ANÁLISE DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO NORMAL E LOGÍSTICO

Dorival LOBATO JUNIOR[1]
Ruben Delly VEIGA[2]

RESUMO: A análise de diagnóstico da qualidade do ajuste é uma importante fase em modelos de regressão, principalmente no que tange a ocorrência de *outliers*, que podem provocar distorções substanciais nos parâmetros estimados do modelo. Neste trabalho, abordou-se dois métodos inovadores nessa área: o método da influência local, que foi proposto, inicialmente, por Cook (1986) e a inclinação máxima, proposta por Billor e Loynes (1993). A proposta de Cook consiste em avaliar a curvatura normal de uma superfície, com base na medida *Likelihood Displacement* (afastamento pela verossimilhança), sob uma pequena perturbação no modelo. Em seguida, apresentou-se a abordagem de Billor e Loynes (1993), os quais aplicaram uma medida afastamento pela verossimilhança modificada. Com isso, a inclinação máxima é utilizada como medida da influência de *outliers* no modelo. Inicialmente a proposta de Billor e Loynes (1993) era aplicada no modelo normal, desenvolveu-se essa proposta também para o modelo logístico. Apresentou-se duas aplicações, nas quais foram utilizadas técnicas de análise de diagnóstico da qualidade do ajuste, por meio de gráficos e das duas propostas: influência local e inclinação máxima. No caso do modelo normal linear simples, a curvatura normal e a inclinação máxima apresentaram mesma sensibilidade na indicação de observações influentes. Já na segunda aplicação, as técnicas utilizadas nesse trabalho, não mostraram uma indicação consistente em relação a alguma observação, porém alguns gráficos coincidiram na indicação da observação 13, como a que mais se destaque, estando afastada das demais.

PALAVRAS-CHAVE: Análise de regressão; regressão logística; técnicas de diagnóstico; influência local; inclinação máxima.

1 Introdução

Segundo Billor e Loynes (1993), um tópico importante na análise de diagnóstico é a detecção de pontos que exercem um peso desproporcional nas estimativas dos parâmetros do modelo, os *outliers*. Para eles a deleção de pontos talvez seja o método mais conhecido para avaliar o impacto da retirada de uma observação particular nas estimativas da regressão.

Cook (1986) propõe um método inovador para avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo, ao invés da avaliação pela

[1] Universidade do Estado do Pará - UEPA, Departamento de Matemática, Estatística e Informática, Belém-PA, Brasil. E-mail: *lobatojr2012@gmail.com.*
[2] Universidade Federal de Lavras - UFLA, Departamento de Estatística, Lavras-MG, Brasil. E-mail*: rdelly@uol.com.br*.

retirada individual ou conjunta de pontos. Essa metodologia, denominada influência local, segundo Paula (2004), tem tido grande aceitação entre os pesquisadores e usuários de regressão.

Billor e Loynes (1993) apresentaram uma outra proposta para avaliar a influência de *outliers* em modelos de regressão normal, aplicaram uma medida afastamento pela verossimilhança modificada, cuja primeira derivada não se anula. Com isso, a inclinação máxima é utilizada como medida da influência de *outliers* no modelo. Tanto a metodologia de Cook (1986) como a de Billor e Loynes (1993) utilizam a medida *likelihood displacement* (afastamento pela verossimilhança) como função objetivo.

A proposta deste trabalho é apresentar algumas técnicas gráficas de análise de diagnóstico e os dois métodos já citados, o método da influência local de Cook (1986) e da inclinação máxima de Billor e Loynes (1993), este último foi proposto originalmente para o modelo normal, desenvolveu-se uma adaptação do mesmo para o modelo de regressão logístico. Mostrou-se o desenvolvimento teórico dos dois métodos e apresentou-se duas aplicações.

Na segunda seção, se desenvolveu o modelo logístico, se definiu alguns resíduos e destacou-se algumas técnicas gráficas, na terceira seção obteve-se várias versões para o cálculo da curvatura normal, na quarta seção aplicamos as técnicas gráficas e a influência local e a inclinação máxima em dois exemplos, um modelo linear e outro modelo logístico. Finalizou-se com as conclusões.

2 Modelos logísticos, resíduos e técnicas gráficas

2.1 Regressão logística

Apresentados por Nelder e Wedderburn (1972), os modelos lineares generalizados (MLG's), permitem que a distribuição da variável resposta pertença à família exponencial de distribuições, significando que cada componente do vetor de observações **Y** tenha a função densidade ou função de probabilidade na forma

$$f(y,\theta_i,\emptyset) = \exp\{\emptyset[y\theta_i - b(\theta_i)] + c(y,\emptyset)\}, \tag{2.1}$$

em que os $\theta_{i's}$ são os parâmetros de interesse e $\emptyset$ é o parâmetro de dispersão e $b(\theta_i) = E[y_i]$. Com relação à componente sistemática do modelo, as variáveis independentes $X_1, X_2, \cdots, X_p$ produzem o vetor $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^t$como preditor linear, que, no caso da regressão normal linear múltipla, é dada por

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots + \beta_p x_{pi} \\ &= \boldsymbol{x_i^t \beta}, com\ i = 1, 2, \ldots, n\end{aligned} \tag{2.2}$$

em que o vetor $\boldsymbol{x}_i^t = (1, x_{1i}, x_{2i}, \ldots, x_{pi})$ é a i-ésima linha da matriz $\boldsymbol{X}$ de variáveis explicativas e $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo. Além disso, a ligação entre a parte sistemática e a média condicional da variável resposta se dá por meio da seguinte relação:

$$\mu_i = g^{-1}(\eta_i), \qquad i = 1, 2, \ldots, n$$

em que g(.) é denominada de função de ligação, que deve ser monótona e diferenciável. No caso da regressão normal, é a função identidade.

No caso da regressão logística simples, assume-se que $nY \sim b(n,p)$. Com isso, a sua função de probabilidade fica estabelecida por

$$\binom{n}{ny} p^{ny}(1-p)^{n-ny} = n\ln(1-p)$$

em que $0 < p < 1$ e $y \in \left\{0, \ldots, \frac{1}{n}, \ldots, 1\right\}$.

Uma função de ligação, denominada logit, é definida por

$$t \longmapsto logit(t) = \ln\frac{t}{1-t}. \tag{2.3}$$

Dessa forma, obtém-se o modelo logístico linear simples

$$logit\, p(x) = \eta \tag{2.4}$$

em que $\eta = \beta_0 + \beta_1 x$ é o preditor linear, e β_0 e β_1 são parâmetros desconhecidos. Segue, de (2.3) e (2.4), que

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{2.5}$$

Para o caso da regressão logística múltipla. Considere-se o modelo geral de regressão logística múltipla

$$p(x) = \frac{e^n}{1+e^n} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}. \tag{2.6}$$

2.2 Resíduos

O resíduo para a i-ésima observação é definido como uma função do tipo $r_i = r(y_i\hat{\mu}_i)$, cujo objetivo é medir a discrepância entre o valor observado e o valor ajustado da i-ésima observação $\widehat{\mu_i}$. A definição mais usual é a de resíduo ordinário**,** dada por

$$r_i = y_i - \widehat{\mu_i}\,. \tag{2.7}$$

Para $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_n)^t$, tem-se que, para o modelo de regressão normal linear,

$$\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y},$$

em que $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ é chamada de matriz *hat*. O elemento h_{ii}, definido a seguir, pertence à diagonal da matriz **H** e desempenha um importante papel na construção de técnicas de diagnóstico

$$h_{ii} = \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i. \quad (2.8)$$

Em particular, $Var(r_i) = (1 - h_{ii})\sigma^2$. Partindo do pressuposto de que todos os pontos exercem a mesma influência sobre os valores ajustados, espera-se que o valor de h_{ii} esteja próximo de $\frac{p}{n}$. Convém, então, examinar aqueles pontos, tais que $h_{ii} \geq \frac{2p}{n}$, que podem ser diagnosticados como pontos que estão localizados em regiões remotas no subespaço gerado pelas colunas da matriz **X** (PAULA, 2004).

Em relação ao modelo de regressão normal linear, tem-se que $r_i \sim N(0, \sigma^2)$. Assim, podem-se obter os resíduos studentizados, denotados por t_i, dividindo-se cada r_i pelo seu respectivo desvio padrão amostral $s(1 - h_{ii})^{1/2}$, em que $s^2 = \sum_{i=1}^{n} \frac{r_i^2}{n-p}$ é o estimador da variância σ^2. Logo,

$$t_i = \frac{r_i}{s(1-h_{ii})^{1/2}},\ i = 1, \ldots, n\ . \quad (2.9)$$

Como r_i e s^2 não são independentes, t_i não segue uma distribuição *t* de *Student*, como se poderia esperar. Paula (2004) mostra que o problema da dependência entre r_i e s^2 pode ser contornado, com algumas substituições na equação (2.9). Dessa forma, tem-se que o novo resíduo studentizado que é dado por

$$t_i^* = t_i\left(\frac{n-p-1}{n-p-t_i^2}\right)^{1/2}, \quad (2.10)$$

segue uma distribuição t_{n-p-1}.

A definição de resíduo studentizado para os MLGs é feita de forma análoga à utilizada na regressão normal linear. Observa-se que, para os MLGs, os estimadores dos parâmetros do modelo são dados por

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\widehat{\boldsymbol{W}}\boldsymbol{z},$$

em que $\boldsymbol{z} = \widehat{\boldsymbol{\eta}} + \widehat{\boldsymbol{W}}^{-1/2}\widehat{\boldsymbol{V}}^{-1/2}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$. Assim, pode-se definir o resíduo ordinário por

$$\boldsymbol{r}^* = \widehat{\boldsymbol{V}}^{-1/2}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}). \quad (2.11)$$

Assim, os resíduos standartizados para os MLGs são definidos por

$$t_{Si} = \frac{\phi^{1/2}(y_i - \hat{\mu}_i)}{\sqrt{\hat{v}_i(1-\hat{h}_{ii})}}, \quad (2.12)$$

em que h_{ii} pertence a diagonal da matriz *hat*.

Os resíduos mais utilizados em MLGs são definidos a partir dos componentes da função desvio (DEMÉTRIO, 2002). A versão padronizada (DAVISON e GIGLI, 1989; MCCULLAGH, 1987) é dada por

$$t_{Di} = \frac{\phi^{1/2} d(y_i;\hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}}, \tag{2.13}$$

em que $d(y_i;\hat{\mu}_i) = \pm\sqrt{2}\{y_i(\hat{\theta}_i^0 - \hat{\theta}_i) + [b(\hat{\theta}_i) - b(\hat{\theta}_i^0)]\}^{1/2}$.

Para os modelos logísticos lineares**,** esse resíduo toma a forma particular

$$t_{Di} = -\frac{\{2n_i|\ln \hat{p}_i|\}^{1/2}}{\sqrt{1-\hat{h}_{ii}}},$$

em $0 < y_i < n_i$. Quando $y_i = 0$, tem-se que, quando $y_i = n_i$, em que $\hat{h}_{ii}$ assume a seguinte forma:

$$\hat{h}_{ii} = n_i\hat{p}_i(1-\hat{p}_i)\boldsymbol{x}_i^t\left(\boldsymbol{X}^t\widehat{\boldsymbol{V}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i$$

com $\widehat{\boldsymbol{V}} = diag[n_i\hat{p}_i(1-\hat{p}_i), \ldots, n_k\hat{p}_k(1-\hat{p}_k)]$ e $n_1 + \cdots + n_k = n$.

O resíduo studentizado t_{Si}, para o modelo logístico linear, é dado pela expressão a seguir:

$$t_{Si} = \frac{(y_i - n_i\hat{p}_i)^2}{\sqrt{(1-\hat{h}_{ii})[n_i\hat{p}_i(1-\hat{p}_i)]}}. \tag{2.14}$$

Para os MLGs a influência exercida pelo valor da i-ésima observação no modelo ajustado é medida ao avaliar o impacto sobre o logaritmo da função de verossimilhança $l(\beta, \boldsymbol{y})$ com a retirada desta observação. Para medir a influência das observações nas estimativas dos coeficientes do modelo logístico, Cook e Weisberg (1982) apresentam a medida

$$LD_i = \frac{(y_i - n_i\hat{p}_i)^2}{(1-\hat{h}_{ii})^2[n_i\hat{p}_i(1-\hat{p}_i)]} \tag{2.15}$$

2.3 Técnicas gráficas e avaliação do modelo ajustado

Uma fase importante da análise de diagnóstico é a interpretação de gráficos utilizados para detectar pontos suspeitos de serem aberrantes e/ou influentes. Para os modelos de regressão normal linear, os gráficos mais utilizados são, segundo Paula (2004):

(i) gráfico de t_i^* contra a ordem das observações; usado para detectar pontos aberrantes;

(ii) gráfico de t_i^2 contra os valores ajustados; indica a ocorrência de heterocedasticidade nos dados;

(iii) gráfico de h_{ii} contra a ordem das observações; também é usado para detectar observações
suspeitas de serem aberrantes;

(iv) gráfico das coordenadas do vetor que produz a maior influência local ($\boldsymbol{\ell}_{max}$) e que é dado pelo autovetor normalizado correspondente ao maior autovalor da matriz $\ddot{\boldsymbol{F}}$, que será definida com mais detalhes na seção 3, mais especificamente na fórmula em (3.10), contra a ordem das observações; destaca pontos que podem ser influentes;

(v) gráfico normal de probabilidade com envelopes; indica se há possíveis afastamentos da normalidade na distribuição dos resíduos. É o gráfico de t_i^* contra os valores esperados das estatísticas de ordem da normal padrão $\left(Z'_{(i)}\, s\right)$.

Para os MLGs, as técnicas gráficas mais recomendadas são as seguintes:

(i) gráfico de t_{si} contra os valores ajustados; detecta a ocorrência de pontos discrepantes em relação aos demais; estes pontos podem ser influentes ou de alavanca;

(ii) gráfico de t_{Di}, contra os valores ajustados ou contra a ordem das observações; detecta pontos influentes;

(iii) gráfico de $\hat{h}_{ii}$ contra os valores ajustados ou contra a ordem das observações; detecta observações suspeitas de serem aberrantes;

(iv) gráfico de LD_i , na equação (2.15), contra a ordem das observações; as informações mais discrepantes neste gráfico podem ser pontos de alavanca e/ou influentes;

(v) gráfico normal de probabilidade para t_{Di} com envelopes; fornece indícios de afastamento da suposição sobre a distribuição adotada para a variável resposta.

Uma das formas de avaliação de modelos ajustados, visa a detecção de pontos que se encontram significativamente mais afastados dos demais. Estes pontos, são usualmente denominados de *outlier*, e formam a base para o desenvolvimento de técnicas de diagnósticos do ajuste de modelos de regressão.

Podemos classificar um *outlier*, conforme sua característica, do seguinte modo (ver Paula, 2004):

- Pontos aberrantes: são pontos que têm certa influência sobre os valores ajustados, embora não estejam muito afastados dos demais pontos.
- Pontos de alavanca ou de alto leverage: são pontos que estão mais afastados dos pontos no subespaço gerado pelas colunas da matriz **X**. Estes pontos não influenciam de forma significativa as estimativas dos parâmetros.
- Pontos influentes: são pontos que ao mesmo tempo são de alto *leverage* e aberrantes.

3 Avaliando a influência local por meio da curvatura normal

3.1 Influência

Considera-se a função de verossimilhança para o parâmetro β a seguir

$$l_{\boldsymbol{\omega}}(\beta, y) = \sum_{j=1}^{n} \omega_j \, l(\beta, y_j)$$

em que $l(\beta, y_j)$ é o logaritmo da função de verossimilhança correspondente à j-ésima observação e ω_j é um tipo de perturbação, tal que $0 \leq \omega_j \leq 1$. Quando $\omega_1 = \omega_2 = \cdots = \omega_n = 1$, significa que não há perturbação no modelo e quando $\omega_j = 0$ significa que a j-ésima observação foi excluída.

A medida de influência mais conhecida é a distância de Cook, definida por

$$D_{\boldsymbol{\omega}} = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}\right)^t (\boldsymbol{X^t X})\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}\right)}{ps^2} \,,$$

em que $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}} = (\boldsymbol{X^t \Delta X})^{-1} \boldsymbol{X^t \Delta y}$ e $\Delta = diag\{\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega}_n\}$. Esta medida mede quanto a perturbação $\boldsymbol{\omega} = (\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega}_n)^{\boldsymbol{t}}$ afasta $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$ de $\widehat{\boldsymbol{\beta}}$, segundo a métrica $\boldsymbol{X^t X}$.

Quando o i-ésimo ponto é excluído, Paula (2004) mostra que a distância de Cook fica expressa por

$$D_i = t_i^2 \, \frac{h_{ii}}{(1-h_{ii})} \frac{1}{p} \quad . \tag{3.1}$$

Para os MLGs a influência exercida pela i-ésima observação no modelo ajustado é medida ao avaliar o impacto sobre o logaritmo da função de verossimilhança $l(\boldsymbol{\beta}, \boldsymbol{y})$ com a retirada desta observação.

Uma medida utilizada para avaliar este impacto é conhecida como afastamento pela verossimilhança (*Likelihood Displacement*) que, segundo Cook e Weisberg (1982), é definida por

$$LD_i \cong 2\left[l\left(\widehat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{y}\right) - l(\widehat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{y})\right], \tag{3.2}$$

em que $l(\widehat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{y})$é a log-verossimilhança sem a i-ésima observação. Em Demétrio (2002) é mostrado que

$$LD_i \cong \left[\frac{\widehat{h}_{ii}}{1-\widehat{h}_{ii}}\right] t_{Si}^2. \tag{3.3}$$

Para medir a influência das observações nas estimativas dos coeficientes do modelo logístico, a versão da medida afastamento pela verossimilhança utilizada é dada por

$$LD_i \cong \frac{(y_i - n_i\hat{p}_i)^2}{\left(1-\hat{h}_{ii}\right)^2[n_i\hat{p}_i(1-\hat{p}_i)]}.$$

3.2 A curvatura normal em um gráfico de influência

Cook (1986) utiliza a medida afastamento pela verossimilhança para avaliar a influência local, que é definida por

$$LD(\boldsymbol{\omega}) = 2\{l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}})\}, \tag{3.4}$$

quando $\boldsymbol{\omega}$ varia numa vizinhança de $\boldsymbol{\omega}_0$. Observe que a equação em (3.4), na verdade, relaciona a estatística $D_i(\boldsymbol{\omega})$ de Cook e a medida afastamento pela verossimilhança definida em (3.2). Decorre daí que um gráfico de $LD(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contém informações essenciais da influência do esquema de perturbação utilizado. Esse gráfico pode ser interpretado como a superfície geométrica $(q+1)-$ dimensional formada pelos valores do vetor

$$\alpha(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ LD(\boldsymbol{\omega}) \end{pmatrix} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \vdots \\ \omega_q \\ LD(\boldsymbol{\omega}) \end{pmatrix},$$

quando $\boldsymbol{\omega}$ varia em Ω. A superfície definida acima é chamada de gráfico de influência.

Nesta seção a atenção centra-se no método de diagnóstico da influência local, avaliado por meio da curvatura normal. Este método, consiste em se avaliar o comportamento da curvatura normal em torno de um ponto $\boldsymbol{\omega}_0$ pertencente ao gráfico de influência $\alpha(\boldsymbol{\omega})$, lembrando que $\boldsymbol{\omega}_0$ corresponde ao ponto em que $l(\widehat{\boldsymbol{\beta}}) = l(\boldsymbol{\beta}|\boldsymbol{\omega}_0)$. Para o caso particular de ter-se $q = 1$, o gráfico de influência se reduz a $\alpha(\boldsymbol{\omega}) = \{\alpha_i(\boldsymbol{\omega})\}, i = 1,2$, o que equivale à curva plana $\alpha(\omega) = (\omega, LD(\omega))$, cuja curvatura é dada por Stoker (1969, p. 26),

$$C = \frac{|\dot{\alpha}_1\ddot{\alpha}_2 - \dot{\alpha}_2\ddot{\alpha}_1|}{\left(\dot{\alpha}_1^{\,2} + \dot{\alpha}_2^{\,2}\right)^{3/2}}, \tag{3.5}$$

em que as notações $\dot{\alpha}$ e $\ddot{\alpha}$ denotam a primeira e a segunda derivadas de α, respectivamente.

Quando $q > 1$, um gráfico de influência é uma superfície em $\mathbb{R}^{q+1}$ e, dessa forma, a noção de curvatura torna-se mais complexa. Estamos interessados em avaliar como a superfície $\alpha(\boldsymbol{\omega})$ desvia-se de seu plano tangente em torno do ponto $\boldsymbol{\omega}_0$ e, para isso, será considerado o comportamento das seções normais da superfície próxima a esse ponto. As seções normais são obtidas pelas interseções entre a superfície $\alpha(\boldsymbol{\omega})$ e os planos contendo o vetor normal ao seu plano tangente em $\boldsymbol{\omega}_0$. As curvaturas dessas seções normais são denominadas de curvaturas normais (STOKER, 1969).

Para obter uma seção normal, considere uma direção arbitrária em $\mathbb{R}^q$, que será representada por um vetor ℓ de comprimento unitário, e uma reta em $\Omega \subset \mathbb{R}^q$ passando por $\boldsymbol{\omega}_0$, que denotaremos por

$$\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\,\ell\,, \text{ com a} \in R.$$

Essa reta gera uma linha projetada (*lifted line*) sobre a superfície $\alpha(\boldsymbol{\omega})$ passando pelo ponto $\alpha(\boldsymbol{\omega}_0)$, que consiste em uma curva sobre essa superfície. Cada direção ℓ especifica uma linha projetada e vice-versa, pois o plano tangente à superfície em $\boldsymbol{\omega}_0$ é gerado pelas colunas da matriz **V** cujos elementos são $\frac{\partial\, \alpha_i(\boldsymbol{\omega})}{\partial\, \boldsymbol{\omega}_j}$, $i = 1,2,\ldots,q+1; j = 1,2,\ldots,q$, com todas as derivadas avaliadas em $\boldsymbol{\omega}_0$.

3.3 A Curvatura normal de uma $lifted\ line$

A $lifted\ line$ (linha projetada) em uma dada direção ℓ consiste na curva plana dada por $\rho(a) = \left(\rho_1, \rho_2\right)^t = [a, LD\boldsymbol{\omega}(a)]^t$ e pode ser vista como o conjunto dos pontos de $\alpha(\boldsymbol{\omega})$, gerado pelos vetores $\boldsymbol{b}_{q+1}$ e $\left(\ell^t, 0\right)^t$. A curvatura normal da curva $\rho(a)$, na direção ℓ, adaptada de (3.5) é escrita da seguinte forma:

$$C_\ell = \frac{\left|\dot{\rho}_1 \ddot{\rho}_2 - \dot{\rho}_2 \ddot{\rho}_1\right|}{\left(\dot{\rho}_1^2 + \dot{\rho}_2^2\right)^{3/2}}, \tag{3.6}$$

avaliada em $\boldsymbol{\omega}_0$, seja $\boldsymbol{b}_{q+1}$ o vetor da base canônica de $\mathcal{R}^{q+1}$, com 1 na última posição e zero nas outras.

Para melhor avaliar esta expressão, deve-se lembrar da suposição de que a função de log-verossimilhança perturbada $l(\boldsymbol{\beta}, \boldsymbol{\omega})$ é de classe C^2 em $(\boldsymbol{\beta}^t, \boldsymbol{\omega}^t)^t$ e, para a superfície $\rho(a) = \left(\rho_1, \rho_2\right)^t = [a, LD\boldsymbol{\omega}(a)]^t$, tem-se o seguinte:

- $\dot{\rho}_1 = \frac{da}{da} = 1$, $\ddot{\rho}_1 = \frac{d^2a}{da^2} = 0$;
- $\frac{d}{da}\boldsymbol{\omega}(a) = \frac{d}{da}\left(\boldsymbol{\omega}_0 + a\,\ell\right)$, logo $\frac{d^2}{da^2} = 0$;

- Com algumas transformações encontra-se:

$$C_{\ell} = \left| \ddot{\rho}_{2} \right| = \left| \ddot{LD}\{\boldsymbol{\omega}(a)\} \right|,$$

avaliada em $\alpha = 0$. Utilizando novamente a regra da cadeia, segue que

$$\ddot{LD}\{\boldsymbol{\omega}(a)\} = \ell^{t} \frac{\partial^2}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{t}} LD\{\boldsymbol{\omega}(a)\}\, \ell,$$

pois a matriz hessiana é simétrica. Avaliando-se a derivada em $a = 0$, obtém-se a curvatura normal do gráfico de influência $\alpha(\boldsymbol{\omega})$, que é dada por

$$C_{\ell} = \left| \ell^{t} \ddot{LD}(\boldsymbol{\omega}_0)\, \ell \right|. \tag{3.7}$$

Pode-se mostrar que,

$$F = \ddot{\Delta}^{t} \ddot{L} \Delta \tag{3.8}$$

em que $\ddot{\boldsymbol{F}}$ é uma matriz quadrada de dimensão $q \times q$, cujos elementos são dados pelas derivadas $\frac{\partial^2 l(\hat{\boldsymbol{\beta}}_{\boldsymbol{\omega}})}{\partial \omega_r \partial \omega_s}$, avaliadas em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, temos $\Delta = \frac{\partial^2 \ell\ (\beta,\omega)}{\partial \beta\ \partial \omega^{t}}$, $\boldsymbol{J} = -\ddot{\boldsymbol{L}}^{-1} + \Delta_{is} = \boldsymbol{0}$ e $J = \frac{\partial \hat{\beta}_{\omega}}{\sigma \omega_r}$.

Observe que, pela definição do afastamento pela verossimilhança, tem-se que portanto, a curvatura normal vista em (4.6) fica resumida por

$$C_{\ell} = 2\left[\ell^{t} \ddot{\boldsymbol{F}}\, \ell \right], \tag{3.9}$$

Uma forma operacional para a curvatura normal, utilizando-se o resultado (3.8) na equação (3.9), obtém-se a seguinte forma de curvatura normal para o gráfico de influência $\boldsymbol{\alpha}(\boldsymbol{\omega})$, na direção de um vetor unitário ℓ:

$$C_{\ell} = 2\left| \ell^{t}\ \Delta^{t}\ \ddot{\boldsymbol{L}}^{-1}\ \Delta\ \ell \right|. \tag{3.10}$$

Com a equação (3.10) pode-se avaliar a influência que pequenas perturbações produzem sobre os componentes do modelo. A direção que produz maior influência local, ℓ_{max}, é o autovetor normalizado correspondente ao maior autovalor C_{max} da matriz $\ddot{\boldsymbol{F}}$. Por meio dele identificamos os fatores mais influentes para o esquema de perturbação em análise. Similarmente, os autovetores associados com os autovalores intermediários

podem também ser utilizados para investigar o comportamento da superfície em direções relativas a curvaturas menos extremas.

A curvatura normal para o modelo de perturbação no modelo normal linear é dada por,

$$C_l = \frac{2}{\hat{\sigma}^2}\ell^t \left[\boldsymbol{D(e)P_X D(e)} + \frac{\boldsymbol{e_{sq}e_{sq}^t}}{2n\hat{\sigma}^2}\right] \ell, \qquad (3.11)$$

em que $\boldsymbol{e_{sq}} = (e_1^2.\cdots, e_n^2)$, $\boldsymbol{P_X = X(X^tX)^{-1}X^t}$, $\boldsymbol{D(e) = diag(e_i)}$, vale ressaltar que não são conhecidas expressões analíticas para autovetores ℓ_{max} da matriz $\ddot{F}$. Segue que, substituindo-se o resultado da matriz $\ddot{F}$ na equação (3.8), obtém-se a curvatura normal para avaliar a perturbação de observações inviduais e regressão logística,

$$C_\ell = 2|\ell^t \boldsymbol{D}\{x\}\boldsymbol{P_W D}\{x\}\ell|. \qquad (3.12)$$

Em que $\boldsymbol{P_W = W(W^tW)^{-1}W^t}$ e $\boldsymbol{D\{x\}} = diag\{x_i\}$.

4 Avaliando a influência local através da inclinação máxima de uma *lifted line*

A abordagem de Cook para diagnóstico de influência local apresenta certas dificuldades práticas, em algumas situações. Billor e Loynes (1993) destacam quatro dessas dificuldades, listadas a seguir:

(i) A escolha de um referencial para comparar com o valor da curvatura máxima obtida não é bem clara. O valor 2 foi sugerido por Cook (1986) como um referencial, ou seja, se a curvatura exceder 2, indica sensibilidade local, porém, esse valor é independente não apenas dos dados e do modelo como também das dimensões n, p e q. Portanto, não parece convincente a escolha desse valor como referencial.

(ii) o cálculo da curvatura máxima (C_{max}). Em situações relativamente simples, não há uma fórmula explícita disponível para a curvatura máxima, por exemplo, no caso de regressão normal linear;

(iii) a falta de invariância da curvatura sob reparametrização do esquema de perturbação. Schall e Dunne (1992) propõem uma solução para o problema de falta de invariância da curvatura máxima, sob reparametrização do esquema de perturbação, usando a relação entre o fator de inflação da variância e a curvatura máxima;

(iv) a falta de definição dos parâmetros. Billor e Loynes (1993) mostram, por meio de um exemplo prático, uma dificuldade decorrente da falta de definição dos parâmetros.

Considere o modelo de regressão normal linear

$$y_i = x_i^t\beta + \sigma_i\varepsilon_i, i = 1, \ldots, n$$

e as duas versões de um esquema de perturbação a seguir:

a) Versão 1:

$$\sigma_i = \begin{cases} \sigma(1+\omega), i = 1 \\ \sigma\ , i \geq 2 \end{cases} \quad (4.1)$$

b) Versão 2:

$$\sigma_i = \begin{cases} \sigma\ , i = 1 \\ \frac{\sigma}{1-\omega}\ , i \geq 2 \end{cases} \quad (4.2)$$

Quando σ não é conhecido, os modelos (4.1) e (4.2) são idênticos. Contudo, Billor e Loynes (1993) constataram a ocorrência de certa diferença no cálculo da curvatura máxima para as duas versões apresentadas.

4.1 Medida de afastamento pela verossimilhança modificada

A medida apresentada por Billor e Loynes (1993) para medir a influência local, que definiremos mais adiante, tem, de imediato, a vantagem de estar de acordo com a falta de definição dos parâmetros, como discutido no início deste capítulo.

Considere a medida afastamento pela verossimilhança modificada, definida por

$$LD^*(\boldsymbol{\omega}) = -2\left[l(\widehat{\boldsymbol{\theta}}) - \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}\right)\right], \quad (4.3)$$

em que $\widehat{\boldsymbol{\theta}}$ **e** $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ são os estimadores de máxima verossimilhança do vetor de parâmetros desconhecidos $\boldsymbol{\theta}$**,** sob os modelos original e perturbado, respectivamente.

Note que, com a medida LD^*, o problema da escolha dos parâmetros desaparece, visto que $l\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}\right)$ é a máxima verossimilhança sob o modelo perturbado, a qual independe de uma particular escolha de parametrização.

Em relação à deleção de observações, Billor e Loynes (1993) levantam a seguinte questão: se nosso interesse é perturbar o modelo, devemos conservar os dados inalterados, mas, no caso de deleção de observações, ocorre, na verdade, uma modificação na dimensão dos dados. Sendo assim, a omissão de um ponto dos dados parece ser mais do que uma simples perturbação, podendo ser mais bem entendida como outro tipo de operação.

Para lidar com essa dificuldade, Billor e Loynes (1993) usam o fato de que as estimativas dos parâmetros $\widehat{\boldsymbol{\beta}}_{(i)}$ e $\hat{\sigma}^2_{(i)}$, baseados no conjunto dos dados sem a i-ésima observação, são iguais às obtidas quando usamos o modelo *outlier mean shift* (deslocamento de média) com o conjunto dos dados completo.

Considere o modelo de regressão linear, com todas as suposições usuais, exceto que a variância dos erros seja conhecida e igual a σ^2. A influência local de Cook foi motivada pela estatística D_i, definida na equação (3.2). Para a medida afastamento pela verossimilhança LD^*, definimos a estatística D_i^* como uma análoga natural da estatística de Cook, D_i. Dessa forma, D_i^* mede a diferença entre $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$, em que $\widehat{\boldsymbol{\beta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\beta}$**.**

Se $l(\boldsymbol{\beta})$ denota a função de log-verossimilhança do modelo original e $l_{ms}(\boldsymbol{\beta}, \phi)$ é a log-verossimilhança do modelo deslocamento de média, dado por

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{d}_i\phi + \boldsymbol{\varepsilon} \quad (4.4)$$

em que $\boldsymbol{d_i}$ é um vetor $n \times 1$, com o i-ésimo elemento igual a 1 e os restantes iguais a zero, defina

$$LD_i^* = -2\left[l\left(\widehat{\boldsymbol{\beta}}\right) - l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right)\right]. \tag{4.5}$$

A função de log-verossimilhança para o modelo deslocamento de média é

$$l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right) = -\frac{1}{2\sigma^2}\sum_{j\neq 1}\left(y_j - \boldsymbol{x}_j^t\boldsymbol{\beta}\right)^2. \tag{4.6}$$

Derivando l_{ms} em relação ao vetor $\boldsymbol{\beta}$**,** e igualando-se a zero a derivada**,** tem-se que

$$\widehat{\boldsymbol{\beta}}_{ms} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} - (\boldsymbol{X}^t\boldsymbol{X})^{-1}\phi\boldsymbol{x}_i.$$

Considerando que $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$**,** obtém-se o estimador de máxima verossimilhança do vetor de parâmetros **β**, para o modelo deslocamento de média

$$\widehat{\boldsymbol{\beta}}_{ms} = \widehat{\boldsymbol{\beta}} - \phi(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i. \tag{4.7}$$

Agora vamos obter o estimador de máxima verossimilhança do parâmetro ϕ. Segue que, derivando-se a equação (4.6) em relação a ϕ, tem-se

$$\frac{\partial l_{ms}}{\partial \phi} = -\frac{1}{\sigma^2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta} - \phi)(-1).$$

Igualando-se a zero este resultado e substituindo-se o valor de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ms}}$ dado na equação (4.7) no resultado, obtém-se que

$$\hat{\phi} = y_i - \hat{\mu}_i + \hat{\phi}h_{ii},$$

em que $\hat{\mu}_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$ é o valor predito para y_i da regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ e $h_{ii} = \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i$ é o elemento (i,i) da matriz hat. Pode-se, ainda, escrever

$$\hat{\phi} = r_i + \hat{\phi}h_{ii} = \frac{r_i}{1-h_{ii}}, \tag{4.8}$$

em que $r_i = y_i - \hat{\mu}_i$ é o i-ésimo resíduo ordinário da regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$**.**

Observe que se pode escrever o estimador de máxima verossimilhança do vetor de parâmetros **β,** sem a i-ésima observação no modelo, como

$$\widehat{\boldsymbol{\beta}}_{(i)} = \left(\boldsymbol{X}^t{}_{(i)}\boldsymbol{X}_{(i)}\right)^{-1}\boldsymbol{X}^t{}_{(i)}\boldsymbol{y}_{(i)}. \tag{4.9}$$

em que $X_{(i)}$ é a matriz de variáveis explicativas sem a i-ésima linha e $Y_{(i)}$ é o vetor de variável resposta sem o i-ésimo elemento. Segue que

$$\boldsymbol{X}^t{}_{(i)}\boldsymbol{X}_{(i)} = \boldsymbol{X}^t\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^t.$$

Aplicando-se o resultado do teorema em (COOK; WEISBERG, 1982) ao segundo membro da equação acima, tem-se que

$$\left(\boldsymbol{X}^t{}_{(i)}\boldsymbol{X}_{(i)}\right)^{-1} = (\boldsymbol{X}^t\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i}.$$

Substituindo-se o resultado acima na equação (4.9) e fazendo algumas transformações, tem-se

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{r_i}{1-h_{ii}}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i. \tag{4.10}$$

Vamos considerar, agora, apenas a parte relevante da log-verossimilhança do modelo original. Assim,

$$l(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2.$$

Segue daí que

$$l\left(\widehat{\boldsymbol{\beta}}\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i{}^2. \tag{4.11}$$

Do mesmo modo, considere-se, agora, apenas a parte relevante da log-verossimilhança do modelo deslocamento de média em (4.6). Tem-se que

$$l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right) = -\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta} - \phi)^2 - \frac{1}{2\sigma^2}\sum_{j\neq 1}\left(y_i - \boldsymbol{x}_j^t\widehat{\boldsymbol{\beta}}_{(i)}\right)^2.$$

Substituindo-se os resultados em (4.8) e (4.10) na equação acima e operando algumas transformações, tem-se

$$l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right) = -\frac{1}{2\sigma^2}\{r_i - r_i\}^2 - \frac{1}{2\sigma^2}\sum_{j\neq 1}\left\{r_j + \frac{r_i}{1-h_{ii}}h_{ij}\right\}^2.$$

em que $h_{ij} = \boldsymbol{x}_j^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i$ é o elemento (i, j) da matriz *hat*. Segue que

$$l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right) = -\frac{1}{2\sigma^2}\{r_i - r_i\}^2 - \frac{1}{2\sigma^2}\sum_{j\neq 1}\left\{r_j + \frac{r_i}{1-h_{ii}}h_{ij}\right\}^2.$$

Somando e subtraindo a parcela $\frac{1}{2\sigma^2}\left\{r_i + \frac{r_i}{1-h_{ii}}h_{ii}\right\}^2$ no resultado acima tem-se que

$$l_{ms}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \phi\right) = \frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(r_j + \frac{r_i}{1-h_{ii}}h_{ij}\right)^2 + \frac{1}{2\sigma^2}\left(r_i + \frac{r_i}{1-h_{ii}}h_{ii}\right)^2. \quad (4.12)$$

Substituindo em (4.11) e (4.12) na equação (4.5), segue que

$$LD_i^* = \frac{1}{\sigma^2}\left\{-2\frac{r_i}{1-h_{ii}}\sum_{j=1}^{n}h_{ij}r_j + \frac{r_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n}h_{ij}^2 + \frac{r_i^2}{(1-h_{ii})^2}\right\}. \quad (4.13)$$

Observe que

$$\begin{aligned}\boldsymbol{Hr} &= \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}\end{aligned}$$

Logo, todos os elementos do produto **Hr** são iguais a zero. Ou seja,

$$\sum_{j=1}^{n}h_{ij}r_j = 0, \forall i. \quad (4.14)$$

Além do mais, como **H** é uma matriz simétrica e idempotente, tem-se que

$$\sum_{j=1}^{n}h_{ij}^2 = h_{ii}. \quad (4.15)$$

Substituindo-se os resultados em (4.14) e (4.15) na equação (4.13), tem-se

$$LD_i^* = \frac{r_i^2}{\sigma^2(1-h_{ii})}. \quad (4.16)$$

Note que a medida acima depende apenas de duas quantidades, os resíduos r_i e os elementos da diagonal principal da matriz *hat.*

Na seção (3), viu-se que uma motivação para a influência local de Cook (1986) é a relação, entre a estatística D_i e a medida de afastamento pela verossimilhança LD, visto na subseção (3.1). Billor e Loynes (1993) consideram uma relação análoga, entre a estatística D_i^* e a medida afastamento pela verossimilhança modificada LD^*, dada por

$$pD_i^*(\boldsymbol{\omega}) = -2\left[l\left(\widehat{\boldsymbol{\beta}}\right) - l\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}\right)\right]. \quad (4.17)$$

Por meio de um raciocínio semelhante ao utilizado por Cook (1986) para estudar a influência local, tem-se que $\left(\boldsymbol{\omega}, LD^*(\boldsymbol{\omega})\right)$ representa uma superfície $q+1$ dimensional e, com isso, é possível estabelecer linhas projetadas (*lifted line*) em várias direções dessa superfície. Considerando a reta em $\mathbb{R}^q$ definida por $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\boldsymbol{\ell}$, em que $\boldsymbol{\omega}_0$ representa um vetor de não perturbação do modelo, $\boldsymbol{\ell}$ é um vetor fixo em $\mathbb{R}^q$ e a é um

escalar, segue que o comportamento de $LD^*(\boldsymbol{\omega}_0 + a\boldsymbol{\ell})$ proporciona informações importantes sobre a influência do esquema de perturbação adotado.

Como a primeira derivada de LD^* avaliada em $\boldsymbol{\omega}_0$ não se anula, exceto para alguns valores particulares do vetor $\boldsymbol{\ell}$, a primeira derivada de LD^* em relação a $\boldsymbol{\ell}$ produz informações valiosas sobre o comportamento local de $LD^*(\boldsymbol{\omega}_0)$. Em particular, a inclinação máxima e a correspondente direção máxima, denotadas, respectivamente, por d^*_{max} e $\boldsymbol{\ell}^*_{max}$, são importantes quando $q \neq 1$, em que q é a dimensão do vetor $\boldsymbol{\omega}$ (BILLOR e LOYNES, 1993).

A direção máxima da função LD^* no ponto $\boldsymbol{\omega}_0$ ocorre na direção do gradiente $\nabla LD^*(\boldsymbol{\omega}_0)$. Logo, a inclinação máxima (d^*_{max}) é a medida por (LIMA, 1981)

$$d^*_{max} = \|\nabla LD^*(\boldsymbol{\omega}_0)\|.$$

Pela regra da cadeia,

$$\frac{\partial LD^*}{\partial a} = \boldsymbol{\ell}^t \left\{ -2 \left[\frac{\partial l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\omega}} - \left(\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \frac{\partial l\boldsymbol{\beta}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}} + \frac{\partial l(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \right) \right] \right\}.$$

De imediato, tem-se que $\frac{\partial l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\omega}} = 0$ e que $\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}}$, quando avaliada em $\widehat{\boldsymbol{\beta}}$ e $\boldsymbol{\omega}_0$, também se anula. Assim, a equação acima fica resumida a

$$\left.\frac{\partial LD^*}{\partial a}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0} = 2\boldsymbol{\ell}^t \left.\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}.$$

Portanto, a inclinação máxima pode ser escrita como

$$d^*_{max} = 2\left\|\nabla l\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{\omega}\right)\right\|. \tag{4.18}$$

Para uma aplicação dessa teoria, vamos considerar o modelo perturbado de regressão normal linear dado a seguir

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

em que $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{W}^{-1}$ e $\boldsymbol{W} = diag(1 + \omega_1, 1, 1, \ldots, 1)$, com o modelo não perturbado sendo obtido quando $\omega_1 = 0$. A inclinação da linha projetada sobre a superfície $(\boldsymbol{\omega}, LD^*(\boldsymbol{\omega}_0))$ é calculada por meio da primeira derivada de $LD^*(\boldsymbol{\omega})$ em relação a ω_1. A log-verossimilhança para o modelo acima fica estabelecida por

$$l(\boldsymbol{\beta}, \boldsymbol{\omega}) = -\frac{1}{2\sigma^2} \sum_{i=2}^{n} (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2,$$

segue daí que

$$\frac{\partial LD^*}{\partial \omega_1} = 2\left[\frac{1}{2(1+\omega_1)} - \frac{(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2}{2\sigma^2}\right].$$

Assim,

$$\left.\frac{\partial LD^*}{\partial \omega_1}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{\omega_1}=\mathbf{0}} = 1 - \frac{r_i^2}{\sigma^2}, \tag{4.19}$$

em que $r_1 = y_1 - \boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}}$ é o resultado ordinário relativo à primeira observação da regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$. Têm-se as seguintes considerações para o resultado do resíduo r_1 acima:

1. se $|r_1| = \sigma$, não existe razão para preferir o modelo perturbado ao não perturbado e vice-versa,
2. se $|r_1| \ll \sigma$ deve-se optar pelo modelo perturbado caso $\omega_1 < 0$, isto é, se $Var(y_1) < \sigma^2$.

Observa-se, ainda, que $1 - \frac{r_i^2}{\sigma^2}$ tem uma cota superior finita quando $|r_1| \to 0$, enquanto sua cota é inferior quando $|r_1| \to +\infty$ e $-\infty$. Essa assimetria está inteiramente de acordo com a teoria vista (BILLOR e LOYNES, 1993).

Agora, considere-se a matriz de perturbação geral dada por

$$\boldsymbol{W}_g = diag(1+\omega_1, 1+\omega_2, \ldots, 1+\omega_n),$$

para o mesmo modelo anterior. Sabe-se que a taxa de variação da função LD^* é medida pelo gradiente

$$\nabla LD^*(\boldsymbol{\omega}_0) = \left(\frac{\partial LD^*}{\partial \omega_1}, \frac{\partial LD^*}{\partial \omega_2}, \ldots, \frac{\partial LD^*}{\partial \omega_n}\right).$$

Utilizando-se um procedimento análogo ao realizado para obter o resultado em (4.19), tem-se que

$$\left.\frac{\partial LD^*(\boldsymbol{\omega})}{\partial \omega_i}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega_0}} = 1 - \frac{r_i^2}{\sigma^2}$$

quando $\omega_i = 0$ com $i = 1, \ldots, n$. Dessa forma, a razão máxima de crescimento de LD^* é na direção do vetor gradiente $\nabla LD^*(\boldsymbol{\omega}_0)$, ou seja,

$$\begin{aligned} d_{max}^* &= \|\nabla LD^*(\boldsymbol{\omega}_0)\| \\ &= \sum_{j=1}^{n}\left(1 - \frac{r_i^2}{\sigma^2}\right)^2. \end{aligned} \tag{4.20}$$

4.2 Avaliando a inclinação máxima d_{max}^{*}

Uma das dificuldades na interpretação dos resultados em um estudo da influência local, na abordagem de Cook (1986), é justamente o de determinar um valor de referência para comparar com a curvatura máxima obtida. Segundo Billor e Loynes (1993), essa questão deve ser respondida pelo julgamento e a intuição em escolher um valor apropriado para servir de referência, e que seja adequado ao problema.

Billor e Loynes (1993) esclarecem que existem várias maneiras para determinar tal referência. Por exemplo, considere-se o modelo de regressão normal linear perturbado pela matriz $\boldsymbol{W}_g = diag(1+\omega_1, 1+\omega_2, \ldots, 1+\omega_n)$, já mencionado anteriormente. Já foi demonstrado que, para esse problema, tem-se

$$d_{max}^{*} = \mathrm{A}^{\frac{1}{2}},$$

em que $A = \sum_{j=1}^{n}\left(1-\frac{r_i^2}{\sigma^2}\right)^2$. Esta expressão é útil para avaliar os momentos de A, porém, para a obtenção de um valor de referência, ela se torna um tanto complicada. Segue daí que, substituindo-se os resíduos r_j em A pelos respectivos erros aleatórios ε_j, os quais são assumidos variáveis aleatórias independentes normalmente distribuídas, com média zero e variância σ^2, pode-se obter a esperança e a variância de A, como segue. Tem-se que

$$A = \sum_{j=1}^{n}\left[\left(\frac{\epsilon_j}{\sigma}\right)^2 - 1\right]^2.$$

Note que

$$\epsilon_j \sim N(0,\sigma^2) \Rightarrow \left(\frac{\epsilon_j}{\sigma}\right)^2 \sim X_{(1)}^2, \forall j$$

em que $X_{(1)}^2$ representa uma distribuição qui-quadrado com um grau de liberdade. Como $E\left(X_{(1)}^2\right) = 1$, segue que

$$E(A) = \sum_{j=1}^{n} Var\left(\frac{\epsilon_j}{\sigma}\right)^2 = 2n, \tag{4.21}$$

uma vez que $Var\left(X_{(1)}^2\right) = 2$

Por meio de algumas proposições em GRAYBLL (1983, P.189), chegamos ao resultado:

$$E[A^2] = 56n + 4n^2 \tag{4.22}$$

Substituindo-se os resultados em (4.21) e (4.22) na fórmula da variância a seguir, tem-se que

$$\begin{aligned} Var[A] &= E[A^2] - (E[A])^2 \\ &= 56n + 4n^2 - (2n)^2 \\ &= 56n\,. \end{aligned} \tag{4.23}$$

Dessa forma, um referencial conveniente seria, então, dado por $E[A] + 2DP[A]$, em que $DP[A]$ representa o desvio padrão de A. Segue que, para o nosso exemplo, poderia se utilizar como referencial a quantidade

$$2n + 4\sqrt{14n}.$$

Para uma melhor aproximação, pode-se usar uma transformação de A. Porém, o valor $2n + 4\sqrt{14n}$ pode ser considerado como uma referência para avaliar a influência local do problema proposto.

4.3 Inclinação máxima no caso de perturbação de observação no modelo normal linear

Considerando apenas a parte relevante da função de log-verossimilhança perturbada como dada em (3.35), tem-se que

$$l(\boldsymbol{\beta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \frac{1}{2\sigma^2}\omega_i \left(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}\right)^2, \tag{4.24}$$

em que ω_i e y_i são as i-ésimas componentes dos vetores $\boldsymbol{\omega}$ e $\boldsymbol{y}$, respectivamente.

Derivando a função (4.24) em relação a ω_i, tem-se que

$$\frac{\partial l}{\partial \omega_i} = \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2.$$

Aplicando-se a derivada acima em $\widehat{\boldsymbol{\beta}}$ e $\hat{\sigma}^2$, tem-se que

$$\left.\frac{\partial l}{\partial \omega_i}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\sigma^2=\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^2} r_i^2.$$

Da subseção (4.1), tem-se que

$$\begin{aligned} \frac{\partial LD^*}{\partial \omega_i} &= 2\left.\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \omega_i}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\sigma^2=\hat{\sigma}^2} \\ &= \frac{r_i^2}{\hat{\sigma}^2}. \end{aligned} \tag{4.25}$$

Vimos ainda, na subseção (4.1), que a direção da inclinação máxima $\boldsymbol{\ell}_{max}^*$ é a direção do gradiente. Assim, pode-se escrever

$$\boldsymbol{\ell}_{max}^* = \nabla LD^*(\boldsymbol{\omega}_0) = \left(\frac{r_1^2}{\hat{\sigma}^2}, \ldots, \frac{r_n^2}{\hat{\sigma}^2}\right). \tag{4.26}$$

Segue que a inclinação máxima é obtida por

$$d^*_{max} = \|\nabla LD^*(\boldsymbol{\omega}_0)\| = \sqrt{\sum_{j=1}^{n}\left(\frac{r_j^2}{\hat{\sigma}^2}\right)^2}. \quad (4.27)$$

Para obter um valor de referência para a inclinação máxima acima, pode-se desenvolver um raciocínio análogo ao realizado na seção (4.2). Seja $A = \sum_{j=1}^{n}\frac{r_1^2}{\hat{\sigma}^2}$, substituindo em A os $r_{j's}$ pelos erros $\varepsilon_{j's}$ que são assumidos variáveis aleatórias independentes normalmente distribuídas com média zero e variância σ^2, segue que

$$E[A] = \sum_{j=1}^{n} E\left[x_j\right] \quad = \sum_{j=1}^{n} 1 = n\,. \quad (4.28)$$

Observe que

$$A^2 = \sum_{j=1}^{n} x_j^2 + \sum_{j\neq 1} x_i x_j.$$

Logo,

$$E[A^2] = \sum_{j=1}^{n} E\left[x_j^2\right] + \sum_{j\neq 1} E\left[x_i x_j\right],$$

pela independência entre x_i e x_j, para todo $i \neq j$, tem-se que

$$E[A^2] = \sum_{j=1}^{n} E\left[x_j^2\right] + \sum_{j\neq 1} E[x_i]E\left[x_j\right]. \quad (4.29)$$

Em JAMES (1981), tem-se que a função geradora de momentos de x_j é dada por

$$M''_{x_j}(t) = 3(1-2t)^{-5/2} \Rightarrow M''_{x_j}(0) = E\left[x_j^2\right] = 3. \quad (4.30)$$

Substituindo-se este resultado em (4.29), obtém-se

$$E[A^2] = \sum_{j=1}^{n} 3 + \sum_{j=1}(n-1) = n^2 + 2n. \quad (4.31)$$

Substituindo-se o resultado das equações (4.28) e (4.31) na fórmula da variância a seguir, obtém-se

$$Var[A] = E[A^2] - (E[A^2])^2$$
$$= 2n.$$

Portanto, pode-se considerar, como um valor de referência para a inclinação máxima obtida em (4.27), o resultado de

$$E[A] + 2\sqrt{Var[A]} = n + 2\sqrt{2n}. \quad (4.32)$$

4.4 A inclinação máxima no caso de perturbação de observação no modelo logístico

Considere a função de log-verossimilhança perturbada para uma amostra aleatória de uma variável aleatória com distribuição Binomial (m_i, p_i), dada por

$$l(\boldsymbol{\beta}, \boldsymbol{y}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i y_i \ln\left[\frac{p_i}{1-p_i}\right] + \sum_{i=1}^{n} \omega_i m_i \ln[1-p_i].$$

Derivando a função acima em relação a ω_i, obtém-se

$$\frac{\partial l}{\partial \omega_i} = y_i \ln\left[\frac{p_i}{1-p_i}\right] + m_i \ln[1-p_i].$$

Avaliando-se a expressão acima em $\hat{p}_i$, tem-se

$$\left.\frac{\partial l}{\partial \omega_i}\right|_{p_i=\hat{p}_i} = y_i \ln\left[\frac{\hat{p}_i}{1-\hat{p}_i}\right] + m_i \ln[1-\hat{p}_i].$$

Dessa forma, as coordenadas do vetor gradiente $\boldsymbol{\ell}^*_{max} = \nabla LD^*(\boldsymbol{\omega}_0)$ são dadas por

$$\begin{aligned}\boldsymbol{\ell}^*_i &= \frac{\partial LD^*}{\partial \omega_i} \\ &= 2\left.\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \omega_i}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\sigma^2=\hat{\sigma}^2} \\ &= 2\left\{y_i \ln\left[\frac{\hat{p}_i}{1-\hat{p}_i}\right] + m_i \ln[1-\hat{p}_i]\right\} \quad (4.33)\end{aligned}$$

A inclinação máxima é obtida pela aplicação imediata de

$$d^*_{max} = \|\nabla LD^*(\boldsymbol{\omega}_0)\|. \quad (4.34)$$

5 Resultados e discussões

5.1 Aplicações

Neste capítulo serão discutidos dois exemplos de análise de dados reais, o software usado foi o pacote estatístico **R**. (R CORE TEAM, 2020). O objetivo principal destas aplicações é colocar em prática as técnicas de diagnóstico, com ênfase para a influência local de Cook e a abordagem de Billor e Loynes, descritas nesta tese.

Utiliza-se também algumas técnicas gráficas, para detectar a fuga das suposições adotadas para os modelos sugeridos e a ocorrência de possíveis *outliers,* estes conceitos foram descritos no capítulo 2, para maiores detalhes consultar Paula (2004).

No primeiro exemplo se usou o modelo de regressão normal linear simples para os dados propostos. A análise foi desenvolvida de quatro formas: técnicas gráficas de diagnóstico, curvatura normal de Cook para regressão normal linear, sob perturbação de observação, inclinação máxima proposta por Billor e Loynes, sob perturbação na variância e sob perturbação de observação.

No segundo exemplo se usou o modelo de regressão logística múltipla para os dados apresentados. Analisamos o modelo proposto sob três aspectos: técnicas gráficas de diagnóstico, curvatura normal de Cook para regressão logística e a inclinação máxima de Billor e Loynes, ambos sob esquema de perturbação de observação.

5.2 Estimativa do número de gansos por bando

Métodos de inspeção aérea são regularmente empregados para estimar a população de gansos, durante o verão no Oeste da Baia de Hudson, no Canadá, Weisberg (1985). Para isso, uma pequena aeronave sobrevoa a região e quando avista um bando de gansos, uma pessoa experiente estima a quantidade dessas aves. Para investigar a eficiência deste método um experimento foi conduzido, onde uma aeronave carregando dois observadores sobrevoou 45 bandos e cada um deles estimou independentemente o número de pássaros por bando. Além disso, uma fotografia foi retirada de cada bando e posteriormente uma contagem do número de gansos foi realizada através desta fotografia.

Os dados coletados neste experimento encontram-se em Weisberg (1985, pg. 102). Aplicamos um modelo normal linear simples para os dados, de modo que Y seja o número de gansos por bando, computados a partir da fotografia aérea e X o número de gansos estimado pelo observador l do experimento. O modelo fica, portanto, dado por

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \cdots, 45 \qquad (5.1)$$

com a suposição de que $\varepsilon_i \sim N(0,\ \sigma^2)$, com os $\varepsilon_{i's}$, mutuamente independentes. As estimativas dos parâmetros fornecidas pelo **R,** foram $\hat{\alpha} = 26{,}650$ e $\hat{\beta} = 0{,}883$, com respectivos erros padrões dados por 8,614 e 0,078. Ambos indicando que os coeficientes se apresentaram significativos ao modelo.

A estimativa da variância é dada por 44,41 e a porcentagem de variação de Y explicada pelo modelo, ou seja, o valor do coeficiente de determinação foi de 74,45%.

O diagrama de dispersão *(plot* de X versus Y*)* mostrado na Figura 1, sugere forte evidência de heterocedasticidade. Assim, podemos esperar que a introdução de uma

pequena perturbação revele alguma espécie de sensibilidade no modelo. Na Figura 2 temos o gráfico normal de probabilidade com envelopes para os resíduos t_{Si}, que avalia a validade da suposição de normalidade da variável resposta, para o modelo adotado em (5.1), pode-se notar por meio deste gráfico, um forte indício de não validade desta suposição, pois há uma quantidade razoável de pontos fora das bandas de confiança do gráfico, configurando talvez a ausência de algum termo extra no modelo.

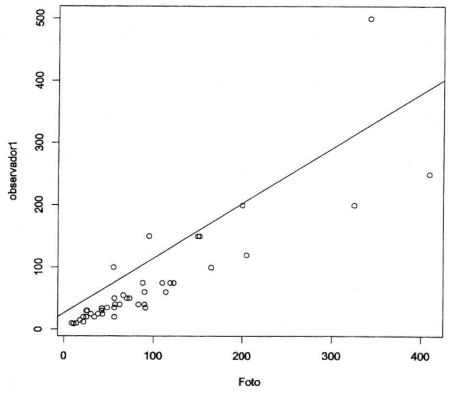

Figura 1 - Número de gansos contados a partir da foto versus o número de gansos estimados pelo observador 1.

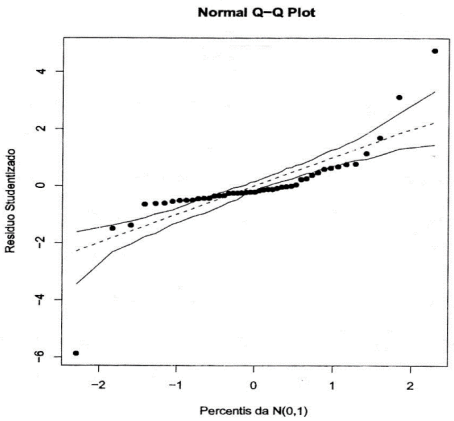

Figura 2 - Gráfico normal de probabilidade com envelopes.

A seguir faremos uma análise de alguns gráficos de diagnósticos, com o objetivo de identificar pontos influentes e de verificar a validade das suposições adotadas para o modelo.

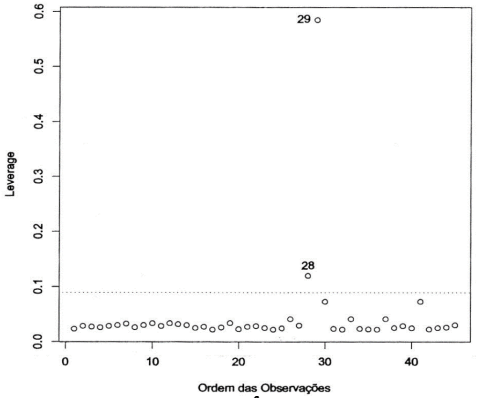

Figura 3 - Leverage $\hat{h}_{ii}$ contra a ordem das observações.

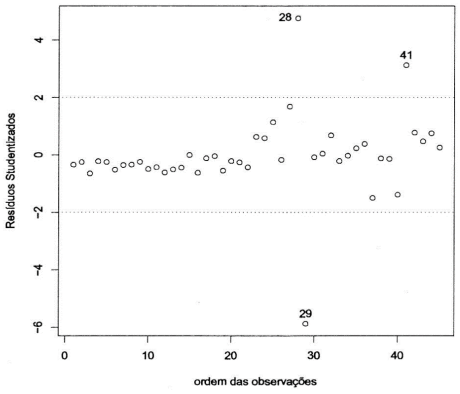

Figura 4 - Resíduos *studentizados* (t_{Si}) contra a ordem das observações.

Um outro gráfico importante a ser interpretado é o gráfico dos valores $\hat{h}_{ii}$ em função das ordens das observações dado na Figura 3, através dele destacamos dois pontos acima da linha $\frac{2*p}{n}$, que correspondem às observações 28 e 29, sendo portanto candidatos a pontos influentes. O gráfico dos resíduos *studentizados* em função das ordens das observações, dado na Figura 4, destaca novamente os pontos 28 e 29, além do ponto 41, sendo estes candidatos a serem pontos aberrantes.

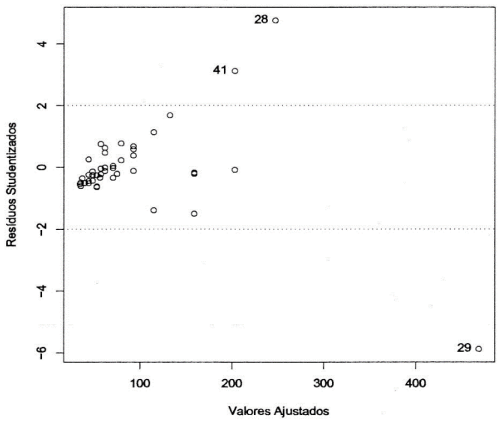

Figura 5 - Resíduos *studentizados* t_{Si} contra os valores ajustados.

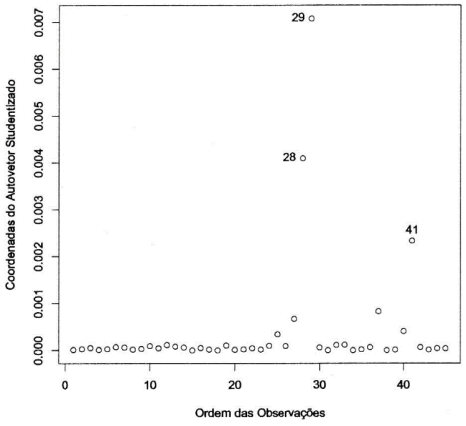

Figura 6 - Gráfico das coordenadas do autovetor d_{max} contra ordem das observações

Na Figura 5 temos o gráfico dos resíduos *studentizados* em função dos valores ajustados, verificamos novamente em destaque os pontos 28, 29 e 41, o gráfico renova a indicação de heterocedasticidade já percebida pelo diagrama de dispersão (Fig.1).

Os mesmos pontos já destacados anteriormente (28, 29 e 41) também aparecem em destaque no gráfico da Figura 6, das coordenadas do autovetor d_{max} em função das ordens das observações.

Dentre os pontos destacados pelos gráficos como candidatos a *outliers,* o que aparece sempre mais remoto é o correspondente à observação 29, daí a escolha deste ponto para sofrer urna pequena perturbação, espera-se com isso que o modelo seja sensível a esta alteração e o modo de medir esta sensibilidade será através da curvatura normal de Cook para modelo normal, dada na equação (3.12). Para introduzir a perturbação nas observações consideramos o vetor $\boldsymbol{\ell}$ da referida fórmula, com 1 em todas as posições com exceção da posição 29 que recebeu o valor 0,9. O valor obtido foi de 17,767, o que indica extrema sensibilidade local.

Usando o esquema de perturbação de observação descrito na subseção (4.3), aplicamos o resultado da inclinação máxima ($\boldsymbol{d}^*_{max}$) dada na equação (4.27), com isso obtemos o resultado 16,573, considerando como referência o valor sugerido pela fórmula em (4.32) que é de 7,381, temos aqui forte indicação de sensibilidade local.

Para o esquema de perturbação da matriz de variância abordado na subseção (4.3), o valor da inclinação máxima é de 17.766, obtido através da fórmula dada em (4.20). Usando o valor de referência para $\boldsymbol{d}^*_{max}$, sugerido na equação (4.32), que é de 13,798, também aqui detectamos forte sensibilidade local.

A abordagem de Billor e Loynes permite que se tenha os valores individuais das coordenadas $(\ell^*_{j's})$ do gradiente $\ell_{max} = \nabla LD^*(\omega_0)$, que o vetor que indica a direção onde se obtém a inclinação máxima. Assim podemos avaliar quais observações contribuem com a maior parcela para o valor de d^*_{max}. Os $\ell^*_{j's}$ individuais, obtidos pela perturbação da matriz de variância, são apresentados na Tabela 1, na qual nota-se que as observações que contribuem com os maiores valores são as de números 28,29 e 41.

Tabela 1 - Valores dos $\ell^*_{j's}$ para os dados de gansos

observação(i)	$\lvert\ell^*_j\rvert$	observação(i)	$\lvert\ell^*_j\rvert$	observação (i)	$\lvert\ell^*_j\rvert$
1	0,8892	16	0,6268	31	0,9975
2	0,9417	17	0,9881	32	0,5387
3	0,5988	18	0,9988	33	0,9586
4	0,9539	19	0,7177	34	0,9997
5	0,9418	20	0,9572	35	0,9452
6	0,7478	21	0,9372	36	0,8507
7	0,8822	22	0,8224	37	1,0793
8	0,8911	23	0,6010	38	0,9875
9	0,9461	24	0,6530	39	0,9834
10	0,7661	25	0,2726	40	0,8199
11	0,8224	26	0,9749	41	6,5283
12	0,6445	27	1,6615	42	0,3999
13	**0,7570**	28	**12,2618**	43	**0,7753**
14	**0,8111**	29	**7,0421**	44	**0,4322**
15	**0,9999**	30	**0,9949**	45	**0,9306**

5.2 Distúrbio coronário versus pressão sanguínea e nível de colesterol

Os dados utilizados nesta seção foram obtidos em Cordeiro (1986, pag.113) e apresentado primeiro por Ku e Kullback (1974), mostram a classificação de 1330 pacientes segundo a ocorrência ou não de distúrbio coronário pelos seguintes níveis de colesterol (NC_i):
$NC_1 =$ menos de 200
$NC_2 =$ de 200 a 219
$NC_3 =$ de 220 a 260
$NC_4 =$ mais de 260
e de pressão sanguínea (PS_i):

$PS_1 =$ menos de 127
$PS_2 =$ de 127 a 146
$PS_3 =$ de 147 a 166
$PS_4 =$ mais de 166.

Considerando como variável resposta $Y =$ proporção de indivíduos que apresentaram distúrbio coronário e como variáveis explicativas NC_i e PS_i , ambas definidas como fator com quatro níveis cada, como descritos acima. Adotamos para os dados o modelo logístico linear múltiplo, dado por

$$y_i = \beta_i + \sum_{j=1}^{4} \beta_{2j}\, PS_j + \sum_{j=1}^{4} \beta_{3j}\, NC_j + e_i \ , \quad i = 1, \cdots, n \tag{5.2}$$

com erro binomial; é assumido que $\beta_{21} = \beta_{31} = 0$. O desvio obtido para este modelo é $D(\boldsymbol{y}, p) = 4{,}775$ (9 graus de liberdade), indicando um ajuste adequado, pois ao nível de significância de 5% $(4{,}755 \ll 16{,}919)$. Na tabela 2, apresentada a seguir, temos as estimativas dos coeficientes do modelo adotado e seus respectivos desvios padrões, todos apresentaram-se significativos ao modelo.

Tabela 2 - Estimativas dos coeficientes referentes ao modelo em (5.2) e seus respectivos desvios padrões

Coeficiente	Estimativas	Desvio Padrão
Constante	-3,495	0,349
PS_2	-0,091	0,451
PS_3	0,562	0,351
PS_4	1, 342	0,343
NC_2	-0,038	0,303
NC_3	0,587	0,328
NC_4	1,204	0,327

Mostraremos agora os gráficos usados para detectar possíveis indicações de transgressão de alguma suposição adotada para o modelo, bem como da ocorrência de algum ponto influente. Podemos observar por meio do gráfico de probabilidade com envelopes, dado na Figura 7, que não há indicação de que a variável resposta não siga uma Binomial (m_i, p_i), pois todos os pontos encontram-se dentro das bandas de confiança do gráfico.

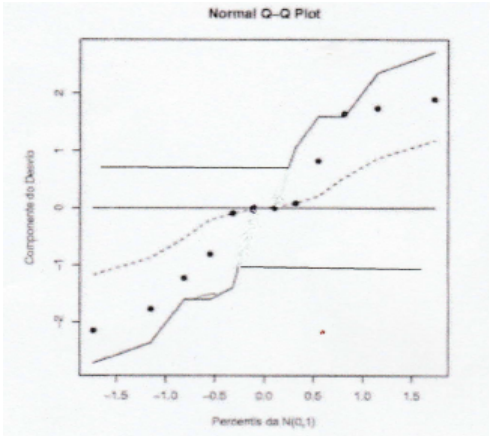

Figura 7 - Gráfico de probabilidade com envelopes para o modelo binomial.

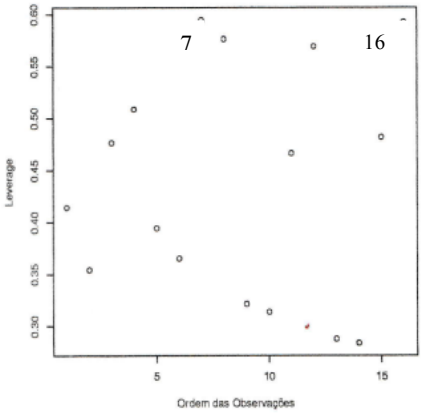

Figura 8 - Leverage $\hat{h}ii$ contra a ordem das observações.

No gráfico dos $\hat{h}_{ii}$ em função da ordem das observações, dado no Fig.8, nenhum dos pontos aparece acima da linha $\frac{2*p}{n}$, não havendo portanto pontos suspeitos de serem aberrantes, contudo os pontos mais afastados correspondem as observações 7 e 16. O gráfico de $\hat{h}_{ii}$ em função dos valores ajustados, mostrado na Figura 9, destaca os pontos correspondentes às observações 12, 15 e 16 como estando mais afastados.

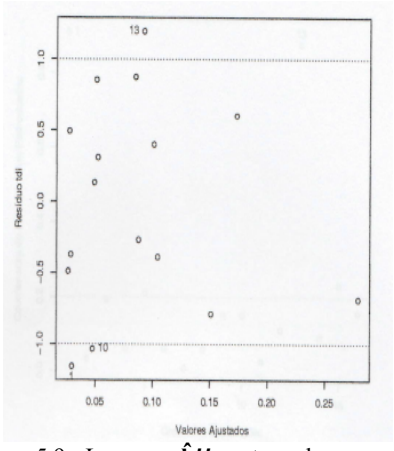

Figura 5.9 - Leverage $\hat{\boldsymbol{h}}\boldsymbol{ii}$ contra valores ajustados.

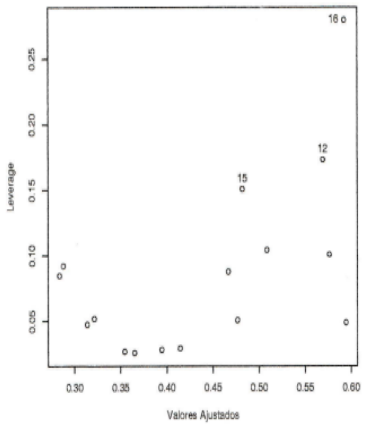

Figura 5.10 - Resíduo t_{Di} contra valores ajustados.

No gráfico do resíduo t_{Di} em função dos valores ajustados, Figura 10, os pontos que se encontram mais afastados correspondem às observações 10 e 13. Na Figura 11 temos o gráfico da medida afastamento pela verossimilhança, LD_i, em função da ordem

das observações, nele podemos perceber que o ponto correspondente á observação 1 parece mais afastado.

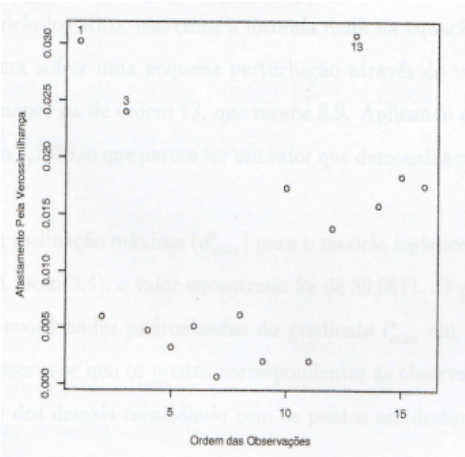

Figura 11 - Gráfico de LD_i contra a ℓ_{max} ordem das Observações .

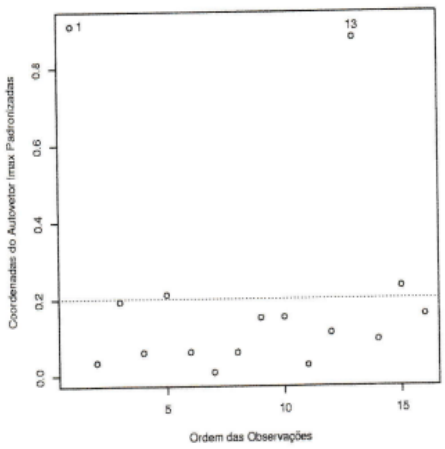

Figura 12 - Coordenadas do autovetor contra a ordem das Observações.

No gráfico das coordenadas do autovetor ℓ_{max} em função da ordem das observações, mostrado na Figura 12, podemos observar que os pontos correspondentes às observações 1 e 13, encontram-se mais afastado.

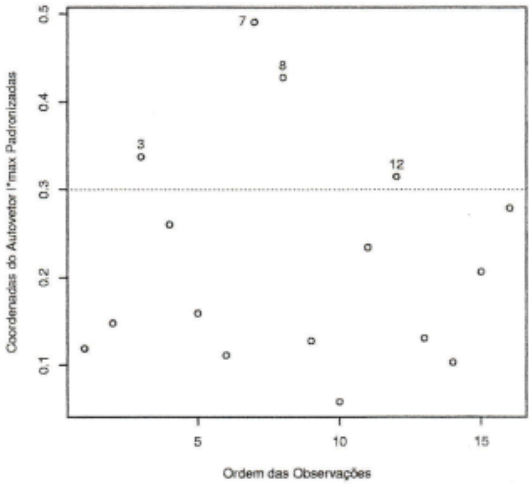

Figura.13 - Coordenadas do autovetor ℓ^*_{max} contra a ordem das observações

O gráfico mostrado na Figura 13, das coordenadas do vetor gradiente ℓ^*_{max} em função da ordem das observações, destaca os pontos correspondentes as observações 7 e 8.

Avaliando a situação dos pontos destacados pelos gráficos não percebemos uma forte evidência de ocorrência de *outliers*, porém os pontos que se apresentaram mais afastados foram aqueles correspondentes às observações 1, 7, 12, 16 e mais evidente a 13. Para calcular a curvatura normal para perturbação de observação em modelo logístico, usou-se a fórmula em (3.12), optamos por perturbar a observação 13, para isso, utilizamos o vetor de perturbação ℓ com 1 em todas as posições menos na observação 13, que recebe o valor 0,9. O resultado obtido para a curvatura normal foi de 9,132, o que parece ser um valor que demonstra certa sensibilidade do modelo..

No cálculo da inclinação máxima (d^*_{max}) no caso de perturbação de observação para o modelo logístico, abordado na subseção (4.4), usamos o resultado obtido na fórmula (4.34), o valor encontrado foi de 89,0611.

As coordenadas padronizadas em valor absoluto do autovetor l_{max} e do gradiente l^*_{max} são mostradas na tabela a seguir:

Tabela 3 - Coordenadas Padronizadas dos Vetores l_{max} e l^*_{max}

Observação (j)	$\lvert l^*_j \rvert$	$\lvert l_j \rvert$
1	0,1184	0,9094
2	0,1478	0,0355
3	0,3374	0,1429
4	0,2605	0,0612
5	0,1591	0,2111
6	0,1109	0,0622
7	0,4904	0,0089
8	0,4273	0,0607
9	0,1278	0,1494
10	0,0577	0,1514
11	0,2340	0,0274
12	0,3148	0,1112
13	0,1310	0.8786
14	0,1030	0,0932
15	0,2063	0,2314
16	0,2788	0,1577

Analisando a tabela (3) podemos destacar às observações 7 e 8 como as que apresentam as maiores coordenadas do gradiente ℓ_{max} enquanto que às observações 1 e 13 apresentam as maiores coordenadas do vetor ℓ^*_{max} . As técnicas utilizadas nesse trabalho, para detectar *outliers,* não mostraram claramente uma indicação consistente em relação a alguma observação, o que talvez indique que para essa aplicação não haja nenhuma observação que possa proporcionar modificações substanciais ao resultado do ajuste, porém alguns gráficos coincidiram na indicação da observação 13, como a que mais se destaque, estando afastada das demais.

6 Conclusões

Tanto a proposta de Cook (1986) como a de Billor e Loynes (1993) foram bastante inovadoras na análise de diagnóstico em modelos de regressão, e trouxeram contribuições significativas, dentre as quais destaca-se o fato de permitirem a análise de diagnósticos de outros modelos além do normal, como por exemplo o modelo logístico.

Um dos pontos fracos da proposta de Cook (1986), destacado no Capítulo 3, é justamente a falta de um valor de referência para comparar com o valor obtido para a curvatura normal. Em resposta a essa e outras dificuldade da versão de Cook (ver Seção 3) Billor e Loynes (1993) propuseram a inclinação máxima como instrumento da análise de diagnóstico em modelos de regressão, e apresentaram sugestões de valores de referência para comparar com o valor da inclinação máxima, sob alguns esquemas de perturbação, do modelo de regressão normal.

Neste trabalho apresentamos o cálculo da inclinação máxima no caso de esquema de perturbação de observação do modelo logístico, porém não sugerimos nenhum valor de referência para efeito de comparação que leve a indicação da sensibilidade do modelo. Destacamos que a natureza de cada problema e o auxílio de outras técnicas de diagnósticos podem ajudar na hora de decidir pela sensibilidade do modelo.

Na aplicação da subseção (5.1), onde usamos o modelo normal linear simples, a curvatura normal e a inclinação máxima apresentaram valores que coincidiram na indicação da sensibilidade do modelo adotado, vale ressaltar que os gráficos indicaram claramente as observações 28, 29 e 41 como mais afastadas.

Na aplicação correspondente a subseção (5.2), em que foi aplicado o modelo logístico múltiplo, não houve observações que evidenciassem possíveis *outliers*, pois houve certa variação dos gráficos utilizados na indicação desses pontos. Quanto ao valor da curvatura normal e da inclinação máxima, a falta de um valor de referência combinada à inexatidão dos gráficos utilizados na detecção de *outliers*, não nos permitiu decidir pela sensibilidade dos modelos, porém alguns gráficos coincidiram na indicação da observação 13, como a que mais se destaque, estando afastada das demais. Espera-se ter contribuído com este trabalho para o enriquecimento da análise de diagnóstico de modelos de regressão normal e logístico.

Agradecimentos

Aos revisores e editores pelas sugestões e comentários que possibilitaram um melhor entendimento do artigo.

LOBATO JUNIOR, D.; VEIGA, R. D. Diagnostic analysis in normal and logistic regression models. *Rev. Bras. Biom*. Lavras, v.38, n.4, p.449-482, 2020.

ABSTRACT: The diagnostic analysis of the quality of the fit is an important phase in regression models, especially with regard to the occurrence of outliers, which can cause substantial distortions in the estimated parameters of the model. In this work, two innovative methods were approached in this area: the method of local influence, which was initially proposed by Cook (1986) and the maximum inclination, proposed by Billor and Loynes (1993). Cook's proposal consists of assessing the normal curvature of a surface, based on the Likelihood Displacement measure (spacing by likelihood), under a small disturbance in the model. Then, Billor and Loynes's (1993) approach was presented, who applied a measure of distance due to the modified likelihood. Thus, the maximum slope is used as a measure of the influence of outliers on the model. Initially Billor and Loynes's (1993) proposal was applied to the normal model, this proposal was also developed for the logistic model. Two applications were presented, in which diagnostic analysis techniques of the quality of the adjustment were used, by means of graphs and the two proposals: local influence and maximum inclination. In the case of the simple linear normal model, the normal curvature and the maximum inclination showed the same sensitivity in the indication of influential observations. In the second application, the techniques used in this work, did not show a consistent indication in relation to any observation, however some graphics coincided with the indication of observation 13, as the one that stands out the most, being away from the others.

KEY-WORDS: Regression analysis; logistic regression; diagnostic techniques; local influence; maximum slope.

Referências

BILLOR, N.; LOYNES, R. M. Local Influence: a new approach. *Communications in Statistics - Theory and Methods*, v.22, p.1595-1661, 1993

COOK, R. D. Assessment of local influence (with discussion*)*. *Journal of the Royal Statistical Society B*, v.48, n.133-169, 1986*.*

COOK, R. D; WEISBERG, S. *Residuals and influence in regression.* New York: Chapman and Hall, 1982

COOK, R. D. Detection of influential observations in linear regressions*.* *Technometrics*, v.15, p.15-18, 1977.

CORDEIRO, G. M. Models of influential observations in linear regressions. *Technometrics*, v.15, p.15-18, 1986.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados em experimentação agronômica*. Piracicaba: ESALQ/USP, 2002.

DRAPER, N R.; SMITH, H. *Applied regression analysis*. 2.ed. New York: John Wiley and Sons, 1981.

GRAYBILL, F. A. *Matrices with applications in statistics.* 2.ed. Belmont: Wadsworth International Group, 1983.

HOFFMAN, K. E.; KUNZE, R. *Álgebra Linear*. 2.ed. São Paulo: LTC: Livros Técnicos e Científicos, 1979.

JAMES, B. R. *Probabilidade: Um Curso em nível intermediário.* Rio de janeiro: Livro Técnico, 1981.

MCCULLAGH, P.; NELDER, J. A. *Generalized Linear Models.* New York: John Wiley and Sons, 1987

NELDER, J. A.; WENDDERBURN, R, W. M. Generalized linear models. *Journal of the Royal Statistical Society* A, v.135, p.370-384, 1972.

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

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2020.

STOKER, J. J. *Differential geometry*. New York: John Wiley and Sons, 1969.

SCHALL, R.; DUNNE, T. T. A note on the relationship between parameter collinearity and local influence. *Biometrika*, v.79, p.399-404, 1992.

Recebido em 19.11.2019

Aprovado após revisão em 29.10.2020

COMANDOS NO PROGRAMA R

A seguir vamos descrever alguns comandos do aplicativo R, que foram utilizados no capítulo 4 deste artigo.

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

Os resíduos *studentizados* t_i e t_i^* podem ser obtidos no R, pela sequência de comandos

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

s← fit.model$sigma

r← resid(fit.model)

t_i←-r / $(s \cdot (1 - h_{ii})$^0,5)

$t_i^* \leftarrow t_i \cdot (\frac{n-p-1}{n-p-t_i^2})$^0,5

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

Comandos para obter os valores h_{ii}, considerar os resultados do ajuste no objeto fit.model

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

X ← model.matrix(fit.model)

H ← solve(t(X)% · %X)

H ← $X\%$ · %$H\%$ · %$t(X)$

$h_{ii} \leftarrow diag(H)$

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

Comandos para obter os resíduos h_{ii} , t_{S_i}, t_{D_i} e LD_i, para os MLGs, aplicativo R

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

X ← model.matrix(fit.model)

n ← nrow(X)

p ← ncol(X)

W ← fit.model$weights

W ← diag(W)

H ← solve(t(X)% · %W%· %X)

H ← sqrt(W)%· %X%· %H%· %t(X)%· %sqrt(W)

h_{ii}← diag(H)

t_{D_i}← resid(fit.model, type=”deviance”) / $sqrt(1 - h_{ii})$

rp ← resid(fit.model, type”Pearson”)

t_{S_i} ← rp / sqrt $(1 - h_{ii})$

$LD_i \leftarrow h_{ii} \cdot (t_{S_i}$^ 2$)/ \ (1 - h_{ii})$

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

Comandos para calcular a curvatura normal no R, considere $\boldsymbol{\ell}$ como o vetor definido na subseção (3.2)

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

l_{ms} ← summary(fit.model)

s ← l_{ms} $sigma

e ← resid(l_{ms})

D_e ← diar(e)

X ← model.matrix(fit.model)
$P_X \leftarrow X\% \cdot \% solve(t(X)\% \cdot \% X)\% \cdot \% t(X)$
$pl \leftarrow D_e\% \cdot \% P_X\% \cdot \% D_e$
$esq \leftarrow e$ ^ 2
$p_2 \leftarrow (esq\% \cdot \% t(esq))$ / $2 \cdot 45 \cdot s$ ^ 2
$p_3 \leftarrow pl + p_2$
$C_{\ell} \leftarrow t(\boldsymbol{\ell})\% \cdot \% p_3\% \cdot \% \boldsymbol{\ell}$
$C_{\ell} \leftarrow 2 \cdot C_{\ell}$ / s ^ 2

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
Comandos para obter d^*_{max} no R
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
l_{max} ← esq / s ^ 2
l_{max} ← l_{max} ^ 2
$d_{max} \leftarrow sum(l_{max})$
d_{max} ← sqtr(d_{max})