UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CÂMPUS DE PRESIDENTE PRUDENTE

LAÍS SILVEIRA YOUSSEF

APLICAÇÃO DO MODELO DE REGRESSÃO DE POISSON PARA O ESTUDO DE CASOS DE ÓBITOS POR SEPTICEMIA NO ESTADO DE SÃO PAULO NOS ANOS DE 2014 A 2018

PRESIDENTE PRUDENTE
2023

LAÍS SILVEIRA YOUSSEF

APLICAÇÃO DO MODELO DE REGRESSÃO DE POISSON PARA O ESTUDO DE CASOS DE ÓBITOS POR SEPTICEMIA NO ESTADO DE SÃO PAULO NOS ANOS DE 2014 A 2018

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da Faculdade de Ciências da Universidade Estadual Paulista "Júlio de Mesquita Filho" para aproveitamento na disciplina Trabalho de Conclusão de Curso. Orientador: Prof. Dr. Mário Hissamitsu Tarumoto.

PRESIDENTE PRUDENTE

2023

Y83a

Youssef, Laís Silveira

Aplicação do modelo de regressão de Poisson para o estudo de casos de óbitos por septicemia no estado de São Paulo nos anos de 2014 a 2018 / Laís Silveira Youssef. -- Presidente Prudente, 2023

47 p. : tabs.

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

Orientador: Mário Hissamitsu Tarumoto

1. Dados de Contagem. 2. Distribuição Poisson. 3. Septicemia. I. Título.

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

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Laís Silveira Youssef

APLICAÇÃO DO MODELO DE REGRESSÃO DE POISSON PARA O ESTUDO DE CASOS DE ÓBITOS POR SEPTICEMIA NO ESTADO DE SÃO PAULO NOS ANOS DE 2014 A 2018

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

Orientador: _____

Prof. Dr. Mário H. Tarumoto
Departamento de Estatística

_____

Prof. Dr. Sérgio M. Oikawa
Departamento de Estatística

_____

Prof. Dr. Fernando Antonio Moala
Departamento de Estatística

Presidente Prudente, 27 de janeiro de 2023.

AGRADECIMENTOS

A realização deste trabalho se fez mais simples com a orientação do Prof. Dr. Mário Hissamitsu Tarumoto, portanto, o agradeço por todo direcionamento e ensinamentos.

Agradeço aos membros da banca, Prof. Dr. Sérgio Oikawa e Prof. Dr. Fernando Antonio Moala, pelo auxílio durante todo o desenvolvimento deste trabalho e por terem sido excelentes professores.

A todos os professores da graduação, agradeço por fazerem parte de uma fase tão importante da minha vida, e serem exemplos de dedicação e conhecimento. Sem vocês minha formação não seria tão completa.

Agradeço principalmente meus pais, que sempre prezaram pela minha educação, por me incentivarem, me darem apoio e estarem presentes de fato.

Aos meus avós, meu irmão, e meu namorado, por fazerem parte da minha formação como pessoa, e aliviarem todos os momentos ruins.

Agradeço a todos os amigos que acompanharam essa jornada de perto, sem esse apoio não teria chegado aqui.

RESUMO

O estudo do comportamento dos óbitos é extremamente relevante para tomada de medidas e entendimento dos problemas e necessidades de uma sociedade. Uma análise exploratória do Sistema de Informações sobre Mortalidade (SIM) do DATASUS, indicou um expressivo número de óbitos com causa primária sendo septicemia, portanto, foi vista a necessidade de analisar essa informação. Foi verificado que o comportamento do número de óbitos por septicemia em relação a outras variáveis contidas na base de dados não é influenciado, assim, sendo um indicativo de que estas variáveis não trarão nenhuma informação adicional a ocorrência de óbitos por septicemia, essas variáveis não foram levadas em consideração no prosseguimento do estudo. A análise leva em consideração a contagem de óbitos por septicemia diários ao longo do tempo observado, e foi realizada utilizando os softwares R e SAS. Na análise de dados de contagem, diversos modelos estatísticos podem ser utilizados, foi considerado o modelo de regressão de Poisson, o que na prática, a utilização desse modelo para dados de contagem, resulta na frequente ocorrência de superdispersão, assim como neste trabalho foi observada essa ocorrência. A superdispersão se dá quando variabilidade é muito maior do que o esperado. Após alguns testes, foi escolhida a equação a ser utilizada para o ajuste do modelo, devido a não linearidade da equação, para que ela fosse maximizada foi necessário o uso de métodos computacionais. Foi notória a falta de convergência da estimativa de três dos cinco parâmetros, o que mostra que os ajustes não foram satisfatórios, isso também pôde ser verificado graficamente e pela análise dos resíduos. Apesar dos ajustes não serem satisfatórios, foi possível observar que os óbitos por septicemia ocorrem com maior frequência em pessoas acima de 60 anos, não se concentra em nenhum município do estado de São Paulo, ou seja, de forma geral ocorre proporcionalmente ao tamanho da população do município, e se comporta igualmente entre os sexos.

Palavras-chave: Distribuição Poisson; Septicemia; Dados de Contagem.

ABSTRACT

The study of the behavior of deaths is extremely relevant for taking measures and understanding the problems and needs of a society. An exploratory analysis of the Mortality Information System (SIM) of DATASUS, indicated an expressive number of deaths with primary cause being sepsis, therefore, it was seen the need to analyze this information. It was found that the behavior of the number of deaths from sepsis in relation to other variables contained in the database is not influenced, thus being an indication that these variables will not bring any additional information to the occurrence of deaths from sepsis, these variables were not taken into consideration in the further study. The analysis considers the daily death count over the observed time and was performed using R and SAS software. In the analysis of count data, several statistical models can be used, the Poisson regression model was considered, which in practice, the use of this model for count data, results in the frequent occurrence of overdispersion, as was observed in this study. Overdispersion occurs when the variability is much larger than expected. After some tests, the equation to be used for model fitting was chosen, due to the non-linearity of the equation, for it to be maximized it was necessary to use computational methods. The lack of convergence of the estimate of three of the five parameters was notorious, which shows that the adjustments were not satisfactory, this could also be verified graphically and by the analysis of the residuals. Although the adjustments were not satisfactory, it was possible to observe that deaths from sepsis occur more frequently in people over 60 years, it is not concentrated in any city in the state of São Paulo, i.e., in general, it occurs proportionally to the size of the population of the city and behaves equally between genders.

Keywords: Poisson Distribution; Sepsis; Counting Data.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

FE – Família Exponencial

IBGE – Instituto Brasileiro de Geografia e Estatística

MLG – Modelo Linear Generalizado

MS – Ministério da Saúde

MV – Máxima Verossimilhança

SAS – Statistical Analysis System

SIM – Sistema de Informações de Mortalidade

V.A. – Variável Aleatória

SUMÁRIO

1 INTRODUÇÃO

Entre os vários conjuntos de dados públicos existentes no Brasil, o DATASUS é uma importante fonte de informações que pode ser usada como meta dados ou dados principais na construção de modelos. Este sistema, conhecido como Departamento de Informática do Sistema Único de Saúde (DATASUS), foi criado no início dos anos 1990 com o intuito de promover ações direcionadas à coleta, processamento e disseminação de informação sobre saúde. Toda produção de dados públicos relacionados a essa área são gerenciados pelos sistemas de informações pertencentes ao DATASUS. Este departamento pertence à Secretaria Executiva do Ministério da Saúde (MS), e tem como principal objetivo estruturar sistemas de informação, integrar dados em saúde e auxiliar na gestão dos diversos níveis de atenção à saúde. Entre os vários sistemas administrativos existentes no DATASUS, neste projeto, serão utilizados os dados dos arquivos disseminados para tabulação do Sistema de Informações sobre Mortalidade do SUS (SIM), disponível em: http://www2.datasus.gov.br/ DATASUS.

Considerando que o DATASUS é uma fonte importante e confiável, será a principal base de dados a ser utilizada neste projeto para a construção de modelos estatísticos. Este órgão é responsável por coletar, processar e disseminar informações sobre saúde no Brasil. Em conjunto com os dados populacionais disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE), será possível avaliar como se encontra a saúde em um local ou região específica a nível nacional.

Entre estas bases de dados, o SIM - Arquivos disseminâveis para tabulação do Sistema de informações de Mortalidade, contém registros de todas as causas de óbitos no Brasil desde 1994, utilizando as informações constatadas no atestado de óbito. Um estudo preliminar apontou que entre as principais causas de óbitos, destaca-se em termos quantitativos, a septicemia, que é caracterizada pela reação exagerada do organismo a um foco infeccioso pré-existente. De acordo Varella (2014) a sepse, como também é conhecida, não necessariamente ocorre apenas em pessoas hospitalizadas, ela pode ser desenvolvida por pessoas saudáveis. Porém, existem alguns fatores de risco como diabetes, câncer, infecção pelo HIV, uso de álcool ou outras drogas, recém-nascidos e idosos.

Os focos infecciosos mais comuns instalam-se normalmente nos pulmões (pneumonia), abdômen (apendicite e peritonite), rins e bexiga (infecções urinárias e renais), na pele (feridas, abcessos e erisipela) e no sistema nervoso central (meningite). Em média, a mortalidade por essa doença no Brasil é de 65%, por outro lado, a mortalidade mundial é de 30% a 40%, apresentado em um estudo por Freitas (2018).

A relevância deste trabalho, está no fato, que a base de dados utilizada, o Sistema de Informação sobre Mortalidade (SIM) do DATASUS, é uma importante fonte de dados, a qual permite entender o comportamento das principais causas de óbitos no país. De acordo com o Ministério da Saúde, o SIM tem a finalidade de reunir dados quantitativos e qualitativos sobre óbitos ocorridos no Brasil, e é considerado uma importante ferramenta de gestão na área da saúde que subsidia a tomada de decisão em diversas áreas da vigilância e assistência à saúde.

As informações obtidas desse Sistema de Informação, podem ser utilizadas para criação de campanhas de prevenção ou educacionais com a finalidade de reduzir a mortalidade. Pois, mesmo a morte sendo algo esperado, quando é notável a sua ocorrência em faixas etárias nas quais não deveria e por razões que podem ser controladas, medidas podem ser tomadas. Fica evidente, então, que estudar os óbitos e suas causas, pode contribuir para a inserção de medidas e criação de projetos para que os problemas possam ser solucionados.

Após observado o grande número de óbitos que possuem como causa primária a septicemia, nota-se a importância de estudar tal fato, e ao obter informações como as possíveis causas e justificativas para uma quantidade tão expressiva de número de óbitos por septicemia, ações podem ser realizadas, para que esse número diminua, ou, ao mínimo, que ele seja compreendido.

Entre as várias possibilidades de estudos de dados desta natureza, a contagem do número de casos diários, semanais ou mensais, considerando-se a localidade, pode ser importante, tendo em vista a possibilidade de comparações entre os períodos do ano ou ainda entre regiões. Para os dados de contagem, como o do caso deste trabalho, a distribuição Poisson pode ser adequada.

Levando-se em consideração a existência de outras informações importantes que podem influenciar no número de casos, o modelo de regressão deve ser considerado. Estudos desta natureza, podem ser analisados para que o

comportamento do número de casos ao longo tempo seja visto. Neste trabalho, estudos exploratórios indicaram que as várias covariáveis consideradas não foram importantes para explicar o comportamento dos dados ao longo do tempo, portanto, a análise poderá ser comparada à análise de séries temporais. Desta forma, este projeto tem o intuito de construir modelos de regressão para dados de contagem observados ao longo tempo, como o modelo de regressão Poisson, considerando-se como um processo de markov. Outra possibilidade de análise, é a utilização da análise de séries temporais, no entanto, não é o escopo do trabalho.

Os programas estatísticos utilizados para analisar os dados foram o RStudio e o SAS Viya. O R é uma linguagem de programação estatística, criada baseada na linguagem S, não mais utilizada, e também um programa gratuito, que está disponível em www.r-project.org. O SAS é uma empresa pioneira em análise de dados de negócios, com seu software é possível realizar todas as etapas do ciclo analítico, disponível em www.sas.com.

2 MODELOS PARA DADOS DE CONTAGEM

2.1 MODELO LINEAR GENERALIZADO

O modelo de regressão linear descreve grande parte dos fenômenos aleatórios, porém, ele está restrito a dados que seguem uma distribuição Normal, variância homogênea e erros independentes. Quando não atingidos os pressupostos, algumas transformações podem ser realizadas para que a normalidade seja atingida e a variância homogeneizada, como por exemplo a transformação Box Cox.

Os modelos lineares generalizados (MLGs), propostos em 1972 por Nelder e Wedderburn, se adequam aos dados em que não é razoável se assumir normalidade.

Os MLGs são caracterizados por:

- Um componente aleatório, responsável por identificar a variável resposta e sua distribuição de probabilidade. Sejam $\mathrm{Y}_1, \ldots, \mathrm{Y}_n$ variáveis aleatórias independentes, sendo Y_i discreta ou contínua, com $\mathrm{E}(\mathrm{Y}_i) = \mu_i$, distribuição de probabilidade pertencente à FE (família exponencial) e com função de probabilidade ou função densidade na forma dada abaixo

$$f_{Y_i}(y|\theta, \phi) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right]$$

  em que $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas, com $\phi^{-1} > 0$ sendo o parâmetro de dispersão e θ_i o parâmetro de posição, sendo ele uma função de μ_i. A FE tem como propriedade que: $\mu_i = \mathrm{E}(\mathrm{Y}_i) = b'(\theta_i) = db(\theta_i)/d\theta_i$ e $\mathrm{Var}(\mathrm{Y}_i) = \phi^{-1}V(\mu_i)$ em que $V_i = V(\mu_i) = d\mu_i/d\theta_i$, é a função de variação, que dentro da família exponencial é a responsável por caracterizar a distribuição, ou, de outra forma, $\mathrm{Var}(\mathrm{Y}_i) = \phi^{-1}b''(\theta_i)$.
- Um componente sistemático em que $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ o vetor que contém valores das variáveis explicativas para cada indivíduo e sendo $x_{i1} = 1$ para todo i.
- Uma função de ligação $g(\cdot)$, sendo ela uma função monótona e diferenciável que relaciona o preditor linear ao valor esperado da distribuição do componente aleatório, de forma simplicada, vincula a média

ao preditor linear. A função de ligação pode, também, modelar uma transformação da média μ_i em função do parâmetro de posição θ_i como uma função linear nos parâmetros

$$g\big(\mathrm{E}(Y_i)\big) = g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

É possível sintetizar a construção de um MLG baseando-se em três principais questões:

1. Qual a distribuição de probabilidade da variável resposta?
2. Quais variáveis explicativas melhor ajudam a descrever o comportamento da variável resposta?
3. Qual é a função de ligação mais adequada?

Segundo Paula (2013) as ligações mais utilizados são logarítmica $g(\mu_i) = \log \mu_i$, raiz quadrada $g(\mu_i) = \sqrt{\mu_i}$ e identidade $g(\mu_i) = \mu_i$.

2.1.1 Estimação dos parâmetros $\beta's$

Como anteriormente mencionado, existem três principais escolhas na aplicação de um MLG: a distribuição de probabilidade, a matriz modelo que contém as variáveis explicativas e a função de ligação. Alguns métodos podem ser utilizados para a estimação dos parâmetros $\beta's$, o principal, que será tratado a seguir, é o método de máxima verossimilhança (MV), que possui propriedades assintóticas importantes, como consistência, eficiência e normalidade.

Para obter a estimativa dos parâmetros, o logaritmo da função de verossimilhança deve ser maximizado, para isso, as seguintes etapas são realizadas:

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi)$$

$$L(\theta) = \prod_{i=1}^{n} \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right]$$

$$L(\theta) = \exp\left[\phi \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi)\right]$$

Aplicando o logaritmo na função de verossimilhança:

$$\ell(\theta) = \ln L(\theta)$$

$$\ell(\theta) = \phi \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi). \quad (2.1)$$

Obtida a função log-verossimilhança e considerando a partição $\theta = (\beta^T, \phi)^T$, (Paula, 2013), a função escore é dada por $U_i(\theta) = \frac{\partial \ell(\theta)}{\partial \beta_i}$, e as estimativas de máxima verossimilhança de β serão dadas a partir da solução do sistema de equações de verossimilhança

$$U_\beta(\theta) = \frac{\partial \ell(\theta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta)}{\partial \beta_j} = 0, j = 1, \dots, p\ .$$

Aplicando a regra da cadeia tem-se:

$$\frac{\partial \ell(\theta)}{\partial \beta_j} = \phi \sum_{i=1}^{n} \left\{ y_i \frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} - \frac{db(\theta_i)}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} \right\}$$

$$= \phi \sum_{i=1}^{n} \left\{ y_i V_i^{-1} \left(\frac{d\mu_i}{d\eta_i}\right) x_{ij} - \mu_i V_i^{-1} \left(\frac{d\mu_i}{d\eta_i}\right) x_{ij} \right\}$$

$$= \phi \sum_{i=1}^{n} \left\{ V_i^{-1} \left(\frac{d\mu_i}{d\eta_i}\right) x_{ij}(y_i - \mu_i) \right\}$$

$$= \phi \sum_{i=1}^{n} \left\{ \sqrt{\frac{\omega_i}{V_i}}(y_i - \mu_i) x_{ij} \right\},$$

em que $\omega_i = \left(\frac{d\mu_i}{d\eta_i}\right)^2 / V_i$. A função escore, então, pode ser escrita na forma matricial

$$\mathrm{U}_{\beta}(\theta) = \phi \mathbf{X}^{T} \mathbf{W}^{\mathbf{1/2}} \mathbf{V}^{\mathbf{-1/2}} (\boldsymbol{y} - \boldsymbol{\mu}), \tag{2.2}$$

em que $\mathbf{X}^{\mathrm{T}} = (X_1, \dots, X_m)$ é uma matriz $m \times p$ de posto completo, $\mathbf{W} = \mathrm{diag}\{\omega_1, \dots, \omega_n\}$ é uma matriz diagonal de pesos denotados por ω_i, $\mathbf{V} = \mathrm{diag}\{v(\mu_1), \dots, v(\mu_m)\}$ é uma matriz diagonal que contém as funções de variância, $\boldsymbol{y} = (\mathrm{y}_1, \dots, y_m)^{\mathrm{T}}$ é o vetor de respostas e $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)^{\mathrm{T}}$ são as respectivas respostas.

Maximizar (2.1) é equivalente a solucionar (2.2), na maioria dos casos isso não pode ser feito analiticamente, portanto, métodos iterativos são utilizados para obter a máxima verossimilhança.

A estimativa de β é normalmente obtida utilizando o método Newton-Raphson (mínimos quadrados iterativos ponderados), que consiste em expandir a função escore U_{β} em torno de um valor inicial $\beta^{(0)}$, tal que

$$\mathrm{U}_{\beta} \cong \mathrm{U}_{\beta}{}^{(0)} + \mathrm{U}'_{\beta}{}^{(0)} (\beta - \beta^{(0)}),$$

em que U'_{β} denota a primeira derivada de U_{β} em relação a β^{T}, sendo $\mathrm{U}'_{\beta}{}^{(0)}$ e $\mathrm{U}_{\beta}{}^{(0)}$, respectivamente, essas quantidades avaliadas em $\beta^{(0)}$. Assim, ao realizar o procedimento acima diversas vezes, chega-se ao processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + \left[\left(-\mathrm{U}'_{\beta}\right)^{-1}\right]^{(m)} \mathrm{U}_{\beta}{}^{(m)}, \; m = 0,1,2 \dots$$

sendo $\left(-\mathrm{U}'_{\beta}\right)^{k}$ a matriz de informação observada com elementos $\left(\frac{-\partial^2 \ell(\theta)}{\partial \beta \partial \beta^{\mathrm{T}}}\right)$, avaliados em $\beta = \beta^{(k)}$. Como a matriz $-\mathrm{U}'_{\beta}$ não pode ser positiva definida, a aplicação do método escore de Fisher substituindo a matriz $-\mathrm{U}'_{\beta}$ pelo correspondente valor esperado $\mathbf{K}_{\beta\beta}$ pode ser mais conveniente, o que resulta no processo iterativo a seguir

$$\beta^{(m+1)} = \beta^{(m)} + \left[\mathbf{K}_{\beta\beta}{}^{-1}\right]^{(m)} \mathrm{U}_{\beta}{}^{(m)}, \; m = 0,1,2 \dots$$

Paula (2013) mostra que ao trabalhar o lado direito da expressão acima chega-se em

$$\beta^{(m+1)} = \left(\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{z}^{(m)},\ m = 0,1,2\ldots \quad (2.3)$$

em que $\boldsymbol{z} = \boldsymbol{\eta} + \mathbf{W}^{\mathbf{1/2}}\mathbf{V}^{\mathbf{-1/2}}(\boldsymbol{y} - \boldsymbol{\mu})$**.** Em um número finito de passos é alcançada a convergência de (2.3), independente dos valores iniciais a serem utilizados. Nota-se que o lado direito de (2.3) não depende de ϕ, portanto, para estimar β não é necessário ser conhecido o valor de ϕ.

2.1.2 Estimação do parâmetro ϕ

A estimação do parâmetro ϕ também pode ser feita pelo método da máxima verossimilhança, sendo assim, sua função escore é dada por

$$U_\phi(\theta) = \frac{\partial \ell(\theta)}{\partial \phi} = \sum_{i=1}^{n}\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n}\frac{\partial c(y_i,\phi)}{\partial \phi}.$$

Igualando a função escore $U_\phi(\theta)$ a zero, é mostrado por Paula (2013) a seguinte solução:

$$\sum_{i=1}^{n} c'\left(y_i,\hat{\phi}\right) = \frac{1}{2}\mathrm{D}(y;\hat{\mu}) - \sum_{i=1}^{n}\{y_i\tilde{\theta}_\iota - b\left(\tilde{\theta}_\iota\right)\},$$

em que $\mathrm{D}(y;\hat{\mu})$ denota o desvio do modelo. Tendo sido os parâmetros estimados anteriormente.

2.2 MODELO DE REGRESSÃO DE POISSON

A distribuição de Poisson é uma distribuição de probabilidade discreta e é utilizada para analisar o número de ocorrências de um evento durante um intervalo fixado de tempo, distância, área ou volume, sendo assim, o modelo de Poisson possui um importante papel na análise de dados de contagem.

Seja Y_i uma variável aleatória que segue distribuição de probabilidade Poisson, $\mathrm{P}(\mu_i)$, com parâmetro $\mu_i > 0$, então, sua função de probabilidade é dada por

$$P(Y_i = y_i) = e^{-\mu_i} \frac{{\mu_i}^{y_i}}{y_i!} \text{ , para } y_i = 0, 1, 2, \ldots \quad (2.4)$$

Uma das propriedades da distribuição de Poisson é que a média e a variância coincidem, tendo assim, $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \mu_i$. Sendo assim, a distribuição de Poisson possui apenas um parâmetro (μ_i), que define a média e a variância.

O modelo de Poisson é um dos inúmeros exemplos de modelos que podem ser citados como sendo modelos lineares generalizados, pois, a distribuição de Poisson pertence à família exponencial, assim, a sua função densidade (2.4) pode ser reescrita da seguinte forma:

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\{[y_i \ln(\mu_i) - \mu_i] - \ln(y_i!)\},$$

em que o parâmetro de dispersão é $\phi = 1$, a função de ligação canônica é $\ln(\mu_i)$, $\mu_i = e^{\theta_i}$, portanto, $b(\theta_i) = \mu_i = e^{\theta_i}$ e $c(y_i, \phi) = -\ln(y_i!)$. Assim, $\mathrm{E}(\mathrm{Y}_i) = b'^{(\theta_i)} = \mu_i = e^{\theta_i}$ e $\mathrm{Var}(\mathrm{Y}_i) = \phi^{-1} b''(\theta_i) = 1 \times de^{\theta_i}/d\theta_i = e^{\theta_i}$, $\mathrm{E}(\mathrm{Y}_i) = \mathrm{Var}(\mathrm{Y}_i)$. Sendo a função de ligação $g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ **,** para a função de ligação canônica $\ln \mu_i$, tem-se $\ln(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, portanto, $\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$.

Para uma única variável explicativa, o modelo loglinear Poisson tem a seguinte forma:

$$\ln(\mu_i) = \beta_0 + \beta_1 x \, .$$

É descrito por Cordeiro e Demétrio (2008) que uma das principais características do modelo de Poisson é gerar uma boa descrição de dados cuja variância é proporcional à média. O que, na prática, resulta na frequente ocorrência de superdispersão quando se utiliza o modelo de regressão de Poisson na análise de dados de contagem.

Para casos de superdispersão, a suposição de que $\phi = 1$ já não é mais válida, sendo necessário buscar alternativas para prosseguir com a análise. Algumas delas: (i) abordagem bayesiana, na qual pode ser assumido que o parâmetro do modelo pode ser representado por uma distribuição de probabilidade; (ii) a estimação

por quase-verossimilhança, incluindo um fator de dispersão diferente da unidade ou uma função de variância alternativa ou (iii) a utilização do modelo Binomial Negativo, entre outras. (TRINDADE, 2014).

2.3 SUPERDISPERSÃO PARA DADOS DE CONTAGEM

É descrito por Agresti (2007) que o fenômeno da superdispersão para os MLGs ocorre quando a variabilidade é muito maior do que o esperado, o que corrobora com o entendimento de Dobson (2002), que define que a superdispersão ocorre quando $Var(Y) > E(Y)$.

De acordo com Cordeiro e Demétrio (2008) a superdispersão pode ser ocasionada por diferentes motivos. A causa pode ser do processo da coleta de dados, pela falta de independência das observações ou pela ausência de covariáveis que possam explicar a heterogeinidade entre as observações. Uma consequência da superdispersão é que os erros padrão das estimativas do modelo estarão incorretos e, também, os desvios serão muito grandes conduzindo à seleção de modelos complexos.

É evidenciado por McCullagh e Nelder (1989): “A superdispersão não é incomum na prática. Na verdade alguns sustentam que a superdispersão é normal na prática e a dispersão nominal é exceção”. Nessa situação, uma alternativa ao modelo de Poisson é dado por Dobson (2002) a utilização do modelo Binomial Negativa.

Tal como a distribuição de Poisson, a distribuição Binomial Negativa assume apenas valores inteiros positivos para a variável resposta. Mas, diferentemente da distribuição de Poisson que possui um parâmetro tanto para a média quanto para a variância, a distribuição Binomial Negativa possui dois parâmetros diferentes, um para a média e um para a variância, o que gera melhor ajuste para casos de superdispersão.

Apesar disso, a interpretabilidade dos coeficientes da regressão de Poisson é uma vantagem. Os coeficientes de regressão podem ser interpretados como uma estimativa do logaritmo do risco relativo, ajustado para os demais preditores do modelo, o que coloca a utilização do modelo Binomial Negativa em desvantagem,

considerando que sua interpretação é mais complexa. Para mais detalhes, ver Cordeiro e Demétrio (2008) cap. 10.9.

2.4 MÉTODOS DE DIAGNÓSTICO PARA MLGs

Os métodos de diagnóstico dos modelos lineares generalizados são semelhantes aos métodos utilizados para os modelos de regressão clássicos. Esta etapa é de extrema importância, pois nela são verificados se pressupostos foram violados, e, se assim ocorreu, o modelo em questão não possui validade.

Na verificação da pressuposição de linearidade para o modelo de regressão clássico utilizam-se os vetores $\mathbf{Y}$ e $\widehat{\boldsymbol{\mu}}$, enquanto para os modelos lineares generalizados utiliza-se a variável dependente ajustada estimada $\hat{\mathbf{z}}$ e o preditor linear $\widehat{\boldsymbol{\eta}}$. A variância residual é substituída por uma estimativa consistente de ϕ e a matriz de projeção $\mathbf{H}$ (matriz chapéu) é definida por

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{1/2}, \tag{2.5}$$

tendo como propriedades $tr(\mathbf{H}) = p$ e $0 \leq h_{ii} \leq 1$.

2.4.1 Resíduos

A importância dos resíduos está no fato de que eles ajudam na detecção de outliers, que precisam ser estudados detalhadamente. O resíduo R_i deve demonstrar a distância entre a observação y_i e o seu valor ajustado $\hat{\mu}_i$.

$$R_i = h_i(y_i, \hat{\mu}_i)$$

sendo h_i uma função adequada com simples interpretação, escolhida baseada na anomalia que deseja se detectar no modelo, como por exemplo, situações em que a variância precisa ser estabilizada, ou em casos de assimetria, no qual a simetria é induzida na distribuição amostral de R_i. A definição referente ao resíduo R_i foi dada por Cox e Snell (1968) e está contida em Cordeiro e Demétrio (2008).

Alguns exemplos dos resíduos que são mais comuns nos MLGs:

- Resíduos ordinários: $r_i = y_i - \hat{\mu}_i$.
- Resíduos de Pearson: $r_i^P = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$, sendo $V(\hat{\mu}_i)$ a função de variância.
- Resíduos de Pearson studentizados: $r_i^{P\prime} = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)(1 - \hat{h}_{ii})}$, sendo h_{ii} o i-ésimo elemento da diagonal da matriz chapéu, definida em (2.5). Os resíduos studentizados $r_i^{P\prime}$ têm, aproximadamente, variância igual a um quando o parâmetro de dispersão, ϕ, tende a zero.

2.4.2 Métodos gráficos

Alguns dos métodos gráficos mais utilizados para os MLGs:

- Resíduos versus valores ajustados: pode mostrar se a variância é heterogênea, deve-se ter uma distribuição dos resíduos em torno de zero com amplitude constante.
- Valores observados ou resíduos versus tempo: mesmo que a variável tempo não esteja incluída no modelo, é importante a construção desse gráfico. Ele pode ajudar na detecção de padrões e de variáveis que sejam altamente correlacionadas com o tempo.
- Gráfico normal e semi-normal de probabilidades (“normal plots” e “half normal plots”): construído da mesma forma que para os modelos de regressão clássicos, porém utiliza-se a distribuição do modelo em questão.

Para maior aprofundamento teórico dos métodos de diagnósticos para os modelos lineares generalizados, ver Cordeiro e Demétrio (2008). Existem materiais disponibilizando os códigos em R para a construção dos gráficos normal e semi-normal de probabilidades.

3 APLICAÇÃO

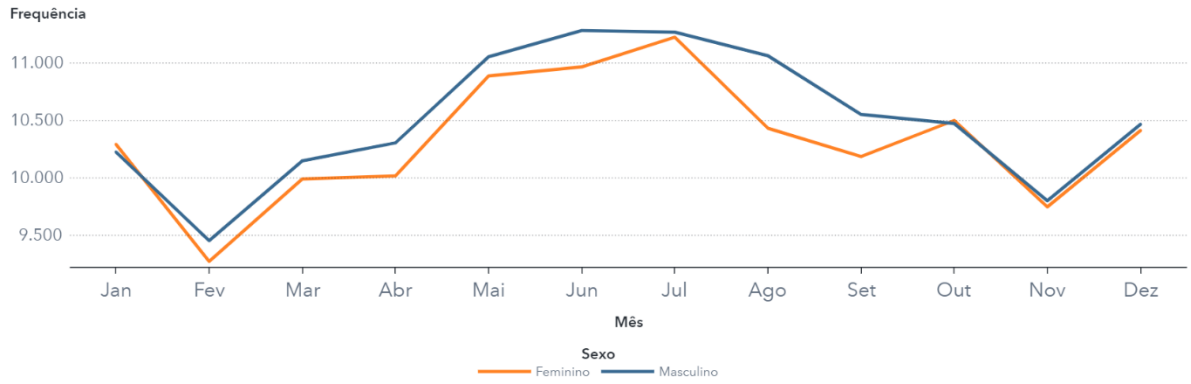
3.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Utilizando os softwares R e SAS foram realizadas análises exploratórias das variáveis disponibilizadas pelo DATASUS. Parte delas referenciava-se à características relacionadas a gravidez, portanto, considerando o tema principal do trabalho, sua relevância era baixa e essas variáveis não foram utilizadas.

Considerando o número de óbitos mensais, cuja causa principal foi a septicemia, as variáveis analisadas foram sexo, município de residência, idade, ocupação e doença de causa secundária do óbito.

Pela Figura 1, que apresenta o número de óbitos mensais por septicemia, parece não haver evidências de que haja diferença no comportamento entre os sexos. Desta forma, nesta análise, vamos considerar que o comportamento do número de óbitos mensais por esta causa, é a mesma entre os dois sexos. No total, dos 250.023 indivíduos analisados, 123.921 eram mulheres (49,56%), 126.091 (50,43%) eram homens e em 11 indivíduos não havia a informação do sexo, portanto, não foi considerada na análise.

Figura 1- Gráfico de Linhas de Óbitos por mês de 2014 a 2018 para os sexos



Fonte: Elaborado pela autora (2022)

Ao longo dos cinco anos analisados, aproximadamente 29% dos óbitos estão concentrados em indivíduos dos quais o município de residência é São Paulo, sendo

este, o que mais concentra óbitos. Estão contidos na Tabela 1 apenas os dez municípios com maiores frequências de óbitos.

Como pode ser observado na Tabela 1, há uma grande discrepância entre o número de óbitos do município com mais frequência para o segundo município com mais frequência. Porém, mesmo sendo notória a diferença, ao ser calculada uma taxa entre o número de óbitos de cada município e a estimativa da população (estimativa populacional de 2019), foram observados resultados muito próximos entres os municípios, corroborando a percepção de não haver diferença no número de óbitos por septicemia entre os municípios.

Tabela 1 - Distribuição de Frequência por Município de Residência 2014 a 2018

Município	**Nº de Óbitos**	**Estimativa 2019 Pop**	**Taxa**
São Paulo	73039	12252023	0,00596
Guarulhos	8057	1379182	0,00584
Campinas	6682	1204073	0,00555
Ribeirão Preto	5835	703293	0,00830
Santo André	5231	718773	0,00728
Osasco	4309	698418	0,00617
São Bernardo do Campo	4052	838936	0,00483
Sorocaba	3461	679378	0,00509
São José dos Campos	3444	721944	0,00477
Jundiaí	3401	418963	0,00812
São José do Rio Preto	3216	460671	0,00698

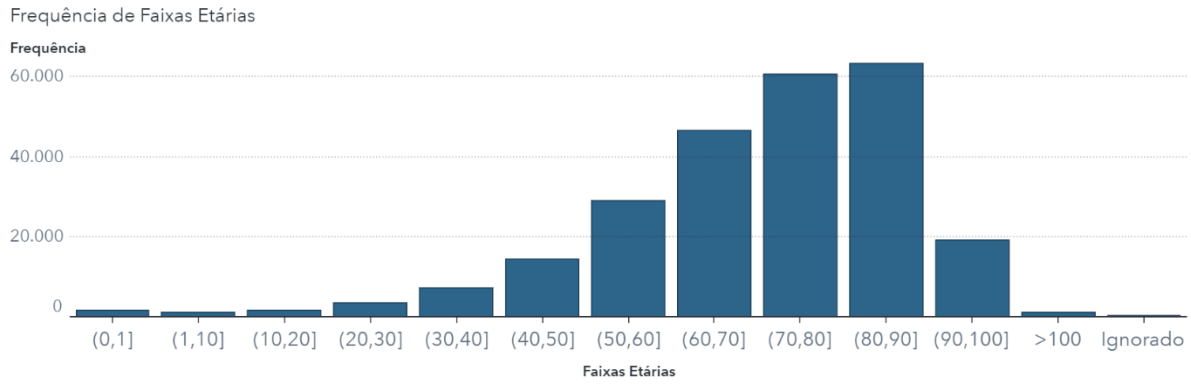
Tabela 1 - Distribuição de Frequência por Município de Residência 2014 a 2018

Município	Nº de Óbitos	Estimativa 2019 Pop	Taxa
Santos	2879	433311	0,00664
Mogi das Cruzes	2646	445842	0,00593
Mauá	2230	472912	0,00472
Bauru	2142	376818	0,00568

Fonte: Elaborado pela autora (2022).

Para facilitar a análise da variável idade, ela foi agrupada em treze categorias: (0,1], (1,10], (10,20], (20,30], (30,40], (40,50], (50,60], (60,70], (80,90], (90,100], >100 e ignorado. Ao analisar todos os indivíduos do estudo, foi constatado que há uma concentração de óbitos por septicemia em pessoas entre 60 e 90 anos, como pode ser verificado na Figura 2.

Figura 2 - Gráfico de Barras das Faixas Etárias de 2014 a 2018

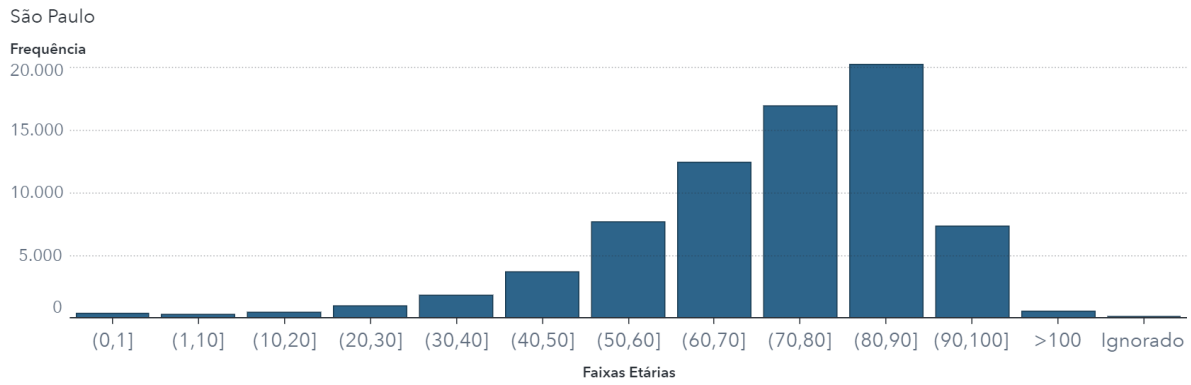


Fonte: Elaborado pela autora (2022).

Para melhor entender a influência da variável idade em relação a outras variáveis dependentes, foram feitos gráficos que pudessem mostrar tendências.

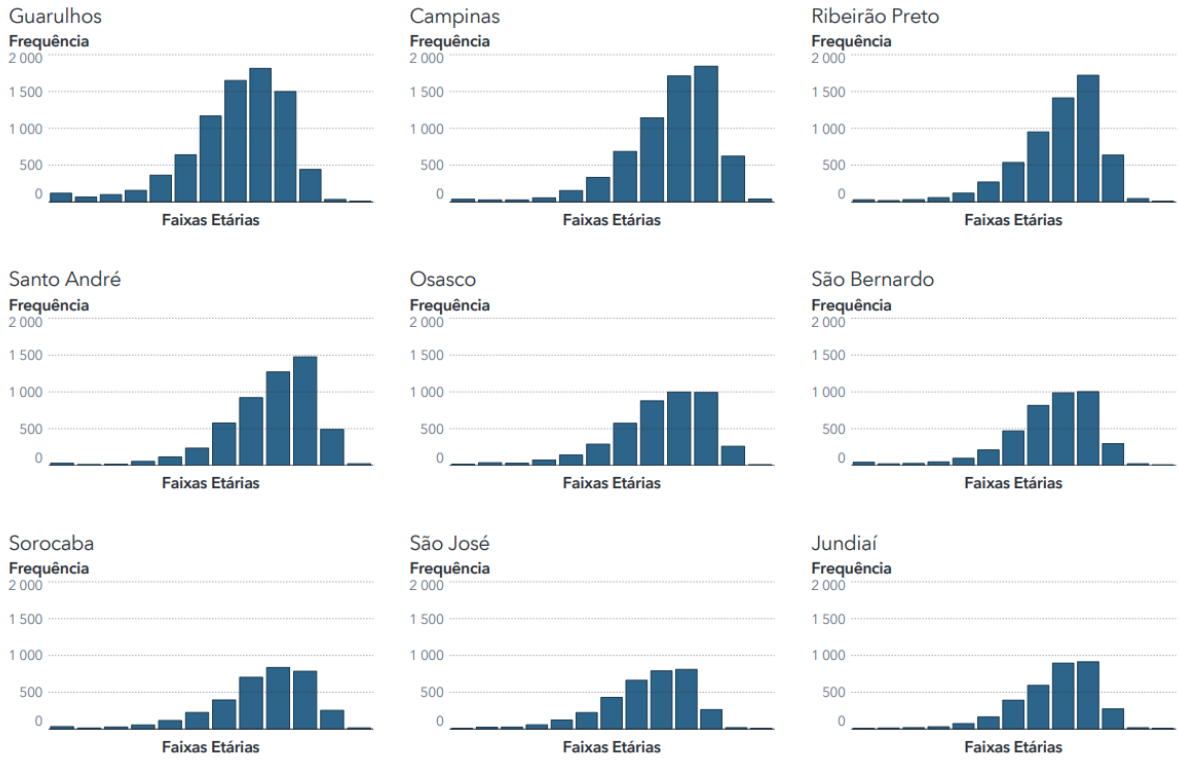
Analisada a idade por município de residência, apenas para os dez municípios com maiores frequências, é notado um comportamento semelhante entre eles. Estando os municípios ordenados na Tabela 1, sendo São Paulo o município com maior frequência e Jundiaí o 10º município de maior frequência.

Figura 3 - Gráfico de Barras das Faixas Etárias de 2014 a 2018 para São Paulo



Fonte: Elaborado pela autora (2022).

Figura 4 - Gráficos de Barras das Idades (Faixas Etárias) para os Municípios com Maior n.º de Óbitos

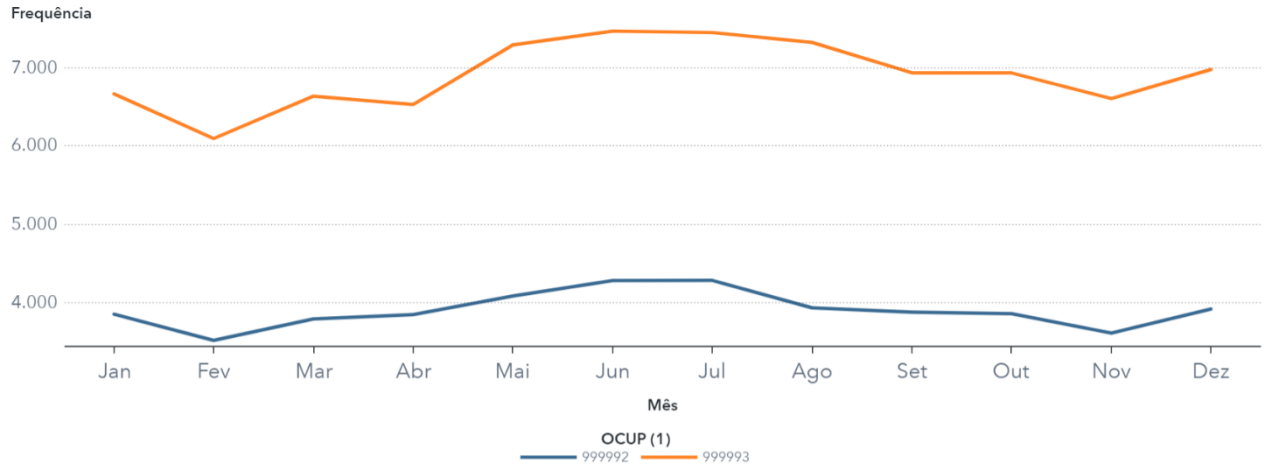


Fonte: Elaborado pela autora (2022).

Nota-se também que a medida que o número de óbitos do município diminui, a diferença entre as faixas etárias (70,80] e (80,90] também diminui.

A ocupação com maior frequência de óbitos foi a definida como Aposentado/Pensionista (código: 999992), concentrando 33,16% dos óbitos, já a ocupação com segunda maior frequência foi a definida como Dona de Casa (código: 999993), concentrando 18,76% dos óbitos. Nota-se a grande diferença entre a ocupação de maior frequência para a ocupação de segunda maior frequência, e é evidente que tais ocupações estão associadas às faixas etárias com maior concentração dos óbitos. Também é possível notar que o número de óbitos acumulado por mês, possui comportamento semelhante para ambas as ocupações citadas acima.

Figura 5 - Gráfico de Linhas para Óbitos por Ocupação por Mês de 2014 a 2018



Fonte: Elaborado pela autora (2022).

Complementando a informação da causa terminal do óbito, é informada também no atestado de óbito a causa antecedente, portanto, estão sendo analisadas as causas antecedentes considerando a causa terminal como septicemia não especificada.

Tabela 2 - Distribuição de Frequência para as principais Doenças Antecedentes

Nome da Doença	**Código CID10**	**Frequência**
Pneumonia por microrganismos não especificada	J18	97.542
Outros transtornos do trato urinário	N39	23.515
Outro	Outro	21.526

Nome da Doença	Código CID10	Frequência
Pneumonia bacteriana não especificada	J15	9.179
Pneumonia devida a sólidos e líquidos	J69	7.567
Peritonite	K65	7.379
Outras septicemias	A41	6.358
Insuficiência respiratória não classificada	J96	6.059
Outros transtornos respiratórios	J98	4.123
Dor abdominal e pélvica	R10	4.074

Fonte: Elaborado pela autora (2022).

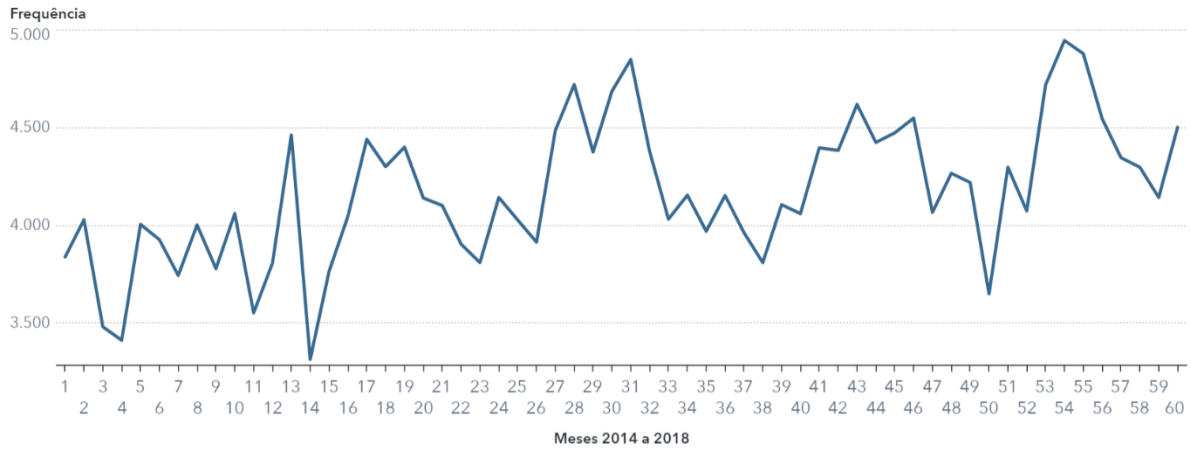
De uma maneira generalizada, é notável a predominância da causa antecedente estar relacionada ao sistema respiratório, tal como pneumonias e transtornos respiratórios. É claro, também, a grande diferença entre o número de óbitos da causa antecedente com maior frequência para o número de óbitos da causa antecedente com a segunda maior frequência. Tal diferença também notada para outras variáveis.

Com base no estudo gráfico realizado, para verificar o comportamento do número mensal de óbitos em relação a algumas variáveis, observa-se que o sexo,

cidade em que ocorreu o óbito, não está diretamente relacionada ao total de óbitos. Desta forma, pode ser uma indicação de que estas variáveis não trarão nenhuma informação adicional a ocorrência de óbitos por septicemia, portanto, neste estudo, cujo objetivo é estudar o número de casos diários, estas variáveis não serão levadas em consideração.

Assim, o prosseguimento do trabalho, inicialmente, será levado em consideração somente a contagem de óbitos por septicemia ao longo do tempo, tendo a seguir, na figura 6 a contagem do numero de óbitos mensais, totalizando 60 meses, e na figura 7 a contagem diária, ambos utilizando o período de 2014 a 2018 e ambos os gráficos foram elaborados no software SAS.

Figura 6 – Gráfico do número de Óbitos por septicemia por mês no período de 2014 a 2018.

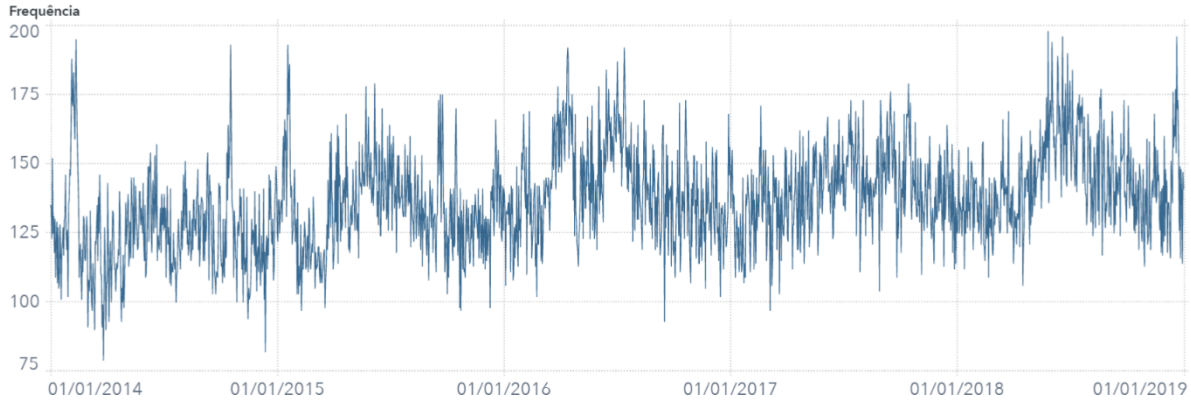


Fonte: Elaborado pela autora (2022).

Como já citado anteriormente, existem casos de superdispersão e subdispersão em modelo para dados de contagem. No caso apresentado na Figura 6, contagem de óbitos por septicemia por mês, foi verificado que a variância da variável contagem de óbitos por septicemia é muito maior que a média, sendo assim, um caso de superdisperção.

A característica de superdispersão também foi notada ao analisar a contagem de óbitos por septicemia por dia.

Figura 7 - Gráfico do número de Óbitos por septicemia por dia no período de 2014 a 2018



Fonte: Elaborado pela autora (2022).

É esperado que ao longo do tempo o número de óbitos aumente, decorrência do aumento do tamanho da população, porém, é possível notar um comportamento anômalo em alguns dias observados. Por exemplo em 11/02/2014 a contagem de óbitos foi igual a 195, número próximo da contagem de óbitos de 27/05/2018 que foi igual a 198.

Quando ordenadas as frequências de óbitos, a situação citada acima é observada de maneira mais clara, delimitando um corte na contagem de óbitos igual a 190, aproximadamente 0,5% dos dias observados possuem contagem maior ou igual a 190, e são observadas datas de todos os anos analisados, com predominância de datas em 2018.

Tabela 3 – Contagem de Óbitos por Dia para Dias com Mais de 189 Óbitos

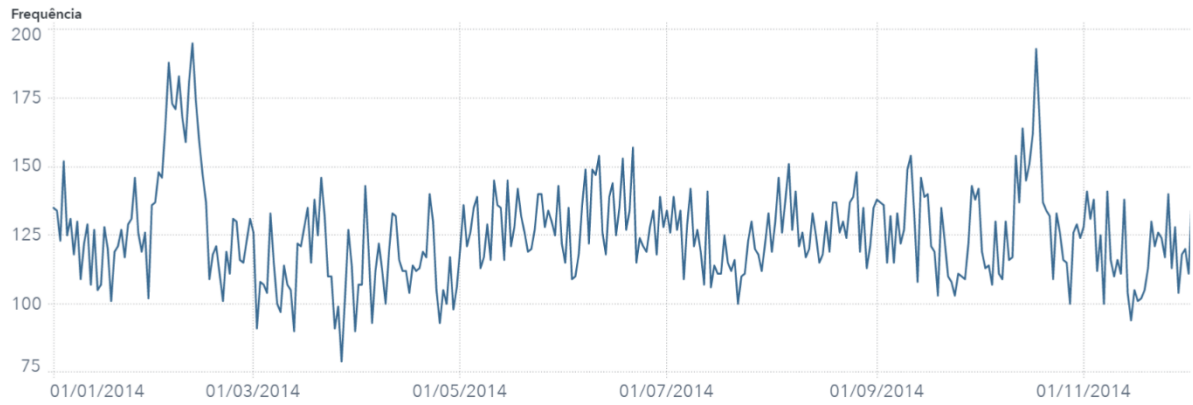
Data	**Contagem de Óbitos**
27/05/2018	198
19/06/2018	196
20/12/2018	196
11/02/2014	195

Data	Contagem de Óbitos
02/06/2018	194
18/10/2014	193
18/01/2015	193
13/04/2016	192
13/07/2016	192
27/06/2018	190

Fonte: Elaborado pela autora (2022).

Para melhor visualização do comportamento da variável, foram feitos gráficos para cada um dos anos, o que facilita visualizar a existência de sazonalidade ou alguma tendência específica, valores atípicos naquele ano, e também, visualizar se o comportamento entre os anos se assemelha.

Figura 8 - Gráfico da Contagem de Óbitos por Dia em 2014

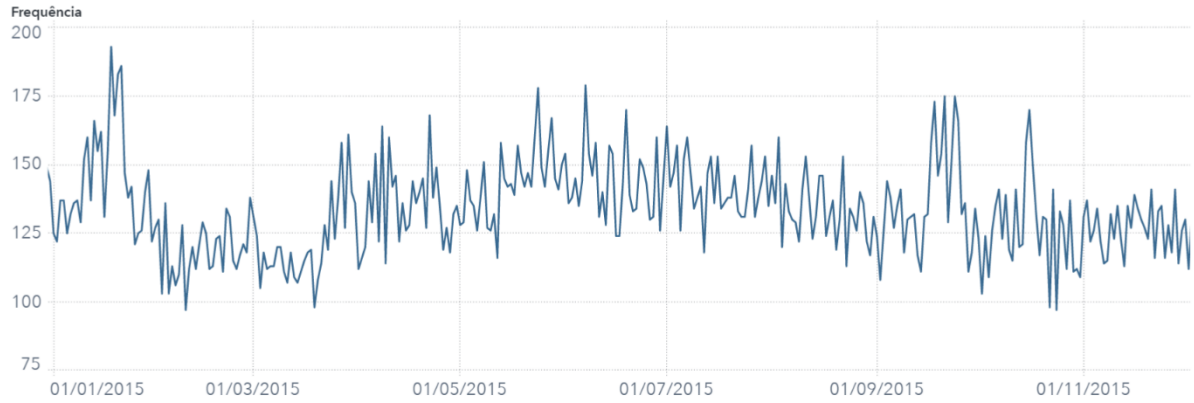


Fonte: Elaborado pela autora (2022).

Nota-se um aumento abrupto da contagem de óbitos do último dia de janeiro que decai na metade de fevereiro, mais especificamente, de 29/01/2014 a 16/02/2014, com a maior contagem sendo de 195 óbitos no dia. Valor próximo à contagem de óbitos do começo de outubro, com a maior contagem sendo de 193 óbitos no dia,

período no qual também foi observado um aumento em 11/10/2014 e decaimento em 23/10/2014.

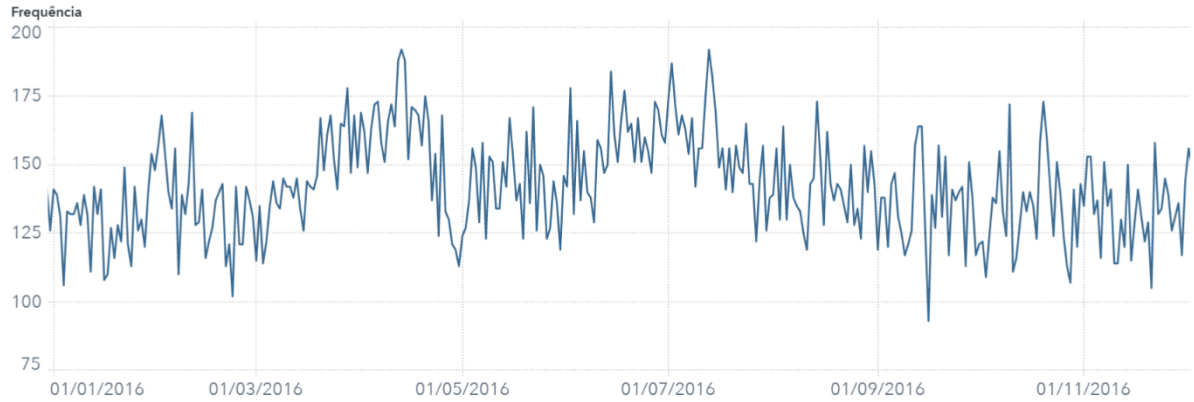
Figura 9 - Gráfico da Contagem de Óbitos por Dia em 2015



Fonte: Elaborado pela autora (2022).

Se assemelhando à 2014, nota-se também um aumento da contagem de óbitos no começo do ano, 16/01/2015 a 25/01/2015. Há um aumento gradativo do número de óbitos ao longo do ano, com aumento em setembro e outubro.

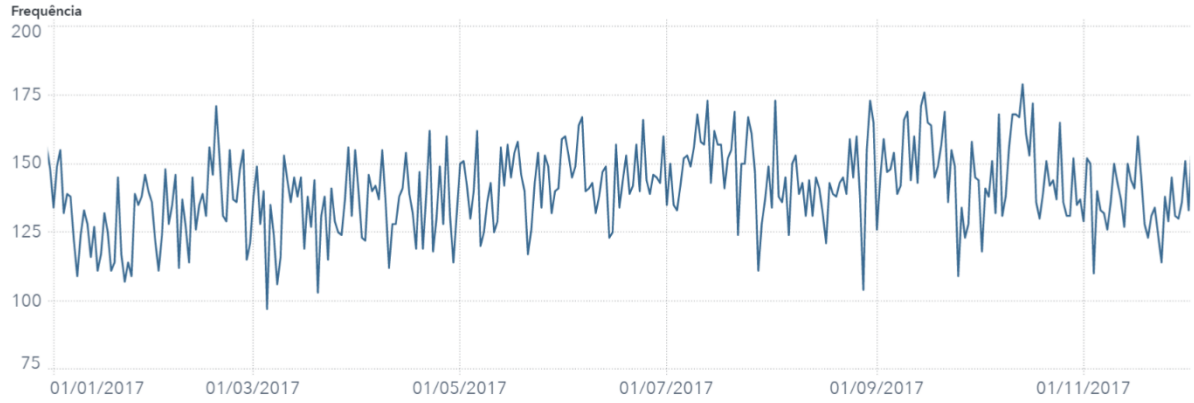
Figura 10 - Gráfico da Contagem de Óbitos por Dia em 2016



Fonte: Elaborado pela autora (2022).

Com picos em diferentes períodos do que observado nos anos anteriores, há um aumento do número de óbitos, de maneira geral, no meio do ano, com maior enfoque em março e abril.

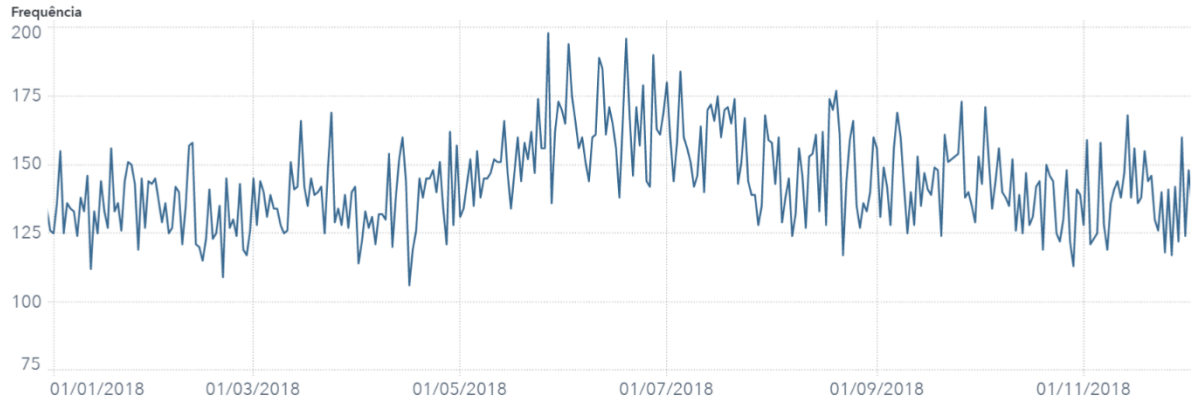
Figura 11 - Gráfico da Contagem de Óbitos por Dia em 2017



Fonte: Elaborado pela autora (2022).

Diferente dos anos anteriores, em 2017 os aumentos e diminuições na contagem demonstram um comportamento ameno ao longo do ano.

Figura 12 - Gráfico da Contagem de Óbitos por Dia em 2018



Fonte: Elaborado pela autora (2022).

É notável o aumento da contagem de óbitos no meio do ano, esse pico se dá entre 18/05/2018 a 30/07/2018.

3.2 MODELAGEM

Para encontrar o modelo que melhor descreve o comportamento de óbitos por septicemia no estado de São Paulo por dia, alguns testes serão feitos, levando em consideração que a única covariável a ser utilizada será o tempo.

Devido ao comportamento gráfico observado na Figura 7, entende-se que o modelo de séries temporais poderia ser aplicado, no entanto, neste projeto, o objetivo é testar algumas funções de ligação utilizando o modelo de regressão de Poisson. Uma tentativa, será a utilização do ajuste da média dada por 3.1, baseada em um ajuste de um modelo de probabilidade apresentado em Rodrigues, Tarumoto e Tzintzun (2019).

$$\ln(\mu_{t_i}) = a\cos(bt_i) + c\sin(dt_i) + e \tag{3.1}$$

$$\mu_{t_i} = e^{a\cos(bt_i) + c\sin(dt_i) + e}$$

assim,

$$P(Y_i = y_i) = e^{-e^{a\cos(bt_i) + c\sin(dt_i) + e}} \frac{\left(e^{a\cos(bt_i) + c\sin(dt_i) + e}\right)^{y_i}}{y_i!}.$$

Para encontrar as estimativas dos estimadores dos parâmetros do modelo, como já mencionado anteriormente, será utilizado o método de máxima verossimilhança. Observe que como um primeiro ensaio, não está sendo levado em consideração a possível dependência entre as observações e a superdispersão.

$$L(a, b, c, d \mid T, Y) = \prod_{i=1}^{n} P(Y_i = y_i)$$

$$L(a, b, c, d \mid T, Y) = \prod_{i=1}^{n} e^{-e^{a\cos(bt_i) + c\sin(dt_i) + e}} \frac{\left(e^{a\cos(bt_i) + c\sin(dt_i) + e}\right)^{y_i}}{y_i!}$$

$$L(a,b,c,d \mid T,Y) = e^{-\sum_{i=1}^{n}(\mathrm{a}\cos(bt_i) - c\sin(dt_i)+e)} \frac{e^{\sum_{i=1}^{n}((a\cos(bt_i) + c\sin(dt_i)+e)^{y_i})}}{\prod_{i=1}^{n} y_i!}$$

$$\ell(a,b,c,d \mid T,Y) = \ln L(a,b,c,d \mid T,Y)$$

$$\ell(a,b,c,d \mid T,Y) = -\sum_{i=1}^{n} e^{a\cos(bt_i) + c\sin(dt_i)+e} + \sum_{i=1}^{n}(a\cos(bt_i) + c\sin(dt_i) + e)^{y_i} - \sum_{i=1}^{n} y_i! \quad (3.2)$$

É necessário que a equação acima seja derivada em relação aos parâmetros e igualada a zero para que as estimativas dos parâmetros do modelo sejam encontradas, porém, como a função encontrada é não linear, serão utilizados métodos computacionais para maximizá-la.

No software R Studio, a função “optim” do pacote “stats”, após a definição da função que será maximizada e parâmetros iniciais, retorna estimativas ótimas para esses parâmetros. Essa função está pré-definida para fazer uma minimização, sendo assim, é definido o negativo do logaritmo da função de verossimilhança, já que, ao minimizar o inverso de uma função, ela é maximizada.

A obtenção das estimativas para o modelo construído dessa forma é instável, já que a convergência das estimativas depende dos parâmetros iniciais. Essa instabilidade possivelmente ocorre devido aos vários máximos locais.

Em uma tentativa de encontrar um resultado, foram geradas cinco mil combinações de valores para os parâmetros iniciais a partir da distribuição uniforme, e a partir desses parâmetros, utilizando o comando “optim”, foram encontradas as melhores combinações de estimativas, seus erros padrão e o resultado do negativo do logaritmo da função de verossimilhança.

Das cinco mil combinações geradas, apenas as cinco melhores serão mostradas aqui, sendo elas denominadas Ajuste 1 a Ajuste 5. É possível observar na Tabela 4 os resultados dos ajustes.

Tabela 4 - Estimativas e Erros Padrão para os Parâmetros dos Ajustes

Parâmetros/Ajustes	Ajuste 1		Ajuste 2		Ajuste 3	
	Estimativa	Erro Padrão	Estimativa	Erro Padrão	Estimativa	Erro Padrão
Parâmetro A	-0,06374	0,00278	-0,06203	0,00278	-0,06092	0,00278
Parâmetro B	0,00210	0,00006	-12,56426	0,00006	-6,28526	0,00006
Parâmetro C	0,00937	0,00286	0,01302	0,00289	-0,00109	0,00877
Parâmetro D	-6,57902	0,00049	-6,49057	0,00029	-5,53598	0,00454
Parâmetro E	4,90668	0,00235	4,90522	0,00232	4,90915	0,00238
Função Verossimilhança	8067,6694		8072,6667		8073,5957	

Fonte: Elaborado pela autora (2022).

Tabela 4 - Estimativas e Erros Padrão para os Parâmetros dos Ajustes

(continuação)

Parâmetros/Ajustes	Ajuste 4		Ajuste 5	
	Estimativa	Erro Padrão	Estimativa	Erro Padrão
Parâmetro A	-0,06485	0,00279	-0,01161	0,00284
Parâmetro B	-6,28525	0,00006	-8,70407	0,00045
Parâmetro C	-0,00117	0,00720	0,18142	0,00688
Parâmetro D	-15,04211	0,00560	12,56747	0,00002
Parâmetro E	4,90881	0,00242	4,78159	0,00551
Função Verossimilhança	8073,9359		8077,2218	

Fonte: Elaborado pela autora (2022).

Os parâmetros “b”, “c”, e “d” apresentam maior heterogeneidade em suas estimativas, evidenciando que os ajustes não foram tão satisfatórios, as estimativas dos parâmetros “a” e “e” são mais homogêneas, o que é uma evidência de que há maior confiança estatística nessas estimativas. O resultado da função de verossimilhança se assemelha muito para os cinco ajustes demonstrados, e, devido ao tamanho da amostra, os erros padrão possuem valores pequenos.

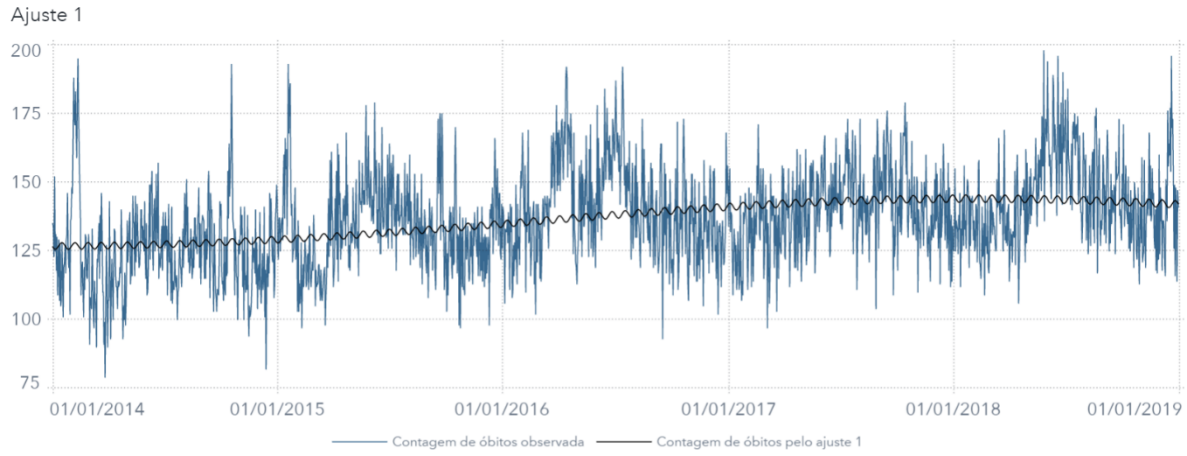
O esperado é que, ao chutar valores iniciais, as estimativas convirjam para a mesma região, porém, fica evidente que, diferentes chutes iniciais estão resultando em estimativas convergindo para regiões muito diferentes, fato que pode ser

observado a partir dos parâmetros “b”, “c”, e “d” e a convergência esperada pode ser observada a partir dos parâmetros “a” e “e”.

A interpretação do parâmetro “e”, que entre os parâmetros apresentou maior homogeneidade de estimativas para cada ajuste, pode ser feita a partir da exponencial da estimativa. Então, aproximadamente tem-se que $e^{4,90} = 134$, valor próximo à média encontrada de 137 óbitos por dia.

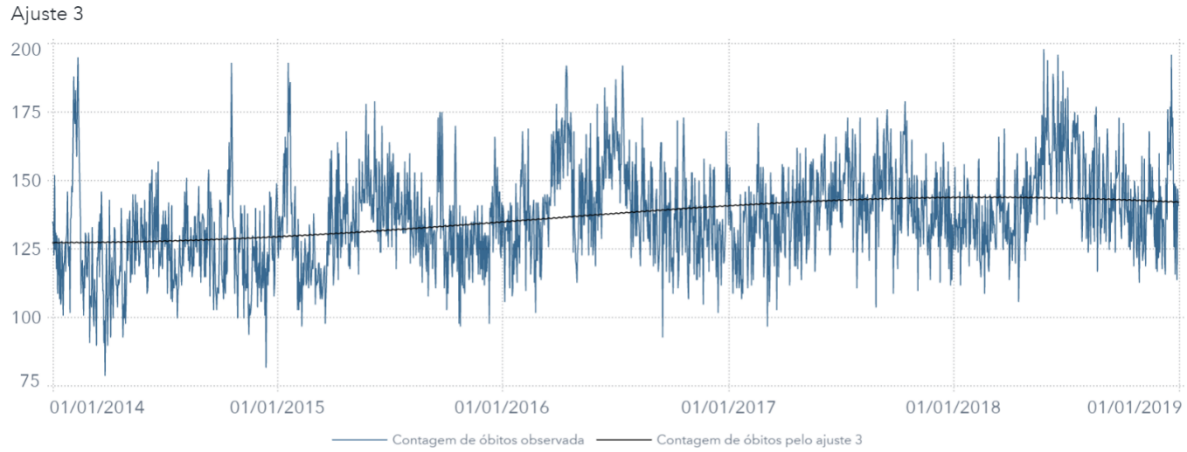
A análise gráfica dos ajustes demonstra as inconsistências analisadas a partir das estimativas dos parâmetros.

Figura 13 – Gráfico da Contagem de Óbitos por Dia Observada e Ajustada (Ajuste 1)



Fonte: Elaborado pela autora (2022).

Figura 14 - Gráfico da Contagem de Óbitos por Dia Observada e Ajustada (Ajuste 3)



Fonte: Elaborado pela autora (2022).

Os modelos ajustados se adequam à tendência temporal observada e estimam valores próximos a média, corroborando com a interpretação do parâmetro “e”, porém, não têm um bom desempenho em estimar os valores mais extremos. Devido a uma maior semelhança no comportamento dos ajustes, apenas o ajuste 1 e o ajuste 3, constam nesse capítulo, os demais gráficos constam no apêndice A.

De forma geral, existem muitos indícios de que os ajustes gerados não foram satisfatórios, isso pode ter se dado devido a não adequação da função de ligação ou a necessidade de se melhorar a técnica de estimação dos parâmetros.

3.3 ANÁLISE DE RESÍDUOS

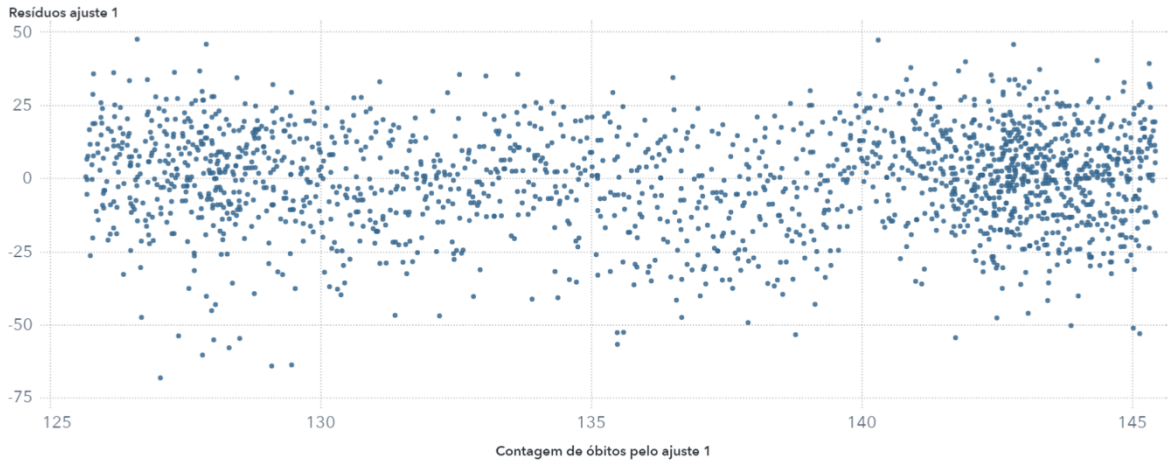
A fim de verificar a qualidade dos cinco melhores ajustes, foram calculados os resíduos ordinários, e para interpretá-los, foram construídos os gráficos dos resíduos versus valores ajustados, a interpretação desse tipo de gráfico se dá pela análise da homogeneidade dos resíduos e de sua amplitude.

Assim como no tópico de modelagem, constam aqui, apenas a análise dos resíduos dos ajustes 1 e ajuste 3, devido à semelhança de interpretabilidade, os gráficos dos resíduos dos ajustes 2, 4 e 5 constam no apêndice.

A partir da análise dos resíduos gerados pelo ajuste 1, Figura 15, conclui-se que os valores estão em torno de zero e com amplitude entre -50 e 50, o que,

considerando os dados analisados, é uma grande variação já que em média ocorrem 137 óbitos por septicemia por dia, sendo assim, ao estimar 50 óbitos a mais ou a menos, produz um erro grande.

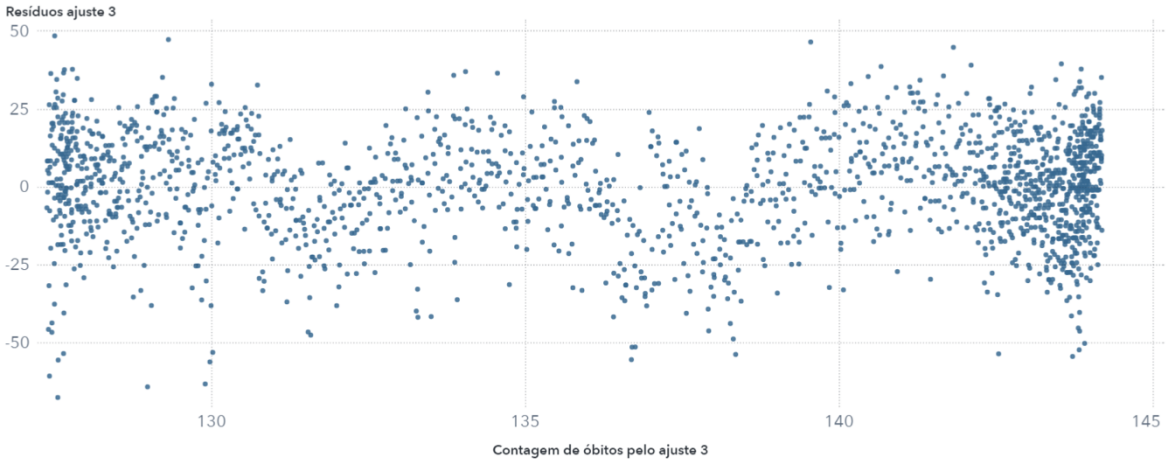
Figura 15 - Gráfico da Análise de Resíduos do Ajuste 1



Fonte: Elaborado pela autora (2022).

A partir da análise dos resíduos gerados pelo ajuste 3, Figura 16, conclui-se que os valores não estão em torno de zero e demonstram uma tendência, sua amplitude entre -50 e 50, assim como na análise de resíduos dos demais ajustes.

Figura 16 - Gráfico da Análise de Resíduos do Ajuste 3



Fonte: Elaborado pela autora (2022).

Assim como evidenciado na análise das estimativas dos parâmetros e seus erros padrão, a análise de resíduos reforça o mau ajuste dos modelos gerados.

4 CONSIDERAÇÕES FINAIS

As possibilidades de análises para dados de contagem são inúmeras, em casos particulares como a superdispersão, exige maior especificidade na escolha da análise.

Os modelos ajustados aos dados deste trabalho de forma geral trouxeram resultados semelhantes, não houve uma boa estimação para valores extremos e as estimativas se concentraram próximas à média.

De forma primitiva, estimar um valor a partir da média observada é o melhor chute quando não se tem recursos mais acurados, porém, a construção de um modelo se faz necessária quando acurácia e significância estatística são desejadas, assim, os ajustes construídos não foram satisfatórios.

Existem alguns fatores que podem ter interferido no resultado dos ajustes, como a não adequação da função de ligação, a necessidade de se utilizar outra técnica de estimação dos parâmetros, ou então a aplicação de outro modelo estatístico, como a utilização da técnica das médias móveis.

Desta forma, existe a perspectiva para trabalhos futuros, este trabalho abre perspectiva para trabalhos futuros, entre elas, realizar um estudo minucioso da função de ligação, incorporação de dependência entre as observações e a superdispersão. Uma possibilidade adicional é considerar os modelos autorregressivos.

O que foi possível observar a partir deste trabalho é o comportamento sazonal dos óbitos por septicemia, o fato da doença atingir de forma equilibrada ambos os sexos, não se concentrar em nenhum município do estado de São Paulo, estando presente de forma proporcional ao tamanho populacional de cada um, e de fato a variável que maior influência no comportamento do número de óbitos é a idade, sendo possível constatar uma maior concentração de óbitos de pessoas acima de 60 anos.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, Alan. (2007). **An Introduction to Categorical Data Analysis**. Disponível em: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf. Acesso em: 02 jul. 2022.

BATISTA, Douglas Toledo. **Modelos para dados de contagem com superdispersão: uma aplicação em um experimento agronômico**. Dissertação (Mestrado) – Curso de Estatística, Universidade de São Paulo Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2015.

Consulta CID 10: **Classificação Internacional de Doenças**. Disponível em: https:// cid10.com.br. Acesso em: 10 jan. 2022

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

CORDEIRO, G. M.; DEMÉTRIO, C.C.B. **Modelos Lineares Generalizados e Extensões**, 2013. Disponível em: https://docs.ufpr.br/~taconeli/CE22518/ LivClarice.pdf. Acesso em: 19 jan. 2022.

CORDEIRO, G. M. **Modelos lineares generalizados**. Recife: UFPE, 1986. 286 p.

DOBSON, A. J**. An introduction to generalized linear models**. 2nd.,Ed., Boca Raton, Chapman Hall, 2010.

FREITAS, Keilla. **Infecção Generalizada Sepse: Causas, Sintomas e Tratamento**, 2013. Disponível em: https://www.drakeillafreitas.com.br/sepse-quando-espera-pode-matar/. Acesso em: 10 dez. 2022.

KUHN L.;DAVIDSON L. L.; DURKIN S. M. (1994). **Use of Poisson Regression and Time Series Analysis for Detecting Changes over Time in Rates of Child Injury following a Prevention Program**. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.890.1237&rep=rep1&type=pdf. Acesso em: 19 jan. 2022.

LIMA, Marcos Alves. **Estudo de caso em um veículo publicitário com dados de contagem longitudinais.** Monografia – Curso de Estatística, Universidade Federal de Juiz de Fora, Juiz de Fora, 2014.

LINDSEY, J. K. (2007). **Applying Generalized Linear Models.Springer**. Disponível em: http://www.leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:lindsey-glm.pdf. Acesso em: 19 jan. 2022.

MCCULLAGH, P. AND NELDER, J.A. **Generalized Linear Models**, second edition. 1989, London: Chapman and Hall.

NELDER, J. AND WEDDERBURN, R. (1972). **Generalized Linear Models**. Journal of the Royal Statistical Society, 135, 370-384. Disponível em: https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf. Acesso em: 20 jun. 2022.

OLSSON, U. (2002). **Generalized Linear Models - An Applied Approach**. Lund: Studentlitteratur.

PAULA, G. A. **Modelos de Regressão com apoio computacional**, 2013. Disponível em: http://www.ime.usp.br/~giapaula/texto-2013.pdf. Acesso em: 20 jun. 2022.

Portal da saúde SUS. Disponível em: https://datasus.saude.gov.br. Acesso em: 14 dez. 2021.

RODRIGUES, ELIANE R.; TARUMOTO, MARIO H.; TZINTZUN, GUADALUPE. **Application of a non-homogeneous Markov chain with seasonal transition probabilities to ozone data**. Journal Of Applied Statistics. Abingdon: Taylor & Francis Ltd, v. 46, n. 3, p. 395-415, 2019.

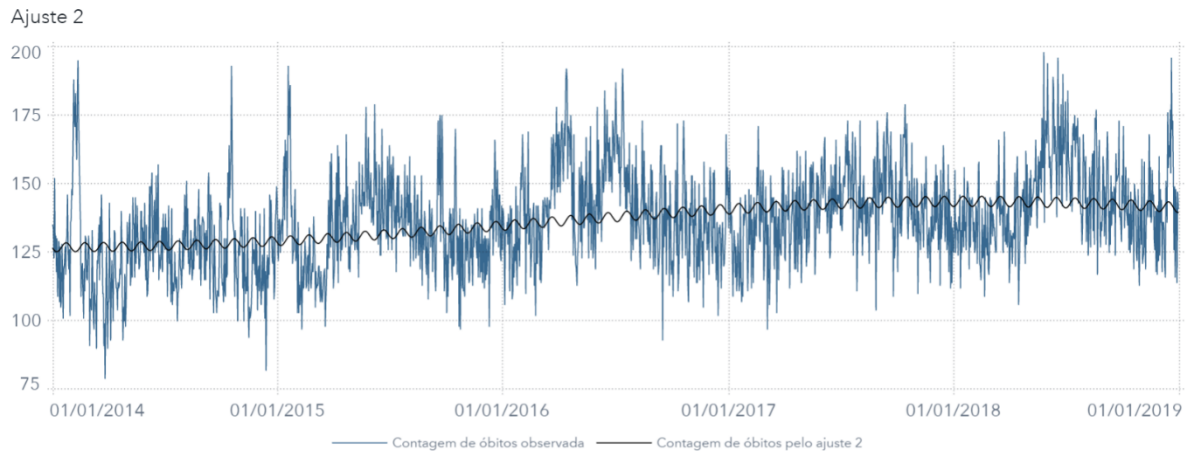
TRINDADE, Daniele de Brito. **Modelagem para dados longitudinais de contagem**. Dissertação (Mestrado) – Curso de Estatística, Universidade Federal de Pernambuco, Recife, 2014.

VARELLA, Maria Helena. **Sepse**, 2014. Disponivel em: https://drauziovarella.uol.com.br/doencas-e-sintomas/sepse-septicemia/. Acesso em: 10 dez. 2021.

APÊNDICE A – MODELAGEM

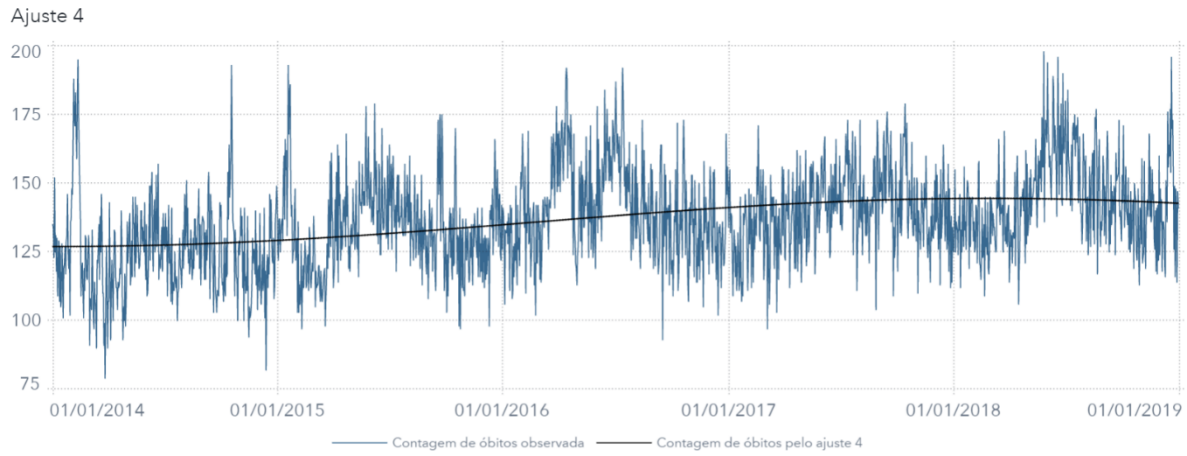
Devido à semelhança entre a interpretabilidade, neste apêndice constam os gráficos dos ajustes 2, 4 e 5.

Figura 17 - Gráfico da Contagem de Óbitos por Dia Observada e Ajustada (Ajuste 2)



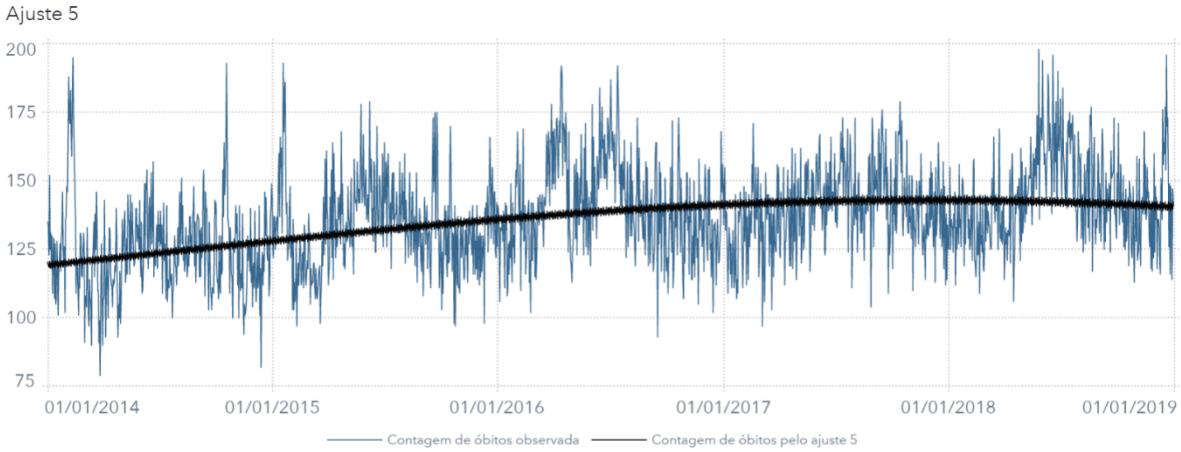
Fonte: Elaborado pela autora (2022).

Figura 18 - Gráfico da Contagem de Óbitos por Dia Observada e Ajustada (Ajuste 4)



Fonte: Elaborado pela autora (2022).

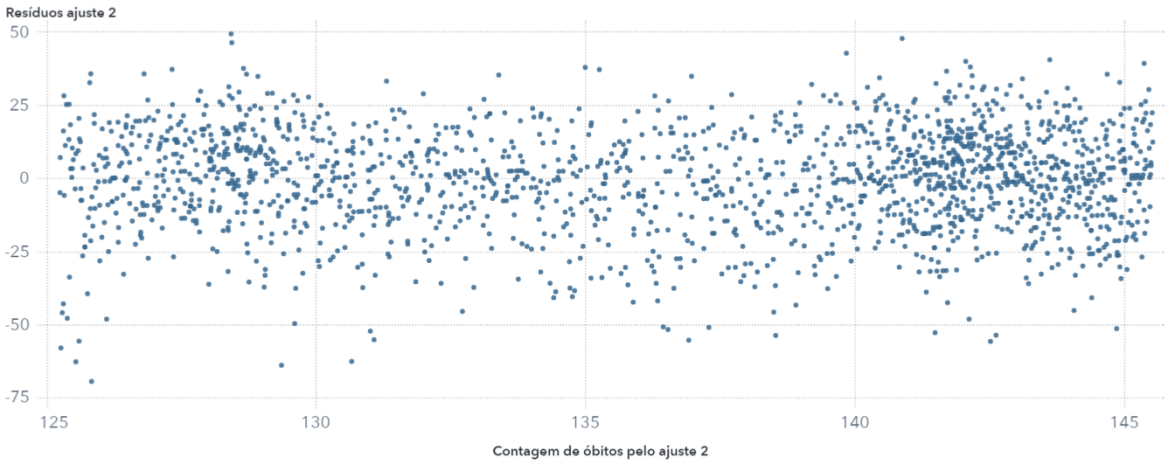
Figura 19 - Gráfico da Contagem de Óbitos por Dia Observada e Ajustada (Ajuste 5)



Fonte: Elaborado pela autora (2022).

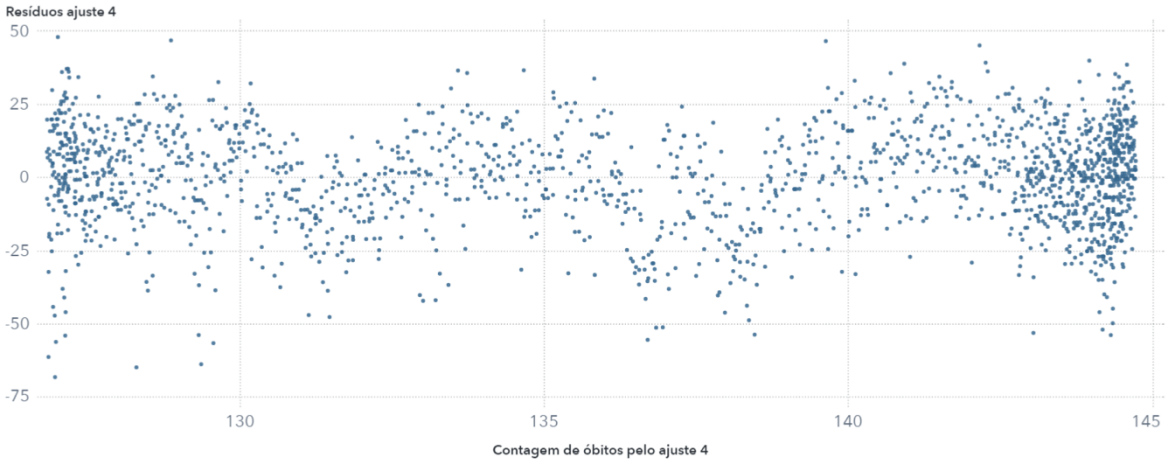
APÊNDICE B – ANÁLISE DE RESÍDUOS

Figura 20 - Gráfico da Análise de Resíduos do Ajuste 2



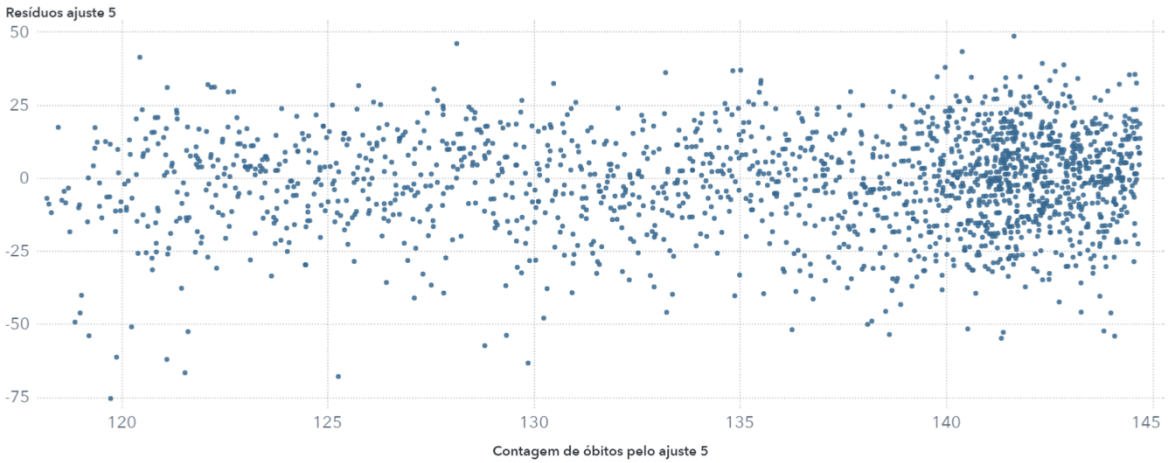
Fonte: Elaborado pela autora (2022).

Figura 21 - Gráfico da Análise de Resíduos do Ajuste 4



Fonte: Elaborado pela autora (2022).

Figura 22 - Gráfico da Análise de Resíduos do Ajuste 5



Fonte: Elaborado pela autora (2022).