DIAGNÓSTICO EM REGRESSÃO NORMAL LINEAR: PRINCÍPIOS E APLICAÇÃO

Artur José LEMONTE[1]

▪ RESUMO: Modelos estatísticos são extremamente usados para extrair e conhecer as características de um conjunto de dados. Modelos, entretanto, são sempre descrições aproximadas de um processo mais complicado. Portanto, considerações da adequacidade de um modelo são extremamente importantes. Apresentamos neste artigo algumas técnicas de diagnóstico em modelos de regressão normais lineares que, de modo geral, podem ser vistas como métodos para estudar a variação na análise dos resultados quando o modelo em estudo é submetido a algum tipo de perturbação. Adicionalmente, apresentamos em detalhes o método de influência local desenvolvido por Cook (1986).

▪ PALAVRAS-CHAVE: Influência local; medidas de diagnóstico; modelo normal linear.

1 Introdução

Análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis. A análise de dados através desta classe de modelos é uma das técnicas mais usadas, existindo uma ampla bibliografia sobre o assunto, por exemplo: Searle (1971), Weisberg (1985), Draper & Smith (1998), Montgomery et al. (2001), entre outros. Em geral, na maioria dos problemas estatísticos que surgem nas áreas de agricultura, ciência política, demografia, ecologia, economia, engenharia, geografia, geologia, história, medicina, psicologia, sociologia, zootecnia, etc, podem ser formulados como modelos de regressão. Deste modo, devido a sua ampla aplicabilidade, a análise de regressão é uma técnica estatística de extrema importância.

Quando as hipóteses usuais estabelecidas para o modelo de regressão linear são verificadas, a estimação dos seus parâmetros é geralmente feita através do método de mínimos quadrados ordinários. Este método possui algumas propriedades

[1]Departamento de Estatística, Universidade de São Paulo – USP, Rua do Matão, 1010, CEP: 05508-090, São Paulo, SP, Brasil. E-mail: *arturlemonte@gmail.com*

estatísticas muito atraentes que fizeram dele um dos mais poderosos e populares métodos de análise de regressão (Gujarati, 2000). Este método é atribuído a Carl Friedrich Gauss, matemático alemão.

Segundo Paula (2004), uma etapa extremamente importante na análise de um ajuste de regressão é a verificação de possíveis afastamentos das suposições feitas para o modelo, especialmente para a parte aleatória e para a parte sistemática, bem como a existência de observações extremas com alguma interferência desproporcional nos resultados do ajuste. Tal etapa, conhecida como *análise de diagnóstico*, tem longa data e iniciou-se com a análise de resíduos para detectar a presença de pontos extremos e avaliar a adequação da distribuição proposta para a variável resposta.

De modo geral, os principais objetivos dos métodos de diagnósticos são: (1) Verificar se há afastamentos significativos das suposições feitas para o modelo. Por exemplo: se os erros não são normalmente distribuidos; se a variância não é constante (heteroscedasticidade); etc. (2) Identificar observações que destoam do conjunto de dados. Tais observações podem ser classificadas em três grupos: (*i*) alavanca: posicionadas em regiões remotas com alta influência no próprio valor ajustado; (*ii*) influentes: com influência desproporcional nas estimativas dos coeficientes; (*iii*) aberrantes: mal ajustadas com resíduo alto. Vale ressaltar que uma observação pode ser classificada em mais de um grupo (Paula, 2004).

O objetivo principal deste artigo é apresentar algumas técnicas de diagnóstico em modelos de regressão normais lineares. Serão apresentadas algumas medidas de influência que são bastante utilizadas na prática, em particular, a distância de Cook. Adicionalmente, será discutido em detalhes o método de influência local proposto em Cook (1986).

Considere o modelo de regressão da forma

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

em que $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ é um vetor n-dimensional representando a variável resposta; $\mathbf{X} = (\mathbf{x}_1\, \mathbf{x}_2 \cdots \mathbf{x}_n)^\top$ é a matriz modelo de dimensão $n \times p$ $(p < n)$, sendo de posto completo, $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ representa sua i-ésima linha, com $i = 1, 2, \ldots, n$ (n é o tamanho da amostra); $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor p-dimensional de parâmetros desconhecidos e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ é um vetor n-dimensional de variáveis aleatórias, em que $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ e $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$. Aqui, estamos assumindo que $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é dado pela equação clássica $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

2 Medidas de diagnóstico

Na prática, pode acontecer que após uma escolha cuidadosa de um modelo e subseqüente ajuste a um conjunto de dados, o resultado obtido seja insatisfatório. Isso pode ocorrer em função de algum desvio sistemático entre valores observados e valores ajustados ou, então, porque um ou mais valores são discrepantes em relação

aos demais. Discrepâncias isoladas podem ocorrer ou porque os pontos estão nos extremos da amplitude de validade da variável regressora, ou porque eles estão realmente errados como resultado de uma leitura errada ou uma transcrição mal feita.

A seguir, algumas medidas de diagnóstico serão apresentadas. Basicamente, o conteúdo apresentado nas Seções 2.1, 2.2 e 2.3.1 segue de Paula (2004). Adicionalmente, grande parte do que está descrito na Seção 2.3.2 foi retirado de Souza (1999).

2.1 Pontos de alavanca

O resíduo ordinário para a i-ésima observação é dado por $e_i = y_i - \widehat{y}_i$, em que e_i mede a discrepância entre o i-ésimo valor observado e o i-ésimo valor ajustado. O sinal de e_i indica a direção dessa discrepância. Seja o vetor de resíduos ordinários definido por $\mathbf{e} = (e_1, e_2, \ldots, e_n)^\top$. Note que,

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{y} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y},$$

em que $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ é a matriz de projeção ortogonal de vetores do $\mathbb{R}^n$ no subespaço gerado pelas colunas da matriz $\mathbf{X}$. $\mathbf{H}$ é conhecida como "matriz chapél", uma vez que $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, isto é, ela coloca um "chapél" em $\mathbf{y}$ (Hoaglin & Welsch, 1978). Os elementos de $\mathbf{H}$ vão ser denotados por h_{ij}.

Observe que $\mathbf{H}$ é simétrica, isto é, $\mathbf{H}^\top = \mathbf{H}$, assim, $h_{ij} = h_{ji}$. $\mathbf{H}$ é também idempotente, ou seja, $\mathbf{HH} = \mathbf{H}$. Se k é qualquer potência inteira, $\mathbf{H}^k = \mathbf{H}$ (Draper & Smith, 1998). Como $\mathbf{H}$ é simétrica e idempotente, tem-se que

$$\text{posto}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \text{tr}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] = \text{tr}[\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}] = \text{tr}(\mathbf{I}_p) = p,$$

em que $\text{tr}(\cdot)$ representa o operador traço de uma matriz. O elemento $h_{ii} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$ desempenha um papel importante na construção de técnicas de diagnóstico. Mostra-se que $1/n \leq h_{ii} \leq 1/c$ (Cook & Weisberg, 1982), em que c é o número de linhas de $\mathbf{X}$ idênticas a $\mathbf{x}_i^\top$. Como $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, o i-ésimo valor ajustado pode ser escrito na forma

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j, \tag{2}$$

e pelo fato de $\mathbf{H}$ ser idempotente, $\sum_{j\neq i} h_{ij} = h_{ii}(1-h_{ii})$. Se $h_{ii} = 1$, temos que $\widehat{y}_i = y_i$, entretanto, a recíproca não é necessariamente verdadeira. Logo, para valores altos de h_{ii}, predomina na expressão (2) a influência de y_i sobre o correspondente valor ajustado. Assim, é muito razoável utilizar h_{ii} como uma medida de influência da i-ésima observação sobre o próprio valor ajustado. Note também que $h_{ii} = \partial\widehat{y}_i/\partial y_i$, ou seja, h_{ii} corresponde à variação em $\widehat{y}_i$ quando y_i é acrescido de um infinitésimo.

Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados, espera-se que h_{ii} esteja próximo de $\text{tr}(\mathbf{H})/n = p/n$. Convém então

examinar aqueles pontos tais que, segundo Belsley et al. (1980), $h_{ii} \geq 2p/n$, sendo conhecidos como pontos de alavanca ou de alto *leverage* (isto é, pontos que têm uma influência desproporcional no próprio valor ajustado) e, geralmente, estão localizados em regiões remotas no subespaço gerado pelas colunas da matriz $\mathbf{X}$.

2.2 Pontos aberrantes

Vimos na seção anterior que $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$. Assim, segue que $\mathrm{E}[(\mathbf{I}_n - \mathbf{H})\mathbf{y}] = (\mathbf{I}_n - \mathbf{H})\mathrm{E}(\mathbf{y}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ e $\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$. Conseqüentemente, $e_i \sim \mathcal{N}\big(0, \sigma^2(1 - h_{ii})\big)$. Adicionalmente, $\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, com $i \neq j$. Observe que os e_i's possuem variâncias distintas, dessa forma, é conveniente expressá-los de uma forma padronizada a fim de permitir uma comparabilidade entre os mesmos. Uma definição natural seria dividir e_i pelo seu respectivo desvio padrão, obtendo-se o resíduo studentizado

$$t_i = \frac{e_i}{s(1 - h_{ii})^{1/2}} = \frac{y_i - \widehat{y}_i}{s(1 - h_{ii})^{1/2}},$$

em que $s^2 = \sum_{i=1}^{n} e_i^2/(n-p)$, $i = 1, 2, \ldots, n$.

Observe que e_i não é independente de s^2. Portanto, t_i não segue distribuição t de Student como se poderia esperar. Tal problema pode ser contornado substituindo s^2 por $s^2_{(i)}$, em que $s^2_{(i)}$[1] é a variância correspondente ao modelo sem a i-ésima observação. Utilizando o resultado bastante conhecido da regressão normal linear,

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{e_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i}{(1 - h_{ii})},$$

é possível mostrar que

$$s^2_{(i)} = s^2\left(\frac{n - p - t_i^2}{n - p - 1}\right). \tag{3}$$

Assim, o novo resíduo studentizado é dado por

$$t_i^* = \frac{e_i}{s_{(i)}(1 - h_{ii})^{1/2}}, \tag{4}$$

em que t_i^* segue distribuição t-Student com $n-p-1$ graus de liberdade. A expressão (4) pode ainda ser simplificada substituindo-se (3) em (4), ou seja,

$$t_i^* = t_i\left(\frac{n - p - 1}{n - p - t_i^2}\right)^{1/2}.$$

Observe que t_i^* é uma transformação monótona de t_i. Assim, podemos usar t_i^* para detectar pontos aberrantes no conjunto de dados, isto é, pontos que apresentam um resíduo muito elevado, indicando que o modelo não descreve satisfatoriamente tal observação. Em geral, considera-se uma observação marginalmente aberrante se $|t_i^*| > 2$.

[1] O índice (i) indica que a i-ésima observação foi excluída.

2.3 Medidas de influência

Segundo Paula (2004), um tópico de grande importância na análise de diagnóstico é a detecção de observações influentes, isto é, pontos que exercem um peso desproporcional nas estimativas do modelo ou até mesmo na significância dos parâmetros. A deleção de pontos talvez seja a técnica mais conhecida para avaliar o impacto da retirada de uma observação particular nas estimativas da regressão. Durante a década de 70 surgiram várias propostas relacionadas com a influência das observações nas estimativas dos coeficientes do modelo normal linear. A distância de Cook (1977) é a mais tradicional medida para detectar pontos influentes e foi originalmente desenvolvida para modelos normais lineares e rapidamente assimilada e estendida para diversas classes de modelos.

Ainda, segundo Paula (2004), um problema que pode ocorrer com a deleção individual de pontos é o que se denomina *masking effect*, ou seja, deixar de detectar pontos conjuntamente discrepantes. Contudo, uma das propostas mais inovadoras na área de diagnóstico em regressão foi apresentada por Cook (1986), que propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo, ao invés da avaliação pela retirada individual ou conjunta de pontos. Essa metodologia, denominada influência local, teve uma grande receptividade entre os usuários e pesquisadores de regressão, havendo inúmeras publicações no assunto em que se aplica a metodologia em classes particulares de modelos ou em que se propõe extensões da técnica.

2.3.1 Influência

A medida de influência mais conhecida, denominada de distância de Cook (Cook, 1977), é da forma

$$D_i = \left\{ \frac{e_i}{s(1-h_{ii})^{1/2}} \right\}^2 \frac{h_{ii}}{(1-h_{ii})} \frac{1}{p} = t_i^2 \frac{h_{ii}}{(1-h_{ii})} \frac{1}{p}, \quad i = 1, 2, \ldots, n. \tag{5}$$

Note que D_i será grande quando o i-ésimo ponto for aberrante (t_i grande) e/ou quando h_{ii} for próximo de um. Adicionalmente, D_i poderá ser não adequado quando e_i for grande e h_{ii} for pequeno, nesse caso, s^2 pode ficar inflacionado e não ocorrendo nenhuma compensação por parte de h_{ii}, D_i pode ficar pequeno.

Uma medida de influência proposta em Belsley et al. (1980) é dada por

$$\text{DFBETA}_i = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{e_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i}{(1-h_{ii})}, \quad i = 1, 2, \ldots, n. \tag{6}$$

Esta medida reflete quanto a retirada do ponto (i) afasta $\widehat{\boldsymbol{\beta}}_{(i)}$ de $\widehat{\boldsymbol{\beta}}$. Os autores sugerem prestar atenção naqueles pontos que apresentam $|\text{DFBETA}_i|$ grande. Outra medida supostamente mais apropriada que D_i foi também proposta em Belsley et al. (1980), definida da forma

$$\text{DFFITS}_i = t_i^* \left\{ \frac{h_{ii}}{(1-h_{ii})} \right\}^{1/2}, \quad i = 1, 2, \ldots, n. \tag{7}$$

Um ponto pode ser considerado influente se $|\text{DFFITS}_i| \geq 2\big\{p/(n-p)\big\}^{1/2}$. Atkinson (1985) propôs uma medida de influência que é um aperfeiçoamento de DFFITS_i. Esta medida é dada por

$$C_i = t_i^* \left\{ \frac{(n-p)}{p} \frac{h_{ii}}{(1-h_{ii})} \right\}^{1/2}, \quad i = 1, 2, \ldots, n. \tag{8}$$

As medidas (5), (6), (7) e (8) apresentadas anteriormente são utilizadas, basicamente, para verificar o quanto a retirada do ponto (i) afasta $\widehat{\boldsymbol{\beta}}_{(i)}$ de $\widehat{\boldsymbol{\beta}}$, ou seja, o quanto o ponto (i) influencia nas estimativas dos parâmetros. Entretanto, Belsley et al. (1980) propuseram uma medida para verificar o quanto muda a estatística-t (usada para testar se o j-ésimo parâmetro é significativo) após a retirada do i-ésimo ponto.

Se a suposição de normalidade é satisfeita, pode-se usar a medida DFTSTAT_{ij} para verificar se o i-ésimo ponto muda a estatística-t. Esta medida é dada por

$$\text{DFTSTAT}_{ij} = \frac{\widehat{\beta}_j}{s\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}} - \frac{\widehat{\beta}_{(i)j}}{s_{(i)}\sqrt{(\mathbf{X}^\top_{(i)}\mathbf{X}_{(i)})^{-1}_{jj}}},$$

em que $\widehat{\beta}_j$ e $\widehat{\beta}_{(i)j}$ correspondem, respectivamente, ao j-ésimo parâmetro estimado com base em todas as observações e sem a i-ésima observação; $(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}$ e $(\mathbf{X}^\top_{(i)}\mathbf{X}_{(i)})^{-1}_{jj}$ denotam, respectivamente, o elemento (j,j) da diagonal principal de $(\mathbf{X}^\top\mathbf{X})^{-1}$ e $(\mathbf{X}^\top_{(i)}\mathbf{X}_{(i)})^{-1}$. Pontos que apresentam $|\text{DFTSTAT}_{ij}|$ grande podem ser considerados influentes em relação à estatística-t, isto é, pontos que mudam o valor da estatística de tal forma a alterar a significância (ou não significância) do parâmetro no modelo.

Existe também uma medida para verificar o quanto a matriz de variâncias e covariâncias de $\widehat{\boldsymbol{\beta}}$, $\text{Cov}(\widehat{\boldsymbol{\beta}})$, é afetada pela exclusão do i-ésimo ponto. Esta medida é dada por

$$\text{COVRATIO}_i = \frac{s^{2p}_{(i)}}{s^{2p}} \frac{\det\{(\mathbf{X}^\top_{(i)}\mathbf{X}_{(i)})^{-1}\}}{\det\{(\mathbf{X}^\top\mathbf{X})^{-1}\}},$$

em que $\det(\cdot)$ representa o determinante de uma matriz. A expressão acima pode ser simplificada na forma (Belsley et al., 1980)

$$\text{COVRATIO}_i = \left\{ \left[\frac{n-p-1}{n-p} + \frac{t_i^{*2}}{n-p} \right]^p (1-h_{ii}) \right\}^{-1}, \quad i = 1, 2, \ldots, n.$$

Deve-se dar atenção àqueles pontos tais que COVRATIO_i não apresenta um valor próximo da unidade, ou seja, pontos que alteram significativamente a variância dos estimadores.

Comportamento de D_i, DFFITS$_i$ e C_i

Nas Figuras 1 e 2, têm-se o comportamento das medidas de diagnóstico D_i, DFFITS$_i$ e C_i. Observe que à medida que o valor de t_i^2 aumenta, as três medidas tendem a indicar que o ponto é influente. Note também que quando h_{ii} está próximo de zero, estas medidas não indicam que o ponto é influente, mesmo para valores altos de t_i^2. Adicionalmente, quando h_{ii} está próximo de um, estas medidas indicam que o ponto é fortemente influente, mesmo para valores pequenos de t_i^2. Note que estas medidas apresentam resultados confiáveis no sentido de indicar que um ponto é influente quando realmente ele é influente (ou caso contrário) para valores de h_{ii} não muito pequenos ou muito grandes.

Pode ser observado que D_i é a menos adequada dentre a três medidas de diagnóstico para valores de h_{ii} pequenos, uma vez que mesmo para valores altos de t_i^2 (isto é, ponto aberrante), esta medida foi relativamente baixa. Portanto, quando o valor de h_{ii} está próximo de zero ou próximo de um, estas medidas de diagnóstico merecem uma atenção especial, pois como mostrado nas Figuras 1 e 2, são fortemente influenciadas pelos valores de h_{ii}.

2.3.2 Influência local

O método de influência local foi proposto por Cook (1986) e consiste em avaliar mudanças nos resultados da análise quando pequenas perturbações são incorporadas ao modelo e/ou aos dados. Se essas perturbações causarem efeitos desproporcionais, pode ser indício de que o modelo está mal ajustado ou que possam existir afastamentos sérios feitas para o mesmo.

A proposta de Cook (1986) tem sido vastamente utilizada na modelagem de regressão. Por exemplo, Beckman et al. (1987) apresentaram estudos de influência local em modelos de análise de variância com efeito misto. Lawrence (1988) investigou a aplicação de influência local em modelos lineares com parâmetros na transformação da resposta. Pettitt & Bin Daud (1989) aplicaram esta metodologia em modelos de regressão de Cox sob riscos proporcionais. Tsai & Wu (1992) investigaram influência local em modelos auto-regressivos de primeira ordem e modelos heteroscedásticos. Escobar & Meeker (1992) adaptaram o método de influência local para modelos de regressão com censura. Paula (1993) aplicou influência local em modelos lineares com restrições nos parâmetros na forma de desigualdades lineares. Kim (1995) e Pan et al. (1997) aplicaram métodos de influência local em regressão multivariada. Galea et al. (1997), Liu (2000), Galea et al. (2003) e Osorio et al. (2007) apresentaram estudos de influência local em modelos de contornos elípticos. Outros trabalhos são O'Hara Hines et al. (1992), Paula (1996), Kwan & Fung (1998), Gu & Fung (1998), Ortega et al. (2003), Rancel & Sierra (2001) e Svetliza & Paula (2001, 2003).

Para um conjunto de dados observados, seja $\ell(\boldsymbol{\theta})$ a função de log-verossimilhança do modelo postulado, em que $\boldsymbol{\theta}$ é um vetor p-dimensional de parâmetros desconhecidos. Seja $\boldsymbol{\omega}$ um vetor q-dimensional de perturbações assumindo valores em um subconjunto aberto $\Omega \subseteq \mathbb{R}^q$; em geral, tem-se que $q = n$.

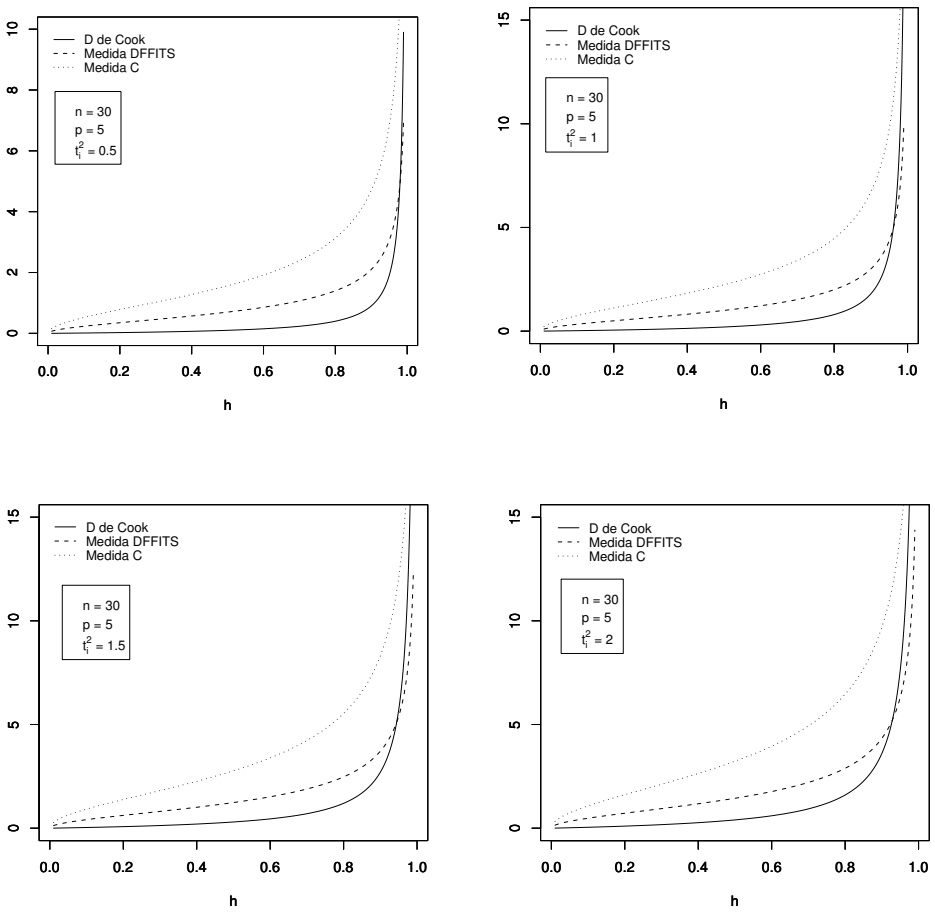

Figura 1 - Comportamento de D_i, DFFITS$_i$ e C_i.

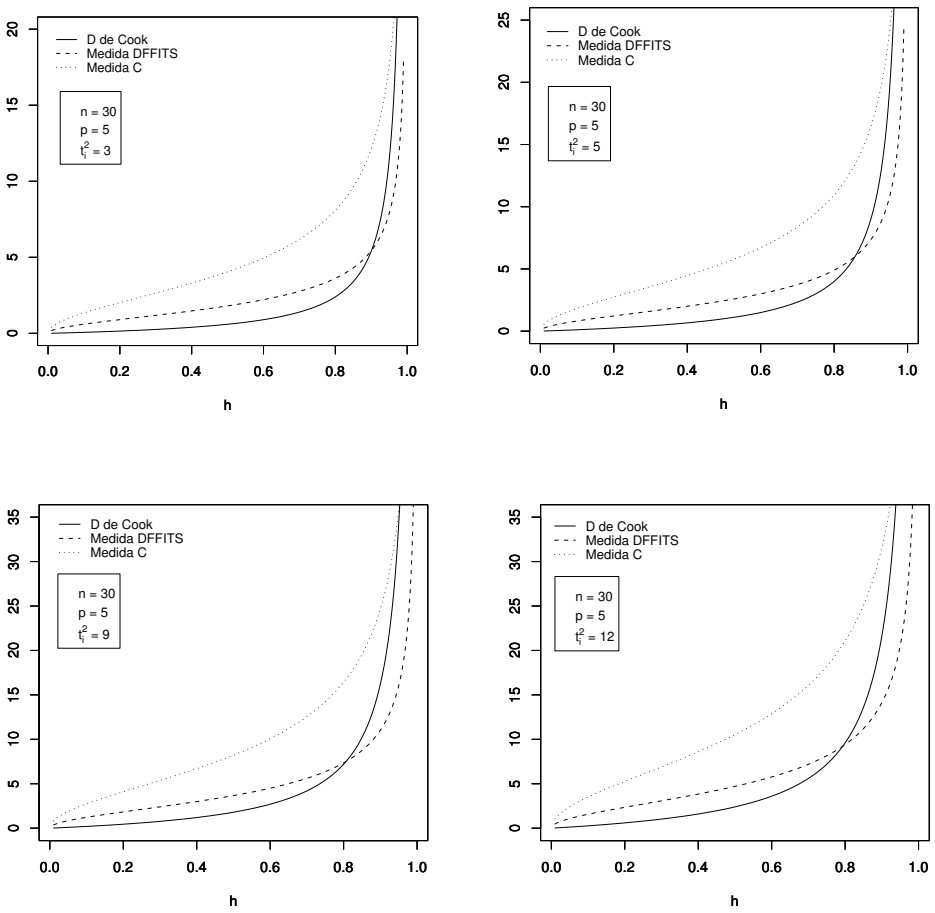

Figura 2 - Comportamento de D_i, DFFITS$_i$ e C_i.

Denote por $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ a função de log-verossimilhança do modelo perturbado. Assuma que a função de log-verossimilhança do modelo perturbado $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ é duas vezes diferenciável em $(\boldsymbol{\theta}^\top, \boldsymbol{\omega}^\top)^\top$, e que o modelo postulado está encaixado no modelo perturbado, isto é, existe $\boldsymbol{\omega}_0 \in \Omega$ tal que $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = \ell(\boldsymbol{\theta})$. Se p e q são pequenos, é suficiente comparar $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ e $\ell(\boldsymbol{\theta})$ graficamente para vários valores de $\boldsymbol{\omega}$ em Ω. Em situações mais gerais o problema é mais complexo.

Seja $\widehat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança (EMV) de $\boldsymbol{\theta}$ sob o modelo postulado e $\widehat{\boldsymbol{\theta}}_\omega$ o EMV sob o modelo perturbado. O objetivo é comparar $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ quando $\boldsymbol{\omega}$ varia em Ω. Se a distância entre eles permanece "pequena" quando $\boldsymbol{\omega}$ varia em Ω, isto indica que existe estabilidade do modelo ajustado no que diz respeito ao esquema particular de perturbação utilizado, e, portanto, ao correspondente aspecto da análise que está sendo monitorado. A diferença entre $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ pode depender da forma de $\ell(\widehat{\boldsymbol{\theta}})$. Se $\ell(\widehat{\boldsymbol{\theta}})$ é suficientemente horizontal, pode-se dizer que $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ são bem "próximos". Entretanto, se $\ell(\widehat{\boldsymbol{\theta}})$ é suficientemente concentrada em torno de $\widehat{\boldsymbol{\theta}}$, então $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ podem parecer muito distantes. Portanto, a comparação direta de $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ pode não ser simples devido a diversos fatores tais como diferença de escala, unidade de medida, erros de medição, etc. Uma proposta para comparar $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_\omega$ foi sugerida por Cook (1986) e consiste no uso da função

$$LD(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\omega)],$$

que é denominada afastamento pela verossimilhança ("likelihood displacement").

Como $\widehat{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ no modelo postulado, segue-se que $LD(\boldsymbol{\omega}) \geq 0$ para $\boldsymbol{\omega} \in \Omega$. Adicionalmente, como $LD(\boldsymbol{\omega}_0) = 0$, pode-se concluir que $\boldsymbol{\omega}_0$ é um ponto de mínimo local da função $LD(\boldsymbol{\omega})$. Uma análise sobre o comportamento geométrico da função $LD(\boldsymbol{\omega})$ quando $\boldsymbol{\omega}$ varia em Ω, pode fornecer informações de características relevantes ao modelo sob investigação; por exemplo, se os valores de $LD(\boldsymbol{\omega})$, para diferentes valores de $\boldsymbol{\omega}$ em Ω, são "bem" próximos, isto indica que existe estabilidade no modelo ajustado sob a perturbação considerada. O gráfico da função $LD(\boldsymbol{\omega})$ pode ser representado pela superfície geométrica $(q+1)$-dimensional formado pelos valores do vetor

$$\boldsymbol{\alpha}(\boldsymbol{\omega}) = (\boldsymbol{\omega}, LD(\boldsymbol{\omega}))^\top.$$

No contexto estatístico essa superfície é chamada de gráfico de influência, uma vez que o gráfico da função $LD(\boldsymbol{\omega})$ mostra a influência do esquema de perturbação; veja Cook (1986).

O estudo de influência local consiste em analisar como a superfície $\boldsymbol{\alpha}(\boldsymbol{\omega}) = (\boldsymbol{\omega}, LD(\boldsymbol{\omega}))^\top$ desvia-se de seu plano tangente em $\boldsymbol{\omega}_0$. Essa análise pode ser feita estudando-se as curvaturas das seções normais da superfície $\boldsymbol{\alpha}(\boldsymbol{\omega})$ em $\boldsymbol{\omega}_0$, que são interseções de $\boldsymbol{\alpha}(\boldsymbol{\omega})$ com o plano contendo o vetor normal a seu plano tangente em $\boldsymbol{\omega}_0$. Se a classe de todas as seções normais da superfície $\boldsymbol{\alpha}(\boldsymbol{\omega})$ em $\boldsymbol{\omega}_0$ tiverem curvatura normal "próxima" de zero, isto é, as seções normais são "próximas" a uma linha reta em torno de $\boldsymbol{\omega}_0$, então $LD(\boldsymbol{\omega})$ é "próximo" de zero em torno de $\boldsymbol{\omega}_0$, o que indica que existe uma estabilidade do modelo ajustado sob o particular esquema de perturbação que está sendo considerado. Por outro lado, se existir uma seção normal

com curvatura, em módulo, não "próximo" de zero, então, nessa direção, tem-se maior oscilação da função $LD(\boldsymbol{\omega})$, e, portanto, sob essa seção normal, tem-se maior influência do particular esquema de perturbação. Projetando perpendicularmente esta seção normal sobre o plano $\mathbb{R}^q$, obtém-se um vetor $\mathbf{d} \in \mathbb{R}^q$ tal que, a reta $\boldsymbol{\omega}_0 + t\mathbf{d}$, $t \in \mathbb{R}$, e a seção normal, pertencem ao mesmo plano e, portanto, será equivalente considerar uma seção normal a $\boldsymbol{\alpha}(\boldsymbol{\omega})$ em $\boldsymbol{\omega}_0$ com um vetor $\mathbf{d} \in \mathbb{R}^q$ para indicar a direção onde existe maior ou menor influência em torno de $\boldsymbol{\omega}_0$.

Para caracterizar o comportamento de $LD(\boldsymbol{\omega})$ em torno de $\boldsymbol{\omega}_0$, considere primeiramente uma direção arbitrária $\mathbf{d}$ em $\mathbb{R}^q$ (sem perda de generalidade, seja $\mathbf{d}$ um vetor unitário, $||\mathbf{d}|| = 1$). Posteriormente, considere o gráfico de $LD(\boldsymbol{\omega}_0 + t\mathbf{d})$ contra $t \in \mathbb{R}$. Por causa de $LD(\boldsymbol{\omega}_0) = 0$, o gráfico de $LD(\boldsymbol{\omega}_0 + t\mathbf{d})$ tem um mínimo local em $t = 0$. A curvatura normal, $C(\boldsymbol{\theta})$, é uma caracterização de $LD(\boldsymbol{\omega}_0 + t\mathbf{d})$ em torno de $t = 0$. Uma sugestão inicial é tomar a direção $\mathbf{d}_{\max}$ que corresponde à maior curvatura, $C_{\max}(\boldsymbol{\theta})$. Os valores de $\mathbf{d}_{\max}$ contêm a influência local das observações nessa direção particular. Logo, o gráfico de $|\mathbf{d}_{\max}|$ contra a ordem das observações pode revelar aqueles pontos com maior influência na vizinhança de $\boldsymbol{\omega}_0$. Tais pontos podem ser responsáveis por mudanças substanciais nas estimativas dos parâmetros sob pequenas perturbações no modelo. Portanto, deve-se olhar com mais cuidado esses pontos a fim de entender melhor a influência dos mesmos e conseqüentemente tentar propor uma forma segura de usar o modelo ajustado.

Cook (1986) mostra que a curvatura normal na direção $\mathbf{d}$ é dada por

$$C(\boldsymbol{\theta}) = 2|\mathbf{d}^\top \boldsymbol{\Delta}^\top \ddot{\boldsymbol{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{\Delta}\mathbf{d}|,$$

em que $\ddot{\boldsymbol{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ e $\boldsymbol{\Delta} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^\top$, avaliados em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. É interessante notar que o máximo de $\mathbf{d}^\top\mathbf{M}\mathbf{d}$, em que $\mathbf{M} = -\boldsymbol{\Delta}^\top \ddot{\boldsymbol{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{\Delta}$, corresponde ao maior autovalor (em valor absoluto) da matriz $\mathbf{M}$. Desta forma, $C_{\max}$ é o maior autovalor da matriz $\mathbf{M}$ e $\mathbf{d}_{\max}$ é um correspondente autovetor de norma igual a um.

É possível também avaliar a influência local apenas para uma parte do vetor de parâmetros $\boldsymbol{\theta}$. Suponha que seja possível particionar $\boldsymbol{\theta}$ da forma $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$. Segundo Cook (1986), se o interesse é calcular a influência local apenas para $\boldsymbol{\theta}_1$, a curvatura normal na direção $\mathbf{d}$ é da forma

$$C(\boldsymbol{\theta}_1) = 2|\mathbf{d}^\top \boldsymbol{\Delta}^\top (\ddot{\boldsymbol{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} - \mathbf{M}_1)\boldsymbol{\Delta}\mathbf{d}|,$$

em que

$$\mathbf{M}_1 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddot{L}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2}^{-1} \end{pmatrix},$$

com $\ddot{L}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^\top$ e, neste caso, $C_{\max}$ é o autovetor de norma igual a um correspondente ao maior autovalor da matriz $\boldsymbol{\Delta}^\top(\ddot{\boldsymbol{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} - \mathbf{M}_1)\boldsymbol{\Delta}$. O mesmo procedimento pode ser feito para avaliar a influência local apenas para $\boldsymbol{\theta}_2$. Em particular, no modelo de regressão normal linear tem-se que $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. Dessa forma, se o interesse está somente em obter a influência local das observações sobre

$\boldsymbol{\beta}$, a curvatura normal na direção $\mathbf{d}$ é dada por $C(\boldsymbol{\beta}) = 2|\mathbf{d}^\top \boldsymbol{\Delta}^\top (\ddot{\boldsymbol{L}}_{\boldsymbol{\theta\theta}}^{-1} - \mathbf{M}_1)\boldsymbol{\Delta}\mathbf{d}|$, sendo

$$\mathbf{M}_1 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddot{L}^{-1}_{\sigma^2\sigma^2} \end{pmatrix},$$

em que $\ddot{L}_{\sigma^2\sigma^2} = \partial^2 \ell(\boldsymbol{\theta})/\partial(\sigma^2)^2$. O gráfico do maior autovetor de $\boldsymbol{\Delta}^\top(\ddot{\boldsymbol{L}}_{\boldsymbol{\theta\theta}}^{-1} - \mathbf{M}_1)\boldsymbol{\Delta}$ contra a ordem das observações pode mostrar aquelas observações com maior influência local sobre $\boldsymbol{\beta}$.

Esquemas de perturbação

Em estatística, não existe uma definição clara de perturbação. Segundo Billor & Loynes (1993), perturbação é qualquer arranjo da mudança da suposição do modelo e/ou dados perturbados para constatar alguma mudança substancial que ocorre nos resultados da análise. A seguir, alguns esquemas de perturbação mais comuns são apresentados.

Ponderação de Casos: A função de log-verossimilhança perturbada tem a forma

$$\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \ell_i(\boldsymbol{\theta}),$$

em que $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, com $0 \leq \omega_i \leq 1$, $i = 1, \ldots, n$. Com este tipo de perturbação, deseja-se avaliar se a contribuição das observações com ponderações diferentes afeta na estimação do parâmetro $\boldsymbol{\theta}$. A ponderação de casos tem sido o esquema de perturbação mais utilizado para análise de influência (Cook, 1987). Este esquema pode ser interpretado como uma ponderação na variância do i-ésimo caso, em especial nos modelos normais lineares (Thomas & Cook, 1989). Quando $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (1, 1, \ldots, 1)^\top$, o modelo perturbado se reduz ao modelo postulado.

Perturbação na Resposta: Este tipo de perturbação considera, em geral, um esquema aditivo de perturbação da resposta em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é alterado através da adição de um vetor $\boldsymbol{\omega}$ de pequenas perturbações. É comum utilizar um fator de escala para padronizar os componentes de $\boldsymbol{\omega}$, por exemplo, a estimativa do desvio padrão de y_i, $\widehat{\sigma}$, de forma que

$$y_i(\omega_i) = y_i + \widehat{\sigma}\omega_i, \quad i = 1, 2, \ldots, n.$$

Quando $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (0, 0, \ldots, 0)^\top$, a variável perturbada se reduz a variável original.

Perturbação nas Covariáveis: Neste caso, considera-se uma perturbação aditiva de uma variável explicativa particular, digamos $\mathbf{x}_t$, adicionando um vetor $\boldsymbol{\omega}$ de pequenas perturbações ponderado por um fator de escala S_t, em que S_t é o desvio padrão da t-ésima covariável modificada, de forma que

$$x_{it}(\omega_i) = x_{it} + S_t\omega_i, \quad i = 1, 2, \ldots, n.$$

Através deste esquema de perturbação é possível acessar a influência individual de cada covariável no processo de estimação do modelo. No entanto, este tipo de perturbação faz sentido somente se a covariável é medida de forma contínua. Quando $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (0, 0, \ldots, 0)^\top$, a variável perturbada se reduz a variável original.

3 Ilustração prática

Os dados utilizados correspondem a uma amostra de 27 imóveis e estão apresentados na Tabela 1 (Paula, 2004). O objetivo do estudo é tentar explicar o preço de venda do imóvel – y (em 1000 dólares) segundo as covariáveis: x_1 – imposto do imóvel (em 100 dólares); x_2 – área do terreno (em 1000 pés quadrados); x_3 – área construída (em 1000 pés quadrados); x_4 – idade da residência (em anos). Todas a análises apresentadas nesta seção foram feitas utilizando o software estatístico R em sua versão 2.7.1.

Primeiramente foi ajustado um modelo de regressão normal linear incluindo todos as covariáveis, ou seja, o modelo a ser ajustado é da forma

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i, \quad i = 1, 2, \ldots, 27.$$

Os coeficientes estimados, erros padrão, estatística-t, p-valores, uma medida de qualidade de ajuste, R^2, e o erro padrão residual, s, estão apresentados na Tabela 2.

Observe que apenas as covariáveis x_1 e x_3 foram significativas. Já as covariáveis x_2 e x_4 não foram significativas, isto é, não estão associadas ao preço de venda do imóvel. Note que o intercepto (Constante) também não foi significativo.

A fim de escolher um modelo mais apropriado aos dados, utilizou-se o critério de seleção AIC. (Esta função está implementada no software R através do comando `stepAIC` da *library* `MASS`.) Após aplicar o método, apenas as covariáveis x_1 e x_3 permaneceram no modelo, além do intercepto. O intercepto não foi significativo, porém, por interpretações práticas será mantido no modelo. É interessante notar que o método selecionou justamente as covariáveis que foram significativas no modelo com todas as covariáveis. As estimativas do modelo selecionado estão na Tabela 3.

Como um modelo aparentemente apropriado aos dados foi selecionado (Tabela 3), o próximo passo é aplicar as medidas de diagnóstico apresentadas na seções anteriores e verificar as suposições iniciais do modelo, bem com se há alguma observação que exerça algum tipo de influência sobre as estimativas dos parâmetros, ou seja, verificar se alguma observação muda a inferência com relação à significância (ou não significância) dos parâmetros.

Na Figura 3, tem-se as medidas de diagnóstico. Há oito gráficos nesta figura: Pontos de Alavanca, Distância de Cook - Di, Medida DFFITSi, Medida Ci, Influência Local, Pontos Aberrantes, Homocedasticidade e Envelope. Este último refere-se a um gráfico de probabilidades normal com bandas de confiança, sendo utilizado para verificar se a distribuição que foi postulada para a variável resposta se verifica, neste caso, a distribuição normal.

Tabela 1 - Dados da ilustração

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
4.9176	3.4720	0.9980	42	25.9	5.0500	5.0000	1.0200	46	30.0
5.0208	3.5310	1.5000	62	29.5	8.2464	5.1500	1.6640	50	36.9
4.5429	2.2750	1.1750	40	27.9	6.6969	6.9020	1.4880	22	41.9
4.5573	4.0500	1.2320	54	25.9	7.7841	7.1020	1.3760	17	40.5
5.0597	4.4550	1.1210	42	29.9	9.0384	7.8000	1.5000	23	43.9
3.8910	4.4550	0.9880	56	29.9	5.9894	5.5200	1.2560	40	37.5
5.8980	5.8500	1.2400	51	30.9	7.5422	4.0000	1.6900	22	37.9
5.6039	9.5200	1.5010	32	28.9	8.7951	9.8900	1.8200	50	44.5
15.4202	9.8000	3.4200	42	84.9	6.0931	6.7265	1.6520	44	37.9
14.4598	12.8000	3.0000	14	82.9	8.3607	9.1500	1.7770	48	38.9
5.8282	6.4350	1.2250	32	35.9	8.1400	8.0000	1.5040	3	36.9
5.3003	4.9883	1.5520	30	31.5	9.1416	7.3262	1.8310	31	45.8
6.2712	5.5200	0.9750	30	31.0	12.0000	5.0000	1.2000	30	41.0
5.9592	6.6660	1.1210	32	30.9					

Tabela 2 - Ajuste do modelo com todos as covariáveis

Covariáveis	Estimativa	Erro padrão	Estatística-t	p-valor
Constante	2.436	4.092	0.595	0.558
x_1	2.078	0.553	3.758	0.001
x_2	0.232	0.507	0.459	0.651
x_3	13.974	2.907	4.808	0.000
x_4	−0.044	0.066	−0.660	0.516
R^2	0.931			
s	4.077			

Tabela 3 - Estimativas do modelo selecionado

Covariáveis	Estimativa	Erro padrão	Estatística-t	p-valor
Constante	0.790	2.279	0.347	0.732
x_1	2.297	0.489	4.698	0.000
x_3	13.933	2.524	5.519	0.000
R^2	0.928			
s	3.982			

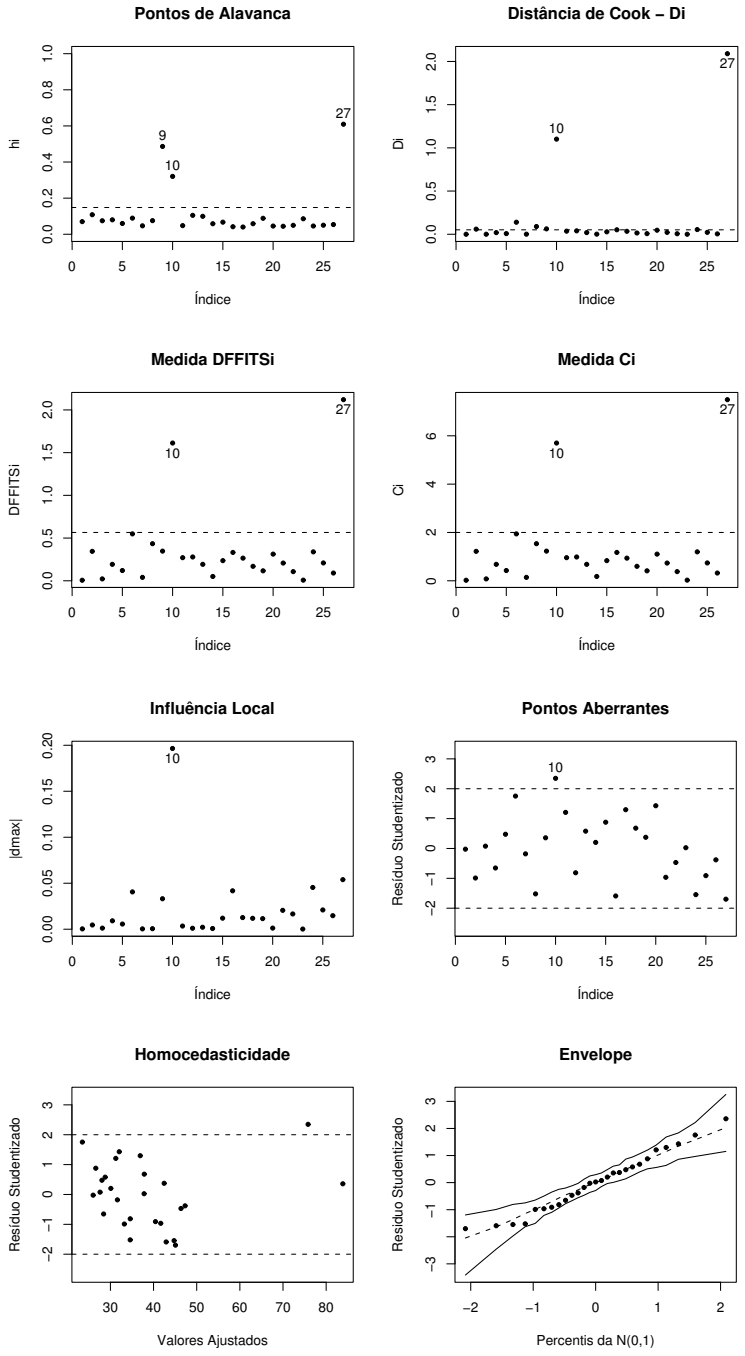

Figura 3 - Medidas de Diagnóstico.

Observe no gráfico Pontos de Alavanca que três pontos (9, 10 e 27) se destacam entre os demais. Eles apresentaram uma "alta" influência (medida através de $\partial\widehat{y}_i/\partial y_i = h_{ii}$) do preço de venda do imóvel observado sobre o preço de venda do imóvel predito. Adicionalmente, note nos gráficos Distância de Cook - Di, Medida DFFITSi, Medida Ci e Influência Local (os quais destacam pontos que possivelmente terão um peso desproporcional nas estimativas dos parâmetros, influentes), duas observações se destacam (10 e 27). Deve-se "olhar" com bastante cuidado estas observações, uma vez que elas podem estar causando o que se denomina *masking effect*, ou seja, mascarando a presença de alguma covariável no modelo. Dessa forma, como as medidas indicaram esses pontos como possíveis influentes, será feito uma análise confirmatória para verificar se tais observações alteram a inferência com respeito a significância dos parâmetros.

Na Tabela 4, tem-se a variação percentual (VP) das estimativas dos parâmetros juntamente com os p-valores quando se retira marginalmente e conjuntamente os pontos 10 e 27. Note nesta tabela que a maior variação percentual foi em relação ao intercepto, porém, a não significância do mesmo continuou inalterada. Adicionalmente, o p-valor para a covariável x_1 (isto é, β_1) ficou inalterado, já o p-valor de x_3 (isto é, β_3) foi levemente alterado, no entanto, não mudando a significância da mesma no modelo. Portanto, chega-se a conclusão que as observações 10 e 27 não alteram a inferência do modelo selecionado e, de fato, as covariáveis x_1 e x_3 devem ser mantidas no modelo.

Tabela 4 - Variação da estimativas e p-valores

	Todas		Sem a obs. 10	
	VP(%)	p-valor	VP(%)	p-valor
Constante	0	0.732	330.06	0.164
x_1	0	0.000	5.62	0.000
x_3	0	0.000	9.77	0.000
	Sem a obs. 27		Sem as obs. 10 e 27	
	VP(%)	p-valor	VP(%)	p-valor
Constante	35.14	0.632	337.33	0.145
x_1	41.73	0.000	31.01	0.000
x_3	32.44	0.015	37.26	0.015

No gráfico Pontos Aberrantes, observe que a observação 10 se destaca, ficando fora dos limites $(-2, 2)$. Este ponto pode ser considerado marginalmente aberrante, uma vez que o resíduo para esta observação é significativamente diferente de zero, indicando que o modelo selecionado não descreve, completamente, esta observação. Observando o gráfico Homocedasticidade, note que, aparentemente, não há indícios de heteroscedasticidade (variância não constante). (É claro que um estudo mais refinado como a aplicação de algum teste de heteroscedasticidade pode ser mais conclusivo em relação à não constância da variância.) Pelo gráfico Envelope, note

que a suposição de normalidade para a variável resposta está satisfeita, uma vez que todos os pontos estão dentro das bandas de confiança. Mais detalhes sobre a construção deste gráfico podem ser encontrados em Atkinson (1985).

Conclusões

Este artigo apresenta uma revisão de técnicas de diagnóstico em modelos de regressão normais lineares. Algumas medidas de diagnóstico foram apresentadas e discutidas, em particular, o método de influência local desenvolvido por Cook (1986) que propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo ou dados. Este método vem sendo amplamente utilizado por usuários na modelagem de regressão e teve uma grande receptividade entre os pesquisadores, havendo inúmeras publicações no assunto em que se aplica a metodologia em classes particulares de modelos ou extensões da técnica. Adicionalmente, apresentamos e discutimos uma ilustração prática aplicando as medidas apresentadas no artigo a um conjunto de dados reais.

A aplicação das técnicas de diagnóstico são de fundamental importância para verificar (validar) a adequacidade de um ajuste de um modelo de regressão, bem como identificar observações que podem influenciar consideravelmente tal ajuste. Portanto, ao aplicar a técnica de análise de regressão a um conjunto de dados reais, deve-se estar ciente de que um ajuste razoável (satisfatório) de um modelo de regressão vem acompanhado de uma boa análise de diagnóstico.

Agradecimentos

O autor agradece ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, e um parecerista pelos comentários.

LEMONTE, A. J. Diagnostic in normal linear regression: principles and applications. *Rev. Bras. Biom.,* São Paulo, v.26, n.2, p.7-26, 2008.

▪ *ABSTRACT: Statistical models are very used to extract and knowing the characteristics of a data set. Models, however, are always approximate descriptions of a process more complicated. So considerations of quality of a model are extremely important. We present in this article some techniques of diagnostic in normal linear regression model, which can be seen as methods to study the variation in the analysis of the results when the model in the study is subject to some kind of disturbance. Additionally, we present in detail the method of local influence developed by Cook (1986, 1987).*

▪ *KEYWORDS: Local influence; measures of diagnostic; normal linear model.*

Referências

ATKINSON, A. C. *Plots, transformations and regressions*. Oxford: Oxford Statistical Science Series, 1985. 282p.

BECKMAN, R. J.; NACHTSHEIM, C. J.; COOK, R. D. Diagnostics for mixed-model analysis of variance. *Technometrics*, Alexandria, v.29, p.413–426, 1987.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. *Residuals and influence in regression*. London: Chapman & Hall, 1980. 292p.

BILLOR, N.; LOYNES, R. M. Local influence: a new approach. *Commun. Stat. - Theory Methods*, New York, v.22, p.1595–1611, 1993.

COOK, R. D. Detection of influencial observations in linear regressions. *Technometrics*, Alexandria, v.19, p.15–18, 1977.

COOK, R. D. Assessment of local influence (with discution). *J. R. Stat. Soc. Ser. B*, London, v.48, p.133–169, 1986.

COOK, R. D. Influence assessment. *J. Appl. Stat.*, Abingdon, v.14, p.117–131, 1987.

COOK, R. D.; WEISBERG, S. *Residuals and influence in regression*. London: Chapman & Hall, 1982. 230p.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. New York: John Wiley & Sons, 1998. 470p.

ESCOBAR, L. A.; MEEKER, W. Q. Assessing influence in regression analysis with censored data. *Biometrics*, Washington, v.48, p.507–528, 1992.

GALEA, M.; PAULA, G. A.; BOLFARINE, H. Local influence in elliptical linear regression models. *Statistician*, London, v.46, p.71–79, 1997.

GALEA, M.; PAULA, G. A.; URIBE-OPAZO, M. On influence diagnostic in univariate elliptical linear regression models. *Stat. Pap.*, New York, v.44, p.23–45, 2003.

GU, H.; FUNG, W. K. Assessing local influence in canonical correlation analysis. *Ann. Inst. Stat. Math.*, Tokio, v.50, p.755–772, 1998.

GUJARATI, D. *Econometria básica*. São Paulo: Makron Books, 2000. 812p.

HOAGLIN, D. C.; WELSCH, R. E. The hat matrix in regression and ANOVA. *Am. Stat.*, Washington, v.32, p.17–22, 1978.

KIM, M. G. Local influence in multivariate regression. *Comm. Stat. - Theory Methods*, New York, v.20, p.1271–1278, 1995.

KWAN, C. W.; FUNG, W. K. Assessing local influence for specific restricted likelihood: Applications to factor analysis. *Psychometrika*, New York, v.63, p.35–46, 1998.

LAWRENCE, A. F. Regression transformation diagnostics using local influence. *J. Am. Stat. Assoc.*, New York, v.84, p.125–141, 1988.

LIU, S. Z. On local influence for elliptical linear models. *Stat. Pap.*, New York, v.41, p.211–224, 2000.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. 3.ed. New York: John Wiley & Sons, 2001. 527p.

O'HARA HINES, R. J.; LAWLESS, J. F.; CARTER, E. M. Diagnostics for a cumulative multinomial generalized linear model with application to grouped toxicological mortality data. *J. Am. Stat. Assoc.*, New York, v.87, p.1059–1069, 1992.

ORTEGA, E. M. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostic in generalized log-gamma regression models. *Comput. Stat. Data Anal.*, Amsterdam, v.42, p.165–186, 2003.

OSORIO, F.; PAULA, G. A.; GALEA, M. Assessment of local influence in elliptical linear models with longitudinal strucuture. *Comput. Stat. Data Anal.*, Amsterdam, v.51, p.4354–4368, 2007.

PAN, J. X.; FANG, K. T.; ROSEN, V. Local influence assessment in the growth curve model with unstructured covariance. *J. Stat. Plann. Infer.*, Amsterdam, v.62, p.263–278, 1997.

PAULA, G. A. Assessing local influence in restricted regression models. *Comput. Stat. Data Anal.*, Amsterdam, v.16, p.63–79, 1993.

PAULA, G. A. Influence diagnostic in proper dispersion models. *Aust. J. Stat.*, Sydney, v.38, p.307–316, 1996.

PAULA, G. A. *Modelos de regressão com apoio computacional*. São Paulo: IME–USP, 2004. 245p.

PETTITT, A. N.; BIN DAUD, I. Case-weight measures of influence for proportional hazards regression. *Appl. Stat.*, Washington, v.38, p.51–67, 1989.

RANCEL, M. M. S.; SIERRA, M. A. G. Regression diagnostics using local influence: a review. *Commun. Stat. - Theory Methods*, New York, v.30, p.799–813, 2001.

SEARLE, S. R. *Linear models*. New York: John Wiley & Sons, 1971. 532p.

SOUZA, F. A. M. *Influência local e análise de resíduos em modelos de regressão von Mises*, 1999. 116f. Tese (Doutorado em estatística), Instituto de Matemática e Estatística - Universidade de São Paulo, São Paulo, 1999.

SVETLIZA, C. F.; PAULA, G. A. On diagnostics in log-linear negative binomial models. *J. Stat. Comput. Simul.*, New York, v.71, p.231–244, 2001.

SVETLIZA, C. F.; PAULA, G. A. Diagnostics in nonlinear negative binomial models. *Commun. Stat. - Theory Methods*, New York, v.32, p.1227–1250, 2003.

THOMAS, W.; COOK, R. D. Assessing influence on regression coefficients in generalized linear models. *Biometrika*, London, v.79, p.741–749, 1989.

TSAI, C. H.; Wu, X. Assessing local influence in linear regression models with first-order autoregressive or heteroscedastic error structure. *Stat. Probab. Lett.*, Amsterdam, v.14, p.247–252, 1992.

WEISBERG, S. *Applied linear regression*. New York: John Wiley & Sons, 1985. 324p.

Recebido em 01.02.2008.

Aprovado após revisão em 12.06.2008.