ALGUMAS CONSIDERAÇÕES SOBRE REGRESSÃO NÃO LINEAR

DAISY GOMES DE SOUZA

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA A OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

ÁREA DE CONCENTRAÇÃO: ESTATÍSTICA

ORIENTADOR: *Prof. Dr. CLÓVIS DE ARAUJO PERES*

- SÃO PAULO, FEVEREIRO DE 1986 -

A meus pais

AGRADECIMENTOS

Ao Prof. Doutor *Clóvis de Araujo Peres* pela sugestão do tema e disponibilidade na orientação.

Ao Prof. Doutor *Carlos Chien Ching Tu* da Escola Politécnica-USP, pela sua colaboração.

Ao Engenheiro *Edson Gomes* da Escola Politécnica-USP, pela colaboração na parte computacional.

À Professora *Silvia Nagib Elian* do IME-USP, pela sua amizade e constante encorajamento.

Aos demais Professores do IME-USP, pela minha formação.

Ao Sr. *Francisco Gomes da Silva* pelo rápido e excelente trabalho de datilografia.

ÍNDICE

CAPÍTULO 1 - INTRODUÇÃO. 1

CAPÍTULO 2 - PROCEDIMENTOS DE OTIMIZAÇÃO NÃO LINEAR. . . . 7

2.1 - Introdução . . . . . .

2.2 - Aspectos Gerais de um Algoritmo. . . . . .

2.3 - Métodos do Gradiente para Minimização Sem Restrições .

2.3.1 - O Método do "Steepest Descent" . . . . . .

2.3.2 - O Método de Newton . . . . . .

2.3.3 - Métodos de "Quasi-Newton". . . . . .

2.4 - Métodos para Mínimos Quadrados Não Lineares. . . . .

2.4.1 - Introdução . . . . . .

2.4.2 - O Método de Gauss Newton . . . . . .

2.4.3 - Modificações do Algoritmo de Gauss Newton. .

2.4.3.1 - Não linearidade. . . . . .

2.4.3.2 - Mal condicionamento. . . . . .

2.4.4 - Métodos Alternativos para Problemas de Resíduos Grandes . . . . . .

CAPÍTULO 3 - ALGUMAS TÉCNICAS DE OBTENÇÃO DE VALORES INICIAIS 47

3.1 - Introdução . . . . . .

3.2 - Algumas Técnicas de Obtenção de Valores Iniciais . .

3.3 - Exemplos . . . . . .

CAPÍTULO 4 - ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA ATRAVÉS DE MÍNIMOS QUADRADOS NÃO LINEARES 59

4.1 - Introdução. . . . . .

4.2 - Resultados Gerais para a Família Exponencial Regular . .

4.3 - Métodos Gerais. . . . . .

4.4 - Exemplos. . . . . .

CAPÍTULO 5 - MEDIDAS DE NÃO LINEARIDADE 84

5.1 - Introdução. . . . . ..

5.2 - Representação Geométrica dos Modelos de Regressão Não Lineares no Espaço Amostral . . . .

5.3 - Medidas de Curvatura de Bates & Watts . . . .

5.4 - Medida do Viés de Box . . . .

CAPÍTULO 6 - AVALIAÇÃO DAS PROPRIEDADES ESTATÍSTICAS DOS ESTIMADORES DE MÍNIMOS QUADRADOS NÃO LINEARES . 107

6.1 - Introdução.. . . . . ..

6.2 - Propriedades dos Estimadores. . . . . ..

6.3 - Avaliação das Propriedades dos Estimadores Através do Uso de Simulação . . . .

REFERÊNCIAS 118

CAPÍTULO 1

INTRODUÇÃO

1.1 - PRELIMINARES

Em Análise de Regressão, algumas vezes as características intrínsicas do problema ou outras considerações indicam que o modelo apropriado para a situação é um modelo não linear.

Um modelo de Regressão é dito "Não Linear", se ele é não linear como função dos seus parâmetros. Nesse estudo estamos interessados no ajuste de modelos que não são possíveis de serem linearizados através do uso de transformações de variáveis. Esses modelos são denominados por Draper & Smith (81) de "intrinsicamente não lineares". Para ilustrar tais modelos consideremos os dois exemplos a seguir:

$$Y_t = \theta_1 e^{\theta_2 X_t} + \theta_3 e^{\theta_4 X_t} + e_t$$

e

$$Y_t = \frac{\theta_1 \theta_3 X_{1t}}{1 + \theta_1 X_{1t} + \theta_2 X_{2t}} + e_t$$

Segundo Draper & Smith (81), quando informações teóricas com relação ao modelo nos levam a um modelo intrinsicamente não linear, nós devemos geralmente preferir ajustar tal modelo, sempre que possível, do que ajustar um modelo linear, talvez menos realístico.

Nas aplicações muitas vezes não se tem um conhecimento direto da forma da relação envolvida entre as variáveis, mas somente determinadas informações ou suposições sobre o modelo

que podem ser representadas por um sistema de equações diferenciais ou equações de diferença, e o modelo então surge como solução desse sistema.

Este é o caso por exemplo do modelo logístico. Ele surge a partir da suposição de que a taxa média de crescimento da população no instante t é a diferença entre a taxa média de natalidade e a taxa média de mortalidade, e que a primeira não depende do tempo t, e nem do tamanho da população (Y(t)), e a segunda é diretamente proporcional ao tamanho da população (o que é uma suposição razoável se considerarmos que com o aumento da população há consequentemente um aumento da competição). Nesse caso a equação de crescimento é:

$$\frac{dy(t)}{dt}/y(t) = K_1 - K_2\, y(t)$$

Integrando essa equação, obtemos:

$$y(t) = \frac{\alpha}{1+\beta\, e^{-Kt}}, \text{ onde } \alpha = \frac{K_1}{K_2},\ \beta = \left(\frac{K_1 - K_2 y_0}{K_2 y_0}\right), \text{ e } y_0 = y(0)$$

(que é uma das formas da função logística).

Além do modelo logístico, existem vários outros modelos de crescimento comumente utilizados, tais como: o de Gompertz, Von Bertalanffy, Michaellis-Menten, Multi-Target, Mitscherlich, etc. Esses modelos são utilizados em muitas áreas diferentes tais como por exemplo: Biologia, Agricultura, Química, Economia, etc.

Além desses modelos existem muitos outros na literatura que são comumente utilizados nas aplicações.

Nos modelos de Regressão Não Lineares o critério de estimação mais comumente utilizado é o critério dos mínimos quadrados.

Nesse trabalho nós tratamos basicamente do problema da obtenção das estimativas de mínimos quadrados e da avaliação das propriedades estatísticas desses estimadores nas aplicações.

Se o modelo de Regressão é:

$$Y_t = f(\underset{\sim}{X}_t, \underset{\sim}{\theta}) + e_t, \quad t=1.,\ldots,n \qquad (1.1)$$

e

$$h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (y_t - f(\underset{\sim}{X}_t, \underset{\sim}{\theta}))^2$$

a soma de quadrados dos resíduos, o estimador de mínimos quadrados $\hat{\underset{\sim}{\theta}}$, por definição é o valor que minimiza a função $h(\underset{\sim}{\theta})$, isto é, $\hat{\underset{\sim}{\theta}}$ é o ponto de mínimo de $h(\underset{\sim}{\theta})$.

Quando o modelo (1.1) é linear, o estimador de mínimos quadrados $\hat{\underset{\sim}{\theta}}$ tem uma fórmula explícita, entretanto no caso não linear, isso geralmente não ocorre e para obter as estimativas desses parâmetros devemos recorrer a procedimentos iterativos (que requerem muito mais cálculos). Para ilustrar consideremos o seguinte modelo:

$$Y_t = \frac{\theta_1 X_t}{X_t + \theta_2} + e_t$$

Nesse caso

$$h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} \left(y_t - \frac{\theta_1 X_t}{X_t + \theta_2}\right)^2$$

e o sistema de "equações normais":

$$\frac{\partial h(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \underset{\sim}{0} \Leftrightarrow \begin{cases} \theta_1 \sum_{t=1}^{n} \dfrac{X_t^2}{(X_t+\theta_2)^2} - \sum_{t=1}^{n} \dfrac{X_t\, y_t}{(X_t+\theta_2)} = 0 \\ \theta_1 \dfrac{X_t^2}{(X_t+\theta_2)^3} - \sum_{t=1}^{n} \dfrac{X_t\, Y_t}{(X_t+\theta_2)^2} = 0 \end{cases}$$

Como podemos observar, esse sistema é não linear em $\underset{\sim}{\theta}$, e $\hat{\underset{\sim}{\theta}}$ não é obtido explicitamente. Isso quer dizer que nós transformamos o problema original de Otimização, num problema de resolver um sistema não linear que apresenta o mesmo grau de dificuldade. Portanto, como a solução desse sistema envolve um algoritmo iterativo, então parece igualmente factível minimizar $h(\underset{\sim}{\theta})$ diretamente, através de algum procedimento iterativo de Otimização.

1.2 - RESUMO HISTÓRICO:

Em Regressão Não Linear, o problema de estimação depende essencialmente da utilização de métodos de Otimização não linear.

Os algoritmos clássicos para minimização de funções não lineares para várias variáveis foram introduzidos por Newton, Cauchy e Gauss. Os algoritmos de Newton e Cauchy podem ser usados para minimizar funções gerais, enquanto que o de Gauss é aplicável somente para problemas de mínimos quadrados.

Nas últimas décadas com o"advento do computator" houve um grande progresso no desenvolvimento dessas técnicas. Consequentemente existe atualmente uma vasta literatura sobre os mais variados métodos para os mais variados tipos de problemas, incluindo o problema de ajuste de dados.

Na década de 60, foram introduzidos vários métodos de Otimização gerais, como os métodos de Quasi-Newton, os métodos de Direções Conjugadas e os métodos do tipo "Simplex". (Box, Davies & Swann (69) apresentam de uma maneira bastante clara os principais métodos que surgiram nesse período).

Na década de 70, as pesquisas se concentraram mais em torno dos métodos de Quasi-Newton para minimização geral.

Os problemas de mínimos quadrados constituem um ramo bastante importante da teoria de Otimização, e atualmente tem havido um grande interesse nesses algoritmos e em reformular problemas que aparentemente não são de mínimos quadrados, como por exemplo os problemas de máxima verossimilhança, de tal modo que eles possam ser revolvidos por algoritmos para mínimos quadrados.

Por outro lado, segundo Ratkowsky (83), apesar da grande variedade de métodos existentes para a obtenção das estimativas de mínimos quadrados, o estudo das propriedades desses estimadores "ainda está na sua infância", e somente nas duas últimas décadas é que alguns resultados mais importantes tem surgido, embora em número muito limitado.

1.3 - APRESENTAÇÃO DOS CAPÍTULOS

No capítulo 2, nós consideramos alguns dos métodos iterativos mais conhecidos em Otimização Não Linear sem restrições, e que utilizam derivadas. Nessa classe nós dividimos os procedimentos em: métodos para funções gerais e métodos especiais para mínimos quadrados.

No capítulo 3, nós apresentamos algumas técnicas para auxiliar na escolha de "valores iniciais", principalmente em modelos de Regressão Não Linear.

No capítulo 4, nós estudamos o problema do ajuste de modelos probabilísticos por máxima verossimilhança "em termos de mínimos quadrados". Esse procedimento consiste em transformar o problema de máxima verossimilhança num problema de mínimos quadrados, de tal modo que as estimativas dos parâmetros possam ser obtidas através de um programa padrão de Regressão Não Linear.

No capítulo 5 estudamos algumas das "medidas de não linearidade" mais recentes; as medidas de curvatura de Bates & Watts (80) e a medida do viés de Box (71).

No capítulo 6, nós avaliamos através de Simulação as propriedades assintóticas dos estimadores de mínimos quadrados nas aplicações.

CAPÍTULO 2

PROCEDIMENTOS DE OTIMIZAÇÃO NÃO LINEAR

2.1 - INTRODUÇÃO

Nos modelos não lineares os estimadores dos parâmetros não tem em geral fórmulas explícitas e as estimativas são obtidas através de técnicas numéricas denominadas técnicas de Otimização Não Linear ou Programação Não Linear.

Em Otimização, a função a ser minimizada (ou maximizada), é denominada "função objetivo", assim no caso de mínimos quadrados por exemplo, a função objetivo é $h(\underset{\sim}{\theta}) = \sum_{i=1}^{n} e_i^2(\theta)$, e no caso de máxima verossimilhança a função objetivo $h(\underset{\sim}{\theta}) = \prod_{t=1}^{n} f(\underset{\sim}{x}_t, \underset{\sim}{\theta})$. Em Otimização os problemas podem ser tratados sempre como problemas de minimização já que $\max_{\underset{\sim}{\theta}}\{h(\underset{\sim}{\theta})\} = \min_{\underset{\sim}{\theta}}\{-h(\underset{\sim}{\theta})\}$.

Para obter a solução dos problemas de Programação Não Linear existe uma grande variedade de algoritmos na literatura, e qual é o melhor vai depender da aplicação específica.

Em Programação Não Linear não existe um procedimento que seja superior a todos os outros para todos os problemas, o que existe são certas classes de algoritmos que são de maneira geral mais eficientes para determinados tipos de problemas (ou equivalentemente, existem certas classes de problemas para os quais determinados algoritmos são de modo geral mais eficientes). Os problemas costumam ser classificados, por exemplo, quanto a diferenciabilidade de $h(\underset{\sim}{\theta})$, (isto é, se

a função objetivo é ou não diferenciável), quanto à estrutura de $h(\underset{\sim}{\theta})$ ($h(\underset{\sim}{\theta})$ é uma soma de quadrados ou não), quanto a presença de vínculos ou restrições, quanto ao tipo de restrições (linear ou não linear), etc., e para cada tipo de problema existe uma teoria especial.

Uma classe bastante importante e que merecerá maior destaque nesse trabalho é a dos problemas de mínimos quadrados (sem restrições), envolvendo funções diferenciáveis, que é o caso mais simples em Regressão Não Linear. Segundo Fletcher(80), os algoritmos que utilizam essa estrutura são em geral bastante eficientes existindo vantagens em utilizar tais métodos em várias circunstâncias, ao invés de um método geral (isto é, um método para funções gerais). Entretanto apesar do fato desses algoritmos serem específicos para mínimos quadrados não quer dizer que eles sejam superiores a qualquer outro algoritmo para resolver todos os problemas; um exemplo típico onde isso ocorre é no caso dos "resíduos grandes" como veremos na seção 2.5.

Antes, porém de apresentarmos alguns dos métodos específicos para mínimos quadrados (seções 2.4 e 2.5); é importante conhecermos alguns conceitos básicos em Programação Não Linear tais como, o de um método descendente, a taxa de convergência de um algoritmo, o critério de parada das iterações, etc., que serão apresentados na seção 2.2. Além disso, na seção 2.3 nós apresentamos alguns dos principais métodos gerais de Otimização Não Linear, que utilizam derivadas. Esses méto-

dos constituem um ramo bastante importante na teoria de Otimização Não Linear, (já que segundo Gill, Murray and Wright (81), os métodos que utilizam derivadas são em geral mais eficientes do que aqueles que não utilizam derivadas).

Em Otimização Não Linear, uma questão importante também é a da localização de pontos ótimos globais ou absolutos; porém nenhum dos algoritmos existentes pode garantir convergência para um ótimo global. Entretanto uma maneira de tentar prevenir possíveis soluções locais é através da escolha de bons valores iniciais já que o ponto para o qual um procedimento converge depende também da escolha do valor inicial. Para ilustrar esse fato consideremos o exemplo da figura abaixo:

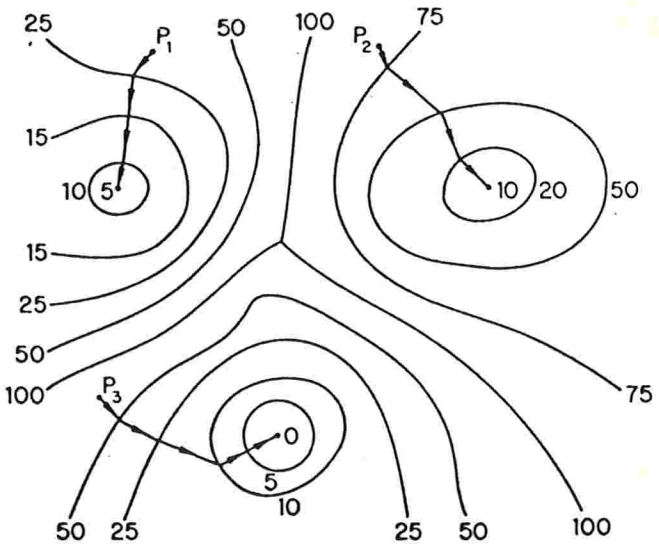

FIG. 2.1.1 - DIFERENTES PONTOS INICIAIS DOS PARÂMETROS PODEM RESULTAR EM DIFERENTES PONTOS MÍNIMOS

Nessa figura temos as curvas de nível de uma função $h(\underset{\sim}{\theta})$ com dois mínimos locais $\underset{\sim}{\theta}_1$ e $\underset{\sim}{\theta}_2$ e um mínimo global θ_3 (onde $h(\underset{\sim}{\theta}_1) = 5$, $h(\underset{\sim}{\theta}_2) = 10$, $h(\underset{\sim}{\theta}_3) = 0$). Tomando P_1 como ponto de partida, o processo convergiu para θ_1 , do mesmo modo partindo-se do ponto P_2, o processo convergiu para um mínimo local θ_2. Entretanto partindo-se de P_3 ocorreu convergência para o mínimo global da função.

Assim, na prática não devemos nos contentar, pelo menos de imediato, com o valor obtido através do processo, mas sim realizar vários testes partindo-se de vários valores iniciais diferentes, e comparar os resultados; esse procedimento ajuda a prevenir possíveis soluções locais do problema.

Como estamos vendo, a solução de um problema de Otimização Não Linear na prática, é tarefa que exige esforço e uma certa habilidade; por esse motivo damos a seguir algumas sugestões de como proceder nas aplicações.

O primeiro passo obviamente, é escolher, segundo algum critério, o algoritmo a ser utilizado, e também uma estimativa inicial, para o processo, que pode ser obtido por exemplo através de experiências anteriores, ou então de algum procedimento como por exemplo a utilização de uma rede de pontos no espaço considerado (No capítulo 3 descrevemos algumas técnicas mais usuais para a obtenção de valores iniciais em modelos de Regressão).

Se porventura o algoritmo escolhido não convergir satisfatoriamente podemos tentar reprocessá-lo usando melhores valores iniciais, ou então utilizar uma modificação desse algo-

ritmo; outra possibilidade é abandonar esse método e utilizar outro diferente; nesse caso é recomendável ter à disposição uma biblioteca de algoritmos que possam resolver o problema. Infelizmente nem sempre existem programas disponíveis para todos os métodos (embora em alguns casos bons programas possam ser obtidos diretamente do autor do método).

Segundo Chambers (73), antigamente muitos métodos de Otimização eram suficientemente simples para que um usuário pudesse escrever seu próprio programa, porém hoje em dia esse procedimento não é mais recomendado porque mesmo métodos confiáveis podem ter um mal desempenho devido a erros de arredondamento, etc., se não forem tomados os devidos cuidados.

Atualmente uma boa opção é utilizar "pacotes", embora também tenham suas limitações. Os pacotes mais conhecidos para Regressão Não Linear são o BMDP, o SAS e o IMSL.

2.2 - ASPECTOS GERAIS DE UM ALGORÍTMO

Um algoritmo iterativo é um procedimento que partindo de um ponto inicial especificado $\underset{\sim}{\theta}^{(0)} = (\theta_1^{(0)},\ldots,\theta_p^{(0)})$, gera uma sequência de pontos $\underset{\sim}{\theta}^{(1)}$, $\underset{\sim}{\theta}^{(2)}$,..., onde cada elemento representa uma estimativa do ponto ótimo $\underset{\sim}{\theta}^*$ da função objetivo, e o que se espera de um bom algoritmo, é que ao ser aplicado ao problema de interesse convirja rapidamente para $\underset{\sim}{\theta}^*$.

Uma grande parte das técnicas de Otimização incluem uma sequência de "pesquisas unidimensionais" no espaço p-dimensio

nal (H). A idéia é a seguinte: partir de um ponto inicial e determinar através de uma determinada regra, uma direção de pesquisa e em seguida mover-se nessa direção a um mínimo da função objetivo. A estrutura da k-ésima iteração é a seguinte:

(i) Determinar a direção de pesquisa $d^{(k)}$;

(ii) Achar $\alpha^{(k)}$ que minimiza $h(\underset{\sim}{\theta}^{(k)}+\alpha\, d^{(k)})$ em relação a α;

(iii) Estabelecer $\underset{\sim}{\theta}^{(k+1)}=\underset{\sim}{\theta}^{(k)}+\alpha^{(k)}d^{(k)}$.

No passo (ii), α é a distância movida ao longo da direção de pesquisa $d^{(k)}$, e é denominada "tamanho do passo".

Esses métodos são chamados "algoritmos de pesquisa unidimensional" e diferentes métodos correspondem a diferentes maneiras de obter $d^{(k)}$.

Uma vez que a direção $d^{(k)}$ tenha sido escolhida o problema agora envolve a solução de um problema univariado. No caso de h ser diferenciável, a equação $\frac{d\,h(\underset{\sim}{\theta}^{(k)}+\alpha d^{(k)})}{d\,\alpha}=0$ deve ser resolvida com a finalidade de obter o ponto de mínimo (isto é, obter um α tal que a derivada direcional de h no ponto $\underset{\sim}{\theta}^{(k)}$ e na direção $d^{(k)}$ seja nula). Entretanto segundo Fletcher (80), esse procedimento tem apenas um valor conceitual já que não é muito eficiente determinar esse mínimo com alta precisão, e o que se faz na prática é obter um "mínimo" aproximado nessa direção, por exemplo através de uma interpolação polinomial (além do uso de uma interpolação polinomial, existem várias outras alternativas que podem ser usadas com essa fina

lidade; Gill, Murray and Wright (81), apresenta uma série de las).

Associado com estas idéias está o conceito de "método descendente". Um método descendente satisfaz, em cada iteração, a seguinte condição:

$$h(\underset{\sim}{\theta}^{(k)} + \alpha^{(k)} d^{(k)}) < h(\underset{\sim}{\theta}^{(k)}) \quad ,$$

para algum $\alpha^{(k)} > 0$. (Essa é uma das condições necessárias para se demonstrar matematicamente a convergência de um algoritmo).

Outro aspecto importante de um algoritmo é a sua taxa de convergência. Segundo Gill, Murray and Wright (81), mesmo que seja possível provar teoricamente que uma sequência converge para um ponto ótimo, esse método será eficiente somente se a convergência ocorrer com certa rapidez, isto é, se a taxa de convergência for relativamente alta.

Essas medidas são teóricas e para vários métodos elas não são possíveis de serem obtidas (esse fato, segundo Gill, Murray and Wright (81), não implica necessariamente que o método não seja eficiente, mas pode simplesmente indicar uma dificuldade algébrica na demonstração).

Segundo Fletcher (80) estudos empíricos baseados em testes computacionais bem selecionados frequentemente dão uma melhor indicação sobre o desempenho de um método na prática,

do que resultados teóricos sobre convergência e taxas de convergência (entretanto, obviamente é sempre recomendável, se possível, aliar a teoria à prática).

Outro aspecto também importante de qualquer algoritmo é o teste de convergência, ou, "critério de parada" das iterações. Um teste natural seria interromper o processo quando $|h(\underset{\sim}{\theta}^{(k)}) - h(\underset{\sim}{\theta}^{*})| < \varepsilon$ ou $|\theta_i^{(k)} - \theta_i^{*}| < \varepsilon_i$, onde $\varepsilon, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ são constantes fixadas arbitrariamente, entretanto esses testes não são aplicáveis porque requerem o conhecimento da solução $\underset{\sim}{\theta}^{*}$.

Um teste bastante utilizado nas aplicações é admitir convergência na k-ésima iteração se:

$$h(\underset{\sim}{\theta}^{(k-1)}) - h(\underset{\sim}{\theta}^{(k)}) \leqslant \varepsilon \qquad (2.2.1)$$

ou

$$|\theta_i^{(k-1)} - \theta_i^{(k)}| < \varepsilon_i \,, \quad i=1,\ldots,p \qquad (2.2.2)$$

Uma alternativa para (2.2.2) é $\|\underset{\sim}{\theta}^{(k-1)} - \underset{\sim}{\theta}^{(k)}\| < \varepsilon'$ (segundo Flethcer (80) esses critérios (2.2.1 e 2.2.2) geralmente funcionam bem somente em algoritmos eficientes, isto é, para os melhores disponíveis).

Outro critério de parada, bastante utilizado nas sub-rotinas existentes, interrompe o processo quando um certo número fixado de iterações é alcançado.

Uma outra alternativa para testar a convergência de uma sequência é interromper o processo na k-ésima iteração se $\|\nabla h(\underset{\sim}{\theta}^{(k)})\| < \varepsilon$.

Além dessas, existem várias outras alternativas práticas com essa finalidade, todas elas tendo vantagens e limitações.

2.3 - MÉTODOS DO GRADIENTE PARA MINIMIZAÇÃO SEM RESTRIÇÕES

Nessa seção nosso objetivo é apresentar algum dos métodos básicos gerais de Otimização sem Restrições e que utilizam derivadas: o método do "Steepest Descent", o método de Newton e os métodos de "Quasi-Newton", que são também conhecidos como "métodos do gradiente".

2.3.1 O MÉTODO DO "STEEPEST DESCENT"

Um dos métodos mais antigos e conhecidos de Otimização de funções de várias variáveis é o método do "Steepest Descent" ou "método do gradiente". Ele foi proposto inicialmente por Cauchy em 1847, e desde então tem sofrido várias modificações. Embora não seja muito eficiente na prática, ele é muito importante do ponto de vista teórico, pois serve como base de comparação para outras técnicas.

O algoritmo do "Steepest Descent" é o seguinte:

$$\underset{\sim}{\theta}^{(k+1)} = \underset{\sim}{\theta}^{(k)} + \alpha^{(k)} \left(-g\left(\underset{\sim}{\theta}^{(k)}\right)\right)$$

onde $g(\theta^{(k)})$ é o vetor gradiente, isto é, $g(\underset{\sim}{\theta}^{(k)}) = (\frac{\partial h}{\partial \theta_1}(\theta^{(k)}), \ldots, \frac{\partial h}{\partial \theta_p}(\theta^{(k)}))$ e $h(\underset{\sim}{\theta})$ é a função objetivo.

A direção $d^{(k)} = -g(\underset{\sim}{\theta}^{(k)})$ do "Steeptest Descent" é uma direção descendente (já que a derivada direcional

$$\frac{d}{d\alpha} h(\underset{\sim}{\theta}^{(k)} - \alpha g(\underset{\sim}{\theta}^{(k)})) = -g(\underset{\sim}{\theta}^{(k)})' . g(\underset{\sim}{\theta}^{(k)}) = - \| g(\underset{\sim}{\theta}^{(k)}) \| < 0).$$

Além disso essa direção é a de maior decréscimo em $h(\underset{\sim}{\theta}^{(k)})$ já que o vetor gradiente $g(\underset{\sim}{\theta}^{(k)})$ aponta na direção de maior crescimento da função em $\underset{\sim}{\theta}^{(k)}$ (Kaplan, W. (71)).

Pelo fato desse método ser descendente, a função objetivo $h(\underset{\sim}{\theta})$ pode ser sempre reduzida a cada iteração escolhendo-se $\alpha^{(k)}$ convenientemente, através de uma pesquisa unidimensional.

Existem várias alternativas que podem ser combinadas com esse algoritmo para obter $\alpha^{(k)}$, e qualquer que seja ela, teremos um método com convergência garantida (isto é, uma convergência demonstrada teoricamente).

Na prática entretanto o método do "Steepest Descent" geralmente exibe um comportamento oscilatório (Box, Davies, Swann, 69), requerendo centenas de iterações para muito pouco progresso em direção à solução. Segundo Dahlquist & Björck (74), o algoritmo de "Steepest Descent" funciona bem somente "longe da solução", isto é, converge rapidamente somente no início do processo.

Fletcher (80), comenta que esse método geralmente termina bem longe da solução em virtude também de erros de arredondamento e que na prática ele não é nem confiável, nem eficien

te, em geral.

Esses fatos concordam de certa forma com os resultados sobre a velocidade de convergência do método. Fritsche (75) demonstra que se a função objetivo é quadrática, isto é,

$$h(\underset{\sim}{\theta}) = \frac{1}{2}\underset{\sim}{\theta}'G\underset{\sim}{\theta} + c'\underset{\sim}{\theta}$$

(onde G é a matriz Hessiana de h, e positiva definida) então o método do "Steepest Descent" que utiliza uma pesquisa unidimensional exata converge linearmente com taxa $(\frac{k-1}{k+1})^2$, onde k é o Número de Condição de G. . Basicamente o que esse resultado informa é que a convergência vai se tornando mais lenta quanto maior for o número de condição, e mais rápida quanto mais próximo o número de condição estiver de 1. Podemos entender esse resultado geometricamente se considerarmos que se o número de condição é exatamente igual a 1, os contornos ou "curvas de nível" de $h(\underset{\sim}{\theta})$ são circulares ou esféricos, e nesse caso a direção do "Steeptest Descent" aponta para o centro, isto é, para o mínimo de $h(\underset{\sim}{\theta})$ (já que a direção do "Steepest Descent" é ortogonal a esses contornos (Kaplan, W. (71)) e portanto o método converge para a solução num só passo. Por outro lado se o número de condição vai aumentando, os contornos vão se tornando elípses mais alongadas e a direção do "Steepest Descent" não aponta mais em geral para o centro, necessitando de muitas iterações para alcançar a solução.

Estes resultados, embora deduzidos para funções quadráticas, valem aproximadamente para funções "suaves", já que numa "pequena vizinhança" do ponto de mínimo a função $h(\underset{\sim}{\theta})$ pode ser muito bem aproximada por uma função quadrática e convexa.

Segundo Box, Davies & Swann (69), as dificuldades com o método do "Steepest Descent" estão essencialmente relacionados com o problema da escala. O método do "Steepest Descent " não é invariante por escala, isso quer dizer que a direção de pesquisa varia com uma mudança de escala e consequentemente a taxa de convergência também muda.

Para ilustrar esse fato consideremos a função

$$h(\underset{\sim}{\theta}) = \frac{1}{2}\theta_1^2 + 2\theta_2^2$$

A direção de pesquisa nesse caso é $-(\theta_1, 4\theta_2)$, entretanto se considerarmos a transformação $\delta_1 = \theta_2$ e $\delta_2 = 2\theta_2$ então

$$h(\delta) = \frac{1}{2}\delta_1^2 + \frac{1}{2}\delta_2^2$$

e a direção de pesquisa será $-(\delta_1, \delta_2)$, que ao contrário da situação acima, aponta para o ponto de mínimo de $h(\underset{\sim}{\theta})$ qualquer que seja δ_1 e δ_2. (O problema do método do "Steepest Descent" é que nem sempre é possível obter transformações adequadas).

2.3.2 MÉTODO DE NEWTON

O método de Newton se baseia na suposição de que na vizinhança do ponto $\underset{\sim}{\theta}^{(k)}$, a função objetivo $h(\underset{\sim}{\theta})$ pode ser aproximada por sua expansão em série de Taylor até 2ª ordem, isto é,

$$h(\underset{\sim}{\theta}) \cong h(\underset{\sim}{\theta}^{(k)}) + g(\underset{\sim}{\theta}^{(k)})' d^{(k)} + \frac{1}{2} d^{(k)} G(\underset{\sim}{\theta}^{(k)}) d^{(k)} \quad (2.3.2.1)$$

onde:

$g(\underset{\sim}{\theta}^{(k)})$ é o vetor gradiente de h em $\underset{\sim}{\theta}^{(k)}$.

$G(\underset{\sim}{\theta}^{(k)})$ é a matriz Hessiana de h em $\underset{\sim}{\theta}^{(k)}$.

$d^{(k)} = \underset{\sim}{\theta} - \underset{\sim}{\theta}^{(k)} = \Delta\underset{\sim}{\theta}^{(k)}$.

A direção de pesquisa $d^{(k)}$ que minimiza $\hat{h}(\underset{\sim}{\theta})$, onde $\hat{h}(\underset{\sim}{\theta})$ é o 2º membro da expressão (2.3.2.1), é a direção do método de Newton, e deve satisfazer o sistema

$$\frac{\partial \hat{h}(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \underset{\sim}{0} \iff G(\underset{\sim}{\theta}^{(k)}) d^{(k)} = -g(\underset{\sim}{\theta}^{(k)})$$

Se a matriz $G(\underset{\sim}{\theta}^{(k)}$ é positiva definida então

$$d^{(k)} = -(G(\underset{\sim}{\theta}^{(k)})^{-1} g(\underset{\sim}{\theta}^{(k)}) \Rightarrow \underset{\sim}{\theta}^{(k+1)} =$$

$$= \underset{\sim}{\theta}^{(k)} - (G(\underset{\sim}{\theta}^{(k)})^{-1} g(\underset{\sim}{\theta}^{(k)}) \quad (2.3.2.2)$$

A expressão (2.3.2.2) é a k+1-ésima iteração do método de Newton na sua forma mais simples. Entretanto frequentemente incorpora-se uma pesquisa unidimensional, obtendo em cada iteração uma constante positiva $\alpha^{(k)}$ tal que

$$h(\underset{\sim}{\theta}^{(k)} + \alpha^{(k)} d^{(k)}) < h(\underset{\sim}{\theta}^{(k)})$$

No caso particular em que $h(\underset{\sim}{\theta})$ é quadrática e convexa, o método de Newton converge para $\underset{\sim}{\theta}^*$ numa única iteração partindo-se de qualquer ponto inicial (já que $G(\underset{\sim}{\theta}^{k})$ não depende de $\underset{\sim}{\theta}$), porém para uma função geral isso não ocorre normalmente. Entretanto se $\underset{\sim}{\theta}^{(k)}$ estiver suficientemente próximo de $\underset{\sim}{\theta}^*$, sabe-se que $h(\underset{\sim}{\theta})$ pode muito bem ser aproximada por uma função quadrática onde $G(\underset{\sim}{\theta}^{(k)})$ é definida positiva, e portanto o método funciona muito bem, (aliás com taxa de convergência quadrática quando $\{\alpha^{(k)}\} \to 1$, segundo Fletcher (80)).

Se por outro lado $\underset{\sim}{\theta}^{(k)}$ não estiver nas vizinhanças de $\underset{\sim}{\theta}^*$, podem ocorrer dificuldades, porque nesse caso, o modelo quadrático não é em geral uma boa aproximação para $h(\underset{\sim}{\theta})$ e $G(\underset{\sim}{\theta}^{(k)})$ pode não ser definida positiva. Assim na prática o algoritmo de Newton deve ser modificado para suportar a possibilidade da matriz Hessiana G não ser definida positiva.

Uma possível modificação é considerar a direção $d^{(k)}$ onde

$$d^{(k)} = \begin{cases} -(G^{(k)})^{-1} g^{(k)}, & \text{se } G^{(k)} \text{ é definida positiva} \\ -g^{(k)}, & \text{caso contrário.} \end{cases}$$

e $G^{(k)} = G(\underset{\sim}{\theta}^{(k)})$ e $g^{(k)} = g(\underset{\sim}{\theta}^{(k)})$.

Outra possibilidade é construir uma matriz $\bar{G}^{(k)}$ definida positiva, com base em $G^{(k)}$, isto é, considerar a direção

$$d^{(k)} = \begin{cases} -(G^{(k)})^{-1} g^{(k)}, & \text{se } G^{(k)} \text{ é definida positiva} \\ -(\bar{G}^{(k)})^{-1} g^{(k)}, & \text{caso contrário} \end{cases}$$

Existem vários métodos que utilizam essa estrutura. Um dos métodos mais conhecidos é o que considera $\overline{G}^{(k)}=(G_k^{(k)}+\lambda I)$, onde I é a matriz Identidade e λ é uma constante escolhida de tal modo que $\overline{G}^{(k)}$ seja positiva definida (Modificações desse tipo foram introduzidas por Levenberg (44) e Marquardt (63)).

Além das modificações apresentadas (Fletcher (80), Gill, Murray and Wrigth (81)), sugerem várias outras alternativas.

Segundo Gill, Murray and Wrigth (81) as propriedades de convergência do método de Newton, fazem dele um algoritmo extremamente atraente, sendo frequentemente tomado como padrão de comparação. Entretanto o método de Newton (tanto na sua forma mais simples como qualquer uma de suas modificações), requer o conhecimento das derivadas segundas de $h(\underset{\sim}{\theta})$, o que segundo Jenrich (79), constitui a sua maior inconveniência ou dificuldade. Assim apesar do método de Newton ser altamente eficiente quando se incorpora uma modificação adequada, em muitos casos ele é impraticável por esse motivo.

2.3.3 MÉTODOS DE QUASI-NEWTON

Na seção anterior vimos que a principal desvantagem do método de Newton (ou Newton Modificado) é a necessidade de obtenção de fórmulas para o cálculo das derivadas segundas para a matriz Hessiana.

Existem entretanto uma classe de métodos denominada: métodos de "Quasi-Newton", ou "variable metric", onde esse pro-

blema não ocorre. Esses métodos são atualmente considerados os mais sofisticados para a solução de problemas gerais sem restrições.

Os métodos de "Quasi-Newton" são semelhantes ao método de Newton exceto pelo fato de que $(G^{(k)})^{-1}$, (a inversa da matriz Hessiana de $h(\underset{\sim}{\theta})$ em $\underset{\sim}{\theta}^{(k)}$), é aproximada por uma matriz simétrica definida positiva $H^{(k)}$, que é corrigida a cada iteração. Essas matrizes $H^{(k)}$ são construidas com base somente nos valores de $h(\underset{\sim}{\theta})$ e suas derivadas primeiras durante o processo, e de tal modo que a sequência $\{H^{(k)}\}$ converge para $\overline{G}^{*-1} = G^{-1}(\underset{\sim}{\theta}^*)$.

Os métodos de "Quasi-Newton" quando comparados com os de Newton têm portanto a vantagem de não utilizar as derivadas segundas; além disso têm a vantagem adicional de que $H^{(k)}$ é sempre positiva definida o que implica que o método é descendente. Além desses fatos o número de multiplicações envolvidas num método de "Quasi-Newton" é bem menor do que no método de Newton (Fletcher (80)).

Os algoritmos de "Quasi-Newton" tem a seguinte estrutura básica:

$$\underset{\sim}{\theta}^{(k+1)} = \underset{\sim}{\theta}^{(k)} - \alpha^{(k)} H^{(k)} g^{(k)}$$

onde:

$g^{(k)}$ é o vetor gradiente em $\underset{\sim}{\theta}^{(k)}$.

$H^{(k)}$ é uma matriz definida positiva que aproxima $G^{k^{-1}}$

$\alpha^{(k)}$ é o tamanho do passo obtido através de uma pesquisa unidimensional.

Nos métodos de "Quasi-Newton", o que se faz geralmente é utilizar fórmulas que permitam calcular $H^{(k+1)}$ a partir de $H^{(k)}$, e uma das maneiras de fazer isso é impor a "condição de Quasi-Newton" descrita em Davidon (59), isto é, fazer com que $H^{(k+1)}$ satisfaça a equação

$$H^{(k+1)}(g^{(k+1)}-g^{(k)}) = \underset{\sim}{\theta}^{(k+1)}-\underset{\sim}{\theta}^{(k)}$$

Se $H^{(k+1)}=H^{(k)}+E^{(k)}$, então existem várias maneiras possíveis de escolher $E^{(k)}$ de tal modo que esta condição seja satisfeita, e para cada escolha teremos um método de "Quasi-Newton" diferente.

A matriz inicial $H^{(0)}$, pode ser qualquer matriz definida positiva. A escolha $H^{(0)} = I$, é normalmente feita, na ausência de uma melhor estimativa, e nesse caso o processo começa como "Steepest Descent" e termina como Newton.

Nessa seção apresentamos alguns dos métodos mais conhecidos de "Quasi-Newton": os métodos de DFP, BFGS e "Broyden Family" (Além desses entretanto, existem vários outros métodos importantes como também centenas de modificações desses algoritmos).

2.3.3.1 O MÉTODO DFP

O método de Davidon-Fletcher-Powell, também conhecido como DFP, foi formulado inicialmente por Davidon (59), e mais tarde aperfeiçoado por Fletcher and Powell (63). Embora o DFP foi o primeiro método de "Quasi-Newton" a surgir, ele tem si

do ainda muito utilizado nas aplicações (Fletcher 80).

O algoritmo de DFP é o seguinte:

$$\underset{\sim}{\theta}^{(k+1)} = \underset{\sim}{\theta}^{(k)} - \alpha^{(k)} H^{(k)} g^{(k)}$$

onde

$$H_{DFP}^{(k)} = H^{(k-1)} + E^{(k-1)}$$

e

$$E^{(k)} = \frac{\Delta\underset{\sim}{\theta}^{(k)} \Delta\underset{\sim}{\theta}^{(k)'}}{\Delta\underset{\sim}{\theta}^{(k)'} \Delta g^{(k)}} - \frac{H^{(k)} \Delta g^{(k)} \Delta g^{(k)'} H^{(k)'}}{\Delta g^{(k)'} H^{(k)} \Delta g^{(k)}} \quad , k=0,1,2,..$$

onde

$$\Delta\underset{\sim}{\theta}^{(k)} = \underset{\sim}{\theta}^{(k+1)} - \underset{\sim}{\theta}^{(k)}$$

$$\Delta g^{(k)} = g(\underset{\sim}{\theta}^{(k+1)}) - g(\underset{\sim}{\theta}^{(k)})$$

e $\alpha^{(k)}$ é obtido através de um procedimento de pesquisa unidimensional.

Esse método apresenta uma série de propriedades teóricas importantes. Assim, por exemplo se a função $h(\underset{\sim}{\theta})$ é quadrática e se for utilizada uma pesquisa unidimensional exata, demonstra-se que o processo termina em no máximo p iterações (onde p é a dimensão de $\mathcal{H}$); além disso para funções gerais, o método é descendente, tem uma taxa de convergência superlinear, requer apenas um número de multiplicações da ordem de $3n^2+n$, e convergência é demonstrada para o caso onde h é estritamente convexa (A demonstração desses resultados pode ser encontrada em Fletcher (80)).

2.3.3.2 O MÉTODO BFGS

Segundo Fletcher (80), o método BFGS tem funcionado muito bem na prática, "talvez até melhor do que o DFP", e atualmente a opinião geral é de que ele seja o melhor método de "Quasi-Newton" disponível.

Nesse método, a matriz $H_{BFGS}^{(k)} = H^{(k-1)} + E^{(k-1)}$, onde:

$$E^{(k)} = \left(1+\frac{\Delta g^{(k)'} H^{(k)} \Delta g^{(k)}}{\Delta\underset{\sim}{\theta}^{(k)'} \Delta g^{(k)}}\right) \frac{\Delta\underset{\sim}{\theta}^{(k)} \Delta\underset{\sim}{\theta}^{(k)'}}{\Delta\underset{\sim}{\theta}^{(k)'} \Delta g^{(k)}} -$$

$$- \left(\frac{\Delta\underset{\sim}{\theta}^{(k)} \Delta g^{(k)'} H^{(k)'} + H^{(k)} \Delta g^{(k)} \Delta\underset{\sim}{\theta}^{(k)'}}{\Delta\underset{\sim}{\theta}^{(k)'} \Delta g^{(k)}}\right)$$

onde

$$\Delta\underset{\sim}{\theta}^{(k)} = \underset{\sim}{\theta}^{(k+1)} - \underset{\sim}{\theta}^{(k)} \quad \text{e} \quad \Delta g^{(k)} = g^{(k+1)} - g^{(k)}$$

Esse método, ao contrário do DFP, não necessita de pesquisas unidimensionais tão acuradas, para que funcione bem. Além disso, do ponto de vista teórico ele tem as mesmas propriedades do DFP (com a vantagem adicional de não requerer uma pesquisa unidimensional exata para demonstrar convergência).

2.3.3.3 O MÉTODO DE "BROYDEN FAMILY"

Uma família uniparamétrica (com parâmetro ϕ) pode ser gerada tomando-se:

$$H_{\phi}^{(k)} = (1-\phi)\, H_{DFP}^{(k)} + \phi\, H_{BFGS}^{(k)}$$

Essa "família" é denominada "Broyden Family" e inclui tanto o DFP (quando $\phi=0$), como o BFGS (quando $\phi=1$), e muitas das propriedades desses métodos são comuns a toda a família.

Uma abordagem bastante completa sobre as propriedades desse método, e a obtenção do valor de ϕ na prática é apresentada por Fletcher (80).

2.4 - MÉTODOS PARA MÍNIMOS QUADRADOS NÃO LINEARES

2.4.1 INTRODUÇÃO

Num problema de mínimos quadrados o objetivo é determinar o ponto de mínimo de $h(\underset{\sim}{\theta})$, onde $h(\underset{\sim}{\theta})$ é uma soma de quadrados de funções não lineares $e_i(\underset{\sim}{\theta})$, isto é,

$$h(\underset{\sim}{\theta}) = \sum_{i=1}^{n} e_i^2(\underset{\sim}{\theta}) \qquad (2.4.1)$$

(Os problemas de mínimos quadrados ponderados podem ser tratados como de mínimos quadrados ordinários simplesmente observando que

$$\sum_{i=1}^{n} w_i \, e_i^2(\underset{\sim}{\theta}) = \sum_{i=1}^{n} (\sqrt{w_i}\, e_i(\underset{\sim}{\theta}))^2$$

Nessa seção nós vamos considerar somente o caso onde os $e_i(\theta)$'s são diferenciáveis (e as derivadas primeiras são facilmente obtidas).

Para minimizar (2.4.1), embora qualquer um dos métodos gerais de Otimização descritos na seção 2.3 podem ser utilizados, os métodos que utilizam essa estrutura especial apresentam certas vantagens sobre esses métodos gerais, segundo Gill, Murray and Wright (81).

Bard (70), através de estudos numéricos compara o desempenho de vários métodos do gradiente com alguns métodos especiais para mínimos quadrados (incluindo modificações do método de Gauss Newton, como por exemplo a de Marquardt) e conclui que os algoritmos que fazem uso dessa estrutura especial são de modo geral muito mais rápidos.

Num problema de mínimos quadrados o vetor gradiente tem a seguinte estrutura:

$$g(\underset{\sim}{\theta}) = \frac{\partial h(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \left(\frac{\partial h(\underset{\sim}{\theta})}{\partial \theta_1}, \ldots, \frac{\partial h(\underset{\sim}{\theta})}{\partial \theta_p}\right) =$$

$$= \left(2 \sum_{t=1}^{n} e_t \frac{\partial e_t(\underset{\sim}{\theta})}{\partial \theta_1}, \ldots, 2 \sum_{t=1}^{n} e_t \frac{\partial e_t(\underset{\sim}{\theta})}{\partial \theta_p}\right) =$$

$$= 2J(\underset{\sim}{\theta})' E(\underset{\sim}{\theta})$$

onde:

$$E(\underset{\sim}{\theta}) = \begin{bmatrix} e_1(\underset{\sim}{\theta}) \\ \vdots \\ e_n(\underset{\sim}{\theta}) \end{bmatrix} \quad \text{e} \quad J(\underset{\sim}{\theta})' = \frac{\partial E(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \begin{bmatrix} \frac{\partial e_1(\underset{\sim}{\theta})}{\partial \theta_1} & \cdots & \frac{\partial e_1(\underset{\sim}{\theta})}{\partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial e_n(\underset{\sim}{\theta})}{\partial \theta_1} & \cdots & \frac{\partial e_n(\underset{\sim}{\theta})}{\partial \theta_p} \end{bmatrix}$$

Além disso a matriz Hessiana:

$$G(\underset{\sim}{\theta}) = \frac{\partial^2 h(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} = 2\left[J(\underset{\sim}{\theta})' J(\underset{\sim}{\theta}) + \sum_{t=1}^{n} e_t(\underset{\sim}{\theta})\, G_t(\underset{\sim}{\theta})\right]$$

onde $G_t(\underset{\sim}{\theta})$ é a matriz Hessiana de $e_t(\underset{\sim}{\theta})$, isto é:

$$G_t(\underset{\sim}{\theta}) = \begin{bmatrix} \dfrac{\partial^2 e_t(\underset{\sim}{\theta})}{\partial\theta_1^2} & \cdots & \dfrac{\partial^2 e_t(\underset{\sim}{\theta})}{\partial\theta_1\partial\theta_p} \\ \vdots & & \vdots \\ \dfrac{\partial^2 e_t(\underset{\sim}{\theta})}{\partial\theta_p\partial\theta_1} & \cdots & \dfrac{\partial^2 e_t(\underset{\sim}{\theta})}{\partial\theta_p^2} \end{bmatrix}$$

No caso de ajuste de funções de regressão por mínimos quadrados (que é o nosso interesse), $e_t(\underset{\sim}{\theta}) = y_t - f(\underset{\sim}{x}_t, \underset{\sim}{\theta})$, $t=1,2,\ldots,n$, e o vetor gradiente $g(\underset{\sim}{\theta})$ é:

$$g(\underset{\sim}{\theta}) = \frac{\partial h(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = -2F(\underset{\sim}{\theta})'E(\underset{\sim}{\theta})$$

onde $F(\underset{\sim}{\theta})$ é a matriz Jacobiana de $f(\underset{\sim}{x}, \underset{\sim}{\theta})$, isto é:

$$F(\underset{\sim}{\theta}) = \begin{bmatrix} \dfrac{\partial f(\underset{\sim}{x}_1, \underset{\sim}{\theta})}{\partial\theta_1} & \cdots & \dfrac{\partial f(\underset{\sim}{x}_1, \underset{\sim}{\theta})}{\partial\theta_p} \\ \vdots & & \vdots \\ \dfrac{\partial f(\underset{\sim}{x}_n, \underset{\sim}{\theta})}{\partial\theta_1} & \cdots & \dfrac{\partial f(x_n, \underset{\sim}{\theta})}{\partial\theta_p} \end{bmatrix}$$

e a matriz Hessiana é

$$G(\underset{\sim}{\theta}) = \frac{\partial^2 h(\underset{\sim}{\theta})}{\partial\underset{\sim}{\theta}\partial\underset{\sim}{\theta}'} = 2\left[F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta}) + B(\underset{\sim}{\theta})\right] \qquad (2.4.2)$$

onde:

$$B(\underset{\sim}{\theta}) = -\sum_{t=1}^{n} e_t(\underset{\sim}{\theta})\, G_t(\underset{\sim}{\theta})$$

e $G_t(\underset{\sim}{\theta})$ é a matriz Hessiana de $f_t(\underset{\sim}{\theta}) = f(\underset{\sim}{x}_t, \underset{\sim}{\theta})$.

Podemos notar através dessas fórmulas que se os $e_t(\underset{\sim}{\theta})$'s forem nulos, ou se os $e_t(\underset{\sim}{\theta})$'s (ou $f_t(\underset{\sim}{\theta})$'s) forem lineares então $B(\underset{\sim}{\theta})$ é uma matriz nula e portanto a matriz Hessiana $G(\underset{\sim}{\theta})$ é $G(\underset{\sim}{\theta}) = 2F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta})$.

Os métodos para mínimos quadrados se baseiam tipicamente na suposição de que $B(\underset{\sim}{\theta}) \cong \underset{\sim}{0}$, isto é, de que a matriz Hessiana $G(\underset{\sim}{\theta})$ pode ser bem aproximada por $2F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta})$. Um dos métodos que utilizam essa estrutura é o método de Gauss Newton que apresentamos a seguir.

2.4.2 O MÉTODO DE GAUSS NEWTON

O algoritmo de Gauss Newton é um dos métodos mais conhecidos para resolver problemas de mínimos quadrados.

A idéia básica do método é considerar aproximações lineares para os $e_t(\theta)$'s em cada iteração (num problema de ajuste de dados isso é equivalente a linearizar em cada iteração, a função de regressão $f(\underset{\sim}{x},\underset{\sim}{\theta})$).

Assim, o problema de minimizar

$$h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (y_t - f(\underset{\sim}{x}_t,\underset{\sim}{\theta}))^2$$

passa a ser o de minimizar na (k+1)-ésima iteração, a função $h^{(k)}(\underset{\sim}{\theta})$, onde

$$h^{(k)}(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (y_t - f^{(k)}(\underset{\sim}{x}_t,\underset{\sim}{\theta}))^2$$

e $f^{(k)}(x_t,\theta)$ é a aproximação de Taylor até 1ª ordem de $f(\underset{\sim}{x}_t,\underset{\sim}{\theta})$ em torno do ponto $\underset{\sim}{\theta}^{(k)}$, isto é:

$$f^{(k)}(\underset{\sim}{x}_t,\underset{\sim}{\theta}) \cong f(\underset{\sim}{x}_t,\underset{\sim}{\theta}^{(k)}) + \sum_{i=1}^{p} \frac{\partial f(\underset{\sim}{x}_t,\underset{\sim}{\theta})}{\partial \theta_i}\Bigg|_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(k)}} \cdot \Delta\theta_i^{(k)},$$

onde $\Delta\theta_i^{(k)} = \theta_i - \theta_i^{(k)}$.

Na forma matricial:

$$h^{(k)}(\underset{\sim}{\theta}) = \| \underset{\sim}{Y} - f(\underset{\sim}{X},\underset{\sim}{\theta}^{(k)}) - F(\underset{\sim}{\theta}^{(k)})\Delta\underset{\sim}{\theta}^{(k)} \|_2^2$$

onde $F(\theta^{(k)})$ é a matriz Jacobiana de $f(\underset{\sim}{X},\underset{\sim}{\theta})$ no ponto $\underset{\sim}{\theta}^{(k)}$, e $\Delta\underset{\sim}{\theta}^{(k)} = \underset{\sim}{\theta} - \underset{\sim}{\theta}^{(k)}$.

Tomando-se $\underset{\sim}{Y} - f(\underset{\sim}{x},\underset{\sim}{\theta}^{(k)}) = E^{(k)}$ temos que:

$$h^{(k)}(\underset{\sim}{\theta}) = \| E^{(k)} - F(\underset{\sim}{\theta}^{(k)})\ \Delta\underset{\sim}{\theta}^{(k)} \|_2^2$$

(onde $E^{(k)}$ "faz o papel" da variável dependente $\underset{\sim}{Y}$, e $F(\theta^{(k)})$ faz o papel da matriz $\underset{\sim}{X}$ dos modelos lineares).

Portanto se $F(\theta^{(k)})$ tiver posto completo, o "valor" $\Delta\underset{\sim}{\theta}$ que minimiza $h^{(k)}(\underset{\sim}{\theta})$ é:

$$\Delta\underset{\sim}{\theta}^{(k)} = (F^{(k)\prime}F^{(k)})^{-1} F^{(k)\prime}E^{(k)} \qquad (2.4.2.1)$$

onde $F^{(k)} = F(\theta^{(k)})$, e o valor de $\underset{\sim}{\theta}$ obtido na (k+1)-ésima iteração será:

$$\underset{\sim}{\theta}^{(k+1)} = \underset{\sim}{\theta}^{(k)} + \Delta\underset{\sim}{\theta}^{(k)}$$

onde $\Delta\underset{\sim}{\theta}^{(k)} = (F^{(k)'} F^{(k)})^{-1} F^{(k)'} E^{(k)}$, $k=0,1,2,...$

(No caso de mínimos quadrados ponderados $\Delta\underset{\sim}{\theta}^{(k)}=(F^{(k)'} W F^{(k)})^{-1} F^{(k)'} W E^{(k)}$

onde $W = \begin{bmatrix} W_1 & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & W_n \end{bmatrix}$). (2.4.2.1.a)

A direção de pesquisa do método de Gauss Newton:

$d^{(k)} = \Delta\underset{\sim}{\theta}^{(k)} = (F^{(k)'} F^{(k)})^{-1} F^{(k)'} E^{(k)}$, pode ser escrita como $d^{(k)} = -(G^{(k)})^{-1} g^{(k)}$ onde $g^{(k)}=-2F^{(k)'} E^{(k)}$ é o vetor gradiente de $h(\underset{\sim}{\theta})$, e $G^{(k)} = 2F^{(k)'} F^{(k)}$ é a matriz Hessiana de $h(\underset{\sim}{\theta})$ quando o 2º termo $B(\underset{\sim}{\theta})$ de (2.4.2) é nulo. Portanto no caso particular da matriz $B(\underset{\sim}{\theta})$ ser $\underset{\sim}{0}$, o método de Gauss Newton coincide com o método de Newton (e por esse motivo ele pode ser visto como uma modificação do método de Newton. (Fletcher 80)).

Nas aplicações, portanto, o algoritmo de Gauss Newton será mais eficiente quanto melhor for a aproximação: $G(\underset{\sim}{\theta}) \cong 2F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta})$. Essa suposição é razoável quando os resíduos $e_t(\underset{\sim}{\theta})$ forem suficientemente pequenos ou então quando o grau de não linearidade do modelo é baixo, isto é, quando o modelo é "próximo" do linear, já que nesses casos $B(\underset{\sim}{\theta}) = \underset{\sim}{0}$.

(Quando o modelo é linear, isto é quando $f(\underset{\sim}{\theta})$ é linear, então o algoritmo de Gauss Newton converge para o ponto de mínimo θ^* numa única iteração partindo-se de qualquer ponto inicial $\underset{\sim}{\theta}^{(0)}$. Entretanto, à medida que o grau de não linearidade de um modelo aumenta, então o algoritmo de Gauss Newton pode não convergir; a função objetivo (soma de quadrados dos resíduos) pode ter mais de um ponto de mínimo e a probabilidade de convergên

cia para o mínimo certo vai diminuindo.

A taxa de convergência do algoritmo básico de Gauss Newton (sem nenhuma modificação) pode ser avaliada através da seguinte desigualdade apresentada por Mc Keown (80)):

$$\frac{\|\underset{\sim}{\theta}^{(k+1)}-\underset{\sim}{\theta}^{*}\|}{\|\underset{\sim}{\theta}^{(k)}-\underset{\sim}{\theta}^{*}\|} \leq \|(F^{*\prime}F^{*})^{-1}B^{*}\| \leq \frac{|\lambda|_{max}(B^{*})}{\lambda_{min}(F^{*\prime}F^{*})} \cong \rho \qquad (2.4.2.2)$$

onde $F^{*}=F(\theta^{*})$, $B^{*}=(\theta^{*})$, $\lambda_{max}(B^{*})$=maior auto valor de B^{*} e $\lambda_{min}(F^{*\prime}F^{*})$=menor auto valor de $F^{*\prime}F^{*}$.

Essa relação mostra que a taxa de convergência, (se ocorrer convergência) não é pior do que a linear, e que se a matriz B^{*} é nula então a taxa de convergência é quadrática. Nos problemas de ajuste de dados entretanto, a matriz B^{*} geralmente não é nula e portanto a taxa de convergência não é melhor do que a linear.

Dessa relação também podemos ver que se B^{*} é "suficientemente grande" (no sentido de que ρ é grande) então ocorrerão provavelmente dificuldades de convergência ao utilizarmos o algoritmo básico de Gauss Newton, e esses problemas (em que o valor de ρ é grande) são denominados "problemas de resíduos grandes") (Mc Keown (80)). Entretanto, um resultado importante demonstrado por Jenrich (69) é que se o tamanho da amostra for suficientemente grande, o algoritmo básico de Gauss Newton convergirá, mesmo em presença de resíduos grandes, se determinadas condições forem satisfeitas.

Outra dificuldade que pode ocorrer com o algoritmo de Gauss Newton é quando existe mal condicionamento. Mal condicionamento é um problema que ocorre em Regressão Linear quando a matriz $\underset{\sim}{X}$ é quase singular (isto é, quando as colunas de $\underset{\sim}{X}$ são quase colineares).

Em Regressão Não Linear esse problema também ocorre ("talvez até com maior frequência" segundo Jenrich (79)), quando $F(\theta^{(k)})$ é quase singular.

O problema de mal condicionamento pode ocorrer simultaneamente com o problema de resíduos grandes. Embora esses dois problemas sejam independentes, (no sentido de que um não afeta o outro), se ocorrerem juntos pode agravar ainda mais a situação, isto é, pode piorar ainda mais o desempenho do algoritmo de Gauss Newton básico, como se pode avaliar através da relação (2.4.2.2) (Essa relação mostra que se $F^{*\prime}F^{*}$ é quase singular, o auto valor mínimo de $F^{*\prime}F^{*}$ é próximo de zero).

Para enfrentar os problemas que ocorrem com o algoritmo de Gauss Newton, temos as seguintes alternativas:

a) utilizar o algoritmo de Gauss Newton com modificações;

b) utilizar outro procedimento numérico;

c) obter melhores valores iniciais.

Na seção 2.4.3 apresentamos algumas modificações do al-

goritmo de Gauss Newton para enfrentar problemas com alto grau de não linearidade (ou problemas de "resíduos grandes") e também problemas de mal condicionamento e na seção 2.4.4 apresentamos alguns procedimentos numéricos alternativos para a solução de problemas de resíduos grandes.

No capítulo 3 apresentamos algumas técnicas para a obtenção de valores iniciais.

2.4.3 MODIFICAÇÕES DO ALGORITMO DE GAUSS NEWTON

Segundo Jenrich (79), o algoritmo de Gauss Newton na sua forma original (isto é, sem modificações), raramente é utilizado. Nessa seção nós apresentamos algumas de suas principais modificações que tratam do problema de alta não linearidade (ou "resíduos grandes") e de mal condicionamento (ou multicolinearidade).

2.4.3.1 NÃO LINEARIDADE

Como já dissemos anteriormente, a suposição de que a matriz $B(\underset{\sim}{\theta}^{(k)}) \cong \underset{\sim}{0}$ nem sempre é válida na prática e nesse caso o problema é denominado: "problema de resíduos grandes". Entretanto a suposição de que $B(\underset{\sim}{\theta}) \cong \underset{\sim}{0}$ pode não ser válida mesmo para resíduos considerados pequenos, se por exemplo $f(\underset{\sim}{\theta})$ for altamente não linear (já que $B(\underset{\sim}{\theta})$ não depende somente dos $e_t(\underset{\sim}{\theta})$, mas também das derivadas segundas).

Uma estratégia para enfrentar um problema altamente não linear é incorporar uma pesquisa unidimensional ao algoritmo básico de Gauss Newton, de tal modo que em cada iteração obtemos uma constante $\alpha^{(k)}$ tal que

$$h(\theta^{(k)} + \alpha^{(k)}\Delta\theta^{(k)}) < h(\theta^{(k)}),\ k=0,1,\ldots,$$

Existem vários procedimentos de pesquisa com esse objetivo. Um procedimento clássico é o Hartley (61), que apresentamos a seguir:

- MODIFICAÇÃO DE HARTLEY

A base da modificação de Hartley é simplesmente a utilização de uma interpolação quadrática, isto é, aproximar a função $h^{(k)}(\lambda) = h(\underset{\sim}{\theta}^{(k)} + \lambda\Delta\underset{\sim}{\theta}^{(k)})$ em cada iteração por uma parábola $h^{(k)}(\lambda) = a^{(k)}\lambda^2 + b^{(k)}\lambda + c^{(k)}$ no intervalo $\{\lambda : 0 \leq \lambda \leq 1\}$ e tomar para $\lambda^{(k)}$ o ponto de mínimo dessa parábola (a idéia é que se $h(\lambda)$ for aproximadamente parabólica nesse intervalo, então $\lambda^{(k)}$ é uma boa aproximação para o ponto de mínimo de $h(\lambda)$).

Um dos procedimentos de Hartley considera, em cada iteração a parábola que passa pelos pontos:

$$(\lambda_1, h^{(k)}(\lambda_1)),\ (\lambda_2, h^{(k)}(\lambda_2)),\ (\lambda_3, h^{(k)}(\lambda_3))$$

onde $\lambda_1 = 0$, $\lambda_2 = \frac{1}{2}$ e $\lambda_3 = 1$.

O ponto de mínimo dessa parábola é dado por:

$$\lambda^{(k)} = -\frac{b^{(k)}}{2a^{(k)}}$$

onde $a^{(k)}$ e $b^{(k)}$ é obtido das equações:

$$\begin{cases} h^{(k)}(0) = c^{(k)} \\ h^{(k)}(1) = a^{(k)} + b^{(k)} + c^{(k)} \\ h^{(k)}\left(\frac{1}{2}\right) = \frac{a^{(k)}}{4} + \frac{b^{(k)}}{2} + c^{(k)} \end{cases}$$

isto é:

$$\lambda^{(k)} = \frac{1}{2} + \frac{1}{4} \cdot \frac{\left[h^{(k)}(0) - h^{(k)}(1)\right]}{\left[h(1) - 2h(\frac{1}{2}) + h(0)\right]}$$

Segundo Gallant (75) e Jenrich (79), no algoritmo de Gauss Newton não faz muita diferença como a pesquisa é feita e que uma regra simples funciona tão bem como qualquer outra.

Um dos esquemas mais simples que tem é o "step halving" (que é utilizado na sub-rotina BMDP3R. Esse procedimento consiste em escolher em cada passo, um número $\alpha^{(k)}$ tal que $\alpha^{(k)}$ seja o primeiro número da sequência $\{1, \frac{1}{2}, \frac{1}{4}, \ldots,\}$ que reduz a soma de quadrados dos resíduos, isto é, tal que

$$h(\theta^{(k+1)}) = h(\underset{\sim}{\theta}^{(k)} + \alpha^{(k)}\Delta\underset{\sim}{\theta}^{(k)}) < h(\underset{\sim}{\theta}^{(k)}) .$$

Assim se $h(\underset{\sim}{\theta}^{(k)} + \Delta\underset{\sim}{\theta}^{(k)}) > h(\underset{\sim}{\theta}^{(k)})$ então calculamos

$$h(\underset{\sim}{\theta}^{(k)} + \frac{1}{2}\Delta\underset{\sim}{\theta}^{(k)}),\ h(\underset{\sim}{\theta}^{(k)} + \frac{1}{4}\Delta\underset{\sim}{\theta}^{(k)}),\ \ldots$$

até que a soma de quadrados seja diminuida.

Para problemas de resíduos grandes, entretanto, existem procedimentos alternativos mais efetivos e que serão discutidos na seção 2.4.4.

2.4.3.2 MAL CONDICIONAMENTO

Uma possível solução para problemas de mal condicionamento é utilizar uma reparametrização conveniente (Draper and Smith (81)). Entretanto segundo Jenrich & Sampson (68) mesmo em Regressão Linear onde reparametrizar é tarefa relativamente mais simples, a maior parte dos pesquisadores preferem manter os parâmetros que são naturais ao problema, e utilizar, por exemplo, um procedimento "Stepwise".

Nessa seção apresentamos uma modificação do método de Gauss Newton, devido a Jenrich & Sampson, que utiliza um procedimento "Stepwise" para enfrentar problemas de mal condicionamento. Esse procedimento é usado pelo BMDP e é denominado "pivotação parcial" (uma técnica muito utilizada em Regressão Linear).

No caso não linear, a idéia é utilizar essa mesma técnica em cada iteração do algoritmo de Gauss Newton (já que em cada passo desse algoritmo temos essencialmente um problema de Regressão Linear onde a matriz $F(\underset{\sim}{\theta}^{k})$ faz o papel da matriz $\underset{\sim}{X}$ de variáveis independentes, $E^{(k)}$ o papel da variável dependente $\underset{\sim}{Y}$, e $\Delta\theta^{(k)}$ o papel de β.

Em linhas gerais, na (k+1)-ésima iteração do algoritmo de Gauss Newton, as "variáveis independentes" (as colunas de $F(\underset{\sim}{\theta}^{(k)})$, entram no modelo de forma "Stepwise", isto é, são selecionadas uma após a outra e de tal modo que a "variável" escolhida em cada passo (do procedimento de "Stepwise") é aquela cuja correlação parcial com a "variável dependente" $E^{(k)}$, dado as "variáveis" já selecionadas, é máxima. Essa "variável", entretanto só entrará efetivamente no modelo se satisfizer o teste de tolerância, isto é, se $T=1-r^2>L$, onde r^2 é o coeficiente de correlação múltipla dessa "variável" com as outras "variáveis independentes" que já entraram no modelo, e L é o limite de tolerância (que no BMDP por exemplo é tomado como 10^{-6}).

Se a "variável" escolhida não satisfizer o critério de tolerância especificado então a componente correspondente de $\Delta\underset{\sim}{\theta}^{(k)} = (\Delta\theta_1^{(k)},\ldots,\Delta\theta_p^{(k)})$ será admitida como nula nessa iteração e o processo continua.

Numericamente este procedimento é executado através de um algoritmo para inversão de matrizes (baseado na pivotação de Gauss Jordan), denominado "Sweeping" (Dixon (85)).

- O MÉTODO DE MARQUARDT

Uma alternativa clássica pará enfrentar problemas mal condicionados é a modificação de Marquardt (63).

Essa técnica consiste em substituir a matriz $(F^{(k)'}F^{(k)})$ do sistema $(F^{(k)'}F^{(k)})\Delta\underset{\sim}{\theta} = F^{(k)'}E^{(k)}$, pela matriz $(F^{(k)'}F^{(k)}+\lambda^{(k)}I)$

onde I é a matriz Identidade e $\lambda^{(k)}$ é uma constante positiva, obtida em cada iteração, através de algum procedimento, (como por exemplo o proposto por Marquardt (63)).

A (k+1)-ésima iteração desse método é portanto:

$$\underset{\sim}{\theta}^{(k+1)} = \underset{\sim}{\theta}^{(k)} + \Delta\underset{\sim}{\theta}^{(k)}$$

onde:

$$\Delta\underset{\sim}{\theta}^{(k)} = (F^{(k)'}F^{(k)} + \lambda^{(k)}I)^{-1} F^{(k)'}E^{(k)}$$

No método de Marquardt a adição de uma constante positiva $\lambda^{(k)}$ aos elementos da diagonal de $F^{(k)'}F^{(k)}$ tem o efeito de aumentar os autovalores de $F^{(k)'}F^{(k)}$ por essa constante, diminuindo portanto o Número de Condição. (Desse modo, pelo menos teoricamente, podemos obter uma matriz $(F^{(k)'}F^{(k)} + \lambda I)$ arbitrariamente bem condicionada escolhendo-se um λ suficientemente grande, mesmo que $F^{(k)'}F^{(k)}$ seja mal condicionada).

O método de Marquardt pode ser visto como uma combinação entre os métodos de Gauss Newton e do "Steepest Descent", já que se $\lambda=0$, a direção de Marquardt coincide com a de Gauss Newton, e para $\lambda\to\infty$ essa direção se aproxima da direção do "Steepest Descent" (como é demonstrado em Marquardt (63)).

Assim o algoritmo de Marquardt aponta aproximadamente na direção do "Steepest Descent" nas iterações onde há mal condicionamento, e na direção de Gauss Newton nas outras iterações.

Nos problemas mal condicionados, o "viés" do método de

Marquardt em direção ao "Steepest Descent" pode ter o efeito de induzir uma convergência lenta, segundo Fletcher (80). Por esse motivo, embora o algoritmo de Marquardt seja frequentemente considerado o melhor método para problemas de mínimos quadrados, muitas vezes ele também pode não ser satisfatório.

O procedimento de Marquardt corresponde a considerar cada iteração do método de Gauss Newton como um problema de "Regressão sobre cristas" (Seber (80)), entretanto o propósito difere, isto é, nesse caso o objetivo não é produzir estimativas alternativas, mas simplesmente lidar com os aspectos computacionais envolvidos no processo de obtenção das estimativas de mínimos quadrados, quando existe mal condicionamento.

O algoritmo de Marquardt é uma das alternativas do SAS (PROC NLIN) para a obtenção das estimativas de mínimos quadrados.

No SAS os $\lambda^{(k)}$'s são obtidos da seguinte maneira: Toma-se como valor inicial $\lambda^{(0)}=10^{-8}$, e na k-ésima iteração se $h(\theta^{(k)})< h(\underset{\sim}{\theta}^{(k-1)})$ então admite

$$\lambda^{(k)} = \frac{\lambda^{(k-1)}}{10} ;$$

se entretanto $h(\theta^{(k)}) > h(\theta^{k-1})$ então $\lambda^{(k)}=10.\lambda^{(k-1)}$. Desse modo se em cada iteração a soma de quadrados $h(\underset{\sim}{\theta}^{(k)})$ é diminuida então $\lambda\to 0$ e estaremos utilizando o método de Gauss Newton, se por outro lado, a soma de quadrados não diminuir então λ é aumentado até que o processo se mova na direção do "Steepest Descent".

- "STAND-UP REGRESSION"

Uma outra alternativa considerada altamente eficaz para problemas de mal condicionamento (e que é utilizada comumente em Regressão Linear (Seber (80)) é empregar alguma forma de decomposição ortogonal da matriz $F(\underset{\sim}{\theta}^{(k)})$ em cada iteração de Gauss Newton, (já que cada iteração de Gauss Newton consiste essencialmente em resolver um problema de regressão linear onde $F(\underset{\sim}{\theta}^{(k)})$ faz o papel da matrix $\underset{\sim}{X}$, e $E^{(k)}$ o papel da variável dependente Y).

A idéia dessas técnicas é evitar a resolução do sistema $F^{(k)'}F^{(k)}\Delta\underset{\sim}{\theta}^{(k)} = F^{(k)'}E^{(k)}$ diretamente, (ou seja evitar a formação da matriz $F^{(k)'}F^{(k)}$, para cálculo da solução $\Delta\underset{\sim}{\theta}^{(k)}$) utilizando alguma forma de decomposição ortogonal da matriz $F^{(k)}$.

Assim por exemplo, se a matriz $F^{(k)}$ puder ser fatorada no produto de uma matriz ortogonal $Q^{(k)}$, por uma matriz $R^{(k)}$ da forma:

$$R^{(k)} = \begin{bmatrix} \tilde{R}^{(k)} \\ \underset{\sim}{0} \end{bmatrix}$$

(onde $\tilde{R}^{(k)}$ é triangular superior) (o que pode ser feito através de uma transformação de Householder (Golub, 65)) então

$$F^{(k)'}F^{(k)}\Delta\underset{\sim}{\theta}^{(k)} = F^{(k)'}E^{(k)} \Longleftrightarrow \tilde{R}^{(k)'}\tilde{R}^{(k)}\Delta\underset{\sim}{\theta}^{(k)} = R^{(k)'}C$$

onde $C = Q^{(k)}.E^{(k)}$.

Portanto:

$\Delta\underset{\sim}{\theta}^{(k)} = \tilde{R}^{(k)^{-1}}.\tilde{C}^{(k)}$, onde $\tilde{C}^{(k)}$ é formada pelos p primeiros elementos de $C^{(k)}$.

Como pudemos ver, a utilização desse procedimento poupou o trabalho de calcular a matriz $F^{(k)'}F^{(k)}$ para resolver o sistema $F^{(k)'}F^{(k)}\Delta\underset{\sim}{\theta}^{(k)} = F^{(k)'}E^{(k)}$, e isso é muito importante pois como se sabe, o número de condição de $(F^{(k)'}F^{(k)})$ é o quadrado do número de condição de $F^{(k)}$, e portanto se $F^{(k)}$ é mal condicionada então $(F^{(k)'}F^{(k)})$ será ainda mais.

A utilização dessas técnicas, portanto, evita a ocorrência de erros desnecessários na determinação da solução $\Delta\underset{\sim}{\theta}^{(k)}$.

Existem vários métodos de decomposição Ortogonal, tais como por exemplo: ortogonalização de Gram-Schmidth, transformação de Householder, transformação de Givens, Decomposição em Valores Singulares, etc., porém nesse trabalho nós não estudaremos esses procedimentos, mas uma abordagem bastante completa sobre esses procedimentos é feita por Lawson & Hanson (74).

- DUPLA PRECISÃO

Em problemas mal condicionados, o uso de dupla precisão é sempre um procedimento bastante seguro.

Segundo Jenrich (79), o uso de dupla precisão em computadores tais como IBM 360 e 370; permite resolver de maneira bastante satisfatória, problemas cuja "Tolerância" mínima é da ordem de 10^{-6}.

Em computadores como o CDC (cuja precisão simples é aproximadamente equivalente à precisão dupla no IBM), o uso de dupla precisão é útil para problemas extremamente mal con-

dicionados.

Para finalizar essa seção é importante salientar que as modificações do algoritmo de Gauss Newton aqui descritas, podem ser combinadas para tentar prevenir simultaneamente , eventuais problemas de mal condicionamento e alta não linearidade.

O BMDP, por exemplo, incorpora simultaneamente (entre outras modificações), a "pivotação parcial" para evitar problemas de mal condicionamento e o "step halving" para problemas altamente não lineares; além disso os cálculos também podem ser processados em dupla precisão em alguns computadores.

2.4.4 MÉTODOS ALTERNATIVOS PARA PROBLEMAS DE RESÍDUOS GRANDES

Quando temos um problema de mínimos quadrados com resíduos grandes então uma opção é ignorar a forma especial do problema (isto é, olhar para $h(\underset{\sim}{\theta})$ como uma função geral) e utilizar um dos métodos para minimização de funções da seção 2.3 (Nesse caso entretanto, a escolha óbvia seria utilizar um dos algoritmos de "Quasi-Newton" da seção 2.3).

Mc Keown (80) observou através de testes computacionais com alguns dos melhores algoritmos de "Quasi-Newton", que em problemas de resíduos grandes esses algoritmos, em geral funcionam melhor do que alguns dos melhores algoritmos

para mínimos quadrados.

Uma outra alternativa para problemas de resíduos grandes, é tentar explorar a forma especial da matriz Hessiana

$$G(\underset{\sim}{\theta}^{(k)}) = \frac{\partial^2 h(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} ,$$

sem ignorar a matriz $B(\underset{\sim}{\theta}^{(k)})$ da relação (2.4.2). Na prática isso significa estimar $B(\underset{\sim}{\theta}^{(k)})$ sem calcular as matrizes das derivadas segundas

$$G_i(\underset{\sim}{\theta}^{(k)}) = \frac{\partial^2 f_i(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}$$

Uma maneira de fazer isso é incluir uma aproximação por "Quasi-Newton" do termo $B(\underset{\sim}{\theta}^{(k)})$, isto é, estimar $B(\underset{\sim}{\theta}^{(k)})$ em cada iteração por um método de "Quasi-Newton"; só que o procedimento nesse caso é bem mais complicado porque somente uma parte da Hessiana em $\underset{\sim}{\theta}^{(k)}$ é aproximada (a outra parte é conhecida exatamente).

As aproximações de "Quasi-Newton" devem satisfazer a seguinte condição:

$$(F^{(k+1)'} F^{(k+1)} + \underset{\sim}{\beta}^{(k+1)}) \Delta\underset{\sim}{\theta}^{(k)} = \Delta g^{(k)}$$

onde

$$\Delta g^{(k)} = g^{(k+1)} - g^{(k)}$$

(Condição de "Quasi-Newton"), e $\underset{\sim}{\beta}^{(k+1)}$ depende de $\beta^{(k)}$ e $F^{(k+1)}$. Segundo Gill, Murray and Wright (81), qualquer uma

das fórmulas apresentadas na seção 2.3 podem ser utilizadas para a construção de $\underset{\sim}{\beta}^{(k+1)}$. Uma fórmula bastante utilizada e que se baseia na fórmula do BFGS, é a seguinte:

$$\underset{\sim}{\beta}^{(k+1)} = \underset{\sim}{\beta}^{(k)} - \frac{1}{\Delta\underset{\sim}{\theta}^{(k)'}(F^{(k)'}F^{(k)}+\underset{\sim}{\beta}^{(k)})\Delta\underset{\sim}{\theta}^{(k)}} \cdot$$

$$\cdot (F^{(k)'}F^{(k)}+\underset{\sim}{\beta}^{(k)})\Delta\underset{\sim}{\theta}^{(k)}\ \Delta\underset{\sim}{\theta}^{(k)'}(F^{(k)'}F^{(k)}+\underset{\sim}{\beta}^{(k)})+$$

$$+ \frac{1}{\Delta g^{(k)'}\Delta\underset{\sim}{\theta}^{(k)}} \cdot \Delta g^{(k)}\Delta g^{(k)'}$$

Existem vários métodos recentes que utilizam uma aproximação de "Quasi-Newton" para a matriz $B(\underset{\sim}{\theta}^{(k)})$, como por exemplo o método de Brown & Dennis (71), o método de Bartholomeu-Biggs (apresentado por Mc Keown (81)) e o método de Dennis & Welch (77). Nesse trabalho entretanto não é nosso objetivo estudar tais métodos, porém Gill, Murray and Wright (81) apresenta uma lista de referência bibliográficas bastante completa sobre esse assunto.

Além dos métodos que utilizam aproximação por "Quasi-Newton" existem também outras alternativas para enfrentar problemas de resíduos grandes. O método de Gill & Murray (78) por exemplo, é um procedimento que contorna simultaneamente possíveis dificuldades com o mal condicionamento.

Outras estrategias possíveis para o problema de resíduos grandes são os métodos híbridos Lavemberg-Marquardt /BFGS, ou então a utilização do método BFGS partindo da matriz ini-

cial $H^{(0)} = 2F^{k'}F^{k}$. Entretanto segundo Fletcher (80), os métodos híbridos não estão suficientemente desenvolvidos para recomendá-los para uso geral, e sobre a 2ª alternativa, ainda há muito o que ser pesquisado.

CAPÍTULO 3

ALGUMAS TÉCNICAS DE OBTENÇÃO DE VALORES INICIAIS

3.1 - INTRODUÇÃO

Qualquer procedimento iterativo de otimização requer para a sua execução a especificação de um valor inicial, isto é de uma estimativa inicial do ponto ótimo.

Em otimização é sempre importante tentar obter bons valores iniciais para o problema. Uma estimativa inicial pobre pode levar o processo a não convergir (dependendo da forma da função objetivo ou do algoritmo utilizado); no caso de existir outros pontos críticos tais como mínimos locais ou pontos de cela, uma estimativa pobre pode levar o processo a convergir para um desses pontos e não para a solução global de interesse.

Em algumas situações nós temos informações de experiências anteriores ou análises semelhantes que podem ser usadas para o "chute" inicial. Em outros casos precisamos utilizar alguma técnica para nos auxiliar na obtenção dessas estimativas.

Existem várias técnicas com essa finalidade e na seção 3.2 apresentamos alguns dos procedimentos que podem ser usados em problemas de análise de regressão.

Nas aplicações, se um determinado procedimento falhar ou então produzir valores iniciais insatisfatórios, devemos recor

rer a outras alternativas na tentativa de encontrar melhores valores iniciais para o problema.

3.2 - TÉCNICAS DE OBTENÇÃO DE VALORES INICIAIS

Nessa seção apresentamos alguns procedimentos utilizados nas aplicações e que funcionam satisfatoriamente em muitos problemas:

1) Uma técnica comumente utilizada para obter estimativas iniciais dos parâmetros de um modelo é "transformar um modelo aproximado". Assim por exemplo se o modelo de regressão é:

$$y_t = \theta_1 e^{\theta_2 x_2 t} + e_t \qquad (3.2.1)$$

ele não pode ser linearizado aplicando-se a transformação logarítmica em ambos os membros de (3.2.1). Entretanto se supusermos que o modelo (3.2.1) tem um erro multiplicativo, isto é, se supusermos que o modelo:

$$y_t = \theta_1 e^{\theta_2 x_2 t} \cdot \varepsilon_t \qquad (3.2.2)$$

é uma aproximação do modelo original (3.2.1), então uma estimativa para θ_1 e θ_2 pode ser obtida facilmente aplicando a transformação logaritmica em (3.2.2).

2) Um procedimento geral para a obtenção de valores iniciais é o método de Hartley & Booker (65).

Esse procedimento consiste em dividir o conjunto de $\underline{n}$ observações $\{(y_t, \underset{\sim}{x}_t),\ t=1,\ldots,n\}$ em $\underline{p}$ grupos (ou conjuntos) $G_1,\ldots,G_p$ com $n_1,\ldots,n_p$ observações respectivamente (onde $\underline{p}$ é o número de parâmetros do modelo e $G_i = \{(y_{t_j}, \underset{\sim}{x}_{t_j}),\ j=1,\ldots n_i\}$) e resolver o sistema não linear:

$$\begin{cases} \overline{y}_1 = \overline{f}_1(\underset{\sim}{\theta}) \\ \vdots \\ \overline{y}_p = \overline{f}_p(\underset{\sim}{\theta}) \end{cases}$$

onde

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{t_j} \quad \text{e} \quad \overline{f}_i(\underset{\sim}{\theta}) = \frac{1}{n_i} \sum_{j=1}^{n_i} f(\underset{\sim}{x}_{tj}, \underset{\sim}{\theta})$$

Um caso particular do método de Hartley & Booker é considerar $n_i=1$, $i=1,\ldots,p$. Esse procedimento é utilizado em Gallant (75), e consiste simplesmente em selecionar $\underline{p}$ pontos (dentre os $\underline{n}$), ao invés de $\underline{p}$ grupos, e resolver o sistema não linear:

$$\begin{cases} y_{t_1} = f(\underset{\sim}{x}_{t_1}, \underset{\sim}{\theta}) \\ \vdots \\ y_{t_p} = f(\underset{\sim}{x}_{t_p}, \underset{\sim}{\theta}) \end{cases}$$

Draper & Smith (81) recomenda que as observações $\underset{\sim}{x}_{ti}$ nesse caso, estejam bem separadas umas das outras, isto é, que as observações sejam representativas do conjunto de dados.

A solução desses sistemas pode ser obtida por exemplo através do método de Gauss Newton para mínimos quadrados, já que achar o ponto de mínimo de $h(\underset{\sim}{\theta}) = \sum_{i=1}^{p} (\overline{y}_i - \overline{f}_i(\underset{\sim}{\theta}))^2$ é equivalente a achar a solução do sistema:

$$\begin{cases} \overline{y}_1 = \overline{f}_1(\underset{\sim}{\theta}) \\ \vdots \\ \overline{y}_p = \overline{f}_p(\underset{\sim}{\theta}) \end{cases}$$

A solução desse sistema, entretanto, requer um valor inicial. Assim, embora o método de Hartley & Booker tenha por finalidade a obtenção de valores iniciais, ele requer também para a sua execução um valor inicial!

Apesar disso, entretanto, esse método é muito importante, pois ele funciona como uma espécie de "filtro" para a obtenção de melhores pontos iniciais.

O método de Hartley & Booker não funcionará, obviamente, se ocorrer convergência para um ponto de cela, ou um mínimo local, ou então quando não houver convergência.

3) Uma técnica alternativa bastante utilizada nas aplicações é considerar uma rede de pontos no espaço paramétrico Θ, tomando para valor inicial por exemplo, o ponto $\underset{\sim}{\theta}^{(0)}$ da rede cujo valor da função objetivo $h(\underset{\sim}{\theta})$ é mínimo (uma alternativa é considerar ao invés de um único ponto, os $\underline{r}$ pontos da rede que produzem os menores valores de $h(\underset{\sim}{\theta})$, e processar o algorítmo de interesse partindo de alguns desses pontos).

Uma estimativa inicial mais precisa pode ser obtida utilizando uma rede "mais fina" na região do espaço paramétrico onde $h(\underset{\sim}{\theta})$ assume os menores valores.

Se o número de parâmetros do modelo é grande, esse

procedimento pode ser bastante complicado. Entretanto se existirem parâmetros que entram linearmente no modelo, o trabalho fica bastante simplificado, já que nesse caso não haverá necessidade de incluí-los na rede. Por exemplo, no modelo $y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_3 e^{\theta_4 x_{3t}} + e_t$, os parâmetros θ_1, θ_2 e θ_3 entram linearmente e assim para cada valor fixado: $\theta_{4(i)}$ de θ_4 (i=1,...,k), as estimativas $\hat{\theta}_{1(i)}$, $\hat{\theta}_{2(i)}$, $\hat{\theta}_{3(i)}$ de θ_1, θ_2 e θ_3 respectivamente, podem ser obtidas por mínimos quadrados lineares, e nesse caso o ponto $\underset{\sim}{\theta}^{(0)}=(\theta_1^{(0)},\theta_2^{(0)},\theta_3^{(0)},\theta_4^{(0)})$ escolhido para valor inicial é aquele que minimiza a função objetivo $h(\underset{\sim}{\theta})$ (dentre os k pontos $\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_k$, onde $\underset{\sim}{\theta}_i=(\hat{\theta}_{1(i)},\hat{\theta}_{2(i)},\hat{\theta}_{3(i)},\theta_{4(i)})$).

4) Uma outra possibilidade para a obtenção de valores iniciais é o procedimento desenvolvido por Ratkowsky (83). Esse procedimento é recomendado principalmente quando a suposição de erro multiplicativo for razoável para o modelo (entretanto também pode ser usado em modelos com estrutura aditiva de erros) e se baseia no criterio de mínimos quadrados ponderados utilizado por Mead (70). Esse critério se baseia em minimizar a expressão:

$$h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} \frac{\sigma^2}{Var(y_t)} (y_t - E(y_t))^2 \qquad (3.2.3)$$

Notamos que se a $Var(y_t)=\sigma^2$, como nos "modelos aditivos" (ou seja, nos modelos com estrutura aditiva de erros) é homocedasticos, então $h(\underset{\sim}{\theta}) = \sum_t (y_t - E(y_t))^2$. Entretanto se $Var(\ell n\ y_t) = \sigma^2$, $h(\underset{\sim}{\theta})$ pode ser aproximador por:

$$h(\underset{\sim}{\theta}) \cong \sum_{t=1}^{n} \frac{1}{\left[E(y_t)\right]^2} (y_t - E(y_t))^2 \qquad (3.2.4)$$

(esse resultado vem do fato de que a $Var(f(y_t))$ pode ser aproximada, segundo Kendall and Stuart (63), por:

$$Var(f(y_t)) \cong \left(\frac{\partial f(y_t)}{\partial y_t} \Big|_{y_t = E(y_t)} \right)^2 . Var(y_t)$$

e portanto se $f(y_t) = \ell n\ y_t$ então $Var(y_t) \cong (E(y_t))^2 Var(\ell n\ y_t)$, que substituindo em (3.2.3) resulta (3.2.4)).

O aspecto interessante dessa técnica é que essa aproximação muitas vezes leva a um problema de mínimos quadrados lineares.

Por exemplo, se $y_t = \frac{1}{\theta_1 + \theta_2 x_t} \cdot e_t$ e $Var(\ell n\ y_t) = \sigma^2$, então $h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (y_t(\theta_1 + \theta_2\ x_t) - 1)^2$ e assim $\underset{\sim}{\theta}_1$ e $\underset{\sim}{\theta}_2$ são obtidas por mínimos quadrados lineares.

3.3 - EXEMPLOS

Exemplo 3.3.1

Consideremos o modelo: $y_t = \frac{1}{\theta_1 + \theta_2\ x_t^{\theta_3}} + e_t$, (3.3.1)

onde $0 < \theta_3 < 1$ e $Var(y_t) = \sigma^2$.

Para obtermos aproximações iniciais para os parâmetros, podemos considerar o seguinte modelo:

$$\frac{1}{y_t} = \theta_1 + \theta_2 \, x_t^{\theta_3} + \frac{1}{e_t} \qquad (3.3.2)$$

(com a suposição aproximada de que $Var(\frac{1}{y_t})$ = constante).

Esse modelo é linear se fixarmos um valor para θ_3. Portanto um procedimento bastante razoável seria considerar uma sequência de valores para θ_3; por exemplo: 0,05, 0,10, 0,15... 0,95, e para cada um desses valores obter as estimativas de mínimos quadrados lineares de θ_1 e θ_2. Nesse caso $h(\underset{\sim}{\theta}) = \Sigma(\frac{1}{y_t} - \theta_1 - \theta_2 x_t^{\theta_3})^2$ como função de θ_3, diminui monotonicamente e depois aumenta (quando θ_3 varia de 0,05 a 0,95), atingindo um mínimo para um certo valor dessa sequência, digamos $\theta_3^{(0)}$. Portanto podemos tomar como valor inicial o ponto $(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})$ (Se desejarmos uma maior precisão podemos considerar intervalos menores para θ_3, por exemplo: 0,025; 0,05; 0.075; 0,01,... 0,975).

As estimativas obtidas dessa maneira, isto é, baseando-se em (3.3.2), podem servir tanto para o "modelo aditivo" (3.3.1) como também para um modelo multiplicativo:

$$y_t = \frac{1}{\theta_1 + \theta_2 \, x_t^{\theta_3}} e_t \quad , \quad \text{onde } Var\,(\ell n\, y_t) = \text{constante}$$

Entretanto para o "modelo multiplicativo", um procedimento mais acurado segundo Ratkwosky (83) é o critério dos mínimos quadrados ponderados" da seção 3.2, que considera a aproximação:

$$h(\underset{\sim}{\theta}) \cong \sum_{t=1}^{n} \left(\frac{y_t}{E(y_t)} - 1 \right)^2 = \sum_{t=1}^{n} (y_t(\theta_1 + \theta_2 x_t^{\theta_3}) - 1)^2 .$$

Assim, para cada θ_3 fixado, podemos obter estimativas de mínimos quadrados lineares para θ_1 e θ_2, repetindo o mesmo procedimento anterior.

Segundo Ratkwosky, esse procedimento (apesar de ser mais adequado para "modelos multiplicativos") também pode ser usado tanto para o modelo (3.3.1) como para o modelo (3.3.2).

Uma outra alternativa baseada em (Gallant (75)) para a obtenção de valores iniciais é resolver o sistema:

$$\begin{cases} y_{t_1} = (\theta_1 + \theta_2 \, x_{t_1}^{\theta_3})^{-1} \\ y_{t_2} = (\theta_1 + \theta_2 \, x_{t_2}^{\theta_3})^{-1} \\ y_{t_3} = (\theta_1 + \theta_3 \, x_{t_3}^{\theta_3})^{-1} \end{cases}$$

onde x_{t_1} , x_{t_2} , x_{t_3} são 3 pontos arbitrariamente escolhidos dentre os $\underline{n}$. A solução desse sistema pode ser obtida por exemplo pelo método de Gauss Newton para a minimização de $h(\underset{\sim}{\theta}) = \sum_{i=1}^{3} (y_{t_i} - (\theta_i + \theta_2 x_{t_i}^{\theta_3}))^2$.

Outra possibilidade é dividir os $\underline{n}$ dados em 3 grupos e resolver o sistema:

$$\begin{cases} \overline{y}_1 = \overline{f}_1(\underset{\sim}{\theta}) \\ \overline{y}_2 = \overline{f}_2(\underset{\sim}{\theta}) \\ \overline{y}_3 = \overline{f}_3(\underset{\sim}{\theta}) \end{cases}$$

onde:

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{t_j} \; ; \quad \overline{f}_i(\underset{\sim}{\theta}) = \frac{1}{n_i} \sum_{j=1}^{n_i} (\theta_1 + \theta_2 \, x_{t_j}^{\theta_3})^{-1}$$

$$i = 1,2,3$$

Exemplo 3.3.2

Consideremos o modelo: $y_t = \theta_1 - \theta_2\ \theta_3^{x_t} + e_t$, (3.3.3)

$0 < \theta_3 < 1$, onde $Var(y_t)$ = constante.

Podemos obter estimativas iniciais para os parâmetros θ_1, θ_2 e θ_3, se utilizarmos o modelo:

$$\ell n(\theta_1 - y_t) = \ell n\ \theta_2 + (\ell n\ \theta_3)x_t + \ell n\ e_t \qquad (3.3.4)$$

(que é resultante da transformação logarítmica no modelo $y_t = \theta_1 - \theta_2\theta_3^{x_t}\ e_t$)

O procedimento mais direto para obter uma estimativa de θ_1 é esboçar o gráfico de Y_t versus X_t, e tomar o valor da assíntota desse gráfico (já que à medida que $x_t \to \infty$, $y_t \to \theta_1$). (Uma outra possibilidade é tomar simplesmente o maior valor observado de Y_t, isto é, $y_{max.}$, ou por exemplo 110% $y_{max.}$).

Obtida uma estimativa de θ_1, o modelo (3.3.4) torna-se linear em $\ell n\ \theta_2$ e $\ell n\ \theta_3$; portanto estimativas de θ_2 e θ_3 podem ser facilmente obtidas.

Nesse procedimento, ao invés de estimar θ_1 pelo gráfico poderíamos também ter considerado uma sequência de valores para θ_1, obtendo para cada valor dessa sequência as estimativas de mínimos quadrados lineares de θ_2 e θ_3, e no final escolher o ponto $\underset{\sim}{\theta} = (\theta_1, \theta_2, \theta_3)$, desse conjunto, que minimiza a função $h(\underset{\sim}{\theta}) = \sum_t (y_t - \theta_1 + \theta_2\theta_3^{x_t})^2$ (no caso do modelo (3.3.3)) ou $h(\theta) = \sum_t (\ell n\ y_t - \ell n(\theta_1 + \theta_2\theta_3^{x_3 t}))^2$ (se o modelo fosse "multiplicativo", isto é, $y = (\theta_1 + \theta_2\theta_3^{x_3 t}) \cdot e_t$).

Uma alternativa bastante simples para a obtenção de valores iniciais do modelo 3.3.3, é simplesmente fixar valores para θ_3 (onde $0<\theta_3<1$), e para cada valor fixado obter as estimativas de θ_1 e θ_2 por mínimos quadrados lineares. Assim o ponto $\underset{\sim}{\theta} = (\theta_1, \theta_2, \theta_3)$ que minimizar $h(\underset{\sim}{\theta})=\Sigma(y_t-\theta_1+\theta_2\theta_3^{x_t})^2$ no caso do modelo (3.3.3) (e $h(\underset{\sim}{\theta})=\Sigma(\ell n\ y_t-\ell n(\theta_1+\theta_2\theta_3^{x_t}))^2$ se o modelo fosse "multiplicativo"), será tomado como valor inicial.

Uma outra alternativa que também pode ser utilizada para obter valores iniciais é a técnica de Hartley & Booker.

Exemplo 3.3.3

Consideremos o modelo: $y_t=\theta_1-\theta_2 e^{-\theta_3^* x_t^{\theta_4}} + e_t$ (3.3.5)

onde $\theta_3^*= e^{-\theta_3}$.

Da mesma forma que no exemplo anterior podemos obter aproximações iniciais dos parâmetros a partir do modelo "linearizado":

$$\log\left(-\log\frac{(\theta_1-y_t)}{\theta_2}\right)= \log\theta_3^*+\theta_4\log x_t + e_t^* \qquad (3.3.6)$$

$$= -\theta_3 + \theta_4\log x_t + e_t^*$$

(que é resultante da transformação logarítmica no modelo:

$$y_t = \theta_1 - \theta_2 e^{-\theta_3^* x_t^{\theta_4}\cdot e_t})$$

Uma estimativa para θ_1 pode ser obtida a partir do gráfico esboçado de y_t versus x_t, observando que à medida que $x_t\to\infty$, $y_t\to\theta_1$ (outra possibilidade é simplesmente tomar o valor máxi-

mo observado de y_t como uma estimativa para θ_1).

Além disso podemos também obter facilmente uma estimativa de θ_2, observando que para x=0 temos que $y_{int}=\theta_1-\theta_2$ assim $\theta_2=\theta_1-y_{int}$, onde y_{int} é o valor do intercepto (que pode ser obtido por exemplo também, a partir do gráfico).

Substituindo as estimativas de θ_1 e θ_2 no modelo (3.3.6) este se torna linear e portanto as estimativas de θ_3 e θ_4 são facilmente obtidas.

Uma outra possibilidade para esse problema é considerar uma sequência de valores para θ_1, e para cada valor dessa sequência, obter uma estimativa de θ_2(através da relação: $\theta_2 = \theta_1 - y_{int}$) e consequentemente as estimativas de θ_3 e θ_4 por mínimos quadrados lineares. O ponto $\underset{\sim}{\theta}=(\theta_1,\theta_2,\theta_3,\theta_4)$, (desse conjunto de pontos), que minimizar $h(\underset{\sim}{\theta})=\Sigma(y_t-\theta_1\theta_2 e^{-\theta_3^{*}\times t^{\theta_4}})^2$ no caso do modelo (3.3.5) (ou de $h(\underset{\sim}{\theta})=\Sigma(\ell n\, y_t-\ell n(\theta_1+\theta_2 e^{-\theta_3^{*}\times t^{\theta_4}}))^2$, no caso de um modelo "multiplicado" da forma $y_t=(\theta_1+\theta_2\, e^{-\theta_3^{*}\times t^{\theta_4}}).e_t$) será a estimativa do ponto inicial.

Uma alternativa para a determinação de θ_3 e θ_4 (sem a utilização do modelo (3.3.6)) é considerar a relação entre o ponto de inflexão do gráfico (esboçado a partir dos dados (x_t, y_t), t=1,...,n) e os parâmetros θ_3 e θ_4. Assim se derivarmos duas vezes a função $f(x_t) = \theta_1-\theta_2\, e^{-\theta_3^{*}\times t^{\theta_4}}$, em relação a x_t, e igualarmos a zero obtemos:

$$x_{infl.} = \left(\frac{\theta_4 - 1}{\theta_3^{*}\ \theta_4}\right)^{1/\theta_4} \qquad (3.3.7)$$

Se substituirmos x_{infl} na relação: $y_t=\theta_1-\theta_2 e^{-\theta_3^{*}\times t^{\theta_4}}$, obte-

mos que

$$y_{infl} = \theta_1 - \theta_2 \, e^{(\frac{1-\theta_4}{\theta_4})} \qquad (3.3.8)$$

Da relação (3.3.8) obtemos $\theta_4 = (\log (\frac{\theta_1 - y_{infl}}{\theta_2}) + 1)^{-1}$ e da relação (3.3.7) obtemos $\theta_3^* = -\log (\frac{\theta_4 - 1}{\theta_4 (x_{infl})^{\theta_4}})$.

Uma outra alternativa para o problema de obter valores iniciais pode ser por exemplo o método geral de Hartley & Booker.

CAPITULO 4

ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA ATRAVÉS DE MÍNIMOS QUADRADOS NÃO LINEARES

4.1 - INTRODUÇÃO

Se tivermos n variáveis aleatórios $Y_1,\ldots,Y_n$ com função de verossimilhança $L(\underset{\sim}{\theta},\underset{\sim}{y})$, então o estimador de máxima verossimilhança de $\underset{\sim}{\theta}$, por definição, é o valor de $\underset{\sim}{\theta}$ que maximiza a função de verossimilhança.

Essa função $L(\underset{\sim}{\theta},\underset{\sim}{y})$ entretanto, muitas vezes não é linear em $\underset{\sim}{\theta}$, e nesses casos o estimador de máxima verossimilhança não tem uma fórmula explícita, tornando necessária a utilização de procedimentos iterativos para a obtenção das estimativas.

Existem vários procedimentos de Otimização Não Linear que podem ser utilizados para a obtenção dessas estimativas; entretanto os métodos clássicos de Newton-Raphson e o "Scoring de Fisher" para a solução de sistemas não lineares são os mais conhecidos e comumente utilizados.

Nessa seção entretanto, nosso objetivo é descrever três maneiras (ou métodos) de transformar um problema de estimação por máxima verossimilhança num problema de ajuste por mínimos quadrados não lineares.

A primeira se refere a problemas em que os dados têm uma distribuição pertencente à família exponencial regular, e as outras duas a problemas em que os dados tem uma distribuição qualquer.

Nessa exposição entretanto, daremos mais ênfase para o primeiro caso, porque se a distribuição envolvida pertence à família exponencial regular, o algoritmo de Gauss-Newton utilizado comumente nos programas de regressão Não Linear, usado de maneira conveniente, coincide com o algoritmo "Scoring de Fisher" para estimação por máxima verossimilhança e em alguns casos com o algoritmo de Newton Raphson; os desvios padrões assintóticos produzidos diferem dos desvios padrões da teoria da Informação, apenas por 1 múltiplo constante; e além disso quando o programa inclui gráficos para análise de resíduos, eles são diretamente aplicáveis para a análise por máxima verossimilhança.

4.2 - RESULTADOS GERAIS PARA A FAMÍLIA EXPONENCIAL REGULAR

Em Regressão Linear, é bastante conhecido que se as observações são normalmente distribuidas, os estimadores de mínimos quadrados coincidem com os estimadores de máxima verossimilhança. No caso não linear Turner, Monroe and Lucas (61), demonstraram a equivalência desses estimadores também supondo normalidade.

Uma questão que surge então é se além da Normal, outras distribuições também levam a mesma conclusão, e nessa linha muitos resultados tem sido publicados.

Um resultado importante estabelecido por Bradley (73) (e demonstrado em seu artigo) para o caso de um modelo de regressão linear: $E(y_t/\underset{\sim}{X}_t) = \underset{\sim}{\theta}'\underset{\sim}{X}_t$ é que se as observações Y_t forem independentes, e tiverem uma distribuição pertencente à família exponencial regular então as equações de verossimilhança coincidem com as equações normais para mínimos quadrados (ponderados

a cada iteração), isto é:

$$\frac{\partial \ell n L(\underset{\sim}{y},\underset{\sim}{\theta})}{\partial \theta_j} = \sum_{t=1}^{n} w_t(\underset{\sim}{\theta})[y_t - \underset{\sim}{\theta}'\underset{\sim}{x}_t]x_{jt} = 0, \quad j=1,\ldots,n$$

onde

$$w_t(\underset{\sim}{\theta}) = \frac{1}{Var(Y_t/\underset{\sim}{X}_t)}$$

depende de $\underset{\sim}{\theta}$ (e portanto os pesos variam de iteração para iteração).

Na forma matricial esse sistema fica:

$$\frac{\partial \ell n L(\underset{\sim}{Y},\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \frac{\partial \underset{\sim}{X}\underset{\sim}{\theta}}{\partial \underset{\sim}{\theta}} \cdot \underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\theta}) \quad (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\theta}) = \underset{\sim}{0}$$

onde

$$\underset{\sim}{\Sigma}(\underset{\sim}{\theta}) = \begin{bmatrix} Var(Y_1/X_1) & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & Var(\underset{\sim}{Y}_n/\underset{\sim}{X}_n) \end{bmatrix}$$

é a matriz de covariância das observações.

Jenrich and Moore (75) extenderam o resultado de Bradley, eliminando a suposição de independência entre as observações Y_t e a suposição de linearidade da função de regressão:

$$E(Y_t/\underset{\sim}{X}_t) = \mu_t(\underset{\sim}{\theta})$$

O resultado é o seguinte:

Se $Y_1,\ldots,Y_n$ são variáveis aleatórias cuja função de verossimilhança é um modelo exponencial regular da forma:

$$L(\underset{\sim}{y},\underset{\sim}{\theta}) = e^{C(\underset{\sim}{\theta})'\underset{\sim}{y}+d(\underset{\sim}{\theta})+g(\underset{\sim}{y})} \qquad (4.2.1)$$

onde:

$$E_{\theta}(\underset{\sim}{Y}) = \mu(\underset{\sim}{\theta})$$

e

$$\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_p)'$$

$$Y = (Y_1,\ldots,Y_n)'$$

$$C(\underset{\sim}{\theta}) = (C_1(\underset{\sim}{\theta}),\ldots,C_n(\underset{\sim}{\theta}))'$$

$$d(\underset{\sim}{\theta}) = \sum_{t}^{n} d_t(\underset{\sim}{\theta})$$

então:

$$\frac{\partial}{\partial\underset{\sim}{\theta}} \ell n\, L(\underset{\sim}{Y},\underset{\sim}{\theta}) = \frac{\partial\mu'}{\partial\underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-}(\underset{\sim}{Y}-\underset{\sim}{\mu}) \qquad (4.2.2)$$

onde:

$\underset{\sim}{\mu} = \mu(\underset{\sim}{\theta}) = E_{\theta}(Y)$

$\underset{\sim}{\Sigma} = \underset{\sim}{\Sigma}(\underset{\sim}{\theta})$ é a matriz de covariâncias das observações

$\underset{\sim}{\Sigma}^{-}$ é tal que $\underset{\sim}{\Sigma}\,\underset{\sim}{\Sigma}^{-}\underset{\sim}{\Sigma} = \Sigma$ (isto é, uma inversa generalizada de $\underset{\sim}{\Sigma}$)

Além disso, se (4.2.2) se mantém para uma função de verossimilhança arbitrária $L(\underset{\sim}{y},\underset{\sim}{\theta})$, então $L(\underset{\sim}{y},\underset{\sim}{\theta})$ é um modelo exponencial.

DEMONSTRAÇÃO:

Seja

$$S(\underset{\sim}{\theta}) = \frac{\partial}{\partial\underset{\sim}{\theta}} \ell n\, L(\underset{\sim}{\theta},\underset{\sim}{Y})$$

o vetor "Score" de Fisher. Então:

$$S(\underset{\sim}{\theta}) = \frac{\partial}{\partial\underset{\sim}{\theta}} C(\underset{\sim}{\theta})'y + \frac{\partial}{\partial\underset{\sim}{\theta}} d(\underset{\sim}{\theta}) \qquad (4.2.3)$$

onde:

$$\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \begin{bmatrix} \frac{\partial C_1(\underset{\sim}{\theta})}{\partial \theta_1} & \cdots & \frac{\partial C_1(\underset{\sim}{\theta})}{\partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial C_n(\underset{\sim}{\theta})}{\partial \theta_1} & & \frac{\partial C_n(\underset{\sim}{\theta})}{\partial \theta_p} \end{bmatrix}$$

e

$$\frac{\partial d(\underset{\sim}{\theta})}{\partial \theta} = \begin{bmatrix} \frac{\partial d(\underset{\sim}{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial d(\underset{\sim}{\theta})}{\partial \theta_p} \end{bmatrix}$$

Como $E(S(\underset{\sim}{\theta})) = \underset{\sim}{0}$, então

$$E\left[\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}' \underset{\sim}{Y} + \frac{\partial d(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right] = \underset{\sim}{0}$$

e portanto

$$\frac{\partial d(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = - \frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}' \mu(\underset{\sim}{\theta})$$

Substituindo essa expressão em (4.2.3) temos que:

$$S(\underset{\sim}{\theta}) = \frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}' (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta})) \qquad (4.2.4)$$

Mas por definição

$$E(Y_t) = \mu_t(\underset{\sim}{\theta}) = \int y_t \, L(\underset{\sim}{y},\underset{\sim}{\theta})d\underset{\sim}{y} \ ,$$

e portanto

$$\frac{\partial \mu(\underset{\sim}{\theta})}{\partial \theta_j} = \int y_t \frac{\partial}{\partial \underset{\sim}{\theta}_j} L(\underset{\sim}{y},\underset{\sim}{\theta})d\underset{\sim}{y} =$$

$$\int y_t \left(\frac{\partial}{\partial \underset{\sim}{\theta}} \ell n L(\underset{\sim}{y},\underset{\sim}{\theta})\right) L(\underset{\sim}{y},\underset{\sim}{\theta})d\underset{\sim}{y} =$$

$$E_{\underset{\sim}{\theta}}(Y_t \, S_j(\underset{\sim}{\theta})) \ ,$$

onde

$$S_j(\underset{\sim}{\theta}) = \frac{\partial}{\partial \underset{\sim}{\theta}} \ell n L(\underset{\sim}{Y},\underset{\sim}{\theta})$$

Em notação matricial:

$$\frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = E_{\underset{\sim}{\theta}}(\underset{\sim}{Y} \, S(\underset{\sim}{\theta})')$$

mas

$$E_{\underset{\sim}{\theta}}(\underset{\sim}{Y} \, S(\underset{\sim}{\theta})') = E_{\underset{\sim}{\theta}}\left(\underset{\sim}{Y}(\underset{\sim}{Y}-\mu(\theta))' \, \frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right) =$$

$$= E_{\underset{\sim}{\theta}}\left(\underset{\sim}{Y}(\underset{\sim}{Y}-\mu(\underset{\sim}{\theta}))'\right)\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = E_{\underset{\sim}{\theta}}(\underset{\sim}{Y}-\mu(\underset{\sim}{\theta}))(\underset{\sim}{Y}-\mu(\underset{\sim}{\theta}))'\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}$$

e portanto

$$\frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \underset{\sim}{\Sigma}\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} \quad , \text{ onde } \underset{\sim}{\Sigma} = \underset{\sim}{\Sigma}(\underset{\sim}{\theta}) \qquad (4.2.5)$$

(Se $\underset{\sim}{\Sigma}$ é inversível então $\frac{\partial C(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \underset{\sim}{\Sigma}^{-1} \frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}$ e (4.2.4) se torna:

$$S(\underset{\sim}{\theta}) = \frac{\partial \mu(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}))$$

Se porém $\underset{\sim}{\Sigma}$ não é inversível e $\underset{\sim}{\Sigma}^{-}$ é uma inversa generalizada de $\underset{\sim}{\Sigma}$, então

$$\underset{\sim}{\Sigma}\underset{\sim}{\Sigma}^{-}(\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}) = (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}))$$

quase certamente, (já que $\underset{\sim}{Y} - \mu(\underset{\sim}{\theta})$ está quase certamente no "range" de $\underset{\sim}{\Sigma}$) e portanto

$$S(\underset{\sim}{\theta}) = \frac{\partial C(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}\underset{\sim}{\Sigma}^{-} (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}))$$

Mas pela expressão (4.2.5):

$$\frac{\partial C(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma} = \frac{\partial \mu(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}}$$

e portanto

$$S(\underset{\sim}{\theta}) = \frac{\partial \mu(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-} (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}))$$

Suponhamos agora que o resultado (4.2.2) seja válido. Então

$$S(\underset{\sim}{\theta}) = \frac{\partial}{\partial \underset{\sim}{\theta}} \ell n L(\underset{\sim}{Y}, \underset{\sim}{\theta}) = \frac{\partial \mu(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-} \underset{\sim}{Y} - \frac{\partial \mu(\underset{\sim}{\theta})'}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-} \mu(\underset{\sim}{\theta})$$

Isso significa que existem funções vetoriais $\alpha(\underset{\sim}{\theta})$ e $\beta_1(\underset{\sim}{\theta}),\ldots,$ $\beta_n(\underset{\sim}{\theta})$ tais que

$$\frac{\partial}{\partial \underset{\sim}{\theta}} \ell n L(\underset{\sim}{y},\underset{\sim}{\theta}) = \alpha(\underset{\sim}{\theta}) + y_1\beta_1(\underset{\sim}{\theta})+\ldots+y_n\beta_n(\underset{\sim}{\theta})$$

Integrando em ambos os lados teremos funções: $d(\underset{\sim}{\theta})$, $C_1(\underset{\sim}{\theta}),\ldots,C_n(\underset{\sim}{\theta})$ e $g(\underset{\sim}{y})$ tais que

$$\ell n(\underset{\sim}{y},\underset{\sim}{\theta}) = y_1C_1(\underset{\sim}{\theta})+\ldots+y_nC_n(\underset{\sim}{\theta}) + d(\underset{\sim}{\theta}) + g(\underset{\sim}{y})$$

o que implica que $L(\underset{\sim}{\theta},\underset{\sim}{y})$ deve ser um modelo exponencial.

Se igualarmos (4.2.2) a zero, temos que:

$$\frac{\partial}{\partial \underset{\sim}{\theta}} \ell n L(\underset{\sim}{Y},\underset{\sim}{\theta}) = \frac{\partial \mu}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{-}(\underset{\sim}{Y}-\underset{\sim}{\mu}) = \underset{\sim}{0}, \text{ onde } \underset{\sim}{\Sigma}^{-}=\underset{\sim}{\Sigma}^{-}(\underset{\sim}{\theta}) \qquad (4.2.6)$$

Isso quer dizer que se $\underset{\sim}{\Sigma}^{-}$ fosse mantida fixa, isto é, não dependesse de $\underset{\sim}{\theta}$, então as "equações de verossimilhança" coincidiriam com as "equações normais" de um ajuste dos $f_t(\underset{\sim}{\theta})=\mu_t(\underset{\sim}{\theta})$ aos dados Y_t's por mínimos quadrados generalizados, usando $\underset{\sim}{\Sigma}^{-}$ como matriz de pesos. No caso particular em que $\underset{\sim}{\Sigma}^{-}$ é uma matriz diagonal, as equações (4.2.6) tomariam a forma das equações normais para mínimos quadrados ponderados.

Entretanto $\underset{\sim}{\Sigma}$ depende de $\underset{\sim}{\theta}$ e portanto varia de iteração para iteração, assim ao invés de um problema de mínimos qua-

drados, temos um problmma de mínimos quadrados "ponderados iterativamente".

O segundo resultado demonstrado por Jenrich and Moore (75), é que a matriz de Informação de Fisher $\tau(\underset{\sim}{\theta})$ pode ser expressa em termos de $\mu(\underset{\sim}{\theta})$ e $\underset{\sim}{\Sigma}(\underset{\sim}{\theta})$ da seguinte maneira:

$$\tau(\underset{\sim}{\theta}) = \text{cov}_{\underset{\sim}{\theta}} S(\underset{\sim}{\theta}) = \frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}' \underset{\sim}{\Sigma}^{-}(\underset{\sim}{\theta}) \frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} \qquad (4.2.7)$$

Esse resultado é importante por duas razões:

1. a primeira é que a direção do algoritmo clássico "Scoring de Fisher" para a obtenção das estimativas de máxima verossimilhança coincide com a direção do algoritmo de Gauss Newton para os mínimos quadrados "ponderados iterativamente".

DEMONSTRAÇÃO:

No método "Scoring", a direção de pesquisa é:

$$d^{(k)} = \left(E\left(\frac{-\partial^2 \ell n L(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}\right)\right)^{-1} \frac{\partial \ell n L(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta}} \quad \text{(Rao, 1965 pág.302)}$$

Portanto

$$d^{(k)} = (\tau(\underset{\sim}{\theta}^{(k)}))^{-1} S(\underset{\sim}{\theta}^{(k)}) =$$

(4.2.2) e (4.2.7)

$$\left(\frac{\partial \mu(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta}}' \underset{\sim}{\Sigma}^{(k)-} \frac{\partial \mu(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta}}\right)^{-1} \frac{\partial \mu(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta}} \underset{\sim}{\Sigma}^{(k)-} (\underset{\sim}{Y} - \mu(\underset{\sim}{\theta}^{(k)})) ,$$

onde $\underset{\sim}{\Sigma}^{(k)-}$ é a inversa generalizada de $\underset{\sim}{\Sigma}$ no ponto $\underset{\sim}{\theta}^{(k)}$.

Assim, se $\Sigma^{(k)}$ é diagonal, essa expressão é exatamente igual a expressão (2.4.2.1.a) (quando $f(\underset{\sim}{\theta}^{(k)})=\mu(\underset{\sim}{\theta}^{(k)})$, $F(\underset{\sim}{\theta}^{(k)})=\frac{\partial \mu(\underset{\sim}{\theta})^{(k)}}{\partial \underset{\sim}{\theta}}$ e $W=\Sigma^{(k)}$, isto é, a direção de pesquisa do método "Scoring" coincide com a direção de pesquisa do algoritmo de Gauss Newton para o ajuste das médias $\mu_t(\underset{\sim}{\theta})$ aos dados Y_t usando $\underset{\sim}{\Sigma}^{(k)}$ como a matriz de pesos.

2. A segunda razão pela qual o resultado (4.2.7) é importante é que ele revela que as matrizes de covariância assintótica dos estimadores de máxima verossimilhança obtidas pelo método "Scoring de Fisher" diferem apenas por um múltiplo $\hat{\sigma}^2$ das matrizes de covariância assintótica dos estimadores de mínimos quadrados obtidos pelo método de Gauss Newton "iterativamente ponderado".

Esse resultado é imediato se considerarmos que para o algoritmo de Gauss Newton "iterativamente ponderado", a estimativa usual para a matriz de covariância assintótica dos estimadores é:

$$\hat{cov}_{\underset{\sim}{\theta}}(\underset{\sim}{\hat{\theta}}) = \hat{\sigma}^2\left(\frac{\partial \mu}{\partial \underset{\sim}{\theta}}(\underset{\sim}{\hat{\theta}})\Sigma^{-}(\underset{\sim}{\hat{\theta}})\frac{\partial \mu}{\partial \underset{\sim}{\theta}}(\underset{\sim}{\hat{\theta}})\right)^{-1},$$

(onde $\hat{\sigma}^2$ é o quadrado médio do resíduo), e portanto por 4.2.7):

$$\hat{cov}_{\underset{\sim}{\theta}}(\underset{\sim}{\hat{\theta}}) = \hat{\sigma}^2(\tau(\underset{\sim}{\hat{\theta}}))^{-1}$$

(Em alguns programas de Gauss Newton, como por exemplo o

BMDP3R, $\hat{\sigma}^2$ é fixado como 1 de tal modo que as duas estimativas coincidem).

O terceiro resultado demonstrado por Jenrich and Moore (75) é para o caso particular de um modelo exponencial regular linear (isto é, quando $C(\underset{\sim}{\theta})$ na expressão (4.2.1) é uma função linear de $\underset{\sim}{\theta}$). Nesse caso o algoritmo "Scoring de Fisher" é exatamente igual ao algoritmo de Newton-Raphson (ou Newton).

DEMONSTRAÇÃO:

A direção do algoritmo "Scoring" é

$$d^{(k)} = (\tau(\underset{\sim}{\theta}^{(k)}))^{-1} . S(\underset{\sim}{\theta}^{(k)}) \qquad (4.2.8)$$

onde:

$$\tau(\underset{\sim}{\theta}) = E \left(- \frac{\partial^2 \ell n L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} \right)$$

mas

$$\frac{\partial^2 \ell n L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} = \frac{\partial}{\partial \underset{\sim}{\theta}} S(\underset{\sim}{\theta}) = \frac{\partial}{\partial \underset{\sim}{\theta}} \left(\frac{\partial}{\partial \underset{\sim}{\theta}} C(\underset{\sim}{\theta})' \underset{\sim}{Y} + \frac{\partial}{\partial \underset{\sim}{\theta}} d(\underset{\sim}{\theta}) \right) =$$

$$= \frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} C(\underset{\sim}{\theta})' \underset{\sim}{Y} + \frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} d(\underset{\sim}{\theta})$$

Mas como $C(\underset{\sim}{\theta})$ é linar por hipótese,

$$\frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} C(\underset{\sim}{\theta}) = \underset{\sim}{0}$$

e portanto

$$\frac{\partial^2 \ell n L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} = \frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} d(\underset{\sim}{\theta})$$

Portanto

$$\tau(\underset{\sim}{\theta}) = E\left(- \frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}, d(\underset{\sim}{\theta}) \right) = - \frac{\partial^2}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}, d(\underset{\sim}{\theta}) =$$

$$= - \frac{\partial^2 \ell n L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} \qquad (4.2.9)$$

Substituindo (4.2.9) em (4.2.8) temos que

$$d^{(k)} = - \left(\frac{\partial^2 \ell n L(\underset{\sim}{\theta}^{(k)})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} \right)^{-1} S(\underset{\sim}{\theta}^{(k)}) ,$$

onde

$$\frac{\partial^2 \ell n L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}$$

é a matriz Hessiana de $\ell n L(\underset{\sim}{\theta})$ e $S(\underset{\sim}{\theta})$ é o vetor gradiente de $\ell n L(\underset{\sim}{\theta})$, isto é, $d^{(k)}$ é exatamente a direção do método de Newton deduzida no capítulo 2.

O que esse resultado nos informa é que se a distribuição envolvida no problema, pertencer à família exponencial regular linear da forma (4.2.1), então os três algoritmos: Gauss Newton "iterativamente ponderado", "Scoring de Fisher" e Newton Raphson são idênticos. Esse fato é muito importante já que o algo

ritmo "Scoring de Fisher" é "robusto" quanto à escolha dos valores iniciais (isto é, não depende muito do ponto inicial para convergir para a solução); enquanto que o algoritmo de Newton como vimos no capítulo 2 apresenta uma convergência rápida perto da solução (converge com taxa quadrática próximo da solução), a aplicação direta desse método, entretanto, exige o cálculo das derivadas segundas, o que nem sempre é facíl de se obter. O algoritmo de Gauss Newton "iterativamente ponderado" entretanto não requer que se escreva, ou mesmo que se conheça explicitamente a função de verossimilhança, mas é suficiente saber apenas que os dados são provenientes de uma distribuição pertencente à família exponencial regular e conhecer as fórmulas das médias das observações Y_t (e suas derivadas primeiras), e das variâncias dos Y_t (como função de $\underset{\sim}{\theta}$).

O quarto resultado e talvez o de maior importância segundo Jenrich (79) é que se o espaço paramétrico (H) for um conjunto aberto e a distribuição envolvida pertencer a uma família exponencial regular linear da forma (4.2.1), então $\underset{\sim}{\theta}$ é um estimador de máxima verossimilhança se e somente se ele é solução de $\frac{\partial \ell n(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \underset{\sim}{0}$. Isso quer dizer que se o algoritmo converge, então $\underset{\sim}{\theta}$ é um ponto de máximo global da função de verossimilhança.

4.3 - MÉTODOS GERAIS

Nessa seção apresentamos dois métodos para a otenção das estimativas de máxima verossimilhança por meio de míni

nos quadrados, para o caso em que a distribuição das observações não pertence necessariamente à família exponencial regular. (Esses métodos, entretanto, requerem que as observações sejam independentes).

O primeiro método é o de Mickey and Britt (74). Nesse método a idéia é considerar o problema equivalente de minimizar a função objetivo $h(\underset{\sim}{\theta}) = -\ell nL(\underset{\sim}{\theta},\underset{\sim}{y})$.

Como as observações são supostas independentes, o problema é minimizar

$$h(\underset{\sim}{\theta}) = -\sum_{t=1}^{n} \ell nL_t(y_t,\underset{\sim}{\theta}) \qquad (4.3.1)$$

Entretanto (4.3.1) pode ser reescrito como:

$$h(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (0 - \sqrt{-\ell n\ L_t(y_t,\underset{\sim}{\theta})})^2$$

Essa expressão, entretanto, pode ser vista como uma soma de quadrados de resíduos do modelo

$$Y_t = f_t^*(\underset{\sim}{\theta}) + e_t$$

onde

$$Y_t = 0,\ t=1,\ldots,n \text{ e } f_t^*(\underset{\sim}{\theta}) = \sqrt{-\ell nL_t(Y_t,\underset{\sim}{\theta})}$$

Portanto as estimativas de máxima verossimilhança podem ser obtidas através da utilização de qualquer algoritmo para mínimos quadrados.

Esse método entretanto é limitado porque, embora ele ge

ralmente produza as estimativas dos parâmetros ele não produz os desvios padrões apropriados para os estimadores; além disso a análise de resíduos fica sem sentido.

O segundo método que apresentamos é o de Jenrich (Dixon (77)).

Jenrich observou que:

$$\frac{\partial}{\partial\theta_i}\ell nL(\underset{\sim}{\theta},\underset{\sim}{y}) = \frac{\partial}{\partial\theta_i}\sum_{t=1}^{n}\ell nL_t(y_t,\underset{\sim}{\theta}) =$$

$$= \sum_{t=1}^{n}\frac{\partial}{\partial\theta_i}\ell nL_t(y_t,\underset{\sim}{\theta})$$

Seja $f_t^*(\underset{\sim}{\theta}) = \ell n\, L_t(y_t,\underset{\sim}{\theta})$; então as equações:

$$\frac{\partial}{\partial\theta_i}\ell nL(\underset{\sim}{\theta},\underset{\sim}{y}) = 0 \iff \sum_{t=1}^{n}\frac{\partial}{\partial\theta_i}f_t^*(\underset{\sim}{\theta}) = 0 \qquad (4.3.2)$$
$$i=1,\ldots,p$$

Entretanto o sistema (4.3.2) é equivalente ao sistema:

$$\sum_{t=1}^{n}(y_t - f_t^*(\underset{\sim}{\theta}))\frac{\partial f^*(\underset{\sim}{\theta})}{\partial\theta_i} = 0$$

(Sistema de equações normais) no caso em que $y_t - f_t^*(\underset{\sim}{\theta}) = 1$, $t=1,\ldots,n$. Em outras palavras isso significa que se admitirmos que a "variável dependente" Y_t, assuma o valor $Y_t = f_t^*(\underset{\sim}{\theta}) + 1$, $t=1,\ldots,n$, então o problema de máxima verossimilhança passa a ser um problema de mínimos quadrados.

Esse procedimento é melhor que o anterior porque, além das estimativas de máxima verossimilhança, nós podemos também

obter uma estimativa para a matriz de covariância assintótica dos estimadores. Porém a sua limitação é que, nesse caso a análise de resíduos fica sem sentido já que $y_t - f_t^*(\underset{\sim}{\theta}) = 1$, $\forall_t$.

4.4 - EXEMPLOS

Nessa seção apresentamos alguns exemplos onde a distribuição das observações é da família exponencial regular da forma (4.2.1).

EXEMPLO 4.4.1 - POISSON

Sejam $Y_1,\ldots,Y_n$ são v.a. independentes onde $Y_t \sim P(\mu_t(\underset{\sim}{\theta}))$ e suponhamos que $\mu_t(\underset{\sim}{\theta}) = n_t\, e^{\theta_1+\theta_2 x_t}$.

Nesse caso

$$E(Y_t) = \mu_t(\underset{\sim}{\theta}) = Var(Y_t) = n_t\, e^{\theta_1+\theta_2 x_t}$$

A função de verossimilhança nesse caso é um modelo exponencial regular linear da forma:

$$L(\underset{\sim}{\theta}, y) = e^{C(\underset{\sim}{\theta})\underset{\sim}{y} + d(\underset{\sim}{\theta}) + g(\underset{\sim}{y})}$$

onde:

$C(\underset{\sim}{\theta}) = (C_1(\underset{\sim}{\theta}),\ldots,C_n(\underset{\sim}{\theta}))$ onde

$C_t(\underset{\sim}{\theta}) = \ell n\, \mu_t(\underset{\sim}{\theta}) = \ell n\, n_t + \theta_1 + \theta_2 x_t$

$d(\underset{\sim}{\theta}) = \sum_{t=1}^{n} d_t(\underset{\sim}{\theta})$, onde $d_t(\underset{\sim}{\theta}) = -\mu_t(\underset{\sim}{\theta})$

$g(\underset{\sim}{y}) = \sum_{t=1}^{n} g(y_t)$, onde $g(y_t) = -\ell n\, y_t!$

A matriz de covariância das observações:

$$\underset{\sim}{\Sigma}(\underset{\sim}{\theta}) = \begin{bmatrix} \mu_1(\underset{\sim}{\theta}) & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \mu_n(\underset{\sim}{\theta}) \end{bmatrix}$$

e portanto a matriz de pesos é:

$$\underset{\sim}{\Sigma}(\underset{\sim}{\theta})^{-1} = \begin{bmatrix} \dfrac{1}{\mu_1(\underset{\sim}{\theta})} & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \dfrac{1}{\mu_n(\underset{\sim}{\theta})} \end{bmatrix}$$

A matriz $\dfrac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = \begin{bmatrix} n_1 e^{\theta_1+\theta_2 x_1} & n_1 x_1 e^{\theta_1+\theta_2 x_1} \\ \vdots & \vdots \\ n_n e^{\theta_1+\theta_2 x_n} & n_n x_n e^{\theta_1+\theta_2 x_n} \end{bmatrix}$

Nesse caso o algoritmo de Gauss Newton para o ajuste do modelo

$$Y_t = \mu_t(\theta) + e_t \quad (\Longleftrightarrow Y_t = n_t\, e^{\theta_1+\theta_2 x_t} + e_t)$$

por mínimos quadrados ponderados iterativamente, onde os pesos são

$$w_t(\underset{\sim}{\theta}) = \frac{1}{\mu_t(\underset{\sim}{\theta})} = \frac{1}{n_t\, e^{\theta_1+\theta_2 x_t}},$$

coincide com o algoritmo "Scoring de Fisher" e com o algoritmo de Newton Raphson para a estimação por máxima verossimilhança.

Além disso podemos obter uma análise completa por máxima verossimilhança: (estimativas dos parâmetros, desvios padrões assintóticos e análise de resíduos).

EXEMPLO 4.4.2 - BINOMIAL - MODELOS DE RESPOSTA QUANTAL

Sejam $Y_1,\ldots,Y_n$, variáveis aleatórias independentes tais que $Y_t \sim B(n_t, \pi_t(\underset{\sim}{\theta}))$.

Nesse caso a função de verossimilhança pertence à família exponencial regular da forma (4.2.1) onde os

$$C_t(\underset{\sim}{\theta}) = \ell n \frac{\pi_t(\underset{\sim}{\theta})}{1+\pi_t(\underset{\sim}{\theta})} \quad , \quad t=1,\ldots,n$$

Além disso

$$E(Y_t) = \mu_t(\underset{\sim}{\theta}) = n_t\pi_t(\underset{\sim}{\theta}),$$

$$Var(Y_t) = n_t\pi_t(\underset{\sim}{\theta})(1-\pi_t(\underset{\sim}{\theta})) \text{ e}$$

$$cov(Y_i,Y_j) = 0, \; i\neq j$$

Nesse caso o algoritmo "Scoring de Fisher" para estimação por máxima verossimilhança coincide com o algoritmo de Gauss Newton iterativamente ponderado para o ajuste das médias $\mu_t(\underset{\sim}{\theta})$ aos dados y_t usando como pesos:

$$w_t(\underset{\sim}{\theta}) = \frac{1}{Var(Y_t)} \,,$$

(já que a inversa da matriz de covariância das observações é:

$$\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\theta}) = \begin{bmatrix} \frac{1}{Var(Y_1)} & \underset{\sim}{0} \\ & \ddots \\ \underset{\sim}{0} & \frac{1}{Var(Y_n)} \end{bmatrix}$$

Na tabela (4.4.2) a seguir os dados são obtidos de um experimento em fundição (Cox (70)).

TABELA 4.4.2

X_t	Y_t	n_t
7	0	55
14	2	157
27	7	159
57	3	16
		N=387

Os dados da 1ª coluna (X_t) representam o tempo de aquecimento; os da 2ª coluna (Y_t), o número de lingotes que ainda não estão prontos para a laminação e os da 3ª coluna (n_t), o número de lingotes testados.

Cox (70) ajustou os dados (Y_t), por um modelo logístico da forma:

$$E(Y_t) = n_t \pi_t(\underset{\sim}{\theta}) = n_t \frac{e^{\theta_1+\theta_2 x_t}}{1+e^{\theta_1+\theta_2 x_t}}$$

Nesse caso

$$C_t(\underset{\sim}{\theta}) = \ell n \frac{\pi_t(\underset{\sim}{\theta})}{1+\pi_t(\underset{\sim}{\theta})} = \theta_1+\theta_2 x_t \quad ,$$

é linear em $\underset{\sim}{\theta}$ e portanto a distribuição dos Y_t pertence à família exponencial linear (e o algoritmo de Gauss Newton "iterativamente ponderado", usando como pesos

$$w_t(\theta) = \frac{1}{n_t \pi_t(\underset{\sim}{\theta})(1-\pi_t(\underset{\sim}{\theta}))} = \frac{(1+e^{\theta_1+\theta_2 x_t})^2}{n_t\, e^{\theta_1+\theta_2 x_t}},$$

coincide com o algoritmo de Newton Raphson.

As fórmulas das derivadas (requeridas pelo algoritmo de Gauss Newton) nesse caso são:

$$\frac{\partial \mu_t(\underset{\sim}{\theta})}{\partial \theta_1} = n_t \frac{e^{\theta_1+\theta_2 x_t}}{(1+e^{\theta_1+\theta_2 x_t})^2} = n_t \pi_t(\underset{\sim}{\theta})(1-\pi_t(\underset{\sim}{\theta})),$$

e

$$\frac{\partial \mu_t(\underset{\sim}{\theta})}{\partial \theta_2} = n_t x_t \frac{e^{\theta_1+\theta_2 x_t}}{(1+e^{\theta_1+\theta_2 x_t})^2} = n_t \pi_t(\underset{\sim}{\theta})(1-\pi_t(\underset{\sim}{\theta}))x_t$$

Dixon (83) apresenta uma análise completa desse exemplo utilizando a sub-rotina BMDP3R. A seguir nós apresentamos alguns desses resultados.

Para esse problema o programa BMDP3R convergiu em seis iterações para os resultados de Cox (70), apesar do uso de um valor inicial arbitrário: $\underset{\sim}{\theta}^{(0)}=(0,0)$, e as estimativas ob-

tidas foram $\hat{\theta}_1 \cong -5,4152$ e $\hat{\theta}_2 = 0,0807$.

As estimativas dos desvios padrões assintóticos (dos estimadores de máxima verossimilhança) foram as seguintes:

$$\sqrt{\hat{Var}(\hat{\theta}_1)} \cong 0,7275 \quad \text{e} \quad \sqrt{\hat{Var}(\hat{\theta}_2)} \cong 0,0224$$

Além desses resultados o programa forneceu também uma estimativa para a matriz de correlações assintóticas dos estimadores. A matriz obtida foi:

$$\begin{bmatrix} 1,0000 & -0,9101 \\ -0,9101 & 1,0000 \end{bmatrix}$$

Os valores preditos (ou frequências esperadas):

$$\hat{y}_t = n_t \cdot \frac{e^{\hat{\theta}_1+\hat{\theta}_2 x_t}}{1+e^{\hat{\theta}_1+\hat{\theta}_2 x_t}} ,$$

e os desvios padrões de $\hat{y}_t$ também foram calculados. Os resultados estão na seguinte tabela:

VALORES OBS.: Y_t	BALORES PREDITOS: $\hat{Y}_t$	DESV.PADRÕES DE $\hat{Y}_t$: $\sqrt{\hat{Var}(\hat{Y}_t)}$
0	0,4271	0,2495
2	2,1322	0,9702
7	6,0132	1,7766
3	3,4275	1,5220

EXEMPLO 4.4.3 - MODELOS MULTINOMIAIS GERAIS

Suponhamos que $Y_1, Y_2, \ldots, Y_n$ têm uma distribuição multinomial com probabilidades: $\pi_1(\underset{\sim}{\theta}), \ldots, \pi_n(\underset{\sim}{\theta})$, onde $\sum_{i=1}^{n} Y_i = N$. Essa distribuição aparece por exemplo na análise de tabelas de contingência.

Verifica-se facilmente que a função de verossimilhança

$$L(y, \underset{\sim}{\theta}) = P(Y_1 = y_1, \ldots, Y_n = y_n) =$$

$$= \frac{N!}{y_1! \ldots y_n!} \pi_1(\underset{\sim}{\theta})^{y_1} \ldots \pi_{n-1}(\underset{\sim}{\theta})^{y_{n-1}} \cdot$$

$$\cdot \left(1 - \sum_{i=1}^{n-1} \pi_i(\underset{\sim}{\theta})\right)^{N - \sum_{i=1}^{n-1} y_i}$$

é um modelo exponencial regular onde

$$C(\underset{\sim}{\theta}) = (C_i(\underset{\sim}{\theta}))$$

e

$$C_i(\underset{\sim}{\theta}) = C(\pi_i(\underset{\sim}{\theta})) = \ell n(\pi_i(\underset{\sim}{\theta}) / (1 - \sum_{i=1}^{n-1} \pi_i(\underset{\sim}{\theta}))), \quad i = 1, \ldots, n-1$$

e além disso $E(Y_i) = \mu_i(\underset{\sim}{\theta}) = N\pi_i(\underset{\sim}{\theta})$ e a matriz de covariância das observações:

$$\underset{\sim}{\Sigma}(\underset{\sim}{\theta}) = \begin{bmatrix} \pi_1(\underset{\sim}{\theta})(1-\pi_1(\underset{\sim}{\theta})) & -\pi_1(\underset{\sim}{\theta})\pi_2(\underset{\sim}{\theta}) & \cdots & -\pi_1(\underset{\sim}{\theta})\pi_n(\underset{\sim}{\theta}) \\ -\pi_2(\underset{\sim}{\theta})\pi_1(\underset{\sim}{\theta}) & \pi_2(\underset{\sim}{\theta})(1-\pi_2(\underset{\sim}{\theta})) & & -\pi_2(\underset{\sim}{\theta})\pi_n(\underset{\sim}{\theta}) \\ \vdots & & \ddots & \\ -\pi_n(\underset{\sim}{\theta})\pi_1(\underset{\sim}{\theta}) & \cdots & & \pi_n(\underset{\sim}{\theta})(1-\pi_n(\underset{\sim}{\theta})) \end{bmatrix}$$

A matriz $\underset{\sim}{\Sigma}(\underset{\sim}{\theta})$ apesar de não ser diagonal (já que as observações não são independentes), tem uma inversa generalizada diagonal da forma:

$$\underset{\sim}{\bar{\Sigma}}(\underset{\sim}{\theta}) = \begin{bmatrix} \frac{1}{N\pi_1(\underset{\sim}{\theta})} & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \frac{1}{N\pi_n(\underset{\sim}{\theta})} \end{bmatrix},$$

como pode ser facilmente verificado.

Assim podemos ajustar os $\mu_i(\underset{\sim}{\theta})=N\pi_i(\underset{\sim}{\theta})$ aos y_i's utilizando como pesos variáveis: $w_i(\underset{\sim}{\theta})=\frac{1}{N\pi_i(\underset{\sim}{\theta})}$, como fizemos no caso da Poisson.

Como um exemplo consideremos a tabela a seguir que contém as frequências observadas para os grupos sanguíneos A-B-O numa população humana (Rao (73), pág. 370).

Seja p, q e r as frequências genotípicas para os genes A, B e O (como a soma dessas frequências é 1 então r=1-p-q e portanto somente p e q são parâmetros desconhecidos, ou seja $\underset{\sim}{\theta}=(p,q)=(\theta_1,\theta_2)$).

FREQUÊNCIAS FENOTÍPICAS PARA OS GRUPOS SANGUÍNEOS A-B-O NUMA POPULAÇÃO HUMANA

TIPO	FREQ. FENOT. (π_i)	Y_i
O	r^2	176
A	p^2+2pr	182
B	q^2+2qr	60
AB	$2pq$	17

TABELA 4.4.4 N=255

Nesse caso: $\pi_1(\underset{\sim}{\theta}) = (1-p-q)^2=(1-\theta_1-\theta_2)^2$

$\pi_2(\underset{\sim}{\theta}) = (p^2+2p(1-p-q))=\theta_1(2-\theta_1-2\theta_2)$

$\pi_3(\underset{\sim}{\theta}) = (q^2+2q(1-p-q))=\theta_2(2-\theta_2-2\theta_1)$

$\pi_4(\underset{\sim}{\theta}) = 2pq \qquad\qquad =2\theta_1\theta_2$

Assim $C_i(\underset{\sim}{\theta}) = \ell n(\pi_i(\underset{\sim}{\theta})/(1-\sum_{i=1}^{3}\pi_i(\underset{\sim}{\theta})))$ não é linear, e portanto a família exponencial resultante também não é linear; (e nesse caso o algoritmo de Gauss Newton é equivalente ao algoritmo "Scoring de Fisher" porém não é equivalente a algoritmo de Newton-Raphson).

Para a utilização do algoritmo de Gauss Newton precisamos fornecer as fórmulas das derivadas $\frac{\partial \mu_t(\underset{\sim}{\theta})}{\partial \theta_j}$ (necessárias para a formação da matriz Jacobiana $\frac{\partial \mu(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}$.

Nesse caso as fórmulas são as seguintes:

$$\frac{\partial \mu_1(\underset{\sim}{\theta})}{\partial \theta_1} = -2N(1-\theta_1-\theta_2) \quad ; \quad \frac{\partial \mu_1(\underset{\sim}{\theta})}{\partial \theta_2} = -2N(1-\theta_1-\theta_2)$$

$$\frac{\partial \mu_2(\underset{\sim}{\theta})}{\partial \theta_1} = 2N(1-\theta_1-\theta_2) \quad ; \quad \frac{\partial \mu_2(\underset{\sim}{\theta})}{\partial \theta_2} = -2N\theta_1$$

$$\frac{\partial \mu_3}{\partial \theta_1}(\underset{\sim}{\theta}) = -2N\theta_2 \quad ; \quad \frac{\partial \mu_3}{\partial \theta_2}(\underset{\sim}{\theta}) = 2N(1-\theta_1-\theta_2)$$

$$\frac{\partial \mu_4}{\partial \theta_1}(\underset{\sim}{\theta}) = 2N\theta_2 \quad ; \quad \frac{\partial \mu_4}{\partial \theta_2}(\underset{\sim}{\theta}) = 2N\theta_1$$

Jenrich and Moore (75) utilizaram a sub-rotina BMDP3R para analisar os dados da tabela 4.4.3. Eles observaram que o processo convergiu em somente três iterações partindo de um valor arbitrário: $\theta_1^{(0)}=\theta_2^{(0)}=0,3$. Os resultados encontrados foram os seguintes:

$$\hat{\theta}_1 = 0,2644 \qquad \sqrt{\hat{Var}(\hat{\Theta}_1)} = 0,01622$$

$$\hat{\theta}_2 = 0,09169 \qquad \sqrt{\hat{Var}(\hat{\Theta}_2)} = 0,1010$$

e a tabela de frequências esperadas ($\hat{Y}_i$) e desvios padrões de $\hat{Y}_i$:

TIPO DE SANGUE	FREQ. ESPER.($\hat{Y}_i$)	$\sqrt{\hat{Var}(\hat{Y}_i)}$
O	179,5	9,82
A	178,2	9,73
B	55,8	6,01
AB	21,4	2,47

CAPÍTULO 5

MEDIDAS DE NÃO LINEARIDADE

5.1 - INTRODUÇÃO

Segundo Guttman & Meeter (65),"medidas de não linearidade são expressões que indicam se o grau de não linearidade num problema de estimação não linear é pequeno o suficiente para justificar a utilização dos resultados usuais da teoria dos modelos lineares como aproximação para os não lineares".

Até recentemente não existia nenhum método facilmente aplicável que quantificasse o comportamento não linear de um "modelo" (ou melhor da combinação modelo/conjunto de dados). A primeira tentativa relevante para medir a não linearidade foi feita por Beale (60), porém a sua medida, segundo Guttman and Meeter (65), tende a subestimar a verdadeira não linearidade.

Mais recentemente, Box (71), desenvolveu uma fórmula para estimar o viés dos estimadores de máxima verossimilhança, e Bates & Watts (80) desenvolveram medidas de não linearidade baseando-se no conceito geométrico de curvatura.

Nesse capítulo nós vamos apresentar as medidas de curvatura de Bates & Watts e a medida do viés de Box, porém antes de apresentá-las nós introduziremos na seção 5.2, o conceito geométrico de curvatura do espaço de estimação (ou "locus de solução") através da representação geométrica dos modelos de Regressão Não Lineares no espaço amostral, em contraste com a

dos modelos lineares, o que será bastante útil para introduzir as medidas de Bates & Watts na seção 5.3.

Ratkowsky (83) fornece um programa para o cálculo das medidas de Bates & Watts e do viés de Box. Este programa já está implantado nos computadores Burroughs e C.D.C. do CCE-USP.

5.2 - REPRESENTAÇÃO GEOMÉTRICA DOS MODELOS DE REGRESSÃO NÃO LINEARES NO ESPAÇO AMOSTRAL

Consideremos o modelo $\underset{\sim}{Y} = f(\underset{\sim}{\theta}) + \underset{\sim}{e}$ onde

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad f(\underset{\sim}{\theta}) = \begin{bmatrix} f(\underset{\sim}{x}_1, \underset{\sim}{\theta}) \\ \vdots \\ f(\underset{\sim}{x}_n, \underset{\sim}{\theta}) \end{bmatrix}$$

$$\underset{\sim}{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, \quad \underset{\sim}{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

e

$$\underset{\sim}{X}_t = (X_{t1}, \ldots, X_{tk}), \quad t=1,\ldots,n$$

onde $E(\underset{\sim}{e}) = \underset{\sim}{0}$ e $Var(\underset{\sim}{e}) = \sigma^2 I$.

Vamos considerar inicialmente o caso onde n=2 e p=1. Esse caso apesar de extremamente simples do ponto de vista experimental, serve muito bem para ilustrar o princípio conceitual envolvido. Porém antes de representarmos graficamente um modelo não linear, é muito importante entendermos o que ocorre no caso linear.

(a) Caso linear

Seja

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad f(\theta) = \underset{\sim}{X}\theta$$

onde

$$\underset{\sim}{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ e } \theta \in R$$

O espaço amostral nesse caso tem dimensão 2, e o espaço de estimação (ou "locus de solução") é um subespaço de dimensão 1 do espaço amostral, e é constituido pelos pontos da forma

$$f(\theta) = \underset{\sim}{X}\theta = \begin{bmatrix} x_1\theta \\ x_2\theta \end{bmatrix}$$

ou seja, é uma reta no R^2 cuja direção é a do vetor $\underset{\sim}{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Uma característica importante dos modelos lineares é que tomando-se valores de θ igualmente espaçados (isto é, $\Delta\theta$ = cte, onde $\Delta\theta = \theta_i - \theta_{i-1}$, i=1,2,3,...) então os pontos correspondentes no locus de solução: $\underset{\sim}{X}\theta_0$, $\underset{\sim}{X}\theta_1$, $\underset{\sim}{X}\theta_2$,... também são igualmente espaçados.

Na figura 5.2.1 a seguir, consideramos os eixos 1 e 2 como base do espaço amostral e tomamos $\underset{\sim}{Y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, $\underset{\sim}{X} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\theta_0 = 1$, $\theta_1 = 2$, $\theta_2 = 3$,...

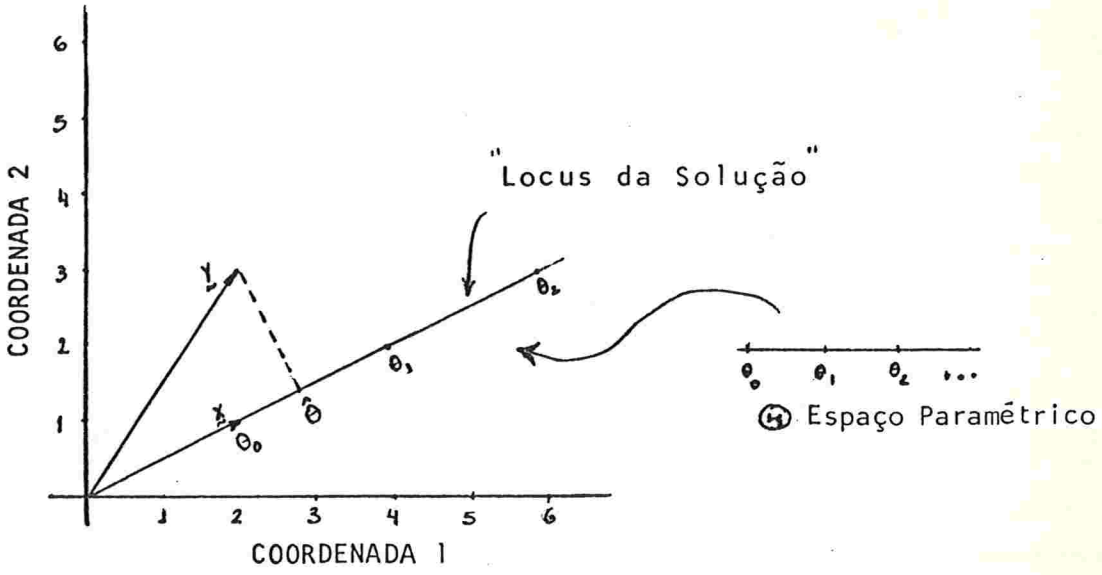

FIG. 5.2.1 - REPRESENTAÇÃO DOS DADOS NO ESPAÇO AMOSTRAL

O locus de solução nesse caso é a reta cuja direção é dada pelo vetor $\underset{\sim}{X} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, e os pontos correspondentes a θ_0, θ_1,... no "locus" são:

$$\underset{\sim}{X}\theta_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \underset{\sim}{X}\theta_1 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \quad \underset{\sim}{X}\theta_3 = \begin{bmatrix} 6 \\ 3 \end{bmatrix} \ldots$$

e portanto são equidistantes (na figura os valores de θ_0, θ_1, θ_2, ... são marcados no lugar de $\underset{\sim}{X}\theta_0$, $\underset{\sim}{X}\theta_1$,...).

A função

$$h(\theta) = \sum_{t=1}^{2} (Y_t - \theta X_t)^2 = \| \underset{\sim}{Y} - \underset{\sim}{X}\theta \|^2$$

representa o quadrado da distância do ponto $\underset{\sim}{Y}$ a um ponto genérico $\underset{\sim}{X}\theta$ do espaço de estimação; assim minimizar $h(\theta)$ em relação a θ corresponde geometricamente a encontrar um ponto no

espaço de estimação cuja distância é a menor possível. Na figura, o ponto $\hat{\theta}$ (que corresponde ao ponto $P=\underset{\sim}{X}\hat{\theta}$ nas coordenadas 1 e 2) é o ponto mais próximo de $\underset{\sim}{Y}$, ou seja, é o estimador de mínimos quadrados de θ.

(b) Consideremos agora o caso não linear (onde n=2 e p=1).

Nesse caso o locus de solução não é mais uma reta como antes, mas uma curva gerada pelos pontos da forma

$$f(\theta) = \begin{bmatrix} f(X_1,\theta) \\ f(X_2,\theta) \end{bmatrix} ,$$

além disso os pontos do locus de solução correspondentes a $\Delta\theta$ = constante, não são mais necessariamente igualmente espaçados.

Para ilustrar consideremos a figura 5.2.2.

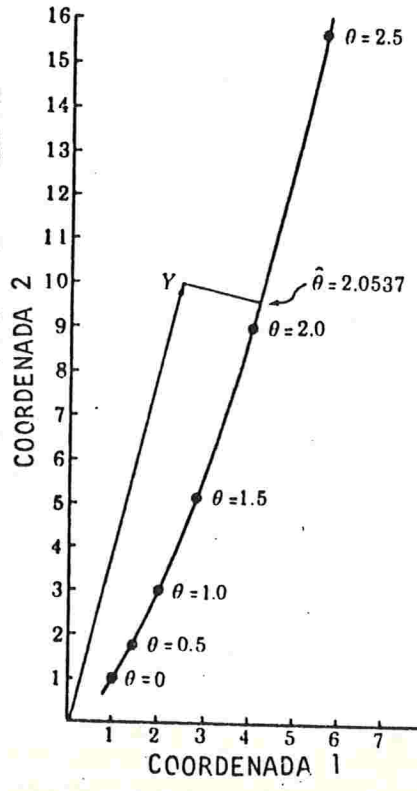

FIG. 5.2.2 - REPRESENTAÇÃO DOS DADOS NO ESPAÇO AMOSTRAL (Figura extraida de Ratkowsky (83))

Nessa figura representamos no espaço amostral o modelo:

$$\underset{\sim}{Y} = f(\theta) + \underset{\sim}{e} \qquad (5.2.1)$$

onde:

$$\underset{\sim}{Y} = \begin{bmatrix} 2,5 \\ 10 \end{bmatrix}, \quad \underset{\sim}{X} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$f(\theta) = \begin{bmatrix} f(X_1,\theta) \\ f(X_2,\theta) \end{bmatrix} = \begin{bmatrix} X_1^{\theta} \\ X_2^{\theta} \end{bmatrix}$$

O locus de solução nesse caso consiste dos pontos da forma

$$f(\theta) = \begin{bmatrix} 2^{\theta} \\ 3^{\theta} \end{bmatrix},$$

onde $\theta \in R$.

Como podemos ver o locus de solução não é mais uma reta mas sim uma curva em torno de $\hat{\theta}$; o valor $\hat{\theta} \cong 2,05$ (correspondente ao ponto $P = (2^{2,05}, 3^{2,05})$ nas coordenadas 1 e 2, cuja distância a $\underset{\sim}{Y}$ é mínima) é a estimativa de mínimos quadrados de θ.

Além disso os pontos do espaço de estimação correspondentes a iguais incrementos $\Delta\theta$ (no caso $\Delta\theta = 0,5$), não são igualmente espaçados como no caso linear mas apresentam incrementos crescentes.

Como vemos, os pontos do locus de solução em torno

de $\hat{\theta}$ e o tipo de espaçamento existente entre os pontos de locus correspondentes a $\Delta\theta$ = constante diferem entre os modelos lineares e os não lineares, e portanto podem ser usados como medidas de não linearidade de um modelo.

Assim, quanto maior a curvatura do locus de solução nas vizinhanças de $\hat{\theta}$, ou seja, quanto mais o locus se afasta da reta tangente em $\hat{\theta}$, maior será o que Bates & Watts (80) definem por "não linearidade intrínsica" do modelo, além disso quanto mais desiguais forem os espaçamentos entre os pontos, maior será o que Bates & Watts definem por "não linearidade causada pela parametrização do modelo".

Os termos "não linearidade devido a parametrização " e "não linearidade intrínsica", assim definidos por Bates & Watts são bastante apropriados já que no primeiro caso o grau de não linearidade depende da maneira como os parâmetros aparecem no modelo podendo assim ser reduzido através de reparametrizações; enquanto que no segundo caso o grau de não linearidade não é alterado através de reparametrizações do modelo.

Para ilustrar esse fato consideremos a seguinte reparametrização do modelo (5.2.1):

$$Y_t = X_t^{\ell n \phi} + e_t \quad , \text{(onde } \phi = e^{\theta}\text{).} \qquad (5.2.2)$$

A representação desse modelo reparametrizado no espaço amostral é apresentada na figura 5.2.3.

Como podemos observar, a curvatura do locus de solu-

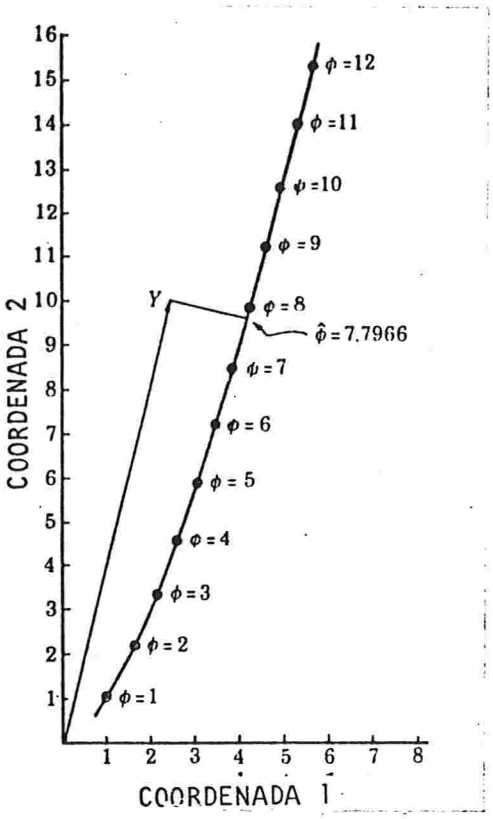

FIG. 5.2.3 - REPRESENTAÇÃO DOS DADOS NO ESPAÇO AMOSTRAL

ção é a mesma da figura 5.2.2, porém o espaçamento entre os pontos correspondentes a $\Delta\theta=1$ agora são praticamente iguais indicando que o grau de não linearidade "causado pela parametrização" foi bastante reduzido e portanto espera-se que o modelo (5.2.2) exiba um comportamento "mais próximo do linear" nas suas propriedades estatísticas do que o modelo (5.2.1), para o mesmo conjunto de dados.

Consideremos agora a situação onde n=3 e p=2. Se o modelo é linear, o locus de solução é um plano (Draper & Smith (81)), e consequentemente a não linearidade intrínsica é zero qualquer que seja o conjunto de dados; além disso se tomarmos retas paralelas e equidistantes no espaço paramétrico então teremos também retas paralelas e equidistantes no locus de solu-

ção e portanto a não linearidade causada pela parametrização também é zero.

Num modelo não linear a situação é diferente, o locus de solução é uma superfície curva em torno de $\hat{\theta}$, e retas paralelas e equidistantes no espaço paramétrico geram curvas nesse locus denominadas "linhas paramétricas", que não são mais semelhantes ao do caso linear, isto é, não são mais retas paralelas e equidistantes. A figura 5.2.4 ilustra o caso onde p=2 e n=3.

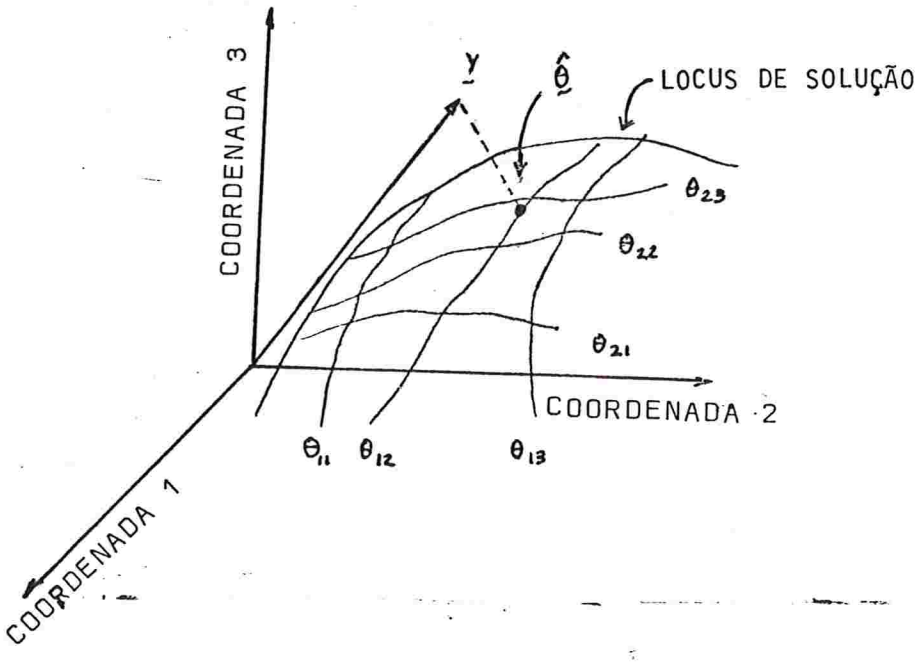

FIG. 5.2.4 - REPRESENTAÇÃO NO ESPAÇO AMOSTRAL DE UM MODELO NÃO LINEAR NO CASO (n=3, p=2)

Quando n>3 e p≥3 é impossível representar graficamente o problema completamente, porém o princípio conceitual per-

manece válido.

Baseando-se nesses princípios Bates & Watts definem medidas quantitativas de não linearidade. Essas medidas são definidas a partir da especificação do modelo e do conjunto de dados e constituem uma ferramenta de grande valor para a avaliação da não linearidade.

5.3 - MEDIDAS DE CURVATURA DE BATES & WATTS

Seja k_h^N a curvatura normal do locus de solução correspondendo a uma particular direção h do espaço paramétrico.

Quando n=3 e p=2, a interpretação de k_h^N é a do inverso do raio do círculo que melhor aproxima o locus de solução na direção do vetor "velocidade instantânea" no ponto $\hat{\theta}$ (isto é, na direção de:

$$\dot{\eta}_h = \frac{d}{dt}\, \eta(\hat{\underset{\sim}{\theta}}+t\underset{\sim}{h}) \Big|_{t=0}$$

onde $\eta(\underset{\sim}{\theta})$ representa o locus de solução).

Bates & Wates (80) denominaram k_h^N de "curvatura intrínsica", já que esta é inerente ao locus de solução, e não depende da particular parametrização do modelo. Além disso eles definem uma medida de "não linearidade intrínsica" como sendo a curvatura normal máxima do locus de solução em $\hat{\underset{\sim}{\theta}}$: $k^N = \max_h\{k_h^N\}$.

Portanto se essa medida é suficientemente baixa, o plano

tangente nesse ponto será uma boa aproximação para a superfície.

Bates & Watts definem k_h^T, como a curvatura causada pela parametrização do modelo correspondendo a uma particular direção $\underset{\sim}{h}$ do espaço paramétrico.

Essa curvatura depende da particular parametrização escolhida e portanto pode ser reduzida consideravelmente através de uma reparametrização conveniente do modelo. Bates & Watts definem também a não linearidade ("aparente") causada pela parametrização, como a curvatura máxima, isto é:

$$K^T = \max_{\underset{\sim}{h}} \{k_h^T\}$$

Essa medida está associada com o fato de que as projeções das linhas paramétricas no plano tangente ao locus de solução não são retas paralelas e equidistantes. Portanto se o valor dessa medida for suficientemente baixo, podemos substituir como uma boa aproximação as linhas paramétricas curvas no plano tangente por uma rede de linhas paralelas e equidistantes.

Bates & Watts definem medidas de curvatura relativa γ_h^N e γ_h^T de tal modo que sejam invariantes por mudança de escala. Isso pode ser feito dividindo as observações $\underset{\sim}{Y}$ e o modelo por uma constante ρ o que implica que

$$\gamma_h^N = \rho k_h^N \quad \text{e} \quad \gamma_h^T = \rho k_h^T.$$

Essas medidas relativas podem ser usadas não somente para comparar diferentes parametrizações de um determinado problema, mas também diferentes conjuntos de dados para o mesmo modelo ou para modelos diferentes.

Um valor bastante sugestivo para ρ é $\hat{\sigma}\sqrt{p}$. Isso porque no caso linear a região de confiança de $(1-\alpha)\%$ para $\underset{\sim}{X}\underset{\sim}{\theta}$, contida no locus de solução é um círculo (quando $p=2$) de centro $P=\underset{\sim}{X}\underset{\sim}{\theta}$ e raio

$$r = \sqrt{h(\hat{\theta})\,\frac{p}{n-p}\,F(p,n-p,1-\alpha)} =$$

$$= \hat{\sigma}\sqrt{p}\,\sqrt{F(p,n-p,1-\alpha)}$$

(Draper & Smith (81)), ou seja $r = \rho\sqrt{F}$

Se o problema é "padronizado" (dividido por ρ), então o raio de curvatura é simplesmente

$$r = \sqrt{F(p,n-p,1-\alpha)}$$

(e portanto a curvatura da região de confiança de $(1-\alpha)\%$ é

$$\frac{1}{\sqrt{F(\alpha,n-p,1-\alpha)}}\ )$$

Os autores utilizam esse raio de curvatura $\sqrt{F}$, como padrão para comparar os raios de curvatura relativos máximos $\frac{1}{\gamma^N}$ e $\frac{1}{\gamma^T}$. Assim se $\frac{1}{\gamma^N}$ é grande quando comparado com $\sqrt{F}$, isto é, se $\gamma^N < \frac{1}{\sqrt{F}}$, então o locus de solução é relativamente plano

sobre a região de confiança, e portanto nós podemos supor com razoável segurança que o locus de solução é relativamente plano (ou seja que a não linearidade intrínsica é pequena).

Para ilustrar esse resultado consideremos a figura 5.3.1.

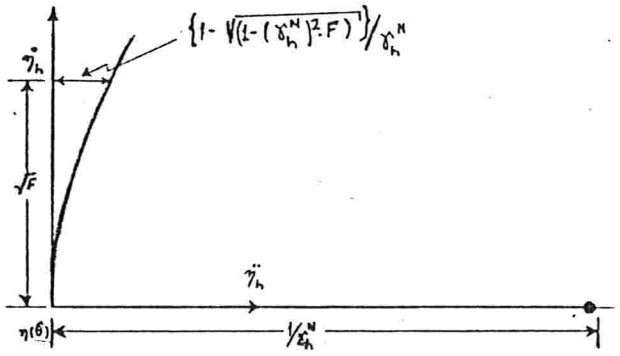

FIG. 5.3.1 - DESVIO ENTRE A APROXIMAÇÃO CIRCULAR E O PLANO TANGENTE

Bates & Watts (78) mostram que o maior desvio entre a aproximação circular (para o locus de solução) e o plano tangente é igual a:

$$(1 - \sqrt{1-(\gamma_h^N)^2 \cdot F}\ )/\gamma_h^N$$

Portanto à medida que o raio da região de confiança $\sqrt{F}$ se aproxima do raio de curvatura $\frac{1}{\gamma_h^N}$, o desvio aumenta até um va

lor máximo $\sqrt{F} = \frac{1}{\gamma_h^N}$, e a partir daí, a aproximação "se quebra".

Para a 2ª componente da não linearidade, os autores também consideram como padrão de comparação a curvatura do intervalo de confiança; assim se

$$\gamma^T < \frac{1}{\sqrt{F}}$$

então a curvatura máxima causada pela parametrização é considerada pequena.

Uma consequência desses resultados é que se um determinado modelo apresenta uma não linearidade intrínsica baixa e uma não linearidade causada pela parametrização também baixa (isto é,

$$\gamma^N < \frac{1}{\sqrt{F}} \quad e \quad \gamma^T < \frac{1}{\sqrt{F}} \quad),$$

então seu comportamento na estimação se aproxima muito do comportanto dos modelos lineares (assim, se considerarmos o modelo (6.1) do capítulo 6, os estimadores de mínimos quadrados $\hat{\underset{\sim}{\theta}}$ serão "quase" não viesados, quase normalmente distribuidos e as variâncias reais dos estimadores serão próximas daquelas obtidas pela matriz de covariâncias assintótica:

$$(F'F)^{-1}\sigma^2$$

onde

$$F = F(\underset{\sim}{\theta}) = \frac{\partial f(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}$$

Por outro lado se a componente intrínsica é baixa ($\gamma^N < \frac{1}{\sqrt{F}}$), mas $\gamma^T > \frac{1}{\sqrt{F}}$, então podemos procurar por reparametrizações no modelo que reduzam a componente devido aos parâmetros.

Além disso se a curvatura intrínsica máxima está muito fora dos limites aceitáveis (isto é, $\gamma^N > \frac{1}{\sqrt{F}}$) então, segundo Ratkowsky (83), existe pouco propósito em procurar por reparametrizações para reduzir a 2ª componente.

Bates & Watts quantificam essas curvaturas usando matrizes de "aceleração" tridimensionais como apresentamos na seção 5.3.2 a seguir.

5.3.2 FÓRMULAS DE CÁLCULO DAS MEDIDAS DE BATES & WATTS

Nessa seção nós apresentamos os passos necessários para o cálculo das medidas de não linearidade de Bates & Watts.

1º Passo:

Seja

$$f(\underset{\sim}{\theta}) = \begin{bmatrix} f_1(\underset{\sim}{\theta}) \\ \vdots \\ f_n(\underset{\sim}{\theta}) \end{bmatrix}$$

onde $f_t(\underset{\sim}{\theta}) = f(\underset{\sim}{x}_t, \underset{\sim}{\theta})$, $t=1,\ldots,n$, a função de regressão.

No 1º passo devemos obter as matrizes V. e V.. onde

$$V. = \left.\frac{\partial f(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}} = \begin{bmatrix} \dot{v}_{11} & \cdots & \dot{v}_{1p} \\ \vdots & & \vdots \\ \dot{v}_{n} & \cdots & \dot{v}_{np} \end{bmatrix}_{n\times p} \qquad (5.3.2.1)$$

onde

$$\dot{v}_{ij} = \left.\frac{\partial f_i(\underset{\sim}{\theta})}{\partial \theta_j}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}}$$

$$V.. = \left.\frac{\partial f(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}\partial \underset{\sim}{\theta}'}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}} \begin{bmatrix} \ddot{\underset{\sim}{v}}_{11} & \cdots & \ddot{\underset{\sim}{v}}_{1p} \\ \vdots & & \vdots \\ \ddot{\underset{\sim}{v}}_{p1} & \cdots & \ddot{\underset{\sim}{v}}_{pp} \end{bmatrix}_{p\times p\times n} \qquad (5.3.2.2)$$

onde

$$\ddot{\underset{\sim}{v}}_{ij} = \begin{bmatrix} \left.\frac{\partial^2 f_1(\underset{\sim}{\theta})}{\partial\theta_i\partial\theta_j}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}} \\ \vdots \\ \left.\frac{\partial^2 f_n(\underset{\sim}{\theta})}{\partial\theta_i\partial\theta_j}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}} \end{bmatrix}_{n\times 1}$$

2º Passo:

Decompor a matriz V. num produto de duas matrizes Q e R (isto é, V.=Q.R), onde Q é uma matriz ortogonal e

$$R = \begin{bmatrix} \tilde{R}_{p\times p} \\ \cdots\cdots \\ \underset{\sim}{0}_{(n-p)\times p} \end{bmatrix}$$

onde $\tilde{R}_{p\times p}$ é triangular superior e inversível (como Q é ortogonal isso é equivalente a encontrar Q e R tais que Q.V=R).

As matrizes Q e R podem ser obtidas a partir da decomposição QR (Businger & Golub, 65) cujo algoritmo é o seguinte:

1ª iteração: Considerar $V^{(1)} = V.$

2ª iteração: Calcular $V^{(2)} = H_1 V.^{(1)} = H_1 V$

3ª iteração: Calcular $V^{(3)} = H_2 V^{(2)} = (H_2 H_1) V$

$\vdots$

$(p+1)^{a}$ iteração: Calcular $V^{(p+1)} = H_p V^{(p)} =$

$= (H_p . H_{p-1} \ldots H_2 H_1) V. = Q.V. = R$

As matrizes H_k são matrizes pxp e são calculadas através do seguinte procedimento

$$H_k = I - \beta_k \, W^{(k)} \, w^{(k)'}$$

onde:

$$\beta_k = \left[\sigma_k \, (\sigma_k + |\dot{v}_{kk}^{(k)}|)\right]^{-1}$$

e

$$\sigma_k = \left(\sum_{i=k}^{n} (\dot{v}_{ik}^{(k)})^2 \right)^{1/2}$$

$$w_i^{(k)} = 0, \text{ se } i<k$$

$$w_i^{(k)} = (\text{sinal de } \dot{v}_{ii}^{(k)}).(\sigma_k + |\dot{v}_{ik}^{(k)}|), \text{ se } i=k$$

$$w_i^{(k)} = \dot{v}_{ik}^{(k)} \text{ , se } i>k$$

3º Passo:

Obter a matriz $U..=L'V..L$ onde $L = \tilde{R}^{-1}_{p \times p}$

Esse procedimento é definido considerando os vetores $\ddot{\underset{\sim}{v}}_{ij}$, como elementos da matriz V.., e portanto U.. é da forma:

$$\begin{bmatrix} \underset{\sim}{U}_{11} & \cdots & \underset{\sim}{U}_{1p} \\ \vdots & & \vdots \\ \underset{\sim}{U}_{p1} & \cdots & \underset{\sim}{U}_{pp} \end{bmatrix}$$

onde $\underset{\sim}{U}_{ij}$ é um vetor nx1.

4º Passo:

Obter a matriz A.. = Q'U.., onde o produto é definido por:

$$Q'U.. = \begin{bmatrix} (Q'\underset{\sim}{U}_{11}) & \cdots & (Q'\underset{\sim}{U}_{1p}) \\ \vdots & & \vdots \\ (Q'\underset{\sim}{U}_{p1}) & \cdots & (Q'\underset{\sim}{U}_{pp}) \end{bmatrix}$$

onde $(Q'\underset{\sim}{U}_{ij})$ é um vetor de dimensão nx1.

Essa matriz é denominada por Bates & Watts de "matriz de aceleração" e ela é da forma:

$$A.. \begin{bmatrix} \begin{bmatrix} a_{11,1} \\ \vdots \\ a_{11,n} \end{bmatrix} & \cdots & \begin{bmatrix} a_{1p,1} \\ \vdots \\ a_{1p,n} \end{bmatrix} \\ \vdots & & \vdots \\ \begin{bmatrix} a_{p1,1} \\ \vdots \\ a_{p1,n} \end{bmatrix} & \cdots & \begin{bmatrix} a_{pp,1} \\ \vdots \\ a_{pp,n} \end{bmatrix} \end{bmatrix}_{pxpxn}$$

onde

$$\begin{bmatrix} a_{11,i} & \cdots & a_{1p,i} \\ a_{21,1} & \cdots & a_{2p,i} \\ \vdots & & \vdots \\ a_{p1,i} & \cdots & a_{pp,i} \end{bmatrix}$$

é denominada "face i" da matriz A.., i=1,...,n.

Seja $A^{T}_{..}$ a matriz constituida das primeiras p faces de A.. e $A^{N}_{..}$, a matriz constituida das últimas n-p faces, ou seja:

$$A^{T}_{..} = \begin{bmatrix} \begin{bmatrix} a_{11,1} \\ \vdots \\ a_{11,p} \end{bmatrix} & \cdots & \begin{bmatrix} a_{1p,1} \\ \vdots \\ a_{1p,p} \end{bmatrix} \\ \vdots & & \vdots \\ \begin{bmatrix} a_{p1,1} \\ \vdots \\ a_{p1,n} \end{bmatrix} & \cdots & \begin{bmatrix} a_{pp,1} \\ \vdots \\ a_{pp,n} \end{bmatrix} \end{bmatrix}_{pxpxp}$$

$$A^{N}_{..} = \begin{bmatrix} \begin{bmatrix} a_{11,p+1} \\ \vdots \\ a_{11,n} \end{bmatrix} & \cdots & \begin{bmatrix} a_{1p,p+1} \\ \vdots \\ a_{1p,n} \end{bmatrix} \\ \vdots & & \vdots \\ \begin{bmatrix} a_{p1,p+1} \\ \vdots \\ a_{p1,n} \end{bmatrix} & \cdots & \begin{bmatrix} a_{pp,p+1} \\ \vdots \\ a_{pp,n} \end{bmatrix} \end{bmatrix}_{pxpx(n-p)}$$

5º Passo:

Obter: $\gamma^N = \max_d \gamma_d^N$ e $\gamma^T = \max_d \gamma_d^T$

onde $\gamma_d^N = \|d' A_{..}^N d\|$ e $\gamma_d^T = \|d' A_{..}^T d\|$ são respectivamente a curvatura intrínsica, e a curvatura devido a parametrização correspondentes a uma particular direção d do espaço paramétrico, e onde $\|d\| = 1$.

Para efetuar esses cálculos em geral não existem fórmulas explícitas, e eles são obtidos numericamente. Bates & Watts (80) propõem o seguinte algoritmo para o cálculo da curvatura intrínsica máxima γ^N:

(i) escolher uma direção inicial d_i.

(ii) calcular

$$g_i = \nabla(\gamma_d^N)^2\Big|_{d_i} = 4(d_i' A_{..}^N d_i)'(A_{..}^N d_i) \quad \text{e}$$

$$\tilde{g}_i = \frac{g_i}{\|g_i\|}$$

(iii) se $\tilde{g}_i' d_i < 1-\varepsilon$, então tomar $d_{i+1} = \tilde{g}_i$ e repetir o procedimento a partir de (ii), caso contrário calcular $\gamma^N = \|d_i' A_{..}^N d_i\|$

Para o cálculo da curvatura máxima causada pela parametrização, basta substituir $A_{..}^N$ por $A_{..}^T$ no mesmo algoritmo.

A idéia desse algoritmo é se mover na direção do gradiente, que é a direção de maior variação da função (Os autores utilizaram ao invés de $\nabla(\gamma_d^N)$, o gradiente do quadrado de γ_d^N, isto é, $\nabla(\gamma_d^N)^2$ já que esses gradientes têm a mesma dire-

ção.

O critério de convergência no passo (iii) verifica se d_i e o gradiente estão aproximadamente na mesma direção (como d_i e $\tilde{g}_i$ são vetores unitários, o produto interno é 1, se eles tiverem exatamente a mesma direção).

Nas aplicações os autores consideram $\varepsilon=0,0001$ e tomam como vetor inicial do processo iterativo, o vetor $d_i^{(0)} = (0,0,\ldots,1)$.

Esse algoritmo, entretanto, segundo Bates & Watts (80), tende a oscilar em torno do ponto ótimo, e por esse motivo eles sugerem a seguinte modificação no passo (iii):

$$d_{i+1} = \frac{3\tilde{g}_i + d_i}{\|3\tilde{g}_i + d_i\|}$$

5.4 - MEDIDA DO VIÉS DE BOX

Box (71) deduziu uma fórmula para medir o viés nos estimadores de máxima verossimilhança dos parâmetros de um modelo de Regressão Não Linear univariada com a suposição de $Var(Y_t) = \sigma^2$. (No caso das observações serem normalmente distribuidas esses resultados também valem para os estimadores de mínimos quadrados).

A fórmula do viés de Box é a seguinte:

$$Viés(\hat{\underset{\sim}{\theta}}) = E(\hat{\underset{\sim}{\theta}}-\underset{\sim}{\theta}) = \frac{-\sigma^2}{2}\left(\sum_{t=1}^{n} \underset{\sim}{E}_t'\underset{\sim}{E}_t\right)^{-1} \sum_{t=1}^{n} \underset{\sim}{E}_t' tr\left(\left(\sum_{t=1}^{n} \underset{\sim}{E}_t'\underset{\sim}{E}_t\right)^{-1} \underset{\sim}{H}_t\right) \quad (5.4.1)$$

onde

$$\underset{\sim}{F}_t' = \begin{bmatrix} \frac{\partial f}{\partial \theta_1}(\underset{\sim}{X}_t, \underset{\sim}{\theta}) \\ \vdots \\ \frac{\partial f}{\partial \theta_p}(\underset{\sim}{X}_t, \underset{\sim}{\theta}) \end{bmatrix}$$

e

$$\underset{\sim}{H}_t = \begin{bmatrix} \frac{\partial^2 f(\underset{\sim}{X}_t, \underset{\sim}{\theta})}{\partial \theta_1^2} & \cdots & \frac{\partial^2 f(\underset{\sim}{X}_t, \underset{\sim}{\theta})}{\partial \theta_1 \partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 f(\underset{\sim}{X}_t, \underset{\sim}{\theta})}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 f(\underset{\sim}{X}_t, \underset{\sim}{\theta})}{\partial \theta_p^2} \end{bmatrix}$$

Nas aplicações $\underset{\sim}{\theta}$ e σ^2 são desconhecidas e as estimativas $\hat{\theta}$ e $\hat{\sigma}^2 = \frac{h(\hat{\theta})}{n-p}$ são utilizadas (onde $h(\hat{\theta})$ é a soma de quadrados dos resíduos).

Bates & Watts (80) demonstraram que o viés de Box está relacionado com a sua medida de não linearidade causada pela parametrização (e portanto o viés pode ser reduzido através de reparametrizações no modelo).

Nas aplicações a importância de utilizar essa fórmula de Box é que ela indica qual (ou quais) dos parâmetros são os maiores responsáveis por um valor alto da não linearidade(causada pela parametrização), de Bates & Watts.

Além da fórmula (5.4.1), Box (71) desenvolveu uma expressão para avaliar o viés dos estimadores de uma nova parametrização em função dos vieses dos estimadores da parametri-

zação antiga. A fórmula é a seguinte:

$$\text{Viés}(\hat{\phi}) = E(\hat{\phi}-\phi) = G'\text{Viés}(\underset{\sim}{\hat{\theta}}) + \frac{1}{2} \text{tr}\,(\underset{\sim}{M}\ \text{cov}(\underset{\sim}{\hat{\theta}}))$$

onde

$$\phi = g(\underset{\sim}{\theta}),\quad G = \left.\frac{\partial g(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right|_{\underset{\sim}{\theta}=\underset{\sim}{\hat{\theta}}},\quad \underset{\sim}{M} = \left.\frac{\partial^2 g(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}\,\partial \underset{\sim}{\theta}'}\right|_{\underset{\sim}{\theta}=\underset{\sim}{\hat{\theta}}},$$

e

$$\text{cov}(\underset{\sim}{\hat{\theta}}) = \sigma^2 (F(\underset{\sim}{\hat{\theta}})'F(\underset{\sim}{\hat{\theta}}))^{-1}$$

onde

$$F(\underset{\sim}{\hat{\theta}}) = \left.\frac{\partial f(\underset{\sim}{X},\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right|_{\underset{\sim}{\theta}=\underset{\sim}{\hat{\theta}}}$$

Essa expressão é importante pois através dela podemos avaliar se a reparametrização sugerida tem ou não valor, sem precisar refazer todos os cálculos.

Box (71) demonstrou também que a variância de $\hat{\phi}$ pode ser obtida através da matriz de covariância assintótica $[F(\theta)'F(\theta)]^{-1}\sigma^2$, através da seguinte fórmula:

$$\text{Var}(\hat{\phi}) = \text{tr}\,((GG')\ \text{cov}(\underset{\sim}{\hat{\theta}}))$$

CAPÍTULO 6

AVALIAÇÃO DAS PROPRIEDADES ESTATÍSTICAS DOS ESTIMADORES DE MÍNIMOS QUADRADOS NÃO LINEARES

6.1 - INTRODUÇÃO

Nesse capítulo nosso principal objetivo é avaliar as propriedades dos estimadores de mínimos quadrados não lineares nas aplicações

Na seção 6.2 nós discutimos as propriedades assintóticas dos estimadores de mínimos quadrados, e na seção 6.3 nós examinamos o comportamento dos estimadores dos parâmetros de um modelo através da utilização de simulação.

6.2 - PROPRIEDADES DOS ESTIMADORES

Consideremos o modelo de regressão:

$$Y_t = f(\underset{\sim}{X}_t, \underset{\sim}{\theta}) + e_t$$

onde $\theta' = (\theta_1, \ldots, \theta_p)$, $X_t = (X_{t1}, \ldots, X_{tk})$, $t=1,\ldots,n$, com a suposição de que $e_t \sim N(0,\sigma^2)$ e são independentes. (6.1)

Quando a função de regressão é linear nos parâmetros, os

estimadores de mínimos quadrados desses parâmetros são não viesados, Normalmente distribuidos e têm variância mínima (independentemente do tamanho da amostra), enquanto que no caso não linear essas propriedades são válidas somente assintoticamente.

É bem conhecido que se as observações são Normalmente distribuidas, os estimadores de mínimos quadrados coincidem com os estimadores de máxima verossimilhança. Portanto sob certas condições de regularidade são assintoticamente não viciados, eficientes e Normalmente distribuidos, isto é,

$$\hat{\underset{\sim}{\theta}} \sim N(\underset{\sim}{\theta}, Var(\hat{\underset{\sim}{\theta}}))$$

onde $Var(\hat{\underset{\sim}{\theta}}) = (F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta}))^{-1}\sigma^2$ e $(F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta}))^{-1}.\sigma^2$ é a inversa de matriz de Informação de Fisher. Nas aplicações $F(\underset{\sim}{\theta})$ é aproximada por

$$F(\hat{\theta}) = \left.\frac{\partial f(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}} ,$$

onde $\hat{\underset{\sim}{\theta}}$ é a estimativa de mínimos quadrados obtida do processo iterativo e σ^2 é aproximado por

$$\hat{\sigma}^2 = \frac{h(\hat{\theta})}{n-p} ,$$

onde $h(\hat{\theta})$ é a soma de quadrados dos resíduos em $\hat{\underset{\sim}{\theta}}$, e portanto uma estimativa para a matriz de covariância assintótica é:

$$\hat{Var}(\hat{\underset{\sim}{\theta}}) = \hat{\sigma}^2(F(\hat{\theta})'F(\hat{\theta}))^{-1}$$

(Segundo Jenrich (69), mesmo quando os resíduos não são Normalmente distribuidos, mas $n \to \infty$, o estimador de mínimos quadrados $\hat{\underset{\sim}{\theta}}$, ainda assim tem aproximadamente uma distribuição

$$N(\underset{\sim}{\theta}, (F(\underset{\sim}{\theta})'F(\underset{\sim}{\theta}))^{-1}\sigma^2)$$

As condições de regularidade e as provas de consistência e normalidade assintóticas nesse caso, são apresentadas em seu artigo. Se os resíduos não forem nem normais nem homocedásticos ainda assim esses resultados se mantém assintoticamente segundo Jenrich (79), e nesse caso

$$\hat{\underset{\sim}{\theta}} \approx N(\underset{\sim}{\theta}, (F(\underset{\sim}{\theta})'W^{-1}F(\underset{\sim}{\theta}))^{-1}\sigma^2) \quad).$$

Esses resultados sobre os modelos não lineares são todos assintóticos. No caso de amostras pequenas as propriedades dos estimadores de mínimos quadrados são desconhecidas. Entretanto nós podemos afirmar que à medida que o tamanho da amostra (n) aumenta, os resultados assintóticos vão se tornando mais aplicáveis.

Na prática qual deve ser o tamanho da amostra necessária para que esses resultados sejam aproximados? Segundo Ratkowsky (83), essa é uma questão complicada já que a validade dessas aproximações depende também do modelo utilizado (existindo modelos em que essas propriedades constituem uma boa aproximação mesmo para amostras consideradas muito peque

nas (por exemplo quando o grau de não linearidade é baixo), enquanto que em outros essas aproximações não são adequadas nem para amostras que poderiam ser consideradas muito grandes na prática.

Quando $\hat{\underset{\sim}{\theta}}$, o estimador de mínimos quadrados de $\underset{\sim}{\theta}$, tem somente um pequeno viés, uma distribuição próxima da Normal, e se as variâncias reais são próximas daquelas dadas pela matriz de covariâncias assintóticas, então dizemos que $\hat{\underset{\sim}{\theta}}$ exibe um comportamento "próximo do linear" (já que o comportamento desses estimadores se aproxima do comportamento dos estimadores de um modelo linear). Nesse caso quanto mais "próximo do linear" for o comportamento de um estimador mais válidos serão os vários testes e intervalos de confiança que fazem analogia com os modelos lineares.

Este é um resultado extremamente importante, e por esse motivo nosso objetivo nesse capítulo é avaliar se o modelo de interesse tem um comportamento próximo do linear, isto é, verificar se as propriedades assintóticas são aplicáveis para o modelo, mesmo em amostras pequenas.

Segundo Ratkowsky (83), o primeiro passo na avaliação dessas propriedades é observar a convergência do algoritmo de Gauss Newton na obtenção das estimativas dos parâmetros. Assim se esse algoritmo, partindo de um ponto inicial arbitrário $\underset{\sim}{\theta}^{(0)}$, convergir rapidamente para um valor $\hat{\underset{\sim}{\theta}}$, e $\hat{\underset{\sim}{\theta}}$ for bem distante de $\underset{\sim}{\theta}^{(0)}$, então isso indica que o modelo tem um comportamento "próximo do linear". Por outro lado se o algo

ritmo apresenta uma convergência lenta mesmo partindo de pontos iniciais próximos da solução, então existem evidências de que o modelo não tenha um comportamento "próximo do linear".

Tendo obtido as estimativas de mínimos quadrados (e a matriz de covariâncias assintóticas), é importante utilizarmos procedimentos sistemáticos para avaliar mais precisamente a não linearidade do modelo, através de medidas de não linearidade como as apresentadas no capítulo 5 (a medida de Box (71), e as medidas de Bates & Watts (80)); além disso podemos também realizar estudos com base em simulação (seção 6.3), (o que segundo Ratkowsky é uma das melhores maneiras de avaliar as propriedades dos estimadores).

Segundo Ratkowsky (83), entretanto, é muito importante combinar essas técnicas, isto é, utilizá-las conjuntamente nas aplicações. Essas análises exigem a utilização de um computador, mesmo para amostras extremamente pequenas. Ratkowsky (83) incorpora esses três procedimentos na sub-rotina NONLIN já implantada no Centro de Computação Eletrônica da USP.

6.3 - AVALIAÇÃO DAS PROPRIEDADES DOS ESTIMADORES ATRAVÉS DO USO DE SIMULAÇÃO

Nessa seção estudaremos o comportamento dos estimado

res de mínimos quadrados através da análise dos resultados obtidos da simulação.

Nesse estudo vamos supor que o modelo de Regressão é do tipo descrito na seção 6.2 (isto é,

$$Y_t = f(\underset{\sim}{X}_t, \underset{\sim}{\theta}) + e_t$$

onde f tem uma forma conhecida e $e_t \sim N(0,\sigma^2)$ e são independentes). Além disso vamos supor que as estimativas de mímimos quadrados dos parâmetros, e a matriz de covariância assintótica tenham sido obtidas a partir de um conjunto de n dados observados.

O interesse nesse estudo é verificar se de fato as propriedades assintóticas dos estimadores de mínimos quadrados constituem uma boa aproximação para o problema, isto é, se o nosso particular modelo tem comportamento "próximo do linear".

Num estudo típico de simulação, a distribuição dos resíduos e_t é especificada e valores $e_1,\ldots,e_n$ são gerados pseudo-aleatoriamente através de uma sub-rotina de computador; além disso um valor para $\underset{\sim}{\theta}$ é fixado, produzindo assim um conjunto de n observações $Y_1, Y_2,\ldots,Y_n$, onde

$$Y_t = f(\underset{\sim}{X}_t, \underset{\sim}{\theta}) + e_t$$

No nosso caso $e_t \sim N(0,\sigma^2)$ e os valores de θ e σ^2 usados para gerar os Y's, são $\hat{\theta}$ (a estimativa de mínimos quadrados) e $\hat{\sigma}^2$

(a variância residual) obtidos do conjunto original de dados.

Simulamos N desses conjuntos (de n observações) e de cada conjunto ajustamos o modelo por mínimos quadrados da maneira usual podendo utilizar como valor inicial do processo iterativo, em cada conjunto, a estimativa $\hat{\theta}$ obtida do conjunto original de dados.

Se ocorrer convergência para cada um desses N conjuntos de dados, teremos N vetores de estimativas do tipo:

$$\underset{\sim}{\hat{\theta}}_k = (\hat{\theta}_{k1}, \hat{\theta}_{k2}, \ldots, \hat{\theta}_{kp}), \quad k=1,\ldots,N$$

A partir dessas estimativas devemos examinar a distribuição conjunta de $\hat{\theta}$. Entretanto a análise de uma distribuição conjunta é tarefa muito complicada, mesmo para o caso mais simples onde p=2, em que devemos examinar se $\hat{\theta}$ tem uma distribuição Normal bivariada.

Ratkowsky (83) utiliza como uma aproximação, os resultados das análises das distribuições marginais de $\underset{\sim}{\hat{\theta}}$ na esperança de que através desse procedimento se possa avaliar aproximadamente, o comportamento conjunto.

A questão do número N de conjuntos que devemos utilizar é arbitrária, entretanto Ratkowsky sugere o valor N=1000, o que considera bastante satisfatório para estudar as propriedades dos estimadores nos modelos de Regressão Não Lineares.

As distribuições marginais podem ser analisadas com base nas N estimativas dos parâmetros correspondentes; tanto atra-

vés de uma análise descritiva, (como por exemplo a construção de histogramas, o cálculo dos momentos amostrais, etc.), como também através de testes como veremos mais adiante.

A tabela 6.3.1, a seguir, apresenta algumas medidas importantes para uma análise das distribuições.

MEDIDAS	FÓRMULAS
(1) Viés (%)	$\frac{\overline{\hat{\theta}}_i - \theta_i}{\theta_i} \cdot 100$
(2) excesso de variância (%)	$\frac{s_i^2 - \overset{*}{\sigma}{}_i^2}{\overset{*}{\sigma}{}_i^2} \times 100$
(3) coeficiente de Skewness	$m_{3i}/\sqrt{m_{2i}^3}$
(4) coeficiente de Kurtósis	m_{4i}/m_{2i}^2

TABELA 6.3.1 - ALGUMAS MEDIDAS ASSOCIADAS À DISTRIBUIÇÃO DE $\hat{\theta}_i$

Uma estimativa do viés de $\hat{\theta}_i$ é dada pela diferença entre a média amostral

$$\overline{\hat{\theta}}_i = \sum_{j=1}^{N} \frac{\hat{\theta}_{ij}}{N} ,$$

e o "verdadeiro" valor do parâmetro (o valor de θ_i usado para gerar os dados). Na 1ª linha da tabela 6.3.1 apresentamos o

viés relativo (como porcentagem do "verdadeiro" valor do parâmetro).

Na 2ª linha da tabela 6.3.1, a fórmula

$$\frac{s_i^2 - \overset{*}{\sigma}_i^2}{\overset{*}{\sigma}_i^2} \cdot 100 ,$$

(onde

$$s_i^2 = \sum_{j=1}^{N} \frac{(\hat{\theta}_{ij} - \bar{\hat{\theta}}_i)^2}{N}$$

é a variância amostral e $\overset{*}{\sigma}_i^2$ é a variância assintótica de $\hat{\theta}_i$ obtida do i-ésimo elemento da diagonal da matriz

$$(F(\hat{\theta})'F(\hat{\theta}))^{-1}\hat{\sigma}^2$$

onde $\hat{\sigma}^2$ é a variância residual obtida do conjunto original de dados), representa a porcentagem na qual s_i^2 excede a variância assintótica.

Nas duas últimas linhas da tabela 6.3.1 apresentamos respectivamente, as fórmulas do coeficinete de Skewness e Kurtósis:

$$g_{1i} = \frac{m_{3i}}{\sqrt{m_{2i}^3}}$$

e

$$g_{2i} = \frac{m_{4i}}{m_{2i}^2}$$

(onde $m_{2i}=s_i^2$, e m_{3i} e m_{4i} representam o terceiro e quarto momento amostral em torno da média).

Além do cálculo dessas medidas nós podemos também realizar testes de hipóteses para verificar se o viés é zero, a variância é mínima, e além disso, se o Skewness e Kurtósis são respectivamente 0 e 3 (os valores típicos da Normal).

Para verificar se o viés é zero podemos utilizar a estatística:

$$Z = \frac{\bar{\theta}_i - \theta_i}{\sqrt{\frac{\overset{*}{\sigma}{}_i^2}{N}}} ,$$

considerando Z Normalmente distribuida com média zero e variância 1.

Para verificar se a suposição de variância mínima é válida, podemos comparar a variância s_i^2 com a variância assintótica obtida, através da estatística:

$$\chi^2 = \frac{(N-1)s_i^2}{\overset{*}{\sigma}{}_i^2}$$

onde χ^2 tem uma distribuição de Qui-Quadrado com N-1 graus de liberdade (como, nos estudos de simulação N é grande, podemos usar a aproximação Normal padrão:

$$Z = \sqrt{2\chi^2} - \sqrt{2(m)-1}$$

onde m=número de graus de liberdade da distribuição de χ^2.

Para testar se o Skewness é zero e a Kurtósis é 3 utilizamos o fato de que g_1 tem aproximadamente uma distribuição $N(0, \frac{6}{N})$ e $g_2 \sim N(3, \frac{24}{N})$ (Snedecor & Cochran (80)).

Se os testes forem todos significativos, então existem evidências de que o estimador de mínimos quadrados não terá um comportamento "próximo do linear". Entretanto em alguns casos podemos reduzir essas diferenças (ou seja diminuir essa não linearidade) através de uma reparametrização; porém a dificuldade que surge é na escolha da reparametrização.

As medidas de não linearidade apresentadas no capítulo 5, apesar de serem úteis para detectar o comportamento não linear, não fornecem nenhuma "pista" neste sentido. Nesses casos o uso de simulação, entretanto, poderá indicar ou sugerir possíveis reparametrizações para o modelo, em algumas situações. Por exemplo se a distribuição de $\hat{\theta}_i$ tem uma cauda mais longa para a direita, então uma possível reparametrização seria

$$\phi_i = \log(\theta_i)$$

Por outro lado se a distribuição tiver uma cauda mais longa para a esquerda uma reparametrização seria

$$\phi_i = e^{\theta_i}$$

REFERÊNCIAS

BATES, D.M. (1979). "Curvature Measures of Nonlinearity". Phd Thesis, Queen's Universdy, Kingston, Canadá.

BATES, D.M. and WATTS, D.G. (1980) - "Relative Curvature Measures of Nonlinearity". J.R. Statist. Soc., Ser. B 42, 1-25.

BARD, Y. (1970) - "Comparison of Gradient Methods for the Solutions of Nonlinear Parameter Estimation". SIAM J.Numer. Anal., 7, 157-186.

BEALE. E.M.L. (1960) - "Confidence Regions in Nonlinear Estimation". J.R. Statis. Soc., Ser B 22, 41-76.

BOX, M.J. (1971) - "Bias in Nonlinear Estimation". J.R. Statist. Soc., B 33, 171-201.

BOX, M.J., DAVIES, D. and SWANN, W.H. (1969) - "Nonlinear Optimization Techniques". Edinburgh: Oliver and Boyd, 60 pp.

BRADLEY, E.L. (1973) - J. Am Stat. Assoc., 68: 199-200.

BROWN, K.M. and DENNIS, J.E. (1971) - "A New Algorition for Nonlinear Least Squares Curve Fitting". in Mathematical Software (Ed. J.R.Rice), Academic Press, New York.

BUSINGER, P. and Golub, G.H. (1965) - "Least Squares by Householder Transformations". Numer. Math. 7, 269-276.

CHAMBERS, J.M. (1973) - "Biometrika". 60: 1-13.

COX, D.R. (1970) - "Analysis of Binary Data". p.86, London: Methuem, 142 pp.

DAHLQUIST, G. and BJÖRCK, A. (1974) - "Numerical Methods". Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

DAVIDON, W.C. (1959) - "Variable Metric Method for Minimization". A.E.C. Research and Development Report, ANL-5990 (Rev.).

DENNIS, J.E., Jr. GAY, D.M. and WELSCH, R.E. (1977) - "An Adaptive Nonlinear Least-Squares Algorithm". Report TR 77-321, Department of Computer Sciences, Cornell University.

DIXON, W.J., BROWN, M.B. eds (1977) - "BMDP Biomedical Computer Programs". p.499-514, Berkley

DIXON, W.J. et al (1985) - "BMDP Statistical Software". University of California Press, Berkley.

BONGARRA, J.J., MOLER, C.B. BUNCH, J.R. and STEWART, G.W. (1979) - "LINPACK Users' Guide". SIAM, Philadelphia.

DRAPER, N.R. and SMITH, H. (1981) - "Applied Regression Analysis". John-Wiley & Sons, Second Edition.

FLETCHER, R. and POWELL, M.J.D. (1963) - "A Rapidly Convergent Descent Method for Minimization". Brit. Computer J., 6, 163-168.

FLETCHER, R. (1980) - "Practical Methods of Optimization". Volume 1, Unconstrained Optimization. John-Wiley & Sons.

FRITZSCHE, H (1978) - "Programação Não Linear - Análise e Métodos". Editora Edgard Blücher Ltda.

GALLANT, A.R. (1975) - "Nonlinear Regression". J.Am.Statist. Assoc., 29, 73-81.

GILL, P.E. and MURRAY, W. (1978) - "Algorithms for the Solution of the Nonlinear Least-Squares Problems". SIAM J. Numer. Anal., 18, pp. 977-992.

GILL, P.E., MURRAY, W. and WRIGHT, M.H. (1981) - "Practical Optimization". Academic Press Inc., London.

GUTTMAN, I. and MEETER, D.A. (1965) - "On Beale's Measures of Non-Linearity". Technometrics 7, 623-637.

HARTLEY, H.O. (1961) - "The Modified Gauss-Newton for the Fitting of Nonlinear Functions by Least-Squares". Technometrics, 3, pp. 269-280.

HARTLEY, H.O. and BOOKER, A. (1965) - "Nonlinear Lead Squares Estimation". The Annals of Mathematical Statistics, 40, pp. 633-643.

JENRICH, R.I. (1969) - "Asymptotic Properties of Nonlinear Least-Squares Estimations". The Annals. of Mathematical Statistics, 20, pp. 633-643.

JENRICH, R.I. and MOORE, R.H. (1975) - "Maximum Likelihood Estimation by Means of Nonlinear Least-Squares". Statistical Computing Section of the Americam Statistical Association, pp. 57-65.

JENRICH, R.I. and RALSTON, M.L. (1979) - "Fitting Nonlinear Models to Data". Ann.Rev.Biophys.Bioeng., 8, pp. 195-238.

JENRICH, R.I. and SAMPSON, P.F. (1968) - "Technometrics". 10, pp. 63-72.

KAPLAN, W. (1971) - "Cálculo Avançado". Vol. 1, Editora Edgard Blücher, Ltda.

KENDALL, M.G. and STUART, A. (1963) - "The Advanced Theory of Statistics". Vol. 1: Distributional Theory, 2nd ed., Charles Griffin, London.

LAWSON, C.L. and HANSON, R.J. (1974) - "Solving Least-Squares Problems". Printice-Hall, Inc., Englewood Cliffs, N.J.

MARQUARDT, D.W. (1963) - SIAM Soc. Ind. Appli. Mathc. J. 11: pp. 431-441.

MEAD, R. (1970) - "Plant Density and Crop Yield". Appl. Statist., 19, pp. 64-81.

Mc KEOWN, J.J. (1980) - "Large Residual Nonlinar Least-Squares Problems in Nonlinear Optimization Theory and Algorithms". edited by Dixon L.C.W., Spedicato E. and Szegö P.G.

MICKEY, M.R. and BRITT, P.M. (1974) - Commun, Stati., 3 pp. 501-511.

RAO, C.R. (1965) - "Linear Statistical Inference and Its Applications". New York: Wiley.

RAO, C.R. (1973) - "Linear Statistical Inference and Its Applications". New York: Wiley.

RATKOWSKY, D.A. (1983) - "Nonlinear Regression Modeling". Marcel Dekker, Inc., New York and Basel.

SNEDECOR, G.W. and COCHRAN, W.G. (1980) - "Statistical Methods". 7th Ed., Iowa State University, Press, Ames., Iowa.

SEBER, G.A.F. (1980) - "The Linear Hypothesis; A General Theory". Charles Griffin & Company Ltd., London.

TURNER, M.E., MONROE, R.S. and LUCAS, H.L. - "Generalized Asymptotic Regression and Nonlinear Data Analysis". Biometrics, 17, pp. 120-143.