R E G R E S S Ã O B I P O N D E R A D A

UM MÉTODO ROBUSTO DE AJUSTE

MÁRIO ALCIDES DE OLIVEIRA SCAFI

Orientador: JOSÉ NORBERTO WALTER DACHS

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas, como requisito parcial para a obtenção do título de Mestre em Estatística

CAMPINAS
Estado de São Paulo - Brasil
Maio, 1979

À meus pais

AGRADECIMENTOS

Agradeço ao Prof. José Norberto Walter Dachs todo apoio e sugestões que me foram dados durante o decorrer deste trabalho.

À minha futura esposa, Elizabete R. de Oliveira, agradeço a revisão do texto original.

Ao Prof. Ronaldo S. Wada, meu particular amigo, agradeço a confecção das figuras constantes do texto.

Devo a Sra. Elisa S. Peron a cuidadosa datilografia deste trabalho.

SUMÁRIO

Neste trabalho é apresentado um método robusto de ajuste de regressões, denominado regressão biponderada. Para a sua apresentação foi necessário introduzir alguns conceitos utilizados em estimação robusta. São os conceitos de resistência e robustez, curvas de influência, emparelhadores e sintonizadores e o conceito de valores aberrantes ou discrepantes. Além disso apresenta-se um estimador robusto para parâmetros de locação, denominado estimador biponderado.

O método apresentado é comparado com o método dos mínimos quadrados e com um outro método robusto de ajuste de regressões através de dois exemplos. São mostradas as razões da preferência da regressão biponderada sobre os outros dois métodos.

No final se apresenta um programa para computadores que possibilita a obtenção das estimativas dos parâmetros por mínimos quadrados, pelo método biponderado ou pela regressão passoponderada - uma versão simplificada do método apresentado.

SUMMARY

In this work a robust method for regression is presented, the biweight. For the understanding of the method it was necessary to present some basic concepts in robust estimation: resistance, robustness, influence curves, matchers, catchers and outlying values or outliers. A robust estimator of location is also presented - the biweight.

The biweight regression is then compared with least-squares and another robust method for regression, using two examples. The reasons for the preference of the biweight against the others are presented.

Finally a computer program is present. This program computes the estimated parameters by least-squares, the biweight estimates or the stepweight estimates - a simplified version of the former method.

ÍNDICE

Página

LISTA DOS QUADROS vii

LISTA DAS FIGURAS ix

I - INTRODUÇÃO 1

II - NOMENCLATURA e DEFINIÇÕES 6

II.1 - Resistência e Robustez 6

II.2 - Curvas de Influência 9

II.2.1 - Curva de Influência para Média 10

II.2.2 - Curva de Influência para S^2 12

II.2.3 - Curva de Influência para a Mediana 13

II.3 - Variáveis e Suportes 15

II.4 - Emparelhadores e Sintonizadores 16

II.4.1 - Emparelhadores 17

II.4.2 - Sintonizadores 20

II.5 - Valores discrepantes ou aberrantes 24

III - ESTIMADORES ROBUSTOS DE LOCAÇÃO E DE ESCALA 29

III.1 - M-, m- e w-estimadores 29

III.1.1 - O estimador Biponderado (biweight) 32

III.1.2 - O estimador Passoponderado (stepweight) . 44

III.2 - Alguns estimadores robustos de escala 48

III.2.1 - Amplitude Interquartis 49

III.2.2 - Mediana dos Desvios Absolutos - MDA 51

III.2.3 - Uma alternativa 52

Página

IV - REGRESSÃO ROBUSTA 54

IV.1 - Notação 55

IV.2 - Regressão Robusta usando o estimador Biponderado ... 58

IV.3 - Regressão Robusta usando o estimador Passoponderado. 62

IV.4 - Critério de Parada 65

IV.4.1 - É certa a convergência? 65

IV.4.2 - Alguns critérios para a determinação da convergência 66

IV.4.3 - Critério adotado 67

IV.4.5 - Como escolher ε e δ? 69

V - EXEMPLOS 76

V.1 - Exemplo 1 77

V.1.1 - Ajuste obtido por Mínimos Quadrados 79

V.1.2 - Ajustes obtidos pela Regressão Biponderada .. 84

V.2 - Exemplo 2 94

V.2.1 - Regressão Biponderada 99

VI - RESUMO E CONCLUSÕES 108

VII- CONSTRUÇÃO DO PROGRAMA PARA COMPUTADORES 111

VII.1 - Fluxograma ilustrativo do programa 113

VII.2 - Colocação dos dados 118

VII.2.1 - Instruções DATA 118

VII.2.2 - Dados referentes às opções 119

VII.2.3 - Dados de dimensionamento 121

VII.2.4 - Leitura do vetor de constantes e da matriz de delineamento 122

Página

VII.2.5 - Leitura de $\tilde{B}_o$ (IID=2) 123

VII.3 - Sub programas 124

VII.3.1 - Sub rotina ORDEM 124

VII.3.2 - Sub rotina ORRES 124

VII.3.3 - Sub rotina AMPIQ (IMD=1) 125

VII.3.4 - Sub rotina CMDN 126

VII.3.5 - Sub rotina XMAD (IMD=2) 126

VII.3.6 - Sub rotina BIPON (IPON=1) 127

VII.3.7 - Sub rotina STPON (IPON=2) 127

VII.3.8 - Sub rotinas para a construção dos gráficos 128

VII.4 - Programa para ajustes de regressão robusta 129

VII.5 - Regressão Biponderada através do SPSS 156

REFERÊNCIAS BIBLIOGRÁFICAS 162

LISTA DOS QUADROS

Quadro nº		Página
3.1.1 -	Resistência e Robustez em Eficiência de alguns estimadores de locação	34
3.1.1.e -	Valores de $\bar{X}$, X' e $\tilde{X}$ para vários valores de x	43
3.1.2.b -	Valores de $\bar{X}$, X' e $\tilde{X}$ para vários valores de x	48
5.1.1 -	Dados para o Exemplo 1	78
5.1.1.1 -	Valores de Y, $\tilde{Y}$, R e RO para o ajuste: $\tilde{Y} = 103,09738 + 1,43996\ X_1 - 0,61395\ X_4$	80
5.1.2.1 -	Valores de Y_9 devido às seis alterações provocadas ...	85
5.1.2.2 -	Estimativas de b_0, b_1 e b_4 obtidas por mínimos quadrados e pela regressão biponderada nos diferentes conjuntos de dados	86
5.1.2.3 -	Variação porcentual entre as estimativas obtidas pela regressão biponderada e as estimativas obtidas por mínimos quadrados com os dados originais	88
5.1.2.4 -	Quantidade de iterações necessárias para se obter convergência com a regressão biponderada, para os vários conjuntos de dados	89
5.2.1 -	Dados referentes a 21 dias de operação de uma planta, convertendo amônia em ácido nítrico	95
5.2.2 -	Resíduos obtidos com as equações (5.2.1.2), (5.2.1.3) e (5.2.1.4) com os dados do quadro 5.2.1 ...	98

Quadro nº | Página

5.2.3 - Variação porcentual entre os parâmetros das equações (5.2.1.5) e (5.2.1.3) e entre os parâmetros das equações (5.2.1.5) e (5.2.1.4) 100

5.2.4 - Valores de Y, $\bar{Y}$ e R para o ajuste:
$\bar{Y} = -37,31424 + 0,81089\ X_1 + 0,54005\ X_2 - 0,07060\ X_3$ 101

LISTA DAS FIGURAS

Figura nº		Página
2.2.1	- Curva de Influência (estilizada) para a Média Aritmética	11
2.2.2	- Curva de Influência (estilizada) para S^2_{n+1}	12
2.2.3	- Curva de Influência (estilizada) para a Mediana	14
2.5.1	- Dados hipotéticos exemplificando como um valor aberrante pode afetar a inclinação de uma reta (y = a + bx), que se pretende ajustar	25
2.5.2	- Dados hipotéticos exemplificando como um valor aberrante pode afetar a estimativa de a, ao se ajustar y = a + bx	26
3.1.1.a	- Curva de Influência (estilizada) para a média $\overline{X}$	38
3.1.1.b	- Curva de Influência (estilizada) para a mediana X'	39
3.1.1.c	- Curva de Influência (estilizada) para o estimador biponderado, $\bar{X}$, (apresentada por MOSTELLER e TUKEY (1977), cap. 14, S é dado por(3.1.1.1))	41
3.1.1.d	- Curva de Influência (estilizada) para o estimador biponderado, $\bar{X}$ (apresentada por MOSTELLER e TUKEY (1977), cap. 14, S=MDA)	42
3.1.2.a	- Curva de Influência (estilizada), para o estimador passoponderado, $\bar{X}$	46
a.1	- Curva de Influência (estilizada) para o M-estimador Seno	72

Figura nº — Página

5.1.1.2 - Gráfico da Distribuição Acumulada dos Resíduos 82

5.1.1.3 - Gráfico de Resíduos vs. Y ajustados 83

5.1.2.5 - Representação de um Desenho Esquemático com os pontos e as regiões correspondentes (DACHS (1978), pág. 14) 90

5.1.2.6 - Desenho Esquemático para os resíduos obtidos ao se fazer ajustes de mínimos quadrados e segundo a regressão biponderada no caso dos dados originais e dos dados alterados 92

5.2.5 - Desenho esquemático para os resíduos do quadro 5.2.4 103

5.2.6 - Gráfico da Distribuição Acumulada dos Resíduos 104

5.2.7 - Gráfico de Resíduos vs. Y ajustados 105

I - INTRODUÇÃO

A idéia de robustez não é tão nova como se pensa. Os primeiros informes datam do século XIX e são de autoria de Legendre, Laplace, Gauss e outros, conforme consta em STIGLER (1973). Entretanto, até pouco tempo atrás, não se deu muita ênfase a estas idéias. Só mais recentemente é que se iniciou o trato mais sistemático deste assunto, já então com o nome robustez, proposto por Box em 1953. HUBER (1964), pode-se dizer que pela primeira vez, apresenta os estimadores robustos para parâmetros de locação. Outro trabalho importante deste início é o de HAMPEL (1968). Com o desenvolvimento natural desta idéia, iniciou-se a busca de estimadores robustos para parâmetros de escala e mais recentemente a idéia foi extendida a métodos robustos de ajustes de regressão. O mais importante trabalho sobre estimação robusta de parâmetros de locação se deve a ANDREWS e outros (1972), onde não somente apresentam cerca de 80 estimadores robustos para parâmetros de locação, como também apresentam estimadores robustos para parâmetros de escala, utilizados pelos estimadores de parâmetros de locação, mas também comentam a

não inclusão de extensões, aparentemente lógicas, como é o caso da extensão aos problemas de regressão.

No caso de regressão robusta pode-se citar FORSYTE(1972), JAECKEL (1972), HUBER (1973), ANDREWS (1974 e 1975), BEATON e TUKEY (1974), HINICH e TALWAR (1975), TUKEY (1975), HILL e HOLLAND (1977), MOSTELLER e TUKEY (1977), TUKEY (1977), BICKEL (1978), DACHS (1978), MORINEAU (1978) e muitos outros. Como se pode perceber há muito pouco tempo é que se começou a tratar mais seriamente deste problema. Dos trabalhos acima citados, serão considerados aqui com mais destaque os de Andrews e os de Tukey.

Apresentar-se-á então um estimador robusto para parâmetros de locação e a sua extensão a problemas de regressão. Foi escolhido um estimador proposto por Tukey, denominado Estimador Biponderado (Biweight, em inglês). Produz resultados muito bons, tanto sob a pressuposição de normalidade como quando isto não se verifica. São basicamente duas as razões desta escolha. A primeira é a qualidade dos resultados obtidos e a segunda é a facilidade (computacional) na obtenção das estimativas. Há muitos outros estimadores robustos propostos para o problema de regressão, mas são muito mais difíceis de se obter (computacionalmente). Como exemplo disto pode-se citar o M-estimador baseado na função seno, apresentado por Andrews, em ANDREWS e outros (1972) (será apresentado neste trabalho no apêndice ao cap. IV), que fornece resultados bastante semelhantes, mas é de obtenção muito mais difícil. Requer um programa de otimização não-linear.

Outras razões para a escolha do estimador biponderado e sua extensão a problemas de regressão, substituindo o método clássico - o método dos mínimos quadrados - podem ser vistos em TUKEY (1975). Dentre elas pode-se citar a fragilidade do método clássico frente a dados com distribuição longe da normal ou frente a presença de valores discrepantes (ou outliers) ou mesmo frente à problemas de heterocedasticia e correlação entre as variáveis independentes. ANDREWS (1974), BEATON e TUKEY (1974), MOSTELLER e TUKEY (1977) e muitos outros reafirmam estas deficiências do método dos mínimos quadrados (além de outras) e comentam a insensibilidade dos métodos robustos a estes problemas; isto é, mesmo face a eles, os resultados que se obtém são bons. ANDREWS (1975) ainda comenta que não há mais razão em apoiar o método dos mínimos quadrados em virtude das facilidades computacionais que apresenta. Com o advento da modernização dos computadores isto não mais se justifica.

Procurou-se apresentar um trabalho onde cada novo conceito, ou cada nova idéia a ser apresentada estivesse fundamentada quase inteiramente nos conceitos anteriormente apresentados. Deste modo, no capítulo II se apresentam vários conceitos, uns já bem conhecidos, outros menos, que serão utilizados quando da introdução do estimador biponderado e de sua extensão a problemas de regressão. Se apresenta conceitos não muito conhecidos como por exemplo as Curvas de Influência (sec. II.2) e os Emparelhadores (matchers em inglês, sec. II.4) e outros já bastante conhecidos como os Valores Aberrantes.

No capítulo III, apresenta-se dois estimadores robustos para parâmetros de locação, o estimador biponderado e o passoponderado,

que apesar de utilizarem medidas robustas de escala, são apresentados antes delas, simplesmente porque esta é a seqüência já encarada como clássica; isto é, primeiro se apresenta estimadores para parâmetros de locação e depois estimadores para parâmetros de escala.

No capítulo IV se apresenta os métodos de regressão robusta, desenvolvidos a partir dos estimadores de locação apresentados no capítulo III. No final do capítulo IV se apresenta, apenas com o intuito de informação, o M-estimador seno, para parâmetros de locação e sua extensão a problemas de regressão. A razão disto é a comparação que será feita no exemplo 2, do capítulo V, entre o desempenho do estimador biponderado e o M-estimador seno, mostrando a proximidade dos resultados obtidos, ratificando assim a idéia inicial de apresentar um método que apresente bons resultados e seja de fácil obtenção.

No capítulo V se apresentará dois exemplos ilustrando o desempenho do estimador biponderado, utilizado para ajustar regressões. O primeiro será processado utilizando o SPSS (Statistical Package for the Social Sciences), de uma maneira que será explicada no capítulo VII; o segundo exemplo será processado com um programa em FORTRAN, cuja listagem e "manual do usuário", estarão no capítulo VII.

No capítulo VI apresentar-se-á as conclusões, não somente baseadas nos exemplos, mas também nos trabalhos que já foram feitos com o estimador biponderado.

No capítulo VII se apresentará os programas para computador e o modo de utilizá-los, não somente para obter regressão biponderada como também regressão passoponderada.

Nunca se pretendeu fazer um trabalho que cobrisse todos os ângulos do assunto. Deste modo continuam alguns problemas em aberto, que podem vir a ser objeto de estudos futuros. Um destes assuntos é a prova formal da ocorrência da convergência na obtenção dos parâmetros da regressão a ser ajustada, que deve ser um problema de difícil solução. Mas por outro lado, partindo de uma boa estimativa inicial, não há porque desconfiar da não convergência do método. Outro problema que apenas foi mencionado, no exemplo 3.1.1 é verificar o que ocorre ao se mudar a medida robusta de escala. Para este problema, MOSTELLER e TUKEY(1977) afirmam que não deve ocorrer grandes modificações, desde que se utilize boas medidas robustas de escala. Também não se verificou o comportamento dos estimadores apresentados ao variar o valor da constante de escalonamento, c (sec. III.1.1). Escolheu-se, para processar os exemplos do capítulo V, c=4, por ser um valor intermediário dentre os sugeridos em várias publicações.

Há também restrições quanto ao programa em FORTRAN a ser apresentado no capítulo VII, no que se refere à precisão das estimativas obtidas quando a matriz dos coeficientes do sistema ($X'P_gX$, sec.IV.2)for mal posta (mal comportada). Mas, está perfeitamente claro, no capítulo VII, o que deve ser feito para trocar esta parte do programa, por um procedimento que apresente melhores resultados.

II - NOMENCLATURA e DEFINIÇÕES

Neste capítulo serão apresentados a nomenclatura e o significado de termos a serem utilizados neste trabalho. Alguns destes conceitos já são bem conhecidos, como é o caso dos valores aberrantes. Outros são pouco conhecidos, como por exemplo as curvas de influência. Serão apresentados com bastante simplicidade, com a intenção de apenas fornecer uma idéia inicial, não se provando nenhum resultado.

II.1 - Resistência e Robustez

Não se tem a intenção de dar definições formais destes conceitos. O propósito é apresentar uma idéia geral de seus significados. HAMPEL (1968) apresenta duas definições, bastante relacionadas, de robustez, para uma sequência de estimadores, e diz serem quase que universalmente aplicáveis.

MOSTELLER e TUKEY (1977) dão uma boa idéia do que seja resistência; é uma propriedade bastante desejável para um estimador. Suponha que se altere uma parte dos dados, possivelmente de modo drástico.

Esta alteração pode afetar substancialmente o estimador que se está usando. Neste caso diz-se que este estimador não é resistente. Por outro lado, se uma alteração em uma pequena parte dos dados não altera substancialmente os resultados de um estimador, ele é dito ser resistente. A média aritmética é um protótipo de estimador não resistente. Se em

$$\frac{1 + 2 + 2 + 3 + \ldots + 23}{101} = 9{,}58$$

troca-se o segundo valor, 2, por 101002, a média irá mudar para:

$$\frac{1 + 101002 + 2 + 3 + \ldots + 23}{101} = 1009{,}58$$

Alterando-se menos de 1% dos dados, a média foi bastante modificada. A mediana é um protótipo, dos mais simples, de estimador resistente. Se no exemplo anterior, com 101 valores, houvesse:

50 valores {1; 2; 2; 3; ... ; 8, 9}
1 valor {9}
50 valores {9,5; 10; ... ; 23}

onde a mediana é 9, fazendo-se a mesma alteração anterior; isto é, de 2 para 101002, a mediana mudaria de 9 para 9,5, não havendo praticamente grandes alterações.

BREIMAN (1973) apresenta uma boa idéia do sentido de robustez. Assuma que se está trabalhando com dados de uma "pequena" família de possíveis distribuições $\{P_\theta\}$, $\theta \in \Theta$ e se está procurando estimadores com um desempenho, ou eficiência, relativamente alta, não sendo

muito sensíveis a pequenos desvios da família de distribuições assumidas como verdadeira. Esta é a idéia da estimação robusta.

HAMPEL (1973) comenta sobre os objetivos principais da estimação robusta. É a construção de salvaguardas contra grandes quantidades de erros grosseiros nos dados, colocando um limite na influência de contaminações escondidas e valores discrepantes (seção II.5), isolando estes valores aberrantes para tratamento em separado (se desejado), mantendo nos modelos paramétricos um comportamento bastante bom.

STIGLER (1973) praticamente resume as duas idéias anteriores ao afirmar:

- "Os cientistas têm-se preocupado com o que poderíamos chamar - robustez - insensividade dos procedimentos a desvios das pressuposições, particularmente da pressuposição de normalidade...".

Deve-se também sempre ter em mente que a eficiência é uma propriedade bastante desejável. MOSTELLER e TUKEY (1977) falam sobre Robustez em Eficiência; isto é, deseja-se grande eficiência em uma variedade de situações em vez de em uma particular situação. Deve-se prestar atenção à menor eficiência que se pode obter em um conjunto razoável de situações. Se esta menor eficiência for alta, pode-se dizer que se tem um bom estimador.

Do que já foi dito, tem-se então: Um estimador é resistente se for pouco afetado pela presença de uma parcela de valores aberrantes e é robusto se sua eficiência é limitada inferiormente por um valor

que não é muito menor do que o da eficiência do "melhor" estimador sob as condições supostas (usualmente de independência, normalidade, homocedasticia, etc.).

II.2 - Curvas de Influência

HAMPEL (1968) apresenta a definição do que denominou Curva de Influência. ANDREWS e outros (1972) se utilizam destas curvas no estudo que fazem sobre estimadores robustos de parâmetros de locação. MOSTELLER e TUKEY (1977) também as utilizam.

Estas curvas podem ser utilizadas como uma ferramenta para o estudo dos estimadores. São formas limites a partir das quais as propriedades assintóticas podem ser determinadas.

HAMPEL (1974) diz que podem ser utilizadas não somente para determinar variâncias assintóticas dos estimadores, mas também para estudar propriedades locais de robustez que são definidas e intuitivamente interpretadas. O estudo destas curvas serve para aprofundar a compreensão dos estimadores e de seu comportamento. Servem também para desenvolver novos estimadores com propriedades robustas pré-especificadas.

Dar-se-á agora alguns exemplos de Curvas de Influência Estilizadas, no sentido apresentado por ANDREWS e outros (1972), por serem obtidas com base em uma amostra finita de valores. Para a obtenção destas curvas procede-se do seguinte modo:

- tem-se n observações $x_1, x_2, \ldots, x_n$.

- acrescenta-se uma observação aos dados, x_{n+1}, e se verifica o comportamento do estimador em questão, fazendo esta (n+1)-ésima observação assumir todos os possíveis valores do seu campo de variação. Além disso suponha também que estes valores da amostra podem assumir qualquer valor real.

II.2.1 - Curva de Influência para Média

Suponha que se tem uma amostra $x_1, x_2, \ldots, x_n$ e que a média seja $\bar{x}_n$. Como se comportará a média, ao se acrescentar uma nova observação, x_{n+1} à amostra? Sabe-se que:

- se $x_{n+1} \geq \bar{x}_n$ então $\bar{x}_{n+1} \geq \bar{x}_n$
- se $x_{n+1} < \bar{x}_n$ então $\bar{x}_{n+1} < \bar{x}_n$

Obviamente

$$\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}$$

que nada mais é que a equação de uma reta, com coeficiente angular igual a 1/(n+1) e intersecção com o eixo das ordenadas em $n\bar{x}_n/(n+1)$. A figura 2.2.1 apresenta esta curva.

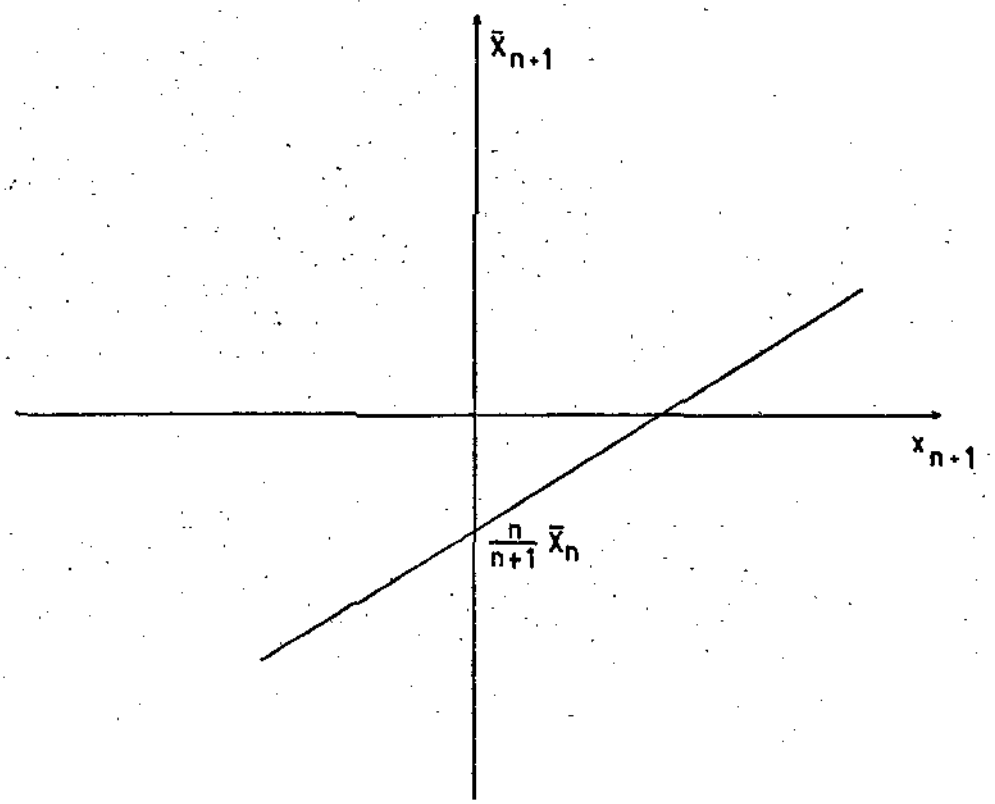
Apenas como curiosidade será apresentada a Curva de Influência para a média, como consta em HAMPEL (1974). Suponha que X seja uma variável aleatória com distribuição F. A média

$$T = \int x dF(x)$$

é definida para todas as medidas de probabilidade onde exista primeiro

momento. Suponha que X tenha média μ, conhecida.

Figura 2.2.1 - Curva de Influência (estilizada) para a Média Aritmética.



- Então a curva de influência de T é definida em F e é dada por:

$$CI_{T,F}(x) = \lim_{\varepsilon \to 0} \left[(1-\varepsilon)\,\mu + x\varepsilon - \mu\right] / \varepsilon = x-\mu,\ x \in R$$

Com base na curva de influência estilizada apresentada na figura 2.2.1 pode-se obter as propriedades da média. Nota-se que é bastante sensível a valores extremos; isto é, se $|x_{n+1}| \to \infty$, então $|\bar{x}_{n+1}| \to \infty$. Isto caracteriza a extrema falta de resistência da média. Além disto, vê-se que a dependência de $\bar{x}_{n+1}$ de x_{n+1} é linear; isto é; $\bar{x}_{n+1}$ varia linearmente com x_{n+1}.

II.2.2 - Curva de Influência para S^2

Tem-se que

$$S_n^2 = \sum_{i=1}^{n} (x_i-\mu)^2/n$$

Para maior simplicidade suponha que $E(X)=\mu=0$, e então S^2 se reduzirá a:

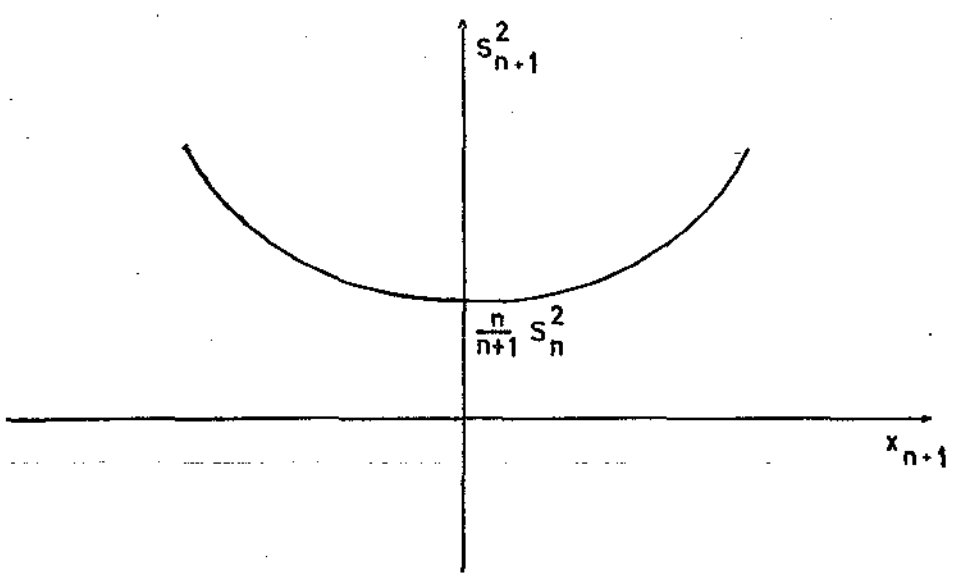
$$S_n^2 = \sum_{i=1}^{n} x_i^2 /n$$

Como é afetada S_n^2, por uma simples observação x_{n+1}, acrescida aos dados? Tem-se que:

$$S_{n+1}^2 = \frac{nS_n^2 + x_{n+1}^2}{n+1} = \frac{1}{n+1} x_{n+1}^2 + \frac{n}{n+1} S_n^2$$

que é uma equação do segundo grau da forma $ax^2 + bx + c$, com $a=1/(n+1)$, $b=0$ e $c=nS_n^2/(n+1)$. A figura 2.2.2 apresenta esta curva.

Figura 2.2.2 - Curva de Influência (estilizada) para S_{n+1}^2.



HAMPEL (1974) também apresenta a curva de influência para S^2. Além das suposições feitas no caso da média, suponha também que a variável X tenha variância σ^2 (conhecida), dada por:

$$T = \int (x-\mu)^2 \, dF(x)$$

e a curva de influência é então:

$$CI_{T,F}(x) = \lim_{\varepsilon \to 0} \left[(1-\varepsilon)\sigma^2 + \varepsilon(x-\mu)^2 - \sigma^2\right]/\varepsilon = (x-\mu)^2 - \sigma^2, \; x \in R$$

Com base na curva apresentada na figura 2.2.2 pode-se obter as propriedades de S^2. Nota-se que este estimador é bastante sensível a valores extremos, pois se $|x_{n+1}| \to \infty$ então $S^2_{n+1} \to \infty$. Se $|x_n| < S_n$ então $S^2_{n+1} < S^2_n$; isto é a variância estimada decresce, no máximo até $nS^2_n/(n+1)$ quando $|x_{n+1}| \to 0$. A curva é ilimitada nos extremos, apresentando a característica da não-resistência de S^2. Pode-se afirmar também que a dependência de S^2_{n+1} de x_{n+1} é quadrática; isto é, S^2_{n+1} varia quadraticamente com x_{n+1}.

II.2.3 - Curva de Influência para a Mediana

Suponha agora que se tenha uma amostra, já ordenada, isto é, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, suponha também, sem perda de generalidade, que $\underline{n}$ é um número par*. Qual será o comportamento da mediana, $\underline{x}'$, ao se acrescentar uma nova observação, x_{n+1}, onde x_{n+1} não está ordenada? Sabe-se que, como $\underline{n}$ é par:

---

* O raciocínio é análogo para n ímpar e a forma da curva de influência (estilizada) é a mesma.

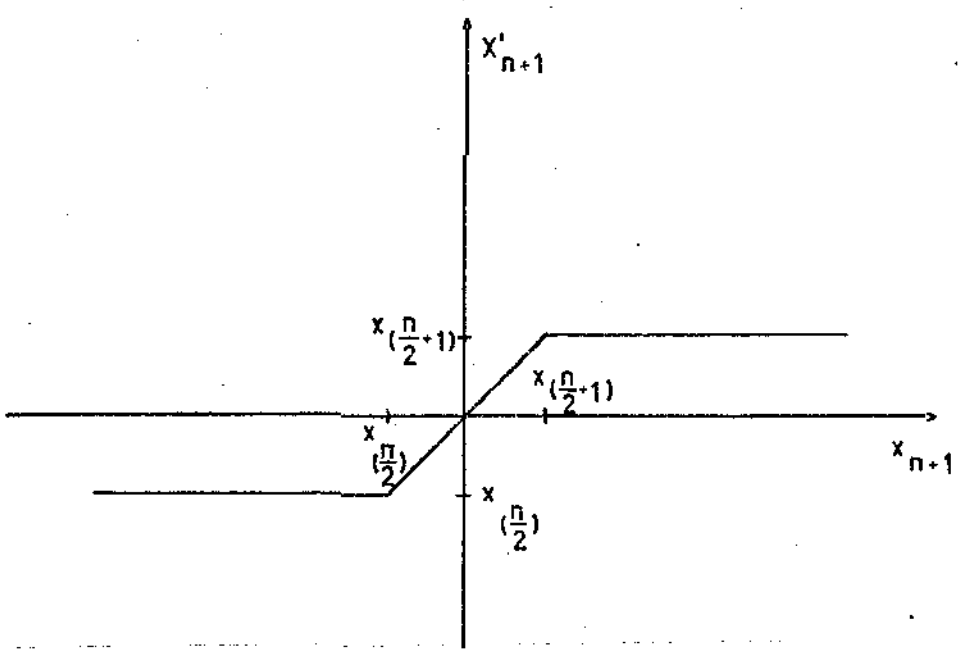
$$x'_n = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right]$$

ao se acrescentar x_{n+1} obtém-se:

$$x'_{n+1} = \begin{cases} x_{\left(\frac{n}{2}\right)} & \text{se } x_{n+1} \leq x_{\left(\frac{n}{2}\right)} \\ x_{\left(\frac{n}{2}+1\right)} & \text{se } x_{n+1} > x_{\left(\frac{n}{2}+1\right)} \\ x_{n+1} & \text{se } x_{\left(\frac{n}{2}\right)} < x_{n+1} \leq x_{\left(\frac{n}{2}+1\right)} \end{cases}$$

A figura 2.2.3 mostra a curva de influência que se obtém.

Figura 2.2.3 - Curva de Influência (estilizada) para a Mediana



Obs.- a Curva de Influência para a mediana, no caso de $x_{\left(\frac{n}{2}\right)} < x_{n+1} \leq x_{\left(\frac{n}{2}+1\right)}$ é uma reta com coeficiente angular igual a 1.

Através da figura 2.2.3 pode-se concluir que a mediana é insensível a valores extremos, caracterizando assim a sua resistência. Por outro lado, na porção central dos dados, reage linearmente aos valores da (n+1)-ésima observação, assim como a média.

ANDREWS e outros (1972) apresentam Curvas de Influência (estilizadas) para diversos estimadores apresentados no seu estudo de estimadores robustos para parâmetros de locação.

Nas secções III.1.1 e III.1.2 e também no apêndice ao capítulo IV, serão apresentadas Curvas de Influência (estilizadas) para, respectivamente, o estimador biponderado, o estimador passoponderado e o M-estimador seno.

II.3 - Variáveis e Suportes

Em regressão se dispõe de um vetor de respostas, ou vetor das observações que contém os valores de uma variável denominada dependente e de uma lista de valores correspondentes a uma ou mais variáveis, denominadas independentes. Por exemplo

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

é o vetor de respostas e $X_1, X_2, \ldots, X_k$ são as variáveis independentes. Pode-se tentar, por exemplo, ajustar uma regressão de Y nas variáveis

$X_1, X_2, \ldots, X_k$, da forma:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \quad ; \quad i=1,2,\ldots,n \qquad (2.3.1)$$

No caso onde houver apenas uma variável independente, X, e se quiser ajustar:

$$y_i = b_1 x_i^1 + b_2 x_i^2 + \ldots + b_k x_i^k \quad ; \quad i=1,2,\ldots,n$$

não se pode chamar $X^1, X^2, X^3, \ldots, X^k$ de variáveis, pois na realidade só há uma variável independente, a variável X. O mesmo vale para:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + b_3 (x_{1i} \cdot x_{2i}) + \ldots \quad ; \quad i=1,2,\ldots,n$$

Em casos como estes surge a necessidade de se dar um novo nome às "variáveis" $X^2, X^3, \ldots, X^k, (X_1 \cdot X_2), \ldots$. BEATON e TUKEY (1974), MOSTELLER e TUKEY (1977), dentre outros, sugerem o nome Suportes (carriers). Com esta denominação pode-se falar da regressão da variável dependente Y nos suportes $X^1, X^2, \ldots, X^k$ e também da regressão de Y nos suportes $X_1, X_2, (X_1 \cdot X_2), \ldots$. Em (2.3.1) tanto se pode falar das variáveis como dos suportes $X_1, X_2, \ldots, X_k$.

Deste ponto em diante somente se usará o termo suporte, significando tanto uma lista de variáveis simples como também uma lista onde as "variáveis" podem ser funções de uma ou mais variáveis simples.

II.4 - Emparelhadores e Sintonizadores

Estes dois termos foram introduzidos por Tukey. Em inglês são, respectivamente, "matcher" e "catcher". Não se tentou dar uma

tradução exata a estes dois nomes. O que se fez foi associar a cada um deles uma palavra que melhor indicasse seus significados.

II.4.1 - Emparelhadores

TUKEY (1975) e MOSTELLER e TUKEY (1977) apresentam este conceito.

Suponha que se deseje ajustar uma regressão de Y nos suportes $X_1, X_2,\ldots,X_k$. Suponha ainda que se deseja ajustar:

$$y_i = b_1 x_{1i} + b_2 x_{2i} +\ldots+b_1 x_{ki} \quad ; \quad i=1,2,\ldots,n \qquad (2.4.1.1)$$

e que se obtêm $\hat{b}_1, \hat{b}_2,\ldots,\hat{b}_k$ como estimativas dos parâmetros.

Então:

$$\hat{y}_i = \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} +\ldots+ \hat{b}_k x_{ki} \quad ; \quad i=1,2,\ldots,n \qquad (2.4.1.2)$$

é o ajuste obtido.

Se houver um conjunto de coeficientes $H=\{h(i); i=1,2,\ldots,n\}$, vai-se chamar a este conjunto de Emparelhador se e somente se;

$$\sum_{i=1}^{n} h(i)\, y_i = \sum_{i=1}^{n} h(i)\, \hat{y}_i \qquad (2.4.1.3)$$

onde y_i e $\hat{y}_i$; $i=1,2,\ldots,n$ são dados, respectivamente, por (2.4.1.1) e (2.4.1.2). Todo e qualquer conjunto que satisfizer (2.4.1.3) será um emparelhador para o ajuste.

Por exemplo, ao se ajustar $y = a+bx$, por mínimos quadrados, $H_1 = \{1,1,\ldots,1\}$ é um emparelhador, pois exigir que:

$$\sum_{i=1}^{n} 1.\, y_i = \sum_{i=1}^{n} 1.\, \hat{y}_i$$

é o mesmo que exigir que:

$$n\,\bar{y} = n\,(\hat{a}+\hat{b}\,\bar{x}) \qquad (2.4.1.4)$$

Também $H_2 = \{x_1,x_2,\ldots,x_n\}$ é um emparelhador, pois exigir que:

$$\sum_{i=1}^{n} x_i\, y_i = \sum_{i=1}^{n} x_i\, \hat{y}_i$$

é o mesmo que exigir que:

$$\sum_{i=1}^{n} x_i\, y_i = \hat{a} \sum_{i=1}^{n} x_i + \hat{b} \sum_{i=1}^{n} x_i^2 \qquad (2.4.1.5)$$

e (2.4.1.4) e (2.4.1.5) podem ser reconhecidas como as equações que formam o chamado Sistema de Equações Normais, que resolvido dá a solução de mínimos quadrados:

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

como pode ser encontrado em vários textos clássicos como por exemplo DRAPER e SMITH (1966), cap. 1; SEARLE (1971), cap. 3 dentre outros.

Ao se ajustar uma regressão linear múltipla, na forma matricial, através de mínimos quadrados, tem-se que a matriz X', a matriz transposta da matriz de delineamento X, é um emparelhador, pois exigir que:

$$X'Y = X'\hat{Y}$$

é o mesmo que exigir:

$$X'Y = X'X\,\hat{B} \Leftrightarrow \hat{B} = (X'X)^{-1}\,X'Y$$

se (X'X) for de posto completo. Este resultado é bem conhecido e é a solução, de mínimos quadrados, como pode ser vista em DRAPER e SMITH(1966), cap. 2; SEARLE (1971), cap. 3 dentre outros.

MOSTELLER e TUKEY (1977) afirmam que os emparelhamentos vêm em "feixes". Se H = {h(i); i=1,2,...,n} e K = {k(i); i=1,2,...,n } são ambos emparelhadores, então:

$$\left[c_h H + c_k K\right] = \left\{\left[c_h h(i) + c_k k(i)\right] \;;\; i=1,2,\ldots,n\right\}$$

onde c_h e c_k são constantes arbitrárias, também é um emparelhador. Em outras palavras, todas as somas ponderadas ou combinações lineares de emparelhadores também são emparelhadores.

Por exemplo, ao se ajustar, por mínimos quadrados, uma regressão linear múltipla, ponderada pela inversa da matriz de variâncias e covariâncias, V, tem-se que $X'V^{-1}$ é um emparelhador, pois exigir que:

$$(X'V^{-1})\,Y = (X'V^{-1})\,\hat{Y}$$

é o mesmo que:

$$(X'V^{-1})Y = (X'V^{-1})X\,\hat{B} \Leftrightarrow \hat{B} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

se $(X'V^{-1}X)$ for de posto completo, e $\tilde{B}$ pode ser reconhecido como a solução de mínimos quadrados, como pode ser visto em SEARLE (1971), cap. 3, dentre outros.

II.4.2 - Sintonizadores

Será apresentado agora, apenas com o intuito de informação, um conjunto especial de emparelhadores, que podem tornar mais fácil a obtenção dos $\hat{b}$'s e que contenham toda a informação sobre $\tilde{B}$. A apresentação será feita de acordo com MOSTELLER e TUKEY (1977).

Suponha que se quer um ajuste da forma de (2.4.1.1), através de um processo que pode ser descrito por emparelhadores. Quanto sobre $\hat{b}_1$, por exemplo, pode ser coletado por um único emparelhador? Se for possível achar um emparelhador $H = \{h(i);\ i=1,2,\ldots,n\}$ tal que:

$$0 = \sum_{i=1}^{n} h(i)\, x_{2i} = \sum_{i=1}^{n} h(i)\, x_{3i} = \ldots = \sum_{i=1}^{n} h(i)\, x_{ki} \qquad (2.4.2.1)$$

de modo que:

$$\sum_{i=1}^{n} h(i)\, y_i = \sum_{i=1}^{n} h(i)\, \tilde{y}_i = \sum_{i=1}^{n} h(i) \left[\hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \ldots + \hat{b}_k x_{ki}\right] =$$

$$= \hat{b}_1 \sum_{i=1}^{n} h(i)\, x_{1i} \qquad (2.4.2.2)$$

então:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} h(i)\, y_i}{\sum_{i=1}^{n} h(i) x_{1i}}, \quad \sum_{i=1}^{n} h(i)\, x_{1i} \neq 0$$

Pode-se dizer que H está "sintonizado" a $\hat{b}_1$, pois, $\hat{b}_2, \hat{b}_3, \ldots, \hat{b}_k$ não aparecem em (2.4.2.2). Se é possível "dessintonizar" todos os $\hat{b}$'s, menos um, pode-se facilmente resolver a equação para este.

Por exemplo, ao se ajustar $y = \mu + b(x-\bar{x})$, $x-\bar{x}$ está sintonizado a $\underline{b}$ e dessintoniza $\underline{\mu}$, como se pode ver a seguir. Seja então o conjunto

$$H = \{x_1-\bar{x},\ x_2-\bar{x},\ldots,x_n-\bar{x}\}$$

Este conjunto é um emparelhador, pois exigir que

$$\sum_{i=1}^{n} h(i)\, y_i = \sum_{i=1}^{n} h(i)\, \hat{y}_i$$

é o mesmo que:

$$\sum_{i=1}^{n} h(i)\, y_i = \sum_{i=1}^{n} (x_i-\bar{x})(\hat{\mu}+\hat{b}(x_i-\bar{x})) = \hat{\mu} \cancel{\sum_{i=1}^{n} (x_i-\bar{x})}^{0} + \hat{b} \sum_{i=1}^{n} (x_i-\bar{x})^2$$

que pode ser reescrito como:

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})\, y_i}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

que pode ser reconhecido como a estimativa obtida por mínimos quadrados, conforme consta em vários textos clássicos.

Um exemplo mais complexo é o seguinte: Suponha que $X_1 \equiv 1$, $X_2 \equiv x$ e $X_3 \equiv x^2$, onde x assume os valores 1,2,3,...,10. Deseja-se ajustar:

$$y = b_1x_1 + b_2x_2 + b_3x_3$$

Então $k(22-11x + x^2)$, onde k é uma constante, está sintonizado a b_3. Para verificar basta ver se a condição (2.4.2.1) é satisfeita; isto é, se:

$$0 = \sum_{i=1}^{10} (22-11x + x^2) \cdot 1 = \sum_{i=1}^{10} (22-11x + x^2) \cdot x$$

lembrando que:

$$\sum_{i=1}^{n} 1 = n$$

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(n+2)}{6}$$

e

$$\sum_{i=1}^{n} i^3 = \left[\frac{n(n+1)}{2}\right]^2$$

verifica-se rapidamente que:

$$\sum_{1}^{10} (22-11x + x^2) = 220 - \frac{11.10.11}{2} + \frac{10.11.12}{6} = 0$$

e

$$\sum_{1}^{10} (22-11x + x^2) \cdot x = 22 \cdot \frac{10.11}{2} - 11 \cdot \frac{10.11.12}{6} + \left(\frac{10.11}{2}\right)^2 = 0$$

e portanto $k\,(22-11x + x^2)$ está sintonizado a b_3.

Se há emparelhadores em número suficiente para garantir uma solução única, pode-se mostrar que sempre haverá um emparelhador

sintonizado com cada $\tilde{b}_i$, i=1,2,...,k. Sejam h_1, h_2,...,h_k, k emparelhadores linearmente independentes, então tudo o que é necessário é achar um conjunto de d's; d_1, d_2,...,d_k que satisfaçam as k-1 equações:

$$\sum_i (d_1h_1 + d_2h_2 + \ldots + d_kh_k)\, x_{ji} = 0,\ j \neq M\ (1 \leq j \leq k;\ 1 \leq M \leq k)$$

que são equivalentes à condição (2.4.2.1) pois dessintonizam todos os $\tilde{b}$'s menos $\tilde{b}_M$, e também devem satisfazer:

$$\sum_i (d_1h_1 + d_2h_2 + \ldots + d_kh_k)\, x_{Mi} = 1^*$$

É necessário que este conjunto de k equações((k-1) + 1) nos d's tenham determinante diferente de zero para que haja solução única. Mais explicitamente:

$$\begin{vmatrix} \sum_i h_1x_1 & \sum_i h_1x_2 & \cdots & \sum_i h_1x_k \\ \sum_i h_2x_1 & \sum_i h_2x_2 & \cdots & \sum_i h_2x_k \\ \vdots & \vdots & & \vdots \\ \sum_i h_kx_1 & \sum_i h_kx_2 & \cdots & \sum_i h_kx_k \end{vmatrix} \neq 0$$

Seja então

$$C_M = d_1h_1 + d_2h_2 + \ldots + d_kh_k$$

o emparelhador C_M não é somente um emparelhador, sintonizado à $\tilde{b}_M$, assim

---

* na realidade este somatório pode valer qualquer constante, diferente de zero. Prefere-se a unidade apenas por facilidade de cálculos.

como qualquer de seus múltiplos poderia ser, ele é mais do que isto; é um Sintonizador para $\hat{b}_M$. Ao se associar y e $\hat{y}$ utilizando este emparelhador se obtém:

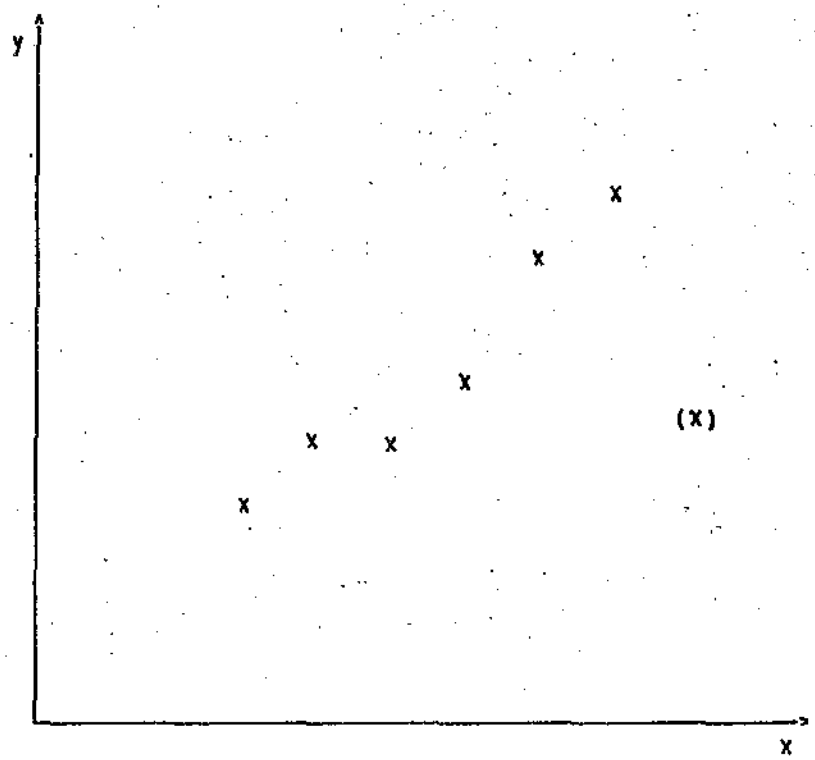
$$\sum_i C_M(i)\, y_i = \hat{b}_M \cdot 1$$

II.5 - Valores discrepantes ou aberrantes

AFIFI e AZEN (1974) afirmam que os valores aberrantes não são erros, mas sim observações que diferem, em magnitude, das restantes e devem ser tomadas como provenientes de uma população, que não a em estudo (nem todos concordam com a última parte desta afirmação).

DANIEL e WOOD (1971) afirmam que grandes conjuntos de dados, e ocasionalmente também os pequenos, às vezes contêm uma pequena quantidade de pontos discrepantes (wild points), algumas vezes chamados "mavericks" ou "outliers", por isto deve-se examinar os dados para encontrar estes ocasionais valores aberrantes. Ao tentar ajustar uma regressão linear simples ($y = a + bx$), dependendo da localização destes valores discrepantes, estes podem afetar as estimativas tanto do ponto de intersecção da reta com o eixo y, $\hat{a}$, como também a inclinação, $\hat{b}$. A figura 2.5.1 apresenta um exemplo (hipotético) de como uma "má" observação, pode afetar a inclinação da reta; isto é, afetar a estimativa $\hat{b}$.

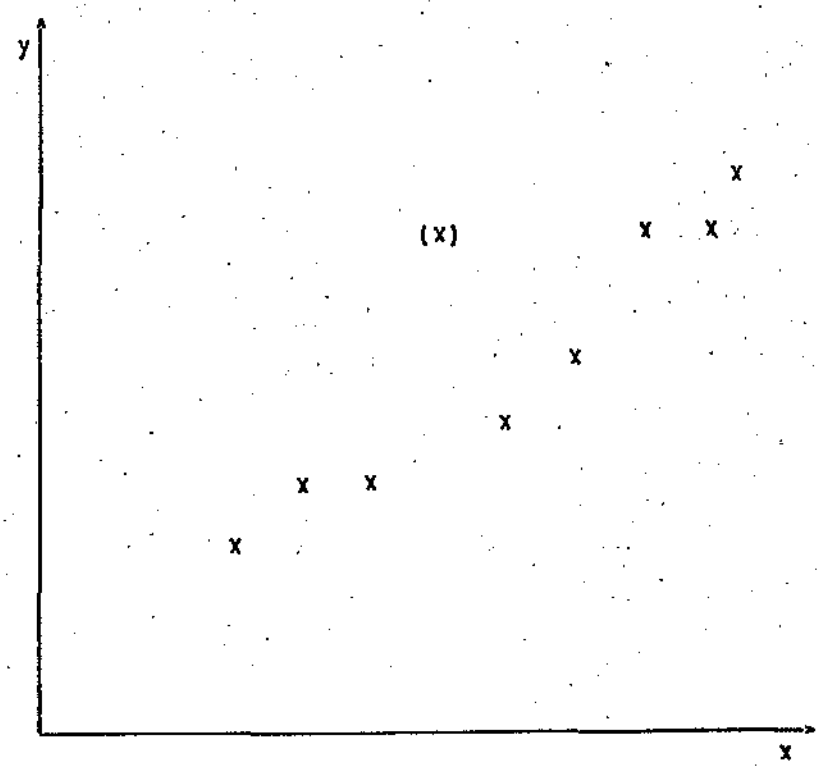
Figura 2.5.1 - Dados hipotéticos exemplificando como um valor aberrante pode afetar a inclinação de uma reta ($y = a + bx$), que se pretende ajustar.



(x) valor aberrante

Neste caso deve-se tomar cuidado pois pode ser que não se tenha, na realidade, um valor aberrante, e, em vez disto, a verdadeira equação é que pode ser uma curva. A figura 2.5.2. apresenta um exemplo (hipotético) com uma "má" observação no centro dos dados. Esta observação pouco influi na inclinação da reta, mas afetará bastante a estimativa $\hat{a}$.

Figura 2.5.2 - Dados hipotéticos exemplificando como um valor aberrante pode afetar a estimativa de a, ao se ajustar y = a + bx.



(x) valor aberrante.

DRAPER e SMITH (1966) fazem uma colocação mais comentada. Uma dada observação, em um conjunto de dados, pode ser considerada um valor aberrante se estiver muito distante das restantes, talvez 3 ou 4 desvios padrões distante da média, por exemplo. Além desta tentativa de quantificar o que pode ou não ser considerado valor aberrante, comentam sobre o tratamento que se deve dar à estas observações. Um valor discrepante tem peculiaridades que o diferenciam do restante dos dados. Então deve-se submetê-los a um exame bastante cuidadoso para verificar as razões destas peculiaridades.

A utilização do desvio-padrão para a rejeição e/ou detecção de valores discrepantes deve ser evitada. Isto porque, para o seu cálculo se utiliza a média amostral, que, como já foi visto anteriormente, é bastante sensível a valores aberrantes, e também o próprio desvio padrão não é uma medida robusta de escala. Ao se calcular o desvio padrão em um conjunto de dados onde há a presença de valores discrepantes, obtém-se um valor "inflacionado". Quando se utiliza do desvio padrão pode-se considerar um valor aberrante como não sendo valor aberrante, pois na presença destes valores o desvio padrão cresce muito, como pode ser visto na seção II.2.2. Ao invés disto, deve-se usar uma medida robusta de escala, como por exemplo as que serão apresentadas na secção III.2. As medidas robustas de escala são pouco ou quase nada afetadas pela presença de valores aberrantes e, então, observações que na realidade são aberrantes e não foram detectadas pelo desvio padrão seriam detectadas. Resumindo, não se deve utilizar procedimentos para a detecção e/ou rejeição de valores discrepantes que sejam afetados por estes, e sim procedimentos baseados em medidas robustas, que não são afetados pelos valores aberrantes.

Já foram propostas muitas regras para a rejeição de valores aberrantes. A rejeição automática destes dados nem sempre é um bom procedimento. Muitas vezes eles podem estar fornecendo informações que as outras observações não podem prestar, devido ao fato de talvez serem resultantes de uma combinação não usual das circunstâncias, que podem ser de interesse vital, e, portanto, requerem investigações posteriores em vez de rejeição. Como uma regra geral, deve-se rejeitar valores

discrepantes somente se forem provenientes de erros de marcação ou medida dos dados, ou erro na montagem e/ou funcionamento dos aparelhos de medição. Caso contrário deve-se proceder à uma cuidadosa investigação sobre as possíveis causas de seu aparecimento. ANSCOMBE (1960), apresenta maiores detalhes sobre a rejeição de valores aberrantes.

Os métodos de regressão a serem apresentados no capítulo IV, de um certo modo, tanto detectam como rejeitam os valores aberrantes. Nos exemplos, do capítulo V, percebe-se claramente este fato.

III - ESTIMADORES ROBUSTOS DE LOCAÇÃO E DE ESCALA

Há vários tipos de estimadores robustos para parâmetros de locação, como por exemplo os L-estimadores, os R-estimadores e os M-estimadores. Neste trabalho somente serão tratados os M-estimadores. Serão apresentados dois estimadores robustos para parâmetros de locação, o estimador biponderado e o passoponderado, obtidos como uma aproximação dos M-estimadores.

Se apresentará também três estimadores robustos para parâmetros de escala, utilizados para a obtenção dos estimadores robustos para parâmetros de locação.

III.1 - M-, m- e w-estimadores

Considere duas funções $\Psi(\mu)$ e $w(\mu)$ relacionadas por:

$$\Psi(\mu) = \mu \cdot w(\mu)$$

onde Ψ é uma função ímpar e w uma função par, em μ.

ANDREWS e outros (1972) definem T, o M-estimador de locação, como a solução de:

$$\sum_{j=1}^{n} \Psi\left(\frac{x_j - T}{S}\right) = 0$$

onde S, uma medida robusta de escala, é estimada tanto simultânea como independentemente de uma equação da forma:

$$\sum_{j=1}^{n} \chi\left(\frac{x_j - T}{S}\right) = 0$$

Por exemplo, Huber, em ANDREWS e outros (1972) propõe uma família de estimadores caracterizada pela função Ψ da forma:

$$\Psi(x; k) = \begin{cases} -k, & x < k \\ x, & -k < x < k \\ k, & x > k \end{cases} \qquad (a)$$

e pela função

$$\chi(x) = \Psi^2(x; k) - \beta(k) \qquad (b)$$

onde

$$\beta(k) = \int \Psi(x; k)^2 \, \Phi(dx)$$

e Φ é a distribuição de uma variável aleatória normal com média igual a zero e variância um. As equações (a) e (b) são então resolvidas simultaneamente para S e T. São resolvidas iterativamente, iniciando, por exemplo com a mediana para T e

$$S = \frac{\text{amplitude interquartís}}{1{,}35}$$

Outros exemplos de M-estimadores podem ser encontrados em ANDREWS e outros (1972), onde não somente são apresentados como também

estudados comparativamente com outros estimadores robustos de locação.

BEATON e TUKEY (1974) definem o M-estimador de locação T, como as soluções tanto de:

$$\sum_{j=1}^{n} \Psi\left(\frac{x_j - T}{cS}\right) = 0 \qquad (3.1.1)$$

como de

$$\sum_{j=1}^{n} w\left(\frac{x_j - T}{cS}\right)(x_j - T) = 0 \qquad (3.1.2)$$

onde c é uma constante de escalonamento.

ANDREWS e outros (1972) definem os M-estimadores de um passo (one step M-estimators). São a primeira aproximação de Gauss-Newton para os M-estimadores. BEATON e TUKEY (1974) apresentam as soluções de (3.1.1) e (3.1.2) obtidas pelo método de Newton-Raphson. No caso de (3.1.1) obtém-se:

$$T_1 = T_0 + \frac{\Sigma\Psi\left(\frac{x_j - T_0}{cS}\right)}{\Sigma\Psi'\left(\frac{x_j - T_0}{cS}\right)} \; cS$$

onde $\Psi'(\mu) = \delta\Psi(\mu)/\delta\mu$. Chamam a esta solução de m-estimador (m^1-estimador) ou M-estimador de um passo.

Para (3.1.2) se obtém, por Newton-Raphson:

$$T_1 = \frac{\Sigma w\left(\frac{x_j - T_0}{cS}\right) x_j}{\Sigma w\left(\frac{x_j - T_0}{cS}\right)}$$

e denominam esta solução w-estimador (w^1-estimador).

Quando somente um (ou poucos) passos são tomados com Ψ' no denominador, BEATON e TUKEY (1974) se referem aos m^1-(m^2-,m^3-,...) estimadores. Analogamente, no caso de (3.1.2), se referem aos w^1-(w^2-,w^3-,...) estimadores.

ANDREWS e outros (1972), BEATON e TUKEY (1974) e HAMPEL (1974) afirmam que:

- os M-estimadores podem ser muito bons
- os m-estimadores, tomando um passo a partir da mediana são de qualidade muito próxima aos M-estimadores. Onde não somente é importante ter $\Psi(\mu) \cong \mu$ quando $\mu \cong 0$, como também $\Psi(\mu)$ ser limitado e também $\Psi(\mu) \to 0$ se $|\mu|$ é grande.

Além disso, BEATON e TUKEY (1974), afirmam que os w-estimadores de locação, tomando um passo a partir da mediana, são de qualidade muito parecidas com os M- e os m- estimadores.

III.1.1 - O estimador Biponderado (biweight)*

BEATON e TUKEY (1974) e MOSTELLER e TUKEY (1977) apresentam o seguinte w-estimador, denominado biponderado:

---

* Biweight - abreviação de bisquared weighted.

$$w(\mu i) = \begin{cases} (1-\mu_i^2)^2 \; ; \; \mu_i^2 \leq 1 \\ 0 \quad\quad ; \; \mu_i^2 > 1 \end{cases} \qquad i=1,2,\ldots,n$$

com

$$\mu_i = \frac{x_i - \bar{x}}{cS} \qquad ; \qquad i=1,2,\ldots,n$$

e

$$\bar{x} = \frac{\sum_{i=1}^{n} w(\mu_i)\, x_i}{\sum_{i=1}^{n} w(\mu_i)}$$

onde c, a constante de escalonamento, é escolhida adequadamente. Por exemplo, MOSTELLER e TUKEY (1977), para distribuições aproximadamente normais, tomando-se S como sendo:

$$S = \frac{1}{2} \times \text{amplitude interquartís}$$

tem-se que:

$$S \cong \frac{4}{3}\sigma$$

e então, $c = 6$, fará com que:

$$cS \cong 4\sigma$$

isto é, dar-se-á peso zero à desvios $(x_i - \bar{x})$ maiores que 4σ.

Como $\bar{x}$ depende de μ e μ depende de $\bar{x}$ há a necessidade de se proceder a um cálculo iterativo para obter a solução. Para se iniciar as iterações há a necessidade de se conhecer uma estimativa inicial de $\bar{x}$ e S (com S não há iteração, apesar de mudar em cada passo). Estas estimativas iniciais podem ser:

$\tilde{x}^0$ = mediana $\{x_1, x_2, \ldots, x_n\}$

$S^0 = \frac{1}{2}$ x amplitude interquartis de $\{x_1, x_2, \ldots, x_n\}$

MOSTELLER e TUKEY (1977), no cap. 10, fazem algumas considerações sobre Resistência e Robustez em Eficiência, comparando a média aritmética, a mediana e o estimador biponderado. Apresentam o seguinte quadro:

Quadro 3.1.1 - Resistência e Robustez em Eficiência de alguns estimadores de locação

Estimador	Tamanho da amostra	Resistente	Eficiência com dados gausianos	Robustez em Eficiência
Média Aritmética	Pequeno	Não	100%	Pobre
	Grande	Não	100%	Muito pobre
Mediana	Pequeno	Sim	Alta	Alta
	Grande	Sim	62%	Moderada
Estimador Biponderado	Pequeno	Razoável	Altíssima	Altíssima
	Grande	Sim	>90%	Alta

Concluem que, exceto para amostras muito pequenas o estimador biponderado tem todas as propriedades desejáveis. Para amostras pequenas, de 3, 4 ou 5 elementos, faz-se melhor ao se escolher a mediana. Na prática, então, tende-se a usar:

- a mediana em exploração, e em outras circunstâncias onde é suficiente uma eficiência moderada numa grande variedade de situações.

- o estimador biponderado, ou algum outro similar, quando é necessário um desempenho muito bom.

- a média somente após um estudo cuidadoso. Quando a tradição ou o significado no campo da aplicação requisitar, quando o custo computacional for exorbitante, quando for necessário linearidade ou quando os dados possuirem caudas curtas e nenhum valor discrepante.

GROSS (1976) inclui em seu estudo sobre Intervalos de Confiança para estimadores robustos o estimador biponderado, chamando-o de estimador biquadrado e identificando-o com as siglas BS74, BS82 e BS90. Define o seguinte:

$$\mu_i = (x_i - \text{mediana})/(c.\text{MDA})^*$$

e também

$$\left.\begin{aligned} \Psi(\mu) &= \mu(1-\mu^2)^2 \\ \Psi'(\mu) &= (1-\mu^2)(1-5\mu^2) \end{aligned}\right\}, \quad \mu^2 \leq 1$$

$$\Psi(\mu) = \Psi'(\mu) = 0, \quad \mu^2 > 1$$

Então a solução é:

---

* MDA - medida robusta de dispersão, a ser apresentada na secção III.2.2.

$$T = \text{mediana} + (c.MDA)\,\frac{\Sigma\Psi(\mu)}{\Sigma\Psi'(\mu)}$$

com variância

$$S^2 = nc^2MDA^2\,\frac{\Sigma\Psi^2(\mu)}{[\Sigma\Psi'(\mu)]^2}$$

onde c = 7,4; 8,2 e 9,0, caracterizando BS74, BS82 e BS90, respectivamente.

BEATON e TUKEY (1974) utilizam c=2, que significa dar peso zero à observações distantes mais de $2,7\sigma$ da mediana (no caso gausiano). Afirmam também que talvez tivessem feito melhor se escolhessem c como sendo 3,4 ou mesmo 5. TUKEY (1975) diz que a constante c deve ser um valor entre 4 e 9.

MOSTELLER e TUKEY (1977) apresentam um intervalo de confiança para os estimadores biponderados baseado no trabalho de GROSS (1976). Seja t* o $\alpha/2$ quantil de uma distribuição t-Student com $0,7(n-1)$ graus de liberdade. Então o intervalo de confiança de $100(1-\alpha)\%$ para os estimadores biponderados, $\hat{x}$, é:

$$(\hat{x}-t^* S_{bi};\ \hat{x} + t^* S_{bi})$$

onde

$$S_{bi}^2 = \frac{\Sigma'(x_i-x')^2(1-\mu_i^2)^4}{\left[\Sigma'(1-\mu_i^2)(1-5\mu_i^2)\right]\left[-1+\Sigma'(1-\mu_i^2)(1-5\mu_i^2)\right]}$$

é a variância assintótica para os estimadores biponderados, Σ' indica a

soma para todo i; i=1,2,...,n tal que $\mu_i^2 \leq 1$ e x' é a mediana dos x_i's. Afirmam que tem um bom desempenho para $n \geq 8$.

No estudo que faz sobre intervalos de confiança, GROSS (1976) afirma que os estimadores biponderados são muito bons, especialmente o BS82 e o BS90, com maiores elogios ao primeiro.

Exemplo 3.1.1 - O seguinte exemplo numérico ilustra o desempenho do estimador biponderado.

Os dados a serem utilizados foram apresentados por MOSTELLER e TUKEY (1977), no capítulo 14. Tem-se o seguinte: Suponha que se tem uma amostra de 10 valores, de uma certa variável em estudo. Os valores, já ordenados, são os seguintes:

10, 7, 3, 3, 3, -2, -5, -5, -6, -8

Nota-se que a soma é, convenientemente, igual a zero. Vai-se comparar o comportamento da média, da mediana e do estimador biponderado ao se acrescentar uma nova observação aos dados originais. Esta décima-primeira observação, x, poderá assumir qualquer valor real. Este exemplo, apesar de ser numérico, servirá também para ilustrar o formato geral da curva de influência para o estimador biponderado. Também se apresentará as curvas de influência para a média e para a mediana. De acordo com o descrito na secção II.2, ter-se-á curvas de influência estilizadas, por terem sido obtidas de um conjunto finito de dados.

A constante de escalonamento, c, foi escolhida, por MOSTELLER e TUKEY (1977) como sendo igual a 6. A medida robusta de escala,

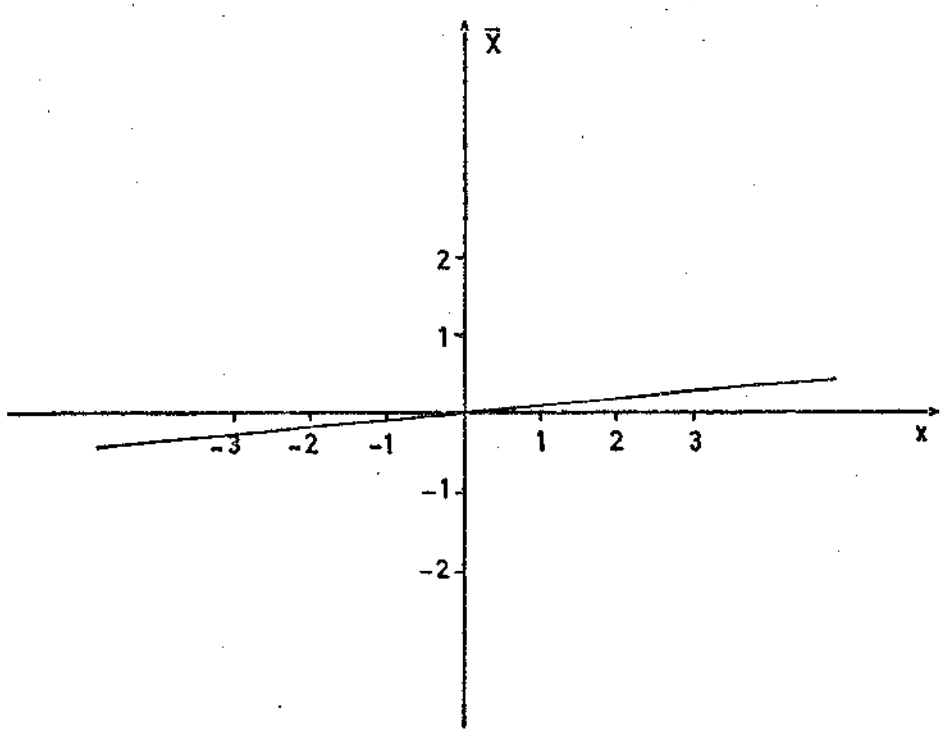
S, será a metade da distância entre x e as observações com valor igual a 3. Esta medida de escala não aparecerá na secção III.2. É uma aproximação da medida de escala a ser apresentada na secção III.2.1.

a média, $\overline{X}$ - como a soma das dez observações originais é igual a zero, a soma das onze observações será igual ao valor da décima-primeira observação; isto é, será igual a x. Consequentemente:

$$\overline{x} = x/11$$

A figura 3.1.1.a apresenta uma reta com inclinação 1/11, passando pela origem, que mostra como $\overline{X}$ responde à mudanças em x.

Figura 3.1.1.a - Curva de Influência (estilizada) para a média $\overline{X}$.

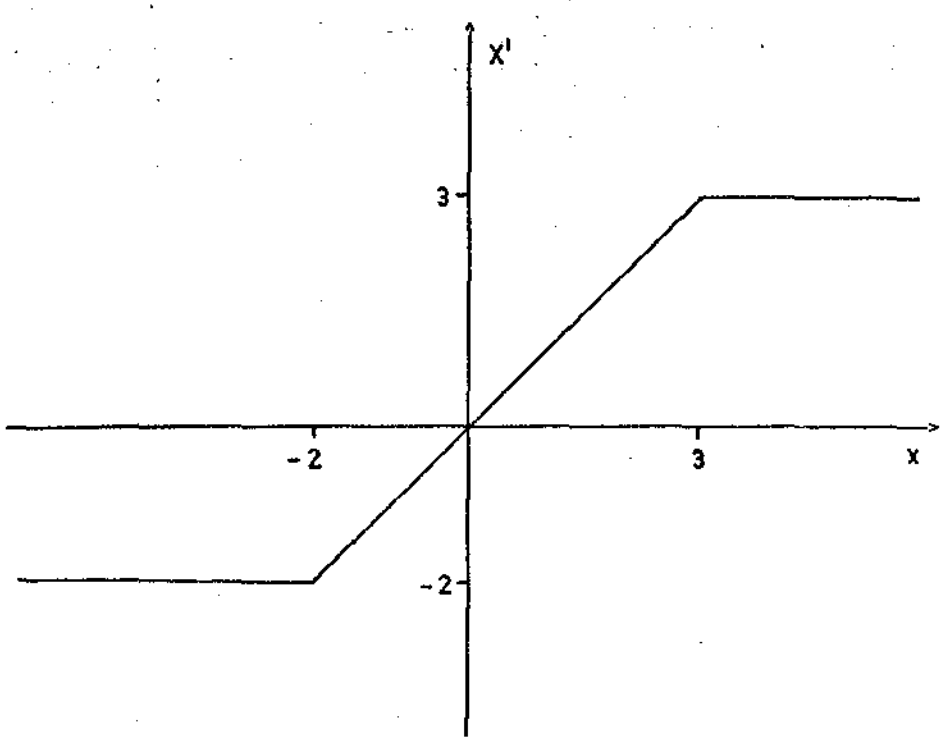


a mediana, X' - como o tamanho da amostra (incluindo x) é um número ímpar (11), a mediana será a observação do "meio" nos dados ordenados, neste caso, será a sexta observação. Então:

$$X' = \begin{cases} -2 & \text{se } x \leq -2 \\ x & \text{se } -2 < x < 3 \\ 3 & \text{se } 3 \leq x \end{cases}$$

A figura 3.1.1.b mostra o comportamento da mediana, X', de acordo com os valores que x assume.

Figura 3.1.1.b - Curva de Influência (estilizada) para a mediana X'.



o estimador biponderado, $\tilde{X}$ - tem-se que:

$$w(\mu_i) = \begin{cases} (1-\mu_i)^2 & \text{se } \mu_i^2 \leq 1 \quad ; \quad i=1,2,\ldots,11 \\ 0 & \text{se } \mu_i^2 > 1 \end{cases}$$

$$\mu_i = \frac{x_i - \hat{x}}{cS} \quad ; \quad i=1,2,\ldots,11$$

sendo c=6 e S como abaixo:

se $-\infty < x \leq -6$	$I = 3-(-6) = 9$	$S = 4,5$	
se $-6 < x < -5$	$I = 3-x$; $4 < S < 4,5$	$S = (3-x)/2$	
se $-5 \leq x \leq 3$	$I = 3-(-5) = 8$	$S = 4$	(3.1.1.1)
se $3 < x < 7$	$I = x-(-5)$; $4 < S < 6$	$S = (x+5)/2$	
se $7 \leq x < \infty$	$I = 7-(-5)$	$S = 6$	

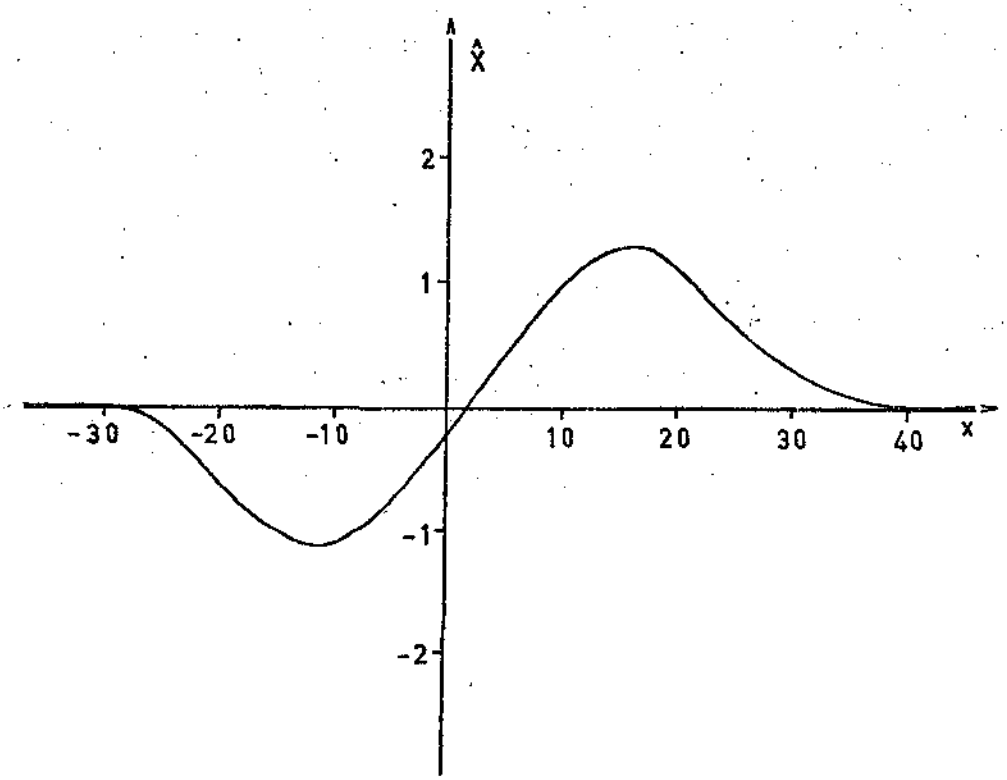
Então

$$\tilde{x} = \frac{\sum_{i=1}^{11} w(\mu_i)x_i}{\sum_{i=1}^{11} w(\mu_i)}$$

MOSTELLER e TUKEY (1977) tomam uma iteração, a partir da mediana e S como acima e constroem a curva de influência (estilizada)que é apresentada na figura 3.1.1.c. Nota-se um comportamento bastante desejável. Quando x assume valores muito negativos ou muito positivos, sua influência no cálculo de $\tilde{x}$ tende a zero e se obtém uma estimativa baseada quase que totalmente nas outras dez observações. Aproximadamente,

pode-se dizer que x tem influência zero quando $x < -27$ ou $x > 36$ e é claro, quando x tende a zero.

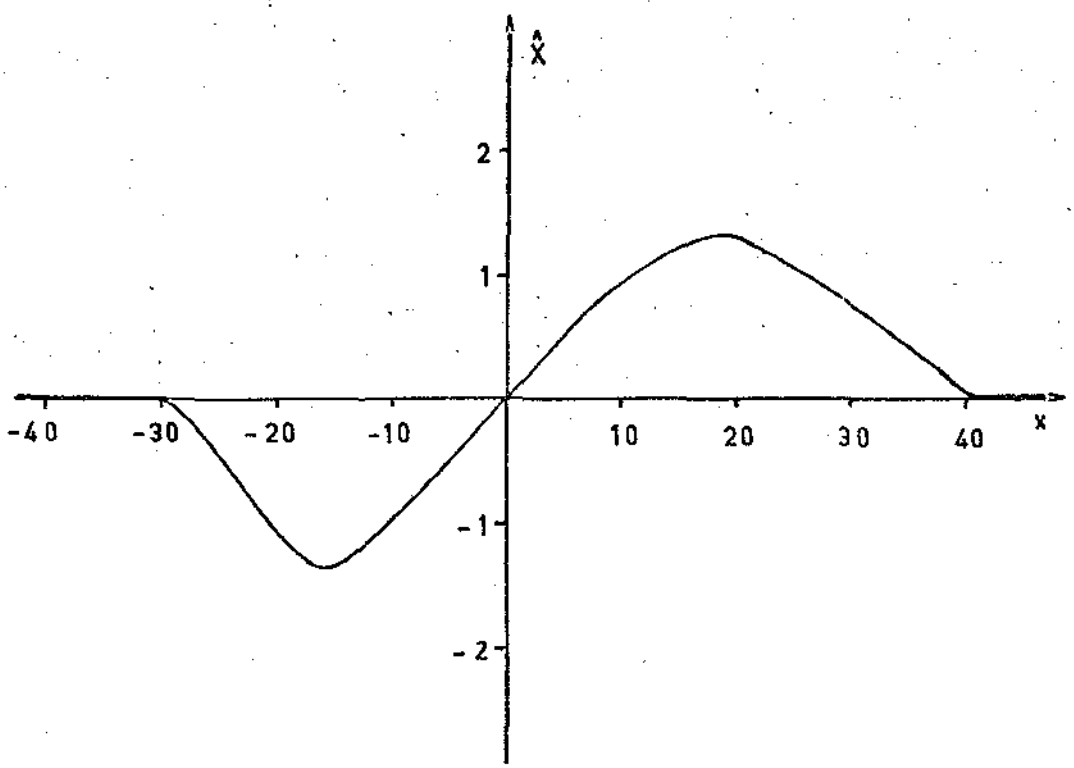
Figura 3.1.1.c - Curva de Influência (estilizada) para o estimador biponderado, $\hat{X}$, (apresentada por MOSTELLER e TUKEY (1977), cap. 14, S é dado por(3.1.1.1)).



Para mostrar a influência da medida robusta de escala, S, utilizada pelo estimador biponderado, MOSTELLER e TUKEY (1977) apresentam a curva de influência (estilizada) para este estimador, utilizando S=MDA, a ser apresentado na secção III.2.2, em vez de S dado por (3.1.1.1). Esta curva será apresentada na figura 3.1.1.d. Nota-se que não há, praticamente, nenhuma diferença entre as curvas apresentadas

pelas figuras 3.1.1.c e 3.1.1.d, evidenciando o fato de que uma boa escolha de S não afeta o desempenho do estimador biponderado. Na secção III.2 serão apresentadas três possíveis (boas) maneiras de se obter S.

Figura 3.1.1.d - Curva de Influência (estilizada) para o estimador biponderado, $\hat{X}$ (apresentada por MOSTELLER e TUKEY (1977), cap. 14, S=MDA)



O quadro 3.1.1.e apresenta os valores para a média, mediana e para o estimador biponderado, que se obtém ao variar x. As estimativas dadas pelo estimador biponderado são obtidas pela primeira iteração a partir da mediana, S dado por (3.1.1.1) e c=6.

Quadro 3.1.1.e - Valores de $\overline{X}$, X' e $\tilde{X}$ para vários valores de x

x	$\overline{X}$	X'	$\tilde{X}$
-35.000	-3.1818182	-2.000	-0.4821151
-30.000	-2.7272727	-2.000	-0.4821151
-25.000	-2.2727273	-2.000	-0.6844982
-20.000	-1.8181818	-2.000	-1.1262822
-15.000	-1.3636364	-2.000	-1.3714335
-10.000	-0.9090909	-2.000	-1.2841345
- 5.000	-0.4545455	-2.000	-1.0439496
0.000	0.0000000	0.000	-0.1319272
5.000	0.4545455	3.000	0.8089473
10.000	0.9090909	3.000	1.1201979
15.000	1.3636364	3.000	1.3904087
20.000	1.8181818	3.000	1.4387364
25.000	2.2727273	3.000	1.2383007
30.000	2.7272727	3.000	0.8385566
35.000	3.1818182	3.000	0.4055475
40.000	3.6363636	3.000	0.2429425
45.000	4.0909091	3.000	0.2429425

Como se pode notar no quadro acima, quando x tende a valores muito longe de zero, tanto positivos como negativos, $\tilde{x}$ é cada vez menos influenciado por ele. Para valores de x entre -15 e 15, $\tilde{x}$ e $\overline{x}$ tem um comportamento parecido, e assumem valores relativamente próximos. Os

resultados obtidos neste exemplo vem ratificar o bom desempenho do estimador biponderado, principalmente no que diz respeito à robustez. Também mostram a característica deste estimador no que se refere ao controle dos valores aberrantes. Quando $|x|$ cresce muito sua influência vai decrescendo até se tornar nula.

III.1.2 - O estimador Passoponderado (stepweight)*

Este estimador está sendo incluido neste estudo com o intuito de oferecer uma alternativa que demande menor quantidade de cálculos. O estimador passoponderado é bastante semelhante, em comportamento, ao estimador biponderado, necessitando uma quantidade de cálculos bastante inferior a este. Na impossibilidade de acesso a um computador, ou mesmo à máquinas de calcular este estimador poderá vir a ser de grande utilidade.

A forma deste w-estimador é a seguinte:

$$w(\mu_i)=\begin{cases} k_1 \;;\; |\mu_i| \leq a_1 \\ k_2 \;;\; a_1 < |\mu_i| \leq a_2 \\ \vdots \\ k_m \;;\; a_{m-1} < |\mu_i| \leq a_m \\ 0 \;;\; a_m < |\mu_i| \end{cases} \qquad i=1,2,\ldots,n \quad (3.1.2.1)$$

---

* stepweight - abreviação para ponderação passo a passo ou ponderação em saltos

onde $k_1 > k_2 > \ldots > k_m > 0$ e $0 < a_1 < a_2 < \ldots < a_m$ são constantes conhecidas. E:

$$\mu_i = \frac{x_i - \tilde{x}}{cS} \quad ; \quad i=1,2,\ldots,n$$

$$\tilde{x} = \frac{\sum_{i=1}^{n} w(\mu_i)x_i}{\sum_{i=1}^{n} w(\mu_i)}$$

Como no caso do estimador biponderado a solução é obtida iterativamente, partindo de valores iniciais para $\tilde{x}$ e S como por exemplo a mediana para $\tilde{x}$ e metade da amplitude interquartís para S. A constante c deve ser escolhida adequadamente, como para o estimador biponderado, mas também se deve levar em conta uma escolha como c=5, para facilitar os cálculos, e, talvez usar o divisor da amplitude interquartís, caso se opte por esta medida de dispersão, como sendo um número de fácil divisão.

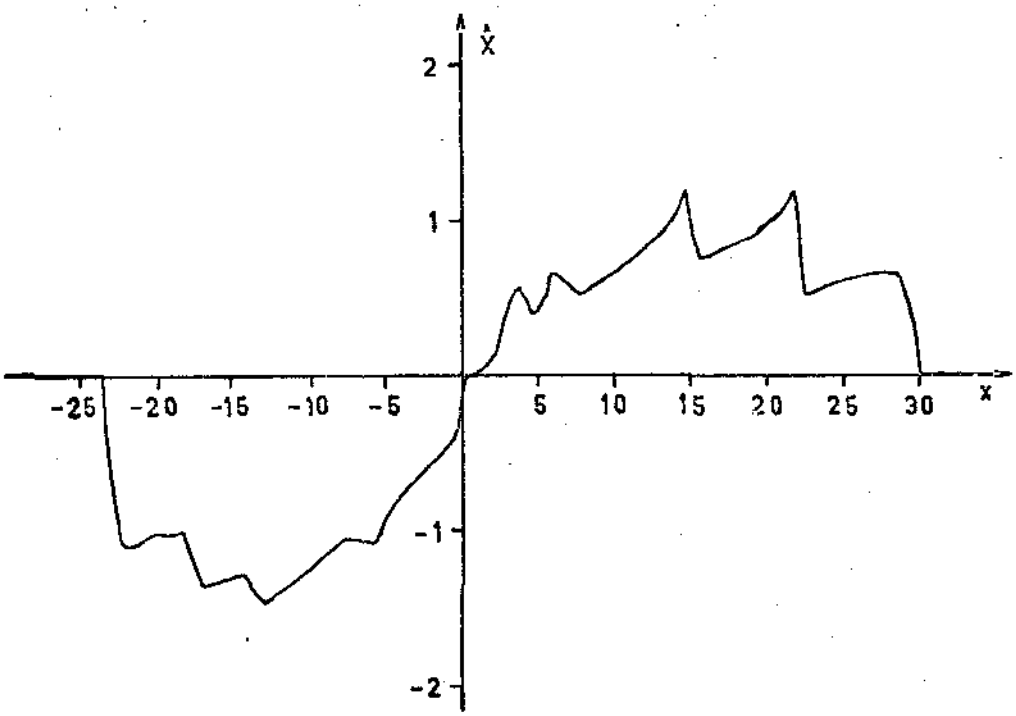
MOSTELLER e TUKEY (1977) sugerem o uso do estimador passo ponderado com a seguinte ponderação:

$$w(\mu_i) = \begin{cases} 4 \; ; \; |\mu_i| \leq 0,2 \\ 3 \; ; \; 0,2 < |\mu_i| \leq 0,4 \\ 2 \; ; \; 0,4 < |\mu_i| \leq 0,6 \\ 1 \; ; \; 0,6 < |\mu_i| \leq 0,8 \\ 0 \; ; \; 0,8 < |\mu_i| \end{cases} \quad ; \quad i=1,2,\ldots,n \qquad (3.1.2.2)$$

Exemplo 3.1.2 - Como na secção anterior, se apresentará agora um exemplo numérico ilustrando o desempenho do estimador pas-

soponderado. Serão utilizados os mesmos dados do exemplo 3.1.1, e se associará os pesos de acordo com(3.1.2.2) A medida robusta de escala, S, será dada por (3.1.1.1) e a constante de escalonamento, c, será igual a 6. As curvas de influência (estilizadas) para a média, $\bar{X}$ e para a mediana X', são as mesmas do exemplo 3.1.1 e estão nas figuras 3.1.1.a e 3.1.1.b, respectivamente. A figura 3.1.2.a, mostra a curva de influência estilizada para o estimador passoponderado.

Figura 3.1.2.a - Curva de Influência (estilizada), para o estimador passoponderado, $\tilde{X}$



Como no exemplo anterior, apesar do exemplo ser numérico, a curva da figura 3.1.2.a. representa bem a forma geral da curva de influência para o estimador passoponderado. Pode-se notar que quando a décima-primeira observação, x, assume valores muito negativos ou muito positivos, sua influência cai a zero, e, é claro, quando x tende a zero também. Pode-se dizer, examinando as figuras 3.1.1.c (também 3.1.1.d) e 3.1.2.a, que os estimadores biponderado e passoponderado tem um comportamento parecido. Então o estimador passoponderado pode ser utilizado como um substituto do estimador biponderado, caso não se disponha de um computador ou mesmo de uma máquina de calcular. Neste caso, como já se afirmou anteriormente, convém também tomar c e S de modo a facilitar ainda mais os cálculos.

O quadro 3.1.2.b apresenta os valores para a média, mediana e para o estimador passoponderado, que se obtém ao variar x (a décima primeira observação). As estimativas são obtidas pela primeira iteração, tomada a partir da mediana, S é dado por 3.1.1.1 e c=6.

Quadro 3.1.2.b - Valores de $\overline{X}$, X' e $\hat{X}$ para vários valores de x

x	$\overline{X}$	X'	$\hat{X}$
-35.000	-3.1818182	-2.000	-0.5277778
-30.000	-2.7272727	-2.000	-0.5277778
-25.000	-2.2727273	-2.000	-0.5277778
-20.000	-1.8181818	-2.000	-1.0540541
-15.000	-1.3636364	-2.000	-1.2894737
-10.000	-0.9090909	-2.000	-1.2564103
- 5.000	-0.4545455	-2.000	-1.2972973
0.000	0.0000000	0.000	-0.0810811
5.000	0.4545455	3.000	0.8717949
10.000	0.9090909	3.000	1.6000000
15.000	1.3636364	3.000	1.7692308
20.000	1.8181818	3.000	1.6842105
25.000	2.2727273	3.000	1.3243243
30.000	2.7272727	3.000	1.4594595
35.000	3.1818182	3.000	0.6666667
40.000	3.6363636	3.000	0.6666667
45.000	4.0909091	3.000	0.6666667

III.2 - Alguns estimadores robustos de escala

Nesta secção serão apresentados alguns estimadores que poderão ser utilizados para a obtenção da medida de escala S, citada na

secção anterior. Dar-se-á mais ênfase ao primeiro, que será o utilizado nos exemplos, por ser o de mais fácil obtenção.

III.2.1 - Amplitude Interquartís

Na realidade não se vai usar pura e simplesmente a amplitude interquartís como uma medida de dispersão. O que se usará será esta amplitude dividida por um número, por exemplo pelo seu valor esperado no caso de se trabalhar sob a pressuposição de normalidade (1,35), como é sugerido em ANDREWS e outros (1972). Obviamente esta não é a única escolha, apesar de parecer ser a mais razoável. Como já se citou, MOSTELLER e TUKEY (1977), em um exemplo com o estimador biponderado utilizam o número 2 como divisor, apesar de desejarem facilidades computacionais.

ANDREWS e outros (1972) apresentam uma maneira pela qual se pode calcular a amplitude interquartís, em um conjunto contendo n observações. É obtida através da seguinte diferença:

$$A_I = h_2 - h_1$$

onde h_1 e h_2 são como se segue:

$$h_1 = \begin{cases} x_{\left(\left[\frac{n}{4}\right]\right)} & \text{se n não é múltiplo de 4} \\ \frac{1}{2}\cdot\left(x_{\left(\frac{n}{4}\right)} + x_{\left(\frac{n}{4}+1\right)}\right) & \text{se n é múltiplo de 4} \end{cases}$$

$$h_2 = \begin{cases} x_{\left(n+1-\left[\frac{n+3}{4}\right]\right)} & \text{se n não é múltiplo de 4} \\ \frac{1}{2}\cdot\left(x_{\left(n+1-\frac{n}{4}\right)} + x_{\left(n-\frac{n}{4}\right)}\right) & \text{se n é múltiplo de 4} \end{cases}$$

onde $x_{(j)}$ é o j-ésimo menor dos x's e [a] é o maior inteiro contido em a, por exemplo [3,5] = 3. TUKEY (1977) dá o nome "hinge" a h_1 e h_2, DACHS (1978) traduz hinge como junta. Neste trabalho serão normalmente referenciado como:

h_1 = primeiro quartil

h_2 = terceiro quartil

ou como, juntas.

É interessante lembrar que o segundo quartil é a mediana.

Finalmente pode-se definir S, a medida robusta de escala, como:

$$S = \frac{AI}{k} = \frac{h_2 - h_1}{k}$$

onde k será usualmente igual a 1,35 que, como já foi dito, é o valor esperado da amplitude interquartís sob a suposição de normalidade $\left(N(0,1)\right)$,

como se pode observar de qualquer tabela da distribuição normal. Se X é uma variável aleatória com distribuição N(0,1), então:

$$P\ [X > k] = P\ [X < -k] = 0,25 \Rightarrow k \cong 0,675$$

e portanto o valor esperado da amplitude interquartis é:

$$AI = k-(-k) = 2k \cong 2x0,675 = 1,35$$

Portanto, usualmente, se usará:

$$S = \frac{AI}{1,35} = \frac{h_2-h_1}{1,35}$$

III.2.2 - Mediana dos Desvios Absolutos - MDA*

MOSTELLER e TUKEY (1977) apresentam esta medida robusta de dispersão, denominada MDA. É a seguinte:

$$MDA = \text{mediana } \{|x_i-x'|\}$$

onde x' é um estimador robusto de locação.

ANDREWS e outros (1972), ANDREWS (1974), dentre outros utilizam várias vezes x' como a mediana do conjunto de dados em estudo. Esta escolha é a mais comum. Sempre que houver referência à esta medida, estará se supondo x' como sendo a mediana. Caso se use outra medida de locação que não a mediana, se fará a observação correspondente.

---

* MAD na abreviação em inglês

HAMPEL (1974) afirma que este é o mais robusto estimador da dispersão. Analogamente ao desvio padrão e desvio médio, ele o denomina Desvio Mediano (median deviation).

III.2.3 - Uma alternativa

MOSTELLER e TUKEY (1977) apresentam outro estimador robusto de dispersão, sugerido por David Lax (1975) em: "An interim report of a Monte Carlo study of robust estimators of widht", Tecnical Report No. 93. Department of Statistcs, Princeton University. É uma medida de escala que demanda uma maior quantidade de cálculos que as duas anteriores. Seja

$$x' = \text{mediana}\left\{x_1, x_2, \ldots, x_n\right\}$$

$$\mu_i = \frac{x_i - x'}{9(MDA)} \qquad ; \qquad i=1,2,\ldots,n$$

Então Lax usa uma medida de escala derivada da variância assintótica dos w-estimadores biponderados;

$$\frac{n\Sigma'(x_i-x')^2(1-\mu_i^2)^4}{\left[\Sigma'(1-\mu_i^2)(1-5\mu_i^2)\right]^2} \qquad (3.2.3.1)$$

onde Σ' indica a soma para todo i tal que $\mu_i^2 \leq 1$; $i=1,2,\ldots,n$. Quando os μ_i^2 são pequenos, o denominador se reduz a aproximadamente $\sum_{i=1}^{n} 1 = n$ e a expressão (3.2.3.1) se reduz a

$$\frac{\Sigma\,(x_i-x')^2}{n}$$

que parece ser um estimador razoável para a variância.

Uma modificação que pode ser reduzida a

$$S^2=\frac{\Sigma\,(x_i-x')^2}{n-1}$$

e é tida como um pouco melhor é:

$$nS^2_{bi}=\frac{n\Sigma'(x_i-x')^2(1-\mu_i^2)^4}{\left[\Sigma'(1-\mu_i^2)(1-5\mu_i^2)\right]\left[-1+\Sigma'(1-\mu_i^2)(1-5\mu_i^2)\right]}$$

Note que $(1-\mu_i^2)^4 = w^2(\mu_i)$ da secção III.1.1. Se é usado $\bar{x}$, o estimador biponderado, em vez da mediana x', não há grande diferença nos resultados.

IV - REGRESSÃO ROBUSTA

Neste capítulo serão apresentados mecanismos de ajuste de regressão que resultarão em ajustes de mínimos quadrados ponderados, obtidos iterativamente. Serão apresentados os métodos de ajuste de regressão usando os estimadores biponderado e passoponderado, apresentados nas secções III.1.1 e III.1.2, respectivamente. Apenas para efeito de informação, se apresentará também um método, baseado no M-estimador seno, apresentado por ANDREWS (1974). Este método será apresentado no final deste capítulo, no apêndice. No capítulo V se apresentará um exemplo numérico onde serão comparados os resultados obtidos através do método baseado no estimador biponderado e do método baseado no M-estimador seno.

Os métodos a serem apresentados resultarão em estimativas mais resistentes/robustas à presença de valores discrepantes do que o método dos mínimos quadrados. Mesmo quando se estiver trabalhando com dados distribuidos normalmente se obterá resultados muito bons.

IV.1 - Notação

Irá se trabalhar usando notação matricial, como, por exemplo, é feito em SEARLE (1971), dentre outros.

Suponha que se tem uma matriz X (nxk), denominada matriz de delineamento ou matriz dos suportes. Seja:

$$X = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

onde x_{ji} é a i-ésima observação do suporte X_j; i=1,2,...,n e j=1,2,...,k. Cada suporte poderá se constituir de uma variável simples ou de uma função de uma ou mais variáveis simples, como já se afirmou na secção II.3. Geralmente se tomará $X_1 \equiv 1$. Isto resultará em uma equação de regressão que não passa pela origem.

Além da matriz X, suponha que se tem dois vetores, Y e B. O primeiro um vetor n-dimensional, denominado vetor das observações ou vetor das respostas:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

O segundo é um vetor k-dimensional, denominado vetor dos parâmetros:

$$B = \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{pmatrix}$$

Nos casos em que $X_1 \equiv 1$, b_1 será o termo constante na equação de regressão obtida.

Suponha que se deseje ajustar:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \quad ; \quad i=1,2,\ldots,n$$

que escrito na forma matricial é:

$$Y = XB \qquad (4.1.1)$$

Seja $\hat{B}$ a estimativa obtida:

$$\hat{B} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ . \\ . \\ . \\ \hat{b}_k \end{pmatrix}$$

O ajuste obtido é então:

$$\hat{Y} = X\hat{B}$$

A estimativa do vetor dos parâmetros, $\hat{B}$, será obtida através de emparelhadores, como descrito na secção II.4. Seja P uma matriz

(nxn) que será denominada matriz de ponderação. A priori esta matriz poderá ser qualquer tipo de matriz. Entretanto no caso particular dos métodos em estudo, esta matriz de ponderação será sempre uma matriz diagonal:

$$P = \begin{pmatrix} p_{11} & 0 & \cdots & 0 \\ 0 & p_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & p_{nn} \end{pmatrix}$$

De modo análogo a secção II.4, tem-se que X'P é um emparelhador pois exigir que:

$$X'PY = X'P\hat{Y}$$

é o mesmo que:

$$X'PY = X'PX\hat{B}$$

e se (X'PX) for de posto completo obtém-se a solução:

$$\hat{B} = (X'PX)^{-1} X'PY \qquad (4.1.2)$$

Passar-se-á agora à apresentação dos métodos de regressão que resultarão em estimativas mais resistentes/robustas à presença de valores discrepantes. Como será visto, o que mudará de um para outro método será apenas a matriz P.

No caso de P≡I(I é a matriz identidade de ordem n), obter-se-á o ajuste de mínimos quadrados como usualmente é feito. Caso $P \equiv V^{-1}$, onde V é a matriz de variâncias e covariâncias dos y's, obter-se-á o ajuste de mínimos quadrados ponderados, como já se viu na secção II.4.

IV.2 - Regressão Robusta usando o estimador Biponderado

Este método é apresentado em BEATON e TUKEY (1974) e MOSTELLER e TUKEY (1977). Não é muito influenciado por um pequeno número de valores aberrantes e também apresenta um razoável desempenho com dados gausianos.

A idéia é associar pesos maiores a observações com resíduos pequenos e ir paulatinamente diminuindo a ponderação a medida que os resíduos vão crescendo, em valor absoluto, chegando mesmo a se dar peso zero à observações com grandes resíduos (maiores que cS, como já foi visto na secção III.1.1). Deste modo tem-se uma maneira de "controlar" os valores discrepantes. O termo "controlar" se refere a dar maior importância aos dados da porção central e menor importância aos dados mais extremos.

Além disto, com este expediente se consegue "detectar" os valores discrepantes. Basta observar quais as observações que apresentam os maiores resíduos (em valor absoluto), facilitando deste modo um estudo mais acurado sobre estes valores, se desejado.

Seja então:

$$p(\mu_i) = \begin{cases} (1-\mu_i^2)^2 & , \quad \mu_i^2 \leq 1 \\ 0 & , \quad \mu_i^2 > 1 \end{cases} \qquad ; \quad i=1,2,\ldots,n$$

e

$$\mu_i = \frac{y_i - \bar{y}_i}{cS} \qquad ; \quad i=1,2,\ldots,n$$

Pode-se agora explicitar a matriz de ponderação, a matriz diagonal P, de dimensão (nxn), com:

$$P(i,j) = \begin{cases} p(\mu_i) & ; \quad i=j \\ 0 & ; \quad i \neq j \end{cases} \qquad ; \quad i,j=1,2,\ldots,n$$

onde P(i,j) representa o i-ésimo elemento da j-ésima coluna. A constante de escalonamento, c e a medida robusta de escala, S, são escolhidas adequadamente, de acordo com o mencionado nas secções III.1.1 e III.1.2 respectivamente. No desenvolvimento dos exemplos do capítulo V e no programa para computadores será utilizado:

$$S = \frac{\text{amplitude interquartis}}{1,35} = \frac{h_2 - h_1}{1,35}$$

com h_1 e h_2 sendo as juntas (DACHS (1978)) como definidas na secção III.2.1 e

$$c=4$$

Como no caso do estimador biponderado, para parâmetros de locação, a solução é obtida iterativamente. Deve-se portanto partir de

uma estimativa inicial dos parâmetros, $\hat{B}_0$. Esta estimativa inicial pode, por exemplo, ser a obtida por mínimos quadrados. DRAPER e SMITH (1966), no cap. 10 comentam sobre os problemas de uma má escolha de $\hat{B}_0$ (no caso de Mínimos Quadrados não lineares). ANDREWS (1974) sugere que $\hat{B}_0$ obtido através do método de regressão pelas medianas poderá ser uma melhor escolha que o obtido por mínimos quadrados, mas não se tratará disto neste trabalho. No caso de uma regressão linear simples, y=a+bx, pode-se mesmo utilizar $\hat{a}$ e $\hat{b}$ obtidos "visualmente" de um gráfico de y versus x. O programa para computadores admite duas opções:

1 - $\hat{B}_0$ obtido pelo método dos mínimos quadrados

2 - $\hat{B}_0$ é lido de cartões (de algum modo já se calculou $\hat{B}_0$ anteriormente e se deseja economizar tempo de máquina)

Suponha então que já se conheça uma estimativa inicial dos parâmetros, $\hat{B}_0$. A partir dela pode-se calcular P_0 e então:

$$\hat{B}_1 = (X'P_0X)^{-1} X'P_0Y$$

se $(X'P_0X)$ for de posto completo. Prosseguindo-se desta maneira obtém-se, de um modo geral:

$$\hat{B}_{g+1} = (X'P_gX)^{-1} X'P_gY$$

se $(X'P_iX)$; i=1,2,...,g forem de posto completo. Pode-se reescrever a expressão acima como:

$$\hat{B}_{g+1} = \hat{B}_g + (X'P_gX)^{-1} X'P_g(Y-X\hat{B}_g)$$

e se é definido:

$$R_g = Y - X\tilde{B}_g$$

como sendo o vetor (n-dimensional) dos resíduos, na g-ésima iteração, tem-se finalmente:

$$\tilde{B}_{g+1} = \tilde{B}_g + (X'P_gX)^{-1} X'P_gR_g$$

onde

$$P_g(i,j) = \begin{cases} p\left(\dfrac{y_i - \tilde{y}_i(g)}{cS_g}\right) & ; \quad i=j \\ & \qquad i,j=1,2,\ldots,n \\ 0 & ; \quad i\neq j \end{cases}$$

$$S_g = \frac{\text{amplitude interquartis dos elementos do vetor } R_g}{1,35}$$

e

$\tilde{y}_i(g)$ é a estimativa de y_i; i=1,2,...,n obtida na g-ésima iteração.

Como em todo processo iterativo, deve-se definir um critério de parada. Pode-se tanto parar o processo após um número máximo pré-fixado de passos como também quando se julga haver obtido uma precisão desejada. Na secção IV.4 será discutido este assunto.

No programa para computadores, a subrotina BIPON, constrói a matriz P como definida anteriormente. Na secção VII.3 há maiores detalhes quanto a esta subrotina.

IV.3 - Regressão Robusta usando o estimador Passoponderado

MOSTELLER e TUKEY (1977) apresentam este método, sendo que a matriz de ponderação P é construida com base em (3.1.2.2).

As vantagens deste método, assim como o seu comportamento semelhante ao biponderado já foram discutidas na secção III.1.2. Além do que já foi comentado, pode-se usar a regressão passoponderada como um estudo preliminar do que se vai obter com o estimador biponderado, ou outro estimador semelhante (como o M-estimador seno, a ser apresentado no apêndice). Este método pode mesmo vir a dar uma idéia da boa ou má escolha da estimativa inicial dos parâmetros, $\hat{B}_0$, ou mesmo da quantidade de passos necessária para se obter boa precisão.

Pode-se pensar num método geral para atribuir a ponderação passoponderada e construir a matriz, diagonal, de ponderação, P. Isto será feito abaixo, e também é o método usado pela subrotina STPON, da secção VII.3. Consiste no seguinte:

- suponha que se quer atribuir pesos de acordo com (3.1.2.1); isto é,

$$w(\mu_i) = \begin{cases} k_1 & ; \ |\mu_i| \leq a_1 \\ k_2 & ; \ a_1 < |\mu_i| \leq a_2 \\ k_3 & ; \ a_2 < |\mu_i| \leq a_3 \\ \vdots & \quad \vdots \\ k_m & ; \ a_{m-1} < |\mu_i| \leq a_m \\ 0 & ; \ a_m < |\mu_i| \end{cases} \quad ; \ i=1,2,\ldots,n$$

- faça $a_1=1/(m+1)$ e $a_j=j/(m+1) = j \cdot a_1$; $j=2,3,\ldots,m$
- escolha um número k_1 e faça $k_j=k_1 \cdot (m-j+1)/m$; $j=2,3,\ldots,m$ e $k_{m+1}=0$

Portanto escolhendo apenas $\underline{m}$ e $\underline{k_1}$, pode-se construir a função que atribue os pesos e então construir a matriz P;

$$P(i,j) = \begin{cases} p(\mu_i) & ; \ i=j \\ 0 & ; \ i \neq j \end{cases} \quad ; \ i,j=1,2,\ldots,n$$

e

$$\mu_i = \frac{y_i - \hat{y}_i}{cS} \quad ; \ i=1,2,\ldots,n$$

Por exemplo ao se escolher $k_1=4$ e $m=4$ obtêm-se

$$\begin{cases} a_1 = 1/5 = 0,2 \\ a_2 = 2.a_1 = 0,4 \\ a_3 = 3.a_1 = 0,6 \\ a_4 = 4.a_1 = 0,8 \end{cases}$$

e

$$\begin{cases} k_1 = 4 \\ k_2 = 4.(4-2+1)/4 = 4.3/4 = 3 \\ k_3 = 4.(4-3+1)/4 = 4.2/4 = 2 \\ k_4 = 4.(4-4+1)/4 = 4.1/4 = 1 \\ k_5 = 0 \end{cases}$$

que é o mesmo que (3.1.2.2).

Obviamente quanto maior for m, maior será a semelhança entre os resultados obtidos pelo estimador biponderado e o passoponderado. Mas, como a intenção deste método é diminuir a quantidade de cálculos, não convém escolher um valor muito grande para m.

De modo análogo à secção anterior, a partir de um $\bar{B}_0$, obtém-se P_0 e S_0. Então

$$\bar{B}_1 = \bar{B}_0 + (X'P_0X)^{-1}X'P_0R_0$$

se $(X'P_0X)$ for de posto completo. Continuando-se com o processo:

$$\bar{B}_{g+1} = \bar{B}_g + (X'P_gX)^{-1}X'P_gR_g$$

se $(X'P_iX)$ for de posto completo; i=1,2,...,g. Prossegue-se com as iterações até obter a precisão desejada ou até completar um número máximo, pré-fixado, de passos, como será visto na secção IV.4.

IV.4 - Critério de Parada

Nesta secção se apresentará uma pequena discussão sobre a ocorrência (ou não) da convergência das estimativas para os parâmetros da regressão múltipla, ponderada, obtida através dos métodos descritos nas secções IV.2 e IV.3.

Apresenta-se também alguns critérios para a verificação da convergência, e, finalmente, o critério adotado e implementado no programa para computador.

IV.4.1 - É certa a convergência?

Vai-se discutir critérios para a verificação da convergência dos métodos de ajuste propostos sem se provar formalmente a ocorrência desta convergência. No método proposto e utilizado por ANDREWS (1974) (a ser apresentado como um apêndice a este capítulo) a convergência é certa, se o valor inicial estiver suficientemente próximo da solução. Como diz Andrews, a solução só depende de um método iterativo para se achar um ponto de máximo para a função cosseno, neste caso particular. Maiores detalhes podem ser obtidos na citada referência.

Com base na convergência acima citada e também na semelhança dos resultados que se obtêm com o estimador biponderado, com o passoponderado e com o M-estimador seno, não há, aparentemente, porque desconfiar da não-convergência dos métodos apresentados nas secções IV.2 e IV.3. Além disto há a evidência apresentada nos exemplos numéricos.Por

mais fraca que seja esta evidência, ela vem ratificar a "certeza" da convergência*.

A prova formal desta convergência não está no propósito deste trabalho. Poderá vir a ser objeto de estudos futuros. Por ora adotar-se-á como certa a convergência.

IV.4.2 - Alguns critérios para a determinação da convergência

Deseja-se fazer um ajuste da forma descrita em (4.1.1); isto é,

$$Y = XB$$

que pode ser escrito alternativamente como:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \quad ; \quad i=1,2,\ldots,n$$

que é uma regressão múltipla nos suportes $X_1, X_2, \ldots, X_k$. Suponha também que o ajuste vai ser obtido através de um método iterativo, como por exemplo, um dos dois descritos anteriormente.

AFIFI e AZEN (1974) afirmam que a maioria dos programas (e "pacotes") utilizam o seguinte critério (em Min. Quad. não lineares):

$$\left| \frac{b_i^{(j)} - b_i^{(j-1)}}{b_i^{(j-1)}} \right| < \delta \quad ; \quad i=1,2,\ldots,k \qquad (4.4.2.1)$$

onde $b_i^{(j)}$; i=1,2,...,k e j=1,2,... é a estimativa do parâmetro b_i na j-ésima iteração e δ é um número arbitrário (tomado tão pequeno quanto

---

* A prova da convergência é um problema em aberto, provavelmente bastante complexo.

se queira), por exemplo, $\delta = 10^{-6}$, como é sugerido em DRAPER e SMITH (1966). Além deste critério, outro que é muito utilizado (segundo AFIFI e AZEN (1974)) é verificar quando a soma de quadrado dos resíduos converge; isto é, verificar quando

$$\left| \text{S.Q.Res.}^{(j)} - \text{S.Q.Res.}^{(j-1)} \right| < \delta_1$$

onde δ_1 é similar a δ anteriormente citado.

Maiores detalhes podem ser obtidos em DRAPER e SMITH(1966) ou em RALSTON e WILF (1960).

IV.4.3 - Critério adotado

Este critério nada mais é que uma pequena modificação do critério (4.4.2.1) apresentado na secção anterior. Lá, quando $\left|b_i^{(j-1)}\right|$, para pelo menos um i; i=1,2,...,k, no denominador, é muito pequeno, ao se dividir $\left|b_i^{(j)}-b_i^{(j-1)}\right|$ por este denominador pode ocorrer que:

$$\left| \frac{b_i^{(j)}-b_i^{(j-1)}}{b_i^{(j-1)}} \right| >> \left| b_i^{(j)}-b_i^{(j-1)} \right| > \delta$$

para um particular índice i, pelo menos. Neste caso pode inclusive já ter ocorrido a convergência, ou então uma boa precisão, e em vez de parar, o processo iria continuar com as iterações. O critério adotado corrige este problema. Foi sugerido por ZAGO (1978) e é o critério que será utilizado no programa para computadores. Consiste do seguinte:

$$\left.\begin{array}{l} 1 \text{ - se } \left|b_i^{(j-1)}\right| < \varepsilon, \text{ então o teste é: } \left|b_i^{(j)} - b_i^{(j-1)}\right| < \delta \\ 2 \text{ - se } \left|b_i^{(j-1)}\right| \geq \varepsilon, \text{ então o teste é: } \left|\dfrac{b_i^{(j)} - b_i^{(j-1)}}{b_i^{(j-1)}}\right| < \delta \end{array}\right\} \; ; \; i=1,2,\ldots,k \qquad (4.4.3.1)$$

onde δ é o mesmo descrito anteriormente e ε também é um número arbritário (tão pequeno quanto se queira, e além disso $\varepsilon > \delta$ (possivelmente $\varepsilon >> \delta$)).

Por exemplo poder-se-ía utilizar $\varepsilon = 10^{-2}$ e $\delta = 10^{-6}$. Caso um (ou mais) dos b_i's não obedeça (4.4.3.1), não se obteve ainda a convergência e procede-se então a mais um passo no processo iterativo, até que se obtenha a convergência.

IV.4.4 - Uma observação

No programa para computadores, a ser apresentado adiante, há dois critérios que determinam o fim do processo iterativo:

- o critério dado por (4.4.3.1)
- caso não ocorra a convergência em um número máximo, pré-fixado, de iterações (NITER) o programa para e imprime o resultado dos dois últimos passos (NITER-1 e NITER).

A razão disto é que se o processo estiver convergindo muito lentamente, não se desejará chegar a um número absurdo de iterações. Pode estar ocorrendo que a estimativa inicial dos parâmetros, $\hat{B}_0$, não tenha sido adequada. Esta parada, após no máximo NITER iterações, permite

tomar conhecimento do que pode estar havendo e também obriga à realização de um estudo mais acurado dos dados que se tem em mãos, antes de se passar à nova tentativa de ajuste, talvez concluindo pela retirada e/ou inclusão de algum(s) suporte(s), ou pela alteração de $\bar{B}_0$. Outra possível razão da lenta convergência pode ser devido à uma escolha muito severa de ε e δ.

IV.4.5 - Como escolher ε e δ?

Esta escolha depende basicamente de dois fatores:

- à precisão da medida dos dados
- o número de casas decimais (corretas) com que se mede os dados

Não se pode ser muito severo na escolha de ε e δ (severo se refere a números muito pequenos) quando não houver certeza da precisão com que se mediu os dados. Pode-se mesmo, neste caso, obter uma rápida convergência das estimativas, mas não há razões que justifiquem esta escolha severa. Ser severo, neste caso, não garante boa qualidade ao ajuste, visto os dados não serem precisos.

Por outro lado, a escolha deve ser feita de acordo com a quantidade de casas decimais dos dados. Por exemplo, se os dados têm precisão de duas casas decimais, não se deve escolher $\delta = 10^{-15}$ e $\varepsilon = 10^{-2}$ que resultariam em "precisão" de 15 casas decimais para os parâmetros. Neste caso, dependendo da confiabilidade dos dados, seria mais lógico, talvez, escolher $\varepsilon = 10^{-2}$ e $\delta = 10^{-3}$ ou $\delta = 10^{-4}$.

Concluindo, pode-se dizer que a escolha de ε e δ está quase que inteiramente baseada no "bom senso", auxiliado pela confiabilidade dos dados em questão.

No programa utilizado para a construção dos exemplos do capítulo V, utilizou-se $\varepsilon = 10^{-2}$ e $\delta = 10^{-5}$. Esta parece ser uma escolha bastante razoável para a maioria dos casos, possivelmente $\delta = 10^{-6}$ também seria uma boa escolha.

Apêndice ao capítulo IV - Regressão Robusta usando o M-estimador seno

Apresenta-se este apêndice com o intuito apenas de informar, visto que irá se comparar, num exemplo, o desempenho deste estimador com o estimador biponderado.

Este estimador foi desenvolvido por Andrews e baseado nos bons resultados obtidos por ANDREWS e outros (1972), ANDREWS (1974) o extende à problemas de regressão. Como estimador de parâmetros de locação apresenta uma alta eficiência com dados Gausianos e bastante robustez sob desvios extremos da pressuposição de normalidade. Ao se extender este estimador para construir um método de ajuste, obtém-se um método bastante resistente à presença de valores discrepantes.

A.1 - O M-estimador Seno, para parâmetros de locação.

Andrews, seguindo sugestões de Jaeckel, Hampel e outros, desenvolveu o seguinte M-estimador, baseado na função seno, apresentado em ANDREWS e outros (1972). Seja Ψ uma função como na secção III.1. Então:

$$\Psi(z) = \begin{cases} \text{seno } (z/c) & ; \quad |z| < c.\ \pi \\ 0 & ; \quad |z| \geq c.\ \pi \end{cases} \qquad (A.1.1)$$

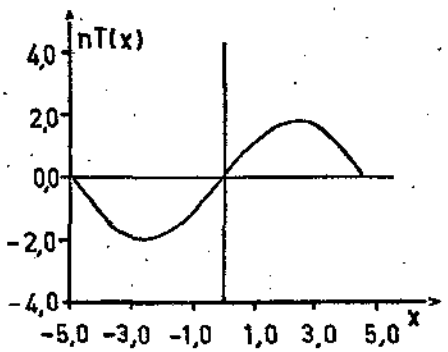
A solução, T, é obtida resolvendo-se

$$\Sigma\Psi \left(\frac{x-T}{S}\right) = 0 \qquad (A.1.2)$$

onde S é uma medida de escala robusta. Originalmente se optou por c=2,1 e S=MDA. O estimador foi identificado com a sigla AMT.

A Curva de Influência (estilizada) para este estimador também é apresentada por ANDREWS e outros (1972). Esta curva está na figura seguinte.

Figura a.1 - Curva de Influência (estilizada) para o M-estimador Seno*



A curva de influência acima foi obtida com uma amostra de 20 valores (19 valores fixos e 1 valor móvel) com distribuição normal. O gráfico representa nT(x) como uma função de x, onde x é a observação móvel e T(x) a solução, para cada x, dada pelo M-estimador seno.

Cabe aqui ressaltar a grande semelhança da curva da figura a.1 com a curva de influência para o estimador biponderado, das figuras 3.1.1.c e 3.1.1.d, apesar de terem sido obtidas através de diferentes amostras. Em ambas nota-se que para observações mais afastadas (valores aberrantes) se associa peso zero. Conforme as observações caminham

---

* ANDREWS e outros (1972), pag. 98.

para a porção central dos dados a C.I. vai aumentando, chegando a um valor máximo, a partir do qual os valores vão diminuindo novamente até zero, quando as observações assumem valores próximos de zero. Com base nisto não é de se estranhar que estes dois estimadores tenham comportamento parecido e apresentem resultados bastante semelhantes.

ANDREWS e outros (1972) e ANDREWS (1974) apresentam a variância assintótica para o M-estimador seno. Pode ser obtida sem maiores dificuldades, para uma variável com densidade f, simétrica com respeito a um ponto μ e com amplitude interquartis σ, pela fórmula:

$$\text{Var}\,(\bar{T}) = \frac{\sigma^2 \int \psi^2 \left(\frac{x-\mu}{\sigma}\right) f(x)\,dx}{\left[\int \psi' \left(\frac{x-\mu}{\sigma}\right) f(x)dx\right]^2} \qquad (*)$$

onde $\psi'(z) = \frac{d\psi(z)}{dz}$. A expressão (*) é bastante parecida com a expressão da variância, para o estimador biponderado, obtida por GROSS (1976), apresentada na secção III.1.1. A variância acima, (*), também depende da constante c. A medida que c cresce o estimador e suas propriedades tendem aos estimadores de mínimos quadrados. ANDREWS (1974) recomenda, para trabalhos exploratórios, c=1,5 ou 1,8, que resultarão em estimadores mais resistentes. Segundo ele, na maioria dos casos o M-estimador seno requer apenas uma iteração, partindo de um ponto inicial $\bar{T}_0$, a mediana dos x_i, já que a expansão de Taylor de primeira ordem, pode ser resolvida, em forma abreviada por:

$$\tan\{(\bar{T}-\bar{T}_0)/cS\} = -\frac{\Sigma \operatorname{sen}\left(\frac{x_i-\bar{T}_0}{c}\right)}{\Sigma \cos\left(\frac{x_i-\bar{T}_0}{c}\right)}$$

se o conjunto dos x_i satisfazendo $|x_i-\bar{T}_0| \leqslant cS$ é o mesmo conjunto que satisfaz $|x_i-\bar{T}| \leqslant cS$ (ambos os somatórios são sobre estes conjuntos).

A.2 - Extensão ao problema da regressão

Os M-estimadores de locação são definidos como sendo as soluções de (A.1.2), onde S pode ser determinado tanto conjunta como independentemente. Isto é equivalente a achar um máximo local da função:

$$\Sigma\ \psi\left(\frac{x_i-T}{S}\right)$$

onde $\Psi(z) = -\frac{d\psi(z)}{dz}$. Nesta segunda forma podem ser extendidos aos modelos de regressão, desde que os (x_i-T) possam ser vistos como os resíduos, r_i, e S como uma medida de escala. A estimativa é definida como os valores dos parâmetros para os quais:

$$\Sigma\ \psi\ (r_i/S)$$

atinge um máximo local. Seja então B um vetor k-dimensional. Os resíduos:

$$r_i(B) = y_i - X_i' B$$

podem ser formados. A medida robusta de escala pode ser (MDA):

$$S(B) = \text{mediana}\ \{|r_i(B)|\}$$

Os parâmetros, no vetor B, podem ser estimados pela localização de um máximo local da função:

$$\Sigma\psi\,(r_i(B)/S(B))$$

onde ψ é a integral de (A.1.1), e é dada por:

$$\psi(z) = \begin{cases} 1 + \cos(z/c) \cdot c & ; \ |z| \leq c\,\pi \\ 0 & ; \ |z| > c\,\pi \end{cases}$$

Um máximo local particular pode ser obtido através de um programa iterativo de otimização que dependerá de um valor inicial $\bar{B}_0$ e do procedimento numérico de maximização utilizado.

ANDREWS (1974) tece comentários sobre a importância do ponto inicial $\bar{B}_0$. No exemplo apresentado no artigo utiliza uma estimativa inicial obtida pelo Método da Regressão pelas Medianas. Além disso trabalha com c=1,5 e S=MDA. Os resultados deste exemplo serão apresentados no capítulo V, onde serão comparados aos resultados obtidos pelo estimador biponderado (exemplo 2 - cap. V).

V - EXEMPLOS

Neste capítulo serão apresentados dois exemplos utilizando o ajuste obtido através do estimador biponderado - a regressão biponderada - apresentada na secção IV.2.

O primeiro exemplo foi processado utilizando o SPSS (Statistical Package for the Social Sciences), existente na UNICAMP. O modo de obter regressão biponderada através do SPSS será explicado na secção VII.5. Neste exemplo se apresentará um conjunto de dados que por não apresentar grandes desvios da pressuposição de normalidade, nem problemas de heterocedasticia, resulta num ajuste de mínimos quadrados muito bom, como poderá ser observado na secção V.1.1. Com base nestes resultados irá se alterar os dados; isto é, os dados serão perturbados artificialmente somando-se valores 5, 10, 20, 50 e 100 a um dos valores observados da variável dependente, Y, escolhido aleatoriamente. Será mostrado que com os dados alterados (e mesmo sem alteração) se obtém, com a regressão biponderada, resultados muito próximos aos obtidos pelo método dos mínimos quadrados para os dados originais.

O segundo exemplo apresentará uma comparação entre os resultados obtidos por mínimos quadrados, pelo M-estimador seno e pelo estimador biponderado. Vai-se mostrar que os resultados obtidos pelos dois métodos robustos estão muito próximos e são bem melhores que o resultado obtido por mínimos quadrados. Este exemplo será processado com um programa em FORTRAN, que será apresentado no capítulo VII.

V.1 - Exemplo 1

Os dados para este exemplo foram obtidos em DRAPER e SMITH (1966), apêndice B, página 366. São também apresentados por HALD (1952) e foram originalmente apresentados por WOODS, STEINOUR e STARKE (1932). São referentes ao calor desenvolvido durante o endurecimento de cimento, de acordo com sua composição química.

Tem-se um conjunto de cinco variáveis; a variável dependente Y, as quatro variáveis independentes, X_1, X_2, X_3 e X_4 e um total de 13 observações. As variáveis são as seguintes:

Y - calor desenvolvido por grama de cimento, medido em calorias

X_1 - quantidade de 3CaO Al_2O_3 (aluminato de tricálcio)

X_2 - quantidade de 3CaO SiO_2 (silicato de tricálcio)

X_3 - quantidade de 4CaO Al_2O_3 Fe_2O_3 (aluminoferrite tetracálcio)

X_4 - quantidade de 2CaO SiO_2 (silicato de dicálcio)

X_1, X_2, X_3 e X_4 são medidas como porcentagem do peso do cimento.

Os dados estão no quadro 5.1.1.

Quadro 5.1.1 - Dados para o Exemplo 1

Y	X_1	X_2	X_3	X_4
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

Não se trabalhará com todas as variáveis independentes. Escolher-se-á um conjunto com duas ou menos variáveis para prosseguir com o exemplo. DRAPER e SMITH (1966), ao analizarem estes dados, no capítulo 6, concluem que as duas melhores regressões que se pode obter, com duas ou menos variáveis são:

$$\hat{Y} = f_1 (X_1, X_2)$$

$$\hat{Y} = f_2 (X_1, X_4)$$

A primeira é a que se obtem por vários procedimentos, tais como regressão para atrás* (backward regression), regressão vai e vem* (stepwise regression) e também a que produz maior R^2 (dentre os ajustes com duas ou menos variáveis). Por outro lado, $\hat{Y}=f_2(X_1,X_4)$ apresenta um R^2 pouco (muito pouco) menor, mas em compensação inclui a variável independente que isoladamente melhor "explica" a variável Y (X_4). Com base nesses resultados optou-se, neste exemplo, pela segunda forma; isto é, vai-se ajustar:

$$\hat{Y}=f_2\ (X_1,X_4)$$

ou seja, deseja-se $\hat{b}_o$, $\hat{b}_1$ e $\hat{b}_4$ tais que

$$\hat{Y} = \hat{b}_o + \hat{b}_1X_1 + \hat{b}_4X_4$$

V.1.1 - Ajuste obtido por Mínimos Quadrados

Com os dados do quadro 5.1.1 se obtém, por mínimos quadrados:

$\hat{b}_o$ = 103,09738

$\hat{b}_1$ = 1,43996

$\hat{b}_4$ = -0,61395

O quadro 5.1.1.1 apresenta os valores de Y, $\hat{Y}$ (Y ajustado), R (resíduos) e RO (resíduos ordenados), para este ajuste.

---

* Tradução segundo DACHS (1978).

Quadro 5.1.1.1 - Valores de Y, $\hat{Y}$, R e RO para o ajuste:

$$\hat{Y} = 103{,}09738 + 1{,}43996\ X_1 - 0{,}61395\ X_4$$

N	Y	$\hat{Y}$	R	RO	Obs.corresp.
1	78,5	76,34095	2,15905	3,76967	6
2	74,3	72,61268	1,68732	2,98279	5
3	104,3	106,65821	-2,35821	2,15905	1
4	87,6	90,08194	-2,48194	1,73031	12
5	95,9	92,91721	2,98279	1,68732	2
6	109,2	105,43033	3,76967	0,62929	9
7	102,7	103,73364	-1,03364	0,13648	11
8	72,5	77,52418	-5,02418	-0,72973	13
9	93,1	92,47071	0,62929	-1,03364	7
10	115,9	117,37417	-1,47417	-1,47417	10
11	83,8	83,66352	0,13648	-2,35821	3
12	113,3	111,56969	1,73031	-2,48194	4
13	109,4	110,12973	-0,72973	-5,02418	8

Como se pode notar do quadro acima, não há resíduos muito grandes, indicando a primeira vista que não deve haver valores aberrantes. Ao realizar uma rápida análise, através do que DACHS (1978) denomina Esquema de Cinco Números, se obtém:

13	
*	-5,02418
j	-1,47417
M	0,13648
j	1,73031
*	3,76967

dj = 1,73031 - (-1,47417) = 3,20448

-1,47417 - (3/2) dj = -6,28089

-1,47417 - dj = -4,67865

1,73031 + (3/2) dj = 6,53703

1,73031 + dj = 4,93479

Toda observação que apresentar um resíduo menor que -6,28089 ou maior que 6,53703 será um Ponto Solto ou valor aberrante, como se pode observar não há este caso no quadro 5.1.1.1. Todo resíduo que estiver entre -4,67865 e -6,28089 ou entre 4,93479 e 6,53703 será correspondente a um Ponto Externo, e nesse caso somente há um valor, -5,02418, correspondente à oitava observação.

Além disso tem-se:

$R^2 = 0,97247$

que é um valor bastante elevado.

Nas figuras 5.1.1.2 e 5.1.1.3 tem-se, respectivamente, o gráfico da distribuição acumulada dos resíduos e o gráfico dos resíduos versus Y ajustados.

Figura 5.1.1.2 - Gráfico da Distribuição Acumulada dos Resíduos.

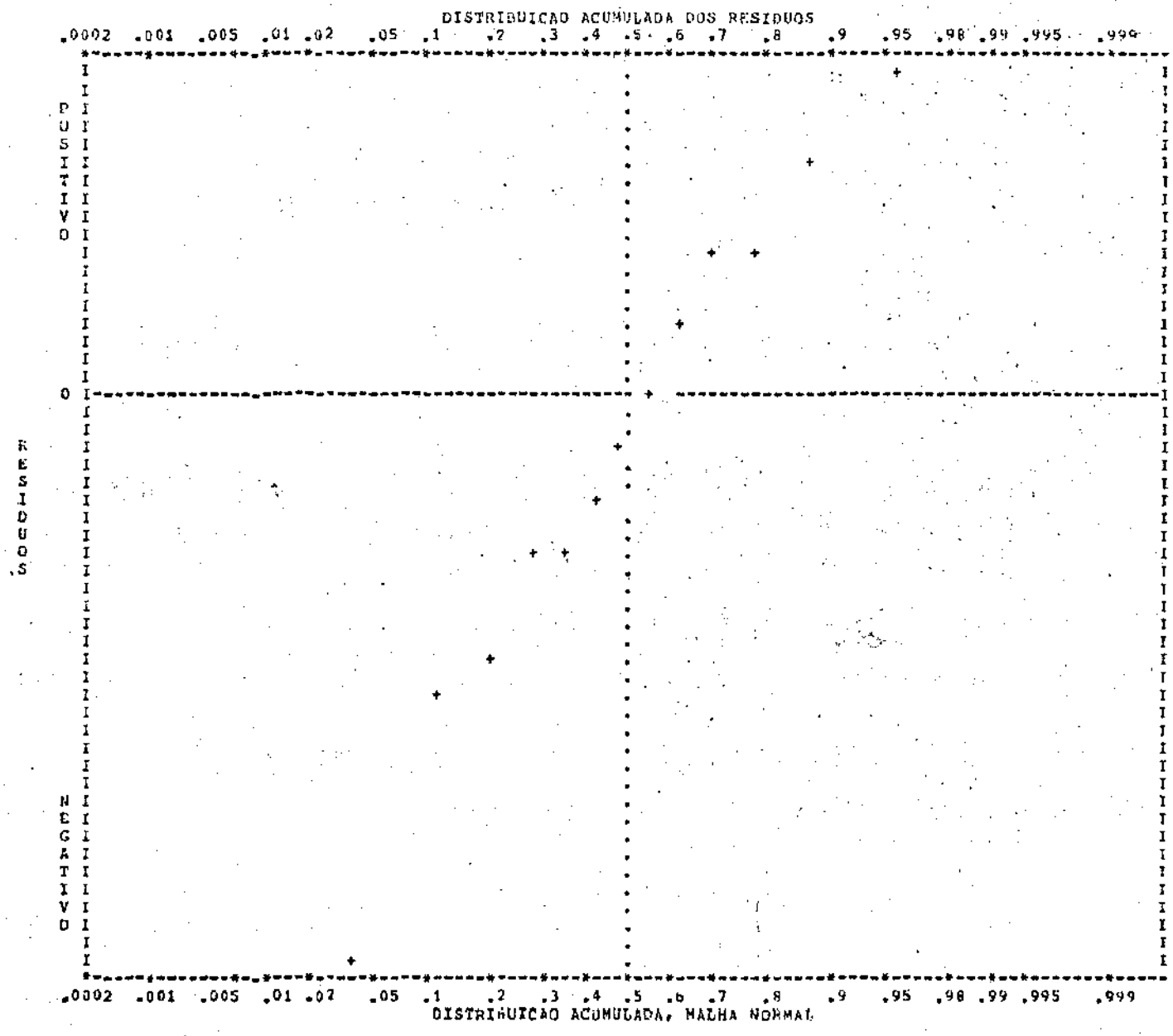
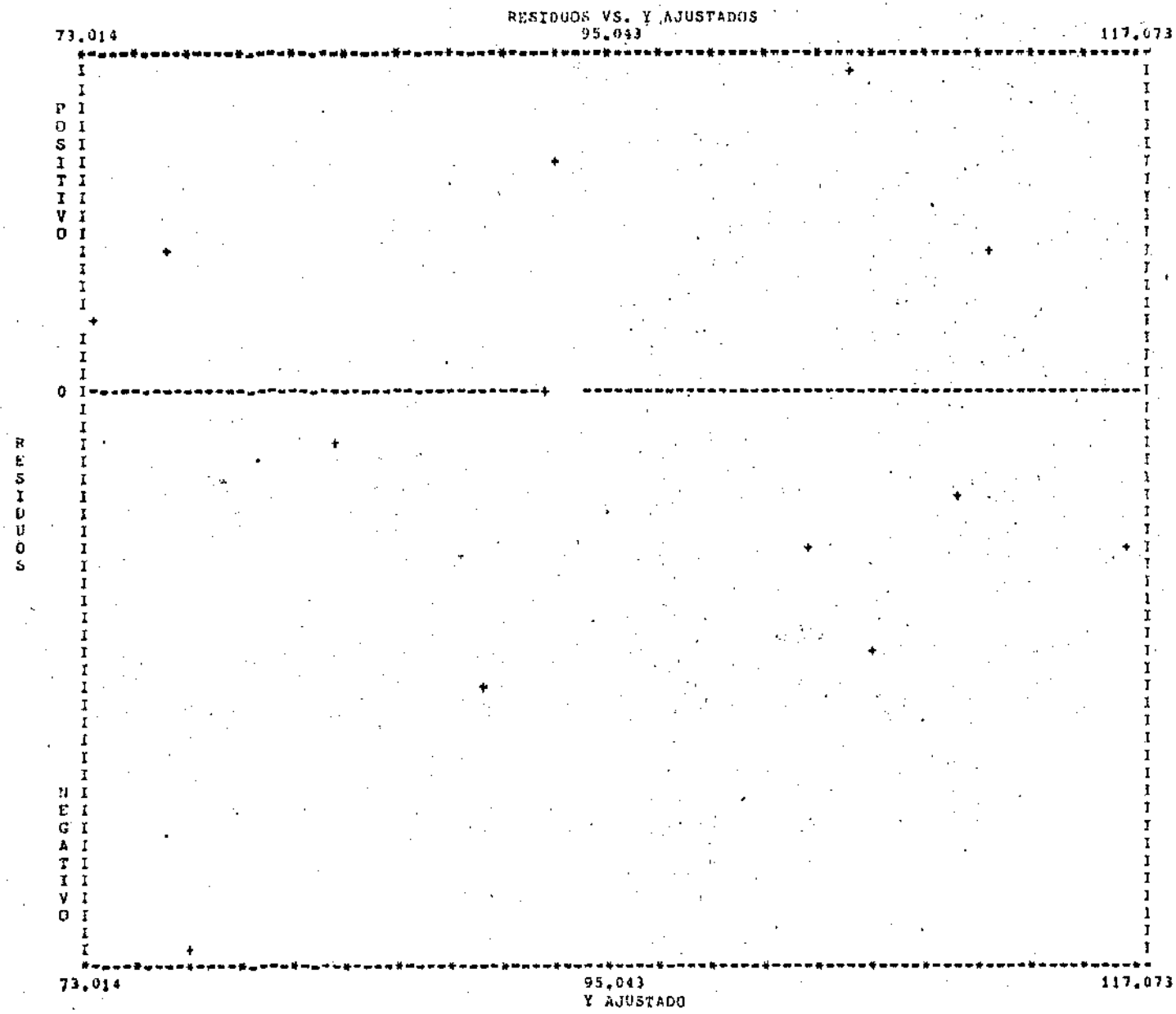


Figura 5.1.1.3 - Gráfico de Resíduos vs. Y ajustados



Das figuras 5.1.1.2 e 5.1.1.3 pode-se dizer que se tem um bom ajuste por mínimos quadrados. A figura 5.1.1.2 mostra que não há grandes desvios da suposição de normalidade, enquanto a figura 5.1.1.3 indica que o modelo ajustado ($Y = b_0+b_1X_1+b_4X_4$) é adequado, também não apresentando problemas aparentes de heterocedasticia. Isso aliado às conclusões anteriores permite que se afirme que o ajuste obtido é muito bom. Então, deve-se esperar, que qualquer outro bom método de ajuste quando aplicado aos dados do quadro 5.1.1 (para o mesmo modelo), produza resultados bastante próximos aos de mínimos quadrados. Isto é o que irá ser mostrado na secção V.1.2 utilizando a regressão biponderada, que como já se afirmou é um método que produz resultados muito bons.

V.1.2 - Ajustes obtidos pela Regressão Biponderada

Foi escolhida, aleatoriamente, a nona observação da variável dependente, Y, para ser artificialmente perturbada. Vai-se realizar os ajustes, segundo a regressão biponderada, em seis diferentes situações; isto é, vai-se alterar os dados de seis modos diferentes. O quadro 5.1.2.1 apresenta estas alterações.

Deve-se observar que a primeira "alteração" na realidade corresponde aos dados originais inalterados. Proceder-se-á então ao ajuste através de regressão biponderada com os dados do quadro 5.1.1, trocando-se Y_9 por Y_9^* do quadro 5.1.2.1.

Quadro 5.1.2.1 - Valores de Y_9 devido às seis alterações provocadas

Y_9	alteração	Y_9^*
93,1	+ 0	93,1
93,1	+ 5	98,1
93,1	+ 10	103,1
93,1	+ 20	113,1
93,1	+ 50	143,1
93,1	+100	193,1

O quadro 5.1.2.2 apresenta os resultados obtidos em cada caso, tanto com o método dos mínimos quadrados como também com a regressão biponderada, onde foi utilizado:

$C = 4$

S = amplitude interquartís/1,35

$\delta = 10^{-5}$

estimativa inicial - mínimos quadrados

Quadro 5.1.2.2 - Estimativas de b_0, b_1 e b_4 obtidas por mínimos quadrados e pela regressão biponderada nos diferentes conjuntos de dados

Regressão	Dados	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_4$
Mínimos Quadrados	Originais	103,09738	1,43996	-0,61395
	$Y_9^*=Y_9+5$	104,63104	1,36114	-0,63265
	$Y_9^*=Y_9+10$	106,16470	1,28231	-0,65135
	$Y_9^*=Y_9+20$	109,23203	1,12467	-0,68874
	$Y_9^*=Y_9+50$	118,43399	0,65173	-0,80092
	$Y_9^*=Y_9+100$	133,77061	-0,13650	-0,98789
Biponderada	Originais	103,16659	1,41356	-0,60695
	$Y_9^*=Y_9+5$	104,49670	1,34505	-0,62199
	$Y_9^*=Y_9+10$	103,39689	1,40191	-0,60962
	$Y_9^*=Y_9+20$	102,96635	1,42469	-0,60499
	$Y_9^*=Y_9+50$	102,96639	1,42469	-0,60499
	$Y_9^*=Y_9+100$	102,96638	1,42469	-0,60499

Ao examinar o quadro 5.1.2.2 nota-se que:

- as estimativas, obtidas por mínimos quadrados, para os tres parâmetros, variam bastante ao mudar o conjunto de dados. Observa-se que a estimativa de b_0 cresce com o aumento do valor de Y_9, a estimativa de b_1 decresce, chegando mesmo a mudar de sinal e a estimativa de b_4 também cresce.

- no caso da regressão biponderada observa-se que praticamente não há variações ao aumentar o valor de Y_9. Com os dados originais e com as alterações de 20, 50 e 100 há diferenças mínimas entre as estimativas, não diferindo muito das estimativas obtidas por mínimos quadrados com os dados originais. Com a alteração de 5 em Y_9 as estimativas obtidas com a regressão biponderada diferem um pouco mais das obtidas por mínimos quadrados com os dados originais, mas não diferem muito das estimativas obtidas por mínimos quadrados com a alteração de 5 em Y_9. Isso era de se esperar pois mesmo com esta alteração ainda se obtém um bom ajuste de mínimos quadrados. A alteração de 10 em Y_9 fica numa posição intermediária. Difere um pouco mais do obtido por mínimos quadrados com os dados originais que no caso das alterações de 20, 50 e 100, mas difere menos que no caso da alteração de 5 em Y_9.

De um modo geral pode-se dizer que a regressão biponderada se mostrou insensível aos erros artificialmente produzidos, enquanto que o método dos mínimos quadrados se mostrou bastante sensível. Mesmo com as alterações a regressão biponderada produziu ótimos resultados, bastante próximos aos obtidos por mínimos quadrados com os dados originais.

Esses fatos são confirmados pelo quadro 5.1.2.3, que apresenta a variação porcentual entre as estimativas obtidas com a regressão biponderada para todos os conjuntos de dados e as estimativas obtidas por mínimos quadrados com os dados originais.

Quadro 5.1.2.3 - Variação porcentual entre as estimativas obtidas pela regressão biponderada e as estimativas obtidas por mínimos quadrados com os dados originais

Dados	Variação porcentual		
	$\tilde{b}_0$	$\tilde{b}_1$	$\tilde{b}_4$
Originais	0,067	1,833	1,140
$Y_9^* = Y_9 + 5$	1,357	6,591	1,310
$Y_9^* = Y_9 + 10$	0,291	2,642	0,706
$Y_9^* = Y_9 + 20$	0,127	1,060	1,459
$Y_9^* = Y_9 + 50$	0,127	1,060	1,459
$Y_9^* = Y_9 + 100$	0,127	1,060	1,459

Do quadro 5.1.2.3 pode-se notar que:

- os resultados obtidos por mínimos quadrados e pela regressão biponderada com os dados originais estão bem próximos. A maior variação porcentual foi de 1,833%, para $\tilde{b}_1$ e é bastante pequena.
- no caso das alterações de 20, 50 e 100 em Y_9 a maior variação porcentual ocorrida com $\tilde{b}_4$, foi de 1,459% e pode ser considerada quase que desprezível. Isso ilustra a resistência da regressão biponderada. Ilustra também o "isolamento" e "controle" de valores aberrantes.
- com a alteração de 5 em Y_9 ocorreram as maiores variações porcentuais. No entanto ao se verificar a variação porcentual entre as estimativas

obtidas por mínimos quadrados com os dados alterados de 5 nota-se que a maior variação é inferior a 2%.

Há ainda um último fato a destacar. Está relacionado com o número de iterações necessárias para se obter a convergência. O quadro 5.1.2.4 apresenta o número necessário de iterações para se obter a convergência.

Quadro 5.1.2.4 - Quantidade de iterações necessárias para se obter convergência com a regressão biponderada, para os vários conjuntos de dados

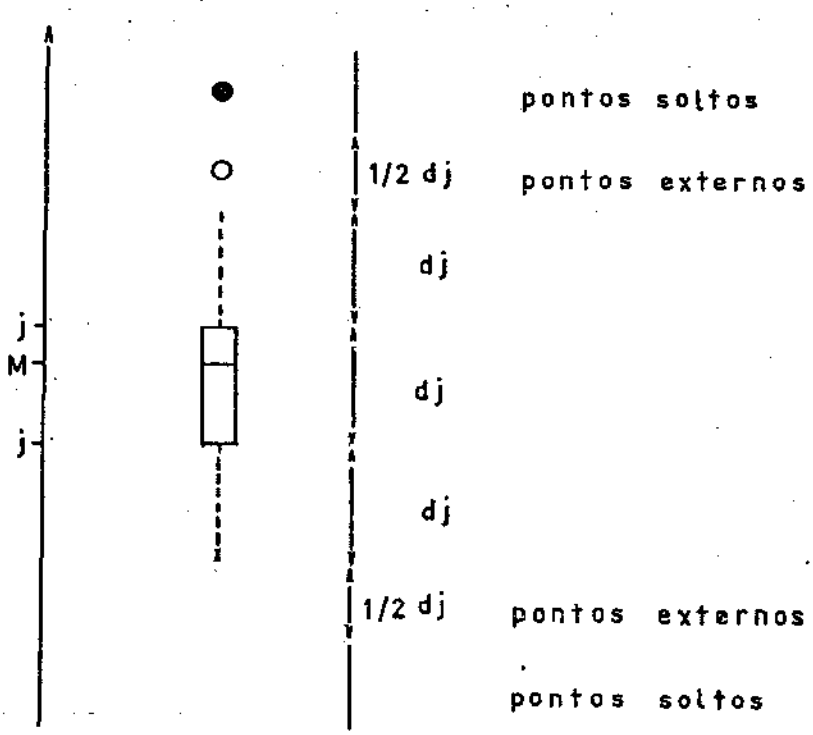
Dados	Iterações
Originais	7
$Y_9^* = Y_9 + 5$	7
$Y_9^* = Y_9 + 10$	14
$Y_9^* = Y_9 + 20$	6
$Y_9^* = Y_9 + 50$	6
$Y_9^* = Y_9 + 100$	7

A não ser no caso da alteração de 10 em Y_9, o número necessário de iterações foi 6 ou 7, apesar da grande "precisão" exigida ($\delta = 10^{-5}$). Isso ilustra um fato já esperado. Quando uma observação ocupa uma posição crítica; isto é, está perto do limite entre ser ou não um valor aberrante, deve-se tomar bastante cuidado ao considerá-la ou ao afastá-la da amostra. A regressão biponderada faz isso "automaticamente",

como se percebe através do número de iterações necessárias para a convergência no caso da alteração de 10 em Y_9.

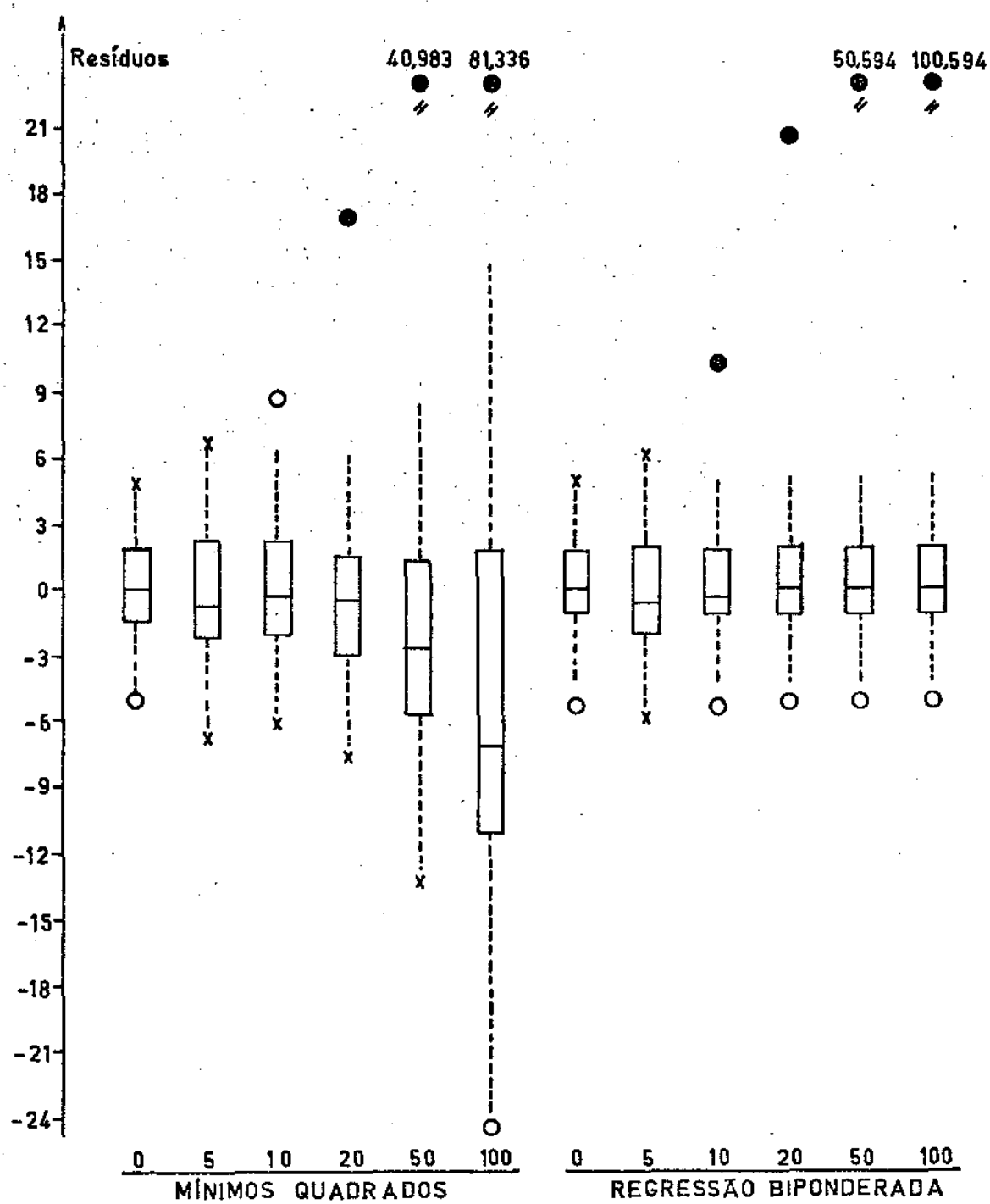
A figura 5.1.2.6 ilustra, de modo a não deixar dúvidas, esses fatos. É construída com base no apresentado por DACHS (1978), lá denominado Desenho Esquemático. Este desenho é feito com o auxílio de um esquema, denominado Esquema de Cinco Números (DACHS (1978)), como o apresentado logo após o quadro 5.1.1.1. Tem-se então os dois extremos, as duas juntas (quartis), a mediana e a quantidade de observações. Com esses dados calcula-se a distância entre juntas, dj, e então passa-se ao desenho. A figura 5.1.2.5 mostra como se deve proceder.

Figura 5.1.2.5 - Representação de um Desenho Esquemático com os pontos e as regiões correspondentes (DACHS (1978), pág. 14)



Os pontos soltos são tomados como valores aberrantes. A regra de tomar dj e (3/2)dj como limites para a determinação de pontos externos e de pontos soltos não é fixa. Admite certas mudanças, como por exemplo tomar (3/2)dj e 2dj (TUKEY (1977)) ou mesmo dj e 2dj como sendo os limites.

Figura 5.1.2.6 - Desenho Esquemático para os resíduos obtidos ao se fazer ajustes de mínimos quadrados e segundo a regressão biponderada no caso dos dados originais e dos dados alterados



De imediato nota-se dois fatos:

- os pontos externos e os pontos soltos aparecem com maior destaque com a regressão biponderada do que com o mētodo dos mīnimos quadrados, para todas as alterações. Isso ilustra a "detecção" dos valores aberrantes. No caso da alteração de 10 em Y_9 o mētodo dos mīnimos quadrados apresenta pontos externos, enquanto que a regressão biponderada apresenta pontos soltos (valores aberrantes).
- a distância entre juntas (amplitude interquartīs) vai crescendo bastante com o aumento da alteração em Y_9, para o mētodo dos mīnimos quadrados enquanto que para a regressão biponderada praticamente não varia. No caso da alteração de 5 em Y_9 ocorre a maior distância entre juntas com a regressão biponderada, mas ainda assim ē menor que para o mētodo dos mīnimos quadrados (com a alteração de 5 no valor de Y_9).

Outro fato de destaque ē a variação que ocorre com o valor da mediana dos resīduos, no caso do mētodo dos mīnimos quadrados. Parece haver uma tendência da mediana a ir diminuindo com o aumento da alteração na nona observação. Com a regressão biponderada não se nota nenhuma tendência.

De um modo geral, a figura 5.1.2.6 ilustra a fragilidade do mētodo dos mīnimos quadrados quando hā valores aberrantes nos dados, destacando o bom desempenho da regressão biponderada com esses dados. Outro ponto que se destaca ē o bom desempenho do mētodo robusto, a regressão biponderada, com dados sem valores discrepantes e/ou mesmo sem grandes desvios da pressuposição de normalidade.

V.2 - Exemplo 2

DANIEL e WOOD (1971), no capítulo 5, consideram com bastante detalhes um exemplo com 21 observações e 3 variáveis independentes. Estes dados também são apresentados por BROWNLEE (1965), DRAPER e SMITH (1966) e ANDREWS (1974). São referentes à transformação de amônia em ácido nítrico, realizada numa planta química. Tem-se então 4 variáveis; a variável dependente Y, as variáveis independentes X_1, X_2 e X_3 e 21 observações, correspondendo ao total de dias em que se mediu a operação da planta.

A variável Y é medida como sendo 10 vezes a porcentagem de amônia que é perdida como óxido nítrico não absorvido; é uma medida indireta da quantidade de amônia que é transformada em ácido nítrico. As variáveis independentes são as seguintes:

X_1 - corrente de ar para a planta

X_2 - temperatura de entrada da água de resfriamento na torre de absorção de óxido nítrico

X_3 - concentração de ácido nítrico no líquido de absorção (codificado para facilidades de cálculo subtraindo-se 50 e então multiplicando o resultado por 10).

Os dados estão no quadro 5.2.1

Quadro 5.2.1 - Dados referentes a 21 dias de operação de uma planta, convertendo amônia em ácido nítrico

Y	X_1	X_2	X_3
42	80	27	89
37	80	27	88
37	75	25	90
28	62	24	87
18	62	22	87
18	62	23	87
19	62	24	93
20	62	24	87
15	58	23	87
14	58	18	80
14	58	18	89
13	58	17	88
11	58	18	82
12	58	19	93
8	50	18	89
7	50	18	86
8	50	19	72
8	50	19	79
9	50	20	80
15	56	20	82
15	70	20	91

DRAPER e SMITH (1966) apresentam esses dados no exemplo D do capítulo 6. Ao apresentarem as Respostas aos Exercícios apresentam o modelo ajustado que lhes pareceu melhor:

$$\hat{Y} = -50,35884 + 0,67115\ X_1 + 1,29535\ X_2 \qquad (5.2.1.1)$$

Comentam que o modelo ajustado é menos satisfatório em $(X_1, X_2, X_3) = (70, 20, 91)$, correspondente a 21a. observação. Dizem que se poderia relutar em utilizar a equação (5.2.1.1) nas vizinhanças deste ponto.

DANIEL e WOOD (1971) ajustam o modelo:

$$\hat{Y} = -39,9197 + 0,71564\ X_1 + 1,29529\ X_2 - 0,15213\ X_3 \qquad (5.2.1.2)$$

e observam de um gráfico da distribuição acumulada dos resíduos, que a observação de número 21 tem um resíduo anormalmente grande. Esta observação pode ter alterado substancialmente os coeficientes do modelo ajustado. Após muito trabalho cuidadoso retiram esta observação e mais outras tres; as de números 1, 3 e 4, apresentando explicações para o comportamento não usual destes pontos. Com base nas 17 observações restantes obtém:

$$\hat{Y} = -37,65245 + 0,79769\ X_1 + 0,57734\ X_2 - 0,06706\ X_3 \qquad (5.2.1.3)$$

dizendo não haver agora maiores problemas com valores aberrantes.

ANDREWS (1974) ajusta estes dados (com as 21 observações) através do método robusto de ajuste apresentado no apêndice do capítulo IV, com c=1,5 e S=MDA (mediana dos desvios absolutos da mediana), obtendo:

$$\hat{Y} = -37,2 + 0,82\ X_1 + 0,52\ X_2 - 0,07\ X_3 \qquad (5.2.1.4)$$

chegando aos mesmos coeficientes, praticamente, da equação (5.2.1.3) obtida por DANIEL e WOOD (1971), sem as observações 1, 3, 4 e 21. Isso ilustra a não influência de uma certa quantidade (nesse caso quase 25%) de valores aberrantes no desempenho do método. Para confirmar ANDREWS (1974) faz um novo ajuste, agora sem as observações consideradas aberrantes e obtêm novamente a equação (5.2.1.4), reafirmando a não influência de valores discrepantes no método de ajuste baseado no M-estimador seno.

O quadro 5.2.2 apresenta os resíduos obtidos com as equações (5.2.1.2), (5.2.1.3) e (5.2.1.4)

Quadro 5.2.2 - Resíduos obtidos com as equações (5.2.1.2), (5.2.1.3) e (5.2.1.4) com os dados do quadro 5.2.1

Y	equação (5.2.1.2)	equação (5.2.1.3)	equação (5.2.1.4)
42	3,24	6,08	6,11
37	-1,92	1,15	1,04
37	4,56	6,44	6,31
28	5,70	8,18	8,24
18	-1,71	-0,67	-1,24
18	-3,01	-1,25	-0,71
19	-2,39	-0,42	-0,33
20	-1,39	0,58	0,67
15	-3,14	-1,06	-0,97
14	1,27	0,35	0,14
14	2,64	0,96	0,79
13	2,78	0,47	0,24
11	-1,43	-2,51	-2,71
12	-0,05	-1,34	-1,44
8	2,36	1,34	1,33
7	0,91	0,14	0,11
8	-1,52	-0,37	-0,42
8	-0,46	0,10	0,08
9	-0,60	0,59	0,63
15	1,41	1,93	1,87
15	-7,24	-8,63	-8,91

Os resíduos sublinhados com um traço são referentes às observações que não entraram no cálculo das estimativas.

V.2.1 - Regressão Biponderada

Devido à pequena quantidade de dígitos apresentada por ANDREWS (1974), apenas 3, na equação (5.2.1.4), não é necessário realizar iterações até se atingir a convergência utilizando um δ muito pequeno.

Deste modo escolheu-se:

$$\delta = 10^{-3}$$

e além disso tomou-se:

$$c = 4$$

e

$$S = \text{amplitude interquartís}/1{,}35$$

Realizado o ajuste obteve-se:

$$\hat{Y} = -37{,}314 + 0{,}811\ X_1 + 0{,}540\ X_2 - 0{,}071\ X_3 \qquad (5.2.1.5)$$

Estes resultados não diferem muito dos obtidos por DANIEL e WOOD (1971) na equação (5.2.1.3) nem dos obtidos por ANDREWS (1974) na equação (5.2.1.4). A variação porcentual entre as estimativas obtidas pela regressão biponderada e as obtidas por mínimos quadrados (com 17 observações) e entre as estimativas obtidas pelos dois métodos robustos estão no quadro 5.2.3.

Quadro 5.2.3 - Variação porcentual entre os parâmetros das equações (5.2.1.5) e (5.2.1.3) e entre os parâmetros das equações (5.2.1.5) e (5.2.1.4)

Parâmetros	Variação Porcentual $\left\|\frac{(5.2.1.3)-(5.2.1.5)}{(5.2.1.5)}\right\|$	$\left\|\frac{(5.2.1.4)-(5.2.1.5)}{(5.2.1.5)}\right\|$
$\bar{b}_0$	0,907	0,306
$\bar{b}_1$	1,641	1,110
$\bar{b}_2$	6,915	3,704
$\bar{b}_3$	5,549	1,408

No quadro 5.2.3 nota-se que realmente quase não há diferenças entre as estimativas obtidas pelos métodos robustos de ajuste, bastando citar o fato da maior variação porcentual ser inferior a 4%. Já entre a regressão biponderada e o método dos mínimos quadrados (com 17 observações) as variações são pouco maiores, sem chegar a serem exageradas. A maior ocorre com $\bar{b}_2$ e é igual a 6,915%.

O quadro 5.2.4 apresenta os valores de Y, $\hat{Y}$ (Y ajustados) e dos resíduos, R, para o ajuste dado pela equação (5.2.1.5), para os dados do quadro 5.2.1.

Quadro 5.2.4 - Valores de Y, $\hat{Y}$ e R para o ajuste:

$$\hat{Y} = -37,31424 + 0,81089\ X_1 + 0,54005\ X_2 - 0,07060\ X_3$$

Y	$\hat{Y}$	R
42	35,85477	6,14523
37	35,92537	1,07463
37	30,64965	6,35035
28	19,77988	8,22012
18	18,69979	-0,69979
18	19,23983	-1,23983
19	19,35630	-0,35630
20	19,35630	0,64370
15	15,99629	-0,99629
14	13,79024	0,20976
14	13,15486	0,84514
13	12,68542	0,31458
11	13,64904	-2,64904
12	13,41252	-1,41252
8	6,66777	1,33223
7	6,87956	0,12044
8	8,40797	-0,40797
8	7,91379	0,08621
9	8,38324	0,61676
15	13,10736	1,89264
15	23,82439	-8,82439

A seguir apresenta-se um Esquema de Cinco Números para os resíduos do quadro 5.2.4:

21	
*	-8,82439
j	-0,69979
M	0,20976
j	1,07463
*	8,22012

dj = 1,07463 -(-0,69979) = 1,7742

Os limites para a detecção de possíveis valores discrepantes são:

-0,69979 - dj = -2,47421 e -0,69979 -(3/2) dj = -3,36142

e

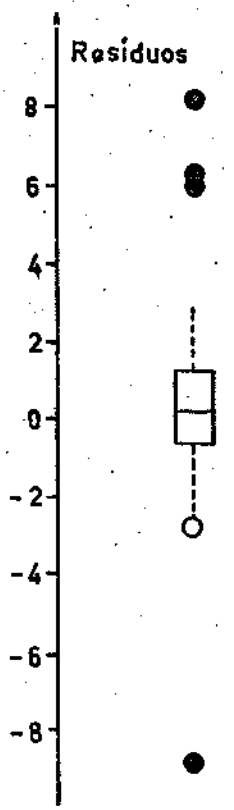
1,07463 + dj = 2,84905 e 1,07463 + (3/2) dj = 3,73626

Com base nesses limites tem-se:

- o ponto de número 13, com resíduo -2,64904, é um ponto externo
- os pontos de números 1, 3, 4 e 21, com resíduos, respectivamente, 6,14523; 6,35035; 8,22012 e -8,82439, são pontos soltos ou valores aberrantes.

O Desenho Esquemático da figura 5.2.5 ilustra melhor a esses fatos:

Figura 5.2.5 - Desenho esquemático para os resíduos do quadro 5.2.4



As figuras 5.2.6 e 5.2.7 apresentam, respectivamente, o gráfico da Distribuição Acumulada dos Resíduos e o gráfico dos Resíduos vs. Y ajustados.

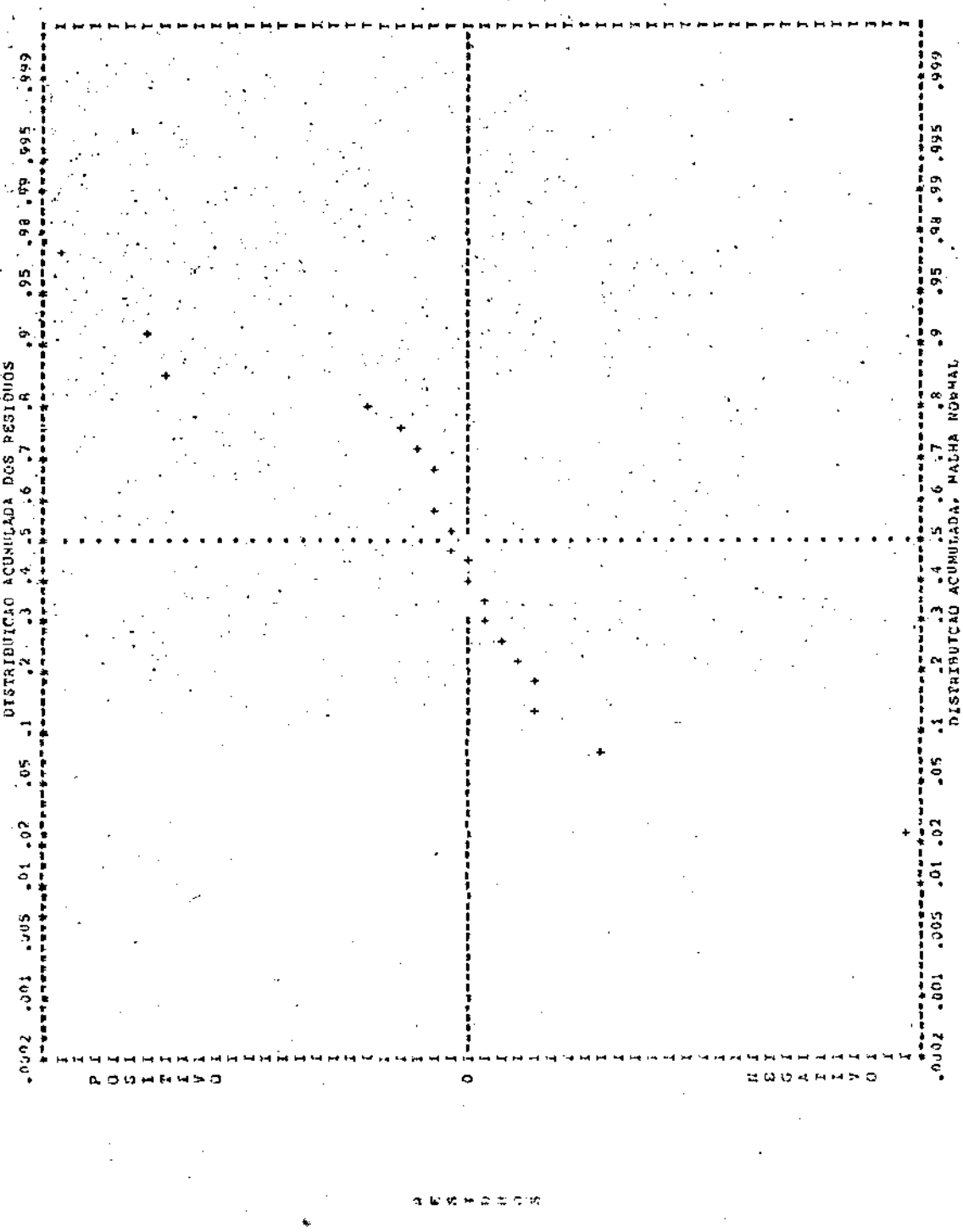
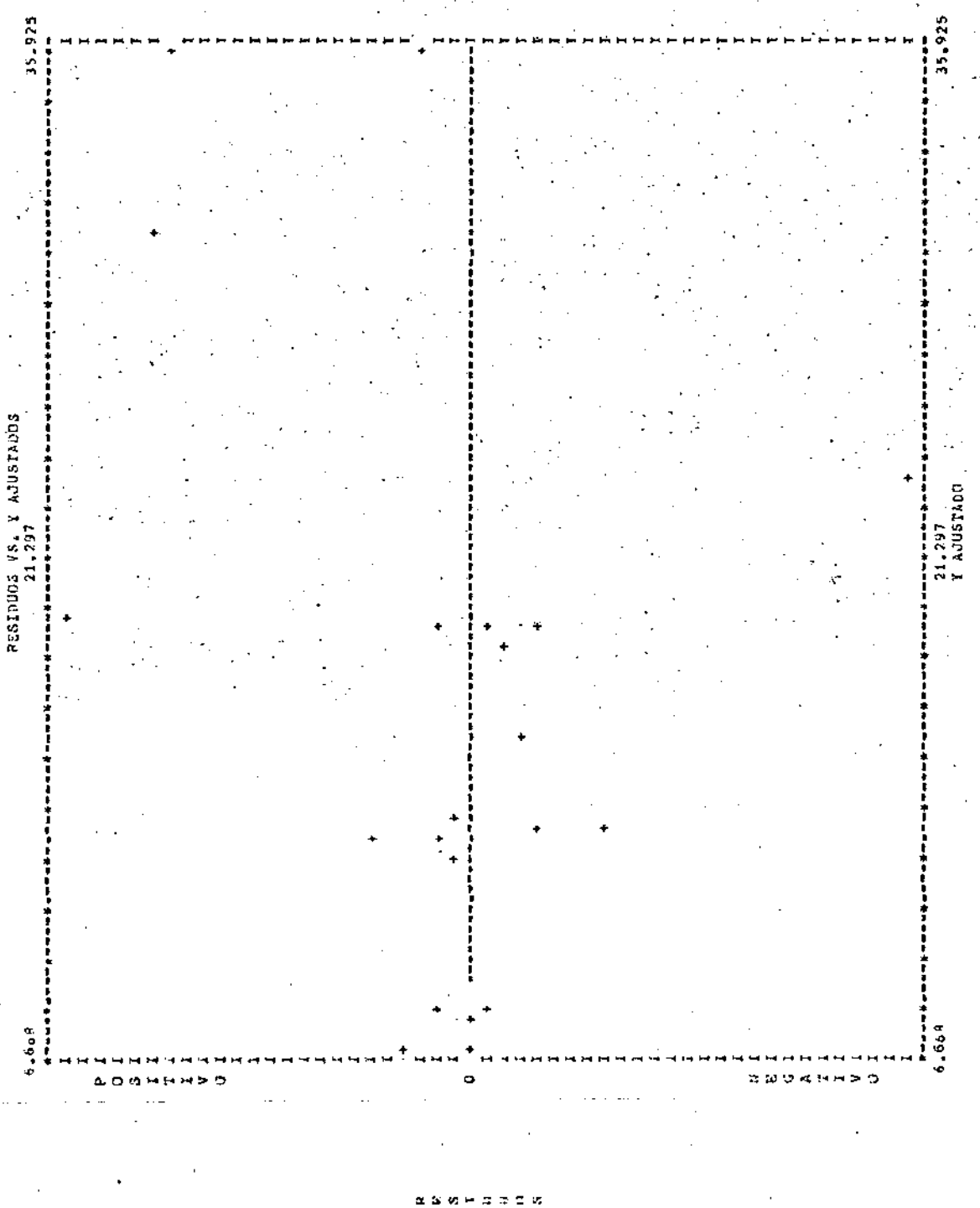


Figura 5.2.6 - Gráfico da Distribuição Acumulada dos Resíduos

Figura 5.2.7 - Gráfico de Resíduos vs. Y ajustados



Na figura 5.2.6 nota-se 4 pontos distantes dos demais.São os pontos referentes às observações de números 1, 3, 4 e 21. Há ainda o resíduo referente a 13a. observação, um pouco mais longe da maioria dos resíduos, mas não tão longe quanto o resíduo das observações discrepantes.

Na figura 5.2.7 nota-se a distância dos pontos referentes às observações 1, 3, 4 e 21 dos demais e também se destaca a presença do ponto referente à observação de número 2. O que ocorre com as observações 1, 3, 4 e 21 já é bem sabido. Para verificar o que ocorre com a observação de número 2 volta-se ao quadro 5.2.1 e observa-se que os valores de (X_1, X_2, X_3), para esta observação são dos mais altos, causando então um valor de $\bar{Y}$ mais alto que os demais. Logo não há nenhum problema aparente com a observação de número 2.

Além destes fatos as figuras 5.2.6 e 5.2.7 não apresentam problemas aparentes de grandes desvios da pressuposição de normalidade e nem problemas de heterocedasticia ou inadequação do modelo.

A regressão biponderada, assim como o método robusto de ajuste de ANDREWS (1974), de imediato, apresentam os resultados que DANIEL e WOOD (1971) tiveram bastante trabalho para obter com o método dos mínimos quadrados. Os resultados finais, nos tres casos, não diferem quase nada, indicando que os tres métodos tem desempenho semelhante. No entanto a quantidade de trabalho com o método dos mínimos quadrados é exorbitantemente maior, tornando-o contra-indicado. Por outro lado o método proposto por ANDREWS (1974), mesmo dando menos trabalho é de obtenção difícil. Deste modo parece lícito propor a regressão biponderada

como melhor alternativa, pois poupa trabalho em relação ao método dos mínimos quadrados, e é de fácil obtenção.

Como curiosidade se realizou o ajuste, com a regressão biponderada, retirando-se as observações 1, 3, 4 e 21, obtendo-se:

$$\hat{Y} = -37,075 + 0,820\ X_1 + 0,514\ X_2 - 0,073\ X_3$$

estimativas estas que praticamente não diferem nada das obtidas com a regressão biponderada para as 21 observações. Isso também ilustra a não influência de alguns valores aberrantes no desempenho da regressão biponderada. Caso houvesse uma porcentagem maior de observações discrepantes provavelmente o desempenho seria inferior.

Na secção VII.4 se apresentará a listagem completa dos resultados obtidos no computador para o ajuste com a regressão biponderada para as 17 observações. Estes resultados virão logo após a listagem do programa, exemplificando a saida deste.

VI - RESUMO E CONCLUSÕES

Foram apresentados procedimentos robustos para a estimação de parâmetros de locação e suas extensões aos problemas de regressão. Indiretamente também foram apresentados estimadores robustos para parâmetros de escala. Foi dada maior ênfase para o ajuste de regressões. Isso tem uma explicação; ANDREWS e outros (1972) praticamente esgotaram o assunto sobre estimadores robustos para parâmetros de locação.

Deu-se grande atenção ao que foi denominado regressão biponderada, mas assim mesmo permanecem muitas dúvidas para serem sanadas no futuro, como é o caso do valor da constante de escalonamento, c e da medida robusta de escala, S.

Todas as qualidades do ajuste através da regressão biponderada, colocadas quando de sua apresentação, no capítulo IV, puderam ser verificadas, facilmente, nos exemplos do capítulo V. Foi possível perceber, sem grandes problemas, a "detecção" e o "controle" de valores aberrantes, sendo a detecção melhor ilustrada no exemplo 2 do capítulo V, e o controle no exemplo 1, no caso, principalmente, da alteração de 10

em Y_9. Ainda no exemplo 1, com os dados originais, foi muito bem ilustrado o bom desempenho da regressão biponderada no caso de não haver valores aberrantes e grandes desvios da suposição de normalidade. Tanto no exemplo 1 como no exemplo 2 percebe-se facilmente que a regressão biponderada é praticamente insensível à presença de uns poucos valores aberrantes nos dados. Note que no caso do exemplo 2 havia quase que 25% dos dados com problemas de discrepância e, ainda assim, se obteve um bom desempenho.

O exemplo 2 serviu também para ilustrar a comparação entre o desempenho do ajuste através da regressão biponderada, do ajuste através do método dos mínimos quadrados e do ajuste através do método proposto por ANDREWS (1974). Como foi dito no capítulo V, parece bastante válido considerar a regressão biponderada como melhor alternativa, evitando o enorme trabalho com o método dos mínimos quadrados e as maiores complicações de obtenção do método de Andrews.

No entanto ainda restam problemas a serem resolvidos. Todas as qualidades apresentadas e ilustradas através dos exemplos para a regressão biponderada não permitem, pelo menos de imediato, verificar a validade do modelo suposto como verdadeiro. O bom desempenho só vem a ser obtido no caso do modelo ajustado estar suficientemente próximo do modelo verdadeiro. Portanto somente se propõe o ajuste através da regressão biponderada após ter certeza de que o modelo a ser ajustado é, pelo menos, próximo do modelo verdadeiro.

O problema da convergência foi pouco citado além de no capítulo IV. Nos exemplos, partindo da estimativa inicial de mínimos

quadrados, não houve problemas, ratificando assim a certeza desta convergência, mesmo não se apresentando uma prova formal.

Quanto à constante de escalonamento, c, quanto menor for seu valor, mais robusto se torna o método de ajuste através da regressão biponderada. Por outro lado, quanto maior o valor de c, partindo de uma estimativa inicial de mínimos quadrados, maior será a semelhança entre os resultados obtidos com a regressão biponderada e com o método dos mínimos quadrados. Quando $c \to \infty$, para $S > 0$, tem-se que, em uma iteração o método converge e as estimativas obtidas são exatamente as mesmas obtidas por mínimos quadrados; isto é, desde que a estimativa inicial seja a obtida por mínimos quadrados.

Com a medida robusta de escala, S, praticamente não se desenvolveu nada. Mas, como foi dito, segundo MOSTELLER e TUKEY (1977) não deve haver muitas diferenças caso se utilize dois bons estimadores robustos para parâmetros de escala, Isso está ilustrado no exemplo da secção III.1.1, inclusive com o auxílio das curvas de influência. No entanto não se apresenta nenhuma prova deste fato.

Finalmente pode-se dizer que este trabalho deve ser encarado apenas como introdutório, restando vários problemas a serem resolvidos ou pelo menos estudados com maiores detalhes. Assim mesmo não há razões para desconfiar das vantagens apresentadas e ilustradas. Parece haver evidências de que o que falta a ser pesquisado (que é muito) exerce pequena influência nos resultados que se obtêm, não invalidando o que foi colocado como qualidade do ajuste através da regressão biponderada e mesmo do estimador biponderado, para parâmetros de locação.

VII - CONSTRUÇÃO DO PROGRAMA PARA COMPUTADORES

O propósito deste capítulo não é somente fornecer um manual para a utilização, sem grandes problemas, do programa para ajustes de regressão robustas, como também providenciar informações necessárias para aqueles que desejarem modificar e/ou complementar o programa. Com esta intenção se apresentará na secção VII.1 o que foi denominado fluxograma ilustrativo do programa. Na secção VII.3, dividida em várias subsecções serão apresentados os subprogramas componentes do programa principal, assim como também uma espécie de "Guia do Usuário", para cada sub programa. Na secção VII.2 se apresentará o modo correto para a colocação dos dados. No final do capítulo se apresentará a listagem completa do programa, incluindo os subprogramas e também o modo de se obter regressão biponderada através do SPSS.

O programa apresenta as seguintes características:

- impressão opcional dos dados
- opção por uma regressão passando ou não pela origem
- possibilidade de definição de novos suportes

- opção na escolha das estimativas iniciais ($\hat{B}_0$)
- impressão opcional das estimativas iniciais
- opção para ajuste de regressões robustas usando um dos dois métodos apresentados no capítulo IV (regressão biponderada ou passoponderada)
- impressão opcional dos resultados de cada iteração
- parada das iterações governada por dois critérios; convergência, de acordo com a secção IV.4.3 ou número máximo de iterações (NITER)
- impressão de dois gráficos, com as estimativas finais, o gráfico da Distribuição Acumulada dos Resíduos e o gráfico de Resíduos vs. Y ajustados
- opção quanto à medida robusta de escala a ser utilizada pelo método de ajuste.

O programa está dimensionado para um máximo de 200 observações e 30 suportes. Caso haja necessidade pode-se alterar as instruções DIMENSION e os valores de NMAX e NVARX na instrução DATA.

Há certas restrições quanto à precisão no cálculo das estimativas. Isto se deve ao fato de se estar utilizando um procedimento para a resolução do sistema AX=B, onde A é uma matriz (KxK) e X e B são vetores K-dimensionais, baseado no processo de diagonalização de Gauss-Jordan. É sabido que este processo não produz bons resultados, quando a matriz A é mal comportada. No entanto é bastante fácil substituir este processo por outro que produza resultados melhores. Na listagem do programa este processo aparece duas vezes. A primeira na parte referente à obtenção das estimativas iniciais, por mínimos quadrados, iniciando na instrução:

C RESOLUÇÃO DO SISTEMA XLX*B=XLY

e findando na instrução:

C FIM DA RESOLUÇÃO DO SISTEMA

O segundo aparecimento é no processo iterativo, iniciando na instrução:

C RESOLUÇÃO DO SISTEMA XLX*BM=XLY

e findando na instrução:

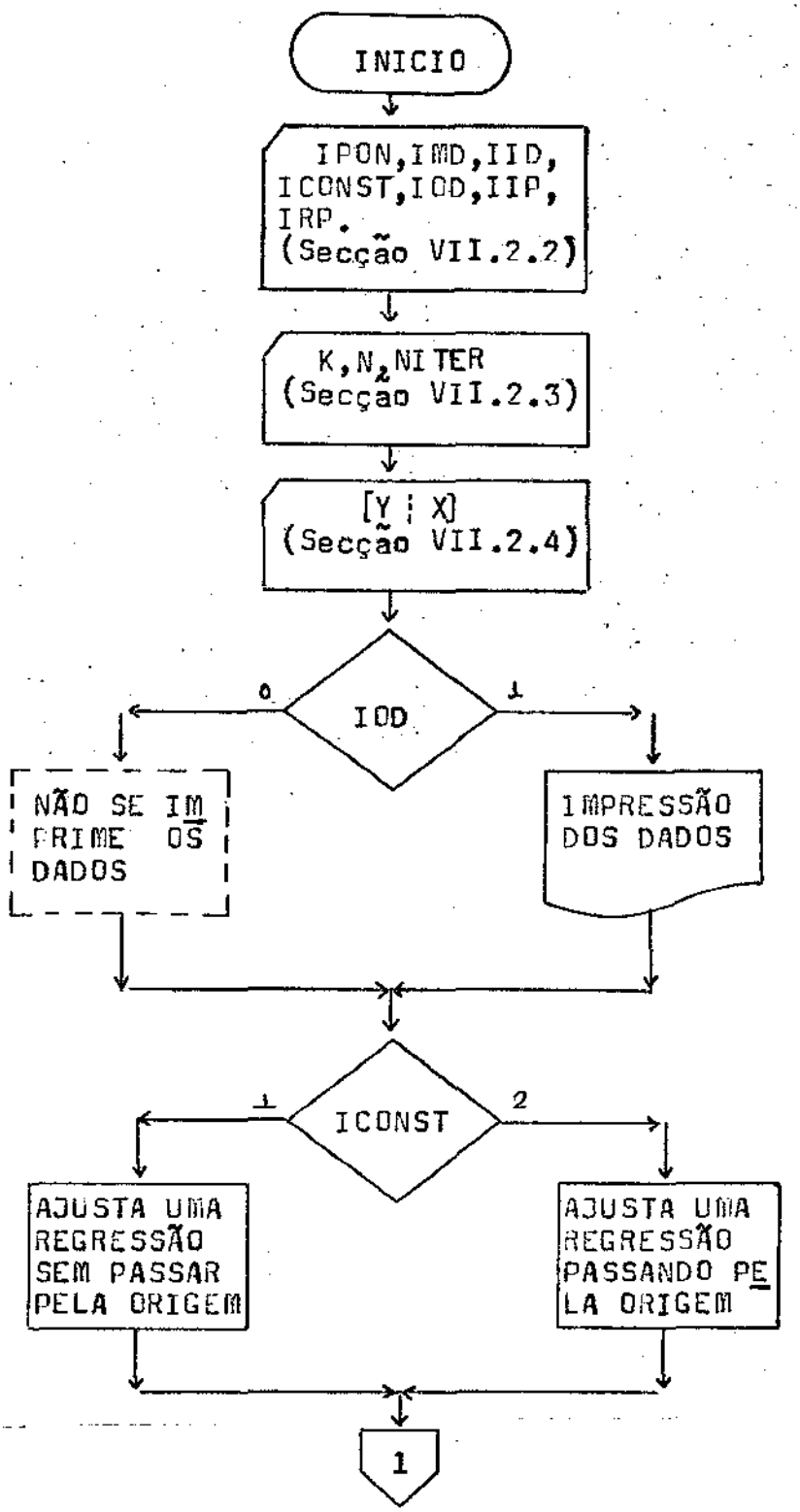
C FIM DA RESOLUÇÃO DO SISTEMA

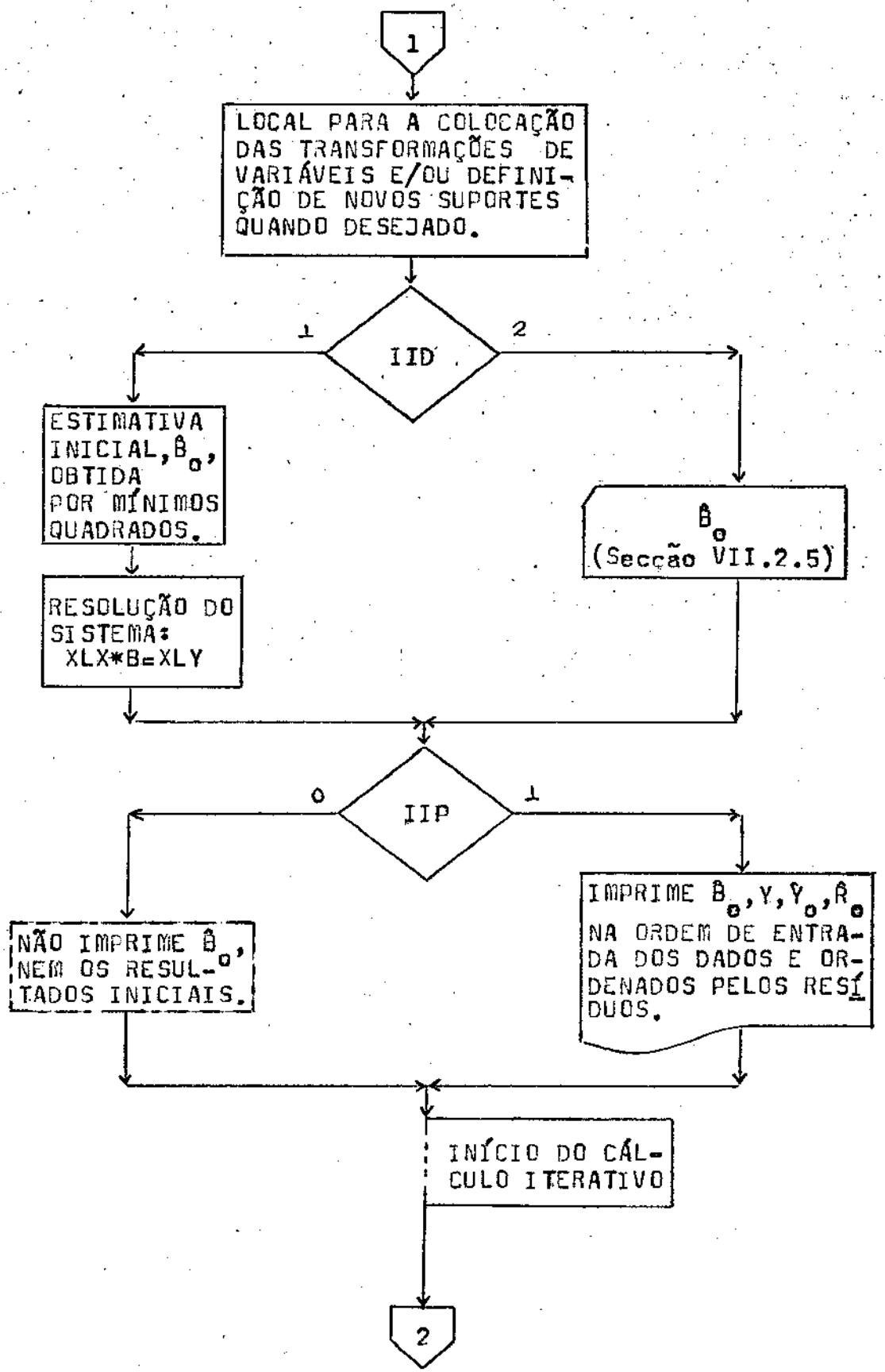
Portanto basta retirar as instruções contidas entre esses limites e substituir pelo processo de resolução de sistemas que se deseja. Pode-se utilizar qualquer número de comando entre 4000 e 4990 e também entre 8000 e 8990.

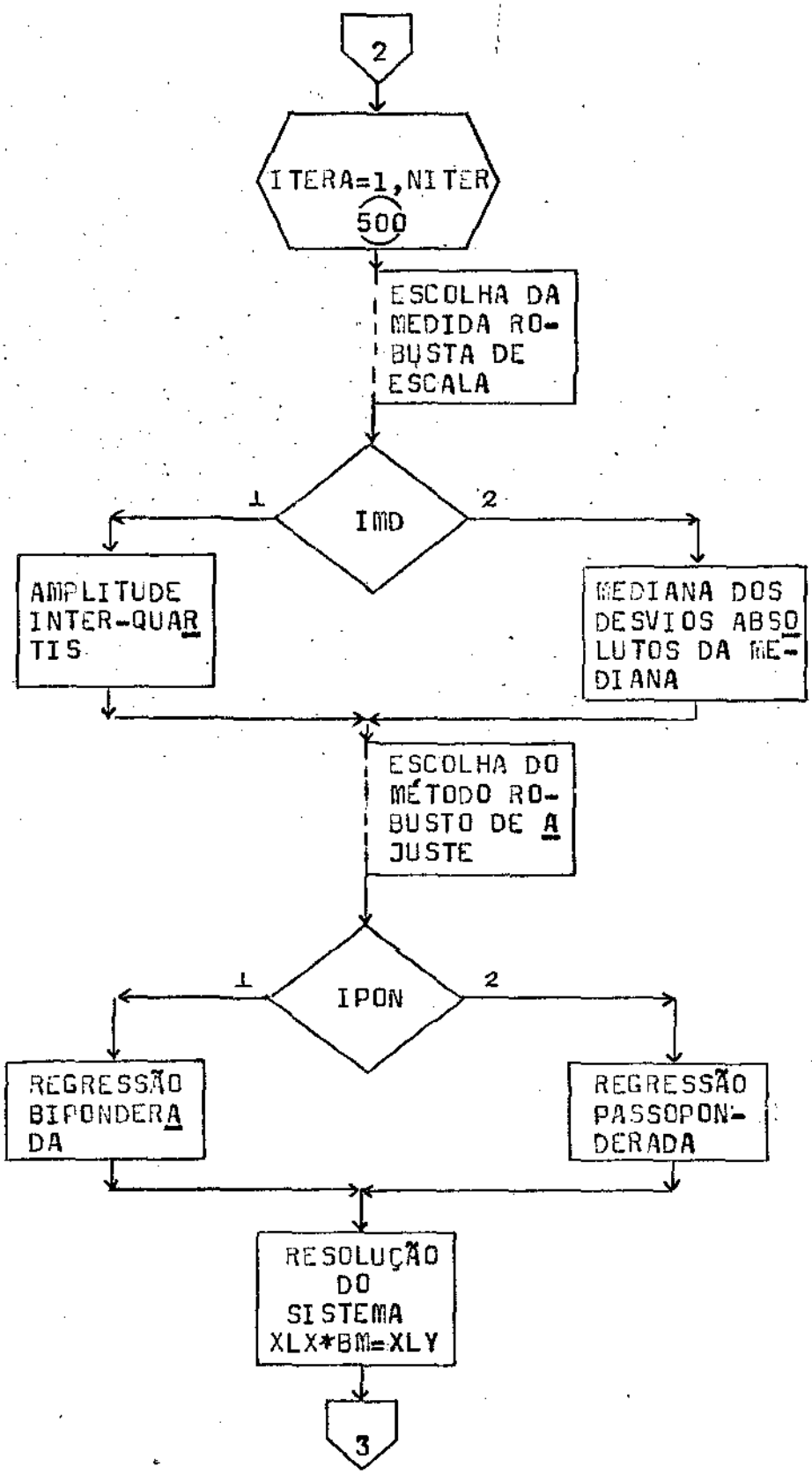
A sub rotina de impressão dos gráficos (sub rotina PLOT) foi construida para apresentar gráficos como os de DANIEL e WOOD (1971). Foi desenvolvida a partir da sub rotina que faz os gráficos no programa apresentado por WOOD (1976).

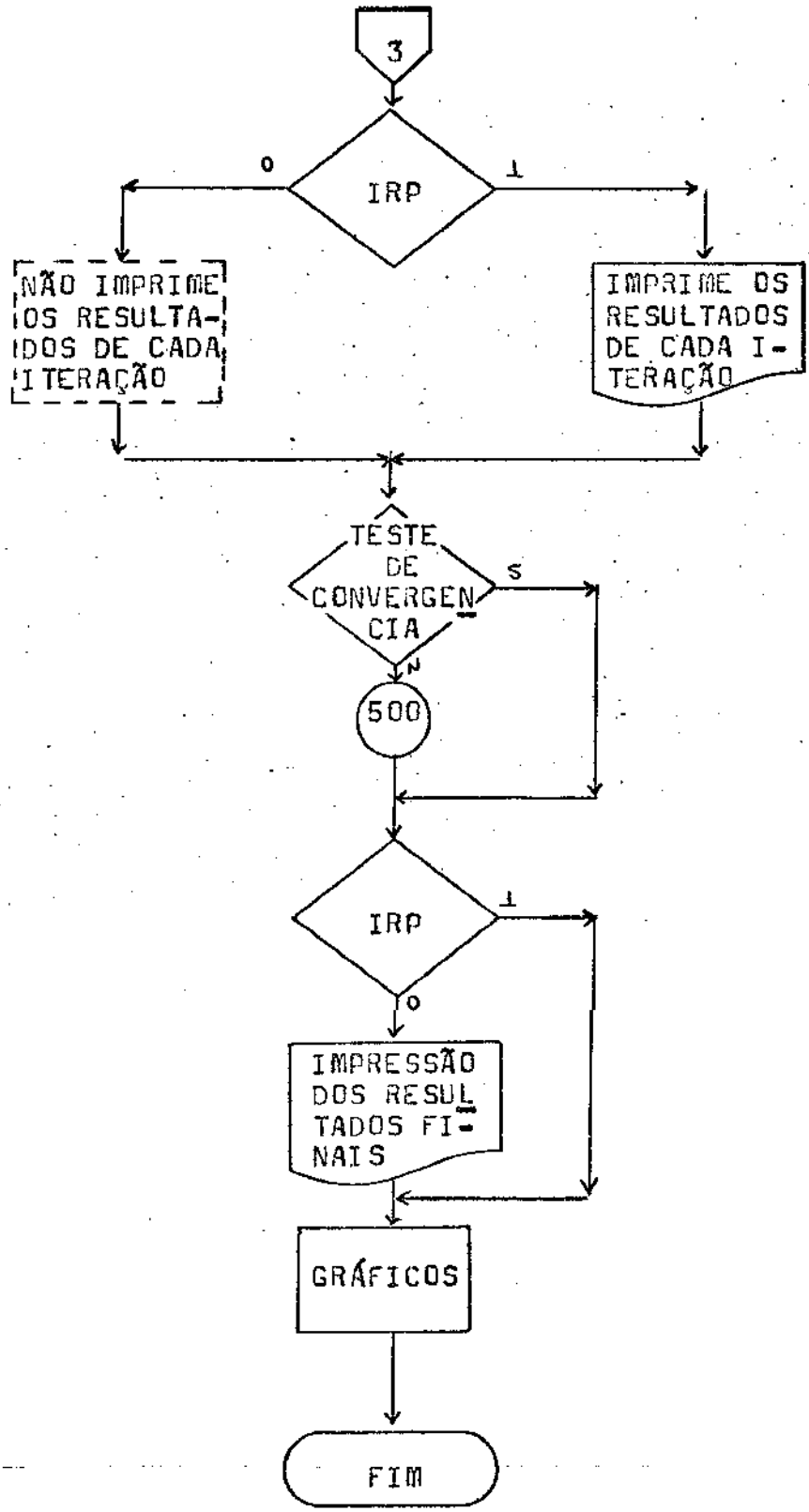
VII.1 - Fluxograma ilustrativo do programa

Apenas com o intuito de facilitar o entendimento do programa, assim como também para facilitar alguma possível modificação, apresenta-se o fluxograma seguinte:


INICIO
IPON,IMD,IID,
ICONST,IOD,IIP,
IRP.
(Secção VII.2.2)
K,N,NITER
(Secção VII.2.3)
[Y ¦ X]
(Secção VII.2.4)
IOD
0
1
NÃO SE IM
PRIME OS
DADOS
IMPRESSÃO
DOS DADOS
ICONST
1
2
AJUSTA UMA
REGRESSÃO
SEM PASSAR
PELA ORIGEM
AJUSTA UMA
REGRESSÃO
PASSANDO PE
LA ORIGEM
1


1
LOCAL PARA A COLOCAÇÃO DAS TRANSFORMAÇÕES DE VARIÁVEIS E/OU DEFINIÇÃO DE NOVOS SUPORTES QUANDO DESEJADO.
IID
1
2
ESTIMATIVA INICIAL, $\hat{B}_o$, OBTIDA POR MÍNIMOS QUADRADOS.
RESOLUÇÃO DO SISTEMA: XLX*B=XLY
$\hat{B}_o$ (Secção VII.2.5)
IIP
0
1
NÃO IMPRIME $\hat{B}_o$, NEM OS RESULTADOS INICIAIS.
IMPRIME $\hat{B}_o, Y, \hat{Y}_o, \hat{R}_o$ NA ORDEM DE ENTRADA DOS DADOS E ORDENADOS PELOS RESÍDUOS.
INÍCIO DO CÁLCULO ITERATIVO
2


2
ITERA=1,NITER
500
ESCOLHA DA MEDIDA ROBUSTA DE ESCALA
IMD
1
2
AMPLITUDE INTER-QUARTIS
MEDIANA DOS DESVIOS ABSOLUTOS DA MEDIANA
ESCOLHA DO MÉTODO ROBUSTO DE AJUSTE
IPON
1
2
REGRESSÃO BIPONDERADA
REGRESSÃO PASSOPONDERADA
RESOLUÇÃO DO SISTEMA XLX*BM=XLY
3


3
IRP
0
1
NÃO IMPRIME OS RESULTADOS DE CADA ITERAÇÃO
IMPRIME OS RESULTADOS DE CADA ITERAÇÃO
TESTE DE CONVERGENCIA
S
N
500
IRP
1
0
IMPRESSÃO DOS RESULTADOS FINAIS
GRÁFICOS
FIM

VII.2 - Colocação dos dados

Nesta secção se apresentará o modo para colocar os dados para o processamento do programa.

VII.2.1 - Instruções DATA

Nestas instruções estão os valores que comumente não serão modificados todas as vezes que se processar o programa. As instruções são as seguintes:

```
DATA  DELTA, EPSON, TOL, CUT, EAI/.1E-5,.1E-1,.1E-6,4.,1.35/
```

onde,

DELTA e EPSON - precisão para a convergência

CUT - constante de escalonamento, utilizada para o cálculo dos estimadores robustos (secções III.1.1, III.1.2, IV.2 e IV.3)

TOL - tolerância para os "zeros", na resolução do sistema. Se houver uma substituição do processo utilizado, provavelmente se tornará desnecessária.

EAI - esperança da amplitude interquartis. O valor 1,35 é o que se obtém no caso de distribuição N(0,1), como já foi dito na secção III.2.1.

A outra instrução DATA é referente ao número máximo de observações no conjunto de dados, NMAX e ao número máximo de suportes (ou variáveis, caso não haja transformações), NVARX. Foi colocada separadamente porque talvez seja mais vezes modificada, por problemas de dimensionamento no programa. A instrução é a seguinte:

```
      DATA  NMAX, NVARX/200,31/
```

Note que para aumentar a capacidade do programa no que se refere a um número maior de observações e/ou suportes, não basta alterar a instrução DATA acima. Há a necessidade de se alterar também as instruções DIMENSION.

VII.2.2 - Dados referentes às opções

Os primeiros valores a serem lidos serão os referentes às opções oferecidas pelo programa. O cartão que os contém deverá ser o primeiro a ser lido. Na instrução READ correspondente tem-se:

```
      READ (2,10) IPON, IMD, IID, ICONST, IOD, IIP, IRP
10    FORMAT(7I)
```

No cartão deve-se então colocar sete valores inteiros, separados entre si por um ou mais espaços em branco. Os valores a serem colocados dependendo das opções escolhidas, poderão ser os seguintes:

IPON - seleciona o método de ajuste a ser utilizado. Pode assumir os valores 1 ou 2. Se:

IPON = 1 - Regressão Biponderada

IPON = 2 - Regressão Passoponderada

IMD - seleciona qual a medida robusta de escala que será utilizada. Há duas opções, referenciadas pelos números 1 e 2. Se:

IMD = 1 - Amplitude interquartís dividida por EAI

IMD = 2 - Mediana dos desvios absolutos (da mediana) - MDA

IID - seleciona o modo de se obter a estimativa inicial dos parâmetros da regressão ($\hat{B}_0$). Há duas opções:

IID = 1 - $\hat{B}_0$ é obtido através do método dos mínimos quadrados

IID = 2 - $\hat{B}_0$ é lido de cartões, tendo sido obtido anteriormente através de meio externo ao programa. O modo de leitura de $\hat{B}_0$ e a posição do(s) cartão(ões) que o contêm serão explicados na secção VII.2.5.

ICONST - seleciona uma equação de regressão, a ser ajustada, passando ou não pela origem. Se:

ICONST = 0 - passa pela origem

ICONST = 1 - não passa pela origem ($X1 \equiv 1$)

As opções seguintes governam as impressões que o programa pode ou não vir à realizar

IOD - governa a impressão dos dados; isto é, do vetor de respostas e da matriz de delineamento. Ressalta-se que a matriz que será impressa é a matriz lida e não a matriz resultante das transformações para a obtenção de novos suportes. A matriz dos suportes, caso se construa, não é impressa pelo programa. Caso se deseje sua impressão deve-se providenciar as instruções necessárias, logo após a definição dos novos suportes. Se:

IOD = 0 - não imprime os dados

IOD = 1 - imprime os dados

IIP - governa a impressão das estimativas iniciais. Se:

IIP = 0 - não imprime $\hat{B}_0$, $\hat{Y}_0$ e $\hat{R}_0$

IIP = 1 - imprime $\hat{B}_0$ e também Y, $\hat{Y}_0$ e $\hat{R}_0$ tanto na ordem de entrada dos dados como também ordenados pelos resíduos. Imprime também o vetor que dá as posições dos elementos do vetor dos resíduos ordenados, no vetor dos resíduos não ordenados (o vetor KO).

IRP - governa a impressão dos resultados de cada iteração. Se:

IRP = 0 - não imprime os resultados de cada iteração

IRP = 1 - imprime os resultados de cada iteração, $\hat{B}_i$ e também Y, $\hat{Y}_i$ e $\hat{R}_i$ tanto na ordem de entrada dos dados, como também ordenados pelos resíduos. Imprime também o vetor KO, anteriormente citado.

VII.2.3 - Dados de dimensionamento

Devem ser colocados no segundo cartão de dados a ser lido. A instrução de leitura correspondente é:

```
      READ (2,20) N, K, NITER
20    FORMAT(3I)
```

Deve-se colocar então tres valores inteiros, separados entre si por pelo menos um espaço em branco. Tem-se que:

N - número de observações ($N \leq NMAX$)

K - quantidade de variáveis independentes ($K \leq NVARX$)

NITER - número máximo de passos para o processo iterativo.

VII.2.4 - Leitura do vetor de constantes e da matriz de delineamento

Esta leitura é feita conjuntamente. A perfuração destes dados é mais facilmente compreendida através de um exemplo. Suponha que se tem N observações e K variáveis independentes. Tem-se então a seguinte matriz, de dimensão (Nx(K+1)):

$$\begin{pmatrix} y_1 & x_{11} & x_{21} & \cdots & x_{K1} \\ y_2 & x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \vdots & & \vdots \\ y_N & x_{1N} & x_{2N} & \cdots & x_{KN} \end{pmatrix}$$

Se $K + 1 \leq 10$, ter-se-á N cartões, cada um com K+1 valores (reais), sendo o primeiro valor de cada cartão referente à variável Y e os K restantes às variáveis X_i; i=1,2,...,K.

Se $11 \leq K+1 \leq 20$, ter-se-á N+N=2N cartões, os primeiros N contendo 10 valores reais, sendo o primeiro valor de cada cartão referente à variável Y e os 9 restantes referentes às variáveis $X_1, X_2, \ldots, X_9$. Os outros (últimos) N cartões conterão, cada um, (K+1)-10=K-9 valores reais, referentes às variáveis $X_{10}, X_{11}, \ldots, X_K$, $K \leq 19$.

Se $21 \leq K+1 \leq 30$, ter-se-á 3N cartões e assim por diante; isto é, se $10J+1 \leq K+1 \leq 10(J+1)$, ter-se-á JxN cartões de dados, como descritos acima.

O formato em que estes dados devem estar nos cartões é o formato livre, formato F, havendo apenas a necessidade de se colocar um (ou mais) espaços em branco para separá-los uns dos outros.

VII.2.5 - Leitura de $\bar{B}_0$ (IID=2)

Caso se faça essa opção, o(s) cartão(ões) referente(s) a estes valores deve(m) ser colocado(s) logo após os dados referentes aos valores de Y e X_i; i=1,2,...,K. A instrução de leitura é a seguinte:

```
210   READ (2,270) (B(I), I=1,K)
270   FORMAT(31F)
```

e os valores $B_1, B_2, \ldots, B_K$ ($K \leq 31$) devem ser colocados em formato F, separados por um (ou mais) espaços em branco entre si.

VII.3 - Sub programas

Esta secção se destina a apresentar, de uma maneira bastante informal, a finalidade e o modo de utilização das diversas sub rotinas que fazem parte do programa. A listagem de cada uma delas está junto com o programa, na secção VII.4.

VII.3.1 - Sub rotina ORDEM

Ordena os elementos de um vetor, V, em ordem crescente ou decrescente (ICRES = 0 ou 1), em valor absoluto ou não (IVABS = 1 ou 2); dados o vetor V, N-dimensional. Na saída tem-se o vetor ordenado VO,não se perdendo o vetor original. É utilizada, direta ou indiretamente pelas sub rotinas AMPIQ, CMDN, XMAD, BIPON, STPON e PLOT, que serão apresentadas mais adiante. É utilizada do seguinte modo:

```
CALL ORDEM (N,V,VO,IVABS,ICRES,VOD)
```

VOD é um vetor N-dimensional, de utilidade interna à subrotina, não devendo ser modificado no CALL ORDEM.

VII.3.2 - Sub rotina ORRES

Ordena os elementos do vetor R, em ordem decrescente, formando o vetor RO. Constrói o vetor KO, que dá as posições dos elementos do vetor RO no vetor R. Ordena o vetor XB, formando o vetor XBO, segundo a ordenação de R em RO. O vetor YY é obtido fazendo:

```
YY = RO - XBO
```

Todos esses vetores são N-dimensionais. Na saida tem-se os vetores R, XB, KO, RO, XBO e YY. Esta sub rotina é chamada do seguinte modo:

CALL ORRES (N, R, XB, KO, RO, XBO, YY)

É utilizada para a obtenção dos resíduos, Y ajustados e Y observados ordenados pelos resíduos (RO, XBO e YY). É utilizada, indiretamente pela sub rotina PLOT, através do vetor LSORT, que é uma reordenação do vetor KO. Portanto caso se queira extrair do programa, para colocar em outro, a construção dos gráficos, haverá a necessidade de se extrair também esta sub rotina.

VII.3.3 - Sub rotina AMPIQ (IMD = 1)

É utilizada para se obter a medida robusta de escala definida na secção III.2.1:

$$S = \frac{\text{amplitude interquartís (dos resíduos)}}{EAI}$$

Dados o vetor V (vetor dos resíduos), N-dimensional, tem-se S=DIS. Utiliza as sub rotinas ORDEM e CMDN. É chamada como abaixo:

CALL AMPIQ (N, V, DIS, VO, EAI)

VO é um vetor utilizado internamente, não devendo ter sua denominação mudada no CALL AMPIQ. É utilizada, indiretamente, pelas sub rotinas BIPON e STPON.

VII.3.4 - Sub rotina CMDN

Calcula a mediana, XMDN, em um conjunto de dados na forma de um vetor ordenado (em ordem crescente ou decrescente), VO, N-dimensional. Como está no programa somente pode ser utilizada após uma chamada prévia da sub rotina ORDEM (ou ORRES). É utilizada do seguinte modo:

```
CALL CMDN (N, VO, XMDN)
```

É utilizada, direta ou indiretamente, pelas sub rotinas AMPIQ, XMAD, BIPON e STPON.

VII.3.5 - Sub rotina XMAD (IMD = 2)

Calcula a medida robusta de escala, MDA, definida na secção III.2.2:

$$MDA = \text{mediana } \{|x_i - x'| \ ; \ i = 1,2,\ldots,n\}$$

com $x' = \text{mediana } \{x_i\}$

Dados o vetor V e sua dimensão N, na saida tem-se MDA = DIS. Utiliza as sub rotinas ORDEM e CMDN. É utilizada como se segue:

```
CALL XMAD (N, V, DIS, D, VO)
```

D e VO são vetores utilizados internamente, não devendo ter suas denominações mudadas no CALL XMAD. É utilizada pelas sub rotinas BIPON e STPON.

VII.3.6 - Sub rotina BIPON (IPON = 1)

Constroi um vetor W que associa pesos de acordo com o estimador biponderado. Este vetor é tomado pelo programa como se fosse a matriz diagonal P, do capítulo IV. Utiliza-se de um vetor, em vez de uma matriz apenas para economizar espaço de máquina. V é um vetor, o vetor dos resíduos, CUT é a constante de escalonamento e DIS é a medida robusta de escala. Ambos os vetores, W e V, são N-dimensionais. Esta sub rotina é chamada como abaixo:

CALL BIPON (N, W, V, CUT, DIS, U)

O vetor U é de uso interno. Não deve sofrer mudanças na sua denominação no CALL BIPON. Esta sub rotina utiliza, direta ou indiretamente as sub rotinas ORDEM, AMPIQ, CMDN e XMAD.

VII.3.7 - Sub rotina STPON (IPON = 2)

Constrói um vetor W que associa pesos segundo o estimador passoponderado. Este vetor faz o papel da matriz P do capítulo IV. V é o vetor dos resíduos. Todos os vetores são N-dimensionais. LQP é o parâmetro que indica o número de ki's que se deseja; corresponde ao número m+1, da secção IV.3. T é a maior ponderação que se deseja associar; corresponde a k_1 da secção IV.3. CUT é a constante de escalonamento e DIS é a medida robusta de escala. O modo de utilização é o seguinte:

CALL STPON (N, W, V, LQP, T, CUT, DIS, A, SS, U)

Os vetores A, SS e U são de uso interno à sub rotina, não devendo sofrer mudanças em sua denominação no CALL STPON. Esta sub rotina utiliza, direta ou indiretamente, as sub rotinas ORDEM, AMPIQ, CMDN e XMAD.

VII.3.8 - Sub rotinas para a construção dos gráficos

Os gráficos são construidos pelas sub rotinas GRID, PACK e PLOT, que foram adaptadas a partir do existente em WOOD (1976). Estas tres sub rotinas possibilitam a construção do gráfico da distribuição acumulada dos resíduos e do gráfico de resíduos vs. Y ajustados. A função de cada uma delas é a seguinte:

GRID - faz a "bordadura" dos gráficos

PACK - coloca o caracter C no bit N da palavra X, onde X é uma palavra simulada do IBM 360, com os bites numerados da esquerda.

PLOT - chama as sub rotinas GRID e PACK. Monta e imprime os gráficos.

As sub rotinas PACK e GRID são chamadas internamente pela sub rotina PLOT, não necessitando de maiores explicações. Para chamar a sub rotina PLOT precisa-se do seguinte:

N - dimensão dos vetores YYCC (o vetor XB; Y ajustados), DELTA (vetor dos resíduos), e

LSORT - uma reordenação do vetor KO

YCMIN - menor elemento do vetor XB

YCMAX - maior elemento do vetor XB

GRIDA e GRIDB - contêm os gráficos e a "bordadura"

A sub rotina PLOT é chamada do seguinte modo:

```
CALL PLOT (N, YYCC, DELTA, LSORT, YCMIN, YCMAX, LSORT (NMAX + 1),
LSORT  (NMAX + 1379))
```

A construção dos gráficos necessita ainda, direta ou indiretamente das sub rotinas ORDEM e ORRES. Como última observação cita-se a necessidade da existência de um compilador FORTRAN extendido (F10) para que seja possível a utilização destas tres sub rotinas que constroem os gráficos. Na secção VII.4 há a maneira correta de executar o programa de modo a obter sem problemas não somente as estimativas dos parâmetros, como também os gráficos.

VII.4 - Programa para ajustes de regressão robusta

As instruções que serão dadas a seguir servirão para que seja possível uma utilização sem contratempos do programa de ajustes de regressão robusta, cuja listagem completa será apresentada a seguir. As instruções são as adequadas para o sistema implantado na UNICAMP, podendo sofrer modificações quando se desejar processar o programa em outros centros. Como já foi mencionado anteriormente há a necessidade de se ter nas instalações onde se pretender processar o programa um compilador FORTRAN extendido. As instruções serão dadas supondo que o usuário tenha acesso aos terminais ligados ao DEC-10 da UNICAMP.

Suponha que já se tenha o programa gravado na área do usuário, com o nome REGRO.F10. Basta então gravar um arquivo de dados, contendo não somente os valores de $Y, X_1,\ldots,X_k$, como também os valores

que selecionam as opções, de acordo com o que foi apresentado na secção VII.2. Há apenas duas restrições com relação a este arquivo:

- o nome principal de ter no máximo 3 caracteres, sendo o primeiro obrigatoriamente um caracter alfabético (letra entre A e Z)

- a extensão de ser CDR

Deste modo, um possível nome poderia ser, por exemplo:

DAD.CDR

outro poderia ser:

EX1.CDR

Com o arquivo de dados e programa já gravados pode-se processar o programa fazendo:

```
SET  CDR  DAD.CRR
EXEC/F10 REGRO.F10
```

no caso do arquivo de dados ser denominado DAD.CDR.

A seguir apresenta-se a listagem completa do programa REGRO.F10, utilizado para obter regressão robusta.

Em seguida apresenta-se a saida do computador com os dados do exemplo 2, no caso onde há 17 observações.

```
C ---   ********************
C
C       SCAFI,M.A.D.(1979) - PROGRAMA PARA AJUSTAR REGRESSAO RO-
C       BUSTA - FM : REGRESSAO BIPONDERADA - UM METODO ROBUSTO DE
C       AJUSTE. DISSERTACAO DE MESTRADO A SER APRESENTADA AO DE-
C       PARTAMENTO DE ESTATISTICA, IMECC-UNICAMP.
C
C ---   ********************
C
C            FAZ AJUSTES TANTO SEGUNDO A REGRESSAO BIPONDERADA
C       COMO TAMBEM SEGUNDO A REGRESSAO PASSOPONDERADA.
C            OFERECE VARIAS OPCOES QUANTO A IMPRESSAO, GOVERNADAS
C       PELAS VARIAVEIS IOD,IIP,IRP. OFERECE OUTRAS OPCOES DE A-
C       CORDO COM AS VARIAVEIS IPON,IMD,IID E ICONST.
C
C ---   ********************
C
C       IPON - OPCAO QUE SELECIONA O METODO ROBUSTO DE AJUSTE A
C              SER UTILIZADO.
C              =1 REGRESSAO BIPONDERADA
C              =2 REGRESSAO PASSOPONDERADA
C       IMD - OPCAO QUE DETERMINA A MEDIDA ROBUSTA DE ESCALA A
C             SER UTILIZADA.
C             =1 (AMPLITUDE INTERQUARTIS)/EAI
C             =2 MDA (MEDIANA DOS DESVIOS ABS. DA MEDIANA)
C       IID - OPCAO PARA A ESTIMATIVA INICIAL DOS PARAMETROS.
C             =1, ESTIMATIVA INICIAL OBTIDA POR MINIMOS QUADRADOS
C             =2, LE DO EM CARTOES
C       ICONST - AJUSTA REGRESSOES PASSANDO OU NAO PELA ORIGEM.
C                =1, NAO PASSA PELA ORIGEM
C                =0, PASSA PELA ORIGEM
C       IOD=1 OU 0 IMPRIME OU NAO OS DADOS.
C       IIP=1 OU 0 IMPRIME OU NAO AS ESTIMATIVAS INICIAIS.
C       IRP=1 OU 0 IMPRIME OU NAO OS RESULTADOS DE CADA ITERACAO.
C
C            O PROGRAMA IMPRIME SEMPRE OS RESULTADOS FINAIS (DES-
C       DE QUE O SISTEMA A SER RESOLVIDO TENHA SOLUCAO).
C            IMPRIME TAMBEM GRAFICO DE RESIDUOS VS. Y AJUSTADOS E
C       GRAFICO DA DISTRIBUICAO ACUMULADA DOS RESIDUOS.
C
C ---   ********************
C
        DIMENSION Y(200),X(200,31),R(200),XO(200),B(31)
        DIMENSION XAT(31,200),VO(200),SO(31),BM(31),RU(200)
        DIMENSION KO(200),LSORT(200),W(200),XBO(200),YY(200)
        DIMENSION XT(31,200),C(31,32),XLX(31,31),XLY(31)
C
        DATA DELTA,EPSON,TOL,CUT,EAI/.1E-2,.1E-1,.1E-6,4.,1.35/
C       DELTA E EPSON - PRECISAO PARA A CONVERGENCIA
C       TOL - TOLERANCIA NO CALCULO DA SOLUCAO DO SISTEMA
C       CUT - CONSTANTE DE ESCALONAMENTO
C       EAI - ESPERANCA DA AMPLITUDE INTERQUARTIS
C
        DATA NMAX,NVARX /200,31/
C       NMAX  - NUMERO MAXIMO DE OBSERVACOES
C       NVARX - NUMERO MAXIMO DE VARIAVEIS INDEPENDENTES, OU SUPORTES
C
C ---   ********************
C       LEITURA DAS OPCOES
        READ(2,10)IPON,IMD,IID,ICONST,IOD,IIP,IRP
```

```
10      FORMAT(7I)
        IF((IPON.LT.1).OR.(IPON.GT.2)) GO TO 2030
2035    IF((IMD.LT.1).OR.(IMD.GT.2)) GO TO 2040
2045    IF((IID.LT.1).OR.(IID.GT.2)) GO TO 2050
2055    IF((ICONST.LT.0).OR.(ICONST.GT.1)) GO TO 2060
2065    IF((IOD.LT.0).OR.(IOD.GT.1)) GO TO 2070
2075    IF((IIP.LT.0).OR.(IIP.GT.1)) GO TO 2080
2085    IF((IRP.LT.0).OR.(IRP.GT.1)) GO TO 2090
        IF(LERRO.EQ.0) GO TO 2100
        GO TO 280
2100    CONTINUE
C
C ---   **********************
C       LEITURA DOS PARAMETROS DE DIMENSIONAMENTO E NUMERO MAXIMO DE ITERA
C       COES
C       N - NUMERO DE OBSERVACOES
C       K - NUMERO DE VARIAVEIS INDEPENDENTES
C       NITER - NUMERO MAXIMO DE ITERACOES
280     READ(2,20)N,K,NITER
20      FORMAT(3I)
        IF((N.GT.NMAX).OR.(K.GE.NVARX)) GO TO 2010
C ---   **********************
C       LEITURA DOS DADOS
C       Y - VETOR DE RESPOSTAS OU VETOR DE CONSTANTES
C       X - MATRIZ DE DELINEAMENTO
        NINC=1
        NVAR=9
        IF(K.LT.NVAR) NVAR=K
        DO 30 I=1,N
30      READ(2,40)Y(I),(X(I,J),J=NINC,NVAR)
40      FORMAT(10F)
60      NREST=K-NVAR
        IF(NREST.LE.0) GO TO 70
        NINC=NVAR+1
        NVAR=NVAR+10
        IF(K.LT.NVAR) NVAR=K
        DO 50 I=1,N
50      READ(2,40)(X(I,J),J=NINC,NVAR)
        GO TO 60
C       FIM DA LEITURA DOS DADOS
C ---   **********************
70      ICONST=ICONST+1
        IOD=IOD+1
        IIP=IIP+1
        IRP=IRP+1
        IF(NITER.LE.0) IPON=3
        IF(ICONST.EQ.1) IPON=4
        GO TO (3000,3010,3050,3030),IPON
3000    WRITE(3,3020)
3020    FORMAT(30X,10('* '),'REGRESSAO PONDERADA ',10('* '),///)
        GO TO 3030
3010    WRITE(3,3040)
3040    FORMAT(28X,10('* '),'REGRESSAO PASSOPONDERADA ',10('* '),///)
        GO TO 3030
3050    WRITE(3,3060)
3060    FORMAT(15X,10('* '),'REGRESSAO ATRAVES DO METODO DOS MINIMOS QUADR
        1ADOS ',10('* '),///)
3030    CONTINUE
        GO TO (80,90),IOD
90      WRITE(3,800)
```

```
800     FORMAT(1X,119('-'))
        WRITE(3,810)
810     FORMAT(/,15X,5('* '),'D A D O S ',5('* '))
        WRITE(3,820)
820     FORMAT(//,2X,'N',3X,'VAR. DEP.',43X,'VARIAVEIS INDEPENDENTES',/,1X
        1,'----',2X,9('-'),4X,10(1X,9('-')))
        NINC=1
        NVAR=10
        IF(K.LT.NVAR) NVAR=K
        DO 100 I=1,N
100     WRITE(3,830)I,Y(I),(X(I,J),J=NINC,NVAR)
830     FORMAT(1X,I3,2X,F9.4,1X,10(1X,F9.4))
120     NREST=K-NVAR
        IF(NREST.LE.0) GO TO 85
        NINC=NVAR+1
        NVAR=NVAR+11
        IF(K.LT.NVAR) NVAR=K
        WRITE(3,825)
825     FORMAT(//,3X,'N',47X,'VARIAVEIS INDEPENDENTES',/,1X,'----',5X,11(1X,
        1,9('-')))
        DO 110 I=1,N
110     WRITE(3,830)I,(X(I,J),J=NINC,NVAR)
        GO TO 120
85      WRITE(3,890)
80      GO TO (130,140),ICONST
140     K=K+1
        DO 150 I=1,N
150     X(I,K)=1.
        L=K
        DO 160 J=1,K
        DO 170 I=1,N
        AUX=X(I,J)
        X(I,J)=X(I,K)
170     X(I,K)=AUX
        L=L-1
        IF(L.LT.2) GO TO 130
160     CONTINUE
130     CONTINUE
C
C ---   **********
C ---   **********
C
C           QUALQUER TRANSFORMACAO DE VARIAVEIS E/OU DEFINICAO  DE  SUPOR
C       TES DEVE SER COLOCADA NESTE PONTO.  NAO ESQUECER DE FAZER:
C       K = NUMERO 'EXATO' DE SUPORTES RESULTANTES
C       X = 'EXATAMENTE' A NOVA MATRIZ DOS SUPORTES
C
C ---   **********
C ---   **********
C
        IF(K.GT.NVARX) GO TO 2020
        IF(LERRO.GE.1) GO TO 5000
        DO 180 I=1,N
        DO 180 J=1,K
        XT(J,I)=X(I,J)
180     XXT(J,I)=X(I,J)
C ---   **********
C       INICIALIZACAO DOS PARAMETROS
        GO TO (200,210),IID
200     WRITE(3,790)
```

```
790     FORMAT(1X,////)
        WRITE(3,800)
C ---   INICIACAO POR MINIMOS QUADRADOS
        WRITE(3,840)
840     FORMAT(/,29X,5('*'),' PARAMETROS DE INICIALIZACAO OBTIDOS POR MINI
        1NOS QUADRADOS ',5('*'),/)
        DO 1210 L1=1,K
        XLY(L1)=0.
        DO 1220 L2=1,K
        XLX(L1,L2)=0.
        DO 1220 I=1,N
1220    XLX(L1,L2)=XT(L1,I)*X(I,L2)+XLX(L1,L2)
        DO 1210 I=1,N
1210    XLY(L1)=XT(L1,I)*Y(I)+XLY(L1)
C
C ---   **********
C       RESOLUCAO DO SISTEMA XLX*B=XLY
C               ADAPTADA A PARTIR DO EXISTENTE EM:
C       PACITTI,T.(1972) - FORTRAN-MONITOR PRINCIPIOS - AO LIVRO TECNI-
C       CO - RIO DE JANEIRO - SEGUNDA EDICAO
C ---   **********
        NX=K-1
        NY=K+1
        DO 4010 I=1,K
        DO 4010 J=1,K
4010    C(I,J)=XLX(I,J)
        DO 4020 I=1,K
4020    C(I,NY)=XLY(I)
        DO 4030 L=1,NX
        LX=L+1
        DO 4040 I=LX,K
        IF(ABS(C(L,L))-ABS(C(I,L)))4050,4040,4040
4050    DO 4040 JX=L,NY
        T=C(L,JX)
        C(L,JX)=C(I,JX)
        C(I,JX)=T
4040    CONTINUE
        PIV=C(L,L)
        DO 4060 JX=L,NY
4060    C(L,JX)=C(L,JX)/PIV
        DO 4030 I=LX,K
        M=0
        DIVA=C(I,L)
        DO 4030 J=L,NY
        C(I,J)=C(I,J)-DIVA*C(L,J)
        IF(J-NY)4070,4080,4080
4070    IF(ABS(C(I,J))-TOL)4030,4030,4090
4090    M=1
        GO TO 4030
4080    IF(M)4100,4100,4030
4100    IF(ABS(C(I,J))-TOL)4130,4130,4140
4030    CONTINUE
        C(K,NY)=C(K,NY)/C(K,K)
        C(K,K)=1.
        DO 4110 I=1,NX
        IX=I+1
        DO 4110 L=IX,K
        DIUB=C(I,L)
        DO 4110 J=L,NY
4110    C(I,J)=C(I,J)-DIUB*C(L,J)
```

```
      DO 4120 I=1,K
4120  B(I)=C(I,NY)
      LERRO=0
      GO TO 4150
4130  WRITE(3,4135)
4135  FORMAT(5X,5('*'),' SOLUCAO INDETERMINADA ',5('*'))
      WRITE(3,800)
      LERRO=1
      GO TO 4150
4140  WRITE(3,4145)
4145  FORMAT(5X,5('*'),' SOLUCAO IMPOSSIVEL ',5('*'))
      WRITE(3,800)
      LERRO=1
4150  CONTINUE
C     FIM DA RESOLUCAO DO SISTEMA
C --- **********************
C
      IF(LERRO.GT.0) GO TO 5000
      DO 225 L1=1,N
      XB(L1)=0.
      DO 225 I=1,K
225   XB(L1)=X(L1,I)*B(I)+XB(L1)
      DO 330 I=1,N
330   R(I)=Y(I)-XB(I)
      GO TO (245,230),IIP
245   WRITE(3,800)
      GO TO 220
230   WRITE(3,850)
850   FORMAT(5X,5('*'),' ESTIMATIVAS INICIAIS ',5('*'))
      GO TO (290,300),ICONST
300   WRITE(3,950)B(1)
950   FORMAT(5X,'CONSTANTE=',F14.8)
      DO 310 I=2,K
      JJ=I-1
310   WRITE(3,860)JJ,B(I)
      GO TO 320
290   DO 240 I=1,K
240   WRITE(3,860)I,B(I)
860   FORMAT(5X,'B(',I2,')= ',3X,F14.8)
C     ++++++++++++++++++++++++++++++++
320   CALL ORRAS(N,R,XP,KO,RO,XBO,YY)
C     ++++++++++++++++++++++++++++++++
      WRITE(3,870)
870   FORMAT(/,1X,20('-'),'ORDEM DE ENTRADA',20('-'),6X,16('-'),'ORDENAD
      1OS PELOS RESIDUOS',16('-'),/,3X,'OBS.',8X,'YOBS',12X,'YEST',14X,'
      2RESIDUOS',8X,'OBS.',9X,'YOBS',12X,'YEST',14X,'RESIDUOS')
      DO 250 I=1,N
250   WRITE(3,880)I,Y(I),XB(I),R(I),KO(I),YY(I),XBO(I),RO(I)
880   FORMAT(3X,I4,2(1X,F),3X,F,8X,I4,2(1X,F),3X,F)
      WRITE(3,800)
      GO TO 220
C --- LE BO, CASO SEJA ESTA A OPCAO, DE CARTOES DE DADOS
210   READ(2,270)(B(I),I=1,K)
270   FORMAT(3IF)
      GO TO (220,235),IIP
235   WRITE(3,790)
      WRITE(3,800)
      WRITE(3,851)
851   FORMAT(/,35X,5('*'),' PARAMETROS DE INICIALIZACAO LIDOS EM CARTOES
      1 ',5('*'),/)
```

```
      GO TO 4150
C     FIM DA INICIALIZACAO DOS PARAMETROS
C ---  ********************
C
220   CONTINUE
      IF(NITER.EQ.0) GO TO 1500
C
C ---  ********************
C     INICIO DO PROCESSO ITERATIVO
C ---  **********.*******
      WRITE(3,790)
      WRITE(3,970)
970   FORMAT(///,32X,10('* '),'PROCESSO ITERATIVO ',10('* '))
      IF(IMD-1)5000,430,440
430   WRITE(3,980)CUT,EAI,DELTA
980   FORMAT(/,1X,'CONSTANTE DE ESCALONAMENTO',5X,F,/,1X,'MEDIDA ROBUSTA
      1 DE ESCALA',6X,'AMPLITUDE INTERQUARTIS/',F,/,1X,'PRECISAO PARA A
      2CONVERGENCIA ',F)
      GO TO 1070
440   WRITE(3,1060)CUT,DELTA
1060  FORMAT(/,1X,'CONSTANTE DE ESCALONAMENTO',5X,F,/,1X,'MEDIDA ROBUSTA
      1 DE ESCALA',6X,'MEDIANA DOS DESVIOS ABSOLUTOS DA MEDIANA',/,1X,'P
      2RECISAO PARA A CONVERGENCIA ',F)
1070  CONTINUE
      DO 500 ITERA=1,NITER
      GO TO (400,410),IMD
400   CALL AMPIO(N,R,DIS,VO,EAI)
      GO TO 420
410   CALL XMAD(N,F,DIS,D,VO)
420   GO TO (510,520),IPON
C ---  REGRESSAO RIPONDERADA (BIWEIGHT)
C     ++++++++++++++++++++++++++
510   CALL BIPON(N,M,R,CUT,DIS,U)
C     ++++++++++++++++++++++++++
      GO TO 530
C ---  REGRESSAO PASSOPONDERADA (STEPWEIGHT)
C     ++++++++++++++++++++++++++++++++++++
520   CALL STPON(N,M,R,A,A,,CUT,DIS,A,SS,U)
C     ++++++++++++++++++++++++++++++++++++
530   DO 535 I=1,K
      DO 535 J=1,N
535   XXT(I,J)=XT(I,J)*W(J)
      DO 1230 L1=1,K
      XLY(L1)=0.
      DO 1240 L2=1,K
      XLX(L1,L2)=0.
      DO 1240 I=1,N
1240  XLX(L1,L2)=XXT(L1,I)*X(I,L2)+XLX(L1,L2)
      DO 1230 I=1,N
1230  XLY(L1)=XXT(L1,I)*Y(I)+XLY(L1)
C
C ---  **********************
C     RESOLUCAO DO SISTEMA XLX*BM=XLY
C ---  **********.*****.*****
      DO 8010 I=1,K
      DO 8010 J=1,K
8010  C(I,J)=XLX(I,J)
      DO 8020 I=1,K
8020  C(I,NX)=XLY(I)
      K11=K+1
```

```
      DO 8030 L=1,NX
      LX=L+1
      DO 8040 I=LX,K
      IF(ABS(C(L,L))-ABS(C(I,L)))8050,8040,8040
8050  DO 8040 JX=L,NY
      T=C(L,JX)
      C(L,JX)=C(I,JX)
      C(I,JX)=T
8040  CONTINUE
      PIV=C(L,L)
      DO 8060 JX=L,NY
8060  C(L,JX)=C(L,JX)/PIV
      DO 8030 I=LX,K
      M=0
      DIVA=C(I,L)
      DO 8030 J=L,NY
      C(I,J)=C(I,J)-DIVA*C(L,J)
      IF(J-NY)8070,8080,8080
8070  IF(ABS(C(I,J))-TOL)8030,8030,8090
8090  M=1
      GO TO 8030
8080  IF(M)8100,8100,8030
8100  IF(ABS(C(I,J))-TOL)8130,8130,8140
8030  CONTINUE
      C(K,NY)=C(K,NY)/C(K,K)
      C(K,K)=1.
      DO 8110 I=1,NX
      IX=I+1
      DO 8110 L=IX,K
      DIUB=C(I,L)
      DO 8110 J=L,NY
8110  C(I,J)=C(I,J)-DIUB*C(L,J)
      DO 8120 I=1,K
8120  BM(I)=C(I,NX)
      LERRO=0
      GO TO 8150
8130  WRITE(3,8135)ITERA
8135  FORMAT(5X,5('*'),' ITERACAO NUMERO ',I4,1X,5('*'))
      WRITE(3,4135)
      WRITE(3,800)
      LERRO=1
      GO TO 8150
8140  WRITE(3,8135)ITERA
      WRITE(3,4145)
      WRITE(3,800)
      LERRO=1
8150  CONTINUE
C     FIM DA RESOLUCAO DO SISTEMA
C --- **********************
C
      IF(LERRO.GT.0) GO TO 5000
      DO 540 L1=1,N
      XB(L1)=0.
      DO 540 I=1,K
540   XB(L1)=A(L1,I)*BM(I)+XB(L1)
      DO 760 L1=1,N
760   R(L1)=Y(L1)-XB(L1)
      GO TO (560,560),IRP
560   WRITE(3,790)
      WRITE(3,800)
```

```
      WRITE(3,890)ITERA
890   FORMAT(/,5X,5('*'),' ITERACAO NUMERO ',I3,1X,5('*'))
      GO TO (700,710),ICONST
710   WRITE(3,950)BM(1)
      DO 720 I=2,K
      JJ=I-1
720   WRITE(3,860)JJ,BM(I)
      GO TO 730
700   DO 570 I=1,K
570   WRITE(3,860)I,BM(I)
730   WRITE(3,870)
C     ++++++++++++++++++++++++++++++
      CALL ORRES(N,R,XB,KO,BO,XBO,YY)
C     ++++++++++++++++++++++++++++++
      DO 580 I=1,N
580   WRITE(3,880)I,Y(I),XB(I),R(I),KO(I),YY(I),XBO(I),BO(I)
      WRITE(3,800)
550   CONTINUE
C ---  TESTE DE CONVERGENCIA
      DO 590 I=1,K
      IF(B(I).LT.EPSON) GO TO 600
      CONV=ABS((BM(I)-B(I))/B(I))
      GO TO 610
600   CONV=ABS(BM(I)-B(I))
610   IF(CONV.GT.DELTA) GO TO 620
590   CONTINUE
      GO TO 1000
620   NIT=NITER-ITERA
      IF(NIT.EQ.1) GO TO 630
      GO TO 640
630   DO 650 I=1,K
650   SO(I)=BM(I)
640   DO 660 I=1,K
660   B(I)=BM(I)
500   CONTINUE
C ---  **********************
C     FIM DO PROCESSO ITERATIVO
C ---  **********************
C
      GO TO (665,1200),IRP
665   WRITE(3,790)
      WRITE(3,800)
      WRITE(3,900)NITER
900   FORMAT(/,1X,5('*'),6X,'O PROCESSO ITERATIVO NAO CONVERGIU EM ',I2,
     1' ITERACOES.  VAI-SE IMPRIMIR O RESULTADO DAS DUAS ULTIMAS ITERAC
     1OES',1X,5('*'),/)
      NIT=NITER-1
      WRITE(3,910)NIT,NITER
910   FORMAT(11X,2(7X,'ITERACAO ',I2))
      GO TO (740,750),ICONST
750   WRITE(3,960)SO(1),B(1)
960   FORMAT(5X,'CONSTANTE=',F14.8,4X,F14.8)
      DO 760 I=2,K
      JJ=I-1
760   WRITE(3,920)JJ,SO(I),B(I)
      GO TO 770
740   DO 670 I=1,K
670   WRITE(3,920)I,SO(I),B(I)
920   FORMAT(5X,'B(',I2,')= ',3X,F14.8,4X,F14.8)
C     ++++++++++++++++++++++++++++++
```

```
770     CALL ORRFS(N,R,XR,KO,RO,XBO,YY)
C       ++++++++++++++++++++++++++++++
        WRITE(3,870)
        DO 666 I=1,N
666     WRITE(3,889)I,Y(I),XB(I),R(I),KO(I),YY(I),XBO(I),RO(I)
        WRITE(3,800)
        GO TO 1500
1200    WRITE(3,790)
        WRITE(3,800)
        WRITE(3,930)NITER
930     FORMAT(/,1X,20('*'),5X,'O PROCESSO ITERATIVO NAO CONVERGIU EM ',I2
        1,' ITERACOES',5X,20('*'),/)
        WRITE(3,800)
        GO TO 1500
1000    WRITE(3,790)
        WRITE(3,800)
        WRITE(3,940)ITERA
940     FORMAT(/,1X,20('*'),5X,'O PROCESSO ITERATIVO CONVERGIU EM',I3,'.I
        1TERACOES',5X,20('*'))
        GO TO (1010,1400),IRP
1010    GO TO (1030,1020),ICONST
1020    WRITE(3,950)BM(1)
        DO 1040 I=2,K
        JJ=I-1
1040    WRITE(3,860)JJ,BM(I)
        GO TO 1050
1030    DO 680 I=1,K
680     WRITE(3,860)I,BM(I)
1050    WRITE(3,870)
C       ++++++++++++++++++++++++++++++
        CALL ORRPS(N,R,XR,KO,RO,XBO,YY)
C       ++++++++++++++++++++++++++++++
        DO 690 I=1,N
690     WRITE(3,889)I,Y(I),XB(I),R(I),KO(I),YY(I),XBO(I),RO(I)
1400    WRITE(3,800)
1500    CONTINUE
C
C ---   **********
C
C       GRAFICOS
C
C ---   **********
C       ++++++++++++++++++++++++++
        CALL ORDER(N,XB,VO,2,1,VOO)
C       ++++++++++++++++++++++++++
        YCMIN=VO(1)
        YCMAX=VO(N)
        N1=N+1
        DO 1700 IA=1,N
        IC=N1-IA
        JC=KO(IC)
1700    LSORT(IA)=JC
C       ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
        CALL PLOT(N,XB,R,LSORT,YCMIN,YCMAX,LSORT(NMAX+1),LSORT(NMAX+1379))
C       ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
C
        GO TO 5000
C ---   **********
C       IMPRESSAO DOS ERROS
C ---   **********
```

```
2010    WRITE(3,2005)
2005    FORMAT(1X,20('*'),'ERRO',20('*'),/,1X,'HA MAIS DE NMAX OBSERVACOES.
        1 OU MAIS DE NVARX VARIAVEIS INDEPENDENTES. ALTERE O DIMENSION, NM
        1AX E NVARX',/,44('*'))
        GO TO 5000
2020    WRITE(3,2015)
2015    FORMAT(1X,20('*'),'ERRO',20('*'),/,1X,'APOS A DEFINICAO DE NOVAS
        1VARIAVEIS E/OU SUPORTES, FICOU-SE COM MAIS DE NVARX SUPORTES. ALT
        1ERE O DIMENSION E NVARX',/,1X,44('*'))
        GO TO 5000
2030    WRITE(3,2031)
2031    FORMAT(1X,5('*'),' ERRO ',5('*'),' IPON SO PODE SER 1 OU 2')
        LERRO =LERRO+1
        GO TO 2035
2040    WRITE(3,2041)
2041    FORMAT(1X,5('*'),' ERRO ',5('*'),' IMD SO PODE SER 1 OU 2')
        LERRO =LERRO+1
        GO TO 2045
2050    WRITE(3,2051)
2051    FORMAT(1X,5('*'),' ERRO ',5('*'),' IID SO PODE SER 1 OU 2')
        LERRO=LERRO+1
        GO TO 2055
2060    WRITE(3,2061)
2061    FORMAT(1X,5('*'),' ERRO ',5('*'),' ICONST SO PODE SER 0 OU 1')
        LERRO=LERRO+1
        GO TO 2065
2070    WRITE(3,2071)
2071    FORMAT(1X,5('*'),' ERRO ',5('*'),' IOD SO PODE SER 0 OU 1')
        LERRO =LERRO+1
        GO TO 2075
2080    WRITE(3,2081)
2081    FORMAT(1X,5('*'),' ERRO ',5('*'),' IIP SO PODE SER 0 OU 1')
        LERRO=LERRO+1
        GO TO 2085
2090    WRITE(3,2091)
2091    FORMAT(1X,5('*'),' ERRO ',5('*'),' IRP SO PODE SER 0 OU 1')
        LERRO=LERRO+1
        GO TO 2100
5000    CONTINUE
        STOP
        END
C
C ---   ********************
C       SUBROTINA ORDEM
C ---   *'**'*****'*'********
C
C               ORDENA OS ELEMENTOS DE UM VETOR, V, EM ORDEM CRESCENTE
C       OU DECRESCENTE ( ICRES = 0  OU  1 ), EM VALOR ABSOLUTO OU  NAO
C       ( IVABS = 1  OU  2 ).  DADOS O VETOR V E SUA DIMENSAO, N.  NA
C       SAIDA TEM-SE O VETOR ORDENADO VO, NAO SE PERDENDO O  VETOR O-
C       RIGINAL.
C
        SUBROUTINE ORDEM(N,V,VO,IVABS,ICRES,VOD)
        DIMENSION V(N),VO(N),VOD(N)
        K=0
        GO TO (60,70),IVABS
60      DO 10 I=1,N
10      VO(I)=ABS(V(I))
        GO TO 40
70      DO 09 I=1,N
```

```
80      VO(I)=V(I)
40      K=K+1
        Z=VO(K)
        DO 20 I=K,N
        IF(Z.LT.VO(I)) GO TO 20
        Z=VO(I)
        KK=I
20      CONTINUE
        A=VO(K)
        VO(K)=Z
        VO(KK)=A
        IF(K.LT.N) GO TO 40
        IF(ICRES.EQ.1) RETURN
        KKK=N
        DO 50 I=1,N
        VOO(KKK)=VO(I)
50      KKK=KKK-1
        DO 90 I=1,N
90      VO(I)=VOO(I)
        RETURN
        END
C
C ---   **********************
C       SUB ROTINA ORRES
C ---   **********************
C
C               ORDENA OS ELEMENTOS DO VETOR R EM ORDEM DECRESCENTE,
C       FORMANDO O VETOR RO.
C               CONSTROI O VETOR KO, QUE DA AS POSICOES DOS ELEMEN-
C       TOS DO VETOR RO NO VETOR R.  ORDENA O VETOR XB, ORMANDO  O
C       VETOR XBO, SEGUNDO A ORDENACAO DE R EM RO.  CONSTROI O VETOR
C       YY=RO-XBO.
C
        SUBROUTINE ORRES(N,R,XB,KO,RO,XBO,YY)
        DIMENSION R(N),XB(N),KO(N),XBO(N),YY(N),RO(N)
        DO 40 I=1,N
        RO(I)=R(I)
40      XBO(I)=XB(I)
        K=0
30      K=K+1
        Z=RO(K)
        DO 10 I=K,N
        IF(Z.GT.RO(I)) GO TO 10
        Z=RO(I)
        KK=I
10      CONTINUE
        A=RO(K)
        RO(K)=Z
        RO(KK)=A
        B=XBO(K)
        XBO(K)=XBO(KK)
        XBO(KK)=B
        IF(K.LT.N) GO TO 30
        DO 50 I=1,N
50      YY(I)=RO(I)+XBO(I)
        DO 60 I=1,N
        DO 70 J=1,N
        IF(RO(I)-R(J))70,80,70
70      CONTINUE
80      KO(I)=J
```

```
60      CONTINUE
        RETURN
        END
C
C ---   **********
C       SUB ROTINA AMPIQ
C ---   *****************
C
C               CALCULA A AMPLITUDE INTERQUARTIS DE UM CONJUNTO DE
C       DADOS NA FORMA DE UM VETOR, V. CHAMA AS SUB ROTINAS ORDEM
C       E CMDN.
C       APOS CALCULAR A AMPLITUDE INTERQUARTIS, AI, DIVIDE-SE ESTA
C       AMPLITUDE POR EAI, QUE E O SEU VALOR ESPERADO SOB AS SUPO-
C       SICOES. NO CASO PARTICULAR DE N(0,1), EAI=1,35.
C
        SUBROUTINE AMPIQ(N,V,DIS,VO,EAI)
        DIMENSION V(N),VO(N)
        CALL ORDEM(N,V,VO,2,1,VOD)
        M4=N/4
        M1=4*M4
        IF(M1.EQ.N) GO  TO 10
        H1=VO(M4)
        MM=(N+3)/4
        MM=N+1-MM
        H2=VO(MM)
        GO TO 20
10      M2=M4+1
        H1=(VO(M4)+VO(M2))/2.
        M3=N+1-M4
        M5=N-M4
        H2=(VO(M3)+VO(M5))/2.
20      AI=H2-H1
        DIS=AI/EAI
        RETURN
        END
C
C ---   **********
C       SUB ROTINA CMDN
C ---   **********
C
C               CALCULA A MEDIANA, XMDN DOS ELEMENTOS DE UM VETOR, VO
C       ORDENADO. SOMENTE PODE SER UTILIZADA APOS UMA  CHAMADA PREVIA
C       DA SUB ROTINA ORDEM.
C
        SUBROUTINE CMDN(N,VO,XMDN)
        DIMENSION VO(N)
        N1=N/2
        N2=N-(2*N1)
        IF(N2)30,10,20
10      N3=N1+1
        XMDN=(VO(N1)+VO(N3))/2.
        GO TO 30
20      N3=(N+1)/2
        XMDN=VO(N3)
30      RETURN
        END
C
C ---   **********
C       SUB ROTINA XMAD
C ---   **********
```

```
C
C               CALCULA A MEDIANA DOS DESVIOS ABSOLUTOS DA MEDIANA,
C       UTILIZADA COMO MEDIDA DE ESCALA.  CHAMA AS SUB ROTINAS  OR-
C       DEM E CMDN.
C
        SUBROUTINE XMAD(N,V,DIS,D,VO)
        DIMENSION V(N),D(N),VO(N)
        CALL ORDEV(N,V,VO,2,1,VOD)
        CALL CMDN(N,VO,XMDN)
        DO 10 I=1,N
10      D(I)=V(I)-XMDN
        CALL ORDEM(N,D,VO,1,1,VOD)
        CALL CMDN(N,VO,XM)
        DIS=XM
        RETURN
        END
C
C ---   **********************
C       SUB ROTINA BIPON
C ---   *********************
C
C               CALCULA UM VETOR W, UTILIZADO COMO SE FOSSE UMA MATRIZ
C       DIAGONAL, QUE ASSOCIA PESOS SEGUNDO O ESTIMADOR BIPONDERADO.
C               V E O VETOR DOS RESIDUOS, DE DIMENSAO NELE.  CUT  E  O
C       PONTO DE CORTE, DEFINIDO QUANDO SE APRESENTA O ESTIMADOR BIPON
C       DERADO (CUT = CONSTANTE DE ESCALONAMENTO).
C
        SUBROUTINE BIPON(NELE,W,V,CUT,DIS,U)
        DIMENSION V(NELE),W(NELE),U(NELE)
        CSK=CUT*DIS
        DO 10 I=1,NELE
        U(I)=V(I)/CSK
        AU=ABS(U(I))
        IF(AU.GT.1.) GO TO 50
        W(I)=(1.-(U(I)**2))**2
        GO TO 10
50      W(I)=0.
10      CONTINUE
        RETURN
        END
C
C ---   **********************
C       SUB ROTINA STPON
C ---   *********************
C
C               MONTA UM VETOR, W, USADO COMO SE FOSSE UMA MATRIZ DIA-
C       GONAL, QUE ASSOCIA PESOS SEGUNDO O ESTIMADOR  PASSOPONDERADO.
C               V E O VETOR DOS RESIDUOS DE DIMENSAO N.
C               CUT E A CONSTANTE DE ESCALONAMENTO OU PONTO DE CORTE.
C               LOP E O PARAMETRO QUE INDICA O NUMERO DE KI'S  QUE  SE
C       DESEJA ASSINALAR ( SECCAO I).6.2 ), LOP.GE.2.
C               T E A MAIOR PONDERACAO QUE SE DESEJA ASSINALAR(T=K1).
C
        SUBROUTINE STPON(N,W,V,LOP,T,CUT,DIS,A,SS,U)
        DIMENSION V(N),W(N),A(LOP),SS(LOP),U(N)
        A(1)=FLOAT(LOP)
        A(1)=1./A(1)
        DO 10 I=2,LOP
10      A(I)=FLOAT(I)*A(1)
        M=LOP-1
```

```
      SS(1)=T
      DO 20 I=2,N
      X=FLOAT(N-I+1)/FLOAT(N)
20    SS(I)=X*T
      SS(LOP)=0.
      CSK=CUT*DIS
      DO 60 I=1,N
      U(I)=V(I)/CSK
      DO 80 K=1,LOP
      IF(ABS(U(I)).GT.A(K)) GO TO 80
      W(I)=SS(K)
      GO TO 60
80    CONTINUE
60    CONTINUE
      RETURN
      END
C
C ---  **********
C      SUB ROTINA GRID
C ---  *********
C
C            COLOCA A "BORDADURA" NOS GRAFICOS.
C            OBTIDA EM:
C      WOOD,F.S.(1976) - NONLINEAR LEAST-SQUARES CURVE FITTING
C      PROGRAM - DECUS PROGRAM LIBRARY.
C
      SUBROUTINE GRID(GRIDA,GRIDB,L1)
      DIMENSION GRIDA(53,26),GRIDB(53,26),GRDA(7),GRDB(4)
      DATA GRDA/4H----,4H-*--,4H--*-,4H---*,4H----,2H*-,2H--/
      DATA GRDB/4HI   ,2H I,4HI---,2H-I/
      DATA BLANK/4H    /,BLINK/4H   ./
      DO 2 IA=1,5
      DO 2 IB=1,5
      IC=(IA-1)*5+IB
      GRIDA(1,IC)=GRDA(IB)
      GRIDA(53,IC)=GRDA(IB)
2     CONTINUE
      GRIDA(1,26)=GRDA(6)
      GRIDA(53,26)=GRDA(6)
      GRIDB(1,1)=GRDA(1)
      GRIDB(1,2)=GRDA(3)
      GRIDB(1,3)=GRDA(5)
      GRIDB(1,4)=GRDA(3)
      GRIDB(1,5)=GRDA(2)
      GRIDB(1,6)=GRDA(2)
      GRIDB(1,7)=GRDA(4)
      GRIDB(1,8)=GRDA(5)
      GRIDB(1,9)=GRDA(1)
      GRIDB(1,10)=GRDA(3)
      GRIDB(1,11)=GRDA(4)
      GRIDB(1,12)=GRDA(1)
      GRIDB(1,13)=GRDA(4)
      GRIDB(1,14)=GRDA(4)
      GRIDB(1,15)=GRDA(4)
      GRIDB(1,16)=GRDA(5)
      GRIDB(1,17)=GRDA(1)
      GRIDB(1,18)=GRDA(3)
      GRIDB(1,19)=GRDA(4)
      GRIDB(1,20)=GRDA(5)
      GRIDB(1,21)=GRDA(2)
```

```
      GRIDB(1,22)=GRDA(2)
      GRIDB(1,23)=GRDA(1)
      GRIDB(1,24)=GRDA(5)
      GRIDB(1,25)=GRDA(1)
      GRIDB(1,26)=GRDA(7)
      DO 40 IA=1,26
      GRIDB(53,IA)=GRIDB(1,IA)
40    CONTINUE
      DO 8 IA=2,52
      DO 6 IB=2,25
      GRIDA(IA,14)=BLANK
      GRIDB(IA,13)=BLANK
6     CONTINUE
      GRIDB(IA,13)=BLINK
8     CONTINUE
      DO 10 IA=2,52
      GRIDA(IA,1)=GRDB(1)
      GRIDA(IA,26)=GRDB(2)
      GRIDB(IA,1)=GRDB(1)
      GRIDB(IA,25)=GRDB(2)
10    CONTINUE
      DO 12 IA=2,25
      GRIDA(L1,IA)=GRDA(5)
      GRIDB(L1,IA)=GRDA(5)
12    CONTINUE
      GRIDA(L1,1)=GRDB(3)
      GRIDA(L1,26)=GRDB(4)
      GRIDB(L1,26)=GRDB(4)
      GRIDB(L1,1)=GRDB(3)
      RETURN
      END
C
C ---  **********************
C      SUB ROTINA PACK
C ---  **********************
C
C            COLOCA O CARACTER "C" NO BYTE "N" DA PALAVRA "X",
C     ONDE X E UMA PALAVRA SIMULADA DO IBM 360, COM OS BYTES NU
C     MERADOS DA ESQUERDA.
C            OBTIDA EM:
C     WOOD,F.S.(1976) - NONLINEAR LEAST-SQUARES CURVE FITTING
C     PROGRAM - DECUS PROGRAM LIBRARY
C
      SUBROUTINE PACK(X,N,C)
      DIMENSION FNT(2)
      IF(N.EQ.1) GO TO 10
      ND=N-1
      NC=9*ND
      ENCODE(10,2,FNT)ND,NC
2     FORMAT(2H(A,I1,5H,A1,R,I1,1H))
      ENCODE(5,FNT,X)X,C,X
      GO TO 20
10    ENCODE(5,1,X)C,X
C*1   FORMAT(A1,R4)
1     FORMAT(A1,R9)
20    RETURN
      END
C
C ---  **********************
C      SUB ROTINA PLOT
```

```
C ---   ********************
C
C               MONTA OS GRAFICOS DE DISTRIBUICAO ACUMULADA DOS RESIDUOS
C       E RESIDUOS VS. Y AJUSTADOS. CHAMA AS SUB ROTINAS GRID E PACK.
C               ADAPTADA A PARTIR DO EXISTENTE EM:
C       WOOD,F.S.(1976) - NON LINEAR LEAST-SQUARES CURVE FITTING
C       PROGRAM - DECUS PROGRAM LIBRARY
C
        SUBROUTINE PLOT(NOOBSV,YYCC,DELTA,LSORT,YCMIN,YCMAX,GRIDA,GRIDB)
        DIMENSION YYCC(1),DELTA(1),LSORT(1)
        DIMENSION GRIDA(53,26),GRIDB(53,26),NEG(8),POS(8),PRAX(51)
        DIMENSION RESID(9)
        DATA RESID/1HR,1HE,1HS,1HI,1HD,1HU,1HO,1HS,1H /
        DATA POS/1HP,1HO,1HS,1HI,1HT,1HI,1HV,1HO/
        DATA NEG/1HN,1HE,1HG,1HA,1HT,1HI,1HV,1HO/
C       DATA IPLUS/4H   +/
        DATA IPLUS/1H+/
        DATA PRAX/.000233, .000302, .000390, .000501, .000641,
       1          .000816, .001035, .001306, .001641, .002052,
       2          .002555, .003167, .003907, .004797, .005868,
       3          .007143, .008656, .010444, .012545, .015003,
       4          .017864, .021176, .025000, .029379, .034380,
       5          .040059, .046479, .053699, .061780, .070781,
       6          .080757, .091759, .103835, .117023, .131357,
       7          .146859, .163543, .181411, .200454, .220650,
       8          .241964, .264347, .287740, .312067, .337243,
       9          .363169, .389739, .416834, .444330, .472097,
       11               .500000/
C       IDENTIFICACAO DA FILA DE IMPRESSAO-KTUV
        KTUV=3
        XNP=NOOBSV
        JJC=LSORT(NOOBSV)
        DELP=DELTA(JJC)
        IIC=LSORT(1)
        YIW=(1.005*DELP-DELTA(IIC))/51.
        XIW=(1.005*(YCMAX-YCMIN)/101.
        DO 2 IA=1,NOOBSV
        JAC=LSORT(IA)
        IF(DELTA(JAC).LT.0) GO TO 2
        L1=INT((DELP-DELTA(JAC))/YIW)+2
        GO TO 4
2       CONTINUE
4       CALL GRID(GRIDA,GRIDB,L1)
        DO 30 IA=1,NOOBSV
        JAC=LSORT(IA)
        LINE=INT((DELP-DELTA(JAC))/YIW)+2
        IF(LINE.LT.2.OR.LINE.GT.52) GO TO 30
        LOCX1=INT((YYCC(JAC)-YCMIN)/XIW)+1
        IF(LOCX1.LT.1.OR.LOCX1.GT.101) GO TO 10
        IF(LOCX1.EQ.1) LOCX1=2
        LOCX2=LOCX1+1
        LCHAR=MOD(LOCX2,4)
        LWORD=LOCX2/4
        IF(LCHAR)8,6,8
6       LCHAR=4
        LWORD=LWORD-1
8       CALL PACK(GRIDA(LINE,LWORD),LCHAR,IPLUS)
10      PRLOC=(IA-.5)/XNP
        IF(PRLOC-.5)12,12,18
12      DO 14 IB=1,51
```

```
      IF(PRLOC-PRAX(IB))16,16,14
14    CONTINUE
      IB=51
16    LOCP1=IB
      GO TO 24
18    DO 20 IB=1,51
      IC=52-IB
      IF(PRLOC+PRAX(IC)-1.)22,22,20
20    CONTINUE
22    LOCP1=IB+50
24    LOCP2=LOCP1+4
      LCHAR=MOD(LOCP2,4)
      LWORD=LOCP2/4
      IF(LCHAR)28,26,28
26    LCHAR=4
      LWORD=LWORD-1
28    CALL PACK(GRIDB(LINE,LWORD),LCHAR,IPLUS)
30    CONTINUE
      WRITE(KTOV,34)
34    FORMAT(1H1,///,54X,'DISTRIBUICAO ACUMULADA DOS RESIDUOS')
      WRITE(KTOV,36)
36    FORMAT(1H ,17X,10(H.0002  .001  .005  .01 .02   .05  .1    .2   .3
     1  .4  .5  .6  .7  .8   .9   .95  .98 .99 .995   .999)
      WRITE(KTOV,38)((GRIDB(IA,JZ),JZ=1,26),IA=1,3)
38    FORMAT(1H ,19X,25A4,A2)
      WRITE(KTOV,40)(POS(IA-3),(GRIDB(IA,JZ),JZ=1,26),IA=4,11)
40    FORMAT(1H ,17X,A1,1X,25A4,A2)
      DO 46 IA=12,22
      IF(IA.EQ.L1) GO TO 42
      WRITE(KTOV,38)(GRIDB(IA,JZ),JZ=1,26)
      GO TO 46
42    WRITE(KTOV,44)(GRIDB(IA,JZ),JZ=1,26)
44    FORMAT(1H ,17X,1H0,1X,25A4,A2)
46    CONTINUE
      DO 54 IA=23,31
      IF(IA.EQ.L1) GO TO 50
      WRITE(KTOV,48)RESID(IA-22),(GRIDB(IA,JZ),JZ=1,26)
48    FORMAT(1H ,13X,A1,5X,25A4,A2)
      GO TO 54
50    WRITE(KTOV,52) RESID(IA-22),(GRIDB(IA,JZ),JZ=1,26)
52    FORMAT(1H ,13X,A1,3X,1H0,1X,25A4,A2)
54    CONTINUE
      DO 58 IA=32,42
      IF(IA.EQ.L1) GO TO 56
      WRITE(KTOV,38)(GRIDB(IA,JZ),JZ=1,26)
      GO TO 58
56    WRITE(KTOV,44)(GRIDB(IA,JZ),JZ=1,26)
58    CONTINUE
      WRITE(KTOV,40)(NEG(IA-42),(GRIDB(IA,JZ),JZ=1,26),IA=43,50)
      WRITE(KTOV,38)((GRIDB(IA,JZ),JZ=1,26),IA=51,53)
      WRITE(KTOV,36)
      WRITE(KTOV,60)
60    FORMAT(1H ,52X,'DISTRIBUICAO ACUMULADA, MALHA NORMAL')
      WRITE(KTOV,62)
62    FORMAT(1H1,///,51X,'RESIDUOS Vs. Y AJUSTADOS')
      YCAVG=(YCMAX+YCMIN)/2.
      WRITE(KTOV,64)YCMIN,YCAVG,YCMAX
64    FORMAT(1H ,13X,F10.3,2(30X,F10.3))
      WRITE(KTOV,38)((GRIDA(IA,JZ),JZ=1,26),IA=1,3)
      WRITE(KTOV,40)(POS(IA-3),(GRIDA(IA,JZ),JZ=1,26),IA=4,11)
```

```
      DO 69 IA=12,22
      IF(IA.EQ.L1) GO TO 66
      WRITE(KTOV,38)(GRIDA(IA,JZ),JZ=1,26)
      GO TO 68
66    WRITE(KTOV,44)(GRIDA(IA,JZ),JZ=1,26)
68    CONTINUE
      DO 72 IA=23,31
      IF(IA.EQ.L1) GO TO 70
      WRITE(KTOV,48)RESID(IA-22),(GRIDA(IA,JZ),JZ=1,26)
      GO TO 72
70    WRITE(KTOV,52)RESID(IA-22),(GRIDA(IA,JZ),JZ=1,26)
72    CONTINUE
      DO 76 IA=32,42
      IF(IA.EQ.L1)  GO TO 74
      WRITE(KTOV,38)(GRIDA(IA,JZ),JZ=1,26)
      GO TO 76
74    WRITE(KTOV,44)(GRIDA(IA,JZ),JZ=1,26)
76    CONTINUE
      WRITE(KTOV,40)(NEG(IA-42),(GRIDA(IA,JZ),JZ=1,26),IA=43,50)
      WRITE(KTOV,39)((GRIDA(IA,JZ),JZ=1,25),IA=51,53)
      WRITE(KTOV,64)YCMIN,YCAVG,YCMAX
      WRITE(KTOV,78)
78    FORMAT(1H ,67X,'Y AJUSTADO')
      RETURN
      END
```

* * * * * * * * * REGRESSAO BIPONDERADA * * * * * * * * * *

*****DADOS*****

N	VAR. DEP.	VARIAVEIS INDEPENDENTES		
1	37.0000	80.0000	27.0000	88.0000
2	18.0000	62.0000	22.0000	87.0000
3	18.0000	62.0000	23.0000	87.0000
4	19.0000	62.0000	24.0000	93.0000
5	20.0000	62.0000	24.0000	93.0000
6	15.0000	58.0000	23.0000	87.0000
7	14.0000	58.0000	18.0000	80.0000
8	14.0000	58.0000	19.0000	89.0000
9	13.0000	58.0000	17.0000	88.0000
10	11.0000	58.0000	18.0000	82.0000
11	12.0000	58.0000	19.0000	93.0000
12	8.0000	50.0000	18.0000	89.0000
13	7.0000	50.0000	18.0000	86.0000
14	8.0000	50.0000	19.0000	72.0000
15	8.0000	50.0000	19.0000	79.0000
16	9.0000	50.0000	20.0000	80.0000
17	15.0000	56.0000	20.0000	82.0000

***** PARAMETROS DE INICIALIZACAO OBTIDOS POR MINIMOS QUADRADOS *****

***** ESTIMATIVAS INICIAIS *****
CONSTANTE= -37.65244910
B(1)= 0.79768513
B(2)= 0.57731180
B(3)= -0.06706032

ORDEM DE ENTRADA

OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.8492820	1.1507177
2	18.0000000	18.6713010	-0.6713006
3	18.0000000	19.2486420	-1.2486424

ORDENADOS PELOS RESIDUOS

OBS.	YOBS	YEST	RESIDUOS
17	15.0000000	13.0658080	1.9341922
12	8.0000000	6.6555911	1.3444089
1	37.0000000	35.8492820	1.1507177

4	19.0000000	19.4236220	-0.4236224	8	14.0000000	13.0370720	0.9629277
5	20.0000000	19.4236720	0.5763776	16	9.0000000	8.4139175	0.5861825
6	15.0000000	16.0579020	-1.0579019	5	20.0000000	19.4236220	0.5763776
7	14.0000000	13.6406159	0.3593849	9	13.0000000	12.5267910	0.4732091
8	14.0000000	13.0370720	0.9629277	7	14.0000000	13.6406150	0.3593849
9	13.0000000	12.5267910	0.4732091	13	7.0000000	6.8567721	0.1432279
10	11.0000000	13.5064950	-2.5064945	15	8.0000000	7.9035361	0.0964639
11	12.0000000	13.3461730	-1.3461728	14	8.0000000	8.3729583	-0.3729583
12	8.0000000	6.6559911	1.3440089	4	19.0000000	19.4236220	-0.4236224
13	7.0000000	6.8567721	0.1432279	2	18.0000000	18.6713010	-0.6713006
14	8.0000000	8.3729583	-0.3729583	6	15.0000000	16.0579020	-1.0579019
15	8.0000000	7.9035361	0.0964639	3	18.0000000	19.2486420	-1.2486424
16	9.0000000	8.4138175	0.5861825	11	12.0000000	13.3461730	-1.3461728
17	15.0000000	13.9658080	1.9341922	10	11.0000000	13.5064950	-2.5064945

---

* * * * * * * * * * PROCESSO ITERATIVO * * * * * * * * * *

CONSTANTE DE ESCALONAMENTO 4.0000000
EDIDA ROBUSTA DE ESCALA AMPLITUDE INTERQUARTIS/ 1.3500000
PRECISAO PARA A CONVERGENCIA 0.0010000

***** ITERACAO NUMERO 1 *****
CONSTANTE= -37.42751140
B(1)= 0.81499580
B(2)= 0.52780841
B(3)= -0.07136924

ORDEM DE ENTRADA

OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9425750	1.0571250
2	18.0000000	18.7049770	-0.7049766
3	18.0000000	19.2327850	-1.2327850
4	19.0000000	19.3323840	-0.3323839
5	20.0000000	19.3323840	0.6676161
6	15.0000000	15.9728020	-0.9728018
7	14.0000000	13.8333370	0.1666627
8	14.0000000	13.1910230	0.8089768
9	13.0000000	12.7345830	0.2654170
10	11.0000000	13.6906010	-2.6906009
11	12.0000000	13.4333590	-1.4333587

ORDENADOS PELOS RESIDUOS

OBS.	YOBS	YEST	RESIDUOS
17	15.0000000	13.1162260	1.8837738
12	8.0000000	6.6710566	1.3289434
1	37.0000000	35.9425750	1.0574250
8	14.0000000	13.1910230	0.8089768
5	20.0000000	19.3323840	0.6676161
16	9.0000000	8.3689876	0.6310124
9	13.0000000	12.7345830	0.2654170
7	14.0000000	13.8333370	0.1666627
13	7.0000000	6.8851613	0.1148387
15	8.0000000	7.9125474	0.0874525
4	19.0000000	19.3323840	-0.3323839

12	8.0000000	6.6710566	1.3289434	14	8.0000000	8.4121251	-0.4121251
13	7.0000000	6.8951613	0.1048387	2	18.0000000	18.7049770	-0.7049766
14	8.0000000	8.4121251	-0.4121251	6	15.0000000	15.9728020	-0.9728018
15	8.0000000	7.9125474	0.0874526	3	18.0000000	19.2327850	-1.2327850
16	9.0000000	8.3689876	0.6310124	11	12.0000000	13.4333590	-1.4333587
17	15.0000000	13.1162260	1.8837738	10	11.0000000	13.6906010	-2.6906009

***** ITERACAO NUMERO 2 *****
CONSTANTE= -37.1317060
B(1)= 0.81911772
B(2)= 0.51688578
B(3)= -0.07269520

ORDEM DE ENTRADA

OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9749850	1.0250139
2	18.0000000	18.7191330	-0.7191327
3	18.0000000	19.2360180	-1.2360184
4	19.0000000	19.3167330	-0.3167331
5	20.0000000	19.3167330	0.6832669
6	15.0000000	15.9595480	-0.9595478
7	14.0000000	13.8839850	0.1160147
8	14.0000000	13.2297290	0.7702714
9	13.0000000	12.7855380	0.2144620
10	11.0000000	13.7385950	-2.7385949
11	12.0000000	13.4558340	-1.4558336
12	8.0000000	6.6767866	1.3232134
13	7.0000000	6.8948722	0.1051278
14	8.0000000	8.4291909	-0.4291909
15	8.0000000	7.9206244	0.0793756
16	9.0000000	8.3648150	0.6351850
17	15.0000000	13.1341310	1.8658692

ORDENADOS PELOS RESIDUOS

OBS.	YOBS	YEST	RESIDUOS
17	15.0000000	13.1341310	1.8658692
12	8.0000000	6.6767866	1.3232134
1	37.0000000	35.9749850	1.0250149
8	14.0000000	13.2297290	0.7702714
5	20.0000000	19.3167330	0.6832669
16	9.0000000	8.3648150	0.6351850
9	13.0000000	12.7855380	0.2144620
7	14.0000000	13.8839850	0.1160147
13	7.0000000	6.8948722	0.1051278
15	8.0000000	7.9206244	0.0793756
4	19.0000000	19.3167330	-0.3167331
14	8.0000000	8.4294909	-0.4294909
2	18.0000000	18.7191330	-0.7191327
6	15.0000000	15.9595480	-0.9595478
3	18.0000000	19.2360180	-1.2360184
11	12.0000000	13.4558340	-1.4558336
10	11.0000000	13.7385950	-2.7385949

***** ITERACAO NUMERO 3 *****
CONSTANTE= -37.0841260
B(1)= 0.82010362

B(2)= 0.51443836
B(3)= -0.07306032

OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9343850	1.0156150
2	18.0000000	18.7233880	-0.7233882
3	18.0000000	19.2378260	-1.2378263
4	19.0000000	19.3139030	-0.3139029
5	20.0000000	19.3139030	0.6860971
6	15.0000000	15.9574120	-0.9574118
7	14.0000000	13.8966420	0.1033577
8	14.0000000	13.2391000	0.7609005
9	13.0000000	12.7977210	0.2022786
10	11.0000000	13.7505220	-2.7505217
11	12.0000000	13.4612970	-1.4612966
12	8.0000000	6.6782708	1.3217292
13	7.0000000	6.8974517	0.1025483
14	8.0000000	8.4347343	-0.4347343
15	8.0000000	7.9233122	0.0766878
16	9.0000000	8.3646903	0.6353097
17	15.0000000	13.1391920	1.8608084

ORDENADOS PELOS RESIDUOS

OBS.	YOBS	YEST	RESIDUOS
17	15.0000000	13.1391920	1.8608084
12	8.0000000	6.6782708	1.3217292
1	37.0000000	35.9843850	1.0156150
8	14.0000000	13.2391000	0.7609005
5	20.0000000	19.3139030	0.6860971
16	9.0000000	8.3646903	0.6353097
9	13.0000000	12.7977210	0.2022786
7	14.0000000	13.8966420	0.1033577
13	7.0000000	6.8974517	0.1025483
15	8.0000000	7.9233122	0.0766878
4	19.0000000	19.3139030	-0.3139029
14	8.0000000	8.4347343	-0.4347343
2	18.0000000	18.7233880	-0.7233882
6	15.0000000	15.9574120	-0.9574118
3	18.0000000	19.2378260	-1.2378263
11	12.0000000	13.4612970	-1.4612966
10	11.0000000	13.7505220	-2.7505217

***** ITERACAO NUMERO 4 *****
CONSTANTE= -37.9771809
B(1)= 0.82034197
B(2)= 0.51387516
B(3)= -0.07315788

OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9869130	1.0130873
2	18.0000000	18.7245390	-0.7245390
3	18.0000000	19.2384140	-1.2384140
4	19.0000000	19.3133420	-0.3133421
5	20.0000000	19.3133420	0.6866579
6	15.0000000	15.9570460	-0.9570463
7	14.0000000	13.8997760	0.1002243
8	14.0000000	13.2413550	0.7586452
9	13.0000000	12.8006370	0.1993626
10	11.0000000	13.7534600	-2.7534599
11	12.0000000	13.4625990	-1.4625986
12	8.0000000	6.6786192	1.3213808
13	7.0000000	6.8980929	0.1019071
14	8.0000000	8.4361783	-0.4361783

ORDENADOS PELOS RESIDUOS

OBS.	YOBS	YEST	RESIDUOS
17	15.0000000	13.1405270	1.8594735
12	8.0000000	6.6786192	1.3213808
1	37.0000000	35.9869130	1.0130873
8	14.0000000	13.2413550	0.7586452
5	20.0000000	19.3133420	0.6866579
16	9.0000000	8.3647904	0.6352096
9	13.0000000	12.8006370	0.1993626
13	7.0000000	6.8980929	0.1019071
7	14.0000000	13.8997760	0.1002243
15	8.0000000	7.9240732	0.0759268
4	19.0000000	19.3133420	-0.3133421
14	8.0000000	8.4361783	-0.4361783
2	18.0000000	18.7245390	-0.7245390
6	15.0000000	15.9570460	-0.9570463

15	8.0000000	7.9240732	0.0759268	3	18.0000000	19.2384140	-1.238
16	9.0000000	8.3647904	0.6352096	11	12.0000000	13.4625986	-1.4625986
17	15.0000000	13.1405270	1.8594735	10	11.0000000	13.7534600	-2.7534599

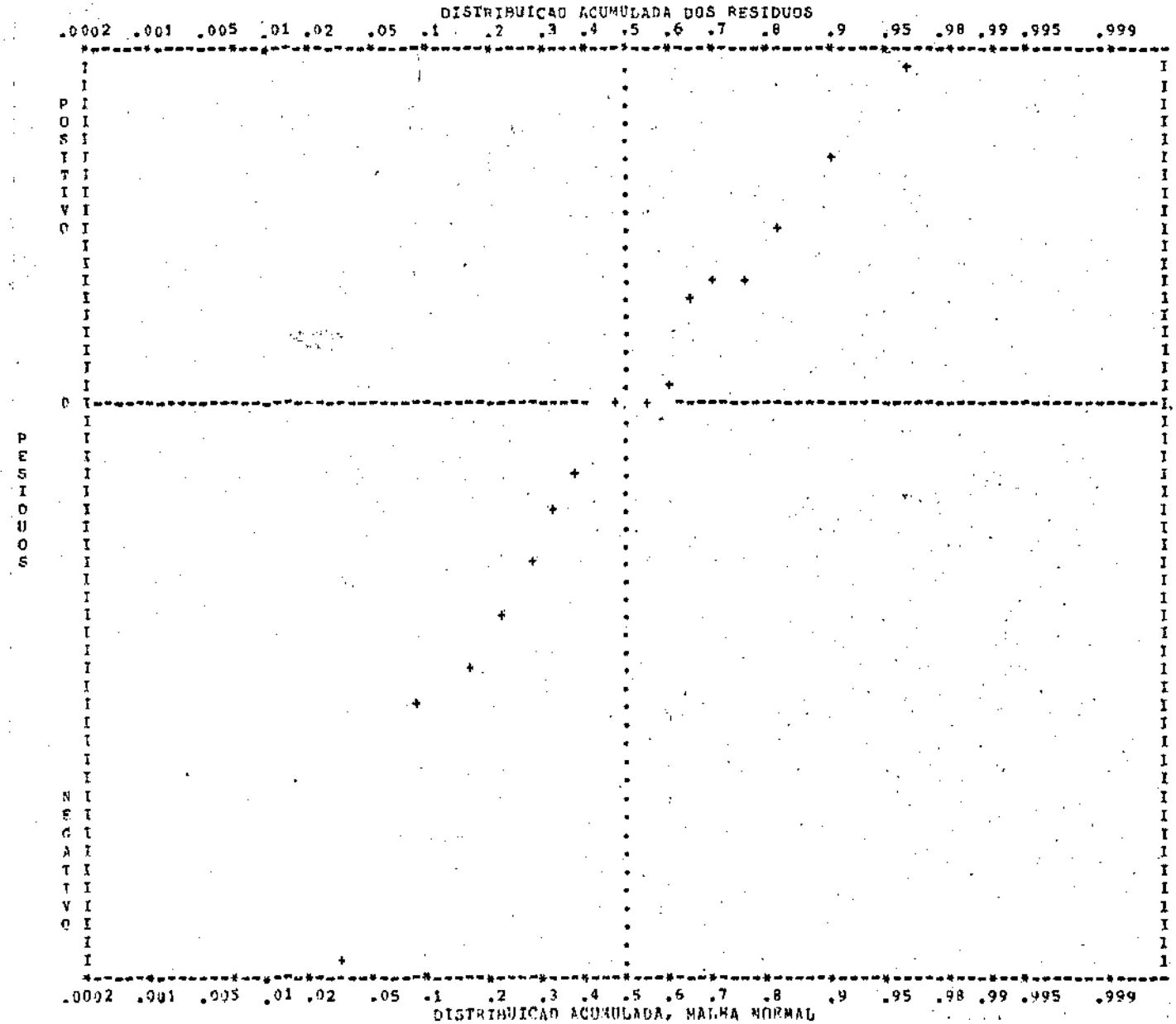
***** ITERACAO NUMERO 5 *****
CONSTANTE= -37.07566600
B(1)= 0.82040038
B(2)= 0.51374335
B(3)= -0.07318011

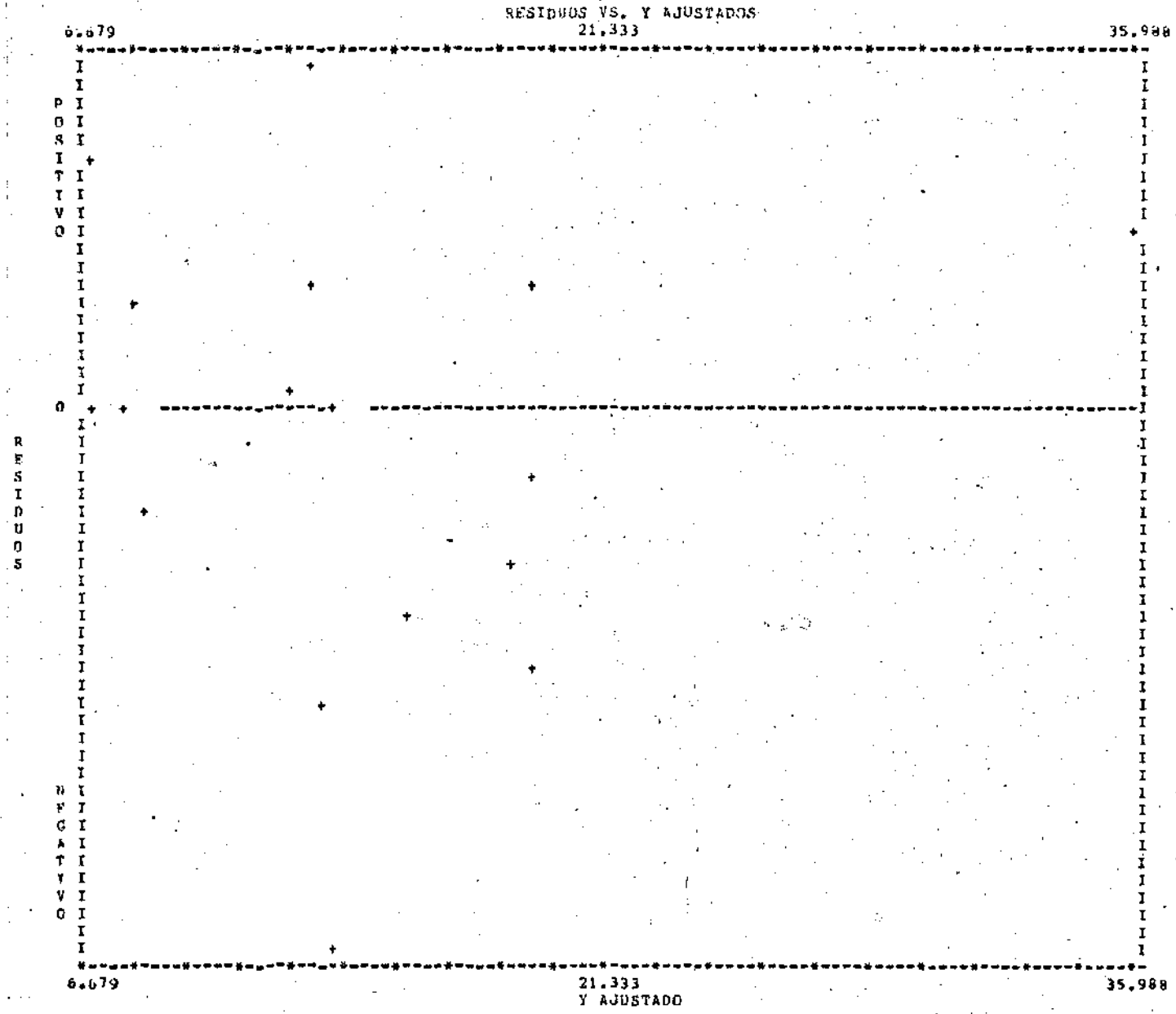
ORDEN DE ENTRADA				ORDENADOS PELOS RESIDUOS			
OBS.	YOBS	YEST	RESIDUOS	OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9875860	1.0124145	17	15.0000000	13.1408540	1.8591462
2	18.0000000	18.7248420	-0.7248416	12	8.0000000	6.6787038	1.3212962
3	18.0000000	19.2385850	-1.2385850	1	37.0000000	35.9875860	1.0124145
4	19.0000000	19.3132480	-0.3132479	8	14.0000000	13.2419070	0.7580934
5	20.0000000	19.3132480	0.6867521	5	20.0000000	19.3132480	0.6867521
6	15.0000000	15.9569840	-0.9569837	16	9.0000000	8.3648114	0.6351886
7	14.0000000	13.9005280	0.0994724	9	13.0000000	12.8013440	0.1986563
8	14.0000000	13.2419070	0.7580934	13	7.0000000	6.8982441	0.1017559
9	13.0000000	12.8013440	0.1986563	7	14.0000000	13.9005280	0.0994724
10	11.0000000	13.7541670	-2.7541674	15	8.0000000	7.9242483	0.0757517
11	12.0000000	13.4629300	-1.4629296	4	19.0000000	19.3132480	-0.3132479
12	8.0000000	6.6787038	1.3212962	14	8.0000000	8.4365090	-0.4365090
13	7.0000000	6.8982441	0.1017559	2	18.0000000	18.7248420	-0.7248416
14	8.0000000	8.4365090	-0.4365090	6	15.0000000	15.9569840	-0.9569837
15	8.0000000	7.9242483	0.0757517	3	18.0000000	19.2385850	-1.2385850
16	9.0000000	8.3648114	0.6351886	11	12.0000000	13.4629300	-1.4629296
17	15.0000000	13.1408540	1.8591462	10	11.0000000	13.7541670	-2.7541674

***** ITERACAO NUMERO 6 *****
CONSTANTE= -37.07519870
B(1)= 0.82041684
B(2)= 0.51370604
B(3)= -0.07319691

ORDEM DE ENTRADA				ORDENADOS PELOS RESIDUOS			
OBS.	YOBS	YEST	RESIDUOS	OBS.	YOBS	YEST	RESIDUOS
1	37.0000000	35.9877640	1.0122356	17	15.0000000	13.1409390	1.8590608
2	18.0000000	18.7249170	-0.7249174	12	8.0000000	6.6787176	1.3212824
3	18.0000000	19.2386230	-1.2386234	1	37.0000000	35.9877640	1.0122356
4	19.0000000	19.3132080	-0.3132081	8	14.0000000	13.2420520	0.7579478
5	20.0000000	19.3132080	0.6867919	5	20.0000000	19.3132080	0.6867919
6	15.0000000	15.9569560	-0.9569560	16	9.0000000	8.3648119	0.6351881
7	14.0000000	13.9007340	0.0992657	9	13.0000000	12.8015330	0.1984669
8	14.0000000	13.2420520	0.7579478	13	7.0000000	6.8982782	0.1017218
9	13.0000000	12.8015330	0.1984669	7	14.0000000	13.9007340	0.0992657
10	11.0000000	13.7543600	-2.7543604	15	8.0000000	7.9242927	0.0757073
11	12.0000000	13.4630110	-1.4630105	4	19.0000000	19.3132080	-0.3132081
12	8.0000000	6.6787176	1.3212824	14	8.0000000	8.4366010	-0.4366010
13	7.0000000	6.8982782	0.1017218	2	18.0000000	18.7249170	-0.7249174
14	8.0000000	8.4366010	-0.4366010	6	15.0000000	15.9569560	-0.9569560
15	8.0000000	7.9242927	0.0757073	3	18.0000000	19.2386230	-1.2386234
16	9.0000000	8.3648119	0.6351881	11	12.0000000	13.4630110	-1.4630105
17	15.0000000	13.1409390	1.8590608	10	11.0000000	13.7543600	-2.7543604

********** O PROCESSO ITERATIVO CONVERGIU EM 6 ITERACOES **********


DISTRIBUICAO ACUMULADA DOS RESIDUOS
.0002 .001 .005 .01 .02 .05 .1 .2 .3 .4 .5 .6 .7 .8 .9 .95 .98 .99 .995 .999
POSITIVO
0
NEGATIVO
RESIDUOS
DISTRIBUICAO ACUMULADA, MALHA NORMAL


RESIDUOS VS. Y AJUSTADOS
6.679
21.333
35.988
POSITIVO
0
NEGATIVO
RESIDUOS
6.679
21.333
35.988
Y AJUSTADO

VII.5 - Regressão Biponderada através do SPSS

Será apresentado agora o modo de obter regressão biponderada através do SPSS (Statistical Package for the Social Sciences). Há a necessidade de se preparar tres programas; dois para o SPSS e um em FORTRAN. Além dos programas deve-se preparar também um arquivo com os dados; os valores de Y, X_1, X_2,...,X_k. Para maior facilidade vai-se explicar o procedimento através de um exemplo. Suponha que se queira realizar um ajuste com os dados do exemplo 1, do capítulo V. Os dados estão no quadro 5.1.1, mas como já foi visto vai se utilizar somente duas variáveis independentes, as variáveis X_1 e X_4. A denominação deste arquivo deve cumprir:

- ter extensão CDR
- o nome principal não pode ter mais de tres caracteres, sendo o primeiro obrigatoriamente alfabético.

Deste modo vai-se escolher o nome DAD.CDR. Este arquivo deverá ser gravado com o formato livre; FREEFIELD no SPSS e formato F no FORTRAN.

Os programas do SPSS serão denominados ESTIN.SPS e BIPON.SPS. O primeiro dará as estimativas iniciais, obtidas por mínimos quadrados, e o segundo dará as estimativas obtidas pela regressão biponderada.

A listagem do primeiro; ESTIN.SPS é a seguinte:

```
RUN NAME                AJUSTE POR MIN. QUAD. - INIC. DOS PARAMETROS
FILE NAME               DADOS2
VARIABLE LIST           Y,X1,X4
INPUT FORMAT            FREEFIELD
INPUT MEDIUM            EX1.CDR
N OF CASES              13
REGRESSION              VARIABLES= Y,X1,X4
                        REGRESSION= Y WITH X1,X4(2)/
                        RESID=0.0

OPTIONS                 8,13
STATISTICS              1,2,3,4,6
READ INPUT DATA
SAVE FILE
FINISH
```

Este programa além de calcular a estimativa inicial dos parâmetros, grava, na área do usuário, o arquivo DADOS2, que será utilizado pelo programa BIPON.SPS e que nada mais é que o arquivo de dados, DAD.CDR, compilado pelo SPSS.

O programa BIPON.SPS necessita, além dos dados, de um arquivo contendo os resíduos. Este arquivo será providenciado pelo programa em FORTRAN, que será denominado RESID.F40. Este segundo programa em SPSS necessita ainda do valor da amplitude interquartís do conjunto dos resíduos, que também será providenciado pelo programa em FORTRAN. Este valor deverá ser colocado no lugar de XXX na expressão:

```
COMPUTE          SK = (4.*   XXX   )/1.35
```

A listagem do programa BIPON.SPS é a seguinte:

```
RUN NAME                REGRESSÃO BIPONDERADA
GET FILE                DADOS2
ADD VARIABLES           R
INPUT FORMAT            FREEFIELD
INPUT MEDIUM            FOR22.DAT
COMPUTE                 SK=(4.*  XXX  )/1.35
COMPUTE                 U=(R/SK)**2
IF                      (U GE 1.)W=0
IF                      (U LT 1.)W=1.-U
COMPUTE                 X0=W
COMPUTE                 X1=W*X1
COMPUTE                 X4=W*X4
COMPUTE                 Y=W*Y
REGRESSION              VARIABLES=Y,X0,X1,X4/
                        REGRESSION=Y WITH X0,X1,X4(2)
                        RESID=0.0
OPTIONS                 8,13,16
STATISTICS              1,2,3,4,6
READ INPUT DATA
FINISH
```

A instrução:

```
INPUT MEDIUM            FOR22.DAT
```

indica que os resíduos, R, estarão gravados num arquivo com o nome FOR22.DAT, criado pelo programa em FORTRAN.

O programa em FORTRAN, RESID.F40, com os valores $\hat{b}_0$, $\hat{b}_1$ e $\hat{b}_4$ calculados ou pelo programa ESTIN.SPS ou pelo programa BIPON.SPS, calcula os resíduos:

$$R = Y - \hat{Y} = Y - X\hat{B}$$

Grava estes resíduos no arquivo FOR22.DAT e ainda calcula a amplitude interquartís dos elementos do vetor dos resíduos, R. A listagem deste programa está logo a seguir. Nela não estão presentes as listagens das

sub rotinas ORDEM, ORRES e AMPIQ, mas podem ser obtidas junto à listagem do programa REGRO.F10, na secção VII.4.

```
      DIMENSION YY(N),VO(N)
      DIMENSION X(N,K),Y(N),XB(N),KO(N),XBO(N),R(N),RO(N)
      N=NUMERO DE OBSERVACOES
      K=NUMERO DE VARIAVEIS INDEPENDENTES
      WRITE(3,1)
1     FORMAT(5X,' ITERACAO XXX ',/)
      BO=VALOR DE BO (CONSTANTE)
      B1=VALOR DE B1
      .
      .
      .
      BK=VALOR DE BK
      READ(2,5)((Y(I),(X(I,J),J=1,K)),I=1,N)
5     FORMAT(K * F)
      WRITE(3,5)BO,B1,...,BK
5     FORMAT(K * F)
      DO 10 I=1,N
      XB(I)=BO+B1*X(I,1)+...+BK*X(I,K)
10    R(I)=Y(I)-XB(I)
      CALL ORRES(N,R,XB,KO,RO,XBO,YY)
      DO 20 I=1,N
20    WRITE(3,30)I,Y(I),XB(I),R(I),YY(I),RO(I),KO(I)
30    FORMAT(1X,I3,3(3X,F),10X,2(3X,F),I3)
      CALL AMPIQ(N,R,DIS,VO,1.35)
      AI=DIS*1.35
      WRITE(3,40)AI
40    FORMAT(F)
      CALL EXIT
      END
```

Os vetores Y, XB, R, YY, RO e KO, que aparecem na listagem do programa acima são, respectivamente:

- o vetor das observações (variável dependente)
- o vetor dos Y ajustados
- o vetor dos resíduos

Estes tres vetores são impressos na ordem em que se gravou os dados; ou seja, na ordem de entrada dos dados. Os tres vetores seguintes, YY, RO e KO são:

RO - vetor dos resíduos ordenados (ordem decrescente)

KO - dá a posição dos elementos do vetor RO no vetor R

YY - vetor que contém os valores do vetor Y, ordenados segundo a ordenação de R em RO

A partir deste ponto será apresentado um processo, dividido em Passos, que possibilitará a obtenção da regressão biponderada através do SPSS. Supõe-se que o usuário tenha acesso aos terminais da UNICAMP ligados ao PDP-10.

Passo 1 - gravar o arquivo de dados; DAD.CDR

gravar os programas do SPSS; ESTIN.SPS e BIPON.SPS

gravar o programa em FORTRAN; RESID.F40

Passo 2 - obtenção das estimativas iniciais

```
RUN NEW: SPSS
LPT:=ESTIN.SPS
```

Passo 3 - obtenção dos resíduos e da sua amplitude interquartís com os valores $\hat{b}_0$, $\hat{b}_1$ e $\hat{b}_4$ obtidos no Passo 2 (ou Passo 4). Antes de processar o programa coloque esses valores no lugar apropriado (ver listagem) do programa RESID.F40. Após isso:

```
SET   CDR   DAD.CDR
EXEC  RESID.F40
```

Passo 4 - obtenção das estimativas dos parâmetros, pela regressão bipon-
derada. Coloque o valor de AI (amplitude interquartís) no lu-
gar de XXX como anteriormente explicado e então faça:

RUN NEW:SPSS

LPT: = BIPON.SPS

Nesta última instrução pode-se trocar LPT por TTY. De posse dos novos valores dos parâmetros e também de posse dos valores anteriores, verifique a convergência, por exemplo, através do apresentado na secção IV.4.3. Se já se obteve a convergência vá para o Passo 5, caso contrário volte ao Passo 2.

Passo 5 - execute pela última vez o programa em FORTRAN para obter a lis-
tagem dos resíduos e dos resíduos ordenados e está terminado o
processo. Para executar o programa

SET CDR DAD.CDR

EXEC RESID.F40

Maiores informações sobre, por exemplo, quais são as opções e estatísticas que se poderia utilizar no SPSS podem ser obtidas em DOZZI (1977), KLECKA (1975) e NIE (1975).

De modo algum retire a opção 16 do programa BIPON.SPS, que não existe nas referências acima e que especifica regressão pela origem. Sem ela se obteria resultados completamente absurdos. Há ainda mais uma referência que pode ser consultada caso se queira maiores informações sobre o SPSS. É disponível a todos que têm acesso ao PDP-10. É obtida, em um terminal, fazendo:

PRINT NEW:SPSS.DOC/PAGES:25

REFERÊNCIAS BIBLIOGRÁFICAS

AFIFI, A. & AZEN, S.P. (1974) Statistical Analysis: A Computer Oriented Approach. Academic Press: Nova Iorque.

ANDREWS, D.F.; BICKEL, P.J.; HAMPEL, F.R.; HUBER, P.J.; ROGRES, W.H. & TUKEY, J.W. (1972) Robust Estimates Of Location: Survey and Advances. Princeton University Press.

ANDREWS, D.F. (1974) A Robust Method for Multiple Linear Regression. Technometrics, 16: 523-33.

ANDREWS, D.F. (1975) Alternative Calculations for Regression and Analysis of Variance. Em: GUPTA, P.R. (editor) Applied Statistics. Proceedings of the Conference at Dailhousie University, Halifax, maio de 1974. North-Holland Pub.Co.: Amsterdã.

ANSCOMBE, F.J. (1960) Rejection of Outliers. Technometrics, 2:123-47.

BEATON, A.E. & TUKEY, J.W. (1974) The fitting of Power Series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16: 147-185.

HAMPEL, F.R. (1973) Robust Estimation: A condensed partial survey. Z. Wahrscheinlichkeitstheorie und verw. Gebiete, 2: 87-104.

HAMPEL, F.R. (1974) The Influence Curve and it's role in Robust Estimation. J.Am.Statist.Assoc., 69: 383-93.

HILL, R.W. & HOLLAND, P.W. (1977) Two robust alternatives to Least-Square Regression. J.Am.Statist.Assoc., 72: 828-33.

HINICH, M.J. & TALWAR, P.O. (1975) A simple method for Robust Regression. J.Am.Statist.Assoc., 70: 113-19.

HUBER, P.J. (1964) Robust Estimation of a location parameter. Ann.Math. Statist., 35: 73-101.

HUBER, P.J. (1973) Robust Regression: Assyntotic conjectures and Monte-Carlo. Ann.Statist., 1: 799-821.

JAECKEL, L.A. (1972) Estimating Regression coefficients by minimizing the dispersion of the residuals. Ann.Math.Statist., 43: 1449-58.

KLECKA, W.R.; NIE, N.H.; HULL, C.H. (1975) SPSS Primer. MacGraw-Hill Book Co.

MORINEAU, A. (1978) Régressions Robustes. Méthodes d'Ajustement et de Validation. Rev.Statist.Appliq., 3: 5-28.

MOSTELLER, F. & TUKEY, J.W. (1977) Data Analysis and Regression. Addison Wesley Pub.Co.: Reading, Mass.

NIE, N.; HULL, C.H.; JENKINS, J.G.; STEINBRENNER, K.; BENT, D.H. (1975) Statistical Package for the Social Sciences. MacGraw-Hill Book Co., 2a. edição.

RALSTON, A. & WILF, H.S. (1960) Mathematical Methods for Digital Computers. John Wiley & Sons: Nova Iorque.

SEARLE, S.R. (1971) Linear Models. John Wiley & Sons: Nova Iorque.

STIGLER, S.M. (1973) Simon Newcomb, Percy Daniell and the history of Robust Estimation 1885-1920. J.Am.Statist.Assoc., 68: 872-79.

TUKEY, J.W. (1975) Instead of Gauss-Markov least-square, what? Em: GUPTA, P.R. (editor) Applied Statistics. Proceedings of the Conference at Dailhousie University, Halifax, maio de 1974. North. Holland Pub.Co.: Amsterdã.

TUKEY, R.W. (1977) Exploratory Data Analysis: EDA. Addison Wesley Pub. Co.: Reading, Mass.

WOOD, F.S. (1976) Nonlinear Least-Squares Curve Fitting Program. DECUS Program Library.

WOODS, H.; STEINOUR, H.H.; STARKE, H.R. (1932) Effect of composition of Portland cement on heat envolved during hardening. Ind.Eng.Che., 24: 1207-14.

ZAGO, J.V. (1978) Comunicação Pessoal. Departamento de Matemática Aplicada, IMECC-UNICAMP.