
UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

MARIA BEATRIZ GALDINO DA SILVEIRA

# APLICAÇÃO DA REGRESSÃO LOGÍSTICA NA ANÁLISE DOS FATORES DE RISCO ASSOCIADOS À HIPERTENSÃO ARTERIAL

CAMPINA GRANDE - PB
2021

MARIA BEATRIZ GALDINO DA SILVEIRA

# APLICAÇÃO DA REGRESSÃO LOGÍSTICA NA ANÁLISE DOS FATORES DE RISCO ASSOCIADOS À HIPERTENSÃO ARTERIAL

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

**Orientador:** Prof.Dr. Sílvio Fernando Alves Xavier Júnior
**Coorientador:** Profa. Dra. Érika Fialho Morais Xavier

**CAMPINA GRANDE - PB**
**2021**

S587a Silveira, Maria Beatriz Galdino da.

Aplicação da regressão logística na análise dos fatores de risco associados à hipertensão arterial [manuscrito] / Maria Beatriz Galdino da Silveira. - 2021.

32 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2021.

"Orientação : Prof. Dr. Sílvio Fernando Alves Xavier Júnior , Departamento de Estatística - CCT."

1. Regressão logística. 2. Estatística. I. Título

21. ed. CDD 519.5

Elaborada por Lucas H. A. da Silva - CRB - 15/898 BC/UEPB

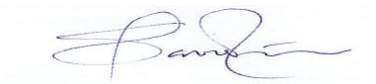
MARIA BEATRIZ GALDINO DA SILVEIRA

# APLICAÇÃO DA REGRESSÃO LOGÍSTICA NA ANÁLISE DOS FATORES DE RISCO ASSOCIADOS À HIPERTENSÃO ARTERIAL

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 07 de outubro de 2021.

**BANCA EXAMINADORA**



Prof. Dr. Sílvio Fernando Alves Xavier Júnior
(Orientador)
Universidade Estadual da Paraíba (UEPB)

Nyedja Fialho Morais Barbosa

Profa. Ma. Nyedja Fialho Morais Barbosa
Universidade Estadual da Paraíba (UEPB)

Ana Patrícia Bastos Peixoto

Profa. Dra. Ana Patrícia Bastos Peixoto de Oliveira
Universidade Estadual da Paraíba (UEPB)

*A Deus, meu bem maior.*

## AGRADECIMENTOS

Agradeço em primeiro lugar a Deus, fonte de todas as bênçãos, inspiração para minha vida, que me permitiu chegar até aqui, superando todas as dificuldades que surgiram em meu caminho.

Agradeço aos meus pais, José Felix e Solange Galdino, pelo amor, incentivo e apoio, sempre se esforçando para fazer o melhor por mim. Sou grata também ao meu irmão João Neto pelo apoio de sempre e a todos da minha família.

Ao meu orientador, Silvio Fernando Alves Xavier Júnior, que atenciosa e pacientemente apoiou esse trabalho, esclarecendo as dúvidas e realizando as correções.

À minha coorientadora, Érika Fialho Morais Xavier, por todo o auxílio, apoio e incentivo para que este trabalho fosse realizado.

Aos professores Nyedja Barbosa e Ana Patrícia Oliveira pela disponibilidade em participar da banca examinadora.

Aos meus amigos e colegas de curso, Rafaella, Viviane, Gilmar e Eduardo, por estarem sempre presentes, me incentivando e ajudando nos momentos que vivemos durante o tempo da graduação, bem como a todos os demais colegas com quem compartilhei experiências durante a trajetória acadêmica.

Agradeço à Universidade Estadual da Paraíba que me permitiu trilhar o árduo, mas satisfatório caminho acadêmico e assim, realizar o sonho da graduação.

A todos os professores do curso por me proporcionarem excelentes aulas e a todos que direta ou indiretamente fizeram parte da minha formação.

*"No futuro, o pensamento estatístico será tão necessário para a cidadania eficiente como saber ler e escrever."*

*(H.G. Wells)*

## RESUMO

A regressão logística é uma técnica importante para modelagem de dados quando se deseja analisar a relação entre uma variável resposta e uma ou mais variáveis independentes. A técnica permite que se estime as chances relacionadas à probabilidade da ocorrência de um evento de interesse. A regressão logística diferencia-se da regressão linear devido à natureza dicotômica da variável dependente e vem sendo utilizada em diversas áreas do conhecimento, incluindo estudos na área da saúde. O presente trabalho utilizou a técnica da regressão logística com o objetivo de analisar a associação entre Hipertensão Arterial e determinados fatores de risco. Os dados utilizados provém da Pesquisa Nacional de Saúde (PNS) do ano de 2019, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em território nacional. Foram ajustados dois modelos, sendo o modelo final composto por sete variáveis com significância estatística de 5%. As técnicas de diagnóstico indicaram um ajuste adequado do modelo, bem como sua precisão para predições. Os resultados apontam que fatores como o aumento da idade, índice de massa corporal (IMC) alto e o diagnóstico positivo para diabetes aumentam as chances de um indivíduo ser hipertenso.



**Palavras-chaves**: Associação. Fatores de risco. Modelo ajustado.

## ABSTRACT

Logistic regression is an important technique for data modeling when you want to analyze the relationship between a response variable and one or more independent variables. The technique allows one to estimate the chances related to the probability of occurrence of an event of interest. Logistic regression differs from linear regression due to the dichotomous nature of the dependent variable and has been used in several areas of knowledge, including health studies. This study used the logistic regression technique to analyze the association between Hypertension and certain risk factors. The data used comes from the Pesquisa Nacional de Saúde (PNS) for the year 2019, carried out by the Instituto Brasileiro de Geografia e Estatística (IBGE) in the country. Two models were adjusted, the final model being composed of seven variables with a statistical significance of 5%. Diagnostic techniques indicated an adequate fit of the model, as well as its accuracy for predictions. The results show that factors such as increasing age, high body mass index (IMC) and a positive diagnosis for diabetes increase the chances of an individual being hypertensive.



**Keywords**: Association. Risk factors. Fit model.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

A Hipertensão Arterial (HA) é considerada uma doença crônica não transmissível caracterizada pela persistente alteração da Pressão Arterial (PA). Conforme as Diretrizes Brasileiras de Hipertensão Arterial (BARROSO et al., 2021) um indivíduo é considerado hipertenso quando a sua pressão arterial sistólica (PAS) é maior ou igual a 140 mmHg e/ou a pressão arterial diastólica (PAD) é maior ou igual a 90 mmHg medida com a técnica correta em pelo menos duas ocasiões diferentes, sem o uso de medicação anti-hipertensiva.

A HA é considerada uma doença multifatorial, ocasionada por fatores genéticos, ambientais e sociais tais como: idade, sexo, etnia, ingestão de sódio e potássio, sedentarismo, álcool e fatores socioeconômicos. A doença também se constitui o principal fator de risco modificável para doenças cardiovasculares, doença renal cônica e morte prematura. Conforme estimativas da Organização Mundial de Saúde (OMS), 22,3% da população mundial com 18 anos ou mais sofria com a doença (MARQUES et al., 2020). Dados do Datasus referentes ao ano de 2017 mostraram que a HA esteve associada a 45% das mortes por doenças cardíacas e a 51% das mortes por doença cerebrovascular no Brasil (BARROSO et al., 2021).

O presente trabalho tem por objetivo ajustar um modelo de regressão logística capaz de realizar adequadamente predições da ocorrência de HA em indivíduos com base em determinadas características. Nesse sentido, visa contribuir juntamente com as análises já existentes sobre fatores de risco associados à doença.

Os métodos de regressão são utilizados quando em uma análise de dados se deseja descrever a relação entre uma variável resposta e uma ou mais variáveis explanatórias. Por meio de uma análise de regressão pode-se explorar tanto a direção (positiva ou negativa), como a magnitude (fraca ou forte) da associação entre a variável dependente ($Y$) e a variável independente ($X$), além de ser possível prever os valores da variável dependente, por meio da variável independente (FIGUEIRA, 2006).

Nos casos onde existem mais de uma variável independente, os métodos de regressão também permitem verificar as contribuições dadas por cada variável para o modelo em geral. Nesse sentido, conforme Hosmer e Lemeshow (2000), os métodos de regressão tem se tornado um componente integrante de qualquer análise de dados que vise explicar a relação entre uma variável resposta e uma ou mais variáveis explanatórias.

A regressão logística difere da regressão linear inicialmente pelo fato de que sua variável resposta é de natureza dicotômica. Sendo assim, o objetivo desta técnica é ajustar um modelo em que a variável dependente representa a probabilidade de um determinado evento ocorrer em função de uma ou mais variáveis independentes, que, por sua vez, podem ser contínuas ou binárias. (FIGUEIRA, 2006). Na área da saúde a regressão logística vem sendo amplamente utilizada com o objetivo de prever a ocorrência de uma doença com base nas características dos pacientes.

# 2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção aborda-se os principais conceitos relacionados à regressão logística. Inicialmente discute-se o desenvolvimento histórico da técnica e em seguida é apresentada a teoria estatística relacionada ao modelo logístico, estimação dos parâmetros e diagnósticos do modelo.

## 2.1 Marco Histórico

O termo ”regressão” como um conceito estatístico foi utilizado primeiramente pelo pesquisador britânico Francis Galton (1822-1911) em estudos sobre hereditariedade. Ao realizar um estudo com sementes de ervilhas, Galton observou que as sementes de ervilhas maiores geraram ervilhas menores e as sementes menores geraram ervilhas maiores. O pesquisador concluiu, então, que as sementes regrediam à média. Ao realizar experimentos sobre a estatura de pais e filhos humanos, Galton percebeu o mesmo efeito obtido no experimento com as sementes (ALVES, 2016).

Em 1885, Galton publicou no Journal of Antropological Institute um artigo intitulado Regressão em direção à Média na Estatura Hereditária, no qual utiliza pela primeira vez o termo regressão linear. Os estudos de regressão foram desenvolvidos e aperfeiçoados posteriormente por outros pesquisadores, como Karl Pearson e Francis Edgeworth e apenas no século XX o Método dos Mínimos Quadrados passou a ser utilizado para estimar os parâmetros do modelo (ALVES, 2016).

A técnica da regressão logística foi descoberta no século XIX em estudos sobre o crescimento das populações e as reações químicas no curso de autocatálise. Em 1845 Pierre-François Verhulst (1804-1849) publicou um artigo na revista Proceedings, no qual define a curva de crescimento populacional por meio de uma função denominada por ele de “logística” (SOUZA, 2006).

Em 1920, Raymond Pearl (1879-1940) e Lowell J. Reed (1886-1866) conseguem chegar novamente à curva logística, ao estudarem a população dos Estados Unidos, mesmo sem conhecer o trabalho de Verhulst. Entretanto, apenas a partir da década de 1950, a técnica passou a ser utilizada mais amplamente em estudos científicos, como observa Cramer (2002) ao quantificar o número de artigos em jornais estatísticos que utilizam as palavras *probit* ou *logit* (SOUZA, 2006).

Os trabalhos de Cox & Snell, *Analysis od Binary Data* (2018) e Hosmer & Lemeshow, *Applied Logistic Regression* (2013) são considerados grandes avanços para os estudos de regressão logística. Um dos exemplos mais famosos de utilização da técnica é o *Framingham Heart Study*, um estudo sobre fatores que podem ocasionar doenças cardiovasculares, realizado com a parceria da Universidade de Boston (MESQUITA, 2014). Atualmente, a regressão logística tem sido utilizada em diversas áreas da pesquisa científica como saúde, economia, marketing e educação, sendo considerada uma importante ferramenta para

análise de variáveis dicotômicas.

Podemos citar os seguintes exemplos de análises que podem ser realizadas utilizando-se o método da regressão logística:

- A ocorrência de uma doença ($Y$) em relação às características dos pacientes ($X_i$);
- A compra de um produto ou serviço ($Y$) em relação a qualidade do produto ($X_1$), preço ($X_2$), anúncio ($X_3$);
- A probabilidade de voto em um determinado partido ($Y$) com base em características sociodemográficas e votos em eleições precedentes do eleitor ($X_i$);
- Análise de crédito ($Y$) em relação à idade ($X_1$), renda ($X_2$), valor do patrimônio ($X_3$), escolaridade ($X_4$).

## 2.2 Regressão Logística

Os modelos de regressão são utilizados em análises estatísticas nas quais se busca descrever as relações entre a variável resposta ($Y$) e a variável explicativa ($X$). Quando se tem apenas uma variável independente, pode-se estabelecer uma regressão linear simples, utilizando o seguinte modelo estatístico:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n, \tag{2.1}$$

onde, $Y_i$ é a variável resposta, $x_i$ a variável explanatória, $\beta_0$ e $\beta_1$ os parâmetros de regressão e $\epsilon_i$ o erro do modelo.

Para os casos em que se têm duas ou mais variáveis explanatórias pode-se modelar os dados por meio da regressão linear múltipla, a qual se constitui uma generalização do modelo anterior:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_h x_{hi} + \epsilon_i, i = 1, \ldots, n. \tag{2.2}$$

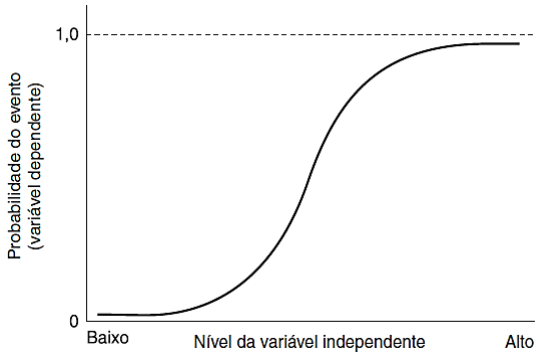
A regressão logística se diferencia dos modelos de regressão linear porque no seu caso a variável dependente é qualitativa e binária. Na regressão logística a variável resposta assume apenas valores 0 e 1, sendo geralmente "1" a ocorrência do evento de interesse e "0" a sua ausência, ou em outros termos, "1" corresponde ao sucesso e "0" ao fracasso. Portanto, o valor da previsão de $Y$ sempre estará no intervalo $0 \geq Y \leq 1$.

Conforme pontua Hair et al (2009), devido a natureza binária da variável dependente, as suposições da regressão linear e múltipla são violadas. Neste sentido, os resíduos seguem distribuição binomial ao invés da normal e a variância não é constante, apresentando heterocedasticidade, além disso, as transformações não são suficientes para corrigir essas

violações. Sendo assim, a regressão logística é um método que se ocupa particularmente com esses problemas.

O termo regressão logística é derivado da transformação realizada com a variável dependente (transformação logit). No modelo de regressão logística, a curva logística é ajustada aos dados permitindo que se calcule a probabilidade de ocorrência de um evento de interesse. A função logística $f(Z) = \frac{1}{1+e^{-(Z)}}$, assume valores entre 0 e 1 para qualquer Z entre -$\infty$ e +$\infty$, apresentando-se como uma curva em formato de "S" conforme demonstra a Figura 1.

Figura 1 – Forma da relação logística entre a variável dependente e as variáveis independentes.




Fonte: (HAIR et al., 2009)

O modelo logístico é definido por:

$$Z = \ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \tag{2.3}$$

onde $p$ é a probabilidade de ocorrência do evento de interesse, $X$ o vetor de variáveis independentes, $\beta_0$, $\beta_1$, $\beta_2$ ... $\beta_k$ os parâmetros do modelo. O termo $\ln\left(\frac{p}{1-p}\right)$ é denominado *logit* e $\left(\frac{p}{1-p}\right)$ é a razão de chances de ocorrência do vento de interesse.

## 2.3 Estimação dos parâmetros

Para os modelos de regressão linear o método mais utilizado para a estimação dos parâmetros é o de mínimos quadrados. Neste método, os coeficientes são estimados pelos valores que minimizam a soma das diferenças quadradas entre os valores observados e os valores preditos. Entretanto, devido a ausência de relação linear em regressão logística, o método de mínimos quadrados não é apropriado para estimar os coeficientes, sendo assim, utiliza-se o método da máxima verossimilhança. O método de máxima verossimilhança consiste em encontrar o valor de $\beta$ que maximiza $L(x_1, x_2, \ldots, x_n)$ para uma determinada amostra.

Para o modelo de regressão logística simples com $Y_i \sim Ber(\pi_i)$, a função de distribuição de probabilidade é dada por:

$$f(y_i, \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{2.4}$$

Logo, a função de verossimilhança será dada por:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \beta \in \mathbb{R}^{(2)}. \tag{2.5}$$

Aplicando o logaritmo, a expressão é definida por:

$$\begin{aligned}
l(\beta) &= \ln\left[L(\beta)\right] && (2.6)\\
&= \ln\left[\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] && (2.7)\\
&= \sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right] && (2.8)\\
&= \sum_{i=1}^{n}\left[y_i \ln(\pi_i) + \ln(1-\pi_i) - y_i\ln(1-\pi_i)\right] && (2.9)\\
&= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right] && (2.10)\\
&= \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 x_1) + \ln\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_1)}\right)\right] && (2.11)\\
&= \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 x_1) - \ln(1+\exp(\beta_0+\beta_1 x_1))\right]. && (2.12)
\end{aligned}$$

Para encontrar o valor de $\beta$ que maximiza $l(\beta)$, deriva-se $l(\beta)$ em relação a cada parâmetro $(\beta_0, \beta_1)$, obtendo-se duas equações:

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial(\beta_0)} &= \sum_{i=1}^{n}\left[y_i - \frac{1}{1+\exp(\beta_0+\beta_1 x_1)}\exp(\beta_0+\beta_1 x_1)\right] && (2.13)\\
\frac{\partial l(\beta)}{\partial(\beta_1)} &= \sum_{i=1}^{n}\left[y_i x_i - \frac{1}{1+\exp(\beta_0+\beta_1 x_1)}\exp(\beta_0+\beta_1 x_1)\right]. && (2.14)
\end{aligned}$$

que igualadas a zero, geram o seguinte sistema de equações:

$$\begin{cases} \sum_{i=1}^{n}(y_i\pi_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i\pi_i) = 0 \end{cases} \tag{2.15}$$

em que $i = 1, \ldots, n$ e $\pi_i = \frac{\exp(\beta_0+\beta_1 x_1)}{1+\exp(\beta_0+\beta_1 x_1)}$

Como as equações não são lineares em $\beta_0$ e $\beta_1$, são necessários métodos iterativos para resolução, estes métodos encontram-se disponíveis em vários programas computacionais.

## 2.4 Razão de chances

A Razão de Chances (*odds ratio* - O.R) é definida como a razão entre a chance de ocorrência de um evento em um grupo e a ocorrência deste evento em outro grupo. Assim, considerando a existência de dois grupos e as respectivas probabilidades de um evento ocorrer em cada um deles dadas por $p$ e $q$ a razão de chances é obtida por:

$$OR = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{p(1-q)}{q(1-p)}. \tag{2.16}$$

Considerando os parâmetros estimados por meio da regressão logística, a razão de chances é calculada exponencializando-se os coeficientes: $\exp(\beta_i)$ (FIGUEIRA, 2006).

## 2.5 Métodos de verificação da qualidade do ajuste

Após a estimação dos coeficientes é necessário verificar a significância das variáveis para o modelo. Para tanto, realiza-se testes de hipóteses. Os mais utilizados são: o teste da Razão da Verossimilhança, o Teste de Wald e o Pseudo $R^2$ de Cox e Snell o Critério de Informação de Akaike (AIC).

### *2.5.1 Teste da Razão de verossimilhança*

Conforme Cabral (2013), por meio desta medida, testa-se se os coeficientes de regressão associados a $\beta$ são todos nulos, com exceção de $\beta_0$. Sendo assim, compara-se os valores observados e os valores esperados usando a função de verossimilhança, conforme a seguinte expressão:

$$D = -2\ln\left[\frac{Função\ de\ máxima\ verossimilhança\ do\ modelo\ corrente}{Função\ de\ máxima\ verossimilhança\ do\ modelo\ saturado}\right]. \tag{2.17}$$

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]. \tag{2.18}$$

Por modelo corrente nos referimos ao modelo com todas as variáveis, já o modelo saturado é aquele com apenas as variáveis de interesse para o estudo. A função D, chamada de *deviance*, é sempre positiva, e quanto menor, melhor é o ajuste do modelo. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \beta_1 = \ldots = \beta_t = 0 \\ H_1 : \exists_{j=1,\ldots,p}\beta_j \neq 0 \end{cases} \tag{2.19}$$

A significância de uma variável independente é estimada comparando-se o valor de D com e sem esta variável da equação. Ao rejeitar a hipótese nula, conclui-se que a variável testada é significativa para o modelo.

### 2.5.2 Teste de Wald

O teste de Wald avalia se cada coeficiente é significativamente diferente de zero. Neste sentido, o teste de Wald avalia se a relação uma determinada variável independente com a variável dependente é estatisticamente significativa. A estatística de teste é dada por:

$$W_j = \frac{\hat{\beta}_j}{var(\hat{\beta}_j)}. \tag{2.20}$$

Há casos em que o teste de Wald costuma não rejeitar a hipótese nula quando esta deveria ser rejeitada (MESQUITA, 2014). Sendo assim, recomenda-se que o teste da razão de verossimilhança seja utilizado nos casos em que houver dúvidas acerca da eficiência do teste de Wald.

### 2.5.3 Pseudo $R^2$ de Cox e Snell

Esta medida é denominada pseudo $R^2$, pelo fato de apresentar semelhanças com o $R^2$ dos modelos de regressão linear. Entretanto, deve-se ressaltar que, apesar da similaridade, a interpretação de ambos é diferente. Existem muitas maneiras de calcular o pseudo $R^2$, sendo o pseudo $R^2$ de Cox e Snell (2018) um dos mais frequentemente usados pelos softwares estatísticos. A medida é definida por:

$$R^2 = 1 - \left(\frac{L(\beta)_0}{L(\beta)_M}\right)^{\frac{2}{n}}, \tag{2.21}$$

onde $n$ é o tamanho da amostra, $L(\beta)_0$ o valor da função verossimilhança para um modelo sem preditores e $L(\beta)_M$ a verossimilhança do modelo sendo estimado.

A medida resulta em um valor que varia de 0 a 1. Esses valores são utilizados para comparar os modelos nos quais as variáveis independentes melhor explicam as variações na variável dependente. Busca-se um modelo que apresente um pseudo $R^2$ mais elevado.

### 2.5.4 Critério de Informação de Akaike (AIC)

O critério de informação de Akaike penaliza os modelos com mais variáveis, apresentando valores menores para modelos mais parcimoniosos. O AIC é definido por:

$$AIC = -2\ln(L_p) + 2[(p+1)+1], \tag{2.22}$$

onde $L_p$ é a função de máxima verossimilhança do modelo e $p$ é o número de variáveis explicativas.

### 2.5.5 Curva ROC (Receiver Operating Characteristic)

A Curva ROC mede a capacidade de predição do modelo, sendo produzida bi-dimensionalmente através das predições de sensibilidade e especificidade. A sensibilidade indica a proporção de verdadeiros positivos e a especificidade a proporção de verdadeiros negativos. A área abaixo da curva ROC, denominada AUC (Area Under the ROC Curve) compara os classificadores da curva em um único valor, indicando a probabilidade do modelo realizar predições corretas (FAWCETT, 2006). O valor apresentado pela AUC é sempre entre 0 e 1, e segundo Hosmer e Lemershow (2013) deve ser considerado aceitável acima de 0,7.

## 2.6 Diagnósticos do modelo

Conforme Souza (2006) é importante que se faça uma análise dos resíduos e diagnósticos do modelo ajustado a fim de detectar possíveis problemas, como por exemplo:

- Presença de observações discrepantes;
- Inadequação das pressuposições para os erros aleatórios ou para as médias;
- Colinearidade entre as colunas da matriz do modelo;
- Forma funcional do modelo inadequada;
- Presença de observações influentes.

Algumas das medidas mais utilizadas para análise dos resíduos e diagnóstico de regressão logística serão apresentadas a seguir.

### 2.6.1 Diagonal da matriz H

Utilizam-se os elementos da matriz $\mathbf{H}$ para verificar pontos extremos no espaço designado. Como tais pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, sua eliminação pode ocasionar mudanças importantes em uma análise estatística.

Tendo em vista que em regressão logística a $Var(\epsilon_i) = \pi_i(1 - \pi_i)$ não é constante, a matriz de projeção para o modelo logístico, utilizando a definição de mínimos quadrados ponderados é dada por:

$$\mathbf{H} = \mathbf{Q}^{\frac{1}{2}} X(\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{\frac{1}{2}}. \tag{2.23}$$

Tal matriz sugere a utilização dos elementos da diagonal principal de $\mathbf{H}$ para detectar a presença de pontos de alavanca no modelo. É importante salientar que conforme

Hosmer e Lemeshow (2013) a matriz de projeção **H** deve ser utilizada com cuidado em regressão logística, pois suas interpretações diferem do caso normal linear. A forma diagonal da matriz $\hat{H}$ é dada por:

$$\hat{h}_{ii} = \hat{\pi}_i(1-\hat{\pi}_i)(x_i^T)\left[I(\hat{\beta})\right]^{-1}(x_i);\ i = 1, 2, \ldots, n. \tag{2.24}$$

### 2.6.2 Resíduo de Pearson

O resíduo de Pearson ajuda a classificar observações que podem ser consideradas *outliers*. O resíduo ordinário, definido como a diferença entre os valores observados e os valores preditos é dado por:

$$r_i = y_i - \hat{\pi}_i. \tag{2.25}$$

Por não ser útil para detectar *outliers*, é necessário transformar esse resíduo a fim de eliminar o efeito de medição da variável resposta e preditora. Os resíduos de Pearson fazem parte da estatística qui-quadrado de Pearson, e a indicação de um bom ajuste para o modelo ocorre quando o valores resultantes são pequenos. O resíduo de Pearson é definido por:

$$(rp)_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}};\ i = 1, 2, \ldots, n. \tag{2.26}$$

### 2.6.3 Resíduo de Deviance

São componentes da *Deviance*, utilizados para detectar os erros no ajuste do modelo. Tais resíduos medem se existem discrepâncias entre o modelo saturado e o modelo restrito em relação às observações $y_i$. O resultado da *deviance* é baseado no logaritmo da verossimilhança, definido por:

$$d_i = \begin{cases} -\sqrt{-2\ln(1-\hat{\pi}_i)}, & se\ \ y_i = 0 \\ \pm\sqrt{2\left[y_i\ln(\frac{y_i}{\hat{\pi}_i}) + (-y_i)\ln(\frac{1-y_i}{1-\hat{\pi}_i})\right]}, & se\ \ 0 < y_i < \\ \sqrt{-2\ln(\hat{\pi}_i)}, & se\ \ y_i = 1 \end{cases} \tag{2.27}$$

### 2.6.4 Resíduo Quantílico Aleatorizado

Os resíduos quantílicos aleatorizados foram propostos por Dunn e Smith (1996) para variáveis respostas que não possuem distribuição Normal. Tais resíduos assumem distribuição Normal se os parâmetros do modelo foram estimados de forma consistente. A abordagem utilizada é semelhante á de Cox e Snell (1968), diferindo no fato de que o

enfoque destes foi em correções da média e variância, enquanto o daqueles na transformação para a normalidade. Os resíduos quantílicos aleatorizados vêm sendo bastante utilizados em trabalhos científicos (PEREIRA, 2019). O resíduo quantílico aleatorizado é dado pela expressão:

$$\hat{r}q_{ij} = \Phi^{-1}\{G_Y(y_{ij}|f(\hat{\alpha}, x_{ij}), \hat{\theta})\},\ i = 1, \ldots, n,\ j = 1, \ldots, m_i, \tag{2.28}$$

onde $\Phi$ e $G_Y$ são as fda da distribuição normal padrão e da distribuição considerada no ajuste, respectivamente.

### 2.6.5 C e Cbar

Avaliam a influência das observações individuais sobre $\beta$, possuindo a mesma ideia da distância de Cook na teoria da regressão linear. Estes diagnósticos são baseados no intervalo de confiança. A medida C, descrita por Pregibon (1981) é definida por:

$$C_i = \frac{(rp_i)^2 h_{ii}}{(1 - h_{ii})^2};\ i = 1, 2, \ldots, n. \tag{2.29}$$

Por sua vez, a medida $\bar{C}$, descrita por Christensen (1997) é definida por:

$$\bar{C}_i = \frac{(rp_i)^2 h_{ii}}{(1 - h_{ii})};\ i = 1, 2, \ldots, n. \tag{2.30}$$

### 2.6.6 DIFCHISQ

Utiliza aproximações lineares e a estatística qui-quadrado de Pearson. Esta medida é adequada para identificar as observações mal ajustadas, que contribuam consideravelmente na diferença entre os dados e os valores preditos (SOUZA, 2006). A medida é definida por:

$$DIFCHISQ_i = \frac{\bar{C}_i}{h_{ii}} = \frac{(rp_i)^2}{1 - h_{ii}};\ i = 1, 2, \ldots, n. \tag{2.31}$$

### 2.6.7 DIFDEV

Esta medida, baseada no resíduo da deviance, é definida por:

$$DIFDEV_i = d_i^2 + \bar{C}_i = d_i^2 + \frac{(rp_i)^2}{h_{ii}(1 - h_{ii})};\ i = 1, 2, \ldots, n. \tag{2.32}$$

A DIFDEV é utilizada para identificar observações que são influentes ou não na estimação do ajuste do modelo, permitindo que se decida posteriormente sobre a sua permanência na análise (SOUZA, 2006).

# 3 APLICAÇÃO

A seguir são apresentadas as etapas da aplicação da regressão logística ao conjunto de dados proposto, buscando-se investigar a associação entre hipertensão arterial e diversos fatores de risco da doença.

## 3.1 Análise descritiva dos dados

Os dados são provenientes da Pesquisa Nacional de Saúde (PNS), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em 2019. A pesquisa é realizada em convênio com o Ministério da Saúde e visa oferecer subsídios para a formulação de políticas públicas nas áreas de promoção, vigilância e atenção à saúde do SUS (Sistema Único de Saúde). Os resultados da pesquisa são disponibilizados pelo IBGE para o conjunto do País, grandes Regiões e Unidades da Federação.

Primeiramente foi realizada a seleção de 11 variáveis relacionadas a hipertensão arterial. Dentre estas incluem-se variáveis demográficas como sexo, idade e cor, variáveis antropométricas como peso e altura, variáveis de estilo de vida como frequência de bebidas alcoólicas, tabagismo e exercícios físicos e consumo de sal e a variáveis clínicas como o diagnóstico de hipertensão e de diabetes. Em seguida, foram excluídas as observações com informações incompletas (NA). Assim, a base de dados final passou a ter 33.457 observações.

Optou-se ainda pelo acréscimo da variável IMC (Índice de Massa corporal), calculado pela divisão do peso (em kg) pelo quadrado da altura (em metros). Assim, as variáveis peso e altura foram desconsideradas no estudo, totalizando-se 10 variáveis para análise. A variável dicotômica Hipertenso, que indica se os indivíduos receberam ou não diagnóstico de hipertensão por algum médico, foi considerada a como variável dependente. As variáveis independentes foram: Sexo, que indica o sexo do indivíduo; Idade em anos; Cor, por autodeclaração, conforme as categorias estabelecidas pelo IBGE, a saber, branca, preta, amarela, parda e indígena; IMC, valor numérico indicando o índice de massa corporal, calculado pela razão entre peso (em Kg) e altura ao quadrado (em metros); Bebida, com categorias indicando a frequência de consumo de bebida alcóolica; Tabagismo, com categorias apresentando a frequência em que o indivíduo fuma; Exercío Físico, indicando a quantidade de dias em que o indicíduo pratica exercício físico ou esporte; Sal, com categorias em relação ao consumo de sal e Diabetes, indicando se o indivíduo recebeu um diagnóstico médico de diabetes ou não Todas as análises foram realizadas no software R (R Core Team, 2020).

As Tabelas 1 e 2 apresentam o resumo dos dados qualitativos e quantitativos, respectivamente. A variável dependente classifica 23,07% de indivíduos como hipertensos e 76,92% como não hipertensos. Dos indivíduos da amostra, 50,84% são do sexo feminino, 47,16% são de cor parda e 6,4% são diabéticos. A média de idade é de 43,62 anos e a

prática de exercícios físicos é de 3,4 dias em média.

Tabela 1 – Estatísticas descritivas das variáveis qualitativas

| Variável | Categorias | Frequência absoluta | Frequência relativa |
|---|---|---|---|
| Hipertenso | Sim | 7.718 | 23,07 |
| | Não | 25.739 | 76,93 |
| Sexo | Homem | 16.447 | 49,16 |
| | Mulher | 17.010 | 50,84 |
| Cor | Branca | 13.546 | 40,49 |
| | Preta | 3.615 | 10,8 |
| | Amarela | 300 | 0,9 |
| | Parda | 15.779 | 47,16 |
| | Indígena | 217 | 0,65 |
| Bebida | Não bebo nunca | 17.561 | 52,49 |
| | Menos de uma vez por mês | 4.634 | 13,85 |
| | Uma vez ou mais por mês | 11.262 | 33,66 |
| Tabagismo | Diariamente | 2.337 | 6,99 |
| | Menos que diariamente | 455 | 1,36 |
| | Não fumo atualmente | 30.665 | 91,65 |
| Consumo de sal | Muito alto | 440 | 1,32 |
| | Alto | 3.262 | 9,75 |
| | Adequado | 20.225 | 60,45 |
| | Baixo | 8.201 | 24,51 |
| | Muito baixo | 1.329 | 3,97 |
| Diabetes | Sim | 2141 | 6,4 |
| | Não | 31316 | 93,6 |

Tabela 2 – Estatísticas descritivas das variáveis quantitativas

| Variável | Média | Mediana | Desvio Padrão |
|---|---|---|---|
| Idade | 43,62 | 42 | 16,48 |
| IMC | 26,41 | 25,91 | 4,51 |
| Exercício físico | 3,4 | 3 | 1,92 |

A Figura 2 apresenta o percentual de indivíduos hipertensos da base de dados em relação às variáveis categóricas. Observa-se que o percentual de mulheres hipertensas (26,6%) é maior em relação ao percentual de homens (19,5%). Também pode-se perceber o elevado percentual de indivíduos hipertensos e diabéticos (61,5%) em relação aos hipertensos não diabéticos (20,4%).

Figura 2 – Gráficos de barras do percentual de indivíduos conforme a variável Hipertenso em relação às variáveis categóricas Sexo, Cor, Bebida, Tabagismo, Sal e Diabetes

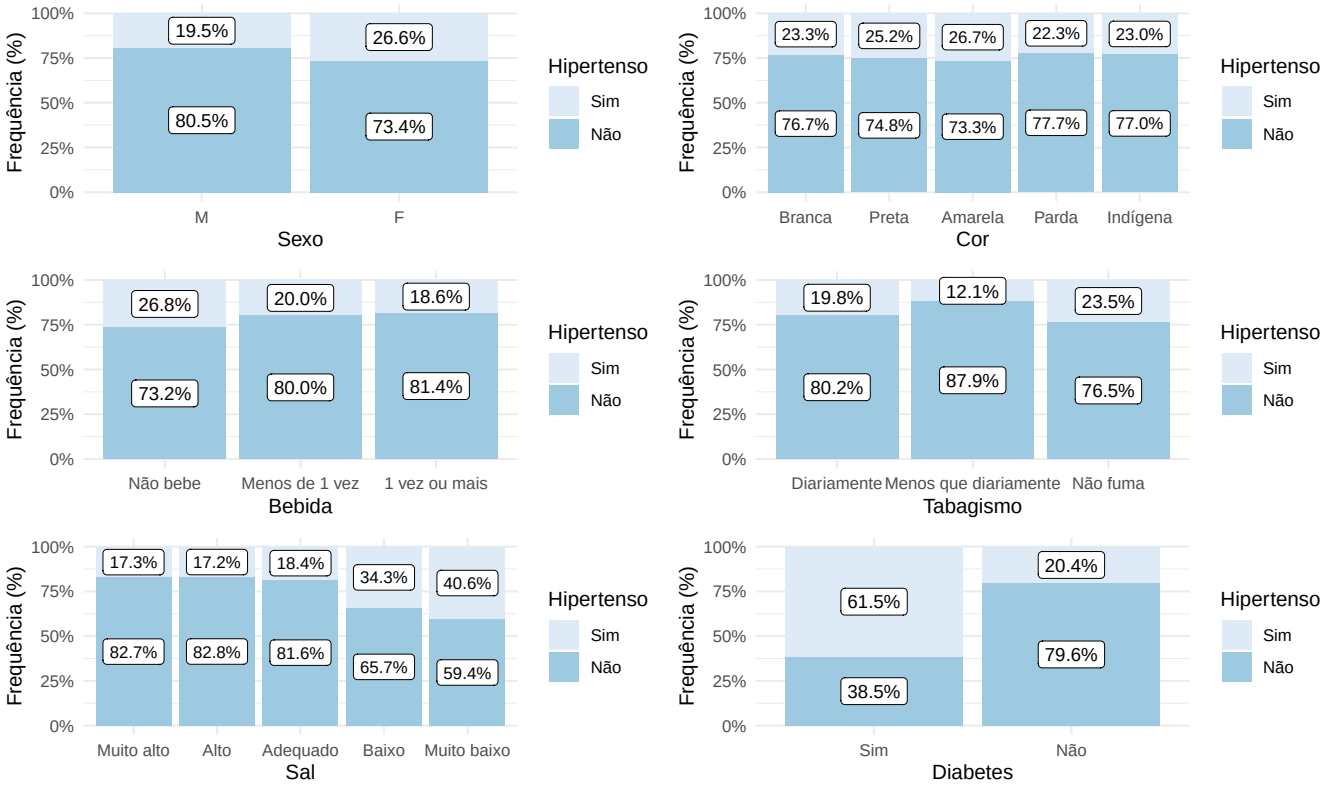
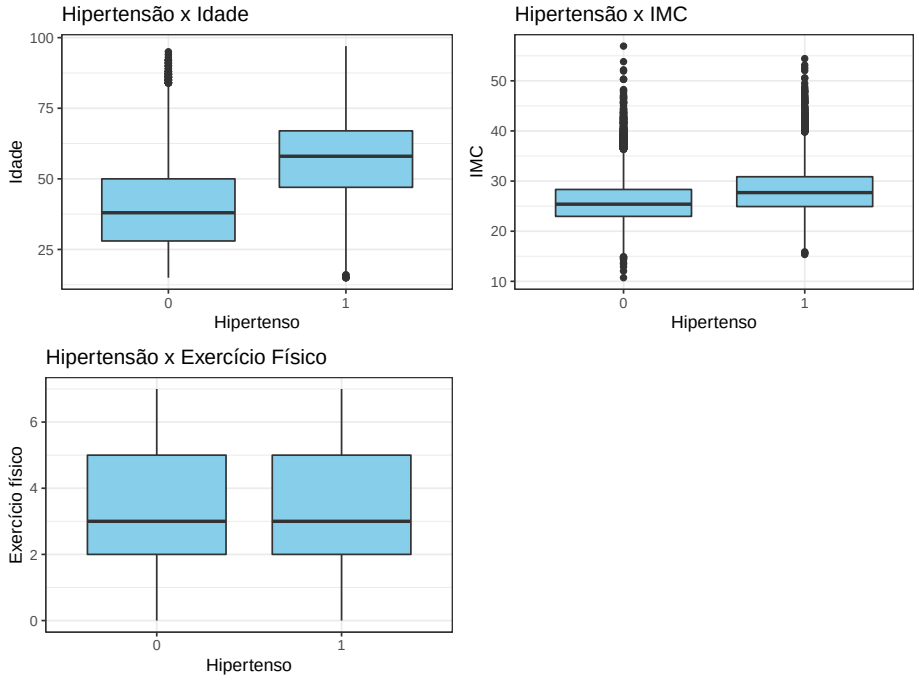



Figura 3 – Boxplot das variáveis Idade, Índice de massa corporal e Exercício Físico em relação à variável Hipertenso



A Figura 3 apresenta o comportamento das variáveis Idade, IMC e Exercício Físico de acordo com a classificação dos indivíduos pela variável Hipertenso. Destaca-se o aumento da idade para os indivíduos classificados como hipertensos, indicando uma possível associação entre as variáveis.

## 3.2 Construção do modelo

Inicialmente a base de dados original foi dividida em duas outras bases de forma aleatória. A primeira base foi denominada base de treino, possuindo 80% dos dados da base original, sendo a base utilizada para a construção dos modelos. Por sua vez, a segunda base, chamada base de teste, passou a ter 20% dos dados, sendo utilizada para a validação do modelo final. Este procedimento permite avaliar a qualidade das previsões do modelo ao testá-lo em dados não utilizados na sua construção.

Foram construídos dois modelos de regressão logística múltipla, tomando-se a variável dicotômica "Hipertenso" como variável dependente a fim de prever a probabilidade de associação de suas classes com base nas variáveis preditoras. O Modelo 1 considerou 9 variáveis independentes: Sexo, Idade, Cor, IMC, Bebida, Tabagismo, Exercício físico, Sal e Diabetes. Para o Modelo 2 foram selecionadas 6 variáveis do Modelo 1, considerando o nível de significância de 5%, sendo excluídas as variáveis Bebida, Tabagismo e Exercício Físico. Para este modelo todas as variáveis apresentaram significância ao nível de 5%. O Critério de Informação de Akaike (AIC) foi de 21.625 para o Modelo 1 e 21.621 para o Modelo 2. Tendo em vista a pequena diferença entre ambos, o Modelo 2 foi escolhido por ser um modelo mais parcimonioso.

## 3.3 Resultados

A Tabela 3 apresenta os coeficientes estimados, o erro padrão e o p-valor para o Modelo 2. Pode-se perceber que a estimativa do coeficiente da variável Idade é positivo (0,072), indicando que o aumento da idade está associada ao aumento da probabilidade de ser hipertenso, o mesmo acontece com a variável IMC (0,120). Já o coeficiente negativo para o consumo adequado de sal (-0,432) indica que os indivíduos classificados nessa categoria serão associados ao diagnóstico negativo de hipertensão.

Por meio dos coeficientes do modelo foi possível verificar a direção da relação entre as variáveis independentes e a variável resposta, porém eles não são adequados para verificar a magnitude destas relações, isto é, o quanto as probabilidades realmente variam. Sendo assim, os coeficientes exponenciados fornecem melhor interpretação para estas relações.

Tabela 3 – Coeficientes estimados, erro padrão e p-valor para as covariáveis do Modelo 2

| Variável | Estimativa | Erro padrão | p-valor |
|---|---|---|---|
| Intercepto | −8,107 | 0,209 | <,001 |
| Idade | 0,072 | 0,001 | <0,001 |
| Sexo2 | 0,268 | 0,034 | <0,001 |
| Cor2 | 0,342 | 0,058 | <0,001 |
| Cor3 | 0,351 | 0,176 | 0,046 |
| Cor4 | 0,230 | 0,037 | <0,001 |
| Cor5 | −0,283 | 0,229 | 0,217 |
| IMC | 0,120 | 0,004 | <0,001 |
| Sal2 | −0,191 | 0,170 | 0,262 |
| Sal3 | −0,432 | 0,162 | 0,008 |
| Sal4 | 0,083 | 0,163 | 0,610 |
| Sal5 | 0,135 | 0,176 | 0,444 |
| Diabetes1 | 0,866 | 0,059 | <0,001 |
| Log de propabilidade | −10.79,490 | | |
| AIC | 21.620,980 | | |

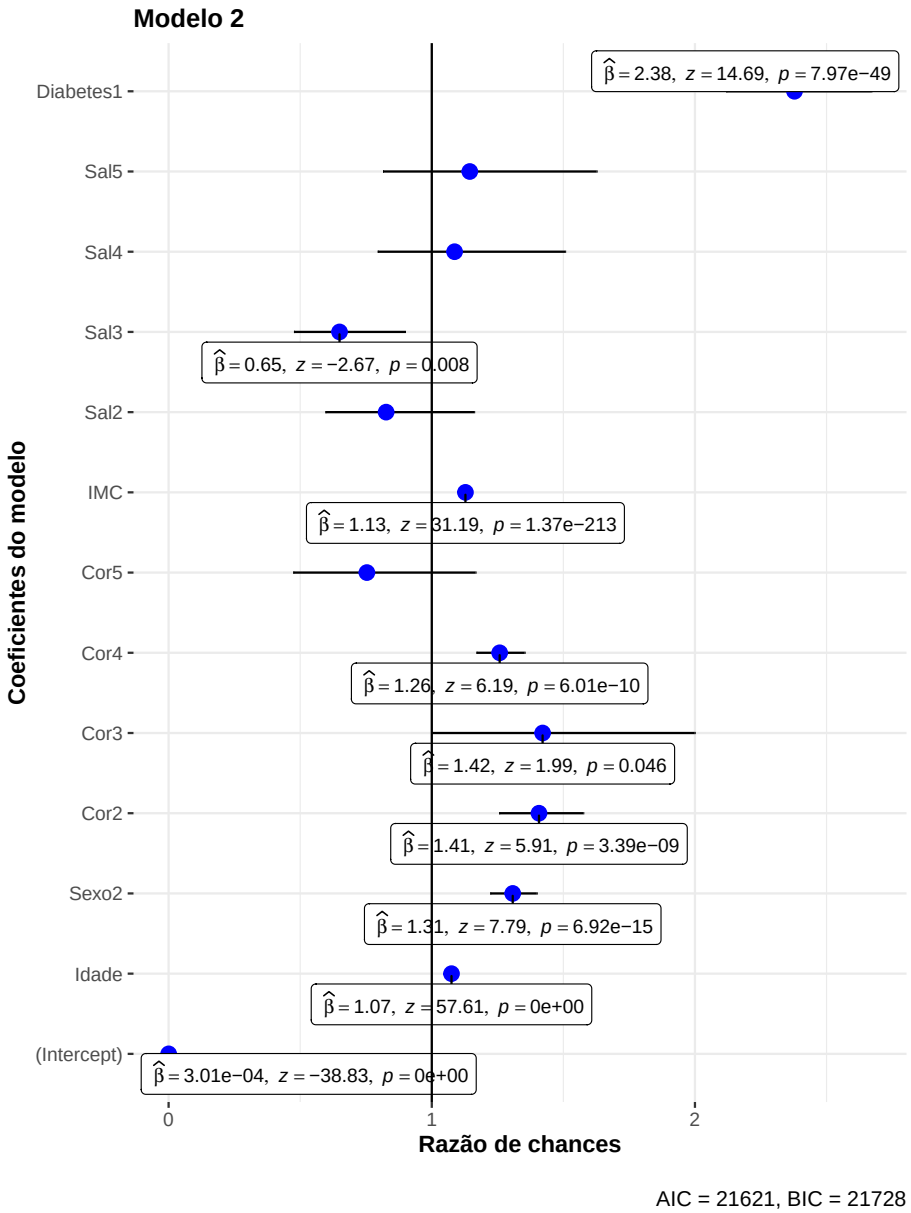
A Tabela 4 apresenta a razão de chances e o intervalo de confiança do Modelo 2. As razões de chance menores que 1 denotam relações negativas, enquanto as maiores que 1 indicam relações positivas.

Tabela 4 – Razão de chances e intervalo de confiança para as variáveis do Modelo 2

| Preditores | Razão de Chances | IC |
|---|---|---|
| (Intercepto) | 0,00 | 0,00 – 0,00 |
| Idade | 1,07 | 1,07 – 1,08 |
| Sexo [2] | 1,31 | 1,22 – 1,40 |
| Cor [2] | 1,41 | 1,26 – 1,58 |
| Cor [3] | 1,42 | 1,00 – 2,00 |
| Cor [4] | 1,26 | 1,17 – 1,35 |
| Cor [5] | 0,75 | 0,47 – 1,17 |
| IMC | 1,13 | 1,12 – 1,14 |
| Sal [2] | 0,83 | 0,60 – 1,16 |
| Sal [3] | 0,65 | 0,48 – 0,90 |
| Sal [4] | 1,09 | 0,79 – 1,51 |
| Sal [5] | 1,14 | 0,81 – 1,63 |
| Diabetes [1] | 2,38 | 2,12 – 2,67 |

Em relação às razões de chances do Modelo 2, percebe-se aumento nas chances de um indivíduo ter o diagnóstico para hipertensão positivo de 7% a cada acréscimo de 1 ano na idade. Indivíduos do sexo feminino apresentam 31% a mais de chance de ser classificados como hipertensos em relação a indivíduos do sexo masculino. Pode-se notar também que o diagnóstico de diabetes positivo aumenta em 138% as chances de classificação como hipertenso. Por outro lado, pessoas que consideram o seu consumo de sal adequado tem 35% menos chance de classificação como hipertenso em relação aos que consideram o consumo de sal muito alto.

Figura 4 – Gráfico da razão de chances, intervalo de confiança e coeficientes do modelo

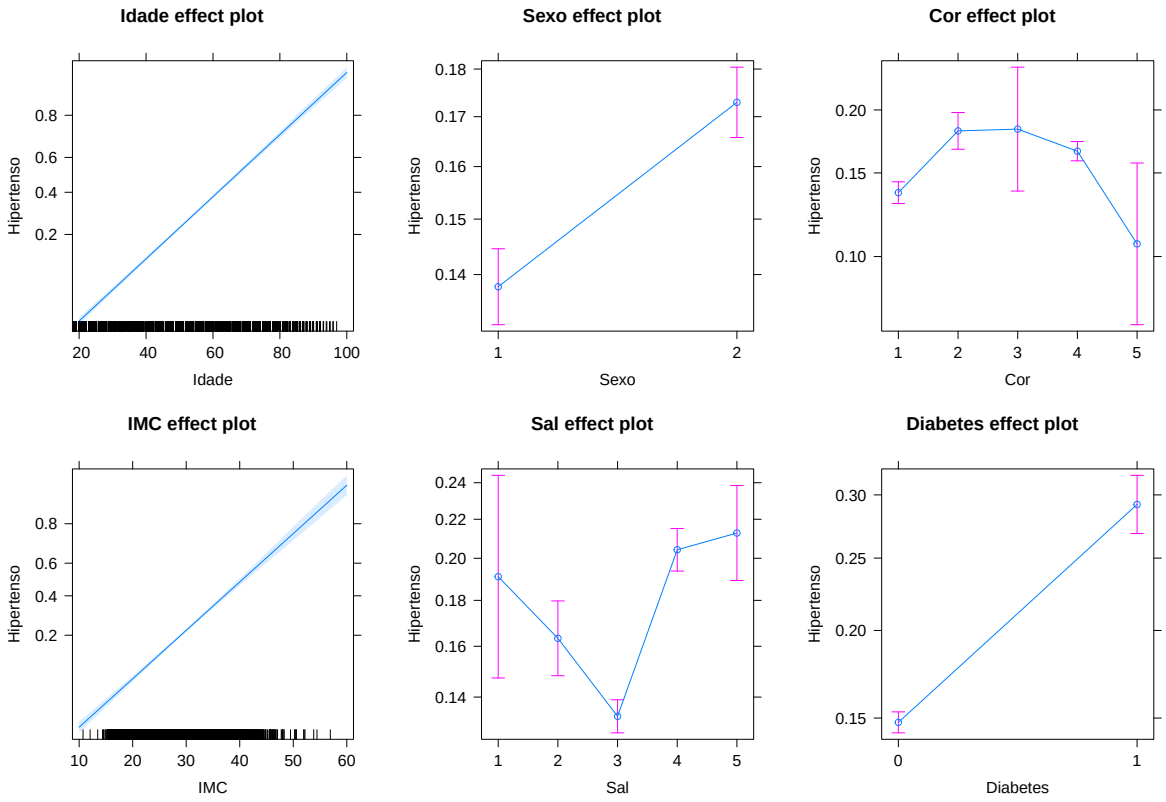



A Figura 4 apresenta graficamente a razão de chances e o intervalo de confiança para as variáveis do Modelo 2. As linhas horizontais representam o intervalo de confiança

e ao cruzarem a linha vertical indicam que a referida variável não é significativa para o modelo. Por sua vez, as linhas que permancem inteiramente dos lados esquerdo ou direito representam variáveis significativas. As variáveis significativas posicionadas ao lado direito da linha horizontal estão relacionadas ao aumento das chances de ser classificado como hipertenso pelo modelo, já as variáveis ao lado esquerdo denotam relação inversa, ou seja, diminuem as chances de classificção como hipertenso.

A Figura 5 apresenta os gráficos dos efeitos de cada covariável do modelo em relação à variável resposta. Tais gráficos proporcionam uma forma visual de observar como cada covariável afeta a variável resposta. Percebe-se que a probabilidade de diagnóstico positivo de hipertensão aumenta conforme aumento da idade e do IMC, bem como probabilidades acentuadas pra as categorias do sexo feminino, cor preta, amarela e parda e diabetes positivo. Além disso, também é possível observar a diminuição da probabilidade para a resposta "adequado" na variável consumo de sal.

Figura 5 – Gráfico dos efeitos das covariáveis Idade, sexo, Cor, Índice de massa corporal, Sal e Diabetes em função da variável resposta Hipertenso




## 3.4 Diagnósticos do Modelo

Por meio da análise dos resíduos é possível verificar se os pressupostos do modelo logístico foram atendidos. Entretanto, nos casos em que a variável resposta não tem distribuição normal, como é o caso da regressão logística, os resíduos podem, muitas

vezes não se aproximar da distribuição normal, mesmo que o modelo se ajuste bem aos dados. Nesse sentido, optou-se pela utilização dos resíduos quantílicos aleatorizados, que apresentam distribuição normal quando o modelo está adequadamente ajustado.

Na Figura 6 o gráfico superior à esquerda apresenta a densidade teórica e empírica dos resíduos quantílicos aleatorizados do Modelo 2, por meio da qual percebe-se que os resíduos estão simétricos em torno de 0, evidenciando a sua a normalidade. O gráfico quantil-quantil na parte superior à direita apresenta os quantis teóricos contra os quantis empíricos, já o gráfico probabilidade-probabilidade, na parte inferior à direita, é uma visualização que representa as probabilidades empíricas e teóricas. Em ambos os gráficos verifica-se a linearidade dos pontos, indicando também a normalidade dos resíduos.

Figura 6 – Gráficos de densidade empírica e teórica, quantil-quantil, função de distribuição cumulativa e probabilidade-probabilidade dos resíduos quantílicos aleatorizados do Modelo 2

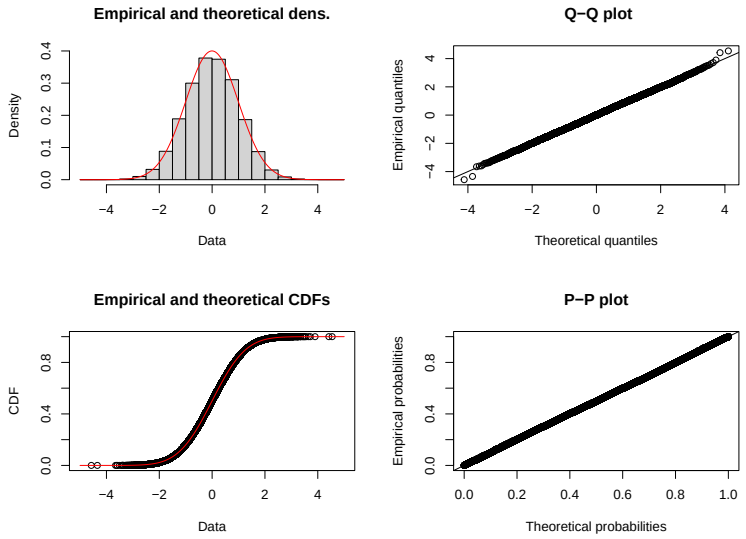
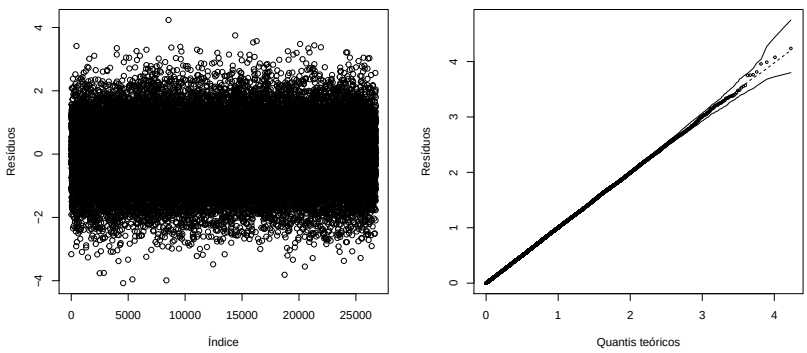



Figura 7 – Gráfico dos resíduos quantílicos aleatorizados e envelope simulado para o Modelo 2



Por meio dos gráficos dos resíduos quantílicos aleatorizados e de envelope simulado (Figura 7), observa-se que não há indícios de heterocedasticidade e que a maioria dos resíduos se encontra dentro das bandas de confiança, indicando a adequação do modelo.

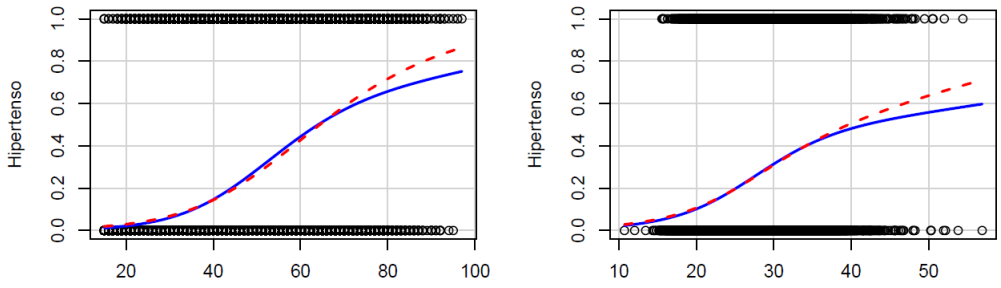
A Tabela 5 apresenta os valores para os fatores de inflação de variância generalizada do Modelo 2, por meio dos quais pode-se observar a ausência de multicolinearidade, tendo em vista que todos estão abaixo do valor 5, considerado como valor limite para a medida.

Tabela 5 – Fator de Inflação de Variância Generalizada do Modelo 2

| Variável | Fator de Inflação de Variância Generalizada |
|---|---|
| Idade | 1,173765 |
| Sexo [2] | 1,734723 |
| Cor | 1,058485 |
| Peso | 1,610961 |
| Altura | 2,216307 |
| Sal | 1,042377 |
| Diabetes | 1,023319 |

Utilizando a base de teste foram realizadas predições para o Modelo 2. Uma forma de investigar a diferença entre os valores observados e os ajustados é por meio do gráficos marginais do modelo (Figura 8). Nestes, a variável resposta está representada em função das variáveis explicativas quantitativas Idade e IMC. Os dados observados e a previsão do modelo são mostrados em linhas azuis e vermelhas, respectivamente. Percebe-se que para as variáveis representadas os valores preditos se aproximam consideravelmente dos valores observados, sendo este um indicativo de um modelo satisfatoriamente ajustado.

Figura 8 – Gráficos marginais do modelo 2




O desempenho do modelo também pode ser analisado por meio da criação de uma matriz de classificação, que represente os níveis de precisão preditiva alcançados pelo modelo logístico. A Tabela 6 apresenta a matriz de classificação do conjunto de dados de teste, com base na predição efetuada pelo Modelo 2.

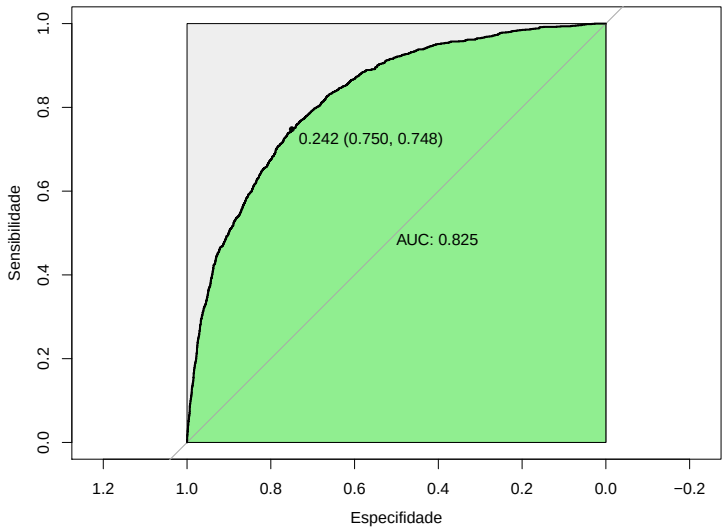
Tabela 6 – Matriz de classificação da variável Hipertenso no conjunto de dados de teste

| | | Observado | |
|---|---|---|---|
| | | Hipertenso | Não Hipertenso |
| Estimado | Hipertenso | 666 | 347 |
| | Não Hipertenso | 877 | 4.800 |

Observa-se que o modelo classificou 666 indivíduos corretamente como hipertensos, resultando em uma sensibilidade de 43,2%, medida esta que representa a proporção de verdadeiros positivos. Já o número de indivíduos corretamente classificados como não hipertensos foi de 4.800. A medida que representa a proporção de verdadeiros negativos, denominada especificidade, apresentou valor de 93,3%. Dessa forma, a acurácia do modelo, medida que representa a proporção das predições corretas sobre o total, foi de 81,7%.

A Figura 9 apresenta a curva ROC associada ao modelo, sendo também uma medida da capacidade de predição deste. Observa-se que a área sob a curva (AUC) é de 82,5%, valor que, segundo Hosmer e Lemeshow (2013), indica uma excelente capacidade preditiva do modelo.

Figura 9 – Curva de Característica de Operação do Receptor(Curva ROC) e Área Sob a Curva (AUC) conforme predições do Modelo 2




Sendo assim, estes resultados levam à conclusão de que o modelo de regressão logística construído nesta análise demonstrou validade externa e significância prática, indicando-nos a aceitação dos seus resultados.

# 4 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma aplicação da técnica da regressão logística para um conjunto de dados com algumas características dos indivíduos, a fim de analisar a associação destas características com o diagnóstico positivo ou negativo de hipertensão arterial. O modelo final contou com sete variáveis independentes e foi considerado bem ajustado conforme as técnicas de diagnóstico, como o teste de Wald e análise de resíduos, apresentando também uma acurácia aceitável para classificações.

Dentre as covariáveis que apresentaram efeitos significativos destacam se o diagnóstico positivo para diabetes, o sexo feminino, a cor preta, amarela e parda, o aumento da idade e do IMC. Tais covariáveis apresentaram relação positiva com a hipertensão e chances elevadas de que os indivíduos que apresentem tais características tenham diagnóstico positivo para hipertensão.

Sendo assim, considera-se que o modelo ajustado pode contribuir com os estudos sobre os fatores de risco associados à hipertensão arterial, bem como também pode ser explorado posteriormente com a inclusão de outras covariáveis que possam estar associadas à doença e não foram analisadas na presente pesquisa.

# REFERÊNCIAS

ALVES, J. M. S. *Dos mínimos quadrados à regressão linear: atividades históricas sobre função afim e estatística usando planilhas eletrônicas*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2016. Citado na página 12.

BARROSO, W. K. S. et al. Diretrizes brasileiras de hipertensão arterial–2020. *Arquivos Brasileiros de Cardiologia*, SciELO Brasil, v. 116, p. 516–658, 2021. Citado na página 11.

CABRAL, C. I. S. *Aplicação do modelo de regressão logística num estudo de mercado*. Tese (Doutorado), 2013. Citado na página 16.

CHRISTENSEN, R. Logistic regression, logit models, and logistic discrimination. *Log-Linear Models and Logistic Regression*, Springer, p. 116–177, 1997. Citado na página 20.

COX, D. R.; SNELL, E. J. *Analysis of binary data*. [S.l.]: Routledge, 2018. Citado 2 vezes nas páginas 12 e 17.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 19.

FAWCETT, T. *Introduction to ROC analysis, 27, 861–874*. 2006. Citado na página 18.

FIGUEIRA, C. V. Modelos de regressão logística. 2006. Citado 2 vezes nas páginas 11 e 16.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009. Citado 2 vezes nas páginas 13 e 14.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado 4 vezes nas páginas 12, 18, 19 e 30.

MARQUES, A. P. et al. Fatores associados à hipertensão arterial: uma revisão sistemática. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 25, p. 2271–2282, 2020. Citado na página 11.

MESQUITA, P. S. B. *Um modelo de Regressão Logística para Avaliação de Programas de Pós-Graduação no Brasil*. Tese (Doutorado) — Master Thesis). Universidade Estadual do Norte Fluminense, Campos dos ..., 2014. Citado 2 vezes nas páginas 12 e 17.

PEREIRA, M. A. A. Modelos não lineares assimétricos com efeitos mistos. Universidade Federal de São Carlos, 2019. Citado na página 20.

PREGIBON, D. Logistic regression diagnostics. *The annals of statistics*, Institute of Mathematical Statistics, v. 9, n. 4, p. 705–724, 1981. Citado na página 20.

SOUZA, É. C. d. *Análise de influência local no modelo de regressão logística*. Tese (Doutorado) — Universidade de São Paulo, 2006. Citado 3 vezes nas páginas 12, 18 e 20.