Universidade de Brasília
Departamento de Estatística

Modelagem do tempo de internação até a morte de pacientes com covid-19 no sistema público de saúde do Distrito Federal via modelo de regressão Log-Normal e Burr-XII

Carolyne Soares de Brito

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Carolyne Soares de Brito

Modelagem do tempo de internação até a morte de pacientes com covid-19 no sistema público de saúde do Distrito Federal via modelo de regressão Log-Normal e Burr-XII

Orientadora: Prof[a]. Juliana Betini Fachini Gomes

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Agradecimentos

Em primeiro lugar, expresso minha gratidão a Deus que me deu a certeza do curso a ser feito e me sustentou durante todo o seu trajeto.

Agradeço aos meus pais pela confiança e apoio durante estes anos de curso, celebrando minhas vitórias e dando suporte nas dificuldades.

Agradeço ao Gabriel, que se tornou mais que uma motivação no curso e na minha vida pessoal, sempre oferecendo incentivo incondicional no meu percurso dentro e fora da universidade.

Agradeço aos meus amigos, como Phillippi, Artur, Maria Clara, Camila e tantos outros, com os quais criei inúmeras memórias e infindáveis chamadas de vídeo que estreitaram laços que vão muito além do vínculo acadêmico.

Agradeço aos professores que tive contato durante a graduação, especialmente as professoras Maria Teresa e Ana Maria Nogales, que expressam seu amor e comprometimento pelo ensino em cada aula e instigam à curiosidade pela profundidade desta área do conhecimento.

Agradeço, por fim, à minha orientadora Juliana Betini, com a qual tive o imenso prazer de cursar três disciplinas na graduação e em cada uma delas fazia transparecer sua dedicação e apreço pela estatística e suas potencialidades. Obrigada por me apoiar, acalmar, inspirar e guiar durante este trabalho.

Resumo

O vírus da covid-19 transformou o Brasil e o mundo com sua velocidade de transmissão e gravidade de casos. Medidas de contenção do vírus, intervenções medicamentosas e até vacinas, que foram disponibilizadas em larga escala no Brasil em um certo período após o primeiro contato do vírus no país, mostraram-se eficazes. Entretanto, mesmo com estas medidas, muitos brasileiros vieram a ser infectados, havendo alta procura de auxílio hospitalar e consequentemente lotação dos mesmos e diversos óbitos pela doença. Dessa forma, o presente estudo tem como objetivo identificar fatores que influenciaram o tempo de hospitalização até a morte de indivíduos que adentraram o sistema público de saúde com causa principal de covid-19, no Distrito Federal de 2020 a 2023, por meio da técnica de análise de sobrevivência. Para isso, fez-se o estudo de distribuições de probabilidade, sobretudo com riscos unimodais devido ao comportamento visto no dados, para definir a distribuição que melhor se adeque à variável tempo. Além disso, considerou-se variáveis explicativas relacionadas aos atributos dos pacientes de forma a serem propostos modelos de regressão provenientes da distribuição selecionada. Os modelos finais, ajustados por meio das distribuições Log-Normal e Burr-XII, obtiveram as variáveis significativas "Idade" do paciente, "Sexo" do paciente, "Período" de internação, "Valor Total" gasto e interação entre as duas últimas. Esses resultados mostram que quanto mais velho for o paciente, menor é sua probabilidade de sobrevivência, assim como homens possuem probabilidade de sobrevivência inferior às mulheres no cenário de hospitalização por covid-19. Ademais, o coeficiente da interação entre o "Período"e o logaritmo do "Valor Total"é positivo, indicando que quanto maior o gasto com o paciente no período de 2022 e 2023, maior a probabilidade de sobreviver do que pacientes nos anos de 2020 e 2021. Uma vez ajustados os modelos, que tiveram seus parâmetros estimados pelo método de máxima verossimilhança, verificou-se o ajuste dos modelos por meio dos resíduos de Cox-Snell.

Palavras-chave: Análise de sobrevivência; Distribuição Log-Normal; Distribuição Burr-XII; Modelos de regressão paramétrico; Covid-19.

Abstract

The covid-19 virus has transformed Brazil and the world with its speed of transmission and severity of cases. Measures to contain the virus, medical interventions, and even vaccines, which became widely available in Brazil a certain period after the virus's first contact in the country, shown to be effective. However, even with this meaures, many brazilians were infected, leading to a high demand for hospital assistance, resulting in the overcrowding of healthcare facilities and numerous deaths from the disease. Therefore, the present study aims to identify factors that influenced the time of hospitalization until death for individuals who entered the public health system with covid-19 as the primary cause of covid-19, in Distrito Federal from 2020 to 2023, using survival analysis techniques. To this end, a study of probability distributions was conducted, especially with unimodal risks due to the observed behavior in the data, to define the distribution that best fits the time variable. Additionally, explanatory variables related to patient attributes were considered to propose regression models derived from the selected distribution. The final models, adjusted using Log-Normal and Burr-XII distributions, yielded significant covariates including the patient's "Age", patient's "Gender", Hospitalization "Period", "Total Value"spent on the patient and the interaction between the last two variables. These results demonstrate that as the patient's age increases, their probability of survival decreases, and, likewise, men have a lower survival probability than women in the context of covid-19 hospitalization. Furthermore, the coefficient of interaction between the "Period"and the logarithm of the "Total Amount"is positive, indicating that the higher the expenditure on the patient during the period of 2022 and 2023, the greater the probability of survival compared to patients in the years 2020 and 2021. Once the models were adjusted, with their parameters estimated using the maximum likelihood method, the fit of the models was assessed through the Cox-Snell residuals.

Keywords: Survival analysis; Log-Normal distribution; Burr-XII distribution; Parametric regression models; Covid-19.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Quando uma doença deixa de afetar apenas uma região e passa a ser disseminada mundialmente com transmissões intercontinentais, sendo capaz de se replicar em seres humanos, sua definição se altera de epidemia para pandemia. Com alertas e esforços gerais para sua contenção e cura, as pandemias vêm atreladas com mortes e sequelas, muitas vezes irreparáveis. O avanço da medicina ainda não impossibilita o aparecimento de doenças a níveis pandêmicos, porém, contribui para diminuição de sua fatalidade e pode proporcionar melhor qualidade de atendimento aos acometidos hospitalizados.

A primeira vez que a Organização Mundial da Saúde (OMS) decretou estado de pandemia no século XXI foi em 2009 em decorrência do então novo tipo de influenza, a gripe H1N1 (SCALERA; MOSSAD, 2009). Observou-se uma doença chegar a níveis pandêmicos novamente apenas em 2020, com a pandemia da covid-19 sendo a segunda do século (MAI; PINTO; FERRI, 2020). Mesmo em meio a protocolos de distanciamento e pesquisas de medicamentos e vacinas, o vírus espalhou-se rapidamente a níveis globais, colocando diversas pessoas em risco.

Com primeiros casos relatados na China, a covid-19 alcançou outros continentes antes de atingir o território Sul-Americano. Uma vez dentro do Brasil, com primeiro caso registrado em Fevereiro de 2020, o aumento de casos gerou uma procura médica elevada, com diversos casos graves que necessitaram de auxílio hospitalar, com medicamentos e internação, o que acarretou superlotação do sistema de saúde em diversas cidades do país (PRADO, 2020). Por ter ocorrido um número expressivo de óbitos ocasionados diretamente pela covid-19 ou por complicações desta, as taxas de mortalidade pela doença foram consideráveis em todo território brasileiro. O estudo de Silva, Jardim e Lotufo (2021), que levou em conta o primeiro ano da pandemia, mostra que a taxa de mortalidade, padronizada por idade, do Distrito Federal é a maior entre as capitais do Centro-Oeste, estando entre as maiores de todas as capitais do país.

Em meio às altas taxas de transmissão do vírus e do acometimento de muitos brasileiros, o Sistema Universal de Saúde (SUS) desempenhou um papel fundamental no combate, prevenção e recuperação da covid-19. Sendo de acesso público, pessoas de todo país tiveram acesso a hospitalizações, a medida em que havia-se vagas, para auxílio medicamentoso e internação de indivíduos acometidos pela doença (BOUSQUAT et al., 2021).

Entretanto, a covid-19 não se apresentava, em muitos casos, como uma doença isolada nos acometidos. Alguns fatores como idade e doenças pré-existentes agiram como estressores na gravidade da condição do paciente, devido à fragilização nos seus sistemas imunológicos, levando pessoas inclusive consideradas "saudáveis"ao óbito precoce. Mesmo

indivíduos que contraíram a doença de forma leve, ou até mesmo assintomáticos, não estavam imunes da responsabilidade coletivamente, pois poderiam servir de vetores do vírus a outros indivíduos, facilitando sua replicação e até mutação (MINUSSI et al., 2020).

Deste modo, o presente trabalho tem como objetivo estudar os fatores que influenciam no tempo até a morte de pacientes hospitalizados pelo SUS no Distrito Federal, em que a causa principal de hospitalização tenha sido a covid-19. Para isso, propõe o estudo de distribuições de probabilidade, com foco nas que possuam riscos unimodais devido ao comportamento visto nos dados, para assim encontrar uma distribuição robusta à variável tempo. E ao considerar a presença de variáveis intrínsecas ao paciente e que possam explicar o comportamento do tempo até a morte de pacientes hospitalizados por covid-19, este trabalho também propõe uma extensão da distribuição de probabilidade selecionada para construir um modelo de regressão.

Para estimar os parâmetros do modelo de regressão será considerado o método de máxima verossimilhança. Uma importante etapa após o ajuste de um modelo a um conjunto de dados consiste em verificar a adequabilidade das suposições feitas ao modelo. Para tanto o resíduo de Cox-Snell será utilizado. Todas as análises serão realizadas por meio *software R*.

Deste modo, este trabalho está organizado da seguinte forma: no Capítulo 2 é descrito o referencial teórico utilizado no trabalho, que inclui conceitos básicos de análise de sobrevivência, funções do tempo de sobrevivência, técnicas não-paramétricas, modelos probabilísticos, estimação de parâmetros, modelos de regressão paramétrico e técnicas de seleção e verificação da adequação do modelo. No Capítulo 3 é detalhada a metodologia utilizada na realização deste trabalho, contextualizando o banco de dados e a estrutura da modelagem realizada. No Capítulo 4 são mostrados os resultados do trabalho, tanto da análise descritiva quanto da modelagem e análise de resíduos. Por fim, o Capítulo 5 apresenta a conclusão do trabalho, na qual revisita-se os resultados dando enfoque nas principais conclusões e sugere possíveis trabalhos relacionados à temática abordada.

2 Referencial Teórico

2.1 Conceitos básicos de análise de sobrevivência

A técnica de análise de sobrevivência tem como objeto de interesse o tempo até a ocorrência de um evento, sendo necessário o acompanhamento ao longo do tempo das unidades observacionais para verificação da ocorrência do evento. Com isso, a variável resposta é composta por duas partes, sendo a primeira delas o tempo até o evento ocorrer ou o último acompanhamento e a segunda uma variável indicadora do acontecimento ou não do evento. Dessa forma, é necessário a definição detalhada do que é o evento de interesse, o tempo inicial do estudo e a escala de medida do tempo, podendo ser anos, dias, horas, etc. (COLOSIMO; GIOLO, 2006). Quando é registrado a ocorrência do evento, tem-se o tempo de falha. Caso contrário tem-se o tempo de censura ou observação parcial da resposta.

Assim, além das observações completas (ou falhas), a análise de sobrevivência também abarca as incompletas, também conhecidas como censuras. Essas são informações que por algum motivo não sofreram o evento de interesse durante o tempo determinado do estudo. Entre os tipos de censura que podem ocorrer estão:

- Censura à direita: O tempo de falha está à direita do tempo do estudo. Este tipo de censura ainda divide-se em três tipos:
  - Censura do tipo I: O estudo termina após um período de tempo fixo pré-estabelecido (t_f) e ao final, pelo menos uma observação do estudo não falhou.
  - Censura do tipo II: O estudo termina após um número pré-estabelecido de falhas (k) ocorrer. Assim, um número fixo ($k \leq n$) vivenciou o evento, enquanto as demais não.
  - Censura aleatória: Abarca todos os casos em que alguma observação não apresentou falha durante o estudo, por motivos não controláveis.
- Censura à esquerda: O tempo de falha está à esquerda do tempo de estudo, isto é, o evento de interesse aconteceu antes da unidade experimental ser analisada.
- Censura intervalar: Os elementos do estudo tem acompanhamentos periódicos, ocasionando que as falhas ocorram em intervalos de tempo com $T \in (L, U]$. Deste modo, quando $U = \infty$ tem-se a censura à direita e quando $L = 0$ tem-se a censura à esquerda, já que ambas são casos particulares da censura intervalar.

Com n obervações no estudo, a variável resposta associada a cada unidade experimental i $(i = 1, ..., n)$ é denotado por (t_i, δ_i), com t_i sendo o tempo, podendo ser tanto

de falha quanto de censura, e δ_i a variável indicadora do evento, de acordo com:

$$\delta_i = \begin{cases} 1, & \text{se a i-ésima observação falhou;} \\ 0, & \text{se a i-ésima observação foi censurada.} \end{cases} \tag{2.1.1}$$

Ao considerar a presença de variáveis explicativas no estudo, os dados da i-ésima observação do estudo são representados por (t_i, δ_i, x_i) em que $x_i = (1, x_1, x_2, ..., x_p)$ é o vetor de variáveis explicativas do modelo.

2.2 Funções do tempo de sobrevivência

O tempo de vida, T, ou tempo de sobrevivência, é uma variável aleatória não-negativa e geralmente contínua, que representa o tempo de falha. Sua distribuição pode ser mensurada pela função de densidade de probabilidade $f(t)$; pela função de sobrevivência $S(t)$ e pela função de risco $h(t)$. Essas funções fornecem informações sobre os dados de modo a ilustrar seu comportamento de diferentes formas.

Na análise de sobrevivência busca-se estimar essas funções de forma a extrair da amostra padrões de sobrevivência que possam ser extrapolados para sua população.

2.2.1 Função densidade de probabilidade

A função densidade de probabilidade, $f(t)$, modela a variável aleatória contínua T tal que:

$$\int_{-\infty}^{\infty} f(t)\, dt = 1. \tag{2.2.1}$$

Assim, $f(t)$ descreve o limite da probabilidade de se experimentar o evento de interesse em um intervalo de tempo ($[t, t + \Delta t]$) por unidade de comprimento do intervalo (Δt). Para tempos contínuos, ela é expressa por (LEE; WANG, 2003):

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{2.2.2}$$

em que a área inferior à curva de $f(t)$ é 1 e que $f(t) \geq 0\ \forall t$.

2.2.2 Função de sobrevivência

A função de sobrevivência, $S(t)$, denota a probabilidade de uma observação não falhar até o tempo t, equivalente à esta sobreviver ao tempo t, denotada por:

$$S(t) = P(T > t) = 1 - P(T \leq 1) = \int_t^{\infty} f(x)\, dx, \tag{2.2.3}$$

em que $S(t)$ é uma função monótona decrescente e contínua (LAWLESS, 1982). Da definição de distribuição acumulada, segue que a probabilidade de uma observação não sobreviver ao tempo t é dada por:

$$F(t) = 1 - S(t) = 1 - P(T > t). \tag{2.2.4}$$

A função $S(t)$ pode ser própria ou imprópria. No caso de ser própria, a função tem as seguintes propriedades: $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = 0$, isto é, a probabilidade de sobreviver pelo menos ao tempo 0 é 1, enquanto a probabilidade de sobreviver a um determinado tempo, que tende ao infinito, é 0. Em termos práticos, este conceito pode ser interpretado de forma que antes do início do estudo, ou seja, em $\lim_{t\to 0} S(t)$ a probabilidade de sobrevivência é 1, logo não há falhas. Já para $\lim_{t\to\infty} S(t)$, é esperado que todas as observações falhem, levando a probabilidade de sobrevivência à 0.

Já para a função de sobrevivência imprópria tem-se que $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = p$, em que p é uma probabilidade, sendo indicados neste caso modelos de sobrevivência com fração de cura.

2.2.3 Função de risco

A função de risco, também conhecida como função taxa de falha, é o limite da probabilidade de uma observação falhar no intervalo de tempo $[t, \Delta t)$, dado que este mesmo indivíduo sobreviveu até o tempo t, tudo isso divido pelo comprimento do intervalo. Esta função é definida por (LAWLESS, 1982):

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2.5}$$

A função de risco descreve a modificação da probabilidade instantânea de falha ao longo do tempo (COX; OAKES, 1984) e também pode ser definida por:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.2.6}$$

Segundo Colosimo e Giolo (2006), a função de risco é mais informativa que a função de sobrevivência. Isso porque diferentes funções de sobrevivência podem possuir formas parecidas, enquanto suas funções de falha correspondentes podem ser extremantes distintas. Assim, percebe-se a importância da modelagem da taxa de falha na análise de sobrevivência, ao poder assumir forma crescente, decrescente, constante ou não monótona.

2.2.4 Função de risco acumulada

Ao utilizar a função de risco, $h(t)$, a função de taxa de falha acumulada também representa o tempo de sobrevivência da seguinte forma:

$$H(t) = -\log(S(t)) = \int_0^t h(u)\,du. \tag{2.2.7}$$

Esta função mede o risco associado a falha até o tempo t, sendo a soma ou integral de todos os riscos dos tempos até o tempo t. Esta função é útil para se chegar à $h(t)$ por meio da estimação não paramétrica, além de ser relacionada com $S(t)$. Em termos práticos, a relação entre $S(t)$ e $h(t)$ se dá pelo aumento de uma implicar na diminuição da outra. Deste modo, quão maior for a probabilidade de sobrevivência em um determinado tempo, menor será seu risco, e vice-versa.

Uma vez introduzidas as principais funções da análise de sobrevivência, as relações das equações (2.2.6), (2.2.3) e (2.2.5) propiciam outras relações importantes:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log(S(t)), \tag{2.2.8}$$

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)\,du\right\}. \tag{2.2.9}$$

2.3 Técnicas não-paramétricas

Uma vez que a análise de sobrevivência lida com dados censurados, medidas de tendência central e variabilidade são impossibilitadas. Dessa forma, técnicas não-paramétricas são utilizadas, para encontrar estatísticas de interesse ao estimar a função de sobrevivência.

2.3.1 Estimador de Kaplan-Meier

Uma das técnicas mais utilizadas para estimar a função de sobrevivência de forma não paramétrica é o estimador de Kaplan-Meier (KAPLAN; MEIER, 1958). Isso porque

$\hat{S}(t)$ é não viesada (se tratando de amostras grandes) e um estimador de máxima verossimilhança não-paramétrico de $S(t)$. Além disso, é um estimador fracamente consistente, significando que a medida que o tamanho da amostra aumenta, o estimador converge para o verdadeiro valor do parâmetro. Ademais, também converge assintoticamente para distribuição Normal (BRESLOW; CROWLEY, 1974).

Na ausência de censura, o estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo t}}{\text{número total de observações no estudo}}, \tag{2.3.1}$$

no qual para obter $\hat{S}(t)$, deve-se primeiramente ordenar as observações. Ademais, $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, sendo n o tamanho da amostra. Caso hajam empates em algum tempo t, o tamanho do respectivo degrau é multiplicado pelo número de empates (COLOSIMO; GIOLO, 2006).

Agora, ao utilizar dados com a presença de censura, a estimativa de Kaplan-Meier passa a ser sequencial, em que o cálculo de cada passo do estimador depende do cálculo do seu antecessor. Existindo n observações e k falhas (com $k \leq n$) nos tempos $t_1 < ... < t_k$, o estimador é definido por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{2.3.2}$$

sendo d_j o número de falhas em t_j e n_j o número de observações sob risco em t_j, ou seja, todas as observações que não falharam e nem foram censuradas até o instante imediatamente anterior à t_j, com $j = 1, 2, ..., k$. Este estimador também é conhecido por estimador produto-limite (KAPLAN; MEIER, 1958).

Logo, das equações (2.3.2) e (2.2.7), o estimador de Kaplan-Meier para a função risco acumulado é expresso por:

$$\hat{H}(t) = -[log(\hat{S}(t))]. \tag{2.3.3}$$

2.3.2 Gráfico do tempo total em teste

Conforme mencionado na Seção 2.2.3, a função taxa de risco pode assumir diversas formas, sendo necessário utilizar metodologias que identifiquem modelos candidatos. Sendo assim, uma das metodologias utilizadas é o gráfico do tempo total em teste, também

conhecido por curva TTT, proposto por Aarset (1987) e é construído a partir de:

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}]}{(\sum_{i=1}^{n} T_i)}, \tag{2.3.4}$$

por r/n, em que $r = 1, ..., n$ e $T_{i:n}$, $i = 1, ..., n$ são as estatísticas de ordem da amostra, nas quais r representa o número de falhas e n o número de observações.

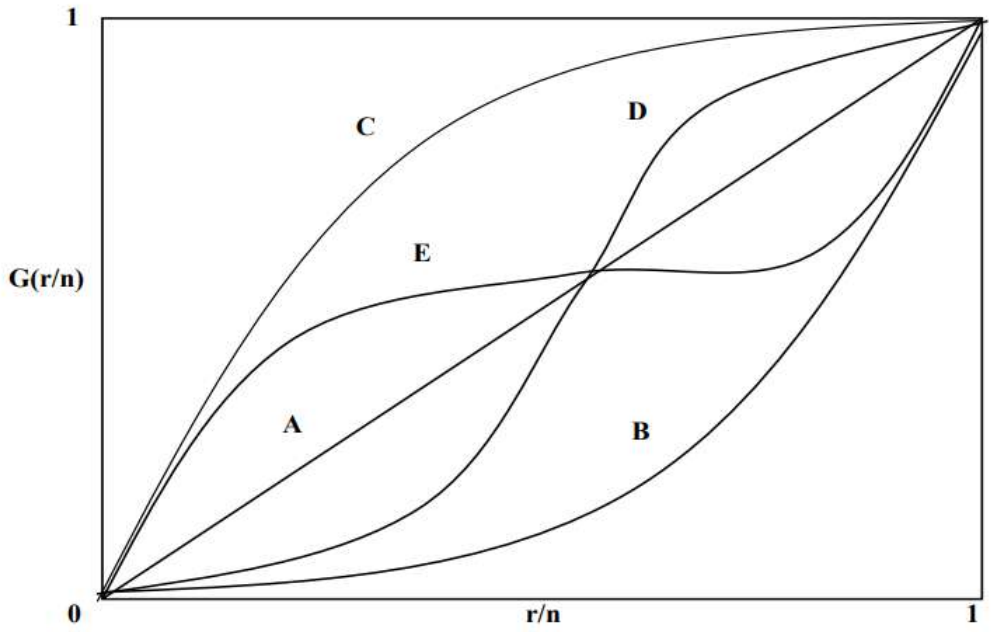

Figura 1: Ilustração de possíveis formatos da curva da função de risco

Fonte: SILVA, 2008

De acordo com a Figura 1, pode-se identificar comportamentos específicos da função de risco, tal que:

- Reta diagonal (A): Indica uma função de risco constante.
- Curva convexa (B): Indica uma função de risco monotonicamente decrescente.
- Curva côncava (C): Indica uma função de risco monotonicamente crescente.
- Curva convexa e depois côncava (D): Indica uma função de risco com formato de U.
- Curva côncava e depois convexa (E): Indica uma função de risco unimodal.

2.3.3 Gráfico da função de risco acumulada

Outra metodologia utilizada para identificar a forma da função taxa de falha dos dados é por meio do estudo do comportamento do gráfico da função de risco acumulada

estimada, $\hat{H}(t)$, em que $\hat{H}(t)$ pode ser calculado por meio do estimador de Kaplan-Meier.

Sua interpretação é o inverso da curva TTT. Logo, também considerando a Figura 1, tem-se:

- Reta diagonal (não necessariamente a reta $y = x$) (A): Indica uma função de risco constante.
- Curva convexa (B): Indica uma função de risco monotonicamente crescente.
- Curva côncava (C): Indica uma função de risco monotonicamente decrescente.
- Curva convexa e depois côncava (D): Indica uma função de risco unimodal.
- Curva côncava e depois convexa (E): Indica uma função de risco com formato de U.

Quando o número de censuras é grande, é indicado avaliar o comportamento da função de risco acumulada, pois essa metodologia utiliza tempos de falha e censura em seu cálculo.

2.3.4 Teste de logRank

Na presença de variáveis regressoras categóricas no conjunto de dados, é de interesse pesquisar o efeito dessas categorias na comparação dos tempos de sobrevivência. Entre os testes não-paramétricos mais utilizados está o teste de logRank, que tem como estatística de teste a diferença entre o número de falhas de cada categoria da variável e o valor esperado de falhas sob a hipótese nula (COLOSIMO; GIOLO, 2006), de acordo com:

$$\begin{cases} H_0 : S_1(t) = S_2(t) \text{ ou não existe diferença entre as curvas de sobrevivência.} \\ H_1 : S_1(t) \neq S_2(t) \text{ ou existe diferença entre as curvas de sobrevivência.} \end{cases}$$

Primeiramente, para cada um dos dois grupos analisados, ordena-se os tempos t_j com $j = 1, 2, ..., k$. Dessa forma, calcula-se os seguintes valores:

- n_j: número total de observações sob risco no tempo imediatamente anterior a t_j;
- n_{1j}: número de observações do grupo 1 sob risco em um tempo imediatamente anterior a t_j;
- d_j: número total de falhas no tempo t_j;
- d_{1j}: número de falhas no grupo 1 no tempo t_j.

Isto posto, sejam $t_{(1)} < t_{(2)} < ... < t_{(k)}$ os tempos de falha distintos da amostra considerando ambos grupos conjuntamente. Ocorrendo d_j falhas no tempo t_j e n_j observações estão sob risco no tempo imediatamente anterior a t_j, tem-se para cada t_j fixo que:

	Grupo 1	Grupo 2	
Falha	d_{1j}	d_{2j}	d_j
Não falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Risco	n_{1j}	n_{2j}	n_j

Tabela 1: Tabela de contingência no tempo t_j

Fixando as marginais de linha e coluna, tem-se que d_{1j} tem distribuição Hipergeométrica com média $E(d_{1j})$ e variância $V(d_{1j})$ dadas por:

$$E(d_{1j}) = d_j \frac{n_{ij}}{n_j} = w_{1j}. \tag{2.3.5}$$

$$V(d_{1j}) = d_j \frac{n_{ij}}{n_j}\left(1 - \frac{n_{ij}}{n_j}\right)\frac{(n_j - d_j)}{(n_j - 1)} = V_{1j}. \tag{2.3.6}$$

A estatística $d_{1j} - w_{1j}$ tem média zero e variância V_{1j}. Consequentemente, a estatística de teste do logRank é:

$$T = \frac{[\sum_{j=1}^{k}(d_{1j} - w_{1j})]^2}{\sum_{j=1}^{k} V_{1j}}, \tag{2.3.7}$$

com distribuição aproximadamente Qui-Quadrado com 1 grau de liberdade.

A utilização deste teste é mais apropriado quando a razão das funções de risco dos grupos comparados for constante, isto é, haver a suposição de riscos proporcionais. Esta suposição pode ser verificada analisando as curvas de sobrevivência ou de risco de ambos grupos, em que a razão das funções de risco deve ser aproximadamente constante, ou seja, as curvas dos riscos de cada função não se cruzam.

Caso não seja verificada a suposição de riscos proporcionais, o teste mais adequado é o teste de Wilcoxon. Seguindo as mesmas hipóteses do teste de logRank, o teste de Wilcoxon tem estatística de teste dada por:

$$T = \frac{[\sum_{j=1}^{k} n_j(d_{1j} - w_{1j})]^2}{\sum_{j=1}^{k} n_j^2 V_{1j}}, \tag{2.3.8}$$

o qual coloca mais peso na proporção inicial do eixo do tempo. Assim como no teste de

logRank, esta estatística de teste tem distribuição aproximadamente Qui-Quadrado com 1 grau de liberdade.

Caso exista evidência de desigualdade entre as curvas de sobrevivência dos dois grupos e os riscos das funções sejam proporcionais, pode ser calculado o risco relativo, de acordo com:

$$RR = \frac{O_1/E_1}{O_2/E_2} = \frac{(\sum_{j=1}^{k} d_{1j})(\sum_{j=1}^{k} d_j - E_{d_{1j}})}{(\sum_{j=1}^{k} E_{d_{1j}})(\sum_{j=1}^{k} d_j - d_{1j})}. \tag{2.3.9}$$

Os testes podem ser generalizados para contemplar variáveis explicativas com mais de duas categorias. Maiores detalhes estão em Colosimo e Giolo (2006).

2.4 Modelos Probabilísticos

Com o objetivo de realizar uma análise paramétrica, é necessário definir a distribuição de probabilidade que melhor representa o tempo de sobrevivência analisado (T). Este tempo se caracteriza por ser contínuo e não-negativo, além de comumente apresentar forte assimetria. Devido às características do tempo de sobrevivência, neste trabalho serão definidas as distribuições Exponencial, Weibull, Log-Logística, Log-Normal, Inversa Gaussiana Reparametrizada, Burr-XII, Kumaraswamy-Log-Normal, Kumaraswamy-Log-Logística, Kumaraswamy-Inversa-Gaussiana-Reparametrizada e Kumaraswamy-Burr-XII.

2.4.1 Distribuição Exponencial

Se a variável aleatória contínua, T, possui distribuição Exponencial, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \tag{2.4.1}$$

com $t \geq 0$, e em que $\alpha > 0$ é o tempo médio de vida e possui a mesma unidade de medida de t.

- Funções de Sobrevivência e Risco

$$S(t) = \left\{-\left(\frac{t}{\alpha}\right)\right\} \text{ e } h(t) = \frac{f(t)}{S(t)} = \frac{1}{\alpha}. \tag{2.4.2}$$

Dessa forma, percebe-se que $h(t)$ não depende do tempo, tendo como comportamento característico de uma função de risco constante.

2.4.2 Distribuição Weibull

Se a variável aleatória contínua, T, possui distribuição Weibull, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \tag{2.4.3}$$

com $t \geq 0$, e em que $\gamma > 0$ é parâmetro de forma e $\alpha > 0$ é o parâmetro de escala com a mesma unidade de medida de t.

- Funções de Sobrevivência e Risco

$$S(t) = \left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \text{ e } h(t) = \frac{f(t)}{S(t)} = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}. \tag{2.4.4}$$

O parâmetro de forma γ é responsável por determinar a forma da função de risco. Logo, para:

- $\gamma < 1$: A função de risco será decrescente.
- $\gamma > 1$: A função de risco será crescente.
- $\gamma = 1$: A função de risco será constante, equivalente à distribuição Exponencial.

2.4.3 Distribuição Log-Logística

Se a variável aleatória contínua, T, possui distribuição Log-Logística, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \left(1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right)^{-2}, \tag{2.4.5}$$

com $t > 0$, e em que $\gamma > 0$ é parâmetro de forma e $\alpha > 0$ é o parâmetro de escala.

- Funções de Sobrevivência e Risco

$$S(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} \text{ e } h(t) = \frac{f(t)}{S(t)} = \frac{\gamma\left(\frac{t}{\alpha}\right)^{\gamma-1}}{\alpha\left[1+\left(\frac{t}{\alpha}\right)^{\gamma}\right]}. \tag{2.4.6}$$

O parâmetro de forma, γ, é responsável por determinar a forma da função de risco e pode assumir as seguintes formas:

- $\gamma > 1$: A função de risco será unimodal.
- $\gamma \leq 1$: A função de risco será decrescente.

2.4.4 Distribuição Log-Normal

Se a variável aleatória contínua, T, possui distribuição Log-Normal, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \tag{2.4.7}$$

em que μ é a média do logaritmo do tempo de falha e σ seu desvio padrão, sendo $t > 0$.

- Funções de Sobrevivência e Risco

As funções de sobrevivência e risco da distribuição Log-Normal não possuem forma analítica explícita, sendo definidas como:

$$S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right) \text{ e } h(t) = \frac{f(t)}{S(t)}, \tag{2.4.8}$$

em que $\Phi(.)$ é a função de distribuição acumulada de uma distribuição Normal padrão. A função de risco da Log-Normal modela formas unimodais, possuindo, deste modo, comportamento decrescente da função de risco para valores grandes de T.

2.4.5 Distribuição Inversa Gaussiana Reparametrizada

A distribuição Inversa Gaussiana tem parâmetros $\mu > 0$ de locação e $\lambda > 0$ de forma, e possui função densidade de probabilidade dada por:

$$f(t) = \left(\frac{\lambda}{2\pi t^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right\}, \tag{2.4.9}$$

em que $t > 0$. Tweedie (1957) mostrou que a esperança e variância são, respectivamente:

$$E(T) = \mu \text{ e } Var(T) = \frac{\mu^3}{\lambda}. \tag{2.4.10}$$

Hashimoto et al. (2023) propôs uma reparametrização na distribuição Inversa Gaussiana, em termos de sua variância, sendo:

$$Var(T) = \frac{\mu^3}{\lambda} = \sigma^2 \Rightarrow \lambda = \frac{\mu^3}{\sigma^2}. \tag{2.4.11}$$

Dessa forma, ao substituir λ na equação (2.4.9), tem-se que se a variável T possui distribuição Inversa Gaussiana Reparametrizada, esta possui as funções:

- Função Densidade de Probabilidade

$$f(t) = \left(\frac{\mu^3}{2\pi\sigma^2 t^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\mu(t-\mu)^2}{2\sigma^2 t}\right\}, \tag{2.4.12}$$

com $t > 0$, $E(T) = \mu$ e $Var(T) = \sigma^2$.

As funções de sobrevivência e risco da distribuição Inversa Gaussiana Reparametrizada não possuem forma analítica explícita. Desta forma, sabendo que $S(t) = 1 - F(t)$, tem-se que:

- Funções de Sobrevivência e Risco

$$S(t) = 1 - \Phi\left(\sqrt{\frac{\mu^3}{\sigma^2 t}}\left(\frac{t}{\mu} - 1\right)\right) - \exp\left(\frac{2\mu^2}{\sigma^2}\right)\Phi\left(-\sqrt{\frac{\mu^3}{\sigma^2 t}}\left(\frac{t}{\mu} + 1\right)\right) \text{ e } h(t) = \frac{f(t)}{S(t)}, \tag{2.4.13}$$

em que $\Phi(.)$ é a função de distribuição acumulada de uma distribuição Normal padrão.

2.4.6 Distribuição Burr-XII

Diferentemente das distribuições citadas anteriormente, a distribuição Burr-XII possui três parâmetros, em que $\alpha > 0$ é o parâmetro de escala, $c > 0$ e $k > 0$ são parâmetros de forma. Se a variável aleatória contínua, T, possui distribuição Burr-XII, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$f(t) = \frac{\frac{kc}{\alpha}\left(\frac{t}{\alpha}\right)^{c-1}}{\left[1+\left(\frac{t}{\alpha}\right)^{c}\right]^{k+1}}, \tag{2.4.14}$$

em que c e k são parâmetros de forma e α parâmetro de escala.

- Funções de Sobrevivência e Risco

$$S(t) = \frac{1}{\left[1+\left(\frac{t}{\alpha}\right)^{c}\right]^{k}} \text{ e } h(t) = \frac{\frac{kc}{\alpha}\left(\frac{t}{\alpha}\right)^{c-1}}{1+\left(\frac{t}{\alpha}\right)^{c}}. \tag{2.4.15}$$

O formato da função de risco, segundo Zimmer, Keats e Wang (1998), pode ser classificado por meio da primeira derivada de $h(t)$.

$$h'(t) = \frac{kct^{(c-2)}\left[c-1-\left(\frac{t}{\alpha}\right)^{c}\right]}{\alpha^{c}\left[1+\left(\frac{t}{\alpha}\right)^{c}\right]^{2}}. \tag{2.4.16}$$

Dessa forma, consideram-se dois casos:

- $c \leq 1$: Para qualquer $t > 0$, tem-se $h'(t) < 0$. Portanto, $h(t)$ tem comportamento decrescente.
- $c > 1$: Calculando $h'(t^*) = 0$, chega-se em $c - 1 - \left(\frac{t^*}{\alpha}\right)^{c} = 0$, tendo seu ponto crítico $t^* = \alpha(c-1)^{\frac{1}{c}}$. Estudando os possíveis casos tem-se que quando $t < t^*$, $h'(t) > 0$ e, portanto, $h(t)$ é crescente. Já, quando $t > t^*$, $h'(t) < 0$, e, portanto, $h(t)$ tem forma decrescente. Dessa forma, vê-se que t^* é ponto de inflexão e $h(t)$ tem forma unimodal.

Além disso, percebe-se que ao tomar-se $k = 1$ na equação (2.4.14), chega-se na função de densidade da distribuição Log-Logística, representada na equação (2.4.5), em que $\alpha = \alpha$ e $c = \gamma$.

2.4.7 Distribuição Kumaraswamy e sua generalização

A distribuição Kumaraswamy possui dois parâmetros, em que $\alpha > 0$, $\beta > 0$ são parâmetros de forma. Se a variável aleatória contínua, T, possui distribuição Kumaraswamy, então, ela será caracterizada pelas seguintes funções:

- Função Densidade de Probabilidade

$$g(t) = \alpha\beta t^{\alpha-1}(1-t^{\alpha})^{\beta-1}. \tag{2.4.17}$$

- Função de Distribuição acumulada

$$G(t) = 1-(1-t^{\alpha})^{\beta}. \tag{2.4.18}$$

Para estas funções, tem-se $0 < t < 1$. Entretanto, em análise de sobrevivência a variável T assume valores não restritos à 1, tal que $t > 0$. Desta forma, é proposta uma nova classe de distribuições por Cordeiro e Castro (2011), fundamentado nos trabalhos de Eugene, Lee e Famoye (2002) e Jones (2009), chamada de Kumaraswamy generalizada.

Considera-se uma função de distribuição acumulada $G(t)$ arbitrária. Desta forma, a distribuição Kumaraswamy generalizada terá as funções:

- Função Densidade de Probabilidade

$$f(t) = abg(t)G(t)^{a-1}[1-G(t)^{a}]^{b-1}. \tag{2.4.19}$$

- Função de Distribuição acumulada

$$F(t) = 1-[1-G(t)^{a}]^{b}, \tag{2.4.20}$$

em que $g(t) = \frac{dG(t)}{dt}$, $a > 0$ e $b > 0$ são dois parâmetros adicionais, os quais funcionam como flexibilizadores dos pesos das caudas, inserindo assimetria. Desta forma, usando a notação K-G para a distribuição Kumaraswamy generalizada, se T é uma variável aleatória com função de densidade de probabilidade 2.4.19, diz-se que $T \sim$ K-G(a, b).

2.5 Estimação de Parâmetros

Para estimar os parâmetros do modelo de interesse, é necessário utilizar algum método de estimação baseado nos dados amostrais. O método de mínimos quadrados não é apropriado para análise de sobrevivência, pois não consegue inserir as observações censuradas em sua estimação. Dessa forma, o método a ser utilizado é o de máxima verossimilhança.

2.5.1 Método da Máxima Verossimilhança

Esse método tem como objetivo encontrar a distribuição, entre todas as definidas pelas possíveis combinações de seus respectivos parâmetros, que possua maior possibilidade de ter gerado os dados observados na amostra.

Dessa forma, sendo $t_1, t_2, ..., t_n$ uma amostra de observações em que todas as observações são falha, ou seja, não são censuradas. Seja ainda que a população da qual a amostra é proveniente tenha função de densidade $f(t_i, \boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros.

Sendo assim, sua função de verossimilhança é definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}). \tag{2.5.1}$$

No contexto em que não existam apenas tempo de falha na amostra, seja $t_1, t_2, ..., t_n$ uma amostra de tempos de sobrevivência em que a variável indicadora δ_i recebe 1 caso t_i seja tempo de falha e 0 caso t_i seja tempo de censura.

Dessa forma, cada elemento da amostra contribui para a função de verossimilhança de acordo com:

$$\begin{cases} f(t_i, \boldsymbol{\theta}), & \text{se } t_i \text{ é tempo de falha.} \\ S(t_i, \boldsymbol{\theta}), & \text{se } t_i \text{ é tempo de censura.} \end{cases} \tag{2.5.2}$$

Considerando todos os mecanismos de censura à direita, sendo r o número de falhas, a função de verossimilhança é dada por:

$$\begin{aligned} L(\boldsymbol{\theta}) &\propto \prod_{i=1}^{r} f(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i, \boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n} [f(t_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i, \boldsymbol{\theta})]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} [h(t_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i, \boldsymbol{\theta})]. \end{aligned} \tag{2.5.3}$$

Em que $f(t_i, \boldsymbol{\theta})$, $S(t_i, \boldsymbol{\theta})$ e $h(t_i, \boldsymbol{\theta})$ são, respectivamente, a função de densidade, a função de sobrevivência e a função de risco do modelo probabilístico considerado.

Para encontrar os valores de $\boldsymbol{\theta}$ que maximizem a função $L(\boldsymbol{\theta})$, ou seja, os estimadores de máxima verossimilhança, aplica-se o logaritmo em $L(\boldsymbol{\theta})$, ou seja,

$$\log(L(\boldsymbol{\theta})) = l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big\{ \delta_i \log[f(t_i, \boldsymbol{\theta})] + (1 - \delta_i) \log[S(t_i, \boldsymbol{\theta})] \Big\}. \tag{2.5.4}$$

Por conseguinte, os estimadores de máxima verossimilhança são encontrados re-

solvendo o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = \mathbf{0}. \tag{2.5.5}$$

Para resolver o sistema de equações (2.5.5) é necessária a aplicação de métodos numéricos, como o de Newton-Raphson. Esse método e outros estão disponíveis no *software R*.

2.5.2 Intervalo de Confiança para os Parâmetros

Além de encontrar as estimativas pontuais dos parâmetros do modelo pelo método de máxima verossimilhança, esse método também permite construir intervalos de confiança para os parâmetros a partir da distribuição assintótica dos estimadores de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ (COLOSIMO; GIOLO, 2006).

Sob certas condições de regularidade e considerando amostras grandes, admite-se que $\hat{\boldsymbol{\theta}}$ tem distribuição assintótica Normal Multivariada com média $\boldsymbol{\theta}$ e matriz de variância e covariância $Var(\hat{\boldsymbol{\theta}})$, ou seja,

$$\hat{\boldsymbol{\theta}} \sim N_q(\boldsymbol{\theta}, Var(\hat{\boldsymbol{\theta}})), \tag{2.5.6}$$

em que q é a dimensão de $\hat{\boldsymbol{\theta}}$.

Ainda sob certas condições de regularidade, a matriz de variância e covariância é aproximadamente o negativo da inversa da informação de Fisher. Logo,

$$Var(\hat{\boldsymbol{\theta}}) \simeq -[I_F(\boldsymbol{\theta})]^{-1} = -\left\{ E\left[\left(\frac{\partial logL(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^2 \right] \right\}^{-1}. \tag{2.5.7}$$

Em casos em que não é possível calcular a esperança descrita acima, utiliza-se a matriz de informação observada $\ddot{L}(\boldsymbol{\theta})$ avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

$$\ddot{L}(\boldsymbol{\theta}) = \frac{\partial^2 logL(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}. \tag{2.5.8}$$

Dessa forma, tem-se que:

$$Var(\hat{\boldsymbol{\theta}}) \simeq -\left[\frac{\partial^2 logL(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}} \right]^{-1}. \tag{2.5.9}$$

Portanto, para a construção do intervalo de confiança de fato, utiliza-se a estimativa do erro padrão de $\hat{\boldsymbol{\theta}}$, sendo os elementos da diagonal principal de $[Var(\hat{\boldsymbol{\theta}})]^{1/2}$. Se

θ é escalar, o intervalo de confiança aproximado ao nível de confiança de $1-\alpha$ é:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta})}. \tag{2.5.10}$$

Caso $\boldsymbol{\theta}$ seja um vetor de parâmetros, constrói-se um intervalo de confiança para cada parâmetro separadamente, sendo necessário obter a estimativa do erro padrão por meio da matriz de variância e covariância $Var(\hat{\boldsymbol{\theta}})$.

2.6 Modelo de Regressão Paramétrico

Em estudos em que é de interesse analisar a presença de variáveis explicativas que representem a heterogeneidade existente na população, pode-se construir modelos de regressão centrados na relação entre os tempos de sobrevivência e as variáveis explicativas de interesse.

Dessa forma, objetiva-se estimar o efeito de $\boldsymbol{\theta}$ de p variáveis explicativas $\mathbf{x}^t = (1, x_1, x_2, ..., x_p)$ sobre o tempo de sobrevivência T. Para tanto, é utilizada uma função de ligação $g(.)$ de forma a relacionar o conjunto $\mathbf{x}^T$ de variáveis explicativas com a variável resposta. O vetor de parâmetros $\boldsymbol{\theta}$, a partir de um conjunto p de covariáveis, é definido por:

$$\boldsymbol{\theta} = g(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{2.6.1}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ é o vetor de coeficientes de regressão.

2.7 Regressão Log-Normal

Sendo T uma variável aleatória com distribuição Log-Normal, definida em (2.4.7), usa-se μ como função de ligação para a distribuição Log-Normal tal que $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$. Dessa forma, a função de ligação torna-se a função identidade $I(.)$. Dessa forma, o modelo de regressão Log-Normal é definido por:

$$f(t|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{\frac{-[\log(t) - \boldsymbol{x}^T\boldsymbol{\beta}]^2}{2\sigma^2}\right\}. \tag{2.7.1}$$

As funções de sobrevivência e risco, respectivamente, são dadas por:

$$S(t|\boldsymbol{x}) = \Phi\left(\frac{\log(t) + \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) \text{ e } h(t|x) = \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})}. \tag{2.7.2}$$

Para estimar os parâmetros do modelo de regressão Log-Normal será utilizado o

método de máxima verossimilhança descrito na Seção 2.5.1.

2.8 Regressão Burr-XII

Sendo T uma variável aleatória com distribuição Burr-XII, definida em (2.4.14) e ao considerar que o parâmetro α depende das variáveis explicativas por meio da relação $\alpha = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$, então, o modelo de regressão Burr-XII é definido por:

$$f(t|\boldsymbol{x}) = \frac{\frac{kc}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{c-1}}{\left[1+\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{c}\right]^{k+1}}, \tag{2.8.1}$$

com funções de sobrevivência e risco dadas por:

$$S(t|\boldsymbol{x}) = \frac{1}{\left[1+\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{c}\right]^{k}} \text{ e } h(t|\boldsymbol{x}) = \frac{\frac{kc}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{c-1}}{1+\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{c}}. \tag{2.8.2}$$

Para estimar os parâmetros do modelo de regressão Burr-XII será utilizado o método de máxima verossimilhança descrito na Seção 2.5.1.

2.9 Seleção de Modelos

Com o intuito de escolher qual modelo entre os possíveis é o que melhor descreve a amostra analisada, serão descritos critérios que possibilitem a comparação entre os melhores modelos ajustados.

2.9.1 Técnica gráfica

Entre as técnicas gráficas que comparam modelos, será apresentada a técnica que é capaz de comparar mais de um modelo ao mesmo tempo. Nela utiliza-se a função de sobrevivência estimada pelo método de Kaplan-Meier e as funções de sobrevivências estimadas dos modelos propostos.

Uma vez estimadas as funções de sobrevivência, constrói-se um gráfico com a curva de sobrevivência de Kaplan-Meier e coloca-se no mesmo gráfico as curvas de sobrevivências estimadas dos modelos de interesse. Seleciona-se o modelo que melhor acompanhar a curva de sobrevivência estimada pelo método de Kaplan-Meier.

2.9.2 Teste da Razão de Verossimilhanças

Para comparar de maneira objetiva dois modelos ou duas distribuições de interesse, pode ser utilizado o Teste da Razão de Verossimilhanças, desde que os modelos ou distribuições sejam encaixados. Desse modo, os modelos ou distribuições a serem comparados tem de estar conectados de forma que um seja caso particular do outro (COLOSIMO; GIOLO, 2006). O teste segue as seguintes hipóteses:

$$\begin{cases} H_0 : \text{O modelo restrito é o adequado } (\boldsymbol{\theta} = \boldsymbol{\theta}_0). \\ H_1 : \text{O modelo completo é o adequado } (\boldsymbol{\theta} \neq \boldsymbol{\theta}_0). \end{cases} \tag{2.9.1}$$

A estatística de teste leva em consideração o logaritmo da função de verossimilhança dos modelos, de acordo com:

$$TRV = -2\left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right] = 2\Big[\log L(\hat{\boldsymbol{\theta}}) - logL(\hat{\boldsymbol{\theta}}_0)\Big], \tag{2.9.2}$$

que sob H_0, tem distribuição aproximadamente Qui-Quadrado com p graus de liberdade, em que p é a subtração do número de parâmetros do modelo completo pelo número de parâmetros do modelo restrito.

2.9.3 Critérios de Informação

Uma forma de comparar mais de dois modelos simultaneamente de forma objetiva é utilizar os critérios de informação, de forma a facilitar a seleção do modelo final de forma parcimoniosa.

- Critério de Akaike - AIC

Este critério leva em consideração o logaritmo da função de verossimilhança no ponto de máximo, aumentada uma penalidade de acordo com o número de parâmetros do modelo, que serve como correção de viés ao comparar modelos com diferente quantidade de parâmetros. Esse critério é definido por:

$$AIC = -2logL(\hat{\boldsymbol{\theta}}) + 2p, \tag{2.9.3}$$

em que p é o número de parâmetros estimados do modelo. De acordo com Anderson e Burnham (2004), é recomendado o utilizar o AIC quando $n/p < 40$. O modelo escolhido deve ser o que apresentar menor AIC dentre os modelos estimados.

- Critério de Akaike Corrigido - AICc

No caso em que a amostra é pequena ($n/p \leq 40$) utiliza-se o AICc em detrimento do AIC. O critério AICc é dado por:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}. \tag{2.9.4}$$

Da mesma forma, escolhe-se o modelo com menor AICc dentre os modelos analisados.

- Critério de Informação Bayesiano - BIC

Este critério traz maior penalidade aos modelos com mais parâmetros em comparação com o critério AIC. Sendo assim, o critério BIC tende a priorizar a escolha de modelos mais parcimoniosos, ou seja, com menos parâmetros. Esse critério é descrito por:

$$BIC = -2logL(\hat{\boldsymbol{\theta}}) + p * \log(n). \tag{2.9.5}$$

Deve-se escolher o modelo que apresentar menor valor de BIC entre os modelos estimados.

2.10 Adequação do modelo

Com objetivo principal de rejeitar modelos inapropriados, norteando o estudo a selecionar o modelo que melhor se adequa aos dados, respeitando as suposições dos pressupostos das técnicas utilizadas, faz-se análise de resíduos dos modelos de interesse (COLOSIMO; GIOLO, 2006).

2.10.1 Resíduos de Cox-Snell

Para se estudar o ajuste geral dos modelos de interesse, calcula-se os resíduos de Cox-Snell de acordo com:

$$\hat{e}_i = \hat{H}(t_i|\mathbf{x}_i), \tag{2.10.1}$$

em que $\hat{H}(.)$ é a função de risco acumulada do modelo ajustado e $\mathbf{x}_i$ o vetor de covariáveis do modelo analisado (COLOSIMO; GIOLO, 2006).

De acordo com Lawless (1982), os resíduos $\hat{e}_i$ são provenientes de uma população homogênea, devendo seguir distribuição exponencial padrão se o modelo for adequado.

Portanto, o gráfico de $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$ deve ser aproximadamente uma reta. Deste modo, dado que $\hat{H}(\hat{e_1}) = -\log(S(\hat{e_i}))$, também pode-se verificar a adequabilidade do modelo por meio do gráfico das curvas de sobrevivência desses resíduos, provenientes do Kaplan-Meier e do modelo exponencial padrão, ou seja, $\exp\{-\hat{e_i}\}$ versus $\hat{S}_K M(\hat{e}_i)$.

3 Metodologia

3.1 Base de Dados

O banco de dados deste estudo é proveniente do TABWIN, localizado no DATASUS. A base utilizada é referente ao Sistema de Informações Hospitalares do SUS (SIHSUS), o qual disponibiliza uma série de informações sobre os indivíduos que foram hospitalizados pelo SUS. Deste modo, foram selecionadas as informações do Distrito Federal, abarcando de janeiro de 2020 até setembro de 2023 (último banco de dados disponível no momento deste estudo).

Cada linha do banco é referente a um indivíduo, que está atrelado à sua respectiva Autorização de Internação Hospitalar (AIH). Nesta autorização constam diversas informações dos pacientes, como sexo, idade, raça/cor, número de filhos, causa principal da hospitalização segundo a CID-10, entre outras. As bases de dados são disponibilizadas por mês, segundo o mês de processamento do paciente no sistema, assim, a base de dados de janeiro de 2021, por exemplo, poderá possuir pacientes que foram internados em Janeiro de 2021 e também pacientes que foram internados em dezembro de 2020, mas apenas processados em janeiro de 2021. Desta forma, a data utilizada neste estudo é a data de entrada do paciente no hospital, e não a de processamento do mesmo.

O banco possui uma variável referente à quantidade de dias que o indivíduo permaneceu no hospital, sendo a variável de tempo do estudo. Nela, um paciente que deu entrada no hospital em um dia, tendo alta ou óbito neste mesmo dia, tem como permanência 0 dias. Para efetuar os cálculos neste estudo, foi acrescentado 1 dia de permanência a todos os pacientes do banco, para que todos tenham estado pelo menos 1 dia de permanência, possibilitando o cálculo das análises de interesse.

Sabe-se que a variável referente ao valor total da AIH não é homogênea, isto porque ao calcular o quanto foi gasto na hospitalização do paciente, a variável utiliza o preço dos medicamentos e procedimentos provenientes de cada localidade. Deste modo, diferentes regiões podem possuir valores distintos para uma mesma intervenção médica aplicada. Contudo, como o estudo abarca sistemas hospitalares do SUS apenas do Distrito Federal, entende-se que esta variabilidade não apresentará discrepâncias consideráveis, sendo mantida a variável no estudo.

O intuito do estudo é modelar o tempo, em dias, até a morte do paciente por covid-19. Desta forma, as CID's (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde) utilizadas para identificação dos indivíduos que deram entrada no hospital com diagnóstico principal de covid-19 foram $B34.2$, referente à infecção por coronavírus de localização não especificada, e a $B97.2$, referente ao coronavírus como

causa de doenças classificadas em outros capítulos, ambas do Capítulo 1 da CID-10 que trata de doenças infecciosas e parasitárias.

3.2 Variáveis

As variáveis utilizadas neste estudo foram escolhidas pela sua ligação direta ao paciente, relação com o tempo até a morte do indivíduo e ausência de informações faltantes no banco de dados. As variáveis escolhidas foram:

- Idade: idade do paciente, em anos, no momento da hospitalização
- Sexo: sexo biológico do paciente, sendo feminino ou masculino
- Ano: ano de internação do paciente
- Valor total da AIH: valor total exato, em reais, gasto referente a todos os aspectos da hospitalização do paciente

A variável resposta é composta por duas variáveis do banco: variável do tempo, em dias, de permanência do paciente no hospital e variável indicadora de morte ou não do paciente. Deste modo, a morte do paciente será considerada falha no estudo, enquanto a sobrevivência será a censura.

3.3 Análise de dados

A análise de dados consiste, primeiramente, na análise descritiva das variáveis de interesse no estudo. Serão utilizados métodos gráficos e testes estatísticos de forma a se entender o comportamento das variáveis. Na análise de variáveis categóricas se utilizará gráficos de barras, enquanto para analisar variáveis quantitativas, os gráficos utilizados serão *boxplots*. Além disso, para entender a relação entre as covariáveis com a variável resposta indicadora de falha, serão utilizados gráficos de colunas justapostas.

Após isto, também será feita análise descritiva utilizando técnicas de análise de sobrevivência, contando com gráficos de Kaplan-Meier além do gráfico do Tempo Total de Teste e de função de risco acumulado, que auxiliarão na definição do melhor tipo de distribuição de probabilidade para modelagem dos dados.

3.4 Modelagem

A modelagem dos dados passará pela análise das distribuições já implementadas no pacote *survival* do *Rstudio*, assim como a implementação manual do *software R* por

meio da função *optim* de distribuições com risco unimodal.

Após a análise descritiva, na qual as variáveis serão avaliadas separadamente, observando as curvas de sobrevivência e diferenças entre suas categorias, será realizado um estudo de qual distribuição de probabilidade se ajusta melhor ao tempo até a morte de pacientes hospitalizados por covid-19. As distribuições analisadas serão as definidas na Seção 2.4.

Uma vez escolhida a melhor distribuição, dentre as estudadas, serão propostos modelos de regressão como definido na Seção 2.6. Os modelos propostos passarão por etapas de seleção de variáveis, de acordo com passo a passo descrito por Colosimo e Giolo (2006), em que serão avaliados modelos com as variáveis explicativas de estudo assim como interações entre elas.

Por fim, para verificar a qualidade do ajuste dos modelos propostos, serão analisados os resíduos dos modelos, de acordo com o procedimento descrito na Seção 2.10.

4 Resultados

4.1 Análise Descritiva

Para entender a relação das variáveis explicativas com a variável resposta, antes da modelagem dos dados propriamente dita, é necessário entender como as variáveis se comportam entre si, por meio da análise descritiva.

4.1.1 Status do paciente

Parte da variável resposta na análise de sobrevivência é a variável indicadora de falha, neste caso, morte do paciente. O gráfico de colunas apresenta os valores absolutos e relativos dos pacientes que foram hospitalizados no DF com diagnóstico principal de covid-19 que receberam alta (sobreviveram), e os que vieram a falecer.

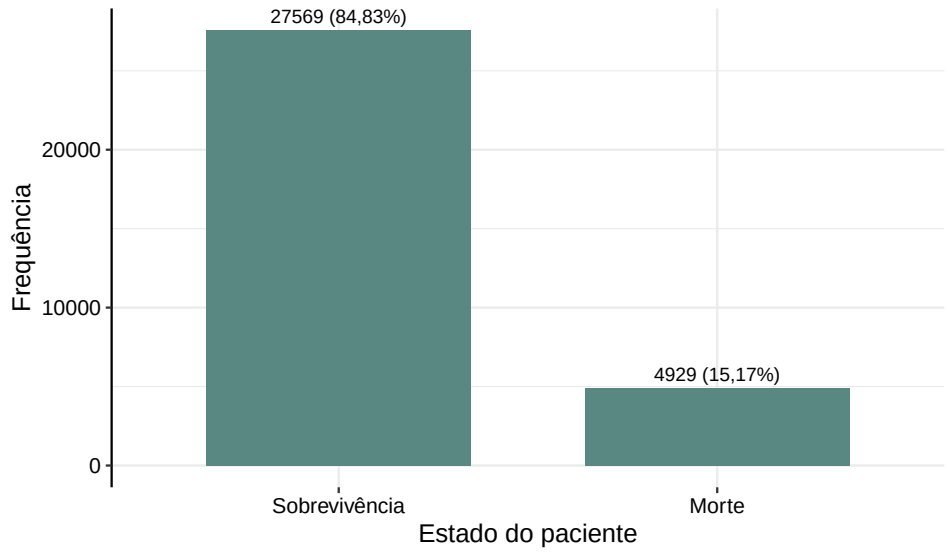

Figura 2: Gráfico de colunas do número de mortes e sobreviventes de covid-19 de 2020 a 2023

Por meio da Figura 2, vê-se que mais de 15% dos pacientes hospitalizados por covid-19 vieram a falecer, considerando todo o período estudado. Dessa forma, percebe-se que quase 85% do banco é constituído por censuras, como também pode ser visto no gráfico de Kaplan-Meier na Figura 3.

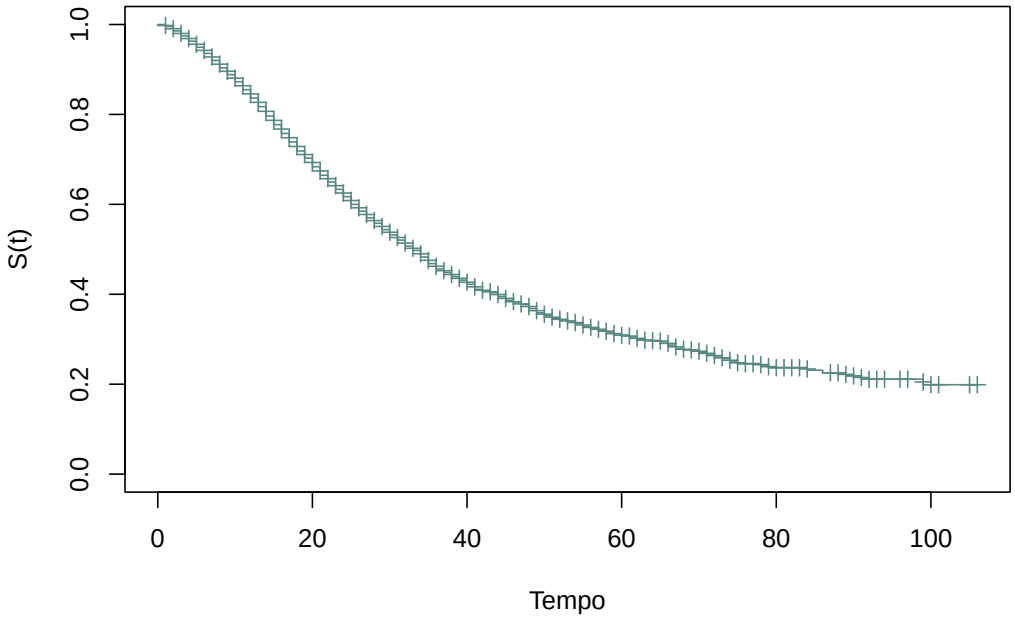

Figura 3: Curva de sobrevivência estimada pelo método de Kaplan-Meier para o tempo até a morte de pacientes hospitalizados por covid-19

A curva de Kaplan Meier indica uma quantidade substancial de censuras ao longo do tempo, como esperado, denotadas pelas linhas verticais na curva. Estas censuras estão aparentemente mais presentes em seu início. A curva apresenta tempo mediano de internação de 32,8 dias, além de decaimento contínuo até estabilizar-se por volta de 0,2.

4.1.2 Ano de internação

Essa variável retrata a disposição dos dados em relação ao ano em que o paciente foi hospitalizado, entre os anos de 2020 a 2023.

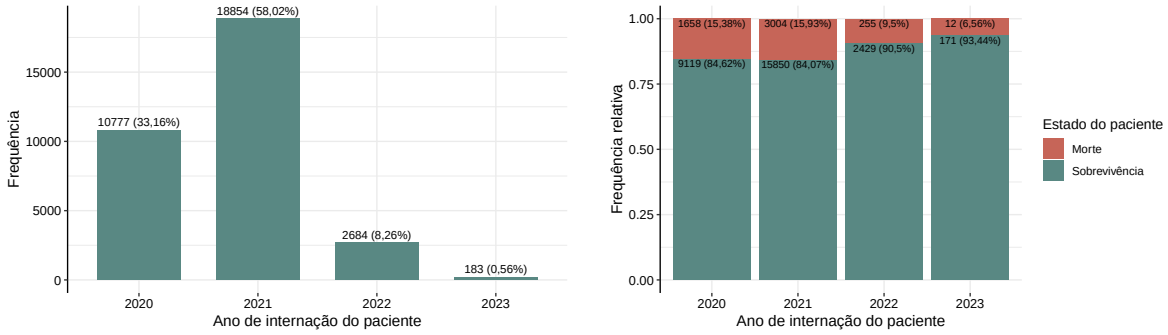

Figura 4: Gráficos de colunas do número de hospitalizações por covid-19 no período de 2020 a 2023 (à esquerda) e do número de mortes e sobreviventes no período de 2020 a 2023 (à direita)

Ao analisar a Figura 4, percebe-se que hospitalizações por covid-19 no Distrito Federal ocorreram em maior quantidade em 2020 e 2021 em comparação com os anos seguintes, representando mais de 91% dos dados. Isso se deve ao primeiro caso da covid-19 no Distrito Federal ter sido registrado em março de 2020, com transmissão sequencial e alta de casos considerável em alguns meses de 2020 e 2021, como registrado na Figura 6. Em meio a diversas medidas de contenção do vírus, assim como recursos de minimização da gravidade de casos da doença, nos anos subsequentes é vista a queda do número de internações pela doença. Dessa forma, os dados analisados neste estudo compreendem mais de 10.000 hospitalizações em 2020, quase 19.000 hospitalizações em 2021, 2684 em 2022 e apenas 183 em 2023.

Observando a relação do ano com a variável indicadora de falha, nota-se que para os anos de 2020 e 2021, as mortes representaram entre 15% e 16%. Já para 2022, as mortes foram 9,5%, enquanto dos hospitalizados em 2023 por covid-19, apenas 12 morreram, representando 6,56%.

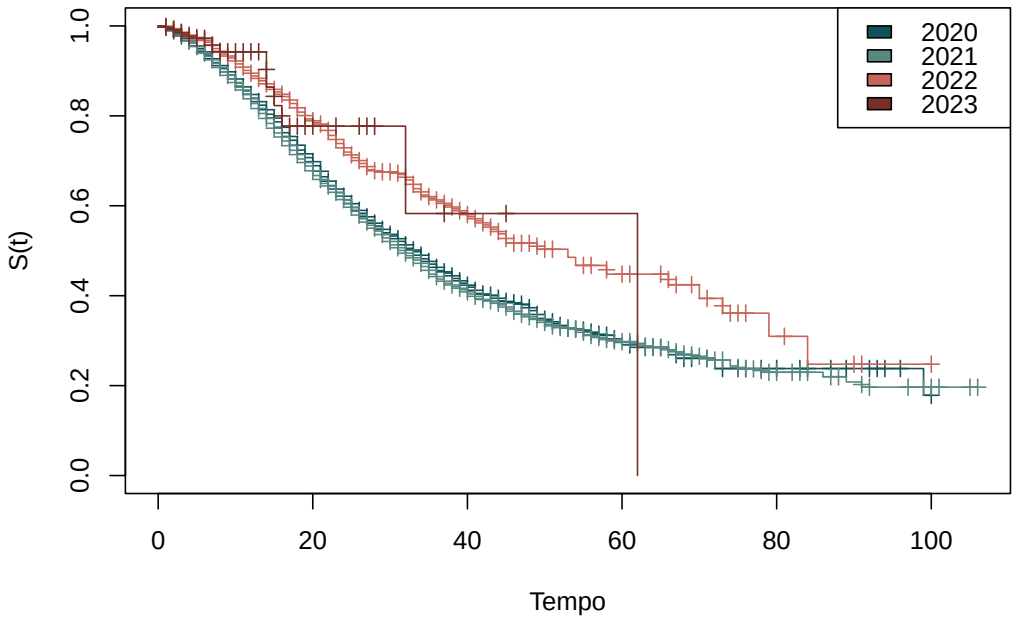

Figura 5: Gráfico de Kaplan-Meier do banco de covid-19 por ano com falha sendo a morte do paciente

Quanto às curvas de sobrevivência, percebe-se na Figura 5 a alta quantidade de censuras nas curvas de todos os anos. No geral, vê-se que as curvas de 2020 e 2021 são inferiores às demais, mostrando que pacientes hospitalizados em 2020 e 2021 tem probabilidade de sobrevivência menor que os dos anos seguintes. Ademais, o formato da curva de 2023 é devido a ter-se dados apenas até setembro de 2023, no momento deste

estudo, mostrando tempo mais curto da curva em comparação aos demais.

Dado que os riscos não são proporcionais, isto é, as curvas de sobrevivência dos grupos se cruzam, será aplicado o teste de Wilcoxon com o fito de se obter mais certeza sobre a existência de diferença ou não entre as curvas de sobrevivência dos anos considerados no estudo. As hipóteses do teste são as seguintes:

$$\begin{cases} H_0 : \text{Não existem diferenças entre as curvas de sobrevivência do ano de internação do paciente.} \\ H_1 : \text{Existem diferenças entre as curvas de sobrevivência do ano de internação do paciente.} \end{cases}$$

Ao analisar os resultados apresentados na Tabela 2 e ao nível de significância de 0,05; rejeita-se a hipótese nula. Dessa forma, afirma-se que pelo menos uma das curvas de sobrevivência dos anos difere das demais.

Tabela 2: Teste de Wilcoxon para a variável Ano

Variável	Estatística de teste	Graus de liberdade	P-valor	Decisão
Ano	58,8	3	$< 0,001$	Rejeita H_0

Aplicando o teste de Wilcoxon 2 a 2, na Tabela 3, percebe-se que o ano de 2023 não difere de nenhum outro ano. Isso pode ser pela pequena quantidade de dados de covid-19 neste ano, até por ser o único ano que só se tem informação sobre os nove primeiros meses, ao invés de todos os meses do ano. Além disso, nota-se que 2020 não difere de 2021, além do que ambos diferem de 2022, mostrando uma clara diferença entre os períodos.

Tabela 3: Teste de Wilcoxon para cada dupla de categorias da variável Ano

Variável	Estatística de teste	Graus de liberdade	P-valor	Decisão
2020:2021	3,1	1	$0,08$	Não rejeita H_0
2020:2022	41,8	1	$< 0,001$	Rejeita H_0
2020:2023	3,1	1	$0,08$	Não rejeita H_0
2021:2022	55	1	$< 0,001$	Rejeita H_0
2021:2023	3,7	1	$0,05$	Não rejeita H_0
2022:2023	0,1	1	$0,8$	Não rejeita H_0

Analisando por mês de internação, em cada ano, o gráfico da esquerda da Figura 6 mostra que os picos de internações por covid-19 aconteceram em março e abril de 2021, com quase 4.000 internações em cada mês, seguidos de queda, mas com valor ainda alto, em maio. Além disso, outros valores de destaque nas hospitalizações estão associados aos meses de julho e agosto de 2020, momentos em que passou-se de 2.000 hospitalizações.

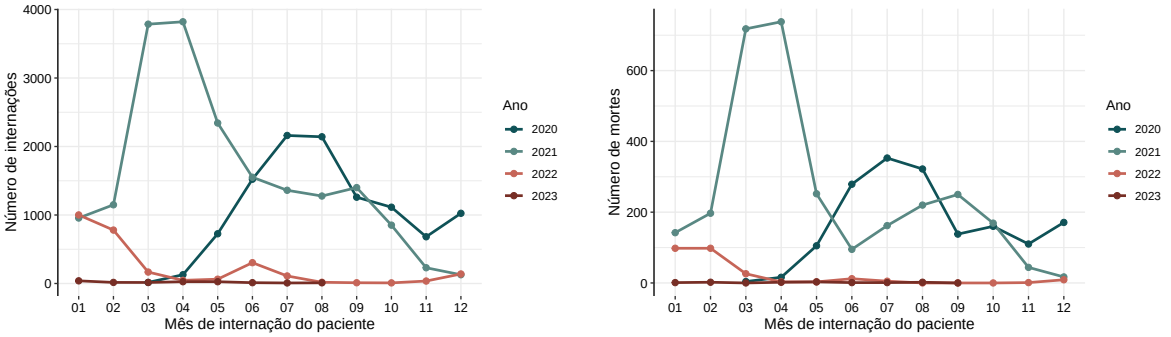

Figura 6: Gráficos de linhas do número de internações (à esquerda) e mortes (à direita) no período de 2020 a 2023 por mês

O gráfico à direita da Figura 6 fornece o número de mortes por mês, em cada ano. Com comportamento similar ao gráfico de hospitalizações, este mostra que os picos de mortes foram nos meses de março e abril de 2021, registrando mais de 700 mortes em cada mês. Além disso, em 2020 também observou-se valores altos, como entre julho e agosto, com mais de 300 mortes em cada mês.

O número de hospitalizações e mortes em 2022 e 2023 foram bem menores em relação aos anos anteriores, com valores quase zerados em alguns meses, devido a escala do gráfico. Vale ressaltar que os dados vão até setembro de 2023, resultando na quebra da linha do ano correspondente.

4.1.3 Sexo

Em relação ao sexo dos pacientes, pelos gráficos apresentados na Figura 7, nota-se números equiparados de internações entre os sexos dos pacientes hospitalizados por covid-19 no DF, com o sexo masculino levemente superior ao feminino no banco geral. Esta mesma tendência é observada nos anos estudados, exceto para 2022, no qual o sexo feminino apresenta valores superiores, mas por menos de 5%.

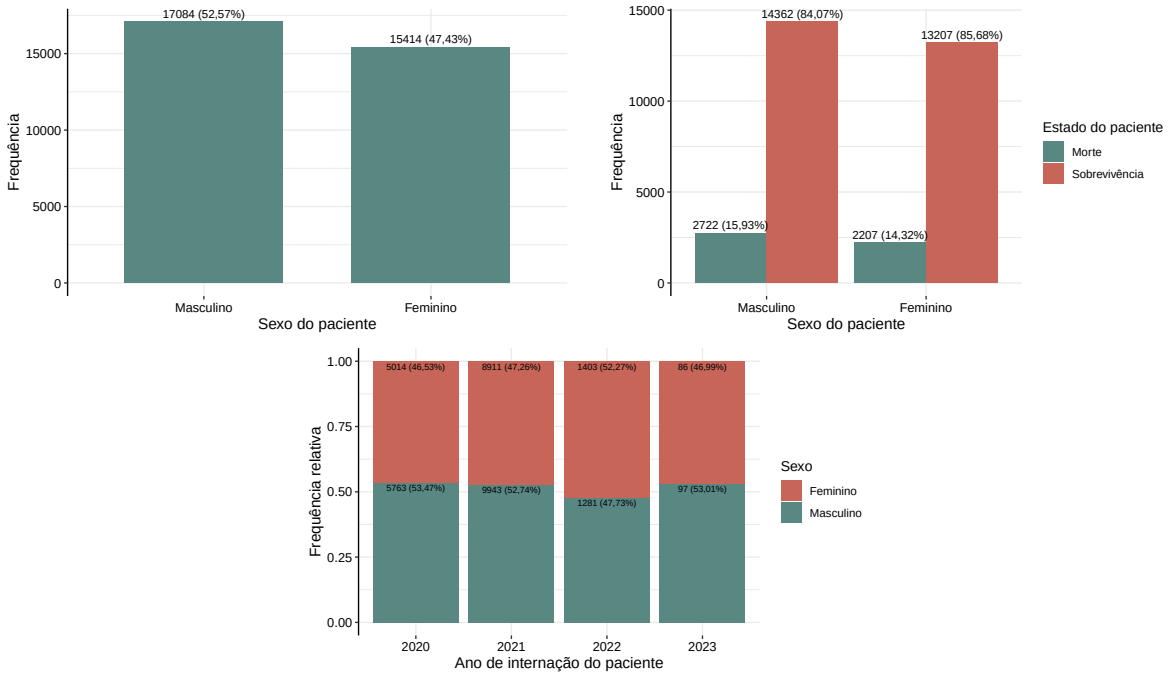

Figura 7: Gráficos de colunas do número de pacientes internados por covid-19 por sexo (superior à esquerda), número de mortes e sobreviventes por sexo (superior à direita) e internações por sexo no período de 2020 a 2023 (inferior)

No segundo gráfico da Figura 7 percebe-se que 15,93% dos homens hospitalizados por covid-19 vieram a falecer, enquanto 14,32% das mulheres hospitalizadas vieram à óbito.

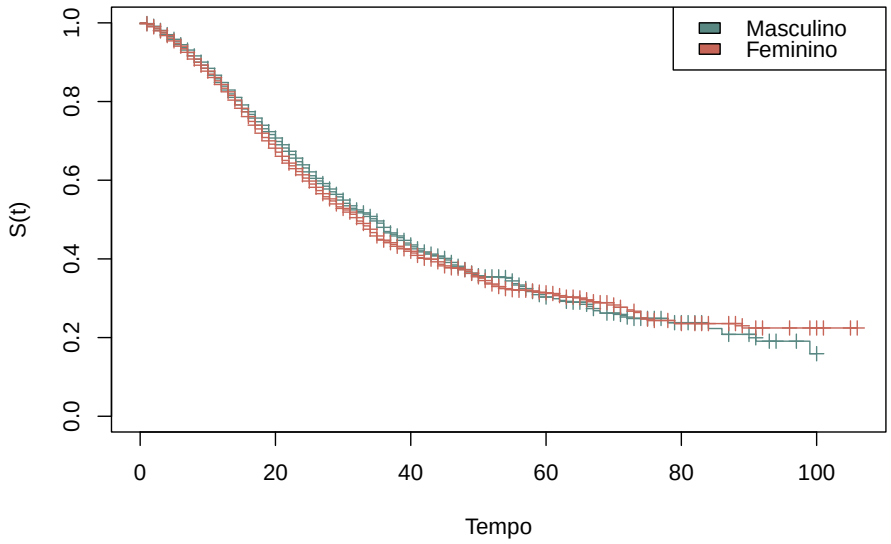

Figura 8: Gráfico de Kaplan-Meier do banco de covid-19 por sexo com falha sendo a morte do paciente

Pela Figura 8 percebe-se que dependendo do instante de tempo, a probabilidade

de sobrevivência entre os sexos se altera, sendo maior para o sexo masculino em alguns períodos e maior para o sexo feminino em outros. Isto se caracteriza por curvas em que a suposições de riscos proporcionais não é verificada. Dessa forma, para saber se há diferenças estatisticamente significativas entras as curvas, realizou-se o teste de Wilcoxon com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não existem diferenças entre as curvas de sobrevivência do sexo do paciente.} \\ H_1 : \text{Existem diferenças entre as curvas de sobrevivência do sexo do paciente.} \end{cases}$$

Os resultados deste teste são apresentados na Tabela 4, os quais evidenciam a rejeição da hipótese nula. Deste modo, afirma-se que há diferença para os sexos entre suas curvas de sobrevivência, o que sustenta os resultados apresentados na Figura 8.

Tabela 4: Teste de Wilcoxon para a variável Sexo

Variável	Estatística de teste	Graus de liberdade	P-valor	Decisão
Sexo	5,5	1	$0,02$	Rejeita H_0

4.1.4 Valor total da AIH

Cada paciente, uma vez hospitalizado, possui uma Autorização de Internação Hospitalar (AIH) vinculada à sua hospitalização. Dessa forma, são atrelados os custos de sua estadia no hospital à AIH.

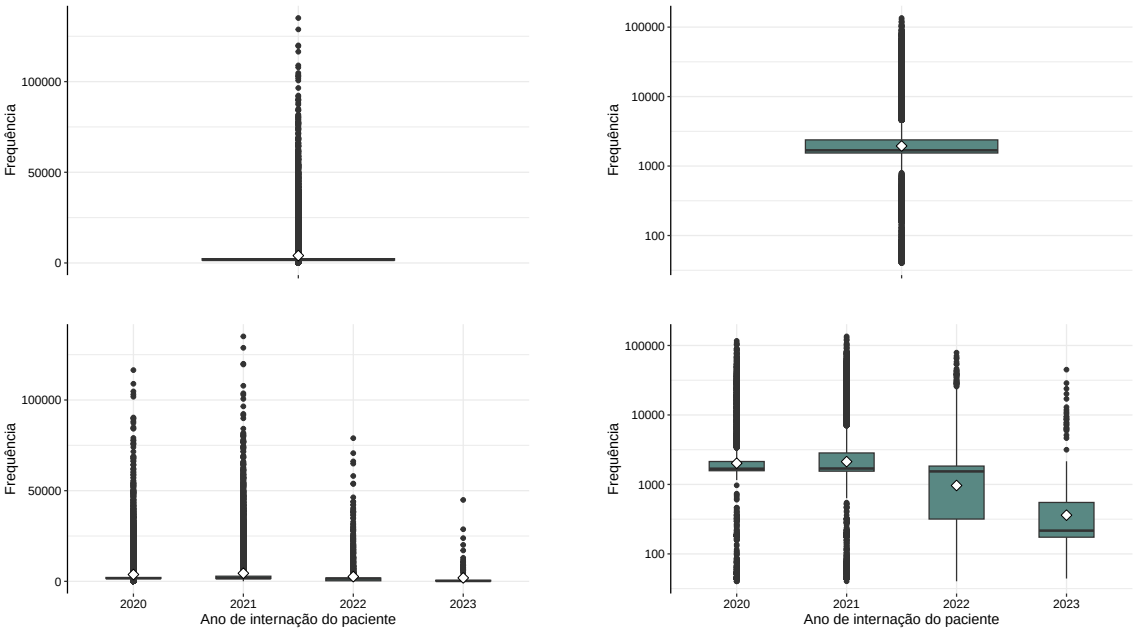

Figura 9: Gráficos *boxplot* do valor total pago da AIH em escala natural (superior à esquerda) e em escala logarítmica (superior à direita) no geral e por ano de internação do paciente escala natural (inferior à esquerda) e em escala logarítmica (inferior à direita)

No primeiro gráfico da Figura 9, percebe-se uma concentração considerável próxima ao 0, em que praticamente todos os pontos visíveis são *outliers*. Isso ocorre por conta da discrepância da escala dos valores analisados. Como forma de contornar isto, por se tratarem de dados estritamente positivos, os gráficos à direita apresentam os *boxplots* em escala logarítmica, no qual a maior parte dos valores se encontra entre 850 e 4.500.

Os gráficos da linha inferior mostram as mesmas informações, mas por ano. Novamente, no gráfico à esquerda praticamente não é possível identificar o comportamento da variável. Já no gráfico à direita em escala logarítmica, nota-se que os valores de 2020 e 2021 estão mais concentrados e superiores, de maneira geral, que os valores de 2022 e 2023. As menores média e mediana vistas são em 2023, enquanto em 2022 por mais que a mediana esteja próxima às de 2020 e 2021, sua média é bem inferior.

Além disso, é interessante observar o comportamento desta variável explicativa em relação à variável indicadora de falha.

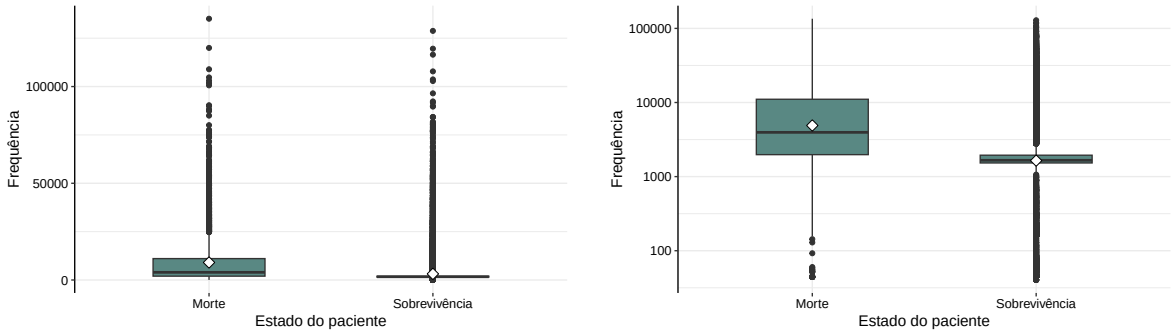

Figura 10: Gráficos *boxplot* do valor total da internação do paciente de covid-19 pelo seu estado, escalas natural (à esquerda) e logarítmica (à direita)

Da Figura 10, percebe-se que para as falhas, no caso morte do paciente, é perceptível tanto para o gráfico em escala natural quanto para escala logarítmica que os valores da AIH são superiores aos valores dos pacientes censurados. Dessa forma, vê-se que entre os pacientes que foram hospitalizados com causa principal de covid-19, os que morreram são, no geral, pacientes que custaram mais do que os que tiveram alta, ou seja, a soma do custo procedimentos médicos aplicados nos pacientes que faleceram foi, em geral, superior aos que sobreviveram.

4.1.5 Idade

Outra informação coletada dos pacientes é a idade, em anos, no momento de sua hospitalização.

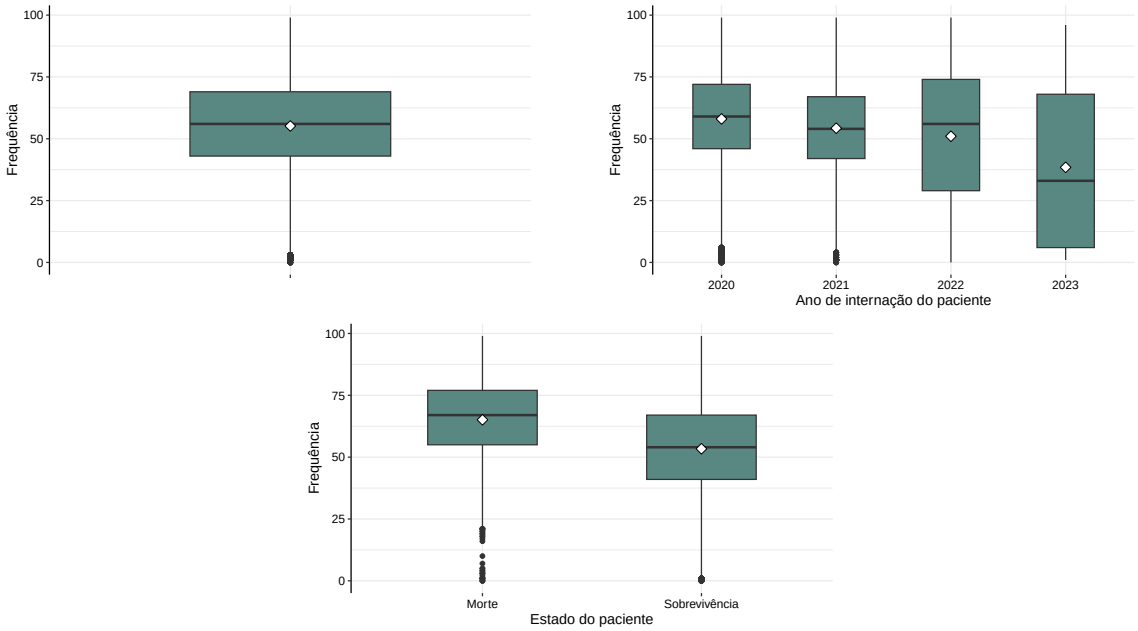

Figura 11: Gráficos *boxplot* da idade do paciente (superior à esquerda) no geral, por ano de internação (superior à direita) e por morte ou sobrevivência do paciente (inferior)

A Figura 11 mostra que variável idade tem comportamento aparentemente simétrico, com alguns *outliers* na parte inferior do *boxplot*. Além disso, sua média e mediana são próximas, com valores acima de 50 anos. Entretanto, quando analisada para cada ano do estudo, percebe-se que seu comportamento difere bastante. Para os anos de 2020 e 2021, a idade dos pacientes apresenta médias mais elevadas que nos anos seguintes. Além disso, nos anos de 2022 e principalmente 2023 percebe-se maior variabilidade, dada por conta da quantidade de dados para estes anos.

Outro gráfico importante é o da idade do paciente pela sua morte ou sobrevivência. É notório como o *boxplot* da idade dos pacientes que vieram a falecer está mais elevado que o dos que sobreviveram. Com todos os quartis e média superiores, o gráfico indica que dos pacientes hospitalizados por covid-19, a idade dos que faleceram, em geral, é maior do que a idade dos que tiveram alta.

4.2 Modelagem

Para iniciar o processo de modelagem é necessário definir a distribuição de probabilidade que melhor se ajusta ao banco de dados em relação a variável resposta. Para auxiliar nessa decisão, primeiramente analisam-se os gráficos TTT e $\hat{H}(t)$, apresentados na Figura 12.

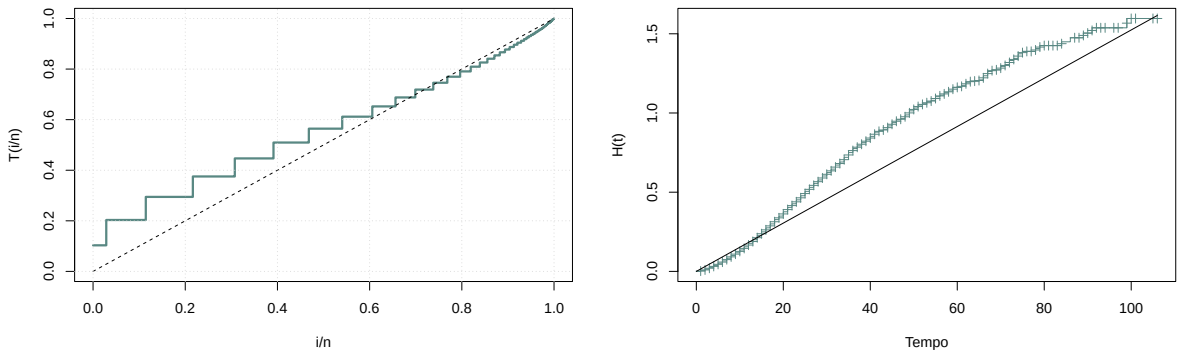

Figura 12: Gráficos TTT (à esquerda) e $\hat{H}(t)$ (à direita) para morte por covid-19 no DF

O gráfico TTT, à esquerda da Figura 12, apresenta comportamento côncavo, seguido de levemente convexo. Este formato de curva é característico de funções de risco unimodais, como a Log-Logística e a Log-Normal. Já o gráfico do risco acumulado $\hat{H}(t)$ apresenta comportamento levemente convexo, seguido de curvatura côncava. Este comportamento também indica funções de risco unimodais.

Por o comportamento côncavo no gráfico do $\hat{H}(t)$, à direita da Figura 12 ter sido suave, também será considerada a distribuição Weibull, por esta ser bem flexível em relação à sua função de risco, que pode assumir valores decrescentes. Por não apresentar comportamento alinhado à reta diagonal, percebe-se que a distribuição Exponencial se ajustará bem aos dados analisados.

Outra ferramenta de auxílio na definição da distribuição é a análise gráfica do Kaplan-Meier juntamente com as distribuições escolhidas. Inicialmente, serão analisadas as distribuições Exponencial, Weibull, Log-Logística e Log-Normal. Essas distribuições possuem até dois parâmetros.

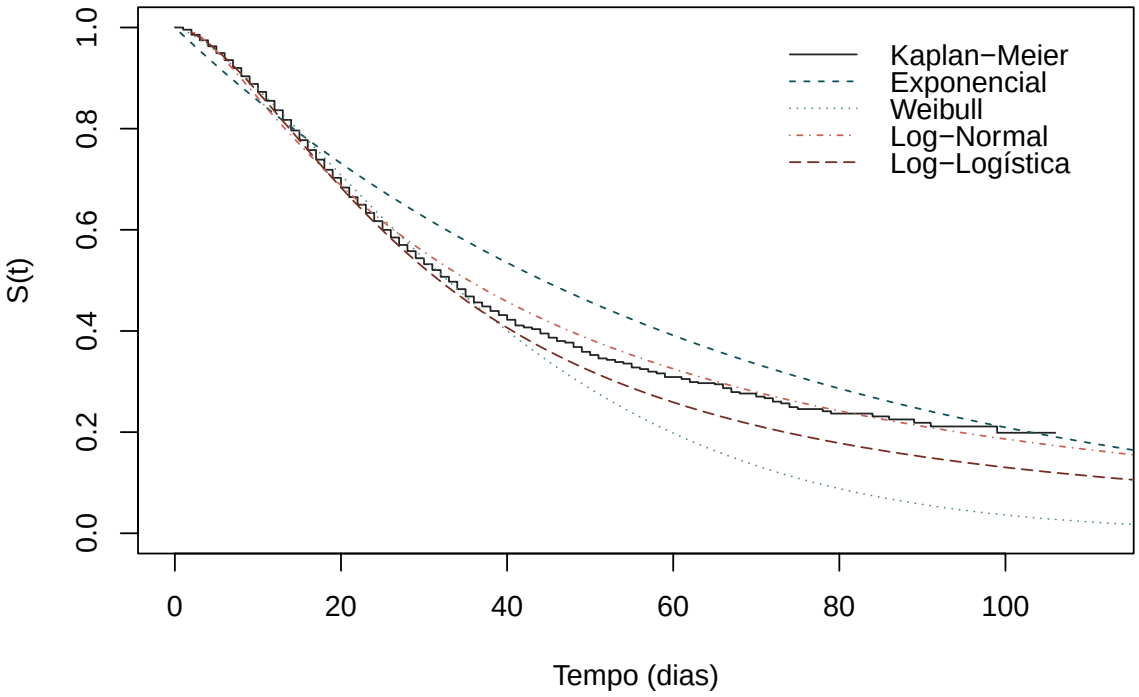

Figura 13: Gráfico de Kaplan-Meier com as distribuições estudadas do banco de covid-19 por ano com censura sendo a alta e falha a morte do paciente

Pela Figura 13, é possível descartar com facilidade as distribuições Exponencial e Weibull pela falta de ajuste às estimativas de Kaplan-Meier. Percebe-se que a distribuição Log-Logística modela bem a parte inicial do gráfico, mas subestima as estimativas de Kaplan-Meier a partir do tempo 40 dias. Dessa forma, há indícios de que a distribuição que melhor se adequa aos dados é a distribuição Log-Normal.

Além dos critérios de seleção gráficos, existem também critérios numéricos, nos quais quanto menor o valor da métrica obtida, melhor é a distribuição para o conjunto de dados.

Tabela 5: Critérios de informação para as distribuições Exponencial, Weibull Log-Normal e Log-Logística

Distribuição	AIC	AICc	BIC
Exponencial	50827,28	50827,28	50835,67
Weibull	49766,95	49766,95	49783,73
Log-Normal	49539,71	49539,71	49556,48
Log-Logística	49508,91	49508,91	49525,68

Para os três critérios obtidos na Tabela 5, os menores valores são vistos para as distribuições Log-Normal e Log-Logística, com pouca diferença entre elas. Dessa forma, por mais que os critérios de seleção numéricos tenham indicado a distribuição Log-Logística, por causa da sua proximidade numérica da distribuição Log-Normal na Tabela 5 e o ajuste visto na Figura 13, a distribuição a ser modelada será a Log-Normal.

Além das quatro distribuições examinadas, existem outras distribuições que possuem risco unimodais e que podem apresentar bom ajuste aos dados. Dessa forma, foram desenvolvidas manualmente no *software* R as distribuições: Inversa Gaussiana Reparametrizada, Burr-XII, Kumaraswamy-Log-Logística, Kumaraswamy-Log-Normal, Kumaraswamy-Inversa-Gaussiana-Reparametrizada e Kumaraswamy-Burr-XII.

De forma a se estimar os parâmetros de cada distribuição, foram considerados como valores iniciais os resultados encontrados nas distribuições de base, isto é, para a Inversa Gaussiana Reparametrizada foram utilizados como chutes iniciais de μ e σ os valores estimados na distribuição Log-Normal. Para a distribuição Burr-XII, foram considerados os estimadores de α e γ encontrados na distribuição Log-Logística, utilizando $k = 1$, já que desta forma tem-se uma como caso particular da outra. Para as distribuições Kumaraswamy foram utilizados os valores das respectivas distribuições, considerando a e b de forma a haver convergência na estimação.

Sendo assim, foram gerados gráficos contendo as distribuições de interesse juntamente ao Kaplan-Meier de forma a entender o ajuste destas aos dados.

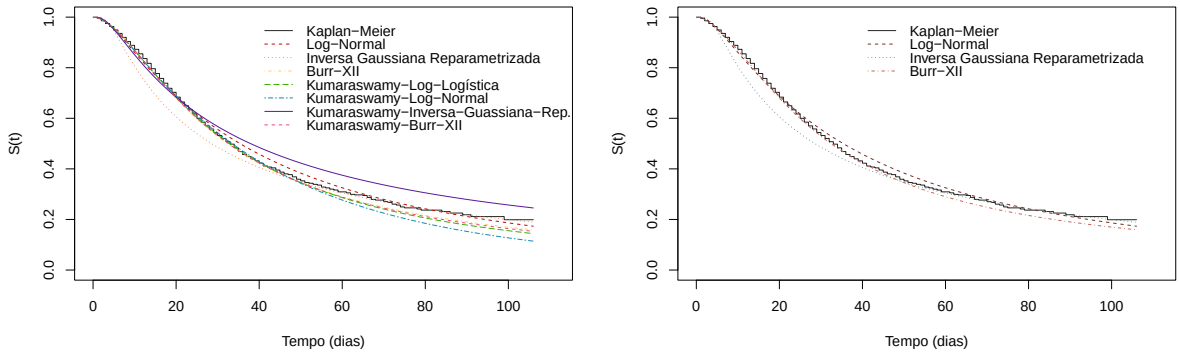

Figura 14: Gráfico de Kaplan-Meier com as distribuições Log-Normal, Inversa Gaussiana Reparametrizada, Burr-XII e da família Kumaraswamy completo (à esquerda) e selecionadas as melhores ajustadas (à direita)

A Figura 14, à esquerda, mostra a comparação do ajuste das distribuições em relação às estimativas de Kaplan-Meier. Percebe-se que a distribuição Inversa Gaussiana Reparametrizada, em laranja, subestima de forma considerável a parte inicial da curva de Kaplan-Meier, oferecendo excelente ajuste apenas na segunda metade do gráfico. A dis-

tribuição Kumaraswamy-Inversa-Gaussiana-Reparametrizada apresenta comportamento inverso, que ajusta com precisão a primeira parte do gráfico, superestimando os valores subsequentes. As outras distribuições Kumaraswamy apresentam comportamento parecido, ajustando bem no início e subestimando os valores finais. O melhor ajuste entre essas foi da distribuição Kumaraswamy-Burr-XII, que ainda assim apresentou comportamento levemente inferior à distribuição Burr-XII, em amarelo, que menos subestimou os valores finais em relação às anteriores.

Deste modo, a Figura 14, à direita, mostra o comportamento das duas distribuições que mais se destacaram na análise anterior, juntamente com a distribuição Log-Normal. É possível perceber, ao analisar a curva de Kaplan-Meier como um todo, que a distribuição Burr-XII apresentou comportamento mais robusto na modelagem do tempo em estudo.

Além do critério visual, também foram calculados os valores dos critérios de informação para as distribuições estuadas, de forma a validar a escolha feita anteriormente.

Tabela 6: Critérios de informação para as distribuições Log-Normal, Inversa Gaussiana Reparametrizada, Burr-XII e da família Kumaraswamy

Distribuição	AIC	AICc	BIC
Log-Normal	49539.71	49539.71	49556.48
Inversa Gaussiana Reparametrizada	51350.17	51350.17	51366.95
Burr-XII	49493.05	49493.05	49518.21
Kumaraswamy-Log-Logística	49484.96	49490.96	49524.51
Kumaraswamy-Log-Normal	49484.96	49484.97	49518.52
Kumaraswamy-Inversa-Gaussiana-Rep.	49806.49	49806.49	49840.04
Kumaraswamy-Burr-XII	49495.33	49495.33	49537.27

Com exceção das distribuições Inversa Gaussiana Reparametrizada e Kumaraswamy-Inversa-Gaussiana-Reparametrizada, todas as outras apresentaram valores notadamente próximos, podendo alguns serem considerados empates técnicos. Segundo o AIC, as melhores distribuições são Kumaraswamy-Log-Logística e Kumaraswamy-Log-Normal. Já para o AICc, a que melhor ajustou foi a Kumaraswamy-Log-Normal. Por fim, a distribuição que apresentou menor BIC foi a Burr-XII. É notório que tanto a distribuição Kumaraswamy-Log-Normal quanto a Burr-XII apresentaram valores inferiores que a distribuição Log-Normal para todas as medidas.

Além das técnicas gráficas e de critérios de informação para selecionar a melhor distribuição, ainda existe um método complementar, feito por meio do teste de razão de verossimilhança. Podendo ser realizado apenas entre distribuições encaixadas, o teste compara as verossimilhanças dos modelos ajustados com as distribuições de interesse,

resultando na escolha da distribuição com mais parâmetros ou na mais restrita. Testes foram realizados em todas as distribuições encaixadas possíveis, para obter-se resultado mais fundamentado para escolha da distribuição final.

Os testes aplicados possuem as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A distribuição mais restrita é a mais adequada (com menos parâmetros).} \\ H_1 : \text{A distribuição mais completa é a mais adequada (com mais parâmetros).} \end{cases} \tag{4.2.1}$$

Percebe-se pela Tabela 11 que em cinco dos seis testes realizados, o melhor modelo foi o com mais parâmetros. Desta forma, a distribuição Burr-XII performou melhor que a Log-Logística, assim como as Kumaraswamy performaram melhor que todas as suas distribuições relacionadas, exceto um caso da Kumaraswamy-Burr-XII. Neste teste percebe-se que o p-valor foi superior ao nível de significância de 5%, informando que a distribuição Burr-XII ficou melhor ajustada aos dados que a distribuição Kumaraswamy-Burr-XII.

Tabela 7: Resultados dos Testes de Razão de Verossimilhança entre distribuições completas e restritas, em cada caso de distribuições encaixadas

Distribuição completa (H_1)	Distribuição restrita (H_0)	TRV	Graus de liberdade	P-valor	Decisão do teste
Burr-XII	Log-Logística	17,86	1	$2,4 \times 10^{-5}$	Rejeita H_0
Kum-Inv-Gau-Rep	Inv-Gau-Rep	1547,69	2	$< 2 \times 10^{-16}$	Rejeita H_0
Kum-Log-Logística	Log-Logística	21,95	2	$1,7 \times 10^{-5}$	Rejeita H_0
Kum-Log-Normal	Log-Normal	58,74	2	$1,7 \times 10^{-13}$	Rejeita H_0
Kum-Burr-XII	Burr-XII	1,72	2	0,42	Não rejeita H_0
Kum-Burr-XII	Log-Logística	19,58	3	2×10^{-4}	Rejeita H_0

Isso posto, foi levado em consideração que o ajuste da distribuição Kumaraswamy-Log-Normal na Figura 14, em azul, subestimou os valores finais do Kaplan-Meier consideravelmente mais que a distribuição Burr-XII, além da distribuição Burr-XII possuir um parâmetro a menos que a distribuição Kumaraswamy-Log-Normal, obtendo menor BIC. Por conseguinte, também dado os testes de razão de verossimilhança empregados, foi escolhida como distribuição final a distribuição Burr-XII.

4.2.1 Seleção de Variáveis

A criação do modelo que descreve o comportamento da variável resposta passa pela seleção das variáveis explicativas que integrarão o modelo final. Dessa forma, existem quatro variáveis candidatas a entrarem no modelo, sendo elas o ano de internação do paciente, o sexo do paciente, o valor total da AIH e a idade do paciente.

Como forma de facilitar a comparação entre os períodos de estudo, levando em consideração a quantidade inferior de dados de 2023 e os testes apresentados na Seção 4.1.2, a variável ano será recategorizada em período pandêmico, anos de 2020 e 2021, e período "pós-pandêmico", anos de 2022 e 2023. Deste modo, o período pandêmico abarca os anos em que houve o pico tanto de casos quanto de mortes pela doença. Já o período "pós-pandêmico" compreende os anos seguintes ao pico, em que o número de casos da doença estava mais contido que no período anterior, mas ainda durante a pandemia, já que a OMS declarou o fim da pandemia apenas em 5 de maio de 2023. O comportamento desta nova variável é mostrado na Figura 15.

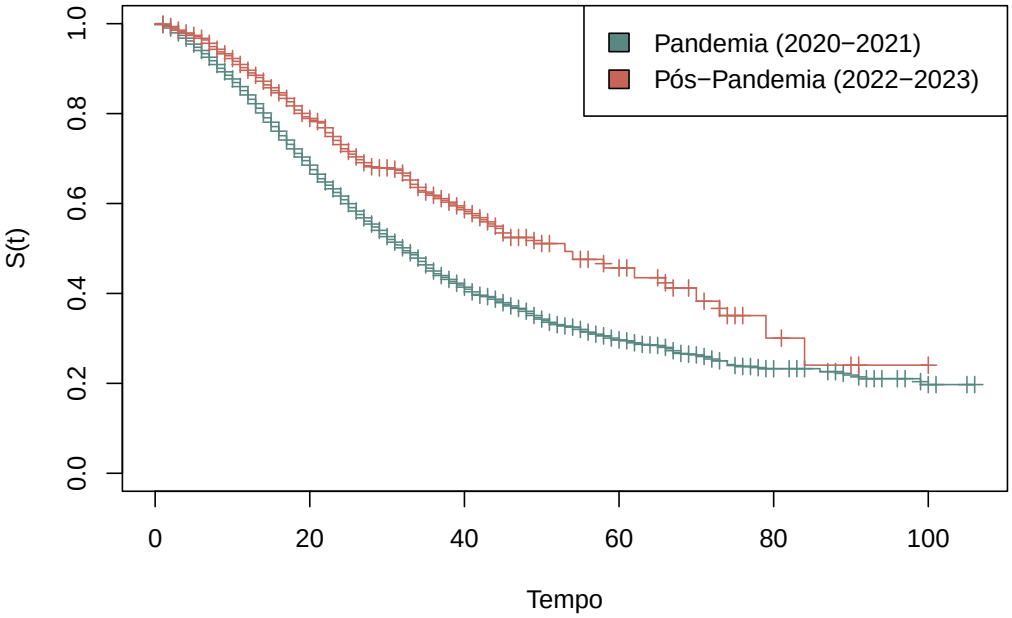

Figura 15: Gráfico de Kaplan-Meier do banco de covid-19 por período com censura sendo a alta e falha a morte do paciente

Percebe-se, pela Figura 15, que a curva de sobrevivência do período "pós-pandemia""é constantemente superior ao do período da pandemia. Desta forma, entende-se que as curvas possuem riscos proporcionais, por não se cruzarem durante todo o período de tempo estudado. Dessa forma, aplica-se o teste de logRank, de acordo com as hipóteses:

$$\begin{cases} H_0 : \text{Não existem diferenças entre as curvas de sobrevivência do sexo do paciente.} \\ H_1 : \text{Existem diferenças entre as curvas de sobrevivência do sexo do paciente.} \end{cases}$$

Os resultados deste teste são apresentados na Tabela 8, pela qual afirma-se que há diferença significativa para os períodos entre suas curvas de sobrevivência, mostrando

que esta separação entre os anos é eficaz em captar diferentes comportamentos das curvas de sobrevivência.

Tabela 8: Teste de logRank para a variável Período

Variável	Estatística de teste	Graus de liberdade	P-valor	Decisão
Período	55,1	1	$< 0,001$	Rejeita H_0

Para fazer a seleção de variáveis, será executado o procedimento indicado e descrito por Colosimo e Giolo (2006), o qual sugere um método manual de seleção, no qual tanto o estatístico quanto o profissional da área possuem controle na entrada e saída das variáveis, de acordo com seus interesses e conhecimentos. Desta forma, serão elencadas as etapas do processo de seleção:

1. Ajustar todos os modelos com as covariáveis isoladamente. Selecionar aquelas significativas ao nível de 10% de significância.

2. Ajustar conjuntamente um modelo com todas as covariáveis significativas no passo 1. Em seguida, excluir do modelo as covariáveis que, conjuntamente, não são significativas, uma de cada vez, a fim de constatar a significância no modelo.

3. Retirar as covariáveis que restaram no passo 2, uma a uma, a fim de verificar se alguma delas pode ser retirada. Nesta etapa, o Teste da Razão de Verossimilhanças é recomendado para confirmar se o modelo com a variável é viável.

4. Com as variáveis restantes do passo 3, incluir as variáveis não significativas no passo inicial e verificar a possibilidade de inclusão de alguma delas. Novamente o uso do TRV é recomendado.

5. Por fim, verificar a possibilidade de incluir interações duas a duas entre as covariáveis. O modelo final será composto pelas covariáveis remanescentes no passo 4 e os termos de interação significativos nesta etapa.

Dessa forma, serão realizados dois ajustes. Primeiramente, será realizado o ajuste do modelo com a distribuição Log-Normal e diante dos resultados, que serão apresentados, serão estimados novos modelos de regressão utilizando a distribuição Burr-XII.

- Distribuição Log-Normal

Na Tabela 9 estão apresentados os resultados dos ajustes dos modelos de variáveis individuais utilizando-se a distribuição Log-Normal, na qual é possível ver o valor da estimativa, assim como seu p-valor.

Tabela 9: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo uma variável explicativa por meio da distribuição Log-Normal

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_{Idade}	-0,0177	0,0007	-26,2042	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0489	0,0212	-2,3050	0,021
$\beta_{\text{Período "Pós-Pandemia"}}$	0,3183	0,0439	7,2444	$4,3 \times 10^{-13}$
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0641	0,0111	-5,7886	$7,1 \times 10^{-9}$

Percebe-se que ao nível de significância de 10%, todas as variáveis são significativas. Dessa forma, segue para a etapa 2, na qual será criado um único modelo com todas as variáveis da etapa anterior.

Tabela 10: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo completo das variáveis selecionadas por meio da distribuição Log-Normal

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_0	5,2011	0,1165	44,66	$< 2 \times 10^{-16}$
β_{Idade}	-0,0187	0,0007	-26,35	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0766	0,0223	-3,44	0,0006
$\beta_{\text{Período "Pós-Pandemia"}}$	0,3574	0,0471	7,58	$3,4 \times 10^{-14}$
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0561	0,0114	-4,94	$7,9 \times 10^{-7}$
$log(scale)$	0,1724	0,0106	16,29	$< 2 \times 10^{-16}$

Os p-valores indicados na Tabela 10 mostram que todas as quatro variáveis explicativas são importantes no modelo para o ajuste da variável resposta. Neste modelo é possível observar a direção da contribuição da variável explicativa para a probabilidade de sobrevivência do paciente. Nela, é possível perceber que quanto maior a idade do paciente, menor é a sua probabilidade de sobreviver à covid-19. De maneira análoga, entende-se que quanto mais foi preciso gastar com o paciente, registrado em sua AIH, menor é a sua probabilidade de sobreviver. Este resultado faz sentido no contexto hospitalar, em que quão mais grave é a situação do paciente, mais medicações e procedimentos possivelmente são aplicados voltados à sua recuperação.

Nas variáveis qualitativas percebe-se que pelo valor negativo do $\hat{\beta}$ relacionado ao sexo, percebe-se que pacientes homens têm menor probabilidade de sobrevivência que pacientes mulheres. Em contraste, o valor positivo do $\hat{\beta}$ do período "Pós-pandemia", isto é, dos anos de 2022 e 2023, mostra que pacientes hospitalizados por covid-19 nestes anos têm probabilidade maior de sobrevivência do que os hospitalizados em 2020 e 2021.

Por conseguinte, avança-se diretamente para a etapa 5, na qual foram testadas todas as interações possíveis entre variáveis quantitativas com qualitativas, das quatro

variáveis regressoras. O teste a ser utilizado é o da Razão de Verossimilhança, sob as seguintes hipóteses:

$$\begin{cases} H_0 : \text{O modelo sem interação é o mais adequado } (\beta_{\text{Interação}} = 0). \\ H_1 : \text{O modelo com interação é o mais adequado } (\beta_{\text{Interação}} \neq 0). \end{cases} \tag{4.2.2}$$

Os resultados dos testes aplicados estão dispostos na Tabela 11.

Tabela 11: Resultados dos Testes de Razão de Verossimilhança entre o modelo completo e os modelos completos com uma interação por meio da distribuição Log-Normal

Interação	Estimativa	$logL(\hat{\boldsymbol{\theta}})$	TRV	Graus de liberdade	P-valor	Decisão do teste
Sexo:Idade	0,00151	-24325,4	1,2	1	0,27	Não rejeita H_0
Sexo:log(Valor Total)	-0.02061	-24325,5	0,98	1	0,32	Não rejeita H_0
Período:Idade	-0,000975	-24325,9	0,15	1	0,69	Não rejeita H_0
Período:log(Valor Total)	0,168649	-24315,2	21,6	1	$3,4 \times 10^{-6}$	Rejeita H_0

Sendo assim, percebe-se que a única interação significativa, comparando o modelo com a interação com o modelo sem a interação, é entre a variável Período e a Valor Total. Com isso, os coeficientes do modelo final com interação estão na apresentados na Tabela 12.

Tabela 12: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo completo com interação

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_0	5,3253	0,1203	44,27	$< 2 \times 10^{-16}$
β_{Idade}	-0,0187	0,0007	-26,40	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0761	0,0223	-3,42	0,0006
$\beta_{\text{Período "Pós-Pandemia"}}$	-0,9073	0,2768	-3,28	0,001
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0711	0,0119	-5,98	$2,2 \times 10^{-9}$
$\beta_{\text{log(Valor Total da AIH):"Pós-Pandemia"}}$	0,1636	0,0355	4,60	$4,1 \times 10^{-6}$
$log(scale)$	0,1725	0,0106	16,30	$< 2 \times 10^{-16}$

O β estimado relacionado à interação entre as variáveis Valor Total da AIH e Período "Pós-pandemia" é positivo, indicando que neste período, o aumento no valor gasto com o paciente aumenta a sua probabilidade de sobrevivência, em relação ao período de 2020 e 2021.

- Distribuição Burr-XII

Utilizando as estimativas dos β's encontradas nas regressões utilizando a distri-

buição Log-Normal como valores iniciais, foram reestimados os coeficientes das regressões por meio da distribuição Burr-XII. Seguindo o mesmo passo a passo, foram ajustados modelos de regressão contendo uma das variáveis explicativas, para estudar sua significância em relação à variável resposta.

Tabela 13: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo uma variável explicativa por meio da distribuição Burr-XII

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_{Idade}	-0,0179	0,0007	-26,1939	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0468	0,0207	-2,2647	0,0235
$\beta_{\text{Período "Pós-Pandemia"}}$	0,3293	0,0443	7,4381	1×10^{-13}
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0619	0,0105	-5,9019	$3,6 \times 10^{-9}$

É notório que, similarmente ao visto na Tabela 9, todas as variáveis apresentaram valores significativos, ao nível de significância de 10%. Desta forma, prossegue-se ao passo de ajustá-las todas conjuntamente em um único modelo.

Tabela 14: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo completo das variáveis selecionadas por meio da distribuição Burr-XII

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_0	5,2570	0,1462	35,95	$< 2 \times 10^{-16}$
β_{Idade}	-0,0197	0,0007	-26,79	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0713	0,0218	-3,26	0,0011
$\beta_{\text{Período "Pós-Pandemia"}}$	0,3965	0,0478	8,29	$1,1 \times 10^{-16}$
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0706	0,0108	-6,54	6×10^{-11}
c	0,0282	0,0290	-	-
k	0,0771	0,0960	-	-

De maneira análoga ao visto na Tabela 10, todas as variáveis explicativas se mostraram significativas quando analisadas conjuntamente no modelo. Também percebe-se que os sinais das estimativas se mantém, gerando a mesma interpretação estudada anteriormente.

Tabela 15: Resultados dos Testes de Razão de Verossimilhança entre o modelo completo e os modelos completos com uma interação por meio da distribuição Burr-XII

Interação	Estimativa	$logL(\hat{\boldsymbol{\theta}})$	TRV	Graus de liberdade	P-valor	Decisão do teste
Sexo:Idade	0,00097	-24287,08	3,3	1	0,069	Não rejeita H_0
Sexo:log(Valor Total)	-0,01845	-24286,9	3,66	1	0,056	Não rejeita H_0
Período:Idade	-0,00352	-24286,48	4,50	1	0,034	Rejeita H_0
Período:log(Valor total)	0,15299	-24278,3	20,95	1	$4,7 \times 10^{-6}$	Rejeita H_0

Percebe-se que ao comparar modelos com uma interação com o modelo sem interações, duas interações se mostraram significativas isoladamente, sendo elas Período com Idade e Período com Valor total. Dessa forma, foi construído um modelo contendo ambas interações e comparou-se, por meio do teste de razão de verossimilhanças, com os modelos com apenas uma interação, gerando os resultados apresentados na Tabela 16.

Tabela 16: Resultados dos Testes de Razão de Verossimilhança entre o modelo completo com duas interações e os modelos completos com uma interação significativa por meio da distribuição Burr-XII

Interação	Estimativa	$logL(\hat{\boldsymbol{\theta}})$	TRV	Graus de liberdade	P-valor	Decisão do teste
Período:Idade	-0,0028	-24277,65	17,66	1	$2,6 \times 10^{-5}$	Rejeita H_0
Período:log(Valor total)	0,1520	-24277,65	1,21	1	0,27	Não rejeita H_0

A Tabela 16 mostra que ao comparar o modelo de duas interações com o modelo apenas com a interação Período:Idade, o modelo com as duas interações se mostrou mais adequado. Entretanto, comparando o modelo com duas interações com o apenas com a interação de Período:log(Valor Total), não rejeita-se H_0, evidenciando que o modelo com apenas a interação de Período:log(Valor Total) é melhor ajustado, ao nível de significância de 5%.

Com isso, o modelo final chegado por meio da distribuição Burr-XII, semelhantemente ao modelo final obtido pela Log-Normal na Tabela 12, tem os seguintes valores:

Tabela 17: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo final obtido por meio da distribuição Burr-XII

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
β_0	5,1014	0,1392	36,66	$< 2 \times 10^{-16}$
β_{Idade}	-0,0190	0,0007	-26,27	$< 2 \times 10^{-16}$
$\beta_{\text{Sexo Masculino}}$	-0,0706	0,0217	-3,25	0,0012
$\beta_{\text{Período "Pós-Pandemia"}}$	-0,8308	0,2791	-2,98	0,0029
$\beta_{\text{log(Valor Total da AIH)}}$	-0,0693	0,0111	-6,26	$3,9 \times 10^{-10}$
$\beta_{\text{log(Valor Total da AIH):"Pós-Pandemia"}}$	0,1530	0,0351	4,35	$1,3 \times 10^{-5}$
c	1,6513	0,0283	-	-
k	0,8506	0,0760	-	-

A Tabela 17 apresenta os coeficientes de regressão do modelo final, mostrando que os coeficientes das variáveis Idade e Sexo são negativos. Desta forma, entende-se que quanto mais velho o paciente, menor é a sua probabilidade de sobrevivência. Além disso, também é evidenciado que a probabilidade de sobrevivência entre os homens é inferior às mulheres hospitalizadas por covid-19. Ademais, o coeficiente da interação entre o período "pós-pandemia" e o log(valor total) é positivo, indicando que quanto maior o gasto com o paciente no período "pós-pandemia", maior a probabilidade de sobreviver do que pacientes no período de pandemia.

4.3 Análise de Resíduos

Uma vez definidos os modelos de cada distribuição, é necessário passar pela etapa de análise de resíduos de forma a verificar a adequabilidade do modelo aos dados.

- Distribuição Log-Normal

Primeiramente, serão analisados os modelos finais sem interação e o modelo com a interação entre Período e log(Valor total).

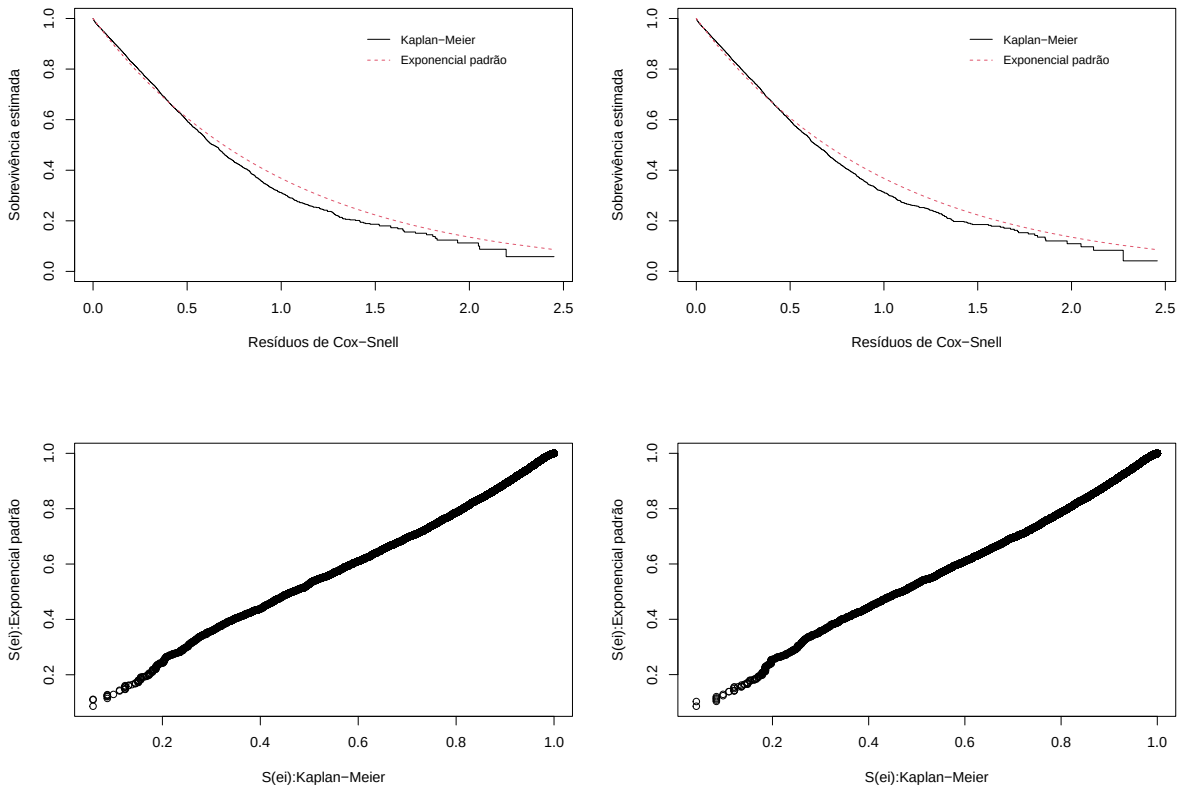

Figura 16: Curvas de sobrevivência estimadas (acima) e resíduos de Cox-Snell estimados por Kaplan-Meier e pelo modelo Exponencial padrão (abaixo) do modelo de regressão Log-Normal completo sem interação (esquerda) e modelo de regressão Log-Normal completo com interação (direita)

A Figura 16 evidencia o bom ajuste da distribuição Log-Normal nos dois modelos, em que, acima, percebe-se a proximidade da curva à da Exponencial padrão, com o ajuste do modelo com interação sendo levemente mais apropriado, ao observar os valores finais da curva. Além disso, os gráficos de resíduos de Cox-Snell mostram que se formou praticamente linhas retas, com apenas alguns desvios na parte inicial do gráfico.

No geral, percebe-se que a distribuição apresentou um bom ajuste aos dados, mas com leve divergência em ambos modelos, observando os gráficos das curvas de sobrevivência estimadas.

- Distribuição Burr-XII

Para a distribuição Burr-XII foram considerados os modelos sem interação e com a interação entre Período e Valor total.

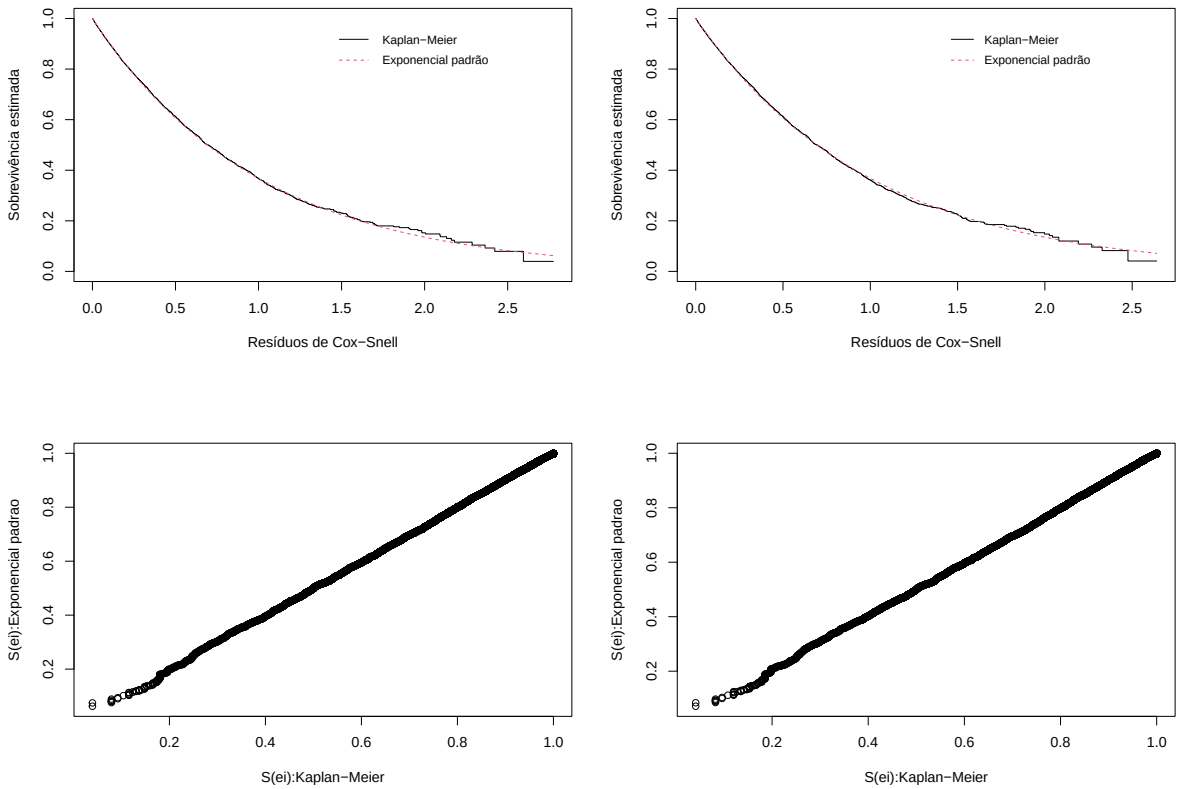

Figura 17: Curvas de sobrevivência estimadas (acima) e resíduos de Cox-Snell estimados por Kaplan-Meier e pelo modelo Exponencial padrão (abaixo) do modelo de regressão Burr-XII completo sem interação (esquerda) e modelo de regressão Burr-XII completo com interação (direita)

É perceptível que o ajuste da distribuição Burr-XII é consideravelmente mais adequado que o da distribuição Log-Normal, principalmente considerando o gráfico de curvas de sobrevivência estimadas. Nestes gráficos, na Figura 17, percebe-se que ambos modelos apresentaram ajustes extremamente próximos ao da curva Exponencial padrão, com a do modelo com interação sendo minimamente melhor ajustado na parte final do gráfico. Também observa-se que os gráficos de Cox-Snell apresentaram comportamentos bem consistentes com o bom ajuste.

Por conseguinte, percebe-se que a distribuição Burr-XII apresentou melhor adequabilidade de ajuste em relação à análise de resíduos.

5 Conclusão

Neste trabalho foram propostos modelos de regressão paramétricos de forma a modelar o tempo até a morte de pacientes hospitalizados com a causa principal de covid-19, de 2020 a 2023 no Distrito Federal. Foram consideradas variáveis idiossincráticas dos pacientes para entender os fatores que influenciassem no tempo até a falha.

Portanto, considerando como falha o óbito do paciente, percebeu-se que durante o período estudado, aproximadamente 15% veio a falecer, sendo um número bastante considerável, condicionando a maior parte da amostra como censura. O gráfico da função sobre o estimador de Kaplan-Meier evidencia que com o aumento dos dias de internação, a probabilidade de sobrevivência decai, com tempo mediano de internação de 32,8 dias. A maior parte da amostra em estudo estavam concentradas nos anos de 2020 e 2021, anos em que a procura hospitalar por conta da covid-19 era de fato altíssima, representando mais de 90% da amostra. Ainda na comparação entre os anos estudados, percebe-se uma clara diferença entre a porcentagem de mortos de cada ano, que em 2020 e 2021 foram por volta de 15%, enquanto para 2022 e 2023 esteve em 9,5% e 6,56%, respectivamente. Testes de comparação de curvas de Wilcoxon encontraram diferenças entre os anos e o teste de logRank, ao comparar os anos de 2020 e 2021 com 2022 e 2023 também apresentou diferença significativa, ao nível de significância de 5%, indicando que o período foi um fator influenciador na probabilidade de morte.

Outro fator estudado foi o sexo do paciente, o qual apresentou curvas parecidas ao longo do gráfico da função sobre o estimador de Kaplan-Meier, mas não suficiente para o teste de Wilcoxon não rejeitar a hipótese de igualdade das curvas de sobrevivência, ao nível de significância de 5%. Além disso, ao se estudar o valor total da AIH do paciente, encontrou-se uma distribuição dos dados extremamente assimétrica, sendo considerado no estudo a transformação logarítmica da variável. No gráfico, em escala logarítmica, é perceptível como a média e a mediana dos pacientes que faleceram foi superior aos que sobreviveram. O último fator analisado foi a idade, a qual já na análise descritiva, por meio do boxplot, percebe-se que a média de idade dos indivíduos que faleceram era superior ao dos que sobreviveram, indicando relação entre a variável resposta e esta covariável.

Deste modo, na investigação da distribuição de sobrevivência que melhor descrevesse a variável resposta, estudou-se primordialmente distribuições com funções de riscos unimodais. Entre as distribuições com até dois parâmetros, foi escolhida a distribuição Log-Normal, que apresentou bom ajuste da função sobre o estimador de Kaplan-Meier, porém, com leve desvio no centro do gráfico ao superestimar o valor ajustado dos dados. Entre as distribuições mais complexas, de três à cinco parâmetros, foi escolhida a distribuição Burr-XII pelo seu ajuste gráfico da função sobre o estimador de Kaplan-Meier,

métricas de informação de AIC, AICc e BIC, além de testes de razão da verossimilhanças entre distribuições encaixadas.

No processo de seleção de variáveis, percebeu-se que para ambas distribuições, todas as variáveis do estudo se mostraram significativas, ao nível de significância de 5%. Deste modo, para ambas distribuições, o modelo final obteve não apenas as quatro variáveis, mas também interação entre o logarítmico do valor total da AIH e o período da internação. Com estimativas próximas, viu-se que a idade interfere na probabilidade de falha, na qual quanto mais velho o paciente, menor sua probabilidade de sobrevivência. Além disso, também viu-se pelo coeficiente negativo do sexo que pacientes do sexo masculino tiveram menor probabilidade de sobrevivência que as mulheres. Outro ponto de destaque foi a interação entre o logarítmico do valor total da AIH e o período da internação. Foi visto que, para o período da pandemia, considerado dos anos de 2020 e 2021, quanto maior o valor do logaritmo do total da AIH, menor era a probabilidade de sobrevivência. Já para o período pós-pandêmico, sendo 2022 e 2023, a probabilidade de sobrevivência aumenta a medida que aumentava o o valor do logaritmo do total da AIH.

A análise de resíduos mostrou como a distribuição Burr-XII possui melhor adequabilidade ao apresentar curva de resíduos mais consistente com a distribuição Exponencial padrão do que a distribuição Log-Normal.

Deste modo, o estudo evidencia a diferença da probabilidade de sobrevivência entre 2020 e 2021 com 2022 e 2023, anos em que a busca hospitalar por causa primária de covid-19 foi bem distinta, além de fatores que modificam o risco, como sexo, idade e o quanto foi gasto na estadia do paciente no hospital. Dessa forma, reforça-se a importância de medidas preventivas e protetivas de doenças virais transmissíveis, de forma a que não haja contaminação em massa que sobrecarregue sistemas de saúde e coloque os infectados em situação ainda mais vulnerável.

Por conseguinte, como proposta para trabalhos futuros, sugere-se a utilização de modelos de sobrevivência com fração de cura, já que este trabalho abarca indivíduos acompanhados por um longo período de tempo, com certa fração deles não experimentando o evento de interesse, estando, de certa forma, imunes à falha. Outro ponto de destaque é que os anos utilizados, de 2020 a 2023, possuem não apenas quantidade de dados distintas, mas também comportamentos nas curvas de sobrevivência diferenciados entre si. Deste modo, uma sugestão é realizar os quatro anos estudados como variável explicativa, considerando o ano de 2023 como referência e os demais anos como variáveis *dummies* na regressão ou realizar a modelagem separada por ano, para entender de maneira mais específica a relação da variável resposta com as variáveis explicativas em cada fase da pandemia.

Referências

AARSET, M. V. How to identify a bathtub hazard rate. *IEEE transactions on reliability*, IEEE, v. 36, n. 1, p. 106–108, 1987.

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, v. 63, n. 2020, p. 10, 2004.

BOUSQUAT, A. et al. Pandemia de covid-19: o sus mais necessário do que nunca. *Revista USP*, n. 128, p. 13–26, 2021.

BRESLOW, N.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics*, JSTOR, p. 437–453, 1974.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006.

CORDEIRO, G. M.; CASTRO, M. de. A new family of generalized distributions. *Journal of statistical computation and simulation*, Taylor & Francis, v. 81, n. 7, p. 883–898, 2011.

COX, D. R.; OAKES, D. *Analysis of survival data*. [S.l.]: CRC press, 1984. v. 21.

EUGENE, N.; LEE, C.; FAMOYE, F. Beta-normal distribution and its applications. *Communications in Statistics-Theory and methods*, Taylor & Francis, v. 31, n. 4, p. 497–512, 2002.

HASHIMOTO, E. M. et al. The re-parameterized inverse gaussian regression to model length of stay of covid-19 patients in the public health care system of piracicaba, brazil. *Journal of Applied Statistics*, Taylor & Francis, v. 50, n. 8, p. 1665–1685, 2023.

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

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.

LAWLESS, J. Statistical methods and model for lifetime data. *Wiley&Sons, New York*, v. 52, 1982.

LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. [S.l.]: John Wiley & Sons, 2003. v. 476.

MAI, F.; PINTO, R. D.; FERRI, C. Covid-19 and cardiovascular diseases. *Journal of cardiology*, Elsevier, v. 76, n. 5, p. 453–458, 2020.

MINUSSI, B. B. et al. Grupos de risco do covid-19: a possível relação entre o acometimento de adultos jovens "saudáveis" e a imunidade. *Brazilian Journal of Health Review*, v. 3, n. 2, p. 3739–3762, 2020.

PRADO, B. Covid-19 in brazil:"so what?". *Lancet*, v. 395, n. 10235, p. 1461, 2020.

SCALERA, N. M.; MOSSAD, S. B. The first pandemic of the 21st century: review of the 2009 pandemic variant influenza a (h1n1) virus. *Postgraduate medicine*, Taylor & Francis, v. 121, n. 5, p. 43–47, 2009.

SILVA, G. A.; JARDIM, B. C.; LOTUFO, P. A. Mortalidade por covid-19 padronizada por idade nas capitais das diferentes regiões do brasil. *Cadernos de Saúde Pública*, SciELO Public Health, v. 37, p. e00039221, 2021.

SILVA, G. O. *Modelos de regressão quando a função de taxa de falha não é monótona e o modelo probabilístico beta Weibull modificada*. Tese (Doutorado) — Universidade de São Paulo, 2008.

TWEEDIE, M. C. Statistical properties of inverse gaussian distributions. i. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 362–377, 1957.

ZIMMER, W. J.; KEATS, J. B.; WANG, F. The burr xii distribution in reliability analysis. *Journal of quality technology*, Taylor & Francis, v. 30, n. 4, p. 386–394, 1998.