UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Estimação Robusta para o Modelo de Regressão Logística

Autor: Natália Bordin Barbieri
Orientador: Professor Dr. Álvaro Vigo

Porto Alegre, 21 de Dezembro de 2012.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Estimação Robusta para o Modelo de Regressão Logística

Autor: Natália Bordin Barbieri

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Álvaro Vigo
Professora Vanessa Bielefeldt Leotti Torman

Porto Alegre, 21 de Dezembro de 2012.

Dedico este trabalho a minha família, a qual constantemente se faz presente em cada dia da minha vida.

“Se não puder se destacar pelo talento, vença pelo esforço.”

Dave Weinbaum

Agradecimentos

Agradeço aos meus pais, Jeferson Barbieri e Marfisa Barbieri, que sempre me apoiaram, incentivaram, e acreditaram que eu era capaz, mesmo em momentos em que nem eu acreditei. Vocês são a minha base, meu tudo. Muito obrigada!

Agradeço aos meus irmãos, Renata e Tomás, por serem sempre esses companheiros queridos, e por me esperarem a cada final de semana em casa. Tenho muito orgulho de ser irmã de vocês. Vocês também são meu tudo! Muito obrigada!

Agradeço ao Lázaro Ribeiro, que acompanhou minha trajetória na UFGRS praticamente desde o começo e foi, além de namorado, um grande companheiro nos momentos em que mais precisei. Muito obrigada!

Agradeço aos amigos que tive a honra de encontrar na UFRGS. Em especial, a Letícia Herrmann, pela amizade que se iniciou já no dia da matrícula, e por ter se tornado uma irmã para mim, presente no meu dia a dia; a Paula Sientchkovski, uma amiga com um coração enorme, que sempre esteve presente e que nunca mediu esforços em prol da nossa amizade, ao Paulo Correa, que sempre me impulsionou e ajudou a concluir etapas, sempre estando presente. Também gostaria de agradecer ao Mateus Becker e ao Andrigo Rodrigues, por terem feito parte da minha história.

Agradeço a todos os amigos que, longe ou perto, entenderam a minha ausência, e me acolheram sempre que pude estar de volta; em especial a Mana Kaefer, Priscila Lawrenz e Carla Schneider.

Agradeço ao Professor Álvaro Vigo, por ter aceitado ser meu orientador, e ter sido um exemplo como pessoa e profissional a se seguir.

Agradeço também a professora Vanessa Torman, por ter aceitado ser minha banca, e ter colaborado de maneira tão positiva e construtiva com o meu trabalho; e a Professora Patrícia Ziegelmann, pela ajuda no decorrer do curso.

Agradeço ao Rodrigo Coster, que mesmo sem me conhecer, ajudou na finalização deste trabalho, mostrando um exemplo de profissional a ser seguido.

Agradeço á aqueles que de alguma maneira colaboraram no meu crescimento profissional, em especial a todos do ELSA-Brasil, pelo apoio nesta etapa final; aos profissionais da Souza Cruz, que possibilitaram diversas experiências, e a oportunidade de participar no grupo de pesquisa da Irani Argimon, onde a

curiosidade era a ferramenta mais importante, e foi o que me fez perceber, logo no início do curso, a importância da Estatística nas mais diferentes áreas.

Resumo

Desfechos dicotômicos são muito comuns em várias áreas do conhecimento, particularmente, na pesquisa clínica e epidemiológica. O modelo de regressão logística tem sido amplamente utilizado para identificar fatores associados com o desfecho, bem como para estimar associações por meio da medida de razão de chances.

Quando existem preditores quantitativos, relativamente comuns em alguns contextos, são necessários cuidados adicionais na etapa de diagnóstico do modelo para minimizar potenciais vieses decorrentes de observações influentes usualmente associadas a observações com valores extremos nos preditores contínuos.

O objetivo do trabalho é apresentar aspectos do diagnóstico do modelo de regressão logística, métodos robustos e procedimentos computacionais para o ajuste do modelo de regressão logística robusta, visando minimizar vieses nas estimativas de associação.

A macro ***robust*** do programa SAS e as funções ***glmrob*** e ***glmRob*** do programa R incorporam estimadores robustos para regressão logística e são ferramentas úteis para minimizar o impacto de valores extremos nos preditores. A partir de exemplos, sintaxes SAS e R mostram, passo a passo, etapas para ajuste do modelo e interpretação dos resultados.

Sumário

1 Introdução

Desfechos dicotômicos são muito comuns em várias áreas de conhecimento, particularmente, na pesquisa clínica e epidemiológica. O modelo de regressão logística tem sido amplamente utilizado para identificar fatores associados com o desfecho, bem como para estimar associações por meio da medida de razão de chances.

Preditores quantitativos também são comuns nesses contextos exigindo do pesquisador cuidados adicionais na etapa de diagnóstico do modelo, no sentido de minimizar potenciais vieses decorrentes de observações influentes usualmente associadas a observações com valores extremos nos preditores contínuos.

Neste trabalho são apresentados aspectos do diagnóstico do modelo de regressão logística bem como de métodos robustos para corrigir potenciais distorções.

Os objetivos do trabalho são apresentados no próximo capítulo. No capítulo 3, são brevemente descritas as definições básicas sobre o modelo de regressão logística e de medidas e procedimentos gráficos para diagnóstico da regressão. O capítulo 4 explora definições básicas sobre robustez, medidas de robustez, estimação robusta e métodos robustos para regressão logística. No capítulo 5 são apresentados procedimentos computacionais para o ajuste do modelo de regressão logística utilizando métodos robustos de estimação. No capítulo 6 foi utilizado um conjunto de dados hipotéticos no contexto epidemiológico para ilustrar, passo a passo, etapas de ajuste do modelo e diagnóstico e, após a identificação de observações potencialmente influentes, o ajuste do modelo de regressão logística robusta. No capítulo 8 são apresentadas as sintaxes utilizadas no R e no SAS para o ajuste dos modelos, bem como, a macro ***robust*** e a função ***glmrob*** para o ajuste do modelo logístico robusto no SAS e R, respectivamente.

2 Objetivos

Objetivo geral

Explorar e divulgar a aplicação de métodos robustos para o modelo de regressão logística para minimizar o impacto de valores extremos (*outliers*) de preditores quantitativos.

Objetivos específicos

- Apresentar conceitos básicos de robustez e de estimação robusta;

- Explorar aspectos computacionais dos programas SAS e R para o ajuste do modelo logístico e das medidas de diagnóstico do modelo para identificar observações influentes;

- Explorar uma macro SAS que utiliza métodos robustos para regressão logística;

- Explorar as funções ***glmrob*** e ***glmRob*** dos pacotes ***robustbase*** e ***robust***, respectivamente, do software R, que utilizam métodos robustos para regressão logística; e,

- Exemplificar a utilização dos métodos robustos para regressão utilizando um conjunto de dados hipotéticos e rotinas computacionais dos programas R e SAS.

3 Regressão Logística

No contexto clínico e epidemiológico é muito comum estudos com desfecho dicotômico. O modelo de regressão logística ainda é uma ferramenta importante para descrever a relação entre resposta e os preditores.

No modelo de regressão logística a variável dependente (Y) geralmente representa a ocorrência ou não de um evento de interesse, que sem perda de generalidade pode ser representado pelas categorias designadas por sucesso $(Y=1)$ ou fracasso $(Y=0)$. Essa variável assume distribuição de probabilidade Bernoulli e a probabilidade de sucesso varia com os valores observados para os preditores $\mathbf{x}=(x_1, x_2,..., x_p)'$. Assim, para o i-ésimo indivíduo, a probabilidade de sucesso é representada por $\pi(\mathbf{x}_i)=P(Y_i=1|\mathbf{x}_i)$ e a probabilidade de fracasso é $P(Y_i=0|\mathbf{x}_i)=1-\pi(\mathbf{x}_i)$, para todo $i=1,2,...,n$. O modelo logístico postula que

$$g(\mathbf{x})=\ln\frac{\pi(\mathbf{x})}{\mathbf{1}-\pi(\mathbf{x})}=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p$$

No modelo acima, foi aplicado a transformação logito, onde obtém-se uma relação linear entre as variáveis e, assim, $g(\mathbf{x})$ tem algumas das propriedades desejáveis de regressão linear. Por exemplo, $g(\mathbf{x})$ é linear em seus parâmetros, pode ser contínua, e pode variar de $(-\infty,+\infty)$ dependendo apenas do intervalo estabelecido para $\mathbf{x}$ (HOSMER E LEMESHOW, 2000).

Independentemente do delineamento epidemiológico que gerou os dados, o modelo de regressão logística estima a medida de associação chamada de razão de chances (RC) – O*ddsRatio (OR)*, em inglês. Esta medida é útil para interpretar a magnitude e direção das associações entre as exposições e a ocorrência do evento que está sendo modelado. Sem perda de generalidade, em um modelo em que não existem termos de interação com a exposição contínua x_j, a razão de chances de ocorrência do evento para um aumento de uma unidade de x_j, considerando os demais preditores constantes, é expressa por

$$e^{\beta_j} = RC\left(X_j = x_j + 1 \times X_j = x_j\right);\ \forall X_i = x_i \text{ fixo e } i \neq j$$

Os parâmetros do modelo usualmente são estimados pelo método da máxima verossimilhança (MV) e a contribuição dos preditores pode ser avaliada pelo teste da razão de verossimilhança (TRV) ou pelo teste de Wald. No entanto, apesar de operacionalmente ser mais simples, o teste de Wald usualmente subestima a contribuição de um preditor (ou conjunto de preditores) na explicação da ocorrência do evento de interesse (HOSMER E LEMESHOW, 2000).

Os estimadores de máxima verossimilhança são sensíveis a presença de valores extremos (*outliers*), e assim a realização de diagnóstico e análise de resíduos no modelo de regressão logística é uma etapa fundamental (Heritier, 2009). A presença de valores extremos pode gerar estimativas instáveis de parâmetros e estimativas de erros padrão inflacionadas, podendo comprometer ou impossibilitar inferências baseadas nos intervalos de confiança ou valores p associados. Métodos para análises de resíduos estão extensamente descritos na literatura, como em Hosmer e Lemeshow (2000), e disponíveis em diversos procedimentos computacionais, tais como SAS, R e SPSS. Nos manuais do SAS, R, e SPSS é possível encontrar uma vasta descrição dos pacotes existentes em cada um deles. Estes procedimentos são abordados com maiores detalhes no Capítulo 5.

É importante ressaltar novamente a importância de detectar potenciais observações influentes no modelo, uma vez que podem gerar estimativas viesadas ou não realistas dos coeficientes de regressão. Como a estimativa de razão de chances depende diretamente dos coeficientes de regressão, as observações influentes possuem o potencial de subestimar ou superestimar a associação ou impacto dos preditores.

Por outro lado, depois de identificar as observações influentes alguma medida de correção deve ser realizada. A decisão simplista de excluir as observações influentes (exceto nos casos que não são plausíveis) ou os preditores com observações influentes poderá limitar a capacidade de generalização do modelo. Além disso, a exclusão arbitraria de observações pode levar a uma diminuição de poder estatístico ou estabilidade das estimativas.

Uma alternativa é atribuir pesos pequenos para essas observações, para que se torne possível encontrar estimativas estáveis e realizar inferências mais próximas da realidade, sem que haja o comprometimento da validade do estudo.

Na regressão logística, assim como em regressão linear, o diagnóstico de ajuste do modelo é realizado através da diferença entre o valor observado e o valor ajustado. No modelo logístico, existem várias maneiras de medir esta diferença. Hosmer e Lemeshow (2000) abordam os resíduos de Pearson e os resíduos da função desvio (*Deviance)*. Heritier (2009) apresenta os resíduos padronizados de *Pearson (Pearson standardized residuals)* e os resíduos padronizados da função desvio (*deviance standardized residuals*) como alternativas. Esses são calculados através dos elementos da diagonal principal da matriz da predição ou matriz chapéu (*Hat matrix*).

Procedimentos gráficos são de extrema importância. Definições formais dessas medidas, e como elas estão implementadas em diferentes procedimentos computacionais, estão disponíveis no manual dos respectivos programas.

A distância de Cook é uma medida de impacto da exclusão da análise de uma determinada observação (Heritier, 2009). É comum analisar as alterações nos coeficientes individuais devido a casos específicos identificados como influentes, com $D_i > 0{,}5$, sendo D_i a distância de Cook. E é sempre importante estudar os casos em que $D_i > 1$.

A medida de influência $DFBETA$ mede o impacto de uma observação particular i, em uma regressão específica estimada $\hat{\beta}_j$. A sua estatística representa a mudança padronizada em $\hat{\beta}_j$ quando a i-ésima observação é excluída da análise, ou seja, avalia a influência de uma dada observação na estimação dos parâmetros.

Outra estatística de influência é $DFFITS_i$. Ela mede o impacto de uma observação sobre o valor de resposta previsto na observação, obtidos a partir da regressão logística múltipla. Esta medida tem relação com a Distância de Cook, uma vez que $D_i = (DFFITS_i)^2 \frac{s(-i)}{tr(H)s^2}$, onde s^2 é soma dos quadrados dos resíduos

obtidos a partir da análise de regressão, incluindo todas as observações. Valores absolutos maiores que $2\sqrt{tr(H/n)}$ devem ser analisados atentamente (HOSMER E LEMESHOW, 2000).

Depois de identificados os potencias pontos de influência, é preciso tomar uma decisão mediante a presença dos mesmos. Para tal, o uso de métodos robustos na regressão logística tem sido utilizado com mais frequencia nos últimos anos. (Farcomeni e Ventura, 2012). No próximo capítulo será abordada as definições de robustez, medidas de robustez, estimação robusta, e estimação robusta na regressão logística.

4 Robustez

A expressão "robustez" usualmente é utilizada para designar que um determinado método de análise estatística não é sensível a pequenas violações (ou desvios) das suposições. A situação mais típica se refere a potenciais desvios da forma da distribuição de probabilidade assumida (*distributional robustness*), mas também pode estar associado a outros tipos de exigências ou suposições, tais como independência, mesma distribuição ou procedimento de aleatorização (HUBER, 1996).

Uma forma relativamente comum de violação da distribuição de probabilidade postulada é a contaminação da amostra com valores extremos (*outliers*); dependendo da quantidade, a cauda da distribuição pode se tornar longa, inflacionando a estimativa do desvio padrão. Assim, na literatura as expressões "*distributional robust*" e "*outlier resistant*" na prática são utilizadas como sinônimos (HUBER, 1996).

Métodos robustos começaram a surgir na década de 60, com o objetivo de minimizar o impacto de valores extremos nas estimativas dos parâmetros. Desde então, o desenvolvimento de métodos robustos tem crescido rapidamente, nas mais diversas áreas (HERITIER, 2009).

A identificação de valores extremos e, quando presentes, a utilização de métodos robustos de análise são aspectos importantes para produzir resultados acurados e precisos. Isto porque a decisão simplista de excluir valores extremos com a aplicação de procedimentos de análise aos dados remanescentes pode ser desastrosa, comprometendo a capacidade de generalização dos resultados (FARCOMENI E VENTURA, 2012)

Mesmo com o avançado e crescente desenvolvimento teórico, o uso de métodos robustos ainda tem sido negligenciado em muitas áreas. Na presença de valores extremos, por exemplo, ainda é comum a exclusão de parte das observações (atípicas) ou a substituição de um procedimento inferencial paramétrico por um método não paramétrico. Outras alternativas frequentemente utilizadas são

os métodos de inferência (testes de hipóteses ou estimação por intervalo) baseados em simulações, tais como *bootstrap* ou *jackknife* (FARCOMENI E VENTURA, 2012).

Entretanto, na presença de valores extremos os métodos robustos podem ser considerados uma escolha melhor, pois as observações podem ser calibradas para ter uma pequena perda de eficiência em relação aos testes paramétricos, e também são mais resistentes a algumas violações das suposições (FARCOMENI E VENTURA, 2012).

A estatística robusta busca produzir estimadores que possam ser considerados consistentes e razoavelmente eficientes, estatísticas de teste com nível estável e poder considerável, quando o modelo não é bem especificado (HERITIER, 2009).

Atualmente métodos robustos estão disponíveis em vários programas de análise estatística, tais como SAS, R e Stata. No programa SPSS é possível instalar um módulo do programa R para análises robustas (IBM SPSS, 2010).

Os procedimentos de análise robusta mais frequentes nos procedimentos computacionais contemplam métodos para estimação e/ou comparações de médias e regressão linear. Entretanto, no contexto clínico e epidemiológico é frequente a presença de desfechos dicotômicos, e existem poucos procedimentos computacionais que incorporam métodos robustos para modelos de respostas dicotômicas.

4.1 Medidas de Robustez

Existem diferentes maneiras dedefinir e medir robustez. Definições formais estão além dos objetivos deste trabalho e podem ser encontradas, por exemplo, em Huber (1996), Heritier et al (2009) ou Farcomeni e Ventura (2012).

Huber (1996) aborda a definição de robustez sob três aspectos: qualitativo, quantitativo e infinitesimal. O conceito qualitativo está embasado no princípio de continuidade fundamental de robustez, que postula que pequenas perturbações na distribuição de probabilidade subjacente deveriam causar pequenas mudanças no desempenho do método estatístico utilizado na análise.

A definição de robustez quantitativa está embasada no conceito de ponto de ruptura (*breakdown point - BP*), e mede em um sentido global, a propriedade de robustez de uma estatística T. O ponto de ruptura é definido como a quantidade máxima de má especificação do modelo probabilístico que um estimador pode resistir antes de "quebrar" (*breakdown*), isto é, antes do estimador produzir resultados absurdos.

A definição infinitesimal é baseada no conceito de função de influência (*influence function*). Considere $X_1, X_2, ..., X_n$ uma amostra aleatória de *n* observações independentes e identicamente distribuídas de uma distribuição F_θ, e $Tn = Tn(X_1, ..., X_n)$ um estimador para o vetor de parâmetros θ. Uma mudança ou contaminação suficientemente pequena no processo que gera os dados provenientes da distribuição F_θ pode resultar em uma mudança arbitrariamente pequena na estimativa do parâmetro. Uma contaminação pequena, neste contexto, significa que os dados observados pertencem a uma vizinhança da função de distribuição F_θ, ou seja, $F_\varepsilon = (1-\varepsilon)F_\theta + \varepsilon G$, em que G é uma função de distribuição arbitrária e $0 \le \varepsilon \le 1$. A função de influência IF é definida como

$$IF(x;T,F_\theta) = \lim_{\varepsilon \to 0} \frac{T(F_{\varepsilon,\Delta}) - T(F_\theta)}{\varepsilon} = \left. \frac{\partial T(F_{\varepsilon,\Delta})}{\partial \varepsilon} \right|_{\varepsilon=0}$$

em que $T(F_{\varepsilon,\Delta}) = T((1-\varepsilon)F_\theta + \varepsilon\Delta_x)$, com Δ_x massa de probabilidade no ponto x, tal que, $\Pr(\Delta_x = x) = 1$.

O supremo da função de influência, chamado de sensibilidade a erros grosseiros (*Gross error sensivity – GES*) mede a pior influência sobre a estatística T. Uma propriedade de robustez desejável é GES finita, ou seja, que a função de influência IF seja limitada (*B-robustness*) (FARCOMENI E VENTURA, 2012).

Outra medida de robustez derivada da função de influência IF é a sensibilidade de deslocamento local (*local-shift sensitivity*), que mede a robustez com respeito a efeitos de arredondamento (HERITIER, 2009).

Diferentes estimadores robustos podem ser comparados utilizando o conceito de ponto de rejeição (*reject point - RP*), sendo bastante utilizado no contexto multivariado. O ponto de rejeição é definido como a distância até o centro dos dados, de maneira que aqueles pontos fora desta distância não têm influência no viés assintótico do estimador. Formalmente, para uma distribuição simétrica centrada em *m,* com função de distribuição F_θ, o ponto de ruptura RP é definido como $\inf\{r > 0 : IF(x;T,F_\theta) = 0 \quad onde \quad \delta(x,m) > r\}$ em que δ é uma medida de distância adequada. Se um estimador tem RP finito, então, pontos muito distantes do centro dos dados recebem peso igual a zero (FARCOMENI E VENTURA, 2012).

A estatística robusta tem como objetivo produzir estimadores consistentes e eficientes, assim como testes de hipóteses com nível de significância e poder estáveis na presença pequenos desvios das suposições do modelo (HERITIER, 2009).

A Figura 1 mostra uma interpretação geométrica da relação existente entre o IF, GES e o BP. É possível observar que o enquanto o GES, através da IF, mede uma aproximação de primeira ordem do viés máximo, o BP mede a máxima quantidade de desvios do modelo que o estimador pode suportar antes que seu viés torna-se demasiadamente grande.

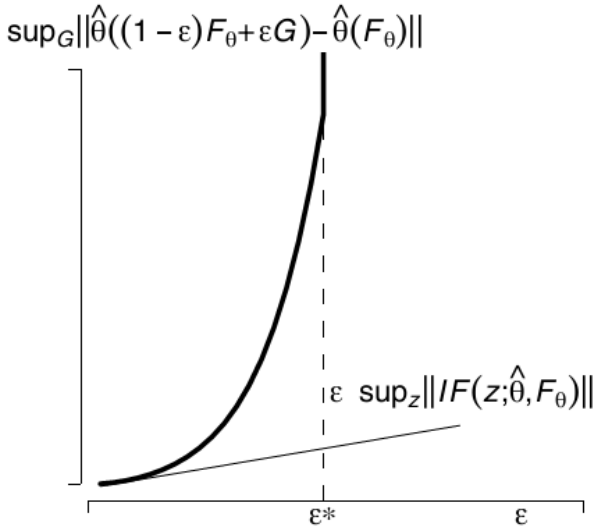

Figura 1: Relação entre IF, GES, e BP, onde ε∗ é o máximo de viés do estimador com má especificação do modelo.
Fonte: Heritier 2009.

4.2 Estimação Robusta

Um estimador é frequentemente membro de uma classe de estimadores que possuem algumas propriedades ótimas, tais como imparcialidade, consistência e eficiência. Os procedimentos clássicos de estimação não têm bom desempenho quando ocorrem pequenas violações. Por exemplo, se $X \sim F_{\boldsymbol{\theta}}$ e $X_1, X_2, ..., X_n$ é uma amostra aleatória desta distribuição, um estimador consistente e eficiente para θ pode ser obtido maximizando o logaritmo da função de verossimilhança; isto é, o estimador de máxima verossimilhança (EMV) representado por $\hat{\boldsymbol{\theta}}_{EMV}$, é a solução de

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}).$$

No caso em que $X \sim N(\mu, \sigma^2)$; $\sigma^2 > 0$, os EMV para $\boldsymbol{\theta} = (\mu, \sigma^2)'$ são dados por $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Assim, $\hat{\boldsymbol{\theta}}_{EMV} = (\overline{X}, S^2)'$ é o EMV para $\boldsymbol{\theta} = (\mu, \sigma^2)'$. No entanto, valores extremos de x (isto é, valores distantes do centro da distribuição μ inflacionam a média e a variância amostral e o estimador

$\hat{\boldsymbol{\theta}}_{EMV} = \left(\overline{X}, S^2\right)'$ é sensível a valores extremos. Como os estimadores baseados no método dos momentos envolvem os momentos amostrais $\overline{X}$ e S^2, em geral produzem estimadores com boas propriedades de robustez local (HERITIER, 2009).

Os métodos de estimação robusta englobam um conjunto de procedimentos que são resistentes a pequenas violações nas exigências do modelo paramétrico assumido. Huber foi o pioneiro nos estudos de estimação robusta, propondo a classe de M-estimadores (*M-Estimators*). A classe de M-estimadores foi estendida para todas as distribuições de probabilidade e generaliza o método da máxima verossimilhança, produzindo estimadores consistentes e assintoticamente normais (HERITIER, 2009).

O estimador de Huber é um caso particular dos M-estimadores. Os M-estimadores estimam as funções da forma:

$$\psi_\beta = \sum_{i=1}^{n} s(x_i)\psi_\beta(r_i) = \sum_{i=1}^{n} w_\beta(r_i) r_i x_i,$$

Onde $r_i = (y_i - x_i^T \beta)/\sigma$ é o i-ésimo resíduo, $s(.)$ e $\psi_\beta(.)$ são funções dadas, e $w_\beta(.)$ são pesos apropriados relacionados com as funções $\psi_\beta(.)$, que tornam $w_\beta(r_i)r_i$ uma função delimitada.

Quando $s(x) = 1$ e $\psi_\beta(.) = \psi_H(.;k)$ tem-se então o estimador de Huber para regressão, que neste caso, $w_\beta(r) = w_\beta^H(r) = \psi_H(r;k)/r = \min(1, k/|r|)$, onde os pesos, que variam de $i = 1,...,n$, podem ser interpretados, e eles automaticamente definem uma medida do quanto a unidade é ou não ou valor extremo. Porém, o estimador de Huber não é robusto em relação aos *bad laverages points* (observações atípicas em relação a X e Y), pois o peso atribuído pelo estimador pode controlar apenas valores extremos.

Uma alternativa é o estimador de Mallows, que através de uma função peso adequada onde $0 \le s(x) \le 1$ e $\psi_\beta(.) = \psi_H(.;k)$, torna-se robusto na presença de qualquer tipo de *outlier*. Quando utilizado $s(x) = 1/|x|$, tem-se então o estimador Hampel-Krasker.

Existem outros estimadores que levam em conta outras medidas de robustez. Como o estimador *Least Median Squares (*LMS), estimador da mínima mediana dos quadrados, que minimiza a mediana dos resíduos, e o *Least Trimmed Squares* (LTS), mínimos quadrados aparados, que minimiza a média aparada dos resíduos, através do conceito de *trimmed mean*. Ambos são baseados na medida de *high break-down point.*

Existe também a classe dos *MM-estimadores (MM-estimators)*, que combina a resistência dos *high break-down point* com a eficiência dos M-estimadores (FARCOMENI E VENTURA, 2012).

Outras classes de estimadores robustos são os R-estimadores (*R-estimators*) e L-estimadores (*L-estimators*) (HUBER, 1996).

Estatísticas robustas têm como objetivo produzir os mesmos resultados na presença ou ausência de *outliers*, devido ao ajuste do modelo a maioria dos dados. Sabemos que muitas vezes *outliers* são originalmente observados e podem ter grandes influências sobre o modelo estimado. Estatísticas robustas têm como propósito lidar com este tipo de discrepância delimitando a influência destes *outliers* para torná-los mais estáveis, evitando assim que os parâmetros do modelo sejam sub ou superestimados (NARGIS, 2005).

Farcomeni e Ventura (2012) apresentam uma revisão de métodos robustos aplicados no contexto da pesquisa clínica e epidemiológica. Além de breve revisão de conceitos de robustez, ilustra a utilização de métodos robustos paraa estimação e comparação de médias, regressão linear, regressão logística e modelo de Cox.

4.3 Estimação Robusta na Regressão Logística

As estimativas dos parâmetros clássicos obtidos por MV podem produzir resultados absurdos devido a pontos de alavancagem ou erro de classificação na resposta (resposta igual a zero ao invés de um, ou vice-versa). Este segundo caso corresponde a um cenário no qual os preditores são erroneamente classificados, embora não distantes na direção X , indicando claramente o resultado oposto (ou

seja, a probabilidade de estimar um valor zero é baixa, mas um valor zero é observado, ou vice-versa). Muitas abordagens para estimação robusta do modelo de regressão logística foram propostas, destacando-se os trabalhos de Pregibon (1982) e Bianco eYohai (1997). Outros métodos deestimação robusta foram derivados para a classe dos modelos lineares generalizados (GLM-*Generalized Linear Models*), tais como os estimadores $OBRE$ *(Optimal Bias-RobustEstimator),* que minimiza o traço da matriz de covariância assintótica sob a restrição de ser uma função de influência limitada, propostos por Künschet al (1989).

Este trabalho explora o estimador do tipo Mallows desenvolvido por Cantoni e Ronchetti (2001), baseado em uma modificação do sistema de equações de estimação, derivadas do estimador de quase-verosimilhança (*quasi-likelihood*). O estimador de Mallows é dado pela solução do sistema de equações de estimação

$$\sum_{i=1}^{n} x_i^T (y_i - \mu_i)\sqrt{V_i} = 0, \tag{1}$$

em que $\mu_i = e^{x_i^T \beta} \big/ (1+e^{x_i^T \beta})$ e $V_i = \mu_i(1-\mu_i), i = 1,...,n$. Cantoni e Ronchetti (2001) sugerem o uso de um esquema de ponderação e a função de Huber $\psi(.;k)$, tal que

$$\sum_{i=1}^{n} w(x_i) x_i^T (\psi_H(r_i;k) - a(\mu_i))\sqrt{V_i} = 0, \tag{2}$$

em que $a(\mu_i) = \psi_H(1-\mu_i \big/ \sqrt{V_i}\,;k)\mu_i + \psi_H\left(-\mu_i \big/ \sqrt{V_i}\,;k\right)(1-\mu_i)$, e $r_i = (y_i - \mu_i)/V_i$ são os resíduos de Pearson.

Note que quando $k \to \infty$ e $w(x_i) = 1$, o lado esquerdo da equação (2) torna-se a função escore do modelo logito, produzindo o EMV clássico. Quando $k < \infty$ e $w(x_i) = 1$, tem-se o estimador de Huber. O termo de correção $a(\mu_i)$ é incluído de modo a assegurar a consistência de Fisher. Note ainda que a equação (2) pode ser vista como uma generalização direta das abordagens robustas para modelos de regressão.

O estimador $\hat{\beta}$, definido como a solução da equação (2), tem função de influência (*IF*) limitada. O efeito de erros de classificação do desfecho é limitado por um valor finito da constante "k", e o efeito de valores extremos na direção dos preditores ***x*** é limitado por uma escolha adequada dos pesos $w(.)$. Uma opção é utilizar $w(x_i) = \sqrt{1 - h_{ii}}$ (onde h_{ii} são os elementos da diagonal da matriz chapéu), ou, quando os preditores forem quantitativos, pode-se utilizar o estimador da matriz de covariâncias de determinante mínimo (*MCD-Minimum Covariance Determinant*) (FARCOMENI E VENTURA, 2012).

Sob nenhuma contaminação, os erros padrões são pouco inflacionados em relação aos EMV, de maneira que se espera uma pequena perda de poder. Por outro lado, sob contaminação (ou seja, na presença de observações atípicas), testes de hipóteses e intervalos de confiança baseados nos EMV não são confiáveis, e associações importantes podem muitas vezes ser mascaradas.

Esses e outros métodos estão implementados nas funções do programa R ***glmrob*** do pacote ***robustbase***, ***glmRob*** do pacote ***robust***, e na macro SAS ***robust*** descritos no Capitulo 5, no qual é ilustrada a utilização de métodos robustos para a regressão logística.

5 Aspectos Computacionais

O modelo de regressão logístico pode ser ajustado em diversos procedimentos computacionais, destacando os programas tradicionais de análise estatística de dados, como SAS, SPSS, STATA e R. Neste trabalho foram abordados alguns aspectos dos programas SAS e R, brevemente descritos nas próximas seções. Este capítulo descreve sucintamente aspectos computacionais destes programas para o ajuste e diagnóstico do modelo logístico. Detalhes podem ser obtidos na documentação dos programas. As sintaxes para o ajuste do modelo e as interpretações dos resultados são exploradas em detalhes no Capítulo 6.

5.1 SAS

No programa SAS o modelo logístico pode ser ajustado em diferentes procedimentos, podendo ser incorporadas diferentes características do delineamento epidemiológico que gerou os dados. Os procedimentos geralmente utilizados são o *PROC LOGISTIC*, *PROC GENMOD* e *PROC GLIMMIX*, porém neste trabalho será abordado somente o primeiro. Informações detalhadas dos métodos disponíveis nestes procedimentos podem ser obtidas na documentação do programa SAS. A documentação completa do SAS está disponível na página de suporte do programa (www.sas.com).

5.1.1 PROC LOGISTIC

O procedimento *PROC LOGISTIC* ajusta modelos de regressão para dados com resposta dicotômica e politômica (nominal ou ordinal). Os parâmetros são estimados pelo método da máxima verossimilhança, utilizando os métodos iterativos de escore de Fisher (*Fisher scoring*) e Newton-Raphson.

Para a análise de resíduos e diagnóstico do modelo estão disponíveis diversas estatísticas e procedimentos gráficos para identificação de observações atípicas ou influentes. Os resíduos estimados são o resíduo de Pearson e o resíduo da função desvio.

A opção *DFBETA* permite fazer uma análise de diagnóstico para cada observação, utilizando a diferença padronizada das estimativas dos parâmetros decorrente da exclusão da observação. Esta análise pode ser visualizada em um painel de gráficos produzido com a especificação da opção *DfBetasPlot*.

As opções *C* e *CBAR* produzem uma análise de diagnóstico do deslocamento do intervalo de confiança, que é uma medida da influência de cada observação nas estimativas dos parâmetros de regressão. Um painel de gráficos é produzido com a especificação da ***dcpplot*.**

Para detectar observações que não estão bem ajustadas, isto é, que contribuem bastante para a discordância entre os valores observados e preditos pelo modelo, utiliza-se as opções *DIFDEV* e *DIFCHISQ*.

No procedimento *PROC LOGISTIC* existem diversas formas de solicitar gráficos e painéis de gráficos para análise de diagnóstico do modelo. Os leverages, por exemplo, podem ser solicitados pela opção ***phat*.** Detalhes sobre medidas e gráficos disponíveis, bem como as correspondentes definições matemáticas e implementação, podem ser obtidos na documentação do procedimento *PROC LOGISTIC*. Exemplos de sintaxes e resultados são explorados no Capítulo 6.

Alternativamente, uma análise gráfica do diagnóstico do modelo pode ser realizada utilizando a macro SAS denominada ***inflogis***. Os gráficos utilizam medidas de influência (*C* e *CBAR*) como tamanho de bolhas, disponibilizando diferentes gráficos baseados nas estatísticas geradas pelas opções *DIFDEV*, *DIFCHISQ*, leverages e matriz de predição. A referida macro e informações adicionais podem ser obtidas na página do autor Michael Frindly, http://www.datavis.ca/sasmac/.

5.1.2 Regressão Logística Robusta

Métodos robustos para o modelo de regressão logística não estão disponíveis nos procedimentos do SAS. Porém, a macro SAS chamada ***robust***, criada por Michael Friendly (disponível na página http://www.datavis.ca/sasmac/robust) disponibiliza alguns métodos, utilizando mínimos quadrados iterativamente reponderados para o ajuste de modelos lineares, por meio dos M-estimadores. Os pesos das observações são determinados por meio

dos métodos de ***Huber***, ***Bisquare***, mínimos valores absolutos (*LAV-Least Absolute Values)* ou mínimos quadrados ordinários (*OLS-Ordinary Least Squares*). A macro permite utilizar dois valores para constante de afinação, especificadas pelo argumento ***tune*** = 6 para o método ***Bisquare*** ou ***tune*** = 2 para o método ***Huber.*** O interesse neste trabalho é explorar o método de Huber. Para o ajuste do modelo logístico, a macro utiliza o procedimento *PROC LOGISTIC*.

5.2 R

Os procedimentos descritos a seguir podem ser encontrados no software R. Os procedimentos descritos nesta seção foram avaliados para a versão 2.15.1. O R é um software livre, e maiores informações podem ser obtidas em sua página (http://www.r-project.org).

5.2.1 Pacotes para Ajuste do Modelo Ordinário eDiagnóstico

É possível ajustar o modelo de regressão logística através da função ***glm***, disponível no pacote ***Stat*** (que já vem com a instalação básica do R). A função ***glm*** utiliza os métodos de máxima verossimilhança e escore de Fisher para a estimação dos parâmetros do modelo. A função ***confint*** estima os intervalos de confiança para os coeficientes de regressão, bem como para a razão de chances. A função ***influence.measures*** disponibiliza diversas estatísticas para realizar o diagnóstico do modelo, como as medidas *DFBETAS, DFFITS*, *Cov.r*, distância de Cook e os valores da diagonal da matriz de predição (*leverage values*), indicando as observações com valores potencialmente influentes.

Uma ferramenta gráfica útil é a função ***influencePlot*** do pacote ***car***, que gera um gráfico dos resíduos padronizados *versus leverages*, salientando as observações atípicas com bolhas de diferentes tamanhos, as quais são proporcionais a distância de Cook (Fox e Weisberg, 2011). As observações potencialmente influentes também são identificadas. Por meio destes gráficos é então possível diagnosticar observações atípicas.

5.2.2 Pacotes para Regressão Logística Robusta

Estão disponíveis para o programa R pacotes com funções específicas para estimação robusta no modelo de regressão logística, tais como a função ***glmrob***, do pacote ***robustbase*** e a função ***glmRob***, do pacote ***robust***. Nessas funções é possível escolher pesos da estimação robusta.

A função ***glmrob*** é usada para ajustar modelos lineares generalizados utilizando métodos robustos para diferentes famílias de distribuições, tais como Binomial, Poisson, Gama e Normal. Com a especificação da opção '*weights*' é possível detectar e ponderar as possíveis observações com valores extremos ou influentes para algum preditor, utilizando os valores da diagonal da matriz de predição ou a distância de *Mahalanobis*. O método ***Mqle*** realiza o ajuste de um modelo linear generalizado por meio dos estimadores do tipo Huber ou Mallows, conforme descritos por Cantoni e Ronchetti (2001).

A função ***glmRob*** do pacote ***robust,*** também pode ser usada para ajustar o modelo logístico, por meio dos estimadores do tipo Mallows, com a especificação da opção ***glmRob.mallows***. No entanto, associado ao uso da função ***glmRob*** tem sido descrito potenciais problemas de estimação quando existem preditores categóricos representados por variáveis de delineamento (*dummies*). Este problema está associado à obtenção de uma matriz singular no processo de estimação pelo método da matriz de covariâncias de determinante mínimo pela função ***mcd*** (pacote ***rrcov)*** utilizada pela função ***glmRob.*** Em face desta limitação, como usualmente preditores categóricos são importantes no contexto clínico e epidemiológico, a função ***glmRob*** será pouco explorada neste trabalho.

No próximo capítulo são mostrados com mais detalhes os procedimentos computacionais descritos acima, utilizando um conjunto de dados do contexto epidemiológico.

6 Aplicação

Este capítulo mostra, passo a passo, o ajuste do modelo de regressão logística utilizando um conjunto de dados hipotéticos. São mostradas rotinas computacionais e resultados dos programas SAS e R, com ênfase em alguns aspectos do ajuste do modelo e métodos robustos. As sintaxes completas estão disponíveis nos anexos.

Os dados retratam um delineamento caso-controle fictício realizado para estimar a associação entre presença de diabetes tipo 2 e inflamação sistêmica. A população em estudo são indivíduos com idade acima de 45 anos residentes em uma determinada comunidade. Neste contexto, 200 indivíduos com diabetes tipo 2 foram selecionados ao acaso da população de diabéticos da referida comunidade, e comparados com 200 indivíduos sem diabetes, selecionados ao acaso da mesma população. Algumas variáveis investigadas, relevantes para exemplificar a utilização de métodos robustos na regressão logística, são descritas no Quadro 1. Aspectos dos métodos utilizados para o diagnóstico de diabetes e para a aferição da exposição, presença de inflamação sistêmica, bem como dos demais preditores, são pouco relevantes neste contexto. Entretanto, embasamento clínico e epidemiológico da plausibilidade da associação podem ser encontrados na literatura (DUNCAN et al, 2003).

Quadro 1 –Descrição das variáveis utilizadas no exemplo hipotético.

Nome da variável	Descrição	Valores ou unidade de medida
DM	Indicador de presença de diabetes tipo 2	0=Não; 1=Sim
INFLAMACAO	Indicador de presença de inflamação sistêmica	0=Não; 1=Sim
SEXOM	Indicador de sexo masculino	0=Feminino; 1=Masculino
RACACOR	Indicador de raça/cor	0=Outra, 1=Branca
HIPERT	Indicador de presença de hipertensão arterial	0=Não; 1=Sim
IMC	Índice de massa corporal	kg/m^2
TRIGT	Triglicerídeos total	mmol/L
RCQ	Resultado da razão entre as medidas de circunferência da cintura (em cm) e do quadril (em cm), multiplicado por 20	-
IDADE	Idade	anos

A subseção 6.1 apresenta uma análise descritiva da amostra e a subseção 6.2, aspectos do ajuste e diagnóstico do modelo de regressão logística. Na subseção 6.3 são explorados aspectos dos métodos robustos.

6.1 Análise Descritiva

A Tabela 1 apresenta uma breve descrição da amostra em estudo, onde os participantes foram divididos entre os que apresentavam diabetes tipo 2, e os que não apresentavam.

Tabela 1: Análise descritiva da amostra

	Com DM (N=200)	**Sem DM (N=200)**
Variável	**Média *(DP)* ou n*(%)***	**Média *(DP)* ou n*(%)***
Inflamação		
Sim	77 *(38,5)*	39 (19,5)
Não	123 *(61,5)*	161 (80,5)
Sexo		
Feminino	125 *(63,5)*	127 (63,5)
Masculino	75 *(37,5)*	73 (36,5)
Raça		
Branca	94 *(47)*	97 (48,5)
Outra	106 *(53)*	103 (51,5)
Hipertensão		
Sim	109 *(54,5)*	64 (32)
Não	91 *(45,5)*	136 (68)
Idade	54,4 *(5,95)*	53,3 (5,61)
IMC *(kg/m2)*	28,8 *(5,41)*	26,5 (5,36)
Triglicerídeos *(mmol/L)*	1,74 *(0,94)*	1,41 (0,83)
RCQ	19,2 *(1,41)*	18,1 (1,6)

Abaixo segue a Figura1 com os *Box Plots* das variáveis IDADE, RCQ, TRIGT e IMC. Para que fosse possível identificar quais são as observações que aparecem como *outliers*, foi utilizada uma função de autoria de*Tal Galili* (disponível

em http://www.r-statistics.com/wp-content/uploads/2011/01/boxplot-with-outlier-label-r.txt), e está incorporada no Anexo 8.2.

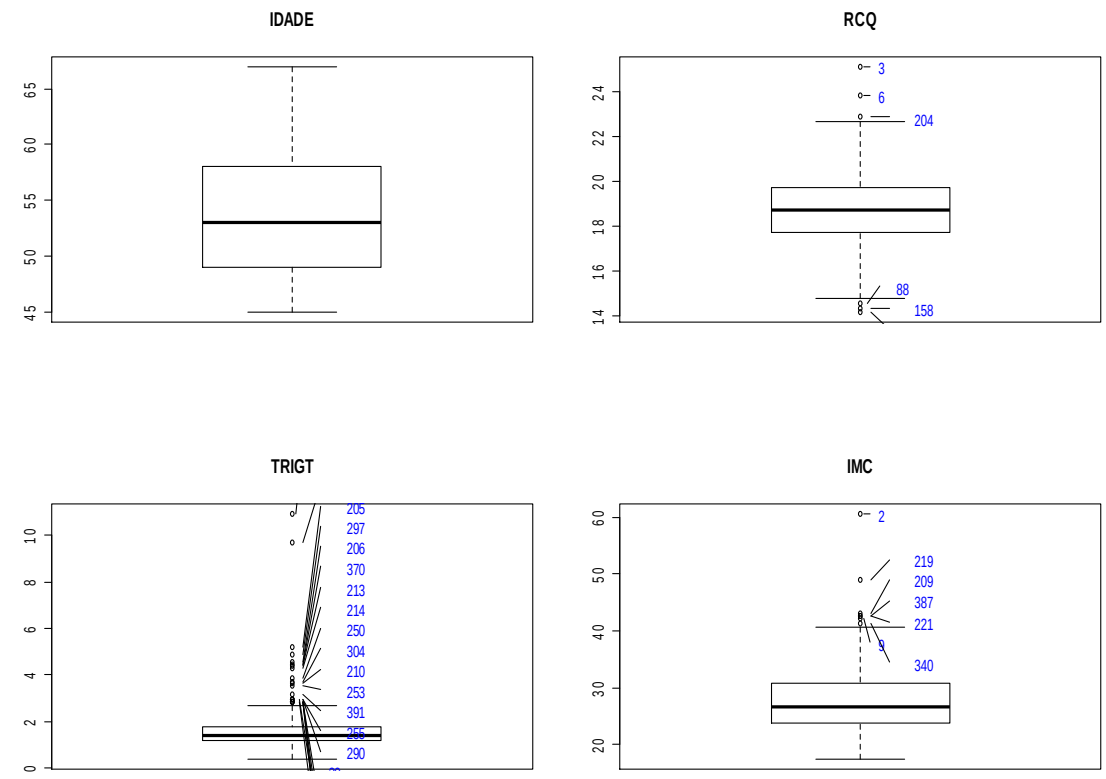

Figura 2: *Boxplots* das variáveis IDADE, RCQ, TRIGT e IMC.

6.2 Regressão Logística Utilizando o Programa SAS

Para ilustrar métodos robustos do modelo de regressão logística foi considerado o modelo multivariável que considera o desfecho presença ou ausência de diabetes tipo 2, a exposição representada pela presença ou ausência de inflamação sistêmica, incluindo também as variáveis sexo, cor de pele/raça, hipertensão arterial, índice de massa corporal, razão cintura quadril e triglicerídeos total . A sintaxe abaixo ajusta o modelo multivariável sem a utilização de métodos robustos. Também requisita estatísticas para o diagnóstico do modelo. A sintaxe completa está no Anexo 8.1.

```
proc logistic data=DM descending plots(only label)=(phat leverage
dpc dfbetas influence);
model DM = INFLAMACAO IDADE SEXOM RACACOR HIPERT IMC RCQ TRIGT / rl;
run;
```

As estimativas dos parâmetros de regressão e de razão de chances são mostradas nos quadros abaixo:

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-10.1235	1.8050	31.4577	<.0001
INFLAMACAO	1	0.6241	0.2544	6.0171	0.0142
IDADE	1	0.0106	0.0201	0.2800	0.5967
SEXOM	1	-0.3566	0.2562	1.9375	0.1639
RACACOR	1	0.00904	0.2558	0.0012	0.9718
HIPERT	1	0.5604	0.2424	5.3453	0.0208
IMC	1	0.0193	0.0251	0.5955	0.4403
RCQ	1	0.4430	0.0921	23.1505	<.0001
TRIGT	1	0.3011	0.1787	2.8405	0.0919

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
INFLAMACAO	1.0000	1.866	1.134	3.073
IDADE	1.0000	1.011	0.972	1.051
SEXOM	1.0000	0.700	0.424	1.157
RACACOR	1.0000	1.009	0.611	1.666
HIPERT	1.0000	1.751	1.089	2.817
IMC	1.0000	1.020	0.971	1.071
RCQ	1.0000	1.557	1.300	1.865
TRIGT	1.0000	1.351	0.952	1.918

As medidas de diagnóstico do modelo foram avaliadas utilizando procedimentos gráficos, mostrados nas figuras 3 e 4. A Figura 3 mostra os resíduos de Pearson e da função desvio, os *leverages* e a medida de deslocamento dos intervalos de confiança, identificados pela ordem das observações no arquivo de dados. A Figura 4 complementa esta análise, com as medidas CBAR, diferença na estatística qui-quadrado e na *deviance* com a exclusão da observação. As observações ordenadas de número #2, #3 e #203 parecem ser observações influentes.

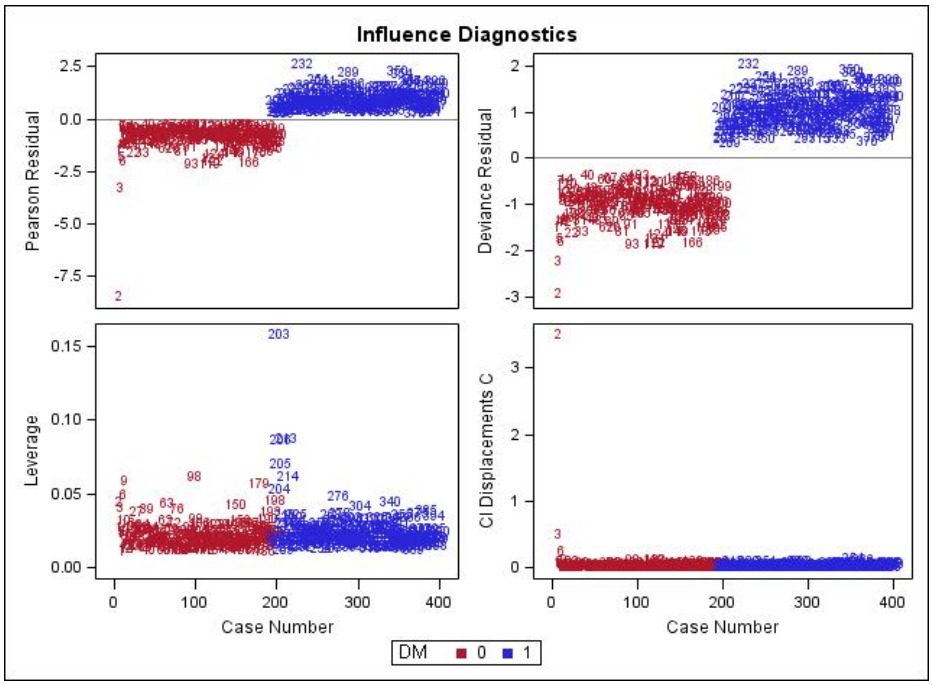

Figura 3 - Diagnóstico de observações influentes.

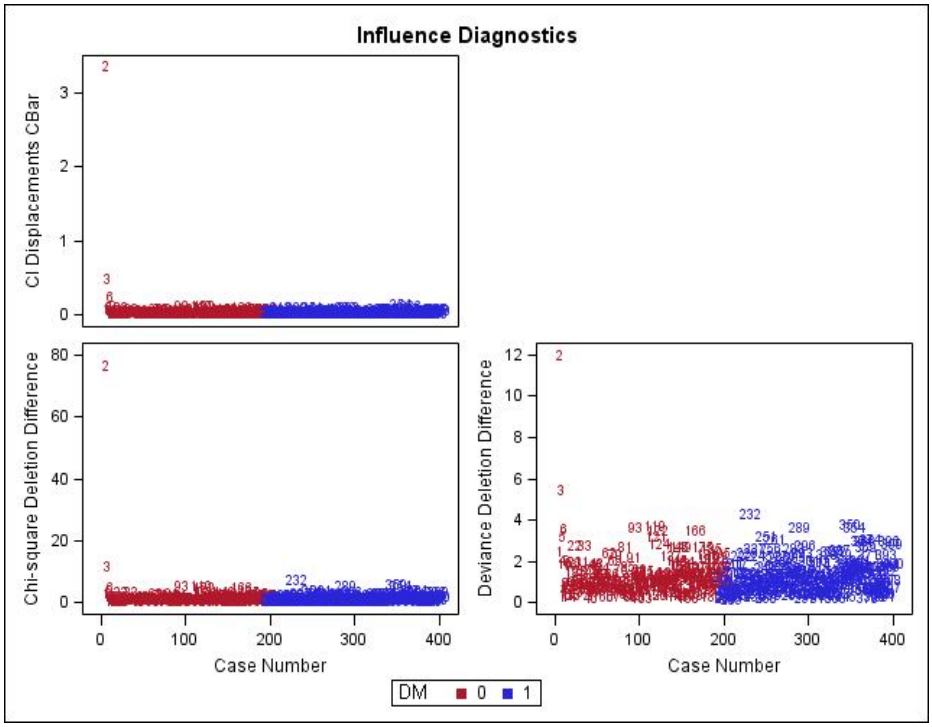

Figura 4 - Diagnóstico de observações influentes.

As Figuras 5, 6 e 7 mostram as medidas de influência baseadas nos *DFBETAS*, sugerindo que a observação #2 pode ser influente com respeito às variáveis IMC e TRIGT, as observações #3 e #6 para a variável RCQ.

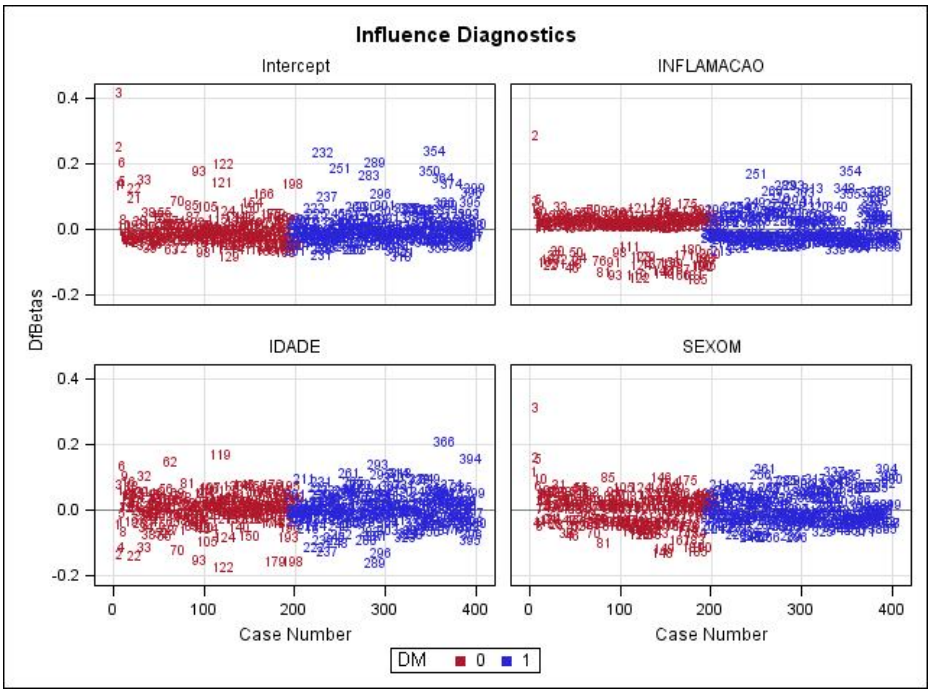

Figura 5- Diagnóstico de observações influentes.

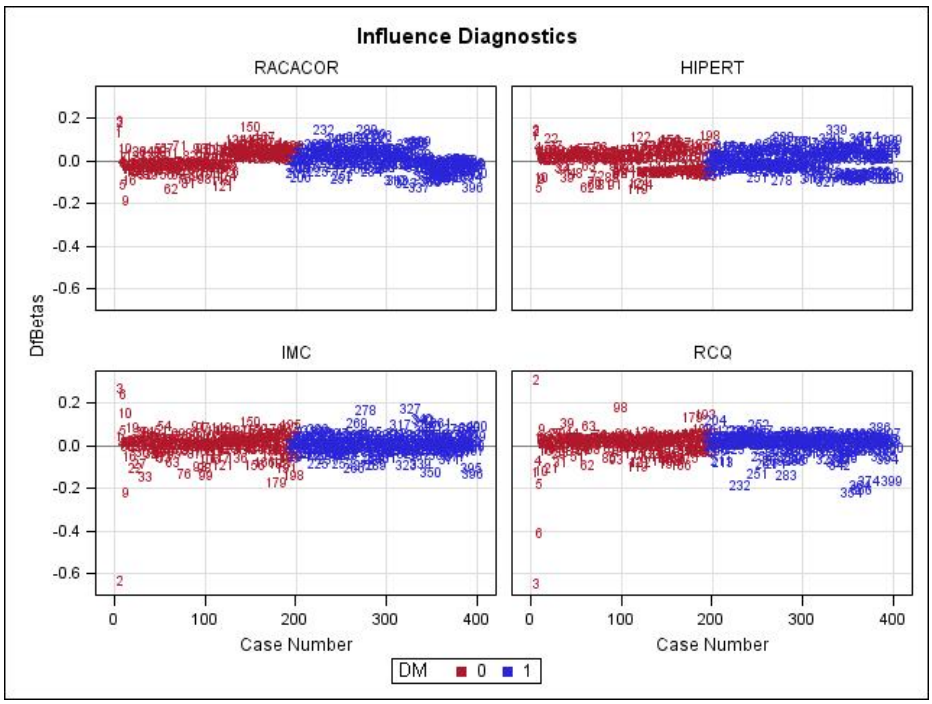

Figura 6- Diagnóstico de observações influentes.

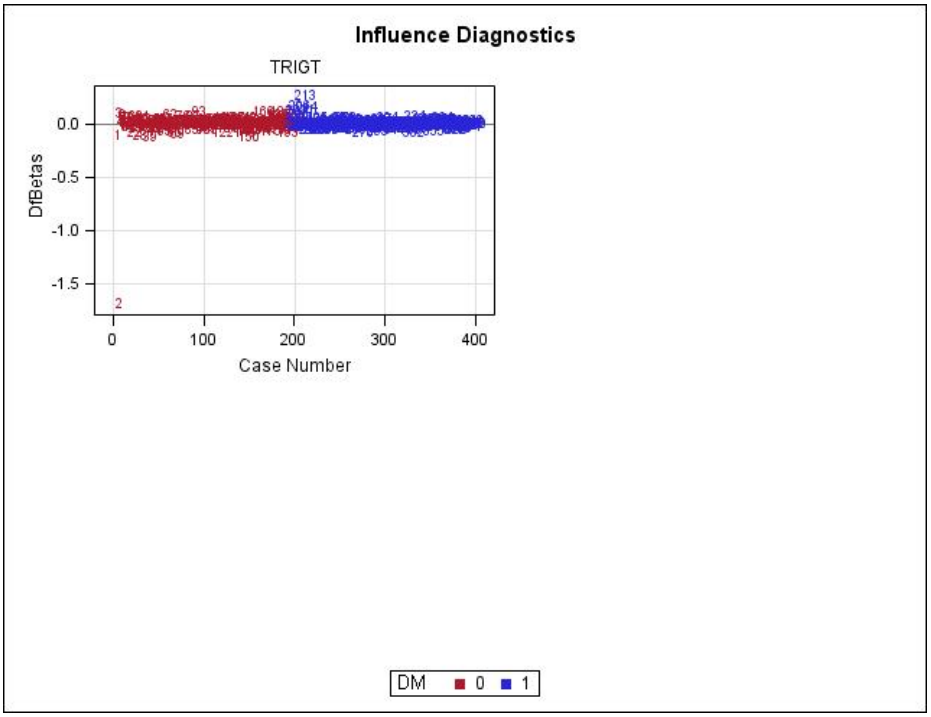

Figura 7 - Diagnóstico de observações influentes.

As Figuras 8 e 9 mostram painéis com diferentes medidas de diagnóstico versus as probabilidades preditas pelo modelo e leverages, respectivamente, identificando as mesmas observações descritas acima como potencialmente influentes. Comportamento similar foi observado na Figura 10.

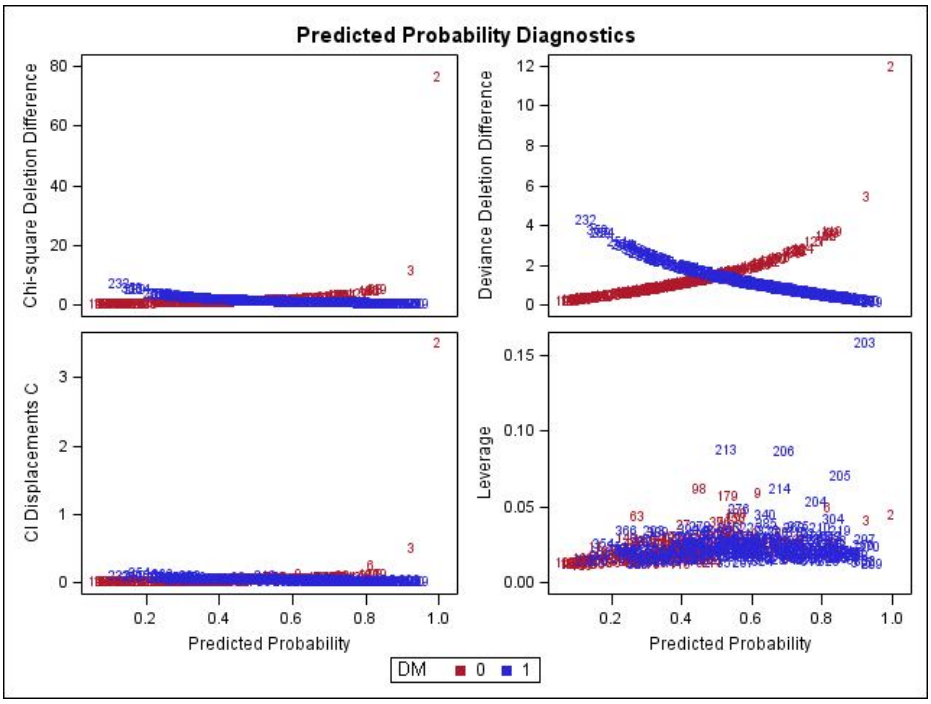

Figura 8 – Diagnóstico da probabilidade predita.

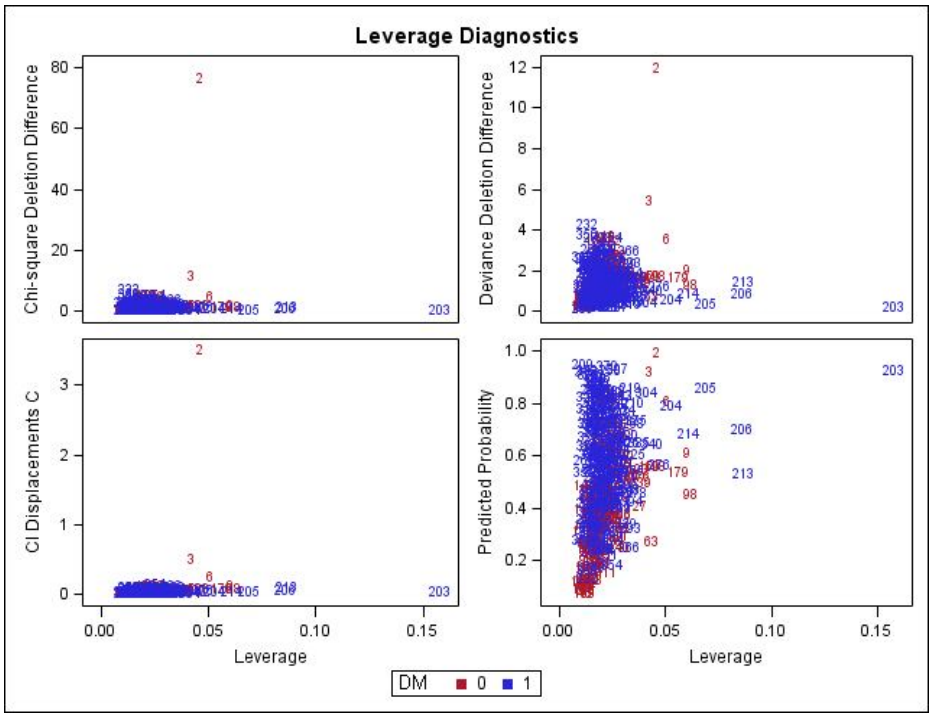

Figura 9 – Gráfico dos *leverages.*

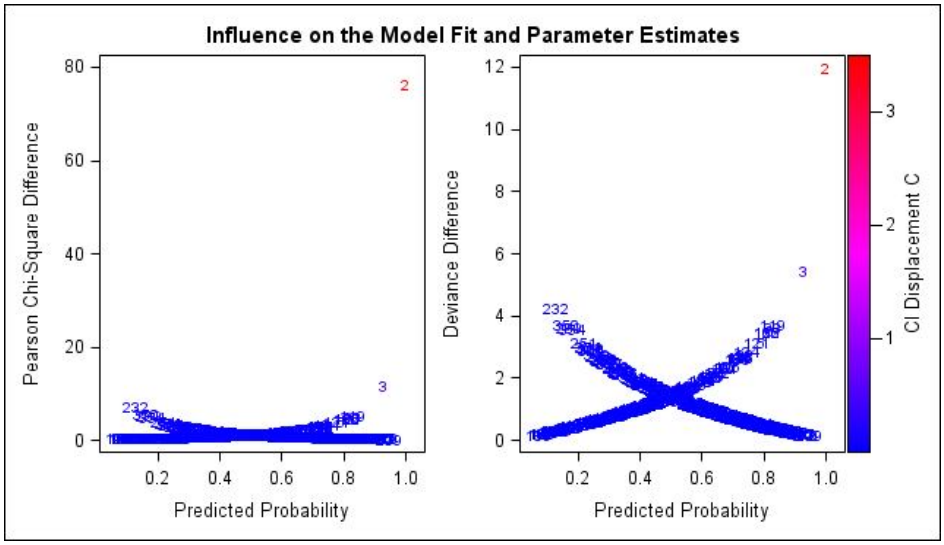

Figura 10 – Influência sobre o ajuste do modelo e estimativas dos parâmetros.

Utilizando a macro ***inflogis*** do programa SAS, foi gerado o gráfico de diagnóstico mostrado na Figura 11, no qual são identificadas as observações #1009 e #2004 como potencialmente influentes, que representam, respectivamente, as observações #2 e #203 do banco de dados ordenado. Esses procedimentos gráficos podem ser considerados complementares para identificação de observações influentes.

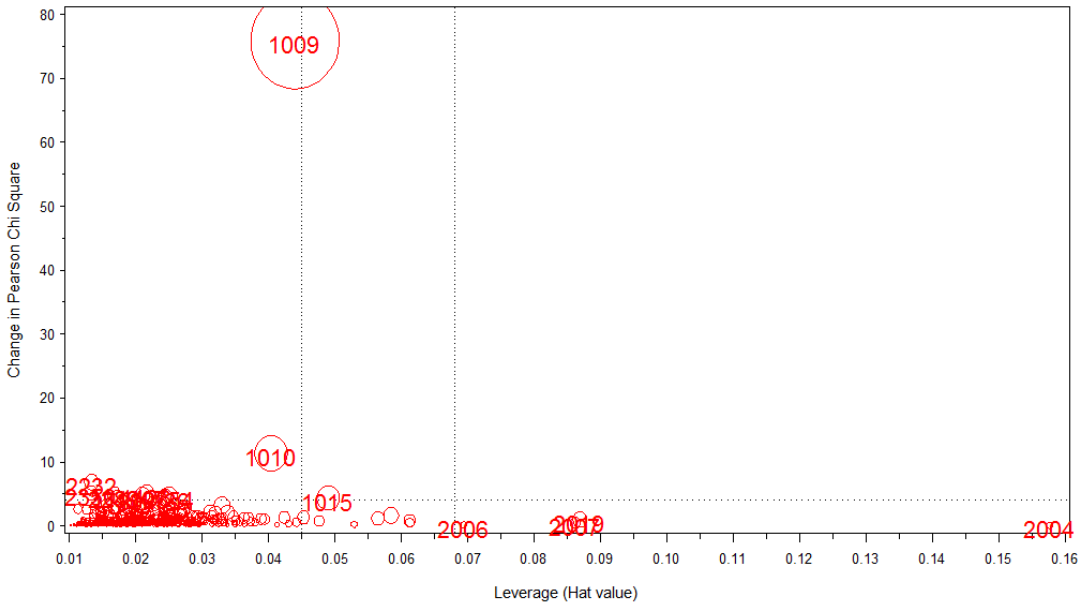

Figura 11 – Gráfico de 'bolhas'.

As observações identificadas como potencialmente influentes precisam ser examinadas quanto à plausibilidade dos valores. No exemplo, todos os valores são biologicamente plausíveis, de modo que para minimizar a influência sobre os parâmetros de regressão do modelo logístico (e, portanto, nas estimativas de razão de chances), o uso de métodos robustos é recomendado.

Para tanto, foi utilizada a macro SAS chamada ***robust*** descrita na Seção 5.1.2. A sintaxe do quadro abaixo ajusta o mesmo modelo multivariável, utilizando o M-estimador de Huber (function=HUBER).

```
%robust(data=DM1, response=DM1, model=INFLAMACAO IDADE SEXOM RACACOR
HIPERT IMC RCQ TRIGT, proc=logistic, FUNCTION=HUBER,
id=ID, iter=10, print=print);
```

Como resultado, pode-se obter as estimativas dos coeficientes e dos *odds ratios* para o modelo de regressão logística ajustado, conforme segue nos quadros abaixo.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-10.2845	1.8161	32.0692	<.0001
INFLAMACAO	1	0.6047	0.2553	5.6094	0.0179
IDADE	1	0.0113	0.0201	0.3138	0.5754
SEXOM	1	-0.3709	0.2572	2.0794	0.1493
RACACOR	1	-0.00662	0.2577	0.0007	0.9795
HIPERT	1	0.5500	0.2432	5.1164	0.0237
IMC	1	0.0231	0.0252	0.8413	0.3590
RCQ	1	0.4388	0.0922	22.6467	<.0001
TRIGT	1	0.3830	0.1916	3.9980	0.0456

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
INFLAMACAO	1.831	1.110	3.020
IDADE	1.011	0.972	1.052
SEXOM	0.690	0.417	1.142
RACACOR	0.993	0.600	1.646
HIPERT	1.733	1.076	2.792
IMC	1.023	0.974	1.075
RCQ	1.551	1.294	1.858
TRIGT	1.467	1.008	2.135

As observações abaixo foram identificadas como sendo influentes pela macro ***robust*** e receberam pesos menores do que as demais observações (coluna_weight_).

Obs	ID	DM1	_fit_	_weight_	_resid_	_hat_	flag	TRIGT	RCQ	IMC	IDADE
2	10092	0.99446	0.72369	-2.76617	0.015169	*		10.9146	19.9040	60.5589	54
3	10102	0.91216	0.93017	-2.15215	0.040159	*		1.0593	25.1508	24.8081	54
2322232	1	0.12777	0.99295	2.01606	0.013281	*		1.3110	15.4182	24.2629	48

6.3 Regressão Logística Utilizando o Programa R

De maneira análoga, é possível ajustar o modelo logístico multivariável no R, por meio da função ***glm***, conforme segue abaixo.

```
DM.glm<- glm(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ + TRIGT,
binomial,data=DM1)
```

Os valores das estimativas do modelo de regressão logística ecorrespondentes razão de chances foram praticamente idênticos aos valores obtidos pelo SAS, como pode ser observado abaixo.

Coefficients:

	Estimate	Std,Error	z value	Pr(>|z|)	
(Intercept)	-10,123527	1,804959	-5,609	2,04e-08	***
INFLAMACAO	0,624062	0,254409	2,453	0,0142	*
IDADE	0,010613	0,020056	0,529	0,5967	
SEXOM	-0,356609	0,256193	-1,392	0,1639	
RACACOR	0,009044	0,255810	0,035	0,9718	
HIPERT	0,560422	0,242397	2,312	0,0208	*
IMC	0,019336	0,025057	0,772	0,4403	
RCQ	0,442951	0,092061	4,812	1,50e-06	***
TRIGT	0,301140	0,178669	1,685	0,0919	.

	OR	2,5%	97,5%
(Intercept)	40,1244	1,0389	0,0012
INFLAMACAO	1,8665	1,1364	3,0874
IDADE	1,0107	0,9717	1,0513
SEXOM	0,7000	0,4218	1,6683
RACACOR	1,0091	0,6107	1,6683
HIPERT	1,7514	1,0901	2,8236
IMC	1,0195	0,9708	1,0712
RCQ	1,5573	1,3079	1,8779
TRIGT	1,3514	0,9877	1,9792

Para evitar redundâncias, serão apresentados apenas alguns dos gráficos fornecidos pelo R, uma vez que a maioria é semelhante aos mostrados na Seção 6.3.

As medidas de diagnóstico do modelo foram realizadas de forma semelhante ao software SAS. Portanto, apresentaremos apenas duas medidas de diagnóstico que podem ser realizadas no R. A primeira delas é obtida pela função ***influence.measures***, que permite identificar quais observações são classificadas como influentes, bem como a estatística utilizada para esta classificação. O comando *summary* lista os resultados da função ***influence.measures***, os quais são mostrados abaixo da sintaxe.

```
inflm.DM<-influence.measures(DM.glm)
which(apply(inflm.DM$is.inf, 1, any))
summary(inflm.DM)
```

	2	203	204	205	206	213	214											
			dfb.	dfb.	dfb.	dfb.	dfb.	dfb.	dfb.	dfb.								
	dfb.1		INFL	IDAD	SEXO	RACA	HIPE	IMC	RCQ	TRIG	dffit		cov.r		cook.d	hat		
2	0,08		0,09	-0,04	0,05	0,06	0,04	-0,20	0,10	-0,53	-59,00	*	0,90	*	0,39	0,04		
203	-0,01		-0,02	0,00	0,00	-0,05	-0,03	-0,01	-0,01	0,17	0,17		1,21	*	0,00	0,16	*	
204	-0,06		-0,02	-0,05	-0,07	-0,04	-0,03	-0,06	0,14	-0,03	0,15		1,07	*	0,00	0,05		
205	0,01		0,03	-0,02	0,02	-0,05	0,02	0,05	0,00	0,12	0,14		1,09	*	0,00	0,07	*	
206	0,03		0,07	-0,03	0,05	-0,10	-0,07	-0,02	-0,03	0,21	0,25		1,10	*	0,01	0,09	*	
213	0,02		-0,07	0,02	-0,07	0,03	-0,06	0,06	-0,09	0,29	0,33		1,09	*	0,01	0,09	*	
214	0,00		-0,04	-0,05	0,01	0,01	-0,03	0,00	0,01	0,18	0,21		1,07	*	0,00	0,06		

A segunda medida de diagnóstico é o gráfico de bolhas que apresenta os valores dos resíduos *studentizados versus* a matriz chapéu, em que o tamanho das 'bolhas' é proporcional a distância de Cook associada a cada observação como mostra a Figura 12. A sintaxe referente a Figura 12 é apresentada no anexo 8.2.

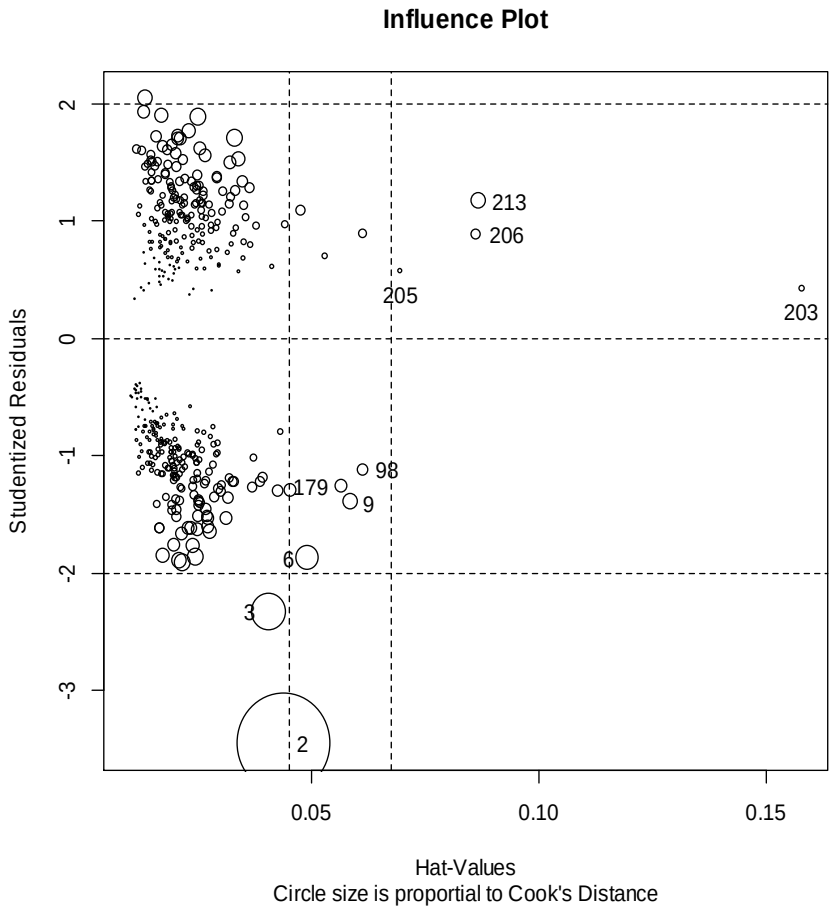

Figura 12–Gráfico de ‘bolhas’

Como complementação desse gráfico, é possível obter os valores dos resíduos studentizados, os valores da diagonal da matriz chapéu e distância de Cook das observações que foram identificadas como possíveis medidas de influência, conforme segue no quadro abaixo.

	StudRes	Hat	CookD
2	-3,455	0,044	0,623
2	-2,326	0,040	0,230
6	-1,869	0,049	0,158
9	-1,399	0,058	0,106
98	-1,113	0,061	0,079
179	-1,258	0,056	0,079
203	0,422	0,158	0,046
204	0,706	0,053	0,042
205	0,572	0,069	0,039
206	0,878	0,086	0,071
213	1,167	0,087	0,102
276	1,088	0,048	0,067

Depois de identificadas as observações influentes faz-se necessário minimizar o impacto das mesmas nas estimativas dos coeficientes de regressão e, consequentemente, nas estimativas de razão de chances. Isto pode ser realizado com os métodos robustos disponíveis na função ***glmrob***, descrita na Seção 5.2.2.

A sintaxe mostrada no quadro abaixo ajusta o modelo de regressão logístico robusto utilizando o estimador de Huber, o qual é especificado pelo valor 1,5 na constante de afinação (*control = glmrobMqle.control(tcc=1.5)),*conforme sugerido por Heritier (2010).

```
DM.glmrob<- glmrob(DM~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ +
                   TRIGT,binomial,data=DM1,method="Mqle",  control = glmrobMqle.control(tcc=1.5))
```

As estimativas dos parâmetros obtidas com a sintaxe acima são mostradas no quadro abaixo, tendo sido observadas 38 observações influentes para as quais o método atribui pesos menores do que 1, para minimizar o impacto sobre os coeficientes de regressão.

Coefficients:

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-11.01541	1.90064	-5.796	6.81e-09 ***
INFLAMACAO	0.49561	0.25924	1.912	0.05590 .
IDADE	0.01469	0.02052	0.716	0.47413
SEXOM	-0.39952	0.26189	-1.526	0.12712
RACACOR	-0.06571	0.26459	-0.248	0.80386
HIPERT	0.49247	0.24643	1.998	0.04567 *
IMC	0.03073	0.02583	1.190	0.23404
RCQ	0.44679	0.09515	4.696	2.66e-06 ***
TRIGT	0.57444	0.22022	2.608	0.00909 **

De maneira similar pode ser ajustado o modelo de regressão logístico robusto utilizando o estimador de Mallows, o qual é obtido especificando a constante de afinação igual a 1,5, e os elementos da diagonal da matriz chapéu para o cálculo dos pesos (*weights.on.x='hat'*). A sintaxe completa e as estimativas de parâmetros são mostradas abaixo.

```
DM.glmrob2 <- glmrob(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ +
TRIGT,family=binomial,control=glmrobMqle.control(tcc=1.5),weights.on.x='hat',data=DM1)
```

Coefficients:

	Estimate	Std. Error	z-value	Pr(>|z|)	
(Intercept)	-11,1773	1,9132	-5,8420	0,0000	***
INFLAMACAO	0,4866	0,2598	1,8730	0,0611	
IDADE	0,0146	0,0206	0,7080	0,4787	
SEXOM	-0,4081	0,2627	-1,5530	0,1203	
RACACOR	-0,0695	0,2654	-0,2620	0,7933	
HIPERT	0,4888	0,2470	1,9790	0,0478	*
IMC	0,0312	0,0259	1,2050	0,2283	
RCQ	0,4551	0,0959	4,7440	0,0000	***
TRIGT	0,5811	0,2221	2,6160	0,0089	**

Os dois métodos robustos utilizados para o ajuste do modelo, Huber e Mallows, identificaram 38 observações influentes mostradas na Tabela 2. As 8 observações que produzem maior impacto nas estimativas dos coeficientes de regressão estão salientadas em azul, e receberam pesos menores que 0,7.

Tabela 2: Observações influentes identificadas pela função ***glmrob*** e respectivos pesos utilizando os métodos de Huber e Mallows

Observação	Huber	Mallows
1	0,8470	0,8342
2	0,0391	0,0373
3	0,4550	0,4416
5	0,8262	0,8182
6	0,7848	0,7707
22	0,8019	0,7986
33	0,8650	0,8570
70	0,9646	0,9615
81	0,9657	0,9682
93	0,7329	0,7302
119	0,6812	0,6786
121	0,7771	0,7697
122	0,6593	0,6547
124	0,9177	0,9204
146	0,8012	0,7921
149	0,8964	0,8969
166	0,7781	0,7746
175	0,8067	0,7975
229	0,9718	0,9695
232	0,5550	0,5487
237	0,8748	0,8723
251	0,7193	0,7093
256	0,9092	0,9050
261	0,7797	0,7737
283	0,8959	0,8849
289	0,6321	0,6271
293	0,8896	0,8773
296	0,8533	0,8510
232	0,8948	0,8923
337	0,9380	0,9360
339	1,0000	0,9992
350	0,6365	0,6320
354	0,6609	0,6507
364	0,8868	0,8783
366	0,8693	0,8593
369	0,9651	0,9646
374	0,8225	0,8135
396	0,8187	0,8162
399	0,8985	0,8904

A função ***glmRob*** mencionada na Seção 5.2.2, também pode ser usada para o ajuste do modelo logístico robusto. Para o estimador tipo Mallows, o modelo pode ser ajustado utilizando a sintaxe abaixo. No entanto, como o modelo especificado possui preditores categóricos (sexo, hipertensão, cor de pele e inflamação) não é possível estimar os parâmetros do modelo, devido a obtenção de uma matriz singular no processo de estimação pelo método da matriz de covariâncias de determinante mínimo pela função ***mcd.***

```
DM.glmRob3<-glmRob(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT +
IMC + RCQ + TRIGT, family=binomial, data=DM,weights=NULL,method="mallows",
model = TRUE, control = glmRob.control)
```

6.4 Comparação dos resultados

A Tabela 3 mostra um resumo dos resultados obtidos utilizando um modelo de regressão logístico não robusto, e também, para os métodos robustos disponibilizados na macro ***robust*** do SAS, e da função ***glmrob*** do R.

No exemplo empírico explorado nesta seção foi possível perceber que a utilização de métodos robustos alterou tanto a estimativa da magnitude de associação quanto dos intervalos de confiança do preditor presença de inflamação sistêmica. Por exemplo, no modelo que não utiliza métodos robustos a razão de chances estimada foi RC=1,87 (IC 95%: 1,13-3,07) ajustando pelas demais variáveis. Comparado ao método de Mallows, houve uma redução de aproximadamente 12,8% na estimativa da magnitude de associação mudando também a significância. Para outros preditores para os quais existem valores extremos, como por exemplo, o nível de triglicerídeos total, o impacto das observações influentes foi ainda maior, em que a razão de chances passa de 1,35 (IC 95%: 0,95-1,92) no modelo logístico não robusto para 1,79 (IC 95%:1,16-2,76) no modelo robusto com estimador tipo Mallows, mudando também a significância estatística.

Isto mostra a importância do diagnóstico da regressão logística para identificação de observações influentes bem como o refinamento do modelo utilizando métodos robustos de estimação.

Os resultados empíricos sugerem um comportamento similar entre os métodos não robusto e robusto do SAS. Porém, foram observadas diferenças entre os resultados obtidos pela macro ***robust*** do SAS e a função ***glmrob*** do R, que são explicadas pelas diferenças dos métodos implementados. Para os estimadores Huber e Mallows disponíveis na função ***glmrob*** não foram observadas diferenças relevantes nos resultados.

Tabela 3: Estimativas de parâmetros, erros padrões, razão de chance e intervalos com 95% de confiança , para o modelo de regressão logística com e sem utilização de métodos robustos.

	Não robusto				**Robusto SAS**				**glmrob - Huber**				**glmrob - Mallows**			
	$\hat{\beta}$	**EP**	**RC**	**IC 95%**	$\hat{\beta}$	**EP**	**RC**	**IC 95%**	$\hat{\beta}$	**EP**	**RC**	**IC 95%**	$\hat{\beta}$	**EP**	**RC**	**IC 95%**
Intercepto	-10,12	1,81			-10,28	1,82			-11,02	1,90			-11,18	1,91		
INFLAMACAO	0,62	0,25	1,87	(1,13-3,07)	0,60	0,26	1,83	(1,11-3,02)	0,50	0,26	1,64	(0,99-2,73)	0,49	0,26	1,63	(0,98-2,71)
IDADE	0,01	0,02	1,01	(0,97-1,05)	0,01	0,02	1,01	(0,97-1,05)	0,01	0,02	1,01	(0,97-1,06)	0,01	0,02	1,01	(0,97-1,06)
SEXOM	-0,36	0,26	0,70	(0,42-1,16)	-0,37	0,26	0,69	(0,42-1,14)	-0,40	0,26	0,67	(0,40-1,12)	-0,41	0,26	0,66	(0,40-1,11)
RACACOR	0,01	0,26	1,01	(0,61-1,67)	-0,01	0,26	0,99	(0,60-1,65)	-0,07	0,26	0,94	(0,56-1,57)	-0,07	0,27	0,93	(0,55-1,57)
HIPERT	0,56	0,24	1,75	(1,09-2,82)	0,55	0,24	1,73	(1,08-2,79)	0,49	0,25	1,64	(1,01-2,65)	0,49	0,25	1,63	(1,00-2,65)
IMC	0,02	0,03	1,02	(0,97-1,07)	0,02	0,03	1,02	(0,97-1,08)	0,03	0,03	1,03	(0,98-1,08)	0,03	0,03	1,03	(0,98-1,09)
RCQ	0,44	0,09	1,56	(1,30-1,87)	0,44	0,09	1,55	(1,29-1,86)	0,45	0,10	1,56	(1,30-1,88)	0,46	0,10	1,58	(1,31-1,90)
TRIGT	0,30	0,18	1,35	(0,95-1,92)	0,38	0,19	1,47	(1,01-2,14)	0,57	0,22	1,78	(1,15-2,73)	0,58	0,22	1,79	(1,16-2,76)

7 Considerações finais

O diagnóstico do modelo é uma etapa crucial no ajuste de modelos de regressão logística para identificação de possíveis problemas. Essa etapa pode ser realizada utilizando vários procedimentos gráficos.

Os métodos robustos implementados na macro ***robust*** do SAS e na função ***glmrob*** do R são ferramentas úteis para minimizar o impacto de observações atípicas ou influentes nos coeficientes de regressão, atribuindo pesos menores. A função ***glmRob*** do R tem limitações quando é necessário o ajuste para preditores categóricos.

Todos os métodos considerados procuram ponderar observações que podem potencialmente sub ou superestimar os parâmetros do modelo. Assim, estes métodos sugerem estimativas de coeficientes e razão de chances mais robustas e resistentes a observações atípicas.

Embora ambos os métodos apresentados tenham identificado diferentes observações como influentes, e ponderado as mesmas com diferentes pesos, ambos são úteis no ajuste do modelo logístico na presença de valores influentes.

As diferenças observadas entre os resultados obtidos pela macro ***robust*** do SAS e a função ***glmrob*** do R, merecem uma investigação mais aprofundada utilizando estudos de simulação para avaliação da eficiência.

Esse estudo descreveu métodos robustos para regressão logística, permitindo que mesmo usuários sem domínio de aspectos computacionais sobre o tema sejam capazes de utilizá-los. Essas ferramentas podem ser muito importantes em situações nas quais existem preditores quantitativos, bastante comuns no contexto de pesquisa clínica e epidemiológica, para minimizar potenciais vieses nas estimativas de associações.

8 Anexos

8.1 Anexo 1 - Sintaxe SAS

```
optionsps=58ls=120nocenternodatenonumberformchar='|----|+|---+=|-/\<>*';
libname L1 'C:\ '; * local onde está disponível o banco de dados
%include'C:\';     * local onde está salvo a macro robust, em formato .sas
%include'C:\';     * local onde está salvo a função inflogis, em formato .sas

libname L1 'I:\2012-2\Monografia\diabetes_versao_final';
%include 'I:\2012-
2\Monografia\diabetes_versao_final\robust_versao_final.sas';
%include 'I:\2012-
2\Monografia\diabetes_versao_final\inflogis_versao_final.sas';
data DM;
     set L1.casocontrole_dm_inflamacao;
run;
proc format;
     valuedmf0='Nao'1='Sim';
     valuesexof0='Masculino'1='Feminino';
     valuecorf1='Branco'0='Nao branco';
     valuehipertf0='Nao'1='Sim';
run;

proc means data=DM max dec=2minmaxmeanstd;
     var INFLAMACAO SEXOM RACACOR HIPERT IDADE IMC RCQ TRIGT;
run;

optionsls=120;
odsgraphics on;

* RL sem robustez;
Proc logistic cdata=DM descendingplots(only
label)=(phatleveragedpcDfBetasinfluence);
model DM = INFLAMACAO IDADE SEXOM RACACOR HIPERT IMC RCQ TRIGT / rl;
run;

%inflogis(data=DM, y=DM, X=INFLAMACAO IDADE SEXOM RACACOR HIPERT IMC RCQ TRIGT,
id=ID, gy=difchisq,
gx=pred hat, bubble=CBAR, lcolor=red, bsize=14);

* Cria variavel DM1 com valores 1=DM presente e 2=DM ausente,
pois a macroo 'robust' modelo o menor valor do desfecho;
data DM1;
     set DM;

     DM1 = 2 - DM;
run;
%robust(data=DM1, response=DM1, model=INFLAMACAO IDADE SEXOM RACACOR HIPERT IMC RCQ
TRIGT,
          proc=logistic, FUNCTION=HUBER, id=ID, iter=10, print=print);

proc print data=resids;
     var ID DM1 _fit_ _weight_ _resid_ _hat_ flag trigtrcqimcidade;
     where _weight_ ne 1;
run;
```

8.2 Anexo 2 - Sintaxe R

```
#Leitura do banco de dados#
setwd('C:\\')
DM1=read.csv('C:\\banco",sep=';',header=T,dec=',')
attach(DM1)

#Análise descritiva:

summary(DM1)
library(psych)   # para o 'describe'
describe(DM1)    # descreve todas as variáveis do banco de dados
t1<- table(INFLAMACAO)/sum(table(INFLAMACAO))
t2<- table(HIPERT)/sum(table(HIPERT))
t3<- table(SEXOM)/sum(table(SEXOM))
t4<- table(RACACOR)/sum(table(RACACOR))

library(plyr)        #Boxplot identificandoos outliers
library(TeachingDemos)
source("http://www.r-statistics.com/wp-content/uploads/2011/01/boxplot-with-outlier-label-r.txt")
par(mfrow=c(2,2))
boxplot.with.outlier.label(IDADE, seq_along(IDADE),main='IDADE')
boxplot.with.outlier.label(RCQ, seq_along(RCQ),main='RCQ')
boxplot.with.outlier.label(TRIGT, seq_along(TRIGT),main='TRIGT')
boxplot.with.outlier.label(IMC, seq_along(IMC),main='IMC')

# Ajustando o modelo de REGRESSÃO LOGÍSTICA:

DM.glm<- glm(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ + TRIGT,
binomial,data=DM1)
summary(DM.glm)       # resumo do modelo ajustado#
confint.default(DM.glm) # intervalos de confiança para o modelo ajustado#
exp(cbind(OR = coef(DM.glm), confint(DM.glm))) # odds ratios

# Medidas de diagnóstico:

tipo.resid1 <- c("deviance", "pearson", "working", "response")
sapply(tipo.resid1, residuals, object = DM.glm)
inflm.DM<-influence.measures(DM.glm) #medidas de influencia
which(apply(inflm.DM$is.inf, 1, any

summary(inflm.DM)  # resumo das medidas de influência
plot(rstudent(DM.glm) ~ hatvalues(DM.glm))  # gráfico de diagnóstico

cutoff<- 4/((nrow(DM)-length(DM.glm$coefficients)-2))  # gráfico da distãncia de Cook do modelo ajustado
plot(DM.glm, which=4, cook.levels=cutoff)

require(car)   #Gráfico de bolhas da distância de cook:
influencePlot(DM.glm, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook's
Distance" )

windows()
par(mfrow=c(2,2))
plot(DM.glm)

plot(fitted(DM.glm), resid(DM.glm),xlab="Fitted values",ylab="Residuals",main="Residuals vs Fitted")

# Ajustando o modelo de REGRESSÃO LOGÍSTICA ROBUSTA:

library(robustbase)
```

```
#Huber
DM.glmrob<- glmrob(DM~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ +
TRIGT, binomial,data=DM1,method="Mqle",control = glmrobMqle.control(tcc=1.5))
sumary.glmrob<- summary(DM.glmrob)
sumary.glmrob$w.r

#Mallows
DM.glmrob2 <- glmrob(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ +
TRIGT,family=binomial,control=glmrobMqle.control(tcc=1.5),weights.on.x='hat',data=DM1)
sumary.glmrob2<- summary(DM.glmrob2)
sumary.glmrob2$w.r

#Mallows através da glmRob # identificamos potencias problemas na função
no uso de preditores binários.
require(robust)
DM.glmRob3<-glmRob(DM ~ INFLAMACAO + IDADE + SEXOM + RACACOR + HIPERT + IMC + RCQ
+ TRIGT, family=binomial, data=DM, weights=NULL,method ="mallows", model = TRUE, control =
glmRob.control)
```

8.3 Anexo 3 – Macro *robust*

```
%macrorobust(
            data=_LAST_,
            response=,        /* response variable                          */
            model=,           /* RHS of model statement                     */
            proc=REG,         /* estimation procedure: GLM, REG, LOGISTIC   */
            class=,           /* class variables (GLM only)                 */
            id=,              /* ID variables                               */
            out=resids,       /* output observations data set               */
            outparm=,         /* output parameters data set                 */
            function=bisquare, /* weight function: BISQUARE, HUBER or LAV   */
            tune=,            /* tuning constant for bisquare/huber         */
            iter=,            /* max number of iterations                   */
            converge=0.05,    /* max change in weight for convergence.      */
/* NB: must have leading 0                    */
            print=no
            );

%let abort=0;
%letproc = %upcase(&proc);
%letdoparm = %index(REG LOGISTIC,&proc) ;%* Getting parameter estimates?;

%if%index(REG LOGISTIC,&proc)
      %then%letoutparm = outest;
      %else%letoutparm = outstat;

%let r=r;
%if&proc = GLM %then%let r=rstudent;
%if&proc = LOGISTIC %then%let r=resdev;

%if%length(&iter)=0%then%do;
      %letiter=10;
      %if&proc = LOGISTIC %then%letiter=4;
      %end;

%let function = %upcase(&function);
%if&tune = %str() %then%do;
      %if&function = BISQUARE %then%let tune = 6;
      %else%let tune = 2;
%end;
%let print = %upcase(&print);
data resids;
      set&data;
      _weight_ = 1;
      lastwt = .;

%do it = 1%to&iter;

      %letpr=noprint;
      %if&print = PRINT %then%let pr=;
      %else%if%index(&print,NOPRINT) %then%letpr=NOPRINT;
      %else%if%index(&print,&it)  %then%letpr=;

      %*-- Remove parmest data set from a prior run;
      %if&it=1%then%do;
      proc datasets nolistnowarn;
            deleteparmest;
      %end;

      %*-- Fit the model, using current weights;
      proc&proc data=resids%if&it >1%then (drop=_resid_ _fit_ _hat_);
            &outparm=parms
            &pr;
            weight _weight_;              %*-- observation weights;
            %if%length(&class)>0& (&proc=GLM or (&proc=LOGISTIC and &sysver>=8))
            %then%do;
                  class&class;
            %end;
            model&response = &model;
            output out=newres&r=_resid_ p=_fit_ h=_hat_;
            title3 "Iteration &it";
      run;
      %if&syserr>4%then%let abort=1; %if&abort %then%goto DONE;

      optionsnonotes;
```

```
      %*-- Find the median absolute residual;
      data resids;
            setnewres;
            absres = abs(_resid_);

      %*-- Find median absolute deviation (MAD);
      proc univariate data=resids noprint;
            varabsres;
            output out=sumry median=mad;

      %*-- Calculate new weights;
      data&out;
            setresids end=eof;
            drop w mad _maxdif_ absreslastwt;
            retain _maxdif_ 0;
            lastwt = _weight_;
            if _n_=1 then set sumry(keep=mad);
            label _weight_ ="&function weight";

            if _resid_ ^= .then do;
                  %*-- scaled residual;
                  w = _resid_ / (&tune * mad);
                  %if&function = BISQUARE    %then %bisquare(w);
                  %else%if&function = HUBER %then %huber(w);
                  %else%if&function = LAV   %then %lav(w);
                  %else _weight_=1; /* OLS */
                  _maxdif_ = max(_maxdif_, abs(_weight_-lastwt));
            end;

            ifeof then do;
            *     file print;
                  put"NOTE: iteration &it  " _maxdif_=;
                  callsymput('maxdif',left(put(_maxdif_,6.4)));
                  end;
      run;
      %*if &doparm %then %do;
      dataparms;
            iter = &it;
            setparms;
            _maxdif_ = input("&maxdif", best.);
      proc append base=parmest new=parms;
      run;
      %*end;

      %if&maxdif<&converge %then%gotofini;
%end;
%fini:;
      dataparmest;
            setparmest;
      %if&doparm%then%do;
            drop _type_
            %if&proc=REG %then _model_ _depvar_  &response;
            ;
            title3 'Iteration history and parameter estimates';
      %end;
      %else%do;
            drop _name_ prob;
            if _type_='SS1' then delete;
            title3 'Iteration history and test statistics';
      %end;
      proc print data=parmest;
            iditer;
            run;

      %if%length(&outparm)>0%then%do;
      data&outparm;
            setparmest end=eof;
            dropiter _maxdif_;
            ifeof then output;
      %end;

%if%index(&print,NO)=0%then%do;
proc print data=&out;
      %if&id ^= %str() | &class ^= %str() %then%do;
            id&class &id;
      %end;
            var&response _fit_ _weight_ _resid_ _hat_ flag;
```

```
            title3 'Residuals, fitted values and weights';
            run;
%end;
title3;
%done:
options notes;
%mend;

%macro bisquare(w);
      if abs(&w) <1
            then do; _weight_ = (1 - &w**2) **2;  flag=' '; end;
            else do; _weight_ = 0;                flag='*'; end;
%mend;

%macro huber(w);
      if abs(&w) <1
            then do; _weight_ = 1;               flag=' '; end;
            else do; _weight_ = 1/abs(&w);       flag='*'; end;
%mend;

%macro lav(w);
      _weight_ = 1/(absres +(absres=0));
%mend;
```

8.4 Macro*inflogis*

```
%macroinflogis(
data=_last_,    /* Name of input data set                  */
   y=,             /* Name of criterion variable             */
trials=,        /* Name of trials variable                 */
   x=,             /* Names of predictors                    */
      class=,         /* Names of class variables (V8+)         */
id=,            /* Name of observation ID variable (char)  */
out=_diag_,     /* Name of the output data set             */
gy=DIFDEV,      /* Ordinate for plot: DIFDEV or DIFCHISQ   */
gx=PRED,        /* Abscissa for plot: PRED or HAT          */
bubble=C,       /* Bubble proportional to: C or CBAR       */
label=INFL,     /* Points to label: ALL, NONE, or INFL     */
infl=%str(difchisq>&dev or &bubble >1 or hat>hcrit1),
dev=4,          /* DIFDEV/DIFCHISQ criterion for inflpts  */
lsize=1.5,      /* obs label size.  The height of other    */
/* text is controlled by the HTEXT= goption*/
lcolor=BLACK,   /* obs label color                         */
lpos=5,         /* obs label position                      */
lfont=,         /* obs label font                          */
bsize=10,       /* bubble size scale factor                */
bscale=AREA,    /* bubble size proportional to AREA or RADIUS */
bcolor=RED,     /* bubble color                            */
bfill=,         /* fill bubbles?  SOLID|GRADIENT              */
refcol=BLACK,   /* color of reference lines                */
reflin=33,      /* line style for reference lines; 0->NONE */
loptions=noprint,/* options for PROC LOGISTIC              */
name=INFLOGIS,
gout=
   );

%let me=INFLOGIS;
%letnv = %numwords(&x);        /* number of predictors */
%letnx = %numwords(&gx);       /* number of abscissa vars */
%letny = %numwords(&gy);       /* number of ordinate vars */
%if&nv = 0%then%do;
%put ERROR: List of predictors (X=) is empty;
%goto done;
%end;

%letgx=%upcase(&gx);
%letgy=%upcase(&gy);
%let label=%upcase(&label);
%let bubble=%upcase(&bubble);
%if not ((%bquote(&bubble) = C)
or   (%bquote(&bubble) = CBAR)) %then%do;
%put BUBBLE=%bquote(&bubble) is not valid. BUBBLE=C will be used;
%let bubble=C;
%end;

%if%length(&class) >0 and &sysver<8%then%do;
      %let class=;
      %put INFLOGIS:  The CLASS= parameter is not supported in SAS &sysver;
      %end;

proc logistic nosimple data=&data &loptions ;
%if%length(&class)>0%then%do;
            class&class;
            %end;
%if%length(&trials)=0%then%do;
model&y         = &x / influence;
%end;
%else%do;
model&y/&trials = &x / influence;
%end;
output out=&out h=hat pred=pred
difdev=difdev
difchisq=difchisq
                    c=c  cbar=cbar
resdev=resdev;
data&out;
set&out;
label difdev='Change in Deviance'
dif chisq='Change in Pearson Chi Square'
hat = 'Leverage (Hat value)'
```

```
studres = 'Studentized deviance residual';
studres = resdev / sqrt(1-hat);
run;

%if%length(&bfill) %then%do;
proc sort data=&out;
      by  descending&bubble;
      run;
%end;
%doi=1%to&ny;
%letgyi = %scan(&gy, &i);
%do j=1%to&nx;
%letgxj = %scan(&gx, &j);
%put&me: Plotting &gyivs&gxj ;

%if&label ^= NONE %then%do;
data _label_;
set&out nobs=n;
lengthxsys $1ysys $1 function $8 position $1 text $16 color $8;
retainxsys'2'ysys'2' function 'LABEL' color "&lcolor" when 'A';
retainhcrit hcrit1;
drophcrit;
*keep &id x y xsysysysfunction position text color size position
hatdifchisqdifdev&bubble;
      x = &gxj;
      y = &gyi;
%if&id ^= %str() %then%do;
text = left( &id );
%end;
%else%do;
text = put(_n_,3.0);
%end;
if _n_=1 then do;
hcrit = 2 * (&nv+1)/n;
        hcrit1 = 3 * (&nv+1)/n;
put "&me: Hatvalue criteria: 2p/n="hcrit4.3', 3p/n=' hcrit1 4.3;
callsymput('hcrit',put(hcrit,4.3));
callsymput('hcrit1',put(hcrit1,4.3));
end;
size=&lsize;
position="&lpos";
      %if%length(&lfont) %then%do;
            style="&lfont";
            %end;
%if&label = INFL %then%do;
/*         if %scan(&gy,1) >&dev
           ordifchisq>&dev
or hat >hcrit
or&bubble > 1
then output;   */
if&infl then output;
%end;
run;
%if&i=1 and &j=1%then%do;
proc print data=_label_;
var&y &x predstudres hat difchisqdifdev&bubble;
            format hat 3.2pred&bubble 4.3studres6.3difdevdifchisq6.3;
%if&id ^= %str() %then%do;
id&id;
%end;
%else%do;
id text;
%put WARNING:  Observations are identified by sequential number (TEXT) because no ID= variable
was specified.;
%end;
%end;
%end;  /* &label ^= NONE */

proc gplot data=&out &GOUT ;
bubble&gyi * &gxj = &bubble /
%if&label ^= NONE %then%do;
annotate=_label_
%end;
frame
vaxis=axis1 vminor=1hminor=1
%if&reflin ^= 0%then%do;
%if (&gyi = DIFDEV) or (&gyi = DIFCHISQ) %then%do;
```

```
vref=&devlvref=&reflincvref=&refcol
%end;
%if (&gxj = HAT) %then%do;
href= &hcrit&hcrit1 lhref=&reflinchref=&refcol
%end;
%end;
bsize=&bsizebcolor=&bcolorbscale=&bscale
            %if%length(&bfill) %then%do;
                  bfill=&bfill
                  %end;
name="&name"
           Des="Logistic influence plot for &y";
axis1 label=(a=90 r=0);
run; quit;
   %gskip;
%end;   /* gx loop */
%end;   /* gy loop */
%done:
quit;
%mend;

%macro numwords(lst);
%leti = 1;
%let v = %scan(&lst,&i);
%do%while (%length(&v) >0);
%leti = %eval(&i + 1);
%let v = %scan(&lst,&i);
%end;
%eval(&i - 1)
%mend;
```

Referências Bibliográficas

HOSMER, D.W.; LEMESHOW S. **Applied logistic regression**. New York: Wiley, 2000.

HERITIER, S. et al.**Robust Methods in Biostatistics**. John Wiley & Sons, 2009.

HUBER, P.J. **Robust Statistical Procedures**. 2ª edição. Germany:Siam, 1996.

FARCOMENI, A.; VENTURA, L. **An overview of robust methods in medical research. Statistical methods in medical research**,21(2):111–33, 2012.

IBM SPSS. **IBM SPSS Statistics. Essentials for R**.Installation Instructions for Windows, 2010.

NARGIS, S.**Robust methods in logistic regression**.University of Canberra. Division of Business L and IS,2005.

VICTORIA-FESER, M.P. **Robust inference with binary data**. Psychometrika.67(1):21–32, 2002.

DUNCAN, B.B. et al. **Low-grade systemic inflammation and the development of type 2 diabetes the atherosclerosis risk in communities study.**Diabetes.52(7):1799–805,2003.

HERITIER S. **Robust Methods in Biostatistics**. The George Institute for Global Health The University of Sydney.ASC Fremantle. 2010.

SAS. **The Power to Know**. Disponível em <www.sas.com>. Acesso em: 16 nov.2012.

R. **The R Project for StatisticalComputing**.Disponível em <www.r-project.org>. Acesso em: 30 nov.2012.

Fox J. e WEISBERGS.**An {R} Companion to Applied Regression**, 2ªEdição. Thousand Oaks CA: Sage. 2011

DATAVIS.Ca. **SAS Graphic Programs and Macros**.Disponível em <www.datavis.ca/sasmac>. Acesso em: 19out. 2012.