APLICAÇÃO DA ANÁLISE DE INFLUÊNCIA LOCAL EM MODELOS DE REGRESSÃO LOGÍSTICA

Édila Cristina de SOUZA[1]
Edwin Moises Marcos ORTEGA[1]
Vicente Garibay CANCHO[2]

▪ RESUMO: Uma etapa importante após a formulação e ajuste de um modelo de regressão é a análise de diagnóstico. A regressão logística tem se constituído num dos principais métodos de modelagem estatística de dados; mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso pode ser modelada pela regressão logística. Os estimadores obtidos pelo método de máxima verossimilhança e Bayesiano foram utilizados. Neste trabalho, consideramos um estudo de diagnóstico no modelo da regressão logística, utilizando a teoria de influência local de Cook (1986). Investigamos a aplicação da técnica de influência local sob diferentes esquemas de perturbação. Como ilustração, apresentamos a aplicação dos resultados desenvolvidos em um conjunto de dados reais.

▪ PALAVRAS-CHAVE: Diagnóstico; influência local; modelo de regressão logística; curvatura; método Bayesiano; Metropolis-Hasting.

1 Introdução

A análise de regressão é uma técnica estatística que tem como objetivo descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas por meio de um modelo que tenha bom ajuste.

Na regressão logística, a variável resposta, normalmente, pode ser dicotômica ou binária, isto é, aquela que apresenta duas possibilidades de resposta (sucesso

[1]Departamento de Ciências Exatas, Universidade de São Paulo - ESALQ/USP, CEP 13418-900 Piracicaba, SP, Brasil. E-mails: *edilacr@yahoo.com.br / edwin@esalq.usp.br*
[2]Departamento de Matemática Aplica e Estatística, Instituo de Ciências Matemáticas e de Computação, Universidade de São Paulo - USP, Caixa Postal 668, CEP: 13560-970, São Carlos, SP, Brasil. E-mail: *gariby@icmc.usp.br*

ou fracasso), como, por exemplo, o objetivo de um ensaio experimental realizado para testar a sobrevivência, ou não, de enxertos de um determinado cultivar, ou então, o efeito (sucesso ou fracasso) de um inseticida quando este é aplicado a um determinado número de insetos.

A modelagem dos dados pode ser feita com base em modelos estatísticos paramétricos supostamente apropriados. A escolha correta de um modelo que se ajuste de forma adequada a um conjunto específico de dados é de grande importância, uma vez que a não tendenciosidade dos resultados da análise depende dessa escolha. Assim, uma etapa importante na análise de um ajuste de regressão logística é o estudo da robustez dos resultados obtidos com relação à presença de pontos extremos. Detectar observações aberrantes e/ou influentes constitui um passo importante na análise do conjunto de dados. Pregibon (1981) aprimorou os métodos de diagnóstico de regressão linear para a regressão logística; desenvolvendo várias medidas para análise de resíduos e diagnóstico, como o resíduo de "*Pearson*" e da "*Deviance*".

Neste trabalho são discutidos alguns procedimentos de diagnóstico aplicados ao modelo de regressão logística; tendo sido utilizadas técnicas que possibilitam medir o quanto pequenas alterações nos dados ou no modelo podem influenciar nos resultados inferencias do problema em estudo.

Técnicas simples são bastante utilizadas para tal propósito e se baseiam na retirada individual de casos. Medidas de influência para cada observação da amostra são construídas pela comparação de estimativas calculadas para o conjunto completo de dados e para o conjunto de dados obtidos, eliminando-se a observação correspondente.

Nesse contexto, Cook (1977) sugere uma medida de influência desenvolvida inicialmente para modelos de regressão linear com erros normais. Também Cook (1986) apresenta a técnica denominada influência local, na qual em vez de retirar uma observação, atribui-se um peso a ela. Nessa última técnica, são introduzidas simultaneamente perturbações em cada um dos casos, sendo a medida de influência construída a partir da função do logaritmo da verossimilhança. Diferentes esquemas de perturbação podem ser aplicados, de acordo com o elemento da análise que o pesquisador deseja monitorar. Essa técnica permite detectar observações conjuntamente influentes, o que constitui uma vantagem em relação ao esquema de retirada de casos, visto que, nesse último, possíveis observações influentes podem não ser detectadas devido a presença de outras observações.

A presença de observações influentes na amostra pode levar a resultados inferenciais completamente diferentes, sendo importante ao pesquisador conhecer e analisar esses casos para decidir pela retirada, ou não, deles do estudo.

Essa metodologia teve uma grande receptividade entre os pesquisadores de regressão, havendo inúmeras publicações sobre o assunto, como, por exemplo, Ortega; Bolfarine e Paula (2003) que aplicam influência local em modelos log-gama generalizados com dados censurados e Hossain e Slam (2003) que aplicam a metodologia em modelos de regressão logística.

Mediante o exposto, o objetivo do presente trabalho é pesquisar e analisar

a aplicação da metodologia de influência local nos modelos de regressão logística onde os parâmetros são estimados utilizando o método de máxima verossimilhança e Bayesiano.

2 Regressão logística

No modelo de regressão logística, a variável resposta (Y) é dicotômica, possui valores 1 e 0, com probabilidade π_i para $Y_i = 1$ e probabilidade $1 - \pi_i$ para $Y_i = 0$, conforme Hosmer e Lemeshow(1989).

Seja um conjunto com p variáveis independentes, denotadas por $\boldsymbol{x}_i^T = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{ip})$, o vetor da i-ésima linha da matriz $(\boldsymbol{X})$ das variáveis explicativas, em que cada elemento da matriz corresponde ao ij-ésimo componente (x_{ij}), em que $i = 1, 2, \ldots, n$ e $j = 0, 1, \ldots, p$, com $x_{i0} = 1$. Denota-se por $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, o vetor de parâmetros desconhecidos e β_j é o j-ésimo parâmetro associado a variável explicativa x_j. No modelo de regressão logística múltipla a probabilidade de sucesso é dada por:

$$\pi_i = \pi(\boldsymbol{x}_i) = P(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

e a probabilidade de fracasso por:

$$1 - \pi_i = 1 - \pi(\boldsymbol{x}_i) = P(Y_i = 0 | \boldsymbol{X} = \boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

No modelo de regressão múltipla assume-se que Y_i tem uma distribuição de Bernoulli com parâmetro de sucesso π_i.

O *"logit"* para o modelo de regressão múltipla é dado pela equação:

$$g(\boldsymbol{x}_i) = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Assim, o logaritmo da função de verossimilhança pode ser escrito como:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln(1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}) \right]. \quad (1)$$

2.1 Estimação dos parâmetros pelo método de máxima verossimilhança

Para poder estimar os parâmetros foi utilizado o método de máxima verossimilhança, no qual encontramos o valor de $\boldsymbol{\beta}$ que maximiza $l(\boldsymbol{\beta})$, foi utilizado o processo iterativo de Newton-Raphson, e para isso fez-se necessário derivar $l(\boldsymbol{\beta})$ em relação a cada parâmetro,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{\exp(x_i^T \boldsymbol{\beta})}{1 + \exp(x_i^T \boldsymbol{\beta})} x_{ij} \right] = \sum_{i=1}^{n} \left[y_i - \pi_i \right] x_{ij}.$$

Dessa maneira, o vetor escore $\boldsymbol{U}(\boldsymbol{\beta})$ pode ser escrito como

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\pi} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\pi}).$$

A matriz de informação de Fischer é dada por:

$$\boldsymbol{I}(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = \boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X},$$

sendo $\boldsymbol{Q} = diag[\boldsymbol{\pi}_i(1 - \boldsymbol{\pi}_i)]$ e $\boldsymbol{X}$ a matriz de dados, e sua inversa $[\boldsymbol{I}(\boldsymbol{\beta})]^{-1}$, a matriz de variâncias e covariâncias das estimativas de máxima verossimilhança dos parâmetros.

A solução para as equações de verossimilhança é obtida usando o método iterativo de Newton Raphson. O conjunto de equações iterativas é dado por:

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + [\boldsymbol{I}(\boldsymbol{\beta}^{(t)})]^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(t)}); t = 0, 1, 2, \ldots \\ &= \boldsymbol{\beta}^{(t)} + [\boldsymbol{X}^T\boldsymbol{Q}^{(t)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\pi}^{(t)}). \end{aligned} \quad (2)$$

sendo que $\boldsymbol{\beta}^{(t)}$ e $\boldsymbol{\beta}^{(t+1)}$ são vetores de parâmetros estimados nos passos t e $t+1$, respectivamente.

Geralmente, não é possível encontrar distribuições exatas para os estimadores, assim sendo trabalha-se com resultados assintóticos considerando-se que o modelo escolhido irá satisfazer as condições de regularidade.

Cox e Hinkley (1986) demostram que, em problemas regulares, a função *Escore* $\boldsymbol{U}(\boldsymbol{\beta}) = \dfrac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$ tem valor esperado igual a zero e a estrutura de covariância é igual à matriz de informação de Fischer $\boldsymbol{I}(\boldsymbol{\beta})$.

Assim, a distribuição assintótica dos $\boldsymbol{\beta}$ é dada por:

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, [\boldsymbol{I}(\boldsymbol{\beta})]^{-1}).$$

Os métodos de inferência são baseados na teoria de máxima verossimilhança. Conforme esta teoria, existem três estatísticas para testar hipóteses relativas aos $\boldsymbol{\beta}$´s, que são deduzidas de distribuições assintóticas de funções adequadas dos $\boldsymbol{\beta}$´s (Demétrio, 2002).

Supondo-se interesse em testar as hipóteses:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$
$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

As três estatísticas são:

i) A estatística da razão da verossimilhança que é dada por:

$$\Lambda = -2\ln\left[\frac{L(\boldsymbol{\beta}_0)}{L(\widehat{\boldsymbol{\beta}})}\right] = 2[l(\widehat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_0)]$$

em que $\widehat{\boldsymbol{\beta}}$ é o estimador da máxima verossimilhança sob todo espaço paramétrico.

ii) A estatística Wald que é dada por:

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

em que $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ é a matriz de informação de Fischer avaliada em $\widehat{\boldsymbol{\beta}}$.

iii) A estatística Escore que é dada por:

$$Es = \boldsymbol{U}^T(\boldsymbol{\beta}_0)[\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}(\boldsymbol{U}(\boldsymbol{\beta}_0))$$

em que $[\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}$ é a inversa da matriz de informação avaliada em $\boldsymbol{\beta}_0$.

Essas três estatísticas são assintoticamente equivalentes e sob H_0 tem-se que:

$$\Lambda, W \text{ e } Es \sim \chi^2_p.$$

Para maiores detalhes ver Allison (1999), Collet (1991), Hinde e Demétrio (1998), Hosmer e Lemeshow (1989), Paula (2004) entre outros.

2.2 Análise Bayesiana para o modelo de regressão logística

O uso do método Bayesiano além de ser uma alternativa de análise, permite a incorporação de conhecimento prévio dos parâmetros por meio de densidades *a priori* informativas. Quando não existe essa informação considera-se prioris não informativas. Na abordagem Bayesiana, a informação referente aos parâmetros do modelo são obtidas pela distribuição marginal *a posteriori*. Neste sentido surgem duas dificuldades: a primeira refere-se à obtenção da distribuição *a posteriori* marginal e a segunda ao cálculo de momentos de interesse. Em ambos os casos são necessárias resoluções de integrais que muitas vezes não apresenta solução analítica. Neste trabalho utilizamos o método de simulação de Monte Carlo via Cadeias de Markov, tais como o amostrador de Gibbs e Metropolis-Hasting. Para representar o grau de conhecimento sobre os parâmetros consideramos as seguintes densidades *a priori* $\beta_j \sim N(\mu_j, \sigma_j^2)$, $j = 0, 1, \ldots, p$, onde μ_j e σ_j^2 são constantes conhecidas baseadas na informação *a priori* do especialista e da análise preliminar dos dados. Assumindo independência entre os parâmetros, a densidade *a priori* conjunta é dada por

$$\pi(\boldsymbol{\beta}) = \Pi_{j=0}^{p} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{ -\frac{1}{2\sigma_{\mathrm{j}}^2}(\beta_{\mathrm{j}} - \mu_{\mathrm{j}})^2 \right\}. \tag{3}$$

Considerando a densidade *a priori* (3) e o logaritmo da função de verossimilhança (1) a densidade *a posteriori* conjunta para os parâmetros é dada por

$$\pi(\boldsymbol{\beta}|D) \propto \exp\left\{ \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{y_i} \boldsymbol{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta} - \ln(1 + \exp^{\boldsymbol{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}})] \right\} \Pi_{\mathrm{j}=0}^{\mathrm{p}} \frac{1}{\sqrt{2\pi\sigma_{\mathrm{j}}}} \exp\left\{ -\frac{1}{2\sigma_{\mathrm{j}}^2}(\beta_{\mathrm{j}} - \mu_{\mathrm{j}})^2 \right\} \tag{4}$$

em que denota D o conjunto de dados observados. No enfoque Bayesiano, inferências são baseadas nas densidades *a posteriori* marginais dos parâmetros envolvidos

no modelo. Pode-se observar que, a densidade *a posteriori* conjunta não é uma densidade padrão, portanto podemos avaliar as densidades marginais por meio de métodos de aproximação, tais como, o método de Laplace ou por métodos de simulação de Monte Carlo via cadeias de Markov (MCMC). Neste trabalho optamos pelo método de simulação por serem de fácil implementação computacional. Para obtermos uma amostra das densidades *a posteriori* marginais de $\boldsymbol{\beta}$ utilizaremos o amostrador de Gibbs que é uma técnica para gerar variáveis aleatórias de uma distribuição marginal sem ter que calcular a forma de densidade. Assim, podemos mostrar que as densidades condicionais marginais são dadas por:

$$\begin{aligned}\pi(\beta_k|\beta_0,\beta_1, \quad \ldots, \quad \beta_{k-1},\beta_{k+1},\ldots,\beta p, D) \propto & \exp\Big\{\beta_{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i x_{ik}}\Big\} \\ & \exp\Big\{-\sum_{\mathrm{i}=1}^{\mathrm{n}}\ln\Big[1+\Big(\prod_{\mathrm{j}=1,\mathrm{j}\neq\mathrm{k}}^{\mathrm{p}}\exp\{\mathrm{x_{ij}}\beta_{\mathrm{j}}\}\Big)\exp\{\mathrm{x_{ik}}\beta_{\mathrm{k}}\}\Big]\Big\} \\ & \exp\Big\{-\frac{1}{2\sigma_{\mathrm{k}}^2}(\beta_{\mathrm{k}}-\mu_{\mathrm{k}})^2\Big\} \qquad (5)\end{aligned}$$

fazendo $\mathrm{c_i} = \prod_{\mathrm{j}=1,\mathrm{j}\neq\mathrm{k}}^{\mathrm{p}}\exp\{\mathrm{x_{ij}}\beta_{\mathrm{j}}\}$ e $\mathrm{w} = \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i x_{ik}}$ as densidades condicionais podem ser escritas como

$$\begin{aligned}\pi(\beta_k|\beta_0,\beta_1, \ \ldots, \ \beta_{k-1},\beta_{k+1},\ldots,\beta p, D) \propto & \exp\Big\{\mathrm{w}\beta_{\mathrm{k}}-\ln\Big[\prod_{\mathrm{i}=1}^{\mathrm{n}}(1+\mathrm{c_i}\exp\{\mathrm{x_{ik}}\beta_{\mathrm{k}}\})\Big]\Big\} \\ & \exp\Big\{-\frac{1}{2\sigma_{\mathrm{k}}^2}(\beta_{\mathrm{k}}-\mu_{\mathrm{k}})^2\Big\} \\ \propto \quad & \exp\Big\{\mathrm{w}\beta_{\mathrm{k}}-\frac{1}{2\sigma_{\mathrm{k}}^2}(\beta_{\mathrm{k}}-\mu_{\mathrm{k}})^2\Big\}\Big[\prod_{\mathrm{i}=1}^{\mathrm{n}}(1+\mathrm{c_i}\exp\{\mathrm{x_{ik}}\beta_{\mathrm{k}}\})\Big]^{-1} \\ \propto \quad & \exp\Big\{-\frac{1}{2\sigma_{\mathrm{k}}^2}\Big[\beta_{\mathrm{k}}-(\sigma_{\mathrm{k}}\mathrm{w}+\mu_{\mathrm{k}})\Big]^2\Big\}\Big[\prod_{\mathrm{i}=1}^{\mathrm{n}}(1+\mathrm{c_i}\exp\{\mathrm{x_{ik}}\beta_{\mathrm{k}}\})\Big]^{-1} \qquad (6)\end{aligned}$$

Observamos claramente que as densidades condicionais marginais não são densidades padrões, portanto, utilizaremos o amostrador de Gibbs (Gelfand e Smith, 1990) com algoritmo de Metropolis-Hasting (Chib e Greenberg, 1995) para gerar as amostras *a posteriori* de $\boldsymbol{\beta}$.

A convergência foi monitorada uitlizando a estatística de Gelman e Rubin apresentada no WinBUGS.

3 Influência local

Ajustando um modelo a um conjunto de dados, deseja-se que as estimativas obtidas a partir do modelo proposto sejam resistentes a pequenas perturbações no modelo ou nas observações.

Enquanto a análise de resíduos estuda os problemas com o modelo ajustado, como presença de observações aberrantes e afastamentos sérios das suposições para a distribuição do erro, uma análise de influência é feita assumindo o modelo como correto, e estuda-se a robustez das conclusões a perturbações nos dados ou no modelo. Uma observação se diz influente quando produz alterações desproporcionais nos resultados da análise ao ser omitida no ajuste do modelo ou submetida a uma pequena perturbação.

Na análise de diagnóstico, considera-se que o modelo postulado é o modelo correto, e comparam-se as estimativas obtidas por meio desse modelo com as estimativas decorrentes de uma pequena perturbação.

Cook (1986) desenvolveu alguns procedimentos de Diagnóstico de Influência Local. Essa metodologia é extensamente discutida por vários pesquisadores para a Regressão Linear, Regressão Não-Linear, Modelos Lineares Generalizados e modelos de Análise de Sobrevivência. Hossain e Islam (2003) analisa os procedimentos de Diagnóstico para os modelos de regressão logística.

Existem na literatura numerosos trabalhos de aplicação da metodologia de Cook (1986), por exemplo, Galea; Bolfarine e Vilca-Labra (2002), Ortega; Bolfarine e Paula (2003) e Hossain e Islam (2003).

3.1 Métodologia de influência local

Dado um conjunto de observações, seja $l(\boldsymbol{\beta})$ o logaritmo da função de verossimilhança correspondente ao modelo postulado, sendo que $\boldsymbol{\beta}$ é um vetor $(p+1) \times 1$ de parâmetros desconhecidos. Perturbações podem ser introduzidas no modelo por de um vetor $\boldsymbol{w}^T = (w_0, w_1, ..., w_n)$ pertencente a um subconjunto aberto Ω de $\mathbb{R}^n$. Geralmente, $\boldsymbol{w}$ pode refletir qualquer esquema de perturbação bem definida, por exemplo, $\boldsymbol{w}$ pode ser usado para introduzir uma menor modificação nas variáveis explicativas ou para perturbar a matriz de covariância nos erros, no modelo de regressão linear. (Galea; Paula e Bolfarine, 1997).

Supondo que o esquema de perturbação esteja definido, denotado por $l(\boldsymbol{\beta}|\boldsymbol{w})$ como logaritmo da função de verossimilhança perturbada, o vetor $\boldsymbol{w}$ expressa um esquema de pesos, existindo um ponto w_0, em que $l(\boldsymbol{\beta}|\boldsymbol{w}_0) = l(\boldsymbol{\beta})$. Dado que $\widehat{\boldsymbol{\beta}}$ é o estimador de máxima verossimilhança obtido por meio de $l(\boldsymbol{\beta})$ e $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ é o estimador de máxima verossimilhança obtido por meio de $l(\boldsymbol{\beta}|\boldsymbol{w})$, o objetivo é comparar $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}$, quando $\boldsymbol{w}$ varia em Ω. Cook (1986) sugere que a comparação entre $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ seja feita por afastamento pelo logaritmo da função de verossimilhança $LD(\boldsymbol{w})$, expresso da seguinte maneira:

$$LD(\boldsymbol{w}) = 2[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}})]. \tag{7}$$

Dessa forma, $LD(\boldsymbol{w})$ contém informação essencial sobre a influência do esquema de perturbação.

A idéia de Cook (1986) é estudar o comportamento da função $LD(\boldsymbol{w})$ numa vizinhança $\boldsymbol{w}_0$, que é o ponto em que as duas verossimilhanças são iguais. Para

isso, o autor considerou a seguinte superfície geométrica:

$$\alpha(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{w} \\ LD(\boldsymbol{w}) \end{pmatrix},$$

que é denominada de gráfico de influência. A idéia principal do autor, foi analisar como $\alpha(\boldsymbol{w})$ desvia-se de seu plano tangente em $\boldsymbol{w}_0$, preocupando-se com o comportamento da função $LD(\boldsymbol{w})$ em torno de $\boldsymbol{w}_0$. O procedimento consiste em selecionar uma direção unitária $\boldsymbol{d}$, e, então, considerar o gráfico de $LD(w_0 + a\boldsymbol{d})$ em função de a, em que $a \in \mathbb{R}$. Esse gráfico é chamado de linha projetada. Desde que $LD(w_0) = 0$, $LD(w_0 + a\boldsymbol{d})$ tem um mínimo local em $a = 0$. Cada linha projetada pode ser caracterizada pela curvatura normal C_d em torno de $a = 0$. Sugere-se considerar a direção $\boldsymbol{d}_{max}$ correspondente à maior curvatura $C_{\boldsymbol{d}_{max}}$. O gráfico das componentes de $\boldsymbol{d}_{max}$ revela os elementos que sob pequenas perturbações, exercem notável influência sobre $LD(\boldsymbol{w})$.

Cook (1986) mostra que a curvatura normal na direção $\boldsymbol{d}$ pode ser expressa da seguinte forma:

$$C_d = 2|\boldsymbol{d}^T \boldsymbol{F} \boldsymbol{d}|, \tag{8}$$

sendo que $\boldsymbol{F} = \boldsymbol{\Delta}^T [\boldsymbol{I}(\widehat{\boldsymbol{\beta}})]^{-1} \boldsymbol{\Delta}$, $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ é a matriz de informação observada sob o modelo postulado e $\boldsymbol{\Delta}$ é a matriz $(p+1) \times n$ definida por:

$$\boldsymbol{\Delta} = \frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \boldsymbol{\beta} \partial \boldsymbol{w}^T} \tag{9}$$

e avaliados em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$.

O resultado na equação (9) pode ser utilizado para avaliar a influência que o esquema de perturbações considerado exerce sobre os componentes do modelo, tais como estimativas dos parâmetros e outros resultados da análise estatística. Segundo Cook (1986), a direção que produz a maior mudança local na estimativa dos parâmetros é dada por $\boldsymbol{d}_{max}$, que corresponde ao autovetor associado ao maior autovalor de $\boldsymbol{\Delta}^T \boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1} \boldsymbol{\Delta}$. O vetor $\boldsymbol{d}_{max}$ é utilizado para identificar as observações que podem estar controlando propriedades importantes na análise dos dados.

3.2 Esquemas de perturbação

Os métodos de diagnóstico para dados perturbados utilizados são: casos ponderados, perturbação, perturbação de uma covariável e perturbação de um subconjunto de covariáveis.

3.2.1 Caso ponderado

Para avaliar a influência das perturbações de casos, o logaritmo da função de verossimilhança perturbada é definida por:

$$l(\boldsymbol{\beta}|\boldsymbol{w}) = \sum_{i=1}^{n} w_i \left[y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln(1 + \exp(x_i^T \boldsymbol{\beta})) \right] \tag{10}$$

Para esse esquema de perturbação, o vetor correspondente à não perturbação é o vetor n-dimensional $\boldsymbol{w}_0 = (1, 1, \ldots, 1)^T$. Nesse caso a i-ésima linha da matriz $\boldsymbol{\Delta}$, é dada por

$$\boldsymbol{\Delta}_i^T = \left[\frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \beta_0 \partial w_i}, \frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \beta_1 \partial w_i}, \ldots, \frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \beta_p \partial w_i} \right]$$

Assim, os elementos da i-ésima linha da matriz $\boldsymbol{\Delta}$, avaliados em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$, para todo $j = 1, 2, \ldots, p$ podem ser expressos da seguinte maneira:

$$\frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \beta_j \partial w_i} = \left[y_i x_{ij} - \frac{\exp\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\}} x_{ij} \right] = (y_i - \pi_i) x_{ij}$$

3.2.2 Variáveis explanatórias

Assim como realizado por Thomas e Cook (1990) e Hossain e Islam (2003), modificou-se a t-ésima coluna da matriz de dados $\boldsymbol{X}$, adicionando um vetor $\boldsymbol{w}$ de pequenas perturbações multiplicadas por um fator de escala v. Neste caso, a perturbação é da forma:

$$x_{it} \longrightarrow x_{it} + v w_i, \;\; i = 1, \ldots, n$$

sendo que, v está atribuindo um peso para cada elemento da perturbação w_i. Como peso utilizou-se a estimativa do desvio padrão da variável X_t. Nesse caso, o logaritmo da função de verossimilhança perturbada é dado por:

$$l(\boldsymbol{\beta}|\boldsymbol{w}) = \sum_{i=1}^{n} \left[\boldsymbol{y}_i \boldsymbol{x}_i^{T*} \boldsymbol{\beta} - \ln(1 + \exp(\boldsymbol{x}_i^{T*} \boldsymbol{\beta})) \right] \tag{11}$$

sendo que,

$$\boldsymbol{x}_i^{T*} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_t (x_{it} + v w_i) + \ldots + \beta_p x_{ip}$$

Assim, os elementos da i-ésima linha da matriz $\boldsymbol{\Delta}$, avaliados em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$, para todo $j = 0, 1, 2, \ldots, p$ podem ser expressos da seguinte maneira:

$$\frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{w})}{\partial \beta_j \partial w_i} = \begin{cases} \left[(y_i - \widehat{\pi}_i) - \widehat{\pi}_i (1 - \widehat{\pi}_i) \widehat{\beta}_t x_{ij} \right] v & \text{para } j = t \\ -\widehat{\pi}_i (1 - \widehat{\pi}_i) x_{ij} \widehat{\beta}_t v & \text{para } j \neq t \end{cases}$$

3.3 Influência local total

Lesaffre e Verbeke (1998) sugeriram avaliar a direção do i-ésimo indivíduo, que é dada pelo vetor $\boldsymbol{d}_i = (0, \ldots, 1, \ldots, 0)$, sendo o i-ésimo elemento igual a um. Nesse caso, a curvatura normal chamada de influência local total do i-ésimo indivíduo, é dada por

$$C_i = 2 |\boldsymbol{\Delta}_i^T [\boldsymbol{I}(\boldsymbol{\beta})]^{-1} \boldsymbol{\Delta}_i| \tag{12}$$

e sugere-se estudar o gráfico de C_i contra a ordem das observações.

4 Aplicação

Os dados utilizados foram cedidos por Paula Roberta Mendes e coletados em clínicas veterinárias da cidade de Lavras, MG. Segundo Mendes, as fichas de atendimento foram previamente avaliadas, registrando 176 animais, porém devido a observações incompletas, foram consideradas neste trabalho, 151 observações.

Nesta aplicação, vamos ajustar um modelo de regressão logística para prever a probabilidade de óbito de cães acometidos por gastroenterite hemorrágica. Para estimar os parâmetros do modelo assim como a parte de influência local foi usado o software R (versão 2.01). Na parte Bayesiana foi utilizado o amostrador de Gibbs do aplicativo Winbugs.

A gastroenterite hemorrágica é uma patologia canina de aparecimento súbito. Os sintomas clínicos mais significantes deste tipo de gastroenterite são vômitos e/ou diarréia podendo conter sangue. O sangue pode apresentar-se de duas formas, em natureza (vermelho vivo) ou digerido (vermelho escuro a acastanhado). Pode ter etiologia viral, bacteriana ou parasitária. Além disso, sabe-se que fatores importantes associados devem ser considerados, como idade, raça, porte (peso), estresse ambiental e condições climáticas (Costa, 1997). O diagnóstico é feito por exclusão de partes, tendo primeiramente que ser consideradas outras causas e patologias de diarréia com sangue, ou seja, úlceras, trauma, tumores ou obstruções gastrointestinais, corpos estranhos, doenças infecciosas e desordens de coagulação. Para avaliação destas outras causas podem ser necessários testes laboratoriais, como hemograma completo, urianálise, radiografias, provas de coagulação e endoscopia ao aparelho gastrointestinal.

As variáveis utilizadas nesta aplicação foram:

y_i: Condição final do animal após o tratamento. (0 = não morreu, 1 = morreu) (**óbito**);

x_{i1}: Sexo do animal (0 = fêmea, 1 = macho) (**sexo**);

x_{i2}: Idade do animal contabilizada a cada seis meses, (1 = cães com menos de seis meses, 2 = cães com sete a doze meses, e assim sucessivamente) (**idade**);

x_{i3}: Quantidade de dias que o animal ficou internado (**diária**);

x_{i4}: Número de vezes que o animal foi consultado na clínica (**atendimento**).

Tabela 1 – Distribuição dos animais após o tratamento conforme o desfecho deste estudo

Óbito	Total de freqüência	Porcentagem(%)
0 (não)	108	71,52
1 (sim)	43	28,48
Total	151	100

Na análise exploratória dos dados pode-se perceber, segundo a Tabela 1, que a variável resposta **óbito**, é a condição final do animal após o tratamento, sendo codificada como: 1 = sim e 0 = não. Dos resultados obtidos, tem-se que dos 151 animais, 43 foram ao óbito, ou seja, 28,48%.

4.1 Utilizando o método de máxima verossimilhança

Ajustando um modelo de regressão logística e testando as hipóteses

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$
$$H_1 : \boldsymbol{\beta} \neq \mathbf{0}$$

tem-se que as estatísticas são dadas pelos resultados apresentados na Tabela 2.

Tabela 2 – Estatísticas da razão da verossimilhança, Escore e Wald

Estatísticas	p-Valor
Razão da verossimilhança(Λ)= 8,2790	0,1025
Escore (Es)= 8,0931	0,0882
Wald (W)= 7,2565	0,1229

Da Tabela 2 pode-se inferir que não foi rejeitada a hipótese nula, considerando um nível de 5% de significância, isto é, nenhuma variável é significativa para o modelo proposto. Entretanto, decidiu-se pela continuidade das análises.

As estimativas de máxima verossimilhança são observadas na Tabela 3, na qual pode ser verificado que, considerando um nível de 5%, nenhuma variável é significativa, e mediante a *deviance* observada conclui-se que o modelo não está bem ajustado.

Tabela 3 – Estimativas dos parâmetros, erro padrão, estatística Wald e p-valor

Efeito	Parâmetro	Estimativa	Erro padrão	Estat. Wald	p-Valor
Intercepto	β_0	-1,5284	0,4342	12,3874	0,0004
sexo	β_1	0,5683	0,3742	2,3063	0,1289
idade	β_2	-0,0143	0,0150	0,9193	0,3377
diaria	β_3	-0,0904	0,1186	0,58070	0,4461
atendime	β_4	0,2866	0,1563	3,3617	0,0667

$Deviance = 172,136$ com 146 g.l.

Na Tabela 4, pelas estimativas das razões de chances, percebe-se que a variável sexo é um fator de risco e a variável diária um fator de proteção em relação a variável óbito dos animais, sendo a chance de um animal ser macho e vir a falecer de 1,765. Entretanto, deve-se ter cuidado com essas interpretações, uma vez que o modelo não está bem ajustado.

Tabela 4 – Estimativas das razões de chances e os limites de confiança

		Limites de confiança	
Efeito	Ponto estimado	Inferior	Superior
Sexo	**1,765**	0,848	3,676
Idade	0,986	0,957	1,015
Diária	**0,914**	0,724	1,153
Atendimento	1,332	0,980	1,809

4.2 Utilizando o método Bayesiano

Para analisarmos os dados de uma perspectiva Bayesiana, é considerando o modelo de regressão logística considerando as 151 observações e as seguintes densidades *a prioris* para os parâmetros do modelo, $\beta_0 \sim N(0, 10^6)$, $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$, $\beta_3 \sim N(0, 10^6)$ e $\beta_4 \sim N(0, 10^6)$. Com essa escolha foram geradas duas cadeias paralelas cada uma com 15.000 iterações; monitorou-se a convergência da amostras Gibbs usando o método de Gelman e Rubin (1992) que utiliza a técnica de análise de variância para determinar se mais iterações são necessárias. Para cada parâmetro as 5.000 primeiras iterações foram descartadas para eliminar o efeito do valores iniciais e daí foram tomadas amostras de 10 em 10 o que totaliza uma amostra final de tamanho 2.000. O amostrador de Gibbs foi implementado usando o aplicativo Winbugs. Na Tabela 5, reportamos o resumo *a posteriori* dos parâmetros do modelo juntamente com os resultados da estimativa dos fatores de redução de escala potêncial $\hat{R}$ (veja, Gelman e Rubin, 1992), para todos os parâmetros. Observamos valores bastante próximos de *um,* o que indica que as cadeias convergiram. Na Tabela 5, observamos também que os intervalos de credibilidade de 95% de β_1, β_2, β_3 e β_4 inclui o valor zero, portanto, podemos concluir que a contribuição das variáveis regressoras associadas a esse coeficiente não é significativa no modelo de regressão logística, resultado similar é obtido na análise clássica do modelo de regressão (veja a Tabela 3).

Tabela 5 – Sumário *a posteriori* para o modelo logístico

	Média	Mediana	DP	Intervalo de credibilidade (95%)	$\hat{R}$
β_0	-1,527	-1,526	0.447	(-2,404,-0,6542)	1,005
β_1	0.5939	0.5975	0.385	(-0.1565, 1.359)	0.9895
β_2	-0.02351	-0.06728	0.01875	(-0.06728, 0.00342)	1.015
β_3	-0.1125	-0.1061	0.125	(-0.3916 , 0.1126)	0.9986
β_4	0.3004	0.2976	0.1587	(-0.007836,0.6252)	0.9994
DIC			182.1		

Na Figura 1, temos as densidades *a posteriori* marginais aproximadas para o modelo de regressão logística.

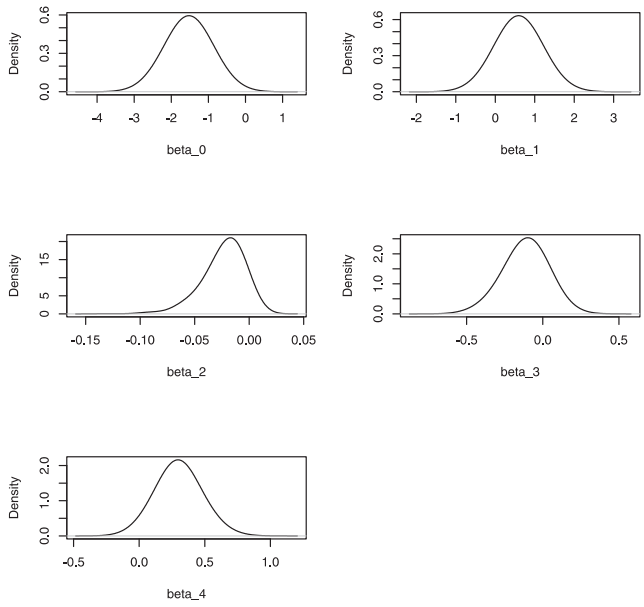

Figura 1 - Densidades marginais *a posteriori* aproximadas dos parâmetros do modelo de regressão logística.

4.3 Influência local

Conforme o esquema de perturbação de casos, temos que: $C_{d_{max}} = 2.746262$.

Na Figura 2, é apresentado o gráfico do autovetor correspondente à $C_{d_{max}}$ e as observações 11, 17, 19 e 76 são as que se destacam das demais.

Já a Figura 3, referente à influência local total, as observações que se destacam são 11, 17, 19, 23, 67 e 76.

4.3.1 Gráfico de envelopes

Nesta parte é apresentado o gráfico de envelopes. Na Figura 4, verifica-se que todos os pontos caem dentro da banda de confiança, apesar de haver uma pequena separação em dois grupos e que os indivíduos 11, 17, 19 e 76 aparecem distante dos demais.

4.4 Reanálise dos dados

Para reanálise dos dados são retirados os possíveis pontos influentes 11, 17, 19 e 76.

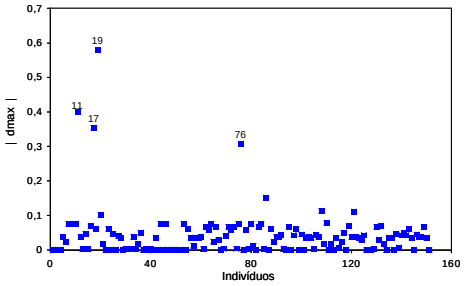

Figura 2 - Gráfico de influência - ponderação de casos.

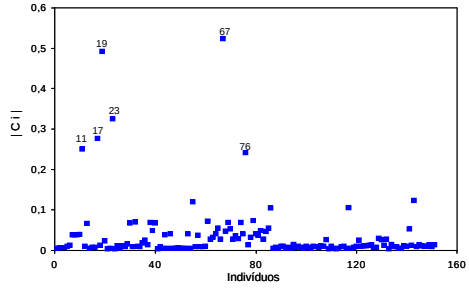

Figura 3 - Gráfico de influência local do i-ésimo indivíduo.

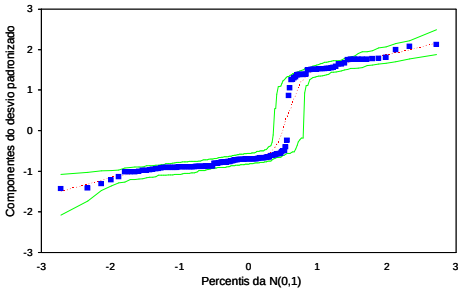

Figura 4 - Gráfico de envelopes para a componente do desvio.

4.4.1 Resultados pelo método de máxima verossimilhança

Os resultados da reanálise são apresentados na Tabela 6.

Tabela 6 – Estatísticas da razão da verossimilhança, Escore e Wald

Estatísticas	p-Valor
Razão da verossimilhança (Λ)= 19,0334	0,0008
Escore (Es)= 15,3500	0,0040
Wald (W)= 12,1444	0,0163

Observando a Tabela 6, verifica-se claramente uma significância alta para rejeitar a hipótese nula, o que significa que pelo menos uma das covariáveis é significativa para o modelo.

Na Tabela 7 são apresentadas as estimativas de máxima verossimilhança.

Tabela 7 – Estimativas dos parâmetros, erro padrão, estatística Wald e p-valor

Efeito	Parâmetro	Estimativa	Erro padrão	Estat. Wald	p-Valor
Intercepto	β_0	-1,5235	0,4584	11,0481	0,0009
Sexo	β_1	0,7562	0,4028	3,5250	0,0604
Idade	β_2	-0,0147	0,0145	1,0273	0,3108
Diária	β_4	-0,7088	0,3055	5,3846	0,0203
Atendimento	β_5	0,2894	0,1729	2,8025	0,0941

$Deviance = 151,056$ com 142 g.l.

Nota-se que, considerando um nível de 5%, a variável diária passa ser significativa. As variáveis sexo e atendimento passariam a ser significativas considerando um nível de 7% e 10% respectivamente. Verifica-se também, que a *deviance* diminuiu, indicando um bom ajuste do modelo.

A tabela das razões de chances estimadas é dada por:

Tabela 8 – Estimativas das razões de chances e limites de confiança

		Limites de confiança	
Efeito	Ponto estimado	Inferior	Superior
Sexo	**2,130**	0,967	4,691
Idade	0,985	0,958	1,014
Diária	**0,492**	0,270	0,896
Atendimento	1,336	0,952	1,874

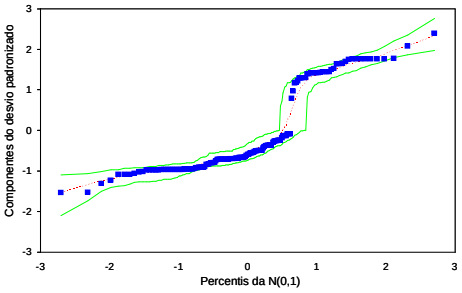

Figura 5 - Gráfico de envelopes para a componente do desvio.

Na Tabela 8 percebe-se que a variável sexo continua sendo um fator de risco e a variável diária um fator de proteção em relação a variável óbito dos animais, sendo que a chance de um animal macho vir a falecer aumentou para 2,130.

No gráfico de envelopes nota-se uma melhor distribuição das observações dentro da banda de confiança, sugerindo ser um ajuste adequado.

4.4.2 Resultados pelo método Bayesiano

Como na subseção 4.2, ajustamos os dados de Paula Roberta Mendes ao modelo de regressão logística sem considerar as observações 11, 17, 19 e 76. Considerando as densidades *a priori* para os parâmetros do modelo, $\beta_0 \sim N(0, 10^6)$, $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$, $\beta_3 \sim N(0, 10^6)$ e $\beta_4 \sim N(0, 10^6)$, geramos duas cadeias separadas de Gibbs, cada uma com 15.000 iterações e utilizamos o método de Gelman e Rubin (1992) para verificar a convergência das cadeias. Para cada parâmetro as 5.000 primeiras iterações foram descartadas para eliminar o efeito dos valores iniciais e foram tomadas amostras de 10 em 10, o que totaliza uma amostra final de tamanho 2.000.

As quantidades *a posteriori* de interesse obtidas a partir das amostras selecionadas são dadas na Tabela 9, na qual observamos que os fatores de redução potencial são menores de 1.1 ($\sqrt{\hat{R}} < 1.1$), indicando a convergência das amostras geradas.

Tabela 9 – Sumário *a posteriori* para o modelo logístico retirando as observações 11, 17, 19 e 76.

	Média	Mediana	DP	Int. cred. (95%)	Int. cred.(90%)	$\hat{R}$
β_0	-1.543	-1.536	0.4588	(-2.466,-0.6585)	(-2.316, -0.7497)	0.9917
β_1	0.7825	0.7855	0.4083	(-0.06175, 0.002584)	(0.1432, 1.484)	1.005
β_2	-0.02209	-0.06175	0.01641	(-0.06175, 0.002584)	$(-0.0539, 5.474 \times 10^{-5})$	1.003
β_3	-0.8465	-0.8043	0.3508	(-1.706 , -0.2973)	(-1.527,-0.3581)	1.010
β_4	0.3042	0.3017	0.176	(-0.05172,0.6592)	(0.02062, 0.5967)	1.006
DIC				161.1		

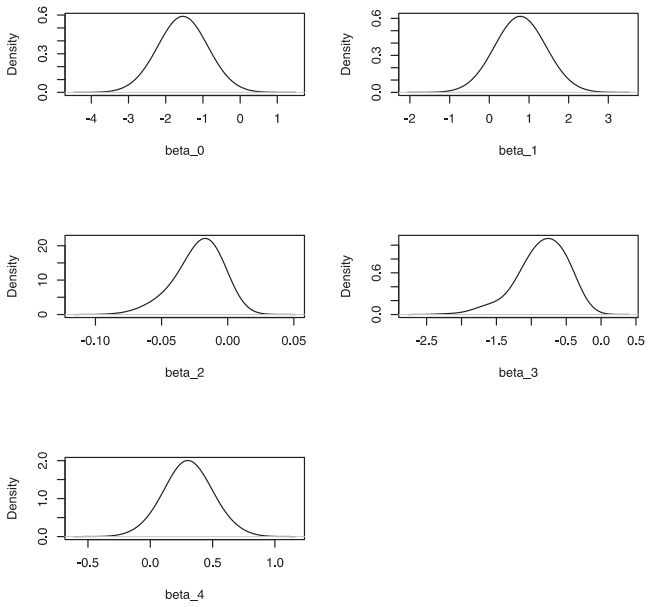

Figura 6 - Densidades marginais *a posteriori* aproximadas dos parâmetros do modelo de regressão logística sem as observaçoes 11, 17, 19 e 76.

Na Figura 6, temos as densidades marginais *a posteriori* aproximadas considerando os 2.000 pontos amostrais gerados.

Da Tabela 9 podemos observar a significância das variáveis sexo, diária e atendimento.

Conclusões

Neste trabalho, discutiu-se a aplicação da teoria de influência local, proposta por Cook (1986), no modelo de regressão logística. Foram obtidas matrizes necessárias para a aplicação da técnica, considerando dois tipos de perturbação nos elementos dos dados e no modelo. Aplicando-se estes resultados em conjunto de dados, obtiveram-se indicações de quais observações ou conjunto de observações influenciam de maneira sensível os resultados da análise. Este fato foi ilustrado por um conjunto de dados reais, sendo verificado que para alguns esquemas de perturbação, a presença de algumas observações pode modificar consideravelmente os níveis de significância de certas covariáveis, tanto utilizando as estimativas de máxima verossimilhança como método Bayesiano.

Finalmente, os resultados das aplicações indicam que o uso da técnica de influência local no modelo de regressão logística é útil na deleção de possíveis

pontos influentes, este fato pode ser verificado tanto utilizando as estimativas de máxima verossimilhança pelo método Bayesiano. Assim, a técnica de influência local pode ser considerada como uma análise complementar em relação às medidas de diagnóstico, propostas por Pregibon (1981).

SOUZA, E. C.; ORTEGA, E. M. M.; CANCHO, V. G. Application of the local influence analysis in logistic regression models. *Rev. Mat. Estat.,* São Paulo, v.24, n.1, p.127-145, 2006.

- *ABSTRACT: An important stage after the formulation and adjustment of a regression model is the diagnosis analysis. Logistic regression is one of the main methods for modeling data and even when the response of interest is not originally of the binary type, some researchers have dichotomized the response in a way that the success probability can be modeled through logistic regression. The estimators obtained for the maximum likelihood and Bayesian methods were used. In this study we consider a study of diagnostic methods with logistic regression, using the local influence technique of Cook (1986). We investigate the application of the local influence technique under different types of disturbance. As an illustration, we show the application of the developed results obtained with real data sets.*

- *KEYWORDS: Logistic regression; diagnostic analysis; local influence; Bayesian method.*

Referências

AGRESTI, A. *An Introduction to categorical data analysis.* New York: John Wiley, 1990. 290p.

ALLISON, P. D. *Logistic regression using the SAS System, theory and application.* Cary: SAS Institute, 1999. 304p.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hasting. *Am. Stat.*, Washington, v.49, p.327-335, 1995.

COLLET, D. *Modelling binary data.* London: Chapman and Hall, 1991, 369p.

COOK, R. R. Assessment of local influence (with discussion). *J. R. Stat. Soc.*, London, v.48, p.133-169, 1986.

COOK, R. R. Detection of influential observations in linear regression. *Technometrics*, Washington, v.19, p.15-118, 1977.

COX, D. R.; HINKLEY, D. V. *Theoretical statistics.* London: Chapman & Hall, 1986. 174p.

COX, D. R.; SNELL, E. J. *Analysis of binary data.*, London: Chapman & Hall, 1989. 236p.

COSTA, S. C. *Regressão Logística aplicada na identificação de fatores de risco para doenças em animais domésticos.* 1997. 104f. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1997.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados em experimentação agronômica.* Piracicaba: CALQ, Departamento Editorial, 2002. 113p.

GALEA, M.; BOLFARINE, H.; VILCA LABRA, F. Influence diagnostics for the structural error-in-variables model under the Student-t distribution. *J. Appl. Stat.*, Abingdon, v.29, p.1191-1204, 2002.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, New York, v.85, p.398-409, 1990.

GELMAN, A.; RUBIN, D. B. Inference from iterarive simulating using multiple sequences (with discussion). *Stat. Sci.*, Hayward, v.7, p.457-511, 1992.

HINDE, I.; DEMÉTRIO, C. *Overdispersion models and estimation*: livro texto de minicurso do 13º SINAPE). Caxambu: ABE, 1998. 73p.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression.* New York: John Wiley, 1989. 307p.

HOSSAIN, M.; ISLAM, M. A. Application of local influence to the linear logistic regression models. *J. Stat. Sci.*, v.51, n.2, p.269-278, 2003.

KLEINBAUM, D. G. *Logistic regression: a self-learning text.* New York: Springer-Verlac, 1994. 278p.

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

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

PREGIBON, D. Logistic regression diagnostics. *Ann. Stat.*, Hayward, v.9, p.705-724, 1981.

THOMAS, W.; COOK, R. D. Assessing influence on predictions from generalized linear models. *Technometrics*, Washington, v.32, p.59-65, 1990.

Recebido em 18.01.2006.

Aprovado após revisão em 16.04.2006.