

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
BACHARELADO EM ESTATÍSTICA

JÉSSICA MARIA MAGALHÃES MORENO

**ANÁLISE BAYESIANA DOS FATORES QUE INFLUENCIAM NO TEMPO ATÉ O
ÓBITO POR COVID-19 NO MUNICÍPIO DE COLÍDER-MT**

CUIABÁ - MT

2024

JÉSSICA MARIA MAGALHÃES MORENO

**ANÁLISE BAYESIANA DOS FATORES QUE INFLUENCIAM NO TEMPO ATÉ O ÓBITO POR
COVID-19 NO MUNICÍPIO DE COLÍDER-MT**

Trabalho de Conclusão de Curso apresentado à banca examinadora do Curso de Bacharelado em Estatística da Universidade Federal de Mato Grosso, como requisito parcial para a obtenção do título de Bacharel em Estatística

Orientador: Prof. Dr. Marcelino Alves Rosa
de Pascoa

CUIABÁ - MT
2024

JÉSSICA MARIA MAGALHÃES MORENO

**ANÁLISE BAYESIANA DOS FATORES QUE INFLUENCIAM NO TEMPO ATÉ O
ÓBITO POR COVID-19 NO MUNICÍPIO DE COLÍDER-MT**

Trabalho de Conclusão de Curso apresentado à banca examinadora do Curso de Bacharelado em Estatística da Universidade Federal de Mato Grosso, como requisito parcial para a obtenção do título de Bacharel em Estatística

APROVADA em _____ de _____ de _____

BANCA EXAMINADORA

Orientador: Prof. Dr. Marcelino Alves Rosa De Pascoa

Examinadora: Prof.^a Dr.^a Juscelia Dias Mendonça

Examinadora: Prof.^a Dr.^a Lia Hanna Martins Morita

CUIABÁ - MT

2024

AGRADECIMENTOS

À Deus, pela força para que eu pudesse prosseguir os meus estudos com perseverança e sabedoria.

Ao meu orientador, **Prof. Dr. Marcelino Alves Rosa De Pascoa**, pela compreensão, paciência e orientação no desenvolvimento deste trabalho.

À **Prof.^a Dr.^a Lia Hanna Martins Morita**, pelos conselhos e incentivo que foram essenciais para a vida acadêmica.

À todos os professores do Departamento de Estatística (DEst-UFMT), pelos ensinamentos e possibilidades que contribuíram para minha formação acadêmica.

Ao Técnico Administrativo do Departamento de Estatística (DEst-UFMT), **Pedro Henrique Delfino**, pela atenção, suporte e amizade.

Por fim, quero agradecer à minha mãe, **Arenil Natalina de Magalhães**, por sua confiança e investimento em minha educação. Seu carinho e dedicação foram a luz que me permitiu seguir em frente em muitos momentos.

RESUMO

A COVID-19 foi declarada uma emergência de saúde pública de importância internacional pela Organização Mundial da Saúde, apresentando desafios significativos para os estados e municípios brasileiros. Este estudo, realizado em Colíder-MT, teve como objetivo identificar uma função de distribuição que melhor se ajustasse ao tempo de hospitalização do paciente até o óbito por complicações da COVID-19, bem como avaliar os fatores associados ao aumento do risco de morte. Utilizando análise de sobrevivência e inferência bayesiana. Pelas análises realizadas, verificou-se que o risco de morte por COVID-19 aumenta cerca de 2,7% a cada ano de idade. Pacientes com pneumopatia apresentam um risco aproximadamente 173% maior de falecer em decorrência da COVID-19 em comparação com aqueles sem essa condição, enquanto pacientes submetidos à ventilação mecânica têm um risco de morte reduzido em cerca de 60%. Espera-se que esses resultados possam orientar pesquisadores, gestores e profissionais da saúde na formulação de estratégias eficazes para o enfrentamento da pandemia.

Palavras-chaves: locação-escala; Weibull; Regressão Log-Weibull; bayesiana.

ABSTRACT

COVID-19 has been declared an international public health emergency by the World Health Organization, presenting significant challenges for Brazilian states and municipalities. This study, conducted in Colíder-MT, aimed to identify a distribution function that best fits the patient's hospitalization time until death from COVID-19 complications and to assess factors associated with increased risk of death. Using survival analysis and Bayesian inference, it was found that the risk of death from COVID-19 increases by roughly 2.7% per year of age. Patients with lung disease face a risk approximately 173% higher of dying from COVID-19 compared to those without this condition. In comparison, patients undergoing mechanical ventilation have a reduced risk of death by about 60%. Hopefully, these results will guide researchers, managers, and healthcare professionals in formulating effective strategies to combat the pandemic.

Keywords: location-scale; Weibull; Log-Weibull Regression; Bayesian.

LISTA DE FIGURAS

Figura 1	– Gráficos ilustrativos de algumas curvas TTT.	13
Figura 2	– TTT-Plot para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	25
Figura 3	– Densidades marginais a <i>posteriori</i> para os parâmetros do modelo Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	26
Figura 4	– Cadeias geradas pelo método MCMC para os parâmetros α e γ referentes ao modelo Weibull.	27
Figura 5	– (a) Densidade estimada do modelo Weibull para os dados de tempo até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT. (b) Estimativas da função de sobrevivência segundo Kaplan-Meier e segundo o modelo Weibull, para os dados de tempo até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT. (c) Risco estimado pelo modelo Weibull, para os dados em estudo.	28
Figura 6	– Densidades marginais a <i>posteriori</i> para os parâmetros do modelo de regressão Log-Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	29
Figura 7	– Cadeias geradas pelo método MCMC para os parâmetros σ e β_s referentes ao modelo de regressão Log-Weibull.	30

LISTA DE TABELAS

Tabela 1 – Estatísticas descritivas para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	24
Tabela 2 – Sumário a <i>posteriori</i> dos parâmetros do modelo Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	24
Tabela 3 – Sumário a <i>posteriori</i> dos parâmetros do modelo de regressão Log-Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.	27

SUMÁRIO

1	INTRODUÇÃO	9
1.1	Justificativa	9
1.2	Objetivo Geral	10
1.3	Objetivos Específicos	10
2	REFERENCIAL TEÓRICO	11
2.1	Conceitos básicos em análise de sobrevivência	11
2.1.1	Distribuição do tempo de sobrevivência	12
2.1.2	O estimador de Kaplan-Meier	13
2.2	Distribuição Weibull	14
2.3	Distribuição Log-Weibull	14
2.4	Modelo de regressão	15
2.4.1	Modelo de regressão locação-escala	15
2.5	Modelo de regressão Log-Weibull	16
2.6	Análise Bayesiana	16
2.7	Cadeias de Markov	17
2.7.1	Análise bayesiana (Modelo Weibull)	20
2.7.2	Análise bayesiana (Modelo Log-Weibull)	21
3	MATERIAL E MÉTODOS	22
3.1	Dados	22
3.2	Descrição de Variáveis	22
3.3	Análise Estatística	22
4	RESULTADOS E DISCUSSÕES	24
5	CONSIDERAÇÕES FINAIS	31
	REFERÊNCIAS	32

1 INTRODUÇÃO

A COVID-19 é uma doença provocada pelo vírus SARS-CoV-2 que foi detectada pela primeira vez no território chinês no final de 2019, e rapidamente se espalhou pelo mundo. Em resposta à rápida disseminação do vírus, a Organização Mundial da Saúde (OMS) declarou a COVID-19 como uma Emergência de Saúde Pública de Importância Internacional em 30 de janeiro de 2020 e, posteriormente, elevou a situação ao status de pandemia em 11 de março de 2020 (Oliveira et al.,2020). A principal forma de contágio do vírus ocorre de indivíduo para indivíduo, por meio do sistema respiratório ou após o contato com superfícies infectadas, seguido de toque nos olhos, nariz ou boca (Chate et al.,2020). No Brasil, em resposta à pandemia, o Ministério da Saúde divulgou uma série de orientações para a comunidade, com o objetivo de esclarecer sobre a transmissão, a prevenção e os procedimentos em caso de infecção pelo vírus. Uma das medidas mais significativas adotadas foi o isolamento social para evitar a propagação do COVID-19 (Duarte et al.,2020).

Localizado no estado de Mato Grosso, o município de Colíder confirmou seu primeiro caso de COVID-19 em 20 de maio de 2020, conforme informações disponíveis no site da Prefeitura (Colíder, 2020). Desde então, a administração local tem se dedicado a combater a pandemia, atualizando regularmente a população sobre o número de casos e o avanço da campanha de vacinação através do “Vacinômetro”. Além disso, a prefeitura tem emitido decretos e boletins informativos para manter os cidadãos bem informados sobre as últimas notícias e os progressos relacionados à COVID-19.

No estudo conduzido por Galvão e Roncalli (2020), foram utilizados dados secundários para realizar uma análise de sobrevivência em indivíduos notificados por COVID-19 no estado do Rio Grande do Norte, Brasil. O objetivo principal foi identificar os fatores de risco e os fatores prognósticos para a ocorrência de óbitos por COVID-19. Para isso, os autores empregaram o método de Kaplan-Meier e o modelo de regressão de Cox dependente do tempo. O estudo revelou que os indivíduos com 80 anos ou mais, do sexo masculino, com cor da pele não branca ou sem informação, que tinham comorbidades ou cuja presença de comorbidades não foi informada, apresentaram maiores riscos de óbito por COVID-19. Pascoa (2021) realizou uma análise utilizando o modelo de Weibull para avaliar a eficácia dos ajustes nos dados da COVID-19 do estado do Mato Grosso. O estudo revelou que a idade dos pacientes era o fator mais significativo no risco de mortalidade. A pandemia de COVID-19 impôs significativos desafios aos estados e municípios brasileiros, evidenciando disparidades no número de casos e óbitos, assim como nas condições de infraestrutura e saúde pública. Dentro desse contexto, o presente estudo foca no município de Colíder, Mato Grosso, e tem como objetivo principal identificar a função de distribuição que mais adequadamente descreve o tempo de internação hospitalar até o óbito por complicações da COVID-19, além de avaliar os fatores que elevam o risco de mortalidade pela doença. Os dados analisados abrangem o período de fevereiro a julho de 2020.

1.1 Justificativa

A pandemia de COVID-19 impactou significativamente diversos municípios brasileiros, cada um com suas particularidades. Em Colíder-MT, as limitações na infraestrutura de saúde e as características socioeconômicas da população conferiram desafios únicos no controle e no tratamento da doença. Este estudo se justifica pela necessidade de entender melhor como essas variáveis influenciam o risco de morte

por COVID-19, proporcionando bases para políticas públicas mais eficazes e direcionadas à realidade local.

1.2 Objetivo Geral

O objetivo geral deste trabalho é identificar quais os fatores predisponentes que contribuem para aumentar o risco de morte por COVID-19, considerando as internações no período compreendido de fevereiro a julho de 2020 no município de Colíder-MT.

1.3 Objetivos Específicos

- Aplicar técnicas de análise de sobrevivência para modelar o tempo desde a hospitalização até o óbito de pacientes com COVID-19;
- Determinar a função de distribuição que melhor se ajusta ao tempo de internação até o óbito, utilizando modelos estatísticos adequados;
- Avaliar o impacto de variáveis demográficas e clínicas, como idade, comorbidades e tratamentos recebidos, no risco de mortalidade.

2 REFERENCIAL TEÓRICO

Nesta seção, é feita uma revisão dos conceitos básicos e essenciais sobre análise de sobrevivência, incluindo a distribuição Weibull e o modelo de regressão Log-Weibull. Também são apresentados os métodos inferenciais para a análise de dados de sobrevivência, incluindo a análise bayesiana.

2.1 Conceitos básicos em análise de sobrevivência

Na análise de sobrevivência, as respostas são definidas pelos tempos de falha ou pela existência de censura. Este método é aplicado quando se pretende examinar dados que correspondem ao até a ocorrência de um evento relevante. Assim, supõe-se T como uma variável aleatória não-negativa e contínua que simboliza o tempo de sobrevivência de um sujeito (ou objeto) e pode ser representada por funções matemáticas.

Existem várias razões para a ocorrência de dados censurados, como: falha dos sujeitos por motivos distintos daqueles de interesse da pesquisa, planejamento do estudo, perda de registros, entre outros. No entanto, em uma análise, é crucial levar em conta todas as informações sobre a variável de interesse, incluindo os registros de informações censuradas. A negligência dessas observações pode resultar em conclusões tendenciosas ou distorcidas ao calcular estatísticas relevantes.

No caso em que todos os sujeitos em estudo falharam, o tempo até a falha de cada sujeito é conhecido. Este caso é denominado de dados completos. Os dados de sobrevivência podem ser representados pelo par (t_i, δ_i) , em que $i = 1, \dots, n$, t_i é o tempo de falha ou de censura do i -ésimo indivíduo e δ_i é a função indicadora de censura, dada por

$$\delta_i = \begin{cases} 1 & , T \leq C; \\ 0 & , T > C. \end{cases}$$

em que $\delta_i = 0$ indica censura e $\delta_i = 1$ indica falha (Colosimo e Giolo, 2006).

Diversos mecanismos de censura são aplicáveis quando se trata de análises estatísticas em estudos que lidam com dados censurados:

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo, sendo que alguns indivíduos não sofreram o evento de interesse, ou seja, não falharam. Desta maneira, o tempo de falha de todos os indivíduos que não foram observados completamente são considerados como censura.
- Censura do tipo II: antes do início do experimento é pré-estabelecida uma quantidade r de indivíduos a apresentarem falhas e o estudo é terminado após r falhas. Assim, para um estudo com n indivíduos, ao final tem-se $n - r$ observações censuradas.
- Censura aleatória: caracterizada pela perda de acompanhamento de um indivíduo no decorrer do estudo sem este ter apresentado a falha. Na prática, este é o tipo de censura que mais aparece, sendo representado por meio de duas variáveis aleatórias (v.a) independentes T_i e C_i , em que T_i representa o tempo de falha do i -ésimo indivíduo e C_i representa o tempo de censura associado a este indivíduo. Neste caso, o tempo de falha da i -ésima observação é dado por $t_i = \min(T_i, C_i)$.

- Censura intervalar: sabe-se que o evento de interesse ocorreu em um certo intervalo de tempo, sendo assim, o tempo de falha não é conhecido exatamente, mas pertence ao intervalo de tempo, com L e U constantes conhecidas.

2.1.1 Distribuição do tempo de sobrevivência

O tempo de vida de um indivíduo proveniente de uma dada população homogênea é representado por uma variável aleatória T não negativa. A distribuição de T pode ser descrita por meio das funções de densidade, probabilidade, sobrevivência ou risco.

A função densidade de probabilidade, $f(t)$, é definida por:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t},$$

em que $F(t) = P[T \leq t]$ é a função de distribuição acumulada (fda) de T . Na análise de sobrevivência, a função $f(t)$ é a função de densidade de probabilidade do tempo de sobrevivência. A $f(t)$ pode ser interpretada como a probabilidade do paciente ou elemento em estudo sofrer a falha no intervalo de tempo Δt e, como toda função $f(t)$, possui duas propriedades:

$$f(t) \geq 0 \quad \text{e} \quad \int_0^{\infty} f(t)dt = 1.$$

Em análise de sobrevivência, a função de sobrevivência, $S(t)$, é definida como a probabilidade de um indivíduo não falhar até o tempo t , ou seja, seu tempo de vida ser superior a t .

A fórmula da função de sobrevivência é dada por:

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx,$$

em que $\lim_{t \rightarrow 0} S(t) = 1$ e $\lim_{t \rightarrow \infty} S(t) = 0$.

A função risco ou taxa de falha, $h(t)$, representa o risco instantâneo no itempo t condicionado à sobrevivência até o tempo t é definida como:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Uma maneira prática de analisar o comportamento da função de risco é através da construção do gráfico do tempo total em teste (TTT), conforme sugerido por Aarset(1987). A curva TTT é gerada ao se criar um gráfico de:

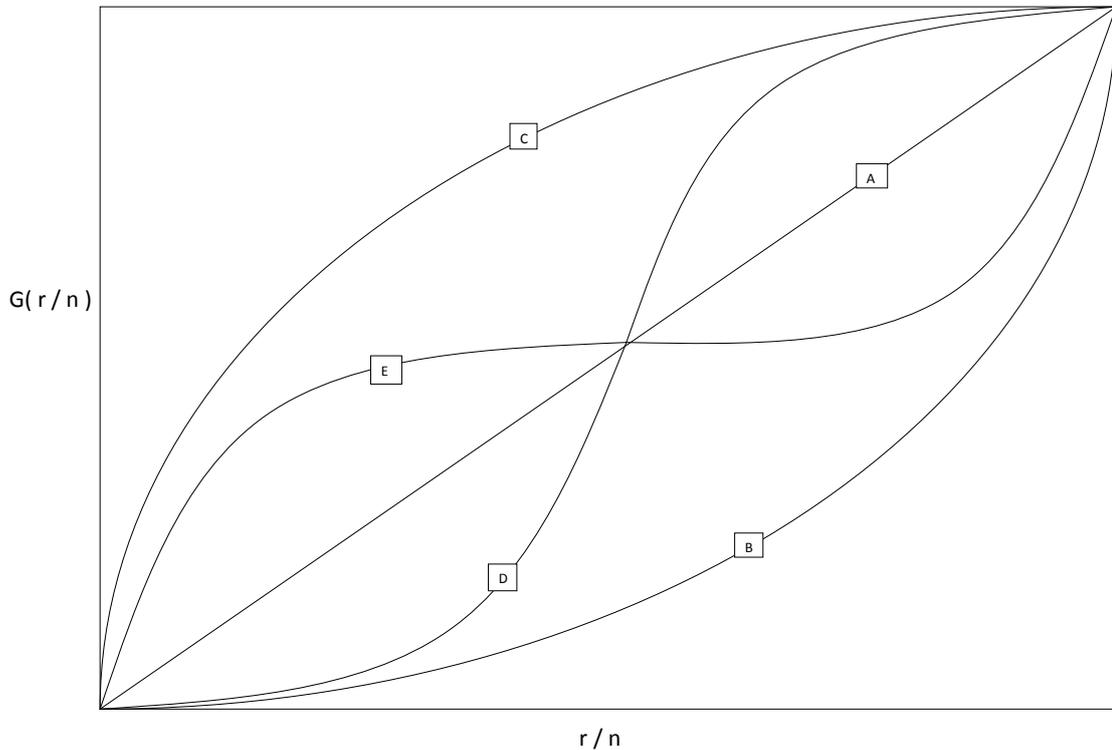
$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^r T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^n T_{i:n}} \quad \text{por} \quad \frac{r}{n},$$

em que n é o tamanho da amostra, $r = 1, \dots, n$ e $T_{i:n}, i = 1, \dots, n$ são estatísticas de ordem da amostra (Mudholkar, Srivastava e Kollia ,1996).

A Figura 1 demonstra as várias formas possíveis da curva TTT (Pascoa, 2012), em que observa-se uma linha diagonal quando o risco é constante (linha A), uma curva convexa quando o risco é decrescente

(curva B) e côncava quando o risco é crescente (curva C). Se uma curva inicialmente convexa torna-se côncava (curva D), indica um risco em forma de “U”, enquanto a situação inversa (curva E) indica um risco unimodal.

Figura 1 – Gráficos ilustrativos de algumas curvas TTT.



Fonte: Pascoa (2012).

2.1.2 O estimador de Kaplan-Meier

Um método não-paramétrico para estimar a função de sobrevivência na presença de censuras foi introduzido por Kaplan e Meier (1958). Este método é comumente conhecido como estimador de Kaplan-Meier ou estimador produto-limite.

Este método consiste na criação de uma série de intervalos de tempo, correspondentes ao número de falhas distintas observadas. Os limites de cada intervalo são definidos pelos tempos de falha que foram registrados na amostra.

Portanto, podemos descrever o estimador de Kaplan-Meier da seguinte maneira:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right),$$

em que $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ representam os k tempos de falha distintos e ordenados, d_j é o número de falhas em $t_{(j)}$, $j = 1, \dots, k$, e n_j é o número de indivíduos sob risco em $t_{(j)}$, abrangendo os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a $t_{(j)}$.

2.2 Distribuição Weibull

A distribuição de probabilidade Weibull foi inicialmente sugerida por Fisher e Tippett em 1928. Posteriormente, em 1939, o físico sueco Waloddi Weibull desenvolveu de forma independente a mesma distribuição enquanto estudava a resistência dos materiais. Após a Segunda Guerra Mundial, devido ao foco na resistência dos materiais, o trabalho de Weibull ganhou destaque e a distribuição passou a ser conhecida pelo seu nome (Bailey (1973); Batista (1989)). A distribuição Weibull tem uma ampla gama de aplicações em diversas áreas da ciência, sendo frequentemente empregada para modelar tempos de sobrevivência em estudos biomédicos e industriais (Colosimo e Giolo, 2006). Considere T como uma variável aleatória que segue a distribuição Weibull. A função densidade de probabilidade de T pode ser expressa da seguinte maneira:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp \left[- \left(\frac{t}{\alpha} \right)^\gamma \right], \quad t \geq 0, \quad (2.1)$$

onde γ é o parâmetro que define a forma, α é o parâmetro que determina a escala e ambos são positivos. Portanto, pode-se afirmar que T segue uma distribuição Weibull com parâmetros α e γ , ou seja, $T \sim \text{Weibull}(\alpha, \gamma)$. Vale ressaltar que quando o parâmetro γ é igual a 1, a distribuição Weibull se reduz à distribuição exponencial. As funções de sobrevivência e risco são caracterizadas, respectivamente, da seguinte maneira:

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

e

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

O parâmetro γ é responsável por definir a forma da função risco. Quando $\gamma < 1$, a função risco apresenta um comportamento decrescente. Por outro lado, se $\gamma > 1$, a função risco é crescente. Caso $\gamma = 1$, a função risco permanece constante.

2.3 Distribuição Log-Weibull

Seja T uma variável aleatória com distribuição Weibull, com função densidade de probabilidade representada pela equação (2.1). Considere a transformação $Y = \log(T)$ e as reparametrizações $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp(\mu)$. Nesse caso, por meio do método do jacobiano, a função densidade de probabilidade de Y pode ser escrita como:

$$f(y) = \frac{1}{\sigma} \exp \left[\left(\frac{y - \mu}{\sigma} \right) - \exp \left(\frac{y - \mu}{\sigma} \right) \right], \quad (2.2)$$

em que $-\infty < y < \infty$, $-\infty < \mu < \infty$ e $\sigma > 0$. A variável transformada com função densidade de probabilidade (2.2) é denominada Log-Weibull, em que $\mu \in \mathbb{R}$ é o parâmetro de locação e $\sigma > 0$ é o parâmetro de escala. Assim, se

$$T \sim \text{Weibull}(\alpha, \gamma) \quad \text{então} \quad Y = \log(T) \sim \text{LWeibull}(\mu, \sigma).$$

A função de sobrevivência para a distribuição Log-Weibull é expressa por:

$$S(y) = \exp \left[- \exp \left(\frac{y - \mu}{\sigma} \right) \right]. \quad (2.3)$$

2.4 Modelo de regressão

Na prática, é comum a ocorrência de situações em que uma ou mais covariáveis estão relacionadas aos tempos de sobrevivência, isto é, os tempos de falha são influenciados por covariáveis. Na área das ciências biomédicas, por exemplo, podem ser a idade, a altura, um tipo de tumor cancerígeno, a quantidade de hemoglobina no sangue e etc.

Considere T uma variável aleatória e seja $\mathbf{x} = (x_1, \dots, x_p)^T$ um vetor formado por p variáveis explanatórias. Um modo de estabelecer uma relação entre T e \mathbf{x} é por meio da utilização de modelos de regressão. No contexto dos dados de sobrevivência, uma maneira de incluir as covariáveis na análise é realizada, utilizando-se a classe de modelos de locação e escala, também conhecidos como modelo de tempo de vida acelerado ou modelo paramétrico. Existe uma vasta literatura sobre modelos de regressão locação-escala, por exemplo, Cox e Oakes (1984), Kalbfleisch e Prentice (2002) e Lawless (2003).

2.4.1 Modelo de regressão locação-escala

Supor que os tempos de vida T são transformados na forma $Y = \log T$. A classe dos modelos de locação-escala corresponde àquela para a qual, dado \mathbf{x} , a variável Y tem uma distribuição com um parâmetro de locação $\mu(\mathbf{x})$ que introduz as covariadas no modelo e um parâmetro de escala $\sigma \geq 0$ constante. Estas quantidades relacionam-se através da expressão:

$$Y = \mu(\mathbf{x}) + \sigma Z,$$

onde e tem uma distribuição que é independente de \mathbf{x} (Tomazella, 1994).

Em geral, $\mu(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$. Note que nos modelos de locação-escala as variáveis independentes atuam linearmente sobre Y e portanto multiplicativamente sobre T , ao passo que nos modelos com riscos proporcionais \mathbf{x} atua multiplicativamente sobre a função de risco de T .

supondo que Y pertence à família de distribuições que se caracteriza pelo fato de ter um parâmetro de locação μ ($-\infty < \mu < \infty$) e um parâmetro de escala σ ($0 < \sigma < \infty$).

As distribuições que pertencem a essa família têm função densidade de probabilidade da forma:

$$f(y; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{y - \mu}{\sigma}\right), \quad -\infty < y < \infty$$

A função de sobrevivência para Y , dado \mathbf{x} , é da forma: $G\left(\frac{y - \mu}{\sigma}\right)$.

Considerando a transformação $Y = \log(T)$, tem-se que o parâmetro de locação μ depende do vetor de covariáveis \mathbf{x} , denotado por $\mu(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$, em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ é o vetor de parâmetros desconhecidos. O modelo de regressão que relaciona Y e o vetor de variáveis explicativas \mathbf{x} é o modelo de locação-escala, que pode ser escrito da seguinte forma:

$$Y = \mu(\mathbf{x}) + \sigma Z,$$

em que Z é um erro aleatório.

Observa-se que esse modelo é log-linear para T , logo, é um modelo de regressão linear para Y . Outra característica desse modelo é que o vetor de variáveis explicativas \mathbf{x} tem efeito multiplicativo em T . Além disso, a função de sobrevivência para Y dado \mathbf{x} , tem a forma $G\left(\frac{y - \mu(\mathbf{x})}{\sigma}\right)$, em que $G(z)$ é a função de sobrevivência de Z . Algumas literaturas com (Lawless, 2003). apresenta mais detalhes dessa classe de

modelos, além de vários modelos usando distribuições de probabilidade para T , comumente usadas em análise de sobrevivência.

Na presença de censura intervalar, os dados observados consistem de um intervalo $(\log(u_i), \log(v_i))$ para cada indivíduo, no qual esses intervalos são conhecidos e inclui y_i com probabilidade um, ou seja, $P(\log(u_i) \leq y_i \leq \log(v_i)) = 1$, e se $\log(v_i) = \infty$, então o logaritmo do tempo é censurado à direita para y_i .

2.5 Modelo de regressão Log-Weibull

O modelo de locação e escala considerando Y dado \mathbf{x} , tem distribuição Log-Weibull com função densidade de probabilidade dada pela expressão (2.2) pode ser representado por:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma Z_i, \quad i = 1, \dots, n, \quad (2.4)$$

onde $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ é um vetor de parâmetros desconhecidos associado a cada variável, $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ é o vetor de variáveis e Z_i é o erro aleatório.

A distribuição da variável aleatória Z pode ser obtida, utilizando-se a transformação de variáveis. A função densidade de probabilidade $f(z)$, definida em (2.5), pode ser expressa como:

$$f(z) = \frac{1}{\sigma} \exp [z - \exp(z)], \quad (2.5)$$

onde $z = \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}$.

A função de sobrevivência para a variável aleatória Z é expressa por:

$$S(z) = \exp [-\exp(z)]. \quad (2.6)$$

2.6 Análise Bayesiana

Dado que $\mathbf{t} = (t_1, t_2, \dots, t_n)^T$ seja um vetor com n observações e $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)^T$. O vetor de parâmetros é uma representação matemática das características de uma população de interesse. Um dos principais objetivos da estatística é estimar os parâmetros populacionais com base em um conjunto de dados amostrais. Uma metodologia amplamente empregada para atingir esse objetivo é a inferência bayesiana (Silva, 2009).

Na abordagem da inferência bayesiana, as n observações estão relacionadas a funções de densidade de probabilidade $f(\mathbf{t}|\boldsymbol{\theta})$, onde $\boldsymbol{\theta}$ indexa a família de distribuições das observações. O valor verdadeiro de $\boldsymbol{\theta}$ é desconhecido, sendo tratado como uma variável aleatória. Nesse contexto, $\boldsymbol{\theta}$ segue uma distribuição associada. Na análise bayesiana, a informação disponível anteriormente sobre $\boldsymbol{\theta}$ desempenha um papel crucial. Essa informação é expressa na forma de uma distribuição de probabilidade, que denotamos como $\pi(\boldsymbol{\theta})$.

A distribuição a *priori* deve expressar, probabilisticamente, o entendimento que possuímos sobre o parâmetro $\boldsymbol{\theta}$. Essa distribuição pode ser considerada informativa ou não informativa. As distribuições a *priori* informativas são aplicadas quando temos algum conhecimento prévio sobre o parâmetro, enquanto as distribuições não informativas são utilizadas quando possuímos pouca ou nenhuma informação sobre o parâmetro (Paulino, Turkman e Murteira, 2003).

A informação *a priori*, combinada com a função dos dados, conhecida como função de verossimilhança $L(\boldsymbol{\theta}|\mathbf{t})$, resulta na distribuição *a posteriori* $\pi(\boldsymbol{\theta}|\mathbf{t})$. Essa distribuição é representada por meio do teorema de Bayes, conforme a seguinte equação:

$$\pi(\boldsymbol{\theta}|\mathbf{t}) = \frac{L(\mathbf{t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{t})d\boldsymbol{\theta}}, \quad (2.7)$$

onde Θ denota o espaço paramétrico de $\boldsymbol{\theta}$ e o denominador é a constante de normalização que não depende de $\boldsymbol{\theta}$.

Assim, o teorema pode ser reescrito como:

$$\pi(\boldsymbol{\theta}|\mathbf{t}) \propto L(\boldsymbol{\theta}|\mathbf{t})\pi(\boldsymbol{\theta}). \quad (2.8)$$

Por meio do teorema de Bayes, é possível atualizar o conhecimento sobre o parâmetro e quantificar esse aumento de informação, o que desempenha um papel crucial na análise bayesiana, onde toda a inferência é baseada na distribuição *a posteriori*.

Considerando o vetor de parâmetros $\boldsymbol{\theta}$, é necessário calcular a distribuição marginal *a posteriori* para cada parâmetro θ_i , onde $i = 1, 2, \dots, d$, obtida através da integração da densidade conjunta *a posteriori*:

$$\pi(\theta_i|\mathbf{t}) = \int \dots \int \pi(\boldsymbol{\theta}|\mathbf{t}) d\theta_i. \quad (2.9)$$

No entanto, a resolução analítica dessa integral é geralmente impraticável. Uma alternativa é a utilização de métodos numéricos, com destaque para os algoritmos de Monte Carlo via Cadeias de Markov (MCMC), que visam simular um passeio aleatório no espaço dos parâmetros $\boldsymbol{\theta}$, convergindo para uma distribuição estacionária que representa a distribuição marginal *a posteriori* de $\boldsymbol{\theta}$ (Paulino, Turkman e Murteira, 2003).

2.7 Cadeias de Markov

Os métodos MCMC (Monte Carlo via Cadeias de Markov) empregam simulação estocástica, considerando as distribuições condicionais completas *a posteriori* de cada parâmetro para gerar amostras que convergem para a densidade marginal, à medida que o número de interações empregado.

A distribuição condicional completa *a posteriori* do parâmetro θ é denotada por $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{t})$, obtida assumindo que, na densidade *a posteriori* conjunta, os demais parâmetros $\boldsymbol{\theta}_{-i}$ são conhecidos, simplificando a expressão, uma vez que as constantes podem ser desconsideradas.

Quando a expressão da condicional completa tem uma forma conhecida, o método de simulação amostrador de Gibbs é utilizado. Quando a forma da condicional completa é desconhecida, recorre-se ao algoritmo Metropolis-Hastings (Hastings, 1970).

O amostrador de Gibbs é um esquema iterativo de amostragem de uma cadeia de Markov, onde o núcleo de transição é formado pelas condicionais completas. O algoritmo do amostrador de Gibbs segue o processo iterativo a seguir:

1. Atribua um valor inicial para $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \dots, \theta_d^0)$.
2. Obtenha um valor para $\boldsymbol{\theta}_{(1)}$ da seguinte forma:
 - (i) Amostre θ_1^1 de $\pi(\theta_1^1|\theta_2^0, \dots, \theta_d^0, \mathbf{t})$.

- (ii) Amostre θ_2^1 de $\pi(\theta_2^1 | \theta_1^1, \theta_3^0, \dots, \theta_d^0, \mathbf{t})$.
- (iii) Amostre θ_d^1 de $\pi(\theta_d^1 | \theta_1^1, \theta_2^1, \dots, \theta_{d-1}^1, \mathbf{t})$.

Continue até concluir uma iteração do esquema de transição de θ^0 para θ^1 .

3. Repita o passo 2 m vezes, até que a convergência seja alcançada.

O algoritmo de Metropolis-Hastings emprega a ideia de gerar uma amostra de uma distribuição auxiliar, a qual é aceita com uma determinada probabilidade. Esse mecanismo de correção assegura a convergência da cadeia para a distribuição de equilíbrio.

Para iniciar o algoritmo de Metropolis-Hastings, é fornecido um valor inicial para θ , denotado como θ^0 . Agora, suponha que o algoritmo esteja na iteração m e deseja-se obter uma nova observação θ_i^m de $f(x) = \pi(x | \theta_{-i}^m, \mathbf{t})$, onde x representa um valor de θ_j .

Proceda da seguinte maneira:

- (i) Amostre um ponto y de $q(z_m, y)$, onde z_m representa o valor atual de θ_j e $q(z_m, y)$ é uma distribuição proposta apropriada (função de transição apropriada).
- (ii) Amostre um ponto u da distribuição uniforme $U[0, 1]$. Se $u \leq \alpha(z_m, y)$, defina $z_{m+1} = y$; caso contrário, mantenha $z_{m+1} = z_m$, onde $\alpha(z_m, y) = \min\left(1, \frac{f(y)q(z_m|y)}{f(z_m)q(y|z_m)}\right)$.

Repita esses passos até que a convergência da cadeia seja alcançada. A distribuição proposta $q(z, y)$ deve ser facilmente simulada e garantir uma convergência rápida para a distribuição alvo $\pi(\theta_i^m | \theta_{-i}^m, \mathbf{t})$ (CHIB e GREENBERG, 1995). A ideia central por trás dos algoritmos MCMC é que, à medida que o número de iterações aumenta, a cadeia gerada se aproxima da distribuição de equilíbrio, ou seja, da densidade marginal desejada de cada parâmetro. Portanto, é essencial monitorar a convergência das sequências geradas pela cadeia de Markov. Existem vários critérios propostos com esse propósito, e uma opção é utilizar o método proposto por (Gelman e Rubin, 1992).

Esses pesquisadores propuseram um método de simulação baseado em várias cadeias de Markov simuladas a partir de diferentes pontos iniciais. Portanto, o método de Gelman e Rubin é baseado na construção de l cadeias independentes, cada uma começando a partir de valores iniciais distintos. Além disso, para cada parâmetro, as primeiras iterações devem ser descartadas para eliminar o efeito do valor inicial, e um espaçamento entre as iterações sucessivas deve ser considerado para evitar problemas de correlação. Dessa forma, obtém-se uma amostra final $\theta^{(v,s)} = (\theta_1^{(v,s)}, \theta_2^{(v,s)}, \dots, \theta_d^{(v,s)})'$.

Onde v denota a v -ésima cadeia e s denota a s -ésima réplica na v -ésima cadeia, com $s = (1, \dots, m)$ e $v = (1, \dots, l)$.

Para avaliar a convergência utilizando o método de Gelman e Rubin, considere que U representa qualquer um dos componentes do vetor de parâmetro θ . Para cada parâmetro de interesse, o cálculo é realizado da seguinte forma:

$$\bar{u}_v = \frac{1}{m} \sum_{s=1}^m U_{v_s} \quad \text{e} \quad S_v^2 = \frac{1}{m-1} \sum_{s=1}^m (U_{v_s} - \bar{u}_v)^2,$$

Agora, calcula-se a média e a variância, respectivamente, de U para cada cadeia $v = 1, \dots, l$. As estimativas são dadas por:

$$B = \frac{m}{l-1} \sum_{v=1}^l (\bar{u}_v - \bar{u}_{..})^2 \quad \text{e} \quad W = \sum_{v=1}^l \frac{S_v^2}{l},$$

onde $\bar{u}_{v..} = \sum_{v=1}^l \bar{u}_v$. A variância de U pode ser estimada como uma média ponderada de W e B , conforme a seguinte equação

$$\widehat{\sigma}^2 = \frac{m-1}{m}W + \frac{1}{m}B.$$

U segue uma distribuição t de Student com parâmetro de localização $\mu^* = \bar{u}_{v..}$, parâmetro de escala $\sqrt{\widehat{V}} = \sqrt{\widehat{\sigma}^2 + \frac{B}{lm}}$ e graus de liberdade $gl = 2\widehat{V}^2 = \frac{\widehat{\text{var}}}{\widehat{V}}$ em que:

$$\begin{aligned} \widehat{\text{var}}(\widehat{V}) &= \left(\frac{m-1}{m}\right)^2 \frac{1}{l} \widehat{\text{var}}(s_v^2) \\ &+ \left(\frac{l+1}{vs}\right)^2 \frac{2}{l-1} B^2 + 2 \frac{(l-1)(m-1)}{l^2 m} [\widehat{\text{cov}}(s_v^2, \bar{u}_v^2) - 2\bar{u}_{v..} \widehat{\text{cov}}(s_v^2, \bar{u}_v^2)] \end{aligned}$$

As variâncias e as covariâncias são estimadas a partir dos l valores amostrais de s_v^2 , \bar{u}_v , e \bar{u}_v^2 .

A convergência é monitorada calculando o fator de redução de escala, definido como:

$$\sqrt{\widehat{R}} = \frac{\sqrt{\widehat{V}}}{W} \frac{gl}{gl-2}$$

O coeficiente de convergência \widehat{R} é usado para medir o fator de redução que o parâmetro de escala da distribuição t de Student para U sofreria se a simulação continuasse. À medida que n aumenta, \widehat{R} converge para 1. Na prática, assume-se que a convergência ocorre quando \widehat{R} está próximo de 1 e as iterações selecionadas para compor a amostra final são independentes e identicamente distribuídas (Gelman e Rubin, 1992). Portanto, quando o coeficiente potencial de redução de escala apresenta um valor elevado, existem razões para acreditar que um maior número de simulações melhoraria a inferência sobre a distribuição desejada.

A distribuição marginal a *posteriori* de um parâmetro θ_i encapsula toda a informação probabilística relacionada a esse parâmetro. No entanto, às vezes é necessário resumir essa informação em alguns valores numéricos. Um caso simples é a estimativa pontual de θ_i , em que a distribuição marginal a *posteriori* é resumida em um único número, $\hat{\theta}_i$. Também é importante fornecer informações sobre a precisão desse número. As medidas de incerteza mais comuns incluem a variância e o coeficiente de variação para a média a *posteriori*, a medida de informação observada de Fisher para a moda a *posteriori* e a distância entre quartis para a mediana a *posteriori* (Ehlers, 2007).

A própria ideia de estimar parâmetros por ponto conduz, no contexto bayesiano, à consideração dos pontos críticos da distribuição a *posteriori* como estimativas. Uma maneira mais informativa de resumir $\pi(\theta|t)$ do que qualquer estimativa pontual é obtida a partir de uma região no espaço paramétrico Θ que contenha uma parte significativa da massa de probabilidade a *posteriori* (Paulino, Turkman e Murteira, 2003).

Definido por Ehlers (2007) C como um intervalo de credibilidade de $100(1 - \alpha)$ para θ se $P(\theta \in C) \geq 1 - \alpha$. Portanto, quanto menor for o tamanho do intervalo, mais concentrada será a distribuição do parâmetro. Ou seja, o tamanho do intervalo fornece informações sobre a dispersão de θ .

É possível construir uma infinidade de intervalos utilizando a definição acima, mas o de maior interesse é aquele com o menor comprimento possível. Os intervalos de comprimento mínimo são obtidos selecionando os valores de θ com a maior densidade a *posteriori*, conhecidos como intervalos de credibilidade de máxima densidade a *posteriori*, ou intervalos HPD.

Conforme descrito por Ehlers (2007), um intervalo de credibilidade C de $100(1 - \alpha)$ para θ é considerado de máxima densidade a *posteriori* se $C = (\theta \in \Theta : \pi(\theta|t) \geq k(\alpha))$, onde $k(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$.

Com base nessa definição, todos os pontos contidos no intervalo de máxima densidade a *posteriori* (intervalo HPD) terão uma densidade maior do que qualquer ponto fora desse intervalo.

2.7.1 Análise bayesiana (Modelo Weibull)

Considere a função densidade de probabilidade (2.1) da distribuição Weibull e sendo T uma variável aleatória com distribuição Weibull com vetor de parâmetros $\theta = (\alpha, \gamma)^T$. Supondo que os dados consistam de n observações independentes $t_i = \min(T_i; C_i)$ para $i = 1, \dots, n$, a função de verossimilhança, considerando a distribuição Weibull, é dada por:

$$L(\theta) = (\gamma\alpha^{-\gamma})^n \exp \left[- \sum_{i=1}^n \left(\frac{t_i}{\alpha} \right)^\gamma \right] \prod_{i=1}^n t_i^{\gamma-1},$$

para os parâmetros α e γ . Para uma análise bayesiana, assume-se a seguinte densidade conjunta a *priori* dada por:

$$\pi(\alpha, \gamma) \propto \pi(\alpha) \times \pi(\gamma),$$

em que $\alpha \sim \Gamma(a_1, b_1)$ e $\gamma \sim \Gamma(a_2, b_2)$, sendo $\Gamma(a_i, b_i)$ a distribuição gama com média $\Gamma(a_i, b_i)$, variância a_i/b_i^2 e função densidade de probabilidade dada por:

$$f(v; a_i, b_i) = \frac{b_i^{a_i} v^{a_i-1} \exp(-vb_i)}{\Gamma(a_i)},$$

em que $v > 0$, $a_i > 0$ e $b_i > 0$. Em que a_i e b_i são números reais positivos fixos e desconhecidos, denominados hiperparâmetros. Assumindo independência entre os parâmetros α e γ , a distribuição conjunta a *posteriori* para α e γ é dada por:

$$\pi(\alpha, \gamma|t) \propto (\gamma\alpha^{-\gamma})^n \exp \left[- \sum_{i=1}^n \left(\frac{t_i}{\alpha} \right)^\gamma \right] \prod_{i=1}^n t_i^{\gamma-1} \times \pi(\alpha, \gamma) \quad (2.10)$$

As densidades marginais a *posteriori* dos parâmetros α e γ não são facilmente obtidas; isso porque a integração da densidade a *posteriori* conjunta (2.10) é muito complexa. Uma alternativa para situações desse tipo é o uso do algoritmo Metropolis-Hasting, descrito na subseção 2.6.

O algoritmo Metropolis-Hasting permite simular observações de distribuições a partir das densidades condicionais a *posteriori*. Assim, para o uso de tal algoritmo consideram-se as seguintes condicionais a *posteriori* completas para os parâmetros α e γ :

$$\pi(\alpha|t, \gamma) \propto \alpha^{-n\gamma} \exp \left[- \sum_{i=1}^n \left(\frac{t_i}{\alpha} \right)^\gamma \right] \times \pi(\alpha)$$

e

$$\pi(\gamma|t, \alpha) \propto \gamma^n \exp \left[- \sum_{i=1}^n \left(\frac{t_i}{\alpha} \right)^\gamma \right] \prod_{i=1}^n t_i^{\gamma-1} \times \pi(\gamma)$$

Para as análises descritivas para os parâmetros do modelo Weibull será utilizado o *software* estatístico R , pacote BOA (*Bayesian Output Analysis*).

2.7.2 Análise bayesiana (Modelo Log-Weibull)

Considere a função densidade de probabilidade (2.2) da distribuição Log-Weibull e sendo Y uma variável aleatória com distribuição Log-Weibull com vetor de parâmetros $\theta = (\sigma, \beta^T)^T$. Supondo que os dados consistam de n observações independentes $y_i = \min\{\log(t_i), \log(c_i)\}$ para $i = 1, \dots, n$, a função de verossimilhança, considerando a distribuição Log-Weibull, é dada por:

$$L(\theta) = \left(\frac{1}{\sigma}\right)^n \exp \left[\sum_{i=1}^n \left(\frac{y_i - X_i^T \beta}{\sigma} \right) - \sum_{i=1}^n \exp \left(\frac{y_i - X_i^T \beta}{\sigma} \right) \right]$$

para os parâmetros σ e β . Para uma análise bayesiana, assume-se a seguinte densidade conjunta a priori dada por:

$$\pi(\sigma, \beta) \propto \pi(\sigma) \times \pi(\beta)$$

em que $\sigma \sim \Gamma(a_1, b_1)$ e $\beta_i \sim N(\mu_s, \sigma_s^2)$, sendo $\Gamma(a_i, b_i)$ a distribuição gama com média $\Gamma(a_i, b_i)$, variância a_i/b_i^2 e função densidade de probabilidade dada por:

$$f(v; a_i, b_i) = \frac{b_i^{a_i} v^{a_i-1} \exp(-vb_i)}{\Gamma(a_i)},$$

em que $v > 0$, $a_i > 0$ e $b_i > 0$, $N(\mu_s, \sigma_s^2)$ denota a distribuição normal com média μ_s , variância σ_s^2 e função densidade de probabilidade dada por:

$$f(x; \mu_s, \sigma_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp \left[-\frac{(x - \mu_s)^2}{2\sigma_s^2} \right],$$

em que $x, \mu \in \mathfrak{R}$ e $\sigma^2 > 0$. Todos os hiperparâmetros são especificados. Assumindo independência entre os parâmetros σ e β , a distribuição conjunta a posteriori para σ e β é dada por:

$$\pi(\sigma, \beta|y) \propto \left(\frac{1}{\sigma}\right)^n \exp \left[\sum_{i=1}^n \left(\frac{y_i - X_i^T \beta}{\sigma} \right) - \sum_{i=1}^n \exp \left(\frac{y_i - X_i^T \beta}{\sigma} \right) \right] \times \pi(\sigma, \beta) \quad (2.11)$$

As densidades marginais a posteriori dos parâmetros σ e β não são facilmente obtidas; isso porque a integração da densidade a posteriori conjunta (2.11) é muito complexa. Uma alternativa para situações desse tipo é o uso do algoritmo Metropolis-Hasting, descrito na subseção 2.6.

O algoritmo Metropolis-Hasting permite simular observações de distribuições a partir das densidades condicionais a posteriori. Assim, para o uso de tal algoritmo consideram-se as seguintes condicionais a posteriori completas para os parâmetros σ e β :

$$\pi(\sigma|y, \beta) \propto \left(\frac{1}{\sigma}\right)^n \exp \left[\sum_{i=1}^n \left(\frac{y_i - X_i^T \beta}{\sigma} \right) - \sum_{i=1}^n \exp \left(\frac{y_i - X_i^T \beta}{\sigma} \right) \right] \times \pi(\sigma)$$

e

$$\pi(\beta|y, \sigma) \propto \exp \left[\sum_{i=1}^n \left(\frac{y_i - X_i^T \beta}{\sigma} \right) - \sum_{i=1}^n \exp \left(\frac{y_i - X_i^T \beta}{\sigma} \right) \right] \times \pi(\beta)$$

Para as análises descritivas para os parâmetros do modelo Weibull, será utilizado o *software* estatístico R (R Development Core Team - version 4.2.3, 2023), com o pacote *Bayesian Optimization Assistant (Boa)* (Smith, 2007; Smith, 2016). Para o estimador de Kaplan-Meier, será utilizado o pacote *Survival Analysis* (Therneau e Grambsch, 2000; Therneau, 2023)

3 MATERIAL E MÉTODOS

3.1 Dados

Nesta seção são abordados os materiais e métodos que foram utilizados no desenvolvimento do estudo. Os dados utilizados no desenvolvimento do trabalho provêm da base de dados construída pela Secretaria de Estado da Saúde de Mato Grosso desde o início da pandemia, que contempla a notificação de casos suspeitos e confirmados de COVID-19 do estado de Mato Grosso, assim como a sua evolução clínica e desfecho, hoje disponibilizada pela Secretaria de Saúde de Mato Grosso (Grosso, 2020).

3.2 Descrição de Variáveis

A variável resposta é o tempo da hospitalização do paciente até o óbito por COVID-19, o período compreendido foi de fevereiro a julho de 2020, do município de Colíder-MT.

As variáveis consideradas neste estudo são as seguintes:

- i) t_i : tempo entre a internação e a morte por COVID-19 (dias);
- ii) $cens_i$: Indicador de censura (0 = censurado, 1 = falha);
- iii) x_{i1} : raça/cor (1 = branco, 0 = não branco);
- iv) x_{i2} : Sexo (0 = feminino, 1 = masculino);
- v) x_{i3} : Hipertensão Arterial Sistêmica - HAS (0 = Não, 1 = Sim);
- vi) x_{i4} : Pneumopatia (0 = Não, 1 = Sim);
- vii) x_{i5} : Diabetes (0 = Não, 1 = Sim);
- viii) x_{i6} : Nefropatia (0 = Não, 1 = Sim);
- ix) x_{i7} : Cardiopatia (0 = Não, 1 = Sim);
- x) x_{i8} : Obesidade (0 = Não, 1 = Sim);
- xi) x_{i9} : Ventilação Mecânica (0 = Não, 1 = Sim);
- xii) x_{i10} : Idade.

3.3 Análise Estatística

O banco de dados é composto por 48 observações. Para entender o comportamento dos dados foi feita uma análise explanatória. O comportamento da forma empírica da função risco se dará por meio da construção do gráfico do tempo total em teste (curva TTT). O risco sugerido pela curva TTT, foi utilizado para confirmar a escolha do modelo Weibull para a análise dos dados. O modelo Weibull foi ajustado por meio do método Bayesiano primeiramente apenas aos tempos de internação. A curva de sobrevivência ajustada obtida a partir do modelo Weibull foi comparada com a sobrevivência estimada por meio do estimador de Kaplan-Meier.

O método de seleção de variáveis *stepwise* foi empregado para identificar as mais relevantes. Posteriormente, o modelo de regressão Log-Weibull foi ajustado por meio do método Bayesiano, para estas variáveis selecionadas.

As análises foram implementadas no *software* R (R Development Core Team - version 4.2.3, 2023), com o pacote *Bayesian Optimization Assistant (Boa)* (Smith, 2007; Smith, 2016). Para o estimador de Kaplan-Meier, será utilizado o pacote *Survival Analysis* (Therneau e Grambsch, 2000; Therneau, 2023)

4 RESULTADOS E DISCUSSÕES

Nesta fase, foi implementada a metodologia descrita na Seção 3 para analisar os dados relativos à COVID-19 no município de Colíder-MT. O conjunto de dados abrange 48 observações, e as variáveis de interesse são: t_i : tempo entre a internação e a morte por COVID-19 (dias), x_{i1} : raça/cor, x_{i2} : Sexo, x_{i3} : Hipertensão Arterial Sistêmica, x_{i4} : Pneumopatia, x_{i5} : Diabetes, x_{i6} : Nefropatia, x_{i7} : Cardiopatia, x_{i8} : Obesidade, x_{i9} : Ventilação Mecânica e x_{i10} : Idade do paciente internado.

Na Tabela 1, estão apresentados os resultados da análise exploratória desses dados. O tempo de óbito mais longo registrado foi de 44 dias, enquanto o mais curto foi de 1 dia. A média do tempo de óbito foi aproximadamente 11 dias. Em relação à idade dos pacientes internados, a idade mais avançada registrada foi de 91 anos, com o paciente permanecendo internado por 1 dia. Por outro lado, a idade mais jovem registrada foi de 36 anos, com o paciente internado por 3 dias.

Tabela 1 – Estatísticas descritivas para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.

Estatística	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
Tempo (dias)	1	4,75	8	10,53	12	44
Idade (anos)	36	55	67,50	66	79,25	91

Na Figura 2 é apresentada a curva TTT, definida na subseção 2.1.1, para os dados em estudo. A Figura 2 indica que a função de risco é monótona crescente. Assim sendo, um possível modelo para ajuste dos dados é o modelo Weibull.

Considerando a análise bayesiana, foram adotadas distribuições *a priori* pouco informativas: $\alpha \sim \Gamma(0, 01; 0, 01)$ e $\gamma \sim \Gamma(0, 01; 0, 01)$.

Assim, foram geradas duas cadeias paralelas e independentes, cada uma com 200.000 iterações para cada parâmetro, utilizando o algoritmo de Metropolis-Hastings. As primeiras 20.000 iterações foram descartadas para eliminar o efeito dos valores iniciais. Além disso, foi considerado um espaçamento de tamanho 10 para evitar problemas de correlação, resultando em uma amostra final de 18.000 observações.

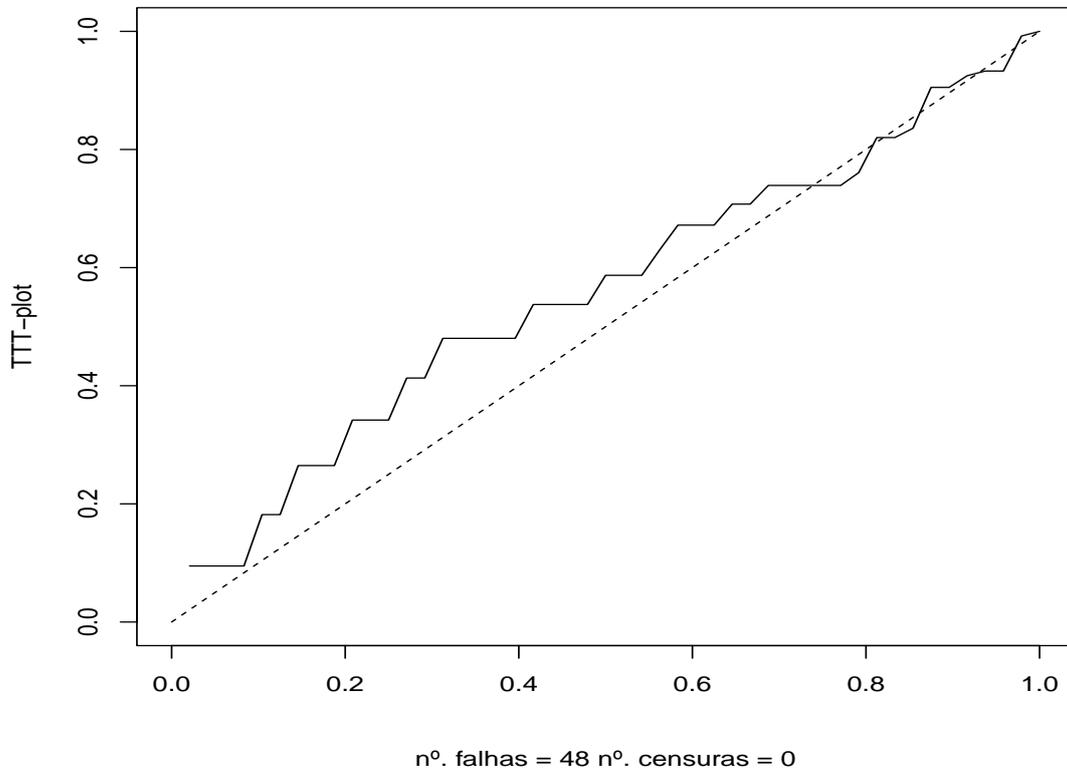
Para monitorar a convergência das amostras, foi utilizado o fator de redução da escala \hat{R} , proposto por Gelman e Rubin (1992). A Tabela 2 apresenta os sumários *a posteriori* para os parâmetros do modelo Weibull, incluindo os valores do fator de redução da escala.

Tabela 2 – Sumário *a posteriori* dos parâmetros do modelo Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.

Parâmetro	Média	Desvio Padrão	HPD (95%)	\hat{R}
α	11,3453	0,0080	(11, 3296; 11, 3609)	0,9901
γ	1,2397	0,0026	(1, 2345; 1, 2452)	0,9816

Observa-se que, em todos os casos, os valores de \hat{R} estão próximos de 1, o que indica a convergência das cadeias. Além disso, na Figura 3, são apresentadas as densidades marginais *a posteriori*, aproximadas por meio de histogramas, com base nas 18.000 observações amostrais. Essas densidades marginais sugerem que a distribuição *a posteriori* apresenta uma tendência à simetria para os parâmetros.

Figura 2 – TTT-Plot para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.



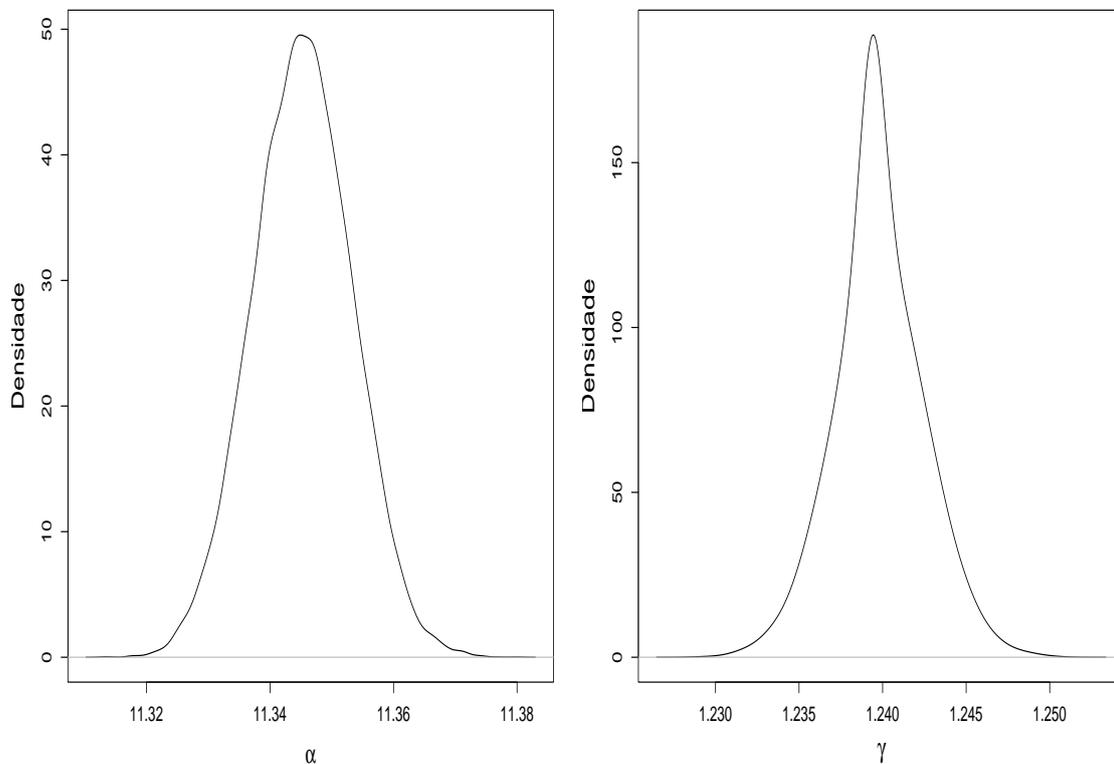
Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

Na Figura 4, são mostradas as cadeias geradas pelo método MCMC para os parâmetros α e γ do modelo Weibull. Observa-se claramente uma rápida convergência dos parâmetros em torno do valor real (média).

Na Tabela 2 são apresentadas as estimativas para o modelo Weibull ajustado apenas aos tempos de internação, em que o parâmetro de forma $\gamma = 1,2397 > 1$ indica, portanto, uma função de risco crescente, isto pode ser confirmado pela Figura 5c, que mostra que o risco estimado pelo modelo Weibull de morte para pacientes internados com COVID-19, cresce com o passar do tempo. Pode-se observar ainda, indícios de que a distribuição se ajusta bem aos dados, de acordo com a Figura 5, em (a) tem-se a densidade de probabilidade estimadas para o modelo Weibull e em (b) o gráfico de Kaplan-Meier com a curva de sobrevivência ajustada obtida a partir do modelo Weibull.

Considerando os indícios de que a distribuição Weibull se ajusta adequadamente aos dados analisados, optou-se por utilizar o modelo de regressão Log-Weibull Para avaliar se as covariáveis tem efeito sobre o tempo e como estes efeitos se dão (aumentam ou diminuem os tempos). Nesse contexto, foi utilizado o método de seleção de variáveis *stepwise* para identificar as mais relevantes no modelo. Os resultados do modelo de regressão Log-Weibull para os dados de tempo até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT, após a

Figura 3 – Densidades marginais *a posteriori* para os parâmetros do modelo Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.



Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

aplicação do método de seleção de variáveis *stepwise*, o subconjunto de variáveis significativas são: x_{i4} (Pneumopatia), x_{i9} (Ventilação Mecânica) e x_{i10} (Idade do paciente internado).

Portanto, o modelo final ajustado de regressão Log-Weibull é o seguinte:

$$y_i = \beta_0 + \beta_4 x_{i4} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \sigma z_i,$$

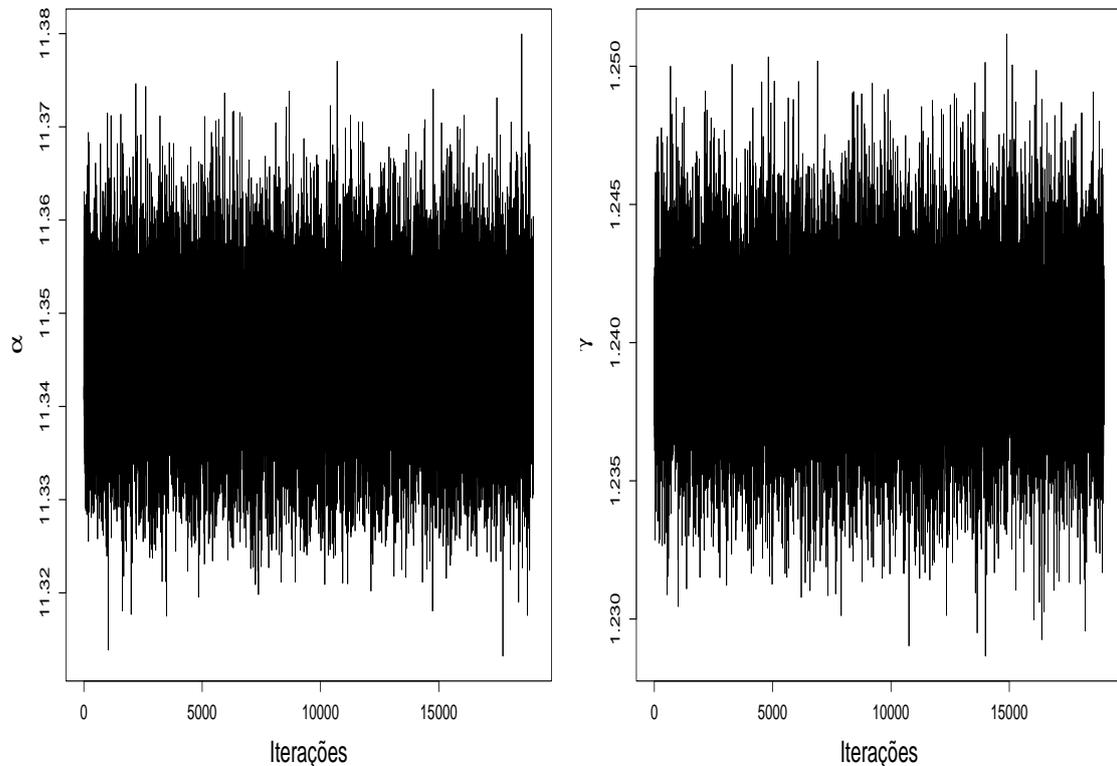
em que os erros z_1, \dots, z_{48} são variáveis aleatórias independentes com função densidade de probabilidade apresentada na equação 2.5.

Considerando a análise bayesiana para o modelo de regressão Log-Weibull, foram adotadas distribuições *a priori* pouco informativas: $\sigma \sim \Gamma(0,01; 0,01)$ e $\beta_s \sim N(0; 100)$.

Assim, foram geradas duas cadeias paralelas e independentes, cada uma com 100.000 iterações para cada parâmetro, utilizando o algoritmo de Metropolis-Hastings. As primeiras 10.000 iterações foram descartadas para eliminar o efeito dos valores iniciais. Além disso, foi considerado um espaçamento de tamanho 10 para evitar problemas de correlação, resultando em uma amostra final de 9.000 observações.

Para monitorar a convergência das amostras, foi utilizado o fator de redução da escala \hat{R} , proposto por Gelman e Rubin (1992). A Tabela 3 apresenta os sumários *a posteriori* para os parâmetros do modelo de regressão Log-Weibull, incluindo os valores do fator de redução da escala.

Figura 4 – Cadeias geradas pelo método MCMC para os parâmetros α e γ referentes ao modelo Weibull.



Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

Tabela 3 – Sumário *a posteriori* dos parâmetros do modelo de regressão Log-Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.

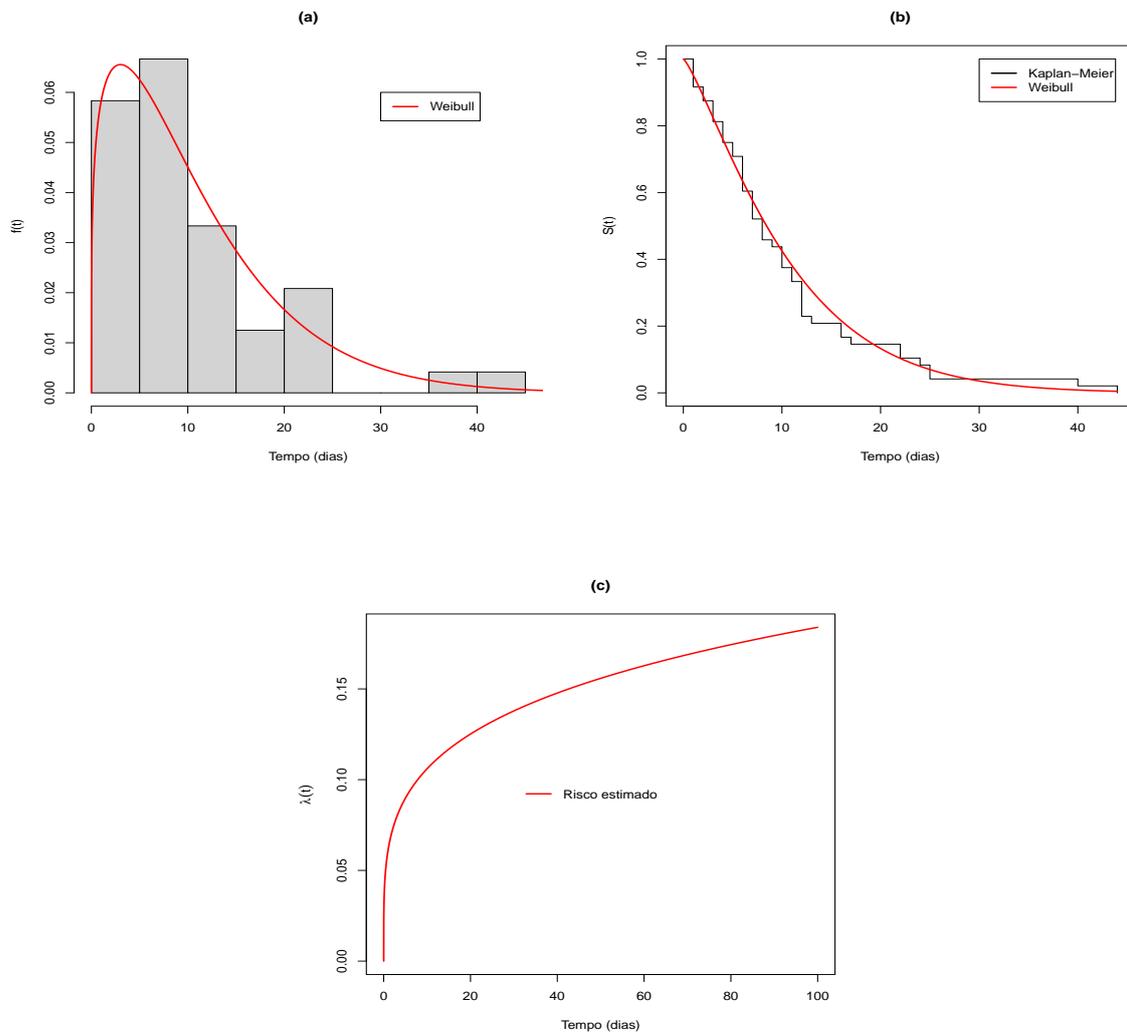
Parâmetro	Média	Desvio Padrão	HPD (95%)	\hat{R}
σ	0,6059	0,0099	(11, 3296; 11, 3609)	1,0315
β_0 : Intercepto	4,1123	0,3160	(3, 5108; 4, 7399)	1,0133
β_4 : Pneumopatia	-1,0057	0,1015	(-1, 2157; -0, 8071)	0,9942
β_9 : Vent. Mec.	0,5087	0,2007	(0, 1209; 0, 8984)	1,0317
β_{10} : Idade	-0,0264	0,0100	(-0, 0461; -0, 0070)	1,0221

Observa-se que, em todos os casos, os valores de \hat{R} estão próximos de 1, o que indica a convergência das cadeias. Além disso, na Figura 6, são apresentadas as densidades marginais *a posteriori*, aproximadas por meio de histogramas, com base nas 9.000 observações amostrais. Essas densidades marginais sugerem que a distribuição *a posteriori* apresenta uma tendência à simetria para os parâmetros.

Na Figura 7, são mostradas as cadeias geradas pelo método MCMC para os parâmetros σ e β_s do modelo de regressão Log-Weibull. Observa-se claramente uma rápida convergência dos parâmetros em torno do valor real (média).

Na Tabela 3, podem ser vistos os valores para as médias *a posteriori* para os parâmetros do

Figura 5 – (a) Densidade estimada do modelo Weibull para os dados de tempo até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT. (b) Estimativas da função de sobrevivência segundo Kaplan-Meier e segundo o modelo Weibull, para os dados de tempo até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT. (c) Risco estimado pelo modelo Weibull, para os dados em estudo.

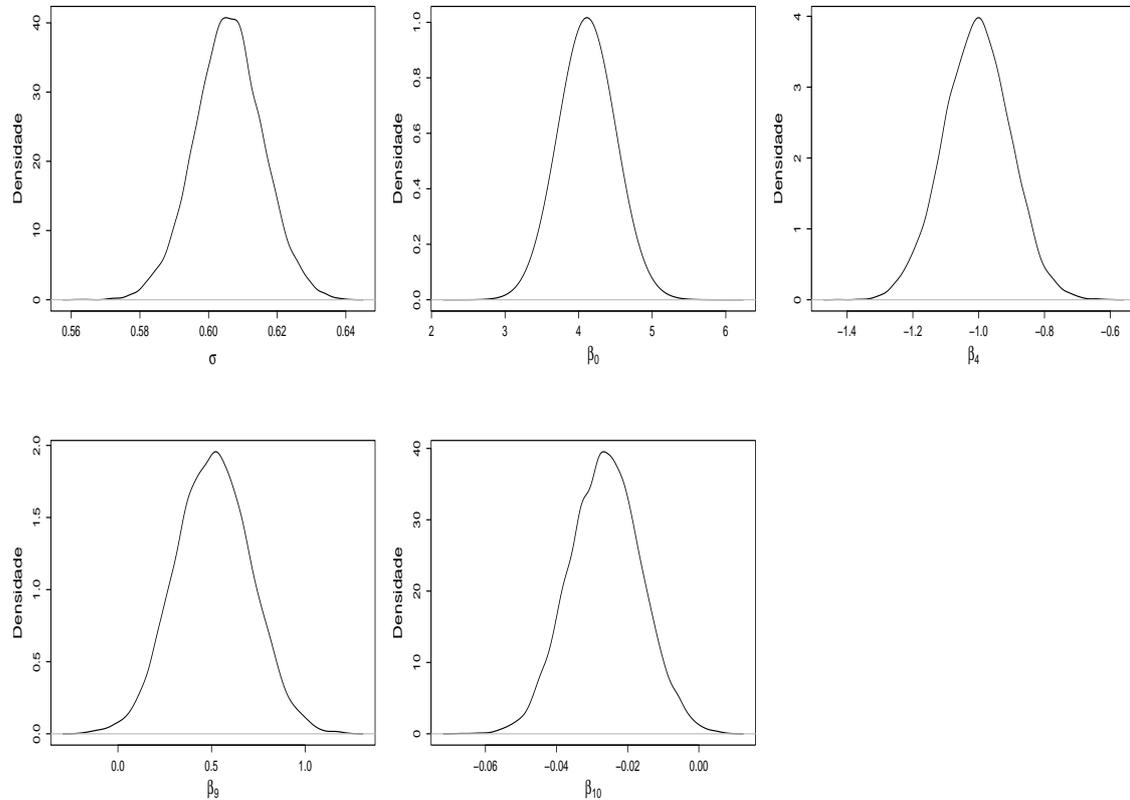


Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

modelo final. Para o cálculo do risco será considerado $\exp(-\beta)$. Com base nisso, pode-se interpretar as estimativas do modelo final da seguinte forma:

- Para cada ano de idade o risco de morte por COVID-19 aumenta aproximadamente 2,7%, esse dado ressalta a vulnerabilidade dos indivíduos mais velhos diante da pandemia [$\exp(0,0264) = 1,0267$];
- Pacientes com pneumopatia enfrentam um risco aproximadamente 173% maior de morrer por COVID-19 em comparação com aqueles sem pneumopatia [$\exp(1,0057) = 2,7338$], esse dado

Figura 6 – Densidades marginais a *posteriori* para os parâmetros do modelo de regressão Log-Weibull para os tempos até a morte dos pacientes internados em decorrência de complicações da COVID-19, no município de Colíder-MT.

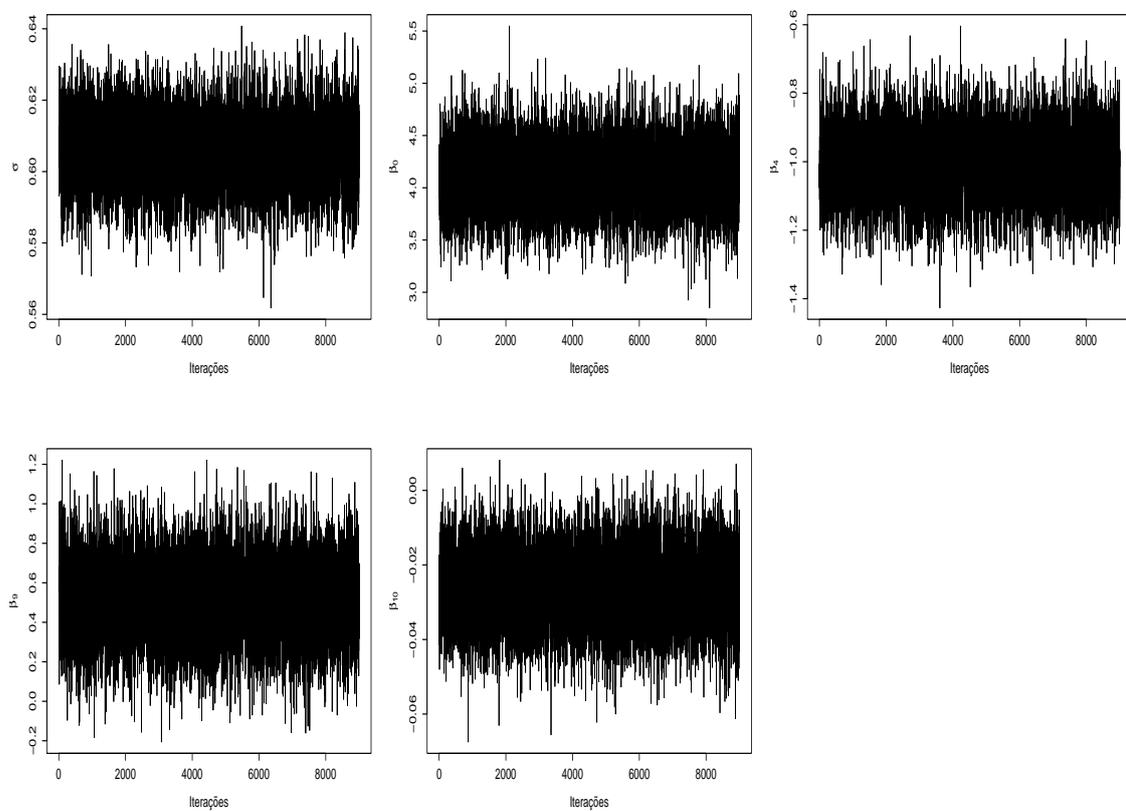


Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

ressalta a gravidade da situação para pessoas com condições pulmonares pré-existentes quando infectadas pelo vírus;

- Pacientes que foram colocados sob ventilação mecânica têm um risco de morte por COVID-19 diminuído em aproximadamente 60% em comparação com aqueles sem ventilação mecânica [$1 - \exp(-0,5087) = 0,3987$], isso destaca a importância crítica do suporte respiratório fornecido pela ventilação mecânica no tratamento de casos graves da doença;
- O risco de morte por COVID-19 aumenta em aproximadamente 65% à medida que aumenta o tempo de internação ($\gamma = 1/\sigma = 1/0,6059 = 1,6504$), esse dado enfatiza a importância de uma gestão eficaz dos casos hospitalizados, especialmente para pacientes que requerem cuidados prolongados.

Figura 7 – Cadeias geradas pelo método MCMC para os parâmetros σ e β_s referentes ao modelo de regressão Log-Weibull.



Fonte: Elaborado pela própria autora utilizando os dados coletados do Painel COVID-19 da Secretaria de Estado de Saúde de Mato Grosso, 2020

5 CONSIDERAÇÕES FINAIS

A técnica estatística de análise de sobrevivência mostrou-se altamente eficaz neste estudo, conseguindo capturar de forma satisfatória o comportamento dos tempos até o óbito e os fatores associados a esse desfecho.

O modelo Weibull revelou-se eficaz para ajustar os dados de COVID-19 no município de Colíder-MT, fornecidos pela Secretaria de Estado da Saúde de Mato Grosso, focalizando especificamente no tempo de internação até o falecimento do paciente, uma vez que o risco de morte tende a aumentar com a prolongação da internação.

Para identificar as variáveis mais relevantes no modelo, empregou-se o método de seleção de variáveis *stepwise*. Verificou-se que pneumopatia, ventilação mecânica e idade foram os principais determinantes, cujos efeitos sobre o tempo foram avaliados pelo modelo de regressão Log-Weibull.

Das análises conduzidas, observou-se que para cada ano de idade, o risco de morte por COVID-19 aumenta aproximadamente 2,7%. Pacientes com pneumopatia enfrentam um risco cerca de 173% maior de falecer devido à COVID-19 em comparação com aqueles sem essa comorbidade, enquanto os pacientes submetidos à ventilação mecânica apresentam um risco de morte reduzido em cerca de 60% em comparação com aqueles sem esse suporte.

Por fim, espera-se que os resultados deste estudo auxiliem pesquisadores, gestores e profissionais da saúde na compreensão dessa epidemia e na formulação de medidas e políticas de saúde voltadas para a população.

REFERÊNCIAS

- AARSET, M. V. How to identify bathtub hazard rate. *Transactions on Reliability, New York*, v. 36, p. 106–108, 1987.
- BAILEY. Weibull model for pinus radiata diameter distribution. *Statistics in Forestry Research Proceedings of the meeting of IUFRO subject group*, v. 6, n. 2, p. 51–59, 1973.
- BATISTA, J. L. F. *A função Weibull como modelo para a distribuição de diâmetros de espécies arbóreas tropicais*. 116 p. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo, Piracicaba - SP, 1989.
- CHATE, R. C. et al. Apresentação tomográfica da infecção pulmonar na covid-19: experiência brasileira inicial. *Jornal Brasileiro de Pneumologia*, v. 46, n. 2, p. 1–4, 2020.
- CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. *The American Statistician, Washington*, v. 49, p. 326–335, 1995.
- COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. [S.l.]: Blucher, 2006.
- COLÍDER, P. M. de. *Colíder*. 2020. Disponível em: <<https://www.colider.mt.gov.br/>>.
- COX, D. R.; OAKES, D. *Analysis of Survival Data*. [S.l.]: New York: Chapman and Hall, 1984. 201 p.
- DUARTE, M. de Q. et al. Covid-19 e os impactos na saúde mental. *Ciência e Saúde Coletiva*, v. 25, n. 9, p. 3401–3411, 2020.
- EHLERS, R. S. *Introdução a Inferência Bayesiana*. [S.l.: s.n.], 2007. 68 p.
- GALVÃO, M. H. R.; RONCALLI, A. G. Fatores associados a maior risco de ocorrência de óbito por covid-19: análise de sobrevivência com base em casos confirmados. *Rev bras epidemiol [Internet]*, 2020.
- GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science, Hayward*, v. 7, p. 457–472, 1992.
- GROSSO, S. de Estado de Saúde de M. *Painel Interativo COVID-19*. 2020. Disponível em: <<http://www.saude.mt.gov.br/painelcovidmt2/>>.
- HASTINGS, R. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika, London*, v. 57, n. 2, p. 97–109, 1970.
- KALBFLEISCH ohn D.; PRENTICE, R. L. *The statistical analysis of failure time data*. [S.l.]: New York: John Wiley, 2002. 439 p.
- KAPLAN, E.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 557–481, 1958.
- LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. [S.l.]: 2nd ed. New York: John Wiley, 2003. 230 p.
- MUDHOLKAR, G. S.; SRIVASTAVA, D. K.; KOLLIA, G. D. A generalization of the weibull distribution with application to the analysis of survival data. *Journal of American Statistical Association*, v. 91, p. 1575–1583, 1996.
- OLIVEIRA, W. K. de et al. Como o brasil pode deter a covid-19. *Epidemiologia e Serviços de Saúde*, v. 29, n. 2, p. 1–8, 2020.

PASCOA, M. A. R. *Estudo de fatores que influenciaram no tempo até o óbito de pacientes com COVID-19 em Mato Grosso no ano de 2020*. Cuiabá - MT: [s.n.], 2021.

PASCOA, M. A. R. de. *Extensões da distribuição gama generalizada: propriedades e aplicações*. 25 p. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo, Piracicaba - SP, 2012. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/11/11134/tde-30052012-091940/pt-br.php>>.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística bayesiana*. [S.l.]: Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

R Development Core Team - version 4.2.3. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org>>.

SILVA, G. O. *Modelos de regressão aplicados quando a função de taxa de falha não segue um padrão monótono, especialmente em cenários que envolvem o modelo probabilístico da distribuição beta Weibull modificada*. 207 p. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo, Piracicaba - SP, 2009.

SMITH, B. J. boa: An r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, v. 21, n. 11, p. 1–37, 2007.

SMITH, B. J. *Boa: Bayesian Optimization Assistant*. [S.l.], 2016. R package version 1.1.8-2. Disponível em: <<https://cran.r-project.org/web/packages/boa/>>.

THERNEAU, T. M. *A Package for Survival Analysis in R*. [S.l.], 2023. R package version 3.5-7. Disponível em: <<https://CRAN.R-project.org/package=survival>>.

THERNEAU, T. M.; GRAMBSCH, P. M. *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000. ISBN 0-387-98784-3.

TOMAZELLA, V. L. D. *Uso De Métodos Bayesianos Precisos Em Modelos De Regressão Com Dados De Sobrevivência*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação - USP, São Carlos-SP, 1994. Disponível em: <<https://doi.org/10.11606/D.55.2018.tde-05072018-084538>>.