UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATÍSTICA

Técnicas de Diagnóstico Aplicadas ao Modelo De Regressão Logística

MARILYN AGRANONIK

Orientadora: Luciana Neves Nunes

Monografia apresentada para a obtenção

do título de Bacharel em Estatística

Porto Alegre, Dezembro de 2005.

RESUMO

A análise de regressão logística está cada vez mais presente nos artigos científicos, principalmente nos relacionados à área médica. Por se tratar de uma análise que tem por objetivo principal encontrar fatores de risco para um desfecho de interesse, faz-se necessário que o modelo esteja bem ajustado. Um modelo pode ser avaliado através de medidas de qualidade de ajuste e técnicas que identifiquem a influência das observações. Para verificar se há alguma observação extrema que cause modificações substancias nas estimativas do modelo é importante realizar a análise de diagnóstico. As técnicas de diagnóstico para modelos de regressão logística foram implementadas por Pregibon (1981) e são bastante semelhantes às técnicas de diagnóstico da regressão linear. No presente trabalho serão apresentadas algumas das técnicas mais importantes para o diagnóstico do modelo de regressão logística. Algumas dessas técnicas usam resíduos calculados a partir do modelo ajustado e identificam através de gráficos os pontos atípicos. Outras medidas de diagnóstico utilizadas nesse trabalho foram: a distância de Cook, DFBETAS, e DFFITTS que trabalham com a exclusão da observação em estudo para avaliar seu impacto nas estimativas da regressão. Para ilustrar o uso dessas técnicas foi utilizado um banco de dados reais, sendo ajustado um modelo de regressão logística para avaliar fatores de risco para sobrepeso ou obesidade.

Agradecimentos

Á professora Luciana Neves Nunes pela orientação e incentivo neste final de curso.

Aos professores do Departamento de Estatística da UFRGS que contribuíram, com seus conhecimentos, para minha formação.

A meus pais pelo amor, apoio e compreensão que sempre me deram.

Aos meus amigos Adriana, Juscelino, Lauren, Michele e Tatiana pela amizade, companheirismo e incentivo ao longo do curso.

Aos professores Marco Antonio Barbieri e Heloisa Bettiol da Faculdade de Medicina de Ribeirão Preto e ao professor Marcelo Goldani (FAMED - UFRGS) por terem cedido os dados utilizados na parte da aplicação deste trabalho.

Agradeço a todos aqueles que de algum modo (direta ou indiretamente) tenham contribuído para a realização deste trabalho.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

Em diferentes áreas de pesquisa, incluindo a área da saúde, é freqüente a situação em que se deseja estudar o comportamento de uma variável resposta em relação a variáveis independentes. As variáveis independentes, também chamadas explicativas, são responsáveis pela variabilidade da variável resposta, ou dependente. Para esses casos, técnicas de modelagem são utilizadas, nas quais se incluem os modelos de regressão.

A variável de interesse, em diversos estudos, é a presença ou ausência de determinada condição. Nesses casos a variável resposta é do tipo binário ou dicotômico podendo ser representada pelos valores 1 (presença) ou 0 (ausência). Quando se deseja estudar esse tipo de situação, o mais indicado é a utilização do modelo de regressão logística linear, mais comumente conhecido por modelo de regressão logística.

Collet (1991) destaca como as principais vantagens de utilizar o modelo de regressão logística:

(1) a conveniência de seu uso do ponto de vista computacional;

(2) a sua interpretação direta em termos do logaritmo da chance de sucesso, sendo útil na análise de dados de estudos epidemiológicos;

(3) o fato de modelos com transformação logística, por suas propriedades, serem mais adequados para a análise de dados que foram coletados retrospectivamente, tal como em estudos de caso-controle.

O modelo logístico foi apresentado primeiramente por Fisher e Yates (1938). Duas décadas depois, Cox (1958) escreveu um importante trabalho sobre o assunto, mas foi a partir da utilização por Truett, Cornfield e Kannel (1967) para analisar os dados do estudo de Framingham sobre doenças coronárias que o modelo de regressão logística passou a

ser mais empregado. Isso fez com que crescesse muito o interesse pelas medidas de risco que podem ser obtidas pela regressão logística. Nos últimos anos, sua utilização tem crescido ainda mais nas diversas áreas do conhecimento, em especial em estudos na área da saúde. Realizando uma busca no PubMed[1] por "logistic regression", em 10 de dezembro de 2005, 42.718 artigos foram encontrados. Se colocarmos essa mesma expressão no Google[2], 2.110.000 resultados aparecem entre artigos, trabalhos acadêmicos, apostilas,... Mesmo em situações em que a variável resposta é contínua, existe uma preferência por categorizá-la, estabelecendo um ponto de corte, justamente para utilizar a regressão logística. A principal razão de "tanto sucesso" é a possibilidade de interpretar os resultados do modelo como razões de chances. Desse modo, é possível atribuir o risco de um indivíduo ou grupo de indivíduos vir a apresentar determinada característica (por exemplo, uma doença) dado que ele pertença à determinada categoria da variável preditora.

Para que um modelo seja considerado adequado, é necessário que:

(1) as principais variáveis explicativas que contribuem para o melhor ajuste dos dados estejam devidamente especificadas e

(2) o modelo não seja influenciado por determinadas características dos dados, como observações que causem alguma mudança nas estimativas dos coeficientes, superestimando-as ou subestimando-as.

Para identificar a ocorrência de observações atípicas, são utilizadas as técnicas de diagnóstico do modelo. Essas técnicas também verificam se as suposições do modelo estão satisfeitas, se há presença de *outliers* e o se o modelo está bem ajustado para o conjunto completo de covariáveis.

---

[1] http://www.pubmed.com/

[2] http://www.google.com/

As técnicas de diagnóstico para modelos de regressão logística foram implementadas por Pregibon (1981) e são bastante semelhantes às técnicas de diagnóstico da regressão linear.

Uma das técnicas de diagnóstico mais usada para modelos de regressão é a análise de resíduos. Um resíduo pode ser definido como a distância entre o valor estimado e o valor observado correspondente da variável dependente (Cox e Snell, 1968). O principal objetivo da análise de resíduos na regressão logística é identificar casos para os quais as estimativas do modelo se distanciam muito dos valores observados, ou casos que exerçam uma influência maior do que deveriam nas estimativas dos parâmetros do modelo. Para detectar pontos de alavanca, utiliza-se a matriz de projeção, ou matriz H, apresentada por Hoaglin e Welsh (1978) para o modelo de regressão normal e adaptada por Pregibon (1981) para o modelo de regressão logística.

Outros métodos de diagnóstico descritos para regressão linear foram adaptados para situações onde a resposta é binária. Esses métodos são usados para verificar o efeito que uma observação produz sobre as estimativas dos parâmetros e incluem técnicas como a distância de Cook (1977), DFBETAS e DFFITS, proposta por Belsley, Kuh & Welsh (1980). Essas técnicas trabalham com a exclusão da observação em estudo para avaliar seu impacto nas estimativas da regressão.

Entretanto, apesar da importância das técnicas de diagnóstico, são poucos os artigos que apresentam seus resultados. Bagley et el (2001), que pesquisou o uso e apresentação da regressão logística na literatura médica, não encontrou, em nenhum dos artigos analisados, as técnicas de diagnóstico utilizadas pelos autores, ou qualquer menção de que elas ao menos tenham sido usadas.

1.1. Objetivos

Os principais objetivos deste trabalho são:

- Descrever os métodos de diagnóstico mais empregados no modelo de regressão logística.
- Aplicar esses métodos de diagnóstico em um conjunto de dados reais.

Para a aplicação, serão utilizados dados referentes a uma coorte de 3470 meninos nascidos em Ribeirão Preto entre 1978 e 1979. Dois momentos serão analisados: ao nascimento e aos 18 anos, quando esses indivíduos foram reavaliados ao se apresentarem para o serviço militar. Para o modelo de regressão logística foi considerada como variável resposta o sobrepeso ou obesidade no início da vida adulta e como preditoras, variáveis relacionadas ao nascimento e condições sociais atuais do indivíduo. As análises foram realizadas no software SPSS (Statistical Package for Social Science).

2. REGRESSÃO LOGÍSTICA

Em estudos epidemiológicos, há interesse em se verificar, por exemplo, quais variáveis são fatores de risco para um paciente apresentar ou não uma doença, ou se determinado tratamento produziu ou não o efeito esperado. Nesses casos, a variável resposta é do tipo binária, podendo ser representada pelos valores 1 (paciente doente) ou 0 (paciente sem a doença). Os fatores de risco são as variáveis independentes ou variáveis explicativas. Para modelar essa relação, entre a resposta e as variáveis independentes, existe o modelo de regressão logística.

2.1 Definição do modelo de regressão logística

Nos modelos de regressão, estuda-se a relação entre a variável resposta (Y) e um conjunto de variáveis independentes (X_1, X_2,...,X_p). A média condicional de Y é definida por $E(Y \mid x)$, onde Y é a variável de interesse (variável dependente) e x representa o vetor de variáveis independentes. Na regressão logística, por conveniência, denota-se a variável Y como 0 ou 1, onde 1 denota a ocorrência do evento de interesse, sendo que Y pode assumir o valor 1, dado x, com probabilidade $\pi(x)$, ou valor 0 com probabilidade $[1-\pi(x)]$. As variáveis independentes x podem ser categóricas ou contínuas.

Essa média, nos modelos de regressão linear, pode ser expressa como uma função linear em:

$$E(Y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \qquad (2.1)$$

Essa expressão permite que $E(Y \mid x)$ assuma qualquer valor se x_i estiver entre $-\infty$ e $+\infty$. Entretanto, para variáveis dicotômicas os valores da média condicional devem ser

maiores do que zero e menores do que um. Isso ocorre porque a esperança de uma variável aleatória discreta é dada por:

$$E(Y \mid x) = 0 \times P(Y = 0 \mid x) + 1 \times P(Y = 1 \mid x) = P(Y = 1 \mid x) = \pi(x) \qquad (2.2)$$

e, portanto, $0 \leq E(Y \mid x) \leq 1$.

A relação entre as variáveis preditoras e a variável resposta não é linear na regressão logística. A curva de regressão logística apresenta forma de S (ver figura 2.1). Para $\beta > 0$, à medida que os valores de x aumentam, $\pi(x)$ tende a 1. Para $\beta < 0$, à medida que os valores de x aumentam, $\pi(x)$ tende a 0. Quanto mais próximo β estiver de zero, a curva de regressão tende a uma linha reta horizontal. Se $\beta = 0$, a variável resposta é independente de x (Agresti, 1990).

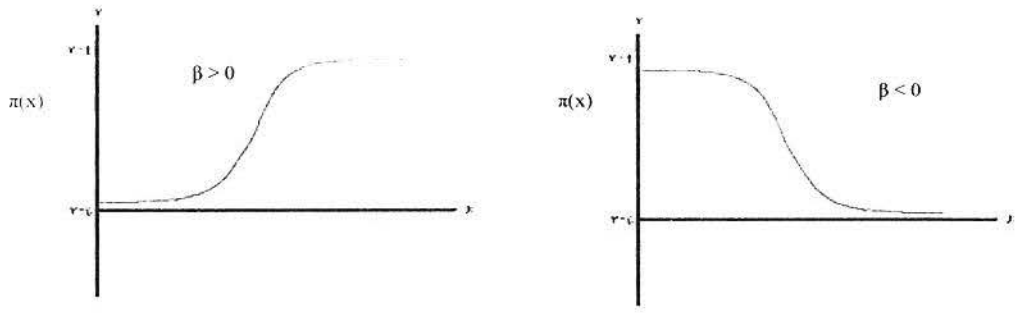


Figura 2.1: Gráfico da função logística.

O modelo de regressão logística é definido por:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}. \qquad (2.3)$$

onde $\pi(x) = E(Y \mid x)$ e p é o número de variáveis independentes consideradas no (2.2)

Através da transformação logito, obtém-se uma relação linear entre essas variáveis:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p . \qquad (2.4)$$

Segundo Hosmer & Lemeshow, a importância dessa transformação é que o g(x) possui muitas das propriedades de um modelo de regressão linear. O logito, g(x), linear nos parâmetros, pode ser contínuo e pode estar entre os valores $-\infty$ e $+\infty$, dependendo dos valores de x.

Na regressão logística, diferentemente da regressão linear, os erros não seguem distribuição Normal. Neste modelo, pode-se expressar o valor da variável de interesse, dado um valor x, por $y = \pi(x) + \varepsilon$, sendo ε o erro associado ao modelo. No modelo de regressão logística, ε só pode assumir dois valores. Se y=1, então $\varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$ e se y=0, então, $\varepsilon = -\pi(x)$ com probabilidade $[1 - \pi(x)]$. Assim, ε tem distribuição com média zero e variância igual a $\pi(x)[1-\pi(x)]$. Ou seja, a distribuição condicional da variável resposta segue distribuição Binomial com probabilidade dada pela média condicional $\pi(x)$. (Hosmer & Lemeshow, 1989).

2.2 Estimação através da função de Verossimilhança

A estimação dos parâmetros do modelo é realizada utilizando a função de máxima verossimilhança. Considere uma amostra com n observações, desfecho dicotômico denotado por y_i, valor da i-ésima observação da j-ésima variável independente denotada por x_{ij}, e o vetor de parâmetros a serem estimados, $\beta' = (\beta_0, \beta_1, ..., \beta_p)$. A função de verossimilhança é definida por:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} . \qquad (2.5)$$

A estimativa de máxima verossimilhança de β é obtida maximizando-se $l(\beta)$. Entretanto, é mais fácil maximizar o logaritmo de $l(\beta)$, definido como:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \quad (2.6)$$

Serão, então, obtidas p+1 equações de máxima verossimilhança para os p+1 parâmetros que desejamos estimar. Essas equações podem ser expressas por:

$$\sum_{i=1}^{n} y_i \ln[\pi(x_i)] = 0 \quad (2.7)$$

e

$$\sum_{i=1}^{n} x_{ij} \ln[y_i - \pi(x_i)] = 0 \quad (2.8)$$

para j = 1, 2,...,p.

As equações de verossimilhança, para a regressão logística, são não-lineares em β, e assim, necessita-se de métodos especiais para solucioná-las. São métodos iterativos e requerem auxilio computacional, disponíveis na maioria dos softwares estatísticos.

Exemplo 1: O presente exemplo analisa os dados de um dos exercícios propostos por Hosmer e Lemeshow (1989). O banco mostra dados sobre a unidade de tratamento intensivo – UTI (*Intensive Care Unit - ICU*) do Baystate Medical Center. A variável dependente é STA que é o estado vital do indivíduo (0, se estiver vivo e 1, se estiver morto). Serão utilizadas duas variáveis preditoras para o estado vital do paciente: a variável contínua pressão sanguínea sistólica (SYS), em mm Hg, e a variável dicotômica tipo de admissão (TYP, 0: eletiva ou 1: emergencial). Serão considerados três modelos:

(1) $g(x) = \beta_0 + \beta_1 TYP$

(2) $g(x) = \nu_0 + \nu_1 SYS$

(3) $g(x) = \tau_0 + \tau_1 TYP + \tau_2 SYS$

Utilizando o SPSS, foram obtidas as estimativas, para o modelo 3, para τ_0, τ_1 e τ_2. Essas estimativas são apresentadas na tabela 2.1.

Tabela 2.1 - Estimativas para os coeficientes estimados para cada variável e seus respectivos erros padrões:

Variável	Coeficiente Estimado ($\hat{\tau}_i$)	Erro Padrão
Constante	-1,330	1,079
Pressão sistólica	-0,014	0,006
Tipo de admissão	2,022	0,750

As estimativas de máxima verossimilhança de τ_0, τ_1 e τ_2 são, respectivamente, $\hat{\tau}_0 = -1{,}330$, $\hat{\tau}_1 = -0{,}014$ e $\hat{\tau}_2 = 2{,}022$. O logito, g(x), pode ser estimado por:

$$\hat{g}(x) = -1{,}330 - 0{,}014 \times TYP + 2{,}022 \times SYS .$$

2.3 Teste de significância dos coeficientes

Após estimar os coeficientes do modelo, é necessário avaliar a significância das variáveis no modelo. Essa avaliação é realizada através de testes de hipóteses que verificam se os coeficientes são significativamente diferentes de zero, ou seja, se existe uma relação significativa entre a variável preditora e o desfecho.

2.3.1 Teste da Razão de Verossimilhança

O teste da razão de verossimilhança é utilizado para verificar a significância de cada um dos coeficientes β_i, i= 1, 2,...,p, do modelo. Este teste compara a diferença entre o

logaritmo da função de verossimilhança do modelo completo (com todas as variáveis, incluindo aquela que está tendo seu coeficiente testado) com o logaritmo da função de verossimilhança do modelo sem a variável. Assim, a estatística do teste, G, é dada por:

$$G = -\ln\left[\frac{\text{verossimilhança do modelo sem a variável}}{\text{verossimilhança do modelo com a variável}}\right]. \qquad (2.9)$$

Sob a hipótese de que β_i é igual a zero, a estatística G segue a distribuição Qui-Quadrado com 1 grau de liberdade.

Usando os mesmos dados do Exemplo 1, para o modelo 1 obtemos, através do SPSS:

[-2*ln("verossimilhança do modelo apenas com a constante")] = 200,161

e [–2*ln("verossimilhança do modelo com a constante e tipo de admissão")] = 185,049.

Podemos calcular G=(200,161-185,049) = 15,112. Comparando o valor de G com o valor da distribuição Qui-Quadrado com 1 grau de liberdade, $\chi^2_{(1);0,05}$ = 3,84 (p < 0,001), pode-se concluir que o tipo de admissão deve ser uma variável significativa na predição do estado vital

O teste da Razão de Verossimilhança também pode ser utilizado para determinar se o modelo total é estatisticamente significativo.

Considerando o modelo 3 do exemplo 1 são obtidos através do SPSS:

[-2*ln("verossimilhança do modelo apenas com a constante")] = 200,161

e –2*ln("verossimilhança do modelo com as duas variáveis") = 179,372.

Então, podemos calcular G = (200,161-179,372) = 20,789 > $\chi^2_{(2);0,05}$ = 5,99

(p<0,001). Conclui-se que pelo menos um dos $\hat{\tau}_i$, i=1, 2, deve ser diferente de zero. Para verificar se os 2 coeficientes são deferentes de zero é necessário testa-los separadamente.

2.3.2 Teste de Wald

O teste de Wald é usado para testar a significância de cada um dos coeficientes do modelo. Este teste compara a estimativa de máxima verossimilhança do parâmetro β_i, i= 1,2,...p, com a estimativa do seu erro padrão, $\hat{EP}(\hat{\beta}_i)$. A estatística W para o teste de Wald é dada por:

$$W = \frac{\hat{\beta}_i}{\hat{EP}(\hat{\beta}_i)}. \qquad (2.10)$$

Sob a hipótese de que β_i é igual a zero, W segue distribuição Normal padrão.

Utilizando o modelo 1, temos, $\hat{\beta}_1 = 2{,}185$ e $\hat{EP}(\hat{\beta}_1) = 0{,}745$. Portanto,

$$W = \frac{\hat{\beta}_i}{\hat{EP}(\hat{\beta}_i)} = \frac{2{,}185}{0{,}745} = 2{,}933.$$

Podemos comparar o valor de W = 2,933 com o valor tabelado z = 1,96 da distribuição normal para um nível de significância de 5%. Assim, temos que W = 2,933 > 1,96 (p=0,005) e podemos concluir que $\hat{\beta}_1 = 2{,}185$ deve ser diferente de zero, ou seja, que a variável estado vital está relacionada com o tipo de admissão, da mesma forma que concluímos no teste da razão de verossimilhança.

É importante ressaltar que alguns autores encontraram problemas na utilização deste teste. Hosmer e Lemeshow (1989) mencionam que Hauck e Donner (1977) verificaram que algumas vezes este teste tende a não rejeitar H_0 quando o coeficiente β_i é

significativo. Menard (1995) adverte que para coeficientes grandes o erro padrão aumenta, reduzindo o valor da estatística de Wald. Esses autores recomendam que se utilize o teste da razão de verossimilhança.

2.4 Interpretação dos Coeficientes

Os coeficientes estimados para as variáveis independentes representam a inclinação ou taxa de mudança de uma função da variável dependente por unidade de mudança para a variável independente. Entretanto, o modelo logístico pode ser escrito em termos da chance de um evento ocorrer.

2.4.1 Para variáveis explicativas dicotômicas:

A chance do desfecho estar presente entre os indivíduos com x=1 é definida como $\pi(1)/[1-\pi(1)]$. De maneira similar, a chance do desfecho estar presente entre os indivíduos com x=0 é definida como $\pi(0)/[1-\pi(0)]$. A razão de chances, ψ, é definida como a razão entre a chance para x=1 e a chance para x =0, e é dada pela equação:

$$\psi = \left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right]. \qquad (2.11)$$

O logaritmo da razão de chances é:

$$\ln(\psi) = \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] = g(1) - g(0) \qquad (2.12)$$

que é a diferença do logito.

Utilizando a expressão do logito dada por (2.4), e usando a parametrização "categoria de referência", obtém-se:

$$g(1) = \beta_0 + \beta_1 \times 1 \qquad \text{e} \qquad g(0) = \beta_0 + \beta_1 \times 0$$

$$\psi = e^{g(1)-g(0)} = e^{\beta_0+\beta_1-\beta_0}$$

$$\psi = e^{\beta_1}. \tag{2.13}$$

A razão de chances é uma medida de associação de quão mais provável (ou improvável) é para o desfecho estar presente entre os indivíduos que apresentam x=1 do que entre aqueles com x=0.

Exemplo: Para o modelo 1, obtém-se a razão de chances estimada, $\hat{\psi} = e^{2,185} = 8,890$. O intervalo com $100\times(1-\alpha)\%$ para $\hat{\psi}$ é obtido por $\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2}\hat{EP}(\hat{\beta}_1)\right]$. Assim, o intervalo com 95% de confiança para a variável tipo de admissão é (2,064; 38,298). Pode-se dizer que, um paciente com admissão emergencial tem 8,89 vezes a chance de morrer do que um paciente com admissão eletiva.

2.4.2 Para variáveis explicativas contínuas:

A razão de chances para variáveis contínuas pode ser obtida de maneira similar a das variáveis categóricas. Mais detalhes podem ser obtidos em Hosmer e Lemeshow (1989). ψ pode ser interpretado como o quanto a chance de y=1 se modifica (aumenta ou diminui) pelo acréscimo de 1 unidade em x.

Exemplo: Considerando o modelo 2, obtém-se $\hat{\nu}_1$= -0,017. Assim, a razão de chances estimada, $\hat{\psi} = e^{-0,017} = 0,983$ (IC95%: 0,972; 0,995). Estima-se que, um indivíduo tem 1,7% (1-0,983$\times$100%) menos chance de morrer pelo acréscimo de 1 unidade na

pressão sistólica.

2.5 Seleção de Modelos

A regressão logística tem como objetivo determinar o modelo que melhor expresse a relação existente entre as variáveis preditoras e a variável resposta. Geralmente são apontados pelo pesquisador alguns fatores que podem ser incluídos no modelo. É necessário, então, utilizar uma boa estratégia de modelagem para descobrir quais dessas variáveis e interações que realmente devem permanecer no modelo. Hosmer e Lemeshow (1989) sugerem quatro etapas para escolher o modelo mais adequado e parcimonioso. Essas etapas são apresentadas a seguir:

(1) O processo de seleção das variáveis deve começar com uma análise bivariada para cada variável explicativa. Para as variáveis independentes categóricas pode-se fazer uma tabela de contingência da variável de interesse versus as k categorias da variável independente. Para verificar se a variável deve entrar no modelo utiliza-se o Teste Qui-Quadrado com K-1 graus de liberdade ou o teste da razão de verossimilhança para a significância do coeficiente de uma variável independente no modelo de regressão logística bivariado, que são equivalentes. Para variáveis contínuas é possível verificar se o coeficiente da variável em questão é significativo através do modelo bivariado ou realizando-se o teste t de Student. Também é necessário verificar se existe linearidade no logito para as variáveis contínuas. Isso pode ser feito através do gráfico de valores preditos versus valores observados da variável contínua. A linearidade no logito também pode ser testada a partir da razão de verossimilhança com descrito em Collet (1994).

(2) Após realizar a análise bivariada, as variáveis que apresentarem $p<0,25$ são consideradas candidatas a entrar no modelo multivariado (Hosmer e Lemeshow,

1989). Para o modelo multivariado existem várias estratégias de modelagem. As mais utilizadas são: método forward, método backward e método stepwise que serão apresentados posteriormente. É necessário também verificar a existência de interações entre as variáveis e de possíveis fatores de confusão.

(3) A importância de cada variável no modelo deve ser verificada após a escolha do modelo multivariado. É necessário realizar o teste de Wald ou o teste da Razão de Verossimilhança para cada variável e deve ser feita uma comparação entre os coeficientes estimados para cada variável no modelo bivariado com os do modelo multivariado. As variáveis que não apresentarem contribuição para o modelo baseado nesses critérios devem ser retiradas do modelo. Fatores de confusão devem ser mantidos no modelo.

(4) Após obter o modelo que contém todas as variáveis consideradas essenciais, deve-se considerar a inclusão de termos de interação entre essas variáveis.

A seguir são apresentados os três métodos de seleção de variáveis citados anteriormente:

2.5.1 Método Forward

Inicia-se com o modelo que contém apenas a constante $E(Y \mid x) = \beta_0$. Para cada variável explicativa ajusta-se o modelo $E(Y \mid x) = \beta_0 + \beta_j x_j$ (j=1,...q). Testa-se $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$. Para os q testes, as variáveis explicativas que tiverem p-valor < p_c (p_c: p critico) são inseridas no modelo uma a uma. Entra, num primeiro momento, aquela variável que tiver o menor p; então, são feitos novos modelos para cada uma das demais variáveis. Cada modelo contém a constante e as variáveis incluídas até o momento. Esse procedimento é repetido até que p-valor >p_c.

2.5.2 Método Backward

Inicia-se com o modelo contendo todas as variáveis de interesse. Testa-se $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$ (j=1,...,q). Para os q testes, as variáveis explicativas que tiverem p-valor > p_c (p_c: p critico) são retiradas do modelo. Repete-se este procedimento até que todas as variáveis restantes no modelo apresentem p-valor < p_c.

2.5.3 Método Stepwise

É uma mistura dos dois procedimentos acima. Inicia-se o processo com o modelo $E(Y \mid x) = \beta_0$. Após duas variáveis terem sido incluídas no modelo, verifica-se se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou retirada do modelo. Geralmente adota-se, para uma variável entrar no modelo p_c<0,25 e para uma variável sair do modelo p_c> 0, 05.

No modelo de regressão logística, entretanto, existem outros fatores que devem ser considerados na seleção do modelo. Por exemplo, existem as variáveis de importância biológica que devem permanecer no modelo mesmo que não sejam estatisticamente significativas. Geralmente é utilizado o método stepwise, respeitando a permanência das variáveis de importância biológica que possam contribuir para uma melhor interpretação dos resultados gerados pelo modelo final. Além disso, quando são incluídas interações significativas entre variáveis, seus efeitos principais de ordem inferior também devem permanecer no modelo.

2.6 Qualidade de Ajuste do Modelo (Goodness-of-Fit)

Após escolher as variáveis (efeitos principais e interações) que compõem o modelo,

deve-se verificar se o modelo é eficiente para descrever a relação entre as variáveis preditoras e a variável resposta. Assim, é necessário verificar o ajuste do modelo. Segundo Hosmer & Lemeshow (1989), pode-se concluir que o modelo está bem ajustado se: (1) as medidas resumo da distância entre y e $\hat{y}$ são pequenas e (2) a contribuição de cada par $(y_i;\hat{y}_i)$, i = 1, 2, 3,...,n para essas medidas resumo é não-sistemática e é pequena em relação à estrutura de erro do modelo. Assim, uma avaliação completa do modelo ajustado envolve tanto o cálculo de medidas resumo da distância entre y e $\hat{y}$ quanto o exame dos componentes individuais dessas medidas.

Algumas dessas medidas serão apresentadas a seguir. Será considerado o modelo com p variáveis explicativas que formam J padrões de covariáveis. Os padrões de covariáveis são formados através das diversas combinações entre os possíveis valores das variáveis independentes. Por exemplo, para duas variáveis independentes, X_1 com duas categorias e X_2 com quatro categorias, existem 2*4 = 8 padrões de covariáveis. Se não há nenhum valor repetido de x entre os casos estudados, J=n.

2.6.1 Qui-Quadrado de Pearson

O Qui-Quadrado de Pearson testa a hipótese de que o modelo está bem ajustado, usando a estatística Qui-Quadrado em função dos resíduos de Pearson. Ou seja, H_0: o modelo está bem ajustado.

A estatística Qui-Quadrado de Pearson é definida por:

$$X^2 = \sum_{j=1}^{J} r(y_j;\hat{\pi}_j)^2 , \qquad (2.14)$$

onde $r(y_j;\hat{\pi}_j) = \dfrac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$ (2.15)

é conhecido como o resíduo de Pearson e $m_j\hat{\pi}_j = \hat{y}_j$, ou seja, são os valores estimados para y.

A estatística X^2 sob a hipótese de que o modelo está bem ajustado segue distribuição Qui-Quadrado com J-(p+1) graus de liberdade. Quanto menor for o valor do X^2 de Pearson, melhor é o ajuste do modelo.

2.6.2 Deviance

Considerando m_j, j= 1,2,...,J, como o número de indivíduos com o mesmo padrão j de covariáveis, ou seja, m_j é o número de indivíduos que apresentam exatamente os mesmos valores para cada uma das p variáveis independentes.

Os resíduos Deviance são definidos por:

$$d(y_j;\hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{(m_j - y_j)}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2} \qquad (2.16)$$

onde o sinal é o mesmo de $(y_j - m_j\hat{\pi}_j)$. Para simplificar a notação $d(y_j;\hat{\pi}_j) = d_j$ será usada.

A estatística Deviance é baseada nesses resíduos:

$$D = \sum_{j=1}^{J} d(y_j;\hat{\pi}_j)^2 . \qquad (2.17)$$

A estatística D, sob a hipótese de que o modelo está bem ajustado, segue distribuição Qui-Quadrado com J-(p+1) graus de liberdade. Assim como para o Qui-Quadrado de Pearson, quanto menor for o valor de D, melhor é o ajuste do modelo.

Para obter o valor da estatística Deviance no SPSS, é necessário obter os resíduos

Deviance através da opção *save* do procedimento *logistic regression*, elevar esses valores ao quadrado e depois somá-los.

Entretanto, tanto para a estatística X^2 como para a D, para valores de J ≈ n, o número de parâmetros do modelo cresce com a mesma taxa que o tamanho da amostra, fazendo com que os valores de p calculados sob a hipótese de que X^2 ou D sigam distribuição $\chi^2_{J-(p+1)}$, estejam incorretos. Mais detalhes podem ser encontrados em Hosmer e Lemeshow (1989). Uma alternativa, nesses casos é o Teste de Hosmer e Lemeshow (1980, 1982).

2.6.3 Teste de Hosmer & Lemeshow

Hosmer & Lemeshow (1980, 1982) propuseram um procedimento que utiliza os valores de probabilidade preditos para criar grupos. As hipóteses são:

$H_0: E[Y] = \frac{e^{(X'\beta)}}{1+e^{(X'\beta)}}$, ou equivalentemente, o modelo está bem ajustado.

$H_1: E[Y] \neq \frac{e^{(X'\beta)}}{1+e^{(X'\beta)}}$, ou equivalentemente, o modelo não está bem ajustado.

Para realizar esse teste, os valores de probabilidade preditos são ordenados. A partir desses valores, são criados g grupos. Os autores recomendam criar g= 10 grupos de aproximadamente mesmo tamanho. Os grupos são criados de maneira que o primeiro tenha probabilidade predita entre 0,0 e 0,1, o segundo, entre 0,1 e 0,2 e assim por diante até que o décimo grupo tenha valores de probabilidade predita entre 0,9 e 1,0. Depois são calculados os valores esperados para cada grupo e comparados com os valores observados usando a estatística Qui-Quadrado de Pearson. Estudos usando simulação mostraram que, para amostras grandes, a estatística do teste tem distribuição aproximada Qui-Quadrado com g-2 graus de liberdade, onde g é o número de grupos utilizados.

Para o modelo do exemplo 1, são apresentados na tabela 2.2 o número de indivíduos observado e esperado sob a hipótese nula para cada um dos 10 grupos, obtidos através do SPSS:

Tabela 2.2 - Número observado e esperado de indivíduos por grupo.

	STA = 0		STA = 1		
	Observado	Esperado	Observado	Esperado	Total
1	19	19,53	1	,47	20
2	19	19,22	1	,78	20
3	21	21,16	2	1,84	23
4	18	16,46	2	3,54	20
5	12	12,63	4	3,37	16
6	14	15,41	6	4,59	20
7	15	14,98	5	5,02	20
8	15	12,96	3	5,04	18
9	15	12,52	3	5,48	18
10	12	15,13	13	9,87	25

Na tabela 2.3 são apresentados o valor de χ^2 e significância estatística para o Teste de Hosmer e Lemeshow com 8 graus de liberdade.

Tabela 2.3 - Teste de Hosmer e Lemeshow

Valor χ^2	GL	p-valor
6,619	8	0,578

Como o Teste de Hosmer e Lemeshow apresenta para o valor de $\chi^2_{calculado} = 6,614$ p = 0,578 não podemos rejeitar H_0, ao nível de significância de 5%. Desse modo, conclui-se que o modelo pode estar bem ajustado.

3. TÉCNICAS DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO LOGÍSTICA

Estatísticas de ajuste do modelo como X^2 de Pearson, Deviance e o Teste de Hosmer & Lemeshow são indicadores da qualidade do modelo como um todo. Entretanto, é necessário verificar também o diagnóstico do modelo para cada ponto, isto é, verificar a dependência do modelo estatístico em relação às várias observações que foram coletadas. Para isso são usadas as técnicas de diagnóstico, que verificam se as suposições do modelo estão satisfeitas e identificam características dos dados, como observações influentes, que causem alguma mudança nas estimativas dos coeficientes, levando a problemas nas conclusões geradas pelo modelo.

Alguns conceitos devem ser definidos antes da utilização das técnicas de diagnóstico:

- *Ponto de alavanca:* Paula (2004) diz que a definição de pontos de alavanca foi motivada pelo estudo da diagonal principal da matriz de projeção $H = X(X'X)^{-1}X$ apresentada por Hoaglin e Welsch (1978). Os pontos de alavanca receberam esse nome por terem um peso desproporcional no próprio valor ajustado. São pontos que apresentam um perfil diferente dos demais no que diz respeito aos valores das variáveis explicativas.

- *Observação influente*: Belsey, Kuh & Welsh (1980) definem observação influente como aquela observação que, individualmente ou junto com outras observações, tem claramente um impacto maior nos valores calculados de várias estimativas do que as demais observações.

Nas figuras 3.1a e 3.1b são identificados cada um desses pontos. A reta de linha contínua representa o modelo estimado com o ponto e a reta de linha pontilhada, o modelo estimado sem o ponto. Na figura 3.1a pode-se ver que o ponto #5 é um ponto de influência, já que ele provoca uma alteração nas estimativas dos coeficientes, fazendo com que a inclinação da reta de regressão se modifique. Já na figura 3.1b, pode-se ver que o ponto #15 está afastado do restante dos pontos em relação a X. Entretanto, a eliminação do mesmo não muda em quase nada as estimativas dos parâmetros, a retas quase se sobrepõem.

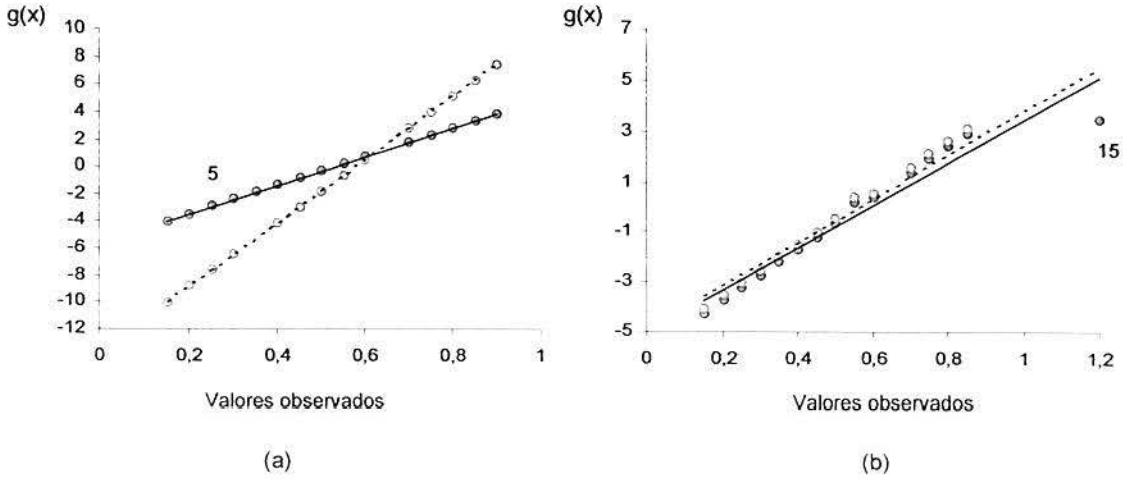


Figura 3.1: Valores observados versus g(x)

Para detectar esses problemas, existem várias técnicas na regressão linear normal que podem ser vistas, por exemplo, em Belsley et al. (1980) e Cook & Weisberg (1982). Pregibon (1981) estendeu os resultados usados nos Modelos Lineares Generalizados para a Regressão Logística. A seguir serão apresentadas algumas destas técnicas de diagnóstico para modelos de regressão logística.

Para as técnicas apresentadas a seguir será considerado o modelo com p variáveis explicativas que formam J padrões de covariáveis. Como foi definido anteriormente, m_j é o número de indivíduos que apresentam exatamente os mesmo valores para cada uma das

p variáveis independentes e segue que $\sum m_j = n$.

3.1 Pontos de Alavanca (leverage values)

No modelo de regressão linear normal, uma medida de alavancagem é dada pelos elementos da diagonal da matriz:

$$H = X(X'X)^{-1}X' \tag{3.1}$$

conhecida como matriz de projeção, ou matriz *chapéu,* pois $\hat{y} = Hy$. Os resíduos $\hat{\varepsilon} = y - \hat{y}$, podem ser expressos em função da matriz H, como $\hat{\varepsilon} = (I - H)y$, onde I é a matriz identidade JxJ. Usando a regressão de mínimos quadrados ponderados como modelo, Pregibon (1981) realizou uma aproximação linear para os valores ajustados definindo a matriz H para a regressão logística por:

$$H = V^{1/2}X(X'VX)^{-1}X'V^{1/2} \tag{3.2}$$

onde V é uma matriz diagonal JxJ com o elemento $v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$.

Desse modo, para o modelo de regressão logística, os pontos de alavanca, denotados por h_j para a j-ésima diagonal de H, são dados por:

$$h_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)](1, x_j')(X'VX)^{-1}(1, x_j')' = v_j \times b_j \tag{3.3}$$

onde $b_j = (1, x_j')(X'VX)^{-1}(1, x_j')'$.

Como a matriz H é simétrica e idempotente, tem-se que: (a) $0 \leq h_j \leq 1$; (b) $tr(H) = \sum_{j=1}^{n} h_j = p + 1$.

Os pontos de alavanca (h_j) medem o quão distante a observação x_j está das demais n-1 observações no espaço definido pelas variáveis explicativas do modelo.

Como se pode ver por (3.3), o elemento h_j só depende dos valores das variáveis explicativas, isto é, da matriz X, e não envolve as observações em y. Se h_j é grande, os valores das variáveis explicativas associadas a j-ésima observação são atípicos, ou seja, estão distantes do vetor de valores médios das variáveis explicativas.

Segundo Cordeiro e Lima Neto (2004), esses pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, ou seja, sua exclusão pode implicar mudanças dentro de uma análise estatística. Paula (2004) afirma que os pontos de alavanca podem ser também informativos com relação à estimativa de β. Entretanto, o exemplo apresentado por ele mostra um ponto de alavanca que, quando eliminado, não muda praticamente nada nas estimativas dos parâmetros.

Alguns autores, como Belsey et al (1980), sugerem que pontos com h_j maior ou igual a $2(p+1)/n$ devem ser investigados, onde p é o número de parâmetros do modelo incluindo os parâmetros das variáveis indicadoras relativas a cada classe de uma variável categórica e a constante. Entretanto, Hosmer e Lemeshow (1989) advertem que quando o número de padrões das covariáveis (J) é muito menor do que n, existe risco de falha para identificar pontos de alavanca. Desse modo, outras medidas devem ser utilizadas para confirmar esse primeiro diagnóstico.

3.2 Resíduos

Um resíduo pode ser definido como a distância entre o valor estimado e o valor observado correspondente da variável Y. A seguir serão definidos os resíduos de Pearson, resíduos Deviance e resíduos Stundetizados, que são úteis para identificar observações

que não estão sendo bem explicadas pelo modelo, como, por exemplo, pontos aberrantes. Esses resíduos serão definidos para cada um dos m grupos com o mesmo padrão de covariáveis. Todas as observações do grupo recebem o mesmo valor predito, independente de terem desfecho 0 ou 1.

3.2.1 Resíduos Padronizados de Pearson

Os resíduos de Pearson, dados por (2.15) são definidos como uma comparação entre a diferença de um valor observado e o seu respectivo valor estimado com o desvio padrão estimado para essa observação (Agresti, 1990). O resíduo padronizado de Pearson para um padrão de covariável x_j é definido por:

$$r_{sj} = r(y_j;\hat{\pi}_j)/\sqrt{1-h_j} \qquad (3.4)$$

3.2.2 Resíduos Deviance

Os resíduos Deviance, dados por (2.16), medem o grau de discordância entre o máximo da função de verossimilhança observada e da estimada. Como a regressão logística usa o principio da máxima verossimilhança, o objetivo é minimizar a soma dos resíduos Deviance.

3.2.3 Resíduos Studentizados

O resíduo studentizado é o quociente entre o resíduo e a estimativa do seu desvio padrão. Entretanto, o desvio-padrão dos resíduos não é constante, sendo diferente para os diversos valores da variável de resposta. Ele é maior para respostas mais próximas da média desta variável. O resíduo studentizado é definido de maneira a garantir a independência entre numerador e denominador na padronização dos resíduos. Assim, define-se o resíduo studentizado por:

$$r_{ij} = \frac{y_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}\sqrt{(1-h_j)}} \qquad (3.5)$$

Segundo Pregibon (1981), medidas como h_j ou resíduos, são úteis para detectar valores extremos, mas não mostram qual o impacto que esses valores têm nos vários aspectos do ajuste do modelo como, por exemplo, nos parâmetros estimados, e nas estatísticas de qualidade do ajuste do modelo. A seguir serão apresentadas três técnicas para detectar medidas de influência e que tem como objetivo quantificar o efeito de cada observação no ajuste do modelo.

3.3 Medidas de Influência

As medidas de influência verificam quanto uma observação contribui em modificações nas estimativas dos parâmetros do modelo de regressão. Geralmente, o efeito que uma observação tem nos parâmetros do modelo é verificado através da exclusão dessa observação. As técnicas de diagnóstico do modelo mais usadas para detectar esse tipo de influência são: distância de Cook, DFFITS e DFBETAS.

3.3.1 Distância de Cook

A distância de Cook (1977) tem como objetivo medir a influência de cada observação nos parâmetros do modelo. Essa medida é calculada através da diferença entre $\hat{\beta}$ e $\hat{\beta}_{(-j)}$, que representam, respectivamente, as estimativas de máxima verossimilhança calculadas usando todos os J padrões de covariáveis e excluindo as observações com padrão j e ainda padronizada pela matriz da covariância de $\hat{\beta}$. Pregibon (1981) mostrou através de uma aproximação linear, que a distância de Cook, para a

regressão logística, é dada por:

$$\Delta\hat{\beta}_j = (\hat{\beta} - \hat{\beta}_{(-j)})'(X'VX)(\hat{\beta} - \hat{\beta}_{(-j)}) = \frac{r_j^2 h_j}{(1-h_j)^2} = \frac{r_{sj}^2 h_j}{(1-h_j)} \qquad \textbf{(3.6)}$$

Alguns autores, como Neter et al (1996) e Cordeiro e Lima Neto (2004) sugerem que observações que apresentam $\Delta\hat{\beta}_j \geq F(p+1; n-p-1)$ podem ser consideradas influentes. Recomenda-se avaliar o que acontece com o ajuste do modelo quando essas observações são excluídas. Cordeiro e Lima Neto (2004) dizem que na prática se o maior valor de $\Delta\hat{\beta}_j$ for muito inferior a 1, então a eliminação de qualquer observação do modelo não irá alterar muito as estimativas dos parâmetros.

3.3.2 DFFITS

Proposta por Besley et al (1980), DFFITS é uma medida alternativa à distância de Cook e tem como objetivo medir a influência da j-ésima observação nos parâmetros de locação e a escala do modelo. O nome DFFITS vem do inglês *Difference in Fit*. DFFITS é uma função do resíduo studentizado r_{tj} e da medida de alavancagem h_j, dada por:

$$DFFITS_j = r_{t_j}\left\{\frac{h_j}{(1-h_j)}\right\}^{1/2} \qquad (3.7)$$

Miazaki e Stangenhaus (1994) sugerem que observações com valores absolutos de DFFITS maiores do que $2(n-p-1)^{1/2}p^{-1/2}$ podem ser considerados possíveis pontos de influência. Neter et al (1996), indicam verificar os casos com valor absoluto de DFFITS superior a um para amostras pequenas e observações com valor absoluto de DFFITS maior ou igual a $2\sqrt{(p+1)/n}$ para amostras grandes.

3.3.3 DFBETAS

$DFBETA_j$ mede, para cada coeficiente de regressão β_j relacionada a um preditor X_j, o quando ele se modifica quando a observação j é excluída. A estatística $DFBETA_j$ é definida por:

$$\hat{\beta} - \hat{\beta}_{(j)} = (X'X)^{-1} x'_j (1 - h_j)^{-1} \hat{\varepsilon}_j \qquad (3.8)$$

O sinal do DFBETA indica se a inclusão de uma observação leva a um aumento ou decréscimo dos coeficientes estimados pela regressão e seu valor absoluto mostra o tamanho dessa diferença em relação ao seu desvio padrão estimado. Segundo Neter et al (1996), deve-se dar maior atenção a observações que apresentarem valores absolutos de DFBETAS superiores a 1, para amostras pequenas e valores absolutos de DFBETAS maiores do que $1/\sqrt{n}$, para amostras grandes.

3.4 Análise Gráfica

Geralmente examinam-se as medidas definidas anteriormente através de gráficos, plotando-as versus os valores preditos ou mesmo versus o número de cada observação. Esses dois tipos de gráficos apresentam resultados similares. Entretanto, há uma mudança quanto à disposição dos pontos. O gráfico que apresenta o número da observação mostra, geralmente, os pontos mais espalhados. Através desses gráficos é possível localizar as observações que estão muito afastadas do restante do conjunto de dados. Também se pode ver quando os pontos de corte sugeridos anteriormente realmente estão separando as observações mais extremas. Deve-se dar mais atenção aos pontos que apresentam valores muito altos para as medidas de diagnóstico.

Exemplo: Para ilustrar o uso destes gráficos, serão usados os dados de um experimento que tem como objetivo avaliar o efeito da taxa e do volume de ar inspirado na ocorrência de vaso-constrição na pele dos dedos da mão (Finney, 1978; Pregibon, 1981). Os dados desse experimento são apresentados na tabela 3.1. A variável resposta é a ocorrência (Y=1) ou não ocorrência (Y=0) de compressão de vasos e as variáveis explicativas são logaritmo natural do volume e logaritmo natural da razão de ar inspirado. O modelo é dado por:

$$\hat{g}(x) = \beta_0 + \beta_1 \ln(volume) + \beta_2 \ln(taxa)$$

Tabela 3.1 - Dados do experimento de Finney sobre vaso constrição na pele dos dedos da mão.

Obs	Volume	Razão	Resposta	Obs	Volume	Razão	Resposta
1	3,70	0,825	1	21	0,40	2,000	0
2	3,50	1,090	1	22	0,95	1,360	0
3	1,25	2,500	1	23	1,35	1,350	0
4	0,75	1,500	1	24	1,50	1,360	0
5	0,80	3,200	1	25	1,60	1,780	1
6	0,70	3,500	1	26	0,60	1,500	0
7	0,60	0,750	0	27	1,80	1,500	1
8	1,10	1,700	0	28	0,95	1,900	0
9	0,90	0,750	0	29	1,90	0,950	1
10	0,90	0,450	0	30	1,60	0,400	0
11	0,80	0,570	0	31	2,70	0,750	1
12	0,55	2,750	0	32	2,35	0,030	0
13	0,60	3,000	0	33	1,10	1,830	0
14	1,40	2,330	1	34	1,10	2,200	1
15	0,75	3,750	1	35	1,20	2,000	1
16	2,30	1,640	1	36	0,80	3,330	1
17	3,20	1,600	1	37	0,95	1,900	0
18	0,85	1,415	1	38	0,75	1,900	0
19	1,70	1,060	0	39	1,30	1,625	1
20	1,80	1,800	1				

As estimativas obtidas para os parâmetros são: $\hat{\beta}_0 = -2{,}875$, $\hat{\beta}_1 = 5{,}179$ e $\hat{\beta}_2 = 4{,}562$. Como o teste de Hosmer e Lemeshow apresenta χ^2=11,09 (p=0,197), podemos concluir que o modelo está bem ajustado.

Na figura 3.2, os valores de h_j aparecem plotados versus os valores preditos e versus o número das observações. Analisando esse gráfico, pode-se perceber que o ponto #31 se distancia dos demais pontos, destacando-se.

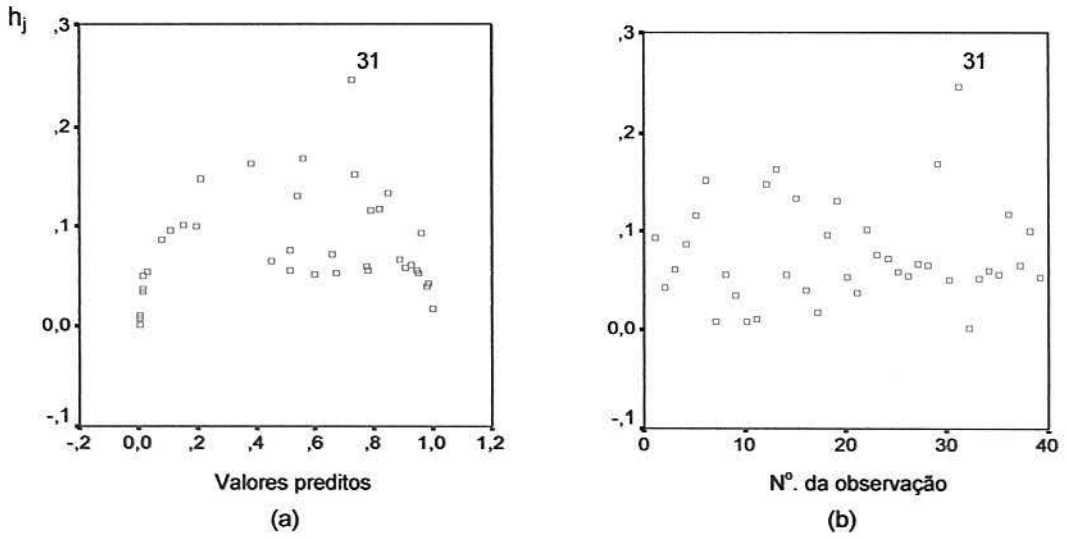


Figura 3.2: Valores de h_j versus valores preditos e versus o número das observações

As figuras 3.3a e 3.3b apresentam os gráficos dos resíduos studentizados versus os valores preditos e versus o número das observações. Pode-se perceber que as observações #4 e #18 estão afastadas do restante das observações. Como sugerido por Hosmer e Lemeshow (1989), ao invés de usar o resíduo Deviance, serão usados resíduo Deviance elevados ao quadrado e denotados por d^2, que pode ser comparado com o valor de $\chi^2_{0,95} = 3{,}84$. Aqui este valor está sendo usado de forma ilustrativa, já que d^2 segue distribuição Qui-Quadrado m-assintoticamente. Nas figuras 3.3c e 3.3d, vemos que as duas observações destacadas no gráfico dos resíduos studentizados, também se apresentam distantes do restante das observações nos gráficos do d^2.

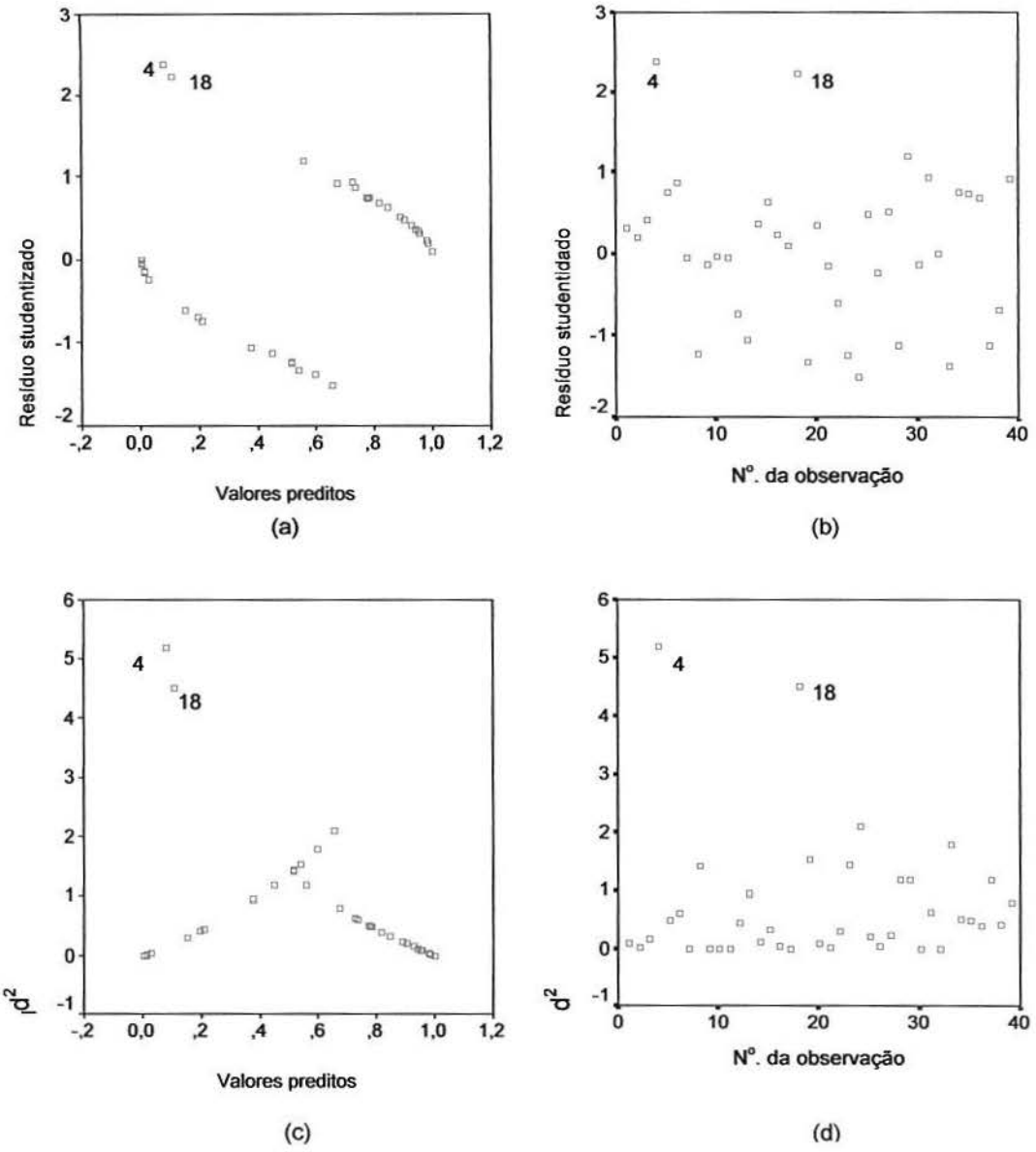
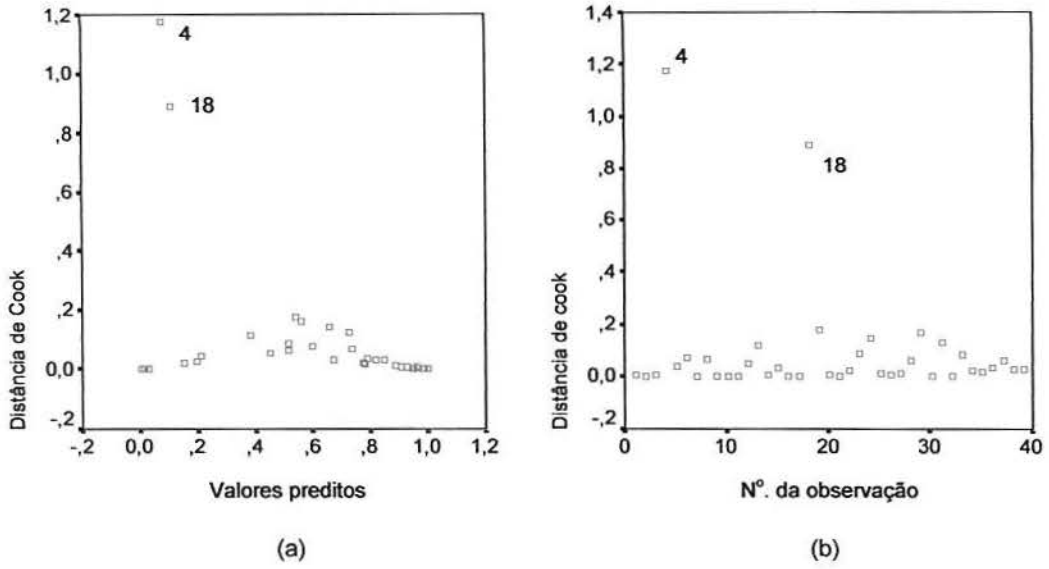


Figura 3.3: Gráfico dos Resíduos studentizados e Resíduos deviance[2]

É interessante observar que as observações #4 e #18 não apresentam valores altos de leverage, indicando que possivelmente não são pontos de alavanca.

Quando se avalia o gráfico da Distância de Cook versus os valores preditos (figura 3.4), pode-se perceber que as observações #4 e #18 estão afastadas do restante do conjunto de dados.



(a) (b)

Figura 3.4: Gráfico da Distância de Cook

Essas duas observações (#4 e #18) se apresentam distantes do restante das observações também nos gráficos do Deviance[2], dos resíduos studentizados e dos DFBETAS (figuras 3.5a-3.5c).

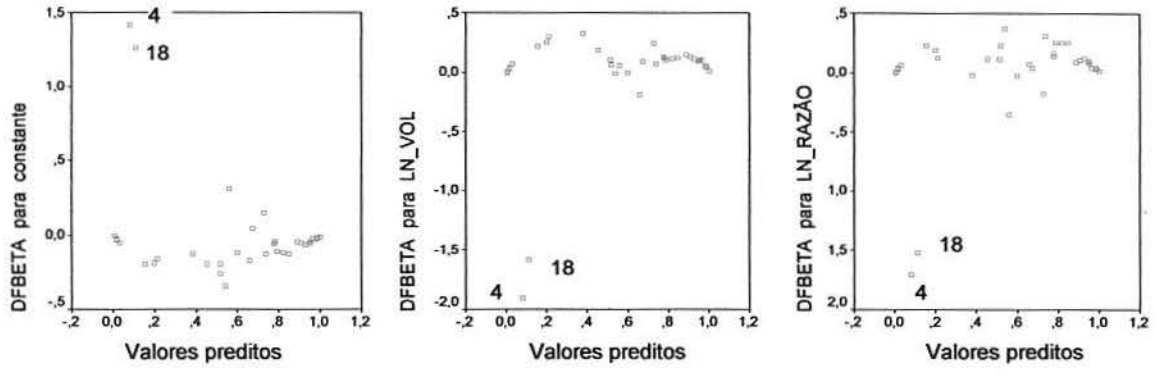


Figura 3.5: Gráfico dos DFBETAS para constante, ln(volume) e ln(razão)

Ao eliminar a observação #4 obtém-se as novas estimativas para os parâmetros do modelo: $\hat{\beta}_0 = -5{,}205$, $\hat{\beta}_1 = 8{,}465$ e $\hat{\beta}_2 = 7{,}463$. As novas estimativas representam modificações de 81%, 63,4% e 63,6% respectivamente.

4. APLICAÇÃO

Neste capítulo, serão analisados os dados da primeira e terceira etapa de um estudo de coorte que incluiu os recém-nascidos no município de Ribeirão Preto no período de 1º de junho de 1978 a 31 de maio de 1979. A terceira etapa do estudo ocorreu em 1996/97, quando os indivíduos tinham 18 anos de idade, com o objetivo de estudar as condições de vida e saúde, inclusive o crescimento, em função de variáveis estudadas ao nascer. Por ter sido realizada no momento do alistamento no serviço militar, as conclusões desta fase da pesquisa restringiram-se aos indivíduos do sexo masculino.

O objetivo do estudo é verificar quais seriam os fatores de risco para o sobrepeso ou obesidade no início da vida adulta. A variável resposta é o sobrepeso ou obesidade definido a partir do índice de massa corporal (IMC). O IMC foi calculado dividindo o peso em quilogramas pela altura ao quadrado, em metros (Keys et al, 1972). Foram considerados com sobrepeso ou obesidade aqueles indivíduos que apresentaram IMC$\geq$25,0. Como possíveis fatores de risco foram considerados: a classe social da mãe, o peso ao nascer, o comprimento ao nascer, a idade gestacional e o grau de instrução atual do indivíduo.

Dos 2083 indivíduos que foram reavaliados, somente 1209 entraram na análise, pois os demais não tinham informação referente ao seu peso ou altura atual, não sendo possível obter seu IMC.

4.1 Análise Univariada

Pode-se observar a distribuição do índice de massa corporal na figura 4.1. A distribuição amostral do IMC é assimétrica à direita. Isso mostra que há poucas pessoas com relação muito alta entre peso e altura.

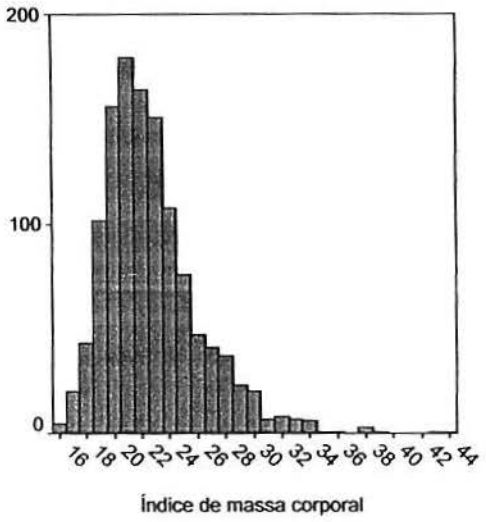


Figura 4.1: Distribuição dos indivíduos aos 18 anos segundo a variável índice da massa corporal.

Para ser utilizada na regressão logística, o IMC foi categorizado em sobrepeso ou obesidade e peso normal. Foram considerados com sobrepeso ou obesidade aqueles que apresentaram IMC igual ou superior a 25, que corresponde ao percentil 80. A distribuição desta nova variável é apresentada na tabela 4.1.

Tabela 4.1 - Distribuição empírica dos indivíduos aos 18 anos segundo índice de massa corporal.

Classificação	n	%
Sobrepeso ou obesidade	240	19,9
Normal	969	80,1
TOTAL	1209	100,0

Pode-se observar que a maioria dos jovens adultos está na faixa de peso normal. Apenas 19,9% deles têm sobrepeso ou obesidade. Pode-se verificar que existe um grande percentual de casos ignorados para o índice de massa corporal. Isso se deve basicamente à falta de informação sobre o peso atual dos indivíduos em estudo.

As variáveis relacionadas a dados perinatais dos adultos jovens são apresentadas na tabela 4.2.

Tabela 4.2 - Dados perinatais dos adultos jovens

Variável	n	média	desvio
Idade gestacional (em semanas)	956	39,3	1,88
Peso ao nascer (em Kg)	1209	3,3	0,50
Comprimento ao nascer (em cm)	1097	49,5	2,11

Em relação à variável idade gestacional pode-se perceber que o grupo estudado se apresenta, em média, dentro do adequado (idade gestacional maior ou igual a 37 semanas). É interessante ressaltar também que essa variável apresenta um número menor de indivíduos estudados. O peso e o comprimento ao nascer também se apresentam, em média, próximos do considerado adequado, 3,3Kg e 50cm, respectivamente.

Tabela 4.3 - Distribuição dos adultos jovens segundo classe social e escolaridade.

Variável	n	%	% válido
Classe social			
Classe baixa	213	17,6	18,7
Classe média	733	60,6	64,2
Classe alta	196	16,2	17,2
Ignorado	67	5,5	
Escolaridade			
Menos de 8 anos	259	21,4	21,5
8-10 anos	502	41,5	41,6
11 anos ou mais	446	36,9	37,0
Ignorado	2	0,2	

A classe social e o grau de escolaridade dos indivíduos em estudo são apresentados na tabela 4.3. Em relação à classe social, pode-se verificar que mais da metade dos adultos jovens pertencia à classe média. Sobre o grau de escolaridade é interessante notar que a há um número maior (41,6%) de indivíduos que tem entre 8 e 10 anos de estudo.

4.2 Análise Bivariada

Nesta seção serão apresentadas as análises da relação entre cada um dos fatores de risco e a variável dependente, sobrepeso ou obesidade. Esta relação será verificada através de regressão logística. A seguir serão apresentados os principais resultados dessa análise.

Para as regressões logísticas que envolviam variáveis com mais de duas categorias foi utilizada a parametrização "categoria de referência" do SPSS, que equivale à construção das variáveis indicadoras necessárias para esses casos.

Na tabela 4.4 são apresentados: o coeficiente estimado $\hat{\beta}$, seu erro padrão, a razão de chances, seu intervalo de 95% de confiança e a significância do Teste de Wald para cada fator de risco em relação ao sobrepeso ou obesidade.

As razões de chance cujo intervalo de confiança não contenha o valor um devem ser consideradas estatisticamente significativas.

Tabela 4.4 - Coeficiente estimado ($\hat{\beta}$), erro padrão ($\hat{EP}(\hat{\beta})$), Razão de chances (RC), intervalo de confiança (IC) e a significância do Teste de Wald referentes a regressões logísticas bivariadas dos fatores de risco em relação ao sobrepeso ou obesidade, 1997, Riberão Preto, SP.

Variável	n	$\hat{\beta}$	$\hat{EP}(\hat{\beta})$	Sig.*	RC	95%-IC para RC
Classe Social	1142					
Classe baixa	213	0,827	0,259	0,001	2,28	(1,38 -3,79)
Classe média	733	0,390	0,228	0,087	1,48	(0,95 -2,31)
Classe alta	196				1,00	
Peso ao nascer	1209	0,373	0,147	0,011	1,45	(1,09 - 1,94)
Comprimento ao nascer	1197	0,079	0,035	0,024	1,08	(1,01 - 1,16)
Idade Gestacional	956	0,015	0,043	0,730	1,02	(0,92 - 1,02)
Escolaridade do indivíduo	1207					
Menos de 8 anos	259	-0,115	0,202	0,468	0,89	(0,60 - 1,32)
8-10 anos	502	0,111	0,162	0,491	1,12	(0,81 - 1,53)
11 anos ou mais	446				1,00	

* significância para o Teste de Wald

Considerando a variável classe social, vemos que, em relação à categoria de referência, classe alta, a classe baixa apresenta risco significativo dos adultos jovens terem sobrepeso ou obesidade. Desse modo, podemos dizer que pertencer à classe alta é um fator de proteção ao sobrepeso ou obesidade. A razão de chances para classe baixa é 2,28, isto significa que um indivíduo que pertence à classe baixa tem 2,28 vezes a chance de vir a apresentar sobrepeso ou obesidade do que aquele que pertence à classe alta. Pertencer á classe média não oferece risco ao sobrepeso ou obesidade.

A chance de ocorrência de sobrepeso ou obesidade em jovens adultos aumenta 45% para cada aumento de 1 Kg no peso ao nascimento.

Em relação a variável comprimento ao nascer, pode-se dizer que quanto maior o

comprimento ao nascer, maior a chance da vir a apresentar sobrepeso ou obesidade no início da vida adulta, já que a ocorrência de sobrepeso ou obesidade em homens aumenta em 8% para cada aumento de 1cm no comprimento ao nascimento.

Quando avaliadas individualmente, nos modelos univariados, a idade gestacional e escolaridade do indivíduo não apresentaram relação estatisticamente significativa com o sobrepeso ou obesidade ao nível de significância de 5%.

4.3 Análise Multivariada: seleção do modelo

Para a construção do modelo multivariado serão considerados os critérios de seleção de modelos, apresentados na seção 2.5.

Inicia-se com o modelo:

$$g(x_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + constante$$

onde: $X_{1i} = \begin{cases} 1, \text{se o indivíduo i é da classe baixa} \\ 0, \text{caso contrário} \end{cases}$ $X_{2i} = \begin{cases} 1, \text{se o indivíduo i é da classe média} \\ 0, \text{caso contrário} \end{cases}$

X_3: peso ao nascer e X_4: comprimento ao nascer

Como, entre as variáveis peso ao nascer e comprimento ao nascer existe correlação bastante significativa (r=0,775; p<0,002), somente o peso ao nascer entrará no modelo, conjuntamente com a classe social. Assim, o modelo final é dado por:

$$g(x_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + constante$$

onde: $X_{1i} = \begin{cases} 1, \text{se o indivíduo i é da classe baixa} \\ 0, \text{caso contrário} \end{cases}$ $X_{2i} = \begin{cases} 1, \text{se o indivíduo i é da classe média} \\ 0, \text{caso contrário} \end{cases}$

e X_3: peso ao nascer

A razão de chances, o intervalo de confiança e a significância do teste de Wald para o modelo apresentado acima relativo aos fatores de risco para sobrepeso ou obesidade são apresentados na tabela 4.5.

Tabela 4.5 - Coeficiente estimado ($\hat{\beta}$), erro padrão ($\hat{EP}(\hat{\beta})$), Razão de chances (RC), intervalo de confiança (IC) e a significância do Teste de Wald referentes à regressão logística multivariavel dos fatores de risco em relação ao sobrepeso ou obesidade, 1997, Riberão Preto, SP.

Variável	n	$\hat{\beta}$	$\hat{EP}(\hat{\beta})$	RC	95%-IC para RC	Sig.*
Classe Social						
Classe baixa	213	0,771	0,260	2,16	(1,30 - 3,60)	0,003
Classe média	733	0,342	0,229	1,41	(0,90 - 2,21)	0,134
Classe alta	196			1,00		
Peso ao nascer	1142	0,395	0,156	1,49	(1,09 - 2,02)	0,011

* significância para o Teste de Wald

Para as próximas análises será considerado, então, o modelo ajustado:

$$\hat{g}(x) = -3{,}142 + 0{,}771 \times CLASS\_B + 0{,}342 \times CLASS\_M + 0{,}395 \times PESO\_NASC\,. \qquad (4.1)$$

4.4 Qualidade do ajuste do modelo

Para verificar a qualidade do ajuste para o modelo dado por (4.1), será utilizado o teste de Hosmer e Lemeshow. Esse teste, com 8 graus de liberdade, apresenta, para o valor de $\chi^2_{calculado}$=6,022, p = 0,645. Assim, não é possível rejeitar H_0, ao nível de significância de 5% e conclui-se que o modelo pode estar bem ajustado.

4.5 Diagnóstico do modelo

Todas as análises do diagnóstico do modelo de regressão logística multivariado, ajustado nesse exemplo, serão realizadas através de análise gráfica.

Através da opção *save* do SPSS, é possível obter os valores de h_j, r_{sj}, d, r_{tj}, $\Delta\hat{\beta}_j$ e DFBETAS. Entretanto, para obter os DFFITS, é necessário calculá-los através do procedimento *COMPUTE* utilizando os valores de h_j e de r_{tj}. A sintaxe é apresentada abaixo:

```
COMPUTE dffit = ABS(sre_1) * SQRT(lev_1 / (1-lev_1)) .
EXECUTE .
```

As variáveis sre_1 e lev_1 representam os resíduos studentizados e os valores de leverage (h_j), respectivamente.

Convém salientar que foi considerado que o SPSS calcula as medidas de diagnóstico através de observações individuais, ao invés de utilizar padrões de covariáveis.

4.5.1 Pontos de Alavanca (leverage values)

A figura 4.2 mostra o gráfico dos valores preditos versus os valores h_j. Pode-se observar que os pontos que se encontram mais distantes dos demais são aqueles acima da reta Y= 0,01. Esses pontos são apresentados na tabela 4.6.

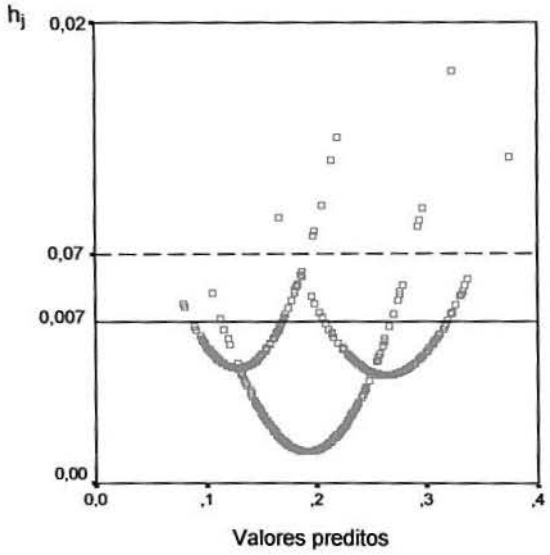


Figura 4.2: h_j versus valores preditos

Tabela 4.6- Características dos indivíduos que tiveram $h_j > 0,01$.

		Peso atual			
Classe	Peso ao nascer	normal	sobrepeso ou obesidade	Total	h_j
Classe baixa	1,850	1		1	0,012
	4,850		1	1	0,014
Classe média	4,780	1		1	0,011
	4,800	1		1	0,011
	4,830	2		2	0,012
	5,150		1	1	0,018
Classe alta	4,330	1		1	0,011
	4,350	1		1	0,011
	4,450	1		1	0,012
	4,600	1		1	0,014
	4,670	1		1	0,015
Total		10	2	12	

As observações apresentadas na tabela 4.6 são aquelas relativas aos indivíduos que apresentaram maiores pesos ao nascer e a um indivíduo que apresentou o menor peso ao nascer, o que concorda com o conceito de pontos de alavanca, que são aqueles

mais afastados do restante dos pontos de uma variável independente.

Se usarmos como base para obter pontos de alavanca os pontos $h_j \geq 2p/n = 0,007$, como foi sugerido por Belsey et al(1980), aumentaria o número de possíveis pontos de alavanca (ver anexo 1). As figuras 4.3a-4.3c mostram os valores de h_j versus peso ao nascer para cada classe social. Podemos observar que os pontos que estão relacionados a valores de h_j maiores do que 0,007 são justamente aqueles que apresentam valores de peso ao nascer muito altos (acima de 4,0 Kg), ou, muito baixos (menores do que 2,5 Kg). Esses valores são os geralmente usados para classificar o peso em baixo peso ao nascer (peso inferior a 2,5 Kg) e macrossomia fetal (peso superior a 4,0 Kg).

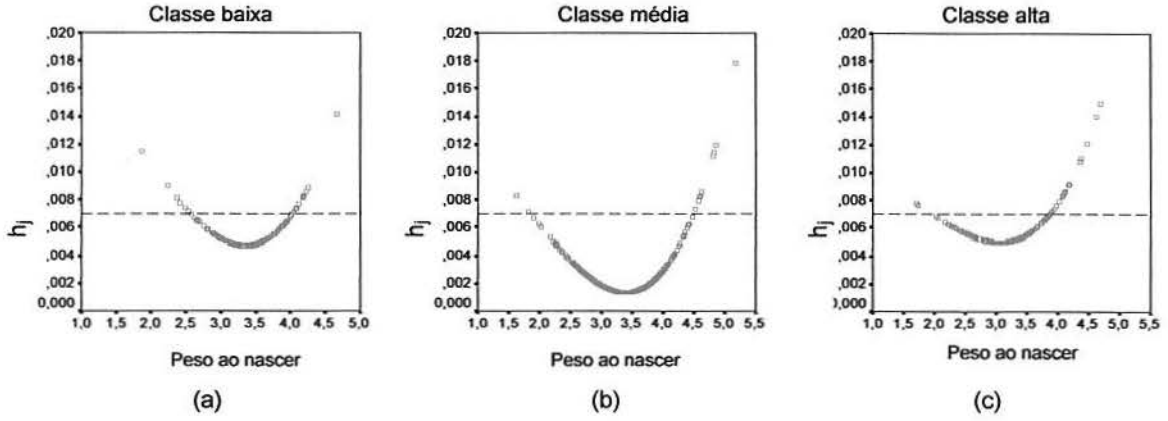


Figura 4.3: Gráficos de peso ao nascer versus h_j segundo classe social.

Eliminando a observação que apresentou maior valor de h_j (=0,018), obtemos as seguintes estimativas para os parâmetros: $\hat{\beta}_0 = -3,028$, $\hat{\beta}_1 = 0,774$, $\hat{\beta}_2 = 0,339$ e $\hat{\beta}_3 = 0,366$. Essas novas estimativas apresentaram uma variação de –3,6%, 0,4% , -0,9% e –7,3%, respectivamente, em relação ao modelo com a observação.

Eliminando a observação que apresentou valor de h_j (=0,00698), ou seja o mais alto abaixo do ponto de corte 0,007, obtemos as seguintes estimativas para os parâmetros: $\hat{\beta}_0 = -3,109$, $\hat{\beta}_1 = 0,776$, $\hat{\beta}_2 = 0,343$ e $\hat{\beta}_3 = 0,391$. Essas novas estimativas apresentaram

uma variação de –1,1%, 0,6% , 0,3% e –1,0%, respectivamente, em relação ao modelo com a observação.

4.5.2 Resíduos

Nas figuras 4.4a e 4.4b, são apresentados gráficos para os resíduos studentidado e Deviance^2 versus os valores preditos através do modelo de regressão logística. Foram considerados para uma análise mais detalhada aquelas observações que apresentaram resíduo studentizado superior a 2, ou resíduo Deviance^2 superior a 4 que é aproximadamente igual ao valor de $\chi^2_{(1);\,0,95}$. Das 11 observações encontradas, 10 eram de casos de indivíduos com sobrepeso ou obesidade que pertenciam à classe alta, mas que não tinham peso ao nascer elevado (os pesos estavam entre 2,150 e 3,100Kg). A 11ª observação era de um indivíduo com sobrepeso ou obesidade da classe média, com peso ao nascer de 2,300Kg. (ver anexo 2).

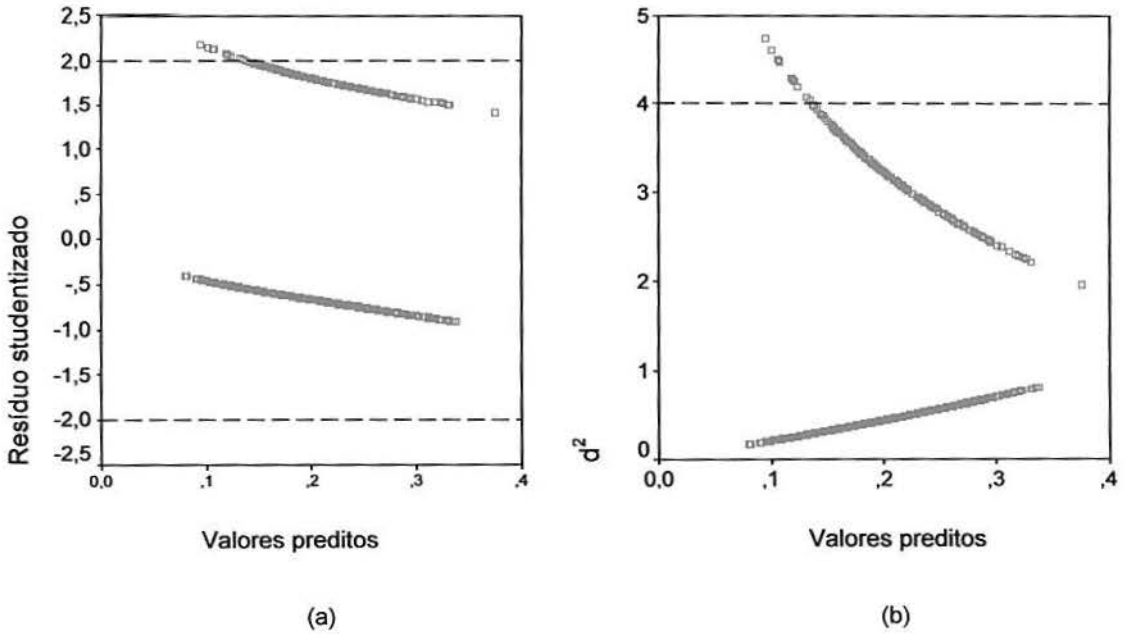


Figura 4.4: Resíduos versus valores preditos.

Eliminando a observação que apresentou maior valor de resíduo studentizado (r_t=2,18) e de Deviance^2 (d^2=4,74), obtemos as seguintes estimativas para os parâmetros:

$\hat{\beta}_0 = -3{,}251$, $\hat{\beta}_1 = 0{,}809$, $\hat{\beta}_2 = 0{,}380$ e $\hat{\beta}_3 = 0{,}421$. Essas novas estimativas apresentaram uma variação de –3,5%, 4,9% , 11,1% e 6,6%, respectivamente, em relação ao modelo com a observação.

4.5.3 Medidas de Influência

Podemos observar na figura 4.5 a relação entre os valores preditos e os valores da distância de Cook. Pode-se observar que todos os pontos estão abaixo da reta Y=0,07, ou seja, abaixo de Y=1 e não haveria nenhum ponto de influência. Entretanto, vale ressaltar que os 7 pontos que apresentam maiores valores da distância de Cook ($\Delta\hat{\beta} \geq 0{,}03845$) apresentam valor de $d^2 \geq 4{,}28$.

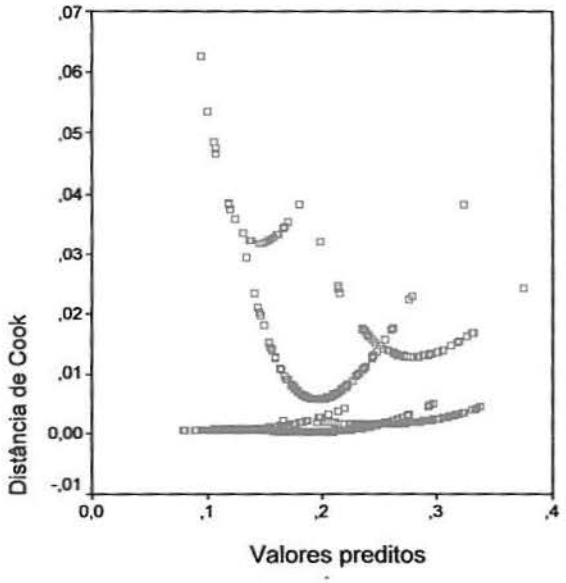


Figura 4.5: Distância de Cook versus valores preditos.

Nas figuras 4.6a-4.6d são apresentados os gráficos para os DFBETAS (para a constante, 6a, para o coeficiente da classe baixa, 6b, para o coeficiente da classe média, 6c e para o coeficiente do peso ao nascer, 6d) versus os valores preditos através do modelo de regressão.

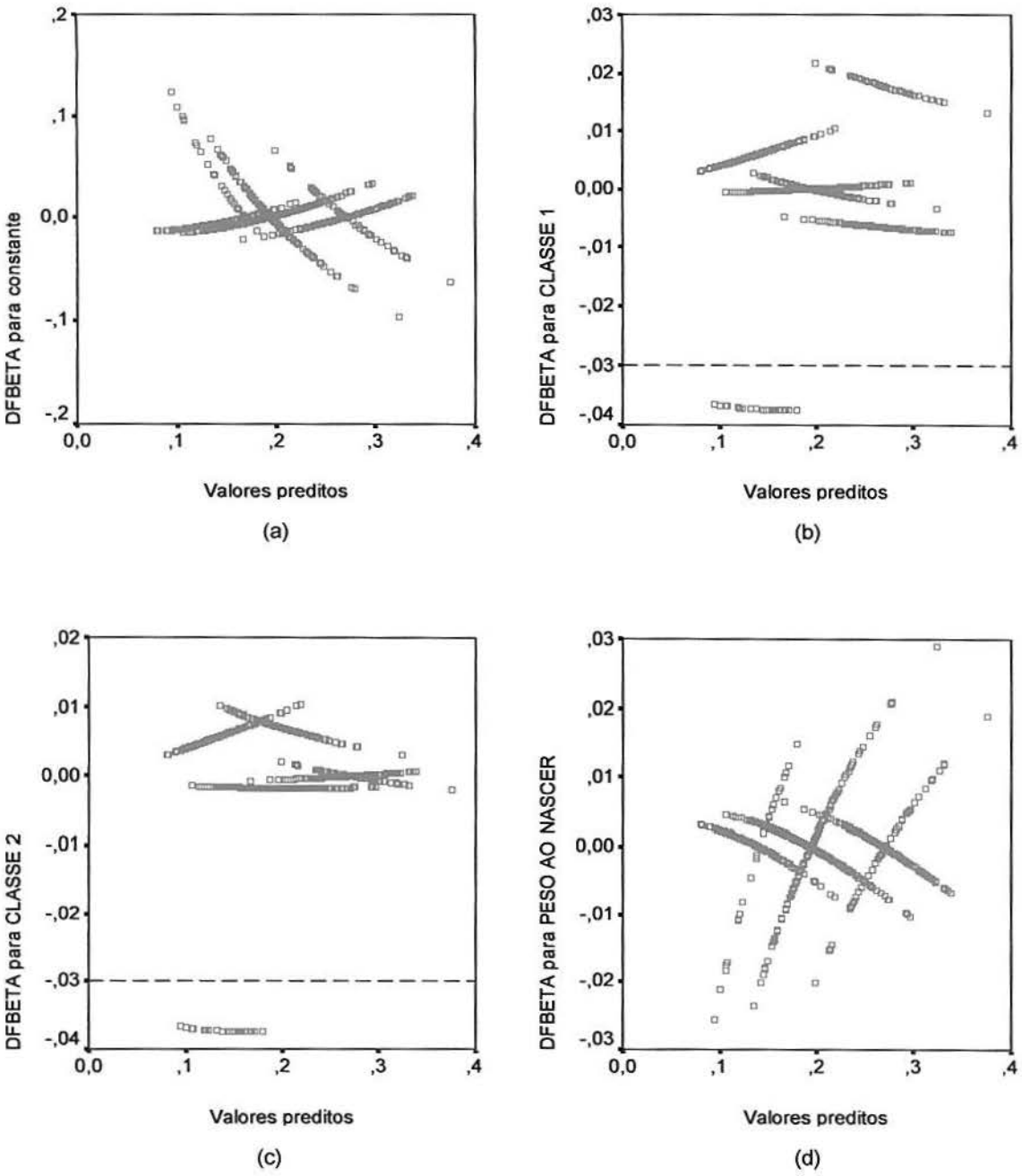


Figura 4.6: DFBETAS versus valores preditos.

Pode-se perceber claramente pelas figuras 4.6b e 4.6c que existem algumas observações que se distanciam das demais. Como sugerido por Neter et al (1996), usando como base para detectar pontos de influência valores absolutos de DFBETAS superiores a $1/\sqrt{n} = 1/\sqrt{1142} = 0{,}03$, essas observações se mantêm fora deste limite. É interessante verificar que essas observações são todas relativas a indivíduos que pertencem à classe

alta e apresentam sobrepeso ou obesidade, independente do peso ao nascer (ver anexo 3).

Após excluir a observação que apresentou maior valor em módulo para os DFBETAS da classe baixa e da classe média, obtemos as seguintes estimativas para os parâmetros: $\hat{\beta}_0 = -3{,}137$, $\hat{\beta}_1 = 0{,}808$, $\hat{\beta}_2 = 0{,}380$ e $\hat{\beta}_3 = 0{,}388$. Essas novas estimativas apresentaram uma variação de –0,2%, 4,8% , 11,1% e -1,8%, respectivamente, em relação ao modelo com a observação.

Na figura 4.7 são apresentados os valores de DFFITS versus os valores preditos. Para identificar possíveis pontos de influência foi traçada uma reta no ponto $2\sqrt{(p+1)/n} = 0{,}118$.

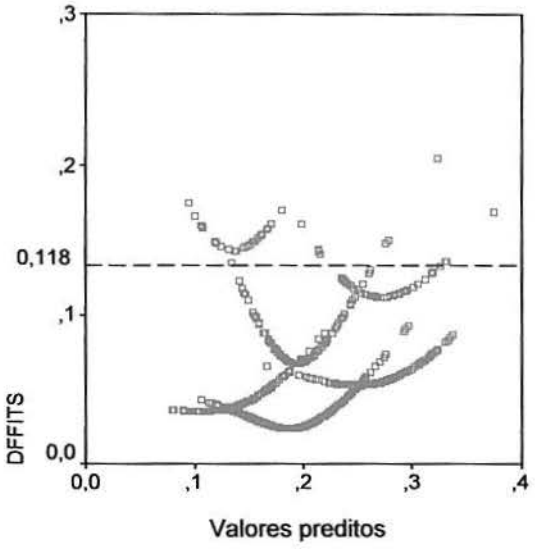


Figura 4.7: DFFITS versus valores preditos.

Assim como ocorreu com os DFBETAS, as observações que correspondem a altos DFFITS são de indivíduos da classe alta que apresentaram sobrepeso ou obesidade no início da vida adulta. Além dessas, também são apontados pelos DFFITS como pontos de influência, observações relativas a indivíduos que pesavam no nascimento menos de

3,070Kg ou mais de 3,800 da classe baixa e indivíduos com peso ao nascer igual ou inferior a 2,450Kg ou superior a 4,300Kg da classe média. (ver anexo 4).

Pode-se perceber que cada técnica apontou para diferentes conjuntos de observações. Alguns pontos como aqueles de indivíduos com sobrepeso ou obesidade cujo peso ao nascer era superior a 4Kg e pertencentes às classes baixa e alta, ou superior a 4,580 e da classe média foram considerados influentes e de alavanca.

É importante ressaltar que o impacto observado nas estimativas dos coeficientes com a retirada tanto de uma observação influente como de um ponto de alavanca foi muito pequeno. Talvez esse impacto pequeno nas estimativas dos parâmetros seja devido aos outros pontos influentes que permaneceram no modelo.

Para mostrar o efeito da retirada de mais de um ponto nas estimativas do modelo serão refeitas as análises de duas formas:

(1) Retirando-se os pontos com $d^2 \geq 4$

(2) Retirando-se as observações de indivíduos com peso ao nascer (PN) $\geq$ 4,5Kg.

As novas estimativas para β_i, após retirar os pontos com $d^2 \geq 4$, são apresentadas na tabela 4.7. Pode-se perceber uma mudança nos valores de significância para o teste de Wald (ver tabela 4.5). A classe média passou a ser significativa neste novo modelo. A constante foi estimada em -4,236. As novas estimativas para β_0, β_1, β_2 e β_3 apresentaram mudanças de -34,8%, 60,3%, 134,2% e 47,8% respectivamente.

Tabela 4.7 - Coeficiente estimado ($\hat{\beta}$),erro padrão ($\hat{EP}(\hat{\beta})$), Razão de chances (RC), intervalo de confiança (IC) e a significância do Teste de Wald referentes à regressão logística multivariada dos fatores de risco em relação ao sobrepeso ou obesidade sem as observações que apresentaram $d^2 \geq 4$.

Variável	n	$\hat{\beta}$	$\hat{EP}(\hat{\beta})$	RC	95%-IC para RC	Sig.*
Classe Social						
Classe baixa	213	1,236	0,299	3,44	(1,92 - 6,20)	< 0,001
Classe média	732	0,801	0,273	2,23	(1,31 - 3,80)	0,003
Classe alta	186			1,00		
Peso ao nascer	1142	0,584	0,163	1,79	(1,30 - 2,47)	< 0,001

* significância para o Teste de Wald

Considerando a retirada de indivíduos que pesavam mais de 4,5Kg ao nascer, obtém-se às estimativas apresentadas na tabela 4.8.

Tabela 4.8 - Coeficiente estimado ($\hat{\beta}$),erro padrão ($\hat{EP}(\hat{\beta})$), Razão de chances (RC), intervalo de confiança (IC) e a significância do Teste de Wald referentes à regressão logística multivariada dos fatores de risco em relação ao sobrepeso ou obesidade sem as observações que apresentaram PN≥4,5Kg.

Variável	n	$\hat{\beta}$	$\hat{EP}(\hat{\beta})$	RC	95%-IC para RC	Sig.*
Classe Social						
Classe baixa	212	0,739	0,261	2,09	(1,26 - 3,49)	2,09
Classe média	723	0,324	0,229	1,38	(0,88 - 2,17)	1,38
Classe alta	194			1,00		1,00
Peso ao nascer	1129	0,381	0,165	1,46	(1,06 - 2,02)	1,46

* significância para o Teste de Wald

A constante foi estimada em –3,058. As novas estimativas para β_0, β_1, β_2 e β_3 apresentaram mudança de 2,7%, -4,2%, -5,3% e -3,5% respectivamente em relação ás estimativas da tabela 4.5. Pode-se notar que com a exclusão das observações com

PN≥4,5Kg o impacto nas estimativas dos parâmetros foi bem maior. Essas observações haviam sido detectadas pelos DFBETAS e DFFITS como observações influentes.

Alguns autores, como Cordeiro e Lima Neto (2004), mencionam que a distância de Cook falha em detectar pontos de influência quando existem muitos destes pontos próximos e talvez por isso nenhum ponto de influência apontado pelos DFFITS e DFBETAS tenha sido detectado por essa medida. Para esses casos, é mais conveniente utilizar uma medida denominada de Influência Local, proposta por Cook (1986). Essa medida verifica a influência exercida simultaneamente por um conjunto de observações no modelo. Mais detalhes podem ser vistos em Paula (2004), Neter el al (1996) e Hossain e Islam (2003).

Entretanto, se todas as observações consideradas como pontos de alavanca ou pontos de influência fossem retiradas do modelo, algumas características presentes na população de indivíduos estudada desapareceriam.

5. CONSIDERAÇÕES FINAIS

É importante fazer o diagnóstico do modelo e verificar quais seriam as observações que poderiam interferir nas estimativas do modelo ajustado através da regressão logística. Entretanto deve se ter muito cuidado quanto à decisão sobre o que fazer com essas observações, que poderiam ser tanto pontos influentes como pontos de alavanca. É necessário lembrar que diferentes autores apontam para diferentes pontos de corte que definem quais são as observações influentes ou pontos de alavanca. Por isso, não é possível simplesmente excluí-las da análise. Cordeiro e Lima Neto (2004), afirmam que se um ponto de grande alavancagem for também influente ele não pode ser retirado da análise. Desse modo, deve-se estar familiarizado com a "natureza" dos dados de maneira a conhecer qual o comportamento dos diversos elementos na população e, quando isso não ocorrer, estar preparado para fazer uma análise mais detalhada da situação para decidir o quão importante é manter ou eliminar uma observação, ou conjunto de observações de determinado padrão ou modificar o modelo de regressão utilizado, de maneira que ele considere essas características dos dados tornando-o mais "eficiente". Segundo Paula (2004), a eliminação de observações deve ser a última opção. Outras soluções seriam: incluir ou eliminar um variável no modelo, fazer uma transformação nas variáveis ou utilizar modelos robustos. Mais detalhes sobre a utilização de modelos robustos podem ser vistos em Gervini (2005).

Os pacotes estatísticos como SPSS, SAS e STATA apresentam rotinas para calcular as medidas de diagnóstico de maneira bem acessível. Entretanto, esses programas não apresentavam a medida DFFITS, sendo necessário calculá-la através da fórmula apresentada na seção 3.3.2. Também é possível desenvolver rotinas para diagnóstico em regressão logística no R, programa de distribuição livre, mas que não foi

usado nesta monografia, ficando como sugestão para um futuro trabalho.

Deve-se levar em conta que na regressão logística a variável resposta é binária, e os erros só podem assumir os valores zero ou um, levando a uma dificuldade maior em analisar essas medidas utilizando as técnicas que inicialmente foram propostas para modelos lineares. Além disso, Pardoe e Cook (2002) afirmam que, muitas vezes, apesar de os resíduos apresentarem dependência em relação as variáveis preditoras, o modelo pode não estar mal ajustado, o que causa complicações na análise de gráficos na regressão logística.

Em última análise, as técnicas de diagnóstico são análises exploratórias dos dados. Por isso, é importante ter bom senso antes de chegar a conclusões sobre os dados ou sobre o modelo adotado.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. AGRESTI, A. (1990). **Categorical data analysis**. John Wiley & sons, New York

2. BAGLEY, S.C.; WHITE, H. & GOLOMB, B.A. (2001). **Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain**. *Journal of Clinical Epidemiology,* 54, 979–985

3. BELSLEY, D.A.; KUH, E. & WELSH, R.E. (1980). **Regression Diagnostics: identifying influential data and sources of collinearity.** John Wiley, New York.

4. COLLET, D. (1991). **Modelling Binary Data**. Chapman and Hall, London.

5. COLLET, D. (1994). **Modelling Survival Data in Medical Research**. Chapman & Hall, London.

6. CORDEIRO, G.M. & LIMA NETO, E.A. (2004) **Modelos Paramétricos**. 16°. SINAPE.

7. COOK, R.D. (1977). **Deletion of Influential Observations in Linear Regression.** *Technometrics,* 19, 15-18.

8. COOK, R.D. & WEINSBERG, S (1982). **Residual and influence in regression**. Chapman & Hall, London.

9. COOK, R.D., and WEISBERG, S. (1997), **Graphs for Assessing the Adequacy of Regression Models.** *Journal of the American Statistical Association,* 92, 490-499.

10. COX, D.R. (1958). **The regression analysis of binary sequences models (with discussion).** *Journal of the Royal Statistical Society,* B, 20, 215–242.

11. COX, D.R. & SNELL, E.J. (1968), **A General Definition of Residuals**, *Journal of the Royal Statistical Society, B,* 30, 248-275.

12. FISHER, R. & YATES, F. (1938) **Statistical Tables for Biological, Agricultural, and Medical Research.** Oliver & Boyd, Edinburgh, 3rd Ed.

13. GERVINI D. (2005). **Robust adaptative estimators for binary regression models.** *Journal of Statistical Planning and Inference,*131:297-311.

14. HAUK, W.W. & DONNER, A. (1977). **Wald's test as applied to hypotesis in logistic**

analysis. *Journal of the American Statistical Association*;72, 851-853.

15. HOAGLIN, D.C. & WELSH, R. (1978). **The Hat Matrix in Regression and ANOVA.** *The American Statistician,* 32, 17-22.

16. HOSMER, D.W. & LEMESHOW, S. (1980). **A goodness-of-fit test for the multiple logistic regression model.** *Communications on Statistics*, A10, 1043-1069

17. HOSMER, D.W. & LEMESHOW, S. (1989). **Applied logistic regression.** John Wiley & sons, New York.

18. HOSSAIN, M. & ISLAM, M. A. (2003). **Application of Local Influence Diagnostics to the Linear Logistic Regression Models**. *Dhaka University Journal of Science*, 51, 269-278.

19. KEYS, D.P.; FIDANZA, F.; KCARVONEN, M.J.; KIMURA, N. & TAYLOR, H.K. (1972) **Indices of relative weight and obesity**. Journal of chronic diseases., n. 25, p. 329-343.

20. LANDWEHR, J.M., PREGIBON, D. & SHOEMAKER, A.C. (1984). **Graphical methods for assessing logistic regression models (with discussion).** *Journal of the American Statistical Association*, 79, 61–71.

21. MENARD, S. (1995). **Applied Logistic Regression Analysis.** Sage University Paper series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

22. MIAZAKI, E.S. & STANGENHAUS G. (1994) **Métodos para detecção de dados atípicos** 11°. SINAPE.

23. NETER, J.; KUTNER, M.H.; NACHTSHEIM, C.J. & WASSERMAN, W. (1996) **Applied linear statistical models.** 3a. Edição. Irwin, Illinois.

24. PARDOE, I. & COOK, D.R. (2002). **A Graphical Method for Assessing the Fit of a Logistic Regression Model.** *The American Statistician,* 56(4), 263–272.

25. PAULA, G.A. (2004). **Modelos de regressão com apoio computacional**. IME-USP

26. PREGIBON, D. (1981) **Logistic regression diagnostic.** *Ann Statist.* 9:705-724.

27. TRUETT, J., CORNFIELD, J. & KANNEL W. (1967). **A multivariate analysis of the risk of coronary hart disease in Framingham.** *Journal of Chronic Diseases*, 20, 511-524.

ANEXOS

ANEXO 1

Quadro A1: Características dos indivíduos que apresentaram valores de $h_j > 0,007$.

CLASSE BAIXA			CLASSE MÉDIA			CLASSE ALTA		
PESO AO NASCER	PESO ATUAL		PESO AO NASCER	PESO ATUAL		PESO AO NASCER	PESO ATUAL	
	normal	sobrepeso ou obesidade		normal	sobrepeso ou obesidade		normal	sobrepeso ou obesidade
1,85	1		1,60	1		1,68	1	
2,21	1		1,80	1		1,72	1	
2,35	1		4,50	1		3,85	1	
2,40		1	4,55	1		3,87	1	
2,47	1		4,57	1		3,88	1	1
2,51	1		4,58		1	3,91	1	
4,02	1	1	4,60		1	3,95	2	
4,05	1		4,78	1		4,00	1	
4,06	1		4,80	1		4,05		1
4,10		3	4,83	2		4,06	1	
4,15	1	1	5,15		1	4,08	1	
4,16		1				4,09	1	
4,17	1					4,10	1	
4,20	2					4,15	2	
4,23	1					4,16	1	
4,65		1				4,33	1	
						4,35	1	
						4,45	1	
						4,60	1	
						4,67	1	
Total	13	8	Total	9	3	Total	21	2

ANEXO 2

Quadro A2: Características dos indivíduos que apresentaram valores de $d^2 > 4$.*

	PESO AO NASCER	n
CLASSE MÉDIA	2,30	1
CLASSE ALTA	2,15	1
	2,34	1
	2,47	1
	2,50	1
	2,52	1
	2,80	1
	2,81	1
	2,85	1
	2,93	1
	3,10	1

* Todos apresentam sobrepeso ou obesidade

ANEXO 3

Quadro A3: Peso ao nascer dos 27 indivíduos que apresentaram valores de DFBETAS>0,030.*

PESO AO NASCER	n	PESO AO NASCER	n
2,15	1	3,45	1
2,34	1	3,46	1
2,47	1	3,51	1
2,5	1	3,52	2
2,52	1	3,57	1
2,80	1	3,60	1
2,81	1	3,65	1
2,85	1	3,70	1
2,93	1	3,72	1
3,10	1	3,80	1
3,23	1	3,83	1
3,25	1	3,88	1
3,40	1	4,05	1

* Todos apresentam sobrepeso ou obesidade e pertencem à classe alta.

ANEXO 4

Quadro A4: Características dos indivíduos que apresentaram valores de DFFITS > 0,118.*

	CLASSE				CLASSE		
PESO AO NASCER	Baixa	Média	Alta	PESO AO NASCER	Baixa	Média	Alta
2,15			1	3,46			1
2,30		1		3,51			1
2,34			1	3,52			2
2,40	1			3,57			1
2,45		2		3,60			1
2,47			1	3,65			1
2,50			1	3,70			1
2,52			1	3,72			1
2,63	1			3,80	1		1
2,64	1			3,83			1
2,67	1			3,85	3		
2,80			1	3,88			1
2,81			1	3,93	1		
2,85			1	4,00	1		
2,93			1	4,02	1		
2,94	2			4,05			1
2,95	1			4,10	3		
2,96	1			4,15	1		
2,98	1			4,16	1		
3,00	2			4,30		1	
3,03	1			4,37		1	
3,07	1			4,39		1	
3,10			1	4,58		1	
3,23			1	4,60		1	
3,25			1	4,65	1		
3,40			1	5,15		1	
3,45			1				

* Todos apresentam sobrepeso ou obesidade