Técnicas robustas de diagnóstico em análise multivariada e em análise de regressão

Agnes Yuka Simidu

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: **Estatística**
Orientadora: **Prof^a. Dr^a. Silvia Nagib Elian**

São Paulo, abril de 2009

Técnicas robustas de diagnóstico em análise multivariada e em análise de regressão

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Agnes Yuka Simidu
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof^a. Dr^a. Silvia Nagib Elian - IME USP.
- Prof. Dr. Rinaldo Artes - IBMEC.
- Prof. Dr. Francisco José de Azevedo Cysneiros - UFPE.

Agradeço a Deus por tudo,

pela minha saúde,

pela minha família e

por ter me dado a oportunidade

de mais esta conquista.

Aos meus pais Luiz e Akemi,
meu irmão Digo e
meu noivo Rodrigo.

Agradecimentos

Em primeiro lugar agradeço à minha orientadora, Prof[a] Silvia Nagib Elian, pela orientação, apoio, confiança, paciência e sobretudo pela amizade durante todo esse período. Professora, muito obrigada por tudo.

Agradeço aos meus pais que estiveram sempre ao meu lado e que desde cedo mostraram a importância dos estudos em minha vida.

Agradeço ao meu noivo Rodrigo por estar sempre ao meu lado, pelo amor, paciência e apoio em todos os momentos.

Agradeço a todos os professores pelos ensinamentos transmitidos no decorrer desses anos de IME USP.

Aos meus amigos Paula, Cátia, Catinha, Marcos, Marcelo, Danillo, Regina e Ayumi que convivem comigo desde os primeiros dias de faculdade, um agradecimento especial. Muito obrigada pela companhia e pelos ótimos momentos vividos.

Aos meus amigos da graduação Roberta, Fabíola, Camila, Danielle, Milena, Eliane e Paula e aos meus amigos do mestrado Carlos Eduardo, Kátia, Márcia e Filipe, agradeço pelo companheirismo dentro e fora da sala de aula. Aos meus amigos de trabalho Cintia, Oswaldo e Tatiana Miamoto agradeço por nunca me deixarem desistir de realizar este objetivo.

Agradeço, por fim, a Deus, por tudo que tenho e por ter tido a oportunidade de chegar até aqui.

Resumo

Observações multivariadas que são claramente atípicas em uma única componente podem freqüentemente serem detectadas através de técnicas univariadas aplicadas a cada variável. No entanto, para dados multivariados, é necessário avaliar cada variável em relação às demais e alguns pontos podem falhar em manter o padrão da relação entre as variáveis existente na maioria dos dados. Devido ao fato dos procedimentos clássicos serem seriamente influenciados por valores atípicos, os métodos robustos produzem uma abordagem complementar alternativa, uma vez que são menos sensíveis à presença de valores atípicos.

Neste trabalho são apresentados métodos de diagnóstico robusto em Análise Multivariada e em Análise de Regressão. O Diagnóstico Robusto em Análise de Regressão foi aplicado em um conjunto de dados reais e realizou-se a comparação com o correspondente método clássico.

Abstract

Multivariate observations which are grossly atypical in a single component can often be detected by applying univariate techniques to each variable. However, for multivariate data, observations are often only found to be atypical when the value for each variable is considered in relation to the other variables and some values fail to maintain the pattern of relationships between the variables evident in the majority of the observations. Since the performance of classical procedures is seriously influenced by atypical values, robust methods which are little influenced by atypical values provide an attractive complementary approach.

In this work are presented robust diagnostic methods in Multivariate and Regression Analysis. The Robust Diagnostic in Regression Analysis was applied to real data and the comparison to the correspondent classical method was done.

Sumário

Capítulo 1

Introdução

Diagnóstico é a etapa da análise estatística que possibilita a identificação de pontos aberrantes. Tais pontos, por diferirem dos demais em suas medidas, podem alterar de forma significativa as conclusões de uma análise estatística. Métodos de diagnóstico podem ser usados ainda para identificar aspectos que não condizem com as suposições iniciais do modelo, sugerindo então ações reparadoras para a análise adequada do ajuste.

Pontos aberrantes são observações que fogem do padrão da maioria dos dados. Ainda quando os pontos são bidimensionais, é possível detectar os pontos aberrantes visualmente, o que não ocorre para dimensões superiores.

O cálculo da Distância de Mahalanobis (DM) é um método clássico utilizado para classificar se uma observação é ou não aberrante. Através dela é possível quantificar o quanto cada ponto x_i, pertencente a um espaço p-dimensional, está distante do centro do conjunto de dados. É definida como:

$$DM_{(i)} = \sqrt{(x_i - T(X)) \cdot S(X)^{-1} \cdot (x_i - T(X))^t} \tag{1.1}$$

em que $T(X)$ é a média aritmética dos dados, $T(X) = \frac{\sum_{i=1}^{n} x_i}{n}$, $S(X)$ é a matriz de covariância amostral usual e n é o tamanho da amostra.

Sabe-se que este método é prejudicado pelo "efeito de mascaramento", já que pontos aberrantes multivariados não necessariamente possuem altos valores de DM. Isto se deve ao fato das estatísticas $T(X)$ e $S(X)$ não serem robustas, ou seja, um pequeno grupo de pontos aberrantes irão atrair $T(X)$ e inflacionar $S(X)$, provocando baixos valores de DM. Com isso, tais pontos aberrantes não seriam diagnosticados por esta medida. Desta maneira, parece natural substituir $T(X)$ e $S(X)$ em (1.1) por estimadores robustos.

Tal fato é destacado por Rousseeuw e Van Zomeren (1990). De acordo com os autores, a análise clássica de diagnóstico nem sempre detectaria observações influentes por ser baseada em vetores de médias e matrizes de covariância amostrais, quantidades essas também afetadas por tais observações.

Para resolver o problema, os autores sugerem o uso de medidas robustas de diagnóstico. De modo geral, tais medidas são obtidas a partir das usuais, substituindo as estimativas de parâmetros de locação e de covariância por estimativas robustas.

O objetivo desta dissertação é estudar a análise de diagnóstico robusto na Análise Multivariada e principalmente na Análise de Regressão, comparando-as com seus respectivos diagnósticos clássicos.

O Capítulo 2 apresenta a análise clássica de diagnóstico. São descritas de forma sucinta as técnicas clássicas de diagnóstico em Componentes Principais, em Análise Discriminante e na Análise de Regressão.

No Capítulo 3 são apresentadas técnicas robustas de diagnóstico em Análise Multivariada, iniciando o estudo através da discussão de estimadores robustos do tipo M. Em seguida são apresentados procedimentos para análise robusta de Componentes Principais e Análise Discriminante Robusta.

O Capítulo 4 é inteiramente dedicado ao estudo de medidas robustas de diagnóstico em Análise de Regressão, baseado na detecção de *outliers* multivariados e pontos de alavanca.

O Capítulo 5 exibe a aplicação da técnica apresentada no capítulo anterior a um conjunto de dados reais, e se constitui na principal contribuição de nosso trabalho. O Capítulo

6 encerra o trabalho com algumas conclusões.

O Apêndice A contém detalhes técnicos para obtenção do elipsóide de volume mínimo, e o Apêndice B contém a programação em R da aplicação em dados reais descrito no Capítulo 5.

Capítulo 2

Análise Clássica de Diagnóstico

2.1 Introdução

Em grandes bancos de dados, embora os modelos utilizados sejam representações parcimoniosas que podem levar a interpretações simples dos dados, é muito importante que existam medidas da possível falta de adequação e sensibilidade dos modelos sob consideração. A utilização de resíduos para expor qualquer falta de adequação de um modelo ajustado na análise de respostas unidimensionais tem sido vastamente reconhecida.

Uma das utilizações dos resíduos unidimensionais é na detecção de *outliers* ou observações extremas, que não são incomuns em grandes bancos de dados. A modelagem robusta ou estimação robusta é uma abordagem que manipula *outliers* minimizando a influência destes no modelo ajustado. Entretanto, localizar um *outlier* com precisão para investigação profunda pode ser de grande valia na análise estatística e procedimentos direcionados especificamente na detecção desses *outliers* podem ser úteis.

Em um particular conjunto de dados com respostas múltiplas, existe à princípio um vetor de resíduos multidimensionais entre os dados observados e os ajustados. Mais do que no caso unidimensional, uma importante questão surge: como expressar esses resíduos multidimensionais? Embora os estudos sejam menos comuns do que no caso univariado,

muitos aspectos podem ser considerados e a discussão nesta seção será concentrada em alguns métodos estatísticos para analisar resíduos multidimensionais.

Existem duas amplas categorias de análise estatística em problemas com respostas múltiplas: (1) a análise da estrutura interna e (2) a análise da sobreposição ou estrutura extra. A primeira categoria inclui técnicas tais como Componentes Principais, Análise Fatorial e Escalonamento Multidimensional, que são úteis no estudo das dependências internas e na redução da dimensionalidade da resposta. Regressão Múltipla Multivariada e Análise de Variância Multivariada são técnicas clássicas para investigação e especificação da dependência das observações multi-resposta e são exemplos da segunda categoria.

Cada categoria de análise gera resíduos multivariados. Por exemplo, a Análise de Componentes Principais Linear pode ser vista como um ajuste de hiperplanos mutuamente ortogonais através da minimização da soma de quadrados dos resíduos ortogonais das observações em relação a cada plano. Portanto, em qualquer estágio, existem resíduos que são desvios perpendiculares dos dados em relação ao hiperplano ajustado. Em contrapartida, ao analisarmos a estrutura de sobreposição (a segunda categoria citada) por Regressão Múltipla Multivariada, esta possui a reconhecida Soma Mínima dos Quadrados do Resíduo.

2.2 Análise de Diagnóstico Clássica em Componentes Principais

A idéia básica da Análise de Componentes Principais é descrever a dispersão de n vetores associados a p variáveis construindo um novo grupo de p combinações lineares ortogonais entre si de maneira que suas variâncias estejam em ordem decrescente de magnitude, sendo que a primeira componente principal é aquela combinação linear que produz variância máxima. A segunda componente principal é aquela que possui a segunda maior variância, sendo ortogonal à primeira, e assim por diante.

Se os elementos de $\mathbf{y}' = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_p})$ denotam as p variáveis originais e as linhas

da matriz $\mathbf{Y_{n\times p}}$ constituem as n observações p-dimensionais, o vetor de médias amostrais e a matriz de covariância podem ser obtidos respectivamente como:

$$\bar{\mathbf{y}}' = (\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, ..., \bar{\mathbf{y}}_\mathbf{p}) = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}'\mathbf{Y}$$

$$\mathbf{S} = (\mathbf{s_{ij}}) = \frac{\mathbf{1}}{\mathbf{n-1}}(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})$$

sendo que $\mathbf{1}'$ é um vetor linha cujos elementos são iguais a 1 e $\bar{Y}$ é uma matriz cujas linhas são iguais a $\bar{\mathbf{y}}'$. A matriz de correlação $\mathbf{R_{pxp}}$ é relacionada a $\mathbf{S}$ por:

$$\mathbf{R} = \mathbf{D}_{\mathbf{1}/\sqrt{\mathbf{s_{ii}}}} \cdot \mathbf{S} \cdot \mathbf{D}_{\mathbf{1}/\sqrt{\mathbf{s_{ii}}}}$$

em que $D_{1/\sqrt{s_{ii}}}$ é uma matriz diagonal pxp cujo i-ésimo elemento da diagonal é $1/\sqrt{s_{ii}}$, para $i = 1, 2, \ldots, p$.

De acordo com Gnanadesikan (1997), a interpretação geométrica da Análise de Componentes Principais é a seguinte: a inversa da matriz de covariância amostral pode ser empregada como a matriz de uma forma quadrática que define uma família de elipsóides centrados no centro amostral de gravidade dos dados, cuja equação é:

$$(\mathbf{y} - \bar{\mathbf{y}})' \cdot \mathbf{S}^{-\mathbf{1}} \cdot (\mathbf{y} - \bar{\mathbf{y}}) = \mathbf{c}. \tag{2.1}$$

Para valores não negativos de c, a expressão (2.1) define uma família de elipsóides no espaço p-dimensional de $\mathbf{y}$. A transformação dos dados via componentes principais fornece a projeção das observações nos eixos principais dos elipsóides desta família. A idéia básica é ilustrada para o caso bidimensional na Figura 2.1.

As coordenadas originais (y_1, y_2) são deslocadas da origem para a média amostral $(\bar{y}_1, \bar{y}_2)$ seguida de uma rotação em torno desta origem, produzindo as coordenadas das componentes principais, z_1 e z_2.

Algebricamente, a Análise de Componentes Principais consiste em obter combinações

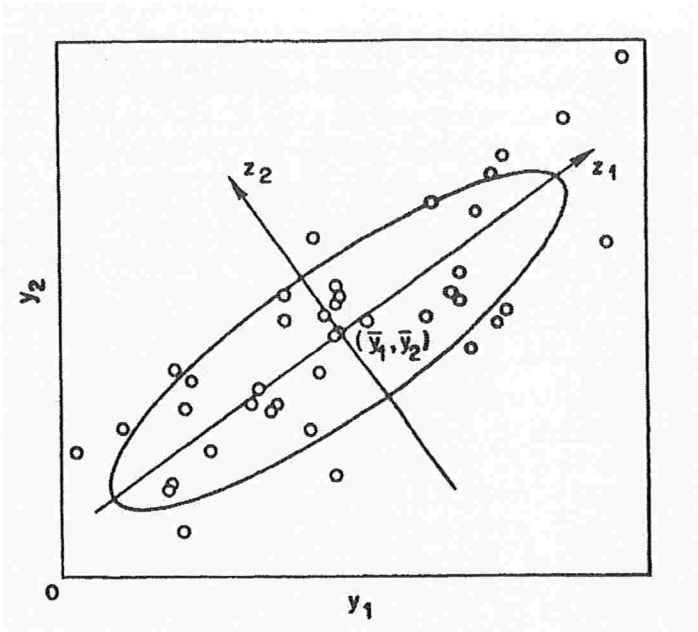

Figura 2.1: Ilustração de componentes principais em dados bivariados.
Fonte: Gnanadesikan (1997).

lineares das variáveis originais a partir dos autovalores e autovetores da matriz de covariância amostral. Mais especificamente, para obter a primeira componente principal z_1, utiliza-se o vetor de coeficientes $\mathbf{a}' = (\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_p})$ de maneira que a combinação linear das componentes de $\mathbf{y}$, $\mathbf{a}'\mathbf{y}$, possua variância máxima na classe de todas as combinações lineares, sujeita à restrição de normalização $\mathbf{a}'\mathbf{a} = \mathbf{1}$. Para um dado vetor $\mathbf{a}$, como a variância amostral de $\mathbf{a}'\mathbf{y}$ é igual a $\mathbf{a}'\mathbf{S}\mathbf{a}$, o problema de encontrar $\mathbf{a}$ torna-se equivalente ao de determinar um vetor não nulo cuja razão $\mathbf{a}'\mathbf{S}\mathbf{a}/\mathbf{a}'\mathbf{a}$ seja máxima. Sabe-se que o valor máximo resultado desta razão é o maior autovalor, c_1, de S, e a solução procurada para $\mathbf{a}$ é o autovetor $\mathbf{a_1}$ de S correspondente ao autovalor c_1.

Depois que a primeira componente principal é determinada, o próximo passo é determinar a segunda combinação linear normalizada ortogonal à primeira tal que, na classe das funções lineares normalizadas de $\mathbf{y}$ ortogonais a $\mathbf{a}_1'\mathbf{y}$, a segunda componente principal tenha maior variância. No próximo estágio, o objetivo é determinar a terceira combinação linear normalizada ortogonal às duas primeiras componentes principais, até que as p componentes

principais tenham sido determinadas. O problema de determinar as p componentes principais é equivalente a determinar os valores estacionários da razão $\mathbf{a'Sa/a'a}$ para variação de todos os vetores não nulos a. Estes valores estacionários são conhecidos como autovalores ($c_1 \geq c_2 \geq \cdots \geq c_p \geq 0$) de S, e as procuradas componentes principais são oriundos de $\mathbf{a'_1y}$, $\mathbf{a'_2y}$, ... e $\mathbf{a'_py}$ onde $\mathbf{a'_iy}$ é o autovetor normalizado de S correspondente ao autovalor c_i, $i = 1, 2, \ldots, p$. Os autovalores ordenados correspondem às variâncias amostrais das combinações lineares das variáveis originais.

Alternativamente, a definição secundária de Análise de Componentes Principais, que é baseada na decomposição espectral da matriz S.

Decomposição Espectral da matriz S: existe uma matriz ortogonal A tal que $S = A \cdot D_c \cdot A'$, onde D_c é uma matriz diagonal com elementos $c_1, c_2, \ldots, c_p$, autovalores de S, na diagonal principal. As colunas de A são os autovetores $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_p}$. As coordenadas das componentes principais, que por conveniência são definidas levando em consideração desvios com relação à média amostral, são então especificadas pela transformação: $z = A' \cdot (\mathbf{y} - \bar{\mathbf{y}})$ e a transformação dos dados via componentes principais é dada por:

$$\mathbf{Z} = \mathbf{A'} \cdot (\mathbf{y} - \bar{\mathbf{y}}). \tag{2.2}$$

A equação (2.2) define uma transformação linear de componentes principais dos dados em termos dos autovetores da matriz de covariância amostral S. Cada linha $\mathbf{a'_j}$, $j = 1, 2, \ldots, p$ de A' fornece uma coordenada da componente principal e cada linha de Z dá os desvios das projeções da amostra original em relação à projeção do centróide amostral $\bar{y}$, em uma coordenada específica.

Quando a Análise de Componentes Principais é vista como um método de ajuste de subespaços lineares ou como uma técnica estatística para detecção e descrição de possíveis singularidades lineares nos dados, o interesse reside especialmente nas projeções dos dados sob as coordenadas das componentes principais correspondentes aos menores autovalores

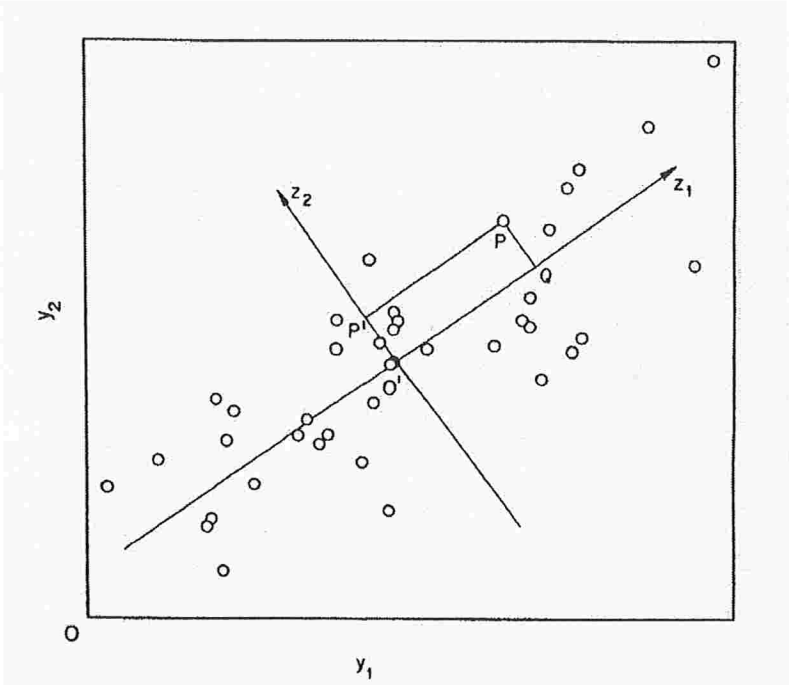

Figura 2.2: Ilustração de resíduos de componentes principais em dados bivariados.
Fonte: Gnanadesikan (1997).

(isto é, às últimas linhas de Z). Então, por exemplo, para $p = 2$, os conceitos essenciais são ilustrados na Figura 2.2, onde y_1 e y_2 denotam as coordenadas originais e z_1 e z_2 denotam as componentes principais oriundas da matriz de covariância dos dados bivariados. A reta de maior proximidade aos dados (avaliada através da soma dos quadrados dos desvios perpendiculares à reta) é o eixo z_1. O resíduo ortogonal de um dado ponto P, como é ilustrado na Figura 2.2, é o vetor $\overline{QP}$, que é equivalente ao vetor $\overline{O'P'}$, onde P' é a projeção de P sobre o eixo z_2, a segunda componente principal. Mais genericamente, com dados p-dimensionais, a projeção sobre a p-ésima componente principal (isto é, aquela com a menor variância) será relevante para estudar o desvio de uma observação em relação ao hiperplano de melhor ajuste. Por outro lado, projeções sobre as q últimas componentes principais serão relevantes para estudar o desvio de uma observação em relação a um subespaço linear ajustado de dimensão $(p - q)$.

Para detectar falta de ajuste de observações individuais, um método sugerido por

Rao (1964) consiste em estudar a soma dos quadrados dos comprimentos das projeções das observações nas q últimas componentes principais. Para cada observação inicial Y_i, $i = (1, \ldots, n)$ o procedimento consiste em calcular:

$$d_i^2 = \sum_{j=p-q+1}^{p} [a_j'(y_i - \bar{y})]^2 = (y_i - \bar{y})'(y_i - \bar{y}) - \sum_{j=1}^{p-q} [a_j'(y_i - \bar{y})]^2$$

e considerar grandes valores de d_i^2 como indicativo de um mau ajuste $(p-q)$-dimensional da i-ésima observação (ou equivalentemente, que a observação é possivelmente um *outlier*).

De acordo com Gnanadesikan (1997), uma técnica gráfica informal utilizada como ferramenta para visualização de outras peculiaridades dos dados seria construir um gráfico de probabilidades da distribuição Gama para os valores $d_i^{2'}$s, utilizando uma estimativa adequada do parâmetro desta distribuição.

Além de calcular a estatística d_i^2 citada, pode-se também estudar as projeções dos dados nas últimas componentes principais de outras formas. Algumas possibilidades são:

1. Gráficos de dispersão bi e tridimensionais dos subconjuntos bivariados e trivariados formados pelas últimas linhas de Z, rotulando os pontos de forma conveniente, por exemplo, em função do tempo ou de algum outro fator de interesse;

2. Gráficos de probabilidades dos valores de cada uma das últimas linhas de Z. Devido à linearidade da transformação envolvida, pode ser razoável que estes valores sejam distribuídos mais próximos à distribuição normal do que os dados originais e o gráfico de probabilidades normal irá fornecer um ponto de partida para a análise. Esta análise seria útil para precisar as coordenadas das últimas componentes principais nas quais a projeção de uma observação pode parecer anormal e fornecer informações adicionais aos gráficos de probabilidades Gama da análise dos $d_i^{2'}$s citada anteriormente;

3. Gráficos dos valores de cada uma das últimas linhas de $\mathbf{Z}$ contra determinadas distâncias no espaço das primeiras componentes principais. Se, por exemplo, a maior parte da

variabilidade de um conjunto de dados penta-dimensional estiver associada às duas primeiras componentes principais, pode ser informativo construir os gráficos das projeções em cada um dos três eixos de componentes principais restantes contra a distância de cada um dos pontos projetados em relação ao centróide no plano bi-dimensional associado às duas primeiras componentes. O ajuste multidimensional é considerado inadequado se a magnitude dos resíduos nas coordenadas associadas aos menores autovalores estiver relacionada com a dos pontos no espaço bidimensional correspondente às duas primeiras componentes principais.

Uma importante questão relacionada à análise sugerida anteriormente é sobre sua robustez. Claramente, se uma observação aberrante é detectada, pode ser natural excluí-la da estimativa inicial de $\mathbf{S}$ (ou $\mathbf{R}$) e então repetir o processo de obtenção e análise dos resíduos da Análise de Componentes Principais. Em algumas circunstâncias pode-se decidir por utilizar uma estimativa robusta da matriz de covariância (ou correlação) inclusive para a análise inicial, na expectativa de que pontos aberrantes tornem-se mais visíveis nas análises de resíduos subseqüentes. Esse assunto será discutido nos próximos capítulos.

2.3 Métodos Clássicos de Diagnóstico em Análise Discriminante

A determinação de pontos influentes em análise discriminante ainda é pouco explorada apesar da vasta literatura existente. Reigada e Elian (2006) apresentam um levantamento das principais medidas de diagnóstico no assunto.

Nesta seção, serão descritas algumas medidas de diagnóstico em análise discriminante com duas populações e matriz de covariância constante. Tais medidas indicam pontos que podem influenciar tanto na regra de alocação quanto na probabilidade de classificação incorreta, elementos básicos da análise discriminante.

Função de Influência em Análise Discriminante

Uma estratégia comum na determinação de pontos influentes na análise de diagnóstico é calcular as estimativas dos parâmetros de interesse com e sem o ponto suspeito e comparar as estimativas obtidas. Nessa linha de estudo, Campbell (1978), citado em Reigada e Elian (2006), sugeriu o uso da função de influência.

O objetivo principal da função de influência teórica é calcular a influência de um particular ponto x no parâmetro de interesse, de modo que valores altos para essa função indicariam que x tem grande influência no parâmetro. Em análise discriminante, a função de influência é determinada excluindo uma observação de apenas um grupo.

O autor considera o caso em que o vetor aleatório $\mathbf{X}$ tem distribuição normal p-variada com vetor de médias μ e matriz de covariância Σ, para $\mathbf{X}$ pertencendo à população π_j, $j = 1, 2$, e utiliza a função de influência definida a seguir. Dadas g populações com funções de distribuição $F_1, F_2, \ldots, F_k, \ldots, F_g$, seja $\Theta = T(F_1, F_2, \ldots, F_k, \ldots, F_g)$ um parâmetro geral, obtido a partir das funções de distribuição F_k, $k = 1, \ldots, g$. Para um valor x do vetor aleatório $\mathbf{X}$, a função de influência $I(x; \theta)$ é definida como:

$I(x;\theta) = \lim_{\varepsilon \to 0} \left(\frac{\widetilde{\theta} - \theta}{\varepsilon} \right)$, em que $0 < \varepsilon < 1$, $\widetilde{\theta} = T(\widetilde{F})$, $\widetilde{F} = (F_1, \ldots, \widetilde{F}_k, \ldots, F_g)$ e $\widetilde{F}_k = (1 - \varepsilon) \cdot F_k + \varepsilon \cdot \delta_x$, sendo δ_x uma função de distribuição que assume probabilidade 1 no ponto x.

No caso, $g = 2$ e consideram-se inicialmente perturbações em Σ^{-1} e μ_j, sendo que $\Sigma = w_1 \Sigma_{F_1} + w_2 \Sigma_{F_2}$; Σ_{F_1} e Σ_{F_2} são as matrizes de covariância para as populações π_1 e π_2 com $w_1 + w_2 = 1$ e $w_j > 0$. Usualmente adotam-se os pesos $w_1 = \frac{n_1}{n_1 + n_2}$ e $w_2 = \frac{n_2}{n_1 + n_2}$.

Tomando a função linear discriminante $y = l'(x) = \delta' \Sigma^{-1} \mathbf{x}$, em que $\delta' = \boldsymbol{\mu_1 - \mu_2}$ e a distância de Mahalanobis $\Delta^2 = (\boldsymbol{\mu_1 - \mu_2})' \boldsymbol{\Sigma^{-1}} (\boldsymbol{\mu_1 - \mu_2})$, calculam-se os parâmetros após a exclusão de um elemento da população π_1 (Reigada (2005)) e usando a notação * sobrescrito para os parâmetros após perturbação tem-se:

$\boldsymbol{\mu_1^*} = \boldsymbol{\mu_1} + \varepsilon(\mathbf{x} - \boldsymbol{\mu_1}) = \boldsymbol{\mu_1} + \varepsilon \mathbf{z}$, com $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu_1}$

$\delta^* = \delta + \varepsilon \mathbf{z}$; $\Sigma^* = (1 - \varepsilon w_1)\Sigma + \varepsilon w_1 \mathbf{z}\mathbf{z}'$ e

$\Sigma^{-1*} \equiv (1 + \varepsilon w_1)\Sigma^{-1} - \varepsilon w_1 \Sigma^{-1} \mathbf{z}\mathbf{z}' \boldsymbol{\Sigma^{-1}}$.

Para construir a função de influência para a distância de Mahalanobis entre as médias amostrais: $\Delta^2 = \delta'\Sigma^{-1}\delta$, após cálculos, obtém-se $\Delta^{2*} \equiv (1+\varepsilon w_1)\Delta^2 + 2\varepsilon\phi - \varepsilon w_1\phi^2$ em que $\phi = \delta'\Sigma^{-1}\mathbf{z}$ e conseqüentemente a função de influência $I(x,\Delta^2) = w_1\Delta^2 + 2\phi - w_1\phi^2$.

Se x pertence à população π_1 que foi perturbada, tem-se $E(\phi) = E(\delta'\Sigma^{-1}\mathbf{z}) = \mathbf{0}$ e $Var(\phi) = \Delta^2$. Portanto $\phi \sim N(0,\Delta^2)$ e $\phi_p = \frac{\phi}{\Delta}$ tem distribuição normal padrão.

Definindo $\phi = \phi_p \cdot \Delta$, a função de influência para Δ^2 pode ser escrita em termos de ϕ_p como:

$$I_p(x,\Delta^2) = w_1\Delta^2 + 2\Delta\phi_p - w_1\Delta^2\phi_p^2.$$

Observa-se assim que a função de influência obtida é uma transformação do vetor aleatório $\mathbf{X}$ e por isso pode ser considerada como uma variável aleatória. Além disso, se $\mathbf{X}$ tem distribuição normal multivariada, verifica-se na página 144 de Reigada (2005) que a distribuição de probabilidades da função de influência para a distância de Mahalanobis (Δ^2) tem assimetria negativa.

Tomando a função de influência para a distância de Mahalanobis com ϕ não padronizado, $I(x,\Delta^2) = w_1\Delta^2 + 2\phi - w_1\phi^2$ e derivando em relação a ϕ, obtém-se o ponto de máximo para $\phi = \frac{1}{w_1}$. Assim, $I_{max}(x,\Delta^2) = w_1\Delta^2 + \frac{1}{w_1}$.

Campbell (1978) sugere o uso de $I_m(x,\Delta^2) = I_{max}(x,\Delta^2) - I(x,\Delta^2)$, dada por $I_m(x,\Delta^2) = w_1^{-1}(1 - 2w_1\Delta\phi_p + w_1^2\Delta^2\phi_p^2)$. Esta variável aleatória é sempre positiva e como $I(x,\Delta^2)$ tem assimetria negativa, segue que $I_m(x,\Delta^2)$ tem distribuição positivamente assimétrica. Multiplicando e dividindo $I_m(x,\Delta^2)$ por ϕ_p^2 obtém-se:

$$I_m(x,\Delta^2) = w_1\Delta^2(\phi_p - w_1^{-1}\Delta^{-1})^2.$$

Como $\phi \sim N(0,1)$, então $(\phi_p - w_1^{-1}\Delta^{-1}) \sim N_\phi(-w_1^{-1}\Delta^{-1},1)$ e dessa forma, $(\phi_p - w_1^{-1}\Delta^{-1})^2 = (\phi_p - w_1^{-1}\Delta^{-1})'1^{-1}(\phi_p - w_1^{-1}\Delta^{-1})$ tem distribuição qui-quadrado não central com 1 grau de liberdade e parâmetro de não centralidade $(w_1^{-1}\Delta^{-1})^2$.

Portanto, $\frac{I_m(x,\Delta^2)}{w_1\Delta^2}$ tem distribuição qui-quadrado não central com um grau de liber-

dade e parâmetro de não centralidade $(w_1^2\Delta^2)^{-1}$. Observa-se então que a distribuição nula de $I_m(x, \Delta^2)$ é qui-quadrado com um grau de liberdade, sendo assim a distribuição esperada na inexistência de pontos discrepantes.

Principais Estatísticas no Diagnóstico em Análise Discriminante

Fung (1992), citado em Reigada e Elian (2006), apresenta duas estatísticas fundamentais no diagnóstico em Análise Discriminante. Uma delas é a medida ϕ, apresentada anteriormente.

Adotando-se a regra discriminante de Fisher em que uma observação x é alocada na população π_1 se $(\bar{x}_1 - \bar{x}_2)'S^{-1}x \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'s^{-1}\bar{x}_1 + \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'s^{-1}\bar{x}_2$, que é equivalente a $\widehat{\phi} \geq -\frac{1}{2}D^2$, em que $\widehat{\phi} = (\bar{x}_1 - \bar{x}_2)'S^{-1}(x - \bar{x}_1)$ e D é a estimativa para a distância de Mahalanobis entre as médias amostrais. Desta forma, observa-se que a medida $\widehat{\phi}$ é importante tanto na discriminação quanto na análise de influência.

A segunda medida apontada pelo autor é $d_{ji}^2 = (x_{ji} - \bar{x}_j)'S^{-1}(x_{ji} - \bar{x}_j)$, $i = 1, \ldots, n_j, j = 1, 2$.

Assim como o resíduo e o ponto de alavanca em regressão, Fung (1992) considera d_{ji}^2 e $\widehat{\phi}_{ji} = \widehat{l}'(x_{ji} - \bar{x}_j)$ como estatísticas básicas para detectar *outliers* e observações influentes em análise discriminante, uma vez que várias medidas de diagnóstico são funções dessas duas estatísticas.

Por facilidade de notação, serão usados os termos d_i^2 e ϕ_i a partir deste ponto.

Fung (1995), citado em Reigada e Elian (2006), mostra que, assintoticamente, d_i^2 tem distribuição χ_p^2 (onde p é a dimensão do vetor $\mathbf{X}$) e $\frac{\widehat{\phi}_i}{D}$ tem distribuição $N(0, 1)$. Tais resultados são fundamentais, pois fornecem um critério de detecção de pontos influentes. Assim, calculando-se estas estatísticas para os elementos amostrais, valores maiores que os respectivos valores críticos da distribuição qui-quadrado e da distribuição normal são indicadores de pontos influentes.

Medidas relacionadas com a probabilidade de classificação incorreta

Em análise discriminante com duas populações, a probabilidade de classificação

incorreta, denotada por PCI, é definida como:

$$PCI = P(1/2) \cdot (1 - q) + P(2/1) \cdot q,$$

sendo que

$P(i/j)$ é a probabilidade de classificar uma observação em π_i quando ela é de π_j;

q é a probabilidade a priori de π_1 e

$(1 - q)$ é a probabilidade a priori de π_2.

Se as probabilidades a priori são iguais, $PCI = \frac{1}{2}P(1/2) + \frac{1}{2}P(2/1)$.

Utilizando o fato que $PCI = \Phi\left(-(\frac{\Delta}{2})\right)$, na qual $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão, Fung (1992) propõe uma medida de diagnóstico com base na diferença das estimativas das probabilidades de classificação incorreta. Tal medida é definida como:

$$DPCI_i = \left[\Phi\left(-\frac{D_{(i)}}{2}\right)\right] - \left[\Phi\left(-\frac{D}{2}\right)\right]$$

em que $D_{(i)}$ é a estimativa da distância de Mahalanobis sem a i-ésima observação do grupo 1. Após cálculos, obtém-se:

$$DPCI_i \equiv \frac{\Phi\left(-\frac{1}{2}D\right)}{4D(n_1 - 1)^2} \cdot \left[(1 - w_1\widehat{\phi}_1)^2\left(d_i^2 - \frac{\widehat{\phi}_i^2}{D^2}\right) + \frac{1}{4}\widehat{\phi}_i^2\right],$$

que depende dos valores de $\widehat{\phi}_i$ e $\widehat{d}_i^2$.

Medidas de diagnóstico na análise discriminante quadrática ou para o caso de g grupos podem ser encontrados em Reigada e Elian (2006).

2.4 Análise Clássica de Diagnóstico em Regressão

A análise de diagnóstico em regressão é mais conhecida na literatura do que as análises descritas nas seções anteriores. Por uniformidade, descreveremos brevemente seus elementos principais.

Os métodos para obtenção de estimativas, testes e outras estatísticas desenvolvidas até hoje contam apenas parte da história da análise de regressão. Estes métodos são utilizados como se o modelo e suas suposições estivessem corretos (satisfeitos), mas em qualquer problema prático as hipóteses são questionadas.

A análise de diagnóstico é a etapa da análise de regressão na qual verificamos os possíveis afastamentos das suposições iniciais do modelo. É nesta etapa também que são identificadas observações extremas que afetam muito os resultados do ajuste. A preocupação inicial da análise de diagnóstico resume-se em duas questões. Primeiro questionamos como o modelo adotado reflete os dados observados. Se o modelo ajustado não produz resíduos "razoáveis", então algum aspecto do modelo pode ser colocado em dúvida. A segunda questão de interesse refere-se ao efeito de cada observação na estimação e em outros aspectos do cálculo das principais estatísticas descritivas. Em alguns bancos de dados estas estatísticas podem sofrer grandes mudanças quando uma observação é retirada. Estas observações são denominadas pontos influentes.

Resíduos e Pontos de Alavanca

Considere o modelo de regressão linear:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

em que $Var(\varepsilon) = \sigma^2 I$, $\mathbf{X}$ é a matriz $n \times p$ e $\mathbf{Y}$, β e ε são vetores de dimensões $n \times 1$, $p \times 1$ e $n \times 1$ respectivamente, com a suposição adicional de normalidade para o vetor aleatório $\mathbf{Y}$.

Nestas condições, o estimador de máxima verossimilhança de β, que coincide com o de Mínimos Quadrados, é dado por:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

em que $Var(\widehat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

A relação entre $\widehat{\mathbf{Y}}$ e $\mathbf{Y}$ é dada por:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

sendo que $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ é denominada matriz de projeção ou matriz chapéu pois transforma o vetor $\mathbf{Y}$ no vetor de respostas ajustadas $\widehat{\mathbf{Y}}$ (Y chapéu). Esta matriz é simétrica e idempotente ($\mathbf{H^2} = \mathbf{H}$).

O vetor de resíduos é definido por:

$$e = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = [\mathbf{I} - \mathbf{H}]\mathbf{Y}.$$

Diferença entre ε e e: os erros ε são variáveis aleatórias não observáveis supostamente independentes com média zero e variância comum σ^2. Os resíduos e são quantidades que podem ser calculadas, possuem média zero e sua matriz de covariância é $(I - \mathbf{H})\mathbf{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H})'$, que será igual a $\sigma^2(I - \mathbf{H})$, pois $\mathbf{H}$ é idempotente ($\mathbf{H^2} = \mathbf{H}$). Como $(I - \mathbf{H})$ não é uma matriz diagonal, os resíduos são variáveis aleatórias correlacionadas.

Para o i-ésimo resíduo teremos:

$$Var(e_i) = \sigma^2(1 - h_{ii}),$$

sendo que h_{ii} é o i-ésimo elemento da diagonal de $\mathbf{H}$ ($h_{ii} < 1$).

Procedimentos de diagnóstico são baseados nos "resíduos computados", os quais gostaríamos que tivessem comportamento igual ao dos resíduos não-mensuráveis. O cumprimento dessa suposição depende da matriz chapéu, já que $\mathbf{H}$ relaciona e a ε e também é considerado no cálculo da variância e covariância de e ($Cov(e_i, e_j) = -\sigma^2 h_{ij}, i \neq j$).

O (i, j)-ésimo elemento de $\mathbf{H}$, denotado por h_{ij} é dado por:

$$h_{ij} = x_i'(X'X)^{-1}x_j = x_j'(X'X)^{-1}x_i = h_{ji}.$$

Os elementos da diagonal de $\mathbf{H}$ são dados por $h_{ii} = x_i'(X'X)^{-1}x_i$.

Como $\widehat{\mathbf{Y}} = \mathbf{HY}$, então

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j \tag{2.3}$$

de modo que h_{ij} mede a influência de y_j em $\widehat{y}_i$ e h_{ii} mede a influência de y_i em $\widehat{y}_i$.

À medida que h_{ii} se aproxima de 1, $\widehat{y}_i$ se aproxima de y_i. Por essa razão, denominamos h_{ii} como sendo a "alavancagem" do elemento i. Para valores altos de h_{ii} na expressão (2.3) predomina a influência de y_i sobre o valor ajustado $\widehat{y}_i$, por isso utilizamos h_{ii} como uma medida de influência da i-ésima observação sobre o próprio valor ajustado. O efeito do i-ésimo elemento amostral no ajuste do modelo de regressão é provavelmente maior se h_{ii} é alto, mas sua importância é incerta, pois depende dos y_j.

Portanto, o elemento $h_{ii} = x_i'(X'X)^{-1}x_i$ desempenha um papel importante na construção de técnicas de diagnóstico. Pontos com h_{ii} alto (próximo de 1) são chamados pontos de alavancagem por possuírem peso desproporcional no próprio valor ajustado; geralmente destoam dos demais em alguma coordenada, trazendo o plano de regressão em sua direção, o que pode ou não distorcer algumas estimativas.

Em resumo, a análise de $\mathbf{H}$ (que não depende de y) pode revelar pontos sensíveis nos quais um valor discrepante em y pode ter um grande impacto no ajuste.

Medidas de Influência

Pontos influentes são aqueles que, quando eliminados, resultam em profundas alterações na análise dos dados. O método de diagnóstico mais utilizado em regressão retira os casos um de cada vez e estuda a influência de cada caso comparando a análise dos dados completos com a análise dos dados sem o caso removido.

Seja $\widehat{\beta}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}$, em que

$\widehat{\beta}_{(i)}$ é o estimador de β calculado sem a observação i;

$\mathbf{X}_{(i)}$ é a matriz $(n-1) \times p$ obtida excluindo-se a linha i de $\mathbf{X}$.

A influência do ponto i pode ser avaliada através da diferença entre as medidas $\widehat{\beta}$ e $\widehat{\beta}_{(i)}$, por exemplo, através da Distância de Cook:

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\beta}_{(\mathrm{i})} - \widehat{\beta})}{p\widehat{\sigma}^2} \tag{2.4}$$

em que $\widehat{\sigma}^2$ trata-se do Quadrado médio do resíduo.

Esta estatística possui diversas propriedades desejáveis. Primeiramente, o contorno em que D_i é constante é um elipsóide. Tais contornos podem ser pensados como definindo a distância entre $\widehat{\beta}_{(i)}$ e $\widehat{\beta}$. Adicionalmente, D_i não depende da parametrização, assim, se as colunas de $\mathbf{X}$ forem modificadas por transformações lineares, a estatística D_i não será influenciada. Finalmente, se definirmos $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}$, então a expressão (2.4) pode ser reescrita como:

$$D_i = \frac{(\widehat{Y}_{(i)} - \widehat{Y})'(\widehat{Y}_{(i)} - \widehat{Y})}{p\widehat{\sigma}^2}.$$

Assim, D_i será a distância euclidiana entre $\widehat{\mathbf{Y}}$ e $\widehat{\mathbf{Y}}_{(i)}$. Altos valores de D_i indicam grande influência da i-esima observação em $\widehat{\beta}$ e nos valores ajustados. A exclusão desse ponto pode resultar em importantes alterações nas conclusões do modelo.

Escrevendo D_i em uma forma mais simplificada temos:

$$D_i = \left[\frac{e_i}{\widehat{\sigma}\sqrt{i - h_{ii}}}\right]^2 \frac{h_{ii}}{1 - h_{ii}} \frac{1}{p}.$$

Observa-se que D_i é crescente em e_i e h_{ii}. Se p é fixado, D_i será determinado por duas diferentes fontes: e_i reflete a falta de ajuste no ponto i e h_{ii} reflete a posição de x_i em relação a $\bar{x}$. D_i será alto se $|e_i|$ for alto ou h_{ii} for alto ou ambos.

Porém, a distância D_i pode não ser adequada quando e_i for grande e h_{ii} pequeno. Desta maneira, Belsey, Kuh e Welsch (1980) propuseram uma estatística muito similar à distância D_i, denominada $DFFITS_i$:

$$DFFITS_i = \sqrt{\frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'(X'X)(\widehat{\beta}_{(i)} - \widehat{\beta})}{\widehat{\sigma}^2_{(i)}}}, i = 1, \ldots, n.$$

Esta estatística difere da estatística D_i por um fator de escala e pela substituição de $\widehat{\sigma}^2$ por $\widehat{\sigma}^2_{(i)}$ (Quadrado médio do resíduo sem a i-ésima observação).

Mais recentemente, Cook (1986) introduziu a técnica de Diagnóstico de Influência Local, que, ao contrário das anteriores, não é baseada na retirada de pontos do conjunto de dados.

A literatura de diagnóstico em regressão é bastante extensa e tem sido amplamente utilizada. Neste capítulo foram apresentadas medidas de diagnóstico clássico em análise multivariada e em análise de regressão. No próximo capítulo iremos nos dedicar ao estudo de técnicas robustas de diagnóstico em Análise Multivariada.

Capítulo 3

Técnicas Robustas de Diagnóstico em Análise Multivariada

3.1 Introdução

A limitação da metodologia de estimação via Mínimos Quadrados com relação a pontos discrepantes conduziu a diversas abordagens alternativas. As técnicas de estimação robusta se constituem em uma abordagem que não depende da distribuição dos dados. Embora os estimadores de mínimos quadrados não dependam da distribuição de probabilidades da variável resposta, sua alta sensibilidade a valores discrepantes, provenientes de distribuições com caudas mais pesadas que a normal, é conhecida. O objetivo da estimação robusta é o de encontrar estimadores eficientes sob determinado modelo, de modo que pequenas alterações na distribuição dos dados produzam pequenas alterações nas estimativas.

A detecção de observações atípicas em um conjunto de dados multivariado pode ser determinada através de gráficos de probabilidade do quadrado da distância de Mahalanobis. Porém, ao invés dos usuais estimadores de máxima verossimilhança, Campbell (1980) sugere a utilização de estimadores M, que são estimadores robustos de médias e covariâncias. Já na ausência de observações discrepantes, os estimadores robustos comportam-se de forma

similar aos estimadores usuais.

Observações multivariadas que são claramente atípicas em uma única componente podem freqüentemente serem detectadas por meio de técnicas univariadas aplicadas a cada variável. No entanto, para dados multivariados, é necessário avaliar cada variável em relação às demais e alguns pontos podem falhar em manter o padrão da relação entre as variáveis existente na maioria dos dados. Devido ao fato dos procedimentos clássicos serem seriamente influenciados por valores atípicos, os métodos robustos produzem uma abordagem complementar alternativa, já que são menos sensíveis à presença de valores atípicos.

Neste capítulo serão discutidas medidas robustas de diagnóstico em análise multivariada. Serão obtidos estimadores de vetores de médias e de matrizes de covariância para um único grupo, pouco influenciados por observações atípicas e analisadas técnicas de detecção de tais observações. Nesse contexto, admite-se que os dados possuem distribuição normal multivariada (originalmente ou após alguma transformação). Na Seção 3.2, serão discutidos os estimadores robustos do tipo M para vetores de médias e matrizes de covariância e sua utilização em conjunto com gráficos de probabilidade do quadrado da distância de Mahalanobis. Procedimentos para análise robusta de componentes principais são propostos na Seção 3.3, e na Seção 3.4, discute-se a Análise Discriminante Robusta.

3.2 Estimação Robusta Multivariada

Consideremos a situação em que o vetor aleatório $\mathbf{X}$ tem uma distribuição normal p-variada com vetor de médias μ e matriz de covariância Σ; $T(X)$ é o vetor $p \times 1$ de médias amostrais e $S(X)$ a matriz de covariância amostral. Nessas condições, o quadrado da distância de Mahalanobis da i-ésima observação x_i com relação à média amostral é definido por:

$DM_i^2 = (x_i - T(X))^t S(X)^{-1}(x_i - T(X))$, $i = 1, \ldots, n$, sendo n o tamanho da amostra.

Healy (1968) e Cox (1968), citados em Campbell (1980), sugeriram uma extensão

dos gráficos de probabilidade de dados univariados para a situação multivariada, construindo o gráfico do quadrado da distância de Mahalanobis de cada observação contra a estatística de ordem de uma distribuição Qui-Quadrado com p graus de liberdade.

O objetivo desta análise é verificar a suposição de normalidade multivariada, além da detecção de observações atípicas. No entanto, vetores de observações multivariadas atípicas tenderão a alterar o vetor de médias, reduzir correlações e possivelmente aumentar variâncias, fazendo com que a distância de Mahalanobis para a observação atípica diminua, obscurecendo sua influência e distorcendo o restante do gráfico.

A distância de Mahalanobis desempenha um importante papel na estimação M multivariada. Do ponto de vista aplicado, os estimadores M podem ser considerados como uma simples modificação dos estimadores clássicos; dão peso total às observações oriundas do conjunto principal do banco de dados, porém peso ou influência reduzida em observações das extremidades da distribuição. Na prática, usualmente, dá-se peso menor às observações com grandes distâncias de Mahalanobis.

Campbell (1980) sugere a utilização dos estimadores M robustos de vetores de médias e matrizes de covariância:

$$\bar{x} = \frac{\sum_{m=1}^{n} w_m x_m}{\sum_{m=1}^{n} w_n},$$

$$V = \frac{\sum_{m=1}^{n} w_m^2 (x_m - \bar{x})(x_m - \bar{x})^t}{\left(\sum_{m=1}^{n} w_m^2 - 1 \right)}, \tag{3.1}$$

em que $w_m = w(d_m) = \omega(d_m)/d_m$ e $d_m = \{(x_m - \bar{x})^t V^{-1} (x_m - \bar{x})\}^{\frac{1}{2}}$ e as soluções para $\bar{x}$ e V são obtidas iterativamente.

No contexto analisado, ω representa uma função de influência limitada, freqüente-

mente linear no conjunto de valores de d_m correspondentes aos dados. Em particular, alguns autores sugerem que a influência e conseqüentemente o peso de uma observação extrema deveria ser nulo, de modo que ω deveria decrescer para valores suficientemente grandes de d_m.

O autor utiliza uma forma bi-paramétrica para ω:

$$\omega(d) = \begin{cases} d, & \text{se } d \leq d_0; \\ d_0 \exp\left\{\frac{-\frac{1}{2}(d-d_0)^2}{b_2^2}\right\}, & \text{se } d > d_0, \end{cases} \tag{3.2}$$

para $d_0 = \sqrt{p} + \frac{b_1}{\sqrt{2}}$.

A motivação da fórmula de d_0 dada na expressão (3.2) é a de que sob hipóteses padrão, por exemplo, de que os dados provêm de uma distribuição normal p-variada, então d^2 tem distribuição Qui-Quadrado com p graus de liberdade ($d_2 \sim \chi_p^2$) e portanto d tem aproximadamente distribuição $Normal(\sqrt{p}, 1/\sqrt{2})$. Assim, b_1 será um quantil da distribuição normal padrão e b_2 é um parâmetro que controla a taxa de decréscimo da função de influência para zero.

Após exaustivos estudos empíricos, Campbell (1980) sugere considerar nas aplicações práticas as três seguintes situações:

(a) $b_1 = \infty$; neste caso, todas as observações possuem peso um, resultando nos estimadores usuais;

(b) $b_1 = 2, b_2 = \infty$; sugerida por Huber (1964), trata-se da forma não-descendente, pois associa valor de ω igual a d_0 a toda observação maior que d_0, qualquer que seja seu valor;

(c) $b_1 = 2, b_2 = 1,25$; uma forma descendente, sugerida por Hampel (1973), de modo que a ponderação decresce a uma velocidade maior do que em (b).

Um posterior estudo de simulação confirmou a eficiência dos valores de b_1 e b_2 propostos.

Se as estimativas robustas forem utilizadas em análises estatísticas subseqüentes, é

importante que elas difiram pouco das estimativas usuais quando aplicadas a dados não contaminados. Com o objetivo de verificar este fato, foram geradas observações de distribuições normais multivariadas, e calculadas as estimativas robustas e usuais das variâncias populacionais. Verificou-se que as estimativas robustas não ultrapassavam 3% das estimativas usuais, situação considerada bastante satisfatória.

Além disso, para distribuição normal multivariada, observou-se que:

$E(d_2) = p$, se $b_1 = \infty$ e

$E(d_2) = kp$, com $k < 1,025$ para $b_1 = 2,0$ e $b_2 = 1,25$.

3.3 Análise Robusta de Componentes Principais

De acordo com Devlin, Gnanadesikan e Kettenring (1981), o uso das matrizes de covariância e correlação como resumos da dispersão e associação é prática comum na análise de dados multivariados. Além disso, tais matrizes freqüentemente constituem a fonte para outras análises. Ainda segundo os autores, a mais clássica e mais utilizada técnica para a análise de dados multivariados é a Análise de Componentes Principais, que é usada para a redução linear da dimensionalidade das observações.

Entretanto, bem como outras técnicas clássicas, os resultados de uma análise de componentes principais podem ser sensíveis a poucas observações discrepantes. Para ilustrar tal fato, um exemplo é retirado de Chen, Gnanadesikan e Kettenring (1974). Os dados são constituídos por observações anuais em 14 variáveis econômicas para 29 companhias químicas e a Figura 3.1 ilustra a projeção dos dados no plano das duas primeiras componentes principais obtidas a partir da matriz de covariância.

O coeficiente de correlação entre as duas primeiras componentes é zero, como deveria ser, porém isto se deve inteiramente a um único ponto no canto inferior direito da Figura 3.1. Com isso, o coeficiente de correlação excluindo esse ponto e calculado com base nos demais 28 pontos é de $0,99$. As componentes principais utilizadas nesta figura são claramente

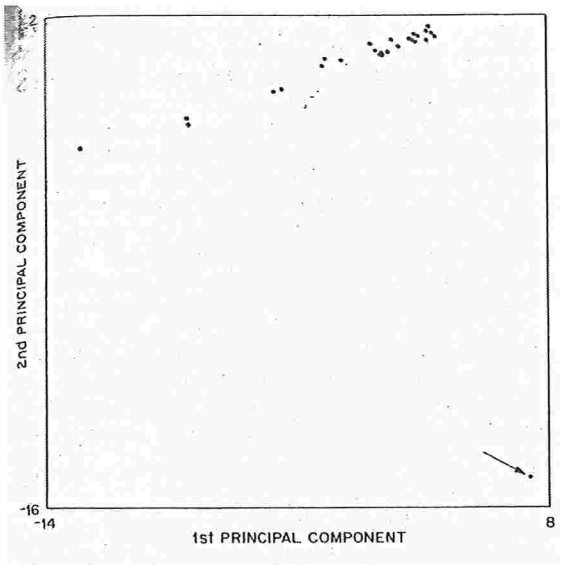

Figura 3.1: Duas primeiras componentes principais do exemplo das companhias químicas.
Fonte: Devlin, Gnanadesikan e Kettenring (1981).

inapropriadas para este conjunto de dados. Certamente um grupo de componentes principais mais apropriado nessas situações pode ser determinado excluindo-se os pontos discrepantes e repetindo-se a análise - freqüentemente uma tarefa difícil em altas dimensões - ou através de uma análise robusta.

De acordo com Ammann (1989), as consequências da presença de *outliers* em dados multivariados são muito mais complexas que no caso univariado, pois além da distorção nas medidas de locação e escala, há ainda uma alteração na estrutura de correlação entre as variáveis. Esta alteração produziria sérios problemas na análise de componentes principais.

Apesar da existência de inúmeros métodos de detecção de *outliers* multivariados, segundo o autor, qualquer ação nesse sentido envolveria a remoção de observações, o que poderia implicar na retirada de pontos importantes. Um outro problema seria que, geralmente, a retirada dessas observações pode ocasionar que outras observações, antes consideradas não problemáticas, tornem-se *outliers*.

Nesse sentido, a Análise Robusta de Componentes Principais mostra-se vantajosa. Segundo o autor, métodos multivariados robustos teriam dois objetivos: a identificação de

outliers multivariados e a possibilidade de uma análise menos sensível a *outliers*.

A Análise de Componentes Principais baseia-se na matriz de covariâncias amostral usual S (ou matriz de correlação R associada) e encontra a combinação linear $y_m = a^t \cdot x_m$ das variáveis originais x_m tal que a variância amostral usual de y_m seja máxima. Conforme apresentado na Seção 2.2, a solução é obtida através de uma decomposição espectral de S, $S = A \cdot D_c \cdot A^t$. Os autovetores a_i obtidos das colunas de A definem combinações lineares, enquanto os elementos da diagonal correspondentes c_i da matriz diagonal de autovalores D_c são as variâncias amostrais das novas variáveis.

Uma forma natural de modificar esta análise é substituir a matriz S pelo estimador robusto definido na expressão (3.1); obtendo a solução via Estimador M para a análise de Componentes Principais Robusta. Uma observação é ponderada de acordo com a distância total DM_i em relação à estimativa robusta de posição.

Por fim, uma estimativa robusta da matriz de covariância ou correlação pode ser encontrada através de $A \cdot D_c \cdot A^t$, como uma estimativa robusta alternativa. Tanto esta abordagem quanto a descrita na seção anterior resultam em matrizes de correlação e covariância positivas definidas. A estimação robusta de cada entrada separadamente nem sempre garante essa propriedade.

Finalizando, uma abordagem prática recomendada pelo autor é determinar as médias, covariâncias, distâncias e pesos associados para $b_1 = \infty$; para $b_1 = 2,0$ e $b_2 = \infty$; e para $b_1 = 2,0$ e $b_2 = 1,25$. Gráficos de probabilidade Normal da raiz cúbica das distâncias ao quadrado (com $b_1 = 2,0$ e $b_2 = 1,25$) irão indicar observações atípicas. Healy (1968) considera este gráfico como sendo o mais adequado na detecção de observações atípicas e para examinar a hipótese de normalidade multivariada. Para mais do que seis ou sete variáveis, uma análise de componentes principais robusta é também útil para identificar observações atípicas. Os pesos w_m^2 associados aos valores de d_m^2 indicam observações atípicas. A experiência mostra que um dado menor do que 0,30 com $b_1 = 2,0$ e $b_2 = 1,25$ (correspondendo aos pesos menores do que 0,60 com $b_2 = \infty$ aproximadamente) sempre indicam uma observação atípica.

Um peso maior do que 0,70 com $b_2 = \infty$ é associado a uma observação razoável.

Ammann (1989) considera que mesmo o procedimento de associar baixos pesos às observações consideradas *outliers* provocaria uma perda de informação. Por esse motivo, sugere uma abordagem Robusta de Componentes Principais baseada no procedimento de Regressão Ortogonal Robusta.

O método de Regressão Ortogonal é aquele que, ao invés de minimizar a soma dos quadrados dos desvios verticais com relação ao plano de regressão, minimiza a soma das distâncias ortogonais. Ammann e Ness (1989) desenvolvem um algoritmo robusto para Regressão Ortogonal, sendo que uma de suas possíveis aplicações é a obtenção de componentes principais robustas.

A aplicação é possível devido à conexão existente entre componentes principais e regressão ortogonal, já que o procedimento de regressão ortogonal fornece o hiperplano ortogonal que passa por $(\bar{x}_1, \cdots, \bar{x}_m, \bar{y})$ e é paralelo às m primeiras componentes principais nas $m + 1$ variáveis.

De forma resumida, os passos básicos do procedimento são:

1. Utilize uma técnica de regressão ortogonal robusta para o conjunto de dados completo. O vetor unitário ortogonal ao hiperplano de regressão obtido será a última componente principal.

2. Projete os dados no hiperplano obtido.

3. Aplique a técnica de regressão ortogonal robusta para os dados projetados obtendo o novo hiperplano de regressão ortogonal e de forma análoga, a próxima componente principal.

4. Continue o processo até que todas as componentes principais, exceto a primeira, sejam obtidas. A primeira componente principal robusta será aquele vetor unitário ortogonal às $p - 1$ componentes já obtidas.

Quanto à regressão ortogonal robusta, necessária na construção do algoritmo que acabamos de descrever, várias técnicas encontram-se descritas na Seção 2 de Ammann e Ness

(1989).

Devlin, Gnanadesikan e Kettenring (1981) utilizam métodos de Monte Carlo para comparar as performances de vários procedimentos robustos para estimação da matriz de correlação e das componentes principais obtidas a partir dela.

A propriedade de que as componentes principais são sensíveis às escalas de medida das variáveis e as dificuldades (incluindo as computacionais) causadas por este motivo são freqüentemente utilizadas como motivo pela escolha da análise de componentes principais utilizando a matriz de correlação ao invés da de covariância.

Definiremos a seguir propriedades gerais de algumas alternativas robustas para a matriz de correlação usual R, apresentadas em Devlin, Gnanadesikan e Kettenring (1981).

Existem muitos métodos que calculam a matriz de correlação robusta, R_*. Uma abordagem básica é estimar cada elemento fora da diagonal de R_* separadamente, através de um coeficiente de correlação robusto r_*. No entanto, a matriz resultante R_* não será necessariamente positiva definida, exceto em casos especiais e esta matriz necessitará de ajustes para atingir esta propriedade. Tal abordagem bivariada de estimadores robustos de correlação obtidos separadamente pode ser considerada contrastante com a abordagem multivariada, que manipula todas as variáveis simultaneamente para obter matrizes de correlação que sejam positivas definidas (se o número de vetores com observações completas é suficientemente grande).

Todos os estimadores robustos de correlação obtidos separadamente serão denotados por r_*, e serão baseados em variáveis padronizadas. Caso não seja especificado anteriormente, a padronização inicial e as subseqüentes estimativas das variâncias são calculadas utilizando médias e variâncias "aparadas" a 5%. Se não existem dados incompletos, então cada elemento r^*_{jk} de R_* será baseado no mesmo número, n, de observações, caso contrário, R_* será obtida utilizando o número de pares completos, n_{jk}, para as variáveis j e k.

Se a matriz R_* obtida não for positiva definida, esta pode ser transformada utilizando-se a expressão (10) de Devlin, Gnanadesikan e Kettenring (1975) com $\Delta = 0,25/\sqrt{n}$, até

que uma matriz positiva definida seja obtida.

A transformação sugerida é:

$$g(r_{jk}^*) = \begin{cases} z^{-1}[z(r_{jk}^*) + \Delta], & \text{se } r_{jk}^* < z^{-1}(\Delta) \\ 0, & \text{se } \mid r_{jk}^* \mid \leq z^{-1}(\Delta) \\ z^{-1}[z(r_{jk}^*) - \Delta], & \text{se } r_{jk}^* > z^{-1}(\Delta) \end{cases}$$

em que z é a transformação de Fisher, $z(r) = \frac{1}{2} \ln \frac{1+r}{1-r}$.

Quando existem dados incompletos, n é substituído por n_{jk} . O resultado final é denotado por R_+^*. Uma estimativa da matriz de covariância pode ser obtida a partir da matriz de correlação, a partir da relação $S = D(s_i) \cdot R \cdot D(s_i)$, com S e R matrizes de covariância e correlação e $D(s_i)$ matriz diagonal de desvios-padrão S_i "aparados".

O Capítulo 10 de Mosteller e Tukey (1977) sugere um procedimento baseado em cálculos de regressão robusta que resultam em uma estimativa robusta para a matriz de covariância S^*. Para variáveis $y' = (y_1, \cdots, y_p)$, o procedimento inicia construindo a regressão de y_j em $(y_1, \cdots, y_{j-1})$, para $j = 2, \cdots, p$, utilizando o procedimento iterativo descrito no Capítulo 14 desta mesma referência. O próximo passo é calcular $z = (I - B^*)y$, onde $B^* = ((b_{jk}^*))$. Para $j \leq k$, $b_{jk}^* = 0$; e para $j > k$, b_{jk}^* é o coeficiente de y_k na regressão de y_j em $(y_1, \cdots, y_{j-1})$.

O terceiro passo é calcular D_*, uma matriz diagonal de estimativas robustas das variâncias dos elementos de z. Então S^*, estimativa robusta da matriz de covariância de Y é definido por:

$S^* = (I - B^*)^{-1} D^* (I - B^{*'})^{-1}$

Esta expressão é consequência do fato que

$$Y = (I - B^*)^{-1} Z \text{ e}$$

$$\widehat{Var}[(I - B^*)^{-1} Z] = (I - B^*)^{-1} \widehat{Var}(Z)(I - B^{*'})^{-1} = (I - B^*)^{-1} D^* (I - B^{*'}).$$

Uma matriz de correlação correspondente, R^*, pode ser obtida através da relação $R = D\left(\frac{1}{s_i}\right) S^* D\left(\frac{1}{s_i}\right)$.

O cálculo da regressão robusta pode ser realizado utilizando Mínimos Quadrados Ponderados Iterativos, considerando a solução dos Mínimos Quadrados Ordinários como ponto de partida. A estratégia aplicada seria utilizar uma ponderação do tipo 'Huber' nas primeiras três iterações, definida como:

$$w_i = \begin{cases} 1, & \text{se } \mid e_i \mid \leq ks^* \\ \frac{ks^*}{|e_i|}, & \text{caso contrário,} \end{cases}$$

onde $k = 2,4388$, e_i é o resíduo para a i-ésima observação e s^*=$mediana(\mid e_i \mid)$.

Com este valor de k, ponderações baixas são atingidas se $\mid e_i \mid$ supera $1,645 \times DP(e_i)$. Os pesos para as próximas iterações são definidos como:

$$w_i = \begin{cases} (1 - \left(\frac{e_i}{cs^*}\right)^2)^2, & \text{se } \left(\frac{e_i}{cs^*}\right)^2 < 1 \\ 0, & \text{caso contrário,} \end{cases}$$

para $c = 6$ ou 9 (sugerido na página 205 de Mosteller e Tukey (1977)). Com estas escolhas de c, o peso zero é obtido se $\mid e_i \mid$ supera quatro ou seis desvios-padrão. O procedimento pára quando a diferença absoluta máxima entre os valores w_i não é superior a 10^{-4} ou após 20 iterações.

A variância de cada elemento z_j de z é estimada através da estatística ns_{bi}^2 definida na página 208 de Mosteller e Tukey (1977).

Uma segunda abordagem multivariada proposta por Devlin, Gnanadesikan e Kettenring (1981) é baseada em procedimentos multivariados de observações aparadas (Gnanadesikan e Kettenring (1972) e Devlin, Gnanadesikan e Kettenring (1975)). Este procedimento também é iterativo. Em cada passo, as distâncias ao quadrado

$$DM_i^2 = (x_i - T(X)^*)' C(X)^{*-1} (x_i - T(X)^*), i = 1, \cdots, n \tag{3.3}$$

dos vetores de observações x_i com relação à estimativa de posição, $T(X)^*$, são medidas na métrica de $C(X)^*$, a estimativa da matriz de covariância usual. Um percentual especificado das observações mais extremas é temporariamente separado e as demais observações são utilizadas para calcular $T(X)^*$ e $C(X)^*$ exatamente como $T(X)$ e $C(X)$, o vetor de médias e a matriz de covariância amostrais.

Para iniciar o processo, $T(X)^*$ e $C(X)^*$ assumem os valores de $T(X)$ e $C(X)$ respectivamente. Em cada estágio, $C(X)^*$ pode ser convertida para R^*, conforme discutido anteriormente. O processo iterativo termina assim que a média da diferença absoluta entre dois valores da transformação de Fisher z dos valores r_{jk}^* em duas iterações sucessivas não supere 10^{-3} ou após 25 iterações.

A vantagem dos procedimentos propostos, ao contrário do método que estima correlações separadamente, é o fato que a matriz R^* obtida é positiva definida.

Os estimadores são formados através de análises bivariadas ou por manipulação simultânea de todas as variáveis utilizando técnicas como Médias e Variâncias Multivariadas Aparadas e Estimação M. Ambos são efetivos ao estimarem as componentes principais, cada técnica possuindo suas particularidades. Porém, os estimadores M podem não funcionar de maneira fácil quando a dimensionalidade é grande e os pontos discrepantes são assimétricos. Na existência de dados incompletos, a abordagem de estimadores robustos de correlação obtidos separadamente torna-se mais atrativa.

3.4 Análise Discriminante Robusta

O interesse na investigação de técnicas alternativas de análise discriminante tem crescido nos últimos anos. Uma primeira motivação para este fato tem sido o reconhecimento de que a análise discriminante clássica (FDL - Função Discriminante Linear), embora adequada

em inúmeras ocasiões, não é eficiente quando utilizada em bancos de dados reais, situação na qual é desejável o desenvolvimento de funções discriminantes específicas. Exemplos de aplicações destes tipos de dados podem ser encontrados quando se deseja classificar uma empresa como propensa ou não propensa a ir à falência, ou quando é necessário classificar se determinado pedido de empréstimo será provável à inadimplência ou mesmo a predição de categorias de classificação.

Segundo Hample, Ronchetti, Rousseeuw e Stahel (1986), um problema comum nos bancos de dados quando se utiliza a análise discriminante é que estes contêm pontos que geralmente são classificados como *outliers*. Bancos de dados que contêm pontos discrepantes são a regra ao invés de exceção, e estima-se que entre 1% e 10% dos pontos em um banco de dados típico são considerados *outliers*. Porém, o método linear padrão, baseado em mínimos quadrados ordinários (MQO) é extremamente sensível a *outliers* e pode produzir resultados não confiáveis e equivocados em sua presença. Conforme já comentado, estimativas da matriz de variância e covariância tornam-se muito distorcidas, conduzindo a estimativas não-ótimas. Outras anomalias distribucionais tais como assimetria também podem causar problemas. Uma resposta alternativa aos problemas na análise discriminante clássica tem sido o desenvolvimento de um grupo alternativo de classificadores não-paramétricos baseados em programação linear.

Esta seção apresenta uma metodologia de análise discriminante robusta proposta por Glorfeld (1990) que pode ser formulada como um problema de programação linear baseado no método de estimação da Mínima Soma de Valores Absolutos (MSVA). Por simplicidade, o foco desta seção será restrito ao problema da análise discriminante linear para duas populações, com probabilidades iguais de pertencerem aos dois grupos e custos iguais de classificação incorreta.

Análise Discriminante Linear para Duas Populações

Segundo Green (1979), existe na literatura certa confusão no que diz respeito aos reais objetivos da análise discriminante linear. A análise discriminante linear para duas

populações possui dois objetivos distintos:

(1) A projeção de um grupo complexo de dados multivariados medidos em duas populações em um espaço linear discriminante unidimensional muito mais simples, que permite visualização gráfica para estudar o grau e a natureza da separação entre os grupos. É útil para a sugestão de possíveis anomalias distribucionais e para a indicação de observações discrepantes;

(2) O desenvolvimento de funções de classificação lineares baseadas na amostra de dados multivariados que permite a classificação de novas observações em um dos dois grupos cometendo erro mínimo.

A maioria das técnicas de análise discriminante linear desenvolvidos até hoje têm como objetivo principal o segundo objetivo citado acima - classificação - e dão pouca atenção ao primeiro objetivo. Uma característica diferenciada da metodologia discriminante da Mínima Soma do Valor Absoluto é a de que seu principal interesse é o primeiro objetivo citado e utiliza a classificação apenas em segundo plano.

A formulação do modelo discriminante clássico para o problema de duas populações com probabilidades iguais de classificação incorreta envolve a construção da combinação linear de p variáveis baseadas em $n_1 = n_2$ observações retirados de cada uma das duas populações, onde a população de cada elemento amostral é conhecida. Se as $n = n_1 + n_2$ observações são independentes, as p variáveis têm distribuição normal multivariada e as matrizes de variância e covariância são iguais para os dois grupos, estes irão diferir apenas no que diz respeito a seus vetores de médias e o procedimento é, nesse sentido, ótimo. A Figura 3.2 descreve a representação clássica.

Segundo Fisher (1938), a função discriminante determina diretamente a projeção unidimensional do espaço e não os limites de classificação. Desta maneira, possui a interessante propriedade de que para as duas populações, um problema com p variáveis é reduzido a um problema univariado de forma muito mais simples. A direção dessa projeção produz um novo grupo de escores discriminantes que possuem variabilidade máxima entre-grupos

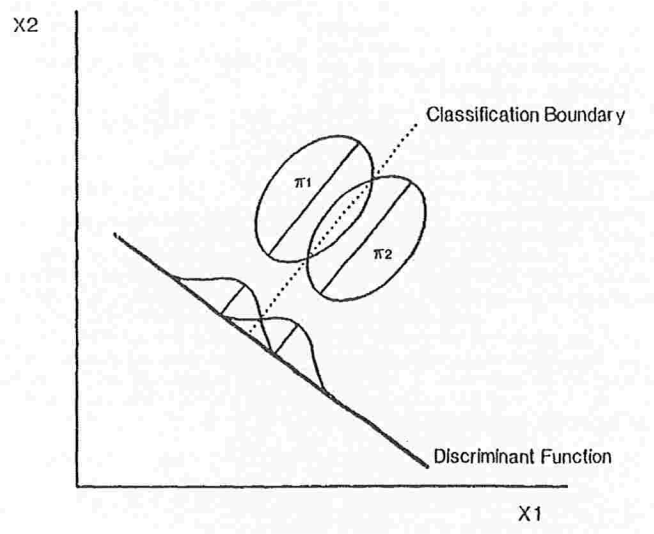

Figura 3.2: Representação da função discriminante linear clássica para duas variáveis.
Fonte: Glorfeld (1990).

quando comparada à variabilidade intra-grupos. Um ponto é classificado no grupo com média mais próxima, fazendo com que o limite de classificação, c, seja igual à média global dos escores projetados. Com base no Teorema Central do Limite, os escores discriminantes lineares compostos podem possuir distribuição aproximadamente normal, mesmo que uma ou mais das p variáveis discriminantes presentes não tenham distribuição normal, indicando robustez da função linear discriminante quanto à violação da hipótese de normalidade.

Na formulação clássica do modelo discriminante, Fisher (1936) mostrou que o problema discriminante de duas populações pode ser formulado em termos da Análise de Regressão via Mínimos Quadrados Ordinários (MQO):

$$Y = W_0 + W_1 X_1 + W_2 X_2 + \ldots + W_p X_p + \varepsilon, \tag{3.4}$$

sendo que Y é uma variável dicotômica n-dimensional que identifica o grupo ao qual o elemento pertence, $X_1, \ldots, X_p$ são as variáveis aleatórias preditoras com n componentes, amostradas de duas populações com vetores de médias μ_1 e μ_2 e matriz de covariância comum Σ; $W_0, \ldots, W_p$ são os parâmetros desconhecidos a serem estimados por $w_0, \ldots, w_p$,

e ε é o erro aleatório. Na formulação clássica, os parâmetros W's são estimados de maneira a minimizar a soma dos quadrados dos resíduos, $\sum e_i^2 = \sum_{i=1}^{n}[y_i - (w_0 + \sum w_i x_{ij})]^2$, em que y_i é a i-ésima componente de Y e x_{ij} é o i-ésimo valor de X_j, $i = 1, \ldots, n$ e $j = 1, \ldots, p$.

A formulação do problema de análise discriminante via regressão vista na expressão (3.4) é um assunto controverso e é considerado por alguns autores como um artificio algébrico. Uma revisão dessa formulação feita por Flury e Riedwyl (1985) mostra a justificativa prática para tal abordagem e sugere este procedimento para aplicações. Os autores mostram que o usual teste-F da regressão é equivalente ao teste T^2-Hotelling para a diferença entre grupos em relação aos centróides e a variável individual p do teste parcial F é um teste-t univariado para a diferença entre as médias ao longo de p dimensões. Os valores estimados, $\widehat{y}_i$, são usados na construção de um histograma para a exibição gráfica dos dados no espaço discriminante.

Se a codificação usual, 'zero' e 'um' for utilizada para indicar as populações, então classifica-se y_i em π_1 se $\widehat{y}_i < 0,5$ e em π_2 caso contrário, onde π_1 e π_2 são as duas populações, respectivamente codificadas como 'zero' e 'um'.

Modelo Robusto da Mínima Soma do Valor Absoluto (MSVA) Segundo Narula e Wellington (1982), o ajuste do modelo de regressão através do método de Mínima Soma de Valores Absolutos (MSVA) tem a mesma formulação do correspondente via Mínimos Quadrados Ordinários, com exceção ao fato de que os erros são medidos como o valor absoluto da diferença entre os valores observados e estimados. Este modelo irá determinar os valores w's tais que a soma dos valores absolutos dos resíduos $\sum \mid e_i \mid = \sum \mid y_i - (w_0 + \sum w_j x_{ij}) \mid$ seja mínima.

Pelo fato do modelo MSVA ser pouco familiar, uma breve formulação será dada a seguir. Utilizando uma representação padrão de programação linear, é possível visualizar cada um dos n vetores observados no modelo linear

$$y_i = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \varepsilon_i, i = 1, \ldots, n \tag{3.5}$$

como uma restrição. Substituindo ε_i na expressão (3.5) por $(\varepsilon_i^+ + \varepsilon_i^-)$, o problema é equivalente a minimizar

$$z = \sum (\varepsilon_i^+ + \varepsilon_i^-), i = 1, \ldots, n, \tag{3.6}$$

sujeito a

$$\begin{cases} y_i - (w_0 + \sum_j w_j x_{ij}) + \varepsilon_i^+ - \varepsilon_i^- = 0, \\ i = 1, \ldots, N, j = i, \ldots, p, \\ \varepsilon_i^+, \varepsilon_i^- \geq 0, i = 1, \ldots, n, \\ w_j, j = 0, \ldots, p \text{ sem restrição de sinal} \end{cases}$$

em que

$$\varepsilon_i^+ = \begin{cases} \varepsilon_i, \text{ se } \varepsilon_i > 0, \\ 0, \text{ caso contrário.} \end{cases}$$

e

$$\varepsilon_i^- = \begin{cases} -\varepsilon_i, \text{ se } \varepsilon_i < 0, \\ 0, \text{ caso contrário.} \end{cases}$$

O valor ótimo de z obtido na expressão (3.6) será o menor valor da soma dos erros absolutos. Um aspecto especialmente interessante da formulação MSVA é a pronta disponibilidade de softwares especializados para resolver o problema da estimação, mesmo para aquelas formulações não familiares. Por exemplo, o software SAS (SAS (1976)) contém um procedimento relativamente eficiente que pode ser empregado por qualquer pesquisador com relativa facilidade.

Para fins de testes estatísticos, a teoria desenvolvida é assintótica, o que permite produzir testes para os parâmetros do modelo de maneira similar ao MQO, desde que se utilize grandes tamanhos de amostra. Diferente da regressão MQO, é necessário assumir apenas que os erros são independentes e identicamente distribuídos, e verifica-se que para qualquer distribuição dos erros para a qual a mediana seja assintoticamente superior à média como um estimador de posição, a estimação MSVA é preferível em relação à MQO. Dielman e Pfaffen-

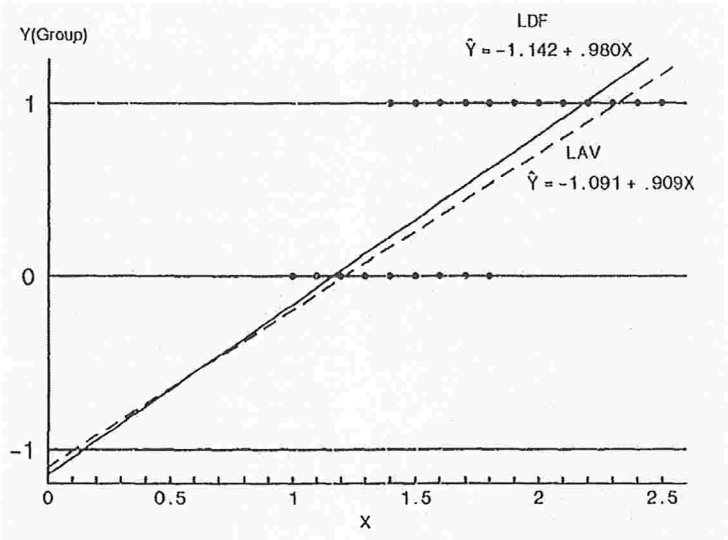

Figura 3.3: Função linear discriminante e Função discriminante da Mínima Soma do Valor Absoluto sem *outliers*.
Fonte: Glorfeld (1990).

berger (1982) dão alguma orientação sobre o tamanho mínimo de amostra necessária para aplicar testes assintóticos. Segundo Teebagy e Chatterjee (1989), para pequenas amostras, testes baseados no método *Bootstrap* podem ser desenvolvidos.

A primeira justificativa para o uso da função discriminante MSVA pode ser compreendida através das Figuras 3.3 e 3.4, que representam o caso discriminante mais simples, onde $p = 1$.

Quando os dados não são contaminados por outliers, a função discriminante usual FDL e a obtida pelo método MSVA fornecerão resultados muito próximos, como descrito na Figura 3.3.

Entretanto, na Figura 3.4 é possível observar que quando existem *outliers*, a função discriminante linear torna-se seriamente distorcida, enquanto a associada ao método MSVA é muito menos afetada.

Tendo usado a função discriminante MSVA para projeção dos dados p-variados em um espaço discriminante univariado, surge a questão de como uma regra de classificação deve ser formulada. Um grande número de alternativas distintas são possíveis. Por exemplo,

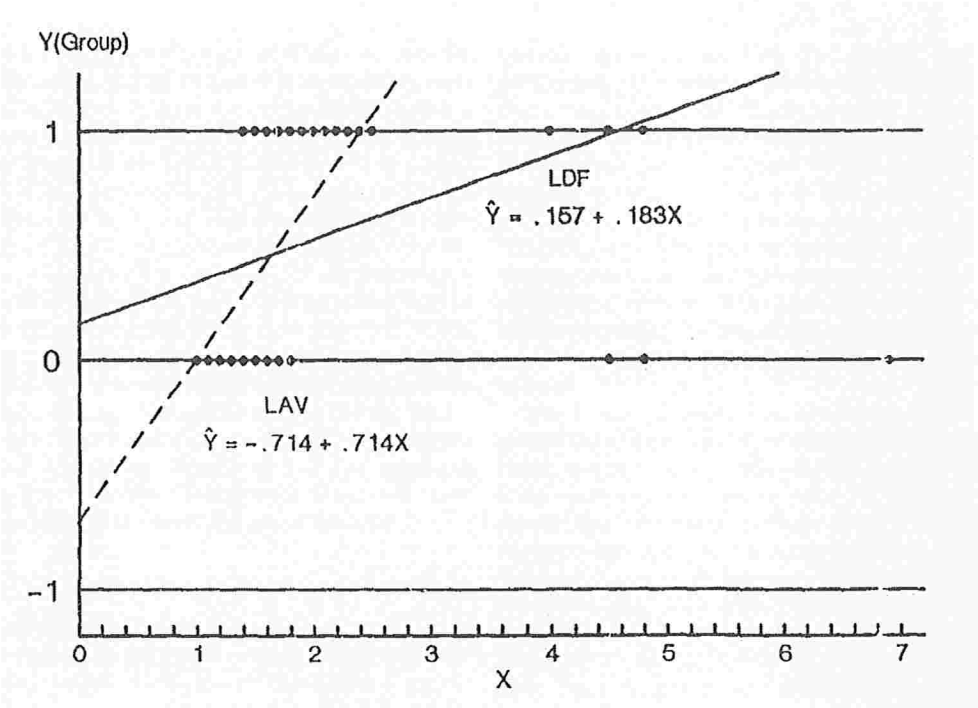

Figura 3.4: Função linear discriminante e Função discriminante do Mínimo Valor Absoluto considerando os *outliers*.
Fonte: Glorfeld (1990).

desde que a função discriminante MSVA esteja próxima da função discriminante via MQO que teria sido desenvolvida se os *outliers* não estivessem presentes nos dados, a classificação MQO do limite pode ser utilizada:

$$w_i = \begin{cases} \text{Se } \widehat{y}_i < 0,5, & \text{classificar em } \pi_1 \\ \text{Se } \widehat{y}_i \geq 0,5, & \text{classificar em } \pi_2. \end{cases}$$

Alternativamente, a mediana global dos escores estimados ($\widehat{y}_i$) através do procedimento MSVA pode ser utilizada como limite de classificação. Limites de classificação podem também ser facilmente determinados minimizando-se o número de erros de classificação.

Através de testes em bancos de dados distintos, a seguinte regra de classificação foi desenvolvida por Glorfeld (1990): classificar y_i em π_1 se $\mid \widehat{y}_i - M_1 \mid < \mid \widehat{y}_i - M_2 \mid$ e em π_2 caso contrário, em que M_1 e M_2 são as medianas dos escores do primeiro e segundo grupos

respectivamente. Este procedimento é equivalente a utilizar a seguinte regra de classificação:

$$\begin{cases} \text{Se } \widehat{y}_i < \frac{M_1+M_2}{2}, & \text{classificar em } \pi_1, \\ \text{caso contrário,} & \text{classificar em } \pi_2. \end{cases}$$

Após a descrição da técnica, como análise de diagnóstico, o autor sugere que as funções discriminantes FDL e MSVA devam ser obtidas para o mesmo conjunto de dados e seus resultados comparados. Se a comparação exibir grandes diferenças, então os resultados obtidos via MSVA podem ser úteis para determinar possíveis erros grosseiros nos dados, assim como fornecer um modelo que melhor se ajuste aos mesmos. Gráficos, coeficientes do modelo, testes estatísticos e a regra de classificação podem ser diretamente comparados para auxiliar na avaliação de problemas com a formulação da função discriminante linear FDL. Detectados pontos discrepantes, estas observações suspeitas podem ser removidas e a análise repetida. Se os resultados das funções discriminantes FDL e MSVA mostram apenas pequenas diferenças, o procedimento da função discriminante linear pode ser mantido.

He e Fung (2000) consideram o problema da análise discriminante com duas populações com médias μ_x e μ_y e matriz de covariância comum Σ. Para a regra de decisão usual de classificar uma observação z na população π_1 se

$$(\mu_x - \mu_y)'\Sigma^{-1}\{z - (\mu_x + \mu_y)/2\} > 0$$

e na população π_2 caso contrário, os autores sugerem o uso de estimadores de alto ponto de ruptura para μ_x, μ_y e Σ. Os autores propõem uma particular forma de obtenção desses estimadores, sugerindo que essa seria uma possível maneira de tornar a análise discriminante de Fisher mais robusta.

Hubert e Van Driessen (2004) propõem o uso dos estimadores robustos de vetores de média e matrizes de covariâncias para situações mais gerais que excluíam o caso de mais do que duas populações, análise discriminante bayesiana, quadrática e linear. Apresentam

ainda estimadores robustos das probabilidades de má classificação.

Neste capítulo foram discutidas medidas robustas de diagnóstico em análise multivariada. Foram apresentados os estimadores robustos do tipo M para vetores de médias e matrizes de covariância, além de procedimentos para análise robusta de componentes principais e análise discriminante. O próximo capítulo é inteiramente dedicado ao estudo de medidas robustas de diagnóstico em análise de regressão.

Capítulo 4

Medidas Robustas de Diagnóstico em Análise de Regressão

4.1 Introdução

O principal objetivo deste capítulo é introduzir técnicas e medidas de diagnóstico robusto na análise de regressão. Conforme comentado no capítulo anterior, medidas de distância como a de Mahalanobis, utilizadas para detectar pontos discrepantes no contexto multivariado, podem sofrer com o problema do mascaramento. Na análise de regressão este problema também pode existir, conforme será discutido a seguir.

4.2 *Outliers Multivariados e Pontos de Alavanca*

Identificação de pontos de alavanca na análise de regressão

Na regressão linear os elementos amostrais são da forma $(\mathbf{x_i}, \mathbf{y_i})$, onde $\mathbf{x_i}$ é p-dimensional e a resposta y_i é unidimensional. Definem-se pontos de alavanca como sendo os casos em que $\mathbf{x_i}$ encontram-se distantes da maioria dos dados no espaço p-dimensional. Conforme já abordado no Capítulo 2, um ponto tem influência significativa quando o valor da sua "alavancagem" for alto. Pontos de alavanca podem ser muito difíceis de serem de-

tectados principalmente quando $\mathbf{x_i}$ tem dimensão superior a dois, pois nesta situação não é possível o embasamento na percepção visual.

No modelo de regressão linear múltiplo dado por $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, geralmente utilizam-se os elementos h_{ii} da diagonal da matriz chapéu $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ como diagnóstico para identificar pontos de alavanca. A literatura sugere que se dê especial atenção a pontos em que o valor de h_{ii}, definido no Capítulo 2, for maior que $\frac{2p}{n}$, sendo p o número de colunas da matriz de planejamento X e n o tamanho da amostra.

De acordo com Rousseeuw e Van Zomeren (1990), a matriz chapéu $\mathbf{H}$, assim como a Distância de Mahalanobis Clássica ($DM_i = \sqrt{(\mathbf{x_i} - \mathbf{T(X)}) \cdot \mathbf{S(X)^{-1}} \cdot (\mathbf{x_i} - \mathbf{T(X)})'}$), é penalizada pelo efeito de mascaramento. Isto pode ser explicado ao se perceber que existe uma relação monótona entre h_{ii} e DM_i associada ao elemento $\mathbf{x_i}$, dada por:

$$h_{ii} = \frac{(DM_i)^2}{n-1} + \frac{1}{n}. \tag{4.1}$$

Desta maneira, h_{ii} não necessariamente irá detectar os pontos de alavanca, ao contrário do que se acredita.

Como ilustração, os autores apresentam a Tabela 4.1, que mostra os valores de h_{ii} para o conhecido conjunto de dados *Stackloss* (Brownlee (1965)). Este banco de dados é obtido da observação da oxidação da amônia em ácido nítrico de uma planta, durante 21 dias. É composto por 21 linhas e 4 variáveis: *AirFlow*: fluxo de ar resfriado, *WaterTemp*: temperatura de entrada da água resfriada, *AcidoConc.*: concentração de ácido e *stack.loss*: basicamente trata-se da quantidade de calor que viaja através de um local ou recipiente sem nenhuma força externa e é a variável resposta. O maior valor de h_{ii} nesta tabela pertence à observação 17. Já a Distância Robusta (DR_i), obtida substituindo-se as estimativas $T(\mathbf{X})$ e $\mathbf{S(X)}$ na equação da Distância de Mahalanobis Clássica (DM_i) por estimadores de "locação" e de "covariância" robustos, identifica as observações 1,2,3 e 21. Sugere-se identificar observações com distâncias superiores a $\sqrt{\chi^2_{p;0,975}}$, em que $\chi^2_{p;0,975}$ é o quantil de ordem $0,975$ da distribuição Qui-Quadrado com p graus de liberdade.

Os estimadores robustos utilizados neste método, denominado EVM (Elipsóide de Volume Mínimo) foram:

$T(\mathbf{X})$: centro do elipsóide de volume mínimo que cobre pelo menos metade das observações $\mathbf{x_i}, \mathbf{i = 1, 2, \ldots, n}$.

De modo geral, a equação do elipsóide é da forma

$$(\mathbf{x_i} - \mathbf{T(X)}) \cdot (\mathbf{S(X)^{-1}}) \cdot (\mathbf{x_i} - \mathbf{T(X)})' \leq \mathbf{a^2},$$

em que $\mathbf{S(X)}$: matriz $p \times p$, positiva semi-definida e a^2 é uma constante positiva.

Esta construção garante que $T(\mathbf{X})$ e $\mathbf{S(X)}$ sejam os estimadores robustos de locação e da matriz de covariância de X.

Os detalhes técnicos para obtenção deste elipsóide estão descritos no Apêndice A.

A Tabela 4.2 apresenta o exemplo oriundo dos dados *Hawkins-Bradu-Kass* (Hawkins, Bradu e Kass (1984)). Este banco de dados artificial consiste de 75 observações em 4 dimensões (uma variável resposta e três explicativas) e fornece um bom exemplo de efeito de mascaramento.

Sabe-se que as 14 primeiras observações desse conjunto de dados são pontos de alavanca, porém apenas as observações 12, 13 e 14 possuem altos valores de h_{ii}. Por outro lado, o cálculo da distância robusta DR_i identifica claramente essas observações. Como consequência, Rousseeuw e Van Zomeren (1990) propõem o uso de distâncias robustas de $\mathbf{x_i}$ como diagnósticos de alavanca, por serem menos sensíveis ao mascaramento que os valores h_{ii}.

Dizemos que $(\mathbf{x_i}, \mathbf{y_i})$ é um ponto de alavanca apenas com relação à discrepância de $\mathbf{x_i}$, não levando em consideração a resposta y_i. Se, além disso, $(\mathbf{x_i}, \mathbf{y_i})$ localiza-se afastado da maioria dos dados, define-se o ponto como sendo um "mau" ponto de alavanca. Este ponto é bastante prejudicial, pois altera a tendência do ajuste da regressão de mínimos quadrados. Em contrapartida, se $(\mathbf{x_i}, \mathbf{y_i})$ segue a tendência do conjunto de dados, será chamado de "bom"

i	DM_i	DR_i	h_{ii}
1	2,25	5,23	0,30
2	2,32	5,27	0,32
3	1,59	4,01	0,17
4	1,27	0,84	0,13
5	0,30	0,80	0,05
6	0,77	0,78	0,08
7	1,85	0,64	0,22
8	1,85	0,64	0,22
9	1,36	0,83	0,14
10	1,75	0,64	0,20
11	1,47	0,58	0,16
12	1,84	0,79	0,22
13	1,48	0,55	0,16
14	1,78	0,64	0,21
15	1,69	2,23	0,19
16	1,29	2,11	0,13
17	2,70	2,07	0,41
18	1,50	2,09	0,16
19	1,59	2,29	0,17
20	0,81	0,64	0,08
21	2,18	3,30	0,28

Tabela 4.1: Distâncias de Mahalanobis (DM_i), Distâncias Robustas (DR_i) e elementos da diagonal da matriz chapéu (h_{ii}) para os dados *Stackloss*.
Fonte: Rousseeuw e Van Zomeren (1990).

ponto de alavanca, pois melhora a precisão das estimativas dos coeficientes da regressão.

A Figura 4.1 mostra esta terminologia em um exemplo de regressão simples. A maioria dos dados são observações regulares, indicados por (a). Os pontos (b) e (d) fogem do padrão linear e, portanto, são chamados de outliers da regressão, mas (c) não. Ambos, (c) e (d) são pontos de alavanca, pois apresentam valores distantes da maioria dos dados para $\mathbf{x_i}$. Porém (c) é um "bom" ponto de alavanca, enquanto (d) é um "mau" ponto de alavanca. O ponto (b) é chamado de *outlier* vertical, pois se trata de um *outlier* de regressão, mas não de um ponto de alavanca.

Para distinguir entre "bons" e "maus" pontos de alavanca é necessário considerar y_i e $\mathbf{x_i}$ e também conhecer o padrão linear definido pela maioria dos dados. Com este objetivo,

i	DM_i	DR_i	h_{ii}	i	DM_i	DR_i	h_{ii}
1	1,92	16,20	0,063	39	1,27	1,34	0,035
2	1,86	16,62	0,060	40	1,11	0,55	0,035
3	2,31	17,65	0,086	41	1,70	1,48	0,052
4	2,23	18,18	0,081	42	1,77	1,74	0,055
5	2,10	17,82	0,073	43	1,87	1,18	0,061
6	2,15	16,80	0,076	44	1,42	1,82	0,041
7	2,01	16,82	0,068	45	1,08	1,25	0,029
8	1,92	16,44	0,063	46	1,34	1,70	0,038
9	2,22	17,71	0,080	47	1,97	1,65	0,066
10	2,33	17,21	0,087	48	1,42	1,37	0,041
11	2,45	20,23	0,094	49	1,57	1,27	0,047
12	3,11	21,14	0,144	50	0,42	0,83	0,016
13	2,66	20,16	0,109	51	1,30	1,19	0,036
14	6,38	22,38	0,564	52	2,08	1,61	0,072
15	1,82	1,54	0,058	53	2,21	2,41	0,079
16	2,15	1,88	0,076	54	1,41	1,26	0,040
17	1,39	1,03	0,039	55	1,23	0,66	0,034
18	0,85	0,73	0,023	56	1,33	1,21	0,037
19	1,15	0,59	0,031	57	0,83	0,93	0,023
20	1,59	1,49	0,048	58	1,40	1,31	0,040
21	1,09	0,87	0,030	59	0,59	0,96	0,018
22	1,55	0,90	0,046	60	1,89	1,89	0,062
23	1,09	0,94	0,029	61	1,68	1,31	0,051
24	0,97	0,83	0,026	62	0,76	1,22	0,021
25	0,80	1,26	0,022	63	1,29	1,17	0,036
26	1,17	0,86	0,032	64	0,97	1,14	0,026
27	1,45	1,35	0,042	65	1,15	1,40	0,031
28	0,87	1,00	0,024	66	1,30	0,78	0,036
29	0,58	0,72	0,018	67	0,63	0,37	0,019
30	1,57	1,97	0,047	68	1,55	1,64	0,046
31	1,84	1,43	0,059	69	1,07	1,17	0,029
32	1,31	0,95	0,036	70	1,00	1,04	0,027
33	0,98	0,73	0,026	71	0,64	0,64	0,019
34	1,18	1,42	0,032	72	1,05	0,52	0,028
35	1,24	1,26	0,034	73	1,47	1,14	0,043
36	0,85	0,86	0,023	74	1,65	0,96	0,050
37	1,83	1,26	0,059	75	1,90	1,99	0,062
38	0,75	0,92	0,021				

Tabela 4.2: Distâncias de Mahalanobis (DM_i), Distâncias Robustas (DR_i) e elementos da diagonal da matriz chapéu (h_{ii}) para os dados *Hawkins − Bradu − Kass*.
Fonte: Rousseeuw e Van Zomeren (1990).

os autores sugerem estimar os parâmetros do modelo de regressão pelo procedimento da Mínima Mediana dos Quadrados do Resíduo (MMQR). Este procedimento proposto por Rousseeuw (1984) minimiza em θ o valor $mediana_{(i=1,\ldots,n)}r_i^2(\theta)$, em que

$$r_i(\theta) = y_i - \mathbf{x_i}\widehat{\theta}$$

é o resíduo da i-ésima observação. Verifica-se que a estimativa $\widehat{\theta}$ de MMQR possui o máximo ponto de ruptura. Segundo Gnanadesikan (1997), ponto de ruptura é a maior fração das observações que podem ser valores extremos em uma amostra, sem alterar o valor do estimador.

Após obter $\widehat{\theta}$, Rousseeuw e Van Zomeren (1990) sugerem calcular também a estimativa do desvio padrão dos erros, dada por:

$$\widehat{\sigma} = k \cdot \sqrt{mediana_{(i=1,\ldots,n)}r_i^2(\theta)},$$

em que k é uma constante positiva. Os resíduos padronizados $r_i/\widehat{\sigma}$ obtidos após esse ajuste podem então ser usados para indicar pontos discrepantes que destoam do padrão linear da maioria dos dados.

É importante observar que as distâncias robustas nas Tabelas 4.1 e 4.2 indicam pontos de alavanca, mas não conseguem distinguir entre os "bons" e "maus", pois y_i não é utilizado. Em contrapartida, os gráficos usuais de resíduos do ajuste MMQR mostram *outliers* de regressão sem informar quais deles são pontos de alavanca. Portanto, parece interessante a construção de um gráfico no qual os resíduos robustos $r_i/\widehat{\sigma}$ são confrontados com as distâncias robustas DR_i. Na Figura 4.2 isto é feito para os dados *Stackloss*. Pontos à direita da linha vertical delimitada por $\sqrt{\chi^2_{3;0,975}} = 3,06$ são considerados pontos de alavanca (por apresentarem valores altos da distância robusta DR_i), sendo que pontos fora das linhas de tolerância horizontal $[-2,5; 2,5]$ são *outliers* de regressão (por apresentarem valores altos para os resíduos padronizados). Neste exemplo, os quatro pontos com os maiores valores

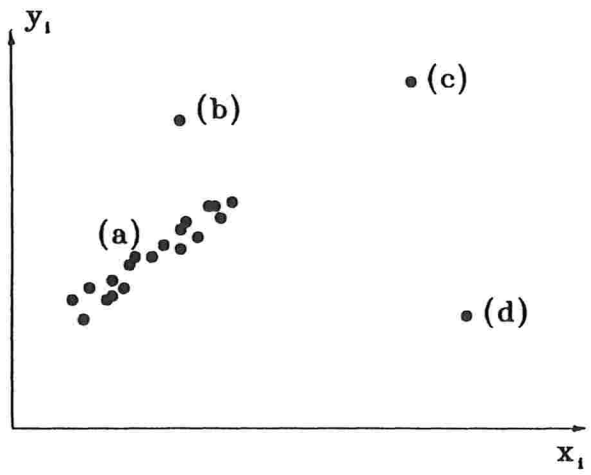

Figura 4.1: Exemplo de regressão linear simples com (a)observações regulares, (b)*outliers* verticais, (c)"bons" pontos de alavanca e (d)"maus" pontos de alavanca.
Fonte: Rousseeuw e Van Zomeren (1990).

de DR_i (1,2,3,21) são também *outliers* de regressão, e conseqüentemente, "maus" pontos de alavanca. A Figura 4.2 também contém um *outlier* vertical (observação 4), que é um *outlier* de regressão com $DR_i < \sqrt{\chi^2_{3;0,975}}$. Os valores de corte utilizados são de certa forma, arbitrários, mas no gráfico é possível reconhecer os casos limítrofes: a observação 21 não está tão longe no espaço dos valores de x, enquanto a observação 2 é apenas um *outlier* de regressão moderado.

Complementando a análise, construiu-se o gráfico dos resíduos de Mínimos Quadrados Usual contra a distância de Mahalanobis não robusta DM_i. Este gráfico encontra-se na Figura 4.3 e seria o correspondente ao da Figura 4.2 em uma análise tradicional. Verifica-se que este não revela nenhum ponto de alavanca ou *outlier* de regressão, pois todos os pontos localizam-se entre as linhas limite e apenas as observações 21 e 17 ficam próximas dessa fronteira. Além disso, como conseqüência da expressão (4.1), a substituição de DM_i por h_{ii} não causaria nenhuma melhoria.

A Figura 4.4 apresenta o gráfico dos resíduos robustos contra as distâncias robustas para os dados $Hawkins - Bradu - Kass$. Claramente nota-se a existência de 14 pontos de alavanca, dos quais 4 são "bons" e 10 são "maus". Este tipo de gráfico apresenta uma

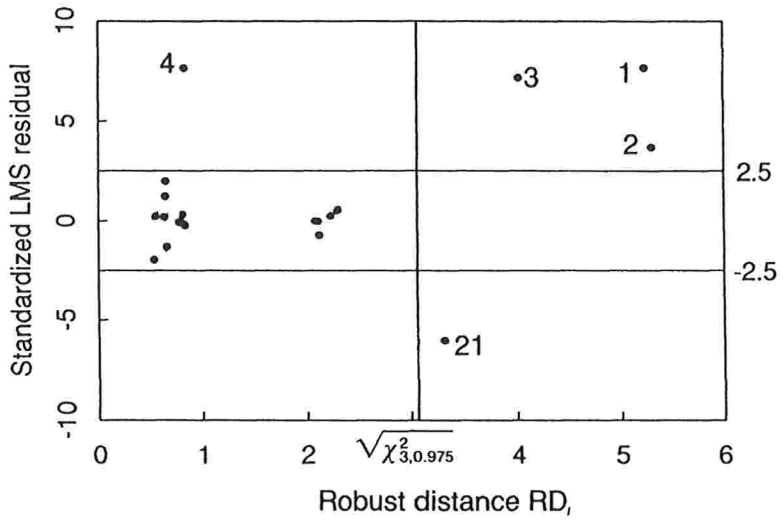

Figura 4.2: Gráfico dos Resíduos Robustos contra as Distâncias Robustas RD_i para os dados *Stackloss*.
Fonte: Rousseeuw e Van Zomeren (1990).

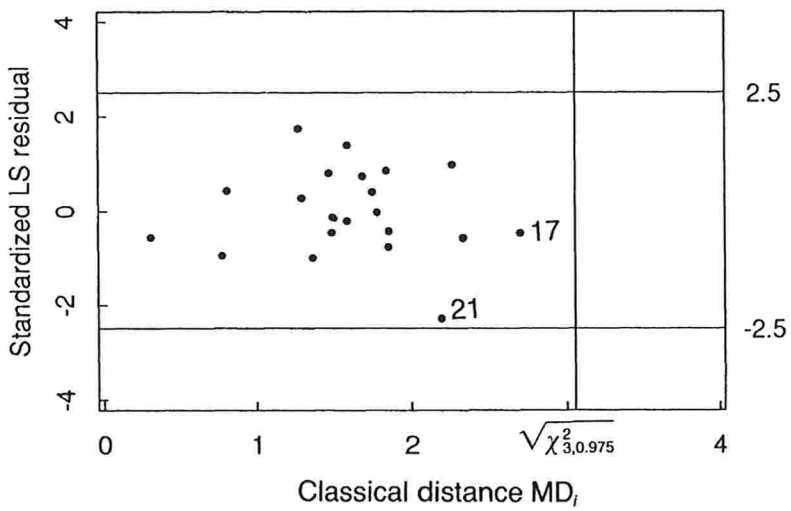

Figura 4.3: Gráfico dos Resíduos de Mínimos Quadrados contra a Distância Clássica de Mahalanobis DM_i para os dados *Stackloss*.
Fonte: Rousseeuw e Van Zomeren (1990).

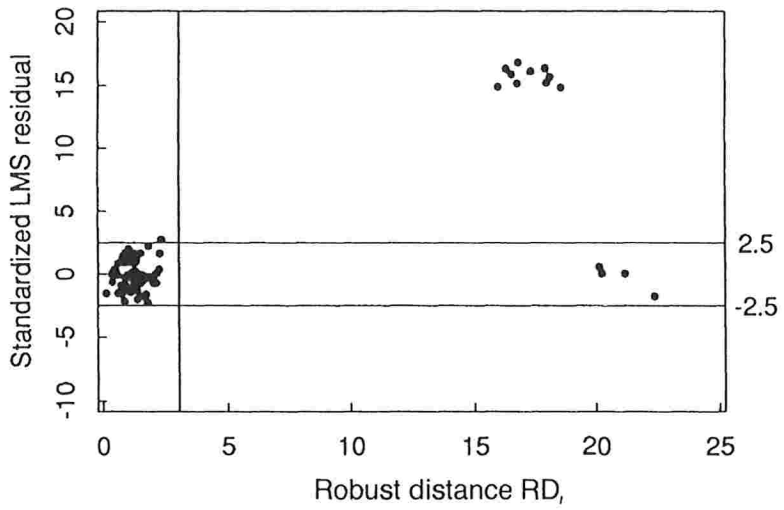

Figura 4.4: Gráfico dos Resíduos Robustos contra as Distâncias Robustas RD_i para os dados *Hawkins – Bradu – Kass*.
Fonte: Rousseeuw e Van Zomeren (1990).

classificação visual dos dados em 4 categorias: as observações regulares com pequenas DR_i e pequenos $r_i/\widehat{\sigma}$; os outliers verticais com pequenas DR_i e grandes $r_i/\widehat{\sigma}$, os "bons" pontos de alavanca com grandes DR_i e pequenos $r_i/\widehat{\sigma}$ e os "maus" pontos de alavanca com grandes DR_i e grandes $r_i/\widehat{\sigma}$. Percebe-se então que um diagnóstico simples nunca é suficiente para identificar estas quatro classificações e medidas conjuntas se fazem necessárias.

Finalizando, vale a pena destacar que o artigo não é a favor da simples remoção dos *outliers*. Ao invés disso, consideram-se gráficos de resíduos e cálculo de distâncias robustas como início da análise. Em alguns casos pode ser útil o retorno aos dados originais e a procura por justificativas para os *outliers*, analisando sua origem e verificando se eventualmente são conseqüências de erros de medida ou transcrição.

Resíduos robustos podem ser utilizados para determinar pesos às observações ou para sugerir transformações nos dados (Caroll e Ruppert (1988); Rousseeuw e Leroy (1987)), assunto que não será abordado no presente trabalho.

Pesquisas futuras nesta linha podem ser realizadas para situações onde uma das variáveis explicativas é discreta (como por exemplo, do tipo 0 ou 1). Além disso, Rousseeuw

e Van Zomeren (1990) acreditam que modelos de regressão polinomial necessitariam de um estudo diferenciado devido às dificuldades de construção de elipsóides se as variáveis explicativas são relacionadas. O mesmo problema surgiria na presença de multicolinearidade.

4.3 Outliers Multivariados e Pontos de Alavanca - Uma confirmação

Conforme já comentado, a identificação de *outliers* multivariados e pontos de alavanca tornam-se difíceis devido ao efeito de mascaramento. Na seção anterior foram citados métodos de estimação robustos com alto ponto de ruptura (Mínima Mediana dos Quadrados do Resíduo - MMQR e Elipsóide de Volume Mínimo - EVM) para detecção destas observações. Entretanto, Fung (1993) levanta o problema que estes métodos tendem a identificar muitas observações como sendo extremas. Uma análise gradativa (passo-a-passo) é proposta pelo autor para confirmação dos *outliers* e pontos de alavanca detectados através destes métodos robustos. São construídas medidas de diagnóstico para observações excluídas que são devolvidas à amostra reduzida. Esta seção apresenta uma descrição do procedimento e sua posterior aplicação a dois exemplos. A limitação da abordagem confirmatória de Atkinson (1986) também é discutida.

Estimadores com alto ponto de ruptura de Rousseeuw e Análise Confirmatória de Atkinson

Medidas de diagnóstico baseados na retirada de observações têm sido desenvolvidas e são úteis na identificação de *outliers* e observações influentes na análise de regressão. Porém, tais medidas podem sofrer pelo efeito de mascaramento, fazendo com que *outliers* múltiplos não sejam detectados. Métodos de retirada simultâneos não possuem este problema, embora sejam pouco utilizados devido às dificuldades computacionais. Na seção anterior foram sugeridos estimadores robustos com alto ponto de ruptura em modelos de regressão para "desmascarar" *outliers* e pontos de alavanca.

Considere o seguinte modelo de regressão:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.2}$$

em que $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})^{\mathbf{t}}$ é uma matriz $n \times p$ de variáveis explicativas, $\mathbf{Y}$ é um vetor de respostas $n \times 1$, β é um vetor de parâmetros e ε é um vetor de erros com matriz de covariância $\sigma^2\mathbf{I_n}$.

Rousseeuw (1984) sugeriu para o modelo dado na expressão (4.2) o uso da regressão da Mínima Mediana dos Quadrados do Resíduo ao invés da utilização dos Mínimos Quadrados usuais. Isto é, utiliza-se o estimador $\widetilde{\beta}$ da Mínima Mediana dos Quadrados do Resíduo, que é obtido minimizando em $\widetilde{\beta}$ o valor $mediana_{(i=1,\ldots,n)}\widetilde{\varepsilon_i}^2(\widetilde{\beta})$ em que $\widetilde{\varepsilon_i}(\widetilde{\beta}) = y_i - \mathbf{x_i^t}\widetilde{\beta}$ é o resíduo da i-ésima observação. Segundo Rousseeuw (1984), este estimador tem um ponto de ruptura de 50%, porém baixa eficiência. A estimativa de escala correspondente é obtida por $\widetilde{\sigma} = k \cdot \sqrt{mediana_{(i=1,\ldots,n)}\widetilde{\varepsilon_i^2}\widetilde{\beta}}$, sendo k uma constante positiva. Rousseeuw e Leroy (1987) sugeriram a utilização de resíduos da Mínima Mediana dos Quadrados dos Resíduos padronizados $\widetilde{\varepsilon_i}/\widetilde{\sigma}$ para detecção dos *outliers*. Os autores utilizam como pontos de corte os valores $-2,5$ e $2,5$.

O método descrito é útil na detecção de *outliers* relacionados à variável resposta, porém *outliers* na direção do eixo x (isto é, os pontos de alavanca) são também de interesse.

Considere a partição $\mathbf{x_i^t} = (\mathbf{1}, \mathbf{z_i^t})$. Foi visto na seção anterior que a distância robusta da i-ésima observação é dada por $DR_i = \sqrt{(\mathbf{z_i} - \mathbf{T(Z)})^{\mathbf{t}} \cdot \mathbf{S(Z)^{-1}} \cdot (\mathbf{z_i} - \mathbf{T(Z)})}$, sendo que $\mathbf{T(Z)}$ e $\mathbf{S(Z)}$ são a média e covariância robustas de $\mathbf{Z}$ obtidas através do método do Elipsóide de Volume Mínimo. Observações que possuem DR_i maior que $\sqrt{\chi^2_{(p-1);0,975}}$ são denominados *outliers* em x ou pontos de alavanca.

Estimadores com alto ponto de ruptura são úteis na detecção e desmascaramento de *outliers* e pontos de alavanca. Porém, Atkinson (1986) demonstrou que um bom ponto de alavanca pode ser classificado como sendo um *outlier* através do método da Mínima Mediana dos Quadrados do Resíduo. Desta forma, o autor sugeriu a utilização de métodos robustos que começam com uma análise exploratória e finalizam com uma análise confirmatória

seguindo o seguinte procedimento:

Seja I o conjunto de índices dos m *outliers* detectados pela abordagem da Mínima Mediana dos Quadrados do Resíduo, que será denominado grupo omitido. Para as $n_0 = n-m$ "boas" observações restantes do grupo $\bar{I}$, considera-se o mesmo modelo linear descrito em (4.2) com $\widehat{\beta}$ sendo a estimativa de Mínimos Quadrados para β, s^2 a estimativa não viesada de σ^2, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ a matriz chapéu de dimensões $n_0 \times n_0$ e h_i os elementos da diagonal dessa matriz. Atkinson (1986) sugeriu que fosse observado a que nível a estimação está sendo afetada pela retirada de uma nova observação i, $i \in \bar{I}$, ou pela recolocação de uma observação i, $i \in I$.

Com esse objetivo, o autor definiu os resíduos da retirada como:

$$t_i = e_i / s_{(i)}\sqrt{(1-h_i)}, i \in \bar{I},$$

e os resíduos da predição por

$$t_i^0 = e_i / s\sqrt{(1+d_i)}, i \in I, \tag{4.3}$$

em que $e_i = y_i - \mathbf{x_i^t}\widehat{\beta}$, $s^2_{(i)}$ é a estimativa da variância do erro através da amostra reduzida, excluindo a observação i, e $d_i = \mathbf{x_i^t}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}, \mathbf{i} \in I$. Atkinson (1986) sugere a representação destes dois tipos de resíduo em um único gráfico para confirmação dos *outliers*, denominado gráfico de re-adicionamento. Para a confirmação de observações influentes, o autor constrói o gráfico da estatística de Cook Modificada:

$$CM_i = (n_0 - p)h_i / p(1-h_i)^{1/2} \mid t_i \mid, i \in \bar{I},$$

e

$$CM_i^0 = (n_0 - p)d_i / p(1+d_i)^{1/2} \mid t_i^0 \mid, i \in I.$$

Nas próximas páginas serão examinadas a utilidade e limitações dos métodos citados por Rousseeuw e Van Zomeren (1990) e Atkinson (1986) através de exemplos. Sugere-se ainda uma nova abordagem confirmatória gradativa proposta por Fung (1993) e demonstra-se a utilidade desta nova abordagem.

Exemplo 1.a: Dados "Hawkins-Bradu-Kass": O banco de dados constituído pela Tabela 4 de Hawkins, Bradu e Kass (1984), já utilizado na seção anterior, consiste de 75 observações e três variáveis explicativas. As primeiras dez observações são consideradas *outliers* e também pontos de alavanca, e conforme apresentado por Rousseeuw e Van Zomeren (1990), seriam denominados de "maus" pontos de alavanca. As quatro observações seguintes são consideradas "bons" pontos de alavanca (e não *outliers*). Os autores identificaram estas 14 observações extremas corretamente através dos métodos da Mínima Mediana dos Quadrados do Resíduo e do Elipsóide de Volume Mínimo.

Fung (1993) observa que, se for seguida estritamente a proposta de Atkinson (1986) e colocada em prática a análise confirmatória através da retirada das dez primeiras observações identificadas pela regressão da Mínima Mediana dos Quadrados do Resíduo, conclui-se que apenas estas dez observações são *outliers* ou influentes. Numa etapa seguinte, após a realização da análise confirmatória excluindo-se as primeiras 14 observações, construiram-se os gráficos de re-adicionamento para os resíduos da retirada e foram calculadas as estatísticas de Cook Modificadas. As Figuras 4.5 e 4.6 apresentam respectivamente os valores dessas medidas para as 75 observações. É evidente que as mesmas 10 observações são confirmadas como *outliers* ou influentes. Porém, ainda não é possível classificar da mesma maneira as outras 4 observações, o que coincide com a análise confirmatória anterior. Assim, o procedimento aponta que apenas as 10 primeiras observações são discrepantes, apesar disto aparentemente não ser verdade.

Exemplo 2.a: Dados "Salinidade": Este exemplo engloba 28 medidas de salinidade e vazão de água em um rio do lago Pamlico, localizado na Carolina do Norte, Estados Unidos. O banco de dados, extraído da página 82 de Rousseeuw e Leroy (1987), consiste de três variáveis

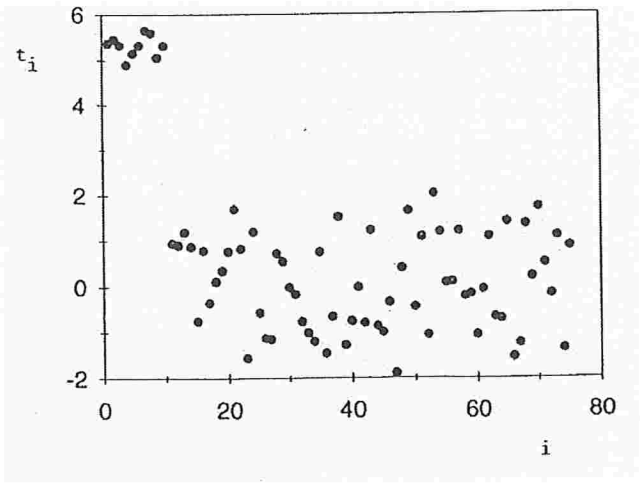

Figura 4.5: Gráfico de re-adicionamento para os resíduos da retirada t_i para os dados *Hawkins – Bradu – Kass*. (As observações excluídas são: $1, \ldots, 14$.)
Fonte: Fung (1993).

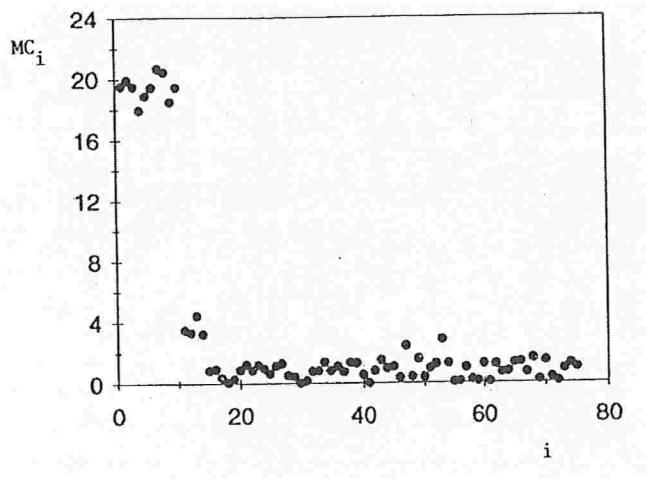

Figura 4.6: Gráfico de re-adicionamento de Atkinson para as estatísticas de Cook Modificadas CM_i para os dados *Hawkins – Bradu – Kass*. (As observações excluídas são: $1, \ldots, 14$.)
Fonte: Fung (1993).

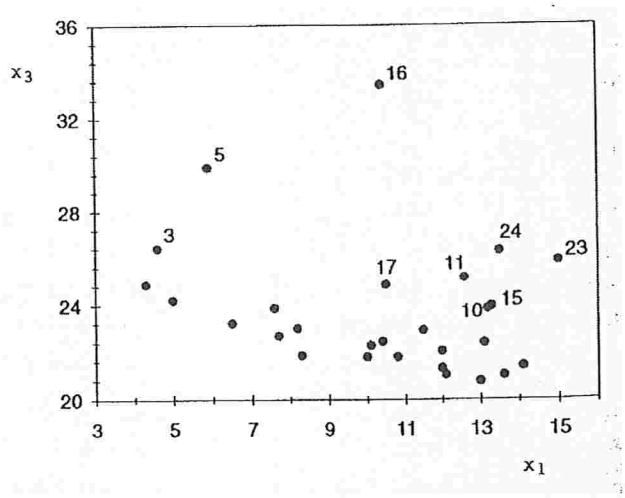

Figura 4.7: Gráfico de dispersão de x_3 contra x_1 para os dados *Salinidade*.Fonte: Fung (1993).

explicativas (x_1, x_2 e x_3). Neste exemplo são utilizados apenas x_1 e x_3 como ilustração, e o diagrama de dispersão dessas variáveis está apresentado na Figura 4.7.

Foram obtidos os resíduos robustos e as distâncias robustas DR_i de acordo com o procedimento da Mínima Mediana dos Quadrados do Resíduo e do Elipsóide de Volume Mínimo. A Figura 4.8 confronta os resíduos robustos contra as distâncias robustas DR_i, gráfico sugerido por Rousseeuw e Van Zomeren (1990). Se for aplicado estritamente o critério de corte sugerido pelos autores, dois *outliers*, cinco "bons" pontos de alavanca e três "maus" pontos de alavanca são identificados. Porém, pelo fato de não ser aconselhável o uso do critério de corte tão rigidamente, Fung (1993) suaviza o critério de classificação e identifica pontos com valores de corte maiores, fora do intervalo $[-5; +5]$ para $\widetilde{\varepsilon}_i$ e maior que 4 (que é maior que o percentil de ordem $99,9\%$ da distribuição qui-quadrado: $\sqrt{\chi^2_{2;0,999}} = 3,72$) para DR_i. Neste caso, seriam detectados apenas três "bons" pontos de alavanca (casos 5, 23 e 24) e um "mau" ponto de alavanca (caso 16).

Além disso, o autor considera que o gráfico da Figura 4.7 não sugere tantos pontos de alavanca e, portanto, o critério proposto por Rousseeuw e Van Zomeren (1990) estaria identificando um número excessivo de pontos de alavanca (e *outliers*).

Segundo o autor, tal fato é consequência da robustez dos procedimentos e portanto

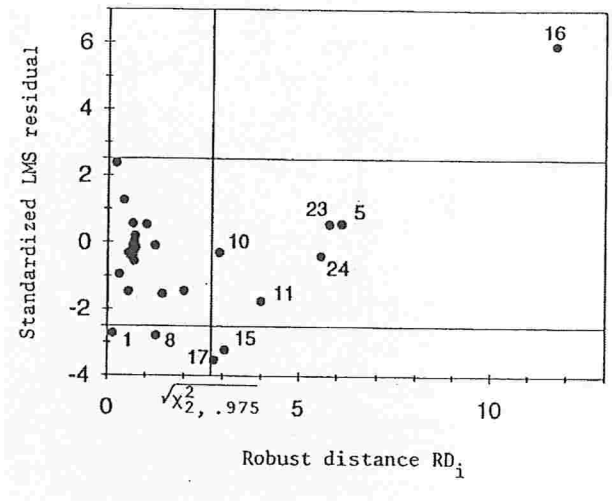

Figura 4.8: Gráfico dos resíduos robustos contra a Distância Robusta RD_i para os dados *Salinidade*(x_1, x_3).
Fonte: Fung (1993).

uma análise confirmatória parece ser necessária. Entretanto, como os exemplos mostram que a abordagem de Atkinson (1986) possui limitações, foi sugerida uma análise confirmatória alternativa.

O problema foi formulado como anteriormente (expressão (4.2)) e os objetivos da análise confirmatória seriam: (1) verificar se as observações do grupo omitido I são realmente *outliers*, e (2) verificar se as observações do grupo restante $\bar{I}$ estão de fato de acordo com o modelo ajustado.

O procedimento proposto consiste em:
Cada observação omitida i, $i \in I$, deve ser recolocada na amostra reduzida e considera-se o seguinte modelo, denominado modelo de re-adicionamento:

$$\mathbf{Y}_{+\mathbf{i}} = \mathbf{X}_{+\mathbf{i}}\beta_{\mathbf{i}} + \varepsilon_{+\mathbf{i}}, \mathbf{i} \in I,$$

que contém $n_0 + 1$ observações, incluindo aquelas do grupo $\bar{I}$ e a observação i, $i \in I$. Calculam-se então as medidas usuais para esta particular observação i sob esta formulação.

A medida "re-adicionada" de alavancagem é dada por:

$$h_{+i} = \mathbf{x_i^T}(\mathbf{X'_{+i}X_{+i}})^{-1}\mathbf{x_i}, \mathbf{i} \in I$$

e as demais medidas são definidas como:

$$t_{+i} = e_i/s\sqrt{(1+d_i)}, i \in I, \text{ que coincide com } \frac{d_i}{1+d_i},$$

e

$$CM_{+i} = (n_0 + 1 - p)h_{+i}/(p(1-h_{+i}))^{1/2} \mid t_{+i} \mid, i \in I.$$

Percebe-se que o resíduo é o mesmo que o encontrado na expressão (4.3), proposto por Atkinson (1986), porém obtido através de um raciocínio distinto. Por outro lado, verifica-se que a estatística de Cook Modificada de Atkinson "re-adicionada" pode ser escrita como $CM_i^0 = \{(n_0 - p)h_{+i}/p\}^{1/2} \mid t_{+i} \mid, i \in I$, que é diferente de CM_{+i}. Com base em suas expressões, é esperado que se comportem de maneiras bem distintas para valores grandes de h_{+i}.

A diferença ocorre basicamente porque Atkinson (1986) considerou a predição condicional em x_i, $i \in I$, já que $d_i = x_i^T(X'X)^{-1}x_i$, $i \in I$. Portanto, verificou-se que x_i ser *outlier* ou não é irrelevante no cálculo de CM_i^0. Isto pode ser evidenciado através do termo que aparece dentro das chaves da equação CM_i^0, que é limitado superiormente por $(n_0 - p)/p$. Dessa forma, CM_i^0, ao contrário de CM_i e CM_{+i}, não pode ser infinito. Ambos, CM_i e CM_{+i}, não são condicionalmente avaliados em x_i, $i \in I$, pois levam $\bar{I}$ em consideração. Assim, o gráfico de "re-adicionamento" de Atkinson que utiliza CM_i^0 pode apenas apontar os *outliers*, mas não os pontos de alavanca que a estatística de Cook Modificada CM_i pode também indicar. De acordo com Fung (1993), a medida CM_{+i} proposta não apresenta estes problemas, e pode, assim como CM_i, detectar *outliers*, pontos de alavanca e observações influentes.

Foram consideradas pelo autor apenas as medidas de "re-adicionamento" h_{+i}, t_{+i} e CM_{+i}. Outras medidas como a estatística de Cook e a estatística de Andrews-Pregibon (Belsey, Kuh e Welsch (1980)) poderiam também ser construídas sob o modelo "re-adicionado".

Definidas as medidas a serem utilizadas, a análise confirmatória proposta é realizada de modo *stepwise* (passo-a-passo). Com base nos gráficos do ponto "re-adicionado", retiram-se observações da amostra reduzida se estas forem *outliers* ou pontos de alavanca, e recolocam-se as observações que não são. O procedimento é repetido até que todas as observações omitidas sejam classificadas como *outliers* ou pontos de alavanca. Podem ser utilizados valores críticos para o teste da estatística t_i para detecção de *outliers* (Weisberg (1985), Tabela E) e a estatística de teste multinormal DM_i (Barnett e Lewis (1984), página 413) como pontos de corte. Pontos de corte para h_i podem ser encontrados através da relação $h_i = (DM_i)^2/(n_0 - 1) + 1/n_o$.

Sugere-se ainda que que estes pontos de corte sejam utilizados principalmente como referência e para comparação, e não de forma rígida.

Para finalizar, Fung (1993) analisa novamente os exemplos anteriores, seguindo o procedimento *stepwise* sugerido.

Exemplo1.b: Re-análise dos dados "Hawkins-Bradu-Kass": No exemplo 1.a foram identificados 14 *outliers* ou pontos de alavanca através dos métodos MMQR e EVM. Porém, o método de Atkinson falhou ao confirmar quatro pontos de alavanca: observações 11, 12, 13 e 14. Este banco de dados foi estudado por Fung (1993) através das medidas propostas na nova análise confirmatória. Gráficos do ponto "re-adicionado" para a medida de alavancagem e a estatística de Cook Modificada são apresentados nas Figuras 4.9 e 4.10. Observa-se claramente que as primeiras 14 observações são confirmadas como pontos influentes de alta alavancagem. Além disso, as 10 primeiras observações são também *outliers*, coincidindo com o que foi observado na Figura 4.5.

Exemplo 2.b: Dados "Salinidade": O método EVM identificou oito ou quatro (dependendo em um corte menor ou maior) pontos de alavanca para o caso de duas variáveis (X_1,X_3).

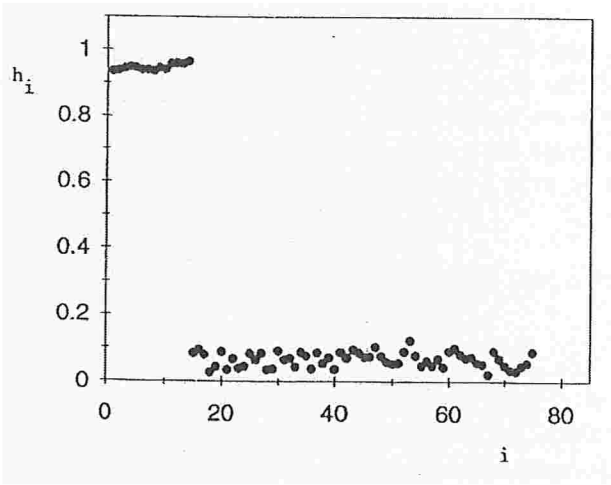

Figura 4.9: Gráfico de re-adicionamento para as medidas de alavancagem h_i para os dados *Hawkins – Bradu – Kass.* (As observações excluídas são: $1, \ldots, 14$.)
Fonte: Fung (1993).

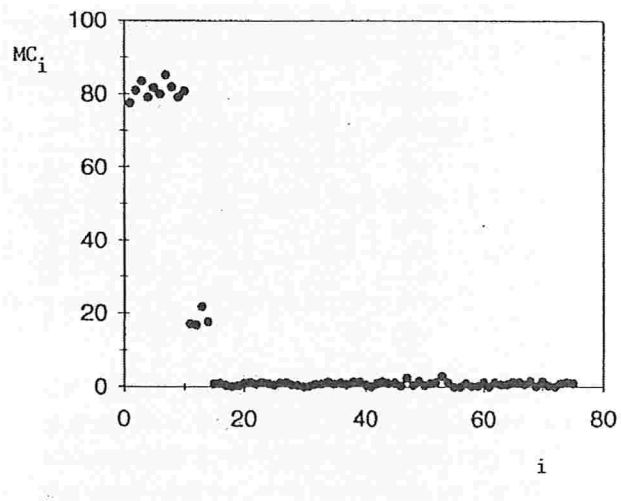

Figura 4.10: Gráfico de re-adicionamento para as estatísticas de Cook Modificadas CM_i para os dados *Hawkins – Bradu – Kass.* (As observações excluídas são: $1, \ldots, 14$.)
Fonte: Fung (1993).

Passo	Número da Observação								Ponto de referência 5%
	5	10	11	15	16	17	23	24	
1	0,77	0,44	0,58	0,46	0,92	0,41	0,75	0,73	0,51
2	0,58		0,31		0,81		0,48	0,46	0,47
3	0,44				0,68		0,29		0,44
4	0,41				0,62				0,43
5					0,51				0,42

Tabela 4.3: Medidas de alavancagem h_{+i} para os dados (X_1, X_3) de *Salinidade*.
Fonte: Fung (1993).

A análise confirmatória realizada considerou inicialmente as oito observações como sendo extremas e a Tabela 4.3 apresenta um resumo dos passos do procedimento. Analisando-se a Figura 4.11, que apresenta o gráfico do ponto "re-adicionado" para a medida de alavancagem, este resultado não se confirma. Na realidade, os valores de alavancagem h_{+i} das observações 10, 15 e 17 são menores do que o ponto de referência e o valor de alavancagem h_i da observação 3 do grupo considerado. Se estas três observações forem classificadas como pontos de alavanca, então assim também deveria ser classificada a observação 3. Além disso, os valores DR_i para estes três pontos, como observado na Figura 4.8, são apenas levemente maiores do que o valor de corte. Desta forma não se classificaria tais pontos como pontos de alavanca. Estas observações são colocadas de volta à amostra reduzida e um novo gráfico dos h_{+i} é construído. A Figura 4.12 mostra claramente que as observações 10, 15 e 17 não possuem grandes valores de alavancagem. Nota-se que os valores de h_{+i} das observações 11 e 24 são pequenos também. O procedimento *stepwise* continua recolocando as observações não-extremas (no caso, as observações 11 e 24) de volta à amostra reduzida. Por fim, obtém-se os valores de alavancagem na Figura 4.13, indicando que o único ponto de alavanca é a observação 16. A mesma conclusão resultou da análise confirmatória, que se iniciou ao utilizar quatro (ao invés de oito) observações como sendo extremas. Tal fato está em concordância com o gráfico de dispersão da Figura 4.7.

A análise da Tabela 4.3 também indica que a observação 5 é um ponto de alavancagem razoável.

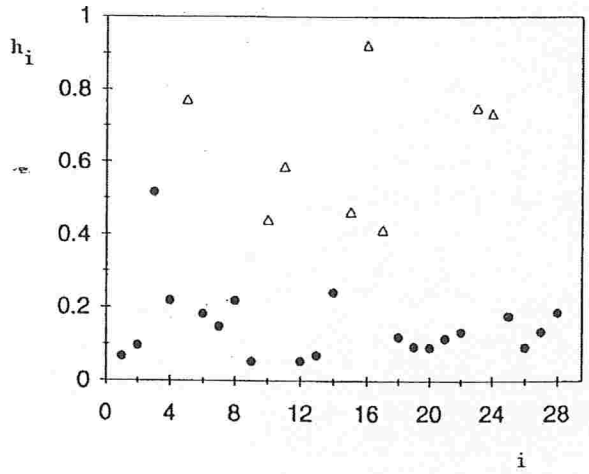

Figura 4.11: Gráfico de re-adicionamento para as medidas de alavancagem h_i para os dados *Salinidade*(x_1, x_3). (Observações excluídas: $5, 10, 11, 15, 16, 17, 23$ e 24 - marcadas com triângulos.) Fonte: Fung (1993).

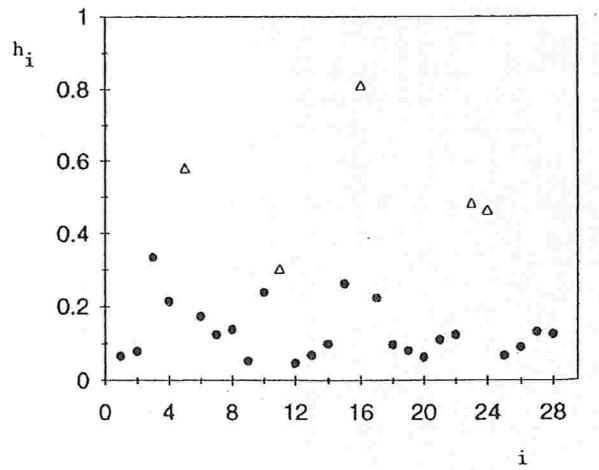

Figura 4.12: Gráfico de re-adicionamento para as medidas de alavancagem h_{+i} para os dados *Salinidade*(x_1, x_3). (Observações excluídas: $5, 11, 16, 23$ e 24 - marcadas com triângulos.) Fonte: Fung (1993).

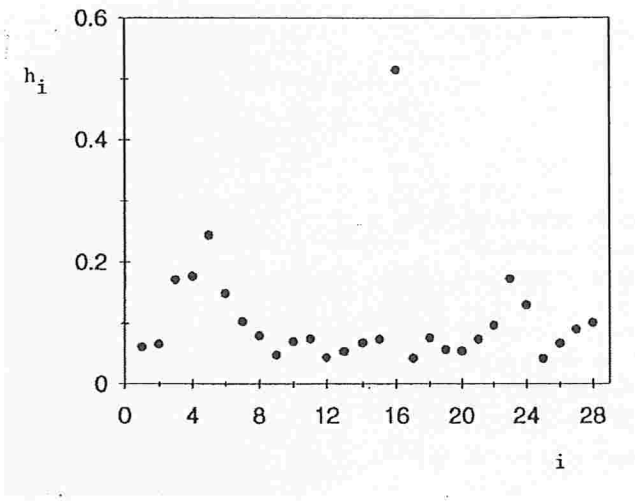

Figura 4.13: Gráfico de índice para as medidas de alavancagem para os dados *Salinidade*(x_1, x_3). Fonte: Fung (1993).

Uma análise confirmatória *stepwise* é também realizada para o caso de cinco *outliers* identificados através do método MMQR (utilizando corte de $\pm 2,5$). Conclui-se finalmente que a observação 16 é o único *outlier* do banco de dados (assim como o método MMQR detectou, com corte de ± 5).

Finalizando, o autor conclui que estimadores com alto ponto de ruptura como Mínima Mediana dos Quadrados do Resíduo e Elipsóide de Volume Mínimo podem resistir até a 50% da presença de *outliers* e pontos de alavanca nos dados. Porém, devido à sua baixa eficiência e tendência de detectar muitos *outliers* e pontos de alavanca, são mais indicados em análises exploratórias. Um segundo estágio, consistindo da análise *stepwise* utilizando medidas de diagnóstico para observações omitidas é recomendável na confirmação de *outliers* e pontos de alavanca.

4.4 Considerações Finais

Existem métodos robustos construídos com a finalidade de detectar *outliers* multivariados e que podem ser aplicados às variáveis independentes do modelo de regressão. Giroldo (2008) apresenta quatro destes métodos e os compara considerando os erros e acertos na classificação

das observações.

No próximo capítulo será apresentada a aplicação do diagnóstico robusto em análise de regressão, que foi realizada em dados reais utilizando uma das quatro técnicas estudadas por Giroldo (2008) em conjunto com a análise proposta por Rousseeuw e Van Zomeren (1990).

Capítulo 5

Aplicação em Dados Reais

O objetivo deste capítulo é aplicar uma das três técnicas apresentadas anteriormente. A técnica escolhida foi o Diagnóstico Robusto em Análise de Regressão. Para este fim, foi utilizado um subconjunto (apenas algumas variáveis) do conjunto de dados obtido no Centro de Estatística Aplicada (CEA) do Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP).

Trata-se do conjunto de dados pertencente ao Relatório de Análise Estatística sobre o projeto: "Efeito da poluição sobre as trocas gasosas de indivíduos jovens de Tibouchina pulchra Cogn. (*Melastomataceæ*) na região de Cubatão, SP" (Aubin, Elian e Alencar (1999)).

Este banco de dados consiste de um monitoramento ecológico (definido como qualquer método que faz uso de reações da vida para identificar ou caracterizar mudanças nas condições ambientais induzidas pela ação humana) realizado na região de Cubatão em SP, que buscava avaliar o impacto da poluição aérea sobre as trocas gasosas de CO_2 (fotossíntese) em regiões com excesso de poluição. A espécie escolhida para a análise foi a *Tibouchinapulchra*, conhecida como manacá-da-serra, da família *Melastomataceæ*, devido à abundância com que aparece na Serra do Mar. Foi realizada uma análise de regressão da fotossíntese (variável resposta) em função das variáveis flureto, nitrogênio e enxofre (variáveis independentes),

uma vez que espera-se que estas variáveis estejam relacionadas à poluição.

Descrição do Experimento: O experimento original completo não será descrito aqui, vamos nos ater somente à parte do mesmo e às variáveis envolvidas na análise de regressão. Foram colocadas 6 mudas de plantas jovens de Tibouchina pulchra, com idade e tamanho semelhantes espalhados na região de Cubatão. A exposição nas áreas foi feita em viveiros construídos em alumínio e cobertos com sombrite de 50% (tela que permite a passagem de somente 50% da luminosidade total do ambiente) para evitar a luminosidade muito forte. Quinzenalmente as plantas eram adubadas com 150ml de solução nutriente e o suprimento de água era feito através de processos específicos e perfeitamente controlados. Este procedimento garantiu a uniformidade de condições de nutrientes, água, luz e solo para todas as plantas. Após 3 semanas do plantio as plantas receberam etiquetas de identificação e tiveram medidos a altura e o diâmetro. No final do período de exposição, as plantas foram levadas para a área de referência (RP), onde permaneceram por 24hs para aclimatação. A fotossíntese foi medida por um aparelho que determinava as trocas gasosas de gás carbônico e água nas folhas. Posteriormente, em laboratório, foram obtidas as medidas de fluoreto, nitrogênio e enxofre, medidos a partir das folhas desidratadas.

Descrição das Variáveis: foram observadas diversas variáveis, somente as variáveis utilizadas neste estudo estão descritas a seguir.

Variável que mede a intensidade das trocas gasosas de CO_2:

- Fotossíntese ($\mu molCO_2/(m^2 s)$).

Variáveis poluentes:

- Fluoreto (μg);
- Nitrogênio (mg);
- Enxofre (mg).

A Tabela 5.1 apresenta as variáveis acima relacionadas em seu formato original (36 observações). Como não foi possível a coleta de todas as variáveis, utilizou-se o banco de

dados desconsiderando as observações faltantes, totalizando 30 observações (Veja Tabela 5.2).

Apresentamos a seguir a análise de diagnóstico robusto para o modelo de regressão da variável fotossíntese em função das variáveis flureto, nitrogênio e enxofre. Toda a análise estatística foi realizada através do *software* R 2.7.2 (R Development Core Team (2007)).

Inicialmente, realizou-se a Análise de Diagnóstico Clássico considerando a *fotossintese* como variável resposta e as variáveis *fluoreto*, *nitrogenio* e *enxofre* como variáveis independentes. Desta forma, considerou-se inicialmente o seguinte modelo:

$$fotossintese_i = \beta_0 + \beta_1 \cdot fluoreto_i + \beta_2 \cdot nitrogenio_i + \beta_3 \cdot enxofre_i + \varepsilon, i = 1, \cdots, 30. \quad (5.1)$$

com as suposições $\varepsilon \sim N(0, \sigma^2)$ e de independência entre as observações.

O resultado do ajuste do modelo (5.1) é apresentado na Tabela 5.3. Esta tabela mostra que a variável *fluoreto* é não significante. A equação (5.2) representa o ajuste do modelo excluindo-se esta variável.

$$fotossintese_i = \beta_0 + \beta_1 \cdot nitrogenio_i + \beta_2 \cdot enxofre_i + \varepsilon, i = 1, \cdots, 30. \quad (5.2)$$

A Tabela 5.4 mostra o resumo do ajuste do modelo de regressão da variável *fotossintese* em função das variáveis *nitrogenio* e *enxofre*. Através desta tabela obtém-se a equação de regressão dada por (5.3).

$$\widehat{fotossintese_i} = 10,57 - 0,47 \cdot nitrogenio_i + 1,28 \cdot enxofre_i, i = 1, \cdots, 30. \quad (5.3)$$

Com o objetivo de aplicação de técnicas descritas no capítulo anterior, foi feita uma análise de diagnóstico para o modelo (5.3). Realizou-se inicialmente a análise de diagnóstico clássica, para posterior comparação com a correspondente análise robusta.

A Figura 5.1 destaca os pontos 14 e 27 como pontos de alavanca, o ponto 14 como

Observação	Fotossintese	Fluoreto	Nitrogenio	Enxofre
1	15,40	18,93	18,11	7,5
2	12,70	17,18	20,11	7,0
3	13,15	20,07	19,63	7,8
4	10,44	18,91	18,11	6,0
5	13,69	21,36	18,11	7,3
6	18,99	13,61	21,11	7,5
7	8,49	8,78	21,63	7,5
8	6,97	11,11	18,11	7,0
9	7,93	7,58	22,15	6,9
10	8,26	9,07	18,11	6,4
11	6,05	7,83	22,15	6,5
12	7,26	7,50	20,63	6,2
13	9,97	19,00	23,15	7,2
14	8,91	17,95	27,15	11,1
15	10,19	25,00	25,15	9,8
16	12,39	21,02	23,15	9,6
17	13,36	19,75	26,15	7,4
18	10,23	20,38	26,67	7,7
19	9,94	14,98	24,63	6,7
20	6,39	13,00	23,15	7,7
21	6,56	15,64	25,15	7,3
22	8,10	9,88	21,63	7,5
23	6,40	11,82	25,67	7,0
24	4,12	10,49	22,63	7,7
25	9,56	105,91	32,19	8,0
26	7,13	118,88	30,67	7,7
27	10,37	84,30	34,19	10,6
28	6,02	102,71	NA	NA
29	5,39	96,96	NA	NA
30	5,72	69,80	NA	NA
31	4,31	51,57	27,15	6,7
32	NA	86,82	33,15	7,4
33	3,51	90,34	28,15	7,3
34	2,36	87,50	33,70	8,3
35	6,77	65,15	NA	NA
36	NA	68,78	NA	NA

Tabela 5.1: Medidas coletadas de *fotossintese*, *fluoreto*, *nitrogenio* e *enxofre* para as 36 plantas.

Observação	Fotossintese	Fluoreto	Nitrogenio	Enxofre
1	15,40	18,93	18,11	7,5
2	12,70	17,18	20,11	7,0
3	13,15	20,07	19,63	7,8
4	10,44	18,91	18,11	6,0
5	13,69	21,36	18,11	7,3
6	18,99	13,61	21,11	7,5
7	8,49	8,78	21,63	7,5
8	6,97	11,11	18,11	7,0
9	7,93	7,58	22,15	6,9
10	8,26	9,07	18,11	6,4
11	6,05	7,83	22,15	6,5
12	7,26	7,50	20,63	6,2
13	9,97	19,00	23,15	7,2
14	8,91	17,95	27,15	11,1
15	10,19	25,00	25,15	9,8
16	12,39	21,02	23,15	9,6
17	13,36	19,75	26,15	7,4
18	10,23	20,38	26,67	7,7
19	9,94	14,98	24,63	6,7
20	6,39	13,00	23,15	7,7
21	6,56	15,64	25,15	7,3
22	8,10	9,88	21,63	7,5
23	6,40	11,82	25,67	7,0
24	4,12	10,49	22,63	7,7
25	9,56	105,91	32,19	8,0
26	7,13	118,88	30,67	7,7
27	10,37	84,30	34,19	10,6
31	4,31	51,57	27,15	6,7
33	3,51	90,34	28,15	7,3
34	2,36	87,50	33,70	8,3

Tabela 5.2: Medidas coletadas de *fotossintese*, *fluoreto*, *nitrogenio* e enxofre para as 30 plantas.

ponto influente e o ponto 6 como aberrante (resíduos padronizados fora do intervalo [-2;2]). Não há indícios de heterocedasticidade.

O Diagnóstico Robusto iniciou-se através da identificação de possíveis *outliers* existentes no banco de dados. O método escolhido foi o *ForwardSearch* (Atkinson e Riani (2004)), selecionado entre 4 métodos estudados por Giroldo (2008). O método é originalmente proposto para determinar *outliers* multivariados e será aplicado para detectar pontos

Coeficientes	Estimativa	Desvio-Padrão	Estatística t	Nível descritivo
(Intercepto)	12,5310	4,9156	2,5490	*0, 0170*
fluoreto	0,0238	0,0314	0,7590	0,4546
nitrogenio	-0,6126	0,2437	-2,5140	*0, 0185*
enxofre	1,3798	0,6028	2,2890	*0, 0304*

Erro padrão do resíduo: 3,288 com 26 g.l.

$R^2 = 0,2877$; $R^2(Ajustado) = 0,2055$

Estatística F: 3,5 com 3 e 26 g.l.; Nível descritivo = 0,02954

Tabela 5.3: Análise de regressão cosiderando as variáveis independentes *fluoreto*, *nitrogenio* e *enxofre* e a variável resposta *fotossintese*.

Coeficientes	Estimativa	Desvio-Padrão	Estatística t	Nível descritivo
(Intercepto)	10,5724	4,1510	2,547	*0, 01688*
nitrogenio	-0,4688	0,1520	-3,085	*0, 00466*
enxofre	1,2787	0,5832	2,192	*0, 03715*

Erro padrão do resíduo: 3,262 com 27 g.l.

$R^2 = 0,27197$; $R^2(Ajustado) = 0,2179$

Estatística F: 5,041 com 2 e 27 g.l.; Nível descritivo = 0,01380

Tabela 5.4: Análise de regressão considerando as variáveis independentes *nitrogenio* e *enxofre* e a variável resposta *fotossintese*.

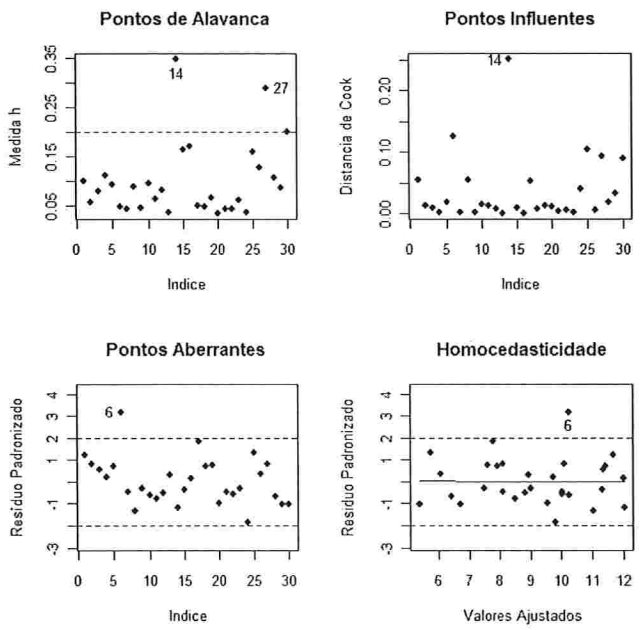

Figura 5.1: Continuação da Análise Clássica de Resíduos para o modelo (5.3).

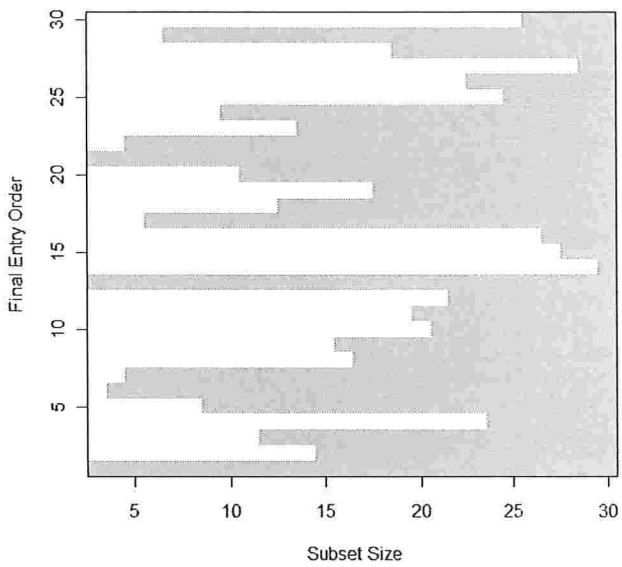

Figura 5.2: Ordem de entrada das observações considerando o modelo (5.3).

discrepantes nas variáveis independentes nitrogênio e enxofre. Este método é um método iterativo robusto, em que inicialmente escolhe-se o conjunto inicial de observações, que no caso do banco de dados utilizado - contendo 30 observações - possui tamanho 3 e é composto pelas observações 1, 13 e 21. O conjunto inicial é escolhido por meio de *boxplots* bivariados (Zani, Riani e Corbellini (1998)). Trata-se de um método gráfico que se baseia na distância de Mahalanobis e pode ser utilizado para detectar *outliers* (o que requer que os dados tenham distribuição aproximadamente Normal) e encontrar grupos de observações semelhantes nos dados. O método está completamente descrito em Giroldo (2008, pág. 24).

Na Figura 5.2, o eixo das ordenadas representa o número da observação e as abcissas informam sobre a ordem de entrada das observações através do comprimento do retângulo branco, que indica o momento em que a observação é incluída no subconjunto. Assim, no primeiro passo, a observação incluída é a 6, no segundo passo é a 7, no terceiro passo é a 22 e assim por diante, sendo que o tamanho do subconjunto passa a ser 4, 5 e 6 com a entrada desses pontos, respectivamente. A Tabela 5.5 traz esta informação para cada passo

Passo	Observação Incluída	Tamanho do subconjunto
2	6	4
3	7	5
4	22	6
5	17	7
6	33	8
7	5	9
8	24	10
9	20	11
10	3	12
11	18	13
12	23	14
13	2	15
14	9	16
15	8	17
16	19	18
17	31	19
18	11	20
19	10	21
20	12	22
21	26	23
22	4	24
23	25	25
24	34	26
25	16	27
26	15	28
27	27	29
28	14	30

Tabela 5.5: Obsevação incluída na amostra em cada passo do método de Atkinson e Riani (2004) considerando o modelo (5.3).

do processo iterativo.

Além disso, também é possível verificar como se comporta a distância de Mahalanobis de cada observação em cada passo do método. O resultado deste estudo encontra-se na Figura 5.3. As curvas presentes neste gráfico correspondem aos valores da distância de Mahalanobis para cada passo do procedimento, definido pelo tamanho da amostra, que encontra-se no eixo das abcissas. Cada curva é associada a uma observação do conjunto de dados. A partir desta figura, é possível verificar que as observações 14 e 27 são as que

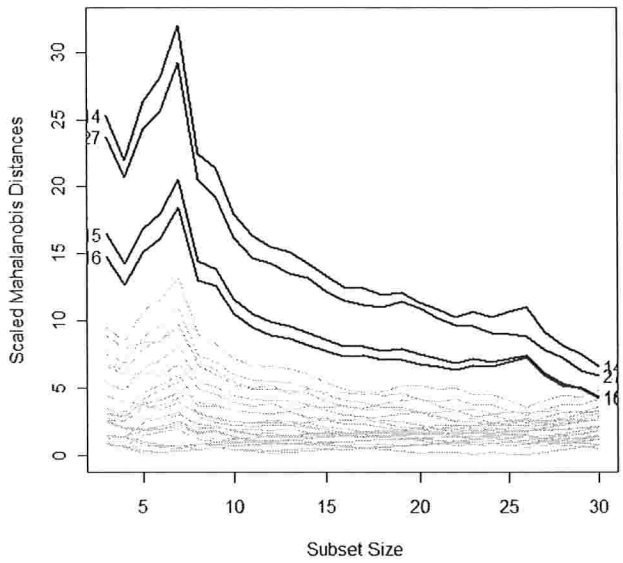

Figura 5.3: Distância de Mahalanobis para o modelo (5.3).

possuem as maiores distâncias de Mahalanobis, seguidas pelas observações 16 e 15. Estas observações são as que possuem as maiores distâncias de Mahalanobis em quase todo o processo iterativo, e por isso podem ser consideradas atípicas pelo método de Atkinson e Riani (2004).

O elipsóide de volume mínimo citado na Seção 4.2 foi obtido através da função 'covfmEllipsesPlot()' do *software R*. Pode-se observar, através da Figura 5.4, que o elipsóide clássico classifica duas observações como *outliers*: 14 e 27. Já o elipsóide robusto classifica além das duas anteriores, também as observações: 4, 15, 16 e 34 como *outliers*. O critério utilizado para classificação das observações como *outliers* foi adotar a confiança de 95% para os elipsóides, ou seja, foi classificada como *outlier* a observação cuja distância foi superior ao valor $\sqrt{\chi^2_{p;0,95}} = \sqrt{5,991} = 2,448$, onde $p = 2$ (variáveis *nitrogenio* e *enxofre*). Este gráfico compara a distância de Mahalanobis clássica (representada pela linha tracejada e obtida utilizando-se as matrizes de média e de covariância amostral dos dados) com a distância de Mahalanobis robusta (representada pela linha tracejada e obtida utilizando-se os estimadores

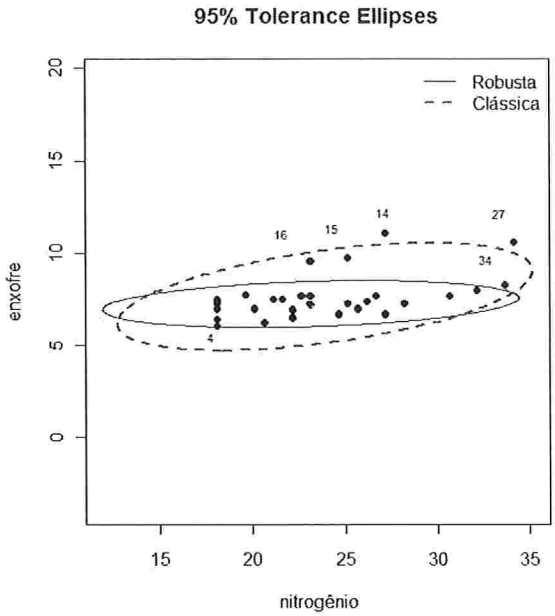

Figura 5.4: Elipsóides Clássico e Robusto das variáveis *nitrogenio* e *enxofre* - modelo (5.3).

EVM de locação e escala) para as 30 plantas. A implementação dos elipsóides encontra-se no Apêndice B.

Para a construção dos elipsóides foi necessário o cálculo das distâncias clássica e robusta dos dados, apresentados na Tabela 5.6. Confrontando estas distâncias em um gráfico (Figura 5.5), observa-se claramente que apenas quatro observações detectadas como *outliers* através do elipsóide robusto localizam-se bem distante dos demais: 14, 15, 16 e 27. Já as observações 4 e 34 localizam-se próximas à massa dos dados. Ao observarmos novamente o elipsóide robusto na Figura 5.4, percebe-se que estes dois últimos pontos estão localizados bem próximos à fronteira do elipsóide. Neste gráfico, adotou-se uma confiança de 97,5% como critério para classficar uma observação como *outlier*. Desta forma, foi classificado como *outlier* a observação cuja distância de Mahalanobis foi superior ao valor $\sqrt{\chi^2_{p;0,975}} = \sqrt{7,378} = 2,716$, onde p=2.

A Figura 5.6 exibe o gráfico dos Resíduos Padronizados de Mínimos Quadrados contra a Distância de Mahalanobis Clássica, discutido na Seção 4.2. Percebe-se que apenas

Observação	DM_i^2	DR_i^2
1	1,9748	1,8329
2	0,7104	0,5297
3	1,3748	2,2483
4	2,3117	5,7313
5	1,7787	1,3682
6	0,4321	0,5823
7	0,2799	0,4578
8	1,6235	1,2391
9	0,3807	0,4208
10	1,8139	3,0257
11	0,8975	1,9755
12	1,4441	3,8529
13	0,1288	0,0070
14	9,1268	53,6467
15	3,8258	23,8043
16	4,0147	21,0365
17	0,4782	0,4346
18	0,4268	1,0459
19	0,9858	1,4044
20	0,0575	0,7910
21	0,2888	0,1840
22	0,2799	0,4578
23	0,8182	0,6760
24	0,1353	0,8592
25	3,6727	4,7296
26	2,7399	2,8517
27	7,4471	40,4517
31	2,1689	2,4030
33	1,5502	1,1975
34	4,8326	7,3086

Tabela 5.6: Tabela contendo o quadrado das Distâncias Clássica e Robusta dos dados - modelo (5.3).

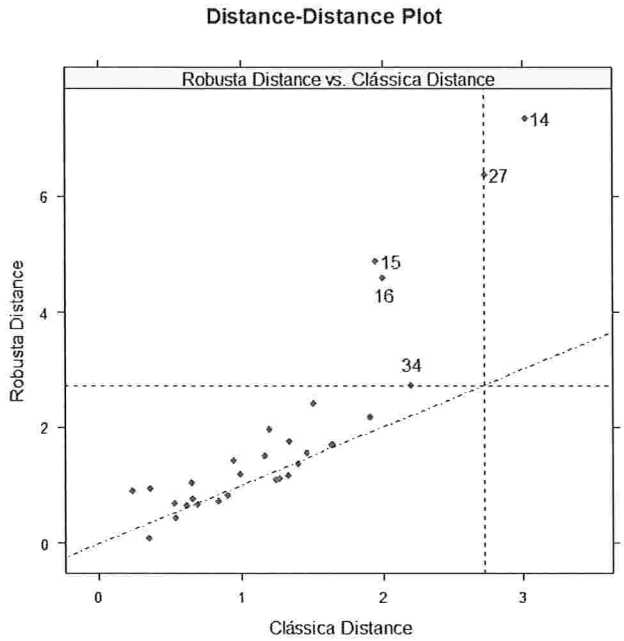

Figura 5.5: Distância Robusta contra Distância Clássica - modelo (5.3).

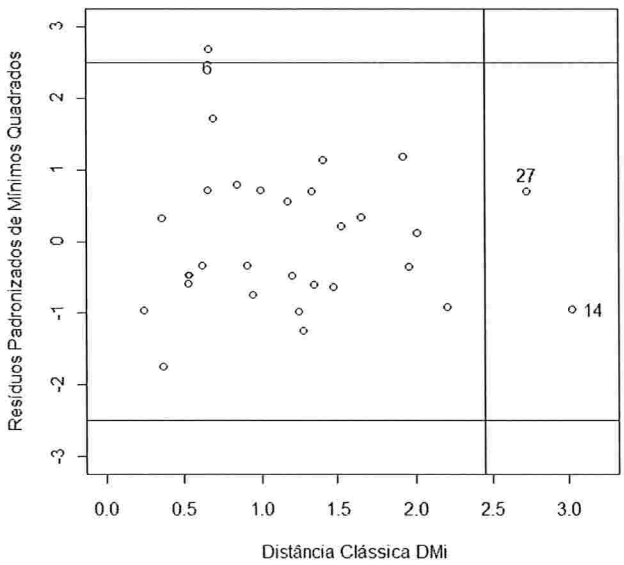

Figura 5.6: Distância Clássica contra Resíduos Padronizados de Mínimos Quadrados - modelo (5.3).

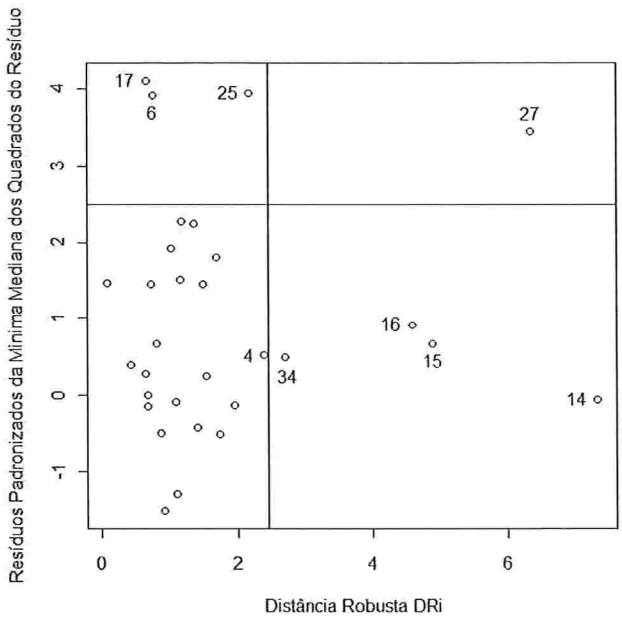

Figura 5.7: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.3).

as observações 6, 14 e 27 encontram-se fora dos limites que identificam pontos de alavanca ou *outliers* (as duas últimas foram identificadas também pelo elipsóide clássico). O mesmo critério utilizado para a detecção dos *outliers* nos elipsóides foi adotado aqui: o de considerar as observações cuja distância é maior do que $\sqrt{\chi^2_{p;0,95}} = \sqrt{5,991} = 2,448$, onde p=2 variáveis (*nitrogenio* e *enxofre*).

Já na Figura 5.7, é possível observar que quando confrontam-se os Resíduos Padronizados da Mínima Mediana ao Quadrado (Resíduos Robustos) contra a Distância de Mahalanobis Robusta, detectam-se muito mais do que apenas três pontos de alavanca ou *outliers*, são eles: observações 6, 17 e 25 (antes não identificados pelo elipsóide robusto) e as mesmas identificadas pelo elipsóide robusto: 4, 14, 15, 16, 27 e 34. O critério para determinar os *outliers* foi o mesmo definido no gráfico anterior, e a implementação destes dois últimos gráficos citados (Figuras 5.6 e 5.7) também encontram-se no Apêndice B.

Além disso, este último gráfico (Figura 5.7) apresenta uma classificação visual dos dados em quatro categorias: (1) observações regulares com baixos valores de DR_i e $r_i/\widehat{\sigma}$, (2)

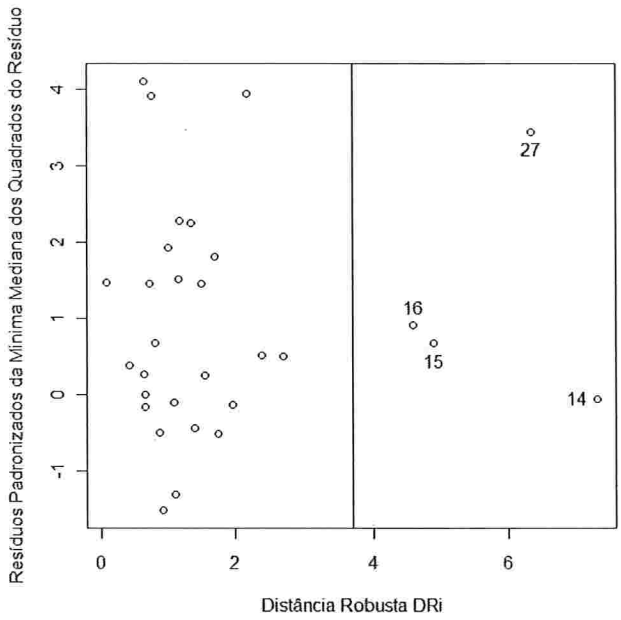

Figura 5.8: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.3).

os *outliers* verticais com baixos valores de DR_i e altos valores de $r_i/\widehat{\sigma}$, (3) os "bons" pontos de alavanca, com altos valores de DR_i e baixos valores de $r_i/\widehat{\sigma}$ e por último, (4) os "maus" pontos de alavanca, com altos valores de DR_i e $r_i/\widehat{\sigma}$. As observações são então distribuídas da seguinte maneira: em (1): todas as observações, com exceção das observações classificadas em (2), (3) e (4); em (2): 6, 17 e 25; em (3): 14, 15, 16 e 34; e em (4): 27. Seguindo a sugestão dada por Fung (1993), que suaviza o critério de classificação adotando valores de corte maiores e percentis de ordem acima do utilizado, pode-se adotar o intervalo [-5;+5] e o percentil de ordem 99,9% da distribuição Qui-Quadrado ($\sqrt{\chi^2_{2;0,999}} = 3,72$). Desta maneira, seriam identificados apenas as observações 14, 15, 16 e 27 como "bons" pontos de alavanca (Observe a Figura 5.8). Caracterizando esses quatro pontos, percebe-se que são as observações que apresentam os quatro maiores valores para a variável *enxofre*.

É importante notar que o diagnóstico não deve ser realizado observando-se apenas um dos gráficos estudados.

Uma vez que as observações 14, 15, 16 e 27 foram identificadas anteriormente como

Observação	Fotossintese	Fluoreto	Nitrogenio	Enxofre
1	15,40	18,93	18,11	7,5
2	12,70	17,18	20,11	7,0
3	13,15	20,07	19,63	7,8
4	10,44	18,91	18,11	6,0
5	13,69	21,36	18,11	7,3
6	18,99	13,61	21,11	7,5
7	8,49	8,78	21,63	7,5
8	6,97	11,11	18,11	7,0
9	7,93	7,58	22,15	6,9
10	8,26	9,07	18,11	6,4
11	6,05	7,83	22,15	6,5
12	7,26	7,50	20,63	6,2
13	9,97	19,00	23,15	7,2
17	13,36	19,75	26,15	7,4
18	10,23	20,38	26,67	7,7
19	9,94	14,98	24,63	6,7
20	6,39	13,00	23,15	7,7
21	6,56	15,64	25,15	7,3
22	8,10	9,88	21,63	7,5
23	6,40	11,82	25,67	7,0
24	4,12	10,49	22,63	7,7
25	9,56	105,91	32,19	8,0
26	7,13	118,88	30,67	7,7
31	4,31	51,57	27,15	6,7
33	3,51	90,34	28,15	7,3
34	2,36	87,50	33,70	8,3

Tabela 5.7: Medidas coletadas de *fotossintese*, *fluoreto*, *nitrogenio* e enxofre para as 26 plantas.

sendo possíveis *outliers*, é interesante que a análise seja refeita excluindo-se esses pontos. A tabela 5.7 contem apenas as 26 observações restantes.

A análise apresentada a seguir refere-se ao diagnóstico robusto para o modelo de regressão da variável fotossíntese em função das variáveis nitrogênio e enxofre para o banco de dados inicial excluindo-se as observações 14, 15, 16 e 27. Desta maneira, esse novo banco de dados constitui-se de 26 observações. O resumo do ajuste do modelo de regressão da *fotossintese* em função do *nitrogenio* e do *enxofre* pode ser observado na Tabela 5.8.

Adotando-se um nível de significância de 1%, verifica-se que tanto *nitrogenio* quanto

Coeficientes	Estimativa	Desvio-Padrão	Estatística t	Nível descritivo
(Intercepto)	4,2327	8,7481	0,63307	0,63307
nitrogenio	-0,5689	0,1772	0,00388	*0, 00388*
enxofre	2,4893	1,3932	1,787	0,08716

Erro padrão do resíduo: 3,375 com 23 g.l.

$R^2 = 0,3102$; $R^2(Ajustado) = 0,2502$

Estatística F: 5,171 com 2 e 23 g.l.; Nível descritivo = 0,01397

Tabela 5.8: Análise de regressão considerando as variáveis independentes *nitrogenio* e *enxofre* e a variável resposta *fotossintese* para o banco de dados contendo 26 observações.

enxofre são significantes e a equação de regressão é dada por (5.4).

$$\widehat{fotossintese}_i = 4,23 - 0,57 \cdot nitrogenio_i + 2,49 \cdot enxofre_i, i = 1, \cdots, 26. \quad (5.4)$$

Observa-se uma grande diferença entre a equação (5.4) e a obtida com as 30 observações (equação (5.3)). Iniciou-se então a análise de diagnóstico para o modelo (5.4). Começando pela análise de diagnóstico clássica, percebe-se através da Figura 5.9 que o ponto 26 destaca-se como ponto de alavanca, os pontos 6 e 26 como pontos influentes e o ponto 6 como aberrante (resíduos padronizados fora do intervalo [-2;2]). Não há indícios de heterocedasticidade.

O Diagnóstico Robusto iniciou-se através da identificação de possíveis *outliers* existentes no banco de dados contendo apenas 26 observações. O método *ForwardSearch* (Atkinson e Riani (2004)), selecionou o conjunto inicial de observações, que no caso do banco de dados utilizado - contendo 26 observações - possui tamanho 3 e é composto pelas linhas 13, 18 e 20 dessa tabela cujas observações são: 13, 21 e 23. Vale lembrar novamente que o método está completamente descrito em Giroldo (2008, pág. 24).

Na Figura 5.10, o eixo das ordenadas representa o número da observação e as abcissas informam sobre a ordem de entrada das observações através do comprimento do retângulo branco, que indica o momento em que a observação é incluída no subconjunto. Assim, no primeiro passo, a observação incluída é a 17, no segundo passo é a 33, no terceiro passo é

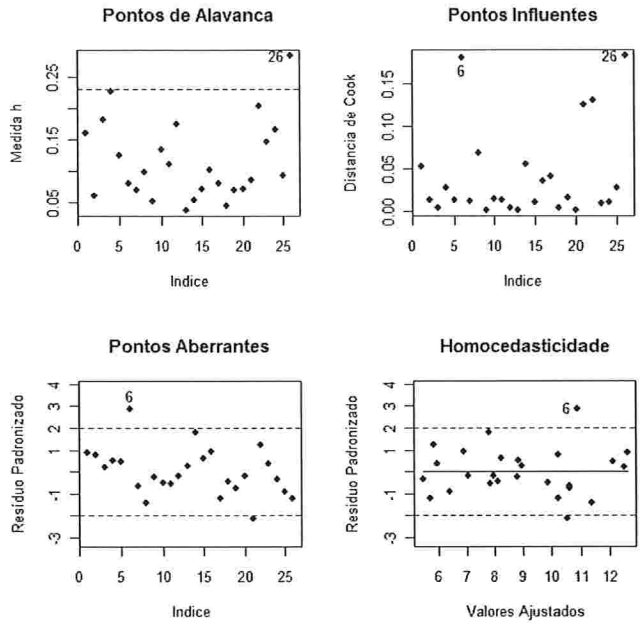

Figura 5.9: Continuação da Análise Clássica de Resíduos para o modelo (5.4).

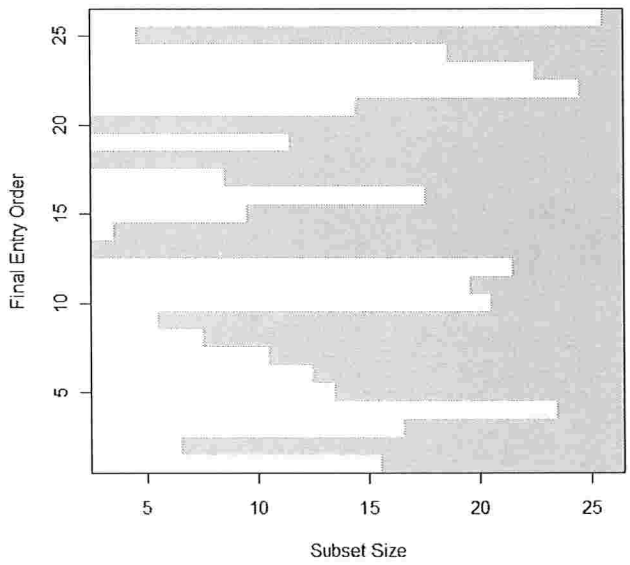

Figura 5.10: Ordem de entrada das observações considerando o modelo (5.4).

Passo	Observação Incluída	Tamanho do subconjunto
2	17	4
3	33	5
4	9	6
5	2	7
6	8	8
7	20	9
8	18	10
9	7	11
10	22	12
11	6	13
12	5	14
13	24	15
14	1	16
15	3	17
16	19	18
17	31	19
18	11	20
19	10	21
20	12	22
21	26	23
22	4	24
23	25	25
24	34	26

Tabela 5.9: Obsevação incluída na amostra em cada passo do método de Atkinson e Riani (2004) considerando o modelo (5.4).

a 9 e assim por diante, sendo que o tamanho do subconjunto passa a ser 4, 5 e 6 com a entrada desses pontos, respectivamente. A Tabela 5.9 traz esta informação para cada passo do processo iterativo.

Além disso, também é possível verificar como se comporta a distância de Mahalanobis de cada observação em cada passo do método. O resultado deste estudo encontra-se na Figura 5.11. As curvas presentes neste gráfico correspondem aos valores da distância de Mahalanobis para cada passo do procedimento definido pelo tamanho da amostra, que encontra-se no eixo das abcissas. Cada curva é associada a uma observação do conjunto de dados. A partir desta figura, é possível constatar que não existe nenhuma observação que apresente valores altos para a distância de Mahalanobis em todo o processo, e por isso não

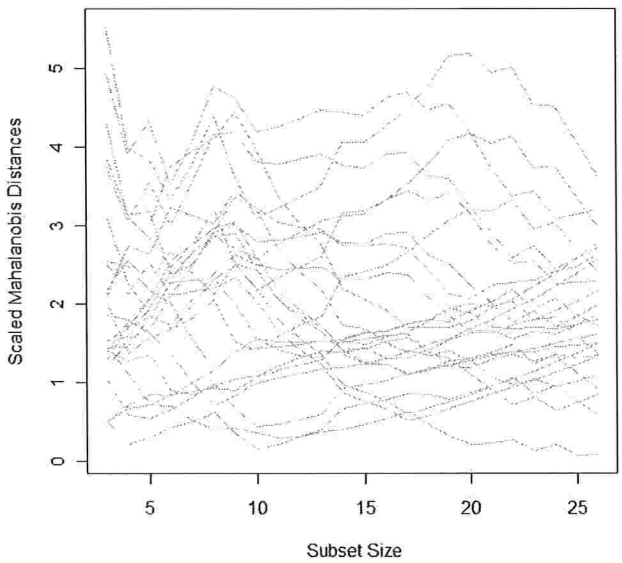

Figura 5.11: Distância de Mahalanobis para o modelo (5.4).

é possível identificar nenhuma observação atípica pelo método de Atkinson e Riani (2004).

O elipsóide de volume mínimo citado na Seção 4.2 foi obtido através da função 'covfmEllipsesPlot()' do *software R*. Pode-se observar, através da Figura 5.12, que o elipsóide clássico classifica apenas a observação 34 como *outlier*. Já o elipsóide robusto classifica além da observação 34, também a observação 25. O critério utilizado para classificação das observações como *outliers* foi adotar a confiança de 95% para os elipsóides, ou seja, foi classificada como *outlier* a observação cuja distância foi superior ao valor $\sqrt{\chi^2_{p;0,95}} = \sqrt{5,991} = 2,448$, onde $p = 2$ (variáveis *nitrogenio* e *enxofre*). Este gráfico compara a distância de Mahalanobis clássica (representada pela linha tracejada e obtida utilizando-se as matrizes de média e de covariância amostral dos dados) com a distância de Mahalanobis robusta (representada pela linha tracejada e obtida utilizando-se os estimadores EVM de locação e escala) para as 26 plantas. A implementação dos elipsóides encontra-se no Apêndice B.

Para a construção dos elipsóides foi necessário o cálculo das distâncias clássica e ro-

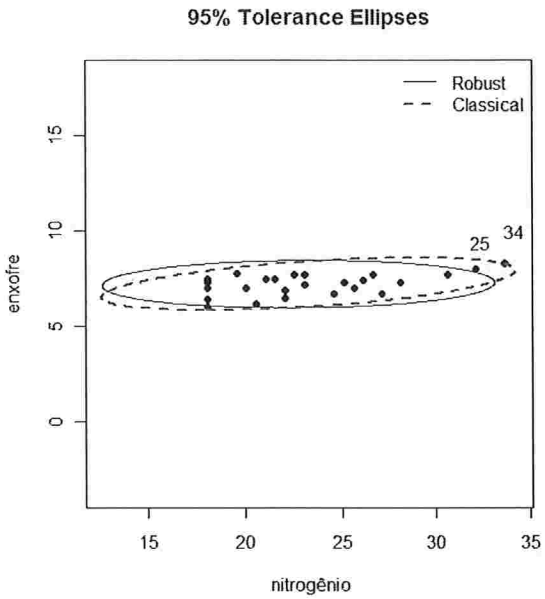

Figura 5.12: Elipsóides Clássico e Robusto das variáveis *nitrogenio* e *enxofre* - modelo (5.4).

busta dos dados, apresentados na Tabela 5.10. Confrontando estas distâncias em um gráfico (Figura 5.13), observa-se claramente que apenas a observação 34 localiza-se bem distante das demais. Neste gráfico, adotou-se uma confiança de 97,5% como critério para classficar uma observação como *outlier*. Desta forma, foi classificado como *outlier* a observação cuja distância de Mahalanobis foi superior ao valor $\sqrt{\chi^2_{p;0,975}} = \sqrt{7,378} = 2,716$, onde p=2.

A Figura 5.14 exibe o gráfico dos Resíduos Padronizados de Mínimos Quadrados contra a Distância de Mahalanobis Clássica, discutido na Seção 4.2. Percebe-se que apenas a observação 34 encontram-se fora dos limites que identificam pontos de alavanca ou *outliers*. O mesmo critério utilizado para a detecção dos *outliers* nos elipsóides foi adotado aqui: o de considerar as observações cuja distância é maior do que $\sqrt{\chi^2_{p;0,95}} = \sqrt{5,991} = 2,448$, onde p=2 variáveis (*nitrogenio* e *enxofre*).

Já na Figura 5.15, é possível observar que quando confrontam-se os Resíduos Padronizados da Mínima Mediana ao Quadrado (Resíduos Robustos) contra a Distância de Mahalanobis Robusta, detectam-se muito mais pontos de alavanca, são eles: observações 1,

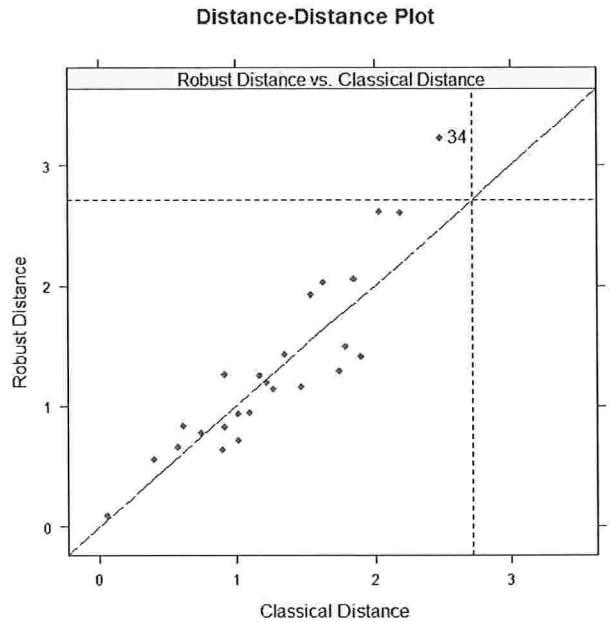

Figura 5.13: Distância Robusta contra Distância Clássica - modelo (5.4).

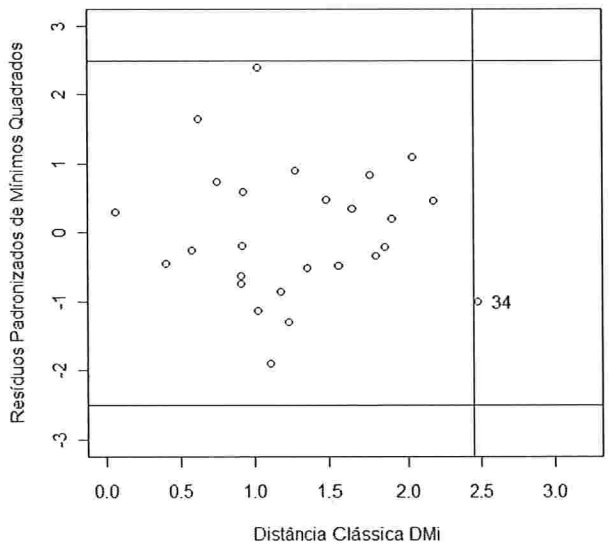

Figura 5.14: Distância Clássica contra Resíduos Padronizados de Mínimos Quadrados - modelo (5.4).

Observação	DM_i^2	DR_i^2
1	9,4214	2,7815
2	0,3106	0,3591
3	12,9743	3,9719
4	22,4886	45,8887
5	4,7734	1,7977
6	1,0846	0,2519
7	0,6659	0,1615
8	2,2408	2,0500
9	0,1089	0,1833
10	5,8644	13,7410
11	3,3696	4,1666
12	11,7441	17,9022
13	0,0001	0,0001
17	0,1475	0,4559
18	0,7158	2,4902
19	2,6226	1,6872
20	1,0853	0,7533
21	0,0257	0,0913
22	0,6659	0,1615
23	0,7081	0,4559
24	1,4773	0,7788
25	17,0648	46,4419
26	7,2703	16,7742
31	10,2788	5,0102
33	1,8852	2,4734
34	37,7093	108,7544

Tabela 5.10: Tabela contendo o quadrado das Distâncias Clássica e Robusta dos dados - modelo (5.4).

3, 4, 5, 6, 17, 18 e 26 (antes não identificados pelo elipsóide robusto) e as mesmas identificadas pelo elipsóide robusto: 25 e 34. O critério para determinar os *outliers* foi o mesmo definido no gráfico anterior, e a implementação destes dois últimos gráficos citados (Figuras 5.14 e 5.15) também encontra-se no Apêndice B.

Além disso, este último gráfico (Figura 5.15) apresenta uma classificação visual dos dados em quatro categorias: (1) observações regulares com baixos valores de DR_i e $r_i/\widehat{\sigma}$, (2) os *outliers* verticais com baixos valores de DR_i e altos valores de $r_i/\widehat{\sigma}$, (3) os "bons" pontos de alavanca, com altos valores de DR_i e baixos valores de $r_i/\widehat{\sigma}$ e por último, (4) os "maus"

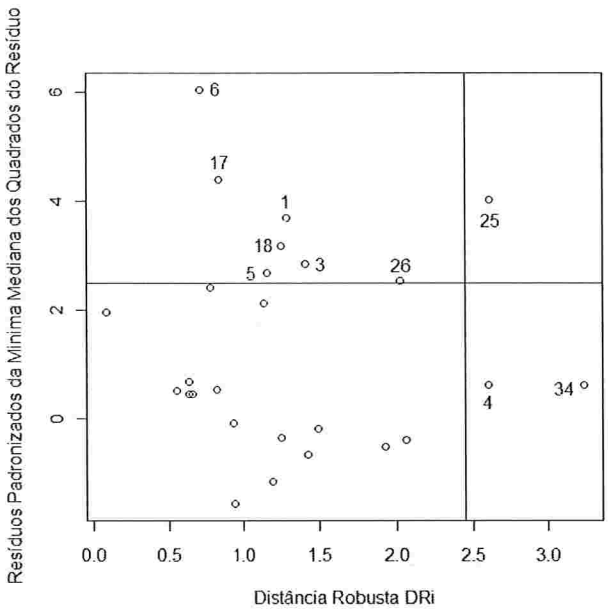

Figura 5.15: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.4).

pontos de alavanca, com altos valores de DR_i e $r_i/\widehat{\sigma}$. As observações são então distribuídas da seguinte maneira: em (1): todas as observações, com exceção das observações classificadas em (2), (3) e (4); em (2): 1, 3, 5, 6, 17, 18 e 26; em (3): 4 e 34; e em (4): 25. Seguindo a sugestão dada por Fung (1993), que suaviza o critério de classificação adotando valores de corte maiores e percentis de ordem acima do utilizado, pode-se adotar o intervalo [-5;+5] e o percentil de ordem 99% da distribuição Qui-Quadrado ($\sqrt{\chi^2_{2;0,99}} = 3,03$). Desta maneira, seriam identificados apenas as observações 6 como *outlier* vertical e 34 como "bom" ponto de alavanca (Observe a Figura 5.16).

Os gráficos de diagnóstico clássico detectaram sempre a observação 34 como sendo *outlier*, que foi a única observação também detectada em todos os gráficos de diagnóstico robusto construídos. Desta maneira, temos que tanto o diagnóstico clássico quanto o robusto levam à mesma conclusão: apenas a observação 34 é considerada *outlier*. Caracterizando este ponto, percebe-se que no banco de dados contendo apenas 26 observações, é a observação que apresenta os maiores valores tanto para a variável *nitrogenio* quanto para a variável *enxofre*.

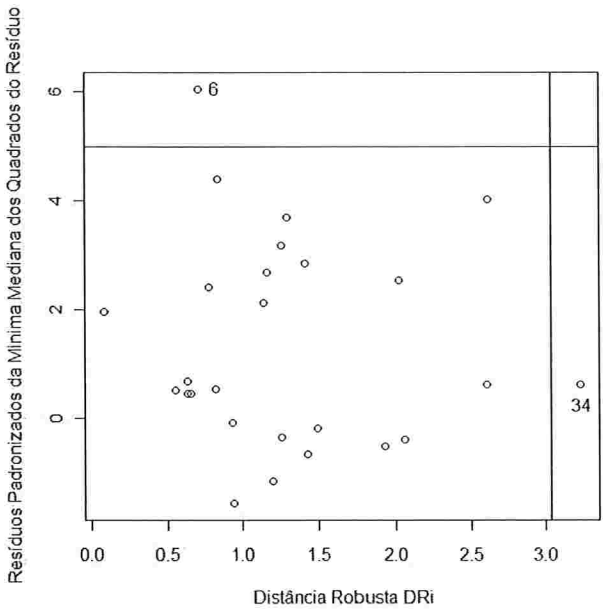

Figura 5.16: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.4).

A análise de diagnóstico robusto apresentada até o momento contemplou apenas duas variáveis explicativas: *nitrogenio* e *enxofre*. Apesar da variável *fluoreto* ter se mostrado não significante, iremos apresentar a análise de diagnóstico robusto com a inclusão desta variável, apenas como ilustração.

Através da Tabela 5.3, obtém-se a equação de regressão da *fotossintese* em função das variáveis independentes *fluoreto*, *nitrogenio* e *enxofre*, dada por:

$$\widehat{fotossintese}_i = 12,53 + 0,02 \cdot fluoreto_i - 0,61 \cdot nitrogenio_i + 1,38 \cdot enxofre_i, i = 1, \cdots, 30. \tag{5.5}$$

Como na análise anterior, realizou-se inicialmente a análise de diagnóstico clássica, para posterior comparação com a correspondente análise robusta.

A Figura 5.17 destaca os pontos 14, 26 e 27 como pontos de alavanca, o ponto 14 como ponto influente e os pontos 6 e 17 como aberrantes (resíduos padronizados fora do intervalo [-2;2]). Não há indícios de heterocedasticidade.

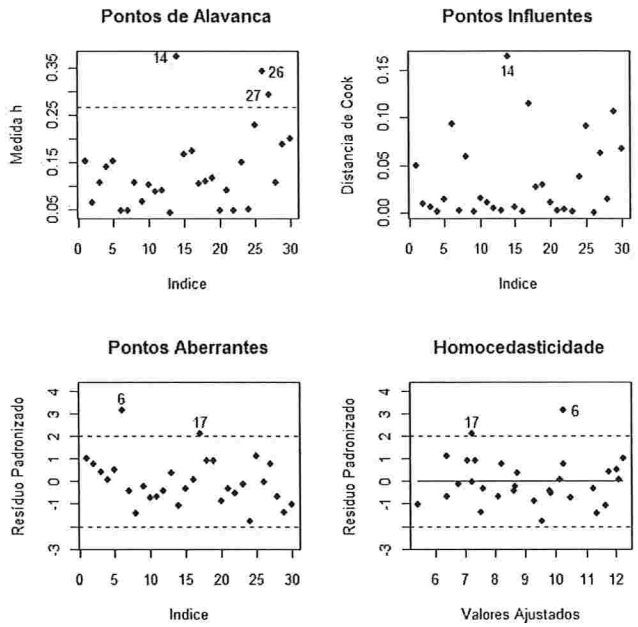

Figura 5.17: Continuação da Análise Clássica de Resíduos para o modelo (5.5).

O Diagnóstico Robusto iniciou-se através da identificação de possíveis *outliers* através do método *ForwardSearch* (Atkinson e Riani (2004)), e foi aplicado para detectar pontos discrepantes nas variáveis independentes *fluoreto*, *nitrogenio* e *enxofre*. O conjunto inicial de observações escolhidos através deste método contém as seguintes observações: 6, 13, 17 e 21.

A Figura 5.18 mostra a distância de Mahalanobis para o modelo de regressão contendo 3 variáveis, no qual o eixo das ordenadas representa o número da observação e as abcissas informam sobre a ordem de entrada das observações através do comprimento do retângulo branco, que indica o momento em que a observação é incluída no subconjunto. Assim, no primeiro passo, a observação incluída é a 22, no segundo passo é a 7, no terceiro passo é a 24 e assim por diante, sendo que o tamanho do conjunto passa a ser 5, 6 e 7 com a entrada desses pontos, respectivamente. A Tabela 5.11 traz esta informação para cada passo do processo iterativo.

Além disso, também é possível verificar como se comporta a distância de Mahalanobis de cada observação em cada passo do método. O resultado deste estudo encontra-se na

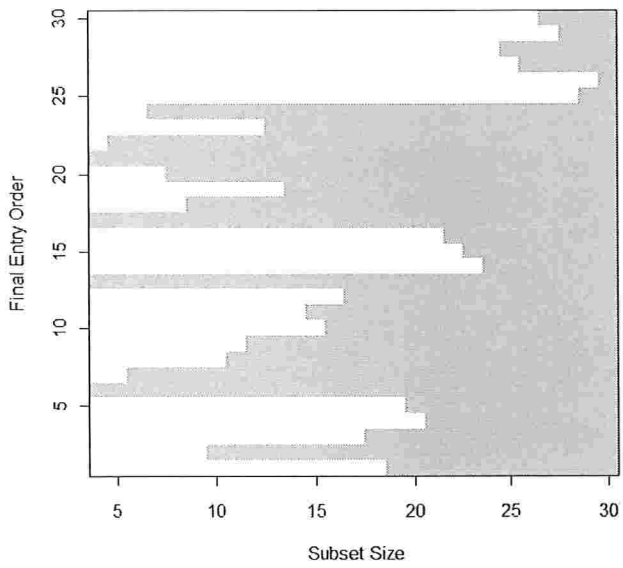

Figura 5.18: Ordem de entrada das observações considerando o modelo (5.5).

Figura 5.19. A partir desta figura, é possível verificar que as observações 27, 26, 25, 34, 33, 31 e 14 são as que possuem as maiores distâncias de Mahalanobis, seguidas pelas observações 16 e 15. Estas observações são as que possuem as maiores distâncias de Mahalanobis em quase todo o processo iterativo, e por isso podem ser consideradas atípicas pelo método de Atkinson e Riani (2004).

O elipsóide de volume mínimo citado na Seção 4.2 pode ser visto na Figura 5.20. A partir desta figura é possível observar 3 conjuntos de elipsóides: o primeiro é exatamente o mesmo comentado anteriormente na Figura 5.4 (para as variáveis *nitrogenio* e *enxofre*). O segundo conjunto de elipsóides confronta as variáveis *nitrogenio* e $fluoreto$ (veja detalhadamente na Figura 5.21), e mostra que o elipsóide clássico classifica apenas a observação 26 como *outlier*, enquanto o elipsóide robusto classifica as observações 25, 27, 31, 33 e 34, além da 26. E por último, o terceiro conjunto de elipsóides confronta as variáveis *enxofre* e *fluoreto* (veja detalhadamente na Figura 5.22), e classifica os *outliers* da mesma maneira que a Figura 5.21, classificando além das já citadas, também as observações 14, 15 e 16 como

Passo	Observação Incluída	Tamanho do subconjunto
2	22	5
3	7	6
4	24	7
5	20	8
6	18	9
7	2	10
8	8	11
9	9	12
10	23	13
11	19	14
12	11	15
13	10	16
14	12	17
15	3	18
16	1	19
17	5	20
18	4	21
19	16	22
20	15	23
21	14	24
22	31	25
23	27	26
24	34	27
25	33	28
26	25	29
27	26	30

Tabela 5.11: Obsevação incluída na amostra em cada passo do método de Atkinson e Riani (2004) considerando o modelo (5.5).

outliers pelo elipsóide robusto. O critério utilizado para classificação das observações como *outliers* foi adotar a confiança de 95% para os elipsóides, ou seja, foi classificada como *outlier* a observação cuja distância foi superior ao valor $\sqrt{\chi^2_{p;0,95}} = \sqrt{7,815} = 2,796$, onde $p = 3$ (variáveis *fluoreto*, *nitrogenio* e *enxofre*). Este gráfico compara a distância de Mahalanobis clássica (representada pela linha tracejada e obtida utilizando-se as matrizes de média e de covariância amostral dos dados) com a distância de Mahalanobis robusta (representada pela linha tracejada e obtida utilizando-se os estimadores EVM de locação e escala) para as 30 plantas.

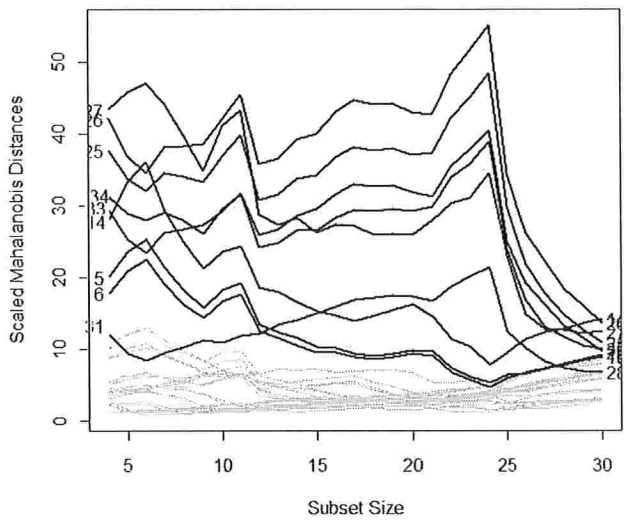

Figura 5.19: Distância de Mahalanobis para o modelo (5.5).

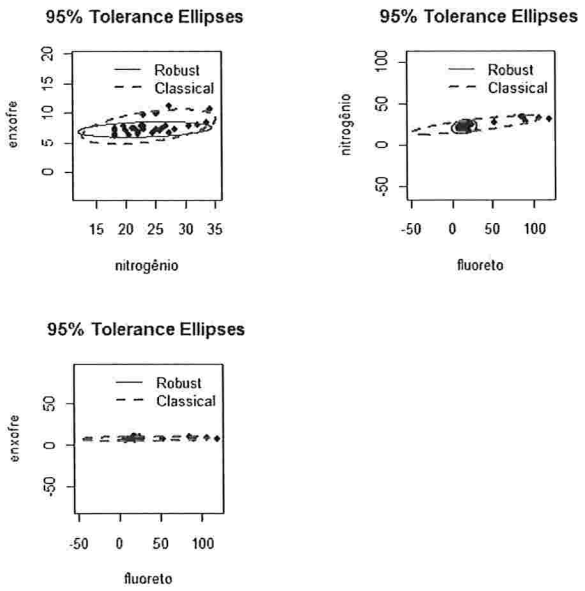

Figura 5.20: Elipsóides Clássico e Robusto das variáveis *fluoreto*, *nitrogenio* e *enxofre* - modelo (5.5).

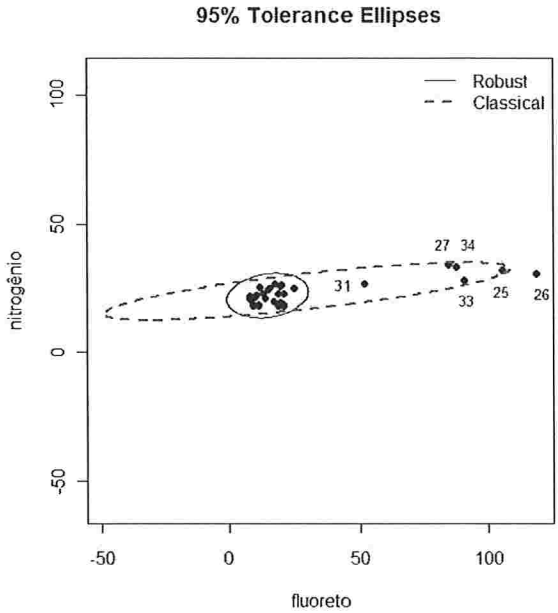

Figura 5.21: Elipsóides Clássico e Robusto das variáveis *nitrogenio* e *fluoreto*.

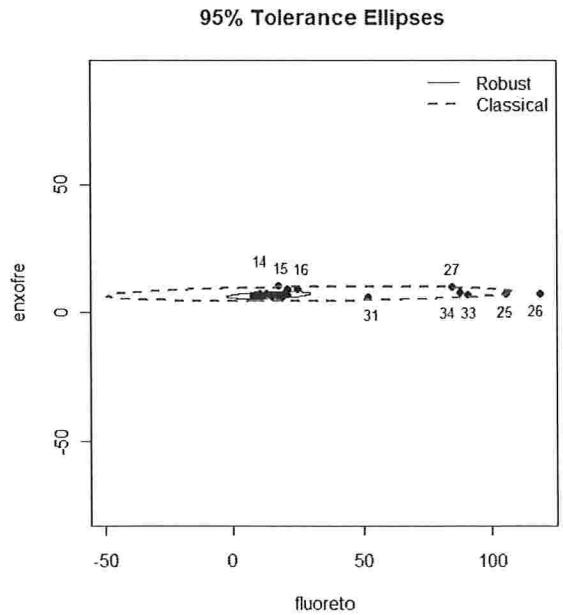

Figura 5.22: Elipsóides Clássico e Robusto das variáveis *enxofre* e *fluoreto*.

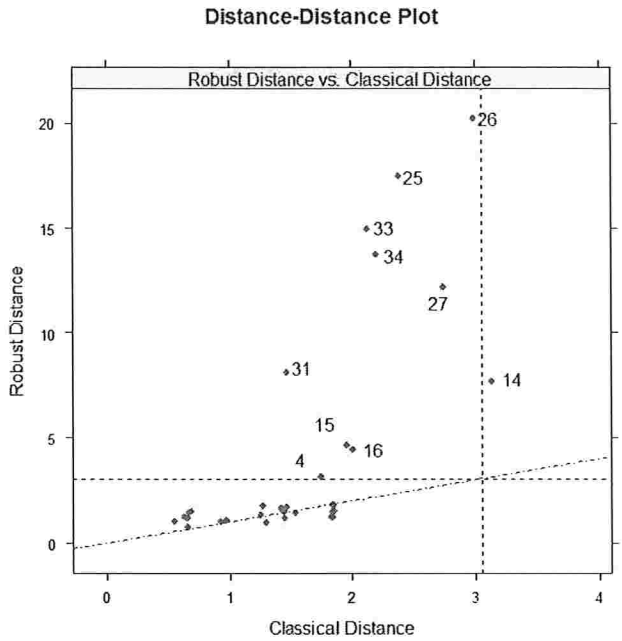

Figura 5.23: Distância Robusta contra Distância Clássica - modelo (5.5).

Para a construção dos elipsóides foi necessário o cálculo das distâncias clássica e robusta dos dados, apresentados na Tabela 5.12. Confrontando estas distâncias em um gráfico (Figura 5.23), observa-se que aproximadamente sete observações detectadas como *outliers* através do elipsóide robusto localizam-se bem distante dos demais: 14, 25, 26, 27, 31, 33 e 34. Já as observações 4, 15 e 16 localizam-se próximas à massa dos dados. Percebe-se que estes três pontos estão localizados bem próximos à fronteira do elipsóide robusto da Figura 5.22 e nem são citados como *outliers* na Figura 5.21. Neste gráfico, adotou-se uma confiança de 97,5% como critério para classificar uma observação como *outlier*. Desta forma, foi classificado como *outlier* a observação cuja distância foi superior ao valor $\sqrt{\chi^2_{p;0,975}} = \sqrt{9,348} = 3,057$, onde p=3.

A Figura 5.24 mostra o gráfico dos Resíduos Padronizados de Mínimos Quadrados contra a Distância de Mahalanobis Clássica, discutido na Seção 4.2. Percebe-se que apenas as observações 6, 14 e 26 encontram-se fora dos limites que identificam pontos de alavanca ou *outliers* (apenas a última foi identificada também pelo elipsóide clássico). O mesmo critério

Observação	DM_i^2	DR_i^2
1	3,4493	2,3399
2	0,8661	1,0571
3	2,1257	2,3214
4	3,0969	9,9547
5	3,4570	3,3255
6	0,4322	0,5665
7	0,4397	2,1064
8	2,1183	1,4055
9	0,9487	1,2009
10	2,0327	2,7762
11	1,5831	1,8214
12	1,6364	3,1402
13	0,2983	1,0801
14	9,8604	58,8446
15	3,8504	21,7957
16	4,0611	19,6552
17	2,0674	2,3920
18	2,1803	2,9325
19	2,3940	2,1522
20	0,4308	1,3823
21	1,6880	0,9542
22	0,3977	1,5793
23	3,3988	1,5850
24	0,4700	2,2798
25	5,6970	304,3517
26	8,9675	407,6538
27	7,5140	148,4961
31	2,1768	65,3021
33	4,5220	223,6707
34	4,8380	188,8859

Tabela 5.12: Tabela contendo o quadrado das Distâncias Clássica e Robusta dos dados - modelo (5.5).

utilizado para a detecção dos *outliers* nos elipsóides foi adotado aqui: o de considerar as observações cuja distâncias são maiores do que $\sqrt{\chi^2_{p;0,95}} = \sqrt{7,815} = 2,796$, onde p=3 variáveis (*fluoreto*, *nitrogenio* e *enxofre*).

Já na Figura 5.25, é possível observar que quando confrontam-se os Resíduos Padronizados da Mínima Mediana ao Quadrado (Resíduos Robustos) contra a Distância de Mahalanobis Robusta, detectam-se muito mais do que apenas três pontos de alavanca ou *outliers*, são eles: observações 4, 6, 8, 17, 18, 19 e 24 (antes não identificados pelos elipsóides robustos) e as mesmas identificadas pelos elipsóides robustos: 14, 15, 16, 25, 26, 27, 31, 33 e 34. O critério para determinar os *outliers* foi o mesmo definido no gráfico anterior.

Além disso, este último gráfico (Figura 5.25) apresenta uma classificação visual dos dados em quatro categorias: (1) observações regulares com baixos valores de DR_i e $r_i/\widehat{\sigma}$, (2) os *outliers* verticais com baixos valores de DR_i e altos valores de $r_i/\widehat{\sigma}$, (3) os "bons" pontos de alavanca, com altos valores de DR_i e baixos valores de $r_i/\widehat{\sigma}$ e por último, (4) os "maus" pontos de alavanca, com altos valores de DR_i e $r_i/\widehat{\sigma}$. As observações são então distribuídas da seguinte maneira: em (1): todas as observações, com exceção das observações classificadas em (2), (3) e (4); em (2): observações 6, 8, 17, 18, 19 e 24; em (3): observações 4, 14, 15, 16, 26, 31, 33 e 34; e em (4): 25 e 27. Seguindo a sugestão dada por Fung (1993), que suaviza o critério de classificação adotando valores de corte maiores e percentis de ordem acima do utilizado, pode-se adotar o intervalo [-5;+5] e o percentil de ordem 99,9% da distribuição Qui-Quadrado ($\sqrt{\chi^2_{3;0,999}} = \sqrt{16,266} = 4,033$). Desta maneira, seriam identificadas as observações 6 e 17 como *outliers* verticais e as observações 14, 15, 16, 25, 26, 27, 31, 33 e 34 como "bons" pontos de alavanca (Observe a Figura 5.26).

Os pontos que aparecem tanto no gráfico dos elipsóides quanto no gráfico da distância de Mahalanobis Robusta contra os resíduos padronizados são os pontos 14, 15, 16, 25, 26, 27, 31, 33 e 34. Ao caracterizarmos estes pontos, percebemos que as observações 14, 15, 16 e 27 foram as mesmas identificadas pelo diagnóstico robusto considerando apenas duas variáveis explicativas e que possuem altos valores para a variável *enxofre*. Já as demais observações,

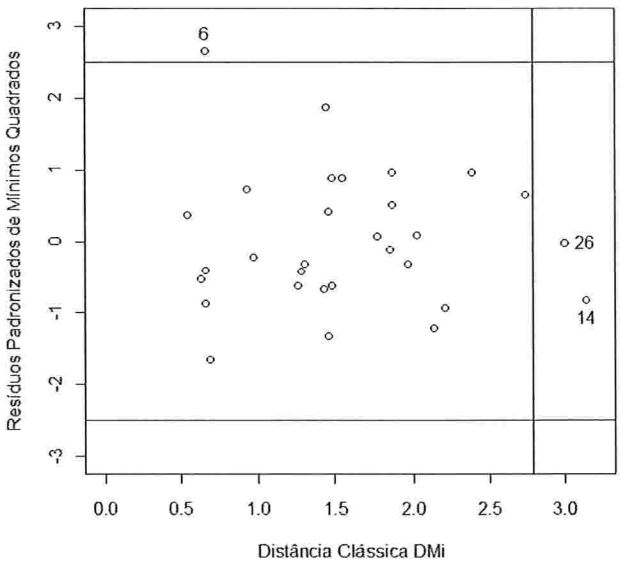

Figura 5.24: Distância Clássica contra Resíduos Padronizados de Mínimos Quadrados - modelo (5.5).

25, 26, 31, 33 e 34 apresentam altos valores para a variável *fluoreto*.

Mais uma vez, é importante notar que o diagnóstico não deve ser realizado observando-se apenas um e sim um conjunto de gráficos estudados.

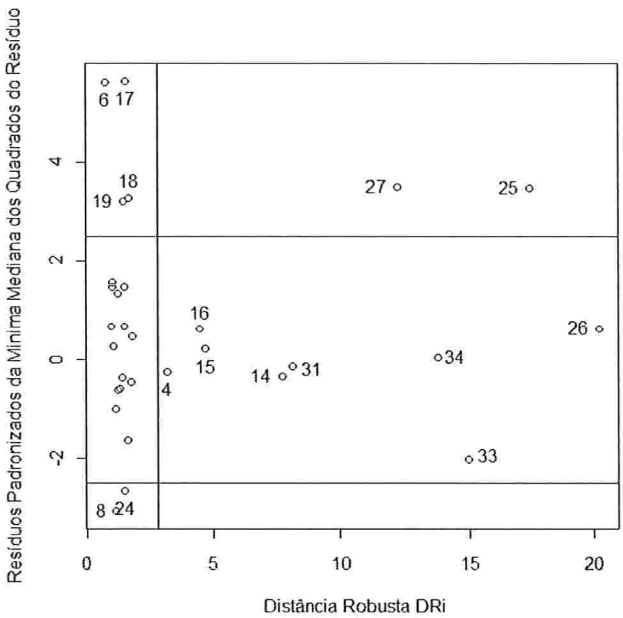

Figura 5.25: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.5).

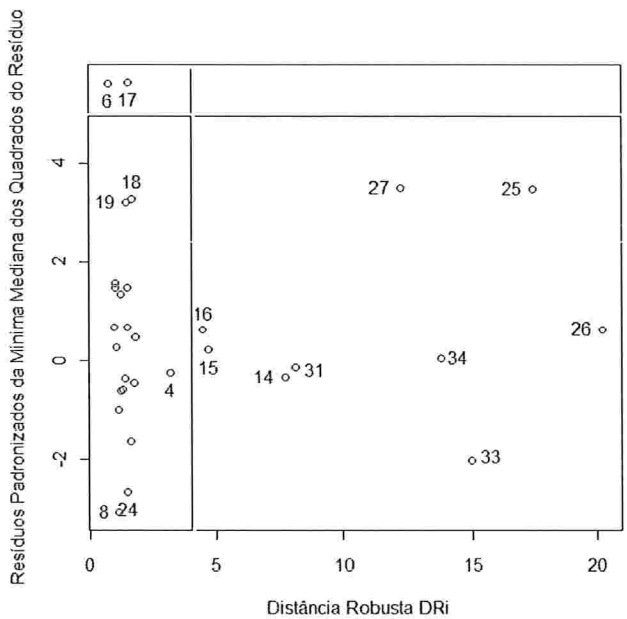

Figura 5.26: Distância Robusta contra Resíduos Padronizados da Mínima Mediana dos Quadrados do Resíduo - modelo (5.5).

Capítulo 6

Conclusões

Neste trabalho foram apresentados métodos de diagnóstico robusto em Análise Multivariada e em Análise de Regressão. Foi visto que este tipo de técnica é útil tanto para verificar as suposições iniciais de um modelo quanto na detecção de observações atípicas.

Aplicou-se a técnica Diagnóstico Robusto em Análise de Regressão em um banco de dados real e realizou-se a comparação com o correspondente método clássico. Analisando todos os gráficos gerados conjuntamente, conclui-se que no método robusto os pontos aberrantes tornam-se mais visíveis do que no método clássico, classificando mais observações como *outliers*. Os gráficos construídos neste trabalho possuem uma rotina desenvolvida no *software* R e que pode ser encontrado no Apêndice B.

Como indicação para continuidade deste trabalho, sugere-se a implementação dos diagnósticos robustos em Análise de Componentes Principais e em Análise Discriminante, que foram apenas apresentados neste trabalho.

Finalizando, gostaríamos de destacar alguns artigos sobre distâncias robustas que, por apresentarem um enfoque distinto do utilizado neste trabalho, não foram aqui abordados.

Hardin e Rocke (2005) analisam o uso da distribuição F ao invés da Qui-Quadrado na detecção de *outliers*. Já Olive (2002) apresenta uma metodologia gráfica alternativa como diagnóstico para distribuições normais multivariadas e elípticas, com interessantes aplicações.

Acreditamos que estes dois trabalhos possam se constituir em um interessante tópico de pesquisa futura.

Apêndice A

Detalhes técnicos para obtenção do Elipsóide de Volume Mínimo (EVM)

O elipsóide de volume mínimo proposto por Rousseeuw e Van Zomeren (1990) tem a seguinte construção:

Suponha o banco de dados $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ contendo n pontos p-dimensionais cujo objetivo é o de estimar o "centro" e a "dispersão" dos pontos através de um vetor $T(\mathbf{X})$ e de uma matriz de covariâncias $S(\mathbf{X})$. Dizemos que os estimadores $\mathbf{T}$ e $\mathbf{S}$ são equivariantes quando:

$$T(\mathbf{x_1A} + \mathbf{b}, \cdots, \mathbf{x_nA} + \mathbf{b}) = \mathbf{T}(\mathbf{x_1}, \cdots, \mathbf{x_n})\mathbf{A} + \mathbf{b}$$

e

$$S(\mathbf{x_1A} + \mathbf{b}, \cdots, \mathbf{x_nA} + \mathbf{b}) = \mathbf{A'S}(\mathbf{x_1}, \cdots, \mathbf{x_n})\mathbf{A}$$

para qualquer vetor linha $\mathbf{b}$ e qualquer matriz não singular A_{pxp}. O vetor de médias amostrais e a matriz de covariância amostral

$$T(\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}$$

e

$$S(\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \mathbf{T(X)})'(\mathbf{x_i} - \mathbf{T(X)})$$

são estimadores equivariantes, porém não são robustos pois são sensivelmente influenciados por *outliers*.

O estimador EVM (Elipsóide de Volume Mínimo) é definido como o par (T,S), onde $T(\mathbf{X})$ é um vetor p-dimensional e $S(\mathbf{X})$ é uma matriz $p \times p$ positiva semi-definida tal que seu determinante é mínimo sujeito à seguinte condição:

$$\#\{i; (\mathbf{x_i} - \mathbf{T})\mathbf{S^{-1}}(\mathbf{x_i} - \mathbf{T})' \leq \mathbf{a^2}\} \geq \mathbf{h},$$

em que # indica o número de elementos do conjunto, $h = [(n + p + 1)/2]$ e $[q]$ é a parte inteira de q. O valor a^2 é uma constante fixada, que pode ser escolhida, por exemplo, como o quantil de ordem 0,50 da distribuição qui-quadrado com p graus de liberdade ($\chi^2_{p;0,50}$). Tal escolha é natural quando espera-se que a maioria dos dados venham de uma distribuição normal.

Verifica-se que o estimador EVM possui ponto de ruptura de aproximadamente 50%, o que significa que $T(\mathbf{X})$ irá permanecer limitada e os autovalores de $S(\mathbf{X})$ irão ficar distantes de zero e infinito quando menos da metade dos dados são substituídos por valores arbitrários (Lopuhaä e Rousseeuw (1991)).

As Distâncias Robustas relativas ao estimador EVM são definidas como:

$$DR_i = \sqrt{(\mathbf{x_i} - \mathbf{T(X)})\mathbf{S(X)^{-1}}(\mathbf{x_i} - \mathbf{T(X)})'}. \tag{A.1}$$

Pode-se calcular a média ponderada:

$$T_1(\mathbf{X}) = \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} \right)^{-1} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i x_i}, \tag{A.2}$$

e a matriz de covariância ponderada:

$$S_1(\mathbf{X}) = \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} - \mathbf{1} \right)^{-\mathbf{1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \mathbf{T_1(X)})'(\mathbf{x_i} - \mathbf{T_1(X)}) \tag{A.3}$$

em que os pesos $\mathbf{w_i} = \mathbf{w(DR_i)}$ dependem das distâncias robustas. Pode-se mostrar que T_1 e S_1 possuem os mesmos pontos de ruptura que os iniciais T e S quando o peso da função $\mathbf{w}$ tende a zero para grandes valores de DR_i (veja Lopuhaä e Rousseeuw (1991)).

Apêndice B

Programas em R

```
#Leitura dos dados#
dados<-read.table(file="C:\\ \\ ... \\ \\DADOS.txt",header=TRUE,sep="\t")
attach(dados) #faz com que o R adicione o banco de dados à sua memoria#
dados

##########################################################
#### Diagnóstico Robusto para duas variáveis explicativas: nitrogenio e enxofre ####
##########################################################

#Selecionando apenas as variáveis de interesse (fotossintese, nitrogenio e enxofre)#
dadosmodelo<-dados[c(4,18,19)] #banco de dados original com 36 linhas#
dadosmodelo

#Excluindo os valores faltantes#
dadosfiltmodelo<-na.omit(dadosmodelo) #banco de dados final 30 linhas #
dadosfiltmodelo

#Matrix X contendo apenas as variáveis explicativas#
dadosX_modelo<-dadosfiltmodelo[c(2,3)]
dadosX_modelo

#Modelo de regressão da fotossintese em função do nitrogenio e do enxofre#
modelo<-lm(fotoss~,data=dadosfiltmodelo)
summary(modelo)

#Detecção de Outliers através do Método Robusto The Forward Search#
#Método Forward Search para o conjunto de dados X do modelo selecionado, utilizando
```

```
boxplots bivariados para definir o conjunto inicial#
#Instalar o pacote "Rfwdmv"#
library(MASS)
library(Rfwdmv)
saida<-fwdmv.init(dadosX_modelo,bsb=bb.subset,scaled=TRUE,monitor="all")

#Conjunto inicial obtido através dos boxplots robustos bivariados #
conj.inicial<-matrix(0,saida$m,1)
n<-30
aux<-1
for(i in 1:n){
if(saida$Unit[i,1]>0){
conj.inicial[aux,1]<-i
aux<-aux+1
}
}
conj.inicial

#Gráfico que mostra a ordem de entrada das observações#
graf1<-fwdmvEntryPlot(saida,entry.order="final",subset.size=-1,psfrag.labels=FALSE)

#Gráfico da Distância de Mahalanobis para cada individuo em cada passo do procedimento#
graf2<-fwdmvDistancePlot(saida,group=NULL,id=TRUE,psfrag.labels=FALSE)

#Elipsóides clássico e robusto#
#Instalar pacote "robust"#
library(robust)
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Gráfico da Distancia Clássica contra a Distancia Robusta#
covfmDistance2Plot(fm)

#Gráfico dos Resíduos Padronizados dos Mínimos Quadrados vs Distância Clássica DMi#
p<-ncol(dadosX_modelo)
n<-nrow(dadosX_modelo)
modelouls<-lm(fotoss .,data=dadosfiltmodelo)
s<-sqrt(sum(modelouls$resˆ2)/(n-p-1))
uls.res<-modelouls$res/s
uls.distance<-sqrt(cov$dist)
plot(uls.distance,uls.res,ylim=c(-3,3),xlim=c(0,3.2),ylab="Resíduos Padronizados de Mínimos
Quadrados",xlab="Distância Clássica MDi")
abline(h=c(-2.5,2.5))
crit.pt<-sqrt(qchisq(.95,p))
abline(v=crit.pt)
```

```
uls.indice<-uls.distance>crit.pt|abs(uls.res)>2.5
identify(uls.distance,uls.res,pts=cbind(uls.distance,uls.res)[uls.indice,])

#Gráfico dos Resíduos Padronizados da Mínima Mediana ao Quadrado vs Distância Robusta DRi#
modelolms<-lmsreg(fotoss .,data=dadosfiltmodelo)
lms.res<-modelolms$residuals/modelolms$scale
mve.distance<-sqrt(covRob$dist)
plot(mve.distance,lms.res,ylab="Resíduos Padronizados da Minima Mediana dos Quadrados do Resíduo",xlab="Distância Robusta DRi")
abline(h=c(-2.5,2.5))
crit.pt<-sqrt(qchisq(.95,p))
abline(v=crit.pt)
lms.indice<-mve.distance>crit.pt|abs(lms.res)>2.5
identify(mve.distance,lms.res,pts=cbind(mve.distance,lms.res)[lms.indice,])

##########################################################
## Diagnóstico Robusto para três variáveis explicativas: fluoreto, nitrogenio e enxofre ##
##########################################################

#Selecionando apenas as variáveis de interesse (fotossintese, nitrogenio e enxofre)#
dadosmodelo<-dados[c(4,17,18,19)] #banco de dados original com 36 linhas#
dadosmodelo

#Excluindo os valores faltantes#
dadosfiltmodelo<-na.omit(dadosmodelo) #banco de dados final 30 linhas #
dadosfiltmodelo

#Matrix X contendo apenas as variáveis explicativas#
dadosX_modelo<-dadosfiltmodelo[c(2,3,4)]
dadosX_modelo

#Modelo de regressão da fotossintese em função do fluoreto, nitrogenio e enxofre#
modelo<-lm(fotoss~.,data=dadosfiltmodelo)
summary(modelo)

#Detecção de Outliers através do Método Robusto The Forward Search#
#Método Forward Search para o conjunto de dados X do modelo selecionado, utilizando boxplots bivariados para definir o conjunto inicial#
#Instalar o pacote "Rfwdmv"#
library(MASS)
library(Rfwdmv)
saida<-fwdmv.init(dadosX_modelo,bsb=bb.subset,scaled=TRUE,monitor="all")

#Conjunto inicial obtido através dos boxplots robustos bivariados #
```

```
conj.inicial<-matrix(0,saida$m,1)
n<-30
aux<-1
for(i in 1:n){
if(saida$Unit[i,1]>0){
conj.inicial[aux,1]<-i
aux<-aux+1
}
}
conj.inicial

#Gráfico que mostra a ordem de entrada das observações#
graf1<-fwdmvEntryPlot(saida,entry.order="final",subset.size=-1,psfrag.labels=FALSE)

#Gráfico da Distância de Mahalanobis para cada individuo em cada passo do procedimento#
graf2<-fwdmvDistancePlot(saida,group=NULL,id=TRUE,psfrag.labels=FALSE)

#Elipsóides clássico e robusto#
#Instalar pacote "robust"#
library(robust)
#Prepara o gráfico para colocar os 3 conjuntos de elipsóides no mesmo gráfico#
par(opar)
par(mfrow=c(2,2))

#Trabalhando apenas com a matriz X das variáveis nitrogenio e enxofre #
dadosX_modelo<-dadosfiltmodelo[c(3,4)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Trabalhando apenas com a matriz X das variáveis fluoreto e nitrogenio #
dadosX_modelo<-dadosfiltmodelo[c(2,3)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Trabalhando apenas com a matriz X das variáveis fluoreto e enxofre #
dadosX_modelo<-dadosfiltmodelo[c(2,4)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Volta ao estilo original, de um gráfico por página#
par(opar)
par(mfrow=c(1,1))
```

```
#Elipsoide do nitrogenio vs enxofre# dadosX_modelo<-dadosfiltmodelo[c(3,4)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Elipsoide do fluoreto vs nitrogenio# dadosX_modelo<-dadosfiltmodelo[c(2,3)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Elipsoide do fluoreto vs enxofre# dadosX_modelo<-dadosfiltmodelo[c(2,4)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmEllipsesPlot(fm)

#Gráfico da Distancia Clássica contra a Distancia Robusta#
dadosX_modelo<-dadosfiltmodelo[c(2,3,4)]
dadosX_modelo
fm<-fit.models(list(Robust="covRob",Classical="cov"),data=(dadosX_modelo))
covfmDistance2Plot(fm)

#Gráfico dos Resíduos Padronizados dos Mínimos Quadrados vs Distância Clássica DMi#
p<-ncol(dadosX_modelo)
n<-nrow(dadosX_modelo)
modelouls<-lm(fotoss .,data=dadosfiltmodelo)
s<-sqrt(sum(modelouls$resˆ2)/(n-p-1))
uls.res<-modelouls$res/s
uls.distance<-sqrt(cov$dist)
plot(uls.distance,uls.res,ylim=c(-3,3),xlim=c(0,3.2),ylab="Resíduos Padronizados de Mínimos
Quadrados",xlab="Distância Clássica MDi")
abline(h=c(-2.5,2.5))
crit.pt<-sqrt(qchisq(.95,p))
abline(v=crit.pt)
uls.indice<-uls.distance>crit.pt|abs(uls.res)>2.5
identify(uls.distance,uls.res,pts=cbind(uls.distance,uls.res)[uls.indice,])

#Gráfico dos Resíduos Padronizados da Mínima Mediana ao Quadrado vs Distância Ro-
busta DRi#
modelolms<-lmsreg(fotoss .,data=dadosfiltmodelo)
lms.res<-modelolms$residuals/modelolms$scale
mve.distance<-sqrt(covRob$dist)
plot(mve.distance,lms.res,ylab="Resíduos Padronizados da Minima Mediana dos Quadrados
do Resíduo",xlab="Distância Robusta DRi")
abline(h=c(-2.5,2.5))
```

```
crit.pt<-sqrt(qchisq(.95,p))
abline(v=crit.pt)
lms.indice<-mve.distance>crit.pt|abs(lms.res)>2.5
identify(mve.distance,lms.res,pts=cbind(mve.distance,lms.res)[lms.indice,])

detach(dados)
```

Referências Bibliográficas

[1] Ammann, L. P. (1989). Robust Principal Components. *Communications in Statistics - Simulation and Computation*, 18, 857-874.

[2] Ammann, L. P. e Van Ness, J. (1989). Standard and Robust Orthogonal Regression. *Communications in Statistics - Simulation and Computation*, 18, 145-162.

[3] Atkinson, A. C. (1986). Masking Unmasked. *Biometrika*, 73, 533-541.

[4] Atkinson, A. C. e Riani, M. (2004). The Forward Search and Data Visualisation. *Computational Statistics*, 19, 29-54.

[5] Aubin, E. C. Q., Elian, S. N. e Alencar, G. P. (1999). *Relatório de análise estatística sobre o projeto: "Efeitos da poluição sobre as trocas gasosas de indivíduos jovens de Tibouchina pulchra Cogn. (Melastomataceæ) na região de Cubatão, SP"*. RAE-CEA-9905. São Paulo.

[6] Barnett, B. e Lewis, T. (1984). *Outliers in Statistical Data*. 2.ed. John Wiley. New York.

[7] Belsey, D. A., Kuh, E. e Welsch, R. E. (1980). *Regression Diagnostics*. John Wiley & Sons. New York.

[8] Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*. 2.ed. John Wiley. New York.

[9] Campbell, N. A. (1978). The Influence Function as an AID in Outlier Detection in Discriminant Analysis. *Applied Statistics*, 27, N°3, 251-258.

[10] Campbell, N. A. (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. *Applied Statistics*, 29, N°3, 231-237.

[11] Caroll, R. J. e Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall. London.

[12] Chen, H. J., Gnanadesikan, R. e Kettenring, J. R. (1974). Statistical Methods for Grouping Corporations. *Sankya*, B, N°36, 1-28.

[13] Cook, R. D. (1986). Assessment of Local Influence (with discussion). *J. R. Statist. Soc.*, B, *N°*48, 133-169.

[14] Cox, D. R. (1968). Notes on Some Aspects of Regression Analysis. *J. R. Statist. Soc.*, A, Nº131, 265-279.

[15] Devlin, S. J., Gnanadesikan, R. e Kettenring, J. R. (1975). Robust Estimation and Outliers Detection with Correlation Coefficients. *Biometrika*, 62, 531-545.

[16] Devlin, S. J., Gnanadesikan, R. e Kettenring, J. R. (1981). Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association*, 76, Nº374, 87-104.

[17] Dielman, T. E. e Pfaffenberger, R. (1982). LAV (Least Absolute Value) Estimation in Linear Regression: A Review. *Optimization in Statistics*, 19. Amsterdam: North-Jolland.

[18] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomy Problems. *Annals of Eugenics*, 7, 179-188.

[19] Fisher, R. A. (1938). The Statistical Utilization of Multiple Measurements. *Annals of Eugenics*, 8, 376-386.

[20] Flury, B. e Riedwyl, H. (1985). T2 Tests, the Linear Two-Group Discriminant Function and Their Computation by Linear Regression. *The American Statistician*, 39, 20-25.

[21] Fung, W. K. (1992). Some Diagnostic Measures in Discriminant Analysis. *Statistics and Probability Letters*, 13, 279-285.

[22] Fung, W. K. (1993). Unmasking *Outliers* and Leverage Points: A Confirmation. *Journal of the American Statistical Association*, 88, Nº422, 515-519.

[23] Fung, W. K. (1995). Diagnostics in Linear Discriminant Analysis. *Journal of the American Statistical Association*, 90, Nº43, 952-956.

[24] Gasko, M. e Donoho, D. (1982). Influential Observation in Data Analysis. *American Statistical Association*, Proceedings of the Business and Economic Statistics Section. 104-109.

[25] Giroldo, F. R. S. (2008). *Alguns Métodos Robustos para Detectar Outliers Multivariados.* IME USP. São Paulo. Dissertação de mestrado. 84p.

[26] Glorfeld, L. W. (1990). A Robust Methodology for Discriminant Analysis Based on Least-Absolut-Value Estimation. *Managerial and Decision Economics*, 11, 267-277.

[27] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2.ed., John Wiley. New York.

[28] Gnanadesikan, R. e Kettenring, J. R. (1972). Robust Estimates, Residuals and Outliers Detection with Multiresponse Data. *Biometrics*, 28, 81-124.

[29] Green, B. F. (1979). The two kinds of Linear Discriminant Functions. *The Journal of Educational Statistics*, 4, 247-263.

[30] Hample, F. R. (1973). Robust Estimation: a Condensed Partial Survey. *Z. Wahrscheinlichkeitstheorie and Verw. Gebiete*, 27, 87-104.

[31] Hample, F. R., Ronchetti, E. M., Rousseeuw, P. J. e Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley. New York.

[32] Hardin, J. e Rocke, D. M. (2005). The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics*, 14, Nº4, 928-946.

[33] Hawkins, D. M., Bradu, D. e Kass, G. V. (1984). Location of Several *Outliers* in Multiple Regression Data Using Elemental Sets. *Technometrics*, 26, 197-208.

[34] He, X. e Fung, W. K. (2000). High Breakdown Estimation for Multiple Populations with Applicatins to Discriminan Analysis. *Journal of Multivariate Analysis*, 72, 151-162.

[35] Healy, M. J. R. (1968). Multivariate Normal Plotting. *Applied Statistics*, 17, 157-161.

[36] Huber, P. J. (1964). Robust Estimation for a Location Parameter. *Annals of Mathematical Statistics*, 35, 73-101.

[37] Hubert, M. e Van Driessen, K. (2004). Fast and Robust Discriminant Analysis. *Computational Statistics & Data Analysis*, 45, 301-320.

[38] Lopuhaä, H. P. e Rousseeuw, P. J. (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals of Statistics*, 19, 229-248.

[39] Mosteller, F. e Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, Mass: Addison-Wesley.

[40] Narula, S. C. e Wellington, J. F. (1982). The Minimum Sum of Absolute Errors Regression: A State of the Art Survey. *International Statistics Review*, 50, 317-326.

[41] Olive, D. J. (2002). Applications of Robust Distances for Regression. *Technometrics*, 44, Nº1, 64-71.

[42] R Development Core Team. (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[43] Rao, C. R. (1964). The Use and Interpretation of Principal Components Analysis in Applied Research. *Sankhya Series*, A, Nº26, 329-358.

[44] Reigada, S. M. B. (2005). *Diagnóstico em Análise Discriminante*, IME USP, São Paulo. Dissertação de Mestrado. 190p.

[45] Reigada, S. M. B. e Elian, S. N. (2006). Diagnóstico em Análise Discriminante. *Revista Brasileira de Estatística*, 67, 45-73.

[46] Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871-880.

[47] Rousseeuw, P. J. e Leroy, A. (1987). *Robust Regression and Outlier Detection.* John Wiley. New York.

[48] Rousseeuw, P. J. e Van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, N°411, 633-639.

[49] SAS. (1976). *SAS Institute Inc.* Cary, NC: SAS Institute Inc. URL: http://www.sas.com.

[50] Teebagy, N. e Chatterjee, S. (1989). Inference in a Binary Response Model with Applicatins to Data Analysis. *Decision Sciences*, 20, 393-403.

[51] Weisberg, S. (1985) *Applied Linear Regression.* 2.ed. John Wiley & Sons.

[52] Zani, S., Riani, M. e Corbellini, A. (1988). Robust Bivariate Boxplots and Multiple *Outlier* Detection. *Computational Statistics and Data Analysis*, 28, 257-270.