

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

Tatiane Gomes de Araújo

**APLICAÇÃO DA REGRESSÃO LOGÍSTICA PARA O ESTUDO DA DOENÇA
ARTERIAL CORONARIANA**

Uberlândia – MG
Novembro/2015

Tatiane Gomes de Araújo

**APLICAÇÃO DA REGRESSÃO LOGÍSTICA PARA O ESTUDO DA DOENÇA
ARTERIAL CORONARIANA**

Trabalho de conclusão de curso de graduação
apresentado à Faculdade de Matemática da
Universidade Federal de Uberlândia (UFU) como
requisito parcial para a obtenção do título de
Bacharel em Estatística.

Universidade Federal De Uberlândia
Faculdade De Matemática
Curso De Graduação Em Estatística

Orientador: Prof. Dr. JANSER MOURA PEREIRA

Uberlândia – MG
Novembro/2015

Tatiane Gomes de Araújo

**APLICAÇÃO DA REGRESSÃO LOGÍSTICA PARA O ESTUDO DA DOENÇA
ARTERIAL CORONARIANA**

Trabalho de conclusão de curso de graduação
apresentado à Faculdade de Matemática da
Universidade Federal de Uberlândia (UFU) como
requisito parcial para a obtenção do título de
Bacharel em Estatística.

Aprovada em: ____ de _____ de _____.

Prof. Dr. Janser Moura Pereira
Orientador

Prof. Dr. José Waldemar da Silva
Convidado

Profa. Dra. Elisângela Rosa Cordeiro
Convidado

Uberlândia – MG
Novembro/2015

"Aos meus pais e irmãos que com muito apoio e carinho, não mediram esforços para que eu chegasse até essa etapa da minha vida."

AGRADECIMENTOS

Agradeço primeiramente a Deus que me deu força e energia para concluir o curso.

Agradeço aos meus pais e irmãos que me deram apoio e incentivo durante todos os anos que estive na faculdade.

Agradeço aos meus familiares e amigos por deixarem a caminhada mais leve.

Agradeço aos meus professores, em especial ao meu orientador, Janser Moura Pereira, pela paciência e pelos ensinamentos.

Agradeço aos meus colegas de curso pelo auxílio e aprendizado em todo percurso.

E a todos, que direta ou indiretamente fizeram parte da minha formação.

*"O coração do homem traça o seu caminho,
mas o SENHOR lhe dirige os passos"*

Pv. 16.9

RESUMO

Este trabalho tem por objetivo avaliar por meio da regressão logística múltipla quais fatores que influenciam no desenvolvimento da doença arterial coronariana. Os dados foram coletados no Setor de Cardiologia do Hospital de Clínicas da Universidade Federal de Uberlândia (HC-UFU) mediante um termo de consentimento assinado pelo paciente e autorização da Comissão de Ética Humana da UFU. Foram incluídos no estudo 181 indivíduos com idade entre 39 e 87 anos, de ambos os sexos. Analisou-se 14 variáveis, que representam possíveis fatores de risco, modificáveis ou não, para o desenvolvimento da doença. Dentre as variáveis estudadas observou-se que apenas duas apresentaram relação significativa com a variável resposta. Para avaliar se o modelo de regressão logística é realmente eficaz foi realizado o Teste de Hosmer-Lemeshow, visando aferir a qualidade do ajuste do modelo.

Palavras-chave: *Logit*, Inferência, Cardiopatia Coronária.

ABSTRACT

This work aims to evaluate through multiple logistic regression which factors that influence in the development of coronary artery disease. The data were collected in the field of Cardiology Clinical Hospital of the Federal University of Uberlândia (HC-UFU) by means of a consent form signed by the patient and authorization of the Human Ethics Committee of the UFU. Were included in the study 181 individuals aged between 39 and 87 years, of both sexes. 14 variables were analyzed, representing possible risk factors, modifiable or not, to the development of the disease. Among the variables studied showed that only two showed a significant relationship with the response variable. To assess whether the logistic regression model is really effective has been conducted the test of Hosmer-Lemeshow, aiming to assess the quality of the fit of the model.

Keywords: *Logit*, Inference, Coronary Heart Disease.

SUMÁRIO

1 INTRODUÇÃO.....	9
2 MATERIAIS E MÉTODOS.....	11
2.1 FORMULAÇÃO MATEMÁTICA.....	12
2.1.1 Regressão logística simples.....	12
2.1.2 Regressão Logística Múltipla.....	16
2.1.3 Estimação dos Parâmetros.....	17
2.1.4 Adequação do modelo.....	18
3 RESULTADOS E DISCUSSÕES.....	21
4 CONCLUSÃO.....	25
REFERÊNCIAS.....	26

1 INTRODUÇÃO

No Brasil e no mundo as doenças cardiovasculares são consideradas um grave problema de saúde. Além de ser uma das principais causas de morbi-mortalidades, são também responsáveis por altos custos em assistência médica (PINHO *et. al.*, 2010).

Dentre as doenças cardiovasculares pode-se destacar a doença arterial coronariana (DAC). A DAC é caracterizada pelo fornecimento inadequado do fluxo sanguíneo ao coração. Placas de gordura formadas por cálcio, colesterol e outras substâncias são depositadas no interior das artérias, provocando um estreitamento (estenose) obstruindo a passagem de sangue e oxigênio ao coração. Essas placas são chamadas de placas ateroscleróticas (ARAÚJO, 2008).

Estudos epidemiológicos indicam que são vários fatores de risco que contribuem para o desenvolvimento da DAC, eles podem ser genéticos ou adquiridos, e associados podem aumentar o risco de progressão da doença. Os fatores mais comuns são: tabagismo, diabetes, hipertensão, gordura abdominal, dislipidemias, sedentarismo, histórico familiar, fatores psicossociais, hábitos alimentares, idade, sexo, entre outros (MAIA *et. al.*, 2007).

Dentre os fatores modificáveis, a hipertensão arterial sistêmica (HAS) é responsável por cerca de 40% dos óbitos por DAC, seguida de diabetes mellitus (DM) e as dislipidemias (MAIA *et. al.*, 2007).

Vários estudos foram feitos visando estimar a prevalência desses fatores em diversas populações a fim de melhor redirecionar os métodos de prevenção em relação a DAC. Em um estudo para identificar os fatores de risco modificáveis, que compreendem aos hábitos adquiridos ao longo da vida, como má alimentação e tabagismo, da DAC em trabalhadores de enfermagem, foram constatados os seguintes fatores: estresse, hipertensão, alto índice de IMC e sedentarismo. Identificar esses fatores é importante para desenvolver ações e educação em saúde relevantes na intervenção preventiva, inclusive desses profissionais (MAIA *et. al.*, 2007).

As intervenções utilizadas no tratamento da DAC são medicamentosas, mudança no estilo de vida e procedimentos cirúrgicos. Os procedimentos cirúrgicos de revascularização miocárdica (CRM) são indicados em alguns casos no tratamento da DAC, garantindo uma atenuação dos sintomas e melhora da capacidade funcional dos pacientes. No entanto, há a possibilidade dessa patologia se desenvolver novamente,

pois a cirurgia não atua nos fatores etiopatogênicos da doença, ou seja, não atua em suas causas ou em seu desenvolvimento (RADAELLI, 2011).

Assim, é de fundamental importância determinar quais fatores influenciam no desenvolvimento da DAC, sejam eles modificáveis ou não. Nesse sentido, os modelos de regressão logística, nos quais a variável resposta é dicotômica, são uma alternativa plausível para elencar quais são as variáveis que podem influenciar no desenvolvimento da DAC.

De acordo com Corrar *et. al.* (2007) a técnica de Regressão Logística foi desenvolvida na década de 1960, devido à necessidade de se realizar previsões ou explicar alguns acontecimentos em que a variável dependente fosse binária. O estudo Framingham Heart Study (DAWBER *et. al.*, 1950), realizado com a colaboração da Universidade de Boston, foi um dos primeiros a destacar esta técnica. Este estudo tinha a finalidade de identificar fatores que contribuem para o desenvolvimento de doenças cardiovasculares. Foram avaliados 5.209 indivíduos entre 30 e 60 anos, moradores da cidade de Framingham, em Massachusetts. Por meio da regressão logística foi possível identificar vários fatores de risco, como: hipertensão arterial, taxas de colesterol elevadas, obesidade, diabetes, sedentarismo e tabagismo.

Embora esta técnica tenha se desenvolvido na área médica, ela também é empregada em outras áreas como economia e marketing e, mais recentemente, no desenvolvimento dos *Credit Scoring* (CORRAS *et. al.*, 2007).

Portanto, o presente trabalho tem por objetivo verificar por meio de ajuste de modelos de regressão logística quais são as variáveis que influenciam o desenvolvimento da DAC.

2 MATERIAIS E MÉTODOS

Os dados foram coletados no Setor de Cardiologia do Hospital de Clínicas da Universidade Federal de Uberlândia (HC-UFU) mediante um termo de consentimento assinado pelo paciente e autorização da Comissão de Ética Humana da UFU, conforme parecer: 189.679. Foram incluídos no estudo 181 indivíduos com idade entre 39 e 87 anos, de ambos os sexos, que apresentaram algum sintoma coronário e que fizeram angiografia coronária. Para cada indivíduo que teve DAC comprovada ou não pela angiografia coronária, foram coletadas as seguintes variáveis:

X_1 : Idade, em anos completos;

X_2 : Sexo, se Masculino = 0 e se Feminino = 1;

X_3 : Sedentarismo, se o paciente não realiza alguma atividade física (SIM = 1 ou NÃO = 0);

X_4 : Tabagismo, se o paciente é dependente do consumo de tabaco (SIM = 1 ou NÃO = 1);

X_5 : Hipertensão arterial sistêmica (HAS), que é uma doença crônica determinada por elevados níveis de pressão sanguínea nas artérias, que leva a um aumento do esforço normal que o coração faz para o sangue circular nos vasos sanguíneos (SIM = 1 ou NÃO = 0);

X_6 : Ant. DAC (ADAC), caso haja algum parente de primeiro grau que tenha DAC (SIM = 1 ou NÃO = 0);

X_7 : Histórico familiar da DAC (HFDAC), caso haja algum histórico familiar do paciente (SIM = 1 ou NÃO = 0);

X_8 : Diabetes *mellitus* (DM), doença causada devido a taxas elevadas de glicose no sangue (SIM = 1 ou NÃO = 0);

X_9 : Colesterol Total (CT), é a soma de todos os colesteróis (HDL, LDL, VLDL) em que o nível considerado bom não pode ser superior a 200mg/dL ($CT < 200 = 0$ ou $CT > 200 = 1$);

X_{10} : Lipoproteína de alta densidade ou "colesterol bom" (HDL), ele absorve o cristais de colesterol depositados nas artérias e os carrega para o fígado, onde é eliminado. Quanto mais elevado os níveis melhor, valores inferiores a 40 mg/dL são considerados baixos ($HDL > 40 = 0$ ou $HDL \leq 40 = 1$);

X_{11} : Lipoproteína de baixa densidade ou "colesterol ruim" (LDL), ele carrega os cristais de colesterol do fígado até as artérias, provocando um acúmulo de

gordura que leva a diminuição da passagem do fluxo sanguíneo. Considera-se acima do normal valores de LDL maiores que 130mg/dL ($LDL < 130 = 0$ ou $LDL > 130 = 1$);

X_{12} : Triglicerídeos (TG) é um tipo de gordura produzido pelo organismo ou adquiridos a partir do consumo de carboidratos ou de alimentos gordurosos e é a forma de armazenamento energético mais importante no organismo. No entanto, valores maiores que 150mg/dL aumentam o risco de doenças cardíacas ($TG < 150 = 0$ ou $TG > 150 = 1$);

X_{13} : Índice de massa corporal (IMC), que é uma medida do nível de gordura corporal reconhecida pela Organização Mundial da Saúde (OMS), calculada pelo peso, em quilogramas, dividido pela altura, em metros, ao quadrado. Nesse estudo para efeito de comparação foi considerado o nível de classificação de obesidade ($IMC < 30 = 0$ ou $IMC > 30 = 1$);

X_{14} : Genótipo: o sistema renina angiotensina (SRA) é um sistema hormonal cuja cascata de formação inclui interações enzima-substratos que resultam na produção de vários peptídeos biologicamente ativos (MOURA, 2007) influenciando o sistema cardiovascular de forma direta e indireta. A enzima conversora de angiotensina (ECA) é uma componente chave desse sistema, seu gene se encontra no cromossomo 17 e o polimorfismo do intron 16 consiste em dois alelos: deleção (D) e inserção (I) que resultam em 3 genótipos: DD, que possuem o maior nível plasmático de ECA; ID, com níveis intermediários e II com níveis inferiores (FRANKEN *et. al.*, 2004) ($GDD = 1$, $GID = GII = 0$).

2.1 FORMULAÇÃO MATEMÁTICA

2.1.1 Regressão logística simples

Primeiramente, considere o modelo de regressão linear simples conforme equação (1):

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (1)$$

em que $Y_i = 0$ ("fracasso") ou $Y_i = 1$ ("sucesso").

Em problemas de regressão modela-se a média condicional, que é o valor médio da variável resposta Y dado os valores da variável independente, x_i , designada por $E(Y_i | X = x_i)$, cujo valor será:

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i, \quad (2)$$

Já que Y_i assume dois resultados possíveis pode-se afirmar que Y_i é uma variável aleatória com distribuição Bernoulli, isto é, com distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1 | x_i) = \pi_i \quad (3)$$

$$Y_i = 0 \rightarrow P(Y_i = 0 | x_i) = 1 - \pi_i \quad (4)$$

Pela definição de esperança matemática tem-se que,

$$E(Y_i | x_i) = \pi_i \quad (5)$$

Igualando (2) e (5) temos,

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i = \pi_i = \pi(x_i) \quad (6)$$

Então, pode-se afirmar que $Y_i = \pi_i + \varepsilon_i$, em que a quantidade ε_i admite o valor $1 - \pi_i$ para $Y_i = 1$, ou $-\pi_i$ para 0, assim ε_i segue uma distribuição Bernoulli com média zero e variância igual a $\pi_i(1 - \pi_i)$, que podem ser verificadas. Sendo $P(\varepsilon_i = -\pi_i) = 1 - \pi_i$ e $P(\varepsilon_i = 1 - \pi_i) = \pi_i$, então o valor esperado de ε_i é (FIGUEIRA, 2006):

$$E(\varepsilon_i) = -\pi_i(1 - \pi_i) + (1 - \pi_i)\pi_i = -\pi_i + \pi_i^2 + \pi_i - \pi_i^2 = 0.$$

E a variância:

$$\begin{aligned} Var(\varepsilon_i) &= E(\varepsilon_i^2) - [E(\varepsilon_i)]^2 = E(\varepsilon_i^2) = (-\pi_i)^2(1 - \pi_i) + (1 - \pi_i)^2 \pi_i \\ &= (-\pi_i)^2(1 - \pi_i) + (1 - \pi_i)^2 \pi_i = \pi_i^2 - \pi_i^3 + (1 - 2\pi_i + \pi_i^2)\pi_i \\ &= \pi_i^2 - \pi_i^3 + \pi_i - 2\pi_i^2 + \pi_i^3 = \pi_i - \pi_i^2 \\ &= \pi_i(1 - \pi_i). \end{aligned}$$

Então, quando a variável resposta é binária, assume o valor 1(um) para representar o “sucesso” ou o “evento de interesse” ou assume o valor 0 (zero) para representar o “fracasso”, a resposta média representará a probabilidade de Y_i ser igual a 1, ao nível da variável preditora x_i (PAGANO & GAUVREAU, 2006). Sendo assim, a princípio poderia considerar um modelo da seguinte forma,

$$\pi_i = \pi(x_i) = \beta_0 + \beta_1 x_i. \quad (7)$$

O modelo (7) caracteriza o modelo de regressão linear simples em que a variável resposta aleatória contínua (Y_i), normalmente distribuída, foi trocada por π_i . Porém, como π_i é uma probabilidade, seus valores variam entre 0 e 1, e o termo $\beta_0 + \beta_1 x_i$ pode

gerar valores fora desse limite. Uma alternativa seria (PAGANO & GAUVREAU, 2006):

$$\pi_i = \pi(x_i) = e^{\beta_o + \beta_1 x_i} \quad (8)$$

A expressão (8) assegura que π_i seja positiva, mas o termo $e^{\beta_o + \beta_1 x_i}$ pode gerar um valor maior que 1. Assim, para satisfazer essa restrição, ajusta-se um modelo do tipo:

$$\pi_i = \pi(x_i) = \frac{e^{\beta_o + \beta_1 x_i}}{1 + (e^{\beta_o + \beta_1 x_i})} \quad (9)$$

O termo acima, chamado de função logística, não gera valores negativos ou maiores que 1, delimitando assim, o valor estimado de π_i de acordo com o intervalo requerido (PAGANO & GAUVREAU, 2006).

Uma propriedade interessante é que a função logística pode ser linearizada por meio da transformação logarítmica. Mas, primeiramente deve-se conhecer o conceito de chance. Define-se como chance, a probabilidade de ocorrência de um evento dividida pela probabilidade da não ocorrência do mesmo evento. A transformação é interpretada como o logaritmo da chance entre π_i e $1 - \pi_i$ (chance do sucesso), ou seja, logaritmo da chance. Esta interpretação é bastante empregada em estudos toxicológicos, epidemiológicos e em outras áreas, sendo definida como (PAGANO & GAUVREAU, 2006):

$$\begin{aligned} g(x_i) &= \ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \ln \left[\frac{\frac{e^{\beta_o + \beta_1 x_i}}{1 + (e^{\beta_o + \beta_1 x_i})}}{\frac{1}{1 + (e^{\beta_o + \beta_1 x_i})}} \right] \\ &= \ln \left[\frac{e^{\beta_o + \beta_1 x_i}}{1 + (e^{\beta_o + \beta_1 x_i})} \cdot \frac{1 + (e^{\beta_o + \beta_1 x_i})}{1} \right] \\ &= \ln [e^{\beta_o + \beta_1 x_i}] = \beta_o + \beta_1 x_i. \quad (10) \end{aligned}$$

Com isso, em vez de assumir que a relação entre π_i e x_i é linear, assume-se que a relação entre $\ln[\pi(x_i)/(1 - \pi(x_i))]$ e x_i é linear. Essa técnica é conhecida como Regressão Logística (CORRAR *et. al.*, 2007).

Portanto, dada uma série de eventos binários, em que as variáveis aleatórias independentes $Y_i, i = 1, \dots, n$, seguem uma distribuição Bernoulli com probabilidade de

sucesso (π_i), ou seja, $Y_i \sim Ber(\pi_i)$ e, ainda, $x_i^T = (1, x_i)$ é a i -ésima linha da matriz X , sendo que $i = 1, \dots, n$. A probabilidade de sucesso do modelo de regressão logístico simples é determinada por (FIGUEIRA, 2006):

$$\pi_i = \pi(x_i) = P(Y_i = 1 | X = x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + (e^{\beta_0 + \beta_1 x_i})}, \quad (11)$$

e a probabilidade de fracasso:

$$1 - \pi_i = 1 - \pi(x_i) = P(Y_i = 0 | X = x_i) = \frac{1}{1 + (e^{\beta_0 + \beta_1 x_i})} \quad (12)$$

em que $\beta = (\beta_0, \beta_1)^T$ é o vetor de parâmetros.

No caso em que a variável independente também é dicotômica, ou seja, $x = 0$ ou $x = 1$ então, a chance de resposta quando $x = 1$ é dada por $\pi(1)/(1 - \pi(1))$ e quando $x = 0$ é dada por $\pi(0)/(1 - \pi(0))$. O logaritmo da chance é dado por:

$$g(1) = \ln \left[\frac{\pi(1)}{1 - \pi(1)} \right] \quad (13)$$

$$g(0) = \ln \left[\frac{\pi(0)}{1 - \pi(0)} \right] \quad (14)$$

Na Tabela 1 é apresentado o resumo das probabilidades de sucesso e fracasso referente à situação quando a variável independente é dicotômica.

Tabela 1 - Resumo das probabilidades de sucesso e fracasso referente à situação quando a variável independente é dicotômica.

Variável resposta (Y)	Variável independente (X)	
	$x=1$	$x=0$
$Y=1$	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + (e^{\beta_0 + \beta_1})}$	$\pi(0) = \frac{e^{\beta_0}}{1 + (e^{\beta_0})}$
$Y=0$	$1 - \pi(1) = \frac{1}{1 + (e^{\beta_0 + \beta_1})}$	$1 - \pi(0) = \frac{1}{1 + (e^{\beta_0})}$
Total	1,0	1,0

Dessa forma, pode-se definir razão das chances ("Odds ratio"), cuja finalidade é identificar a probabilidade associada à ocorrência de determinado evento, como (CORRAR *et. al.*, 2007):

$$\psi = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+(e^{\beta_0+\beta_1})}}{\frac{e^{\beta_0}}{1+(e^{\beta_0})}} \bigg/ \frac{\frac{1}{1+(e^{\beta_0+\beta_1})}}{\frac{1}{1+(e^{\beta_0})}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \quad (15)$$

Então, o logaritmo da expressão (15) é dado por:

$$\ln(\psi) = \ln(e^{\beta_1}) = \beta_1 \quad (16)$$

2.1.2 Regressão Logística Múltipla

O modelo de regressão logístico pode ser generalizado para o caso em que há mais de uma variável preditora.

Dado um conjunto com p variáveis independentes, em que o vetor da i -ésima linha da matriz (X) das variáveis explicativas é designado por: $(x_{i0}, x_{i1}, \dots, x_{ip})$, onde cada elemento da matriz corresponde ao ij -ésimo componente (x_{ij}) , sendo que $i = 1, \dots, n$ e $j = 1, \dots, p$, com $(x_{i0}) = 1$. Tem-se que o vetor de parâmetros desconhecidos é dado por $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ e β_j é o j -ésimo parâmetro associado à variável explicativa x_j . Assim, a probabilidade de sucesso no modelo de regressão múltipla passa a ser expressa por (HOSMER & LEMESOWH, 1989):

$$\pi_i = \pi(x_i) = P(Y_i = 1 | X = x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + \left(e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}} \right)} = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad (17)$$

E a probabilidade de fracasso por:

$$1 - \pi_i = 1 - \pi(x_i) = P(Y_i = 0 | X = x_i) = \frac{1}{1 + \left(e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}} \right)} = \frac{1}{1 + \exp(x_i^T \beta)} \quad (18)$$

Assumindo que $Y_i \sim \text{Ber}(\pi_i)$, a transformação logit agora resulta em:

$$g(x_i) = \ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = x_i^T \beta = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (19)$$

E o logaritmo da função de verossimilhança, para estimar os parâmetros do modelo de regressão logístico múltiplo, passa a ser escrito:

$$l(\beta) = \sum_{i=1}^n [y_i x_i^T \beta - \ln(1 + \exp(x_i^T \beta))] \quad (20)$$

2.1.3 Estimação dos Parâmetros

Para estimar os parâmetros β_0 e β_1 do modelo de regressão logística simples utiliza-se o método da máxima verossimilhança. Os estimadores de máxima verossimilhança de β_0 e β_1 são os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de verossimilhança. Considere uma amostra independente (x_i, y_i) de tamanho n , em que x_i é o valor da variável independente da i -ésima observação, com $i = 1, \dots, n$ e y_i é o valor da variável resposta dicotômica. Sendo que $Y_i \sim \text{Ber}(\pi_i)$, a distribuição de probabilidade de Y_i é dada por (ABREU, 2004):

$$f(y_i, \pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (21)$$

Como as observações y_i são independentes, a função de máxima verossimilhança é dada por;

$$\prod_{i=1}^n f(y_i, \pi_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad y_i \in [0, 1]$$

$$L(\beta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad (22)$$

em que $\beta = (\beta_0, \beta_1)^T$ é o vetor de parâmetros.

Por meio da máxima verossimilhança o que se deseja é determinar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximiza $L(\beta)$. Logo, aplicando o logaritmo na expressão (22) tem-se:

$$l(\beta) = \ln(L(\beta)) = \ln \left(\prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \right)$$

$$= \sum_{i=1}^n [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)]$$

$$= \sum_{i=1}^n \left[y_i \ln \left(\frac{\pi_i}{1 - \pi_i} \right) + \ln(1 - \pi_i) \right] \quad (23)$$

Substituindo as equações (11) e (12) em (23) tem-se:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n \left[y_i (\beta_o + \beta_1 x_i) + \ln \left(\frac{1}{1 + \exp(\beta_o + \beta_1 x_i)} \right) \right] \\ &= \sum_{i=1}^n \left[y_i (\beta_o + \beta_1 x_i) - \ln(1 + \exp(\beta_o + \beta_1 x_i)) \right] \end{aligned} \quad (24)$$

Derivando a expressão (24) em relação a cada parâmetro tem-se:

$$\frac{\partial l}{\partial \beta_o} = \sum_{i=1}^n \left[y_i - \frac{1}{1 + \exp(\beta_o + \beta_1 x_i)} (\beta_o + \beta_1 x_i) \right] \quad (25)$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^n \left[y_i x_i - \frac{1}{1 + \exp(\beta_o + \beta_1 x_i)} (\beta_o + \beta_1 x_i) x_i \right] \quad (26)$$

No entanto, não existe solução analítica para os estimadores $\hat{\beta}_o$ e $\hat{\beta}_1$ dos parâmetros β_o e β_1 (PAGANO & GAUVREAU, 2006). Esse procedimento é feito com recursos computacionais, e o método iterativo utilizado pelo *software* R é o de Newton-Raphson.

O valor ajustado para o i-ésimo valor é dado por:

$$\pi_i = \frac{e^{\hat{\beta}_o + \hat{\beta}_1 x_i}}{1 + (e^{\hat{\beta}_o + \hat{\beta}_1 x_i})} \quad (27)$$

A função resposta ajustada é dada por:

$$\pi = \frac{e^{\hat{\beta}_o + \hat{\beta}_1 x}}{1 + (e^{\hat{\beta}_o + \hat{\beta}_1 x})} \quad (28)$$

Analogamente, obtêm-se os estimadores dos parâmetros para um modelo de regressão logística múltipla.

2.1.4 Adequação do modelo

Após estimar o modelo é necessário verificar se de fato ele é o melhor para representar o que se deseja medir, se ele expressa a realidade e, conseqüentemente, se proporciona uma melhoria nas decisões (CORRAR *et. al.*, 2007).

No presente trabalho foi utilizado o teste de Hosmer e Lemeshow para verificar a adequação do modelo; o critério de informação de Akaike (AIC) para a seleção do

modelo; teste de Wald para avaliar a significância dos parâmetros dos modelos. Portanto, ajustou-se um modelo de regressão logística múltipla selecionando as variáveis que irão compor o modelo por meio do critério de Stepwise. Após cada etapa de incorporação de uma variável, temos uma etapa em que uma das variáveis já selecionadas pode ser descartada. O procedimento chega ao final quando nenhuma variável é incluída ou descartada. A estatística adotada pelo critério de Stepwise para adicionar ou remover uma variável foi o critério de informação de AIC. Para selecionar as variáveis a compor o modelo de regressão logístico foi utilizado o critério de Stepwise. O ideal é encontrar critérios que balanceiam as duas ideias seguintes, de tendência oposta: (i) por um lado, para melhorar o processo de previsão, maior número possível de variáveis é desejado, pela melhoria de ajustamento decorrente; (ii) por outro lado, devido aos eventuais custos de obtenção dos x_i 's e às dificuldades de interpretação e de manuseio de um modelo grande, o número de variáveis deve ser tão pequeno quanto possível (DRAPER & SMITH, 1998).

Para verificar a adequação do modelo, ou seja, bondade do ajuste do modelo utilizou-se o teste de Hosmer e Lemeshow. O teste de Hosmer e Lemeshow consiste em dividir o número de observações em aproximadamente 10 grupos e depois compara as frequências preditas com as observadas. Sendo a hipótese H_0 de que não existem diferenças entre os valores preditos e observados, a um nível de significância α , ou seja, o modelo proposto ajusta-se bem aos dados. Calcula-se as frequências esperadas para $Y=1$, que é a soma das probabilidades estimadas de todos componentes do grupo e para $Y=0$ que é dada por 1 menos a probabilidade do outro grupo (CORRAR *et. al.*, 2007). A estatística do teste segue aproximadamente uma distribuição Qui-quadrado com $g-2$ graus de liberdade dada por:

$$\hat{C} = \sum_{k=1}^g \frac{(O_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)} \underset{aprox.}{\sim} \chi_{(g-2)}^2 \quad (29)$$

sendo:

n_k' o número de indivíduos no k -ésimo grupo;

c_k é o número de combinações de níveis dentro do k -ésimo decil;

$O_k = \sum_{j=1}^{c_k} y_i$ o número total de combinações de níveis dentro do k -ésimo decil;

$$\bar{\pi}_k = \sum_{i=1}^{c_k} \frac{\bar{\pi}_i m_i}{n_i'}$$

O critério de informação de Akaike (AIC) é um método de seleção de modelos. O AIC foi desenvolvido por meio dos estimadores de máxima verossimilhança (EMV), para decidir qual o modelo mais adequado quando se utiliza muitos modelos com quantidades diferentes de coeficientes. A decisão quanto ao melhor modelo ajustado é realizada escolhendo o menor valor de AIC. Sobral & Barreto (2011) define AIC como:

$$AIC = -2l(\beta) + 2k \quad (30)$$

em que $l(\beta)$ é o logaritmo da função de verossimilhança do modelo e k é o número de parâmetros.

A estatística de Wald tem como finalidade aferir o grau de significância de cada coeficiente da equação de regressão logística, inclusive a constante, ou seja, tem por objetivo verificar se cada parâmetro estimado é significativamente diferente de zero. A estatística de Wald segue uma distribuição Qui-quadrado e quando a variável dependente tem um único grau de liberdade pode ser calculada elevando-se ao quadrado a razão entre coeficiente que está sendo testado e o respectivo erro padrão, conforme segue:

$$Wald = \left(\frac{\beta_j}{SE_{\beta_j}} \right)^2 \quad (31)$$

em que β_j é o coeficiente de uma variável independente incluída no modelo; SE é o erro padrão (standard error) (CORRAR *et. al.*, 2007).

Os resultados dos testes são apresentados na seção seguinte. Cabe ressaltar que todas as análises estatísticas foram realizadas no freeware R (R Core Team, 2015).

3 RESULTADOS E DISCUSSÕES

As variáveis independentes são: ADAC, genótipo DD, idade, sexo, sedentarismo, tabagismo, HAS, HFDAC, DM, IMC, LDL, CT, TG e HDL, sendo a variável resposta a DAC. Para selecionar as variáveis a compor o modelo de regressão logístico foi utilizado o critério de Stepwise. Na primeira etapa ajustou-se o modelo de regressão logística com todas as variáveis independentes e retirou-se uma variável de cada vez. O descarte da variável realizou-se por meio do teste de Wald ao nível de 5 % de significância, ou seja, descartou-se a variável cujo parâmetro foi não significativo. Nesse processo foram ajustados 14 (quatorze) modelos. A única variável que permaneceu no modelo foi a variável idade. Os resultados da primeira etapa são apresentados na Tabela 2.

Tabela 2 - Estatística referentes aos modelos ajustados na primeira etapa por meio do critério de Stepwise.

Modelo	AIC	Variável a ser excluída	⁽¹⁾ p-valor	⁽²⁾ HL
1º - Modelo Completo	87,8102	ADAC	0,9939	--
2º	90,1162	CT	0,8493	--
3º	97,4868	DM	0,9406	--
4º	94,4923	LDL	0,8403	--
5º	98,1839	HDL	0,9290	--
6º	96,1918	TG	0,5263	--
7º	229,8005	Sedentarismo	0,4376	--
8º	228,4037	HAS	0,3177	--
9º	227,4022	GDD	0,2439	--
10º	226,7726	Sexo	0,1773	--
11º	226,6105	Tabagismo	0,1453	--
12º	232,4234	HFDAC	0,1538	--
13º	232,4585	IMC	0,0996	--
14º - Modelo Final de 1ª Etapa (Idade)	233,2568	--	0,0013	0,8997

⁽¹⁾: p-valor referente à estatística de Wald; ⁽²⁾: p-valor referente à estatística do teste de Hosmer e Lemeshow.

Ao final dessa etapa o modelo estimado foi composto apenas pela variável idade, a única variável significativa ao nível de 5% pelo teste de Wald, cujo valor do AIC é 233,26. Embora não seja o menor valor AIC observado, optou-se por esse modelo, pois a variável que o compõe foi significativa. Além disso, ao nível de significância de 5%, pelo teste de Hosmer e Lemeshow, não rejeita-se a hipótese H_0 de que não existem diferenças entre os valores preditos e observados, ou seja, o modelo da expressão (32) estimado na primeira etapa ajusta-se bem aos dados.

O modelo estimado na primeira etapa é:

$$\ln \left[\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right] = -2,8894 + 0,0543Idade \quad (32)$$

Na segunda etapa será acrescentada ao modelo da expressão (32), uma variável por vez, cujo propósito é verificar se existe outra variável significativa na presença da variável idade. Caso exista mais de uma variável significativa nessa etapa, a escolha será pelo modelo em que as duas variáveis são significativas e de menor AIC, desde que o modelo seja adequado. Na Tabela 3 são apresentados os resultados dos modelos ajustados na segunda etapa para a seleção da variável que será acrescentada ao modelo da expressão (32).

Tabela 3 - Estatística referentes aos modelos ajustados na segunda etapa por meio do critério de Stepwise.

Modelo	AIC	Variável acrescentada	⁽¹⁾ p-valor	⁽²⁾ HL
1º	224,8256	ADAC	Intercepto Idade ADAC 0,0107 0,0049 0,0108	0,9069
2º	195,7481	LDL - n.s.	0,2405	--
3º	97,3063	TG - nada significativo	0,6620	--
4º	209,0737	HDL - n.s.	0,7740	--
5º	232,4585	IMC - n.s.	0,0996	--
6º	94,0327	CT - nada significativo	0,6440	--
7º	234,7069	DM - n.s.	0,4612	--
8º	233,8688	Sedentarismo - n.s.	0,3649	--
9º	233,9578	GDD - n.s.	0,2566	--
10º	234,5295	HAS - n.s.	0,3929	--
11º	233,3848	Sexo - n.s.	0,1728	--
12º	227,7873	Tabagismo - n.s.	0,2180	--
13º	232,8597	HFDAC - n.s.	0,1218	--

⁽¹⁾: p-valor referente à estatística de Wald; ⁽²⁾: p-valor referente à estatística do teste de Hosmer e Lemeshow; n.s.: apenas a variável discriminada foi não significativa; nada significativo: nenhum parâmetro significativo (intercepto, idade e a variável).

Após esse processo, observa-se na Tabela 3 que a única variável que será incorporada ao modelo da expressão (32) é ADAC, pois seu p-valor é significativo ao nível de 5%. Então, o novo modelo ajustado é:

$$\ln \left[\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right] = -2,7077 + 0,04846Idade + 1,9537ADAC \quad (33)$$

Na terceira etapa será acrescentada ao modelo da expressão (33), uma variável por vez, cujo propósito é verificar se existe outra variável significativa na presença das

variáveis idade e ADAC. Caso exista mais de uma variável significativa nessa etapa, a escolha será pelo modelo em que as três variáveis são significativas e de menor AIC, desde que o modelo seja adequado. Na Tabela 4 são apresentados os resultados dos modelos ajustados na terceira etapa para a seleção da variável que será acrescentada ao modelo da expressão (33).

Tabela 4 - Estatística referente ao modelo final ajustado na terceira etapa do critério de Stepwise.

Modelo	AIC	Variável Acrescentada	⁽¹⁾ p-valor
1º	185,1815	LDL – n.s.	0,3582
2º	92,5339	TG – nada significativo	0,8700
3º	197,6497	HDL – n.s.	0,9082
4º	225,7441	IMC – n.s.	0,3016
5º	89,6374	CT – nada significativo	0,9020
6º	226,8180	DM – n.s.	0,9308
7º	226,2889	Sedentarismo – n.s.	0,4641
8º	225,1908	GDD – n.s.	0,2036
9º	226,5624	HAS – n.s.	0,6076
10º	224,5743	Sexo – n.s.	0,1354
11º	220,6815	Tabagismo – n.s.	0,2431
12º	225,4305	HFDAC – n.s.	0,2379

⁽¹⁾: p-valor referente à estatística de Wald; n.s.: apenas a variável discriminada foi não significativa; nada significativo: nenhum parâmetro significativo (intercepto, idade, ADAC e a variável).

Ao final dessa etapa é possível verificar que nenhuma outra variável foi incorporada ao modelo expresso em (33), pois todas as variáveis foram não significativas pelo teste de Wald, ao nível de significância de 0,05. Portanto, o modelo final será o modelo da expressão (33). Os resultados sobre o ajuste do modelo final, com as estimativas de máxima verossimilhança, erro padrão, estatística de Wald, e a razão de chances, AIC e teste de Hosmer e Lemeshow estão apresentados na Tabela 5.

Tabela 5 - Estatística sobre o modelo de regressão logística múltipla estimado.

Parâmetros	Estimativa	Erro Padrão	Wald	ORR	⁽¹⁾ p-valor	AIC	⁽²⁾ HL
Intercepto	-2,7077	1,0607	6,5161		0,0107	224,8256	0,9069
Idade	0,0485	0,0172	7,9288	1,0496	0,0049		
ADAC	1,9537	0,7668	6,4923	7,0548	0,0108		

⁽¹⁾: p-valor referente à estatística de Wald; ⁽²⁾: p-valor referente à estatística do teste de Hosmer e Lemeshow.

Com base na Tabela 5, ao nível de significância de 5%, há evidências de que todas as variáveis contribuem de forma significativa para o modelo. No conjunto de variáveis, a que alcançou maior nível de significância foi a idade. Além disso, a variável idade exerce um efeito positivo, ou seja, quanto mais elevada a idade, maior será a probabilidade de que ela venha desenvolver a doença. A variável ADAC também exerce

um efeito positivo, pois a estimativa do parâmetro associado a variável em questão é positivo. Conclui-se que o paciente que tiver algum parente de 1º grau com doença arterial coronariana, as chances de se ter a doença é 7,0548 vezes mais provável quando comparado com uma pessoa que não tem algum parente de 1º com doença arterial coronariana.

Observa-se também que o modelo (33) apresenta menor AIC (224,8256) quando comparado com o modelo (32) (AIC = 233,26), e ao nível de significância de 5%, pelo teste de Hosmer e Lemeshow, não rejeita-se a hipótese H_0 de que não existem diferenças entre os valores preditos e observados, ou seja, o modelo (33) ajusta-se bem aos dados.

Assim, substituindo os coeficientes da equação do modelo de regressão logística, pelos coeficientes estimados e apresentados na Tabela 5, tem-se que a probabilidade de o paciente ter DAC é dada por:

$$\hat{\pi}_i = \frac{\exp(-2,7077 + 0,04846x_1 + 1,9537x_2)}{1 + \exp(-2,7077 + 0,04846x_1 + 1,9537x_2)} \quad (34)$$

E a probabilidade de o paciente não ter DAC é dada por:

$$1 - \hat{\pi}_i = \frac{1}{1 + \exp(-2,7077 + 0,04846x_1 + 1,9537x_2)} \quad (35)$$

em que x_1 é a idade e x_2 é ADAC.

Supondo que uma pessoa tenha 40 anos e histórico familiar da DAC, a probabilidade de ela desenvolver a doença arterial coronariana é de 76,58%.

Embora os dados analisados não tenham apresentado fatores de risco modificáveis, cabe ressaltar a importância de prevenção desses fatores para evitar o desenvolvimento da doença arterial coronariana e, também, de outras doenças.

4 CONCLUSÃO

Neste trabalho foi possível estudar a relação da doença arterial coronariana e vários fatores de risco da DAC através da regressão logística. No modelo ajustado apenas as variáveis idade e antecedentes da DAC foram significativas. Esses fatores de risco são não modificáveis, ou seja, não podem ser alterados. Não se observou a influência de valores modificáveis como sedentarismo, tabagismo, IMC, CT, HDL, LDL, TG, DM e HAS e dos fatores não modificáveis GDD, sexo e HFDAC.

É importante salientar que, mesmo não sendo significativos no presente trabalho, os fatores de risco modificáveis merecem atenção especial e devem ser prevenidos, pois associados ou não, podem contribuir para o desenvolvimento da DAC. O estudo Framingham Heart Study (DAWBER *et. al.*, 1950) reconhecido no mundo todo como uma verdade científica e outros estudos apontam fatores como a hipertensão arterial, taxas de colesterol elevadas, obesidade, diabetes, sedentarismo e tabagismo, como fatores de risco para a DAC.

REFERÊNCIAS

ABREU, H. J. *Aplicação da análise de sobrevivência em um problema de Credit Scoring e comparação com a regressão logística*. 2004 118 p. Dissertação (Mestrado) – Universidade Federal de São Carlos, São Carlos, 2004.

ARAÚJO, M. C. *Efeitos do exercício físico regular e suplementação de licopeno sobre marcadores de estresse oxidativo na doença arterial coronariana*. 2008. 78 f. Dissertação (Mestrado em Ciências da Saúde) – Universidade do Extremo Sul Catarinense, Criciúma, 2008.

CORRAR, S. L.; PAULO, E.; DIAS FILHO, J.M. (Coord.). **Análise multivariada para cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2007, 539p.

DAWBERT, T. R.; MEADORS, G. F.; MOORE JR., F. E. Epidemiological Approaches to Heart Disease: The Framingham Study. National Heart Institute, National Institutes of Health, Public Health Service, Federal Security Agency, Washington, DC. *Epidemiology, Session of the Epidemiology, Health Officers, Medical Care, and Statistics Sections of the American People Health Association, at the Seventy-eight Annual Meeting*, St. Louis, Missouri, 03 de nov. 1950.

DRAPER, N.R.; SMITH, H. **Applied regression analysis**. 3.ed. New York: John Wiley e Sons, 1998. 706p.

FIGUEIRA, C. V. *Modelos de regressão logística*. 2006 149 p. Dissertação (Mestrado em Matemática) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

FRANKEN, R. A.; BELLESSO, M.; CAVAZIN, A. M.; POLÔNIO, I. B.; MATTHEUCCI, E.; VARGA, J. Associação do polimorfismo da enzima conversora de angiotensina com dados ecocardiográficos em jovens normotensos filhos de hipertensos. *Revista da associação médica brasileira*. São Paulo, 50(1), 62-7, 2004.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. New York: John Wiley, 1989, 307 p.

MAIA, C. O.; GOLDMEIER, S.; MORAES, M. A.; BOAZ, M. R.; AZZOLIN, K. Fatores de risco modificáveis para doença arterial coronariana nos trabalhadores de enfermagem. *Acta Paulista de Enfermagem*. São Paulo, v. 20, n. 2, abr/jun, 2007.

MOURA, M. M. *Reflexos cardiovasculares em camundongos com alteração na expressão do receptor da Angiotensina-(1-7)*, MAS. 2007. 213 f. Tese (Doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Belo Horizonte, 2007.

PAGANO, M.; GAUVREAU, K. **Princípios de Bioestatística**. 2 ed. São Paulo: Thomson, 2006. 506 p.

PINHO, R. A.; ARAÚJO M. C.; GHISI G. L. M.; BENTTI, M. Doença Arterial Coronariana, Exercício Físico e Estresse Oxidativo. *Arquivos Brasileiros de Cardiologia*. São Paulo, v.94, n. 4, abr/2010.

R Core Team (2015). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

RADAELLI, G. *Relação entre uso de inibidor da enzima conversora de angiotensina e desfechos no pós-operatório de cirurgia de revascularização miocárdica*. 2011. 69 f. Dissertação (Mestrado em Medicina e Ciência da Saúde) - Faculdade de Medicina, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2011.

SOBRAL, T. E. L.; BARRETO, G. Análise dos critérios de informação para a seleção de ordem em modelos auto regressivos. *Conferência brasileira de Dinâmica, Controle e Aplicações*. Águas de Lindóia-SP v.1, n. único, 2011.