UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

Vinícius Aparecido Otálora Pereira

MODELOS LINEARES GENERALIZADOS: ABORDAGENS "CLÁSSICA" E "BAYESIANA" COM APLICAÇÃO NA DOENÇA ARTERIAL CORONARIANA

PRESIDENTE PRUDENTE

2022

Vinícius Aparecido Otálora Pereira

MODELOS LINEARES GENERALIZADOS: ABORDAGENS “CLÁSSICA” E “BAYESIANA” COM APLICAÇÃO NA DOENÇA ARTERIAL CORONARIANA

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/UNESP para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientador: Prof. Dr. Fernando Antonio Moala.

Coorientador: Prof. Dr. Sérgio Minoru Oikawa

PRESIDENTE PRUDENTE

2022

P436m

Pereira, Vinícius Aparecido Otálora

Modelos lineares generalizados: abordagens "clássica" e "bayesiana" com aplicação na doença arterial coronariana / Vinícius Aparecido Otálora Pereira. -- Presidente Prudente, 2022

52 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Fernando Antonio Moala

Coorientador: Sérgio Minoru Oikawa

1. Estatística. 2. Doença arterial coronariana. 3. Modelos lineares generalizados. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Vinícius Aparecido Otálora Pereira

MODELOS LINEARES GENERALIZADOS: ABORDAGENS "CLÁSSICA" E "BAYESIANA" COM APLICAÇÃO NA DOENÇA ARTERIAL CORONARIANA

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Dr. Fernando Antonio Moala

Departamento de Estatística

_____

Prof. Dr. Manoel Ivanildo Silvestre Bezerra

Departamento de Estatística

_____

Prof. Dr. José Gilberto Spasiani Rinaldi

Departamento de Estatística

Presidente Prudente, 23 de março de 2022.

RESUMO

A doença arterial coronariana (DAC) é uma doença cardíaca que ocorre pela barragem nas artérias coronárias, causado geralmente, por placas de gordura. Com o objetivo de identificar fatores que expliquem o evento de um indivíduo vir a doença arterial coronariana em futuros 10 anos, o trabalho apresenta uma aplicação de regressão logística com base na teoria de modelos lineares generalizados (MLG), e utiliza duas abordagens para a estimação dos parâmetros (“clássica” e bayesiana). Na estimação “clássica”, a técnica utilizada é a de máxima verossimilhança com a ajuda do método numérico Newton – Raphson para encontrar as soluções das equações de verossimilhança. Na estimação bayesiana, são utilizadas prioris normais independentes, com médias 0 e variância alta (semelhante ao uso de prioris não informativas). O método de MCMC é utilizado na aplicação como forma de se obter as densidades a posterioris marginais.

Palavras-chaves: Modelos lineares. Análise de regressão logística. Doenças cardíacas. Artérias coronarianas.

ABSTRACT

Coronary heart disease (CHD) is a heart disease that occurs by the damming in the coronary arteries, usually caused by fatty plaques. With the objective of identifying factors that explain the event of an individual coming to coronary artery disease in the future 10 years, the work presents an application of logistic regression based on the theory of generalized linear models (GLM), and uses two approaches for the estimation. parameters ("classical" and Bayesian). In the "classical" estimation, the technique used is the one of maximum likelihood with the help of the Newton – Raphson numerical method to find the solutions of the likelihood equations. In Bayesian estimation, independent normal priors are used, with means 0 and high variance (similar to the use of non-informative priors). The MCMC method is used in the application as a way to obtain marginal posterior densities.

Keywords: Linear models. Logistic regression analysis. Cardiac diseases. Coronary arteries.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O sistema cardiovascular ou sistema circulatório é composto pelo coração e vasos sanguíneos, sua estrutura contempla complexas redes de ligações entre vasos, órgãos e tecidos. Dentre as funções do sistema cardiovascular destacam-se, a circulação do sangue, o transporte de gases, nutrientes e mecanismos de defesa.

No Brasil e no mundo, as doenças cardiovasculares (DCV) estão entre as que mais impactam no número de morbimortalidade e gastos hospitalares, segundo a Sociedade Brasileira de Cardiologia (SBC), as doenças cardiovasculares são responsáveis por mais mortes do que diversos tipos de doenças e infecções, tais como, câncer, doenças respiratórias, AIDS, entre outras. A estimativa do número de óbitos realizada pela SBC para este ano, se aproxima à 400 mil cidadãos.

A doença arterial coronariana (DAC), caracterizada pela deficiência do fluxo sanguíneo no coração distribuído pelas artérias coronárias, é um problema que afeta sobretudo, países desenvolvidos onde as populações possuem faixa etária elevada. Com relação aos óbitos relacionados a doenças cardiovasculares, a mortalidade da DAC pode corresponder a cerca de 80% das mortes. Segundo pesquisas, essa insuficiência sanguínea está diretamente ligada ao estreitamento das artérias (estenose), fomentado por placas ateroscleróticas (placas de gordura) ((PINHO et.al., 2010);(MAIA et.al., 2007)).

Estudos epidemiológicos apontam que os fatores que mais corroboram no desenvolvimento e progressão da doença, denominados fatores de riscos (FR), são variados e podem ser genéticos e / ou obtidos. Hipertensão arterial sistêmica (HAS), tabagismo, dislipidemias, diabetes mellitus (DM), obesidade, sedentarismo, idade e o sexo são alguns dos fatores, com destaque para HAS, responsável por 40% dos óbitos ocasionados pela DAC (MAIA et. al., 2007).

Dentre os processos tomados na prevenção da DAC, está a identificação dos fatores de riscos, pois quando precoce, o reconhecimento dos mesmos, possibilita elaborações de estratégias preventivas eficientes que visem dar qualidade de vida a população. Dados de estudos anteriores indicam que um indivíduo aos 50 anos de idade, sem a manifestação de FR, têm probabilidade próxima à 0,6 de desenvolver um caso coronariano (MAIA et. al.,2007).

A aplicação da teoria de modelos lineares com foco na previsão de características de interesse, ou na explicação de fenômenos baseados em uma ou mais variáveis independentes, são diversas e estão entre as técnicas estatísticas mais utilizadas no momento. Áreas do conhecimento como ecologia, agronomia, zootecnia, economia, medicina, biologia, sociologia e engenharia usufruem da teoria de modelos lineares com grande frequência em pesquisas científicas.

Em modelos lineares, um dos principais procedimentos na análise estatística é a estimação de parâmetros desconhecidos, adotando-se nas abordagens tanto a inferência Clássica quanto a Bayesiana. Na abordagem Clássica, os dois métodos mais utilizados são, máxima verossimilhança e mínimos quadrados. Embora o tratamento Clássico ainda seja o mais empregado, a análise Bayesiana tem ganhado forças com avanços computacionais cada vez maiores.

A principal diferença entre as duas abordagens está na forma de estimação dos parâmetros. Na inferência Clássica a estimação se baseia estritamente na informação fornecida pelo conjunto de dados, necessitando assim de grandes amostras para produzir resultados confiáveis e consistentes. Já na segunda abordagem, os parâmetros desconhecidos do modelo são assumidos como sendo variáveis aleatórias e, neste caso, especifica-se uma distribuição de probabilidade para os mesmos, a partir de conhecimentos já obtidos, estudos anteriores, opiniões de especialista ou qualquer outra fonte relevante. A distribuição escolhida para o vetor de parâmetros é conhecida como distribuição a priori e além de adicioná-la na análise, a inferência Bayesiana tem como objetivo principal, a obtenção de distribuições a posteriori, resultado pela combinação entre distribuição a priori e função de verossimilhança (conjunto de dados).

Anteriormente a década de 90, implementar os métodos Bayesianos em problemas complexos para modelos lineares era totalmente inviável, pois a determinação da distribuição a posteriori sem meios computacionais não era tarefa simples. Somente com o progresso dos computadores mais modernos foi possível a obtenção de distribuições a posteriori de maneira otimizada e com boas aproximações, utilizando-se métodos numéricos como o conhecido MCMC (Método de Monte Carlo via Cadeias de Markov).

Dessa forma, levando em consideração a discussão levantada, o presente trabalho tem como objeto de estudo, identificar fatores que expliquem o evento de um indivíduo vir a desenvolver a doença em futuros 10 anos e estimar seus riscos através da teoria de modelos lineares generalizados, com função de ligação “logit”, utilizando-se abordagens Clássica e Bayesiana.

2 Modelos Lineares Generalizados

O método de mínimos quadrados com distribuição dos erros Normais (curva Gaussiana), desenvolvido por C. F. Gauss, foi estabelecido há muito tempo como “regra”, sendo o alicerce para o desenvolvimento de muitas aplicações em estatística. Por exemplo, os modelos lineares de Gauss-Markov Normal foram a “base” para muitos procedimentos na análise de dados de natureza contínua. Citam-se: Análise de Regressão Linear (Simples e Múltipla), Análise de Variância (ANOVA), Análise de Covariância (ANCOVA), Análise de Dados Longitudinais, etc.

Na época, as propriedades ótimas dessa teoria tornavam conveniente o uso de transformações (Box-Cox) para adaptar dados originalmente coletados em outras escalas de medida (tais como, naturezas discretas, proporções, etc.), e assim, obter um ganho na simplicidade de cálculos. Na pesquisa biológica é possível encontrar vários exemplos de dados "Não-Normais" que, através da metodologia a ser descrita neste trabalho, podem ser tratados "naturalmente" na sua escala original.

A unificação dos procedimentos de modelagem de dados, denominada Modelos Lineares Generalizados (MLG) foi proposta por Nelder e Wedderburn (1972) que, através da Análise de *Deviance*, generalizaram a Análise de Variância para dados com distribuição Normal. E ainda, incorporaram os modelos Log-Lineares para a análise de tabelas de contingência, análise de regressão Logística para dados binários e regressão Poisson para dados de contagens, interligando as análises de dados discretos e contínuos.

Os modelos lineares generalizados são uma extensão dos modelos clássicos de regressão, segundo (CORDEIRO, 2007) nos clássicos tem-se a estrutura,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

onde $\boldsymbol{X\beta}$ é a componente estrutural do modelo, sendo $\boldsymbol{X_{nxp}}$ uma matriz de dimensão n por p (matriz das covariáveis), associada a um vetor de parâmetros desconhecidos, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^T{}_{px1}$ e $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T_{nx1}$ o vetor de erros aleatórios, com $\varepsilon_i \sim N(0, \sigma^2),\ para\ i = 1, \dots, n.$ As relações estabelecidas, implicam que a distribuição do vetor de respostas $\boldsymbol{Y_{nx1}}$ seja normal, $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ e o vetor de médias igual ao preditor linear, dessa forma, $E(\boldsymbol{Y}) = E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta} = \boldsymbol{\mu}$. Essa ligação entre a média e o preditor é o mais simples possível, sendo chamada de função de ligação identidade.

A extensão mencionada é dada em dois seguimentos. Primeiro, o componente aleatório do modelo não necessita aderir a distribuição Normal,

podendo seguir qualquer distribuição contida na família exponencial (Normal, Binomial, Poisson, Gama, entre outras). Segundo, embora a linearidade se mantenha, a função que associa o valor esperado e o vetor de covariáveis, nomeada por função de ligação, suporta ser qualquer função diferenciável (TURKMAN et. al., 2000).

Dessa maneira, os modelos lineares generalizados são compostos na seguinte estrutura:

Componente aleatório: Resumidamente, o componente aleatório é a variável resposta do modelo.

Dado um conjunto de variáveis aleatórias $Y_1, \dots, Y_n$ independentes, onde a distribuição pertencente à família exponencial, com médias $\mu_1, \dots, \mu_n$ e $E(Y_i) = \mu_i \; para \; i = 1, \dots, n$. A função de probabilidade ou densidade de Y_i fica definida como,

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \qquad (2.2)$$

Sendo, θ e ϕ parâmetros desconhecidos e $b(.)$ e $c(.)$ funções conhecidas.

Segundo os autores usados como referência neste trabalho, Turkman (2000), Cordeiro (2001) e Paula (2013), os primeiros momentos de Y podem ser escritos em função de $b(.)$.

$$E(Y) = \mu = \; b'(\theta) \Rightarrow \theta = b'^{(-1)}(\mu), \qquad (2.3)$$

$$Var(Y) = \phi b''(\theta) \Rightarrow \sigma^2 = \phi b''\left(b'^{-1}(\mu)\right) = \phi v(\mu). \qquad (2.4)$$

Os dois resultados, em geral, implicam em uma relação entre a média e a variância, dada por $\sigma^2 = \phi v(\mu)$, onde $v(.)$ é chamada de função de variância.

Componente sistemático: É o preditor linear do modelo, estabelecido pela relação entre as covariáveis e vetor de parâmetros, sendo $\boldsymbol{X}_{nxp}$ a matriz das variáveis explicativas do modelo, $\boldsymbol{\beta}_{px1} = \left(\beta_1, \dots, \beta_p\right)^T$ o vetor de parâmetros desconhecidos e $\boldsymbol{\eta}_{nx1} = (\eta_1, \dots, \eta_n)^T$ o vetor dos preditores.

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{X_i^T\beta} \; ou \; \boldsymbol{\eta = X\beta} \qquad (2.5)$$

Função de ligação: Responsável por relacionar o valor esperado da componente aleatória μ_i e a estrutura linear do modelo η_i.

$$\eta_i = g(\mu_i), \tag{2.6}$$

diferentes funções podem ser escolhidas como função de ligação, no entanto, é interessante que a imagem da função escolhida esteja restrita ao domínio esperado da variável resposta. Em situações onde a função de ligação $g(.)$ transforma o valor esperado do componente aleatório (μ_i), no parâmetro canônico (θ_i), ela é denominada, função de ligação canônica.

$$g(\mu_i) = \theta_i = \eta_i, \tag{2.7}$$

para maior entendimento a tabela abaixo lista algumas funções de ligação.

Tabela 1 - Funções de ligação canônicas

Distribuição	Nome da função	Função
Normal	**Identidade**	$\boldsymbol{\eta = \mu}$
Binomial	**Logit**	$\boldsymbol{\eta = ln(\frac{\mu}{1-\mu})}$
Poisson	**Logarítmica**	$\boldsymbol{\eta = ln(\mu)}$
Gama	**recíproca**	$\boldsymbol{\eta = \mu^{-1}}$

2.1 Exemplos de modelos

A construção de um MLG se dá primeiramente pela natureza do conjunto de dados estudado, antes da escolha de um modelo é de suma importância verificar características como, assimetria, intervalo de variação, média, desvios e outras informações que possam indicar possíveis distribuições para as observações analisadas. Seguem abaixo, dois exemplos de modelos com diferentes distribuições de dados:

2.1.1 Modelos de Regressão para Dados Binários

Quando a variável estudada tem característica binária com probabilidade de sucesso μ, uma possível distribuição para representá-la é a de Bernoulli, ou seja, $Y \sim Ber(\mu)$ e $P(Y = 1) = \mu$, com função de probabilidade igual a,

$$f(y|\mu) = \mu^y (1-\mu)^{1-y}, para\ y = 0,1\ , \tag{2.8}$$

a função de probabilidade (2.8) pertence à família exponencial, portanto pode ser escrita da seguinte forma

$$f(y|\mu) = \exp\left\{y\ log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right\}, \tag{2.9}$$

onde,

$$\theta = \log\left(\frac{\mu}{1-\mu}\right);\ \phi = 1$$

a função de ligação é a logit, e para o modelo de regressão logística, com $Y_i \sim Ber(\mu_i)$ e $\log\left[\frac{\mu_i}{1-\mu_i}\right] = \boldsymbol{x_i'\beta}$,

$$\mu_i = \frac{\exp(\boldsymbol{x_i'\beta})}{1+\exp(\boldsymbol{x_i'\beta})}\ para\ i = 1,2\ , \dots\ , n \tag{2.10}$$

2.1.2 Modelos para Dados de Contagem

Em dados de contagem trabalha-se geralmente com a distribuição de Poisson, $Y \sim Poisson(\lambda)$, cuja a função de probabilidade é,

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\{y\log(\lambda) - [\lambda + \log(y!)]\}, \tag{2.11}$$

a função de ligação é a logarítmica, logo, $Y_i \sim Poisson(\lambda_i)$ e $\log(\lambda_i) = \boldsymbol{x_i'\beta}$, resultando em,

$$\lambda_i = \exp(\boldsymbol{x_i'\beta})\ , para\ i = 1, \dots, n. \tag{2.12}$$

3 Estimação dos parâmetros do MLG

Seja um MLG conforme o definido no capítulo 2 na qual as observações são estabelecidas em um vetor $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$, o logaritmo da verossimilhança em função do vetor de parâmetros $\boldsymbol{\beta}$ (considerando o parâmetro de dispersão ϕ conhecido), é representado pela equação:

$$l(\boldsymbol{\beta}, \phi) = \frac{1}{\phi} \sum_{i=1}^{n} [y_i \theta_i - \mathrm{b}(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi), \tag{3.1}$$

o vetor escore é obtido através da derivada de primeira ordem da função de verossimilhança (ou log-verossimilhança) em relação ao vetor de parâmetros desejado na estimação.

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[\frac{1}{\phi} \sum_{i=1}^{n} [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi)\right] \\ &= \frac{1}{\phi} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} [y_i \theta_i - b(\theta_i)]\right) + 0 \\ &= \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} [y_i - b'(\theta_i)\,] \end{aligned} \tag{3.2}$$

O estimador de máxima verossimilhança (EMV) $\widehat{\boldsymbol{\beta}}$ para o vetor de parâmetros $\boldsymbol{\beta}$ é obtido encontrando-se as raízes da equação (3.2). Em geral as equações resultantes em MLG não são lineares e por tanto, não comportam solução analítica. Dessa forma, aplicando o método de Newton multivariado (similar ao caso univariado) na solução da equação $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\boldsymbol{\partial l(\beta)}}{\boldsymbol{\partial \beta}} = \mathbf{0}$,

$$\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k+1})} = \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})} + \{(-\boldsymbol{U}')^{(-1)}\}^{(\boldsymbol{k})} \boldsymbol{U}^{(\boldsymbol{k})}, \tag{3.3}$$

sendo $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})}$ e $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k+1})}$os vetores de parâmetros estimados em suas respectivas k - ésima e (k+1) - ésima iterações, $\boldsymbol{U}^{(\boldsymbol{k})}$ o vetor escore, e $\{(-\boldsymbol{U}')^{(-1)}\}^{(\boldsymbol{k})}$a inversa da negativa da matriz de segundas derivadas de $l(\boldsymbol{\beta})$, na qual, seus elementos são oferecidos por $-\partial^2 l(\boldsymbol{\beta})/\partial\beta_s \partial\beta_t$ (CORDEIRO,2007).

Segundo Paula (2013), como a matriz $-\boldsymbol{U}'(\boldsymbol{\beta})$ pode não ser positiva definida, a aplicação do método escore de Fisher, que consiste em substituir a

matriz de informação observada, $-\boldsymbol{U}'(\boldsymbol{\beta})$, pela matriz de informação esperada de Fisher, $\boldsymbol{K}(\boldsymbol{\beta})$, pode ser mais conveniente, resultando no seguinte processo iterativo.

$$\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k+1})} = \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})} + \left\{\boldsymbol{K}^{(-1)}\right\}^{(\boldsymbol{k})}\boldsymbol{U}^{(\boldsymbol{k})}, \tag{3.4}$$

onde $\boldsymbol{K}$ possui elementos dados por

$$K_{jk} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right] = E\left[\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j}\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_k}\right]$$

$$K_{jk} = -\frac{1}{\phi}\sum_{i=1}^{n}\frac{1}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ij}x_{ik} = -\frac{1}{\phi}\sum_{i=1}^{n} w_i x_{ij}x_{ik}, \tag{3.5}$$

com $w_i = \frac{1}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$

Trabalhando Matricialmente,

$$\boldsymbol{K}(\boldsymbol{\beta}) = E\left\{-\frac{\boldsymbol{\partial^2 L(\beta)}}{\boldsymbol{\partial\beta\partial\beta^T}}\right\} = \phi^{-1}\boldsymbol{X^T W X}, \tag{3.6}$$

sendo $\boldsymbol{W}$, uma matriz diagonal de pesos, $\boldsymbol{W} = diag\{w_1, \dots, w_n\}$

Para maior entendimento dos resultados (3.5) e (3.6), os materiais de Paula (2013) e Cordeiro (2007) podem ser consultados.

Segundo Demétrio (2001), o método usual para iniciar o processo iterativo é especificar um chute inicial para $\boldsymbol{\beta}^{(\boldsymbol{0})}$ e sucessivamente alterá-lo até que a convergência seja atingida, portanto, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\boldsymbol{m+1})}$.

Diferentes critérios são capazes de verificar a convergência e pausar o processo iterativo, sendo um deles,

$$\sum_{r=1}^{p}\left(\frac{\hat{\beta}_r^{\,(k+1)} - \hat{\beta}_r^{\,(k)}}{\hat{\beta}_r^{\,(k)}}\right)^2 < \varepsilon, \tag{3.7}$$

em que ε é um número positivo suficientemente pequeno, $\varepsilon > 0$ (CORDEIRO,2007).

3.1 Estimação do parâmetro de dispersão ϕ

Na ocasião do parâmetro de dispersão não ser conhecido, se faz necessário uma estimação, podendo ser feita por máxima verossimilhança com solução da equação $\frac{\partial \boldsymbol{U}(\boldsymbol{\phi},\boldsymbol{\beta})}{\partial \boldsymbol{\phi}} = 0$. Demétrio (2001), também cita a estimação de ϕ com base na estatística X^2 de Pearson.

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\mu_i)} \sim \chi^2_{n-p}$$

$$\hat{\phi} = \frac{\mathrm{X}^2}{n-p}, \tag{3.8}$$

4 Inferência para o MLG

4.1 Distribuição assintótica de $\widehat{\boldsymbol{\beta}}$

Excluindo o caso particular de modelos lineares onde a variável estudada segue distribuição Normal, a determinação de distribuições exatas para os estimadores e estatísticas utilizadas no ajuste de modelos, em geral, é uma tarefa de alto grau de complexidade, portanto, ferramentas como o teorema central do limite e outras propriedades de inferência estatística, são utilizadas na busca de resultados assintóticos para essas distribuições.

Em situações onde a amostra é suficientemente grande, a distribuição assintótica do vetor de parâmetros $\widehat{\boldsymbol{\beta}}$ é normal p-variada, ou seja, $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N_p}(\boldsymbol{\beta}, \boldsymbol{K^{-1}})$, onde $\boldsymbol{K}$ é a matriz de informação de Fisher definida em (3.6), $\boldsymbol{K^{-1}}$a sua inversa e $\widehat{\boldsymbol{K}}^{-1} = \phi\left(\boldsymbol{X^T \widehat{W} X}\right)^{-1}$ uma estimativa consistente (CORDEIRO,2007).

Segundo Demétrio (2001), os erros padrões dos estimadores $\hat{\beta}_1, \ldots, \hat{\beta}_p$ são iguais às raízes quadradas dos elementos da matriz $\widehat{\boldsymbol{K}}^{-1}$,isto é, $s(\hat{\beta}_j) = k_{jj}$. Assim, os intervalos de confiança assintóticos com $(1-\alpha)\%$ de confiança são dados por

$$IC(\beta_j; 1-\alpha) = \hat{\beta}_j \pm z_{\alpha/2}\sqrt{s(\hat{\beta}_j)}\,, \tag{4.1}$$

a correlação entre $\hat{\beta}_j$ e $\hat{\beta}_k$ também pode ser obtida através da matriz $\widehat{\boldsymbol{K}}^{-1}$

$$\rho_{jk} = \frac{C\hat{o}v(\hat{\beta}_j, \hat{\beta}_k)}{\sqrt{V\hat{a}r(\hat{\beta}_j)}\sqrt{V\hat{a}r(\hat{\beta}_k)}} = \frac{k_{jk}}{k_{jj}k_{kk}}\,, \tag{4.2}$$

4.2 Função de Deviance

Um modelo linear bem elaborado, deve sempre prezar pelo equilíbrio, entre qualidade de ajuste, gasto computacional e boa interpretabilidade (das estimativas e métricas de qualidade). O equilíbrio mencionado, possui relação com a quantia de variáveis utilizadas no modelo, pois quando se trabalha com um número baixo de variáveis explicativas, geralmente o modelo se ajusta com dificuldade aos dados, equitativamente problemas podem ocorrer ao se construir um modelo exagerado em variáveis explanatórias.

Levando em consideração a quantidade de variáveis incluídas em um modelo, três importantes classificações de modelos são feitas, sendo elas:

- Modelo nulo: contém apenas o intercepto, é o modelo mais simples que se pode fazer, ele atribui a média a todas observações;

- Modelo saturado: contém tantos parâmetros quanto observações;

- Modelo proposto: Qualquer modelo desenvolvido intermediário entre nulo e saturado.

Nelder e Wedderburn (1972) introduziram a análise de deviance com o intuito de mensurar a diferença de ajuste entre um modelo proposto com p variáveis e um modelo saturado com n (com $p < n$), através de uma função desvio, concedida por:

$$S(y, \hat{\mu}) = 2[\hat{l}_p - \hat{l}_n]$$

onde $\hat{l}_n$ e $\hat{l}_p$ equivalem aos logaritmos das verossimilhanças maximizados nos modelos saturado e proposto. Operando os logaritmos $\hat{l}_n$ e $\hat{l}_p$ tem-se então

$$S(y, \hat{\mu}) = \phi^{-1} D(y, \hat{\mu}) = 2\phi^{-1} \sum_{i=1}^{n} [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)], \qquad (4.3)$$

sendo $\tilde{\theta}_i$ e $\hat{\theta}_i$ as estimativas do parâmetro θ_i com base nos modelos saturado e proposto, $S(y, \hat{\mu})$ e $D(y, \hat{\mu})$ são as funções de desvio (CORDEIRO,2007).

A deviance é sempre maior ou igual a zero e conforme se aumenta a quantidade de covariáveis no componente sistemático, decresce até chegar a zero para o modelo saturado (DEMÉTRIO, 2001). Modelos que indicam bons ajustes, resultam em menores valores para o desvio, como já comentado, uma forma de diminuir a deviance é implementar mais variáveis independentes no componente estrutural do modelo, porém, o aumento abusivo de variáveis pode torná-lo extremamente complexo ou até não informativo, tendendo ao saturado. No caso de um modelo com poucas variáveis explicativas e baixa deviance, é possível concluir que a estrutura construída se ajusta bem aos dados, entretanto, na prática, a procura é por um modelo de complexidade e desvio moderado.

Os graus de liberdade para o teste de adequação do modelo alicerçado no desvio, é fomentado pela diferença entre o tamanho da amostra n e o posto da matriz de covariáveis $\boldsymbol{X}$, resultando em $(n-p)$ graus de liberdade. Assim, o valor obtido de $S(y,\hat{\mu})$ é comparado com o percentil da distribuição adequada. Os resultados assintóticos para a distribuição de $S(y,\hat{\mu})$ quando possíveis, convergem em uma qui-quadrado, ou seja, $S(y,\hat{\mu}) \sim \chi^2_{n-p}$, com caso particular na distribuição normal, em que χ^2_{n-p} é a distribuição exata do teste.

4.3 Análise de Deviance

Assim como nos modelos clássicos de regressão onde existe a análise de variância (ANOVA), que visa comparar modelos distintos (dois ou mais) e verificar os efeitos de fatores e suas interações, a análise de deviance (ANODEV) existe para os modelos lineares generalizados.

Uma importante observação a ser feita é que os modelos comparados nas duas análises (ANOVA e ANODEV), são diferidos pela quantidade de parâmetros, onde o último modelo avaliado possui um maior número, e todos os outros que o antecedem são casos particulares do mesmo, ou seja, sofrem uma redução gradativa no número de parâmetros.

Segundo Demétrio (2001), sejam dois modelos M_p e M_q com p e q parâmetros $(p<q)$. Considerando ϕ conhecido, a estatística $D_p - D_q$ com $(q-p)$ graus de liberdade, pode ser interpretada como a discrepância dos dados, explicada pelos componentes que estão em M_q e não estão em M_p, com distribuição assintótica χ^2

$$S_p - S_q = \frac{1}{\phi}\left(D_p - D_q\right) \sim \chi^2_{q-p}, \tag{4.4}$$

essa estatística é a mesma utilizada no teste de razão de verossimilhança (explicada nas próximas secções). No caso de ϕ ser desconhecido, deve-se obter uma estimativa consistente, sendo (3.8) uma das possíveis estimativas. Dessa forma, a comparação dos modelos baseia-se na estatística,

$$F = \frac{(D_p - D_q)/(q-p)}{\hat{\phi}} \sim F_{q-p,n-m}, \tag{4.5}$$

4.4 Teste de Hipóteses

Os métodos de inferência nos modelos lineares generalizados, sustentam-se na teoria de máxima verossimilhança. De acordo com essa teoria, existem três estatísticas para testar hipóteses sobre os parâmetros $\beta's$, que são deduzidas de distribuições assintóticas de funções adequadas das estimativas dos $\beta's$ (DEMÉTRIO, 2001).

Com apoio no interesse em testar as hipóteses:

$$\begin{cases} H_0: \boldsymbol{\beta} = \boldsymbol{\beta_0} \\ H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta_0}, \end{cases} \tag{4.6}$$

Nas hipóteses formuladas em (4.6), $\boldsymbol{\beta}$ representa o vetor de parâmetros do modelo ajustado, e pode conter um ou mais elementos na hipótese testada, pois nem sempre se têm vontade de testar todos os parâmetros do modelo de uma única vez, vale ressaltar que os valores de $\boldsymbol{\beta_0}$ são fixados de acordo com o objetivo do teste. Para maior entendimento, seguem três exemplos de possíveis hipóteses a serem formuladas.

$$\begin{cases} H_0: \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ H_1: \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \end{cases} \quad \begin{cases} H_0: \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ H_1: \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{cases} \quad \begin{cases} H_0: (\beta_2) = (0) \\ H_1: (\beta_2) \neq (0) \end{cases}$$

dessa forma, as três estatísticas são:

a) Razão de verossimilhança:

$$\Lambda = -2 \ln \left[\frac{L(\boldsymbol{\beta_0})}{L(\widehat{\boldsymbol{\beta}})} \right] = 2\left[l(\widehat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta_0}) \right], \tag{4.7}$$

sendo $\hat{\beta}$ o estimador de máxima verossimilhança

Uma maneira muito simples de entender a utilidade do teste de razão de verossimilhança (TRV), é analisar a importância de uma única variável e seu coeficiente β para explicar a característica de interesse do modelo, por exemplo, ao testar o parâmetro β_j, tem-se as seguintes hipóteses e estatística:

$$\begin{cases} H_0: (\beta_j) = (0) \\ H_1: (\beta_j) \neq (0) \end{cases}$$

$$\Lambda = -2 \ln \left[\frac{verossimilhança\ sem\ a\ variável}{verossimilhança\ com\ a\ variável} \right]$$

se a diferença entre as verossimilhanças for pequena, é possível se dizer que o parâmetro β_j é estatisticamente insignificante para explicar a variável resposta do modelo.

b) Wald:

$$W = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_o}\right)^T \boldsymbol{K}\left(\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_o}\right), \tag{4.8}$$

sendo $\boldsymbol{K}\left(\widehat{\boldsymbol{\beta}}\right)$ a matriz de informação de Fisher em função de $\widehat{\boldsymbol{\beta}}$

No caso de um teste de hipóteses uniparamétrico como apresentado no exemplo de entendimento para a razão de verossimilhança, a estatística W assume a forma:

$$W = \frac{(\beta_j - 0)}{S(\beta_j)} \sim N(0,1)$$

tendo, sob H_0 e ϕ conhecido, distribuição assintótica normal padrão.

c) Escore:

$$Es = \boldsymbol{U^T}(\boldsymbol{\beta_0})\ \boldsymbol{K^{-1}}(\boldsymbol{\beta_0})\ \left(\boldsymbol{U}(\boldsymbol{\beta_0})\right), \tag{4.9}$$

Sob a condição de H_0 e sendo ϕ conhecido, as três estatísticas são assintoticamente equivalentes, com distribuição χ_p^2.

5 Seleção de variáveis

Durante o processo de elaboração de um modelo linear, uma importante etapa é a seleção de covariáveis. Um bom modelo, preza pelo equilíbrio entre a explicação do fenômeno de interesse e complexidade das equações trabalhadas. Em ocasiões em que a quantidade de variáveis explicativas à disposição, não é grande, a escolha é mais simples, e pode ser feita por meio da análise de deviance, utilizando testes de hipóteses. Na situação contrária (grandes quantidades de covariáveis), geralmente se faz uso de algoritmos de seleção. Uma observação a ser feita, é que a análise exploratória uni e bivariada nunca deve ser descartada e é de suma importância para o entendimento das variáveis independentes e suas relações com a variável dependente a ser explicada pelo modelo.

5.1 Critério de informação de Akaike

O AIC (Akaike Information Criterion) é uma métrica que quantifica a qualidade do ajuste de um modelo, levando em consideração a quantidade de parâmetros estimados para “penalizar” o resultado.

$$AIC = -2\ln\big(L(\boldsymbol{\beta})\big) + 2p$$

onde $\ln\big(L(\boldsymbol{\beta})\big)$ é o logaritmo da verossimilhança maximizada em $\boldsymbol{\beta}$**.**

A vantagem de usar o AIC como métrica na comparação de possíveis modelos, é que eles podem não ser encaixados, ou seja, ao comparar modelos com diferentes parâmetros, os que possuírem menos não precisam ser uma restrição daquele que tiver mais, e a escolha do modelo se dá pelo menor AIC.

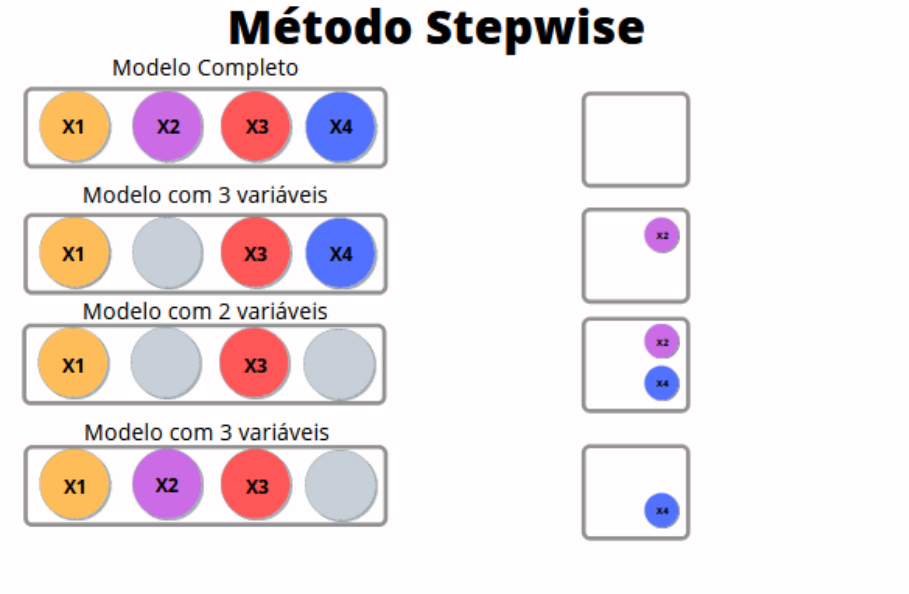
5.2 Algoritmos de seleção

Como já mencionado, em situações onde há muitas covariáveis disponíveis para incluir no modelo, deve-se escolher uma combinação que se ajuste bem aos dados e não seja complexa. Tentar eleger a melhor combinação de variáveis manualmente torna-se totalmente inviável nesses casos, e por isso, algoritmos de seleção como, Forward, Backward e Stepwise são indispensáveis. Os algoritmos de seleção consistem em adicionar e / ou retirar variáveis do modelo seguindo algum critério escolhido, como AIC, razão de verossimilhança ou qualquer outra regra que possa identificar diferenças entre os modelos, procurando assim, o grupo de explanatórias que otimize a relação entre ajuste e complexidade.

5.3 Stepwise

Stepwise é o método que combina os outros dois algoritmos citados (backward e forward). Nesse método, o critério de seleção não se restringe apenas a uma direção (frente ou atrás), o modelo de início é completo, e assim como no backward, o primeiro passo é a remoção da variável menos significativa (a que resultar maior AIC, por exemplo), porém, a cada nova fase, o algoritmo considera a possibilidade tanto de entrada quanto de saída das variáveis.

Figura 1 -Algoritmo de seleção Stepwise



6 Inferência Bayesiana

Como já mencionado anteriormente, a ideia principal da inferência bayesiana, é tratar o parâmetro desconhecido (ex: θ) como uma variável aleatória, e ter como objeto de estudo a distribuição a posteriori $p(\theta|y)$, resultante da combinação entre a priori (distribuição de probabilidade pré-estabelecida para o parâmetro θ, $p(\theta)$) e função de verossimilhança (informação do conjunto de dados).

6.1 Teorema de Bayes

A combinação entre a informação do parâmetro, incrementada de maneira probabilística (priori) e a verossimilhança, é estabelecida pelo Teorema de Bayes,

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(\theta)\, p(y|\theta) d\theta}, \tag{6.1}$$

Nota-se que $p(y)$ não depende de θ, portanto, representa uma constante na determinação da distribuição posteriori.

$$p(\theta|y) = Kp(y|\theta)p(\theta), \tag{6.2}$$

A notação é a mesma tanto para os casos contínuos quanto discretos, porém vale destacar a diferença da constante normalizadora nas duas situações.

$$K^{-1} = p(y) = \begin{cases} \int p(y|\theta)p(\theta) d\theta\,, caso\ contínuo; \\ \sum p(y|\theta)p(\theta), caso\ discreto. \end{cases} \tag{6.3}$$

O fato de $p(y)$ não depender de θ e ser tratada apenas uma constante, levou à uma simplificação das relações estabelecidas,

$$p(\theta|y) \propto p(y|\theta)p(\theta)\,, \tag{6.4}$$

lê-se, que ‘a posteriori é proporcional à priori vezes a verossimilhança’.

6.2 Problema Geral da Inferência Bayesiana

Segundo (Ehlers, 2005), a distribuição a posteriori pode ser resumida em termos de esperanças de funções específicas do parâmetro θ, como segue o exemplo abaixo.

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta$$

Dessa forma, o problema geral da bayesiana consiste em calcular integrais de alta complexidade, e na maioria dos casos essas integrais não possuem solução analítica, por isso o uso de métodos de aproximações numéricas, como MCMC, se tornam indispensáveis dentro da inferência bayesiana.

6.3 Método de Monte Carlo via Cadeias de Markov (MCMC)

Os métodos de Monte Carlo via Cadeias de Markov são vistos como uma boa alternativa aos métodos não iterativos, principalmente para os problemas mais complexos. O MCMC tem como objetivo, simular um passeio aleatório dentro do espaço de θ até que a distribuição estacionária seja encontrada, é nesse momento em que a distribuição a posteriori é estudada (Ehlers, 2005). Existem dois principais algoritmos (não determinísticos) nos métodos de MCMC, Metropolis – Hastings e Amostrador de Gibbs, para o presente trabalho, utilizaremos o Metropolis.

As cadeias de Markov utilizadas no MCMC, nada mais são, resumidamente, do que um processo estocástico $T_0, \dots, T_n$, onde a distribuição de $(T_n|T_{n-1}, \dots, T_0)$ depende apenas de T_{n-1}, ou seja, $P(T_n|T_{n-1}, \dots, T_0) = P(T_n|T_{n-1})$

Com relação as cadeias, os métodos de MCMC exigem que elas sejam, homogêneas, irredutíveis e aperiódicas.

- **Homogênea:** A probabilidade de um estado para outro no processo é invariante;

- **Irredutível:** Todos os estados podem ser alcançados a partir de qualquer outro da cadeia em finitas iterações;

- **Aperiódica:** Sem estados absorventes.

6.4 Algoritmo Metropolis - Hastings

Segundo (Ehlers, 2005), os algoritmos de Metropolis – Hastings usam o mesmo princípio dos métodos de rejeição, ou seja, um valor é gerado a partir de uma distribuição auxiliar e aceito com base em uma probabilidade. Esse mecanismo garante a convergência da cadeia para a distribuição estacionária (distribuição a posteriori no caso).

O algoritmo se inicia com um valor inicial para θ, gerado a partir da distribuição proposta $q(.|\theta)$ (auxiliar). O valor inicial θ' é aceito com probabilidade $\alpha(\theta, \theta')$,

$$\alpha(\theta, \theta') = min\left\{1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right\}, \tag{6.5}$$

O algoritmo pode ser descrito em alguns passos, sendo eles:

1. Inicialize o algoritmo com um chute inicial para θ_0;

2. A partir da distribuição $q(.|\theta)$, gere um novo valor θ';

3. Calcule $\alpha(\theta, \theta')$ e gere um valor aleatório $u \sim U(0,1)$;

4. Se $\alpha \geq u$, aceite o valor gerado θ' e faça $\theta_{t+1} = \theta'$, caso contrário determine $\theta_{t+1} = \theta_t$;

5. Repita o processo a partir do passo 2 e incremente uma unidade no contator, de t para $t + 1$;

6.5 Distribuição a priori e a posteriori

Para os componentes do vetor de parâmetros $\boldsymbol{\beta}$, foram escolhidas prioris normais e independentes com precisões baixas.

$$\beta_j \sim N(0,10^5), j = 1,2,\dots,n.$$

Logo, a distribuição a posteriori conjunta é proporcional ao produto da priori com a função de verossimilhança:

$$p(\beta|y) \propto \prod_{j=1}^{n} exp\left\{\frac{-(\beta_j - 0)^2}{2 \times 10^5}\right\} \times exp\left\{\frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i,\phi)\right\} \quad (6.6)$$

Como $y_i \sim Ber(p_i)$

$$p(\beta|y) \propto \prod_{j=1}^{n} exp\left\{\frac{-(\beta_j - 0)^2}{2 \times 10^5}\right\} \times exp\left\{\sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\log(1-p_i)\right\} \quad (6.7)$$

onde,

$$\theta = \log\left(\frac{p}{1-p}\right);\ p = \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \Rightarrow \log\left(\frac{p}{1-p}\right) = X'\beta$$

6.6 Intervalos de credibilidade

Os intervalos de credibilidade têm como objetivo, propor uma região em que seja mais "provável" que o parâmetro se encontre. Como os parâmetros são tratados como aleatórios e a posteriori é de fato uma distribuição de probabilidade, definir um intervalo para a estimativa pontual do parâmetro, é relativamente simples. Assim, qualquer região R pode ser uma região de credibilidade para θ se

$$\int_R p(\theta|y)d\theta = 1-\alpha, \quad (6.8)$$

A probabilidade de $\theta \subset R$, com base na distribuição a posteriori é $(1-\alpha)$.

7 Aplicação

7.1 Conjunto de dados

A base de dados foi encontrada na plataforma digital Kaggle, tendo sido disponibilizadas por um usuário no meio do ano de 2019. Segundo as informações disponíveis no site, o conjunto de dados refere-se à um estudo cardiovascular que estava em andamento na cidade de Framingham do estado de Massachusetts, sobre a doença arterial coronariana (DAC), com o objetivo principal de identificar fatores que explicassem o evento de um indivíduo vir a desenvolver a doença em futuros 10 anos. A amostra coletada na pesquisa foi composta por 4238 indivíduos com idades entre 30 e 70 anos, onde 15 variáveis foram observadas, listadas a seguir.

y_i: Indica se o indivíduo desenvolveu DAC em 10 anos observados, $(y_i = 0)$ para os que não desenvolveram e $(y_i = 1)$ para os que desenvolveram **(TenYearCHD)**;

x_{i1}: Classifica o gênero biológico da pessoa (0 = Feminino, 1 = Masculino) **(male)**;

x_{i2}: Idade em anos completos **(age)**;

x_{i3}: Especifica se a pessoa é atual fumante ou não (0 = Não, 1 = Fumante) **(currentSmoker)**;

x_{i4}: Quantidade de cigarros consumidos por dia (em média) **(cigsPerDay)**;

x_{i5}: Identifica se o indivíduo faz uso de medicamentos para controle da pressão arterial (0 = não faz uso, 1 = faz uso) **(BPMeds)**;

x_{i6}: Identifica se o indivíduo já teve um caso de AVC ao menos uma vez (0 = Não, 1 = Sim) **(prevalenteStroke)**;

x_{i7}: Identifica se o indivíduo possui ou não diabetes (0 = Sem diabetes, 1 = Com diabetes) **(diabetes)**;

x_{i8}: Quantifica o colesterol total da pessoa **(totChol)**;

x_{i9}: Medição da pressão arterial sistólica **(sysBP)**;

x_{i10}: Medição da pressão arterial diastólica **(diaBP)**;

x_{i11}: Índice de massa corporal do indivíduo, dado pela razão entre massa corporal e altura ao quadrado, $IMC = massa/Altura^2$ **(BMI)**;

x_{i12}: Frequência cardíaca **(heartRate)**;

x_{i13}: Medição dos níveis de glicose **(glucose)**;

x_{i14}: Indica se o indivíduo possui Hipertensão (0 = Não, 1 = Sim) **(prevalentHyp)**.

7.2 Renomeação das variáveis

Levando em consideração que a coleta de dados se deu nos Estados Unidos, entende-se como natural que as variáveis tenham sido nomeadas e descritas em inglês, porém, com intuito de facilitar o entendimento e manejo das mesmas, novos nomes foram dados a cada uma.

Tabela 2 - Variáveis renomeadas

Original	TenYearCHD	male	age	currentSmoker	cigsPerDay
Renomeada	**DAC**	**sexo**	**idade**	**fumante**	**tabagismo**
Original	BPMeds	prevalentStroke	diabetes	totChol	sysBP
Renomeada	**Medicamentos**	**AVC**	**diabetes**	**colesterol**	**sistolica**
Original	diaBP	BMI	heartRate	glucose	prevalentHyp
Renomeada	**diastolica**	**IMC**	**FreqCard**	**glicose**	**hipertenso**

7.3 Imputação de dados

Na base de dados havia observações faltantes, portanto, valores foram inferidos com o objetivo de preencher o vazio e diminuir a falta de informação. No processo de imputação, cada variável recebeu uma "regra de preenchimento" específica, correlacionando uma ou mais variáveis que pudessem explicar as observações vazias. Regras utilizadas:

- **FreqCard**: Mediana;
- **Medicamentos**: Com relação ao uso de medicamentos, só os indivíduos hipertensos que faziam uso, cerca de 10% dos hipertensos utilizavam medicamentos para pressão, assim, o valor inferido para os casos nulos foi 0 (assumiu-se que mesmo sendo hipertenso, seria mais provável que o indivíduo não utilizasse medicamentos);
- **IMC**: Mediana;
- **Glicose**: Mediana, diferindo pessoas diabéticas e não diabética;
- **Tabagismo**: Identificou-se se o indivíduo era fumante ou não, caso não fosse, 0, para os fumantes, as medianas separadas por sexo foram inferidas.
- **Colesterol**: foi ajustado um modelo de regressão, $Colesterol_i = \beta_0 + \beta_1 IMC_i + \beta_2 Sistólica_i + \beta_3 Idade_i + \beta_4 FreqCard_i$

Tabela 3 - Valores faltantes

Variável	**Tabagismo**	**Medicamentos**	**Colesterol**	**FreqCard**	**IMC**	**Glicose**
Valores Nulos	29	53	50	1	19	388

7.4 Agrupamento de variáveis

Dentre as variáveis quantitativas disponíveis no banco de dados, duas foram agrupadas em classes, sendo elas, tabagismo (original: CigsPerday) e IMC (original: BMI). A seguir, as tabelas evidenciam as relações estabelecidas no agrupamento das variáveis.

Tabela 4 - Agrupamento da variável tabagismo

Classificação	**Qtd. de cigarros p/dia**
Não fuma	0
Grau I	0< Cigarros ≤ 10
Grau II	10 < Cigarros ≤ 20
Grau III	<20

Tabela 5 - Agrupamento da variável IMC

Classificação	**IMC**
Abaixo	< 18,5
Ideal	18,5 ≤ IMC < 25
Sobrepeso	25 ≤ IMC < 30
Obesidade	IMC ≥ 30

7.5 Regressão logística

No problema apresentado, a variável resposta possui estrutura binária (desenvolveu ou não desenvolveu a doença em 10 anos), por isso, o modelo escolhido para se ajustar aos dados foi o de regressão logística.

Resgatando o modelo explicado na secção 2.1 do capítulo 2, $Y_i \sim Ber(\mu_i)$ com $\ln\left({}^{\mu_i}/_{1-\mu_i}\right) = \boldsymbol{x_i'\beta}$,

$$\mu_i = \frac{\exp(\boldsymbol{x_i'\beta})}{1 + \exp(\boldsymbol{x_i'\beta})}$$

e

$$1 - \mu_i = \frac{1}{1 + \exp(\boldsymbol{x_i'\beta})}$$

onde, $\mu_i = P(\boldsymbol{Y} = 1|\boldsymbol{X} = \boldsymbol{x_i'})$ e $1 - \mu_i = P(\boldsymbol{Y} = 0|\boldsymbol{X} = \boldsymbol{x_i'})$

7.6 Interpretação dos coeficientes (variáveis explicativas dicotômicas)

Na regressão logística, os parâmetros estimados não são retratados como na regressão linear, justamente pela diferença na função de ligação. Em logística, os $\beta's$ são interpretados com relação às chances de o evento de interesse ocorrer.

A maneira mais simples de se entender a chance de ocorrência, é no contexto de uma variável explicativa dicotômica (exemplo - sexo [(0) feminino e (1) masculino]). A chance de o evento estudado ocorrer no sexo masculino (x=1) é dada por $\left(\mu(1)\middle/1 - \mu(1)\right)$ e $\left(\mu(0)\middle/1 - \mu(0)\right)$ no feminino (x=0). Sendo assim, razão das chances ("Odds ratio"), denotada por Ψ, é calculada pela razão entre as chances de (x=1) e (x=0).

$$\Psi = \frac{\mu(1)/[1 - \mu(1)]}{\mu(0)\ /[1 - \mu(0)]}$$

o logaritmo de "Ψ" fica expresso por:

$$\ln(\Psi) = \ln\left(\frac{\mu(1)/[1 - \mu(1)]}{\mu(0)/[1 - \mu(0)]}\right) = \ln\left(\frac{\mu(1)}{[1 - \mu(1)]}\right) - \ln\left(\frac{\mu(0)}{[1 - \mu(0)]}\right)$$

na situação de uma regressão logística simples, o $\ln(\Psi)$ é dado por,

$$\ln(\Psi) = \ln\left(\frac{\mu(1)}{[1 - \mu(1)]}\right) - \ln\left(\frac{\mu(0)}{[1 - \mu(0)]}\right) = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

dessa maneira, é possível associar Ψ como,

$$\Psi = \exp\{\ln(\Psi)\} = e^{\beta_1}$$

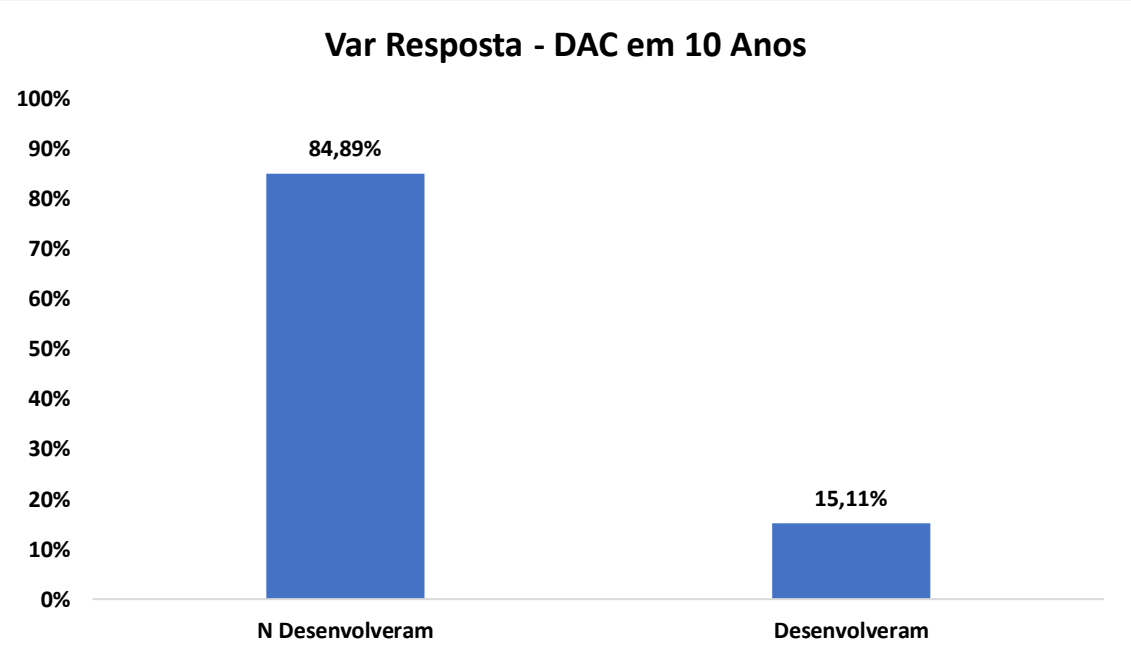
Considerando o mesmo exemplo da variável sexo, se $e^{\beta_1} = 2$, interpreta-se, que a chance de o evento ocorrer no sexo masculino, é duas vezes maior do que no feminino, ou então, pode-se dizer que para o sexo masculino, as chances aumentam em 100%.

7.7 Análise Descritiva

A análise descritiva foi feita a fim de conhecer melhor o conjunto de dados, entender a característica de natural de cada variável, detectar possíveis outliers, buscar relações entre as variáveis independentes e a dependente e observar a existência ou não de correlações entre duas ou mais explicativas.

Alguns outliers foram identificados nos dados, ex: pessoas que consumiam mais de 40 cigarros por dia, níveis de glicose acima de 300 mg/dl, frequência cardíaca acima de 110 bpm, porém, os únicos excluídos, foram as pessoas que tinham os níveis de colesterol total acima de 400 mg/dl (10 observações), resultando em 4228 observações.

Figura 2 - Gráfico de barras da variável resposta



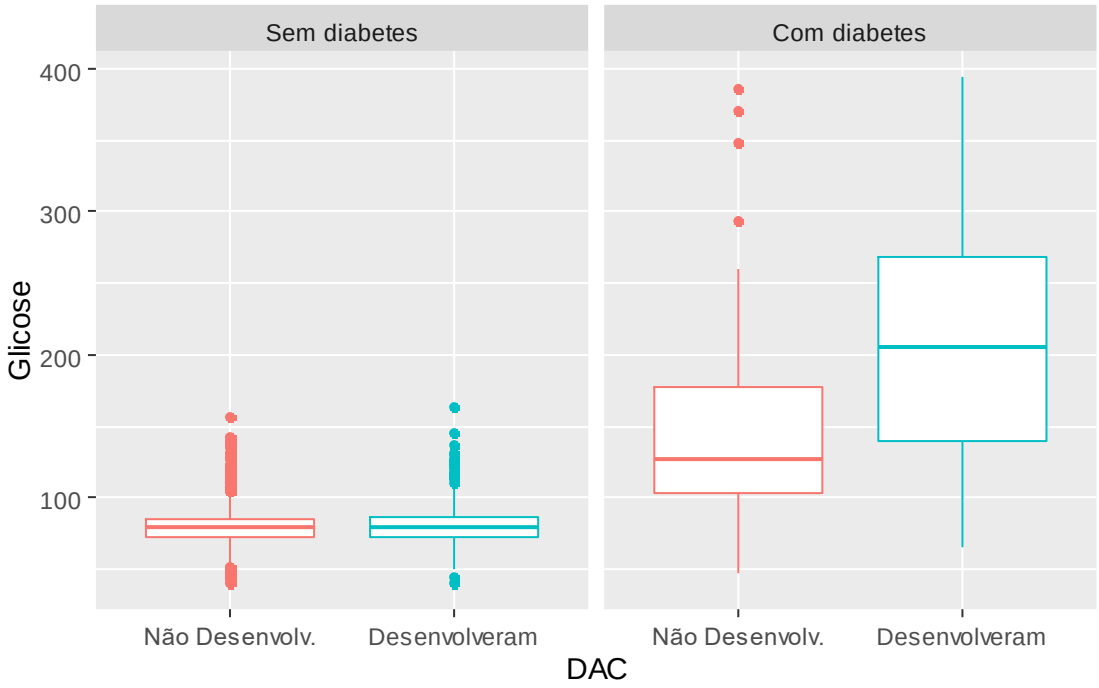
Na figura 4, é apresentado o gráfico de barras, em é possível se observar a distribuição da DAC, na qual, 85% dos indivíduos analisados (3589) não desenvolveram e 15% (639) desenvolveram.

Tabela 6 - Relação entre as variáveis quantitativas e a resposta

Variável	Status	Min	1°Q	Mediana	Média	3°Q	Max	C.V(%)
Idade	Desenvolveram DAC	35	48	**55**	54	61	70	14,8%
	Não Desenvolveram	32	42	**48**	49	55	70	17,3%
Colesterol	Desenvolveram DAC	107	214	**241**	243	271	380	17,8%
	Não Desenvolveram	113	205	**232**	235	261	398	18,1%
Sistólica	Desenvolveram DAC	83,5	125	**139**	144	158	295	18,6%
	Não Desenvolveram	83,5	116	**127**	130	141	243	15,7%
Diastólica	Desenvolveram DAC	48	78	**86**	87	95	140	16,2%
	Não Desenvolveram	50	74	**81**	82	88	143	13,8%
FreqCard	Desenvolveram DAC	50	68	**75**	76	84	120	15,9%
	Não Desenvolveram	44	68	**75**	76	82	143	15,8%
Glicose	Desenvolveram DAC	40	73	**79**	89	89	394	45,0%
	Não Desenvolveram	43	72	**79**	81	85	386	22,5%

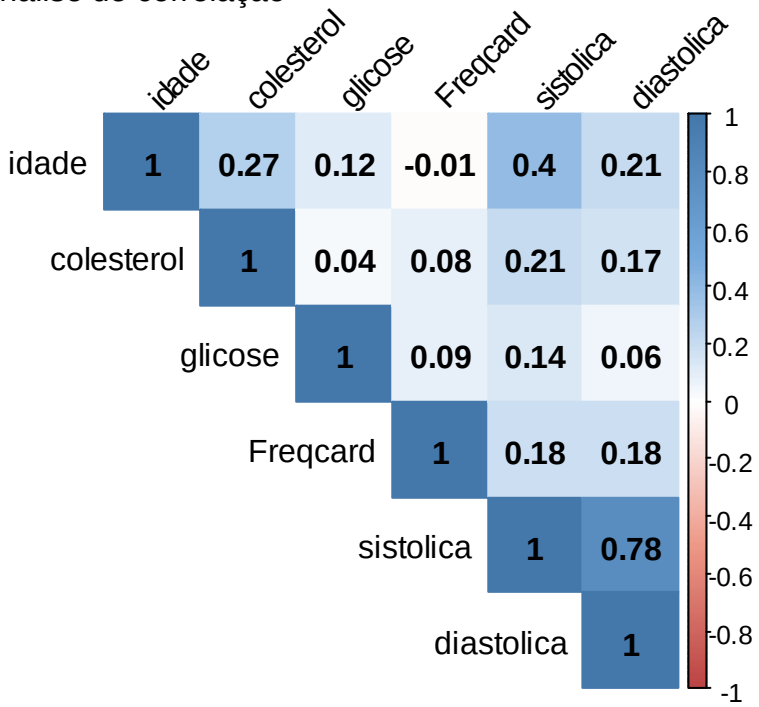
Na tabela 6, destacam-se, idade, colesterol e a pressão arterial sistólica, com medianas aparentemente maiores dentre os que desenvolveram a doença. Vale destacar também o coeficiente de variação alto na glicose, provocado provavelmente pelos indivíduos diabéticos.

Figura 3 - Box-plot da glicose relacionada com a diabetes e a variável resposta



Pode-se observar uma enorme diferença nos níveis de glicose estratificando os indivíduos pela diabetes. Além disso, o box-plot salienta o contraste entre os diabéticos que desenvolveram a doença arterial coronariana e os que não.

Figura 4 - Análise de correlação



Verifica-se que fora o par (sistólica, diastólica; com $\rho = 0{,}78$), nenhuma outra combinação de variáveis apresentou forte correlação.

Tabela 7 - Relação entre as variáveis qualitativas e a resposta

Variável	Categoria	Qtde	Distribuição	Desenvolv. DAC	Percentual
Sexo	Masculino	1814	42,90%	341	18,80%
	Feminino	2414	57,10%	298	12,34%
Fumante	Sim	2088	49,39%	331	15,85%
	Não	2140	50,61%	308	14,39%
Tabagismo	Não Fuma	2140	50,61%	308	14,39%
	Grau I	628	14,85%	73	11,62%
	Grau II	1001	23,68%	161	16,08%
	Grau III	459	10,68%	97	21,13%
Hipertenso	Sim	1309	0,59%	321	24,52%
	Não	2919	99,41%	318	10,89%
AVC	Sim	25	2,55%	11	44,00%
	Não	4203	97,45%	628	14,94%
Diabetes	Sim	108	2,55%	39	36,11%
	Não	4120	97,45%	600	14,04%
IMC	Abaixo	57	1,35%	8	12,06%
	Ideal	1865	44,11%	225	12,06%
	Sobrepeso	539	41,79%	301	17,03%
	Obesidade	1767	12,75%	105	19,48%

Na tabela 7 estão as variáveis qualitativas, e relacionando-as com a DAC verifica-se que:

- Cerca de 19% dos homens vieram a ter o quadro coronariano, indicando uma pré-disposição do sexo masculino quanto ao desenvolvimento da doença;
- O fator fumar ou não, mostrou provável dependência ao grau de tabagismo com relação ao risco de DAC;
- Os percentuais de quadros coronarianos em situações de hipertensão, histórico de AVC e diabetes soam como alarmes para o desenvolvimento da doença, porém, é importante salientar as baixas quantidades de indivíduos com histórico de AVC (25) e diabetes (108).
- Com relação ao IMC, destacam-se os quadros de sobrepeso e obesidade.

Tabela 8 - Uso de medicamentos em hipertensos correlacionados a DAC

Variável	Categoria	Quantidade	Distribuição	DesenvolveramDAC	Percentual
Medicamentos	Sim	136	10,38%	44	32,35%
	Não	1173	89,61%	277	23,61%

A tabela 8 evidencia um maior percentual de hipertensos que desenvolveram a doença sob o uso de medicamentos, muito provavelmente porque os indivíduos que faziam uso de fármacos já se encontravam em quadros mais graves de hipertensão.

8 Aplicação do Modelo

Ajustando um modelo e testando as hipóteses

8.1 Modelo Completo

Antes de propor um modelo específico, o modelo completo (com todas as variáveis) foi testado, a fim de saber se ao menos um parâmetro seria diferente de zero.

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X'\beta}$$

Estrutura do modelo,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 sex(masc) + \beta_2 Idade + \beta_3 fumante(sim) + \beta_4 Tab(GrauI)$$
$$+ \beta_5 Tab(GrauII) + \beta_6 Medicamentos + \beta_7 AVC + \beta_8 Hiperten.$$
$$+\beta_9 Diabetes + \beta_{10} Colesterol + \beta_{11} Sistolica + \beta_{12} Diastolica$$
$$+\beta_{13} IMC(ideal) + \beta_{14} IMC(sobrepeso) + \beta_{15} IMC(obesidade)$$
$$+\beta_{16} Freqcard + \beta_{17} Glicose$$

Hipótese testada,

$$\begin{cases} H_0: \boldsymbol{\beta} = \mathbf{0} \\ H_1: \boldsymbol{\beta} \neq \mathbf{0}' \end{cases}$$

Ou,

$$\begin{cases} H_0: \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{17} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \\ H_1: \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{17} \end{pmatrix} \neq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \end{cases}$$

8.2 Teste de Razão de Verossimilhança

$$\Lambda = -2\ln\left[\frac{L(\boldsymbol{\beta_0})}{L(\widehat{\boldsymbol{\beta}})}\right] = 2\left[l(\widehat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta_0})\right] = Deviance_{mod_{nulo}} - Deviance_{mod_{cheio}}$$

$$= 3591 - 3188{,}1 = 402{,}9 \sim \mathrm{X}^2_{(18-1)}$$

Tabela 9 - Estatística do teste

Estatística	g.l.	χ^2	p-valor
Razão de Verossimilhança $(\Lambda) = 402{,}9$	17	$27{,}58$	$2{,}2\ x\ 10^{-16}$

O baixo valor do p-valor rejeita a hipótese nula, verificando-se assim, a existência de ao menos um parâmetro significativo (diferente de zero).

Na tabela 10 são apresentadas as contribuições de cada variável para o modelo com base na deviance.

Tabela 10 - Contribuição de cada variável

efeito	g.l.	deviance	g.l residual	Dev. residual	p-valor
Nulo	----	-------	4227	3591	--------
Sexo	1	33,284	4226	3557,7	$\mathbf{0,79\ x\ 10^{-8}}$
Idade	1	224,192	4225	3333,5	$\mathbf{0,22\ x\ 10^{-15}}$
Fumante	1	12,617	4224	3320,9	$\mathbf{0,38\ x\ 10^{-3}}$
Tabagismo	2	16,455	4222	3304,5	$\mathbf{0,26\ x\ 10^{-8}}$
Medicamento	1	13,701	4221	3290,8	$\mathbf{0,21\ x\ 10^{-3}}$
AVC	1	5,570	4220	3285,2	$\mathbf{0,18\ x\ 10^{-1}}$
Hipertenso	1	42,805	4219	3242,4	$\mathbf{0,60\ x\ 10^{-10}}$
Diabetes	1	13,871	4218	3228,5	$\mathbf{0,19\ x\ 10^{-3}}$
Colesterol	1	2,336	4217	3226,2	$\mathbf{0,126}$
Sistólica	1	24,379	4216	3201,8	$\mathbf{0,79\ x\ 10^{-6}}$
Diastólica	1	0,248	4215	3201,5	$\mathbf{0,61}$
IMC	3	2,103	4212	3199,4	$\mathbf{0,55}$
FreqCard	1	0,097	4211	3199,3	$\mathbf{0,75}$
Glicose	1	11,298	4210	3188,1	$\mathbf{0,77\ x\ 10^{-3}}$

A maioria das variáveis mostraram-se significativas quanto a contribuição produzida, com exceção das seguintes, colesterol, pressão diastólica, IMC, e frequência cardíaca.

8.3 Modelo Stepwise

Aplicando o algoritmo de seleção stepwise, tem-se o modelo composto por:

- Sexo;
- Idade;
- Tabagismo;
- AVC;

- Hipertenso;
- Sistólica;
- Glicose;

Estrutura do modelo escolhido pelo stepwise,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 sex(masc) + \beta_2 Idade + \beta_3 Tab(GrauI) + \beta_4 Tab(GrauII) + \beta_5 Tab(GrauIII) + \beta_6 AVC + \beta_7 Hiperten. + \beta_8 Sistolica + \beta_9 Glicose$$

$$\Lambda = Deviance_{mod_{stepwise}} - Deviance_{mod_{cheio}} \sim \mathrm{X}^2_{(18-10)}$$

$$\Lambda = 3193 - 3188 = 5$$

Tabela 11 - Comparação entre os dois modelos pelo TRV

Estatística	χ^2	p-valor
Razão de Verossimilhança $(\Lambda) = 5$	15,50	0,75

O p_valor indica que o modelo selecionado pelo algoritmo se ajusta tão bem quanto o modelo com todas as variáveis, de maneira mais simples, pode-se dizer que as demais variáveis, presentes apenas no modelo cheio, contribuem pouco no ajuste.

8.4 Modelo Proposto

Apesar das variáveis: colesterol, IMC, diabetes e frequência cardíaca, apresentarem um p_valor pouco significativo e não terem sido selecionadas pelo algoritmo stepwise, são variáveis que fazem sentido no contexto, portanto, foram testadas separadamente uma a uma como possíveis variáveis do modelo final. A variável “Diastólica”, referente à pressão diastólica, não entrou no modelo por ser altamente correlacionada com a sistólica.

Sendo assim, os modelos propostos, ficaram definidos pelas seguintes variáveis:

- Sexo;
- Idade;
- Tabagismo;
- AVC;
- Hipertenso;
- Sistólica;
- Glicose;
- (Diabetes, Colesterol, IMC e/ou FreqCard);

Estrutura dos modelos,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 sex(masc) + \beta_2 Idade + \beta_3 Tab(GrauI) + \beta_4 Tab(GrauII) + \beta_5 Tab(GrauIII) + \beta_6 AVC + \beta_7 Hiperten. + \beta_8 Sistolica + \beta_9 Sistolica + \beta_{10} Glicose + \beta_{11} Nova\ variável\ testada$$

Abaixo, na tabela 12 são apresentados os testes de razão de verossimilhança verificando a existência ou não de diferenças significativas entre o modelo selecionado pelo algoritmo e os novos modelos propostos.

$$\Lambda = Deviance_{mod_{stepwise}} - Deviance_{mod_{+diabetes}} \sim \mathrm{X}^2_{(11-10)}$$

$$\Lambda = 3193{,}16 - 3193 = 0{,}16$$

$$\Lambda = Deviance_{mod_{stepwise}} - Deviance_{mod_{+Colesterol}} \sim \mathrm{X}^2_{(11-10)}$$

$$\Lambda = 3193{,}16 - 3191{,}6 = 1{,}56$$

$$\Lambda = Deviance_{mod_{stepwise}} - Deviance_{modelo_{+IMC}} \sim \mathrm{X}^2_{(13-10)}$$

$$\Lambda = 3193{,}16 - 3191{,}2 = 1{,}96$$

$$\Lambda = Deviance_{mod_{stepwise}} - Deviance_{mod_{+FreqCard}} \sim \mathrm{X}^2_{(11-10)}$$

$$\Lambda = 3193{,}16 - 3192{,}9 = 0{,}26$$

Tabela 12 - Comparação entre os modelos pelo TRV

Estatística	χ^2	p-valor
Razão de Vero. Teste Diabetes $(\Lambda) = 0{,}16$	3,84	0,68
Razão de Vero. Teste Colesterol $(\Lambda) = 1{,}56$	3,84	0,22
Razão de Vero. Teste IMC $(\Lambda) = 1{,}96$	7,81	0,59
Razão de Vero. Teste FreqCard $(\Lambda) = 0{,}26$	3,84	0,63

Os baixos resultados de p_valor indicam que as inclusões das variáveis testadas são pouco significativas, ou seja, contribuem pouco no ajuste.

Sendo assim, o modelo final, escolhido para se ajustar ao conjunto de dados foi o selecionado pelo algoritmo stepwise.

Na tabela 13 são apresentadas as estimativas pontuais realizadas por máxima verossimilhança, erros padrões e as estatísticas Wald para testar a significância de cada parâmetro.

Estatística Wald,

$$W_j = \frac{\beta_j - 0}{s(\beta_j)}$$

Tabela 13 - Métricas do modelo final

Efeito	Parâmetro	Estimativa	Desv.Padrão	W (z-valor)	p-valor
intercepto	β_0	-8,10	0,44	-18,289	$2{,}0\ x\ 10^{-16}$
Sexo (masc)	β_1	0,48	0,097	4,906	$9{,}3\ x\ 10^{-7}$
Idade	β_2	0,065	0,006	10,928	$2{,}0\ x\ 10^{-16}$
Tab. (Grau I)	β_3	0,061	0,146	0,418	$0{,}675$
Tab. (Grau II)	β_4	0,45	0,118	3,851	$1{,}18\ x\ 10^{-4}$
Tab. (Grau III)	β_5	0,75	0,147	5,095	$3{,}48\ x\ 10^{-7}$
AVC (1)	β_6	0,99	0,44	2,267	$0{,}0233$
Hipertenso(1)	β_7	0,24	0,126	1,943	$0{,}0519$
Sistólica	β_8	0,013	0,002	4,993	$5{,}93{,}x\ 10^{-6}$
Glicose	β_9	0,0077	0,0016	4,753	$2{,}01\ x\ 10^{-6}$
Deviance = 3193,16 com 4218 g.l. e AIC = 3213,1					

Tabela 14 - Razão das chances (Odds ratio)

Efeito	Estimativa pontual	Limite inferior (2,5%)	Limite superior (97,5%)
Sexo (masc)	1,615	1,339	1,9771
Idade	1,067	1,055	1,080
Tabagismo (Grau I)	1,063	0,7934	1,410
Tabagismo (Grau II)	1,57	1,2488	1,984
Tabagismo (Grau III)	2,12	1,5858	2,831
AVC (1)	2,70	1,1256	6,406
Hipertenso (1)	1,27	0,9970	1,639
Sistólica	1,013	1,0081	1,018
Glicose	1,0077	1,0045	1,011

Pelas estimativas pontuais apresentadas na tabela 14, pode-se interpretar os resultados como:

- Variáveis categóricas:
    - **Sexo**: A chance de indivíduos do sexo masculino virem a desenvolver quadros coronarianos em futuros 10 anos, é cerca de 61,5% maior do que a do sexo feminino.
    - **Tabagismo**: Pode-se de observar que as chances aumentam conforme o grau de tabagismo é maior, pessoas classificadas com 3° de tabagismo têm 112% mais chances de desenvolverem um quadro coronariano (em futuros 10 anos), com relação a quem não fuma.
    - **AVC**: Ter ao menos um quadro de AVC aumenta em 170% a chance de desenvolver DAC em futuros 10 anos.
    - **Hipertensão**: Indivíduos hipertensos têm 27% mais chances de desenvolverem um quadro coronariano em relação aos que não possuem hipertensão.
- Variáveis quantitativas:
    - **Idade**: Para cada aumento de um ano de vida, têm-se um aumento de 6,7% na chance de desenvolver DAC em futuros 10 anos.
    - **Sistólica**: Para cada aumento de uma unidade nos níveis médios de pressão sistólica, têm-se um aumento de 1,3% na chance de desenvolver DAC em futuros 10 anos.
    - **Glicose**: Para cada aumento de uma unidade nos níveis médios de glicose, têm-se um aumento de 0,77% na chance de desenvolver DAC em futuros 10 anos.

8.5 Aplicação Bayesiana

Para estimação dos parâmetros foi utilizado o algoritmo de MCMC mencionado em (6.5), com 45.000 iterações, onde as primeiras 10.000 foram descartadas. Como já mencionado em (6.7), as prioris utilizadas foram distribuições normais independentes com precisões baixas.

Tabela 15 – Estimativas (pontual) dos modelos

		MCMC (Metropolis)		**Max. Verossimilhança**	
Efeito	**Parâmetro**	**Estimativa**	**Desv.Padrão**	**Estimativa**	**Desv.Padrão**
intercepto	β_0	-8,05	0,42	-8,10	0,44
Sexo (masc)	β_1	0,47	0,096	0,48	0,097
Idade	β_2	0,064	0,006	0,065	0,006
Tab. (Grau I)	β_3	0,053	0,147	0,061	0,146
Tab. (Grau II)	β_4	0,45	0,116	0,45	0,118
Tab. (Grau III)	β_5	0,75	0,149	0,75	0,147
AVC (1)	β_6	0,98	0,44	0,99	0,44
Hipertenso(1)	β_7	0,25	0,122	0,24	0,126
Sistólica	β_8	0,013	0,002	0,013	0,002
Glicose	β_9	0,0076	0,0016	0,0077	0,0016

Tabela 16 – Intervalos de credibilidade

Efeito	**Parâmetro**	**Estimativa**	**2.5%**	**97.5%**
intercepto	β_0	-8,05	-8,94	-7,27
Sexo (masc)	β_1	0,47	0,288	0,666
Idade	β_2	0,064	0,053	0,077
Tab. (Grau I)	β_3	0,053	-0,237	0,339
Tab. (Grau II)	β_4	0,45	0,221	0,675
Tab. (Grau III)	β_5	0,75	0,447	1,033
AVC (1)	β_6	0,98	0,104	1,859
Hipertenso(1)	β_7	0,25	0,012	0,488
Sistólica	β_8	0,013	0,008	0,0178
Glicose	β_9	0,0076	0,004	0,0108

Tabela 17 - Razão das chances (Odds ratio)

Efeito	Parâmetro	Estimativa	2.5%	97.5%
Sexo (masc)	β_1	1,617	1,334	1,947
Idade	β_2	1,067	1,054	1,080
Tab. (Grau I)	β_3	1,066	0,788	1,404
Tab. (Grau II)	β_4	1,581	1,248	1,965
Tab. (Grau III)	β_5	2,134	1,564	2,811
AVC (1)	β_6	2,962	1,109	6,421
Hipertenso(1)	β_7	1,294	1,012	1,629
Sistólica	β_8	1,013	1,008	1,018
Glicose	β_9	1,007	1,004	1,011

Como o esperado, as estimativas obtidas por meio da inferência resultaram valores muito próximos aos resultados da máxima verossimilhança. Quando o tamanho da amostra é grande, a função de verossimilhança tem muito “peso” na distribuição a posteriori, além disso, o uso de prioris com baixas precisões se assemelham ao de prioris não informativas.

A interpretação das estimativas pontuais, são praticamente as mesmas feitas nos resultados anteriores.

- Variáveis categóricas:
    - **Sexo**: A chance de indivíduos do sexo masculino virem a desenvolver quadros coronarianos em futuros 10 anos, é cerca de 61,7 % maior do que a do sexo feminino.
    - **Tabagismo**: Pode-se de observar que as chances aumentam conforme o grau de tabagismo é maior, pessoas classificadas com 3° de tabagismo têm 113% mais chances de desenvolverem um quadro coronariano (em futuros 10 anos), com relação a quem não fuma.
    - **AVC**: Ter ao menos um quadro de AVC aumenta em 196% a chance de desenvolver DAC em futuros 10 anos.
    - **Hipertensão**: Indivíduos hipertensos têm 29% mais chances de desenvolverem um quadro coronariano em relação aos que não possuem hipertensão.
- Variáveis quantitativas:
    - **Idade**: Para cada aumento de um ano de vida, têm-se um aumento

de 6,7% na chance de desenvolver DAC em futuros 10 anos.

- **Sistólica**: Para cada aumento de uma unidade nos níveis médios de pressão sistólica, têm-se um aumento de 1,3% na chance de desenvolver DAC em futuros 10 anos.
- **Glicose**: Para cada aumento de uma unidade nos níveis médios de glicose, têm-se um aumento de 0,70% na chance de desenvolver DAC em futuros 10 anos.

Figura 5 -Trajetórias das cadeias geradas e densidade dos parâmetros β_0 a β_5

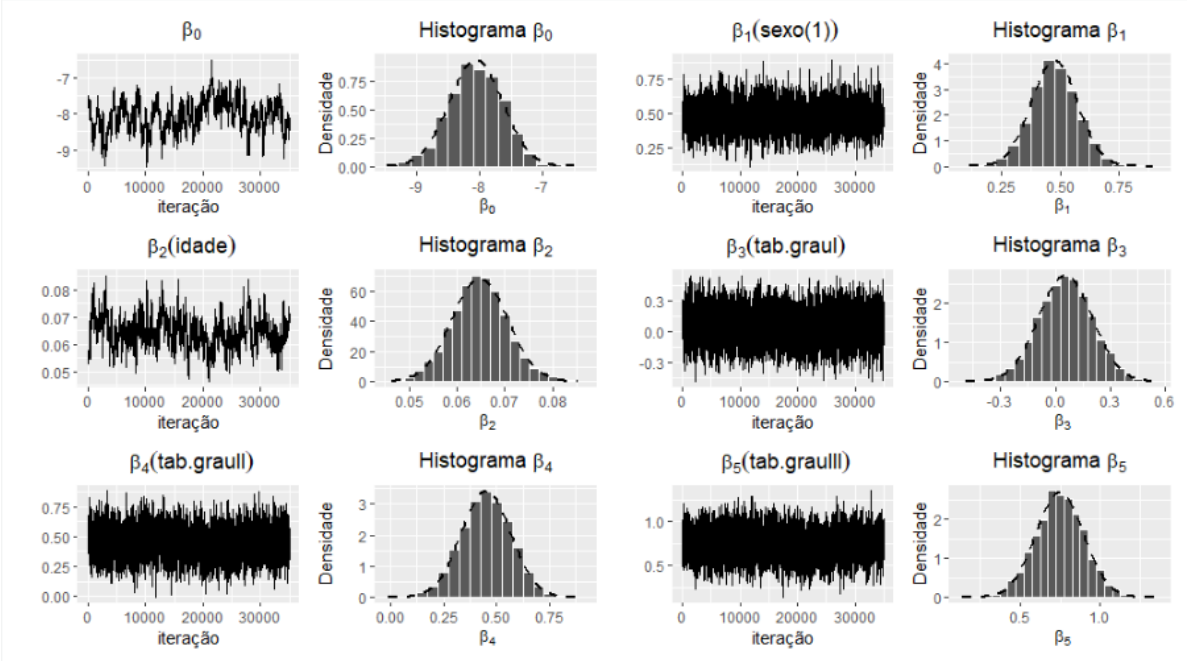
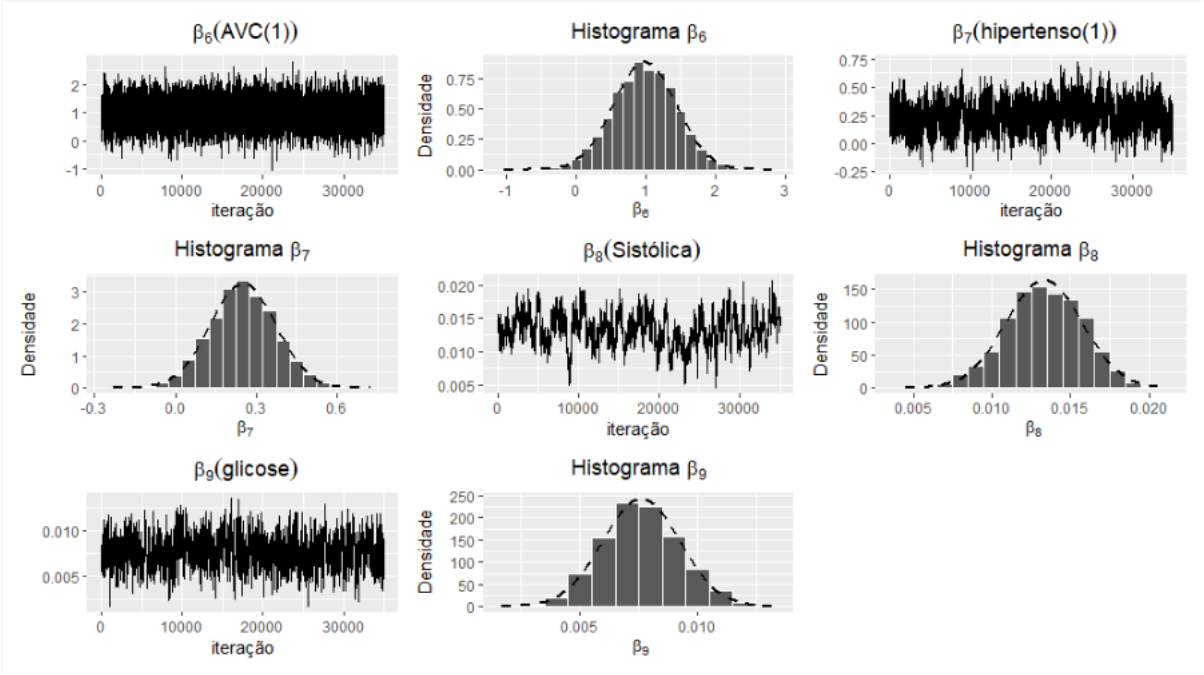


Figura 6 - Trajetórias das cadeias geradas e densidade dos parâmetros β_6 a β_9



As cadeias geradas pelo método MCMC mostraram convergência razoavelmente aceitáveis ao longo das iterações, e as distribuições obtidas para cada parâmetro, mostraram (graficamente) boa aderência à distribuição normal.

Figura 7 - Distribuições das razões de chances β_0 a β_5

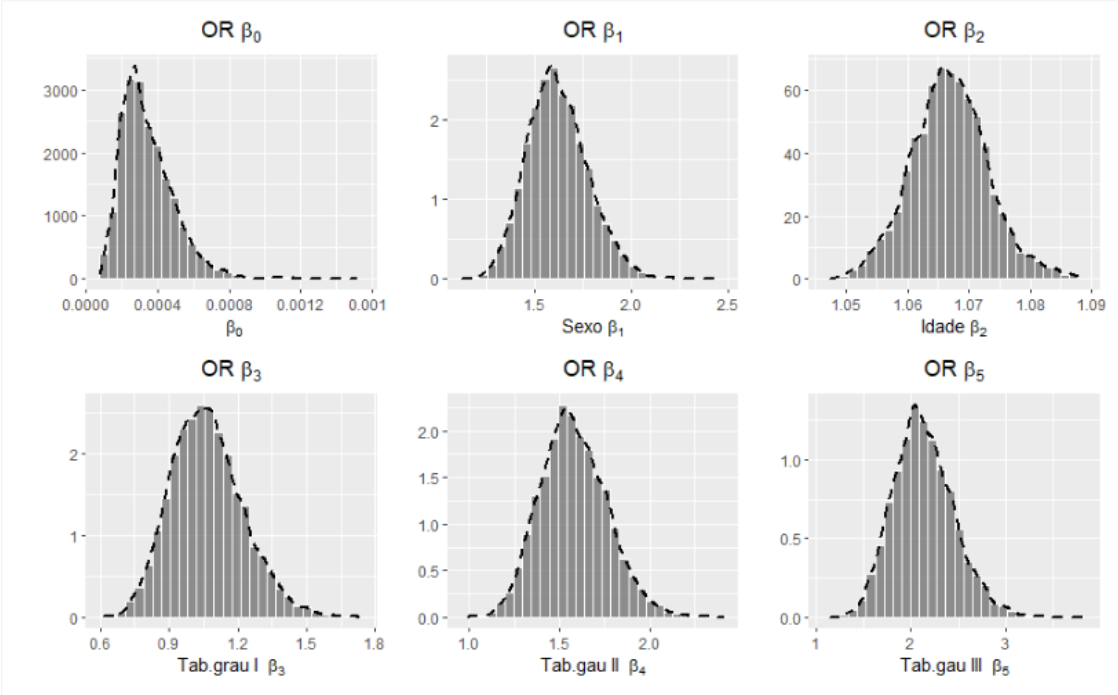
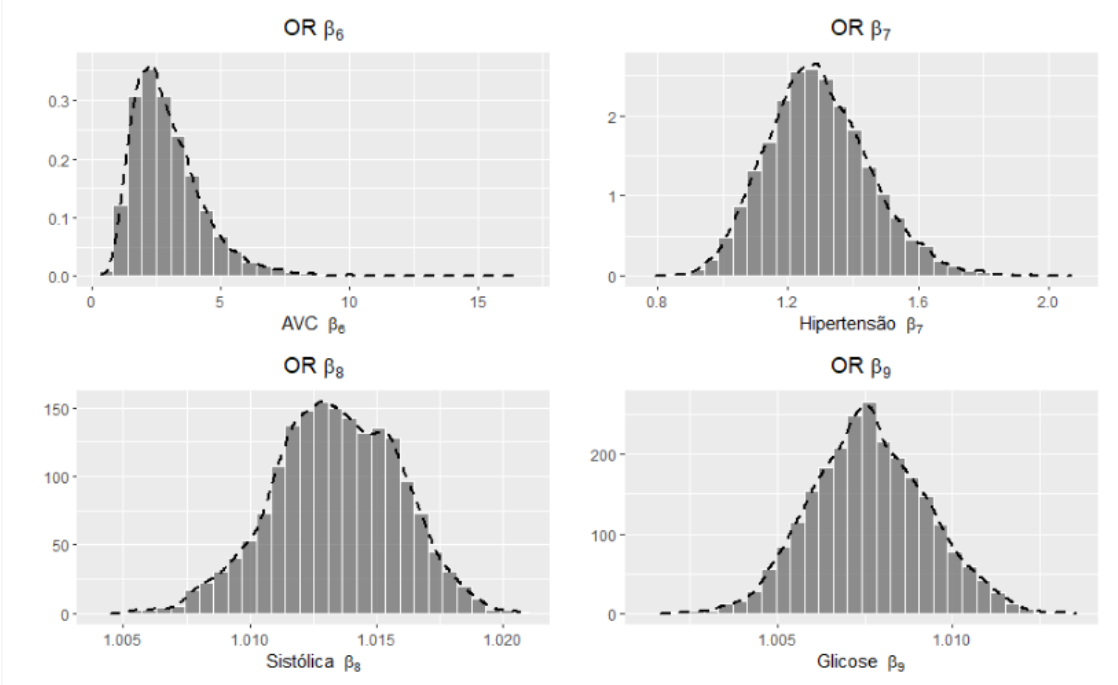


Figura 8 - Distribuições das razões de chances β_6 a β_9



Algumas densidades estimadas para razões de chances mostraram tendências assimétricas, o que já era esperado, por terem domínios limitados à zero.

9 Conclusão

Os modelos estatísticos vêm se mostrando uma ótima ferramenta em diferentes estudos científicos, principalmente nos relacionados a saúde. A regressão logística mostra uma via simples e eficiente de aplicação, facilitando a interpretação de resultados e auxiliando na tomada de decisão. Um bom modelo, pode ajudar a identificar fatores de riscos, ou até mesmo, fatores que mitiguem as chances de desenvolvimento de determinada doença.

No presente trabalho, é apresentado a utilização da regressão logística para identificar o desenvolvimento da doença arterial coronariana em futuros 10 anos, com duas formas de estimação dos parâmetros, a abordagem clássica (máxima verossimilhança) e bayesiana (utilizando o método MCMC). Os resultados mostraram que os modelos além identificar fatores de riscos, puderam mensurá-los com relação às chances. O trabalho também mostrou, uma forma simples de se implementar e interpretar a inferência bayesiana a partir do método de MCMC.

O trabalho também mostra um exemplo prático da teoria estatística, os resultados confirmam que para grandes amostras, os resultados das duas abordagens “clássica” e bayesiana tendem a se aproximar, principalmente com a utilização de prioris pouco informativas, como no caso de normais independentes com média zero e precisões baixas.

Referências

CORDEIRO, G. M. **Modelos Lineares Generalizados, Minicurso para o 12° SEAGRO e a 52ª Reunião Anual da RBRAS UFSM**, Santa Maria, RS, 2007.

DEMÉTRIO, C. G. **Modelos Lineares Generalizados em Experimentação Agronômica,46ª Reunião Anual da RBRAS**, Piracicaba, 2001

EHLERS, R. S. **INTRODUÇÃO À INFERÊNCIA BAYESIANA**.ed.5ª, 2007.

Kaggle, **Logistic regression To predict heart disease**. Disponível em:<https://www. kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression>. Acesso em: 20 Mai. 2020.

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

MENDONÇA, T. S. **Modelos de regressão logística clássica, bayesiana e redes neurais para Credit Scoring**. Dissertação (Mestrado), UFScar, São Carlos, 2008.

MONFARDINI, F. **Modelos Lineares Generalizados Bayesianos para Dados Longitudinais.** Dissertação (Mestrado). Programa de Matemática Aplicada e Computacional, UNESP, Presidente Prudente, SP, Fev. 2016.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models. Journal of the Royal Statístical Society. Série A (General).** vol. 135, No. 3, p. 370 - 384, 1972.

PAULA, Gilberto A. **MODELOS DE REGRESSÃO com apoio computacional**. Instituto de Matemática e Estatística. Universidade de São Paulo, São Paulo, Fev. 2013.

PINHO, Ricardo Aurino de et al. **Doença arterial coronariana, exercício físico e estresse oxidativo. Arq. Bras. Cardiol.** São Paulo, v. 94, n.4, p.549-555, Apr. 2010. Disponível em < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X 2010000400018&lng=en&nrm=iso >. Acesso em 10 Mar. 2020.

PIRES, M. C. **Abordagem Bayesiana para Modelos de Regressão Logística com Erros e Classificações Repetidas**. Dissertação (Mestrado). Programa de Pós-Graduação em Estatística, UFMG, Belo - Horizonte, MG, Maio 2011.

SANTOS, M. A.; MOALA, F.A.; TACHIBANA, V. M. **Predição do risco de óbito de infarto do miocárdio usando Regressão logística Bayesiana**. Departamento de Matemática, Estatística e Computação, UNESP, Presidente Prudente, SP.

Sociedade Brasileira de Cardiologia, **CARDIÔMETRO, MORTES POR DOENÇAS CARDIOVASCULARES NO BRASIL**. Disponível em : < http://www.cardiometro.com.br>. Acesso em 15 de Abril. 2020.

TURKMAN, Maria A. A.; SILVA, G. L. **Modelos Lineares Generalizados: da teoria à prática. In: VIII Congresso Anual da Sociedade Portuguesa de Estatística**, Lisboa. 2000.

World Health Organization, **Cardio vascular disease (CVDs)**. Disponível em : < https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) >. Acesso em: 12 Mar. 2020.