UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

FRANCISCO WELLINGTON PINTO DE SOUSA

FATORES QUE INFLUENCIAM A QUALIDADE DE VIDA DOS IDOSOS POR MEIO DA REGRESSÃO LOGÍSTICA: UM ESTUDO DE CASO NO ESTADO DO CEARÁ

FORTALEZA

2020

FRANCISCO WELLINGTON PINTO DE SOUSA

FATORES QUE INFLUENCIAM A QUALIDADE DE VIDA DOS IDOSOS POR MEIO DA REGRESSÃO LOGÍSTICA: UM ESTUDO DE CASO NO ESTADO DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. João Welliandre Carneiro Alexandre

FORTALEZA

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S696f Sousa, Francisco Wellington Pinto de.
Fatores que influenciam a qualidade de vida dos idosos por meio da regressão logística : um estudo de caso no Estado do Ceará / Francisco Wellington Pinto de Sousa. – 2020.
80 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Estatística, Fortaleza, 2020.
Orientação: Prof. Dr. João Welliandre Carneiro Alexandre.

1. Regressão logística. 2. Qualidade de vida. 3. Idosos . I. Título.

CDD 519.5

FRANCISCO WELLINGTON PINTO DE SOUSA

FATORES QUE INFLUENCIAM A QUALIDADE DE VIDA DOS IDOSOS POR MEIO DA REGRESSÃO LOGÍSTICA: UM ESTUDO DE CASO NO ESTADO DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em: 12/11/2020

BANCA EXAMINADORA

Prof. Dr. João Welliandre Carneiro Alexandre (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Sílvia Maria de Freitas
Universidade Federal do Ceará (UFC)

Profa. Dra. Maria Jacqueline Batista
Universidade Federal do Ceará (UFC)

A todas as pessoas que fazem e fizeram parte da minha vida, e que de alguma maneira colaboraram com o meu crescimento e evolução pessoal ao longo desta caminhada.

AGRADECIMENTOS

Aos meus pais, Antônio Martins de Sousa e Maria Eridan Pinto de Oliveira *(in memorian)*, que apesar de todas as dificuldades vividas, sempre acreditaram no meu potencial e me motivaram a seguir neste caminho.

Aos amigos e amigas que fiz durante a graduação. Sem a ajuda de vocês em vários momentos desta caminhada, eu não teria conseguido concluir esta graduação. Vocês são as pessoas mais maravilhosas e inteligentes que eu conheço.

Aos professores que tive ao longo deste período, que em vários momentos me estimularam a sair da “zona de conforto" e ir em busca de evolução constante.

Ao Prof. Dr. João Welliandre Carneiro Alexandre, por me orientar nesta monografia e por toda sua paciência durante o desenvolvimento deste trabalho.

“O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis."

(José de Alencar)

RESUMO

O processo de envelhecimento da população é um fenômeno que vem ocorrendo nas últimas décadas de maneira acelerada em diversos países do mundo, inclusive o Brasil. É uma questão relevante em função das necessidades geradas por esse processo, que precisam ser cobertas por políticas em diversas áreas, que devem levar em consideração aspectos importantes na construção dos fatores que têm impacto na expectativa de vida, na saúde e na independência, fatores ligados à qualidade de vida dos idosos. Neste sentido, o objetivo deste estudo é investigar, por meio do uso do método de regressão logística, quais fatores possuem maior impacto na qualidade de vida dos idosos atendidos pelo Ambulatório de Geriatria do Hospital Universitário Walter Cantídio (HUWC), vinculado à Universidade Federal do Ceará (UFC). Os dados utilizados nessa monografia foram obtidos de maneira secundária, a partir da dissertação de Carvalho (2019), em que houve a aplicação de um questionário baseado em um estudo internacional acerca da qualidade de vida dos idosos, feito em doze países latino-americanos, incluindo o Brasil, em que o estado do Ceará foi incluído. Neste estudo, 95 idosos atendidos pelo HUWC-UFC responderam a um questionário contendo 26 questões do *World Health Organization Quality of Life Questionnaire* (WHOQOL), instrumento de avaliação da qualidade de vida na terceira idade proposto pela Organização Mundial de Saúde. Foi verificado, através do p-valor obtido, a adequabilidade do modelo proposto, e após isto, feita uma análise diagnóstica e, com o auxílio da curva ROC, os idosos foram classificados de acordo com a sua qualidade de vida em Bom ou Ruim. Os resultados obtidos indicaram que um idoso que convive em um ambiente saudável, está bem-informado do que acontece ao seu redor, está satisfeito com sua capacidade física e suas relações sociais e pessoais possuem maior chance de estar satisfeito com sua qualidade de vida.

Palavras-chave: qualidade de vida; estatística; idosos; regressão logística.

ABSTRACT

The population aging process is a phenomenon that has been occurring in the last decades in an accelerated way in several countries of the world, including Brazil. It is a relevant issue due to the needs generated by this process, which need to be covered by policies in several areas, which must take into account important aspects in the construction of factors that have an impact on life expectancy, health and independence, factors linked the quality of life of the elderly. In this sense, the objective of this study is to investigate, through the use of the logistic regression method, which factors have the greatest impact on the quality of life of the elderly assisted by the Geriatrics Outpatient Clinic of the Walter Cantídio University Hospital, linked to the Federal University of Ceará. The data used in this monograph were obtained in a secondary way, from Carvalho's dissertation (2019), in which there was the application of a questionnaire based on an international study on the quality of life of the elderly, carried out in twelve Latin American countries, including Brazil, in which the state of Ceará was included. In this study, 95 elderly people attended by the HUWC-UFC answered a questionnaire containing 26 questions from WHOQOL, an instrument for assessing quality of life in old age proposed by the World Health Organization. The adequacy of the proposed model was verified through the p-value obtained, and after that, a diagnostic analysis was made and, with the aid of the ROC curve, the elderly were classified according to their quality of life in Good or Bad. The results obtained indicated that an elderly person who lives in a healthy environment, is well informed of what happens around them, is satisfied with their physical capacity and their social and personal relationships have a greater chance of being satisfied with their quality of life.

Keywords: quality of life; statistics; seniors; logistic regression.

LISTA DE FIGURAS

Figura 1 – Demonstrativo da evolução da expectativa de vida no Brasil. 18
Figura 2 – Demonstrativo da expectativa de vida por estado brasileiro. 19
Figura 3 – Pirâmide Etária do Brasil projetada para 2020, comparada à 2010. 19
Figura 4 – Pirâmide Etária do Brasil projetada para 2020, dividida por gêneros. 20
Figura 5 – Gráfico quantil-quantil Normal com envelope simulado sob a suposição de componente aleatória binomial para o modelo proposto com as 73 observações. 49
Figura 6 – Gráficos para a análise diagnóstica do modelo. 49
Figura 7 – Curva ROC obtida para o modelo em estudo. 54
Figura 8 – Curva ajustada do modelo final adotado. 57

LISTA DE TABELAS

Tabela 1 – Tabela de classificação Ilustrativa. 27
Tabela 2 – Informações retratadas em uma matriz de confusão. 37
Tabela 3 – Resultados obtidos no teste χ^2 para as variáveis avaliadas. 42
Tabela 4 – Resultados do ajuste dos modelos logísticos simples. 44
Tabela 5 – Resultados do ajuste do modelo logístico múltiplo. 45
Tabela 6 – Lista dos melhores modelos selecionados, segundo o critério do AIC. 46
Tabela 7 – Comparação entre os valores obtidos aplicando os critérios de seleção de variáveis aos modelos selecionados via `bestglm`. 46
Tabela 8 – Estatísticas de ajuste dos modelos 1 e 4. 47
Tabela 9 – Resultados do ajuste do modelo 1 selecionado na Tabela 6. 47
Tabela 10 – Resultados do Teste de Hosmer-Lemeshow para o modelo proposto. 48
Tabela 11 – Comparação entre frequências observadas e esperadas do modelo selecionado. 48
Tabela 12 – Valores de $\hat{\pi}$, h, LD e d das seis observações destacadas. 50
Tabela 13 – Coeficientes dos modelos ajustados sem as observações destacadas individualmente. 50
Tabela 14 – Tabela de classificação para o modelo com ponto de corte em 0,5. 52
Tabela 15 – Tabela de classificação para o modelo com ponto de corte em 0,5651. 53
Tabela 16 – Estimativas dos parâmetros do modelo adotado. 55

LISTA DE ABREVIATURAS E SIGLAS

HUWC	Hospital Universitário Walter Cantídio
UFC	Universidade Federal do Ceará
WHOQOL	*World Health Organization Quality of Life Questionnaire*
IBGE	Instituto Brasileiro de Geografia e Estatística
OMS	Organização Mundial da Saúde
SBGG	Sociedade Brasileira de Geriatria e Gerontologia
AIC	Critério de Informação de Akaike
AICc	Critério de Informação de Akaike Corrigido
BIC	Critério de Informação Bayesiano
ROC	*Receiver Operating Characteristic*
AUC	*Area Under the Curve*

LISTA DE SÍMBOLOS

R^2	Coeficiente de Determinação
χ^2	Qui-Quadrado
ln	Logaritmo natural
n	Número de observações
q	Número de variáveis
p	Probabilidade
$\mathbb{P}$	Função Probabilidade
Bin	Distribuição Binomial
β	Parâmetros do modelo
α	Nível de Significância
$\hat{\pi}$	Probabilidade Estimada
$\mathscr{H}_0$	Hipótese Nula
$\mathscr{H}_1$	Hipótese Alternativa

SUMÁRIO

1 INTRODUÇÃO 15
2 REFERENCIAL TEÓRICO 18
2.1 Panorama do envelhecimento da população brasileira 18
2.1.1 Conceituação da Qualidade de Vida 20
2.1.2 Fatores que afetam a qualidade de vida na Terceira Idade 21
2.1.3 Políticas Públicas para a Terceira Idade 21
3 PROCEDIMENTO METODOLÓGICO 23
3.1 Descrição do Estudo 23
3.1.1 Codificação das Respostas 25
3.2 Tratamento de dados faltantes 26
3.3 Teste Qui-Quadrado 27
3.4 Regressão Logística 28
3.5 Estimação dos parâmetros 29
3.6 Critério de Seleção de Variáveis 31
3.6.1 Critério de Informação de Akaike 31
3.6.2 Critério de Informação de Akaike Corrigido 32
3.6.3 Critério de Informação Bayesiano 32
3.7 Estatística *Deviance* 33
3.8 Coeficientes de Determinação 33
3.9 Outras estatísticas de ajuste do modelo 35
3.9.1 Erro Quadrático Médio 35
3.9.2 Coeficiente de Variação 35
3.10 Teste de Wald 36
3.11 Teste de Hosmer e Lemeshow 36
3.12 Classificação e Curva ROC 37
3.13 Razão de Chances 39
3.14 Resíduos 39
4 APLICAÇÃO E RESULTADOS 42
4.1 Testes de Independência 42
4.2 Regressão Logística Simples 43

4.3 **Regressão Logística Múltipla** 45
4.4 **Seleção de Variáveis** 45
4.5 **Adequabilidade do Modelo** 47
4.5.1 ***Análise Diagnóstica*** 48
4.5.2 ***Classificação*** 52
4.5.3 ***Interpretação*** 54
5 **CONCLUSÕES** 58
5.1 **Considerações Finais** 58
5.2 **Sugestões para Trabalhos Futuros** 59
REFERÊNCIAS 60
APÊNDICE A–CÓDIGOS NO R 63
ANEXO A–WHOQOL ABREVIADO 77

1 INTRODUÇÃO

Este trabalho tem por temática uma investigação dos fatores que influenciam a qualidade de vida dos idosos, utilizando-se da técnica da Regressão Logística, sendo baseado na dissertação de mestrado intitulado “Fatores que influenciam na Qualidade de Vida dos Idosos atendidos em um Hospital Universitário", de autoria de Carvalho (2019).

Carvalho (2019) cita que o processo de envelhecimento da população é um fenômeno que vem ocorrendo de maneira acelerada em diversos países do mundo, assim como no Brasil. Esse processo, caracterizado inicialmente pela queda nas taxas de mortalidade e posteriormente pela queda nos índices de natalidade, é uma questão relevante em função dos desafios que se impõem.

Conforme pode ser visto em Carvalho (2019), pela Legislação vigente no Brasil, uma pessoa é considerada idosa caso possua acima de 60 anos de idade. Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) referentes à 2019 apontam uma população idosa de cerca de 29,4 milhões no Brasil, o que corresponde a 14,3% da população total do país. As projeções do IBGE estimam que em 2060 a proporção de população idosa em relação ao total da população do país será de cerca de 25%, o que corresponderá a cerca de 58 milhões de habitantes. A expectativa de vida dos brasileiros em 2019 é de 72 anos para os homens e de 79 anos para as mulheres, com uma expectativa média de 75,5 anos. Segundo o Instituto, a melhoria na expectativa de vida deve-se à avanços importantes na medicina, o aumento da rede de saneamento básico, aumentos na renda média e no nível de escolaridade dos brasileiros, dentre outros fatores importantes.

Carvalho (2019) ainda afirma que, segundo a Organização Mundial da Saúde (OMS), um país é considerado envelhecido a partir do momento em que a população de determinado país possui 14% de população acima dos 60 anos. Como exemplos de países envelhecidos, a França levou 115 anos nesse processo, enquanto a Suécia levou 85 anos. Segundo a Sociedade Brasileira de Geriatria e Gerontologia (SBGG), as projeções para o Brasil indicam que o país será considerado envelhecido em 2032, com 32,5 milhões de idosos, em uma população total de cerca de 226 milhões de habitantes, dados referentes ao ano de 2019. Entretanto, conforme os dados citados anteriormente e contrariando as projeções, o Brasil atingiu com alguns anos de antecedência o estágio de país envelhecido, e esse cenário aumenta mais ainda a pressão em relação a políticas públicas voltadas para os idosos.

Para Cardoso e Costa (2010), este processo determina um novo perfil de morbimor-

talidade, que se caracteriza por um aumento de doenças crônicas, que no entanto não interferem necessariamente na independência dos idosos, desde que estas doenças estejam devidamente controladas. Para o mesmo autor, esse quadro é um desafio, uma vez que demandam ações de planejamento, gerência e prestação de serviços, tornando extremamente importante o conhecimento das necessidades e condições de vida desse grupo etário.

Carvalho (2019) cita que o envelhecimento humano é um fenômeno que afeta a sociedade de uma maneira geral, em que pese ser um processo natural, dinâmico e que acarrete perdas em vários âmbitos, e que geram um conjunto de necessidades que são especificadas por uma série de variáveis, tais como gênero, faixa de renda e o lugar onde vivem. Ainda segundo a autora, este conjunto de necessidades possui importância fundamental no processo construtivo dos fatores que impactam direta ou indiretamente na expectativa de vida, na saúde, na independência e na qualidade de vida dos idosos.

Para Anderson *et al.* (1998), a saúde e a qualidade de vida dos idosos, mais do que em outras faixas etárias, sofrem a influência de fatores físicos, psicológicos, sociais e culturais. Segundo os mesmos autores, avaliar e promover a saúde dos idosos significa considerar variáveis de distintos campos do saber, numa atuação interdisciplinar e multidimensional. Para Vecchia *et al.* (2005), o conceito de qualidade de vida é subjetivo, dependente do nível sociocultural, da faixa etária e das aspirações pessoais do indivíduo.

O Estatuto do Idoso, criado a partir da Lei nº 10.741, sancionada em 01 de outubro de 2003 (SENADO FEDERAL, 2003), é a ferramenta mais poderosa na garantia dos direitos dos idosos, lhes garantindo os direitos fundamentais a todos os humanos e assegurando-lhes oportunidades e facilidades para o exercício de seus direitos.

Neste contexto, medir os fatores que influenciam a qualidade de vida dos idosos é de fundamental importância, uma vez que conhecidos quais fatores têm maior ou menor influência, pode-se direcionar um maior suporte a esses fatores, desenvolvendo e fortalecendo ações e políticas públicas que contribuam com a qualidade de vida dos idosos. Uma ferramenta que pode ser utilizada para este objetivo é o modelo de regressão logística.

A Regressão Logística é uma área particular dos modelos de regressão cuja aplicação se dá em situações onde a variável resposta é qualitativa, ou seja, podem assumir níveis ou categorias, com as variáveis explicativas (também chamadas de variáveis independentes ou regressoras) podendo ser tanto qualitativas quanto quantitativas (GONZALEZ, 2018). As variáveis em estudo neste trabalho, tanto as variáveis explicativas quanto a variável resposta

são dicotômicas ou binárias, ou seja, assumem dois níveis ou categorias distintos (Sim e Não, Positivo e Negativo, Sucesso e Insucesso, 0 e 1, etc.). Ainda segundo Gonzalez (2018), mesmo quando a variável não é dicotômica, ou seja, pode assumir múltiplos níveis, é possível torná-la dicotômica, e assim permitir o uso da regressão logística.

Este trabalho tem por base a dissertação de mestrado de Carvalho (2019). O objetivo geral é investigar, por meio do uso do método de regressão logística, quais fatores possuem maior impacto na qualidade de vida dos idosos. Os objetivos específicos são:

i. Ajustar o modelo de regressão logística mais adequado para descrever o evento em estudo;
ii. Investigar, por meio do uso da regressão logística, quais fatores possuem maior impacto na qualidade de vida dos idosos.

Este trabalho está estruturado em cinco capítulos. No capítulo introdutório, traz-se uma breve explanação acerca da construção deste trabalho. No capítulo dois serão abordados, de uma maneira mais detalhada, os fatores que afetam a qualidade de vida dos idosos; também será apresentada a Regressão Logística. No capítulo três, será abordada a metodologia de pesquisa. O capítulo quatro será dedicado ao procedimento de modelagem e análise dos dados. Finalmente, no capítulo cinco, será trazida a conclusão obtida a partir dos resultados da análise.

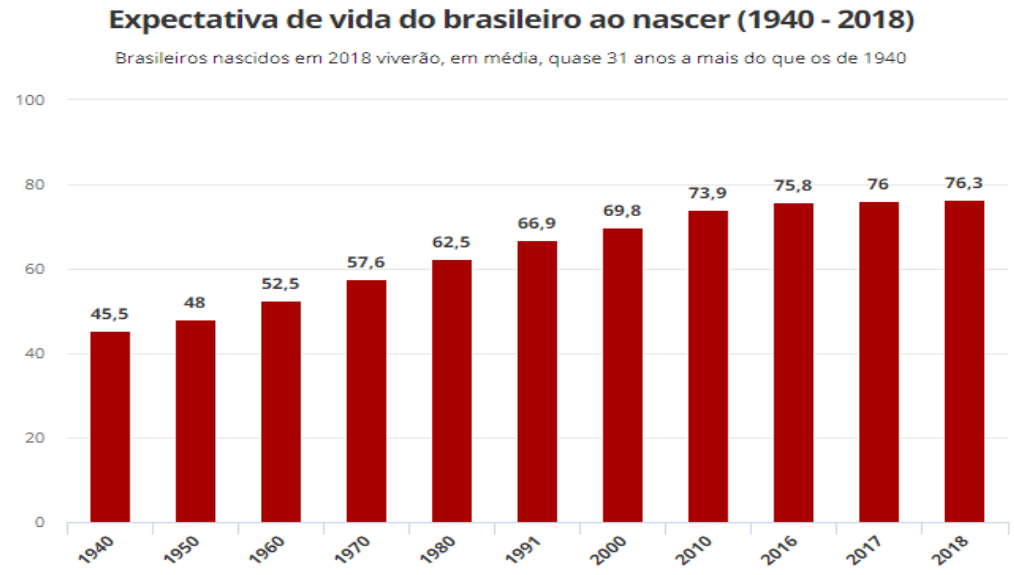
2 REFERENCIAL TEÓRICO

Neste Capítulo serão abordados de uma maneira mais detalhada os fatores que afetam a qualidade de vida dos idosos. A base para este capítulo foi o trabalho de Carvalho (2019).

2.1 Panorama do envelhecimento da população brasileira

Dados do IBGE divulgados no portal BrasilPrev (2019) indicam um aumento na expectativa de vida dos brasileiros. Em 2020, a expectativa média projetada é de 76,7 anos. Para 2040, a expectativa média é de 79,9 anos, chegando a 81,2 anos em 2060. Em termos comparativos, nos anos 1940 a expectativa de vida dos brasileiros era de 45,5 anos. Este aumento pode ser visualizado na Figura 1. Esses números são reflexo da evolução significativa em diversos aspectos, como saúde, renda e escolaridade. No entanto, estes números também apontam para novos desafios, e exigem mudanças nas formas de pensamento e vivência da velhice na sociedade.

Figura 1 – Demonstrativo da evolução da expectativa de vida no Brasil.

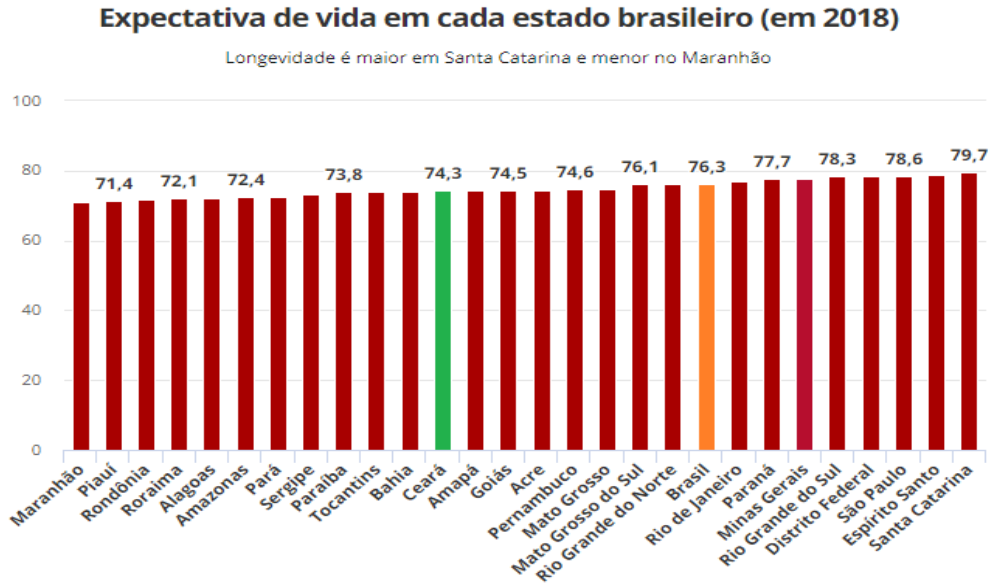


Fonte: G1 (2019), com base em dados do IBGE (2018b).

A Figura 2 retrata a expectativa de vida por estado brasileiro em 2018, e refletem alguns dos desafios relacionados ao aumento da expectativa de vida. De acordo com os dados do IBGE (2018b), Santa Catarina possui a melhor expectativa de vida do país, com média de 79,7 anos, alavancado por alto nível de escolaridade de sua população, uma renda média razoavelmente boa e um bom acesso aos serviços de saúde. Na ponta oposta desta lista, tem-se o estado do Maranhão, um dos mais pobres do Brasil, com índices de escolaridade baixos, acesso

a serviços de saúde incipiente e a menor renda média entre todos os estados brasileiros.

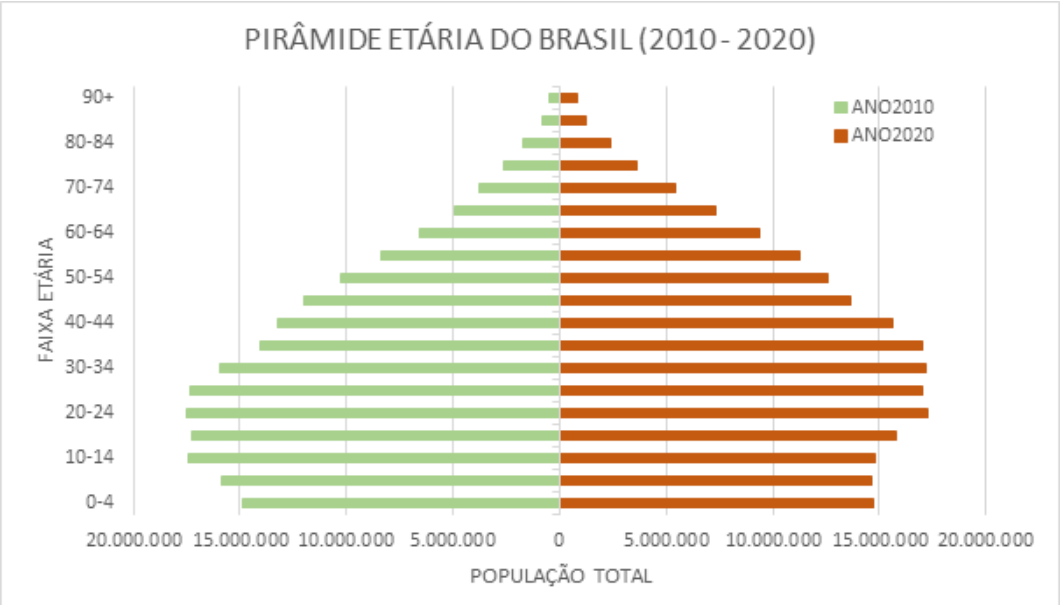
Figura 2 – Demonstrativo da expectativa de vida por estado brasileiro.



Fonte: G1 (2019), com base em dados do IBGE (2018b).

O envelhecimento da população brasileira no período de 2010 a 2020 é retratado na Figura 3. Essa mudança pode ser observada pelo estreitamento da base da pirâmide (o que indica a redução da porcentagem de população jovem), enquanto ocorre o alargamento do topo da pirâmide (indicando o aumento a porcentagem de população mais velha).

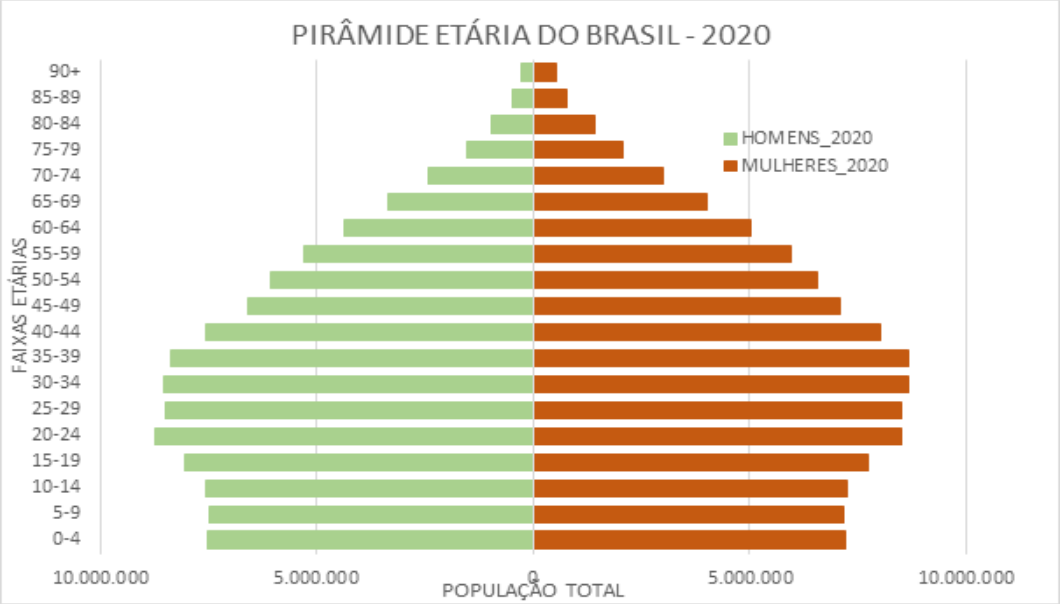
Figura 3 – Pirâmide Etária do Brasil projetada para 2020, comparada à 2010.



Fonte: Elaborada pelo autor, com base em dados obtidos do IBGE (2018a).

A Figura 4 mostra a pirâmide etária por gênero do Brasil projetada para o ano de 2020, e onde é perceptível o movimento de envelhecimento da população brasileira.

Figura 4 – Pirâmide Etária do Brasil projetada para 2020, dividida por gêneros.



Fonte: Elaborada pelo autor, com base em dados obtidos do IBGE (2018a).

De acordo com o portal G1 (2019), segundo o IBGE, a diferença da expectativa de vida entre homens e mulheres vai se acentuando conforme a faixa etária vai evoluindo em direção ao topo da pirâmide. Esse fenômeno é conhecido por “sobremortalidade masculina", e pode ser justificado por fatores externos ou não-naturais que possuem maior incidência sobre a população masculina, notadamente relacionados ao processo de urbanização do Brasil, além de fatores biológicos, comportamentais e até sociais. A sobremortalidade masculina é um dos fatores, que, junto com o índice de mortalidade infantil e o nível de acesso da população idosa aos serviços de saúde, impedem uma maior expectativa de vida da população brasileira.

2.1.1 Conceituação da Qualidade de Vida

De acordo com a Organização Mundial da Saúde (BIBLIOTECA VIRTUAL DA SAÚDE, 2013), qualidade de vida é “a percepção do indivíduo de sua inserção na vida, no contexto da cultura e sistemas de valores nos quais ele vive e em relação aos seus objetivos, expectativas, padrões e preocupações”. Segundo a mesma biblioteca, “envolve o bem estar espiritual, físico, mental, psicológico e emocional, além de relacionamentos sociais, como família e amigos e, também, saúde, educação, habitação saneamento básico e outras circunstâncias da vida”.

2.1.2 Fatores que afetam a qualidade de vida na Terceira Idade

Silva *et al.* (2017) citam que a qualidade de vida dos idosos é influenciada por fatores ligados à questões físicas (saúde), psicológicas (mental), sociais, econômicas e ambientais.

Com base nos autores acima, o autor deste trabalho entende que, no âmbito da saúde, a manutenção de hábitos saudáveis e um bom acesso a serviços de saúde é fundamental na qualidade de vida dos idosos. Na parte psicológica, é importante que o idoso tenha acesso à atividades que estimulem o uso contínuo da mente, visando preservar a saúde mental na terceira idade. No contexto social, é importante estimular os idosos a manterem interações sociais com outras pessoas, visando evitar problemas ligados à solidão. Em termos econômicos, idosos com melhor renda tendem a possuir melhores possibilidades de acesso a serviços de saúde, dentre outros aspectos influentes na sua qualidade de vida. Por fim, sob a ótica ambiental, o idoso conviver em um ambiente em condições adequadas às suas necessidades é um dos fatores mais importantes para uma boa qualidade de vida na terceira idade.

2.1.3 Políticas Públicas para a Terceira Idade

Segundo Caldas *et al.* (2008), as Políticas Públicas são a totalidade de ações, metas e planos que os governos (nacionais, estaduais ou municipais) traçam para alcançar o bem-estar da sociedade e o interesse público. No contexto do envelhecimento da população, estas políticas, especialmente nas áreas da saúde, lazer e assistência social, de previdência social e segurança têm por objetivo resolver inúmeros questionamentos que acompanham este processo, ao garantir direitos sociais fundamentais para este crescente segmento da população.

Para Porto (2002), a atenção ao idoso como política pública é relacionada à reivindicação por parte de movimentos sociais, bem como ao processo de desnevolvimento sócio-econômico-cultural. Para a autora, a Constituição Federal de 1988 foi um grande marco neste sentido, ao introduzir em seus dispositivos o conceito de Seguridade Social, mudando o foco assistencialista para uma ampliação do conceito de cidadania. O artigo 230 da Constituição determina que "A família, a sociedade e o Estado têm o dever de amparar as pessoas idosas, assegurando sua participação na comunidade, defendendo sua dignidade e bem-estar e garantindo-lhes o direito à vida."

Partindo deste princípio, estabeleceu-se a Política Nacional do Idoso (Lei n. 8.842, de 4 de Janeiro de 1994), que tem por objetivo promover uma longevidade com maior qualidade

de vida, baseada em ações em múltiplas áreas, normatizando direitos sociais dos idosos, e garantindo autonomia, integração e participação efetiva como instrumento do exercício pleno da cidadania (PORTO, 2002).

O Estatuto do Idoso (Lei n. 10.741, de 1 de Outubro de 2003) trouxe uma novidade significativa neste aspecto, ao definir princípios de proteção integral e regular direitos específicos destinados a este segmento, além de determinar prioridade absoluta, tornando-se a ferramenta mais poderosa na garantia dos direitos dos idosos, ao lhes garantir os direitos fundamentais a todos os humanos e assegurar-lhes oportunidades e facilidades para o exercício de seus direitos (BRASIL, 2003).

As políticas públicas para a Terceira Idade no Brasil, apesar dos avanços construídos ao longo dos últimos anos, ainda estão em um estágio bastante atrasado, algo que atinge principalmente idosos cuja escolaridade e renda sejam mais baixos (AGÊNCIA CÂMARA DE NOTÍCIAS, 2019). Segundo o mesmo portal, há uma forte demanda por parte da sociedade civil em relação a políticas que sejam continuadas e que integrem diversos setores, haja vista a proporção de idosos que precisam de algum tipo de auxílio para exercer atividades básicas.

3 PROCEDIMENTO METODOLÓGICO

Neste capítulo será abordada a coleta e organização dos dados utilizados neste trabalho e o ferramental adotado para o tratamento, modelagem e análise.

3.1 Descrição do Estudo

Esta aplicação e análise é originada de dados secundários advindos do estudo elaborado por Carvalho (2019), que fez parte de uma ampla pesquisa internacional acerca da qualidade de vida de pessoas da terceira idade (idosos), cuja pesquisa tratou dos idosos que são atendidos no Ambulatório de Geriatria do HUWC-UFC, e baseado no *World Health Organization Quality of Life Questionnaire* (WHOQOL-Bref), questionário proposto pela OMS com o intuito de ser um instrumento de avaliação da qualidade de vida na terceira idade.

Segundo o Instituto PHD (2019), dados secundários são dados geralmente gratuitos e de fácil acesso, seja pela internet ou meios impressos, que indicam tendências e fornecem um útil panorama geral sobre um segmento, um público-alvo ou um cenário político, econômico e social. Segundo o mesmo instituto, sua utilização, embora complementar, não é menos importante. A pesquisa de dados secundários, também chamada "pesquisa documental" e *Desk Research*, é o passo inicial do levantamento de dados, contribuindo para uma análise mais profunda acerca de um determinado assunto, ou para levantamento de hipóteses em relação ao estudo a ser executado.

De acordo com Carvalho (2019), por meio de convite e aceitação foram selecionadas pessoas acima dos 60 anos de ambos os gêneros, saudáveis e atendidas nos diversos níveis de serviços de saúde. No polo do Ceará participaram da pesquisa 95 idosos, entre os quais foi aplicado o questionário WHOQOL-bref, em uma versão que contém 26 questões e que está presente no Anexo A. Essas questões se referem às duas semanas anteriores à aplicação de questionário, e são distribuídas da seguinte maneira:

- A questão 1 é referente à qualidade de vida sob um aspecto geral;
- A questão 2 é referente à satisfação pessoal com a saúde;
- As questões 3 a 9 se referem a sensações pessoais;
- As questões 10 a 14 referem-se à capacidade de execução de atividades rotineiras;
- As questões 15 a 25 se referem à satisfação com relação a aspectos cotidianos do idoso;
- A questão 26 é referente a experimentar certos tipos de sentimentos.

As questões de 3 a 26 são classificadas de acordo com os seguintes domínios:

- Domínio físico: são questões relacionadas à capacidade física dos idosos. São as questões 3, 4, 10, 15 a 18;
- Domínio psicológico: são questões relacionadas à capacidade mental dos idosos. São as questões 5 a 7, 11, 19 e 26;
- Relações Sociais: são questões relacionadas à interação e integração social dos idosos. São as questões 20 a 22;
- Meio Ambiente: são questões relacionados aos ambientes de convivência dos idosos. São as questões 8, 9, 12 a 14, 23 a 25.

Esses domínios serão as variáveis explanatórias, ou seja, fatores independentes que têm por objetivo explicar a variável em estudo (resposta), que poderão ser visualizados no questionário que está no Anexo A. Para fins de facilitar a análise dos dados, a variável resposta será denotada por *QV* (Como você avaliaria sua qualidade de vida?), que possui relação com o objetivo principal deste trabalho. As variáveis explicativas serão denotadas como descrito no Quadro 1:

Quadro 1 – Notação das variáveis explicativas relacionadas à qualidade de vida na Terceira Idade.

Notação	Variáveis Explicativas
Q3	Em que medida você acha que sua dor (física) impede você de fazer o que você precisa?
Q4	O quanto você precisa de algum tratamento médico para levar sua vida diária?
Q5	O quanto você aproveita a vida?
Q6	Em que medida você acha que a sua vida tem sentido?
Q7	O quanto você consegue se concentrar?
Q8	Quão seguro(a) você se sente em sua vida diária?
Q9	Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?
Q10	Você tem energia suficiente para seu dia-a- dia?
Q11	Você é capaz de aceitar sua aparência física?
Q12	Você tem dinheiro suficiente para satisfazer suas necessidades?
Q13	Quão disponíveis para você estão as informações que precisa no seu dia-a-dia?
Q14	Em que medida você tem oportunidades de atividade de lazer?
Q15	Quão bem você é capaz de se locomover?
Q16	Quão satisfeito(a) você está com o seu sono?
Q17	Quão satisfeito(a) você está com sua capacidade de desempenhar as atividades do seu dia-a-dia?
Q18	Quão satisfeito(a) você está com sua capacidade para o trabalho?
Q19	Quão satisfeito(a) você está consigo mesmo?
Q20	Quão satisfeito(a) você está com suas relações pessoais (amigos, parentes, conhecidos, colegas)?
Q21	Quão satisfeito(a) você está com sua vida sexual?
Q22	Quão satisfeito(a) você está com o apoio que você recebe de seus amigos?
Q23	Quão satisfeito(a) você está com as condições do local onde mora?
Q24	Quão satisfeito(a) você está com o seu acesso aos serviços de saúde?
Q25	Quão satisfeito(a) você está com o seu meio de transporte?
Q26	Com que frequência você tem sentimentos negativos tais como mau humor, desespero, ansiedade, depressão?

Fonte: Elaborado pelo autor, baseado no questionário WHOQOL presente no Anexo A.

3.1.1 Codificação das Respostas

As respostas ao questionário seguem uma escala de Likert de 5 níveis, na qual quanto maior for a pontuação melhor será a avaliação em relação a determinada variável, com níveis que vão do extremo-negativo ao extremo-positivo, com uma posição de neutralidade ou indiferença, descritas no Quadro 2:

Quadro 2 – Escalas de Respostas do Questionário aplicado na pesquisa.

Questões	Níveis				
	1	2	3	4	5
1 e 15	muito ruim	Ruim	nem ruim nem boa	boa	muito boa
2, 16 a 25	muito insatisfeito	Insatisfeito	nem satisfeito, nem insatisfeito	satisfeito	muito satisfeito
3 a 14	nada	muito pouco	mais ou menos (médio)	bastante (muito)	extremamente (completamente)
26	nunca	algumas vezes	frequentemente	muito frequentemente	sempre

Fonte: Elaborado pelo autor, baseado no questionário WHOQOL presente no Anexo A.

Adotou-se, nesta monografia, o seguinte critério para a codificação dos dados, em todas as variáveis: os níveis 1 e 2 (respostas negativas), além do nível 3 (neutralidade) foram codificadas como valor 0 (ausência do fator em questão), enquanto os níveis 4 e 5 (respostas positivas) receberam valor 1 (presença do fator em questão).

Segundo o The WHOQOL Group (1995), as questões 3, 4 e 26 do questionário possuem escalas negativas, ou seja, pontuações mais altas indicam piores avaliações para a variável em questão. Deste modo, no processo de tabulação dos dados, a pontuação referente a essas questões deve ser codificada de maneira invertida.

A codificação dos dados originais foi realizada de forma direta, por meio do Microsoft Excel (2016), e os dados codificados foram armazenados em uma planilha auxiliar, para utilização nos passos seguintes.

3.2 Tratamento de dados faltantes

Paralelamente ao passo anterior, foi percebido que, dentre as 95 observações coletadas no estudo original, em 22 (ou 23,1% do total) foi notada a ausência de resposta em pelo menos uma das perguntas do questionário. Assim, o autor deste trabalho optou pela retirada destas observações, e a análise dos dados prosseguiu com base nas 73 observações restantes. Sempre que se fizer necessário mencionar uma observação específica, esta observação será identificada por uma ID, presente no banco de dados original.

Nas seções seguintes, será feita uma breve explanação do ferramental teórico utlizado neste trabalho, visando facilitar o entendimento do que será feito posteriormente, e por fazer parte do processo de escolha e validação do modelo. A execução das análises foi feita por meio do *software* R (R CORE TEAM, 2019), versão 3.6.1.

3.3 Teste Qui-Quadrado

De acordo com Agresti (2007), o objetivo do teste χ^2 é determinar, com base na diferença entre os valores que são esperados e os valores que são obtidos, o nível de associação entre duas variáveis, também podendo avaliar o grau de dependência entre elas. Considere a Tabela 1:

Tabela 1 – Tabela de classificação Ilustrativa.

Status	Previsão		Total
	D	D^C	
D	n_{11}	n_{12}	n_{1*}
D^C	n_{21}	n_{22}	n_{2*}
Total	n_{*1}	n_{*2}	n

Fonte: Elaborado pelo autor.

A probabilidade de alocação de um indivíduo aleatoriamente em uma célula é dada por:

$$\hat{\pi}_{ij} = \frac{n_{i*} \times n_{*j}}{n^2}.$$

Pode-se pensar no evento de uma observação pertencente ou não à célula (i, j) sendo que n_{ij} representa o número de vezes que as observações foram alocadas na célula (i, j) com probabilidade $\hat{\pi}_{ij}$. Assim,

$$n_{ij} \sim Bin(n, \hat{\pi}_{ij}),$$

e como consequência, o número esperado de n_{ij} é dado por

$$\mathbb{E}[n_{ij}] = n\hat{\pi}_{ij} = \frac{n_{i*} \times n_{*j}}{n} = E_{ij}.$$

O objetivo é testar:

$\mathscr{H}_0$: Não há associação entre as variáveis.

$\mathscr{H}_1$: Há associação entre as variáveis.

A estatística de teste é dada por:

$$Q = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((l-1)(c-1)),$$

com l e c representando, respectivamente, o número de linhas e de colunas da tabela de classificação. Quanto maiores as diferenças entre n_{ij} e E_{ij}, mais provavelmente a hipótese nula será

rejeitada. O $p-$valor é dado por:

$$p-\text{valor} = \mathbb{P}(Q > \chi^2((l-1)(c-1))),$$

Assim, se o $p-$valor $< \alpha$, rejeita-se a hipótese nula.

Sabe-se que o teste χ^2 não possui bom desempenho para amostras não tão grandes, sendo recomendado que a frequência esperada das células seja ≥ 5, para uma boa aproximação deste teste (AGRESTI, 2007).

3.4 Regressão Logística

De acordo com Fávero *et al.* (2009), a regressão logística é uma técnica utilizada para descrever o comportamento entre uma variável dependente binária e variáveis explicativas qualitativas ou quantitativas, destinando-se a a investigar o efeito das variáveis pelas quais os indivíduos, objetos ou sujeitos estão expostos sobre a probabilidade de ocorrência de determinado evento de interesse. Segundo os mesmos autores, diferentemente da regressão múltipla, a regressão logística não pressupõe a existência de homogeneidade de variância e normalidade dos resíduos. Assim, considere uma variável binária, onde tem-se:

$$R_i = \begin{cases} 1, & \text{se o indivíduo possui determinada característica;} \\ 0, & \text{caso contrário.} \end{cases}$$

A probabilidade de sucesso, ou seja, $\mathbb{P}(R_i = 1) = p$, indica que R_i segue uma distribuição de Bernoulli com parâmetro p, média p e variância $p(1-p)$. Em uma sequência de n provas independentes, tem-se que

$$Y = \sum_{i=1}^{n} R_i \sim Bin(n, p).$$

A ideia é obter uma estimativa para p por meio de fatores que estejam diretamente ligados a ele, ou seja,

$$f(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q,$$

na qual $\beta_0, \beta_1, \ldots, \beta_q$ é o conjunto de parâmetros do modelo, $x_1, x_2, \ldots, x_q$ as variáveis explicativas e $f(p)$ a variável-resposta. A premissa inicial é que $0 < p < 1$, já que está-se trabalhando com probabilidades. Então, precisa-se de um modelo ou transformação que respeite esta premissa. Pensando nisto, a melhor transformação é dada por

$$f(p) = \ln\left(\frac{p}{1-p}\right).$$

Note que $\frac{p}{1-p}$ é chamado de *odds*, ou seja, o quanto é possível um evento ocorrer em relação à chance deste mesmo evento não ocorrer. Essa transformação é conhecida por *logit*. Perceba que (FÁVERO *et al.*, 2009)

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \Leftrightarrow \ln\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$$
$$\Rightarrow \quad logit(p) = \beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k.$$

O $logit(p)$ é diretamente proporcional a p, ou seja, quanto maior for p, maior será o valor de $logit(p)$. Esse modelo equivale a fazer um ajuste de modelo linear no qual a variável resposta é $\ln\left(\frac{p}{1-p}\right)$. É a esse modo de ajustar modelos que denomina-se regressão logística.

3.5 Estimação dos parâmetros

A obtenção do vetor de parâmetros estimados $\hat{\beta}$ será via máxima verossimilhança, da seguinte forma. Conforme citado em Hosmer-Jr *et al.* (2013), a função de verossimilhança é dada por:

$$l(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

sendo

$$p_i = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}},\ \beta_i \in \mathbb{R}.$$

A intenção é estimar o valor de β que maximiza $l(\beta)$. Uma maneira de se obter isto é utilizando o logaritmo em $l(\beta)$ e obtendo a função de log-verossimilhança $L(\beta)$, definido em Hosmer-Jr *et al.* (2013) como segue:

$$
\begin{aligned}
L(\beta) = \ln(l(\beta)) &= \ln\left[\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln(p_i) + \ln(1-p_i) - y_i \ln(1-p_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}}{1-\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}}\right) + \ln\left(1-\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}\right)\right] \\
&= \sum_{i=1}^{n}\left[y_i \ln(e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}) - \ln(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k})\right] \\
L(\beta) &= \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k) - \ln(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k})\right]
\end{aligned}
$$

Feito isto, deriva-se $L(\beta)$ em relação a cada um dos β_i do modelo, obtendo-se:

$$
\frac{\partial L(\beta)}{\partial L(\beta_0)} = \sum_{i=1}^{n}\left[y_i - \left(\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}\right)\right] = \sum_{i=1}^{n}\left[y_i - p_i\right]
$$

$$
\begin{aligned}
\frac{\partial L(\beta)}{\partial L(\beta_j)} &= \sum_{i=1}^{n}\left[y_i x_i - \left(x_i \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}\right)\right] \\
&= \sum_{i=1}^{n} x_i\left[y_i - \left(\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}\right)\right] \\
&= \sum_{i=1}^{n} x_i\left[y_i - p_i\right]
\end{aligned}
$$

Portanto, para encontrar $\hat{\beta}$, é necessário encontrar as soluções das equações

$$
\sum_{i=1}^{n}\left[y_i - p_i\right] = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_i\left[y_i - p_i\right] = 0.
$$

Para se resolver as equações acima, se faz necessário o uso do método iterativo de Newton-Raphson, uma vez que são não-lineares. O passo-a-passo, também adotado em Carelli (2017), é feito como segue:

1. Após fazer a derivação, obtém-se o vetor escore:

$$
U(\beta) = X^\top Y - X^\top \pi = X^\top[Y - \pi],
$$

na qual $Y = (y_1, y_2, \ldots, y_n)^\top$ e $\pi = (p_1, p_2, \ldots, p_n)^\top$.

2. A matriz de informação de Fisher é dada por

$$I(\beta) = \mathbb{E}\left[-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^\top}\right] = X^\top QX,$$

sendo $Q = diag[p_i(1-p_i)]_{n\times n}$, $X_{n\times p}$ é a matriz dos dados e $[I(\beta)]^{-1}_{p\times p}$ é a matriz de variâncias-covariâncias das estimativas.

3. O conjunto de equações iterativas é dado por

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + [I(\beta^{(t)})]^{-1}U(\beta^{(t)}) \\ &= \beta^{(t)} + [X^\top Q^{(t)}X]^{-1}X^\top[Y-\pi^{(t)}],\end{aligned}$$

para $t = 0,1,2,\ldots$, $\beta^{(t)}$ e $\beta^{(t+1)}$ são os vetores estimados nos passos t e $t+1$, respectivamente. Normalmente o valor inicial é um vetor $\beta = (0,0,\ldots,0)^\top$, e vai se repetindo o processo até que atinja um nível de estabilidade, representado por uma margem de tolerância, na qual o resultado obtido na iteração $t+1$ não difira significativamente do resultado da iteração t.

$$tol = \begin{cases} |\beta^{(t+1)} - \beta^{(t)}| < 0,0001, & \text{se } \beta^{(t)} \leq 0,01; \\ \left|\dfrac{\beta^{(t+1)} - \beta^{(t)}}{\beta^{(t)}}\right| < 0,0001, & \text{se } \beta^{(t)} > 0,01. \end{cases}$$

3.6 Critério de Seleção de Variáveis

Esta Seção trata dos critérios de seleção de variáveis que serão adotados neste trabalho, com o objetivo de auxiliar no processo de escolha do melhor modelo para os dados em estudo.

3.6.1 Critério de Informação de Akaike

Com o objetivo de selecionar o melhor modelo, Akaike (1974) desenvolveu o Critério de Informação de Akaike (AIC), que pode ser obtido através da minimização da informação de Kullback e Leibler (1951). A Informação de K-L é uma medida que calcula a distância entre o modelo verdadeiro e um outro aproximado. Dessa forma, Akaike (1974) mostrou que, assintoticamente, o viés é dado por q, em que q é o número de parâmetros. Assim, foi desenvolvido uma estimativa da informação K-L, utilizando a função de log-verossimilhança no seu ponto máximo, adicionando uma penalidade relacionada ao número de parâmetros. Dessa maneira, esse critério pode ser empregado quando se tem interesse em testar modelos

e compará-los a respeito de qual se adequa melhor à um conjunto de dados, havendo, porém, apenas modelos melhores e piores uns que os outros, entretanto, jamais deve ser usado como um teste de hipóteses no sentido de rejeitar um ou outro modelo, uma vez que o AIC não possui significância e nem $p-$valor. O critério para avaliar se um modelo se adequa bem, em relação a outros modelos propostos, à um conjunto de dados é verificar se o seu AIC é menor que o dos outros modelos.

Ainda segundo Akaike (1974), o AIC é dado por:

$$AIC = -2L(\hat{\beta}) + 2q,$$

em que $L(\hat{\beta})$ é o logaritmo da função de log-verossimilhança e q o número de parâmetros. Burnham e Anderson (2004) recomendam utilizar o AIC apenas quando $n > 40q$, em que n é o número de observações.

3.6.2 Critério de Informação de Akaike Corrigido

Tendo em vista o problema do AIC para amostras pequenas, Hurvich e Tsai (1989) propuseram uma correção para este método de seleção, do qual é recomendado para pequenas amostras e dados que sejam modelados pela distribuição normal. O Critério de Informação de Akaike Corrigido (AICc) é dado por:

$$AICc = AIC + \frac{2(q+1)(q+2)}{n-q-2}$$

Davison (2001) admitiu que o AICc pode aumentar a probabilidade de se escolher um modelo apropriado para o evento estudado, mas deve-se considerar que este critério foi desenvolvido para populações normalmente distribuídas, logo utilizar para populações modeladas por outras distribuições é incerto. Além disso, é importante ressaltar que para $n \to \infty$ o AICc converge para o AIC.

3.6.3 Critério de Informação Bayesiano

O Critério de Informação Bayesiano (BIC) foi desenvolvido por Schwarz (1978), e assim chamado pelo fato de Schwarz ter usado um argumento Bayesiano para prová-lo. É um critério de seleção de modelos definido em termos da probabilidade a posteriori e de maneira análoga ao AIC, o BIC também penaliza modelos com muitas variáveis. Além disso, o BIC é um resultado assintótico derivado sob a hipótese de que os dados observados são regidos por

modelos probabilísticos que pertencem à Família Exponencial (para maiores informações acerca da Família Exponencial, ver CASELLA e BERGER (2011)).

Assim, o BIC é dado por:

$$BIC = -2\ln f_X(x_n|\beta) + q\ln n,$$

em que $f_X(x_n|\theta)$ é o modelo probabilístico escolhido, q o número de parâmetros e n o tamanho da amostra.

O pacote `bestglm` (MCLEOD *et al.*, 2020) do *software* R traz a implementação de algoritmos que se utilizam dos critérios de informação para a seleção do melhor modelo para os dados em estudo. A função `bestglm` produz uma lista de possíveis modelos selecionáveis, adotando-se como critério os melhores valores para o AIC (ou algum outro critério para a seleção de variáveis), e sugere o modelo dentre os selecionáveis mais adequado, dentro do critério adotado.

3.7 Estatística *Deviance*

O *deviance* (ou desvio) do modelo logístico é equivalente à soma de quadrados dos resíduos do modelo de regressão linear, e definida em Hosmer-Jr *et al.* (2013) como:

$$Dev = -2\ln\left[\frac{L_{red}}{L_{sat}}\right],$$

na qual L_{red} e L_{sat} representam, respectivamente, a verossimilhança do modelo reduzido e a verossimilhança do modelo saturado. A estatística *deviance* segue distribuição χ^2 com $n-q$ graus de liberdade, e quanto menor for melhor será o ajuste do modelo.

3.8 Coeficientes de Determinação

Segundo o portal Beta Analítica (2019), o coeficiente de determinação R^2 é uma medida adotada em regressão linear com o objetivo de avaliar a qualidade de um ajuste, e esta mesma interpretação pode ser aplicada aos modelos de regressão logística, podendo ser utilizada para comparar o desempenho entre dois ou mais modelos selecionáveis. No caso de modelos lineares generalizados, são conhecidos por pseudo-R^2, pois, apesar de variarem entre 0 e 1, a interpretação é análoga ao R^2 usado na regressão linear. Existe uma ampla variedade de pseudos-R^2 utilizados, alguns dos mais empregados são descritos abaixo, citados em Smith e McKenna (2013).

Seja L_0 o valor da função de verossimilhança para um modelo nulo e L_1 seja a verossimilhança para o modelo que está sendo estimado. O R^2 de McFadden é definido como

$$R^2_{MF} = 1 - \frac{\ln L_1}{\ln L_0}.$$

O R^2 de Cox-Snell é dado por

$$R^2_{CS} = 1 - \left(\frac{\ln L_0}{\ln L_1}\right)^{2/n},$$

em que n é o tamanho da amostra.

O R^2 de Nagelkerke é uma versão do R^2 de Cox-Snell adaptada para fornecer resultados entre 0 e 1:

$$R^2_N = \frac{1 - \left(\frac{\ln L_0}{\ln L_1}\right)^{2/n}}{1 - (\ln L_0)^{2/n}}.$$

O R^2 de Efron é dado por

$$R^2_E = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}.$$

Note que o R^2 de Efron não usa a verossimilhança, é um coeficiente baseado na razão entre a soma de quadrados dos resíduos e a somas de quadrados das somas dos desvios em relação à média. Ainda segundo o portal Beta Analítica (2019), o pseudo-R^2 só tem real utilidade para comparar múltiplos modelos no caso de estes modelos originarem-se no mesmo banco de dados, bem como não é possível comparar resultados obtidos em diferentes métodos por conta dos diferentes padrões de cálculo.

O pacote `performance` do *software* R traz a implementação destes e de vários outras formas de se obter o pseudo-R^2. Entre duas equações logísticas igualmente válidas, deve-se preferir o que apresente o coeficiente mais elevado.

Para modelos lineares generalizados, existe o coeficiente D^2, equivalente ao R^2, que é dado por

$$D^2 = \frac{Dev_0 - Dev_1}{Dev_0},$$

em que Dev_0 e Dev_1 são, respectivamente, o *deviance* do modelo nulo e o *deviance* do modelo em estudo. Weisberg (1980) sugeriu um ajuste no coeficiente D^2, similar ao que acontece no R^2:

$$D^2_{aj} = 1 - \left[\frac{n-1}{n-q}\right](1 - D^2).$$

Note que este ajuste é uma penalização baseada no número de observações e no número de parâmetros do modelo. Da mesma forma do pseudo-R^2, deve-se preferir o ajuste com maior coeficiente D^2.

3.9 Outras estatísticas de ajuste do modelo

A função `accuracy` do pacote `rcompanion` do *software* R retorna uma série de estatísticas relacionadas aos ajustes de modelo, tais como o erro quadrático médio, o valor do pseudo-R^2, e o coeficiente de variação.

3.9.1 Erro Quadrático Médio

Em Särndal *et al.* (2003), o erro quadrático médio de um estimador é obtido por meio da fórmula:

$$EQM(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2,$$

na qual

$$B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Dentro da função `accuracy`, Mangiafico (2016) utiliza a seguinte forma de calcular o erro quadrático médio:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

ou seja, o autor calcula o erro quadrático médio a partir da média dos quadrados dos resíduos obtidos no modelo. Sob este critério, o melhor ajuste é o que possui menor erro quadrático médio.

3.9.2 Coeficiente de Variação

Särndal *et al.* (2003) citam que a razão entre o erro padrão do estimador e o valor esperado é chamada de erro padrão relativo ou coeficiente de variação do estimador. Esta medida é bastante útil para comparar a variabilidade de dois ajustes ou conjuntos de dados, sendo uma medida adimensional, e geralmente expresso na forma percentual:

$$CV(\hat{\theta}) = \frac{[V(\hat{\theta})]^{1/2}}{\mathbb{E}(\hat{\theta})} \times 100.$$

Mangiafico (2016) utiliza, dentro da função `accuracy`, o seguinte método de obtenção do coeficiente de variação:

$$CV = \frac{\sqrt{EQM}}{\bar{Y}} \times 100,$$

em que o erro quadrático médio é calculado pelo mesmo autor da forma citada na Seção 3.9.1. O melhor ajuste é o que possui menor coeficiente de variação.

3.10 Teste de Wald

O princípio do teste de Wald é testar a significância de uma parâmetro dentro de um modelo. O teste de hipóteses, segundo Agresti (2007), é dado por:

$$\mathscr{H}_0 : \beta_i = 0$$
$$\mathscr{H}_1 : \beta_i \neq 0$$

Segundo Hosmer-Jr *et al.* (2013), a estatística de teste é obtida utilizando a razão entre a estimativa do parâmetro $(\hat{\beta}_i)$ e seu respectivo erro-padrão estimado, ou seja:

$$W_i = \frac{\hat{\beta}_i}{\hat{EP}(\hat{\beta}_i)}.$$

O $p-$valor, neste caso, é dado por:

$$p - \text{valor} = 2\mathbb{P}(|Z| > W_i),$$

com Z representando a variável aleatória de uma distribuição normal-padrão. Rejeita-se a hipótese nula quando $p-\text{valor} < \alpha$.

3.11 Teste de Hosmer e Lemeshow

O teste de Hosmer e Lemeshow (1980) é um teste de bondade de ajuste, baseado na avaliação das distâncias entre as probabilidades observadas e as probabilidades ajustadas. Neste teste, as probabilidades são ordenadas e divididas em g grupos de comprimento aproximadamente igual. Hosmer e Lemeshow (1980) sugerem utilizar $g = 10$.

Na literatura há pouca informação sobre a forma de se escolher o valor de g. No entanto, Hosmer e Lemeshow (1980) utlizaram em suas simulações $g > p+1$, com p sendo o número de covariáveis do modelo ajustado, mas deve-se utilizar no mínimo 3 grupos, pois para

$g < 3$ é impossível calcular a estatística do teste. O objetivo é testar:

$\mathscr{H}_0$: Não há diferença entre o que foi esperado e o que foi observado.

$\mathscr{H}_1$: Há diferença entre o que foi esperado e o que foi observado.

A estatística de Hosmer e Lemeshow (1980) é dada por:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k^* \bar{\pi}_k)^2}{n_k^* \bar{\pi}_k (1 - \bar{\pi}_k)},$$

onde O_k representa a frequência observada de observações com $Y = 1$, n_k^* representa a frequência total de observações e $\bar{\pi}_k$ é a probabilidade média estimada do grupo k.

A estatística $\hat{C}$ segue uma distribuição χ^2 com $g-2$ graus de liberdade. Rejeita-se a hipótese nula se $p-$valor $> \alpha$.

3.12 Classificação e Curva ROC

Segundo Margotto (2010), o ponto de corte *PC* é um valor definido entre 0 e 1, que pode ser selecionado de maneira arbitrária entre os valores possíveis para a variável de decisão, a qual classifica observações com valores abaixo como negativo (ausência do fator em questão), e valores acima como positivo (presença do fator em questão).

Há vários critérios para a escolha de um ponto de corte. No entanto alguns são arbitrários, podendo induzir a uma baixa taxa de detecção, fazendo com que o modelo se torne ineficiente. Para cada ponto de corte há uma tabela de classificação como a Tabela 2, chamada comumente de *matriz de confusão*, e que permite visualizar o desempenho de um modelo (ou outro tipo de algoritmo) de classificação.

Tabela 2 – Informações retratadas em uma matriz de confusão.

Status	Previsão		Total
	D	D^C	
D	VP	FP	$VP+FP$
D^C	FN	VN	$FN+VN$
Total	$VP+FN$	$FP+VN$	n

Fonte: Elaborado pelo autor.

Tabelas como a Tabela 2 trazem informações importantes, listadas abaixo:

– As indicações VP, VN, FP e FN são, respectivamente, verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo, que são as possibilidades de classificação de cada observação de acordo com o que é observado em relação ao que é previsto no modelo;

- Acurácia, ou taxa global de classificação, que é a capacidade de acerto do modelo, independente do modo como as observações são classificadas: $ACC = \frac{VP+VN}{n}$;
- Especificidade, que é a capacidade de previsão correta da ausência do evento em estudo: $ESPEC = \frac{VN}{VN+FN}$;
- Sensibilidade, que é a capacidade de previsão correta da presença do evento em estudo: $SENS = \frac{VP}{VP+FP}$;
- Taxa de Falso-Positivo: $TFP = \frac{FP}{VP+FP}$;
- Taxa de Falso-Negativo: $TFN = \frac{FN}{VN+FN}$.

Deve-se escolher o ponto de corte sob algum critério que torne o modelo eficiente. Um dos critérios mais adotados é o que o modelo seja simultaneamente altamente específico e altamente sensível (AGRESTI, 2007), ou seja,

$$PC = max\{SENS + ESPEC\},$$

obtendo assim um modelo bastante eficaz. No entanto, é importante atentar para os pontos extremos (muito próximos de 0 ou de 1), pois podem gerar somas altas por altos valores de *SENS* ou de *ESPEC*, tornando o modelo ineficiente.

O ponto de corte tem influência nos níveis de sensibilidade e de especificidade, o que pode ser desvantajoso. Porém, analisando vários valores de ponto de corte $(0 < PC < 1)$, tem-se uma visão mais ampla acerca da preditividade do modelo. Para cada ponto de corte avaliado tem-se um par (*SENS*, *ESPEC*), utilizados na construção da curva *Receiver Operating Characteristic* (ROC), que é um gráfico de pontos (*SENS*, $1 - ESPEC$), unidos por segmentos de reta. Quanto mais pontos de corte uma curva ROC possuir, maior a precisão do modelo adotado.

Isto posto, para Hosmer-Jr *et al.* (2013), a área sob a curva ROC, que varia de 0,5 a 1, fornece uma medida de capacidade para que seja efetuada a discriminação entre os indivíduos que experimentam o resultado de interesse versus os que não experimentam. Quanto maior for o valor dessa área, melhor será a capacidade preditiva do modelo ajustado.

Hosmer-Jr *et al.* (2013) apresenta uma regra interpretativa da *Area Under the Curve* (AUC), relacionada ao poder preditivo do modelo ajustado, da maneira que segue:

- Se $AUC = 0,5$, não indica discriminação;
- Se $0,7 \leq AUC < 0,8$, indica uma discriminação aceitável;
- Se $0,8 \leq AUC < 0,9$, indica uma discriminação excelente;
- Se $AUC \geq 0,9$, indica discriminação excepcional.

Segundo Margotto (2010), em termos práticos, um $AUC = 0,5$ não é interessante, uma vez que o poder de discriminação do modelo não seria muito melhor do que jogar uma moeda honesta, e um $AUC = 1$ (classificador perfeito) é de difícil alcance. Assim, valores de AUC no intervalo $(0,5;1)$ já podem ser considerados bons, e melhores o quanto mais próximos de 1 estiverem.

3.13 Razão de Chances

De acordo com Hosmer-Jr *et al.* (2013), a razão de chances, também chamada de *odds ratio*, é a razão entre os *odds* de duas categorias A e B com probabilidades r e s, respectivamente, e indica o quão provável um evento é em uma categoria em relação ao mesmo evento na outra categoria. O *odds ratio* é calculado como segue:

$$OR = \frac{odds(A)}{odds(B)} = \frac{r/1-r}{s/1-s} = \frac{r(1-s)}{s(1-r)}.$$

A interpretação do *OR* pode ser feita de duas maneiras:

- Em números absolutos, um $OR = 1$ indica que o evento é equiprovável em ambos os grupos. Valores de *OR* maiores que 1 indicam que o evento tem maior probabilidade de ocorrer no grupo A em relação ao grupo B, enquanto que valores de *OR* menores que 1 indicam o contrário.
- Em termos percentuais, temos que $ORp = (OR - 1) \times 100\%$, na qual um $ORp > 0\%$ aponta um aumento na probabilidade de o evento ocorrer no grupo A em relação ao grupo B, enquanto que valores de $ORp < 0\%$ indicam o contrário. O evento é equiprovável em ambos os grupos se $ORp = 0\%$.

No modelo de regressão logística o *odds ratio* também pode ser interpretado como a contribuição de cada variável para o modelo, ou seja, é uma maneira de quantificar o quanto haverá de variação na probabilidade de ocorrência do evento quando $X_i = 1$ em relação a $X_i = 0$.

3.14 Resíduos

Assim como os testes feitos anteriormente, a análise de resíduos é um componente importante na avaliação da qualidade do ajuste do modelo, uma vez que, segundo Collett (2003), existem várias razões para a inadequabilidade de um modelo, tais como problemas de especificação, modelagem sem variáveis que poderiam compor o modelo, pontos de alavanca, influentes ou outliers, ou alguma violação de pressupostos do modelo.

Segundo Carelli (2017), o cálculo dos resíduos é a base para a obtenção destas medidas. Para o ajuste de regressão logística com dados binários, o resíduo de Pearson pode ser utilizado, sendo definido como:

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}.$$

Outra medida que pode ser adotada é o resíduo *deviance*, definido em Carelli (2017) como:

$$d_i = \begin{cases} -\sqrt{-2[y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)}, & \text{se sinal}(y_i - \hat{p}_i) = -1; \\ +\sqrt{-2[y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)}, & \text{se sinal}(y_i - \hat{p}_i) = +1. \end{cases}$$

Carelli (2017) cita que, por maior facilidade de interpretação, pode-se adotar as versões padronizadas para os resíduos de Pearson e *deviance*, obtidas, respectivamente, como:

$$r_{rPi} = \frac{r_i}{\sqrt{1-h_{ii}}} \quad \text{e} \quad r_{dPi} = \frac{d_i}{\sqrt{1-h_{ii}}}.$$

Outro ponto importante a ser avaliado é a presença de pontos muito influentes no modelo, medida por meio da distância de Cook, obtida em Paula (2013) da seguinte forma:

$$LD_i = \frac{r_{rPi}^2 h_{ii}}{(1-h_{ii})}.$$

Além destes, um outro modo de avaliar pontos influentes é a detecção de pontos de alavanca (*leverage points*). Para obter os pontos de alavanca, em Hosmer-Jr *et al.* (2013), precisa-se calcular a matriz *hat H*, da forma que segue:

$$H = \hat{Q}^{1/2} X (X^\top \hat{Q} X)^{-1} X^\top \hat{Q}^{1/2},$$

com a matriz $\hat{Q} = diag[\hat{p}_i(1-\hat{p}_i)]$. Calculada a matriz H, define-se os pontos de alavanca como $h = diag(H)$. Para efeito de classificação, uma observação é considerada influente se $h_i > 2(q+1)/n$.

Corroborando com Carelli (2017), o ferramental gráfico é essencial e que auxiliam no processo de análise e identificação das medidas diagnósticas. O gráfico envelope simulado, por exemplo, indica se o componente aleatório do modelo (no caso da regressão logística, a distribuição binomial) é o mais adequado. Já os gráficos de pontos de alavanca (h_{ii}) ou da distância de Cook auxiliam na identificação dos pontos de alavanca, enquanto que o gráfico de resíduos *versus* as probabilidades estimadas podem apontar discrepâncias ou inadequações no componente sistemático do modelo. Hosmer-Jr *et al.* (2013) recomendam cuidado na

análise destas medidas, uma vez que as propriedades destas na regressão logística não são necessariamente as mesmas presentes nos ajustes lineares.

Conforme é recorrente em trabalhos com esta temática, a análise diagnóstica do modelo proposto é feita por meio de programas computacionais que possam ser utilizados no *software* R. Os programas utilizados neste trabalho foram elaborados por Paula (2013), estando disponíveis em `https://www.ime.usp.br/~giapaula/textoregressao.htm`.

4 APLICAÇÃO E RESULTADOS

Neste Capítulo o foco será dado ao procedimento de modelagem e análise dos dados, realizadas por meio do *software* R (R CORE TEAM, 2019).

4.1 Testes de Independência

A etapa inicial da análise é o teste de independência em todas as variáveis avaliadas, que considera o nível de significância $\alpha = 0,05$. Para isto, foi-se aplicado o teste χ^2 para fazer a verificação individual do grau de associação entre cada uma das variáveis explicativas com a variável-resposta, conforme descrito na Seção 3.3. Todas possuem frequências maiores ou iguais a 5, permitindo o uso do teste em todas as variáveis. Os resultados da aplicação do teste estão descritos na Tabela 3. As variáveis destacadas em negrito obtiveram $p-$valores significativos, indicando que as mesmas possuem associação com a variável-resposta.

Tabela 3 – Resultados obtidos no teste χ^2 para as variáveis avaliadas.

	Nivel 0	Nível 1	$p-$valor
Q3	49	24	0,8612
Q4	63	10	0,9888
Q5	51	22	0,0774
Q6	22	51	**0,0196**
Q7	37	36	0,0759
Q8	34	39	**0,0156**
Q9	33	40	**0,0003**
Q10	43	30	0,1434
Q11	35	38	**0,0075**
Q12	68	5	1,0000
Q13	48	25	**0,0173**
Q14	62	11	0,7561
Q15	24	49	**0,0199**
Q16	41	32	0,6472
Q17	27	46	**0,0099**
Q18	31	42	**0,0020**
Q19	24	49	**0,0009**
Q20	17	56	**0,0338**
Q21	47	26	0,1101
Q22	27	46	0,5283
Q23	20	53	0,1948
Q24	43	30	1,0000
Q25	27	46	0,1085
Q26	15	58	0,1135

Fonte: Elaborado pelo autor.

Pelos resultados do teste χ^2, as variáveis *Q6, Q8, Q9, Q11, Q13, Q15, Q17, Q18, Q19* e *Q20* seguirão para a próxima fase da construção do modelo. Assim, das 24 variáveis iniciais neste estudo, 10 serão avaliadas para a construção do modelo final.

4.2 Regressão Logística Simples

Nesta etapa começa a modelagem dos dados com as variáveis que foram selecionadas pelo teste de independência aplicado anteriormente. As questões selecionadas estão listadas abaixo (o questionário completo está no Anexo A):

I. O quanto você tem sentido algumas coisas?

   6. (*Domínio Psicológico*) Em que medida você acha que a sua vida tem sentido?

   8. (*Domínio Meio Ambiente*) Quão seguro(a) você se sente em sua vida diária?

   9. (*Domínio Meio Ambiente*) Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?

II. O quão completamente você tem sentido ou é capaz de fazer certas coisas?

   11. (*Domínio Psicológico*) Você é capaz de aceitar sua aparência física?

   13. (*Domínio Meio Ambiente*) Quão disponíveis para você estão as informações que precisa no seu dia-a-dia?

III. O quão bem ou satisfeito você se sentiu a respeito de vários aspectos de sua vida?

   15. (*Domínio Físico*) Quão bem você é capaz de se locomover?

   17. (*Domínio Físico*) Quão satisfeito(a) você está com sua capacidade de desempenhar as atividades do seu dia-a-dia?

   18. (*Domínio Físico*) Quão satisfeito(a) você está com sua capacidade para o trabalho?

   20. (*Domínio Relações Sociais*) Quão satisfeito(a) você está com suas relações pessoais (amigos, parentes, conhecidos, colegas)?

IV. Com que frequência você sentiu ou experimentou certas coisas?

   Não houve questões selecionadas.

Para cada uma das variáveis citadas, foi ajustado um modelo de regressão logística simples, tendo como variável-resposta a questão *QV* (*Como você avaliaria sua qualidade de vida?*). Para efeito de avaliação, usou-se o p-valor do Teste de Wald. O objetivo do teste é avaliar se os modelos univariados são significativos, e assim como na seção anterior, o interesse do teste é selecionar as variáveis cujo p-valor $\leq 0,05$. Os resultados da aplicação estão contidos na Tabela 4.

Tabela 4 – Resultados do ajuste dos modelos logísticos simples.

	$\hat{\beta}$	$EP(\hat{\beta})$	$\exp(\hat{\beta})$	p-valor
$Q6$	1,3683	0,5435	3,9286	0,0118
$Q8$	1,2905	0,4949	3,6346	0,0091
$Q9$	1,9315	0,5261	6,9000	0,0002
$Q11$	1,4240	0,5003	4,1538	0,0044
$Q13$	1,4040	0,5513	4,0714	0,0109
$Q15$	1,3257	0,5269	3,7647	0,0119
$Q17$	1,4191	0,5154	4,1333	0,0059
$Q18$	1,6582	0,5141	5,2500	0,0013
$Q19$	1,9169	0,5642	6,8000	0,0007
$Q20$	1,3863	0,5996	4,0000	0,0208

Fonte: Elaborado pelo autor.

Os valores de $\hat{\beta}$ obtidos são os coeficientes dos parâmetros estimados, $\exp(\hat{\beta})$ é a razão de chances, $\hat{EP}$ é o erro-padrão estimado e o p-valor é o nível de significância associados a cada um dos ajustes logísticos simples. Pela exposição dos resultados da Tabela 4, tem-se que todos os modelos univariados são significativos, por apresentarem p-valor $\leq 0,05$.

Nota-se que todas as variáveis têm coeficientes positivos. Isso significa que todas as variáveis possuem influência positiva em relação à variável resposta, ou seja, elas têm o efeito de aumentar a probabilidade de o idoso estar satisfeito com determinado aspecto de sua vida, o que corrobora com a tese de que uma resposta positiva à essas perguntas indica satisfação com a sua qualidade de vida.

Como exemplo desta interpretação, seja a questão 9, "Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?", que obteve o maior valor de razão de chance dentre as variáveis em estudo $(\exp(\hat{\beta}) = 6,90)$. Esse resultado indica que, ao dar uma resposta positiva a esta questão, o idoso tem aproximadamente 7 vezes mais chance de estar satisfeito com a sua qualidade de vida, em relação a um idoso que tenha dado uma resposta negativa a esta mesma questão.

Também é possível interpretar esse resultado em termos percentuais, utilizando-se da fórmula $(\exp(\hat{\beta}) - 1) \times 100\%$, em que tem-se $(6,90 - 1) \times 100\% = 590\%$, apontando que uma resposta positiva a essa pergunta aumenta em 590% a probabilidade de estar satisfeito com a sua qualidade de vida. Essa maneira de interpretação também se aplica às demais variáveis estudadas.

4.3 Regressão Logística Múltipla

Neste tópico, foi feito um modelo de regressão logística múltipla englobando todas as variáveis selecionadas. Para a construção dessa análise foram aceitas todas as variáveis cujo p-valor $\leq 0,10$ com o objetivo de incluir o maior número de variáveis na construção do modelo. Os resultados estão apresentados na Tabela 5, e os valores destacados em negrito foram apontados como significativos no modelo.

Tabela 5 – Resultados do ajuste do modelo logístico múltiplo.

	$\hat{\beta}_i$	$EP(\hat{\beta}_i)$	z	$p-$valor (Wald)
Intercepto	-2,9114	0,9183	-3,17	**0,0015**
$Q6$	-0,1341	0,8433	-0,16	0,8736
$Q8$	-0,2747	0,7344	-0,37	0,7084
Q9	1,7126	0,7195	2,38	**0,0173**
$Q11$	0,3652	0,6967	0,52	0,6002
Q13	1,1451	0,6663	1,72	**0,0857**
$Q15$	0,6898	0,8409	0,82	0,4120
$Q17$	-0,5101	0,9409	-0,54	0,5877
$Q18$	0,8639	0,8938	0,97	0,3338
$Q19$	0,8028	0,7487	1,07	0,2836
$Q20$	0,8563	0,7848	1,09	0,2752

Fonte: Elaborado pelo autor.

Os valores dos $\hat{\beta}$ obtidos são os coeficientes dos parâmetros estimados, EP é o erro-padrão, z é o valor obtido a partir do quantil da distribuição normal padrão e o p-valor é o nível de significância associados ao teste de Wald para cada um dos parâmetros do modelo múltiplo. Percebe-se pelos resultados da Tabela 5 que, no modelo inicial com todas as variáveis presentes, algumas variáveis deixam de ser significativas na presença de outras, considerando p-valor $> 0,10$. Assim, se faz necessário o uso de técnicas de seleção de variáveis.

4.4 Seleção de Variáveis

Assim como feito em Carelli (2017) e Freitas (2018), todos os subconjuntos listados pelo pacote `bestglm` foram avaliados, sendo esta análise importante para verificar a existência de efeitos simultâneos em alguns destes subconjuntos. O Quadro 6 apresenta 10 possíveis modelos, com a letra P significa que a variável está presente no modelo e NP, caso não esteja presente.

Tabela 6 – Lista dos melhores modelos selecionados, segundo o critério do AIC.

Modelo	$Q6$	$Q8$	$Q9$	$Q11$	$Q13$	$Q15$	$Q17$	$Q18$	$Q19$	$Q20$	AIC
1	NP	NP	P	NP	P	NP	NP	P	NP	P	81,53
2	NP	NP	P	NP	P	NP	NP	NP	P	P	81,54
3	NP	NP	P	NP	P	P	NP	NP	P	NP	81,75
4	NP	NP	P	NP	P	NP	NP	P	NP	NP	81,92
5	NP	NP	P	NP	P	P	NP	NP	NP	NP	81,93
6	NP	NP	P	NP	P	NP	NP	P	P	P	81,96
7	NP	NP	P	NP	NP	NP	NP	P	NP	P	82,05
8	NP	NP	P	NP	P	NP	NP	NP	P	NP	82,20
9	NP	NP	P	NP	P	NP	NP	P	P	NP	82,20
10	NP	NP	P	NP	P	P	NP	NP	P	P	82,23

Fonte: Elaborado pelo autor.

Conforme citado na seção 3.6.1, para valores de $n < 40q$ o AIC se torna um critério problemático para a escolha do melhor modelo. Assim, o AIC não é o critério mais apropriado para fazer a seleção de variáveis neste caso, e se faz necessário adotar outros critérios para fazer a seleção das variáveis, tais como o AICc e o BIC. A Tabela 7 traz uma comparação dos valores obtidos em cada um dos critérios, com os respectivos números de parâmetros de cada modelo selecionável.

Tabela 7 – Comparação entre os valores obtidos aplicando os critérios de seleção de variáveis aos modelos selecionados via `bestglm`.

Modelo	q	AIC	BIC	AICc
1	5	83,530	94,982	84,425
2	5	83,545	94,997	84,440
3	5	83,745	95,197	84,641
4	4	83,922	93,084	84,511
5	4	83,930	93,092	84,518
6	6	83,956	97,699	85,229
7	4	84,052	93,214	84,640
8	4	84,198	93,360	84,786
9	5	84,202	95,654	85,097
10	6	84,226	97,969	85,499

Fonte: Elaborado pelo autor.

Nota-se pelo que foi apresentado na Tabela 7 que, em que pese o AICc ser uma correção do AIC para ser utilizado para amostras pequenas, os resultados obtidos sob este critério não diferem muito do que foi obtido na aplicação do AIC, selecionando o Modelo 1. No entanto, ao analisar os resultados obtidos na aplicação do BIC, o Modelo 4 aparenta ser o mais adequado para o prosseguimento da análise. Entretanto, o Modelo 1 possui um parâmetro a mais em relação ao Modelo 4.

Para ajudar na decisão entre os dois modelos, será utilizada a função `accuracy` do

pacote `rcompanion` do *software* R, descrita na Seção 3.9. Estes dados estão descritos na tabela 8:

Tabela 8 – Estatísticas de ajuste dos modelos 1 e 4.

Modelo	EQM	Pseudo-R^2	$CV\%$
1	0,164	0,337	74,0
4	0,173	0,301	75,9

Fonte: Elaborado pelo autor.

Os resultados mostrados na Tabela 8 indicam que o ajuste representado no Modelo 1 é melhor que o ajuste do Modelo 4. Os coeficientes D^2 para os modelos são, respectivamente, 0,268 e 0,244, com D^2_{aj} de 0,225 e 0,211, também apontando para um melhor ajuste do Modelo 1 em relação ao Modelo 4. Com base nestes critérios, definiu-se pelo Modelo 1 para o prosseguimento da análise.

Os resultados obtidos para o Modelo 1 estão na Tabela 9:

Tabela 9 – Resultados do ajuste do modelo 1 selecionado na Tabela 6.

	$\hat{\beta}_i$	EP	z	$p-$valor (Wald)
Intercepto	-2,5448	0,8253	-3,08	0,0020
$Q9$	1,6885	0,5791	2,92	0,0035
$Q13$	0,9875	0,6334	1,56	0,1190
$Q18$	1,1115	0,5900	1,88	0,0596
$Q20$	1,1251	0,7411	1,52	0,1290

Fonte: Elaborado pelo autor.

O modelo possui um desvio de 73,529 e 68 graus de liberdade. Avaliando o Teste Razão de Verossimilhança, comparando com o modelo geral com as 10 variáveis, obteve-se um $p-$valor de 0,8352, indicando que a redução das variáveis não implicou em perda de informação. Deste modo, prossegue-se a análise com o Modelo 1.

4.5 Adequabilidade do Modelo

Os resultados para o teste de Hosmer-Lemeshow aplicado ao modelo proposto estão na Tabela 10:

Tabela 10 – Resultados do Teste de Hosmer-Lemeshow para o modelo proposto.

Estatística $\hat{C}$	Graus Liberdade	$p-$valor
1,8165	8	0,9861

Fonte: Elaborado pelo autor.

Assim, para um nível de significância $\alpha = 0,05$, verifica-se que o teste de hipóteses descrito na seção 3.11 para o modelo proposto obteve um $p-$valor não-significativo, sinalizando uma adequação do ajuste do modelo, o que se evidencia também pela estatística $\hat{C} = 1,8165$ com 8 graus de liberdade e 10 grupos. Além disso, o teste dividiu o modelo em 8 intervalos de probabilidade estimados para verificação da situação esperada com a observada, conforme a Tabela 11, e na qual o valor 0 representa uma qualidade de vida *Ruim*, enquanto o valor 1 indica uma qualidade de vida *Boa*.

Tabela 11 – Comparação entre frequências observadas e esperadas do modelo selecionado.

Int. Prob. Est.	0 obs.	1 obs.	0 esp.	1 esp.
$[0,0728-0,093]$	7,00	1,00	7,42	0,58
$(0,093-0,195]$	9,00	1,00	8,08	1,92
$(0,195-0,412]$	2,00	2,00	2,62	1,38
$(0,412-0,424]$	5,00	3,00	4,61	3,39
$(0,424-0,567]$	4,00	5,00	3,91	5,09
$(0,567-0,686]$	1,00	4,00	1,68	3,32
$(0,686-0,799]$	4,00	14,00	3,75	14,25
$(0,799-0,914]$	1,00	10,00	0,94	10,06

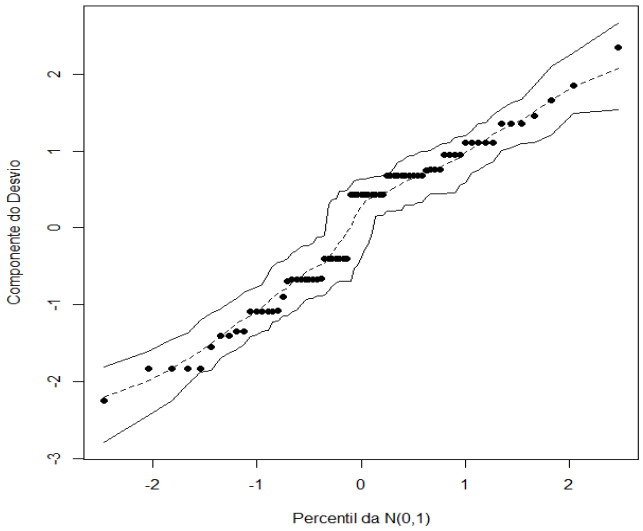
Fonte: Elaborado pelo autor.

Os resultados descritos na Tabela 11 indicam uma boa concordância entre os valores observados em comparação aos valores esperados, assim seguindo em análise o modelo proposto.

4.5.1 Análise Diagnóstica

O passo inicial para a análise de resíduos é a construção do envelope simulado para o modelo proposto, conforme a Figura 5. Nota-se que todos os pontos estão dentro dos limites do envelope, sem nenhum comportamento atípico de o que se espera ao elaborar este tipo de gráfico.

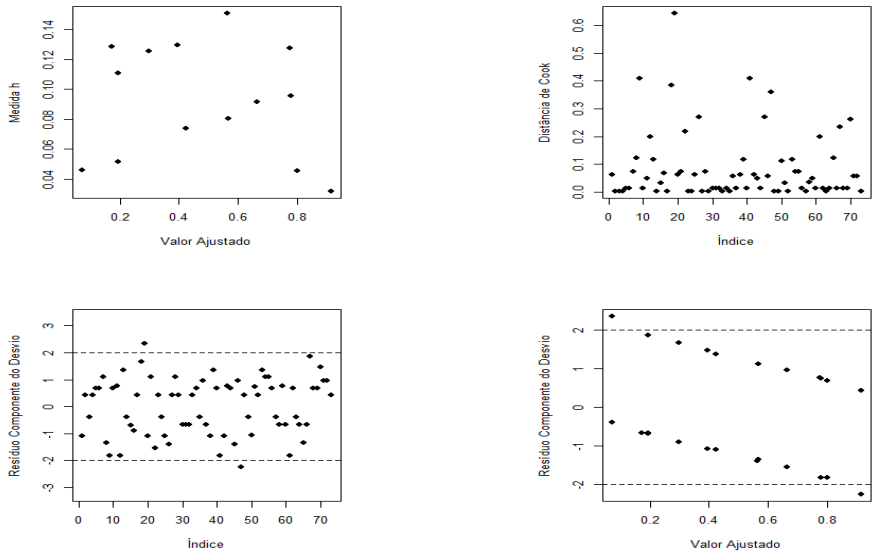
Figura 5 – Gráfico quantil-quantil Normal com envelope simulado sob a suposição de componente aleatória binomial para o modelo proposto com as 73 observações.



Fonte: Elaborado pelo autor.

Na sequência, foram construídos outros 4 gráficos, mostrados na Figura 6, nos quais estão representados, respectivamente, os possíveis pontos de alavanca, a Distância de Cook, os resíduos deviance e uma comparação entre o resíduo deviance e o respectivo valor ajustado.

Figura 6 – Gráficos para a análise diagnóstica do modelo.



Fonte: Elaborado pelo autor.

Partindo dos gráficos da Figura 6, foram selecionadas algumas observações, por se destacarem em valores absolutos em cada um dos gráficos, fazendo-se necessário fazer a análise diagnóstica inicial retirando-se individualmente cada uma das observações destacadas. A Tabela 12 traz um resumo das observações que foram destacadas, com as probabilidades estimadas $(\hat{\pi})$, valor de alavanca (h), distância de Cook (LD) e resíduo do desvio (d).

Tabela 12 – Valores de $\hat{\pi}$, h, LD e d das seis observações destacadas.

ID	h	LD	d	$\hat{\pi}$
#10	0,10	0,41	-1,83	0,78
#21	0,05	0,64	2,34	0,07
#32	0,15	0,27	-1,40	0,56
#48	0,10	0,41	-1,83	0,78
#53	0,15	0,27	-1,40	0,56
#55	0,03	0,36	-2,25	0,91

Fonte: Elaborado pelo autor.

Na sequência, a Tabela 13 apresenta os valores dos ajustes obtidos nos modelos sem a presença individual das observações destacadas na tabela anterior, comparando com os resultados do ajuste do modelo com todas as observações e com o ajuste do modelo sem as seis observações destacadas.

Tabela 13 – Coeficientes dos modelos ajustados sem as observações destacadas individualmente.

	Sem #32	Sem #53	Sem #10	Sem #48	Sem #55	Sem #21
Intercepto	-2,3464	-2,3464	-2,6214	-2,6214	-2,7878	-3,3266
$Q9$	1,7738	1,7738	1,8518	1,8518	1,8616	1,9317
$Q13$	0,8985	0,8985	1,2748	1,2748	1,2603	1,0972
$Q18$	1,2382	1,2382	0,8931	0,8931	1,2157	1,2995
$Q20$	0,8350	0,8350	1,2547	1,2547	1,2400	1,6385
AIC	81,6420	81,6420	80,1320	80,1320	78,2420	77,6050
$\hat{C}$	1,7290	1,7290	3,0581	3,0581	2,6113	–
Deviance	71,6147	71,6147	70,1316	70,1316	68,2423	67,6053
GL	67	67	67	67	67	67
$p-$valor	0,9882	0,9882	0,9307	0,9307	0,9563	–

Fonte: Elaborado pelo autor.

Observando as Tabelas 12 e 13, nota-se que:

– Pontos de Alavanca: no geral, são valores relativamente baixos. Entretanto, as observações #32 e #53 estão acima do limite aceitável para o modelo em estudo, sendo considerados pontos de alavanca. Pelos dados da Tabela 12, ambas as observações individualmente têm a mesma influência sobre o modelo, e a exclusão de ambas não modificaram significativamente a estimação dos parâmetros do modelo, além de não ter impacto significativo nas demais estatísticas associadas.
– Distância de Cook: Os valores mais destacados sob este critério são 0,64 (observação #21) e 0,41 (observações #10 e #48). A retirada individual das observações #10 e #48 possuem o mesmo peso para o ajuste do modelo, enquanto a remoção da observação #55 representou melhoria significativa no ajuste inicial.
– Resíduo Componente do Desvio: os maiores valores absolutos pertencem às observações #21 e #55, ultrapassando os limites pré-definidos, indicando que estas observações podem ter comportamento distinto das demais. Observa-se que são idosos que deram uma resposta no questionário, porém no modelo preditivo ajustado obtiveram probabilidades inversas às respostas dadas na variável QV, indicando que as respostas dadas podem estar contradizendo ao que foi respondido às variáveis explicativas. Não se pode garantir que estas observações devam ser removidas do modelo.
– Resíduo Componente do Desvio *versus* Valor Ajustado: para este gráfico, não se nota nenhum comportamento anormal. No entanto, pelo fato da variável resposta do modelo ser binária, há uma maior dificuldade de interpretação (CARELLI, 2017), o que é corroborado pelo autor deste trabalho.

O resultado do ajuste após a remoção da observação #21 chama a atenção, uma vez que a remoção desta observação em particular teve alta influência no teste de Hosmer e Lemeshow, induzindo o algoritmo a gerar um intervalo de probabilidade sem observações e inviabilizando a estimativa do respectivo teste com a configuração que está sendo adotada.

Mesmo destacando todos estes pontos, no modelo proposto em que não se exclui nenhuma observação não há indícios estatísticos de inadequabilidade, com AIC de 83,53, estatística $\hat{C}$ com $p-$valor de 0,9861 e um Desvio de 73,53 com 68 graus de liberdade. A Figura 5 apresentou o gráfico quantil-quantil Normal com envelopes simulados sob a suposição de componente aleatório binomial sem nenhum comportamento atípico, corroborando com o que foi exposto. Assim, o autor deste trabalho não julgou necessário fazer uma análise diagnóstica

com a remoção das seis observações destacadas.

Antes de decidir o modelo final, e objetivando a otimização do modelo, fez-se uma análise mais detalhada dos dois pontos que foram destacados pelo resíduo componente do desvio. Avaliou-se o comportamento do modelo com as 73 observações, em relação ao mesmo modelo sem a presença destas observações. Avaliando graficamente a situação, percebe-se que os pontos destacados possuem comportamento bastante distinto dos demais resíduos, extrapolando o intervalo pré-definido $[-2,2]$.

A remoção das observações #21 e #55 conjuntamente fez com que todos os parâmetros do modelo sejam significativos, com um desvio de 61,72 e $p-$valor da estatística $\hat{C}$ de 0,9921. No entanto, mesmo com as observações destacadas, conclui-se que não há indícios de inadequabilidade no modelo com todas as observações presentes, sendo que a remoção destas observações pode induzir à estimação de falsos comportamentos. Assim, o modelo proposto com as 73 observações será o modelo final deste trabalho.

4.5.2 Classificação

Para o modelo de regressão logística múltipla escolhido, considera-se inicialmente o método mais comum deste tipo de análise, classificando como *Boa* as probabilidades estimadas ($\hat{\pi}$) com valores maiores ou iguais 0,5, e de *Ruim*, caso contrário. O resultado obtido para o modelo estimado está descrito na Tabela 14:

Tabela 14 – Tabela de classificação para o modelo com ponto de corte em 0,5.

Observado	Predito		Total
	$\hat{Y}=0$	$\hat{Y}=1$	
$Y=0$	23	10	33
$Y=1$	7	33	40
Total	30	43	73

Fonte: Elaborado pelo autor.

Pelos resultados obtidos na Tabela 14, tem-se que a taxa geral de classificação do modelo é de $100[(23+33)/73] = 76,71\%$, com nível de especificidade de $100(23/33) = 69,69\%$ e sensibilidade de $100(33/40) = 82,5\%$.

Ou seja, significa que $69,69\%$ dos casos de resposta negativa em relação à qualidade de vida na terceira idade são classificados de maneira correta, demostrando uma certa facilidade para especificar essa situação. Por outro lado, $82,5\%$ dos casos de resposta positiva em relação à

qualidade de vida na terceira idade são classificados corretamente, demostrando grande facilidade na previsão desse comportamento. Também tem-se uma taxa de $100-69,69=30,31\%$ de falsos positivos, e $100-82,5=17,5\%$ de falsos negativos.

A taxa de falsos positivos de $30,31\%$ é algo que chama a atenção, por poder estar atrelada ao tamanho amostral utlizado, ou ao comportamento geral dos idosos que participaram desta pesquisa, que independente da maneira como avaliam a sua qualidade de vida, responderam positivamente às questões. Outro fator que pode ter influência nessa taxa é o ponto de corte escolhido. A codificação das respostas foi descrita na seção 3.2.

A manutenção do ponto de corte em 0,5 indica que a taxa de $30,31\%$ de falsos positivos pode atrapalhar a classificação estimada dos idosos. Sendo assim, faz-se necessário adotar um critério de escolha de um ponto de corte, conforme descrito na Seção 3.12.

Para auxiliar nesta decisão, usou-se a função `coords()` do pacote `pROC` do *software* R. A função retorna uma lista com possíveis valores para o ponto de corte, sendo possível visualizar também os valores de sensibilidade, especificidade e acurácia para cada um destes pontos. Utilizando-se desta ferramenta, o autor deste trabalho determinou que o ponto de corte ideal para o modelo é de $0,5651$.

Tabela 15 – Tabela de classificação para o modelo com ponto de corte em 0,5651.

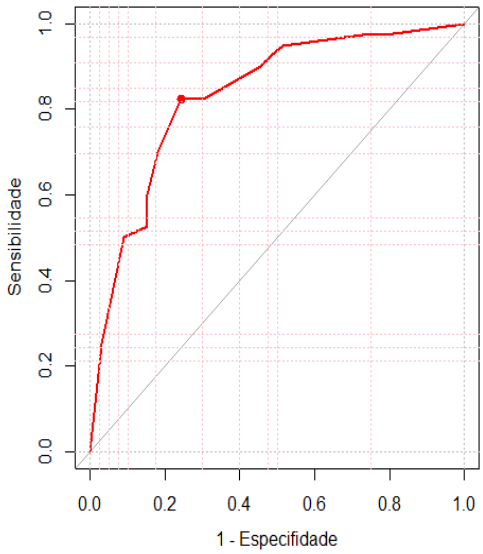
Observado	Predito		Total
	$\hat{Y}=0$	$\hat{Y}=1$	
$Y=0$	25	8	33
$Y=1$	7	33	40
Total	30	43	73

Fonte: Elaborado pelo autor.

Pelos resultados obtidos na Tabela 15, observa-se uma redução da taxa de falsos positivos, de $30,31\%$ para $24,25\%$, e tem-se uma taxa geral de classificação do modelo de $100[(25+33)/73]=79,42\%$. O nível de especificidade para este ponto de corte é $100(25/33)=75,75\%$ e a sensibilidade se mantém em $82,5\%$.

Para finalizar o estudo do modelo final, foi obtida a curva ROC que está apresentada na Figura 7, que simula uma curva que pode ser atingida por meio de uma classificação, de maneira aleatória, como *Bom* ou *Ruim*.

Figura 7 – Curva ROC obtida para o modelo em estudo.



Fonte: Elaborado pelo autor.

A área sob a curva do gráfico da Figura 7 teve o valor de 0,833, bastante próximo de 1, indicando que o modelo possui ótimo índice de previsões corretas. Em termos práticos, pelo fato de não ter sido excluída nenhuma observação além das que foram removidas no tratamento inicial, os resultados obtidos podem ser considerados satisfatórios.

4.5.3 Interpretação

O objetivo deste tópico é apresentar o resultado final do modelo obtido por meio da regressão logística, assim como identificar quais fatores dos domínios físico, psicológico, social e ambiental influenciam a qualidade de vida dos idosos atendidos pelo HUWC. As variáveis que entraram na composição do modelo estão listadas abaixo:

I. O quanto você tem sentido algumas coisas?

   *Q*9. (*Domínio Meio Ambiente*) Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?

II. O quão completamente você tem sentido ou é capaz de fazer certas coisas?

   *Q*13. (*Domínio Meio Ambiente*) Quão disponíveis para você estão as informações que precisa no seu dia-a-dia?

III. O quão bem ou satisfeito você se sentiu a respeito de vários aspectos de sua vida?

   *Q*18. (*Domínio Físico*) Quão satisfeito(a) você está com sua capacidade para o trabalho?

   *Q*20. (*Domínio Relações Sociais*) Quão satisfeito(a) você está com suas relações pessoais

(amigos, parentes, conhecidos, colegas)?

IV. Com que frequência você sentiu ou experimentou certas coisas?

Não houve questões selecionadas.

A Tabela 16 traz as estimativas das variáveis identificadas.

Tabela 16 – Estimativas dos parâmetros do modelo adotado.

Variáveis	$\hat{\beta}_i$	$\exp(\hat{\beta}_i)$
Intercepto	-2,5448	-
$Q9$	1,6885	5,4112
$Q13$	0,9875	2,6844
$Q18$	1,1115	3,0389
$Q20$	1,1251	3,0806

Fonte: Elaborado pelo autor.

Todas as variáveis do modelo têm coeficientes positivos. Isto significa que todas as variáveis possuem influência positiva em relação à variável resposta, ou seja, elas têm o efeito de aumentar a probabilidade de o idoso estar satisfeito com determinado aspecto de sua vida.

Tomando por exemplo desta interpretação, seja a questão “Quão saudável é o seu ambiente físico?" ($Q9$). Conclui-se que o idoso que deu uma resposta positiva a esta questão tem aproximadamente 5,4 vezes mais chance de possuir boa qualidade de vida de que um outro idoso que tenha respondido a esta mesma questão de maneira negativa.

Uma característica importante da utilização da regressão logística é a sua capacidade preditiva. Para demonstrar isto, assim como em Freitas (2018) serão usadas situações fictícias para fazer a simulação da probabilidade.

Seja o evento Q:{O idoso possuir boa qualidade de vida na terceira idade}. Assim, tem-se que:

$$\mathbb{P}(Q) = \frac{1}{1+e^{-(-2.5448+(1.6885\times Q9)+(0.9875\times Q13)+(1.1115\times Q18)+(1.1251\times Q20))}},$$

em que:

- $Q9$ recebe valor 1 se o idoso considerar que o seu ambiente físico é *Saudável*, e 0 caso considere *Não-saudável*;
- $Q13$ recebe valor 1 se o idoso considerar que há *Muita Disponibilidade* de informações necessárias para o seu dia-a-dia, e 0 caso considere que há *Pouca Disponibilidade*;
- $Q18$ recebe valor 1 se o idoso estiver *Satisfeito* com a sua capacidade para o trabalho, e 0 caso se estiver *Insatisfeito*;

– $Q20$ recebe valor 1 se o idoso esteja *Satisfeito* com as suas relações pessoais, e 0 caso contrário.

Supondo três idosos, A, B e C, que responderam o questionário da seguinte maneira: o idoso A respondeu positivamente a todas as questões, recebendo valor 1 nos coeficientes. O idoso B respondeu negativamente às questões $Q9$ e $Q20$ e positivamente às demais, enquanto o idoso C deu respostas opostas às do idoso B, sempre sendo atribuído valor 1 aos coeficientes com respostas positivas. Assim,

$$\mathbb{P}(Q_A) = \frac{1}{1+e^{-(-2.5448+(1.6885\times 1)+(0.9875\times 1)+(1.1115\times 1)+(1.1251\times 1))}} \approx 0,91$$

$$\mathbb{P}(Q_B) = \frac{1}{1+e^{-(-2.5448+(1.6885\times 1)+(0.9875\times 0)+(1.1115\times 0)+(1.1251\times 1))}} \approx 0,56$$

$$\mathbb{P}(Q_C) = \frac{1}{1+e^{-(-2.5448+(1.6885\times 0)+(0.9875\times 1)+(1.1115\times 1)+(1.1251\times 0))}} \approx 0,39$$

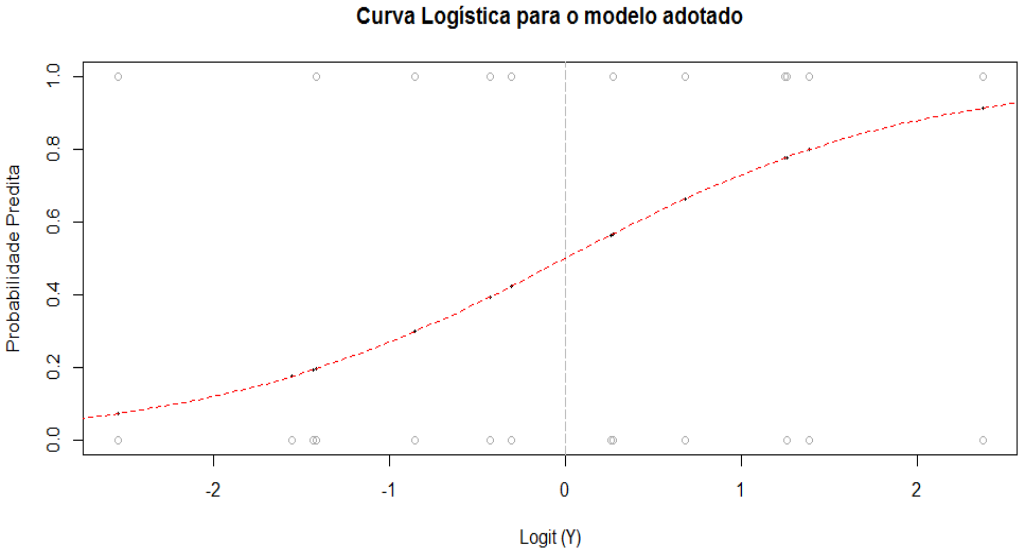
Pelos resultados acima, tem-se que o Idoso A, ao responder positivamente a todas as questões ($Q9, Q13, Q18$ e $Q20$), tem 91% ($\mathbb{P}(Q_A) \approx 0,91$) de chance de ter uma boa qualidade de vida.

Já o Idoso B, ao responder que tem pouca disponibilidade de informações necessárias ao seu dia-a-dia e está insatisfeito com a sua capacidade para o trabalho, mas está satisfeito com suas relações pessoais e vieem um ambiente físico saudável possui chance de 56% ($\mathbb{P}(Q_B) \approx 0,56$) de ter boa qualidade de vida.

Por fim, o Idoso C, ao responder que seu ambiente físico é não-saudável e que está insatisfeito com suas relações pessoais, mas possui muita disponibilidade de informações necessárias ao seu dia-a-dia e está satisfeito com sua capacidade para o trabalho tem chance de possuir boa qualidade de vida de 39% ($\mathbb{P}(Q_C) \approx 0,39$).

A curva logística ajustada do modelo final é mostrada na Figura 8:

Figura 8 – Curva ajustada do modelo final adotado.



Fonte: Elaborado pelo autor.

A curva apresentada na Figura 8, representando as probabilidades preditas em relação ao que foi observado, indica o efeito que alterações nas variáveis explicativas têm sobre a variável resposta, tendo comportamento similar à de uma função de distribuição acumulada de uma variável aleatória contínua.

5 CONCLUSÕES

Neste Capítulo serão apresentadas algumas considerações a partir dos resultados obtidos da análise dos dados.

5.1 Considerações Finais

O objetivo geral deste trabalho foi investigar, por meio da regressão logística, quais fatores têm influência na qualidade de vida dos idosos que são atendidos no Ambulatório de Geriatria do HUWC.

De início, foi preciso retirar algumas observações com respostas faltantes, e a análise foi feita a partir das observações restantes. Após isto, foi verificado o nível de relação das variáveis explicativas com a variável resposta utilizando-se neste processo o teste χ^2, não sendo necessário o uso de outros testes. Com base neste teste, das 24 variáveis independentes iniciais, 10 possuíam relação de dependência com a variável resposta.

Na sequência, as dez variáveis explicativas foram avaliadas individualmente criando-se modelos logísticos univariados, onde se comprovou que todas possuem influência sobre a variável resposta *QV* (Como você avaliaria sua qualidade de vida?). Após isto iniciou-se a construção do modelo logístico múltiplo, objetivando identificar quais fatores, dentro dos domínios físico, psicológico, social e ambiental, possuem influência na qualidade de vida dos idosos. O modelo foi definido com a presença de 4 destas 10 variáveis.

A seguir, foi verificada a adequabilidade do modelo proposto, aplicadas as técnicas de diagnóstico e estabelecida uma classificação ideal que faça a distinção, com base nas probabilidades estimadas, da qualidade de vida dos idosos entre *Boa* ou *Ruim*, e avaliado o poder preditivo do modelo criado, com resultados considerados satisfatórios.

O Modelo Logístico Múltiplo final foi composto com as seguintes variáveis:

I. O quanto você tem sentido algumas coisas?

   9. (*Domínio Meio Ambiente*) Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?

II. O quão completamente você tem sentido ou é capaz de fazer certas coisas?

   13. (*Domínio Meio Ambiente*) Quão disponíveis para você estão as informações que precisa no seu dia-a-dia?

III. O quão bem ou satisfeito você se sentiu a respeito de vários aspectos de sua vida?

18. (*Domínio Físico*) Quão satisfeito(a) você está com sua capacidade para o trabalho?

20. (*Domínio Relações Sociais*) Quão satisfeito(a) você está com suas relações pessoais (amigos, parentes, conhecidos, colegas)?

Todas as variáveis remanescentes no modelo possuem influência positiva sobre a qualidade de vida dos idosos, o que é um comportamento esperado, uma vez que são fatores relevantes para uma boa qualidade de vida dos idosos, e apontam para o desenvolvimento de projetos e políticas que visem fortalecer estes fatores.

Cada uma das variáveis selecionadas no modelo final é ligada a um fator importante na qualidade de vida dos idosos. Assim, um idoso que convive em um ambiente saudável, é bem-informado sobre o que acontece ao seu redor, está satisfeito com a sua capacidade física e com suas relações sociais ou pessoais tem uma grande chance de possuir uma boa qualidade de vida, o que corrobora com a tese inicial deste trabalho.

Assim, a técnica de regressão logística se mostrou bastante contributiva para o objetivo geral deste estudo, com interpretabilidade clara e um bom poder preditivo, ao ser possível prever a probabilidade de o idoso estar satisfeito com a sua qualidade de vida, com base nas quatro variáveis componentes do modelo logístico múltiplo.

5.2 Sugestões para Trabalhos Futuros

Ressalta-se que os resultados obtidos nesta pesquisa não podem ser generalizados, sugerindo-se aplicar a mesma metodologia em um tamanho amostral maior, com o objetivo de obter um modelo mais preciso. Também é sugerido fazer um estudo logístico multivariado para verificar a influência destas mesmas variáveis em relação às variáveis-resposta *Como você avaliaria sua qualidade de vida?* e *Quão satisfeito(a) você está com a sua saúde?*.

REFERÊNCIAS

AGÊNCIA CÂMARA DE NOTÍCIAS. **Brasil está atrasado nas políticas públicas para idosos, dizem especialistas**. 2019. Disponível em: http://www.camara.leg.br/noticias/602947-brasil-esta-atrasado-nas-politicas-publicas-para-idosos-dizem-especialistas/. Acesso em: 30 jul. 2020.

AGRESTI, A. **An introduction to categorical data analysis**. 2. ed. New York: John Wiley & Sons, 2007.

AKAIKE, H. Anew look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ANDERSON, M. I. P.; ASSIS, M. d.; PACHECO, L. C.; SILVA, E. d.; MENEZES, I. S.; DUARTE, T.; STORINO, F.; MOTTA, L. Saúde e qualidade de vida na terceira idade. **Textos envelhecimento**, v. 1, n. 1, p. 23–43, 1998.

BETA ANALÍTICA. **Existe R^2 para modelos lineares generalizados (GLMS)?** 2019. Disponível em: https://betaanalitica.com.br/existe-r\%C2\%B2-para-modelos-lineares-generalizados-glms/. Acesso em: 31 out. 2020.

BIBLIOTECA VIRTUAL DA SAÚDE. **Dicas de Saúde em 5 passos**. 2013. Disponível em: https://bvsms.saude.gov.br/bvs/dicas/260_qualidade_de_vida.html. Acesso em: 31 out. 2020.

BRASIL. Lei nº 10.741, de 1º de outubro de 2003. dispõe sobre o estatuto do idoso e dá outras providências. **Diário Oficial da União**, p. 1–1, 2003.

BRASILPREV. **Aumento da expectativa de vida demanda melhor planejamento financeiro**. 2019. Disponível em: http://www2.brasilprev.com.br/Empresa/SalaDeImprensa/Releases/Paginas/Aumentodaexpectativadevidademandamelhorplanejamentofinanceiro. Acesso em: 31 out. 2020.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. **Sociological Methods and Research**, v. 33, p. 261–304, May 2004.

CALDAS, R. W.; LOPES, B.; AMARAL, J. N. **Políticas Públicas**: conceitos e práticas. [*S. l.*: *s. n.*], 2008. v. 7.

CARDOSO, J. H.; COSTA, J. S. D. Características epidemiológicas, capacidade funcional e fatores associados em idosos de um plano de saúde. **Ciência & Saúde Coletiva**, SciELO, v. 15, p. 2871 – 2878, Set. 2010.

CARELLI, D. S. C. **Predição de aprovação em um curso em Tecnologia da Informação no Instituto Metrópole Digital da UFRN**: uma aplicação da análise de regressão logística. 102 f. Monografia (Centro de Ciências Exatas e da Terra) – Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, 2017.

CARVALHO, D. A. **Fatores que influenciam na qualidade de vida dos idosos atendidos em um hospital universitário**. 64 f. Dissertação (Mestrado Profissional em Políticas Públicas e Gestão da Educação Superior) – Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal do Ceará, Fortaleza, 2019.

CASELLA, G.; BERGER, R. L. **Inferência estatística**: Tradução da 2ª edição norte-americana. 2. ed. São Paulo: Centage Learning, 2011.

COLLETT, D. **Modeling Binary Data**. Boca Ratón, FL: Chapman & Hall/CRC Press, 2003.

DAVISON, A. C. Biometrika centenary: Theory and general methodology. **Biometrika**, v. 88, p. 13–52, 2001.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FREITAS, P. C. S. **Análise da influência dos fatores extrínsecos e intrínsecos na motivação dos alunos de um curso de ensino superior por meio da regressão logística**. 103 f. Monografia (Graduação em Estatística - Centro de Ciências) – Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, 2018.

G1. **Expectativa de vida do brasileiro ao nascer é de 76,3 anos em 2018, diz IBGE**. 2019. Disponível em: https://g1.globo.com/bemestar/noticia/2019/11/28/expectativa-de-vida-do-brasileiro-ao-nascer-foi-de-763-anos-em-2018-diz-ibge.ghtml. Acesso em: 10 out. 2020.

GONZALEZ, L. A. **Regressão logística e suas aplicações**. 46 f. Monografia (Graduação em Ciência da Computação - Centro de Ciências Exatas e Tecnológicas) – Universidade Federal do Maranhão, São Luís, 2018.

HOSMER, D. W.; LEMESHOW, S. Goodness of fit tests for the multiple logistic regression model. **Communications in statistics-Theory and Methods**, Taylor & Francis, Nova York, v. 9, n. 10, p. 1043–1069, 1980.

HOSMER-JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. New York: John Wiley & Sons, 2013. v. 2.

HURVICH, C. M.; TSAI, C. L. Regression and time series model selection in small samples. **Biometrika**, v. 76, p. 297–307, 1989.

IBGE. **Projeções da População**. 2018. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?=&t=resultados. Acesso em: 31 jul. 2020.

IBGE. **Tábuas Completas de Mortalidade**. 2018. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9126-tabuas-completas-de-mortalidade.html. Acesso em: 31 out. 2020.

INSTITUTO PHD. **O que é melhor**: pesquisa de dados primários ou secundários? 2019. Disponível em: https://www.institutophd.com.br/o-que-e-melhor-pesquisa-de-dados-primarios-ou-secundarios/. Acesso em: 03 nov. 2020.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **The Annals of Mathematical Statistics**, JSTOR, v. 22, n. 1, p. 79–86, 1951.

MANGIAFICO, S. **Rcompanion**: functions to support extension education program evaluation. [*S. l.*], 2016. R package version 2.3.25. Disponível em: https://CRAN.R-project.org/package=rcompanion.

MARGOTTO, P. R. **Curva roc**: como fazer e interpretar no spss. Brasília: s.ed., 2010.

MCLEOD, A.; XU, C.; LAI, Y. **Bestglm**: best subset glm and regression utilities. [*S. l.*], 2020. R package version 0.37.3. Disponível em: https://CRAN.R-project.org/package=bestglm.

PAULA, G. A. **Modelos de regressão com apoio computacional**. 2013. Disponível em: https://www.ime.usp.br/~giapaula/textoregressao.htm/. Acesso em: 26 set. 2020.

PORTO, M. **A política nacional do idoso: um Brasil para todas as idades**. 2002. Disponível em: https://www.comciencia.br/dossies-1-72/reportagens/envelhecimento/texto/env02.htm. Acesso em: 10 out. 2020.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna, Austria, 2019. Disponível em: https://www.R-project.org/.

SÄRNDAL, C.-E.; SWENSSON, B.; WRETMAN, J. **Model assisted survey sampling**. 1. ed. New York: Springer Science & Business Media, 2003.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, California, v. 6, n. 2, p. 461–464, 1978.

SENADO FEDERAL. **Estatuto do idoso**. Brasília, 2003.

SILVA, J. A. S.; SOUZA, L. E. A.; GANASSOLI, C. Qualidade de vida na terceira idade: prevalência de fatores intervenientes. **Revista da Sociedade Brasileira de Clínica Médica**, Universidade Federal do Pará, v. 15, n. 3, p. 146–149, 2017.

SMITH, T. J.; MCKENNA, C. M. A comparison of logistic regression pseudo r2 indices. **Multiple Linear Regression Viewpoints**, v. 39, n. 2, p. 17–26, 2013.

THE WHOQOL GROUP. The world health organization quality of life assessment (whoqol): position paper from the world health organization. **Social Science Medicine**, v. 41, p. 1403–1409, 1995.

VECCHIA, R. D.; RUIZ, T.; BOCCHI, S. C. M.; CORRENTE, J. E. Qualidade de vida na terceira idade: um conceito subjetivo. **Revista Brasileira de Epidemiologia**, SciELO Public Health, v. 8, p. 246–252, 2005.

WEISBERG, S. **Applied linear regression**. New York: John Wiley, 1980.

APÊNDICE A – CÓDIGOS NO R

```
#### CÓDIGO ANÁLISE DOS DADOS - MONOGRAFIA
library(readxl)
library(xtable)
library(MASS)
library(bestglm)
library(pROC)
library(ResourceSelection)
library(rcompanion)

#### Essa é a versão do código onde não serão utlizados
#### observações com respostas faltantes. Atentar para a
#### identação ao copiar o código para o R.

#### Leitura da Base de Dados
dados_qv <- read_excel("dados_qv.xlsx")
View(dados_qv)

#### Removendo os NA's (observações que contenham alguma resposta vazia)
dados2 = na.omit(dados_qv)

#### As questões 1 e 2 são variáveis resposta.
#### Trabalharemos com a questão 1 de variável resposta, retiramos a
#### questão 2 e as demais serão as variáveis explicativas.
#### Para identificar melhor os pontos de análise de diagnóstico,
#### manter a coluna id.

ID = dados2[,1]
dados2 = dados2[,-c(1,3)]

#### Matriz de tabulação das variáveis explicativas em estudo
ctg = matrix(0,ncol(dados2)-1,2)
```

```
for(i in 1:(ncol(dados2)-1)){
  ctg[i,] = table(dados2[,i+1])
  i=i+1
}
colnames(ctg) = c("Nivel 0","Nível 1")
rownames(ctg) = c("Q3","Q4","Q5","Q6","Q7","Q8","Q9",
                  "Q10","Q11","Q12","Q13","Q14","Q15","Q16","Q17","Q18",
                  "Q19","Q20","Q21","Q22","Q23","Q24","Q25","Q26")
ctg

#### Como todas as variáveis possuem pelo menos 5 respostas em cada nível,
#### podemos aplicar unicamente o Teste QuiQuadrado

pvalor = 0
tqq = as.matrix(dados2)
for(i in 2:ncol(tqq)){
  pvalor[i]=(chisq.test(as.factor(tqq[,i]),as.factor(tqq[,1]))$p.value)
  i = i+1
}
pvalor = round(pvalor,4)

#### Eliminando as variáveis que foram excluídas no teste QQ:
#### Exclui as variáveis em que p > 0.05 (não-significativas)
varmodelo = dados2[,which(pvalor <= 0.05)]

#### Tabelas de coeficientes modelos univariados
# Coeficientes (a exponencial dos valores é que são os coeficientes do modelo)
TCF = matrix(0,nrow = ncol(varmodelo)-1,ncol = 3,byrow = T)
rownames(TCF) = colnames(varmodelo)[-1]
varmodelo = as.data.frame(varmodelo)
for(i in 2:ncol(varmodelo)){
```

```
  TCF[i-1,] = summary(glm(
  as.factor(varmodelo[,1]) ~ as.factor(varmodelo[,i]),
  family = binomial(), data = data.frame(varmodelo)))$coef[2,-3]
}

colnames(TCF) = c('B1.est','EP','p-valor')
exp.B1.est = exp(TCF[,1])
muv = cbind(TCF[,1],exp.B1.est,TCF[,-1])
colnames(muv) = c('B1.est','exp(B1.est)','EP','p-valor')
round(muv,4)

# variáveis com o p <= 0.05 serão aceitas.
TCF[,3] <= 0.05 # Todas as variáveis são aceitas no modelo

#########################################################################
#######        Seleção do Modelo Logístico Múltiplo           ########
#########################################################################

#### Tabelas de coeficientes modelos múltiplos
attach(varmodelo)
reglog = glm(w1 ~ ., family = binomial(), data = varmodelo)
summary(reglog)

exp(reglog$coefficients)

seleção = subset(varmodelo, select = c(w6,w8,w9,w11,w13,w15,w17,
w18,w19,w20,w1))

#####
MelhoresMLG = bestglm(Xy = seleção, family = binomial,
                      IC = "AIC", method = "exhaustive",
                      TopModels = 10, RequireFullEnumerationQ = T)
```

```
MelhoresMLG
summary.bestglm(MelhoresMLG)

##### Presença de variável não significativa
summary.glm(MelhoresMLG$BestModel)

##### Melhores modelos propostos
MelhoresMLG$BestModels

############################################################
####          Verificação dos modelos propostos    #########
####          Testando os modelos selecionados     #########
####               pelo comando bestglm            #########
############################################################

####################### Modelo 1 ####################################
fit.a = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = varmodelo)
summary.glm(fit.a)
anova(reglog, fit.a, test = "Chisq")

####################### Modelo 2 ####################################
fit.b = glm(w1 ~ w9+w13+w19+w20, family = binomial(), data = varmodelo)
summary.glm(fit.b)
anova(reglog, fit.b, test = "Chisq")

####################### Modelo 3 ####################################
fit.c = glm(w1 ~ w9+w13+w15+w19, family = binomial(), data = varmodelo)
summary.glm(fit.c)
anova(reglog, fit.c, test = "Chisq")

####################### Modelo 4 ####################################
fit.d = glm(w1 ~ w9+w13+w18, family = binomial(), data = varmodelo)
```

```
summary.glm(fit.d)
anova(reglog, fit.d, test = "Chisq")

###################### Modelo 5 ###################################
fit.e = glm(w1 ~ w9+w13+w15, family = binomial(), data = varmodelo)
summary.glm(fit.e)
anova(reglog, fit.e, test = "Chisq")

###################### Modelo 6 ###################################
fit.f = glm(w1 ~ w9+w13+w18+w19+w20, family = binomial(),
data = varmodelo)
summary.glm(fit.f)
anova(reglog, fit.f, test = "Chisq")

###################### Modelo 7 ###################################
fit.g = glm(w1 ~ w9+w18+w20, family = binomial(), data = varmodelo)
summary.glm(fit.g)
anova(reglog, fit.g, test = "Chisq")

###################### Modelo 8 ###################################
fit.h = glm(w1 ~ w9+w13+w19, family = binomial(), data = varmodelo)
summary.glm(fit.h)
anova(reglog, fit.h, test = "Chisq")

###################### Modelo 9 ###################################
fit.i = glm(w1 ~ w9+w13+w18+w19, family = binomial(),
data = varmodelo)
summary.glm(fit.i)
anova(reglog, fit.i, test = "Chisq")

###################### Modelo 10 ##################################
fit.j = glm(w1 ~ w9+w13+w15+w19+w20, family = binomial(),
```

```
data = varmodelo)
summary.glm(fit.j)
anova(reglog, fit.j, test = "Chisq")

###### Estatísticas D2 e D2.aj
D2 = c(Dsquared(fit.a),Dsquared(fit.b),Dsquared(fit.c),
      Dsquared(fit.d),Dsquared(fit.e),Dsquared(fit.f),
      Dsquared(fit.g),Dsquared(fit.h),Dsquared(fit.i),Dsquared(fit.j))
D2

D2.aj = c(Dsquared(fit.a,adjust = T),Dsquared(fit.b,adjust = T),
  Dsquared(fit.c,adjust = T),Dsquared(fit.d,adjust = T),
  Dsquared(fit.e,adjust = T),Dsquared(fit.f,adjust = T),
          Dsquared(fit.g,adjust = T),Dsquared(fit.h,adjust = T),
  Dsquared(fit.i,adjust = T),Dsquared(fit.j,adjust = T))
D2.aj

###### Comparação entre ajustes
Tav = accuracy(list(fit.a,fit.b,fit.c,fit.d,fit.e,
                    fit.f,fit.g,fit.h,fit.i,fit.j), plotit=FALSE)
Tav

#########################################################################
#####           Adequabilidade do Modelo Proposto           #########
#####             Estatística de Hosher-Lemeshow            #########
#########################################################################

hl.a = hoslem.test(fit.a$y, fitted(fit.a));hl.a
Chat.a = hl.a$statistic; Chat.a #Estatistica da Adequabilidade

pi.a = round(fit.a$fitted.values,2);pi.a # Probabilidade estimada - fitted.value
desvio.a = fit.a$deviance; desvio.a #Desvio = 'deviance' do modelo
```

```
gl.a = fit.a$df.residual; gl.a

# Odds Ratios dos parâmetros do modelo
logitor(w1 ~ w9+w13+w18+w20, data = varmodelo)

# Intervalos de Confiança para os parâmteros
exp(cbind(OR=coef(fit.a), confint(fit.a)))

##### Tabela de valores observados x esperados
oe.a = cbind(hl.a$observed, hl.a$expected); oe.a

#########################################################################
########           Diagnóstico do Modelo Proposto           #########
#########################################################################

fit.model=fit.a
source("envel_bino.R")
source("diag_bino.R")

### Localizando os pontos a serem estudados mais detidamente
# Tabela completa
trs = cbind(ID,h,di,td,pi.a)

## Pontos de alavanca
p = 4
n = nrow(varmodelo)
hii = 2*(p+1)/n # hii = 0.1369863
h > hii
h[h > hii]
trs[h > hii,] # Pontos de alavanca: obs. 26 e 45 (ID's 32 e 53)

## Distância de Cook, estabelecendo uma linha de corte arbitrária de 0,40
```

```
trs[di > 0.40,]
## Observações a serem destacadas: 9,19,41
## ID's 10, 21, 48

## Resíduos Componentes do Desvio
trs[abs(td) > 2,]
## RCD: a observação 19 e 47 (ID 21 e 55) tem valores absolutos maiores
## que os limites, merecendo uma maior atenção
## RCD x VA: aparentemente não há comportamento anormal.

## Observações a serem atentadas:
## 9,19,26,41,45,47 (ID's 10, 21, 32, 48, 53, 55)
trs[c(9,19,26,41,45,47),]

#########################################################################
############              Avaliação dos Modelos           ##############
############          sem as Observações destacadas       ##############
#########################################################################

# Sem as observações 10 individualmente
vm09 = varmodelo[-09,]
fit09 = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = vm09)
summary.glm(fit09)

##### Estatística de Hosher-Lemeshow
hl09 = hoslem.test(fit09$y, fitted(fit09)); hl09
Chat09 = hl09$statistic #Estatistica da Adequabilidade

pi09 = round(fit09$fitted.values,2) # Probabilidade estimada - fitted.value
desvio09 = fit09$deviance #Desvio = 'deviance'do modelo
gl09 = fit$df.residual
```

```
cat("Estatística C:",Chat09," com deviance de ",desvio09," e ",gl09,"
graus de liberdade.\n")
cat("-----------------------------------\n")

# Sem as observações 21 individualmente
vm19 = varmodelo[-19,]
fit19 = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = vm19)
summary.glm(fit19)

##### Estatística de Hosher-Lemeshow
hl19 = hoslem.test(fit19$y, fitted(fit19)); hl19
Chat19 = hl19$statistic #Estatistica da Adequabilidade

pi19 = round(fit19$fitted.values,2) # Probabilidade estimada - fitted.value
desvio19 = fit19$deviance #Desvio = 'deviance'do modelo
gl19 = fit$df.residual

cat("Estatística C:",Chat19," com deviance de ",desvio19," e ",gl19,"
graus de liberdade.\n")
cat("-----------------------------------\n")

# Sem as observações 32 individualmente
vm26 = varmodelo[-26,]
fit26 = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = vm26)
summary.glm(fit26)

##### Estatística de Hosher-Lemeshow
hl26 = hoslem.test(fit26$y, fitted(fit26)); hl26
Chat26 = hl26$statistic #Estatistica da Adequabilidade

pi26 = round(fit26$fitted.values,2) # Probabilidade estimada - fitted.value
desvio26 = fit26$deviance #Desvio = 'deviance'do modelo
```

```
gl26 = fit$df.residual

cat("Estatística C:",Chat26," com deviance de ",desvio26," e ",gl26,"
graus de liberdade.\n")
cat("-----------------------------------\n")

# Sem as observações 48 individualmente
vm41 = varmodelo[-41,]
fit41 = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = vm41)
summary.glm(fit41)

##### Estatística de Hosher-Lemeshow
hl41 = hoslem.test(fit41$y, fitted(fit41)); hl41
Chat41 = hl41$statistic #Estatistica da Adequabilidade

pi41 = round(fit41$fitted.values,2) # Probabilidade estimada - fitted.value
desvio41 = fit41$deviance #Desvio = 'deviance'do modelo
gl41 = fit$df.residual

cat("Estatística C:",Chat41," com deviance de ",desvio41," e ",gl41,"
graus de liberdade.\n")
cat("-----------------------------------\n")

# Sem as observações 53 individualmente
vm45 = varmodelo[-45,]
fit45 = glm(w1 ~ w9+w13+w18, family = binomial(), data = vm45)
summary.glm(fit45)

##### Estatística de Hosher-Lemeshow
hl45 = hoslem.test(fit45$y, fitted(fit45)); hl45
Chat45 = hl45$statistic #Estatistica da Adequabilidade
```

```
pi45 = round(fit45$fitted.values,2) # Probabilidade estimada - fitted.value
desvio45 = fit45$deviance #Desvio = 'deviance'do modelo
gl45 = fit$df.residual

cat("Estatística C:",Chat45," com deviance de ",desvio45," e ",gl45,"
graus de liberdade.\n")
cat("-----------------------------------\n")

# Sem as observações 55 individualmente
vm47 = varmodelo[-47,]
fit47 = glm(w1 ~ w9+w13+w18+w20, family = binomial(), data = vm47)
summary.glm(fit47)

##### Estatística de Hosher-Lemeshow
hl47 = hoslem.test(fit47$y, fitted(fit47)); hl47
Chat47 = hl47$statistic #Estatistica da Adequabilidade

pi47 = round(fit47$fitted.values,2) # Probabilidade estimada - fitted.value
desvio47 = fit47$deviance #Desvio = 'deviance'do modelo
gl47 = fit$df.residual

cat("Estatística C:",Chat47," com deviance de ",desvio47," e ",gl47,"
graus de liberdade.\n")
cat("-----------------------------------\n")

#########################################################################
#######          Tabela de Classificação e Curva ROC             ######
#########################################################################

ModeloFinal = fit.a

###.TABELA DE CLASSIFICAÇÃO E CURVA DE ROC - Adaptado de:
```

```
##DIEGO SILVA CAMPOS CARELLI.
#Disponível em:
#_____________________________________________________________________#
### Carelli, D. S. C (2017). "Predição de aprovação em um curso em tecnologia
### da informação do Instituto Metrópole Digital da UFRN: uma aplicação
### da análise de regressão logística" .
#_____________________________________________________________________#
#---------- INÍCIO DO CÓDIGO ADAPTADO DE DIEGO SILVA CAMPOS CARELLI ------#
S = predict(ModeloFinal, type = "response");round(S,2)
g = roc(w1 ~ ModeloFinal$fitted.values, data = dados2);g

coords(g, "all", ret=c("threshold", "specificity", "sensitivity", "accuracy"),
transpose = F)

plot(g, main = "", xlab = "1 - Especifidade", ylab = "Sensibilidade",
     col = "red", legacy.axes = TRUE, lwd = 2, identify.lwd = 2,
     identify.col = "red", identify.lty = 2)

# Adaptação para identificar as coordenadas do ponto de corte na curva ROC
points(0.7575758,0.825,col = "blue", pch = 16,lwd = 5)
abline(h = g$specificities, v = g$sensitivities, col = "pink",lty = 3)
## 0.5651211   0.7575758       0.825 0.7945205

### Ponto de corte 1: 0.50
coords(g, 0.50, transpose = T)
threshold = 0.50
curva_roc(threshold,print=TRUE)

# Taxa de Classificação do Modelo
100*(56/73)  # 76,71%

# Taxa de Especificidade
```

```
100*(23/33)  # 69,69%

# Taxa de Sensibilidade
100*(33/40)  # 82,5%

### Ponto de corte 2: 0.5651211
coords(g, 0.5651211, transpose = T)
threshold = 0.5651211
curva_roc(threshold,print=TRUE)

# Taxa de Classificação do Modelo
100*(58/73)  # 79,45%

# Taxa de Especificidade
100*(25/33)  # 75,75%

# Taxa de Sensibilidade
100*(33/40)  # 82,5%
#------------ FIM DO CÓDIGO ADAPTADO DE DIEGO SILVA CAMPOS CARELLI-----------#

##########################################################################
######          Teste de Poder Preditivo do Modelo             ##########
##########################################################################

##### TESTES PREDITIVOS
##### Simulação com três idosos, A,B e C, que responderam ao questionário
#### de três formas distintas
w9p = c(1,1,0) #1,1,0
w13p = c(1,0,1) #1,0,1
w18p = c(1,0,1) #1,0,1
w20p = c(1,1,0) #1,1,0
```

```
modelo = 1/(1+exp(-1*(-2.5448+(1.6885*w9p)+(0.9875*w13p)+(1.1115*w18p)+(1.1251*w20p
modelo

resposta = cbind(w9p,w13p,w18p,w20p,modelo)
resposta

#########################################################################
#########          Curva logística do modelo final            ##########
#########################################################################

Sx = function(x) 1/(1+exp(-x)) # Curva teórica da regressão logística
plotGLM(fit.a,lwd = 3, main = "Curva Logística para o modelo adotado", plot.values
ylab = "Probabilidade Predita")
curve(Sx, xlim = c(-3,3),ylim = c(0,1), col = "red", main = "Curva de Regressão Log
#### FIM DO CÓDIGO #####
```

ANEXO A – WHOQOL ABREVIADO

Instrumento de Avaliação de Qualidade de Vida

The World Health Organization Quality of Life - WHOQOL-bref

Instruções

Este questionário é sobre como você se sente a respeito de sua qualidade de vida, saúde e outras áreas de sua vida. Por favor responda a todas as questões. Se você não tem certeza sobre que resposta dar em uma questão, por favor, escolha entre as alternativas a que lhe parece mais apropriada.
Esta, muitas vezes, poderá ser sua primeira escolha. Por favor, tenha em mente seus valores, aspirações, prazeres e preocupações. Nós estamos perguntando o que você acha de sua vida, tomando como como referência as duas últimas semanas. Por exemplo, pensando nas últimas duas semanas, uma questão poderia ser:

	nada	Muito pouco	médio	muito	completamente
Você recebe dos outros o apoio de que necessita?	1	2	3	4	5

Você deve circular o número que melhor corresponde ao quanto você recebe dos outros o apoio de que necessita nestas últimas duas semanas. Portanto, você deve circular o número 4 se você recebeu "muito" apoio como abaixo.

	nada	Muito pouco	médio	muito	completamente
Você recebe dos outros o apoio de que necessita?	1	2	3	④	5

Você deve circular o número 1 se você não recebeu "nada" de apoio. Por favor, leia cada questão, veja o que você acha e circule no número e lhe parece a melhor resposta.

		muito ruim	Ruim	nem ruim nem boa	boa	muito boa
1	Como você avaliaria sua qualidade de vida?	1	2	3	4	5
		muito insatisfeito	Insatisfeito	nem satisfeito nem insatisfeito	satisfeito	muito satisfeito
2	Quão satisfeito(a) você está com a sua saúde?	1	2	3	4	5

As questões seguintes são sobre **o quanto** você tem sentido algumas coisas nas últimas duas semanas.

		nada	muito pouco	mais ou menos	bastante	extremamente
3	Em que medida você acha que sua dor (física) impede você de fazer o que você precisa?	1	2	3	4	5

4	O quanto você precisa de algum tratamento médico para levar sua vida diária?	1	2	3	4	5
5	O quanto você aproveita a vida?	1	2	3	4	5
6	Em que medida você acha que a sua vida tem sentido?	1	2	3	4	5
7	O quanto você consegue se concentrar?	1	2	3	4	5
8	Quão seguro(a) você se sente em sua vida diária?	1	2	3	4	5
9	Quão saudável é o seu ambiente físico (clima, barulho, poluição, atrativos)?	1	2	3	4	5

As questões seguintes perguntam sobre **quão completamente** você tem sentido ou é capaz de fazer certas coisas nestas últimas duas semanas.

		nada	muito pouco	médio	muito	completamente
10	Você tem energia suficiente para seu dia-a- dia?	1	2	3	4	5
11	Você é capaz de aceitar sua aparência física?	1	2	3	4	5
12	Você tem dinheiro suficiente para satisfazer suas necessidades?	1	2	3	4	5
13	Quão disponíveis para você estão as informações que precisa no seu dia-a-dia?	1	2	3	4	5
14	Em que medida você tem oportunidades de atividade de lazer?	1	2	3	4	5

As questões seguintes perguntam sobre **quão bem ou satisfeito** você se sentiu a respeito de vários aspectos de sua vida nas últimas duas semanas.

		muito ruim	ruim	nem ruim nem bom	bom	muito bom
15	Quão bem você é capaz de se locomover?	1	2	3	4	5
		muito insatisfeito	Insatisfeito	nem satisfeito nem insatisfeito	satisfeito	Muito satisfeito
16	Quão satisfeito(a) você está com o seu sono?	1	2	3	4	5
17	Quão satisfeito(a) você está com sua capacidade de desempenhar as atividades do seu dia-a-dia?	1	2	3	4	5
18	Quão satisfeito(a) você está com sua capacidade para o trabalho?	1	2	3	4	5

19	Quão satisfeito(a) você está consigo mesmo?	1	2	3	4	5
20	Quão satisfeito(a) você está com suas relações pessoais (amigos, parentes, conhecidos, colegas)?	1	2	3	4	5
21	Quão satisfeito(a) você está com sua vida sexual?	1	2	3	4	5
22	Quão satisfeito(a) você está com o apoio que você recebe de seus amigos?	1	2	3	4	5
23	Quão satisfeito(a) você está com as condições do local onde mora?	1	2	3	4	5
24	Quão satisfeito(a) você está com o seu acesso aos serviços de saúde?	1	2	3	4	5
25	Quão satisfeito(a) você está com o seu meio de transporte?	1	2	3	4	5

As questões seguintes referem-se a **com que freqüência** você sentiu ou experimentou certas coisas nas últimas duas semanas.

		nunca	Algumas vezes	freqüentemente	muito freqüentemente	sempre
26	Com que freqüência você tem sentimentos negativos tais como mau humor, desespero, ansiedade, depressão?	1	2	3	4	5

Alguém lhe ajudou a preencher este questionário?
.....

Quanto tempo você levou para preencher este questionário?
.....

Você tem algum comentário sobre o questionário?

OBRIGADO PELA SUA COLABORAÇÃO