X}'\mathbf{X})^{-1}$ représente la trace de $(\mathbf{X}'\mathbf{X})^{-1}$, s est le rang de la matrice $\mathbf{X}$ et λ_i les valeurs propres non nulles de la matrice $(\mathbf{X}'\mathbf{X})$. Dans l'expression (1.5), on constate que la variance totale devient exagérément grande lorsqu'une ou plusieurs valeurs propres sont très petites. Pour traiter ce problème, Hoerl (1962) fut le premier à proposer comme alternative l'estimateur ridge. Repris et développé par Hoerl et Kennard (1970a,b) l'estimateur ridge a la forme suivante

$$\boldsymbol{\beta}_{ridge} = (\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{1.6}$$

où $k \geq 0$. En d'autres termes, une constante a été ajoutée à chaque élément de la diagonale de $(\mathbf{X}'\mathbf{X})$. Souvent, de façon à pouvoir comparer plus directement les paramètres à estimer, on standardise les variables

en leur soustrayant leur moyenne et en les divisant par leur écart-type. Dans ce paragraphe, cette standardisation est implicite.

La variance totale de l'estimateur ridge devient ainsi

$$Var(\boldsymbol{\beta}_{ridge}) = \sigma^2 Tr(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1} = \sigma^2 \sum_{i=1}^{s} \frac{1}{\lambda_i + k} \qquad (1.7)$$

En comparant (1.5) et (1.7), on constate que la variance totale de l'estimateur ridge est plus petite que celle de l'estimateur des moindres carrés. En fait ce gain de variabilité se fait au détriment du biais. L'estimation des paramètres par la méthode des moindres carrés est connue pour produire des estimateurs non biaisés. Par contre, la procédure d'estimation ridge fournit des estimateurs avec biais, c'est-à-dire qu'en faisant la moyenne de ces estimateurs sur tous les échantillons possibles d'une population, cette moyenne ne sera pas égale à la valeur des paramètres de la population. Un résultat important est celui de Hoerl et Kennard (1970a); ils ont en effet montré qu'il existe toujours une constante $k > 0$ telle que l' erreur quadratique moyenne de l'estimateur ridge est plus petite que celle de l'estimateur des moindres carrés. L'erreur quadratique moyenne (MSE) étant définie par

$$MSE(\widehat{\boldsymbol{\beta}}) = Var(\widehat{\boldsymbol{\beta}}) + [Biais(\widehat{\boldsymbol{\beta}})]^2$$

$$\text{où } Biais(\widehat{\boldsymbol{\beta}}) = \mathbf{E}(\widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\beta}}$$

Le principal problème de la procédure d'estimation ridge est celui du choix de k. Une procédure graphique décrite dans Hoerl et Kennard (1970b) souvent proposée consiste à obtenir les estimateurs ridge pour différentes valeurs de k. On trace ensuite dans un diagramme les différents coefficients obtenus en fonction de k. Les coefficients correspondants aux variables colinéaires sont très instables et changent rapidement lorsque k augmente, puis se stabilisent. On choisit alors une valeur de k pour laquelle ces coefficients se sont stabilisés. Cette méthode est subjective et peut dépendre de l'échelle utilisée pour k. Cependant, depuis 1970, différentes méthodes pour déterminer k sont apparues dans la littérature. Dans ce travail, nous limiterons le choix de k à la proposition faite par Hoerl, Kennard et Baldwin (1975). En effet, nous utiliserons

$$k = \frac{p\hat{\sigma}^2}{\boldsymbol{\beta}'_{L_2}\boldsymbol{\beta}_{L_2}} \text{ avec } \hat{\sigma}^2 = \frac{\mathbf{e}'_{L_2}\mathbf{e}_{L_2}}{n-p}$$

où p représente le nombre de variables explicatives. Ce choix de k, souvent évoqué dans la littérature, semble approprié dans le sens où l'estimateur ridge correspondant a une erreur quadratique moyenne plus faible que l'estimateur L_2 en présence du problème de la multicolinéarité.

1.3 Hypothèses dans un modèle de régression

L'utilisation de la méthode des moindres carrés dans un modèle de régression linéaire nécéssite certaines hypothèses, notamment sur les erreurs. En effet, l'estimateur des moindres carrés doit sa popularité en partie au fait qu'il possède la variance minimale parmi tous les estimateurs linéaires non biaisés. Cette propriété d'optimalité n'est cependant valable que sous certaines hypotèses, à savoir

i) Les éléments du vecteur d'erreurs ϵ sont indépendants;

ii) Le vecteur d'erreurs ϵ a une moyenne de 0 et une variance σ^2.

Si l'on rajoute l'hypothèse supplémentaire :

iii) Les erreurs sont distribuées selon une *loi normale* $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$

alors, l'estimateur des moindres carrés est aussi l'*estimateur du maximum de vraissemblance*. En effet, supposons les erreurs $\epsilon_i, i = 1, \ldots, n$ indépendantes et distribuées selon une loi normale $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{\epsilon_i^2}{2\sigma^2})$$

Dans ce cas, la *fonction de vraissemblance* est définie comme le produit

$$\prod_{i=1}^{n} f(\epsilon_i) = \frac{1}{\sigma^n(2\pi)^{n/2}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2)$$

D'où, pour σ fixé, maximiser la fonction de vraissemblance revient à minimiser[1] $\sum \epsilon_i^2$. A noter que cette propriété fournit une justification théorique supplémentaire pour utiliser la méthode des moindres carrés. L'hypothèse de normalité des erreurs permet également de faire des tests d'hypothèse comme les tests de Student ou de Fischer et donc de construire des intervalles de confiance basés sur les distributions t ou F.

[1]Sauf mention du contraire, la somme s'effectue pour i allant de 1 à n.

Mentionnons enfin que les estimateurs L_1 sont des estimateurs du maximum de vraissemblance lorsque les erreurs suivent une *loi de Laplace* (double exponentielle). En effet, si les erreurs sont indépendantes et suivent une loi de Laplace

$$f(\epsilon_i) = \frac{1}{2\sigma} exp(-\frac{\mid \epsilon_i \mid}{\sigma})$$

alors la fonction de vraissemblance est donnée par

$$\prod_{i=1}^{n} f(\epsilon_i) = \frac{1}{(2\sigma)^n} exp(-\frac{1}{\sigma} \sum_{i=1}^{n} \mid \epsilon_i \mid)$$

et sa maximisation revient à minimiser $\sum \mid \epsilon_i \mid$. Cette propriété est importante dans la justification théorique de l'utilisation de la méthode basée sur la norme L_1.

L'utilisation adéquate d'un modèle de régression linéaire est soumise à certaines conditions et limitations. Il est important d'examiner ces conditions avant de procéder à l'estimation des paramètres ou d'utiliser certaines procédures inférentielles, comme les tests d'hypothèses. Ainsi, outre les hypothèses i)-iii) faites sur les erreurs, il faut prendre en considération les aspects suivants :

1. La condition la plus importante est la spécification correcte du modèle. En d'autres termes, le modèle doit décrire de manière adéquate le comportement des données.

2. Le modèle de régression doit être linéaire par rapport aux paramètres. Il faut cependant garder à l'esprit qu'il peut être non-linéaire par rapport aux variables (le carré, la racine ou l'inverse d'une variable par exemple peut être nécéssaire à la spécification du modèle).

Finalement, le fait d'avoir découvert une relation entre certaines variables ne signifie pas qu'il s'agit d'une relation de cause à effet. Ce genre de relations ne peuvent être établies sur la seule base d'une analyse de régression. Les raisons pour lesquelles les variables explicatives ont une influence causale doivent être trouvées en dehors de l'analyse de régression. D'autre part, le choix des variables explicatives devra être suggéré par la théorie ou par la connaissance que l'on a du problème.

Il faut donc retenir que l'analyse de régression ne peut pas prouver la causalité, mais seulement être utilisée comme aide dans la confirmation ou la réfutation d'hypothèses causales.

1.4 Importance de l'estimation L_1 en économie

Pour mieux comprendre l'importance et l'intérêt qu'ont suscité ces dernières décennies les estimateurs L_1, en particulier en économie, rappelons que jusqu'au milieu du XXème siècle, il n'y a eu que peu de tentatives d'utilisation de ces estimateurs. Ce manque d'intérêt est dû principalement aux raisons suivantes

1. Difficultés de calcul pour obtenir les valeurs numériques des estimateurs L_1 (il n'existe en effet pas de formule analytique analogue à celle des moindres carrés).

2. Absence de théorie asymptotique pour l'estimation L_1, et donc absence des procédures statistiques inférentielles usuelles pour le modèle de régression linéaire.

3. Manque d'évidence de la supériorité des estimateurs L_1 sur les estimateurs des moindres carrés dans le cas des petits échantillons.

Or, depuis que Charnes, Cooper et Ferguson (1955) ont montré que le problème d'estimation L_1 est un problème de programmation linéaire, ces estimateurs ont connu un regain d'intérêt considérable. Du point de vue numérique, Barrodale et Roberts (1973, 1974) ont développé un algorithme du simplexe modifié permettant ainsi d'obtenir rapidement les valeurs numériques des estimateurs L_1. Depuis lors et jusqu'à aujourd'hui, de très nombreux auteurs se sont penchés sur l'efficacité des algorithmes. Grâce à leurs recherches ainsi qu'à la rapidité des ordinateurs actuels, il est aujourd'hui pratiquement aussi facile d'obtenir les estimateurs L_1 que ceux des moindres carrés.

D'autre part, c'est Bassett et Koenker (1978) qui ont développé la théorie asymptotique des estimateurs L_1, conduisant ainsi à rendre l'estimation L_1 plus attractive. En effet, il est désormais possible de faire des tests d'hypothèses, de calculer des intervalles de confiance ou encore de dresser des tableaux d'analyse de variance. Récemment, McKean

et Schrader (1987) ont présenté une analyse de variance aussi complète et unifiée que celle des moindres carrés. Ces nouvelles connaissances ont conduit les spécialistes de différents domaines à s'intéresser de plus près aux estimateurs L_1, même si ces derniers ne sont pas souvent utilisés dans les disciplines économiques. Par exemple, de nombreux phénomènes économiques se heurtent au problème de la distribution des erreurs à grande variance. On y trouve la distribution des revenus personnels, des emplois, des prix spéculatifs, des prix de stock, des taux d'intérêt parmi d'autres. Or, les estimateurs L_1, par rapport à ceux des moindres carrés, doivent en partie leur popularité au fait que l'hypothèse de normalité des erreurs n'est pas nécéssaire. Au contraire, ils sont particulièrement adaptés aux problèmes dans lesquels la distribution des erreurs présente une courbe dont les extrémités ne tendent pas rapidement vers 0 (fat-tailed distributions). En économie, certaines distributions de ce type, comme la distribution de Cauchy ou Laplace, y jouent un rôle important. Par conséquent, les estimateurs L_1 peuvent être appropriés dans ce genre de cas. Il faut de plus mentionner que, même si la majorité des erreurs dans le modèle suivent une distribution normale, il arrive souvent qu'un petit nombre d'observations suivent une distribution différente. Dans ce cas, on dit que l'échantillon est contaminé par des valeurs aberrantes. Puisque les estimateurs L_1 sont peu sensibles aux données aberrantes, ils sont particulièrement adaptés à ce genre de situations. Finalement, Appa et Smith (1973) relèvent que le critère d'estimation L_1 intéresse particulièrement les économètres qui ont souvent à estimer les paramètres d'un modèle linéaire avec un nombre relativement restreint d'observations.

1.5 Conclusion

Le but de cette introduction est de permettre au lecteur de se familiariser avec les problèmes rencontrés en analyse de régression. Elle permet également de fournir les notations qui seront utilisées ultérieurement.

Bien que les problèmes rencontrés en analyse de régression soient nombreux, nous nous sommes limités à décrire ceux qui font l'objet de cette étude : les données aberrantes et le problème de la multicolinéarité. En présence de ces deux types de problèmes, l'utilisation de la méthode des moindres carrés, bien que largement répandue, peut

s'avérer nettement moins performante qu'une méthode alternative. Les méthodes d'estimation permettant de traiter chacun de ces problèmes pris séparément ont été décrites : il s'agit de l'estimation L_1 et de l'estimation ridge. Le recours à l'une ou l'autre de ces méthodes selon la nature du problème envisagé est justifié par les propriétés respectives de ces estimateurs. Une solution élégante pour traiter ces deux types de problèmes serait de combiner les propriétés des estimateurs L_1 et ridge. La recherche d'une telle solution fait également partie de cette étude et est présentée au Chapitre 4 consacré à la comparaison des différents estimateurs. La méthodologie adoptée est suivante : en premier lieu le comportement des estimateurs en présence de données aberrantes et du problème de la multicolinéarité est étudié sur la base de la simulation. La simulation présente en effet l'avantage de pouvoir comparer ces estimateurs dans le cas où la distribution des erreurs n'est pas normale. Il suffit pour cela de générer certaines lois de probabilité dont on connait leur capacité à produire des données aberrantes, comme par exemple les *lois de Laplace, Cauchy, Normale contaminée.* La simulation permet également de faire varier le degré de multicolinéarité en introduisant un coefficient prenant ses valeurs entre 0 et 1 permettant ainsi de passer du cas où il n'y a pas de multicolinéarité à celui où celle-ci est parfaite. Ainsi connaissant la vraie valeur des paramètres, il est possible de comparer les valeurs obtenues par les différentes méthodes d'estimation avec ces vraies valeurs. De manière à obtenir la comparaison la plus objective possible, différents critères de comparaison seront adoptés.

L'étude porte également sur des données déjà étudiées dans la littérature, notamment au Chapitre 5, dans le cadre de la sélection de variables, ce qui permet de comparer les nouveaux résultats à ceux déjà publiés.

CHAPITRE 2

DECOUVERTE DES ESTIMATEURS L_1 ET L_2

2.1 Introduction

Le problème fondamental de trouver la meilleure équation linéaire (ou non linéaire) permettant d'exprimer la relation entre une variable dépendante et une ou plusieurs variables indépendantes a été depuis des siècles le centre d'intérêt des scientifiques. Les données sont des observations sujettes à des erreurs aléatoires. Le choix des meilleures mesures de tendance centrale et de dispersion de ces observations dépend de la distribution des erreurs aléatoires.

Si l'on suppose que les valeurs des variables indépendantes sont connues exactement et que les erreurs dans les observations de la variable dépendante sont distribuées normalement, alors il est bien connu que la moyenne est la meilleure mesure de tendance centrale, l'écart-type la meilleure mesure de dispersion et la méthode des moindres carrés (estimation L_2) la meilleure méthode de régression.

Cependant d'autres hypothèses conduisent à des choix différents. Certains utilisateurs ont tendance à faire l'hypothèse de normalité des erreurs sans se préocupper des conséquences lorsque cette hypothèse n'est pas satisfaite. Le problème se pose en particulier lorsque les observations sont contaminées par des observations aberrantes provenant de distributions ayant des moyennes différentes ou des variances plus grandes. Dans ce cas, des méthodes alternatives à celle de l'estimation L_2 sont

plus adéquates. Parmi ces méthodes alternatives, celle de l'estimation L_1 joue un rôle important du fait qu'elle a l'avantage d'être moins sensible aux données aberrantes. La première partie de ce chapitre est consacrée à l'évolution chronologique de la méthode d'estimation L_1 depuis son origine avec l'approche de Boscovich (1757) jusqu'au milieu du $XX^{ème}$ siècle avec celle de Charnes, Cooper et Fergusson (1955). La seconde partie relate brièvement la découverte de la méthode des moindres carrés, postérieure à celle de l'estimation L_1, et à la controverse liée à sa paternité.

2.2 L'estimation L_1 (1757-1955)

Parmi les estimateurs robustes, les estimateurs L_1 ont probablement l'histoire la plus ancienne. En effet, Ronchetti (1987) mentionne qu'on en retrouve des traces dans l'oeuvre de Galilée (1632), intitulée "Dialogo dei massimi sistemi". Le problème était alors de déterminer la distance de la terre à une étoile récemment découverte à cette époque. C'est cependant à Boscovich (1757) que l'on reconnait généralement l'introduction du critère d'estimation L_1 (Harter (1974a), Ronchetti (1987), Dielman (1992)).

L'un des problèmes qui excita le plus la curiosité des hommes de science du $XVIII^{ème}$ siècle fut celui de la détermination de l'ellipticité de la terre. C'est dans ce contexte, près d'un demi-siècle avant l'annonce par Legendre (1805) du principe des moindres carrés et vingt ans avant la naissance de Gauss en 1777, que Roger Joseph Boscovich (1757) proposa une procédure pour déterminer les paramètres du modèle de régression linéaire simple $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$. Pour obtenir la droite $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ décrivant au mieux les observations, il proposa le critère de l'estimation L_1

$$\min \sum_{i=1}^{n} \mid y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \mid$$

en imposant que la droite estimée passe par le centroïde des données $(\overline{x}; \overline{y})$ en ajoutant la condition

$$\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0.$$

Boscovich justifia cette approche de la manière suivante. Le critère de l'estimation L_1 comme étant nécéssaire pour que la solution soit aussi proche que possible des observations, et la condition supplémentaire pour que les erreurs positives et négatives soient de probabilité égales. En effet cette condition signifie que la somme des erreurs positives et négatives doit être la même. De plus, elle peut se mettre sous la forme $\overline{y} - \widehat{\beta}_0 - \widehat{\beta}_1\overline{x} = 0$. D'où

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \tag{2.1}$$

et le problème se réduit alors à minimiser

$$\sum_{i=1}^{n} \mid (y_i - \overline{y}) - \widehat{\beta}_1(x_i - \overline{x}) \mid \tag{2.2}$$

Par conséquent, la détermination de la "droite de Boscovich" satisfaisant les deux critères revient à déterminer la pente $\widehat{\beta}_1$ de l'équation (2.2), puis à évaluer l'ordonnée à l'origine $\widehat{\beta}_0$ par l'équation (2.1). Ce n'est cependant que trois ans plus tard, en 1760, que Boscovich donna une procédure géométrique permettant de résoudre l'équation (2.2). Cette procédure est décrite en détails dans un article d'Eisenhart (1961).

Sept ans avant de s'intéresser aux estimateurs L_1, Laplace (1786) proposa une procédure permettant d'estimer les paramètres d'un modèle de régression linéaire simple en se basant sur le critère L_∞. En d'autres termes, il proposa une solution pour trouver $(\widehat{\beta}_0, \widehat{\beta}_1)$ qui minimise

$$\max_{1\leq i\leq n} \mid y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \mid = \max_{1\leq i\leq n} \mid \widehat{e}_i \mid$$

Dans une publication ultérieure, Laplace (1793) proposa une procédure qu'il qualifie lui-même de plus simple. Cette procédure, basée sur les deux critères qu'avait proposé Boscovich en 1757, a l'avantage d'etre analytique alors que celle proposée par Boscovich était géométrique. L'intérêt de cette procédure analytique réside dans la facilité à obtenir les paramètres $\widehat{\beta}_0$ et $\widehat{\beta}_1$ lorsque le nombre d'observations augmente. Cette solution analytique de Laplace est élégante et mérite d'être rappelée ici. En adoptant les notations suivantes

$$Y_i = y_i - \overline{y} \text{ et } X_i = x_i - \overline{x}$$

le problème revient à trouver la valeur de β qui minimise la fonction

$$f(\beta) = \sum_{i=1}^{n} \mid Y_i - \beta X_i \mid \tag{2.3}$$

Notons que les valeurs X_i peuvent être supposées non nulles ($X_i \neq 0$) puisque $f(\beta) = \sum \mid Y_i \mid + \sum \mid Y_i - \beta X_i \mid$, la première somme étant prise pour les X_i nuls et la seconde somme pour les X_i non nuls. Le minimum de la fonction f étant atteint pour la même valeur de β que celle rendant la seconde somme minimale. La fonction (2.3) peut s'écrire

$$f(\beta) = \sum_{i=1}^{n} f_i(\beta) \text{ avec } f_i(\beta) = \mid Y_i - \beta X_i \mid$$

Chaque fonction $f_i(\beta)$ est continue, linéaire par morceaux et convexe. Elle est formée de deux droites avec un minimum en $(\frac{Y_i}{X_i}; 0)$. Sa pente est donnée par

$$f_i^{'}(\beta) = \begin{cases} - \mid X_i \mid & \text{si } \beta < \frac{Y_i}{X_i} \\ + \mid X_i \mid & \text{si } \beta > \frac{Y_i}{X_i} \end{cases}$$

Pour étudier la pente de $f(\beta)$, il s'agit d'ordonner en ordre croissant les rapports $\frac{Y_i}{X_i}$ de manière à ce que :

$$\frac{Y_1}{X_1} \leq \frac{Y_2}{X_2} \leq \ldots \leq \frac{Y_n}{X_n}$$

Ceci peut toujours être fait en renumérotant les observations. Ces rapports $\frac{Y_k}{X_k}$ seront désignés par $\beta_{(k)}$, $k = 1, \ldots, n$.
Pour $\beta < \beta_{(1)}$, chacune des fonctions $f_i(\beta)$ a une pente de $- \mid X_i \mid$ et par conséquent la pente de la fonction (2.3) est donnée par

$$f^{'}(\beta) = -\sum_{i=1}^{n} \mid X_i \mid$$

En chaque point $\frac{Y_k}{X_k}$, la pente de $f(\beta)$ augmente de $2 \mid X_k \mid$, $k = 1, \ldots, n$. $f(\beta)$ étant continue, linéaire par morceaux et convexe, elle atteint son minimum lorsque sa pente change de signe, c'est-à-dire pour $\beta_{(r)}$ tel que

$$-\sum_{i=1}^{n} \mid X_i \mid + 2\sum_{k=1}^{r-1} \mid X_k \mid < 0 \text{ et } -\sum_{i=1}^{n} \mid X_i \mid + 2\sum_{k=1}^{r} \mid X_k \mid \geq 0$$

Dans le cas où

$$-\sum_{i=1}^{n} \mid X_i \mid +2\sum_{k=1}^{r} \mid X_k \mid = 0$$

la solution n'est *pas unique*; dans ce cas, pour toute valeur $\widehat{\beta}$ telle que $\beta_{(r)} \leq \widehat{\beta} \leq \beta_{(r+1)}$, $f(\beta)$ est minimale. Notons encore que $\widehat{\beta} = \frac{Y_r}{X_r}$ est appelée *médiane pondérée* des $\frac{Y_i}{X_i}$, avec poids $\mid X_i \mid$. Ainsi, dans le cas où la droite L_1 doit satisfaire le second critère de Boscovich, elle passe par le centroïde des données et par l'une des observations au moins.

C'est à Gauss (1809) que l'on doit une étape importante de la caractérisation des estimateurs L_1. Contrairement à Boscovich, il étudia la méthode consistant à minimiser la somme des erreurs en valeur absolue sans la restriction que leur somme soit nulle (appliquant uniquement le critère 1). A cette époque, Gauss ne semblait d'ailleurs pas savoir que cette restriction avait été introduite par Boscovich, puisqu'il l'attribue à Laplace. D'autre part, Gauss s'intéressa à l'estimation L_1 dans un modèle de régression linéaire multiple, en cherchant le vecteur de paramètres $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ qui minimise

$$\sum_{i=1}^{n} \mid y_i - x_{i1}\widehat{\beta}_1 - \ldots - x_{ip}\widehat{\beta}_p \mid = \sum_{i=1}^{n} \mid \widehat{e}_i \mid$$

Il mentionna que cette méthode fournit nécéssairement p résidus nuls et qu'elle n'utilise les autres $(n-p)$ résidus que dans la détermination du choix des p résidus nuls. De plus, il mentionne que la solution obtenue par cette méthode n'est pas modifiée si la valeur des y_i est augmentée ou diminuée sans que les résidus changent de signe. Gauss remarqua également que la méthode consistant à

$$\text{minimiser } \sum_{i=1}^{n} \mid \widehat{e}_i \mid \text{ avec la restriction que } \sum_{i=1}^{n} \widehat{e}_i = 0$$

fournit nécéssairement $(p-1)$ résidus nuls. Dans le cas de la régression linéaire simple $(p = 2)$ traitée par Laplace avec la restriction que la somme des résidus soit nulle, on obtient effectivement une droite passant par l'une des observations, c'est-à-dire qu'un des résidus est nul.

Bloomfield et Steiger (1983) prouvent ce résultat et indiquent qu'il pourrait bien être l'un des premiers en programmation linéaire, mais pas assez profond pour que Gauss le démontre.

Avec l'annonce par Legendre (1805) de la méthode des moindres carrés et son développement par Gauss (1809, 1823, 1828) et Laplace (1812) basé sur la théorie des probabilités, la méthode d'estimation L_1 joua un rôle secondaire durant la plus grande partie du $XIX^{ème}$ siècle. Ce n'est qu'en 1887 que cette méthode refait surface grâce au travail d'Edgeworth. En effet, Edgeworth supprima la restriction faite par Boscovich que la somme des résidus soit nulle. Il présenta d'un point de vue géométrique une procédure générale décrite ici dans le cas de la régression linéaire simple ($p = 2$). Les n observations sont notées $P_1(x_1; y_1), \ldots, P_n(x_n; y_n)$. En posant $m_0 = 1$ et en traitant P_{m0} comme l'origine (en soustrayant P_1 des autres observations), la procédure de Laplace peut s'appliquer. Or la droite ainsi forcée de passer par P_{m0} contiendra l'une des autres observations, disons P_{m_1}. Traitant cette nouvelle observation comme l'origine, on trouve une droite passant par une autre observation P_{m_2} et ainsi de suite. Cette procédure ne requiert pas plus de $r = n - 1$ étapes, chacune représentant le calcul d'une médiane pondérée, puisqu'elle se termine lorsque $P_{m_r} = P_{m_{r-2}}$. Lorsque $p > 2$, l'algorithme , bien que plus compliqué, est analogue et revient à fixer $(p - 1)$ des paramètres à estimer puis à utiliser la procédure de Laplace pour déterminer la valeur optimale du paramètre restant. Notons finalement que l'algorithme décrit ci-dessus dans le cas de la régression linéaire simple présente certains défauts. Par exemple, il se peut que sur l'une des droites obtenues, il y ait trois observations (d'indice i_1, i_2 et i_3) conduisant la procédure à faire un cycle de la façon suivante : $i_1 \rightarrow i_2 \rightarrow i_3 \rightarrow i_1 \ldots$ sans conduire pour autant à diminuer la somme des résidus en valeur absolue comme l'indique Sposito (1976). Karst (1958) mentionne que cette procédure peut s'arrêter prématurément. C'est le cas lorsque par exemple la droite forcée à passer par P_{i_1} contient l'observation P_{i_2} et vice-versa, mais n'est pas optimale. Une implentation en langage Fortran de cet algorithme a été faite par Sadovski (1974) permettant d'éviter les problèmes décrits ci-dessus.

Rhodes (1930) trouva la solution graphique d'Edgeworth difficile à appliquer en pratique. Il proposa alors une procédure itérative que l'on peut résumer ainsi

1. Choisir arbitrairement $p - 1$ équations.

2. Les utiliser pour éliminer les $p-1$ premiers paramètres du problème.

3. Utiliser la procédure de Laplace pour estimer le paramètre restant.

4. Associer l'équation correspondant au point 3 à l'ensemble des $p-1$ équations.

5. Si l'ensemble des p équations se répète p fois, on s'arrête. Sinon, on élimine l'équation la plus ancienne et l'on retourne au point 2.

Cette procédure reste cependant difficile à utiliser dans des problèmes pratiques compte tenu des moyens de l'époque.

C'est grâce au travail de Charnes, Cooper et Fergusson (1955) que l'intérêt porté aux estimateurs L_1 a été le plus stimulé. Comme alternative à la méthode des moindres carrés, ils proposèrent l'utilisation de la programmation linéaire pour calculer les estimateurs L_1. Le contexte dans lequel ils développèrent la méthode d'estimation L_1 basée sur la programmation linéaire est économique. En effet, ils s'intéressèrent aux salaires à verser aux employés d'une entreprise : selon la position hiérarchique qu'un individu occupe, son salaire correspond à un niveau donné, étant entendu qu'il y a un salaire minimum et maximum. Cependant, d'autres facteurs peuvent influencer la performance des employés dans leur travail. Les facteurs retenus sont ici au nombre de neuf et correspondent respectivement à l'efficacité à travailler avec les autres, l'acceptation de responsabilités, l'initiative, l'expérience, le niveau d'éducation, la capacité à s'exprimer, l'aptitude à planifier, l'aptitude mentale et l'aptitude à exécuter les tâches. En procédant à des auditions soigneusement préparées et en faisant passer des tests au personnel de l'entreprise, des points (entre 0 et 5) ont été distribués pour chacun des neuf facteurs. Les salaires étant exprimés comme combinaison linéaire de ces facteurs, le but était de minimiser la somme des déviations en valeur absolue entre les niveaux de salaire selon la hiérarchie et les salaires estimés en fonction des neuf facteurs pouvant influencer la performance des employés. Pour résoudre ce problème, Charnes, Cooper et Fergusson (1955) ont montré que le problème de régression linéaire multiple basé sur la norme L_1 peut se mettre sous la forme d'un problème de programmation linéaire; pour cela, ils considèrent les résidus comme la différence de deux variables non négatives. En posant $e_i = u_i - v_i$, où $u_i, v_i \geq 0$ représentent les déviations positives et négatives respectivement, le problème devient

$$\begin{aligned} &\text{minimiser} \sum_{i=1}^{n}(u_i + v_i) \\ &\text{s.c.} \sum_{j=1}^{p} \widehat{\beta}_j x_{ij} + u_i - v_i = y_i \\ &\text{et } u_i, v_i \geq 0, i = 1, \ldots, n. \end{aligned}$$

où $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ sont sans restriction de signe. Ainsi, le problème de régression linéaire multiple basé sur la norme L_1 peut être formulé comme un problème de programmation linéaire avec $2n + p$ variables et n contraintes. Cette formulation du problème correspond à la forme primale et a pu être résolu en utilisant la méthode du simplexe. Cependant, il a rapidement été reconnu que la structure de ce type de problème pouvait être prise en compte pour améliorer la performance des algorithmes. En effet, Wagner (1959) proposa une formulation de ce problème basée sur le dual, ce qui permit à Barrodale et Young (1966) de mettre au point un algorithme relativement rapide. Par la suite, de nombreuses publications furent consacrées à l'élaboration d'algorithmes de plus en plus performants, dont le développement sera repris dans le chapitre suivant.

2.3 La découverte de l'estimation L_2

La découverte de l'estimation L_2 (méthode des moindres carrés) mérite d'être rappelée ici puisqu'elle fut à l'origine de l'une des plus grandes disputes dans l'histoire de la statistique. Adrien Marie Legendre (1805) publia le premier la méthode des moindres carrés. Il donna une explication claire de la méthode en donnant les équations normales et en fournissant un exemple numérique.

Selon Stigler (1981), Robert Adrain, un américain, publia la méthode vers la fin de l'année 1808 ou au début de l'année 1809. Selon Stigler (1977, 1978), il se pourrait que Robert Adrain ait "découvert" cette méthode dans l'ouvrage de Legendre (1805). Cependant, quatre ans après la publication de Legendre, Gauss (1809) a le courage de réclamer la paternité de la méthode des moindres carrés, en prétendant l'avoir utilisée depuis 1795. La revendication de Gauss déclencha l'une des plus grandes disputes scientifiques dont les détails sont présentés et résumés dans un article de Plackett (1972).

Bien que le doute subsiste, plusieurs faits troublants semblent indiquer que Gauss a effectivement utilisé la méthode des moindres carrés avant 1805. En particulier, Gauss prétend qu'il a parlé de cette méthode à certains astronomes (Olbers, Lindenau et von Zach) avant 1805. De plus, dans une lettre de Gauss datant de 1799, il est fait mention de "ma méthode", sans qu'un nom y soit donné. Il semble difficile de ne pas le croire, vu l'extraordinaire compétence reconnue à Gauss comme mathématicien.

Il reste cependant une question très importante : quelle importance attachait Gauss à cette découverte ? La réponse pourrait être que Gauss, bien que jugeant cette méthode utile, n'a pas réussi à communiquer son importance à ses contemporains avant 1809. En effet, dans sa publication de 1809, Gauss est allé bien plus loin que Legendre dans ses développements autant conceptuels que techniques. C'est dans cet article qu'il lie la méthode des moindres carrés à la loi normale (Gaussienne) des erreurs. Il propose également un algorithme pour le calcul des estimateurs. Son travail a d'ailleurs été discuté par plusieurs auteurs comme Seal (1967), Eisenhart (1968), Goldstine (1977), Sprott (1978) et Sheynin (1979).

Gauss a certainement été le plus grand mathématicien de cette époque, mais c'est Legendre qui a cristallisé l'idée de la méthode des moindres carrés sous une forme compréhensible par ses contemporains.

CHAPITRE 3

METHODES DE CALCUL DES ESTIMATEURS L_1

3.1 Introduction

Depuis la publication de l'article de Charnes, Cooper et Ferguson (1955), dans lequel ils ont montré que l'algorithme du simplexe pouvait être utilisé pour le calcul des estimateurs L_1, ces derniers ont connu un regain d'intérêt considérable. Les algorithmes permettant d'obtenir les estimateurs L_1 seront abordés dans ce chapitre. Ils se scindent en deux catégories générales :

- les algorithmes basés sur la programmation linéaire
- les algorithmes basés sur la méthode de descente.

Les algorithmes basés sur la programmation linéaire sont décrits dans le prochain paragraphe dans le cadre de l'estimation L_1, en supposant la méthode du simplexe connue du lecteur. Ils sont ensuite illustrés par celui de Barrodale et Roberts, puisqu'il est représentatif de ce type d'algorithmes. Nous illustrerons les algorithmes basés sur la méthode de descente par l'algorithme de Wesolowsky. De nombreux algorithmes ont été élaborés dans le domaine des méthodes de calcul des estimateurs L_1. Nous consacrerons également un paragraphe au choix de l'algorithme le plus rapide.

3.2 Algorithmes de programmation linéaire

Le calcul des estimateurs L_1 s'est vu considérablement facilité à partir de la seconde moitié de ce siècle. En effet, grâce au développement de la programmation linéaire et surtout à la reconnaissance du problème d'estimation L_1 comme problème de programmation linéaire, des progrès considérables ont pu être réalisés dans ce domaine. La majorité des développements algorithmiques récents sont basés sur la programmation linéaire qui présente le double avantage de la rapidité d'algorithmes spécialement conçus pour calculer ces estimateurs ainsi que de la possibilité d'obtenir toutes les solutions du problème lorsque le vecteur de paramètres n'est pas unique. De plus, les techniques de postoptimalité et les résultats sur la stabilité pour la programmation linéaire peuvent être appliqués aux estimateurs L_1. Il est ainsi possible de caractériser les effets de changements dans les données sur ces estimateurs, comme l'ont proposé Arthanari et Dodge (1993), approche reprise et développée par Dupacova (1992).

En principe, la paternité de la reconnaissance du problème d'estimation L_1 comme problème de programmation linéaire revient à Charnes, Cooper et Ferguson (1955). Cependant, Bloomfield et Steiger (1980) mentionnent que Harris (1950) l'a également observé mais ne citent pas l'article de Charnes, Cooper et Ferguson de 1955 qui est pourtant de très loin la référence la plus souvent donnée. Toujours est-il que cette nouvelle formulation a été l'un des tournants majeurs dans le calcul des estimateurs L_1. Par la suite et jusqu'à aujourd'hui, de nombreux auteurs se sont penchés sur l'amélioration des algorithmes basés sur la programmation linéaire. Les algorithmes les plus performants ont été comparés et les résultats obtenus figurent dans le paragraphe consacré au choix de l'algorithme le plus rapide.

Voyons à présent comment le problème de l'estimation L_1 peut se mettre sous la forme d'un problème de programmation linéaire. Rappelons qu'il s'agit de trouver le vecteur de paramètres $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ qui minimise $\sum \mid e_i \mid$, où $y_i = \sum \beta_j x_{ij} + e_i$, $i = 1, \ldots, n$. Posons

$$e_i = p_i - n_i \ , \ i = 1, \ldots, n$$

où p_i et n_i représentent les déviations positives et négatives respectivement. Il sont appelés *résidus complémentaires* puisque si $e_i \geq 0$ alors $p_i = e_i$ et $n_i = 0$, sinon $p_i = 0$ et $n_i = -e_i$. Posons également

			0	...	0	−1	...	−1	...	−1
$\mathbf{c_B}$	base	$\mathbf{R_B}$	$\boldsymbol{\pi_1}$	...	$\boldsymbol{\eta_p}$	$\mathbf{p_1}$	...	$\mathbf{n_1}$	...	$\mathbf{n_n}$
−1	$\mathbf{p_1}$	y_1	x_{11}	...	$-x_{1p}$	1	...	−1	...	0
−1	$\mathbf{p_2}$	y_2	x_{21}	...	$-x_{2p}$	0	...	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋱	⋮
−1	$\mathbf{p_n}$	y_n	x_{n1}	...	$-x_{np}$	0	...	0	...	−1
	$z_j - c_j$	$-\sum y_i$	$-\sum x_{i1}$	...	$\sum x_{ip}$	0	...	2	...	2

Tableau 3.1: Tableau initial du simplexe pour l'estimation L_1.

$$\widehat{\beta}_j = \pi_j - \eta_j \ , \ j = 1, \ldots, p$$

où π_j et η_j sont des *paramètres complémentaires*, c'est-à-dire si $\widehat{\beta}_j \geq 0$ alors $\pi_j = \widehat{\beta}_j$ et $\eta_j = 0$, sinon $\pi_j = 0$ et $\eta_j = -\widehat{\beta}_j$. Dès lors le problème peut être exprimé comme un problème de programmation linéaire

$$\text{minimiser } z = \sum_{i=1}^{n} (p_i + n_i)$$

$$\text{s.c. } \sum_{j=1}^{p} (\pi_j - \eta_j) x_{ij} + p_i - n_i = y_i \ , \ i = 1, \ldots, n$$

$$p_i, n_i \geq 0 \ , \ i = 1, \ldots, n$$

$$\pi_j, \eta_j \geq 0 \ , \ j = 1, \ldots, p.$$

Le tableau initial du simplexe correspondant à ce problème est indiqué dans le tableau (3.1). Le nom des colonnes dans ce tableau correspond aux notations introduites pour l'estimation L_1 et le vecteur des variables de base est noté $\mathbf{R}_B$. Dans la pratique, le tableau (3.1) est présenté (respectivement implanté sur ordinateur) sous sa *forme condensée,* avec seulement $(p+1)$ colonnes, $\mathbf{R}_B$ et π_j, $j = 1, \ldots, p$, comme indiqué dans le tableau (3.2).

La mémorisation de ces $(p+1)$ vecteurs est suffisante pour appliquer l'algorithme du simplexe puisque

base	$\mathbf{R_B}$	$\boldsymbol{\pi_1}$	...	$\boldsymbol{\pi_p}$
$\mathbf{p_1}$	y_1	x_{11}	...	x_{1p}
$\mathbf{p_2}$	y_2	x_{21}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{p_n}$	y_n	x_{n1}	...	x_{np}
$z_j - c_j$	$-\sum y_i$	$-\sum x_{i1}$	...	$-\sum x_{ip}$

Tableau 3.2: Tableau initial du simplexe sous sa forme condensée.

$$\mathbf{p}_i + \mathbf{n}_i = 0 \ , \ i = 1, \ldots, n$$

$$\pi_j + \eta_j = 0 \ , \ j = 1, \ldots, p.$$

De plus, en notant par $z(p_i), z(n_i), z(\pi_j)$ et $z(\eta_j)$ les valeurs des termes $(z_j - c_j)$ correspondants, les relations suivantes sont satisfaites

$$z(\mathbf{p}_i) + z(\mathbf{n}_i) = 2 \ , \ i = 1, \ldots, n$$

$$z(\pi_j) + z(\eta_j) = 0 \ , \ j = 1, \ldots, p.$$

Il est ainsi possible à chaque itération de la méthode du simplexe de déduire les quantités éliminées à partir de celles mémorisées. Finalement, mentionnons que si l'un des termes y_i est négatif dans le tableau du simplexe initial (ou l'un des termes R_{Bi} dans le tableau correspondant à l'une des itérations), la restriction de non-négativité imposée aux éléments R_{Bi} dans l'algorithme du simplexe peut être satisfaite en changeant de signe tous les éléments de la ligne i et en remplaçant la variable de base par son complémentaire hors base.

Voyons à présent comment fonctionne cette méthode sur un exemple en prenant le modèle de régression linéaire simple : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ et les données $(x_i;y_i)$: (1;1), (2;1), (3;2), (4;3) et (5;2). Le but est de trouver les deux paramètres $\widehat{\beta}_0$ et $\widehat{\beta}_1$ obtenus par la méthode d'estimation L_1. En utilisant les notations introduites ci-dessus, le problème s'écrit de la manière suivante

$$\text{minimiser} \sum_{i=1}^{5} (p_i + n_i)$$

base	$\mathbf{R_B}$	π_1	π_2
$\mathbf{p_1}$	1	1	1
$\mathbf{p_2}$	1	1	2
$\mathbf{p_3}$	2	1	3
$\mathbf{p_4}$	3	1	4
$\mathbf{p_5}$	2	1	5*
$z_j - c_j$	-9	-5	-15

Tableau 3.3: Tableau condensé initial du simplexe (* pivot).

s.c. $(\pi_1 - \eta_1) + (\pi_2 - \eta_2)x_i + p_i - n_i = y_i$, $i = 1, \ldots, 5$

$p_i, n_i \geq 0$, $i = 1, \ldots, 5$

$\pi_j, \eta_j \geq 0$, $j = 1, 2$.

Résolvons ce problème avec la méthode du simplexe en faisant les calculs sur les tableaux condensés. Le tableau initial est représenté dans le tableau (3.3).

A l'aide du critère d'entrée et de sortie, nous trouvons l'élément pivot qui est indiqué dans le tableau du simplexe condensé. Nous pouvons dès lors calculer l'itération suivante qui figure dans le tableau (3.4). Comme il reste un $(z_j - c_j) < 0$, nous calculons l'itération suivante en échangeant cette fois le vecteur $\mathbf{p_2}$ contre le vecteur π_1. Le nouveau tableau figure dans le tableau (3.5). A noter que dans le tableau (3.5), le vecteur $\mathbf{n_2}$ remplace le vecteur $\mathbf{p_2}$. Ceci provient du fait que $z(\mathbf{p_2}) = 10/3$ et donc $z(\mathbf{n_2}) = -4/3$ puisque $z(\mathbf{p_2}) + z(\mathbf{n_2}) = 2$. Seul le vecteur pour lequel $(z_j - c_j)$ est négatif apparaît dans le tableau. Finalement, il reste une itération pour obtenir une solution optimale. Le tableau (3.6) correspond à cette itération. Comme il n'y a plus de $(z_j - c_j) < 0$ dans le tableau (3.6), la solution est optimale. Le fait que l'un des $(z_j - c_j) = 0$ indique que la solution peut ne pas être unique. Dans ce cas, il est possible de procéder à une nouvelle itération qui ne changera pas la valeur de la fonction objectif mais qui fournit d'autres paramètres. Dans notre cas, la solution optimale est donnée par $\widehat{\beta}_0 = 3/4$ et $\widehat{\beta}_1 = 1/4$. L'équation de la droite L_1 est donc : $y = 3/4 + 1/4x$. Graphiquement, ces trois itérations correspondent à la recherche d'une droite qui rend la somme des déviations en valeur absolue plus petite à chaque itération.

Après la première itération, la droite passe par l'origine et l'observation 5 ($\mathbf{p_5}$ et $\mathbf{n_5}$ n'apparaissent pas dans la base et sont donc nuls).

base	$\mathbf{R_B}$	π_1	$\mathbf{p_5}$
$\mathbf{p_1}$	3/5	4/5	−1/5
$\mathbf{p_2}$	1/5	3/5*	−2/5
$\mathbf{p_3}$	4/5	2/5	−3/5
$\mathbf{p_4}$	7/5	1/5	−4/5
π_2	2/5	1/5	1/5
$z_j - c_j$	-3	-2	3

Tableau 3.4: Tableau du simplexe après la première itération (* pivot).

base	$\mathbf{R_B}$	$\mathbf{n_2}$	$\mathbf{p_5}$
$\mathbf{p_1}$	1/3	4/3*	1/3
π_1	1/3	−5/3	−2/3
$\mathbf{p_3}$	2/3	2/3	−1/3
$\mathbf{p_4}$	4/3	1/3	−2/3
π_2	1/3	1/3	1/3
$z_j - c_j$	-7/3	-4/3	5/3

Tableau 3.5: Tableau du simplexe après la deuxième itération (* pivot).

Après la deuxième itération, la droite passe par les observations 2 et 5 et après la troisième itération la droite passe par les observations 1 et 5. Cette dernière droite est la droite L_1 et la somme des déviations en valeur absolue vaut 2. A noter qu'en procédant à une nouvelle itération, les paramètres trouvés auraient été : $\widehat{\beta}_0 = 1/2$ et $\widehat{\beta}_1 = 1/2$. En refaisant une itération, on retombe sur les paramètres trouvés initialement. Ainsi, dans cet exemple, il y a deux droites L_1 extrêmes.

3.3 L'algorithme de Barrodale et Roberts

L'algorithme du simplexe pour résoudre le problème de l'estimation L_1 décrit dans la section précédente n'est pas optimal du point de vue de sa vitesse. Barrodale et Roberts (1973) ont proposé une modification

de cet algorithme permettant de réduire considérablement le nombre d'itérations. Encore aujourd'hui, l'algorithme implanté dans la plupart des logiciels permettant de calculer ces estimateurs est celui de Barrodale et Roberts. Pour illustrer cet algorithme, nous décrivons tout d'abord les modifications à apporter à l'algorithme usuel du simplexe, puis nous appliquerons cet algorithme à l'exemple de la section précédente.

Par rapport à l'algorithme usuel du simplexe, les modifications de l'algorithme de Barrodale et Roberts sont les suivantes. L'algorithme se décompose en deux parties. La première consiste à ne faire entrer dans la base que les vecteurs $\boldsymbol{\pi}_j$ (ou $\boldsymbol{\eta}_j$) selon le critère usuel de la méthode du simplexe. Cette première partie comprend au maximum p étapes (correspondant au nombre de paramètres). La seconde partie de l'algorithme consiste à ne faire entrer que les vecteurs $\mathbf{p}_i$ (ou $\mathbf{n}_i$) dans la base. Ici encore le critère d'entrée reste classique.

La modification essentielle qui s'applique aux deux parties de l'algorithme réside dans le critère de sortie de la base. Dans les deux cas, le vecteur qui doit quitter la base est choisi parmi les vecteurs de base $\mathbf{p}_i$ et $\mathbf{n}_i$ en sélectionnant le vecteur qui améliore le plus la valeur de la fonction objectif. La règle classique pour déterminer le vecteur qui doit quitter la base est modifiée de la manière suivante : on commence par appliquer la règle classique pour déterminer la ligne pivot parmi les vecteurs de base $\mathbf{p}_i$ et $\mathbf{n}_i$. Si, en additionnant deux fois la valeur du pivot à la valeur $(z_j - c_j)$ de la colonne pivot, on obtient un résultat non négatif alors on procède à une transformation du tableau (itération). Sinon on additionne deux fois la ligne pivot à la ligne des $(z_j - c_j)$, on multiplie la ligne pivot par -1 et on remplace le vecteur $\mathbf{p}_i$ (ou $\mathbf{n}_i$) correspondant à la ligne pivot par son complémentaire $\mathbf{n}_i$ (ou $\mathbf{p}_i$). Cette opération permet d'améliorer la fonction objectif tout en gardant une solution réalisable de base et sans procéder à une itération. On répète cette procédure jusqu'à l'obtention d'un pivot qui ne peut plus être rejeté et l'on procède à une transformation classique du simplexe.

Appliquons à présent l'algorithme de Barrodale et Roberts à l'exemple de la section précédente. Le tableau initial du simplexe figure dans le tableau (3.7). Le pivot classique du simplexe est (5*). En additionnant deux fois cette ligne pivot à la dernière, on trouve

$z_j - c_j$	-5	-3	-5

La valeur $(z_j - c_j)$ correspondant à la colonne pivot étant strictement négative (-5), on ne procède pas à une itération mais on multiplie la cin-

quième ligne par -1 et on remplace le vecteur $\mathbf{p}_5$ par $\mathbf{n}_5$. Il s'agit dès lors de trouver le pivot suivant qui est (2**). La dernière ligne devient ainsi

$z_j - c_j$	-3	-1	-1

. Ici encore la valeur de $(z_j - c_j)$ correspondant à la colonne pivot est strictement négative (-1). On multiplie donc la deuxième ligne par -1 et on remplace $\mathbf{p}_2$ par $\mathbf{n}_2$. La recherche du nouveau pivot conduit à (3***). Dans ce cas, la valeur $(z_j - c_j)$ correspondant à la colonne pivot devient positive (5). On procède par conséquent à une itération de la méthode du simplexe avec ce pivot, en faisant entrer π_2 dans la base à la place de $\mathbf{p}_3$. On obtient ainsi le tableau (3.8) après la première itération.

Dans ce nouveau tableau, la recherche du pivot classique fournit (2/3*). En additionnant deux fois cette valeur au $(z_j - c_j)$ correspondant, on trouve une valeur positive (2/3). On effectue donc une itération directement avec ce pivot. Le résultat est indiqué dans le tableau (3.9) qui correspond à une solution optimale. Dans cet exemple, la deuxième partie de l'algorithme (introduction des vecteurs $\mathbf{p}_i$ ou $\mathbf{n}_i$ dans la base) n'est pas nécéssaire puisqu'à la fin de la première partie on trouve une solution optimale. A noter que la solution optimale trouvée correspond à la deuxième droite L_1 extrême, c'est-à-dire : $y = 1/2 + 1/2x$.

3.4 Algorithmes de descente

Contrairement à la méthode basée sur l'algorithme du simplexe, le problème d'estimation L_1 peut être résolu par une méthode dite de descente. Ce terme provient du fait que la fonction objectif à minimiser est formée d'arêtes. La méthode consiste alors à descendre le long de ces arêtes pour finalement atteindre le minimum. Dans cette section, deux algorithmes seront présentés. L'un permettant de résoudre le problème de régression linéaire simple et l'autre celui de la régression linéaire multiple. Ces deux algorithmes sont dus à Wesolowsky (1981). Comme pour les algorithmes basés sur la programmation linéaire, les algorithmes de descente sont nombreux mais pas tous efficaces du point de vue du temps de calcul. Les deux algorithmes présentés ici ont l'avantage d'être relativement rapides et simples puisque les règles de "descente" se ramènent à une optimisation univariée dont nous avons déjà parlé dans l'historique de l'estimation L_1.

Voyons à présent l'algorithme de descente permettant de résoudre le

base	$\mathbf{R_B}$	$\mathbf{p_1}$	$\mathbf{n_5}$
$\mathbf{n_2}$	1/4	3/4	−1/4
$\boldsymbol{\pi_1}$	3/4	5/4	1/4
$\mathbf{p_3}$	1/2	−1/2	1/2
$\mathbf{p_4}$	5/4	−1/4	3/4
$\boldsymbol{\pi_2}$	1/4	−1/4	−1/4
$z_j - c_j$	-2	1	0

Tableau 3.6: Tableau du simplexe après la troisième et dernière itération (solution optimale).

base	$\mathbf{R_B}$	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$
$\mathbf{p_1}$	1	1	1
$\mathbf{p_2}$	1	1	2**
$\mathbf{p_3}$	2	1	3***
$\mathbf{p_4}$	3	1	4
$\mathbf{p_5}$	2	1	5*
$z_j - c_j$	-9	-5	-15

Tableau 3.7: Tableau initial du simplexe (*,**,*** éléments pivots).

base	$\mathbf{R_B}$	$\boldsymbol{\pi_1}$	$\mathbf{p_3}$
$\mathbf{p_1}$	1/3	2/3*	-1/3
$\mathbf{n_2}$	1/3	-1/3	2/3
$\boldsymbol{\pi_2}$	2/3	1/3	1/3
$\mathbf{p_4}$	1/3	-1/3	-4/3
$\mathbf{n_5}$	4/3	2/3	5/3
$z_j - c_j$	-7/3	-2/3	1/3

Tableau 3.8: Tableau du simplexe après une itération (* pivot).

base	R_B	p_1	p_3
π_1	1/2	3/2	-1/2
n_2	1/2	1/2	1/2
π_2	1/2	-1/2	1/2
p_4	1/2	1/2	-3/2
n_5	1	-1	2
$z_j - c_j$	-2	1	0

Tableau 3.9: Tableau du simplexe après deux itérations (solution optimale).

problème de l'estimation L_1 dans le cas du modèle de régression linéaire simple. Le problème d'optimisation est le suivant

$$\text{minimiser} \sum_{i=1}^{n} \mid y_i - \beta_0 - \beta_1 x_i \mid$$

Considérons l'un des termes de cette somme : $z_j = \mid y_j - \beta_0 - \beta_1 x_j \mid$. Dans le plan (β_0, β_1), on a

$$z_j = y_j - \beta_0 - \beta_1 x_j \text{ si } \beta_0 \leq y_j - \beta_1 x_j$$
$$\text{et } z_j = -(y_j - \beta_0 - \beta_1 x_j) \text{ si } \beta_0 \geq y_j - \beta_1 x_j$$

Il s'agit donc de deux plans se rejoignant le long d'une arête dont la projection est $\beta_0 = y_j - \beta_1 x_j$. Ainsi un point (x_j, y_j) correspond à une arête. La surface est donc formée de morceaux de plans avec des arêtes dont la projection sur le plan (β_0, β_1) sont les droites $\beta_0 = y_j - \beta_1 x_j$. Comme la fonction objectif est convexe, le minimum doit se trouver non seulement sur une arête mais à l'intersection d'au moins deux arêtes. Si l'on peut trouver le point minimum le long de chaque arête, on peut se déplacer le long des arêtes pour atteindre le point minimum (qui n'est pas forcément unique) qui est garanti par la convexité de la fonction objectif.

Il reste à trouver le minimum le long d'une arête. Supposons qu'une arête est définie par le point $(x_j; y_j)$. Alors

$$\beta_0 = y_j - \beta_1 x_j$$
$$\text{ou } \beta_1 = y_j / x_j - \beta_0 / x_j$$

En substituant cette dernière expression dans la fonction objectif à minimiser, le problème devient

$$\text{minimiser} \sum_{x_i \neq x_j}^{n} \mid 1 - x_i/x_j \mid \left| \frac{y_i - y_j x_i/x_j}{1 - x_i/x_j} - \beta_0 \right| \tag{3.1}$$

Ce problème est connu puisqu'il s'agit de la recherche de la médiane pondérée. En effet, il s'agit de résoudre un problème du type

$$\text{minimiser} \sum_{i=1}^{n} w_i \mid a_i - t \mid$$

par rapport à t, où a_i sont des constantes positives ou négatives et w_i des constantes positives. Rappelons que la solution de ce problème, notée $\widehat{t}$, s'obtient en ordonnant les constantes $a_{(i)}$ en ordre croissant. Les poids $w_{(i)}$ représentent les poids ordonnés correspondants. On trouve $\widehat{t} = a_{(j)}$, où j est l'indice pour lequel les deux inégalités suivantes sont respectées

$$\sum_{i=1}^{j} w_{(i)} \geq \sum_{i=j+1}^{n} w_{(i)} \text{ et } \sum_{i=1}^{j-1} w_{(i)} < \sum_{i=j}^{n} w_{(i)}$$

L'algorithme permettant d'obtenir $\widehat{\beta}_0$ et $\widehat{\beta}_1$ est décrit ci-dessous.

Etape 1: Poser $k = 1$. Choisir les valeurs initiales $\beta_0^{(1)}$ et $\beta_1^{(1)}$. Un choix possible est la valeur des paramètres obtenus par la méthode des moindres carrés. Choisir l'indice j pour lequel $\mid y_j - \beta_0 - \beta_1 x_j \mid$ est minimum.

Etape 2: Poser $k = k + 1$. Résoudre le problème (3.1) de médiane pondérée en retenant l'indice i pour lequel le terme $(y_i - y_j x_i/x_j)/(1 - x_i/x_j)$ est la médiane pondérée.

Etape 3: a) Si $\beta_0^{(k)} - \beta_0^{(k-1)} = 0$ et $k \geq 3$, passer à l'étape 4. Sinon poser $j = i$ et retourner à l'étape 2.

b) Si $\beta_0^{(k)} - \beta_0^{(k-1)} \neq 0$, poser $j = i$ et retourner à l'étape 2.

Etape 4: Les estimateurs L_1 sont donnés par

$$\widehat{\beta}_0 = \beta_0^{(k)} \text{ et } \widehat{\beta}_1 = y_j/x_j - \beta_0^{(k)}/x_j$$

Voyons à présent l'algorithme de descente dans le cas de la régression multiple. Le problème est le suivant

$$\text{minimiser} \sum_{i=1}^{n} \mid y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi} \mid \qquad (3.2)$$

où $x_{1i} = 1, \forall i$. Considérons un ensemble de p points $(x_{1j}^I, \ldots, x_{pj}^I, y_j^I)$ choisis de telle manière que le système d'équations

$$\begin{aligned} y_1^I &= \beta_1 x_{11}^I + \beta_2 x_{21}^I + \ldots + \beta_p x_{p1}^I \\ &\vdots \\ y_p^I &= \beta_1 x_{1p}^I + \beta_2 x_{2p}^I + \ldots + \beta_p x_{pp}^I \end{aligned}$$

possède une solution unique. Ici, une "arête" est formée de tout sous-ensemble J formant $p-1$ équations. Pour obtenir une méthode d'optimisation le long d'une arête, posons

$$\mathbf{B} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{Y}_I = \begin{pmatrix} y_1^I \\ \vdots \\ y_p^I \end{pmatrix}, \mathbf{X}_I = \begin{pmatrix} x_{11}^I & \ldots & x_{p1}^I \\ \vdots & \vdots & \vdots \\ x_{1p}^I & \ldots & x_{pp}^I \end{pmatrix}$$

$$\text{et } \mathbf{Y}_J = \begin{pmatrix} y_1^J \\ \vdots \\ y_{p-1}^J \end{pmatrix}, \mathbf{X}_J = \begin{pmatrix} x_{11}^J & \ldots & x_{p1}^J \\ \vdots & \vdots & \vdots \\ x_{1,p-1}^J & \ldots & x_{p,p-1}^J \end{pmatrix}$$

Soit $\mathbf{B}_\theta$ le vecteur $\mathbf{B}$ dans lequel β_θ a été supprimé, $\mathbf{X}_\theta$ la colonne θ dans $\mathbf{X}_j$ et $\mathbf{X}_{j\theta}$ la nouvelle matrice obtenue de $\mathbf{X}_j$ en supprimant $\mathbf{X}_\theta$. Dans ce cas, pour β_θ donné, on a

$$\mathbf{B}_\theta = \mathbf{X}_{j\theta}^{-1} \mathbf{Y}_J - \beta_\theta \mathbf{X}_{j\theta}^{-1} \mathbf{X}_\theta.$$

Notons les éléments de $\mathbf{B}_\theta$ par

$$\beta_m = r_m - s_m \beta_\theta \ , \ m \neq p \qquad (3.3)$$

En substituant (3.3) dans (3.2), on obtient le problème suivant

$$\text{minimiser} \sum_{i=1}^{n} \mid x_{\theta i} - \sum_{m \neq \theta}^{p} s_m x_{mi} \mid \left| \frac{y_i - \sum_{m \neq \theta}^{p} r_m x_{mi}}{x_{\theta i} - \sum_{m \neq \theta}^{p} s_m x_{mi}} - \beta_1 \right| \qquad (3.4)$$

$$\text{avec } x_{\theta i} \neq \sum_{m \neq \theta}^{p} s_m x_{mi}.$$

Le problème (3.4) n'est rien d'autre qu'un problème de médiane pondérée! Celle-ci est obtenue pour l'un des indices i dans (3.4). Si cet indice n'apparaît pas dans l'ensemble des indices J, alors l'équation correspondante entre dans le système d'équations tandis que l'une des équations est éliminée. On obtient ainsi une nouvelle arête. A noter qu'il y a $(p-1)$ arêtes possibles et l'on choisit celle qui améliore le plus la fonction objectif. Le minimum est atteint lorsqu'aucune arête ne permet d'améliorer la fonction objectif.

Pour terminer cette section sur les algorithmes de descente, nous décrivons l'algorithme permettant de résoudre le problème de l'estimation L_1 dans le cas de la régression multiple.

Etape 1: Poser $k = 1$ et $l = 0$. Choisir $\beta_1, \beta_2, \ldots, \beta_p$ par exemple par la méthode des moindres carrés. Soit $I^{(1)} = \{j_1^{(1)}, \ldots, j_p^{(1)}\}$ un ensemble de p indices correspondants à p points choisis séquentiellement de la manière suivante. Le point avec le plus petit résidu au carré est choisi de façon à ce que

$$\mathbf{X}_I^{(1)} = \begin{pmatrix} 1 & x_{2j_1}^{(1)} & \ldots & x_{pj_1}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2j_p}^{(1)} & \ldots & x_{pj_p}^{(1)} \end{pmatrix}$$

est non-singulière. Trouver $\beta_m^{(1)}$ pour $m = 1, \ldots, p$ en résolvant $\mathbf{Y}_I^{(1)} = \mathbf{X}_I^{(1)}\mathbf{B}$. Poser $I^{(k)} = \{j_1^{(k)}, \ldots, j_p^{(k)}\}$ et $J = \{j_2^{(k)}, \ldots, j_p^{(k)}\}$.

Etape 2: Poser $k = k+1$. Obtenir β_θ (avec le plus petit θ) en résolvant (3.4). Poser $\beta_\theta^{(k)} = \beta_\theta$. Soit i l'indice qui définit la médiane pondérée.

Etape 3: a) Si $\beta_\theta^{(k)} - \beta_\theta^{(k-1)} = 0$ et $l > p$, passer à l'étape 4. Sinon

poser $I = \{j_2^{(k-1)}, \ldots, j_p^{(k-1)}, i\}$, $l = l + 1$ et $\beta_m^{(k)} = \beta_m^{(k-1)}$ pour tout m; passer à l'étape 2.

b) Si $\beta_\theta^{(k)} - \beta_\theta^{(k-1)} \neq 0$, poser $l = 0$. Calculer $\beta_m^{(k)}$ pour $m \neq \theta$. Poser $I^{(k)} = \{j_2^{(k-1)}, \ldots, j_p^{(k-1)}, i\}$ et retourner à l'étape 2.

Etape 4: Calculer $\beta_m^{(k)}$ pour $m \neq \theta$. Les estimateurs L_1 s'obtiennent en posant $\widehat{\beta}_1 = \beta_1^{(k)}, \widehat{\beta}_2 = \beta_2^{(k)}, \ldots, \widehat{\beta}_p = \beta_p^{(k)}$.

3.5 Choix de l'algorithme le plus rapide

Comme le montre la section précédente, de nombreux algorithmes ont été développés pour résoudre le problème de l'estimation L_1 en régression linéaire. Cependant, certains algorithmes sont plus efficaces que d'autres du point de vue du temps de calcul. Des études ont été menées pour déterminer l'algorithme le plus rapide, notamment par Dielman et Pfaffenberger (1984), Gentle, Narula et Sposito (1987) et Müller (1992) qui confirme les résultats obtenus dans l'étude de Dielman et Pfaffenberger. Dans ces différentes études, la distinction entre les problèmes de régression linéaire simple et multiple a été faite. En effet, certains algorithmes n'ont été élaborés que pour résoudre les problèmes de régression linéaire simple.

Dans l'étude de Dielman et Pfaffenberger, cinq algorithmes ont été retenus pour le cas de la régression simple. Il s'agit des algorithmes de Barrodale et Roberts (1974), Armstrong et Kung (1978), Abdelmalek (1980a), Wesolowsky (1981) et Klingman et Mote (1982). Ces algorithmes sont notés respectivement : (BR), (AK), (A), (W1) et (KM). L'efficacité du point de vue du temps de calcul est mesurée par le rapport du temps CPU total de deux algorithmes. Ainsi, l'algorithme (A) est 86.7% aussi efficace que celui de (BR), ceci en résolvant 7 problèmes différents et en faisant varier le nombre n d'observations. L'algorithme (W1) a été comparé à celui de (AK) en générant aléatoirement 25 problèmes de taille différente. L'efficacité de l'algorithme (AK) par rapport à l'algorithme (W1) dépend fortement du nombre d'observations. Les résultats figurent dans le tableau (3.10).
Ainsi l'algorithme (W1) devient plus efficace que (AK) lorsque le nombre d'observations augmente. Les algorithmes (KM) et (AK) ont aussi été comparés en fonction du nombre d'observations. L'efficacité de (AK)

n=10	n=250	n=500	n=750	n=1000
174%	46.6%	38.7%	27.8%	24.5%

Tableau 3.10: Efficacité de l'algorithme (AK) par rapport à (W1).

par rapport à (KM) pour différentes valeurs de n est donnée dans le tableau (3.11).

n=50	n=100	n=250	n=500
160%	69.6%	69.7%	50.7%

Tableau 3.11: Efficacité de l'algorithme (AK) par rapport à (KM).

En résumé, Dielman et Pfaffenberger recommandent l'algorithme (W1) lorsque $n \leq 50$. Lorsque $n > 50$, ils préconisent soit l'algorithme (AK) soit (KM). Ils précisent que dans tous les cas, les temps pour obtenir les paramètres estimés sont très raisonnables (moins de 2 secondes dans la plupart des cas).

Pour le cas de la régression multiple, les quatre algorithmes suivants ont été retenus parce qu'ils ne posaient pas de problème de convergence et que le temps CPU pour obtenir les paramètres estimés était raisonnable; il s'agit des algorithmes de Barrodale et Roberts (1974), Armstrong, Frome et Kung (1979), Bloomfield et Steiger (1980) et Wesolowsky (1981), les deux derniers étant des algorithmes de descente. Ces algorithmes sont notés respectivement : (BR), (AFK), (BS) et (W2). L'algorithme (AFK) a été comparé à celui de (BR) pour différentes valeurs de n et en faisant varier le nombre p de paramètres. Le tableau (3.12) donne l'efficacité de l'algorithme (BR) par rapport à celui de (AFK) selon le nombre de paramètres à estimer.

Il semble donc que l'algorithme (AFK) soit plus efficace (environ deux fois plus rapide) lorsque le nombre d'observations est relativement élevé même si l'efficacité est moins importante lorsque le nombre de paramètres augmente. L'algorithme (W2) a été comparé à celui de

p=5	p=10	p=15	p=20
41.3%	42.5%	42.7%	47.9%

Tableau 3.12: Efficacité de l'algorithme (BR) par rapport à (AFK).

(AFK) pour différentes valeurs de n et de p. Pour chaque comparaison, 25 problèmes ont été générés aléatoirement et le temps moyen CPU calculé. L'efficacité de l'algorithme (AFK) par rapport à (W2) figure dans le tableau (3.13) en fonction du nombre de paramètres à estimer.

p=3	p=6	p=10	p=20	p=25
67.2%	111.9%	160.6%	289.7%	300.5%

Tableau 3.13: Efficacité de l'algorithme (AFK) par rapport à (W2).

L'algorithme (W2) semble plus rapide que celui de (AFK) jusqu'à $p = 4$, ensuite (AFK) l'emporte facilement sur (W2). Finalement, l'algorithme (BS) a été comparé à celui de (BR). Pour chaque combinaison de n et p, dix ensembles de données ont été générés. L'efficacité de l'algorithme (BR) par rapport à (BS) a été calculée en fonction du nombre d'observations n et figure dans le tableau (3.14).

n=100	n=300	n=600	n=1200	n=1800
119%	112%	90%	68%	61%

Tableau 3.14: Efficacité de l'algorithme (BR) par rapport à (BS).

L'algorithme (BS) est clairement plus rapide que (BR) pour des valeurs de n relativement grandes ($n > 1000$). Puisque la comparaison de l'algorithme (AFK) et (BR) ne s'est faite que pour $p = 5, 10, 15$ et 20, il n'est pas possible de comparer l'efficacité de (AFK) à celle de (BS). Cependant, puisque (AFK) résout les problèmes L_1 environ deux fois plus vite que (BR), il est raisonnable de supposer que (AFK) est plus rapide que (BS) pour n relativement petit ($n \leq 200$) et aussi rapide que (BS) pour $n > 200$. En résumé, l'algorithme (W2) semble être le plus performant si $p \leq 4$. Si le nombre d'observations est très grand ($n \geq 2000$), l'algorithme (BS) peut être préférable. L'algorithme (AFK) est certainement le plus efficace lorsque $p \geq 5$. Dans l'étude de Gentle, Narula et Sposito (1987), les résultats comparatifs sont semblables. Ils indiquent que les algorithmes spécialement conçus pour résoudre les problèmes de régression linéaire simple n'offrent pas d'avantages supplémentaires par rapport à ceux développés pour la régression multiple. Ils mentionnent également que pour des problèmes dans lesquels le nombre d'observations est relativement faible, le choix devrait se faire

entre les algorithmes (AFK) et (BS). A noter que l'algorithme (W2) ne faisait pas partie de l'étude. Ils concluent en indiquant que l'algorithme (AFK) semble être le meilleur.

3.6 Conclusion

Les méthodes de calcul des estimateurs L_1 ont été présentées dans ce chapitre. Certains algorithmes comme celui de Barrodale et Roberts permettent de résoudre efficacement ce type de problèmes. Il nous a paru important de faire figurer cet algorithme dans ce chapitre puisqu'il est encore largement utilisé dans les logiciels statistiques les plus répandus. Outre les algorithmes basés sur la programmation linéaire, il existe une deuxième catégorie d'algorithmes basés sur les méthodes de descente. Nous avons retenu l'algorithme de Wesolowsky, l'un des plus rapides de sa catégorie. Finalement, la dernière partie de ce chapitre a été consacrée au choix de l'algorithme le plus rapide en indiquant les études réalisées pour comparer l'efficacité des algorithmes les plus importants. Bien que certains algorithmes soient plus performants que d'autres, leur efficacité dépend du nombre n d'observations ainsi que du nombre p de paramètres à estimer. Il semble qu'aucun algorithme soit le plus efficace dans tous les cas. L'extraordinaire développement de l'informatique (vitesse et puissance de calcul) permet aujourd'hui de résoudre le problème de l'estimation L_1 rapidement. D'un point de vue pratique, l'obtention des paramètres estimés par la méthode L_1 est presque aussi rapide que par celle des moindres carrés (pas de temps d'attente pour l'utilisateur). Si le problème du temps de calcul était primordial il y a quelques années, il tend aujourd'hui à s'estomper du fait de la vitesse accrue des ordinateurs actuels, même si dans certains cas comme la simulation, la rapidité de l'algorithme reste très importante.

CHAPITRE 4

COMPARAISON DES DIFFERENTS ESTIMATEURS

4.1 Introduction

Ce chapitre est consacré à l'étude de l'impact des deux problèmes de la multicolinéarité et des données aberrantes sur différents estimateurs. Nous évoquerons dans un premier temps les problèmes liés aux données aberrantes puis ceux liés à la multicolinéarité. Comme l'estimateur L_1 est une alternative possible à l'estimateur L_2 en présence de données aberrantes et l'estimateur ridge une alternative en présence du problème de la multicolinéarité, il est naturel d'espérer qu'une combinaison des estimateurs L_1 et ridge se comporte mieux lorsque les deux problèmes surviennent simultanément. Nous étudierons donc un cas particulier des ridge M-estimateurs pour construire un estimateur L_1-ridge. Le paragraphe suivant sera consacré aux propriétés de ce nouvel estimateur.

Notre intérêt portant essentiellement sur les paramètres estimés par différentes méthodes, nous discuterons des critères de comparaison de ces estimateurs. Grâce à la simulation qui nous permet de connaître les vraies valeurs des paramètres, nous pourrons étudier le comportement des estimateurs L_1, L_2 et ridge en présence du problème des données aberrantes. Pour simuler les données aberrantes, nous générerons les erreurs selon différentes lois dont on connait leur capacité à produire

des données aberrantes. Il s'agit en fait de distributions non normales des erreurs comme la distribution de Laplace ou celle de Cauchy. Nous étudierons ensuite le comportement de ces estimateurs en présence du problème de la multicolinéarité. Ici aussi, grâce à la simulation, il est possible de créer artificiellement ce problème en générant des variables explicatives dont la corrélation est connue à l'avance.

Une part importante de ce chapitre sera consacrée à l'étude de l'estimateur L_1-ridge. Nous avons procédé dans un premier temps à une étude préliminaire permettant de le comparer à certains estimateurs reconnus pour leur bon comportement en présence de l'un ou l'autre des problèmes envisagés. Finalement, dans une étude plus approfondie, le comportement de l'estimateur L_1-ridge est comparé à celui des estimateurs L_1, L_2 et ridge pour différents degrés de multicolinéarité et différentes distributions d'erreurs.

4.2 Problèmes liés aux données aberrantes

Une hypothèse importante faite en analyse de régression est que le modèle est approprié pour toutes les observations. Or, dans la pratique, il n'est pas rare qu'une (ou plusieurs) observation(s) possèdent une valeur de la variable dépendante qui ne corresponde pas au modèle adapté à la plupart des observations. On parle dans ce cas de données aberrantes qui peuvent se caractériser par des résidus qui sont, en valeur absolue, considérablement plus grands que les autres. Les problèmes liés aux données aberrantes dépendent également de la méthode d'estimation utilisée. Comme nous l'avons déjà mentionné, la méthode la plus couramment utilisée est celle des moindres carrés. Cette méthode peut s'avérer particulièrement sensible aux données aberrantes. L'effet de ces données aberrantes sur le modèle de régression peut se vérifier facilement en retirant ces observations et en recalculant la nouvelle équation de régression. Cette étude porte essentiellement sur l'effet que les données aberrantes (et la multicolinéarité) ont sur les paramètres estimés. Dans le cas de la méthode des moindres carrés, non seulement les paramètres peuvent être très sensibles aux données aberrantes, mais également les différentes statistiques de la régression linéaire telles que les statistiques t (test de Student) ou F (test de Fisher), le coefficient de détermination R^2 ou encore la variance estimée des résidus s^2. Pour illustrer les

problèmes que l'on peut rencontrer en présence de données aberrantes, examinons l'exemple suivant. Soit le modèle de régression linéaire simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, 10$. Fixons à l'avance les paramètres à estimer en posant $\beta_0 = \beta_1 = 1$ de manière à connaître les vraies valeurs des paramètres. Pour la variable indépendante, générons 10 valeurs selon une loi normale de moyenne 0 et d'écart-type 1. Avant de pouvoir calculer les valeurs de la variable dépendante, générons 10 valeurs pour le terme d'erreurs selon une loi reconnue pour produire des données aberrantes. Nous avons choisi ici celle de Cauchy standard. Les calculs ont été faits avec le logiciel S-Plus et les valeurs obtenues pour x_i, y_i et ε_i sont données dans le tableau (4.1).

i	x_i	y_i	ε_i
1	1.765	4.078	1.313
2	0.058	0.822	-0.236
3	0.269	1.607	0.338
4	1.383	2.244	-0.139
5	-0.814	-0.312	-0.498
6	-0.363	1.385	0.748
7	0.002	1.193	0.191
8	-0.524	2.982	2.506
9	0.334	1.642	0.308
10	-1.234	5.147	5.381

Tableau 4.1: Observations et résidus du problème de régression.

Le résidu correspondant à l'observation 10 est particulièrement grand comme celui de l'observation 8 (dans une moindre mesure). Voyons à présent les résultats obtenus avec un logiciel statistique (ici minitab), respectivement avec toutes les observations, sans l'observation 10 et sans les observations 8 et 10. Les nombres entre parenthèses dans le tableau (4.2) représentent les valeurs p correspondantes. Les résultats montrent combien les valeurs aberrantes peuvent influencer les différentes statistiques dans une analyse de régression. Avec toutes les observations, les paramètres estimés par la méthode des moindres carrés sont sensiblement différents des vrais paramètres. En effet, $\widehat{\beta}_0 = 2.064$ et $\widehat{\beta}_1 = 0.176$ alors que les vraies valeurs des paramètres sont $\beta_0 = \beta_1 = 1$, selon la construction adoptée ici.

	avec toutes les obs.	sans l'obs. 10	sans les obs. 8 et 10
$\hat{\beta}_0$	2.064	1.501	1.140
$\hat{\beta}_1$	0.176	1.014	1.348
t pour $\hat{\beta}_0$	3.83 (0.005)	4.36 (0.003)	5.54 (0.001)
t pour $\hat{\beta}_1$	0.29 (0.780)	2.46 (0.043)	5.67 (0.001)
F	0.08 (0.780)	6.06 (0.043)	32.18 (0.001)
R^2	1%	46.4 %	84.3 %
s	1.696	0.991	0.539

Tableau 4.2: Résultats obtenus avec la méthode des moindres carrés.

En retirant l'observation 10, on trouve $\widehat{\beta}_0 = 1.501$ et $\widehat{\beta}_1 = 1.014$ et en retirant les observations 8 et 10, on trouve $\widehat{\beta}_0 = 1.140$ et $\widehat{\beta}_1 = 1.348$, ce qui montre que les paramètres estimés se stabilisent autour des vraies valeurs lorsque les valeurs aberrantes sont supprimées. D'autre part, les valeurs des statistiques t sont particulièrement affectées par les observations aberrantes, notamment par l'observation 10. On voit par exemple que la valeur p pour tester l'hypothèse nulle $\beta_1 = 0$ passe de 0.780 à 0.043 lorsque l'observation 10 est éliminée. Ainsi, à un seuil de 5%, on ne peut pas rejeter l'hypothèse nulle lorsque l'observation 10 est présente alors qu'on peut la rejeter (et donc conclure que la pente de la droite, β_1, est significativement non nulle) lorsque l'observation 10 est éliminée. Dans cet exemple, l'effet le plus important des données aberrantes est l'impact sur le coefficient de détermination R^2. En effet, en présence des deux données aberrantes, la valeur de R^2 est seulement de 1%. En retirant l'observation 10, cette valeur passe à 46.4% et en retirant les deux observations 8 et 10 elle passe à 84.3%. Le coefficient de détermination peut donc s'avérer très sensible aux données aberrantes. Finalement l'écart-type estimé des résidus s est également influencé puisqu'il passe de 1.696 avec toutes les observations à 0.991 lorsque l'observation 10 est éliminée et à 0.539 lorsque les observations 8 et 10 sont supprimées. Dans cet exemple, s est environ trois fois plus élevé avec les observations aberrantes que sans. Il s'agit là d'un phénomène typique des données aberrantes qui ont tendance à "gonfler" la variance des résidus et par conséquent l'estimation de leur écart-type.

4.3 Problèmes liés à la multicolinéarité

Dans le modèle de régression linéaire multiple $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ où la matrice $\mathbf{X}$ est formée des p colonnes $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$, on définit la multicolinéarité en termes de dépendance linéaire des colonnes de $\mathbf{X}$. On dit que les vecteurs $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ sont linéairement dépendants s'il existe des constantes non toutes nulles $c_1, c_2, \ldots, c_p$ telles que

$$\sum_{j=1}^{p} c_j \mathbf{x_j} = 0 \tag{4.1}$$

Si la relation (4.1) se vérifie exactement pour un sous-ensemble des colonnes de $\mathbf{X}$, alors le rang de la matrice $\mathbf{X'X}$ est inférieur à p et l'inverse $(\mathbf{X'X})^{-1}$ n'existe pas. Cependant, dans la pratique la relation (4.1) est rarement satisfaite exactement mais elle peut l'être approximativement et on dit dans ce cas que le problème de multicolinéarité existe. A noter que si les variables explicatives et la variable dépendante sont standardisées, la matrice $\mathbf{X'X}$ est la $p \times p$ matrice des corrélations entre les variables explicatives et le $p \times 1$ vecteur $\mathbf{X'y}$, le vecteur des corrélations entre les variables explicatives et la variable dépendante.

Les problèmes de multicolinéarité peuvent provenir de différentes sources comme par exemple la manière de récolter les données, d'imposer des contraintes sur le modèle, de choisir certaines variables explicatives pour construire le modèle ou encore d'être en présence d'un modèle dans lequel il y a plus de variables que d'observations. La multicolinéarité peut avoir des effets importants sur les coefficients estimés par la méthode des moindres carrés. Pour s'en convaincre, voyons l'impact que peut avoir la multicolinéarité sur les coefficients estimés par la méthode des moindres carrés dans le cas où il y a deux variables explicatives $\mathbf{x}_1$ et $\mathbf{x}_2$ et une variable dépendante $\mathbf{y}$. On sait que l'estimateur des moindres carrés est donné par $\boldsymbol{\beta}_{L_2} = (\mathbf{X'X})^{-1}\mathbf{X'y}$. Lorsque les données sont standardisées, les paramètres estimés sont

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix} = \begin{pmatrix} \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^2)} \\ \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}^2)} \end{pmatrix}$$

où r_{12} est le coefficient de corrélation entre $\mathbf{x}_1$ et $\mathbf{x}_2$, r_{1y} le coefficient de

corrélation entre $\mathbf{x}_1$ et $\mathbf{y}$, r_{2y} le coefficient de corrélation entre $\mathbf{x}_2$ et $\mathbf{y}$. En posant

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{pmatrix}$$

on constate que lorsque la multicolinéarité entre $\mathbf{x}_1$ et $\mathbf{x}_2$ est importante, le coefficient de corrélation r_{12} devient proche de 1 (en valeur absolue) et la variance des estimateurs devient arbitrairement grande

$$V(\widehat{\beta}_j) = C_{jj}\sigma^2 \longrightarrow \infty \text{ lorsque } \mid r_{12} \mid \longrightarrow 1.$$

La covariance entre $\widehat{\beta}_1$ et $\widehat{\beta}_2$ devient également arbitrairement grande

$$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = C_{12}\sigma^2 \longrightarrow \pm\infty \text{ selon que } r_{12} \longrightarrow \pm 1.$$

Lorsque le nombre de variables est supérieur à deux, les effets de la multicolinéarité sont semblables. On peut montrer que les éléments diagonaux de la matrice $\mathbf{C}$ sont donnés par

$$C_{jj} = \frac{1}{1 - R_j^2} \ , j = 1, \ldots, p \tag{4.2}$$

où R_j^2 est le coefficient de détermination multiple obtenu en faisant jouer le rôle de la variable dépendante à $\mathbf{x}_j$ et en prenant les $(p-1)$ autres variables explicatives comme variables indépendantes. Ainsi lorsqu'il existe un degré de multicolinéarité important entre la variable $\mathbf{x}_j$ et un sous-ensemble des autres $(p-1)$ variables explicatives, alors la valeur de R_j^2 sera proche de 1 et donc C_{jj} deviendra arbitrairement grand. Comme la variance de $\widehat{\beta}_j$ est donnée par $V(\widehat{\beta}_j) = C_{jj}\sigma^2 = (1-R_j^2)^{-1}\sigma^2$, la variance des estimateurs des moindres carrés augmente avec la multicolinéarité. C'est la raison pour laquelle on appelle les éléments de (4.2) les *facteurs d'inflation de la variance.* Ils s'avèrent particulièrement utiles dans les techniques de diagnostics pour détecter la multicolinéarité. Il existe évidemment d'autres diagnostics pour détecter ce problème. L'une des mesures les plus simples consiste à inspecter les éléments r_{ij} hors de la diagonale principale de la matrice $\mathbf{X}'\mathbf{X}$. Ces termes r_{ij} représentent la corrélation entre les variables $\mathbf{x}_i$ et $\mathbf{x}_j$ et par conséquent $\mid r_{ij} \mid$ sera proche

de 1 si $\mathbf{x}_i$ et $\mathbf{x}_j$ sont presque linéairement dépendantes. Remarquons que l'inspection des coefficients de corrélation r_{ij} entre les variables explicatives ne permet de détecter que la dépendance linéaire entre paires de variables. Il se peut qu'aucun coefficient de corrélation $\mid r_{ij} \mid$ ne soit proche de 1 et que la multicolinéarité soit tout de même importante; cette situation se rencontre lorsqu'il y a plus de deux variables explicatives impliquées dans une dépendance linéaire presque exacte. Il n'est donc en général pas suffisant de ne s'intéresser qu'aux termes r_{ij} si l'on cherche à détecter le problème de la multicolinéarité.

Il existe de plus des techniques qui permettent de mesurer l'intensité du problème de la multicolinéarité dans les données. En ordonnant les valeurs propres de la matrice $\mathbf{X}'\mathbf{X}$ obtenues en résolvant l'équation caractéristique $\mid \mathbf{X}'\mathbf{X}-\lambda\mathbf{I} \mid =0$, on trouve $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$. Lorsqu'il y a une ou plusieurs dépendances presque exactes dans les données, alors une ou plusieurs valeurs propres sont petites. A l'inverse, une ou plusieurs petites valeurs propres impliquent certaines dépendances linéaires presque exactes parmi les colonnes de $\mathbf{X}$. On peut définir l'*indice de conditionnement* donné par

$$\kappa = \frac{\lambda_p}{\lambda_1} \tag{4.3}$$

qui est parfois défini dans la litérature par sa racine carrée ("condition number"), comme dans Weisberg (1985), Chatterjee et Hadi (1988) ou Belsley, Kuh and Welsh (1980). Ces derniers formulent en page 105 une appréciation quantitative du problème de multicolinéarité en fonction de la racine carrée de l'indice de conditionnement (4.3). Bien qu'en la matière il soit difficile d'être catégorique, ils ont mené une étude empirique permettant d'indiquer un faible problème de multicolinéarité lorsque cette valeur est de l'ordre de 5 à 10. Par contre, ils parlent de multicolinéarité modérée à forte pour des valeurs entre 30 et 100. Ainsi, avec la définition (4.3), une valeur de κ inférieure à 100 n'est pas indicatrice d'un grand problème de multicolinéarité. Lorsque la valeur de κ est plus grande que 100 mais inférieure à 900, le problème de multicolinéarité peut être considéré comme relativement modéré. Finalement, pour des valeurs de l'ordre de plusieurs milliers, le problème de multicolinéarité peut être considéré comme sévère.

4.4 Combinaison des estimateurs L_1 et ridge

Le but de ce travail est non seulement de comparer différents estimateurs entre eux, mais également de combiner les estimateurs les plus performants lorsque les deux problèmes des données aberrantes et de la multicolinéarité se posent simultanément. Dans les paragraphes consacrés à la comparaison des estimateurs L_1, L_2 et ridge présentés, nous avons pu constater que l'estimation L_1 est la méthode la plus performante lorsque la loi des erreurs n'est pas normale comme dans le cas de la loi de Laplace ou celle de Cauchy. D'autre part, quand la loi des erreurs est normale mais que le degré de multicolinéarité est élevé, l'estimation ridge est la plus performante. Il est donc naturel, pour combattre simultanément les deux problèmes évoqués ci-dessus, d'essayer de combiner les estimateurs L_1 et ridge qui se comportent bien en présence de chacun des problèmes pris séparément. Pour cela, nous introduisons tout d'abord les M-estimateurs puis leur combinaison avec les estimateurs ridge, comme introduits par Nyquist (1985). Le cas des estimateurs L_1-ridge a pu être traité sans passer par une méthode itérative, ce qui permet de calculer ces estimateurs rapidement. Huber (1973) a proposé les estimateurs du maximum de vraissemblance, appelés M-estimateurs et définis comme la solution du problème de minimisation suivant

$$\sum_{i=1}^{n} \rho((y_i - \mathbf{x}_i'\beta)/s)$$

où ρ est une fonction convexe des résidus à spécifier et s un paramètre d'échelle robuste estimé par

$$s = median \mid (y_i - \mathbf{x}_i'\beta_{L_1}) - median(y_i - \mathbf{x}_i'\beta_{L_1}) \mid /0.6745.$$

Le facteur $1/0.6745$ est introduit de manière à ce que s soit un estimateur consistant de σ lorsque les erreurs sont distribuées normalement selon $\mathcal{N}(0, \sigma^2)$. En général les M-estimateurs ne sont malheureusement pas invariants par rapport à l'échelle utilisée, d'où l'introduction du paramètre d'échelle s pour les rendre invariants. Remarquons que les estimateurs L_1 et L_2 sont des cas particuliers des M-estimateurs pour lesquels il n'est pas nécéssaire d'introduire le paramètre s. En effet, le choix $\rho(u) = |u|$ conduit aux estimateurs L_1 alors que le choix $\rho(u) = u^2$

fournit les estimateurs L_2. Ainsi dans ce qui suit, le cas général des M-estimateurs sera traité avec le paramètre d'échelle s, alors que pour les cas particuliers où $\rho(u) = |u|$ et $\rho(u) = u^2$, ce paramètre sera supprimé.

Pour trouver les M-estimateurs, il s'agit de résoudre le système d'équations suivant

$$\sum_{i=1}^{n} x_{ij}\psi((y_i - \mathbf{x}_i'\boldsymbol{\beta})/s) = 0 \ , \ j = 1, \ldots, p \tag{4.4}$$

où la fonction ψ est la dérivée de la fonction ρ. La résolution du système d'équations (4.4), qui n'est en général pas un système linéaire, se fait par un algorithme itératif, appelé technique itérative des moindres carrés pondérés. Nyquist (1985) utilise cette technique pour calculer les ridge M-estimateurs définis comme la solution du problème de minimisation

$$\sum_{i=1}^{n} \rho((y_i - \mathbf{x}_i'\boldsymbol{\beta})/s) + \sum_{j=1}^{p} \rho(\sqrt{k}\beta_j/s). \tag{4.5}$$

Lorsque ρ est dérivable, le système d'équations à résoudre est

$$-\sum_{i=1}^{n} x_{ij}\psi((y_i - \mathbf{x}_i'\boldsymbol{\beta})/s) + \sqrt{k}\psi(\sqrt{k}\beta_j)/s) = 0 \ , \ j = 1, \ldots, p. \tag{4.6}$$

Soit l'estimateur $\boldsymbol{\beta}^{(l)}$ obtenu à la l-ème itération. En se basant sur la technique itérative des moindres carrés pondérés utilisée par Holland et Welsch (1977), les poids $w_i^{(l)}$ pour les n observations et $v_j^{(l)}$ pour les p équations supplémentaires sont calculés par les deux expressions suivantes

$$w_i^{(l)} = \frac{\psi((y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(l)})/s)}{((y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(l)})/s)}$$

$$\text{et } v_j^{(l)} = \frac{\psi(\sqrt{k}\beta_j^{(l)}/s)}{(\beta_j^{(l)}/s)}$$

L'estimateur $\boldsymbol{\beta}^{(l+1)}$ à l'itération suivante est la solution du système d'équations linéaires

$$\sum_{i=1}^{n} w_i^{(l)} x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(l+1)}) + v_j^{(l)}\boldsymbol{\beta}^{(l+1)} = 0 \ , \ j = 1, \ldots, p.$$

Sous forme matricielle, ce nouvel estimateur se calcule par

$$\boldsymbol{\beta}^{(l+1)} = (\mathbf{X}'\mathbf{W}^{(l)}\mathbf{X} + \mathbf{V}^{(l)})^{-1}\mathbf{X}'\mathbf{W}^{(l)}\mathbf{y}$$

où $\mathbf{W}^{(l)}$ et $\mathbf{V}^{(l)}$ sont deux matrices diagonales d'ordre $(n \times n)$ et $(p \times p)$. Cet algorithme peut être utilisé dans les cas où $\rho(u) = |u|$ et $\rho(u) = u^2$. Notons qu'avec $\rho(u) = u^2$, on obtient l'estimateur ridge en résolvant l'équation (4.6). En effet, dans ce cas $\psi(u) = 2u$ et l'on obtient

$$-2\sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + 2k\beta_j = 0 \ , \ j = 1, \ldots, p$$

ou sous forme matricielle

$$\begin{gathered} -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2k\boldsymbol{\beta} \cdot \mathbf{I} = \mathbf{0} \\ -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2k\boldsymbol{\beta} \cdot \mathbf{I} = \mathbf{0} \\ (\mathbf{X}'\mathbf{X} + k\mathbf{I})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \\ \boldsymbol{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \end{gathered}$$

D'autre part, nous proposons d'étudier les estimateurs obtenus en introduisant $\rho(u) = |u|$ dans (4.5). Ces estimateurs seront appelés estimateurs L_1-ridge. Dans ce cas la technique itérative des moindres carrés pondérés peut présenter quelques difficultés, notamment dans le calcul des poids $w_i = \frac{1}{|r_i|}$, où r_i sont les résidus de l'estimation L_1. En effet, certains résidus étant nuls, il faut introduire un test permettant d'éviter la division par 0 et remplacer ces résidus nuls par des valeurs arbitrairement petites, choisies par l'utilisateur. De plus, il faut introduire un critère d'arrêt de l'algorithme lorsque les estimateurs obtenus à deux itérations successives sont suffisamment proches, laissant l'utilisateur devant un autre choix. Il est cependant possible d'éviter l'utilisation de cette technique en considérant les estimateurs L_1-ridge, définis par le problème de minimisation suivant

$$\sum_{i=1}^{n} \mid y_i - \mathbf{x}_i'\boldsymbol{\beta} \mid + \sqrt{k}\sum_{j=1}^{p} \mid \beta_j \mid \qquad (4.7)$$

et en définissant

$$\bar{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ et } \bar{\mathbf{X}} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \\ \sqrt{k} & 0 & \dots & 0 \\ 0 & \sqrt{k} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{k} \end{pmatrix} \text{ avec } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

le problème (4.7) revient au problème d'estimation L_1 qui consiste à minimiser

$$\sum_{i=1}^{n+p} | \bar{y}_i - \bar{\mathbf{x}}_i' \boldsymbol{\beta} |$$

sur les données augmentées $\bar{\mathbf{y}}$ et $\bar{\mathbf{X}}$. Cet estimateur sera noté $\boldsymbol{\beta}_{L_1-ridge}$.

4.5 Propriétés de l'estimateur L_1-ridge

Le mécanisme du comportement de l'estimateur L_1-ridge est loin d'être trivial et semble difficile à traiter algébriquement. Cela n'est pas surprenant puisqu'il n'existe pas de formule analytique pour l'estimateur L_1 dans un modèle de régression linéaire multiple. La programmation linéaire s'est avérée l'un des outils les plus performants pour calculer les estimateurs L_1. Dans ce sens l'estimateur L_1-ridge peut être calculé efficacement en ayant recours à la programmation linéaire paramétrique en introduisant le paramètre non-négatif $\lambda = \sqrt{k}$ dans la fonction (4.7). Cependant, en régression linéaire simple, des formules exactes peuvent être développées. La technique utilisée pour trouver ces formules est illustrée dans ce paragraphe. De plus un important résultat pour la régression linéaire multiple a pu être trouvé et démontré à la fin de ce paragraphe. Ce résultat donne une condition suffisante pour laquelle tous les paramètres sont nuls et trouve son application dans le Chapitre 5 consacré à la sélection de modèles.

Nous nous penchons tout d'abord sur le comportement mathématique de l'estimateur L_1-ridge dans le cas univarié. La fonction objectif à minimiser est donnée par

$$f(\beta_0) = \sum_{i=1}^{n} \mid y_i - \beta_0 \mid + \lambda \mid \beta_0 \mid \tag{4.8}$$

où $\lambda \geq 0$ est le paramètre qui joue le rôle de $\sqrt{k}$ dans l'estimation L_1-ridge. L'estimateur correspondant, noté $\widehat{\beta}_0(\lambda)$, est la valeur de β_0 qui minimise (4.8). Pour $\lambda = 0$, la solution est bien connue puisqu'elle se réduit à l'estimateur L_1, c'est-à-dire la médiane des observations y_i.

$$\widehat{\beta}_0(0) = \underset{i}{med}\, y_i.$$

Lorsque les observations ont été ordonnées en ordre croissant de manière à ce que $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, la solution est donnée par

$$\widehat{\beta}_0(0) = y_{(r)}$$

où $r = \dfrac{n+1}{2}$ lorsque n est impair (dans ce cas, la médiane est unique). Pour n pair, la solution n'est pas unique et chaque valeur $\widehat{\beta}_0(0)$ satisfaisant $y_{(r)} \leq \widehat{\beta}_0(0) \leq y_{(r+1)}$ avec $r = \dfrac{n}{2}$ minimisera (4.8). Dans ce cas, $y_{(r)}$ et $y_{(r+1)}$ sont respectivement appelées la médiane inférieure et supérieure.

Pour illustrer comment des formules exactes peuvent être obtenues pour l'estimateur L_1-ridge dans (4.8) en fonction de λ, nous supposerons n impair et les observations y_i strictement positives, c'est-à-dire $y_{(1)} > 0$. Les autres cas à considérer, notamment lorsque toutes les observations sont négatives ($y_{(n)} < 0$) et lorsque $y_{(1)} < 0$ et $y_{(n)} > 0$ peuvent être étudiés exactement de la même façon que ci-dessous.

La difficulté pour résoudre le problème de minimisation (4.8) provient de la non-différentiabilité de cette fonction aux points $\beta_0 = 0$ ainsi que $\beta_0 = y_i, i = 1, \ldots, n$. Cependant, $f(\beta_0)$ est une fonction continue, convexe et linéaire par morceaux. Elle atteint son minimum lorsque la pente de la tangente change de signe, passant d'une pente strictement négative à positive. Le problème revient ainsi à étudier cette fonction et sa pente sur les intervalles décrits dans le tableau (4.3).

Intervalle	Fonction $f(\beta_0)$	Pente $f'(\beta_0)$
$]-\infty;0[$	$\sum_{1}^{n} y_i - n\beta_0 - \lambda\beta_0$	$-n-\lambda$
$]0;y_{(1)}[$	$\sum_{1}^{n} y_i - n\beta_0 + \lambda\beta_0$	$-n+\lambda$
$]y_{(1)};y_{(2)}[$	$\sum_{2}^{n} y_{(i)} - \sum_{1}^{1} y_{(i)} + (2-n+\lambda)\beta_0$	$2-n+\lambda$
$]y_{(2)};y_{(3)}[$	$\sum_{3}^{n} y_{(i)} - \sum_{1}^{2} y_{(i)} + (4-n+\lambda)\beta_0$	$4-n+\lambda$
$]y_{(3)};y_{(4)}[$	$\sum_{4}^{n} y_{(i)} - \sum_{1}^{3} y_{(i)} + (6-n+\lambda)\beta_0$	$6-n+\lambda$
$\vdots$	$\vdots$	$\vdots$
$]y_{(r-2)};y_{(r-1)}[$	$\sum_{r-1}^{n} y_{(i)} - \sum_{1}^{r-2} y_{(i)} + (2r-4-n+\lambda)\beta_0$	$2r-4-n+\lambda$
		$=-3+\lambda$
$]y_{(r-1)};y_{(r)}[$	$\sum_{r}^{n} y_{(i)} - \sum_{1}^{r-1} y_{(i)} + (2r-2-n+\lambda)\beta_0$	$2r-2-n+\lambda$
		$=-1+\lambda$
$]y_{(r)};y_{(r+1)}[$	$\sum_{r+1}^{n} y_{(i)} - \sum_{1}^{r} y_{(i)} + (2r-n+\lambda)\beta_0$	$2r-n+\lambda$
		$=1+\lambda$
$]y_{(r+1)};y_{(r+2)}[$	$\sum_{r+2}^{n} y_{(i)} - \sum_{1}^{r+1} y_{(i)} + (2r+2-n+\lambda)\beta_0$	$2r+2-n+\lambda$
		$=3+\lambda$
$\vdots$	$\vdots$	$\vdots$
$]y_{(n)};\infty[$	$-\sum_{1}^{n} y_i + n\beta_0 + \lambda\beta_0$	$n+\lambda$

Tableau 4.3: $f(\beta_0)$ et $f'(\beta_0)$ sur les différents intervalles de λ.

L'estimateur L_1-ridge s'obtient en examinant le changement de signe de la pente. Une propriété importante de cet estimateur est qu'il reste constant sur chaque intervalle de λ indiqué dans le tableau (4.3). Il ne change que pour certaines valeurs de λ. Par exemple, l'estimateur L_1 qui correspond à la médiane des observations $y_{(r)}$ avec $r = (n+1)/2$, est obtenu pour toutes les valeurs de λ telles que $0 \leq \lambda < 1$. Ensuite, l'estimateur L_1-ridge change à $y_{(r-1)}$ et reste constant pour $1 \leq \lambda < 3$. Ce processus continue jusqu'à ce que l'estimateur L_1-ridge soit nul. Dans ce qui suit, les valeurs de λ en lesquelles l'estimateur L_1-ridge change seront appelées points de changement. Les points de changement et les estimateurs L_1-ridge correspondants sont donnés dans le tableau (4.4).

Points de changement λ	Estimateur L_1-ridge $\widehat{\beta}_0(\lambda)$
0	$y_{(r)}$
1	$y_{(r-1)}$
3	$y_{(r-2)}$
5	$y_{(r-3)}$
7	$y_{(r-4)}$
$\vdots$	$\vdots$
$n-4$	$y_{(2)}$
$n-2$	$y_{(1)}$
n	0

Tableau 4.4: Solutions pour les estimateurs L_1-ridge lorsque n est impair.

L'estimateur L_1-ridge dans ce cas est une fonction décroissante de λ avec des points de changement régulièrement espacés sur les entiers impairs (lorsque n est impair!). Une condition suffisante pour que l'estimateur L_1-ridge soit nul est que $\lambda > n$. Cette condition n'est cependant pas nécéssaire comme on peut facilement le vérifier en considérant des valeurs de y telles que $y_{(1)} < 0$ et $y_{(n)} > 0$. Finalement, le nombre maximal de points de changement dans (4.8) ne peut pas excéder $[n/2]+1$, une propriété qui ne se généralise pas à des modèles plus compliqués comme celui étudié ci-après.

Le problème d'estimation L_1-ridge correspondant à un modèle de

régression linéaire simple sans ordonnée à l'origine, est défini en minimisant la fonction objectif suivante

$$f(\beta_1) = \sum_{i=1}^{n} \mid y_i - \beta_1 x_i \mid + \lambda \mid \beta_1 \mid \tag{4.9}$$

Les solutions exactes des estimateurs L_1-ridge sont données dans le cas d'observations strictement positives (aussi bien pour y que pour x; les autres cas s'obtiennent de manière analogue). Pour cela nous supposerons, sans perte de généralité, que les observations ont été indicées dans l'ordre croissant des valeurs de $y_i/x_i = t_i$ de manière à ce que $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$. La fonction dans (4.9) est différentiable presque partout, à l'exception des points $\beta_1 = 0$ et $\beta_1 = t_{(i)}, i = 1, \ldots, n$. La même technique que précédemment peut être utilisée pour résoudre le problème, à savoir l'examen du signe de la pente de (4.9) dans les intervalles considérés dans le tableau (4.5). Introduisons encore les notations

$$\begin{aligned} S_k &= \sum_{i=1}^{k} x_{(i)} - \sum_{i=k+1}^{n} x_{(i)} \text{ pour } k = 1, \ldots, n-1 \\ \text{et } S_0 &= -\sum_{i=1}^{n} x_i \ , \ S_n = \sum_{i=1}^{n} x_i \end{aligned}$$

L'estimateur L_1, noté $\widehat{\beta}_1(0) = t_{(r)}$ pour l'indice $1 \leq r \leq n$ satisfait la double condition suivante $S_{r-1} < 0$ et $S_r \geq 0$. Les points de changement et les estimateurs L_1-ridge correspondants figurent dans le tableau (4.6). Dans ce cas, l'estimateur L_1-ridge est une fonction décroissante de λ. Cependant, la décroissance de cette fonction ne se généralise pas au cas de la fonction objectif relative au modèle de régression multiple. Le nombre maximal de points de changement peut cette fois être égal à n, le nombre d'observations. Finalement, une condition suffisante pour que l'estimateur soit nul dans ce cas est que $\lambda > \sum_{i=1}^{n} x_i$.

Lorsqu'on considère l'estimation L_1-ridge en régression linéaire simple avec ordonnée à l'origine, la fonction suivante doit être minimisée

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \mid y_i - \beta_0 - \beta_1 x_i \mid + \lambda \mid \beta_0 \mid + \lambda \mid \beta_1 \mid \tag{4.10}$$

Intervalle	Pente $f'(\beta_1)$
$]-\infty; 0[$	$-\sum_{i=1}^{n} x_i - \lambda$
$]0; t_{(1)}[$	$-\sum_{i=1}^{n} x_i + \lambda$
$]t_{(1)}; t_{(2)}[$	$\sum_{i=1}^{1} x_{(i)} - \sum_{i=2}^{n} x_{(i)} + \lambda$
$]t_{(2)}; t_{(3)}[$	$\sum_{i=1}^{2} x_{(i)} - \sum_{i=3}^{n} x_{(i)} + \lambda$
$]t_{(3)}; t_{(4)}[$	$\sum_{i=1}^{3} x_{(i)} - \sum_{i=4}^{n} x_{(i)} + \lambda$
$\vdots$	$\vdots$
$]t_{(k)}; t_{(k+1)}[$	$\sum_{i=1}^{k} x_{(i)} - \sum_{i=k+1}^{n} x_{(i)} + \lambda$
$\vdots$	$\vdots$
$]t_{(n-1)}; t_{(n)}[$	$\sum_{i=1}^{n-1} x_{(i)} - \sum_{i=n}^{n} x_{(i)} + \lambda$
$]t_{(n)}; \infty[$	$\sum_{i=1}^{n} x_i + \lambda$

Tableau 4.5: Pente $f'(\beta_1)$ sur les différents intervalles de λ.

Bien que les résultats obtenus pour les deux modèles précédents sont utiles dans ce problème et peuvent être appliqués dans certaines situations, le cas général est beaucoup plus ardu à résoudre mathématiquement. Tous les points de changement dans (4.10) n'ont pas pu être trouvés algébriquement, sauf dans quelques cas particuliers. On retrouve ici les mêmes difficultés rencontrées dans l'estimation L_1 pour trouver une formule analytique pour ces estimateurs. Par contre, il est encore possible de trouver une condition suffisante pour que les deux paramètres soient nuls. Ce résultat est démontré dans le théorème 1 ci-dessous.

Comme la démonstration de ce théorème repose sur plusieurs résultats relatifs à des inégalités (dont l'inégalité triangulaire), ces derniers sont donnés sous la forme de lemmes avant d'énoncer le théorème 1.

Points de changement λ	Estimateur L_1-ridge $\hat{\beta}_1(\lambda)$
0	$t_{(r)}$
$\mid S_{r-1} \mid = \sum_{i=r}^{n} x_{(i)} - \sum_{i=1}^{r-1} x_{(i)}$	$t_{(r-1)}$
$\mid S_{r-2} \mid = \sum_{i=r-1}^{n} x_{(i)} - \sum_{i=1}^{r-2} x_{(i)}$	$t_{(r-2)}$
$\mid S_{r-3} \mid = \sum_{i=r-2}^{n} x_{(i)} - \sum_{i=1}^{r-3} x_{(i)}$	$t_{(r-3)}$
$\vdots$	$\vdots$
$\mid S_2 \mid = \sum_{i=3}^{n} x_{(i)} - \sum_{i=1}^{2} x_{(i)}$	$t_{(2)}$
$\mid S_1 \mid = \sum_{i=2}^{n} x_{(i)} - \sum_{i=1}^{1} x_{(i)}$	$t_{(1)}$
$\mid S_0 \mid = \sum_{i=1}^{n} x_i$	0

Tableau 4.6: Solutions pour l'estimateur L_1-ridge en régression linéaire simple.

Lemme 1

$\mid a - b \mid \geq \mid a \mid - \mid b \mid, \forall a, b \in R.$

Ce résultat se vérifie facilement en considérant les quatre cas pour lesquels a et b sont positifs ou négatifs.

Le second lemme, plus connu sous le nom d'inégalité triangulaire généralisée, s'obtient par récurrence après avoir démontré l'inégalité triangulaire simple $\mid a + b \mid \leq \mid a \mid + \mid b \mid, \forall a, b \in R$.

Lemme 2

$\mid a_1 + a_2 + \ldots + a_n \mid \leq \mid a_1 \mid + \mid a_2 \mid + \ldots + \mid a_n \mid, \forall a_i \in R.$

Le troisième lemme est une généralisation du lemme 1.

Lemme 3

$\mid a_1 - a_2 - \ldots - a_n \mid \geq \mid a_1 \mid - \mid a_2 \mid - \ldots - \mid a_n \mid, \forall a_i \in R.$

Preuve : En appliquant d'abord le lemme 1 puis le lemme 2, on

obtient le résultat énoncé, à savoir

$$\begin{aligned} \mid a_1 - a_2 - \ldots - a_n \mid &= \mid a_1 - (a_2 + a_3 + \ldots + a_n) \mid \\ &\geq \mid a_1 \mid - \mid a_2 + a_3 + \ldots + a_n \mid \\ &\geq \mid a_1 \mid - (\mid a_2 \mid + \mid a_3 \mid + \ldots + \mid a_n \mid) \\ &= \mid a_1 \mid - \mid a_2 \mid - \mid a_3 \mid - \ldots - \mid a_n \mid . \end{aligned}$$

Théorème 1

Si $\lambda > max(n, \sum_{i=1}^{n} \mid x_i \mid)$, alors $\widehat{\beta}_0(\lambda) = \widehat{\beta}_1(\lambda) = 0$.

Démonstration

Ce résultat provient de la continuité de la fonction dans (4.10) et du fait que $f(\beta_0, \beta_1) > f(0,0), \forall \beta_0, \beta_1 \neq 0$ comme démontré ci-dessous en appliquant n fois le lemme 3 et l'hypothèse du théorème 1.

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^{n} \mid y_i - \beta_0 - \beta_1 x_i \mid + \lambda \mid \beta_0 \mid + \lambda \mid \beta_1 \mid \\ &\geq \sum_{i=1}^{n} \mid y_i \mid - n \mid \beta_0 \mid - \mid \beta_1 \mid \sum_{i=1}^{n} \mid x_i \mid + \lambda \mid \beta_0 \mid + \lambda \mid \beta_1 \mid \\ &= \sum_{i=1}^{n} \mid y_i \mid + \underbrace{(\lambda - n)}_{>0} \mid \beta_0 \mid + \underbrace{(\lambda - \sum_{i=1}^{n} \mid x_i \mid)}_{>0} \mid \beta_1 \mid \\ &> \sum_{i=1}^{n} \mid y_i \mid = f(0,0). \end{aligned}$$

Le théorème 1 combine les résultats obtenus pour les deux modèles de régression linéaire simple avec un paramètre étudiés au début de ce paragraphe. Ce résultat peut d'ailleurs se généraliser au cas d'un modèle de régression linéaire multiple avec ordonnée à l'origine dont la fonction objectif à minimiser est donnée dans (4.11).

$$f(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \mid y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} \mid + \lambda \sum_{j=0}^{p} \mid \beta_j \mid \quad (4.11)$$

La condition suffisante pour que tous les paramètres soient nuls dans

(4.11) est donnée dans le théorème 2.

Théorème 2

$$\lambda > max(n, \sum_{i=1}^{n} \mid x_{i1} \mid, \ldots, \sum_{i=1}^{n} \mid x_{ip} \mid) \Longrightarrow \widehat{\beta}_0(\lambda) = \ldots = \widehat{\beta}_p(\lambda) = 0.$$

Démonstration

En utilisant les mêmes arguments que dans la démonstration ci-dessus, on obtient le résultat

$$\begin{aligned} f(\beta_0, \beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} \mid y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} \mid + \lambda \sum_{j=0}^{p} \mid \beta_j \mid \\ &\geq \sum_{i=1}^{n} \mid y_i \mid - n \mid \beta_0 \mid - \mid \beta_1 \mid \sum_{i=1}^{n} \mid x_{i1} \mid - \\ &\quad - \ldots - \mid \beta_p \mid \sum_{i=1}^{n} \mid x_{ip} \mid + \lambda \sum_{j=0}^{p} \mid \beta_j \mid \\ &= \sum_{i=1}^{n} \mid y_i \mid + \underbrace{(\lambda - n)}_{>0} \mid \beta_0 \mid + \underbrace{(\lambda - \sum_{i=1}^{n} \mid x_{i1} \mid)}_{>0} \mid \beta_1 \mid + \\ &\quad + \ldots + \underbrace{(\lambda - \sum_{i=1}^{n} \mid x_{ip} \mid)}_{>0} \mid \beta_p \mid \\ &\geq \sum_{i=1}^{n} \mid y_i \mid = f(0, 0, \ldots, 0). \end{aligned}$$

la dernière inégalité étant stricte si tous les paramètres ne sont pas nuls.

4.6 Choix des critères de comparaison

De manière à pouvoir étudier le comportement des différents estimateurs que nous nous proposons d'examiner dans ce travail, nous aurons recours à la simulation. En effet, grâce à la simulation, les vraies valeurs des paramètres sont connues et nous pouvons donc les comparer aux paramètres estimés par différentes méthodes d'estimation. Les estimateurs ridge, contrairement aux estimateurs L_2, sont en général biaisés mais peuvent avoir une variance plus petite que celle des estimateurs L_2

lorsque la distribution des erreurs n'est pas normale. L'erreur quadratique moyenne (somme de la variance et du biais au carré) est une mesure particulièrement bien adaptée pour pouvoir comparer les différents estimateurs. Les rapports obtenus en divisant l'erreur quadratique moyenne de deux estimateurs permettent d'indiquer quel estimateur est le plus performant. En effet pour savoir si l'estimateur A est plus performant que l'estimateur B, on calcule pour les différents paramètres le rapport : $mse(A)/mse(B)$. Si ce rapport est inférieur à 1, l'estimateur A est le plus performant tandis qu'une valeur supérieure à 1 indique que l'estimateur B est le plus performant.

Mathématiquement, si l'on répète dans la simulation l fois l'expérience et que le paramètre estimé est noté $\widehat{\beta}$ et le vrai paramètre β, on calcule

$$\text{moyenne} : \bar{\beta} = \frac{1}{l}\sum_{i=1}^{l}\widehat{\beta}_i$$

$$\text{écart-type} : s(\widehat{\beta}) = \sqrt{\frac{\sum_{i=1}^{l}(\widehat{\beta}_i - \bar{\beta})^2}{l-1}}$$

$$\text{biais} : Biais(\widehat{\beta}) = \bar{\beta} - \beta$$

$$\text{erreur quadratique moyenne} : MSE(\widehat{\beta}) = s^2(\widehat{\beta}) + Biais^2(\widehat{\beta})$$

D'autres critères peuvent être adoptés pour comparer les différents estimateurs. Connaissant les vraies valeurs des paramètres, il est possible de mesurer si un estimateur est plus proche qu'un autre des vraies valeurs des paramètres. Pour cela les deux critères suivants ont été retenus

$$\text{Coefficient d'erreur absolue} : CE_1(\widehat{\beta}) = \sum_{j=0}^{p} \mid \widehat{\beta}_j - \beta_j \mid$$

$$\text{Coefficient d'erreur quadratique} : CE_2(\widehat{\beta}) = \sum_{j=0}^{p}(\widehat{\beta}_j - \beta_j)^2$$

La comparaison entre deux estimateurs A et B se fera en calculant le pourcentage de cas pour lesquels l'estimateur A a produit un coefficient d'erreur absolue (respectivement quadratique) plus petit que l'estimateur B. Une valeur de 50 % indique un comportement analogue des deux estimateurs. Une valeur supérieure à 50 % indique une meilleure performance de l'estimateur A et inversement.

Finalement, pour investiguer la stabilité des différents estimateurs nous avons retenu comme critère la variance de l'angle formé par l'hyperplan estimé et le vrai hyperplan.

4.7 Plan de simulation

Comme l'étude du comportement des différents estimateurs dans ce chapitre se fera sur la base de la simulation, il est nécéssaire de présenter l'expérience qui a été menée. Comme il s'agit à la fois de simuler des lois d'erreurs susceptibles de produire des données aberrantes et de faire varier le degré de multicolinéarité, le plan de simulation retenu est le suivant. Le modèle utilisé est donné par

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots, n$$

et les vraies valeurs des paramètres ont été fixées à 1 en donnant les valeurs $\beta_0 = \beta_1 = \ldots = \beta_p = 1$. Les variables explicatives sont générées de manière à ce que le coefficient de corrélation entre paires de variables soit égal à ρ pour pouvoir faire varier le degré de multicolinéarité en choisissant ρ entre 0 et 1

$$x_{ij} = \sqrt{1-\rho} z_{ij} + \sqrt{\rho} z_{ip+1} \text{ avec } i = 1, \ldots, n \text{ et } j = 1, 2, \ldots, p$$

les termes z_{ij} étant des nombres aléatoires distribués selon une loi normale $\mathcal{N}(0,1)$ dans le logiciel S-Plus version 3.2. Ainsi, les variables explicatives ont pour moyenne 0, pour variance 1 et la covariance entre paires de variables est ρ. Par conséquent, ρ correspond également au coefficient de corrélation entre paires de variables.

Une fois générées pour une taille d'échantillon n, les variables explicatives sont fixées durant l'expérience. La taille de l'échantillon n est le premier facteur de l'expérience. Nous avons retenu $n = 25$ et $n = 100$. Le second facteur est le nombre de variables explicatives.

Nous avons choisi $p = 2$, $p = 5$ et $p = 10$. Le troisième facteur ρ est particulièrement important puisqu'il permet de faire varier le degré de multicolinéarité. Pour couvrir les cas où la multicolinéarité est faible à modérée, dix valeurs de 0.09 à 0.99 ont été choisies par pas de 0.1; les cas où la multicolinéarité est forte à sévère sont étudiés en donnant à ρ quatre valeurs allant de 0.999 à 0.999999. Pour chaque valeur de ρ, l'indice de conditionnement κ a été calculé pour savoir quelle est l'importance du problème de la multicolinéarité. Le dernier facteur est la distribution des erreurs générée dans S-Plus. Il s'agit des distributions suivantes

a) standard normale $\mathcal{N}(0,1)$.

b) normale contaminée : $\mathcal{N}(0,1)$ et $\mathcal{N}(0,5)$ avec probabilité 0.85 et 0.15 respectivement.

c) standard Laplace : $\mathcal{L}(0,1)$.

d) standard Cauchy.

e) Exp(1)+ 0.3 $\mathcal{N}(0,1)$.

La première distribution $\mathcal{N}(0,1)$ est le cas classique de la normalité des erreurs. Elle permet de savoir si les résultats sont cohérents avec la théorie. Dans le cas où il n'y a pas de données aberrantes ni de problème de multicolinéarité, l'estimation L_2 est connue pour être la meilleure. Il est ainsi possible de savoir si les calculs de la simulation ont été correctement effectués. Les trois distributions b), c) et d) sont destinées à créer le problème des données aberrantes, plus particulièrement celle de Cauchy. La distribution de Laplace joue un rôle important étant donné que les estimateurs L_1 sont les estimateurs du maximum de vraissemblance lorsque les erreurs suivent cette loi. La dernière distribution, somme d'une exponentielle de paramètre 1 et d'une loi normale, a été choisie parce qu'elle n'est pas symétrique et que le comportement des différents estimateurs est peu connu lorsque la loi des erreurs n'est pas symétrique. Finalement, pour chaque traitement, l'expérience est répétée 500 fois. Le générateur de nombres aléatoires utilisé est celui implanté dans S-plus et décrit en pp. 124 et 125 de Venables et Ripley (1994). Le seed peut être choisi dans S-plus grâce à la commande set.seed. Pour refaire les expériences menées dans cette étude, il faut donner le seed en tapant set.seed(32). Cette commande est particulièrement utile puisqu'elle permet de sélectionner (pour tout nombre compris entre 1 et 1000) l'un des seeds présélectionnés qui assure une période d'environ 6.6×10^{14}.

4.8 Estimateurs L_1, L_2 et ridge en présence de données aberrantes

Pour étudier le comportement des estimateurs L_1, L_2 et ridge en présence de données aberrantes à partir de la simulation effectuée, nous nous intéressons au cas où il n'y a pas le problème de la multicolinéarité ($\rho = 0$). Pour $n = 25$ et trois modèles avec respectivement p=2, 5 et 10 variables explicatives, nous avons relevé dans le tableau (4.7) les valeurs des paramètres estimés par ces méthodes lorsque les erreurs suivent une loi de Cauchy, reconnue pour produire des données aberrantes de $\mathbf{y}$.

Au vu de ces résultats, on constate que l'estimation L_1 se comporte particulièrement bien avec des paramètres estimés proches de 1, indiquant une très faible influence des données aberrantes sur l'estimation de ces paramètres. Ces résultats illustrent parfaitement la robustesse des estimateurs L_1 face aux données aberrantes. Les résultats de la colonne correspondant à l'estimation L_2 montrent combien les paramètres estimés peuvent être différents des vraies valeurs des paramètres. Si dans le cas $p = 2$, la situation ne semble pas si mauvaise (ordonnée à l'origine proche de 1, $\widehat{\beta}_1$ légèrement surestimé et $\widehat{\beta}_2$ sous-estimé), l'estimation L_2 est loin d'être précise dans les cas où $p = 5$ et $p = 10$. Certains paramètres ayant même un signe erroné. Dans notre cas, pour $p = 5$ et $p = 10$, l'estimation L_2 est particulièrement sensible aux données aberrantes. La dernière colonne du tableau (4.7) fournit les résultats pour l'estimation ridge. Celle-ci n'étant pas spécialement destinée à être utilisée en présence de données aberrantes, elle ne semble pas mieux adaptée que l'estimation L_2. Ici également le cas $p = 2$ est trompeur, ne révélant pas une sensibilité très importante aux données aberrantes, contrairement aux cas $p = 5$ et $p = 10$. Finalement, nous présentons dans le tableau (4.8) les résultats obtenus lorsque la distribution des erreurs est normale. On constate que les deux méthodes d'estimation L_1 et L_2 fournissent des paramètres estimés très proches des vraies valeurs des paramètres, alors que ceux obtenus par l'estimation ridge sont pour la plupart sous-estimés. Ces résultats correspondent bien au fait que les estimateurs ridge présentent un biais et qu'ils ont tendance à être réduits vers 0.

Loi : Cauchy	Paramètres	Est. L_1	Est. L_2	Est. ridge
$p = 2$	β_0	1.00	1.03	1.07
	β_1	0.978	1.86	0.928
	β_2	0.959	0.605	0.542
$p = 5$	β_0	1.01	0.743	0.727
	β_1	0.992	2.41	1.49
	β_2	0.999	-0.616	0.179
	β_3	0.996	2.47	1.15
	β_4	1.03	-0.429	-0.495
	β_5	1.03	2.01	0.772
$p = 10$	β_0	0.959	-0.55	-0.194
	β_1	0.976	-1.14	-0.965
	β_2	1.09	1.90	0.898
	β_3	0.957	1.04	0.754
	β_4	1.07	0.299	0.697
	β_5	0.936	-1.23	-0.787
	β_6	1.04	3.17	2.02
	β_7	0.975	2.46	1.80
	β_8	0.979	3.36	1.99
	β_9	1.02	0.0997	0.71
	β_{10}	0.985	-2.47	-1.59

Tableau 4.7: Estimateurs L_1, L_2 et ridge pour $p = 2, 5, 10$ lorsque les erreurs sont distribuées selon une loi de Cauchy.

Loi : Normale	Paramètres	Est. L_1	Est. L_2	Est. ridge
$p = 2$	β_0	0.999	0.993	0.956
	β_1	1.01	1.00	0.960
	β_2	1.03	1.02	0.972
$p = 5$	β_0	0.994	0.998	0.970
	β_1	1.00	0.997	0.935
	β_2	1.03	1.03	0.992
	β_3	1.01	1.00	0.945
	β_4	0.979	0.978	0.934
	β_5	0.983	0.986	0.942
$p = 10$	β_0	0.992	0.999	0.963
	β_1	1.00	0.992	0.935
	β_2	1.03	1.03	1.03
	β_3	1.01	1.00	0.934
	β_4	0.972	0.976	0.904
	β_5	1.00	0.995	0.926
	β_6	0.985	0.981	0.943
	β_7	0.990	0.999	0.872
	β_8	1.01	1.00	0.907
	β_9	1.01	1.00	0.996
	β_{10}	0.993	1.01	0.863

Tableau 4.8: Estimateurs L_1, L_2 et ridge pour $p = 2, 5, 10$ lorsque les erreurs sont distribuées selon une loi Normale.

4.9 Estimateurs L_1, L_2 et ridge en présence de la multicolinéarité

A partir de la simulation effectuée, nous pouvons étudier le comportement des estimateurs L_1, L_2 et ridge en présence du problème de la multicolinéarité uniquement. Ainsi, nous nous intéressons à différentes valeurs de ρ lorsque la loi des erreurs est normale $\mathcal{N}(0,1)$. Pour voir comment se comportent ces trois estimateurs, nous les comparons du point de vue de l'erreur quadratique moyenne en calculant les rapports mentionnés dans le paragraphe relatif au choix des critères de comparaison. Le tableau (4.9) contient ces différents rapports pour chaque paramètre lorsque le problème de multicolinéarité est peu important avec $\rho = 0.09$, la taille de l'échantillon égale à 100 et le nombre de variables explicatives $p = 10$. La première colonne du tableau (4.9) permet de comparer les estimateurs L_1 et L_2. Une valeur des rapports supérieure à 1 indique que l'estimateur L_2 est plus performant. On constate que les rapports pour les différents paramètres sont semblables et tous plus grands que 1. L'estimation L_2 est donc plus performante que l'estimation L_1 dans ce cas. Ces résultats ne sont pas surprenants puisque les erreurs sont distribuées selon une loi normale et qu'il n'y a pas de gros problèmes de multicolinéarité. La seconde colonne permet de comparer les estimateurs L_1 et ridge. Ici aussi les rapports sont semblables et tous plus grands que 1, indiquant une meilleure performance de l'estimation ridge. Ces résultats nous confortent dans l'idée que les estimateurs ridge se comportent mieux que les estimateurs L_1 lorsqu'on est en présence de multicolinéarité (même faible comme ici) et sans problème de données aberrantes. Finalement la dernière colonne du tableau (4.9) permet de comparer les estimateurs L_2 et ridge. Tous les rapports sont très proches de 1 mais légèrement inférieurs à 1, favorisant donc un peu l'estimation ridge. Il est particulièrement intéressant d'étudier le comportement de ces rapports lorsque le degré de multicolinéarité augmente. Le tableau (4.10) fournit les résultats pour $\rho = 0.89$. La première colonne qui permet de comparer l'estimation L_1 et L_2 diffère peu du cas où $\rho = 0.09$. Les estimateurs L_2 se comportent donc toujours mieux que les estimateurs L_1 lorsque le degré de multicolinéarité augmente. La seconde colonne permettant de comparer les estimateurs L_1 et ridge contient des rapports plus grands que 1 et encore plus grands que dans le cas où $\rho = 0.09$.

Paramètres	$\frac{\text{mse}(L_1)}{\text{mse}(L_2)}$	$\frac{\text{mse}(L_1)}{\text{mse(ridge)}}$	$\frac{\text{mse(ridge)}}{\text{mse}(L_2)}$
β_0	1.592	1.621	0.982
β_1	1.634	1.683	0.971
β_2	1.600	1.640	0.976
β_3	1.515	1.554	0.975
β_4	1.510	1.550	0.974
β_5	1.565	1.595	0.981
β_6	1.675	1.705	0.976
β_7	1.643	1.684	0.976
β_8	1.590	1.630	0.975
β_9	1.590	1.630	0.975
β_{10}	1.630	1.670	0.976

Tableau 4.9: Rapports de l'erreur quadratique moyenne des différents estimateurs pour $\rho = 0.09$.

Paramètres	$\frac{\text{mse}(L_1)}{\text{mse}(L_2)}$	$\frac{\text{mse}(L_1)}{\text{mse(ridge)}}$	$\frac{\text{mse(ridge)}}{\text{mse}(L_2)}$
β_0	1.615	1.663	0.971
β_1	1.640	1.980	0.828
β_2	1.537	1.947	0.789
β_3	1.431	1.657	0.864
β_4	1.382	1.796	0.815
β_5	1.539	1.824	0.844
β_6	1.672	2.010	0.832
β_7	1.634	1.930	0.846
β_8	1.640	1.994	0.822
β_9	1.534	1.851	0.829
β_{10}	1.611	1.966	0.820

Tableau 4.10: Rapports de l'erreur quadratique moyenne des différents estimateurs pour $\rho = 0.89$.

Les estimateurs ridge se comportent donc encore mieux face aux estimateurs L_1 lorsque ρ augmente. De même, la troisième colonne indique des rapports inférieurs à 1 et plus petits que pour $\rho = 0.09$. Ainsi les estimateurs ridge se comportent aussi mieux face aux estimateurs L_2 à mesure que le problème de la multicolinéarité devient plus important. Une étude plus approfondie de ces rapports est nécéssaire pour comprendre le comportement de ces trois estimateurs. Pour savoir pour quelles valeurs de ρ le problème de la multicolinéarité devient important, nous avons calculé l'indice de conditionnement κ correspondant à chaque valeur de ρ examinée. Le tableau (4.11) contient les résultats pour $n = 100$ et $p = 10$, indiquant un problème de multicolinéarité peu important pour les valeurs de ρ inférieures à 0.99, puis nettement plus important jusqu'à $\rho = 0.9999$ et finalement très sévère pour les deux dernières valeurs de ρ. Nous avons vu dans les tableaux (4.9) et (4.10) que les rapports obtenus étaient très semblables pour les différents paramètres. Ainsi nous ne présentons dans le tableau (4.11) que les résultats pour l'un des paramètres (β_1). Nous pouvons ainsi comparer les trois méthodes d'estimation selon le degré de multicolinéarité lorsque les erreurs sont distribuées selon une loi normale. La comparaison entre les estimateurs L_1 et L_2 nous montre que les estimateurs L_2 sont toujours meilleurs que les estimateurs L_1 avec des rapports supérieurs à 1 pour toutes les valeurs de ρ envisagées. Cette supériorité est pratiquement constante pour tous les degrés de multicolinéarité. La distribution des erreurs étant normale, c'est logiquement les estimateurs L_2 qui l'emportent. De même, les estimateurs ridge sont meilleurs que les estimateurs L_1. Dans ce cas, les rapports ne sont pas constants mais augmentent avec la multicolinéarité, donnant un net avantage aux estimateurs ridge lorsque le problème de la multicolinéarité devient important. Finalement la dernière colonne du tableau (4.11) nous apprend que les estimateurs ridge, comparés aux estimateurs L_2, sont d'autant plus performants que le degré de multicolinéarité est élevé.

Voyons encore comment se comportent ces trois estimateurs lorsque la distribution des erreurs est censée favoriser l'estimation L_1, c'est-à-dire pour une loi de Laplace. Les résultats obtenus sont donnés dans le tableau (4.12). En comparant les estimateurs L_1 et L_2, on constate que les rapports sont tous inférieurs à 1, indiquant un meilleur comportement des estimateurs L_1 quelque soit le degré de multicolinéarité.

ρ	κ	$\frac{\text{mse}(L_1)}{\text{mse}(L_2)}$	$\frac{\text{mse}(L_1)}{\text{mse(ridge)}}$	$\frac{\text{mse(ridge)}}{\text{mse}(L_2)}$
0.09	2.05	1.634	1.683	0.971
0.19	2.65	1.627	1.676	0.971
0.29	3.26	1.602	1.660	0.965
0.39	3.93	1.618	1.676	0.965
0.49	4.72	1.615	1.692	0.955
0.59	5.69	1.637	1.735	0.944
0.69	7.0	1.637	1.765	0.928
0.79	9.04	1.634	1.812	0.902
0.89	13.2	1.640	1.980	0.828
0.99	46.0	1.629	3.984	0.409
0.999	146	1.600	6.400	0.250
0.9999	466	1.571	6.786	0.232
0.99999	1460	1.727	7.364	0.235
0.999999	4630	1.636	7.182	0.228

Tableau 4.11: Rapports obtenus avec une loi Normale.

ρ	κ	$\frac{\text{mse}(L_1)}{\text{mse}(L_2)}$	$\frac{\text{mse}(L_1)}{\text{mse(ridge)}}$	$\frac{\text{mse(ridge)}}{\text{mse}(L_2)}$
0.09	2.05	0.851	0.891	0.956
0.19	2.65	0.843	0.902	0.935
0.29	3.26	0.853	0.912	0.935
0.39	3.93	0.863	0.931	0.926
0.49	4.72	0.873	0.951	0.918
0.59	5.69	0.865	0.971	0.891
0.69	7.0	0.873	1.010	0.864
0.79	9.04	0.873	1.078	0.809
0.89	13.2	0.867	1.224	0.708
0.99	46.0	0.875	2.729	0.321
0.999	146	0.833	3.722	0.224
0.9999	466	0.889	3.778	0.235
0.99999	1460	0.875	3.75	0.233
0.999999	4630	0.875	3.875	0.226

Tableau 4.12: Rapports obtenus avec une loi de Laplace.

La comparaison entre les estimateurs L_1 et ridge est particulièrement intéressante, montrant un meilleur comportement des estimateurs L_1 lorsque le problème de la multicolinéarité est peu important puis une meilleure performance des estimateurs ridge lorsque le degré de multicolinéarité devient plus prononcé. Finalement, le comportement des estimateurs ridge face aux estimateurs L_2 est similaire au cas où les erreurs sont distribuées normalement. Ainsi aucune de ces méthodes d'estimation n'est la plus performante lorsque les deux problèmes des données aberrantes et de la multicolinéarité apparaissent simultanément.

4.10 Investigations préliminaires de l'estimateur L_1-ridge

Pour savoir si l'estimateur L_1-ridge construit au paragraphe traitant de la combinaison des estimateurs L_1 et ridge permet de limiter les effets à la fois des données aberrantes et de la multicolinéarité, il nous a semblé important de le comparer au préalable à d'autres estimateurs utilisés lorsque ces problèmes se posent. Pour le problème des données aberrantes, nous avons retenu certains estimateurs de la classe des L-estimateurs qui sont des des combinaisons de statistiques d'ordre. Ces estimateurs ont joué un rôle important dans le développement des méthodes robustes, notamment dans les modèles univariés pour lesquels les quantiles sont des L-estimateurs. Une généralisation pour les modèles de régression linéaire multiple a été introduite par Koenker et Bassett (1978), basée sur l'analogue des quantiles en p dimensions. L'estimateur correspondant des quantiles de régression est défini comme la solution du problème de minimisation suivant

$$\sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i'\boldsymbol{\beta})$$

où $\mathbf{x}_i'$ est la i-ème ligne de la matrice $\mathbf{X}$ et $\rho_\theta(u)$ la fonction définie par $\rho_\theta(u) = \theta u^+ + (1-\theta)u^-$, avec u^+ et u^- la partie positive et négative de u respectivement. Cet estimateur est noté ici par $\boldsymbol{\beta}_{rq\theta}$ avec θ remplacé par le quantile en question. Les détails de l'algorithme permettant de calculer cet estimateur peuvent être trouvés dans Koenker et d'Orey (1987). A noter le cas particulier important $\boldsymbol{\beta}_{rq50}$ correspondant à la médiane, et qui n'est autre que l'estimateur L_1.

Un autre estimateur intéressant qui a été introduit par Pfaffenberger et Dielman (1990) est en fait l'estimateur ridge calculé avec une valeur de k basée sur l'estimateur L_1 plutôt que sur celui des moindres carrés. Dans ce paragraphe, il sera noté $\boldsymbol{\beta}_{rkPD}$, où

$$k_{PD} = \frac{p\hat{\sigma}^2}{\boldsymbol{\beta}'_{L_1}\boldsymbol{\beta}_{L_1}} \text{ avec } \hat{\sigma}^2 = \frac{\mathbf{e}'_{L_1}\mathbf{e}_{L_1}}{n-p},$$

alors que l'estimateur ridge de Hoerl, Kennard et Baldwin introduit au Chapitre 1 s'obtient avec

$$k_{HKB} = \frac{p\hat{\sigma}^2}{\boldsymbol{\beta}'_{L_2}\boldsymbol{\beta}_{L_2}} \text{ avec } \hat{\sigma}^2 = \frac{\mathbf{e}'_{L_2}\mathbf{e}_{L_2}}{n-p}.$$

La stratégie adoptée pour mener ces premières investigations du comportement de l'estimateur L_1-ridge face aux quantiles de régression et à l'estimateur ridge avec k_{PD} et k_{HKB} est basée sur un programme de simulation écrit dans Splus. De nombreuses situations ont été étudiées en faisant varier les différents facteurs de l'expérience. Ces facteurs n'ont pas tous la même importance. Par exemple, les résultats obtenus en répétant l'expérience 500 ou 1000 fois sont assez semblables, comme ceux obtenus avec $n = 25$ ou $n = 100$. Par contre, les deux facteurs relatifs aux problèmes des données aberrantes et de la multicolinéarité sont les deux plus importants. Il s'agit du coefficient de corrélation ρ et de la distribution des erreurs.

Pour savoir comment réagit l'estimateur L_1-ridge par rapport aux estimateurs cités ci-dessus en présence de données aberrantes, nous avons retenu la distribution de Cauchy avec $\rho = 0$, dans un modèle avec $p = 10$ variables explicatives lorsque la taille de l'échantillon vaut $n = 100$. Le vrai vecteur des paramètres a été choisi avec $\beta_0 = \beta_1 = \cdots = \beta_{10} = 1$ et le nombre de fois que l'expérience a été répétée est de 500. Dans le tableau (4.13), la moyenne des paramètres estimés est donnée pour chaque estimateur considéré. La variance de ces paramètres apparaît dans le tableau (4.14). Ces résultats indiquent qu'en présence du problème des données aberrantes les quantiles de régression sont particulièrement bien estimés puisque leur moyenne est très proche de 1 pour tous les paramètres, à l'exception de l'ordonnée à l'origine qui est logiquement sous-estimée pour les quantiles inférieurs à $\theta = 50\%$ et surestimée pour les quantiles supérieurs à $\theta = 50\%$.

Par.	β_{rq10}	β_{rq20}	β_{rq50}	β_{rq80}	β_{rq90}	β_{rkPD}	$\beta_{L_1-ridge}$
β_0	-3.29	-0.679	1.00	2.74	5.36	0.384	0.767
β_1	0.937	0.980	0.991	0.987	0.964	0.464	0.899
β_2	1.18	1.05	1.02	1.01	0.949	0.490	0.915
β_3	1.01	1.01	0.996	0.987	0.978	0.416	0.814
β_4	1.07	1.04	1.01	0.990	0.914	0.437	0.830
β_5	0.921	0.980	1.00	0.989	0.997	0.477	0.872
β_6	1.01	0.990	1.00	1.03	1.06	0.396	0.810
β_7	1.01	1.02	0.995	0.988	0.964	0.388	0.803
β_8	0.925	0.989	1.00	1.02	1.05	0.417	0.799
β_9	0.949	0.995	1.01	1.05	1.14	0.477	0.876
β_{10}	0.963	0.999	0.990	0.968	0.989	0.434	0.835

Tableau 4.13: Moyenne des paramètres estimés pour la distribution de Cauchy avec $\rho = 0$.

Par.	β_{rq10}	β_{rq20}	β_{rq50}	β_{rq80}	β_{rq90}	β_{rkPD}	$\beta_{L_1-ridge}$
β_0	4.73	0.306	0.0399	0.391	4.46	0.218	0.0694
β_1	1.02	0.163	0.0362	0.153	0.908	0.235	0.0487
β_2	1.37	0.185	0.0514	0.202	1.11	0.239	0.0707
β_3	0.761	0.127	0.0330	0.116	0.758	0.219	0.0554
β_4	0.968	0.169	0.0497	0.196	1.07	0.235	0.0752
β_5	0.915	0.123	0.0300	0.151	0.957	0.228	0.0439
β_6	1.03	0.161	0.0373	0.157	1.05	0.231	0.0589
β_7	0.888	0.143	0.0343	0.154	1.20	0.243	0.0534
β_8	1.08	0.187	0.0396	0.195	0.977	0.234	0.0691
β_9	0.990	0.163	0.0402	0.184	1.25	0.236	0.0613
β_{10}	1.28	0.181	0.0396	0.172	1.12	0.221	0.0621

Tableau 4.14: Variance des paramètres estimés pour la distribution de Cauchy avec $\rho = 0$.

A noter le très bon comportement du cas particulier $\beta_{rq50} = \beta_{L1}$ à la fois en termes de moyenne et de variance. Cet estimateur sera par conséquent retenu pour une étude comparative plus complète. Quant à l'estimateur ridge β_{rkPD}, il est clairement biaisé avec des paramètres estimés tendant fortement vers 0. Dans une moindre mesure, les paramètres estimés par $\beta_{L_1-ridge}$ sont également légèrement sous-estimés par rapport à la vraie valeur des paramètres. Cependant la variance de l'estimateur $\beta_{L_1-ridge}$ est nettement plus petite que celle de l'estimateur β_{rkPD}. Finalement l'estimateur $\beta_{L_1-ridge}$ ayant une variance notablement plus petite que les quantiles de régression (sauf pour le cas particulier de l'estimateur β_{L_1}), il sera naturellement retenu pour une étude plus poussée.

Pour connaître le comportement de ces estimateurs lorsque les variables explicatives sont corrélées, la distribution standard normale $\mathcal{N}(0,1)$ a été générée pour le terme d'erreurs dans le cas où ρ est proche de 1. Comme l'estimateur ridge β_{rkHKB} a été développé pour être utilisé en présence du problème de la multicolinéarité (sans données aberrantes), il a été inclus dans les tableaux (4.15) et (4.16). Ces deux tableaux donnent respectivement la moyenne et la variance des paramètres estimés par les différentes méthodes lorsque $\rho = 0.99$, les autres facteurs étant les mêmes que dans l'expérience précédente. Les résultats obtenus montrent que les paramètres estimés sont proches des vraies valeurs des paramètres pour tous les estimateurs, à l'exception de l'ordonnée à l'origine pour les quantiles de régression inférieurs et supérieurs à $\theta = 50\%$, comme déjà mentionné. Par contre, des différences sensibles apparaissent pour la variance des paramètres estimés selon la méthode d'estimation utilisée. Ici encore, parmi les quantiles de régression, celui correspondant à l'estimateur β_{L_1} se montre le plus performant. A noter que l'estimateur L_1-ridge a une variance plus petite que l'estimateur L_1. Finalement, le meilleur estimateur du point de vue de la variance est β_{rkHKB}. Il sera donc retenu pour être comparé à l'estimateur $\beta_{L_1-ridge}$ dans le prochain paragraphe.

Pour conclure ces investigations préliminaires, nous avons encore étudié la stabilité des estimateurs relativement peu sensibles aux données aberrantes β_{L_1}, β_{rkPD} et $\beta_{L_1-ridge}$ en calculant la variance de l'angle (exprimé en degrés) entre l'hyperplan estimé et le vrai hyperplan.

Comme l'estimateur des moindres carrés β_{L_2} est connu pour son

Par.	β_{rq10}	β_{rq20}	β_{rq50}	β_{rq80}	β_{rq90}	β_{rkHKB}	$\beta_{L_1-ridge}$
β_0	-0.227	0.182	0.996	1.81	2.22	0.992	0.992
β_1	1.09	1.07	1.07	1.10	1.19	1.05	1.07
β_2	0.997	0.973	0.913	0.947	0.816	0.958	0.918
β_3	1.13	1.07	1.07	1.06	0.975	1.04	1.04
β_4	0.922	0.954	1.00	1.02	0.955	0.997	1.00
β_5	0.959	1.04	0.982	0.978	1.03	1.01	0.987
β_6	0.942	0.913	0.931	1.01	0.977	0.980	0.955
β_7	0.900	1.02	1.04	1.03	1.05	1.01	1.03
β_8	0.991	0.968	1.05	0.958	1.02	0.999	1.04
β_9	1.09	1.06	0.934	0.883	0.901	0.969	0.961
β_{10}	0.977	0.937	1.01	1.02	1.10	0.992	0.994

Tableau 4.15: Moyenne des paramètres estimés pour la distribution standard normale avec $\rho = 0.99$.

Par.	β_{rq10}	β_{rq20}	β_{rq50}	β_{rq80}	β_{rq90}	β_{rkHKB}	$\beta_{L_1-ridge}$
β_0	0.0339	0.023	0.0191	0.0248	0.0356	0.0108	0.0183
β_1	3.13	1.92	1.57	2.17	3.09	0.394	0.975
β_2	3.37	2.24	1.76	2.54	3.25	0.398	1.01
β_3	2.36	1.59	1.10	1.59	2.21	0.383	0.727
β_4	3.16	2.12	1.59	2.29	3.68	0.410	0.958
β_5	2.53	1.87	1.19	1.65	2.37	0.346	0.755
β_6	2.93	2.07	1.62	1.88	2.89	0.420	0.868
β_7	2.51	1.77	1.52	1.74	2.49	0.409	0.910
β_8	3.10	2.29	1.77	1.94	2.97	0.434	1.01
β_9	2.89	2.10	1.57	1.94	2.76	0.433	0.931
β_{10}	2.63	2.13	1.48	1.87	2.79	0.358	0.894

Tableau 4.16: Variance des paramètres estimés pour la distribution standard normale avec $\rho = 0.99$.

comportement instable en présence du problème de la multicolinéarité, il a été inclus dans cette étude pour vérifier la cohérence des calculs effectués. Les distributions d'erreur retenues sont la distribution normale contaminée, celle de Laplace, celle de Cauchy et la distribution asymétrique. L'expérience a été menée avec $n = 100$ observations et $p = 2$ variables explicatives pour chacune des 14 valeurs de ρ indiquées dans le plan de simulation. Cette notion de stabilité peu mieux être perçue sur un graphique mettant en relation la variance de l'angle, la distribution d'erreurs et le degré de multicolinéarité. La figure (4.1) dans laquelle la variance de l'angle est tracée en fonction de ρ permet de visualiser les résultats pour l'estimateur β_{L_1} selon les différentes distributions d'erreurs générées. Une échelle logarithmique $(-\log(1-\rho))$ a été utilisée en abscisse pour pouvoir distinguer les résultats pour les valeurs de ρ proches de 1.

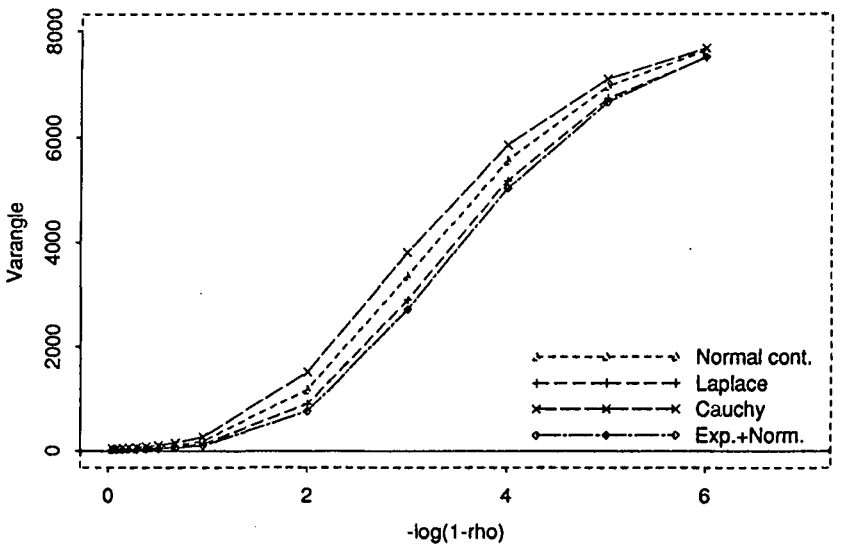

Figure 4.1: Stabilité de l'estimateur L_1.

Les comportements respectifs des estimateurs β_{L_2}, β_{rkPD} et $\beta_{L_1-ridge}$ apparaissent dans les figures (4.2) , (4.3) et (4.4). Il faut tout d'abord constater que pour les quatre estimateurs la stabilité diminue (la variance de l'angle augmente) à mesure que le degré de multicolinéarité

augmente. Cependant, cette mesure de la stabilité laisse apparaître des différences notoires entre ces estimateurs. L'estimateur β_{L_1} comme l'estimateur β_{L_2} devient particulièrement instable dès que le degré de multicolinéarité devient sévère. Il est cependant plus stable que β_{L_2} et β_{rkPD} lorsque le degré de multicolinéarité est faible à modéré, notamment lorsque la distribution des erreurs est celle de Cauchy. Pour toutes les distributions d'erreur, l'estimateur $\beta_{L_1-ridge}$ se comporte aussi bien que l'estimateur β_{L_1} lorsque le degré de multicolinéarité est faible et est nettement plus stable que celui-ci en présence d'un sévère problème de multicolinéarité.

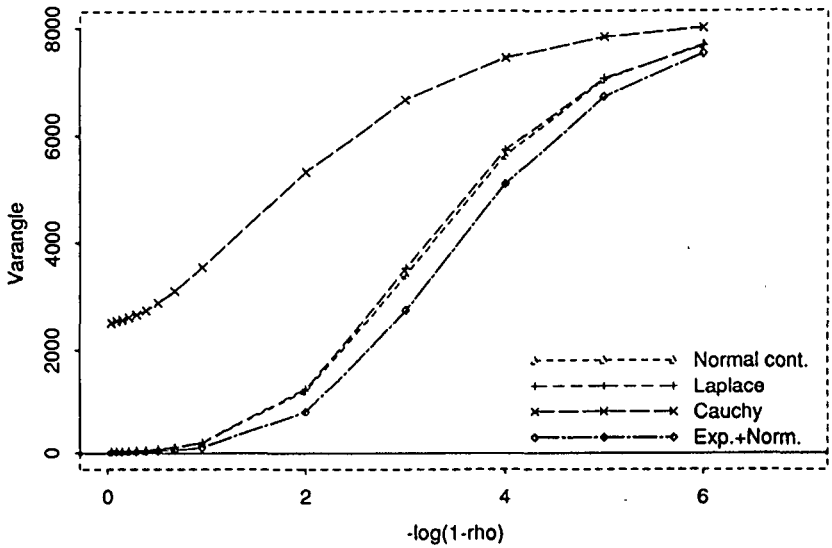

Figure 4.2: Stabilité de l'estimateur L_2.

Finalement, l'estimateur $\beta_{L_1-ridge}$ est toujours plus stable que l'estimateur β_{rkPD}, quelque soit la distribution des erreurs. Ces résultats nous conduisent à retenir l'estimateur $\beta_{L_1-ridge}$ pour une étude plus approfondie sur son comportement lorsque les deux problèmes des données aberrantes et de la multicolinéarité se posent simultanément.

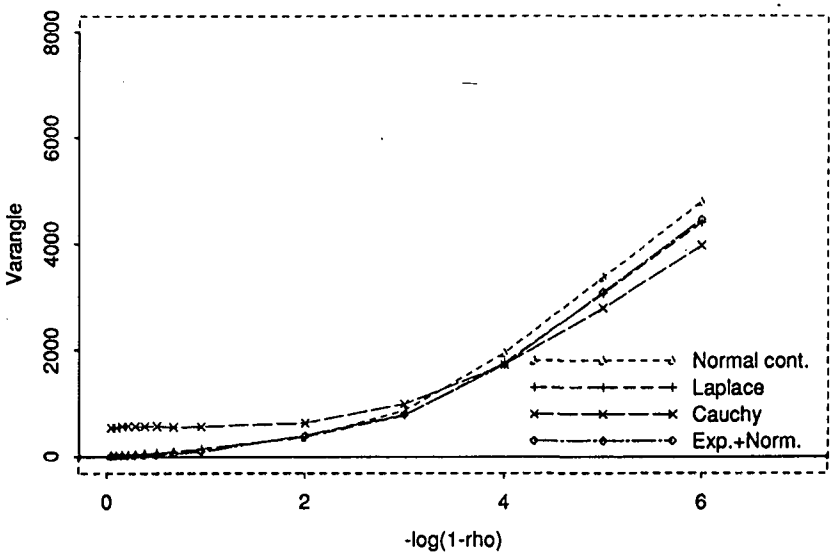

Figure 4.3: Stabilité de l'estimateur ridge avec k_{PD}.

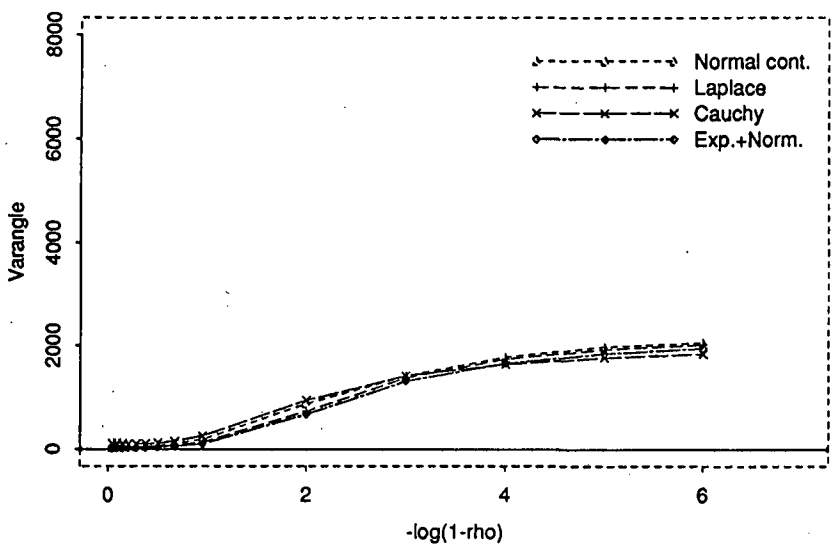

Figure 4.4: Stabilité de l'estimateur L_1-ridge.

4.11 Comportement de l'estimateur L_1-ridge

Ce paragraphe est consacré à l'étude du comportement de l'estimateur L_1-ridge par rapport aux estimateurs L_1, L_2 et ridge respectivement selon différents critères. Pour savoir si l'estimateur L_1-ridge fournit des paramètres estimés plus proches des vraies valeurs des paramètres, nous avons calculé le pourcentage de cas pour lesquels celui-ci a produit un coefficient d'erreur absolue ou quadratique plus petit que l'estimateur L_1, L_2 et ridge respectivement. Les résultats permettant de comparer l'estimateur L_1-ridge et L_1 selon le coefficient d'erreur absolue sont donnés dans le tableau (4.17) pour les différentes valeurs de ρ et les cinq distributions d'erreur dans le cas où $p = 10$. Une valeur supérieure à 0.5 indique une meilleure performance de l'estimateur L_1-ridge par rapport à l'estimateur L_1, alors qu'une valeur proche de 0.5 indique un comportement semblable de ces deux estimateurs. On constate une très nette différence selon le degré de multicolinéarité, l'estimateur L_1-ridge produisant des paramètres estimés toujours plus proches des vraies valeurs des paramètres lorsque le problème de multicolinéarité est très important. Dans le cas où les erreurs sont distribuées normalement et que le degré de multicolinéarité est peu important ($\rho < 0.99$), les deux estimateurs se comportent pratiquement de la même manière. Par contre, en présence du problème des données aberrantes, dans les cas où les erreurs sont distribuées selon une loi normale contaminée, une loi de Laplace et surtout dans le cas de la loi de Cauchy, l'estimateur L_1 est plus performant que l'estimateur L_1-ridge. Ceci correspond bien au fait que les estimateurs L_1 sont particulièrement bien adaptés pour le problème des données aberrantes lorsque le problème de la multicolinéarité ne se pose pas. Par contre, l'estimateur L_1-ridge se comporte très bien lorsque les deux problèmes se posent simultanément.
Finalement, dans le cas de la loi asymétrique, l'estimateur L_1-ridge se comporte mieux que l'estimateur L_1 pour toutes les valeurs de ρ et fournit des paramètres estimés toujours plus proches des vraies valeurs des paramètres lorsque le problème de la multicolinéartité est très sévère. La figure (4.5) permet de visualiser les résultats plus facilement que dans un tableau. Le graphe représente les pourcentages obtenus en fonction de ρ. Comme pour les figures précédentes, une échelle logarithmique a été utilisée pour pouvoir distinguer les résultats pour les valeurs de ρ proches de 1.

ρ	Norm.	Norm. Cont.	Laplace	Cauchy	Exp. + Norm.
0.09	0.506	0.468	0.474	0.278	0.576
0.19	0.508	0.472	0.494	0.376	0.586
0.29	0.456	0.480	0.508	0.380	0.586
0.39	0.476	0.480	0.494	0.426	0.594
0.49	0.512	0.482	0.498	0.434	0.616
0.59	0.526	0.500	0.474	0.414	0.594
0.69	0.510	0.500	0.456	0.432	0.586
0.79	0.498	0.496	0.460	0.470	0.586
0.89	0.524	0.504	0.480	0.588	0.558
0.99	0.892	0.904	0.912	0.964	0.858
0.999	1.00	1.00	1.00	1.00	1.00
0.9999	1.00	1.00	1.00	1.00	1.00
0.99999	1.00	1.00	1.00	1.00	1.00
0.999999	1.00	1.00	1.00	1.00	1.00

Tableau 4.17: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur absolue plus petit que l'estimateur L_1.

La figure (4.6) permet de visualiser les résultats pour le coefficient d'erreur quadratique.
Les résultats permettant de comparer l'estimateur L_1-ridge avec l'estimateur L_1 selon le coefficient d'erreur quadratique sont globalement semblables à ceux obtenus avec le coefficient d'erreur absolue. Lorsque le degré de multicolinéarité est très important, l'estimateur L_1-ridge fournit des paramètres estimés toujours plus proches des vraies valeurs des paramètres que l'estimateur L_1, quelque soit la distribution des erreurs. Dans le cas où le degré de multicolinéarité est peu important, les deux estimateurs se comportent de façon assez semblables pour les cinq distributions. A noter qu'avec le coefficient d'erreur quadratique, contrairement au coefficient d'erreur absolue, l'estimateur L_1-ridge se comporte légèrement mieux que l'estimateur L_1 lorsque la loi est celle de Laplace ou Cauchy. Ainsi, du point de vue du critère utilisé, le coefficient d'erreur absolue semble quelque peu favoriser l'estimateur L_1 dans ce cas. Cependant, dans les cas pour lesquels il y a les deux problèmes des données aberrantes et de la multicolinéarité, l'estimateur L_1-ridge se comporte toujours mieux que l'estimateur L_1 indépendemment du critère utilisé.

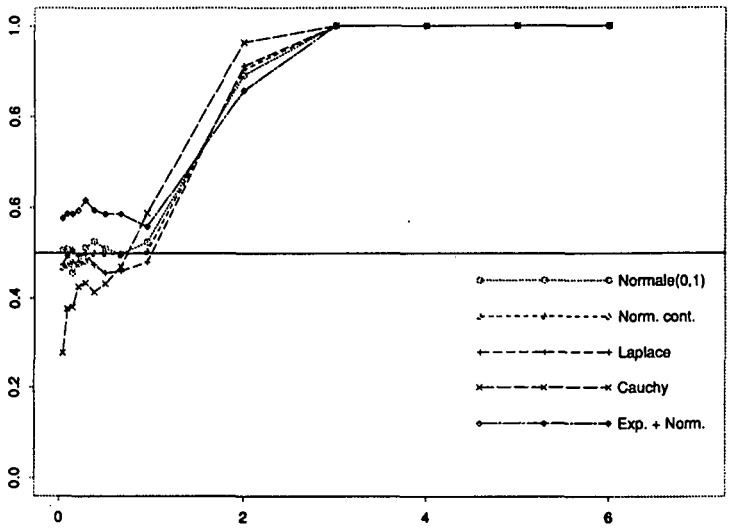

Figure 4.5: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur absolue plus petit que l'estimateur L_1.

Voyons à présent comment se comporte l'estimateur L_1-ridge face à l'estimateur L_2. La figure (4.7) permet de comparer le comportement de l'estimateur L_1-ridge et L_2 en fonction du critère du coefficient d'erreur absolue. L'estimation L_2 étant reconnue pour être la meilleure méthode d'estimation lorsque les erreurs sont normalement distribuées et qu'il n'y a pas de problème de multicolinéarité, il est parfaitement logique que cet estimateur l'emporte sur l'estimateur L_1-ridge pour les faibles valeurs de ρ lorsque la distribution des erreurs est normale. En effet, dans ce cas les pourcentages de cas pour lesquels l'estimateur L_1-ridge a fournit un coefficient d'erreur absolue plus petit que l'estimateur L_2 sont plus petits que 0.5; par contre, même pour la distribution d'erreur normale, l'estimateur L_1-ridge est plus proche des vraies valeurs des paramètres lorsque le degré de multicolinéarité devient important. Pour toutes les autres distributions d'erreur, l'estimateur L_1-ridge est plus performant que l'estimateur L_2 quelque soit le degré de multicolinéarité. Dans le cas de la distribution de Cauchy, l'estimateur L_1-ridge fournit toujours des paramètres estimés plus proches des vraies valeurs des paramètres

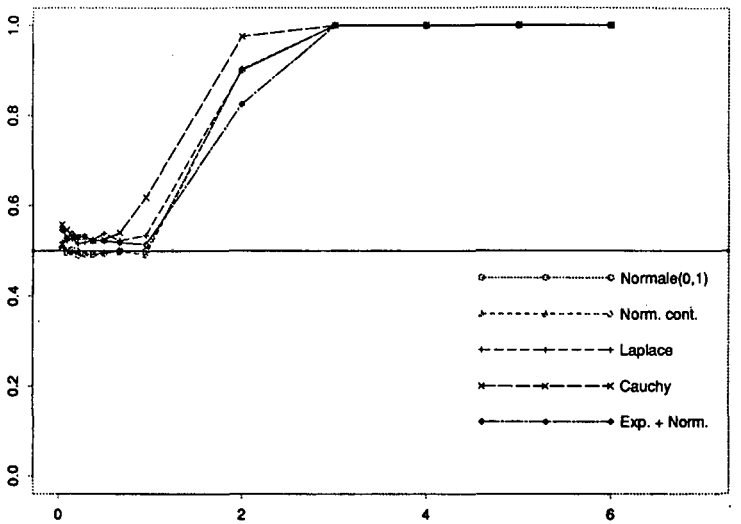

Figure 4.6: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur quadratique plus petit que l'estimateur L_1.

que l'estimateur L_2. On retrouve ainsi dans ce nouvel estimateur la propriété d'être peu affecté par les données aberrantes. Remarquons encore que l'estimteur L_2 semble être plus affecté que l'estimateur L_1-ridge lorsque la distribution des erreurs est asymétrique. La figure (4.8) correspond au critère du coefficient d'erreur quadratique. les résultats sont comparables à ceux obtenus pour le coefficient d'erreur absolue. L'estimateur L_1-ridge se comportant mieux que l'estimateur L_2 lorsque le problème de la multicolinéarité s'accentue. De même, lorsque la distribution des erreurs n'est pas normale, l'estimateur L_1-ridge l'emporte sur l'estimateur L_2.

A noter la supériorité de l'estimateur L_2 lorsque le problème de la multicolinéarité est peu marqué et que les erreurs suivent une loi normale. Cette supériorité est cependant moins prononcée qu'avec le coefficient d'erreur absolue qui favorise plus l'estimation L_1-ridge. Finalement nous retiendrons de cette comparaison, le très bon comportement de l'estimateur L_1-ridge en présence des deux problèmes des données aberrantes et de la multicolinéarité.

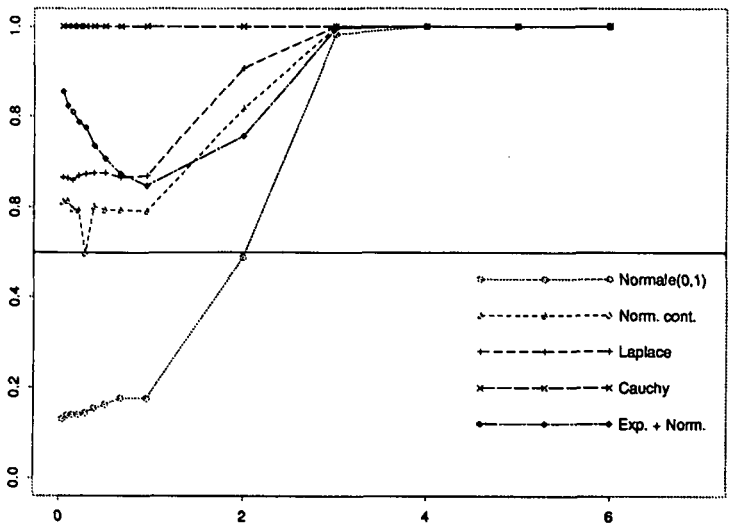

Figure 4.7: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur absolue plus petit que l'estimateur L_2.

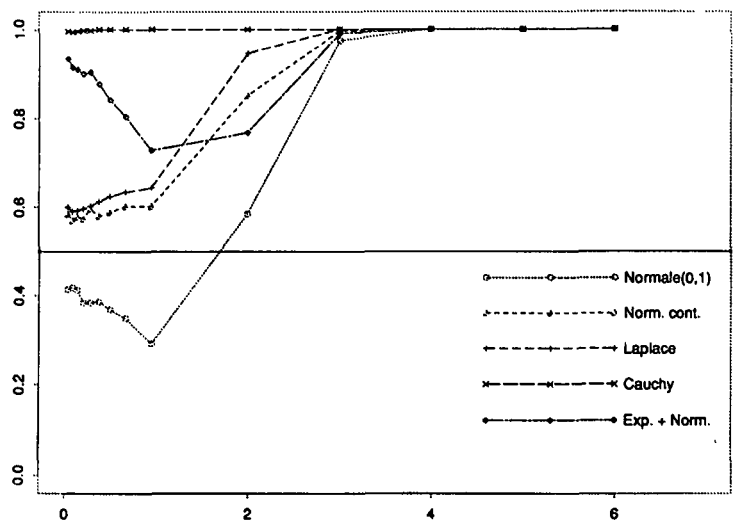

Figure 4.8: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur quadratique plus petit que l'estimateur L_2.

Il nous reste à comparer la performance de l'estimateur L_1-ridge face à l'estimateur ridge. Les résultats pour le coefficient d'erreur absolue sont représentés dans la figure (4.9). On constate, comme dans le cas des estimateurs L_2, que l'estimateur ridge se comporte mieux que l'estimateur L_1-ridge lorsque les erreurs sont distribuées selon une loi normale. Cependant, lorsque la multicolinéarité devient très sévère, l'estimateur L_1-ridge l'emporte. Pour les autres distributions d'erreur, sauf pour celle de Cauchy, les résultats présentent une structure particulière, indiquant une performance d'autant meilleure de l'estimateur ridge à mesure que ρ augmente puis inversement. Ainsi, sur un intervalle, l'estimateur ridge l'emporte sur l'estimateur L_1-ridge. Il est possible que le choix retenu pour k dans l'estimation ridge soit à l'origine de ce phénomène; en effet, la valeur optimale de k n'étant pas connue mais estimée à partir des données, il n'est pas étonnant que le comportement de l'estimateur ridge passe par une valeur optimale en fonction de ρ. Cependant, à mesure que le problème de la multicolinéarité devient plus important, l'estimateur L_1-ridge l'emporte sur l'estimateur ridge. Ici encore, dans le cas où les erreurs suivent une loi de Cauchy, l'estimateur L_1-ridge est toujours plus proche des vraies valeurs des paramètres.

Finalement, la figure (4.10) fournit les résultats pour le coefficient d'erreur quadratique indiquant un comportement analogue de ces deux estimateurs par rapport aux résultats obtenus avec le coefficient d'erreur absolue. Dans ce cas également, la supériorité de l'estimateur ridge est moins marquée que pour le coefficient d'erreur absolue pour les faibles valeurs de ρ. De plus les remarques faites dans le cas du coefficient d'erreur absolue sur le bon comportement de l'estimateur ridge pour certaines valeurs de ρ s'appliquent également ici. Cependant, l'estimateur L_1-ridge se comporte mieux que l'estimateur ridge lorsque le problème de la multicolinéarité est sévère. De plus, lorsqu'il n'y a pas de problème de multicolinéarité mais que les erreurs sont distribuées selon une loi de Cauchy, l'estimateur L_1-ridge est aussi plus performant que l'estimateur ridge. Ainsi, l'estimateur L_1-ridge est particulièrement bien adapté lorsque les deux problèmes des données aberrantes et de la multicolinéarité apparaissent simultanément.

Pour être aussi complet que possible dans la comparaison de l'estimateur L_1-ridge avec les trois autres estimateurs, nous les comparons à présent selon le critère de l'erreur quadratique moyenne. Cette com-

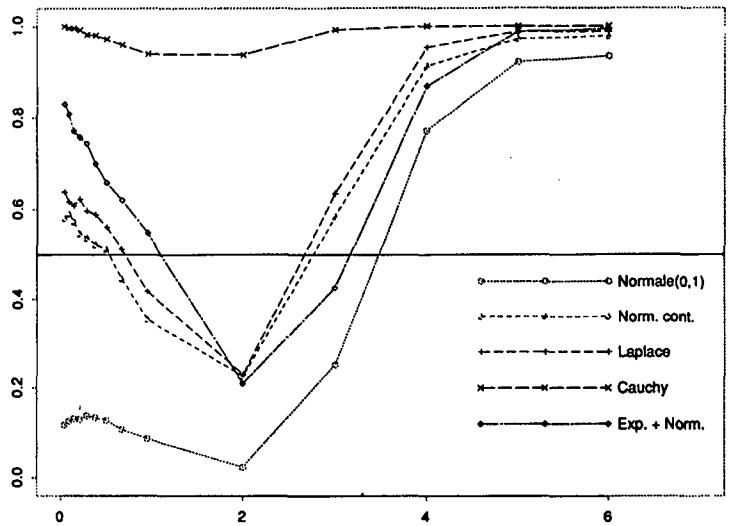

Figure 4.9: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur absolue plus petit que l'estimateur ridge.

paraison se fera sur la même base que celle utilisée pour comparer les estimateurs L_1, L_2 et ridge en calculant les rapports respectifs de l'erreur quadratique moyenne. Rappelons que le comportement de ces rapports pour les différents paramètres étant très semblables, nous ne présentons les résultats que pour l'un de ces paramètres, notamment β_1, dans le cas où $n = 100$. Ainsi en calculant le rapport de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur celle de chacun des autres estimateurs, il est possible de comparer la performance de l'estimateur L_1-ridge à celle des autres estimateurs selon le degré de multicolinéarité et pour les différentes distributions d'erreur. Une valeur inférieure à 1 indiquant une meilleure performance de l'estimateur L_1-ridge. Les calculs ont été faits pour $p = 2$, $p = 5$ et $p = 10$. Cependant, seuls les cas $p = 2$ et $p = 10$ sont présentés ici du fait que les résultats sont relativement peu différents selon les valeurs de p. Les résultats sont également présentés graphiquement de manière à pouvoir directement comparer les différents estimateurs. La figure (4.11) permet de comparer l'estimateur L_1-ridge et l'estimateur L_1 pour $p = 2$. La droite tirée à la hauteur 1 permet de

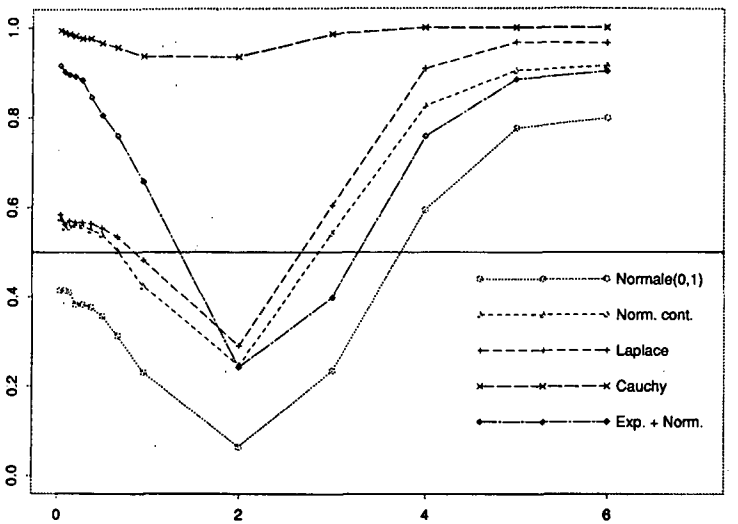

Figure 4.10: Pourcentage de cas où l'estimateur L_1-ridge a produit un coefficient d'erreur quadratique plus petit que l'estimateur ridge.

mieux distinguer dans quels cas l'estimateur L_1-ridge est le plus performant.

On constate une très nette différence selon le degré de multicolinéarité. En effet, pour toutes les distributions d'erreur, sauf celle de Cauchy, les estimateurs L_1-ridge et L_1 se comportent pratiquement de manière identique lorsque le degré de multicolinéarité est peu élevé, avec des rapports proches de 1. A noter un comportement plus performant de l'estimateur L_1 en présence de données aberrantes générées par la loi de Cauchy lorsque le problème de la multicolinéarité est peu important. Cependant, même pour cette distribution, l'estimateur L_1-ridge l'emporte sur l'estimateur L_1 lorsque le degré de multicolinéarité augmente, indiquant une performance nettement supérieure de l'estimateur L_1-ridge en présence de données aberrantes et du problème de la multicolinéarité. Pour les autres distributions, le comportement de l'estimateur L_1-ridge face à l'estimateur L_1 est également d'autant meilleur que le degré de multicolinéarité augmente. Ainsi, ce nouvel estimateur est nettement moins sensible que l'estimateur L_1 en présence de la multicolinéarité.

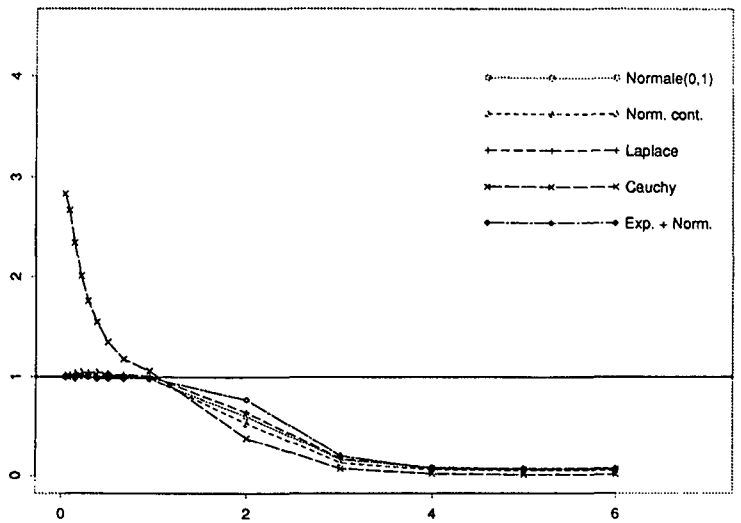

Figure 4.11: Rapports de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur l'estimateur L_1 ($p = 2$).

La figure (4.12) fournit les résultats dans le cas où $p = 10$. La différence la plus marquante avec le cas $p = 2$ est le comportement de ces deux estimateurs dans le cas de la loi de Cauchy. Cette fois, l'estimateur L_1-ridge se comporte pratiquement de la même manière que l'estimateur L_1 pour les faibles valeurs de ρ. Ceci peut s'expliquer du fait que ρ représente le coefficient de corrélation entre paires de variables. Ainsi, en augmentant le nombre de variables explicatives, le degré de multicolinéarité est plus important pour une valeur fixée de ρ. On retiendra de la comparaison entre ces deux estimateurs que la supériorité de l'estimateur L_1-ridge dépend essentiellement du problème de la multicolinéarité et peu de la distribution des erreurs. En effet, ces deux estimateurs se comportent globalement de façon analogue lorsque le degré de multicolinéarité est peu important, alors que l'estimateur L_1-ridge devient de plus en plus performant à mesure que le degré de multicolinéarité augmente.

Voyons à présent comment se comporte l'estimateur L_1-ridge face à l'estimateur L_2. La figure (4.13) représente les résultats pour $p = 2$. Les

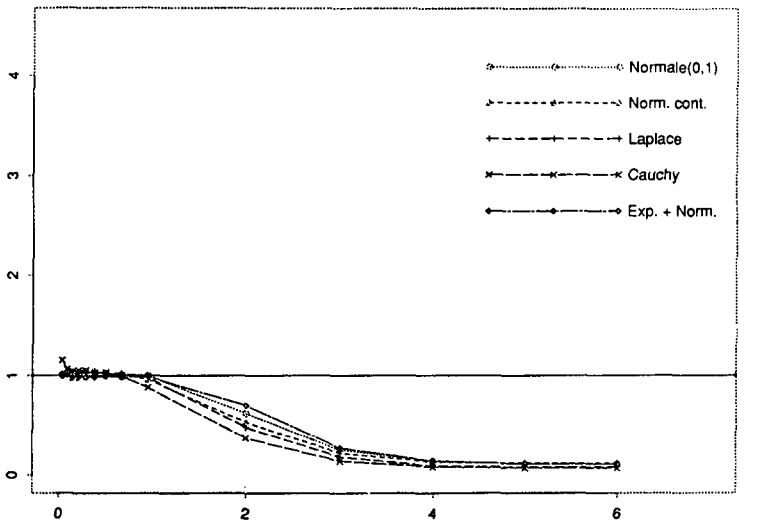

Figure 4.12: Rapports de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur l'estimateur L_1 ($p = 10$).

résultats doivent cette fois être nuancés selon la distribution d'erreur; en effet, lorsque la distribution des erreurs est normale et que le problème de la multicolinéarité est peu important, l'estimateur L_2 est plus performant que l'estimateur L_1-ridge; ce résultat confirme que l'estimateur L_2 est le plus performant lorsqu'il n'y a ni problème de données aberrantes ni problème de multicolinéarité. Par contre, l'estimateur L_1-ridge devient plus performant que l'estimateur L_2 dès que le problème de multicolinéarité devient sévère. Pour les autres distributions, l'estimateur L_1-ridge est meilleur que l'estimateur L_2, notamment dans le cas de la loi de Laplace et celle de Cauchy.
Pour cette dernière, l'estimateur L_1-ridge est toujours supérieur à l'estimateur L_2 quelque soit le degré de multicolinéarité. L'estimateur L_1-ridge se comporte une fois de plus particulièrement bien en présence des deux problèmes évoqués. De plus, face à l'estimateur L_2, il a tendance à se comporter d'autant mieux que le degré de multicolinéarité augmente. La figure (4.14) permet de comparer l'estimateur L_1-ridge et L_2 lorsque $p = 10$. Les résultats sont très semblables à ceux obtenus

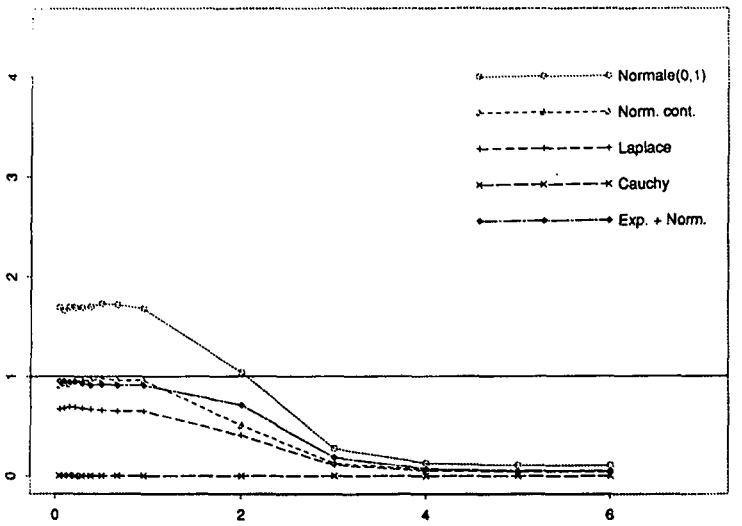

Figure 4.13: Rapports de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur l'estimateur L_2 ($p = 2$).

pour $p = 2$ et appellent les mêmes commentaires du point de vue de la performance de ces deux estimateurs. A noter toutefois que lorsque le problème de la multicolinéarité est peu prononcé, l'estimateur L_1-ridge se comporte mieux que l'estimateur L_2 lorsque les erreurs suivent une loi normale contaminée. Pour la loi de Cauchy, l'estimateur L_1-ridge est toujours plus performant que l'estimateur L_2 quelque soit le degré de multicolinéarité, l'estimateur L_1-ridge se montrant particulièrement peu sensible au problème des données aberrantes.

Il nous reste à comparer le comportement de l'estimateur L_1-ridge par rapport à l'estimateur ridge. Les résultats obtenus pour les différentes valeurs de p étant semblables (comme dans le cas de la comparaison avec les estimateurs L_2), nous donnons dans la figure (4.15) les résultats pour $p = 10$. On y retrouve un comportement qui a déjà pu être observé en comparant ces deux estimateurs selon les critères des coefficients d'erreur absolue et quadratique. Lorsque la distribution des erreurs est normale, l'estimateur ridge se comporte mieux que l'estimateur L_1-ridge à mesure que ρ augmente puis inversement. De même pour les autres

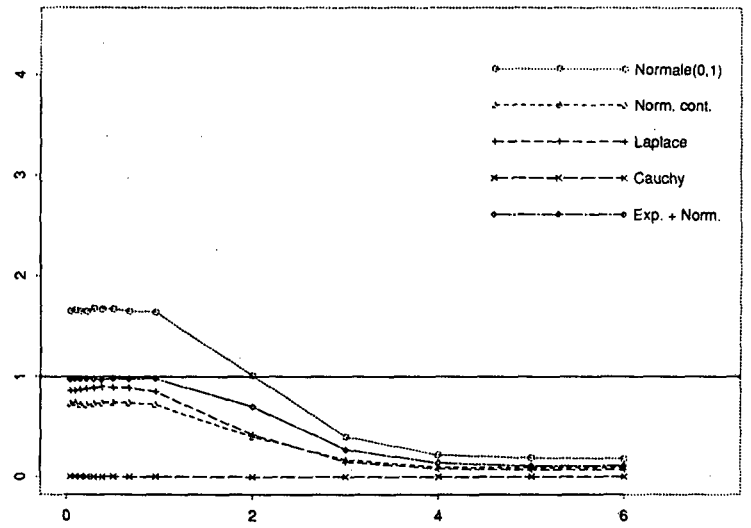

Figure 4.14: Rapports de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur l'estimateur L_2 ($p = 10$).

distributions, excepté pour celle de Cauchy, l'estimateur ridge est plus performant sur un intervalle. Par contre, l'estimateur ridge est particulièrement sensible aux données aberrantes comme dans le cas de la loi de Cauchy, pour laquelle l'estimateur L_1-ridge l'emporte quelque soit le degré de multicolinéarité. D'autre part, lorsque le problème de multicolinéarité est sévère, l'estimateur L_1-ridge l'emporte sur l'estimateur ridge, plus particulièrement lorsque la loi est celle de Laplace ou la loi normale contaminée. On retrouve ici la bonne performance de l'estimateur L_1-ridge lorsque les erreurs ne sont pas distribuées normalement.

Dans le cas de la loi normale, les deux estimateurs se comportent pratiquement de la même manière lorsque le degré de multicolinéarité est très élevé. Ainsi, l'estimateur L_1-ridge se comporte particulièrement bien lorsque les deux problèmes des données aberrantes et de la multicolinéarité se posent simultanément. Par contre, lorsqu'il n'y a que le problème de la multicolinéarité qui se pose, le recours à l'estimateur ridge reste une alternative raisonnable.

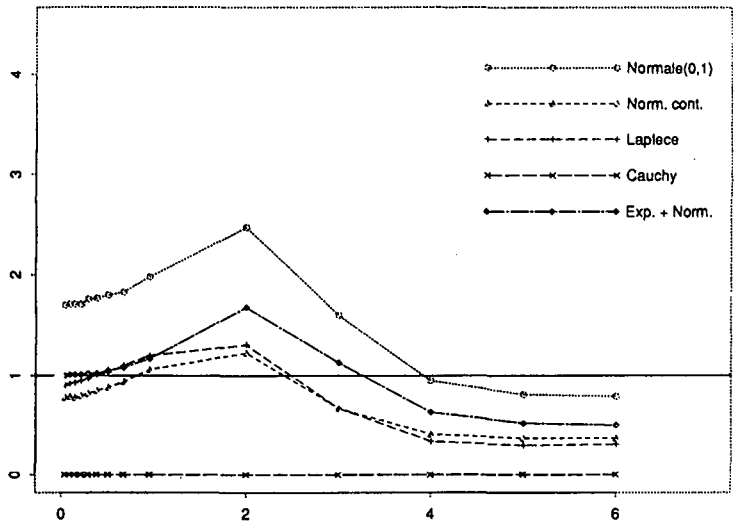

Figure 4.15: Rapports de l'erreur quadratique moyenne de l'estimateur L_1-ridge sur l'estimateur ridge ($p = 10$).

4.12 Conclusion

L'étude comparative menée dans ce chapitre nous a permis de mieux comprendre le comportement des différents estimateurs lorsque les deux problèmes des données aberrantes et de la multicolinéarité se posaient. Pour chaque problème pris séparément, il existe des méthodes d'estimation mieux adaptées que l'estimation L_2, notamment l'estimation L_1 dans le cas des données aberrantes et l'estimation ridge lorsque le degré de multicolinéarité est important. Cependant, lorsque les deux problèmes se posent simultanément, aucune de ces méthodes n'est la plus appropriée. En effet, l'estimateur L_1 est sensible au problème de la multicolinéarité et est moins performant que l'estimateur ridge dans ce cas. D'autre part, l'estimateur ridge, comme l'estimateur L_2, peut être fortement influencé par les données aberrantes. Une approche naturelle pour se protéger de ces deux problèmes lorsqu'ils apparaissent simultanément consiste à combiner l'estimateur L_1 et ridge. L'estimateur ainsi obtenu, appelé estimateur L_1-ridge, peut être calculé sans passer par une tech-

nique itérative des moindres carrés pondérés mais directement sur les données augmentées. Une large place a été consacrée au comportement de ce nouvel estimateur par rapport à d'autres estimateurs. Dans une étude préliminaire, nous avons comparé l'estimateur L_1-ridge aux quantiles de régression ainsi qu'à l'estimateur ridge obtenu avec k_{PD} et k_{HKB} respectivement. Les résultats ont montré que l'estimateur L_1-ridge se comporte mieux que les autres estimateurs lorsque les deux problèmes des données aberrantes et de la multicolinéarité se posent simultanément. En présence du seul problème des données aberrantes, les quantiles de régression avec $\theta = 50\%$ (estimateur L_1!) sont les mieux adaptés. Lorsque seul le problème de la multicolinéarité se pose, l'estimateur ridge avec k_{HKB} est le plus performant. Nous avons donc mené une étude plus approfondie en retenant les estimateurs L_1, L_2 et ridge pour comparer leurs performances à celles de l'estimateur L_1-ridge (avec k_{HKB}). Les résultats obtenus permettent de considérer la méthode d'estimation L_1-ridge comme une alternative particulièrement bien adaptée lorsque les deux problèmes se posent simultanément. En effet, l'estimateur L_1-ridge se comporte mieux que l'estimateur L_1 lorsque le degré de multicolinéarité est élevé. De même, il est beaucoup moins influencé par les données aberrantes que l'estimateur ridge, en particulier quand la distribution des erreurs suit une loi de Cauchy. De plus, lorsque le problème de multicolinéarité est sévère, l'estimateur L_1-ridge est également plus performant. Par rapport à l'estimateur L_2, cet estimateur se comporte d'autant mieux à mesure que le degré de multicolinéarité augmente, quelque soit la distribution des erreurs. Il est également plus performant en présence du problème des données aberrantes lorsque le degré de multicolinéarité est faible. Ainsi pour chaque problème pris séparément, l'estimateur L_1-ridge est mieux adapté que l'estimateur L_2. La supériorité de l'estimateur L_1-ridge est encore plus marquée lorsque les deux problèmes se posent simultanément. Par contre, l'estimateur L_2 reste le meilleur estimateur lorsqu'aucun des deux problèmes ne se pose. Ainsi, la méthode d'estimation la mieux adaptée dépend de la nature du problème rencontré. Ici, dans le cas où les deux problèmes des données aberrantes et de la multicolinéarité se posent simultanément, l'estimateur L_1-ridge est celui qui se comporte le mieux face aux estimateurs L_1, L_2 et ridge respectivement.

Finalement, il serait intéressant de calculer les estimateurs ridge et L_1-ridge en fonction de k, la valeur de k estimée à partir des données n'étant pas forcément optimale. De ce point de vue, les résultats trouvés dans ce chapitre, notamment le théorème 2 démontré dans le paragraphe consacré aux propriétés des estimateurs L_1-ridge peuvent s'avérer particulièrement utiles et facilement appliqués en posant $\lambda = \sqrt{k}$. Cette approche sera reprise dans le chapitre relatif aux applications.

CHAPITRE 5

APPLICATIONS

5.1 Introduction

Ce chapitre est consacré aux nouvelles applications des différents estimateurs étudiés dans ce travail. Dans les modèles de régression linéaire multiple, la méthode des moindres carrés est la plus couramment utilisée. Cependant, dans les ensembles de données réelles, plusieurs problèmes peuvent se poser, notamment celui de la multicolinéarité et des données aberrantes. De nombreux progrès ont été faits dans le développement de diagnostics pour détecter les données aberrantes. L'estimation L_1 est d'ailleurs l'une des méthodes reconnues pour cela. Dans le cas où le problème de multicolinéarité se pose, certaines techniques ont également été développées, comme par exemple l'estimation ridge qui permet d'obtenir des estimateurs plus stables en réduisant les paramètres estimés vers 0 (selon la valeur de k). La combinaison des estimateurs L_1 et ridge introduite dans le chapitre précédent fournit également des paramètres estimés qui ont tendance à être réduits vers 0 lorsque k augmente. Contrairement aux paramètres estimés par la méthode ridge qui ne sont en général pas nuls, certains paramètres obtenus par l'estimation L_1-ridge s'annulent en augmentant k. Il est donc possible de sélectionner les variables correspondant aux paramètres non nuls de l'estimation L_1-ridge et d'éliminer celles pour lesquelles ces paramètres sont nuls. Il s'agit là d'une application particulièrement intéressante pour la construction d'un modèle. Nous étudierons dans ce chapitre deux applications sur des données de type économique et comparerons les résultats

avec ceux obtenus par les outils conventionnels de sélection d'un modèle. Certains logiciels comme Minitab (version 9.1) permettent d'évaluer toutes les combinaisons possibles des variables explicatives. S'il y a p variables explicatives, il y a 2^p sous-ensembles possibles de variables explicatives correspondant aux différents modèles. La commande *breg* (best regression) de Minitab sélectionne les deux meilleurs sous-ensembles de chaque taille, c'est-à-dire ceux pour lesquels le coefficient de détermination est le plus grand. Il est ainsi possible de comparer les résultats obtenus avec cette méthode à ceux obtenus en appliquant les estimateurs L_1-ridge aux données.

5.2 Sélection L_1-ridge de modèles (prix de vente des maisons)

Le premier exemple se rapporte à un ensemble de données relatif aux prix de vente de maisons (Narula et Wellington (1977), Weisberg (1985)). Il s'agit de prédire Y, le prix de vente d'une maison (en milliers de dollars), en fonction des variables explicatives décrites ci-dessous

X_1 : Taxes (en centaines de dollars)
X_2 : Nombre de salles de bain
X_3 : Grandeur du terrain (en milliers de pieds carrés)
X_4 : Surface habitable (en milliers de pieds carrés)
X_5 : Nombre de garages
X_6 : Nombre de pièces
X_7 : Nombre de chambres à coucher
X_8 : Age de la maison (en années)
X_9 : Type de construction (brique (1), brique et bois (2), aluminium et bois (3), bois (4))
X_{10} : Style (2 étages (1), 1 étage et demi (2), ranch (3))
X_{11} : Nombre de cheminées

Les données relatives à ces variables correspondent à $n = 27$ maisons vendues en Pennsylvanie (USA) et figurent dans le tableau (5.1). Avant d'analyser ces données avec les différentes méthodes de régression linéaire, examinons l'indice de conditionnement pour savoir s'il y a un problème de multicolinéarité. Pour cela, nous calculons les valeurs propres de la

matrice $\mathbf{X}'\mathbf{X}$ en les ordonnant de manière décroissante. Dans ce qui suit les données ont été au préalable standardisées, ce qui permet d'avoir des variables du même ordre de grandeur et d'éviter les problèmes d'unités. Les valeurs propres sont les suivantes

$$\lambda_1 = 139.07,\ \lambda_2 = 48.48,\ \lambda_3 = 26.76,\ \lambda_4 = 23.47,\ \lambda_5 = 18.99,$$

$$\lambda_6 = 12.73,\ \lambda_7 = 7.97,\ \lambda_8 = 4.50,\ \lambda_9 = 2.17,\ \lambda_{10} = 1.09,\ \lambda_{11} = 0.78.$$

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
4.9176	1.0	3.4720	0.9980	1.0	7	4	42	3	1	0	25.9
5.0208	1.0	3.5310	1.5000	2.0	7	4	62	1	1	0	29.5
4.5429	1.0	2.2750	1.1750	1.0	6	3	40	2	1	0	27.9
4.5573	1.0	4.0500	1.2320	1.0	6	3	54	4	1	0	25.9
5.0597	1.0	4.4550	1.1210	1.0	6	3	42	3	1	0	29.9
3.8910	1.0	4.4550	0.9880	1.0	6	3	56	2	1	0	29.9
5.8980	1.0	5.8500	1.2400	1.0	7	3	51	2	1	1	30.9
5.6039	1.0	9.5200	1.5010	0.0	6	3	32	1	1	0	28.9
15.4202	2.5	9.8000	3.4200	2.0	10	5	42	2	1	1	84.9
14.4598	2.5	12.8000	3.0000	2.0	9	5	14	4	1	1	82.9
5.8282	1.0	6.4350	1.2250	2.0	6	3	32	1	1	0	35.9
5.3003	1.0	4.9883	1.5520	1.0	6	3	30	1	2	0	31.5
6.2712	1.0	5.5200	0.9750	1.0	5	2	30	1	2	0	31.0
5.9592	1.0	6.6660	1.1210	2.0	6	3	32	2	1	0	30.9
5.0500	1.0	5.0000	1.0200	0.0	5	2	46	4	1	1	30.0
8.2464	1.5	5.1500	1.6640	2.0	8	4	50	4	1	0	36.9
6.6969	1.5	6.9020	1.4880	1.5	7	3	22	1	1	1	41.9
7.7841	1.5	7.1020	1.3760	1.0	6	3	17	2	1	0	40.5
9.0384	1.0	7.8000	1.5000	1.5	7	3	23	3	3	0	43.9
5.9894	1.0	5.5200	1.2560	2.0	6	3	40	4	1	1	37.5
7.5422	1.5	4.0000	1.6900	1.0	6	3	22	1	1	0	37.9
8.7951	1.5	9.8900	1.8200	2.0	8	4	50	1	1	1	44.5
6.0931	1.5	6.7265	1.6520	1.0	6	3	44	4	1	0	37.9
8.3607	1.5	9.1500	1.7770	2.0	8	4	48	1	1	1	38.9
8.1400	1.0	8.0000	1.5040	2.0	7	3	3	1	3	0	36.9
9.1416	1.5	7.3262	1.8310	1.5	8	4	31	4	1	0	45.8
12.0000	1.5	5.0000	1.2000	2.0	6	3	30	3	1	1	41.0

Tableau 5.1: Données relatives au prix de vente de maisons.

L'indice de conditionnement s'obtient en calculant le rapport de la plus grande valeur propre à la plus petite $\kappa = \frac{\lambda_1}{\lambda_{11}} = \frac{139.066}{0.781} = 178.06$.

Le problème de multicolinéarité semble relativement important ici; en effet, une valeur plus grande que 100 pour l'indice de conditionnement peut être révélatrice d'un problème de multicolinéarité. Dans ce problème de régression multiple, comme cela arrive relativement souvent en économie, les variables explicatives choisies sont corrélées entre elles. Il n'est d'ailleurs pas étonnant de rencontrer ce problème en examinant ces variables; par exemple, la surface habitable (X_4) est certainement liée au nombre de pièces (X_6), au nombre de chambres à coucher (X_7) et au nombre de salles de bain (X_2). Voyons à présent les résultats obtenus par la méthode des moindres carrés. Le tableau (5.2) contient les paramètres estimés pour les données standardisées, leur écart-type, la valeur t du test de Student et la valeur p correspondante. Ces résultats appellent plusieurs remarques. En ce qui concerne les paramètres, il est surprenant de trouver des valeurs négatives pour les paramètres relatifs aux variables X_6 et X_7. En effet, le prix de vente d'une maison devrait logiquement être positivement associé au nombre de pièces et de chambres à coucher. Il s'agit ici d'un problème lié à celui de la multicolinéarité pour lequel il n'est pas rare de trouver des coefficients avec un signe erroné. Mais le problème le plus important est certainement celui du test de Student qui permet de déterminer si les coefficients sont significativement non nuls (à un seuil de 5%, une valeur p supérieure à 0.05 indique un coefficient qui n'est pas significativement non nul). On constate que seul le coefficient correspondant à la variable X_4 est significativement non nul à un seuil de 5% (la conclusion est d'ailleurs la même jusqu'à un seuil de 20%). Ici aussi, ces résultats sont une conséquence directe de la multicolinéarité, puisque la variance (et donc l'écart-type) des paramètres estimés a tendance à être exagérément grande en présence du problème de la multicolinéarité, rendant ainsi les valeurs t plus petites. Il est toujours difficile de traiter ce type de problème directement. La méthode de régression L_1-ridge permet de calculer les coefficients de l'équation de régression en fonction du paramètre k, comme indiqué dans le chapitre 4.

Contrairement à l'estimation ridge qui a tendance à réduire uniformément vers 0 tous les paramètres estimés, l'estimation L_1-ridge se comporte différemment. Certains paramètres deviennent rapidement nuls en augmentant légèrement k, alors que d'autres restent non nuls. Ce comportement peut ainsi être exploité pour déterminer les variables à retenir dans

Variable	Coef.	Stdev.	t-ratio	p
X_1	0.1700	0.1802	0.94	0.361
X_2	0.2439	0.2361	1.03	0.318
X_3	0.04311	0.09504	0.45	0.657
X_4	0.5477	0.2013	2.72	0.016
X_5	0.06914	0.08304	0.83	0.418
X_6	-0.0868	0.2234	-0.39	0.703
X_7	-0.0165	0.2174	-0.08	0.941
X_8	-0.07317	0.09384	-0.78	0.448
X_9	0.08348	0.06364	1.31	0.209
X_{10}	0.0422	0.1181	0.36	0.726
X_{11}	0.09032	0.08778	1.03	0.320

Tableau 5.2: Résultats obtenus par la méthode des moindres carrés.

le modèle, notamment celles dont les coefficients sont relativement stables et non nuls. Les résultats de l'estimation L_1-ridge sont donnés dans le tableau (5.3) pour des valeurs de k variant entre 0 et 1 comme il est en général préconisé pour l'estimation ridge. Une variante de cette méthode fournissant les mêmes conclusions serait d'utiliser le théorème 2 démontré à la fin du chapitre 4 pour déterminer la valeur de k pour laquelle tous les coefficients sont nuls, puis de réduire k vers 0. La première ligne du tableau (5.3) fournit les valeurs des paramètres estimés par la méthode L_1 du fait que $k = 0$. On constate les mêmes problèmes de signe que pour l'estimation par la méthode des moindres carrés. Par contre, il est intéressant de voir le comportement de chaque paramètre en fonction de k (seules les valeurs de k pour lesquelles il y a un changement dans les valeurs des paramètres apparaissent ici). Visiblement, certains coefficients deviennent rapidement nuls. Il s'agit dans l'ordre des coefficients relatifs aux variables X_{10}, X_7, X_3, X_6, X_5 et X_8. A noter que dans le cas des coefficients relatifs à X_3 et à X_7, ils peut arriver qu'ils n'aient pas une valeur exactement nulle pour certaines valeurs de k. Deux autres coefficients, notamment ceux correspondant aux variables X_9 et X_{11} ne sont pas exactement nuls mais décroissent rapidement vers des valeurs proches de 0.

k	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
.00	.37	.22	-.07	.41	.08	-.14	.05	-.11	.06	-.03	.14
.01	.26	.25	-.04	.44	.09	-.08	0	-.11	.09	0	.12
.02	.28	.34	-.02	.35	.08	-.09	0	-.07	.07	0	.11
.03	.25	.35	0	.36	.08	-.08	0	-.06	.07	0	.10
.04	.22	.26	0	.44	.05	-.08	0	-.10	.06	0	.09
.06	.19	.38	0	.38	.07	0	-.08	-.06	.08	0	.06
.09	.28	.28	0	.36	.04	0	0	-.06	.04	0	.07
.12	.33	.24	0	.29	.02	0	0	-.08	.03	0	.09
.26	.36	.21	0	.29	.01	0	0	-.08	.02	0	.09
.32	.36	.18	.01	.31	0	0	0	-.08	.01	0	.08
.37	.41	.25	0	.22	0	0	0	-.05	.01	0	.07
.42	.45	.24	0	.20	0	0	0	-.01	.01	0	.06
.43	.45	.24	0	.19	0	0	0	-.01	.01	0	.06
.51	.45	.24	0	.19	0	0	0	-.01	.01	0	.06
.63	.46	.23	.01	.18	0	0	0	-.01	.01	0	.05
.72	.44	.24	.02	.18	0	0	0	0	.01	0	.04
1.0	.44	.24	.02	.18	0	0	0	0	.01	0	.04
∞	0	0	0	0	0	0	0	0	0	0	0

Tableau 5.3: Paramètres estimés par la méthode L_1-ridge.

Par contre, les trois coefficients qui correspondent aux variables X_1, X_2 et X_4 sont très nettement non nuls. Il arrive même que certains de ces coefficients augmentent avant de se stabiliser. Ces variables semblent donc les plus importantes et devraient être prises en considération dans la sélection d'un modèle avec moins de variables. De manière à pouvoir comparer les résultats obtenus avec cette méthode, nous avons utilisé le logiciel Minitab pour établir toutes les régressions possibles. Le tableau (5.4) contient les deux modèles possédant la plus grande valeur du coefficient de détermination R^2 selon la taille des sous-ensembles des variables explicatives. Certaines variables apparaissent clairement dans pratiquement tous les modèles. Ils s'agit des variables X_4, X_2 et X_1 respectivement. D'autres n'apparaissent que très rarement, comme les variables X_3, X_5, X_6, X_7 et X_{10}. Finalement, les variables X_8, X_9 et X_{11} sont présentes dans les modèles avec plus de 4 ou 5 variables. Il faut remarquer que le coefficient de détermination dépasse 90% à partir de modèles avec seulement deux variables (X_1 et X_4) ou (X_2 et X_4). Dans le cas d'un modèle avec trois variables, on retrouve le modèle à trois variables (X_1, X_2 et X_4) obtenu avec la méthode L_1-ridge, avec un très bon coefficient de détermination de 93.5 %.

Var.	R^2	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X 10	X 11
1	86.3				X							
1	85.4		X									
2	92.8	X			X							
2	90.3		X		X							
3	93.5	X	X		X							
3	93.2	X			X							X
4	93.8	X	X		X			X				
4	93.8	X	X		X				X			
5	94.4	X			X				X	X		X
5	94.3	X	X		X				X			X
6	94.7	X	X		X				X	X		X
6	94.5	X			X		X		X	X		X
7	94.8	X	X		X	X			X	X		X
7	94.8	X	X		X				X	X	X	X
8	95.0	X	X		X	X		X	X	X		X
8	95.0	X	X		X	X	X		X	X		X
9	95.0	X	X	X	X	X		X	X	X		X
9	95.0	X	X		X	X	X		X	X	X	X
10	95.1	X	X	X	X	X	X		X	X	X	X
10	95.1	X	X	X	X	X	X	X	X	X		X
11	95.1	X	X	X	X	X	X	X	X	X	X	X

Tableau 5.4: Meilleurs sous-ensembles pour la régression.

Les résultats obtenus par la méthode basée sur l'estimation L_1-ridge conduisent ainsi à des conclusions très semblables à la méthode consistant à examiner toutes les régressions possibles (dans ce cas, $2^{11} = 2048$ régressions ont été examinées, seules les deux meilleures pour chaque sous-ensemble de variables apparaissent dans le tableau (5.4)). Nous avons vu dans le chapitre 4 que l'estimation L_1-ridge se ramenait à un problème d'estimation L_1 sur les données augmentées. Lorsque le nombre de variables devient très important, il n'est plus possible d'examiner toutes les régressions possibles, alors que l'estimation L_1-ridge peut toujours s'appliquer dans ce cas. D'autre part, le coefficient de détermination peut être sensible aux données aberrantes alors que l'estimation L_1-ridge est moins sensible à ce problème.

Avant de passer au deuxième exemple d'application, donnons encore les résultats obtenus pour le modèle avec les trois variables X_1, X_2 et X_4 dans le tableau (5.5).

Variable	Coef.	Stdev.	t-ratio	p
X_1	0.3760	0.1131	3.32	0.003
X_2	0.2106	0.1442	1.46	0.158
X_4	0.4261	0.1255	3.40	0.002

Tableau 5.5: Résultats obtenus pour le modèle à trois variables.

Cette fois, les paramètres estimés sont très significativement non nuls pour les variables X_1 (Taxes) et X_4 (Surface habitable), alors que pour X_2 (Nombre de salles de bain) cette conclusion est moins marquée. Les deux variables X_1 et X_4 semblent ainsi particulièrement importantes pour expliquer la variation du prix de vente d'une maison. La sélection d'un bon modèle en régression linéaire est fondamentale. Dans cet exemple d'application des estimateurs L_1-ridge, nous avons pu réduire considérablement le nombre de variable (réduisant ainsi le problème lié à la multicolinéarité) sans pour autant sacrifier le pouvoir explicatif du modèle. Finalement, en appliquant la méthode des moindres carrés sur les données non standardisées, on obtient l'équation de régression suivante pour le modèle avec les trois variables X_1, X_2 et X_4

$$\hat{Y} = -0.626 + 1.867X_1 + 7.113X_2 + 10.921X_4.$$

5.3 Sélection L_1-ridge de modèles (taux d'accidents)

Le deuxième exemple d'application des estimateurs L_1-ridge pour sélectionner un modèle se fera sur un ensemble de données avec 13 variables pour expliquer le taux d'accidents (Y) dans l'etat du Minnesota (Weisberg (1985)). Les données comprennent 39 observations faites sur des tronçons d'autoroute. Les variables retenues sont les suivantes

X_1 : longueur du tronçon (en miles)
X_2 : trafic moyen quotidien (en milliers de véhicules)
X_3 : pourcentage du volume de camions par rapport au volume total
X_4 : vitesse limite autorisée (en miles/heure)
X_5 : largeur de la piste (en pieds)
X_6 : largeur de la piste d'arrêt d'urgence (en pieds)
X_7 : nombre de changements de pistes libres (par mile sur le tronçon)
X_8 : nombre de changements de pistes signalées (par mile)
X_9 : nombre de points d'entrée sur l'autoroute (par mile sur le tronçon)
X_{10} : nombre total de pistes (dans les deux directions)
X_{11} : 1 s'il s'agit d'une autoroute fédérale inter-état, 0 sinon
X_{12} : 1 s'il s'agit d'une artère principale d'autoroute, 0 sinon
X_{13} : 1 s'il s'agit d'une artère majeure d'autoroute, 0 sinon.

Le tableau (5.6) contient l'ensemble de ces données. Les paramètres estimés par la méthode des moindres carrés sur ces données avec les valeurs p correspondant au test de Student sont donnés ci-dessous

Paramètre	β_0	β_1	β_2	β_3	β_4	β_5
estimé	13.66	-0.065	-0.004	-0.100	-0.124	-0.134
Valeur p	0.058	0.064	0.906	0.391	0.142	0.825

β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}
0.014	-0.475	0.713	0.067	0.027	0.543	-1.01	-0.548
0.931	0.714	0.186	0.130	0.926	0.756	0.370	0.579

On constate qu'avec le modèle comprenant l'ensemble des treize variables explicatives, aucun des paramètres estimés par la méthode des moindres carrés n'est significativement non nul à un seuil de 5%. En effet aucune valeur p n'est inférieure à 0.05.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	Y
4.99	69	8	66	12	10	1.20	0	4.6	8	1	0	0	4.68
16.11	73	8	60	12	10	1.43	0	4.4	4	1	0	0	2.86
9.75	49	10	60	12	10	1.54	0	4.7	4	1	0	0	3.02
10.65	61	13	65	12	10	0.94	0	3.8	6	1	0	0	2.29
20.01	28	12	70	12	10	0.65	0	2.2	4	1	0	0	1.61
5.97	30	6	55	12	10	0.34	1.84	24.8	4	0	1	0	6.87
8.57	46	8	55	12	8	0.47	0.70	11.0	4	0	1	0	3.85
5.24	25	9	55	12	10	0.38	0.38	18.5	4	0	1	0	6.12
16.79	43	12	50	12	4	0.95	1.39	7.6	4	0	1	0	3.29
8.26	23	7	50	12	5	0.12	1.21	8.2	4	0	1	0	5.88
7.03	23	6	60	12	10	0.29	1.85	5.4	4	0	1	0	4.20
13.28	20	9	60	12	2	0.15	1.21	11.2	4	0	1	0	4.61
5.40	18	14	60	12	8	0	0.56	16.2	2	0	1	0	4.80
2.96	21	8	60	12	10	0.34	0	5.4	4	0	1	0	3.85
11.75	27	7	66	12	10	0.26	0.60	7.9	4	0	1	0	2.69
8.86	22	9	60	12	10	0.68	0	3.2	4	0	1	0	1.99
9.78	19	9	60	12	10	0.20	0.10	11.0	4	0	1	0	2.01
5.49	9	11	50	12	6	0.18	0.18	8.9	2	0	1	0	4.22
8.63	12	8	65	13	6	0.14	0	12.4	2	0	1	0	2.76
20.31	12	7	60	12	10	0.05	0.99	7.8	4	0	1	0	2.56
40.09	15	13	55	12	8	0.05	0.12	9.6	4	0	1	0	1.89
11.81	8	8	60	12	10	0	0	4.3	2	0	1	0	2.34
11.39	5	9	50	12	8	0	0.09	11.1	2	0	1	0	2.83
22.00	5	15	60	12	7	0	0	6.8	2	0	1	0	1.81
3.58	23	6	40	12	2	0.56	2.51	53.0	4	0	0	1	9.23
3.23	13	6	45	12	2	0.31	0.93	17.3	2	0	0	1	8.60
7.73	7	8	55	12	8	0.13	0.52	27.3	2	0	0	1	8.21
14.41	10	10	66	12	6	0	0.07	18.0	2	0	0	1	2.93
11.54	12	7	45	12	3	0.09	0.09	30.2	2	0	0	1	7.48
11.10	9	8	60	12	7	0	0	10.3	2	0	0	1	2.57
22.09	4	8	45	11	3	0	0.14	18.2	2	0	0	1	5.77
9.39	5	10	55	13	1	0	0	12.3	2	0	0	1	2.90
19.49	4	13	55	12	4	0	0	7.1	2	0	0	1	2.97
21.01	6	12	56	10	8	0	0.10	14.0	2	0	0	1	1.84
27.16	2	10	55	12	3	0.04	0.04	11.3	2	0	0	1	3.78
14.03	3	8	50	12	4	0.07	0	16.3	2	0	0	1	2.76
20.53	1	11	55	11	4	0	0	9.6	2	0	0	1	4.27
20.06	3	11	60	12	8	0	0	9.0	2	0	0	0	3.05
12.91	1	10	55	12	3	0	0	10.4	2	0	0	0	4.12

Tableau 5.6: Ensemble de données relatives au taux d'accidents dans l'etat du Minnesota (USA).

Le tableau (5.7) fournit les résultats pour les données standardisées. On notera que les valeurs p du test de Student ne sont pas modifiées par cette standardisation. Que ce soit sur les données brutes ou standardisées, les conclusions sont les mêmes, à savoir qu'aucun coefficient n'est significativement non nul (à un seuil de 5%). On retrouve ainsi un problème semblable à celui rencontré dans l'application précédente.

Pour déterminer quelles sont les variables importantes dans ce modèle, nous comparons les résultats obtenus avec l'estimation L_1-ridge et ceux obtenus en examinant toutes les régressions possibles (8192 ici, puisqu'il y a 13 variables). Les paramètres estimés par la méthode L_1-ridge en fonction de k sont donnés dans le tableau (5.8). Certains coefficients sont très rapidement nuls et indiquent quelles variables sont peu importantes

Variable	Coef.	Stdev.	t-ratio	p
X_1	-0.2481	0.1279	-1.94	0.064
X_2	-0.0378	0.3181	-0.12	0.906
X_3	-0.1187	0.1360	-0.87	0.391
X_4	-0.3645	0.2406	-1.52	0.142
X_5	-0.0307	0.1372	-0.22	0.825
X_6	0.0216	0.2479	0.09	0.931
X_7	-0.0984	0.2656	-0.37	0.714
X_8	0.2276	0.1675	1.36	0.186
X_9	0.3124	0.1997	1.56	0.130
X_{10}	0.0183	0.1945	0.09	0.926
X_{11}	0.0297	0.2947	0.31	0.756
X_{12}	-0.2575	0.2819	-0.91	0.370
X_{13}	-0.1318	0.2346	-0.56	0.579

Tableau 5.7: Résultats obtenus par la méthode des moindres carrés.

dans le modèle. Il s'agit respectivement des variables X_2, X_6, X_{10}, X_{11}, X_{13}, X_7 et X_5. Les variables à retenir sont celles pour lesquelles les coefficients associés sont non nuls. D'un point de vue quantitatif, les coefficients les plus importants sont ceux liés aux variables X_4 et X_9. Viennent ensuite respectivement les variables X_8, X_1, X_3 et X_{12}. Ainsi, avec cette méthode, le modèle sélectionné pourrait être celui comprenant ces six variables. D'autres modèles comprenant un sous-ensemble de ces variables peuvent également être envisagés. Il serait alors souhaitable d'inclure les deux variables X_4 et X_9 dans cette sélection. Comme dans l'application précédente, nous présentons les résultats obtenus en examinant toutes les régressions possibles dans le tableau (5.9) de façon à pouvoir comparer les résultats. Rappelons que sur toutes les régressions examinées, seuls les deux meilleurs modèles du point de vue du coefficient de détermination sont retenus pour chaque taille des sous-ensembles des variables explicatives. On constate que le coefficient de détermination pour le modèle comprenant toutes les variables est de 76.1%. Ce coefficient n'augmente pratiquement plus à partir d'un modèle avec cinq ou six variables. A partir d'un modèle avec au moins quatre variables, le coefficient de détermination dépasse 70%. Il est donc raisonnable d'envisager un modèle avec entre quatre et six variables.

k	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}
.00	-.19	-.01	-.05	-.42	-.16	.03	-.17	.27	.23	.05	.13	-.26	-.13
.01	-.19	0	-.05	-.41	-.15	.03	-.16	.27	.24	.05	.12	-.26	-.13
.02	-.19	0	-.05	-.41	-.15	.03	-.16	.27	.24	.05	.11	-.26	-.13
.03	-.19	0	-.04	-.36	-.18	0	-.13	.26	.26	.01	.09	-.26	-.13
.04	-.19	0	-.04	-.36	-.18	0	-.11	.26	.26	.01	.09	-.22	-.13
.06	-.18	0	-.05	-.36	-.17	0	-.12	.27	.26	.01	.10	-.22	-.13
.07	-.18	0	-.03	-.37	-.18	0	-.03	.27	.26	0	0	-.23	-.13
.08	-.18	0	-.04	-.37	-.18	0	-.03	.27	.26	0	0	-.23	-.13
.12	-.17	0	-.04	-.41	-.10	0	-.06	.26	.07	0	0	-.25	-.13
.14	-.18	0	-.06	-.38	-.09	0	-.06	.18	.15	0	0	-.22	-.14
.15	-.17	0	-.06	-.39	-.09	0	-.05	.17	.15	0	0	-.21	-.02
.16	-.18	0	-.06	-.38	-.09	0	-.05	.18	.14	0	0	-.21	0
.18	-.18	0	-.06	-.39	-.09	0	-.05	.18	.14	0	0	-.21	0
.44	-.17	0	-.06	-.39	-.09	0	-.05	.18	.14	0	0	-.21	0
.57	-.18	0	-.08	-.36	-.14	0	0	.16	.17	0	0	-.14	0
.61	-.18	0	-.10	-.34	0	0	0	.15	.19	0	0	-.12	0
.70	-.18	0	-.11	-.29	0	0	0	.14	.22	0	0	-.08	0
.87	-.19	0	-.11	-.28	0	0	0	.14	.24	0	0	-.06	0
.95	-.17	0	-.09	-.26	0	0	0	.14	.26	0	0	-.04	0
∞	0	0	0	0	0	0	0	0	0	0	0	0	0

Tableau 5.8: Paramètres estimés par la méthode L_1-ridge.

Les résultats du tableau (5.9) sont particulièrement utiles pour déterminer quelles variables sont importantes dans la construction d'un modèle. Dans le cas où il n'y a qu'une variable explicative, les deux meilleurs modèles sont ceux comprenant soit la variable X_9 (Nombre de points d'entrée sur l'autoroute) soit la variable X_4 (Vitesse limite autorisée), c'est-à-dire les deux variables les plus importantes déterminées par la méthode L_1-ridge. Cependant le coefficient de détermination reste relativement faible dans ce cas, indiquant que d'autres variables sont importantes pour expliquer le taux d'accidents. En effet, outre ces deux variables, d'autres variables apparaissent dans pratiquement tous les modèles; à partir des modèles à deux variables, X_1 et X_3 sont pratiquement toujours retenues, alors qu'à partir des modèles à quatre et cinq variables, X_8 et X_{12} sont retenues. On retrouve ainsi les mêmes variables qu'avec la méthode L_1-ridge; le modèle avec les six variables X_1, X_3, X_4, X_8, X_9 et X_{12} fait également partie des deux meilleurs modèles à six variables avec un coefficient de détermination de 75.2%. Ce modèle a l'avantage d'avoir relativement peu de variables et permet d'expliquer aussi bien le taux d'accidents que le modèle avec l'ensemble des treize variables explicatives.

Var.	R^2	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X 10	X 11	X 12	X 13
1	56.6									X				
1	46.4				X									
2	65.2	X								X				
2	63.3			X						X				
3	70.1	X			X					X				
3	67.7			X	X					X				
4	71.8	X			X				X	X				
4	71.7	X		X	X					X				
5	74.5	X			X				X	X			X	
5	73.0	X			X				X	X		X		
6	75.2	X		X	X				X	X			X	
6	74.8	X			X		X		X	X			X	
7	75.6	X		X	X				X	X			X	X
7	75.3	X		X	X		X		X	X			X	
8	75.8	X		X	X			X	X	X			X	X
8	75.7	X		X	X	X			X	X			X	X
9	75.9	X		X	X	X		X	X	X			X	X
9	76.0	X		X	X			X	X	X		X	X	X
10	76.0	X		X	X	X		X	X	X		X	X	X
10	76.0	X		X	X		X	X	X	X		X	X	X
11	76.0	X		X	X	X	X	X	X	X		X	X	X
11	76.0	X	X	X	X	X		X	X	X		X	X	X
12	76.0	X	X	X	X	X		X	X	X	X	X	X	X
12	76.0	X	X	X	X	X	X	X	X	X		X	X	X
13	76.1	X	X	X	X	X	X	X	X	X	X	X	X	X

Tableau 5.9: Meilleurs sous-ensembles pour la régression.

5.4 Conclusion

Nous avons vu dans ce chapitre deux applications des estimateurs L_1-ridge. Pour les deux ensembles de données considérées, les résultats obtenus avec ces estimateurs permettent de sélectionner un modèle très satisfaisant avec les variables à disposition. En comparant ces résultats avec ceux obtenus en considérant toutes les régressions possibles, on retrouve les variables sélectionnées selon le critère du coefficient de déter-

mination. Il semble donc que les estimateurs L_1-ridge permettent de sélectionner les variables importantes dans les problèmes de régression comprenant un nombre élevé de variables explicatives et pour lesquels le problème de la multicolinéarité est souvent source d'inconvénients pour une utilisation directe de la méthode des moindres carrés. Comme nous l'avons vu dans le chapitre précédent, l'estimation L_1-ridge est également nettement moins sensible au problème des données aberrantes, ce qui est un avantage important par rapport à la méthode des moindres carrés. Cet aspect du problème ne s'est pas posé dans les ensembles de données considérées mais nous avons vu comment le coefficient de détermination pouvait être affecté par des données aberrantes. Dans ce cas, les résultats obtenus en considérant toutes les régressions possibles (qui sont basées sur la méthode des moindres carrés) peuvent être trompeurs. La méthode présentée dans ce chapitre devrait permettre d'éviter ce type de problèmes. L'un des avantages de la méthode L_1-ridge par rapport à la méthode classique de la sélection du meilleur modèle est le temps de calcul. Les algorithmes à disposition aujourd'hui pour le calcul des estimateurs L_1 et donc des estimateurs L_1-ridge sont performants. Pour l'utilisateur, dans les exemples étudiés dans ce chapitre, les résultats pour l'estimation L_1-ridge, comme pour l'estimation L_2 sont immédiats. Par contre, lorsqu'il s'agit d'examiner toutes les régressions possibles (2^p, où p est le nombre de variables explicatives), le temps de calcul peut devenir très important. Le logiciel Minitab demande d'ailleurs à l'utilisateur une confirmation lorsqu'il y a plus de 14 variables explicatives en l'avertissant que cela peut prendre beaucoup de temps! Mentionnons encore que s'il est possible de comparer les estimateurs L_1, L_2, ridge et L_1-ridge comme nous l'avons fait en utilisant la simulation, il n'est guère possible de les comparer de manière objective sur des ensembles de données puisque les vraies valeurs des paramètres ne sont pas connues. Quant à l'estimation ridge, elle fournit des paramètres qui ne sont pas nuls mais qui tendent vers 0 de manière relativement uniforme, rendant ainsi l'interprétation d'un modèle particulièrement difficile. Par contre, certains paramètres estimés par la méthode L_1-ridge sont très rapidement nuls, alors que d'autres restent non nuls indiquant ainsi les variables à prendre en considération dans le modèle. Il s'agit là d'une application originale des estimateurs L_1-ridge, qui permet ainsi de sélectionner un modèle.

CHAPITRE 6

CONCLUSION ET RECHERCHES FUTURES

La recherche effectuée dans ce travail nous a permis de mieux comprendre le comportement des estimateurs L_1, L_2, ridge et L_1-ridge lorsque les deux problèmes de la multicolinéarité et des données aberrantes se posaient simultanément. Nous avons ainsi pu comparé ces différentes méthodes d'estimation dans les modèles de régression linéaire multiple, en présence de ces deux problèmes. Alors que l'on trouve de nombreuses publications sur chaque problème pris séparément, la voie consistant à les étudier simultanément n'a pas connu le même succès. Notre approche est basée sur une combinaison des deux méthodes de régression L_1 et ridge, la première étant reconnue pour être peu sensible aux données aberrantes et la seconde utilisée lorsque le problème de la multicolinéarité se pose. Cette méthode d'estimation appelée L_1-ridge semble fournir une approche bien adaptée en présence de ces deux problèmes. En effet, par rapport aux estimateurs L_1, L_2 et ridge il est le plus performant lorsque les deux problèmes apparaissent simultanément. L'estimateur L_1-ridge a de plus l'avantage de pouvoir être calculé facilement sur les données augmentées et n'exige pas le recours à un algorithme itératif comme celui des moindres carrés pondérés. Sur le plan théorique, nous avons trouvé des formules pour calculer cet estimateur dans des modèles de régression simple. En régression linéaire multiple, nous avons démontré une propriété importante qui caractérise cet estimateur. Finalement nous avons vu comment cette méthode pouvait être appliquée

dans la sélection de modèles, une démarche particulièrement importante en analyse de régression.

Ces résultats nous conduisent à penser que de nouvelles voies de recherche peuvent être envisagées. En effet, l'idée consistant à combiner deux méthodes d'estimation prévues pour chaque problème respectivement offre certainement des perspectives de recherche intéressantes : de nombreuses méthodes d'estimation ont été proposées dans la littérature, notamment en statistique robuste, comme par exemple les L-estimateurs, M-estimateurs, R-estimateurs, l'estimateur des moindres carrés tronqués ou encore l'estimateur de la moindre médiane des carrés. La combinaison de l'un de ces estimateurs robustes avec l'estimateur ridge pourrait ainsi fournir de nouveaux estimateurs dont le comportement en présence des deux problèmes évoqués serait ensuite comparé à celui de l'estimateur L_1-ridge. D'autre part, une autre voie de recherche prometteuse et importante est celle liée à la sélection de variables permettant de construire un modèle. Dans ce cas, il serait également possible de comparer les différentes méthodes en simulant des ensembles de données pour lesquels certains coefficients de l'équation de régression sont nuls indiquant ainsi quelles variables ne doivent pas être retenues dans le modèle. Finalement, ces estimateurs construits à partir de deux méthodes d'estimation pourraient également être utilisés pour détecter les données aberrantes lorsque le problème de multicolinéarité se pose. Il faut cependant noter qu'il est en général recommandé de traiter d'abord le problème de la multicolinéarité avant de s'intéresser aux autres problèmes qui peuvent survenir dans les données. Dans ce sens, la sélection d'un bon modèle est cruciale; c'est pourquoi nous avons mis l'accent sur cet aspect du problème dans le chapitre consacré aux applications. Il existe ainsi plusieurs directions de recherche possibles qui permettraient de faire avancer les connaissances dans ce domaine.

Pour conclure, il nous semble important d'insister sur le fait que la méthode d'estimation la mieux adaptée dépend essentiellement de la nature des problèmes rencontrés. Lorsque les deux problèmes de la multicolinéarité et des données aberrantes se posent simultanément, les résultats obtenus en comparant les différentes méthodes indiquent que l'estimation L_1-ridge proposée ici est la mieux adaptée. L'estimateur L_1-ridge a en outre l'avantage, grâce aux propriétés qui le caractérisent, de pouvoir s'appliquer à la sélection de modèles.

BIBLIOGRAPHIE

[1] Abdelmalek, N.N. (1980). L_1 Solution of Overdetermined Systems of Linear Equations. *ACM Transactions on Mathematical Software*, **6**, 220-227.

[2] Adrain, R. (1808). Research concerning the probabilities of the errors which happen in making observations. *Analyst*, **1**, 93-109.

[3] Appa, G. and Smith, C. (1973). On L_1 and Chebyshev Estimation. *Mathematical Programming*, **5**, 73-87.

[4] Armstrong, R.D. and Kung, M.T. (1978). Algorithm AS 132 : Least Absolute Value Estimates for a Simple Linear Regression Problem. *Applied Statistics*, **27**, 363-366.

[5] Armstrong, R.D., Frome, E.L. and Kung, D.S. (1979). A Revised Simplex Algorithm for the Absolute Deviation Curve Fitting Problem. *Commun. Statist.-Simula. Computa.*, **B8(2)**, 175-190.

[6] Arthanari, T.S. and Dodge, Y. (1981). *Mathematical Programming in Statistics*. John Wiley, Interscience Division, New York.

[7] Arthanari, T.S. and Dodge, Y. (1993). *Mathematical Programming in Statistics*. John Wiley Classics Library Edition, New York.

[8] Barrodale, I. and Young, A. (1966). Algorithms for Best L_1 and L_∞ Linear Approximations on a Discrete Set. *Numerische Mathematik*, **8**, 295-306.

[9] Barrodale, I. and Roberts, F.D.K. (1973). An Improved Algorithm for Discrete L_1 Linear Approximation. *SIAM J.Numer.Anal.*, **10**, 839-848.

[10] Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478 : Solution of an Overdetermined System of Equations in the l_1 norm [F4]. *Communications of the ACM*, **17**, 319-320.

[11] Bassett, G.W. and Koenker, R.W. (1978). Asymptotic Theory of Least Absolute Error Regression. *Journal of the American Statistical Association*, **73**, 618-622.

[12] Belsley, D., Kuh, E. and Welsh, R.E. (1980). *Regression Diagnostics*. John Wiley, New York.

[13] Bloomfield, P. and Steiger, W. (1980). Least Absolute Deviations Curve-Fitting. *SIAM J.Sci.Stat.Comput.*, **1**, 290-301.

[14] Bloomfield, P. and Steiger, W.L. (1983). *Least absolute deviations, Theory, Applications and Algorithms*. Birkäuser, Boston.

[15] Boscovich, R.J. (1757). De litteraria expeditione per pontificam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii*, **4**, 353-396.

[16] Boscovich, R.J. (1760). De recentissimis graduum dimensionibus, et figura, ac magnitudine terrae inde derivanda. Philosophiae Recentioris, a Benedicto Stay in Romano Archigynasis Publico Eloquentare Professore, versibus traditae, Libri X, cum adnotianibus et Supplementis P. Rogerii Boscovich, S.J., Tomus II, pp. 406-426, esp. 420-425. Romae.

[17] Charnes, A., Cooper, W.W. and Fergusson, R.O. (1955). Optimal Estimation of Executive Compensation by Linear Programming. *Management Science*, **2**, 138-151.

[18] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity analysis in linear regression*. John Wiley, New York.

[19] Dielman, T.E. and Pfaffenberger, R.C. (1984). Computational algorithms for calculating least absolute value and Chebyshev estimates for multiple regression. *American Journal of Mathematical and Management Sciences*, **4**, 169-197.

[20] Dielman, T.E. (1992). Computational algorithms for least absolute value regression. In L_1-Statistical Analysis and Related Methods, Dodge, Y. editor, 311-326. North-Holland, Amsterdam.

[21] Edgeworth, F.Y. (1887). On Observations Relating to Several Quantities. *Philosophical Magazine, London*, **5th serie**, 222-223.

[22] Eisenhart, C. (1961). *Boscovich and the Combination of Observations. Roger Joseph Boscovich, S.J., F.R.S., 1711-1787 : Studies of his Life and Work on the 250th Anniversary of his Birth.* (L.L. White, Ed.). London : Allen and Unwin, Ltd., 200-212.

[23] Eisenhart, C. (1968). Gauss, Carl Friedrich. In *International Encyclopedia of the Social Sciences*, 74-81. Reprinted 1978 in *International Encyclopedia of Statistics.* (With additions), **1**, 378-386. Macmillan and Free Press, New York.

[24] Galilei, Galileo (1632). *Dialogo sopra i due massimi sistemi del mondo : Ptolemaico e Copernicano.* Landini, Florence. (English translation, Dialogue concerning the two chief world systems, Ptolemaic and Copernican, by Stillman Drake. Univ. of Calif. Press, Berkeley, 1953.

[25] Gauss, C.F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Frid. Perthes et I.H. Besser, Hamburgi. Reprinted 1906 in *Werke*, Band VII, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 1-280.

[26] Gauss, C.F. (1823). Theoria combinationis observationum erroribus minimis obnoxiae. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores*, **5**. Reprinted 1880 in *Werke*, Band IV, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 3-53, 95-104.

[27] Gauss, C.F. (1828). Supplementum theoriae combinationis observationum erroribus minimis obnoxiae. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores*, **6**. Reprinted 1880 in *Werke*, Band IV, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 57-93, 104-108.

[28] Gentle, J.E., Narula, S.C. and Sposito, V.A. (1987). Algorithms for Unconstrained L_1 Linear Regression. In *Statistical Data Analysis Based on the L_1- Norm*, edited by Y.Dodge, Elsevier/North-Holland, Amsterdam, 83-94.

[29] Goldstine, H. (1977). *A history of Numerical Analysis from the 16th through the 19th Century*. Springer, New York.

[30] Harter, H.L. (1974a). The Method of Least Squares and some Alternatives I. *International Statistical Review*, **42**, 147-174.

[31] Harter, H.L. (1974b). The Method of Least Squares and some Alternatives II. *International Statistical Review*, **42**, 235-264.

[32] Harter, H.L. (1975a). The Method of Least Squares and some Alternatives III. *International Statistical Review*, **43**, 1-44.

[33] Harter, H.L. (1975b). The Method of Least Squares and some Alternatives IV. *International Statistical Review*, **43**, 125-190.

[34] Harter, H.L. (1975c). The Method of Least Squares and some Alternatives V. *International Statistical Review*, **43**, 269-278.

[35] Harter, H.L. (1976). The Method of Least Squares and some Alternatives VI. *International Statistical Review*, **44**, 113-159.

[36] Hoerl, A.E. (1962). Application of ridge analysis to regression problems. *Chem. Eng. Progress*, **58**, 54-59.

[37] Hoerl, A.E. and Kennard, R.W. (1970a). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67.

[38] Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82.

[39] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge Regression: Some Simulations. *Communications in Statistics*, **4**, 105-123.

[40] Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics*, **A 6**, 813-827.

[41] Huber, P.J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Mathematical Statistics*, **1**, 799-821.

[42] Karst, O.J. (1958). Linear Curve Fitting Using Least Deviations. *American Statistical Association Journal*, **53**, 118-132.

[43] Klingman, D. and Mote J. (1982). Generalized Network Approaches for solving Least Absolute Value and Tchebycheff Regression Problems. *TIMS/Studies in Management Sciences*, **19**, 53-66.

[44] Koenker, R.W. and Bassett, G.W. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

[45] Koenker, R.W. and d'Orey, V. (1987). Computing regression quantiles. *Appl. Statist.*, **36**, 383-393..

[46] Laplace, P.S. (1786). Mémoire sur la figure de la terre. Mémoires de l'Académie royale des Sciences de Paris, **Année 1783**, 17-46. Reprinted in Oeuvres complètes de Laplace, Vol. 11, pp. 3-32. Gauthier-Villars, Paris, 1895.

[47] Laplace, P.S. (1793). Sur quelques points du système du monde. *Mémoires de l'Académie royale des Sciences de Paris*, **Année 1789**, 1-87. Reprinted in *Oeuvres complètes de Laplace*, **11**, 477-458. Gauthier-Villars, Paris, 1895.

[48] Laplace, P.S. (1812). *Théorie analytique des Probabilités.* Third edition with new introduction and three supplements. Paris 1820 ; reprinted as Vol. VII of *Ouvres de Laplace.* Paris, 1847. National edition, Gauthier-Villars. Paris, 1886.

[49] Legendre, A.M. (1805). Nouvelles méthodes pour la détermination des orbites et des comètes. Courcier, Paris. (Appendice sur la méthode des moindres carrés, pp. 72-80).

[50] McKean, J.W. and Schrader, R.M. (1987). Least absolute errors analysis of variance. In *Statistical Data Analysis Based on the L_1-norm and Related Methods, Dodge.Y. editor, 297-305.* North Holland, Amsterdam.

[51] Müller, M. (1992). *A comparative study of L_1-norm based simple and multiple regression algorithms.* Report, Postgrade in Statistics, University of Neuchâtel, 1-56.

[52] Narula, S.C. and Wellington, J.F. (1977). An Algorithm for the Minimum Sum of Weighted Absolute Errors Regression. *Commun. Statist.-Simula. Computa.*, **B6(4)**, 341-352.

[53] Nyquist, H. (1985). Ridge type M-estimators. In *Linear Statistical Inference.* Edited by Calinski, T. and Klonecki, W., Springer, Berlin (1985), 246-258.

[54] Pfaffenberger, R.C. and Dielman, T.E. (1990). A Comparison of Regression Estimators when Both Multicollinearity and Outliers are Present. In: *Robust Regression: Analysis and Applications*, edited by Lawrence, K.D. and Arthur, J.L., Marcel Dekker, Inc., New York and Basel (1990), 243-270.

[55] Plackett, R.L. (1972). The discovery of the method of least squares. *Biometrika*, **59**, 239-251. Reprinted in *Studies in History of Statistics and Probability.* (M.G. Kendall and R.L. Plackett, eds.) Griffin, London (1977).

[56] Rhodes, E.C. (1930). Reducing Observations by the Method of Minimum Deviations. *Philosophical Magazine, 7th serie, London*, **9**, 974-992.

[57] Ronchetti, E. (1987). Bounded Influence Inference in Regression: A Review. In *Statistical Data Analysis Based on the L_1- Norm*, edited by Y.Dodge, Elsevier/North-Holland, Amsterdam, 65-80.

[58] Sadovski, A.N. (1974). Algorithm AS74 : L_1 Norm Fit of a Straight Line. *Appl. Stat.*, **23**, 244-248.

[59] Seal, H.L. (1967). The historical development of the Gauss linear model. *Biometrika*, **54**, 1-24. Reprinted in *Studies in History of Statistics and Probability.* (E.S. Pearson and M.G. Kendall, eds.) Griffin, London.

[60] Sheynin, O.B. (1979). C.F. Gauss and the theory of errors. *Archive for History of Exact Science*, **20**, 21-72.

[61] Sposito, V.A. and Smith, W.C. (1976). On a Sufficient Condition and a Necessary Condition for L_1 Estimation. *Appl. Stat.*, **25**, 154-157.

[62] Sprott, D.A. (1978). Gauss's contributions to statistics. *Historia Mathematica*, **5**, 183-203.

[63] Stigler, S.M. (1977). An attack on Gauss, published by Legendre in 1820. *Historia Mathematica*, **4**, 31-35.

[64] Stigler, S.M. (1978). Francis Ysidro Edgeworth, Statistician. *Journal of the Royal Statistical Society*, **Serie A, 141**, 287-322.

[65] Stigler, S.M. (1978). Mathematical statistics in the early states. *Annals of Statistics.*, **6**, 239-265.

[66] Stigler, S.M. (1981). Gauss and the Invention of Least Squares. *Annals of Statistics*, **9**, 465-474.

[67] Usow, K.H. (1967). On L_1 Approximation I: Computation for Continuous Functions and Continuous Dependence. *SIAM J.Numer.Anal.*, **4**, 70-88.

[68] Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus.* Statistics and Computing, Springer-Verlag, New York.

[69] Wagner, H.M. (1959). Linear Programming Techniques for Regression Analysis. *Journal of the American Statistical Association*, **54**, 206-212.

[70] Weisberg, S. (1985). *Applied linear regression.* John Wiley, New York.

[71] Wesolowski, G.O. (1981) A New Descent Algorithm for the Least Absolute Value Regression Problem. *Commun. Statist.-Simula. Computa.*, **B10(5)**, 479-491.