Regressão logística politômica: revisão teórica e aplicações

Hélio Radke Bittencourt

Resumo

O tradicional modelo de regressão logística tornou-se um método padrão de análise na área das ciências da saúde, especialmente Epidemiologia, pois é capaz de estabelecer uma relação de dependência entre uma única variável-resposta binária e um conjunto de variáveis independentes quantitativas ou qualitativas. A técnica é considerada uma abordagem parcialmente não-paramétrica, não exigindo suposições sobre o comportamento probabilístico dos dados de entrada. Neste trabalho uma extensão da regressão logística para variáveis-resposta politômicas é apresentada, bem como uma revisão sobre os aspectos teóricos mais importantes e aplicações da técnica com a utilização de bancos de dados reais.

Palavras-chave: Regressão Logística, Análise Discriminante.

Abstract

The traditional logistic regression model became a standard method in the medical and biological sciences, especially in epidemiology, because allows modeling of binary response variables only and a set of quantitative or qualitative independent variables. Logistic regression can be regarded as a partially parametric approach, since it assumes nothing about the probability distribution of variables. This paper describes an extension of the logistic regression to polytomous response variables, as well as presents a revision about the most important theoretical aspects and gives some results obtained when using real databases.

Key-Words: Logistic Regression, Discriminant Analysis.

1 Introdução

Variáveis qualitativas nominais são aquelas que mais limitam a possibilidade de utilização de técnicas estatísticas, especialmente quando o número de categorias excede dois. É muito comum a utilização de tabelas de contingência para verificação de associação entre variáveis nominais que, geralmente, são acompanhadas do bem conhecido teste Qui-quadrado. A prova não-paramétrica do Qui-quadrado, as-

Hélio Radke Bittencourt - Laboratório de Eestatística - Ulbra

sim como as medidas de associação derivadas, pode ser adequada para um grande número de casos, entretanto só permite a análise simultânea de duas variáveis e, no caso de variáveis quantitativas, é necessária a prévia categorização, implicando em perda de informação.

De acordo com Allison (1999) existem pesquisadores utilizando inadequadamente a técnica de regressão linear para tratamento de variáveis-resposta qualitativas nominais e ordinais o que, na sua opinião, se deve ao desconhecimento de técnicas mais avançadas.

O presente estudo inicia com uma revisão do modelo de regressão logística tradicional e, em seguida, é apresentada a extensão da técnica para variáveis politômicas, resultados práticos e as considerações finais.

2 O modelo de regressão logística tradicional

De acordo com Hosmer e Lemeshow (1989) a regressão logística, em sua forma tradicional, consiste de um modelo que relaciona um conjunto de *p* variáveis independentes X_1 , X_2, ... , X_p a uma variável dependente Y que assume apenas dois possíveis estados, digamos 0 ou 1. O modelo logístico permite a estimação direta da probabilidade de ocorrência de um evento (Y = 1):

$$P(Y=1)=\frac{\exp\left(\beta_0+\beta_1x_1+\ldots+\beta_px_p\right)}{1+\exp\left(\beta_0+\beta_1x_1+\ldots+\beta_px_p\right)}$$

e, conseqüentemente,

$$P(Y=0)=1-P(Y=1)=\frac{1}{1+\exp\left(\beta_0+\beta_1x_1+\ldots+\beta_px_p\right)};$$

onde β_i são os parâmetros do modelo, estimados pelo método de máxima verossimilhança.

A transformação que está por trás do modelo logístico é a chamada transformação *logit*, denotada por *g*(*x*). É uma função linear nos parâmetros β, contínua e que pode variar de $-\infty$ a $+\infty$:

$$logit(x)=g(x)=ln\left[\frac{P(Y=1)}{1-P(Y=1)}\right]=$$

$$\beta_0+\beta_1x_1+\ldots+\beta_px_p$$

Hosmer e Lemeshow (1989) dizem que há pelo menos duas razões para utilização do modelo logístico na análise de variáveis-resposta dicotômicas: 1) de um ponto de vista matemático, é extremamente flexível e fácil de ser utilizado; 2) permite uma interpretação de resultados bastante rica e direta. A Figura 1 apresenta a função logística com o seu característico formato em 'S' e a relação linear entre uma única variável *x* e o *logit* g*(x)*.

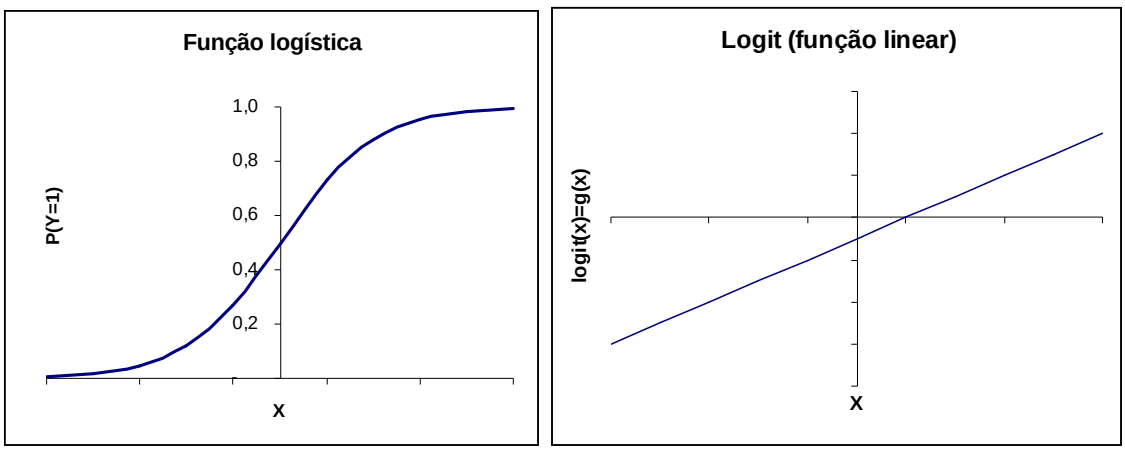

Figura 1: Função logística e a relação logit

Anderson (1982) enfatiza que a discriminação logística pode ser aplicada a uma grande variedade de famílias de distribuições, pois a suposição de linearidade do *logit* é válida numa grande quantidade de funções de distribuição de probabilidade, como por exemplo, a distribuição normal multivariada.

2.1 Histórico do Modelo Logístico e Principais Aplicações

É difícil precisar exatamente o ano no qual o modelo logístico foi utilizado pela primeira vez, mas Cox e Snell (1989) e Hosmer e Lemeshow (1989) concordam que o modelo de regressão logística ganhou reconhecimento após o trabalho de Truett, Cornfield e Kennel (1967) que analisava o risco de doença coronária em um grande projeto conhecido por *"Framingham heart study"*. Esse trabalho ganhou fama e até hoje é considerado um marco inicial dos estudos envolvendo regressão logística nas áreas da saúde. McLachlan (1992) também afirma que as primeiras aplicações do modelo logístico foram em estudos prospectivos de doenças coronárias. Contudo, nessas aplicações, os autores realizaram o processo de estimação de parâmetros sob a suposição de normalidade, que se torna desnecessária quando a estimação é feita por máxima verossimilhança via métodos numéricos. O procedimento de estimação em um contexto mais genérico foi proposto por Day and Kerridge (1967) e por Walker and Duncan (1967).

Hosmer e Lemeshow (1989) afirmam que o modelo de regressão logística tornou-se um método padrão de análise de regressão de dados dicotômicos, especialmente nas ciências da saúde. De fato, aplicações da regressão logística são comumente encontradas em periódicos da área de saúde, tais como *The American Journal of Epidemiology*, *The American Journal of Public Health*, *The International Journal of Epidemiology* e *The Journal of Chronic Diseases*.

A literatura sobre regressão logística é muito vasta, tendo apresentado um crescimento muito rápido. Além das inúmeras aplicações na área da saúde, a regressão logística também tem sido utilizada no campo da econometria, administração e educação. Por esse motivo, encontramos artigos envolvendo regressão logística em periódicos de diversas áreas.

3 Regressão logística politômica

O modelo de regressão logística, originalmente desenvolvido para variáveis-resposta binárias, é extensível para variáveis-resposta politômicas (três ou mais categorias). O entendimento da regressão logística politômica fica mais simples se for utilizado como exemplo introdutório um modelo cuja variável-resposta Y assume apenas três níveis, digamos 0, 1 e 2, assim como descrito em Hosmer e Lemeshow (1989). Agora, o modelo logístico terá duas funções *logit*: a razão entre Y=1 e Y=0 e a razão entre Y=2 e Y=0. Nesse caso, o nível Y=0 foi assumido como base.

$$g_1(x) = ln\left[\frac{P(Y=1)}{P(Y=0)}\right] =$$

$$\beta_{10} + \beta_{11}x_1 + \ldots + \beta_{1p}x_p$$

$$g_2(x) = ln\left[\frac{P(Y=2)}{P(Y=0)}\right] =$$

$$\beta_{20} + \beta_{21}x_1 + \ldots + \beta_{2p}x_p$$

A partir das funções lineares $g_i(x)$, cujos parâmetros são estimados por máxima verossimilhança, é possível calcular as probabilidades condicionais de ocorrência de cada categoria da variável-resposta Y dado um vetor de observações **x**, conforme segue:

$$P(Y=0\,|\,\mathbf{x}) = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=1\,|\,\mathbf{x}) = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=2\,|\,\mathbf{x}) = \frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

Krzanowski (1988) afirma que, no momento em que as probabilidades *a posteriori* $P(Y=y|\mathbf{x})$ do modelo logístico são utilizadas para se estabelecer uma regra de alocação, a abordagem é chamada de *discriminação logística.* Na área de Reconhecimento de Padrões esse é o termo mais utilizado, conforme se pode verificar em McLachlan (1992) e Bittencourt e Clarke (2002).

A generalização do modelo logístico para variáveis-resposta com *k* níveis (*k*>*2*) é direta, permitindo sua utilização para discriminação entre *k* classes. Na regressão logística politômica a probabilidade de uma dada observação **x** pertencer a uma das classes y_i é estimada diretamente por meio da seguinte expressão:

$$P(Y=y_i\,|\,\mathbf{x}) = \frac{exp\{g_i(x)\}}{1+\sum_{j=1}^{k-1} exp\{g_j(x)\}}$$

$$i = 1, 2, \ldots, k\text{-}1$$

onde a função *logit*, assumindo o nível y_k como base, é dada por

$$g_i(x) = ln\left[\frac{P(Y=y_i\,|\,\mathbf{x})}{P(Y=y_k\,|\,\mathbf{x})}\right] =$$

$$\beta_{i0} + \beta_{i1}x_1 + \ldots + \beta_{ip}x_p$$

$$i = 1, 2, \ldots, k\text{-}1$$

$$g_k(x) = 0 \cdot$$

Considerando y_1, y_2,..., y_k categorias exaustivas e exclusivas da variável Y, podemos afirmar que $\sum_{i=1}^{k} P(y_i\,|\,\mathbf{x}) = 1$. Portanto, a probabilidade de uma observação **x** pertencer a classe y_k, denotada por $P(y_k|\mathbf{x})$, pode ser obtida por diferença:

$$P(y_k\,|\,\mathbf{x}) = 1 - \sum_{i=1}^{k-1} P(y_i\,|\,\mathbf{x})$$

A utilização do modelo logístico para discriminação de classes pode ser direta. A regra de classificação para alocar uma observação **x** numa das classes y_i é muito simples:

$$\mathbf{x} \in y_i \quad se \quad P(y_i\,|\,\mathbf{x}) > P(y_j\,|\,\mathbf{x}) \quad \forall\, j \neq i$$

O modelo logístico necessita da estimação de *k-1* vetores de parâmetros $\boldsymbol{\beta}_i' = [\beta_1, \beta_2, \ldots, \beta_p]$, correspondentes a *k-1* categorias da variável Y. A *k*-ésima categoria é assumida como base. O processo de estimação dos parâmetros em regressão logística está baseado na maximização da função de verossimilhança $\ell(\mathbf{x}, \boldsymbol{\beta})$. Para tornar possível a realização desse procedimento são necessárias *n* amostras de treinamento $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, cujas classes a que pertencem são conhecidas.

Os vetores solução $\mathbf{b}_i$ que maximizam a função $\ell(\mathbf{x}, \boldsymbol{\beta})$ são aqueles que tornam máxima a probabilidade da particular amostra de treinamento $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ ter sido selecionada. Sendo assim, sob a hipótese da amostra ser representativa da população em estudo, obtém-se um modelo que maximiza as chances de classificar todas observações da população nas classes y_i as quais realmente pertencem. Como as equações derivadas da função de verossimilhança são não lineares, há necessidade da utilização de métodos numéricos para

encontrar uma solução. Esses processos são iterativos e estão disponíveis em alguns *softwares* estatísticos. No presente estudo foi utilizado o procedimento *Multinomial Logistic Regression* disponível no *software* SPSS versão 10.0 e que se encontra-se descrito em Norusis e SPSS Inc. (1999). O procedimento CATMOD do sistema SAS também pode ser utilizado.

3.1 Testes de Significância

O primeiro teste de significância importante na regressão logística é o teste da razão de verossimilhança, onde a hipótese de que pelo menos um dos parâmetros β_{ij} é diferente de zero (exceto os interceptos – parâmetros β_{i0}) é testada. Esse teste faz uma comparação entre o valor da função de verossimilhança para o modelo contendo apenas os interceptos e a verossimilhança do modelo final com todos os parâmetros estimados. A estatística de teste *D*, chamada de *deviance*, tem uma distribuição qui-quadrado e é calculada da seguinte forma:

$$D = -2\ln\left(\frac{\ell(\boldsymbol{\beta}_0)}{\ell(\mathbf{x},\boldsymbol{\beta})}\right) = -2\ln \ell(\boldsymbol{\beta}_0) - 2\ln \ell(\mathbf{x},\boldsymbol{\beta})$$

$$\sim \chi^2_{(k-1)p}$$

onde,

$\ell(\boldsymbol{\beta}_0)$ é o valor da função de verossimilhança apenas com os interceptos

$\ell(\mathbf{x},\boldsymbol{\beta})$ é o valor da função de verossimilhança para o modelo final

k é o número de categorias da variável-resposta Y

p é o número de variáveis independentes (*x*) incluídas no modelo

Para a realização de testes de significância individuais para os parâmetros β_{ij}, é bastante comum a utilização da bem-conhecida estatística de *Wald*, onde a hipótese nula é a de que o particular coeficiente β_{ij} é igual a zero. A estatística *W* de *Wald* é definida como o quadrado da razão entre a estimativa de máxima verossimilhança para o coeficiente e seu respectivo erro-padrão (EP). Essa estatística tem uma distribuição assintoticamente qui-quadrado com um único grau de liberdade:

$$W = \left(\frac{\hat{\beta}_{ij}}{EP\left(\hat{\beta}_{ij}\right)}\right)^2 \sim \chi^2_1$$

As saídas dos programas estatísticos SPSS e SAS apresentam os testes da razão de verossimilhança e de *Wald*.

3.2 Interpretação de parâmetros

A interpretação dos parâmetros estimados no modelo de regressão logística torna-se similar ao caso da regressão múltipla tradicional. No caso de uma variável resposta com *k* níveis, o *k*-ésimo nível será assumido como base e, portanto, pode-se estabelecer *k-1* funções *logit*, contrastando cada nível contra o nível base, conforme segue:

$$g_i(x) = ln\left[\frac{P(y_i \mid x)}{P(y_k \mid x)}\right] = \beta_{i0} + \beta_i' x$$

$$1 \le i \le k-1$$

Aplicando a função exponencial nos dos lados da igualdade:

$$e^{g_i(x)} = \frac{P(y_i \mid x)}{P(y_k \mid x)} = e^{\beta_{i0} + \beta_i' x}$$

$$1 \le i \le k-1$$

Assim, um incremento de uma unidade na variável x_j causará um aumento de $e^{\beta_{ij}}$ unidades na razão entre as proba-

bilidades da observação **x** pertencer a classe y_i em relação à classe y_k. Portanto, quando x_j aumenta em uma unidade, a classe y_i torna-se $e^{\beta_{ij}}$ vezes mais provável que y_k.

4 Aplicações

Nos itens subsequentes são apresentados dois exemplos de aplicação da regressão logística politômica, enfatizando a utilização prática e a interpretação dos modelos estimados. Os bancos de dados utilizados foram encontrados a partir do trabalho de Aeberhard et al. (1994) que fez um comparativo entre uma grande quantidade de métodos de reconhecimento de padrões utilizando dados reais e simulados.

4.1 Reconhecimento de Vinhos *(Aplicação nº 1)*

O banco de dados *Wine Recognition Data* encontra-se disponível na *home page* do Departamento de Informação e Ciências da Computação da Universidade da Califórnia – Irvine e deve-se a Forina et al. (1988). Trata-se do resultado de uma análise química realizada com vinhos provenientes de uma mesma região da Itália, mas derivados de três diferentes cultivares (y_1, y_2, y_3) Um total de 13 características de cada amostra de vinho foi analisado. Os tamanhos amostrais para os três diferentes tipos de cultivares são 59, 71 e 48, considerados suficientes para estimativas confiáveis.

Por razões didáticas, apenas três variáveis, dentre as 13 disponíveis, serão consideradas no presente exemplo (x_1: teor alcoólico, x_2: total de fenóis e x_3: intensidade da cor). A saída do *software* SPSS 10.0 é apresentada na Figura 2.

O teste da razão de verossimilhança resultou altamente significativo (*Deviance = 67,257)* indicando que o modelo estimado pode ser útil na discriminação dos três tipos de cultivares. Os valores *Pseudo R-Square* são uma espécie de coeficiente de determinação (R^2), mas com uma interpretação mais complexa, entretanto segue a regra básica: quanto maior, melhor é o ajuste do modelo. Dentre as três medidas apresentadas dá-se preferência a de *Nagelkerke*, visto ser uma medida no intervalo [0;1]. Nesse caso a medida resultou 0,941, muito próxima do valor máximo.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	386,630			
Final	67,257	319,372	6	,000

Pseudo R-Square

Cox and Snell	,834
Nagelkerke	,941
McFadden	,826

Parameter Estimates

CLASS		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1	Intercept	-29,539	15,974	3,419	1	,064			
	ALCOHOL	1,558	1,255	1,541	1	,214	4,749	,406	55,566
	PHENOLS	7,757	1,909	16,515	1	,000	2338,084	55,477	98538,407
	COLOR	-1,413	,508	7,729	1	,005	,243	8,987E-02	,659
2	Intercept	33,421	14,621	5,225	1	,022			
	ALCOHOL	-2,471	1,196	4,270	1	,039	8,454E-02	8,116E-03	,880
	PHENOLS	6,208	1,829	11,524	1	,001	496,881	13,790	17903,238
	COLOR	-2,919	,668	19,093	1	,000	5,397E-02	1,457E-02	,200

Classification

Observed	Predicted			Percent Correct
	1	2	3	
1	54	5	0	91,5%
2	3	65	3	91,5%
3	1	2	45	93,8%
Overall Percentage	32,6%	40,4%	27,0%	92,1%

Figura 2 – Saída da regressão logística politômica no software SPSS 10.0 para o exemplo do Reconhecimento de Vinhos

As estimativas para os parâmetros do modelo também são apresentadas na Figura 2. O número de parâmetros a ser estimado é de (*k-1*)(*p+1*) parâmetros. No exemplo há três classes e três variáveis (*k=p=3*), o que leva a um total de oito parâmetros. As duas funções *logit* estimadas foram as seguintes:

$g_1(x) = -29{,}539 + 1{,}558x_1 + 7{,}757x_2 - 1{,}413x_3$ (Cultivar Tipo 1)

$g_2(x) = 33{,}421 - 2{,}471x_1 + 6{,}208x_2 - 2{,}919x_3$ (Cultivar Tipo 2).

O Cultivar Tipo 3 foi considerado como base e, portanto, $g_3(x) = 0$. De acordo com o teste de Wald apenas dois parâmetros estimados não resultaram significativos ao nível de 5% (Sig. > 0,05), entretanto verifica-se que todas as variáveis têm coeficientes significativos em pelo menos uma das equações, o que constitui uma situação altamente desejável. Geralmente não há interesse nos testes de hipóteses das constantes (b_{i0}). A aplicação do modelo é simples, sendo necessário inserir os valores de *x* nas funções *logit* para obtenção das probabilidades de pertencer as classes:

$$P(Y = y_1 \mid \mathbf{x}) = \frac{exp\{g_1(x)\}}{1 + exp\{g_1(x)\} + exp\{g_2(x)\}}$$

$$P(Y = y_2 \mid \mathbf{x}) = \frac{exp\{g_2(x)\}}{1 + exp\{g_1(x)\} + exp\{g_2(x)\}}$$

$$P(Y = y_3 \mid \mathbf{x}) = 1 - P(Y = y_1 \mid \mathbf{x}) - P(Y = y_2 \mid \mathbf{x})$$

De acordo com o modelo estimado, um vinho com graduação alcoólica de 12,0º, fenóis totais de 2,5 e intensidade da cor de 6,0 – $\mathbf{x}' = [12{,}0 ; 2{,}5 ; 6{,}0]$ – teria as seguintes probabilidades de classificação:

$P(Y = y_1 \mid \mathbf{x}) \cong 0{,}135$

$P(Y = y_2 \mid \mathbf{x}) \cong 0{,}740$

$P(Y = y_3 \mid \mathbf{x}) \cong 0{,}125$.

Portanto, um vinho com tais características seria classificado como proveniente do Cultivar Tipo 2. A Figura 2 apresenta a tabela de classificação utilizando todas as 178 observações do conjunto de dados, onde percebe-se uma habilidade satisfatória do modelo para classificação, com taxa de acerto de 92,1%.

Ainda explorando a Figura 2, percebe-se a presença da coluna *Exp(B)* e seu respectivo intervalo de confiança. Numa rápida inspeção visual percebe-se que os intervalos de confiança são muito amplos, ocasionados pelos grandes erros-padrão das estimativas. Por meio da interpretação da coluna *Exp(B)* chega-se a interpretações do tipo: a cada aumento de uma unidade na graduação alcoólica, espera-se um aumento de 0,406 a 55,566 vezes na probabilidade do vinho ser proveniente do Cultivar Tipo 1 em relação à probabilidade do vinho pertencer ao Cultivar Tipo 3.

4.2 As Íris de Fisher *(Aplicação nº 2)*

O banco de dados das Íris de Fisher é, sem dúvida, um dos mais famosos conjuntos de observações na área de classificação e discriminação e encontra-se disponível em diversas páginas da Internet. Esse sucesso deve-se ao importante trabalho publicado por Fisher (1936) no qual a análise discriminante foi abordada. Trata-se de um caso onde há três espécies de flores (y_1: Íris Setosa, y_2: Íris Versicolor e y_3: Íris Virgínica) e quatro variáveis independentes (x_1: comprimento da sépala, x_2: largura da sépala, x_3: comprimento da pétala, x_4: largura da pétala). O banco de dados é composto de 150 observações, sendo 50 para cada tipo de flor. A saída do *software* SPSS 10.0 para esse problema é apresentada na Figura 3.

O teste da razão de verossimilhança resultou altamente significativo (*Deviance=11,899)* indicando que pelo menos uma das quatro características pode ser utilizada para discriminação dos três tipos de flores. O valor do coeficiente de determinação de *Nagelkerke* foi praticamente máximo: 0,99.

As estimativas para os dez parâmetros do modelo também são apresentadas na Figura 3. Ocorreram problemas numéricos devido a uma separação completa da classe Íris Setosa, o que comprometeu a parte inferencial do modelo (testes de significância) provocando erros padrão visivelmente "inflados". Também verificou-se alta correlação entre as variáveis independentes o que provoca aumento nos erros padrões e prejudica o procedimento de estimação. Verifica-se que, apesar da ocorrência de tais problemas, as estimativas encontradas continuam sendo úteis, como prova a taxa de classificação correta de 98,7% apresentada na tabela de classificação.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	329,584			
Final	11,899	317,685	8	,000

Pseudo R-Square

Cox and Snell	,880
Nagelkerke	,990
McFadden	,964

Parameter Estimates

Class		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Iris Setosa	Intercept	30,291	2272547	,000	1	1,000			
	SEP_LEN	14,670	667745,9	,000	1	1,000	2350569	,000	,[a]
	SEP_WID	14,474	313392,1	,000	1	1,000	1931522	,000	,[a]
	PET_LEN	-31,522	310957,5	,000	1	1,000	2,043E-14	,000	,[a]
	PET_WID	-43,107	,000	.	1	.	1,900E-19	1,900E-19	1,900E-19
Iris Versicolour	Intercept	42,638	25,708	2,751	1	,097			
	SEP_LEN	2,465	2,394	1,060	1	,303	11,766	,108	1284,293
	SEP_WID	6,681	4,480	2,224	1	,136	797,026	,123	5181847,602
	PET_LEN	-9,429	4,737	3,962	1	,047	8,033E-05	7,457E-09	,865
	PET_WID	-18,286	9,743	3,523	1	,061	1,144E-08	5,828E-17	2,246

a. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

Classification

Observed	Predicted			
	Iris Setosa	Iris Versicolour	Iris Virginica	Percent Correct
Iris Setosa	50	0	0	100,0%
Iris Versicolour	0	49	1	98,0%
Iris Virginica	0	1	49	98,0%
Overall Percentage	33,3%	33,3%	33,3%	98,7%

Figura 3 – Saída da regressão logística politômica no software SPSS 10.0 para o exemplo Fisher Iris Data

As duas funções *logit* estimadas foram as seguintes:

$g_1(x) = 30{,}291 + 14{,}670x_1 + 14{,}474x_2 - 31{,}522x_3 - 43{,}107x_4$ (Setosa)

$g_2(x) = 42{,}638 + 2{,}465x_1 + 6{,}681x_2 - 9{,}429x_3 - 18{,}286x_4$ (Versicolor).

Considerando uma observação $\mathbf{x}' = [4 ; 3{,}5 ; 4 ; 2]$, obtém-se as seguintes probabilidades:

$$P(Y = y_1 \mid \mathbf{x}) \cong 0$$

$$P(Y = y_2 \mid \mathbf{x}) \cong 0{,}831$$

$$P(Y = y_3 \mid \mathbf{x}) \cong 0{,}169$$

nesse caso, uma flor com tais características seria classificada como Íris Versicolor porque a maior probabilidade está associada à classe y_2.

5 Considerações finais

A regressão logística politômica consiste de uma poderosa ferramenta para análise de variáveis qualitativas nominais, apresentando algumas características bastante interessantes e desejáveis em técnicas de modelagem estatística. A primeira

característica refere-se ao fato da regressão logística não fazer suposições sobre o comportamento probabilístico das variáveis independentes. A segunda consiste da possibilidade de estimação direta da probabilidade de uma observação pertencer a determina classe. Por fim, é possível testar a significância de um grande número de variáveis independentes e, assim, eleger as variáveis que contribuem mais para a separabilidade entre as classes.

Como em todas técnicas estatísticas, também há problemas na regressão logística politômica, conforme se pode observar no item 4.2. Um dos principais problemas se dá em casos de separabilidade completa entre as classes, o que inviabiliza uma solução única nas equações de verossimilhança. Segundo Hosmer e Lemeshow (1989) esse problema ocorre principalmente com amostras pequenas acompanhadas de um grande número de variáveis independentes, sendo muito improvável haver separação completa em modelos estimados a partir de amostras substanciais. Uma forma simples de identificar o problema é verificar se há presença de erros padrão exageradamente grandes nas estimativas. Outro problema que ocorre freqüentemente é chamado de colinearidade e se refere à presença de correlação entre as variáveis independentes. A colinearidade é facilmente identificada numa matriz de correlação. Uma solução eficiente para o problema é escolher apenas uma variável quando houver um par de variáveis altamente correlacionadas. A presença de colinearidade também ocasiona erros-padrão grandes.

No item 4.2 os dois problemas mencionados foram identificados e, realmente, os erros padrão das estimativas, especialmente para a classe Íris Setosa, foram muito exagerados. Percebe-se que, mesmo com a ocorrência dos problemas, o modelo apresentou boa habilidade preditiva. Não há como resolver o problema da separabilidade completa entre as classes, mas o problema da colinearidade seria facilmente resolvido com a exclusão de uma ou duas variáveis. No item 4.1 tem-se um exemplo "bem-comportado" onde não ocorreram problemas e, portanto, toda parte inferencial pode ser aproveitada.

Como última consideração, sugere-se que a regressão logística seja utilizada sempre que houver necessidade de entender algum fenômeno onde a variável independente é do nível nominal. No caso de ocorrência de problemas o pesquisador pode optar por técnicas mais simples, caso não haja como resolvê-los, porque a parte inferencial será necessariamente sacrificada. Se o interesse for único e exclusivamente de discriminação entre classes, a amostra pode ser dividida em duas partes: uma para estimação e outra para validação. Mesmo havendo problemas numéricos o modelo pode ser respaldado pelos resultados da amostra de validação.

Referências bibliográficas

AEBERHARD, S; COOMANS, D. e DE VEL, O. (1994) Comparative Analysis of Statistical Pattern Recognition Methods in High Dimensional Settings. Pattern Recognition. Vol. 27, No. 8, p. 1065-77.

ALLISON, P. D. (1999) *Logistic Regression using the SAS System: Theory and Application.* Cary, NC: SAS Institute Inc.

ANDERSON, J. A. (1982) Logistic Discrimination. In *Handbook of Statistics* (Vol. 2) P. R. Krishnaiah and L. Kanal (Eds.) Amsterdam: North-Holland, p. 169-191.

BITTENCOURT, H. R. e CLARKE, R.T. (2002) Use of Logistic Discrimination to Classify Remotely-Sensed -Digital Images. In.: 12TH PORTUGUESE CONFERENCE ON PATTERN RECOGNITION. Proceedings... Aveiro, Portugal: Associação Portuguesa de Reconhecimento de Padrões. BULL, S. and DONNER, A. (1987) The efficiency of multinomial logistic regression compared with multiple group

discriminant analysis. *Journal of the American Satistical Association*. vol. 82, p. 1118-1122.

COX, D.R. and SNELL, E. J. (1989). *The Analysis of Binary Data.* Second Edition. London: Chapman and Hall.

DAY, N. and KERRIDGE, D. (1967) A general maximum likelihood discriminant. *Biometrics*, vol. 23, p. 313-324.

FISHER, R. A. (1936) The use of multiple measures in taxonomic problems. *Annals Eugenica*, vol. 7(II), p. 179-188.

FORINA, M. LEARD, R. ARMANINO C. LAUTER, S. (1988) *Parvus – an extendible package for data exploration, classification and correlation*. Institute of Pharmaceutical and Food Analysis and Technologies, Genoa – Italy.

HOSMER, D. and LEMESHOW, S.. (1989) *Applied Logistic Regression*. New York: John Wiley & Sons.

KRZANOWSKY, W. J. (1988) *Principles of Multivariate Analysis*. Oxford: Clarendon Press.

McLACHLAN, G. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.

NORUSIS, M. J. and SPSS Inc. (1999) *SPSS Regression Models 10.0*. Chicago, IL: SPSS Inc.

TRUETT, J. CORNFIELD, J. and KANNEL, W. (1967) A multivariate analysis of the risk of coronary heart disease in Framinghan. *Journal of Chronic Diseases*. v. 20, p. 511-524.

WALKER, S. H. and DUNCAN, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika* vol. 54, p. 167-169.