RT-MAE 2010-07

ANÁLISE DE INFLUÊNCIA NA REGRESSÃO EM CRISTAS

by

Koki Fernando Oikawa
and
Silvia Nagib Elian

Palavras-Chave: Estimador de Regressão em Cristas, Diagnóstico, análise de Influência Local, Função de Verossimilhança.

Classificação AMS: 62J05.

- Outubro de 2010 -

Análise de Influência na Regressão em Cristas

Koki Fernando Oikawa [1]

Silvia Nagib Elian[2]

[1] Faculdade Capital

[2] Departamento de Estatística
Instituto de Matemática e Estatística,
Universidade de São Paulo
C.P 66281-São Paulo, Brasil
E-mail: selian@ime.usp.br

Resumo

Modelos de Regressão em Cristas apresentam características próprias e problemas específicos. São geralmente utilizados para contornar o problema da multicolinearidade, consequência da existência de relações lineares entre as variáveis explicativas.

O objetivo do presente trabalho é apresentar e discutir as medidas de diagnóstico e a correspondente análise de influência quando é utilizado o procedimento de regressão em cristas.

Apresentaremos inicialmente medidas de influência específicas para o contexto de regressão em cristas. Serão analisadas ainda medidas de influência local neste mesmo contexto. Finalmente, os procedimentos descritos serão aplicados a um conjunto de dados reais.

1.Medidas de Influência na Regressão em Cristas

Consideremos o modelo de regressão linear

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

onde $\mathbf{y}$ é um vetor de variáveis aleatórias observáveis, $\mathbf{1}$ é um vetor contendo o valor 1 em todas as posições, β_0 é um parâmetro desconhecido, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ é uma matriz $n \times p$ centralizada e padronizada de constantes conhecidas $(\mathbf{1}'\mathbf{x}_i = 0,\ \mathbf{x}_i'\mathbf{x}_i = 1,\ i = 1, ..., p)$, $\boldsymbol{\beta}_1$ é um vetor de parâmetros desconhecidos e ε é um vetor de erros não observáveis com $E(\boldsymbol{\varepsilon}) = 0$ e $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Se $\mathbf{Z} = (\mathbf{1}, \mathbf{X})$, o estimador de mínimos quadrados de $\boldsymbol{\beta}$ $[\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')]$ é $\mathbf{b} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$, o vetor de respostas ajustadas fica $\hat{\mathbf{y}} = \mathbf{Z}\mathbf{b}$ e o estimador de σ^2 é $s^2 = \mathbf{e}'\mathbf{e}/(n - p - 1)$, sendo que $\mathbf{e}$ é o vetor de resíduos $(\mathbf{y} - \hat{\mathbf{y}})$.

O estimador em cristas surgiu como uma forma de contornar o problema de multicolinearidade que pode aparecer em dados amostrais.

O problema principal quando se utiliza o estimador de mínimos quadrados na presença de multicolinearidade é que, embora este seja não viciado, sua variância é grande.

Por outro lado, o estimador em cristas é viciado, mas seu erro quadrático médio pode ser menor que o do estimador não viciado $\hat{\boldsymbol{\beta}}$ de mínimos quadrados, devido ao decréscimo na variância (Montgomery e Peck, 1982, pg. 311) .

O estimador em cristas é definido como

$$\mathbf{b}^{*} = \left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^{*}\right)^{-1}\mathbf{Z}'\mathbf{y}$$

onde $\mathbf{I}^{*} = diag(0, 1, ..., 1)$ de dimensão $p+1$

Existem inúmeros critérios para a determinação do valor de k e vários deles estão descritos em Oishi (1983).

A análise de diagnóstico em modelos de regressão quando os parâmetros são estimados pelo procedimento de mínimos quadrados é bastante conhecida. No entanto, para o procedimento de regressão em cristas, a literatura não se mostra tão rica. Apresentaremos a seguir algumas medidas de diagnóstico para esse caso, extraídas do trabalho de Walker e Birch (1988).

Ao utilizarmos o estimador em cristas, o vetor de valores ajustados será

$$\begin{aligned}\hat{\mathbf{y}}^{*} &= \mathbf{Z}\mathbf{b}^{*} \\ &= \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^{*}\right)^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned}$$

Portanto, a matriz $\mathbf{H}^{*} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^{*}\right)^{-1}\mathbf{Z}'$ assume uma função similar à da matriz "hat" na estimação por mínimos quadrados. O *i-ésimo* valor previsto pode ser escrito em termos dos elementos de $\mathbf{H}^{*}$ como

$$\hat{y}_i^{*} = \sum_{j=1}^{n} h_{ij}^{*} y_j .$$

Consequentemente, $\partial\hat{y}_i^{*} / \partial y_i = h_{ii}^{*} \equiv h_i^{*}$ e com isso, os elementos da diagonal da matriz "hat" do estimador em cristas podem ser interpretados, assim como no caso de mínimos quadrados, como um valor de alavancagem.

Uma versão alternativa para a distância de Cook adaptada também ao contexto de regressão em cristas é dada pela expressão

$$D_i^* = \left(1/\left((p+1)s^2\right)\right)\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)' \mathbf{Z}'\mathbf{Z}\left(\mathbf{b}^* - \mathbf{b}^*(i)\right),$$

em que $\mathbf{b}^*(i)$ é o estimador em cristas calculado sem a *i-ésima* observação.

A medida D_i^* também pode ser escrito como

$$D_i^* = \left(1/(p+1)s^2\right)\left(\hat{\mathbf{y}}^* - \hat{\mathbf{y}}^*(i)\right)\left(\hat{\mathbf{y}}^* - \hat{\mathbf{y}}^*(i)\right),$$

sendo que $\hat{\mathbf{y}}^* = \mathbf{Z}\mathbf{b}^*(i)$.

2 -Análise de Influência Local na Regressão em Cristas

O método da influência local foi desenvolvido por Cook (1986) e é aplicável apenas em procedimentos de estimação via função de verossimilhança.

Seja $L(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança para um modelo inicial, sendo $\boldsymbol{\theta}$ um vetor $p \times 1$ de parâmetros desconhecidos com estimador de máxima verossimilhança dado por $\hat{\boldsymbol{\theta}}$.

São introduzidos distúrbios no modelo através do vetor $\mathbf{w}$, $m \times 1$, $\mathbf{w} \subset \Omega$, $\Omega \subset \Re^m$, onde Ω representa um conjunto aberto de possíveis pequenos distúrbios. Do ponto de vista prático, $\mathbf{w}$ refletiria qualquer esquema de perturbação, por exemplo, nas variáveis explicativas ou na matriz de covariâncias da variável resposta do modelo de regressão.

Seja $L(\boldsymbol{\theta} \mid \mathbf{w})$ o logaritmo da função de verossimilhança que corresponde ao modelo que sofreu perturbação e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ o estimador de máxima verossimilhança correspondente a esse modelo. Supondo que exista um ponto $\mathbf{w}_0$ em $\boldsymbol{\Omega}$ que representa a ausência de perturbação nos dados, de modo que $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{w}_0)$, e assumindo que $L(\boldsymbol{\theta} \mid \mathbf{w})$ seja duplamente diferenciável e contínua em uma vizinhança de $(\hat{\boldsymbol{\theta}}', \mathbf{w}'_0)$, o deslocamento de verossimilhanças de Cook é definido como

$$LD(\mathbf{w}) = 2\left[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_{\mathbf{w}})\right],$$

e compara as estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$, podendo, assim, avaliar a influência dos distúrbios $\mathbf{w}$. Grandes valores de $LD(\mathbf{w})$ indicam que $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ diferem consideravelmente em relação ao contorno da função de verossimilhança sem perturbação $L(\boldsymbol{\theta})$.

Esse método é baseado no estudo do comportamento local de um gráfico de influência $\alpha(\mathbf{w}) = (\mathbf{w}', LD(\mathbf{w}))$ ao redor de $\mathbf{w}_0$. O procedimento consiste em considerar $\mathbf{w}$ como $\mathbf{w}(a) = \mathbf{w}_0 + a\mathbf{d}$, $a \in \Re$ e $\mathbf{d}$ um vetor direção de comprimento unitário.

Cook (1986) sugere investigar a direção na qual a medida de influência $LD(\mathbf{w})$ muda localmente mais rapidamente, que é a curvatura máxima de LD, dada por

$$C_{\max} = \max_{|\mathbf{d}|=1} 2\left|\mathbf{d}'\mathbf{F}\mathbf{d}\right|,$$

em que $\mathbf{F}$ é uma matriz $m \times m$ definida por

$$\mathbf{F} = \Delta' \mathbf{Q}^{-1} \Delta,$$

Δ é a matriz $p \times m$ $(p = \dim(\boldsymbol{\theta}), m = \dim(\mathbf{w}))$ com elementos

$$\Delta_{ij} = \frac{\partial^2 L(\boldsymbol{\theta} \mid \mathbf{w})}{\partial \boldsymbol{\theta}_i \partial \mathbf{w}_j},$$

avaliados em $\hat{\boldsymbol{\theta}}$ e $\mathbf{w}_0$, e $-\mathbf{Q}$ representa a matriz de informação observada do modelo sem distúrbios $\mathbf{Q} = \left[\partial^2 L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j\right]$, avaliada em $\hat{\boldsymbol{\theta}}$. Verifica-se que a maximização de $|\mathbf{d}'\mathbf{F}\mathbf{d}|$, sujeito à restrição que $\mathbf{d}'\mathbf{d} = 1$, resulta em $\mathbf{d}_{max}$, que representa o autovetor correspondente ao maior autovalor absoluto C_{max} de $\mathbf{F}$. A direção do vetor $\mathbf{d}_{max}$ seria aquela que produziria a maior mudança local nas estimativas dos parâmetros.

Cook (1986) sugere como referência geral uma curvatura igual a 2, sendo que curvaturas maiores que esse valor indicariam notável sensibilidade local.

Billor e Loynes (1999) propuseram ainda uma medida alternativa, descrita por

$$LD^*(\mathbf{w}) = -2\left[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_w \mid \mathbf{w})\right].$$

A quantidade $LD^*(\mathbf{w})$ compararia então as funções de verossimilhança das duas situações consideradas, com e sem perturbação. Para $m \geq 2$, os autores sugerem o uso da medida

$$l_{max} = \left|\nabla LD^*(\mathbf{w}_0)\right|,$$

onde $\nabla LD^*(\mathbf{w}_0)$ é o vetor gradiente da função LD^* em $\mathbf{w}_0$

Para o cálculo das medidas de influência, de acordo com essa abordagem, os autores escrevem o estimador em cristas como um estimador de pseudo-máxima verossimilhança.

Para tal fim, consideraram um modelo de regressão linear múltipla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (2.1)$$

onde $\mathbf{X}$ é uma matriz conhecida $n \times p$ padronizada, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros conhecidos, $\boldsymbol{\varepsilon}$ é o vetor de erros $p \times 1$ independentes e com distribuição normal com média zero e variância desconhecida σ^2. Admitiu-se adicionalmente que, nesse modelo, o termo constante não foi incluído.

Marquardt (1970) demonstrou que o estimador em cristas é equivalente ao estimador de mínimos quadrados quando os dados são suplementados por um conjunto de dados fictícios tomados de acordo com a matriz de planejamento ortogonal Hk e a variável resposta $\mathbf{Y}$ sendo zero em cada ponto fictício adicionado.

O modelo aumentado com matriz de planejamento $(n+p)\times p$

$$\mathbf{X}_a = \begin{pmatrix} \mathbf{X} \\ (k\mathbf{I})^{1/2} \end{pmatrix}$$

e o vetor $(n+p)\times 1$ de variáveis resposta $\mathbf{Y}_a' = (\mathbf{Y}'\ 0')$ pode ser escrito como

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

onde $\boldsymbol{\varepsilon}_a$ representa um vetor aleatório cujas componentes são variáveis aleatórias independentes e normalmente distribuídas com média zero e variância σ^2. A função

densidade de $\mathbf{Y}_a$ será denominada função pseudo-densidade e a correspondente função de pseudo-verossimilhança será descrita por

$$L_p(\beta) = \frac{n+p}{2}\log 2\pi - \frac{n+p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2 + k\beta'\beta\right].$$

O estimador de máxima pseudo-verossimilhança é resultante de $\frac{\partial L_p(\beta)}{\partial \beta} = 0$.

Como

$$\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\beta\right)^2 + k\beta'\beta$$

pode ser escrito na forma

$$\begin{aligned}&(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + k\beta'\beta \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\beta,\end{aligned}$$

derivando-se essa expressão com relação a β e igualando a zero, obtem-se

$$2[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\beta - 2\mathbf{X}'\mathbf{y} = 0.$$

Resolvendo essa equação, em decorrência da utilização da matriz aumentada, obtêm-se a solução $\hat{\beta}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, que é o estimador em cristas. Uma vez que o estimador em cristas é o estimador de máxima pseudo-verossimilhança para o modelo considerado, a medida de influência local de Cook pode ser aplicada na regressão em cristas.

Considerando o modelo originalmente descrito em (2.1), que supõe homogeneidade na variância do erro, ou seja, $\text{var}(\varepsilon) = \sigma^2\mathbf{I}$ e que pequenos distúrbios

são introduzidos na variância de ε_i por meio de um vetor de distúrbios $\mathbf{w}$, $n \times 1$, onde σ^2 é suposto conhecido, a função de pseudo-verossimilhança com distúrbios para o modelo aumentado é

$$L_p(\boldsymbol{\beta}) = cons\tan te - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \mathbf{w}_i^* + k\boldsymbol{\beta}'\boldsymbol{\beta}\right] + \frac{1}{2}\sum_{i=1}^{n}\log \mathbf{w}_i^*$$

onde $\mathbf{w}_i^* = 1 + \mathbf{w}_i$, $\mathbf{w}_i$ sendo o *i-ésimo* componente do vetor $n \times 1$ de distúrbios $\mathbf{w}$.

Para o cálculo da curvatura máxima C_{max} é necessária a obtenção dos componentes individuais da matriz da informação observada $-\mathbf{Q}$ e da matriz

$$\boldsymbol{\Delta} = \left[\frac{\partial^2 L_p(\boldsymbol{\beta} \mid \mathbf{w})}{\partial \boldsymbol{\beta}_i \partial \mathbf{w}_j}\right],$$

avaliados em $\hat{\boldsymbol{\beta}}$ e $\mathbf{w}_0$.

Nessa situação, as matrizes $\mathbf{Q}$ e $\boldsymbol{\Delta}$ são dadas por

$$-\mathbf{Q} = \frac{(\mathbf{X}'\mathbf{X} + k\mathbf{I})}{\sigma^2}$$

$$\boldsymbol{\Delta} = \frac{\mathbf{X}'\mathbf{D}(\mathbf{e}^*)}{\sigma^2}$$

onde $\mathbf{e}^*$ é o vetor de resíduos em cristas, isto é, $\mathbf{e}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*$, $\hat{\boldsymbol{\beta}}^*$ é o estimador em cristas e $\mathbf{D}(\mathbf{e}^*) = diag(\mathbf{e}_1^*, ..., \mathbf{e}_n^*)$. A curvatura é obtida como:

$$C_d = 2\left|\mathbf{d}'\mathbf{F}\mathbf{d}\right|$$
$$= 2\left|\mathbf{d}'\Delta'\mathbf{Q}^{-1}\Delta\mathbf{d}\right|$$
$$= \frac{2\left|\mathbf{d}'\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*)\mathbf{d}\right|}{\sigma^2}.$$

Após cálculos, verifica-se que a curvatura máxima de LD é dada por

$$C_{max} = \frac{2\lambda^*_{max}}{\sigma^2}$$

onde λ^*_{max} é o maior autovalor de

$$\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*).$$

Já, para LD^*, a máxima inclinação será dada por

$$\left|\nabla LD^*_{\mathfrak{R}}\right| = \left[\sum_{i=1}^{n}\left(1-\frac{(e_i^*)^2}{\sigma^2}\right)^2\right]^{1/2}.$$

Com relação a uma interpretação adequada dessas medidas, Cook (1986) sugere que $C_{max} = 2$ pode ser usado como um valor limite. Contudo, Billor e Loynes (1993) apontaram o valor $\sqrt{2n+4(14n)^{1/2}}$ como relevante na determinação de influência local.

3-Aplicação

As técnicas de análise apresentadas foram aplicadas ao conjunto de dados do projeto: Relação Estrutura-Atividade de Anestésicos Locais N,N [Dimetilamina] Etil Benzoatos Para-Substituídos, (André, Elian e Bruscato, 1997) em Oikawa(2008). O projeto é da área farmacológica e investiga o efeito de diversos tipos de anestésicos

locais sobre o coração de ratos. O interesse desse estudo consistiu em verificar quais características físico-químicas da molécula de determinada droga influenciam mais em sua potência tóxica, definida como a dose de droga necessária para ocorrer uma redução de 30% na freqüência do átrio. Para tal, foram utilizados setenta e dois ratos, homogêneos entre si, divididos em quatorze grupos, contendo de três a oito ratos. Cada grupo foi submetido a uma droga diferente e a potência tóxica calculada após a realização de um experimento, descrito no referido trabalho.

Foram consideradas as variáveis independentes

- **B4**: largura do comprimento substituinte a partir do eixo da ligação, perpendiculares a ele (medida em Ângstron).
- **F**: componente de campo (adimensional);
- **R**: componente de ressonância (adimensional);
- **SIGMA**: constante de Hammet – combinação linear das duas anteriores (adimensional);
- **LOG PAPP**: logaritmo do coeficiente de partição óleo-água medido (adimensional);

e a variável resposta é dada por:

- **POTÊNCIA**: $-\log(DE_{30})$, onde DE_{30} é a dose de droga necessária para ocorrer uma redução de 30% na freqüência do átrio em relação ao controle (adimensional).

Na análise da relação entre a variável resposta e as variáveis independentes utilizou-se um modelo de regressão linear múltipla. No entanto, foi detectada a presença de multicolinearidade através do cálculo do Fator de Inflação da Variância e

do número condicional, que é obtido pela razão $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$, onde λ_{max} é o maior autovalor da matriz $(\mathbf{X'X})$, na sua forma de correlação, enquanto que λ_{min} é o menor autovalor dessa matriz. Os autovalores da matriz $(\mathbf{X'X})$ obtidos foram: 3,1756; 1,0058; 0,6851; 0,1267 e 0,0066. $\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{3,1756}{0,0066} = 481,15$, o que sugere existência de forte multicolinearidade nos dados.

Como forma de contornar o problema da multicolinearidade, um modelo de regressão em cristas foi ajustado. Para isso, foi utilizado o traço como critério de escolha para o valor de k, com k variando de zero a dois.

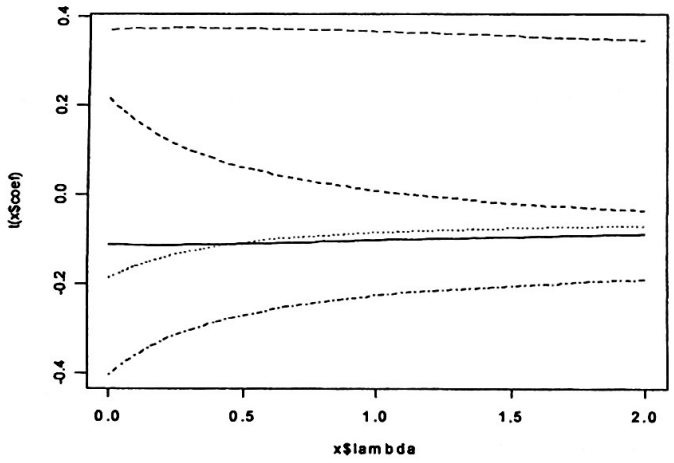

Figura 3.1 – Traço das estimativas dos coeficientes de regressão das variáveis: B4, SIGMA. F, R e LOG.PAPP

Através da Figura 3.1, percebe-se que a partir de $k = 1$ os coeficientes tendem a se estabilizar. Assim, esse valor foi escolhido obtendo-se o modelo de regressão em cristas

$$\hat{Y} = 3{,}27 - 0{,}07 \cdot \text{B4} + 0{,}02 \cdot \text{SIGMA} - 0{,}44 \cdot \text{F} - 0{,}81 \cdot \text{R} + 0{,}30 \cdot \text{LOG.PAPP}$$

Posteriormente foram aplicadas algumas das técnicas de diagnóstico descritas com o auxílio de programas desenvolvidos no pacote computacional R.

Calculando-se os elementos da diagonal principal da matriz $\mathbf{H}^* = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'$ verificou-se que as observações 64, 65 e 66 eram as mais influentes, com valor $h_i^* = 0{,}172$ (Figura 3.2). Correspondiam a três ratos com os maiores valores de SIGMA (0,72) e foram os únicos a apresentarem valores negativos em LOG.PAPP (-0,70). Apresentavam também os maiores valores da variável F (0,54) e da variável R (0,22).

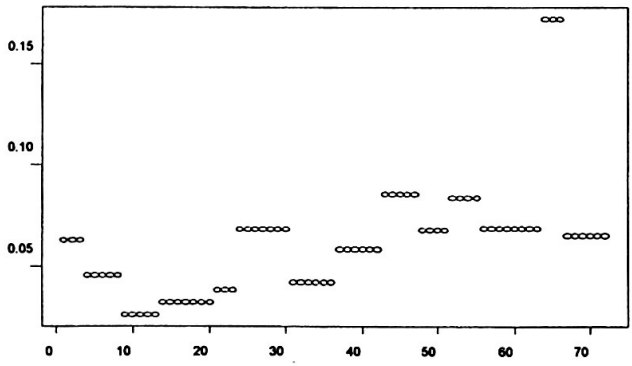

Figura 3.2 - Valores da diagonal principal da matriz H*

Não foram detectados pontos influentes por meio da medida D_i^*. Se o procedimento adotado fosse o de mínimos quadrados, sete seriam as observações influentes: 21, 44, 64, 65, 69, 70 e 71.

A curvatura máxima obtida foi $C_{max} = \frac{2 \cdot \lambda_{max}^*}{\hat{\sigma}^2} = \frac{2 \cdot 0{,}0525435}{0{,}03663602} = 2{,}87$. Dessa forma, podemos concluir por uma sensibilidade moderada nos dados, de acordo com o critério de Cook $(C_{max} > 2)$.

O autovetor associado a λ_{max}^* também fornece informação sobre a influência dos pontos,de modo que as coordenadas com maiores valores correspondem aos pontos mais influentes. Segundo esse critério, foram detectados os ratos de números: 44, 43, 46, 49, 21, 48 e 45, todos com componente de λ_{max}^* maiores que $|0{,}2|$, como pode ser verificado na Figura 3.3.

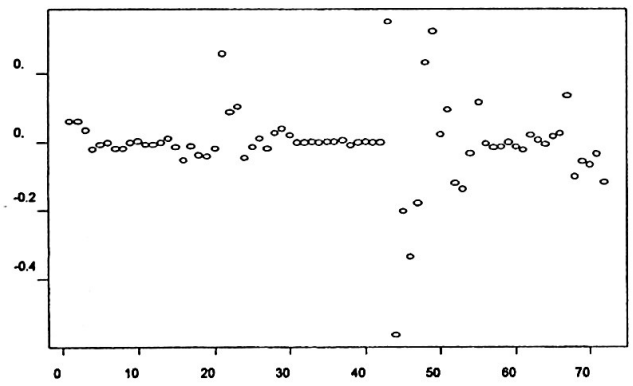

Figura3.3- Análise da Influência pelas componentes do autovetor associado a λ_{max}^*

A inclinação máxima l_{max} obtida foi

$$\left|\nabla LD^{\bullet}_{\gamma_1}\right| = \left[\sum_{i=1}^{n}\left(1-\frac{(e_i^{\bullet})^2}{\hat{\sigma}^2}\right)^2\right]^{1/2},$$

$$\left|\nabla LD^{\bullet}_{\gamma_1}\right| = 13{,}53.$$

Assim, como $l_{max} = 13{,}53 < 16{,}46 = \sqrt{2n+4(14n)^{1/2}}$, esta medida não sugere sensibilidade local para os dados . Os valores absolutos individuais de l_i , em que $l_i = \left(1-\frac{(e_i^{\bullet})^2}{\hat{\sigma}^2}\right)$, encontram-se na Figura 3.4 e Tabela 3.1.

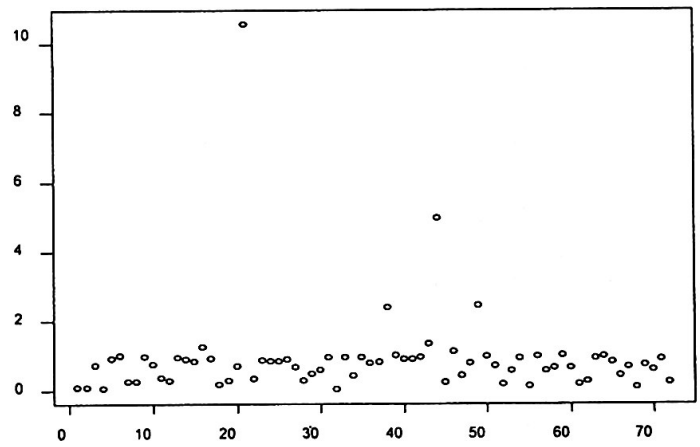

Figura 3.4 – Valores absolutos de l_i

Tabela 3.1– Valores absolutos individuais de l_i

Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$
1	0,10	16	1,25	31	0,94	46	1,10	61	0,16
2	0,10	17	0,91	32	0,04	47	0,41	62	0,27
3	0,72	18	0,17	33	0,95	48	0,77	63	0,92
4	0,05	19	0,29	34	0,41	49	2,43	64	0,97
5	0,91	20	0,69	35	0,95	50	0,98	65	0,80
6	0,99	21	10,5	36	0,76	51	0,70	66	0,42
7	0,24	22	0,32	37	0,81	52	0,18	67	0,68
8	0,24	23	0,85	38	2,39	53	0,54	68	0,10
9	0,97	24	0,82	39	1,00	54	0,91	69	0,72
10	0,74	25	0,83	40	0,88	55	0,12	70	0,59
11	0,37	26	0,89	41	0,89	56	0,96	71	0,90
12	0,28	27	0,68	42	0,95	57	0,57	72	0,23
13	0,95	28	0,29	43	1,34	58	0,63		
14	0,89	29	0,46	44	4,96	59	0,99		
15	0,84	30	0,60	45	0,24	60	0,63		

Com base neles, detectamos quatro observações apresentando valores perceptivelmente maiores que as demais: 21, 38, 44 e 49, sendo que as observações 21 e 44 já haviam sido detectadas pelo método de mínimos quadrados e também pelas componentes de λ^{*}_{max} Dessa maneira, a análise de diagnóstico mostrou-se plenamente satisfatória pois todos os pontos diagnosticados correspondiam a elementos atípicos, o que evidenciou a extrema importância das técnicas utilizadas.

Referências:

André, C.D.S.; Elian, S.N.; Bruscato, A.(1997) *Relatório de Análise Estatística sobre o projeto: "Relação Estrutura- Atividade de Anestésicos locais N,N [dimetilamina] etilBenzoatos parasubstituídos"*. São Paulo, IME-USP, 38p.

Billor, N. and Loynes, R. M. (1993). "Local Influence: A New Approach", *Communications in Statistics – Theory and Methods*, 22, pp. 1595-1611.

Cook, R. D. (1986). "Assessment of Local Influence (wit discussion)". *Journal of the Royal Statistical Society*, Series B, 48, pp. 133-169.

Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation". *Technometrics*, 12, 591-612.

Montgomery, D. C. and Peck, E. A. (1982). "Introduction to Linear Regression Analysis", New York: John Wiley.

Oishi, J. (1983). "Regressão Sobre Cristas". *Dissertação de Mestrado*. IME-USP.

Oikawa, K. F. (2008). "Análise de Influência na Regressão em Cristas". *Dissertação de Mestrado*. IME-USP.

Walker, E. and Birch, J. B. (1988). "Influence Measures in Ridge Regression". *Technometrics*, 30, 221-227.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2010-01 - BUENO, V.C. Component importance for a coherent system under a hyperminimal distribution. 09p. (RT-2010-01)

2010-02 - ABADI, M., ESTEVES, L.G., SIMONIS, A. Pointwise approximationos for sums of non-identically distributed Bernoulli trials. 12p. (RT-MAE-2010-02)

2010-03 - BUENO, V.C. A Series Representation of a Coherent System. 16p. (RT-MAE-2010-03)

2010-04 - POLETO, F.Z., PAULINO, C.D., MOLENBERGHS, G. SINGER, J.M. *Inferential implications of over-parameterization: a case study in incomplete categorical data. 33p. (RT-MAE-2010-04)*

2010-05 - TSUNEMI, M.H., ESTEVES, L.G., LEITE, J.G., WECHSLER, S. A Bayesian nonparametric model for Taguchi's on-line quality monitoring procedure for attributes. 23P. (RT-MAE-2010-05)

2010-06 - BUENO, V.C. Dynamic signatures of a coherent system. 18p. (RT-MAE-2010-06)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05314-970 - São Paulo, Brasil