COMPARAÇÃO DA REGRESSÃO LOGÍSTICA E ANÁLISE DISCRIMINANTE

Miranda Albino Martins Muaualo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Basílio de Bragança Pereira

Rio de Janeiro
Fevereiro de 2013

COMPARAÇÃO DA REGRESSÃO LOGÍSTICA E ANÁLISE DISCRIMINANTE

Miranda Albino Martins Muaualo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

Prof. Basílio de Bragança Pereira, Ph.D.

Prof. Edilson Fernandes Arruda, D.Sc.

Profa. Marina Silva Paez, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2013

Muaualo, Miranda Albino Martins

Comparação da Regressão Logística e Análise Discriminante/Miranda Albino Martins Muaualo. – Rio de Janeiro: UFRJ/COPPE, 2013.

XII, 57 p.: il.; 29, 7cm.

Orientador: Basílio de Bragança Pereira

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Produção, 2013.

Referências Bibliográficas: p. 44 – 46.

1. Regressão Logística e Análise Discriminante. 2. Eficiência Assintótica Relativa. 3. Classificação de indivíduos. 4. Classificações corretas. I. Pereira, Basílio de Bragança. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

Dedico esta dissertação aos meus pais, Albino Martins e Rosalina, e meus tios, Alfredo Armando Malico e Maria, que me propiciaram uma vida digna onde eu pudesse crescer acreditando que tudo é possível, desde que sejamos honestos, íntegros de caráter e tendo a convicção de que a perseverança seja uma ação contínua em nossas vidas. Além de extenso amor e carinho, também me proporcionaram os conhecimentos de procurar sempre em Deus à benção e força maior para o meu reavivamento como crente.

Agradecimentos

Gostaria de expressar o meu agradecimento ao pai Divino por tudo que tem me proporcionado, e a todos quantos de várias formas me foram suavizando a longa viagem que me encaminhou até ao trabalho que a seguir apresento. Por detrás das nossas realizações pessoais, além de um esforço pessoal, esconde-se normalmente um número considerável de contribuições, apoios, sugestões, críticas vindas de muitas pessoas. A sua importância assume no caso presente uma valia preciosa que, sem elas, teria sido muito difícil chegar a quaisquer resultados dignos de menção. Por isso, agradeço à CNPq/MCT-MZ pelo apoio financeiro, pois, sem o mesmo não seria possível percorrer toda viagem. Aos Professores Basílio de Bragança Pereira e Edilson Fernandes de Arruda pelo apoio acadêmico, críticas construtivas, disponibilidade de todos os momentos, e pelo exemplo de Professores afáveis e sapientes acompanhar-me-á como uma referência pelo resto da minha vida.

Aos Professores do Programa de Pós-Graduação em Engenharia de Produção da COPPE/UFRJ da área de Pesquisa Operacional e do Instituto de Matemática da UFRJ do Departamento de Métodos estatísticos pelo ensinamento e paciência. À equipe de trabalho em especial aos seguranças, faxineiros, e equipe da secretaria do PEP e da PO pela prontidão e simpatia. Não poderia me esquecer da Andréia Lima da Silva Moreira e Roberta de Mattos Arruda da PO e PEP respectivamente, pela simpatia e carinhoso atendimento. Mencionar o nome delas aqui constitui um preito de justiça e de homenagem sentida por parte do autor deste trabalho.

À Dra. Emília Matos, ao Cristian e Juan Dell' Oso pela ajuda durante todo processo de aquisição dos códigos e do processamento dos dados, e em especial ao Cachimo pela força dada através do companheirismo sempre que precisei. À Gladys por ter me ajudado na ideia de fazer de otimismo a maneira de viver.

Á Direção do DMI-UEM pela permissão que deu em continuar a estudar. Aos Professores Manuel Álves, Carvalho Madivate, ao Dr. Lino Marques, aos meus amigos Nordino, Teodósio, Cláudio, Francisco da Conceição, Nerito e Paulo pela força e incentivo. Aos meus irmãos e familiares pela força que sempre me deram, e aos meus colegas de turma de mestrado, pela colaboração de modo que este curso fosse realidade. Ao mano Waite e à minha família que desde a concepção da ideia inicial até a sua fase final, me deram todo o apoio moral, endereço a minha gratidão.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

COMPARAÇÃO DA REGRESSÃO LOGÍSTICA E ANÁLISE DISCRIMINANTE

Miranda Albino Martins Muaualo

Fevereiro/2013

Orientador: Basílio de Bragança Pereira

Programa: Engenharia de Produção

Este trabalho compara a regressão logística e análise discriminante, e a regressão logística multinomial e análise discriminante com múltiplos grupos, em termos das suas eficiências. Essa comparação é feita de duas formas.

Primeiro verifica-se se as variáveis explicativas satisfazem as suposições da análise discriminante, ou seja, se as observações provêm de populações com distribuição normal multivariada, e se há homogeneidade das matrizes de variâncias e covariâncias. Uma vez satisfeitas tais suposições, essas variáveis são utilizadas para a estimação dos modelos de interesse e faz-se a comparação da eficiência relativa e da eficiência assintótica relativa desses modelos com duas populações, mediante as suas probabilidades de classificações corretas e valores da eficiência assintótica da regressão logística comparada com análise discriminante apresentados neste trabalho. Após isso, faz-se a comparação da eficiência relativa dos modelos de quatro populações através das probabilidades de classificações corretas.

A segunda e última forma é feita utilizando as variáveis que violam as suposições da análise discriminante, a fim de verificar a variação da eficiência relativa com e sem inclusão das variáveis que satisfazem as referidas suposições. Nesta última análise, somente é feita a comparação da eficiência relativa.

Os resultados deste trabalho revelam que em ambas as formas, a regressão logística ou a regressão logística multinomial é mais eficiente em relação à análise discriminante de dois ou mais de dois grupos respectivamente. Porém, quando as variáveis explicativas satisfazem as suposições da análise discriminante, a regressão logística é assintoticamente menos eficiente que a análise discriminante. Para outros membros da família exponencial a regressão logística é mais eficiente.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS

Miranda Albino Martins Muaualo

February/2013

Advisor: Basílio de Bragança Pereira

Department: Production Engineering

This work compares logistic regression and discriminant analysis, and multinomial logistic regression and discriminant analysis with multiple groups in terms of their efficiencies. This comparison is done in two ways.

Firstly, we verify whether the explanatory variables satisfy the assumptions of discriminant analysis, ie, if the observations come from populations with multivariate normal distribution, and if there is homogeneity in both variance and covariance matrices. Once assumptions are satisfied these variables are used for the estimation of models of interest and we compare the relative efficiency and the asymptotic relative efficiency of these models with two populations both by their probabilities of correct classification and by the values of the asymptotic efficiency of logistic regression compared with discriminant analysis. Then, we make a comparison of the relative efficiency of models across four populations with respect to the probabilities of correct classifications.

The second and final analysis is made by using the variables that violate the assumptions of discriminant analysis, in order to verify the variation of relative efficiency with and without the inclusion of variables that meet these assumptions. In this analysis, only a comparison of the relative efficiencies is performed.

The results of this study show that in both forms, the logistic regression or multinomial logistic regression is more efficient than discriminant analysis of two or more than two groups respectively. But, when the explanatory variables satisfy the assumptions of the discriminant analysis, logistic regression is asymptotically less efficient than discriminant analysis. For other exponential family members the logistic regression is more efficient.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A análise discriminante e a regressão logística são métodos utilizados para classificação de indivíduos para certo grupo. Ambos os métodos são da análise multivariada de dados e de grande interesse em aplicações biomédicas, econômicas e na área de ensino, mas diferem nas suas suposições e no tipo de variáveis explicativas.

A forma mais poderosa de análise em estudos longitudinais é a abordagem da função discriminante (TRUETT et al., 1967), proposta por CORNFIELD, GORDON e SMITH (1961) (apud GORDON, 1974). Apesar disso, muitos pesquisadores preferem utilizar o método iterativo dos mínimos quadrados ponderados, equivalente ao método da máxima verossimilhança, e que não requer suposições rígidas (PRESS e WILSON, 1978). Esse método foi proposto por WALKER e DUNCAN (1967) e DAY e KARREDGE (1967).

Na maioria das vezes, a estimação da função discriminante linear de Fisher é feita através de um conjunto de observações que provêm de duas populações com distribuição normal multivariada, e com homogeneidade das matrizes de variâncias e covariâncias. Para isso, substituem-se os estimadores dos parâmetros da máxima verossimilhança incondicional nas expressões dos coeficientes (β_0, β'). Além disso, usando as mesmas suposições da análise discriminante, pode-se estimar a regressão logística (EFRON, 1975).

Nesses procedimentos, a classificação de indivíduos é feita com base na função discriminante estimada (EFRON, 1975). Se o escore discriminante for maior que zero, ou seja, se $\lambda(x) > 0$, o indivíduo é classificado para a população 1; caso contrário, o indivíduo é classificado para a população 0. Esse método de classificação minimiza a probabilidade esperada de erro de classificação (EFRON, 1975).

A estimação da regressão logística com base nas suposições da análise discriminante não é comum em aplicações. Porém, no caso de serem satisfeitas a distribuição normal multivariada das variáveis explicativas e a homogeneidade das matrizes das variâncias e covariâncias, a regressão logística é assintoticamente menos eficiente do que a análise discriminante (EFRON, 1975).

Por outro lado, quando a normalidade multivariada não é satisfeita, a regressão logística é mais robusta, ou seja, a regressão logística pode ser utilizada nas situações em que a normalidade multivariada não é satisfeita. Ao passo que, quando a homogeneidade das matrizes de variâncias e covariâncias não é satisfeita, TRUETT et al. (1967) apontam que a expressão do expoente da regressão logística é quadrática em vez de linear nas variáveis explicativas.

PRESS e WILSON (1978), baseando-se em dados coletados de pacientes com câncer de mama (entre 1955 e 1963), e em dados de mudança populacional coletados para os 50 estados dos EUA, constataram que, em ambos os casos, a regressão logística superou a análise discriminante em termos da proporção das classificações corretas.

Atualmente, constatou-se que a regressão logística é mais flexível e robusta, e os métodos adaptativos não-lineares de aprendizagem (modelos aditivos generalizados, árvores de classificação, *MARS-Multivariate Adaptive Regression Splines*) não superam um modelo de regressão logística relativamente simples (POHAR et al., 2004, ENNIS et al., 1998).

No caso de comparação da eficiência dos dois modelos, EFRON (1975) definiu primeiro os procedimentos da análise discriminante e da regressão logística, e com esses procedimentos obteve a expressão da medida da eficiência relativa. Para a obtenção dessa expressão, baseou-se nas probabilidades de erro de classificação (taxa de erro na situação ótima, ou seja, quando $\lambda(x) = 0$, e taxa de erro quando $n \to \infty$), tendo em conta o procedimento da análise discriminante baseado nas estimativas de máxima verossimilhança e da regressão logística baseada na verossimilhança condicional com respeito a (β_0, β) que foi apresentada na forma da família exponencial.

O fato da verossimilhança da regressão logística poder ser escrita na forma da família exponencial (EFRON, 1975) esclarece porque os coeficientes logísticos estimados não são assintoticamente enviesados, são consistentes e o modelo apresenta uma estatística suficiente com vetor média e matriz de covariâncias. Esses parâmetros comprovam que este modelo provém da amostra de duas populações que têm distribuição normal. É nesse contexto, que EFRON (1975) definiu a expressão da eficiência assintótica relativa, tendo em conta as suposições da análise discriminante e a aleatoriedade da taxa de erro.

Como os resultados de EFRON (1975) baseiam-se em duas populações, BULL e DONNER (1987a e 1987b) estenderam esses resultados para mais de duas populações. A análise da eficiência assintótica feita por BULL e DONNER (1987a) consistiu em dois casos específicos: nos quais os β_{sj} foram avaliados quando são iguais à zero, e supondo que a eficiência assintótica relativa vai atingir um máximo nos β_{sj}, para distâncias fixas entre populações e frequências fixas da população.

Primeiro, foi considerado o caso de coeficientes logísticos iguais à zero, através de teste de hipóteses. Porém, para casos em que houvesse duas variáveis explicativas não correlacionadas, foi considerado que os coeficientes também são iguais a zero.

No segundo caso, foram considerados vetores de coeficientes logísticos diferentes de zero, também mediante o teste de hipóteses.

Nessa análise, constatou-se que no caso de vetores de coeficientes logísticos nulos, a presença de mais um grupo de resposta pode ter uma influência substancial na eficiência assintótica relativa, sobretudo para frequências iguais de grupo de resposta. E no caso de vetores de coeficientes logísticos diferentes de zero, os autores constataram que a eficiência relativa assintótica geralmente diminui na presença de um grupo adicional de resposta (BULL e DONNER, 1987a).

Neste trabalho, abordamos a teoria sobre a regressão logística (RL) e análise discriminante linear de Fisher (ADL), e comparamos empiricamente a eficiência relativa e assintótica relativa de ambos os métodos.

1.1 Estrutura do trabalho

No capítulo 2 é apresentada a teoria sobre a análise discriminante, as suposições, os métodos de seleção das variáveis, e o método de estimação da função discriminante para duas ou mais de duas populações.

No capítulo 3 faz-se a abordagem sobre a regressão logística. Nela é apresentado o método de estimação da regressão logística, a regressão logística multinomial, testes estatísticos utilizados, e os métodos de classificação de indivíduos nos modelos de interesse. No capítulo 4 desenvolve-se a teoria sobre a comparação da eficiência. No capítulo 5 é feita a aplicação, a presentação dos resultados e sua interpretação. A discussão, conclusões e recomendações são apresentadas no capítulo 6, e por fim são apresentadas as referências bibliográficas.

Capítulo 2

Análise discriminante

A análise discriminante é uma técnica multivariada apropriada quando a variável dependente é nominal, e as variáveis independentes são métricas. Quando a variável dependente com duas classes é envolvida no estudo, a análise discriminante é dita de dois grupos. Em outros casos, uma variável dependente com mais de duas classes é envolvida no estudo, podendo chamar-se de análise discriminante com múltiplos grupos ou análise discriminante múltipla.

Por exemplo, num estudo pode ser envolvida uma variável dependente com as classificações do tipo tem a doença e não tem a doença ou baixo, médio e alto. Nesse caso, a análise discriminante de dois grupos é apropriada para classificação dos sujeitos ou análise discriminante para múltiplos grupos respectivamente.

Muitas das vezes, uma mistura do tipo de variáveis explicativas, binárias e contínuas, pode enviesar a estimação da função discriminante (BULL e DONNER, 1987a).

2.1 Suposições da análise discriminante

As duas principais suposições da análise discriminante são: a normalidade multivariada das variáveis explicativas, e a homogeneidade das matrizes de variâncias e covariâncias (EFRON, 1975, PRESS e WILSON, 1978, BULL e DONNER, 1987a).

A suposição da normalidade multivariada não é satisfeita em aplicações, o que se faz é verificar se as distribuições marginais não são muito longe da normalidade. Por outro lado, HALPERIN et al. (1971) são mais rigorosos ao afirmarem que mesmo com a aproximação baseada nas distribuições, não se garante a normalidade.

A homogeneidade das matrizes de variâncias e covariâncias é avaliada através da estatística de teste M-Box (MARDIA et al., 1979). Este teste é muito sensível à dimensão das amostras, ou seja, amostras grandes conduzem geralmente à rejeição da H_0 mesmo que as diferenças entre as matrizes de variâncias e covariâncias sejam muito pequenas. Não obstante, a análise discriminante é uma técnica bastante

robusta à violação dos pressupostos desde que a dimensão de menor grupo seja superior ao número de variáveis em estudo, e que as médias dos grupos não sejam proporcionais às suas variâncias.

2.2 Métodos de seleção das variáveis

Os métodos de seleção de variáveis para a estimação da função discriminante são o simultâneo e o *stepwise*. O método *stepwise* consiste na inclusão das variáveis independentes na função discriminante, com base em seu poder discriminatório uma por vez. Assim, seleciona-se a melhor variável discriminante em cada passo, e em seguida, as variáveis que não são úteis na discriminação entre os grupos são eliminadas. Por conseguinte, é identificado um conjunto reduzido de variáveis.

Este método torna-se menos estável e generalizável à medida que a proporção entre o tamanho da amostra e as covariáveis diminui abaixo do nível recomendado de 20 observações por variável explicativa. Por outro lado, é apropriado quando se pretende considerar um número relativamente grande de variáveis explicativas para inclusão na função, e é considerado como alternativo ao método simultâneo.

O método simultâneo consiste na obtenção da função discriminante considerando juntas todas as covariáveis. Desse modo, a função discriminante é obtida com base no conjunto todo de variáveis independentes sem considerar o poder discriminatório de cada uma delas.

Neste trabalho, utilizou-se o método simultâneo. Com este método, foram incluídas todas as variáveis independentes na análise, sem interesse de ver os resultados intermediários baseados apenas nas variáveis mais discriminantes.

2.3 Estimação das funções discriminantes

Após a seleção das variáveis explicativas, o objetivo seguinte é de utilizá-las em combinação, para melhorar o poder discriminatório de qualquer variável individual, de forma a criar a função discriminante.

O primeiro passo na obtenção da função discriminante é a escolha do método de seleção de variáveis (simultâneo ou *stepwise*). Em seguida, determina-se o número de funções a serem obtidas. Geralmente considera-se p-variáveis e g grupos para estabelecer $m = min\,(g-1, p)$ funções discriminantes que são combinação linear das p-variáveis.

As variáveis utilizadas para estimação da função discriminante devem provir de uma das duas populações que satisfazem as suposições da análise discriminante, ou seja, $X \sim N_p\,(\mu_j, \Sigma)$ com probabilidade $\pi_j\,(j = 0, 1)$. Com essas suposições satisfeitas, conhecendo os parâmetros π_1, $\pi_0 = 1 - \pi_1$, μ_1, μ_0, e Σ (π_1 é o risco de

certo indivíduo pertencer ao grupo 1, π_0 é o risco de certo indivíduo pertencer ao grupo 0, μ_1 e μ_0 são as médias do grupo 1 e grupo 0, e Σ é a matriz de covariâncias), e sendo que $\mathbb{X}$ é a matriz de covariáveis a função discriminante linear de Fisher para duas populações será obtida pela seguinte expressão (EFRON, 1975):

$$\lambda(X) = \beta_0 + \beta'\mathbb{X} \tag{2.1}$$

Para estimar os coeficientes da função discriminante linear utilizam-se os seguintes estimadores de máxima verossimilhança (POHAR et al., 2004, EFRON, 1975, TRUETT et al., 1967):

$$\hat{\mu}_1 = \bar{X}_1 \equiv \frac{\sum\limits_{yj=1} X_j}{n_1}, \qquad \hat{\mu}_0 = \bar{X}_0 \equiv \frac{\sum\limits_{yj=0} X_j}{n_0} \tag{2.2}$$

$$\hat{\pi}_1 = \frac{n_1}{n}, \qquad \hat{\pi}_0 = \frac{n_0}{n}, \qquad n = n_0 + n_1$$

$$\hat{\Sigma} = \frac{\left[\sum\limits_{yj=1}(X_j - \bar{X}_1)(X_j - \bar{X}_1)' + \sum\limits_{yj=0}(X_j - \bar{X}_0)(X_j - \bar{X}_0)'\right]}{n}$$

onde n_1 e n_0 são o número de indivíduos da população 1 e da população 0 respectivamente, yj indica que população X_j provém, e substituindo os estimadores de máxima verossimilhança nas equações seguintes:

$$\beta_0 \equiv log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\left(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0\right) \qquad \beta' \equiv (\mu_1 - \mu_0)'\Sigma^{-1} \tag{2.3}$$

obtêm-se os coeficientes β_0 e β' estimados, e o procedimento de discriminação normal será expresso na forma (EFRON, 1975):

$$\hat{\lambda}(X) = \hat{\beta}_0 + \hat{\beta}'\mathbb{X} \tag{2.4}$$

No caso da análise discriminante com múltiplos grupos, consideram-se $J + 1$ populações p-dimensionais, tendo em conta com as suposições da análise discriminante, tais que (BULL e DONNER, 1987a):

$$X \sim N_p(\mu_j, \Sigma) \quad \text{com probabilidade} \quad \pi_j \, (j = 0, 1, \ldots, J)$$

onde $\pi_j > 0$ para todo j e $\sum_{j=0}^{J} \pi_j = 1$, $\pi_0 = 1 - \sum_{j=1}^{J} \pi_j$. Observando n realizações independentes da variável aleatória X denotada pelos pares (z_i, x_i) $i = 1, 2, \ldots, n$ onde z_i indica em que população o x_i provem, e x_i é um vetor das medições. Dadas as suposições, o procedimento da discriminação normal, baseado na verossimilhança total $L_1 = \Pi_{i=1}^{n} f(z_i, x_i)$ produz estimativas de máxima

verossimilhança dos parâmetros normais π_j, μ_j $(j = 1, \ldots, J)$, μ_0 e Σ. Assim, os parâmetros β_{0j} e β_j são estimados pelas formas (BULL e DONNER, 1987a, HAGGSTROM, 1983):

$$\hat{\beta}_{0j} = log\left(\frac{\pi_j}{\pi_0}\right) - \frac{1}{2}\left(\hat{\mu}_j'\hat{\Sigma}^{-1}\hat{\mu}_j - \hat{\mu}_0'\hat{\Sigma}^{-1}\hat{\mu}_0\right)$$

$$\hat{\beta}_j = \hat{\Sigma}^{-1}(\hat{\mu}_j - \hat{\mu}_0) \qquad j = (1, \ldots, J). \tag{2.5}$$

Portanto, os coeficientes da função discriminante são estimados de modo que a variabilidade dos escores da função discriminante linear de Fisher seja máxima entre os grupos e mínima dentro dos grupos, ou seja, de modo que:

$$\lambda_i = \frac{SQF(\lambda_i)}{SQE(\lambda_i)} \tag{2.6}$$

seja máxima. Onde $SQE(\lambda_i)$ é a soma dos quadrados dentro dos grupos ou a variabilidade dentro dos grupos e $SQF(\lambda_i)$ é a soma dos quadrados entre os grupos ou a variabilidade entre os grupos.

Capítulo 3

Regressão Logística

Os modelos da regressão logística são utilizados quando a variável dependente Y é do tipo nominal dicotômica $(0, 1)$. As covariáveis associadas à regressão logística podem ser todas contínuas, todas categóricas ou uma combinação dos dois tipos. Em geral, estes modelos são utilizados para modelar a ocorrência em termos probabilísticos, de uma das duas realizações das classes da variável.

Para uma variável dependente binária Y e uma variável independente X, sendo $\hat{\pi}_j = P[Y_j = 1]$ função usada para estimar a probabilidade de uma determinada realização $j(j = 1, \ldots, n)$ da variável resposta ser um sucesso, o modelo de regressão logística binária é dado por (TRUETT et al., 1967):

$$\hat{\pi}_j = \frac{exp(\beta_0 + \beta_1 X_j)}{1 + exp(\beta_0 + \beta_1 X_j)} \tag{3.1}$$

Na presença de mais de uma covariável $(X_1, \ldots, X_m)$ no estudo, o modelo é chamado de regressão logística múltipla e é dado pela seguinte expressão:

$$\hat{\pi}_j = \frac{exp(\beta_0 + \beta_1 X_{1j} + \cdots + \beta_p X_{pj})}{1 + exp(\beta_0 + \beta_1 X_{1j} + \cdots + \beta_p X_{pj})} \tag{3.2}$$

Por conseguinte, a probabilidade de uma determinada realização $j(j = 1, \ldots, n)$ da variável resposta ser o insucesso, pode ser estimada da seguinte forma:

$$P[Y_j = 0] = 1 - P[Y_j = 1] = (1 - \hat{\pi}_j) = \frac{1}{1 + exp(\beta_0 + \beta_1 X_{1j} + \cdots + \beta_p X_{pj})} \tag{3.3}$$

Da expressão (3.2), obtém-se a seguinte forma matricial do modelo de regressão logística:

$$\hat{\pi} = \frac{exp(\mathbb{X}\beta)}{1 + exp(\mathbb{X}\beta)} \tag{3.4}$$

onde:

$$\mathbb{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix} \tag{3.5}$$

ou seja, $\mathbb{X}$ é a matriz das covariáveis, cuja primeira coluna é o vetor das unidades;

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{3.6}$$

ou seja, β é o vetor dos $p+1$ coeficientes de regressão logística; e $\hat{\pi}$ é o vetor de probabilidades estimadas.

Ao representar graficamente um conjunto de dados de Y e X, sendo que Y toma apenas os valores $Y = 1$ (sucesso) ou $Y = 0$ (insucesso), é fácil constatar que a variação de Y não tem sido linear e aditiva, e que o valor esperado de Y é $\hat{Y} = E(Y|X) = \Sigma(Y|n) = P\{Y = 1\} = \pi$. Não obstante, o modelo de regressão logística pode ser ajustado recorrendo à regressão não linear. A solução tradicional consiste em linearizar a função (3.4) com a transformação $Logit(\hat{\pi})$ que é dada por COX e SNELL (1989) da seguinte forma:

$$Logit(\hat{\pi}) = log\frac{\hat{\pi}}{1-\hat{\pi}} = \beta\mathbb{X} \tag{3.7}$$

onde β é o vetor coluna dos parâmetros, $\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)$, $\mathbb{X}$ é a matriz das covariáveis. Assim, o modelo da regressão logística para múltiplas covariáveis é:

$$Logit(\hat{\pi}_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_p X_{pj}. \tag{3.8}$$

A função *Logit* é chamada por função de ligação nos modelos lineares generalizados, e ela permite linearizar a variável dependente, podendo ser modelada em função de um modelo linear. A outra função de ligação comumente usada nos modelos com variáveis dependentes qualitativas binárias é a função de distribuição normal, e neste caso o modelo chama-se modelo *Probit*.

A razão de verossimilhança $\frac{\hat{\pi}}{(1-\hat{\pi})}$ da função *Logit*, que também é denominada razão de chances ou simplesmente chances, é a razão entre a probabilidade de sucesso (π) e a probabilidade de insucesso ($1-\pi$), ou seja, as chances (risco) de se observar o sucesso ($Y = 1$) em relação ao insucesso ($Y = 0$). A razão de verossimilhança é designada por razão de desigualdades (HAIR et al, 2009).

3.1 Estimação do modelo de regressão logística

A regressão logística é comumente estimada através do método iterativo dos mínimos quadrados ponderados análogo ao método da máxima verossimilhança sugerido por WALKER e DUNCAN (1967) e DAY e KERRIDGE (1967) (EFRON, 1975, BULL e DONNER, 1987a e 1987b). Esta abordagem utiliza estimativa da máxima verossimilhança com base na probabilidade condicional, dadas as variáveis explicativas.

A probabilidade de um evento ocorrer é expressa em (3.2) (EFRON, 1975, BROOKS et al., 1988). Além disso, a natureza das variáveis dependentes, $Y_i's$, da regressão logística é binária (0 ou 1). Por isso, cada uma das observações é um ensaio de Bernoulli, por conseguinte, a equação da máxima verossimilhança é derivada a partir da seguinte distribuição de probabilidade da variável dependente (BROOKS et al., 1988, HOSMER e LEMESHOW, 2000):

$$P(Y = y_i) = \pi^{y_j}(1-\pi)^{1-y_j}. \tag{3.9}$$

Como as n observações de Y são independentes, a função de verossimilhança que dá a probabilidade conjunta $P(Y = y_1, Y = y_2, \ldots, Y = y_n)$ de se observarem os valores $(y_1, y_2, \ldots, y_n)$ amostrados é (BROOKS et al., 1988, HOSMER e LEMESHOW, 2000):

$$l(\beta) = \Pi_{j=1}^{n} \pi_j^{y_j}(1-\pi_j)^{1-y_j} \tag{3.10}$$

Assim, o logaritmo da verossimilhança (3.10) é (BROOKS et al., 1988):

$$ln\left[l(\beta)\right] = \sum_{j=1}^{n} y_j ln(\pi_j) + \sum_{j=1}^{n}(1-y_j)ln(1-\pi_j). \tag{3.11}$$

As equações de máxima verossimilhança são obtidas achando a derivada de (3.11) com respeito aos $\beta's$ (BROOKS et al., 1988):

$$\frac{\partial ln\left[l(\beta)\right]}{\partial \beta_0} = \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \pi_j = 0 \qquad \frac{\partial ln\left[l(\beta)\right]}{\partial \beta} = \sum_{j=1}^{n} y_j x_j - \sum_{j=1}^{n} x_j \pi_j = 0. \tag{3.12}$$

O algoritmo utilizado para resolver as equações (3.12) é o de Newton-Rapson, e os $\beta's$ são estimados iterativamente. A função de verossimilhança (3.10) considera que as covariáveis são quantitativas. Na presença de covariáveis qualitativas, o *log* da função de verossimilhança é (HOSMER e LEMESHOW, 2000):

$$LL = \sum_{j=1}^{J} y_j Ln(\hat{\pi}_j) + (n_j - y_j)Ln(1-\hat{\pi}_j). \tag{3.13}$$

3.2 Regressão Logística Policotômica

O modelo de regressão logística múltipla utiliza uma variável aleatória dependente binária, ou seja, assume somente dois valores (0 e 1). Este modelo pode ser generalizado de modo que a sua variável aleatória dependente apresente mais de duas classes. Nesse caso, chama-se regressão logística policotômica ou regressão logística multinomial.

A regressão logística multinomial considera $p+1$ variáveis explicativas denotadas por $X = (X_0, X_1, X_2, \ldots, X_p)$ onde $X_0 = 1$ e uma variável aleatória dependente Y de natureza nominal policotômica que pode assumir as classes $j = 0, 1, 2, \ldots, q$.

Seja $\pi_j(X) = P(Y = j|X)$ em um conjunto fixo X de variáveis explicativas, com $\sum_j \pi_j(X) = 1$. A probabilidade da variável dependente Y tomar o valor de qualquer uma das $q + 1$ classes da regressão logística multinomial é dada por:

$$P(Y = 0|X) = \frac{exp\left(\beta_{00} + \beta_{01}X_1 + \cdots + \beta_{0p}X_p\right)}{exp\left(\beta_{00} + \beta_{01}X_1 + \cdots + \beta_{0p}X_p\right) + \cdots + exp\left(\beta_{q0} + \beta_{q1}X_1 + \cdots + \beta_{qp}X_p\right)}$$

$$P(Y = 1|X) = \frac{exp\left(\beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p\right)}{exp\left(\beta_{00} + \beta_{01}X_1 + \cdots + \beta_{0p}X_p\right) + \cdots + exp\left(\beta_{q0} + \beta_{q1}X_1 + \cdots + \beta_{qp}X_p\right)}$$

$$\cdots \quad \cdots \quad \cdots$$

$$P(Y = q|X) = \frac{exp\left(\beta_{q0} + \beta_{q1}X_1 + \cdots + \beta_{qp}X_p\right)}{exp\left(\beta_{00} + \beta_{01}X_1 + \cdots + \beta_{0p}X_p\right) + \cdots + exp\left(\beta_{q0} + \beta_{q1}X_1 + \cdots + \beta_{qp}X_p\right)}$$

A sua forma matricial é:

$$P(Y = 0|X) = \frac{exp\left(\mathbb{X}\beta_0\right)}{exp\left(\mathbb{X}\beta_0\right) + exp\left(\mathbb{X}\beta_1\right) + \cdots + exp\left(\mathbb{X}\beta_q\right)} \tag{3.14}$$

$$P(Y = 1|X) = \frac{exp\left(\mathbb{X}\beta_1\right)}{exp\left(\mathbb{X}\beta_0\right) + exp\left(\mathbb{X}\beta_1\right) + \cdots + exp\left(\mathbb{X}\beta_q\right)} \tag{3.15}$$

$$\cdots \quad \cdots \quad \cdots$$

$$P(Y = q|X) = \frac{exp\left(\mathbb{X}\beta_q\right)}{exp\left(\mathbb{X}\beta_0\right) + exp\left(\mathbb{X}\beta_1\right) + \cdots + exp\left(\mathbb{X}\beta_q\right)} \tag{3.16}$$

Assim, o modelo de regressão logística multinomial consiste num conjunto de k modelos logísticos corrigidos. Dado que o sistema é indeterminado, é necessário normalizá-lo relativamente a uma categoria da variável dependente, e um dos coeficientes referentes a uma das classes tem de ser igualado a 0. Desse modo, as chances de ocorrer uma das classes da variável dependente em relação à classe de referência 0 são:

$$\frac{P(Y = 1|X)}{P(Y = 0|X)} = exp\left(\mathbb{X}\beta_1\right) \cdots \quad \cdots \quad \cdots \frac{P(Y = q|X)}{P(Y = 0|X)} = exp\left(\mathbb{X}\beta_q\right) \tag{3.17}$$

Os modelos $Ln\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \mathbb{X}\beta_1 \cdots \quad \cdots \quad \cdots Ln\left[\frac{P(Y=q|X)}{P(Y=0|X)}\right] = \mathbb{X}\beta_q$ descrevem simultaneamente os efeitos de $X's$ sobre estes q *logits*, e os efeitos variam de acordo com a resposta emparelhada com a linha de base.

As medidas da qualidade do ajustamento da regressão logística multinomial são as mesmas da regressão logística multipla. Porém, a razão das chances relativo à classe de referência 0 é calculada de seguinte modo:

$$Exp(\beta_{ci}) = \frac{P\left(Y = c|X_i = x_i + 1\right)/P\left(Y = 0|X_i = x_i + 1\right)}{P\left(Y = c|X_i = x_i\right)/P\left(Y = 0|X_i = x_i\right)} \tag{3.18}$$

onde $c(c = 1, \ldots, q)$ é a classe da variável dependente relativamente a variável independente $i(i = 1, \ldots, p)$.

A razão das chances na regressão logística multinomial é sempre em relação à classe de referência. Se for necessário calcular a razão das chances relativamente à outra classe, por exemplo, da classe 2 relativamente à classe 1, estas podem obter-se considerando a razão entre as probabilidades de se observarem as duas classes de interesse.

3.3 Testes estatísticos

Nesta subseção, são apresentados os testes estatísticos que foram utilizados para verificar as suposições da análise discriminante, avaliar a significância dos coeficientes individualmente do modelo ajustado, a significância global dos coeficientes, e a qualidade dos modelos de interesse.

3.3.1 Testes para verificar as suposições da análise discriminante

Os testes utilizados para verificar se as suposições da análise discriminante foram satisfeitas são o teste Shapiro-Wilk e teste M-Box (HAIR et al, 2009). A estatística de teste Shapiro-Wilk é utilizada para testar a hipótese de que a amostra provém de uma população normal. Se o p-valor associado à estatística deste teste for menor que o nível de significância selecionado, rejeita-se a hipótese nula, caso contrário conclui-se que a amostra provém de uma população normal.

A estatística de teste M-Box é utilizada para testar se as matrizes das variâncias e covariâncias são homogêneas (HAIR et al, 2009). A hipótese testada é de que há homogeneidade das matrizes das variâncias e covariâncias. Se o p-valor associado à estatística deste teste for menor que o nível de significância selecionado, rejeita-se a hipótese nula de que as matrizes das variâncias e covariâncias são homogêneas.

3.3.2 Teste de Λ de Wilks

O teste de Λ de Wilks é utilizado para testar se os centroides dos grupos (a média de todas as m funções discriminantes para cada um dos g grupos) são iguais (MARDIA et al., 1979):

$$H_0 : \mu_{ij} = \mu_{kl} \quad vs. \quad H_a : \exists(i,j);(k,l) : \mu_{ij} \neq \mu_{kl}$$

$$(i \neq k \quad e \quad i,k = 1,\ldots,m; j \neq l \quad e \quad j,l = 1,\ldots,g).$$

A estatística de teste de Λ de Wilks é calculada a partir do determinante da matriz da soma dos quadrados e produtos cruzados dentro dos grupos, e do determinante da matriz da soma dos quadrados e produtos cruzados totais (MARDIA et al., 1979):

$$\Lambda = \frac{|W|}{|T|}. \tag{3.19}$$

Sob a hipótese nula (H_0), como a distribuição do Λ é desconhecida, a transformação proposta é (MARDIA et al., 1979):

$$\chi^2 = -\left(n - \frac{p+g}{2} - 1\right) ln\Lambda \tag{3.20}$$

que possui a distribuição Qui-quadrado com p(g-1) graus de liberdade. Rejeita-se a hipótese nula se p-valor for menor ou igual ao nível selecionado.

A estatística de teste de Λ de Wilks também pode ser utilizada, para seleção das variáveis potencialmente discriminantes. Neste caso, ela pode ser transformada para uma estatística que tem aproximadamente uma distribuição F-Snedecor com p e $(N - p - 1)$ graus de liberdade.

Esta estatística somente é válida se as suposições da análise discriminante forem satisfeitas, e se nenhuma das variáveis pode ser combinação linear de quaisquer outras, ou seja, serem multicolineares.

3.3.3 Teste da Razão de Verossimilhança

A estatística de teste da razão de verossimilhança é utilizada para comparar a verossimilhança do modelo nulo ou modelo reduzido (modelo com a constante) com a verossimilhança do modelo completo (HOSMER e LEMESHOW, 2000), ambos modelos da regressão logística. O teste da razão de verossimilhança testa as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad vs \quad H_a : \exists_i : \beta_i \neq 0 \quad (i = 1,\ldots,p).$$

A estatística de teste da razão de verossimilhança segue assintoticamente uma

distribuição χ^2 com número de graus de liberdade igual ao número de restrições (p) sob a hipótese nula. O cálculo da estatística de teste é feito a partir das estatísticas de teste para o modelo nulo e para o modelo completo da seguinte forma (BUSE, 1982):

$$G^2 = \chi_0^2 - \chi_c^2 = -2LL_0 - (-2LL_c) = -2ln\left[\frac{L_0}{L_c}\right] \sim \chi_{(p)}^2. \tag{3.21}$$

Geralmente, calcula-se a probabilidade de se observar o valor de G^2 ou um valor mais extremo, $p - valor = P\left(G^2 \geq \chi_{(p)}^2\right)$. Se $p - valor \leq \alpha$ rejeita-se a H_0 concluindo-se que pelo menos uma das variáveis independentes do modelo consegue predizer o $Logit(\pi)$.

3.3.4 Teste de Wald

A estatística de teste de Wald é utilizada para testar se cada coeficiente estimado da regressão logística é igual a zero, condicionado pelos valores estimados dos outros coeficientes, ou seja, testa a hipótese nula H_0 de que o parâmetro β da regressão logística é igual a zero contra a hipótese alternativa de que é diferente de zero (BUSE, 1982):

$$H_0 : \beta_i = 0|\beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \quad vs \quad H_a : \beta_i \neq 0|\beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \quad (i = 1, \ldots, p).$$

A estatística de teste Wald é dada da seguinte forma:

$$z_i = \frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)} \tag{3.22}$$

onde $\hat{\beta}_i$ é o estimador de β_i e $S\hat{E}(\hat{\beta}_i)$ é o erro-padrão da estimativa de β_i. Sob a hipótese nula, tem-se que a estatística z_i apresenta aproximadamente uma distribuição t-student, e quando a amostra é grande se aproxima assintoticamente da distribuição normal padrão. Quando $p - valor \leq \alpha$ rejeita-se a hipótese nula H_0 de que o parâmetro β da regressão logística é igual a zero.

3.3.5 Teste de ajustamento do modelo de regressão logística

Para avaliar a qualidade do ajuste do modelo de regressão logística utiliza-se uma medida global de qualidade de ajuste denominada função desvio ou deviance. A função desvio pode ser utilizada para obter uma estatística global de qualidade de ajuste do modelo. A hipótese nula H_0 de que o modelo se ajusta aos dados pode ser testada pela seguinte estatística (HOSMER e LEMESHOW, 2000):

$$D = -2ln\left[\frac{L_C}{L_S}\right] \tag{3.23}$$

onde L_C é a verossimilhança do modelo ajustado, modelo completo ou não, e L_S é a verossimilhança do modelo saturado. Para amostra grande, sob a H_0, a estatística tem distribuição assintótica Qui-quadrado com $n - p - 1$ graus de liberdade, e rejeita-se a hipótese nula se $p - valor \leq \alpha$.

O Critério de Informação de Akaike (AIC) é utilizado adicionando uma penalidade na função log verossimilhança, a fim de selecionar o melhor modelo, com maior valor da função verossimilhança maximizada, portanto, um modelo mais complexo. O critério de informação de Akaike é dado por (REID, 2010):

$$AIC = -2logL(\hat{\theta}) + 2p \tag{3.24}$$

onde p é o número de parâmetros estimados. O modelo é selecionado se tiver menor valor de AIC. O termo $2p$ é uma penalidade para ajustamento de modelos com maior número de parâmetros.

3.4 Classificação de indivíduos por recurso à regressão logística e análise discriminante

A classificação de indivíduos com base na função discriminante estimada é feita de modo que a probabilidade de má classificação seja mínima. Neste método de classificação utiliza-se o valor do escore discriminante: se $\lambda(x) > 0$ o indivíduo é classificado para população 1, caso contrário, o indivíduo é classificado para população 0 (EFRON, 1975).

Por outro lado, se $\lambda(x) = 0$ o risco de o indivíduo ser classificado para a população 1 é igual ao risco de ser classificado para a população 0. Por conseguinte, o indivíduo não será classificado para população 1 nem para população 0.

A forma análoga a este método de classificação, utiliza o risco $\hat{\pi}_j$ de cada um dos indivíduos pertencer ao grupo 1 (sucesso) comparativamente ao grupo de referência 0 (insucesso). Neste caso, se o risco do sujeito ser classificado para a população 1 for superior a 0.5, o indivíduo é classificado no grupo 1, caso contrário, o indivíduo é classificado no grupo 0.

Em alguns casos pode-se alterar o valor de $\hat{\pi}_j$ a fim de ter melhor precisão de classificação dos sujeitos. Assim, reduzimos o valor 0.5 se pretendermos que o modelo seja mais preciso ou rigoroso na forma de classificar os indivíduos, ou acrescê-lo se pretendermos que o modelo seja menos preciso.

A classificação de indivíduos recorrendo à regressão logística multinomial é feita com base na avaliação do risco do indivíduo pertencer à certa classe. Assim, os indivíduos são classificados em mais de duas classes da variável dependente e calcula-se a probabilidade de um determinado indivíduo j pertencer a cada uma das c classes

da variável dependente. Desse modo, o indivíduo é classificado para a classe na qual a sua probabilidade de ocorrência for maior.

3.4.1 Taxa de erro de classificação

A taxa de erro é a probabilidade de má classificação de indivíduos sob suposições da análise discriminante (EFRON, 1975).

As três formas de estimação da taxa de erro são: as probabilidades quando os parâmetros são estimados, U-método de Jack-knifing (*Jacknife*), e o método de resubstituição (MARDIA et al., 1979).

No método das probabilidades quando os parâmetros são estimados, estimam-se as taxas de erro de classificação p_{ij} para duas populações da seguinte forma (EFRON, 1975):

$$p_{ij} = \pi_1 prob\left\{X \in R_0 | X \sim \pi_p\left(\mu_1, \Sigma\right)\right\} + \pi_0 prob\left\{X \in R_1 | X \sim \pi_p\left(\mu_0, \Sigma\right)\right\} \quad (3.25)$$

onde R_0 e R_1 é a partição do espaço p-dimensional E^p, na qual o indivíduo é classificado para a população 0 se x pertence à região 0 e para a população 1 se pertencer à região 1. Se os parâmetros das distribuições subjacentes são estimados a partir dos dados, então temos probabilidades estimadas p_{ij}.

Este método tende a ser otimista sobre a probabilidade de erro de classificação, ou seja, tende a subestimar a probabilidade de má classificação verdadeira quando n é pequeno, e o mesmo acontece no método de resubstituição. Além disso, um dos problemas na estimação da função discriminante é o fato de as mesmas observações serem usadas para definir a regra discriminante, e para avaliar a sua precisão (MARDIA et al., 1979).

O método que resolve esse problema é o *Jacknife*. Pois, esse método consiste em considerar as amostras, abandonando uma observação de cada vez, ou seja, ajusta-se o modelo sem a observação cujo grupo se quer prever e depois usar esse modelo para classificar o caso deixado fora (MARDIA et al., 1979). Esses autores são mais rigorosos ao afirmarem que no caso de duas populações com a mesma matriz de covariância, esta abordagem leva a estimativas mais confiáveis das probabilidades de má classificação do que qualquer um dos dois anteriores.

Capítulo 4

Comparação da eficiência da RL e AD

Na análise sobre a eficiência de cada modelo em estudo, utiliza-se a matriz de classificação ou matriz de confusão na qual se ilustra a percentagem de casos corretamente classificados. Essa percentagem permite verificar a capacidade preditiva, e comparar a eficiência dos modelos de interesse.

Para avaliar a qualidade da classificação feita pelo modelo individualmente, no geral compara-se a percentagem global de classificações corretas obtidas pelo modelo, com a percentagem proporcional de classificações corretas por acaso. A percentagem proporcional de classificações corretas por acaso é calculada a partir do número de indivíduos observados em cada uma das k classes da variável dependente (C_i) pela seguinte expressão:

$$CC_pA(\%) = 100 \times \sum_{i=1}^{k} \left(\frac{C_i}{N}\right)^2 \tag{4.1}$$

onde $CC_pA(\%)$ é a percentagem proporcional de classificações corretas por acaso.

Se a percentagem de casos corretamente classificados pelo modelo for superior em pelo menos 25% à percentagem de classificação proporcional por acaso, considera-se que o modelo tem boas propriedades classificatórias.

A eficiência classificatória do modelo também pode ser avaliada pela sensibilidade e especificidade. A sensibilidade é a percentagem de classificações corretas na classe sucesso (1) da variável dependente, ou seja, é a capacidade de um instrumento reconhecer os verdadeiros positivos em relação ao total de doentes. A especificidade é a percentagem de classificações corretas na classe insucesso (0) do modelo, ou seja, é o poder de distinguir os verdadeiros negativos em relação ao total de doentes.

Em estudos biomédicos, a percentagem de observações originalmente classificadas na classe insucesso (0), mas que o modelo classifica na classe sucesso (1) relativamente ao número de insucessos (0) original é comumente chamado por taxa de falsos positivos. Além disso, a percentagem de observações originalmente

classificadas na classe sucesso (1), mas que o modelo classifica na classe insucesso (0) relativamente ao número de sucessos (1) original é denominado por taxa de falsos negativos. As taxas de falsos positivos e de falsos negativos utilizam-se para avaliar a eficiência do modelo, e é a denominada taxa (probabilidade) de erro de classificação.

Os falsos positivos e falsos negativos indicam o número de falhas do modelo estimado, ao passo que o número de casos de sensibilidade e de especificidade, indica o acerto do modelo.

4.1 Eficiência assintótica relativa da RL comparada com a análise discriminante

Os valores numéricos das medidas da eficiência relativa e da eficiência assintótica relativa (Eff_∞) são obtidos a partir da integral A_i (EFRON, 1975, p. 895-897), utilizando a regra de Simpson. Esses valores são substituídos nas equações 3.19 de EFRON (1975). Desse modo, com base nas suposições da análise discriminante, e mediante o valor da eficiência assintótica relativa (Eff_∞) são comparados os modelos da regressão logística e da análise discriminante. Um dos elementos utilizados, que através do qual se pode obter a eficiência assintótica relativa dos modelos em estudo, é a distância de Mahalanobis (EFRON, 1975).

A distância de Mahalanobis Δ^2 é uma medida de distância baseada nas covariâncias ou correlações entre variáveis com as quais, distintos padrões podem ser identificados e analisados. É uma estatística útil para determinar a similaridade entre uma amostra desconhecida e uma conhecida. Desse modo, os dados a comparar deverão ter o mesmo número de variáveis, ou seja, o mesmo número de colunas, mas não necessariamente o mesmo número de elementos (o número de linhas pode ser diferente).

A distância de Mahalanobis é dada por $\Delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$, onde μ_1 e μ_0 são as médias do grupo 1 e 0 respectivamente, e Σ é a matriz das covariâncias.

O quadrado da distância de Mahalanobis é usado de modo a separar o mais possível os grupos. Assim, uma variável é adicionada na análise se a sua distância Δ^2 aumentar significativamente, caso contrário é removida. Essa distância é utilizada para ilustrar como varia a eficiência dos dois modelos de interesse em função da probabilidade de um indivíduo pertencer a certo grupo. A probabilidade de certo indivíduo ser classificado para o grupo de sucesso ou insucesso ($\hat{\pi}_1 \quad ou \quad \hat{\pi}_0$) é estimada como foi apresentado em (2.2).

Na tabela 4.1 seguinte é fácil observar que uma vez obtido o valor da distância de Mahalanobis, em função do valor estimado da probabilidade de certo indivíduo ser classificado para o grupo de sucesso ou insucesso, pode-se obter o valor da eficiência

assintótica relativa (EFRON, 1975):

Tabela 4.1: Eficiência da RL comparada com a análise discriminante

π_1 ou π_0	Δ	Eff_∞	Eff_1	q	A_0	A_1	A_2
0.5	2	0.899	0.899	1	0.450	0	0.266
0.6	2	0.892	0.906	1.024	0.458	-0.038	0.273
0.667	2	0.879	0.913	1.070	0.465	-0.067	0.287
0.75	2	0.855	0.915	1.177	0.488	-0.108	0.319
0.9	2	0.801	0.804	1.697	0.589	-0.253	0.487
0.95	2	0.801	0.706	2.233	0.674	-0.375	0.667
0.5	2.5	0.786	0.786	1	0.307	0	0.154
0.6	2.5	0.778	0.794	1.013	0.311	-0.025	0.158
0.667	2.5	0.762	0.806	1.038	0.319	-0.044	0.167
0.75	2.5	0.733	0.819	1.096	0.337	-0.074	0.188
0.9	2.5	0.660	0.750	1.379	0.423	-0.181	0.304
0.95	2.5	0.650	0.637	1.671	0.501	-0.282	0.441
0.5	3	0.641	0.641	1	0.197	0	0.084
0.6	3	0.633	0.649	1.008	0.200	-0.016	0.087
0.667	3	0.618	0.662	1.023	0.206	-0.027	0.092
0.75	3	0.589	0.682	1.057	0.219	-0.046	0.104
0.9	3	0.511	0.667	1.225	0.282	-0.117	0.175
0.95	3	0.492	0.588	1.400	0.344	-0.189	0.265
0.5	3.5	0.486	0.486	1	0.120	0	0.044
0.6	3.5	0.479	0.493	1.005	0.122	-0.009	0.045
0.667	3.5	0.467	0.505	1.014	0.125	-0.016	0.048
0.75	3.5	0.442	0.526	1.035	0.134	-0.027	0.055
0.9	3.5	0.370	0.550	1.142	0.176	-0.07	0.095
0.95	3.5	0.348	0.516	1.252	0.220	-0.116	0.147
0.5	4	0.343	0.343	1	0.069	0	0.022
0.6	4	0.338	0.348	1.003	0.070	-0.005	0.022
0.667	4	0.328	0.358	1.009	0.072	-0.009	0.024
0.75	4	0.309	0.375	1.024	0.077	-0.014	0.027
0.9	4	0.252	0.416	1.094	0.103	-0.039	0.048
0.95	4	0.230	0.416	1.168	0.131	-0.065	0.076

FONTE: EFRON, 1975.

Para a obtenção dos valores da eficiência assintótica relativa da regressão logística comparada com a análise discriminante, utilizamos a expressão $\frac{1}{Eff_\infty}$, onde Eff_∞ é obtido na tabela 4.1. As expressões de Eff_1, q, e de $A_i (i = 0, 1, 2)$ encontram-se em 3.11, 3.21 e 3.22 de EFRON (1975). Observando os valores da tabela, é notório

que a análise discriminante é mais eficiente em relação a regressão logística, pois, todos os valores numéricos da eficiência são inferiores a 1. Para outros membros da família exponencial a regressão logística será mais eficiente.

A família de distribuições com função de probabilidade (densidade) $p(x|\theta)$ pertence a família exponencial com r parâmetros se $p(x|\theta)$ pode ser escrito da seguinte forma:

$$p(x|\theta) = a(x)exp\left\{\sum_{j=1}^{r} U_j(x)\phi_j(\theta) + b(\theta)\right\} \tag{4.2}$$

Quando um único X é observado, pelo critério de fatoração, $U_1(X), \ldots, U_r(X)$ são estatísticas suficientes para θ. Para um tamanho n de amostra de $\mathbf{X}$ pode ser escrito da seguinte forma:

$$p(\mathbf{x}|\theta) = \left[\prod_{i=1}^{n} a(x_i)\right] exp\left\{\sum_{j=1}^{r}\left[\sum_{i=1}^{n} U_j(x_i)\right]\phi_j(\theta) + nb(\theta)\right\} \tag{4.3}$$

onde $a(\mathbf{x}) = \prod_{i=1}^{n} a(x_i)$ e $U_j(\mathbf{X}) = \sum_{i=1}^{n} U_j(X_i)$, $j = 1, \ldots, r$. Desse modo, $\mathbf{T} = (T_1, \ldots, T_r)$ com $T_j = U_j(\mathbf{X})$, $j = 1, 2, \ldots, r$ é uma estatística suficiente para θ. A família exponencial inclui várias distribuições das quais mencionamos as distribuições: normal, bernoulli, poisson, binomial e gama.

EFRON (1975) ilustra que a função de verossimilhança da regressão logística pode ser escrita na forma da família exponencial seguinte:

$$f_{\beta_0,\beta}\left(y_1, \ldots, y_n|x_1, \ldots, x_n\right) = exp\left[(\beta_0, \beta')\mathbf{T} - \psi(\beta_0, \beta)\right] \tag{4.4}$$

onde $\mathbf{T} \equiv \sum_{j=1}^{n} \begin{pmatrix} 1 \\ x_j \end{pmatrix} y_j$ e $\psi(\beta_0, \beta) = \sum_{j=1}^{n} log(1 + exp(\beta_0 + \beta' x_j))$. $\mathbf{T}$ é a estatística suficiente que tem vetor média e matriz de covariâncias. É evidente que a regressão logística estimada provém de uma amostra da população com distribuição normal, pois a distribuição normal é membro da família exponencial com um parâmetro dimensional. Portanto, a análise discriminante é mais eficiente que a regressão logística, somente se os dados têm distribuição normal.

Capítulo 5

Aplicação

Nesta seção são aplicados os métodos descritos no trabalho, a fim de comparar os modelos da regressão logística (RL) ou regressão logística multinomial (RLM) e análise discriminante (AD) com dois grupos ou mais de dois grupos respectivamente. Para isso, foram utilizados dois bancos de dados.

O primeiro banco de dados é referente à 40 pacientes chagásicos e 19 indivíduos controle, os quais tinham ecocardiograma normal, radiografia de tórax, desempenho do exercício normal para a idade, gênero, índice de massa corporal, e outras variáveis descritas na tabela 5.1. O estudo foi realizado na clínica de Chagas doença ambulatorial do Hospital Universitário Clementino Fraga Filho da Universidade Federal do Rio de Janeiro(UFRJ), e a seleção dos pacientes foi feita entre Outubro de 2003 e Agosto de 2007.

O segundo e último banco de dados é referente à 82 pacientes renais crônicos (casos de sobrevida), submetidos a tratamentos hemodialíticos, provenientes de quatro centros de estado do Rio de Janeiro (três do município de Niteroi e um do município de Rio Bonito) (ALVES, 2012). Os pacientes foram recrutados no período de 15 de Julho de 1997 à 15 de Julho de 1998, e a maioria era atendida pelo sistema único de saúde.

As variáveis envolvidas como a Idade, PASpre descritas na tabela 5.2, são utilizadas no presente estudo para analisar as capacidades preditivas dos modelos de regressão logística e multinomial comparados com a análise discriminante com duas e quatro populações respectivamente. Os dois bancos de dados utilizados neste trabalho, foram formados em Excel. Posteriormente, os dados foram processados através do software R, e somente a estatística do teste M-Box foi processada em SPSS, pois este teste não está incluso em R. A outra estatística processada através deste software é a do teste Lambda de Wilks.

Para comparar os modelos de regressão logística e análise discriminante com dois grupos, foram utilizados os dois bancos de dados, tendo sido considerados como desfechos as variáveis classebin e censura. Ao passo que, a comparação entre a

regressão logística multinomial e análise discriminante com quatro grupos foi feita mediante o banco de dados de pacientes renais crônicos (sobrevida), sendo que o desfecho foi a variável CID.

No presente trabalho, primeiro comparamos modelos com dois grupos (classes) cujas variáveis explicativas satisfazem as suposições da análise discriminante. Após isso, comparamos também, os modelos com dois grupos cujas variáveis violam tais suposições, a fim de analisar a variação da eficiência com a inclusão das variáveis que satisfazem ou que violam as suposições da análise discriminante. Do mesmo modo, comparamos os modelos de regressão logística multinomial e análise discriminante com quatro grupos.

Para comparar a eficiência relativa desses modelos, utilizou-se a matriz de classificação, comumente denominada de matriz de confusão na qual se identificou a percentagem de casos corretamente classificados, a sensibilidade, a especificidade, e a probabilidade de erro de classificação de cada modelo.

O nosso interesse é de comparar a eficiência assintótica relativa da regressão logística e análise discriminante com dois grupos. Para isso, calculamos as estimativas das distâncias de Mahalanobis e do risco de certo indivíduo ser classificado para o grupo de chagas (1) ou morte (1). Com base nessas estimativas e mediante a tabela 4.1, foram obtidos os valores da eficiência assintótica relativa desses modelos.

As estatísticas de testes Shapiro-Wilk, M-Box, Lambda de Wilks, teste da razão de verossimilhança, Wald, Qui-quadrado, função desvio ou deviance, F-Snedecor, foram utilizadas para verificar os pressupostos, a significância e o ajuste dos modelos aos dados, com nível de significância (α) selecionado de 5%. O critério de informação usado para a seleção dos modelos foi o AIC. As variáveis utilizadas neste estudo são apresentadas nas tabelas 5.1 e 5.2 seguintes.

5.1 Descrição das variáveis

Tabela 5.1: Variáveis de pacientes chagásicos.

Variável	Descrição da variável	Categorias
classebin	Desfecho	0-Controle; 1-chagas
Sex	Gênero	0-Feminino; 1-Masculino
hm20	Cintilografia com MIBG-IODO-123, com imagens planares de 20 minutos.	
hm3	Cintilografia com MIBG-IODO-123, com imagens planares de 3 horas.	
Washout	Washout	
Spect	Single photon emission computed tomography.	0-normal; 1-alterado
ECG	Eletrocardiograma	0-normal; 1-alterado
age	Idade	

FONTE: LANDESMANN et al., 2011

Tabela 5.2: Variáveis de pacientes renais crônicos (sobrevida).

Variável	Descrição da variável	Categorias
Censura	Desfecho	0-sobrevida; 1-morte
CID	Classificação Internacional de doença (desfecho).	0-pacientes vivos; 1-doença aterotrombótica vascular; 2-doença infecciosa; 3-outras causas de óbito.
Idade	Idade	
Gênero	Gênero	0-Feminino; 1-Masculino
AVE	Acidente Vascular Encefálico	0-Não; 1-Sim
DIC	Doença Isquêmica Cardíaca	0-Não; 1-Sim
Diabetes	Diabetes	0-Não; 1-Sim
Tabagismo	Tabagismo	0-Não; 1-Sim
ECA	Enzima Conversora da Angiotensina I (Genótipo)	1-DD; 2-II; 3-DI
ANGO	Genótipo	1-MM; 2MT; 3-TT.
RendaMen	Renda Mensal	1-Menor que dois salários mínimos; 2-De 2 à 5 SM (inclusive); 3-Maior que 5 SM.
IMC	Índice de massa corporal	
Trig	Triglicerídeos	
ColTotal	Colesterol total	
PASpre	Pressão Arterial sistólica	
HdlC	HdlC	
Hematocrito	Hematocrito	

FONTE: ALVES, 2012.

5.2 Teste da normalidade multivariada das variáveis explicativas

Para comparar modelos da análise multivariada de dados é imprescindível que esses modelos tenham as mesmas condições, ou seja, tenham mesmas covariáveis, sejam modelos com mesmo número de grupos na variável dependente e se use o mesmo software para sua estimação. Assim, dado que os modelos de regressão logística não têm restrições em termos do tipo de covariáveis a utilizar, envolvemos no estudo covariáveis de cada banco de dados apropriadas para a análise discriminante a fim de estimar e comparar os modelos de interesse.

Relativamente ao pressuposto da distribuição normal multivariada, não existe nenhum teste para verificar este pressuposto, mas, de uma forma geral aceita-se que se cada uma das variáveis possui distribuição normal, a distribuição das variáveis é multivariada apesar de nem sempre esta extensão da normalidade univariada para a multivariada ser válida.

Neste trabalho, utilizamos os histogramas a fim de ter uma ideia sobre a forma da distribuição de cada uma das variáveis explicativas, ou seja, se a curva é simétrica ou assimétrica. Após isso, utilizamos a estatística do teste Shapiro-Wilk para a obtenção das conclusões referentes ao mesmo pressuposto.

As figuras abaixo representam histogramas das variáveis explicativas dos casos chagásicos e de sobrevida (pacientes renais crônicos), com uma curva simétrica, das quais se pode observar que as três primeiras figuras são dos casos chagásicos. Ao passo que as duas últimas são dos casos de sobrevida:

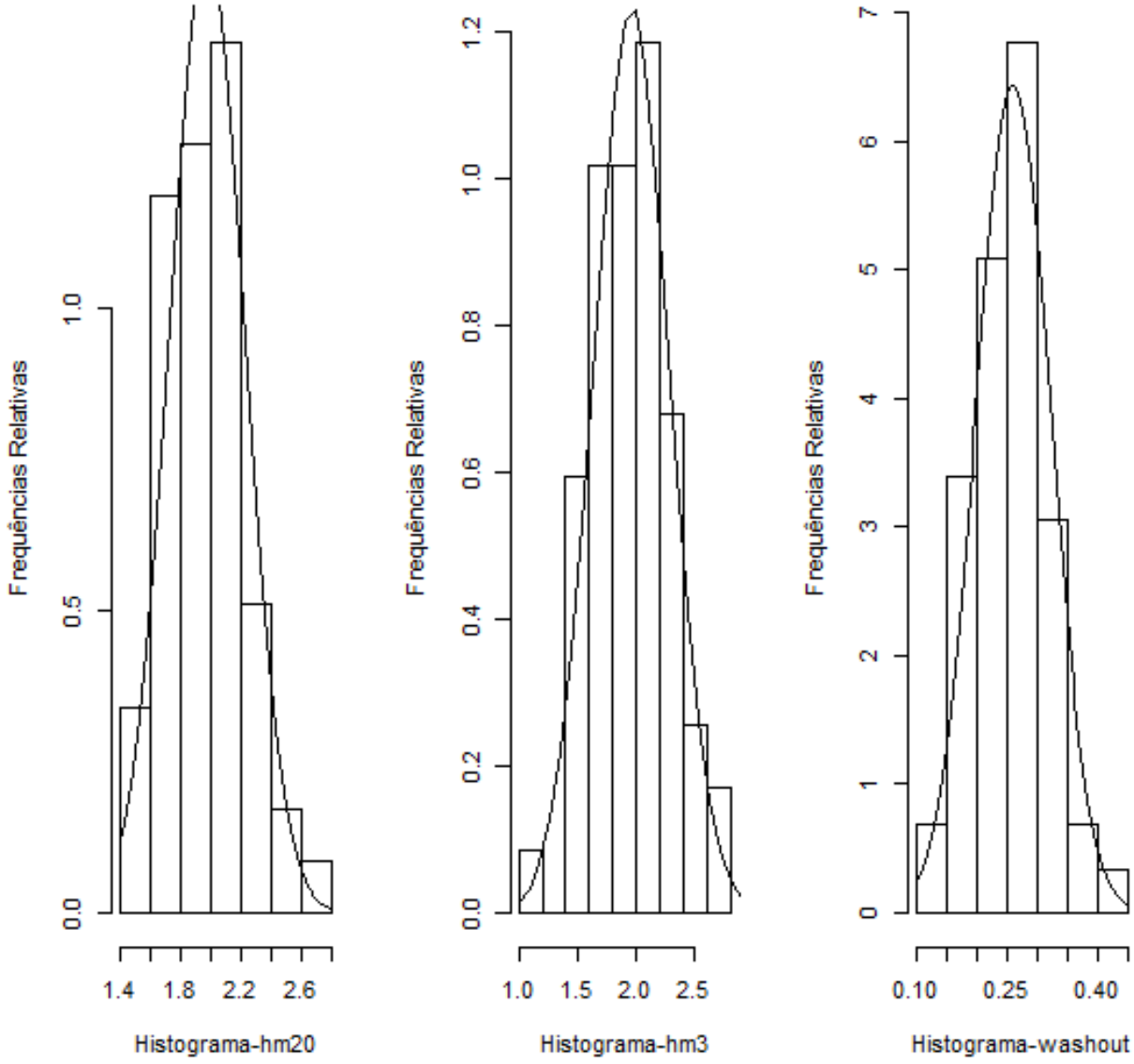

Figura 5.1: Histogramas das variáveis dos casos de chagas

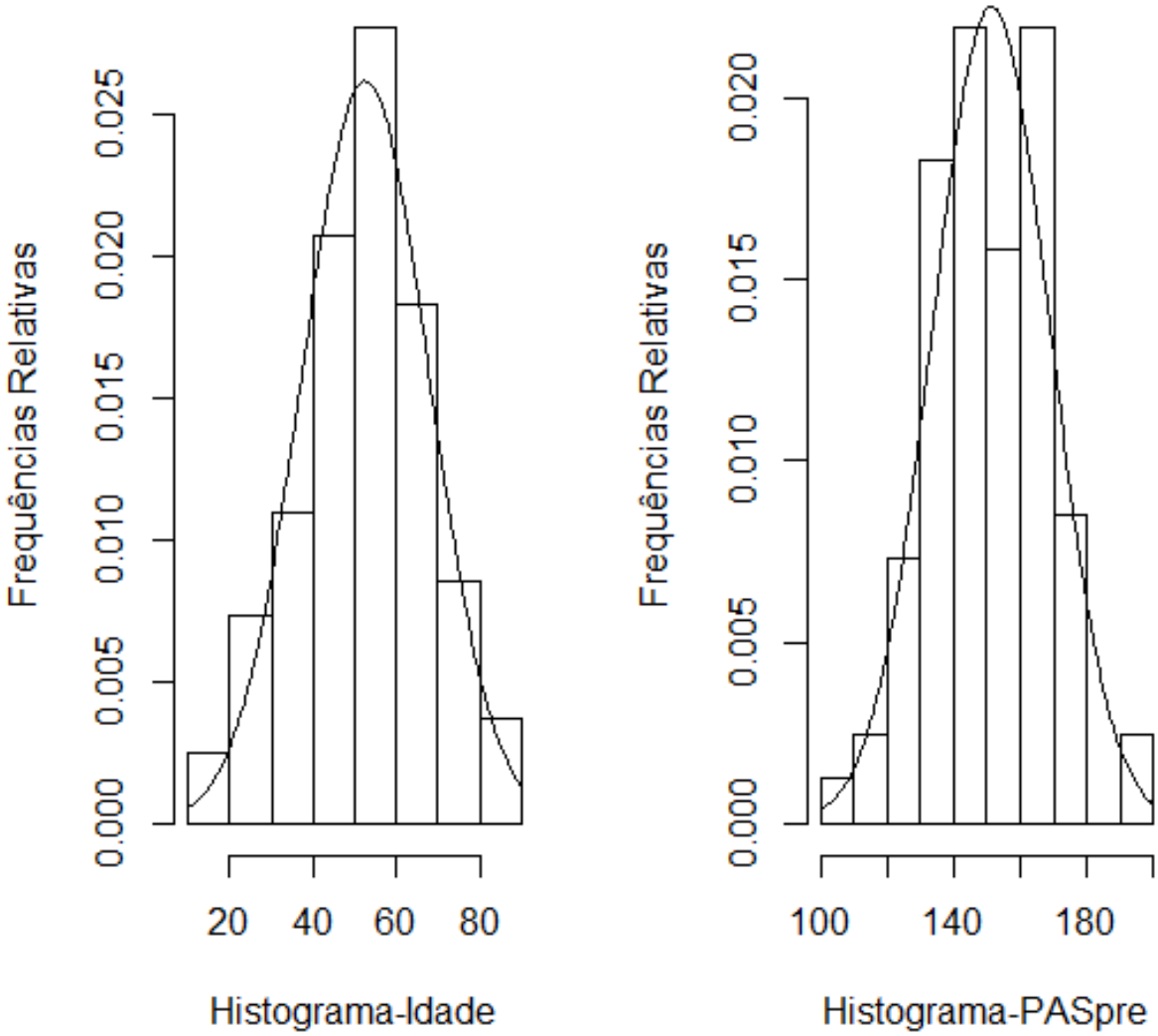

Figura 5.2: Histogramas das variáveis dos pacientes renais crônicos

Uma vez observado que os histogramas tendem a ter uma curva simétrica, ou seja, tendem a ter uma distribuição normal, foi utilizada a estatística do teste Shapiro-Wilk para verificar se a normalidade é satisfeita em cada variável de estudo:

Tabela 5.3: Teste de Shapiro-Wilk para as variáveis explicativas de chagas e pacientes renais crônicos (sobrevida).

Variável	Estatística	p-valor
age	0.9544	0.02702
hm20	0.9777	**0.3507**
hm3	0.9923	**0.9716**
Washout	0.9877	**0.816**
Variável	Estatística	p-valor
Idade	0.9869	**0.5733**
IMC	0.865	4.115e-07
Trig	0.9074	2.045e-05
ColTotal	0.9683	0.03998
HdlC	0.8679	5.249e-07
PASpre	0.9929	**0.9367**
Hematocrito	0.9675	0.03536

De acordo com os resultados obtidos (tabela 5.3), pode-se constatar que a um nível de significância de 5%, a variável age não satisfaz a hipótese de que a amostra dos casos de chagas provém de uma população com distribuição normal. Porém, todas as restantes variáveis contínuas satisfazem essa suposição. No caso das variáveis de sobrevida, apenas as variáveis Idade e PASpre satisfazem a suposição de que a amostra provém de uma população normal, a um nível de significância de 5%.

5.3 Teste de igualdade das matrizes de variâncias e covariâncias

Nas tabelas seguintes são apresentadas as estatísticas do teste M-Box para verificar se é satisfeita a suposição da igualdade das matrizes de variâncias e covariâncias:

Tabela 5.4: Casos de chagas com todas variáveis

M-Box	F	df1	df2	p
8.953	0.812	10	6047.302	0.617

Tabela 5.5: Casos de chagas cujas variáveis têm distribuição normal

M-Box	F	df1	df2	p
6.842	1.061	6	8232.645	0.384

Tabela 5.6: Casos de sobrevida com todas variáveis

M-Box	F	df1	df2	p
36.600	1.184	28	21304.248	0.230

Tabela 5.7: Casos de sobrevida cujas variáveis têm distribuição normal

M-Box	F	df1	df2	p
2.439	0.791	3	3995308	0.499

Tabela 5.8: Casos de sobrevida com quatro grupos e todas variáveis

M-Box	F	df1	df2	p
107.729	3.408	28	14730.124	0.000

Tabela 5.9: Casos de sobrevida com quatro grupos cujas variáveis têm distribuição normal

M-Box	F	df1	df2	p
13.612	1.368	9	2561.887	0.197

Com base nos resultados das estatísticas do teste M-Box, constatamos que a um nível de significância de 5% a homogeneidade das matrizes de variâncias e covariâncias é satisfeita nos casos de chagas e de sobrevida com duas populações, quando as variáveis explicativas seguem ou não a distribuião normal. Porém, na presença de quatro grupos (com desfecho CID), quando são envolvidas todas as variáveis explicativas, a homogeneidade das matrizes das variâncias e covariâncias não é satisfeita.

Contudo, todas as variáveis explicativas envolvidas no presente estudo que satisfazem a hipótese de que a amostra provém de uma população com distribuição normal, satisfazem as suposições da análise discriminante.

5.4 Modelos de RL e ADL de casos de chagas

Nesta subseção apresentamos os coeficientes estimados da regressão logística e análise discriminante com dois grupos dos casos de chagas, cujas variáveis satisfazem as suposições da análise discriminante. Além disso, comparamos os resultados sobre a eficiência desses modelos.

Tabela 5.10: Modelo de regressão logística

Variáveis	coef	Std. Error	z-value	Pr(>|z|)	exp(coef)
(Intercept)	9.6180	3.7548	2.561	0.01042	1.503e+04
hm20	2.0335	2.3605	0.861	0.38898	7.640975
hm3	-6.2818	2.3151	-2.713	0.00666	1.870079e-03
Washout	-0.8206	5.7712	-0.142	0.88693	0.4401637

Tabela 5.11: Estatística do teste da razão de verossimilhança

Chi-square	df	p
55.27121	3	6.009748e-12

Tabela 5.12: Teste de ajustamento do modelo de RL

Chi-square	df	p	AIC
55.27121	55	0.4643781	63.271

A tabela 5.10 ilustra os coeficientes originais, erro padrão, as estatísticas Z (estatísticas de teste Wald), o p-valor associado, e os coeficientes exponenciados para o modelo de regressão logística. A estatística do teste da razão de verossimilhança é apresentada na tabela 5.11, e a medida do ajuste do modelo é ilustrada na tabela 5.12.

Os resultados obtidos apontam que de forma conjunta, os coeficientes são estatisticamente significativos (Tabela 5.11), e o modelo selecionado pelo Critério de Informação de Akaike (AIC=63.271) se ajusta melhor do que o modelo nulo (Tabela 5.12). Porém, o teste da significância estatística individual dos coeficientes (na tabela 5.10) sugere que apenas o coeficiente logístico da variável hm3 e o intercepto são estatisticamente significativos ($p - valor = 0.00666 < \alpha = 0.05$ e $p - valor = 0.01042 < \alpha = 0.05$ respectivamente).

Tabela 5.13: Modelo discriminante linear

Variáveis	coef	df	Sum Sq	Mean Sq	F	Pr(>|z|)
hm20	1.126933	1	1.722	1.7223	10.046	0.00249
hm3	-4.421945	1	1.692	1.6925	9.872	0.00270
Washout	-2.252922	1	0.037	0.0368	0.215	0.64484
Resíduos	...	55	9.430	0.1714	...	...

Tabela 5.14: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.732	17.311	3	0.001

A tabela 5.13 resume a informação sobre os coeficientes da função discriminante,

estatísticas do teste de significância dos coeficientes associadas ao p-valor, e a estatística do teste de significância da função discriminante (tabela 5.14).

Vale lembrar, que a suposição da homogeneidade das matrizes de variâncias e covariâncias, das variáveis explicativas envolvidas para a estimação deste modelo é satisfeita (tabela 5.5), e a estatística do teste de Lambda de Wilks (na tabela 5.14) sugere que a função discriminante estimada é estatisticamente significativa ($p - valor = 0.001 < 0.05$). Porém, a variável Washout não é estatisticamente significativa ($p - valor = 0.64484 > \alpha = 0.05$, na tabela 5.13). Contudo, de todas as variáveis métricas envolvidas no modelo discriminante, somente a variável Washout estatisticamente não contribui na classificação de indivíduos para o grupo.

Importa destacar que, a significância estatística é uma forma de avaliar a confiabilidade dos resultados estatísticos, e não significa que o efeito é importante ou que tem qualquer utilidade na tomada de decisão. Por isso, continuamos fazendo a análise sobre a eficiência das classificações dos modelos de interesse.

Tabela 5.15: Matriz de classificação de análise discriminante e regressão logística.

	Classificação da análise discriminante			Classificação da regressão logística		
	Previsto		Percentagem	Previsto		Percentagem
Observado	0	1	Correta (%)	0	1	Correta (%)
0	19	0	100	10	9	52.6
1	21	19	47.5	4	36	90
% Total			64.4			77.9

A regressão logística classificou corretamente 46 dos 59 indivíduos (chagásicos e controle), correspondentes a uma percentagem de 77.9% de classificações corretas, ao passo que, a análise discriminante classificou corretamente 38 dos 59 indivíduos, o que corresponde a uma percentagem de 64.4% de classificações corretas. Estes resultados revelam que existe uma diferença de capacidades preditivas dos dois modelos (tabela 5.15).

As taxas de erro de classificação da regressão logística e análise discriminante ($1 - 0.779 = 0.221$ e $1 - 0.644 = 0.356$ respectivamente), comprovam a existência de diferenças em termos do quão bem a pertinência à grupo é prevista pelos modelos. Pois, constatamos que 22.1% dos indivíduos foram incorretamente classificados em ambos os modelos, mas, a análise discriminante classificou incorretamente, mais 13.5% dos indivíduos que o modelo de regressão logística classificou corretamente.

Além disso, os resultados dos casos que foram mal classificados por cada procedimento têm alguma sobreposição (tabela 5.15). Quatro (4) indivíduos foram mal classificados em ambos os procedimentos (regressão logística e análise discriminante). Portanto, todos os 4 foram pacientes de chagas(1), mas foram

classificados como indivíduos controle(0). A análise discriminante classificou incorretamente, mais 17 indivíduos chagásicos(1) como indivíduos do grupo controle(0), que a regressão logística classificou corretamente. Ao passo que, a regressão logística classificou incorretamente 9 indivíduos de controle(0) como chagásicos(1), que a análise discriminante classificou corretamente.

Essas diferenças refletem-se nos valores da sensibilidade e especificidade de cada modelo. A regressão logística tem sensibilidade de 52.6% e especificidade de 90%, ao passo que a análise discriminante tem sensibilidade de 100% e especificidade de 47.5%. Portanto, o modelo de regressão logística classificou corretamente muitos casos chagásicos, enquanto que a análise discriminante classificou corretamente todos os indivíduos controle.

Em termos de capacidade preditiva ou propriedades classificatórias, constatamos que o procedimento de regressão logística tem boas propriedades classificatórias em relação à análise discriminante, ou seja, a regressão logística foi relativamente mais eficiente que a análise discriminante para estes resultados (de chagas e controle).

5.5 Modelos de RL e ADL de casos de sobrevida

A seguir são apresentados os resultados dos modelos em estudo, para casos de sobrevida, cujas variáveis satisfazem as suposições da análise discriminante. Após a obtenção dos modelos, foi também feita a comparação da eficiência dos mesmos.

Tabela 5.16: Modelo de regressão logística

Variáveis	coef	Std. Error	z-value	Pr(>|z|)	exp(coef)
(Intercept)	-1.44217	2.16781	-0.665	0.506	0.236414
Idade	0.01842	0.01502	1.226	0.220	1.018589
PASpre	0.00411	0.01281	0.321	0.748	1.004118

Tabela 5.17: Estatística do teste da razão de verossimilhança

Chi-square	df	p
111.6462	2	0.0000

Tabela 5.18: Teste de ajustamento do modelo de RL

Chi-square	df	p	AIC
111.6462	79	0.009181891	117.65

A tabela 5.16 apresenta os coeficientes originais, erro padrão, as estatísticas Z (estatísticas de teste Wald), o p-valor associado, os coeficientes exponenciados para modelo de regressão logística. A estatística de teste da razão de verossimilhança é apresentada na tabela 5.17, e a tabela 5.18 ilustra a medida de ajuste do modelo.

Com estes resultados, constatamos que de forma conjunta, os coeficientes são estatisticamente significativos, pois $p - valor = 0 < 0.05$ (Tabela 5.17). Por outro lado, o modelo selecionado pelo Critério de Informação de Akaike (AIC= 117.65) não se ajusta bem aos dados ($p - valor = 0.009181891 < 0.05$, na tabela 5.18). As estatísticas de teste de significância individual dos coeficientes (Tabela 5.16) sugerem que nenhum dos coeficientes da regressão logística é estatisticamente significativo, pois os p-valores são maiores que o nível de significância selecionado ($\alpha = 0.05$).

Tabela 5.19: Modelo discriminante linear

Variáveis	coef	df	Sum Sq	Mean Sq	F	Pr(>|z|)
Idade	0.06501019	1	0.367	0.3674	1.451	0.232
PASpre	0.01446778	1	0.025	0.0254	0.100	0.752
Resíduos	...	79	19.997	0.2531	...	

Tabela 5.20: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.981	1.537	2	0.464

As tabelas 5.19 e 5.20 apresentam os coeficientes do modelo, as estatísticas do teste de significância dos coeficientes da análise discriminante associadas aos p-valores, e a estatística do teste de significância da função discriminante.

Com base no resultado da estatística do teste M-Box apresentado na tabela 5.7, constatamos que há homogeneidade das matrizes de variâncias e covariâncias ($p - valor = 0.499 > \alpha = 0.05$). Uma vez que essas observações provêm de uma população com distribuição normal multivariada, conclui-se que elas não violam as suposições da análise discriminante.

Constatamos também que a estatística de teste de Lambda de Wilks (tabela 5.20) revela que a função discriminante estimada não é estatisticamente significativa ($p - valor = 0.464 > \alpha = 0.05$), e nenhuma das variáveis é estatisticamente significativa (tabela 5.19).

O nosso objetivo é comparar a eficiência dos modelos de interesse. Para isso, apresentamos na tabela 5.21 os seguintes resultados:

Tabela 5.21: Matriz de classificação de análise discriminante e regressão logística.

	Classificação da análise discriminante			Classificação da regressão logística		
	Previsto		Percentagem	Previsto		Percentagem
Observado	0	1	Correta (%)	0	1	Correta (%)
0	15	23	39.47	12	26	31.58
1	29	15	34.09	9	35	79.55
% Total			36.59			57.32

Conforme os resultados da tabela 5.21, a regressão logística classificou corretamente 47 dos 82 indivíduos (sobrevida e morte), correspondentes a 57.32% de classificações corretas. Ao passo que, a análise discriminante classificou corretamente 30 dos 82 indivíduos, o que corresponde a 36.59% de classificações corretas. Estes resultados, também revelam que existe uma diferença de capacidades preditivas dos dois modelos.

Com base nas taxas de erro de classificação da regressão logística e análise discriminante ($1 - 0.5732 = 0.4268$ e $1 - 0.3659 = 0.6341$ respectivamente), constatamos que há diferenças em termos da eficiência relativa, pois, observamos que 42.68% dos indivíduos foram incorretamente classificados em ambos os modelos. Porém, a análise discriminante classificou incorretamente mais 20.73% dos indivíduos que o modelo da regressão logística classificou corretamente.

Na mesma tabela (tabela 5.21), observa-se também que os resultados dos casos que foram mal classificados por cada procedimento têm alguma sobreposição. Nove(9) indivíduos foram mal classificados em ambos os modelos. Os 9 são mortos(1), mas foram classificados como sobreviventes(0). A análise discriminante classificou incorretamente mais 20 indivíduos mortos(1) como sobreviventes(0), que a regressão logística classificou corretamente.

Ambos os modelos também classificaram incorretamente 23 indivíduos sobreviventes(0) como mortos(1), mas, a regressão logística classificou incorretamente mais 3 indivíduos sobreviventes(0) como mortos(1), que a análise discriminante classificou corretamente. Por conseguinte, a sensibilidade da regressão logística é de 31.58% e especificidade de 79.55%, ao passo que a análise discriminante tem sensibilidade de 39.47% e especificidade de 34.09%.

Com estes resultados, observamos também que o procedimento da regressão logística tem boas propriedades classificatórias em relação à análise discriminante, ou seja, a regressão logística é relativamente mais eficiente que a análise discriminante.

Na tabela seguinte, é apresentada a distância de Mahalanobis estimada dos casos chagásicos, a fim de comparar a eficiência assintótica relativa da regressão logística e análise discriminante:

Tabela 5.22: Distância de Mahalanobis para os casos de chagas

Distância de Mahalanobis	
Δ^2	2.949153

Observamos então, que a $\Delta = 1.71731 \approx 2$ (tabela 5.22). Dado que a $\pi_1 = \frac{n_1}{n} = \frac{40}{59} = 0.678$, de acordo com o valor da eficiência assintótica apresentado por EFRON (1975) na tabela 4.1 (0.879), a eficiência assintótica relativa da regressão logística (1/0.879) está entre 1/8 e 1/7 em relação ao procedimento da discriminação normal.

Este resultado revela que o procedimento da regressão logística é assintoticamente menos eficiente em relação à análise discriminante quando as observações satisfazem as suposições da análise discriminante.

Tabela 5.23: Distância de Mahalanobis para os casos de sobrevida

Distância de Mahalanobis	
Δ^2	1.97561

Com base no resultado da tabela 5.23, como a $\Delta = 1.40556$ e $\pi_1 = \frac{n_1}{n} = \frac{44}{82} = 0.54$, da tabela 4.1 observamos que $Eff_\infty = 0.899$, pois essa tabela não apresenta valores da distância de mahalanobis abaixo de 2. Por conseguinte, constatamos que a regressão logística é $0.11(\frac{1}{0.899})$ assintoticamente menos eficiente em relação a análise discriminante. Para este resultado, a eficiência da regressão logística está entre 1/9 e 1/8 quando comparada com o procedimento da discriminante normal.

5.6 Modelos de RL e ADL dos casos de chagas cujas variáveis violam as suposições

Nesta e na seguinte subseção comparamos os resultados dos modelos obtidos através das variáveis explicativas que violam as suposições da análise discriminante, a fim de verificar como a eficiência desses modelos varia quando comparada com a dos modelos cujas suas variáveis explicativas satisfazem essas suposições.

Tabela 5.24: Modelo de regressão logística

Variável	coef	Std. Error	z-value	Pr(>|z|)	exp(coef)
(Intercept)	-2.72035	1.11744	-2.434	0.0149	0.0658515
age	0.06997	0.02264	3.091	0.0020	1.0724726

Tabela 5.25: Estatística do teste da razão de verossimilhança

Chi-square	df	p
62.43921	1	2.775558e-15

Tabela 5.26: Teste de ajustamento do modelo de RL

Chi-square	df	p	AIC
62.43921	57	0.2890776	66.439

De acordo com os resultados apresentados nas tabelas 5.24 e 5.25, o modelo obtido é estatisticamente significativo. Além disso, este modelo se ajusta melhor aos dados (Tabela 5.26). Pode ser observado também, que o modelo discriminante estimado (Tabela 5.27) é estatisticamente significativo, pois o $p - valor = 0.001 < 0.05$ (Tabela 5.28).

Tabela 5.27: Modelo discriminante linear

Variável	coef	df	Sum Sq	Mean Sq	F	Pr(>|z|)
age	0.07294405	1	2.427	2.4273	13.23	0.000593
Resíduos	...	57	10.454	0.1834	...	...

Tabela 5.28: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.812	11.797	1	0.001

Tabela 5.29: Matriz de classificação de análise discriminante e regressão logística.

	Classificação da análise discriminante			Classificação da regressão logística		
	Previsto		Percentagem	Previsto		Percentagem
Observado	0	1	Correta (%)	0	1	Correta (%)
0	19	0	100	8	11	42.11
1	21	19	47.5	5	35	87.5
% Total			64.4			72.88

Na tabela 5.29 observamos que o modelo de regressão logística apresenta 72.88% das classificações corretas. Este resultado supera o valor das classificações corretas obtidas pela análise discriminante, em 8.48%.

Mediante os valores da sensibilidade e especificidade, claras diferenças podem ser observadas. Pois, a regressão logística tem sensibilidade igual a 42.11% e

especificidade de 87.5%. Enquanto que a análise discriminante tem 100% de sensibilidade e 47.5% de especificidade.

Portanto, o modelo de regressão logística classificou incorretamente 57.89% dos indivíduos controle(0) que a análise discriminante classificou corretamente, ao passo que 12.5% dos indivíduos chagásicos(1) são incorretamente classificados em ambos modelos, mas, a análise discriminante classificou incorretamente, mais 40% dos indivíduos chagásicos(1) que o modelo de regressão logística classificou corretamente.

Estes resultados, também revelam que o modelo de regressão logística foi mais eficiente em relação à análise discriminante, embora as suposições sejam violadas.

5.7 Modelos de RL e ADL dos casos de sobrevida cujas variáveis violam as suposições

Tabela 5.30: Modelo de regressão logística

Variáveis	coef	Std. Error	z-value	Pr(>|z|)	exp(coef)
(Intercept)	1.524	1.775	0.858	0.391	4.5894159
IMC	-1.674e-02	4.423e-02	-0.379	0.705	0.9833947
Trig	-1.279e-06	2.898e-03	0.000	1.000	0.9999987
ColTotal	-2.684e-03	6.034e-03	-0.445	0.656	0.9973198
HdlC	3.535e-03	1.807e-02	0.196	0.845	1.0035414
Hematocrito	-2.192e-02	4.867e-02	-0.450	0.652	0.9783147

Tabela 5.31: Estatística do teste da razão de verossimilhança

Chi-square	df	p
112.403	5	0.0000

Tabela 5.32: Teste de ajustamento do modelo de RL

Chi-square	df	p	AIC
112.403	76	0.00423311	124.4

Os coeficientes estimados do modelo de regressão logística não são estatisticamente significativos individualmente (tabela 5.30), mas, de forma conjunta os coeficientes são estatisticamente significativos (tabela 5.31), pois o p-valor=0.00<0.05. Ao observar a tabela 5.32, constatamos que este modelo, não se ajusta melhor aos dados ($p - valor = 0.00423311 < 0.05$).

Tabela 5.33: Modelo discriminante linear

Variáveis	coef	df	Sum Sq	Mean Sq	F	Pr(>|z|)
IMC	-8.198183e-02	1	0.069	0.06875	0.259	0.612
Trig	-3.194787e-05	1	0.010	0.01014	0.038	0.846
ColTotal	-1.307046e-02	1	0.070	0.06985	0.263	0.610
HdlC	1.717904e-02	1	0.008	0.00774	0.029	0.865
Hematocrito	-0.1068859	1	0.050	0.05009	0.189	0.665
Resíduos	...	76	20.184	0.26557	...	...

Tabela 5.34: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.99	0.789	5	0.978

O modelo estimado de análise discriminante apresentado na tabela 5.33, também não é estatisticamente significativo (Tabela 5.34). Não obstante, analisamos a seguir a eficiência relativa dos modelos.

Tabela 5.35: Matriz de classificação de análise discriminante e regressão logística.

	Classificação da análise discriminante			Classificação da regressão logística		
	Previsto		Percentagem	Previsto		Percentagem
Observado	0	1	Correta (%)	0	1	Correta (%)
0	15	23	39.47	9	29	23.68
1	29	15	34.09	10	34	77.27
% Total			36.59			52.44

Os resultados apontam que o modelo de regressão logística é mais eficiente quando comparado com o da análise discriminante. Pois, 52.44% dos casos são corretamente classificados por este modelo (Tabela 5.35). Este resultado é superior em relação ao do procedimento da análise discriminante, em 15.85% dos casos.

5.8 Modelos de RLM e AD com múltiplos grupos

Nesta subseção apresentamos os resultados obtidos, dos modelos da regressão logística multinomial e análise discriminante com quatro grupos dos casos de sobrevida com suposições satisfeitas, e com violação das mesmas. Além disso, fez-se a comparação desses modelos, em termos da sua eficiência relativa.

5.8.1 Modelos de RLM e AD com quatro grupos cujas suposições são satisfeitas

Tabela 5.36: Modelo de regressão logística multinomial

Variáveis	$Coef_1$	$Coef_2$	$Coef_3$	$exp(Coef_1)$	$exp(Coef_2)$	$exp(Coef_3)$
(Intercept)	-1.893793	-3.966236	-1.943035	0.1504999	0.0189446	0.1432684
Idade	-0.021696322	0.007888469	0.024032067	0.9785374	1.0079197	1.0243232
PASpre	0.008597136	0.012597899	0.003570943	1.008634	1.012678	1.003577

Tabela 5.37: Estatística do teste da razão de verossimilhança

Chi-square	df	p
51.44144	2	6.755152e-12

Tabela 5.38: Teste de ajustamento do modelo de RLM

Chi-square	df	p	AIC
51.44144	79	0.9930858	202.4561

A tabela 5.37 ilustra o teste da significância dos modelos de regressão logística multinomial apresentados na tabela 5.36. Estes modelos são estatisticamente significativos, e melhor se ajustam aos dados (Tabela 5.38). Ao passo que, a tabela 5.39 é referente aos modelos da análise discriminante estimados. Estes, não são estatisticamente significativos (Tabela 5.40).

Tabela 5.39: Modelo discriminante com quatro grupos

Variáveis	LDA1	LDA2	df	Sum Sq	Mean Sq	F	Pr(>|F|)
Idade	6.578554e-02	-0.00525142	1	4.21	4.213	2.194	0.143
PASpre	-2.475771e-05	-0.05574098	1	0.14	0.137	0.072	0.790
Resíduos	...	...	79	151.71	1.920		

Tabela 5.40: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.953	3.737	6	0.712
0.996	0.318	2	0.853

Tabela 5.41: Matriz de classificação de análise discriminante com quatro grupos e regressão logística multinomial.

Observado	Análise discriminante com quatro grupos					Regressão Logística Multinomial				
	Previsto					Previsto				
	0	1	2	3	% corr.	0	1	2	3	% corr.
0	1	3	11	21	2.78	26	0	0	10	72.2
1	1	1	2	3	14.29	5	0	0	2	0
2	0	0	4	3	57.14	5	0	0	2	0
3	5	3	15	9	28.13	18	0	0	14	43.75
% Total					18.29					48.78

No que tange a eficiência dos modelos, observamos que com a inclusão das variáveis explicativas que satisfazem as suposições da análise discriminante, o modelo da regressão logística multinomial é mais eficiente em relação à análise discriminante com quatro grupos (tabela 5.41). Pois, 48.78% dos indivíduos foram corretamente classificados pelo modelo, enquanto que, a análise discriminante classificou corretamente apenas 18.29% dos mesmos.

5.8.2 Modelos de RLM e AD com quatro grupos cujas suposições são violadas

Tabela 5.42: Modelo de regressão logística multinomial

Variáveis	$Coef_1$	$Coef_2$	$Coef_3$	$exp(Coef_1)$	$exp(Coef_2)$	$exp(Coef_3)$
(Intercept)	-3.4844348	0.9552706	2.1438075	0.03067109	2.59937381	8.53186123
IMC	0.05857277	0.01078115	-0.04212795	1.0603221	1.0108395	0.9587471
Trig	0.00090099	-0.00304591	-0.00104750	1.0009014	0.9969587	0.9989530
ColTotal	0.00494889	-0.00884208	-0.00170173	1.0049612	0.9911969	0.9982997
HdlC	-0.03300272	-0.07240132	0.009446633	0.9675359	0.9301575	1.0094914
Hematocrito	0.01701209	0.06704781	-0.04279662	1.0171576	1.0693466	0.9581062

Tabela 5.43: Estatística do teste da razão de verossimilhança

Chi-square	df	p
50.40271	5	1.146237e-09

Tabela 5.44: Teste de ajustamento do modelo de RLM

Chi-square	df	p	AIC
50.40271	76	0.9896606	213.3692

A estatística de teste da razão de verossimilhança apresentada na tabela 5.43 sugere que a um nível de significância de 5%, os modelos ilustrados na tabela 5.42 são estatisticamente significativos. Estes modelos melhor se ajustam aos dados (tabela 5.44).

Tabela 5.45: Modelo discriminante com quatro grupos

Variáveis	LDA1	LDA2	LDA3	df	Sum Sq	Mean Sq	F	Pr(>|F|)
IMC	0.0986021	-0.0544526	0.03109	1	2.40	2.4006	1.21	0.276
Trig	0.0014025	-0.0044588	0.00106	1	0.72	0.7204	0.36	0.549
ColTotal	0.0012605	-0.0152524	-0.00187	1	0.53	0.5267	0.27	0.609
HdlC	-0.0461509	-0.0311779	-0.03155	1	0.25	0.2531	0.13	0.722
Hematocrito	0.0909584	0.0744681	-0.16848	1	0.82	0.8179	0.41	0.524
Resíduos	...	...	...	76	151.34	1.9913		

Tabela 5.46: Estatística do teste Lambda de Wilks

Lambda de Wilks	Chi-square	df	Pr(>|z|)
0.877	10.002	15	0.820
0.958	3.250	8	0.918
0.997	0.222	3	0.974

Os modelos da análise discriminante com quatro grupos (Tabela 5.45), cujas variáveis explicativas violam as suas suposições, não são estatisticamente significativos, pois, todos os p-valores da tabela 5.46 são maiores que o nível de significância selecionado ($\alpha = 0.05$).

Tabela 5.47: Matriz de classificação de análise discriminante com quatro grupos e regressão logística multinomial.

	Análise discriminante com quatro grupos					Regressão Logística Multinomial				
Observado	Previsto					Previsto				
	0	1	2	3	% corr.	0	1	2	3	% corr.
0	1	3	11	21	2.78	23	0	0	13	63.89
1	1	1	2	3	14.29	6	0	0	1	0
2	0	0	4	3	57.14	6	0	0	1	0
3	5	3	15	9	28.13	15	0	0	17	53.13
% Total					18.29					48.78

A eficiência da regressão logística multinomial, quando comparada com a da análise discriminante com quatro grupos (Tabela 5.47) cujas variáveis violam as suposições em destaque, é superior, e constatamos que não houve alteração do valor da mesma eficiência se compararmos com o caso em que as variáveis satisfazem essas suposições.

Capítulo 6

Discussão e conclusões

Na introdução deste trabalho, foi abordado que num estudo desenvolvido por EFRON (1975) sobre a eficiência assintótica relativa da regressão logística comparada com a análise discriminante consideram-se apenas as variáveis que satisfazem as suposições da análise discriminante. Além disso, a variável dependente é de duas classes.

Os modelos utilizados para classificação de indivíduos neste caso são a análise discriminante com dois grupos e a regressão logística binária. Estes modelos são bastante úteis em várias áreas, sobretudo na área de saúde. A utilidade desses modelos na área biomédica varia desde a identificação dos efeitos que mais contribuem para a classificação dos pacientes para certo grupo até a estimação de o risco de um paciente ser classificado para certo grupo em análise.

O estudo que foi desenvolvido neste trabalho que visa analisar as capacidades preditivas dos modelos de interesse num contexto assintótico ou simplesmente relativo é de extrema importância para a identificação de um modelo que melhor classifica os indivíduos. Em concordância com os objetivos deste trabalho, primeiro abordamos nele a teoria sobre os métodos de classificação de indivíduos, verificamos se as variáveis satisfazem as suposições em causa, e com essas variáveis estimamos os modelos que foram comparados.

Decorre então, que do estudo comparativo que apresentamos na seção anterior, os resultados revelam que no caso em que as variáveis explicativas satisfazem as suposições da análise discriminante, o método de regressão logística é mais eficiente em relação à análise discriminante de dois grupos em termos de classificação de indivíduos, embora essa eficiência tenha decrescido em ambos os métodos com a utilização de uma amostra com um tamanho maior (com o caso de sobrevida, n=82) em relação ao da primeira (casos chagásicos, n=59). Os resultados análogos a estes, foram observados no estudo de PRESS e WILSON (1978), embora não tivessem sido analisados com relação aos tamanhos da amostra. Relativamente à análise da eficiência assintótica, o método de regressão logística é assintoticamente menos

eficiente quando comparado com a análise discriminante.

Os valores obtidos neste trabalho da eficiência assintótica relativa da regressão logística comparada com a análise discriminante estão entre 1/8 e 1/7 para os casos de chagas ($n = 59$) e entre 1/9 e 1/8 no caso da amostra de pacientes renais crônicos (caso sobrevida, n=82). Estes resultados indicam que a regressão logística é assintoticamente menos eficiente em relação à análise discriminante.

Este resultado numérico da eficiência assintótica foi menor que o resultado obtido por EFRON (1975), revelando a evidência de que a análise discriminante é assintoticamente mais eficiente.

No caso em que as suposições são violadas, a inclusão dessas variáveis explicativas que violam as suposições em causa, gera um modelo de regressão logística também mais eficiente em termos de classificação de indivíduos em relação à análise discriminante. Além disso, também a eficiência relativa decresce com a utilizaçâo de uma amostra de tamanho maior em relação ao da primeira.

Não obstante tenha-se constatado que o método da regressão logística multinomial foi mais eficiente em relação ao método discriminante com mais de dois grupos, a mesma eficiência relativa não variou na presença de variáveis explicativas que violam as suposições em causa.

Com base nos resultados deste trabalho, conclui-se que quando as suposições da análise discriminante são satisfeitas, a regressão logística ou regressão logística multinomial é mais eficiente em relação à análise discriminante de dois ou mais grupos respectivamente. Além disso, o método de regressão logística é assintoticamente menos eficiente que a análise discriminante somente se as variáveis explicativas têm distribuição normal multivariada. Para outros membros da família exponencial a regressão logística é mais eficiente.

No caso da violação das suposições da análise discriminante, tanto para duas populações quanto para mais de duas, a regressão logística é robusta mantendo-se relativamente mais eficiente em relação à análise discriminante.

6.1 Recomendações

A utilização da regressão logística é vista neste trabalho, como uma forma conveniente em aplicações, dado que ela não impõe que as suposições da análise discriminante sejam satisfeitas, e por ser uma técnica robusta na classificação de indivíduos.

Para trabalhos futuros recomenda-se: a aplicação das técnicas deste trabalho com vários bancos de dados para duas e mais de duas populações; a variação de número de classes para o caso dos modelos de regressão logística multinomial e análise discriminante de mais de dois grupos, a fim de analisar a eficiência assintótica relativa

para modelos com múltiplos grupos; e a utilização de outras técnicas estatísticas de classificação de indivíduos, como é o caso de árvores de classificação.

Referências Bibliográficas

ALVES, M. *Sobrevida de pacientes renais crônicos submetidos a tratamento hemodialítico e associação com os polimorfismos dos genes da ECA e do angiotensinogênio*. DSc. Tese, UFRJ/Programa de Pós-graduação em Medicina (Cardiologia) do Departamento de Clínica Médica da Faculdade de Medicina e do Instituto do Coração Edson Saad, Rio de Janeiro-Brasil, 2012.

BULL, S. B., DONNER, A. "The efficiency of multinimial logistic regression compared with multiple group discriminant analysis", *Journal of the American Statistical Association*, v. 82, n. 400, pp. 1118–1122, 1987a.

BULL, S. B., DONNER, A. "Derivation of Large Sample Efficiency of Multinomial Logistic Regression Compared to Multiple Group Discriminant Analysis". In: eds. I. B. MacNeill, Umphreyi, G. J. (Eds.), *in Proceedings of the Symposia in Statistics and Festschrift in Honour of V. M. Joshi. Advances in the Statistical Sciences-Biostatistics*, v. 5, pp. 177–197, Boston: D. Reidel, 1987b.

BUSE, A. "The Likelihood Ratio, Wald and Lagrange Multiplier Tests: An Expository Note", *The American Statistician*, v. 36, n. 3, pp. 153–157, 1982.

BROOKS, C. A., ET AL. "The robustness of the logistic risk function", *Communications in Statistics-Simulation*, v. 17, n. 1, pp. 1–24, 1988.

COX, D. R., SNELL, E. J. *Analysis of Binary Data*. 2 ed. London, Chapman and Hall, 1989.

DAY, N. E., KERRIDGE, D. "A general maximum likelihood discriminant", *Biometrics*, v. 23, n. 2, pp. 313–323, 1967.

EFRON, B. "The efficiency of logistic regression compared to normal discriminant analysis", *Journal of the American Statistical Association*, v. 70, n. 352, pp. 892–898, 1975.

ENNIS, M., ET AL. “A comparison of statistical learning methods on the gusto data base”, *Statistics in Medicine*, v. 17, pp. 2501–2508, 1998.

GORDON, T. “Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies”, *J Chron Dis.*, v. 27, n. 3, pp. 97–102, 1974.

HAIR, J. F., ET AL. *Análise multivariada de dados*. 6 ed. São Paulo, Bookman, 2009.

HAGGSTROM, G. W. “Logistic Regression and Discriminant Analysis by Ordinary Least Squares”, *Journal of Business and Economic Statistics*, v. 1, n. 3, pp. 229–238, 1983.

HALPERIN, M., ET AL. “Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches”, *Journal of Chronic Diseases*, v. 24, pp. 125–158, 1971.

HOSMER, D. W., LEMESHOW, S. *Applied Logistic Regression*. 2 ed. New York, John Wiley and Sons, 2000.

LANDESMANN, M. C. P., ET AL. “Iodine-123 Metaiodobenzylguanidine Cardiac Imaging as a Method to Detect Early Sympathetic Neuronal Dysfunction in Chagasic Patients With Normal or Borderline Electrocardiogram and Preserved Ventricular Function”, *Clinical Nuclear Medicine*, v. 36, n. 9, pp. 757–761, 2011.

MARDIA, K. V., ET AL. *Multivariate Analysis*. London. New York. Toronto. Sydney. San Francisco, Academic Press. A series of Monographs and Textbooks, 1979.

PRESS, S. J., WILSON, S. “Choosing between logistic regression and discriminant analysis”, *Journal of the American Statistical Association*, v. 73, n. 364, pp. 699–705, 1978.

POHAR, M., ET AL. “Comparison of logistic regression and linear discriminant analysis: a simulation study”, *Metodološki zvezki*, v. 1, n. 1, pp. 143–161, 2004.

REID, N. “Likelihood inference”, *John Wiley and Sons, inc*, v. 2, pp. 517–525, 2010.

TRUETT, J., ET AL. “A multivariate analysis of the risk of coronary heart disease in Framinghan”, *Jornal of Chronic Diseases.*, v. 20, n. 7, pp. 511–524, 1967.

WALKER, S. H., DUNCAN, D. B. "Estimation of the probability of an event as a function of several independent variables", *Biometrika*, v. 54, n. 1/2, pp. 167–169, 1967.

Apêndice A

Programa utilizado no processamento de dados

```
#########Histogramas das variáveis explicativas dos casos de chagas##############
par(mfrow=c(1,3))
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
hist(analises$hm20, main=" ", xlab="Histograma-hm20", ylab="Frequências Relativas", prob = T)
curve(dnorm(x, mean= 1.970339, sd=0.2484848), 1.4, 2.8, add=T)
hist(analises$hm3, main=" ", xlab="Histograma-hm3", ylab="Frequências Relativas", prob = T)
curve(dnorm(x, mean=1.960339 , sd=0.3230297), 1, 10, add=T)
hist(analises$washout, main=" ", xlab="Histograma-washout", ylab="Frequências Relativas", prob = T)
curve(dnorm(x, mean=0.26, sd=0.06197886), 0.10, 0.45, add=T)

#####Histogramas das variáveis explicativas dos casos de sobrevida################
par(mfrow=c(1,2))
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
hist(analises$Idade, main=" ", xlab="Histograma-Idade", ylab="Frequências Relativas", prob = T)
curve(dnorm(x, mean= 52.60976, sd=15.24674), 10, 90, add=T)
hist(analises$PASpre, main=" ", xlab="Histograma-PASpre", ylab="Frequências Relativas", prob = T)
curve(dnorm(x, mean= 151.3293, sd= 17.69847), 100, 200, add=T)

####Teste da normalidade das variáveis explicativas para os dois bancos de dados#####
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
shapiro.test(analises$age)
shapiro.test(analises$hm20)
shapiro.test(analises$hm3)
shapiro.test(analises$washout)
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
shapiro.test(analises$Idade)
shapiro.test(analises$IMC)
shapiro.test(analises$Trig)
shapiro.test(analises$ColTotal)
```

```
shapiro.test(analises$HdlC)
shapiro.test(analises$PASpre)
shapiro.test(analises$Hematocrito)
##############################Teste M-Box ##############################
```
O teste M-Box não está incluído em R. Mas o código está disponível em http://www.statmethods.net/stats/anovaAssumptions.html
```
#############Modelo de regressão logística de casos de chagas cujas variáveis satisfazem os pressupostos da análise discriminante ############################
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
analises$spect <- factor(analises$spect)
analises$sex <- factor (analises$sex)
analises$ecg <- factor (analises$ecg)
reglog<-glm(classebin~hm20 + hm3 + washout,data=analises, family=binomial(link="logit"))
summary(reglog)
########################Coeficientes exponenciados ######################
exp(coef(reglog))
################Estatística do teste da razão de verossimilhança################
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,3)
######Teste de ajustamento do modelo de regressão logística####################
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,55)

###Modelo de análise discriminante cujas variáveis satisfazem as suposições#########
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
classebin <- c(rep("chagas", 40), rep("controle", 19))
classebin <- factor(classebin)
library(MASS)
LDA <- lda(classebin ~ hm20 + hm3 + washout, data=analises, na.action="na.omit", CV=FALSE)
LDA
```

```
attach(analises)
summary(aov(classebin ~ hm20 + hm3 + washout, data=analises))
################Matriz de classificação da Regressão Logística ################
proba <- predict(reglog, newdata=analises, type="response")
pred <- ifelse(proba < 0.5, 0, 1)
pred <- factor (pred)
mc <- table (analises$classebin, pred)
print(mc)
###########Matriz de classificação da discriminante linear#####################
ct <- table(analises$classebin, classebin)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))

##Modelo da RL para casos de sobrevida cujas variáveis satisfazem as suposição da análise discriminante#########################################################
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
analises$genero <- factor(analises$genero)
analises$AVE <- factor (analises$AVE)
analises$DIC <- factor (analises$DIC)
analises$Diabetes <- factor (analises$Diabetes)
analises$Tabagismo <- factor (analises$Tabagismo)
analises$ECA <- factor (analises$ECA)
analises$ANGo <- factor (analises$ANGo)
reglog<-glm(censura~Idade + PASpre,data=analises, family=binomial(link="logit"))
summary(reglog)
############################## Coeficientes exponenciados#################
exp(coef(reglog))

###################Estatística do teste da razão de verossimilhança############
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,2)
```

```
#############Teste de ajustamento do modelo de regressão logística#############
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,79)

###############Modelo da ADL de caso sobrevida cujas variáveis satisfazem os pressupostos#################################################################
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
censura <- c(rep("morte", 44), rep("sobrevida", 38))
censura <- factor(censura)
library(MASS)
LDA<-lda(censura ~ Idade + PASpre, data=analises)
LDA
attach(analises)
summary(aov(censura ~ Idade + PASpre, data=analises))

#############Matriz de classificação da Regressão logistica e de ADL############
proba <- predict(reglog, newdata=analises, type="response")
pred <- ifelse(proba < 0.5, 0, 1)
pred <- factor (pred)
mc <- table (analises$censura, pred)
print(mc)
#########################################################################
ct <- table(analises$censura, censura)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))
#######Distância de Mahalanobis para casos de chagas e caso de sobrevida#########
################################Chagásicos################################
x<-read.csv2("C:\\banco1carolina1.csv",head=T,sep=";")
stopifnot(mahalanobis(x, 0, diag(ncol(x))) == rowSums(x*x))
Sx <- cov(x)
D2 <- mahalanobis(x, colMeans(x), Sx)
y<-mean(D2)
sqrt(y)
```

```
##########################Sobrevida#####################################
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
g<-cbind(analises$Idade, analises$PASpre)
stopifnot(mahalanobis(g, 0, diag(ncol(g))) == rowSums(g*g))
Sx <- cov(g)
D2 <- mahalanobis(g, colMeans(g), Sx)
y<-mean(D2)
y
sqrt(y)

###########Modelo de Regressão logistica de casos de chagas cujas variáveis violam as suposições da análise discriminante#########################################
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
reglog<-glm(classebin~age ,data=analises, family=binomial(link="logit"))
summary(reglog)
############################Coeficiente exponenciados####################
exp(coef(reglog))
#################Estatística do teste da razão de verossimilhança###############
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,1)
###############Teste de ajustamento do modelo de regressão logística###########
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev,57)
#######Modelo de ADL de casos de chagas cujas variáveis violam suas suposições###
analises<-read.csv2("C:\\banco1carolina.csv",head=T,sep=";")
classebin <- c(rep("chagas", 40), rep("controle", 19))
classebin <- factor(classebin)
library(MASS)
LDA <- lda(classebin ~ age, data=analises, na.action="na.omit", CV=FALSE)
LDA
attach(analises)
summary(aov(classebin ~ age, data=analises))
```

```
###########Matriz de classificação do modelo da Regressão Logística#############
proba <- predict(reglog, newdata=analises, type="response")
pred <- ifelse(proba < 0.5, 0, 1)
pred <- factor (pred)
mc <- table (analises$classebin, pred)
print(mc)
##################Matriz de classificação da discriminante linear#################
ct <- table(analises$classebin, classebin)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))

####Modelo de RL dos casos de sobrevida cujas variáveis violam as suposições######
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
analises$genero <- factor(analises$genero)
analises$AVE <- factor (analises$AVE)
analises$DIC <- factor (analises$DIC)
analises$Diabetes <- factor (analises$Diabetes)
analises$Tabagismo <- factor (analises$Tabagismo)
analises$ECA <- factor (analises$ECA)
analises$ANGo <- factor (analises$ANGo)
analises$RendaMensal <- factor (analises$RendaMensal)
analises$CID <- factor (analises$CID)
reglog<-glm(censura~IMC + Trig + ColTotal + HdlC + Hematocrito,data=analises,
family=binomial(link="logit"))
summary(reglog)

####################Coeficientes exponenciados##############
exp(coef(reglog))
#######################Teste da Razão de verossimilhança e de Ajustamento#####
reglog.dev<-sum(resid(reglog,type="deviance")^2)
reglog.dev
1-pchisq(reglog.dev, 5)
1-pchisq(reglog.dev, 76)
```

```
###Modelo de ADL dos casos de sobrevida cujas variáveis violam as suposições######
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
censura <- c(rep("morte", 44), rep("sobrevida", 38))
censura <- factor(censura)
library(MASS)
LDA<-lda(censura ~ IMC + Trig + ColTotal + HdlC + Hematocrito, data=analises)
LDA
attach(analises)
summary(aov(censura ~ IMC + Trig + ColTotal + HdlC + Hematocrito, data=analises))
############Matriz de classificação da regressão logística######################
proba <- predict(reglog, newdata=analises, type="response")
pred <- ifelse(proba < 0.5, 0, 1)
pred <- factor (pred)
mc <- table (analises$censura, pred)
print(mc)
################Matriz de classificação da ADL ###################
ct <- table(analises$censura, censura)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))
#Modelo de RLM para quatro grupos cujas variáveis satisfazem as suposições da ADL#
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
analises$Diabetes <- factor (analises$Diabetes)
analises$genero <- factor(analises$genero)
analises$AVE <- factor (analises$AVE)
analises$DIC <- factor (analises$DIC)
analises$CID <- factor (analises$CID)
analises$Diabetes <- factor (analises$Diabetes)
analises$Tabagismo <- factor (analises$Tabagismo)
analises$ECA <- factor (analises$ECA)
analises$ANGo <- factor (analises$ANGo)
analises$RendaMensal <- factor (analises$RendaMensal)
require(nnet)
analises$CID2<-relevel(analises$CID, ref="0")
```

```
mlogit.model<- multinom(CID2~Idade+PASpre, data = analises, reflevel="0") #a categoria basal é zero
summary(mlogit.model)
############################coeficientes exponenciados####################
exp(coef(mlogit.model))
###Estatística de teste da Razão de verossimilhança e de ajuste do modelo##########
mlogit.model.dev<-sum(resid(mlogit.model,type="deviance")^2)
mlogit.model.dev
1-pchisq(mlogit.model.dev,2)
1-pchisq(mlogit.model.dev,79)
####Modelo de ADL para quatro grupos cujas variáveis satisfazem as suposições#####
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
CID <- c(rep("outras causas", 32),rep("doença infecciosa", 7), rep("doença aterotrombotica vascular", 7), rep("Pacientes vivos", 36))
CID <- factor(CID)
library(MASS)
LDA<-lda(CID ~ Idade+PASpre, data=analises)
LDA
attach(analises)
summary(aov(CID ~ Idade + PASpre, data=analises))
####################Matriz de classificação da RLM###################
dIdade <- data.frame(analises$CID , PASpre = mean(analises$PASpre))
proba <- predict(mlogit.model, newdata=analises, type="probs")
sum(proba[1,])
sum(proba[2,])
nr=82
nc=4
cid11=analises$CID
clasif=matrix(0,4,4)
for(i in 1:82){
   k=cid11[i]
   mm=max(proba[i,])
  for(j in 1:4){
    if(proba[i,j]==mm){clasif[k,j]=clasif[k,j]+1}
```

```
  }
}
clasif
##########Matriz de classificação da ADL para quatro grupos#########
ct <- table(analises$CID, CID)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))
ct <- table(analises$CID, CID)
x<-matrix(prop.table(ct,1))
matrix(x, 4, 4)
#################Modelo de RLM cujas variáveis violam as suposições###########
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
analises$Diabetes <- factor (analises$Diabetes)
analises$genero <- factor(analises$genero)
analises$AVE <- factor (analises$AVE)
analises$DIC <- factor (analises$DIC)
analises$CID <- factor (analises$CID)
analises$Diabetes <- factor (analises$Diabetes)
analises$Tabagismo <- factor (analises$Tabagismo)
analises$ECA <- factor (analises$ECA)
analises$ANGo <- factor (analises$ANGo)
analises$RendaMensal <- factor (analises$RendaMensal)
require(nnet)
analises$CID2<-relevel(analises$CID, ref="0")
mlogit.model<- multinom(CID2~IMC + Trig + ColTotal + HdlC + Hematocrito, data =
analises, reflevel="0") #a categoria basal é zero
summary(mlogit.model)
exp(coef(mlogit.model))

##################Testes da razão de verossimilhança e de ajustamento##########
mlogit.model.dev<-sum(resid(mlogit.model,type="deviance")^2)
mlogit.model.dev
1-pchisq(mlogit.model.dev,5)
1-pchisq(mlogit.model.dev,76)
```

```
######Modelo de ADL para quatro grupos cujas variáveis violam as suposições#####
analises<-read.csv2("C:\\banco2mauro.csv",head=T,sep=";")
CID <- c(rep("outras causas", 32),rep("doença infecciosa", 7), rep("doença aterotrombotica vascular", 7), rep("Pacientes vivos", 36))
CID <- factor(CID)
library(MASS)
LDA<-lda(CID ~ IMC + Trig + ColTotal + HdlC + Hematocrito, data=analises)
LDA
attach(analises)
summary(aov(CID ~IMC + Trig + ColTotal + HdlC + Hematocrito, data=analises))
######Matriz de classificação da Regressão Logística Multinomial###############
dIMC <- data.frame(analises$CID , ColTotal = mean(analises$ColTotal))
proba <- predict(mlogit.model, newdata=analises, type="probs")
sum(proba[1,])
sum(proba[2,])
nr=82
nc=4
cid11=analises$CID
clasif=matrix(0,4,4)
for(i in 1:82){
   k=cid11[i]
   mm=max(proba[i,])
  for(j in 1:4){
     if(proba[i,j]==mm){clasif[k,j]=clasif[k,j]+1}
  }
}
clasif
##############Matriz de classificação da ADL com quatro grupos cujas variáveis violam as suposições#############################################################
ct <- table(analises$CID, CID)
diag(prop.table(ct,1))
sum(diag(prop.table(ct)))
ct <- table(analises$CID, CID)
x<-matrix(prop.table(ct,1))
matrix(x, 4, 4)
#########################################################################
```