Universidade de Brasília
Departamento de Estatística

Análise de sobrevivência em câncer de mama com MLGs: Estudo dos dados do GBSG para identificação de fatores de risco

Renan Luiz da Silva Nascimento

Brasília
18/12/2023

Sumário

1 Introdução

A análise de sobrevivência é uma área crucial na ciência médica e estatística, desempenhando um papel fundamental na compreensão dos fatores que influenciam a duração do tempo até um evento de interesse, como a mortalidade ou a ocorrência de um determinado desfecho clínico.

Este trabalho visa explorar e analisar os dados do German Breast Cancer Study Group (GBSG) por meio da aplicação de análise de sobrevivência. O conjunto de dados GBSG é uma valiosa fonte de informações sobre pacientes com câncer de mama, oferecendo insights valiosos sobre variáveis clínicas e demográficas que podem influenciar o tempo de sobrevida desses pacientes.

O cancêr é uma das principais causas de morte e uma barreira importante ao aumento da esperança de vida em todos os países do mundo (SUNG et al., 2021), de acordo com estimativas da Organização Mundial da Saúde (OMS) em 2019, o cancêr é a primeira ou segunda principal causa de morte antes dos 70 anos de idade em 112 paises.

A censura é um elemento-chave nesse contexto, pois nem todos os pacientes experimentarão o evento de interesse (no caso, a morte) durante o período de observação do estudo. Alguns podem ser censurados, ou seja, seus tempos de sobrevida não são totalmente observados, seja porque o estudo terminou antes que o evento ocorresse para eles ou porque foram perdidos no acompanhamento.

Compreender os fatores que afetam a sobrevida das pacientes com câncer de mama é essencial para melhorar as estratégias de tratamento e prognóstico. A análise de sobrevivência, ao considerar a censura nos dados, permite estimar a probabilidade de sobrevivência ao longo do tempo, mesmo quando não se tem informações completas para todos os casos.

Os objetivos deste estudo incluem a construção e avaliação de modelos de regressão de sobrevivência utilizando diferentes variáveis clínicas e demográficas disponíveis no conjunto de dados GBSG. Além disso, pretendemos selecionar o modelo mais adequado com base em critérios estatísticos, como o teste de razão de verossimilhança (TRV), para melhor compreender os fatores que influenciam a sobrevivência nesse contexto clínico.

2 Metodologia

A metodologia para esse trabalho utilizando análise de sobrevivência com dados do GBSG para câncer de mama irá incluir algumas etapas.

2.1 Análise descritiva

Será realizado uma inspeção inicial das variáveis disponíveis, incluindo informações sobre tempo de sobrevida, status do evento, idade das pacientes, tamanho do tumor, grau do tumor, número de linfonodos invadidos, status da cápsula dos linfonodos e grau tumoral.

Irão ser criados gráficos descritivos, também usaremos a estimativa de Kaplan-Meier para visualizar a função de sobrevivência ao longo do tempo para diferentes grupos, gráfico da função de risco acumulado e gráfico do tempo total em teste (TTT), também será observado o TTT para cada cováriavel de grupo, afim de observamos funções de densidade de probabilidade que podem ser usadas para construir modelos de regressão paramétrico.

2.2 Seleção da distribuição

Considerando as funções de densidade de probabilidade possiveis, será feita uma seleção de qual densidade será usada para analisar a variável "Tempo". Será utilizados distribuições que estão disponíveis na função survreg no R e para fazer essa seleção serão consideradas medidas como AIC, AICc e BIC.

2.3 Método de estimação

Dada a função de densidade de probabilidade escolhida do modelo de regressão, é possivel encontrar os estimadores de máxima verossimilhança dos parâmetros. A função de verossimilhança é representada por

$$L(\theta) = \prod_{i=1}^{n} f(t_i \mid x_i)^{\delta_i} \cdot S(t_i \mid x_i)^{1-\delta_i}. \tag{2.3.1}$$

Onde:

- t_i é o tempo de sobrevivência ou falha para o indivíduo i;
- x_i são as covariáveis associadas ao indivíduo i;
- δ_i é o indicador de censura (0 para censurado, 1 para não censurado) do evento de sobrevivência para o indivíduo i;
- $f(t_i \mid x_i)$ é a função de densidade de probabilidade condicional do tempo de sobrevivência para o indivíduo i dado o vetor de covariáveis x_i;
- $S(t_i \mid x_i)$ é a função de sobrevivência condicional do tempo de sobrevivência para o indivíduo i dado o vetor de covariáveis x_i.

2.4 Seleção de variáveis para o modelo

Para a seleção de variáveis no modelo de regressão de sobrevivência, adotaremos uma estratégia de seleção de modelos derivada da proposta de Collett (1994), que segue uma série de passos e utiliza o teste de razão de verossimilhança para avaliar a significância das variáveis.

O teste de razão de verossimilhança segue a seguinte estrutura

$$TRV = 2\left[\log L(\hat{\theta}_{bG}) - \log L(\hat{\theta}_{bM})\right]. \tag{2.4.1}$$

Onde temos que:

- $logL(\hat{\theta}_{bG})$ é o logaritmo da verossimilhança do modelo completo ou mais robusto.
- $logL(\hat{\theta}_{bM})$ é o logaritmo da verossimilhança do modelo mais simples

Onde as hipóteses são:

- **Hipótese Nula (H0):** Não há diferença significativa entre os modelos, ou seja, o modelo mais robusto não é estatisticamente melhor que o modelo mais simples.
- **Hipótese Alternativa (Ha):** Existe uma diferença significativa entre os modelos, indicando que o modelo mais robusto é estatisticamente melhor ao modelo mais simples.

2.5 Análise de resíduos

Após a obtenção do modelo de regressão de sobrevivência, realizaremos uma análise dos resíduos para avaliar a adequação e a robustez do modelo. Serão utilizados gráficos de resíduos, para verificar os pressupostos de um modelo de regressão.

3 Resultados

3.1 Descrição do Banco de Dados

O conjunto de dados que será utilizado é o German Breast Cancer Study Group (GBSG), que é uma fonte fundamental de informações sobre pacientes com câncer de mama, coletadas em estudos clínicos para investigar fatores prognósticos e terapêuticos.

O conjunto de dados inclui uma variedade de variáveis, tais como:

- **id:** Identificador único atribuído a cada paciente no conjunto de dados.
- **age:** Representa a idade do paciente em anos.
- **meno:** Indica o status menopausal do paciente (0 = pré-menopausa, 1 = pós-menopausa).
- **size:** Refere-se ao tamanho do tumor em milímetros.
- **grade:** Indica o grau do tumor.
- **nodes:** Representa o número de linfonodos positivos.
- **pgr:** Refere-se à concentração de receptores de progesterona no sangue.
- **er:** Indica a concentração de receptores de estrogênio no sangue.
- **hormon:** Indica se o paciente recebeu terapia hormonal (0 = não, 1 = sim).
- **rfstime:** Representa o tempo, em dias, até a ocorrência de recorrência, óbito ou o último acompanhamento.
- **status:** Indica o status do paciente no ponto final (0 = vivo sem recorrência, 1 = recorrência ou óbito).

3.2 Análise descritiva

Iniciamos nossa análise visualizando o histograma da variável "tempo", que representa o tempo até o evento de recorrência, óbito ou último acompanhamento. Este histograma nos permite observar a distribuição dos tempos de sobrevivência e identificar padrões ou sugestões quanto à sua distribuição.

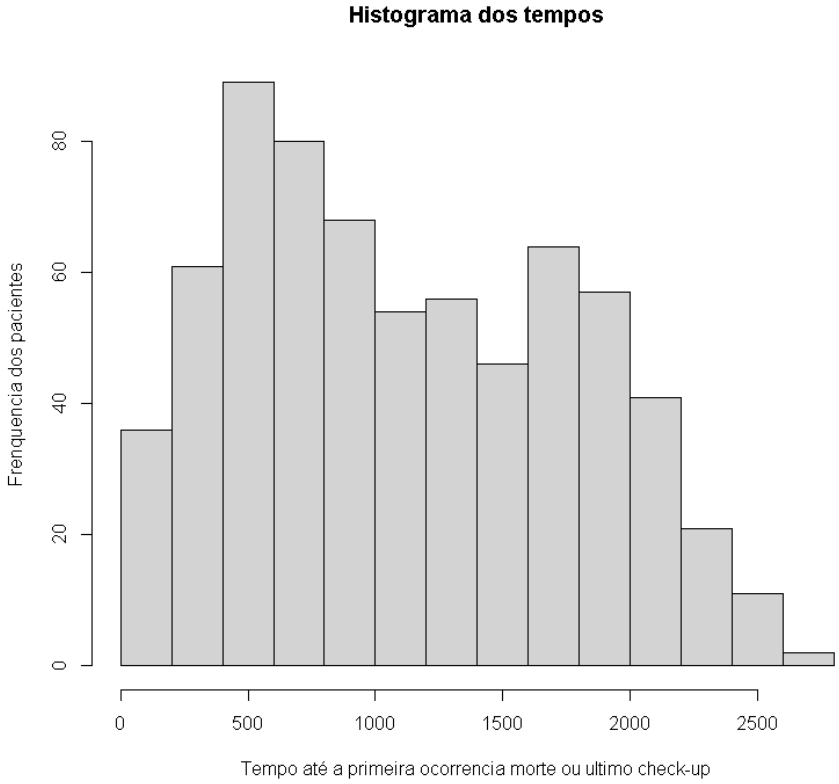

Figura 1: Histograma dos tempos

Parece haver um aumento constante inicialmente nos números de censura, atingindo um pico na faixa de 400 a 600 dias, com mais de 80 casos. Em seguida, observa-se um declínio gradual até os 1600 dias, seguido por um leve aumento até os 1700 dias, seguido novamente por um declínio até o final do estudo.

Partindo para realizar uma estimativa da curva de sobrevida utilizando o método de Kaplan-Meier para o conjunto de dados como um todo, nos dá a possibilidade de gerar um gráfico da função de sobrevivência, exibindo a probabilidade de um paciente sobreviver além de determinado tempo e também podemos representar graficamente a função de risco acumulada, que mostra a taxa acumulada de falhas ao longo do tempo. A fórmula de Kaplan-Meier segue a seguinte estrutura

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right). \tag{3.2.1}$$

Onde:

- d_j é o número de falhas em t_j, onde $j = 1, \ldots, k$.
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

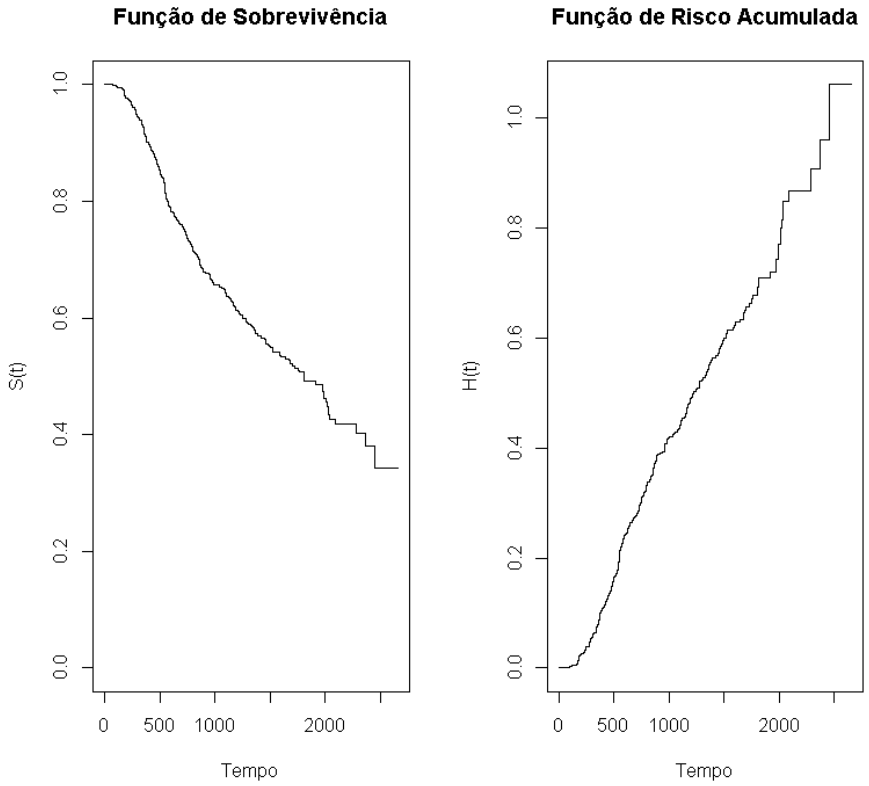

Figura 2: Função de sobrevivência e de risco acumulada ao longo do tempo

Uma função de taxa de falha unimodal ou decrescente poderia se ajustar aos dados, porém é importante analisar os dados mais detalhadamente e realizar testes estatísticos para determinar qual modelo se ajusta melhor aos padrões observados nos números de censura ao longo do tempo, vamos agora avançar para o gráfico do tempo total em teste (TTT) para verificar se encontramos evidências distintas das observadas até o momento.

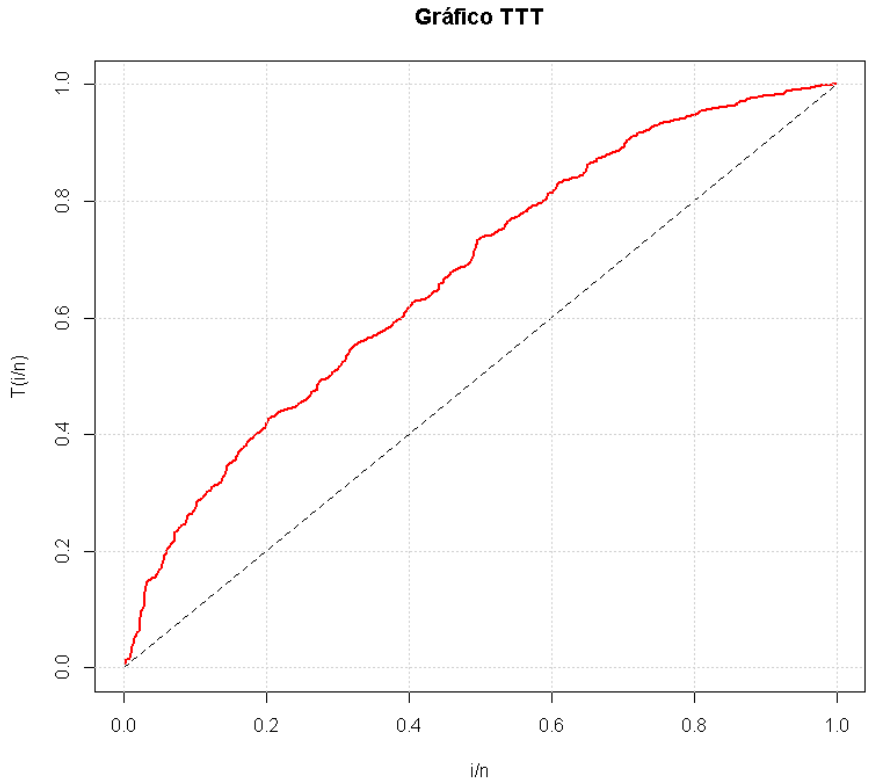

Figura 3: Gráfico TTT

Pelo TTT, parece que uma função de risco crescente pode ser uma escolha adequada. A divergência entre as candidatas a funções pode ser explicada pela presença significativa de censuras nos dados. É importante notar que o gráfico TTT não considera as censuras, o que pode influenciar a interpretação e a escolha do modelo apropriado para descrever o comportamento dos eventos ao longo do tempo.

Após examinarmos o gráfico TTT geral, é fundamental aprofundar nossa análise ao considerar as covariáveis categóricas. Esta etapa visa investigar as relações entre as características específicas dos pacientes e os tempos de ocorrência dos eventos. Esta etapa tem por objetivo selecionar quais variáveis categóricas (covariáveis) prosseguirão na análise. Neste conjunto de dados, há três variáveis categóricas: menopausa, grau do tumor e terapia hormonal.

Primeiramente, iremos realizar uma análise visual examinando a função de sobrevivência de cada grupo de cada variável. Essa abordagem visual permite uma compreensão inicial das diferenças nas curvas de sobrevivência entre os grupos em estudo. Posteriormente,a utilização do teste de Wilcoxon ou do teste logrank se torna pertinente para avaliar diferenças entre as curvas de sobrevivência dentro dos grupos. Esses testes são particularmente relevantes para comparar as distribuições de eventos ao longo do tempo dentro de cada grupo específico, oferecendo uma análise estatística sobre possíveis dispari-

dades nas curvas de sobrevivência. Vamos iniciar essa análise com a variável menopausa, considerando os grupos pré-menopausa (0) e pós-menopausa (1).

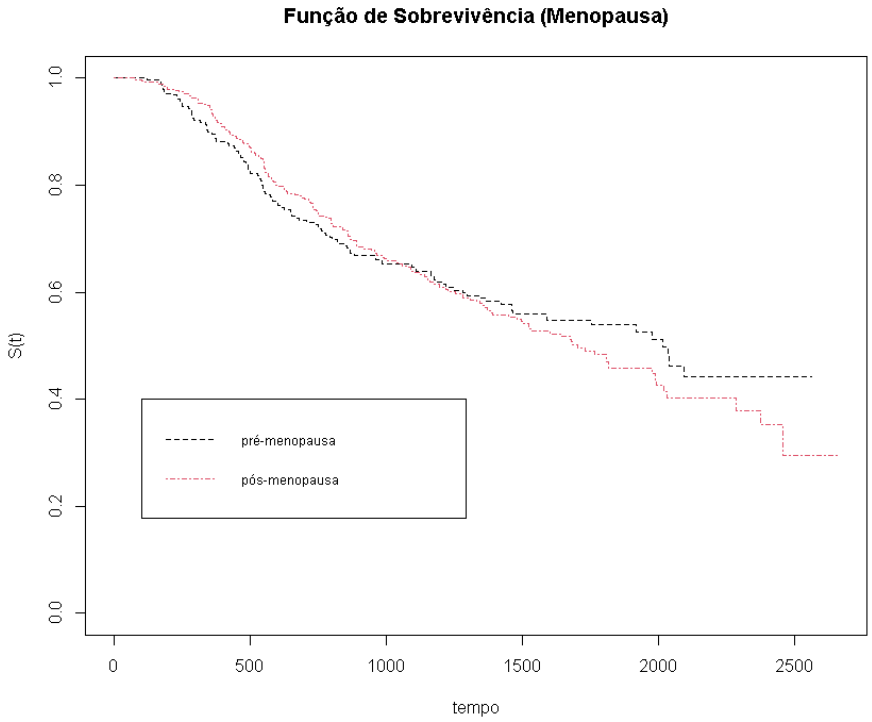

Figura 4: Função de sobrevivência dos grupos menopausa ao longo do tempo

A observação visual das curvas indica uma possível similaridade entre o grupo pré-menopausa e o grupo pós-menopausa.

O próximo passo é comparar os grupos definidos pela variável "grau do tumor". Este procedimento visa examinar as possíveis diferenças nos tempos de sobrevivência entre os diferentes graus de tumor apresentados pelas pacientes. Ao comparar as curvas de sobrevivência entre esses grupos específicos, buscamos identificar se o grau do tumor possui influência nos desfechos de sobrevivência das pacientes com câncer de mama neste conjunto de dados.

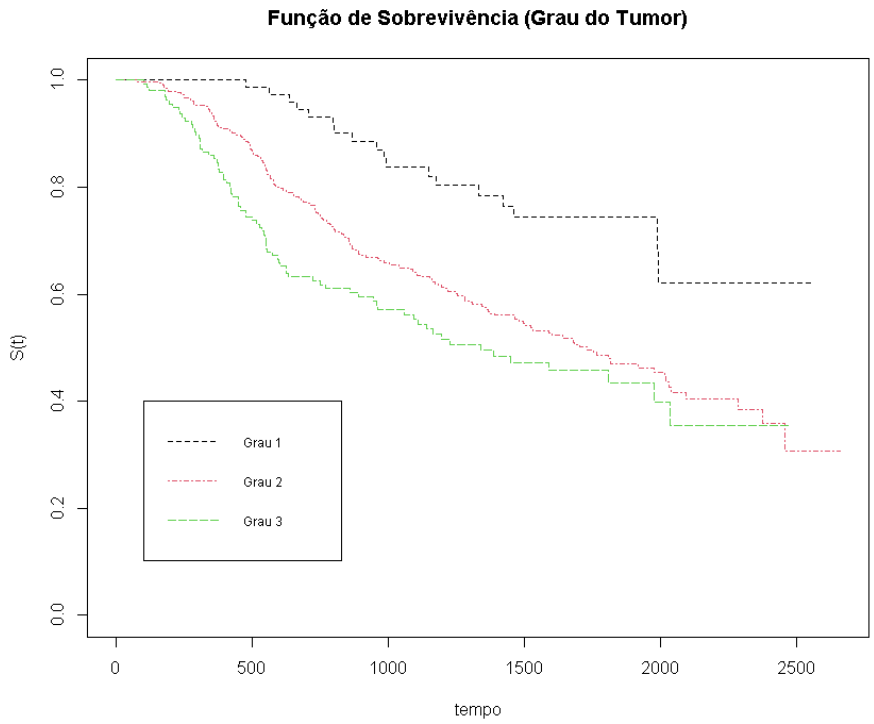

Figura 5: Função de sobrevivência dos grupos grau do tumor ao longo do tempo

Analisando o gráfico dos grupos definidos pelo grau do tumor comprovou uma tendência que já era esperada na maioria dos eventos observados. Foi notada uma maior probabilidade de sobrevivência para pacientes com grau de tumor 1 em relação aos graus 2 e 3, ao longo do período observado. Visualmente, a curva representativa do grau 1 apresenta-se distinta das curvas correspondentes aos graus 2 e 3, sugerindo uma possível diferença estatística entre essas categorias de tumor. Esse resultado preliminar indica que o grau do tumor pode desempenhar um papel significativo nos desfechos de sobrevivência das pacientes com câncer de mama neste conjunto de dados, corroborando a importância prognóstica dessa característica tumoral específica.

Avançando para a investigação da influência da variável "terapia hormonal"nos desfechos de sobrevivência das pacientes com câncer de mama neste conjunto de dados. A visualização das curvas de sobrevivência dos grupos definidos pela terapia hormonal permitirá explorar se a aplicação desse tratamento está associada a diferenças nos tempos de sobrevivência.

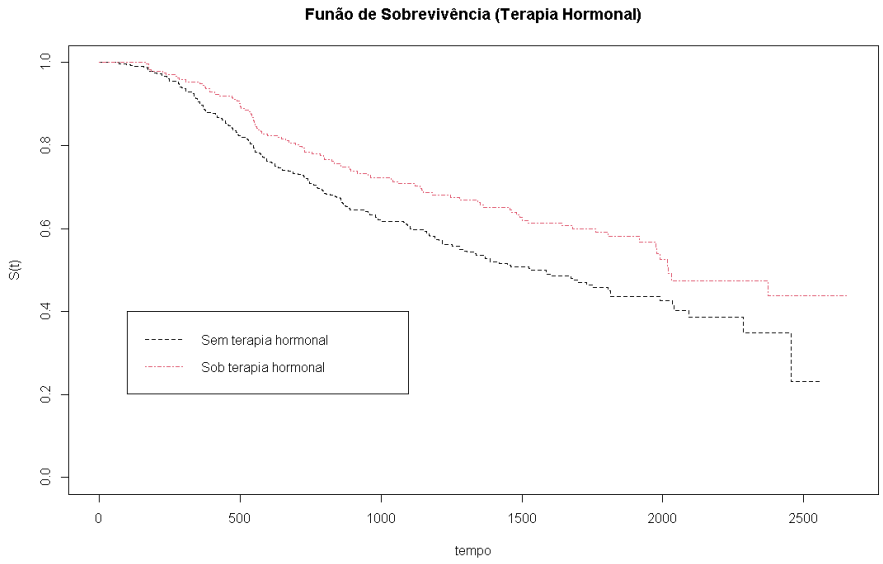

Figura 6: Função de sobrevivência dos grupos terapia hormonal do tempo

As curvas indicam uma pequena divergência, sugerindo que o grupo de pacientes sob terapia hormonal tende a apresentar uma maior probabilidade de sobrevivência ao longo de todo o período de estudo. No entanto, é fundamental realizar uma análise estatística para confirmar se essas diferenças observadas das variáveis menopausa, grau do tumor e terapia hormonal são estatisticamente significativas e não apenas uma variação ao acaso nos dados. Para isso temos o teste de Wilcoxon e o logrank.

O teste Wilcoxon pode ser descrito pela fórmula a seguir

$$S = \frac{\left[\sum_{j=1}^{n} n_j(d_{1j} - w_{1j})\right]^2}{\sum_{j=1}^{n} n_j^2 V_{1j}}. \tag{3.2.2}$$

O teste logrank pode ser descrito pela fórmula a seguir

$$T = \frac{\sum_{j=1}^{k}(d_{2j} - w_{2j})^2}{\sum_{j=1}^{k} V_j^2}. \tag{3.2.3}$$

Onde temos as seguintes hipóteses para ambos os testes:

- **Hipótese Nula (H0):** As funções de sobrevivência dos grupos são iguais ao longo do tempo.
- **Hipótese Alternativa (Ha):** As funções de sobrevivência dos grupos são diferentes ao longo do tempo.

O critério utilizado nesse trabalho foi o de permanecer com as variáveis que

apresentaram valores p inferiores a 0,25 em pelo menos um dos testes de comparação das curvas de sobrevivência (COLOSIMO; GIOLO, 2006).

Tabela 1: Testes (p-valor)

Covariável	**logrank**	**Wilcoxon**
Menopausa	0,28 (0,5966)	0,0004 (0,9828)
Grau do tumor	21,09 (< 0,001)	25,58 (< 0,001)
Terapia hormonal	8,56 (0,0034)	8,71 (0,0032)

Com base nos resultados apresentados na Tabela 1, verifica-se que as variáveis grau do tumor e terapia hormonal atenderam ao critério estabelecido previamente para inclusão na etapa de modelagem. Em contrapartida, a variável menopausa não atendeu aos critérios estabelecidos e será descartada da etapa de modelagem. Portanto, as variáveis grau do tumor e terapia hormonal serão consideradas como candidatas relevantes para a construção do modelo preditivo.

3.3 Seleção da distribuição

Entramos agora em uma fase crucial da modelagem, buscando identificar a distribuição mais adequada para os dados de sobrevivência. Inicialmente, iremos basear nossa investigação nas análises visuais realizadas anteriormente, especialmente no gráfico TTT e na função de risco acumulada. Dada a observação de um possível padrão unimodal ou decrescente nos desfechos de sobrevivência, a distribuição log-logística surge como uma escolha inicial viável, pois essa distribuição pode acomodar ambas as características. Além disso, para cenários onde a suposição é de uma taxa de falha crescente, consideraremos a distribuição Weibull, amplamente utilizada para modelar dados de sobrevivência em situações onde a taxa de falha aumenta ou diminui ao longo do tempo.

Além dessas distribuições, consideraremos a lognormal como uma alternativa, pois é conhecida por sua aplicabilidade em modelar variáveis positivas e assimétricas, o que pode ser relevante para os dados de sobrevivência. A inclusão dessas distribuições na análise permitirá uma avaliação abrangente das suposições subjacentes aos dados de sobrevivência, possibilitando a seleção da distribuição mais apropriada que melhor represente os padrões observados nos desfechos de sobrevivência das pacientes com câncer de mama neste estudo.

A seleção da melhor distribuição para modelar os dados de sobrevivência pode ser uma tarefa desafiadora. Para essa finalidade, será feita uma análise visual do gráfico de sobrevivência das distribuições, e também utilizaremos métricas estatísticas como o Critério de Informação de Akaike (AIC), o AIC corrigido (AICc) e o Critério de Informação

Bayesiano (BIC). Essas métricas são ferramentas valiosas na seleção de modelos, levando em consideração a qualidade do ajuste do modelo e a complexidade do mesmo.

Tabela 2: Valores de AIC, AICc e BIC para diferentes distribuições

Distribuição	**AIC**	**AICc**	**BIC**
Log-logística	5259,894	5259,912	5268,956
Weibull	5278,553	5278,57	5287,614
Log-normal	5241,77	5241,782	5250,832

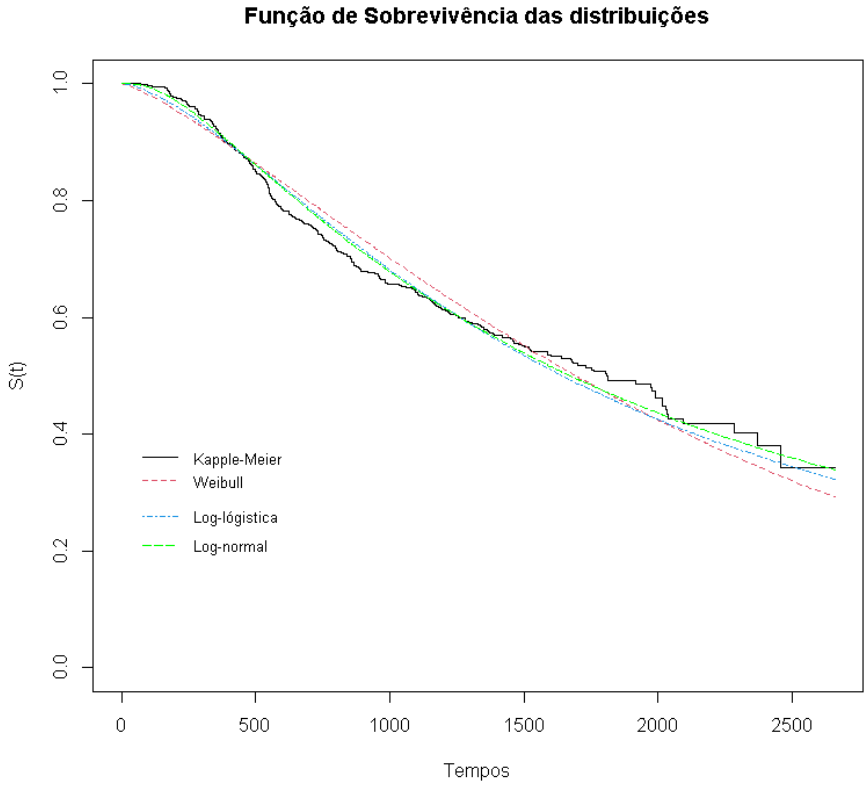

Figura 7: Função de sobrevivência da Weibull, Log-logistica, Log-normal e de Kaplan-Meier, ajustadas aos dados

Com base nas métricas de informação como AIC, AICc e BIC, juntamente com a análise gráfica das distribuições testadas, optamos por modelar os dados de sobrevivência usando a distribuição log-normal.

Para a distribuição log-normal, temos a seguinte função de densidade

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right], \quad t \geq 0. \tag{3.3.1}$$

em que μ é a média do logaritmo do tempo de falha assim como σ é o desvio-padrão.

Temos a seguinte função de sobrevivência

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right). \tag{3.3.2}$$

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

Além disso, empregaremos um modelo de escala logarítmica, onde definimos a variável de interesse Y como o logaritmo do tempo de sobrevivência $Y = \log(T)$. Essa abordagem permite interpretar os coeficientes do modelo sem a necessidade de qualquer reparametrização adicional, facilitando a interpretação dos resultados e fornecendo uma visão direta do impacto das variáveis preditoras sobre a escala original do tempo de sobrevivência.

3.4 Seleção do modelo

No processo de seleção do modelo para a análise dos dados, é fundamental considerar a abordagem que melhor se adequa à complexidade e ao contexto do estudo em questão. Optamos por não utilizar métodos automatizados de seleção de variáveis, como os métodos forward, backward ou stepwise. A escolha fundamenta-se na consideração de 7 covariáveis potencialmente relevantes para descrever os padrões da resposta, o que resulta em um número considerável de possíveis modelos tornando impraticável ajustar todos esses modelos para selecionar o que melhor explique a resposta.

Os métodos automatizados, embora amplamente disponíveis nos pacotes estatísticos, tendem a identificar um conjunto específico de covariáveis, deixando de lado outras combinações igualmente válidas para explicar a resposta. Em vista dessas limitações dos métodos automáticos, optamos por adotar uma abordagem mais participativa e proativa no processo de seleção do modelo. Neste estudo, os passos utilizados no processo de seleção são apresentados a seguir (COLLETT, 1994):

1. **Ajuste inicial com uma única covariável:** São ajustados modelos individuais contendo cada covariável separadamente. São incluídas no modelo aquelas covariáveis que se mostram estatisticamente significativas ao nível de 0,05. Iremos utiliza o teste da razão de verossimilhanças para todos os passos para avaliar a significância das variáveis.

2. **Ajuste simultâneo das covariáveis significativas:** As covariáveis identificadas como significantes na etapa anterior são ajustadas conjuntamente. Em presença de certas covariáveis, outras podem perder significância. Modelos reduzidos são ajustados, excluindo uma covariável por vez. Apenas as covariáveis que alcançam significância permanecem no modelo.

3. **Ajuste do novo modelo com covariáveis retidas:** Um novo modelo é ajustado com as covariáveis retidas no passo 2. Covariáveis excluídas no passo anterior retornam ao modelo para confirmar a ausência de significância estatística.

4. **Inclusão de covariáveis significativas anteriores:** Covariáveis significativas identificadas no passo 3 são incluídas no modelo junto com aquelas do passo 2. Neste estágio, retornam às covariáveis excluídas no passo 1 para verificar sua significância estatística.

5. **Análise de Possível Exclusão de Covariáveis no Modelo:** Um modelo é ajustado incluindo todas as covariáveis significativas identificadas no passo 4. Nesta etapa, é testada a possibilidade de remover alguma covariável do modelo.

6. **Inclusão e Avaliação de Termos de Interação:** Após a seleção das covariáveis no passo 5, o modelo final será ajustado para os efeitos principais. Para a conclusão da modelagem, será explorada a possibilidade de inclusão de termos de interação. Serão testadas todas as combinações possíveis de interações entre pares de covariáveis incluídas no modelo.

Para a construção do modelo, serão testadas as seguintes variáveis como preditoras:

Tabela 3: Indicador e Variável

Indicador	**Variável**
X1	Grau do tumor
X2	Terapia hormonal
X3	Idade
X4	Tamanho do tumor
X5	Quantidade de linfonodos
X6	Níveis de progesterona
X7	Níveis de estrogênio

Tabela 4: Resultados dos testes

Passo	**Modelo**	**Estatística**	**p-valor**
Passo 1	X1	32.78	<0,001
	X2	9,54	0,002
	X3	1,88	0,17
	X4	17,65	<0,001
	X5	58,36	<0,001
	X6	34,73	<0,001
	X7	6,52	0,011
Passo 2	X1+X2+X4+X5+X6+X7	-	-
	X2+X4+X5+X6+X7	12,89	<0,001
	X1+X4+X5+X6+X7	10,57	0,001
	X1+X2+X5+X6+X7	3,96	0,046
	X1+X2+X4+X6+X7	36,84	<0,001
	X1+X2+X4+X5+X7	19,12	<0,001
	X1+X2+X4+X5+X6	0,002	0,97
Passo 3	X1+X2+X4+X5+X6	-	-
	X1+X2+X4+X5+X6+X7	0,002	0,97
Passo 4	X1+X2+X4+X5+X6	-	-
	X1+X2+X4+X5+X6+X3	0,43	0,511
Passo 5	X1+X2+X4+X5+X6	-	-
	X1+X2+X4+X5	21,09	<0,001
	X1+X2+X4+X6	36,86	<0,001
	X1+X2+X5+X6	3,98	0,046
	X1+X4+X5+X6	10,69	0,001
	X2+X4+X5+X6	12,89	<0,001

Partiremos para o próximo passo. será realizado uma abordagem que nos permitirá avaliar minuciosamente a influência de cada interação nas respostas do modelo, identificando quais interações mantêm a significância estatística e quais não contribuem significativamente para a explicação dos dados. Este processo gradual de remoção de interações uma a uma nos ajudará a refinar e simplificar o modelo, focando nas interações mais relevantes e robustas, garantindo uma representação mais concisa e confiável dos fatores que influenciam as variáveis de interesse em nosso estudo.

Nosso modelo completo segue a seguinte estrutura:

$$X1 + X2 + X4 + X5 + X6 + X1 * X2 + X1 * X4 + X1 * X5 + X1 * X6 + X2 * X4 + X2 * X5 + X2 * X6 + X4 * X5 + X4 * X6 + X5 * X6 \quad (3.4.1)$$

Tabela 5: Resultados dos testes

Passo	**Modelo**	**Estatística**	**p-valor**
Passo 6	Modelo completo	-	-
	Modelo sem X1X2	2,80	0,09
	Modelo sem X1X4	1,34	0,25
	Modelo sem X1X5	9,15	0,002
	Modelo sem X1X6	5,35	0,02
	Modelo sem X2X4	0,10	0,75
	Modelo sem X2X5	1,90	0,17
	Modelo sem X2X6	3,46	0,06
	Modelo sem X4X5	5,70	0,02
	Modelo sem X4X6	0,00	0,98
	Modelo sem X5X6	0,17	0,68

Após esta etapa, chegamos a um modelo que inclui uma seleção específica de variáveis e interações que parecem ser significantes para explicar os dados. O modelo atual incorpora os termos X1, X2, X4, X5, X6, além das interações X1X5, X1X6, X4X5 e X5X6. No entanto, iremos verificar se as interações que foram removidas inicialmente do modelo realmente deveriam ser descartadas. Na próxima etapa iremos adicionar as interações que foram retiradas nesse novo modelo para verificar se de fato devem ser desconsideradas. Basicamente, iremos repetir o passo 3 para esse novo modelo.

Tabela 6: Resultados dos testes

Modelo	**Estatística**	**p-valor**
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6	-	-
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X1X2	3,45	0,18
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X1X4	0,72	0,70
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X2X4	0,25	0,62
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X2X5	3,18	0,07
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X2X6	4,26	0,04
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X4X6	0,01	0,92

Identificamos que o fator de interação entre terapia hormonal e progesterona (X2X6) demonstrou uma significância estatística relevante. Esta descoberta ressalta a importância dessa interação na explicação dos dados observados, indicando que a combinação entre terapia hormonal e níveis de progesterona pode ter um impacto substancial nos resultados que estamos estudando. Assim, considerando essa relevância estatística, é recomendável incluir o termo de interação X2X6 no modelo.

A próxima etapa consiste em um processo de refinamento adicional do modelo, no qual removeremos as interações uma a uma para verificar sua contribuição significativa. Esse procedimento nos permitirá avaliar minuciosamente a importância individual de cada interação no contexto do modelo expandido.

Tabela 7: Resultados dos testes

Modelo	**Estatística**	**p-valor**
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6+X2X6	-	-
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X5X6	4,26	0,04
X1+X2+X4+X5+X6+X1X5+X1X6+X4X5+X2X6	-0,12	0,73
X1+X2+X4+X5+X6+X1X5+X1X6+X5X6+X2X6	-5,07	0,02
X1+X2+X4+X5+X6+X1X5+X4X5+X5X6+X2X6	5,87	0,053
X1+X2+X4+X5+X6+X1X6+X4X5+X5X6+X2X6	-3,48	0,18

Identificamos indicações para a remoção das interações entre grau do tumor e quantidade de linfonodos (X1X5) e progesterona e quantidade de linfonodos (X5X6) do modelo. No entanto, ao retirarmos a interação entre grau do tumor e progesterona (X1X6), observamos que o p-valor obtido foi de 0,053, o que está bastante próximo do nível de significância estabelecido para este trabalho (0,05). Diante dessa proximidade e considerando a potencial relevância dessa interação, decidimos manter essa interação no modelo

final.

No desdobramento dos resultados, chegamos a um conjunto de três interações que permaneciam no modelo: grau de tumor e progesterona (X1X6), tamanho do tumor e quantidade de linfonodos (X4X5), e terapia hormonal e progesterona. No entanto, ao analisar mais profundamente a interação entre tamanho do tumor e quantidade de linfonodos (X4X5), observamos um padrão inesperado: o coeficiente dessa interação é positivo. Isso implicaria que, conforme a interação aumenta, a chance de sobrevivência também aumenta. Essa conclusão parece contraditória à lógica clínica, pois quando observamos os coeficientes individuais de cada variável, tanto o tamanho do tumor quanto a quantidade de linfonodos apresentam coeficientes negativos, indicando uma associação inversa com a chance de sobrevivência. Diante dessa inconsistência e da contradição com o conhecimento clínico estabelecido, optamos por remover essa interação do nosso modelo final, buscando garantir que as relações presentes na análise reflitam de forma coerente e sensata os padrões observados nos dados.

Os coeficientes estimados estão expressos na escala logarítmica do tempo, isto é, para $Y = \log(T) = X\beta + \sigma\nu$. Além de que o modelo esta ajustado a niveis de grau de tumor 1 e não estar sob terapia hormonal.

Tabela 8: Coeficientes do Modelo

Covariável	**Estimativa**	**p-valor**
Intercepto	8.336771	<0.001
Grau de tumor 2	-0.797588	0.00037
Grau de tumor 3	-0.873088	0.00021
Terapia hormonal 1	0.185998	0.09646
Tamanho do tumor	-0.006246	0.04642
Qtd de linfonodos	-0.048337	<0.001
Progesterona	-0.000319	0.69428
Terapia hormonal 1*Progesterona	0.001614	0.05215
Grau de tumor 2*Progesterona	0.001928	0.03800
Grau de tumor 3*Progesterona	0.000351	0.77044

3.5 Análise de resíduos

Agora serão examinados os resíduos do modelo para avaliar se eles aderem a uma distribuição exponencial padrão. Será investigado essa distribuição ao invés de uma distribuição normal padrão por conta da natureza dos dados de sobrevivência e da estrutura do modelo usado, que frequentemente estão associados a uma distribuição exponencial dos resíduos.

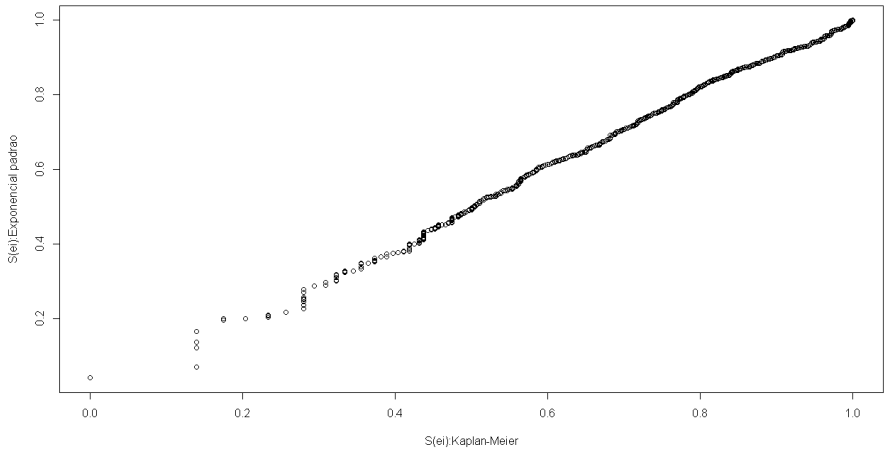

Figura 8: Gráfico de Resíduos vs Distribuição Exponencial Padrão

Olhando a Figura 8, sugere que os resíduos Cox-Snell seguem uma distribuição exponencial, o que é coerente com as suposições do modelo. Esse padrão reforça a consistência das relações identificadas entre as variáveis explicativas e a variável de resposta.

Na avaliação da adequação do modelo, recorremos aos resíduos de Cox-Snell como uma ferramenta essencial. A expectativa é que esses resíduos sigam a tendência da curva estimada de Kaplan-Meier dos erros. Em outras palavras, buscamos verificar se os resíduos padronizados apresentam uma distribuição que se assemelha à curva esperada, representada pela estimativa de Kaplan-Meier dos erros.

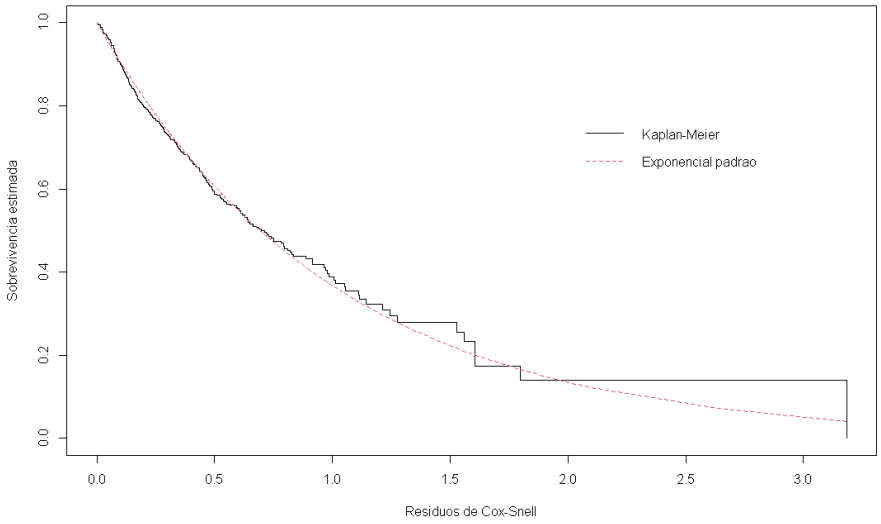

Figura 9: Adequação do Modelo: Resíduos de Cox-Snell vs Curva de Kaplan-Meier dos Erros

Observamos que o modelo parece se ajustar bem aos dados observados. Os resíduos de Cox-Snell exibem um comportamento que acompanha de perto a curva de Kaplan-Meier dos erros, sugerindo uma concordância entre a distribuição esperada dos resíduos e a distribuição observada na curva de sobrevivência estimada. Na parte final há uma pequena diferença entre as curvas sugerindo que um modelo com fração de cura pudesse ser uma boa opção para melhorar a modelagem dos dados.

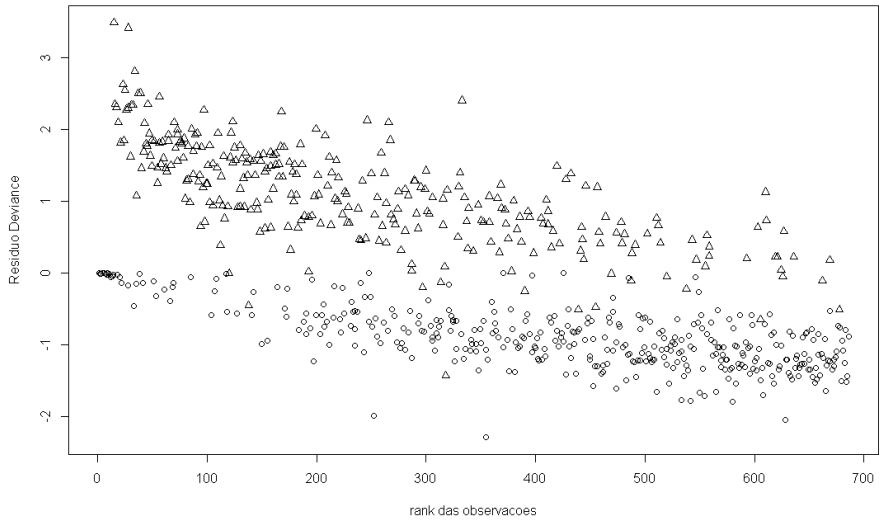

Figura 10: Resíduos Deviance vs Rank das Observações

3.6 Interpretação dos coeficientes

- **Grau do tumor:** Ambos os graus mais elevados de tumor (grau 2 e grau 3) estão associados a um menor log do tempo de sobrevivência em comparação com o grau 1 mantendo constantes todas as outras variáveis do modelo. Isso sugere que os pacientes com grau 2 ou 3 tendem a ter um tempo de sobrevivência menor do que aqueles com grau 1.

- **Terapia Hormonal:** Os pacientes que receberam terapia hormonal tendem a ter um tempo de sobrevivência maior em comparação com aqueles que não receberam essa terapia, mantendo constantes todas as outras variáveis do modelo. Essa diferença positiva no log do tempo de sobrevivência pode indicar um possível efeito benéfico da terapia hormonal na expectativa de vida dos pacientes com câncer de mama.

- **Tamanho do tumor e quantidade de linfonodos:** Ambos os coeficientes negativos indicam que, à medida que o tamanho do tumor ou o número de linfonodos aumenta, o log do tempo de sobrevivência diminui mantendo constantes todas as outras variáveis do modelo, sugerindo uma associação adversa entre essas variáveis e o tempo de sobrevivência dos pacientes com câncer de mama. Essas descobertas podem indicar a importância do tamanho do tumor e do envolvimento dos linfonodos na prognóstico da doença.

- **Progesterona:** Para cada aumento unitário no nível de progesterona, observamos um pequeno aumento no log do tempo de sobrevivência mantendo constantes todas as outras variáveis do modelo.

- **Terapia Hormonal*Progesterona:** A interação entre terapia hormonal e progesterona, o coeficiente positivo indica que para cada unidade de aumento na progesterona, associada à terapia hormonal, espera-se um pequeno aumento no log do tempo de sobrevivência.

- **Grau do tumor*Progesterona:** Também indicam um pequeno aumento no log do tempo de sobrevivência à medida que a progesterona aumenta para o grau 2 e 3 de tumor.

Referências

COLLETT, D. *Modelling Survival Data in Medical Research*. Springer US, 1994. (CRC Monographs on Statistics & Applied Probability). ISBN 9780412448805. Disponível em: ⟨https://books.google.com.br/books?id=hAQpAQAAMAAJ⟩.

COLOSIMO, E.; GIOLO, S. *Análise de sobrevivência aplicada*. Edgard Blücher, 2006. (ABE - Projeto Fisher). ISBN 9788521203841. Disponível em: ⟨https://books.google.com.br/books?id=g0-uOgAACAAJ⟩.

GOMES, J. B. F. *Notas de Aula de Análise de Sobrevivência*. 2023. Universidade de Brasília.

SUNG, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, v. 71, n. 3, p. 209–249, 2021. Disponível em: ⟨https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660⟩.