UNIVERSIDADE FEDERAL DE ALFENAS

DAIANE DE OLIVEIRA GONÇALVES

MODELOS PARAMÉTRICOS DE SOBREVIVÊNCIA APLICADOS A DADOS DE CÂNCER

Alfenas/MG

2020

DAIANE DE OLIVEIRA GONÇALVES

MODELOS PARAMÉTRICOS DE SOBREVIVÊNCIA APLICADOS A DADOS DE CÂNCER

Dissertação apresentada ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, área de concentração em Estatística Aplicada e Biometria da Universidade Federal de Alfenas-MG, como parte dos requisitos para a obtenção do título de Mestre.
Linha de Pesquisa: Modelagem Estatística e Estatística Computacional.
Orientadora: Prof. Dra. Natália da Silva Martins Fonseca.
Coorientador: Prof. Dr. Fabricio Goecking Avelar.

Alfenas/MG

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal de Alfenas

Gonçalves, Daiane de Oliveira.
G635m Modelos paramétricos de sobrevivência aplicados a dados de câncer / Daiane de Oliveira Gonçalves -- Alfenas/MG, 2020.
75 f. : il. --

Orientadora: Natália da Silva Martins Fonseca.
Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, 2020.
Bibliografia.

1. Distribuição (Teoria da probabilidade). 2. Pulmões (Câncer). 3. Taxa de Sobrevida. I. Fonseca, Natália da Silva Martins. II. Título.

CDD-519.5

Ficha Catalográfica elaborada por Fátima dos Reis Goiatá
Bibliotecária-Documentalista CRB6/425

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Alfenas / UNIFAL-MG
Programa de Pós-graduação em Estatística Aplicada e Biometria
Rua Gabriel Monteiro da Silva, 700. Alfenas - MG CEP 37130-001
Fone: (35) 3701-9600
https://www.unifal-mg.edu.br/ppgeab/

DAIANE DE OLIVEIRA GONÇALVES

"MODELOS PARAMÉTRICOS DE SOBREVIVÊNCIA APLICADOS A DADOS DE CÂNCER"

A Banca Examinadora, abaixo assinada, aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Área de Concentração: Estatística Aplicada e Biometria

Aprovado em: 06 de fevereiro de 2020.

Profa. Dra. Natália da Silva Martins Fonseca Assinatura: _____
Instituição: UNIFAL-MG

Prof. Dr. Filidor Edilfonso Vilca Labra Assinatura: _____
Instituição: UNICAMP

Prof. Dr. Marcelo Ângelo Cirillo Assinatura: _____
Instituição: UFLA

Dedico esse estudo a Deus e Nossa Senhora Aparecida, ao meu pai Sebastião e a minha mãe Maria.

AGRADECIMENTOS

Agradeço a Deus e a Nossa Senhora Aparecida, por estarem sempre comigo, iluminando-me e guiando meus passos.

Agradeço aos meus pais Maria e Sebastião, pelo amor, carinho e apoio em todos os momentos.

À minha irmã Bruna, pela amizade, conselhos e incontáveis ajudas.

Ao meu namorado Gabriel, por seu amor, paciência e compreensão.

Aos meus avós que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa, colocando-me sempre em suas orações. Em especial meu agradecimento com saudades do vovô Oliveiros, que agora está com Deus, por todo o amor, confiança e ensinamentos.

Agradeço à minha querida orientadora, Natália da Silva Martins Fonseca, por toda a sua dedicação, apoio, confiança e ensinamentos. Gostaria de expressar toda a minha gratidão e admiração por sua competência e amizade.

Ao meu coorientador, professor Fabricio Goecking Avelar, por seu apoio e amizade, além de sua ajuda e dedicação.

Agradeço à Universidade Federal de Alfenas e aos Professores e funcionários do Instituto de Ciências Exatas, pela atenção e contribuição para a minha formação.

Aos membros das bancas do exame de qualificação e da banca de defesa de mestrado, professores Filidor Edilfonso Vilca Labra e Marcelo Angelo Cirillo, pela participação na banca, sugestões e contribuições para o desenvolvimento deste estudo.

Agradeço aos meus colegas de turma Iasmine, Alice, Lara, Claudiana, Saditt, Frank, Nalva e Valdeline pela amizade e companheirismo.

Agradeço a todos que, de alguma forma, contribuíram para a realização desse estudo.

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

RESUMO

O câncer tem se tornado uma das doenças responsáveis por uma grande quantidade de óbitos em todo o mundo. Pois, de acordo com a Organização Mundial da Saúde essa patologia é uma das doenças malignas mais comuns no mundo. Deste modo, seria de interesse avaliar os fatores capazes de influenciar no tempo até o óbito de pacientes portadores dessa doença. Para tanto, esse estudo tem por objetivo a apresentação de uma revisão sobre os principais modelos paramétricos de sobrevivência utilizados para descrever o tempo de vida de pacientes com essa patologia bem como os fatores a ele associados. E, a partir desses modelos realizar um ajuste dos mesmos a um conjunto de dados provenientes de pacientes com câncer de pulmão. Os modelos paramétricos ajustados foram Weibull, Gama, Gama Generalizado e Log-normal, sendo o melhor modelo selecionado por meio dos critérios de informação AIC, SBC e GD. Após a seleção do melhor modelo, nesse estudo o Gama Generalizado, realizou-se a análise residual e avaliou-se a acurácia do modelo. Verificou-se que o modelo Gama Generalizado foi o que melhor se ajustou ao conjunto de dados, sendo assim utilizado para descrever o tempo até a ocorrência do óbito dos pacientes portadores de câncer de pulmão e os fatores a ele associados. E, concluiu-se que os fatores que contribuíram para tempo de sobrevida dos pacientes foram gênero, idade, estadiamento (TNM), histologia e cigarro.

Palavras-chaves: Distribuições (Teoria da Probabilidade); Pulmões (Câncer); Taxa de sobrevida.

ABSTRACT

Cancer has become one of the diseases responsible for a large number of deaths worldwide. According to the World Health Organization, this pathology is one of the most common malignant diseases in the world. Thus, it would be of interest to evaluate the factors capable of influencing the time until the death of patients with this disease. Therefore, this study aims to present a review of the main parametric models of survival used to describe the life time of patients with this pathology as well as the factors associated with it. And, based on these models, fitted them to a set of data from patients with lung cancer. The adjusted parametric models were Weibull, Gamma, Generalized Gamma and Log-normal, with the best model being selected using the AIC, SBC and GD information criteria. After selecting the best model, in this study the Generalized Gamma, residual analysis was performed and the accuracy of the model was evaluated. It was found that the Generalized Gamma model was the one that best fitted the data set, being thus used to describe the time until the death of patients with lung cancer and the factors associated with it. And, it was concluded that the factors that contributed to the patients' survival time were gender, age, staging (TNM), histology and smoking.

Key words: Distributions (Probability Theory); Lungs (Cancer); Survival rate.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

$S(\cdot)$	- Função de sobrevivência.
$F(\cdot)$	- Função de distribuição acumulada.
$\lambda(\cdot)$	- Função taxa de falha.
$\Lambda(\cdot)$	- Função taxa de falha acumulada.
T	- Variável aleatória não negativa representando o tempo de falha.
$t_j, j = 1, \cdots, p$	- p tempos distintos observados e ordenados de falha.
$d_j, j = 1, \cdots, p$	- Número de falhas observadas em t_j.
$n_j, j = 1, \cdots, p$	- Número de indivíduos sob risco em t_j.
$\widehat{q}_j, j = 1, \cdots, p$	- Proporção de indivíduos vivos após t_{j-1} que sobrevivem além de t_j.
μ, σ, r	- Parâmetros das distribuições consideradas.
$E(T)$	- Esperança da variável aleatória T.
$Var(T)$	- Variância da variável aleatória T.
x	- Vetor de covariáveis.
β	- Vetor de parâmetros desconhecidos.
Γ	- Função Gama.
$L(\boldsymbol{\theta})$	- Função de verossimilhança para um vetor de parâmetros $\boldsymbol{\theta}$).
n	- Número de observações.
p	- Número de falhas.
δ_i	- Variável indicadora de falha ou censura.
C	- Tempo de censura.
$g(c)$	- Função densidade de C.
$G(c)$	- Função de sobrevivência de C.
k	- Número de parâmetros a serem estimados.

$\widehat{e_i}$ - Resíduos de Cox-Snell.

$\widehat{m}_i$ - Resíduos martingal.

$\widehat{d_i}$ - Resíduos deviance.

SUMÁRIO

1 INTRODUÇÃO

O câncer é uma patologia causada pela multiplicação desordenada das células, ocasionada por mutações nos genes, sendo estes responsáveis por codificar as proteínas reguladoras do ciclo celular. Esta patologia tem se tornado uma das doenças responsáveis por uma grande quantidade de óbitos em todo o mundo. Estima-se que o câncer causou a morte de 9,6 milhões de pessoas em todo o mundo durante o ano de 2018.

De acordo com a Organização Mundial da Saúde (OMS) o câncer de pulmão é uma das doenças malignas mais comuns. Estima-se que entre os novos casos de câncer, 13% sejam de pulmão. Esse tipo de câncer é a principal causa de mortalidade por câncer no Brasil, como ocorre na maioria dos países, segundo Instituto Nacional de Câncer José Alencar Gomes da Silva (2017).

Paralelamente às preocupações causadas pelo efeito desse tipo de doença, o estudo do comportamento das mesmas vêm se tornado cada vez mais importante, visto que através desses estudos é possível ter acesso a novas terapias, a melhores condições de vida e, consequentemente, proporcionar o aumento da sobrevida dos pacientes. A partir do momento que as causas que influenciam na sobrevida são identificadas, uma área muito utilizada nesse processo é a análise de sobrevivência.

Em termos estatísticos, o estudo relacionado ao tempo de vida de pacientes que manifestam esse câncer, bem como, a sua evolução, a classe de modelos de análise de sobrevivência se destaca para considerar os dados relacionados ao tempo de ocorrência de um determinado evento de interesse, o qual é denominado por tempo de falha. Uma característica elementar dos dados de sobrevivência é existência de observações censuradas, as quais podem ser entendidas como observações parciais da variável resposta, em outras palavras, são observações que por algum motivo não foram verificadas e/ou avaliadas.

Para que uma análise estatística resulte em conclusões fidedignas, é de interesse que todas as observações provenientes do estudo sejam utilizadas na análise. Nesse contexto, as técnicas de análise de sobrevivência permitem que essas observações incompletas (censuradas) também sejam consideradas. Pois carregam informações sobre o tempo até a ocorrência do evento de interesse do elemento amostral em estudo.

Na análise de sobrevivência, devido a necessidade da incorporação de observações cen-

suradas, várias técnicas e modelos foram desenvolvidos a fim de descrever essas informações, dentre os quais encontram-se técnicas não-paramétricas, modelos semi-paramétricos e paramétricos. Embora exista uma predileção dos pesquisadores das áreas médicas na utilização de modelos semi-paramétricos, principalmente pela maior simplicidade da interpretação dos seus parâmetros, em muitos casos os modelos paramétricos apresentam estimativas mais precisas, tendo em vista as boas propriedades que os estimadores apresentam, sendo assim mais indicado para descrever o comportamento da variável de interesse.

Tendo em vista a grande quantidade de estudos que utilizam apenas técnicas não-paramétricas ou modelos semi-paramétricos no estudo de dados de câncer. E, considerando que os modelos paramétricos descrevem melhor os fatores que influenciam no tempo de vida dos pacientes, tem-se a pretensão de realizar uma revisão bibliográfica dos modelos paramétricos, assim como utilizá-los na modelagem de um conjunto de dados de câncer de pulmão e avaliar a qualidade do modelo ajustado.

Assim, o objetivo deste estudo consiste na realização de uma aplicação dos principais modelos paramétricos utilizados para descrever dados de sobrevida de pacientes com câncer, com o propósito de selecionar o que melhor descreve o conjunto de dados; e por fim a análise de resíduo como uma análise complementar aos critérios de informação utilizados comumente na escolha do modelo.

2 REFERENCIAL TEÓRICO

Nesta seção é apresentado o referencial teórico utilizado como base para a execução deste estudo. Na subseção 2.1 é detalhado o conteúdo da aplicação. A subseção 2.2 apresenta alguns conceitos e resultados da análise de sobrevivência. Na subseção 2.3 é apresentado o método de estimação de Kaplan-Meier. A subseção 2.4 se refere às distribuições mais utilizadas na análise de sobrevivência. Na subseção 2.5 é apresentado o processo de estimação dos parâmetros dos modelos. Na seção 2.6 é apresentado o pacote GAMLSS. E por fim, a subseção 2.7 apresenta a avaliação da qualidade do modelo.

2.1 CÂNCER

O câncer é uma patologia causada por uma multiplicação desordenada das células, ocasionada por mutações nos genes, sendo estes responsáveis por codificar as proteínas reguladoras do ciclo celular. Tais mutações permitem que as células apresentem diferentes características e, por isso, são denominadas de células cancerígenas. Essas células apresentam capacidade de se espalharem para outros tecidos e/ou órgãos do corpo, iniciando-se o processo conhecido como metástase. Elas ainda apresentam capacidade de multiplicar-se mesmo na ausência de fatores ou sinais de proteínas que as estimulem, além de não se submeterem a apoptose (morte celular programada), (BERNARDES et al., 2019).

Bray et al. (2018), utilizando as estimativas GLOBOCAN 2018 de incidência e mortalidade por câncer produzidas pela Agência Internacional para Pesquisa sobre o Câncer, com foco na variabilidade geográfica em 20 regiões do mundo, estimaram a ocorrência de 18,1 milhões de novos casos de câncer (17 milhões excluindo câncer de pele não melanoma) e de 9,6 milhões de mortes por câncer (9,5 milhões excluindo câncer de pele não melanoma) em todo o mundo em 2018. Os autores afirmam que o câncer de pulmão é o mais diagnosticado (11,6% do total de casos) e a principal causa de morte por câncer (18,4% do total de mortes por câncer), seguido de perto pelo câncer de mama feminino (11,6%), câncer colorretal (10,2%) e câncer de próstata (7,1%) para incidência e câncer colorretal (9,2%), câncer de estômago (8,2%) e câncer de fígado (8,2%) para mortalidade.

Entre os homens, o câncer de pulmão é o mais diagnosticado e a principal causa de morte, seguido pelo câncer de próstata e colorretal por incidência, e câncer de fígado e estômago por mortalidade. Para as mulheres, o câncer de mama é o mais diagnosticado e a principal causa de

morte por câncer, seguido pelo câncer colorretal e de pulmão por incidência e pela mortalidade. O câncer do colo do útero ocupa o quarto lugar tanto na incidência quanto na mortalidade. No geral, os 10 principais tipos de câncer são responsáveis por mais de 65% dos casos e mortes por câncer recém-diagnosticados (BRAY et al., 2018).

De acordo com o Instituto Nacional de Câncer José Alencar Gomes da Silva (2017), a estimativa mundial revela que, em 2012, houve um predomínio do sexo masculino tanto na incidência (53%) quanto na mortalidade (57%). Para o Brasil, em 2018 e 2019, estimou-se a ocorrência de 600 mil casos novos de câncer, para cada ano.

De acordo com Santos (2018), as estimativas de novos casos de câncer para o Brasil se assemelham às de países desenvolvidos, porém ainda possui altas taxas de cânceres associados a infecções, que são característicos de países em desenvolvimento. O autor atribui essa realidade às desigualdades regionais, como as diferenças na expectativa de vida, condições socioeconômicas, e até mesmo o acesso aos serviços de saúde para diagnóstico oportuno e tratamento adequado.

Em relação a distribuição da incidência por região geográfica, tem-se que as regiões Sudeste e Sul, juntas, concentram 70% da incidência, cujo padrão dos tipos de cânceres são semelhantes aos países desenvolvidos, com predominância dos cânceres de mama feminina, próstata, pulmão, cólon e reto. A região Sul também é destacada pela alta incidência de câncer de pulmão, principalmente no Rio Grande do Sul. Na região Norte existe mais incidência de câncer do colo do útero e câncer de estômago, dessa forma possuindo semelhanças com países menos desenvolvidos. Também os Estados do Amazonas, Amapá e Maranhão assemelham-se aos países menos desenvolvidos, possuindo incidência de câncer do colo do útero.

2.1.1 Câncer de pulmão

De acordo com a OMS o câncer de pulmão está entre as doenças malignas mais comuns. Estima-se que entre os novos casos de câncer, 13% sejam de pulmão. Este câncer é o segundo tipo de câncer de maior incidência em homens e o quarto tipo de câncer de maior incidência em mulheres no país, segundo Instituto Nacional de Câncer José Alencar Gomes da Silva (2017).

Segundo Araujo et al. (2018), o câncer de pulmão é a principal causa de mortalidade por câncer no Brasil, característica semelhante com a maioria dos países.

De acordo com Instituto Nacional de Câncer José Alencar Gomes da Silva (2017), para o Brasil, estimam-se 18.740 casos novos de câncer de pulmão entre homens e de 12.530 nas

mulheres para cada um dos anos de 2018 e 2019. Desse modo, o risco estimado é de 18,16 novos casos a cada 100 mil homens, ocupando a segunda posição na lista de tumores mais frequentes. No caso das mulheres o risco estimado é de 11,81 para cada 100 mil, ficando com a quarta posição na lista de tumores mais frequentes.

Stewart e Wild (2014), relatam que o câncer de pulmão é um dos tipos de câncer mais agressivos, possuindo uma razão mortalidade/incidência de, aproximadamente, 0,87. Para a obtenção da razão mortalidade/incidência, dividiu-se o número de casos de mortalidade pelo de incidência, ambos referentes ao período compreendido entre 2001 a 2014. Os autores afirmam que a detecção desse tipo de câncer é feita quando o mesmo se encontra em estágio avançado, pois os sintomas não são observados nos estágios iniciais. Desse modo a sobrevida em cinco anos é baixa para maioria das população mundial, sendo em média de 10% a 15%.

A principal causa de câncer de pulmão é o tabagismo, que causa cerca de sete milhões de mortes por ano, incluindo as mortes por câncer (AMERICAN CANCER SOCIETY, 2015).

Segundo Alwan et al. (2011), até 2020, o número de mortes causadas pelo tabagismo aumentará para 7,5 milhões, representando 10% de todas as mortes. O mesmo autor estimou que o tabagismo cause cerca de 71% do câncer de pulmão, 42% das doenças respiratórias crônicas e quase 10% das doenças cardiovasculares.

De acordo com Instituto Nacional de Câncer José Alencar Gomes da Silva (2017), o tabagismo tem influência na mortalidade por câncer de pulmão e o controle do tabaco tem sido o principal modo de redução desse tipo de doença. Para esse fim foi criado o Programa Nacional de Controle do Tabagismo do Brasil, a partir do qual pôde ser observado uma tendencia à redução da incidência e da mortalidade por câncer, principalmente de câncer de pulmão, relacionado ao tabaco.

Segundo Green et al. (1993), a incidência do câncer de pulmão tem um aumento substancial a partir dos 50 anos. Assim, a idade pode ser um fator influente na sobrevivência de pacientes com câncer de pulmão.

Ramalingam et al. (1998) concluíram que a presença de maior quantidade de mulheres e negros no grupo de indivíduos mais jovens com câncer de pulmão sugere a existência de suscetibilidade dessas populações à agentes cancerígenos.

Levando em consideração os fatos mencionados, torna-se importante o conhecimento do

comportamento do câncer de pulmão, dos fatores a eles associados e que podem influenciar na sobrevida dos pacientes, a partir do momento que as causas que influenciam na sobrevida são identificadas. Uma área da Estatística utilizada nesse tipo de estudo é a análise de sobrevivência, a qual será abordada na seção 2.2.

2.2 ANÁLISE DE SOBREVIVÊNCIA

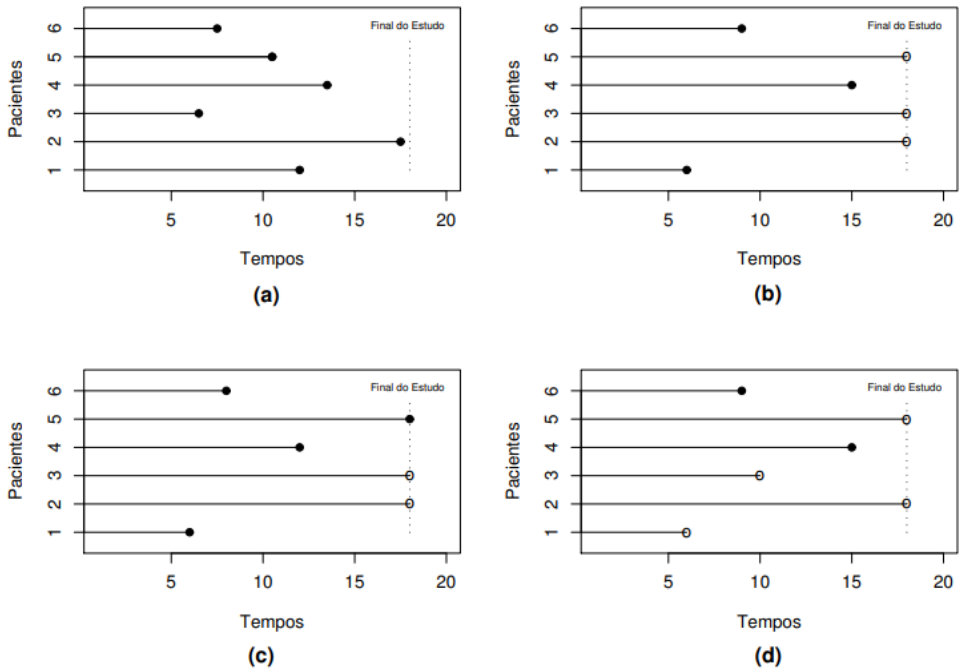
Segundo Hermeto (2014), a análise de sobrevivência é uma área da estatística que foi explorada de modo mais frequente a partir das últimas décadas do século XX, conquistando o seu maior desenvolvimento e notoriedade por volta da década de oitenta desse mesmo século. Embora as aplicações da análise de sobrevivência estejam fortemente relacionadas com a medicina, as técnicas deste procedimento estatístico podem ser utilizadas nas mais variadas áreas, tais como engenharia, visto no trabalho de Wang, Mahboub e Hancher (2005), sociologia, encontrado em Blossfeld, Hamerle e Mayer (2014), entre outras.

De acordo com Colosimo e Giolo (2006), a análise de sobrevivência diz respeito ao estudo de dados relacionados ao tempo de ocorrência de um determinado evento de interesse, partindo de um tempo inicial, o qual deve ser predefinido, até o tempo final. Este tempo é denominado de tempo de falha, isto é, tempo de ocorrência do evento de interesse, como por exemplo, o tempo até a morte do paciente portador de câncer, o tempo até a prisão de um assassino ou ainda tempo de vida útil de um equipamento eletrônico, entre outros. Há sempre a possibilidade do evento de interesse não ocorrer com alguns indivíduos/objetos do estudo. Quando esse evento de interesse não ocorre, tem-se observações incompletas, as quais são definidas como censura.

Souza (2015) descreve que a principal característica dos dados de sobrevivência é existência de observações censuradas, as quais podem ser entendidas como aquelas observações parciais da variável resposta, isto é, que por algum motivo o acompanhamento do indivíduo/objeto em estudo é interrompido e o evento de interesse não é observado.

O tipo de censura mais conhecido é censura à direita, a qual pode ser entendida através dos mecanismos de censura apresentados na Figura 1, em que o tempo de ocorrência do evento de interesse está à direita do tempo registrado, ou seja, o tempo de falha é maior que o tempo observado.

Figura 1 – Ilustração de alguns mecanismos de censura em que • representa a falha e ○ representa a censura. (a) Dados completos, (b) Dados com censura tipo I, (c) Dados com censura tipo II e (d) Dados com censura aleatória.



Fonte: (COLOSIMO; GIOLO, 2006).

Na Figura 1 (a) observa-se os dados completos, isto é, todos os pacientes experimentaram o evento antes do final do estudo. Na Figura 1 (b) alguns pacientes não experimentaram o evento até o final do estudo, caracterizando os dados com censura tipo I. Na Figura 1 (c) tem-se a presença de dados com censura tipo II, em que o estudo foi finalizado após a ocorrência de um número pré-estabelecido de falhas e na Figura 1 (d) observa-se a ilustração dos dados com censura aleatória, onde o acompanhamento de alguns pacientes foi interrompido por alguma razão e alguns pacientes não experimentaram o evento até o final do estudo.

De acordo com Hosmer e Lemeshow (1999), também podem ocorrer a censura à esquerda, quando o tempo registrado é maior que o tempo de falha, em outras palavras, quando o evento de interesse havia ocorrido quando o objeto em estudo foi observado. E também a censura intervalar, a qual acontece em estudos onde os objetos ou indivíduos são acompanhados em visitas periódicas e, desse modo a informação conhecida é apenas que o evento de interesse aconteceu em um determinado intervalo de tempo, pois o tempo de falha exato não é conhecido.

Machin, Cheung e Parmar (2006) destacam a importância de se utilizar todos os resultados provenientes de um estudo de sobrevivência, mesmo que entre estes existam observações

censuradas que, embora sejam incompletas, carregam informações sobre o tempo de vida do indivíduo ou objeto em estudo, e se excluídos podem levar a conclusões equivocadas.

Segundo Botelho, Silva e Cruz (2009), na análise de sobrevivência tem-se que a variável dependente ou resposta é, comumente, o tempo até ocorrência de determinado evento de interesse. Caso não haja censuras, pode-se usar as técnicas de estatística clássica na análise dos dados, porém caso exista a presença de dados censurados, tais técnicas não podem ser utilizadas, pois estas necessitam de todos os tempos de falha. Desse modo, tem-se a necessidade de utilizar um dos métodos de análise de sobrevivência, os quais permitem agregar as informações contidas nos dados censurados.

Conforme Bustamante-Teixeira, Faerstein e Latorre (2002), foi após as décadas de 1950 e de 1960 que surgiram as primeiras propostas de estimadores das probabilidades de sobrevida que introduziam os dados censurados. Entre os estimadores não-paramétricos utilizados na análise de sobrevivência destaca-se o estimador de Kaplan-Meier, o qual foi proposto por Kaplan e Meier (1958), e será apresentado na subseção 2.3.

De acordo com Hermeto (2014), a variável aleatória que descreve o tempo no qual um indivíduo ou objeto de estudo participante de um experimento de sobrevivência leva para manifestar o evento de interesse é uma variável aleatória contínua e pode ou não seguir uma distribuição de probabilidade conhecida. A partir do momento que se tem o conhecimento de que a variável aleatória segue uma distribuição de probabilidade conhecida, pode-se utilizar as funções de sobrevivência, função densidade e função de risco estabelecidas para cada distribuição.

Alan (1980) apresenta alguns modelos paramétricos muito utilizados para descrever dados de sobrevivência, tais como as distribuições Exponencial, Weibull, Log-normal, Gama e de Valores Extremos.

Outros modelos, também muito empregados na análise de dados de sobrevida, são os semi-paramétricos, como por exemplo o desenvolvido por Cox (1972), comumente utilizado em estudos das áreas médicas.

Segundo Zhu et al. (2011), um modelo semi-paramétrico que também tem sido amplamente utilizado é o de Cox, principalmente por explicar a relação entre a sobrevida e as covariáveis. O autor exemplifica essa utilização com um estudo em que foi feito para avaliar o efeito de fatores clínico-patológicos e demográficos na sobrevida de pacientes com câncer de estômago, as covariáveis que foram consideradas significativas no estudo foram a idade, o

gênero, o histórico familiar de câncer gástrico ou características diagnósticas.

Segundo Moghimi-Dehkordi et al. (2008), mesmo que a preferência dos pesquisadores em ciências médicas seja pela utilização de modelos semi-paramétricos, os modelos paramétricos podem fornecer estimativas mais precisas, tendo em vista as boas propriedades que os estimadores apresentam, sendo assim também indicado para descrever o comportamento da variável de interesse.

Algumas das principais distribuições utilizadas na análise de sobrevivência, tais como, Weibull, Log-normal, Gama e Gama Generalizada, serão apresentadas na seção 2.4. Para o entendimento da utilização tanto do método de estimação de Kaplan-Meier, apresentado na subseção 2.3, quanto dos modelos paramétricos apresentados, faz-se necessário a definição das funções de sobrevivência, taxa de falha e taxa de falha acumulada.

De acordo com Hosmer e Lemeshow (1999), a variável aleatória não-negativa que representa o tempo de falha, simbolizada por T, é usualmente determinada em análise de sobrevivência pela sua função de sobrevivência. A função de sobrevivência é definida como a probabilidade de uma observação não falhar até um certo tempo t, isto é, a probabilidade de uma observação sobreviver ao tempo t. O que pode ser expresso da seguinte maneira:

$$S(t) = P(T \geq t) = 1 - F(t), \tag{2.1}$$

em que $F(t)$ é a função de distribuição acumulada.

A probabilidade de ocorrência de falha em um intervalo de tempo $[t_1, t_2)$ pode ser apresentada em termos da função de sobrevivência como:

$$S(t_1) - S(t_2). \tag{2.2}$$

Segundo Colosimo e Giolo (2006), a taxa de falha no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Logo, de maneira geral, adotando o intervalo como $[t, t + \Delta t)$, a taxa de falha no tempo t condicional à sobrevivência até o tempo t, assumindo Δt bem pequeno,

pode ser definida como:

$$\lambda(t) = \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{2.3}$$

desenvolvendo a expressão (3.4), tem-se que

$$\begin{aligned}
\lambda(t) &= \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\
&= \frac{P[(t \leq T < t + \Delta t) \cap (T \geq t)]}{\Delta t P(T \geq t)} \\
&= \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} \\
&= \frac{P(T \geq t) - P(T \geq t + \Delta t)}{\Delta t P(T \geq t)} \\
&= \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.
\end{aligned} \tag{2.4}$$

Desse modo, tem-se que a função de taxa de falha instantânea de T é definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{2.5}$$

ou também como:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \\
&= \lim_{\Delta t \to 0} \frac{1 - F(t) - [1 - F(t + \Delta t)]}{\Delta t S(t)} \\
&= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\
&= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\
&= \frac{f(t)}{S(t)},
\end{aligned} \tag{2.6}$$

ou ainda,

$$\begin{aligned}\lambda(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{-f(t)}{1-F(t)} \\ &= -\frac{d[log(1-F(t))]}{dt} \\ &= -\frac{d}{dt}\left[log(1-F(t))\right] \\ &= -\frac{d}{dt}\left[log\ S(t)\right]. \end{aligned} \tag{2.7}$$

A função taxa de falha acumulada é definida por:

$$\Lambda(t) = \int_0^t \lambda(u)du. \tag{2.8}$$

Uma relação importante entre a função taxa de falha acumulada e a função de sobrevivência pode ser observada a seguir:

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(u)du \\ &= -log\ S(t). \end{aligned} \tag{2.9}$$

Para encontrar uma estimativa para a função de sobrevivência, pode-se utilizar o estimador não-paramétrico de Kaplan-Meier, descrito na subseção 2.3.

2.3 MÉTODO DE ESTIMAÇÃO DE KAPLAN-MEIER

Segundo Colosimo e Giolo (2006), o estimador de Kaplan-Meier para a função de sobrevivência, ou estimador limite-produto, proposto por Kaplan e Meier (1958), é um dos mais utilizados nas pesquisas das áreas médicas. Este estimador é baseado no procedimento geral descrito a seguir, onde $t_j,\ j = 1, \cdots, p$ são os p tempos distintos e ordenados de falha, d_j é o número de falhas em $t_j,\ j = 1, \cdots, p$, e n_j é o número de indivíduos sob risco em t_j:

a) A escala de tempo é dividida em intervalos adequadamente escolhidos de acordo com os tempos de falha, $(0, t_1), (t_1, t_2), \cdots, (t_{p-1}, t_p)$;

b) Para cada intervalo (t_j, t_{j+1}) estima-se $\widehat{q}_j = d_j/n_{j-1}$, a proporção de indivíduos vivos após t_{j-1} que sobrevivem além de t_j;

c) Se t é um ponto de divisão a proporção $S(t)$ na população sobrevivente após t é estimada pelo produto do $\widehat{q}_j$ estimado para todos os intervalos anteriores a t.

De acordo com Kleinbaum e Klein (2010), a condição para a obtenção do estimador é que dentro de cada intervalo, as falhas e censuras sejam separadas de uma maneira conhecida. Desse modo, é possível supor que nenhum intervalo contém falhas e censuras. Assim, se o número sob observação logo após t_{j-1} é denotado por n_j, e d_j as falhas observadas no intervalo (t_{j-1}, t_j), a estimativa é então:

$$\widehat{q}_j = \frac{(n_j - d_j)}{n_j}. \tag{2.10}$$

O estimador de Kaplan-Meier é dado por:

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right). \tag{2.11}$$

Segundo Kaplan e Meier (1958), outra propriedade importante do estimador de Kaplan-Meier é que este converge assintoticamente para uma Distribuição Normal. Esse estimador é consistente e não-viciado para amostras grandes, e é estimador de máxima verossimilhança de $S(t)$. A demonstração de que o estimador de Kaplan-Meier para $S(t)$ é um estimador de máxima verossimilhança será apresentada a seguir:

Suponhamos que d_j observações falham no tempo t_j, para $j = 1, \cdots, p$, e m_j observações são censuradas no intervalo $[t_j, t_j + 1)$, nos tempos $t_{j1}, \cdots, t_{jm_j}$. A probabilidade de falha no tempo t_j é:

$$S(t_j) - S(t_j+), \tag{2.12}$$

com $S(t_j+) = \lim_{\Delta t \to 0^+} S(t + \Delta t))$, $j = 1, \cdots, p$. Contudo, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_{jl} para $l = 1, \cdots, m_j$ é:

$$P(T > t_{jl}) - S(t_j+)S(t_{jl}+), \tag{2.13}$$

Desse modo, a função de verossimilhança pode ser escrita como:

$$L(S(\cdot)) = \prod_{j=0}^{p} \left\{ [S(t_j) - S(t_j+)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}+) \right\}. \tag{2.14}$$

Logo, pode-se mostrar que $S(t)$, que maximiza $L(S(\cdot))$, é a expressão (2.11).

Segundo Araujo (2008) esta definição do estimador de máxima verossimilhança é uma generalização do conceito usual utilizado em modelos paramétricos, em que se tem tantos parâmetros quanto falhas distintas.

2.4 DISTRIBUIÇÕES MAIS UTILIZADAS NA ANÁLISE DE SOBREVIVÊNCIA

As distribuições Weibull, Log-normal, Gama e Gama Generalizada são algumas das principais distribuições utilizadas na análise de sobrevivência e as quais foram ajustadas aos dados de sobrevivência neste estudo, apresentadas nas subseções a seguir.

2.4.1 Distribuição Weibull

Segundo Colosimo e Giolo (2006), a função densidade de probabilidade da Distribuição Weibull para a variável aleatória T é dada por:

$$f(t;\mu,\sigma) = \frac{\sigma}{\mu^{\sigma}} t^{\sigma-1} \exp\left\{ -\left(\frac{t}{\mu}\right)^{\sigma} \right\}, \tag{2.15}$$

em que $t \geq 0, \;\; \mu > 0$ é o parâmetro de escala e $\sigma > 0$ é o parâmetro de forma, que dependem do vetor de covariáveis $\boldsymbol{x}$, ou seja $\mu = \exp(\boldsymbol{x}^T \boldsymbol{\beta_\mu})$ e $\sigma = \exp(\boldsymbol{x}^T \boldsymbol{\beta_\sigma})$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)$ é o vetor de parâmetros desconhecidos e $\boldsymbol{x} = (x_1, x_2, \cdots, x_q)$, sendo q o número de variáveis explicativas. Tem-se também que a Esperança e a Variância da variável aleatória T são dadas, respectivamente por:

$$E[T] = \mu \Gamma\left[1 + \sigma^{-1}\right], \tag{2.16}$$

e

$$Var[T] = \mu^2 \left[\Gamma\left[1 + 2\sigma^{-1}\right] - \Gamma\left[1 + \sigma^{-1}\right]^2 \right], \tag{2.17}$$

em que $\Gamma(k)$ a função gama, definida por $\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx$.

A função de sobrevivência $S(t)$ para a Distribuição Weibull é determinada da seguinte forma:

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t)) \\ &= 1 - \int_0^t \frac{\sigma}{\mu^\sigma} \exp\left\{-\left(\frac{s}{\mu}\right)^\sigma\right\} ds, \end{aligned}$$

realizando a seguinte substituição:

$$u = -\left(\frac{s}{\mu}\right)^\sigma \qquad \Rightarrow \qquad du = -\left(\frac{\sigma \cdot s^{\sigma-1}}{\mu^\sigma}\right) ds,$$

voltando para a função de sobrevivência, tem-se então:

$$\begin{aligned} S(t) &= 1 - \int_0^{-\left(\frac{t}{\mu}\right)^\sigma} -\exp(u)\, du \\ &= 1 + \left[\exp(u)\right]_0^{-\left(\frac{t}{\mu}\right)^\sigma} \\ &= 1 + \exp\left\{-\left(\frac{t}{\mu}\right)^\sigma\right\} - \exp(0) \\ &= 1 + \exp\left\{-\left(\frac{t}{\mu}\right)^\sigma\right\} - 1 \\ &= \exp\left\{-\left(\frac{t}{\mu}\right)^\sigma\right\}. \end{aligned} \tag{2.18}$$

A função de taxa de falha para a Distribuição Weibull é determinada da seguinte forma:

$$\lambda(t) = \frac{\sigma}{\mu^\sigma} t^{\sigma-1}. \tag{2.19}$$

2.4.2 Distribuição Log-normal

Lee e Wang (2003) apresentam a função densidade de probabilidade da Distribuição Log-normal em relação a variável aleatória T da seguinte forma:

$$f(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi} t \sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}, \tag{2.20}$$

em que $t > 0$, μ é a média do logaritmo do tempo de falha e σ é o desvio-parão, que dependem do vetor de covariáveis $\boldsymbol{x}$, ou seja $\mu = \boldsymbol{x}^T\boldsymbol{\beta_\mu}$ e $\sigma = \exp(\boldsymbol{x}^T\boldsymbol{\beta_\sigma})$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)$ é o vetor de parâmetros desconhecidos e $\boldsymbol{x} = (x_1, x_2, \cdots, x_q)$, sendo q o número de variáveis explicativas.

A Esperança e a Variância da variável aleatória T seguindo a Distribuição Log-normal são dadas, respectivamente por:

$$E[T] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}, \tag{2.21}$$

e

$$Var[T] = \exp\left\{2\mu + \sigma^2\right\}\left(\exp\left\{\sigma^2\right\} - 1\right). \tag{2.22}$$

A função de sobrevivência $S(t)$ para a Distribuição Log-normal é representadas por:

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right), \tag{2.23}$$

com $\Phi(\cdot)$ sendo a função de distribuição acumulada de uma normal padrão, pois se T tem distribuição Log-normal, então $X = log(t)$ tem distribuição Normal, com parâmetros μ e σ, pois:

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= P(\exp(X) \leq t) \\ &= P(X \leq \log(t)) \\ &= P\left(\frac{X - \mu}{\sigma} \leq \frac{\log(t) - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{\log(t) - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{\log(t) - \mu}{\sigma}\right). \end{aligned} \tag{2.24}$$

A Função de taxa de falha $\lambda(t)$ para a Distribuição Log-normal é dada por:

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.25}$$

2.4.3 Distribuição Gama

Segundo Gross e Clark (1975), a função densidade de probabilidade da Distribuição Gama em relação a variável aleatória T é dada por:

$$f(t;\mu,\sigma) = \frac{1}{\Gamma(\mu)\sigma^{\mu}} t^{\mu-1} \exp\left\{-\frac{t}{\sigma}\right\}, \tag{2.26}$$

em que $t > 0$, $\mu > 0$ é chamado de parâmetro forma e $\sigma > 0$ de escala. Os parâmetros dependem do vetor de covariáveis $\boldsymbol{x}$, ou seja $\mu = \exp(\boldsymbol{x}^T\boldsymbol{\beta_\mu})$ e $\sigma = \exp(\boldsymbol{x}^T\boldsymbol{\beta_\sigma})$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)$ é o vetor de parâmetros desconhecidos e $\boldsymbol{x} = (x_1, x_2, \cdots, x_q)$, sendo q o número de variáveis explicativas.

A Esperança e a Variância da distribuição Gama são dadas, respectivamente por,

$$E[T] = \mu\sigma \quad \text{e} \quad Var[T] = \mu\sigma^2. \tag{2.27}$$

As funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$ para a Distribuição Gama são dadas respectivamente por:

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t)) \\ &= 1 - \int_0^t \frac{1}{\Gamma(r)\mu^r} u^{r-1} \exp\left\{-\frac{u}{\mu}\right\} du \end{aligned} \tag{2.28}$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.29}$$

2.4.4 Distribuição Gama Generalizada

A distribuição Gama Generalizada foi proposta por Stacy (1962), podendo ser caracterizada por seus três parâmetros, σ, r e μ, em que todos estes são positivos. A a função densidade

de probabilidade da distribuição Gama Generalizada em relação a variável aleatória T é dada por:

$$f(t;\mu,\sigma,r) = \frac{r}{\Gamma(\mu)\sigma^{r\mu}} t^{r\mu-1} \exp\left\{-\left(\frac{t}{\sigma}\right)^r\right\}, \tag{2.30}$$

em que $\sigma > 0$ é o parâmetro de escala, $\mu > 0$ e $r > 0$ são os parâmetros de forma, que dependem do vetor de covariáveis $\boldsymbol{x}$, ou seja $\sigma = \exp(\boldsymbol{x}^T\boldsymbol{\beta_\sigma})$, $\mu = \exp(\boldsymbol{x}^T\boldsymbol{\beta_\mu})$ e $r = \boldsymbol{x}^T\boldsymbol{\beta_r}$, em que $\boldsymbol{\beta_m} = (\beta_0, \beta_1, \cdots, \beta_q)$ é o vetor de parâmetros desconhecidos, sendo $m = \mu, \sigma$ e r, e $\boldsymbol{x} = (x_1, x_2, \cdots, x_q)$, sendo q o número de variáveis explicativas.

Esta é uma importante distribuição visto que possui como casos particulares: a distribuição Exponencial quando $\gamma = k = 1$; a distribuição Gama quando $\gamma = 1$ e a distribuição Weibull quando $k = 1$.

A Esperança e a Variância da distribuição Gama Generalizada são dadas, respectivamente por,

$$E[T] = \frac{\sigma\Gamma\left(\frac{r\mu+1}{r}\right)}{\Gamma(\mu)} \quad \text{e} \quad Var[T] = \frac{\sigma^2}{\Gamma(\mu)}\left\{\Gamma\left(\frac{r\mu+2}{r}\right) - \frac{\left[\Gamma\left(\frac{r\mu+1}{r}\right)\right]^2}{\Gamma(\mu)}\right\}. \tag{2.31}$$

As funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$ para a Distribuição Gama Generalizada são dadas respectivamente por:

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t)) \\ &= 1 - \int_0^t \frac{r}{\Gamma(\mu)\sigma^{r\mu}} s^{r\mu-1} \exp\left\{-\left(\frac{s}{\sigma}\right)^r\right\} ds \end{aligned} \tag{2.32}$$

e

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{t^{r\mu-1}}{S(t)}. \tag{2.33}$$

em que $F(t)$ é a função de distribuição acumulada da Gama Generalizada.

2.5 ESTIMAÇÃO DOS PARÂMETROS DOS MODELOS

No ajuste de uma distribuição de probabilidade a um conjunto de dados faz-se necessário a estimação dos seus respectivos parâmetros. Para a estimação dos parâmetros alguns métodos foram propostos, tais como método dos quadrados mínimos, proposto de maneira independente por Gauss (1809) e Legendre (1806), segundo Garnés, Sampaio e Dalmolin (1997), e método da máxima verossimilhança, cuja formulação original foi feita por Fisher (1922).

O método dos quadrados mínimos é uma técnica que busca encontrar o melhor ajuste para um conjunto de dados, com a finalidade de minimizar a soma dos quadrados das diferenças entre os valores preditos e os dados observados. Essas diferenças são chamadas resíduos. Segundo Buckley e James (1979) este método é uma alternativa para o modelo de Cox quando a suposição de riscos proporcionais é violada.

O método da máxima verossimilhança, que segundo Zanakis e Kyparisis (1986), é considerado um dos mais confiáveis, pois ele têm as propriedades desejáveis de consistência, normalidade assintótica e eficiência assintótica.

Segundo Colosimo e Giolo (2006), o de estimação da máxima verossimilhança é apropriado para estimar os parâmetros das distribuições de probabilidades utilizadas para descrever o tempo de vida é o da máxima verossimilhança.

Portanto, o método dos quadrados mínimos procura encontrar os estimadores que reduzem os resíduos, enquanto o método da máxima verossimilhança busca encontrar os estimadores que maximizam a plausibilidade da amostra ter ocorrido. A escolha de qual método de estimação utilizar fica a critério do pesquisador. No presente estudo, optou-se pelo método de máxima verossimilhança, visto que este está implementado no pacote a ser utilizado.

2.5.1 Método de Máxima Verossimilhança

Para o entendimento do método da máxima verossimilhança, considere uma amostra $t_1, t_2, ..., t_n$, de observações de uma dada população de interesse em que todas são não-censuradas.

Suponha que a população seja caracterizada pela sua função de densidade $f(t)$. Assegurada a pressuposição de independência entre as n observações, a função de verossimilhança para

um vetor de parâmetros $\boldsymbol{\theta}$ pode ser expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}). \tag{2.34}$$

É possível notar que a função de verossimilhança mostra que a contribuição de cada observação não-censurada é a sua função de densidade. Mas na presença de observações censuradas esse processo não irá fornecer o resultado correto, pois a contribuição de cada observação censurada não é sua função de densidade. De acordo com Tableman e Kim (2003), para que se possa utilizar o método da Máxima Verossimilhança, deve-se separar as n observações em dois conjuntos: um contendo as p observações não-censuradas e um contendo as $n - p$ observações censuradas. A função de verossimilhança para o caso em que se tem obervações censuradas pode ser definida como a seguir:

1. **Censura do tipo I**: tem-se p falhas, pois estas foram as observações não-censuradas, e $n - p$ censuras observadas ao fim do experimento e, desse modo, $L(\boldsymbol{\theta})$ assume a seguinte forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta}) \prod_{i=p+1}^{n} S(t_i; \boldsymbol{\theta}). \tag{2.35}$$

em que,

$$\prod_{i=p+1}^{n} S(c; \boldsymbol{\theta}) = [S(c; \boldsymbol{\theta})]^{n-p}, \tag{2.36}$$

em que as censuras ocorrem em $T = C$, sendo T o tempo de falha e C o tempo de censura.

2. **Censura do tipo II**: p, sendo p o número de falhas, é fixo e somente os p menores tempos são observados. Desse modo, segue que:

$$L(\boldsymbol{\theta}) = \frac{n!}{(n-p)!} \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta}) \prod_{i=p+1}^{n} S(t_i; \boldsymbol{\theta}). \tag{2.37}$$

em que,

$$\prod_{i=p+1}^{n} S(c; \boldsymbol{\theta}) = [S(c; \boldsymbol{\theta})]^{n-p}, \tag{2.38}$$

com T_p é o maior tempo observado. Note que $\dfrac{n!}{(n-p)!}$ é uma constante e, assim pode ser desprezado, levando em conta que não envolve qualquer parâmetro de interesse. Logo

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta}) \prod_{i=p+1}^{n} S(t_i; \boldsymbol{\theta}). \tag{2.39}$$

3. **Censura do tipo aleatória:** considerando T o tempo de falha e C o de censura, para $i-1, 2, ..., n$, os dados observados consistem nos pares (t_i, δ_i), em que $t_i = min(T_i, C_i)$ e $\delta_i = 1$ se $T_i \leq C_i$ ou $\delta_i = 0$ se $T_i \geq C_i$. Considerando T e C independentes e supondo que $g(c)$ e $G(c)$ as funções de densidade e sobrevivência de C, respectivamente, se para o i-ésimo indivíduo,

(a) for observada uma censura, tem-se que:

$$P(t_i = t, \delta_i = 0) = P(C_i = t, T_i > C_i) = P(C_i = t, T_i > t) = g(t)S(t; \boldsymbol{\theta}). \tag{2.40}$$

(b) e, se for observada uma falha, tem-se que:

$$P(t_i = t, \delta_i = 1) = P(C_i = t, T_i \leq C_i) = P(C_i = t, T_i \geq t) = f(t; \boldsymbol{\theta})G(t). \tag{2.41}$$

Desse modo,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta})G(t_i) \prod_{i=p+1}^{n} g(t_i)S(t_i; \boldsymbol{\theta}). \tag{2.42}$$

Com isso, supondo que o mecanismo de censura não carregue informações sobre os parâmetros, os termos $G(t)$ e $g(t)$ podem ser ignorados, visto que não envolvem $\boldsymbol{\theta}$, logo a função de verossimilhança será:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta}) \prod_{i=p+1}^{n} S(t_i; \boldsymbol{\theta}). \tag{2.43}$$

Portanto, tem-se então que a função de verossimilhança para todos os mecanismos de censura, com excessão de constantes, é representada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} f(t_i; \boldsymbol{\theta}) \prod_{i=p+1}^{n} S(t_i; \boldsymbol{\theta}). \tag{2.44}$$

ou

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [f(t_i; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{\theta})]^{1-\delta_i}. \tag{2.45}$$

em que δ_i é a variável indicadora de falha ou censura,

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ é um tempo de falha} \\ 0, \text{ se } t_i \text{ é um tempo censurado} \end{cases} \tag{2.46}$$

De acordo com Hosmer e Lemeshow (1999), os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam a função de verossimilhança $L(\boldsymbol{\theta})$, isto é, que maximizam a plausibilidade da amostra ter ocorrido. Para realizar a maximização do vetor de parâmetros deve-se calcular as derivadas parciais em relação a cada um de seus parâmetros, e em seguida igualar as derivadas parciais a zero.

Para simplificar os cálculos, nota-se que é mais vantajoso trabalhar com soma ao invés de produtos, pois é mais fácil derivar soma do que produto de funções. Para transformar funções envolvendo produtos em somas, pode-se utilizar o logaritmo dessas funções. Desse modo, para obter-se o máximo da função $L(\boldsymbol{\theta})$ é conveniente aplicar o logaritmo natural à essa função. Portanto, os estimadores de máxima verossimilhança dos parâmetros são encontrados resolvendo-se:

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0. \tag{2.47}$$

Seja uma amostra aleatória $t_1, \cdots, t_n$ de uma variável aleatória T com distribuição de Weibull, sendo que t_i, $i = 1, \cdots, n$, indica o tempo de censura ou de falha e $\boldsymbol{\theta} = (\mu, \sigma)$. Desse modo, a função de verossimilhança pode ser escrita da seguinte forma:

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{t}) &= \prod_{i=1}^{n} \left[f\left(t_i \mid \boldsymbol{\theta}\right)\right]^{\delta_i} \left[S\left(t_i \mid \boldsymbol{\theta}\right)\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[\frac{\sigma}{\mu^\sigma} {t_i}^{\sigma-1} \exp\left\{-\left(\frac{t_i}{\mu}\right)^\sigma\right\}\right]^{\delta_i} \left[exp\left\{-\left(\frac{t_i}{\mu}\right)^\sigma\right\}\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[\frac{\sigma}{\mu^\sigma} {t_i}^{\sigma-1}\right]^{\delta_i} \left[\exp\left\{-\left(\frac{t_i}{\mu}\right)^\sigma\right\}\right]^{\delta_i} \left[\exp\left\{-\left(\frac{t_i}{\mu}\right)^\sigma\right\}\right] \left[\exp\left\{-\left(\frac{t_i}{\mu}\right)^\sigma\right\}\right]^{\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{\sigma}{\mu^\sigma} {t_i}^{\sigma-1}\right]^{\delta_i} \left[\exp\left\{-\left(\frac{t}{\mu}\right)^\sigma\right\}\right]. \end{aligned} \tag{2.48}$$

Tomando o logarítimo de $L(\boldsymbol{\theta})$, tem-se:

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta})) \\
&= \log\left\{\prod_{i=1}^{n}\left[\frac{\sigma}{\mu^{\sigma}}{t_i}^{\sigma-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\mu}\right)^{\sigma}\right\}\right]\right\} \\
&= \sum_{i=1}^{n}\log\left\{\left[\frac{\sigma}{\mu^{\sigma}}{t_i}^{\sigma-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t}{\mu}\right)^{\sigma}\right\}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\log\left[\frac{\sigma}{\mu^{\sigma}}{t_i}^{\sigma-1}\right]^{\delta_i} + \log\left[\exp\left\{-\left(\frac{t_i}{\mu}\right)^{\sigma}\right\}\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\delta_i\log\left[\frac{\sigma}{\mu^{\sigma}}{t_i}^{\sigma-1}\right] - \left(\frac{t_i}{\mu}\right)^{\sigma}\right\} \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\sigma) - \delta_i\log(\mu)^{\sigma} + \delta_i\log(t_i)^{\sigma-1} - \left(\frac{t_i}{\mu}\right)^{\sigma}\right\} \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\sigma) - \sigma\delta_i\log(\mu) + (\sigma-1)\delta_i\log(t_i) - \mu^{-\sigma}{t_i}^{\sigma}\right\} \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\sigma) - \sigma\delta_i\log(\mu) + \sigma\delta_i\log(t_i) - \delta_i\log(t_i) - \mu^{-\sigma}{t_i}^{\sigma}\right\} \\
&= \log(\sigma)\sum_{i=1}^{n}\delta_i - \sigma\left[\log(\mu)\right]\sum_{i=1}^{n}\delta_i + \sigma\sum_{i=1}^{n}\delta_i\log(t_i) - \sum_{i=1}^{n}\delta_i\log(t_i) - \mu^{-\sigma}\sum_{i=1}^{n}{t_i}^{\sigma},
\end{aligned}
$$

como p é o número de falhas, tem-se que $\sum_{i=1}^{n}\delta_i = p$, portanto:

$$
l(\boldsymbol{\theta}) = p\left[\log(\sigma)\right] - p\sigma\left[\log(\mu)\right] + \sigma\sum_{i=1}^{n}\delta_i\log(t_i) - \sum_{i=1}^{n}\delta_i\log(t_i) - \mu^{-\sigma}\sum_{i=1}^{n}{t_i}^{\sigma}. \tag{2.49}
$$

Derivando-se a expressão (2.49) em relação à ambos os parâmetros μ e σ tem-se:

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial\mu} &= -\frac{p\sigma}{\mu} + \sigma\mu^{-\sigma-1}\sum_{i=1}^{n}{t_i}^{\sigma} \\
&= -\frac{p\sigma}{\mu} + \frac{\sigma\mu^{-\sigma}\sum_{i=1}^{n}{t_i}^{\sigma}}{\mu} \\
&= \frac{\sigma}{\mu}\left(-p + \mu^{-\sigma}\sum_{i=1}^{n}{t_i}^{\sigma}\right)
\end{aligned} \tag{2.50}
$$

e

$$
\frac{\partial l(\boldsymbol{\theta})}{\partial\sigma} = \frac{p}{\sigma} - p\left[\log(\mu)\right] + \sum_{i=1}^{n}\delta_i\log(t_i) + \mu^{-\sigma}\left[\log(\mu)\right]\sum_{i=1}^{n}{t_i}^{\sigma} - \mu^{-\sigma}\sum_{i=1}^{n}{t_i}^{\sigma}\log(t_i). \tag{2.51}
$$

Igualando ambas as expressões (2.50) e (2.50) a zero, obtém-se:

$$\begin{aligned}
\frac{\widehat{\sigma}}{\widehat{\mu}}\left(-p+\widehat{\mu}^{-\widehat{\sigma}}\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\right)&=0\\
-p+\widehat{\mu}^{-\widehat{\sigma}}\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}&=0\\
\widehat{\mu}^{-\widehat{\sigma}}\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}&=p\\
\widehat{\mu}^{-\widehat{\sigma}}&=\frac{p}{\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}}\\
\widehat{\mu}^{\widehat{\sigma}}&=\frac{\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}}{p}\\
\widehat{\mu}&=\left(\frac{\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}}{p}\right)^{\frac{1}{\widehat{\sigma}}}
\end{aligned}\tag{2.52}$$

e

$$\frac{p}{\widehat{\sigma}}-p\left[\log(\widehat{\mu})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+\widehat{\mu}^{-\widehat{\sigma}}\left[\log(\widehat{\mu})\right]\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}-\widehat{\mu}^{-\widehat{\sigma}}\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)=0$$

$$\frac{p}{\widehat{\sigma}}-p\left[\log(\widehat{\mu})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+\frac{p}{\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}}\log(\widehat{\mu})\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}-\frac{p}{\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}}\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)=0$$

$$\frac{p}{\widehat{\sigma}}-p\left[\log(\widehat{\mu})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+p\left[\log(\widehat{\mu})\right]-p\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)=0$$

$$\frac{p}{\widehat{\sigma}}+\sum_{i=1}^{n}\delta_i\log(t_i)-p\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)=0$$

$$p\left[\frac{1}{\widehat{\sigma}}+\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{p}-\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)\right]=0$$

$$\frac{1}{\widehat{\sigma}}+\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{p}-\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)=0$$

$$\frac{1}{\widehat{\sigma}}=\sum_{i=1}^{n}{t_i}^{\widehat{\sigma}}\log(t_i)-\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{p}.\tag{2.53}$$

As soluções deste sistema não-linear fornecem os estimadores de máxima verossimilhança dos parâmetros, porém este não possui solução analítica, logo métodos numéricos como o

de Newton-Rapshon são necessários para encontrar tais soluções. A seguir, aplicou-se o método de Newton-Raphson, o qual usa a matriz de derivadas segundas da função de log-verossimilhança. Seja $U(\boldsymbol{\theta})$ a função escore apresentada na (2.47). Tem-se que, para o estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$,

$$U(\widehat{\boldsymbol{\theta}}) = 0, \tag{2.54}$$

expandindo $U(\widehat{\boldsymbol{\theta}})$ em série de Taylor em torno de um $\boldsymbol{\theta}_0$, tem-se que:

$$0 = U(\widehat{\boldsymbol{\theta}}) \cong U(\boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)H(\boldsymbol{\theta}_0), \tag{2.55}$$

ou

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - H(\boldsymbol{\theta}_0)^{-1}U(\boldsymbol{\theta}_0). \tag{2.56}$$

onde, $H(\boldsymbol{\theta})$ é a matriz de derivadas parciais de segunda ordem negativas de $l(\boldsymbol{\theta})$, também conhecida como matriz Hessiana.

Da expressão (2.56) pode ser obtido o processo iterativo de Newton-Raphson:

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - H(\boldsymbol{\theta}_j)^{-1}U(\boldsymbol{\theta}_j). \tag{2.57}$$

em que é iniciado com o valor $\boldsymbol{\theta}_0$ e então um novo valor de $\boldsymbol{\theta}_1$ é obtido a partir da expressão (2.57) e assim consecutivamente, até que o processo se estabilize.

A matriz Hessiana é definida por:

$$H_{ij}(\mu,\sigma) = \begin{bmatrix} \frac{\partial^2}{\partial\mu^2}l(\mu,\sigma) & \frac{\partial^2}{\partial\mu\partial\sigma}l(\mu,\sigma) \\ \frac{\partial^2}{\partial\sigma\partial\mu}l(\mu,\sigma) & \frac{\partial^2}{\partial\sigma^2}l(\mu,\sigma) \end{bmatrix}$$

Considerando a distribuição Weibull, temos que os elementos (i,j) da matriz Hessiana são dados por:

$$\frac{\partial^2}{\partial\mu^2}l(\mu,\sigma) = \frac{p\sigma}{\mu^2} + \sigma(-\sigma-1)\mu^{-\sigma-2}\sum_{i=1}^{n} {t_i}^{\sigma},$$

$$\frac{\partial^2}{\partial\sigma^2}l(\mu,\sigma) = -\frac{p}{\sigma^2} - \sum_{i=1}^{n}\left(\frac{t_i}{\mu}\right)^{\sigma}\left[\log\left(\frac{t_i}{\mu}\right)\right]^2,$$

e

$$\frac{\partial^2}{\partial\mu\partial\sigma}l(\mu,\sigma) = \frac{\partial^2}{\partial\sigma\partial\mu}l(\mu,\sigma) = -\frac{p}{\mu} + \frac{1}{\mu}\sum_{i=1}^{n}\left(\frac{t_i}{\mu}\right)^{\sigma} + \frac{\sigma}{\mu}\sum_{i=1}^{n}\left(\frac{t_i}{\mu}\right)^{\sigma}\log\left(\frac{t_i}{\mu}\right).$$

Os elementos (i, j) da matriz U foram apresentados nas expressões (2.50) e (2.51) considerando a distribuição Weibull. A matriz U é dada por:

$$U_{ij}(\mu,\sigma) = \begin{bmatrix} \frac{\partial}{\partial\mu}l(\mu,\sigma) \\ \frac{\partial}{\partial\sigma}l(\mu,\sigma) \end{bmatrix}$$

De acordo com Bolfarine e Sandoval (2010), o estimador de máxima verossimilhança é obtido quando $\mid \boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j \mid < \varepsilon$, em que ε é o erro na estimação, ou seja, quando a diferença entre as iterações é menor que um erro ε.

O procedimento para a estimação dos parâmetros das demais distribuições é análogo ao apresentado para a distribuição Weibull.

2.6 PACOTE GAMLSS

Rigby e Stasinopoulos (2005) desenvolveram uma nova classe de modelos estatísticos de regressão paramétricos ou semi-paramétricos, denominada de modelos aditivos generalizados para posição, escala e forma (GAMLSS), com a finalidade de superar as falhas dos modelos propostos até então. Tais como os modelos lineares generalizados (GLM) proposto por Nelder e Wedderburn (1972) e os modelos aditivos generalizados (GAM), proposto por Hastie e Tibshirani (1990). Ambas são técnicas de modelagem de regressão univariada muito utilizadas na prática.

Tanto o GLM quanto o GAM assumem que a distribuição da variável resposta pertence à família exponencial e apenas sua média é modelada a partir das variáveis explicativas. Porém, no GAMLSS a variável resposta pode ter qualquer distribuição e, além disso, permite a modelagem não apenas da média, mas de todos os parâmetros da distribuição condicional da variável resposta. De acordo com Stasinopoulos e Rigby (2007), todos os parâmetros da distribuição podem ser modelados como funções paramétricas (lineares ou não-lineares) e/ou funções não-paramétricas

suavizadas de variáveis explicativas (splines cúbicas, splines penalizadas, loess) e/ou efeitos aleatórios.

Segundo Florencio (2010), os GAMLSS são paramétricos pela necessidade de uma distribuição paramétrica para a variável resposta e também semi-paramétricos por consentirem que a modelagem dos parâmetros da distribuição e das funções das variáveis explicativas possam envolver o uso de funções de suavização não-paramétricas.

De acordo Paiva, Freire e Cecatti (2008), no pacote GAMLSS é possível utilizar tanto distribuições contínuas (Normal, Log-normal, Exponencial, Gamma, Beta, BoxCox Power Exponencial, BoxCox t, BoxCox Cole Green, Gumbell, Johnson's SU, Weibull, etc.), como distribuições discretas (Poisson, Binomial, Beta Binomial, etc.) ou mistas. Desse modo, por meio do pacote GAMLSS pode ser realizado o ajuste de variáveis de diversas distribuições de probabilidade, entre elas estão as distribuições abordadas na seção 3.4.

O estimador de máxima verossimilhança é utilizado nesse pacote para realizar o ajuste dos parâmetros dos modelos, como mostra Stasinopoulos e Rigby (2007). Existem dois algoritmos básicos utilizados na busca das estimativas dos parâmetros: o algoritmo CG, sigla referindo as iniciais de Cole e Green, o qual é uma generalização do algoritmo de Cole e Green (1992) e o algoritmo RS, sigla referindo as iniciais de Rigby e Stasinopoulos, que é uma generalização do algoritmo usado por Rigby e Stasinopoulos (1996). Também existe a opção *mixed*, ou seja, um método que mistura os algoritmos CG e RS, começando com RS e terminando o procedimento com o CG.

Segundo Florencio (2010), o algorítimo CG utiliza a primeira derivada e o valor esperado ou aproximado das derivadas de segunda ordem e das derivadas cruzadas do logaritmo da função de verossimilhança em relação aos parâmetros da distribuição, enquanto o algorítimo RS não utiliza o valor esperado das derivadas cruzadas no ajuste da média e da dispersão de modelos aditivos.

De acordo com Vasconcelos (2017), o algoritmo CG pode ser melhor para distribuições com estimativas de parâmetros correlacionados. Enquanto o algoritmo RS é um método mais rápido quando se trata de um grande conjunto de dados e também não necessita de valores iniciais precisos nos parâmetros do modelo para que a convergência seja garantida.

Maiores detalhes sobre os algoritmos CG e RS podem ser verificados no trabalho de Nakamura (2016).

O pacote GAMLSS possibilita realizar a seleção de modelos automaticamente. Esse pacote utiliza o critério de informação de Akaike (AIC), descrito em Akaike (1974), definido da seguinte forma:

$$AIC = -2\left[log(L(\widehat{\boldsymbol{\theta}}))\right] + 2k,$$

em que $L(\widehat{\boldsymbol{\theta}})$ é o máximo da função de verossimilhança e k é o número de parâmetros a serem estimados no modelo.

Para a comparação de dois ou mais modelos, o pacote além de retornar os valores do AIC, também retorna os valores do critério Bayesiano de Schwarz (SBC) disponível em Schwarz et al. (1978), o qual é dado por:

$$SBC = -2\left[log(L(\widehat{\boldsymbol{\theta}}))\right] + k\left[log(n)\right],$$

em que k é o número de parâmetros e n é o tamanho da amostra. Para que se tenha um melhor modelo, basta verificar aquele que possui o menor AIC ou SBC.

O pacote GAMLSS também retorna o Global Deviance (GD), o qual também pode ser utilizado para avaliar o modelo que possui melhor ajuste. Esse critério mede a falta de ajuste entre o modelo ajustado e o dados reais previstos e, portanto, é preferido um modelo com menor desvio global da previsão. O GD é dado por:

$$GD = -2\left[log(L(\widehat{\boldsymbol{\theta}}))\right].$$

Para Florencio (2010) uma vantagem da utilização do GAMLSS refere-se à facilidade de acesso a programas de livre distribuição, como o ambiente de programação no software R CORE TEAM (2018). O GAMLSS está implementado em pacotes no R CORE TEAM (2018) e possibilita o ajuste de muitas distribuições de probabilidade, entre elas as distribuições apresentadas neste estudo.

2.7 QUALIDADE DO MODELO

De acordo com Vasconcelos (2017), a análise de resíduos é uma técnica muito importante para a avaliação da qualidade de um modelo, visto que permite detectar as observações que

podem ser discrepantes, bem como identificar as suposições do ajuste que podem afetar os resultados inferenciais, tais como normalidade, homocedasticidade e independência.

Segundo Ramires et al. (2017), para verificar os desvios das suposições do ajuste e a presença de observações discrepantes, pode-se utilizar as ferramentas de diagnóstico do pacote GAMLSS. A primeira técnica do GAMLSS compõe-se do resíduos quantílicos normalizados randomizados, de acordo com Rigby e Stasinopoulos (2009), a verificação da qualidade do modelo é observada quando os resíduos seguem uma distribuição normal padrão.

O GAMLSS também retorna os gráficos dos resíduos versus valores ajustados, resíduos versus índice, ou seja, o resíduo referente a cada um dos indivíduos numerados, os quais podem indicar um bom ajustamento para o modelo caso não exista padrão para os resíduos.

Outra técnica que pode ser utilizada, descrita por Buuren e Fredriks (2001), é *W*orm Plot (WP). Este gráfico de resíduos é utilizado para a identificação de regiões ou intervalos de uma variável explicativa dentro da qual o modelo não se ajusta adequadamente aos dados. A inadequação de um modelo pode ser percebida no momento em que alguns pontos plotados estão fora das faixas de confiança do gráfico. De acordo com Rigby et al. (2019), este gráfico também pode identificar discrepâncias, tais como *outliers*, os quais podem atrapalhar o ajuste do modelo.

Uma ferramenta utilizada para a verificação da validade da pressuposição de normalidade dos resíduos é o gráfico quantil-quantil ou qq-norm, proposto Wilk e Gnanadesikan (1968).

Em particular, se intenção for investigar se densidade é normal, o gráfico qq-plot será chamado de qq-norm. Esta técnica consiste em verificar se os resíduos são normalmente distribuídos, pressuposição necessária para adequação de um modelo. Caso os pontos do deste gráfico formem aproximadamente uma linha reta, tem-se a sugestão de que os resíduos são normalmente distribuídos. Caso haja pontos que se desviem do comportamento linear, a normalidade pode ser considerada suspeita.

Para ter mais garantia de que os resíduos são normalmente distribuídos, pode-se realizar testes de normalidades, entre eles encontra-se o teste de Shapiro-Wilk, proposto por Shapiro e Wilk (1965).

A avaliação da adequação do modelo também pode ser verificada por técnicas gráficas que utilizam os diferentes resíduos, tais como resíduos de Cox-Snell, proposto por Cox e Snell (1968), e resíduos deviance, introduzidos por Therneau, Grambsch e Fleming (1990).

O resíduo de Cox-Snell pode ser aplicado para qualquer modelo paramétrico, sendo representado por e_i, para o i-ésimo indivíduo cujo tempo de sobrevivência foi observado, sendo este de censura ou de falha, é definido como:

$$\widehat{e_i} = \widehat{\Lambda}(t_i) = -\log\ \widehat{S}(t_i), \qquad i = 1, 2, ..., n$$

em que $\widehat{S}(t)$ é a função de sobrevivência estimada no tempo t e $\widehat{\Lambda}(\cdot)$ a função de taxa de falha acumulada obtida do modelo ajustado.

De acordo com Lee e Wang (2003), uma importante propriedade dos resíduos de Cox-Snell é que se o modelo selecionado se ajusta aos dados, $\widehat{e_i}$ segue uma distribuição exponencial padrão, de modo que o gráfico do resíduo de Cox-Snell $\widehat{e_i}$ versus a taxa de falha acumulada $\widehat{\Lambda}(\widehat{e_i})$ seja uma linha reta com inclinação 1.

Os resíduos deviance são definidos como:

$$\widehat{d_i} = \text{sinal}(\widehat{m}_i)\left[-2\left(\widehat{m}_i + \delta_i log(\delta_i - \widehat{m}_i)\right)\right]^{1/2},$$

em que $\widehat{m}_i$ são os resíduos martingal, definidos por:

$$\widehat{m}_i = \delta_i - \widehat{e}_i,$$

com δ_i sendo a variável indicadora de falha ou censura e $\widehat{e}_i$ os resíduos de Cox-Snell. Os resíduos de desvio são distribuídos simetricamente em torno de zero quando o modelo ajustado é adequado (SUNG; TANAKA, 2004).

3 METODOLOGIA

Para atingir os objetivos propostos, foi realizado um estudo quantitativo e exploratório mediante a análise dos dados de 56 pacientes com câncer de pulmão tratados na Santa Casa de Misericórdia da cidade de uma cidade do sul de Minas Gerais, no período de junho de 2014 a julho de 2017. Para a análise desse conjunto de dados admitiu-se como variável resposta o tempo de vida dos pacientes com câncer de pulmão juntamente com oito variáveis explicativas, sendo essas: gênero, idade, histologia, cigarro, TNM, terapia realizada, ocorrência metástase e cirurgia.

As categorias das variáveis explicativas, contidas nos prontuários, são apresentadas na Tabela 1.

Tabela 1 – Descrição e classificações das variáveis explicativas obtidas nos prontuários dos pacientes com câncer de pulmão, no momento do diagnóstico, avaliadas no estudo do tempo até o óbito dos pacientes com câncer de pulmão.

Variáveis do Prontuário	Categorias	Descrição
histologia	SQC	carcinoma de células escamosas
	ADC	adenocarcinoma
	SCLC	carcinoma de células pequenas
	outros	
	sem informação	
TNM	Estágio I	
	Estágio II	
	Estágio III	
	Estágio IV	
	não disponível	
terapia	nenhuma	
	QT	quimioterapia
	CIR	cirurgia
	RT	radioterapia
	CQR	cirurgia + quimioterapia + radioterapia
	CRT	cirurgia + radioterapia
	QTCIR	quimioterapia + cirurgia
	QR	quimioterapia + radioterapia
	não disponível	
metástase	presença	possui metástase
	ausência	não possui metástase

Fonte: Da autora.

Na Tabela 1 o TNM é um sistema de classificação tumoral que a cada tipo de câncer é atribuída uma letra ou número para descrever T indica as características do tumor primário,

N indica metástase para os linfonodos regionais, quando o câncer que se espalhou para os linfonodos próximos, e M refere-se à presença ou ausência de metástases distantes, quando o câncer se espalhou para partes distantes do organismo, o qual fornece o estadiamento do câncer de pulmão.

Os dados desse estudo são caracterizados pela presença de censura, deste modo tem-se como variável resposta o tempo de vida desde o diagnóstico do câncer até o óbito. Em que a falha é o óbito do paciente, representado por 1, e 0 representa a censura (não óbito do paciente). As informações contidas nesse conjunto de dados foram coletadas no momento do preenchimento da ficha de cadastro de cada paciente.

As categorias das variáveis explicativas são apresentadas na Tabela 2.

Tabela 2 – Descrição e classificações das variáveis explicativas, no momento do diagnóstico, avaliadas no estudo do tempo até o óbito dos pacientes com câncer de pulmão.

Variável	Descrição	Categorias
X_1	gênero	Feminino Masculino
X_2	idade	idade, em anos, dos indivíduos variando de 46 a 87 anos
X_3	histologia	SQC ADC SCLC
X_4	cigarro	fumante não fumante
X_5	TNM	estágio II estágio III estágio IV
X_6	terapia	CIR QR QT QTCIR
X_7	metástase	presença ausência

Fonte: Da autora.

Para a estimação dos parâmetros dos modelos ajustados, foi utilizado o método de estimação de Máxima Verossimilhança apresentada na subseção 3.4.1. Os cálculos foram

realizados utilizando-se o sistema computacional estatístico R, versão 3.5.1, (R Core Team, 2018) com o auxílio do pacote GAMLSS proposto por Stasinopoulos e Rigby (2007).

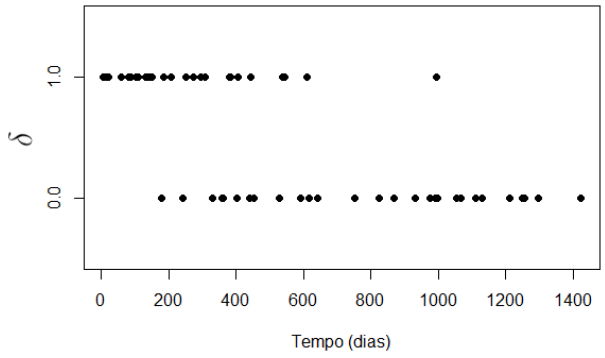
Para realizar o ajuste dos modelos com o auxílio do pacote GAMLSS foi considerado o método RS, visto que este não possui tantas restrições quanto os demais, e deste modo pode-se garantir a convergência.

A seleção do melhor modelo foi realizada por meio dos critérios AIC, SBC e GD, descritos na subseção 3.6. Na avaliação da qualidade do modelo ajustado foram realizadas análises gráficas e testes para verificar a independência e a distribuição dos resíduos, descritos na subseção 3.7, considerando um nível de significância de 5%. A qualidade do modelo, também, foi avaliada por meio dos resíduos de Cox-Snell, deviance e análise do *worm plot*.

4 RESULTADOS E DISCUSSÕES

Inicialmente foi realizada uma análise descritiva do conjunto de dados em estudo. Foi possível constatar que 52% das observações são censuradas e 48% representam as observações de falha. Desse modo, verifica-se que a maioria das informações são cesuradas, o que caracteriza uma deficiência nos dados, visto que maioria das observações não apresentaram falha. Na Figura 2 são apresentados os tempos de falha ou de censura dos pacientes com câncer de pulmão.

Figura 2 – Tempos de censura $(\delta = 0)$ ou de falha $(\delta = 1)$ observados nos pacientes com câncer de pulmão.



Fonte: Da autora.

Na Figura 2 nota-se que maioria dos tempos de falha ocorreram até, aproximadamente, 600 dias e uma única observação apresentou falha em 995 dias. Os tempos de censura variam de 200 a 1423 dias, sendo estas censuras à direita. Também é possível notar que as falhas concentram-se mais no inicio dos tempos (a esquerda do gráfico). Observa-se, ainda, que este comportamento não é verificado para os tempos de censura, estando estes ao londo de todo o período observado.

Foi realizada uma análise dos tempos, em dias, separados em dois grupos, o primeiro contendo os tempos censurados e o segundo os tempos de falha. Os resultados podem ser observados na Tabela 3.

Tabela 3 – Análise descritiva dos tempos censurados e de falha dos pacientes com câncer de pulmão.

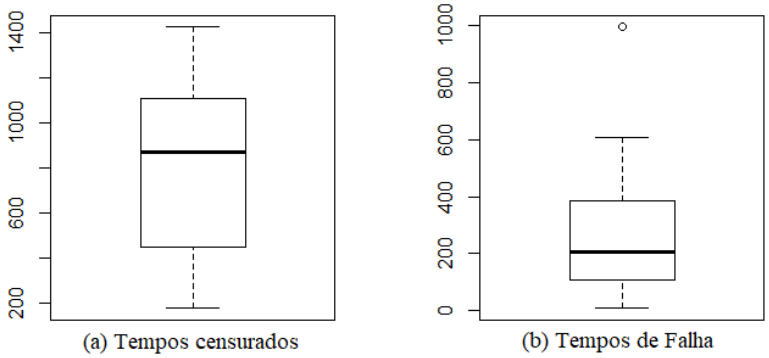
δ	Média	Mediana	Desvio Padrão	Mínimo	Máximo	Coeficiente de Variação (%)
0	807,4	868	362,7	178,0	1423,0	44,9
1	263,9	207,0	224,0	7,0	995,0	84,9

Fonte: Da autora.

Pela Tabela 3 observa-se que o tempo médio de vida dos pacientes foi menor para aqueles que apresentaram falha. O menor tempo até a morte do paciente foi 7 dias, sendo este um

tempo de falha. O maior tempo registrado foi um censurado, sendo este 1423 dias. Os tempos censurados estão entre 178 e 1423 dias, com desvio padrão de 362,7 dias. Nota-se por meio do Coeficiente de Variação que o grupo dos tempos de falha é mais heterogênio que o grupo dos tempos censurados. Enquanto os tempos de falha estão entre 7 e 995 dias, com desvio padrão de 224,0 dias. Na Figura 3 é apresentado o gráfico Boxplot referente aos tempos de falha e censurados.

Figura 3 – Gráfico Boxplot dos 29 pacientes apresentando censura (a) e 27 pacientes apresentando falha (b), ambos pertencentes ao grupo de pacientes com câncer de pulmão.

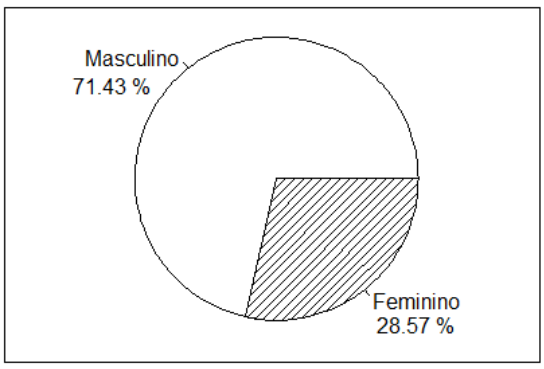


Fonte: Da autora.

Verifica-se que a mediana das observações censurados é maior que a mediana das observações de falha. Em ambos os tempos é possível constatar uma distribuição assimétrica. Em relação aos tempos censurados não há presença de *outliers*, porém analisando os tempos de falha observa-se a presença de um possível ponto discrepante ou candidato a *outlier*. A variabilidade dos tempos censurados é maior que dos tempos de falha.

Na Figura 4 observa-se o percentual de pacientes do gênero masculino e do gênero feminino com câncer de pulmão.

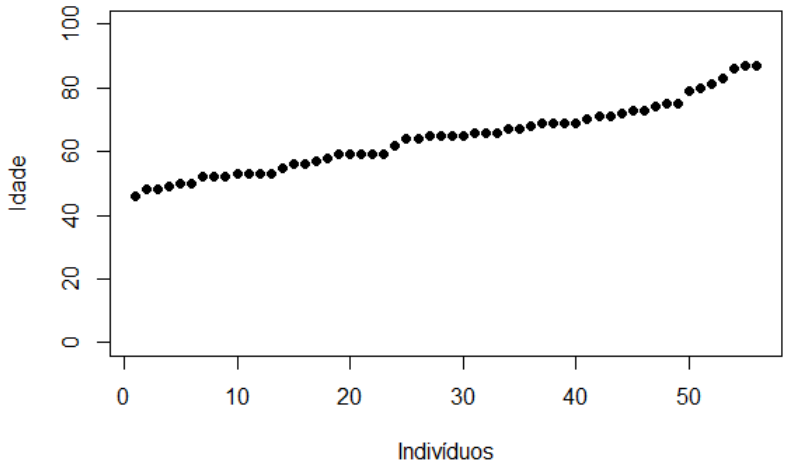
Figura 4 – Gráfico de setores representando o gênero dos pacientes com câncer de pulmão.



Fonte: Da autora.

Dentre os indivíduos observados, verificou-se que maior parte dos pacientes com câncer de pulmão eram do gênero masculino. Na Figura 5 pode ser observado o gráfico de dispersão da idade, em ordem crescente, dos pacientes tratados neste estudo.

Figura 5 – Gráfico dispersão da idade em ordem crescente dos pacientes com câncer de pulmão.

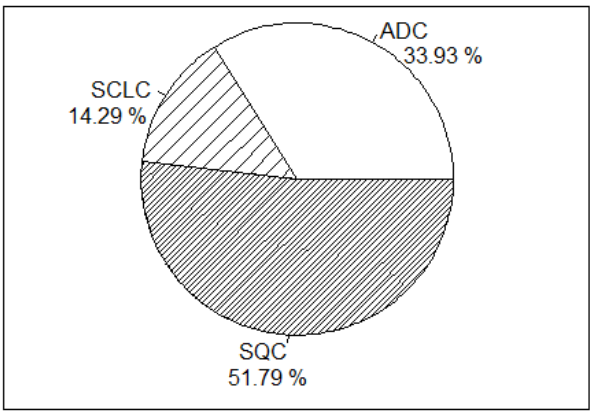


Fonte: Da autora.

Os indivíduos observados apresentavam idade entre 46 e 87 anos, como pode-se notar na Figura 5, sendo a idade média dos indivíduos 64 anos. O paciente mais jovem tinha 46 anos e apresentou tempo de falha de 609 dias, enquanto que o paciente com mais idade possuía 87 anos e tempo de falha de 445 dias.

Em relação a variável explicativa histologia, pôde-se observar na Figura 6 a porcentagem dos indivíduos que possuíam SQC (carcinoma de células escamosas), ADC (adenocarcinoma) e SCLC (carcinoma de células pequena), não houve indivíduos sem confirmação histológica ou que tiveram outro tipo de histologia.

Figura 6 – Gráfico de setores representando a histologia dos pacientes com câncer de pulmão.

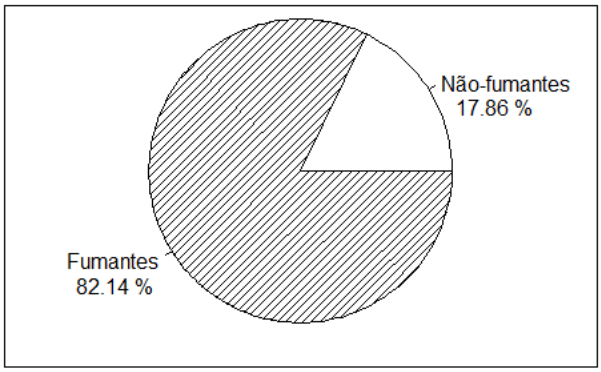


Fonte: Da autora.

Observa-se na Figura 6 que a maioria (51,79%) dos pacientes com a doença em estudo apresentavam no momento do diagnóstico a histologia carcinoma de células pequenas.

Para a variável explicativa cigarro (retorna se o paciente é fumante ou não fumante) observa-se, no momento do diagnóstico, que mais de 80% eram fumantes. Esse resultado pode ser observado na Figura 7, e verifica-se ainda que todos os indivíduos tiveram a informação sobre o uso de cigarro disponível.

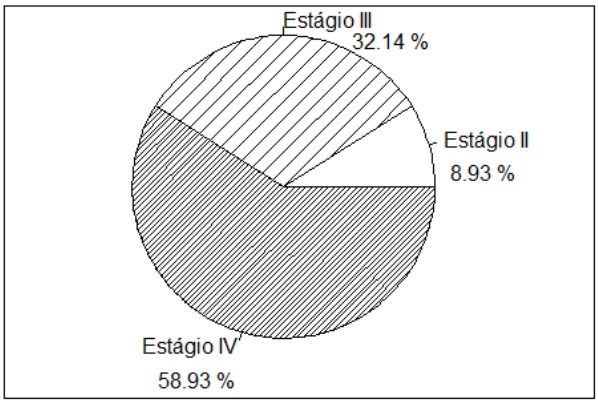
Figura 7 – Gráfico de setores representando o uso de cigarro pelos pacientes com câncer de pulmão no momento do diagnóstico.



Fonte: Da autora.

Em relação a variável explicativa TNM, a qual fornece o estadiamento da doença, observou-se que nenhum dos indivíduos apresentavam câncer de pulmão no estágio I, os percentuais sobre os demais estadiamentos podem ser observados na Figura 8. Todos os indivíduos tiveram a informação sobre o TNM disponibilizada.

Figura 8 – Gráfico de setores representando o estadiamento da doença no momento do diagnóstico.



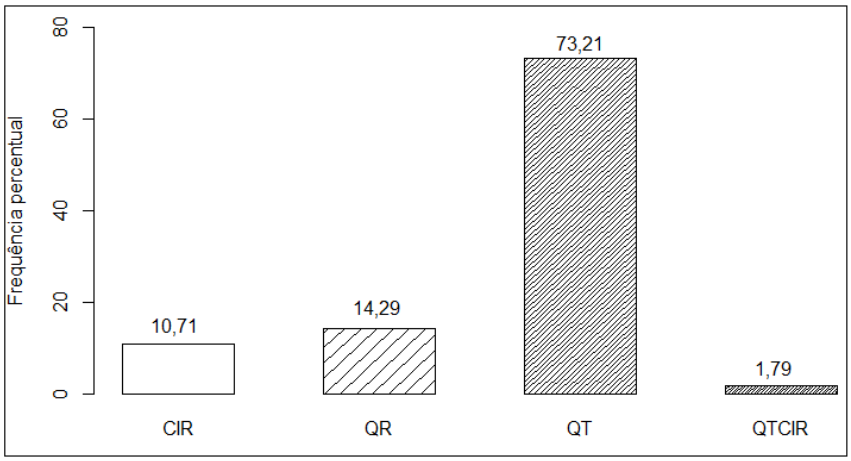
Fonte: Da autora.

Pode-se observar na Figura 8 que maior parte (58,93%) dos indivíduos adoecidos possuíam câncer de pulmão no estágio IV. Esse resultado é algo característico dessa doença, pois

segundo Barros et al. (2006), na maioria dos casos o diagnóstico é tardio, ou seja, quando a doença se encontra avançada localmente e/ou disseminada, tendo em vista que tumores iniciais não costumam produzir sintomas que justifiquem a investigação.

Observou-se também o tipo de terapia realizada pelos pacientes no momento do diagnóstico é a quimioterapia, como pode ser visto na Figura 9. Nenhum indivíduo teve como terapia inicial a radioterapia, cirurgia juntamente com a quimioterapia e radioterapia, cirurgia juntamente com a radioterapia. Todos os indivíduos tiveram essa informação disponibilizada.

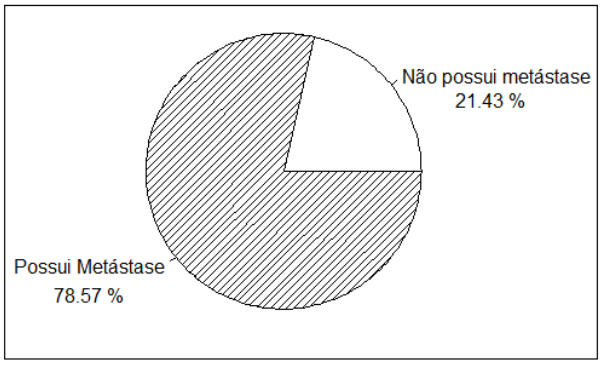
Figura 9 – Gráfico de barras representando a terapia dos pacientes com câncer de pulmão no momento do diagnóstico, sendo QT (quimioterapia), CIR (cirurgia), QTCIR (quimioterapia + cirurgia) e QR (quimioterapia + radioterapia).



Fonte: Da autora.

Analisando a variável explicativa metástase, observou-se que todos os indivíduos apresentavam a informação sobre a presença de metástase disponível. O percentual de presença e ausência de metástase pode ser visualizado na Figura 10.

Figura 10 – Gráfico de setores representando a presença de metástase nos pacientes adoecidos no momento do diagnóstico.

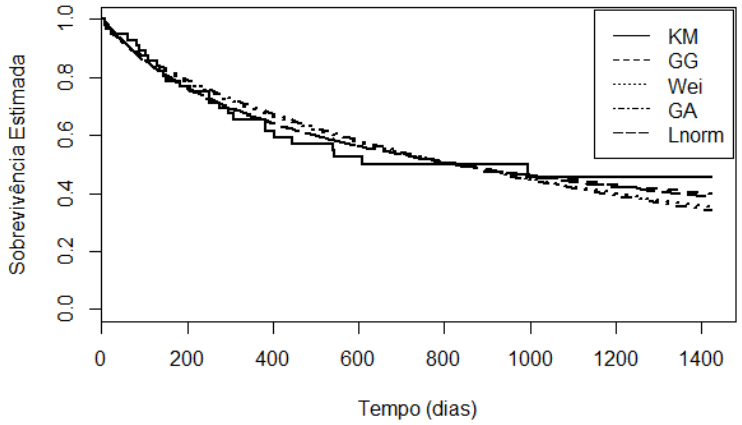


Fonte: Da autora.

Pela Figura 10 verifica-se que maioria (78,57%) dos pacientes observados apresentavam metástase no momento do diagnóstico, fato relacionado com o diagnóstico tardio da doença.

Após a análise descritiva, realizou-se o ajuste do modelo ao tempo de sobrevida dos pacientes diagnosticados com câncer de pulmão por meio das técnicas de análise de sobrevivência. Primeiramente, foram plotadas as curvas de sobrevivência das distribuições Weibull, Log-normal, Gama e Gama Generalizado, juntamente com a curva de Kaplan-Meier, apresentados na Figura 11. O modelo Exponencial não foi investigado visto que este não consegue descrever a variável resposta deste estudo, pois este modelo possui taxa de falha constante e, é sabido que a taxa de falha para pacientes portadores dessa patologia não é constante.

Figura 11 – Funções de Sobrevivência estimadas e curva de Kaplan-Meier



Fonte: Da autora.

Na Figura 11, por meio das distâncias da curva de Kaplan-Meier com as demais funções, pode-se observar que as funções de sobrevivência estimadas estão próximas da curva de Kaplan-Meier, e portanto torna-se difícil selecionar o modelo que melhor se ajustou utilizando a análise gráfica.

Tendo em vista que a análise gráfica não é suficiente para escolher um modelo que descreva o tempo até o óbito do paciente, na Tabela 4 são apresentados os valores estatísticos dos critérios GD, AIC e o SBC, utilizados para avaliar qual o melhor modelo.

Tabela 4 – Tabela contendo os valores estatísticos do AIC, do SBC e do GD dos modelos completos analisados.

Modelo	AIC	SBC	GD
Weibull	421,253	433,405	409,253
Gama	406,959	423,162	390,959
Gama Generalizada	372,292	392,545	352,292
Log-normal	420,694	432,846	408,694

Fonte: Da autora.

Pode-se observar que o modelo Gama Generalizado apresenta os menores valores estatísticos GD, o AIC e o SBC, disponíveis na Tabela 4. Logo tem-se que este é o modelo mais apropriado para descrever a variável resposta em estudo.

Na Tabela 5 são apresentadas as estimativas, erros padrões e os valores de p obtidos para o modelo Gama Generalizado ajustado através do GAMLSS. Os parâmetros foram selecionados considerando-se um nível de 5% de significância.

Tabela 5 – Tabela contendo as estimativas dos parâmetros, os nomes das variáveis explicativas, com seus respectivos erros padrões e valores p.

	Variáveis Explicativas	Parâmetros	Estimativas	Erros padrões	valores p
$log(\mu)$	intercepto	β_{01}	8,668	0,356	$< 0,0001$
	gênero (feminino)	β_{11}	-0,401	0,103	$< 0,0001$
	idade	β_{21}	-0,022	0,004	$< 0,0001$
	estágio III	β_{61}	-0,856	0,179	$< 0,0001$
	estágio IV	β_{71}	-1,350	0,170	$< 0,0001$
$log(\sigma)$	intercepto	β_{02}	-1,090	0,069	$< 0,0001$
r	intercepto	β_{03}	28,770	6,379	$< 0,0001$
	histologia (SCLC)	β_{33}	-20,848	4,148	$< 0,0001$
	histologia (SQC)	β_{43}	-0,399	3,565	$0,9113$
	cigarro (fumante)	β_{53}	-16,880	6,108	$0,0079$

Fonte: Da autora.

Após eleger o modelo Gama Generalizado como sendo o modelo que melhor descreve a variável de interesse, tem-se que:

$$\begin{aligned}\mu =& \exp\{8,668 - 0,401(\text{se } X_1 = \text{feminino}) - 0,022X_2 - 0,856(\text{se } X_5 = \text{III}) \\ & - 1,350(\text{se } X_5 = \text{IV})\},\end{aligned}$$

$$\sigma = \exp\{-1,090\}$$

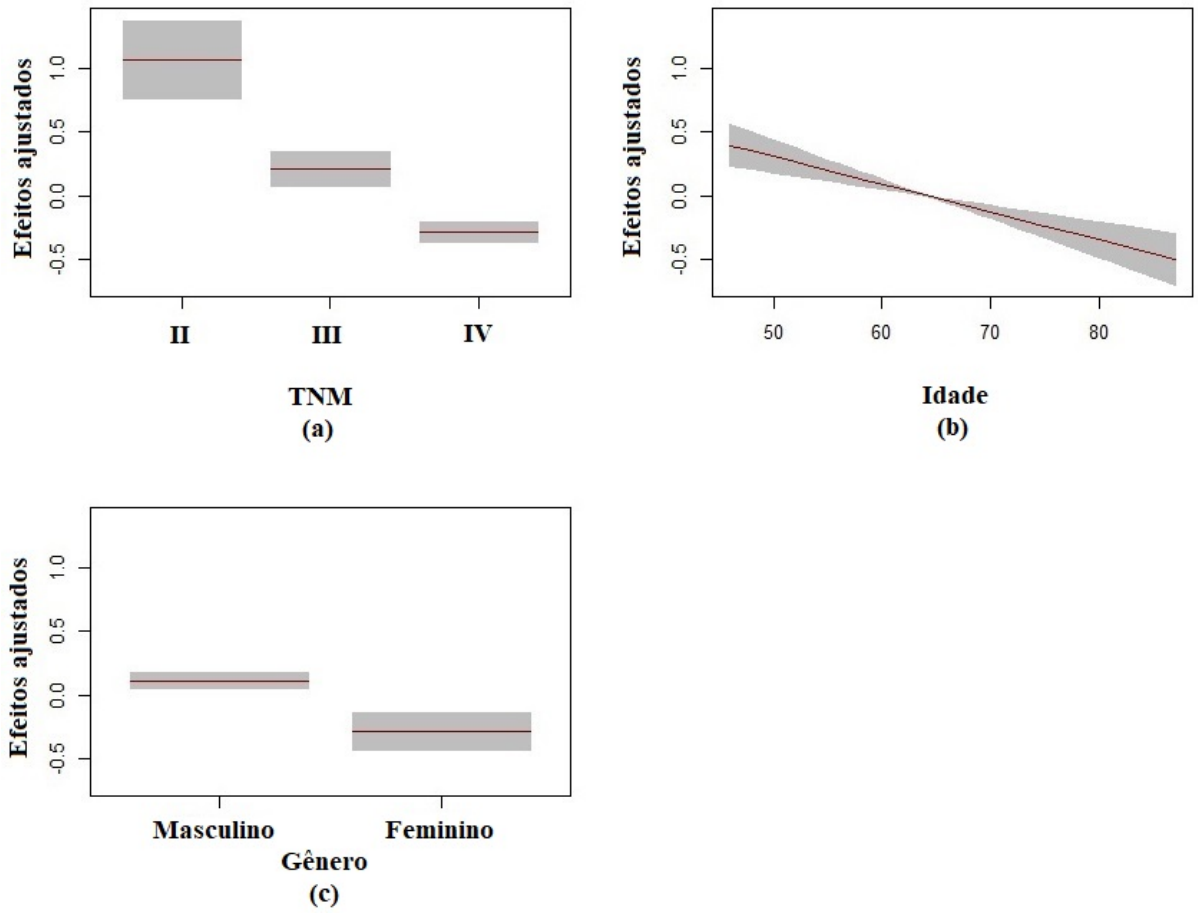
e

$$\begin{aligned}r =& 28,700 - 20,848(\text{se } X_3 = \text{SCLC}) - 0,399(\text{se } X_3 = \text{SQC}) \\ & - 16,880(\text{se } X_4 = \text{fumante}),\end{aligned}$$

Por meio dos resultados obtidos e apresentados na Tabela 5, verifica-se quais variáveis associadas aos parâmetros que influenciam no tempo de vida dos pacientes com câncer de pulmão

(valor-p $< 5\%$), são estas gênero, idade, TNM, histologia e cigarro. Sendo que gênero, idade e TNM tiveram efeito em μ como pode ser visto na Figura 12.

Figura 12 – Grafico com os efeito parcial das variáveis explicativas em μ.



Fonte: Da autora.

Na Figura 12 (a) pode-se analisar o efeito da variável explicativa TNM em μ, e notar que quanto mais avançado é o estágio do câncer menor é seu efeito no referido parâmetro em relação ao estágio II, ou seja, se o paciente possui câncer de pulmão no estágio III o valor do parâmetro tem redução de 0,856 e se o paciente possui câncer de pulmão no estágio IV o valor do parâmetro tem redução de 1,350.

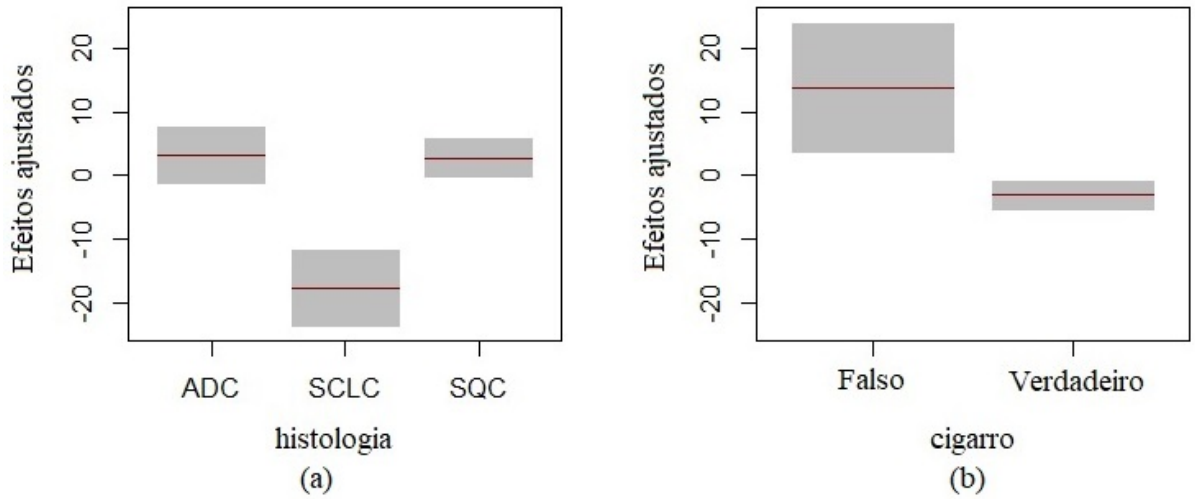
Na Figura 12 (b) pode-se observar o efeito da variável explicativa idade em μ, em que quanto maior a idade do paciente no momento do diagnóstico do câncer de pulmão menor é seu efeito no parâmetro, ou seja, a cada ano de vida a mais que o paciente possui no momento do diagnóstico, ocasiona uma redução de 0,022 no referido parâmetro.

Na Figura 12 (c) é exposto o efeito da variável explicativa gênero para o parâmetro μ,

e assim pode-se concluir que gênero feminino têm menor efeito no parâmetro em questão em relação ao gênero masculino, isto é, um paciente com câncer de pulmão do gênero feminino reduz o valor do parâmetro μ em 0,401.

Na Figura 13, pode-se observar o efeito parcial das variáveis explicativas histologia e cigarro em r.

Figura 13 – Gráfico com os efeito parcial das variáveis explicativas em r.



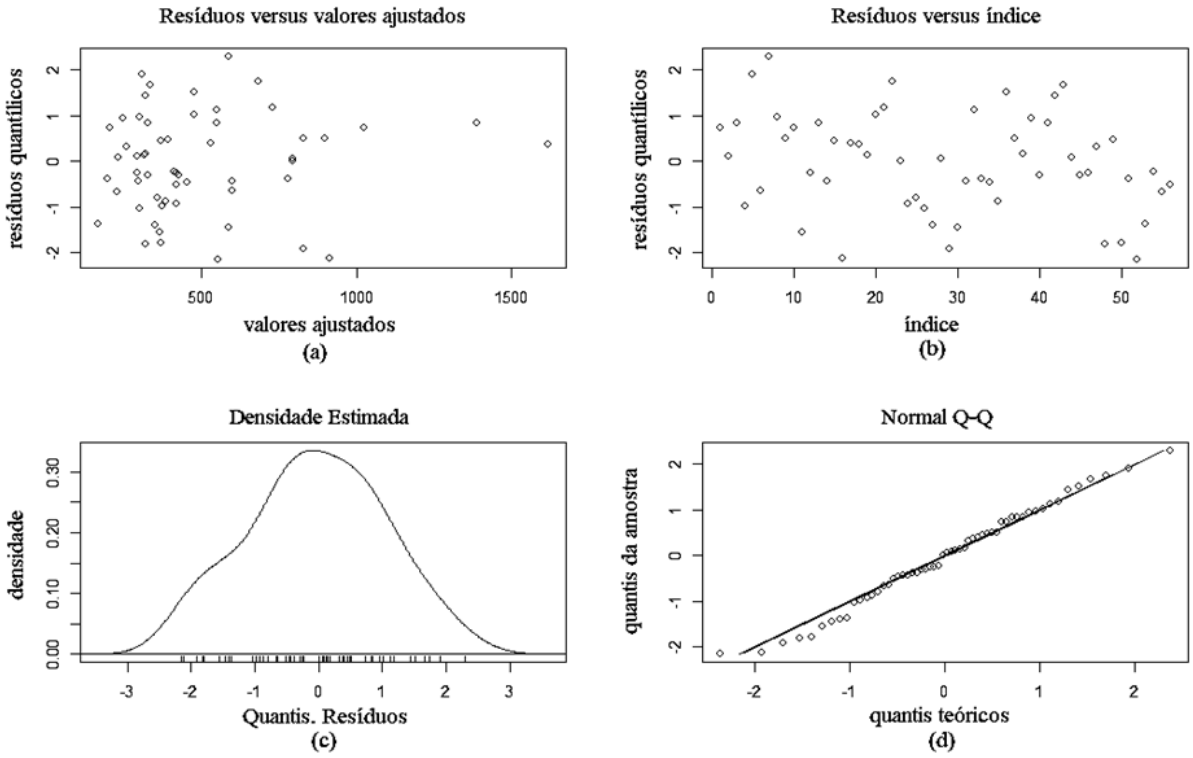
Fonte: Da autora.

Na Figura 13 (a) pode-se analisar o efeito da variável explicativa histologia em r, e notar que tanto a histologia SCLC quando SQC possuem menor efeito sobre o parâmetro r em relação à histologia ADC, mais precisamente considerando um paciente com câncer de pulmão que possui a histologia SCLC tem redução de 20,848 no valor do parâmetro e o paciente com a mesma doença que possui a histologia SQC tem redução de 0,399 no valor do parâmetro r.

Na Figura 13 (b) pode-se observar o efeito da variável explicativa cigarro em r, em que o paciente que faz uso do cigarro possui efeito negativo em relação ao paciente que não faz uso do cigarro, sendo que esse efeito diminui 16,880 o valor do parâmetro r.

Finalizado o ajuste do modelo, deu-se início a avaliação da qualidade do modelo Gama Generalizado. A Figura 14 apresenta alguns gráficos que auxiliam na verificação da qualidade e das suposições do modelo ajustado.

Figura 14 – Resíduo do modelo de regressão Gama Generalizado.



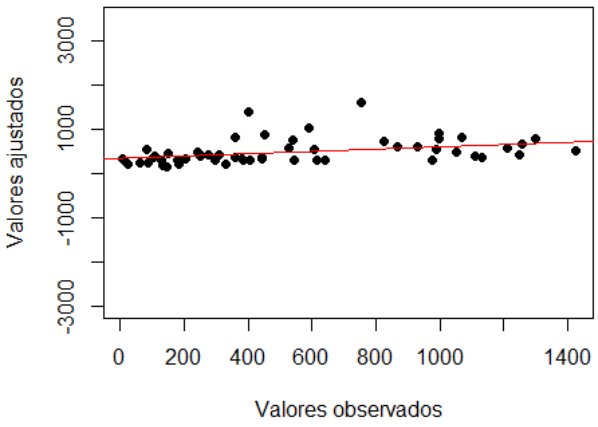
Fonte: Da autora.

Na Figura 14 (a) nota-se que os resíduos não apresentam um padrão e na Figura 14 (b) tem-se que os pontos do gráfico, aparentemente, distribuem-se de maneira aleatória, ambas as informações sugerem que os resíduos são independentes. Na Figura 14 (c) pode-se notar que os resíduos aparentemente seguem uma distribuição Normal. Também espera-se que os pontos do gráfico dos quantis amostrais dos resíduos versus os quantis teóricos dos resíduos sigam o comportamento de uma reta. Observamos na Figura 14 (d) uma situação satisfatória. Portanto, existem indícios de que os erros são normalmente distribuídos.

A normalidade residual ainda foi avaliada por meio do teste de Shapiro-Wilk, pelo qual se obteve um valor $p = 0,774$. Deste modo, conclui-se que os resíduos apresentam distribuição normal, evidenciando um bom ajuste do modelo.

Na Figura 15 é apresentado o gráfico de valores observados versus valores ajustados.

Figura 15 – Gráfico dos Valores observados versus Ajustados pelo modelo Gama Generalizado.

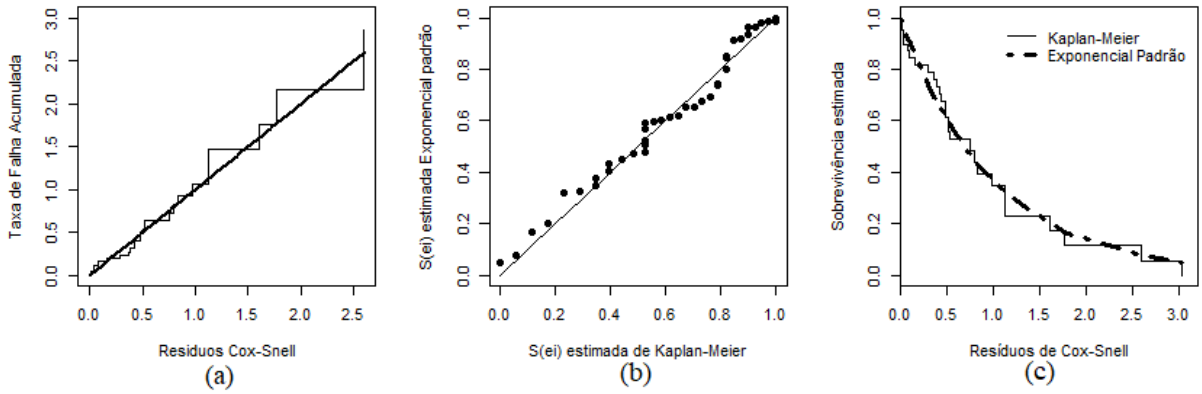


Fonte: Da autora.

O gráfico de valores observados versus valores ajustados, Figura 15, mostra que os pontos estão em torno de uma reta, sugerindo a adequabilidade do modelo proposto.

Também para avaliar o ajuste do modelo Gama Generalizado aos dados, foram utilizados os resíduos de Cox-Snell. Os gráficos de resíduos de Cox-Snell podem ser observados na Figura 16.

Figura 16 – Análise gráfica dos resíduos de Cox-Snell do modelo Gama Generalizado.



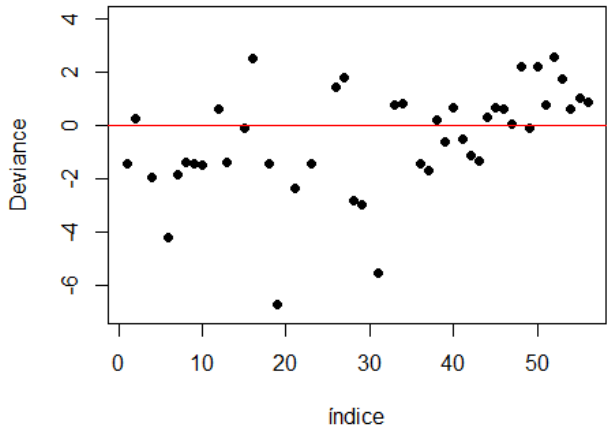
Fonte: Da autora.

Na Figura 16 (a) pode-se observar o gráfico do resíduo de Cox-Snell $\widehat{e_i}$ versus a taxa de falha acumulada $\widehat{\Lambda}(\widehat{e_i})$, a qualidade do modelo pode ser verificada quando se observa aproximadamente uma reta com inclinação 1. Na Figura 16 (b) tem-se o gráfico da função de sobrevivência dos resíduos estimada por Kaplan-Meier versus função de sobrevivência do modelo exponencial padrão, a qualidade do modelo pode ser verificada quando se observa aproximadamente uma reta. Na Figura 16 (c) nota-se a curva de sobrevivência do resíduo de Cox-Snell estimada por Kaplan-Meier e a curva de sobrevivência do modelo exponencial. Como as duas curvas estão

próximas há indícios da boa qualidade do ajuste do modelo Gama Generalizado, como é descrito em Lee e Wang (2003).

Na Figura 17 pode-se observar pelo gráfico de resíduos deviance versus o índice que os resíduos apresentam um comportamento aleatório em torno de zero, o que também indica a adequabilidade do modelo Gama Generalizado.

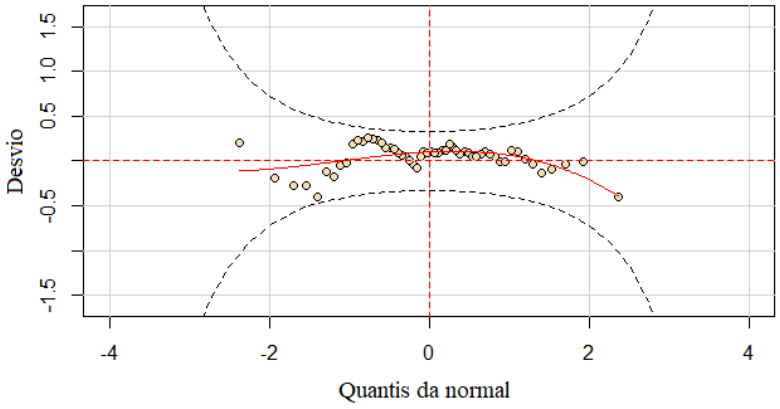
Figura 17 – Gráfico os resíduos deviance versus índice do modelo Gama Generalizado.



Fonte: Da autora.

Na Figura 17 um dos pontos observados, aparentemente, está mais distante que os demais da reta, podendo indicar a presença de um ponto discrepante. Para verificar se esse ponto seria um *outlier* foi observado o Worm-Plot, na Figura 18, e nenhum resíduo ultrapassou os limites de confiança, isto é, a totalidade dos resíduos está na região de aceitação, indicando que o candidato a *outlier* não gera influência sobre o modelo, e assim a existência de um ponto discrepante no conjunto de dados pode ser descartada, e, desse modo, conclui-se pela boa qualidade do modelo ajustado.

Figura 18 – Gráfico Worm-Plot do modelo de regressão Gama Generalizado.



Fonte: Da autora.

Neste estudo identificou-se que os indivíduos acometidos com câncer de pulmão apresentavam uma idade entre 46 anos e 87 anos. Sendo a faixa etária de 50 a 70 anos a com maior incidência da referida doença de acordo com o Instituto Nacional do Câncer (Brasil) (2010). Deste modo, a faixa etária dos pacientes analisados nesse estudo assemelha-se a faixa etária dos pacientes brasileiros.

O Instituto Nacional de Câncer José Alencar Gomes da Silva (2017) ressalta que a maior ocorrência de câncer de pulmão se dá em indivíduos do gênero masculino, resultados esses compatíveis com os encontrados para esse estudo, em outras palavras, os resultados encontrados nos pacientes com câncer de pulmão tratados na Santa Casa de Misericórdia de uma cidade do sul de Minas Gerais condizem com os resultados nacionais.

Assim como no presente estudo, Khaksar et al. (2017) também realizaram um estudo para determinar a sobrevida de pacientes com carcinoma pulmonar de células não pequenas. Os objetivos dos autores foram encontrar fatores prognósticos e determinar um modelo eficiente para descrever a sobrevida dos pacientes. Foram ajustados modelos semi-paramétricos e paramétricos aos dados e os autores verificaram a eficiência do modelo paramétrico Log-normal, pelo critério AIC. Por meio dos resultados os autores concluíram que os fatores tipo de terapia e idade foram significativos para a sobrevida dos pacientes.

Verifica-se, ainda, por meio do estudo de Khaksar et al. (2017) que as variáveis explicativas desse estudo assemelham-se as variáveis explicativas que vem sendo utilizadas na análise de dados em que a variável resposta é o tempo até a ocorrência do óbito por câncer de pulmão. Visto que os autores analisaram as variaveis explicativas idade, tipo de tratamento (cirurgia, quimioterapia ou quimioterapia mais radioterapia), estágio do câncer no diagnóstico (II, III ou IV), histórico de tabagismo, duração do tabagismo, gênero, tipo histopatológico (adenocarcinoma, carcinoma espinocelular ou carcinoma de células grandes), local de residência dos pacientes, ocorrência de morte, período de tempo entre o diagnóstico e o início do tratamento (em meses) e a duração tempo entre o diagnóstico e a morte (em meses).

Zhu et al. (2011) realizaram a comparação do modelo semi-paramétrico de Cox com o modelo paramétrico Weibull para o estudo da sobrevivência de dados de pacientes com câncer gástrico. Por meio dos resultados provenientes dos modelos ajustados os autores verificaram que o histórico familiar, grau histológico, profundidade da invasão do tumor, localização e idade foram fatores prognósticos significativos para o tempo de sobrevida dos pacientes. Concluíram,

ainda, que o modelo paramétrico Weibull apresentou um melhor ajuste que o modelo semi-paramétrico. Desta maneira, esse estudo corrobora com a hipótese de que modelos paramétricos podem fornecer melhores resultados.

Assim como nos trabalhos de Khaksar et al. (2017) e Zhu et al. (2011), no presente estudo também foi identificado que a variável idade foi significativa, e desse modo, é possível perceber que a idade influencia na sobrevivência dos pacientes em diferentes tipos de canceres. De acordo com os mesmos autores os modelos que melhor se ajustaram aos dados foram os paramétricos.

Após a apresentação dos resultados e discussão, realizou-se a verificação da aplicabilidade do modelo ajustado. Para tanto, foram consideradas situações hipotéticas utilizando a função de sobrevivência do modelo Gama Generalizado ajustado para avaliar a sobrevida dos pacientes diagnosticados com câncer de pulmão.

À vista disso, considerando duas pacientes do gênero feminino, estando uma com 50 anos e outra com 80 anos, e ambas possuindo histologia SQC, estágio IV e sendo fumantes, tem-se que a sobrevivência no tempo $t = 550$ dias foi de 55,55% para a paciente com 50 anos e de 25,93% para a paciente com 80 anos. Desse modo, verifica-se que a idade mais avançada retorna uma sobrevida menor para a paciente em questão. Caso sejam consideradas as mesmas características, com as idades de 20 e 95 anos essas sobrevivências seriam de 73,34% e 6,05%, respectivamente.

Neste momento, considerando-se dois pacientes do gênero masculino, um com 50 anos e outro com 80 anos, e ambos possuindo histologia SQC, estágio IV e sendo fumantes, tem-se que a sobrevivência no tempo $t = 550$ dias foi de 67,42% para o paciente com 50 anos e de 45,68% para a paciente com 80 anos. Logo, verifica-se que a idade mais avançada influencia de modo a diminuir a sobrevida do paciente do gênero masculino com câncer de pulmão.

Tomando agora dois pacientes, um do gênero masculino e uma do gênero feminino, ambos com 64 anos de idade, histologia SQC, estágio IV e fumantes, tem-se que a sobrevivência no tempo $t = 550$ dias foi de 58,64% para o paciente do gênero masculino e 43,58% para paciente do gênero feminino. Dessa maneira, é possível verificar que ao se considerar pacientes em iguais condições de diagnóstico, a sobrevivência do paciente do gênero masculino foi maior do que do gênero feminino.

5 CONCLUSÕES

O modelo Gama Generalizado foi o que melhor descreveu o tempo de sobrevida dos pacientes portadores de câncer de pulmão e as covariáveis a ele associados. As variáveis explicativas que foram significativas no tempo de vida desses pacientes foram gênero, idade, TNM, histologia e cigarro. As variáveis explicativas que tiveram efeito no parâmetro μ do modelo, foram gênero, idade e TNM, e as variáveis explicativas que tiveram efeito no parâmetro r foram histologia e cigarro. O parâmetro σ não sofreu efeito de nenhuma covariável.

6 CONSIDERAÇÕES FINAIS

Como trabalhos futuros pretende-se escrever um artigo com o conteúdo deste estudo e realizar um novo estudo utilizando a abordagem Bayesiana.

Entre os frutos deste estudo tem-se o artigo de revisão sistemática intitulado “Modelos de sobrevida aplicados a dados de câncer: revisão sistemática”, no qual é possível encontrar os artigos escritos entre os anos de 2000 e 2018 que utilizam modelos paramétricos no ajuste de dados de câncer. Também tem-se o trabalho intitulado “Aplicação de modelos de sobrevida no tempo de cura de hanseníse”, o qual foi apresentado em formato de pôster no XXXIX O Congresso Nacional de Matemática Aplicada e Computacional que aconteceu em em Uberlândia-MG. E o trabalho intitulado “Utilização do modelo Weibull na descrição do tempo até a cura de hanseníase: uma abordagem Bayesiana”, que foi apresentado no VI Workshop em Análise de Sobrevivência e Aplicações realizado na cidade de Piracicaba-SP.

REFERÊNCIAS

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

ALAN, J. G. Parametric methods in the analysis of survival data. *Microelectron Reliab*, v. 34, p. 477–481, 1980.

ALWAN, A. et al. *Global status report on noncommunicable diseases 2010.* Geneva: World Health Organization, 2011.

AMERICAN CANCER SOCIETY. Cancer facts and figures. Atlanta, 2015.

ARAUJO, L. H. et al. Câncer de pulmão no brasil. *Jornal Brasileiro de mologia*, v. 44, n. 1, 2018.

ARAUJO, M. *Análise de Sobrevivência do Tomateiro–A Phytophthora infestans*. 2008. 53 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) — Universidade Federal de Viçosa, Viçosa, 2008.

BARROS, J. A. et al. Diagnóstico precoce do câncer de pulmão: o grande desafio. *Jornal Brasileiro de Pneumologia*, v. 32, p. 221–227, 2006.

BERNARDES, N. B. et al. Câncer de mama x diagnóstico/breast cancer x diagnosis. *ID on lube Revista Multidisciplinar e de psicologia*, v. 13, n. 44, p. 877–885, 2019.

BLOSSFELD, H. P.; HAMERLE, A.; MAYER, K. U. *Event history analysis: Statistical theory and application in the social sciences*. Nova York: Psychology Press, 2014.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. 2. ed. São Paulo: Sociedade Brasileira de Matemática, 2010.

BOTELHO, F.; SILVA, C.; CRUZ, F. Epidemiologia explicada–análise de sobrevivência. *Acta Urológica*, v. 26, n. 4, p. 33–38, 2009.

BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, Wiley Online Library, v. 68, n. 6, p. 394–424, 2018.

BUCKLEY, J.; JAMES, I. Linear regression with censored data. *Biometrika*, Oxford University Press, v. 66, n. 3, p. 429–436, 1979.

BUSTAMANTE-TEIXEIRA, M. T.; FAERSTEIN, E.; LATORRE, M. R. Técnicas de análise de sobrevida. *Cadernos de Saúde Pública*, SciELO Public Health, v. 18, p. 579–594, 2002.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

COLOSIMO, E.; GIOLO, S. *Análise de sobrevivência aplicada*. São Paulo: Blücher, 2006.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

COX, D. R.; SNELL, E. J. A general definition of residuals. *J.R. Stat. Soc. B*, v. 30, n. 2, p. 248–275, 1968.

FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, The Royal Society London, v. 222, n. 594-604, p. 309–368, 1922.

FLORENCIO, L. de A. *Engenharia de avaliações com base em modelos GAMLSS*. 2010. 125 f. Dissertação (Mestrado em Estatística) — Departamento de Estatística, Universidade Federal de Pernambuco, Recife, 2010.

GARNÉS, S. J. dos A.; SAMPAIO, R. J. B. de; DALMOLIN, Q. Ajustamento paramétrico por mínimos quadrados com análise na estabilidade da solução. *Boletim de Ciências Geodésicas*, v. 2, n. 1, 1997.

GREEN, L. S. et al. Bronchogenic cancer in patients under 40 years old: the experience of a latin american country. *Chest*, Elsevier, v. 104, n. 5, p. 1477–1481, 1993.

GROSS, A. J.; CLARK, V. *Survival distributions: reliability applications in the biomedical sciences*. New York: John Wiley & Sons, 1975.

HASTIE, T.; TIBSHIRANI, R. *Generalized Additive Models*. Taylor & Francis, 1990. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412343902. Disponível em: <https://books.google.com.br/books?id=qa29r1Ze1coC>.

HERMETO, R. T. *Análise de Sobrevivência na Modelagem do Tempo de Vida de Redes de Sensores sem Fio*. 2014. 63 f. Dissertação (Mestrado em Engenharia de Teleinformática) — Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, Fortaleza, 2014.

HOSMER, D. W.; LEMESHOW, S. Applied survival analysis: regression modelling of time to event data. *Eur Orthodontic Soc*, p. 561–2, 1999.

Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2018: incidência de câncer no Brasil*. Rio de Janeiro: INCA, 2017.

Instituto Nacional do Câncer (Brasil). *Câncer no Brasil: dados dos registros de base populacional*. Rio de Janeiro: INCA, 2010.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KHAKSAR, E. et al. Cox regression and parametric models: Comparison of how they determine factors influencing survival of patients with non-small cell lung carcinoma. *Asian Pacific Journal of Cancer Prevention: APJCP*, Shahid Beheshti University of Medical Sciences, v. 18, n. 12, p. 3389, 2017.

KLEINBAUM, D. G.; KLEIN, M. *Survival analysis*. New York: Springer, 2010.

LEE, E.; WANG, J. *Statistical Methods for Survival Data Analysis*. New Jersey: John Wiley & Sons, 2003.

MACHIN, D.; CHEUNG, Y. B.; PARMAR, M. *Survival analysis: a practical approach*. Chichester, U.K.: John Wiley & Sons, 2006.

MOGHIMI-DEHKORDI, B. et al. Statistical comparison of survival models for analysis of cancer data. *Asian Pac J Cancer Prev*, v. 9, n. 3, p. 417–20, 2008.

NAKAMURA, L. R. *Advances on the Birnbaum-Saunders distribution*. 2016. 94 f. Tese (Doutorado em Estatística e Experimentação Agronômica) — Universidade de São Paulo, Piracicaba, 2016.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, v. 12, n. 3, p. 289–310, 2008.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 2 jan. 2020.

RAMALINGAM, S. et al. Lung cancer in young patients: analysis of a surveillance, epidemiology, and end results database. *Journal of clinical oncology*, v. 16, n. 2, p. 651–657, 1998.

RAMIRES, T. G. et al. Predicting the cure rate of breast cancer using a new regression model with four regression structures. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, 2017.

RIGBY, R.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, Springer, v. 6, n. 1, p. 57–65, 1996.

RIGBY, R.; STASINOPOULOS, D. A flexible regression approach using gamlss in r. *London Metropolitan University, London*, 2009.

RIGBY, R. et al. *Distributions for modeling location, scale, and shape: using GAMLSS in R*. New York: CRC Press, 2019.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

SANTOS, M. de O. Estimativa 2018: Incidência de câncer no brasil. *Revista Brasileira de Cancerologia*, v. 64, n. 1, p. 119–120, 2018.

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

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality. *Biometrika*, London, v. 52, n. 3, p. 591–611, 1965.

SOUZA, R. C. *Análise de sobrevivência na presença de censura informativa: uma abordagem Bayesiana*. 2015. 104 f. Dissertação (Mestrado em Estatística) — Departamento de Estatística, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

STACY, E. A generalization of the gamma distribution. *Annals of Mathematical Statistics*, v. 33, p. 1187–1192, 1962.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007.

STEWART, B. W.; WILD, C. P. *World cancer report: 2014*. Lyon: IARC Press, 2014.

SUNG, J.; TANAKA, Y. A numerical study on statistical diagnostics on cox proportional hazards models for survival data analysis. *Journal of the Faculty of Environmental Science and Technology*, Okayama University, v. 9, n. 1, p. 37–44, 2004.

TABLEMAN, M.; KIM, J. S. *Survival analysis using S: analysis of time-to-event data*. Boca Raton: Chapman and Hall/CRC, 2003.

THERNEAU, T. M.; GRAMBSCH, P. M.; FLEMING, T. R. Martingale-based residuals for survival models. *Biometrika*, Oxford University Press, v. 77, n. 1, p. 147–160, 1990.

VASCONCELOS, J. C. S. *Modelo Linear Parcial Generalizado Simétrico*. 2017. 71 f. Dissertação (Mestrado em Ciências) — Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2017.

WANG, Y.; MAHBOUB, K. C.; HANCHER, D. E. Survival analysis of fatigue cracking for flexible pavements based on long-term pavement performance data. *Journal of transportation engineering*, American Society of Civil Engineers, v. 131, n. 8, p. 608–616, 2005.

WILK, M. B.; GNANADESIKAN, R. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, Oxford University Press, v. 55, n. 1, p. 1–17, 1968.

ZANAKIS, S. H.; KYPARISIS, J. A review of maximum likelihood estimation methods for the three-parameter weibull distribution. *Journal of statistical computation and simulation*, Taylor & Francis, v. 25, n. 1-2, p. 53–73, 1986.

ZHU, H. P. et al. Application of weibull model for survival of patients with gastric cancer. *BMC gastroenterology*, BioMed Central, v. 11, n. 1, p. 1, 2011.

APÊNDICE A – Rotina do R

```
# Carregando os pacotes utilizados

library(flexsurv)
library(gamlss)
library(gamlss.cens)
require(nortest)
library(RColorBrewer)

# lendo os dados

dados<-read.table("dados.txt", h=TRUE)

head(dados)

tempos<-dados$Tempo # variável resposta
cens<-dados$Censura # variável indicadora de falha ou censura

## Variáveis explicativas

x1<-dados$Gender
x2<-dados$Age
x3<-dados$Histologia
x4<-dados$Cigarro
x5<-dados$TNM
x6<-dados$Terapia
x7<-dados$Metastase

########## análise descritiva #########
```

```
# box-plot dos tempos de censura e de falha

TempoCensurado<-c(330,442,361,...,358)

TempoDeFalha<-c(543,405,103,...,309)

par(mfrow=c(1,2))
boxplot(TempoCensurado)
boxplot(TempoDeFalha)

# gráfico censura versus tempo

plot(tempos,cens,ylim=c(-0.5,1.5),ylab = "Censura",
xlab="Tempo (dias)",pch=16)

# gráficos das variáveis explicativas

color <- colorRampPalette(c("black"))

# gráfico da idade

plot(sort(x2),ylim=c(0,100),ylab = "Idade",xlab="Indivíduos",
pch=16)

# gráfico de setores da variável explicativa Histologia

Q1 <- prop.table(table(dados$Histologia))*100
names(Q1) <- c("ADC","SCLC","SQC")
pie(Q1,col=color(3),density = c(0,10,30))
text(locator(n=1),paste(round(Q1[1],digits=2),"%"))
text(locator(n=1),paste(round(Q1[2],digits=2),"%"))
text(locator(n=1),paste(round(Q1[3],digits=2),"%"))
```

```
box(which = "figure", col="black", lwd=20)
box(which = "outer", col="white",lwd=18)

# gráfico de setores da variável explicativa Cigarro

Q1 <- prop.table(table(dados$Cigarro))*100
names(Q1) <- c("Não-fumantes","Fumantes")
pie(Q1,col=color(1),density = c(0,20))
text(locator(n=1),paste(round(Q1[1],digits=2),"%"))
text(locator(n=1),paste(round(Q1[2],digits=2),"%"))
box(which = "figure", col="black", lwd=20)
box(which = "outer", col="white",lwd=18)

# gráfico de setores da variável explicativa TNM

Q1 <- prop.table(table(dados$TNM))*100
names(Q1) <- c("Estágio II","Estágio III","Estágio IV")
pie(Q1,col=color(3),density = c(0,10,30))
text(locator(n=1),paste(round(Q1[1],digits=2),"%"))
text(locator(n=1),paste(round(Q1[2],digits=2),"%"))
text(locator(n=1),paste(round(Q1[3],digits=2),"%"))
box(which = "figure", col="black", lwd=20)
box(which = "outer", col="white",lwd=18)

# gráfico de barras da variável explicativa Terapia

barplot(prop.table(table(dados$Terapia))*100,
        space=.8, width=c(.1,.1,.1,.1), col=color(1),
        xlab="", ylab="Frequência percentual",ylim=c(0,80),
        density = c(0,10,30,50))
text(locator(n=4),c("10,71", "14,29", "73,21",  "1,79"))
box(which = "figure", col="black", lwd=1)
```

```
# gráfico de setores da variável explicativa Metastase

Q1 <- prop.table(table(dados$Metastase))*100
names(Q1) <- c("Não possui metástase","Possui Metástase")
pie(Q1,col=color(1),density = c(0,20))
text(locator(n=1),paste(round(Q1[1],digits=2),"%"))
text(locator(n=1),paste(round(Q1[2],digits=2),"%"))
box(which = "figure", col="black", lwd=20)
box(which = "outer", col="white",lwd=18)

########## curvas de kaplan-meier ##########

# estimador de kaplan meier

km  <- survfit(Surv(tempos,cens) ~ 1, data= dados)
tempo <- km$time
sobrev.km <- km$surv
plot (km, conf.int=F,  xlab="Tempo (dias)",lwd=2,
      ylab="S(t) Sobrevivência Estimada")

## ajuste dos modelos modelos

# ajuste Gama generalizado
m1<- flexsurvreg(Surv(tempos,cens)~1, data = dados,
dist="gengamma")

# ajuste Weibull
m2<- flexsurvreg(Surv(tempos,cens)~1, data = dados,
dist="weibull")

# ajuste gama
```

```
m3<- flexsurvreg(Surv(tempos,cens)~1, data = dados,
dist="gamma")

# ajuste log normal
m5<- flexsurvreg(Surv(tempos,cens)~1, data = dados,
dist="lnorm")

# Gráfico de kaplan-meier e das curvas estimadas dos modelos

plot (km, conf.int=F,lty = 1,  xlab="Tempo (dias)",lwd=2,
      ylab="Sobrevivência Estimada")
plot(m1,cl=F,lty = 2,col="black",add=TRUE,lty.ci=0)
plot(m2,cl=F,lty = 3,col="black",add=TRUE,lty.ci=0)
plot(m3,cl=F,lty = 4,col="black",add=TRUE,lty.ci=0)
#plot(m4,cl=F,lty = 5,col="black",add=TRUE,lty.ci=0)
plot(m5,cl=F,lty = 5,col="black",add=TRUE,lty.ci=0)
legend(1180,1.03 , c("KM","GG", "Wei","GA","Lnorm"), lty= 1:5)

# utilizando o GAMLSS

########## análise Wibull ##########

nuloWeibull<-gamlss(Surv(tempos, cens,type="right" )~1,
                    family=cens(WEI),method=RS())
Weibull<-stepGAICAll.A(nuloWeibull, scope=list(lower=~1,
                      upper=~x1+x2+x3+x4+x5+x6+x7))

summary(Weibull)

term.plot(Weibull)
term.plot(Weibull,parameter="sigma")
```

```
plot(density(Weibull$residuals))
qqnorm(Weibull$residuals,pch=16)
qqline(Weibull$residuals)
wp(Weibull)

########## análise Log-Normal ##########

nuloLogNormal<-gamlss(Surv(tempos, cens,type="right")~1,
                      family=cens(LOGNO),method=RS())
LogNormal<-stepGAICAll.A(nuloLogNormal, scope=list(lower=~1,
                      upper=~x1+x2+x3+x4+x5+x6+x7))
summary(LogNormal)

term.plot(LogNormal)
term.plot(LogNormal,parameter="sigma")

plot(density(LogNormal$residuals))
qqnorm(LogNormal$residuals,pch=16)
qqline(LogNormal$residuals)
wp(LogNormal)

########## análise Gamma ##########

nuloGama<-gamlss(Surv(tempos, cens,type="right")~1,
                 family=cens(GA),method=RS())
Gama<-stepGAICAll.A(nuloGama, scope=list(lower=~1,
                 upper=~x1+x2+x3+x4+x5+x6+x7))
summary(Gama)

term.plot(Gama)
term.plot(Gama,parameter="sigma")
```

```
plot(density(Gama$residuals))
qqnorm(Gama$residuals,pch=16)
qqline(Gama$residuals)
wp(Gama)

########## análise Gamma Generalizada ##########

nuloGamaGeneralizada<-gamlss(Surv(tempos, cens)~1,
                      family=cens("GG",type="right"),
                      method=RS())
GamaGeneralizada<-stepGAICAll.A(nuloGamaGeneralizada,
                  scope=list(lower=~1,
                  upper=~x1+x2+x3+x4+x5+x6+x7))

summary(GamaGeneralizada)

# visto que o modelo Gama Generalizado possuiu menor AIC, SBC
e GD foram realizadas as análises referentes a esse modelo

# influencia das variaveis explicativas nos parâmetros

par(mfrow=c(2,2))
term.plot(GamaGeneralizada)
term.plot(GamaGeneralizada,parameter='sigma')
par(mfrow=c(1,2))
term.plot(GamaGeneralizada,parameter='nu')

# normalidade dos resíduos

par(mfrow=c(2,2))
plot(density(GamaGeneralizada$residuals))
qqnorm(GamaGeneralizada$residuals,pch=16)
```

```
qqline(GamaGeneralizada$residuals)
abline(0, 1, col = "red")

# normalidade Shapiro
shapiro.test(GamaGeneralizada$residuals)


# gráficos do modelo final

plot(GamaGeneralizada)

# graficos de diagnósticos

# Quantis dos Resíduos x Ordem das observacções

par(mfrow = c(1,1))
plot(GamaGeneralizada$residuals, main="Quantis dos Resíduos
     x Ordem das observacções",xlab="Ordem",
     ylab="Quantil dos Resíduos")

#gráfico dos valores observados x valores ajustados

plot(tempos,fitted(GamaGeneralizada),
     xlab = "Valores observados", ylab = "Valores ajustados",
     ylim = c(-2000,3500),pch=16)
abline(lsfit(tempos,fitted(GamaGeneralizada)), col="red")

# Residuos Deviance

plot(GamaGeneralizada$residuals,ylab ="Quantile residuals")
rm= cens+log(1-pGG(tempos,GamaGeneralizada$mu.fv,
              GamaGeneralizada$sigma.fv,
```

```
                GamaGeneralizada$nu.fv))
ResiduosDeviance = sign(rm)*( -2*(rm+ log(cens-rm)) )^(0.5)
plot (ResiduosDeviance, xlab = "índice" ,ylab = "Deviance",
                pch =16 , ylim =c( -7 ,4) )
abline(0,0,col="red")

# Resíduo de Cox-Snell

ei<- -log(1-pGG(tempos, GamaGeneralizada$mu.fv,
                    GamaGeneralizada$sigma.fv,
                    GamaGeneralizada$nu.fv))

## # estimador de kaplan meier do resíduo de Cox-Snell
ekmrCoxSnell<-survfit(Surv(ei,dados$Censura)~1)
tCoxSnell<-ekmrCoxSnell$time
st<-ekmrCoxSnell$surv
sexp<-exp(-tCoxSnell)

par(mfrow=c(1,3))

plot(tCoxSnell, -log(st),xlab="Residuos Cox-Snell",
     ylab="Taxa de Falha Acumulada",type="s",ylim=c(0,3),
     xlim=c(0,2.6))
a <- seq(0, 2.6,length=100)
lines(a,a,lwd=2)

plot(st,sexp,xlab="S(ei) estimada de Kaplan-Meier",
     ylab="S(ei) estimada Exponencial padrão",pch=16,
     xlim=c(0,1), ylim = c(0,1))
z1<-seq(0,1,0.1)
lines(z1,z1)
```

```
lm(sexp~st) # por meio desse comando verifica-se o quão
próximo a reta ajustada está da reta com inclinação 1

plot(ekmr,conf.int=F,mark.time=F, xlab="Resíduos de Cox-Snell",
     ylab="Sobrevivência estimada",xlim=c(0,3.2),ylim=c(0,1.0))
lines(t,sexp,lwd=3,lty=4)
legend(1.0,1.0,lty=c(1,4),lwd=c(1,3),c("Kaplan-Meier",
       "Exponencial Padrão"), bty="n",col=c("black","black"))

# Worm-plot

wp(GamaGeneralizada)
```