ESTIMADORES DE REGRESSÃO COM ALTO PONTO DE RUPTURA E DETECÇÃO DE MÚLTIPLAS OBSERVAÇÕES DISCREPANTES

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra. Ysela Dominga Agüero Palacios e aprovada pela Comissão Julgadora.

Campinas, 22 de Fevereiro de 1994.

Profa. Dra. Gabriela Stangenhaus

Dissertação apresentada ao Instituto de Matemática, estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em ESTATÍSTICA.

A minha família.

AGRADECIMENTOS

- A Profa. Dra. Gabriela Stangenhaus pela competente orientação e incentivo.

- Ao Prof. Dr. Euclides Lima Filho pelas excelentes aulas de regressão do ponto de vista geométrico.

- A Profa. Mg. Edith Seier por ser responsável de grande parte de minha formação académica.

- Ao CNPq e UNICAMP pelo apoio financiero.

- A os amigos que direta ou indiretamente contribuiram para a elaboração deste trabalho através de críticas e sugestões ou simplesmente estiveram de meu lado dando apoio.

- À baixinha Bárbara de Holanda Texeira por aminorar a saudade da família e da patria.

RESUMO

Os métodos de diagnósticos de observações discrepantes na análise de regressão linear múltipla baseiam-se na eliminação de apenas uma observação de cada vez. Existem também métodos nos quais se elimina mais de uma observação de cada vez, mais são pouco aplicados devido aos problemas combinatorios envolvidos.

Por outro lado, existem conjuntos de dados com um padrão de múltiplas observações discrepantes os quais não são revelados pelos métodos de eliminação de uma observação de cada vez. Nestes casos dizemos que aconteceu um problema de "mascaramento". Neste trabalho estudamos métodos exploratórios de identificação de múltiplas observações discrepantes usando estimadores com alto ponto de ruptura, isto é, estimadores que além de não serem afetados no caso de existir múltiplas observações discrepantes no conjunto de dados, sejam úteis para identificá-los.

SUMARIO

CAPíTULO 1 : Introdução 1

CAPíTULO 2 : Alguns Conceitos Básicos de Robustez 6

2.1 Introdução 6
2.2 Função de Influência 7
2.2.1 Aproximação da Função de Influência para amostras Finitas 7
2.3 Robustez Qualitativa 9
2.4 Ponto de Ruptura 10
2.5 Relação entre Função de Influência, Robustez Qualitativa e Ponto de Ruptura 16

CAPíTULO 3 : Observações Discrepantes e Diagnósticos Baseados no Método de Mínimos Quadrados 18

3.1 Introdução 18
3.2 Observações Discrepantes 20
3.2.1 Objetivos da Detecção de Observações Discrepantes . 22
3.2.2 Origem das Observações Discrepantes 22
3.3 Observações discrepantes na Análise de Regressão Linear Múltipla 23
3.4 Diagnósticos em regressão Baseados no método de Mínimos Quadrados 25
3.4.1 A matriz de Projeção H 26
3.4.2 Análise de Resíduos 29
3.4.3 Outros Diagnósticos de Influência 30
3.4.4 Sumário 33

CAPÍTULO 4 : Obtenção de Estimadores com Alto Ponto de Ruptura 35

4.1 Introdução 35
4.2 Definição de Estimadores de Regressão Linear Norma L_p 36
4.2.1 Existência e Unicidade dos estimadores Norma L_p 37
4.2.2 Representação Geométrica dos estimadores Norma L_p 39
4.3 Relação entre os estimadores Norma L_p, M-estimadores e de Máxima verossimilhança 43
4.4 Construção de Estimadores com Alto Ponto de Ruptura . 45
4.4.1 Exemplos de Estimadores com Alto Ponto de Ruptura 47
4.5 Sumário 53

CAPÍTULO 5 : Elipsóide de Volume Mínimo e sua Utilização na . . . Identificação de Observações Discrepantes 55

5.1 Introdução 55
5.2 Identificação de Observações Discrepantes em conjuntos de Dados Multivariados 56
5.3 Estimação do Elipsóide de Volume Mínimo 60
5.3.1 Ponto de Ruptura do estimador Elipsóide de Volume Mínimo 62
5.4 Diagnósticos de Observações Discrepantes Baseados em Métodos Robustos 63
5.5 Sumário 67

CAPÍTULO 6 : Algoritmos 68

6.1 Introdução 68
6.2 Conjuntos Elementares 69
6.2.1 Algumas Observações sobre os Conjuntos Elementares 70
6.2.2 Número de Conjuntos Elementares Investigados pelos Algoritmos de Reamostragem quando n e k são grandes 72
6.3 Aplicações do Método de reamostragem de Conjuntos . . . Elementares 76
6.3.1 Estimação de Parâmetros no modelo de Regressão Linear Múltipla. 77

6.3.2 Estimação de Parâmetros no modelo de Posição Multivariado 78
6.4 Algoritmos Baseados em Conjuntos Elementares 80
6.4.1 Programa Computacional PROGRESS. 80
6.4.2 Programa Computacional MINVOL 82
6.4.3 Algoritmo Exato de Stromberg 83
6.4.4 Algoritmo de Conjuntos Factíveis, FSA 86
6.5 Sumário 90

CAPíTULO 7 : Aplicação 92

COMENTARIOS FINAIS 113

APÊNDICE 114

BIBLIOGRAFIA 120

CAPITULO 1

INTRODUÇÃO

Consideremos o modelo de regressão linear múltipla

$$Y = \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_k X_k + \varepsilon \qquad (1.1)$$

onde Y é uma variável aleatória (resposta), X_1, X_2, ..., X_k são as variáveis regressoras ou variáveis independentes, θ_1, θ_2, ..., θ_k são parâmetros desconhecidos, $\theta_j \in \mathbb{R} \quad \forall \quad j = 1, 2, \ldots, k$ e ε é um erro aleatório com distribuição de probabilidade F, onde $F \in \{F_\tau : \tau \in \mathbb{T}\}$.

Seja o conjunto de observações $Z = \left\{z_1, z_2, \ldots, z_n\right\}$, onde z_i é um vetor k+1 dimensional com elementos $z_i = \left(y_i, x_{i1}, \ldots, x_{ik}\right)$ que se relacionam segundo o modelo (1.1). Expressando-o matricialmente temos

$$y = X\theta + \varepsilon, \qquad (3.1)$$

onde $y_{n\times 1}$ é o vetor de respostas observadas, $X_{n\times k}$ a matriz do modelo, $\theta_{k\times 1}$ o vetor de parâmetros desconhecidos, $\varepsilon_{n\times 1}$ o vetor de erros

aleatórios independentes, e identicamente distribuídos com distribuição acumulada F. Supoe-se que o vetor de resíduos ou erros é resultado de variações aleatórias na população considerada. Estas variações podem ser efeitos devidos ao modelo suposto, erros de medição ou observação, variáveis regressoras não consideradas no modelo, etc.

Em geral X_1, X_2, ..., X_k são variáveis aleatórias, no entanto em regressão essas variáveis são consideradas fixas e todos os resultados que se derivam são válidos sob esta suposição.

Existe uma infinidade de métodos de estimação dos parâmetros do modelo (1.1). Alguns desses métodos consistem em obter estimações minimizando uma certa função do vetor de resíduos. Boscovich em 1757 propôs um método de estimação dos parâmetros de um modelo de regressão linear simples minimizando a soma dos valores absolutos dos resíduos,

$$\underset{\theta \in \mathbb{R}^2}{\text{Min}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{2} \theta_j x_{ij} \right| . \qquad (1.3)$$

Este método não foi muito usado naquela época devido às dificuldades computacionais, mesmo para modelos relativamente simples. Legendre em 1805 e posteriormente Gauss em 1808 propuseram estimadores para os parâmetros do modelo de regressão linear minimizando a soma de quadrados dos resíduos, isto é,

$$\underset{\theta \in \mathbb{R}^k}{\text{Min}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \theta_j x_{ij} \right)^2 . \qquad (1.4)$$

Este método ainda é muito popular pela facilidade dos cálculos e pelo fato de que quando os resíduos são variáveis aleatórias independentes, e identicamente distribuídas com distribuição normal, o estimador de Gauss e Legendre denominado *estimador de mínimos quadrados* é o estimador não viciado de mínima variância, aliás, ele é equivalente ao estimador de máxima verossimilhança. Na prática, essas suposições nem

sempre são atendidas. Por exemplo, pode existir no conjunto de dados uma certa proporção de observações que não provêm da distribuição normal mas de uma distribuição com caudas mais longas (Laplace, Cauchy, etc.), ou que existam problemas de multicolinearidade, dependência dos erros, etc. O não cumprimento das suposições manifesta-se em muitos casos através da discrepância de uma ou mais observações com respeito ao conjunto de dados. Foi amplamente comprovado que estas observações discrepantes podem distorcer os resultados da análise e consequentemente as conclusões. Portanto, precisamos utilizar métodos de diagnósticos para verificarmos se as suposições estabelecidas na formulação do modelo estão realmente sendo atendidas.

Nos últimos anos desenvolveram-se métodos de diagnósticos de observações discrepantes e influentes baseados nos resíduos do ajuste e características da matriz X. Assim por exemplo Belsley, Kuh e Welsch (1980), Cook e Weisberg (1982) e outros, propuseram e estudaram diferentes tipos de diagnósticos baseados no método de mínimos quadrados. Muitos desses diagnósticos já formam parte dos pacotes computacionais desenvolvidos para análise de regressão. Porém, a maioria deles são úteis somente para identificar observações discrepantes individuais, e em muitos casos práticos apenas uma observação fortemente discrepante pode mascarar outras que também estão influenciando no ajuste. Este problema é conhecido como *efeito de mascaramento* e existem métodos de diagnósticos baseados no ajuste de mínimos quadrados para estudar estes casos, mas eles são pouco factíveis de serem implementados computacionalmente.

Uma forma alternativa para detecção de múltiplas observações discrepantes e influentes, que produz bons resultados, é através da utilização de métodos robustos. Existem diferentes critérios de robustez que são usados tanto para construir, quanto para avaliar estimadores (Hampel (1971)).

Um dos objetivos deste trabalho é estudar estimadores robustos, no sentido de não serem muito afetados se o conjunto de dados

contém uma certa proporção de observações discrepantes. Esta propriedade do estimador é chamada de *ponto de ruptura*.

Pela estrutura dos dados envolvidos na análise de regressão, a discrepância de uma observação pode dever-se às discrepâncias na direção das variáveis regressoras, da variável resposta, ou em ambas direções. Logo, precisamos de métodos com alto ponto de ruptura tanto para estimação dos parâmetros de regressão quanto para obtermos estatísticas que detectam discrepâncias na direção da matriz do modelo X.

Outro objetivo, é apresentar métodos exploratórios robustos para identificação de múltiplas observações discrepantes. Estes diagnósticos são muito úteis e podem ser usados conjuntamente com aqueles produzidos pelo método de mínimos quadrados.

Os diagnósticos baseados em estimadores com alto ponto de ruptura são menos usados quando comparados com aqueles baseados no método de mínimos quadrados. Isso pode se dever à dificuldade computacional envolvida no cálculo dos primeiros. Um dos principais problemas a serem resolvidos para que esses diagnósticos robustos façam parte das ferramentas de análise do pesquisador é a produção de algorítmos que utilizem pouco espaço e tempo computacional.

No capítulo 2, introduzimos alguns conceitos de robustez que mesmo não sendo utilizados no capítulos seguintes, consideramos necessários para uma melhor comprensão do estimador aqui eleito para detecção de observações discrepantes.

No capítulo 3 abordamos o problema das observações discrepantes na análise de regressão linear múltipla. Neste mesmo capítulo, fizemos uma ligeira revisão dos métodos de diagnósticos de observações discrepantes que geralmente encontram-se disponíveis no pacotes computacionais usados para análise de regressão pelo método de mínimos quadrados.

No capítulo 4 estudamos as características dos estimadores de

regressão norma L_p. Em seguida apresentamos uma família de estimadores de regressão linear múltipla com alto ponto de ruptura, cujos membros são obtidos mudando a geometria das bolas associadas aos estimadores norma L_p. Um membro desta família é o estimador mínima mediana dos quadrados dos resíduos (LMS), proposto por Rousseeuw (1984) e posteriormente estudado e implementado computacionalmente por Rousseeuw e Leroy (1987) e outros.

No capítulo 5 apresentamos métodos robustos de estimação dos parâmetros de posição e dispersão multivariados. A finalidade do mesmo é apresentar uma distância robusta alternativa à distância de Mahalanobis e consequentemente aos elementos da diagonal da matriz de projeção usados na análise de regressão pelo método de mínimos quadrados.

No capítulo 6, apresentamos alguns algoritmos que foram desenvolvidos para obtenção de estimativas com alto ponto de ruptura.

No capítulo 7, apresentamos um exemplo comparando os métodos exploratórios baseados no ajuste de mínimos quadrados e os robustos.

Finalmente incluimos um apêndice com alguns dos conceitos que são usados no transcurso deste trabalho.

CAPITULO 2

ALGUNS CONCEITOS BÁSICOS DE ROBUSTEZ

2.1 INTRODUÇÃO

O termo robustez introduzido por George Box (1953) é usado para denominar os procedimentos estatísticos que são resistentes diante de desvios das suposições feitas na formulação do modelo matemático. Em um sentido amplo, e não formal, a estatística robusta está relacionada com o fato de que muitas suposições (normalidade, independência, linearidade, etc.), que geralmente são feitas na formulação de um modelo estatístico, são somente aproximações da realidade e que com frequência não se verificam nos problemas práticos.

Antes de começarmos o estudo de estimadores robustos úteis na detecção de dados discrepantes na análise de regressão, conceituaremos matematicamente o que significa dizer que um estimador é robusto em relação à presença de dados discrepantes, o que é robustez diante de pequenos desvios do modelo suposto, e o que significa dizer que um estimador é mais robusto que outro. Hampel (1971) apresenta os conceitos de Robustez Qualitativa, Função de Influência e Ponto de

ruptura, nos quais nos basaremos para tentarmos responder às perguntas anteriores.

2.2. FUNÇÃO DE INFLUÊNCIA

Seja um funcional definido em algum conjunto de medidas de probabilidade cujos elementos, F, estão definidos sobre $\mathbb{R}$, e seja δ_z uma medida de probabilidade com massa um no ponto $z \in \mathbb{R}$.

A função de influência é definida como

$$IC(z;\ T,\ F) = \lim_{\varepsilon \longrightarrow 0} \left\{ \frac{T\left((1-\varepsilon)F + \varepsilon\delta_z\right) - T(F)}{\varepsilon} \right\}, \qquad (2.1)$$

nos pontos z do espaço amostral onde o limite existe. As mixturas de F e δ_z são escritas como $(1-\varepsilon)F + \varepsilon\delta_z$, para todo $\varepsilon \in (0,1)$.

O número IC(z; T, F) fornece uma idéia da velocidade com que o valor de T muda quando o modelo suposto F, é contaminado por uma distribuição com massa um no ponto z. Logo, do ponto de vista da robustez, preferimos estimadores definidos por funcionais T com IC(z; T, F) o mais próximo possível de zero para todo ponto z.

A função de Influência é também utilizada para construir estimadores robustos. Hampel e outros (1986), apresentaram exemplos de estimadores de posição construídos a partir desta função.

2.2.1 Aproximação da Função de Influência para Amostras Finitas

A definição de função de influência é assintótica, pois está

baseada em funcionais que são consistentes em F. Por outro lado, existem estimadores cujas funções de influência assintóticas não podem ser facilmente calculadas. Nesses casos podemos usar versões para amostras finitas. Uma destas versões é a *Curva de Sensibilidade* a qual é muito usada na análise exploratória de dados e é baseada na diferença

$$t_n\left(z_1, z_2, \ldots, z_{n-1}, z\right) - t_{n-1}\left(z_1, z_2, \ldots, z_{n-1}\right),$$

que indica como uma nova informação "z" influi no estimador $t = t_n$. Esta expressão servirá para dar uma definição formal de curva de sensibilidade.

DEFINIÇÃO 2.2.1

Sejam t_{n-1} e t_n estimadores de θ baseados em amostras de tamanho n-1 e n, respectivamente. Ambos são estimadores definidos pelo mesmo funcional (somente os tamanhos amostrais são diferentes). A Curva de sensibilidade de t_n com respeito a t_{n-1} é definida por

$$SC_{n-1}(z) = \frac{\left(t_n\left(z_1, z_2, \ldots, z_{n-1}, z\right) - t_{n-1}\left(z_1, z_2, \ldots, z_{n-1}\right)\right)}{1/n} \qquad (2.2)$$

Considerando as distribuições empíricas de probabilidade (Def. A1), temos

$$t_{n-1}\left(z_1, z_2, \ldots, z_{n-1}\right) = t\left(F_{n-1}\right),$$

$$t_n\left(z_1, z_2, \ldots, z_{n-1}, z\right) = t\left(\frac{n-1}{n} F_{n-1} + \frac{1}{n} I_{\{z_n = z\}}\right).$$

Logo a curva de sensibilidade pode ser expressa como

$$SC_{n-1}(z) = \frac{t\left(\frac{n-1}{n} F_{n-1} + \frac{1}{n} I_{\{z_n = z\}} \right) - t\left(F_{n-1}\right)}{1/n} \quad . \qquad (2.3)$$

Observamos que estimadores robustos no sentido de resistência a um único valor discrepante são os que tem SC_{n-1} limitada. O estimador de regressão de mínimos quadrados é um exemplo de estimador com função de influência não limitada.

Vejamos agora o que entendemos por robustez diante de pequenos desvios do modelo suposto, ou seja, o que entendemos por robustez qualitativa.

2.3 ROBUSTEZ QUALITATIVA

Suponhamos que a verdadeira distribuição de probabilidade da amostra aleatória $z_1, z_2, \ldots, z_n$, não é a distribuição suposta F, mas uma distribuição Q "próxima" de F. Diremos que o estimador t_n é qualitativamente robusto, se as distribuições de probabilidade do estimador sobre F e Q, denotadas por $\Omega_F\left(t_n\right)$ e $\Omega_Q\left(t_n\right)$, estão uniformemente "próximas" para todo n.

A expressão A "próximo" de B, leva à noção de distância, e neste caso específico à distância entre distribuições, isto é, entre pontos do conjunto de medidas de probabilidade definidas sobre $\mathbb{R}$. Na teoria de espaços métricos encontramos vários tipos de medidas de distância entre distribuições. As mais usadas na teoria de robustez são as métricas de Levy, Prohorov e Kolmogorov. Para distribuições de probabilidade definidas sobre $\mathbb{R}$, as métricas de Levy e Prohorov são equivalentes, e sob certas condições, a métrica de Levy é equivalente

à métrica de Kolmogorov (Huber (1981), pag. 34). Segue então, uma definição mais formal de robustez qualitativa.

DEFINIÇÃO 2.3.1

Seja π^* uma métrica sobre $\mathfrak{F}(\mathbb{R})$, o conjunto de medidas de probabilidade definidas sobre $\mathbb{R}$. Dizemos que t_n é qualitativamente robusto em $F \in \mathfrak{F}(\mathbb{R})$ se para todo $\varepsilon > 0$, existe $\delta > 0$ tal que

$$\pi^*(F, Q) \leq \delta \Rightarrow \pi^*\left[\mathcal{L}_F\left(t_n\right), \mathcal{L}_Q\left(t_n\right)\right] \leq \varepsilon \qquad \forall\, n, \qquad (2.4)$$

onde $Q \in \mathfrak{F}(\mathbb{R})$.

A expressão (2.4) é uma condição necessária de robustez qualitativa, mas não é suficiente. Essa medida detecta apenas a ausência deste tipo de robustez mas não garante que um procedimento estatístico é qualitativamente robusto. Além disso, a robustez qualitativa está fortemente associada à escolha da métrica, π^*.

2.4 PONTO DE RUPTURA

Ponto de ruptura é uma medida global de robustez que fornece uma idéia da tolerância do estimador a observações discrepantes. Hampel (1971), apresenta uma definição assintótica, de natureza matemática, a qual não será usada aqui. Utilizaremos uma versão para amostras finitas, introduzida por Donoho e Huber (1983) baseada na definição de vício.

Consideremos uma amostra de n observações $Z = \left\{z_1, \ldots, z_n\right\}$, e Z^* um conjunto construído substituindo m observações de Z por

valores arbitrários. Definimos b(m; t, Z), o *vício máximo* causado pela contaminação do conjunto de dados, por

$$b(m;\ t,\ Z) = \underset{z^*}{\text{Sup}} \parallel t(Z^*) - t(Z) \parallel, \tag{2.5}$$

onde o supremo é obtido sobre todas as possíveis amostras contaminadas. Se b(m; t, Z) é infinito, então as m observações discrepantes têm um efeito arbitrariamente grande no estimador.

DEFINIÇÃO 2.4.1

Seja $t = t_n$ um estimador aplicado na amostra Z. O ponto de ruptura, $\varepsilon_n^*(t, Z)$, desse estimador é dado por

$$\varepsilon_n^*(t,\ Z) = \underset{1 \le m \le n}{\text{Min}} \left\{ \frac{m}{n}\ ,\ b(m;\ t,\ Z) \text{ é infinito} \right\}. \tag{2.6}$$

Da expressão (2.6) temos que o ponto de ruptura do estimador é igual à proporção mínima de observações discrepantes contidas na amostra, que torna o vício do estimador infinito.

Do ponto de vista da robustez, interessam os estimadores com alto ponto de ruptura. O valor máximo do ponto de ruptura é 50%, pois se o número de dados contaminados for maior do que este valor teremos que trocar as definições de observações "boas" e discrepantes para esse conjunto de dados.

No caso do estimador de mínimos quadrados temos que uma única observação discrepante pode afetar fortemente a estimativa. Assim, o ponto de ruptura desse estimador é

$$\varepsilon_n^*(t,\ Z) = 1/n \qquad \text{e} \qquad \varepsilon_n^*(t,\ Z) \longrightarrow 0 \text{ quando } n \longrightarrow \infty\ .$$

Outro conceito que será de utilidade nos capítulos seguintes

é a definição de ponto de ruptura do estimador da matriz de dispersão de um conjunto de dados multivariado, X.

DEFINIÇÃO 2.4.2

Definimos o ponto de ruptura do estimador da matriz de dispersão, C_n, por

$$\varepsilon^* \left(C_n, X\right) = \underset{1\le m\le n}{\text{Min}} \left\{ \frac{m}{n} : \underset{X^*}{\text{Sup}}\, D\left(C_n(X), C_n(X^*)\right) \text{ é infinito} \right\}, \quad (2.7)$$

onde o supremo é calculado sobre todos os conjuntos X^* resultantes da substituição de m observações do conjunto de dados $X = \left\{x_1, \ldots, x_n\right\}$, $x_i \in \mathbb{R}^K$, $k \ge 1$ por valores arbitrarios. O ponto de ruptura de C_n é dado pela menor fração de dados discrepantes que podem fazer com que o maior autovalor, $\lambda_1\left(C_n\right)$, seja muito grande e/ou que o menor autovalor, $\lambda_k\left(C_n\right)$, assuma valores próximos de zero.

A distância entre as matrizes utilizadas na construção da expressão da direita de (2.7) é definida por

$$D(A, B) = \text{Max} \left\{ \lambda_1(A) - \lambda_1(B), \lambda_k^{-1}(A) - \lambda_k^{-1}(B) \right\},$$

com $\lambda_1(A) \ge \lambda_2(A) \ge \ldots \ge \lambda_k(A)$ autovalores da matriz A.

Propriedades desejáveis dos estimadores são a equivariância e invariância sob transformações lineares nas variáveis, pois facilitam a manipulação dos dados. Além disso, podemos encontrar limites superiores para o ponto de ruptura dos estimadores que têm essas propriedades. Os teoremas seguintes relacionam o ponto de ruptura com as propriedades de equivariância sob transformações lineares, tanto para estimadores de posição multivariados quanto de regressão linear múltipla.

TEOREMA 2.4.1

Seja $Z = \left\{z_1, \ldots, z_n\right\}$ com elementos $z_i = \left(y_i, x_{i1}, \ldots, x_{ik}\right)$. Se t_n é um estimador de regressão que tem a propriedade de equivariância sob transformações afins, então

$$\varepsilon^*\left(t_n, Z\right) \leq \frac{[(n-k)/2] + 1}{n} \quad , \tag{2.8}$$

onde [u] é o enteiro menor o igual a u.

PROVA

Suponhamos que

$$\varepsilon^*\left(t_n, Z\right) > \frac{[(n-k)/2] + 1}{n} \quad .$$

Então existe uma constante finita b tal que

$$t(Z^*) \in B(t(Z),b) = \left\{\theta \in \mathbb{R}^k \; : \; \|t(Z) - \theta\| \leq b \right\},$$

onde Z^* é um conjunto contaminado contendo pelo menos $q = n-[(n-k)/2]-1$ ou equivalentemente $q = n - [(n+k+1)/2] - 1$ pontos de Z.

Seja $v \neq 0$ um vetor de dimensão k, tal que

$$x_1 v = 0, \ \ldots, \ x_{k-1} v = 0$$

Por outro lado, temos que

$$2q - (k-1) = \begin{cases} n & \text{se } (n+k+1) \text{ é par} \\ n-1 & \text{se } (n+k+1) \text{ é impar} \end{cases}$$

ou seja, $2q-(k-1) \leq n$. Podemos então construir o conjunto Z^{**} substituindo os primeiros $2q-(k-1)$ pontos de Z por

$$\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots\ldots \left(x_{k-1}, y_{k-1}\right), \left(x_k, y_k - x_k\tau v\right), \ldots,$$

$$\left(x_q, y_q - x_q\tau v\right), \ldots, \left(x_k, y_k\right), \ldots, \left(x_q, y_q\right),$$

para algum $\tau > 0$. Para esta amostra, a estimação

$$t(Z^{**}) \in B(t(Z), b), \tag{2.9}$$

pois Z^{**} contém q pontos de Z.

Por outro lado, podemos construir o conjunto $Z^{*} = Z^{**} + \tau v$ com elementos

$$\left(x_1, y_1\right), \ldots., \left(x_{k-1}, y_{k-1}\right), \left(x_k, y_k\right), \ldots, \left(x_q, y_q\right), \left(x_k, y_k - x_k\tau v\right),$$

$$\ldots, \left(x_q, y_q - x_q\tau v\right).$$

Esta amostra também contém pelo menos q pontos de Z, logo a estimativa

$$t(Z^{*}) \in B(t(Z), b).$$

Pela propriedade de invariância sob transformações de regressão podemos escrever

$$t(Z^{*}) = t(Z^{**}) + \tau v. \tag{2.10}$$

De (2.9) e (2.10) temos que

$$t(Z^{*}) \in B(t(Z) + \tau v, b),$$

o qual é uma contradição pois para τ suficientemente grande

$$B(t(Z), b) \cap B(t(Z) + \tau v, b) \longrightarrow \emptyset. \qquad \blacksquare$$

TEOREMA 2.4.2

Seja $X = \left\{x_1, \ldots, x_n\right\}$, um conjunto de dados em $\mathbb{R}^k$; se t_n é um estimador de posição equivariante sob translação, então

$$\varepsilon^* \left(t_n, X\right) \leq \frac{[(n+1)/2]}{n} , \qquad (2.11)$$

onde [u] é o enteiro menor ou igual a u.

Observação.- A prova do teorema tem a mesma construição que a prova do teorema 2.4.1 e pode ser encontrada em Lopunhaä e Rousseeuw (1991).

Algumas propriedades dos estimadores, que facilitam o trabalho analítico e numérico, são dados pelo seguinte lema.

LEMA 2.4.1

Seja $X = \left\{x_1, \ldots, x_n\right\}$ um conjunto de observações em $\mathbb{R}^k$; $t_n(X) \in \mathbb{R}^k$ e $C_n(X) \in PDS(k)$, estimadores de posição e dispersão, respectivamente; onde PDS(k) é a classe de todas as matrizes de ordem k definidas positivas e simétricas.

i) Se t_n é invariante sob translação,

$$\varepsilon^*\left(t_n, X + v\right) = \varepsilon^*\left(t_n, X\right), \qquad \forall\, v \in \mathbb{R}^k \qquad (2.12)$$

ii) Se t_n é invariante sob rotação,

$$\varepsilon^*\left(t_n, AX\right) = \varepsilon^*\left(t_n, X\right), \qquad \forall\, A \in \mathbb{R}^n x \mathbb{R}^k \qquad (2.13)$$

iii) Se C_n é invariante sob translação

$$\varepsilon^*\left(C_n,\ X + B\right) = \varepsilon^*\left(C_n,\ X\right), \qquad \forall\ B \in \mathbb{R}^n \times \mathbb{R}^k \tag{2.14}$$

iv) Se C_n é invariante sob rotação

$$\varepsilon^*\left(C_n,\ AX\right) = \varepsilon^*\left(C_n,\ X\right), \qquad \forall\ A \in \mathbb{R}^n \times \mathbb{R}^k \tag{2.15}$$

2.5 RELAÇÃO ENTRE FUNÇÃO DE INFLUÊNCIA , ROBUSTEZ QUALITATIVA E PONTO DE RUPTURA

Os três critérios de robustez apresentados são alguns dos muitos que podem ser encontrados na literatura de robustez. A escolha de um desses critérios para avaliar ou construir um estimador vai depender da necessidade de robustecer determinado aspecto da análise de dados.

Existe uma relação entre os aspectos de robustez mencionados neste capítulo, como veremos a seguir. Um estimador com ponto de ruptura zero é qualitativamente não robusto (por exemplo, o estimador de mínimos quadrados no modelo linear). A inversa não necessariamente é verdadeira, isto é, um estimador com ponto de ruptura alto nem sempre é qualitativamente robusto. Por exemplo, a mediana tem ponto de ruptura 1/2 mas é qualitativamente robusto, sob a distribuição F, somente no caso que $t = F^{-1}(1/2)$ assume apenas um valor. Por outro lado, o estimador de mínimos quadrados aparados com $\alpha \in (0,1/2)$, dado pelo funcional

$$T(F) = \frac{1}{(1-2\alpha)} \int F^{-1}(t)dt,$$

é um exemplo de estimador com ponto de ruptura diferente de zero ($\varepsilon_n^* = \alpha$), que é qualitativamente robusto.

A função de influência e o ponto de ruptura são medidas quantitativas de robustez, o que permite estabelecer comparações entre estimadores. Na medida do possível, procura-se construir estimadores que tenham função de influência limitada, próxima de zero, e alto ponto de ruptura.

Podemos encontrar uma grande quantidade de estimadores que são robustos segundo alguns dos critérios mencionados, os quais são estabelecidos de acordo com a utilização do estimador. Por exemplo, na anãlise de regressão, o interesse é a utilização de estimadores que sejam robustos no sentido de não serem afetados por múltiplas observações discrepantes e que, além disso, ponham em evidência estas observações, ou seja, estamos interessados em estimadores com alto ponto de ruptura.

No capítulo seguinte, trataremos sobre observações discrepantes na análise de regressão linear múltipla, apresentaremos os diagnósticos de dados discrepantes, baseados no método de estimação de mínimos quadrados, que são geralmente usados.

CAPITULO 3

OBSERVAÇÕES DISCREPANTES E DIAGNÓSTICOS BASEADOS NO MÉTODO DE MÍNIMOS QUADRADOS

3.1 INTRODUÇÃO

A confiabilidade de um conjunto de dados obtidos sob condições similares está baseada na relação entre as observações do conjunto. Eventualmente, alguns dados podem apresentar um comportamento discrepante do padrão seguido pela maioria. Observações que, na opinião do pesquisador, encontram-se fora da massa dos dados têm sido chamados de "outliers", "observações discordantes", "dados aberrantes", "dados contaminantes", "observações surpreendentes", dados discrepantes", "dados grosseiros", etc.. Não existe uniformidade em relação ao significado exato destes termos, a pesar de estudios desse tipo de observações ter-se iniciado há quase 200 anos, e da grande quantidade de artigos escritos sobre o assunto. A denominação mais comum é "outlier", mas neste caso usaremos "*observação discrepante*".

A necessidade de identificar essas observações discrepantes

deve-se ao fato das mesmas poderem distorcer a informação que a massa dos dados deve fornecer sobre o fenômeno em estudo. Por outro lado, essas observações discrepantes poderiam também ser importantes "*mensagens*" de que as suposições que formulamos em relação à fonte que gerou o conjunto de dados não são corretas e, em consequência, precisamos mudar nossa concepção do fenômeno sob estudo. Na seção 2 apresentaremos, brevemente, possíveis origens das observações discrepantes em um conjunto de dados e as razões do interesse em estudá-las.

A estrutura particular dos dados na análise de regressão indica se a discrepância de uma observação é atribuida à discrepância na direção da variável resposta, das variáveis regressoras ou em relação ao modelo especificado. Consequentemente, os métodos de identificação de observações discrepantes deverão considerar essas características dos dados. Na seção 3, descreveremos as diferentes formas em que podem apresentar-se as observações discrepantes na análise de regressão linear.

Uma forma comumente usada para identificar observações discrepantes na análise de regressão linear é através da análise dos resíduos e dos diagnósticos de influência, os quais orientam a atenção sobre observações que têm uma influência maior do que as outras na estimação dos parâmetros do modelo. Existe, atualmente, uma grande quantidade de estatísticas de diagnósticos as quais não somente identificam as observações discrepantes como também estudam o efeito de observações individuais ou grupos de observações em determinadas fases da análise de dados. A maioria desses métodos baseiam-se no resíduos associados ao ajuste pelo método de mínimos quadrados.

Na seção 4 apresentaremos alguns resultados derivados do ajuste de mínimos quadrados tais como a matriz de projeção, os diferentes tipos de resíduos e alguns diagnósticos de influência comumente usados nos pacotes computacionais dirigidos à análise de regressão.

3.2 OBSERVAÇÕES DISCREPANTES

Um dos artigos mais completos referentes ao tema dos dados discrepantes é apresentado por Beckman e Cook (1983). Eles resumem em três as diferentes denominações usadas nos artigos e dão as definições de observação discordante, contaminante e discrepante.

Observação discordante Qualquer observação que parece surpreendente para o pesquisador. Isso implica que ele tem pelo menos uma visão informal do modelo teórico, ou supos um determinado modelo.

Observação contaminante Qualquer observação que não é uma realização da distribuição suposta. A discrepância destas observações pode não ser tão obvia.

Observação discrepante ("outlier") refere-se à observação contaminante ou discordante.

Essa definição de observação discrepante é um tanto ambígua, pois o que realmente caracteriza a "observação discrepante" é o seu impacto no observador. Entretanto uma observação pode ser contaminante e não ser percebida. Para visualizar melhor as diferenças, vejamos o gráfico 3.2.1, que foi apresentado por Barnett e Lewis (1978). O conjunto de dados $Z = \left\{z_1, z_2, \ldots, z_n\right\}$ corresponde a uma amostra aleatória que supoe-se foi obtida desde uma distribuição F. Ordenando o conjunto de dados observamos que: as observações $z_{(1)}$ e $z_{(n)}$ são extremas; a observação $z_{(1)}$ é discordante entanto que $z_{(n)}$ não apresenta comportamento suspeito. Por outro lado, $z_{(2)}$ é discordante contaminante, i.é, encontra-se fora da massa dos dados e não é realização da distribuição F; $z_{(4)}$ é contaminante pois não é realização da distribuição F. Esta observação dificilmente será identificada pelos métodos de diagnósticos pois não é discordante. Finalmente as

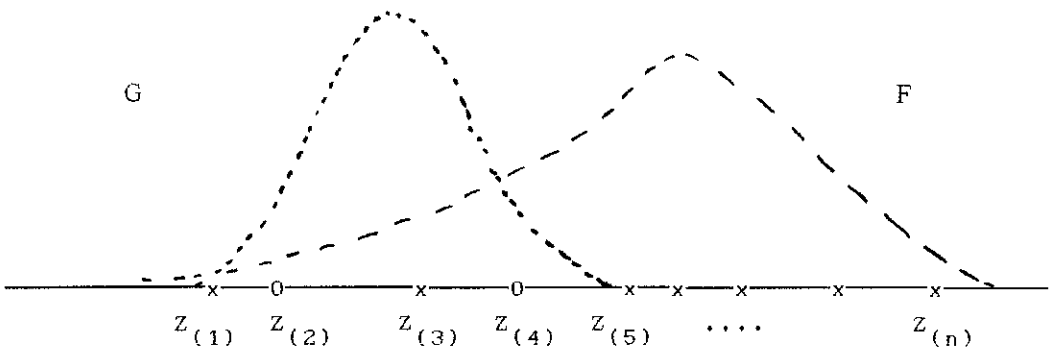

Figura 3.2.1 O conjunto $\{z_1, z_2, \ldots, z_n\}$ é uma amostra obtida a partir de uma variável aleatória com distribuição F.

observações $z_{(3)}$, $z_{(5)}$, ..., $z_{(n)}$ provém da distribuição F e não apresentam comportamento suspeito. Podemos então concluir que:

- Uma observação contaminante não necessariamente é discrepante. Ela será discrepante se além de contaminante for também discordante.

- Uma observação extrema será discrepante se aparece como discordante sob o modelo suposto.

Diremos então que, observação discrepante é uma observação que parece *inconsistente* com os demais dados.

3.2.1 Objetivos da Detecção de Observações Discrepantes

A detecção de observações discrepantes pode ser realizada com distintos objetivos, tais como:

a) Interesse especial nas observações discrepantes, por se mesmas. Neste caso podemos desejar identificar os dados discrepantes para:
   . Estudo posterior
   . Obter nova informação importante contida em variáveis concomitantes, que poderiam em outro caso não ser percebidas.
   . Sua incorporação no conjunto de dados através de uma revisão do modelo ou método de estimação.
   . Reconhecimento de uma debilidade inerente aos dados.

   Nestes casos o problema estatística envolve a obtenção de inferências acerca das observações que forem detectadas como suspeitas.

b) Dar uma linha base para julgar observações como discrepantes. Neste caso o interesse concentra-se na detecção de fenômenos raros, mais do que na estimação de características comuns.

c) Verificação das suposições feitas na formulação do modelo. Em análises de dados estruturados onde se estabelecem modelos probabilísticos, a presença de observações discrepantes quase sempre indica falhas no modelo, nos dados, ou ambos.

Nos casos a) e b) a estimação dos parâmetros somente é necessária para investigar a discordância de certas observações. Em c) o objetivo é a predição ou outras inferências que serão válidas para os dados genuinos.

3.2.2 Origem das Observações Discrepantes

As causas da presença de observações discrepantes em um conjunto de dados podem ser agrupadas em três categorias. Estas são: debilidade do modelo suposto, problemas na obtenção dos dados e variabilidade natural dos dados. Subentende-se pelas duas primeiras categorias que as observações discrepantes são julgadas com base em um modelo explícito ou implícito em mente.

As debilidades no modelo suposto incluem causas tais como: uma variável resposta na escala errada, ou o modelo teórico suposto não é adequado. A primeira causa pode levar a uma transformação na resposta, enquanto que a última pode conduzir a substituir o modelo atual.

Os problemas na obtenção dos dados referem-se somente às observações e não ao modelo como um todo. Estas causas podem indicar que as observações discrepantes sejam tratadas individualmente. Exemplos deste tipo de dados são os erros de medição, observação ou registro.

A variabilidade natural dos dados refere-se a dados que podem aparecer como discrepantes, entretanto eles são observações genuínas da distribuição considerada.

3.3 OBSERVAÇÕES DISCREPANTES NA ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA

Consideremos o conjunto de observações $Z = \left\{z_1, z_2, \ldots, z_n\right\}$, onde $z_i = \left(y_i, x_{i1}, \ldots, x_{ik}\right)$; $k \geq 1$, relacionadas segundo o modelo (1.2). Neste contexto podem surgir observações discrepantes nas variáveis regressoras, na variável resposta, ou em ambas direções.

Frequentemente, as variáveis regressoras $X_1, \ldots, X_k$ são

quantidades observadas sujeitas a variabilidade. Mesmo em situações onde o experimento é planejado podem acontecer erros imprevisíveis. Estas observações são conhecidas como dados discrepantes na direção da matriz do modelo, X, e terão influência na estimação. A influência pode ser "boa" ou "ruim" dependendo do valor da variável resposta com a qual se combina. Estes dados são chamados de *pontos de alavanca* e podem ser resultados de erros nas linhas da matriz do modelo, cobertura inadequada das regiões no espaço das variáveis regressoras, heterocedasticidade ou não aditividade. Notemos que os pontos de alavanca "ruins" são discrepantes na direção de X, mas as observações discrepantes na direção de X, não necessariamente são pontos de alavanca "ruins".

As respostas discrepantes caracterizam-se por possuirem resíduos verticais notavelmente grandes. Chamaremos observações discrepantes da regressão àquelas que se afastam do padrão seguido pela maioria das observações.

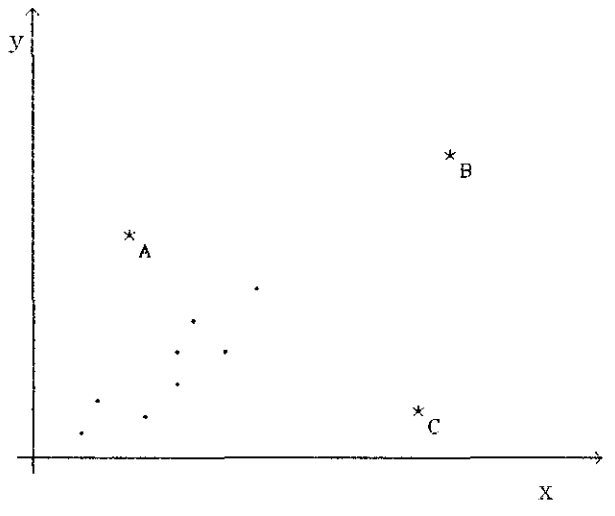

Figura 3.3.1 Pontos Discrepantes na direção de y, (x,y) e de x representados pelos pontos A, B, e C, respectivamente.

No gráfico (3.3.1), a massa dos pontos apresenta uma relação aproximadamente linear, entretanto, os pontos "A" e "C" se afastam do padrão seguido pela maioria. A observação "A" é discrepante na direção de y, mas está próxima do centro da variável x. Este ponto apresentará resíduo grande. O ponto "B" é discrepante na direção da variável x mas possue resíduo pequeno. Este é um exemplo de ponto de alavanca "bom". O ponto "C" é discrepante na direção da variável x e pelo valor de y com o qual se combinou poderá deslocar a reta na sua direção dependendo do método de estimação utilizado, portanto é um ponto de alavanca "ruim". O resíduo correspondente não é necessariamente grande.

Os métodos de estimação de mínimo valor absoluto, e de mínimos quadrados são muito afetados por pontos de alavanca tais como o ponto "C" da figura. Dado que o objetivo desses métodos é minimizar a soma de resíduos absolutos e a soma dos quadrados dos resíduos no primeiro e segundo caso, respectivamente , o hiperplano ajustado será deslocado na direção do ponto de alavanca. O resíduo correspondente ao ponto "C" pode não ser notavelmente grande e portanto dificilmente detectado na análise de resíduos. Entretanto, outras observações que seguem a direção da maioria dos dados podem apresentar resíduos relativamente grandes.

3.4 DIAGNÓSTICOS EM REGRESSÃO BASEADOS NO MÉTODO DE MÍNIMOS QUADRADOS

Sob a suposição de que os erros do modelo (1.1), ε_i, i=1,...n são variáveis aleatórias independentes com esperança zero e variância σ^2, o teorema de Gauss-Markov garante que o estimador

$$\hat{\theta} = (X^tX)^{-1}X^ty, \qquad (3.1)$$

é o único estimador não viciado com variância mínima na classe dos estimadores lineares. Se além disso os erros do modelo têm distribuição

normal com os parâmetros antes mencionados, o vetor $\hat{\theta}$ definido em (3.1) é o estimador de máxima verossimilhança dos parâmetros do modelo de regressão e é não viciado de mínima variância.

Seguem os métodos de diagnósticos de observações discrepantes mais usados na análise de regressão linear múltipla, baseados no método de estimação de mínimos quadrados. Estes diagnósticos já formam parte da maioria dos pacotes computacionais que fazem análise de regressão linear. Os métodos que apresentaremos investigam o efeito individual, isto é, a *influência* de cada observação em diferentes fases da análise de dados. Ao final do capítulo comentaremos brevemente sobre os métodos de diagnóstico de múltiplas observações discrepantes.

3.4.1 A Matriz de Projeção H

A partir de (3.1) podemos expressar o vetor de respostas preditas, como

$$\hat{y} = X(X^tX)^{-1}X^ty. \qquad (3.2)$$

A matriz

$$H = X(X^tX)^{-1}X^t, \qquad (3.3)$$

é o *operador de projeção ortogonal no espaço coluna de* X. Os elementos da diagonal de H, denotados por h_{ii} são muito usados para diagnosticar a influência de observações individuais.

Versões escalares dos valores preditos, sua variância e a variância dos resíduos são dados por

$$\hat{y}_i = h_{ii}\, y_i + \sum_{j \neq i}^{n} h_{ij}\, y_j \quad ; \qquad i = 1,2, \ldots,n. \qquad (3.4)$$

$$V\left(\hat{y}_i\right) = h_{ii}\sigma^2 \quad ; \qquad i = 1,2,\ldots,n. \tag{3.5}$$

$$V\left(e_i\right) = \left(1- h_{ii}\right)\sigma^2 \quad ; \qquad i = 1,2,\ldots,n. \tag{3.6}$$

$$V\left(\hat{\theta}_j\right) = c_{jj}\sigma^2 \quad ; \qquad j = 1,2,\ldots,k, \tag{3.7}$$

onde c_{jj} é o j-ésimo elemento da diagonal de $(X^tX)^{-1}$.

Pelas propriedades de simetria e idempotencia da matriz de projeção temos que

$$h_{ii} = \sum_{j\neq i}^{n} h_{ij}^2 + h_{ii}^2 = \sum_{j=1}^{n} h_{ij}^2,$$

$$0 \leq h_{ii} \leq 1,$$

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ji} = 1.$$

Notemos que se $h_{ii} \longrightarrow 1$, então $h_{ij} \longrightarrow 0$; $\forall\ i \neq j$ e $\hat{y}_i \longrightarrow y_i$, isto é, a i-ésima observação é ajustada exatamente pelo hiperplano de regressão dado pela expressão (3.4). Em consequência, também teremos que a variância dos resíduos tende a zero em (3.6) e a variância do valor predito tende a σ^2 em (3.5). Vemos portanto que se h_{ii} está próximo a 1, a observação correspondente será altamente influente na determinação da regressão. Estes valores podem ser vistos como uma medida da distância do vetor linha $x_i \in X$ ao centroide dos dados.

O posto da matriz H é dado por

$$\text{Posto(H)} = \text{Posto (X)} = \text{traço (H)} = k,$$

e o tamanho médio dos elementos da diagonal é k/n. Em experimentos delineados o ideal é usar pontos no espaço das variáveis regressoras que sejam igualmente influentes, isto é, que cada ponto tenha h_{ii}

próximo a k/n. Em geral, X não é delineado e não é possível controlar os valores de h_{ii}. Precisamos então, de algum critério para decidir quando um valor de h_{ii} pode ser considerado suficientemente grande para chamar nossa atenção. Hoaglin e Welsch (1978) e Cook e Weisberg (1982) sugerem que com $h_{ii} \geq 2k/n$, a i-ésima observação seja considerada potencialmente influente.

Dado que o estimador de mínimos quadrados é equivariante sob mudanças de posição na matriz do modelo, sem perda de generalidade, podemos supor que o centroide dos vetores linha $x_i \in \mathbb{R}^k$ é zero. Logo, considerando um modelo com intercepto temos

$$h_{ii} = \frac{1}{n} + x_i (X^t X)^{-1} x_i^t \ . \qquad (3.8)$$

Outra forma de detectar observações potencialmente influentes é construindo o menor conjunto convexo contendo a totalidade dos vetores linha da matriz do modelo, denominado *contorno das variáveis regressoras* (RVH). Para isso, precisamos encontrar o maior elemento da diagonal de H, isto é, $\underset{1 \leq i \leq n}{\text{Max}} \ h_{ii} = h_{max}$ e calcular

$$\frac{1}{n} + x_i (X^t X)^{-1} x_i^t \leq h_{max} \ . \qquad (3.8)$$

Se o volume do elipsóide em (3.9) é muito maior do que o volume que se obteria considerando o seguinte h_{ii} máximo, então a observação com h_{max} será considerada uma observação influente ou ponto de alavanca.

Uma desvantagem em utilizarmos h_{ii} ; i=1,2,...,n, como medida de influência é que estes sofrem o efeito de mascaramento, isto é, múltiplas observações discrepantes podem não ser detectadas por este método de diagnóstico. Este fato será mais claramente visualizado no capítulo 5, onde veremos a relação dos h_{ii} com a distância de Mahalanobis.

3.4.2 Análise de Resíduos

Um método comum de detecção de observações discrepantes consiste em examinar os resíduos padronizados

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad ; \qquad i=1,2,\ldots,n, \tag{3.10}$$

onde

$$\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 /(n-k). \tag{3.11}$$

Estes resíduos são denominados "studentizados" externamente pois sob a suposição de normalidade dos erros do modelo, ε_i, a distribuição dos r_i será aproximadamente t-Student. Um critério para decidir quando um resíduo pode ser considerado notoriamente grande é usar o fato de que resíduos com $r_i > 2.5$ são muito raros nesta distribuição, exceto para graus de liberdade pequenos.

Outro tipo de resíduo padronizado, conhecido como resíduo "studentizado" internamente o qual é obtido excluindo a i-ésima observação no cálculo da variância estimada dos resíduos, é dado por

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \quad ; \qquad i=1,2,\ldots,n, \tag{3.12}$$

onde

$$\hat{\sigma}^2_{(i)} = \frac{(n-k)\,\hat{\sigma}^2 - e_i/\left(1-h_{ii}\right)}{n-k} \; . \tag{3.13}$$

Note que e_i e $\hat{\sigma}^2_{(i)}$ são independentes pois a i-ésima observação não participa do ajuste do modelo. Logo a variável t_i definida em (3.12) tem distribuição t-Student, sempre que os erros do modelo (1.1) apresentem distribuição normal. A expressão (3.12) é usada como um teste estatístico para estabelecer se a i-ésima observação

está de acordo com o modelo. Este tipo de resíduo é mais efetivo do que os r_i definidos em (3.10) para detecção de observações discrepantes individuais.

Considerando que os resíduos podem não ser notoriamente grandes, pelo fato de que as observações discrepantes deslocam o hiperplano na sua direção, e que h_{ii} não detecta múltiplas observações discrepantes na direção das variáveis regressoras, podemos concluir que, tanto os resíduos "studentizados" internamente quanto os "studentizados" externamente não são confiáveis na presença de múltiplas observações discrepantes.

3.4.3 Outros Diagnósticos de Influência

Para avaliar a influência potencial de uma observação em determinada fase da análise de dados faremos o ajuste com e sem a i-ésima observação. Empregaremos o sub-indice (i) para significar que a i-ésima observação foi retirada do conjunto de dados. Logo temos

$$\hat{\theta}_{(i)} = \left(X_{(i)}^t X_{(i)}\right)^{-1} X_{(i)}^t y_{(i)}, \qquad (3.14)$$

$$\hat{y}_{(i)} = X_{(i)}\left(X_{(i)}^t X_{(i)}\right)^{-1} X_{(i)}^t y_{(i)}, \qquad (3.15)$$

onde $X_{(i)}$ é a matriz do modelo e, $y_{(i)}$ é o vetor de respostas desconsiderando a i-ésima linha da matriz X e, a i-ésima resposta observada, y_i. Os elementos dos vetores $\hat{\theta}_{(i)}$ e $\hat{y}_{(i)}$ em (3.14) e (3.15) são denotados por $y_{(i),i}$ e $\hat{\theta}_{(i),j}$; $i = 1,2,\ldots,n$, $j = 1,2,\ldots,k$, respectivamente.

A seguir determinaremos a magnitude das diferenças entre as estatísticas derivadas dos dois ajustes (3.1) e (3.14).

i) Estatística DFBETAS

Mede a influência da i-ésima observação na estimativa do j-ésimo coeficiente de regressão. Esta influência será avaliada a partir da diferença $\hat{\theta}_{(i)} - \hat{\theta}$ padronizada pela variância do vetor de coeficientes, $V(\hat{\theta})$, cuja versão escalar é dada em (3.7). A variância dos erros é dada por (3.13). A estatística DFBETAS para o j-ésimo coeficiente de regressão é

$$DFBETAS_{(i),j} = \frac{a_{ij}\, t_i}{\left(c_{jj}(1-h_{ii})\right)^{1/2}} \quad ; \quad \begin{matrix} i=1,2,\ldots,n \\ j=1,2,\ldots,k \end{matrix} . \qquad (3.16)$$

c_{jj} foi definido na expressão (3.7) e a_{ij} é o (i,j)-ésimo elemento da matriz $(X^tX)^{-1}X^t$, e t_i é definida em (3.12).

Observação.- O ponto de corte sugerido neste caso é $2/n^{1/2}$.

ii) Estatística DFFITS

Esta estatística proposta por Belsley, Kuh e Welsch (1980), indica a mudança na i-ésima resposta estimada, $\hat{y}_i$, quando realizamos o ajuste desconsiderando a i-ésima observação. Esta mudança será avaliada através da diferência $\hat{y}_i - \hat{y}_{(i),i}$, padronizado pela variância do estimador, $V\left(\hat{y}_i\right)$, dada em (3.5). A variância dos erros é dada por (3.13).

Expressando-a em termos dos resíduos "studentizados" internamente temos

$$DFFITS_i^2 = t_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right); \qquad i=1,2,\ldots,n. \qquad (3.17)$$

Uma regra empírica para decidirmos se uma observação influi em determinada resposta estimada é usando o ponto de corte $2\{k/(n-k)\}^{1/2}$, ou também $2(k/n)^{1/2}$ para os casos em que $k \ll n$.

iii) Distância de COOK

Proposta por Cook (1977) é definida como a distância, padronizada, entre os vetores de coeficientes estimados, $\hat{\theta}_{(i)}$ e $\hat{\theta}$,

$$COOK = \frac{\left(\hat{\theta}_{(i)} - \hat{\theta}\right)\left(X^tX\right)\left(\hat{\theta}_{(i)} - \hat{\theta}\right)}{k\,\sigma^2}, \qquad (3.18)$$

onde k é o número de parâmetros e σ^2 é a variância dos erros que será estimada por (3.11). Para a i-ésima observação temos

$$COOK_i = \frac{1}{k}\, r_i^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right) ; \qquad i=1,2,\ldots,n. \qquad (3.19)$$

Um valor grande de COOK indica que a i-ésima observação influi fortemente nos coeficientes de regressão estimados e sua eliminação pode produzir mudanças importantes nas conclusões.

Observação.- "Grande" neste caso será relativo ao grupo.

A distância definida em (3.19) é equivalente à estatística $DFFITS^2$ dada em (3.17), as duas calculam a distância euclideana entre os vetores de estimativas com e sem a i-ésima observação, a diferença está nos fatores de escala pois usam versões diferentes da variância dos erros.

Os pontos de corte sugeridos para as diferentes estatísticas de influência são na verdade puramente referenciais. Na prática o analista é quem decide quando o valor da estatística pode ser

considerado notoriamente grande, baseado na sua experiência e conhecimento dos dados. Por exemplo, uma forma de agirmos poderia ser a seguinte: calcular as estatísticas de influência para todas as observações e considerar como grandes aquelas observações que têm valores notoriamente maiores do que as outras.

As estatísticas de influência apresentadas combinam resíduos "studentizados" externamente ou internamente com os elementos da diagonal da matriz H. Logo precisamos analisar cada um dos termos da estatística para termos uma idéia mais clara do que significa um valor grande ou pequeno da estatística de influência que se está considerando.

3.4.4. Sumário

Resumindo os resultados desta seção temos

- A identificação de dados discrepantes realiza-se com a finalidade de ressaltar observações suspeitas, estudando seu efeito nas diferentes fases da análise. Se observações discrepantes são identificadas e, estas não afetam a fase da análise de dados a qual é de interesse, então, permanecerão no conjunto de dados.

- Os métodos de diagnósticos que apresentamos nésta seção são usados para detectar uma observação discrepante de cada vez. Na presença de múltiplas observações discrepantes estes métodos não necessariamente os identificam, pois uma observação discrepante pode mascarar outras. Pode também ocorrer que dados seguindo o padrão da maioria sejam diagnosticados como discrepantes. Este fenômeno é denominado *efeito de* "swamping".

- Os métodos de diagnósticos de múltiplas observações discrepantes são uma solução para o problema de mascaramento. As estatísticas para estes diagnósticos são calculadas retirando grupos de m > 1 dados.

de possíveis subconjuntos de (n-m) observações a serem analizadas cresce rapidamente com n. Dado que não conhecemos o número exato de observações discrepantes existentes no conjunto de dados, a análise deverá ser realizada com m = 2, 3, ..., n-k. Isso pode levar a um trabalho computacional muito grande e quase impossível de ser realizado em algumas situações.

Em consequência, precisamos de métodos de detecção de múltiplas observações discrepantes que não sofram o efeito de mascaramento.

CAPITULO 4

OBTENÇÃO DE ESTIMADORES COM ALTO PONTO DE RUPTURA

4.1. INTRODUÇÃO

Os estimadores de mínimos quadrados e mínima soma de desvios absolutos têm em comum a forma de construição da função objetivo a minimizar, isto é, aplica-se um funcional sobre o vetor de resíduos observados. Eles são membros de uma clase geral de estimadores denominados *estimadores norma* L_p, e os funcionais aplicados pertencem à classe de normas vetoriais L_p. Estes estimadores serão apresentados na seção 4.2, incluindo as condições sob as quais existe uma solução para o problema de minimização e os casos em que esta solução é única. Nesta mesma seção apresentaremos a geometria da função objetivo do problema de minimização e exporemos a relação entre os estimadores norma L_p, M-estimadores e de máxima verossimilhança, com a finalidade de explicar as razões pelas quais os estimadores norma L_p são afetados pelas observações discrepantes.

Na seção 4.3, estudaremos modificações da função objetivo dos

estimadores norma L_p, afim de obtermos estimadores robustos com alto ponto de ruptura.

4.2 DEFINIÇÃO DE ESTIMADORES DE REGRESSÃO LINEAR NORMA L_p

Uma forma geral de apresentar o problema de estimação dos parâmetros do modelo (1.1) é através da minimização da distância do vetor de observações $y \in \mathbb{R}^n$ a um hiperplano gerado pela combinação linear dos vetores $x_1, x_2, \ldots, x_k$; $x_j \in \mathbb{R}^n$. Esta distância será medida por uma métrica convenientemente escolhida. Como qualquer norma pode ser deduzida a partir de uma métrica, cada norma L_p dará lugar a um estimador (definição A5). Um estimador de regressão norma L_p é um vetor $\tilde{\theta} \in \mathbb{R}^k$ que minimiza a norma L_p do vetor de resíduos, isto é,

$$W(\tilde{\theta}) = \min_{\theta} W(\theta) \qquad \text{para } \theta \in \mathbb{R}^k,$$

onde

$$W(\theta) = \begin{cases} \displaystyle\sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} x_{ij}\,\theta_j \right|^p & 1 \le p < \infty, \\ \displaystyle\max_{1 \le i \le n} \left| y_i - \sum_{j=1}^{k} x_{ij}\,\theta_j \right| & p = \infty. \end{cases} \tag{4.1}$$

Como casos particulares, são conhecidos seguintes estimadores:

i) Estimador norma L_1, o qual é uma generalização da expressão (1.3) para regressão linear múltipla.

$$\min_{\theta \in \mathbb{R}^k} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} x_{ij}\,\theta_j \right|. \tag{4.2}$$

Este estimador é conhecido também como estimador LAV (Least absolute value).

ii) Estimador norma L_2 ou estimador de mínimos quadrados; o qual foi dado na expressão (1.4).

iii) Estimador norma L_∞, também conhecido como estimador minimax ou estimador de Chebyshev,

$$\underset{\theta \in \mathbb{R}^k}{\text{Min}} \left\{ \underset{1 \le i \le n}{\text{Max}} \left| y_i - \sum_{j=1}^{k} x_{ij}\, \theta_j \right| \right\}. \qquad (4.3)$$

4.2.1 Existência e Unicidade dos Estimadores Norma L_p

Para cada vetor $\tilde{\theta}$ que minimiza a função objetivo, $W(\theta)$, existe um vetor $\hat{y} = X\,\tilde{\theta}$. Todos os vetores $\hat{y}$ assim definidos formam um conjunto que denotaremos por $\mathbb{P}_G(y)$. Dado que a matriz X é de posto completo, o vetor $\tilde{\theta}$ é unicamente determinado para cada vetor $y \in \mathbb{P}_G(y)$. Os vetores $x_1, x_2, \ldots, x_k$, y e $\hat{y}$ são elementos do espaço vetorial norma L_p.

Seja G o conjunto de todos os vetores que são combinações lineares dos vetores $x_1, x_2, \ldots, x_k$, isto é,

$$G = \left\{ g \in L_p \,:\, g = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k \;;\; \theta_j \in \mathbb{R},\; j = 1,2,\ldots,k \right\}$$

é um subespaço linear do espaço L_p gerado pelos vetores $x_1, x_2, \ldots, x_k$. Logo, os vetores $\hat{y}$ são aqueles que satisfacem:

$$\underset{g \in G}{\text{Min}} \; \| y - g \| = \| y - \hat{y} \|. \qquad (4.4)$$

definindo $\mathbb{P}_g(y)$ formalmente temos

$$\mathbb{P}_G(y) = \left\{ \hat{y} \in G : \| y - \hat{y} \| = \inf_{g \in G} \| y - g \| \right\}. \qquad (4.5)$$

O subespaço linear G é de dimensão finita.

Precisamos agora garantir que o conjunto $\mathbb{P}_G(y)$ seja não vazio, isto é, que exista no mínimo uma solução para o problema de otimização. Cheney (1966) demostrou que dado um espaço linear normado, existe pelo menos um ponto com distância mínima a partir do ponto fixado. Baseados neste resultado podemos dizer que no espaço normado L_p existe no mínimo um vetor $\hat{y} \in \mathbb{P}_G(y)$, isto é, existe pelo menos uma estimativa norma L_p para o conjunto de dados. O teorema a seguir apresenta os valores de p para os quais $\mathbb{P}_G(y)$ tem apenas um elemento.

TEOREMA 4.2.1

Seja G um subespaço linear do espaço norma L_p, $1 < p < \infty$, então existe exatamente um elemento $\hat{y}$ em $\mathbb{P}_G(y)$.

PROVA

A prova do teorema baseia-se no fato de que os espaços lineares com norma L_p, $1<p<\infty$, são estritamente convexos (Definição A7).

Suponhamos que $\hat{y}$ e $\hat{y}' \in \mathbb{P}_G(y)$. Logo, pela definição de $\mathbb{P}_G(y)$ (expressão 4.5), temos

$$\min_{g \in G} \| y - g \| = \| y - \hat{y} \|,$$

e

$$\min_{g \in G} \| y - g \| = \| y - \hat{y}' \|.$$

Se

$$\delta = \| y - \hat{y} \| = \| y - \hat{y}' \|,$$

então, pela desigualdade triangular temos

$$\delta \leq \| y - (\hat{y} + \hat{y}')/2 \| \leq \frac{1}{2} \| y - \hat{y} \| + \frac{1}{2} \| y - \hat{y}' \| = \delta.$$

Logo, pela convexidade estrita, existe algum $z \in L_p$ tal que

$$y - \hat{y} = \alpha z \qquad \text{e} \qquad y - \hat{y}' = \beta z.$$

Mas dado que

$$\| y - \hat{y} \| = \| y - \hat{y}' \|,$$

então

$$\alpha = \beta \; ; \quad \alpha, \beta \in \mathbb{R}^+.$$

Portanto,

$$\hat{y} = \hat{y}'. \qquad \blacksquare$$

O teorema 4.2.1 não se aplica para $p = 1$ e $p = \infty$, pois os espaços lineares L_1 e L_∞ não são estritamente convexos.

4.2.2 Representacão Geométrica dos Estimadores Norma L_p

Para facilitar as interpretações, usaremos o modelo linear com duas variáveis independentes e três observações

$$G = \left\{ g : g = \theta_1 x_1 + \theta_2 x_2 \; ; \; \theta_i \in \mathbb{R}; \; i=1,2 \right\}.$$

As figuras 4.2.1 e 4.2.2 representam as bolas associadas aos estimadores de mínimos quadrados e mínimo valor absoluto, respectivamente.

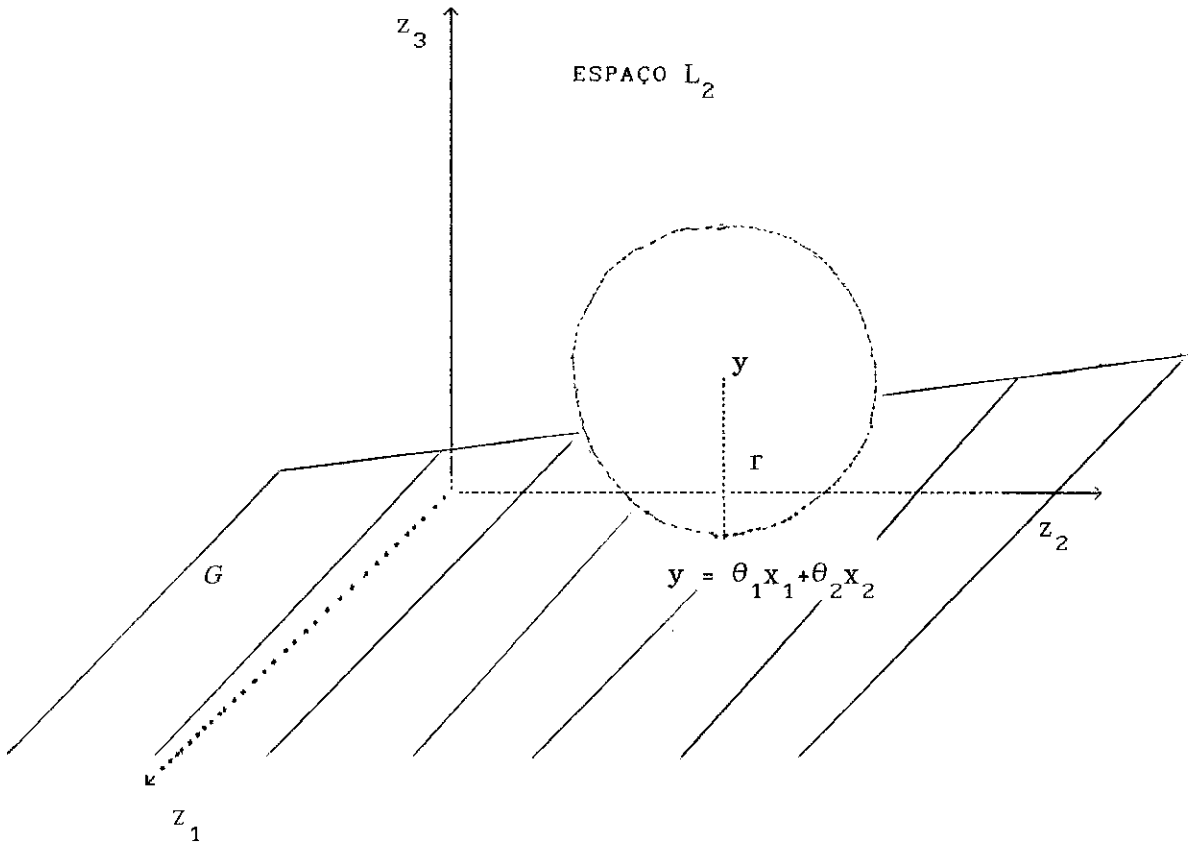

Figura 4.2.1 O espaço L_2 e o conjunto de todos os vetores com distância r partindo do vetor y.

Consideremos todos os vetores que estão a uma certa distância r do vetor $y \in \mathbb{R}^3$. Estamos então considerando no espaço L_2 o conjunto

$$\left\{ s \in L_2 : \left(\sum_{j=1}^{3} \left(y_i - s_i \right)^2 \right)^{1/2} = r \right\} .$$

o qual forma uma esfera com centro em $y \in \mathbb{R}^3$ e raio r. Para um raio suficientemente grande, a esfera toca o plano $g \in G$. Quando $n > 3$ o subespaço G será formado por hiperplanos e o conjunto de vetores com distância r do vetor y será uma hiperesfera. Para um raio suficientemente grande a hiperesfera toca o hiperplano $g \in G$ mais próximo em apenas um ponto. Este ponto tangente corresponde a $\hat{y} = X \tilde{\theta}$. Logo, sempre existe uma única estimativa norma L_2.

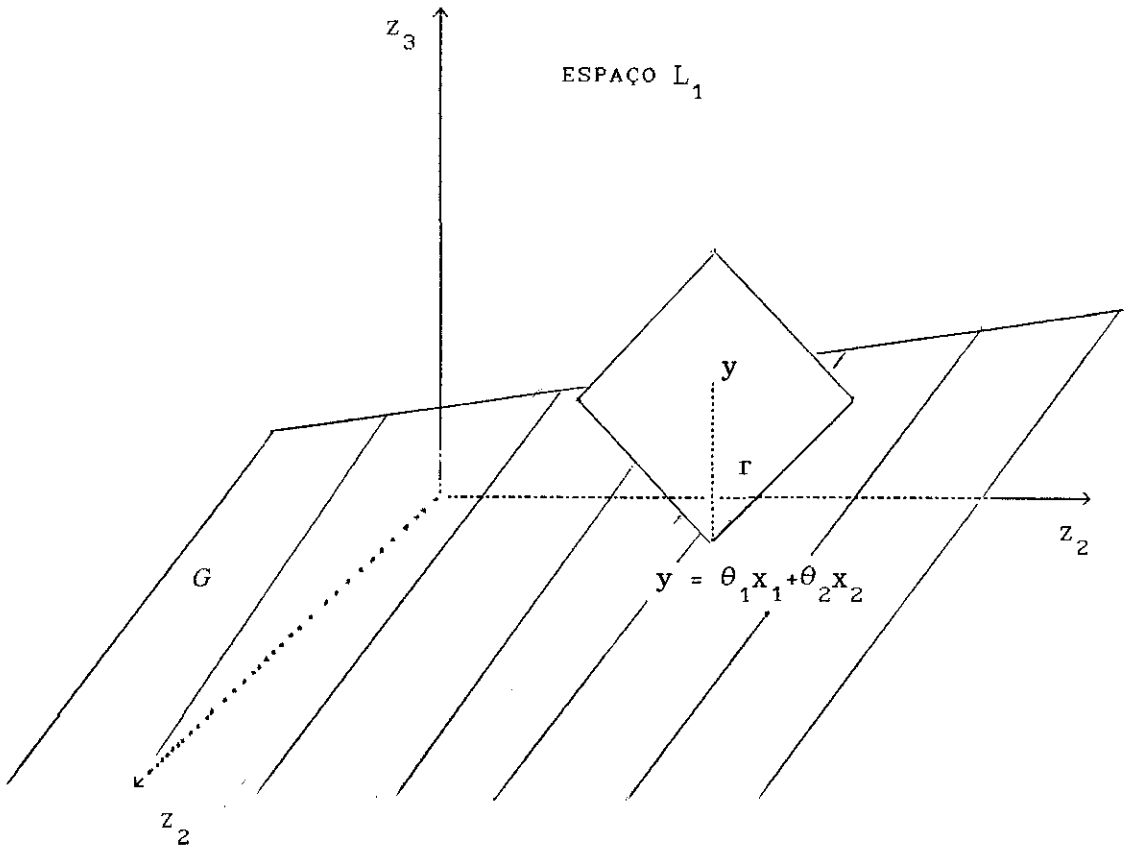

Figura 4.2.2 O espaço L_1 e o conjunto de todos os vetores com distância r partindo do vetor y.

O conjunto de pontos que estão a uma certa distância r a partir do vetor y, segundo a métrica L_1 é dado por

$$\left\{ s \in L_1 : \sum_{j=1}^{3} |y_i - s_i| = r \right\} .$$

Este conjunto tem a forma de um rombóide com centro em $y \in \mathbb{R}^3$ e com diagonais de comprimento 2r paralelas aos eixos de coordenadas. Para r suficientemente grande, o rombóide toca o plano $g \in G$. Em consequência, está garantida a existência de no mínimo uma estimativa norma L_1. O romboide e o plano $g \in G$ podem interceptar-se em um único ponto ou em uma aresta, ou em uma face do rombóide. Logo, as estimativas não necessariamente são únicas.

A figura 4.2.3 mostra as características da bola associada ao estimador de Chebyshev.

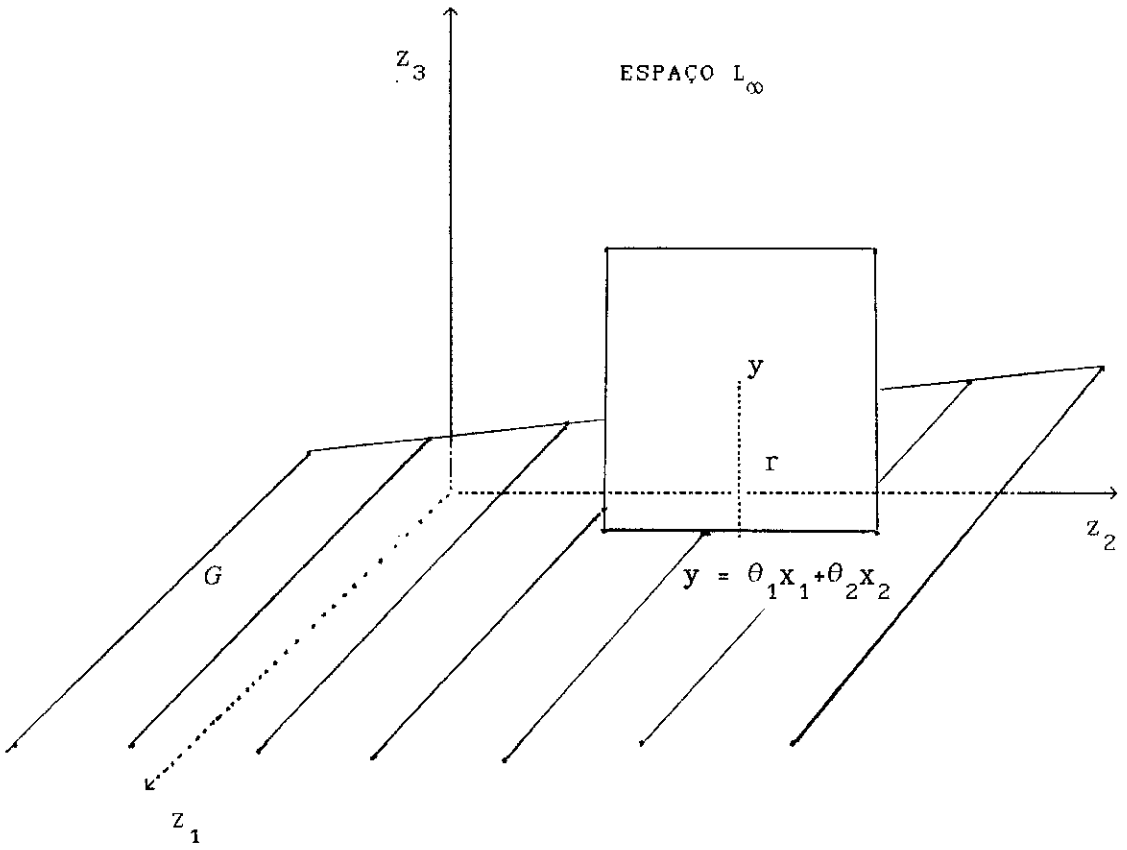

Figura 4.2.3 O espaço L_∞ e o conjunto de todos os vetores com distância r partindo do vetor y.

O conjunto dos vetores que estão a uma distância r do vetor $y \in \mathbb{R}^3$, segundo a métrica L_∞, será dado por

$$\left\{ s \in L_\infty : \max_{1 \le i \le 3} \left(y_i - s_i \right) = r \right\},$$

os mesmos formam um cubo com centro em $y \in \mathbb{R}^k$ e lados de comprimento 2r, paralelos aos eixos coordenados. Como no caso dos estimadores norma L_1 a existência de no mínimo uma solução está garantida, porém não se garante a unicidade.

4.3 RELAÇÃO DOS ESTIMADORES NORMA L_p, M-ESTIMADORES E DE MÁXIMA VEROSSIMILHANÇA

Huber (1964) e outros propuseram uma clase geral de estimadores denominados M-*estimadores*, os quais foram posteriormente generalizados para modelos de regressão linear múltipla..

O M-estimador de regressão foi definido como aquele vetor $\tilde{\theta}$ que minimiza a função

$$\sum_{i=1}^{n} \rho \left(y_i - \sum_{j=1}^{k} x_{ij}\, \theta_j \right), \qquad \theta_j \in \mathbb{R}. \tag{4.6}$$

A função $\rho(t)$ é escolhida convenientemente com o objetivo de minimizar a importância dos resíduos grandes. Em geral, pede-se que ρ seja não negativa, simétrica, convexa e que $\rho(t) \longrightarrow \infty$ quando $t \longrightarrow \infty$. Se considerarmos as funções ρ que são deriváveis teremos $\rho'(t) = \psi(t)$.

As estimativas serão obtidas resolvendo o sistema de equações

$$\sum_{i=1}^{n} \psi \left(y_i - \sum_{j=1}^{k} x_{ij}\, \theta_j \right) = 0. \tag{4.7}$$

Se definirmos $\rho(t) = -\log f(t)$, onde $f(t)$ é função densidade dos erros, ε_i, obtemos o estimador de máxima verossimilhança do vetor de parâmetros θ. Se $\rho(t) = t^p$ temos que os estimadores norma L_p são M-estimadores. Podemos então concluir que, os estimadores de máxima verossimilhança e norma L_p são subconjuntos da classe dos M-estimadores. Interessa-nos agora, saber se existe interseção entre estes dois subconjuntos. A seguir, verificaremos que para uma certa família de distribuições de probabilidade, os estimadores de máxima verossimilhança e norma L_p coincidem.

Turner (1960) considerou a seguinte família de funções de densidade de probabilidade

$$f(y) = \frac{\gamma}{2\delta\Gamma(1/\gamma)} \exp\left\{-\frac{|y - \theta|^{\gamma}}{\delta^{\gamma}}\right\}, \qquad (4.8)$$

para $y \in \mathbb{R}$; $\theta \in \mathbb{R}$; $\delta > 0$; $\gamma > 0$; $\Gamma(.)$ denota a função Gamma. Mudando o valor de γ em (4.8) obtêm-se diferentes funções de densidade de probabilidade. Por exemplo, se :

i) $\gamma = 1$, temos a distribuição de Laplace ou exponencial dupla.

$$f(y) = \frac{1}{2\delta} \exp\left\{-\frac{|y - \theta|}{\delta}\right\}; \qquad y \in \mathbb{R};\ \theta \in \mathbb{R};\ \delta > 0. \qquad (4.9)$$

ii) $\gamma = 2$, a distribuição é normal com esperança θ e variância $\delta^2/2$.

$$f(y) = \frac{1}{\delta\sqrt{\pi}} \exp\left\{-\frac{(y - \theta)^2}{\delta^2}\right\}; \qquad y \in \mathbb{R};\ \theta \in \mathbb{R};\ \delta > 0. \qquad (4.10)$$

iii) $\gamma \longrightarrow \infty$, no limite a distribuição é uniforme.

$$f(y) = \frac{1}{2\delta}, \qquad |y - \theta| \leq \delta. \qquad (4.11)$$

A função de verossimilhança para uma amostra aleatória i.i.d. de tamanho n, com função de densidade f dada em (4.8) será

$$\ell(\theta, \delta; y) = \ln \gamma^n - \ln 2\delta\, \Gamma(1/\gamma) - \frac{1}{\delta^{\gamma}} \sum_{i=1}^{n} |y_i - \theta|^{\gamma}.$$

O estimador de máxima verossimilhança de θ será aquele vetor $\tilde{\theta}$ que minimiza

$$\sum_{i=1}^{n} |y_i - \theta|^{\gamma}.$$

No caso da distribuição uniforme ($\gamma \longrightarrow \infty$), o estimador de máxima verossimilhança será

$$\tilde{\theta} = \lim_{\gamma \to \infty} \sum_{i=1}^{n} |y_i - \theta|^{\gamma},$$

ou

$$\tilde{\theta} = \max_{0 \leq i \leq n} |y_i - \theta|.$$

Voltando ao modelo de regressão linear temos que os estimadores de máxima verossimilhança e norma L_p são equivalentes quando a componente ε do modelo (1.1) é uma variável aleatória cuja distribução corresponde a um dos membros da família (4.8).

No modelo de posição, $Y = \theta + \varepsilon$, a estimativa norma L_1 de θ é a mediana amostral, e as estimativas norma L_2 e L_∞ são a média aritmética e a amplitude média, respectivamente. Lembrando que a média aritmética é sensível à presença de dados discrepantes, enquanto que a mediana amostral é resistente a essas observações, segue que, o estimador norma L_1 é preferível quando a distribuição dos resíduos tem caudas longas. Por outro lado, quando a distribuição dos resíduos tem extremos bem definidos, como a distribuição uniforme, é de se esperar que o estimador L_∞, o qual é mais sensível a resíduos grandes do que a norma L_1, seja o adequado.

4.4 CONSTRUÇÃO DE ESTIMADORES COM PONTO DE RUPTURA MÁXIMO

Das propriedades dos estimadores de regressão norma L_p, vimos que estes estimadores não são resistentes a pontos de alavanca. A geometria da bola associada a estes estimadores nos fornece a explicação deste fato. As bolas associadas aos estimadores norma L_p são limitadas tornando-as muito sensíveis a mudanças nos dados. Isso se

reflete no baixo ponto de ruptura dos estimadores norma L_p.

Nosso interesse agora é mudar a geometria das bolas associadas aos estimadores norma L_p, afim de que admitam uma certa porcentagem de observações discrepantes sem alterar demasiadamente o valor da estimação, ou seja, pretendemos construir estimadores com ponto de ruptura o mais alto quanto possível.

para mudarmos a forma geométrica das bolas associadas aos estimadores norma L_p, podemos dividir o conjunto de pontos no espaço L_p em duas partes: a primeira com resíduos que chamaremos de "pequenos", que correspondem aos pontos localizados perto do hiperplano ajustado, e a outra com resíduos "grandes", ou seja, aqueles pontos que se afastam do hiperplano ajustado. A classificação dos resíduos como "pequenos" ou "grandes" depende do ponto de ruptura desejado e do número de dados. Por exemplo, se desejamos um estimador com ponto de ruptura de 20% e temos 5 observações classificaremos 4 observações com os resíduos "pequenos" e uma com o resíduo "grande".

Outro aspecto que devemos levar em consideração na construção dos estimadores com alto ponto de ruptura é que a função objetivo a ser utilizada proporcione *separação total* dos dados, isto é, o subconjunto de dados com resíduos "pequenos" deve estar bem identificado. Aplicaremos um dos estimadores norma L_p sobre o conjunto de dados com resíduos "pequenos" para obtermos as estimativas e os resíduos.

Resumindo, os passos para construirmos estimadores com alto ponto de ruptura são os seguintes:

- Selecionar o ponto de ruptura desejado, ($\varepsilon^* \leq 0.50$).
- Definir um critério de divisão dos dados em função do ponto de ruptura desejado, de modo que possamos identificar claramente um grupo de resíduos "pequenos" e outro de "grandes".
- Estimar o vetor de parâmetros θ aplicando um funcional previamente

selecionado sobre o conjunto de dados com resíduos "pequenos". O funcional neste caso será uma norma L_p, $p \geq 1$.

O ponto de ruptura pretendido não é sempre atingido na prática, mas podemos conseguir valores muito próximos aos desejados.

4.4.1 Exemplos de Estimadores com Alto Ponto de Ruptura

Os estimadores que apresentaremos a seguir foram construídos seguindo os três passos antes mencionados.

a) Estimador de Mínima Mediana dos Quadrados dos Resíduos (LMS)

Rousseeuw (1984) propôs o estimador denominado "*estimador de mínima mediana dos quadrados dos resíduos*" (LMS), que tem ponto de ruptura 50%. Este estimador foi definido como o vetor θ que minimiza

$$W(\theta) = \underset{1 \leq i \leq n}{\text{Mediana}} \left(e_i\right)^2; \qquad \theta \in \mathbb{R}^k, \tag{4.13}$$

ou seja,

$$\tilde{\theta} = \underset{\theta}{\text{Min}} \ \underset{1 \leq i \leq n}{\text{Mediana}} \left(e_i\right)^2; \qquad \theta \in \mathbb{R}^k \tag{4.14}$$

Onde a mediana é definida como a estatística de ordem [n/2] + 1 dos resíduos $e_i^2 = \left(y_i - x_i\theta\right)^2$; i=1,2,...,n. Podemos melhorar o ponto de ruptura do estimador (4.14) considerando a q-ésima estatística de ordem dos quadrados dos resíduos. Assim

$$\tilde{\theta} = \underset{\theta}{\text{Min}} \ e^2_{q:n} \ ; \qquad \theta \in \mathbb{R}^k, \tag{4.15}$$

onde q = [n/2] + [(k+1)/2]. Tanto em (4.14) quanto em (4.15) o estima-

dor tem ponto de ruptura assintótico 50%, porém em amostras finitas o ponto de ruptura de (4.15) é ligeiramente maior.

A formulação (4.15) é equivalente a

$$\tilde{\theta} = \underset{\theta}{\text{Min}} \quad \underset{1\le i\le q}{\text{Máximo}} \; e^{2}_{i:n} \qquad ; \qquad \theta^{k} \in \mathbb{R}. \tag{4.16}$$

Na expressão (4.16) observamos claramente que o estimador LMS formulado em (4.15) é um membro da família de estimadores com alto ponto de ruptura, construído aplicando a norma do máximo sobre o conjunto de resíduos "pequenos". O critério de divisão dos resíduos em "grandes e "pequenos" é dado pela q-ésima estatística de ordem dos quadrados dos resíduos do ajuste.

A formulação do problema de estimação LMS de (4.14) como um problema de estimação minimax (ou de Chebyshev) foi de grande utilidade pois facilitou a busca de algoritmos exatos para resolver o problema de otimização.

A figura 4.4.1 apresenta a geometria da bola associada ao estimador LMS. A representação é feita no espaço $\mathbb{R}^3$ e

$$G = \left\{ g \; : \; g = \theta_1 x_1 + \theta_2 x_2 \; ; \; \theta_i \in \mathbb{R}; \; i=1,2 \right\}.$$

A característica mais notável da bola no gráfico 4.4.1 é que a mesma possui contornos que se extendem ao longo dos eixos até o infinito. Assim, por exemplo, os pontos $(0,0,\infty)$ e $(0,0,0)$ estão a uma distância zero do centro e encontram-se dentro da bola LMS. Isto mostra que, pontos que estão próximos, de acordo com a métrica associada ao estimador LMS, podem estar muito longe sob métricas convexas limitadas.

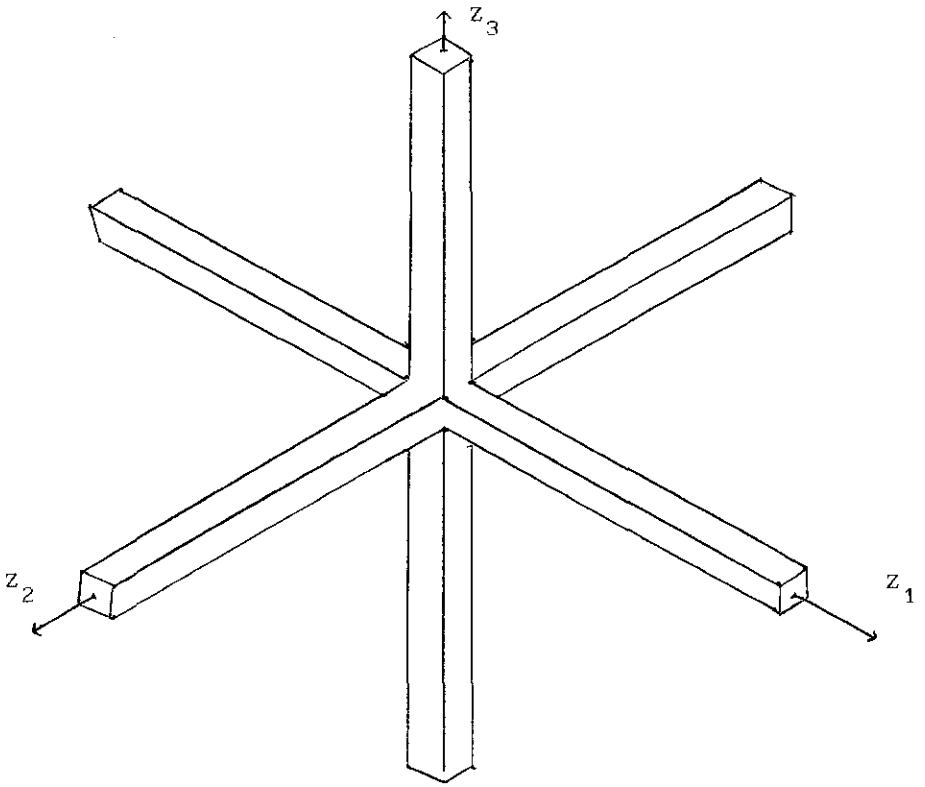

Figura 4.4.1 Geometria da bola associada ao estimador LMS.

A figura 4.4.2 apresenta uma visão bidimensional da bola associada ao estimador LMS, considerando a interseção da bola no espaço $\mathbb{R}^3$ com um plano bidimensional passando pela origem e sendo paralela a um eixo.

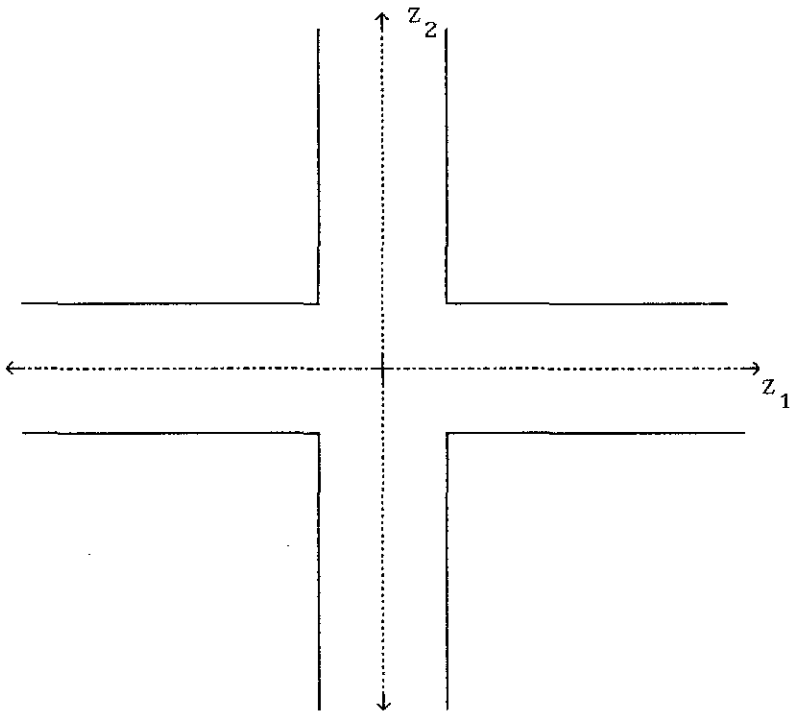

Figura 4.4.2 Interseção da bola associada ao estimador LMS e um plano que passa pela origem e que é paralela a um eixo.

Geometricamente, a solução LMS corresponde a encontrar a faixa mais estreita, que cobre pelo menos a metade das observações. A amplitude dessa faixa é medida na direção vertical e, espera-se que pelo menos h = [n/2] + [(k+1)/2] + 1 pontos estejam nela contidos. O estimador será robusto no sentido de resistência a observações discrepantes sempre que pelo menos 50% das observações estejam entre os dois hiperplanos, $X\theta \pm \rho$, onde ρ é a amplitude da faixa.

O estimador LMS tem as propriedades de equivariância sob transformações lineares na resposta (de regressão), e sob transformações afins na matriz do modelo. Além disso, tem o maior ponto de ruptura entre todos os estimadores equivariantes de regressão.

Como o estimador LMS é gerado pela norma de Chebyshev, está garantido que existe pelo menos uma estimativa, porém, esta pode não ser única.

Um resultado interessante que mostra que o ponto de ruptura do estimador LMS converge assintóticamente para 1/2 é dado pelo teorema seguinte.

TEOREMA 4.4.1 (Rousseeuw e Leroy (1987), pp. 118-120)

Se k > 1, e as observações estão em *posição geral* (Def. A9), então o ponto de ruptura do estimador LMS é

$$\varepsilon_n^*(t, Z) = ([n/2]-k+2)/n. \qquad (4.17)$$

O teorema seguinte mostra que sob certas condições existe um hiperplano que interpola um subconjunto de observações que estão em posição geral.

TEOREMA 4.4.2 (Rousseeuw e Leroy (1987), pp. 122-123)

Se k > 1, e existe θ tal que pelo menos n-[n/2]+k-1 das observações satisfaçam exatamente o modelo (1.1) além disso, os dados estão em posição geral, então, a solução LMS é igual a θ, quaisquer que sejam as outras observações.

Observação.- Este estimador não é eficiente, mas como estamos interessados em métodos exploratórios para detecção de observações atípicas, esta característica do estimador LMS não é muito importante em nosso estudo.

b) Estimador de Mínima Soma dos Quadrados dos Resíduos Aparados (LTS).

Rousseeuw e Leroy (1987), propuseram um estimador denominado Estimador de Mínima Soma Aparada dos quadrados dos Resíduos (LTS),

definido como

$$\tilde{\theta} = \underset{\theta}{\text{Min}} \sum_{i=1}^{q} \left(e_i\right)^2 \; ; \qquad \theta \in \mathbb{R}^k, \tag{4.18}$$

onde $e_i = \left(y_i - x_i\theta\right)$, e $q = [n/2] + [(k+1)/n]$. Este estimador tem ponto de ruptura 1/2 e é obtido utilizando o seguinte critério de divisão:

. Obter a q-ésima estatística de ordem dos resíduos.

. Considerar como "pequenos" os resíduos menores o iguais a $e^2_{q:n}$.

O funcional é dado pela norma L_2, isto é, o estimador com ponto de ruptura 1/2 é dado por aquele valor de $\theta \in \mathbb{R}^k$ que minimiza a soma de quadrados dos resíduos, calculado sobre o conjunto de dados com resíduos "pequenos".

Na figura 4.4.3 observamos que a geometria da bola associada ao estimador LTS, é similar à correspondente ao estimador LMS sendo que seus contornos não são quadrados mais sim tubos circulares.

Este estimador é mais eficiente do que o estimador LMS, além disso, tem convergência assintótica de ordem $O\left(n^{1/2}\right)$ a uma distribuição normal. No entanto, ele é computacionalmente mais lento do que o estimador LMS, tornando-o pouco utilizado como método exploratório.

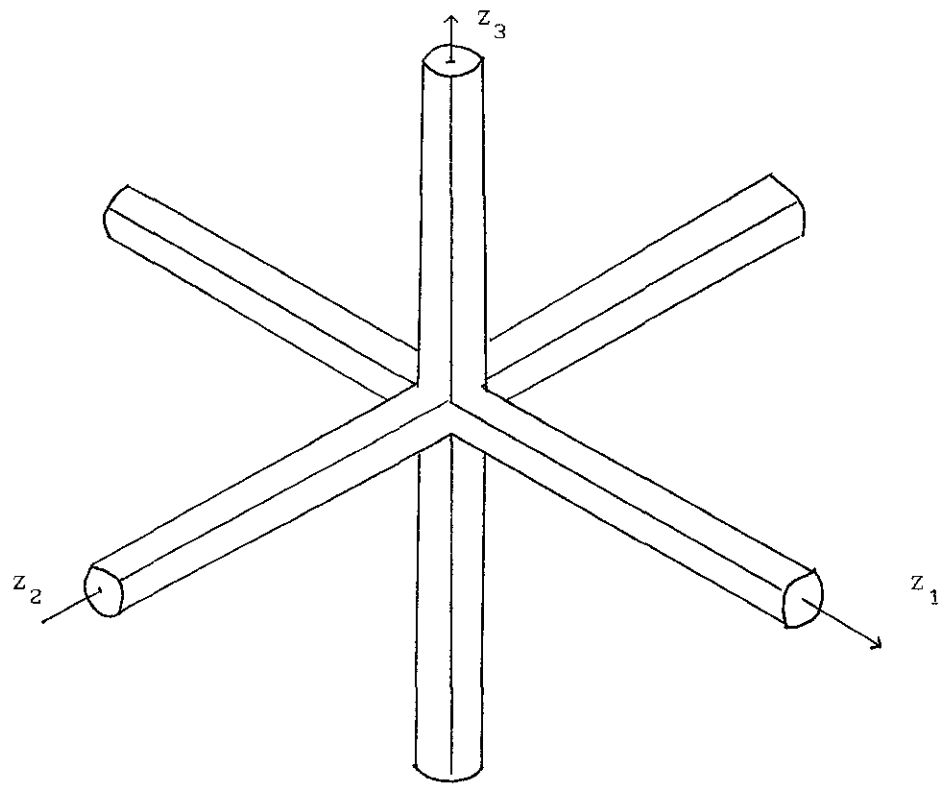

Figura 4.4.3 Geometria da bola associada ao estimador LTS.

4.5 SUMARIO

Resumindo os resultados deste capítulo temos:

- Os estimadores de regressão norma L_p são ótimos, se a distribução de ε no modelo de regressão (1.1) é um membro da família de distribuições definidas em (4.8).

- O estimador norma L_1 é robusto com respeito às observações discrepantes, porém ele é muito afetado pelos pontos de alavanca. Já o estimador de mínimos quadrados é afetado tanto pelas observações discrepantes na resposta quanto pelos pontos de alavanca.

- A falta de robustez dos estimadores norma L_p deve-se ao fato de que as bolas associadas a estes estimadores são limitadas.

- Os estimadores com alto ponto de ruptura são construídos modificando a geometria das bolas associadas aos estimadores norma L_p, fazendo-as não limitadas. Estes estimadores não provocam o efeito de mascaramento, portanto, os diagnósticos serão mais confiáveis do que àqueles produzidos pelo estimador de mínimos quadrados.

- O estimador LMS é o mais conhecido membro desta família de estimadores com alto ponto de ruptura. Os resíduos do ajuste calculados empregando este estimador são utilizados para detectar observações discrepantes da regressão.

No capítulo seguinte estudaremos métodos de identificação de observações discrepantes na matriz do modelo.

CAPITULO 5

ELIPSÓIDE DE VOLUME MÍNIMO E SUA UTILIZAÇÃO NA IDENTIFICAÇÃO DE DADOS DISCREPANTES

5.1 INTRODUÇÃO

O problema de identificação de observações discrepantes, quando ocorre o problema de mascaramento, foi resolvido apenas em parte no capítulo anterior. Os estimadores de regressão com alto ponto de ruptura identificam observações que são discrepantes, porém não fornecem qualquer informação com respeito ao motivo da discrepância, ou seja, se a mesma é devida à resposta, as variáveis regressoras ou ambas. Neste capítulo apresentaremos métodos de detecção de pontos de alavanca que não sofrem o efeito de mascaramento.

Considerando que a matriz do modelo, X, está formada por n vetores linhas de dimensão k, a busca de métodos robustos para identificar pontos de alavanca será tratada no contexto da análise multivariada de dados.

A necessidade de caracterizar um ponto discrepante no espaço das variáveis regressoras levou-nos ao emprego de métodos de

sub-ordenamento dos pontos através de uma medida de distância a qual nos diz o quão longe encontra-se um ponto do seu centroide, considerando, a dispersão do conjunto total de pontos no espaço das variáveis regressoras. Quando os parâmetros de posição e dispersão necessários para obtermos a distância são estimados pela média e covariância amostrais, a distância assim definida é chamada de *distância de Mahalanobis*.

Na seção 2 verificaremos que a distância de Mahalanobis de um vetor pertencente à matriz do modelo, é uma função monótona crescente do elemento correspondente da diagonal da matriz de projeção. Esta relação explica a vulnerabilidade dos elementos da diagonal da matriz de projeção quando o conjunto de dados apresenta múltiplas observações discrepantes.

Na seção 3, apresentaremos estimadores robustos para os parâmetros de posição e dispersão do elipsóide de volume mínimo que tem ponto de ruptura 50%, consequentemente a medida de distância obtida a partir desses estimadores será robusta.

Finalmente, na seção 4 daremos um método de diagnósticos de observações discrepantes que consiste em um gráfico bidimensional relacionando às distâncias robustas e os resíduos do ajuste, obtidos empregando um dos estimadores de regressão com alto ponto de ruptura propostos no capítulo 4.

5.2 IDENTIFICAÇÃO DE OBSERVAÇÕES DISCREPANTES EM CONJUNTOS DE DADOS MULTIVARIADOS

No caso univariado, caracterizamos uma observação discrepante como uma observação extrema que está notoriamente distante da massa de dados, percebendo-se a discrepância por simples inspeção. Em conjuntos de dados bivariados ainda podemos identificar estas observações com

ajuda dos gráficos de dispersão. Porém, em dimensões maiores do que 2 a detecção de observações discrepantes é mais complicada e precisamos de métodos mais elaborados para identifica-las.

A idéia de observação extrema utilizada na caracterização de observações discrepantes univariadas surge naturalmente de algum tipo de "ordenamento" dos dados, mas na análise multivariada não existe uma única forma de ordenamento a qual seja clara e definida. O máximo que podemos fazer é estabelecer um tipo de sub-ordenamento. Uma forma de sub-ordenamento muito usada na identificação de observações que se afastam da massa de dados, consiste em empregar alguma medida de distância univariada, tal como

$$D\left(x;\ x_0,\ \Gamma\right) = \left(x - x_0\right)\ \Gamma^{-1}\ \left(x - x_0\right)^t, \tag{5.1}$$

onde x_0 representa um parâmetro de posição, e Γ está relacionado com a dispersão dos dados.

Suponhamos que o conjunto de dados $X = \left\{x_1,\ x_2,\ \ldots,\ x_n\right\}$, $x_i \in \mathbb{R}^k$, corresponde a uma amostra aleatória de uma distribuição normal multivariada com parâmetros média e variância $\mu_{k\times 1}$ e $\Sigma_{k\times k}$ respectivamente. Neste caso os estimadores consistentes não viciados de mínima variância que empregaremos para estimar μ e Σ serão a média de variância amostrais dados por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left(x_1,\ \ldots,\ x_k\right) \tag{5.2}$$

e

$$C = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^t\left(x_i - \bar{x}\right), \tag{5.3}$$

respectivamente. A distância de Mahalanobis (MD_i) é dado por

$$MD_i = \left(x_i - \bar{x}\right) C^{-1} \left(x_i - \bar{x}\right)^t, \qquad i=1,2,\ldots,n. \tag{5.4}$$

Estas distâncias têm distribuição chi-quadrado com k graus de liberdade. Para decidirmos se uma observação encontra-se afastada da massa dos dados, comparamos sua distância de Mahalanobis com um quantil da distribuição χ^2_k previamente fixado. Por exemplo, se escolhermos o quantil 95% podemos esperar que somente 5% dos dados que provém de uma distribuição normal multivariada estejam fora do elipsóide

$$\left(x_i - \bar{x}\right) C^{-1} \left(x_i - \bar{x}\right)^t \leq b, \tag{5.5}$$

onde b é o quantil 95%. Esses 5% das observações podem ser vistos como discrepantes.

Notemos que tanto o elipsóide (5.5) quanto o chamado contorno das variáveis regressoras definido na expressão (3.9) utilizam somente a matriz do modelo em sua construição. A seguir mostraremos a relação existente entre a distância de Mahalanobis e os elementos da diagonal da matriz H, isto é, mostraremos a relação entre os elipsóides definidos nas expressões (3.9) e (5.5).

Consideremos um modelo linear com intercepto

$$Y = \theta_0 + \theta_1 X_1 + \ldots. + \theta_k X_k + \varepsilon. \tag{5.6}$$

A matriz do modelo pode ser escrita como

$$X_{nx(k+1)} = \left(1_{nx1}\ V_{nxk}\right),$$

onde 1_{nx1} é um vetor coluna de uns; $V_{nxk} = \left(x_1,\ldots,x_k\right)$, $x_i \in \mathbb{R}^n$ é a matriz de observações correspondente às k variáveis regressoras. A matriz de desvios das observações contidas nas linhas de V com relação a sua média aritmética a qual denotaremos por $\mathfrak{X}$, é dada por

$$\mathcal{X}_{n\times k} = V_{n\times k} - 1_{n\times 1}\,\bar{x}_{1\times k},$$

calculamos o produto X^tX fazendo

$$\begin{pmatrix} 1 & \mathcal{X} \end{pmatrix}^t \begin{pmatrix} 1 & \mathcal{X} \end{pmatrix} = \begin{bmatrix} n & 0 \\ 0 & \mathcal{X}^t\mathcal{X} \end{bmatrix},$$

onde

$$\mathcal{X}^t\mathcal{X} = \begin{bmatrix} \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2 & \cdots & \sum_{i=1}^{n}\left(x_{ik} - \bar{x}\right)\left(x_{i1} - \bar{x}_1\right) \\ \vdots & & \vdots \\ \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)\left(x_{ik} - \bar{x}_k\right) & \cdots & \sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2 \end{bmatrix}.$$

Seja $\mathcal{x}_i$ a i-ésima linha da matriz $\mathcal{X}$, logo podemos escrever o i-ésimo elemento da diagonal da matriz de projeção, H, como

$$h_{ii} = \begin{pmatrix} 1 & \mathcal{x}_i \end{pmatrix} \begin{bmatrix} n & 0 \\ 0 & \mathcal{X}^t\mathcal{X} \end{bmatrix} \begin{pmatrix} 1 \\ \mathcal{x}_i \end{pmatrix}$$

$$= \frac{1}{n} + \mathcal{x}_i\left(\mathcal{X}^t\mathcal{X}\right)^{-1}\mathcal{x}_i^t. \qquad (5.5)$$

Multiplicando e dividindo por n-1 no segundo termos da direita de (5.5) obtemos

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1}\,\mathcal{x}_i\left(\frac{1}{n-1}\,\mathcal{X}^t\mathcal{X}\right)^{-1}\mathcal{x}_i^t \qquad (5.6)$$

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1} MD_i^2 \qquad (5.7)$$

Tem sido amplamente comprovado que a média e variância amostrais tem ponto de ruptura zero, isto é, uma única observação discrepante desloca a média, atraindo-a e aumenta a variância. Consequentemente, os elipsóides (3.9) e (5.3) também serão deslocados e aumentados para aproximar-se dessa observação discrepante. Se o conjunto de dados contém múltiplas observações discrepantes, pode acontecer que estas sejam mascaradas, apresentando assim distâncias de Mahalanobis relativamente pequenas ou pelo menos não notoriamente grandes. Concluindo, tanto a distância de Mahalanobis quanto os elementos da diagonal da matriz H são úteis para identificação de observações discrepantes individuais, porém na presença de múltiplas observações discrepantes sofrem o efeito de mascaramento. O baixo ponto de ruptura da distância de Mahalanobis, devido ao emprego da média e covariância amostrais, torna necessária a busca de estimadores robustos dos parâmetros x_0 e Γ da distância definida em (5.1).

Estamos interessados em estimadores robustos que não provoquem o efeito de mascaramento mais que mantenham as propriedades de equivariância sob transformações afins, possuidas pela média e covariância amostrais. Para atender à primeira condição empregaremos estimadores com alto ponto de ruptura. Encontrar, contudo, estimadores que além disso sejam equivariantes sob transformações afins não é uma tarefa fácil. Sahel (1981) e Donoho (1982), independentemente, propuseram os primeiros estimadores multivariados com ponto de ruptura alto equivariantes sob transformações afins. Outro estimador que reune as duas condições é obtido através do elipsóide de volume mínimo que apresentaremos a seguir.

5.3 ELIPSÓIDE DE VOLUME MÍNIMO

Rousseeuw (1985), propôs o estimador elipsóide de volume mínimo (MVE) e provou que este é equivariante sob transformações afins e possui ponto de ruptura 50%.

DEFINIÇÃO 5.4.1

Seja $X = \left\{x_1, x_2, \ldots, x_n\right\}$, $x_i \in \mathbb{R}^k$, $n \geq k+1$. Definimos o estimador elipsóide de volume mínimo como o par (t,C), onde $t \in \mathbb{R}^k$, $C \in PDS(k)$ tal que

$$\text{Min } \det(C),$$

Sujeito a:

$$\# \left\{ i : \left(x_i - t\right) C^{-1} \left(x_i - t\right)^t \leq b \right\} \geq q \ . \qquad (5.8)$$

Det(.) denota o determinante, a constante b será discutida mais adiante, $t = t_n(X)$ e $C = C_n(X)$ determinam o centróide e a estrutura de covariâncias do elipsóide de volume mínimo cobrindo pelo menos q dados, onde $q = [(n+k+1)/2]$. O escalar b é fixado e não tem influência no cálculo de t, mas é muito importante na determinação da magnitude de C.

Se b é escolhido no dominio de uma distribuição de probabilide do tipo elíptica, obteremos estimadores consistentes de x_0 e Γ. Sob a suposição de que o conjunto de dados provém de uma distribuição normal com parâmetros $\mu \in \mathbb{R}^k$ e $\Sigma \in PDS(k)$, b será obtido a partir de

$$P_{\mu,\Sigma} \left\{ \left(x_i - \mu\right) \Sigma^{-1} \left(\bar{x}_i - \mu\right)^t \leq b \right\} = \frac{1}{2} \ .$$

Como a variável aleatória $\left(x_i - \mu\right) \Sigma^{-1} \left(\bar{x}_i - \mu\right)^t$ tem distribuição χ^2_k, segue que b corresponde a um percentil da distribuição chi-quadrado com k graus de liberdade, cobrindo no máximo 50% da área total da curva.

Se cada subconjunto de q observações de X contem pelo menos k+1 observações em posição geral, então existe no mínimo uma solução (t,C) para o problema de minimização da definição 5.3.1. Mesmo que algum subconjunto de q observações esteja contido em um hiperplano de dimensão menor, pode-se ainda definir $t \in \mathbb{R}^k$ como o centro do elipsóide de volume mínimo, dentro deste hiperplano, cobrindo pelo menos q observações.

5.3.1 Ponto de Ruptura do Estimador Elipsóide de Volume Mínimo

Rousseeuw (1985) mostrou que o ponto de ruptura do estimador MVE é ([n/2] - k + 1)/n, sempre que na restrição (5.8) da definição 5.3.1 tenhamos q = [n/2] + 1. Posteriormente, Davies (1987) e Lopunhaä e Rousseeuw (1991) propuseram uma forma de melhorar o ponto de ruptura deste estimador fazendo que o mesmo alcançasse o limite superior para estimadores de dispersão equivariantes apresentado no teorema 2.4.1. A modificação consiste em introduzir o número de regressores, k , na expressão de q. Assim o ponto de ruptura dos estimadores de covariância equivariantes será mo máximo [(n-k+1)/2]/n (Davies (1987)). Entretanto, os estimadores de posição podem ter um limite superior maior do que [(n-k+1)/2]/n. Tanto o estimador de posição quanto o de dispersão os quais contituem o estimador elipsóide de volume mínimo, alcançãm este ponto de ruptura como veremos no teorema seguinte.

TEOREMA 5.4.1

Seja $X = \left\{x_1, x_2, \ldots, x_n\right\}$, um conjunto de $n \geq k+1$ pontos, onde $x_i \in \mathbb{R}^k$, e X está em posição geral. sejam t_n e C_n estimadores elipsóide de volume mínimo de posição e dispersão.

i) Se k=1

$$\varepsilon^*(t_n, X) = \frac{\left[\frac{n+1}{2}\right]}{n} \tag{5.9}$$

e

$$\varepsilon^*(C_n, X) = \frac{\left[\frac{n}{2}\right]}{n} .$$

ii) Quando $k \geq 2$

$$\varepsilon^*(t_n, X) = \varepsilon^*(C_n, X) = \frac{\left[\frac{n-k+1}{2}\right]}{n} . \tag{5.10}$$

PROVA (Ver Lopunhaä e Rousseeuw (1991)).

5.4 DIAGNÓSTICOS DE OBSERVAÇÕES DISCREPANTES BASEADOS EM MÉTODOS ROBUSTOS

Um método para diagnósticar observações discrepantes, frequentemente usado na análise de regressão pelo método de mínimos quadrados, consiste em construir gráficos de resíduos do ajuste versus os elementos da diagonal da matriz de projeção, ou estes mesmos resíduos versus as distâncias de Mahalanobis ou, simplesmente, os resíduos versus os índices das observações. Estes gráficos podem também ser construídos relacionando as distâncias robustas e os resíduos do ajuste, calculados utilizando os estimadores de regressão robustos

descritos no capítulo anterior. Os gráficos assim construídos serão mais ilustrativos do que a análise dos resíduos e as distâncias separadamente, dado que estes combinam a informação das observações discrepantes da regressão com os pontos de alavanca, possibilitando que estruturas não necessariamente reveladas pelo método de mínimos quadrados sejam descobertas.

A seguir descreveremos os passos para a construição e interpretação destes gráficos de diagnósticos de regressão.

1) Estimar os parâmetros do modelo utilizando um estimador com alto ponto de ruptura. O estimador mais comumente usado é o LMS.

2) Obter os resíduos do ajuste

$$e_i = \left(y_i - \hat{y}_{LMS}\right).$$

3) Calcular uma estimativa de dispersão dos erros. Por exemplo, se usarmos o estimador de regressão LMS, o estimador robusto de dispersão será

$$s = \underset{1\leq i\leq n}{\text{Mediana}} \left(y_i - \hat{y}_{LMS}\right)^2. \qquad (5.15)$$

Rousseeuw e Leroy (1987) propuseram calcular estimativas robustas da dispersão efetuando os seguintes calculos:

i) Obter a estimativa de dispersão inicial

$$s^0 = (1.4826)\left(1 + \frac{5}{n-k-1}\right)^2 s.$$

A expressão $\left(1 + \frac{5}{n-k-1}\right)^2$ é um fator de correção para amostras finitas e o valor 1.4826 é igual ao quantil 75% de uma distribuição normal. Este último fator é empregado para a obtenção de estimativas consistentes.

ii) Obter pesos para cada uma das observações, isto é, calcular

$$w_i = \begin{cases} 1 & \text{se } \left(e_i/s^0\right) \leq 2.5 \\ 0 & \text{caso contrario.} \end{cases}$$

iii) O estimador de escala final será

$$s_{LMS} = \left(\frac{\sum_{i=1}^{n} w_i e_i}{\sum_{i=1}^{n} w_i - p} \right)^{1/2} .$$

O estimador s_{LMS} tem ponto de ruptura 50%, ou seja, s_{LMS} não tende a zero nem a infinito para menos de 50% de contaminação.

4) Padronizar os resíduos

$$e_i^* = \frac{e_i}{s_{LMS}} \quad ; \qquad i=1,2,\ldots,n.$$

5) Determinar um ponto de corte para decidir quando um resíduo padronizado pode ser considerado discrepante. Rousseeuw e Leroy (1987) sugerem o uso dos pontos -2.5 e 2.5.

6) Calcular as estimativas robustas dos parâmetros de posição e dispersão k - variados, (x_0, Γ).

7) Determinar as distâncias robustas, RD_i^2, empregando a expressão (5.1).

8) Fixar o ponto de corte para decidir quando uma distância será considerado grande. Um critério poderia ser considerar como ponto de corte o quantil $\chi^2_{k,.975}$.

Observação.- Os pontos de corte tanto no passo (5) quanto no passo (8) são arbitrários. O pesquisador é quem decide quando um resíduo padronizado ou uma distância podem ser considerados como sendo grandes.

A figura 5.4.1 ilustra os diferentes tipos de observações que podem ocorrer em um conjunto de dados

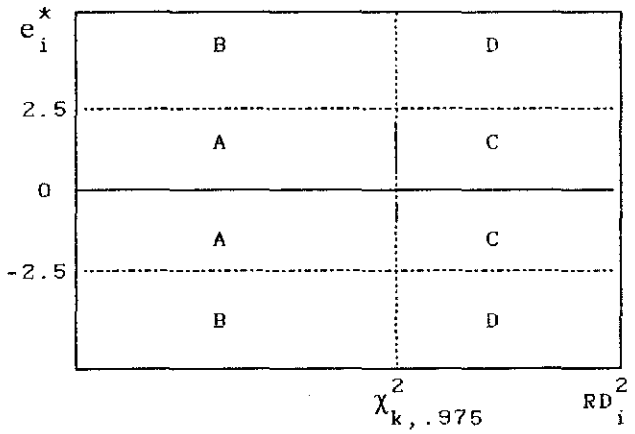

Figura 5.4.1 Possíveis tipos de observações em um conjunto de dados relacionados segundo o modelo de regressão linear Múltipla.

A) Observações regulares apresentam resíduos padronizados e distâncias robustas pequenas.

B) Observações discrepantes da regressão apresentam resíduos grandes e distâncias pequenas.

C) Pontos de alavanca "bons" possuem distâncias robustas grandes, mas os resíduos robustos não notáveis.

D) Pontos de alavanca "ruins" apresentam resíduos e distâncias robustas grandes.

5.5 SUMÁRIO

Um resumo deste capítulo é o seguinte:

- Os eleemtos da diagonal da matriz de projeção está funcionalmente relacionados com as distâncias de Mahalanobis.

- A distância de Mahalanobis, tradicionalmente usada como método de subo-rdenamento de conjuntos de dados multivariados, é afetado pelo efeito de mascaramento. Consequentemente, os elementos da diagonal da matriz de projeção o mesmo problema.

- Distâncias robustas podem ser calculados utilizando estimadores robustos dos parâmetros de posição e dispersão da expressão (5.1).

- O estimador elipsóide de volume mínimo (MVE), é um exemplo de estimador com ponto de ruptura máximo, que é usado conjuntamente com o estimador de regressão LMS para fazer diagnósticos de observações discrepantes.

No capítulo seguinte apresentaremos alguns algoritmos que são usados para calcular as distâncias e resíduos robustos.

CAPITULO 6

ALGORITMOS DE REAMOSTRAGEM BASEADOS EM CONJUNTOS ELEMENTARES

5.1 INTRODUÇÃO

A estimação de parâmetros no modelo de regressão linear múltipla ou no modelo de posição multivariado correspondem, na prática, à obtenção da solução de um problema de minimização de certa função, $W(\theta)$, que envolve os resíduos do ajuste. A solução deste problema pode ser obtida por métodos analíticos, quando a função objetivo é convexa, contínua, e diferenciável como no caso da estimação de mínimos quadrados. Se a função é não diferenciável ainda podemos resolver o problema empregando métodos iterativos.

Infelizmente, as funções objetivo associadas aos estimadores com alto ponto de ruptura são não convexas e têm múltiplos mínimos locais. Além disso, na maioria dos casos, elas são não diferenciáveis, portanto em geral, os métodos usuais de otimização falham.

Neste capítulo mencionaremos brevemente os diferentes algoritmos propostos para resolver o problema de estimação com alto ponto de ruptura.

Na seção 6.2, apresentaremos o método de reamostragem de conjuntos elementares, que tem servido de base para o desenvolvimento de uma grande quantidade de algoritmos orientados à estimação quando a função objetivo apresenta os problemas antes mencionados.

Nas seções 6.4.1 e 6.4.2, apresentaremos os programas computacionais PROGRESS e MINVOL para estimação LMS e MVE, respectivamente. Estes programas baseiam-se nos algoritmos a serem apresentados nas seções 6.3.1 e 6.3.2.

Outro algoritmo que foi desenvolvido especificamente para estimação LMS, e que utiliza a estimação de Chebyshev além dos conjuntos elementares, foi proposto por Stromberg (1991). Este algoritmo, que apresentaremos na seção 6.4.3, foi o primeiro a proporcionar uma solução exata para o problema de estimação LMS.

O algoritmo que apresentaremos na seção 6.4.4, foi desenvolvido por Hawkins (1993, (a)) introduzendo uma melhora no algoritmo exato de Stromberg no sentido que permite examinar um número razoavelmente maior de conjuntos elementares do que o algoritmo de Stromberg e também produz soluções exatas. Este algoritmo serviu de base para o desenvolvimento do programa computacional FSA (Feasible Set Algorithm) Para estimação LMS, MVE e LTS.

6.2 CONJUNTOS ELEMENTARES

Consideremos o conjunto de dados $Z_{n\times(k+1)} = \left(X_{n\times k} \quad y_{n\times 1}\right)$, cujos elementos se relacionam segundo o modelo (1.1). Particionando a matriz Z como

$$Z = \begin{pmatrix} X_j & y_j \\ X_{n-j} & y_{n-j} \end{pmatrix} \qquad (6.1)$$

onde $J = \left\{ J_1, \ldots, J_k \right\}$ é um subconjunto de índices com k elementos, obtido do conjunto {1,2,...., n}, e considerando

$$S = \left\{ \{1,2,\ldots,k\}, \ldots, \{J_1, \ldots, J_k\}, \ldots, \{n-k+1,\ldots, n\} \right\}, \qquad (6.2)$$

o conjunto de possíveis subconjuntos de índices que podemos obter com n e k fixados, temos que, cada subconjunto $\left(X_J \quad y_J \right)$, $J \subset S$, onde

$$Z_J = \begin{pmatrix} x_{J_1,1} & \ldots & x_{J_1,k} & y_{J_1} \\ \vdots & & & \\ x_{J_k,1} & \ldots & x_{J_k,k} & y_{J_k} \end{pmatrix}, \qquad (6.3)$$

é denominado *"Conjunto Elementar"*.

O método de reamostragem de conjunto elementares foi usado por Bradu e Hawkins (1982) para identificar dados discordantes em tabelas de dupla entrada (caso especial do modelo linear). Rousseeuw (1984) também propôs empregar este recurso para resolver o problema de estimação LMS. Hawkins, Bradu e Kass (1984) o empregaram na identificação de observações discrepantes.

Os algoritmos que apresentaremos nas seções seguintes utilizam conjuntos elementares, tanto para estimação dos coeficientes do modelo (1.1) quanto para estimar os parâmetros de posição e dispersão no modelo de posição multivariado.

6.2.1 Algumas Observações sobre Conjuntos Elementares

A seguir apresentaremos algumas observações interessantes sobre os conjuntos elementares.

i) A escolha dos subconjuntos de dados com o menor tamanho possível,

isto é, igual ao número de parâmetros, deve-se ao fato que a proporção de subconjuntos contendo pelo menos uma observação discordante cresce rapidamente em função do tamanho do subconjunto. Logo, mantendo os subconjuntos tão pequenos quanto possível maximiza-se o número de estimativas, $\tilde{\theta}_j$, que não serão influenciadas pelas observações discordantes, tornando-se utilizáveis para obter a estimativa final.

A idéia de reamostragem repetida de subconjuntos dos dados tem uma analogia conceitual com o bootstrap (Efron (1979)), sendo que no caso dos subconjuntos elementares as sub-amostras são de tamanho mínimo.

ii) As estimativas finais obtidas usando algoritmos baseados em conjuntos elementares mantêm as propriedades de alto ponto de ruptura, invariância e equivariância dos estimadores.

iii) As estimativas de mínimos quadrados (norma L_2) calculadas sobre o conjunto total de dados será igual à média das estimativas calculadas sobre todos os conjuntos elementares, ponderados pelo volume ao quadrado de cada subconjunto, X_J. Assim temos que:

- O vetor de coeficientes estimados é

$$\hat{\theta} = \frac{\sum_{J=1}^{S} \det^2 \left(X_J\right) \theta_J}{\sum_{J=1}^{S} \det^2 \left(X_J\right)} . \qquad (6.4)$$

- Os resíduos e_i da regressão pelo método de mínimos quadrados são dados por

$$e_i = \frac{\sum_{J=1}^{S} \det^2 \left(X_J\right) e_{iJ}}{\sum_{J=1}^{S} \det^2 \left(X_J\right)} ; \quad i=1,2,\ldots,n. \qquad (6.4)$$

• A soma de quadrados dos erros é obtida fazendo

$$SCE = \frac{\sum_{J=1}^{S} \det^2 \left(X_J\right) \sum_{i \in J} e_{iJ}^2}{\sum_{J=1}^{S} \det^2 \left(X_J\right)} \qquad (6.6)$$

6.2.2 Número de Conjuntos Elementares Examinados pelos Algoritmos de Reamostragem quando n e k são Grandes

Se o número de elementos do conjunto de sub-índices, S, é pequeno, todos os conjuntos elementares podem ser examinados exaustivamente. Entretanto, o número de conjuntos elementares cresce rapidamente com n e k, tornando a busca da estimativa do vetor de parâmetros no modelo de regressão linear, ou das médias e covariâncias no modelo de posição multivariado, muito trabalhosa e lenta. Em alguns casos sua realização é impossível pelo excessivo tempo e espaço computacional requerido.

Bradu e Hawkins (1993) estudaram o efeito do tamanho da amostra de subconjuntos elementares na probabilidade de obtermos um conjunto elementar que produz estimativas muito próximas daquelas que seriam obtidas se empregarmos todos os possíveis conjuntos elementares.

Definimos um conjunto elementar "limpo" como aquele que não contem observações discrepantes, e um conjunto elementar contaminado como aquele que contem no mínimo uma observação discrepante. Suponhamos que o conjunto de n dados tem m observações discrepantes, então teremos $\binom{n-m}{k}$ conjuntos elementares "limpos" e $\binom{n}{k}-\binom{n-m}{k}$ conjuntos elementares contaminados. Portanto, a probabilidade de obtermos um conjunto elementar "limpo" por amostragem aleatória simples sem reposição será

$$\pi = \frac{\binom{n-m}{k}}{\binom{n}{k}} . \qquad (6.7)$$

Consideremos agora a extração de uma amostra aleatória simples de N conjuntos elementares. Diremos que a estimação de parâmetros teve *éxito* se em alguma etapa encontrarmos pelo menos um conjunto elementar "limpo". A probabilidade de éxito é

$$p_E = 1 - (1 - \pi)^N \qquad (6.8)$$

Rousseeuw e Leroy (1987) obtiveram uma expressão ligeiramente diferente para p_E, partindo da proporção de observações contaminadas, e apresentaram uma tabela com o tamanho de amostra que se precisaria para garantir uma probabilidade de 95%, ou maior, de obtermos um conjunto elementar "limpo", para $1 \leq k \leq 10$ e frações de contaminação de 5% e 50%.

Os conjuntos elementares contaminados produzem estimativas não confiáveis, pois elas estarão influenciadas pelas observações discrepantes contidas no conjunto elementar, mesmo que os resíduos correspondentes a esse conjunto elementar sejam iguais a zero. Por outro lado, os conjuntos elementares "limpos" não necessariamente produzem estimativas confiáveis.

Tomando em consideração que o objetivo é detectar observações discrepantes, precisaremos que o modelo ajustado tenha a capacidade de discriminar claramente as observações discrepantes do resto dos dados. Este requerimento é chamado de *"Separação total"* e foi mencionado na seção 4.4.

Definiremos um subconjunto elementar como "bom" se ele exibe separação total, caso contrário será considerado "ruim". Na prática, os conjuntos elementares contaminados e os não contaminados "ruins" produzem estimativas não confiáveis. Portanto, não vale a pena fazer distinção entre eles pois em ambos os casos as estimativas não são

confiáveis.

Como estamos tratando de amostragem aleatória e soluções aproximadas, temos de investigar se em qualquer estagio do processo de otimização, os conjuntos elementares "bons" vão produzir valores da função objetivo, W, menores do que aqueles produzidos pelos subconjuntos elementares "ruins". A seguir discutiremos a possibilidade de responder afirmativamente a essa questão.

Denotemos por E_1 e E_2 os subconjuntos de conjuntos elementares "bons" e "ruins", respectivamente As proporções de conjuntos "bons" e "ruins" serão denotados por P_1 e P_2. Os valores da função objetivo, W, sob o subconjunto E_1 não são necessariamente menores do que os valores de W sob o conjunto E_2. O que podemos afirmar, escolhendo adequadamente a função objetivo, é que seus valores sob E_1 são *estocasticamente* menores que aqueles obtidos sob E_2, isto é, que

$$F_1(W) \geq F_2(W) \quad ; \quad \text{para qualquer } W, \tag{6.9}$$

onde F_1 e F_2 são as distribuições acumuladas de W sob E_1 e E_2, respectivamente.

Suponhamos também que

$$M_1 < M_2, \tag{6.10}$$

onde

$$M_i = \inf_{E_i} W \quad ; \qquad i=1,2. \tag{6.11}$$

Precisamos então encontrar N suficientemente grande de modo a garantir que (6.10) se verifique com uma probabilidade tão próxima de um quanto desejarmos, o que significará que encontramos um subconjunto bom. A condição $M_1 < M_2$ é também necessária pois se $M_1 > M_2$ os papéis de E_1 e E_2 estariam trocados.

Estudaremos agora as consequências de seleccionarmos uma amostra de N conjuntos elementares. Denotaremos por U_1 e U_2 os eventos

"mínimo de W" sob os conjuntos elementares "bons" e "ruins", respectivamente, na amostra. Diremos que ocorreu um sucesso quando $U_1 < U_2$. A probabilidade de suceso, $P\left(U_1 < U_2\right)$, depende de N, P_1, F_1 e F_2. A determinação da expressão matemática da probabidade anterior nos permitirá atingir o tamanho mínimo de N, maximizando $P\left(U_1 < U_2\right)$.

Bradu e Hawkins (1993) mostraram que

$$P\left(U_1 < U_2\right) = 1 - NP_2 \int \left(1-F(u)\right)^{N-1} dF_2(u) - \lambda \qquad (6.12)$$

ou

$$P\left(U_1 < U_2\right) = NP_1 \int \left(1-F(u)\right)^{N-1} dF_1(u) + \lambda \qquad (6.13)$$

onde $F(u) = N_1P_1 + N_2P_2$ e λ é um fator de correção de continuidade devido a F_1 e F_2 serem discretas, porém muito próximas ás distribuições contínuas.

Usando as expressões (6.12) e (6.13) podemos verificar que a probabilidade de termos $U_1 < U_2$ depende principalmente do extremo esquerdo das distribuições de interesse. Para isso, obteremos limites inferiores para a probabilidade de $U_1 < U_2$.

Seja $r > 0$ e $D_r = \left\{u : F_1(u) \geq F_2(u)\right\}$, e seja b tal que $(-\infty, b) \subset D_r$. Então para $u \in D_r$ temos

$$F(u) = P_1F_1(u) + P_2F_2(u) \leq \left(P_1 + P_2 / r\right) F_1(u).$$

Logo,

$$P\left(U_1 < U_2\right) \geq \int_{\substack{D \\ r}} NP_1 \left\{ 1 - \left\{P_1 + P_2 / r \right\} F_1(u)\right\}^{N-1} dF_1(u)$$

$$\geq \left\{P_1 / \left(P_1 + P_2/r\right)\right\} \int_{(-\infty, b)} -d\left\{1 - \left\{P_1 + P_2 / r\right\} F_1(u)\right\}^N$$

$$= \left\{P_1 / \left(P_1 + P_2/r\right)\right\} \left\{1 - \left\{1 - \left\{P_1/\left(P_1 + P_2/r\right) \right\} F_1(b)\right\}\right\}. \qquad (6.14)$$

A existência de tais valores r e b implica que quando N cresce, $P\left(U_1 < U_2\right)$ tem um limite inferior tão próximo de $P_1/\left(P_1 + P_2/r\right)$ quanto desejarmos. No caso particular em que b é menor do que o Min U_2 e $F_1(b) > 0$, então, o limite inferior (6.14) é atingido para qualquer $r > 0$. Portanto, fazendo $r \longrightarrow \infty$

$$P\left(U_1 < U_2\right) \geq 1 - \left(1 - P_1 F_1(b)\right)^N$$

$$\lim_{N \to \infty} P\left(U_1 < U_2\right) \geq 1 - \lim_{N \to \infty} \left(1 - P_1 F_1(b)\right)^N = 1. \qquad (6.15)$$

Concluimos que, tomando N suficientemente grande, a probabilidade de $U_1 < U_2$ pode estar tão próximo de um quanto desejarmos.

6.3 APLICAÇÕES DO MÉTODO DE REAMOSTRAGEM DE CONJUNTOS ELEMENTARES

Nesta seção apresentaremos alguns algoritmos baseados em conjuntos elementares para estimação de parâmetros no modelo de regressão linear múltipla e o estimador elipsóide de volume mínimo para os parâmetros de posição e dispersão no modelo de posição multivariado.

6.3.1 Estimação de Parâmetros no Modelo de Regressão Linear Múltipla

Considerando conjuntos elementares de tamanho igual ao número de coeficientes do modelo, ajustaremos uma única equação de regressão linear múltipla para cada conjunto elementar. Os passos a seguir são:

1. Obter todos os possíveis conjuntos elementares, isto é, todos os possíveis subconjuntos Z_J, $J \subset S$.

2. para cada $J \subset S$ repetir os passos seguintes:

   a) Calcular

$$\theta_J = \left(X_J^t X_J\right)^{-1} X_J^t y_J$$

$$= X_J^{-1} y_J, \qquad (6.16)$$

   sempre que $\det\left(X_J\right) \neq 0$ (não singular). Esta é uma condição mais fraca do que a posição geral. Se assumirmos que os n vetores de dados estão em posição geral, o vetor de coeficientes, θ_J, do hiperplano que passa por exatamente k pontos sempre existe.

   b) Obter os resíduos do ajuste

$$e_i\left(\theta_J\right) = y_i - \sum_{I=1}^{k} x_{iI}\theta_J , \qquad i=1,2,\ldots,n; \quad J \subset S. \qquad (6.17)$$

   Se $i \in J$, $e_i\left(\theta_J\right) = 0$ e se $i \notin J$ então $e_i\left(\theta_J\right)$ é um resíduo predito. Logo, para cada subconjunto $J \subset S$ calculamos o vetor $e\left(\theta_J\right)$.

3. A estimativa final do vetor de parâmetros que denotaremos com $\tilde{\theta}$, é o vetor θ_J que minimiza $W(\theta)$, ou seja, $\tilde{\theta} = \theta_E$ se

$$W\left(\theta_E\right) = \underset{J \subset S}{\text{Min}}\ W\left(\theta_J\right). \qquad (6.18)$$

W(.) pode ser a função objetivo de qualquer estimador. Em particular podem ser os estimadores com alto ponto de ruptura definidos na seção 4.4, os quais têm múltiplos mínimos locais.

6.3.2 Estimação de Parâmetros no Modelo de Posição Multivariado

Rousseeuw e van Zomeren (1991) propuseram o seguinte algoritmo baseado no método de reamostragem de conjuntos elementares, para calcular estimativas dos parâmetros de posição e dispersão do modelo de posição multivariado.

1. Obter todas as possíveis amostras de tamanho k+1. O conjunto S definido em (6.2) muda ligeiramente pois teremos $\binom{n}{k+1}$ possíveis subconjuntos com k+1 elementos.

2. Para cada $J \subset S$ repetir os seguintes passos:

   a) Calcular

$$\bar{x}_J = \frac{1}{k+1} \sum_{i \in J} x_i \qquad (6.19)$$

$$C_j = \frac{1}{k} \sum_{i \in J} \left(x_i - \bar{x}_j\right)^t \left(x_i - \bar{x}_j\right). \qquad (6.20)$$

   b) Calcular as distâncias de Mahalanobis empregando $\bar{x}_j$ e C_j

$$D_{ji}^2 = \left(x_i - \bar{x}_j\right)^t C^{-1} \left(x_i - \bar{x}_j\right), \qquad i=1,2,\ldots,n. \qquad (6.21)$$

   c) Obter m_j^2, a q-ésima estatística de ordem dos D_{Ji}^2 obtidos em (6.21) e calcular

$$V_J = m_J^{2k} \det\left(C_J\right). \qquad (6.22)$$

3. Guardar o par $\left(\bar{x}_L, C_L\right)$ correspondente ao subconjunto $L \in S$ para o qual

$$V_L = \underset{J \in S}{\text{Min}}\ V_J \qquad (6.23)$$

4. Calcular a estimativa elipsóide de volume mínimo (MVE) para posição e escala

$$t = \bar{x}_L$$

$$C = \left(\chi^2_{k,.50}\right)^{-1} m_L\, C_L\, c^2_{n,k}\ ; \qquad (6.24)$$

onde $c^2_{n,k}$ é um fator de correção para amostras pequenas que explicaremos a seguir.

Um critério para decidirmos se uma observação pode ser considerada discrepante é comparar as distâncias robustas baseadas na estimação MVE com um quantil da distribuição χ^2_k.

Rousseeuw e van Zomeren (1991) perceberam usando dados reais, que a porcentagem de cobertura do elipsóide de volume mínimo é menor do que o esperado segundo a distribuição χ^2_k. Os autores realizaram estudos de simulação considerando hiperplanos de regressão com valores de k iguais a 2, 3, 4 e tamanhos de amostra, n, iguais a 20, 50, 100 e 50 replicações de cada modelo. Construiram gráficos Q-Q relacionando as médias, medianas e amplitudes interquartílica dos quantis simulados e os correspondentes à distribuição χ^2_k. Este estudo levou às seguintes conclusões:

. Sob a suposição de normalidade, a distribuição empírica das distâncias robustas tem caudas mais pesadas do que a distribuição χ^2_k, o que explica o fato da probabilidade de cobertura do elipsóide de volume

mínimo ser menor do que o esperado para a distribuição chi-quadrado com k graus de liberdade.

. Quando o tamanho de amostra cresce, a proporção de cobertura melhora. Logo, para conjuntos grandes de dados a aproximação pode produzir melhores resultados. Entretanto, verificou-se no estudo que a convergência é lenta.

Baseados nestes resultados, realizaram um estudo de simulação considerando diversas combinações de n e k para construir gráficos Q-Q, determinando os valores de corte reais correspondentes ao quantil 97.5% da distribuição χ^2_k. O gráfico desses valores contra o fator 1/(n-k) mostrou uma relação linear. Essa relação foi utilizada para obter o coeficiente de correlação para amostras pequenas,

$$c^2_{n,k} = \left(1 + \frac{15}{(n-k)}\right)^2, \qquad (6.25)$$

utilizado no calculo da matriz de covariâncias estimadas em (6.24).

6.4 ALGORITMOS BASEADOS EM CONJUNTOS ELEMENTARES

Nesta seção apresentaremos alguns dos algoritmos propostos para resolver o problema de estimação com alto ponto de ruptura.

6.4.1 Programa Computacional PROGRESS

Desenvolvido por Leroy e Rousseeuw (1984), este programa encontra-se amplamente descrito no livro de Rousseeuw e Leroy (1987). Produz diagnósticos de observações discrepantes para o modelo de regressão (1.1) tanto pelo tradicional método de mínimos quadrados quanto pelo método robusto LMS.

O algoritmo para ajuste LMS baseia-se no método de

reamostragem de conjuntos elementares descrito na seção 6.3.1. A solução LMS obtida é aproximada, exceto para os modelos de posição e regressão linear simples em que as soluções são exatas.

Para o modelo de posição emprega-se o algoritmo proposto por Rousseeuw ((1984), teorema 2), que consiste em :

a) Obter as estatísticas de ordem dos dados

$$y_{1:n} \leq y_{2:n} \leq \ldots \leq y_{n:n}$$

b) Calcular as diferênças

$$y_{q:n} - y_{1:n} \ , \ y_{q+1:n} - y_{2:n} \ , \ \ldots \ , \ y_{n:n} - y_{n-q-1:n} \ ,$$

onde $q = [n/2] + 1$.

c) Escolher a menor das diferenças calculadas em (b). O ponto médio do intervalo correspondente a essa diferença é a estimativa LMS do parâmetro de posição, θ_0.

Se considerarmos um modelo linear com intercepto, o PROGRESS utiliza o algoritmo acima descrito para ajustar o intercepto. O ajuste será feito como segue:

i) Obter as estimativas LMS finais empregando o algoritmo da seção 6.3.1, isto é, obter $\tilde{\theta} = (\tilde{\theta}_0, \tilde{\theta}_1, \ldots, \tilde{\theta}_k)$.

ii) Calcular os números

$$y_i - \sum_{i=1}^{k} x_{ij} \tilde{\theta}_j \qquad ; \qquad i=1,2,\ldots,n.$$

iii) Repetir os passos (a) a (c) do algoritmo acima descrito e substituir $\tilde{\theta}_o$ pela estimativa LMS obtida a partir dos valores calculados no passo (ii).

Este procedimento faz o valor da função (4.13) decrescer, pois o algoritmo unidimensional produz o intercepto estimado ótimo, condicionado aos valores $\tilde{\theta}_1$, $\tilde{\theta}_2$, ..., $\tilde{\theta}_k$.

Para calcular as estimativas LMS exatas no modleo linear simples, utilizam-se os algoritmos propostos por Steele e Steiger (1986). Eles foram os primeiros que propuseram algoritmos exatos para estimação LMS em modelos do tipo (1.1). O intercepto deste modelo também é ajustado empregando o algoritmo acima descrito para o modelo de posição.

O PROGRESS oferece a possibilidade de examinarmos todos os conjuntos elementares ou utilizarmos uma amostra de N deles. Rousseeuw e Leroy ((1987), pag.198) apresentam uma tabela com os tamanhos de amostra de conjuntos elementares necessários para garantir uma probabilidade de pelo menos 95% de obtermos um subconjunto elementar "limpo" para diferentes valores de k e porcentagens de contaminação, ε%.

6.4.2 Programa Computacional MINVOL

O algoritmo para calcular o estimador elipsóide de volume mínimo, MVE, foi primeiramente implementado por Rousseeuw e van Zomeren (1987), em um programa computacional denominado PROCOVIEV (RObust COvariance and Identification of Extreme Values). Posteriormente, Dallall e Rousseeuw (1992) desenvolveram o programa MINVOL, que calcula estimativas para os parâmetros de posição e dispersão de um modelo de posição multivariado. O par de estimativas (t,C) é obtido usando a definição de estimador elipsóide de volume mínimo apresentado no capítulo 5. O programa computacional baseia-se no algoritmo de reamostragem de conjunto elementares descrito na subseção 6.3.2.

Dallall e Rousseeuw (1992), desenvolveram um programa denominado LMSMVE, que reune os algoritmos para estimação LMS do PROGRESS e MVE do MINVOL, isto é, o programa produz distâncias robustas para

identificação de ponto de alavanca e resíduos do ajuste LMS para identificação de observações discrepantes da regressão. Além disso, produz o gráfico bidimensional descrito na seção 5.4, que permite identificar claramente os 4 tipos de observações que podem ocorrer em uma análise de regressão.

Os programas PROGRESS e MINVOL, já foram incorporados nos pacotes computacionais S-PLUS e BMDP.

Hawkins e Simonoff (1992) propuseram outro algoritmo basado em conjuntos elementares, denominado MVELMS, que oferece estimativas LMS e MVE. O algoritmo baseia-se em uma proposta de Cook e Hawkins (1990) que consiste em utilizar o método Simplex de programação linear para realizar os cálculos das estimativas LMS e MVE simultaneamente e de maneira rápida e eficiente.

Diferentemente do programa PROGRESS, o MVELMS ajusta o intercepto da equação de regressão para cada conjunto elementar.

6.4.3 Algoritmo Exato de Stromberg

Stromberg (1991) foi o primeiro a propôr um algoritmos exato para estimação LMS com mais de uma variável regressora. Sua proposta baseia-se no fato de que ao modificar ligeiramente a função objetivo do estimador LMS definido em (4.14), expressando-a como o q-ésimo quantil (expressão 4.16), podemos formular o problema de estimação LMS como um problema de estimação minimax ou de Chebyshev.

A seguir apresentaremos alguns resultados referentes à estimação Chebyshev que foram adaptados por Stromberg para desenvolver seu algoritmo exato.

Um resultado importante que permite distinguir as soluções minimax dos outros pontos do espaço $\mathbb{R}^k$ é dado pelo teorema de caracterização (Cheney (1966), pag. 35), que diz o seguinte:

"*Dado um ponto* $\theta \in \mathbb{R}^k$ *e uma função sinal*

$$s_i = \text{Sinal}\left\{ e_i(\theta) \right\} ; \quad i=1,2,\ldots,n,$$

$$M = \left\{ i : e_{i:n}(\theta) = W(\theta) \right\},$$

onde

$$W(\theta) = \underset{1 \le i \le q}{\text{Max}} \; e_{i:n}; \qquad q = [n/2] + [(k+1)/2].$$

Então, o ponto θ *minimiza* $W(\theta)$*, se e somente se, a origem do espaço de parâmetros encontra-se no contorno convexo do conjunto*

$$\left\{ s_i x_i : i \in M, \; x_i \in \mathbb{R}^k \right\}."$$

Outro teorema (Cheney, pag. 36) garante que as estimativas minimax, para um conjunto de $n \geq k$ dados, são as estimativas minimax de um apropriado sub-sistema de k+1 equações. Este teorema dá lugar a duas questões a serem resolvidas para obtermos as estimativas minimax. A primeira é como obter o "subconjunto apropriado", e a segunda é como calcular a estimação minimax nesse subconjunto. O primeiro problema foi resolvido por Stromberg investigando todos os possíveis subconjuntos de tamanho k+1 (conjuntos elementares), isto é, o conjunto de índices dos conjuntos elementares tem agora $\binom{n}{k+1}$ elementos. A determinação da solução minimax para cada subconjunto foi feita utilizando o método de La Vallée Poussin, que descreveremos a seguir.

Suponha que de algum modo podemos obter um vetor $\vec{\theta}_c \in \mathbb{R}^k$, uma função sinal s_i e um número δ (erro), tal que as seguintes condições sejam cumpridas:

$$\text{a) } e_{Ji}\left(\tilde{\theta}_c \right) = s_i \delta \quad ; \quad i=1,2,\ldots,k+1 \qquad J \subset S. \tag{6.26}$$

b) $0 \in \mathcal{H}\left\{s_1 x_1,\ s_2 x_2,\ \ldots,\ s_{k+1} x_{k+1}\right\}$.

Pelo teorema de caracterização e pela condição (b) temos que $\tilde{\theta}_c$ é uma solução de (a).

Para resolver o sistema de k+1 equações com k+1 incógnitas precisamos conhecer os sinais s_i. Isto é resolvido usando um teorema provado em Cheney (pag. 40):

"Dado um hiperplano $g \in \mathbb{R}^k$, os pontos de g que minimizam duas normas monótonas diferentes tem componentes com sinais iguais (ou podem ser assim escolhidas em caso de não unicidade)".

Dado que as normas L_p são monótonas, podemos obter os sinais dos resíduos do ajuste de mínimos quadrados calculados sobre os subconjuntos de k+1 dados e utilizá-los para resolver o sistema (a).

Finalmente o resultado que enunciaremos a seguir (Cheney, pag. 41) permite obter o valor de δ e a solução minimax exata a partir dos conjuntos elementares de tamanho k+1.

"Para cada conjunto elementar J com k+1 observações, calculamos as estimativas de mínimos quadrados θ_J e o vetor de sinais, s_J com elementos

$$s_{Ji} = \text{Sinal}\left\{e_i\left(\theta_J\right)\right\}\ ; \quad i=1,2,\ldots,k+1, \qquad J \subset S.$$

O resíduo máximo do ajuste de Chebyshev para as k+1 observações do conjunto elementar J é dado por

$$\delta = \frac{\sum_{i=1}^{k+1} e_i^2\left(\theta_J\right)}{\sum_{i=1}^{k+1}\left|e_i\left(\theta_J\right)\right|}\ . \qquad (6.27)$$

Logo, as k+1 observações terão resíduos iguais a δ *e o vetor de coeficientes correspondente a estimativa Chebyshev será*

$$\theta_J = \left(X_J^t X_J\right)^{-1} X_J \left(y_J - \delta s_J\right). \qquad (6.28)$$

A estimativa de Chebyshev para o conjunto total de dados será

$$\tilde{\theta}_{exato} = \underset{J \subset S}{\text{Min}} \; e^2_{q:n}\left(\theta_J\right)''.$$

Este algoritmo é computacionalmente tratável para tamanho de amostra e número de parâmetros pequenos, porém quando n e k são grandes, a quantidade de conjuntos elementares a serem investigados torna-o computacionalmente ineficiente.

6.4.4 Algoritmo de Conjuntos Factíveis (FSA)

Hawkins (1993, (a)) propôs uma melhora no algoritmo de Stromberg no sentido de buscar a solução exata somente entre aqueles subconjuntos de k+1 observações que cumprem a condição necessária para ser um ótimo. Este algoritmo não examina todos os $\binom{n}{k+1}$ possíveis conjuntos elementares, mas garante que a solução exata será atingida quase certamente (com probabilidade um), quando o número de subconjuntos examinados cresce.

A proposta de Hawkins consiste em calcular as estimativas de Chebyshev dos conjuntos elementares de tamanho k+1 aprimorando a composição do conjunto elementar base. Isso é feito substituindo uma observação por uma outra que produz uma melhor estimativa de Chebyshev. O mecanismo usado baseia-se no algoritmo de programação linear de Barrodale e Phillips (1975), que é uma modificação do método Simplex aplicado à formulação do dual do problema de Chebyshev.

Formulação do Problema de Programação Linear

O problema consiste em determinar o vetor θ tal que

$$\text{Min } e^*$$

Sujeito a :

$$\max_{1\le i\le n} \left| y_i - x_i\theta \right| \le e^*. \qquad (6.29)$$

A restrição pode ser expressa como

$$\begin{aligned} \left(y_i - x_i\theta \right) &\le e^* \\ -\left(y_i - x_i\theta \right) &\le e^* \\ \theta \in \mathbb{R}^k,\ e^* &\ge 0. \end{aligned} \qquad ; \quad i=1,2,\ldots,n$$

Esta formualação tem 2n restrições; entretanto se resolvermos o problema dual o número de restrições se reduz a k+1 como veremos a seguir.

O problema dual de (6.29) é

$$\text{Maximizar} \quad -yc^- + yc^+,$$

sujeito a :

$$-x^t c^- + x^t c^+ = 0$$

$$\sum_{i=1}^{k} \left(c^- + c^+ \right) \le 1 \qquad (6.30)$$

$$c^-,\ c^+ \ge 0, \qquad i=1,2,\ldots,n.$$

A determinação dos custos c^-, c^+ é equivalente a estabelecer os pontos que apresentam resíduos máximos. Assim, a i-ésima observação

tem desvio máximo se e somente se $c^- > 0$ ou $c^+ > 0$. Por outro lado, se $c_i^+ \left(c_i^-\right) > 0$, o valor y_i cai acima ou abaixo do hiperplano ótimo.

Os custos marginais utilizados na tabela simplex para o problema dual correspondem às distâncias verticais ao hiperplano factível definido pela base dual atual. A variável dual selecionada para entrar na base é aquela que apresenta o custo marginal mais negativo. Isso é equivalente a escolher o ponto que possui o maior desvio com respeito ao hiperplano factível atual.

A seguir decreveremos uma versão simplificada do algoritmo de soluções factíveis de Hawkins.

1. Começar com um hiperplano factível inicial (base inicial). Este hiperplano será obtido escolhendo um conjunto elementar com k+1 observações e realizando o ajuste de Chebyshev descrito na seção 6.4.3. Logo, o vetor de coeficientes estimados será dado por (6.27). As k+1 observações empregadas para obter o ajuste de Chebyshev têm resíduos iguais a δ, o qual é calculado usando (6.27). Os resíduos das n-(k+1) observações restantes são obtidas a partir do vetor de estimativas minimax.

2. O algoritmo simplex produz custos marginais para cada observação. Os custos para as observações que cumprem as restrições de cobertura são positivos, enquanto que os custos marginais negativos, correspondem às observações com resíduos maiores do que àqueles produzidos pelo hiperplano base atual.

3. Diferentemente do algoritmo de Barrodale e Phillips (1975), neste algoritmo entrará na base a observação cujo resíduo apresente o q-ésimo menor custo marginal, onde $q = [n/2] + [(k+1)/2]$. Suponhamos que m é o número total de observações com resíduos menores ou iguais a δ. Nesta etapa podem apresentar-se três casos:

   i) $q < m$, isto é, a solução atual pode não ser a solução LMS, exceto em casos degenerados, e não precisa ser investigada, mas será incluida na lista de soluções factíveis por que:

. Em casos degenerados o ajuste de Chebyshev pode produzir resíduos iguais a δ para mais de k+1 observações. Logo, a solução LMs exata poderá ter $m > q$.

. Se o valor de δ para este ajuste com $q < m$ é pequeno, pode ser útil como um limite superior a ser atingido pelas soluções factíveis.

ii) Se $q = m$, a solução atual satisfaz a condição necessária para a solução LMS ótima e é, portanto, uma solução factível. Seu valor δ é guardado e quando termina o processo, a solução factível com o menor valor δ será a solução LMS exata.

iii) Se $m < h$, a solução obtida não é ótima, isto é, o ajuste de Chebyshev para esta base não cobre a quantidade mínima requerida de observações.

4. Escolher alguma coluna com custo marginal negativo e introduzi-la na base, retirando a observação identificada pelas regras de pivotamento simplex. Neste algoritmo não escolhemos a observação com custo mais negativo como se faz em Barrodale e Phillips, usa-se a observação com o q-ésimo menor custo marginal.

5. Repetir os passos 2 a 4 até obter uma solução factível. Se o valor da função objetivo para a solução factível encontrada e maior do que qualquer solução previamente encontrada, esta solução converte-se no novo candidato para ser a solução LMS, e será guardada.

6. Repetir os passos 1 a 5 para um número razoavelmente grande de subconjuntos.

Todas as soluções investigadas nos passos 2 a 4 satisfazem a condição para ser uma solução LMS exata, portanto são soluções factíveis. Para ter utilidade prática o algoritmo deverá convergir ao ótimo global para uma grande quantidade de hiperplanos iniciais. Este conjunto de estimativas iniciais as quais convergem às soluções factíveis denomina-se *domínio de atração*. Logo um grande domínio de

factíveis denomina-se *domínio de atração*. Logo um grande domínio de atração significa que uma proporção alta de subconjuntos seleccionados aleatoriamente levam à solução factível.

6.5 SUMÁRIO

- Exceto o algoritmo FSA o qual foi desenvolvido para a estimação LTS, os algoritmos mencionados neste capítulo referem-se, principalmente, à estimação LMS e MVE. Porém muitas das conclusões podem ser extendidas à maioria dos estimadores com alto ponto de ruptura.

- Quando temos modelos com mais de um preditor, o algoritmo de reamostragem descrito na seção 6.3.1 produz soluções aproximadas.

- Na estimação de parâmetros com alto ponto de ruptura, a minimização da estatística $W(\theta)$ teria de ser feita sobre todo o espaço de parâmetros $\theta \in \mathbb{R}^k$. Dado que isso não é possível, reduzimos a busca a um conjunto finito de vetores θ, que são determinados pelos conjuntos elementares. Por outro lado, quando o número $\binom{n}{k}$ de possíveis conjuntos elementares é muito grande, a busca se reduz a uma amostra de tamanho N de conjuntos elementares.

- Hawkins demostrou que as estimativas LMS, LTS e de mínimos quadrados, quando obtidas empregando uma amostra de conjuntos elementares, convergem a aquelas calculadas sobre todos os conjuntos elementares. Por outro lado, as estimativas obtidas sobre todos os conjuntos elementares convergem às estimativas que se obteriam considerando o conjunto total de dados. Consequentemente, as estimativas produzidas pelos algoritmos baseados em conjuntos elementares são boas aproximações daquelas obtidas sobre todo o espaço de parâmetros. Hawkins sugere que estes resultados poderiam ser extendidos para

outros estimadores. Portanto para fins de análise exploratória, o uso de algoritmos baseados em conjuntos elementares podem ser considerados adequados.

CAPITULO 7

APLICAÇÄO

Um tipo de pesquisa em geologia consiste em realizar estudos de prospecção para saber se em determinada área geográfica existem indícios de existência de petróleo. A existência ou não de petróleo, na região em estudo, está ligada à maior ou menor permeabilidade da rocha, isto é, a sua capacidade de deixar que os líquidos a transpassem. Esta característica somente pode ser medida no laboratório através da análise de amostras de rochas (testemunhos). Entretanto, a medição da permeabilidade no laboratório é demorada e principalmente dispendiosa. Simultaneamente à obtenção dos testemunhos realizam-se medições de um conjunto de características das rochas tais como: teor de argila, densidade atômica, porosidade de densidade, porosidade neutrônica, potencial espontâneo, profundidade. raios gamma, resistividade esférica, resistividade profunda, resistividade rasa, porosidade de densidade, diferença de porosidade de densidade e porosidade neutrônica, raios gamma, argilosidade, resistividade micrônica, saturação de agua, etc.. Estas medições realizam-se a cada 20 centímetros de profundidade da rocha.

O interesse do pesquisador é formular um modelo para estimar a permeabilidade da rocha em função das medições das variáveis antes

mencionadas.

Os dados a serem analizados foram gentilmente cedidos pelo Eng. Armando Paulo Barros, aluno de Pos-graduação em Geologia, UNICAMP.

Logo após um processo de seleção das variáveis regressoras, isto é, das características que podem explicar a permeabilidade, encontramos que as variáveis

RMSFL : Resistividade esférica
VSH : Argilosidade
PHID : Porosidade de densidade
DPHI : Diferença porosidade de densidade e porosidade neutrônica

são relevantes para explicar LNKHL, o logaritmo da permeabilidade de laboratório (KHL = Permeabilidade de laboratório). A transformação logaritmica fez-se necessária para estabilizar a variância.

A seguir analizaremos o conjunto de dados da tabela 7.1, empregando os métodos de diagnósticos tradicionais e os robustos apresentados nos capítulos 3 e 5, respectivamente. Para construir os diagnósticos robustos utilizaremos os estimadores com alto ponto de ruptura LMS e MVE. Para cada uma das estatísticas de diagnósticos a serem utilizadas, consideraremos os pontos de corte sugeridos ao apresentar as estatísticas de diagnósticos correspondentes. Notemos que os pontos de corte generalmente usados para a análise dos resíduos são de ±2.0, mas acontece que os métodos robustos têm a tendência a declarar mais observações discrepantes,das que realmente podem existir no conjunto de dados. Por esta razão e para facilitar a comparabilidade entre os métodos, consideraremos uma observação como discrepante se seu resíduo padronizado não pertence ao intervalo (-2.5, 2.5). Por outro lado, consideraremos uma observação como sendo discrepante na direção das variáveis regressoras se sua distância de mahalanobis ou sua distância robusta excedem o quantil 97.5% de uma distribuição chi-quadrada com 4 graus de liberdade (número de variáveis).

Tabela 7.1 Medições da permeabilidade de laboratório (KHL) e das variáveis RMSFL, VSH, PHID, e DPHI de uma amostra de rocha em um ponto no nordeste do Brasil.

OBS.	RMSFL	VSH	PHID	DPHI	KHL	LNKHL
1	11.9	0.039	17.5	2.6	65.3	4.17899
2	15.0	0.105	12.3	5.3	45.2	3.81110
3	14.6	0.110	10.5	5.6	2.3	0.83291
4	11.2	0.119	10.4	9.7	260.0	5.56068
5	11.3	0.254	10.1	5.7	5.0	1.60944
6	6.3	0.100	19.5	-1.3	9.9	2.29253
7	7.8	0.199	15.6	3.5	30.8	3.42751
8	7.1	0.369	13.5	8.2	33.9	3.52342
9	7.1	0.331	13.2	8.2	81.7	4.40305
10	7.9	0.246	17.6	1.7	6.4	1.85630
11	8.6	0.193	18.4	-0.4	14.0	2.63906
12	10.0	0.210	16.5	2.8	15.8	2.76001
13	10.2	0.140	16.8	1.5	1.5	0.40547
14	13.6	0.051	18.6	-0.9	6.0	1.79176
15	25.1	0.124	20.9	-0.8	154.3	5.03890
16	26.6	0.131	20.4	0.2	35.2	3.56105
17	42.4	0.163	20.7	0.4	102.3	4.62791
18	27.2	0.146	21.5	1.4	247.6	5.51181
19	25.7	0.111	21.4	-0.3	336.5	5.81860
20	21.5	0.114	22.0	-0.1	105.2	4.65586
21	52.9	0.080	22.1	-0.5	425.2	6.05256
22	21.6	0.071	20.4	-0.7	63.3	4.14789
23	41.1	0.102	23.3	-1.9	549.8	6.30955
24	23.6	0.099	23.4	-2.0	262.6	5.57063
25	32.5	0.018	23.5	-4.3	455.5	6.12140
26	19.5	0.209	18.7	3.1	60.3	4.09933
27	18.4	0.212	17.4	5.5	24.8	3.21084
28	13.2	0.654	13.6	8.1	16.5	2.80336
29	7.7	0.747	15.0	10.0	0.2	-1.60944
30	9.1	0.193	20.2	-0.9	18.8	2.93386
31	14.9	0.589	16.5	6.9	45.5	3.81771
32	20.7	0.122	21.5	0.7	201.1	5.30380
33	22.0	0.155	22.7	-0.5	22.8	3.12676
34	19.8	0.167	20.1	1.3	96.4	4.56851
35	20.2	0.190	18.8	1.7	25.0	3.21888

Inicialmente ajustamos um modelo com intercepto mas o coeficiente θ_0 resultou não significante. Assim, o modelo finalmente considerado foi

$$LNKHL = \theta_1 RMSFL + \theta_2 VSH + \theta_3 PHID + \theta_4 DPHI + \varepsilon$$

A tabela 7.2, a seguir, apresenta os coeficientes de regressão do modelo ajustado pelo método dos mínimos quadrados. Notar que todos os coeficientes estimados são significativamente diferentes de zero.

Tabela 7.2 Coeficientes de regressão estimados, desvio padrão, estatistica T, e probabilidade de rejeição, associados ao ajuste de mínimos quadrados.

Variável	Coeficiente estimado	Desvio padrão	Estatística T	Prob > T
RMSFL	0.071362	0.0249612	2.859	0.0075
VSH	-5.502811	1.9267419	-2.856	0.0076
PHID	0.160751	0.0344509	4.666	0.0001
DPHI	0.232072	0.0863397	2.688	0.0115

$R^2 = 0.9161$

$s^2 = 1.55297$

O ajuste robusto dos dados usando o pacote computacional PROGRESS forneceu as estimativas LMS dos coeficientes de regressão, os quais são mostrados na tabela 7.3 a seguir

Tabela 7.3 Coeficientes de regressão estimados, associados ao ajuste LMS.

Variável	Coeficiente estimado
RMSFL	0.10882
VSH	-2.91651
PHID	0.13497
DPHI	0.24449

R^2 = 0.96109*

s_{LMS} = 1.11655

* Coeficiente de determinação baseado em medianas (Rousseeuw e Leroy (1987), Pag. 45).

As estimativas LMS (tabela 7.3) apresentam ligeiras diferenças com respeito às estimativas de mínimos quadrados (Tabela 7.2) exceto para a variável argilosidade (VSH).

A análise dos resíduos do ajuste de mínimos quadrados revela a observação 29 como sendo discrepante, e as observações 3 e 13 como suspeitas, porém, sem exceder os pontos de corte fixados (Tabela 7.4). Entretanto, o ajuste robusto revela claramente a discrepância das 3 observações mencionadas. Este fato pode ser visualizado nas figuras 7.1 e 7.2.

Por outro lado, na tabela 7.5 vemos que os elementos da diagonal da matriz de projeção e as distâncias de Mahalanobis (colunas 1 e 2, respectivamente) revelam as observações 4, 21 e 29 como sendo discrepantes na direção das variáveis regressoras. No entanto, as distâncias robustas (MVE), mostram as observações 3, 5, 17, 28 e 31, além daquelas identificadas pelo método não robusto.

O programa computacional MINVOL usado para o cálculo das distâncias, tanto a de Mahalanobis quanto as robustas, foi obtido da

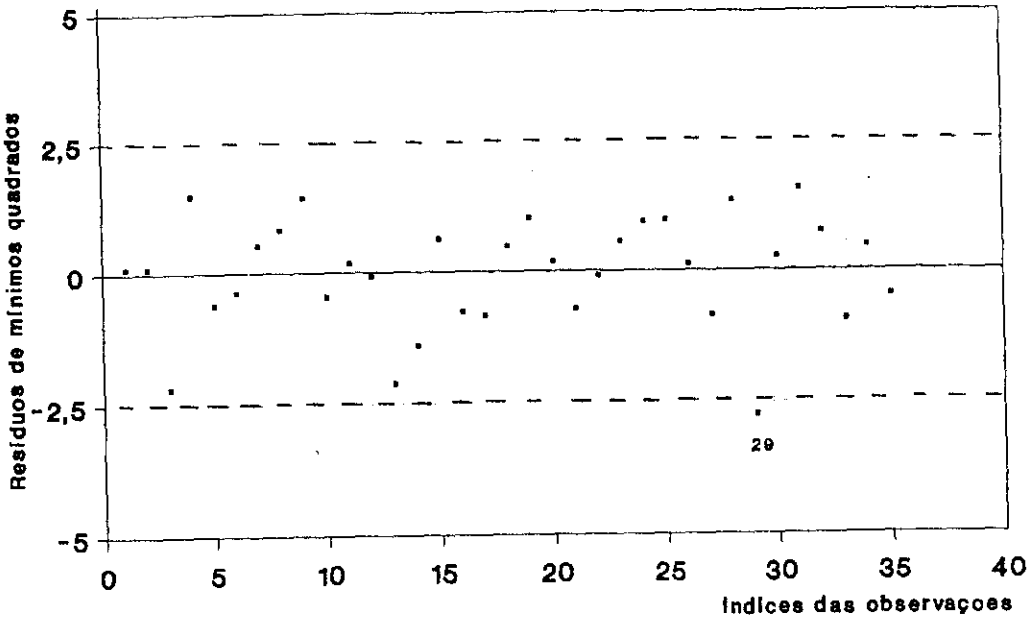

Figura 7.1. Resíduos padronizados associados ao ajuste de ajuste de mínimos quadrados.

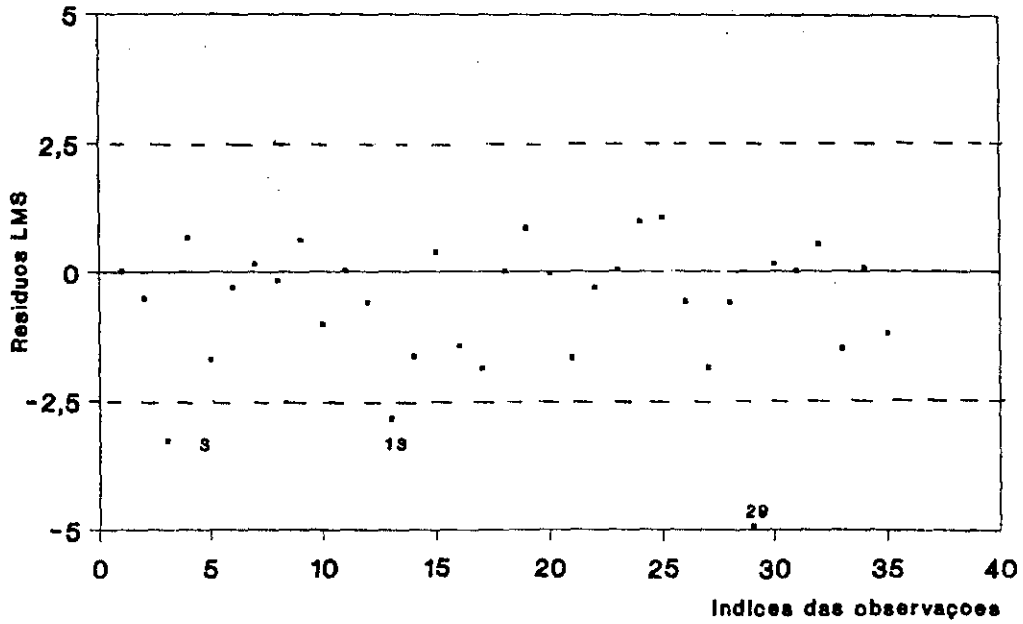

Figura 7.2. Resíduos padronizados associados ao ajuste LMS.

biblioteca de programas computacionais "The StatLib Collection of Applied Statistics Algorithms". O endereço electrônico da biblioteca é

STATLIBD@STAT.CMU.EDU

Obtivemos as estatísticas de diagnósticos de observações discrepantes individuais baseadas no ajuste de mínimos quadrados, os quais são apresentados na Tabela 7.6. Os valores de corte para cada tipo de diagnóstico aparecem na parte superior da coluna correspondente. Notemos que as observação 29 excede amplamente o valor de corte de todas as estatísticas de diagnósticos exceto para o DFFITS da variável resistividade esférica (RMSFL) . As observações 3 e 4 apresentam estatísticas DFFITS que excedem o valor de corte e têm uma maior influência nas estimativas correspondentes as variáveis argilosidade (VSH) e diferênça porosidade de densidade e porosidade neutrônica (DPHI).

A figura 7.3, relacionando os resíduos padronizados do ajuste de mínimos quadrados e as distâncias de Mahalanobis mostra a observação 29 como sendo um ponto de alavanca "ruim", e as observações 4 e 21 como pontos de alavanca "bons" . Por outro lado, a figura 7.4, relacionando os resíduos robustos padronizados e as distâncias robustas mostra as observações 3 e 29 como sendo pontos de alavanca "ruins" e a observação 13 como sendo discordante na direção da variável resposta. As observações 4, 5, 17, 21, 28 e 31 são exibidas como sendo pontos de alavanca "bons".

Tabela 7.4 Resíduos "studentizados" externamente e resíduos padronizados associados ao ajuste de mínimos quadrados e resíduos LMS padronizados.

Nº de Obs. (i)	Ajuste de Mínimos Quadrados		Ajuste LMS
	Res. "Student." (t_i)	Res. Padron. (r_i)	Res. Padron. (e_i)
1	0.1077	0.109	0.00
2	0.0931	0.095	-0.42
3	-2.3684	-2.210	-2.89
4	1.5234	1.492	0.82
5	-0.6093	-0.616	-1.47
6	-0.3766	-0.382	-0.37
7	0.5271	0.533	0.18
8	0.8241	0.828	-0.00
9	1.4713	1.444	0.72
10	-0.4755	-0.482	-0.96
11	0.1850	0.188	-0.11
12	-0.0809	-0.082	-0.56
13	-2.2807	-2.140	-2.62
14	-1.4187	-1.396	-1.64
15	0.6136	0.623	0.04
16	-0.7688	-0.774	-1.57
17	-0.8222	-0.827	-2.15
18	0.4820	0.488	-0.24
19	1.0078	1.008	0.48
20	0.1912	0.194	-0.27
21	-0.7167	-0.722	-2.09
22	-0.0969	-0.099	-0.52
23	0.5479	0.554	-0.49
24	0.9463	0.948	0.56
25	0.9616	0.963	0.46
26	0.1064	0.108	-0.62
27	-0.8390	-0.843	-1.67
28	1.3469	1.330	-0.49
29	-3.1330	-2.764	-4.24
30	0.2608	0.265	0.00
31	1.6104	1.571	0.00
32	0.7180	0.724	0.30
33	-0.9268	-0.929	-1.57
34	0.4368	0.443	-0.12
35	-0.4776	-0.484	-1.23

Tabela 7.5 Elementos da diagonal da matriz de projeção, distâncias de Mahalanobis e distâncias robustas.

Nº de Obs. (i)	Diagonal da Mat. de projeção (h_{ii})	Distâncias de Mahalanobis (MD_i)	Distâncias robustas (RD_i)
1	0.1210	1.898	2.591
2	0.1108	2.089	3.006
3	0.1146	2.792	3.757
4	0.3549	3.376	4.106
5	0.0551	3.052	3.777
6	0.1463	2.117	2.276
7	0.0552	1.066	1.027
8	0.1095	1.760	1.235
9	0.1154	1.808	1.042
10	0.0739	1.384	1.883
11	0.0986	2.012	2.156
12	0.0436	0.874	0.985
13	0.0492	1.213	1.078
14	0.0683	1.578	1.235
15	0.0503	0.910	0.813
16	0.0459	0.768	0.980
17	0.2001	2.460	3.786
18	0.0489	1.602	1.094
19	0.0481	0.902	0.562
20	0.0528	1.389	1.087
21	0.3624	3.376	4.840
22	0.0497	0.868	0.555
23	0.1590	2.106	2.582
24	0.0787	1.465	1.135
25	0.1263	1.986	1.235
26	0.0346	0.864	0.421
27	0.0775	2.003	1.235
28	0.2919	3.125	6.119
29	0.3487	3.465	5.935
30	0.1278	1.903	2.362
31	0.2079	2.497	4.853
32	0.0492	1.514	1.197
33	0.0584	1.442	1.235
34	0.0355	0.774	0.372
35	0.0296	0.204	0.734

Nota.- O ponto de corte para as distâncias de Mahalanobis e robustas é $\left(\chi^2_{4,.975}\right)^{1/2} = 3.3381$.

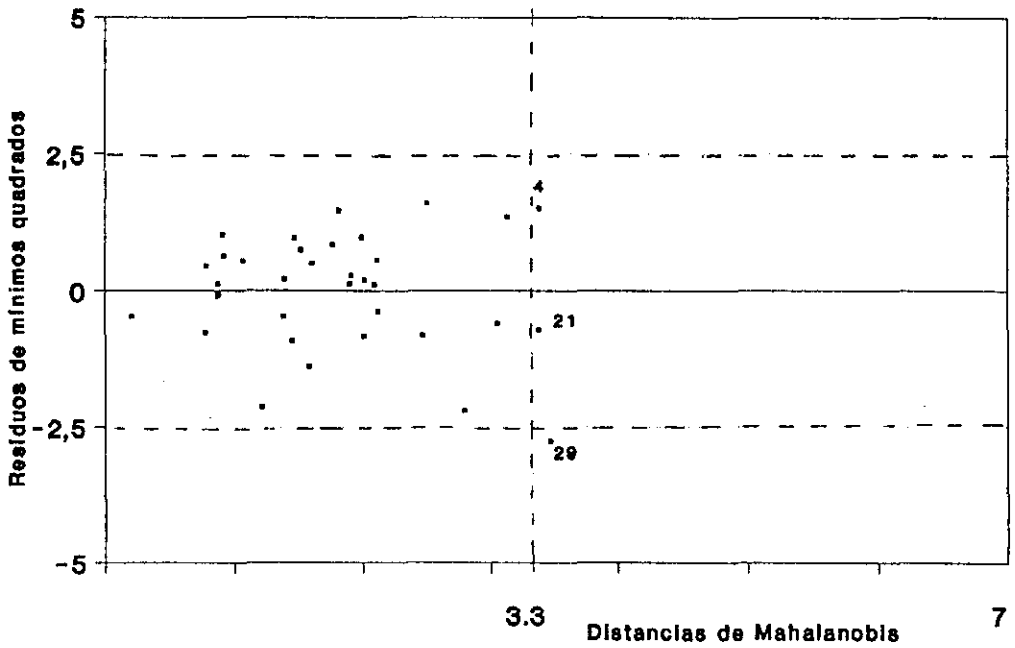

Figura 7.3. Resíduos padronizados associados ao ajuste de ajuste de mínimos quadrados vs dist. de Mahalanobis.

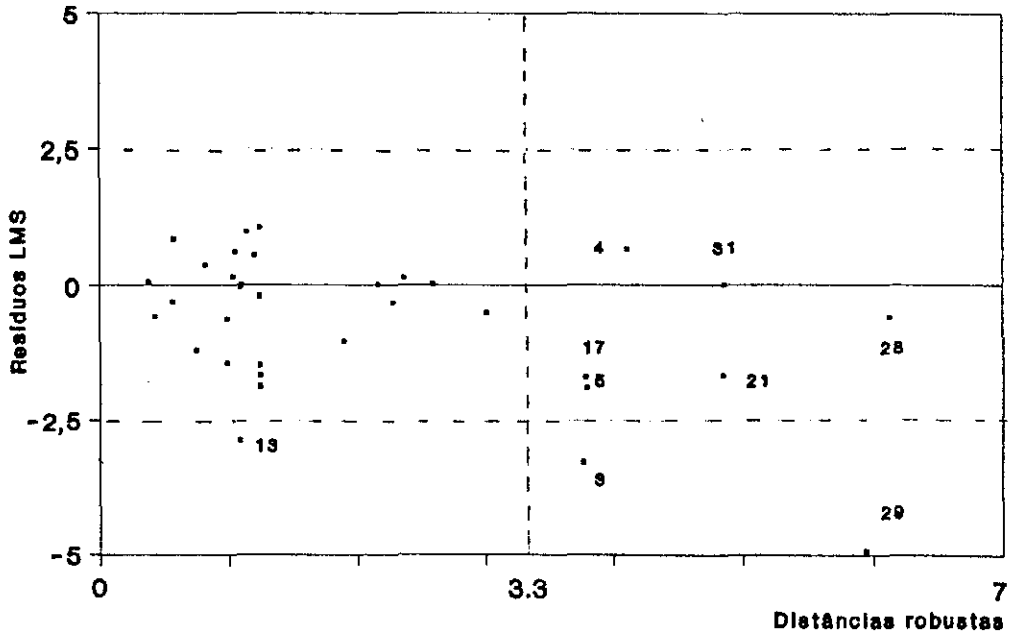

Figura 7.4. Resíduos padronizados associados ao ajuste LMS versus distancias robustas.

Tabela 7.6 Diagnósticos DFFITS, de COOK e DFBETAS do método de estimação de mínimos quadrados.

No de Obs.	DFFITS	COOK	D F B E T A S (0.338)			
(i)	(0.686)	(1.0)	RMSFL	VSH	PHID	DPHI
1	0.0400	0.000	-0.0190	-0.0324	0.0323	0.0263
2	0.0329	0.000	0.0042	-0.0238	0.0067	0.0307
3	-0.8522	0.158	-0.1991	0.5684	-0.0627	-0.7959
4	1.1299	0.306	0.0396	-0.8399	0.2419	1.1148
5	-0.1472	0.006	-0.0391	-0.0092	0.0256	-0.0923
6	-0.1559	0.006	0.1276	0.0109	-0.1364	0.0431
7	0.1275	0.004	-0.0850	-0.0386	0.1009	0.0479
8	0.2890	0.021	-0.0595	-0.0296	0.0597	0.1738
9	0.5314	0.068	-0.1118	-0.1383	0.1406	0.3800
10	-0.1344	0.005	0.0948	-0.0348	-0.0886	0.0406
11	0.0612	0.001	-0.0417	0.0180	0.0396	-0.0311
12	-0.0173	0.000	0.0109	0.0018	-0.0128	-0.0027
13	-0.5186	0.059	0.3651	0.1482	-0.4629	-0.0644
14	-0.3842	0.036	0.2329	0.1252	-0.3341	0.0316
15	0.1419	0.005	0.0236	0.0197	0.0228	-0.0543
16	-0.1686	0.007	-0.0649	-0.0005	-0.0031	0.0213
17	-0.4112	0.043	-0.3684	-0.0842	0.2605	0.0409
18	0.1093	0.003	0.0365	-0.0266	0.0138	0.0217
19	0.2266	0.013	0.0355	-0.0214	0.0581	-0.0340
20	0.0452	0.001	-0.0124	-0.0099	0.0286	-0.0041
21	-0.5403	0.074	-0.4970	-0.0264	0.3347	-0.0033
22	-0.0222	0.000	0.0035	0.0061	-0.0125	0.0019
23	0.2382	0.015	0.1751	0.0476	-0.1023	-0.0702
24	0.2766	0.019	-0.0599	0.0163	0.1331	-0.1248
25	0.3656	0.033	0.0925	0.0464	0.0179	-0.1905
26	0.0201	0.000	0.0006	-0.0039	0.0061	0.0075
27	-0.2432	0.015	-0.0162	0.1275	-0.0599	-0.1961
28	0.8648	0.182	0.2083	0.6923	-0.4150	-0.2350
29	-2.2923	1.023	-0.0564	-1.6887	0.6739	0.4707
30	0.0999	0.003	-0.0691	0.0268	0.0664	-0.0516
31	0.8251	0.162	0.1368	0.6548	-0.3134	-0.2466
32	0.1634	0.007	-0.0470	-0.0564	0.1104	0.0162
33	-0.2308	0.013	0.0562	-0.0254	-0.1137	0.0898
34	0.0838	0.002	-0.0157	-0.0102	0.0442	0.0009
35	-0.0834	0.002	-0.0051	-0.0053	-0.0212	0.0022

Vejamos agora o que acontece se retiramos a observação 29 do conjunto de dados. Os coeficientes estimados pelo método de mínimos quadrados são apresentados na tabela 7.7, a seguir.

Tabela 7.7 Coeficientes de regressão estimados, desvio padrão, estatistica T, e probabilidade de rejeição, associados ao ajuste de mínimos quadrados. (Sem observação 29).

Variável	Coeficiente estimado	Desvio padrão	Estatística T	Prob > T
RMSFL	0.072603	0.0220288	3.296	0.0025
VSH	-2.631765	1.9313592	-1.363	0.1831
PHID	0.140267	0.0310939	4.511	0.0001
DPHI	0.196214	0.0770391	2.547	0.0162

$R^2 = 0.9365$

$s^2 = 1.20912$

A retirada da observação 29 fez decrescer a diferença entre as estimativas robusta e de mínimos quadrados correspondente a variável argilosidade (VSH), a qual deixou de ser significante no modelo (Tabela 7.7). Por outro lado, as observações 3 e 13 aparecem agora mais claramente com sendo discrepantes, isto é, tanto os resíduos do ajuste de mínimos quadrados, quanto do ajuste robusto (Tabela 7.8) mostram estas mesmas observações como sendo discrepantes. Logo, o conjunto de dados tem na verdade 3 observações atípicas. Notar que a discrepância das observações 3 e 13 estava sendo mascarada pela forte discordância da observação 29.

As distâncias de Mahalanobis das observações 3 e 21 ainda excedem o ponto de corte, mas agora também a observação 28 aparece como sendo discrepante, esta observação estava mascarada pela discrepância da observação 29 (Tabela 7.9). No entanto, a observação 21 mantem a distância de Mahalanobis excedendo o valor de corte, mas sua distância robusta decresceu.

Tabela 7.8 Resíduos "studentizados" externamente e resíduos padronizados associados ao ajuste de mínimos quadrados e resíduos LMS padronizados (Sem Obs. 29).

Nº de Obs. (i)	Ajuste de Mínimos Quadrados		Ajuste LMS
	Res. "Student." (t_i)	Res. Padron. (r_i)	Res. Padron. (e_i)
1	0.4355	0.441	-0.23
2	0.2214	0.225	-0.84
3	-2.6603	-2.426	-3.54
4	2.0660	1.929	0.00
5	-1.0137	-1.013	-1.69
6	-0.3704	-0.376	-0.04
7	0.4713	0.478	0.21
8	0.4585	0.465	-0.15
9	1.2780	1.265	0.55
10	-0.8230	-0.827	-0.71
11	0.0190	0.019	0.36
12	-0.2540	-0.258	-0.47
13	-2.6811	-2.441	-2.72
14	-1.4342	-1.410	-1.56
15	0.7110	0.717	0.45
16	-0.8585	-0.862	-1.40
17	-1.0215	-1.021	-1.94
18	0.5817	0.588	-0.06
19	1.2240	1.214	0.84
20	0.3025	0.307	0.00
21	-0.6544	-0.661	-1.92
22	-0.0390	-0.040	0.30
23	0.6868	0.693	0.00
24	1.1711	1.164	1.09
25	1.3409	1.323	1.11
26	0.0021	0.002	-0.56
27	-1.0305	-1.029	-1.98
28	0.0827	0.084	0.00
30	0.1192	0.121	0.53
31	0.6723	0.678	0.52
32	0.9012	0.904	0.53
33	-1.0794	-1.076	-1.28
34	0.4532	0.459	0.11
35	-0.6572	-0.664	-1.09

Tabela 7.9 Elementos da diagonal da matriz de projeção, distâncias de Mahalanobis e distâncias robustas (Sem observação 29).

No de Obs. (i)	Diagonal da Mat. de projeção (h_{ii})	Distâncias de Mahalanobis (MD_i)	Distâncias robustas (RD_i)
1	0.1299	1.906	2.506
2	0.1120	2.110	2.869
3	0.1152	2.820	3.659
4	0.3584	3.329	3.459
5	0.0644	3.004	3.711
6	0.1466	2.095	0.932
7	0.0568	1.044	0.932
8	0.1304	1.988	2.704
9	0.1285	1.947	2.352
10	0.0812	1.402	1.221
11	0.1020	1.976	1.053
12	0.0462	0.866	0.842
13	0.0492	1.195	0.907
14	0.0712	1.644	1.633
15	0.0503	0.886	0.932
16	0.0459	0.739	0.901
17	0.2006	2.418	2.705
18	0.0490	1.588	0.932
19	0.0486	0.874	0.664
20	0.0535	1.361	1.024
21	0.3641	0.334	3.227
22	0.0519	0.897	1.126
23	0.1593	2.071	1.994
24	0.0794	1.434	0.967
25	0.1309	2.035	2.170
26	0.0360	0.921	0.896
27	0.0780	2.024	1.582
28	0.4398	3.702	6.926
30	0.1307	1.878	0.932
31	0.3144	3.171	5.922
32	0.0499	1.488	1.177
33	0.0585	1.438	1.045
34	0.0356	0.774	0.497
35	0.0309	0.226	0.764

Nota.- O ponto de corte para as distâncias de Mahalanobis e robustas é o quantil $\left(\chi^2_{4,.975}\right)^{1/2} = 3.3381$.

Tabela 7.10 Diagnósticos DFFITS, de COOK e DFBETAS do método de estimação de mínimos quadrados. (Sem observação 29).

No de Obs.	DFFITS	COOK	D F B E T A S (0.343)			
(i)	(0.686)	(1.0)	RMSFL	VSH	PHID	DPHI
1	0.1683	0.007	-0.0779	-0.1369	0.1378	0.1123
2	0.0787	0.002	0.0098	-0.0538	0.0173	0.0735
3	-0.9599	0.192	-0.2224	0.5941	-0.0830	-0.8949
4	1.5143	0.520	0.0501	-1.0574	0.3471	1.4925
5	-0.2660	0.018	-0.0672	-0.0343	0.0630	-0.1377
6	-0.1535	0.006	0.1257	0.0126	-0.1326	0.0409
7	0.1157	0.003	-0.0758	-0.0214	0.0843	0.0395
8	0.1776	0.008	-0.0322	0.0190	0.0179	0.0862
9	0.4908	0.059	-0.0950	-0.0321	0.0873	0.3056
10	-0.2447	0.015	0.1633	-0.0880	-0.1350	0.0807
11	0.0064	0.000	-0.0043	0.0022	0.0037	-0.0033
12	-0.0559	0.001	0.0341	-0.0014	-0.0366	-0.0063
13	-0.6098	0.077	0.4289	0.1470	-0.5292	-0.0729
14	-0.3972	0.038	0.2372	0.1496	-0.3475	0.0197
15	0.1637	0.007	0.0271	0.0188	0.0263	-0.0616
16	-0.1883	0.009	-0.0723	0.0014	-0.0042	0.0230
17	-0.5117	0.065	-0.4583	-0.1050	0.3223	0.0543
18	0.1321	0.004	0.0439	-0.0314	0.0177	0.0269
19	0.2767	0.019	0.0426	-0.0362	0.0750	-0.0366
20	0.0720	0.001	-0.0197	-0.0177	0.0461	-0.0052
21	-0.4951	0.062	-0.4537	-0.0055	0.2922	-0.0079
22	0.0091	0.000	-0.0014	-0.0031	0.0053	-0.0005
23	0.2990	0.023	0.2192	0.0458	-0.1225	-0.0850
24	0.3440	0.029	-0.0748	0.0019	0.1681	-0.1478
25	0.5205	0.066	0.1276	0.0106	0.0451	-0.2488
26	0.0004	0.000	0.0000	-0.0000	0.0001	0.0001
27	-0.2998	0.022	-0.0204	0.1264	-0.0669	-0.2346
28	0.0733	0.001	0.0151	0.0622	-0.0369	-0.0224
30	0.0462	0.001	-0.0315	0.0140	0.0283	-0.0244
31	0.4553	0.053	0.0661	0.3843	-0.1932	-0.1488
32	0.2066	0.011	-0.0595	-0.0738	0.1406	0.0237
33	-0.2690	0.018	0.0653	-0.0299	-0.1278	0.1046
34	0.0871	0.002	-0.0162	-0.0064	0.0435	-0.0000
35	-0.1174	0.004	-0.0074	-0.0178	-0.0234	0.0066

Na tabela 7.10, vemos que as observações 3 e 4 ainda têm estatísticas DFFITS excedendo os pontos de corte e influenciando as estimações dos coeficientes correspondentes às variáveis VSH e DPHI. No entanto, as observações próximas de 29 tais como 28 e 31 deixaram de ter influência notável.

Concluindo, a retirada da observação 29 não resolve o problema pois não é a única observação discrepante "ruim". A análise dos dados retirando as observações 3 e 29 também não produz melhoras notáveis, isto é, a variável VSH permanece sendo não significante (tabela 7.11). Além disso, tanto os resíduos do ajuste robusto quanto

Tabela 7.11 Coeficientes de regressão estimados, desvio padrão, estatística T, e probabilidade de rejeição, associados ao ajuste de mínimos quadrados. (Sem observações 3 e 29).

Variável	Coeficiente estimado	Desvio padrão	Estatística T	Prob > T
RMSFL	0.077071	0.02015794	3.823	0.0006
VSH	-3.678057	1.80457050	-2.038	0.0507
PHID	0.142620	0.02836807	5.027	0.0001
DPHI	0.259043	0.07411486	3.495	0.0015

$R^2 = 0.9489$

$s^2 = 1.0054$

os resíduos do ajuste de mínimos quadrados mostram a observação 13 como sendo discrepante (Tabela 7.12).

Reanalizando os dados sem as observações 3, 13 e 29, as quais foram identificadas pelo método de diagnóstico robusto (figura 7.4), os resultados são

Tabela 7.12 Resíduos "studentizados" externamente e resíduos padronizados associados ao ajuste de mínimos quadrados e resíduos LMS padronizados. (Sem Obs. 3 e 29).

Nº de Obs. (i)	Ajuste de Mínimos Quadrados		Ajuste LMS
	Res. "Student." (t_i)	Res. Padron. (r_i)	Res. Padron. (e_i)
1	0.2491	0.253	0.00
2	-0.0901	-0.092	0.00
4	1.5011	1.470	1.60
5	-1.3008	-1.286	-1.11
6	-0.2866	-0.291	-0.91
7	0.4323	0.438	0.18
8	0.2997	0.305	0.40
9	1.1507	1.144	1.16
10	-0.8263	-0.831	-1.21
11	0.1715	0.174	-0.48
12	-0.3186	-0.324	-0.61
13	-3.0828	-2.771	-2.86
14	-1.5795	-1.540	-1.98
15	0.8004	0.805	0.00
16	-0.9886	-0.989	-1.54
17	-1.2347	-1.224	-1.74
18	0.5291	0.536	-0.08
19	1.3168	1.300	0.50
20	0.3092	0.314	-0.38
21	-0.9426	-0.944	-1.50
22	0.0163	0.017	-0.64
23	0.7406	0.746	-0.29
24	1.3684	1.348	0.33
25	1.5869	1.547	0.29
26	-0.1048	-0.107	-0.47
27	-1.4158	-1.392	-1.36
28	0.2014	0.205	0.00
30	0.3109	0.316	-0.45
31	0.8333	0.838	0.38
32	0.9286	0.931	0.26
33	-1.1436	-1.138	-1.79
34	0.4541	0.460	-0.11
35	-0.7653	-0.771	-1.20

Tabela 7.13 Coeficientes de regressão estimados, desvio padrão, estatistica T, e probabilidade de rejeição, associados ao ajuste de mínimos quadrados (Sem observações 3, 13 e 29).

Variável	Coeficiente estimado	Desvio padrão	Estatística T	Prob > T
RMSFL	0.068416	0.01794689	3.812	0.0007
VSH	-3.953331	1.58936229	-2.487	0.0191
PHID	0.157822	0.02542827	6.207	0.0001
DPHI	0.265033	0.06520197	4.065	0.0004

$R^2 = 0.9618$

$s^2 = 0.7775$

Notar que todas as variáveis são significantes, o coeficiente de determinação melhorou e as variâncias estimadas tanto dos erros quanto dos coeficientes estimados são menores com respecto às análises anteriores, além disso, todos os resíduos encontram-se dentro do intervalo (-2.5, 2.5). Mesmo se considerarmos um intervalo de amplitude menor, por exemplo (-2, 2), os resíduos permanecem dentro do intervalo.

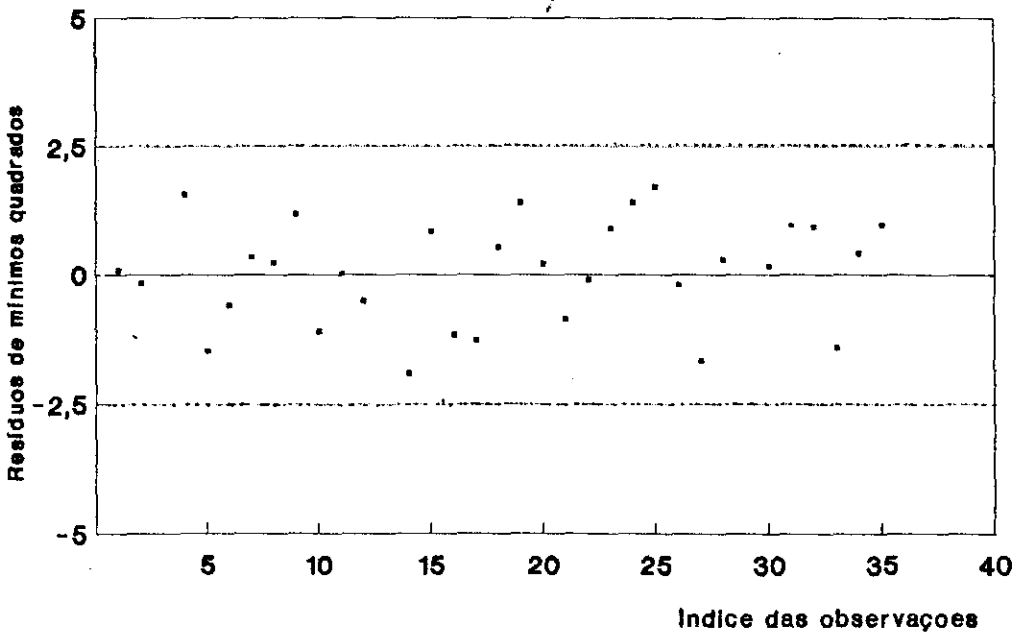

Figura 7.5 Resíduos padronizados associados ao ajuste de mínimos quadrados (sem Obs. 3, 13 e 29).

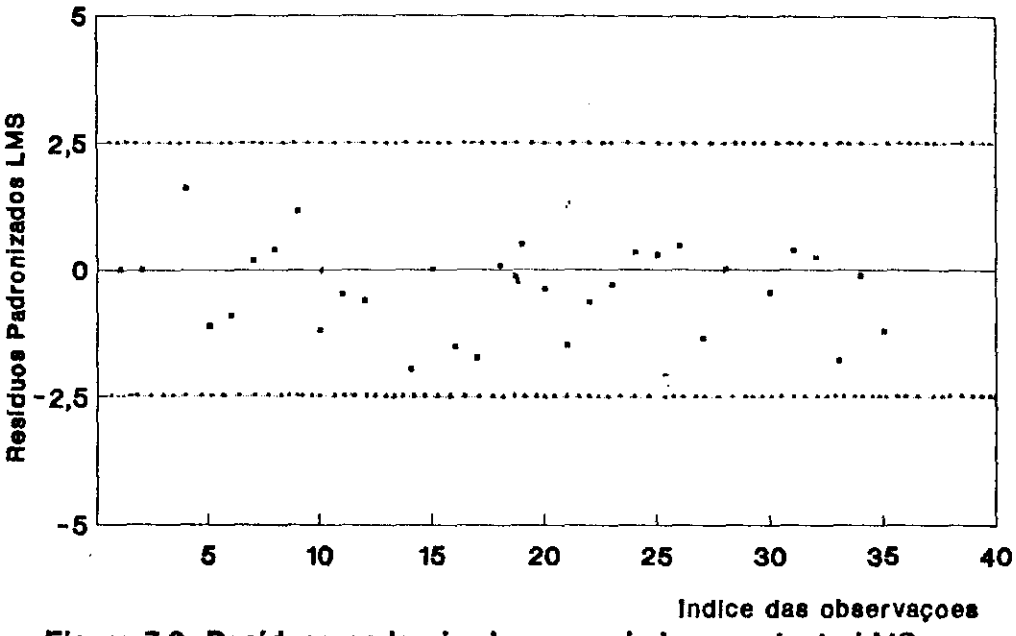

Figura 7.6 Resíduos padronizados associados ao ajuste LMS (sem Obs. 3, 13 e 29).

Tabela 7.14 Resíduos "studentizados" externamente e resíduos padronizados associados ao ajuste de mínimos quadrados e resíduos LMS padronizados. (Sem Obs. 3,13 e 29).

No de Obs. (i)	Ajuste de Mínimos Quadrados		Ajuste LMS
	Res. "Student." (t_i)	Res. Padron. (r_i)	Res. Padron. (e_i)
1	0.0817	0.083	0.00
2	-0.1740	-0.177	0.00
4	1.5871	1.546	1.60
5	-1.5198	-1.485	-1.11
6	-0.5802	-0.587	-0.91
7	0.3349	0.340	0.18
8	0.2311	0.235	0.40
9	1.1925	1.184	1.16
10	-1.1192	-1.114	-1.21
11	0.0186	0.019	-0.48
12	-0.5049	-0.512	-0.61
14	-2.0305	-1.926	-1.98
15	0.8400	0.844	0.00
16	-1.1856	-1.177	-1.54
17	-1.2871	-1.272	-1.74
18	0.5331	0.540	-0.08
19	1.4226	1.397	0.50
20	0.2179	0.222	-0.38
21	-0.8624	-0.866	-1.50
22	-0.0953	-0.097	-0.64
23	0.8965	0.900	-0.29
24	1.4276	1.402	0.33
25	1.7678	1.704	0.29
26	-0.2062	-0.210	-0.47
27	-1.7379	-1.678	-1.36
28	0.2873	0.292	0.00
30	0.1498	0.152	-0.45
31	0.9495	0.951	0.38
32	0.9192	0.922	0.26
33	-1.4500	-1.422	-1.79
34	0.4068	0.413	-0.11
35	-0.9547	-0.956	-1.20

Tabela 7.15 Elementos da diagonal da matriz de projeção, distâncias de Mahalanobis e distâncias robustas (Sem observações 3, 13 e 29).

Nº de Obs. (i)	Diagonal da Mat. de projeção (h_{ii})	Distâncias de Mahalanobis (MD_i)	Distâncias robustas (RD_i)
1	0.1404	1.869	3.780
2	0.1268	2.461	3.268
4	0.4034	3.397	4.389
5	0.0684	3.388	3.788
6	0.1542	2.091	2.219
7	0.0603	1.123	1.099
8	0.1368	1.915	2.520
9	0.1374	1.887	1.981
10	0.0847	1.391	1.193
11	0.1080	1.983	1.144
12	0.0485	1.923	0.815
14	0.0741	1.735	2.296
15	0.0509	0.850	0.835
16	0.0464	0.695	1.020
17	0.2030	2.369	4.017
18	0.0511	1.569	1.169
19	0.0494	0.823	0.543
20	0.0555	1.361	0.978
21	0.3712	3.272	4.078
22	0.0534	0.892	1.078
23	0.1596	1.995	2.644
24	0.0818	1.442	0.936
25	0.1323	1.983	1.099
26	0.0385	0.867	1.099
27	0.0875	1.955	1.328
28	0.4410	3.580	9.660
30	0.1387	1.861	1.099
31	0.3152	3.115	8.206
32	0.0523	1.482	1.099
33	0.0603	1.486	1.131
34	0.0371	0.740	0.517
35	0.0318	0.166	1.136

Nota.- O ponto de corte para as distâncias de Mahalanobis e robustas é

o quantil $\left(\chi^2_{4,.975}\right)^{1/2} = 3.3381$.

Tabela 7.16 Diagnósticos DFFITS, de COOK e DFBETAS do método de estimação de mínimos quadrados (Sem observações 3, 13, 29).

No Obs. (i)	DFFITS (0.686)	COOK (1.0)	D F B E T A S (0.354) RMSFL	VSH	PHID	DPHI
1	0.0330	0.000	-0.0147	-0.0271	0.0268	0.0225
2	-0.0663	0.001	-0.0088	0.0467	-0.0149	-0.0624
4	1.3051	0.404	0.0653	-0.9345	0.3023	1.2880
5	-0.4117	0.040	-0.1056	-0.0282	0.0876	-0.2276
6	-0.2477	0.016	0.2042	0.0162	-0.2131	0.0681
7	0.0848	0.002	-0.0549	-0.0181	0.0625	0.0306
8	0.0920	0.002	-0.0159	0.0050	0.0111	0.0473
9	0.4759	0.056	-0.0857	-0.0562	0.0926	0.3084
10	-0.3405	0.029	0.2312	-0.1205	-0.1912	0.1127
11	0.0065	0.000	-0.0044	0.0023	0.0038	-0.0034
12	-0.1140	0.003	0.0698	0.0002	-0.0763	-0.0150
14	-0.5744	0.074	0.3491	0.2121	-0.5050	0.0244
15	0.1945	0.010	0.0280	0.0218	0.0341	-0.0703
16	-0.2616	0.017	-0.0968	0.0073	-0.0102	0.0233
17	-0.6497	0.103	-0.5813	-0.1211	0.4070	0.0504
18	0.1237	0.004	0.0396	-0.0335	0.0189	0.0308
19	0.3244	0.025	0.0442	-0.0467	0.0933	-0.0340
20	0.0528	0.001	-0.0154	-0.0134	0.0345	-0.0027
21	-0.6626	0.111	-0.6068	0.0049	0.3889	-0.0307
22	-0.0226	0.000	0.0040	0.0077	-0.0135	0.0007
23	0.3906	0.038	0.2844	0.0581	-0.1595	-0.1041
24	0.4261	0.044	-0.1026	0.0072	0.2120	-0.1814
25	0.6903	0.111	0.1556	0.0255	0.0625	-0.3302
26	-0.0412	0.000	-0.0011	0.0050	-0.0112	-0.0153
27	-0.5381	0.068	-0.0395	0.2484	-0.1267	-0.4320
28	0.2552	0.017	0.0517	0.2139	-0.1274	-0.0777
30	0.0601	0.001	-0.0416	0.0189	0.0367	-0.0322
31	0.6443	0.104	0.0892	0.5365	-0.2687	-0.2098
32	0.2160	0.012	-0.0645	-0.0803	0.1494	0.0308
33	-0.3674	0.032	0.0978	-0.0410	-0.1794	0.1390
34	0.0799	0.002	-0.0161	-0.0079	0.0414	0.0025
35	-0.1731	0.008	-0.0081	-0.0205	-0.0387	0.0034

COMENTARIOS FINAIS

Os métodos de diagnósticos de múltiplas observações discrepantes baseados no ajuste de mínimos quadrados, podem avaliar a influência de grupos de observações. Porém, estes métodos raramente são usados pela dificuldade em determinar quantas e quais são as observações que devem ser consideradas como conjuntamente influentes. Por outro lado, os métodos robustos em geral, e particularmente os estimadores com alto ponto de ruptura, têm a tendência a declarar mais observações discrepantes das que realmente existem no conjunto de dados. Consequentemente, o método exploratório estudado neste trabalho pode servir para identificar grupos de observações suspeitas e posteriormente comprovar a discrepância destes grupos de observações utilizando os diagnósticos de múltiplas observações discrepantes baseados no ajuste de mínimos quadrados. Atkinson (1986) mencionou a importância destes métodos confirmatorios e posteriormente Fung (1993) retomou esta idéia e propôs novas técnicas exploratórias.

APENDICE

DEFINIÇÃO A1

Dada uma amostra aleatória unidimensional $\left\{z_1, \ldots, z_n\right\}$. Definimos a *Função de Distribuição Empírica* como

$$P_n\left(z_i\right) = \frac{1}{n} \sum_{i=1}^{n} I_{\left\{Z \leq z_i\right\}} \quad ; \qquad z_i \in \mathbb{R},$$

independentemente do ordem das observações.

DEFINIÇÃO A2

Definimos um *funcional estatístico* ou *função estatística* como

$$T : D(T) \longrightarrow \mathbb{R},$$

onde D(T) é um conjunto de distribuições de probabilidade definidas sobre $\mathbb{R}$. Denotaremos por $\mathcal{F}(\mathbb{R})$ o conjunto de todas as distribuições de probabilidade definidas sob $\mathbb{R}$.

Exemplos de funcionais estatísticos são:

a) Esperança

$$T(F) = \int_{\mathbb{R}} xd(F(x)).$$

$$D(T) = \left\{F : F \text{ distribuição sobre } \mathbb{R} \text{ com momento de } 1^{a} \text{ ordem finito}\right\},$$

$D(T) \subset \mathfrak{F}(\mathbb{R})$.

b) Variância

$$T(F) = \int_{\mathbb{R}} \left(x - \int_{\mathbb{R}} xd(F(x))\right)^{2} F(x),$$

$$D(T) = \left\{F : F \text{ distribuição sobre } \mathbb{R} \text{ com momento de } 2^{a} \text{ ordem finito}\right\},$$

$D(T) \subset \mathfrak{F}(\mathbb{R})$.

.

DEFINIÇÃO A3

Dizemos que o funcional T é *consistente* se é contínuo e

$$T(F) = \lim_{n \longrightarrow \infty} T\left(F_{n}\right),$$

onde F_{n} é a função de distribuição empírica acumulada.

DEFINIÇÃO A4

Uma propriedade que é transformada adequadamente sob transformações e chamada de *equivariante sob transformações.*

Estamos interessados na propriedade de equivariância sob rotação, translação e sob transformações afins de $t_n(X)$, $C(X)$ e $T(X,y)$, estimadores de posição e dispersão multivariados e do estimador de regressão, respectivamente. Vejamos em que consistem estas propriedades:

a) Dizemos que $T(X)$ é equivariante sob translação se:

$$t\ (X+v) = t\ (X) + v\ ; \qquad v \in \mathbb{R}^k,$$

onde

$$X+v = \left\{x_1+v,\ x_2+v,\ldots,x_n+v\right\}\ .$$

b) $T(X)$ é equivariante sob transformações afins se:

$$t_n(AX+v) = At_n(X) + v\ ; \qquad v \in \mathbb{R}^k,\quad A_{kxk} \text{ matriz não singular}$$

onde

$$AX+v = \left\{Ax_1+v,\ Ax_2+v,\ldots,Ax_n+v\right\}\ .$$

c) O estimador de dispersão, $C_n(X)$, é equivariante sob transformações afins se:

$$C_n(AX+v) = AC_n(X)A^t\ ; \qquad v \in \mathbb{R}^k,\quad A_{kxk} \text{ matriz não singular}$$

d) Dizemos que o estimador de regressão $t_n(X,\ y)$ é equivariante de regressão se:

$$t_n(X,\ y+Xv) = t_n(X,\ y) + v\ ; \qquad v \in \mathbb{R}^k,$$

onde

$$y + Xv = \left\{ y_i + x_1 v,\ y_2 + x_2 v,\ \ldots,\ y_n + x_n v \right\}.$$

e) Dizemos que o estimador de regressão $t_n(X,y)$ é equivariante sob transformações afins se :

$$t_n(AX,\ y) = A^{-1}\ t_n(X,\ y)\ ; \qquad A_{k\times k} \text{ matriz não singular}$$

Isto quer dizer que transformações nas variáveis regressoras, transformam t_n no mesmo sentido, devido a que

$$\hat{y} = x_i t_n\left(X\right) = \left(x_i A\right)\left(A^{-1} t_n\right).$$

Assim, podemos mudar o sistema de coordenadas das variáveis regressoras sem afetar as estimativas.

f) $t_n(X,y)$ é equivariante sob trocas de escala se:

$$t_n(X,cy) = c\,t_n(X,y); \qquad c \in \mathbb{R}^n.$$

DEFINIÇÃO A5

Qualquer espaço normado pode torna-se um espaço métrico a partir da definição de distância

$$d(x,y) = \| x-y \| \ ; \quad x,\ y \in \mathbb{R}^n.$$

DEFINIÇÃO A6

Uma norma L_p do vetor $w \in \mathbb{R}^n$ é definida como

$$\| w \|_p = \begin{cases} \left(\sum_{i=1}^{n} w_i^p \right)^{1/p} & 1 \leq p < \infty \\ \underset{1 \leq i \leq n}{\text{Max}} \; w_i & p = \infty. \end{cases}$$

DEFINIÇÃO A7

Um espaço linear normado $\mathbb{E}$ é dito *estritamente convexo* se:

$$\| y - z \| = \| y \| + \| z \| \quad ; y, z \in \mathbb{E}$$

o qual implica a existência de algum vetor $u \in \mathbb{E}$ tal que

$$y = \alpha u \quad e \quad z = \beta u \quad ; \alpha, \beta \in \mathbb{R}^+.$$

PROPOSIÇÃO A1

Para $1 < p < \infty$, os espaços L_p são estritamente convexos (Prova ver Köthe (1960, § 25.2)).

PROPOSIÇÃO A2

Um subespaço linear de dimensão finita, de um espaço linear normado, contem pelo menos um ponto de distância mínima desde um ponto fixo (Cheney (1966)).

DEFINIÇÃO A8

Dizemos que $\|.\|$ é uma *norma monótona* em $\mathbb{R}^k$ se os vetores $x, y \in \mathbb{R}^k$ estão relacionados pelas desigualdades

$$x_i \leq y_i \quad ; \qquad i=1,2,\ldots,n.$$

Então

$$\|x\| \leq \|y\|$$

Observação.- As normas L_p têm a propriedade de monotonia.

DEFINIÇÃO A9

Um conjunto de vetores $\{z_1, \ldots, z_n\}$, $z_j \in \mathbb{R}^{k+1}$ estão em *posição geral* se :

i) Não existem mais de k+1 pontos contidos em um hiperplano de dimensão menor de k.

ii) $n \geq k + 1$.

Observação.- A posição geral é conhecida também com o nome de *condição de Haar.*

BIBLIOGRAFIA

ATKINSON, A.C. (1986). *"Masking Unmasked"*. Biometrica. Vol 73, No 3.

BARNETT, V. & LEWIS, T. (1978). *"Outliers in Statistical Data"*. John Wiley & Sons. New York.

BARRODALE, I. & PHILLIPS, C. (1975). *"Algorithm 495: Solution of an Overdetermined Sistem of Linear Equations in the Chebyshev Norm."* ACM Transactions on Mathematical Software, 1, 254-270.

BASSET, Jr. G.W. (1991, (a)). *"how Estimates Differ when R^2 is Close to One"*. Department of Economics. University of Illinois at Chicago.

BASSET, Jr. G.W. (1991, (b)). *"Equivariant Monotonic 50% Breakdown Estimation"*. The American Statistician, Vol 43, No 2.

BECKMAN, R.J. & COOK, R.D (1983). *"Outlier.....s"*. Technometrics © Vol. 25, No 2.

BELSLEY, D.A. & KUH, E. & WELSCH, R.E. (1980). *"Regression Diagnostics: Identifying Influential Data and Sources of Colinearity"*. Wiley: New York.

BOX, G.E.P. (1953). *"Non Normality and Tests on Variance"*. Biometrica, 40 318-335.

BRADU, D. (1992). *"Recents developments in Elemental Regression Methods"*. Department of Statistics, University of South Africa.

BRADU, D. & HAWKINS, D.M. (1982). *"Location of Outliers in two-way tables Using tetrads"*. Technometrics ©, 24, 103-108.

BRADU, D. & HAWKINS, D.M. (1993). *"Sample Size Requeriments for Múltiple Outlier Location Techniques Based on Elemental Sets"*. Computational Statistics & Data Analysis 16 257-270. North Holland.

CHENEY, E.W. (1966). *"introduction of Aproximation Theory"*. McGraw-hill, New York.

COOK, R.D. (1977). *"detection of Influential Observations in Linear Regression"*. Technometrics, ©. 19 15-18.

COOK, R.D. & WEISBERG, S. (1982). *"residuals and Influence in Regression"*. New York: Chapman and Hall.

COOK, R.D. & HAWKINS, D.M. (1990). Comment on : *"Unmasking Multivariate Outliers and Leverage Points"*. Journal of American Statistical Association, 85, pp. 640-644.

DALLALL, E.D. & ROUSSEEUW, P.J. (1992). "LMSMVE : *A Program for Leat Median of Squares Regression and Robust Distances"*. Computers and Biomedical research 25, 384-391.

DAVIES, P.(1987). *Asymptotic Behaviour of S-estimates of Multivariate Location Parameters and Dispersion Matrices"*. Annals of Statistics 15 1269-1292.

DONOHO, D.L. (1982). *"Breakdown Properties of Multivariate Location Estimators"*. Qualifying Paper, Harvard University, Boston, M.A.

DONOHO, D.L. ; HUBER, P.J. (1983). *"The Notion of Breakdown Point"*. In a Festschrift for Erich L. Lehmann. Eds. P.J. Bickel, K.A. Docksum, J.L. Hodges. Wadsworth, Belmont, Calif. pp. 157-184.

EFRON, B. (1979). *"Bootstrap Methods: Another Look at the Jacknife"*. Annals of Statistics, 7, 1-26.

FUNG, Wing-Kam (1993). *"Unmasking Outliers and Leverage Points: A Confirmation"*. Journal of the American Statistical Association, Vol. 88, No 422, Theory and Methods.

HAMPEL, F.R. (1971). *"A General Qualitative Definition of Robustness"*. The Annals of Mathematical Statistics. Vol 42, No 6, pp 1887-96.

HAMPEL, F.R. (1974). *"The Influence Curve and its Role in Robust Estimation"*. Journal of the American statistical Association. Vol. 69 No 346, Theory and Methods.

HAMPEL, F. R. ; RONCHETTI, E.M. ; ROUSSEEUW, P.J.; STHAEL, W.A. (1986). *"Robust statistics: The Approach Based on Influence Functions"*. John Wiley & Sons. New York.

HAWKINS, D.M.; BRADU, D.; KASS, G. (1984). *"Location of Several Outliers in Multiple Regression Data Using Elemental Sets"*. Technometrics ©. Vol 26, No 3.

HAWKINS, D.M. & SIMONOFF, J. (1992), *"High Breakdown Regression and Multivariate estimation". U. Minnesota, U. New York.*

HAWKINS, D.M. (1993, (a)). *"The Feasible Set Algorithms for Least Median of Squares Regresion"*. Computational Statistics & Data Analysis 16 81-101. North Holland.

HAWKINS, D.M. (1993, (b)) *"The accuracy of Elemental Set Aproximation for Regression"*. Journal of the American Statistical Association. Vol. 88, No 422. Theory and Methods.

HETTMANSPERGER, T.P. & SHEATER, S.J. (1992). *"A Cautionary Note on the Method of Least Median Squares"*. The American Statistician. Vol. 46, No 2.

HOAGLIN, D.C. & WELSCH, R.E. (1978). *"The Hat Matrix in Regression and ANOVA"*. American Statistician, 32, 17-22.

HUBER, P.J. (1964). *"Robust Estimation of a Location Parameter"*. Annals of Mathematical statistics. Vol. 35, 73-101.

HUBER, P.J. (1972). *"Robust Statistics : A Review"*. Annals of Mathematical Statistics. Vol 43, 1041-1067.

HUBER, P.J. (1981). *"Robust Statistics"*. John Wiley & Sons New York.

HUBER, P.J. (1991). *"Between Robustness and Diagnostics"*. In Directions in Robust Statistics and Diagnostics : Part I. Eds. W. Stahel and S. Weisberg. New York, Springer-Verlag.

KÖTHE, G. (1960). *"Topologische Lineare Raümi I"*. Springer-Verlag. Berlin.

LEROY, M.A., ROUSSEEUW, P.J. (1984). "PROGRESS: *A Program for Robust Regression Analysis"*. Technical Report # 201. Center for Statistics and O.R. University of Brussels, Belgium.

LOPUNHAÄ, H.P., ROUSSEEUW, P.J. (1991). *"Breakdown Points of Affine Equivariant estimators of Multivariate Location and Covariance Matrices"*. The Annals of Statistics. Vol 19, No 1 pp 229-248.

NIQUIST, H. (1980). *"Recents Studies on L_p-Norm Estimation "* Statistical Research Report. University of UMEÅ. Sweden.

ROUSSEEUW, P.J. (1984). *"Least Median of Squares Regression"*. JASA, Vol. 79, No 388, Theory and Methods.

ROUSSEEUW, P.J. (1985). *"A Regression Diagnostic for Multiple Outliers and Leverage Points"*. Abstract in IMS Bulletin, 14 pp. 399.

ROUSSEEUW, P.J. (1992). *"A Diagnostic Plot for Regression Outliers and Leverage Points"*. Statistical Software Newsletter. U.I.A.

Vesaliuslaan 24. B2650 Edegem, Belgium.

ROUSSEEUW, P.J. & LEROY, M.A. (1987). *"Robust Regression and Outliers Detection"*. John Wiley & Sons.

ROUSSEEUW, P.J. & van ZOMEREN, B.(1987). *"Identification of Multivariate Outliers and Leverage Points by means of Robust Covariance Matrices"*. Technical Report, Faculty of Mathematics and Informatics. Delft University of Technology of the Netherlands.

ROUSSEEUW, P.J. & van ZOMEREN, B.(1990). *"Unmasking Multivariate Outliers and Leverage Points"*. JASA, Vol. 85, No 411, Theory and Methods,

ROUSSEEUW, P.J. & van ZOMEREN, B.(1991). *"Robust Distances: Simulations and Cutoff Values"*. In Directions in Robust Statistics and Diagnostics: Part II. Eds. W Stahel and Weisberg. New York. Springer-Verlag.

ROUSSEEUW, P.J. & BASSETT, Jr. G.W. (1991). *"Robustness of the p-subset Algorithms for Regression with High Breakdown Point"*. In Directions in Robust Statistics and Diagnostics: Part II. Eds. W. Stahel and Weisberg. New York, Springer-Verlag.

SEN, A. & SRIVASTAVA, M. (1990). *"Regression Analysis: Theory, Methods and Aplications"*. New York, Springer-Verlag.

STAHEL, J.M. (1981). *"Robuste Schätsungen : Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen"* Ph. D. Thesis, ETH Zurich, Switzerland.

STEELE, J.M. & STEIGER, W.L. (1986). *"Algorithms and Complexity for Leat Median of Squares Regression"*. Discrete Applied Mathematics 14 pp 93-100.

STIGLER, S.M. (1986). *"The History of Statistics. The Measurement of Uncertainty Before* 1900". The Belknap Press of Harvard University Press.

STROMBERG. A.J. (1991). *"Computing the Exact Value of The Least Median Squares estimate and Stability Diagnostics in Multiple Linear Regression"* Technical Report No 561. Department of Statistics. University of Minnesota.

TURNER, M. (1960). *"On Heuristic Estimation Methods"*. Biometrics 16 pp. 299-301.