UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

VICTOR HUGO MENDES SILVA

ANÁLISE DO TEMPO DE VIDA EM MULHERES COM CÂNCER DE MAMA NO ESTADO DE SÃO PAULO

PRESIDENTE PRUDENTE

2023

VICTOR HUGO MENDES SILVA

ANÁLISE DO TEMPO DE VIDA EM MULHERES COM CÂNCER DE MAMA NO ESTADO DE SÃO PAULO

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso II.
Orientador: Prof. Dr. Mário Hissamitsu Tarumoto.

PRESIDENTE PRUDENTE
2023

Silva, Victor Hugo Mendes
S586a Análise do tempo de vida em mulheres com câncer de mama no estado de São Paulo / Victor Hugo Mendes Silva. -- Presidente Prudente, 2023
88 p.

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

1. Análise de Sobrevivência. 2. Câncer de mama. 3. Modelo de Cox. I. Título.

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

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

VICTOR HUGO MENDES SILVA

ANÁLISE DO TEMPO DE VIDA EM MULHERES COM CÂNCER DE MAMA NO ESTADO DE SÃO PAULO

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

Orientador: _____

Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística

Orientador: _____

Prof. Dr. Klaus Schlünzen Junior
Departamento de Estatística

Orientador: _____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

Presidente Prudente, 28 de Janeiro de 2023.

RESUMO

O câncer de mama é o tipo de câncer mais comum em mulheres no Brasil e no mundo. Somente no estado de São Paulo foram estimados 18.280 casos, no ano de 2020. Neste trabalho, um dos principais objetivos, foi o de estimar o tempo de internação pelo SUS das pacientes com câncer de mama. As análises deste trabalho foram feitas através dos dados disponibilizados pelo SIHSUS (Sistema de Informações Hospitalares do SUS), no site do DATASUS, entre os anos de 2016 e 2020. O campo da estatística trabalhado foi a Análise de Sobrevivência, onde o tempo de sobrevivência foi medido a partir da internação da paciente com o câncer. Os modelos de regressão paramétricos e o modelo de Cox foram os métodos mais desenvolvidos no quesito interpretação e conclusão dos resultados, apesar de ser feita também uma análise não-paramétrica com o estimador de Kaplan-Meier e estimar as sobrevivências a partir dos modelos paramétricos: Exponencial, Weibull e Log-normal. O modelo de regressão escolhido possui distribuição Log-normal, com as seguintes covariáveis inseridas: idade da paciente, tipo de UTI utilizada, IDHM da cidade do hospital em que foi internada e a região da neoplasia. A partir das estimativas dos coeficientes do modelo foi concluído que quanto maior for a idade da paciente, menor será sua longevidade, se o IDHM da cidade do hospital for maior, a sobrevivência da paciente tende a ser maior e possuir lesão invasiva é tipo mais grave do câncer de mama. Para o modelo de Cox, foram selecionadas as covariáveis: idade da paciente, IDHM da cidade do hospital em que foi internada e a região da neoplasia. Para verificar a adequação do modelo, foi feita a análise gráfica da proporção dos riscos, a análise dos resíduos de Schoenfeld, resíduos Martingal e Deviance. Os resultados dos ajustes do modelo de Cox sugerem que pacientes com mais de 80 anos de idade possuem um risco de 78% maior de vir a falecer comparado com pacientes com menos de 80 anos e que pacientes com lesão invasiva possuem um risco de falecer 28% maior em relação a pacientes com câncer de mama em outra região, como no mamilo ou na região interna.

Palavras-chave: Câncer de Mama; Análise de Sobrevivência; Modelo de regressão; Modelo de Cox.

ABSTRACT

Breast cancer is the most common type of cancer in women in Brazil and worldwide. In the state of São Paulo alone, 18.280 cases were estimated in the year of 2020. In this work, one of the main objectives was to estimate the length of hospitalization by the SUS of patients hospitalized with breast cancer. The analyzes of this work were carried out using data made available by SIHSUS (Sistema de Informações Hospitalares do SUS), on the DATASUS website, between the years 2016 and 2020. The field of statistics worked was Survival Analysis, where survival time was measured from the admission of the patient with cancer. The regression models and the Cox model were the most developed methods in terms of interpretation and conclusion of the results, although a non-parametric analysis was also carried out with the Kaplan-Meier estimator and survival estimates from the parametric models: Exponential, Weibull and Log-normal. The regression model chosen has a Log-normal distribution, with the following covariates inserted: patient's age, type of ICU used, HDI of the city of the hospital where she was admitted and the region of the neoplasm. Based on the estimates of the model's coefficients, it was concluded that the greater the age of the patient, the shorter her longevity, if the HDI of the city where the hospital is located is higher, the patient's survival tends to be longer and having an invasive lesion is the more dangerous type of breast cancer. For the Cox model, the following covariates were selected: patient's age, HDI of the city of the hospital where she was admitted and the region of the neoplasm. To verify the adequacy of the model, a graphical analysis of the proportion of risks, the analysis of Schoenfeld's residuals, Martingal and Deviance residues was performed. The results of the Cox model fit suggest that patients over 80 years of age have a 78% greater risk of dying compared to patients younger that 80 years and that patients with an invasive lesion have a 28% greater risk of dying in relation to patients with breast cancer in another region, such as the nipple or internal region.

Keywords: Breast Cancer; Survival Analysis; Regression Model; Cox Model.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O câncer de mama é uma doença caracterizada pela multiplicação descontrolada de células da mama, gerando células anormais e formando um tumor. O câncer de mama hoje é um problema relevante de saúde pública, tendo sido estimados 2,1 milhões de casos novos de câncer e 627 mil óbitos pela doença, segundo as estatísticas mundiais do Globocan 2018 (BRAY et al., 2018), sendo a neoplasia maligna mais incidente entre as mulheres.

De acordo com o Instituto Nacional de Câncer (INCA) em *Dados e Números Sobre Câncer de Mama* (INCA, 2019), para cada ano do triênio 2020-2022, o número de novos casos estimados de câncer de mama feminina no Brasil é de 66.280. Para o ano de 2020, somente no estado de São Paulo foram estimados 18.280 casos, com uma taxa de aproximadamente 78 a cada 100 mil mulheres. O INCA afirma que a incidência do câncer de mama tende a aumentar a partir dos 40 anos, assim como a mortalidade. Ainda de acordo com o relatório, o risco de a mulher vir a óbito é 10 vezes maior quando se tem 60 anos em comparação a menos de 40 anos de idade.

A probabilidade de desenvolvimento do câncer de mama pode aumentar devido a condições individuais, como excesso de gordura corporal; estilo de vida, com pouca atividade física, alto consumo de bebida alcoólica e uso de tabaco. Certos fatores de risco não podem ser alterados, como os hereditários, reprodutivos e hormonais, alguns tipos de doença benigna da mama, raça e idade, por exemplo (APOSTOLOU; FOSTIRA, 2013; LEVY-LAHAD; FRIEDMAN, 2007).

Os fatores prognósticos, que são características do paciente ou do tumor, podem servir de preditor da sobrevida do paciente. Os principais fatores prognósticos são o tamanho do tumor juntamente com a condição dos linfonodos axilares, a idade do paciente e até mesmo o nível socioeconômico, as classes sociais mais baixas são prejudicadas, pois têm o seu diagnóstico estabelecido numa fase avançada da doença (ABREU; KOIFMAN, 2002).

Assim, nas últimas décadas o campo da Análise de Sobrevivência tem ganhado uma atenção especial, existem técnicas que fazem um papel importante nas diferentes áreas de pesquisa, e uma dessas é a área de Medicina. Um estudo de sobrevivência de mulheres com

diagnóstico de câncer de mama no município do Rio de Janeiro, analisado por Santos (2013), fazendo o uso do modelo de regressão de Cox, destaca que as covariáveis consideradas significativas ao modelo ajustado foram o tipo de tratamento e a idade da paciente.

Para este trabalho, o tempo, medido em dias, do momento que a paciente é internada na UTI com câncer de mama até sua saída da UTI, foi considerada a variável de interesse. Caso a paciente venha a falecer, é considerado como censura. No propósito de identificar fatores associados à sobrevida de pacientes com câncer de mama, a metodologia estatística utilizada foram o uso de técnicas não-paramétricas, como a estimação da curva de sobrevivência por Kaplan-Meier e suas quantidades básicas; modelos paramétricos e de regressão, com a inserção de covariáveis estatisticamente significativas para estimar a sobrevida da paciente para diferentes situações, com ênfase nas distribuições exponencial, Weibull e log-normal. E por fim, utilizado o modelo de regressão de Cox, que é um modelo semi-paramétrico e de riscos proporcionais, que possui componentes tanto paramétricos como não-paramétricos também.

A ideia central do projeto foi visualizar como se comporta a sobrevivência das pacientes com câncer de mama após sua internação, utilizando as principais metodologias no campo da Análise de Sobrevivência, sendo elas: Análise Não-Paramétrica e Paramétrica, Modelos de Regressão e Modelo de riscos proporcionais de Cox. Foi dada a atenção maior aos resultados e interpretações para o Modelo de Regressão e Modelo de Cox, pois se trata de métodos mais complexos e sofisticados.

Desta forma, este relatório está estruturado da seguinte maneira: No primeiro capítulo é apresentada a introdução ao tema do estudo. No segundo e terceiro capítulo são apresentados o objeto de estudo e as ferramentas, assim como as referências bibliográficas utilizadas. O quarto capítulo informa mais detalhes sobre a base de dados e se inicia com a análise exploratória. Os capítulos posteriores apresentam as análises não-paramétrica, modelos paramétricos, modelos de regressão, modelo de regressão de Cox e a conclusão dos resultados, respectivamente.

2 METODOLOGIA

Neste trabalho foi feita uma revisão teórica sobre análise de sobrevivência e suas técnicas estatísticas, utilizando métodos não-paramétricos, modelos paramétricos, modelos de regressão, e o uso do modelo de riscos proporcionais de Cox. Para a aplicação prática, foi usado um banco de dados disponibilizado conjuntamente pelo Ministério da Saúde e DATASUS, em Sistema de Informações do SUS, SIHSUS. Foi feito o *download* da base de dados para cada mês do ano, possuindo os nomes RDSP(ano-mês). As análises para obtenção dos resultados foram feitas no software R 3.6.1 (TEAM, 2020), por meio dos pacotes Survival versão 3.2 (THERNEAU, 2020) e flexsurv versão 1.1.1 (JACKSON, 2019).

O banco de dados é composto por 38.756 mulheres internadas e diagnosticadas com câncer de mama, registradas em hospitais do SUS (Sistema Único de Saúde) no estado de São Paulo nos anos de 2016 a 2020. Para cada paciente registrado, existe um total de 11 variáveis, sendo elas, dias de permanência na UTI, idade, caráter de internação, região da neoplasia maligna, raça, morte da paciente, e entre outras.

Dentre as 38.756 observações de mulheres internadas com a doença, a censura nos dados (morte da paciente) ocorreu em 2.443 delas, sendo aproximadamente 6,3% dos casos. As idades das pacientes seguem aproximadamente uma distribuição normal com média de 54 anos, e 84% das observações são de pacientes com idade entre 31 a 70 anos.

Diante disto, foi feita a construção de tabelas e gráficos para a análise exploratória dos dados, curva de sobrevivência não-paramétrica de Kaplan-Meier, comparação de curvas de sobrevida para cada variável e aplicar o teste *log-rank*, para não somente investigar esta diferença nas estimativas das sobrevivências ao longo do tempo, mas também ter uma noção de possíveis variáveis significantes. Ajuste de modelos paramétricos, com ênfase nas distribuições exponencial, Weibull e log-normal, usando métodos gráficos de comparação e Teste da razão de verossimilhança (TRV) para identificar qual se adequa melhor aos dados. Em modelos de regressão, selecionar as variáveis que possuem significância estatística para estimar as sobrevivências das pacientes, comparar o ajuste de outros modelos de interesse com o TRV, validar a adequação do mesmo por meio de análise de resíduos e realizar a interpretação dos coeficientes e comparar estimativas de sobrevida para alguns cenários. Ajustar um modelo de riscos proporcionais de Cox, analisar sua adequação e comparar as curvas de sobrevivências e

de riscos das pacientes para diferentes cenários, de acordo com as variáveis inseridas no modelo.

3 EMBASAMENTO TEÓRICO

Neste capítulo foi apresentado os conceitos de Análise de Sobrevivência e suas técnicas estatísticas, que servirão como referência para analisar e aplicar tais técnicas aos dados de mulheres com câncer de mama. Grande parte dos assuntos e metodologias abordados neste capítulo foi baseado nos livros *Análise de Sobrevivência Aplicada* (COLOSIMO; GIOLO, 2006) e *Análise de Sobrevivência: Teoria e Aplicações em Saúde* (CARVALHO et al., 2011).

3.1 ANÁLISE DE SOBREVIVÊNCIA

A Análise de Sobrevivência é um método estatístico usado para analisar dados em que a variável resposta é o tempo até a ocorrência de um evento de interesse, que, para este trabalho, é o momento em que a paciente recebe alta da UTI, e tal evento deve ser previamente especificado. Porém, há casos em que nem todas as observações do estudo são obtidas, havendo algum tipo de censura. As técnicas de análise de sobrevivência utilizam tanto as observações obtidas quanto as observações censuradas, sendo um diferencial dos demais métodos estatísticos e essencial para responder as questões que o trabalho visa buscar.

3.1.1 Funções do Tempo de Sobrevivência

Seja T uma variável aleatória não-negativa, na maioria das vezes contínua, onde representa o tempo de falha. Sua função densidade de probabilidade, que fornece a probabilidade de que o evento de interesse ocorra no intervalo de tempo $[t, t + \Delta t]$, pode ser escrita da forma:

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

sendo $f(t) \geq 0$ para todo t e área sob a curva igual a 1.

A função de sobrevivência é definida como a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, é escrito como:

$$S(t) = P(T \geq t) = 1 - F(t),$$

em que $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$ e $F(t)$ é a função distribuição acumulada.

A função taxa de falha (ou risco) descreve a forma em que a taxa instantânea de falha muda com o tempo, e é útil para descrever a distribuição do tempo de vida dos pacientes. A função taxa de falha de T pode ser definida da seguinte forma:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

ou pode ser escrita em função de $f(t)$ e $F(t)$, como:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}.$$

A função taxa de falha acumulada é definida por:

$$H(t) = \int_0^t h(u)du,$$

segundo Colosimo e Giolo (2006), a função taxa de falha acumulada não possui uma interpretação direta, mas pode ser útil para estimar $h(t)$, principalmente na estimação não-paramétrica, em que $h(t)$ é difícil de ser estimada.

3.1.2 Censura

A censura é a presença de uma observação incompleta ou parcial do indivíduo de estudo. Pode ocorrer por diversas razões, seja pela perda de acompanhamento do indivíduo ao longo do estudo, a não ocorrência do evento de interesse até o fim do experimento, ou até mesmo pelo fato do indivíduo ter morrido por uma causa diferente do estudo.

Existem diferentes tipos de censura, dentre elas, censura do tipo I, censura do tipo II e censura intervalar. Em estudos clínicos, por exemplo, a censura do tipo I ocorre quando o estudo é finalizado e existem indivíduos que ainda não participaram do evento de interesse. A censura do tipo II ocorre quando o estudo é finalizado após ter ocorrido o evento de interesse em um

número pré-estabelecido de indivíduos, ou seja, aqueles que não ocorreram o evento de interesse serão considerados como censura. E por fim, tem-se a censura intervalar, que acontece quando não se sabe o tempo exato da ocorrência do evento de interesse, porém, sabe-se o intervalo que ocorreu.

A variável aleatória resposta em análise de sobrevivência é representada pelo par $(t_{i,}\, \delta_i)$, em que o indivíduo varia de i $(i = 1, 2, \dots, n)$. Sendo que, t_i representa o tempo de falha ou censura e δ_i, a variável indicadora de falha ou censura, da forma:

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

Em casos que houver covariáveis sendo medidas no i-ésimo indivíduo, por exemplo, $x_i = (sexo_i,\ idade_i,\ tratamento_i)$, os dados ficam representados por $(t_i,\ \delta_i,\ x_i)$.

3.2 MODELOS NÃO-PARAMÉTRICOS

As funções de sobrevivência podem ser estimadas de três formas: utilizando métodos não-paramétricos, paramétricos e semi-paramétricos. Os modelos não-paramétricos trazem a ideia de não haver nenhuma suposição de distribuição de probabilidade ao tempo de sobrevivência. Estes modelos possuem caráter descritivo, e podem ser usados para auxiliar na escolha de um modelo paramétrico adequado.

Os estimadores mais conhecidos da função de sobrevivência $S(t)$ são o de Kaplan-Meier, Nelson-Aalen e a Tabela de Vida ou Atuarial. Neste projeto, por questão de conveniência, foi abordado somente o estimador de Kaplan-Meier.

3.2.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, também conhecido como estimador limite-produto, é uma técnica não-paramétrica utilizada para estimar a função de sobrevivência. Este estimador é uma adaptação da função de sobrevivência empírica, sendo assim, é a função de sobrevivência estimada na ausência de censura.

Este estimador considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra. O procedimento para se obter a estimativa de Kaplan-Meier requer uma sequência de passos, em que o próximo depende do anterior. Os passos são gerados a partir de intervalos definidos pela ordenação dos tempos de falha de forma que cada um deles começa em um tempo observado e termina no próximo tempo. O estimador de Kaplan-Meier é definido como:

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

sendo $t_1 < t_2 < \ldots < t_k$ os k tempos distintos e ordenados de falha, d_j o número de falhas em cada t_j, e n_j o número de indivíduos sob risco em t_j. De acordo com Kaplan e Meier (1958) *apud* Colosimo e Giolo (2006), $\widehat{S(t)}$ é o estimador de máxima verossimilhança de $S(t)$, é um estimador não viciado para grandes amostras e converge assintoticamente para um processo gaussiano.

3.3 MODELOS PARAMÉTRICOS

Os modelos paramétricos (ou probabilísticos) se baseiam em supor uma distribuição de probabilidade para os tempos de sobrevivência. Alguns modelos paramétricos ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas, entre esses modelos, o Exponencial, sendo um caso particular da Weibull, a própria Weibull e Log-Normal. Segundo Colosimo e Giolo (2006), a escolha de um modelo probabilístico para descrever o tempo de falha deve ser feita com cuidado, visto que a utilização de um modelo inadequado pode acarretar erros grosseiros nas estimativas dessas quantidades.

3.3.1 Distribuição Weibull

A distribuição Weibull vem sendo frequentemente usada em estudos biomédicos e industriais. A sua popularidade se deve ao fato dela apresentar uma grande variedade de formas para sua função taxa de falha. Ela será monótona, podendo ser crescente, decrescente ou constante.

Para uma variável aleatória T com distribuição Weibull, sua função densidade de probabilidade é dada por:

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

onde $t \geq 0$, γ é o parâmetro de forma e α, o de escala, ambos positivos. O parâmetro α possui a mesma unidade de medida de t, e γ não tem unidade. As funções de sobrevivência e de risco são, respectivamente:

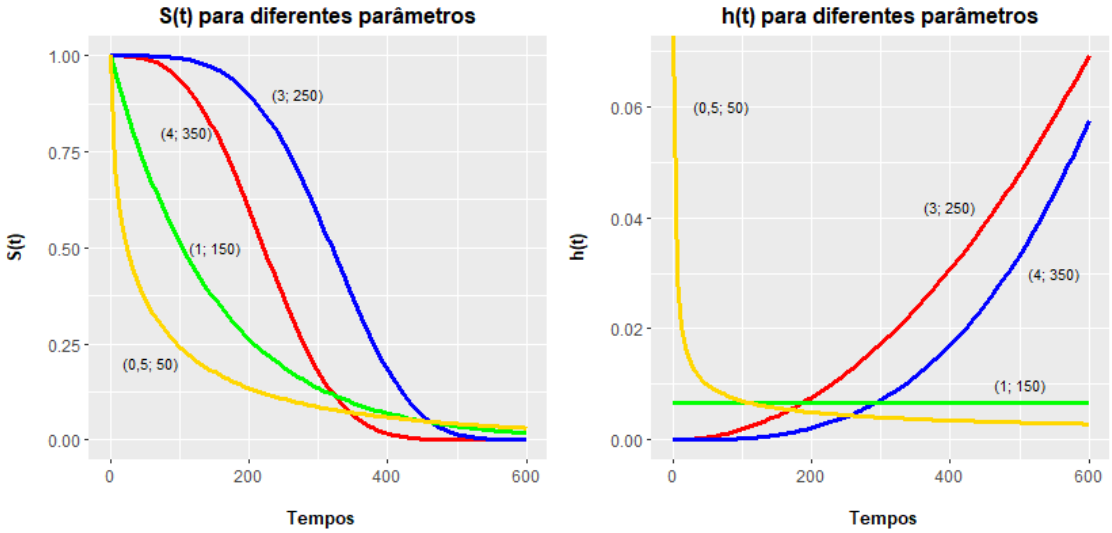
$$S(t) = e^{-\left(\frac{t}{\alpha}\right)^{\gamma}},$$

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

para $t \geq 0$, α e $\gamma > 0$. Quando $\gamma = 1$, a função de sobrevivência, $S(t)$, assume a distribuição exponencial, resultando em $S(t) = e^{-\left(\frac{t}{\alpha}\right)}$, e a função taxa de falha se torna constante no tempo, pois $h(t) = \frac{1}{\alpha}$. Nesse sentido, a distribuição exponencial é um caso particular da distribuição Weibull.

A Figura 1 apresenta diferentes formatos da função de sobrevivência e função de risco de acordo com os parâmetros γ e α, note que, quando $\gamma = 1$ (curva de cor verde), tem-se as funções de sobrevivência e risco com distribuição exponencial, conforme mencionado.

Figura 1 – Forma típica das funções de sobrevivência e de risco da distribuição Weibull, para alguns valores dos parâmetros (γ, α)



Fonte: Elaborado pelo autor, 2022.

3.3.2 Distribuição Log-Normal

A distribuição Log-normal é muito utilizada para caracterizar os tempos de vida de produtos e indivíduos. Existe uma relação entre a Log-normal e a Normal, assim como o nome sugere, o logaritmo de uma variável com distribuição Log-normal de parâmetros μ e σ tem distribuição Normal com média μ e desvio padrão σ.

A função densidade de probabilidade de uma variável aleatória T, com distribuição Log-normal é dada por:

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

onde $t > 0$, μ é a média do logaritmo do tempo de falha e σ é o desvio padrão. As funções de sobrevivência e de taxa de falha de uma distribuição Log-normal são dadas por, respectivamente:

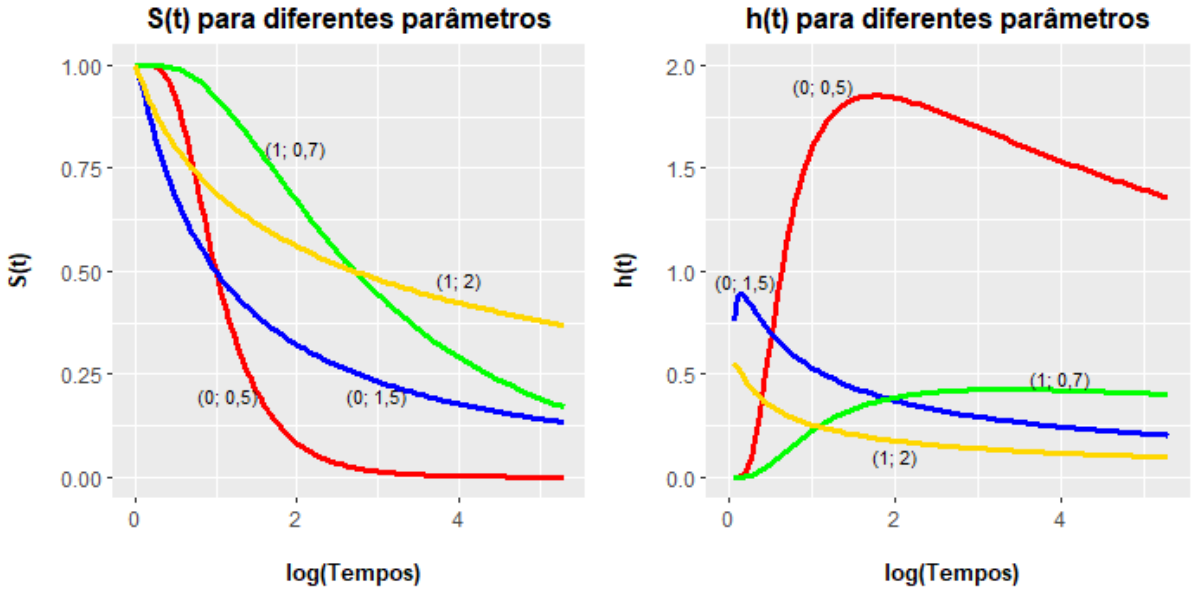
$$S(t) = \phi\left(\frac{-\log(t) + \mu}{\sigma}\right),$$

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

em que $\phi(\cdot)$ é a função distribuição acumulada de uma distribuição Normal padrão. A característica da função taxa de falha da distribuição Log-normal é que começa crescente, atinge um valor máximo e depois decresce.

A Figura 2 apresenta a forma de funções de sobrevivência e de taxa de falha da distribuição log-normal para alguns valores de μ e σ.

Figura 2 – Forma típica das funções de sobrevivência e de taxa de falha da log-normal para valores dos parâmetros (μ, σ)



Fonte: Elaborado pelo autor, 2022.

3.3.3 Distribuição Gama Generalizada

A distribuição Gama Generalizada (GG) foi introduzida por Stacy (1962) *apud* Colosimo e Giolo (2006), e é caracterizada por três parâmetros, τ, k e $\alpha > 0$. Sua função densidade é dada por:

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

sendo $t > 0$ e $\Gamma(k)$ é a função gama. Para esta distribuição tem-se um parâmetro de escala, sendo o α, e dois de forma, τ e k. Importante notar que se $\tau = k = 1$ tem-se $T \sim Exp(\alpha)$, para $k = 1$ tem-se $T \sim Weib(\tau, \alpha)$ e para $\tau = 1$ tem-se $T \sim Gama(k, \alpha)$. As funções de sobrevivência e de taxa de falha são dadas por, respectivamente:

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

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

em que $\gamma(k, x) = \int_0^x w^{k-1} e^{-w}\, dw$ é a razão da função gama incompleta, definida por $\gamma_1(k, x) = \gamma(k, x)\Gamma(k)$. Outras propriedades da distribuição gama generalizada podem ser encontradas em Lawless (1980).

3.4 ESTIMAÇÃO DOS PARÂMETROS

Existem alguns métodos de estimação conhecidos na literatura estatística. Em dados de sobrevivência que possuem censuras, é de extrema importância o processo de estimação ser capaz de incorporá-las. Colosimo e Giolo (2006) afirmam que o método de máxima verossimilhança é uma opção apropriada na área de análise de sobrevivência, pois incorpora as censuras dos dados, é relativamente simples de ser entendido e possui propriedades ótimas para grandes amostras.

3.4.1 Estimação de Máxima Verossimilhança

Suponha uma amostra de observações $t_1, t_2, \ldots, t_n$ de uma certa população de interesse em que todas são não-censuradas e a população é caracterizada pela sua função de densidade $f(t)$. A função de verossimilhança para um parâmetro genérico θ desta população é expressa por:

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

A dependência de f em θ é preciso ser mostrada, pois L é função de θ. O parâmetro θ não necessariamente representa apenas um parâmetro, também pode ser representado como um vetor de parâmetros. Por exemplo, no caso da distribuição Weibull, $\theta = (\gamma,\ \alpha)$.

Em casos de censura do tipo I, tem-se r falhas e $n - r$ censuras observadas no término do experimento e, sendo assim, $L(\theta)$ assume a forma:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta)\,.$$

Em casos de censura do tipo II, r é fixo e somente os r menores tempos são observados. Segue que $L(\theta)$ assume a forma:

$$L(\theta) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta)\,,$$

em que $\prod_{i=r+1}^{n} S(t_i; \theta) = [S(t_r; \theta)]^{n-r}$ com t_r o maior tempo observado. Como $\frac{n!}{(n-r)!}$ é uma constante, pode ser desprezado, pois não envolve nenhum parâmetro de interesse. Então:

$$L(\theta) \propto \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta)\,.$$

3.5 TESTE LOG-RANK

Um método de comparação de curvas de sobrevivência muito utilizado na área é o teste Log-rank, proposto por Mantel em 1966. Este teste é particularmente apropriado quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante. A estatística deste teste é a diferença entre o número observado de falhas em cada grupo e uma quantidade que corresponde ao número esperado de falhas sob a hipótese nula. Para este teste, as hipóteses a serem testadas são:

H_0: As curvas de sobrevivência são idênticas.

H_1: As curvas de sobrevivência não são idênticas (para pelo menos um tempo t).

A estatística Log-rank é dada da forma:

$$\text{Log} - \text{rank} = \sum \frac{\left(\sum O_{jt} - \sum E_{jt}\right)^2}{\sum E_{jt}} \sim {\chi^2}_{k-1},$$

onde $\sum O_{jt}$ representa a soma do número observado de eventos no j-ésimo grupo, $\sum E_{jt}$ representa a soma do número esperado de eventos no j-ésimo grupo e k representa o número de grupos sendo comparados. A hipótese H_0 é rejeitada se Log-rank $> {\chi^2}_{k-1}$.

3.6 COMPARAÇÃO DE MODELOS

Na comparação de modelos, é bastante útil usar técnicas gráficas para discriminar se um modelo é melhor que outro. Fazer a comparação da curva do modelo proposto com a curva de Kaplan-Meier ou a linearização da função de sobrevivência dos modelos, por exemplo. Outra forma de discriminar modelos é por meio de testes de hipóteses.

As hipóteses a serem testadas são:

H_0: O modelo de interesse é adequado.

H_1: O modelo de interesse não é adequado.

Colosimo e Giolo (2006) afirmam que este teste é geralmente realizado utilizando-se a estatística da razão de verossimilhanças em modelos encaixados (Cox e Hinkley, 1974). Portanto, deve ser identificado um modelo geral de modo que os modelos de interesse sejam casos particulares. A estatística da razão de verossimilhanças é dada por:

$$TRV = -2\log\left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)}\right] = 2\left[logL(\hat{\theta}_G) - logL(\hat{\theta}_M)\right],$$

em que $logL(\hat{\theta}_G)$ e $logL(\hat{\theta}_M)$ são, respectivamente, o logaritmo da função de verossimilhança do modelo geral e modelo de interesse. Sob H_0, TRV tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros $\hat{\theta}_G$ e $\hat{\theta}_M$.

3.7 MODELOS DE REGRESSÃO

Os estudos na área médica muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Segundo Carvalho et al. (2011), o efeito de covariáveis sobre o tempo de sobrevivência é estimado através de um modelo de regressão, no qual o tempo de sobrevivência é a variável resposta e $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ é o vetor de covariáveis (variáveis independentes) possivelmente associadas ao tempo. O modelo contém um componente aleatório, que descreve probabilisticamente o comportamento do tempo de sobrevivência, e um componente sistemático, que descreve a relação entre os parâmetros da distribuição de probabilidade e as covariáveis. Entre os vários modelos, um que descreve este relacionamento é dado da forma:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)g(\boldsymbol{x\beta}),$$

em que β's são os coeficientes a serem estimados, $g(\cdot)$ é uma função positiva e contínua das covariáveis e $\lambda_0(t)$ é o risco basal de um indivíduo que possui $g(\boldsymbol{x\beta}) = 1$.

3.7.1 Modelo de Regressão Exponencial

No modelo de regressão exponencial, utilizado quando é assumido que o risco é constante ao longo do tempo, o parâmetro α depende das covariáveis $\boldsymbol{x}$ da forma:

$$\alpha(\boldsymbol{x}) = e^{\boldsymbol{x\beta}},$$

sendo β's as estimativas dos efeitos das covariáveis e α o parâmetro que define o risco exponencial. As funções de risco e sobrevivência para o modelo de regressão exponencial são definidas, respectivamente, como:

$$\lambda(t|\boldsymbol{x}) = \alpha(\boldsymbol{x}) = e^{\boldsymbol{x\beta}},$$

$$S(t|\boldsymbol{x}) = e^{-\alpha(\boldsymbol{x})t} = \exp(-\exp(\boldsymbol{x\beta})\, t).$$

3.7.2 Modelo de Regressão Weibull

Utilizar a distribuição Weibull no contexto da modelagem de sobrevivência significa que o tempo T segue uma distribuição Weibull, e o parâmetro α é modelado pelas covariáveis. De acordo com Carvalho et al. (2011), existem várias formas de incluir covariáveis na distribuição Weibull, mas essa é a mais utilizada. As funções de risco e sobrevivência para o modelo de regressão Weibull são dadas, respectivamente, por:

$$\lambda(t|\boldsymbol{x}) = \gamma t^{\gamma-1}\alpha(\boldsymbol{x})^{\gamma} = \gamma t^{\gamma-1}\exp(\boldsymbol{x\beta})^{\gamma},$$

$$S(t|\boldsymbol{x}) = \exp(-(\alpha(\boldsymbol{x})t)^{\gamma}) = \exp(-(\exp(\boldsymbol{x\beta})t)^{\gamma}).$$

3.7.3 Modelo de Regressão Log-normal

Quando T segue uma distribuição Log-normal, a função de sobrevivência para o modelo de regressão Log-normal é dada por:

$$S(t|\boldsymbol{x}) = \phi\left(\frac{-\log(t) + \boldsymbol{x\beta}}{\sigma}\right).$$

3.8 RESÍDUOS

Para avaliar o quão bem o modelo se ajusta aos dados, é fundamental que os resíduos do modelo sejam analisados. Desta forma, pode-se confirmar se o modelo está apropriado ou não para uso. Técnicas Gráficas são bastante utilizadas para examinar diferentes aspectos do modelo, como por exemplo, a distribuição dos erros.

3.8.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell (1968) são utilizados para examinar o ajuste global do modelo. Estes resíduos são quantidades determinadas por:

$$\widehat{e_i} = \widehat{\Lambda}(t_i|x_i),$$

em que $\widehat{\Lambda}(\cdot)$ é a função de risco acumulado do modelo ajustado. Para os modelos de regressão Exponencial, Weibull e Log-normal, os resíduos Cox-Snell são dados, respectivamente, por:

$$\widehat{e_i} = t_i e^{\{-x_i'\hat{\beta}\}},$$

$$\widehat{e_i} = \left[t_i e^{\{-x_i'\hat{\beta}\}}\right]^{\hat{\gamma}},$$

$$\widehat{e_i} = -\log\left[1 - \Phi\left(\frac{\log(t_i) - x_i'\hat{\beta}}{\hat{\sigma}}\right)\right].$$

Segundo Lawless (1982) *apud* Colosimo e Giolo (2006), os resíduos $\widehat{e_i}$ vêm de uma população homogênea e devem seguir uma distribuição Exponencial padrão se o modelo for adequado. Assim, o gráfico $\widehat{e_i}$ *versus* $\widehat{\Lambda}(\widehat{e_i})$ deve ser aproximadamente uma reta com inclinação 1, quando o modelo exponencial for adequado, uma vez que $\widehat{\Lambda}(\widehat{e_i}) = -\log\left[\hat{S}(\widehat{e_i})\right]$.

3.8.2 Resíduos Padronizados

Assim como os resíduos Cox-Snell, os resíduos padronizados são úteis para examinar o ajuste global do modelo, sendo que estes resíduos padronizados são quantidades calculadas por:

$$\hat{v}_i = \frac{\log(t_i) - x_i'\hat{\beta}}{\hat{\sigma}}.$$

Se o modelo de regressão Exponencial for adequado, estes resíduos devem ser uma amostra censurada da distribuição Valor Extremo padrão. De modo análogo, se o modelo Log-normal for apropriado, os resíduos devem ser uma amostra censurada da distribuição Normal padrão.

3.9 MODELO DE REGRESSÃO DE COX

Segundo Colosimo e Giolo (2006), o modelo de regressão de Cox é o mais utilizado em estudos clínicos por sua versatilidade. O modelo de regressão de Cox (Cox, 1972) abriu uma

nova fase na modelagem de dados clínicos. O artigo de Cox (1972), onde o modelo é apresentado, foi neste período o segundo artigo mais citado na literatura estatística.

O modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis. O modelo de Cox assume sua forma mais simples quando uma única covariável é um indicador de grupos. Supondo uma comparação do tempo de falha de dois grupos, em que os indivíduos são selecionados aleatoriamente para receber um tratamento padrão (grupo 0) ou um tratamento especial (grupo 1). Sendo $\lambda_0(t)$ e $\lambda_1(t)$ as funções de taxa de falha do primeiro e segundo grupo, respectivamente, e assumindo proporcionalidade entre as duas funções, tem-se que:

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K,$$

em que K é a razão das taxas de falha, constante para todo tempo t de acompanhamento do estudo. Se x é a variável indicadora de grupo, onde:

$$x = \begin{cases} 0 & \text{se grupo 0,} \\ 1 & \text{se grupo 1,} \end{cases}$$

e $K = \exp\{\beta x\}$, então, $\lambda(t) = \lambda_0(t)\exp\{\beta x\}$, que é a expressão para uma única covariável do modelo de Cox. Ou seja:

$$\lambda(t) = \begin{cases} \lambda_1(t) = \lambda_0(t)\exp\{\beta\}, & \text{se } x = 1 \\ \lambda_0(t), & \text{se } x = 0. \end{cases}$$

De forma geral, considerando p covariáveis, de modo que $\mathbf{x}$ seja um vetor com os componentes $\mathbf{x} = (x_1, \ldots, x_p)'$. A expressão geral do modelo de regressão de Cox é dada por:

$$\lambda(t) = \lambda_0(t) g(\mathbf{x}'\beta),$$

em que g é uma função não-negativa que deve ser especificada, tal que $g(0) = 1$. O modelo é composto pelo produto de dois componentes, um não-paramétrico e outro paramétrico. O

componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. O componente paramétrico é frequentemente utilizado na forma multiplicativa:

$$g(\mathbf{x}'\beta) = \exp\{\mathbf{x}'\beta\} = \exp\{\beta_1 x_1 + \cdots + \beta_p x_p\},$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis. Esta forma garante que $\lambda(t)$ seja sempre não-negativa.

Este modelo também é denominado por modelos de riscos proporcionais, pois a razão das taxas de falha de dois indivíduos diferentes é constante no tempo. Portanto, a razão das funções de taxa de falha para os indivíduos i e j dada por:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\mathbf{x}'_i\beta\}}{\lambda_0(t)\exp\{\mathbf{x}'_j\beta\}} = \exp\{\mathbf{x}'_i\beta - \mathbf{x}'_j\beta\},$$

não depende do tempo. Se um indivíduo no início do estudo possui um risco de morte igual a três vezes o risco de um segundo indivíduo, logo, esta razão de riscos será a mesma para todo o período de acompanhamento.

4 BASE DE DADOS

Para a realização das análises, a base de dados utilizada para o trabalho possui 38.576 observações de pacientes com câncer de mama, em um período de 5 anos, de 2016 a 2020. Destas observações, houve 2.443 mortes (censura), aproximadamente 6,3% dos casos. Os dados foram retirados do SIHSUS (Sistema de Informações Hospitalares do SUS), no site do DATASUS.

A base contém um total de 11 variáveis, sendo a maioria categórica. Na Tabela 1 são apresentados os nomes e a descrição de cada variável.

Tabela 1 – Descrição das variáveis da base de dados

VARIÁVEL	DESCRIÇÃO
COMPLEX	Complexidade do Procedimento
CAR_INT	Caráter da Internação
MARCA_UTI	Tipo de UTI utilizada
IDADE	Idade da paciente em anos
CANC	Região da Neoplasia
IDH_HOSP	IDHM da cidade do hospital
RACA	Raça da paciente
PROC_REA	Procedimento realizado
ESPEC	Especialidade do leito
DIAS_PERM	Dias de permanência na UTI
MORTE	Falecimento da paciente

Fonte: Informe Técnico SIHSUS.

4.1 ANÁLISE EXPLORATÓRIA

Em toda análise estatística, é fundamental realizar uma análise exploratória ou, análise descritiva dos dados para que seja possível ter uma noção de como os dados se comportam. Diante disso, foram construídos histogramas, *Box-Plots* e tabelas de frequência em torno da variável de interesse, o tempo de internação da paciente na UTI até o recebimento de sua alta do hospital.

Pode-se ver na Tabela 2 que a maioria das pacientes estão entre 31 a 70 anos de idade, representando aproximadamente 84% dos dados, e conforme a idade aumenta, a taxa proporcional de falha decresce levemente. Quanto ao tempo de permanência na UTI, 99,6% das pacientes ficam até um mês internadas. Em relação ao tipo/região da neoplasia, a mais frequente é a lesão invasiva, representando 47% das observações. Em relação a raça da paciente, cerca de 67% são brancas, e de maneira geral possuem taxas de falhas proporcionalmente parecidas. Por fim, a maioria das internações é de forma eletiva, com 1% destas observações sendo a morte da paciente, porém, quando a internação é urgente, 22% das pacientes acabam falecendo.

Tabela 2 – Análise Descritiva dos dados de câncer de mama (continua)

VARIÁVEL	CATEGORIA	N	N %	FALHA	CENS.	FALHA %
IDADE	0: 1 a 30 anos	668	2%	632	36	95%
	1: 31 a 50 anos	12.773	33%	12.120	653	95%
	2: 51 a 70 anos	19.641	51%	18.404	1.237	94%
	3: mais de 70 anos	5.494	14%	4.977	517	91%
RACA	0: Amarela	448	1%	422	26	94%
	1: Branca	25.964	67%	24.309	1.655	94%
	3: Parda	7.910	21%	7.394	516	93%
	4: Preta	2.186	6%	2.032	154	93%
	5: Desconhecida	2.068	5%	1.976	92	96%
CANC	1: Lesão Invasiva	18.130	47%	16.766	1.364	92%
	2: Mamilo	8.615	22%	8.028	587	93%
	3: Região Externa	9.695	25%	9.277	418	96%
	4: Região Interna	2.136	6%	2.062	74	97%

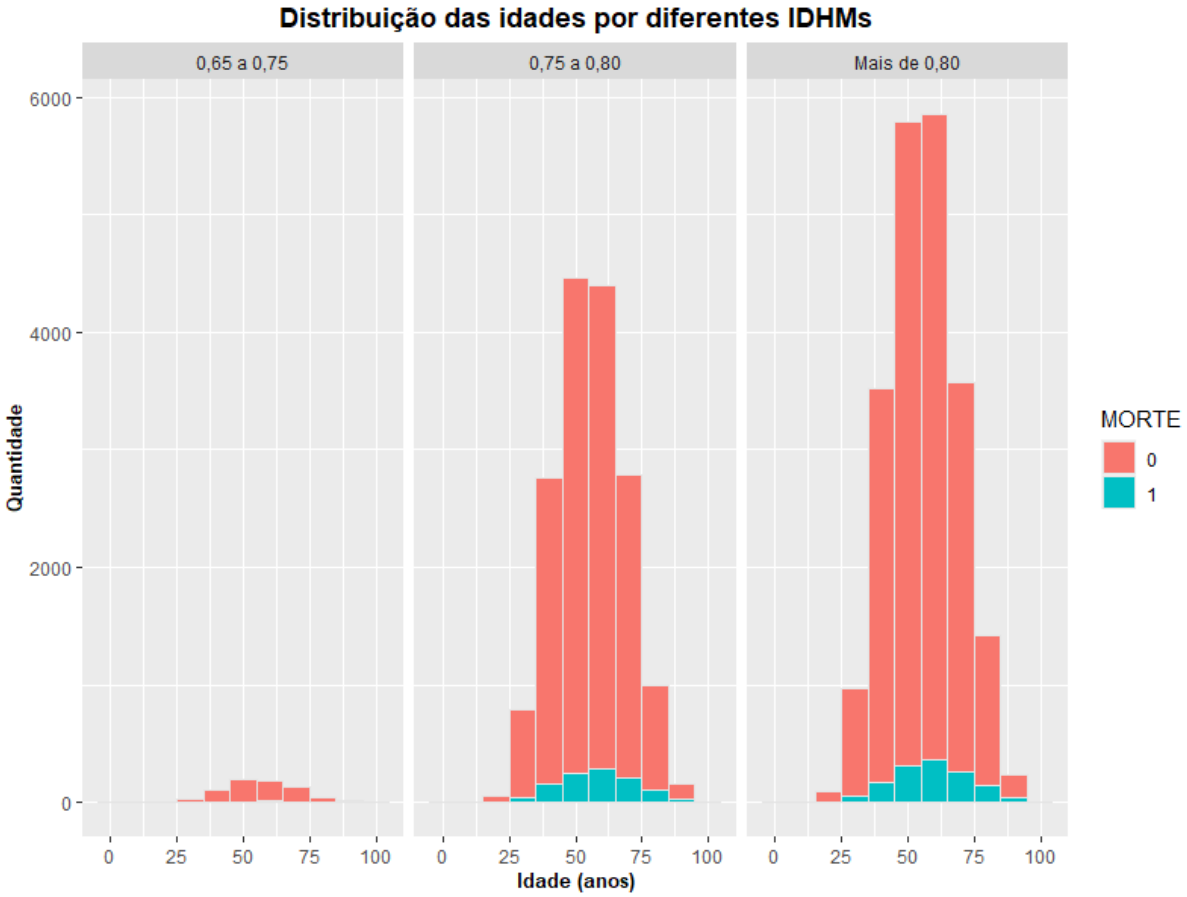
Tabela 2 – Análise Descritiva dos dados de câncer de mama (conclusão)

VARIÁVEL	CATEGORIA	N	N %	FALHA	CENS.	FALHA %
IDH_HOSP	0: 0,65 a 0,75 Pts	716	2%	678	38	95%
	1: 0,75 a 0,80 Pts	16.404	43%	15.333	1.071	93%
	2: mais de 0,80 Pts	21.456	56%	20.122	1.334	94%
DIAS_PERM	0: 1 a 15 dias	37.824	98%	35.650	2.174	94%
	1: 16 a 30 dias	598	1,6%	385	213	64%
	2: mais de 30 dias	154	0,4%	98	56	64%
ESPEC	1: Clínica Cirúrgica	29.966	78%	29.928	38	99,9%
	3: Clínica Médica	8.191	21%	5.971	2.220	73%
	4: Cuidados Prolongados	348	1%	163	185	47%
	9: Trat. realizado em Hospital Dia	71	0,2%	71	0	100%
CAR_INT	1: Eletiva	28.799	75%	28.548	251	99%
	2: Urgência/Emergência	9.777	25%	7.585	2.192	78%
MARCA_UTI	0: Leito sem especialidade	37.503	97%	35.307	2.196	94%
	75: UTI Adulto II	726	2%	564	162	78%
	76: UTI Adulto III	347	1%	262	85	76%

Fonte: Elaborado pelo autor, 2022.

Outra forma de diagnosticar o comportamento dos dados é graficamente, por meio dos histogramas. Na Figura 3 é possível visualizar a distribuição das idades das pacientes separando-as por diferentes níveis de IDHM (Índice de Desenvolvimento Humano Municipal) da cidade do hospital, e sendo preenchidas pela variável de interesse. Neste caso, a morte da paciente é dada por azul (Morte = 1), indicando censura e a alta da UTI, na cor vermelha (Morte = 0).

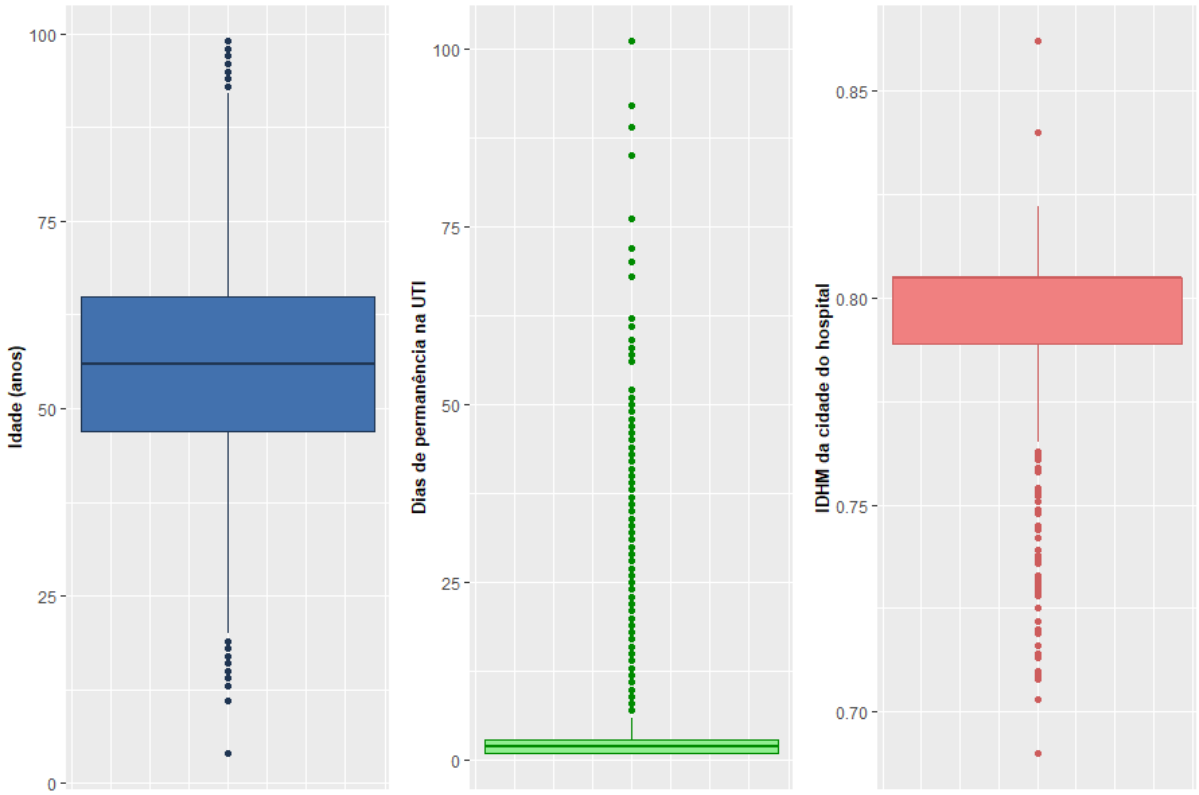
Figura 3 – Distribuição das idades das pacientes separadas por diferentes níveis de IDHM e preenchidas pela variável de interesse



Fonte: Elaborado pelo autor, 2022.

Para checar a presença de *Outliers*, observações que se diferenciam das demais, seja por causas naturais ou inconsistências na obtenção dos dados, é importante visualizar os gráficos *Box-plot* para cada variável. Pode-se verificar na Figura 4 que as pacientes menores de 20 anos e com mais de 85 anos de idade são observações atípicas, assim como que mais de duas semanas de permanência na UTI é considerado raro, isso pode ser confirmado na Tabela 2, onde 2% das observações passam de 15 dias na UTI. Alguns hospitais de cidades com IDHM menor que 0,77 e maior que 0,82 estão sendo considerados *Outliers*.

Figura 4 – Box-Plots das idades, dias de permanência na UTI e IDHMs das cidades do hospital, respectivamente



Fonte: Elaborado pelo autor, 2022.

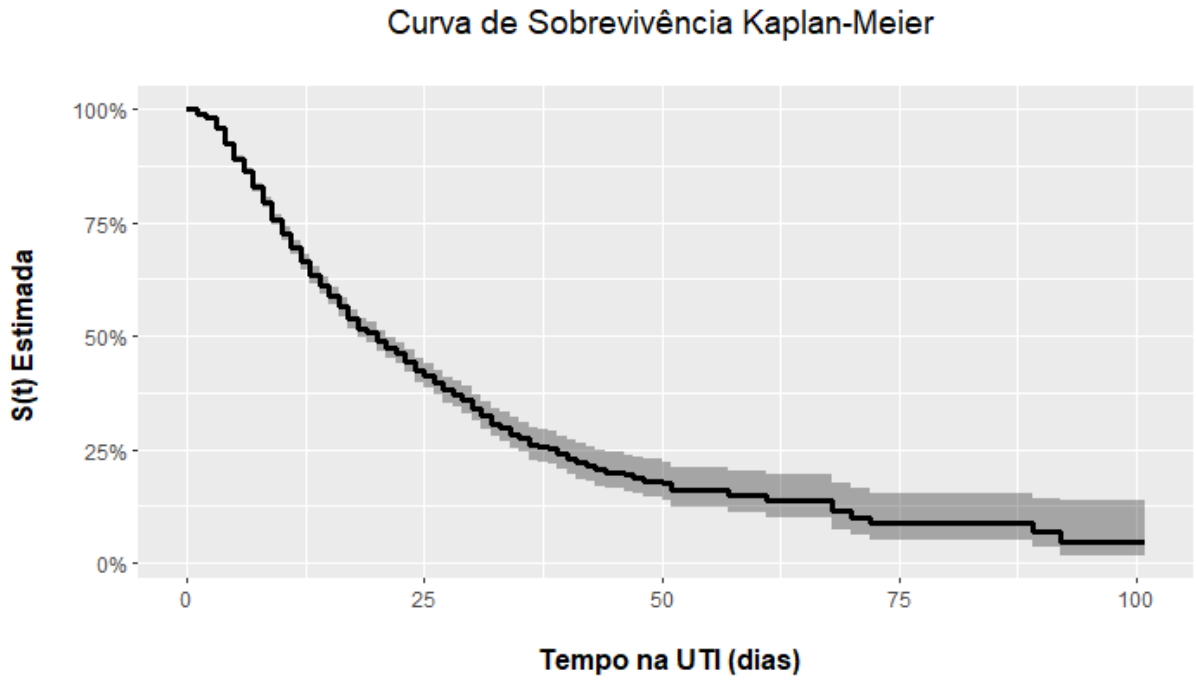
5 ANÁLISE NÃO-PARAMÉTRICA

Neste capítulo foram introduzidas as técnicas de análise de sobrevivência aos dados de forma não paramétrica. Portanto, foi construída a curva de Kaplan-Meier, estimação das quantidades básicas, como o tempo médio de vida, tempo mediano de vida e tempo médio de vida restante. E por fim, comparar curvas de sobrevivências com diferentes variáveis utilizando o método de Log-rank.

5.1 ESTIMADOR DE KAPLAN-MEIER

Aplicando o estimador de Kaplan-Meier ao conjunto de dados e construindo a curva de sobrevivência estimada, com intervalos de confiança de 95%, pode-se verificar pela Figura 5 uma curva acentuada no primeiro mês de internação e certa constância após 45 dias.

Figura 5 – Curva de sobrevivência estimada Kaplan-Meier para o tempo de internação na UTI para pacientes com câncer de mama no estado de São Paulo, entre 2016 e 2020



Fonte: Elaborado pelo autor, 2022.

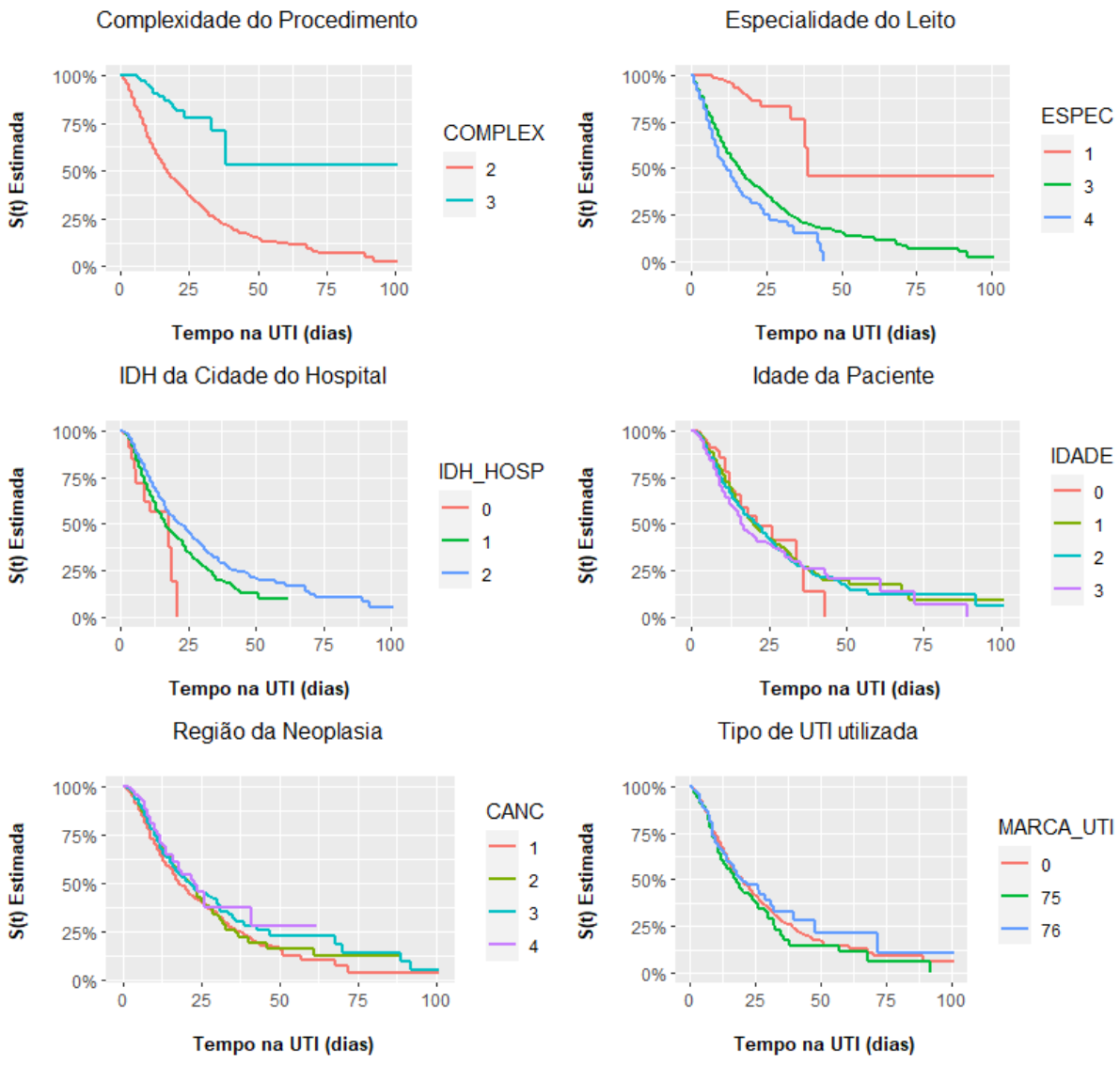
O tempo médio de internação das pacientes na UTI é de aproximadamente 29 dias, e o tempo mediano é de aproximadamente 19 dias, ou seja, 50% dos pacientes permanecem na UTI até o 19º dia. Fazendo o cálculo de vida média restante para os tempos t = 7, 15 e 30 é obtido

os resultados 27 dias, 28 dias e 28 dias, respectivamente, neste caso, pacientes que sobreviverem até uma semana, duas semanas e um mês, possuem um tempo de internação restante de, em média, de aproximadamente 28 dias.

5.2 COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

Além das estimativas gerais de Kaplan-Meier, é possível também calcular estimativas para diferentes estratos por variável. Neste sentido, foi construído graficamente as estimativas de sobrevivência para a maioria das variáveis do estudo. Na Figura 6 é observado que as variáveis COMPLEX, ESPEC e IDH_HOSP parecem possuir diferentes curvas de sobrevivência, mas para confirmar este fato, foi utilizado o teste de *Log-rank*.

Figura 6 – Curvas de sobrevivência para diferentes variáveis pelo método de Kaplan-Meier



Fonte: Elaborado pelo autor, 2022.

Realizando o teste *Log-rank* para comparação de curvas com todas as variáveis do estudo, sendo utilizado um nível de significância $\alpha = 5\%$, verifica-se na Tabela 3 que, praticamente todas as variáveis possuem pelo menos uma curva que difere estatisticamente entre a(s) outra(s), exceto a variável MARCA_UTI e RACA, com p-valores 0,10 e 0,07, respectivamente.

Tabela 3 – Teste Log-rank com todas as variáveis do estudo

VARIÁVEL	**ESTATÍSTICA LOG-RANK**	**P-VALOR**
COMPLEX	1.038,0	*p < 0,0001*
CAR_INT	1.062,0	*p < 0,0001*
MARCA_UTI	4,1	*p = 0,10*
IDADE	40,7	*p < 0,0001*
CANC	54,2	*p < 0,0001*
IDH_HOSP	59,7	*p < 0,0001*
RACA	8,8	*p = 0,07*
PROC_REA	2.101,0	*p < 0,0001*
ESPEC	1.806,0	*p < 0,0001*

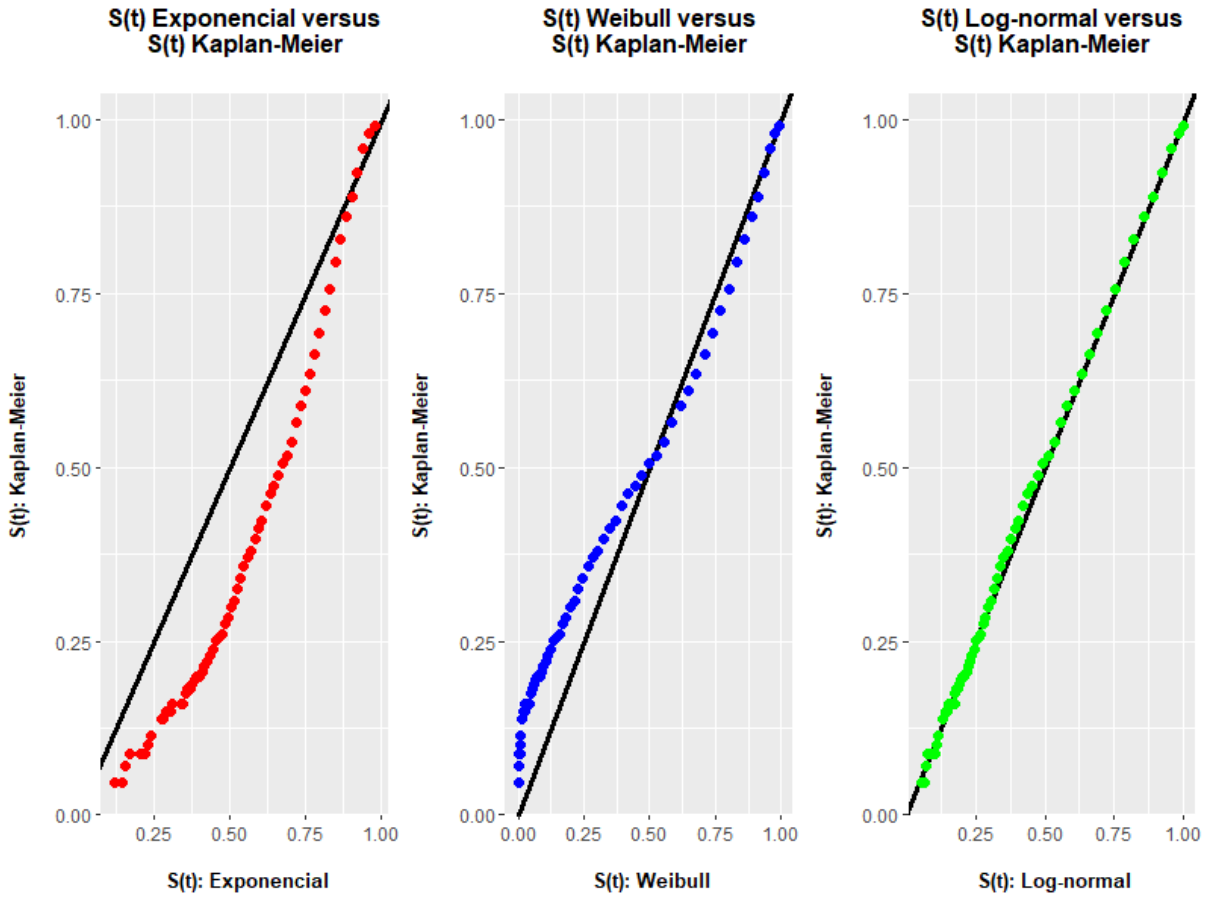
Fonte: Elaborado pelo autor, 2022.

6 MODELAGEM PARAMÉTRICA

Neste capítulo foram utilizados modelos probabilísticos para estimar as probabilidades de sobrevivência das pacientes com câncer de mama. Com isso, foi comparado por meio de técnicas gráficas e pelo TRV (Teste da Razão de Verossimilhança) qual distribuição se ajusta melhor em relação as estimativas geradas por Kaplan-Meier, assim como as funções Taxa de Falha Acumulada. As distribuições estudadas foram a Exponencial, Weibull e a Log-normal.

Uma forma simples de verificar o ajuste do modelo com as estimativas de Kaplan-Meier é colocando as estimativas de cada modelo no eixo horizontal contra as estimativas de Kaplan-Meier no eixo vertical, ou vice-versa. Se o ajuste estiver adequado, então os pontos seguirão uma reta $y = x$. Na Figura 7 é possível ver que o modelo probabilístico que melhor se ajustou foi o modelo Log-normal, porque os pontos se encontram mais próximos da reta.

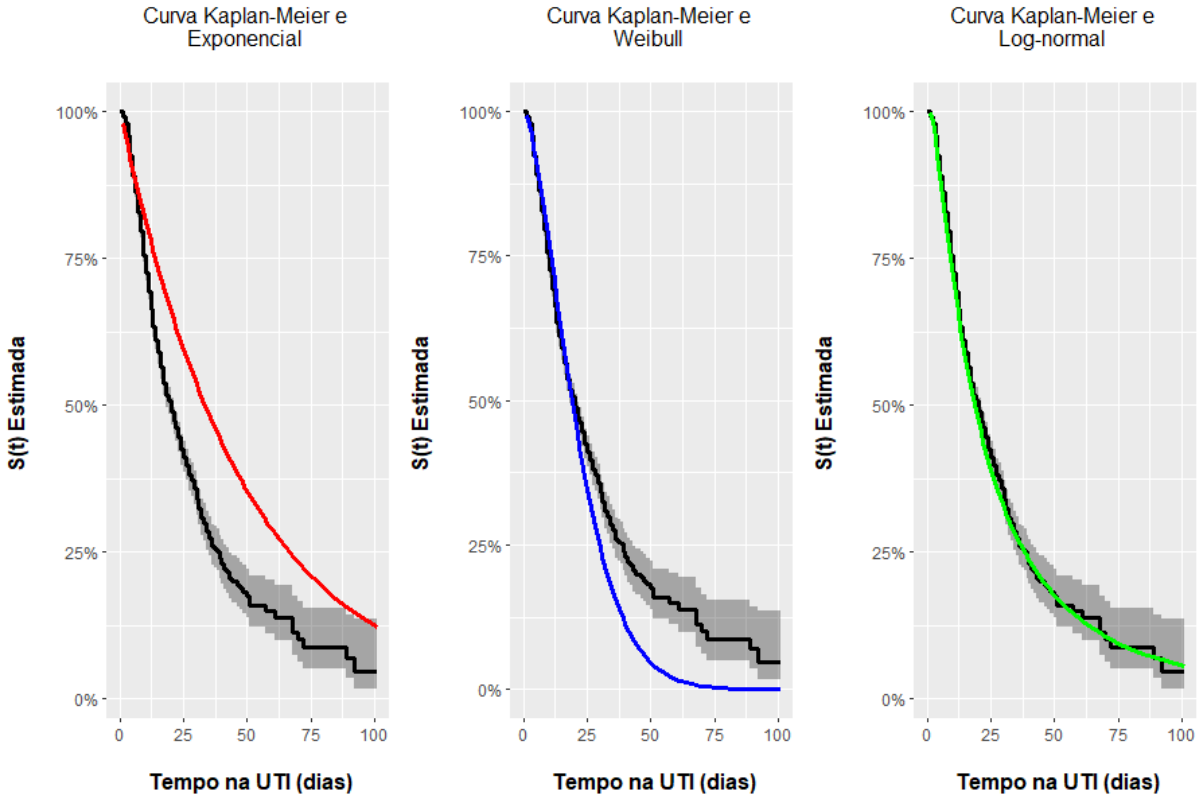
Figura 7 – Gráficos do ajuste de cada modelo contra as estimativas de Kaplan-Meier



Fonte: Elaborado pelo autor, 2022.

Na Figura 8 está sendo comparado o ajuste de cada modelo diretamente com a curva de sobrevivência de Kaplan-Meier mostrada na Figura 5, onde a linha preta são as estimativas Kaplan-Meier contendo o I.C. de 95% em cinza, e as linhas coloridas são as estimativas de cada modelo probabilístico. Reafirmando a conclusão da Figura 7, as estimativas do modelo Log-normal foi o que performou melhor, enquanto o modelo Exponencial ficou com estimativas superestimadas e o modelo Weibull obteve estimativas subestimadas após um mês, aproximadamente.

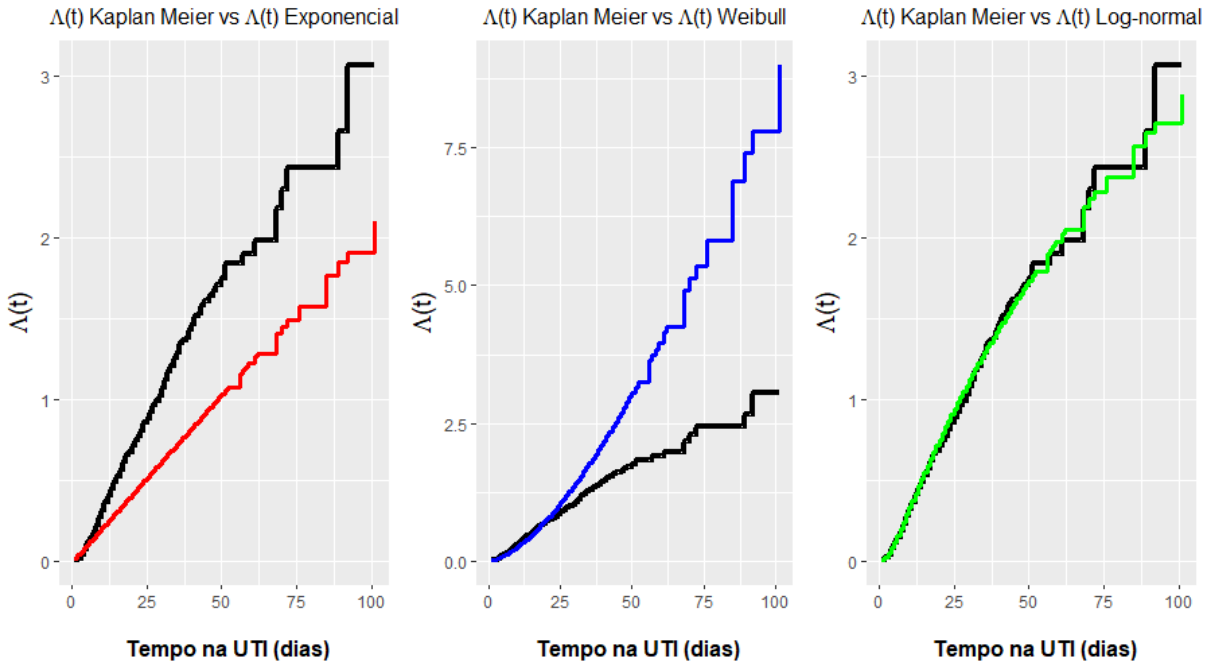
Figura 8 – Comparação das curvas de sobrevivência de cada modelo com a curva Kaplan-Meier



Fonte: Elaborado pelo autor, 2022.

As taxas de falha apresentam certa constância ao longo do tempo, na Figura 9 é mostrado a função de taxa de falha acumulada estimada por Kaplan-Meier comparando-a com as estimativas dos três modelos estudados. Como esperado, as funções de taxa de falha acumulada do modelo Log-normal são próximas das estimadas por Kaplan-Meier, o modelo Exponencial apresenta taxas de falha menores ao longo do tempo, enquanto o modelo Weibull apresenta taxas de falha exorbitantes após 40 dias.

Figura 9 – Comparação da função de taxa de falha acumulada de Kaplan-Meier *versus* cada modelo do estudo



Fonte: Elaborado pelo autor, 2022.

Para não somente deixar a subjetividade de lado ao comparar graficamente o ajuste de cada modelo, foi aplicado o Teste da Razão de Verossimilhança em modelos encaixados, portanto, o modelo geral para comparação foi o Gama Generalizada (GG). A hipótese alternativa é a de que o modelo de interesse ajustado é diferente do modelo geral. Portanto, se assumimos que o modelo geral é o melhor, a escolha seria pelo modelo que mais se assemelha ao modelo geral. A Tabela 4 mostra os resultados da estatística TRV para cada modelo e seus respectivos p-valores.

Tabela 4 – Teste da Razão de Verossimilhanças para cada modelo de interesse

MODELO	ESTATÍSTICA TRV	P-VALOR
EXPONENCIAL	1.643,0	$p < 0,0001$
WEIBULL	414,0	$p < 0,0001$
LOG-NORMAL	0,4225	$p = 0,5157$

Fonte: Elaborado pelo autor, 2022.

Utilizando um nível de significância padrão de $\alpha = 5\%$, temos um p-valor de 0,5157 para o modelo Log-normal, sendo o único modelo de interesse que não é rejeitada a hipótese que o mesmo seja adequado. Portanto, pelo TRV, pode-se concluir que o modelo mais indicado para estimar as sobrevivências das pacientes é o Log-normal.

7 MODELO DE REGRESSÃO

Neste capítulo, foi abordado os modelos de regressão com o intuito de incluir variáveis que possuem significância estatística para estimar as probabilidades de sobrevivência. Diante disso, foi apresentado o método de seleção das variáveis e seus critérios, análise de resíduos do modelo, interpretação dos coeficientes e predições para determinadas situações de acordo com os coeficientes estimados.

7.1 SELEÇÃO DE VARIÁVEIS

O método de seleção de variáveis utilizado neste capítulo foi o *Forward Stepwise*, que consiste em iniciar com somente o intercepto e inserir uma variável por vez e testar sua significância. Portanto, foi usado um nível de significância de $\alpha = 5\%$, a distribuição para o ajuste foi a gama generalizada, e o critério de comparação foi o TRV (Teste da Razão de Verossimilhança).

Para fins de simplificação com a nomenclatura das variáveis, considere as variáveis X1 = COMPLEX, X2 = CAR_INT, X3 = MARCA_UTI, X4 = IDADE, X5 = CANC, X6 = IDH_HOSP, X7 = RACA, X8 = PROC_REA e X9 = ESPEC. Na Tabela 5 é possível ver que a ordem de seleção das variáveis foi: Idade da paciente, Região da Neoplasia, Tipo de UTI utilizada e IDHM da cidade do hospital. Note que houve a checagem de interação dupla entre as quatro variáveis na etapa 5, a única interação significante foi X4 * X3 (IDADE e MARCA_UTI).

Tabela 5 – Seleção de variáveis usando o modelo Gama Generalizado (continua)

ETAPAS	MODELO	TRV	P-VALOR
Etapa 1	Nulo	-	-
	X1	1.608,0	*p < 0,0001*
	X2	3.876,0	*p < 0,0001*
	X3	0,5224	*p = 0,4698*
	X4	50,0	*p < 0,0001*
	X5	63,0	*p < 0,0001*
	X6	4.421,0	*p < 0,0001*
	X7	3,82	*p = 0,056*
	X8	1,85	*p = 0,1736*
	X9	1.627,0	*p < 0,0001*

Tabela 5 – Seleção de variáveis usando o modelo Gama Generalizado (conclusão)

Etapa 2	X4 + X1	1.268,0	*p < 0,0001*
	X4 + X2	1.264,0	*p < 0,0001*
	X4 + X3	0,1521	*p = 0,6966*
	X4 + X5	1.240	*p < 0,0001*
	X4 + X6	58,0	*p < 0,0001*
	X4 + X7	3,7	*p = 0,055*
	X4 + X8	1,7	*p = 0,19*
	X4 + X9	1.619,0	*p < 0,0001*
Etapa 3	X4 + X5 + X1	1.128,0	*p < 0,0001*
	X4 + X5 + X2	79,0	*p < 0,0001*
	X4 + X5 + X3	1.177,0	*p < 0,0001*
	X4 + X5 + X6	2.521,0	*p < 0,0001*
	X4 + X5 + X7	1.173,0	*p < 0,0001*
	X4 + X5 + X8	1.177,0	*p < 0,0001*
	X4 + X5 + X9	405,0	*p < 0,0001*
Etapa 4	X4 + X5 + X3 + X1	1.255,0	*p < 0,0001*
	X4 + X5 + X3 + X2	1.260,0	*p < 0,0001*
	X4 + X5 + X3 + X6	57,0	*p < 0,0001*
	X4 + X5 + X3 + X7	4,9	*p = 0,027*
	X4 + X5 + X3 + X8	0,69	*p = 0,4053*
	X4 + X5 + X3 + X9	1.587,0	*p < 0,0001*
Etapa 5	Interação X4 * X5	2,2	*p = 0,1404*
	Interação X4 * X3	21,0	*p < 0,0001*
	Interação X4 * X6	0,056	*p = 0,8127*
	Interação X5 * X3	1,6	*p = 0,2041*
	Interação X5 * X6	0,002	*p = 0,9683*
	Interação X3 * X6	2,3	*p = 0,1273*

Fonte: Elaborado pelo autor, 2022.

Sendo assim, o modelo final de regressão com distribuição gama generalizada ficou da seguinte forma: X4 + X5 + X3 + X6 + X4 * X3. Porém, a fim de trabalhar com um modelo com menos parâmetros, foi feito o T.RV. comparando com as outras três distribuições de interesse, Exponencial, Weibull e Log-normal. A Tabela 6 indica que o modelo mais adequado para estimar as probabilidades de sobrevivência é o modelo Log-normal.

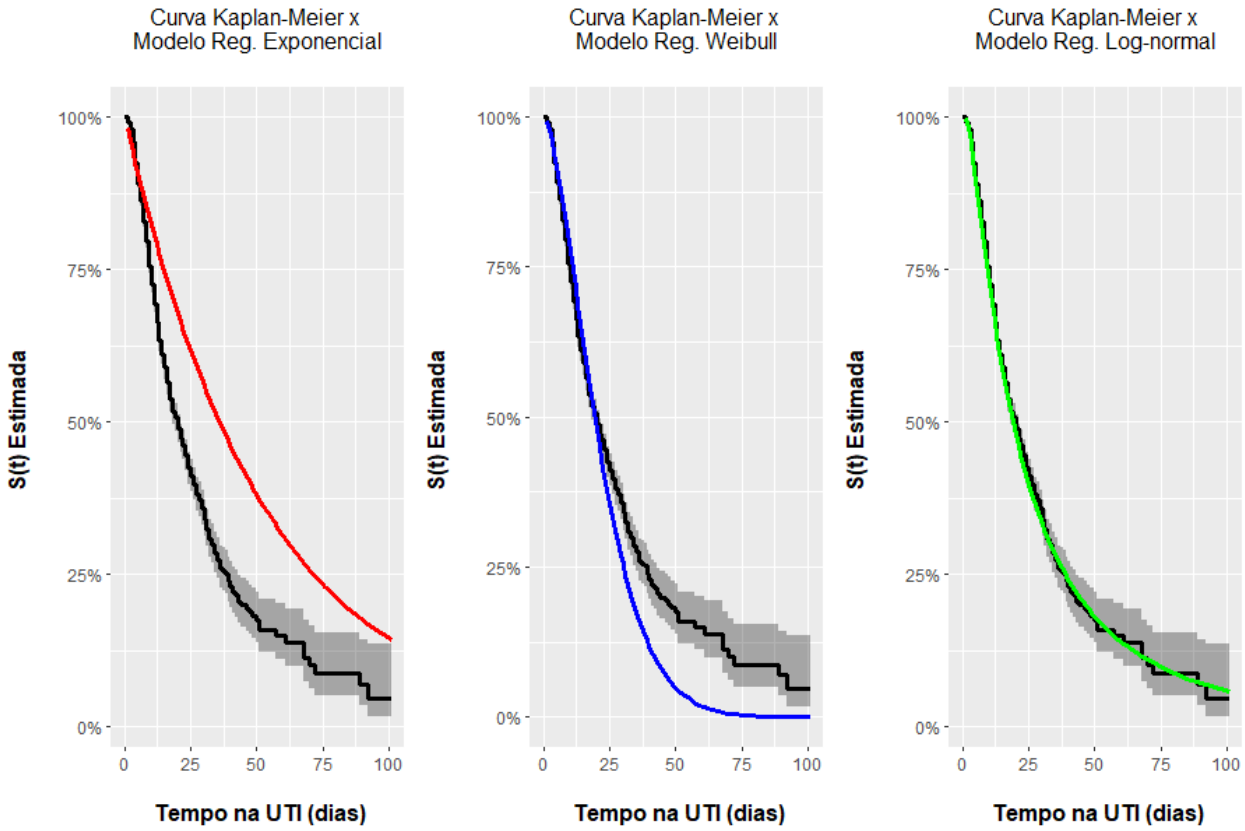
Tabela 6 – Comparando modelos com o Teste da Razão de Verossimilhança

MODELO	ESTATÍSTICA TRV	P-VALOR
EXPONENCIAL	1.607,0	*p < 0,0001*
WEIBULL	415,0	*p < 0,0001*
LOG-NORMAL	0,3656	*p = 0,5454*

Fonte: Elaborado pelo autor, 2022.

Comparando graficamente os resultados da Tabela 6, foi construído o ajuste do modelo de regressão de cada distribuição contra as estimativas de Kaplan-Meier, que pode ser visto nos resultados da Figura 10.

Figura 10 – Comparação dos ajustes de cada modelo de regressão *versus* a curva K.M.



Fonte: Elaborado pelo autor, 2022.

As estimativas do modelo de regressão exponencial ficam superestimadas na maior parte do tempo, o modelo Weibull ajusta-se muito bem até 25 dias, mas fica com estimativas subestimadas após o 25º dia, já o modelo Log-normal obteve estimativas muito próxima das estimativas de Kaplan-Meier ao longo do tempo.

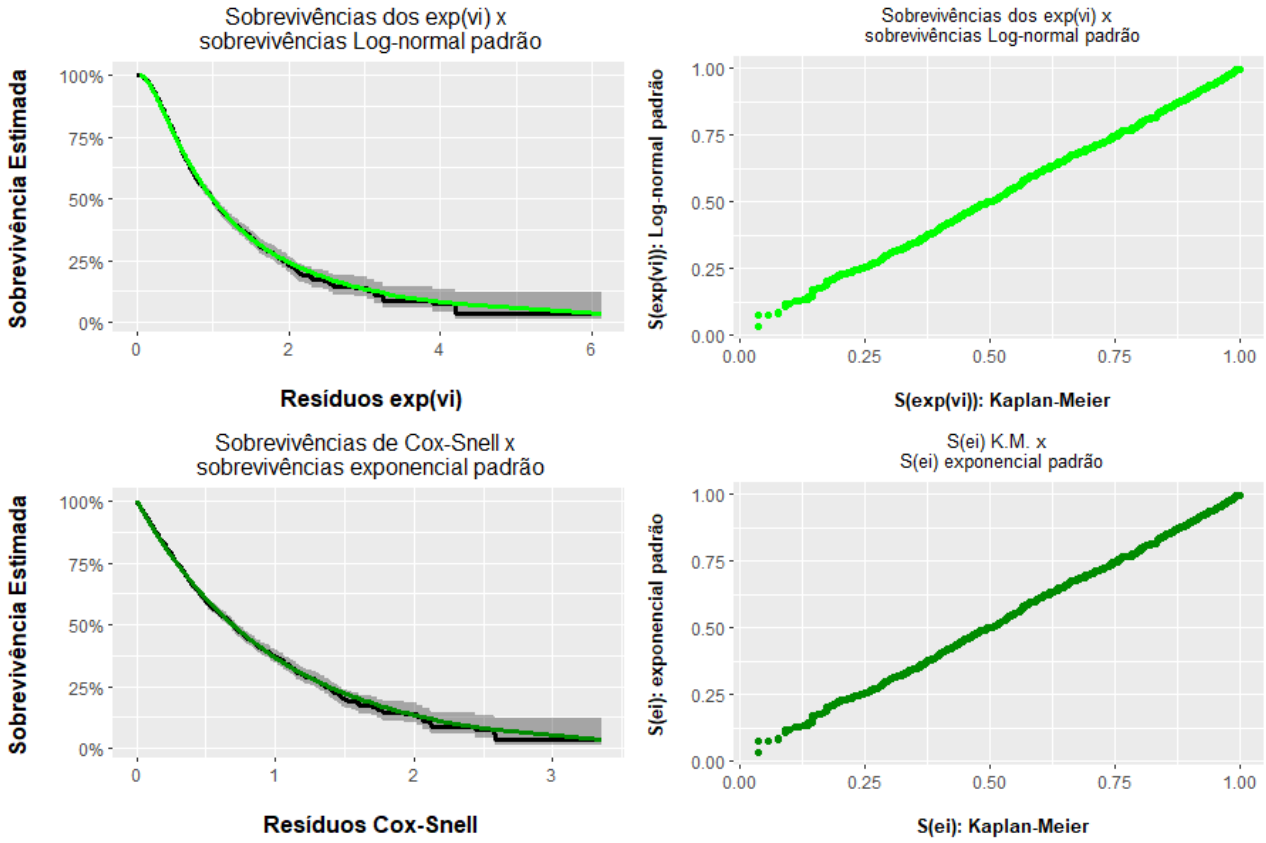
7.2 ANÁLISE DE RESÍDUOS

Nesta seção foi avaliado o ajuste e adequação do modelo de regressão final, com distribuição Log-normal, a partir da análise de resíduos. Se as sobrevivências dos resíduos $\exp(\hat{v}_i)$, estimadas pelo método Kaplan-Meier forem próximas das estimativas dos resíduos pelo modelo log-normal padrão, onde $\hat{v}_i$ são os resíduos padronizados, então o modelo pode

ser considerado adequado. De forma parecida, se as estimativas de sobrevivência pelo método de K.M. dos resíduos de Cox-Snell forem próximas das estimativas pelo modelo exponencial padrão, é considerado a adequação do modelo.

A Figura 11 apresenta os dois cenários, na primeira linha os resultados dos resíduos padronizados e na segunda linha os resultados dos resíduos Cox-Snell. Tanto os resíduos padronizados como os de Cox-Snell apresentam resultados satisfatórios, portanto, pode-se dizer que o modelo de regressão está adequado para uso.

Figura 11 – Análise de resíduos do modelo final, utilizado resíduos padronizados e Cox-Snell



Fonte: Elaborado pelo autor, 2022.

7.3 INTERPRETAÇÃO DOS COEFICIENTES

Nesta seção houve uma breve interpretação dos coeficientes estimados do modelo de regressão final, e de forma gráfica, foram comparadas as curvas de sobrevivência para diferentes situações. A Tabela 7 mostra as estimativas de cada coeficientes, erro padrão, p-valor e o $e^{\widehat{\beta}}$.

Tabela 7 – Estimativas dos coeficientes do modelo de regressão Log-normal

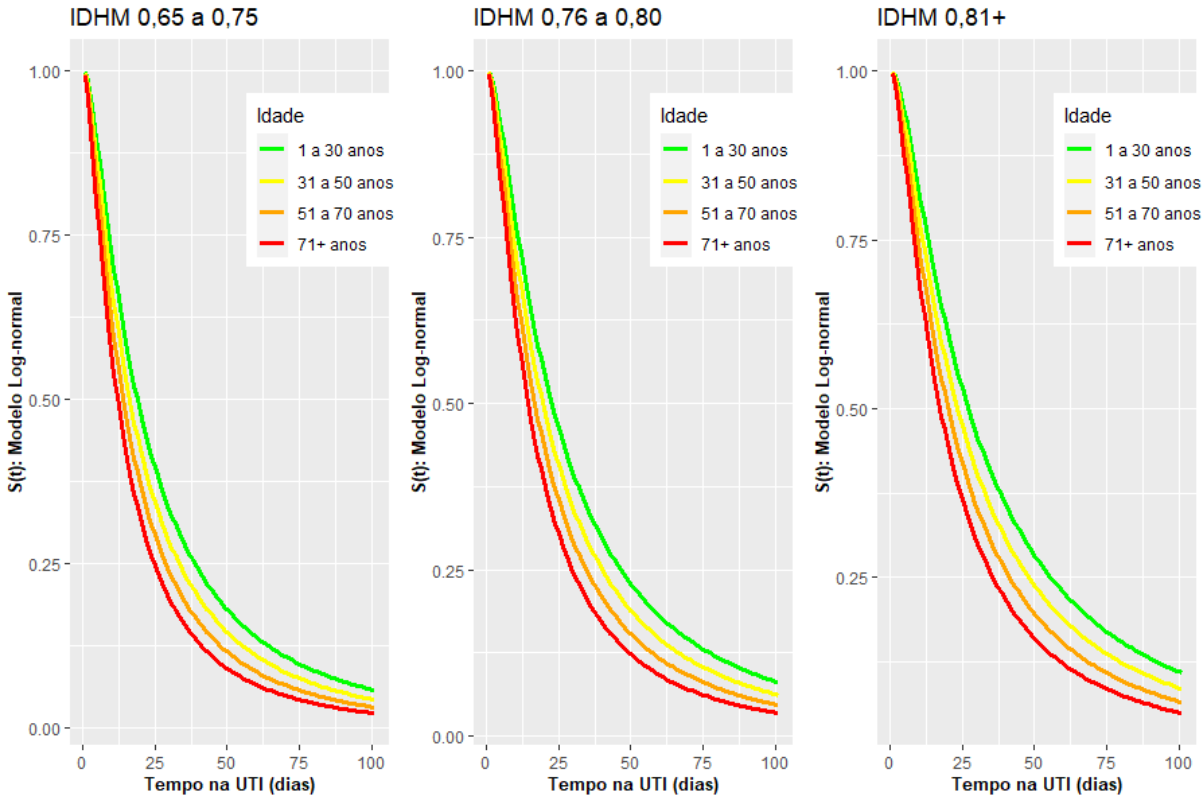
VARIÁVEL	ESTIMATIVA	ERRO PADRÃO	P-VALOR	$e^{\hat{\beta}}$
INTERCEPTO	2,72	0,06	*p < 0,0001*	-
IDADE	-0,1484	0,018	*p < 0,0001*	0,862
MARCA_UTI	-0,001	0,002	*p < 0,0001*	0,991
IDH_HOSP	0,18	0,023	*p < 0,0001*	1,197
CANC	0,1085	0,014	*p < 0,0001*	1,114
IDADE * MARCA_UTI	0,004	0,001	*p < 0,0001*	1,004

Fonte: Elaborado pelo autor, 2022.

Com os resultados da Tabela 7, vemos que os coeficientes das variáveis IDADE e MARCA_UTI são negativos, ou seja, quanto maior for o valor fixado para esta variável, menores serão as probabilidades de sobrevivência, em relação aos coeficientes positivos, quanto maior for o valor fixado, maior serão as probabilidades de sobreviver. Por exemplo, em IDADE temos $e^{\hat{\beta}}$ = 0,86, isso significa que a cada classe de idade o tempo mediano de vida das pacientes diminuem em 14%, em outras palavras, pacientes com idade de 31 a 50 anos possuem 14% de redução no tempo mediano de vida, comparado com pacientes de 1 a 30 anos. Para CANC, pacientes que possuem neoplasia na parte central da mama (CANC = 2) possuem um tempo mediano de vida 11% ($e^{\hat{\beta}} = 1{,}114$) maior que pacientes com neoplasia com lesão invasiva (CANC = 1).

Supondo uma situação com pacientes de diferentes faixas de idade e hospitais em diversas cidades com IDHMs diferentes, todas com neoplasia no mamilo (CANC = 2) e fixando MARCA_UTI = 0 (Leito sem especialidade), a Figura 12 revela que a probabilidade de sobrevivência de 15 dias de uma paciente jovem em um hospital de uma cidade com IDHM maior que 0,80 é de 71,2%, enquanto uma paciente idosa em um hospital em uma cidade com IDHM entre 0,65 e 0,75 é de 42,1%, neste caso, é o exemplo mais extremo de se comparar. A Figura 12 apresenta todos os cenários partindo de diferentes faixas de idade da paciente e IDHMs da cidade do hospital.

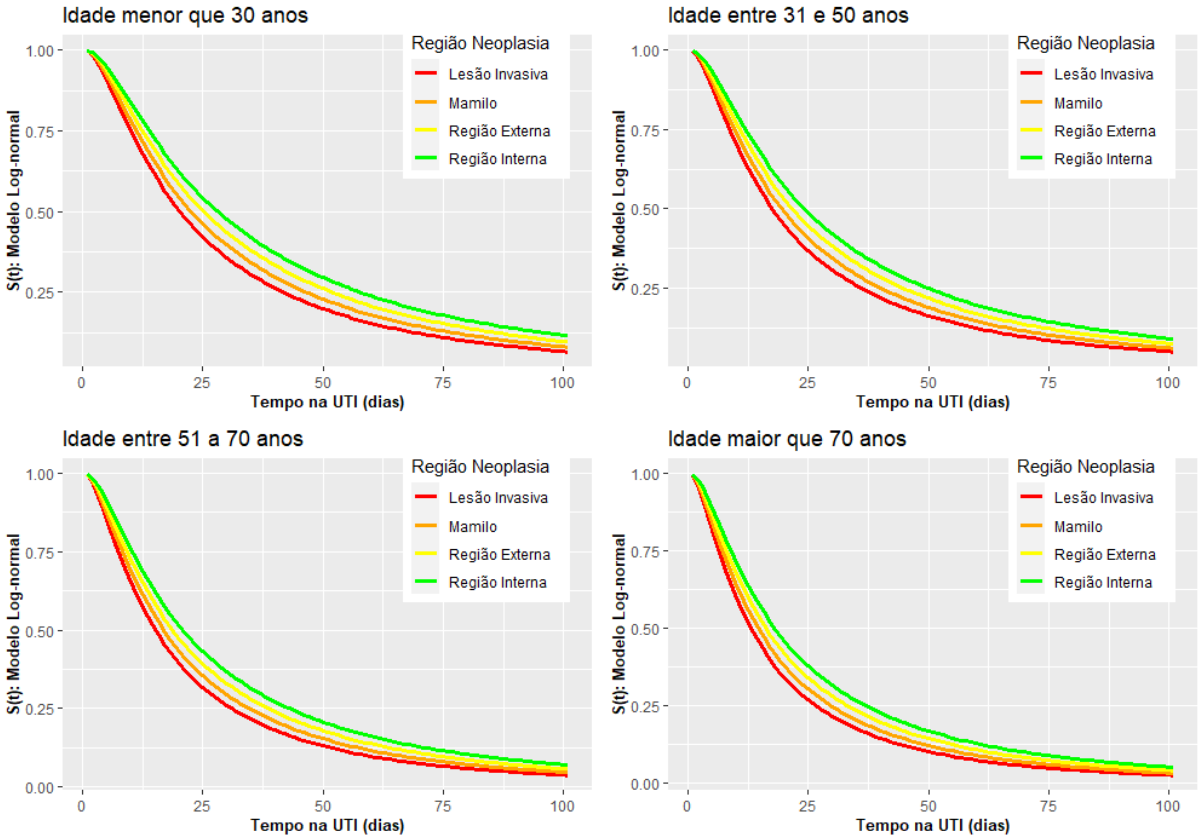
Figura 12 – Comparação de curvas de sobrevivência de pacientes com diferentes idades e IDHMs



Fonte: Elaborado pelo autor, 2022.

Considerando outra situação, visando comparar os tipos de neoplasias em diferentes faixas de idade das pacientes, fixando IDH_HOSP = 1 (IDHM entre 0,75 e 0,80) e também MARCA_UTI = 0, de acordo com a Figura 13, temos que a probabilidade de sobrevivência de uma semana para uma paciente jovem, com menos de 30 anos, e com neoplasia na região interna, é de 90,5%, e comparando com uma paciente idosa, com neoplasia no mamilo, a probabilidade é de 75,4%. Assim como a idade da paciente é um fator importante para sua sobrevivência, a região da neoplasia também é. De maneira geral, as regiões mais agravantes são: lesão invasiva, mamilo, região externa e região interna, respectivamente, que pode ser visto na Figura 13.

Figura 13 – Comparação de sobrevivências entre faixas de idade e região de neoplasia



Fonte: Elaborado pelo autor, 2022.

8 MODELO DE REGRESSÃO DE COX

Foi abordado neste capítulo a modelagem de regressão por meio do modelo de Cox, que possui dois componentes, um não-paramétrico e outro paramétrico. Além do ajuste do modelo de Cox, foi mostrado as novas estratificações de algumas variáveis da base de dados, seleção das variáveis para o ajuste, adequação do modelo, análise dos resíduos a partir dos resíduos *martingal* e *deviance*, interpretação dos coeficientes e, por fim, os gráficos de sobrevivências e riscos estimados.

Antes da modelagem, foram feitas algumas estratificações diferentes das anteriores, de modo que, a suposição básica do modelo de Cox seja atendida, que as taxas de falha sejam aproximadamente proporcionais ao longo do tempo. As variáveis com alterações na estratificação foram: IDADE, RACA, CANC, ESPEC e IDH_HOSP; a Tabela 8 mostra com mais detalhes tais estratificações.

Tabela 8 – Estratificações de algumas variáveis da base de dados

VARIÁVEL	ESTRATO
IDADE	0: 1 a 80 anos. 1: Mais de 80 anos.
RACA	0: Outras. 1: Branca.
CANC	0: Outras. 1: Lesão Invasiva.
ESPEC	0: Outras. 1: Clínica Cirúrgica.
IDH_HOSP	0: 0 a 0,8 Pts. 1: Mais de 0,8 Pts.

Fonte: Elaborado pelo autor, 2022.

8.1 SELEÇÃO DE VARIÁVEIS (MODELO DE COX)

O método utilizado para a seleção de variáveis para o ajuste do modelo de Cox é o mesmo do capítulo 7, para modelos de regressão, sendo o *Forward Stepwise*. Novamente,

considere as mesmas nomenclaturas para cada variável do estudo usadas no capítulo 7. Pode-se verificar pela Tabela 9 que as variáveis selecionadas para o modelo de regressão de Cox foram a Idade da paciente (X4), Região da neoplasia (X5) e IDHM da cidade do hospital (X6), não houve interação dupla significativa entre as três variáveis do modelo, considerando um nível de significância de $\alpha = 5\%$. Portanto, o modelo de regressão de Cox final é: X4 + X5 + X6.

Tabela 9 – Seleção de variáveis usando o modelo de regressão de Cox

ETAPA	MODELO	TRV	P-VALOR
Etapa 1	Nulo	-	-
	X1	1.395,0	*p < 0,0001*
	X2	1.224,0	*p < 0,0001*
	X3	1,64	*p = 0,2010*
	X4	42,0	*p < 0,0001*
	X5	47,0	*p < 0,0001*
	X6	55,0	*p < 0,0001*
	X7	5,9	*p = 0,0200*
	X8	0,154	*p = 0,6950*
	X9	2.307,0	*p < 0,0001*
Etapa 2	X4 + X1	1.388,0	*p < 0,0001*
	X4 + X2	1.220,0	*p < 0,0001*
	X4 + X3	0,96	*p = 0,3270*
	X4 + X5	48,0	*p < 0,0001*
	X4 + X6	58,0	*p < 0,0001*
	X4 + X7	4,9	*p = 0,0250*
	X4 + X8	0,21	*p = 0,6440*
	X4 + X9	2.294,0	*p < 0,0001*
Etapa 3	X4 + X5 + X1	1.388,0	*p < 0,0001*
	X4 + X5 + X2	1.241,0	*p < 0,0001*
	X4 + X5 + X3	1,3	*p = 0,2571*
	X4 + X5 + X6	47,0	*p < 0,0001*
	X4 + X5 + X7	4,7	*p = 0,0310*
	X4 + X5 + X8	0,07	*p = 0,7925*
	X4 + X5 + X9	2.281,0	*p < 0,0001*
Etapa 4	X4 + X5 + X6 + X4 * X5	0,34	*p = 0,8539*
	X4 + X5 + X6 + X4 * X6	2,98	*p = 0,0845*
	X4 + X5 + X6 + X5 * X6	0,41	*p = 0,5222*

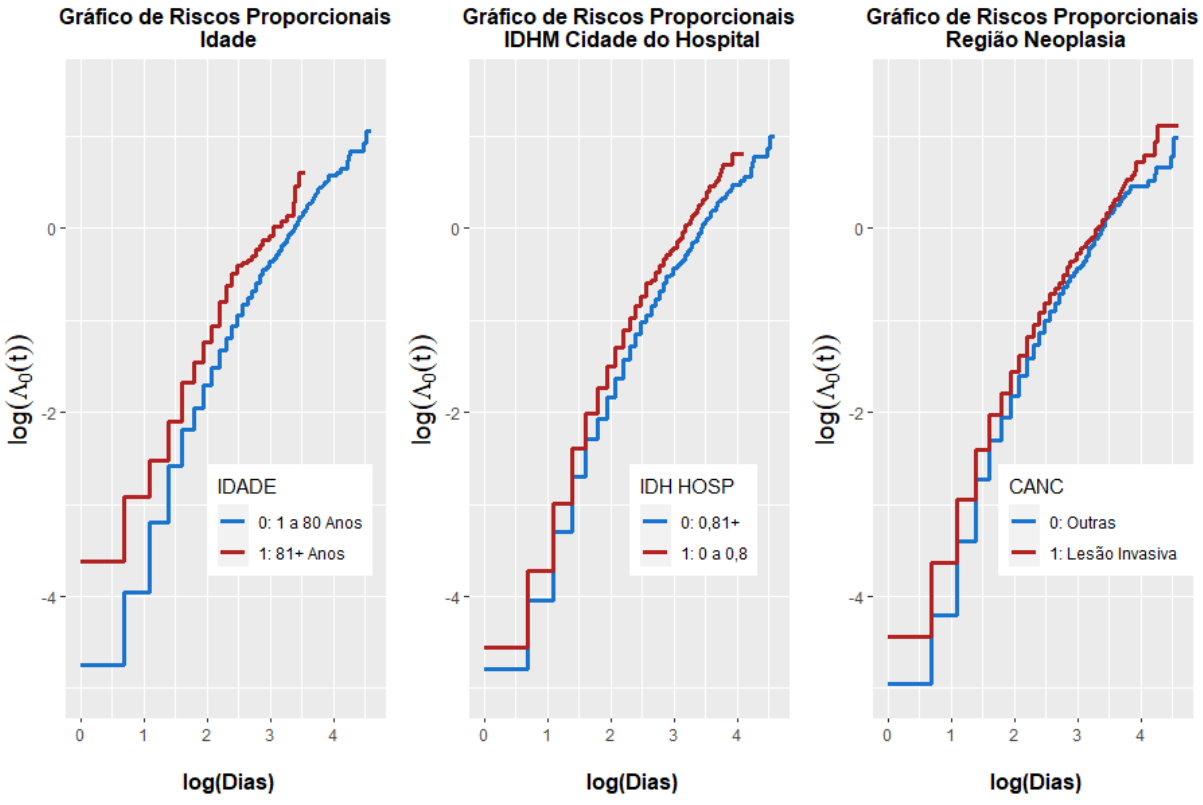
Fonte: Elaborado pelo autor, 2022.

8.2 ADEQUAÇÃO DO MODELO

Nesta seção foi feita a verificação da adequação do modelo de Cox final, com técnicas gráficas e pelo teste de resíduos padronizados de Schoenfeld, a fim de avaliar a suposição de riscos proporcionais ao longo do tempo. Após isto, foi mostrado os gráficos dos resíduos *martingal* e *deviance* do modelo ajustado, para verificar se há a presença de pontos atípicos.

A maneira tradicional de verificar a principal suposição é fazendo o gráfico do logaritmo da taxa de falha acumulada, $\widehat{\Lambda}_{0j}(t)$, sobre o tempo, ou até mesmo o logarítmo do tempo. A Figura 14 mostra este gráfico para cada variável do modelo, pode-se observar que as curvas não se cruzam para nenhuma das variáveis, mesmo havendo alguns desvios em relação a proporção dos riscos, não há evidências de que os desvios possam sugerir uma séria violação da suposição.

Figura 14 – Gráfico dos riscos proporcionais para cada variável do modelo de Cox



Fonte: Elaborado pelo autor, 2022.

A Tabela 10 e a Figura 15 apresentam, para as mesmas variáveis, os testes de proporcionalidade dos riscos e os gráficos dos resíduos padronizados de Schoenfeld, respectivamente. Analisando os p-valores na Tabela 10, a variável CANC sugere uma possível

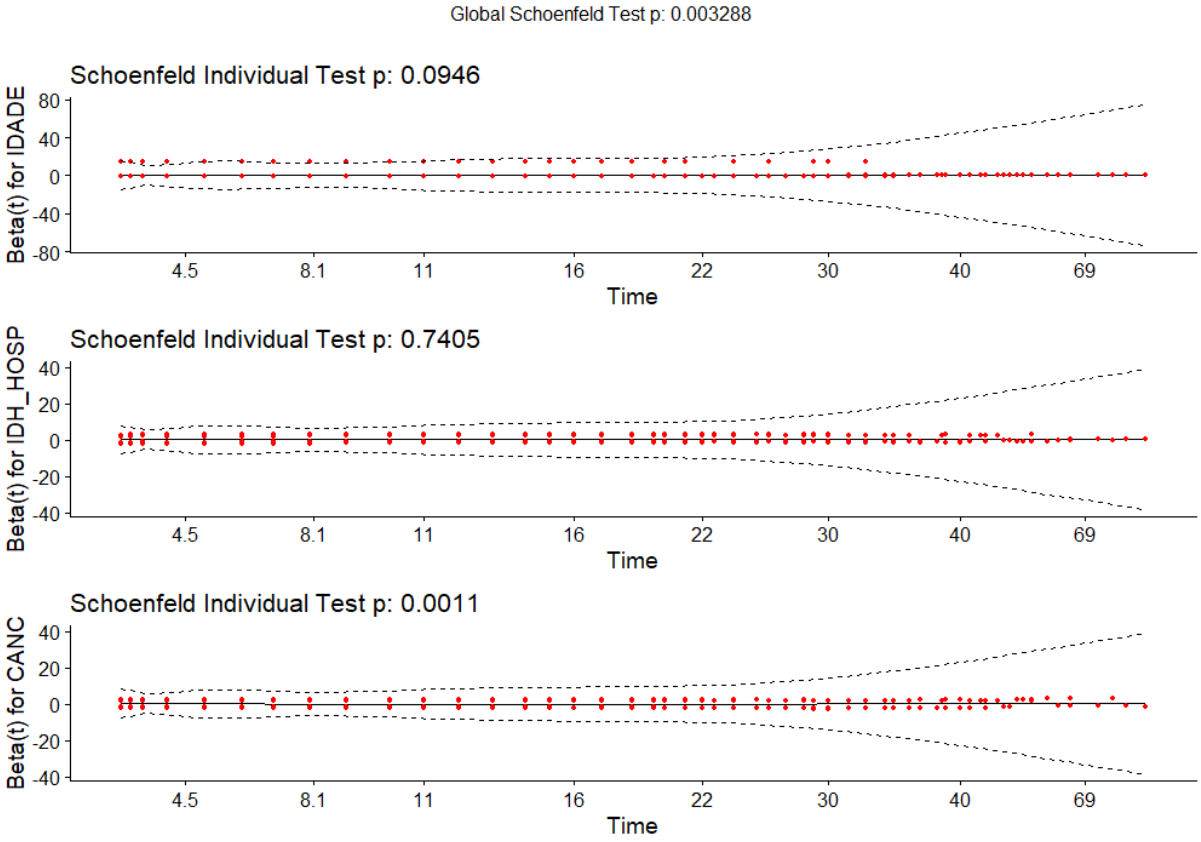
violação da suposição de riscos proporcionais ($p = 0{,}006$), no entanto, foi visto na Figura 12 que as curvas de taxa de falha acumulada não se cruzam em nenhum momento, desta forma, não há evidências de séria violação da suposição de riscos proporcionais.

Tabela 10 – Testes da proporcionalidade dos riscos do modelo ajustado

VARIÁVEL	χ^2	P-VALOR
IDADE	2,830	p = 0,0925
IDH_HOSP	0,045	p = 0,8233
CANC	7,442	p = 0,0064
GLOBAL	10,464	p = 0,0150

Fonte: Elaborado pelo autor, 2022.

Figura 15 – Gráfico dos resíduos padronizados de Schoenfeld

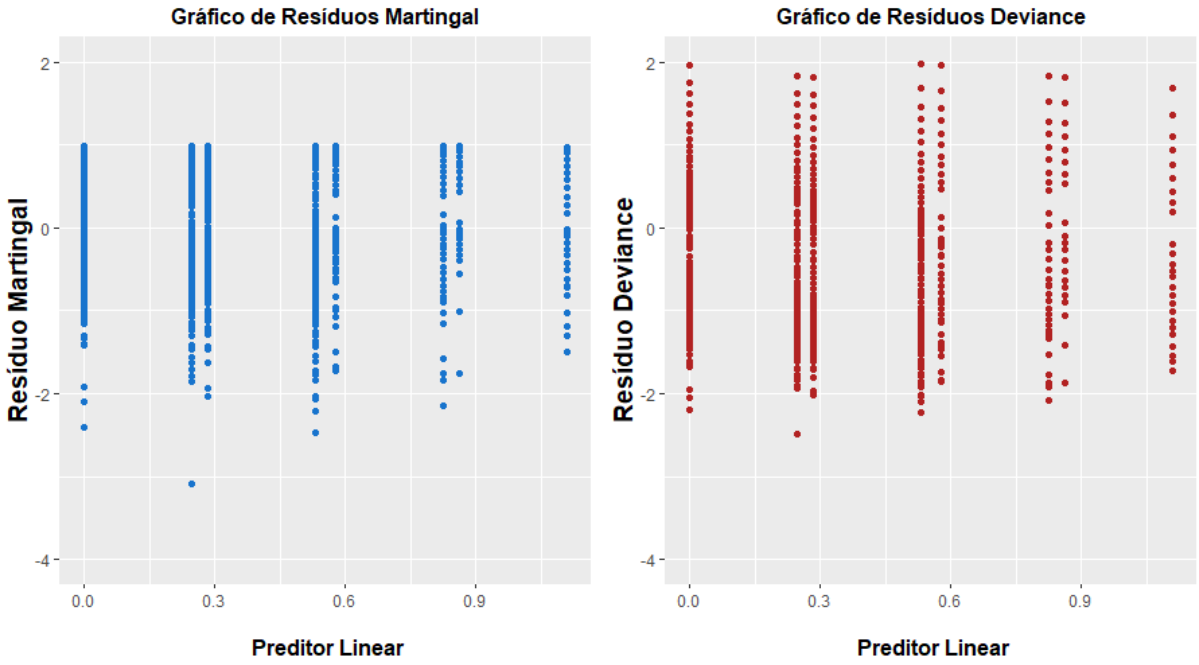


Fonte: Elaborado pelo autor, 2022.

Para a checagem de pontos atípicos, foi construído os gráficos de resíduos de *martingal* e *deviance* do modelo ajustado, respectivamente. Na Figura 16 é possível ver que não há pontos que possam ser considerados atípicos (*outliers*), talvez com uma exceção ao resíduo *martingal* de valor igual a $-3{,}08$. Para este resíduo em específico, se tem, contudo, um correspondente

resíduo *deviance* de $-2{,}48$, o qual é um valor aceitável dentro da variação observada para estes resíduos.

Figura 16 – Resíduos *martingal* e *deviance versus* preditor linear do modelo de Cox ajustado



Fonte: Elaborado pelo autor, 2022.

O comportamento dos resíduos *deviance* em torno de zero, visto na parte direita da Figura 16, fornece indicativos favoráveis à adequação do modelo de Cox ajustado.

8.3 INTERPRETAÇÃO E ANÁLISE DOS RESULTADOS

A partir das verificações de adequação do modelo ajustado, é possível seguir com as interpretações dos coeficientes, assim como as análises das sobrevivências e riscos estimados pelo modelo.

Diante disto, os resultados obtidos do ajuste do modelo de riscos proporcionais de Cox com as variáveis IDADE, IDH_HOSP e CANC, encontram-se na Tabela 11. É possível afirmar que pacientes com mais de 80 anos de idade possuem um risco de falecer aproximadamente 78,4% maior do que pacientes com menos de 80 anos, além disso, pode-se dizer com 95% de confiança estatística que o risco varia entre 53% e 108%. Para as outras duas variáveis, a interpretação é análoga, pacientes internados em cidades com IDHM menor que 0,80 pontos,

possuem um risco de 33% a mais de morte, comparado a pacientes internados em cidades com IDHM maior que 0,80 pontos. Por fim, há um risco de 28% maior de falecimento em pacientes com neoplasia do tipo “Lesão Invasiva” do que outros tipo/região (Mamilo, região interna e região externa), ou um risco que varia entre 18% e 39%, com um nível de confiança de 95%.

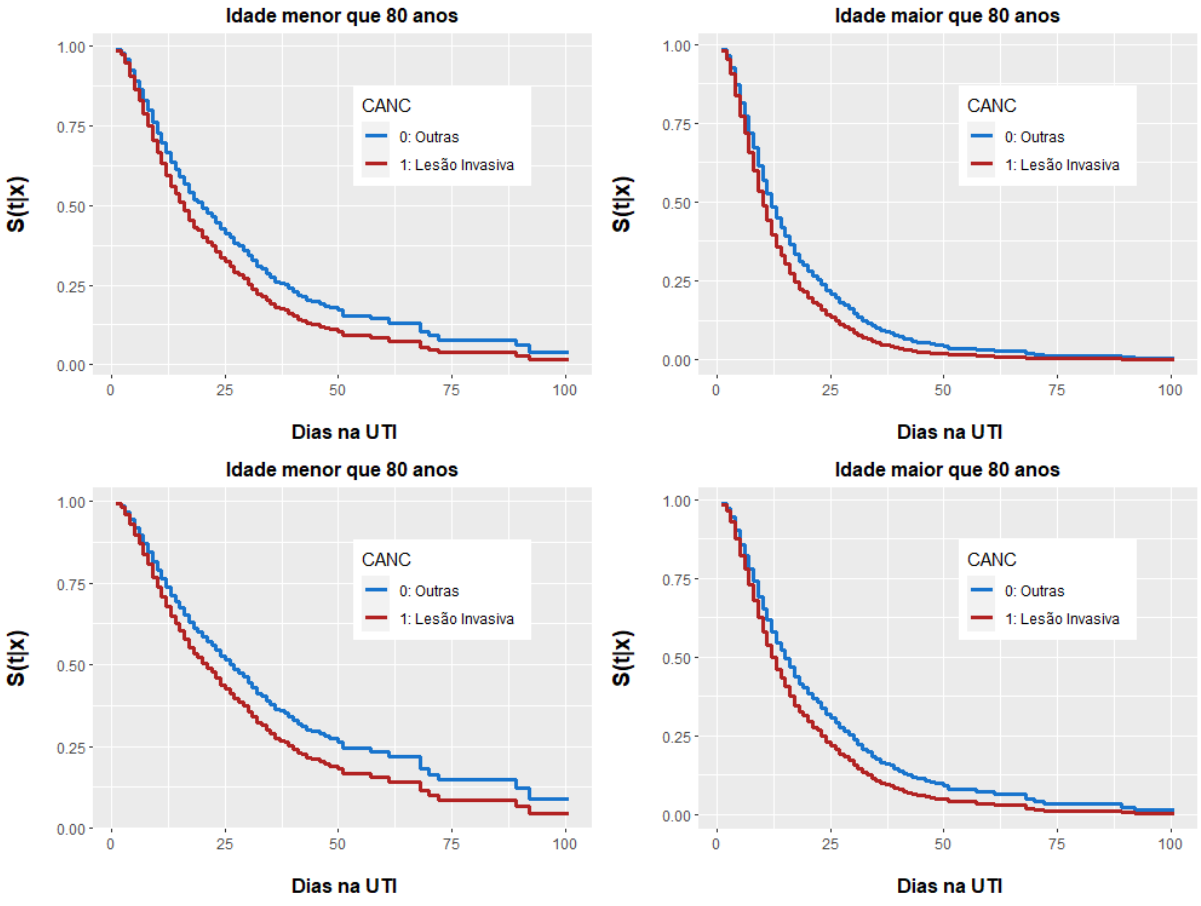
Tabela 11 – Resultado do ajuste do modelo de Cox e correspondentes razões de risco (RR)

VARIÁVEL	ESTIMATIVA	P-VALOR	RR	$IC_{95\%}(RR)$
IDADE	0,579	$p < 0,0001$	1,784	(1,53; 2,08)
IDH_HOSP	0,284	$p < 0,0001$	1,329	(1,23; 1,44)
CANC	0,247	$p < 0,0001$	1,280	(1,18; 1,39)

Fonte: Elaborado pelo autor, 2022.

A Figura 17 mostra os resultados das sobrevivências estimadas do modelo de Cox para cada cenário entre as três variáveis, importante notar que cada gráfico compara a idade com a

Figura 17 – Sobrevivências estimadas pelo modelo de Cox

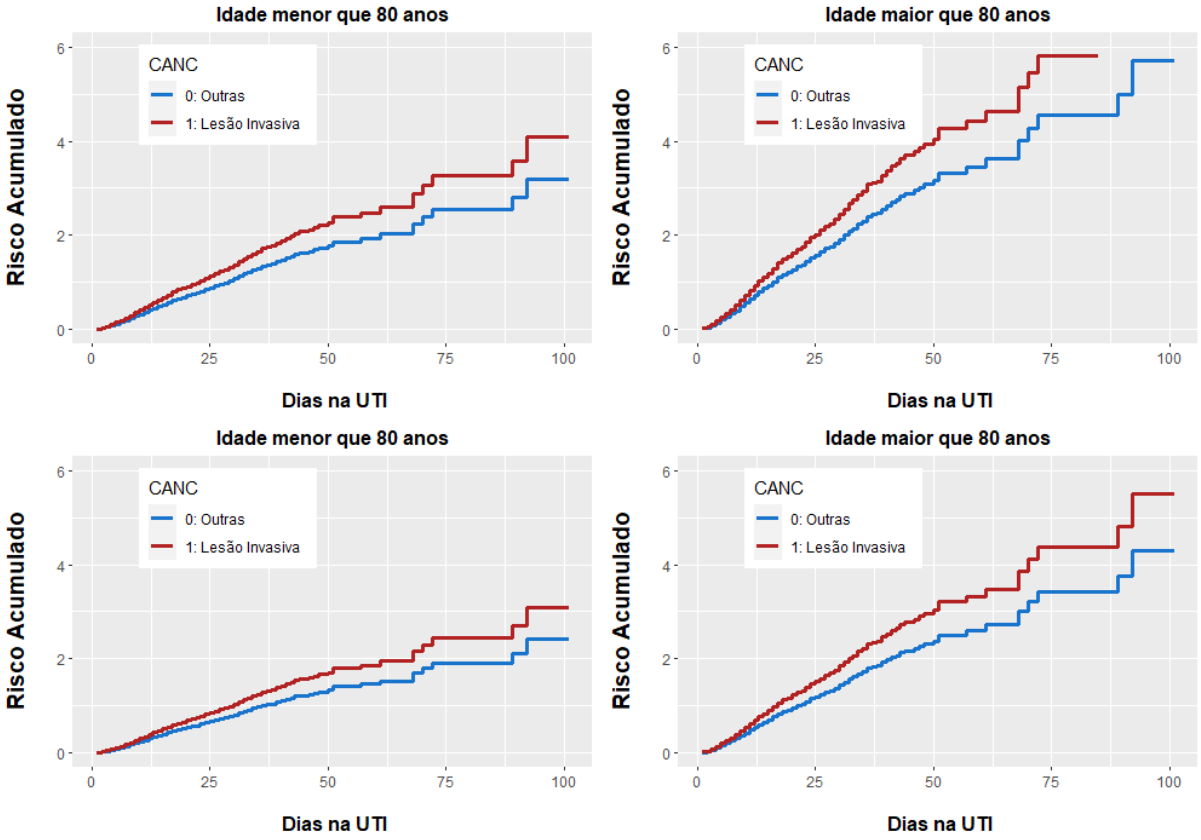


Fonte: Elaborado pelo autor, 2022.

região da neoplasia, e os gráficos na parte de cima são fixados IDH_HOSP = 1 (0 a 0,80 pontos), na parte de baixo são fixados IDH_HOSP = 0 (mais de 0,80 pontos). É claramente notável a diferença das curvas de sobrevivência comparando-se a idade da paciente, as curvas em que IDH_HOSP = 1 são um pouco mais acentuadas do que as curvas de sobrevivência em que IDH_HOSP = 0. A probabilidade estimada de sobrevivência de duas semanas de uma paciente com mais de 80 anos e com neoplasia sendo Lesão Invasiva, fixando IDH_HOSP = 1, é de aproximadamente 33%, comparando com o mesmo cenário, mas com IDH_HOSP = 0, a probabilidade resulta em 43%, aproximadamente. Se for comparado o cenário mais extremo, paciente com idade maior que 80 anos, com Lesão Invasiva e IDH_HOSP = 1, contra uma paciente com menos de 80 anos, com outro tipo/região de neoplasia e IDH_HOSP = 0, as probabilidades de sobrevivência estimadas para uma semana, são respectivamente, de 66% e 87%, aproximadamente.

Na Figura 18 são apresentados os gráficos de taxa de risco acumulado para os diferentes cenários. Como esperado, os resultados apresentados de forma gráfica, corroboram com a discussão dos resultados da Tabela 11, ou seja, as pacientes com mais de 80 anos possuem um risco significantemente maior do que pacientes com menos de 80 anos de idade. Pacientes com idade maior que 80 anos, com Lesão Invasiva e IDH_HOSP = 1 possuem quase o dobro de risco de falecer no 50º dia de internação, comparando pacientes com as mesmas condições, mas com menos de 80 anos.

Figura 18 – Riscos estimados pelo modelo de Cox



Fonte: Elaborado pelo autor, 2022.

9 CONSIDERAÇÕES FINAIS

Devido a magnitude do câncer de mama no Brasil e no mundo, sendo o câncer mais comum entre as mulheres, este trabalho buscou conhecer como se comporta(m) a(s) curva(s) de sobrevivência(s) de mulheres que foram internadas em hospitais públicos do SUS e entender quais covariáveis afetam a longevidade das pacientes.

Foi analisado o conjunto de dados disponibilizado pelo SIHSUS (Sistema de Informações Hospitalares do SUS), no site do DATASUS, entre os anos de 2016 a 2020, totalizando 18.280 observações. Para atingir o objetivo do trabalho, foram traçados alguns objetivos específicos: Além de realizar a análise exploratória, conhecer a curva de sobrevivência das pacientes de forma não-paramétrica, com o estimador de Kaplan-Meier. Definir um modelo paramétrico que possui melhor ajuste em relação a curva de K.M. e identificar fatores que afetam a sobrevida das pacientes utilizando modelos de regressão e modelo de Cox.

De maneira geral, os resultados obtidos neste trabalho foram satisfatórios. A análise exploratória foi breve, porém, com bastante informação sobre como cada covariável se comporta em relação a censura e evento de interesse, o tempo em dias em que a paciente fica internada na UTI até seu recebimento de alta do hospital. Com o estimador de Kaplan-Meier, foi possível conhecer a curva de sobrevivência das pacientes de forma não-paramétrica. Na análise paramétrica, foi visto que o modelo mais adequado para estimar as sobrevivências das pacientes é com a distribuição Log-normal. O modelo de regressão final ficou com as covariáveis inseridas: idade da paciente, tipo de UTI utilizada, IDHM do município do hospital e a região da neoplasia, também com distribuição Log-normal. Para avaliar a adequação do modelo, foram analisados os resíduos padronizados e de Cox-Snell do modelo. Na seção do modelo de Cox, foram feitas algumas alterações nas faixas de idade e IDHM, para a região da neoplasia, foi denotado com ou sem lesão invasiva. Estas modificações foram realizadas para que os pressupostos do modelo fossem atendidos. O modelo de Cox final ficou com as covariáveis inseridas: idade da paciente, região da neoplasia e IDHM do município do hospital. Os métodos de avaliação da adequação do modelo de Cox foram feitas por: análises gráficas, verificando se a razão dos riscos fica proporcional ao longo do tempo e os resultados dos resíduos de Schoenfeld, Martingal e Deviance.

Na base de dados disponibilizada pelo SIHSUS, ainda há informações faltantes que poderiam ser cruciais para desenvolver o trabalho de estimar as sobrevivências das pacientes com câncer de mama. O tamanho do tumor na mama da paciente, se ela é fumante ou não, número de filhos e uso de outros medicamentos, por exemplo. O custo de processamento/manutenção para conseguir tais informações não seriam altos (talvez, exceto o tamanho do tumor) e agregariam bastante para futuros estudantes e pesquisadores.

REFERÊNCIAS

ABREU, E. de; KOIFMAN, S. Fatores prognósticos no câncer da mama feminina. **Revista Brasileira de Cancerologia**, v. 48, n. 1, p. 113–31, 2002.

APOSTOLOU, P.; FOSTIRA, F. Hereditary breast cancer: the era of new susceptibility genes. **BioMed Research International**, v. 2013.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, R. L.; TORRE, L. A.; JEMAL, A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. **CA**: a Cancer Journal for Clinicians, v. 68, p. 394 – 424, 2018.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevivência**: teoria e aplicações em saúde. 2. ed. [S.l.]: Editora Fiocruz, 2011.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. 1. ed. [S.l.]: Editora Edgard Blücher, 2006.

COX, D. R. Regression Models and Life Tables (with discution). **Journal Royal Statistical Society**, B, 34, p. 187 – 220, 1972.

COX, D. R.; HINKLEY, D. V. **Theoretical Statistics**. 1. ed. [S.l.]: Chapman and Hall, 1974.

COX, D. R.; SNELL, E. J. A General Definition of Residuals. **Journal of the Royal Statistical Society B**, 30, p. 248 – 275, 1968.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA. **Estimativa 2020**: incidência do Câncer no Brasil. Rio de Janeiro: INCA, 2019. Disponível em: https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//.... Acesso em: 31 agosto 2022.

INSTITUTO NACIONAL DE CÂNCER, inca.gov.br, 2020. Disponível em: inca.gov.br/controle-do-cancer-de-mama/conceito-e-magnitude. Acesso em: 5 jan. 2021.

JACKSON, C. **Flexsurv**. R package version 1.1.1. 2019.

KAPLAN, E. L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, New York, v. 53, p. 457 – 481, 1958.

LAWLESS, J. F. Inference in the Generalized Gamma and Log Gamma Distributions. **Technometrics**, v. 22, p. 409 – 419, 1980.

LEVY-LAHAD, E.; FRIEDMAN, E. Cancer risks among BRCA1 and BRCA2 mutation carriers. **British Journal of Cancer**. v. 96, n. 1, p. 11 – 15, 2007.

PAULA, G. A. **Modelos de Regressão**: com apoio computacional. Instituto de Matemática e Estatística Universidade de São Paulo. [S.l.], 2013. Disponível em: ime.usp.br/~giapaula/texto_2013.pdf. Acesso em: 5 jan. 2021.

RDSP(ano-mês): banco de dados. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Acesso em 31 agosto 2022.
SANTOS, R. d. S. **Sobrevivência de mulheres com diagnóstico de Câncer de Mama no município do Rio de Janeiro**. Dissertação (Mestrado) – Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro, 2013.

SCHILITHZ, A. O. C.; ALMEIDA, L. M.; GUIMARÃES, M. T. C.; SOUZA, M. C.; ASSIS, M. **A Situação do Câncer de Mama no Brasil**: síntese de dados dos sistemas de informação. Rio de Janeiro, Instituto Nacional de Câncer, 2019. Disponível em: inca.gov.br/sites/ufu.sti.inca.local/files//media/document//a_situacao_ca_mama_brasil_2019.pdf. Acesso em: 5 jan. 2021.

STACY, E. W. A Generalization of the Gamma Distribution. **Institute of mathematical Statistics**, v. 33, p. 1187 – 1192, 1962.

TEAM, R. C. **A language and environment for statistical computing**. Vienna, Autria: R Foundation for Statistical Computing. Disponível em: www.R-project.org.

THERNEAU, T. M. **Survival**. R package version 3.2-7. 2020.

APÊNDICE – Códigos utilizados no software R para a elaboração do trabalho

```
library(read.dbc)
library(dplyr)
library(lattice)
library(survival)
library(ggplot2)
library(xlsx)
library(readxl)
library(survminer)
library(devtools)
library(ggfortify)
library(lmtest)
library(flexsurv)
library(gridExtra)

# Gráfico das idades por diferentes IDHMs

ggplot(dados, aes(IDADE, fill = MORTE)) +
  geom_histogram(binwidth = 10, col = "grey90") +
  facet_wrap(~ IDH_TEXTO) +
  ylab('Quantidade') + xlab('Idade (anos)') +
  ggtitle('Distribuição das idades por diferentes IDHMs') +
  theme(plot.title = element_text(hjust = 0.5, face = 'bold', colour = 'black'),
axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
axis.title.y = element_text(face = 'bold', colour = 'black', size = 10))

# Box-Plot das covariáveis contínuas da base de dados

grid.arrange(

ggplot(dados, aes(y = IDADE)) +
  geom_boxplot(fill = "#4271AE", colour = "#1F3552") +
  ylab('Idade (anos)') +
 # ggtitle('Box-plot das idades das pacientes') +
  theme(plot.title = element_text(hjust = 0.5),
  axis.ticks.x=element_blank(),
  axis.title.x=element_blank(),
  axis.text.x=element_blank(),
  axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)),

ggplot(dados, aes(y = DIAS_PERM)) +
  geom_boxplot(fill = "lightgreen", colour = "green4") +
  ylab('Dias de permanência na UTI') +
 # ggtitle('Box-plot dos dias de permanência na UTI') +
  theme(plot.title = element_text(hjust = 0.5),
        axis.ticks.x=element_blank(),
        axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)),
```

```
ggplot(dados, aes(y = IDH_HOSP)) +
  geom_boxplot(fill = "lightcoral", colour = "indianred") +
  ylab('IDHM da cidade do hospital') +
 # ggtitle('Box-plot dos IDHs dos hospitais das cidades') +
  theme(plot.title = element_text(hjust = 0.5),
        axis.ticks.x=element_blank(),
        axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)),
  ncol = 3)

# Estimativas de Kaplan Meier

ekm = survfit(Surv(tempo, morte) ~ 1)

autoplot(ekm, surv.colour = 'blue', surv.size = 1, censor = FALSE,
         main = 'Curva Kaplan-Meier', xlab = 'Dias', ylab = 'S(t)')

autoplot(ekm, surv.size = 1.2, censor = FALSE, surv.colour = 'black') +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Curva de Sobrevivência
Kaplan-Meier \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 12))

# Calculando Tempo médio e Mediano

test_tm = as.data.frame(list(tempo = tempo, morte = morte))

t  = select(subset(test_tm[order(test_tm$tempo), ], morte == 1), tempo)
t  = distinct(t)
t  = as.numeric(unlist(t))

surv = as.numeric(sort(c(1, ekm$surv), decreasing = T))

tm = c(NULL)

for ( i in 1:55 ){
  tm[i] = surv[i]*(t[i+1] - t[i])
}

sum(tm)

# Gráfico das curvas de sobrevivência para diferentes covariáveis

grid.arrange(

autoplot(survfit(Surv(tempo, morte) ~ complexidade), surv.size = 1, conf.int = FALSE,
censor = FALSE) +
```

```
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Complexidade do
Procedimento \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'COMPLEX'),

autoplot(survfit(Surv(espec9$DIAS_PERM, espec9$MORTE) ~ espec9$ESPEC), surv.size =
1, conf.int = FALSE, censor = FALSE) +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Especialidade do Leito \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'ESPEC'),

autoplot(survfit(Surv(tempo, morte) ~ idh), surv.size = 1, conf.int = FALSE, censor =
FALSE) +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'IDH da Cidade do Hospital
\n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'IDH_HOSP'),


autoplot(survfit(Surv(tempo, morte) ~ idade_agr), surv.size = 1, conf.int = FALSE, censor =
FALSE) +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Idade da Paciente \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'IDADE'),


autoplot(survfit(Surv(tempo, morte) ~ cancer), surv.size = 1, conf.int = FALSE, censor =
FALSE) +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Região da Neoplasia \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'CANC'),


autoplot(survfit(Surv(tempo, morte) ~ uti_util), surv.size = 1, conf.int = FALSE, censor =
FALSE) +
  labs(x = '\n Tempo na UTI (dias)', y = 'S(t) Estimada \n', title = 'Tipo de UTI utilizada \n') +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 10)) +
  scale_color_discrete(name = 'MARCA_UTI'),
```

```
  ncol = 2, nrow = 3 )

# Fazendo as comparações das curvas utilizando o Teste Log-Rank

survdif_complex = survdiff(Surv(tempo, morte) ~ complexidade)
survdif_cancer  = survdiff(Surv(tempo, morte) ~ cancer)
survdif_carint  = survdiff(Surv(tempo, morte) ~ car_int)
survdif_raca    = survdiff(Surv(tempo, morte) ~ raca)
survdif_idade   = survdiff(Surv(tempo, morte) ~ idade_agr)
survdif_idh     = survdiff(Surv(tempo, morte) ~ idh)
survdif_uti     = survdiff(Surv(tempo, morte) ~ uti_util)
survdif_espec   = survdiff(Surv(tempo, morte) ~ espec)
survdif_proc    = survdiff(Surv(tempo, morte) ~ proc)

#### Modelagem Paramétrica ####

# Para modelar com Weibull, não pode-se ter Tempos = 0

dados2 = subset(dados, DIAS_PERM > 0)

ekm = survfit(Surv(dados2$DIAS_PERM, dados2$MORTE) ~ 1)

# Weibull

weib = survreg(Surv(DIAS_PERM, MORTE) ~ 1, dist = 'weibull', dados2)

alpha_weib = exp( weib$coefficients[1])
gama_weib  = 1/weib$scale

time = ekm$time
st   = ekm$surv

St_weib = exp( - (( time / alpha_weib )^gama_weib ) )

# Exponencial

expo       = survreg(Surv(DIAS_PERM, MORTE) ~ 1, dist = 'exponential', dados2)
alpha_expo = exp( expo$coefficients[1] )

St_expo    = exp( - time / alpha_expo)

# Log-Normal

log_norm = survreg(Surv(DIAS_PERM, MORTE) ~ 1, dist = 'lognorm', dados2)

St_ln    = pnorm((-log(time) + 2.948)/ 1.123852)
```

```
# Método Gráfico

par(mfrow = c(1, 3))

plot(st, St_expo, pch = 16, ylim = range(c(0,1)), xlim = range(c(0,1)),
    xlab = 'S(t): Kaplan-Meier', ylab = 'S(t): Exponencial', col = 'green')
lines(c(0,1), c(0,1), lwd = 3, lty = 1)

plot(st, St_weib, pch = 16, ylim = range(c(0,1)), xlim = range(c(0,1)),
    xlab = 'S(t): Kaplan-Meier', ylab = 'S(t): Weibull', col = 'blue')
lines(c(0,1), c(0,1), lwd = 3, lty = 1)

plot(st, St_ln, pch = 16, ylim = range(c(0,1)), xlim = range(c(0,1)),
    xlab = 'S(t): Kaplan-Meier', ylab = 'S(t): Log-Normal', col = 'red')
lines(c(0,1), c(0,1), lwd = 3, lty = 1)

# Gráficos

par(mfrow = c(1, 3))

plot(ekm, conf.int = TRUE, xlab = 'Tempos', ylab = 'S(t)', lwd = 1.5)
lines(c(0, time), c(1, St_expo), lwd = 3, col = 'green')

plot(ekm, conf.int = TRUE, xlab = 'Tempos', ylab = 'S(t)', lwd = 1.5)
lines(c(0, time), c(1, St_weib), lwd = 3, col = 'blue')

plot(ekm, conf.int = TRUE, xlab = 'Tempos', ylab = 'S(t)', lwd = 1.5)
lines(c(0, time), c(1, St_ln), lwd = 3, col = 'red')

## Modelo de Regressão Paramétrico ##

# Método Forward Stepwise #

# PASSO 1

fw_1_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ 1,          data = dados_models, dist =
'gengamma')
fw_1_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ COMPLEX,     data =
dados_models, dist = 'gengamma')
fw_1_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ CAR_INT,     data = dados_models,
dist = 'gengamma')
fw_1_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ MARCA_UTI,   data =
dados_models, dist = 'gengamma')
fw_1_4 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2,  data =
dados_models, dist = 'gengamma')
fw_1_5 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ CANC2_AUX,   data =
dados_models, dist = 'gengamma')
fw_1_6 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDH_HOSP_AUX, data =
dados_models, dist = 'gengamma')
```

```
fw_1_7 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ RACA_COR,     data =
dados_models, dist = 'gengamma')
fw_1_8 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ PROC_REA_AUX, data =
dados_models, dist = 'gengamma')
fw_1_9 = flexsurvreg(Surv(DIAS_PERM, MORTE)   ~ ESPEC,        data = dados_models,
dist = 'gengamma')

# TRV PASSO 1 #tcc 5

lrtest(fw_1_1, fw_1_0) # p < 0.0001
lrtest(fw_1_2, fw_1_0) # p < 0.0001
lrtest(fw_1_3, fw_1_0) # p < 0.4698
lrtest(fw_1_4, fw_1_0) # p < 0.0001
lrtest(fw_1_5, fw_1_0) # p < 0.0001
lrtest(fw_1_6, fw_1_0) # p < 0.0001
lrtest(fw_1_7, fw_1_0) # p < 0.05055
lrtest(fw_1_8, fw_1_0) # p < 0.1736
lrtest(fw_1_9, fw_1_0) # p < 0.0001

# PASSO 2

fw_2_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + COMPLEX,
data = dados_models, dist = 'gengamma')
fw_2_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CAR_INT,      data
= dados_models, dist = 'gengamma')
fw_2_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI,
data = dados_models, dist = 'gengamma')
fw_2_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX,
data = dados_models, dist = 'gengamma')
fw_2_4 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + IDH_HOSP_AUX,
data = dados_models, dist = 'gengamma')
fw_2_5 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + RACA_COR,
data = dados_models, dist = 'gengamma')
fw_2_6 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
PROC_REA_AUX, data = dados_models, dist = 'gengamma')
fw_2_7 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + ESPEC,        data =
dados_models, dist = 'gengamma')

# TRV PASSO 2

lrtest(fw_2_0, fw_1_4) # p < 0.0001
lrtest(fw_2_1, fw_1_4) # p < 0.0001
lrtest(fw_2_2, fw_1_4) # p < 0.6966
lrtest(fw_2_3, fw_1_4) # p < 0.0001
lrtest(fw_2_4, fw_1_4) # p < 0.0001
lrtest(fw_2_5, fw_1_4) # p < 0.05477
lrtest(fw_2_6, fw_1_4) # p < 0.1899
lrtest(fw_2_7, fw_1_4) # p < 0.0001

# PASSO 3
```

```
fw_3_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
COMPLEX,      data = dados_models, dist = 'gengamma')
fw_3_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
CAR_INT,      data = dados_models, dist = 'gengamma')
fw_3_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI,    data = dados_models, dist = 'gengamma')
fw_3_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
IDH_HOSP_AUX, data = dados_models, dist = 'gengamma')
fw_3_4 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
RACA_COR,     data = dados_models, dist = 'gengamma')
fw_3_5 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
PROC_REA_AUX, data = dados_models, dist = 'gengamma')
fw_3_6 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
ESPEC,        data = dados_models, dist = 'gengamma')

# TRV PASSO 3

lrtest(fw_2_3, fw_3_0) # p < 0.0001
lrtest(fw_2_3, fw_3_1) # p < 0.0001
lrtest(fw_2_3, fw_3_2) # p < 0.0001
lrtest(fw_2_3, fw_3_3) # p < 0.0001
lrtest(fw_2_3, fw_3_4) # p < 0.0001
lrtest(fw_2_3, fw_3_5) # p < 0.0001
lrtest(fw_2_3, fw_3_6) # p < 0.0001

# PASSO 4

fw_4_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + COMPLEX,      data = dados_models, dist = 'gengamma')
fw_4_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + CAR_INT,      data = dados_models, dist = 'gengamma')
fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + IDH_HOSP_AUX, data = dados_models, dist = 'gengamma')
fw_4_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + RACA_COR,     data = dados_models, dist = 'gengamma')
fw_4_4 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + PROC_REA_AUX, data = dados_models, dist = 'gengamma')
fw_4_5 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI + ESPEC,        data = dados_models, dist = 'gengamma')

# TRV PASSO 4

lrtest(fw_3_2, fw_4_0) # p < 0.0001
lrtest(fw_3_2, fw_4_1) # p < 0.0001
lrtest(fw_3_2, fw_4_2) # p < 0.0001
lrtest(fw_3_2, fw_4_3) # p < 0.012
lrtest(fw_3_2, fw_4_4) # p < 0.4053
lrtest(fw_3_2, fw_4_5) # p < 0.0001
```

```
# PASSO 4.1 (INTERAÇÕES)

t0_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + IDADE_AUX2 * CANC2_AUX, data =
dados_models, dist = 'gengamma')
t1_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + IDADE_AUX2 * MARCA_UTI, data =
dados_models, dist = 'gengamma')
t2_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + IDADE_AUX2 * IDH_HOSP_AUX, data =
dados_models, dist = 'gengamma')
t3_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX * MARCA_UTI, data =
dados_models, dist = 'gengamma')
t4_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX * IDH_HOSP_AUX, data =
dados_models, dist = 'gengamma')
t5_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + CANC2_AUX
+ MARCA_UTI + IDH_HOSP_AUX + MARCA_UTI * IDH_HOSP_AUX, data =
dados_models, dist = 'gengamma')

# TRV PASSO 4.1

lrtest(t0_fw_4_2, fw_4_2) # p < 0.1404
lrtest(t1_fw_4_2, fw_4_2) # p < 0.0001
lrtest(t2_fw_4_2, fw_4_2) # p = 0.8127
lrtest(t3_fw_4_2, fw_4_2) # p = 0.2041
lrtest(t4_fw_4_2, fw_4_2) # p < 0.9683
lrtest(t5_fw_4_2, fw_4_2) # p < 0.1273

## MODELO FINAL : t5_fw_4_2

# MODELOS DE INTERESSE

weib_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'weibull')
ln_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI
+ IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'lognormal')
llog_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'llog')

lrtest(fw_4_2, weib_fw_4_2) # p < 0.0001
lrtest(fw_4_2, ln_fw_4_2)   # p = 0.2752
lrtest(fw_4_2, llog_fw_4_2) # p < 0.0001


# MODELOS DE INTERESSE COM INTERAÇÃO
```

```
t_exp_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX + IDADE_AUX2 * MARCA_UTI, data
= dados_models, dist = 'exponential')
t_weib_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX + IDADE_AUX2 * MARCA_UTI, data
= dados_models, dist = 'weibull')
t_ln_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX + IDADE_AUX2 * MARCA_UTI, data
= dados_models, dist = 'lognormal')
t_llog_fw_4_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 +
MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX + IDADE_AUX2 * MARCA_UTI, data
= dados_models, dist = 'llog')


lrtest(t1_fw_4_2, t_exp_fw_4_2) # p < 0.0001
lrtest(t1_fw_4_2, t_weib_fw_4_2) # p < 0.0001
lrtest(t1_fw_4_2, t_ln_fw_4_2)   # p = 0.5454
lrtest(t1_fw_4_2, t_llog_fw_4_2) # p < 0.0001



# PASSO 5

fw_5_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + COMPLEX,      data = dados_models, dist =
'gengamma')
fw_5_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + CAR_INT,      data = dados_models, dist =
'gengamma')
fw_5_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + RACA_AUX,     data = dados_models, dist =
'gengamma')
fw_5_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + PROC_REA_AUX, data = dados_models, dist =
'gengamma')
fw_5_4 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC,        data = dados_models, dist = 'gengamma')

# TRV PASSO 5

lrtest(fw_4_2, fw_5_0) # p < 0.0001
lrtest(fw_4_2, fw_5_1) # p < 0.0001
lrtest(fw_4_2, fw_5_2) # p < 0.9028
lrtest(fw_4_2, fw_5_3) # p < 0.3924
lrtest(fw_4_2, fw_5_4) # p < 0.0001

# PASSO 6
```

```
fw_6_0 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC + COMPLEX, data = dados_models, dist =
'gengamma')
fw_6_1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC + CAR_INT, data = dados_models, dist =
'gengamma')
fw_6_2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC + RACA_AUX, data = dados_models, dist =
'gengamma')
fw_6_3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC + PROC_REA_AUX, data = dados_models,
dist = 'gengamma')

# TRV PASSO 6

lrtest(fw_5_4, fw_6_0) # p < 0.0001
lrtest(fw_5_4, fw_6_1) # p < 0.0001
lrtest(fw_5_4, fw_6_2) # p < 0.3853
lrtest(fw_5_4, fw_6_3) # p < 0.0001

# TESTE

mod = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'llog')
mod_geng = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI
+ IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'gengamma')

mod1 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + COMPLEX +
IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'lognormal')
mod2 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + ESPEC, data = dados_models, dist = 'lognormal')
mod3 = flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ COMPLEX, data = dados_models, dist
= 'lognormal')
mod4= flexsurvreg(Surv(DIAS_PERM,  MORTE)  ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX, data = dados_models, dist = 'lognormal')


mf1 = c(1, as.data.frame(summary(mod1))[['est']])
mf2 = c(1, as.data.frame(summary(mod_geng))[['est']])
mf3 = c(1, as.data.frame(summary(t_ln_fw_4_2))[['est']])
mf4 = c(1, as.data.frame(summary(mod3))[['est']])

# GRÁFICO COM OS MODELOS

plot(ekm_dados_models, conf.int = TRUE, xlab = 'Tempos', ylab = 'S(t)', lwd = 1.5)

lines(c(0, ekm_dados_models$time), c(mf1), lwd = 1, col = 'green')
lines(c(0, ekm_dados_models$time), c(mf2), lwd = 1, col = 'yellow')
lines(c(0, ekm_dados_models$time), c(mf3), lwd = 1, col = 'red')
lines(c(0, ekm_dados_models$time), c(mf4), lwd = 1, col = 'blue')
```

```
# MODELO FINAL IDADE_AUX2 + MARCA_UTI + IDH_HOSP_AUX + CANC2_AUX

model_fw = flexsurvreg(Surv(DIAS_PERM,  MORTE) ~ IDADE_AUX2 + MARCA_UTI +
IDH_HOSP_AUX + CANC2_AUX + IDADE_AUX2 * MARCA_UTI, data =
dados_models, dist = 'lognormal') #

# ANÁLISE RESÍDUOS MODELO FW

Xb_model = model_fw$coefficients[1]                          +
        model_fw$coefficients[3] * dados_models$IDADE_AUX2   +
        model_fw$coefficients[4] * dados_models$MARCA_UTI    +
        model_fw$coefficients[5] * dados_models$IDH_HOSP_AUX +
        model_fw$coefficients[6] * dados_models$CANC2_AUX    +
        model_fw$coefficients[7] * dados_models$IDADE_AUX2 *
dados_models$MARCA_UTI


sigma_model = exp(model_fw$coefficients[2])

resid_model = exp(((log(dados_models$DIAS_PERM) - Xb_model)/sigma_model))

ekm_resid_model = survfit(Surv(resid_model, MORTE) ~ 1, data = dados_models)

Sln_model = pnorm(-log(ekm_resid_model$time))


# GRÁFICO DOS RESÍDUOS

par(mfrow = c(1,2))

plot(ekm_resid_model, xlab = 'Resíduos (ei*)', ylab = 'S^(t)', lwd = 1.5)
lines(ekm_resid_model$time, Sln_model, lwd = 2, col = 'blue')
legend('bottomleft', lty = c(1, 1), c('Kaplan-Meier', 'Log-normal padrão'), col = c('black',
'blue'), cex = 0.6, bty = 'n', lwd = c(1.5, 2))

plot(ekm_resid_model$surv, Sln_model, pch = 16, xlab = 'S(ei*): Kaplan-Meier', ylab =
'S(ei*): Log-normal padrão', col = 'blue')


#### RESÍDUOS COX-SNELL ####

ei = -log(1 - pnorm(((log(dados_models$DIAS_PERM) - Xb_model)/sigma_model)))

ekm1 = survfit(Surv(ei, dados_models$MORTE) ~ 1)

sexp = exp(-ekm1$time)
```

```
par(mfrow = c(1, 2))

plot(ekm1, conf.int = T, mark.time = F, xlab = 'Resíduos Cox-Snell', ylab = 'Sobrevivência
Estimada', lwd = 1.5)
lines(ekm1$time, sexp, lwd = 2, col = 'blue')

plot(ekm1$surv, sexp, xlab = 'S(ei): Kaplan-Meier', ylab = 'S(ei): Exponencial Padrão', pch =
16, col = 'blue')

## COMPARANDO CURVAS DE SOBREVIDA ##

## COMPARANDO CURVAS DE SOBREVIVÊNCIA DE DIFERENTES IDHs COM
IDADE SEGMENTADA ##
## MARCA_UTI FIXADO = 0 E REGIÃO NEOPLASIA NO MAMILO (= 2) ##

## S(t) = pnorm((-ln(t) + mu/(x'B))/sigma)

tempo = c(1:max(dados_models$DIAS_PERM))

# IDH baixo, jovem

xb1 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 0 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 0 +
       model_fw$coefficients[6] * 2

# IDH baixo, adulto

xb2     = model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 1 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 0 +
       model_fw$coefficients[6] * 2

# IDH baixo, meia idade

xb3 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 2 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 0 +
       model_fw$coefficients[6] * 2

# IDH baixo, idoso

xb4 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 3 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 0 +
       model_fw$coefficients[6] * 2
```

```
# IDH médio, jovem

xb5 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 0 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 2

# IDH médio, adulto

xb6 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 1 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 2

# IDH médio, meia idade

xb7 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 2 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 2

# IDH médio, idoso

xb8 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 3 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 2

# IDH alto, jovem

xb9 =    model_fw$coefficients[1]    +
       model_fw$coefficients[3] * 0 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 2 +
       model_fw$coefficients[6] * 2

# IDH alto, adulto

xb10 =    model_fw$coefficients[1]    +
        model_fw$coefficients[3] * 1 +
        model_fw$coefficients[4] * 0 +
        model_fw$coefficients[5] * 2 +
        model_fw$coefficients[6] * 2

# IDH alto, meia idade
```

```
xb11 =    model_fw$coefficients[1]    +
        model_fw$coefficients[3] * 2 +
        model_fw$coefficients[4] * 0 +
        model_fw$coefficients[5] * 2 +
        model_fw$coefficients[6] * 2

# IDH alto, idoso

xb12 =    model_fw$coefficients[1]    +
        model_fw$coefficients[3] * 3 +
        model_fw$coefficients[4] * 0 +
        model_fw$coefficients[5] * 2 +
        model_fw$coefficients[6] * 2


curva1  = pnorm((- log(tempo) + xb1)/exp(model_fw$coefficients[2]))
curva2  = pnorm((- log(tempo) + xb2)/exp(model_fw$coefficients[2]))
curva3  = pnorm((- log(tempo) + xb3)/exp(model_fw$coefficients[2]))
curva4  = pnorm((- log(tempo) + xb4)/exp(model_fw$coefficients[2]))
curva5  = pnorm((- log(tempo) + xb5)/exp(model_fw$coefficients[2]))
curva6  = pnorm((- log(tempo) + xb6)/exp(model_fw$coefficients[2]))
curva7  = pnorm((- log(tempo) + xb7)/exp(model_fw$coefficients[2]))
curva8  = pnorm((- log(tempo) + xb8)/exp(model_fw$coefficients[2]))
curva9  = pnorm((- log(tempo) + xb9)/exp(model_fw$coefficients[2]))
curva10 = pnorm((- log(tempo) + xb10)/exp(model_fw$coefficients[2]))
curva11 = pnorm((- log(tempo) + xb11)/exp(model_fw$coefficients[2]))
curva12 = pnorm((- log(tempo) + xb12)/exp(model_fw$coefficients[2]))

curvas_df = data.frame(tempo, curva1, curva2, curva3, curva4, curva5,
                       curva6, curva7, curva8, curva9, curva10, curva11, curva12)


plot1 = ggplot(curvas_df, aes(tempo)) +
        geom_line(aes(y = curva1, color = '1 a 30 anos'),  size = 1.2) +
        geom_line(aes(y = curva2, color = '31 a 50 anos'), size = 1.2) +
        geom_line(aes(y = curva3, color = '51 a 70 anos'), size = 1.2) +
        geom_line(aes(y = curva4, color = '70+ anos '),    size = 1.2) +
          ylab('S^(t): Modelo Log-normal') + xlab('Tempo (dias)')            +
          ggtitle('IDH Baixo do hospital: Idades Segmentadas')               +
          labs(color = 'Idade')                                              +
          scale_colour_manual(values = c('green', 'yellow', 'orange', 'red')) +
          theme(legend.position = c(0.8, 0.8))

plot2 = ggplot(curvas_df, aes(tempo)) +
        geom_line(aes(y = curva5, color = '1 a 30 anos'),  size = 1.2) +
        geom_line(aes(y = curva6, color = '31 a 50 anos'), size = 1.2) +
        geom_line(aes(y = curva7, color = '51 a 70 anos'), size = 1.2) +
        geom_line(aes(y = curva8, color = '70+ anos '),    size = 1.2) +
          ylab('S^(t): Modelo Log-normal') + xlab('Tempo (dias)')            +
```

```
        ggtitle('IDH Médio do hospital: Idades Segmentadas')              +
        labs(color = 'Idade')                                             +
        scale_colour_manual(values = c('green', 'yellow', 'orange', 'red')) +
        theme(legend.position = c(0.8, 0.8))

plot3 = ggplot(curvas_df, aes(tempo)) +
      geom_line(aes(y = curva9, color = '1 a 30 anos'),  size = 1.2) +
      geom_line(aes(y = curva10, color = '31 a 50 anos'), size = 1.2) +
      geom_line(aes(y = curva11, color = '51 a 70 anos'), size = 1.2) +
      geom_line(aes(y = curva12, color = '70+ anos '),    size = 1.2) +
        ylab('S^(t): Modelo Log-normal') + xlab('Tempo (dias)')           +
        ggtitle('IDH Alto do hospital: Idades Segmentadas')               +
        labs(color = 'Idade')                                             +
        scale_colour_manual(values = c('green', 'yellow', 'orange', 'red')) +
        theme(legend.position = c(0.8, 0.8))

grid.arrange(plot1, plot2, plot3, ncol = 3)

## COMPARANDO CURVAS DE SOBREVIVÊNCIA DE DIFERENTES IDADES COM
REGIÃO NEOPLASIA SEGMENTADA ##
## MARCA_UTI FIXADO = 76 E IDH HOSPITAL "MÉDIO" (= 1) ##

# Idade Jovem, Regiao 1

xb_1 = model_fw$coefficients[1]    +
     model_fw$coefficients[3] * 0 +
     model_fw$coefficients[4] * 0 +
     model_fw$coefficients[5] * 1 +
     model_fw$coefficients[6] * 1 +
     model_fw$coefficients[7] * 0 * 0

# Idade Jovem, Regiao 2

xb_2 = model_fw$coefficients[1]    +
     model_fw$coefficients[3] * 0 +
     model_fw$coefficients[4] * 0 +
     model_fw$coefficients[5] * 1 +
     model_fw$coefficients[6] * 2 +
     model_fw$coefficients[7] * 0 * 0

# Idade Jovem, Regiao 3

xb_3 = model_fw$coefficients[1]    +
     model_fw$coefficients[3] * 0 +
     model_fw$coefficients[4] * 0 +
     model_fw$coefficients[5] * 1 +
     model_fw$coefficients[6] * 3 +
     model_fw$coefficients[7] * 0 * 0

# Idade Jovem, Regiao 4
```

```
xb_4 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 0 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 4 +
      model_fw$coefficients[7] * 0 * 0


# Idade Adulto, Regiao 1

xb_5 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 1 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 1 +
      model_fw$coefficients[7] * 0 * 1

# Idade Adulto, Regiao 2

xb_6 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 1 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 2 +
      model_fw$coefficients[7] * 0 * 1

# Idade Adulto, Regiao 3

xb_7 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 1 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 3 +
      model_fw$coefficients[7] * 0 * 1

# Idade Adulto, Regiao 4

xb_8 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 1 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 4 +
      model_fw$coefficients[7] * 0 * 1


# Idade média, Regiao 1

xb_9 = model_fw$coefficients[1]    +
```

```
      model_fw$coefficients[3] * 2 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 1 +
      model_fw$coefficients[7] * 0 * 2

# Idade média, Regiao 2

xb_10 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 2 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 2 +
      model_fw$coefficients[7] * 0 * 2

# Idade média, Regiao 3

xb_11 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 2 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 3 +
      model_fw$coefficients[7] * 0 * 2

# Idade média, Regiao 4

xb_12 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 2 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 4 +
      model_fw$coefficients[7] * 0 * 2


# Idade idosa, Regiao 1

xb_13 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 3 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 1 +
  model_fw$coefficients[7] * 0 * 3

# Idade idosa, Regiao 2

xb_14 = model_fw$coefficients[1]    +
      model_fw$coefficients[3] * 3 +
      model_fw$coefficients[4] * 0 +
      model_fw$coefficients[5] * 1 +
      model_fw$coefficients[6] * 2 +
```

```
  model_fw$coefficients[7] * 0 * 3

# Idade idosa, Regiao 3

xb_15 = model_fw$coefficients[1]     +
       model_fw$coefficients[3] * 3 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 3 +
  model_fw$coefficients[7] * 0 * 3

# Idade idosa, Regiao 4

xb_16 = model_fw$coefficients[1]     +
       model_fw$coefficients[3] * 3 +
       model_fw$coefficients[4] * 0 +
       model_fw$coefficients[5] * 1 +
       model_fw$coefficients[6] * 4 +
  model_fw$coefficients[7] * 0 * 3



curva_1  = pnorm((- log(tempo) + xb_1)/exp(model_fw$coefficients[2]))
curva_2  = pnorm((- log(tempo) + xb_2)/exp(model_fw$coefficients[2]))
curva_3  = pnorm((- log(tempo) + xb_3)/exp(model_fw$coefficients[2]))
curva_4  = pnorm((- log(tempo) + xb_4)/exp(model_fw$coefficients[2]))
curva_5  = pnorm((- log(tempo) + xb_5)/exp(model_fw$coefficients[2]))
curva_6  = pnorm((- log(tempo) + xb_6)/exp(model_fw$coefficients[2]))
curva_7  = pnorm((- log(tempo) + xb_7)/exp(model_fw$coefficients[2]))
curva_8  = pnorm((- log(tempo) + xb_8)/exp(model_fw$coefficients[2]))
curva_9  = pnorm((- log(tempo) + xb_9)/exp(model_fw$coefficients[2]))
curva_10 = pnorm((- log(tempo) + xb_10)/exp(model_fw$coefficients[2]))
curva_11 = pnorm((- log(tempo) + xb_11)/exp(model_fw$coefficients[2]))
curva_12 = pnorm((- log(tempo) + xb_12)/exp(model_fw$coefficients[2]))
curva_13 = pnorm((- log(tempo) + xb_13)/exp(model_fw$coefficients[2]))
curva_14 = pnorm((- log(tempo) + xb_14)/exp(model_fw$coefficients[2]))
curva_15 = pnorm((- log(tempo) + xb_15)/exp(model_fw$coefficients[2]))
curva_16 = pnorm((- log(tempo) + xb_16)/exp(model_fw$coefficients[2]))


curvas_df2 = data.frame(tempo, curva_1, curva_2, curva_3, curva_4, curva_5,
                        curva_6, curva_7, curva_8, curva_9, curva_10, curva_11,
                        curva_12, curva_13, curva_14, curva_15, curva_16)


plot_1 = ggplot(curvas_df2, aes(tempo)) +
         geom_line(aes(y = curva_1, color = 'Lesão Invasiva'), size = 1.2) +
         geom_line(aes(y = curva_2, color = 'Mamilo'),         size = 1.2) +
         geom_line(aes(y = curva_3, color = 'Região Externa'), size = 1.2) +
         geom_line(aes(y = curva_4, color = 'Região Interna'), size = 1.2) +
```

```
        ylab('S(t): Modelo Log-normal') + xlab('Tempo na UTI (dias)')                +
        ggtitle('Idade menor que 30 anos')                +
        labs(color = 'Região Neoplasia')                                +
        scale_colour_manual(values = c('red', 'orange', 'yellow', 'green')) +
        theme(legend.position = c(0.8, 0.8),
            axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
            axis.title.y = element_text(face = 'bold', colour = 'black', size = 10))

plot_2 = ggplot(curvas_df2, aes(tempo)) +
        geom_line(aes(y = curva_5, color = 'Lesão Invasiva'),  size = 1.2) +
        geom_line(aes(y = curva_6, color = 'Mamilo'),          size = 1.2) +
        geom_line(aes(y = curva_7, color = 'Região Externa'),  size = 1.2) +
        geom_line(aes(y = curva_8, color = 'Região Interna'), size = 1.2) +
        ylab('S(t): Modelo Log-normal') + xlab('Tempo na UTI (dias)')                +
        ggtitle('Idade entre 31 e 50 anos')                +
        labs(color = 'Região Neoplasia')                                +
        scale_colour_manual(values = c('red', 'orange', 'yellow', 'green')) +
        theme(legend.position = c(0.8, 0.8),
            axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
            axis.title.y = element_text(face = 'bold', colour = 'black', size = 10))

plot_3 = ggplot(curvas_df2, aes(tempo)) +
        geom_line(aes(y = curva_9, color = 'Lesão Invasiva'), size = 1.2) +
        geom_line(aes(y = curva_10, color = 'Mamilo'),         size = 1.2) +
        geom_line(aes(y = curva_11, color = 'Região Externa'), size = 1.2) +
        geom_line(aes(y = curva_12, color = 'Região Interna'), size = 1.2) +
        ylab('S(t): Modelo Log-normal') + xlab('Tempo na UTI (dias)')                +
        ggtitle('Idade entre 51 a 70 anos')                +
        labs(color = 'Região Neoplasia')                                +
        scale_colour_manual(values = c('red', 'orange', 'yellow', 'green')) +
        theme(legend.position = c(0.8, 0.8),
            axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
            axis.title.y = element_text(face = 'bold', colour = 'black', size = 10))

plot_4 = ggplot(curvas_df2, aes(tempo)) +
        geom_line(aes(y = curva_13, color = 'Lesão Invasiva'), size = 1.2) +
        geom_line(aes(y = curva_14, color = 'Mamilo'),         size = 1.2) +
        geom_line(aes(y = curva_15, color = 'Região Externa'), size = 1.2) +
        geom_line(aes(y = curva_16, color = 'Região Interna'), size = 1.2) +
        ylab('S(t): Modelo Log-normal') + xlab('Tempo na UTI (dias)')                +
        ggtitle('Idade maior que 70 anos')                +
        labs(color = 'Região Neoplasia')                                +
        scale_colour_manual(values = c('red', 'orange', 'yellow', 'green')) +
        theme(legend.position = c(0.8, 0.8),
            axis.title.x = element_text(face = 'bold', colour = 'black', size = 10),
            axis.title.y = element_text(face = 'bold', colour = 'black', size = 10))

grid.arrange(plot_1, plot_2, plot_3, plot_4, nrow = 2, ncol = 2)

# Modelo de Cox #
```

```
# Modificando as covariáveis de forma que atenda o pressuposto
# Com Riscos Proporcionais

dados_cox$IDADE_AUX2 = ifelse(dados_cox$IDADE > 80, 1, 0)

dados_cox$ESPEC = ifelse(dados_cox$ESPEC == 1, 1, 0)

dados_cox$CANC2_AUX = ifelse(dados_cox$CANC2_AUX == 1, 1, 0)

#dados_cox$MARCA_UTI = ifelse(dados_cox$MARCA_UTI <= 75, 0, 1)

dados_cox$IDH_HOSP_AUX = ifelse(dados_cox$IDH_HOSP_AUX <= 1, 1, 0)

dados_cox$RACA_AUX = ifelse(dados_cox$RACA_AUX == 1, 1, 0)

c0 = coxph(Surv(DIAS_PERM, MORTE) ~ 1,           data = dados_cox, x = T, method =
'breslow')
c7 = coxph(Surv(DIAS_PERM, MORTE) ~ COMPLEX,      data = dados_cox, x = T, method
= 'breslow')
c6 = coxph(Surv(DIAS_PERM, MORTE) ~ CAR_INT,      data = dados_cox, x = T, method
= 'breslow')
c4 = coxph(Surv(DIAS_PERM, MORTE) ~ MARCA_UTI,    data = dados_cox, x = T,
method = 'breslow')
c1 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2,   data = dados_cox, x = T,
method = 'breslow')
c3 = coxph(Surv(DIAS_PERM, MORTE) ~ CANC2_AUX,    data = dados_cox, x = T,
method = 'breslow')
c5 = coxph(Surv(DIAS_PERM, MORTE) ~ IDH_HOSP_AUX, data = dados_cox, x = T,
method = 'breslow')
c8 = coxph(Surv(DIAS_PERM, MORTE) ~ RACA_AUX,     data = dados_cox, x = T,
method = 'breslow')
c9 = coxph(Surv(DIAS_PERM, MORTE) ~ PROC_REA_AUX, data = dados_cox, x = T,
method = 'breslow')
c2 = coxph(Surv(DIAS_PERM, MORTE) ~ ESPEC,        data = dados_cox, x = T, method =
'breslow')

# TRV PASSO 1

lrtest(c0, c1) # p < 0.0001
lrtest(c0, c2) # p < 0.0001
lrtest(c0, c3) # p < 0.0001
lrtest(c0, c4) # p = 0.2008
lrtest(c0, c5) # p < 0.0001
lrtest(c0, c6) # p < 0.0001
lrtest(c0, c7) # p < 0.0001
lrtest(c0, c8) # p = 0.006
lrtest(c0, c9) # p = 0.6949

# PASSO 2
```

```
c2_6 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + COMPLEX,      data =
dados_cox, x = T, method = 'breslow')
c2_5 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CAR_INT,      data =
dados_cox, x = T, method = 'breslow')
c2_3 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + MARCA_UTI,    data =
dados_cox, x = T, method = 'breslow')
c2_2 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX,    data =
dados_cox, x = T, method = 'breslow')
c2_4 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + IDH_HOSP_AUX, data =
dados_cox, x = T, method = 'breslow')
c2_7 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + RACA_AUX,     data =
dados_cox, x = T, method = 'breslow')
c2_8 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + PROC_REA_AUX, data =
dados_cox, x = T, method = 'breslow')
c2_1 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + ESPEC,        data =
dados_cox, x = T, method = 'breslow')


# TRV PASSO 2

lrtest(c1, c2_1) # p < 0.0001
lrtest(c1, c2_2) # p < 0.0001
lrtest(c1, c2_3) # p = 0.3269
lrtest(c1, c2_4) # p < 0.0001
lrtest(c1, c2_5) # p < 0.0001
lrtest(c1, c2_6) # p < 0.0001
lrtest(c1, c2_7) # p < 0.01
lrtest(c1, c2_8) # p = 0.644

# PASSO 3

c3_7 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
COMPLEX,      data = dados_cox, x = T, method = 'breslow')
c3_4 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX + CAR_INT,
data = dados_cox, x = T, method = 'breslow')
c3_2 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
MARCA_UTI,    data = dados_cox, x = T, method = 'breslow')
c3_3 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
IDH_HOSP_AUX, data = dados_cox, x = T, method = 'breslow')
c3_5 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
RACA_AUX,     data = dados_cox, x = T, method = 'breslow')
c3_6 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
PROC_REA_AUX, data = dados_cox, x = T, method = 'breslow')
c3_1 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX + ESPEC,
data = dados_cox, x = T, method = 'breslow')


# TRV PASSO 3

lrtest(c2_2, c3_7) # p < 0.0001
```

```
lrtest(c2_2, c3_1) # p < 0.0001
lrtest(c2_2, c3_2) # p = 0.2571
lrtest(c2_2, c3_3) # p < 0.0001
lrtest(c2_2, c3_4) # p < 0.0001
lrtest(c2_2, c3_5) # p = 0.006
lrtest(c2_2, c3_6) # p = 0.7925

# PASSO 4 ITERAÇÕES

c4_1 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
IDH_HOSP_AUX + IDADE_AUX2 * CANC2_AUX,    data = dados_cox, x = T, method =
'breslow')
c4_2 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
IDH_HOSP_AUX + IDADE_AUX2 * IDH_HOSP_AUX, data = dados_cox, x = T, method
= 'breslow')
c4_3 = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE_AUX2 + CANC2_AUX +
IDH_HOSP_AUX + CANC2_AUX * IDH_HOSP_AUX,  data = dados_cox, x = T, method =
'breslow')

# TRV PASSO 4

lrtest(c3_3, c4_1) # p = 0.8539
lrtest(c3_3, c4_2) # p = 0.08451
lrtest(c3_3, c4_3) # p = 0.5222


# MODELO FINAL COX

names(dados_cox)[names(dados_cox) == 'IDADE']        = 'IDADEE'
names(dados_cox)[names(dados_cox) == 'CANC2_AUX']    = 'CANC'
names(dados_cox)[names(dados_cox) == 'IDH_HOSP_AUX'] = 'IDH_HOSP'
names(dados_cox)[names(dados_cox) == 'IDADE_AUX2']   = 'IDADE'


cox_model = coxph(Surv(DIAS_PERM, MORTE) ~ IDADE + IDH_HOSP + CANC, data =
dados_cox, x = T, method = 'breslow')
cox.zph(cox_model)

ggcoxzph(cox.zph(cox_model))


# PLOT GRÁFICO

fit1 = coxph(Surv(DIAS_PERM[IDADE_AUX2 == 0], MORTE[IDADE_AUX2 == 0]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
fit2 = coxph(Surv(DIAS_PERM[IDADE_AUX2 == 1], MORTE[IDADE_AUX2 == 1]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
fit3 = coxph(Surv(DIAS_PERM[IDH_HOSP_AUX == 0], MORTE[IDH_HOSP_AUX ==
0]) ~ 1, data = dados_cox, x = T, method = 'breslow')
```

```
fit4 = coxph(Surv(DIAS_PERM[IDH_HOSP_AUX == 1], MORTE[IDH_HOSP_AUX == 1]) ~ 1, data = dados_cox, x = T, method = 'breslow')
fit5 = coxph(Surv(DIAS_PERM[CANC2_AUX == 0], MORTE[CANC2_AUX == 0]) ~ 1, data = dados_cox, x = T, method = 'breslow')
fit6 = coxph(Surv(DIAS_PERM[CANC2_AUX == 1], MORTE[CANC2_AUX == 1]) ~ 1, data = dados_cox, x = T, method = 'breslow')

ss_1 = survfit(fit1)
s0_1 = round(ss_1$surv, digits = 5)
H0_1 = -log(s0_1)

ss_2 = survfit(fit2)
s0_2 = round(ss_2$surv, digits = 5)
H0_2 = -log(s0_2)

ss_3 = survfit(fit3)
s0_3 = round(ss_3$surv, digits = 5)
H0_3 = -log(s0_3)

ss_4 = survfit(fit4)
s0_4 = round(ss_4$surv, digits = 5)
H0_4 = -log(s0_4)

ss_5 = survfit(fit5)
s0_5 = round(ss_5$surv, digits = 5)
H0_5 = -log(s0_5)

ss_6 = survfit(fit6)
s0_6 = round(ss_6$surv, digits = 5)
H0_6 = -log(s0_6)

df_coxph1 = data.frame(log(ss_1$time), log(H0_1))
colnames(df_coxph1) = c('x','y')
df_coxph2 = data.frame(log(ss_2$time), log(H0_2))
colnames(df_coxph2) = c('x','y')
df_coxph3 = data.frame(log(ss_3$time), log(H0_3))
colnames(df_coxph3) = c('x','y')
df_coxph4 = data.frame(log(ss_4$time), log(H0_4))
colnames(df_coxph4) = c('x','y')
df_coxph5 = data.frame(log(ss_5$time), log(H0_5))
colnames(df_coxph5) = c('x','y')
df_coxph6 = data.frame(log(ss_6$time), log(H0_6))
colnames(df_coxph6) = c('x','y')

## PLOT ##

grid.arrange(

ggplot() +
  geom_step(data = df_coxph1, aes(x = x, y = y, color = '0: 1 a 80 Anos'), size = 1.2) +
```

```
  geom_step(data = df_coxph2, aes(x = x, y = y, color = '1: 81+ Anos'), size = 1.2) +
scale_y_continuous(limits = c(-5, 1.5)) +
  labs(x = '\n log(Dias)', y = expression(log(Lambda[0]*(t))), title = 'Gráfico de Riscos
Proporcionais\n Idade', color = 'IDADE') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.7,0.3)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_coxph3, aes(x = x, y = y, color = '0: 0,81+'), size = 1.2) +
  geom_step(data = df_coxph4, aes(x = x, y = y, color = '1: 0 a 0,8'), size = 1.2) +
scale_y_continuous(limits = c(-5, 1.5)) +
  labs(x = '\n log(Dias)', y = expression(log(Lambda[0]*(t))), title = 'Gráfico de Riscos
Proporcionais\n IDHM Cidade do Hospital', color = 'IDH HOSP') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.7,0.3)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_coxph5, aes(x = x, y = y, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_coxph6, aes(x = x, y = y, color = '1: Lesão Invasiva'), size = 1.2) +
scale_y_continuous(limits = c(-5, 1.5)) +
  labs(x = '\n log(Dias)', y = expression(log(Lambda[0]*(t))), title = 'Gráfico de Riscos
Proporcionais\n Região Neoplasia', color = 'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.7,0.3)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),

  ncol = 3, nrow = 1 )

## Resíduos Martingale e Resíduos Deviance

par(mfrow = c(1, 2))

rd = resid(cox_model, type = 'deviance')
rm = resid(cox_model, type = 'martingale')

pl = cox_model$linear.predictors

df_res = data.frame(rd, rm, pl)

# PLOT RESÍDUOS
```

```
grid.arrange (
ggplot(data = df_res, aes(x = pl, y = rm), size = 1.2,) +
  geom_point(colour = 'dodgerblue3') + scale_y_continuous(limits = c(-4, 2)) +
  labs(x = '\n Preditor Linear', y = 'Resíduo Martingal', title = 'Gráfico de Resíduos Martingal')
+
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14)),

ggplot(data = df_res, aes(x = pl, y = rd), size = 1.2,) +
  geom_point(colour = 'firebrick') + scale_y_continuous(limits = c(-4, 2)) +
  labs(x = '\n Preditor Linear', y = 'Resíduo Deviance', title = 'Gráfico de Resíduos Deviance')
+
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14)),

  ncol = 2, nrow = 1 )


#### GRÁFICO DE SOBREVIVÊNCIA E DE RISCO ( COX )

Ht = basehaz(cox_model, centered = F)
tempos = Ht$time

H0 = Ht$hazard
S0 = exp(-H0)

round(cbind(tempos, S0, H0), digits = 5)
tt = sort(tempos)

aux1 = as.matrix(tt)
n = nrow(aux1)

aux2 = as.matrix(cbind(tempos, S0))

S00 = rep(max(aux2[ , 2]), n)

  for ( i in 1:n) {

    if(tt[i] > min(aux2[ , 1])) {

      i1 = aux2[, 1] <= tt[i]
      S00[i] = min(aux2[i1, 2]) }
    }

ts0 = cbind(tt, S00)

b = cox_model$coefficients
```

```
## COMBINAÇÕES DE TODAS AS CURVAS (8) ##


#dados_cox$IDADE_AUX2 = ifelse(dados_cox$IDADE > 80, 1, 0) #CERTO

#dados_cox$CANC2_AUX = ifelse(dados_cox$CANC2_AUX == 1, 1, 0) #CERTO

#dados_cox$IDH_HOSP_AUX = ifelse(dados_cox$IDH_HOSP_AUX <= 1, 1, 0) # CERTO

# IDADE MENOR QUE 80, IDH < 0,8, CANC: OUTRAS
st1 = S00 ^ (exp(b[2] * 1))
# IDADE MENOR QUE 80, IDH < 0,8, CANC: LESAO INVASIVA
st2 = S00 ^ (exp(b[2] * 1 + b[3] * 1))
# IDADE MAIOR QUE 80, IDH < 0,8, CANC: OUTRAS
st3 = S00 ^ (exp(b[1] * 1 + b[2] * 1 ))
# IDADE MAIOR QUE 80, IDH < 0,8, CANC: LESAO INVASIVA
st4 = S00 ^ (exp(b[1] * 1 + b[2] * 1 + b[3] * 1))

# IDADE MENOR QUE 80, IDH > 0,8, CANC: OUTRAS
st5 = S00 ^ (exp(0))
# IDADE MENOR QUE 80, IDH > 0,8, CANC: LESAO INVASIVA
st6 = S00 ^ (exp(b[3] * 1))
# IDADE MAIOR QUE 80, IDH > 0,8, CANC: OUTRAS
st7 = S00 ^ (exp(b[1] * 1))
# IDADE MAIOR QUE 80, IDH > 0,8, CANC: LESAO INVASIVA
st8 = S00 ^ (exp(b[1] * 1 + b[3] * 1))

df_stcox = data.frame(tt, st1, st2, st3, st4, st5, st6, st7, st8)

## PLOT

grid.arrange(

ggplot() +
  geom_step(data = df_stcox, aes(x = tt, y = st1, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_stcox, aes(x = tt, y = st2, color = '1: Lesão Invasiva'), size = 1.2) +
  labs(x = '\n Dias na UTI', y = 'S(t|x)\n', title = 'Idade menor que 80 anos', color = 'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.7,0.7)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_stcox, aes(x = tt, y = st3, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_stcox, aes(x = tt, y = st4, color = '1: Lesão Invasiva'), size = 1.2) +
  labs(x = '\n Dias na UTI', y = 'S(t|x)\n', title = 'Idade maior que 80 anos', color = 'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
```

```
      axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
      axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
      legend.position = c(0.7,0.7)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_stcox, aes(x = tt, y = st5, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_stcox, aes(x = tt, y = st6, color = '1: Lesão Invasiva'), size = 1.2) +
  labs(x = '\n Dias na UTI', y = 'S(t|x)\n', title = 'Idade menor que 80 anos', color = 'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
      axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
      axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
      legend.position = c(0.7,0.7)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_stcox, aes(x = tt, y = st7, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_stcox, aes(x = tt, y = st8, color = '1: Lesão Invasiva'), size = 1.2) +
  labs(x = '\n Dias na UTI', y = 'S(t|x)\n', title = 'Idade maior que 80 anos', color = 'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
      axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
      axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
      legend.position = c(0.7,0.7)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),

  nrow = 2, ncol = 2 )


## GRÁFICO DAS HAZARDS

Ht1 = -log(st1)
Ht2 = -log(st2)
Ht3 = -log(st3)
Ht4 = -log(st4)
Ht5 = -log(st5)
Ht6 = -log(st6)
Ht7 = -log(st7)
Ht8 = -log(st8)

df_htcox = data.frame(tt, Ht1, Ht2, Ht3, Ht4, Ht5, Ht6, Ht7, Ht8)


## PLOT

grid.arrange(

ggplot() +
  geom_step(data = df_htcox, aes(x = tt, y = Ht1, color = '0: Outras'), size = 1.2) +
```

```
  geom_step(data = df_htcox, aes(x = tt, y = Ht2, color = '1: Lesão Invasiva'), size = 1.2) +
scale_y_continuous(limits = c(0, 6)) +
  labs(x = '\n Dias na UTI', y = 'Risco Acumulado\n', title = 'Idade menor que 80 anos', color =
'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.3,0.8)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_htcox, aes(x = tt, y = Ht3, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_htcox, aes(x = tt, y = Ht4, color = '1: Lesão Invasiva'), size = 1.2) +
scale_y_continuous(limits = c(0, 6)) +
  labs(x = '\n Dias na UTI', y = 'Risco Acumulado\n', title = 'Idade maior que 80 anos', color =
'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.3,0.8)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_htcox, aes(x = tt, y = Ht5, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_htcox, aes(x = tt, y = Ht6, color = '1: Lesão Invasiva'), size = 1.2) +
scale_y_continuous(limits = c(0, 6)) +
  labs(x = '\n Dias na UTI', y = 'Risco Acumulado\n', title = 'Idade menor que 80 anos', color =
'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.3,0.8)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),


ggplot() +
  geom_step(data = df_htcox, aes(x = tt, y = Ht7, color = '0: Outras'), size = 1.2) +
  geom_step(data = df_htcox, aes(x = tt, y = Ht8, color = '1: Lesão Invasiva'), size = 1.2) +
scale_y_continuous(limits = c(0, 6)) +
  labs(x = '\n Dias na UTI', y = 'Risco Acumulado\n', title = 'Idade maior que 80 anos', color =
'CANC') +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = 'bold', colour = 'black'),
        axis.title.x = element_text(face = 'bold', colour = 'black', size = 12),
        axis.title.y = element_text(face = 'bold', colour = 'black', size = 14),
        legend.position = c(0.3,0.8)) +
  scale_colour_manual(values = c('dodgerblue3', 'firebrick')),

  ncol = 2, nrow = 2 )
```

```
par(mfrow = c(1, 3))

# IDADE

fit1 = coxph(Surv(DIAS_PERM[IDADE_AUX2 == 0], MORTE[IDADE_AUX2 == 0]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
ss_1 = survfit(fit1)
s0_1 = round(ss_1$surv, digits = 5)
H0_1 = -log(s0_1)

plot(log(ss_1$time), log(H0_1), xlim = c(0, 5), xlab = 'Tempos (dias)', ylab =
expression(log(Lambda[0]*(t))),
    bty = 'n', type = 's', col = 'blue', lwd = 2)

fit2 = coxph(Surv(DIAS_PERM[IDADE_AUX2 == 1], MORTE[IDADE_AUX2 == 1]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
ss_2 = survfit(fit2)
s0_2 = round(ss_2$surv, digits = 5)
H0_2 = -log(s0_2)

lines(log(ss_2$time), log(H0_2), type = 's', lty = 1, col = 'red', lwd = 2)

# IDH HOSP

fit1 = coxph(Surv(DIAS_PERM[IDH_HOSP_AUX == 0], MORTE[IDH_HOSP_AUX ==
0]) ~ 1, data = dados_cox, x = T, method = 'breslow')
ss = survfit(fit1)
s0 = round(ss$surv, digits = 5)
H0 = -log(s0)

plot(log(ss$time), log(H0), xlim = c(0, 5), xlab = 'Tempos (dias)', ylab =
expression(log(Lambda[0]*(t))),
    bty = 'n', type = 's', col = 'blue', lwd = 2)

fit2 = coxph(Surv(DIAS_PERM[IDH_HOSP_AUX == 1], MORTE[IDH_HOSP_AUX ==
1]) ~ 1, data = dados_cox, x = T, method = 'breslow')
ss = survfit(fit2)
s0 = round(ss$surv, digits = 5)
H0 = -log(s0)

lines(log(ss$time), log(H0), type = 's', lty = 1, col = 'red', lwd = 2)

# CANC2_AUX
```

```
fit1 = coxph(Surv(DIAS_PERM[CANC2_AUX == 0], MORTE[CANC2_AUX == 0]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
ss = survfit(fit1)
s0 = round(ss$surv, digits = 5)
H0 = -log(s0)

plot(log(ss$time), log(H0), xlim = c(0, 5), xlab = 'Tempos (dias)', ylab =
expression(log(Lambda[0]*(t))),
    bty = 'n', type = 's', col = 'blue', lwd = 2)

fit2 = coxph(Surv(DIAS_PERM[CANC2_AUX == 1], MORTE[CANC2_AUX == 1]) ~ 1,
data = dados_cox, x = T, method = 'breslow')
ss = survfit(fit2)
s0 = round(ss$surv, digits = 5)
H0 = -log(s0)

lines(log(ss$time), log(H0), type = 's', lty = 1, col = 'red', lwd = 2)
```