Research, Society and Development, v. 10, n. 16, e20101622964, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i16.22964

Aplicação da regressão logística na análise dos dados dos fatores de risco associados à hipertensão arterial

Application of logistic regression in the analysis of risk factor associated with arterial hypertension

Aplicación de la regresión logística en el análisis de factores de riesgo asociados a la hipertensión arterial

Recebido: 08/11/2021 | Revisado: 18/11/2021 | Aceito: 22/11/2021 | Publicado: 28/11/2021

Maria Beatriz Galdino da Silveira
ORCID: https://orcid.org/0000-0002-3810-4126
Universidade Estadual da Paraíba, Brasil
E-mail: maria.silveira@aluno.uepb.edu.br
Nyedja Fialho Morais Barbosa
ORCID: https://orcid.org/0000-0003-1813-320X
Universidade Estadual da Paraíba, Brasil
E-mail: nyedja@servidor.uepb.edu.br
Ana Patrícia Bastos Peixoto
ORCID: https://orcid.org/0000-0003-0690-1144
Universidade Estadual da Paraíba, Brasil
E-mail: anapatricia@servidor.uepb.edu.br
Érika Fialho Morais Xavier
ORCID: https://orcid.org/0000-0002-8217-7891
Fundação Oswaldo Cruz, Brasil
E-mail: erika.xavier@fiocruz.br
Sílvio Fernando Alves Xavier Júnior
ORCID: https://orcid.org/0000-0002-4832-0711
Universidade Estadual da Paraíba, Brasil
E-mail: silvio@servidor.uepb.edu.br

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-chave: 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 studies in the health area. This study used the logistic regression technique to analyze the association between Hypertension and certain risk factors. The data used comes from the National Health Survey (PNS) for the year 2019, carried out by the Brazilian Institute of Geography and Statistics (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 (BMI) and a positive diagnosis for diabetes increase the chances of an individual being hypertensive.
Keywords: Association; Risk factors; Fitted model.

Resumen

La regresión logística es una técnica importante para el modelado de datos cuando desea analizar la relación entre una variable de respuesta y una o más variables independientes. La técnica permite estimar las posibilidades relacionadas

con la probabilidad de que ocurra un evento de interés. La regresión logística se diferencia de la lineal por la naturaleza dicotómica de la variable dependiente y se ha utilizado en varias áreas del conocimiento, incluidos estudios en el área de la salud. Este estudio utilizó la técnica de regresión logística para analizar la asociación entre Hipertensión y ciertos factores de riesgo. Los datos utilizados provienen de la Encuesta Nacional de Salud (PNS) del año 2019, realizada por el Instituto Brasileño de Geografía y Estadística (IBGE) en el país. Se ajustaron dos modelos, estando el modelo final compuesto por siete variables con una significancia estadística del 5%. Las técnicas de diagnóstico indicaron un ajuste adecuado del modelo, así como su precisión para las predicciones. Los resultados muestran que factores como la edad avanzada, el índice de masa corporal (IMC) alto y un diagnóstico positivo de diabetes aumentan las posibilidades de que una persona sea hipertensa.
Palabras clave: Asociación; Factores de riesgo; Modelo de ajuste.

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). Vital et al (2020) concluíram que o estresse ocupacional é um dos principais fatores que afetam o trabalhador no ambiente de trabalho levando em algumas situações a hipertensão arterial. Borges et al (2020) verificaram as possíveis correlações entra a vitamina D e a pressão arterial. Nascimento et al (2021) investigaram a associação de variáveis antropométricas e hemodinâmicas à presença de HA em indivíduos com características sedentárias.

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).

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 (2013), 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 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). Os trabalhos de Cox & Snell, *Analysis od Binary Data* (2018) e Hosmer e 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 (Ahmad et al, 2014; Heo & Ryu, 2018; Andriani & Chamidah, 2019; Johnson et al., 2010), economia (Cruz & Mapa, 2013), marketing e educação (Bozpolat, 2016; Constantin, 2015; Koç & Yeniaj, 2013), sendo considerada uma importante ferramenta para análise de variáveis dicotômicas.

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.

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.

2. Metodologia

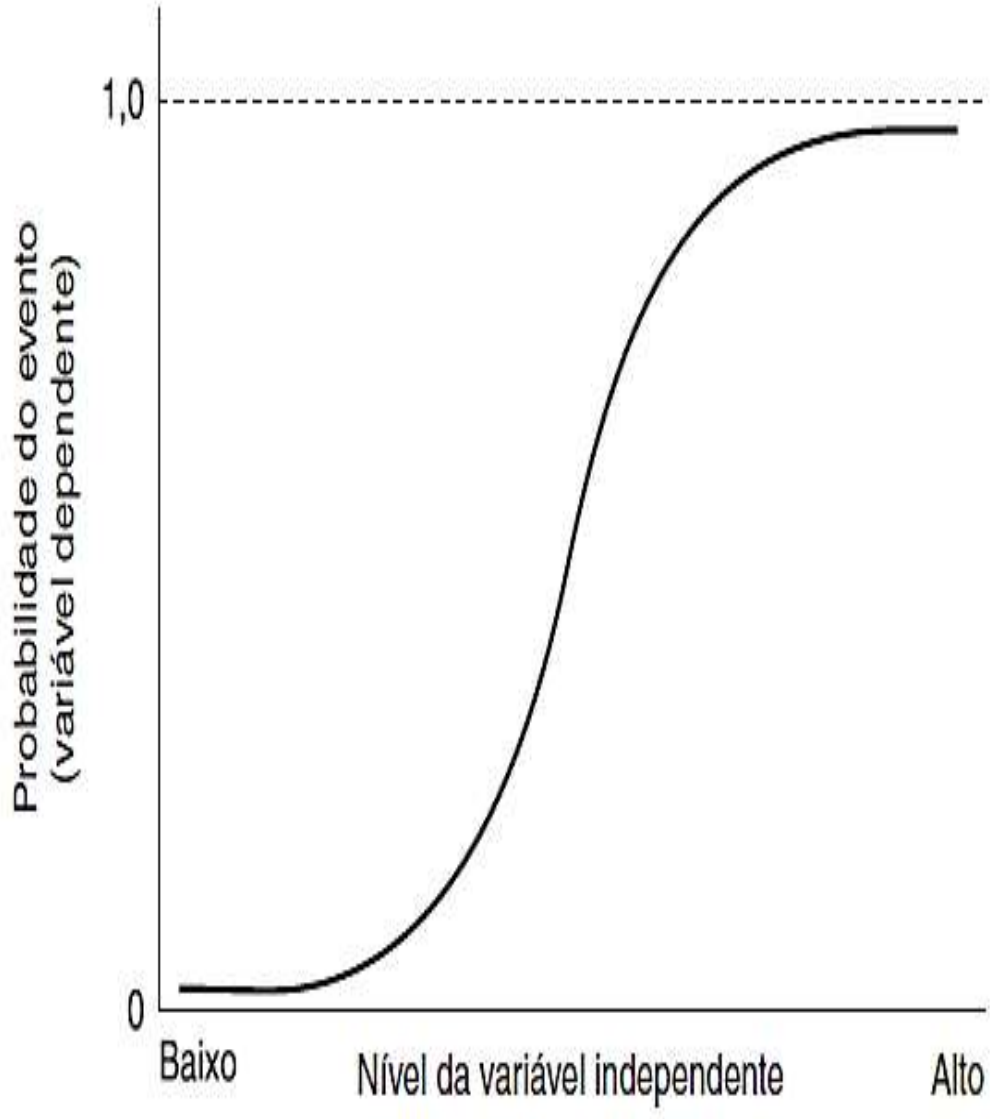
2.1 Regressão Logística

A regressão logística se diferencia dos modelos de regressão linear porque 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 $0 \leq Y \leq 1$.

De acordo com 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 pode ser visto na Figura 1.

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



Fonte: Hair Jr. et al (2009).

O modelo logístico é definido por

$$Z = \ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_k X_k, \quad (2.1),$$

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

2.2 Estimação dos Parâmetros

Nos 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 β que maximiza $L\ (x_1, x_2,..., x_n)$ para uma determinada amostra.

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

$$f(y_i, \pi_i) = \pi^{y_i}(1-\pi_i)^{1-y_i}. \quad (2.2)$$

Dessa forma, a função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi_i)^{1-y_i}, \beta \in R^2 \quad (2.3).$$

Aplicando-se o logaritmo na função de verossimilhança, tem-se que

$$\log(L(\beta)) = \log\left(\prod_{i=1}^{n} \pi^{y_i}(1-\pi_i)^{1-y_i}\right) = ... = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_1) - \log(1+\exp(\beta_0 + \beta_1 x_1)) \quad (2.4).$$

Para encontrar o valor β que maximiza a Equação (3.4) deriva-se a equação em função de β_0 e β_1.

2.3 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:

$$O.R = (p/1-p)/(q/1-q) = p(1-q)/q(1-p) \quad (2.5).$$

Considerando os parâmetros estimados por meio da regressão logística, a razão de chances é calculada exponencializando-se os coeficientes: exp (β_1) (Figueira, 2006).

2.4 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, o Pseudo R^2 de Cox e Snell e o Critério de Informação de Akaike (AIC).

2.4.1 Teste da Razão de verossimilhança

Conforme Cabral (2013), por meio desta medida, testam-se se os coeficientes de regressão associados a β são todos nulos, com exceção de β_0. Sendo assim, compara-se os valores observados e os valores esperados usando a função de verossimilhança, conforme a seguinte expressão:

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

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\pi_i}{1-y_i}\right)\right] \quad (2.6).$$

O modelo corrente refere-se 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: $H_0: \beta_1 = \ldots = \beta_t = 0$ *versus* $H_1: \exists_{j=1,\ldots,p}\, \beta_j \neq 0$.

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.

2.4.2 Teste de Wald

O teste de Wald verifica 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:

$$W_j = \frac{\widehat{\beta_j}}{var\left(\widehat{\beta_j}\right)} \quad (2.7).$$

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.

2.4.3 Pseudo R² de Cox e Snell

Esta medida é denominada pseudo R² pelo fato de apresentar semelhanças com o R² dos modelos de regressão linear. Entretanto, ressalta-se que apesar da similaridade, a interpretação de ambos é diferente. Existem muitas maneiras de calcular o pseudo R², sendo o pseudo R² de Cox & Snell (2018) um dos mais frequentemente usados pelos softwares estatísticos. A medida é definida por:

$$R^2 = 1 - \left(\frac{L(\beta)_0}{L(\beta)_M}\right)^{\frac{2}{n}} \quad (2.8),$$

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 R² mais elevado.

2.4.4 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:

$$AIC = -2\ln(L_p) + 2\left[(p+1)+1\right] \quad (2.9),$$

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

2.4.5 Curva ROC (Receiver Operating Characteristic)

A Curva ROC avalia 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 & Lemershow (2013) deve ser considerado aceitável acima de 0,7.

2.5 Diagnóstico 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.

2.5.1 Diagonal da Matriz H

Utilizam-se os elementos da matriz **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 $Var(\varepsilon_i) = \pi_i(1-\pi_i)$, não é constante, a matriz de projeção para o modelo logístico, utilizando a definição de mínimos quadrados ponderados é dada por:

$$H = Q^{\frac{1}{2}} X (X^T Q X)^{-1} X^T Q^{\frac{1}{2}} \quad (2.10).$$

Tal matriz sugere a utilização dos elementos da diagonal principal de **H** para detectar a presença de pontos de alavanca no modelo. É importante salientar que conforme Hosmer & Lemeshow (1989, apud Souza, 2006) 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 $\hat{\mathbf{H}}$ é dada por:

$$\hat{h}_{ii} = \hat{\pi}_i(1-\hat{\pi}_i) x_i^T [I(\beta)]^{-1} (x_i); i = 1,2,\ldots,n. \quad (2.11)$$

2.5.2 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:

$$r_i = y_i - \hat{\pi}_i \quad (2.12).$$

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:

$$(rp) = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}; i = 1,2,\ldots,n \quad (2.13).$$

2.5.3 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:

$$d_i = \begin{cases} -\sqrt{2\ln(1-\hat{\pi}_i)} & se\ y_i = 0 \\ \pm\sqrt{2\left[y_i \ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right]} & se\ 0 < y_i < 1 \\ \sqrt{2\ln(\hat{\pi}_i)} & se\ y_i = 1 \end{cases} \quad (2.14).$$

2.5.4 Resíduo Quantílico Aleatorizado

Os resíduos quantílicos aleatorizados foram propostos por Dunn & 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 a de Cox & Snell (2018), 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:

$$rq_{ij} = \Phi^{-1}\left(G_Y\left(y_{ij} : f\left(\hat{\alpha}, x_{ij}, \hat{\theta}\right)\right)\right), i = 1, \ldots, n;\ j = 1, \ldots, m_i, \quad (2.15)$$

onde Φ e G_Y são as funções de distribuição acumuladas da distribuição normal padrão e da distribuição considerada no ajuste, respectivamente.

2.5.5 C e Cbar ($\overline{C}$)

Avaliam a influência das observações individuais sobre β, 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:

$$C_i = \frac{(r\,p_i)^2 h_{ii}}{(1-h_{ii})^2};\ i = 1,2,\ldots,n \quad (2.16).$$

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

$$\overline{C}_i = \frac{(r\,p_i)^2 h_{ii}}{(1-h_{ii})};\ i = 1,2,\ldots,n \quad (2.17).$$

2.5.6 DIFCHISQ e DIFDEV

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:

$$DIFCHISQ_i = \frac{\overline{C}_i}{h_{ii}} = \frac{(r\,p_i)^2}{1-h_{ii}};\ i = 1,2,\ldots,n \quad (2.18).$$

A DIFDEV baseia-se no resíduo da *deviance*, é definida por:

$$DIFDEV_i = d_i^2 + \overline{C}_i = d_i^2 + \frac{(r\,p_i)^2}{h_{ii}(1-h_{ii})}; i = 1,2,\ldots,n \quad (2.19).$$

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. Resultados e Discussão

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 relizou-se 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 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) calculada através 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; 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		**Frequência absoluta**	**Frequência relativa(%)**
Hipertenso	Sim	7718	23,07
	Não	25739	76,93
Sexo	Homem	16447	49,16
	Mulher	17010	50,84
Cor	Branca	13546	40,49
	Preta	3615	10,8
	Amarela	300	0,9
	Parda	15779	47,16
	Indígena	217	0,65
Bebida	Não bebo nunca	17561	52,49
	Menos de uma vez por mês	4634	13,85
	Uma vez ou mais por mês	11262	33,66
Tabagismo	Diariamente	2337	6,99
	Menos que diariamente	455	1,36
	Não fumo atualmente	30665	91,65
Consumo de sal	Muito alto	440	1,32
	Alto	3262	9,75
	Adequado	20225	60,45
	Baixo	8201	24,51
	Muito baixo	1329	3,97
Diabetes	Sim	2141	6,4
	Não	31316	93,6

Fonte: Autores (2021).

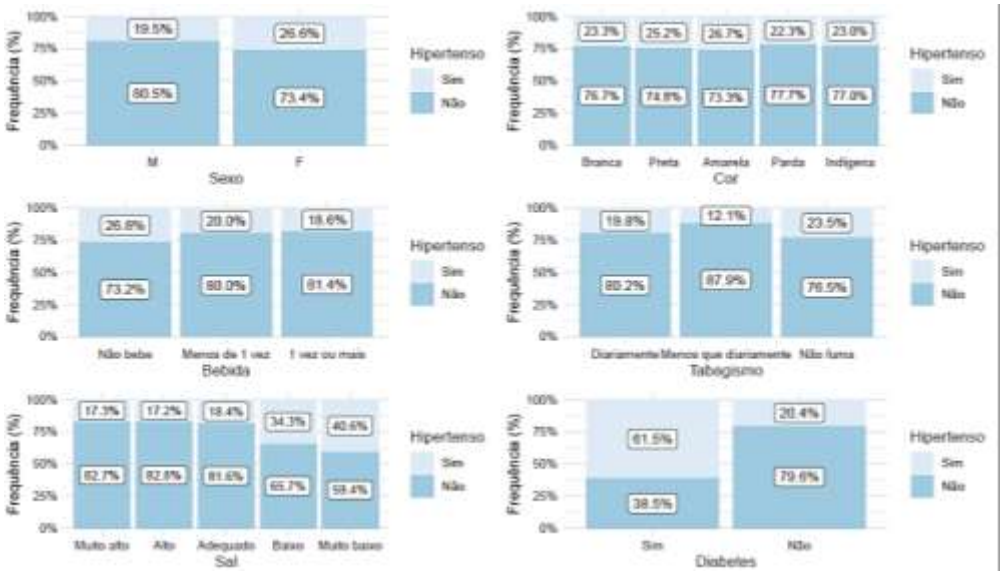
Tabela 2 - Estatísticas descritivas das variáveis quantitativas

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

Fonte: Autores (2021).

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%).

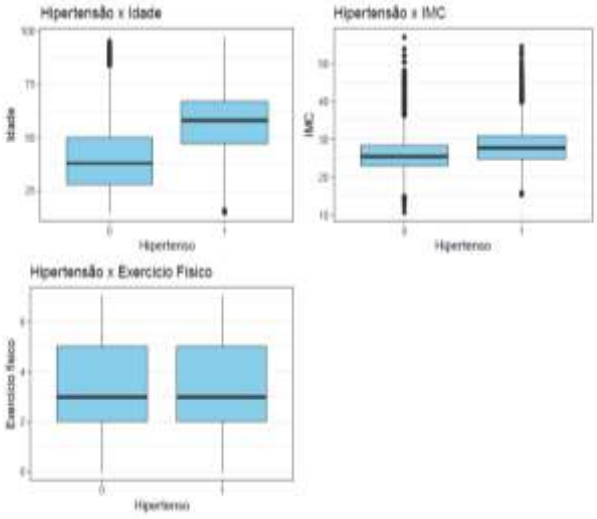
Figura 2 - Gráficos de barras do percentual de indivíduos conforme a variável Hipertenso em relação às variáveis categóricas Sexo, Cor, Bebida, Tabagismo, Sal e Diabetes.



Fonte: Autores (2021).

Na Figura 3 tem-se 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.

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



Fonte: Autores (2021).

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.

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 positiva (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 verificou-se 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	-8,11	0,21	0
Idade	0,07	0,01	0
Sexo2	0,27	0,03	0
Cor2	0,34	0,06	0
Cor3	0,35	0,18	0,05
Cor4	0,23	0,04	0
Cor5	-0,28	0,23	0,22
IMC	0,12	0	0
Sal2	-0,19	0,17	0,26
Sal3	-0,43	0,16	0,01
Sal4	0,08	0,16	0,61
Sal5	0,14	0,18	0,44
Diabetes1	0,87	0,06	< 0,001
Log - probabilidades	0		
AIC	21620,98		

Fonte: Autores (2021).

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.

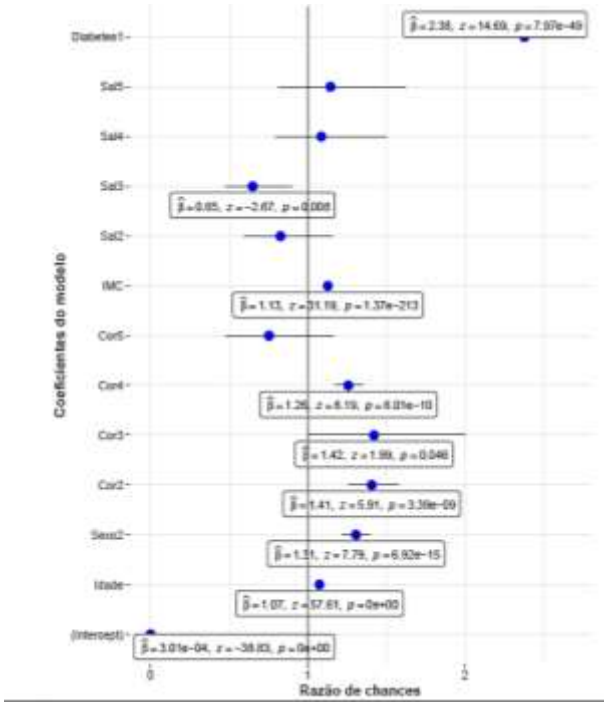
Preditores	Razão de Chances	Intervalo de Confiança (IC)
(Intercepto)	0	0
Idade	1,07	0
Sexo2	1,31	0
Cor2	1,41	0
Cor3	1,42	0
Cor4	1,26	0
Cor5	0,75	0
IMC	1,13	0
Sal2	0,83	0
Sal3	0,65	0
Sal4	1,09	0
Sal5	1,14	0
Diabetes1	2,38	0

Fonte: Autores (2021).

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.

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 permanecem 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 classificação como hipertenso.

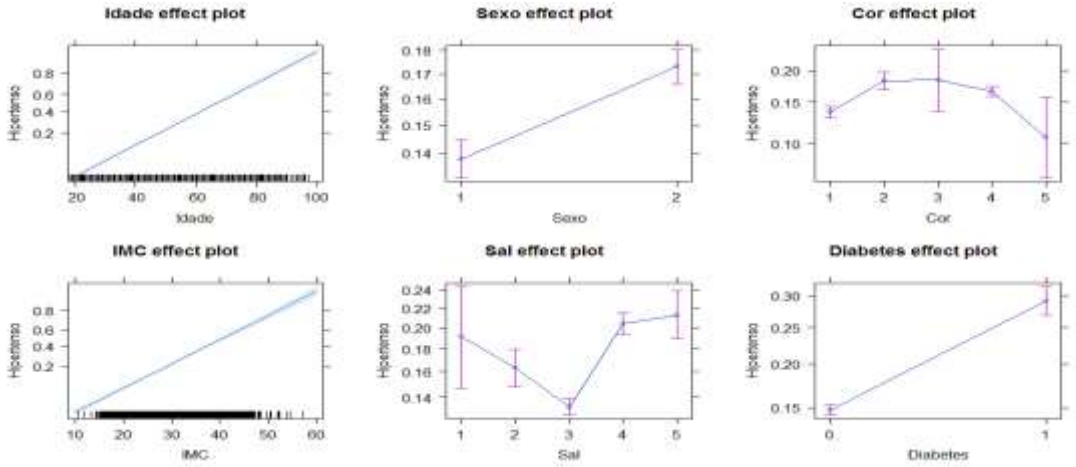
Figura 4 - Gráfico da razão de chances, intervalo de confiança e coeficientes do modelo.



Fonte: Autores (2021).

A Figura 5 apresenta os gráficos dos efeitos de cada covariável do modelo em relação à 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.

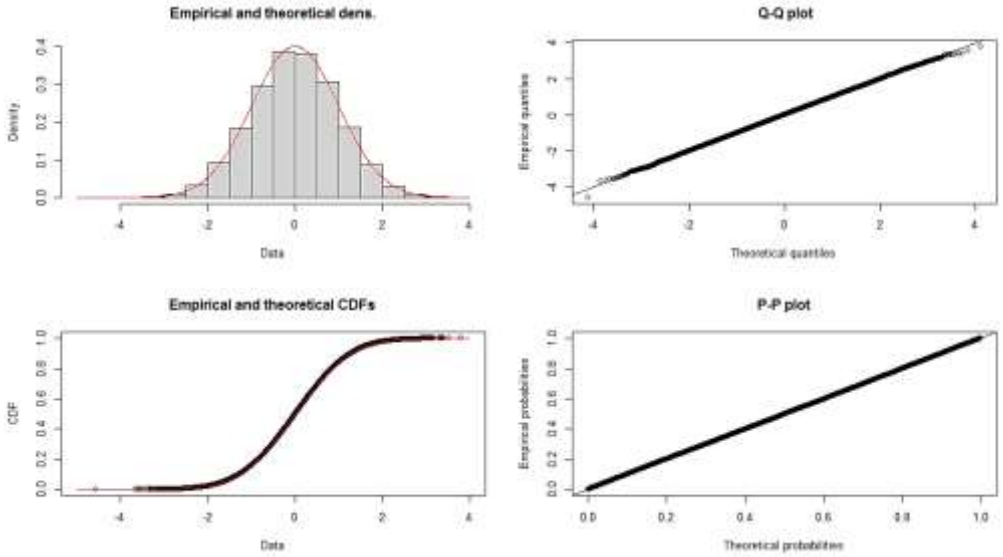


Fonte: Autores (2021).

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.

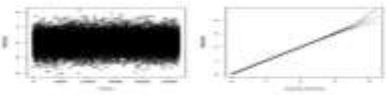
Figura 6 - Gráficos de densidade empírica e teórica, quantil-quantil, função de distribuição cumulativa e probabilidade-probabilidade dos resíduos quantílicos aleatorizados do Modelo 2.



Fonte: Autores (2021).

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 encontra-se dentro das bandas de confiança, indicando a adequação do modelo.

Figura 7- Gráfico dos resíduos quantílicos aleatorizados e envelope simulado para o modelo.



Fonte: Autores (2021).

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.

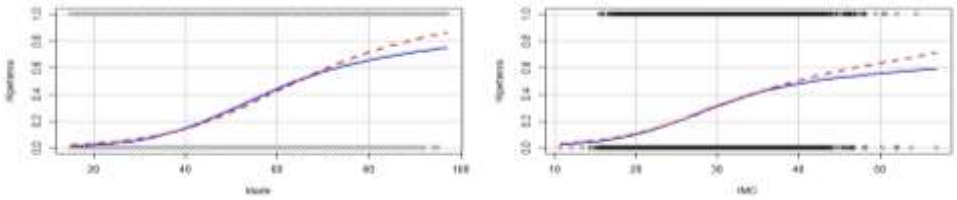
Tabela 5 – Fator de Inflação de Variância Generalizada do Modelo 2.

Variável	Fator de Inflação de Variância Generalizada
Idade	1,173765
Sexo2	1,734723
Cor	1,058485
Peso	1,610961
Altura	2,216307
Sal	1,042377
Diabetes	1,023319

Fonte: Autores (2021).

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.



Fonte: Autores (2021).

O desempenho do modelo 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

Fonte: Autores (2021).

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.

Fonte: Autores (2021).

4. Conclusão

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. Futuras pesquisas podem explorar a inclusão de outras covariáveis que possam estar associadas à doença e não foram analisadas na presente pesquisa.

Agradecimentos

Os autores agradecem aos revisores anônimos pelas sugestões e contribuições no processo de revisão do artigo.

Referências

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

Ahmad, W. M. A. W., Nawi, M. A. B. A., Aleng, N., Halim, N., Mamat, M., Hamzah, M., & Ali, Z. (2014). Association of hypertension with risk factors using logistic regression. *Applied Mathematical Sciences*, *8*(52), 2563-2572.

Andriani, P., & Chamidah, N. (2019, August). Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia. In *Journal of Physics: Conference Series* 1306(1), 012027. IOP Publishing.

Barroso, W. K. S., Rodrigues, C. I. S., Bortolotto, L. A., Mota-Gomes, M. A., Brandão, A. A., Feitosa, A. D. D. M., ... & Nadruz, W. (2021). Diretrizes Brasileiras de Hipertensão Arterial–2020. *Arquivos Brasileiros de Cardiologia*, *116*, 516-658.

Borges, A. C. do N., Bezerra, J. B., Alencar, V. Y. C., Silva, K. de A., Costa, A. A. A., Oliveira, B. G. S., Costa, A. L., Portela, J. V. F., & Bezerra, F. das C. L. (2020). Vitamin D linked to high blood pressure. *Research, Society and Development*, *9*(1), e110911691. https://doi.org/10.33448/rsd-v9i1.1691

Bozpolat, E. (2016). Investigation of the Self-Regulated Learning Strategies of Students from the Faculty of Education Using Ordinal Logistic Regression Analysis. *Educational Sciences: Theory and Practice*, *16*(1), 301-318.

Cabral, C. I. S. (2013). *Aplicação do modelo de regressão logística num estudo de mercado* (Dissertação de Mestrado). Universidade de Lisboa, Lisboa, Portugal.

Christensen, R. (1997). Logistic Regression, Logit Models, and Logistic Discrimination. *Log-Linear Models and Logistic Regression*, 116-177.

Constantin, C. (2015). Using the Logistic Regression model in supporting decisions of establishing marketing strategies. *Bulletin of the Transilvania University of Brasov. Economic Sciences. Series 8*(2), 4.

Cox, D. R., & Snell, E. J. (2018). *Analysis of binary data*. Routledge.

Cruz, C. J. F., & Mapa, D. (2013). An early warning system for inflation in the Philippines using Markov-switching and logistic regression models. *Theoretical and Practical Research in Economic Fields*, *2*, 137-152.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236-244.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Figueira, C. V. (2006). *Modelos de regressão logística* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora.

Heo, B. M., & Ryu, K. H. (2018). Prediction of Prehypertenison and Hypertension Based on Anthropometry, Blood Parameters, and Spirometry. *International journal of environmental research and public health*, *15*(11), 2571. https://doi.org/10.3390/ijerph15112571.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Johnson, S., Corsten, M. J., McDonald, J. T., & Chun, J. (2010). Socio-economic factors and stage at presentation of head and neck cancer patients in Ottawa, Canada: A logistic regression analysis. *Oral oncology*, *46*(5), 366-368.

Koç, A. A., & Yeniay, Ö. (2013). A comparative study of artificial neural networks and logistic regression for classification of marketing campaign results. *Mathematical and Computational Applications*, *18*(3), 392-398.

Marques, A. P., Szwarcwald, C. L., Pires, D. C., Rodrigues, J. M., Almeida, W. D. S. D., & Romero, D. (2020). Fatores associados à hipertensão arterial: uma revisão sistemática. *Ciência & Saúde Coletiva*, *25*, 2271-2282.

Mesquita, P. S. B. (2014). *Um modelo de Regressão Logística para Avaliação de Programas de Pós-Graduação no Brasil* (Dissertação de Mestrado). Universidade Estadual do Norte Fluminense, Campos dos Goytacazes, Brasil.

Nascimento, R. L. do, Carvalho, F. O.., Araujo, F. de S.., Melo-Marins, D. de., Carneiro, M. V. O., Saraiva, L. C., Moreira, S. R., & Nascimento Junior, J. R. A. (2021). Anthropometric and hemodynamic indicators associated with arterial hypertension in sedentary people. *Research, Society and Development*, *10*(7), e25310716603. https://doi.org/10.33448/rsd-v10i7.16603.

Pereira, M. A. A. (2019). *Modelos não lineares assimétricos com efeitos mistos* (Tese de Doutorado). Universidade Federal de São Carlos, São Carlos, Brasil.

Pregibon, D. (1981). Logistic regression diagnostics. *The annals of statistics*, *9*(4), 705-724.

Souza, É. C. D. (2006). *Análise de influência local no modelo de regressão logística* (Tese de Doutorado). Universidade de São Paulo, São Paulo, Brasil.

Vital, T. G., Silva, I. de O., & Paz, F. A. do N. (2020). Arterial hypertension and work-related risk factors: a literature review. *Research, Society and Development*, *9*(7), e905975085. https://doi.org/10.33448/rsd-v9i7.5085.