REGRESSÃO BIPONDERADA PASSO-A-FRENTE

Este exemplar corresponde a redação final da tese devidamente corrigida pela Sra. Inês Carvalho de Azevedo e aprovada pela Comissão Julgadora.

Campinas, 18 de novembro de 1987

Prof. Dr. José Norberto Walter Dachs

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística

A meus pais

A G R A D E C I M E N T O S

Ao prof. Dr. J. Norberto W. Dachs por sua orientação.

Ao Arthur pelos sorrisos mais deliciosos.

As amigas Irene, Karla e Lourdes pelo apoio e ajuda.

Ao Amorim por tudo.

Ao CNPq, Fapesp e Capes pelo apoio financeiro.

I N T R O D U Ç Ã O

Os procedimentos de seleção de variáveis na construção de modelos empíricos de regressão têm encontrado utilização crescente, graças principalmente ao barateamento e popularização do uso, e mesmo da posse, de computadores, e à disponibilidade cada vez maior de "softwares" específicos de boa qualidade.

Do ponto de vista Estatístico, todavia, o ferramental disponível para testes de significância é inadequado, podendo a utilização generalizada dos testes F de significância levar a distorções graves e, pior, desconhecidas.

Por outro lado, a preocupação, a partir dos anos 60, com a robustez dos estimadores, encontrou forte ressonância na área de regressão, com severas e bem fundamentadas restrições aos métodos dos mínimos quadrados sendo levantadas. Também aqui, o barateamento e popularização dos recursos computacionais, tanto em termos de "hardware" como de "software", abriram novas fronteiras para a pesquisa sobre estimadores, já agora livre das pesadas limitações de cálculo que fizeram no passado, os estimadores de mínimos quadrados, com a elegância algébrica e estatística de suas soluções, um competidor imbatível. Os estimadores de mínimos desvios absolutos, por exemplo, têm se constituído mais recentemente numa ativa frente de pesquisas.

Em 1977, Mosteller e Tukey propuseram a abordagem do problema de seleção de variáveis em análise de regressão, com a utilização de estimadores robustos. O estimador biponderado (Beaton e Tukey (1974)) contempla a situação em que os erros aleatórios possuem distribuição com caudas mais pesadas que as da normal. Em particular, no contexto de normais contaminadas, de óbvio interesse prático, o estimador biponderado de Tukey parece particularmente atraente.

Neste trabalho nós empregamos, num contexto específico, o estimador biponderado ao problema de seleção de variáveis, com atenção dirigida às questões dos níveis reais de significância do teste F usual, e da resistência dos estimadores à distribuição dos erros aleatórios de caudas mais pesadas que as da normal.

Nosso objetivo principal foi desenvolver uma metodologia para o estudo comparativo entre o estimador biponderado e o mínimos quadrados, aplicados ao problema da construção de modelos de regressão.

No capítulo 1 fazemos uma apresentação da teoria clássica de Regressão, introduzindo os conceitos básicos que serão utilizados nos trabalhos.

O estimador biponderado ao contrário do estimador de mínimos quadrados, não possui uma expressão algébrica fechada, sendo determinada de forma iterativa. Para uma introdução didática, de forma a se permitir uma melhor compreensão do mesmo, apresentamos no capítulo 2 exemplos de aplicação do estimador biponderado desde seu contexto mais simples : a estimação do parâmetro de locação, até sua aplicação no ajuste de modelos de Regressão Linear Múltipla. Em todos casos desenvolvemos um estudo comparativo entre o estimador biponderado e alguns competidores. Embora apresentando eficiência menor do que 1 com relação aos mínimos quadrados nos casos de normais puras ou de baixa contaminação, o estimador biponderado apresentou boa eficiência nos demais contextos estudados, caracterizados por média e pesada contaminação.

No capítulo 3 aplicamos uma variação de proposta de Mosteller e Tukey, a regressão biponderada passo-à-frente, ao problema da seleção de variáveis na construção de modelos de regressão. Trabalhando com vários níveis de contaminação, estudamos o desempenho do estimador biponderado e de mínimos quadrados em termos dos níveis de significância reais do teste F, e à frequência com que se chega, em cada caso, ao modelo final correto.

Simulações Monte Carlo são feitas para diversas características da distribuição dos erros aleatórios, desde a normal pura até situações caracterizadas por pesada contaminação. Os resultados, apresentados em forma de tabela, permitem um estudo do desempenho de cada um dos dois estimadores no contexto particular em que foram empregados.

No capítulo 4 fazemos uma análise comparativa dos dois estimadores, com base nos dados tabulados apresentados nos capítulos 2 e 3. Esta análise traz evidências de que o estimador biponderado pode ser uma alternativa vantajosa aos mínimos quadrados em alguns contextos particulares de interesse prático.

Í N D I C E

CAPÍTULO 1

REGRESSÃO

1.1 - INTRODUÇÃO

A moderna atividade humana se caracteriza, em diversas áreas, por uma intensa e permanente produção de informação. Esta informação às vezes passa despercebida e se perde, às vezes é coletada e registrada de maneira mais ou menos planejada e sistemática. A análise correta da informação coletada e armazenada pode, não raramente, fornecer elementos importantes ao aprofundamento da compreensão dos processos envolvidos, sejam eles de natureza física, política, econômica, social, entre outros.

A informação, dependendo da situação considerada, pode ser gerada de forma contínua. Em determinados processos industriais por exemplo, variáveis relativas às características de operação do processo e do produto final, como temperatura, pH, índice de viscosidade, etc., são coletados e registrados continuamente. Às vezes, ainda, a informação é produzida de forma discreta, em instantes bem definidos de tempo. De qualquer forma, o estabelecimento de um processo sistemático de coleta e registro de informação deve atender a objetivos claramente definidos. Conforme apontado em Draper e Smith(1982) com frequência um processo sistemático de coleta e registro de informação é mantido por simples questão de hábito ou tradição, quando os objetivos originais para tal procedimento já foram há muito esquecidos ou abandonados.

Assim, a coleta e registro da informação não deve ser vista como tendo uma finalidade em si, mas como uma etapa intermediária de um processo mais amplo, objetivando um entendimento mais detalhado do fenômeno considerado.

Esta compreensão muitas vezes é procurada objetivando orientar decisões imediatas, ou o estabelecimento de planos a médio e longo prazos. Em outros contextos, entretanto, busca-se fundamentação empírica para teorias existentes, ou indicações de novas direções promissoras para investigação.

Exemplo 1.1 - Um grande hospital escola atende diariamente a centenas de pacientes, distribuídos nas diversas especialidades da atividade médica. Num campo onde a realização de pesquisa de forma planejada está fortemente limitada por restrições óbvias de natureza ética, estes pacientes oferecem, a cada uma daquelas especialidades, uma inesgotável base empírica para estudos e pesquisas. O acompanhamento cuidadoso de cada paciente, e a análise conjunta de um grande número de casos semelhantes pode permitir a descoberta de regularidades importantes possibilitando, assim, o avanço de passos cruciais no entendimento e domínio dos processos biológicos envolvidos.

O acompanhamento cuidadoso referido acima inclue, certamente, o registro de um grande número de dados objetivos sobre cada paciente. Estes dados podem se referir à história pregressa, aos hábitos, ao ambiente social, às atividades profissionais, à base genética, à história patológica, ao perfil psicológico, ao estado físico e patológico atual e à evolução destes no tempo, às intervenções médicas efetuadas, etc, para cada paciente.

No planejamento das rotinas de coleta de dados a serem implantadas, a escolha das variáveis que devem ser anotadas não se baseia sempre na certeza, mas frequentemente na suspeita da possibilidade da importância desta para a compreensão do processo biológico sendo acompanhado. Assim o número de variáveis sendo monitoradas é em geral inflacionado, podendo vir a ser realmente grande. Multiplique-se pelo número de pacientes acompanhados, geralmente alto para diversas áreas importantes, e se tem uma volumosa base de dados, em contínuo processo de expansão.

O computador pode aqui desempenhar, certamente, um papel importante no processamento eficiente de tantos dados. Contudo só os métodos de análise exploratória de dados, da seleção de variáveis relevantes, da construção de modelos, etc, permitirão peneirar para fora do processo de coleta as variáveis irrelevantes, simplificando o mesmo. Mais importante ainda, aqueles métodos estatísticos de análise podem permitir, pelo ajuste adequado de modelos, a explicitação de relações importantes que possibilitarão avanços no entendimento do processo biológico em questão.

Neste contexto, em geral as variáveis consideradas podem ser agrupadas em dois tipos : variáveis explicativas e variáveis respostas. Por exemplo, no acompanhamento de gestantes de alto risco, variáveis como idade, peso, pressão arterial, tipo de atividade profissional, grupo racial, nível de renda, hábitos alimentares, etc, referentes à gestante, podem ser consideradas como possíveis variáveis explicativas para respostas tais como : peso, nota de Apgar, etc, referentes ao recém-nascido. Busca-se então, com base em dados históricos acumulados, construir um modelo que permita prever o nível de uma determinada variável resposta, como função de um elenco adequado de variáveis explicativas.

Este processo iterativo de acúmulo e análise de dados, e construção de modelos empíricos, tem sido fator muito importante no progresso verificado em áreas como a pesquisa do câncer, dos transplantes de órgãos, entre outras.

Exemplo 1.2 - Na área de Engenharia de Alimentos é muito comum o processamento de determinadas matérias primas para a sua utilização em alimentos industrializados através da extrusão. Este processamento é feito através de um extrusor, um aparelho que possue vários fatores a serem ajustados pelo operador.

Muitos estudos de laboratório têm sido feitos de forma a determinar como certas características do produto final são explicadas como função das condições de operação do extrusor. Como em geral ocorre em estudos controlados, de laboratório, tem-se aqui uma situação onde determinadas variáveis são mantidas constantes, e um certo conjunto de variáveis são ajustadas, sequencialmente, em pontos de operação fixados através de um delineamento experimental bem definido.

No processo de extrusão de farinha desengordurada de tremoço doce, por exemplo, pode-se estar controlando as seguintes variáveis : temperatura da extrusão, umidade da farinha, diâmetro da matriz e rotação da rosca. Neste contexto, pode-se estudar diversas características, em particular a mastigabilidade do produto extrudado como função destas variáveis sob controle.

O que se busca é ter a variável resposta como função das variáveis controladas de forma a ser possível determinar como a mastigabilidade, por exemplo, responde a perturbações nos valores das variáveis controladas.

Nos exemplos acima, procura-se construir modelos empíricos com base em dados acumulados, seja casualmente, seja através de um delineamento experimental rigorosamente estabelecido, que permita representar uma certa variável resposta, aproximadamente, como uma função de um certo conjunto convenientemente selecionado de variáveis explicativas.

Assim, representando por Y a variável resposta, da qual se tem n observações, e por X1, X2, ..., Xp , as p variáveis explicativas selecionadas, podemos representar sumariamente os dados disponíveis por

$$Yi, Xi1, \ldots, Xip \quad , \text{ para } i = 1, \ldots, n$$

onde Xij é o valor de Xj correspondente à i-ésima observação, Yi.

Em seguida estabelece-se uma conjectura básica com respeito à existência de uma relação funcional entre Y e as p variáveis explicativas X, através do modelo

$$Y = f(X1, \ldots, Xp) + \varepsilon$$

com f pertencendo a uma certa família de funções, e ε o erro aleatório para o qual se pressupõe algumas propriedades de regularidade. O que se busca, em geral, é uma relação funcional que permita, com base no conhecimento dos valores de X1, ..., Xp , prever um valor aproximado para Y.

A escolha da família funcional f, e a determinação do elemento daquela família de funções que melhor se ajusta ao particular conjunto de dados, serão alguns dos temas a serem abordados nas seções seguintes.

1.2- CONCEITUAÇÃO E NOTAÇÃO MATRICIAL

Em diversas situações, ao se tentar ajustar um modelo, como descrito na seção anterior, a forma funcional de f já é conhecida por considerações de natureza teórica. O levantamento de dados experimentais ou observacionais permitirá, além da checagem do modelo teórico adotado, a estimação dos parâmetros envolvidos.

Em uma outra situação mais comum, a forma funcional verdadeira de f não só é desconhecida, mas ainda de pouco interesse. Nestes casos busca-se simplesmente uma aproximação satisfatória que, baseada unicamente nos dados empíricos, permitirá uma certa capacidade predi-

tiva para Y, dentro de uma faixa limitada e bem definida de valores das variáveis X.

Neste segundo contexto é comumente utilizado o modelo geral de regressão linear. Esta classe de modelos é atraente por sua simplicidade e por aproximar geralmente bastante bem as funções, quando restritas a regiões limitadas, como é o caso em diversas situações práticas de interesse.

No modelo geral de regressão linear a equação ajustada é da forma :

$$Yi = \beta 0 + \beta 1 Xi1 + \beta 2 Xi2 + \ldots + \beta p Xip + \varepsilon i \qquad (1)$$

onde Yi = i-ésima observação conhecida também chamada de variável dependente e,

Xij = valor da j-ésima variável para a i-ésima observação, também conhecido.

Os β's são os p+1 parâmetros desconhecidos que se quer estimar, e os εi's são os erros aleatórios, também desconhecidos.

A denominação linear provém do fato de que a relação é linear nos parâmetros β's. As variáveis podem ser, e muitas vezes são, transformações, não necessariamente lineares, das variáveis originais do problema.

Existem maneiras de se trabalhar com esses modelos para especificações bastante gerais do comportamento dos erros aleatórios, como por exemplo com os chamados modelos lineares generalizados, apresentados por Nelder e Wedderburn (1972). Há um texto de Cordeiro (1986), em português, sobre esse tipo de abordagem. Aqui serão considerados apenas os casos em que esses erros são variáveis aleatórias de médias zero, não correlacionados e de variância constante, como especificado de

forma precisa nas próximas seções. Para aplicar as técnicas de seleção é necessária ainda a suposição de normalidade dos erros. As técnicas de seleção com uso do método de mínimos quadrados dependem fortemente dessa suposição.

O problema pode ser formulado em forma matricial :

$$Y = X\beta + \varepsilon$$

onde

$$Y' = (Y1 \ldots Yn)$$

é o vetor de observações, e

$$X = \begin{vmatrix} 1 & X11 & \ldots & X1p \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & Xn1 & \ldots & Xnp \end{vmatrix}$$

é a matriz dos valores das variáveis X's. A primeira coluna de 1's corresponde à inclusão da constante no modelo e cada uma das colunas restantes corresponde aos valores de uma destas variáveis para as n observações,

$$\beta' = (\beta o \ldots \beta p)$$

é o vetor de parâmetros, a serem estimados e

$$\varepsilon' = (\varepsilon 1 \ldots . \varepsilon n)$$

é o vetor de erros, desconhecido.

Neste trabalho será considerado apenas o caso em que a matriz X tem as p+1 colunas linearmente independentes. Esta situação de posto completo da matriz X é conhecida usualmente como problema de regressão linear múltipla.

1.3 - CRITÉRIOS PARA AJUSTE DO MODELO

Tendo-se caracterizado o modelo, devemos procurar a maneira de se obter as estimativas dos parâmetros envolvidos. É natural que se deva buscar estimativas que, de alguma forma, minimizem os erros de ajuste, isto é, a diferença entre os valores observados e respectivos valores ajustados de Y. Deste princípio básico decorrem inúmeros critérios de ajuste.

Por exemplo, pode-se adotar o critério de se procurar minimizar a soma dos desvios absolutos, ou ainda, o de minimizar o erro absoluto máximo. Alguns critérios possuem maior apelo intuitivo que outros. Contudo, o critério com maior apelo intuitivo, não é necessariamente aquele que implica num tratamento matemático mais simples. Por exemplo, o critério de mínimos desvios absolutos é mais intuitivo do que o critério de mínimos quadrados, no entanto este possibilita um tratamento matemático mais simples e elegante.

A elegância do tratamento matemático e a simplicidade relativa dos cálculos é um dos fortes motivos pelos quais o critério de mínimos quadrados tem sido o mais difundido para ajuste de modelos de regressão. Além disso, sob a suposição de normalidade dos erros, frequentemente satisfatória em problemas de regressão, as estatísticas envolvidas têm, pelo menos nos contextos mais simples, distribuições conhecidas.

Mais recentemente, com o avanço da indústria de computadores permitindo acesso cada vez mais generalizado e barato a amplos recur-

sos de cálculo, a simplicidade dos cálculos associados ao método de mínimos quadrados vem deixando de ser um fator tão decisivo. Métodos mais dependentes de cálculos complexos, como o dos mínimos desvios absolutos, têm apresentado apelo crescente. A este respeito ver Bloomfield e Steiger (1983). Por outro lado, o desenvolvimento dos métodos estatísticos computacionalmente intensivos, vem oferecendo opções para testes de hipóteses estatísticas que compensam, satisfatoriamente, a simplicidade e elegância das propriedades estatísticas dos estimadores de mínimos quadrados, no contexto de erros aleatórios distribuídos normalmente.

1.4 - O CRITÉRIO DE MÍNIMOS QUADRADOS E SUAS PROPRIEDADES

O método dos mínimos quadrados permite um tratamento matemático e, sob certas suposições gerais, estatístico, simples e elegante. Por estes motivos ele ganhou amplo predomínio sobre seus competidores. De forma esquemática, o método de mínimos quadrados consiste em tomar para estimativas dos β's, valores que minimizem

$$\sum_{i=1}^{n} (Yi - \hat{Y}i)^2$$

onde $\hat{Y}i$ é o valor ajustado para Yi com a estimativa b, dado por

$$\hat{Y}i = bo + bl.Xli + \ldots + bp.Xpi \ .$$

Denominando de resíduos as diferenças entre o valor das observações Yi e os valores correspondentes ajustados, $\hat{Y}i$, ou seja

$$ei = Yi - \hat{Y}i$$

o que se busca são, pois, valores dos b's, que minimizem a soma de quadrados dos resíduos. Em notação matricial,

$$SQRes = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - X\beta)'(Y - X\beta) \quad (2)$$

Derivando-se a expressão acima em relação a β, igualando-se a zero e representando por b o vetor com as estimativas, tem-se :

$$(\partial/\partial\beta) \quad (Y - X\beta)'(Y - X\beta) = 0$$

O que leva às equações normais

$$X'Xb = X'Y,$$

cuja solução, quando (X'X) é não singular, é dada por

$$b = (X'X)^{-1}X'Y,$$

que, pode-se mostrar, corresponde ao mínimo de (2). Esta é conhecida como a solução de mínimos quadrados para β. No caso de regressão (matriz X de posto completo) esta solução é única (Scheffé (1959)).

As propriedades estatísticas dos estimadores de mínimos quadrados dependerão, naturalmente, do comportamento estatístico dos erros. Neste ponto introduz-se duas suposições simplificadoras razoáveis sobre este comportamento. A primeira assume que

$$i) \quad E(\varepsilon) = 0 \ ,$$

ou seja, que Y é uma aproximação não tendenciosa de $f(X1,\ldots,Xp)$. A segunda diz respeito à variância de ξ. Neste caso faz-se, de início, a suposição mais simples possível, ou seja, a de que

$$ii) \quad Var(\varepsilon) = \sigma^2 I \ ,$$

isto é, a variância é constante independente dos valores das variáveis X's, e os erros são não correlacionados.

No modelo (1), assumindo-se (i) e (ii) valem as seguintes propriedades para os elementos do vetor b :

$$E(b) = \beta$$

$$Var(b) = \sigma^2 (X'X)^{-1}$$

O teorema de Gauss-Markov garante que os elementos de b são os estimadores lineares não viciados de mínima variância (BLUE) dos elementos de β (Scheffé (1959)).

Se, além das duas suposições anteriores, tem-se ainda que é satisfeita a hipótese de normalidade :

$$iii) \qquad \varepsilon \sim N(0, \sigma^2 I)$$

então os estimadores b's são também os estimadores de máxima verossimilhança para os β's.

Após obtido o ajuste deve-se procurar maneiras de se verificar a qualidade deste. Existem muitas abordagens para esse estudo, entre elas a análise dos resíduos - incluindo confecção de gráficos dos mesmos contra valores ajustados e as variáveis X's -, o uso de técnicas de diagnóstico, como as descritas por Belsley, Kuh e Welsch (1980) e muitas outras. No contexto em que vamos trabalhar, com seleções automáticas de variáveis, usam-se, geralmente, apenas as mais simples, constituídas por exame de R2, o coeficiente de correlação múltipla ao quadrado, exame da tabela de análise de variância geral e testes de hipóteses.

O coeficiente de correlação múltipla ao quadrado, R2, é definido como :

$$R2 = \frac{SQReg}{SQTot}$$

onde SQReg é a soma de quadrados da regressão dada por

$$SQReg = \sum_{i=1}^{n} (\hat{Y}i - \bar{Y})^2$$

e SQTot é a soma de quadrados total corrigida

$$SQTot = \sum_{i=1}^{n} (Yi - \bar{Y})^2$$

sendo que

$$SQReg + SQRes = SQTot$$

Estes valores de somas de quadrados e os correspondentes graus de liberdade, bem como os quocientes das somas de quadrados por estes últimos, chamados de quadrados médios, são comumente agrupados em uma tabela de análise de variância como a que é apresentada abaixo.

Uma Tabela de Análise de Variância

fonte de variação	graus de liberdade	soma de quadrados	quadrado médio
Regressão	p	SQReg	QMReg = SQReg/p
Resíduo	n-p-1	SQRes	QMRes = SQRes/(n-p-1)
TOTAL	n-1	SQTot	

Válida a suposição (iii), pode-se fazer inferências estatísticas sobre cada parâmetro individualmente, ou sobre combinações lineares dos mesmos, uma vez que se dispõe de estatísticas com distribuições conhecidas.

Uma questão frequente neste contexto consiste em testar se o coeficiente de uma determinada variável é zero, ou seja, testar

$$H0 : \beta i = 0$$
$$\text{vs}$$
$$H1 : \beta i \neq 0$$

para algum i = 1, ..., p

Seja dii o i-ésimo elemento da diagonal de $(X'X)^{-1}$. Temos então que sob a hipótese nula :

$$\frac{bi}{\sigma \sqrt{dii}} \sim N(0,1)$$

Como σ^2 é desconhecido é necessário ter uma estimativa S^2 para este parâmetro. Como $E(SQRes) = E(\sum_{i=1}^{n}(Yi - \hat{Y}i)^2) = (n - p - 1)$ temos que $S = SQRes / (n-p-1)$, é um estimador não viciado de σ^2. Como S^2 e cada um dos bi são independentes e S^2/σ^2 tem distribuição qui-quadrado com n-p-1 graus de liberdade, então, sob a hipótese nula H0 :

$$\frac{bi}{S \sqrt{dii}}$$

tem distribuição t de Student com n-p graus de liberdade, pois é o quociente entre duas variáveis aleatórias independentes, com numerador distribuído segundo uma N(0,1) , e o quadrado do denominador distribuído segundo uma χ^2 com n-p graus de liberdade.

Uma maneira equivalente de se realizar o mesmo teste baseia-se no acréscimo na soma de quadrados dos resíduos devido à restrição $\beta i = 0$, dividido pelo quadrado médio dos resíduos. Com as suposições feitas, sob a hipótese nula, temos que esta estatística segue uma dis-

tribuição F de Snedecor com 1 e n-p graus de liberdade. Esta estatística pode ser obtida a partir da seguinte tabela de análise de variância :

fonte de variação	graus de liberdade	soma de quadrados	quadrado médio
X(i)	p-1	SQ (X(i))	QM (X(i)) = SQ (X(i))/(p-1)
Xi	1	ASQ ($\beta i = 0$)	ASQ ($\beta i = 0$)
RESÍDUO	n-p-1	SQRes	QMRes = SQRes/(n-p-1)
TOTAL	n-1	SQTot	

onde

SQ (X(i)) = soma de quadrados da regressão com a matriz X não tendo a i-ésima coluna.

$$ASQ\ (\beta i = 0) = SQReg - SQ\ (X(i))$$

$$SQTot = \sum_{i=1}^{n} (Yi - \bar{Y})^2 \quad e$$

$$SQ\ (X(i)) + ASQ\ (\beta i = 0) + SQRes = SQTot$$

e é dada por

$$\frac{ASQ\ (\beta i = 0)}{S^2} = \frac{bi \cdot bi}{S^2 \cdot dii}$$

Esta estatística é exatamente igual ao quadrado da anterior.

1.5 - MÍNIMOS QUADRADOS PONDERADOS

Muitas vezes dispõe-se de um conjunto de dados onde nem todas as observações têm a mesma confiabilidade. Considere por exemplo a situação comum em que o valor da variável Y é determinado através de análises laboratoriais. Neste caso, pode ocorrer que as análises sejam feitas por dois ou mais técnicos diferentes, sendo um deles mais experiente na operação dos instrumentos envolvidos, produzindo, consequentemente, resultados mais acurados e confiáveis. Parece claro que esta diferença de confiabilidade dos dados deva ser transmitida ao ajuste de forma que uma observação mais confiável venha a ter um peso maior na determinação dos valores estimados dos parâmetros.

Em termos das suposições básicas sobre ε, esta diferença no grau de confiabilidade associado a valores diferentes de Y pode ser modelada, relaxando-se a suposição (iii) de variância constante. Em termos matriciais, podemos dizer que $Var(Y) = V\sigma^2$, onde V é uma matriz diagonal, positiva definida mas diferente da identidade.

Avançando mais um passo, podemos considerar o caso em que $Var(Y) = V\sigma^2$, onde V é uma matriz positiva definida qualquer, não necessariamente diagonal. Isto ocorrerá no caso em que os ε's forem correlacionados.

Tem-se então que o modelo e suas hipóteses são agora :

$$Y = X\beta + \varepsilon$$

com

a) $E(\varepsilon) = 0$

b) $Var(\varepsilon) = V\sigma^2$

e

c) $\varepsilon \sim N\ (0, \sigma^2 V)$

onde V é uma matriz positiva definida suposta conhecida. Existe então uma única matriz P, simétrica, não singular tal que :

$$P'P = PP = P^2 = V$$

Considerando $\varepsilon^* = P^{-1}\varepsilon$

tem-se que

$$E(\varepsilon^*) = 0$$

e

$$Var(\varepsilon^*) = E(\varepsilon^*\varepsilon^{*\prime}) = E(P^{-1}\varepsilon\ \varepsilon' P^{-1}) = P^{-1}PPP^{-1}\sigma^2 = I\sigma^2$$

Como ε^* é uma combinação linear dos elementos de ε e ε é normalmente distribuído, tem-se que ε^* também é normalmente distribuído.

Pré-multiplicando-se a equação (1) por P^{-1} chega-se a um novo modelo :

$$P^{-1}Y = P^{-1}X\beta + P^{-1}\varepsilon$$

ou definindo $Z = P^{-1}Y$, $Q = P^{-1}X$, e $\varepsilon^* = P^{-1}\varepsilon$, reescreve-se :

$$Z = Q\beta + \varepsilon^*$$

onde * satisfaz às hipóteses (i), (ii) e (iii). Então, pelos resultados apresentados na seção anterior, o estimador b de mínimos quadrados para β, neste modelo mais geral, pode ser obtido a partir das equações

$$Q'Qb = Q'Z$$

ou seja

$$X'P^{-1\prime}P^{-1}Xb = X'P^{-1\prime}P^{-1}Y$$

$$X'V^{-1}Xb = X'V^{-1}Y \quad ,$$

denominadas equações de Aitken, e que levam à solução

$$b = (X'V^{-1}X)^{-1} X'V^{-1}Y$$

pois $(X'V^{-1}X)$ é não singular.

Neste caso a tabela de análise de variância fica

fonte de variação	graus de liberdade	soma de quadrados	quadrado médio
Regressão	p	SQReg	QMReg = SQReg/p
Resíduo	n-p-1	SQRes	QMRes = SQRes/(n-p-1)
TOTAL	n-1	SQTot	

onde

$$SQReg = \sum_{i=1}^{n} (\hat{Y}i - \bar{Y})^2$$ sendo

$\hat{Y}i = b0 + \sum_{i=1}^{p} bjXij$ com bj's soluções das equações de Aitken

$$SQTot = \sum_{i=1}^{n} (Yi - \bar{Y})^2$$

e

$$SQReg + SQRes = SQTot$$

1.6 - SELEÇAO DE VARIAVEIS

O problema básico de Análise de Regressão consiste em, dada uma variável Y e uma série de p variáveis X1, ... , Xp ajustar, com base em dados empíricos, um modelo que aproxime Y como uma função das variáveis X. Em particular, o problema pode consistir em encontrar p constantes b1, ... , bp de forma a que o modelo Y = bo + b1.X1 + ... +

bp.p, tenha a maior aproximação possível de Y em termos da soma dos quadrados dos desvios.

Este problema tem solução elementar, dadas certas condições básicas. Contudo em diversas situações práticas de interesse, o número p de variáveis independentes X é muito grande, incluindo possivelmente algumas que não devam entrar no modelo, seja por possuir muito pouco poder explicativo, seja por serem redundantes com outras variáveis mais importantes. O problema aqui consiste em, antes de ajustar o modelo, decidir quais variáveis X devem participar do mesmo.

A busca por um subconjunto das variáveis X que participarão do modelo pode ser feita de forma mais ou menos artesanal, quando p não for muito grande. O número de modelos possíveis quando se dispõe de p possíveis variáveis explicativas é de 2^p, indo desde aquele que não inclue nenhuma até o que inclue todas as p variáveis. Para p pequeno, não é impraticável testar, um a um, todos os modelos possíveis e escolher aquele que melhor atende a certos critérios de ajustamento e parcimônia previamente estabelecidos. O princípio da parcimônia, na construção de modelos de regressão, em geral, requer o emprego de um número menor possível de variáveis.

Em estudos na área de saúde, por exemplo, pode-se querer construir um modelo, com base em dados empíricos acumulados, que permita estimar o tempo de recuperação após um certo tipo de cirurgia, como função de algumas variáveis relativas ao quadro geral passado e presente de um determinado paciente. Em certos processos industriais pode-se ter uma grande quantidade de dados históricos disponíveis sobre as condições de operação do processo para diversos grupos de itens produzidos, e se desejar estudar alguma característica do produto como função das condições de operação da máquina.

Em ambos os casos pode-se dispor, de início, de uma grande quantidade de variáveis candidatas ao modelo. O problema consiste em escolher um subconjunto destas variáveis que produza um modelo que

possa ser considerado bom, no sentido de ser parcimonioso e com bom poder explicativo.

Deve-se lembrar aqui que o critério de ajuste baseado somente no coeficiente de determinação R2 é inadequado, pois este sempre aumenta com a inclusão de novas variáveis. Contudo, sabe-se que num processo de inclusão sequencial de variáveis no modelo, a partir de certo ponto as estimativas dos desvios padrões dos estimadores crescem rapidamente. Além disso, em termos de uso prático, é inconveniente trabalhar com modelos que incluam um número muito grande de variáveis no ajuste.

Deve-se buscar um compromisso, um ponto intermediário satisfatório, entre parcimônia e explicabilidade. Em geral existem várias alternativas possíveis de combinações das variáveis que produzem resultados bastante parecidos. Pode não fazer muito sentido falar em obter o "melhor" modelo. O que se quer é chegar a um conjunto de bons modelos para, em seguida, escolher entre os mesmos o mais conveniente, inclusive segundo critérios subjetivos, como de melhor interpretabilidade física, ou por incluir uma ou mais variáveis que por alguma razão se quer no modelo.

Os métodos para se chegar ao modelo adequado são vários. Os mais comumente empregados são :

- Método Artesanal - Consiste em usar primeiro uma ou duas variáveis que por alguma razão devam estar no modelo e ir incluindo outras, a partir desse ponto, usando critérios de decisão baseados, por exemplo, em gráficos de resíduos de regressões parciais de Larsen e McCleary (1972).

- Método de Todas as Regressões - Consiste em fazer todas as 2^p regressões possíveis e usar algum critério de escolha como o do Cp de Mallows ou o do Quadrado Médio do Resíduo entre outros.

- Métodos da seleção automática.

É fácil perceber as dificuldades associadas ao método artesanal quando o número de variáveis candidatas for grande. Na realidade sua utilidade já é bem limitada quando se tem mais do que cinco ou seis variáveis. Em geral, na prática, o processo de seleção de variáveis incluirá aspectos de todos os três tipos de abordagem acima, como sugerido por exemplo por Mosteller e Tukey (1977) e Dachs (1978).

Neste trabalho serão considerados apenas alguns métodos de seleção automática muitas vezes designados como processos "stepwise". Na realidade convém distinguir três maneiras de fazer seleção automática: os métodos para a frente (forward), os para trás (backward) e os por passos (vai e vem, stepwise propriamente dito). Estes métodos serão melhor apresentados na seção a seguir.

Uma boa apresentação de diversos métodos de seleção, com discussão sobre os mesmos, pode ser encontrada em Hocking (1976). Uma comparação entre diversos procedimentos com relação à qualidade de predição e estimação de parâmetros pode ser encontrada em Hoerl, Schuenemeyer e Hoerl (1986).

1.7 - SELEÇÃO PASSO-A-FRENTE

Os procedimentos automáticos citados anteriormente são bastante difundidos, mas na sua utilização deve-se sempre ter em mente suas limitações. Esses procedimentos, como poderá ser visto mais adiante, são baseados em estatísticas cujas distribuições dependem da distribuição dos erros. Os valores críticos para os testes de hipóteses associados são retirados da distribuição F de Snedecor, embora se saiba que, neste caso, esta não seja a distribuição verdadeira para aquelas estatísticas.

De forma a melhor situar as estatísticas de teste envolvidas apresentamos a seguir um detalhamento de três procedimentos de seleção automática, um dos quais, o passo-à-frente, de especial interesse neste trabalho.

i) O procedimento passo-à-frente .

Esse procedimento consiste em incluir variáveis no modelo uma a uma. A variável candidata a ser incluída, em um determinado p asso, é aquela cuja inclusão no modelo forneça uma maior explicabilida de da variabilidade na variável resposta. Uma medida disto é obtida através do acréscimo na soma de quadrados do resíduo provocado pela retirada desta variável no modelo (ASQ ($\beta i = 0$)).

O processo se inicia com apenas a constante no modelo e inclue uma variável por vez, até qua todas as variáveis tenham sido incluídas ou que seja satisfeito um critério de parada fixado a priori. A variável considerada para inclusão é aquela que gera o maior valor F entre aquelas variáveis que não tenham ainda sido incluídas. Este valor é calculado da seguinte maneira :

$$Fi = \frac{ASQ(\beta i = 0)}{QMRes}$$

onde estes valores são obtidos da tabela de análise de variância encontrada na seção 2

De maneira esquemática temos o procedimento descrito a seguir

Considere-se que m, $0<m<p$, variáveis já tenham sido incluídas no modelo, e supondo, sem perda de generalidade, que estas variáveis sejam X1, ..., Xm.

i - Toma-se i = m+1

ii - Ajusta-se o modelo de regressão pelo método de mínimos quadrados, para Y, X1, X2, ..., Xm, Xi.

iii - Calcula-se o acréscimo na soma de quadrados do resíduo devido à restrição ($\beta i = 0$) denominado ASQ ($\beta i = 0$).

iv - i = i + 1

v - Se i + 1 > p vá para (vi) caso contrário vá para (ii)

vi - Definindo-se $Fi = bi^2 / (S^2.dii)$ determina-se $F = Max\ (Fi)$ e o i correspondente onde $m<i<p+1$

bi = estimativa de mínimos quadrados de βi

dii = i-ésimo elemento da diagonal de $(X'X)^{-1}$ e

S^2 = estimativa para σ^2 a partir da SQRes do ajuste com X1, X2, ..., Xm, Xi

vii - Compara-se F = Fimax, com o valor de referência da tabela da distribuição F de Snedecor com 1 e n-m-2 graus de liberdade. Se F for maior que o valor de referência, inclue-se a variável Ximax no modelo e inicia-se novo estágio. Se for menor, encerra-se o processo de seleção de variáveis, com o modelo final envolvendo as variáveis X1, X2, ..., Xm

ii) O procedimento passo-à-trás.

Este procedimento é reverso do anterior, o passo-à-frente, pois ele faz a seleção do modelo por exclusão de variáveis ao invés da inclusão. O processo se inicia com todas as variáveis incluídas no modelo e estas serão excluídas ou não com base nas mesmas estatísticas Fi do procedimento anterior. Neste caso a variável que é ou não excluída é aquela que origina o menor F entre todas as variáveis que ainda estão no modelo. Temos então, que a variável i é retirada do modelo que se encontra com p variáveis incluídas se

$$F = \min_i \frac{SQResp(i) - SQResp}{QMResp} = \frac{ASQ(\beta i = 0)}{QMResp} < Fsai$$

onde SQResp = soma de quadrados do resíduo com o modelo com as p variáveis

QMResp = quadrado médio do resíduo com o modelo com as p variáveis

SQResp(i) = soma de quadrados do resíduo com o modelo com as p variáveis a menos da variável i.

Fsai = valor de corte fixado a priori.

iii) O procedimento passo-à-passo.

Esta técnica vem a ser uma combinação das duas técnicas apresentadas anteriormente. Inicialmente tem-se o modelo apenas com a constante. Passa-se então a tentar incluir variáveis usando um passo-à-frente, e após ter sido feita a inclusão de uma variável, verifica-se se há alguma variável a ser excluída com um passo-à-trás.

Um dos problemas encontrados na utilização destas técnicas consiste em escolher o valor de corte adequado para a inclusão e exclusão de variáveis. A distribuição de referência utilizada é a F de Snedecor, embora seja fato sabido de que esta não seja a distribuição real das estatísticas em questão.

A questão da distribuição exata da estatística F tem recebido uma atenção regular na literatura graças ao interesse pelos métodos de seleção de variáveis. Chamando a atenção para o problema, Draper, Guttman e Kanemasu (1971), mostram que os testes baseados na distribuição F de Snedecor, como distribuição de referência para as estatísticas F, subestimam a cauda da distribuição verdadeira, produzindo níveis de significância reais maiores que os nominais. Estas distorções podem ser realmente graves. Posteriormente, uma série de artigos como Diehr e Hoflin (1974), Rencher e Pun (1980), Wilkinson e Dallal (1981), trazem resultados baseados em simulações Monte Carlo, em contextos particulares, levando a conclusões semelhantes.

Os três procedimentos descritos acima não levam necessariamente ao mesmo modelo final. É comum ainda a situação onde o modelo selecionado não necessariamente é aquele que dá a menor soma de quadrados do resíduo para um dado tamanho de subconjunto. Entretanto, as diferenças entre os procedimentos, não são significativas na prática como demonstrado por Berk (1978).

Já que as diferenças nos modelos encontrados não são significativas, faz sentido dar preferência ao passo-à-frente já que os aspectos computacionais envolvidos são mais simples e eficientes do que diversos outros procedimentos. Além disso, o critério de parada para o procedimento é fácil de ser especificado.

Convém ressaltar aqui, que procedimentos automáticos devem ser utilizados com a cautela devida. O fato da seleção ser feita à revelia do analista faz com que o modelo encontrado possa ser de difícil

interpretação física. Do mesmo conjunto de variáveis candidatas se poderia conseguir um outro subconjunto também bom que fizesse muito mais sentido físico.

A seleção automática por passos tem um papel fundamental em fazer uma boa triagem inicial quando se tem um conjunto grande de variáveis disponíveis. De posse de um conjunto menor a análise por métodos mais detalhados é possível e deve ser sempre procurada.

CAPÍTULO 2

O ESTIMADOR BIPONDERADO

2.1 - INTRODUÇAO

No capítulo anterior, apresentou-se o critério de mínimos quadrados, o critério mais difundido, como uma maneira para se ajustar um modelo. Este capítulo trata de um outro critério para ajuste de modelos, o estimador biponderado em regressão. De forma a torná-lo mais compreensível ele será apresentado a partir de um contexto elementar : o estimador biponderado de Tukey para parâmetros de locação (Beaton e Tukey (1974) e Scafi (1979)).

Na análise estatística de um conjunto de dados (y1, ..., yn), proveniente de uma distribuição F, uma das questões fundamentais consiste na determinação, a partir da informação disponível, do parâmetro de locação daquela distribuição.

A média aritmética, $\bar{Y} = (\sum_{i=1}^{n} yi)/n$, é o estimador de locação mais utilizado. Além de seu forte apelo intuitivo, e da facilidade dos cálculos envolvidos, $\bar{Y}$ possue ainda a propriedade de ser o estimador de mínimos quadrados. Isto quer dizer que $\bar{Y}$ é a solução para

$$\min_{T} \sum_{i=1}^{n} (yi - T)^2 .$$

Desta propriedade decorrem outras muito importantes. Em particular, para uma ampla família de distribuições, que inclue a Normal, o estimador de mínimos quadrados é não viciado uniformemente de mínima variância para a esperança de Y.

Contudo o critério de mínimos quadrados possui competidores. Na verdade, pode-se estabelecer um número ilimitado de alternativas com algum sentido físico. Por exemplo, o critério dos mínimos desvios absolutos faz tanto ou mais sentido intuitivo que o dos mínimos quadrados, tendo inclusive sido considerado antes (Bloomfield e Steiger(1983)).

Historicamente, o critério de mínimos quadrados ganhou predominância total sobre seus competidores devido à simplicidade de cálculo dos estimadores derivados, e das convenientes propriedades estatísticas destes.

Mais recentemente, com o barateamento e generalização do uso de computadores, as razões por trás do amplo predomínio do critério de mínimos quadrados enfraqueceram muito. Consequentemente tem crescido o interesse por critérios mais sofisticados, que implicam em cálculos mais complexos e consequentemente em um uso intensivo de recursos computacionais.

Destes novos critérios têm surgido estimadores de locação que se constituem em alternativas vantajosas à média amostral, em diversas situações especiais. Por exemplo, em presença da possibilidade de erros grosseiros, a média amostral é muito pouco estável, sendo em geral vantajosamente substituída pela mediana amostral.

2.2 - ESTIMADORES DE LOCAÇÃO

O parâmetro de locação de uma determinada população é foco de muitas atenções em situações práticas, o que justifica a frequência com que questões relacionadas à estimação deste parâmetro são abordados na literatura.

Porque não utilizarmos sempre a média aritmética dos dados para estimar a média populacional? A média é muito sensível a valores aberrantes, o que faz com que um único erro grosseiro, possa influir fortemente na qualidade de uma estimativa. Este problema sugere, então, a busca de alternativas que façam com que erros grosseiros venham a ter pouca influência, pouco "peso", no valor da estimativa para o parâmetro de locação.

Para várias distribuições comumente encontradas, não temos garantido que a média amostral seja um estimador não viciado uniformemente de mínima variância para a média populacional. Um destes casos é o de uma normal contaminada por uma outra normal de variância maior. Temos que à medida que cresce a variância da normal contaminante, a média também vai tendo a sua variância aumentada. Têm sido bastante estudados estimadores para, nestes casos, estimar também a média populacional mas com uma variância menor do que a da média amostral.

Uma maneira natural de se tratar de problemas como os dois citados, o da possibilidade da presença de erros grosseiros e o da normal contaminada, consiste em estabelecer uma função de peso para as observações, decrescente com a distância ao centro dos dados. Um exemplo extremo é a mediana. Esta dá peso não nulo no máximo às duas observações mais centrais. Para um estudo amplo e abrangente sobre estimação de um parâmetro de locação ver Andrews et al.(1972). Um texto introdutório em português é Bustos(1981).

2.3 - ESTIMADORES DO TIPO M

No estudo de estimadores de locação procurou-se agrupá-los e classificá-los por propriedades e características semelhantes. Dentre as diferentes classes temos os estimadores do tipo M. Uma estatística T é dita ser um estimador do tipo M para θ se :

$\sum_{i=1}^{n} \rho(\, yi;T\,)$ for mínimo ou equivalentemente se

$\sum_{i=1}^{n} \Psi(\, yi;T\,) = 0$ onde $\rho\,(y;\theta)$ é uma função arbitrária e

$\Psi\,(y;\theta) = (\partial/\partial\theta)\,\rho\,(y;\theta)$

Para se ter asseguradas boas propriedades assintóticas para T, fazem-se necessárias algumas suposições de regularidades para a função ρ. A este respeito ver Huber (1981).

No caso de um estimador de locação, o problema é então encontrar T tal que

$\sum_{i=1}^{n} \rho\,((yi - T)/S)$ seja mínimo ou analogamente,

$\sum_{i=1}^{n} \Psi\,((yi - T)/S) = 0$, sendo S uma medida de escala.

A expressão $\sum_{i=1}^{n} \Psi((yi - T)/S)$ pode ser reescrita como

$$\sum_{i=1}^{n} wi.((yi - T)/S)$$

onde $wi = \dfrac{\Psi((yi - T)/S)}{((yi - T)/S)}$,

o que nos possibilita ver T como sendo uma média ponderada das observações: $T = \dfrac{\sum_{i=1}^{n} wi\ yi}{\sum_{i=1}^{n} wi}$, com os pesos wi's calculados a partir dos dados.

Seja $f(Y;\theta)$ a função de densidade de probabilidade de Y. Quando temos $\rho(yi;\theta) = -\log f(yi;\theta)$ o estimador encontrado para θ, vem a ser o estimador de máxima verossimilhança, de onde provém a denominação da classe.

Vemos assim que a média amostral, o estimador de mínimos quadrados para a média populacional, é um caso particular de estimador do tipo M, com $\rho(x) = x^2$ e $\Psi(x) = 2x$. O peso de cada observação neste caso é igual a 1.

Outros exemplos de estimadores do tipo M são :

i) Estimador linear por partes (Andrews et al.(1972))

Neste caso temos :

$$\Psi(x) = -\Psi(-x) = \begin{cases} x & \text{se } 0<x<a \\ a & \text{se } a<x<b \\ (c-x)/(c-b).a & \text{se } b<x<c \\ 0 & \text{se } x>c \end{cases}$$

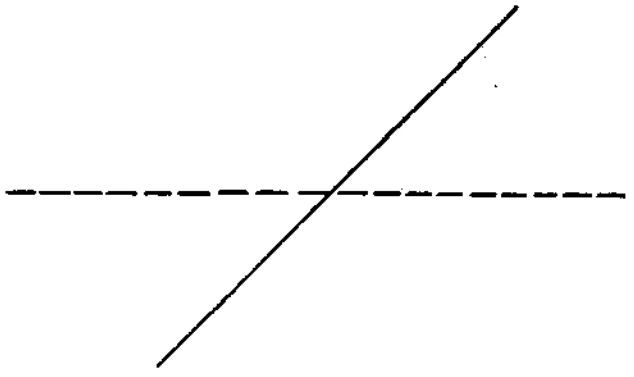
onde para cada escolha de $a<b<c$ temos um diferente estimador

ii) Estimador seno (Andrews et al.(1972)).

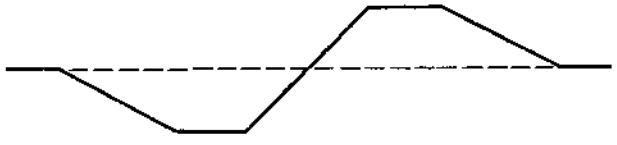
$$\Psi(x) = \begin{cases} \text{sen}(x) & \text{se } -\pi<x<\pi \\ 0 & \text{c.c.} \end{cases}$$

De forma a se evitar distorções devido a escala sempre usamos como argumentos as observações padronizadas.

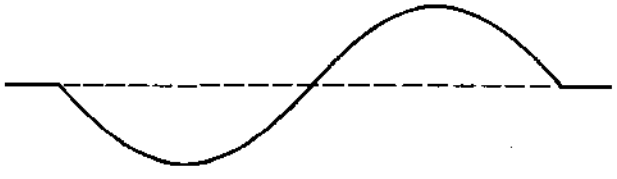
Uma maneira de se comparar diferentes estimadores quanto ao peso dado às observações é estudar suas curvas de influência. Estas curvas mostram como varia o valor da estimativa, quando uma observação



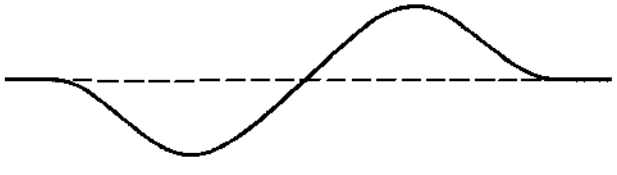
média



linear por partes



seno



biponderado

Figura 2.1 - Curvas de influência de alguns estimadores de locação

percorre todos os valores possíveis, enquanto as demais permanecem fixadas. Esta é uma forma de se quantificar a sensibilidade de um estimador a variações no valor de uma observação. As curvas de influência para alguns estimadores de locação podem ser observadas na figura 2.1.

2.4 - O ESTIMADOR BIPONDERADO

A grande maioria dos estimadores do tipo M de locação proposto na literatura segue o princípio de dar a cada observação pesos decrescentes com a distância desta ao "miolo" dos dados. Esses estimadores diferenciam-se entre si pela forma com que estabelecem estes pesos. É natural que para se ter uma medida de distância coerente com a estrutura própria dos dados, deve-se conhecer um parâmetro de escala da população, ou pelo menos uma estimativa satisfatória deste, a qual denominaremos de S.

Um estimador do tipo M de particular interesse neste trabalho é o estimador biponderado proposto por Beaton e Tukey(1974). No contexto de estimação de parâmetro de locação ele é definido por :

$$\rho'(x) = \Psi(x) = \begin{cases} x(1-x^2)^2 & \text{se } |x| < 1 \\ 0 & \text{caso contrário} \end{cases}$$

com $x = (y - T)/c.S$, onde T é uma medida de locação e S é uma medida de escala das observações, e c uma constante positiva arbitrária.

$$\text{Consequentemente temos } wi = \begin{cases} (1-ui^2)^2 & \text{se } |ui| < 1 \\ 0 & \text{caso contrário} \end{cases}$$

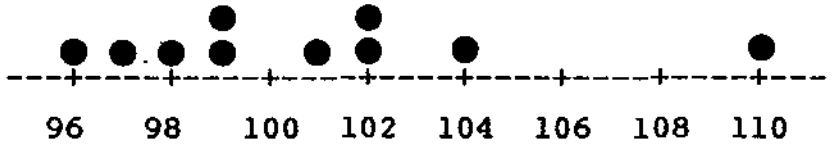
com $ui = (yi - T)/c.S$ sendo T, c e S como definidos acima.

O princípio dos pesos decrescentes com a distância é satisfeito. A função de pesos é suave e contínua, sendo nula a partir de certo ponto.

Vejamos um exemplo de como funciona o estimador biponderado. Vamos neste caso utilizar a constante c = 3.0 e para estimar o utilizamos S = (amplitude interquartis)/1.35.

Exemplo 2.1 : Considere a seguinte amostra já ordenada de tamanho 10

96 97 98 99 99 101 102 102 104 110



Neste caso tem-se então :

$$S = (3^{o}\text{quartil} - 1^{o}\text{quartil})/1.35 = (103 - 97.5)/1.35 = 4.074$$

Para valor inicial do processo iterativo se utilizará a média

$$\bar{Y} = 100.8$$

Na tabela a seguir vê-se os pesos atribuídos às observações nas diferentes iterações :

yi	$u_i^{(1)}$	$w_i^{(1)}$	$w_i^{(1)}.y_i$	$w_i^{(2)}$	$w_i^{(2)}.y_i$	$w_i^{(3)}$
96	-0.3927	0.7154	68.67	0.7827	75.14	0.7979
97	-0.3109	0.8160	79.16	0.8716	84.55	0.8837
98	-0.2291	0.8978	87.98	0.9391	92.03	0.9475
99	-0.1473	0.9571	94.75	0.9824	97.26	0.9868
99	-0.1473	0.9571	94.75	0.9824	97.26	0.9868
101	0.0164	0.9995	100.95	0.9904	100.03	0.9865
102	0.0982	0.9808	100.04	0.9547	97.38	0.9468
102	0.0982	0.9808	100.04	0.9547	97.38	0.9468
104	0.2618	0.8676	90.23	0.8114	84.39	0.7966
110	0.7527	0.1878	20.66	0.1229	13.52	0.1087

Como estimativa para o parâmetro de locação após a primeira iteração tem-se

$$T^{(1)} = \left(\sum_{i=1}^{n} wi.yi\right) / \sum_{i=1}^{n} wi = 837.26/8.36 = 100.15$$

A partir desta nova estimativa calcula-se novos pesos para continuar o processo iterativo. Para este exemplo o critério de parada consiste em interromper o processo iterativo quando

$$T^{(i+1)} - T^{(i)} < 0.01 \qquad \text{onde}$$

$T^{(i)}$ = valor de T após a i-ésima iteração.

Deve-se observar como o peso dado à 10ª observação diminui até ficar bem menor que os pesos das demais observações.

Neste caso, a convergência ocorre após 3 iterações e o valor final obtido para a estimativa é 99.90

Para ilustração comparamos agora, via Monte Carlo, o desempenho de 3 estimadores de locação: média, mediana e biponderado, no contexto de uma normal contaminada. Consideramos a Normal contaminada com desvio padrão 1, e Normais contaminantes com desvios padrões iguais a 2, 5 e 10. As taxas de contaminação consideradas foram de 10 e 25%, e os tamanhos amostrais considerados foram 10 e 50. Em todos os casos empregamos 500 repetições Monte Carlo. Assim,

$$Y1, Y2, \ldots, Yn \quad iid \sim f$$

onde $f(y) = p\sigma\varphi((y-100)/\sigma) + (1-p)\varphi(y-100)$ onde $\varphi(x)$ é a densidade da normal padrão e $(p,\sigma) \in (0.10, 0.25) \times (2, 5, 10)$.

Os resultados podem ser apreciados na tabela 2.1 e nos histogramas a seguir (figura 2.2). Nas tabelas temos as estimativas das eficiências relativas da mediana e do estimador biponderado com respeito à média. Como as distribuições utilizadas são simétricas, temos que a mediana e o estimador biponderado são estimadores não viciados para a média populacional. Para compararmos então o desempenho dos diferentes estimadores basta nos fixarmos nas estimativas das variâncias. A tabela 2.1 apresenta, para cada caso considerado, a eficiência relativa da mediana e do estimador biponderado, com relação à média amostral, respectivemente. Os histogramas são para o caso em que $n = 10$, $\sigma = 10$, $p = 0.10$.

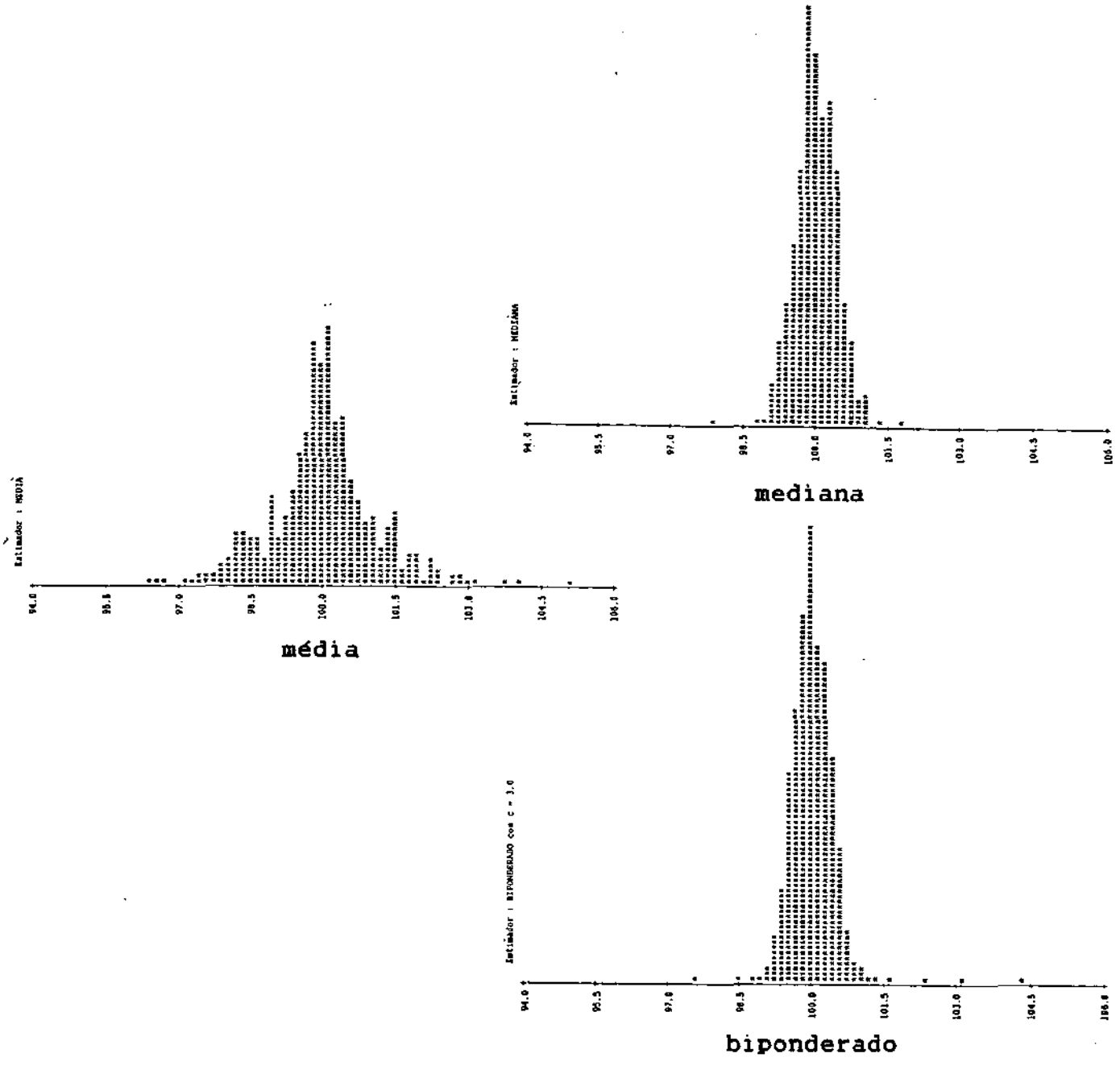


Figura 2.2 - Histogramas para a média amostral, mediana amostral e estimador biponderado de locação baseados em 500 repetições Monte Carlo com Y1,...,Y10 iid $\sim$ 0.10x10xφ((y-100)/10) + 0.9xφ(y-100).

Tabela 2.1 - Eficiência relativa da mediana e do estimador biponderado com respeito à média amostral, para amostras de tamanho 10 e 50, percentagem de contaminação p, e desvio padrão da normal contaminante σ, com d.p. da normal contaminada igual a 1. O valor acima é referente à mediana e o valor abaixo é referente ao estimador biponderado.

		n = 10		n = 50	
		p=0.10	p=0.25	p=0.10	p=0.25
	2	0.813	0.847	0.737	1.067
		0.897	0.960	1.000	0.941
s					
i	5	2.006	2.980	1.594	2.480
g		2.504	2.321	1.759	4.769
m					
a	10	6.567	8.782	4.939	19.374
		5.117	2.924	8.579	20.483

Pode-se observar, que à medida que o desvio padrão da contaminante aumenta, a eficiência tanto da mediana como do estimador biponderado melhora. Quando p e σ são ambos grandes, temos que se a amostra é pequena a mediana apresenta um desempenho bem superior ao estimador biponderado. No entanto, com uma amostra maior, o estimador biponderado volta a desempenhar melhor

2.5 - ESTIMADORES DO TIPO M EM REGRESSÃO

Em regressão, como já foi visto anteriormente, temos que os estimadores de mínimos quadrados ordinários para as componentes de β são os estimadores lineares não viciados uniformemente de mínima variância. Entretanto, para erros com distribuições de caudas mais pesadas que a distribuição normal, ou ainda quando estamos diante de valores aberrantes, o critério de mínimos quadrados pode nos levar a estimativas distorcidas para os parâmetros. Surge então, a idéia de se buscar alternativas, que forneçam estimativas melhores nos casos em que não temos um bom desempenho do estimador de mínimos quadrados e que, quando os erros sejam realmente normais ainda tenhamos estimativas razoáveis. Visando este objetivo parece natural estendermos a idéia dos estimadores do tipo M de locação ao contexto de regressão linear múltipla.

No modelo geral de regressão

$$Yi = \beta o + \sum_{j=1}^{n} xij\ \beta j + \varepsilon i \qquad i = 1, \ldots, n$$

onde Yi's são as n observações conhecidas, Xij é o valor da i-ésima variável na j-ésima observação também conhecido, βj's são p+1 parâmetros desconhecidos que se quer estimar e os εi's são variáveis aleatórias independentes e identicamente distribuídas. Em notação matricial temos o modelo como sendo :

$$Y = X\ \beta + \varepsilon \quad \text{onde}$$

$Y' = (Y1\ Y2\ \ldots\ Yn)$

$X' = (1\ X1\ X2\ \ldots\ Xp)$ sendo

$$\mathbb{1}' = (\,1 \quad 1 \quad \ldots \quad 1\,)$$

$$X1' = (X11\ X21\ \ldots\ Xn1)$$

$$\vdots$$

$$Xp' = (X1p\ X2p\ \ldots\ Xnp)$$

$$\beta' = (\beta 0\ \beta 1\ \ldots\ \beta p)$$

$$\varepsilon' = (\varepsilon 1\ \varepsilon 2\ \ldots\ \varepsilon n)\ .$$

A idéia é que ao invés de se minimizar $\sum_{i=1}^{n} (yi - \sum_{j=0}^{p} xij\beta j)^2$, a soma de quadrados dos resíduos, que fornece as estimativas de mínimos quadrados ordinários para β, procure-se um mínimo para

$$\sum_{i=1}^{n} \rho(yi - \sum_{j=0}^{p} xij\beta j)$$

onde ρ possa ser alguma outra função. Sendo $\Psi(y;\theta) = (\partial/\partial\theta)\rho(y;\theta)$ isto equivale a solucionar o seguinte sistema de equações em b0, b1, ..., bp, as respectivas estimativas para $\beta 0$, $\beta 1$, ..., βp.

$$\sum_{i=1}^{n} \Psi(yi - \sum_{j=0}^{p} xijbj).xik = 0 \qquad k = 0, 1, \ldots, p \qquad (2.1)$$

onde $yi - \sum_{j=0}^{p} xijbj$ vem a ser ei, o i-ésimo resíduo.

Como, em geral, a escala não é conhecida introduzimos uma estimativa S para esta de forma a (2.1) ser invariante por escala.

Ficamos então com o problema de resolver o seguinte sistema de equações :

$$\sum_{i=1}^{n} \Psi(ei/S)\ .\ xik = 0 \qquad k = 0, 1, \ldots, p$$

Assim, se está diante de um problema cuja solução deve ser obtida através de um processo iterativo, pois, para se ter um ajuste do modelo, deve-se ter uma estimativa S para σ, o desvio padrão dos erros, e para se obter esta estimativa é necessário estar de posse de um ajuste do modelo.

2.6 - O ESTIMADOR BIPONDERADO EM REGRESSÃO

A partir da conceituação do estimador biponderado de locação, surge uma maneira natural de aplicá-lo no contexto de regressão linear.

Como já foi visto anteriormente o estimador biponderado de locação é uma média ponderada das observações onde os pesos dados às mesmas dependem de uma medida robusta de escala dos dados. No caso de regressão, a variabilidade é encontrada nos erros. Parece natural, que se dê pesos menores àquelas observações cujo erro é grande. Para se verificar, a magnitude do erro, nada mais natural do que utilizarmos resíduos. A partir dos resíduos calculamos pesos para as diferentes observações e, de posse desses, procedemos como em mínimos quadrados ponderados.

Temos novamente aqui um processo iterativo pois os resíduos são obtidos a partir de um ajuste do modelo. O ajuste do modelo pressupõe uma definição dos pesos para as observações, que por sua vez são obtidos a partir dos resíduos.

Utilizando a função $\Psi(x)$ do estimador biponderado de locação no contexto de regressão temos o seguinte processo :

Passo 1 - Obtenção de um vetor de estimativas $b^{(i)}$ para as componentes de β.

Passo 2 - A partir de $b^{(i)}$ obter $S^{(i)}$, uma estimativa para o D.P. dos erros.

Passo 3 - a partir de $S^{(i)}$ calcular wj's, os pesos das observações yj's da seguinte maneira :

$$wj = \begin{cases} (1 - uj^2)^2 & \text{se } |uj| < 1 \\ 0 & \text{caso contrário} \end{cases}$$

$$\text{onde} \quad ui = \frac{yj - \hat{y}j}{c*S}$$

sendo c = constante pré-fixada

$\hat{y}j$ = o valor ajustado para yj tendo $\hat{\beta} = b^{(i)}$

Passo 4 - De posse dos wj's fazer um ajuste por mínimos quadrados ponderados, com os pesos sendo os wj's, obtendo assim um novo vetor de estimativas $b^{(i+1)}$ para a componentes de β .

Passo 5 - Se $|bj^{(i+1)} - bj^{(i)}| >$ para algum j retornar ao passo 1 caso contrário encerrar o processo com $\hat{\beta} = b^{(i+1)}$.

O vetor de valores iniciais para o vetor de estimativas b a ser usado no processo iterativo deve ser escolhido com alguma atenção ao contexto em que se está trabalhando. Quando estamos com um caso em que, por exemplo, um único ponto pode fazer com que as estimativas de mínimos quadrados para β estejam bastante distorcidas, o mais indicado é utilizar um outro ponto inicial. O critério de mínimos desvios absolutos fornece um bom ponto de partida nestes casos.

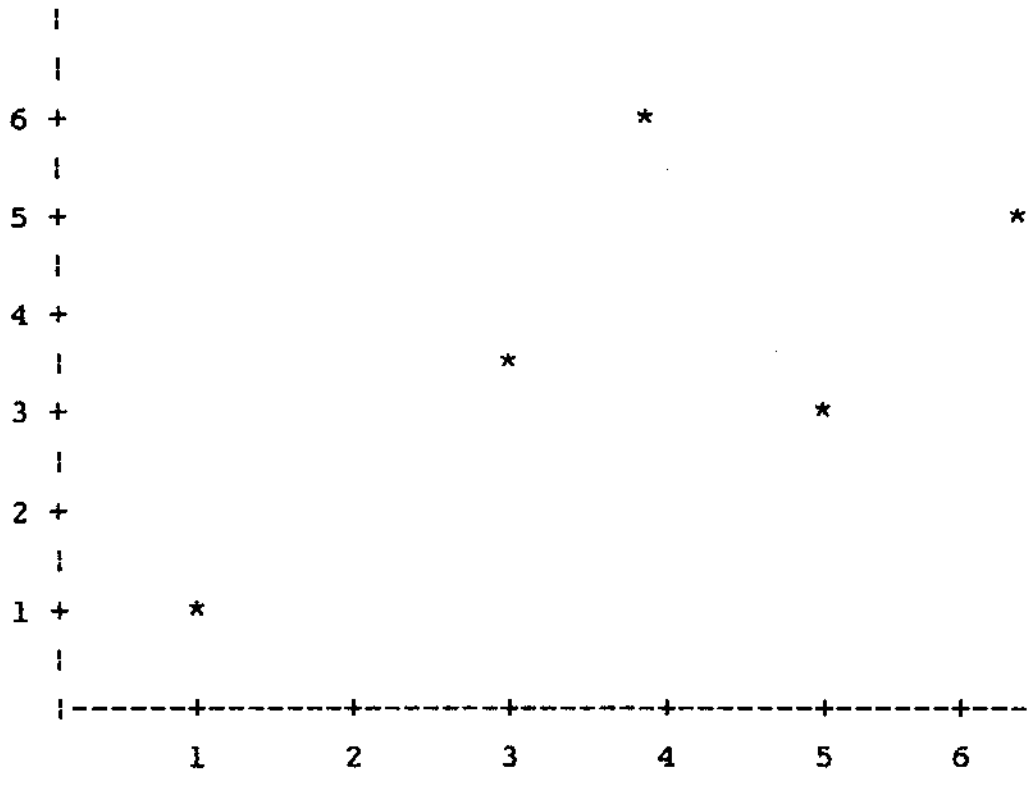
Exemplo 2.2 : Vejamos como funciona o estimador biponderado no caso de regressão linear simples. Considere o seguinte conjunto de pontos.

xi	1.0	3.0	4.0	5.0	6.5
yi	1.0	3.5	6.0	3.0	5.0

Vamos utilizar o estimador biponderado para se ajustar o modelo $Y = a + bX$. Utilizaremos como estimativas iniciais, as estimativas de mínimos quadrados, $c = 3.0$ e $S =$ (amplitude interquartis)/1.35. O quadro abaixo mostra os valores obtidos na primeira iteração.

yi	$\bar{y}i$	ei	wi
1.0	1.87	-0.87	0.761
3.5	3.13	0.37	0.955
6.0	3.76	2.24	0.025
3.0	4.39	-1.39	0.452
5.0	5.34	-0.34	0.961

O processo é repetido até que tenhamos a convergência nas estimativas dos dois parâmetros. Neste caso temos 5 iterações. A reta de mínimos quadrados e a reta do biponderado podem ser visualizadas no gráfico a seguir.



Uma comparação do estimador biponderado em regressão com outros estimadores do tipo M pode ser encontrada em Rocke e Shanno (1986).

De forma a se verificar o desempenho do estimador biponderado em comparação com o estimador de mínimos quadrados ordinários optou-se por uma abordagem via simulações de Monte Carlo, já que o desempenho do estimador biponderado, depende da estrutura probabilística dos erros e também da estrutura da matriz X. Nas simulações efetuadas neste trabalho o modelo ajustado é da forma :

$$Y = X\beta + \varepsilon \qquad \text{onde ,}$$

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 4 & 2 \\ 1 & 5 & 3 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 4 & 2 \\ 1 & 5 & 3 \end{bmatrix} \qquad \beta = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$$

$\varepsilon \sim g$ onde $g(x) = p\sigma\ (x/\sigma) + (1-p)\ (x)$

sendo (x) a densidade da normal padrão.

Para p utilizou-se os valores de 0.05, 0.10 e 0.25 e para σ os valores foram 2, 3, 4, 5 e 10. Foram feitas simulações para todas as combinações de p e σ. O número de repetições de Monte Carlo para cada combinação foi de 2000.

Para estas simulações o estimador foi utilizado da seguinte maneira : para b0 foi utilizado o estimador de mínimos quadrados ordinários. Para c foram utilizados os valores 3.0 e 5.0. Como S temos (amplitude interquartis)/1.35 que é uma medida robusta de escala para o D.P. dos erros. O processo é interrompido ou quando se verifica a convergência, ou quando já houveram 15 iterações.

Como o estimador biponderado é em cada passo um estimador de mínimos quadrados ponderados temos que ele é não viciado assim como o estimador de mínimos quadrados. Para comparar então o desempenho dos dois estimadores devemos nos fixar na variância de cada um. Usaremos para este fim, as estimativas das eficiências dos estimadores nos diferentes casos, que foram obtidas com as estimativas das variâncias dos estimadores calculadas com as 2000 repetições.

Os resultados obtidos podem ser apreciados nas tabelas 2.2, 2.3, 2.4 e 2.5.

Na tabela 2.2 pode-se observar as médias e variâncias das 2000 estimativas, por mínimos quadrados e pelo biponderado para c=3.0, de βo, $\beta 1$ e $\beta 2$.

As médias das estimativas são em geral bastante próximas dos valores verdadeiros em ambos os casos, conforme esperado. A maior diferença ocorre para p=0.25 e $\sigma = 5$, quando a média das estimativas por mínimos quadrados de $\beta 2$ iguala a 3.783. Esta foi a única diferença significativa ao nível 1%. Das diferenças, apenas 2 são significadivas ao nível 5%, as médias de b2, para o estimador de mínimos quadrados em $p = 0.25$ e $\sigma = 3$, e para o estimador biponderado em $p = 0.05$ e $\sigma = 5$. O número de diferenças significativas não foi, contudo, significativo, dado o alto número, 96, de comparações feitas.

As variâncias são menores para b1, seguidas pelas de bo, e bem maiores para b2. Tais diferenças, contudo, podem ser explicadas pelas diferenças nos delineamentos. As variâncias, tanto para os Mínimos Quadrados, quanto para o Biponderado crescem com p e com σ, o que não poderia ser diferente, dado o crescimento da variância efetiva de Y com aqueles parâmetros. Contudo, cumpre notar que, para o caso do estimador biponderado o crescimento é dramaticamente atenuado, indicando exatamente sua propriedade de proteger as estimativas contra as caudas pesadas da normal contaminada, às quais o estimador de mínimos quadrados se revela particularmente vulnerável. Para a normal pura, nos casos em $\sigma = 2$, e no caso em que $\sigma = 3$ e $p = 0.05$, o estimador por mínimos quadrados apresentou variâncias de 5 a 20% menores. Contudo nos contextos de contaminação mais pesada, seja por uma alta taxa de contaminação p, ou por um σ grande, as vantagens do biponderado são esmagadoras, como no caso de $p = 0.10$ e $\sigma = 10$, em que V(b2) foi estimada como 4.975 e 1.197, para os estimadores de mínimos quadrados e biponderado, respectivamente.

Tabela 2.2 - Média e variância de estimativas de $\beta 0$, $\beta 1$ e $\beta 2$, obtidas por mínimos quadrados ordinários e estimador biponderado com c = 3.0, com base em 2000 repetições Monte Carlo quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$

c = 3.0

	$\varepsilon \sim N(0,1)$			
	M.Q.		BIP.	
	med.	var.	med.	var.
b0	4.009	0.682	4.007	0.860
b1	3.988	0.477	3.993	0.603
b2	4.011	1.695	4.003	2.139

		p = 0.05				p = 0.10				p = 0.25			
		M.Q.		BIP.		M.Q.		BIP.		M.Q.		BIP.	
		med.	var.	med.	var.	med.	var.	med.	var.	med.	var.	med.	var.
σ = 2	b0	3.987	0.8282	4.001	1.033	3.951	0.952	3.963	1.083	3.984	1.204	3.970	1.262
	b1	3.986	0.548	3.968	0.634	4.037	0.608	4.022	0.696	4.000	0.824	4.002	0.895
	b2	4.029	1.986	4.051	2.359	3.960	2.218	3.978	2.510	4.015	2.973	4.015	3.174
σ = 3	b0	4.019	0.999	4.029	1.023	3.941	1.328	3.966	1.202	4.076	2.136	4.048	1.888
	b1	3.979	0.659	3.990	0.686	4.050	0.851	4.024	0.766	4.041	1.493	4.039	1.246
	b2	4.024	2.328	3.998	2.430	3.944	3.096	3.973	2.750	3.895	5.191	3.909	4.355
σ = 4	b0	3.988	1.174	3.990	0.989	3.994	1.757	4.007	1.285	3.954	3.475	3.947	2.435
	b1	3.991	0.848	4.000	0.711	3.998	1.139	4.003	0.808	4.025	2.113	3.972	1.625
	b2	4.028	3.029	4.009	2.501	4.016	4.115	4.004	2.952	3.975	7.596	4.071	5.596
σ = 5	b0	3.984	1.529	3.999	1.028	4.027	2.337	3.995	1.344	4.118	5.223	4.051	3.075
	b1	4.042	1.031	4.052	0.657	4.010	1.704	4.002	0.940	4.095	3.451	4.028	2.258
	b2	3.943	3.683	3.914	2.451	3.974	5.849	4.000	3.319	3.783	12.482	3.928	7.763
σ = 10	b0	4.045	4.248	3.997	1.241	4.067	7.791	4.018	1.748	4.050	18.062	4.002	9.281
	b1	3.999	2.918	3.982	0.768	3.988	4.975	3.979	1.197	3.976	12.310	3.994	6.187
	b2	3.979	9.993	4.031	2.710	3.964	17.487	4.020	4.213	3.995	44.047	4.001	22.125

Tabela 2.3 - Média e variância de estimativas de $\beta 0$, $\beta 1$ e $\beta 2$, obtidas por mínimos quadrados ordinários e estimador biponderado com c = 5.0, com base em 2000 repetições Monte Carlo quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$

c = 5.0

	$\varepsilon \sim N(0,1)$			
	M.Q.		BIP.	
	med.	var.	med.	var.
b0	4.009	0.653	4.014	0.664
b1	3.989	0.464	3.992	0.490
b2	4.011	1.648	4.004	1.740

		p = 0.05				p = 0.10				p = 0.25			
		M.Q.		BIP.		M.Q.		BIP.		M.Q.		BIP.	
		med.	var.	med.	var.	med.	var.	med.	var.	med.	var.	med.	var.
σ = 2	b0	3.967	0.801	3.988	0.809	3.993	0.878	3.989	0.862	4.053	1.268	4.056	1.250
	b1	4.010	0.551	4.009	0.561	3.983	0.642	3.981	0.627	3.987	0.826	3.998	0.812
	b2	3.989	1.905	3.991	1.947	4.031	2.253	4.036	2.203	4.001	2.995	3.981	2.932
σ = 3	b0	3.988	0.980	3.999	0.890	3.984	1.273	4.007	1.097	3.923	2.024	3.950	1.791
	b1	3.990	0.670	3.993	0.604	4.012	0.870	4.009	0.718	4.024	1.324	4.021	1.134
	b2	4.023	2.372	4.014	2.118	3.980	3.111	3.979	2.573	4.000	4.720	3.992	4.026
σ = 4	b0	3.997	1.301	3.988	1.023	4.012	1.845	4.007	1.298	3.986	3.265	3.967	2.568
	b1	4.003	0.851	4.001	0.622	4.019	1.200	4.016	0.829	3.976	2.238	3.983	1.667
	b2	4.004	3.083	4.012	2.280	3.968	4.357	3.978	2.995	4.056	7.748	4.044	5.763
σ = 5	b0	4.012	1.430	4.015	1.005	3.914	2.283	3.973	1.385	3.937	4.916	3.954	3.687
	b1	4.000	0.998	4.016	0.662	3.954	1.498	3.964	0.823	3.959	3.488	3.949	2.443
	b2	3.995	3.557	3.966	2.383	4.120	5.338	4.075	2.890	4.100	12.315	4.103	8.560
σ = 10	b0	4.001	4.398	4.016	1.385	4.053	7.535	4.030	2.762	4.134	16.960	4.150	9.552
	b1	4.014	3.207	3.987	0.994	3.922	5.553	3.999	1.770	4.118	11.662	4.105	6.380
	b2	3.971	10.910	4.006	3.391	4.103	19.580	3.988	5.865	3.745	41.554	3.759	22.066

Resultados análogos, mas agora tomando o estimador biponderado com c = 5.0, são apresentados na tabela 2.3. Novamente as médias não diferem significativamente dos valores verdadeiros. Uma diferença significativa ao nível 5% só foi verificada uma vez, tendo ocorrido no caso p = 0.10, σ = 5, para b2 por mínimos quadrados.

Os mesmos padrões com respeito às variâncias, verificados para o caso de c = 3, ocorrem aqui. Novamente, o estimador biponderado se revela bastante estável com respeito ao aumento do peso das caudas.

A tabela 2.4 de uma forma sumariza os resultados mais interessantes apresentados na tabela 2.2. Tem-se aqui as eficiências relativas do estimador biponderado com c = 3, com respeito ao estimador de mínimos quadrados.

Vê-se aqui que diante da normalidade dos erros, o estimador apresenta um desempenho apenas satisfatório, sendo de 0.79 para bo, dl e b2. A medida que p ou σ aumentam, o desempenho relativo do estimador biponderado vai melhorando. Para p = 0.05, a partir de σ = 4 tem-se uma eficiência maior que 1. Para p = 0.10 e p = 0.25 a eficiência é maior do que 1 já a partir de σ = 3.

A tabela 2.5 apresenta resultados análogos, agora referentes aos dados apresentados na tabela 2.3. Neste caso, mesmo com erros distribuídos normalmente o estimador biponderado já apresenta um desempenho bastante bom, sendo superior, para os três estimativas, a 0.94.

Com erros distribuídos segundo normais contaminadas a eficiência só é inferior a 1 no caso em que p = 0.05 e σ = 2. A tendência no valor da eficiência apresentada para variações no valor de p ou σ é semelhante a da tabela 2.4.

Tabela 2.4 - Eficiência relativa amostral do estimador biponderado com $c = 3.0$ quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

$c = 3.0$

$\varepsilon \sim N(0,1)$		
b0	b1	b2
0.792	0.792	0.793

	p = 0.05			p = 0.10			p = 0.25		
	b0	b1	b2	b0	b1	b2	b0	b1	b2
$\sigma = 2$	0.802	0.863	0.842	0.879	0.874	0.884	0.954	0.921	0.937
$\sigma = 3$	0.976	0.960	0.958	1.104	1.111	1.126	1.132	1.198	1.192
$\sigma = 4$	1.186	1.192	1.211	1.367	1.410	1.588	1.427	1.300	1.357
$\sigma = 5$	1.488	1.569	1.503	1.740	1.813	1.762	1.699	1.528	1.608
$\sigma = 10$	3.423	3.801	3.687	4.457	4.157	4.151	1.946	1.990	1.991

Tabela 2.5 - Eficiência relativa amostral do estimador biponderado com $c = 5.0$ quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

$c = 5.0$

$\varepsilon \sim N(0,1)$		
b0	b1	b2
0.984	0.948	0.947

	p = 0.05			p = 0.10			p = 0.25		
	b0	b1	b2	b0	b1	b2	b0	b1	b2
$\sigma = 2$	0.990	0.981	0.978	1.019	1.023	1.023	1.015	1.018	1.021
$\sigma = 3$	1.101	1.110	1.120	1.160	1.212	1.209	1.130	1.168	1.172
$\sigma = 4$	1.273	1.368	1.352	1.422	1.447	1.455	1.271	1.342	1.345
$\sigma = 5$	1.422	1.507	1.493	1.648	1.820	1.847	1.333	1.428	1.439
$\sigma = 10$	3.176	3.226	3.217	2.729	3.138	3.339	1.776	1.828	1.883

CAPÍTULO 3

SELEÇÃO DE VARIAVEIS BIPONDERADA PASSO-A-FRENTE

3.1 - INTRODUÇÃO

Como já foi citado anteriormente, é muito comum a situação em que, dispondo-se de um conjunto de p variáveis explicativas potenciais: Xl, ..., Xp, se deseje selecionar um subconjunto destas variáveis para se construir um modelo de regressão que resulte num compromisso adequado entre parcimônia e explicabilidade.

Os métodos usuais para a seleção de variáveis são formulados com base em estatísticas obtidas a partir do ajuste pelo critério de mínimos quadrados. Nas situações em que os mínimos quadrados não apresentam um bom desempenho, como é o caso em que os erros têm distribuição com cauda mais pesada do que a da normal, o processo de seleção do modelo sofre algum prejuízo, que se transfere para o modelo selecionado. Seria então de interesse considerar alternativas menos sensíveis a valores extremos, na seleção das variáveis.

3.2 - PROPOSTA DE MOSTELLER E TUKEY

A proposta de Mosteller e Tukey(1977) combina o emprego de estimadores robustos com métodos já existentes de seleção de variáveis. Neste trabalho explora-se estas idéias, aplicando o estimador biponderado num contexto de seleção passo-a-frente de variáveis.

O esquema básico proposto por Mosteller e Tukey sugere uma abordagem específica para o problema. Em particular, tendo-se um con-

junto de dados, aplica-se o procedimento passo-a-frente tradicional, isto é, com as estimativas para β sendo obtidas por mínimos quadrados. Tendo-se chegado por esta via a um subconjunto Xil, ..., Xik, das p variáveis originais, estima-se β utilizando-se o estimador biponderado em regressão. De posse dos pesos na última iteração do estimador biponderado faz-se um passo-à-frente com os ajustes obtidos por mínimos quadrados ponderados. De uma maneira esquemática tem-se o procedimento a seguir :

1 - Escolhe-se pesos para as observações, se desejado

2 - Faz-se um passo-à-frente utilizando-se mínimos quadrados ponderados com pesos fixos e escolhe-se um subconjunto de variáveis.

3 - Aplica-se o estimador biponderado com as variáveis explicativas sendo aquelas escolhidas em (2). Isto resultará em (a) um ajuste, (b) resíduos e (c) pesos da última iteração. Examina-se os resíduos e pesos cuidadosamente. Caso pareçam razoáveis, de posse dos pesos, retorna-se a (1).

Para se decidir a respeito do número de repetições necessárias eles sugerem que a experiência guiará cada tipo de problema

3.3 - BIPONDERADO PASSO-A-FRENTE

A proposta de Mosteller e Tukey é bastante atraente mas sugere, de imediato, uma maneira alternativa de se combinar estas duas técnicas. O passo-à-frente, empregando os mínimos quadrados ordinários, tem os acréscimos na soma de quadrados do resíduo, devido à restrição ($\beta i = 0$) fortemente influenciados pelas estimativas de σ^2, cuja

variabilidade, por sua vez, sofre forte influência das caudas da distribuição dos erros.

Daí a questão: Haveria alguma maneira de se precaver contra erros grosseiros nas estimativas de σ^2? Uma idéia que surge naturalmente, consiste da utilização do princípio de seleção passo-a-frente com as estimativas para σ^2 e as para β sendo obtidas de uma maneira menos sensível a valores extremos.

A proposta é de que em cada estágio do procedimento passo-à-frente, as estatísticas utilizadas para se testar $\beta i = 0$ contra $\beta i \neq 0$, sejam calculadas com b e S, as estimativas de β e σ respectivamente, obtidas a partir de um ajuste através do estimador biponderado. Teremos, desta forma, o procedimento descrito a seguir.

Em cada estágio, para cada variável Xi, ainda não incluída no modelo, faz-se o ajuste, mantendo-se as variáveis já incluídas no modelo até o presente estágio e adicionando-se a variável Xi, utilizando-se o estimador biponderado. Calcula-se o acréscimo correspondente na soma de quadrados de resíduos devido a restrição ($\beta i = 0$). Verifica-se então se o máximo destes acréscimos é significativo ou não.

Considerando-se um estágio do processo de seleção em que m, $0<m<p$, variáveis já tenham sido selecionadas e incluídas no modelo - suponha-se sem perda de generalidade, que as variáveis já selecionadas sejam X1, X2, ..., Xm - o procedimento pode ser descrito de maneira esquemática, como se segue :

i - Toma-se i = m+1

ii - Ajusta-se o modelo de regressão envolvendo-se X1, X2, ...,Xm e Xi utilizando-se o estimador biponderado

iii - Calcula-se o acréscimo na soma de quadrados dos resíduos devido à restrição $\beta i = 0$, denominado ASQ ($\beta i = 0$)

iv - i = i + 1

v - Se i+1 > p vá para (vi), caso contrário vá para (ii)

vi - Definindo-se $Fi = bi^2 / S^2.dii$, determina-se $F = \max_{m<i<p+1} (Fi)$ e o i correspondente, onde

bi - a estimativa para βi, pelo estimador biponderado

dii - i-ésimo elemento da diagonal de $(X'PX)^{-1}$ e

S - estimativa para σ a partir dos resíduos obtidos do ajuste com o estimador biponderado.

vii - Compara-se F = Fimax, com o valor de referência da tabela da distribuição F de Snedecor com (1,n-m-2) graus de liberdade. Se F for maior que o valor de referência, inclue-se a variável Ximax no modelo e inicia-se novo estágio. Se for menor, encerra-se o processo de seleção de variáveis, com o modelo final envolvendo as variáveis X1, X2, ..., Xm.

3.4 - ASPECTOS ESTUDADOS E RESULTADOS

O nosso objetivo nesste trabalho é comparar o desempenho de dois métodos : o passo-à-frente usual e o biponderado passo-à-frente. Para isto focalizaremos três aspectos : o desempenho do estimador biponderado, que já foi visto no capítulo 2, o nível de significância

real do teste para a entrada de uma variável e o modelo final selecionado.

Os aspectos estudados neste trabalho, são bastante relacionados com a estrutura da matriz X utilizada e à estrutura probabilística dos erros. Dada a complexidade implícita numa abordagem analítica desta questão, a abordagem natural do problema é via simulações Monte Carlo.

No nosso caso, para o estudo comparativo proposto, consideramos sempre n = 10, gerando os dados segundo o modelo :

$$Y = 4.0 + 4.0\ X1 + 4.0\ X2 + \varepsilon$$

O delineamento adotado, envolvendo ainda a variável X3, é dado pela matriz X .

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 3 & 2 & 1 \\ 1 & 4 & 2 & 1 \\ 1 & 5 & 3 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 2 & 1 & -1 \\ 1 & 3 & 2 & -1 \\ 1 & 4 & 2 & -1 \\ 1 & 5 & 3 & -1 \end{bmatrix}$$

Como o estimador biponderado é indicado para situações em que os erros têm distribuição com cauda mais pesada do que a distribuição normal, consideramos os erros distribuídos segundo normais contaminadas, isto é, com função de densidade de probabilidade f dada por :

$$f(x) = p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$$

onde $\varphi(x)$ é a densidade da normal padrão. Desta forma, p é a fração de contaminação e σ o desvio padrão da contaminante. O desvio padrão da contaminada foi fixado em 1. Neste trabalho consideramos todos os (p,σ) em (0.05, 0.10, 0.25) X (2, 3, 4, 5, 10)

De posse das amostras aplicamos os três métodos e comparamos os resultados.

Para cada combinação de p e σ foram geradas 2000 amostras independentes. O biponderado passo-a-frente foi realizado com o programa cuja listagem se encontra no anexo 1. Os outros programas necessários foram todos obtidos através de modificações neste programa base. A linguagem empregada foi Pascal e o computador utilizado foi o VAX.

Para cada distribuição dos erros obtivemos uma tabela que informa com que frequência cada uma das três variáveis foi incluída a 1%, 5% e 10%. A tabela tem o seguinte formato :

		α nominal		
		0.10	0.05	0.01
variável	1			
	2			
	3			

Para cada distribuição dos erros obteve-se também o número de amostras que resultaram em cada modelo final possível

Foram geradas amostras para mínimos quadrados ordinários, para o biponderado com c igual a 3.0 e a 5.0. Os resultados são apresentados nas tabelas 3.1, 3.2, 3.3, 3.4, 3.5 e 3.6

Na tabela 3.1 observa-se a frequência com que cada uma das variáveis foi incluída no modelo aos níveis nominais α de significância de 0.10, 0.05 e 0.01, quando o ajuste é obtido com mínimos quadrados ordinários,

Quando os erros são distribuídos normalmente tem-se um poder, no teste para entrada de variáveis, muito bom para a variável X1 chegando a 1.00 para α = 0.10. A variável X2 apresenta um bom poder mas ao contrário da variável X1 há uma diferença muito grande no poder à medida que α diminui. O nível de significância real se aproxima bastante do nominal, o que se pode observar a partir das frequências da variável X3. Os níveis de significância observados são de 9.8%, 4.1% e 0.6%, correspondentes aos níveis nominais de 10%, 5% e 1%, respectivamente. A diferença no último caso tem nível de significância de 3.9%.

A menos do caso em que σ - o desvio padrão da contaminante - é igual a 10, quando a fração de contaminação é de 0.05, o nível de significância real não difere significativamente do nominal. Quando σ = 10, o nível real ficou sempre bem abaixo do nominal, embora para α = 0.01 as diferenças observadas não houvessem sido piores que para valores menores de σ. Para σ = 10 e p = 0.05, por exemplo, a variável 3 entrou apenas 52 vezes, ao nível α = 0.05, em 2000 repetições Monte Carlo, caracterizando um nível de significância nominal de 2.6%. Podemos então sumarizar dizendo que no contexto estudado, uma baixa taxa de contaminação (5%), e um desvio padrão da contaminante 10 vezes maior que o da contaminada, os níveis de significância reais tendem a ser menores que o níveis nominais, implicando numa posição mais conservativa que o arbitrado.

Tabela 3.1 - Frequência de inclusão das variáveis X1, X2 e X3 aos níveis nominais de significância α com o ajuste obtido por mínimos quadrados ordinários e $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$

MQO

		0.10	0.05	0.01
N(0,1)	X1	1.000	0.999	0.981
	X2	0.875	0.758	0.420
	X3	0.098	0.041	0.006

		p = 0.05			p = 0.10			p = 0.25		
α		0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$\sigma = 2$	X1	0.997	0.989	0.958	0.996	0.984	0.953	0.985	0.961	0.909
	X2	0.830	0.710	0.409	0.794	0.650	0.361	0.697	0.562	0.310
	X3	0.096	0.047	0.011	0.098	0.051	0.012	0.087	0.032	0.004
$\sigma = 3$	X1	0.980	0.971	0.938	0.968	0.948	0.909	0.905	0.870	0.817
	X2	0.796	0.670	0.389	0.743	0.619	0.365	0.588	0.476	0.302
	X3	0.088	0.038	0.004	0.089	0.038	0.010	0.084	0.031	0.003
$\sigma = 4$	X1	0.965	0.952	0.926	0.927	0.903	0.867	0.853	0.816	0.777
	X2	0.775	0.667	0.384	0.710	0.602	0.376	0.539	0.452	0.327
	X3	0.091	0.040	0.007	0.080	0.037	0.006	0.089	0.033	0.006
$\sigma = 5$	X1	0.957	0.944	0.919	0.898	0.897	0.850	0.802	0.777	0.753
	X2	0.745	0.636	0.389	0.668	0.568	0.351	0.478	0.416	0.318
	X3	0.100	0.040	0.010	0.078	0.033	0.004	0.078	0.031	0.003
$\sigma = 10$	X1	0.885	0.876	0.850	0.801	0.785	0.752	0.655	0.639	0.588
	X2	0.714	0.618	0.394	0.634	0.577	0.403	0.461	0.427	0.359
	X3	0.069	0.026	0.006	0.056	0.017	0.004	0.073	0.026	0.001

Por outro lado, ainda com p = 0.05, com relação à variável X1, há uma queda gradual de poder à medida que o σ cresce, mas a queda não chega a ser muito grave, indo de 0.997, 0.989 e 0.958, para σ = 2, para 0.885, 0.876 e 0.850, para σ = 10, correspondentes a α = 0.10, 0.5 e 0.1, respectivamente. Fixados σ e p, temos que o poder não difere muito de um α para outro, indicando uma nítida separação entre a distribuição de b1 e sua distribuição sob a hipótese nula.

Com relação à variável X2, verifica-se uma marcante deterioração no poder com o crescimento de σ. Para p = 0.05, o poder cai de 0.830, 0.710 e 0.409, para 0.714, 0.636 e 0.394, quando σ cresce de 2 para 10, para os níveis nominais de 10%, 5% e 1%, respectivamente.

Para taxas de contaminação mais altas, com p = 0.10 e 0.25, o nível de significância real tende a se afastar mais do nominal, sempre no sendito de esquemas mais conservadores que o arbitrado. O poder para a variável X2 decai substancialmente chegando a 0.318 para p = 0.25, σ = 5 e α = 0.01. No caso da variável X1 também se verifica uma acentuada queda de poder quando σ aumenta. Para p = 0.25, este vai de 0.985, 0.961 e 0.909 para 0.655, 0.639 e 0.588, quando σ cresce de 2 para 10, para os níveis de significância nominais de 0.10, 0.05 e 0.01, respectivamente.

Na tabela 3.2 observa-se a frequência com que as variáveis foram incluídas no modelo, aos níveis de significância nominais especificados, quando o ajuste é feito utilizando-se o estimador biponderado com c = 3.0. Novamente, o número de repetições Monte Carlo empregado foi de 2000.

Os níveis de significância reais, que podem ser analisados a partir das frequências relativas a variável X3, são na grande maioria dos casos substancialmente superiores aos níveis de significâncias nominais. Quando os erros são distribuídos normalmente, tem-se por exemplo que os níveis reais observados são 0.173, 0.104 e 0.035 correspon-

Tabela 3.2 - Frequência de inclusão das variáveis X1, X2 e X3 aos níveis nominais de significância α com o ajuste sendo feito com o estimador biponderado com $c = 3.0$ e $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

biponderado c = 3.0

		0.10	0.05	0.01
N(0,1)	X1	1.000	0.995	0.954
	X2	0.792	0.668	0.382
	X3	0.173	0.104	0.035

		p = 0.05			p = 0.10			p = 0.25		
α		0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$\sigma = 2$	X1	0.996	0.984	0.937	0.990	0.975	0.929	0.979	0.956	0.900
	X2	0.739	0.622	0.373	0.740	0.606	0.363	0.660	0.521	0.317
	X3	0.171	0.103	0.030	0.160	0.103	0.039	0.150	0.089	0.027
$\sigma = 3$	X1	0.993	0.983	0.941	0.984	0.969	0.915	0.950	0.924	0.879
	X2	0.745	0.607	0.344	0.717	0.577	0.334	0.595	0.471	0.301
	X3	0.148	0.089	0.027	0.156	0.098	0.035	0.137	0.082	0.023
$\sigma = 4$	X1	0.990	0.979	0.938	0.972	0.955	0.908	0.920	0.893	0.847
	X2	0.740	0.605	0.350	0.717	0.574	0.342	0.561	0.445	0.299
	X3	0.171	0.105	0.029	0.156	0.097	0.027	0.120	0.066	0.021
$\sigma = 5$	X1	0.993	0.978	0.938	0.974	0.956	0.907	0.893	0.861	0.820
	X2	0.742	0.613	0.363	0.672	0.536	0.320	0.521	0.423	0.298
	X3	0.155	0.096	0.033	0.143	0.085	0.028	0.109	0.056	0.020
$\sigma = 10$	X1	0.981	0.970	0.927	0.956	0.943	0.899	0.820	0.793	0.747
	X2	0.740	0.606	0.363	0.660	0.525	0.298	0.485	0.408	0.282
	X3	0.143	0.085	0.032	0.118	0.072	0.026	0.083	0.042	0.009

dentes aos níveis nominais de 0.10, 0.05 e 0.01 respectivamente. A diferença entre os níveis nominais e os níveis reais vai diminuindo à medida que p ou σ aumentam. Assim, ficam evidências de que o esquema biponderado com c = 3 inclue variáveis no modelo mais liberalmente, sendo menos conservativo que o arbitrado. Quando p = 0.25 e σ = 10 ou 5 e p = 0.10 e σ = 10 tem-se que a diferença já não é tão marcante.

O poder para a variável X1 decresce à medida que p ou σ aumentam. Porém, a menos de quando p = 0.25 e σ = 10 o poder nunca é inferior a 0.80 qualquer que seja o nominal.

Para a variável X2, também há um decréscimo no poder à medida que p ou σ aumentam, mas para esta variável o poder chega a ser bem baixo (0.282). Para p e σ constantes, verifica-se uma queda acentuada do poder quando α diminue de 0.10 para 0.01. Por exemplo, para p = 0.05 e σ = 3, o poder cai de 0.745 para 0.344 quando α vai de 0.10 a 0.01.

Na tabela 3.3 observa-se a frequência com que cada uma das variáveis foi incluída no modelo aos níveis de significância nominais especificados, quando o ajuste é obtido com o estimador biponderado com c = 5.0.

Com os erros distribuídos normalmente, o nível de significância real do teste é bem próximo do nominal. O poder para a variável X1 é bastante bom. Para a variável X2, embora o poder seja bom para α = 0.10 (0.809), este fica bastante baixo quando α = 0.01 (0.363).

O nível de significância real do teste F para inclusão de variáveis pode ser observado a partir das frequências relativas à variável X3. Na parte triangular esquerda superior da tabela o nível real se aproxima bastante bem do nominal enquanto qu na parte triangular inferior direita tem-se uma aproximação razoável. Retirando-se (p,σ) = (0.25,5), (0.25,10) e (0.10,10) tem-se que o nível de significância real sempre se encontra dentro do intervalo (0.87,1.12) para α = 0.10, (0.45,0.63) para α = 0.05 e (0.012,0.019) para α = 0.01.

Tabela 3.3 - Frequência de inclusão das variáveis X1, X2 e X3 aos níveis nominais de significância α com o ajuste obtido com o estimador biponderado com c = 5.0 e $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

biponderado c = 5.0

		0.10	0.05	0.01
N(0,1)	X1	0.998	0.991	0.949
	X2	0.809	0.656	0.363
	X3	0.105	0.053	0.015

		p = 0.05			p = 0.10			p = 0.25		
α		0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
σ = 2	X1	0.995	0.983	0.943	0.994	0.983	0.935	0.978	0.957	0.904
	X2	0.786	0.639	0.367	0.754	0.601	0.333	0.809	0.656	0.301
	X3	0.111	0.062	0.019	0.107	0.063	0.017	0.105	0.053	0.015
σ = 3	X1	0.996	0.986	0.939	0.984	0.965	0.919	0.946	0.917	0.858
	X2	0.751	0.607	0.337	0.716	0.572	0.331	0.589	0.476	0.294
	X3	0.098	0.059	0.015	0.097	0.055	0.016	0.090	0.045	0.012
σ = 4	X1	0.992	0.982	0.946	0.977	0.962	0.917	0.909	0.878	0.832
	X2	0.752	0.619	0.337	0.683	0.545	0.317	0.566	0.462	0.288
	X3	0.107	0.059	0.015	0.098	0.052	0.013	0.092	0.045	0.013
σ = 5	X1	0.987	0.975	0.924	0.961	0.944	0.897	0.888	0.856	0.812
	X2	0.757	0.621	0.346	0.696	0.555	0.319	0.547	0.444	0.296
	X3	0.104	0.054	0.017	0.094	0.047	0.013	0.082	0.036	0.005
σ = 10	X1	0.982	0.975	0.931	0.948	0.933	0.889	0.809	0.787	0.738
	X2	0.752	0.600	0.348	0.665	0.519	0.307	0.493	0.416	0.288
	X3	0.087	0.050	0.014	0.085	0.041	0.007	0.073	0.034	0.008

O poder do teste para a variável X1 é bom e relativamente estável. No pior dos casos tem-se o poder igual a 0.738 (para $(p,\sigma) = (0.25,10)$ e $\alpha = 0.01$) sendo que na maioria dos casos ele é superior a 0.90. A diminuição no poder à medida que p ou σ aumentam se dá de uma maneira suave.

O poder para a variável X2 nunca é muito bom atingindo um máximo de 0.809. Nos piores casos ($(p,\sigma) = (0.25,5)$, $(0.25,10)$ e $(0.10,10)$), tem-se que para $\alpha = 0.10$ o poder é bem baixo sendo até 0.493. Além disso, observa-se um decréscimo muito grande no poder à medida que α diminui. O poder chega a ser tão baixo quanto 0.288. Embora quando p = 0.05 o poder também seja baixo tem-se neste caso uma estabilidade maior com relação a aumento de σ, quando comparado com p = 0.10 ou 0.25

A tabela 3.4 mostra quantas vezes, das 2000 repetições Monte Carlo, se chegou a cada um dos modelos finais possíveis. Por exemplo, para p = 0.05 e $\sigma = 10$, o modelo final incluiu exatamente as variáveis X1 e X2, 1081 (54.05%) vezes, quando o ajuste foi obtido por mínimos quadrados ordinários. Para simplicidade representamos, por exemplo, por X1X2 o modelo que inclue apenas as variáveis X1 e X2.

Com os erros distribuídos normalmente tem-se que apenas os modelos X1, X1X2, X1X3 e X1X2X3 foram escolhido alguma vez. O modelo correto, X1X2 foi escolhido 77.6% das vezes.

Para p = 0.05 o modelo correto é o mais escolhido, para todos os valores de σ. Porém, a frequência com que se chega ao modelo correto vai decrescendo à medida em que σ cresce, em favor de um crescimento das frequências de ocorrências de X1 e de X2. Assim, X1X2 é escolhido 73.6%, 70.0%, 65.9%, 61.6% e 54.1% das vezes, para $\sigma = 2$, 3, 4, 5 e 10, respectivamente. Convém observar aqui que o delineamento tem uma parcela de contribuição neste fato.

Tabela 3.4 - Número de amostras (em 2000) com um dado modelo final selecionado com o ajuste obtido por mínimos quadrados ordinários e $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

MQO

		variáveis no modelo final						
		X1	X2	X3	X1X2	X1X3	X2X3	X1X2X3
N(0,1)		242	0	0	1562	8	0	188
	σ							
p = 0.05	2	329	6	0	1473	12	0	180
	3	386	39	0	1400	23	1	151
	4	432	68	0	1318	19	3	160
	5	482	86	0	1232	29	1	170
	10	557	224	0	1081	14	6	117
p = 0.10	2	403	9	0	1392	10	0	186
	3	490	64	0	1269	24	1	152
	4	546	146	0	1148	34	0	126
	5	629	200	0	1016	35	5	115
	10	701	383	0	801	29	13	70
p = 0.25	2	572	30	0	1224	35	1	138
	3	770	182	0	881	54	8	105
	4	859	274	0	689	64	21	93
	5	972	369	0	503	73	27	56
	10	996	617	1	221	61	53	31

Tabela 3.5 - Número de amostras (em 2000) com um dado modelo final selecionado com o ajuste obtido com o estimador biponderado com c = 3.0 quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

biponderado c = 3.0

		variáveis no modelo final						
		X1	X2	X3	X1X2	X1X3	X2X3	X1X2X3
N(0,1)		383	1	0	1270	33	0	313
	σ							
p = 0.05	2	475	8	0	1176	47	1	293
	3	472	13	0	1219	38	1	257
	4	484	19	0	1155	36	2	304
	5	478	14	0	1198	39	1	270
	10	496	37	0	1182	25	1	259
p = 0.10	2	477	21	0	1182	43	0	277
	3	516	31	0	1142	50	2	259
	4	528	51	0	1109	39	6	267
	5	603	51	0	1061	53	1	231
	10	638	83	0	1043	41	4	190
p = 0.25	2	617	40	0	1043	64	2	234
	3	741	91	0	894	70	10	194
	4	819	149	0	792	59	12	169
	5	888	197	0	697	70	17	131
	10	959	327	0	536	59	21	85

Tabela 3.6 - Número de amostras (em 2000) com um dado modelo final selecionado com o ajuste obtido com o estimador biponderado com c = 5.0 quando $\varepsilon \sim p\sigma\varphi(x/\sigma) + (1-p)\varphi(x)$.

biponderado c = 5.0

		variáveis no modelo final						
		X1	X2	X3	X1X2	X1X3	X2X3	X1X2X3
N(0,1)		366	5	0	1420	16	0	193
	σ							
p = 0.05	2	406	10	0	1362	23	0	199
	3	475	8	0	1321	23	1	172
	4	469	16	0	1301	28	0	186
	5	473	25	0	1295	14	1	192
	10	480	37	0	1309	17	0	157
p = 0.10	2	474	13	0	1300	18	0	195
	3	539	31	0	1237	30	1	162
	4	605	44	0	1155	29	2	165
	5	581	77	0	1154	27	2	159
	10	650	100	0	1079	24	4	142
p = 0.25	2	599	44	0	1147	40	1	169
	3	770	104	0	946	52	5	123
	4	818	174	0	824	50	9	125
	5	845	206	0	786	61	19	83
	10	938	340	2	557	55	21	68

O aumento de p deteriora esta capacidade de se chegar ao modelo correto. Para p = 0.25, as frequências acima decaíram para 61.2%, 44,1%, 34.4%, 25.1% e 11.1%, respectivamente. Verifica-se, paralelamente, um aumento nas frequências a que se chega aos modelos X1 e X2.

Para p = 0.25 e σ = 10, verificou-se inclusive uma ocorrêcia bizarra quando, por uma vez, se chegou ao modelo que inclui apenas a variável X3. Outra situação curiosa é que para σ = 10, o modelo Y = c, que não inclue nenhuma das variáveis, ocorre 1, 2 e 20 vezes, para p = 0.05, 0.10 e 0.25, respectivamente.

A tabela 3.5 é a equivalente à tabela anterior, agora usando o estimador biponderado com c = 3.0, verificando-se um padrão semelhante com as frequências de ocorrência de cada um dos modelos, à medida que p e σ variam. O modelo correto, X1X2, ocorre com mais frequência nas situações de contaminação leve (p e σ pequenos) à medida que p e σ crescem a frequência do modelo alternativo cai, em favor de outros modelos razoáveis, como X1 e X2. Também aqui, para σ = 10, modelo Y = c ocorreu algumas vezes : exatamente 1 e 13 vezes, respectivamente para p = 0.10 e p = 0.25.

A tabela 3.6 traz resultados análogos, agora para o estimador biponderado com c = 5.0. Os padrões já observados nos dois casos anteriores persistem. A queda na frequência de X1X2 quando σ aumenta, para p = 0.05 não é significativa.

CAPÍTULO 4

COMENTÁRIOS E CONCLUSÕES

4.1 - DESEMPENHO COMPARATIVO DOS ESTIMADORES

Conforme se pode observar das tabelas 2.2 a 2.5, os dois estimadores estudados, o de mínimos quadrados e o biponderado, em suas duas versões, com c = 3.0 e com c = 5.0, não apresentam diferenças de comportamento com respeito ao valor médio esperado das estimativas. Este resultado já era esperado, dado o fato de que ambos os estimadores são não viciados.

Com respeito à variância dos estimadores, contudo, verifica-se uma forte diferenciação entre o estimador de mínimos quadrados e o biponderado, à medida que σ^2 cresce. As duas opções do biponderado não apresentam grande diferenciação entre si.

Contudo é interessante notar que, embora diferindo muito pouco em termos de eficiência relativa, as duas opções do biponderado possuem contextos específicos em que cada uma apresenta melhor desempenho. Assim, de 48 situações diferentes em que foram comparados (os 3 parâmetros vezes 3 valores de p, vezes 5 valores de σ, mais os 3 parâmetros no caso da normal pura), apresentados nas tabelas 2.4 e 2.5, a versão c = 5.0 foi melhor 26 vezes, contra 22 da opção c = 3.0. Este equilíbrio, contudo, não deixa de revelar que a opção c = 3.0 foi sistematicamente melhor nos contextos de contaminação pesada, caracterizada por p e σ grandes. A opção c = 5.0 ganhou sistematicamente no contexto oposto.

Comparando em conjunto o biponderado com o estimador de mínimos quadrados, vemos que a variância deste cresce rapidamente à medida

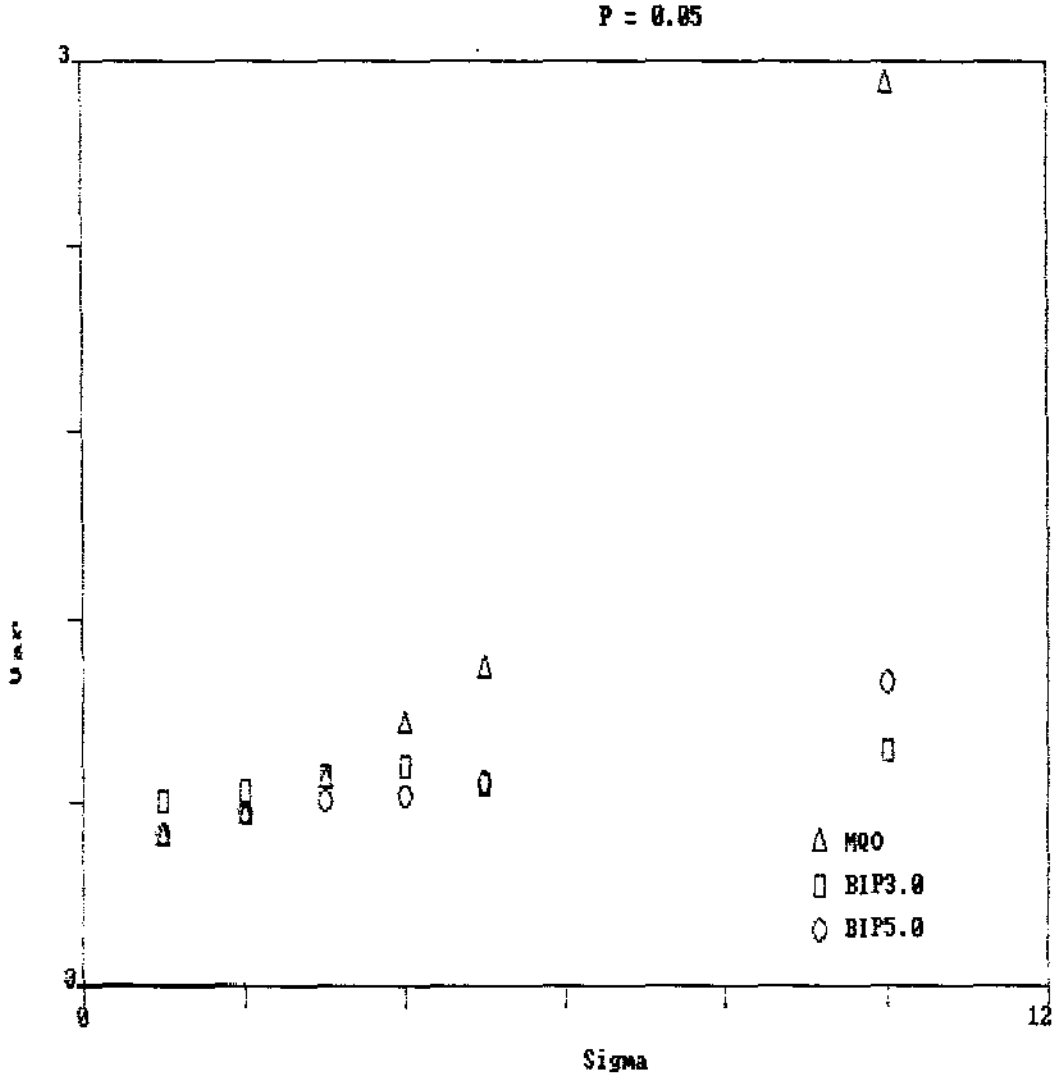


Figura 4.1a - Variâncias dos estimadores de β1 para p = 0.05.

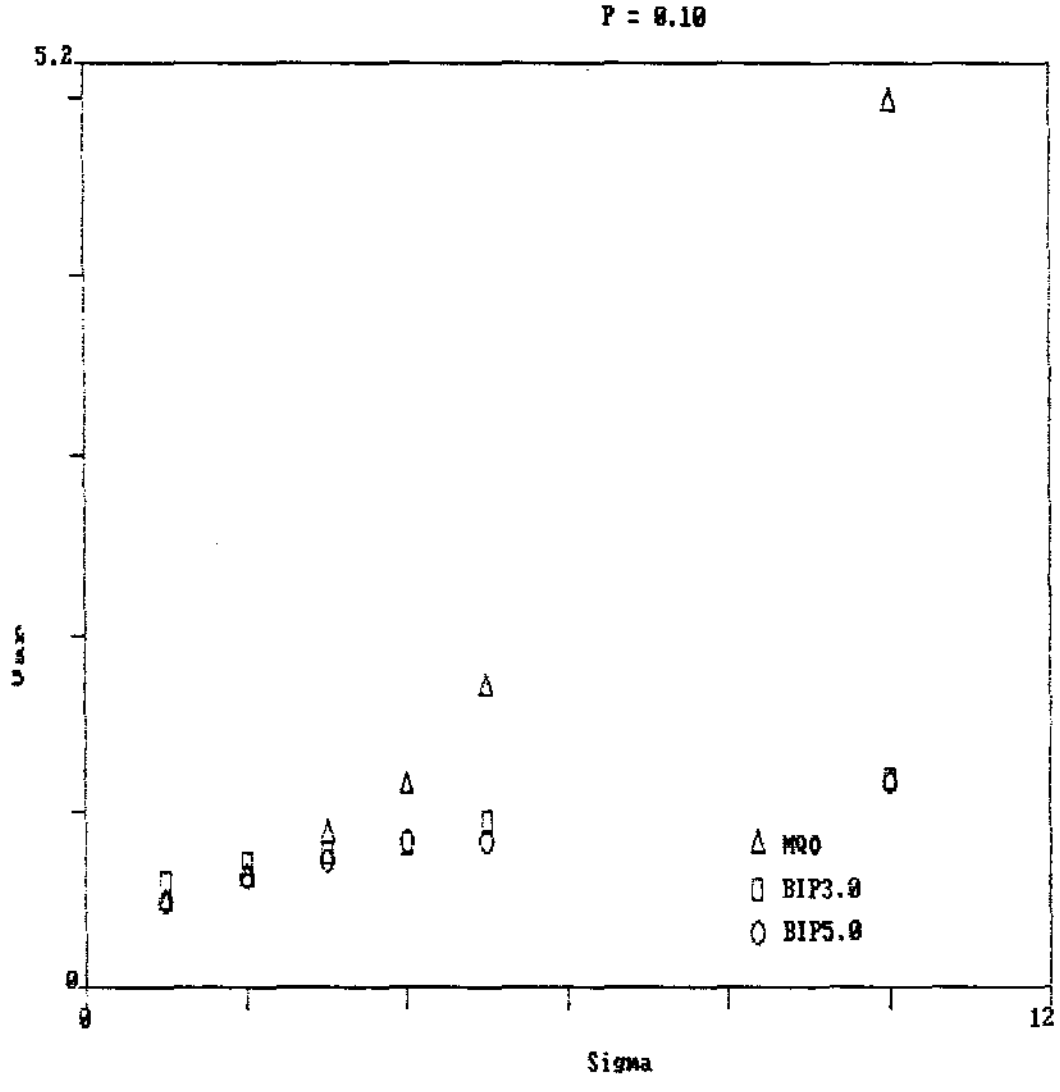


Figura 4.1b - Variâncias dos estimadores de $\beta 1$ para p = 0.10.

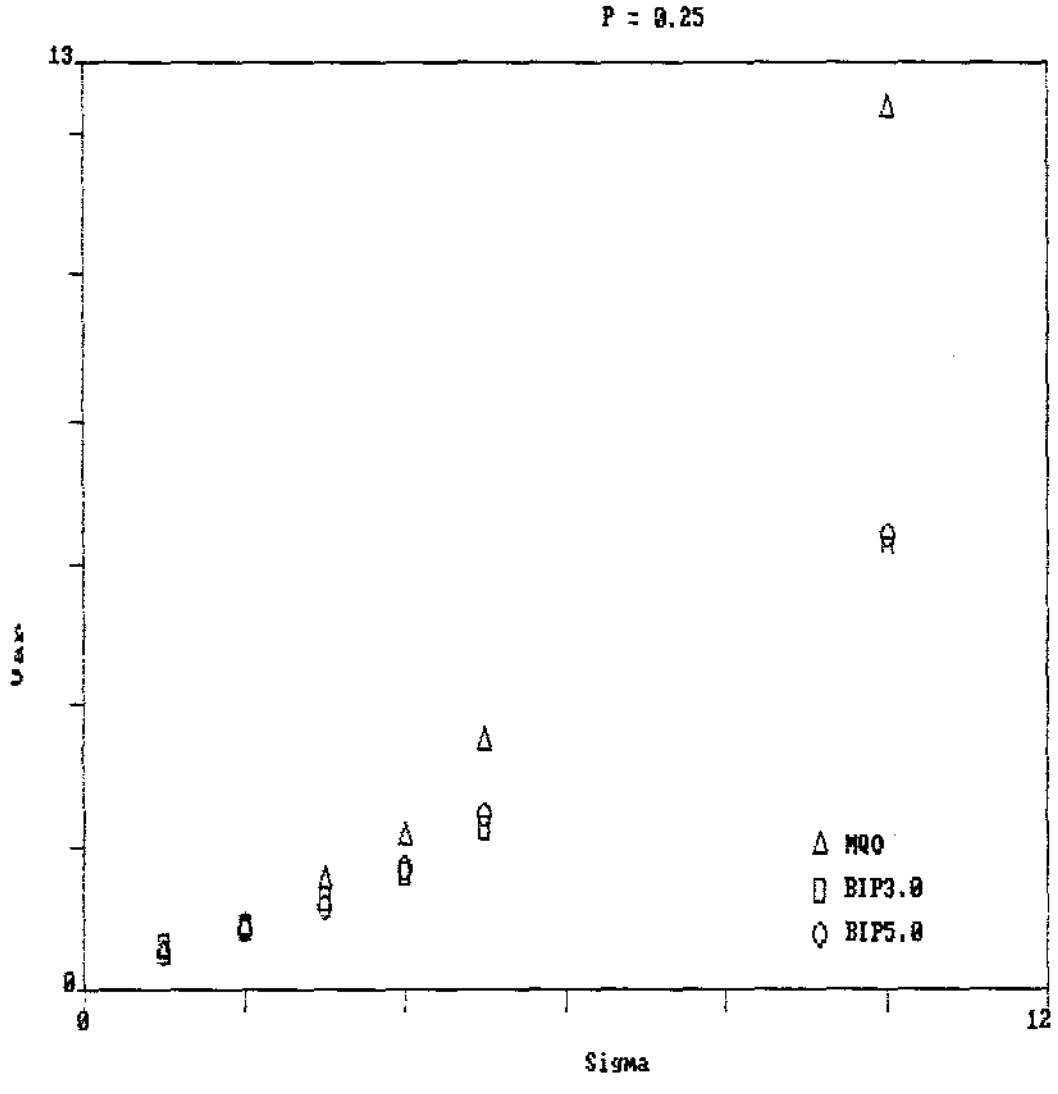


Figura 4.1c - Variâncias dos estimadores de β1 para p = 0.25.

que σ aumenta, destacando-se fortemente das variâncias para as duas opções utilizadas do biponderado. Este padrão se repete para os três valores de p.

A medida que p cresce, cresce também as variâncias envolvidas. Para p = 0.05 e p = 0.10, o crescimento de σ não afeta significativamente as variâncias dos biponderados, ao contrário do que ocorre para o estimador de mínimos quadrados, conforme vimos. Tal fato revela a característica estabilizadora do biponderado com relação a crescimento dos pesos das caudas das distribuições dos erros.

Para p = 0.25, contudo, já se verifica um crescimento das variâncias dos biponderados à medida que o cresce. Tal crescimento é bastante acentuado, embora ainda bem inferior ao crescimento das variâncias do estimador de mínimos quadrados.

Tais fatos são ilustrados na figura 4.1. Nela estão representadas as variâncias de b1, para o estimador de mínimos quadrados e para as duas versões consideradas do biponderado, como função de σ . Cada gráfico corresponde a um dos valores de p.

Os gráficos para as variâncias de bo e de b2 apresentariam padrões muito semelhantes e não são, por isto, apresentados.

4.2 - PODER E NIVEL DE SIGNIFICANCIA

Observando as tabelas 3.1 a 3.3, que apresentam as frequências com que as variáveis foram incluídas aos níveis de significância nominais de 0.10, 0.05 e 0.01, pode-se tirar conclusões a respeito do poder e do nível de significância reais dos testes envolvidos no processo empregado para seleção de variáveis.

Verificando-se, por exemplo, a frequência com que a variável X3 foi incluída no modelo, obtem-se indicações sobre o nível de signi-

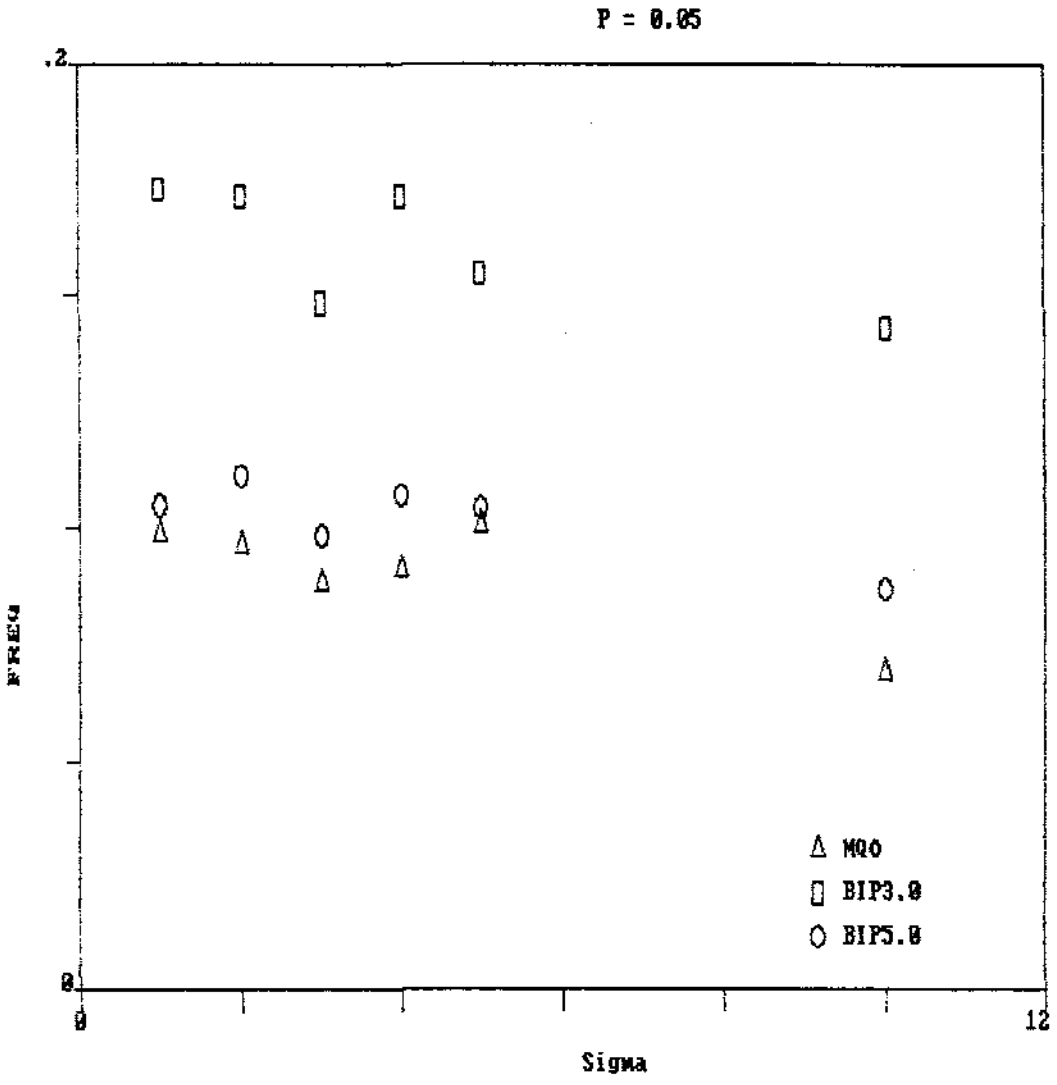


Figura 4.2a - Nível de significância real associado a variável X3 para $\alpha = 0.10$ e $p = 0.05$.

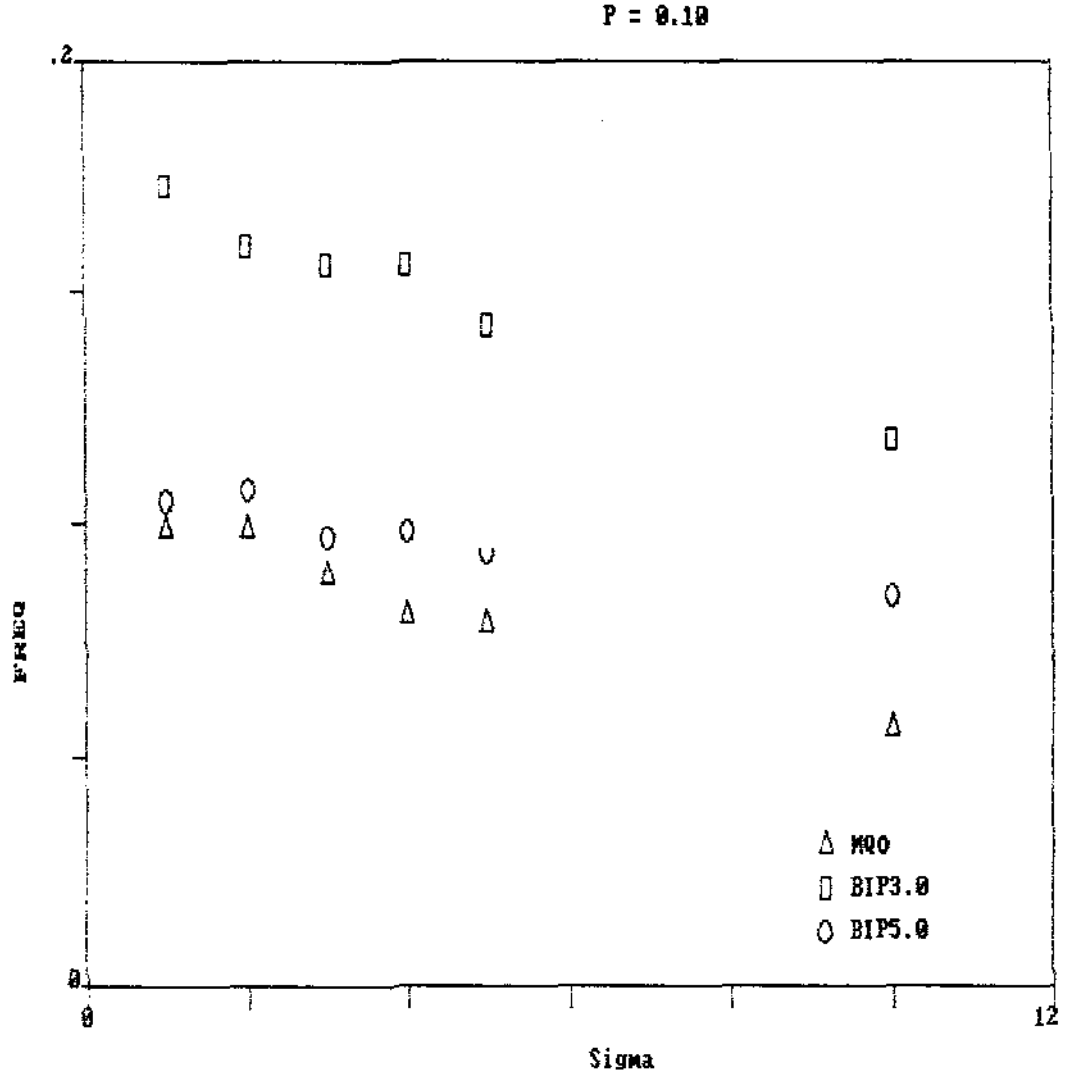


Figura 4.2b - Nível de significância real associado a variável X3 para $\alpha = 0.10$ e $p = 0.10$.

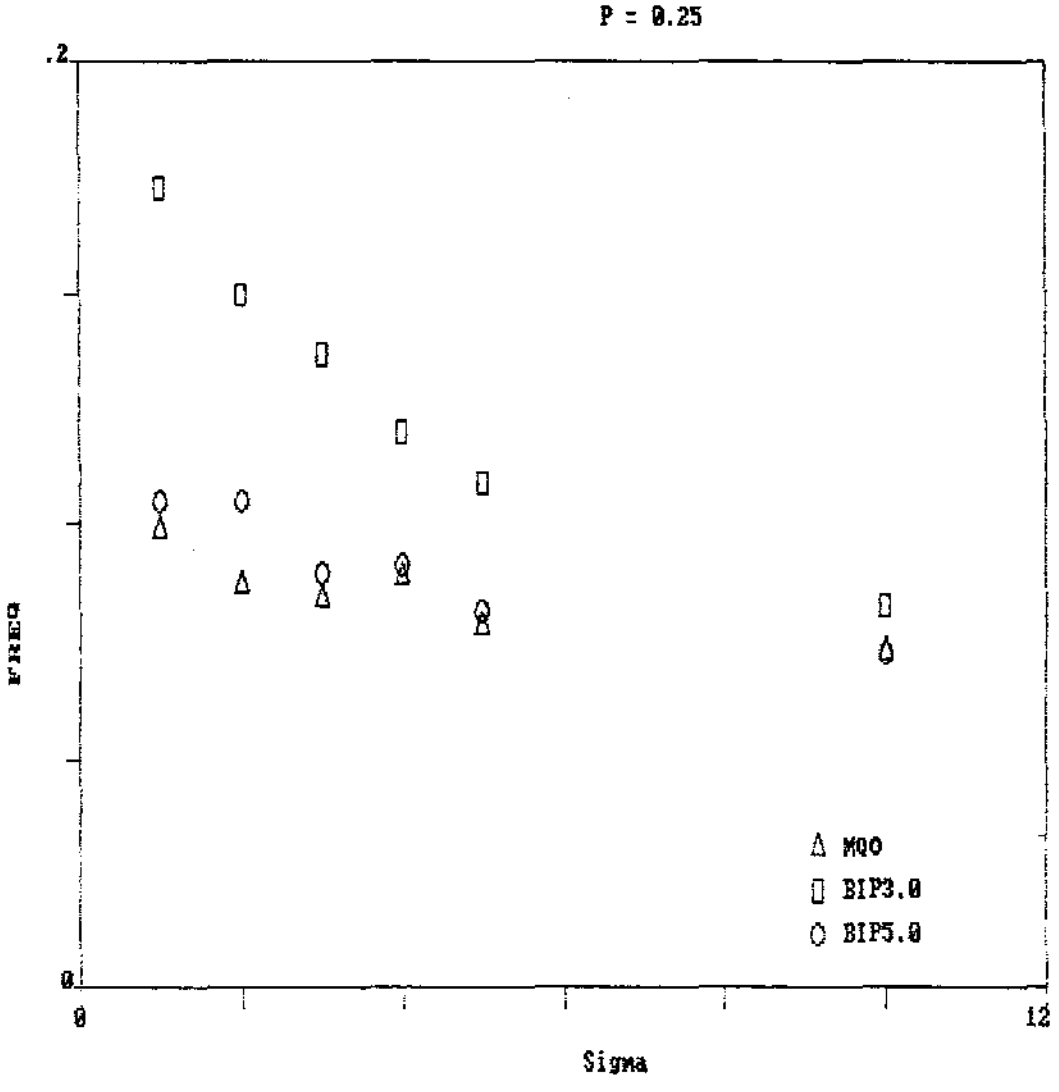


Figura 4.2c - Nível de significância real associado a variável X3 para $\alpha = 0.10$ e $p = 0.25$.

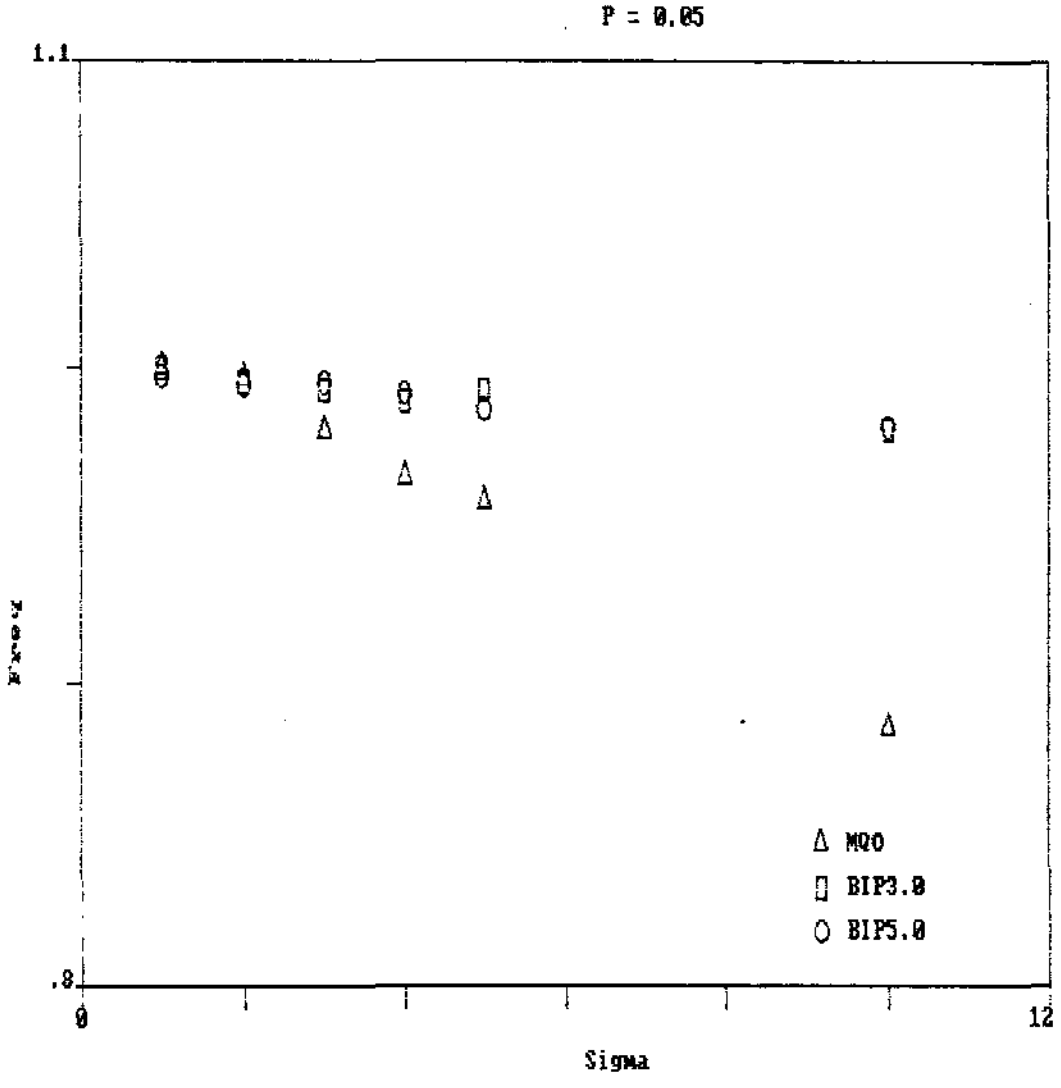


Figura 4.3a - Poder associado à variável X1 para p = 0.05.

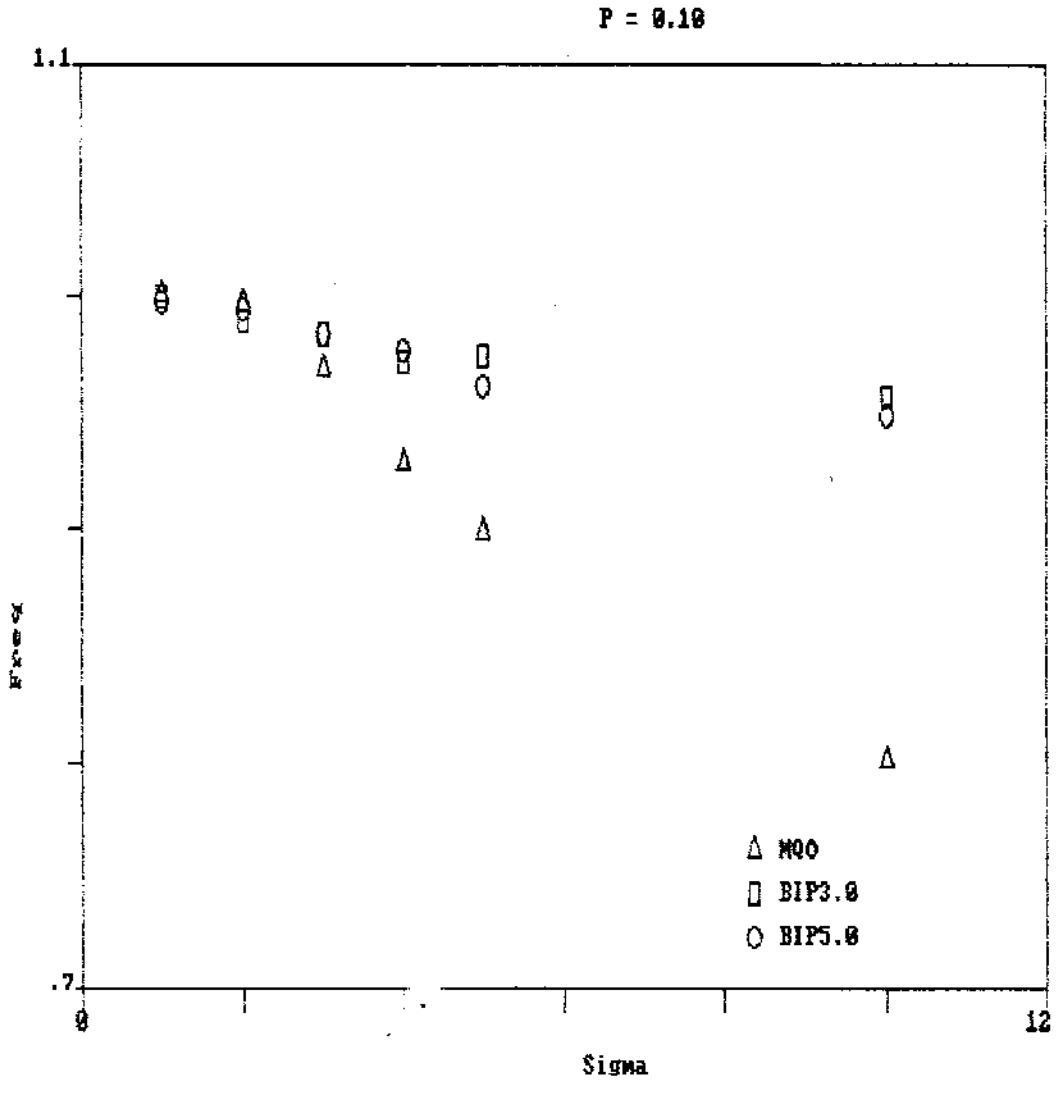


Figura 4.3b - Poder associado à variável X1 para p = 0.10.

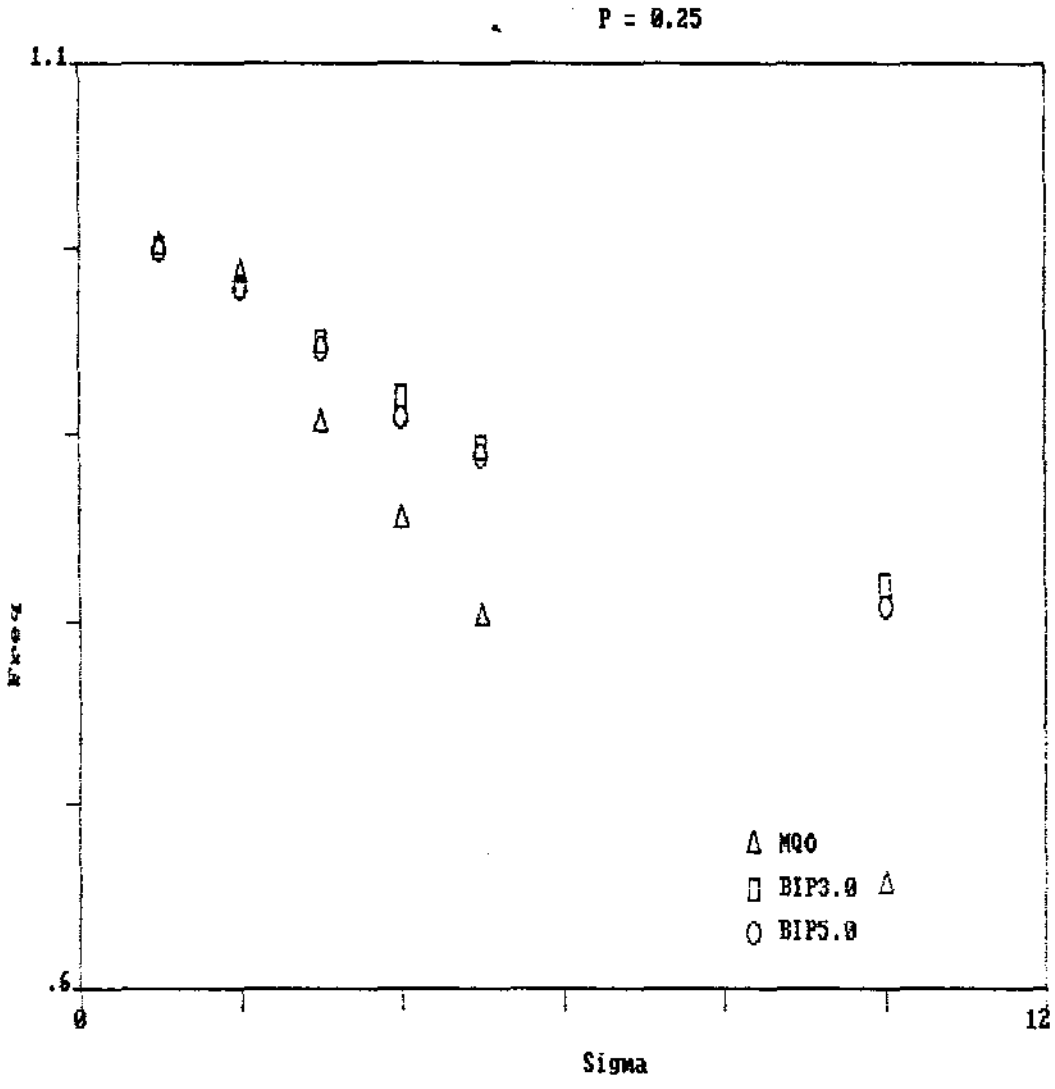


Figura 4.3c - Poder associado à variável Xl para p = 0.25.

ficância associado aos testes para inclusão de variáveis já que $\beta 3 = 0$.

Com as estimativas do biponderado com c = 3.0 o nível de significância real está bastante afastado do nominal, fato este que não ocorre para os mínimos quadrados ordinários e biponderado com c = 5.0. Na maior parte dos casos os dois últimos não apresentam um resultado muito diferente. Temos que próximo a normalidade, isto é, com p e σ pequenos, o estimador de mínimos quadrados ordinários apresenta um nível real menor que o nível nominal. Tal comportamento já era esperado, dado que no processo de seleção automática empregado, toma-se para critério de entrada o F máximo. Já para o biponderado com c = 5.0, o nível de significância real é sempre maior que o verificado para o MQO. Embora também neste caso o nível de significância vá reduzindo à medida que p ou σ aumentam, a variação é menor que a verificada para o MQO.

Observando as frequências de ocorrência das variáveis X1 e X2 no modelo final, pode-se ter uma idéia sobre o poder dos teste envolvidos, já que tanto $\beta 1$ como $\beta 2$ são diferentes de 0. Em termos da variável 1 temos que os mínimos quadrados ordinários, à exceção dos contextos de proximidade da normalidade, tem um desempenho inferior ao biponderado. Entre c = 3.0 e c = 5.0 não há muita diferença. Além do fato de o biponderado ter um poder consistentemente maior do que mínimos quadrados ordinários, deve-se observar que a deterioração não é tão grande quando p ou σ aumentam.

Com a variável X2, não se observa muita diferença no poder entre os três métodos.

As figuras 4.2 e 4.3 ilustram parcialmente os fatos comentados acima. As figuras 4.2 mostram as variações, com σ , dos níveis de significância reais associados a X3, para $\alpha = 0.10$, para o MQO e as duas versões do biponderado, para cada valor de p. As figuras 4.3 fazem o mesmo com relação ao poder relativo à variável X1.

4.3 - MODELO FINAL

Nas tabelas 3.4, 3.5 e 3.6 observa-se o número de vezes em que se obteve cada um dos modelos finais possíveis, para o ajuste obtido através de mínimos quadrados ordinários, estimador biponderado com c = 3.0 e estimador biponderado com c = 5.0 respectivamente.

Com relação ao modelo correto, X1X2, tem-se que, fixado p, para σ menores os resultados obtidos por mínimos quadrados ordinários são melhores enquanto que para σ maiores os resultados com os dois estimadores biponderados superam os de mínimos quadrados. Comparando-se apenas os estimadores biponderados tem-se que, com c = 5.0, o modelo correto é selecionado mais vezes do que para c = 3.0.

O ajuste por mínimos quadrados leva a uma maior incidência do modelo X2X3 do que o biponderado principalmente quando p = 0.25.

O ajuste que menos leva à escolha do modelo X1X2X3 vem a ser o de mínimos quadrados.

Os resultados com o estimador biponderado se assemelham mais entre si do que com os de mínimos quadrados.

B I B L I O G R A F I A

- ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. & TUKEY, J. W. (1972), Robust Estimates of Location, Princeton University Press.

- BEATON, A. E. & TUKEY, J. W. (1974), The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data, Technometrics, 16, 147-185.

- BERK, K. N. (1978), Comparing Subset Regression Procedures, Technometrics, 20, no.1, 1-6.

- BELSLEY, D. A., KUH, E. & WELSCH, R. E. (1980), Regression Diagnostics, John Wiley & Sons.

- BLOOMFIELD, P. & STEIGER, W. L. (1983), Least Absolute Deviations - Theory, Applications and Algorithms - Birkhauser

- BUSTOS, O. (1981), Estimação Robusta no Modelo de Posição, Notas do 13º Colóquio Brasileiro de Matemática.

- CORDEIRO, G. M. (1986), Modelos Lineares Generalizados, Notas do VII SINAPE.

- DACHS, J. N. W. (1978), Análise de Dados e Regressão, IMECC UNICAMP

- DIEHR, G. & HOFLIN, D. R. (1974), Approximating the Distribuition of the Sample R2 in Best Subset regression, Technometrics, 16, no.2, 317-320.

- DRAPER, N. R., GUTTMAN, I. & KANEMASU, H. (1971), Biometrika, 58, no.2, 295-298.

- DRAPER, N. R. & SMITH, H. (1981), Applied Regression Analysis, 2nd. ed., John Wiley & Sons.

- HOCKING, R. R. (1976), The Analysis aand Selection of Variables in Linear Regression, Biometrics, 32, no.1, 1-50.

- HOERL, R. W., SCHUENEMEYER, J. H. & HOERL, A. E. (1986), A Simulation of Biased Estimation and Subset Regression Technics, Technometrics, 28, no.4, 369-380.

- HUBER, P. J. (1981), Robust Statistics, John Wiley & Sons.

- LARSEN, W. A. & MCCLEARY, S. J. (1972), The Use of Partial Residual Plots in Regression Analysis, Technometrics, 14, 781-790.

- MOSTELLER, F. & TUKEY, J. W. (1977), Data Analysis and Regression - Addison-Wesley

- NELDER, J. A. & WEDDERBURN, R. W. M. (1972), Generalized Linear Models, J. R. Statist. Society A, 135, 370-384.

- RENCHER, A. C. & PUN, F. C. (1980), Inflation of R2 in Best Subset Regression, Technometrics, 22, no.1, 49-53.

- ROCKE, D. M. & SHANNO, D. F. (1986), The Scale Problem in Robust Regression M-Estimates, J. Statist. Comput. Simul., 24, 47-69.

- SCAFI, M. A. O. (1979), Regressão Biponderada - Um Método Robusto de Ajuste, Tese de Mestrado, UNICAMP.

- SCHEFFÉ, H. (1959), The Analysis of Variance, John Wiley & Sons.

- WILKINSON, L. & DALLAL, G. E. (1981), Teste of Significance in Forward Selection Regression With an F-to-Enter Stopping Rule, Technometrics, 23, no.4, 377-380.

A P Ê N D I C E

Apresenta-se aqui a listagem do programa em Pascal utilizado para se obter os resultados. Este programa gera as amostras e faz a regressão biponderada para o conjunto de dados gerados. Para os outros resultados necessários utilizou-se programas obtidos a partir de modificações neste.

```
PROGRAM  programa_completo (INPUT,OUTPUT) ;


(*****                                          *****)
(** Declaracao de variaveis e constantes **)
(*****                                          *****)


CONST
  p = 3 ;    (** numero de variaveis a menos da media **)
  n = 10 ;   (** numero de observacoes **)


TYPE
  matriz1003 = ARRAY[1..10,0..4] of REAL ;
  matriz0303 = ARRAY[0..4,0..4] of DOUBLE ;
  vetor03    = ARRAY[0..4] of DOUBLE ;
  vetor10    = ARRAY[1..10] of DOUBLE ;
  vetboo     = ARRAY[0..10] of BOOLEAN ;
  BYTE       = 0..255 ;

VAR
  reserv   : matriz1003 ;(** matriz X **)
  ce       : REAL ;
  tab      : REAL ;(** valor F de entrada de variaveis **)
  s        : DOUBLE ;(** amplitude interquartis **)
  acres    : vetor10 ;(** vetor dos valores de F devido a inclusao das
                         variaveis **)
  nvinc    : INTEGER ;(** numero de variaveis ja incluidas **)
  varinc   : ARRAY[0..3] of INTEGER ;(** indicador das variaveis que ainda
                                        nao foram incluidas **)
  ii       : INTEGER ;(** contador **)
  dmin     : vetor10 ;(** tolerancias  das variaveis **)
  varmod   : ARRAY[0..3] of INTEGER ;(** -1 qdo. variavel no modelo 1 caso
                                        contrario **)
  inc      : BOOLEAN ;(** diz se a variavel deve ou nao ser incluida **)
  jj       : INTEGER ;(** contador **)
  achou    : BOOLEAN ;(** variavel aux. para atualizar o vetor varinc **)
  kk       : INTEGER ;(** contador **)
  pronto   : BOOLEAN ;(** indica se o processo de inclusoes ja terminou **)
  rej      : ARRAY[1..3,1..3] of INTEGER ;(**indica a qtos por cento a
                                             variavel entrou **)
  ll       : INTEGER ;(** conta o numero de amostras geradas **)
  modfim   : ARRAY[0..7] of INTEGER ;(** diz quantas amostras terminaram em
                                         cada modelo **)
  residuo  : vetor10 ; (** vetor de residuos **)
  sigma    : REAL ;(** o desvio padrao da contaminante **)
  pcont    : REAL ;(** percentagem de contaminacao **)
  semente  : INTEGER ;(** semente da sequencia de numeros aleatorios **)
  F        : ARRAY[1..10,1..3] of REAL ;(** matriz dos valores F
                                            tabelados **)
```

```
(*****                                                *****)
(*** Declaracao dos valores tabelados de F ***)
(*****                                                *****)

PROCEDURE declara_F ;

BEGIN
  F[1,1] := 39.863 ; F[1,2] := 161.45 ; F[1,3] := 4052.2 ;
  F[2,1] := 8.5263 ; F[2,2] := 18.513 ; F[2,3] := 98.503 ;
  F[3,1] := 5.5383 ; F[3,2] := 10.128 ; F[3,3] := 34.116 ;
  F[4,1] := 4.5448 ; F[4,2] := 7.7086 ; F[4,3] := 21.198 ;
  F[5,1] := 4.0604 ; F[5,2] := 6.6079 ; F[5,3] := 16.258 ;
  F[6,1] := 3.7759 ; F[6,2] := 5.9874 ; F[6,3] := 13.745 ;
  F[7,1] := 3.5894 ; F[7,2] := 5.5914 ; F[7,3] := 12.246 ;
  F[8,1] := 3.4579 ; F[8,2] := 5.3177 ; F[8,3] := 11.259 ;
  F[9,1] := 3.3603 ; F[9,2] := 5.1174 ; F[9,3] := 10.561 ;
  F[10,1] := 3.2850 ; F[10,2] := 4.9646 ; F[10,3] := 10.044 ;
END ;



(*****                                                *****)
(*** Inicializa a geracao de numeros aleatorios ***)
(*****                                                *****)


[EXTERNAL,asynchronous] FUNCTION MTH$RANDOM(VAR seed : INTEGER) :REAL;
                        EXTERN;


(*****                    *****)
(*** Gera uma amostra ***)
(*****                    *****)


PROCEDURE  cria_dados ;

VAR
  A    : ARRAY[1..3] of REAL ;
  i,j  : INTEGER ;
  erro : REAL ;

PROCEDURE  cria_erro (VAR  normal : REAL)  ;

VAR
  uni1      : REAL ;
  uni2      : REAL ;


BEGIN
  uni1 := MTH$RANDOM(semente) ;
  uni2 := MTH$RANDOM(semente) ;
  uni1 := SQRT(2*LN(1/(1-uni1))) ;
  uni2 := 3.14159*(2*uni2-1) ;
  normal := uni1*SIN(uni2) ;
  IF MTH$RANDOM(semente) < pcont THEN normal := sigma*normal ;
END ;
```

```
BEGIN
  FOR i := 1 TO 5 DO
  BEGIN
    reserv[i,1] := i ;
    reserv[i+5,1] := i ;
  END ;
  FOR i := 1 TO 5 DO
    reserv[i,3] := 1 ;
  FOR i := 6 TO 10 DO
    reserv[i,3] := -1 ;
  FOR i := 1 TO 2 DO
  BEGIN
    reserv[2*i-1,2] := i ;
    reserv[2*i,2] := i ;
    reserv[2*i+4,2] := i ;
    reserv[2*i+5,2] := i ;
  END ;
  reserv[5,2] := 3 ;
  reserv[10,2] := 3 ;
  A[1] := 4 ;
  A[2] := 4 ;
  A[3] := 0 ;
  FOR i := 1 TO n DO
  BEGIN
    reserv[i,0] := 1 ;
    reserv[i,p+1] := 4 ;
    FOR j := 1 TO p DO
      reserv[i,p+1] := reserv[i,p+1] + A[j]*reserv[i,j] ;
    cria_erro (erro) ;
    reserv[i,p+1] := reserv[i,p+1] + erro ;
  END ;
END ;



(*****                                                    *****)
(*** Regressao biponderada com as variaveis necessarias ***)
(*****                                                    *****)


PROCEDURE  regressao_biponderada ;


VAR
  pesos    : vetor10 ;(** pesos das obs. para o processo iterativo **)
  i        : BYTE ;(** contador **)
  nint     : INTEGER ;(** numero de iteracoes **)
  fim      : BOOLEAN ;(** indica se o processo iterativo ja convergiu **)
  bini     : ARRAY[0..10] of DOUBLE ;(** est. dos parametros no inicio do
                                         i-esimo passo **)
  bchap    : vetor03 ;(** est. dos parametros no fim do i-esimo passo **)
  j        : BYTE ;(** contador **)
  xaum     : matriz0303 ;(** matriz ampliada ja sweepada no zero **)
```

```
(*****                                          *****)
(*** Faz o sweep nas linhas necessarias ***)
(*****                                          *****)


PROCEDURE  sweepa_tudo ;

VAR
  ivar    : ARRAY[0..21] of INTEGER ;(** -1 qdo. a linha ja foi sweepada 1
                                        caso contrario **)
  k       : BYTE ;(** contador **)
  l       : BYTE ;(** contador **)


PROCEDURE  sweep(r : INTEGER) ;

VAR
  m       : BYTE ;(** contador **)
  t       : BYTE ;(** contador **)
  d,b,c   : DOUBLE ;
  colin   : vetboo ;(** indica se ha colinearidade **)

BEGIN
  FOR m := 1 TO p DO
    colin[m] := FALSE ;
  d := xaum[r,r] ;
  IF ((d < dmin[r]) AND (ivar[r] = 1)) THEN colin[r] := TRUE ;
  IF NOT colin[r] THEN
  BEGIN
    FOR m := 0 TO p+1 DO
    BEGIN
      IF m <> r THEN
      BEGIN
        IF m > r THEN b := ivar[m]*ivar[r]*xaum[r,m]/d ;
        IF m < r THEN b := xaum[m,r]/d ;
        FOR t := m TO p+1 DO
        BEGIN
          IF t <> r THEN
          BEGIN
            IF t < r THEN c := ivar[t]*ivar[r]*xaum[t,r] ;
            IF t > r THEN c := xaum[r,t] ;
            xaum[m,t] := xaum[m,t] - b*c ;
          END ;
        END ;
      END ;
    END ;
    IF r <> 0 THEN FOR m := 0 TO r-1 DO
      xaum[m,r] := -xaum[m,r]/d ;
    FOR m := r+1 TO p+1 DO
      xaum[r,m] := xaum[r,m]/d ;
    xaum[r,r] := 1/d ;
    ivar[r] := -ivar[r] ;
  END ;
END ;
```

```
BEGIN
  ivar[0] := -1 ;
  FOR k := 1 TO p+1 DO
    ivar[k] := 1 ;
  k := 0 ;
  l := i ;
  WHILE k < nvinc DO
  BEGIN
    IF (varmod[l] = -1) THEN
    BEGIN
      sweep(l) ;
      k := k +1 ;
    END ;
    l := l + 1 ;
  END ;
  IF inc = TRUE THEN
  BEGIN
    sweep(ii) ;
  END
  ELSE
    sweep(varinc[ii]) ;
  FOR k := 0 TO p DO
    bchap[k] := xaum[k,p+1] ;
END ;



(*****                                    *****)
(*** Calcula a matriz (X'X|X'Y) ***)
(*****                                    *****)


PROCEDURE  cria_matriz ;

VAR
  k     : BYTE ;(** contador **)
  l     : BYTE ;(** contador **)
  xx    : vetor03 ;(**  vetor auxiliar para guardar os dados de uma
                       observacao **)
  m     : BYTE ;(** contador **)

BEGIN
  FOR k := 1 TO p+1 DO
  BEGIN
    xaum[0,k] := reserv[1,k] ;
    FOR l := k TO p+1 DO
      xaum[k,l] := 0 ;
  END ;
  FOR k := 2 TO n DO
  BEGIN
    FOR l := 1 TO p+1 DO
    BEGIN
      xx[l] := reserv[k,l] ;
    END ;
```

```
    FOR l := i TO p+1 DO
    BEGIN
      FOR m := 1 TO p+1 DO
        xaum[l,m] := xaum[l,m] + ((xx[l]-xaum[0,l])/k)*((xx[m]-
                     xaum[0,m])*(k-1)) ;
    END ;
    FOR l := i TO p+1 DO
      xaum[0,l] := ((k-1)*xaum[0,l] + xx[l])/k ;
  END ;
  xaum[0,0] := i/n ;
  FOR k := 0 TO p+1 DO
  BEGIN
    IF k = 0 THEN
    BEGIN
      FOR l := k+1 TO p+1 DO
        xaum[l,k] := - xaum[k,l] ;
    END
    ELSE
    BEGIN
      FOR l := k+1 TO p+1 DO
        xaum[l,k] := xaum[k,l] ;
    END ;
  END ;
END ;



(*****                                                  *****)
(*** Calcula os residuos apos uma iteracao ***)
(*****                                                  *****)


PROCEDURE  calc_residuo ;

VAR
  v     : BYTE ; (** contador **)
  x     : BYTE ; (** contador **)

BEGIN
  FOR v := 1 TO n DO
    residuo[v] := 0 ;
  IF NOT inc THEN
  BEGIN
    FOR v := 1 TO n DO
    BEGIN
      FOR x := 0 TO p DO
      BEGIN
        IF ((varmod[x] = -1) OR (x = varinc[i,j])) THEN
        residuo[v] := residuo[v] + reserv[v,x]*xaum[x,p+1] ;
      END ;
      residuo[v] := reserv[v,p+1] - residuo[v] ;
    END ;
  END
  ELSE
```

```
  BEGIN
    FOR v := 1 TO n DO
    BEGIN
      FOR x := 0 TO p DO
      BEGIN
        IF ((varmod[x] = -1) OR (x = ii)) THEN
        residuo[v] := residuo[v] + reserv[v,x]*xaum[x,p+1] ;
      END ;
      residuo[v] := reserv[v,p+1] - residuo[v] ;
    END ;
  END ;
END ;



(*****                                              *****)
(*** Faz o ajuste por minimos quadrados ***)
(*****                                              *****)


PROCEDURE  minimos_quadrados ;


BEGIN
  cria_matriz ;
  sweepa_tudo ;
  calc_residuo ;
END ;



(*****                                              *****)
(*** Calcula a amplitude interquartis ***)
(*****                                              *****)


PROCEDURE  calc_S ;

VAR
  k,l      : BYTE ;
  ind,ord  : INTEGER ;
  yor      : vetor10 ;
  h1,h2    : DOUBLE ;
  aux      : DOUBLE ;
  ninv     : BOOLEAN ;

BEGIN
  FOR k := 1 TO n DO
    yor[k] := residuo[k] ;
  l := n-1 ;
  repeat
    ninv := false ;
    for k := 1 to l
        do begin
             if yor[k] > yor[k+1]
```

```
                then begin
                       ninv := true ;
                       aux := yor[k] ;
                       yor[k] := yor[k+1] ;
                       yor[k+1] := aux ;
                     end ;
          end ;
    l := l - 1 ;
  until not ninv ;

  k := n DIV 4 ;
  IF n MOD 4 = 0 THEN
  BEGIN
    h1 := (yor[k] + yor[k+1])/2 ;
    h2 := (yor[n+1-k] + yor[n-k])/2 ;
  END
  ELSE
  BEGIN
    l := (n+3) DIV 4 ;
    h1 := yor[l] ;
    h2 := yor[n+1-l] ;
  END ;
  s := (h2-h1)/1.35 ;
END ;



(*****                                                              *****)
(*** Calcula os pesos a serem utilizados na proxima iteracao ***)
(*****                                                              *****)


PROCEDURE  calc_Pi ;

VAR
  k           : BYTE ;(** contador **)
  pesos_k     : DOUBLE ;
  ce_vezes_s  : DOUBLE ;
  varinc_ii   : REAL ;

BEGIN
  ce_vezes_s := ce*s ;
  varinc_ii := varinc[i,i] ;
  FOR k := 1 TO n DO
  BEGIN
    pesos_k := residuo[k] ;
    pesos_k := (pesos_k)/ce_vezes_s ;
    IF SQR(pesos_k) > 1 THEN
      pesos_k := 0
    ELSE
      pesos_k := SQR(1 - SQR(pesos_k)) ;
    pesos[k] := pesos_k ;
  END ;
END ;
```

```
(*****                                  *****)
(*** Calcula a matriz (X'PX|X'PY) ***)
(*****                                  *****)


PROCEDURE  calc_XPX ;

VAR
  k     : BYTE ;(** contador **)
  l     : BYTE ;(** contador **)
  xx    : vetor03 ;(** vetor auxiliar para guardar os dados de uma
                      observacao **)
  m     : BYTE ;(** contador **)

BEGIN
  FOR k := 1 TO p+1 DO
  BEGIN
    xaum[0,k] := reserv[1,k] ;
    xaum[0,k] := SQRT(pesos[1])*xaum[0,k] ;
    FOR l := k TO p+1 DO
      xaum[k,l] := 0 ;
  END ;
  FOR k := 2 TO n DO
  BEGIN
    FOR l := 1 TO p+1 DO
    BEGIN
      xx[l] := reserv[k,l] ;
      xx[l] := SQRT(pesos[k])*xx[l] ;
    END ;
    FOR l := 1 TO p+1 DO
    BEGIN
      FOR m := l TO p+1 DO
        xaum[l,m] := xaum[l,m] + ((xx[l]-xaum[0,l])/k)*((xx[m]-
                     xaum[0,m])*(k-1)) ;
    END ;
    FOR l := 1 TO p+1 DO
      xaum[0,l] := ((k-1)*xaum[0,l] + xx[l])/k ;
  END ;
  xaum[0,0] := 1/n ;
  FOR k := 0 TO p+1 DO
  BEGIN
    IF k = 0 THEN
    BEGIN
      FOR l := k+1 TO p+1 DO
        xaum[l,k] := - xaum[k,l] ;
    END
    ELSE
    BEGIN
      FOR l := k+1 TO p+1 DO
        xaum[l,k] := xaum[k,l] ;
    END ;
  END ;
END ;
```

```
(****                                                    *****)
(*** Verifica se ja houve a convergencia ***)
(*****                                                   *****)


PROCEDURE  teste ;

VAR
  k      : BYTE ;

BEGIN
  fim := TRUE ;
  FOR k := 0 TO p DO
  BEGIN
    IF ((varmod[k] = -1) OR (k = varinc[i])) THEN
    BEGIN
      IF (bini[k] = 0) THEN
      BEGIN
        IF ABS(bchap[k] - bini[k]) > 0.0001 THEN
          fim := FALSE
      END
      ELSE
      BEGIN
        IF ABS((bchap[k] - bini[k])/bini[k]) > 0.0001 THEN
          fim := FALSE ;
      END ;
    END ;
  END ;
END ;



(*****                                                                  *****)
(*** Calcula o valor da estatistica F para a variavel candidata ***)
(*****                                                                  *****)


PROCEDURE  calc_acres (q : INTEGER) ;

BEGIN
  acres[q] := xaum[q,p+1]*xaum[q,p+1] ;
  acres[q] := acres[q]/xaum[q,q] ;
  acres[q] := acres[q]*(n-nvinc-2) ;
  acres[q] := acres[q]/xaum[p+1,p+1] ;
END ;


BEGIN
  nint := 0 ;
  minimos_quadrados ;
  calc_S ;
  calc_Pi ;
  fim := false ;
  WHILE (NOT fim) AND (nint < 15)
    DO BEGIN
         nint := nint + 1 ;
```

```
          FOR j := 0 TO p DO
            bini[j] := bchap[j] ;
          calc_XPX ;
          sweepa_tudo ;
          teste ;
          IF NOT fim THEN
          BEGIN
            calc_residuo ;
            calc_S ;
            calc_Pi ;
          END ;
        END ;
  calc_acres(varinc[ii]) ;
END ;



(*****                                                              *****)
(*** Verifica qual variavel levou a uma maior estatistica F ***)
(*****                                                              *****)


PROCEDURE  acres_max ;

VAR
  j       : BYTE ;(** contador que vai indicar qual var deu o maior acres **)
  max     : DOUBLE ;

BEGIN
  ii := 0 ;
  max := 0 ;
  FOR j := 1 TO p DO
  BEGIN
    IF varmod[j] = 1 THEN
    BEGIN
      IF acres[j] > max THEN
      BEGIN
        max := acres[j] ;
        ii := j ;
      END ;
    END ;
  END ;
END ;


(*****                                                          *****)
(*** Verifica se o maximo e maior do que o valor tabelado ***)
(*****                                                          *****)


PROCEDURE  testa_inclusao ;

VAR
  signif,j   : BYTE ;
```

```
BEGIN
  signif := 0 ;
  FOR j := 1 TO 3 DO
  BEGIN
    IF acres[ii] > F[n-nvinc-2,j] THEN
    BEGIN
      signif := signif + 1 ;
    END ;
  END ;
  IF signif > 0 THEN
  BEGIN
    inc := TRUE ;
    pronto := FALSE ;
  END
  ELSE
  BEGIN
    inc := FALSE ;
    pronto := TRUE ;
  END ;
  IF inc = TRUE THEN
  BEGIN
    rej[ii,signif] := rej[ii,signif] + 1 ;
  END ;
END ;


(*****                                                    *****)
(*** Ajeita os indices apos a inclusao de uma variavel ***)
(*****                                                    *****)


PROCEDURE conserta_indices ;

BEGIN
  regressao_biponderada ;
  varmod[ii] := -1 ;
  jj := 1 ;
  achou := FALSE ;
  REPEAT
    IF varinc[jj] <> ii THEN
      jj := jj + 1
    ELSE
    BEGIN
      achou := TRUE ;
      kk := jj ;
    END ;
  UNTIL achou ;
  nvinc := nvinc + 1 ;
  IF (kk <= p-nvinc) THEN
  BEGIN
    FOR jj := kk TO p-nvinc DO
      varinc[jj] := varinc[jj + 1] ;
  END ;
END ;
```

```
(*****                                                    *****)
(*** Verifica qual foi o modelo final escolhido ***)
(*****                                                    *****)


PROCEDURE modelo_final ;

BEGIN
  IF (nvinc = 3) THEN modfim[7] := modfim[7] + 1 ;
  IF (nvinc = 2) THEN
    IF (varmod[1] = 1) THEN modfim[6] := modfim[6] + 1
    ELSE
    IF (varmod[2] = 1) THEN modfim[5] := modfim[5] + 1
    ELSE modfim[4] := modfim[4] + 1;
  IF (nvinc = 1) THEN
    IF (varmod[1] = -1) THEN modfim[1] := modfim[1] + 1
    ELSE
    IF (varmod[2] = -1) THEN modfim[2] := modfim[2] + 1
    ELSE modfim[3] := modfim[3] + 1 ;
END ;



(** PROGRAMA PRINCIPAL **)


BEGIN
  FOR ii := 1 TO 3 DO
    FOR jj := 1 TO 3 DO
      rej[ii,jj] := 0 ;
  mm := 0 ;
  FOR ii := 0 TO 7 DO
    modfim[ii] := 0 ;
  WRITELN('entre o valor de c   ') ;
  READLN(ce) ;
  WRITELN('entre o D.P. da contaminante   ') ;
  READLN(sigma) ;
  WRITELN('entre a percentagem de contaminacao   ') ;
  READLN(pcont) ;
  WRITELN('entre a semente da sequencia aleatoria') ;
  READLN(semente) ;
  declara_F ;
  FOR ll := 1 TO 2000 DO
  BEGIN
    WRITELN('amostra de numero ',ll) ;
    cria_dados ;
    nvinc := 0 ;
    varmod[0] := -1 ;
    FOR ii := 1 TO p DO
    BEGIN
      varinc[ii] := ii ;
      dmin[ii] := 0.00000001 ;
      varmod[ii] := 1 ;
    END ;
    REPEAT
      inc := FALSE ;
```

```
      FOR ii :=1 TO p-nvinc DO
      BEGIN
        regressao_biponderada ;
      END ;
      acres_max ;
      testa_inclusao ;
      IF inc = TRUE THEN conserta_indices ;
    UNTIL (pronto OR (p = nvinc)) ;
    modelo_final ;
  END ;
  WRITELN('Estes sao os resultados com sigma = ',sigma:4:2,' e a
          contaminacao sendo de ',pcont:4:2) ;
  WRITELN ;
  WRITELN('Este e o vetor de modelos finais') ;
  WRITELN ;
  FOR jj := 1 TO 7 DO
    WRITELN(modfim[jj]) ;
  WRITELN ;
  WRITELN('Esta e a tabela com os niveis de significancia de
          entrada das variaveis') ;
  WRITELN ;
  FOR ii := 1 TO 3 DO
  BEGIN
    FOR jj := 1 TO 3 DO
      WRITE(' ',rej[ii,jj]) ;
    WRITELN ;
  END ;
END .
```