Análise de diagnóstico
em regressão logística

Cecília Aparecida Vaiano Farhat

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE
MESTRE EM ESTATÍSTICA

Área de Concentração : Estatística

Orientadora : Profa. Dra. Silvia Nagib Elian

- São Paulo, fevereiro de 2003 -

Análise de diagnóstico em regressão logística

Este exemplar corresponde à redação
final da dissertação devidamente
corrigida e defendida por
Cecília Aparecida Vaiano Farhat
e aprovada pela comissão julgadora.

São Paulo, fevereiro de 2003.

Banca examinadora:

- Profa. Dra. Silvia Nagib Elían (Orientadora) - IME-USP
- Prof. Dr. Gilberto Alvarenga Paula - IME-USP
- Profa. Dra. Arminda Lucia Siqueira - UFMG

Aos meus filhos.

Agradecimentos

Agradeço primeiramente a minha orientadora, professora Doutora Silvia Nagib Elian, pela amizade, pelas sugestões, dedicação e eficiência na orientação desta dissertação.

Agradeço aos professores do departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Paulo com quem tive a satisfação de estudar.

Agradeço também ao professor Dr. José Eluf Neto, do Departamento de Medicina Preventiva da Faculdade de Medicina da Universidade de São Paulo, com quem tive a oportunidade de contar, pelo conjunto de dados reais cedidos gentilmente.

A todos vocês meu muito obrigada.

Resumo

Este trabalho tem por objetivo apresentar a análise de ajuste de modelos de regressão logística. Inicialmente, dedicamo-nos ao estudo de técnicas que propiciam uma análise global do ajuste do modelo. Numa etapa seguinte, introduzimos as principais medidas de diagnóstico para a detecção de pontos discrepantes e um estudo dos principais gráficos para esse tipo de análise. Concluimos o trabalho com a aplicação das técnicas descritas a um conjunto de dados reais.

Abstract

In this dissertation our objective is to present the analysis of the goodness-of-fit of the Logistic Regression Models. First, we describe some techniques to make a global analysis of the fit. Further, the main diagnostics measures in Logistic Regression are introduced and discussed. The last part of the work is devoted to show graphical methods of diagnostics. Finally, we conclude the dissertation with a real data analysis.

Conteúdo

1	Introdução	1
2	Modelos de Regressão Logística	3
2.1	Modelo de Regressão Logística Simples	3
2.1.1	Teste de significância dos parâmetros e teste do ajuste do modelo	10
2.2	Modelo de Regressão Logística Múltipla	14
2.2.1	Estimação dos Parâmetros do Modelo de Regressão Logística Múltipla	16
2.3	Testes de significância dos parâmetros do modelo	18
3	Avaliação do Ajuste do Modelo	23
3.1	Introdução	23
3.2	Testes de ajuste do modelo	25
3.3	Exemplo	32

3.4	Considerações Finais	39
4	Medidas de Diagnóstico em Regressão Logística	43
4.1	Introdução	43
4.2	Medidas de diagnóstico	44
4.3	Cálculo das medidas de diagnóstico para o exemplo da Seção	
3.3		55
5	Análise Gráfica para diagnóstico	61
5.1	Introdução	61
5.2	Gráficos Específicos	62
5.3	Gráficos para o exemplo	75
6	Exemplo com dados reais	82
7	Considerações Finais e Sugestões	94
A	Apêndice	96
	Bibliografia	113

Capítulo 1

Introdução

O uso de modelagem estatística, especialmente a regressão logística, tem se tornado uma prática quase rotineira nos dias de hoje. Aparentemente, isso se relaciona com a popularidade de pacotes estatísticos de utilização muitas vezes fácil e rápida, que permitem o manejo de múltiplas variáveis simultaneamente.

A Análise de regressão logística apresenta características próprias por tratar da situação em que a variável resposta é discreta e a distribuição dos erros não é normal, não se enquadrando portanto na análise de Regressão Linear tradicional.

Neste caso, a variável de interesse (Y), habitualmente designada por variável resposta, possui somente dois valores possíveis, em que atribui-se por conveniência o código 1 geralmente ao acontecimento de interesse (“sucesso”), e 0 ao outro valor, ligado ao acontecimento complementar (“fracasso”). As demais variáveis envolvidas no experimento são chamadas de variáveis

explicativas, podendo ou não ser binárias.

- Neste trabalho nos dedicaremos à Análise de diagnóstico em regressão logística, que tem como objetivo a obtenção de um modelo adequado para descrever a relação entre Y e as variáveis explicativas. Para isso, iniciaremos com uma apresentação do modelo de regressão logística no Capítulo 2.

Abordaremos no Capítulo 3 procedimentos para analisar a qualidade do ajuste global de um modelo inicialmente suposto adequado. As estatísticas de diagnóstico apresentadas no Capítulo 4 possibilitarão verificar a existência de pontos atípicos, que podem prejudicar o ajuste ou ter uma influência exagerada nas características apresentadas pelo modelo final. Esta análise será complementada com a construção de gráficos, presentes no Capítulo 5. Finalmente, no Capítulo 6, encerramos o trabalho com um exemplo envolvendo um conjunto de dados reais, no qual aplicamos grande parte das técnicas descritas.

Capítulo 2

Modelos de Regressão Logística

2.1 Modelo de Regressão Logística Simples

Regressão Logística é uma ferramenta de análise estatística paramétrica que está se tornando um dos principais métodos de modelagem de dados com resposta binária, isto é, que admitem apenas dois possíveis resultados. Comumente é chamado *sucesso* o resultado mais importante da resposta que representaria a presença de uma particular característica de interesse. Assim esta técnica é utilizada diante de uma situação em que temos uma variável resposta dicotômica (binária) e desejamos verificar sua relação com uma ou mais variáveis explicativas que podem ser qualitativas ou quantitativas. Mesmo quando a resposta não é originalmente do tipo binária, caso haja interesse, podemos dicotomizá-la, de modo que a probabilidade de sucesso possa ser modelada através da regressão logística.

Embora a Análise de Regressão Logística seja conhecida desde os anos

50, foi através de Cox (1970) (vide também Cox e Snell, 1989) que a mesma tornou-se popular entre os usuários de Estatística.

A variável resposta binária, que toma valores 0 e 1, será denotada por Y e é definida como:

$$Y = \begin{cases} 1 & \text{se o elemento possui a característica de interesse;} \\ 0 & \text{caso contrário.} \end{cases}$$

A variável explicativa, que consideraremos não aleatória, será representada por X .

Da forma como foi definida, podemos notar que a variável Y tem distribuição de Bernoulli com probabilidade de sucesso $\pi(x)$, em que $\pi(x) = P(Y = 1|X = x)$ é a probabilidade desconhecida da existência da característica de interesse, associada ao valor x da variável independente (ou dado o valor de X). O modelo de regressão logística simples, também chamado de função resposta logística admite que:

$$\pi(x) = P(Y = 1|x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (2.1)$$

ou de uma forma equivalente a (2.1), é dado por:

$$P(Y = 1|x) = [1 + e^{-\beta_0 - \beta_1 x}]^{-1}$$

em que, β_0 e β_1 são parâmetros desconhecidos.

Sejam os pares: $\{(Y_1, x_1), (Y_2, x_2), \dots, (Y_n, x_n)\}$, em que, $(Y_1, Y_2, \dots, Y_n)$ são variáveis aleatórias independentes tais que Y_i tem distribuição de *Bernoulli* com probabilidade de sucesso $\pi(x)$ e x_i é o valor da variável explicativa associada ao elemento cujo o valor da resposta é Y_i . Nesse contexto, $(x_1, x_2, \dots, x_n)$

são n constantes conhecidas, logo, X não é variável aleatória. Temos então que cada Y_i tem a seguinte distribuição de probabilidades:

Y_i	probabilidade
1	$P(Y_i = 1) = \pi(x_i)$
0	$P(Y_i = 0) = 1 - \pi(x_i)$

e, pela definição de esperança de variáveis aleatórias discretas,

$$E(Y_i|X = x_i) = \pi(x_i), \quad (2.2)$$

para $i = 1, 2, \dots, n$, ou seja, o valor esperado irá sempre representar a probabilidade de $Y_i = 1$, $\pi(x_i)$, já definida em (2.1). Como

$$\pi(x_i) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}}, \quad (2.3)$$

em que $g(x_i) = \beta_0 + \beta_1 x_i$, temos que

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} (1 + e^{g(x_i)})$$

e como consequência,

$$\ln \frac{\pi(x_i)}{1 - \pi(x_i)} = g(x_i). \quad (2.4)$$

A quantidade $\ln \frac{\pi(x_i)}{1 - \pi(x_i)}$ é denominada transformação logito, e $g(x_i)$ é a função resposta logito. Podemos notar que quando x está variando de $-\infty$ a $+\infty$, $g(x_i)$ apresenta a mesma variação.

Mostramos que $\pi(x_i)$ é a média, dada em (2.2). A ligação entre a média e o preditor linear $g(x_i)$, não é a identidade como nos modelos de regressão linear. Portanto, fazendo a transformação logito, encontramos uma função

da média que é igual a $g(x_i)$ e que é uma função linear nos parâmetros. Sendo assim, garante-se a concavidade da função de verossimilhança, que será definida posteriormente e como consequência, muitos resultados assintóticos são obtidos mais facilmente. Além disso, a concavidade da função de verossimilhança implica na unicidade da estimativa de máxima verossimilhança de β_0 e β_1 , quando essa existe (Paula (1999), [pg.9]).

Funções de Ligação são oriundas da Teoria de Modelos Lineares Generalizados dos quais os Modelos de Regressão Logística são um caso particular.

A análise de regressão logística é regida pelas mesmas linhas da análise de regressão linear, apesar de seus modelos não terem as suposições de normalidade e variância constante.

O que distingue o modelo de regressão logística do modelo de regressão linear é que a variável resposta na regressão logística é binária. Isto se reflete na escolha do modelo paramétrico e nas suposições. Adicionalmente os resultados inferenciais da regressão logística são assintóticos, enquanto na regressão normal linear esses resultados são exatos. Salvo estas diferenças, os métodos empregados na análise de regressão logística seguem os mesmos princípios dos utilizados na regressão linear. Através da Tabela 2.1 ilustramos essas diferenças entre os modelos regressão linear e de regressão logística.

Tabela 2.1 - Modelos de Regressão Linear Simples e Regressão Logística Simples.

Regressão Linear Simples	Regressão Logística Simples:
$E(Y X = x) = \beta_0 + \beta_1 x$	$E(Y X = x) = \pi(x)$
$-\infty < E(Y X = x) < +\infty$	$0 \leq E(Y X = x) \leq 1$
$y = \beta_0 + \beta_1 x + \epsilon$	$y = \pi(x) + \epsilon$
$\epsilon \sim N(0, \sigma^2)$ e	$\epsilon = \begin{cases} (1 - \pi(x)) & \text{com probabilidade } \pi(x) \\ -\pi(x) & \text{com probabilidade } 1 - \pi(x). \end{cases}$
$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$	$E(\epsilon) = 0, Var(\epsilon) = \pi(x)(1 - \pi(x))$
	$y \sim Bernoulli(\pi(x))$.

Observamos então que, no modelo de regressão logística simples:

1. A distribuição dos erros não é normal, pois, para uma variável binária, cada erro $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$ irá assumir apenas dois valores. No Modelo de Regressão Linear, considera-se os erros normalmente distribuídos.
2. As variâncias dos erros não são constantes para todo x , não valendo portanto a suposição de homocedasticidade.
3. A quantidade $E(Y_i)$ é uma probabilidade, sendo assim um número pertencente ao intervalo $[0, 1]$.

Nos Modelos de Regressão Logística, os parâmetros β_0 e β_1 são usualmente estimados por máxima verossimilhança.

Como cada Y_i é uma variável aleatória com distribuição de Bernoulli, em que

$$P(Y_i = 1|x_i) = \pi(x_i) \text{ e } P(Y_i = 0|x_i) = 1 - \pi(x_i),$$

podemos representar essa distribuição de probabilidade da seguinte maneira:

$$P(Y_i = y_i|x_i) = f(y_i|x_i) = [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}, \quad y_i = 0, 1, \quad i = 1, \dots, n.$$

Notemos que $f(1|x_i) = \pi(x_i)$ e $f(0|x_i) = 1 - \pi(x_i)$ representam a probabilidade que $Y_i = 1$ ou 0 , dado que $X = x_i$.

Portanto, como as observações Y_i são independentes, a função distribuição de probabilidades conjunta de $Y_1, \dots, Y_n$ será:

$$g(y_1, y_2, \dots, y_n) = \prod_{i=1}^n f(y_i|x_i) = \prod_{i=1}^n [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}.$$

Temos que $E(Y_i|x_i) = \pi(x_i)$ é a função de β_0 e β_1 , definida em (2.1). Então a função de verossimilhança de β_0 e β_1 com base na amostra $(x_1, y_1), \dots, (x_n, y_n)$ pode ser expressa como:

$$L(\beta_0, \beta_1) = \prod_{i=1}^n f(y_i|x_i) = \prod_{i=1}^n [\pi(x_i)]^{y_i}[(1 - \pi(x_i))]^{1-y_i}, \quad y_i = 0, 1. \quad (2.5)$$

O logaritmo neperiano da função de verossimilhança é

$$\ln(L(\beta_0, \beta_1)) = l(\beta_0, \beta_1) = \sum_{i=1}^n [y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i))]. \quad (2.6)$$

Os estimadores de máxima verossimilhança $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros do modelo de regressão logística são aqueles que maximizam o logaritmo da função de verossimilhança dada em (2.6). Assim,

$$\frac{\partial(l(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^n \left[\frac{y_i}{\pi(x_i)} \Delta_i - \frac{(1 - y_i)}{1 - \pi(x_i)} \Delta_i \right],$$

em que

$$\Delta_i = \frac{\partial \pi(x_i)}{\partial \beta_0} = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} - \frac{e^{\beta_0 + \beta_1 x_i}}{(1 + e^{\beta_0 + \beta_1 x_i})^2} e^{\beta_0 + \beta_1 x_i} = \pi(x_i)[1 - \pi(x_i)].$$

Como consequência, a equação de verossimilhança $\frac{\partial(l(\beta_0, \beta_1))}{\partial \beta_0} = 0$ é equivalente a

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0.$$

Para β_1 ,

$$\frac{\partial(l(\beta_0, \beta_1))}{\partial \beta_1} = \sum_{i=1}^n \left[\frac{y_i}{\pi(x_i)} \Lambda_i - \frac{(1 - y_i)}{1 - \pi(x_i)} \Lambda_i \right],$$

em que

$$\Lambda_i = \frac{\partial \pi(x_i)}{\partial \beta_1} = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} x_i - \frac{e^{\beta_0 + \beta_1 x_i}}{[1 + e^{\beta_0 + \beta_1 x_i}]^2} e^{\beta_0 + \beta_1 x_i} x_i = x_i \pi(x_i)[1 - \pi(x_i)]$$

e $\frac{\partial(l(\beta_0, \beta_1))}{\partial \beta_1} = 0$ implica em

$$\sum_{i=1}^n x_i [y_i - \pi(x_i)] = 0.$$

Podemos verificar que essas equações não são lineares em β_0 e β_1 , e que não é possível obter a solução explícita desse sistema. Então, utiliza-se métodos iterativos para resolvê-los, determinando assim as estimativas de máxima verossimilhança $\hat{\beta}_0$ e $\hat{\beta}_1$. Em particular, as soluções dessas equações podem ser obtidas usando o processo iterativo de mínimos quadrados reponderados, que discutiremos posteriormente.

Obtidos os estimadores de máxima verossimilhança de β_0 e β_1 , pela propriedade da invariância, podemos obter o estimador de máxima verossimilhança de $\pi(x)$. Desta forma, uma vez encontradas as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$,

substituímos estes valores na função resposta logística, para obter o valor ajustado $\hat{\pi}(x_i)$ que é dado por $\hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$ e é a estimativa de máxima verossimilhança de $\pi(x_i)$, de acordo com o modelo (2.1).

Se utilizarmos a transformação logito dada por (2.4), podemos expressar o valor ajustado da função resposta logito por: $\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

2.1.1 Teste de significância dos parâmetros e teste do ajuste do modelo

O modelo (2.1) possui apenas dois parâmetros, β_0 e β_1 , pois admite que

$$\pi_i = P(Y = 1|X = x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}, i = 1, 2, \dots, n,$$

ou seja, que todos os $\pi(x_i)$ são funções desses parâmetros.

Muitas vezes, o conjunto de dados observado apresenta réplicas na variável independente X . Desta forma, dispomos de L ($L \leq n$) distintos valores de X , organizados da seguinte maneira:

- m_1 elementos na amostra com $X = x_1$,
- m_2 elementos na amostra com $X = x_2$,
- $\vdots$
- m_L elementos na amostra com $X = x_L$,

sendo que $\sum_{j=1}^L m_j = n$.

Nesse caso, o modelo geral, também denominado saturado será

$$P(Y = 1|X = x_j) = \pi(x_j), j = 1, 2, \dots, L,$$

em que os $\pi(x_j)$ são parâmetros não especificados com $0 \leq \pi(x_j) \leq 1$, de modo que o número de parâmetros é $L < n$.

Sob esse modelo, $\hat{\pi}(x_j) = \frac{y_j}{m_j}$, $j = 1, 2, \dots, L$, em que y_j é o número de valores de Y iguais a 1 dentre os m_j associados ao valor de $X = x_j$.

Assim, podemos construir um teste de ajustamento do modelo (2.1) se o compararmos com o modelo saturado. Esta comparação é feita através da estatística da razão de verossimilhanças, cuja expressão é:

$$D = -2\ln\left[\frac{L(\hat{\beta}_0, \hat{\beta}_1)}{L_0}\right] \quad (2.7)$$

ou

$$D = 2[\ln L_0 - \ln(L(\hat{\beta}_0, \hat{\beta}_1))].$$

em que L_0 é a função de verossimilhança do modelo saturado calculada em $\hat{\pi}_0(x)$, sendo $\hat{\pi}_0(x)$ o estimador de máxima verossimilhança de π , sob tal modelo e $L(\hat{\beta}_0, \hat{\beta}_1)$ é a função de verossimilhança (2.5) calculada nos estimadores de máxima verossimilhança de β_0 e β_1 . A forma matemática desta função será dada a seguir. Nestas condições, estaremos testando H_0 : o modelo é $\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$ contra H_a : o modelo é o saturado.

Com a notação introduzida, após cálculos verificamos que o logaritmo da função de verossimilhança do modelo (2.1) é

$$\ln L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{j=1}^L [y_j \ln(\hat{\pi}(x_j)) + (m_j - y_j) \ln(1 - \hat{\pi}(x_j))]$$

com

$$\hat{\pi}(x_j) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}.$$

Para o modelo saturado, lembrando que $\hat{\pi}_0(x_j) = \frac{y_j}{m_j}$, então o logaritmo de L_0 fica

$$\ln L_0 = \sum_{j=1}^L [y_j \ln(\frac{y_j}{m_j}) + (m_j - y_j) \ln(\frac{m_j - y_j}{m_j})].$$

Como consequência,

$$D = -2 \sum_{j=1}^L \{y_j \ln(\frac{m_j \hat{\pi}(x_j)}{y_j}) + (m_j - y_j) \ln(\frac{1 - \hat{\pi}(x_j)}{1 - \frac{y_j}{m_j}})\}$$

ou

$$D = 2 \sum_{j=1}^L \{y_j \ln(\frac{y_j}{m_j \hat{\pi}(x_j)}) + (m_j - y_j) \ln(\frac{m_j - y_j}{m_j - m_j \hat{\pi}(x_j)})\}, \quad (2.8)$$

em que $\hat{\pi}(x_j) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}$.

Destacamos que, para conjuntos em que $y_j = 0$ ou $y_j = m_j$, o correspondente termo de D é definido como $-2m_j \ln(1 - \hat{\pi}(x_j))$ quando $y_j = 0$ e $-2m_j \ln(\hat{\pi}(x_j))$ quando $y_j = m_j$.

Quando temos grandes amostras, sob a hipótese nula de ajuste do modelo de regressão logística, a estatística (2.7) tem aproximadamente distribuição qui-quadrado com $(L - 2)$ graus de liberdade ($D \sim \chi^2_{L-2}$).

A estatística D é chamada de desvio ou função desvio do modelo (2.1), e é definida como duas vezes a diferença entre os logaritmos das funções de verossimilhança maximizadas do modelo saturado e do modelo com parâmetros desconhecidos β_0 e β_1 . Essa estatística tem um papel importante, pois pode ser utilizada como uma medida de *qualidade* do ajuste do modelo (2.1),

quanto maior seu valor, pior é o ajuste de tal modelo. Compara-se seu valor observado com os percentis da distribuição qui-quadrado com $(L - 2)$ graus de liberdade (χ^2_{L-2}). Se tal valor for muito superior a $(L - 2)$, que é a média dessa distribuição, entende-se que há evidências de que o modelo não fornece um bom ajuste (Paula, (1999) [pg.16]). No caso binomial quando L é fixo e n_j tende a infinito para cada $j = 1, 2, \dots, L$, sob a hipótese de que o modelo é verdadeiro, D segue uma distribuição χ^2_{L-2} . Isso não vale quando n é grande e $n_j\pi_j(1 - \pi_j)$ permanece limitado.

Para verificar se realmente existe alguma relação funcional de π_i com a variável explicativa X , devemos, a partir de uma amostra aleatória, estimar os coeficientes β_0 e β_1 e testar a significância do coeficiente da variável independente (β_1) no modelo. Isso envolve a formulação das hipóteses estatísticas $H_0 : \beta_1 = 0$ contra $H_a : \beta_1 \neq 0$, com estatística do teste da razão de verossimilhança

$$G = D_1 - D_2, \quad (2.9)$$

em que D_1 é o desvio do modelo

$$\pi_i = \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}}, \quad (2.10)$$

sem a variável independente e D_2 é o desvio do modelo (2.1).

Sob H_0 , $G \sim \chi^2_1$. Fixado um nível de significância α , se a estatística G for maior que o quantil de ordem $1 - \alpha$ da distribuição qui-quadrado com 1 grau de liberdade, rejeitamos H_0 , isto é, concluímos que existe relação funcional da variável explicativa com a variável resposta ou há evidência de que a variável independente é significativa para o modelo. A manutenção desse modelo está sujeita a um teste de ajuste. Quando a comparação de

D com χ^2_{L-2} é um procedimento informal para verificação da qualidade do ajuste. Testes mais rigorosos serão apresentados no próximo capítulo.

Verifica-se também que 2.9 pode ser expressa como

$$G = -2\ln\left[\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0, \hat{\beta}_1)}\right],$$

em que $L(\hat{\beta}_0)$ é a função de verossimilhança do modelo (2.10) calculada no estimador de máxima verossimilhança de β_0 , sob esse modelo.

2.2 Modelo de Regressão Logística Múltipla

Nesta seção, vamos generalizar o modelo de regressão logística simples para o caso de mais de uma variável independente. Neste caso, admite-se que a probabilidade de sucesso de uma variável com resposta binária é função de um conjunto de variáveis explicativas (ou variáveis independentes). Então ao considerarmos variáveis independentes no modelo logístico, associado ao i -ésimo elemento amostral, temos o vetor de variáveis que denotaremos por $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{ip})'$ em que cada componente x_{ik} , $k = 1, 2, \dots, p$ e $i = 1, 2, \dots, n$, corresponde a um valor fixo conhecido de cada variável independente, com $x_{i0} = 1$. No modelo de regressão logística múltipla, a probabilidade de sucesso é dada por:

$$\pi_i = P(Y_i = 1 | \mathbf{X} = \mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})} \quad (2.11)$$

ou

$$\pi_i = P(Y_i = 1 | \mathbf{X} = \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})}, i = 1, 2, \dots, n$$

em que β_k é o k -ésimo parâmetro desconhecido associado à variável explicativa X_k , $k = 1, 2, \dots, p$ e

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

A expressão (2.11), que quando $p = 1$ se reduz a (2.1), define o chamado modelo de regressão logística múltipla que, por conveniência, denominaremos somente de modelo de regressão logística. Outra forma de apresentá-lo é:

$$g(\mathbf{x}_i) = \ln\left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (2.12)$$

sendo que $g(\mathbf{x})$ é o logito.

Nesse caso,

$$E(Y_i | \mathbf{X} = \mathbf{x}_i) = \pi(\mathbf{x}_i) = \frac{\exp(g(\mathbf{x}_i))}{1 + \exp(g(\mathbf{x}_i))},$$

para $i = 1, 2, \dots, n$.

Para uma particular situação, os dados observados numa amostra aleatória de n indivíduos são formados pela variável resposta Y_i (0 ou 1) e pelos valores de x_{ik} das variáveis explicativas, $k = 1, 2, \dots, p$ e $i = 1, 2, \dots, n$. Considerando as variáveis aleatórias Y_i , $i = 1, 2, \dots, n$, a função de verossimilhança associada ao modelo é definida por um produto de n distribuições de Bernoulli (π_i).

As variáveis preditoras ou variáveis explicativas podem ser quantitativas ou qualitativas e neste último caso, devem ser representadas por variáveis indicadoras ou variáveis *dummy*.

2.2.1 Estimação dos Parâmetros do Modelo de Regressão Logística Múltipla

Vamos assumir que temos uma amostra aleatória $(x_{i1}, x_{i2}, \dots, x_{ip}, Y_i)$, para $i = 1, 2, \dots, n$, sendo que $(x_{i1}, x_{i2}, \dots, x_{ip})$ são valores de variáveis não aleatórias $X_1, X_2, \dots, X_p$. O método de estimação usado neste caso é o mesmo que no modelo de regressão logística simples. A função de verossimilhança é idêntica à dada na equação (2.5), com uma pequena mudança em π_i , que está definido em (2.11). Então teremos $(p+1)$ equações que são obtidas diferenciando o logaritmo da função de verossimilhança com respeito aos coeficientes $\beta_0, \beta_1, \dots, \beta_p$. As equações de verossimilhança podem ser expressas da seguinte forma:

$$\sum_{i=1}^n [Y_i - \pi_i(x_i)] = 0$$

e

$$\sum_{i=1}^n x_{ik} [Y_i - \pi_i(x_i)] = 0$$

para $k = 1, 2, \dots, p$.

Assim como no modelo de regressão logística simples, o sistema formado por essas equações é resolvido através de processos iterativos. Se $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ são as soluções, então $\pi(x_i)$ é estimado por

$$\hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})}.$$

As variâncias e covariâncias assintóticas dos estimadores dos parâmetros do modelo são obtidas a partir de resultados da teoria de estimadores de máxima verossimilhança (ver Rao (1973)). A matriz de Informação (ou In-

formação Esperada), $I(\beta)$, é uma matriz $(p+1)$ por $(p+1)$, com elemento de ordem (j,u) igual a $E[-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_u}]$.

No caso do modelo de regressão logística múltipla, verifica-se que

$$E[-\frac{\partial^2 l(\beta)}{\partial \beta_j^2}] = -\frac{\partial^2 l(\beta)}{\partial \beta_j^2}$$

e

$$E[-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_u}] = -\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_u}$$

para u e $j = 0, 1, 2, \dots, p$ em que $-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_u} = \iota(\beta)$ é o elemento de ordem (u, j) da matriz de informação observada.

Além disso, verifica-se que a matriz de variância e covariância dos estimadores dos parâmetros do modelo é, para n grande, aproximadamente igual a $[I(\beta)]^{-1}$, que indicaremos por $\Sigma(\beta)$. Portanto, vamos usar a notação $\sigma^2(\hat{\beta}_j)$ para denotar o j -ésimo elemento da diagonal dessa matriz, isto é, a variância de $\hat{\beta}_j$, e $\sigma(\hat{\beta}_j, \hat{\beta}_u)$ para indicar os elementos fora da diagonal principal, que são as covariâncias entre $\hat{\beta}_j$ e $\hat{\beta}_u$. A estimativa da matriz de variância e covariância de $\hat{\beta}$, denotada por $\hat{\Sigma}(\hat{\beta})$, é obtida avaliando $\Sigma(\beta)$ em $\hat{\beta}$. Os correspondentes estimadores dessas variâncias e covariâncias serão indicados por $\hat{\sigma}^2(\hat{\beta}_j)$ e $\hat{\sigma}(\hat{\beta}_j, \hat{\beta}_u)$, respectivamente, $j, u = 0, 1, 2, \dots, p$ e são os elementos de $\hat{\Sigma}(\hat{\beta})$. Partindo disso, podemos ter as estimativas dos erros-padrão de $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, denotados por:

$$\hat{SE}(\hat{\beta}_j) = [\hat{\sigma}^2(\hat{\beta}_j)]^{\frac{1}{2}}$$

para $j = 0, 1, 2, \dots, p$.

Uma aproximação para $I(\beta)$ que será particularmente útil quando discutirmos a avaliação da qualidade do ajuste do modelo de regressão logística

é $I(\hat{\beta}) = X' \vartheta X$, em que X é uma matriz n por $(p+1)$, contendo os dados de cada elemento amostral, e ϑ é uma matriz diagonal n por n com i -ésimo elemento da diagonal dado por $\hat{\pi}_i(1 - \hat{\pi}_i)$. Assim,

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}_{n \times (p+1)}$$

e

$$\vartheta = \begin{pmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1 - \hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \hat{\pi}_n(1 - \hat{\pi}_n) \end{pmatrix}_{n \times n}.$$

2.3 Testes de significância dos parâmetros do modelo

Ajustado o modelo de regressão logística, temos interesse em construir testes de hipóteses para seus parâmetros, com o objetivo de verificar a significância das variáveis independentes, ou seja a inclusão ou exclusão dessas variáveis explicativas no modelo.

No modelo de regressão logística simples, (Seção 2.1), o primeiro passo foi avaliar a significância da variável através da estatística G dada em (2.9). No caso de regressão logística múltipla, a única diferença é que estamos trabalhando com um vetor com $(p + 1)$ parâmetros, $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$. O

desvio ou função desvio do modelo é definido como duas vezes a diferença entre os logaritmos das funções de verossimilhança maximizadas do modelo saturado e do modelo com vetor de parâmetros $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$, tendo a mesma formulação que (2.8) para o modelo (2.11). Este valor será definido com detalhes no Capítulo 3.

Podemos ter interesse em testar hipóteses relativas aos p parâmetros $\beta_1, \beta_2, \dots, \beta_p$ ou somente a um conjunto de q deles ($q < p, q = 1, 2, \dots, p - 1$). Em particular, são importantes hipóteses do tipo $\beta_1 = \beta_2 = \dots = \beta_q = 0$, que uma vez aceitas, implicariam na eliminação das correspondentes variáveis explicativas $X_1, X_2, \dots, X_q$. Neste último caso define-se o modelo completo, que seria o modelo com todas as p variáveis em questão e o modelo reduzido seria um modelo sem as q variáveis explicativas.

Então, um dos testes pode ser o da razão de verossimilhanças, em que a estatística G testaria a significância dos p coeficientes das variáveis independentes no modelo da mesma maneira que no caso do modelo de regressão logística simples. Dessa forma, para testar $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$, utilizaríamos a estatística

$$G = D_1 - D_2,$$

em que D_1 e D_2 são respectivamente os desvios dos modelos (2.10) e (2.11).

Rejeita-se H_0 ao nível de significância α se $G \geq \chi^2_{1-\alpha, p}$, em que $\chi^2_{1-\alpha, p}$ é o quantil de ordem $1 - \alpha$ da distribuição qui-quadrado com p graus de liberdade.

Podemos ainda verificar a conveniência da exclusão de uma ou mais variáveis no modelo para um melhor ajuste com menos variáveis indepen-

dentes. Testaríamos então a hipótese $H_0: \beta_1 = \beta_2 = \dots = \beta_q = 0$, através da estatística G^* , a seguir, que compara o valor da estatística D (2.7 ou 2.8) que encontramos através do modelo com e sem as variáveis independentes em questão. Isso significa que estamos verificando a permanência ou não dessas variáveis no modelo, logo estamos formulando as hipóteses estatísticas H_0 : o modelo sem as variáveis é conveniente versus H_a : o modelo com as variáveis é conveniente. Então, a estatística G^* é dada da seguinte forma

$$G^* = -2\ln\left[\frac{L_r}{L_0}\right] + 2\ln\left[\frac{L_c}{L_0}\right]$$

ou

$$G^* = D(L_r) - D(L_c).$$

em que,

- L_0 é a função de verossimilhança do modelo saturado
- L_r é a função de verossimilhança do modelo reduzido
- L_c é a função de verossimilhança do modelo completo
- D_r é o desvio do modelo reduzido
- D_c é o desvio do modelo completo,

todas calculadas nas estimativas de máxima verossimilhança sob cada modelo.

Como consequência, G^* pode ser escrita como

$$G^* = -2\ln\left[\frac{L_r}{L_c}\right]. \quad (2.13)$$

Observamos então que a estatística G^* corresponde a uma diferença dos desvios. Se n é razoavelmente grande, sob H_0 , G^* tem aproximadamente distribuição qui-quadrado com q graus de liberdade (χ_q^2), em que q é o número

de parâmetros que estão sendo testados. Portanto, se $G > \chi^2_{(1-\alpha, q)}$, rejeita-se H_0 , isto é, há indícios de que o modelo sem as correspondentes variáveis independentes não é conveniente. Note que a estatística do teste da razão de verossimilhanças pode ser dada por:

$$\mathfrak{R} = -2\ln\left[\frac{L_r}{L_c}\right]. \quad (2.14)$$

Portanto verifica-se que a estatística G^* e a do teste da razão de verossimilhanças (2.14) são iguais.

Considera-se um bom modelo aquele que utiliza um menor número de parâmetros, portanto, uma vez aceita H_0 : $\beta_1 = \beta_2 = \dots = \beta_q = 0$, adotamos o modelo reduzido contendo somente as variáveis cujos coeficientes são significantes.

Veremos dois outros testes equivalentes ao teste da razão de verossimilhanças para avaliar a significância dos parâmetros do modelo que são: o teste de Wald e o teste do escore. Estes testes serão apresentados a seguir, para o modelo de regressão logística múltipla, em que o vetor de parâmetros é $\beta = (\beta_0, \beta_1, \dots, \beta_p)$.

Para testar $H_0 : \beta^* = \beta^0$, em que $\beta^* = (\beta_1, \beta_2, \dots, \beta_p)$ e $\beta^0 = (\beta_1^0, \beta_2^0, \dots, \beta_p^0)$, a estatística de Wald é dada por

$$W = (\hat{\beta}_* - \beta^0)' [\Sigma^*(\hat{\beta}_*)]^{-1} (\hat{\beta}_* - \beta^0)$$

em que $\hat{\beta}_* = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ e $\Sigma^*(\hat{\beta}_*) = \text{Var}(\hat{\beta}_*)$ é a submatriz de $\Sigma(\hat{\beta})$ correspondendo, a $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$. Em particular, no teste H_0 : $\beta_1 = \beta_2 = \dots = \beta_p = 0$,

$$W = \hat{\beta}_*' [\Sigma^*(\hat{\beta}_*)]^{-1} \hat{\beta}_*$$

ou

$$W = \hat{\beta}'_* [X^{*'} \vartheta X^*] \hat{\beta}_*$$

em que X^* é a matriz X eliminada a coluna de uns. Sob $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$, W tem distribuição qui-quadrado com p graus de liberdade.

O teste de escore, também conhecido como teste de Rao, é baseado na função escore, isto é, na primeira derivada do logarítmo da função de verossimilhança com respeito a β . Para testar $H_0 : \beta = \beta^0$, a estatística de teste é dada da seguinte maneira:

$$Sr = \bigcup (\beta^0)' [\hat{\Sigma}_0(\hat{\beta})] \bigcup (\beta^0), \quad (2.15)$$

em que $\bigcup(\beta^0) = \frac{\partial l(\beta)}{\partial(\beta)}|_{\beta=\beta^0}$ é o valor da função escore no ponto β^0 e $\Sigma(\beta)^0$ denota a matriz de variância e covariância assintótica de $\hat{\beta}$ calculada no ponto $\beta = \beta^0$, sendo que $Var(\hat{\beta}) = \sigma(\beta) = [I(\beta)]^{-1}$.

Um passo importante após a estimação dos parâmetros do Modelo de Regressão Logística é a avaliação da qualidade do ajuste. Esse assunto será discutido com detalhes no próximo capítulo.

Capítulo 3

Avaliação do Ajuste do Modelo

3.1 Introdução

Após ajustar o modelo de regressão logística múltipla a uma amostra de n observações independentes é necessário estimar os valores do vetor de parâmetros β e testá-los, para verificar quais variáveis independentes permanecerão no modelo. O método de estimação dos parâmetros desconhecidos em regressão logística já foi abordado no capítulo anterior, sendo que posteriormente testamos a significância dos parâmetros, utilizando a estatística G . Passamos então a analisar a qualidade do ajuste do modelo *final* obtido através de uma análise de resíduos.

Em geral, numa análise de regressão, se temos uma amostra aleatória de n valores da variável resposta, representados no vetor $\mathbf{Y}' = (y_1, y_2, \dots, y_n)'$, obtemos os valores preditos do modelo ajustado $\hat{y}_i, i = 1, 2, \dots, n$, representados em $\hat{\mathbf{Y}}' = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)'$. Então, uma possível maneira de avaliarmos

a qualidade do ajuste do modelo seria através do cálculo de uma medida de distância entre Y e $\hat{Y}$. Através dessa medida, verificamos se esses valores são próximos, isto é, se a discrepância entre o valor observado e o estimado é pequena, sendo assim, haverá uma indicação de que o modelo está sendo bem ajustado. Existem outros tipos de medidas de distâncias mais rigorosas, mas não entraremos nesse aspecto no momento. Se essas diferenças são grandes, há uma clara indicação da existência de problemas com o modelo. Para cada elemento amostral, definimos o resíduo, $(y_{\text{observado}} - y_{\text{estimado}})$, e assim, o resíduo da i -ésima observação procura medir a discrepância ou a diferença entre o valor observado e o valor ajustado (estimado) da i -ésima observação através do modelo proposto.

Pregibon (1981) destaca que, após o ajuste do modelo de regressão logística, mas antes de se fazer inferência com base no mesmo, o passo seguinte deveria ser uma análise do modelo estimado, através de algum tipo de avaliação.

O papel do diagnóstico em regressão é prover métodos rotineiros para analisar modelos. Isto requer um entendimento do modelo e da natureza do método de ajuste, no caso, o da máxima verossimilhança.

Admitiremos que nosso modelo de regressão logística foi ajustado através do método da máxima verossimilhança e que já dispomos das estimativas. De acordo com Pregibon (1981), deseja-se então obter medidas de diagnóstico úteis e informativas que complementaríamos nossa análise. Essas medidas deveriam possibilitar a identificação de observações que não são bem explicadas pelo modelo ou ainda aquelas que são mais discrepantes no ajuste.

3.2 Testes de ajuste do modelo

Em regressão logística, há algumas possibilidades de medir as diferenças entre os valores observados e os estimados. Consideraremos inicialmente o caso em que temos réplicas, isto é, dispomos de $L \leq n$ possíveis valores de $\mathbf{x}_j = (1, x_{j1}, x_{j2}, \dots, x_{jp})$, $j = 1, 2, \dots, L$ e que para cada x_j , existem m_j , elementos amostrais com valor de x igual a x_j . Seja y_j o número de elementos amostrais em que $y = 1$ dentre os m_j com $x = x_j$. Se $\pi_j = P(Y = 1 | \mathbf{X} = \mathbf{x}_j)$, o valor previsto para y_j com base em $\hat{\pi}_j$ é dado por

$$m_j \hat{\pi}_j = m_j \left(\frac{\exp(\hat{g}(\mathbf{x}_j))}{1 + \exp(\hat{g}(\mathbf{x}_j))} \right), j = 1, 2, \dots, L$$

em que $\hat{g}(\mathbf{x}_j) = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \dots + \hat{\beta}_p x_{jp}$.

Vamos considerar apenas duas medidas das diferenças entre o valor observado e o previsto: o resíduo de Pearson e o resíduo componente do desvio, cuja análise será apresentada no próximo capítulo. Para um particular conjunto j de valores das variáveis explicativas, o resíduo de Pearson é definido da seguinte forma:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}, j = 1, 2, \dots, L \quad (3.1)$$

e a estatística baseada no resíduo de Pearson é:

$$\chi^2 = \sum_{j=1}^L r^2(y_j, \hat{\pi}_j) = \sum_{j=1}^L \frac{(y_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}. \quad (3.2)$$

O resíduo componente do desvio é definido como

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)} \right) \right] \right\}^{\frac{1}{2}} \quad (3.3)$$

em que o sinal é o mesmo de $(y_j - m_j \hat{\pi}_j)$. Para um conjunto com $y_j = 0$, o resíduo componente do desvio é dado por

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |\ln(1 - \hat{\pi}_j)|}$$

e para $m_j = y_j$, será

$$d(y_j, \hat{\pi}_j) = \sqrt{2m_j |\ln(\hat{\pi}_j)|}.$$

A estatística baseada no resíduo componente do desvio é:

$$D = \sum_{j=1}^L d^2(y_j, \hat{\pi}_j), \quad (3.4)$$

que testa o ajuste do modelo proposto com relação ao modelo saturado.

Uma indicação de mau ajuste seria quando os valores das estatística D ou χ^2 forem grandes. A distribuição dessas estatísticas acima, supondo que o modelo estimado está correto em todos os aspectos, é aproximadamente qui-quadrado com graus de liberdade igual a $(L - (p + 1))$. Para o desvio, temos o fato que D é a estatística da razão de verossimilhança que testa a adequação do modelo saturado com L parâmetros contra o modelo estimado com $(p + 1)$ parâmetros. O problema surge quando temos $L \approx n$, pois a distribuição obtida acima é válida assintoticamente, e portanto o número de parâmetros estará crescendo com o tamanho da amostra. McCullagh e Nelder (1989) examinaram o valor esperado de χ^2 e D quando os parâmetros do modelo foram estimados. Os autores mostram que, nesta situação, o valor esperado destas estatísticas é menor que o número de graus de liberdade, $(L - (p + 1))$. Propoem, ainda, um fator de correção que pode ser usado quando $m_j \hat{\pi}_j$ e $m_j(1 - \hat{\pi}_j)$ excedem um para cada j . A propósito, usar $(L - (p + 1))$ graus

de liberdade quando $L \approx n$, proporciona uma razoável estimativa do valor da esperança de χ^2 e D, quando o modelo estimado está *correto*.

Hosmer e Lemeshow (1980) e Lemeshow e Hosmer (1982) propuseram um teste de ajuste em que são formados grupos baseados nos valores estimados das probabilidades $\pi(\mathbf{x}) = P(Y|X = \mathbf{x})$. Supondo $L = n$, são formados grupos de observações segundo um dos seguintes métodos:

1. percentis das probabilidades estimadas $\hat{\pi}(\mathbf{x})$ e
2. valores fixos das probabilidades estimadas.

Com o primeiro método, os autores sugerem a formação de g grupos de mesmo tamanho (aproximadamente), da seguinte forma:

grupo 1: formado pelos $n'_1 = \frac{n}{g}$ elementos amostrais com os menores valores de $\hat{\pi}(\mathbf{x})$,

grupo 2: formado pelos $n'_2 = \frac{n}{g}$ elementos amostrais com os segundos menores valores de $\hat{\pi}(\mathbf{x})$,

e assim, sucessivamente, até a obtenção do último grupo

grupo g: formado pelos $n'_g = \frac{n}{g}$ elementos amostrais com os maiores valores de $\hat{\pi}(\mathbf{x})$.

Se $\frac{n}{g}$ não for inteiro, tomar o inteiro mais próximo.

No segundo método, os autores sugerem a formação de g grupos com pontos de corte para $\hat{\pi}(\mathbf{x})$: $\frac{1}{g}, \frac{2}{g}, \dots, \frac{g-1}{g}$, assim, temos

grupo 1: formado pelos elementos amostrais com $0 \leq \hat{\pi}(\mathbf{x}) \leq \frac{1}{g}$

grupo 2: formado pelos elementos amostrais com $\frac{1}{g} < \hat{\pi}(\mathbf{x}) \leq \frac{2}{g}$

e assim, sucessivamente, até

grupo g: formados pelos elementos amostrais com $\frac{g-1}{g} < \hat{\pi}(\mathbf{x}) \leq 1$.

Na maioria dos casos, utiliza-se o valor de $g = 10$.

Para cada método, os autores obtêm o valor de uma estatística C para avaliar a qualidade do ajuste do modelo, que tem como base a estatística qui-quadrado de Pearson, isto é, $\sum \frac{(O-E)^2}{E}$, em que, O : valores observados, E : valores esperados.

Então, a fórmula para o cálculo de C é dada por:

$$C = \sum_{k=1}^g \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (3.5)$$
$$o_k = \sum_{l=1}^{n'_k} y_l$$

em que,

- o_k : número de uns (em Y) no k -ésimo grupo;
- n'_k : número de diferentes valores de $\mathbf{x}$ no k -ésimo grupo;
- m_l : número de elementos amostrais com $x = x_l$ e

$$\bar{\pi}_k = \sum_{l=1}^{n'_k} \frac{m_l \hat{\pi}_l}{n'_k}$$

é a média das probabilidades estimadas dentro do k -ésimo grupo, $k = 1, 2, \dots, g$.

Hosmer e Lemeshow (1980, [pg.1053]) verificam através de simulações que quando $L = n$ e o modelo de regressão logística estimado está correto, a

distribuição assintótica da estatística C pode ser bem aproximada por uma distribuição qui-quadrado com $(g-2)$ graus de liberdade.

Uma alternativa para o denominador na equação (3.5) é obtida se considerarmos o_k como soma de variáveis aleatórias independentes e não identicamente distribuídas, de modo que a soma dos quadrados das diferenças entre a frequência esperada e a observada no k -ésimo grupo fica padronizada por:

$$\sum_{l=1}^{n'_k} \hat{\pi}_l(1 - \hat{\pi}_l).$$

Neste caso, a estatística do teste fica

$$C' = \sum_{k=1}^g \frac{(o_k - n'_k \bar{\pi}_k)^2}{\sum_{l=1}^{n'_k} \hat{\pi}_l(1 - \hat{\pi}_l)}.$$

Verifica-se que o valor de C' é um pouco maior que o de C . Sendo essa diferença muito pequena, os autores sugerem o uso de C . Grandes valores da estatística C indicam que o modelo de regressão logística não é apropriado. A um nível de significância α , a hipótese H_0 : o modelo logístico é apropriado é rejeitada se e somente se

$$C > \chi^2_{(1-\alpha, g-2)},$$

em que $\chi^2_{(1-\alpha, g-2)}$ é o quantil de ordem $1 - \alpha$ da distribuição qui-quadrado com $g - 2$ graus de liberdade.

Como mencionamos na Seção 2.1, uma maneira rápida e informal de examinar o ajuste do modelo é comparar o valor observado do desvio com $(L - 2)$ graus de liberdade no modelo de regressão logística simples e no modelo de regressão logística múltipla, se esse valor do desvio for muito

superior a $(L - p - 1)$, o ajuste do modelo é considerado inadequado. Então o uso de C seria uma técnica mais elaborada para examinar a qualidade do ajuste do modelo.

Finalizando, apresentaremos uma outra maneira de avaliar o ajuste do modelo de regressão logística, proposta por Tsiatis (1980). Os possíveis valores das variáveis explicativas são divididos em g grupos e com base nessa divisão, define-se o vetor de funções indicadoras I , tal que $I_i = (I_i^1, I_i^2, \dots, I_i^g)$, sendo que

$$I_i^k = \begin{cases} 1 & \text{se } (x_{i0}, x_{i1}, \dots, x_{ip}) \text{ pertence ao } k\text{-ésimo grupo} \\ 0 & \text{caso contrário,} \end{cases}$$

para $i = 1, 2, \dots, n$ e $k = 1, 2, \dots, g$. A análise é baseada no teste de escore eficiente (Rao, 1973 pag. 418) para o parâmetro γ do modelo

$$\ln \frac{\pi(x)}{1 - \pi(x)} = \beta' \mathbf{X} + \gamma' \mathbf{I} \quad (3.6)$$

em que I é a função indicadora e $\gamma' = (\gamma_1, \dots, \gamma_g)$ é um vetor de g parâmetros.

A hipótese de que o modelo de regressão logística é adequado equivale a $H_0: \gamma = 0$. Para testar essa hipótese, a estatística proposta por Tsiatis é dada por:

$$T = \chi' V^{-1} \chi \quad (3.7)$$

em que

- χ é um vetor g -dimensional (vetor da função escore), dado por

$$U(\gamma) = \chi' = \left(\frac{\partial l}{\partial \gamma_1}, \dots, \frac{\partial l}{\partial \gamma_g} \right);$$

- l é o logaritmo neperiano da função de verossimilhança do modelo (3.6);

- $Var(\gamma) = V = A - BC^{-1}B'$: é uma matriz g por g , sendo que os elementos de A , B e C são respectivamente

$$A_{jj'} = -\frac{\partial^2 l}{\partial \gamma_j \partial \gamma_{j'}}, j, j' = 1, \dots, g$$

$$B_{jj'} = -\frac{\partial^2 l}{\partial \gamma_j \partial \beta_{j'}}, j = 1, \dots, g, j' = 0, \dots, p$$

$$C_{jj'} = -\frac{\partial^2 l}{\partial \beta_j \partial \beta_{j'}}, j, j' = 0, \dots, p \text{ e}$$

- V^{-} é a inversa generalizada de V .

Esses elementos são calculados no ponto $\gamma = 0$ e $\beta = \hat{\beta}$ em que $\hat{\beta}$ é o estimador de máxima verossimilhança de β no modelo sob H_0 , $ln \frac{\pi(x)}{1-\pi(x)} = \beta' X$.

Sob H_0 , a estatística T tem assintoticamente distribuição qui-quadrado com g graus de liberdade.

Após cálculos, os autores verificam que essa estatística é uma forma quadrática no vetor de frequências observadas menos as esperadas nos g grupos. Com isso temos

$$\chi' = (O_1 - E_1, O_2 - E_2, \dots, O_g - E_g),$$

em que

- O_k é o número de uns (elementos com a característica de interesse) dentro do grupo k e
- E_k é a soma das probabilidades estimadas dentro de cada grupo, $k = 1, 2, \dots, g$.

Verifica-se também que as matrizes A , B e C podem ser obtidas da seguinte forma:

- a matriz A é diagonal g por g com elementos da diagonal principal

$$\sum_{\xi_k} \hat{\pi}_i(1 - \hat{\pi}_i), k = 1, \dots, g$$

em que ξ_k denota o conjunto de índices i tal que $(x_{i0}, x_{i1}, \dots, x_{ip})$ pertence ao grupo k e $\hat{\pi}_i = \frac{\exp(\hat{\beta}' \mathbf{x}_i)}{1 + \exp(\hat{\beta}' \mathbf{x}_i)}$

- a matriz B , g por $(p+1)$, é dada por

$$B = \begin{pmatrix} \sum_{\xi_1} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_1} x_{i1} \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{\xi_1} x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{\xi_2} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_2} x_{i1} \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{\xi_2} x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\xi_g} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_g} x_{i1} \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{\xi_g} x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) \end{pmatrix}_{g \times (p+1)}$$

e finalmente a matriz C que é simétrica $(p+1)$ por $(p+1)$, fica,

$$C = \begin{pmatrix} \sum_{i=1}^n \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{i=1}^n x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{i=1}^n x_{i1} \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{i=1}^n x_{i1} x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} \hat{\pi}_i(1 - \hat{\pi}_i) & \cdots & \sum_{i=1}^n x_{ip}^2 \hat{\pi}_i(1 - \hat{\pi}_i) \end{pmatrix}_{(p+1) \times (p+1)}$$

A seguir, procuraremos ilustrar num exemplo algumas das análises apresentadas.

3.3 Exemplo

Analisaremos a seguir o conjunto de dados apresentado em Neter, Kutner, Nachtsheim e Wasserman (1996), relativo à variável resposta compra de carro (sim ou não) em função de algumas características da família.

Uma montadora de veículos contrata uma empresa de pesquisa de mercado na condução de um estudo piloto para analisar a probabilidade de uma família comprar um carro novo durante o próximo ano. Uma amostra casual de 33 famílias foi selecionada e foram coletados as seguintes informações:

- dados da renda familiar anual (X_1 , em milhares de dólares) e
- idade do automóvel mais velho da família (X_2 , em anos).

Foi conduzida uma entrevista 12 meses mais tarde para determinar se, durante o ano, houve a compra de um novo veículo ($Y = 1$) ou não ($Y = 0$). Para análise desses dados aplicou-se o modelo de regressão logística múltipla (2.11), sendo que $\pi_i = P(Y_i | X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}$ é a probabilidade da i -ésima família obter um automóvel novo, $i = 1, 2, \dots, 33$.

Utilizamos o programa estatístico *S - Plus* 2000, para o ajuste do modelo com base nos dados presentes na Tabela 3.1.

A partir dos dados analisados, obtemos a Tabela 3.2, com as estimativas dos parâmetros do modelo ajustado. O valor da função desvio do modelo completo, isto é, do modelo com as duas variáveis independentes em questão, foi de $D = 36,69$ (30 graus de liberdade), indicando um ajuste razoável, pois, o desvio é um pouco maior que o número de graus de liberdade. Vamos retirar cada variável independente do modelo para verificarmos a sua importância.

Tabela 3.1 - Dados relativos às 33 famílias.

família	Y_i	x_{i1}	x_{i2}	$\hat{\pi}(x_{i1})$	família	Y_i	x_{i1}	x_{i2}	$\hat{\pi}(x_{i1})$
1	0	32,0	3	0,36	2	0	45,0	2	0,49
3	1	60,0	2	0,65	4	0	53,0	1	0,58
5	0	25,0	4	0,29	6	1	68,0	1	0,73
7	1	82,0	2	0,83	8	1	38,0	5	0,42
9	0	67,0	2	0,72	10	1	92,0	2	0,88
11	1	72,0	3	0,76	12	0	21,0	5	0,26
13	0	26,0	3	0,30	14	1	40,0	4	0,44
15	0	33,0	3	0,37	16	0	45,0	1	0,49
17	1	61,0	2	0,66	18	0	16,0	3	0,22
19	1	18,0	4	0,23	20	0	22,0	6	0,26
21	0	27,0	3	0,31	22	1	35,0	3	0,39
23	1	40,0	3	0,44	24	0	10,0	4	0,18
25	0	24,0	3	0,28	26	1	15,0	4	0,21
27	0	23,0	3	0,27	28	0	19,0	5	0,24
29	1	22,0	2	0,26	30	0	61,0	2	0,66
31	0	21,0	3	0,26	32	1	32,0	5	0,36
33	0	17,0	1	0,22					

Retirando a variável idade do automóvel mais velho (X_2) do modelo completo, obtemos o modelo reduzido. Na Tabela 3.3 encontram-se as estimativas dos parâmetros desse modelo. Com isso desejamos verificar se a inclusão dessa variável é importante para a explicação da variável resposta. Após sua retirada, observamos que o desvio aumentou um pouco, indo para $D = 39,30$, com 31 graus de liberdade. Com isso, G^* dada em (2.13) é

$G^* = 39,30 - 36,69 = 2,61$, que é a diferença dos valores das funções desvios do modelo reduzido e completo. Para testarmos a importância de X_2 , comparamos o valor de G^* obtido com 3,8, que é o valor crítico da distribuição qui-quadrado com 1 grau de liberdade e nível de significância $\alpha = 0,05$. Assim, concluímos que não há evidência estatística de que a variável X_2 seja significativa no modelo. Repetimos esse procedimento para a variável renda familiar (X_1) e obtemos o desvio do modelo que envolve somente X_2 . O valor obtido, $D = 44,99$ para 31 graus de liberdade indica um ajuste ruim, pois notamos um desvio maior que o do modelo que inclui somente X_1 .

Tabela 3.2 - Estimativas dos parâmetros do modelo completo

Efeito	Parâmetro	Estimativa	desvio-padrão estimado
constante	β_0	-4,7393	2,1016
renda familiar	β_1	0,068(p=0,016)	0,0281
automóvel mais velho	β_2	0,599(p=0,124)	0,3900

Tabela 3.3 - Estimativas dos parâmetros do modelo reduzido

Efeito	Parâmetro	Estimativa	desvio-padrão estimado
constante	β_0	-1,9808	0,8561
renda familiar	β_1	0,043(p=0.031)	0,0201

Então propomos a utilização do modelo com apenas a variável renda familiar, isto é, o modelo reduzido. Concluímos que a probabilidade de uma família adquirir um automóvel novo no próximo ano esta diretamente relacionada com a renda familiar anual, isto é conforme aumenta a renda familiar

anual, a chance de adquirir um novo carro aumenta também. Além disso, estima-se que essa probabilidade de compra de carro para uma família com renda familiar x é

$$\hat{\pi}(x) = \frac{e^{-1,98+0,043x}}{1 + e^{-1,98+0,043x}}.$$

Selecionado o modelo, vamos medir a qualidade do seu ajuste, através da estatística C , dada em (3.5).

Apresentamos na Tabela 3.4 os grupos formados com as observações da Tabela 3.1 para o cálculo da estatística C proposta por Hosmer e Lemeshow (1989).

Tabela 3.4 - Quantidades para o cálculo de C .

Grupo	Observações	O_i	n'_i	$\bar{\pi}_i$
1	18,19,24,26,33	2	5	0,212
2	12,20,28,29	1	4	0,255
3	5,25,27,31	0	4	0,275
4	1,13,21,32	1	4	0,333
5	8,15,22,23	3	4	0,405
6	2, 4,14,16	1	4	0,500
7	3, 9,17,30	2	4	0,673
8	6, 7,10,11	4	4	0,800

Foram formados 8 grupos, sendo sete com quatro observações cada e um grupo com cinco observações. O método para formação dos grupos foi o das menores probabilidades estimadas do modelo selecionado para as maiores,

que encontram-se na quarta coluna da Tabela 3.1 ($\hat{\pi}(x_1)$), sendo cinco observações para o primeiro grupo e as demais distribuídas nos outros sete grupos, de modo que se tenha as mesmas quantidades de observações em cada um deles. Os valores necessários para o cálculo de C encontram-se na Tabela 3.4 e seu cálculo a seguir:

$$C = \frac{(2 - 5.0, 212)^2}{5.0, 212.(1 - 0, 212)} + \dots + \frac{(4 - 4.0, 800)^2}{4.0, 800.(1 - 0, 800)},$$

$$C = 1, 0578 + 0, 0005 + 1, 5172 + 0, 1241 + 1, 9757 + 1 + 0, 5440 + 1 = 7, 2194.$$

O nível descritivo associado ao valor de C numa distribuição qui-quadrado com $(g - 2) = 6$ graus de liberdade é dado por 0,3010, indicando que o ajuste é adequado. Utilizando a estatística C' obtemos,

$$C' = 1, 0597 + 0, 0005 + 1, 5182 + 0, 1246 + 1, 9817 + 1, 0103 + 0, 5455 + 1, 0220 = 7, 2625.$$

Notemos que C' é um pouco maior que C , sendo a diferença muito pequena, conforme já comentado anteriormente.

Calcularemos a seguir a estatística de Tsiatis, (3.7), para avaliarmos a qualidade do ajuste do modelo. Inicialmente para o cálculo desta estatística, vamos definir as regiões ou grupos:

$$R_1 = \{i \mid 8 \leq x_{i1} < 24\}$$

$$R_2 = \{i \mid 24 \leq x_{i1} < 40\}$$

$$R_3 = \{i \mid 40 \leq x_{i1} < 56 \}$$

$$R_4 = \{i \mid 56 \leq x_{i1} < 72\}$$

$$R_5 = \{i \mid 72 \leq x_{i1} \leq 92\}.$$

As regiões foram construídas de modo a ter amplitudes aproximadamente iguais.

O vetor χ e as matrizes A , B e C são dados por

$$\chi' = (O_1 - E_1, O_2 - E_2, O_3 - E_3, O_4 - E_4, O_5 - E_5),$$

$$A = \begin{pmatrix} \sum_{\xi_1} \hat{\pi}_i(1 - \hat{\pi}_i) & 0 & 0 & 0 & 0 \\ 0 & \sum_{\xi_2} \hat{\pi}_i(1 - \hat{\pi}_i) & 0 & 0 & 0 \\ 0 & 0 & \sum_{\xi_3} \hat{\pi}_i(1 - \hat{\pi}_i) & 0 & 0 \\ 0 & 0 & 0 & \sum_{\xi_4} \hat{\pi}_i(1 - \hat{\pi}_i) & 0 \\ 0 & 0 & 0 & 0 & \sum_{\xi_5} \hat{\pi}_i(1 - \hat{\pi}_i) \end{pmatrix}$$

$$B = \begin{pmatrix} \sum_{\xi_1} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_1} x_i \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{\xi_2} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_2} x_i \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{\xi_3} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_3} x_i \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{\xi_4} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_4} x_i \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{\xi_5} \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{\xi_5} x_i \hat{\pi}_i(1 - \hat{\pi}_i) \end{pmatrix}$$

$$C = \begin{pmatrix} \sum_{i=1}^n \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{i=1}^n x_i \hat{\pi}_i(1 - \hat{\pi}_i) \\ \sum_{i=1}^n x_i \hat{\pi}_i(1 - \hat{\pi}_i) & \sum_{i=1}^n x_i^2 \hat{\pi}_i(1 - \hat{\pi}_i) \end{pmatrix}.$$

Lembrando que $T = \chi' V^{-1} \chi$, em que $V = A - BC^{-1}B'$, após cálculos algébricos encontramos $T = 1,127$. Como a estatística T assintoticamente tem distribuição qui-quadrado com 3 graus de liberdade sob a hipótese de que o modelo proposto é adequado e $T_{crítico, \alpha=0,05} = 9,488$, então não rejeitamos H_0 , isto é, o modelo de regressão logística com renda familiar como variável independente é conveniente.

3.4 Considerações Finais

Em algumas situações pode ser possível excluir uma sub-amostra das observações, desenvolver um modelo baseado nos elementos restantes e então testar sua qualidade com base nos elementos excluídos. Em outras situações, obtém-se uma nova amostra para avaliar a qualidade de um modelo já ajustado. Esses processos são chamados de validação, e têm especial importância quando o modelo é usado para prever valores da variável resposta para observações futuras. Algumas medidas de diagnóstico que serão discutidas nas próximas seções, como $\Delta\chi^2, \Delta D, \Delta\hat{\beta}$ imitam esta idéia, pois são calculadas com base na exclusão de elementos amostrais. Com um novo conjunto de dados, é possível uma avaliação mais completa.

Uma vez ajustado o modelo de regressão logística com base numa amostra inicial, o processo exige uma segunda amostra para validação com n_v observações $(y_i, x_i), i = 1, 2, \dots, n_v$ divididas em J_v conjuntos de variáveis independentes. Seja y_j o número de respostas um dentre os m_j elementos com conjunto de variáveis independentes $\mathbf{x} = \mathbf{x}_j$, para $j = 1, 2, \dots, J_v$. A análise é feita com base em

$$\hat{\pi}_j = \frac{\exp\{\hat{g}(\mathbf{x}_j)\}}{1 + \exp\{\hat{g}(\mathbf{x}_j)\}},$$

a estimativa da probabilidade π_j para o valor das variáveis explicativas em x_j presentes na amostra de validação. Essa quantidade será utilizada no cálculo das medidas de ajuste já discutidas, χ^2 , D e C , através do processo de validação.

O cálculo da estatística qui-quadrado de Pearson da equação (3.2) é

feito com as respectivas substituições dos seus valores para os elementos da amostra de validação. No caso, χ^2 é calculado como a soma dos J_v termos independentes, ou seja,

$$\chi^2 = \sum_{j=1}^{J_v} \frac{(y_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}.$$

Se cada m_j é grande o suficiente para usar a aproximação normal para a distribuição binomial, então, sob a hipótese de que o modelo ajustado é correto, a estatística obtida terá uma distribuição $\chi^2_{J_v}$.

De acordo com Hosmer e Lemeshow, usualmente os valores dos m_j são pequenos, com a maior parte deles igual a 1. Nesse caso, a aproximação descrita anteriormente não pode ser empregada e os autores sugerem que o valor da estatística seja comparado com a esperança da distribuição $\chi^2_{J_v}$, dada por J_v .

Similares considerações valem no cálculo e interpretação de D para a amostra de validação. O modelo estimado é considerado um ajuste *pobre* para os dados de validação caso essas duas estatísticas (χ^2 e D) excedam J_v .

Pode-se obter equivalentemente a estatística C de (3.5) para a amostra de validação. Assim, admitindo a construção de g grupos,

$$C_v = \sum_{k=1}^g \frac{(o_k - e_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (3.8)$$

em que,

n_k é aproximadamente igual a $\frac{n_v}{g}$ que é o número de sujeitos no k -ésimo conjunto,

$o_k = \sum_{j=1}^{n_k} m_j y_j$ é o número de uns dentro do k -ésimo grupo,

$e_k = \sum_{j=1}^{n_k} m_j \hat{\pi}_j$ é a estimativa da frequência esperada do número de uns no k -ésimo grupo sob a hipótese de que o modelo é correto e $\bar{\pi}_k = \frac{\sum m_j \hat{\pi}_j}{n_k}$.

O índice v enfatiza que estamos calculando as estatísticas para a amostra de validação. Sob a hipótese de que o modelo ajustado é correto, se todos os e_k são suficientemente grandes, de modo que cada termo de C_v tenha distribuição qui-quadrado com 1 grau de liberdade, temos que C_v tem distribuição qui-quadrado com g graus de liberdade. Além da análise de C_v através do cálculo do p-valor, recomenda-se que cada termo em C_v seja examinado para avaliar o ajuste dentro de cada grupo.

Para complementar a análise do ajuste, existe ainda a possibilidade do uso de medidas do tipo R^2 adaptadas para modelos de regressão logística.

Se L e L_p são os valores das funções de verossimilhança calculadas nas estimativas de máxima verossimilhança do modelo (2.10) e (2.11), respectivamente, uma possível medida da qualidade do ajuste do modelo com p variáveis independentes é

$$R^2 = 1 - \frac{\log L_p}{\log L}.$$

Como L_p e L estão compreendidas entre 0 e 1, grandes valores de R^2 indicam um valor de L_p muito maior que o de L e portanto, a conveniência da introdução das p variáveis independentes no modelo. Além disso, como é usual para medidas dessa natureza, $0 \leq R^2 \leq 1$.

Verifica-se ainda a existência de uma relação direta entre R^2 e G , a estatística que testa a hipótese $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$. Devido a esta relação, Hosmer e Lemeshow (1989) consideram que a medida R^2 tem pouco

a informar sobre o ajuste do modelo. Outras medidas de ajuste para modelos de regressão logística encontram-se em Windmeijer (1995).

As quantidades χ^2 , D e C introduzidas neste capítulo, são medidas globais do ajuste do modelo.

A análise de diagnóstico, que será introduzida no próximo capítulo, permite o estudo do ajuste para cada elemento amostral separadamente. Caso existam muitas observações aberrantes, elas podem mascarar o resultado final, prejudicando a conclusão do estudo em questão. Esses pontos são detectados através de um estudo dos resíduos, com a utilização de técnicas de diagnóstico específicas para regressão logística.

Capítulo 4

Medidas de Diagnóstico em Regressão Logística

4.1 Introdução

Apresentaremos nesta seção algumas maneiras de detectar observações discrepantes numa amostra, isto é, vamos examinar os resíduos do modelo ajustado através de medidas para verificarmos a existência de possíveis pontos aberrantes ou discrepantes.

Este estudo é baseado na discrepância entre o valor observado e o valor ajustado da i -ésima observação. Para isso, conforme comentamos, verifica-se o comportamento de algumas medidas, e a este procedimento é usual chamar de Análise de diagnóstico. Observações muito distintas das demais podem ter uma influência desproporcional nos resultados e conseqüentemente na conclusão da análise que está sendo feita.

Em Regressão Logística, assim como na linear, temos uma coleção de possíveis gráficos de diagnóstico, testes para avaliar a qualidade do modelo estimado (Capítulo 3) e medidas de diagnóstico.

Tais métodos de diagnóstico são muito importantes numa análise de dados e devem fazer sempre parte da mesma. Observações discrepantes são facilmente detectadas se empregarmos os diversos tipos de resíduos e a correspondente técnica de diagnóstico corretamente. Trataremos nesta seção da detecção de pontos discrepantes.

4.2 Medidas de diagnóstico

O resíduo, uma medida de discrepância, compara os valores observados (y_i) com os valores preditos pelo modelo ($\hat{y}_i$) para cada observação na amostra, ou seja, é uma função do tipo $r_i = r(y_i, \hat{y}_i)$. A definição mais simples e comum de resíduo é

$$r_i = y_i - \hat{y}_i,$$

que é o resíduo ordinário, mas existem outras possíveis formas.

Antes de aceitarmos um modelo ajustado, devemos entre outras coisas, fazer uma avaliação através do estudo dos resíduos, que é um procedimento bastante utilizado na regressão linear. Neste capítulo, estudaremos com detalhe os resíduos de Pearson e o resíduo componente do desvio, já mencionados no capítulo anterior, bem como formas para analisá-los. Tal análise foi introduzida por Pregibon (1981) que, num trabalho teórico, estende as técnicas de diagnóstico de regressão linear para regressão logística.

O objetivo dessa seção é descrever algumas técnicas de diagnóstico em regressão logística. Neste desenvolvimento, vamos assumir que o modelo estimado ou ajustado contem p variáveis explicativas $(X_1, X_2, \dots, X_p)$.

Em regressão linear, assume-se que a variância de Y_j não depende da média condicional, $E(Y_j|\mathbf{x}_j)$. Contudo, em regressão logística, esta variância é uma função da média condicional, pois $\text{var}(Y_j|\mathbf{x}_j) = m_j E(Y_j|\mathbf{x}_j)[1 - E(Y_j|\mathbf{x}_j)] = m_j \pi(\mathbf{x}_j)[1 - \pi(\mathbf{x}_j)]$. Neste caso, são utilizados os resíduos definidos nas equações (3.1) e (3.3), respectivamente

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}, j = 1, 2, \dots, L$$

e

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

definidos para cada vetor de variáveis explicativas $\mathbf{x}_j = (1, x_1, \dots, x_p)$, $j = 1, 2, \dots, L$.

Tais valores podem ser analisados como uma medida de influência do j -ésimo caso nas estatísticas de ajuste χ^2 e D , definidas em (3.2) e (3.4) respectivamente.

Seja $\mathbf{X}$ a matriz contendo os L valores do conjunto de variáveis explicativas associadas às p variáveis independentes, sendo que a primeira coluna é composta por *uns* para refletir a presença do intercepto do modelo. A matriz $\mathbf{X}$ é chamada de matriz de planejamento.

Em regressão linear, a matriz de projeção é conhecida e definida como $H = X(X'X)^{-1}X'$, de modo que $\hat{y} = Hy$. O vetor de resíduos da regressão linear, $y - \hat{y}$, expresso em termos da matriz de projeção é $(I - H)y$, em que

I é uma matriz identidade L por L .

No modelo de regressão linear para mínimos quadrados ponderados, temos que $\text{Var}(\epsilon) = \sigma^2 W$ em que W é a matriz de pesos, diagonal e o estimador de máxima verossimilhança de β é dado por $\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}y$, conseqüentemente a matriz de projeção é dada por $H = W^{-1/2}X(X'W^{-1}X)^{-1}X'W^{-1/2}$, em que $W^{-1/2}$ é a matriz tal que $W^{-1} = W^{-1/2}W^{-1/2}$.

A obtenção da matriz de projeção para o modelo de regressão logística, bem como a extensão dos resultados de diagnóstico da regressão linear para a logística foram propostos por Pregibon (1981). Esta generalização foi possível devido à formulação do estimador de máxima verossimilhança de β no modelo de regressão logística como um estimador do tipo $(X'W^{-1}X)^{-1}X'W^{-1}z$. Para encontrar esse estimador, utiliza-se métodos iterativos de cálculo de estimadores de máxima verossimilhança. Inicialmente lineariza-se a função escore através do uso de expansão de Taylor, obtendo-se

$$U(\beta) \cong U(\beta^0) + (\beta - \beta^0) \frac{\partial U(\beta^0)}{\partial(\beta)}$$

em que $U(\beta) = \frac{\partial l(\beta)}{\partial \beta}$ e β^0 é uma estimativa inicial ou valor inicial. Daí, como conseqüência do fato que $U(\hat{\beta}) = 0$ e utilizando-se essa equação iterativamente, teremos

$$\beta^{m+1} = \beta^m + \{-\partial U(\beta^m)\}^{-1}U(\beta^m)$$

em que o superescrito m indica uma quantidade avaliada na m -ésima interação, para $m = 0, 1, 2, \dots$, que é um processo iterativo de Newton Raphson (Andersen, 1991). O processo é repetido até que $|\frac{\beta_k^{m+1} - \beta_k^m}{\beta_k^m}| < \epsilon, \forall k = 1, 2, \dots, p$.

Podemos também usar o Método *Scoring* de Fisher (McCullagh e Nelder, 1989), isto é, substituir $-\frac{\partial U(\beta)}{\partial(\beta)}$ pelo correspondente valor esperado, $E[-\frac{\partial U(\beta)}{\partial(\beta)}]$ que coincide com $I(\beta)$. Deste modo,

$$\beta^{m+1} = \beta^m + \{I(\beta^m)\}^{-1}U(\beta^m) \quad (4.1)$$

sendo que $I(\beta) = X'WX$ é a matriz de informação de Fisher e W é a matriz diagonal, com o j -ésimo elemento da diagonal dado por $m_j\pi_j(1 - \pi_j)$, $j = 1, 2, \dots, L$.

Paula (1999) obtém a expressão de $U(\beta)$ para modelos da família exponencial. Em particular, para a distribuição de Binomial, segue que $U(\beta) = X'(\mathbf{y} - \mathbf{n}\pi(x))$ e assim,

$$\begin{aligned} \hat{\beta}^{m+1} &= \hat{\beta}^m + (X'W^mX)^{-1}X'(\mathbf{y} - \mathbf{n}\hat{\pi}^m(x)) \\ &= \hat{\beta}^m + (X'W^mX)^{-1}X'W^m(W^m)^{-1}(\mathbf{y} - \mathbf{n}\hat{\pi}^m(x)) \\ &= (X'W^mX)^{-1}X'W^mX\hat{\beta}^m + (X'W^mX)^{-1}X'W^m(W^m)^{-1}(\mathbf{y} - \mathbf{n}\hat{\pi}^m(x)) \\ &= (X'W^mX)^{-1}X'W^m[X\hat{\beta}^m + (W^m)^{-1}(\mathbf{y} - \mathbf{n}\hat{\pi}^m(x))] \end{aligned}$$

emque

$$\mathbf{y} = (y_1, y_2, \dots, y_L)'$$

$$\mathbf{n} = (n_1, n_2, \dots, n_L)'$$

W^m é a matriz W estimada na m -ésima interação e

$\hat{\pi}^m(x)$ é um vetor contendo as estimativas de $\pi(\mathbf{x}_j)$, $j = 1, 2, \dots, L$ na m -ésima interação. Por exemplo, no modelo de regressão logística simples

$$\hat{\pi}^m(x_1) = \frac{\exp(\hat{\beta}_0^m + \hat{\beta}_1^m x_1)}{1 + \exp(\hat{\beta}_0^m + \hat{\beta}_1^m x_1)}.$$

Desta forma

$$\hat{\beta}^{m+1} = (X'W^mX)^{-1}X'W^m\mathbf{z}^m, \quad (4.2)$$

em que

$$z^m = X\hat{\beta}^m + (W^m)^{-1}(y - n\hat{\pi}^m(x)),$$

e de modo geral $\hat{\beta} = (X'WX)^{-1}X'Wz$.

Como consequência, a matriz de projeção neste caso é dada por

$$H = V^{1/2}X(X'VX)^{-1}X'V^{1/2} \quad (4.3)$$

em que V é uma matriz diagonal L por L com o elemento geral dado por $v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$.

Em regressão linear, os elementos da diagonal da matriz de projeção são chamados de valores alavanca ou pontos alavanca, são proporcionais à distância entre x_j e a média dos x_j na amostra. Este conceito de distância é importante em regressão linear, os pontos que estão distantes da média podem ser de grande influência nos valores das estimativas dos parâmetros. A extensão desse conceito de pontos alavanca para regressão logística requer adicional discussão e esclarecimento.

Seja h_j o j -ésimo elemento da diagonal da matriz H definida na equação (4.3). A partir dessa expressão, verifica-se facilmente que

$$h_j = v_j x_j' (X'VX)^{-1} x_j = m_j \hat{\pi}(x_j) [1 - \hat{\pi}(x_j)] x_j' (X'VX)^{-1} x_j = v_j b_j \quad (4.4)$$

em que

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

Hosmer e Lemeshow (1989) mostram através de um exemplo que o comportamento de h_j numa regressão logística é diferente do comportamento dessa mesma medida na regressão linear. Os autores fazem um gráfico que

descreve os valores de pontos alavanca (h_j) em função das probabilidades ajustadas $\hat{\pi}(\mathbf{x}_j)$ para uma amostra de 100 observações. A medida h_j pode ser interpretada de maneira similar à da regressão linear normal para $0, 1 \leq \hat{\pi}_j \leq 0, 9$. Mas, quando $\hat{\pi}_j$ é pequeno ou alto, h_j torna-se em geral pequeno, o que pode dificultar a detecção de pontos que estejam mais afastados no subespaço gerado pelas colunas da matriz X (Paula, 1999).

Uma quantidade que cresce com a distância da média é $b_j = \mathbf{x}_j'(\mathbf{X}'V\mathbf{X})^{-1}\mathbf{x}_j$. Então, se somente essa distância interessar, podemos nos focar em b_j , fazendo um gráfico de b_j em função das probabilidades estimadas.

Consideremos um conjunto fixo de valores das variáveis explicativas com m_j elementos, $y_j = 0$ e probabilidades estimadas através do modelo logístico $\hat{\pi}_j$. O resíduo de Pearson para o i – ésimo elemento desse grupo é dado por

$$r_i = \frac{(0 - \hat{\pi}_j)}{\sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}} = -\sqrt{\frac{\hat{\pi}_j}{1 - \hat{\pi}_j}}, i = 1, 2, \dots, m_j$$

e o resíduo de Pearson baseado em todos os elementos com este conjunto de variáveis explicativas é

$$r_j = \frac{(0 - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j(1 - \hat{\pi}_j)}} = -\sqrt{m_j} \sqrt{\frac{\hat{\pi}_j}{1 - \hat{\pi}_j}},$$

que aumenta em valor absoluto quando m_j cresce. Grandes valores deste resíduo indicam um mau ajuste do modelo para aquela observação. Se $m_j = 1$ e $\hat{\pi}_j = 0, 5$, então $r_j = -1$ que não é um grande resíduo. De outro modo, se temos $m_j = 16$ elementos com mesmo conjunto de variáveis explicativas, então $r_j = -4, 0$, que é considerado um grande resíduo.

Andersen(1991,[pgs. 288 a 290]), utilizando resultados desenvolvidos

para a estimação de máxima verossimilhança dos parâmetros da distribuição multinomial mostra que

$$Var(y_j - m_j \hat{\pi}(\mathbf{x}_j)) \approx m_j \pi(\mathbf{x}_j) [1 - \pi(\mathbf{x}_j)] (1 - h_j),$$

que sugere que o resíduo de Pearson não tem variância igual a 1. Verifica-se ainda que os r'_j s têm variâncias diferentes.

De acordo com o autor, um alto valor de h_j , além do discutido anteriormente, indicaria um elemento amostral para o qual a variância do resíduo $y_j - m_j \hat{\pi}(\mathbf{x}_j)$ se desviaria mais da variância de uma variável aleatória com distribuição Binomial que os demais dados. Como consequência, este ponto pode apresentar uma influência inadequada no ajuste do modelo.

Então, se denotarmos por $r(y_j, \hat{\pi}_j)$ o resíduo de Pearson, isto é, $r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}$, padronizando-o teremos o resíduo “studentizado” r_{sj} , que tem a forma

$$r_{sj} = \frac{r(y_j, \hat{\pi}_j)}{\sqrt{1 - h_j}}, \quad (4.5)$$

e que considera a presença de observações com alto valor de h_j . Um valor grande do resíduo de Pearson num ponto com alto h_j , resultará num altíssimo valor do resíduo “studentizado”. Observamos ainda que este resíduo informa se o valor de y_j difere significativamente do esperado $m_j \hat{\pi}_j$. Já o termo h_j é um indicador de como a variância do resíduo é influenciada pela j -ésima observação.

Jennings (1986) discute a dependência da distribuição dos resíduos d_j nos valores de $E(y|\mathbf{x}_j) = \pi(\mathbf{x}_j)$, fato que não ocorre com os resíduos usuais no modelo de regressão linear. Tal dependência fica evidente na tabela apresentada pelo autor, que mostra os valores das médias de d_j e d_j^2 para vários

valores de $\pi(\mathbf{x}_j)$. Verifica-se que $E(d_j) \neq 0$ de modo geral e $E(d_j^2)$ não é constante, sendo que ambos dependem de $\pi(\mathbf{x}_j)$. Como consequência, características exibidas pelos gráficos de $d(y_j, \hat{\pi}_j)$ dependeriam do valor de $\pi(\mathbf{x}_j)$.

Usando a aproximação

$$\frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \approx \sqrt{2} \sqrt{y_j \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j) \ln\left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)}\right)},$$

Andersen (1991) mostra que $\text{var}(d_j) = 1 - h_j$. Este fato sugere o uso do resíduo componente do desvio padronizado proposto por Paula (1999),

$$t_j = \pm \sqrt{\frac{2}{1 - h_j}} \left\{ y_j \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j) \ln\left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)}\right) \right\}^{1/2},$$

com o mesmo sinal de $y_j - m_j \hat{\pi}_j$. Para $y_j = 0$ ou $y_j = m_j$, utiliza-se respectivamente os valores

$$t_j = - \frac{\sqrt{2m_j |\ln(1 - \hat{\pi}_j)|}}{\sqrt{1 - h_j}}$$

e

$$t_j = \frac{\sqrt{2m_j |\ln(\hat{\pi}_j)|}}{\sqrt{1 - h_j}}.$$

Antigamente, medidas de diagnóstico na regressão logística eram obtidas através do uso de programas de regressão linear, devido a formulação (4.2). Atualmente praticamente todos os pacotes computacionais estatísticos tradicionais têm já implantado em sua rotina o tópico de diagnóstico em regressão logística, facilitando assim o cálculo das medidas de diagnóstico. Encontram-se no apêndice programas para obtenção das estatísticas de diagnóstico em alguns pacotes estatísticos.

Outras medidas de diagnóstico são construídas para verificar o efeito da deleção de todos os indivíduos com um particular vetor de variáveis ex-

plicativas nas estimativas dos coeficientes $\beta_0, \beta_1, \dots, \beta_p$ e também nos valores das estatísticas χ^2 e D dadas em (3.2) e (3.4) respectivamente.

A análise da mudança no valor das estimativas dos coeficientes é análoga ao estudo da medida proposta por Cook (1977) para a regressão linear. Para verificar a influência desses pontos retirados, calcula-se a diferença padronizada entre $\hat{\beta}$ e $\hat{\beta}_{-j}$, em que $\hat{\beta}$ e $\hat{\beta}_{-j}$ representam os respectivos estimadores de máxima verossimilhança de β usando todas as observações e excluindo os m_j elementos com vetor de variáveis explicativas $\mathbf{x}_j$. A padronização é feita através da estimativa da matriz de covariância de $\hat{\beta}$ e a medida é dada por

$$\Delta\hat{\beta}_j = (\hat{\beta} - \hat{\beta}_{-j})'(X'VX)(\hat{\beta} - \hat{\beta}_{-j}).$$

Pregibon (1981), utilizando uma aproximação linear, mostra que esta medida se reduz a

$$\Delta\hat{\beta}_j \approx \frac{r^2(y_j, \hat{\pi}_j)h_j}{(1 - h_j)^2} = \frac{r_{sj}^2 h_j}{1 - h_j}. \quad (4.6)$$

O autor mostra também que $\Delta\hat{\beta}_j$ pode ser obtida a partir de uma classe geral de medidas oriundas da abordagem de curvas de influência para os estimadores de mínimos quadrados de β em modelos de regressão linear (Cook e Weisberg, 1982). Esta classe de medidas, específicas para estimadores de mínimos quadrados, pode ser utilizada devido à formulação dada em (4.2).

Novamente aqui, podemos observar a grande influência de um ponto com alto valor de h_j e do resíduo de Pearson.

Usando uma aproximação similar, verifica-se que, de modo geral, o valor da estatística qui-quadrado de Pearson decresce devido à retirada de

elementos com vetor de variáveis explicativas $\mathbf{x}_j$ e que esse decréscimo vale

$$\Delta X_j^2 \approx \frac{r^2(y_j, \hat{\pi}_j)}{1 - h_j} = r_{sj}^2,$$

coincidindo com o quadrado do valor do resíduo de Pearson “studentizado”, também utilizado para avaliar a presença de observações aberrantes.

De modo análogo, é calculada a quantidade ΔD_j para avaliar a alteração no desvio. Verifica-se que esta medida é dada por

$$\Delta D_j \approx d_j^2 + \frac{r^2(y_j, \hat{\pi}_j)h_j}{1 - h_j}.$$

Pregibon (1981) acrescenta que ΔD_j tem assintoticamente distribuição qui-quadrado com 1 grau de liberdade. Esta informação pode ser utilizada na detecção de pontos discrepantes.

Estas estatísticas de diagnóstico são que acabamos de apresentar são conceitualmente atraentes. Elas permitem identificar vetores de variáveis independentes que tornam o ajuste do modelo pobre, no sentido de fornecer grandes valores de ΔX_j^2 ou ΔD_j). Além disso permitem detectar elementos que têm grande influência nas estimativas dos parâmetros (grandes valores de $\Delta \hat{\beta}_j$). Uma vez identificados estes elementos, pode-se decidir sobre sua manutenção ou não na análise.

Consideraremos inicialmente a medida ΔX_j^2 . Esta medida será pequena quando y_j e $m_j \hat{\pi}(\mathbf{x}_j)$ são próximos. Isto é mais provável de acontecer quando $y_j = 0$ e $\hat{\pi}(\mathbf{x}_j) < 0,1$ ou $y_j = m_j$ e $\hat{\pi}(\mathbf{x}_j) > 0,9$. Similarmente, ΔX_j^2 será maior quanto mais y_j for afastado de $\hat{\pi}(\mathbf{x}_j)$ e isto ocorreria se tivermos o valor de $y_j = 0$ e $\hat{\pi}(\mathbf{x}_j) > 0,9$ ou $y_j = m_j$ e $\hat{\pi}(\mathbf{x}_j) < 0,1$. Este caso, em geral, não apresenta um grande valor de $\Delta \hat{\beta}_j$ pois $\Delta \hat{\beta}_j = \Delta X_j^2 h_j / (1 - h_j)$

e quando $\hat{\pi}(\mathbf{x}_j) < 0,1$ ou $\hat{\pi}(\mathbf{x}_j) > 0,9$, h_j é próximo de zero. A medida de influência $\Delta\hat{\beta}_j$ será grande quando ambos ΔX_j^2 e h_j forem no mínimo moderados. Isto é mais provável de ocorrer quando $0,1 < \hat{\pi}(\mathbf{x}_j) < 0,3$, ou $0,7 < \hat{\pi}(\mathbf{x}_j) < 0,9$, que são os intervalos em que h_j é grande. Na região em que $0,3 < \hat{\pi}(\mathbf{x}_j) < 0,7$, em geral ΔX_j^2 ou h_j não são grandes.

A Tabela 4.1 apresentada em Hosmer e Lemeshow (1989) mostra um sumário dessas observações. Segundo os autores esta tabela resume o que pode ser esperado, não o que exatamente vai acontecer numa particular situação. Por esta razão, deve ser usada apenas como um guia para futuras análises das estatísticas de diagnóstico.

Tabela 4.1 - Valores prováveis das estatísticas de diagnóstico ΔX_j^2 , $\Delta\hat{\beta}_j$ e h_j em cada uma das cinco regiões definidas pelas probabilidades estimadas.

Medida de diagnóstico	Probabilidades estimadas				
	0,0 - 0,1	0,1 - 0,3	0,3 - 0,7	0,7 - 0,9	0,9 - 1,0
ΔX_j^2	grande	moderado	moderado	moderado	grande
$\Delta\hat{\beta}_j$	ou pequeno		para pe- queno		ou pequeno
h_j	pequeno	grande	moderado	grande	pequeno
	pequeno	grande	moderado	grande	pequeno
			para pe- queno		

Paula (1999, [pg.119]) sugere que gráficos de ΔX_j^2 e $\Delta\hat{\beta}_j$ contra as pro-

habilidades ajustadas $\hat{\pi}$ podem ser informativos a respeito do posicionamento dos pontos discrepantes com relação às probabilidades estimadas.

Hosmer e Lemeshow (1989) destacam que na análise da qualidade do ajuste nos modelos de regressão logística, devemos confiar mais na avaliação gráfica, já que as distribuições das estatísticas de diagnóstico sob a hipótese de que o modelo se ajusta são conhecidas apenas em casos muito particulares. Como exemplo, citam o resíduo de Pearson, $r(y_j, \hat{\pi}_j)$, que é frequentemente comparado com quantis da distribuição normal padrão. No entanto, este procedimento seria válido somente quando m_j é suficientemente grande para justificar que a distribuição normal produz uma adequada aproximação da distribuição binomial. Conforme salientado por Jennings (1986), se $m_j = 1$ então r_j terá somente dois valores possíveis e assim, a aproximação normal não é razoável. Além disso, quando as variáveis explicativas de um conjunto de dados são contínuas, em geral, L será igual a n e m_j igual a 1. Como consequência, resultados que admitem m_j grandes não podem ser aplicados.

Na próxima seção, calcularemos as medidas de diagnóstico apresentadas para o exemplo do Capítulo 3. Métodos gráficos para a avaliação do ajuste do modelo de regressão logística serão introduzidos no próximo capítulo.

4.3 Cálculo das medidas de diagnóstico para o exemplo da Seção 3.3

No exemplo da Seção 3.3 havíamos concluído que o modelo reduzido, modelo com a variável renda familiar (X_1) foi o melhor na explicação da compra

de um carro novo para o próximo ano e também avaliamos a qualidade do modelo ajustado que se mostrou adequado. Vamos agora refazer a análise calculando as medidas de diagnóstico mencionadas na seção anterior, para o modelo completo, isto é, com as variáveis X_1 e X_2 .

Para o cálculo dessas medidas, vamos utilizar os pacotes estatísticos SPSS 8.0, Minitab 13 e S-plus 2000. Encontram-se no apêndice os respectivos programas (ou caminhos) que auxiliarão nos cálculos.

O programa SPSS 8.0 não fornece diretamente todas essas medidas, mas elas podem ser calculadas como: “Influence Cook’s” ($\Delta\hat{\beta}_j$), “Influence Leverage values” (h_j), “Residuals Standardized” (r_{sj}), “Residuals Desvio” ($d(y_j, \pi_j)$). No Minitab, encontram-se todos os resíduos mencionados e com fácil acesso. No programa S-plus 2000 somente os resíduos de Pearson e de desvio são diretos e os demais resíduos devem ser calculados na Seção comando, conforme apêndice.

Inicialmente calculamos as probabilidades estimadas de acordo com o modelo completo e as correspondentes medidas de diagnóstico. Na Tabela 3.2, encontram-se as estimativas dos parâmetros do modelo completo e na Tabela 4.2 apresentamos as medidas de diagnóstico referentes ao modelo completo.

Tabela 4.2 - Valores das medidas de diagnóstico para o exemplo da Seção 3.3 referente ao modelo completo.

observação	$r(y_j, \hat{\pi}_j)$	r_{sj}	$d(y_j, \pi_j)$	ΔX_j^2	ΔD_j	$\Delta \hat{\beta}_j$	h_j	$\hat{\pi}_j$
1	-0,68	-0,69	-0,87	0,48	0,78	0,02	0,05	0,32
2	-0,78	-0,81	-0,98	0,65	1,00	0,05	0,08	0,38
3	0,77	0,81	0,97	0,65	0,99	0,07	0,10	0,63
4	-0,76	-0,82	-0,95	0,67	1,00	0,10	0,15	0,37
5	-0,72	-0,74	-0,92	0,55	0,87	0,03	0,06	0,34
6	0,79	0,86	0,99	0,75	1,09	0,14	0,12	0,61
7	0,37	0,39	0,50	0,15	0,27	0,02	0,07	0,88
8	0,66	0,72	0,85	0,51	0,80	0,09	0,19	0,70
9	-1,65	-1,75	-1,62	3,05	2,96	0,38	0,10	0,73
10	0,26	0,28	0,36	0,08	0,14	0,01	0,04	0,94
11	0,38	0,40	0,52	0,16	0,29	0,02	0,09	0,87
12	-0,85	-0,91	-1,04	0,82	1,19	0,11	0,10	0,42
13	-0,55	-0,57	-0,73	0,32	0,55	0,02	0,06	0,23
14	0,83	0,87	1,03	0,75	1,11	0,06	0,09	0,59
15	-0,70	-0,72	-0,89	0,51	0,82	0,02	0,05	0,33
16	-0,58	-0,62	-0,76	0,38	0,63	0,06	0,17	0,25
17	-0,42	-0,47	-0,41	0,22	0,21	0,05	0,10	0,64
18	-0,39	-0,41	-0,54	0,17	0,30	0,01	0,07	0,13
19	1,76	1,82	1,68	3,29	3,03	0,23	0,06	0,24
20	-1,19	-1,36	-1,33	1,85	2,20	0,58	0,23	0,58
21	-0,57	-0,59	-0,75	0,35	0,59	0,02	0,05	0,25
22	1,33	1,36	1,43	1,85	2,11	0,08	0,04	0,36
23	1,12	1,15	1,28	1,32	1,69	0,06	0,05	0,44
24	-0,43	-0,45	-0,59	0,20	0,36	0,02	0,06	0,16

continuação								
observação	$r(y_j, \hat{\pi}_j)$	r_{sj}	$d(y_j, \pi_j)$	ΔX_j^2	ΔD_j	$\Delta \hat{\beta}_j$	h_j	$\hat{\pi}_j$
25	-0,52	-0,53	-0,69	0,28	0,49	0,02	0,06	0,21
26	1,94	2,01	1,77	4,05	3,41	0,30	0,06	0,21
27	-0,50	-0,52	-0,67	0,27	0,46	0,02	0,06	0,20
28	-0,79	-0,85	-0,99	0,71	1,06	0,09	0,09	0,39
29	2,79	2,91	2,08	8,47	5,03	0,74	0,10	0,11
30	-1,34	-1,41	-1,44	1,99	2,25	0,21	0,10	0,64
31	-0,47	-0,48	-0,63	0,23	0,41	0,02	0,06	0,18
32	0,81	0,87	1,00	0,76	1,11	0,12	0,15	0,60
33	-0,22	-0,23	-0,31	0,05	0,10	0,00	0,14	0,05

Notamos que, com exceção de h_j , todas as medidas de diagnóstico para as observações de números 26 e 29 têm valores altos em comparação às demais. A observação de número 29 correspondente a uma família com renda baixa e com o automóvel mais velho de apenas 2 anos mas que adquiriu um automóvel novo. Já a de número 26 também tem baixa renda e adquiriu um automóvel novo naquele ano, mas seu automóvel mais velho é mais antigo, tem 4 anos. Então são famílias com renda baixa mas houve a compra de um automóvel. Percebemos assim que a renda é um fator importante juntamente com a idade do automóvel mais velho.

Detectados esses dois pontos aberrantes, eles foram retirados e reajustamos o modelo completo, refazendo a análise. Inicialmente optou-se por retirar o elemento de número 29 ($n=32$) e após retiramos também a observação 26 ($n=31$). Os resultados encontram-se na Tabela 4.3.

Tabela 4.3 - Resumo do ajuste dos modelos

Observação retirada	modelo	função desvio (D)	g.l.	estatística G^*
nenhuma	completo (I)	36,69	30	
29	completo (II)	31,57	29	
29	reduzido só com X_1	36,49	30	4,92
29	reduzido só com X_2	43,20	30	11,63
29 e 26	completo (III)	27,22	28	
29 e 26	reduzido só com X_1	32,64	29	5,42
29 e 26	reduzido só com X_2	41,38	29	14,16

Analisando a Tabela 4.3 notamos que G^* é maior que 3,841, que é o valor crítico da distribuição qui-quadrado com 1 grau de liberdade e um nível de significância de 5%, indicando assim que cada uma das variáveis é significativa para o modelo.

Com a eliminação das duas observações, 26 e 29, temos as novas estimativas para β_1 e β_2 , respectivamente, 0,1129 (desvio-padrão estimado = 0,0409) e 1,0372 (desvio-padrão estimado = 0,4958) com variação de $|1 - \frac{0,1129}{0,0677}|100\% = 67\%$ e $|1 - \frac{1,0372}{0,5986}|100\% = 73\%$, indicando uma alta influência destas observações nas estimativas dos parâmetros. O desvio do modelo reduziu para $D = 27,22$ (28 graus de liberdade), variação de $(1 - \frac{27,22}{36,69})100\% = 26\%$, sendo que somente com a retirada da observação de número 29 havia uma variação de 14%. Logo, escolheremos o modelo (III) como modelo final.

Vamos avaliar a qualidade desse novo modelo, encontrando a Estatística C , com base nos cálculos presentes na Tabela 4.4.

Tabela 4.4 - Quantidades para o cálculo de C .

Grupo	Observações	O_i	n'_i	$\bar{\pi}_i$
1	31,18,24,29	0	4	0,040
2	26,25,13,16	1	4	0,090
3	19,21,1,15	1	4	0,150
4	5,4,22,2	1	4	0,235
5	27,12,23,14	3	4	0,398
6	6, 20, 30, 3	3	4	0,633
7	17,28,8,9	2	4	0,735
8	11, 7,10	3	3	0,967

Desta forma, obtemos

$$C = \frac{(0 - 4.0, 040)^2}{4.0, 040.(1 - 0, 040)} + \dots + \frac{(3 - 3.0, 967)^2}{3.0, 967.(1 - 0, 967)} = 5, 2861.$$

Como sob a hipótese de adequação do modelo, C tem uma distribuição qui-quadrado com $(g - 2)$ graus de liberdade, obtemos o valor crítico de 12, 6, supondo um nível de significância de 5%. Logo concluímos que o modelo *III* é adequado.

Observamos então a importância da análise de diagnóstico, pois pode, entre outras coisas levar à retirada de observações, e também à seleção de outro modelo mais adequado à realidade dos dados. Métodos gráficos, juntamente com as medidas de diagnóstico, são fundamentais, podem e devem ser usados para examinar falhas no modelo ajustado e também detectar pontos afastados dos demais. Tais métodos serão descritos no próximo capítulo.

Capítulo 5

Análise Gráfica para diagnóstico

5.1 Introdução

Em regressão linear, é comum a construção de gráficos para diagnóstico e que são muito úteis para detectar observações aberrantes.

Para regressão logística, devemos tomar o mesmo cuidado, isto é, examinar as relações entre a resposta e as variáveis explanatórias para verificar se melhorias podem ser feitas, através dos gráficos de diagnósticos. Com isso podemos ajudar a detectar um modelo mais apropriado aos dados.

Esta análise envolve a construção de vários tipos de gráficos de diagnóstico para verificar diferentes aspectos do modelo estimado, pois cada tipo de gráfico identifica um particular desvio do modelo. No momento da

decisão devemos levar em consideração somente quando vários gráficos de resíduos identificar em sempre as mesmas observações.

Existe uma grande quantidades de gráficos possíveis, então restringimos a atenção aos mais significativos para a análise de diagnóstico no modelo de regressão logística, por exemplo, os gráficos ΔX_j^2 , ΔD_j e $\Delta \hat{\beta}_j$ contra $\hat{\pi}_j$. Esses gráficos fornecem uma avaliação da contribuição de cada ponto nos valores das estatísticas de diagnóstico em função das probabilidades estimadas. Por exemplo, grandes valores de ΔX_j^2 sugerem que as observações são candidatas a pontos aberrantes. Com o gráfico de ΔD_j podemos avaliar a alteração no desvio e o gráfico de $\Delta \hat{\beta}_j$ mostra a influência de cada observação nas estimativas dos coeficientes.

Além desses, existem gráficos mais específicos, que serão descritos a seguir.

5.2 Gráficos Específicos

Landwehr, Pregibon e Shoemaker (1984) propõe e discutem três métodos gráficos que auxiliam na avaliação do ajuste do modelo de regressão logística. Tais métodos são generalizações de gráficos já existentes adaptados para levar em consideração o caráter binário da variável resposta. A seguir apresentaremos esses métodos.

Gráfico do Desvio Local Médio Os autores comentam que o desvio é uma medida global que avalia a “discordância” entre os dados observados e

os dados ajustados através do modelo. No entanto, basear toda uma análise de ajuste num único número seria inadequado. Por isso, sugerem que o valor de D seja particionado em duas componentes, uma correspondente ao erro puro e a outra, devido à falta de ajustamento. Quando existem réplicas, a componente do erro puro no desvio é facilmente obtida. Na inexistência de réplicas perfeitas, utiliza-se um método de agrupamento com base nas variáveis independentes. O procedimento proposto é o seguinte

1. Dividir as n observações em K grupos com n_k observações em cada um deles, $k = 1, 2, \dots, K$. Não há limite para o tamanho dos grupos. No entanto, como serão feitos ajustes locais, é melhor o uso de grupos pequenos e de pontos homogêneos segundo as variáveis independentes. Grupos de tamanho $n_k = 1$ podem ser omitidos, pois não contribuem com nenhuma informação para a variação local.
2. Formar uma matriz n por K , em que Z_{ik} , o elemento de ordem i, k dessa matriz será 1 ou 0 conforme a i -ésima observação está ou não no k -ésimo grupo.
3. Calcular a estimativa local $\hat{p}^L$ através do ajuste do modelo $\text{logito}(p^L) = Z\gamma + X\beta$. Este ajuste admite diferentes logitos para cada um dos K grupos. Se $p_k^L = \pi_k(Y = 1|X = x)$ é a probabilidade de $Y = 1$ dado $X = x$, no k -ésimo grupo, então estamos admitindo que $\log(\frac{p_k^L}{1-p_k^L}) = \gamma_k + x\beta$, $k = 1, 2, \dots, K$. Quando dispomos de uma única variável independente, o procedimento é equivalente a ajustar K retas paralelas com diferentes interceptos para os logitos.
4. Usar o modelo ajustado no item 3 para encontrar a contribuição de cada observação no desvio local, obtendo $d(\hat{p}_{kj}^L; y_{kj})$, e somar os desvios de cada

um dos grupos, obtendo $D_k^L = \sum d(\hat{p}_{kj}^L; y_{kj})$.

5. Reordenar os K grupos de maneira que $S_1 \leq S_2 \leq \dots \leq S_K$, em que S_i é uma medida de heterogeneidade do grupo. Por exemplo, se utilizarmos um método de agrupamento hierárquico, S_i poderia ser a altura da árvore na qual o grupo é formado. Assim, os pontos do grupo 1 são mais similares que os do grupo 2, que são mais similares que os do grupo 3 e assim por diante.
6. Calcula-se uma estimativa do erro puro aproximado, $\bar{D}^L(t) = \frac{\sum_{k=1}^t D_k^L}{\sum_{k=1}^t (N_k - 1)}$, $t = 1, 2, \dots, K$. Os valores $\bar{D}^L(t)$ representam a média dos desvios locais dos t grupos mais homogêneos.
7. Constrói-se o gráfico de $Y = \bar{D}^L(t)$ contra seu número de graus de liberdade $X = \sum_{k=1}^t (N_k - 1)$, para $t = 1, \dots, K$. Constrói-se no mesmo gráfico a reta $Y = \bar{D}$, em que $\bar{D} = \frac{D(\hat{p}; y)}{n-p}$ é o desvio global médio e observa-se a posição dos pontos $(\sum_{k=1}^t (N_k - 1), \bar{D}^L(t))$ em relação a essa reta.

Este procedimento foi denominado Gráfico Desvio Local. A falta de ajuste estará caracterizada quando a reta do item 7, que fornece uma medida de falta de ajuste dos dados com relação ao modelo de regressão logística, está consistentemente acima dos pares (X, Y) obtidos através do ajuste dos modelos locais. Quando os pontos se distribuem em torno da reta $Y = \bar{D}$, temos indicações de modelo apropriado.

Gráfico de Probabilidades Empíricas Landwehr, Pregibon e Shoemaker (1984) apresentam o Gráfico de Probabilidades Empíricas, que é recomendado para avaliar o ajuste do modelo escolhido.

Em regressão linear, o gráfico normal de probabilidades é uma ferramenta útil para detectar “outliers” e examinar a suposição de normalidade

dos erros. A idéia básica seria colocar os valores dos resíduos padronizados ordenados no eixo das ordenadas contra os correspondentes quantis ordenados da distribuição normal padrão no eixo das abcissas. Quando o modelo assumido ajusta corretamente os dados, este gráfico aproxima-se de uma reta. Uma configuração não linear desses pontos pode resultar de uma distribuição de erros não normal ou da presença de “outliers”.

Para regressão logística, o gráfico análogo é o dos resíduos contra os quantis de sua distribuição. Um resíduo natural nesse caso é o desvio $d_j = d(y_j, \hat{\pi}_j)$, padronizado pelo seu erro padrão. Argumentos intuitivos sugerem o uso da distribuição qui-quadrado com 1 grau de liberdade como referência, mas esse fato não foi apoiado nos estudos empíricos dos autores. Por esse motivo, propõe o uso do resíduo $y_j - \hat{\pi}_j$ e a obtenção da distribuição de referência para os resíduos através de simulação. O procedimento é baseado em $\hat{\pi}$ obtido sob o modelo proposto, gerando dessa forma a distribuição que o resíduo teria se tal modelo fosse o correto.

Os passos básicos para a construção desse gráfico são:

1. Obter os resíduos $r_j = y_j - \hat{\pi}_j$, $j = 1, \dots, n$ do modelo ajustado $\text{logito}(\hat{\pi}) = \mathbf{X}\hat{\beta}$.
2. Ordenar r_j , gerando os valores $r_{(j)}$, $j = 1, \dots, n$.
3. Simular dados y^* do modelo ajustado, isto é, obter Y_j^* com distribuição Binomial $(1, \hat{\pi}_j)$, $j = 1, \dots, n$.
4. Ajustar o modelo $\text{logit}(\hat{\pi}^*) = \mathbf{X}\hat{\beta}^*$ para os y^* .
5. Calcular os resíduos $r_j^* = y_j^* - \hat{\pi}_j^*$ em que $\hat{\pi}_j^*$ é obtido através do modelo do passo 4. Ordenar os valores dos r_j^* .
6. Repetir os passos de 3 a 5 m vezes independentemente. Seja $r_{i(j)}^*$ o j-ésimo

resíduo ordenado na i -ésima interação, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$.

7. Para cada $j = 1, 2, \dots, n$, calcula-se um valor típico dos resíduos ordenados nas m réplicas, por exemplo, $T_{(j)} = \text{med}_i\{r_{i(j)}^*\}$, $i = 1, \dots, m$, em que $\text{med}_i\{r_{i(j)}^*\}$ representa a mediana das m observações.

8. Constroi-se o gráfico dos resíduos originais ordenados contra os resíduos típicos ordenados obtidos na simulação, isto é, o gráfico dos pares $(T_{(j)}, r_{(j)})$, $j = 1, \dots, n$.

9. Obtem-se o gráfico com os limites superior e inferior de confiança para cada um dos $T_{(j)}$ utilizando-se os quantis dos m valores $\{r_{i(j)}^*\}$, $i = 1, 2, \dots, m$, de acordo com o coeficiente de confiança desejado. Assim, por exemplo, se $m = 25$ e tomarmos os limites inferior e superior como respectivamente $r_{2(j)}^*$ e $r_{24(j)}^*$, estaríamos utilizando um coeficiente de confiança de $\frac{22}{26}100\% = 84,6\%$.

O procedimento descrito no passo 9 corresponde à construção dos chamados “envelopes” de simulação. De acordo com Atkinson (1985, [pg.34]), um problema na análise de gráficos em que o comportamento esperado é uma reta seria a verificação de quão bem os pontos se aproximam da mesma, ou seja, se as irregularidades apresentadas representam algo além de flutuações aleatórias.

Uma dificuldade adicional com gráficos que envolvem distribuições acumuladas é o fato da ordenação das observações introduzir dependência entre seus valores. Com isso, alguns desvios do esperado forneceria uma falsa impressão de inadequação. Já a simulação proporciona um comportamento de referência das flutuações e os envelopes uma idéia do que seria razoável esperar. →

Quando o modelo que estamos ajustando é adequado aos dados, o gráfico dos pontos $(T_{(j)}, r_{(j)})$, $j=1, 2, \dots, n$ se aproxima de uma reta pela origem com inclinação um. Além disso, através desse gráfico são construídos empiricamente intervalos de confiança para os resíduos, ou seja, as “bandas de confiança”, sendo que as observações não devem ultrapassar esses limites e deverão ter um comportamento aleatório em toda sua extensão. Se isso acontecer, podemos dizer que o modelo adotado está bem ajustado, caso contrário, ou seja, se os pontos ultrapassarem as bandas de confiança, provavelmente serão pontos com algum tipo de discrepância e deverão ser analisados ou o modelo não é adequado.

Paula (1999) fornece o programa para a construção do gráfico de probabilidade empíricas para modelos lineares generalizados. Apresentamos no apêndice o procedimento para modelos de regressão logística.

Gráfico de Resíduos Parciais O terceiro gráfico apresentado por Landher, Pregibon e Schoemaker (1984) é uma adaptação do conhecido Gráfico de Resíduos Parciais para análise do ajuste do modelo de regressão logística. Na regressão linear, este gráfico é útil para avaliar a necessidade de introdução de funções não lineares das variáveis independentes no modelo.

Vamos supor que ajustou-se um modelo da forma

$$E(y) = X\beta + \gamma z,$$

quando o verdadeiro modelo é

$$E(y) = X\beta + f(z),$$

com f desconhecida. O resíduo parcial do modelo ajustado é definido como $r_{par} = r_{y|x,z} + z\hat{\gamma}$, em que $r_{y|x,z}$ é o resíduo usual do modelo $E(y) = X\beta + \gamma z$ e $\hat{\gamma}$ é a estimativa de γ neste modelo.

Se f é linear, $E(r_{par}) = z\gamma$. Se f é não linear, o gráfico de r_{par} contra z pode auxiliar na determinação de f . Embora nesse caso $E(r_{par}) \approx f(z)$ a tendência exibida no gráfico de resíduos parciais fornecerá uma indicação da possível forma de f .

A definição de resíduos parciais para modelos de regressão logística utiliza a formulação do estimador de máxima verossimilhança de β dada em (4.2),

$$\hat{\beta}(t+1) = (X'V(t)X)^{-1}X'V(t)y^*(t)$$

com

$$y^*(t) = X\hat{\beta}(t) + V^{-1}(t)r(t)$$

e

$$r(t) = y - \hat{\pi}(t).$$

Se z é a variável independente contra a qual desejamos construir o gráfico de resíduos parciais no modelo de regressão logística, sejam,

$$\text{logito}(\pi) = X\beta + z\gamma$$

e $\hat{\beta}$, $\hat{\gamma}$, $\hat{\pi}$ e V os estimadores de máxima verossimilhança de β , γ , π e V nesse modelo.

Encarando y^* como uma observação, por analogia ao modelo de regressão linear, o resíduo parcial na regressão logística é definido como

$$y^*(t) - X\hat{\beta}(t) + \hat{\gamma}z = V^{-1}(t)r(t) + \hat{\gamma}z.$$

Lembrando a forma geral da matriz V , temos

$$r_{par} = \frac{y - \hat{\pi}}{\hat{\pi}(1 - \hat{\pi})} + z\hat{\gamma}$$

e o gráfico de resíduos parciais é o gráfico de r_{par} contra z .

Devido a natureza binária da variável y , o gráfico de resíduos parciais consistirá de duas nuvens de pontos separadas, uma correspondente a $y = 0$ e a outra a $y = 1$. Por esse motivo os autores usam o método de suavização proposto por Cleveland (1979) com o objetivo de facilitar a determinação da tendência exibida por esse gráfico.

Dado um conjunto de pares (x_i, y_i) , $i = 1, 2, \dots, n$ dispostos num gráfico, o procedimento de construção de novos pares $(x_i, \hat{y}_i)$, em que $\hat{y}_i$ representa uma estimativa de alguma medida de locação da distribuição de y dado o valor da variável $X = x_i$ é denominado suavização ou alisamento. Admite-se nesse caso que $y_i = g(x_i) + \epsilon_i$ em que g é uma função de x e ϵ_i uma variável aleatória com média 0 e variância 1. Nesse contexto, $\hat{y}_i$ seria uma estimativa de $g(x_i)$.

Cleveland (1979) apresenta o método de suavização de Regressão Robusta Localmente Ponderada. Tal método, além de proceder à suavização do gráfico, tem a vantagem adicional da robustez, ou seja, fornece uma proteção com relação a pontos discrepantes.

Para a obtenção de $\hat{y}_i$ é utilizada a função de pesos tricúbica

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{para } |x| < 1 \\ 0 & \text{para } |x| \geq 1. \end{cases}$$

São fixados os valores $0 < f \leq 1$ e r , o inteiro mais próximo de fn . Para cada x_i , o peso $w_k(x_i)$, da k -ésima observação, $k = 1, \dots, n$, no cálculo de $\hat{y}_i$ é dado por

$$w_k(x_i) = W\left(\frac{x_k - x_i}{h_i}\right), \quad (5.1)$$

em que h_i é a distância entre x_i e o r -ésimo vizinho mais próximo de x_i .

Como $W(x) = 0$ para $|x| \geq 1$, somente as r observações mais próximas de x_i contribuirão para o cálculo de $\hat{y}_i$ e como W é uma função não crescente, as mais próximas terão peso maior.

O procedimento de regressão localmente ponderada robusta é formado pelos seguintes passos:

1. Para cada observação i , $i = 1, 2, \dots, n$, ajusta-se modelo de regressão polinomial de grau d por mínimos quadrados ponderados com pesos dados por $w_k(x_i)$. Obtem-se então

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \dots + \hat{\beta}_d x_i^d,$$

em que $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_d$ minimizam

$$\sum_{k=1}^n w_k(x_i) (y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k - \dots - \hat{\beta}_d x_k^d)^2.$$

Esta operação consiste no método de Regressão Localmente Ponderada e pode ser executado isoladamente. Se acrescida dos próximos passos, fica definido o procedimento de Regressão Localmente Ponderada Robusta.

2. Calcula-se os resíduos do modelo ajustado $e_i = y_i - \hat{y}_i$ e $s = \text{mediana } |e_i|$, $i = 1, 2, \dots, n$.

Com o objetivo de robustez define-se os pesos

$$\delta_k = B\left(\frac{e_k}{6s}\right)$$

em que

$$B(x) = \begin{cases} (1 - x^2)^2, & \text{para } |x| < 1 \\ 0 & \text{para } |x| \geq 1 \end{cases}$$

3. Obtem-se um novo valor de $\hat{y}_i$, para cada i , através do ajuste de um polinômio de grau d por mínimos quadrados ponderados com peso $\delta_k w_k(x_i)$ no ponto (x_k, y_k) .
4. Repete-se os passos 2 e 3 num total de t vezes. O último $\hat{y}_i$ obtido será o valor ajustado após suavização.

Os passos 2 a 4 do procedimento correspondem à obtenção de um suavizador robusto, no sentido de que uma pequena fração de pontos discrepantes não distorceria os resultados. O autor apresenta ainda métodos para a escolha de f , d , t e W .

Finalizado o procedimento, analisa-se o gráfico dos pontos $(x_i, \hat{y}_i)$ obtidos após suavização.

Conforme comentamos anteriormente, Landwehr, Pregibon e Shoemaker (1984) aplicam o procedimento de suavização para os resíduos parciais do modelo de regressão logística. Os autores utilizam apenas o passo 1 do procedimento ($t = 0$) correspondendo à regressão localmente ponderada não robusta.

Em um de seus exemplos, cem observações foram geradas de acordo com o modelo $\text{logito}(\pi) = -1 + x_1 + x_2 + 2x_2^2$, em que x_1 assume os valores

$-1,00$, $0,02$ e $1,00$ e os valores de x_2 foram selecionados independentemente identicamente distribuído a partir de uma distribuição uniforme no intervalo $[-1, 1]$. Assumindo que não conhecemos o modelo, ajustou-se o modelo linear em x_1 e x_2 :

$$\text{logito}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

com desvio $D = 131,95$ para 98 graus de liberdade, e foram feitos gráficos de resíduos parciais em função de x_1 e x_2 . Estes gráficos estão exibidos nas Figuras 5.1 e 5.2. Nelas podemos notar as duas nuvens de pontos formadas pelos resíduos parciais e a curva correspondendo ao alisamento proposto por Cleveland (1979).

A Figura 5.1 sugere basicamente uma dependência linear em x_1 . O alisamento da Figura 5.2, no entanto, sugere que a dependência em x_2 é não linear e que um termo adicional tal como x_2^2 ou $|x_2|$ deveria ser incluído no modelo.

Figura 5.1: Gráfico de resíduos parciais em função da variável X_1 ,
Landwehr, Pregibon e Shoemaker (1984).

Figura 5.2: Gráfico de resíduos parciais em função da variável X_2 ,
Landwehr, Pregibon e Shoemaker (1984).

Fowlkes (1987) utiliza técnicas de alisamento para analisar o ajuste dos modelos de regressão logística. Segundo o autor, neste caso, os gráficos de diagnóstico são fortemente afetados pelo caráter binário da variável resposta. Como exemplo, comenta que o gráfico do resíduo de Pearson contra as probabilidades ajustadas, de modo geral, resulta em duas curvas, uma para $y = 0$ e outra para $y = 1$ seja o modelo correto ou não. A proposta do autor consiste em proceder ao alisamento da variável binária Y , ao contrário do sugerido por Landwehr, Pregibon e Shoemaker, que aplica a técnica de alisamento no resíduo. Além disso, o método de alisamento é um pouco diferente,

pois utiliza uma média ponderada das observações ao invés de uma regressão ponderada e não aplica nenhum procedimento que garanta a robustez. O valor alisado, $\hat{y}_i$, associado ao valor x_i da variável independente é dado por

$$\hat{y}_i = \sum_k \left(\frac{w_k y_k}{\sum w_k} \right),$$

com as somas para k abrangendo os $\frac{n}{10}$ elementos mais próximos de y_i segundo o valor de x_i e w_k é a função definida em (5.1).

No caso do modelo de regressão logística múltipla, com

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=1}^p \beta_j x_{ij}, i = 1, 2, \dots, n,$$

o valor $\hat{y}_i$ associado à observação y_i deve ser obtido após alisamento de acordo com as variáveis independentes.

Admitindo que

$$E(\hat{y}_i) = \pi_i$$

e expandindo $\log \frac{\hat{y}_i}{1 - \hat{y}_i}$ em série de Taylor, obtemos

$$\log \frac{\hat{y}_i}{1 - \hat{y}_i} \cong \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \frac{\hat{y}_i - \pi_i}{\pi_i(1 - \pi_i)}.$$

Supondo que x_p é a variável independente z contra a qual desejamos construir o gráfico dos resíduos parciais, então a expressão anterior fica

$$\log \frac{\hat{y}_i}{1 - \hat{y}_i} \cong \sum_{l=1}^{p-1} \beta_l x_{il} + z_i \gamma + \frac{\hat{y}_i - \pi_i}{\pi_i(1 - \pi_i)}.$$

Substituindo os valores dos β_l , γ e π_i por suas estimativas de máxima verossimilhança, o resíduo parcial fica

$$r_i^* = z_i \hat{\gamma} + \frac{\hat{y}_i - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)},$$

em que

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \sum_{l=1}^p \hat{\beta}_l x_{il}}}{1 + e^{\hat{\beta}_0 + \sum_{l=1}^p \hat{\beta}_l x_{il}}}.$$

O gráfico dos resíduos parciais é então construído com os pares (z_i, r_{i*}) , $i = 1, 2, \dots, n$. Conforme comentado anteriormente, caso z esteja linearmente relacionada com $\log \frac{y_i}{1-y_i}$, espera-se que esse gráfico esteja próximo de uma reta com coeficiente angular aproximadamente igual a $\hat{\gamma}$. Não linearidade nesse gráfico seria indicativo da necessidade de uma transformação não linear em z .

No mesmo artigo, o autor comenta sobre a suavização para a construção do Gráfico do Desvio Local Médio. No cálculo dessa medida, é utilizado o valor

$$d_i^S = -2\{y_i \log(\hat{p}_i^S) + (1 - y_i) \log(1 - \hat{p}_i^S)\}$$

em que, $\hat{p}_i^S$ é o valor suavizado da variável y_i de acordo com as p variáveis explicativas x_j , $j = 1, 2, \dots, p$.

5.3 Gráficos para o exemplo

Dando continuidade à análise do exemplo da Seção 3.3, na Figura 5.3 temos o gráfico das probabilidades empíricas para o modelo completo, que sugere um modelo quase adequado.

As Figuras 5.3.a, b e c apresentam respectivamente os gráficos das medianas, limite inferior e limite superior dos resíduos ordenados obtidos na simulação em função dos resíduos do modelo ajustado ordenados.

Apesar de todos os pontos estarem dentro da banda de confiança, notamos uma pequena separação em dois grupos, o que deve sugerir uma análise de diagnóstico mais detalhada.

Figura 5.3: Gráfico das probabilidades empíricas, modelo completo

Apresentamos na Seção 4.2 as medidas de diagnóstico mais utilizadas e na Seção 4.3 o cálculo dessas medidas para o exemplo. Agora ilustraremos através de gráficos essas medidas com o objetivo de uma visualização rápida dos pontos que se distanciam da maioria. Esses gráficos são básicos e fazem parte da saída padrão do pacote estatístico “S-PLUS 2000”.

Nos gráficos das Figuras 5.4 e 5.5, notamos apenas o destaque da observação de número 20, comum a ambos, mas que não se destaca nos demais gráficos. Tal fato não acontece com a observação de número 29, que embora não apresente um alto valor do alavanca, mostra-se discrepante em todos os

demais gráficos.

Figura 5.4: Gráfico do alavanca em função das probabilidades estimadas.

Figura 5.5: Gráfico da distância de Cook $\Delta\hat{\beta}$ em função das probabilidades

estimadas.

Podemos observar também nos gráficos das Figuras 5.6 a 5.9, que as observações de números 26 e 29 apresentam alto valor do resíduo de Pearson e têm grande influência no valor das estatísticas χ^2 e D. Tal fato já tinha sido notado na Seção 4.3.

Figura 5.6: Gráfico resíduo de Pearson padronizado em função das probabilidades estimadas.

Notamos que os gráficos das Figuras 5.7 a 5.9 apresentam o mesmo comportamento, com as mesmas observações em destaque. A curva crescente corresponde às observações para $Y = 0$ e a curva decrescente refere-se a $Y = 1$. Isto é consequência do resultado

$$\Delta X_j^2 \approx \frac{r^2(y_j, \hat{\pi}_j)}{1 - h_j} = \begin{cases} \frac{(1 - \hat{\pi}_j)^2}{1 - h_j} & \text{se } y = 1 \\ \frac{(0 - \hat{\pi}_j)^2}{1 - h_j} & \text{se } y = 0. \end{cases}$$

Figura 5.7: Gráfico de ΔX_j^2 em função das probabilidades estimadas.

Figura 5.8: Gráfico de ΔD_j em função das probabilidades estimadas.

Figura 5.9: Gráfico de ΔX_j^2 com bolhas proporcionais ao tamanho de $\Delta \hat{\beta}$ em função das probabilidades estimadas.

Na Figura 5.10 temos novamente o gráfico das probabilidades empíricas para o modelo sem as observações de números 26 e 29. Podemos notar uma melhor distribuição dos pontos dentro das bandas de confiança, sugerindo um ajuste mais adequado. Lembramos que, conforme comentado no capítulo anterior, as famílias de números 26 e 29 são incomuns pois tendo baixa renda, adquiriram automóvel novo.

Figura 5.10: Gráfico das probabilidades empíricas do modelo sem as observações 26 e 29 em função das probabilidades estimadas.

Diante do exposto, e também da análise de ajuste feita para este modelo através da estatística C (Seção 4.3), concluímos pela manutenção do modelo *III*, sem as observações de números 26 e 29.

No próximo capítulo, repetiremos a análise feita nesse exemplo para um conjunto com dados reais.

Capítulo 6

Exemplo com dados reais

Neste capítulo, aplicaremos algumas das técnicas apresentadas anteriormente a um conjunto de dados reais gentilmente cedido pelo Prof. Dr. José Eluf Neto, do Departamento de Medicina Preventiva da Faculdade de Medicina da Universidade de São Paulo. Tais dados foram coletados em São Paulo e analisados em sua tese de doutorado defendida em 1993 na London School of Hygiene and Tropical Medicine na University of London (Eluf Neto, 1993).

O estudo foi desenvolvido para investigar fatores de risco para câncer invasivo cervical. A amostra continha 217 controles (sem a doença) e 199 casos (com a doença)(os dados encontram-se no final do apêndice). Os casos foram obtidos em quatro hospitais em São Paulo, dentre pacientes com diagnósticos confirmados por histologia e que não tinham nenhum tipo de tratamento iniciado. Os controles eram pacientes de hospitais, mas que nunca tiveram a doença em estudo ou doenças associadas, nem suspeita de qualquer outro tipo de câncer.

O pesquisador teve critérios próprios de inclusão e exclusão no estudo, para os casos e controles. As informações das variáveis foram obtidas através de aplicação de questionário e medidas marcadas pelo Papilomavírus humano (HPV). A variável resposta é a ocorrência (caso, $Y = 1$) ou ausência (controle, $Y = 0$) de câncer no colo do útero e dispunha-se inicialmente de um conjunto de 15 variáveis independentes.

O modelo adotado pelo pesquisador foi de regressão logística. Ressaltamos que em estudos retrospectivos para a regressão logística ser válida é necessário que o procedimento amostral de caso e controle não dependa das variáveis explicativas. Para a seleção das variáveis independentes ou fatores de risco do modelo final utilizou-se o seguinte procedimento. Primeiramente, foram ajustadas regressões logísticas simples para analisar a importância de cada variável independente isoladamente, sempre considerando idade e fator sócioeconômico. Após encontrar as variáveis significativas individualmente foi utilizada a regressão logística múltipla com essas variáveis para verificar conjuntamente quais são os fatores de risco associados a câncer de colo de útero. As variáveis independentes selecionadas são as descritas a seguir:

Variáveis categóricas

1. Risco de HPV3, correspondendo à detecção de papilomavírus humano, com as categorias negativo, HPV tipo não identificado, HPV 16/18/31/33 e ausente (sem informação).
2. Número de exames de Papanicolau realizados anteriormente, agrupada nas categorias 0 = nenhum exame, 1= 1 a 2 exames, 2= 3 a 5 exames e 3=

maior ou igual a 6 exames.

3. Facilidades, definida como o número de elementos dentre geladeira, televisão, água encanada e esgoto presentes na casa da paciente. Esta variável foi categorizada em possuir os quatro elementos, três, dois e um ou nenhum.

Variáveis quantitativas

4. Paridade, consistindo no número de filhos da paciente

5. Idade, medida em anos.

Com o objetivo de aplicar as técnicas de diagnóstico, reajustamos o modelo final obtendo

$$\log \frac{\hat{\pi}(\mathbf{x})}{1 - \hat{\pi}(\mathbf{x})} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \dots + \hat{\beta}_{12} x_{12}$$

$$\log \frac{\hat{\pi}(\mathbf{x})}{1 - \hat{\pi}(\mathbf{x})} = -0.107 + 0.116x_1 - 0.016x_2 + 1.231x_3 + 0.983x_4 - 0.309x_5 - 0.448x_6 - 0.315x_7 - 0.371x_8 + 0.074x_9 - 0.383x_{10} + 0.017x_{11} + 0.199x_{12},$$

em que X_1 é a variável paridade, X_2 é a idade e as variáveis $X_3, X_4, \dots, X_{12}$ são variáveis indicadoras definidas como

$$(X_3, X_4, X_5) = \begin{cases} (1, 0, 0) & \text{para HPV negativo} \\ (0, 1, 0) & \text{para HPV não identificado} \\ (0, 0, 1) & \text{para HPV 16/18/31/33} \\ (0, 0, 0) & \text{para ausente ou sem informação} \end{cases}$$

$$(X_6, X_7, X_8, X_9) = \begin{cases} (1, 0, 0, 0) & \text{para nenhum exame de papanicolau anteriormente} \\ (0, 1, 0, 0) & \text{para número de exame de papanicolau 1 ou 2} \\ (0, 0, 1, 0) & \text{para número de exame de papanicolau 3 a 5} \\ (0, 0, 0, 1) & \text{para número de exame de papanicolau maior ou igual a 6} \\ (0, 0, 0, 0) & \text{para número de exame de papanicolau ausente} \end{cases}$$

$$(X_{10}, X_{11}, X_{12}) = \begin{cases} (1, 0, 0) & \text{para quatro itens facilidades} \\ (0, 1, 0) & \text{para três itens facilidades} \\ (0, 0, 1) & \text{para duas itens facilidades} \\ (0, 0, 0) & \text{para uma ou zero itens facilidades.} \end{cases}$$

O valor do desvio do modelo foi $D=315$ com 403 graus de liberdade, indicando, numa análise informal, um ajuste adequado. Vamos medir a qualidade do ajuste do modelo, através da estatística C , dada em (3.5).

Na Tabela 6.1 apresentaremos os grupos formados com as observações para o cálculo da estatística C . Foram formados 8 grupos, pois teremos o mesmo número de observações em cada um deles, sendo cada grupo constituído de 52 observações, perfazendo um total de 416 observações, correspondendo a 217 controles e 199 casos.

Tabela 6.1 - Quantidades para o cálculo de C .

Grupo	O_i	n'_i	$\overline{\pi}_i$
1	0	52	0,034
2	5	52	0,090
3	10	52	0,169
4	17	52	0,297
5	27	52	0,525
6	41	52	0,813
7	49	52	0,930
8	50	52	0,969

O cálculo da estatística C é dado a seguir

$$C = \frac{(0 - 52 \cdot 0,034)^2}{52 \cdot 0,034 \cdot (1 - 0,034)} + \dots + \frac{(50 - 52 \cdot 0,969)^2}{52 \cdot 0,969 \cdot (1 - 0,969)} =$$

$$= 1,804 + 0,024 + 0,205 + 0,219 + 0,007 + 0,210 + 0,124 + 0,098 = 2,690$$

e o nível descritivo associado ao valor de C numa distribuição qui-quadrado com $(g - 2) = 6$ graus de liberdade é dado por $p = 0,8466$, indicando que o ajuste é adequado.

A Figura 6.1.d corresponde ao gráfico das probabilidades empíricas para o resíduo componente do desvio padronizado do modelo ajustado.

Figura 6.1: Gráfico de probabilidades empíricas

As Figuras 6.1.a, b e c apresentam respectivamente os gráficos das medianas, limite inferior e limite superior dos resíduos ordenados obtidos na simulação em função dos resíduos do modelo ajustado ordenados.

O gráfico da Figura 6.1.a exibe um comportamento linear, sugerindo uma reta com coeficiente angular um. Dada a grande quantidade de pontos no gráfico da Figura 6.1.d, fica difícil verificar se existem elementos fora das faixas. Por isso, construímos as variáveis: $td_i - LI_i$ e $LS_i - td_i$, $i = 1, 2, 3, \dots, n$, em que td_i , LI_i e LS_i são respectivamente os resíduos padronizados ordenados do modelo e os limites inferior e superior das bandas de confiança construídas. Valores negativos dessas variáveis representam situações em que o resíduo padronizado está fora da faixa. Após os cálculos executados notamos que não houve nenhum valor negativo, logo todos os resíduos padronizados estão dentro da faixa. Concluímos então que o modelo ajustado é adequado aos

dados.

Dando continuidade à análise, serão calculadas as medidas de diagnóstico mencionadas neste trabalho, e seus valores estarão representados graficamente. Esses gráficos são informativos a respeito do posicionamento dos pontos aberrantes com relação às probabilidades ajustadas ($\hat{\pi}$).

Na Figura 6.2 temos o gráfico do alavanca contra as probabilidades estimadas.

Figura 6.2: Gráfico do alavanca em função das probabilidades estimadas

Podemos notar dois pontos que se destacam dos demais, com altos valores do *leverage*, correspondendo às mulheres de identificação: 406 (controle) e 202 (caso). Uma tendência similar é exibida na Figura 6.3, relativa à Distância de Cook, indicando que esses pontos têm uma grande influência nas estimativas dos coeficientes β . Analisando-se os dados, observamos que

as duas mulheres têm em comum a variável número de papanicolau com valor ausente pois ambas não sabiam o número de exames de papanicolau realizados e não tinham confirmado o tipo de HPV, uma era HPV não identificado e a outra HPV sem informação.

Figura 6.3: Gráfico da distância de Cook $\Delta\hat{\beta}$ em função das probabilidades estimadas

Na Figura 6.4 encontra-se o gráfico do resíduo de Pearson padronizado em função das probabilidades estimadas. Podemos notar uma cauda maior para as mulheres do grupo controle ($y = 0$), indicando que esses elementos tendem a apresentar os resíduos padronizados com valores absolutos grandes e probabilidades estimadas grandes também. As seis observações com os maiores resíduos (em valor absoluto) serão comentadas a seguir.

Figura 6.4: Gráfico do Resíduo de Pearson padronizado em função das probabilidades estimadas

Os gráficos de ΔX_j^2 , ΔD_j e ΔX_j^2 com bolhas proporcionais a $\Delta \hat{\beta}_j$, nas Figuras 6.5 a 6.7, respectivamente, são similares. Neles destacam-se os pontos: 85, 207, 209, 311, 316 e 336, comuns aos três gráficos. Notamos que as cinco primeiras observações são mulheres que não têm câncer de colo de útero, mas tem o HPV 16/18/31/33, que é considerado grave e uma indicação para câncer de colo de útero. Tal fato poderia ser consequência de uma falha no exame. Sendo assim, é estranho que tenham esse tipo de HPV e não desenvolveram câncer. Já a última observação (336) tem câncer de colo de útero e o exame de HPV foi negativo, o que também é inesperado. Notamos então que essas mulheres que acabamos de destacar são bem diferentes da maioria ou do que poderia ser considerado de “normal” e isso foi detectado pelas medidas de diagnóstico utilizadas.

Notamos ainda, na Figura 6.7, a existência de alguns pontos com ΔX_j^2 alto, no entanto, como os círculos associados não são grandes, tais pontos tem pouca influência nas estimativas dos coeficientes β .

Figura 6.5: Gráfico de ΔX_j^2 em função das probabilidades estimadas

Figura 6.6: Gráfico de ΔD_j em função das probabilidades estimadas

Figura 6.7: Gráfico de ΔX_j^2 com bolhas proporcionais ao tamanho de $\Delta \hat{\beta}$ em função das probabilidades estimadas

A análise de diagnóstico para esse conjunto de dados mostrou-se plenamente satisfatória, pois as medidas calculadas e os gráficos apresentados foram fundamentais para encontrar e descrever elementos amostrais discrepantes, após o ajuste do modelo.

No próximo capítulo serão feitas as considerações finais sobre este trabalho e serão dadas algumas sugestões para futuras pesquisas.

Capítulo 7

Considerações Finais e Sugestões

Neste trabalho procuramos apresentar os principais conceitos em medidas de diagnósticos em regressão logística e aplicações do mesmo, com o objetivo de mostrar o muito que se faz necessário para uma análise de um modelo e para detectar pontos diferentes da maioria dos dados.

Esta teoria não é muito aplicada no nosso cotidiano, devido à complexidade matemática dos métodos envolvidos, praticamente inviáveis sem o auxílio do computador. O que temos notado é que a teoria vem sendo desenvolvida e explorada num ritmo que ainda não vem sendo acompanhado pelo desenvolvimento de programas computacionais eficientes para agilizarem sua utilização numa escala maior. Finalizando, destacaremos alguns artigos sobre diagnóstico em regressão logística que, por apresentarem um enfoque distinto do utilizado nesse trabalho, não foram aqui abordados.

Johnson (1985) constroi medidas de diagnóstico na regressão logística avaliando a influência de cada elemento amostral na estimativa das probabilidades $\hat{\pi}(x)$ e também na classificação de observações futuras, ou seja, na previsão da variável resposta Y , com valores 0 ou 1. Para quantificar a diferença entre $\hat{\pi}(x)$ com e sem a i -ésima observação, é utilizada a medida de divergência de Kullback-Leibler.

Cook e Croos-Dabrera (1998) estudam a eficiência do gráfico de resíduos parciais na determinação de transformações nas variáveis independentes em modelos lineares generalizados. Neste caso, as variáveis independentes são aleatórias e utilizam-se distribuições condicionais. É introduzindo o gráfico “CERES” (Combinação de Esperanças Condicionais e Resíduos) e a utilidade desse gráfico na detecção de transformações também é analisada. A utilização de tais gráficos na regressão logística é analisada como caso particular.

O artigo de Pardoe e Cook (2002) seria o mais atual referente ao assunto. Os autores propõe uma metodologia gráfica com base em uma estrutura Bayesiana na avaliação do ajuste do modelo de regressão logística. Tais gráficos seriam facilmente obtidos graças aos “softwares” disponíveis e, segundo os autores, são gráficos mais intuitivos que os gráficos tradicionais.

Podemos então notar que existem várias áreas referentes a diagnóstico em regressão logística que podem ainda ser exploradas.

Apêndice A

Apêndice

Apresentaremos a seguir os passos básicos para o cálculo das medidas de diagnóstico em alguns pacotes estatísticos.

- No Minitab 13, a sequência de comandos é a seguinte

Stat > Regression > Binary Logistic Regression > Storage.

Como consequência dessas operações podem ser obtidas as medidas de diagnósticos descritas neste trabalho.

- No S-plus 2000, para obter os resíduos de Pearson e o desvio devemos acessar

Statistics > Regression > Logistic > Results.

Os demais resíduos devem ser calculados abrindo a seção de comandos (Commands) e programando conforme a fórmula de cada um deles. Seguem a seguir os passos para os comandos do exemplo da Seção 3.3.

```

1. ajuste <- glm(formula = Yi ~ renda.familiar + automovel.mais.velho,
family = binomial(link = logit), data = Ch14pr10pl, na.action = na.exclude,
control = list(epsilon = 0.0001, maxit = 50, trace = F))
2. X<-model.matrix(ajuste)
3. V<-ajuste$weights
4. v<-diag(V)
5. Hat<-sqrt(v)%*%X%*%solve(t(X)%*%v%*%X)%*%t(X)%*%sqrt(v)
6. h<-diag(Hat)
7. rp<-resid(ajuste,type="pearson") (Resíduo de Pearson= $r(y_j, \hat{\pi}_j)$ )
8. rps<-rp/sqrt(1-h) (Resíduo de pearson padronizado =  $r_{sj}$ )
9. deltabeta<-(rps2*h)/(1-h) ( $\Delta\hat{\beta}_j$ )
10. deltaqui<-rp2/(1-h) ( $\Delta X_j^2$ )
11. rdev<-resid(ajuste) (Resíduo desvio (default))
12. deltadev<-rdev2+rp2*h/(1-h) ( $\Delta D_j$ ).

```

Apresentaremos a seguir o programa para construção do gráfico de Probabilidades Empíricas para a distribuição Binomial. Para mais detalhes, vide Paula (1999).

```

1. X<-model.matrix(ajuste)
2. n<-nrow(X)
3. e<-matrix(0,n,100)
4. for(i in 1:100){
+ dif<-runif(n)-fitted(ajuste)
+ dif[dif>=0]<-0
+ dif[dif<0]<-1

```

```

+ nresp<-dif
+ fit<-glm(nresp ~ X, family=binomial)
+ w<-fit$weights
+ W<-diag(w)
+ H<-solve(t(X)%*%W%*%X)
+ H<-sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
+ h<-diag(H)
+ e[,i]<-sort(resid(fit,type="desvio")/sqrt(1-h))}
#
e1<-numeric(n)
e2<-numeric(n)
for(i in 1:n){
+ eo<-sort(e[,i])
+ e1[i]<-eo[5]
+ e2[i]<-eo[95]}
#
td<-sort(rdev*sqrt(1/(1-h)))
med<-apply(e,1,mean)
faixa<-range(td,e1,e2)
#

```

par (pty="s") (abrirá uma janela *GSD2* para construção do gráfico)

ou

par(mfrow=c(2,2))(abrirá uma janela para formação de 4 gráficos)

plot(td, med, xlab="Componentes do Desvio",ylab="Valores medianos",

```

ylim=faixa)
plot(td,e1,xlab="Componentes do Desvio", ylab="L.I.da Simulação", type="l",
ylim=faixa, lty=1)
plot(td,e2,xlab="Componentes do Desvio", ylab="L.S.da Simulação", type="l",
ylim=faixa, lty=1)
plot(td, med, xlab="Componentes do Desvio", ylab="Valores medianos e os
limites", ylim=faixa)
par(new=T)
plot(td,e1,axes=F,xlab=,ylab=,type="l",ylim=faixa,lty=1)
par(new=T)
plot(td,e2,axes=F,xlab=,ylab=,type="l",ylim=faixa,lty=1)

```

Dados amostrais do capítulo 6

Y	Idade	Facilidades	Paridade	N.exames de papanicolau	Risco de HPV3
0	64	0	6	2	0
1	51	0	8	2	1
0	48	1	11	0	1
1	49	3	12	0	1
0	41	0	3	3	0
1	45	0	5	2	2
0	51	0	1	2	0
1	42	0	6	0	2
0	58	0	8	2	0
1	76	0	2	1	2
0	57	0	5	0	0
1	46	0	3	0	1
0	64	3	2	0	0
1	39	3	1	0	1
0	56	2	6	0	0
1	53	1	10	0	2
0	41	0	2	3	0
1	52	1	3	0	2
0	52	0	7	1	0
1	34	0	7	0	2
0	57	0	3	2	1
1	60	1	16	0	2
0	31	0	1	1	0
1	45	1	8	0	2
1	38	3	4	0	2
1	68	0	3	0	0
1	46	0	6	0	0
1	41	1	5	0	2
1	56	0	0	0	1
1	43	0	9	0	2
1	43	0	2	0	0
1	66	3	7	0	2
1	40	2	10	0	2
1	54	0	13	1	2
1	56	2	0	0	2
1	61	0	9	1	2
1	45	3	2	0	2
1	44	0	8	0	0
1	49	3	9	0	0
1	64	0	0	0	2
1	36	0	9	0	9

1	42	3	8	0	2
1	53	1	7	0	1
1	65	0	6	1	1
1	53	2	10	0	2
1	57	0	10	0	2
1	42	0	5	0	9
1	61	0	10	0	1
1	47	0	10	0	2
1	41	2	2	0	2
1	75	2	10	0	2
1	45	1	7	1	2
1	70	0	5	0	2
1	40	0	3	0	2
1	63	2	6	1	2
1	66	0	1	1	1
1	30	0	1	1	0
1	68	0	12	0	1
1	75	0	5	0	1
1	61	3	5	1	2
1	65	1	5	0	2
1	48	0	6	0	2
1	42	0	3	0	2
1	33	0	4	1	2
1	56	2	13	1	2
1	59	0	2	0	1
1	58	0	5	0	2
1	40	1	4	1	2
1	67	0	13	0	2
1	66	1	5	0	2
1	52	1	9	1	2
1	57	3	17	0	2
1	66	1	12	0	0
1	67	3	11	0	2
1	44	2	11	0	2
1	50	2	5	2	2
1	49	1	7	1	2
1	69	3	8	2	2
1	47	1	6	0	2
1	48	0	5	0	2
0	61	0	3	3	0
0	43	3	6	0	0
0	37	0	2	3	0
0	44	0	2	1	0
0	56	0	9	0	2
0	55	0	17	3	0

0	72	0	9	0	0
0	64	0	1	2	0
0	62	0	7	0	0
0	74	3	10	0	0
0	42	0	3	3	0
0	44	1	11	1	9
0	43	1	4	1	9
0	52	3	8	0	0
0	58	1	4	0	0
0	40	1	5	0	1
0	77	0	2	1	2
0	46	1	2	1	2
0	69	0	2	1	0
0	50	0	3	2	0
0	38	2	3	3	0
0	58	0	3	0	0
0	48	0	4	3	0
0	77	3	19	0	0
0	51	1	2	3	0
0	43	0	6	3	0
0	34	0	3	1	0
0	67	0	4	0	0
0	49	0	6	3	1
0	63	0	4	3	0
0	35	0	2	0	0
0	40	0	1	1	0
0	62	0	5	3	0
0	38	0	1	3	0
0	59	0	4	1	0
0	52	1	6	3	1
0	54	0	7	2	0
0	46	1	3	1	0
0	54	0	14	3	0
0	51	0	2	0	0
0	47	0	2	3	0
0	36	0	2	3	0
0	54	0	3	1	2
0	52	0	4	2	0
0	65	0	5	0	9
0	51	0	4	1	0
0	57	0	4	2	0
0	46	0	4	1	0
0	76	0	1	0	0
0	50	2	2	0	0
0	63	1	1	0	0

0	38	0	5	2	0
0	68	0	7	0	0
0	60	0	1	2	1
0	47	1	4	1	1
0	27	0	2	0	0
0	36	0	2	3	0
0	42	1	7	3	0
0	33	2	4	2	0
1	46	1	8	0	0
0	46	0	2	3	0
1	46	0	7	0	2
0	68	0	2	1	0
1	57	1	13	0	2
0	35	0	2	1	0
1	44	1	10	1	2
0	26	0	2	1	0
1	59	2	4	3	2
0	38	2	9	0	1
1	59	0	9	0	1
0	35	1	2	2	2
1	36	0	3	0	2
0	29	0	2	2	1
1	72	1	5	0	0
0	66	0	8	1	0
1	62	0	4	0	2
0	32	0	6	1	9
1	36	1	6	1	1
0	30	2	2	0	1
1	46	0	4	0	2
0	62	0	6	3	0
0	50	0	6	1	9
1	32	0	4	1	2
0	51	0	4	1	2
1	39	0	4	0	2
0	30	0	1	0	0
1	51	0	7	0	0
0	42	0	0	0	0
1	59	0	12	1	0
0	38	3	2	0	9
1	40	2	4	3	2
0	35	0	2	3	9
0	55	2	11	1	0
0	34	0	1	1	0
1	50	3	6	0	2
0	38	0	3	1	0

1	49	2	11	0	2
0	59	1	10	1	0
1	74	0	1	0	9
0	66	3	8	0	0
1	57	0	4	0	0
0	50	0	9	3	0
0	29	0	0	0	0
0	58	0	10	1	0
1	52	3	9	0	2
0	40	0	2	2	9
1	30	1	4	2	2
0	68	2	1	1	0
1	28	1	4	0	9
0	39	0	1	3	0
1	54	2	5	0	2
0	40	1	3	2	0
1	39	0	12	0	0
0	43	1	2	3	9
1	63	0	0	0	2
0	63	3	15	0	0
1	41	0	1	0	0
0	72	0	1	0	0
0	57	0	0	3	2
1	46	3	9	1	2
0	53	0	2	0	0
1	62	0	8	9	1
1	48	0	9	1	2
0	57	3	11	1	9
1	51	3	11	0	0
1	37	0	5	0	2
0	76	2	5	0	2
1	45	2	5	0	2
0	67	0	7	0	2
1	55	0	9	0	2
0	75	0	5	0	0
1	54	1	11	1	2
0	59	0	5	1	0
1	76	3	14	0	1
0	49	1	8	1	0
1	39	0	7	0	2
0	52	0	6	0	9
1	36	0	5	1	2
0	44	2	4	0	0
1	29	0	1	0	2
0	70	0	2	1	9

1	49	0	12	2	2
0	40	1	3	3	0
1	60	3	20	0	2
0	61	1	13	3	0
1	76	1	13	1	2
0	56	3	9	0	0
0	60	0	3	1	0
0	57	2	2	3	0
0	40	0	3	3	0
1	51	2	4	1	9
0	62	0	3	3	0
1	39	0	8	1	1
1	54	1	2	3	2
0	43	2	10	0	0
1	64	1	12	0	2
0	55	0	3	2	9
1	51	0	5	0	2
0	53	0	1	3	0
0	56	0	5	2	9
1	44	3	8	0	1
0	76	0	5	1	0
1	45	1	6	0	2
0	63	0	11	2	0
1	50	0	3	3	2
1	62	2	8	0	2
0	46	1	6	1	0
1	47	0	4	1	1
0	54	0	2	3	0
1	60	1	11	0	2
0	73	0	5	0	0
1	55	2	9	0	1
0	73	0	5	0	1
1	42	1	3	0	2
0	45	2	1	0	0
1	64	2	8	0	1
0	64	0	0	1	1
1	58	0	10	0	2
0	65	0	7	1	0
1	30	0	3	2	2
0	71	0	2	0	0
1	58	1	3	0	2
0	53	0	3	1	0
1	57	3	9	1	0
0	58	2	8	1	0
1	55	3	8	1	2

0	47	2	5	0	9
1	72	0	3	1	2
0	71	0	9	1	0
1	44	0	7	0	2
0	49	0	2	1	2
1	46	3	5	0	0
0	55	0	6	3	0
1	62	0	6	0	9
0	65	0	3	2	0
1	45	0	2	2	9
0	44	0	2	3	0
1	34	3	5	2	0
0	56	0	7	0	0
1	46	2	7	0	9
1	36	2	4	0	2
0	44	0	4	2	0
1	61	0	10	0	1
0	77	1	11	0	0
0	69	0	5	3	9
1	48	0	3	1	0
1	67	0	4	0	0
0	73	1	8	0	0
1	71	0	5	1	9
0	62	0	3	1	0
1	60	3	7	0	2
0	47	1	5	0	1
1	42	1	3	2	2
0	48	0	12	1	0
1	47	0	3	1	0
0	47	1	6	2	0
1	66	0	7	0	2
0	62	0	8	0	0
1	64	0	6	0	2
0	61	0	6	0	0
1	59	3	13	0	2
0	51	0	2	3	0
1	56	3	11	0	1
0	42	1	3	1	0
1	45	3	7	0	0
1	59	0	5	0	9
0	48	0	4	2	0
1	33	0	2	3	2
0	39	1	4	2	0
1	42	1	6	1	1
0	40	1	7	0	2

0	48	0	2	1	0
0	61	0	9	3	0
1	29	2	0	0	1
0	60	0	4	2	0
0	43	2	5	0	2
1	44	0	4	0	0
0	61	0	5	1	1
0	68	1	9	1	9
1	46	0	10	1	2
0	52	0	4	0	0
1	65	0	5	1	2
0	62	0	4	3	0
1	66	1	8	0	1
0	40	3	7	0	0
1	69	0	4	0	2
0	65	3	5	0	0
1	71	0	7	0	1
0	49	0	2	3	1
1	43	1	8	1	2
0	48	0	6	1	0
1	64	3	5	0	0
0	57	0	1	0	0
1	78	1	5	1	1
0	62	0	6	3	0
1	45	0	2	2	0
0	62	0	2	0	0
1	74	0	3	0	2
0	48	0	3	2	0
1	64	1	3	0	1
0	70	0	0	0	0
1	48	0	10	1	9
0	61	1	0	1	0
1	44	1	10	0	2
0	55	0	2	3	0
0	66	0	4	1	0
1	50	1	7	1	2
0	47	0	0	0	0
1	60	0	10	0	2
0	49	0	4	2	0
1	65	0	4	0	9
0	55	1	6	1	0
1	73	2	9	0	1
0	43	0	1	2	1
1	50	3	9	0	0
1	56	3	2	0	2

0	63	0	5	1	0
1	40	0	8	1	2
0	65	1	11	1	0
1	78	1	11	0	1
0	48	2	7	3	0
1	64	2	14	0	2
0	49	0	3	2	0
1	32	0	3	0	9
0	49	3	13	0	0
1	46	0	4	3	1
0	46	3	0	1	9
1	49	0	4	1	2
0	48	1	11	0	0
1	38	1	3	0	2
0	49	0	7	3	0
1	56	2	14	1	0
0	40	1	12	0	1
1	38	0	3	3	2
0	47	0	11	0	0
1	47	3	4	0	0
0	48	0	6	0	0
1	44	2	10	0	1
0	47	3	10	0	0
1	49	1	7	0	2
0	48	0	7	1	0
1	60	1	10	0	2
0	65	0	5	1	0
1	65	0	3	1	2
0	36	1	0	1	1
1	53	1	4	0	2
1	64	2	11	2	1
0	36	0	0	3	9
1	46	3	8	0	2
0	47	2	2	3	0
1	46	0	8	0	2
0	32	0	0	1	0
1	66	0	5	0	2
1	35	0	3	0	2
0	45	1	12	0	9
1	68	0	5	0	0
0	55	0	4	3	0
0	67	0	3	2	0
1	40	2	6	0	2
0	59	0	2	3	0
1	47	0	1	0	2

0	45	0	5	2	0
0	63	1	9	1	9
1	55	0	8	0	1
1	43	0	8	1	2
0	58	0	6	9	9
1	47	1	7	2	2
0	46	1	5	2	1
0	34	0	7	0	0
0	43	0	0	1	9
0	66	0	0	0	9
0	46	0	0	3	0
0	68	0	7	0	9
0	35	1	3	0	9
0	71	0	2	0	9
0	56	0	2	0	9

Bibliografia

- [1] Andersen, E. B. (1991). *The Statistical Analysis of Categorical Data*. 2ed - Springer Verlag. 531 p.
- [2] Atkinson, A.C.(1985). *Plots, Transformations and Regression*. Oxford Science Publications. 282 p.
- [3] Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.
- [4] Cook, R.D.(1977).Detection of Influential Observations in Linear Regression. *Technometrics*, 19: 15-18.
- [5] Cook, R.D. and Croos-Dabrera R. (1998).Partial Residual Plots in Generalized Linear Models. *Journal of the American Statistical Association*, 93,442, 730-739.
- [6] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London. 230 p.

- [7] Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London. 142 p.
- [8] Cox, D.R. and Snell, E.J. (1989). *The Analysis of Binary Data*, 2ed. Chapman and Hall, London. 236 p.
- [9] Eluf Neto, J. (1993). Risk Factors for Invasive Cervical Cancer in Brazil. *Thesis degree of Doctor*, University of London. 232 p.
- [10] Everitt, B.S. (1994). *Statistical Analysis using S-Plus*. Chapman and Hall, London. 143 p.
- [11] Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, 74, 3, p 503-515.
- [12] Hosmer, D.W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.
- [13] Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley Sons, New York. 307 p.
- [14] Jennings, E. D. (1986). Outliers and Residual distributions in Logistic Regression. *Journal of the American Statistical Association*, 81, 396, 987-990.
- [15] Johnson, W. (1985). Influence measures for Logistic Regression: Another point of view. *Biometrika*, 72, 59-65.

- [16] Landwehr, J.M., Pregibon D. and Shoemaker A.C. (1984). Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association*, 79, 385, 61-71.
- [17] Lemeshow, S. and Hosmer, D.W. (1982). The use of goodness-of-fit statistics in the development of logistic regression models. *Amerian Journal of Epidemiology*, 115, 92-106.
- [18] McCullagh, P. and Nelder (1989). *Generalized Linear Models*. 2 ed Chapman Hall, London. 511 p.
- [19] Neter, J., Kutner, M.H., Nachtsheim C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models*. 4nd. edition - IRWIN. 1.408 p.
- [20] Pardoe, I. and Cook, R.D. (2002). A Graphical Method for Assessing the Fit of a Logistic Regression Model. *The American Statistical*, vol.56, 4, 263-272.
- [21] Paula, G.A. (1999). *Modelos de Regressão*. Instituto de Matemática e Estatística (IME) - USP. Não publicado. 226 p.
- [22] Pregibon, D. (1981). Logistic regression diagnostics. *The Annals os Statistics* 9, 705-724.
- [23] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2 ed. New York - Wiley. 625 p.
- [24] S-Plus (2000). Pacote Estatístico Profissional - Release 3.
- [25] Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*. 67, 250-251.

- [26] Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews*. 14(1), 101-116.