ISSN: 2317-0840

Modelos GAMLSS: Aplicação em casos de Síndrome Respiratória Aguda Grave com ênfase em Influenza e outras causas

Miriam Lecília Farias Ribeiro[1†], Letícia Souza de Oliveira[1]; Josimar Mendes de Vasconcelos[1]

[1]*Programa de Pós-Graduação em Biometria e Estatística Aplicada (PPGBEA), Universidade Federal Rural de Pernambuco (UFRPE); SEDE; DEINFO/PPGBEA; Recife, Pernambuco*

Resumo: *Um dos maiores causadores da Síndrome Respiratória Aguda Grave (SRAG) é a Influenza, uma doença respiratória provocada por diversas cepas do vírus. A disseminação global da Influenza tem representado um desafio significativo desde 2009. Neste contexto, a presente pesquisa propõe fazer uma análise da SRAG com ênfase na Influenza e outras causas (vírus respiratórios e outros agentes etiológicos), a fim de investigar e compreender a relação entre variáveis preditoras (comorbidades) e a variável de resposta específica (evolução) nos casos registrados no Brasil. A partir da análise dos resultados encontrados, referente ao período de 2020 a agosto de 2022, notou-se que os dados apresentaram um ajuste insatisfatório ao modelo de Regressão Linear Múltipla e os Modelos Lineares Generalizados, o que motivou a utilização dos Modelos Aditivos Generalizados para Locação, Escala e Forma. Para analisar esses dados, utilizamos distribuições clássicas de probabilidade no contexto dessa modelagem: binomial, binomial negativo, geométrico e Poisson. Dessas distribuições trabalhadas, a distribuição binomial mostrou-se eficaz, permitindo mapear a adequação desta modelagem. Foram realizadas previsões nos conjuntos de treinamento e teste, a fim de verificar possíveis superajustamentos no modelo obtido. Essa abordagem proporcionou uma compreensão mais profunda do conjunto de dados e das relações entre as variáveis estudadas. Pacientes asmáticos mostraram uma forte e significativa associação com a taxa de mortalidade, enquanto aqueles com condições hepáticas crônicas demonstraram menor risco de óbito. Esses resultados fornecem estratégias importantes para intervenção e cuidado das populações afetadas por essas condições de saúde.*

Palavras-chave: *Modelagem Estatística; Regressão GAMLSS; SRAG; Influenza.*

GAMLSS Models: Application in cases of Severe Acute Respiratory Syndrome with emphasis on Influenza and other causes

Abstract: *One of the major causes of Severe Acute Respiratory Syndrome (SARS) is Influenza, a respiratory disease caused by several strains of the virus. The global spread of Influenza has represented a significant challenge since 2009. In this context, this research proposes to analyze SARS with an emphasis on Influenza and other causes (viruses and other etiological agents), in order to investigate and understand the relationship between predictor variables (comorbidities) and the specific response variable (evolution) in the cases registered in Brazil. From the analysis of the results found, referring to the period from 2020 to August 2022, it was noted that the data indicated an unsatisfactory fit to the Multiple Linear Regression model and the Generalized Linear Models, which motivated the use of Generalized Additive Models for Location, Scale and Shape. To analyze these data, we used classical probability distributions in the context of this modeling: binomial, negative binomial, geometric and Poisson. Of these distributions worked, the binomial distribution proved to be effective, allowing us to map the adequacy of this modeling. Our training and test sets were performed in order to verify possible overfitting in the obtained model. This approach provided a deeper understanding of the data set and the relationships between the studied variations. Asthmatic patients demonstrated a strong and significant association with the mortality rate, while those with chronic liver conditions had a lower risk of death. These results provide important strategies for intervention and care of populations affected by these health conditions.*

Keywords: *Statistical Modeling; GAMLSS Regression; SRAG; Influenza.*

†Autor correspondente: leciliariber@gmail.com

Manuscrito recebido em: 01/08/2024
Manuscrito revisado em:15/10/2024
Manuscrito aceito em: 31/10/2024

Sigmae, Alfenas, v.13, n.4, p.265-281, 2024.

68ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras)

Introdução

As condições de saúde pública, como a Síndrome Respiratória Aguda Grave (SRAG), tem desafiado continuamente os sistemas de saúde global. A SRAG, uma condição que abrange casos de Síndrome Gripal (SG) com subsequente comprometimento da função respiratória, tem gerado considerável preocupação devido a sua capacidade de evoluir para quadros graves, frequentemente resultando em hospitalizações e, em casos mais extremos, levando ao óbito (Brasil, 2020).

A pandemia recente trouxe à tona a urgência de compreender profundamente a SRAG, principalmente pela sua relação com eventos de saúde pública de escala global. O entendimento dos fatores de risco, padrões de evolução clínica e estratégias de manejo eficazes torna-se imperativo para enfrentar não apenas as situações de crise imediata, mas também para promover estruturas de saúde resilientes e sistemas de resposta preparados para desafios similares no futuro (Freitag *et al.*, 2021; Ribeiro *et al.*, 2022).

Essa situação pandêmica atual desperta reflexões sobre desafios históricos enfrentados com a Influenza, uma doença respiratória provocada por diversas cepas do vírus influenza. Havendo assim a caracterização de subtipos de vírus, como os influenza A, B, C e posteriormente suas variantes. A transmissão ocorre principalmente através de gotículas respiratórias expelidas por pessoas infectadas ao tossir, espirrar ou falar. Essas gotículas podem ser inaladas por pessoas próximas ou depositadas em superfícies, onde o vírus pode permanecer ativo por um tempo, podendo ser transferido para as mãos e, em seguida, para o rosto de outras pessoas. O contato com essas superfícies contaminadas e posterior contato com as mucosas do nariz, boca ou olhos pode resultar na infecção pelo vírus influenza (Brasil, 2009).

No âmbito da modelagem estatística, modelagens como a Regressão Linear Múltipla (RLM), Modelos Lineares Generalizados (MLGs) e os Modelos Aditivos Generalizados para Locação, Escala e Forma (em inglês Generalized Additive Models for Location, Scale and Shape, GAMLSS) vem se destacando devido a sua flexibilidade e capacidade de lidar com dados complexos.

Na aplicação de RLM fazemos algumas suposições, ou seja: o modelo de regressão é linear nos parâmetros, não deve haver correlação entre as variáveis independentes. Os resíduos são normalmente distribuídos, os resíduos devem ter variância constante (homocedasticidade) e os resíduos não devem apresentar autocorrelação. Todavia, nem sempre as suposições são satisfeitas, por exemplo os resíduos não terem distribuição normal.

Neste caso, quando não satisfazer a suposição de normalidade, recorremos aos MLGs que podem ser empregadas por meio de modelos que pertencem à família exponencial, por exemplo as distribuições binomial, binomial negativa, geométrica, Poisson e dentre outras clássicas. Entretanto, pode ocorrer dos modelos de probabilidade não pertencer à família exponencial ou a variável resposta e as covariáveis não serem lineares, daí podemos recorrer ao Modelos Aditivos Generalizados (GAMs) que é uma extensão dos MLGs que permite incorporar relações não lineares. Nos modelos GAMs, a relação de cada preditor com a média da variável resposta não é direta, mas é feita através de uma função densidade de probabilidade (ou de massa de probabilidade).

Por outro lado, os GAMLSS vem sendo introduzido por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou, Rigby e Stasinopoulos (2002) visa superar algumas limitações

encontradas nos MLGs e GAMs, conforme discutido por Nelder e Wedderburn (1972) e Hastie e Tibshirani (1990), respectivamente. Trata-se de uma técnica de modelagem estatística univariada que permite o ajuste de uma ampla família de distribuições contínuas e discretas para a variável resposta e possibilita a modelagem explícita, utilizando funções paramétricas e/ou não-paramétricas, de todos os parâmetros da distribuição da variável resposta em relação às variáveis explanatórias.

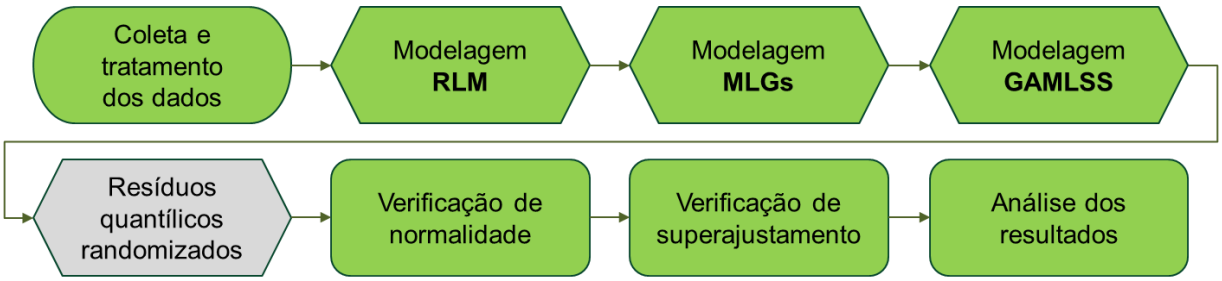
Finalmente, através dos testes de hipóteses verificamos a normalidade dos residuos randomizados. Consequentemente aplicamos o teste da Raiz do Erro Quadrático Médio (RMSE) para descartar possíveis superajustamentos.

Neste contexto, o presente artigo propõe fazer uma análise da SRAG com ênfase na Influenza e outras causas, aplicando os GAMLSS como uma abordagem analítica para investigar e compreender a relação entre variáveis preditoras (comorbidades) e a variável resposta específica (evolução), buscando o melhor modelo de regressão com o menor número de variáveis independentes para que tenhamos bons valores previstos da evolução.

Metodologia

A Figura 1 apresenta os passos metodológicos utilizados para a análise dos dados em questão.

Figure 1: Methodological diagram of data analysis.



Source: from the authors (2024).

Como mostra a Figura 1 o processo iniciou-se com a coleta e tratamento dos dados. A seguir, os dados foram submetidos a diferentes técnicas de modelagem: a RLM, MLGs e os GAMLSS. Após a modelagem analisou detalhadamente os resíduos quantílicos randomizados, que são essenciais para avaliar a adequação dos modelos. Em seguida, realizou-se a verificação da normalidade e a verificação de superajustamento, garantindo que os modelos não estejam ajustando-se excessivamente aos dados de treinamento, comprometendo sua capacidade de generalização. Por fim, a metodologia desta pesquisa culmina com a análise dos resultados, em que são extraídas conclusões a partir dos modelos ajustados.

Dados coletados para análise

Os dados foram coletados através do OpenDataSUS do Ministério da Saúde, no Brasil, referente ao período de 2020 a agosto de 2022 para trabalhar no contexto qualitativo longitudinal (Samperi *et al.*, 2013).

O Ministério da Saúde (MS), por meio da Secretaria de Vigilância em Saúde (SVS), desenvolve a vigilância da SRAG no Brasil, desde a pandemia de Influenza A(H1N1). A partir disso, a vigilância de SRAG foi implantada na rede de vigilância, que anteriormente

atuava exclusivamente com a vigilância sentinela de Síndrome Gripal (SG). Em 2020, a vigilância da COVID-19 foi incorporada na rede de vigilância da Influenza e outros vírus respiratórios (DataSUS, 2022).

A plataforma que coletamos os dados tem como finalidade disponibilizar o legado dos bancos de dados epidemiológicos de SRAG, desde o início da sua implantação (2009) até o ano corrente (2022). Atualmente, o sistema oficial para o registro dos casos e óbitos por SRAG é o Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe) (DataSUS, 2022).

Os dados coletados da vigilância de SRAG no Brasil estão sujeitos a alterações decorrentes da investigação, ou mesmo correções de erros de digitação, pelas equipes de vigilância epidemiológica que desenvolvem o serviço nas três esferas de gestão. Os bancos de dados de SRAG disponibilizadas no portal passam por tratamento que envolve a anonimização, em cumprimento a Lei 13.709/2018 (DataSUS, 2022).

De acordo com Brasil (2022), os dados relacionados à SRAG devem ser lidos como mostra o Quadro 1 a seguir.

Chart 1: Data analyzed from the Severe Acute Respiratory Syndrome database.

COLUNAS	NOME DO CAMPO	DESCRIÇÃO	CATEGORIA
CARDIOPATI	Fatores de risco/ Doença Cardiovascular Crônica	Paciente possui Doença Cardiovascular Crônica?	1-Sim 2-Não
HEMATOLOGI	Fatores de risco/ Doença Hematológica Crônica	Paciente possui Doença Hematológica Crônica?	
SIND_DOWN	Fatores de risco/ Síndrome de Down	Paciente possui Síndrome de Down?	
HEPATICA	Fatores de risco/ Doença Hepática Crônica	Paciente possui Doença Hepática Crônica?	
ASMA	Fatores de risco/ Asma	Paciente possui Asma?	
DIABETES	Fatores de risco/ Diabetes mellitus	Paciente possui Diabetes *mellitus*?	
NEUROLOGIC	Fatores de risco/ Doença Neurológica Crônica	Paciente possui Doença Neurológica?	
PNEUMOPATI	Fatores de risco/ Outra Pneumatopatia Crônica	Paciente possui outra pneumopatia crônica?	
IMUNODEPRE	Fatores de risco/ Imunodeficiência ou Imunodepressão	Paciente possui Imunodeficiência ou Imunodepressão (diminuição da função do sistema imunológico)?	
RENAL	Fatores de risco/ Doença Renal Crônica	Paciente possui Doença Renal Crônica?	
OBESIDADE	Fatores de risco/ Obesidade	Paciente possui obesidade?	
CLASSI_FIN	Classificação final do caso	Diagnóstico final do caso. Se tiver resultados divergentes entre as metodologias laboratoriais, priorizar o resultado do RTPCR	1-SRAG por influenza 2-SRAG por outro vírus respiratório 3-SRAG por outro agente etiológico, qual: 4-SRAG não especificado 5-SRAG por COVID-19

Souce: SIVEP-Gripe - Adapted by the authors (2023).

O banco de dados obtido no site do Ministério da Saúde, inicialmente, apresentou 946.404 dados para o Brasil, contemplando informações do período 30 de dezembro de 2019

a 05 de agosto de 2022. No entanto, foi realizado um tratamento nesse banco visando deixar apenas os casos sinalizados pelo número 1 (sim) e 2 (não) na coluna "Evolução do caso", e para isso foi feito a exclusão do número 3 (óbito por outras causas) e do número 9 (ignorado).

Os dados tratados passaram por outro tratamento visando deixar somente os casos sinalizados pelo número 1 (SRAG por influenza), 2 (SRAG por outro vírus respiratório) e 3 (SRAG por outro agente etiológico) na coluna "Classificação final do caso", e para isso fizemos a exclusão do número 4 (SRAG não especificado) e 5 (SRAG por COVID-19).

O próximo passo do tratamento foi renomear os casos sinalizados pelo número 2 (não) na coluna "Evolução do caso" por 0, e unificar e renomear os casos sinalizados pelo número 2 (SRAG por outro vírus respiratório) e 3 (SRAG por outro agente etiológico) na coluna "Classificação final do caso" por 0. Portanto, os dados analisados consistiram nos casos assinalados com o número 1 (sim) e 0 (não) para a coluna "Evolução do caso", assim como nos casos assinalados com 1 (SRAG por influenza) e 0 (SRAG por outras causas) para a coluna "Classificação final do caso". Após o tratamento dos dados, considerou apenas 339.751 elementos para o Brasil, contemplando informações do período de 30 de dezembro de 2019 a 05 de agosto de 2022.

O modelo GAMLSS

Quando não há um encaixe adequado dos dados em uma modelagem por RLM, MLGs ou GAMs, uma abordagem alternativa é a aplicação do método GAMLSS. Objetivando superar algumas das limitações associadas aos modelos descritos anteriormente, Rigby & Stasinopoulos (2005), propuseram uma nova classe de modelos estatísticos de regressão (semi) paramétricos, denominada de GAMLSS. O modelo GAMLSS é dado da seguinte forma:

$$y \perp D(y \mid \mu, \sigma, \nu, \tau)$$
$$\eta_1 = g_1(\mu) = x_1 s_1 + S_{11}(x_{11}) + \cdots + S_1 j_1(x_1 + j_1)$$
$$\eta_2 = g_2(\sigma) = x_2 s_2 + S_{21}(x_{21}) + \cdots + S_2 j_1(x_2 + j_2)$$
$$\eta_3 = g_3(\nu) = x_3 s_3 + S_{31}(x_{31}) + \cdots + S_3 3(x_3 + j_3)$$
$$\eta_4 = g_4(\tau) = x_4 s_4 + S_{41}(x_{41}) + \cdots + S_4 j_4(x_4 + j_4)$$

em que $D(\cdot)$ é em uma distribuição de probabilidade com 4 parâmetros. Sendo μ o parâmetro de Locação, σ o parâmetro de Escala, ν e τ são parâmetros de formas da distribuição. η_i representa o preditor linear de cada parâmetro indexados as variáveis.

A estrutura da regressão GAMLSS os p parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ de uma função densidade de probabilidade$f\ (y \mid line\theta)$, em que são modelados utilizando termos aditivos. Presume-se que as observações y_i, $i = 1, 2, \ldots, n$ são independentes e condicionais a θ^i, com função densidade de probabilidade $f\ (y_i \mid \theta^i)$, em que $\theta^{it} = \theta_{i1}, \theta_{i2}, \ldots, \theta_{ip}$, é um vetor de p parâmetros relacionado às variáveis explanatórias e efeitos aleatórios. Destaca-se que quando os valores assumidos pelas covariáveis são estocásticos ou as observações y_i dependem de seus valores ocorridos, então $f\ (y_i \mid \theta)$, é interpretada como sendo condicional a estes valores.

No modelo GAMLSS, os parâmetros μ, σ, ν e τ fornecem um ajuste flexível aos dados. Dependendo da natureza dos dados podem ser incluídos na modelagem. Em alguns casos, apenas os parâmetros de locação, μ, e o de escala, Σ, são suficientes para capturar as características principais dos dados, em particular se a distribuição for simétrica e a forma padrão. No caso dos parâmetros ν e τ, parâmetros de formas, permitem um ajuste flexível aos dados, podendo ou não ser incluídos conforme necessário. Em alguns casos, apenas o

parâmetro de locação (μ) e o de escala (σ) são suficientes para capturar as características principais dos dados, especialmente se a distribuição for simétrica e com uma forma padrão. Já ν e τ são parâmetros de forma que permitem ao modelo capturar aspectos adicionais, como assimetria e curtose, que ajustam a distribuição para dados mais complexos e não simétricos. Esses parâmetros podem ser ajustados de forma independente, de acordo com as características dos dados, permitindo que o modelo GAMLSS se adapte a uma grande variedade de distribuições e padrões.

Modelo Binomial

A distribuição de regressão binomial pertence à família exponencial e é utilizada para modelar dados de contagem ou eventos binários (sucesso/fracasso). Neste modelo, a variável resposta segue uma distribuição binomial, sendo particularmente útil em situações onde se deseja modelar a probabilidade de ocorrência de um evento.

Seja X uma variável aleatória discreta binomial, então a função de massa de probabilidade é dada por

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0,1,2,\dots,n, \tag{1}$$

em que $f(x)$, representa a probabilidade de obter exatamente x sucessos em n tentativas; p, é a probabilidade de sucesso em uma tentativa individual e $1-p$, é a probabilidade de fracasso (Magalhães, 2015).

Quando se usa a regressão binomial dentro da estrutura dos GAMLSS, a abordagem permite modelar a probabilidade do evento (como na regressão binomial padrão), mas também possibilita a modelagem de outras características da distribuição binomial como a dispersão, assimetria e curtose.

Os GAMLSS permitem escolher diferentes funções de ligação para cada um desses parâmetros, de forma independente. Isso oferece a capacidade de adaptar o modelo para melhor representar a relação entre os preditores e os parâmetros da distribuição.

No modelo de regressão binomial, a função de ligação é a função *logit*, que relaciona a média da distribuição Binomial à combinação linear das variáveis explicativas. Sendo dada por

$$logit(p) = log\left(\frac{p}{1-p}\right), \tag{2}$$

em que p representa a probabilidade de sucesso em um evento binomial. Esses são alguns dos principais elementos matemáticos e estatísticos associados ao modelo de regressão binomial, que é fundamental para analisar dados binários ou de contagem, em que a distribuição subjacente é binomial.

Testes

Avaliação da normalidade

Os testes Kolmogorov-Smironff e Lilliefors, são ferramentas estatísticas amplamente utilizadas na metodologia de pesquisa para avaliar a normalidade de uma distribuição de dados ou comparar duas distribuições amostrais. Quando aplicado para validar a normalidade, os testes ajudam a determinar se os dados seguem uma distribuição Normal. Quando utilizado para comparar duas distribuições, os testes examinam se as duas amostras são provenientes da mesma distribuição (Gonzalez, *et al.*, 1977).

Superajustamento

O RMSE avalia possíveis casos de superajustamento (*overfitting*). O RMSE é uma métrica estatística que mede a média quadrática das diferenças entre os valores previstos por um modelo e os valores observados. Sua aplicação visa verificar a capacidade do modelo em generalizar para dados não utilizados no treinamento. A sua formulação é dada por

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (3)$$

em que $\hat{y}_i$, é calculado pela média quadrática das diferenças entre os valores previstos; y_i, são os valores reais para cada observação i onde n é o número total de observações (Li e Yang, 2008).

A fim de verificar um possível superajustamento para o melhor modelo de distribuição, foi-se aplicado o RMSE, onde as previsões foram geradas para o conjunto de treinamento (‘dados_treino’) e o conjunto de teste (‘dados_teste’) usando a função *predict*. Em seguida, o RMSE foi calculado para ambos os conjuntos, representando a diferença média entre as previsões do modelo e os valores reais.

Resultados e Discussão

Nesta seção discorremos sobre a análise numérica dos dados através do *Software R*. Iniciamos com uma análise descritiva dos dados para depois fazermos as modelagens estatísticas seguindo o desenho de estudo do artigo.

A análise descritiva da variável "idade" revela uma ampla variação, de 0 a 105 anos. O primeiro quartil é 34 anos, a mediana é 58 anos, e a média é 52,52 anos, indicando que a maioria da população acometida está abaixo dos 58 anos. O terceiro quartil, com 74 anos, sugere que 75% das pessoas têm até 74 anos. A diferença entre a média e a mediana sugere uma leve assimetria à direita, possivelmente devido a alguns valores extremos.

Regressão Linear Múltipla

O passo seguinte após a análise inicial é investigar como os dados se comportam no Critério de Informação de Akaike (AIC), Critério de Informação Bayesiano (BIC), coeficiente de determinação (R^2), R^2 ajustado, R^2 não centrado e no teste RESET, respectivamente, na Tabela 1.

Table 1: AIC, BIC, R2, Adjusted R2, Uncentered R2 and RESET test - SRAG.

AIC	BIC	R^2	R^2 Ajustado	R^2 Não Centrado	RESET test
153119,53859	153247,54562	0,02400	0,02393	0,02400	78,09, df1 = 1, df2 = 139609, p-value < 2,2e-16

Souce: from the authors (2024).

Observa-se na Tabela 1 que o AIC (critério que penaliza modelos mais complexos) tem o valor aproximadamente de 153119,54, isto indica uma parcimônia relativa do modelo, visto que valores menores são favoráveis. O BIC (medida mais rigorosa para modelos complexos) registra aproximadamente 153247,55. O R^2 revela que o modelo explica apenas 2,4% da variabilidade na variável resposta, corroborado pelo R^2 ajustado ligeiramente inferior, em 0,0239, considerando a penalização para modelos mais complexos. O R^2 não centrado

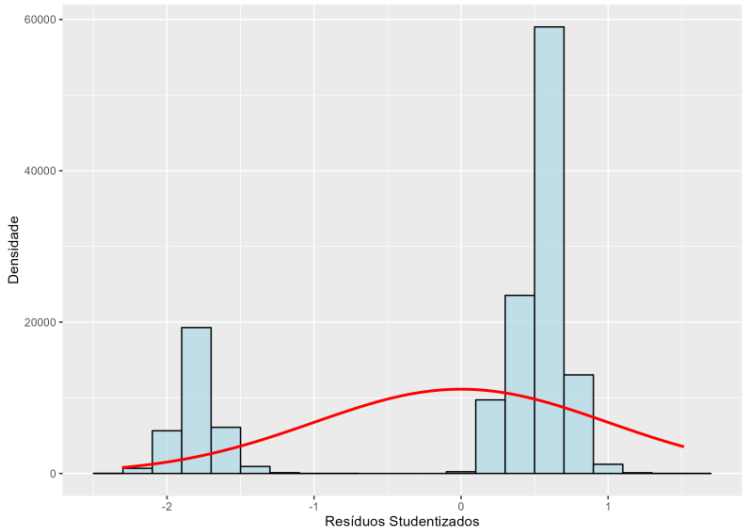
oferece uma abordagem alternativa para avaliar a explicação do modelo. O teste RESET fornece evidências estatísticas de que a equação de regressão pode ser aprimorada (ou seja, há erro de especificação no modelo), sugerindo a necessidade de revisão ou inclusão de variáveis para melhor descrever os dados.

Logo após, verificamos que há significância estatística das comorbidades investigadas em relação à variável de interesse. As variáveis CARDIOPATI, HEMATOLOGI, HEPATICA, ASMA, DIABETES, NEUROLOGIC, PNEUMOPATI, IMUNODEPRE, RENAL e OBESIDADE demonstram fortes evidências estatísticas ($p < 0,001$). Os altos valores da estatística F e baixos *p-valores* indicam que essas comorbidades têm um efeito estatisticamente significativo na variável de interesse. Além disso, destaca-se a variável SIND_DOWN que embora apresente significância estatística a um nível de 0,05, possui um impacto estatisticamente menor comparado às outras variáveis.

Todavia, na análise de diagnóstico percebemos que não satisfaz algumas suposições do modelo. Por exemplo, os valores quantílicos estão muito afastados da reta ajustada do *qqplot* e os resíduos do modelo não aderem estritamente a uma distribuição Normal. Esse desvio da normalidade nos resíduos sugere possíveis violações da pressuposição de que os resíduos devem seguir uma distribuição Normal para que o modelo de RLM seja adequado. Essa falta de aderência à normalidade pode indicar a presença de padrões não capturados pelo modelo ou influências de variáveis não consideradas, levando a resíduos que não estão distribuídos conforme o esperado, o que pode afetar a precisão das inferências feitas a partir do modelo estatístico.

Enfim, na análise dos resíduos studentizados percebemos que não aderem a uma distribuição Normal (veja a Figura 2), indicando a presença de padrões ou valores atípicos nos dados que não foram adequadamente explicados pelo modelo. Esses resíduos studentizados representam as discrepâncias entre os valores observados e os valores previstos pelo modelo estatístico. Quando eles não seguem a distribuição Normal, isso sugere que o modelo pode não estar capturando completamente certos padrões ou influências presentes nos dados, como *outliers* ou estruturas mais complexas, indicando a necessidade de revisão ou aprimoramento do modelo estatístico para acomodar essas características peculiares dos dados.

Figure 2: Distribution of studentized residues - SRAG.



Source: from the authors (2023).

GAMLSS

A escolha do melhor modelo se deu através do AIC, "Plot" e dos testes de normalidade. O primeiro passo da análise por GAMLSS foi investigar como os AICs da Tabela 2 se encaixam em cada modelo discreto.

Table 2: AIC – SRAG.

Modelos discretos		AIC
M1	Geométrico	336916,1
M2	Binomial Negativo	270120,5
M3	Poison	270118,5
M4	Binomial	148545,2

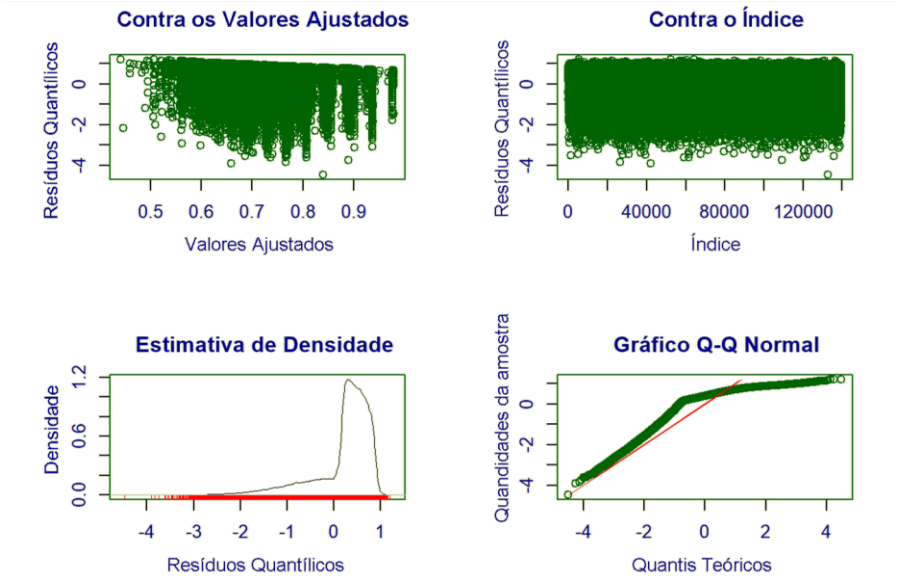
Source: from the authors (2023).

Analisando os dados de AIC dos modelos discretos da Tabela 2, nota-se que o modelo Binomial possui o menor valor de AIC (148545,2), seguido pelo modelo Poisson (270118,5), Binomial Negativo (270120,5) e, por fim, o modelo Geométrico (336916,1). Esses valores indicam que, dentre os modelos testados, o modelo Binomial apresenta o melhor ajuste aos dados. Isso sugere que, considerando apenas os critérios de AIC, o modelo Binomial é o mais adequado para descrever os dados em comparação com os demais modelos analisados.

Comportamento dos dados ao modelo Geométrico (M1)

A primeira aplicação ao modelo discreto foi na distribuição Geométrica, como pode ser visto na Figura 3.

Figure 3: Geometric distribution by GAMLSS - SRAG.



Source: from the authors (2023).

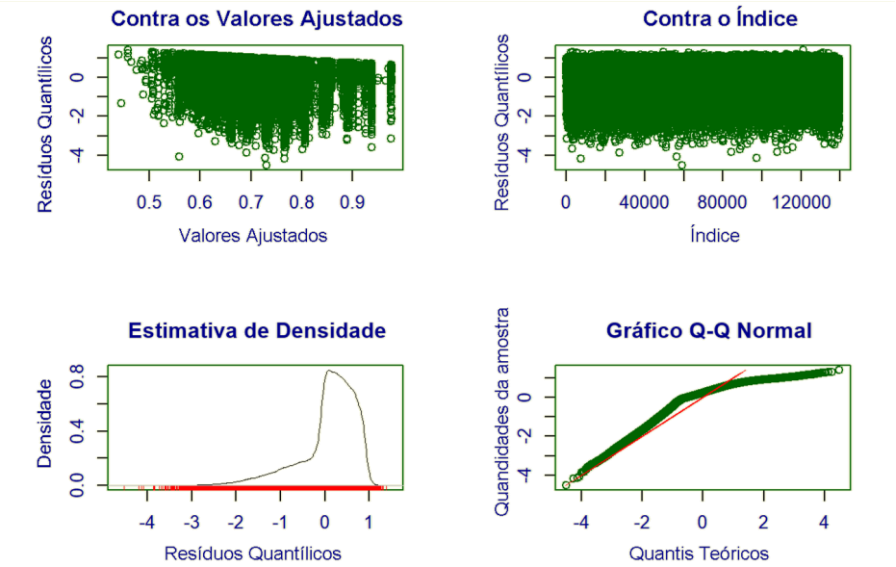
Observa-se na Figura 3 que nos gráficos “Valores Ajustados” e “Índice”, os pontos estão concentrados na parte superior e central, indicando uma tendência dos resíduos para valores mais elevados e sugerindo a possibilidade de padrões não capturados pelo modelo nos dados. O gráfico “Estimativa de Densidade” exibiu uma forma em 'u' invertido para baixo e para a direita, com os pontos formando uma linha paralela próxima ao eixo x, isso indica baixa dispersão dos resíduos, embora a forma em 'u' possa apontar para uma possível assimetria e presença de valores extremos no conjunto de dados com a utilização desse modelo. O “QQ Normal” mostrou uma distribuição dos resíduos que não aderiu estritamente à linha esperada para uma distribuição Normal, com uma parte dos pontos alinhada à linha e outra parte distante, sugerindo uma potencial violação da suposição de normalidade.

Comportamento dos dados ao modelo Binomial Negativa (M2)

A segunda aplicação ao modelo discreto foi na distribuição Binomial Negativa, como pode ser visto na Figura 4.

Nota-se que os gráficos “Valores Ajustados”, “Índice”, “Estimativa de Densidade” e “QQ Normal” da Figura 4 se comportaram de forma parecida com a vista no modelo anterior (Geométrica).

Figure 4: Negative Binomial Distribution by GAMLSS - SARS.



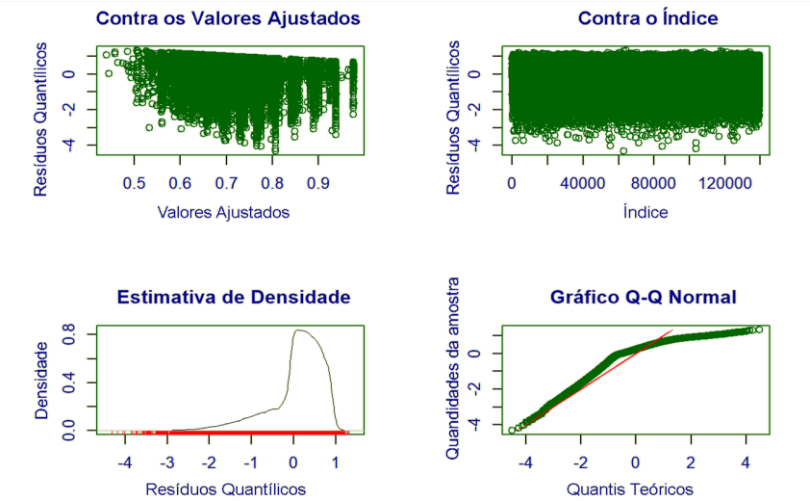
Source: from the authors (2023).

Comportamento dos dados ao modelo Poisson (M3)

A terceira aplicação ao modelo discreto foi na distribuição Poisson, como pode ser visto na Figura 5.

Vê-se que os gráficos “Valores Ajustados”, “Índice”, “Estimativa de Densidade” e “QQ Normal” da Figura 5 se comportaram de forma parecida com a vista no modelo anterior (Geométrica).

Figure 5: Poisson distribution by GAMLSS - SRAG.

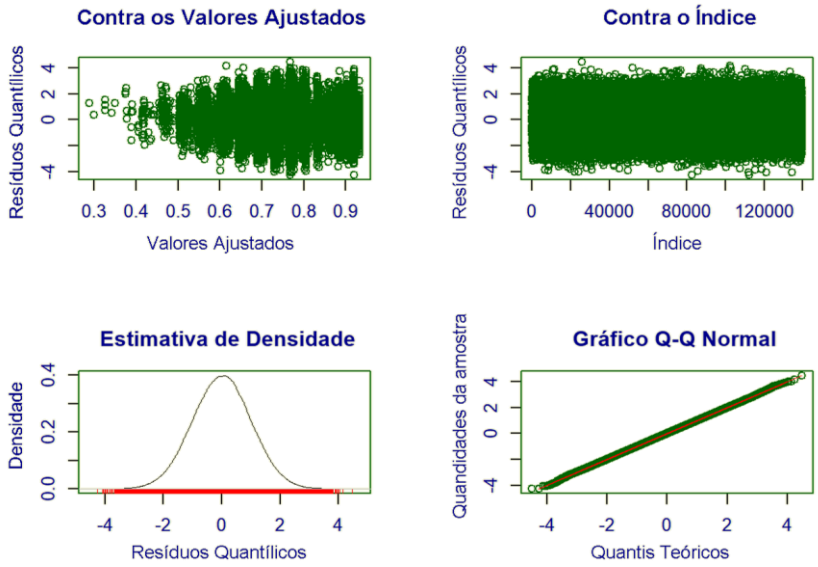


Source: from the authors (2023).

Comportamento dos dados ao modelo Binomial (M4)

A quarta e última aplicação ao modelo discreto foi na distribuição Binomial, como pode ser visto na Figura 6.

Figure 6: Binomial distribution by GAMLSS - SARS.



Source: from the authors (2023).

Percebe-se na Figura 6 que o gráfico “Valores Ajustados” apresenta uma concentração de pontos mais centrada. No gráfico “Índice”, a uniformidade na distribuição dos pontos ao longo do índice dos dados sugere uma distribuição homogênea dos resíduos, o que é considerado favorável. A “Estimativa de Densidade” exibe uma distribuição dos resíduos próxima à normalidade, centrada em torno do zero, com os pontos alinhados em uma linha próxima ao eixo x, indicando baixa dispersão. O gráfico “QQ Normal”, com a maioria dos pontos alinhados à linha, sugere que os resíduos seguem uma distribuição aproximada da Normal.

Resíduos quantílicos randomizados

Sendo verificado os Resíduos quantílicos randomizados a seguir na Tabela 3.

A análise dos resíduos quantílicos randomizados da Tabela 3 para os quatro modelos revela que o modelo Binomial (M4) é o mais adequado para os dados em questão. Este modelo apresenta a média dos resíduos mais próxima de zero, a menor assimetria, uma curtose próxima do ideal e o maior coeficiente de correlação de Filliben, indicando um bom ajuste dos resíduos à distribuição normal. Portanto, o modelo Binomial (M4) é recomendado para a modelagem dos dados com base nessas métricas.

Table 3: Randomized quantile residuals - SRAG.

Descrição	Geométrica (M1)	Binomial Negativa (M2)	Poisson (M3)	Binomial (M4)
Média	0,2245225	0,09255322	0,09103308	-0,001798499
Variância	0,3807319	0,4166499	0,416646	0,9986145
Coeficiente de assimetria	-1,874093	-1,466622	-1,47494	-0,003872364
Coeficiente de curtose	6,807245	5,557564	5,598073	2,997375
Coeficiente de correlação de Filliben	0,8977943	0,9398472	0,9394275	0,9999943

Source: from the authors (2023).

Verificação da normalidade

Sendo verificado a normalidade a seguir na Tabela 4.

Os resultados dos testes Kolmogorov-Smirnov e Lilliefors da Tabela 4 indicam que o modelo Binomial (M4) tem a melhor adequação dos resíduos à distribuição normal, conforme evidenciado pelos menores valores de D e P altos. Esses resultados reforçam a conclusão de que o modelo Binomial (M4) é o mais adequado para modelar os dados, proporcionando uma melhor qualidade de ajuste e uma distribuição mais normal dos resíduos.

Tabela 4: Kolmogorov-Smirnov and Lilliefors heads - SRAG.

Descrição	Geométrica (M1)		Binomial Negativa (M2)		Poisson (M3)		Binomial (M4)	
	KS	Lilliefors	KS	Lilliefors	KS	Lilliefors	KS	Lilliefors
Dados	M1$residuals		M2$residuals		M3$residuals		M4$residuals	
D	0,31649	0,20924	0,21439	0,1392	0,21429	0,13908	0,0017648	0,0017904
P-value	< 2,2e-16	<2,2e-16	< 2,2e-16	<2,2e-16	< 2,2e-16	<2,2e-16	0,7772	0,3425

Source: from the authors (2023).

Superajustamento do modelo de distribuição Binomial (M4)

Com base nos resultados obtidos, não parece haver indícios de superajustamento no modelo Binomial (M4). O processo de avaliação do ajuste do modelo envolveu a realização de previsões nos conjuntos de treinamento e teste, seguida pelo cálculo do RMSE como métrica de desempenho, podendo ser vista na Tabela 5.

Table 5: RMSE: Binomial distribution - SARS.

Métrica	Conjunto de Treino	Conjunto de Teste
RMSE	0,418	0,420
Superajustamento	Não evidenciado	Não evidenciado

Source: from the authors (2023).

O fato de os valores do RMSE da Tabela 5 serem muito próximos pode indicar que o modelo está generalizando bem para dados não vistos. Não há uma disparidade significativa entre o desempenho no conjunto de treino e o conjunto de teste.

Relações entre as variáveis independentes e a variável dependente

Os resultados do modelo GAMLSS da distribuição Binomial com a função de ligação logit, oferecem uma visão detalhada das relações entre as variáveis independentes (comorbidades) e a variável dependente (evolução). Os coeficientes estimados e a Razão de Chances são apresentados nas Tabelas 6 e 7, respectivamente.

Percebe-se na Tabela 6 que a variável ASMA destaca-se como a mais influente, com uma estimativa significativamente alta (1,0256010) e um valor p extremamente baixo (1,742882e-238). Isso sugere que a presença de ASMA está fortemente associada a um aumento significativo nas chances de óbito.

Em seguida, a variável NEUROLOGIC também apresenta uma influência considerável, com uma estimativa negativa robusta (-0,3995408) e um valor p muito baixo (1,060620e-96), indicando uma associação significativa com um menor risco de óbito. Outras variáveis, como HEPATICA, IMUNODEPRE, RENAL e CARDIOPATI, também exibem associações significativas, embora com diferentes direções de influência. Ao analisar a Razão de Chances e os intervalos de confiança (IC) a 95% da Tabela 7.

Table 6: Coefficients of the GAMLSS - SRAG Model.

	Estimativa	Error Padro	Valor T	Pr(>|t|)
(Intercept)	1,4183437	0,01157933	122,489300	0,000000e+00
CARDIOPATI	-0,2255728	0,01320131	17,087148	2,156939e-65
HEMATOLOGI	-0,1204170	0,04431574	2,717251	6,583475e-03
SIND_DOWN	0,1857928	0,08155150	2,278227	2,271455e-02
HEPATICA	-0,5817051	0,04019828	-14,470898	2,002917e-47
ASMA	1,0256010	0,03104001	33,041258	1,742882e-238
DIABETES	-0,1990766	0,01407517	-14,143813	2,192212e-45
NEUROLOGIC	-0,3995408	0,01913174	-20,883665	1,060620e-96
PNEUMOPATI	-0,1717500	0,01901999	-9,029977	1,738147e-19
IMUNODEPRE	-0,3899863	0,02312758	16,862392	9,840820e-64
RENAL	-0,3509015	0,02250555	15,591777	9,210993e-55
OBESIDADE	0,1814572	0,02963420	6,123237	9,193654e-10

Source: from the authors (2023).

Table 7: Odds Ratio and Confidence Intervals - SARS.

	Razão de Chances	IC_95% .2.5	IC_95% .97.5	Variável
(Intercept)	4,1302736	4,0375925	4,2250822	(Intercept)
CARDIOPATI	0,7980590	0,7776748	0,8189775	CARDIOPATI
HEMATOLOGI	0,8865507	0,8127967	0,9669971	HEMATOLOGI
SIND_DOWN	1,2041728	1,0262946	1,4128810	SIND_DOWN
HEPATICA	0,5589445	0,5165969	0,6047634	HEPATICA
ASMA	2,7887711	2,6241676	2,9636996	ASMA
DIABETES	0,8194871	0,7971891	0,8424089	DIABETES
NEUROLOGIC	0,6706279	0,6459467	0,6962522	NEUROLOGIC
PNEUMOPATI	0,8421897	0,8113721	0,8741778	PNEUMOPATI
IMUNODEPRE	0,6770661	0,6470605	0,7084632	IMUNODEPRE
RENAL	0,7040531	0,6736722	0,7358040	RENAL
OBESIDADE	1,1989633	1,1313089	1,2706635	OBESIDADE

Source: from the authors (2023).

Nota-se na Tabela 7 que algumas associações se destacam como as mais significativas, assim como na tabela anterior, a presença de ASMA apresenta um impacto substancial, com um Razão de Chances de 2,79 (IC 95%: 2,62 a 2,96), indicando que indivíduos com asma têm significativamente maiores chances de óbito. Vale destacar a condição de SÍNDROME DE DOWN, que também está associada ao aumento das chances de óbito, com um Razão de Chances de 1,20 (IC 95%: 1,03 a 1,41).

Por outro lado, a presença de HEPATICA está associada a uma redução significativa em relação às demais condições nas chances de óbito, com um Razão de Chances de 0,56 (IC 95%: 0,52 a 0,60).

Considerações finais

Este artigo buscou fazer uma análise aprofundada da SRAG com ênfase na Influenza, empregando e aplicando os GAMLSS como uma abordagem analítica para investigar e compreender a relação entre variáveis preditoras e a variável de resposta específica, analisando por fim os resultados encontrados (casos registrados no Brasil).

A partir da análise dos resultados encontrados, referente ao período de 2020 a agosto de 2022, notou-se que os dados apresentaram um ajuste insatisfatório ao modelo de RLM e MLGs, o que motivou a utilização dos GAMLSS para a análise. Onde dos 04 modelos discretos aplicados dos 09 disponíveis (Binomial, Binomial Negativo, Geométrico e Poisson) o que teve o melhor ajuste aos dados utilizados no estudo, permitindo mapear e determinar a adequação foi o modelo Binomial.

O modelo discreto Binomial (modelo 4) apresentou o melhor ajuste, pois: teve o menor valor de AIC (148545,2); os resultados dos testes (KS e Lilliefors) sinalizaram que não há evidências significativas para afirmar que os resíduos não seguem uma distribuição Normal e não pareceu haver indícios de superajustamento.

Os resultados da análise das relações entre as variáveis independentes e a variável dependente destacaram a variável ASMA como um fator crucial na predição do óbito, mostrando uma associação forte e significativa com um aumento substancial nas chances de mortalidade. Destaca-se que a presença de HEPATICA está associada a uma redução significativa nas chances de óbito. Outras variáveis exibem associações significativas com diferentes direções de influência. Esses resultados fornecem estratégias importantes para

intervenção e cuidado das populações afetadas por essas condições de saúde.

Sugere-se para trabalhos futuros, utilizar os MLGs e/ou os GAMLSS para uma análise em escalas regionais e estaduais do Brasil ou até mesmo de outro território nacional, fazendo uma abordagem analítica para investigar e compreender a relação entre variáveis preditoras e a variável de resposta específica de outras categorias da SRAG.

Agradecimentos

Agradecemos à Universidade Federal Rural de Pernambuco (UFRPE), ao Departamento de Estatística e Informática (DEINFO), ao Programa de Pós-Graduação em Biometria e Estatística (PPGBEA), à Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

AKANTZILIOTOU, C., RIGBY, R. A., & STASINOPOULOS, D. M. (2002). The R Implementation of Generalized Additive Models for Location, Scale and Shape. In M. Stasinopoulos & G. Touloumi (Eds.), Statistical Modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling (pp. 75–83). Chania, Greece.

BRASIL. (2009). Estratégia Nacional de Vacinação contra o Vírus Influenza Pandêmico (H1N1) 2009. Ministério da Saúde. Brasília, DF. Recuperado de https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_nacional_vacinacao_influenza.pdf

BRASIL. (2022). Dicionário de dados. Ministério da Saúde. Secretária de Vigilância em Saúde. Sistema de Informação de Vigilância Epidemiológica da Gripe. Recuperado de https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/pdfs/dicionario_de_dados_srag_hosp_17_02_2022.pdf

BRASIL. (2020). Protocolo de Manejo Clínico. Ministério da Saúde. Brasília, DF. Recuperado de https://www.saude.ms.gov.br/wp-content/uploads/2020/03/Protocolo-Manejo-Clinico_APS_versao04.pdf

DATASUS. (2022). Ministério da Saúde. SRAG - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. OpenDataSUS. Recuperado de https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022

FREITAG, V. L., ANTONIO, M. G. D., LOUREIRO, L. H., & PEREIRA, R. M. S. (2021). COVID 19 e a propagação de fake news sobre a contaminação pelo dióxido de carbono com o uso de máscaras faciais: Um estudo de reflexão. Research, Society and Development, 10(10), e104101018696. https://doi.org/10.33448/rsd-v10i10.18696

GONZALEZ, TEOFILO; SAHNI, SARTAJ; FRANTA & WILLIAM R. (1977). An efficient algorithm for the Kolmogorov-Smirnov and Lilliefors tests. ACM Transactions on Mathematical Software (TOMS), v. 3, n. 1, p. 60-64.

HASTIE, T. J., & TIBSHIRANI, R. J. (1990). Generalized Additive Models. Chapman and Hall.

LI, S., & YANG, B. (2008). Region-based multi-focus image fusion. https://doi.org/10.1016/B978-0-12-372529-5.00009-3

MAGALHÃES, M. N. (2015). Probabilidade e Variáveis Aleatórias. EDUSP.

NELDER, J. A., & WEDDERBURN, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society, A, 135, 370–384.

RIBEIRO, M. L. F., CORDEIRO, N. M., & ALVES, D. A. N. DA S. (2022). Aplicação de modelagem preditiva via árvore de decisão em casos de Síndrome Respiratória Aguda Grave (SRAG), com ênfase na Doença por Corona Vírus 2019 (COVID-19) no Brasil para o período de 2020 a 2022. Research, Society and Development, 11(15), e01111536173. https://doi.org/10.33448/rsd-v11i15.36173

RIGBY, R. A., & STASINOPULOUS, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. Applied Statistics, 54, 507–554.

RIGBY, R. A., & STASINOPULOUS, D. M. (2001). The GAMLSS project: a Flexible Approach to Statistical Modelling. In B. Klein & L. Korsholm (Eds.), New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling (pp. 249–256). Odense, Denmark.

SAMPERI, R. H., COLLADO, C. F., & LUCIO, M. DEL P. B. (2013). Metodologia Científica. AMGH Editora.