Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Marcos Vinicius Bueno

Análise de sobrevivência para dados de HIV/AIDS: abordagem paramétrica

MARINGÁ – PARANÁ
2022

Marcos Vinicius Bueno

Análise de sobrevivência para dados de HIV/AIDS: abordagem paramétrica

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof. Dr. Robson Marcelo Rossi

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

MARINGÁ – PARANÁ

2022

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Bueno, Marcos Vinicius

B928a Análise de sobrevivência para dados de HIV/AIDS : abordagem paramétrica / Marcos Vinicius Bueno. -- Maringá, PR, 2022.
98 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Robson Marcelo Rossi.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2022.

1. Análise de sobrevivência (Biometria). 2. Dados de HIV/AIDS. 3. Pesquisa médica - Métodos estatísticos. I. Rossi, Robson Marcelo, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 519.5

Jane Lessa Monção - CRB 9/1173

MARCOS VINICIUS BUENO

Análise de sobrevivência para dados de HIV/AIDS: abordagem paramétrica

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Robson Marcelo Rossi
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Edson Zangiacomi Martinez
Universidade de São Paulo – PBE/USP

Profª. Dra. Vera Lucia Damasceno Tomazella
Universidade Federal de São Carlos – DES/UFSCar

Maringá, 28 de março de 2022.

Agradecimentos

Primeiramente, agradeço a Deus pelo dom da vida e por ter me dado perseverança para atingir meus objetivos.

Ao meu orientador Prof. Dr. Robson Marcelo Rossi, por toda paciência, dedicação, conhecimentos partilhados e contribuições para elaboração deste trabalho.

Aos membros da banca de qualificação/defesa pelas sugestões para o aperfeiçoamento deste trabalho.

Aos meus pais, Maria e Eduardo, ao meu irmão, Eduardo Junior e a minha namorada, Thalia, por todo apoio e por sempre acreditarem em mim.

Aos meus professores e amigos, Ms. Adriana Strieder Philippsen e Dr. Marcos Vinicius de Oliveira Peres, pelo incentivo para ingressar em um curso de mestrado.

A todos professor e colegas de turma do Programa de Pós-Graduação em Bioestatística da UEM, com os quais pude adquirir novos conhecimentos ao longo do curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo auxílio financeiro.

Resumo

A síndrome da imunodeficiência adquirida (AIDS) é uma doença crônica causada pelo vírus da imunodeficiência humana (HIV), presente no Brasil desde a década de 1980. Quando o interesse é analisar a variável tempo até a ocorrência do evento de interesse, na presença de censuras, a Análise de Sobrevivência é a principal ferramenta estatística utilizada. Neste contexto, é possível encontrar na literatura uma grande quantidade de trabalhos envolvendo dados de HIV/AIDS utilizando abordagens não paramétrica e semi-paramétrica via modelos de regressão de Cox, porém, poucos trabalhos são apresentados considerando a paramétrica, isto é, que a variável resposta segue uma distribuição de probabilidade. Partindo disso, o objetivo deste trabalho foi analisar dados de tempo de vida (em semanas) observados de pacientes com HIV/AIDS no Estado do Mato Grosso do Sul, Brasil, observados entre 2009 e 2018, por meio de abordagem paramétrica. No estudo inicial, objetivou-se verificar dentre os principais modelos probabilísticos conhecidos na literatura, qual seria o mais adequado para ajustar os dados, e a partir disso, comparar o ajuste com a distribuição Weibull, modelo padrão nesse tipo de análise, utilizando uma abordagem Bayesiana. Por meio do logaritmo da função de verossimilhança e do critério de informação de Akaike (AIC), verificou-se que a distribuição Log-Normal foi a mais adequada, e utilizando o critério de *Deviance* Bayesiano (DIC), notou-se também que o modelo selecionado apresentou melhor ajuste que a distribuição Weibull para a representação dos dados. No segundo estudo, objetivou-se construir modelos de regressão, por meio das abordagens frequentista e Bayesiana. Utilizando a distribuição Log-Normal, foram estimados os parâmetros do modelo associados às variáveis explicativas: sexo, raça, escolaridade e uso de drogas injetáveis. A partir das interpretações desses parâmetros, é possível discutir a necessidade de políticas de prevenção e diagnóstico precoce focadas em grupos que apresentam um menor tempo mediano de sobrevida após a notificação da infecção pelo vírus do HIV. Em outras análises, objetivou-se verificar a adequação dos dados a distribuições mais complexas, de modo que, em uma das análises foram incluídos parâmetros adicionais para a fração de cura e inflação de zeros. Em ambas as análises, a distribuição Kumaraswamy Gama Generaliza foi selecionada com base no valor do AIC, na qual, posteriormente também foi ajustado um modelo de regressão considerando as variáveis explicativas, de modo que, em comum com a segunda análise, o sexo feminino, a raça branca, a escolaridade superior a 8 anos foram associados a uma maior probabilidade de sobrevivência.

Palavras-chave: dados censurados; modelos de longa duração; modelos probabilísticos; tempos de falha.

Abstract

Acquired immunodeficiency syndrome (AIDS) is a chronic disease caused by the human immunodeficiency virus (HIV), present in Brazil since the 1980s. When the interest is to analyze the time variable until the event of interest in the presence of censored lifetimes, the Survival Analysis is the main statistical tool used. In this context, it is possible to find in the literature a large number of studies involving HIV/AIDS data using non-parametric and semi-parametric approaches via Cox regression models, however, few studies are presented considering the parametric, that is, that the response variable follows a probability distribution. Based on this, the objective of this study was to analyze survival time data (in weeks) of patients with HIV/AIDS in the State of Mato Grosso do Sul, Brazil, observed between 2009 and 2018, using a parametric approach. In the initial study, the objective was to verify, among the main probabilistic models known in the literature, which would be the most appropriate to adjust the data, and from this, to compare the fit with the Weibull distribution, the standard model in this type of analysis, using an approach Bayesian. Through the logarithm of the likelihood function and the Akaike information criterion (AIC), it was found that the Log-Normal distribution was the most appropriate, and using the Bayesian Deviance criterion (DIC), it was also noted that the selected model presented a better fit than the Weibull distribution for data representation. In the second study, the objective was to build regression models using frequentist and Bayesian approaches. Using the Log-Normal distribution, the model parameters associated with the explanatory variables were estimated: gender, race, education, and injecting drug use. Based on the interpretations of these parameters, it is possible to discuss the need for prevention and early diagnosis policies focused on groups that have a lower median survival time after notification of HIV infection. In other two analyses, the objective was to verify the adequacy of the data to more complex distributions. In the second analysis, additional parameters were included for a fraction of cure, in addition to the inflation of zeros. In both analyses, the Kumaraswamy Gama Generaliza distribution was selected based on the value of the AIC, in which a regression model was later adjusted considering the explanatory variables, so that, in common with the second analysis, female gender, white race, higher education at eight years were associated with a higher probability of survival.

Keywords: censored data; failure times; long-term models; probabilistic models.

Lista de ilustrações

Figura 1 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Exponencial com diferentes valores para α. 21
Figura 2 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Weibull com diferentes valores para γ e α. 22
Figura 3 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Gama com diferentes valores para k e α. 23
Figura 4 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Gama Generalizada com diferentes valores para k, γ e α. 24
Figura 5 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Log-Normal com diferentes valores para μ e σ. 26
Figura 6 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Log-Logística com diferentes valores para γ e α. 27
Figura 7 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Burr XII com diferentes valores para γ, α e θ. 28
Figura 8 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Beta-Weibull com diferentes valores para α, β, λ e γ. 29

Figura 9 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Kumaraswamy Gama Generalizada com diferentes valores para seus parâmetros. 30
Figura 10 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Weibull Generalizada Exponenciada de Potência com diferentes valores para seus parâmetros. . 32
Figura 11 – Curva de sobrevivência empírica via Kaplan-Meier. 41
Figura 12 – Curvas de sobrevivência estimadas via Kaplan-Meier e procedimentos Bayesianos para as distribuições Weibull e Log-Normal. 42
Figura 13 – Curvas de sobrevivência dos modelos Weibull e Log-Normal estimadas via métodos frequentistas e Bayesianos, comparadas com as estimativas de Kaplan-Meier. 43
Figura 14 – Análise de resíduos de Cox-Snell para os modelos Weibull e Log-Normal 44
Figura 15 – Curva de sobrevivência empírica de HIV/AIDS. 52
Figura 16 – Resíduos de Cox-Snell do modelo de regressão Bayesiano estimadas pelo método de Kaplan-Meier e pelo Exponencial padronizado (à esquerda) e respectivas curvas de sobrevivência estimadas (à direita). . . 54
Figura 17 – Gráficos de tempo total em teste (TTT-plot) e $S(t)$ via Kaplan-Meier, respectivamente. 66
Figura 18 – Valores preditores gerados pelas funções de sobrevivência e estimativas de Kaplan-Meier. 67
Figura 19 – Comparação entre as curvas de sobrevivência empírica (Kaplan-Meier) e estimada pelo melhor modelo (KGG). 69
Figura 20 – Resíduos de Cox-Snell para os dados trabalhados. 69
Figura 21 – TTT-plot e função de sobrevivência estimada via Kaplan-Meier, respectivamente. 76
Figura 22 – Comparação entre as curvas de sobrevivência empírica e ajustada, e comparação entre os resíduos de Cox-Snell e função de risco acumulado, respectivamente. 78
Figura 23 – Comparações entre as curvas de Kaplan-Meier e a função de sobrevivência estimada para cada modelo regressão, por variável explicativa. . 79
Figura 24 – Análise de resíduos de Cox-Snell para o modelo de regressão ZCKGG. 81
Figura 25 – Comparações entre as estimativas de Kaplan-Meier e valores preditos para os modelos probabilísticos sem as observações com 521,57 semanas. 95
Figura 26 – Resíduos de Cox-Snell para a distribuição KGG considerando os dados sem observações com 521,57 semanas. 96

Lista de tabelas

Tabela 1 – Distribuições de probabilidade e função de sobrevivência para as distribuições Weibull e Log-Normal. 39
Tabela 2 – Resultados do logaritmo das funções de verossimilhança e AIC para as distribuições consideradas. 41
Tabela 3 – Estimativas *a posteriori* dos parâmetros e seus intervalos com 95% de credibilidade $[P_{2,5\%}; P_{97,5\%}]$. 42
Tabela 4 – Resumo descritivo para as variáveis explicativas entre 2009 e 2018. . . 49
Tabela 5 – Função densidade de probabilidade dos modelos probabilísticos considerados. 51
Tabela 6 – Resultados para o AIC e $\log L(\theta)$ para os modelos probabilísticos considerados. 53
Tabela 7 – Estimativas dos parâmetros do modelos de regressão frequentista e Bayesiano. 55
Tabela 8 – Distribuições de probabilidade consideradas para os ajustes dos dados. 63
Tabela 9 – Estimativas para os parâmetros dos modelos. 68
Tabela 10 – Resumo descritivo para as variáveis explicativas. 73
Tabela 11 – Distribuições de probabilidade consideradas para os ajustes dos dados. 74
Tabela 12 – Estimativas dos parâmetros das distribuições consideradas. 77
Tabela 13 – Estimativas dos parâmetros (e erro padrão) dos modelos de regressão. 80
Tabela 14 – Teste de Wald para a significância das variáveis explicativas do modelo de regressão ZCKGG. 81
Tabela 15 – Estimativas para os parâmetros dos modelos considerando dados sem censuras com 521,57 semanas. 94

Sumário

1 Introdução 12

2 Fundamentação Teórica 15

2.1 Conceitos de Análise de Sobrevivência 15

2.1.1 Tempo de Falha 15

2.1.2 Censura e Truncamento 16

2.1.3 Representação de Dados de Sobrevivência 17

2.1.4 Principais Funções para Dados de Sobrevivência 17

2.1.4.1 Função de Sobrevivência 17

2.1.4.2 Função Taxa de Falha 18

2.1.4.3 Função Taxa de Falha Acumulada 18

2.1.5 Estimador de Kaplan-Meier 18

2.1.6 Modelos Paramétricos em Análise de Sobrevivência 20

2.1.7 Distribuições de Probabilidade 20

2.1.7.1 Distribuição Exponencial 20

2.1.7.2 Distribuição Weibull 21

2.1.7.3 Distribuição Gama 22

2.1.7.4 Distribuição Gama Generalizada 23

2.1.7.5 Distribuição Log-Normal 25

2.1.7.6 Distribuição Log-Logística 26

2.1.7.7 Distribuição Burr XII 27

2.1.7.8 Distribuição Beta-Weibull 28

2.1.7.9 Distribuições da Família Kumaraswamy 29

2.1.7.10 Distribuição Weibull Generalizada Exponenciada de Potência 31

2.1.8 Inflação de Zeros e Fração de cura 32

2.1.8.1 Modelos com Inflação de Zeros 32

2.1.8.2 Modelos com Fração de Cura 33

2.1.8.3 Modelos com Inflação de Zeros e Fração de Curados . . . 33

2.2 Métodos de Estimação 34

2.2.1 O Método da Máxima Verossimilhança 34

2.2.1.1 Intervalo de Confiança 35

2.2.1.2 Função de Ligação 35
2.2.2 Inferência Bayesiana 35
3 Análise de sobrevivência para dados de HIV/AIDS: uma abordagem Bayesiana 37
3.1 Introdução 37
3.2 Materiais e Métodos 38
3.3 Resultados e discussão 40
3.4 Conclusões 44
4 Fatores associados ao tempo de sobrevida para pacientes com HIV/AIDS no Estado Mato Grosso do Sul: abordagem paramétrica 45
4.1 Introdução 46
4.2 Materiais e Métodos 48
4.3 Resultados e Discussão 52
4.4 Conclusões 57
5 Modelos estatísticos para a análise de sobrevivência de pacientes com HIV/AIDS 58
5.1 Introdução 59
5.2 Materiais e Métodos 60
5.3 Resultados e discussão 65
5.4 Conclusões 69
6 Modelo de fração de cura inflacionado de zero na identificação de fatores associados à sobrevida de pacientes com HIV/AIDS no Estado do Mato Grosso do Sul, Brasil 70
6.1 Introdução 71
6.2 Materiais e Métodos 72
6.3 Resultados e Discussão 76
6.4 Conclusões 81
7 Considerações finais e passos futuros 83

Referências 85

Apêndices 92
APÊNDICE A Modelos estatísticos para análise de sobrevivência de pacientes com HIV/AIDS desconsiderando censuras 93
APÊNDICE B Principais códigos em R 97

Capítulo 1

Introdução

A síndrome da imunodeficiência adquirida (AIDS) é uma doença crônica causada pelo vírus da imunodeficiência humana (HIV), cujo o primeiro caso no Brasil foi registrado na década de 80, e considerada inicialmente como uma doença observada apenas em grandes metrópoles, que afetava principalmente homossexuais, bissexuais masculinos, hemofílicos e usuários de drogas injetáveis. Com o passar dos anos, a AIDS deixou de ser uma doença de segmentos populacionais e disseminou-se sobre a população geral (SZWARCWALD et al., 2000).

Martins et al. (2014) afirmam que, além do impacto na saúde, o HIV amplia males da sociedade e deficiências dos sistemas sociais, de modo que a resposta para a epidemia tem proporcionado uma oportunidade de fortalecimento do tecido social, reforçando os serviços essenciais para os segmentos mais vulneráveis da sociedade.

Além disso, os mesmos autores discorrem sobre problemas sociais causados pela infecção por HIV, bem como a discriminação, onde relatam-se que em diversos países, pessoas vivendo com HIV perderam o emprego e acessos aos cuidados de saúde ou outros serviços públicos.

Segundo Candido et al. (2021), o quadro epidêmico de infecção pelo vírus do HIV e pela AIDS vem sofrendo diversas modificações ao longo do tempo. O envolvimento de populações socialmente mais vulneráveis, o aumento do número de notificações em municípios de pequeno e médio porte e o crescente número de mulheres infectadas pelo vírus são marcas da epidemia no Brasil. Essas características indicam que o sistema de saúde brasileiro apresenta deficiências relacionadas à prevenção e ao tratamento da infecção pelo HIV.

No contexto de análise de sobrevivência, são comumente encontrados na literatura trabalhos que analisam dados de HIV/AIDS, porém utilizando a abordagem semi-paramétrica por meio do modelo de risco proporcional de Cox.

Medeiros et al. (2017) avaliam a sobrevida de pessoas com HIV/AIDS atendidas por um hospital de João Pessoa (PB) e discutem sobre fatores associados à diminuição da sobrevida. De maneira similar, Melo, Donalisio e Cordeiro (2017) descrevem fatores que influenciam no tempo de sobrevida dos pacientes das região dos Campos Gerais, no Paraná.

Zaslavsky, Goulart e Ziegelmann (2019), realizam uma análise com dados de pacientes com HIV segundo o deslocamento em busca do tratamento, visando verificar se existe diferença na sobrevida. Os pacientes foram divididos em três grupos: sem a necessidade; que fazem tratamento em outra região; que realizam o tratamento em outro país da trile fronteira Brasil, Paraguai e Argentina.

Entretanto, trabalhos envolvendo dados de HIV/AIDS em um contexto de análise de sobrevivência paramétrica são mais escassos, principalmente para dados brasileiros. Um exemplo de trabalho utilizando esta abordagem pode ser visto em Nakhaee e Law (2011), que aplicam os modelos Exponencial, Weibull, Log-Normal e Log-Logístico para dados de HIV/AIDS da Austrália, de modo que o modelo Weibull apresentou um melhor ajuste, e com isso, são incluídas variáveis explicativas em um modelo de regressão.

Em Million (2018), é realizada uma análise via modelo de Cox, além de apresentar um modelo de regressão Log-Normal para dados de pacientes de HIV/AIDS com tratamento anti-retroviral altamente ativo atendidos em um determinado hospital da Etiópia. Por meio dos modelos, são discutidos os fatores que apresentam um maior risco de morte para a doença e sobre a utilização do modelo paramétrico para realizar previsões para o tempo de sobrevivência dos pacientes com HIV/AIDS.

O objetivo deste trabalho foi analisar dados de HIV/AIDS para o Estado do Mato Grosso do Sul, Brasil, observados entre 2009 e 2018, no contexto de análise de sobrevivência, por meio de técnicas paramétricas.

O presente trabalho está dividido da seguinte forma: no Capítulo 2 é apresentada uma fundamentação teórica sobre os principais conceitos abordados ao longo do trabalho, no Capitulo 3 apresenta-se uma discussão sobre um modelo probabilístico que expresse o comportamento da variável resposta por meio de uma abordagem frequentista e Bayesiana, o Capitulo 4 são apresentados modelos de regressão utilizando as variáveis explicativas associadas, por meio das abordagens frequentista e Bayesiana, o Capítulo 5 é apresentada uma discussão referente a um modelo probabilístico para a variáve respos-

tas, considerando diversas distribuições de probabilidade em um contexto frequentista, e por fim, no Capítulo 6 é discutidos sobre um modelo probabilístico adequado aos dados, incorporando a inflação de zeros e fração de cura, bem como construído um modelo de regressão utilizando as variáveis explicativas associadas, por meio da abordagem frequentista.

Capítulo 2

Fundamentação Teórica

2.1 Conceitos de Análise de Sobrevivência

Segundo Colosimo e Giolo (2006), a análise de sobrevivência é a área da Estatística que estuda o tempo até a ocorrência de um evento de interesse, o qual é denominado tempo de falha. Alguns exemplos relacionados à área da saúde são o tempo até o óbito de um paciente causado por uma determinada doença em estudo, bem como o tempo até a cura ou recidiva de uma doença. Na engenharia, o termo análise de sobrevivência é comumente denominado como análise de confiabilidade.

A ocorrência do evento de interesse estudado, é denominada falha ou desfecho. Carvalho et al. (2011) descrevem que o termo falha surgiu na análise de confiabilidade, no qual busca-se modelar o tempo até a ocorrência da falha de algum componente ou equipamento.

Uma característica de grande importância para dados de sobrevivência é a presença de censuras. As censuras são classificadas como observações que não apresentaram o evento de interesse do estudo, seja pelo término do estudo, pela perca de informação do paciente durante o período de acompanhamento ou ainda pelo óbito do paciente devido à outras causas.

2.1.1 Tempo de Falha

Colosimo e Giolo (2006) caracterizam o tempo de falha em três elementos, sendo estes:

- O tempo inicial do estudo, em que os indivíduos precisam ser comparáveis a partir do tempo que iniciaram sua participação no estudo;

- A escala de medida, que geralmente corresponde à escala de tempo real. Em alguns casos, principalmente na área da engenharia, são usados outros tipos de escala, como, por exemplo, o número de ciclos ou a quilometragem percorrida por um veículo;
- O evento de interesse, que deve ser definido precisamente antes da realização do estudo.

2.1.2 Censura e Truncamento

São denominadas de censuras as observações que possuem informações incompletas ou parciais, as quais não apresentam o evento de interesse durante o estudo. Existem diferentes tipos de censuras, que segundo Bastos e Rocha (2006), podem ser classificas como da seguinte forma, de acordo com a observação dos dados:

- Censura à direita: ocorre quando o tempo de acompanhamento do indivíduo é menor que o do tempo até evento de interesse, ou seja, o paciente não apresenta o evento de interesse;
- Censura à esquerda: ocorre quando o tempo de falha é menor que o tempo observado, ou seja, o evento de interesse já ocorreu quando o indivíduo foi observado. Este tipo de censura é menos frequente que a censura à direita;
- Censura intervalar: ocorre quando não se conhece-se o tempo em que se registou o evento de interesse, porém, sabe-se que ele ocorreu dentro de um determinado intervalo;
- Censura aleatória: ocorre quando o indivíduo é perdido de seguimento por uma causa alheia.

Além destas, podem ser classificadas de acordo com o planejamento do experimento em:

- Censura do tipo I: ocorre quando o estudo tem uma data final pré-determinada e o paciente não apresenta o evento de interesse ao fim do estudo;
- Censura do tipo II: ocorre quando é estabelecido que o estudo será terminado após uma determinada quantidade de falhas, e o indivíduo não apresentou o evento de interesse após já terem ocorrido a quantidade de falhas estipuladas no estudo.

Uma representação do mecanismo de censura pode ser construída por meio de duas variáveis aleatórias, T e C, que representam o tempo de falha e o tempo de censura associados a um paciente, respectivamente:

$$\delta = \begin{cases} 1, & \text{se} \quad T \leq C \\ 0, & \text{se} \quad T > C. \end{cases}$$

Carvalho et al. (2011) definem o truncamento quando indivíduos são excluídos do estudo por motivo relacionado à ocorrência do evento, ou seja, ocorre quando o estudo considera somente indivíduos que apresentaram o evento de interesse dentro de um intervalo temporal pré-estabelecido (T_A, T_B). O trucamento pode ser classificado como à esquerda ou à direita, em que o primeiro inclui somente observações que ocorreram após o limite inferior do intervalo de tempo pré-estabelecido $(T \geq T_A)$, e o segundo inclui somente indivíduos que sofreram o evento de interesse, ou seja, a data de ocorrência do evento é menor que o limite superior do intervalo de tempo $(T \leq T_B)$.

2.1.3 Representação de Dados de Sobrevivência

De maneira geral, os dados de sobrevivência são representados pelo par (t_i, δ_i), de maneira que t_i é o tempo de falha ou censura e δ_i é o indicador de falha ou censura, para cada indivíduo i $(i = 1,2,...,n)$ presente no estudo. A representação da indicação de censura, na maioria das vezes, é dada por:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é tempo de falha} \\ 0, & \text{se } t_i \text{ é tempo de censura.} \end{cases}$$

2.1.4 Principais Funções para Dados de Sobrevivência

Em análise de sobrevivência, para uma variável aleatória não-negativa T, sendo esta o tempo até a ocorrência do evento de interesse, informações podem ser observadas com base nas funções de sobrevivência e taxa de falha.

2.1.4.1 Função de Sobrevivência

A função de sobrevivência descreve a probabilidade de que um indivíduo do estudo não apresente o evento de interesse até um tempo t, em outras palavras, essa função indica a probabilidade do indivíduo sobreviver em um determinado tempo t. Em notação, temos: $S(t) = P(T > t)$.

Sabe-se que a função de distribuição acumulada representa a probabilidade do evento de interesse ocorrer até um tempo t, ou seja, $F(t) = P(T \leq t)$. Com isso, tem-se a

seguinte relação:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

2.1.4.2 Função Taxa de Falha

A função taxa de falha, ou função de risco, é definida como o risco de que um indivíduo sofra o evento de interesse dentro de um intervalo t e $t + \Delta t$, tal que:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[(t \leq T \leq t + \Delta t) | T \geq t\right]}{\Delta t}.$$

Carvalho et al. (2011), afirmam que o valor obtido por $h(t)$ não é uma probabilidade, de modo que sua unidade de medida é dada pelo inverso do tempo da variável aleatória em estudo, ou seja, tempo^{-1}, podendo assumir qualquer valor real maior que zero, não sendo restrita para o intervalo $[0,1]$.

Por meio de algumas relações entre funções citadas por Lee e Wang (2003), a função de risco pode ser escrita por meio da relação entre as funções de densidade de probabilidade $(f(t))$ e de sobrevivência, em um determinado tempo t, da seguinte forma:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t)).$$

2.1.4.3 Função Taxa de Falha Acumulada

A função taxa de falha acumulada, ou função de risco acumulado, fornece a taxa de falha acumulada de um indivíduo. Segundo Colosimo e Giolo (2006), embora essa função não tenha uma interpretação direta, ela pode ser útil na avaliação da função taxa de falha, em que isso ocorre essencialmente na estimação não paramétrica quando $h(t)$ é difícil de ser estimada. A função de risco acumulado é definida por:

$$\Lambda(t) = \int_0^t h(u)du = -\ln S(t).$$

2.1.5 Estimador de Kaplan-Meier

A presença de dados com censura é um problema para uma análise estatística descritiva, dificultando o cálculo de medidas de tendência central e de variabilidade. A partir de dados censurados envolvendo tempo de vida, o principal componente em uma análise descritiva é encontrar uma estimativa para a função de sobrevivência, e a partir dela, estimar as estatísticas de interesse (COLOSIMO; GIOLO, 2006).

Apresentado por Kaplan e Meier (1958), o estimador de produto-limite (não paramétrico), ou ainda estimador de Kaplan-Meier, é a principal solução para estimar a função de

sobrevivência na presença de observações censuradas, por meio de conceitos de probabilidade condicional e de independência de eventos.

A expressão geral do estimador de Kaplan-Meier é dada por:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i} \right),$$

em que:

- $t_1 < t_2 < ... < t_k$ são k tempos de falha distintos e ordenados;
- d_i é o número de falhas para o tempo t_i, em que $i = 1,2,...,n$;
- n_i é a quantidade de indivíduos em risco no tempo t_i, ou seja, não apresentaram o evento de interesse e também não foram censurados até o tempo anterior.

Dentre as principais propriedades do estimador (empírico) de Kaplan-Meier, destacam-se o fato dele ser não-viciado para amostras grandes, ser um estimador de máxima verosimilhança para a função de sobrevivência, ser fracamente consistente e convergir assintoticamente para um processo gaussiano.

Segundo Colosimo e Giolo (2006), este estimador, assim como outros, está sujeito a variações que precisam ser descritas através de estimações intervalares. A variância assintótica do estimador é expressa da seguinte forma:

$$\widehat{Var}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{i:t_i<t} \frac{d_i}{n_i(n_i - d_i)}.$$

Partindo de que o estimador de Kaplan-Meier para $S(t)$ tem distribuição Normal, um intervalo de confiança para o estimador, com nível de significância α, é dado por:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))}.$$

Colosimo e Giolo (2006), afirmam que para valores de tempos extremos, esse intervalo de confiança pode apresentar o limite inferior menor que 0 ou o limite superior maior que 1, gerando uma incoerência com a definição de probabilidade. Em casos desse tipo, deve-se realizar uma transformação na função de sobrevivência, como $\hat{U}(t) = \ln\left[-\ln \hat{S}(t)\right]$, proposta por Kalbfleisch e Prentice (2002), em que o intervalo de confiança para $S(t)$ com nível de significância α, é dado por:

$$\left[\hat{S}(t)\right]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))}\right\}}.$$

2.1.6 Modelos Paramétricos em Análise de Sobrevivência

Por meio de uma abordagem não paramétrica, como por exemplo, os estimadores de Kaplan-Meier e Nelson-Aalen (COLOSIMO; GIOLO, 2006), é possível fazer estimações para as funções de sobrevivência e de taxa de falha a partir dos dados, sem a necessidade de determinar se as observações são geradas por um modelo probabilístico.

Em contrapartida, os modelos paramétricos consistem em descrever os dados partindo da ideia que a variável resposta, tempo até ocorrência do evento de interesse, segue uma distribuição de probabilidade. Dessa forma, é possível calcular as estimativas dos parâmetros de interesse e fazer previsões dentro do escopo do modelo.

Segundo Carvalho et al. (2011), a partir de modelos paramétricos, é possível comparar o efeito de (co)variáveis no modelo, visto que as estimativas de Kaplan-Meier permitem apenas comparar ou testar a igualdade entre as curvas de sobrevivência.

2.1.7 Distribuições de Probabilidade

Existe uma diversidade de modelos probabilísticos utilizados na análise de sobrevivência. Os mais comuns são: Exponencial, Weibull, Gama, Gama Generalizado, Log-Normal, Log-Logístico, Bur XII, Beta-Weibull, Weibull Generalizada Exponenciada de Potência e algumas distribuições da família Kumaraswamy descritos a seguir.

2.1.7.1 Distribuição Exponencial

O modelo Exponencial é o mais simples para descrever dados de tempo de falha, possuindo apenas um parâmetro e é caracterizada por possuir a função de taxa de falha constante.

Considerando que uma variável aleatória não negativa, $T > 0$, segue distribuição Exponencial $(T \sim Exp(\alpha))$ com parâmetro de forma $\alpha > 0$, que representa o tempo médio de vida, tem sua função densidade de probabilidade, sobrevivência e taxa de falha, respectivamente, definidas por

$$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\},$$

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

e

$$h(t) = \frac{1}{\alpha}.$$

Algumas formas para essas funções considerando diferentes valores de α, podem ser observadas na Figura 1.

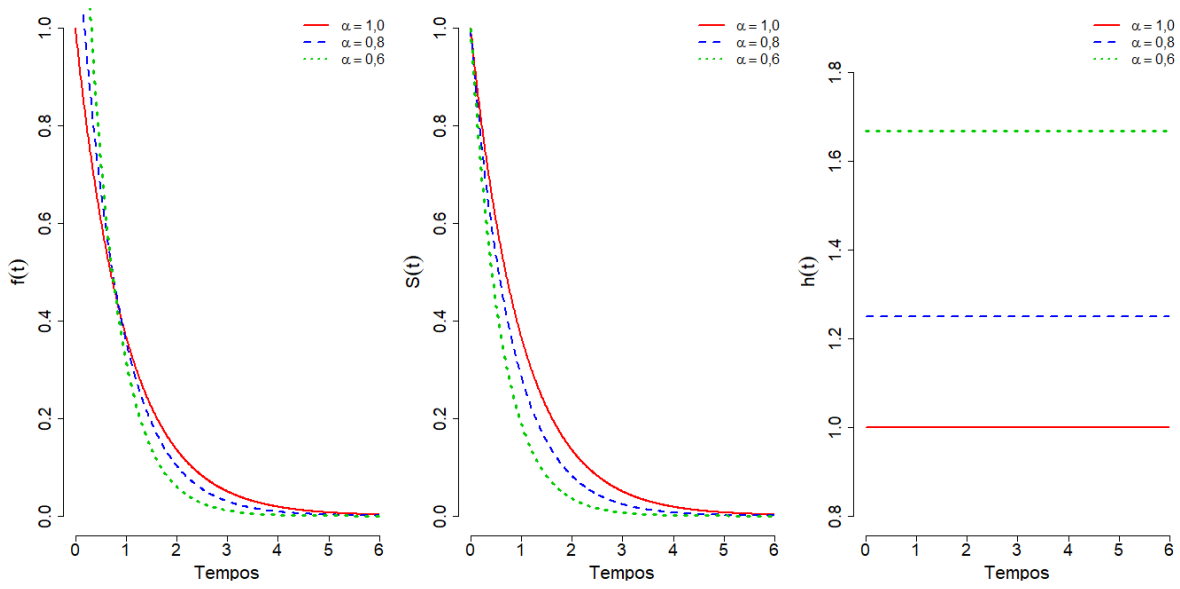

Figura 1 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Exponencial com diferentes valores para α.

2.1.7.2 Distribuição Weibull

Proposta por Weibull (1939), o modelo possui ampla utilização em aplicações práticas devido a sua capacidade de apresentar uma grande quantidade de formas. Segundo Colosimo e Giolo (2006), uma característica importante de seu uso em análise de sobrevivência é o fato de que sua função de taxa de falha é monótona, ou seja, pode ser crescente, decrescente ou constante.

Considerando que uma variável aleatória, $T > 0$, tem distribuição Weibull $(T \sim W(\alpha,\gamma))$ com parâmetros de forma $\gamma > 0$ e de escala $\alpha > 0$, sua função densidade de probabilidade, sobrevivência e taxa de falha são, respectivamente, dadas por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\},$$

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\},$$

e

$$h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}.$$

É possível observar que para $\gamma = 1$, tem-se a distribuição Exponencial, sendo esta um caso particular da Weibull.

Algumas formas para o modelo Weibull podem ser observadas na Figura 2 para diferentes valores dos parâmetros.

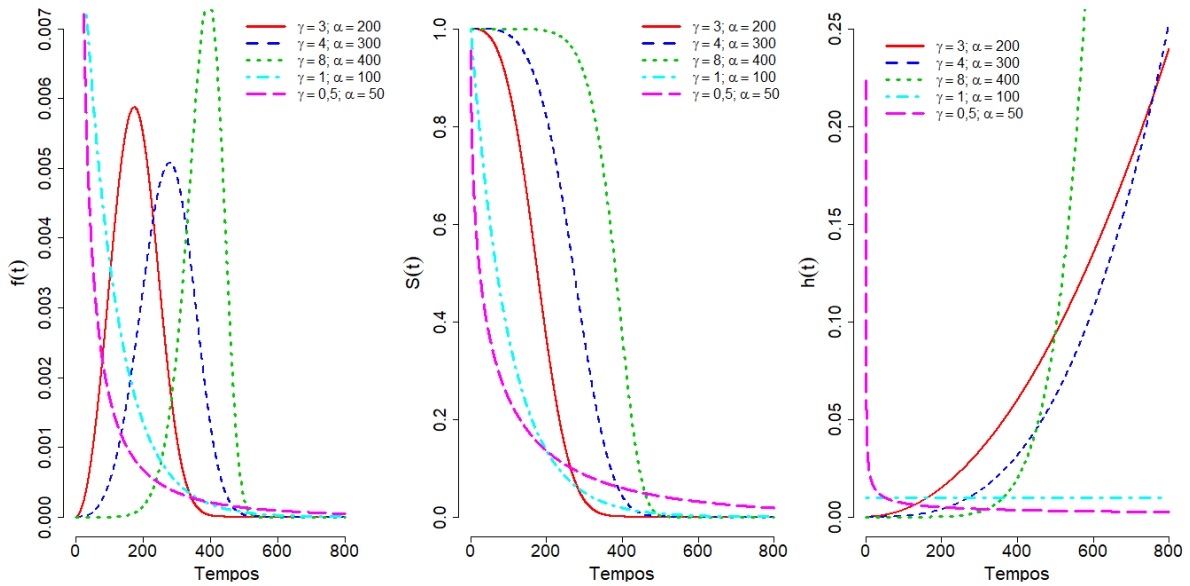

Figura 2 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Weibull com diferentes valores para γ e α.

2.1.7.3 Distribuição Gama

Usada por Brown e Flood (1947) para descrever o comportamento do tempo de vida de copos de vidro circulando em uma cafeteria, tem sido muito utilizada em estudos de duração de componentes, devido a sua adequação para estas áreas relacionadas. Também possui frequente utilização em modelagens que envolvem fatores aleatórios, como modelos de fragilidade. Entretanto, segundo Colosimo e Giolo (2006), sua abordagem na área médica para descrever tempos de vida é relativamente recente.

Considerando que variável aleatória, $T > 0$, tem distribuição Gama $(T \sim G(\alpha,k))$ com parâmetros de forma $k > 0$ e de escala $\alpha > 0$, sua função densidade de probabilidade, de sobrevivência e risco (em que as duas últimas não possuem forma analítica fechada), são dadas, respectivamente por:

$$f(t) = \frac{t^{k-1}}{\Gamma(k)\alpha^k} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\},$$

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du,$$

$$h(t) = \frac{f(t)}{S(t)}.$$

Algumas formas para as funções de probabilidade, sobrevivência e taxa de falha podem ser vistas na Figura 3, com diferentes valores para os parâmetros.

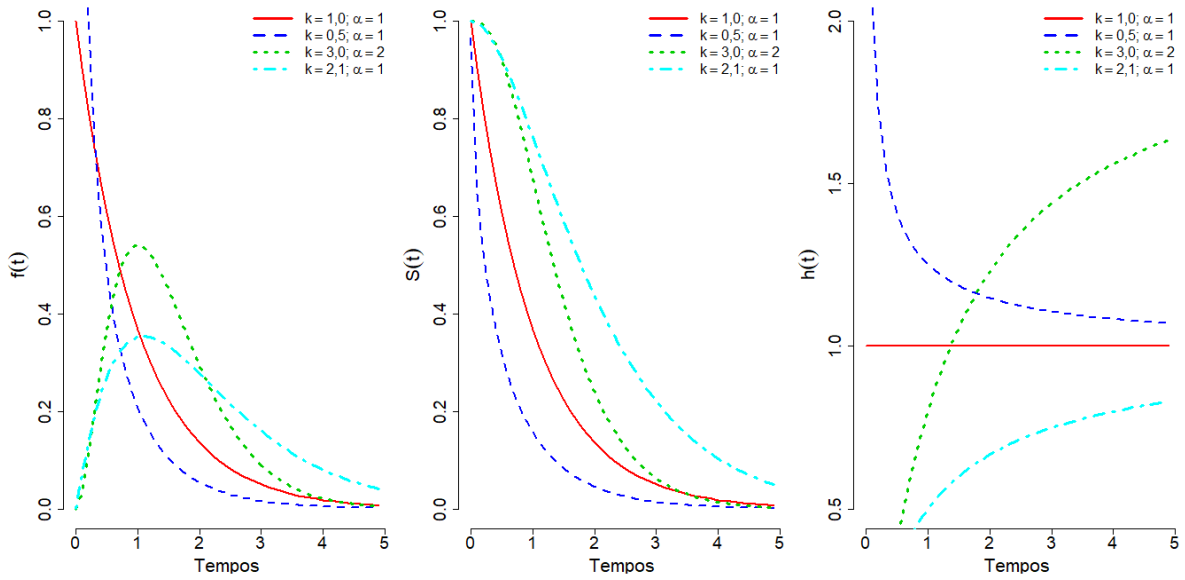

Figura 3 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Gama com diferentes valores para k e α.

2.1.7.4 Distribuição Gama Generalizada

Stacy et al. (1962) apresentam uma generalização da distribuição Gama. O modelo conta com a adição de mais um parâmetro, tornando a distribuição mais flexível.

Considerando que uma variável aleatória, $T > 0$, tem distribuição Gama Generalizada $(T \sim GG(\alpha,\gamma,k))$ com parâmetros de forma $\gamma > 0$ e $k > 0$, e parâmetro de escala $\alpha > 0$, sua função densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma^k}} t^{\gamma^k} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}.$$

Colosimo e Giolo (2006), apresentam algumas distribuições como casos particulares do modelo:

- para $\gamma = k = 1$ tem-se $T \backsim Exp(\alpha)$,
- para $k = 1$ tem-se $T \backsim W(\gamma,\alpha)$,

- para $\gamma = 1$ tem-se $T \backsim G(\alpha,k)$.

A função de sobrevivência e de risco para o modelo Gama Generalizado não possuem forma analítica fechada, as quais podem ser obtidas pelas mesmas técnicas descritas na Subseção 2.1.7.3.

Algumas formas para as funções de probabilidade, sobrevivência e taxa de falha podem ser vistas na Figura 4, com diferentes valores para os parâmetros.

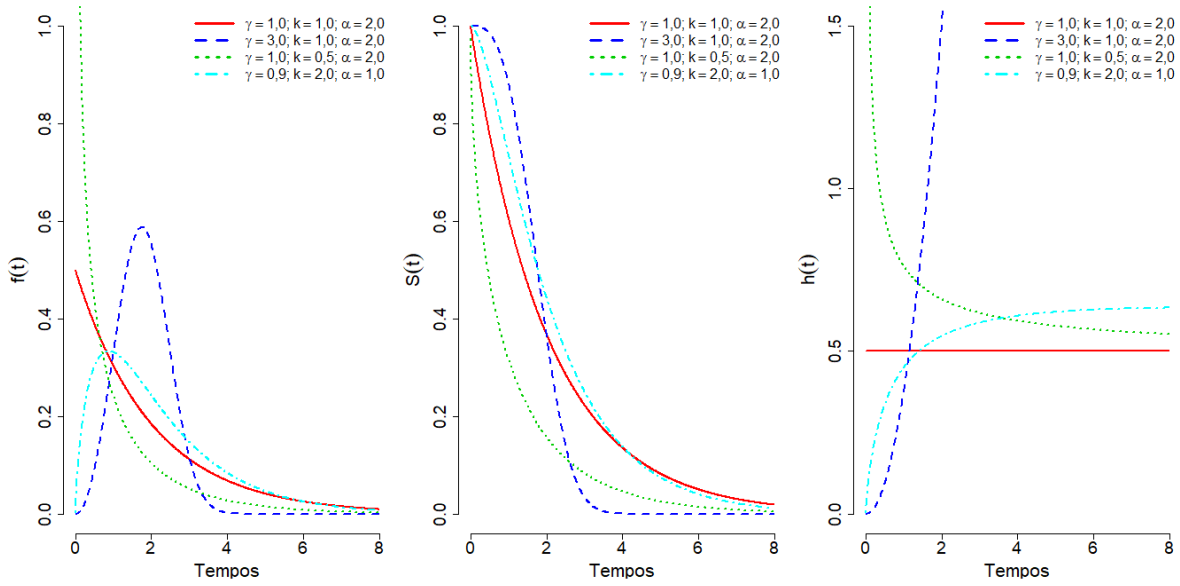

Figura 4 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Gama Generalizada com diferentes valores para k, γ e α.

Uma outra parametrização da distribuição GG é apresentada por Prentice (1974), (também conhecida como Log-Gama), em que se $g \backsim$ Gama$(Q^2, -1)$, e $w = (Q^2 g)/Q$, então $t = \exp(\mu + \sigma w)$ segue uma distribuição GG com parâmetros μ, σ e Q, e função de densidade:

$$f(t) = \frac{|Q|(Q^{-2})^{Q^{-2}}}{\sigma t \Gamma(Q^{-2})} \exp(Q^{-2}) \left[Qw - \exp(Qw)\right].$$

Cox et al. (2007) discutem sobre a aplicação da distribuição em análise de sobrevivência. A função de sobrevivência para a distribuição GG pela parametrização de Prentice é dada por:

$$S(t) = \begin{cases} 1 - I(g,\mu) & \text{se} \quad Q > 0 \\ 1 - \Phi\left(\dfrac{\log(t) - \mu}{\sigma}\right) & \text{se} \quad Q = 0 \end{cases},$$

em que $I(g,\mu) = \int_0^u \exp(-t)/\Gamma(\gamma)$ e $\Phi(.)$ é a distribuição Normal acumulada padrão. A função de taxa de falha pode ser obtida de maneira genérica, usando a relação entre a função de densidade e de sobrevivência.

2.1.7.5 Distribuição Log-Normal

O modelo probabilístico Log-Normal é amplamente conhecido na literatura e utilizado em várias áreas, inclusive na análise de sobrevivência. O modelo apresentado por Aitchison e Brown (1957) permite que os dados sejam analisados segundo uma distribuição Normal de probabilidade, entretanto, deve-se considerar o logaritmo natural dos dados ao invés dos dados originais.

Um exemplo para aplicação da distribuição Log-Normal pode ser visto em Santos, Almeida e Oliveira (2019) para modelar o tempo de vida de pacientes portadores da doença Mieloma Múltiplo.

Considerando que uma variável aleatória, $T > 0$, tem distribuição Log-Normal ($T \sim LN(\mu,\sigma)$) com média e desvio-padrão do logaritmo natural do tempo de falha definidos, respectivamente, por μ e σ, sua função densidade de probabilidade e de sobrevivência são, respectivamente, dadas por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right\}$$

e

$$S(t) = \Phi\left(\frac{-\ln(t)+\mu}{\sigma}\right),$$

em que $\Phi(.)$ é a função de probabilidade acumulada de uma Normal padrão. Já sua função de risco não apresenta forma analítica fechada.

Algumas formas para as funções de distribuição probabilidade, sobrevivência e taxa de falha podem ser vistas na Figura 5, com diferentes valores para os parâmetros.

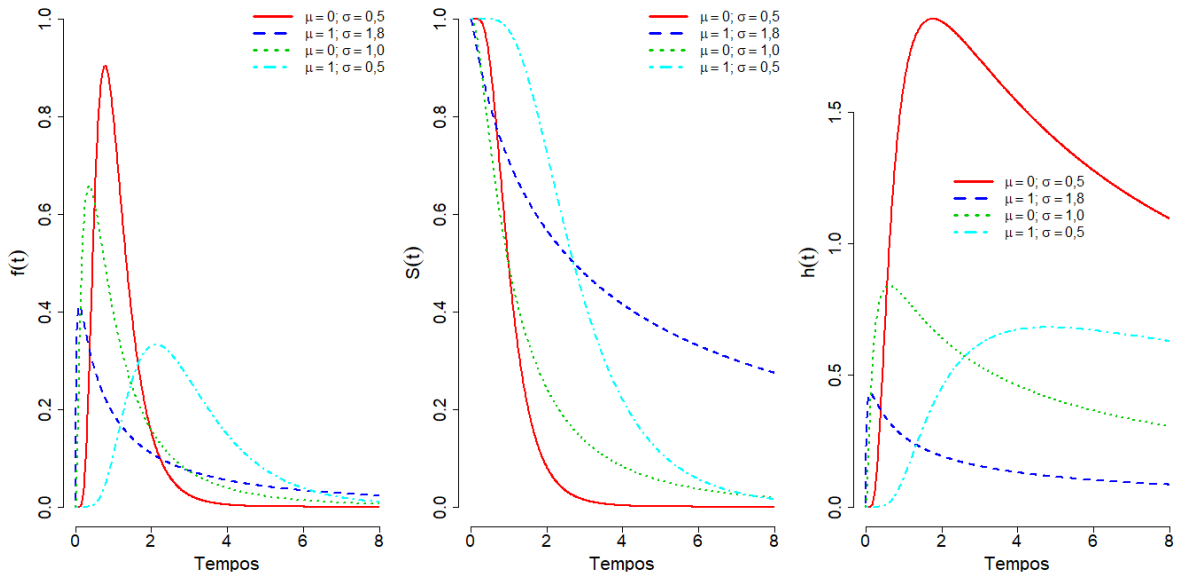

Figura 5 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Log-Normal com diferentes valores para μ e σ.

2.1.7.6 Distribuição Log-Logística

Segundo Colosimo e Giolo (2006), a distribuição Log-Logística é muito semelhante em forma à distribuição Log-Normal, entretanto, é mais adequada para o uso em análise de sobrevivência devido a sua melhor tratabilidade ao lidar com as observações censuradas, que ocorrem com frequência nesse tipo de análise estatística. Os mesmos autores ilustram a distribuição em um modelo de regressão para dados de homens com câncer de pulmão inoperável.

Considerando que uma variável aleatória não negativa, $T > 0$, tem distribuição Log-Logística $(T \sim LL(\alpha,\gamma))$ com parâmetros de forma $\gamma > 0$ e de escala $\alpha > 0$, sua função densidade de probabilidade, sobrevivência e taxa de falha são dados, respectivamente, por:

$$f(t) = \frac{(\gamma/\alpha)\,(t/\alpha)^{\gamma-1}}{(1+(t/\alpha)^{\gamma})^{2}},$$

$$S(t) = \frac{1}{1+(t/\alpha)^{\gamma}}$$

e

$$h(t) = \frac{\gamma/\alpha\,(t/\alpha)^{\gamma-1}}{(1+(t/\alpha)^{\gamma})}.$$

Algumas formas para as funções de probabilidade, sobrevivência e taxa de falha podem ser vistas na Figura 6, com diferentes valores para os parâmetros.

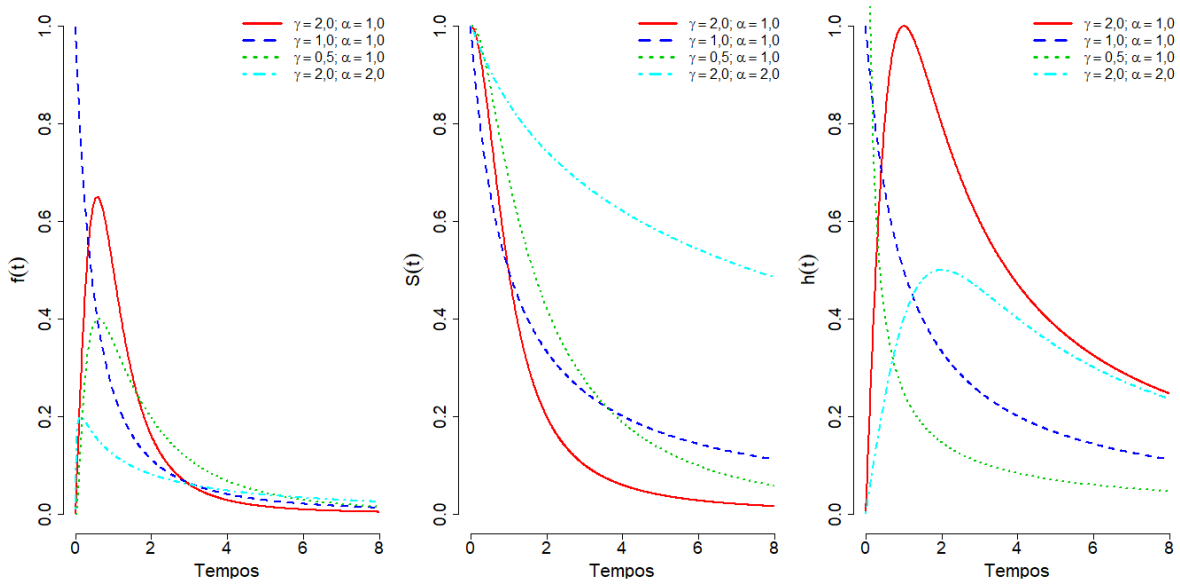

Figura 6 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Log-Logística com diferentes valores para γ e α.

2.1.7.7 Distribuição Burr XII

Segundo Paranaíba (2012), a distribuição Burr XII é comumente encontrada para modelar dados de sobrevivência, podendo assumir uma função de risco monótona e unimodal. Considerando uma variável aleatória não negativa, $T > 0$, segue uma distribuição Burr XII $(T \sim BurrXII(\theta,\alpha,\gamma))$, sua função de densidade de probabilidade, de sobrevivência e de taxa de falha são expressas, respectivamente, por:

$$f(t) = \frac{\gamma\alpha t^{\alpha-1}}{\theta^{\alpha}}\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)},$$

$$S(t) = \left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma},$$

e

$$h(t) = \frac{\gamma\alpha\left(\frac{t}{\theta}\right)^{\alpha-1}}{\theta\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]},$$

em que θ é o parâmetro de escala e α e γ são os parâmetros de forma, todos estritamente maiores que zero. Na Figura 7 são apresentas algumas formas para a função de densidade de probabilidade, sobrevivência e taxa de falha para a distribuição Burr XII.

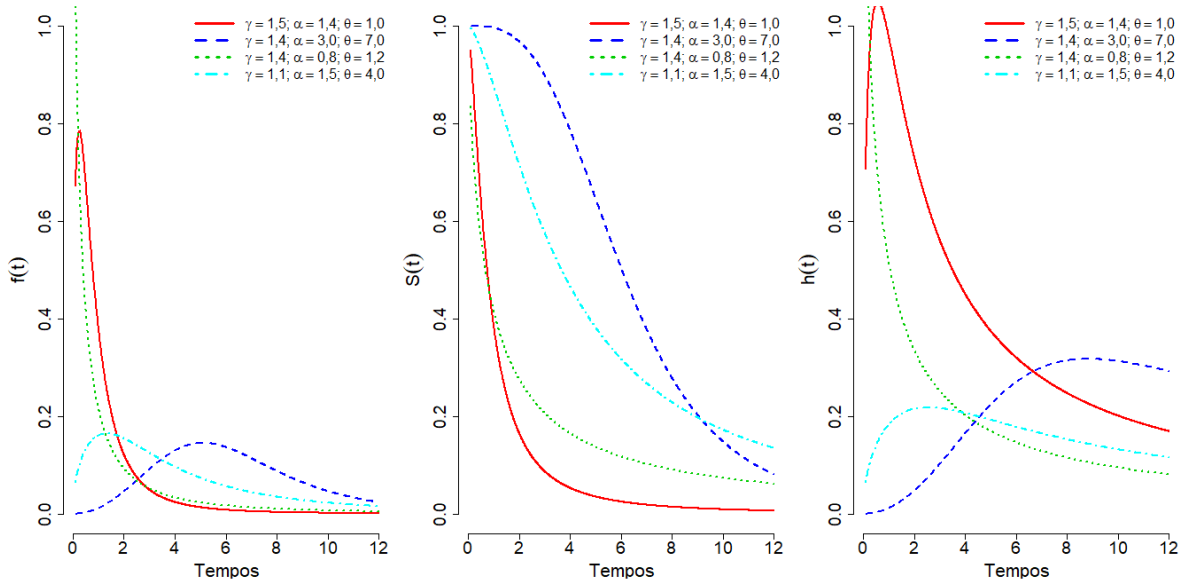

Figura 7 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Burr XII com diferentes valores para γ, α e θ.

2.1.7.8 Distribuição Beta-Weibull

De acordo com Peres (2016), a distribuição Beta-Weibull é uma generalização de alguns modelos probabilísticos já descritos na literatura, sendo versátil em análise de sobrevivência, pois sua função de risco pode assumir curvas monótonas e não monótonas, de acordo com os valores dos parâmetros.

Considerando uma variável aleatória não negativa, $T > 0$, segue uma distribuição Beta-Weibull $(T \sim BW(\lambda,\alpha,\beta,\gamma))$, sua função de densidade de probabilidade, de sobrevivência e de risco são definidas da seguinte forma:

$$f(t) = \frac{\gamma t^{\gamma-1}}{B(\alpha,\beta)\lambda^{\gamma}} \exp\left[-\beta\left(\frac{t}{\lambda}\right)^{\gamma}\right]\left\{1-\exp\left[-\left(\frac{t}{\lambda}\right)^{\gamma}\right]\right\}^{\alpha-1},$$

$$S(t) = 1 - \frac{1}{B(\alpha,\beta)}\int_0^{1-\exp\{-(t/\lambda)\}} w^{\alpha-1}(1-w)^{\beta-1}dw,$$

e

$$h(t) = \frac{f(t)}{S(t)},$$

em que $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, λ é o parâmetro de escala e α, β e γ são parâmetros de forma, todos estritamente positivos. Algumas formas de suas funções em análise de sobrevivência são apresentadas na Figura 8.

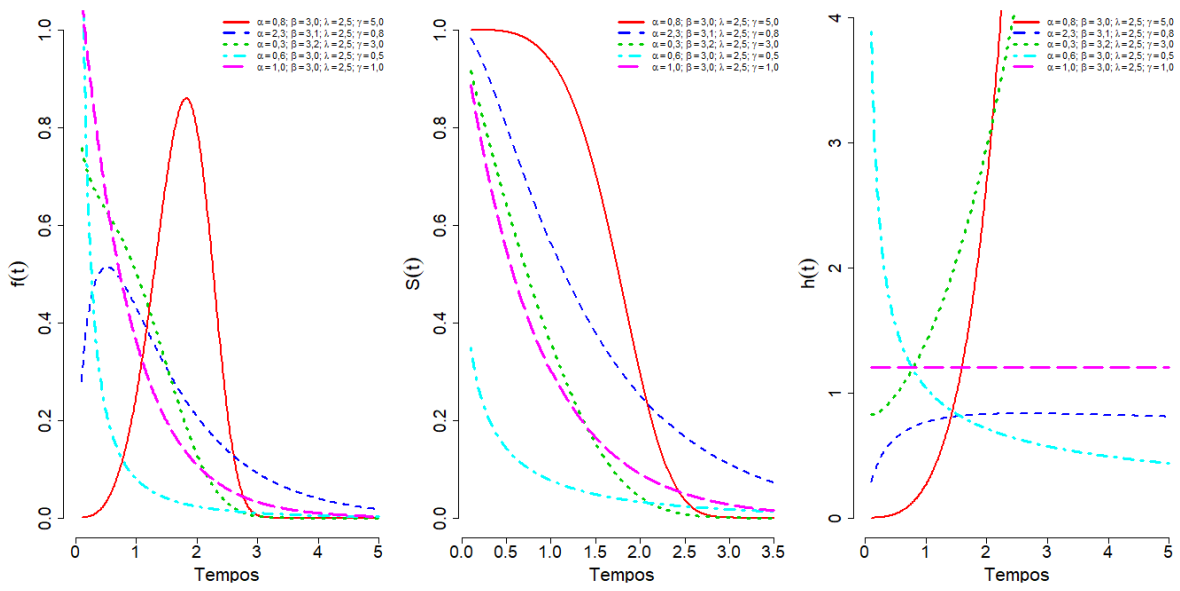

Figura 8 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Beta-Weibull com diferentes valores para α, β, λ e γ.

2.1.7.9 Distribuições da Família Kumaraswamy

A distribuição Kumaraswamy Gama Generalizada é um modelo probabilístico recente, apresentado por Pascoa, Ortega e Cordeiro (2011). De acordo com os autores, a distribuição possui grande flexibilidade para descrever dados de sobrevida, de modo que sua função de risco pode assumir comportamento monótono e não monótono.

Considerando uma variável aleatória não negativa, $T > 0$, segue uma distribuição Kumaraswamy Gama Generalizada $(T \sim KGG(\alpha,\tau,\varphi,k,\lambda))$, sua função de densidade de probabilidade, sobrevivência e risco são definidas por:

$$f(t) = \frac{\lambda\varphi\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\tau k-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}\left(1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi-1},$$

$$S(t) = \left(1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi},$$

e

$$h(t) = \frac{\lambda\varphi\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\tau k-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}\left(1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{-1},$$

em que $\gamma_1(k,x) = \gamma(k,x)/\Gamma(k) = \left(\int_0^x w^{k-1}e^{-w}dx\right)/\Gamma(k)$, α é o parâmetro de escala e τ, φ, k e λ são parâmetros de forma, todos estritamente positivos. Para $k = 1$, tem-se a distribuição Kumaraswamy Weibull, apresentada por Cordeiro, Ortega e Nadarajah (2010).

Na Figura 9, são apresentadas algumas formas características da função de densidade de probabilidade, de sobrevivência e de taxa de falha para a distribuição Kumaraswamy Gama Generalizada.

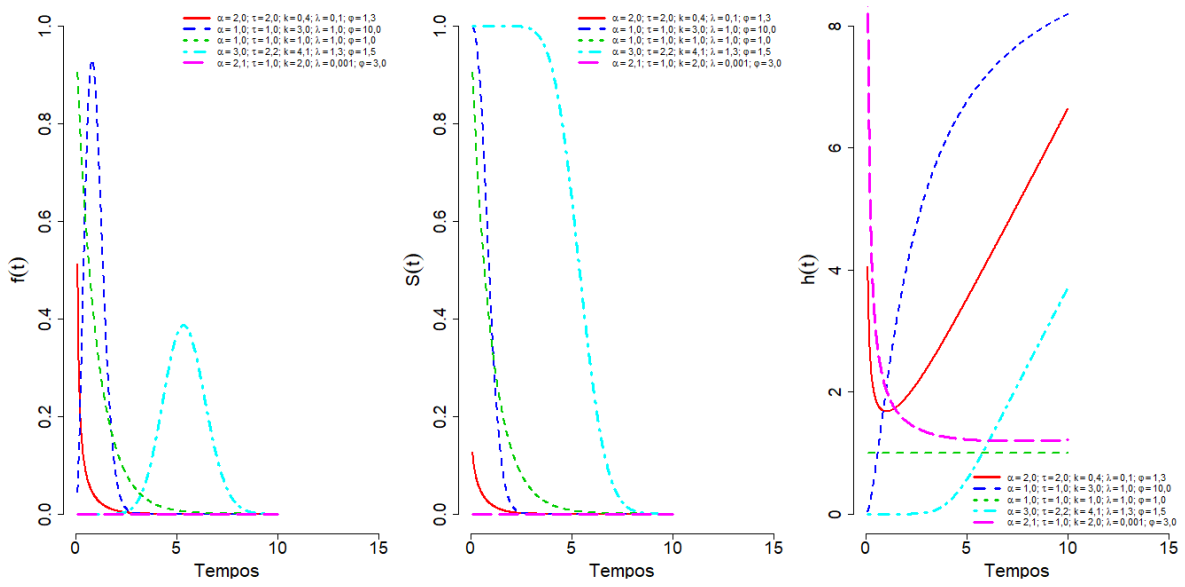

Figura 9 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Kumaraswamy Gama Generalizada com diferentes valores para seus parâmetros.

Uma outra distribuição proveniente da família Kumaraswamy Generalizada é a Kumaraswamy Burr XII, proposta por Paranaíba et al. (2013), que pode assumir função de risco monótona e não monótona.

Considerando uma variável aleatória não negativa, $T > 0$, segue uma distribuição Kumaraswamy Burr XII $(T \sim KBXII(\alpha,\beta,\lambda,\gamma,\theta))$, sua função de densidade de probabilidade, de sobrevivência e de risco são definidas por:

$$f(t) = \alpha\gamma\beta\lambda\theta^{-\alpha}t^{(\alpha-1)}\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{(\lambda-1)}$$
$$\times\left[1-\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{\lambda}\right]^{(\beta-1)},$$

$$S(t) = 1-\left\{1-\left[1-\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{\lambda}\right]^{(\beta)}\right\},$$

e

$$h(t) = \alpha\gamma\beta\lambda\theta^{-\alpha}t^{(\alpha-1)}\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{(\lambda-1)}$$
$$\times\left[1-\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{\lambda}\right]^{(-1)},$$

sendo α, β e λ e γ parâmetros de forma e θ o parâmetro de escala, todos estritamente positivos.

2.1.7.10 Distribuição Weibull Generalizada Exponenciada de Potência

A partir da família exponenciada, Pena-Ramirez et al. (2018) propem a distribuição Weibull Generalizada Exponenciada de Potência. De acordo com os autores, o modelo probabilístico é bastante flexível para os tipos comuns de função de risco.

Considerando uma variável aleatória não negativa, $T > 0$, segue uma distribuição Weibull Generalizada Exponenciada de Potência $(T \sim EPGW(\alpha,\beta,\beta,\gamma,\lambda))$, sua função densidade de probabilidade, sobrevivência e risco são definidas como:

$$f(t) = \alpha\beta\lambda\gamma t^{\gamma-1}\frac{(1+\lambda t^{\gamma})^{\alpha-1}\exp\left[1-(1+\lambda t^{\gamma})^{\alpha}\right]}{\left\{1-\exp\left[1-(1+\lambda t^{\gamma})^{\alpha}\right]\right\}^{1-\beta}},$$

$$S(t) = 1-\left[1-\exp\left\{1-(1+\lambda t^{\gamma})\right\}\right]^{\beta},$$

e

$$h(t) = \frac{f(t)}{S(t)},$$

em que α, β e γ são os parâmetros de forma e λ é o parâmetro de escala, todos estritamente positivos.

A Figura 10 apresenta algumas formas das funções características em análise de sobrevivência para a distribuição.

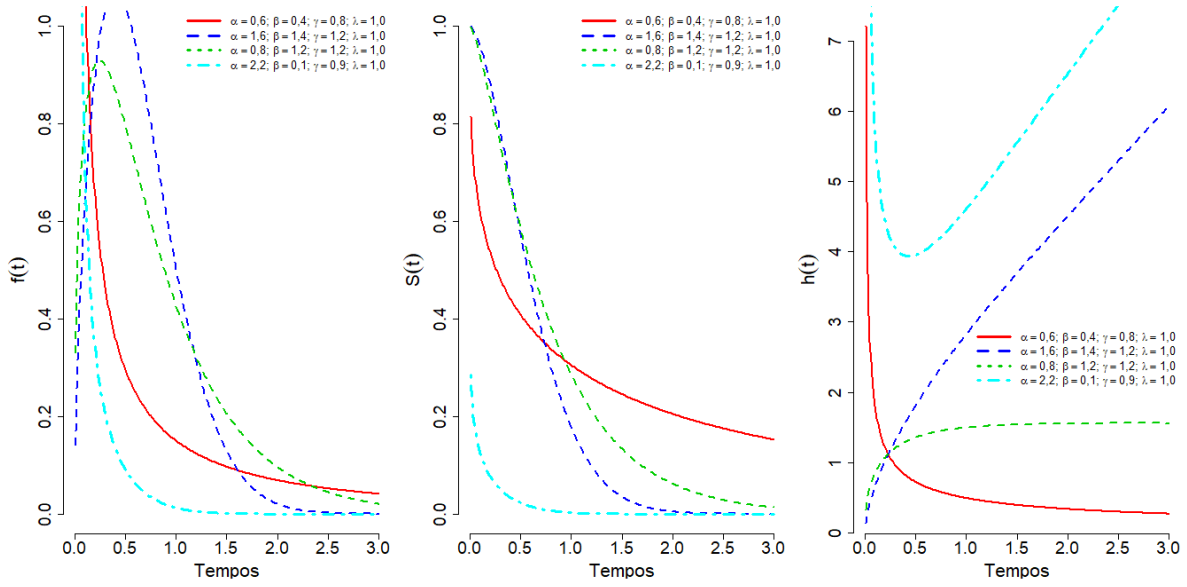

Figura 10 – Formas típicas para as funções de probabilidade, sobrevivência e taxa de falha, respectivamente, para a distribuição Weibull Generalizada Exponenciada de Potência com diferentes valores para seus parâmetros.

2.1.8 Inflação de Zeros e Fração de cura

2.1.8.1 Modelos com Inflação de Zeros

Grande parte das distribuições de probabilidade, tanto as amplamente conhecidas na literatura, bem como distribuições apresentadas mais recentemente, podem ser usadas para descrever fenômenos ou experimentos que produzem resultados reais não nulos.

Entretanto, ao analisar conjuntos de dados com valores contínuos, é possível que a situação de estudo apresente valores observáveis de zeros para a variável resposta, e quando a frequência é alta, é dito então que os dados possuem excesso de zeros, ou que são inflacionados de zeros. Ospina e Ferrari (2010) afirmam que essa proporção de zeros não permite que as observações sejam provenientes de uma distribuição contínua. Dessa forma, para analisar dados com excessos de zeros, devem ser utilizados modelos inflacionados, em que os dados são distribuídos de forma contínua com suporte no modelo probabilístico, mas que incluem observações em um dos seus extremos (HASHIMOTO, 2013).

Considerando T uma varável aleatória que representa o tempo até a ocorrência do evento de interesse, com excessos de zeros, segundo Calsavara et al. (2019), a função de sobrevivência é dada da seguinte forma:

$$S(t) = (1 - p_0)S_0(t),$$

onde $0 < p_0 < 1$ é o parâmetro que modela a proporção de tempos zero, $S_0(t)$ é a função de sobrevivência própria.

Desse modo, a função de probabilidade, em que $f_0(t)$ é a função de densidade de probabilidade própria, é dada da seguinte forma:

$$f(t) = \begin{cases} p_0 & \text{se} \quad t = 0 \\ (1 - p_0) f_0(t) & \text{se} \quad t \neq 0 \end{cases}$$

2.1.8.2 Modelos com Fração de Cura

Os modelos de sobrevivência com fração de cura podem ser utilizados para modelar dados de sobrevida onde observações possuem um longo tempo de sobrevivência. Nessa abordagem, parte-se da ideia de que a população em estudo é formada por indivíduos que são suscetíveis a experimentar o evento de interesse, bem como indivíduos que não são suscetíveis, podendo ser classificados como "imunes" ao evento de interesse (MALLER; ZHOU, 1996).

Nessa abordagem, a função de sobrevivência é dada da seguinte forma:

$$S(t) = p_1 + (1 - p_1) S_0(t),$$

em que $0 < p_1 < 1$ é a proporção de indivíduos que não estão suscetíveis a ocorrência do evento de interesse e $S_0(t)$ é a função de sobrevivência própria.

Já a função de densidade de probabilidade é definida da seguinte forma:

$$f(t) = (1 - p_1) f_0(t),$$

em que p_1 é a proporção de indivíduos "imunes" ao evento de interesse e $f_0(t)$ é a função de densidade de probabilidade própria.

Alguns trabalhos abordando modelos com fração de cura podem ser observados em Achcar, Barros e Mazucheli (2012), Martinez et al. (2013) e Peres, Santos e Oliveira (2020).

2.1.8.3 Modelos com Inflação de Zeros e Fração de Curados

Considerando dados que possuem inflação de zeros e também uma proporção de curados, ambas características podem ser incorporados nas análises, como pode ser observado em Júnior, Moreira e Louzada (2017) e Souza et al. (2021), definindo as funções de densidade de probabilidade e sobrevivência (em modelos de mistura), respectivamente, como:

$$f(t) = \begin{cases} p_0 & \text{se} \quad t = 0, \\ (1 - p_0 - p_1) f_0(t) & \text{se} \quad t \neq 0, \end{cases}$$

e

$$S(t) = p_1 + (1 - p_0 - p_1)S_0(t),$$

em que p_0 é a proporção de zeros, p_1 é a proporção de indivíduos que não apresentaram o evento de interesse, de modo que $0 < (p_0,p_1) < 1$, e $f_0(t)$ e $S_0(t)$ são, respectivamente, as funções de densidade de probabilidade e de sobrevivência próprias.

2.2 Métodos de Estimação

2.2.1 O Método da Máxima Verossimilhança

Exitem vários métodos estatísticos conhecidos na literatura que permitem realizar a estimação dos parâmetros de distribuições de probabilidade. Entretanto, na análise de sobrevivência, a presença de censuras acaba tornando alguns métodos de estimação inadequados.

O método da máxima verossimilhança apresentado por Fisher (1912), consiste em buscar a distribuição, dentre todas as possíveis de acordo com os valores dos parâmetros, que tenha a maior probabilidade de ter gerado os dados que estão sendo estudados. Em termos matemáticos, é de interesse encontrar os valores de um vetor $\boldsymbol{\theta}$ que maximize a função de verosimilhança $L(\boldsymbol{\theta})$.

Para uma melhor compreensão do método da máxima verossimilhança e visualização de seu uso para a estimação de parâmetros, ver Ramachandran e Tsokos (2015).

Em análise de sobrevivência, a contribuição de cada observação não censurada para $L(\boldsymbol{\theta})$ é sua função de densidade, já as observações censuradas, apenas informam que tempo de falha é maior que o tempo de censura observado, logo, sua contribuição para $L(\boldsymbol{\theta})$ é dada pela função de sobrevivência, $S(t)$. Assumindo uma amostra aleatória (t_i,δ_i), em que $i = 1,...,n$, a função de verossimilhança pode ser expressa por:

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

Os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ são os valores responsáveis por maximizar a função de verossimilhança $L(\boldsymbol{\theta})$, ou de forma equivalente, o logaritmo da função de verossimilhança, $\log L(\boldsymbol{\theta})$, em que estes são encontrados resolvendo o seguinte sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

2.2.1.1 Intervalo de Confiança

De acordo com Peres (2016), considerando que $\hat{\theta} \sim N(0,1)$, ou seja, o estimador do parâmetro θ_k segue uma distribuição Normal com média 0 e variância 1, então um intervalo de confiança assintótico para um parâmetro θ_k com nível de significância α é dado por:

$$\hat{\theta_k} \pm Z_{1-\alpha/2}\sqrt{Var(\hat{\theta_k})},$$

em que $Z_{1-\alpha/2}$ é o quantil de uma distribuição Normal padronizada com coeficiente de confiança assintótico α, e $Var(\hat{\theta_k})$ é a variância de $\hat{\theta_k}$, estimada pela matriz hessiana:

$$-\left[E(H(\hat{\theta}))\right]^{-1} = \begin{bmatrix} Var(\theta_1) & Cov(\theta_1,\theta_2) & \dots & Cov(\theta_1,\theta_k) \\ Cov(\theta_2,\theta_1) & Var(\theta_2) & \dots & Cov(\theta_2,\theta_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\theta_k,\theta_1) & Cov(\theta_k,\theta_2) & \dots & Var(\theta_k) \end{bmatrix}.$$

2.2.1.2 Função de Ligação

Na implementação de (co)variáveis ao modelo, se faz necessária a utilização de uma função de ligação, relacionando as (co)variáveis a seus respectivos parâmetros, de modo que, uma função monótona $g(\theta)$ relaciona o parâmetro θ a uma função linear.

A função de ligação mais utilizada é a Identidade, dada por: $g(\theta) = \theta$, entretanto Peres (2016) apresenta outras comuns utilizadas na análise de sobrevivência, além de discutir sobre a escolha da função de ligação adequada, de modo que é necessário conhecer o domínio do espaço paramétrico em questão.

2.2.2 Inferência Bayesiana

Em modelos frequentistas, os parâmetros de um modelo são tradados como escalares ou um vetor desconhecido fixo. Em contrapartida, em modelos Bayesiano, o escalar ou vetor aleatório desconhecido é quantificado em termos de probabilidade, seguindo uma distribuição *a priori*. A partir das informações *a priori* e dos dados amostrais, é possível modelar e atualizar as estimativas *a posteriori* por meio das regras de Bayes (PAULINO; TURKMAN; MURTEIRA, 2003; ROSSI, 2011).

De acordo com Achcar et al. (2019), considerando uma amostra aleatória $y = (y_1, y_2, ..., y_n)$ de dados independentes e identicamente distribuídos com uma distribuição conjunta dada pela densidade $f(y|\theta)$, representada pela função de verossimilhança, $L(y|\theta)$, e uma distribuição *a priori* para θ, dada por $\pi(\theta)$, por meio da regra de Bayes, tem-se a distribuição *a posteriori* para θ dado y,

$$\pi(\theta|y) \propto L(y|\theta)\pi(\theta).$$

Segundo Rossi (2011), a teoria frequentista muitas vezes é limitada em alguns aspectos, como a aceitação de algo não real (por exemplo, considerar de que os dados não normais seguem uma distribuição Normal) e a obtenção de estimativas absurdas em alguns casos, de modo que, o método Bayesiano pode ser uma alternativa para estas situações. Além disso, a utilização de informações prévias sobre o parâmetro de interesse podem gerar estimativas *a posteriori* mais coerentes com a realidade da variável de estudo, e a utilização das distribuições *a priori* não informativas permitem realizar comparações com os resultados frequentistas.

Martinez e Achcar (2014) também ressaltam algumas vantagens da utilização de métodos Bayesianos, de modo que estes podem ser uma saída para situações em que os modelos frequentistas apresentam problemas de significância ou convergência das estimativas dos parâmetros, função de verossimilhança muito complexas e problemas de identificabilidade.

Na abordagem Bayesiana, deseja-se obter-se as estimativas dos parâmetros *a posteriori*, além de seus intervalos de credibilidade, entretanto, de acordo com Assunção (2019), essas estimativas nem sempre são fáceis de serem obtidas, sendo necessária a utilização de algoritmos para se obter uma amostra da distribuição *a posteriori* e assim, estimar esperanças com respeito a esta distribuição via processos estocásticos.

Os algoritmos mais utilizados, os de Monte Carlo via Cadeias de Markov (MCMC), são usados para gerar uma amostra (aproximada) grande de uma certa distribuição, quando não se consegue gerar diretamente desta distribuição. A construção é dada por meio de uma cadeia de Markov de modo que uma série de propriedades devem ser satisfeitas para que o algoritmo tenha as propriedades desejadas (ASSUNÇÃO, 2019).

Outra abordagem para inferência Bayesiana é dada pelo método de Aproximação de Laplace Aninhada Integrada (*Integrated Nested Laplace Approximation* - INLA), proposta por Rue, Martino e Chopin (2009), sendo que uma das principais características é ser uma alternativa ao método MCMC em termos de acurácia e tempo computacional. De uma maneira geral, este método combina aproximação de Laplace para integrais com técnicas numéricas para matrizes esparsas.

Uma discussão detalhada sobre o processo de estimação pode ser observado nos trabalhos de Rue, Martino e Chopin (2009), Caballero (2013) e Rue et al. (2017), além do site (https://www.r-inla.org) do pacote *INLA* para o ambiente estatístico R.

Capítulo 3

Análise de sobrevivência para dados de HIV/AIDS: uma abordagem Bayesiana

Resumo

Este artigo[1] teve como objetivo estimar os parâmetros das funções de sobrevivência de algumas distribuições para dados sobre HIV/AIDS do Mato Grosso do Sul, Brasil, entre 2009 e 2018. Foram comparados alguns ajustes frequentistas e com base nos valores do logaritmo da função de verossimilhança e do Critério de Informação de Akaike (AIC), selecionou-se o melhor modelo para a realização de um ajuste Bayesiano, além de realizar uma comparação deste com o modelo padrão em análise de sobrevivência, o Weibull. Foi possível calcular as estimativas dos parâmetros, bem como comparar os dois modelos por meio dos valores do Critério de Informação da *Deviance* Bayesiana (DIC). Verificou-se que o modelo selecionado (Log-Normal) apresentou um melhor ajuste aos dados, comparado ao modelo padrão.

Palavras-chave: inferência Bayesiana; modelos probabilísticos; tempos de falha.

3.1 Introdução

A análise de sobrevivência consiste em uma análise estatística de dados, cuja variável resposta corresponde à quantidade de tempo observado de um instante inicial até a

[1] Este artigo foi publicado na revista Brazilian Journal of Development: BUENO, M. V.; ROSSI, R. M. Análise de sobrevivência para dados de HIV/AIDS: uma abordagem Bayesiana. *Brazilian Journal of Development*, Curitiba, v.7, n.6, p. 55797 - 55805 jun. 2021. DOI: 10.34117/bjdv7n6-133.

ocorrência, ou não, do evento de interesse. Segundo Bastos e Rocha (2006), algumas datas importantes para o desenvolvimento desse tipo de análise podem ser citadas, como a publicação do livro "*Natural and Political Observations upon the Bill of Mortality*" por John Graunt em 1662 e o desenvolvimento da primeira tabela de mortalidade por Edmund Halley em 1693.

Na abordagem Bayesiana, o parâmetro em estudo é tratado como uma variável aleatória, quantificada em termos de probabilidade, desse modo, diz-se que o parâmetro segue uma distribuição *a priori*. A partir dessas distribuições e dos dados amostrais, são encontradas as estimativas *a posteriori* por meio da regra de Bayes. Existem inúmeras aplicações em diferentes áreas de estudo dessa abordagem, como exemplificado em (ROSSI, 2011).

Segundo Brito, Castilho e Szwarcwald (2001), a AIDS destaca-se entre as enfermidades infecciosas, devido a sua grande magnitude e a extensão de seus danos causados às populações, a qual desde sua origem, suas características e repercussões são amplamente discutidas pela comunidade científica.

Vários trabalhos podem ser observados envolvendo dados de HIV/AIDS utilizando técnicas não paramétricas, como em Müller e Borges (2020), Melo, Donalisio e Cordeiro (2017), Medeiros et al. (2017), entre outros. Entretanto, ainda há uma escassez de trabalhos para essa mesma temática de dados utilizando abordagens paramétricas, tanto frequentistas quanto Bayesianas.

Partindo disso, este trabalho teve como objetivo estimar os parâmetros para a função de sobrevivência de algumas distribuições e apresentar o melhor modelo sob a ótica frequentista quando aplicados à dados de HIV/AIDS, assim como fazer uma breve comparação ao modelo padrão de análise de sobrevivência, o Weibull, sob o ponto de vista Bayesiano.

3.2 Materiais e Métodos

O conjunto de dados utilizado pertence a um estudo epidemiológico, transversal e analítico apresentado por Werle (2021), realizado com 9.021 notificações por HIV/AIDS no Estado do Mato Grosso do Sul, Brasil, no período de 2009 a 2018, sobre o tempo (em semanas) de acompanhamento até a morte por AIDS. O estudo originário atendeu as normas nacionais e internacionais de ética em pesquisa com seres humanos, e foi aprovado com o parecer n° 3.789.678.

A caracterização dos dados de sobrevivência é apresentada por meio dos tempos de falha ou de censuras. O tempo de falha é a medida de tempo entre a exposição do indivíduo ou objeto ao determinado fator estudado, já definido previamente para a realização do estudo, até o evento de interesse (falha) ocorrer. As censuras representam observações

incompletas ou parciais, devido a não ocorrência do evento de interesse ou a perda de acompanhamento do paciente durante o período de estudo (COLOSIMO; GIOLO, 2006).

Para a análise dos dados foi considerando como critério de exclusão, observações que apresentavam tempo de falha igual a '0' (zeros, inconsistências de registros). Foram removidas um total de 309 observações e, dessa forma, o banco de dados final conteve 8.712 observações.

Utilizando os pacotes *survival* e *flexsurv*, do ambiente estatístico R (R Core Team, 2022), e partindo de uma abordagem frequentista, inicialmente foram testadas algumas distribuições de probabilidade amplamente conhecidas na literatura para a representação dos dados, a saber: Exponencial, Weibull, Gama, Gama Generalizada, Log-Normal e Log-logística. Neste processo inicial, as estimativas dos parâmetros foram obtidas pelo método da máxima verossimilhança. Foi escolhida a distribuição que melhor representasse os dados com base nos valores do logaritmo da função de verossimilhança e do Critério de Informação de Akaike (AIC).

A partir destes resultados, objetivou-se comparar, por meio de técnicas Bayesianas, o melhor modelo avaliado com um modelo seguindo a distribuição Weibull, considerada padrão na literatura e possuindo grande usabilidade em estudos biomédicos. Considerou-se então duas situações, sendo uma em que a variável aleatória tempo (em semanas) até a falha segue uma distribuição Log-Normal e outra em que a mesma segue uma distribuição Weibull. As distribuições de probabilidade e suas respectivas funções de sobrevivência associadas são apresentadas na Tabela 1, em que $\Phi(.)$ é a função de distribuição acumulada de uma Normal padrão.

Tabela 1 – Distribuições de probabilidade e função de sobrevivência para as distribuições Weibull e Log-Normal.

Distribuição	Função Densidade de Probabilidade	Função de Sobrevivência
Weibull	$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}$	$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \ t > 0$
Log-Normal	$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(t)-\mu)^2}{2\sigma^2}\right\}$	$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right), \ t > 0$

Considerando a distribuição Weibull sabe-se que γ e α são parâmetros de forma e escala, respectivamente. O parâmetro γ não possui unidade de medida e α possui a mesma unidade de medida de t. Já no modelo Log-Normal, os parâmetros μ e σ são, respectivamente a média e o desvio-padrão do logaritmo natural do tempo de falha.

O estimador de (KAPLAN; MEIER, 1958), também conhecido como estimador produto-limite, é uma adaptação da função empírica de sobrevivência (probabilidade de que um indivíduo sobreviva em um tempo t, ou seja, $P(T \geq t)$), que permite calcular as esti-

mativas para os tempos de sobrevivência. Por se tratar de uma abordagem não paramétrica, as estimativas são calculadas sem a necessidade de fazer a suposição de uma distribuição de probabilidade para a variável resposta. Sua expressão geral é definida por $\hat{S}(t) = \prod_{i:t_i<t}\left(1 - \frac{d_i}{n_i}\right)$, em que d_i é o número de falhas no tempo i e n_i o número de observações vivas e não censuradas no tempo i.

A partir da função de sobrevivência estimada pelo método do produto-limite, caso seja interesse do estudo, é possível obter as demais funções básicas em análise de sobrevivência, como por exemplo, a função de taxa acumulada $\Lambda(t) = -\ln S(t)$ (CARVALHO et al., 2011).

Para a modelagem Bayesiana, em ambas as situações, foram consideradas distribuições *a priori* não informativas para os parâmetros, conforme a parametrização do OpenBUGS, Spiegelhalter et al. (1994), programa específico para tal análise. No modelo Weibull, atribuiu-se as seguintes distribuições *a priori* $\gamma \sim$ Gama$(0{,}001; 0{,}001)$ e $\lambda \sim N(0; 0{,}000001)$, onde $\alpha = \exp(-\lambda/\gamma)$. No modelo Log-Normal, foram atribuídas $\mu \sim N(0; 0{,}000001)$ e $\tau \sim$ Gama$(0{,}001; 0{,}001)$, onde $\sigma = 1/\sqrt{\tau}$. Para uma abordagem mais detalhada, ver Ibrahim, Chen e Sinha (2001).

Para obtenção das estimativas *a posteriori* dos parâmetros, foi utilizado o pacote *BRugs* (vinculado ao OpenBugs) do R, o qual foram gerados 110.000 valores simulados em um processo Monte Carlo via Cadeias de Markov (MCMC), com saltos de tamanho 50 e descarte inicial de 10.000 valores. A verificação da convergência das cadeias foi realizada pelo critério de Heidelberger e Welch, Heidelberger e Welch (1983), via pacote *coda* do R. A comparação entre os modelos Bayesianos foi feita pelo Critério de Informação da *Deviance* Bayesiana (DIC).

Para a verificação da qualidade dos ajustes, utilizaram-se os resíduos de Cox-Snell. Segundo Lawless (2003) os resíduos devem ser provenientes de uma população homogênea e seguir uma distribuição Exponencial padronizada caso o modelo seja adequado.

3.3 Resultados e discussão

O gráfico das estimativas de Kaplan-Meier, permitiu a observação do comportamento da função da sobrevivência empírica para os dados (Figura 11). A Tabela 2 apresenta os resultados para o logaritmo das funções de verossimilhança e o AIC das distribuições consideradas. De acordo com os resultados, é possível observar que a distribuição Log-Normal é a mais adequada dentre as avaliadas, visto que apresentou o menor valor do AIC (=12.792,54).

A partir da escolha do melhor ajuste (Log-Normal) sob a ótica frequentista paramétrica,

Tabela 2 – Resultados do logaritmo das funções de verossimilhança e AIC para as distribuições consideradas.

Distribuição	$\log(L(\theta))$	AIC
Exponencial	-7.127,07	14.256,14
Weibull	-6.418,45	12.840,91
Gama	-6.420,42	12.844,85
Gama Generalizada	-6.403,17	12.812,35
Log-Normal	**-6.394,27**	**12.792,54**
Log-Logística	-6.416,60	12.837,20

na sequência, por meio do procedimento Bayesiano, um resumo das estimativas *a posteriori* deste modelo Log-Normal e do modelo padrão (Weibull) são apresentadas na Tabela 3.

Observa-se que para o modelo Log-Normal, a média e a mediana das estimativas dos parâmetros *a posteriori* não apresentam grande diferença, indicando distribuição simétrica *a posteriori*.

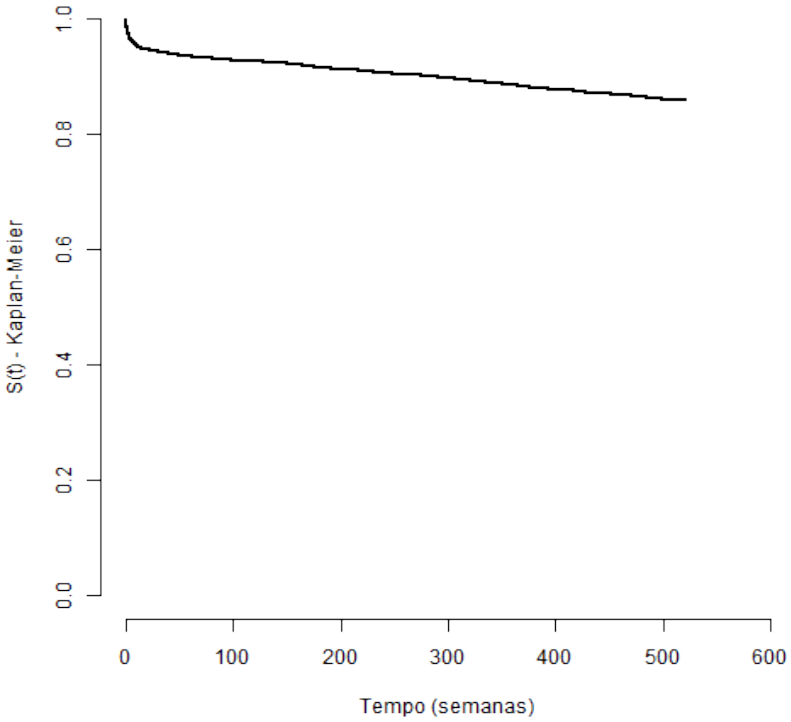

Figura 11 – Curva de sobrevivência empírica via Kaplan-Meier.

Pode-se observar pelos valores dos DIC e pelas comparações entre a curva de sobrevivência empírica de Kaplan-Meier e as funções de sobrevivência paramétricas com

Tabela 3 – Estimativas *a posteriori* dos parâmetros e seus intervalos com 95% de credibilidade $[P_{2,5\%}; P_{97,5\%}]$.

Distribuição	Parâmetro	Média	Desvio-Padrão	Mediana	DIC	$P_{2,5\%}$	$P_{97,5\%}$
Weibull	α	119.100	27.980	115.900	12.840	74.400	184.700
	γ	0,363	0,012	0,362		0,340	0,387
Log-Normal	μ	12,818	0,265	12,816	12.790	12,304	13,337
	σ	5,770	0,169	5,764		5,450	6,101

estimativas Bayesianas para os parâmetros (Figura 12), que as duas distribuições estão ajustadas aos dados de forma similar. As comparações das curvas de sobrevivências frequentistas e Bayesianas com a função de sobrevivência empírica para cada modelo (Figura 13), indicam que as curvas estimadas por métodos Bayesianos e frequentistas também tiveram ajustes similares.

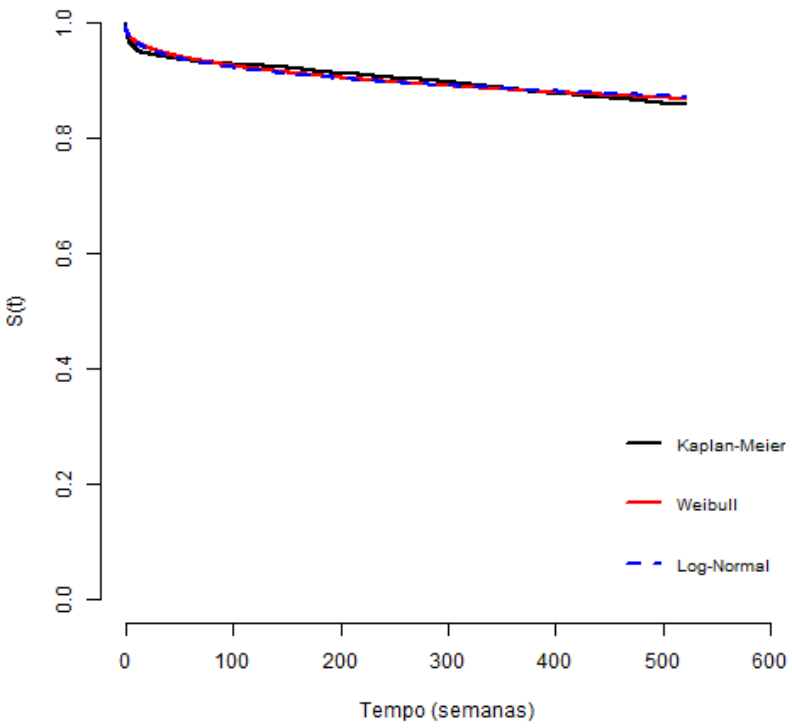

Figura 12 – Curvas de sobrevivência estimadas via Kaplan-Meier e procedimentos Bayesianos para as distribuições Weibull e Log-Normal.

Essa similaridade pode ser observada nos resultados dos cálculos de sobrevivência para os tempos t. Por exemplo, calculando a probabilidade de que um indivíduo sobreviva a pelo menos 170 semanas, são obtidos os valores de 91,03%, 90,84% e 91,82% segundo o modelo Weibull, Log-Normal e as estimativas de Kaplan-Meier, respectivamente.

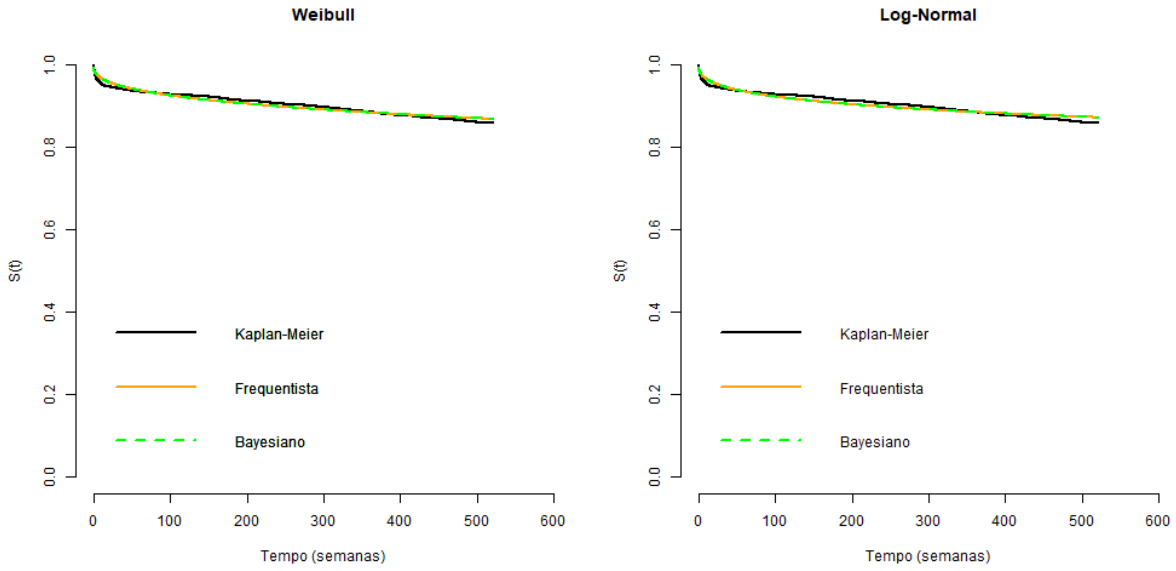

Figura 13 – Curvas de sobrevivência dos modelos Weibull e Log-Normal estimadas via métodos frequentistas e Bayesianos, comparadas com as estimativas de Kaplan-Meier.

A análise de resíduos de Cox-Snell indica que ambas distribuições abordadas não apresentaram um ajuste adequado aos dados (Figura 14). Talvez, algumas justificativas para tal fato, sejam a possibilidade de os dados terem sido inseridos no sistema do SUS de forma irregular ou a própria natureza dos dados, e neste caso, discute-se sobre a necessidade de buscar outas distribuições de probabilidade mais flexíveis e com uma maior quantidade de parâmetros, como por exemplo a Kumaraswamy Gama Generalizada, apresentada por Pascoa, Ortega e Cordeiro (2011), possa melhorar a qualidade dos ajustes.

A inclusão de (co)variáveis no ajuste de modelos, não abordado no presente estudo, é algo ainda a ser investigado, devido ao grande interesse nas áreas da saúde e de políticas públicas. Além de que, podem ter sua parcela de colaboração no bom ajuste do modelo aos dados.

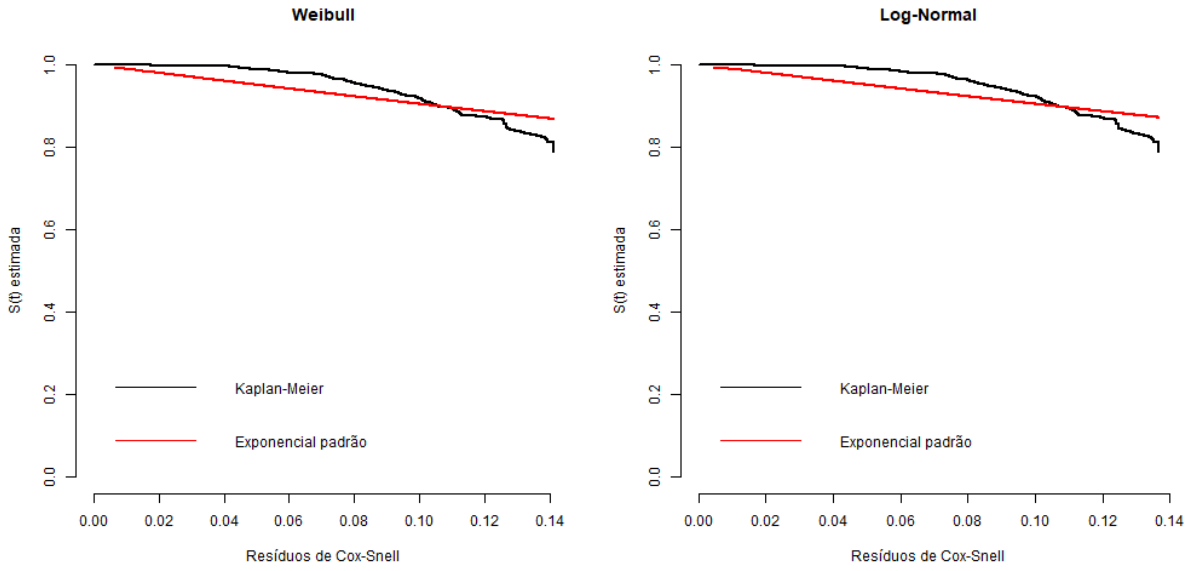

Figura 14 – Análise de resíduos de Cox-Snell para os modelos Weibull e Log-Normal

3.4 Conclusões

O modelo Log-Normal foi o que apresentou melhor ajuste aos dados de HIV/AIDS. Entretanto, acredita-se que para uma melhor performance dos ajustes, deva-se considerar modelos probabilísticos mais flexíveis, permitindo ajustar a curva de sobrevivência de uma forma adequada.

Capítulo 4

Fatores associados ao tempo de sobrevida para pacientes com HIV/AIDS no Estado Mato Grosso do Sul: abordagem paramétrica

Resumo

Neste artigo[1], objetivou-se ajustar algumas distribuições de probabilidade para o tempo de sobrevida em pacientes com HIV/AIDS no Mato Grosso do Sul, Brasil, acompanhados entre 2009 e 2018, por meio de métodos frequentistas e Bayesianos. A partir de modelos de regressão, foi possível estimar o efeito de variáveis explicativas sobre a variável resposta. Com base nos valores do critério de informação de Akaike (AIC) e o no logaritmo de máxima verossimilhança, foi verificado que a distribuição Log-Normal foi a mais adequada para o ajuste aos dados. Foram construídos dois modelos de regressão com base nas metodologias descritas, de modo que ambos convergiram para uma mesma interpretação das variáveis explicativas: sexo, raça, escolaridade, e uso de drogas injetáveis. Verificou-se que o tempo mediano até a morte por HIV/AIDS é, aproximadamente: 2,1 vezes maior para indivíduos do sexo feminino, 1,8 vezes maior para indivíduos brancos, 5,4 vezes maior para indivíduos com escolaridade superior a 8 anos e 5,5 vezes maior para indivíduos que não fazem uso de drogas injetáveis. Por meio das interpretações dos coeficientes dos parâme-

[1] Este artigo foi aceito para publicação na Revista Brasileira de Biometria, previsto para 30 de setembro de 2022. BUENO, M. V.; ROSSI, R. M. Factors associated with survival time for patients with HIV/AIDS in the State of Mato Grosso do Sul: Parametric approach. Revista Brasileira de Biometria, v. 40, n. 3, 2022.

tros do modelo, pode-se discutir a necessidade de políticas de prevenção e diagnóstico precoce focadas em grupos que apresentam um menor tempo mediano de sobrevida após a notificação da infecção pelo vírus do HIV.

Palavras-chave: análise de sobrevivência; modelos probabilísticos; tempos de falha; inferência bayesiana.

4.1 Introdução

A análise de sobrevivência consiste em uma área da estatística que estuda a duração do tempo até a observação de um evento de interesse, em que este tempo é denominado tempo de falha. Alguns exemplos podem ser apresentados como o tempo até a morte, cura ou recidiva de uma doença em um paciente.

Segundo Colosimo e Giolo (2006), a principal características de dados de sobrevivência é a presença de observações parciais das respostas, chamadas de censuras. Desse modo, o termo análise de sobrevivência refere-se a situações médicas que envolvem dados censurados.

Em um contexto histórico, algumas datas importantes para o desenvolvimento deste tipo de análise estatística são apresentadas por Bastos e Rocha (2006), sendo estas a publicação do livro "*Natural and Political Observations upon the Bill of Mortality*", por John Graunt em 1962 e o desenvolvimento da primeira tabela de mortalidade em 1693 por Edmund Halley, semelhantes às utilizadas atualmente em demografia.

Por meio de modelos de regressão, é possível verificar o efeito de variáveis explicativas sobre a variável resposta, neste caso, o tempo. Na análise de sobrevivência, tem-se na literatura duas classes de modelos propostos, sendo estes o modelo semiparamétrico, também denominado simplesmente de modelo de Cox, e o paramétrico.

Para a utilização dos modelos paramétricos, é necessário realizar a suposição de que a variável resposta segue uma distribuição de probabilidade. O modelo é composto por um componente aleatório, que descreve em termos de probabilidade o comportamento do tempo de sobrevivência e um componente sistemático, que descreve a relação entre os parâmetros do modelo probabilístico e as (co)variáveis explicativas (CARVALHO et al., 2011).

Em contrapartida à modelagem frequentista, tem-se a Bayesiana, que tem ganhado uma maior relevância ao longo dos anos na literatura devido ao avanço do poder computacional para a realização das análises. Na inferência bayesiana, considera-se que os parâmetros em estudo são aleatórios e caracterizados por uma distribuição *a priori*, e a partir dela e dos dados amostrais, são encontradas suas estimativas *a posteriori* por meio

da regra de Bayes (ROSSI, 2011).

A síndrome da imunodeficiência adquirida (AIDS) é uma doença crônica que causa danos ao sistema imunológico, causada pelo vírus HIV. O Boletim Epidemiológico de HIV/AIDS de 2020 apresenta dados referentes a HIV e AIDS no Brasil. Segundo este, no ano de 2019 foram diagnosticados 41.909 novos casos de HIV e 37.308 casos de AIDS. De 1980 a junho de 2020, um total de 1.011.617 casos de AIDS foram identificados no Brasil. Em relação a mortalidade no país, entre 1980 e 2019, 349.784 óbitos tiveram o HIV/AIDS como causa básica, os quais, 57,7% pertencem a região Sul, 17,8% a região Nordeste, 5,3% a região Centro-Oeste e 5,3% a região Norte (BRASIL, 2020).

Em uma discussão sobre os aspectos clínicos, Neto et al. (2021) discorrem que a infecção por HIV pode ser classificada em três fases, sendo estas: *Infecção Aguda pelo HIV*, a qual ocorre entre as três primeiras semanas de infecção e apresenta sintomas inespecíficos como febre, cefaleia, astenia, adenopatia, entre outros; *Latência Clínica*, em geral assintomática e com duração de anos, gerando queda gradual de LT-CD4+ (células de defesa), com aparecimento intermitente de infecções ou reativação de infecções antigas; *AIDS*, onde ocorrem manifestações de imunodeficiência avançada, em que seus indicativos são o aparecimento de infecções oportunistas ou neoplasias.

Ribeiro, Fonseca e Pereira (2019), apresentam uma análise exploratória de dados espaciais sobre a incidência de AIDS por Estado no Brasil entre 1992 e 2017, permitindo visualizar a evolução da doença ao longo dos anos por meio de técnicas gráficas. Entre os resultados apresentados, é discutido sobre a aparição dos primeiros casos nos anos 80 na região Sudeste, seu espalhamento para as regiões Sul e Centro-Oeste nos anos 90 e para o Norte e Nordeste a partir do ano de 2000.

Embora Colosimo e Giolo (2006) relate que a análise de sobrevivência tenha sido uma das áreas estatísticas que mais cresceu nas últimas duas décadas do século passado, ainda há uma escassez de trabalhos para dados de HIV/AIDS em um contexto paramétrico, tanto para uma abordagem frequentista quanto Bayesiana.

Neste contexto, o presente trabalho teve como objetivo ajustar uma distribuição de probabilidade para o tempo (em semanas), desde o diagnóstico, até o óbito de pacientes com HIV/AIDS para dados do Mato Grosso do Sul, Brasil, acompanhados entre 2008 e 2019, bem como verificar qual (quais) variável (is) explicativa (as), dentre algumas consideradas na literatura, são relevantes para o modelo e indicam associação à variável resposta utilizando os procedimentos frequentistas e Bayesianos no contexto com a análise de sobrevivência.

4.2 Materiais e Métodos

O conjunto de dados analisados referem-se a um estudo epidemiológico, transversal e analítico, realizado com dados provenientes das notificações por HIV/AIDS, no período de 2009 a 2018. Os dados são abordados por Werle (2021), contendo inicialmente um total de 9.021 indivíduos para a Unidade Federativa do Mato Grosso do Sul, Brasil. Ao verificar o conjunto de dados, foram encontradas observações que apresentavam o mesmo dia para notificação e morte (0 dias até a morte por HIV/AIDS), além de informações perdidas para algumas variáveis explicativas. Dessa forma, optou-se por remover essas observações, tal que o conjunto de dados analisado contou com 4.681 observações. O estudo originário dos dados atendeu as normas nacionais e internacionais de ética em pesquisa com seres humanos, e foi aprovada com parecer n° 3.789.678. No presente estudo, dispensa-se a apreciação pelo Comitê de Ética em Pesquisa (CEP), do Conselho Nacional de Saúde (CNS) e da Comissão Nacional de Ética em Pesquisa (CONEP) baseado na Resolução 466/2012 CEP-CONEP.

Além da resposta, tempo (em semanas), e o indicador de censura, o conjunto de dados conta com informações sobre algumas variáveis explicativas: faixa etária; sexo; raça; escolaridade; transmissão sexual; transmissão vertical; uso de drogas injetáveis; transfusão para hemofílicos; transfusão de sangue e acidente com material biológico. Na Tabela 4, são apresentadas as classes das variáveis explicativas, bem como as frequências absolutas e relativas de cada uma.

A principal característica de dados de sobrevivência é a presença de observações parciais da resposta, chamadas de censuras. Uma observação censurada indica que o indivíduo em questão, por alguma razão, não apresentou o evento de interesse, como exemplo, ter morrido por outra causa, ter mudado de cidade ou o estudo ter terminado antes do evento do evento de interesse ocorrer. A presença de uma observação censurada indica que o tempo de falha é superior àquele observado (COLOSIMO; GIOLO, 2006).

No contexto de análise de sobrevivência, o tempo de falha geralmente é especificado pelas funções de sobrevivência e de taxa de falha. A função de sobrevivência é definida como a probabilidade de que o indivíduo observado sobreviva a um determinado tempo t, em termos probabilísticos tem-se $S(t) = P(T \geq t)$. Consequentemente, a função de distribuição acumulada permite encontrar a probabilidade de que o indivíduo não sobreviva ao tempo t, ou seja, $F(t) = 1 - S(t)$.

Tabela 4 – Resumo descritivo para as variáveis explicativas entre 2009 e 2018.

Variáveis	Classes	Frequência	%	Censuras (%)
Sexo	Masculino	2.953	63,08	90,45
	Feminino	1.728	36,92	91,49
Raça/Cor	Branca	2.068	44,18	92,12
	Não branca	2.613	55,82	89,82
Escolaridade (em anos)	Maior que 8	1.986	43,43	94,31
	Menor ou igual a 8	2.695	57,57	88,27
Uso de drogas injetáveis	Sim	153	3,27	80,33
	Não	4.528	96,73	91,19

A função de risco, ou função de taxa de falha, descreve a forma em que a taxa instantânea de falha muda ao decorrer do tempo. É necessário ressaltar que a função de risco fornece o valor de uma taxa, e não uma probabilidade, podendo assumir qualquer valor positivo diferente de 0. Matematicamente, a função de risco é expressa da seguinte forma: $h(t) = \lim_{\Delta t \to 0} \dfrac{P(t \leq T < t + \Delta t | T \geq T)}{\Delta t}$. A partir desta, a função de risco acumulado pode ser encontrada integrando a função de risco em um intervalo $[0,t]$.

Dentre as funções de sobrevivência é possível verificar algumas relações matemáticas importantes entre as funções de sobrevivência, como por exemplo a função de risco, expressa por $h(t) = \dfrac{f(t)}{S(t)}$. Essas relações são descritas e discutidas em Lee e Wang (2003), e mostram que a partir do conhecimento de uma, é possível obter as demais.

Embora a função de risco também seja de grande importância no contexto de análise de sobrevivência, neste trabalho, optou-se por fazer uma discussão apenas da função de sobrevivência.

Uma importante técnica não paramétrica em análise de sobrevivência é o estimador de Kaplan e Meier (1958), também conhecido como estimador limite-produto. Este estimador é uma adaptação da função empírica de sobrevivência, e utiliza conceitos de probabilidade condicional e independência de eventos, e definido como: $\prod_{i:t_i<t} \left(1 - \dfrac{d_i}{n_i}\right)$, em que considera-se $t_1 < t_2 < ... < t_k$, como os k tempos de falha distintos e ordenados, d_i o número de falhas em um tempo t_i e n_i o número de indivíduos em risco em t_i.

O modelo semiparamétrico riscos proporcionais de Cox (1972), permite estimar os efeitos das variáveis explicativas sem a necessidade de realizar suposições referentes a distribuição de probabilidade do tempo de sobrevivência. Segundo Carvalho et al. (2011), este modelo é dito semiparamétrico pois não assume qualquer distribuição probabilística para a função de risco basal, $h_0(t)$. De forma genérica, considerando p variáveis explicativas e o vetor de componentes $\boldsymbol{x} = (x_1,...,x_p)'$, a expressão geral do modelo de

regressão de Cox é dada por: $h(t|\boldsymbol{x}) = h_0(t)g(\boldsymbol{x'\beta})$. Colosimo e Giolo (2006) afirmam que o componente paramétrico é frequentemente usado na seguinte forma multiplicativa: $g(\boldsymbol{x'\beta}) = \exp\{\boldsymbol{x'\beta}\} = \exp\{\beta_1 x_1 + ... + \beta_p x_p\}$.

Métodos não paramétricos e/ou modelos de regressão de Cox são mais comuns em análise de dados de pacientes com HIV/AIDS, como vistos em Colosimo e Vieira (1996), Medeiros et al. (2017), Müller e Borges (2020) e Melo, Donalisio e Cordeiro (2017). De modo geral, o objetivo dessas análises é buscar associações entre os variáveis explicativas e a doença sem considerar que os dados seguem uma distribuição de probabilidade.

Em contrapartida, tem-se que os modelos de regressão paramétricos. Para ajustar tais modelos, é necessário supor uma distribuição de probabilidade para a variável resposta, que define a função de sobrevivência. Para os dados abordados neste trabalho, foram consideradas as seguintes distribuições de probabilidade: Exponencial, Weibull, Gama, Gama Generalizada, Log-Normal e Log-Logística.

Em contrapartida à metodologia frequentista tem-se a Bayesiana, em que o parâmetro (considerado aleatório) é quantificado em termos de probabilidade, em formalmente diz-se que este segue uma distribuição *a priori*, em que, a partir das informações amostrais e *a priori*, pode-se modelar e atualizar as estimativas dos parâmetros *a posteriori* por meio da regra de Bayes. Para um melhor discussão sobre a Inferência Bayesiana, ver Box e Tiao (1973) e Rossi (2011). Aplicações da inferência Bayesiana no contexto de análise de sobrevivência podem ser vistas em Santos e Achcar (2011), Brunello e Nakano (2015), entre outras.

Para a seleção do melhor modelo probabilístico, serão utilizados os critérios de informação de Akaike (AIC), além do logaritmo da função de máxima verossimilhança $(\log L(\theta))$.

A função de densidade de probabilidade para cada modelo probabilístico são apresentados na Tabela 5, de modo que suas funções de sobrevivência, risco e risco acumulado podem ser obtidas com base nas relações entre funções apresentadas em Lee e Wang (2003).

Neste trabalho, para a verificação da distribuição de probabilidade para a variável resposta, bem como o modelo de regressão paramétrico, foram utilizados o ambiente estatístico R (R Core Team, 2022), bem como os pacotes estatísticos *survival* (THERNEAU, 2020), e *flexsurv* (JACKSON, 2016). Para a estimação dos parâmetros dos modelos de regressão, utilizou-se o método da máxima verossimilhança para o modelo frequentista e o método de aproximação de Laplace aninhada integrada, via pacote INLA (RUE; MARTINO; CHOPIN, 2009), para o modelo Bayesiano.

A partir do melhor modelo de probabilidade ajustado aos dados, dentre os testados,

Tabela 5 – Função densidade de probabilidade dos modelos probabilísticos considerados.

Modelo	Função densidade de probabilidade
Exponencial	$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$
Weibull	$f(t) = \frac{\gamma}{\alpha^{\gamma}} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$
Gama	$f(t) = \frac{t^{k-1}}{\Gamma(k)\alpha^{k}} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$
Gama Generalizada	$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma^{k}}} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$
Log-Normal	$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^{2}\right\}$
Log-Logística	$f(t) = \frac{\gamma}{\alpha^{\gamma}} r^{\gamma-1}\left[1+\left(\frac{t}{\alpha}\right)^{\gamma}\right]^{-2}$

objetiva-se realizar dois modelos de regressão, abordando as metodologias frequentista e Bayesiana.

No contexto de análise de sobrevivência, a função de verossimilhança para os modelos de regressão é definida por:

$$L(\theta) = \prod_{i=1}^{n} \left[f(t_i|x_i)\right]^{\delta_i} \left[S(t_i|x_i)\right]^{1-\delta_i}, \tag{4.1}$$

sendo δ_i o indicador de censura.

A interpretação dos coeficientes estimados para os modelos Log-Normal e Weibull é discutida em Colosimo e Giolo (2006), de modo que para uma variável binária, é necessário fazer uso da razão de tempos medianos:

$$\frac{t_{0,5}(x=1,\hat{\beta})}{t_{0,5}(x=0,\hat{\beta})} = e^{\hat{\beta}} \tag{4.2}$$

Essa interpretação também pode ser estendida para (co)variáveis categóricas e contínuas, de modo que uma discussão rigorosa da interpretação das quantidades estimadas pode ser vista em Hosmer e Lemeshow (1999).

A análise de resíduos constitui-se como uma etapa de grande importância na construção de um modelo de regressão, visto que avaliar a adequação do modelo ajustado é parte fundamental da análise de dados. Para a verificação dos ajustes dos modelos, utilizou-se os resíduos de Cox e Snell (1968).

A técnica de verificação consiste em calcular os resíduos com base na função de taxa de falha acumulada obtida por meio do modelo estimado $(\hat{e}_i = \hat{\Lambda}(t_i|x_i))$ e observar se esses seguem uma distribuição Exponencial padronizada (LAWLESS, 2003).

4.3 Resultados e Discussão

Uma análise inicial dos dados consistiu em verificar a curva de sobrevivência estimada por Kaplan-Meier (Figura 15), na qual observa-se uma queda brusca na sobrevivência por HIV/AIDS nas primeiras semanas.

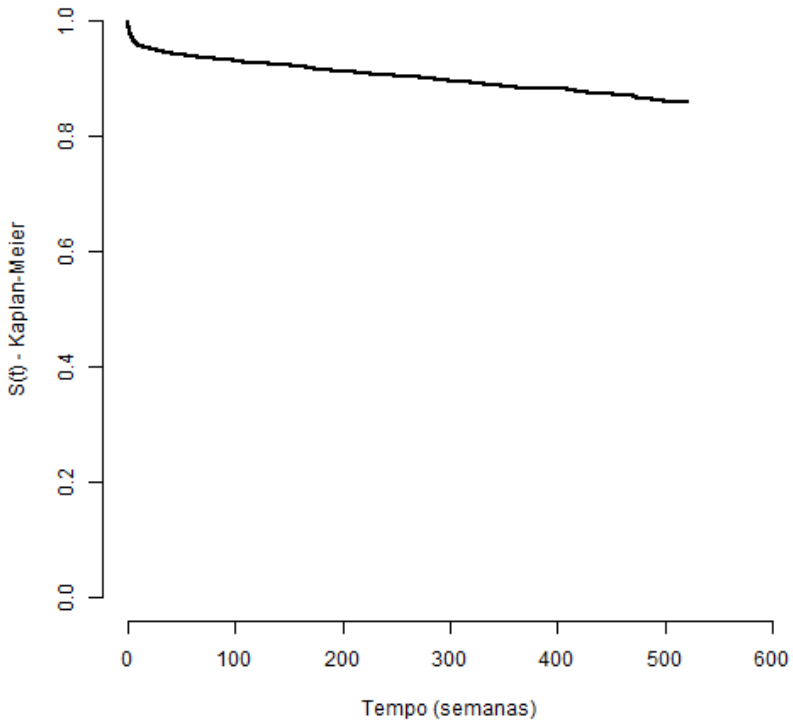

Figura 15 – Curva de sobrevivência empírica de HIV/AIDS.

Após ajustes frequentistas, considerando as distribuições apresentadas na Tabela 5, os resultados (Tabela 6) apontam para uma melhor adequabilidade da distribuição Log-Normal para a variável resposta, visto que este modelo apresentou o menor AIC (6.879,91) e o maior $\log L(\theta)$ (-3.437,95). Com base nisto, considerou-se então que a variável resposta segue uma distribuição de probabilidade Log-Normal para o próximo passo da análise.

Em especial, o modelo de regressão Log-Normal é dado por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu(\boldsymbol{x}'\boldsymbol{\beta})}{\sigma}\right)^2\right\}, \quad t > 0, \tag{4.3}$$

de modo que σ é o parâmetro de dispersão e μ é o parâmetros de locação, de forma que $\mu(\boldsymbol{x}'\boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, por meio da função de ligação identidade.

Tabela 6 – Resultados para o AIC e $\log L(\theta)$ para os modelos probabilísticos considerados.

Modelo	AIC	$\log L(\theta)$
Exponencial	7.538,97	-3.768,48
Weibull	6.902,78	-3.449,39
Gama	6.904,64	-3.450,32
Gama Generalizada	6.888,81	-3.441,40
Log-Normal	**6.879,91**	**-3.437,95**
Log-Logística	6.901,08	-3.448,54

Santos e Nakano (2015), utilizam o modelo de regressão Log-Normal para uma comparação com o modelo de Cox, trabalhando com dados sobre o tempo de permanência de trabalhadores no mercado de trabalho no Distrito Federal.

Outra aplicação do modelo de regressão Log-Normal pode ser observada em Silva et al. (2018), no qual os autores no contexto de análise de sobrevivência, trabalhando com dados sobre a evasão dos alunos do curso de Estatística da Universidade Federal da Paraíba.

Em um contexto Bayesiano, para o modelo Log-Normal, as seguintes distribuições *a priori* são consideradas *default* segundo a parametrização do *INLA* no R:

$\tau \sim$ Log-Gama$(1; 0{,}00005)$;

$\beta_0 \sim N(0; 0{,}001)$;

$\beta_j \sim N(0; 0{,}001), \;\; j = 1, 2, ..., p,$

em que $\sigma = \dfrac{1}{\sqrt{\tau}}$, β_0 é o intercepto e β_j são os parâmetros associados a cada variável explicativa.

Werle (2021), por meio de um modelo de regressão de Cox nos dados, realizou a verificação da significância dos parâmetros associados às variáveis explicativas, ao nível de 5% de significância. Com base nos resultados discute-se sobre o tempo até o óbito para os indivíduos de determinados grupos presentes no estudo. São apontadas as variáveis explicativas significativas para o modelo final (selecionado com base nas variáveis explicativas indicadas como significativas no trabalho original dos dados) ajustado: sexo, transmissão sexual, raça (trabalhada em outra codificação), escolaridade e uso de drogas injetáveis.

Com base nos resultados apresentados por Werle (2021), foram testados diferentes modelos, em que foi verificado que a presença da variável explicativa transmissão sexual fazia com que a variável explicativa sexo apresentasse um efeito contrário ao observado nas curvas de sobrevivência empíricas. Desse modo, os modelos finais contaram com as seguintes variáveis explicativas: sexo, raça, escolaridade e uso de drogas injetáveis.

Foram construídos modelos de regressão paramétricos para os dados em questão, utilizando a distribuição Log-Normal, considerando os métodos frequentista e Bayesiano. Os estimativas dos parâmetros de ambos os resultados são apresentados na Tabela 7. Os resultados evidenciam a similaridade entre ambos os ajustes.

Os resíduos de Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo Exponencial padronizado, bem como a curva de sobrevivência estimada para o modelo Bayesiano são apresentados na Figura 16. Optou-se por não apresentar os gráficos de resíduos de Cox-Snell para o modelo frequentista, visto que apresentam comportamento análogo ao Bayesiano.

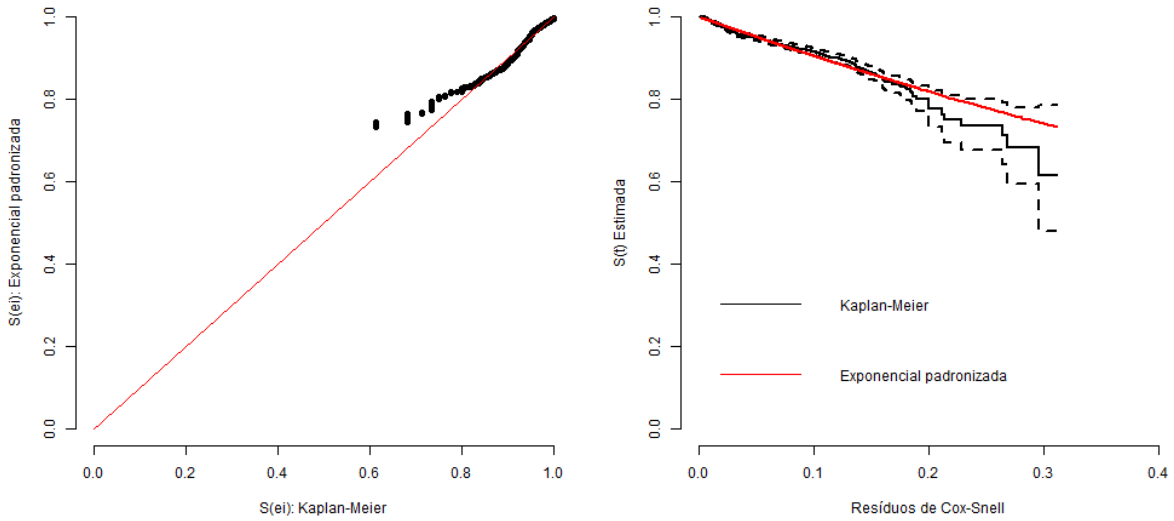

Figura 16 – Resíduos de Cox-Snell do modelo de regressão Bayesiano estimadas pelo método de Kaplan-Meier e pelo Exponencial padronizado (à esquerda) e respectivas curvas de sobrevivência estimadas (à direita).

Nota-se que ambos os modelos apresentaram resultados análogos, além de que, não apresentam problemas de falta de ajuste, visto que as curvas não se distanciam da curva de sobrevivência de Kaplan-Meier para os resíduos.

De acordo com a Equação 4.2, pode-se interpretar a razão dos tempos medianos de sobrevivência para o modelo Log-Normal. Ambos modelos apresentam interpretações extremamente parecidos, de modo que, segundo o modelo Bayesiano, o tempo mediano até a morte por HIV/AIDS é aproximadamente:

- 2,1 vezes maior para o indivíduos do sexo feminino;
- 1,8 vezes maior para indivíduos de raça branca;

Tabela 7 – Estimativas dos parâmetros do modelos de regressão frequentista e Bayesiano.

Parâmetro	Classes	E.M.V. (E.P.)	IC 95%	$e_F^{\hat{\beta}}$	M$_P$ (E.P.)	ICr 95%	$e_B^{\hat{\beta}}$
β_0	-	13,527 (0,467)	(12,611 ; 14,443)	-	13,596 (0,469)	(12,714 ; 14,547)	-
σ	-	5,345 (0,218)	(4,934 ; 5,790)	-	5,419 (0,078)	(4,155 ; 6,962)	-
β_1: Sexo	Masculino	-	-	-	-	-	-
	Feminino	0,757 (0,290)	(0,189 ; 1,326)	2,133	0,762 (0,293)	(0,192 ; 1,340)	2,142
β_2: Raça	Branca	-	-	-	-	-	-
	Não branca	-0,589 (0,281)	(-1,141 ; -0,038)	0,554	-0,592 (0,284)	(-1,153 ; -0,038)	0,553
β_3: Escolaridade (em anos)	Maior que 8	-	-	-	-	-	-
	Menor ou igual a 8	-1,677 (0,307)	(-2,280 ; -1,074)	0,187	-1,691 (0,310)	(-2,312 ; -1,094)	0,184
β_4: Drogas injetáveis	Não	-	-	-	-	-	-
	Sim	-1,701 (0,629)	(-2,932 ; -0,470)	0,182	-1,716 (0,633)	(-2,954 ; -0,466)	0,180

E.M.V.: Estimativas de máxima verossimilhança; E.P.: Erro padrão; IC: Intervalo de confiança; $e_F^{\hat{\beta}}$: Razão dos tempos medianos frequentistas; M$_P$: Média *a posteriori*; ICr: Intervalo de credibilidade; $e_B^{\hat{\beta}}$: Razão dos tempos medianos Bayesianos.

- 5,4 vezes maior para indivíduos com escolaridade superior a 8 anos;
- 5,5 vezes maior para indivíduos que não fazem uso de drogas injetáveis.

Segundo os resultados apresentados, a menor escolaridade e o uso de drogas injetáveis foram associados à uma menor sobrevida, de maneira similar os resultados apresentados pelo de Cox em Werle (2021).

A variável explicativa sexo é comum no modelo de Cox com dados de um hospital da cidade de João Pessoa (PB), Brasil, discutidos por Melo, Donalisio e Cordeiro (2017), onde o sexo feminino foi relacionado a uma sobrevida mais prolongada, em comum com a resultados apresentados neste estudo.

De maneira similar, em comparação com os resultados do modelo de Cox apresentados por Müller e Borges (2020), para dados da Região de Campos Gerais (PR), Brasil, os autores discutem que o sexo feminino, a raça branca e uma maior escolaridade foram associados a uma maior sobrevida, de maneira similar aos resultados encontrados neste trabalho para o modelo paramétrico.

Com base nos modelos estimados, é possível calcular estimavas para o risco e sobrevivência de pacientes notificados com HIV do Estado do Mato Grosso do Sul de acordo com as variáveis explicativas.

O desenvolvimento de novas distribuições de probabilidade é de suma importância da área da Estatística, pois podem promover maior flexibilidade nas aplicações. Baseados na distribuição Kumaraswamy, apresentada por Jones (2009), Pascoa, Ortega e Cordeiro (2011) introduzem e discutem as propriedades da distribuição Kumaraswamy Gama Generalizada. A importância deste modelo probabilístico segundo os autores, está relacionada a sua capacidade de modelar funções de taxa de falha monótonas e não monótonas, além de que, sua função de densidade de probabilidade possui a capacidade de se ajustar a dados distorcidos que não podem ser ajustados por outras distribuições já conhecidas, em razão da flexibilidade de suas cadas, possibilitando assim seu uso em diversas áreas (PASCOA; ORTEGA; CORDEIRO, 2011). Utilizando um conjunto de dados com informações sobre 143 crianças nascidas no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, entre os anos de 1995 e 2001, expostas ao vírus HIV por transmissão vertical, os pesquisadores utilizam técnicas paramétricas, no contexto frequentista e Bayesiano para evidenciar a eficiência do modelo.

Além disso, na Figura 15 observa-se que o comportamento das estimativas empíricas de sobrevivência indicam a presença de uma fração de sobreviventes de longa duração (possivelmente imunes ao evento de interesse), indicando que a inclusão de um parâmetro de fração de cura poderia melhorar a qualidade dos ajustes.

Dito isso, temos motivação suficiente para investigar tal distribuição, dentre outras, com a presença de um parâmetro para modelar a fração de curados, em novas pesquisas, considerando os dados de HIV/AIDS aqui apresentados.

4.4 Conclusões

O modelo de regressão Log-Normal foi o mais adequado para o ajuste dos dados de sobrevivência de pacientes com HIV/AIDS do Estado do Mato Grosso do Sul, Brasil, considerando métodos frequentistas e Bayesianos, cujas estimativas dos parâmetros foram análogas.

Com base nas interpretações dos coeficientes dos parâmetros das variáveis preditoras, pode-se discutir a necessidade de políticas de prevenção e diagnóstico precoce focadas em grupos que apresentam um menor tempo mediano de sobrevida após a notificação da infecção pelo vírus do HIV.

Capítulo 5

Modelos estatísticos para a análise de sobrevivência de pacientes com HIV/AIDS

Resumo

Considerando o tempo de sobrevida, em semanas, de pacientes com HIV/AIDS no Estado do Mato Grosso do Sul, acompanhados durante o período entre 2009 e 2018, o presente trabalho teve como objetivo ajustar alguns modelos estatísticos para a análise de sobrevivência dos dados descritos. Foram consideradas as distribuições Weibull, Burr XII, Gama, Gama Generalizada, Log-Normal, Beta-Weibull, Kumaraswamy Gama Generalizada, Kumaraswamy Weibull, Kumaraswamy Burr XII e Weibull Generalizada Exponenciada de Potência. A distribuição Kumaraswamy Gama Generalizada apresentou um ajuste mais plausível para dados, com base nos valores do critério de informação de Akaike e de uma inspeção visual entre as estimativas de Kaplan-Meier e as funções de sobrevivência estimadas.

Palavras-chave: dados censurados, modelos probabilísticos, tempo de falha.

5.1 Introdução

A análise de sobrevivência é definida como a área da estatística que estuda o tempo partindo de um período inicial até a ocorrência de um evento de interesse. Segundo Bastos e Rocha (2006), dois eventos são marcados como importantes para o surgimento deste tipo de análise estatística, sendo eles, a publicação do livro *"Natural and Political Observations upon the Bill of Mortality"* em 1962, de John Graunt, que continha informações sobre óbitos classificados em causa da morte, data da morte, sexo e idade, e a criação da primeira tabela de mortalidade, por Edmund Halley em 1963.

Os dados de sobrevivência são caracterizados por meio do tempo de falha da observação e pela presença ou não de censuras. O tempo de falha é a variável que representa a quantidade de tempo do indivíduo ou objeto ao fator de exposição, já definido antes da realização do estudo, até o evento de interesse (falha) ocorrer. As censuras representam as observações incompletas ou parciais, em que o evento de interesse não ocorreu ou houve a perda do acompanhamento do paciente durante o estudo (COLOSIMO; GIOLO, 2006).

A modelagem estatística é um dos instrumentos mais importantes em estudos epidemiológicos. Na literatura, são propostos dois mecanismos de geração de incertezas para a justificativa da utilização de modelos probabilísticos, sendo estes o pressuposto da existência de uma população infinita, com parâmetros populacionais desconhecidos, os quais se pretende estimar através de uma amostra coletada, e o pressuposto de que a mensuração das características dos indivíduos presentes na população obedece a uma lei estocástica comum a todos esses indivíduos (CARVALHO et al., 2011).

Identificada em 1981, a síndrome da imunodeficiência adquirida (AIDS), causada pela infecção do vírus da imunodeficiência humana (HIV), destaca-se entre as enfermidades infecciosas por sua grande magnitude e extensões de danos causados em populações. No Brasil, era considerada inicialmente como uma doença restrita a grandes metrópoles que atingia principalmente indivíduos, na grande maioria, homossexuais masculinos e hemofílicos. Ao passar do tempo, houve uma disseminação geográfica e registrados aumentos na transmissão heterossexual e de usuários de drogas injetáveis (BRITO; CASTILHO; SZWARCWALD, 2001).

Algumas informações ao nível global são apresentadas no site do Programa das Nações Unidas Sobre HIV/AIDS (UNAIDS, 2020), revelando que cerca de 38 milhões de pessoas viviam com HIV no ano de 2019, das quais cerca de 7,1 milhões não sabiam possuir a doença, e havendo um total de, aproximadamente, 690.000 óbitos relacionados à AIDS para o mesmo ano. Outros dados importantes são relacionadas à quantidade de pessoas com acesso à terapia anti-retroviral (cerca de 26 milhões acessando até o final de junho de

2020) e sobre novas infecções (cerca de 1,7 milhões de pessoas infectadas em 2019).

Segundo dados do Boletim Epidemiológico de 2020 (BRASIL, 2020), 342.459 casos de infecção pelo vírus HIV no Brasil foram relatados ao SINAN (Sistema Nacional de Notificações) entre os anos de 2007 e 2020, dos quais 25.966 (7,6%) foram registrados na região Centro-Oeste. Apenas para o ano de 2019, registrou-se 41.919 casos de infecções, dos quais 3.802 (9,1%) correspondem a região Centro-Oeste, sendo a região com os menores percentuais em um comparativo com as demais. Entretanto, em relação ao coeficiente de mortalidade no ano de 2019, os Estados do Mato Grosso do Sul e Mato Grosso apresentaram, respectivamente, os coeficientes de mortalidade (óbitos por 100 mil habitantes) de 5,3 e 4,4, fazendo parte das onze Unidades Federativas com coeficiente de mortalidade superior ao nacional, de 4,1.

Com intuito de estimar a taxa de sobrevivência de pacientes com AIDS desde o início da doença até a morte, Poorolajal et al. (2016) realizam uma metanálise na base de dados do PubMed, Web of Science e Scopus, pesquisados para fevereiro de 2016, além de verificar a lista de referências dos estudos e dados da Sociedade Internacional de AIDS (IAS). Ao total, os autores incluíram 57 estudos (25 de coortes prospectivos e 2 de retrospectivos), envolvendo um total de 294.662 participantes. Os principais resultados foram: a probabilidade de sobrevivência de 2, 4, 6, 8, 10 e 12 anos, da progressão do diagnóstico de HIV para o início da AIDS, em 82%, 72%, 64%, 57%, 26% e 19%, respectivamente; a probabilidade de sobrevivência de 2, 4, 6 e 10 anos, da progressão do início da AIDS para morte em pacientes recendo o tratamento HAART (Terapia anti-retroviral), em 87%, 86%, 78%, 78% e 61%; a probabilidade de sobrevivência de 2, 4 e 6 anos, de progressão do início da AIDS para morte em pacientes que não recebiam o tratamento HAART, em 48%, 26% e 16%, respectivamente.

Generalizações e/ou modificações nas distribuições de probabilidade são abordadas frequentemente na literatura, com a finalidade de apresentar um ajuste mais flexível à dados por meio de uma abordagem paramétrica, em diferentes métodos de estimação, entretanto há uma escassez de trabalhos para dados de HIV/AIDS.

Diante disto, o presente trabalho teve como objetivo verificar a adequação de alguns modelos estatísticos para a análise de sobrevivência de pacientes infectados pelo vírus HIV no Estado do Mato Grosso do Sul, Brasil.

5.2 Materiais e Métodos

O conjunto de dados analisados é provenientes de Werle (2021), contendo informações sobre ao tempo (em semanas) de acompanhamento de indivíduos com HIV/AIDS, a partir

da confirmação do diagnóstico positivo da doença, para a Unidade Federativa do Mato Grosso do Sul, entre os anos de 2009 e 2018.

O estudo originário dos dados atendeu as normas nacionais e internacionais de ética em pesquisa com seres humanos, e foi aprovado com parecer n° 3.789.678. No presente estudo, dispensa-se a apreciação pelo Comitê de Ética em Pesquisa (CEP), do Conselho Nacional de Saúde (CNS) e da Comissão Nacional de Ética em Pesquisa (CONEP) baseado na Resolução 466/2012 CEP-CONEP.

O conjunto de dados possui originalmente informações referentes ao tempo até a falha ou censura de 9.021 pacientes infectados com o vírus HIV. Foi utilizado como critério de exclusão indivíduos com tempo de falha igual a 0 (os quais possivelmente receberam o diagnóstico após o óbito), visto que as distribuições de probabilidade consideradas neste estudo não possuem suporte para incorporar essas observações no modelo. Dessa forma, o conjunto de dados contou com um total de 8.712 pacientes, dos quais 90,67% são censuras, ou seja, tempos que não foi observado o evento de interesse.

O gráfico de tempo total em teste (TTT-plot), proposto por Barlow e Campo (1975), possibilita a verificação gráfica para a identificação da forma da função de taxa de falha, permitindo assim a escolha de modelos mais apropriados para a variável tempo em estudo. Neste sentido, foi construído o TTT-plot para os dados descritos, com objetivo de se verificar o comportamento da função taxa de falha é compatível com a distribuição escolhida.

O estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), é uma adaptação da função de sobrevivência empírica (a probabilidade de que uma observação não apresente o evento de interesse até um determinado tempo t). Por se tratar de uma abordagem não paramétrica, as estimativas são calculadas sem a suposição de uma distribuição de probabilidade para o tempo de sobrevivência, e dessa forma, não existem parâmetros a serem estimados. A partir das estimativas de Kaplan-Meier, foi possível observar o comportamento que a função de sobrevivência deve seguir.

Métodos não paramétricos e/ou modelos de regressão de Cox (1972) são comuns em análise de dados de pacientes com HIV/AIDS (COLOSIMO e VIEIRA, 1996; MELO, DONALISIO e CORDEIRO, 2017; MEDEIROS et al., 2017; MÜLLER e BORGES, 2020 entre outros) e, em geral, buscam associações entre fatores e a doença. Entretanto, mesmo que escassos, há estudos de cunho teórico e de aplicação, através de métodos paramétricos, que enfatizam o uso de novas famílias de distribuições de probabilidade e que promovam melhores inferências.

A função de sobrevivência é fundamental para a realização de uma análise de sobrevivência, pois fornece informações cruciais resumidas das observações da amostra. Ela é

definida como a probabilidade de que um indivíduo sobreviva mais que um tempo t especificado, ou seja, $S(t) = P(T \geq t)$. Outra função importante no estudo de sobrevivência é a função taxa de falha, ou função de risco. Esta função permite calcular a probabilidade do evento de interesse ocorrer em um determinado tempo t, definida por: $h(t) = f(t)/S(t)$. Consequentemente, a função de distribuição acumulada pode ser definida como a probabilidade do indivíduo não sobreviver o tempo t, ou seja, $F(t) = 1 - S(t)$ (KLEINBAUM; KLEIN, 2011).

Por meio dos gráficos das estimativas de Kaplan-Meier e do TTT-plot, foram testadas algumas distribuições de probabilidade (Tabela 8), buscando uma função de sobrevivência que melhor descreva o comportamento das estimativas de Kaplan-Maier.

A distribuição Weibull, descrita em detalhes por Weibull (1939), é amplamente conhecida na literatura. Segundo Colosimo e Giolo (2006), sua popularidade em aplicações deve-se ao fato da distribuição possuir uma grande variedade de formas, bem como sua função taxa de falha ser monótona.

Segundo Paranaíba (2012), a distribuição Burr XII, proposta por Burr (1942), em particular, é muito utilizada em dados de sobrevivência e para modelagem de funções de taxa de falha monótonas e unimodais. Uma aplicação desta distribuição pode ser observada para dados do tempo até a primeira postura de codornas em Peralta et al. (2017).

Uma generalização da distribuição Gama foi proposta por Stacy et al. (1962), adicionando mais um parâmetro de forma, tornando a distribuição mais flexível. Devido a uma melhor adequabilidade, optou-se por usar a parametrização de Prentice (1974), em que se $\gamma \sim$ Gama$(Q^{-2},1)$, e $w = \frac{\log(Q^2\gamma)}{Q}$, então $t = \exp(\mu + tw)$ segue uma distribuição Gama generalizada (conhecida como Log-Gama). Sua aplicação em análise de sobrevivência é discutida por Cox et al. (2007).

Outra distribuição comumente utilizada para caracterizar tempos de vida de indivíduos e de produtos é a Log-Normal. Nesse modelo probabilístico, os dados podem ser analisados segundo uma distribuição Normal, porém, considerando o logaritmo natural dos dados em vez dos dados originais (COLOSIMO; GIOLO, 2006).

A distribuição Beta-Weibull, proposta por Lambert et al. (2007), é uma extensão da distribuição Weibull convencional, buscando se ajustar de modo mais satisfatório a conjuntos de dados reais. Segundo Peres (2016), a distribuição Beta-Weibull é uma generalização de algumas distribuições já descritas na literatura, sendo elas a Weibull, a Exponencial Weibull e a Beta-Exponencial.

Tabela 8 – Distribuições de probabilidade consideradas para os ajustes dos dados.

Modelo	Função densidade de probabilidade	Autor
Weibull (W)	$f(t) = \frac{\tau}{\alpha^\tau} t^{\tau-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\tau\right\}$	Weibull (1939)
Burr XII (BXII)	$f(t) = \frac{\gamma\alpha t^{\alpha-1}}{\theta^\alpha}\left[1+\left(\frac{t}{\theta}\right)^\alpha\right]^{-(\gamma+1)}$	Burr (1942)
Gama (G)	$f(t) = \frac{1}{\Gamma(\gamma)\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$	Lukacs (1955)
Log- Normal (LN)	$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}$	Aitchison e Brown (1957)
Gama Generalizada (GG)	$f(t) = \frac{\lvert Q\rvert (Q^{-2})^{Q^{-2}}}{\sigma t \Gamma(Q^{-2})} \exp\left[Q^{-2}(Qw - \exp(Qw))\right]$	Prentice (1974)
Beta-Weibull (BW)	$f(t) = \frac{\gamma t^{\gamma-1}}{B^{(1)}(\alpha,\beta)\lambda^\gamma}$ $\exp\left[-\beta\left(\frac{t}{\lambda}\right)^\gamma\right]\left\{1-\exp\left[-\left(\frac{t}{\lambda}\right)^\gamma\right]\right\}^{\alpha-1}$	Lambert et al. (2007)
Kumaraswamy Weibull (KW)	$f(t) = \frac{\lambda\varphi\tau}{\alpha}\left(\frac{t}{\alpha}\right)^{\tau-1} \exp\left[-\left(\frac{t}{\alpha}\right)^\tau\right]\left\{\gamma_1^{(2)}\left[1,\left(\frac{t}{\alpha}\right)^\tau\right]\right\}^{\lambda-1}$ $\times\left(1-\left\{\gamma_1^{(2)}\left[1,\left(\frac{t}{\alpha}\right)^\tau\right]\right\}^\lambda\right)^{\varphi-1}$	Cordeiro, Ortega e Nadarajah (2010)
Kumaraswamy Gama Generalizada (KGG)	$f(t) = \frac{\lambda\varphi\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\tau k-1} \exp\left[-\left(\frac{t}{\alpha}\right)^\tau\right]\left\{\gamma_1^{(2)}\left[k,\left(\frac{t}{\alpha}\right)^\tau\right]\right\}^{\lambda-1}$ $\times\left(1-\left\{\gamma_1^{(2)}\left[k,\left(\frac{t}{\alpha}\right)^\tau\right]\right\}^\lambda\right)^{\varphi-1}$	Pascoa, Ortega e Cordeiro (2011)
Kumaraswamy Burr XII (KBXII)	$f(t) = \alpha\gamma\beta\lambda\theta^{-\alpha}t^{(\alpha-1)}\left[1+\left(\frac{t}{\theta}\right)^\alpha\right]^{-(\gamma+1)}$ $\times\left\{1-\left[1+\left(\frac{t}{\theta}\right)^\alpha\right]^{-\gamma}\right\}^{(\lambda-1)}\left[1-\left\{1-\left[1+\left(\frac{t}{\theta}\right)^\alpha\right]^{-\gamma}\right\}^\lambda\right]^{(\beta-1)}$	Paranaíba et al. (2013)
Weibull Gen. Exp. de Potência (EPWG)	$f(t) = \alpha\beta\lambda\gamma t^{\gamma-1}\frac{(1+\lambda t^\gamma)^{\alpha-1}\exp\left[1-(1+\lambda t^\gamma)^\alpha\right]}{\left\{1-\exp\left[1-(1+\lambda t^\gamma)^\alpha\right]\right\}^{1-\beta}}$	Pena-Ramirez et al. (2018)

$^{(1)}B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, $^{(2)}\gamma_1(k,x) = \gamma(k,x)/\Gamma(k) = \left(\int_0^x w^{k-1}e^{-w}dw\right)/\Gamma(k)$.

Baseados na distribuição Kumaraswamy, apresentada por Jones (2009), Pascoa, Ortega e Cordeiro (2011) introduzem e discutem as propriedades da distribuição Kumaraswamy Gama Generalizada (KGG). A importância deste modelo probabilístico segundo os autores, está relacionada a sua capacidade de modelar funções de taxa de falha monótonas e não monótonas. Além disso, sua função de densidade de probabilidade possui a capacidade de se ajustar a dados distorcidos que não podem ser ajustados por outras distribuições já conhecidas, devido à flexibilidade de suas caudas, permitindo assim seu uso em diversas áreas (PASCOA; ORTEGA; CORDEIRO, 2011).

Utilizando um conjunto de dados com informação sobre 143 crianças nascidas no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, entre 1995 e 2001, expostas ao vírus HIV por transmissão vertical (transmissão do vírus pela mãe soropositiva, podendo ocorrer durante a gestação, trabalho de parto ou amamentação), Pascoa, Ortega e Cordeiro (2011) utilizam técnicas paramétricas, no contexto frequentista e bayesiano para mostrar a eficiência do modelo Kumaraswamy Gama Generalizada, realizando comparações entre ajustes com outras distribuições de probabilidade.

Pascoa, Ortega e Cordeiro (2011) ainda discutem alguns submodelos da distribuição Kumaraswamy Gama Generalizada, sendo eles a Gama Generalizada Exponenciada, a Kum-Weibull, a Kum-Gama, a Kum-Qui-Quadrado e a Kum-Qui-Quadrado em escala. Além da Kum-Gama Gerealizada, o submodelo Kum-Weibull $(k = 1)$, apresentada por Cordeiro, Ortega e Nadarajah (2010), também foi considerada para verificação do ajuste das distribuições aos dados.

Também com base na distribuição Kumaraswamy, Paranaíba et al. (2013) apresentam a distribuição Kumaraswamy Burr XII. Segundo os autores, a distribuição de cinco parâmetros possui grande flexibilidade, e foi discutida por meio de simulações para diferentes tamanhos de amostras e condissem sobre as possibilidades de aplicações mais amplas para análise de tempos de vida.

A distribuição Exponenciada Weibull Generalizada de Potência foi recentemente implementada por Pena-Ramirez et al. (2018). O modelo foi obtido a partir da aplicação da família exponenciada à distribuição Weibull Generalizada de Potência, sendo flexível para modelar tipos comuns de função de taxa de risco, sendo bem competitiva com as conhecidas distribuições Weibull. Uma aplicação em análise de sobrevivência pode ser vista em Aldahlan et al. (2020).

Exitem vários métodos estatísticos conhecidos para realizar a estimação dos parâmetros de uma distribuição, entre eles, o método dos mínimos quadrados é muito popular em regressões lineares, entretanto é inadequado para realizar as estimações na presença de censuras. Em análise de sobrevivência, é comum utilizar-se o método da máxima verossi-

milhança, que consiste em escolher soluções para os parâmetros que maximizem a função de verossimilhança. Para maiores detalhes sobre o método da máxima verossimilhança em um contexto geral, ver Ramachandran e Tsokos (2015).

No contexto de análise de sobrevivência, uma expressão geral para função de verossimilhança que engloba todos os mecanismos de censura, com exceção de constantes, é definida por:

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

onde δ_i é a variável indicadora de falha, tal que assume o valor 1 para dados não censurados e 0 para censuras (COLOSIMO; GIOLO, 2006).

Para verificar a adequação do modelo ajustado aos dados, serão utilizados os resíduos de Cox-Snell, Cox e Snell (1968), que resíduos são calculados por meio do oposto do logaritmo natural da função de sobrevivência estimada. Segundo Lawless (2003), para que o modelo seja adequado, os resíduos devem vir de uma população homogênea e seguir uma distribuição Exponencial padrão.

A escolha do melhor modelo será feito por meio da análise dos valores do logaritmo da função de verosimilhança e do Critério de Informação de Akaike pela quantidade de parâmetros (AIC), proposto por Sugiura (1978), além de comparações gráficas entre as estimativas de Kaplan-Meier e a função de sobrevivência ajustada.

Referente a parte computacional, utilizou-se o ambiente estatístico R (R Core Team, 2022), bem como os pacotes: *survival*, (THERNEAU, 2020); *flexsurv*, (JACKSON, 2016); *maxLik*, (HENNINGSEN; TOOMET, 2011).

5.3 Resultados e discussão

A busca de distribuições candidatas para o ajuste foi feita por meio de uma análise inicial dos gráficos empíricos TTT-plot e $S(t)$ estimada, nota-se que a função de risco descreve uma curva decrescente, indicando um comportamento decrescente para a mesma (Figura 17).

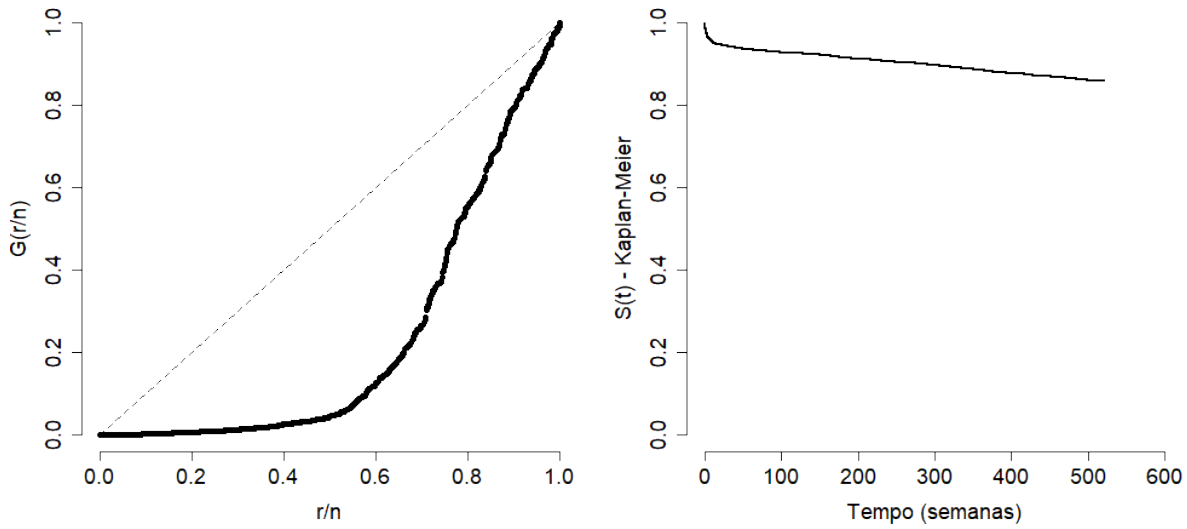

Figura 17 – Gráficos de tempo total em teste (TTT-plot) e $S(t)$ via Kaplan-Meier, respectivamente.

Observando as comparações entre cada função de sobrevivência e as estimativas de Kaplan-Meier (Figura 18), os valores do AIC e do logaritmo da função de verossimilhança, com base no menor e maior valor, respectivamente (Tabela 9), a distribuição KGG foi considerada a mais plausível para representar os dados abordados. Na Figura 19 é apresentada uma comparação entre a função de sobrevivência estimada por Kaplan-Meier e a função de sobrevivência via o modelo KGG.

Considerando o melhor modelo probabilístico selecionado (KGG), e visto que a distribuição possui suporte para comportar funções de risco monótonas (como a observada na Figura 17), os resíduos de Cox-Snell foram construídos para verificar a adequação dos modelos ajustados (Figura 20). Pode-se notar que de maneira geral, a distribuição apresentou um ajuste aceitável aos dados.

Realizando uma observação visual das estimativas de Kaplan-Meier (Figura 19), nota-se que houve um forte decrescimento nos tempos inicias, indicando uma grande quantidade de mortes relacionadas ao vírus HIV nesse período. Além disso, as curvas empíricas de sobrevivência, via Kaplan-Meier, não apresentam tendências de estabilização, dessa forma, não é possível indicar uma taxa de sobrevivência constante para os indivíduos infectados com HIV após um determinado tempo t, com base no estudo de aproximadamente 500 semanas de acompanhamento. A continuidade do acompanhamento dos pacientes poderia fornecer informações sobre a taxa de sobrevivência para uma maior quantidade de tempo.

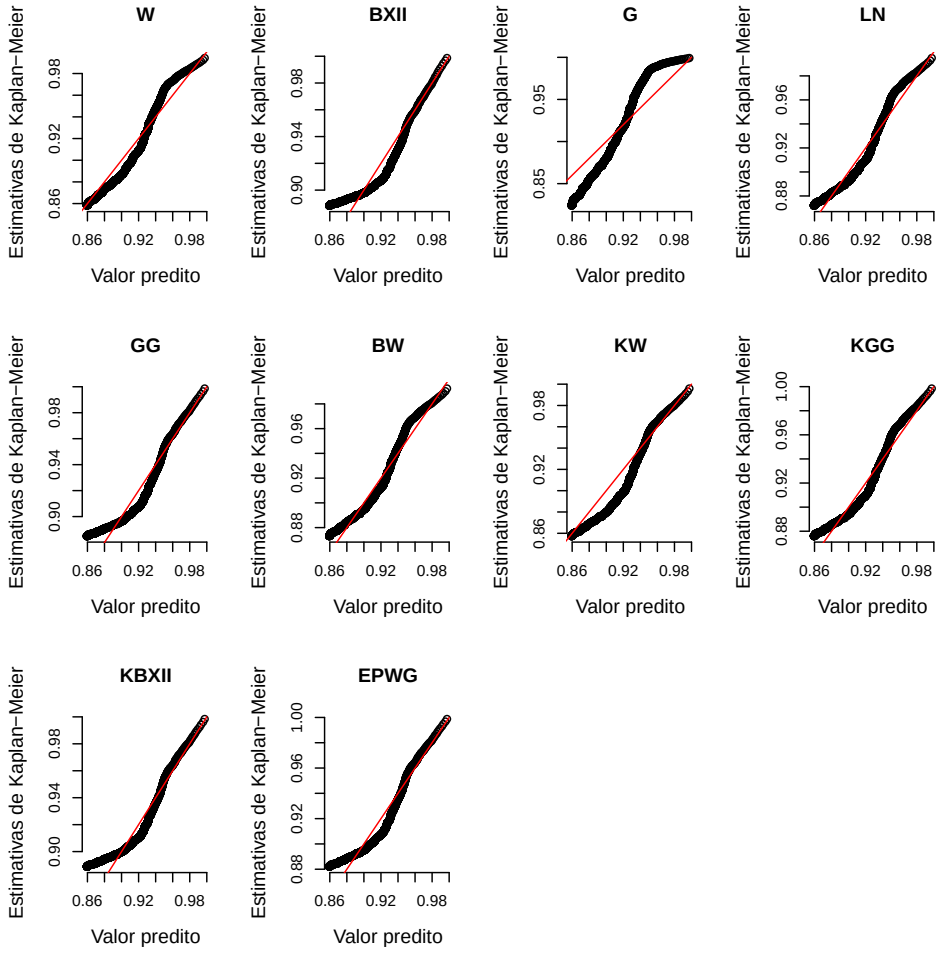

Figura 18 – Valores preditores gerados pelas funções de sobrevivência e estimativas de Kaplan-Meier.

Observa-se também que existem indicativos da presença de uma fração de sobreviventes de longa duração, de modo que a inclusão de um parâmetro de fração de cura, não abordada neste estudo, poderia melhorar a qualidade dos ajustes. Dito isto, temos motivação para investigar modelos probabilísticos envolvendo fração de cura, em novas pesquisas, para os dados aqui apresentados.

Tabela 9 – Estimativas para os parâmetros dos modelos.

Modelo	Estimativas (erro padrão)					$\log(L(\theta))$	AIC
W (τ,α)	0,4013	60.000,00				-6.428,61	12.858,23
	(0,0055)	(1,0490)					
BURRXII(θ, α, γ)	0,6898	1,4032	0,0127			-6.375,01	12.756,03
	(0,1214)	(0,1687)	(0,0019)				
G (γ,α)	0,6586	0,0001				-6.640,88	13.285,76
	(0,0301)	(0,0001)					
LN (μ,σ)	12,7552	5,7236				-6.398,59	12.801,18
	(0,2517)	(0,1611)					
GG (μ, σ, Q)	6,0508	2,2445	-4,625			-6.364,11	12.734,21
	(0,9218)	(0,4329)	(0,6550)				
BW (λ,α,β,γ)	0,4607	1,1882	0,0146	0,3215		-6.417,17	12.842,35
	(0,4390)	(0,1116)	(0,0048)	(0,0147)			
KMW (α, τ,k, λ, φ)	0,0046	0,2752	1,0000	10,1792	0,0069	-6.373,80	12.755,61
	(0,0001)	(0,0001)		(0,0008)	(0,0003)		
KGG ($\alpha, \tau, k, \lambda, \varphi$)	0,0002	0,2219	1,6830	33,4208	0,0063	**-6.347,18**	**12.704,35**
	(0,0003)	0,0001	(0,0002)	(0,0022)	(0,0004)		
KBXII(λ, β, α,γ,θ)	9,9657	0,0184	0,4025	2,9570	0,3952	-6.365,48	12.740,96
	(2,7978)	(0,0025)	(0,0386)	(0,2659)	(0,1411)		
EPWG ($\lambda, \alpha, \beta, \gamma$)	7,2247	0,0246	1,5060	1,2736		-6.355,48	12.718,96
	(1,2757)	(0,0058)	(0,1233)	(0,1870)			

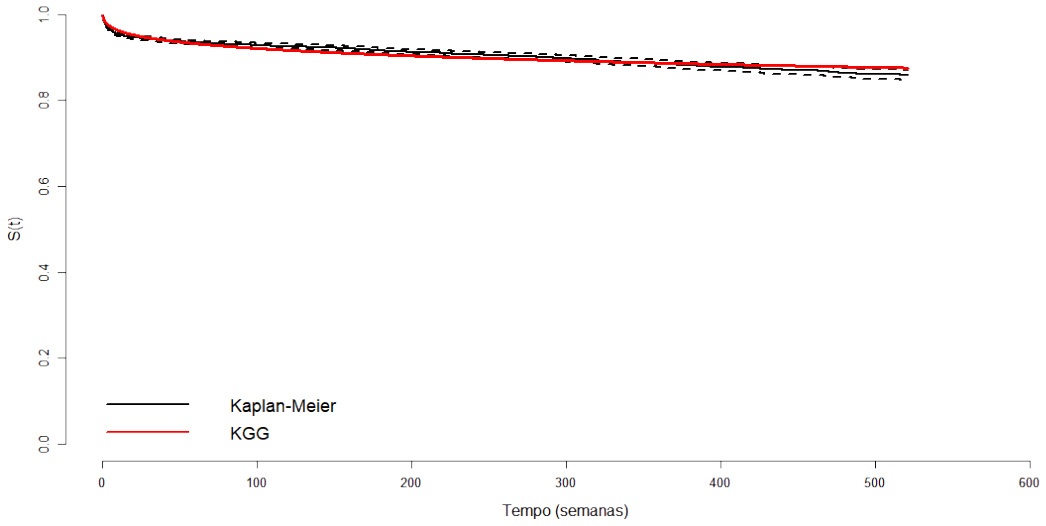

Figura 19 – Comparação entre as curvas de sobrevivência empírica (Kaplan-Meier) e estimada pelo melhor modelo (KGG).

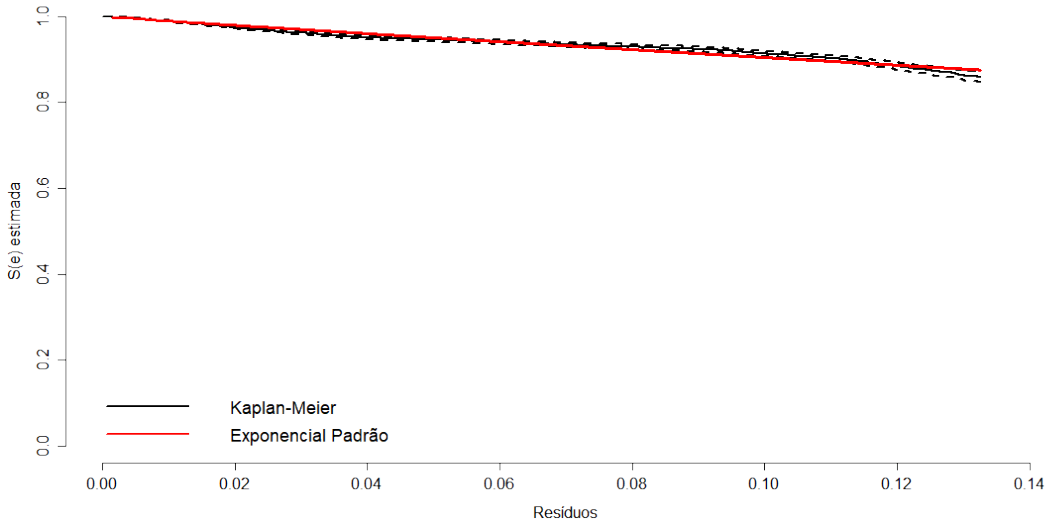

Figura 20 – Resíduos de Cox-Snell para os dados trabalhados.

5.4 Conclusões

O presente estudo sobre dados de tempos (em semanas) de sobrevida de pacientes com HIV/AIDS, evidenciou que os modelo probabilístico Kumaraswamy Gama Generalizada foi o mais adequado para a representação dos dados, sendo está uma das distribuições mais recentes entre as abordadas.

Capítulo 6

Modelo de fração de cura inflacionado de zero na identificação de fatores associados à sobrevida de pacientes com HIV/AIDS no Estado do Mato Grosso do Sul, Brasil

Resumo

Neste trabalho, objetivou-se ajustar algumas distribuições de probabilidade para descrever o tempo de sobrevida em pacientes com HIV/AIDS no Mato Grosso do Sul, Brasil, acompanhados entre 2009 e 2018. Nas distribuições abordadas, fez-se necessária a implementação de parâmetros para modelar os tempos nulos (inflação de zeros) e os de longa duração (fração de cura). Com base no Critério de Akaike, a distribuição Kumaraswamy Gama Generalizado foi a que melhor se ajustou aos dados, e então foi utilizada em um modelo de regressão que continha as variáveis explicativas: sexo, raça e escolaridade. Baseado nesse ajuste, o sexo feminino, a raça branca, e a escolaridade superior a oito anos foram associados a um maior tempo de sobrevida. Com base nessas interpretações, pode-se discutir a necessidade de políticas de prevenção e diagnóstico precoce de HIV focadas para grupos específicos associados a menor sobrevida.
Palavras-chave: análise de sobrevivência, dados censurados, modelos inflacionados de zeros, modelos de longa duração.

6.1 Introdução

A síndrome da imunodeficiência humana (AIDS) é uma doença crônica, sendo o terceiro estágio da infecção pelo vírus da imunodeficiência humana (HIV), que se faz presente no Brasil desde a década de 80. Segundo Brito, Castilho e Szwarcwald (2001), inicialmente a doença era presente em grandes centros urbanos e afetava principalmente homossexuais masculinos e indivíduos hemofílicos.

De acordo com Melo, Almeida e Donalisio (2021), o perfil epidemiológico da infecção pelo vírus HIV sofreu alterações ao longo das décadas, deixando de ser majoritariamente concentrada em grandes centros urbanos, se espalhando dessa forma para municípios de interior. Além da disseminação para municípios de pequeno e médio porte, o envolvimento de populações mais socialmente vulneráveis e o crescente número de mulheres infectadas pelo vírus, são marcas das modificações sofridas ao longo do tempo no país (CANDIDO et al., 2021).

Uma análise exploratória de dados espaciais sobre a incidência de AIDS para cada Estado brasileiro entre os anos de 1992 e 2017 é apresentada por Ribeiro, Fonseca e Pereira (2019), permitindo observar a evolução da taxa de incidência da doença ao longo dos anos. De acordo com os autores, mesmo havendo uma epidemia distribuída mais equilibradamente pelo país em tempos atuais, ainda há incidência considerável nos estados do Rio de Janeiro, Santa Catarina e Rio Grande do Sul, e a região Norte passou a fazer parte das áreas de concentração da epidemia.

Segundo dados do Boletim Epidemiológico de HIV/AIDS de 2020 (BRASIL, 2020), no Brasil, entre 2007 e junho de 2020, foram notificados no Sistema de Informação de Agravos de Notificação (SINAM) um total de 342.459 casos de infecção pelo vírus HIV, dos quais 44,4%, 20%, 19%, 9% e 7,6% pertencem, respectivamente, às regiões Sudeste, Sul, Nordeste, Norte e Centro-Oeste.

A análise de sobrevivência consiste na área da Estatística responsável pelo estudo do tempo até a ocorrência de um determinado evento de interesse. Nesse contexto, vários trabalhos podem ser encontrados no literatura analisando dados de HIV/AIDS como em Colosimo e Vieira (1996), Melo, Donalisio e Cordeiro (2017), Medeiros et al. (2017), Zaslavsky, Goulart e Ziegelmann (2019), Müller e Borges (2020), Melo et al. (2021), além de diversos outros. Entretanto, a maior parte dos trabalhos utilizando dados neste contexto, partem de uma abordagem semi-paramétrica por meio do modelo de regressão de Cox (1972), sem a necessidade de fazer a suposição de uma distribuição de probabilidade para a variável resposta.

Embora sejam mais escassos, alguns trabalhos de análise de sobrevida no contexto

paramétrico envolvendo dados de HIV/AIDS podem ser vistas em Nakhaee e Law (2011), Pascoa, Ortega e Cordeiro (2011), Million (2018) e Bueno e Rossi (2021).

Ainda no contexto de análise de sobrevida, pode existir dentro da população em estudo uma proporção de indivíduos que não são suscetíveis a apresentar o evento de interesse, de modo que os modelos probabilísticos comuns não são capazes de estimar essa proporção de indivíduos. Desse modo, se faz necessário o uso dos modelos com fração de cura, em que esses geralmente são imunes ou curados do desfecho de interesse. Martinez et al. (2013), Peres, Santos e Oliveira (2020), Moraes, Previdelli e Silva (2021), entre outros, são exemplos de trabalhos com uso dessa abordagem.

Uma outra situação, não tão comum na prática, é a ocorrência de observações de tempos nulos (ou zeros) e, na alta frequência, são denominados dados inflacionados de zeros, sendo necessário então, utilizar as distribuições que acomodem tal situação, como observado em Hashimoto (2013), Júnior, Moreira e Louzada (2017), Calsavara et al. (2019), entre outros.

Neste contexto, objetivou-se neste trabalho realizar modelagens paramétricas, considerando diferentes distribuições para a resposta do tempo de sobrevida dos pacientes com HIV/AIDS, na presença de fração de cura e inflação de zero para o ajuste aos dados.

6.2 Materiais e Métodos

Os dados analisados neste trabalho são provenientes de um estudo epidemiológico apresentado por Werle (2021). O banco de dados contém originalmente informações referente a 9.021 indivíduos diagnosticados com HIV/AIDS no Estado do Mato Grosso do Sul, Brasil, observados entre 2009 e 2018. Foram removidas 4.263 observações que continham valores perdidos, dessa forma, o conjunto de dados utilizados conteve observações de 4.758 pacientes. O estudo originário dos dados atendeu as normas nacionais e internacionais de ética em pesquisa com seres humanos, e foi aprovada com parecer n° 3.789.678. No presente estudo, dispensa-se a apreciação pelo Comitê de Ética em Pesquisa (CEP), do Conselho Nacional de Saúde (CNS) e da Comissão Nacional de Ética em Pesquisa (CONEP) baseado na Resolução 466/2012 CEP-CONEP.

Na Tabela 10, é apresentado um resumo descritivo das variáveis explicativas consideradas, bem como frequências absolutas e relativas, além da proporção de censuras.

A análise de sobrevivência é um ramo da Estatística utilizada quando o tempo é o objeto de interesse, podendo este ser interpretado como o tempo até a ocorrência de um evento de interesse ou o risco de ocorrência de um evento de interesse por uma determinada unidade de tempo (CARVALHO et al., 2011). As funções de sobrevivência, de risco e de

risco acumulado são algumas funções características nesta área de estudo, de modo, uma discussão sobre estas pode ser observada em Lee e Wang (2003).

Tabela 10 – Resumo descritivo para as variáveis explicativas.

Variáveis	Classes	Freq.	%	Censuras (%)	Zeros (%)	Média	EPM
Sexo	Masculino	3.004	63,14	89,04	1,70	460,25	3,19
	Feminino	1.754	36,86	90,25	1,48	471,27	3,66
Raça/Cor	Branca	2.094	44,01	90,97	1,24	473,78	3,32
	Não branca	2.664	55,99	88,32	0,98	456,80	3,44
Escolaridade (em anos)	> 8	2.008	42,20	93,38	1,10	483,62	3,21
	≤ 8	2.750	57,80	86,65	2,00	452,39	3,36

Freq.: Frequência; Média: média restrita proveniente das estimativas de Kaplan-Meier; EPM: erro padrão da média restrita.

Colosimo e Giolo (2006), sugerem inicialmente realizar uma análise descritiva dos dados por meio de uma estimação da função de sobrevivência, e a partir dela, estimar as demais estatísticas de interesse. Existem várias técnicas para estimar a função de sobrevivência na presença de dados censurados, entretanto, a mais usual é por meio do estimador de Kaplan-Meier (ou estimador produto-limite). Proposto por Kaplan e Meier (1958), o estimador é definido como: $\hat{S}(t) = \prod_{j:t_j<t}(1 - d_j/n_j)$, de modo que d_j é o número de falhas e n_j o número de indivíduos em risco em cada tempo t_j.

Os modelos probabilísticos em análise de sobrevivência partem da suposição de que a variável resposta segue uma distribuição de probabilidade. O gráfico de tempo total em teste (TTT-plot), proposto originalmente por Barlow e Campo (1975) e generalizado por Aarset (1985), permite identificar a forma da função de risco, possibilitando considerar modelos mais adequados aos dados em questão.

Na Tabela 11 são apresentadas as distribuições consideradas neste trabalho. Sabe-se que, originalmente, estas não possuem suporte para observações de tempo igual a 0 e portanto, análises estatísticas em dados caracterizados pela presença de excessos de zeros são realizadas por meio dos modelos inflacionados, como apresentado em Hashimoto (2013).

Considerando T uma varável aleatória que representa o tempo até a ocorrência do evento de interesse, com excessos ou inflação de zeros, segundo Calsavara et al. (2019), a função de sobrevivência é dada da seguinte forma:

$$S(t) = (1 - p_0)S_0(t),$$

onde $0 < p_0 < 1$ é o parâmetro que modela a proporção de tempos zero e $S_0(t)$ é a função de sobrevivência própria.

Tabela 11 – Distribuições de probabilidade consideradas para os ajustes dos dados.

Modelo	Função densidade de probabilidade	Autor
Weibull (W)	$f(t) = \frac{\tau}{\alpha^{\tau}} t^{\tau-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}$	Weibull (1939)
Burr XII (BXII)	$f(t) = \frac{\gamma\alpha t^{\alpha-1}}{\theta^{\alpha}} \left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}$	Burr (1942)
Log- Normal (LN)	$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^{2}\right\}$	Aitchison e Brown (1957)
Beta-Weibull (BW)	$f(t) = \frac{\gamma t^{\gamma-1}}{B^{(1)}(\alpha,\beta)\lambda^{\gamma}}$ $\exp\left[-\beta\left(\frac{t}{\lambda}\right)^{\gamma}\right]\left\{1-\exp\left[-\left(\frac{t}{\lambda}\right)^{\gamma}\right]\right\}^{\alpha-1}$	Lambert et al. (2007)
Kumaraswamy Weibull (KW)	$f(t) = \frac{\lambda\varphi\tau}{\alpha}\left(\frac{t}{\alpha}\right)^{\tau-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1^{(2)}\left[1,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}$ $\times\left(1-\left\{\gamma_1^{(2)}\left[1,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi-1}$	Cordeiro, Ortega e Nadarajah (2010)
Kumaraswamy Gama Generalizada (KGG)	$f(t) = \frac{\lambda\varphi\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\tau k-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1^{(2)}\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}$ $\times\left(1-\left\{\gamma_1^{(2)}\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi-1}$	Pascoa, Ortega e Cordeiro (2011)
Kumaraswamy Burr XII (KBXII)	$f(t) = \alpha\gamma\beta\lambda\theta^{-\alpha}t^{(\alpha-1)}\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}$ $\times\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{(\lambda-1)}\left[1-\left\{1-\left[1+\left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{\lambda}\right]^{(\beta-1)}$	Paranaíba et al. (2013)
Weibull Gen. Exp. de Potência (EPWG)	$f(t) = \alpha\beta\lambda\gamma t^{\gamma-1}\frac{(1+\lambda t^{\gamma})^{\alpha-1}\exp\left[1-(1+\lambda t^{\gamma})^{\alpha}\right]}{\left\{1-\exp\left[1-(1+\lambda t^{\gamma})^{\alpha}\right]\right\}^{1-\beta}}$	Pena-Ramirez et al. (2018)

$^{(1)}B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, $^{(2)}\gamma_1(k,x) = \gamma(k,x)/\Gamma(k) = \left(\int_0^x w^{k-1}e^{-w}dw\right)/\Gamma(k)$.

Desse modo a função de densidade de probabilidade $f(t)$ é dada da seguinte forma:

$$f(t) = \begin{cases} p_0, & \text{se} \quad t = 0 \\ (1 - p_0) f_0(t), & \text{se} \quad t \neq 0. \end{cases}$$

Os modelos de sobrevivência com fração de cura podem ser utilizados para modelar dados de sobrevida onde observações possuem um longo tempo de sobrevivência, de modo que, isso acontece quando parte dos indivíduos não recebem o evento de interesse, ou este não é observado durante o tempo de estudo. Nessa abordagem, parte-se da ideia de que a população em estudo é formada por indivíduos que são suscetíveis a experimentar o evento de interesse, bem como indivíduos que não são suscetíveis (MALLER; ZHOU, 1996).

Nessa abordagem a função de sobrevivência populacional é dada pela modelo de mistura padrão:

$$S(t) = p_1 + (1 - p_1) S_0(t),$$

em que $0 < p_1 < 1$ é a proporção de indivíduos que não estão suscetíveis a ocorrência do evento de interesse (fração de cura) e $S_0(t)$ é a função de sobrevivência própria.

De acordo com Júnior, Moreira e Louzada (2017), considerando modelos que abordam tanto a inflação de zeros, quanto a fração de cura, suas funções de sobrevivência e densidade de probabilidade são definidas, respectivamente por:

$$S(t) = p_1 + (1 - p_0 - p_1) S_0(t)$$

e

$$f(t) \begin{cases} p_0, & \text{se} \quad t = 0 \\ (1 - p_0 - p_1) f_0(t), & \text{se} \quad t \neq 0. \end{cases}$$

As estimativas dos parâmetros, considerando o método da máxima verossimilhança, no contexto de análise de sobrevivência, com presença de inflação de zeros e fração de cura, são encontradas maximizando a função de verossimilhança dada por:

$$L(p_0,p_1,\mathrm{D},\boldsymbol{\phi}) = \prod_{i:t_i=0} p_0 \prod_{i:t_i>0} \left\{ \left[(1 - p_0 - p_1) f_0(t_i|\boldsymbol{\phi})\right]^{\delta_i} \left[p_1 + (1 - p_1 - p_0) S_0(t_i|\boldsymbol{\phi})\right]^{1-\delta_1} \right\},$$

em que $\boldsymbol{\phi}$ é o vetor de parâmetros, $\mathrm{D} = \{t_i,\delta_i\}$, tal que $\delta_i = 1$ indica uma observação que apresentou o evento de interesse e $\delta_i = 0$ indica censura.

Após a estimação dos parâmetros dos modelos, feitas com o ambiente estatístico R (R Core Team, 2022), a seleção do melhor foi feita com base no critério de Akaike (AIC).

Além disso, utilizou-se os resíduos de Cox e Snell (1968) para a verificação da qualidade do ajuste, em que, segundo Lawless (2003), para que modelo apresentar um ajuste adequado, os resíduos de Cox-Snell devem seguir uma distribuição Exponencial padrão.

6.3 Resultados e Discussão

Observa-se a presença da fração de cura nos dados por meio das estimativas de Kaplan-Meier (Figura 21). O gráfico TTT (Figura 21) indica que a função de risco apresenta o comportamento decrescente.

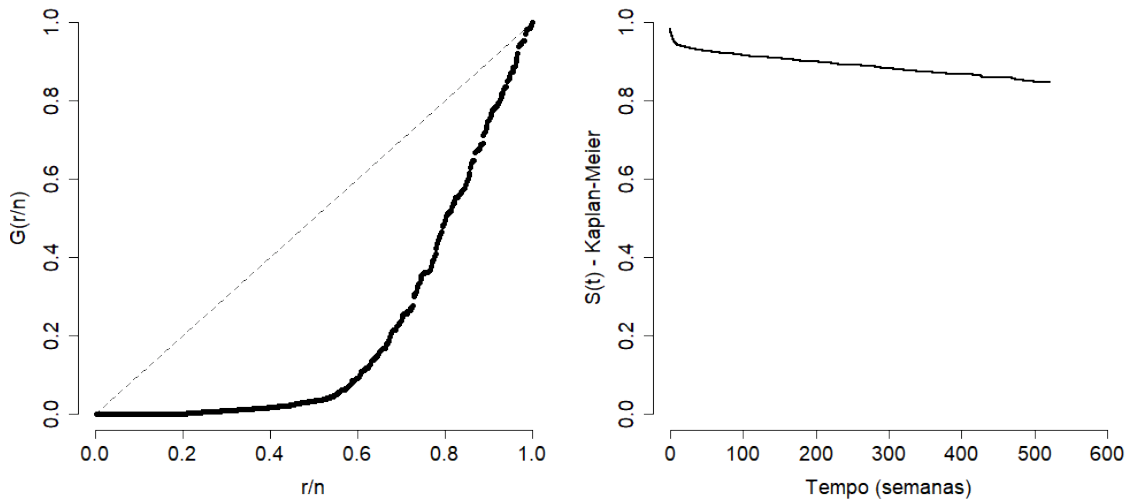

Figura 21 – TTT-plot e função de sobrevivência estimada via Kaplan-Meier, respectivamente.

As estimativas dos parâmetros para as distribuições consideradas (com a adição dos parâmetros de inflação de zeros e fração de cura, indicados por ZC antes da sigla da distribuição), bem como o AIC, são apresentados na Tabela 12, indicando que o modelo ZCKGG foi o mais adequado para a representação dos dados, visto que apresentou o menor valor para o AIC (7.596,49). Além disso, a distribuição KGG possui suporte para função de risco decrescente, conforme foi observado no comportamento do TTT-plot (Figura 21).

Diversos trabalhos que utilizam dados inflacionados de zeros ou com fração de cura podem ser encontrados na literatura, como em Calsavara et al. (2019), onde são introduzidos os modelos defectivos Gompertz e Gaussiana inversa com inflação de zeros e Moraes, Previdelli e Silva (2021) apresentam uma modelagem para a fração de cura considerando um modelo Weibull, sob a ótica Bayesiana, para dados de câncer de mama observados entre 2004 e 2016 no Paraná, Brasil.

Tabela 12 – Estimativas dos parâmetros das distribuições consideradas.

Distribuição (Parâmetros)	Estimativas (Erro Padrão)							AIC
ZCW(p_0,p_1,τ,α)	0,015	0,747	0,435	989,606				7.641,44
	(0,001)	(0,020)	(0,017)	(1,438)				
ZCBXII$(p_0,p_1,\theta,\alpha,\gamma)$	0,013	0,567	1,202	1,466	0,028			7.650,73
	(0,001)	(0,039)	(0,312)	(0,153)	(0,003)			
ZCLN$(p_0,p1,\mu,\sigma)$	0,015	0,674	7,343	3,976				7.617,01
	(0,002)	(0,087)	(1,168)	(0,403)				
ZCBW$(p_0,p_1,\lambda,\alpha,\beta,\gamma)$	0,040	0,596	0,990	0,863	0,035	0,420		7.769,48
	0,004	0,087	0,142	0,327	0,013	0,103		
ZCKW$(p_0,p_1,\tau,\alpha,\lambda,\varphi)$	0,015	0,725	0,411	0,081	1,524	0,018		7.631,06
	(0,001)	(0,052)	(<0,001)	(<0,001)	(<0,001)	(0,005)		
ZCKGG$(p_0,p_1,\tau,\alpha,\lambda,\varphi,k)$	0,015	0,709	0,303	0,024	16,041	0,030	0,696	**7.596,49**
	(0,002)	(0,054)	(0,004)	(0,002)	(0,005)	(0,007)	(0,003)	
ZCKBXII$(p_0,p_1,\lambda,\beta,\alpha,\gamma,\theta)$	0,014	0,583	1,351	0,081	1,124	0,394	0,285	7.684,43
	(0,002)	(0,035)	(0,160)	(0,009)	(0,090)	(0,040)	(0,056)	
ZCEPWG$(p_0,p_1,\alpha,\beta,\lambda,\gamma)$	0,015	0,646	0,104	1,291	0,0820	0,720		7.617,64
	(0,002)	(0,090)	(0,040)	(0,186)	(0,402)	(0,138)		

Em especial, para dados de pacientes infectados com HIV, um exemplo de trabalho utilizando fração de cura pode ser visto em Freitas et al. (2021), onde é apresentada uma aplicação de um modelo de sobrevivência paramétrico Weibull Discreto Exponenciado com fração de cura, por meio de abordagens frequentista e Bayesiana.

Algumas aplicações que utilizam ambas as abordagens podem ser vistas em Júnior, Moreira e Louzada (2017) para dados de risco de crédito, em Souza et al. (2021) para dados de câncer invasivo do colo de útero, entre outros

A partir da escolha da melhor distribuição para os dados, realizou-se uma comparação entre a função de sobrevivência ajustada e empírica (via Kaplan-Meier), e também entre os resíduos de Cox-Snell e a função de risco acumulada. A observação destas duas representações gráficas indicam que o modelo ajustado está (satisfatoriamente) adequado para descrever os dados abordados (Figura 22).

Para a construção do modelo de regressão ZCKGG, propusemos adicionar o vetor de parâmetros ($\boldsymbol{\beta}$) associados às variáveis explicativas no parâmetro φ, um dos quatro parâmetros de forma (τ,λ,φ,k) da distribuição. Desse modo, o modelo de regressão é

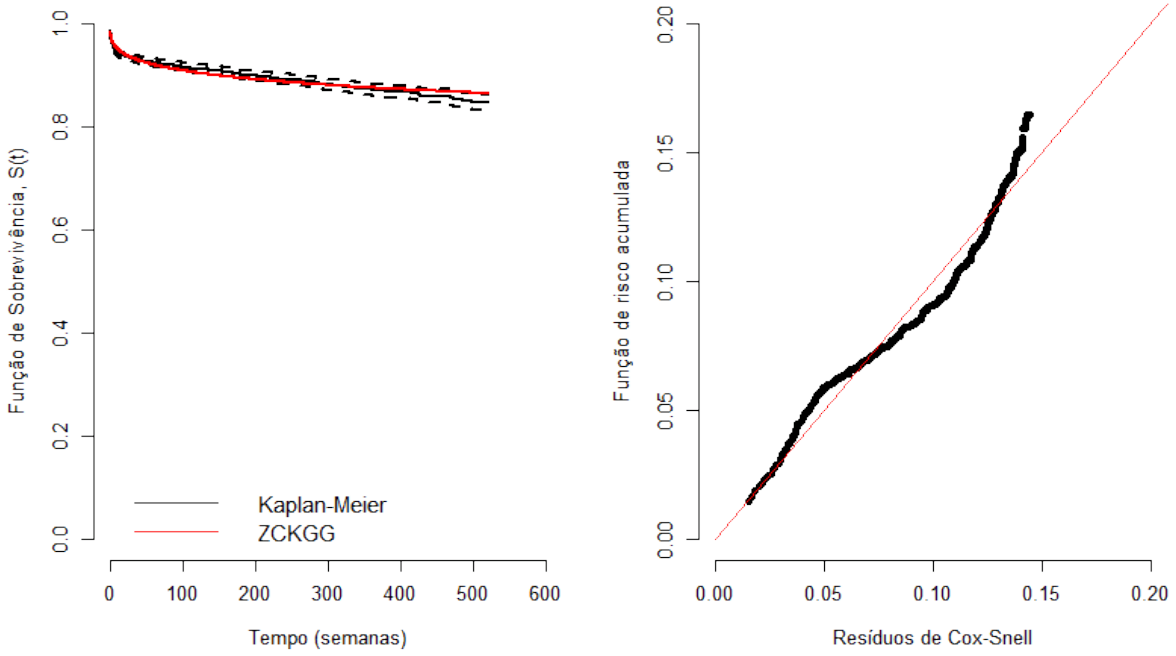

Figura 22 – Comparação entre as curvas de sobrevivência empírica e ajustada, e comparação entre os resíduos de Cox-Snell e função de risco acumulado, respectivamente.

dado por:

$$f(t) = (1 - p_0 - p_1)\,\frac{\lambda\varphi(\boldsymbol{x}'\boldsymbol{\beta})\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k\tau-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}$$
$$\times\left(1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi(\boldsymbol{x}'\boldsymbol{\beta})-1},$$

em que $\varphi(\boldsymbol{x}'\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$, que x_p representa a *p*-ésima variável explicativa.

Para construção do modelo de regressão, optou-se por remover as variáveis relacionadas ao modo de infecção pelo vírus que fizeram parte do modelo indicado por Werle (2021), devido a forte com relação com outras variáveis ou que não apresentaram melhora no modelo de regressão paramétrico. As estimativas dos parâmetros do modelo de regressão ZCKGG são apresentadas na Tabela 13.

Visando verificar a adequação das variáveis explicativas, separadamente, ao modelo ZCKGG, comparações entre a curva de sobrevivência empírica (via Kaplan-Meier) e a estimada via modelos de regressão foram construídas (Figura 23).

Para verificação da importância do parâmetro de fração de cura, um novo modelo de regressão KGG (Kumaraswamy Gama Generalizada) foi construído, incorporando apenas

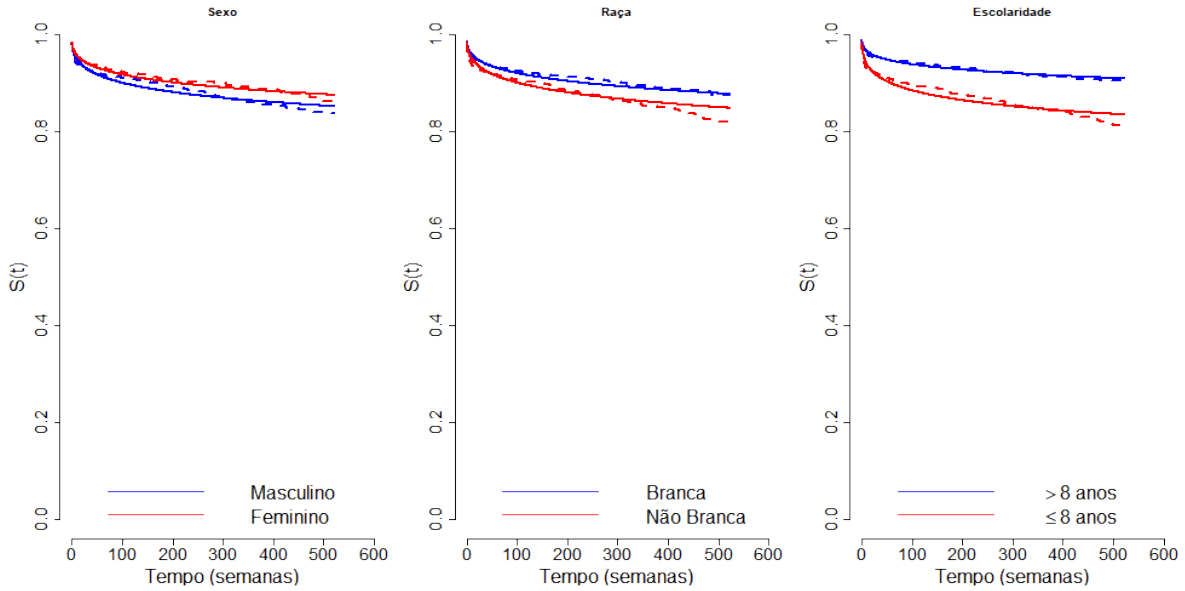

Figura 23 – Comparações entre as curvas de Kaplan-Meier e a função de sobrevivência estimada para cada modelo regressão, por variável explicativa.

a inflação de zeros (ZKGG), o qual as estimativas também são apresentadas na Tabela 13. O valor do AIC é favorável, por ser o menor valor, ao modelo com a fração de cura (AIC = 7.582,45) em comparação ao segundo modelo (AIC = 7.887,34). Através da substituição das estimativas nos parâmetros nos modelos de regressão, em ambos, o sexo feminino foi associado à maior sobrevida, já a raça não branca e a escolaridade igual ou inferior a 8 anos foram associadas a diminuição da sobrevida.

Os efeitos das variáveis explicativas sexo, raça e escolaridade coincidem com os apresentados pelo modelo de regressão de Cox apresentado por Müller e Borges (2020), para dados de HIV/AIDS referentes à Região de Campos Gerais (PR), Brasil.

Considerando os valores do AIC podemos afirmar que há evidências que o modelo com fração de cura é mais adequado para modelar esses dados, proporcionado dessa forma uma estimativa da proporção de indivíduos que não irão vir a experimentar o evento de interesse, interpretado pelo parâmetro p_1. Dessa forma, segundo os modelos ZCKGG, com e sem variáveis explicativas, a proporção de indivíduos que não irão a óbito por HIV/AIDS no Estado do Mato Grosso do Sul são de, respectivamente, 63,16% e 70,90%. Acredita-se que essa alta proporção deva-se aos avanços na medicina desenvolvidos para o tratamento de HIV/AIDS nas últimas décadas. Além disso, a utilização do parâmetro de fração de cura para os dados abordados, como esperado, vai de encontro a alguns trabalhos na literatura, como em Varshney et al. (2018) e Freitas et al. (2021).

Escolhido o melhor modelo de regressão (ZCKGG), foram construídos os resíduos de

Tabela 13 – Estimativas dos parâmetros (e erro padrão) dos modelos de regressão.

Parâmetro	Classe	Modelo	
		ZCKGG	ZKGG
p_0	-	0,0151 (0,0018)	0,0126 (0,0015)
p_1	-	0,6316 (0,1431)	-
τ	-	0,1750 (0,5138)	0,1657 (0,0152)
α	-	0,8571 (1,2255)	3,0047 (0,7987)
λ	-	5,0539 (2,6459)	1,3489 (0,1364)
k	-	1,0096 (1,9157)	1,3963 (0,1529)
β_0	-	-1,7556 (9,1419)	-2,5603 (0,287)
β_1: Sexo	Masculino	-	-
	Feminino	-0,3097 (0,1301)	-1,8194 (0,1584)
β_2: Raça	Branca	-	-
	Não Branca	0,2442 (0,1217)	0,2352 (0,0978)
β_3: Escolaridade (anos)	> 8	-	-
	≤ 8	0,6916 (0,1404)	0,5748 (0,1031)
AIC		7.582,74	7.887,34

Cox-Snell para verificação da qualidade do ajuste do modelo (Figura 24). Embora alguns pontos tenham apresentado um comportamento diferente do esperado, a interpretação gráfica dos resíduos indica um modelo aceitável devido a complexidade da modelagem.

O teste de Wald, construído via o pacote *maxLik* (HENNINGSEN; TOOMET, 2011), para verificar a significância das variáveis explicativas presentes no modelo final, ao nível de 5% de significância, cuja a hipótese nula é de que a estimativa do parâmetro é igual a 0, indica que apenas o intercepto (β_0) não foi marginalmente significativo entre as variáveis explicativas do modelo de regressão ZCKGG (Tabela 14). Nesse modelo α é o parâmetro de escala e τ, λ, k e $\varphi(\boldsymbol{x'\beta}) = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}$ são responsáveis por modelar a forma da função de sobrevivência.

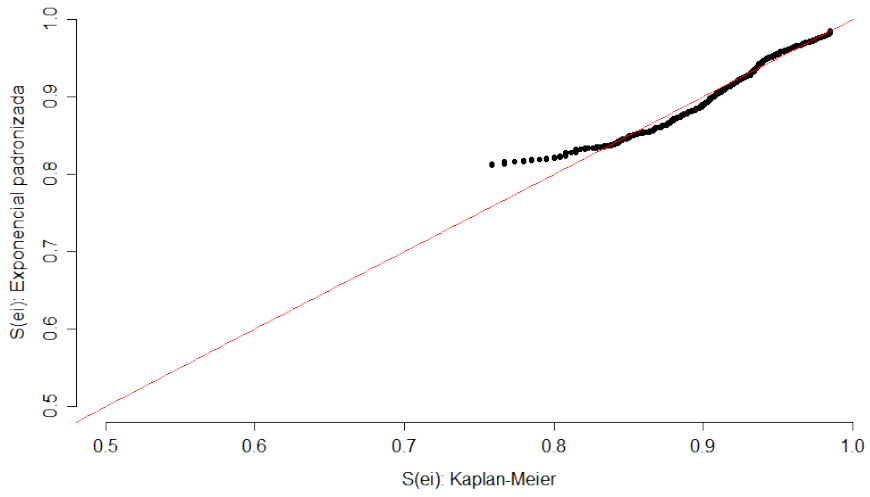

Figura 24 – Análise de resíduos de Cox-Snell para o modelo de regressão ZCKGG.

Tabela 14 – Teste de Wald para a significância das variáveis explicativas do modelo de regressão ZCKGG.

Parâmetro	Estatística do teste	Valor p
β_0	-0,192	0,8477
β_1	-2,381	0,0173
β_2	2,007	0,0448
β_3	4,927	8,36 $\times 10^{-7}$

O parâmetro da inflação de zeros $(p_0 = 0{,}0151)$ fornece a probabilidade (aproximadamente 1,51%) de que indivíduos apresentem o evento de interesse 0 semanas após o diagnóstico.

O parâmetro da fração de cura $(p_1 = 0{,}6316)$ indica que 63,86% da população em estudo não experimentarão o evento de interesse. Possivelmente, essa alta proporção deve-se ao fato de que os indivíduos mantiveram o tratamento corretamente após o diagnóstico soropositivo para o HIV.

6.4 Conclusões

Dentre os modelos probabilísticos testados, a distribuição Kumaraswamy Gama Generalizada com fração de cura e inflação de zeros foi a mais adequada para representar os dados. Além disso, as comparações entre os modelos de regressão com e sem o parâmetro de fração de cura, mostram que a presença da fração de cura resultou em um ajuste melhor, corroborando com a curva de sobrevivência estimada via Kaplan-Meier.

A identificação de grupos com menor tempo de sobrevida pode auxiliar em futuras políticas de prevenção e diagnóstico precoce para esses grupos, visando impedir possíveis mortes ou casos de evolução da infecção viral.

Capítulo 7

Considerações finais e passos futuros

Esse trabalho apresentou algumas distribuições amplamente conhecidas na literatura, além de algumas recentes, bem suas funções no contexto de análise de sobrevivência, além de uma breve introdução sobre as abordagens Bayesianas no Capitulo 2.

No Capítulo 3 foi apresentada uma aplicação ao conjunto de dados de HIV/AIDS para o Estado do Mato Grosso do Sul, Brasil, observados entre 2009 e 2018, utilizando técnicas paramétricas para a verificação do melhor modelo, dentre os apresentados, e após isso, foram construídos e comparados os modelos Bayesianos para o melhor modelo probabilístico (Log-Normal) e o modelo padrão de análise de sobrevivência (Weibull). Verificou-se pela análise de resíduos que ambos não apresentaram um ajuste adequado aos dados, discutindo-se sobre a necessidade da inclusão das variáveis explicativas ou da utilização de distribuições mais flexíveis.

No Capítulo 4 foi verificado novamente qual o melhor modelo para o ajuste aos dados, após todas as observações perdidas das variáveis explicativas serem removidas, e construídos os modelos de regressão, frequentista e Bayesiano, para o melhor modelo (Log-Normal). Verificou-se que ambos modelos tiveram ajustes similares e as análises de resíduos apresentaram resultados satisfatórios. Por meio das estimativas dos parâmetros, foi realizada uma comparação dos tempos de sobrevida medianos, e desse modo, permitindo discutir-se a necessidade de políticas de prevenção e diagnóstico precoce focadas em grupos com um menor tempo mediano de sobrevida.

Ao comparar algumas distribuições de probabilidade, tanto conhecidas na literatura quanto descobertas mais recentemente, verificou-se que a distribuição Kumaraswamy Gama Generalizada (KGG) teve o ajuste mais adequado para os dados em questão (Capítulo 5). Em uma outra análise, forma inseridos parâmetros para modelar a inflação de zeros e fra-

ção de cura nos dados, e novamente, a distribuição Kumaraswamy Gama Generalizada (ZCKGG) foi selecionada. Um modelo de regressão foi construído para esta distribuição, onde foi observado que o sexo feminino foi associado ao aumento da sobrevida, e que raça não branca e escolaridade igual ou inferior a oito anos foram associados à diminuição do tempo de sobrevida (Capítulo 6).

Pesquisas futuras podem proporcionar ajustes de modelos de regressão ZCKGG com as variáveis explicativas em outros parâmetros de forma, além de construir modelos Bayesianos para verificar a concordância com os modelos frequentistas.

Referências

AARSET, M. V. The null distribution for a test of constant versus"bathtub"failure rate. *Scandinavian Journal of Statistics*, JSTOR, p. 55–61, 1985.

ACHCAR, J. A.; BARROS, E. A. C.; MAZUCHELI, J. Cure fraction models using mixture and non-mixture models. *Tatra Mt. Math. Publ*, v. 51, p. 1–9, 2012.

ACHCAR, J. A.; BARROS, E. A. C.; SOUZA, R. M.; MARTINEZ, E. Z. *Uma introdução aos métodos bayesianos aplicados à análise de dados*. Timburi: CIA DO EBOOK, 2019.

AITCHISON, J.; BROWN, J. A. The Lognormal distribution with special reference to its uses in economics. Cambridge Univ. Press, 1957.

ALDAHLAN, M. A.; JAMAL, F.; CHESNEAU, C.; ELBATAL, I.; ELGARHY, M. Exponentiated power generalized Weibull power series family of distributions: Properties, estimation and applications. *PloS One*, Public Library of Science San Francisco, CA USA, v. 15, n. 3, p. e0230004, 2020.

ASSUNÇÃO, G. O. *Comparação entre algoritmos MCMC para Inferência Bayesiana em modelos dicotômicos da TRI*. Dissertação (Mestrado em Estatística) — Universidade Federal de Minas Gerais, 2019.

BARLOW, R. E.; CAMPO, R. A. *Total Time on Test Processes and Applications to Failure Data Analysis.* [S.l.], 1975.

BASTOS, J.; ROCHA, C. Análise de sobrevivência: conceitos básicos. *Arquivos de Medicina*, ArquiMed-Departamento de Edições Científicas da AEFMUP, v. 20, n. 5-6, p. 185–187, 2006.

BOX, G. E.; TIAO, G. C. *Bayesian inference in statistical analysis*. [S.l.]: John Wiley, 1973.

BRASIL. *Boletim epidemiológico: HIV/AIDS*. Brasília: Ministério da Saúde, 2020.

BRITO, A. M. d.; CASTILHO, E. A. d.; SZWARCWALD, C. L. AIDS e infecção pelo HIV no Brasil: uma epidemia multifacetada. *Revista da sociedade brasileira de medicina tropical*, SciELO Brasil, v. 34, n. 2, p. 207–217, 2001.

BROWN, G. W.; FLOOD, M. M. Tumbler mortality. *Journal of the American Statistical Association*, Taylor & Francis, v. 42, n. 240, p. 562–574, 1947.

BRUNELLO, G. H. V.; NAKANO, E. Y. Inferência bayesiana no modelo Weibull discreto em dados com presença de censura. *TEMA (São Carlos)*, SciELO Brasil, v. 16, n. 2, p. 97–110, 2015.

BUENO, M. V.; ROSSI, R. M. Análise de sobrevivência para dados de HIV/AIDS: uma abordagem Bayesiana. *Brazilian Journal of Development*, Curitiba, v. 7, n. 6, p. 55797–55805, 2021.

BURR, I. W. Cumulative frequency functions. *The Annals of Mathematical Statistics*, JSTOR, v. 13, n. 2, p. 215–232, 1942.

CABALLERO, T. V. *Uma Comparação entre Métodos de Aproximações Determinísticas e Estocástica para Inferência Bayesiana em Modelos Dinâmicos Lineares Generalizados*. Tese (Mestrado em Estatística) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

CALSAVARA, V. F.; RODRIGUES, A. S.; ROCHA, R.; LOUZADA, F.; TOMAZELLA, V.; SOUZA, A. C.; COSTA, R. A.; FRANCISCO, R. P. Zero-adjusted defective regression models for modeling lifetime data. *Journal of Applied Statistics*, Taylor & Francis, v. 46, n. 13, p. 2434–2459, 2019.

CANDIDO, P. G. G.; AMADOR, B. M.; SILVA, F. F.; SANTOS, F. S.; PINHEIRO, L. M. d. L.; FILHO, A. B. d. O. Adherence to antiretroviral therapy among women living with HIV/AIDS in the interior of the Brazilian state of Pará: cross-sectional study. *São Paulo Medical Journal*, SciELO Brasil, v. 139, n. 2, p. 99–106, 2021.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. *Análise de sobrevivência: teoria e aplicações em saúde*. 2. ed. Rio de Janeirol: Fiocruz, 2011.

COLOSIMO, E.; VIEIRA, A. O modelo de regressao de Cox com covariável dependente do tempo: Uma aplicaçao envolvendo pacientes infectados pelo HIV. *Revista Brasileira de Estatıstica*, v. 54, n. 57, p. 139–152, 1996.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Editora Blucher, 2006.

CORDEIRO, G. M.; ORTEGA, E. M.; NADARAJAH, S. The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, Elsevier, v. 347, n. 8, p. 1399–1429, 2010.

COX, C.; CHU, H.; SCHNEIDER, M. F.; MUNOZ, A. Parametric survival analysis and taxonomy of hazard functions for the generalized Gamma distribution. *Statistics in Medicine*, Wiley Online Library, v. 26, n. 23, p. 4352–4374, 2007.

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. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968.

FISHER, R. A. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, v. 41, p. 155–160, 1912.

FREITAS, B. C. L.; PERES, M. V. de O.; ACHCAR, J. A.; MARTINEZ, E. Z. Classical and Bayesian inference approaches for the exponentiated discrete Weibull model with censored data and a cure fraction. *Pakistan Journal of Statistics and Operation Research*, College of Statistical and Actuarial Sciences, v. 17, n. 2, p. 467–481, 2021.

HASHIMOTO, E. M. *Modelo de regressão Gama-G em análise de sobrevivência*. 177 p. Tese (Doutorado em Ciências) — Universidade de São Paulo, Piracicaba, 2013.

HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. *Operations Research*, INFORMS, v. 31, n. 6, p. 1109–1144, 1983.

HENNINGSEN, A.; TOOMET, O. maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, v. 26, n. 3, p. 443–458, 2011.

HOSMER, D. W.; LEMESHOW, S. *Applied survival analysis*. New York: John Wiley and Sons, 1999.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. *Bayesian Survival Analysis*. New York: Springer, 2001.

JACKSON, C. lexsurv a platform for parametric survival modeling in R. *Journal of Statistical Software*, v. 70, n. 8, p. 1–33, 2016.

JONES, M. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, Elsevier, v. 6, n. 1, p. 70–81, 2009.

JÚNIOR, M. R. O.; MOREIRA, F.; LOUZADA, F. The zero-inflated promotion cure rate model applied to financial data on time-to-default. *Cogent Economics & Finance*, Cogent OA, v. 5, n. 1, p. 1395950, 2017.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. 2. ed. [S.l.]: John Wiley & Sons, 2002.

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.

KLEINBAUM, D. G.; KLEIN, M. *Survival analysis*. 3. ed. Nova York: Springer, 2011.

LAMBERT, P. C.; THOMPSON, J. R.; WESTON, C. L.; DICKMAN, P. W. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, Oxford University Press, v. 8, n. 3, p. 576–594, 2007.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2003. v. 362.

LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. 3. ed. New Jersey: John Wiley & Sons, 2003.

LUKACS, E. A characterization of the gamma distribution. *The Annals of Mathematical Statistics*, JSTOR, v. 26, n. 2, p. 319–324, 1955.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. New York: Wiley, 1996.

MARTINEZ, E. Z.; ACHCAR, J. A. Trends in epidemiology in the 21st century: time to adopt Bayesian methods. *Cadernos de Saúde Pública*, SciELO Brasil, v. 30, p. 703–714, 2014.

MARTINEZ, E. Z.; ACHCAR, J. A.; JÁCOME, A. A.; SANTOS, J. S. Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 112, n. 3, p. 343–355, 2013.

MARTINS, T. A.; KERR, L.; KENDALL, C.; MOTA, R. M. S. Cenário epidemiológico da infecção pelo HIV e AIDS no mundo. *Rev Fisioter S Fun*, v. 3, n. 1, p. 4–7, 2014.

MEDEIROS, A. R. C.; LIMA, R.; MEDEIROS, L. B. d.; MORAES, R. M. d.; VIANNA, R. P. d. T. Análise de sobrevida de pessoas vivendo com HIV/AIDS. *Revista de Enfermagem UFPE Online*, v. 11, n. 1, p. 47–56, 2017.

MELO, G. C. d.; CARVALHO, A. C. A.; MOREIRA, A. d. S.; PAIXÃO, J. T. d. S. Tempo de sobrevida e distância para acesso a tratamento especializado por pessoas vivendo com HIV/Aids no estado de Alagoas, Brasil. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 24, 2021.

MELO, M. C. d.; ALMEIDA, V. C. d.; DONALISIO, M. R. Trend incidence of HIV-AIDS according to different diagnostic criteria in Campinas-SP, Brazil from 1980 to 2016. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 26, p. 297–307, 2021.

MELO, M. C. d.; DONALISIO, M. R.; CORDEIRO, R. C. Sobrevida de pacientes com AIDS e coinfecção pelo bacilo da tuberculose nas regiões Sul e Sudeste do Brasil. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 22, p. 3781–3792, 2017.

MILLION, W. Statistical modeling for the survival of HIV/AIDS patients treated with highly active anti-retroviral therapy (HAART): A case study at Dilchora Hospital, Dire Dawa, Ethiopia. *Journal of Biometrics & Biostatistics*, v. 9, n. 5, p. 1–10, 2018.

MORAES, T. E. N. T. de; PREVIDELLI, I.; SILVA, G. L. da. A Bayesian Weibull analysis of breast cancer data with long-term survivors in Paraná State, Brazil. *Revista Brasileira de Biometria*, v. 39, n. 2, p. 293–310, 2021.

MÜLLER, E. V.; BORGES, P. K. de O. Sobrevida de pacientes HIV/AIDS em tratamento antirretroviral e fatores associados na Região dos Campos Gerais, Paraná: 2002-2014. *Brazilian Journal of Development*, v. 6, n. 5, p. 28523–28542, 2020.

NAKHAEE, F.; LAW, M. Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: data from Australia. *Eastern Mediterranean Health Journal*, v. 17, n. 3, p. 231–237, 2011.

NETO, L. F. d. S. P.; PERINI, F. d. B.; ARAGÓN, M. G.; FREITAS, M. A.; MIRANDA, A. E. Protocolo brasileiro para infecções sexualmente transmissíveis 2020: infecção pelo HIV em adolescentes e adultos. *Epidemiologia e Serviços de Saúde*, SciELO Public Health, v. 30, p. e2020588, 2021.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. *Statistical Papers*, Springer, v. 51, n. 1, p. 111, 2010.

PARANAÍBA, P. F. *Caracterização e extensoes da distribuição Burr XII: Propriedades e Aplicações*. 142 p. Tese (Doutorado em Estatística e Experimentação Agronômica) — Universidade de São Paulo, 2012.

PARANAÍBA, P. F.; ORTEGA, E. M.; CORDEIRO, G. M.; PASCOA, M. A. d. The kumaraswamy burr xii distribution: theory and practice. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 11, p. 2117–2143, 2013.

PASCOA, M. A. D.; ORTEGA, E. M.; CORDEIRO, G. M. The Kumaraswamy Generalized Gamma distribution with application in survival analysis. *Statistical Methodology*, Elsevier, v. 8, n. 5, p. 411–433, 2011.

PAULINO, C. D. M.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003.

PENA-RAMIREZ, F. A.; GUERRA, R. R.; CORDEIRO, G. M.; MARINHO, P. R. The exponentiated power generalized weibull: Properties and applications. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 90, n. 3, p. 2553–2577, 2018.

PERALTA, D.; MAZUCHELI, J.; EMANUELLI, I.; ROSSI, R. M. Aplicação da distribuição BURR XII na análise do tempo até a primeira postura de codornas. *Revista Brasileria de Biometria*, v. 35, n. 1, p. 1–16, 2017.

PERES, M. V. D. O.; SANTOS, F. S. D.; OLIVEIRA, R. P. D. Estimation of survival and hazard curves of mixture Mirra cure rate model: Application to gastric and breast cancer data. *Biom Biostat Int J*, v. 9, n. 4, p. 132–137, 2020.

PERES, M. V. de O. *Aplicações das Distribuições Weibull Modificada e Beta-Weibull na Presença de Frações de Cura sob o Enfoque Frequentista e Bayesiano*. 94 p. Dissertação (Mestrado em Bioestatística) — Universidade Estadual de Maringá, Maringá, 2016.

POOROLAJAL, J.; HOOSHMAND, E.; MAHJUB, H.; ESMAILNASAB, N.; JENABI, E. Survival rate of AIDS disease and mortality in HIV-infected patients: a meta-analysis. *Public Health*, Elsevier, v. 139, p. 3–12, 2016.

PRENTICE, R. L. A log Gamma model and its maximum likelihood estimation. *Biometrika*, Oxford University Press, v. 61, n. 3, p. 539–544, 1974.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

RAMACHANDRAN, K. M.; TSOKOS, C. P. *Mathematical statistics with applications in R*. 2. ed. [S.l.]: Academic Press, 2015.

RIBEIRO, R. A.; FONSECA, F. F.; PEREIRA, G. F. M. evolução da AIDS no B: uma análise espacial. *Revista do Seminário Internacional de Estatística com R*, v. 4, n. 2, 2019.

ROSSI, R. M. *Introdução aos métodos Bayesianos na análise de dados zootécnicos com uso do WinBUGS e R*. Maringá: Eduem, 2011.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, v. 71, p. 319–392, 2009.

RUE, H.; RIEBLER, A.; SØRBYE, S. H.; ILLIAN, J. B.; SIMPSON, D. P.; LINDGREN, F. K. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, Annual Reviews, v. 4, p. 395–421, 2017.

SANTOS, C. A.; ACHCAR, J. A. A Bayesian analysis in the presence of covariates for multivariate survival data: an example of application. *Revista Colombiana de Estadística*, Universidad Nacional de Colombia., v. 34, n. 1, p. 111–131, 2011.

SANTOS, D. F. dos; ALMEIDA, P. L. R. de; OLIVEIRA, T. A. de. Distribuiçao Log-Normal e Log-Normal com fraçao de cura para dados de sobrevivência. *Sigmae*, v. 8, n. 2, p. 323–330, 2019.

SANTOS, R. de O.; NAKANO, E. Y. Análise do tempo de permanência de trabalhadores no mercado de trabalho do Distrito Federal via modelo de riscos proporcionais de Cox e log-normal. *Revista Brasileira de Biometria*, v. 33, n. 4, p. 570–584, 2015.

SILVA, A. de O.; FILHO, A. G. C. G.; SILVA, C. R. da; LEITE, D. R. A.; SILVA, L. C. M. da; FREITAS, W. W. da L. Modelos de sobrevivência aplicados à evasão dos alunos de Estatística da UFPB. *Revista InterScientia*, v. 6, n. 2, p. 134–145, 2018.

SOUZA, H. C. C. de; LOUZADA, F.; RAMOS, P. L.; JÚNIOR, M. R. de O.; PERDONÁ, G. S. C. A Bayesian approach for the zero-inflated cure model: an application in a Brazilian invasive cervical cancer database. *Journal of Applied Statistics*, Taylor & Francis, p. 1–17, 2021.

SPIEGELHALTER, D.; THOMAS, A.; BEST, N.; GILKS, W.; LUNN, D. *BUGS - Bayesian Inference using Gibbs Sampling*. Cambridge: MRC Biostatistics Unit, 1994.

STACY, E. W. et al. A generalization of the Gamma distribution. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 33, n. 3, p. 1187–1192, 1962.

SUGIURA, N. Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-theory and Methods*, Taylor & Francis, v. 7, n. 1, p. 13–26, 1978.

SZWARCWALD, C. L.; BASTOS, F. I. A.; ESTEVES, M. A. P.; ANDRADE, C. L. T. d. A disseminação da epidemia de AIDS no Brasil no período de 1987-1996: uma análise espacial. *Cadernos de Saúde Pública*, SciELO, v. 16, p. S07 – S19, 2000.

THERNEAU, T. M. *A Package for Survival Analysis in R*. [S.l.], 2020. R package version 3.2-3. Disponível em: <https://CRAN.R-project.org/package=survival>.

UNAIDS. *Global HIV & AIDS STATISTICS — 2020 fact sheet*. 2020. Disponível em: <https://www.unaids.org/en/resources/fact-sheet>. Acesso em: 04 fev. 2021.

VARSHNEY, M.; GROVER, G.; RAVI, V.; THAKUR, A. Cure fraction model for the estimation of long-term survivors of HIV/AIDS patients under antiretroviral therapy. *J Commun Disc*, v. 5, n. 3, p. 1–10, 2018.

WEIBULL, W. A statistical theory of strength of materials. *IVB-Handl.*, 1939.

WERLE, J. E. *HIV/AIDS em Mato Grosso do Sul*: análise de tendência, distribuição espacial e sobrevida dos casos. 75 p. Dissertação (Mestrado em Enfermagem) — Universidade Federal do Mato Grosso do Sul, Campo Grande, 2021.

ZASLAVSKY, R.; GOULART, B. N. G.; ZIEGELMANN, P. K. Cross-border healthcare and prognosis of HIV infection in the triple border Brazil-Paraguay-Argentina. *Cadernos de Saúde Pública*, SciELO Public Health, v. 35, p. e00184918, 2019.

Apêndices

APÊNDICE A

Modelos estatísticos para análise de sobrevivência de pacientes com HIV/AIDS desconsiderando censuras

Durante as análises feitas no Capítulo 5, considerou-se verificar se a remoção de pacientes que não apresentaram o evento de interesse desde a entrada no estudo (3.651 dias, ou 521,57 semanas), apresentaria uma melhor qualidade nos ajustes. A Tabela 15 apresenta as estimativas dos parâmetros dos modelos, em que novamente o modelo KGG foi o melhor. Entretanto as Figuras 25 (comparação entre os resultados do modelo e as estimativas de Kaplan-Meier) e 26 (análise de resíduos de Cox-Snell para o modelo KGG) indicam que a remoção dessas observações resultou na piora da qualidade dos modelos.

Tabela 15 – Estimativas para os parâmetros dos modelos considerando dados sem censuras com 521,57 semanas.

Modelo	Estimativas (erro padrão)					$\log(L(\theta))$	AIC
W (τ,α)	0,3801	56.188,99				-6.303,86	12.611,73
	(0,0051)	(1,0490)					
BXII (θ, α, γ)	194,3857	0,4500	0,1651			-6.298,81	12.603,62
	(1,0500)	(0,01366)	(0,0058)				
G (γ,α)	0,5867	0,0001				-6.417,85	12.839,71
	(0,0316)	(0,00003)					
LN (μ,σ)	11,9317	5,4101				-6.283,36	12.570,73
	(0,2468)	(0,1608)					
GG (μ, σ, Q)	6,998	9,1320	-3,6200			-6.261,50	12.528,99
	(1,0343)	(0,0346)	(0,6361)				
BW (λ,α,β,γ)	0,0008	16,4390	0,6935	0,0684		-6.284,03	12.576,06
	(0,0007)	(0,1403)	(0,1924)	(0,0034)			
KW (α, τ,k, λ, φ)	0,0503	0,2110	1,0000	5,7328	0,0289	-6.268,61	12.545,22
	(0,0793)	(0,0665)	FX$^{(1)}$	(1,2708)	(0,0343)		
KGG ($\alpha, \tau, k, \lambda, \varphi$)	0,0055	0,5735	0,2779	43,3572	0,0070	**-6.239,64**	**12.489,27**
	(0,0004)	(0,0009)	(0,0008)	(0,0018)	(0,0003)		
KBXII(λ, β, α,γ,θ)	5,2648	0,0133	0,4972	4,1314	1,4363	-6.263,65	12.537,29
	(0,2648)	(0,0006)	(0,0279)	(0,0755)	(0,0586)		
EPWG ($\lambda, \alpha, \beta, \gamma$)	6,7937	0,0276	1,5325	1,2581		-6.251,79	12.511,58
	(0,8488)	(0,0049)	(0,1070)	(0,1363)			

$^{(1)}$: não apresenta erro padrão devido a k ser um parâmetro fixo.

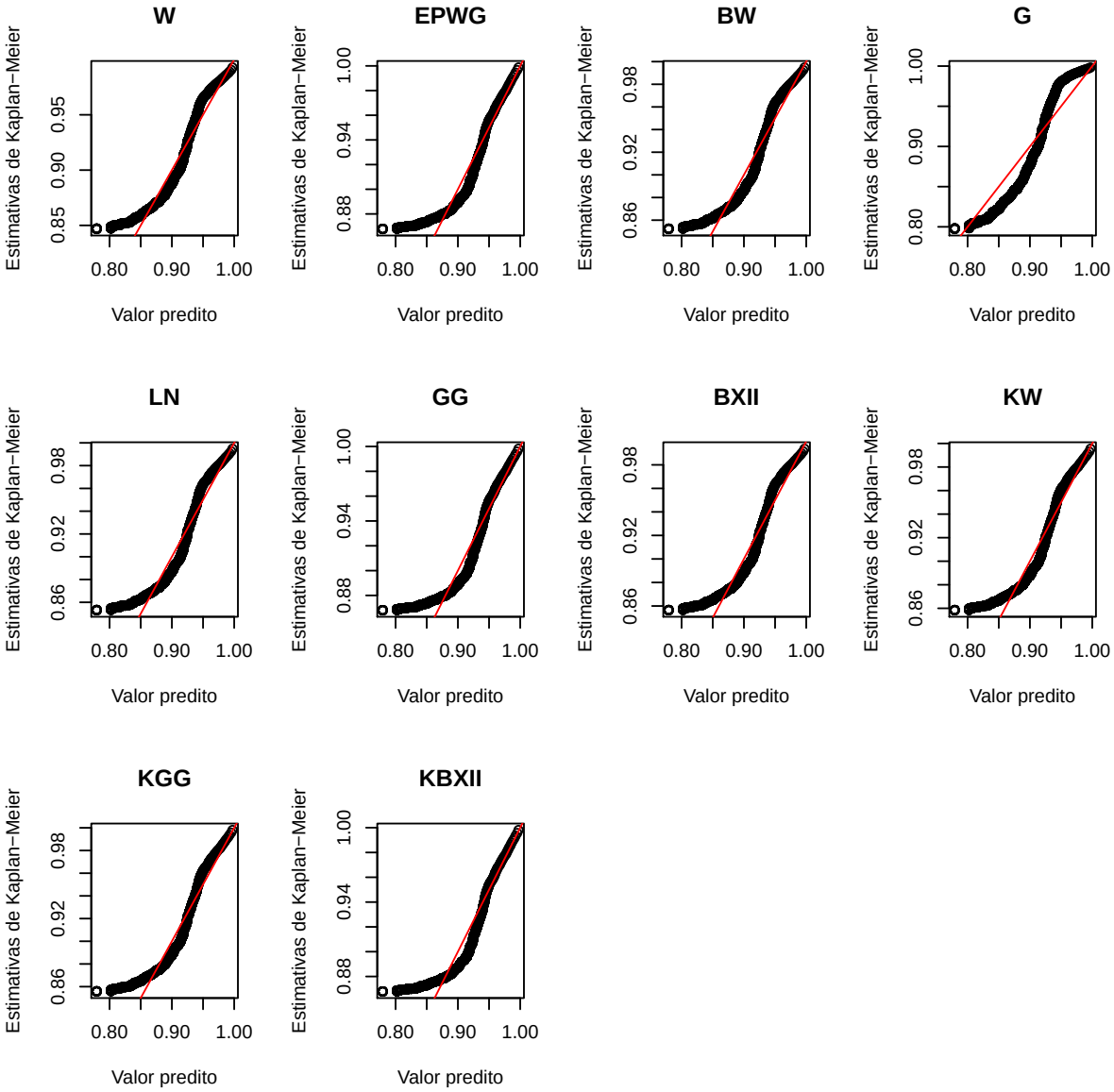

Figura 25 – Comparações entre as estimativas de Kaplan-Meier e valores preditos para os modelos probabilísticos sem as observações com 521,57 semanas.

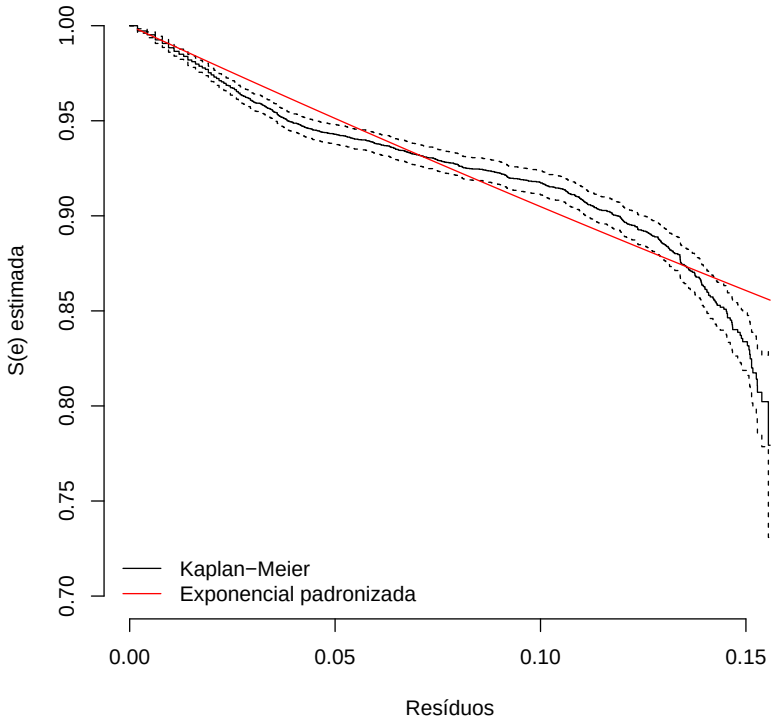

Figura 26 – Resíduos de Cox-Snell para a distribuição KGG considerando os dados sem observações com 521,57 semanas.

APÊNDICE B

Principais códigos em R

Códigos em R para a função de densidade de probabilidade, função de sobrevivência e do logaritmo da função de verosimilhança obtenção das estimativas do modelo KGG (Capítulo 5).

```
library(flexsurv); library(maxLik)

fkumgg = function(t, tau, alpha, k, lambda, phi)
{
gr   = pgengamma.orig(t, shape = tau, scale = alpha, k = k)
dens = ((lambda*phi*tau)/(alpha*gamma(k)))*(t/alpha)^(tau*k-1)*
exp(-(t/alpha)^tau)*(gr)^(lambda-1)*(1-(gr)^lambda)^(phi-1)
return(dens)
}


Skumgg = function(t, tau, alpha, k, lambda, phi)
{
gr = pgengamma.orig(t, shape=tau, scale=alpha, k=k)
st = (1-gr^lambda)^phi
return(st)
}
```

```
Lkumgg = function(par)
{
tau    = par[1]
alpha  = par[2]
k      = par[3]
lambda = par[4]
phi    = par[5]

if(par[1]<=0)return(-Inf)
if(par[2]<=0)return(-Inf)
if(par[3]<=0)return(-Inf)
if(par[4]<=0)return(-Inf)
if(par[5]<=0)return(-Inf)

A = cen*log(fkumgg(t,tau,alpha,k,lambda,phi))
B = (1-cen)*log((Skumgg(t,tau,alpha,k,lambda,phi)))
L = exp(A+B)
logL = sum(log(L))

if(is.na(logL)==TRUE){return(-Inf)}else{return(logL)}
}
```

Código em R para obtenção das estimativas do modelo ZCKGG (Capítulo 6).

```
Lkumggcom0cura = function(par)
{
p0     = par[1]
tau    = par[2]
alpha  = par[3]
k      = par[4]
lambda = par[5]
phi    = par[6]
fc     = par[7]
if(par[1]<=0)return(-Inf)
if(par[2]<=0)return(-Inf)
if(par[3]<=0)return(-Inf)
if(par[4]<=0)return(-Inf)
if(par[5]<=0)return(-Inf)
if(par[6]<=0)return(-Inf)
if(par[7]<=0)return(-Inf)
if(par[1]+par[7]>1)return(-Inf)
gr   = pgengamma.orig(t, shape=tau, scale=alpha, k=k)
dens = ((lambda*phi*tau)/(alpha*gamma(k)))*(t/alpha)^(tau*k-1)*
exp(-(t/alpha)^tau)*(gr)^(lambda-1)*(1-(gr)^lambda)^(phi-1)
res  = dens*(1-p0-fc)
res[which(t==0)] = p0
f0   = res
gr   = pgengamma.orig(t, shape=tau, scale=alpha, k=k)
st   = (1-gr^lambda)^phi
S0   = fc+(1-p0-fc)*st
L    = ((f0)^cen*(S0)^(1-cen))
logL = sum(log(L))
if(is.na(logL)==TRUE){return(-Inf)}else{return(logL)}
}
```