Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Davi Dantas Erthal 200016741

Eduardo Moreira Araújo 202043700

Júlia Garcia Ribeiro 202017960

Análise de Sobrevivência: estudo sobre câncer de pulmão em estágio avançado diagnosticado em veteranos do serviço militar americano

Brasília, DF

18 de Dezembro de 2023

Sumário

1 Introdução

Os modelos de sobrevivência são fundamentais em estatística quando o objetivo é analisar o tempo até a ocorrência de determinado evento de interesse. O fenômeno sob estudo pode envolver o tempo até a falha de um equipamento, o tempo até a morte de um paciente ou o tempo até a implementação de uma política pública, por exemplo. Nesse sentido, a análise de sobrevivência incorpora técnicas estatísticas poderosas e que podem ser aplicadas em diversas áreas como a médica, indústria, entre outras. Uma característica importante dessa classe de modelos é a possibilidade de trabalharmos com dados censurados, nos quais uma parte dos elementos ou indivíduos do estudo não experimentam o evento de interesse (COLOSIMO; GIOLO, 2006).

Na análise de sobrevivência duas abordagens são consideradas. A primeira é a modelagem paramétrica, na qual nos baseamos em distribuições de probabilidade capazes em explicar a variável tempo e a probabilidade de sobrevivência ou falha. Na abordagem não paramétrica, denominada modelo de Cox, não é atribuído uma distribuição específica, entretanto, há o pressuposto de riscos proporcionais. A abordagem selecionada depende, fundamentalmente, da análise exploratória dos dados.

Diante do exposto, a análise de sobrevivência nos permite desenvolver modelos de regressão robustos que nos ajudam a entender o tempo de sobrevivência e a influência de covariáveis sobre tal fenômeno. Análises estatísticas nesse campo nos auxiliam, por exemplo, no estudo sobre câncer de pulmão. Nessa perspectiva, o presente trabalho tem como objetivo estudar fatores de influência no tempo de sobrevivência de pacientes com câncer de pulmão em estágio avançado, diagnosticado em veteranos do serviço militar americano por volta de 1980.

De acordo com o sistema de assistência médica dos Estados Unidos *Veteran Administration* (2023), estima-se que mais de 8000 veteranos são diagnosticados e tratados por câncer de pulmão todos os anos. Se não identificado cedo, a doença entra em um estágio avançado e o paciente não pode ser mais operado. Nesse sentido, estudos estatísticos são importantes para avaliar se novos tratamentos podem aumentar o tempo de sobrevivência dos indivíduos ou até mesmo ajudar na cura do paciente.

A estrutura do relatório está dividida em três grandes partes: Na primeira desenvolvemos a **Metodologia**, na qual discutimos teoricamente conceitos inicias de análise de sobrevivência, a caracterização da distribuição Weibull e os métodos de estimação, como realizamos a análise descritiva tradicional e por métodos de sobrevivência e como ocorreu a modelagem paramétrica dos dados. Na segunda parte apresentamos os **Resultados** com uma análise descritiva dos dados, o ajuste de um modelo de regressão pela distribuição Weibull e a interpretação do modelo final. Por fim, resumimos os resultados e apresentamos sugestões e limitações do estudo

na parte de **Considerações Finais**.

A manipulação, visualização e análise dos dados foram realizadas por meio do software R, através da interface RStudio na versão 4.3.2. Os pacotes utilizados foram: *tidyverse*, *survminer*, *survival* e *AdequacyModel*. A aplicação foi realizada a partir de um banco de dados disponibilizado pela Administração de Veteranos dos Estados Unidos em um estudo acerca do efeito de um novo tratamento para pacientes com câncer de pulmão em estágio avançado e inoperável. O código em R e os dados podem ser baixados na íntegra a partir do endereço eletrônico disponibilizado em anexo.

2 Metodologia

2.1 Análise de sobrevivência: conceitos introdutórios

A análise de sobrevivência é utilizada quando o tempo for o objeto de interesse, seja esse interpretado como o tempo até a ocorrência de um evento ou o risco de ocorrência de um evento por unidade de tempo. Nesse sentido, a concepção é estudar determinados fenômenos a partir de uma análise de dados de natureza longitudinal (CARVALHO et al., 2011). O tempo de interesse pode ser chamado tanto de tempo de falha quanto tempo de sobrevivência, a depender do objeto e área de estudo.

Nos dados de sobrevivência é comum a perda da informação temporal completa, no entanto, esses dados não são descartados, visto que ainda fornecem informações sobre o tempo em que os indivíduos estiveram expostos ao risco e omiti-los pode acarretar em conclusões viesadas na análise estatística. A essas observações parciais é dado o nome de censura. Essa característica é o que difere a modelagem de dados em sobrevivência de outros modelos estatísticos, uma vez que precisamos incorporar informações parciais e incompletas de elementos ou indivíduos que, por alguma razão, não experimentaram o evento de interesse (COLOSIMO; GIOLO, 2006). As censuras podem ser classificadas da seguinte forma:

- **Censura à direita:** Ocorre quando o tempo de ocorrência do evento está à direita do tempo registrado.
  - **Tipo I:** O estudo é finalizado após um período de tempo fixo.
  - **Tipo II:** O estudo é finalizado após o evento de interesse ocorrer em um número fixo (k $\leq$ n) de indivíduos.
  - **Aleatória:** Casos em que as observações não experimentam o evento de interesse por motivos não controláveis, tais como: perda de acompanhamento, mudança de endereço, morte por motivos não associados ao estudo, entre outros.
- **Censura à esquerda:** Ocorre quando o tempo registrado é maior que o tempo de falha, ou seja, o evento de interesse já aconteceu quando o indivíduo foi observado.
- **Censura intervalar:** Ocorre em estudos nos quais os elementos têm acompanhamento periódico de forma que o evento de interesse ocorre em um intervalo de tempo.

Segundo Colosimo e Giolo (2006), conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e de censura, que correspondem à variável resposta $(t_i\,, \delta_i)$. Geralmente, também são medidas covariáveis que podem influenciar no

tempo de sobrevivência ou censura, o que nos levará ao desenvolvimento de um modelo de regressão. O tempo de falha é constituído pelo tempo inicial, estabelecido no início do estudo, pela escala de medida e pelo evento de interesse, ou seja, a falha, que pode ocorrer por um único motivo ou vários, nesse caso denominados *riscos competitivos* (PRENTICE et al., 1978).

A informação da variável resposta associada a cada indivíduo é representada pela indicadora abaixo:

$$\delta_i = \begin{cases} 1 \text{ caso } t_i \text{ seja tempo de falha} \\ 0 \text{ caso } t_i \text{ seja tempo de censura} \end{cases}$$

Ademais, considerando uma única variável aleatória contínua do tempo de falha (ou de sobrevivência) não-negativa, representada por T, podemos defini-la por meio de algumas funções fundamentais, conforme Lawless (2003):

Seja $f(t)$ a densidade de probabilidade (f.d.p) de T e $F(t)$ a função de distribuição acumulada (f.d.a), temos:

$$F(t) = P(T \leq t) = \int_0^t f(x)\, dx \tag{2.1.1}$$

A função de distribuição acumulada nos retorna a probabilidade do tempo de sobrevivência no intervalo $(0, t]$. Por sua vez, a função densidade de probabilidade, de acordo com o teorema fundamental do cálculo e a noção de derivada, corresponde ao limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo por unidade de tempo. A probabilidade de um indivíduo sobreviver até o tempo t, ou seja, não falhar até determinado tempo, é dada pela seguinte função de sobrevivência:

$$S(t) = P(T \geq t) = \int_t^\infty f(x)\, dx \tag{2.1.2}$$

S(t) é uma função monótona decrescente e contínua, com $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$. Outra função muito importante é a de taxa de falha (ou de risco), que especifica a taxa instantânea de falha no tempo t dado que o indivíduo sobreviveu até o tempo t, definida como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.1.3}$$

Em suma, a função de risco tende a ser mais informativa do que a função de sobrevivência em termos de forma. As funções de risco (taxa de falha) podem variar drasticamente entre si, entretanto, diferentes funções de sobrevivência podem apre-

sentar formas bem similares (COLOSIMO; GIOLO, 2006). Esse fato será essencial para a identificação de um modelo adequado para se ajustar aos dados. Também é útil definir a acumulada da função de risco, uma vez que ela será fundamental para obter $h(t)$ na estimação não paramétrica. Vejamos:

$$H(t) = P(T \leq t) = \int_0^t h(x)\, dx \tag{2.1.4}$$

É possível, ainda, relacionar as funções apresentadas a partir de suas propriedades. Essas características nos permitem, a partir de uma única função, obter as demais sem grande esforço. Além das fórmulas apresentadas acima, podemos listar mais algumas, conforme observamos abaixo:

$$F(t) = 1 - S(t) \tag{2.1.5}$$

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{\partial[\log S(t)]}{\partial t} \tag{2.1.6}$$

$$S(t) = \exp[-H(t)] \tag{2.1.7}$$

2.2 Distribuição Weibull: abordagem em sobrevivência

Uma das etapas da análise de dados em sobrevivência é identificar, no caso paramétrico, uma distribuição de probabilidade adequada para a modelagem do tempo de sobrevivência. Veremos mais adiante no capítulo de resultados que a distribuição Weibull será a mais indicada para desenvolvermos um modelo de regressão em relação ao tempo, logo, precisamos entender um pouco melhor a distribuição Weibull e o método de estimação utilizado neste trabalho.

Se a variável aleatória tempo (T) segue uma distribuição Weibull, então, podemos definir analiticamente as funções densidade de probabilidade $f(t)$, a função de sobrevivência $S(t)$ e a função de risco $h(t)$. Considerando $\gamma > 0$ o parâmetro de forma e $\alpha > 0$ o parâmetro de escala com a mesma unidade de medida de t, em que $t \geq 0$, Dobson e Barnett (2008) nos apresenta a seguinte configuração:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right] \tag{2.2.1}$$

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right] \tag{2.2.2}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \tag{2.2.3}$$

A distribuição Weibull é bastante utilizada em sobrevivência por sua flexibilidade em modelar funções de risco mónotonas decrescentes, crescentes ou constantes. Ao olharmos para a fórmula de $h(t)$, é possível observar que a função de risco é decrescente para $\gamma < 1$, constante para $\gamma = 1$ e crescente para $\gamma > 1$. Cabe observar que quando temos $\gamma = 1$ caímos na distribuição exponencial, caso particular da Weibull.

O próximo passo é discutirmos como o processo de estimação ocorre para o modelo Weibull. Nesse sentido, precisamos definir algumas notações importantes. Vamos considerar n o número de observações, x_i um vetor de variáveis explicativas, t_i o tempo de sobrevivência e δ_i uma indicação de censura em que $\delta_i = 1$ o tempo de sobrevivência é não censurado e $\delta_i = 0$ se o tempo é censurado. O método de estimação clássico em análise de sobrevivência é por máxima verossimilhança, dessa forma, vamos incorporar a informação sobre os dados não censurados a partir da função densidade da Weibull e a informação de observações censuradas é atribuída à função de sobrevivência. Portanto, obtemos a seguinte função de verossimilhança:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{2.2.4}$$

É possível obter a função log-verossimilhança e substituir a função densidade pela função de risco. Dessa forma, conseguimos o seguinte resultado:

$$l = \sum_{i=1}^{n} [\delta_i \log h(t_i) + \log S(t_i)] \tag{2.2.5}$$

$$l = \sum_{i=1}^{n} \left[\delta_i \log \left(\frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \right) - \left(\frac{t}{\alpha} \right)^\gamma \right] \tag{2.2.6}$$

A função log-verossimilhança dependerá do parâmetro de forma γ da distribuição de probabilidade e dos parâmetros da componente linear $X^T\beta$, considerando que vamos atribuir $\alpha = \exp(X^T\beta)$. Esses parâmetros podem ser estimados por diversos métodos iterativos como o Fisher Scoring ou Newton-Rapshon, considerando que os coeficientes de regressão e o parâmetro γ serão estimados a partir da derivada da log-verossimilhança. O método de Newton-Rapshon é um dos mais utilizados e a inversa da matriz de informação utilizada no processo de iteração nos retorna uma estimativa assintótica da matriz de variância e covariância dos parâmetros estimados (Dobson e Barnett, 2008).

A distribuição Weibull faz parte da classe de modelos denominada de tempo de falha acelerado (*accelerated failure time*). Nesse tipo de modelagem, considera-se o logaritmo do tempo como variável resposta e há a inclusão de um termo de erro que segue alguma distribuição de probabilidade (SAIKIA; BARMAN, 2017). Os efeitos das variáveis explicativas estão relacionadas diretamente com a função de sobrevivência, conforme vemos a seguir:

$$S(t \mid x) = s_0(\exp(-X^T\beta)t)$$

Aqui, $S(t \mid x)$ é a função de sobrevivência no tempo t e $s_0(\exp(-X^T\beta)t)$ é a função de sobrevivência no nível de referência das variáveis explicativas no tempo t. Em suma, se $X^T\beta$ decresce, então o tempo até a falha acelera e a função sobrevivência decai mais rapidamente, do contrário, $X^T\beta$ cresce e o tempo até a falha desacelera e a função sobrevivência decai mais lentamente (DOBSON; BARNETT, 2008).

Os modelos AFT são também chamados de modelos de locação-escala e, quando utilizados o logaritmo do tempo de sobrevivência, temos a seguinte relação:

$$\log T = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \sigma\epsilon$$

Nessa estrutura β_0 é o intercepto e σ o parâmetro de escala. Além disso, ϵ corresponde aos erros do modelo e segue uma distribuição de probabilidade especificada. Se T segue uma distribuição Weibull, então ϵ e, consequentemente $\log T$, segue uma distribuição do valor extremo (SAIKIA; BARMAN, 2017). Neste trabalho não vamos fazer a transformação logarítmica da variável resposta diretamente, entretanto, é importante entender a classe dos modelos de tempo de falha acelerado e observar que a função *Survreg*, que será utilizada no R, trabalha com esse tipo de modelagem e transformação. Nesse sentido, vamos considerar o exponencial dos coeficientes no momento de interpretação dos resultados para voltarmos para a escala dos tempos ao invés da escala logarítmica.

2.3 Análise descritiva: operacionalização

2.3.1 Técnicas tradicionais

As medidas estatísticas, como mínimo, máximo, quartis, média e mediana, são essenciais para proporcionar uma compreensão abrangente da distribuição dos dados quantitativos. Nesse sentido, vamos listar as medidas resumo que foram utilizadas na análise descritiva.

O mínimo e máximo identificam os extremos do conjunto de dados.

$$\text{Mínimo:} \quad X_{\min} = \min(X) \tag{2.3.1}$$

$$\text{Máximo:} \quad X_{\max} = \max(X) \tag{2.3.2}$$

Enquanto os quartis oferecem insights sobre a dispersão e a presença de valores atípicos. A mediana é robusta em relação a outliers, representando o ponto central dos dados.

$$Q_1 = \frac{n+1}{4}\text{-ésimo valor} \tag{2.3.3}$$

$$Q_2 = \frac{2(n+1)}{4}\text{-ésimo valor (Mediana)} \tag{2.3.4}$$

$$Q_3 = \frac{3(n+1)}{4}\text{-ésimo valor} \tag{2.3.5}$$

A média fornece uma medida de tendência central, representando o valor médio.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.3.6}$$

Para variáveis qualitativas, a descrição baseia-se em frequências e porcentagens, oferecendo uma visão clara da distribuição das categorias. Entender a distribuição de frequências é crucial para identificar padrões, preferências ou desequilíbrios nas respostas qualitativas.

As tabelas de contagem sintetizam a distribuição, enquanto as medidas resumo e gráficos proporcionam uma representação visual, facilitando a interpretação. Os histogramas revelam padrões de concentração, e os boxplots ajudam a identificar a presença de outliers (MORETTIN; BUSSAB, 2010).

2.3.2 Técnicas em análise de sobrevivência

Kaplan-Meier

O estimador de Kaplan-Meier é uma técnica não paramétrica usada para estimar a função de sobrevivência em dados de tempo de vida censurados. A função de sobrevivência, denotada por S(t), representa a probabilidade de um indivíduo sobreviver além do tempo t (COLOSIMO; GIOLO, 2006). A fórmula para o estimador de Kaplan-Meier é dada por:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.3.7}$$

Onde:

- t_i são os tempos de eventos observados,
- d_i são os números de eventos ocorridos em t_i,
- n_i representa o número de indivíduos sob risco no tempo t_i.

Os eventos ocorridos podem significar falha ou sobrevivência, por exemplo. O estimador de Kaplan-Meier é também conhecido como estimador produto. Ao utilizarmos tal ferramenta é possível identificar padrões nos dados e obter uma função de sobrevivência empírica, discutir possíveis tipos de modelagem, comparar grupos a partir de variáveis categóricas, etc.

Teste Log-Rank / Wilcoxon

Os testes de log-rank ou de Wilcoxon são utilizados para comparar as curvas de sobrevivência de dois ou mais grupos (COLOSIMO; GIOLO, 2006). Assim, temos as hipóteses:

$$\begin{cases} H_0 : \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1 : \text{Existe diferença entre as curvas de sobrevivência} \end{cases}$$

Deve-se agrupar e ordenar os tempos distintos observados entre os grupos, t_j, $j = 1, 2, ..., k$. Assim obtemos as quantidades:

- n_j representa o número total de indivíuos sob risco em um tempo imediatamente anterior a t_j,
- n_{1j} representa o número de indivíduos do grupo 1 sob risco em um tempo imediatamente anterior a t_j,

- d_j representa o número total de falhas no tempo t_j,
- d_{1j} representa o número de falhas do grupo 1 no tempo t_j;

A média $E[d_1 j]$ e variância $V(d_{1j})$ dadas por:

$$E[d_{1j}] = d_j \frac{n_{1j}}{n_j} = w_{1j} \tag{2.3.8}$$

$$V[d_{1j}] = d_j \frac{n_{1j}}{n_j}(1 - \frac{n_{1j}}{n_j})\frac{n_j - d_j}{n_j - 1} \tag{2.3.9}$$

A estatística $d_{1j} - w_{1j}$ tem média zero e variância V_{1j}. Assim, a estatística do teste é definida como:

$$T = \frac{[\sum_{j=1}^{k} u_j(d_{1j} - w_{1j})]^2}{\sum_{j=1}^{k} u_j^2 V_{1j}} \tag{2.3.10}$$

Que tem aproximadamente distribuição Qui-Quadrado com 1 grau de liberdade. De acordo com o valor de u_j temos os testes:

- $u_j = 1$ Teste de Log-Rank, coloca mesmo peso para todo o eixo do tempo,
- $u_j = n_j$ teste de Wilcoxon, utiliza peso igual ao número de indivíduos sob risco, coloca mais peso na proporção inicial do eixo do tempo. Em situações que não temos a proporcionalidade dos riscos esse teste é mais adequado do que o teste de Log-Rank.

Gráfico TTT

O gráfico TTT exibe a função acumulativa inversa do estimador de Kaplan-Meier. Ele representa a probabilidade de ocorrer um evento antes de um determinado tempo, proporcionando uma visão dinâmica da taxa de eventos ao longo do tempo.

Para criar o gráfico TTT:

- No eixo horizontal, representamos o tempo.
- No eixo vertical, representamos a probabilidade de um evento ocorrer antes de um determinado tempo.

Cada ponto no gráfico TTT indica um evento ocorrido, e a linha do gráfico se move para baixo sempre que um evento acontece, refletindo a diminuição da

probabilidade de sobrevivência. Portanto, o gráfico TTT mostra a evolução da probabilidade cumulativa de sobrevivência ao longo do tempo (COLOSIMO; GIOLO, 2006). Esse tipo de gráfico é importante para avaliar o comportamento da função de risco e, consequentemente, na escolha da distribuição de probabilidade para a modelagem do tempo de sobrevivência.

Gráfico de Risco Acumulado

O gráfico de risco acumulado é outra ferramenta visual na análise de sobrevivência que destaca a taxa acumulada de eventos ao longo do tempo. Ele é particularmente útil quando há mais de dois grupos de interesse, permitindo comparar a evolução das taxas de eventos entre esses grupos.

Para criar o gráfico de risco acumulado:

- No eixo horizontal, representamos o tempo.
- No eixo vertical, representamos a taxa acumulada de eventos.

Cada linha no gráfico de risco acumulado representa um grupo ou uma categoria de interesse. À medida que o tempo progride, as linhas mostram como a taxa de eventos acumulados evolui para cada grupo. Se há diferenças notáveis entre as linhas, isso indica variações nas taxas de eventos entre os grupos.

Ambos os gráficos, TTT e de risco acumulado, são ferramentas valiosas para entender a dinâmica temporal dos eventos e para comparar diferentes grupos em estudos de sobrevivência. Eles complementam as análises estatísticas tradicionais, proporcionando uma representação visual intuitiva das diferenças nas experiências de eventos ao longo do tempo.

Teste de Schoenfeld

O Teste de Resíduos de Schoenfeld é utilizado para testar se os riscos são proporcionais entre grupos. Trata-se de uma ferramenta que nos retorna um indício de utilizar ou não um modelo paramétrico (COLOSIMO; GIOLO, 2006).

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

O uso dos resíduos padronizados de Schoenfeld para avaliar a suposição de riscos proporcionais é baseado em um resultado que considera o modelo expresso pela seguinte fórmula:

$$h(t; x) = h_o(t) \exp^{X^T \beta}$$

2.4 Modelagem paramétrica: aspectos metodológicos

A escolha do modelo probabilístico a ser utilizado é uma etapa fundamental na análise paramétrica de dados de tempo de sobrevivência. A primeira técnica consiste em comparar os gráficos da função de sobrevivência dos modelos propostos com a função de sobrevivência empírica estimada por Kaplan-Meier. O modelo selecionado é o que melhor acompanhar a curva de sobrevivência estimada pelo método de Kaplan-Meier (COLOSIMO; GIOLO, 2006).

O próximo passo é utilizar os critérios de parcimônia Critério de Akaike (AIC), Critério de Akaike corrigido (AICc) e Critério de Informação Bayesiano (BIC). O melhor ajuste é representado pelo modelo com os menores valores em tais medidas. Considerando p o número de parâmetros estimados, n o número de observações e $\log(L(\hat{\theta}))$ o logaritmo da função de verossimilhança no ponto de máximo, temos:

Critério de Akaike (AIC)

$$AIC = -2\log(L(\hat{\theta})) + 2p \tag{2.4.1}$$

Critério de Akaike corrigido (AICc)

$$AICc = AIC + \frac{2p(p+1)}{n-p-1} \tag{2.4.2}$$

Critério de informação Bayesiano (BIC)

$$BIC = -2\log(L(\hat{\theta})) + p\log(n) \tag{2.4.3}$$

Após a escolha da distribuição de probabilidade para modelar o tempo de sobrevivência ajusta-se um modelo de regressão aos dados, tentando observar o efeito das covariáveis sob estudo em relação à variável resposta.

Sob o contexto de modelos de sobrevivência, há a necessidade de se avaliar a significância dos estimadores $\hat{\beta}_i$, $i = 1, 2..., p$. Assim, tem-se o teste de hipóteses com $H_0 : B_j = 0$ e $H_1 : B_j \neq 0$, isto é, para o caso em que rejeita-se a hipótese nula, tem-se que o coeficiente de regressão é significantemente diferente de zero (DOBSON; BARNETT, 2008). O teste de Wald é um dos mais utilizados e, sob hipótese nula, tem aproximadamente distribuição qui-quadrado com 1 grau de liberdade e a seguinte estatística de teste:

$$W_T^2 = \left(\frac{\hat{\beta}_j - \beta^{(0)}}{ep(\hat{\beta}_j)} \right)^2 \sim \chi^2_{(1)} \tag{2.4.4}$$

Para a comparação entre diferentes modelos candidatos, além dos critérios de parcimônia, também é adequado utilizar o Teste da Razão de Verossimilhança (TRV) para modelos encaixados. Sob hipótese nula, temos aproximadamente uma distribuição qui-quadrado com p graus de liberdade e a seguinte estatística de teste:

$$TRV = 2[\log(L(\hat{\theta})) - \log(L(\hat{\theta_0}))] \tag{2.4.5}$$

Nessa etapa rodamos manualmente um algoritmo análogo ao método stepwise. Dessa forma, começamos a modelagem observando cada covariável de forma individual, depois introduzimos pares de variáveis, grupos de variáveis e o modelo completo. Além disso, modelos mais complexos com interações também foram ajustados.

Para a interpretação dos coeficientes de regressão consideramos o exponencial das estimativas. Isso ocorre pois, no modelo baseado na distribuição Weibull, a exponencial do coeficiente estimado nos retorna quanto o tempo mediano aumenta ou reduz na comparação entre categorias de uma variável ou no acréscimo de uma unidade em covariáveis quantitativas. Como esse valor é uma razão entre tempos medianos, também é possível interpretar como uma razão de chances de sobrevivência (COLOSIMO; GIOLO, 2006).

Por fim, devemos considerar a qualidade do ajuste dos modelos. Nesse sentido, a avaliação do modelo final é realizada a partir da análise de resíduos de Cox-Snell, martingal e deviance. Essas técnicas são usadas como um meio de rejeitar modelos claramente inapropriados e não para provar que um particular modelo está correto (KLEIN; MOESCHBERGER, 2003).

Os resíduos de Cox-Snell auxiliam a examinar o ajuste global do modelo. O gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, auxiliam nesse processo. Esses resíduos são definidos como a função de risco acumulada do modelo ajustado e podem ser escritos por:

$$\hat{e}_i = \hat{H}(t_i \,|\, x_i) = \hat{H}(y_i \,|\, x_i) \tag{2.4.6}$$

Uma vez que os resíduos são usados para identificar discrepâncias entre um modelo ajustado e o conjunto de dados, é conveniente buscar uma definição para resíduos que levem em consideração a contribuição de cada observação sobre a medida

de qualidade de ajuste (COLOSIMO; GIOLO, 2006). Nesse sentido, temos os resíduos de Martingal:

$$\hat{r}_m = \delta_i - \hat{H}(t_i \,|\, x_i) = \delta_i - \hat{e}_i \tag{2.4.7}$$

Para tornar os resíduos de martingal mais simétricos em torno de zero, calculamos os resíduos deviance. Se o modelo for apropriado, esses resíduos devem apresentar um comportamento aleatório em torno de zero (COLOSIMO; GIOLO, 2006).

OBS: Cabe ressaltar que em todas as análises e testes realizados neste trabalho o nível de significância escolhido foi de 10%.

2.5 Banco de dados: seleção e contextualização

A base de dados escolhida para o presente estudo é referente a dados sobre câncer de pulmão da Administração de Veteranos, fornecidos por volta de 1980. A Administração de Veteranos é um sistema de assistência médica para veteranos militares dos Estados Unidos que atende, principalmente, pessoas com problemas ocasionados pela própria prestação de serviço militar. O sistema atende mais de 9 milhões de pessoas todos os anos. Esta base possui dados sobre um experimento envolvendo homens com câncer de pulmão inoperável avançado que foram randomizados para quimioterapia padrão ou teste. O ponto principal para a comparação da terapia foi a hora da morte.

Apenas 9 dos 137 tempos de sobrevivência foram censurados. Como é comum em tais estudos, houve muita heterogeneidade entre pacientes em, por exemplo, extensão da doença e patologia, tratamento prévio da doença, antecedentes demográficos e estado de saúde inicial. Os dados no apêndice incluem informações sobre uma série de covariáveis que medem alguns aspectos desta heterogeneidade, além da variável resposta do estudo que envolve o tempo e a indicação de censura:

Tempo: Tempo em dias desde o início do experimento até o dia da morte do paciente. Tipo da variável: quantitativa contínua.

Censura: 0 = houve censura e 1 = não houve censura. Tipo da variável: qualitativa nominal.

Score: Uma medida aleatória de avaliação do status de desempenho do paciente: 10-30 totalmente hospitalizado, 40-60 confinamento parcial, 70-90 capaz de cuidar de si mesmo. Tipo da variável: quantitativa discreta. Embora tenhamos tais interpretações para faixas de score, a variável não foi categorizada.

Meses diagnóstico Tempo em meses desde o diagnóstico até a randomização. Tipo da variável: quantitativa contínua.

Idade: Idade do paciente em anos. Tipo da variável: quantitativa contínua.

Terapia anterior Terapia prévia; 0 = não, 10 = sim. Tipo da variável: qualitativa nominal.

Célula: Tipo histológico de tumor: escamoso, células pequenas, adeno, células grandes. Tipo da variável: qualitativa nominal.

Tratamento: 0 = padrão, 1 = teste. Tipo da variável: qualitativa nominal.

3 Resultados

3.1 Análise descritiva I: técnicas tradicionais

3.1.1 Univariada

Para as variáveis Tempo, Score, Meses Diagnóstico e Idade serão utilizadas medidas resumo, como máximo, mínimo, quartis, média e mediana, para resumir as distribuições além de histogramas e boxplots para melhor visualização.

Tabela 1: Medidas Resumo

Medida	Tempo	Score	Meses Diagnóstico	Idade
Mínimo	1	10	1	34
1° Quartil	25	40	3	51
Mediana	80	60	5	62
Média	121,60	58,57	8,77	58,31
3° Quartil	144	75	11	66
Máximo	999	99	87	81

Figura 1: Histogramas das variáveis Tempo, Score, Meses Diagnóstico e Idade

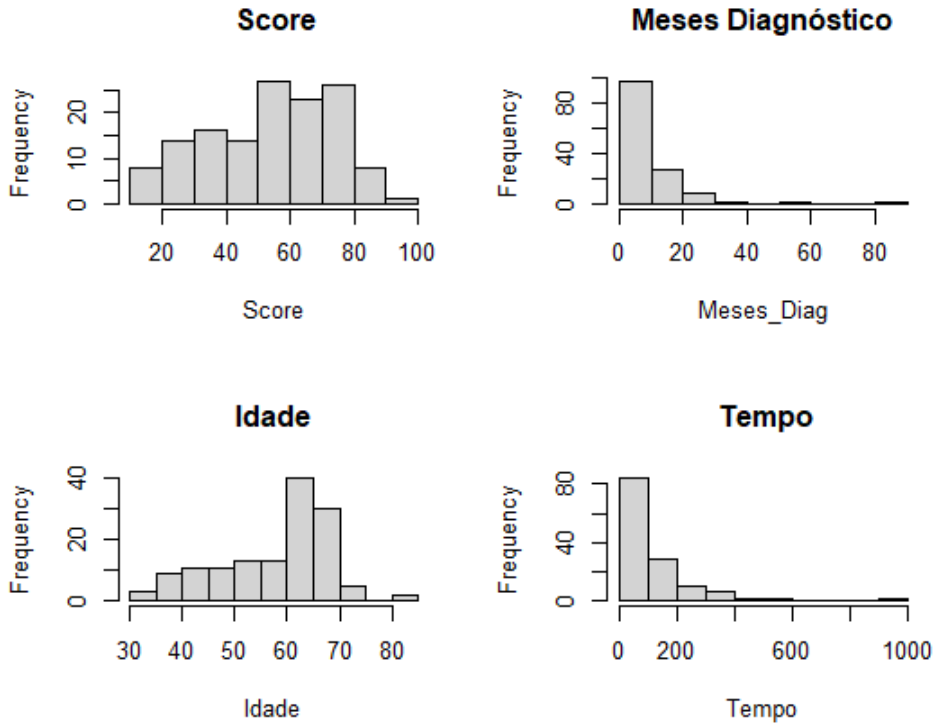
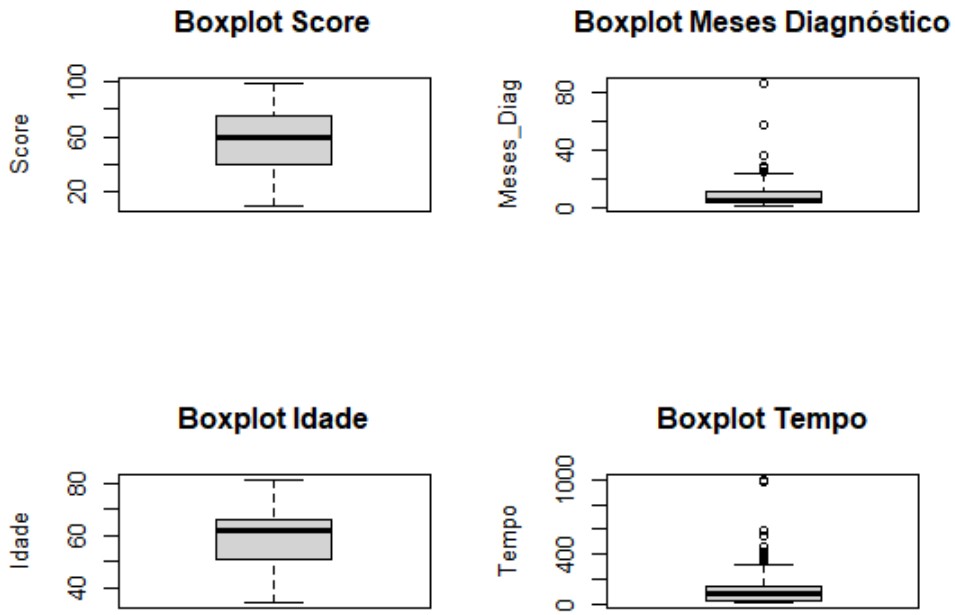


Figura 2: Boxplots das variáveis Tempo, Score, Meses Diagnóstico e Idade



Principais informações levantadas:

- Predominância de idades entre 60 e 70 anos.
- Assimetria à direita do tempo de sobrevivência. Quantidade significativa de mortes antes dos 200 dias.
- Score de desempenho com maiores frequências entre 50 e 80. No geral, a maior parte dos indivíduos estão parcialmente hospitalizados.

Para as variáveis Censura, Tratamento, Célula e Terapia Anterior as distribuições serão apresentadas por meio de frequências absolutas e relativas. Aqui vale ressaltar a baixa presença de censura nos dados de acordo com a natureza do estudo e o fato de 71% dos indivíduos não terem feito algum tipo de terapia anterior.

Tabela 2: Tabela de Frequência para a Variável *Censura*

Censura	**Frequência**	**Frequência Relativa**
0 = Censura	9	0.066
1 = Falha	128	0.934
Total	137	1.00

Tabela 3: Tabela de Frequência para a Variável *Tratamento*

Tramento	**Frequência**	**Frequência Relativa**
1	69	0.503
2	68	0.497
Total	137	1.00

Tabela 4: Tabela de Frequência para a Variável *Célula*

Célula	**Frequência**	**Frequência Relativa**
1	35	0.255
2	48	0.351
3	27	0.197
4	27	0.197
Total	137	1.00

Tabela 5: Tabela de Frequência para a Variável *Terapia Anterior*

Terapia Anterior	**Frequência**	**Frequência Relativa**
0	97	0.71
10	40	0.29
Total	137	1.00

3.1.2 Bivariada

A análise bivariada com o tempo é essencial para entender como diferentes variáveis podem influenciar a taxa de eventos ao longo do período de estudo. Nesta seção, examinaremos a relação entre cada variável e o tempo, explorando padrões e identificando possíveis associações.

É importante notar que, para esta análise, estaremos focando exclusivamente nas observações que não foram censuradas. Isso garantirá que nossa análise seja direcionada apenas aos eventos ocorridos, permitindo uma avaliação mais precisa das relações temporais.

Figura 3: Gráficos de dispersão de Tratamento, Célula e Terapia Anterior por Tempo

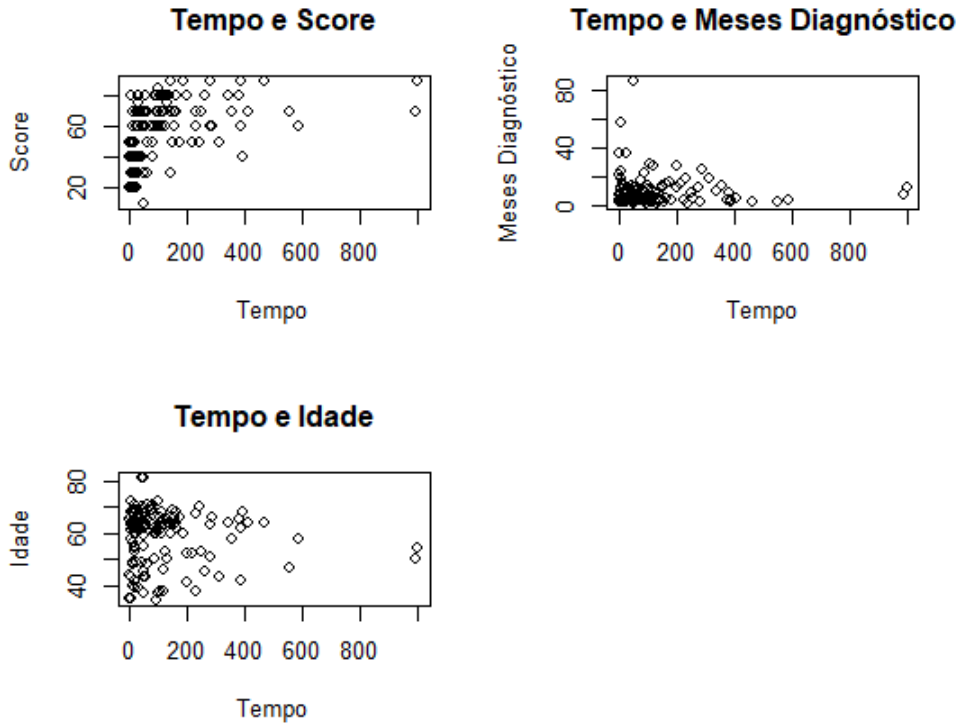
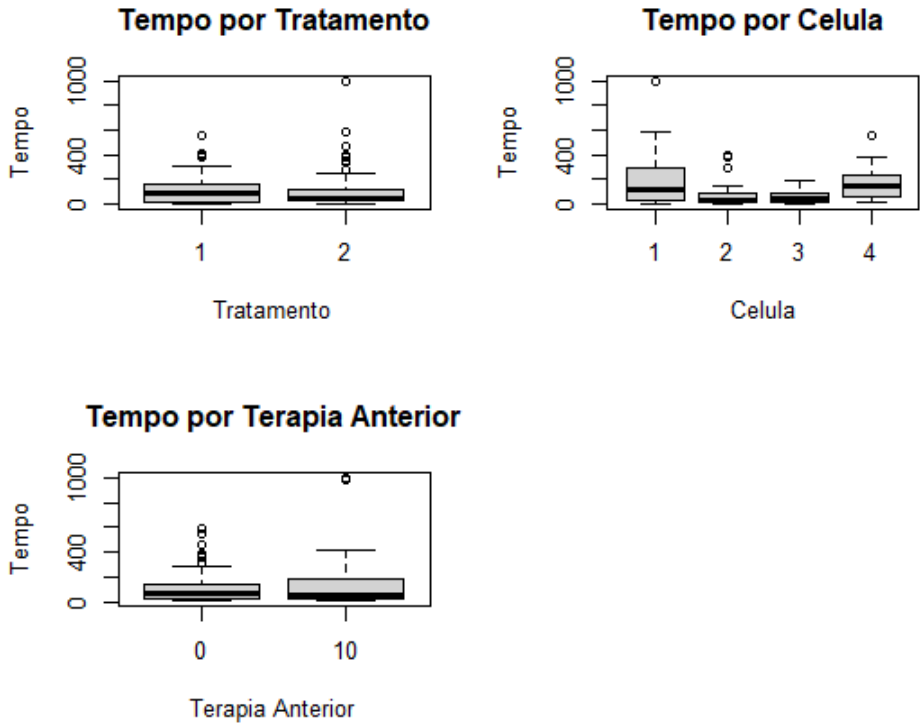


Figura 4: Boxplots de Tratamento, Célula e Terapia Anterior por Tempo



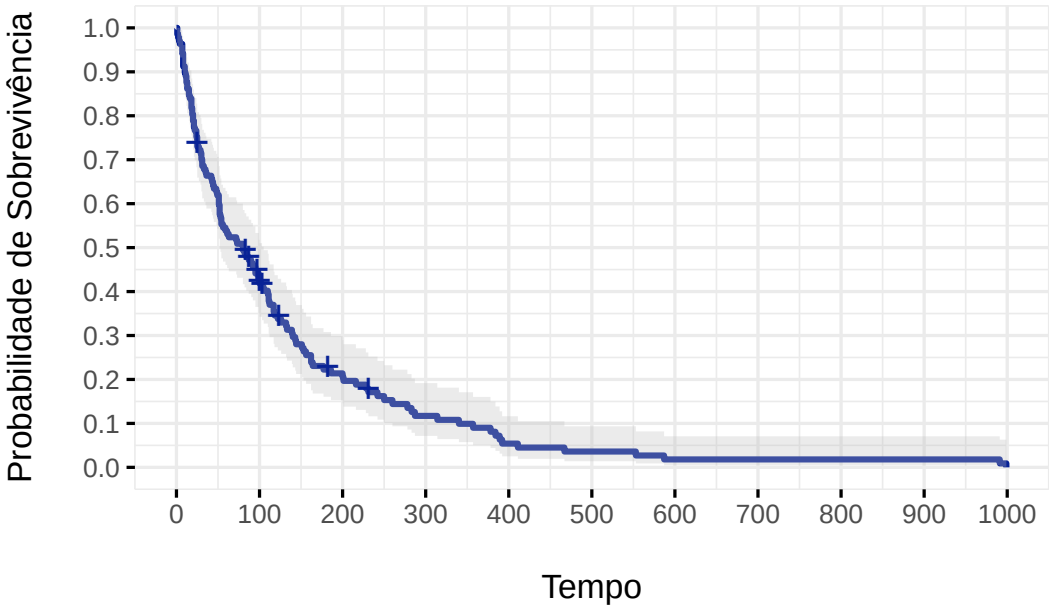
Principais informações levantadas:

- Como a variável tempo é bem assimétrica e possui outliers, torna-se difícil analisar esta variável em conjunto com outras, sobretudo, no gráfico de dispersão, no qual não foi possível observar associação.
- Apesar da assimetria e das observações discrepantes, parece haver diferença entre as células pequena e adeno em relação as células escamosa e grande. O tempo médio e mediano de sobrevivência aparentam ser menores para tumores pequenos e do tipo adeno.
- A análise descritiva utilizando técnicas de sobrevivência será mais eficiente em nos retornar hipóteses do que a análise descritiva tradicional, conforme esperado.

3.2 Análise descritiva II: modelos de sobrevivência

3.2.1 Modelo de sobrevivência geral

Figura 5: Função de Sobrevivência de Kaplain Meier



Ao iniciar a análise descritiva a fim de encontrar a distribuição de probabilidade da variável tempo do estudo, deve-se iniciar com o gráfico da função de sobrevivência de Kaplain Meier, em que é possível entender como foram os tempos de falha e como as censuras estão distribuídas ao longo do estudo.

Pelo gráfico 5 é possível observar que há mais censuras até o período de 100 dias. Nota-se que até próximo dia 200, a queda da probabilidade de sobrevivência é muito significativa. Além disso, é importante destacar que o gráfico mostra que a probabilidade de sobrevivência chega a 0 e antes disso, há um momento de estabilidade desta probabilidade de sobrevivência, do dia 600 até próximo ao dia 1000. Ao todo há 9 observações censuradas e não há indícios de que um modelo de frações de cura seja o mais adequado, visto que a curva decai até o nível 0 em termos de probabilidade. Não existe uma especificação clara no estudo, entretanto, as censuras identificadas aparentam ser à direita aleatória.

Figura 6: Função de Risco Acumulado - Nelson Aalen

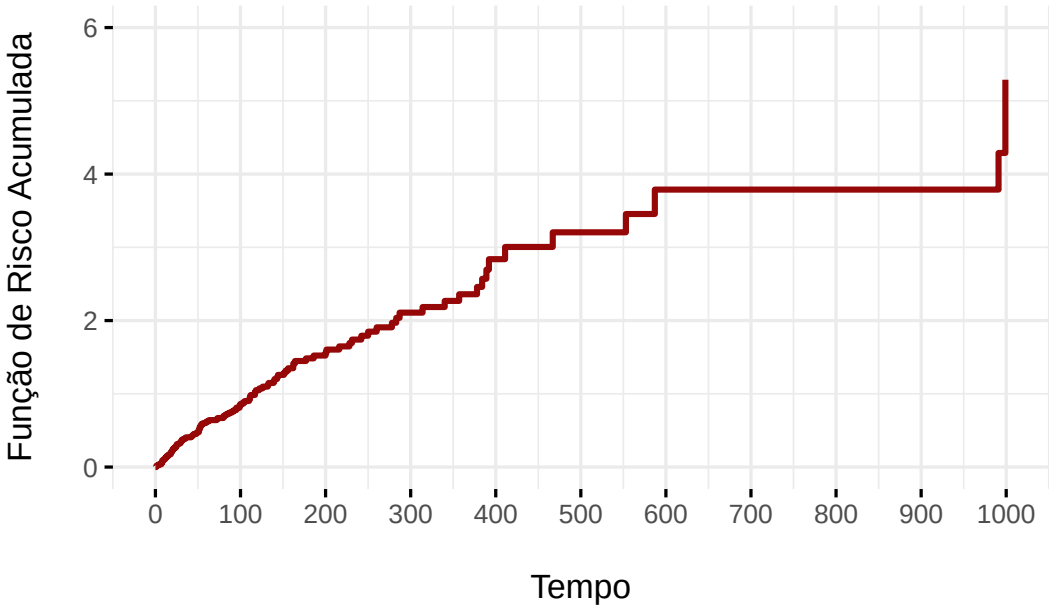
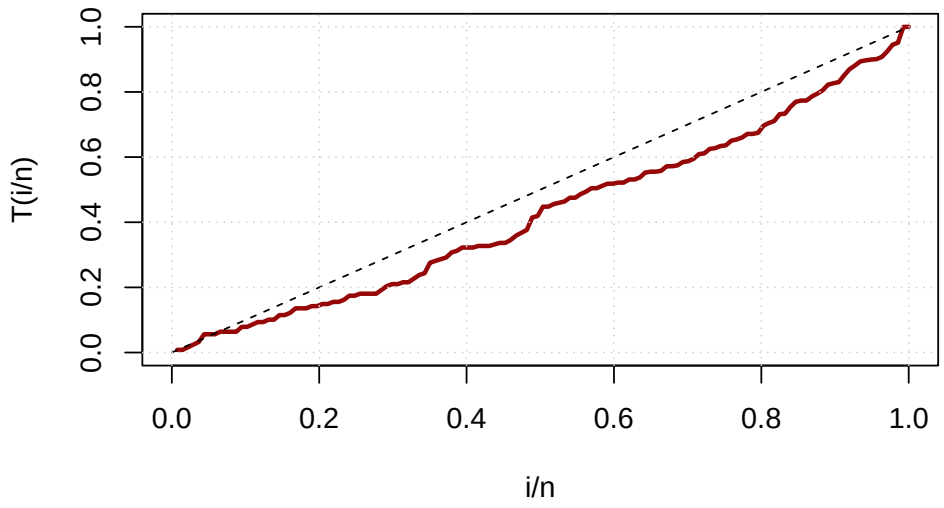


Figura 7: Gráfico TTT



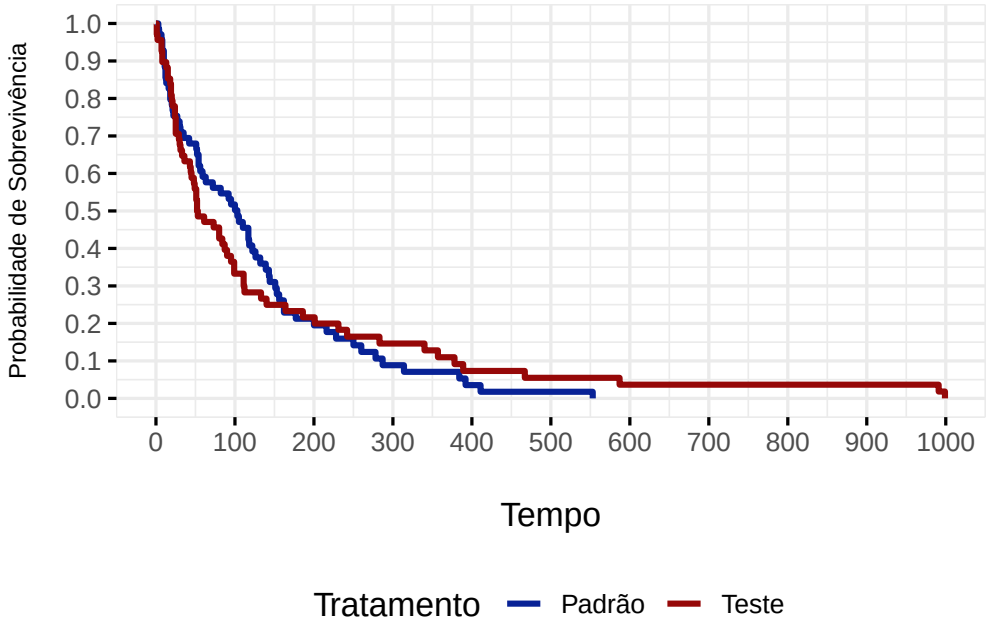
Ao observar o gráfico da função de risco acumulada por Nelson Aalen e o gráfico TTT, é possível já haver indícios da distribuição de probabilidade do estudo, que será fundamental para a construção do modelo paramétrico.

O gráfico da função de risco acumulada sugere uma curva de risco unimodal, que é indício das distribuições log-normal e log-logística. Entretanto, a curva observada no gráfico TTT é convexa, ou seja, indica uma função de risco monotonicamente decrescente. Tais características sugerem a distribuição Weibull ou log-logística. Mais a frente neste estudo será feito mais técnicas para entender e descobrir a real

distribuição para a análise de sobrevivência. É importante destacar que o gráfico TTT é um ótimo indicador da distribuição, ainda mais neste caso em que não se tem muita censura no estudo (gráfico TTT é sensível a grandes quantidades de censura).

3.2.2 Modelo de sobrevivência por Tratamento

Figura 8: Função de Sobrevivência de Kaplain Meier por Tratamento

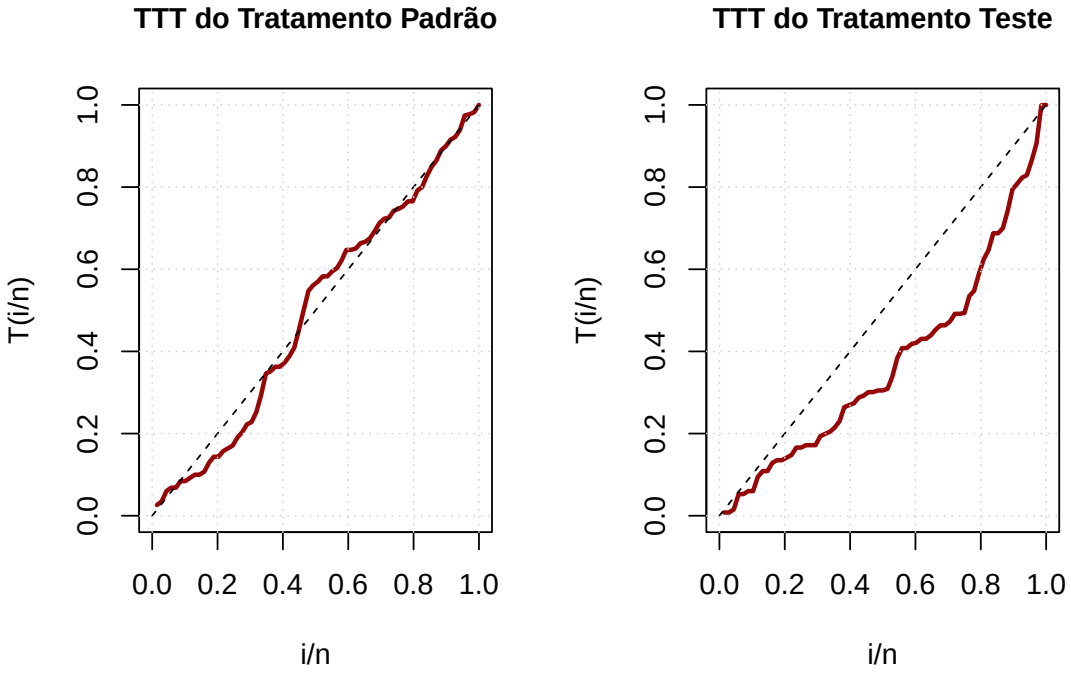


Observando agora a função de sobrevivência pela variável Tratamento, tem-se o gráfico acima que mostra a função de Kaplain Meier pelos dois grupos de tratamento: Quimioterapia Padrão e Quimioterapia Teste.

Comparando os tratamentos, o padrão chega a uma probabilidade de sobrevivência igual a 0 aproximadamente 450 dias antes de quando este mesmo cenário ocorre com o tratamento teste. Apesar de que no tratamento teste nos primeiros dias há uma queda mais abrupta na probabilidade de sobrevivência, o tratamento teste mantém uma probabilidade de sobrevivência em até 1000 dias, enquanto no tratamento padrão o paciente sobrevive até aproximadamente 550 dias.

Não parece haver uma diferença significativa entre as curvas de sobrevivência dos dois grupos, de forma geral. A maior diferença ocorre em um intervalo de tempo muito pequeno em torno de $t = 100$, entretanto, não é um padrão. Além disso, as curvas se cruzam mais de uma vez, logo, não existe indicação de riscos proporcionais.

Figura 9: Gráfico TTT por Tratamento



Pelos gráficos TTT acima, o tratamento padrão indica a forma U (banheira) para a função de risco, porém o tratamento teste tem uma curva convexa que indica função de risco decrescente que pode ser modelada, por exemplo, pela distribuição Weibull.

A seguir será feito um teste para testar se há diferenças entre as curvas de sobrevivência da variável tratamento: Padrão e Teste. As curvas se cruzam ao longo do gráfico de função de sobrevivência de Kaplain Meier. Logo, o teste utilizado é o de Wilcoxon ao nível de significância de 10% , com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1 : \text{Existe diferença entre as curvas de sobrevivência} \end{cases}$$

Quadro 1: Resultado do teste de Wilcoxon

P-valor	**Decisão do teste**
0,4	Não Rejeita H_0

Com base nos resultados do quadro acima, não rejeita-se a hipótese nula. Portanto, não há diferença entre as curvas de sobrevivência dos tratamentos.

A seguir será realizado o Teste de Resíduos de Schoenfeld para testar se os riscos são proporcionais entre os tratamentos e será um indício de utilizar ou não um

modelo paramétrico. Logo, estas são as hipóteses ao nível de 10% de significância:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

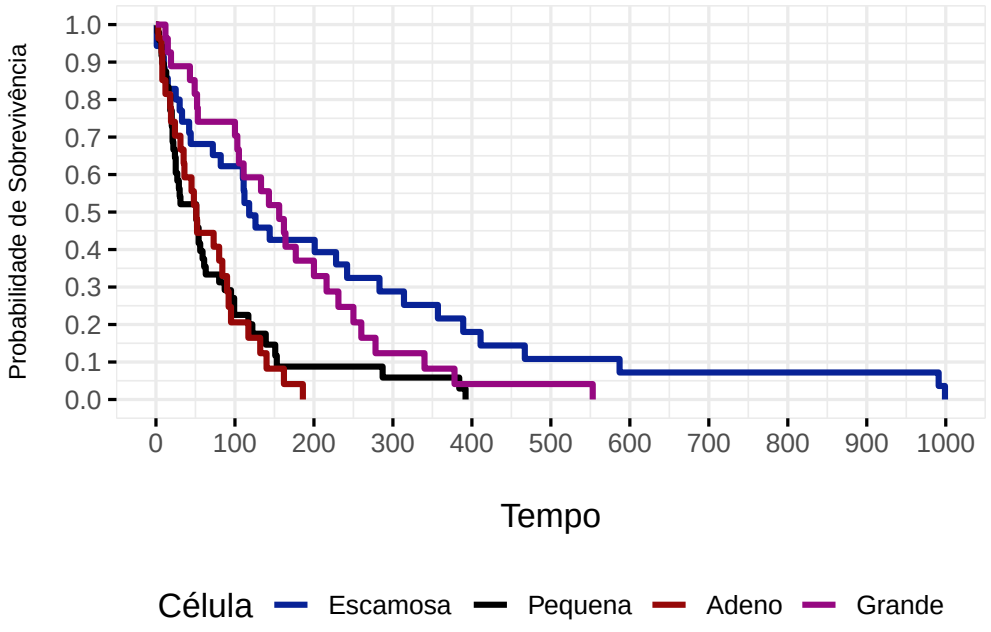
Quadro 2: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
0,06	Rejeita H_0

Dado o resultado observado no quadro acima, rejeita-se a hipótese nula e afirma-se que os riscos não são proporcionais. O teste de Wilcoxon realmente é mais indicado do que o log-rank e o pressuposto de riscos proporcionais para um modelo não paramétrico de Cox não parece ser adequado.

3.2.3 Modelo de sobrevivência por Célula

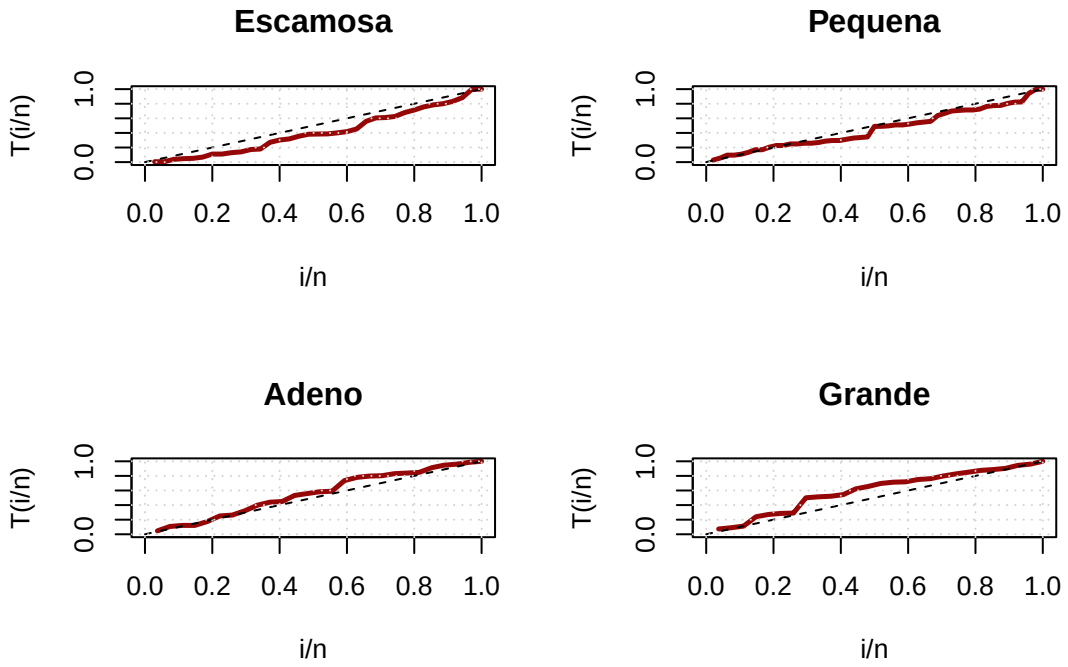
Figura 10: Função de Sobrevivência de Kaplain Meier por Célula



Analisando o gráfico acima por célula, é possível notar que o tipo de célula que possui uma probabilidade de sobrevivência maior em relação às outras é o tipo de célula escamosa, pois esta probabilidade só zera ao final de 1000 dias de tratamento. Por outro lado, o paciente que possui o tipo de célula adeno tem uma probabilidade de sobrevivência muito baixa em relação aos outros, pois esta probabilidade zera ao

final de 200 dias de tratamento. Logo, os pacientes que possuem esta célula têm alta probabilidade de morte em relação aos demais. Em suma, parece haver dois grupos distintos, formados pelos tumores escamoso e grande no primeiro e tumores pequenos e adeno no segundo.

Figura 11: TTT por Célula



Para identificar a distribuição de probabilidade por célula observa-se os gráficos TTT acima. Analisando-os é possível identificar que para as células escamosa e pequena, o gráfico possui uma curva convexa, ou seja, monotonicamente decrescente. Enquanto para células do tipo adeno e grande, a curva é côncava, ou seja, monotonicamente crescente. Em ambas as situações a distribuição Weibull é sugerida pelos gráficos. É importante salientar que essa distribuição foi sugerida também pelos gráficos anteriores, porém nesta atual análise que houve uma divisão maior das variáveis, esta foi a única distribuição sugerida em todos os gráficos TTT.

A seguir será feito um teste para testar se há diferenças entre as curvas de sobrevivência da variável célula. As curvas se cruzam ao longo do gráfico de função de sobrevivência de Kaplain Meier. Logo, o teste utilizado é o de Wilcoxon ao nível de significância de 10% , com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1 : \text{Existe diferença entre as curvas de sobrevivência} \end{cases}$$

Quadro 3: Resultado do teste de Wilcoxon

P-valor	**Decisão do teste**
<0,001	Rejeita H_0

Com base nos resultados do quadro acima, rejeita-se a hipótese nula. Portanto, há diferença entre as curvas de sobrevivência dados os tipos de células.

A seguir será realizado o Teste de Resíduos de Schoenfeld para testar se os riscos são proporcionais entre os tipos de células cancerígenas. Logo, estas são as hipóteses ao nível de 10% de significância:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

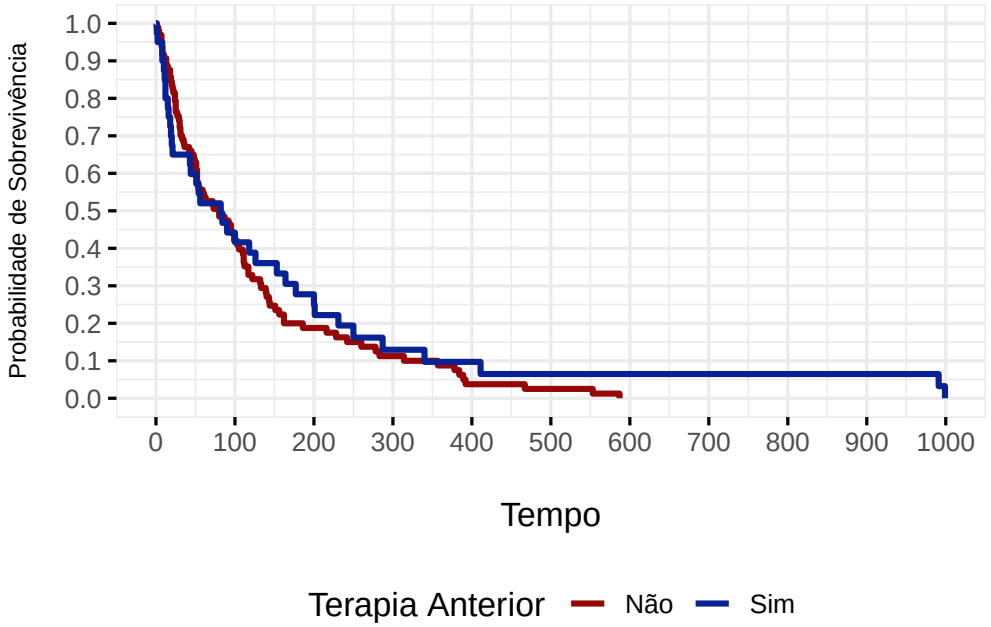
Quadro 4: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
0,031	Rejeita H_0

Os riscos não são proporcionais, o que indica novamente que não é possível o uso do modelo de Cox.

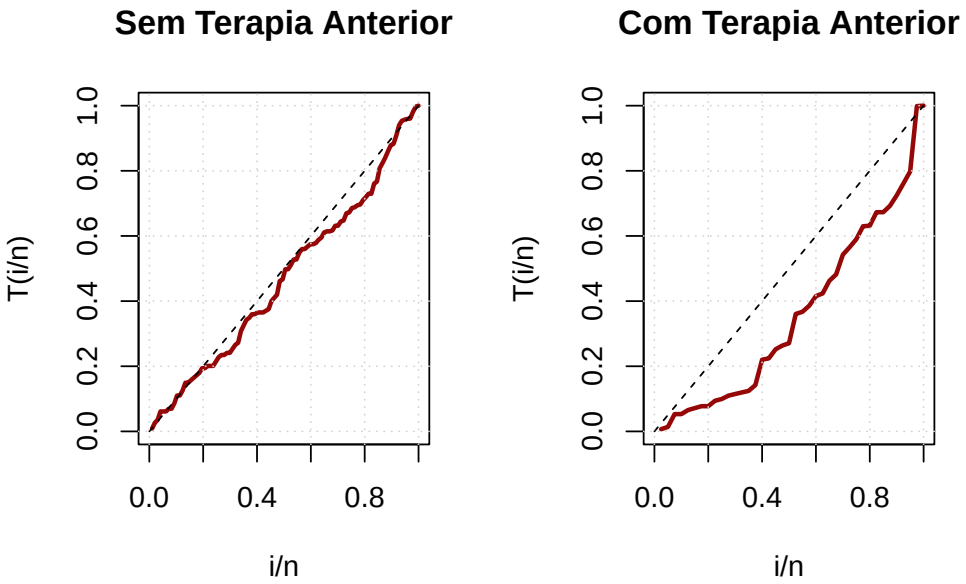
3.2.4 Modelo de sobrevivência por Terapia Anterior

Figura 12: Função de Sobrevivência de Kaplain Meier por Terapia Anterior



Pela função de sobrevivência indicada pelo gráfico por Terapia Anterior, nota-se claramente como que o paciente que fez terapia anterior não possui uma probabilidade muito maior de sobrevivênica do que o paciente que não fez. Como os pacientes possuem câncer em estágio avançado e inoperável, eventuais terapias anteriores não auxiliaram no fator de cura, logo, possuem efeito próximo à não terapia.

Figura 13: TTT por Terapia Anterior



Em ambos os gráficos a curva é convexa e indica uma possível modelagem pela distribuição Weibull.

A seguir será feito um teste para testar se há diferenças entre as curvas de sobrevivência da variável terapia anterior. As curvas se cruzam ao longo do gráfico de função de sobrevivência de Kaplain Meier. Logo, o teste utilizado é o de Wilcoxon ao nível de significância de 10% , com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1 : \text{Existe diferença entre as curvas de sobrevivência} \end{cases}$$

Quadro 5: Resultado do teste de Wilcoxon

P-valor	**Decisão do teste**
0,8	Não Rejeita H_0

Com base nos resultados do quadro acima, a hipótese nula não é rejeitada. Logo, não há diferença entre as curvas de sobrevivência dados a Terapia Anterior.

A seguir será realizado o Teste de Resíduos de Schoenfeld para testar se os riscos são proporcionais entre os grupos de terapia anterior. Logo, estas são as hipóteses ao nível de 10% de significância:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

Quadro 6: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
0,083	Rejeita H_0

Com base no pvalor a hipótese nula é rejeitada e portanto, os riscos não são proporcionais.

3.2.5 Modelo de sobrevivência por Score

Para avaliar o uso do modelo paramétrico ou modelo de Cox foi feito o teste de Resíduos de Schoenfeld para a variável quantitativa Score. Abaixo estão são as hipóteses ao nível de 10% de significância:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

Quadro 7: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
<0,001	Rejeita H_0

Os riscos não são proporcionais para o modelo envolvendo o Score.

3.2.6 Modelo de sobrevivência por Meses de diagnóstico

Agora a variável em questão é a quantidade de meses desde o diagnóstico do paciente. Para esta situação, foi feito o modelo de Cox e o teste para a avaliação dos resíduos de Schoenfeld com nível de significância de 10%:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

Quadro 8: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
0,91	Não Rejeita H_0

Neste caso, os riscos são proporcionais, sendo este um pressuposto atendido para o modelo de Cox.

3.2.7 Modelo de sobrevivência por Idade

Será realizado novamente o teste Teste de Resíduos de Schoenfeld para a avaliação dos riscos proporcionais do modelo com a variável Idade do paciente ao nível de significância de 10%:

$$\begin{cases} H_0 : \text{Os riscos são proporcionais} \\ H_1 : \text{Os riscos não são proporcionais} \end{cases}$$

Quadro 9: Resultado do teste de Schoenfeld

P-valor	**Decisão do teste**
0,2	Não Rejeita H_0

Assim como na seção anterior, aqui os riscos são proporcionais. Porém, analisando os resultados anteriores envolvendo as outras variáveis, o resultado mais visto foi de que os riscos não são proporcionais. Logo, o modelo de Cox não é o ideal para este estudo e sim o modelo paramétrico, no qual atribuimos uma distribuição de probabilidade para modelar o tempo de sobrevivência. A principal distribuição de probabilidade candidata é a Weibull, entretanto, é pertinente testar as distribuições log-normal e log-logística.

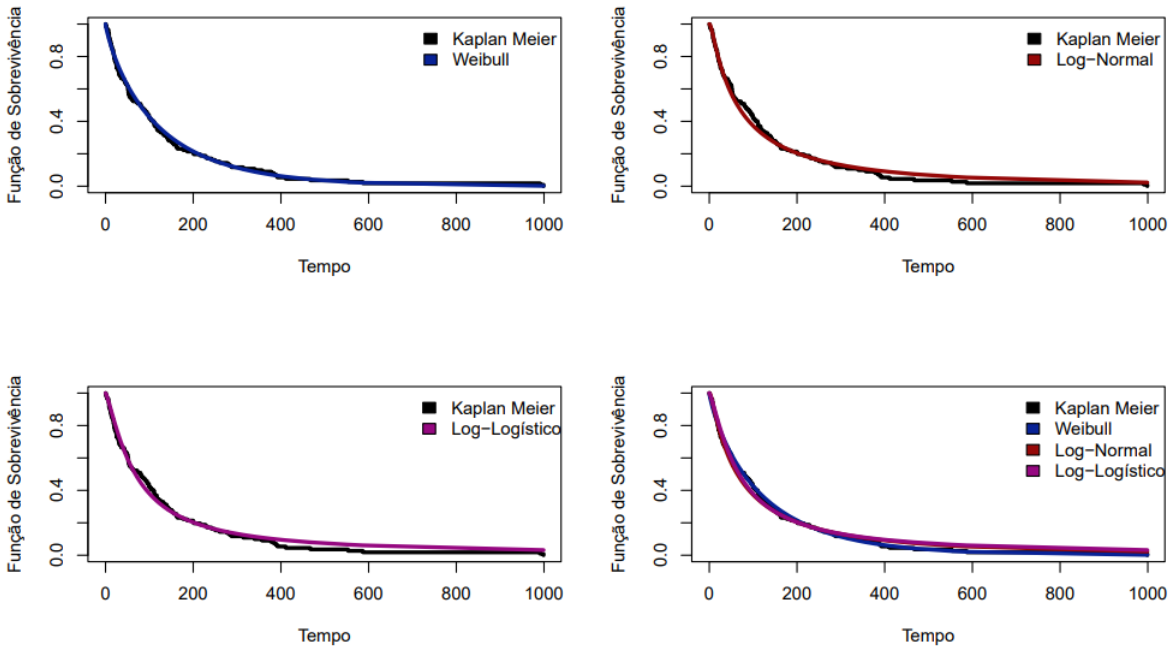
3.3 Ajuste e seleção do modelo probabilístico

3.3.1 Modelo paramétrico

Após a análise descritiva vimos que um modelo de riscos proporcionais de Cox não é adequado para os dados sobre câncer de pulmão. Nesse sentido, o ajuste de um modelo de regressão paramétrico faz mais sentido. A próxima etapa consistiu na comparação entre distribuições de probabilidade que nos possibilitam modelar o tempo de sobrevivência. O candidato mais forte para a modelagem foi a distribuição Weibull, entretanto, também testamos as distribuições log-normal e log-logística.

Baseando-se no método gráfico de comparação da função de sobrevivência do modelo proposto com a função de sobrevivência estimada por Kaplan-Meier, temos a seguinte configuração:

Figura 14: Comparação entre distribuições de probabilidade e Kaplan-Meie



É possível verificar na figura acima cada um dos ajustes entre as curvas das distribuições de probabilidade e a curva de Kaplan-Meier. Em suma, a Weibull apresenta melhores resultados, visto que sua curva é extremamente próxima à curva de sobrevivência empírica. Por outro lado, as distribuições Log-normal e Log-logística também apresentam resultados satisfatórios, entretanto, existe um afastamento em relação à curva de Kaplan Meier em alguns trechos, principalmente após o tempo de 400 dias.

Para verificarmos se a distribuição Weibull realmente apresenta um melhor ajuste, podemos calcular algumas medidas de parcimônia como o Critério de Akaike (AIC),

Critério de Akaike Corrigido (AICc) e o Criterio de Informação Bayesiano (BIC). Vejamos os resultados abaixo:

Tabela 6: Critérios para classificação e seleção de modelos - AIC, AICc e BIC

Modelo	**AIC**	**AICc**	**BIC**
Weibull	1500,18	1500,27	1506,02
Log-Normal	1502,95	1503,04	1508.79
Log-Logística	1504,53	1504,62	1510,37

Como nas três medidas os menores valores correspondem à distribuição Weibull, observa-se excelente adequabilidade do ajuste de um modelo regressão no qual o tempo de sobrevivência segue uma distribuição Weibull. Cabe observar que os parâmetros de forma γ e de escala α foram estimados e as estimativas encontradas foram, respectivamente, $\hat{\gamma} = 0,852$ e $\hat{\alpha} = 120,680$. Novamente, obtivemos resultados que coincidem com a análise descritiva realizada previamente, uma vez que a distribuição Weibull surgiu como a principal candidata para a modelagem dos dados e os gráficos indicaram predominantemente uma função de risco decrescente, com um parâmetro de forma menor que 1.

3.3.2 Seleção do modelo final

Inicialmente, ajustou-se 6 modelos. A ideia foi verificar individualmente o efeito de cada uma das variáveis em relação ao tempo de sobrevivência dos pacientes com câncer de pulmão inoperável. Os resultados iniciais mostraram que as variáveis *Célula* (indica o tipo de tumor) e *Score* (indica a condição do paciente) foram consideradas significativas. Por outro lado, as variáveis *Tratamento* (nível de referência "Padrão"), *Meses de diagnóstico*, *Idade do paciente* e *Terapia Anterior* (nível de referência "Não") não apresentaram significância estatística. Vejamos os principais resultados abaixo:

Tabela 7: Modelo 1 - Efeito individual da variável *Tratamento*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	4,769	0,147	32,25	<0,001
Trat. Teste	0,047	0,207	0,23	0,818

Tabela 8: Modelo 2 - Efeito individual da variável *Célula*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	5,406	0,188	28,78	<0,001
Pequena	-1,083	0,240	-4,50	<0,001
Adeno	-1,216	0,275	-4,43	<0,001
Grande	-0,263	0,275	-0,96	0,34

Tabela 9: Modelo 3 - Efeito individual da variável *Score*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	2,645	0,295	8,95	<0,001
Score	0.035	0,005	7,26	<0,001

Tabela 10: Modelo 4 - Efeito individual da variável *Meses de diagnóstico*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	4,884	0,143	34,15	<0,001
Meses diag	-0,010	0,010	-0,99	0,321

Tabela 11: Modelo 5 - Efeito individual da variável *Idade*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	5,489	0,662	8,30	<0,001
Idade	-0,012	0,011	-1,07	0,284

Tabela 12: Modelo 6 - Efeito individual da variável *Terapia anterior*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	4,727	0,124	38,06	<0,001
Terapia "Sim"	0,229	0,227	1,01	0,312

O próximo passo foi testar modelos mais complexos, nos quais fomos introduzindo pares de variáveis, grupos de variáveis e assim por diante até a testagem do modelo completo com todas as variáveis do estudo. Em suma, o melhor modelo condiz com a análise individual acima, ou seja, o modelo de regressão encontrado envolve as variáveis *Célula* e *Score* conjuntamente.

Tabela 13: Candidato 1 - Modelo com *Célula* + *Score*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	3,480	0,340	10,22	<0,001
Pequena	-0,708	0,226	-3,13	0,002
Adeno	-1,108	0,253	-4,39	<0,001
Grande	-0,322	0,250	-1,29	0,198
Score	0,029	0,005	6,31	<0,001

No modelo acima percebemos que a variável *Score* é extremamente significativa, ou seja, o grau de desempenho e condição do paciente exerce influência sobre o tempo de sobrevivência dos indivíduos com câncer de pulmão. Quanto maior o score, maiores a chance de sobrevivência, uma vez que o nível de assistência necessário não será tão intenso (de totalmente hospitalizado para sem necessidade de hospitalização). Com relação à variável célula, vemos que existe diferença entre as células escamosa ("referência") e os tumores do tipo pequeno e adeno. Por outro lado, a diferença constatada não é significativa entre células escamosas e grandes.

Ao realizar o teste da razão de verossimilhança, observou-se que este modelo é preferível em relação aos modelos mais simples 2 e 3. No primeiro caso, a estatística de teste foi de $35,28$ com 1 grau de liberdade e $pvalor < 0,001$, ou seja, rejeitamos a hipótese nula e ficamos com o modelo candidato 1 em relação ao modelo 2. No segundo caso, a estatística de teste foi de $19,04$ com 3 graus de liberdade e $pvalor < 0,001$, ou seja, rejeitamos a hipótese nula e ficamos com o modelo candidato 1 em relação ao modelo 3.

O modelo acima já nos retorna informações importantes sobre o estudo, entretanto, se fez necessário testar outros modelos com interações entre variáveis. A concepção é verificar se o efeito de uma variável se altera a depender do nível ou categoria de uma outra variável. Apesar da diferença entre os tratamentos tradicional e de teste não ter sido considerada significativa, verificamos que existe certa

interação entre o tratamento e o tipo de tumor. Nesse sentido, o segundo modelo candidato para a análise é apresentado a seguir:

Tabela 14: Candidato 2 - Modelo com *Célula*, *Score*, *Tratamento* e Interação

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	3,352	0,347	9,64	<0,001
Pequena	-0,306	0,298	-1,02	0,305
Adeno	-1,091	0,392	-2,79	0,005
Grande	0,043	0,347	0,13	0,900
Score	0,028	0,004	6,26	<0,001
Trat. Teste	0,317	0,331	0,96	0,338
Teste*Pequena	-1,020	0,432	-2,36	0,018
Teste*Adeno	-0,076	0,513	-0,15	0,882
Teste*Grande	-0,756	0,481	-1,57	0,116

Os dois modelos candidatos apresentados acima apresentam resultados extremamente próximos. Quando realizado o teste da razão de verossimilhança, considerando que temos modelos aninhados, a estatística de teste foi de $8,29$ com 4 graus de liberdade e $pvalor = 0,08$, ou seja, rejeitamos a hipótese nula com um nível de significância de 10% e ficamos com o modelo candidato 2 em relação ao modelo candidato 1. Por outro lado, se calculamos algumas medidas de parcimônia como o Critério de Akaike (AIC), Critério de Akaike Corrigido (AICc) e o Criterio de Informação Bayesiano (BIC), observamos que os dois modelos são semelhantes, entretanto, o AICc e o BIC são menores no modelo candidato 1.

Tabela 15: Critérios para classificação e seleção de modelos - AIC, AICc e BIC

Modelo	**AIC**	**AICc**	**BIC**
Candidato 1	1445,03	1445,67	1462,55
Candidato 2	1444,73	1446,47	1473.93

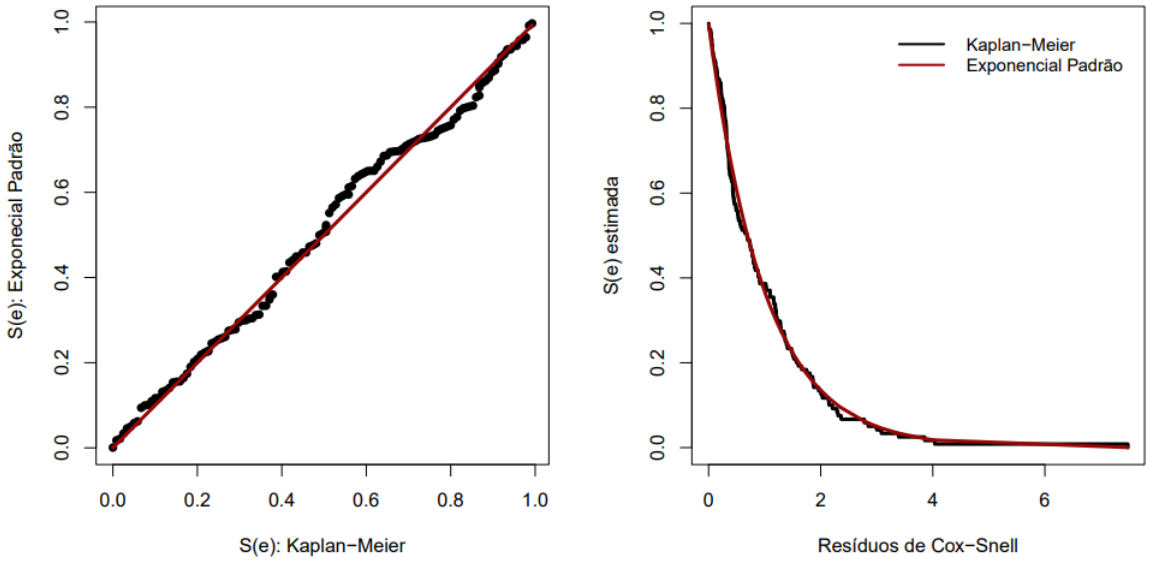
Em síntese, os dois modelos são válidos. A ação mais adequada é realizar uma pesquisa mais aprofundada sobre as variáveis do estudo e consultar especialistas na área caso tenhamos que escolher entre um modelo ou outro. Em ambos os casos, a análise de resíduos também apresenta semelhança entre os candidatos, ou seja, não conseguimos distinguir qual dos modelos apresenta melhor ajuste pelos critérios da análise dos resíduos.

Apesar do exposto sobre a semelhança entre os modelos candidatos, optou-se neste trabalho pelo modelo candidato 2. A justificativa é que este modelo é mais completo e nos permite introduzir a variável *tratamento*. Como estamos lidando com pacientes que não podem ser operados, o risco de morte é alto e um tratamento

está sendo testado para tentar amenizar tal problema, a introdução de uma variável que compara os tratamentos padrão e de teste é essencial para o estudo. Nessa perspectiva, estabelecer uma diferença entre tratamentos é um dos principais objetivos da Administração de Veteranos.

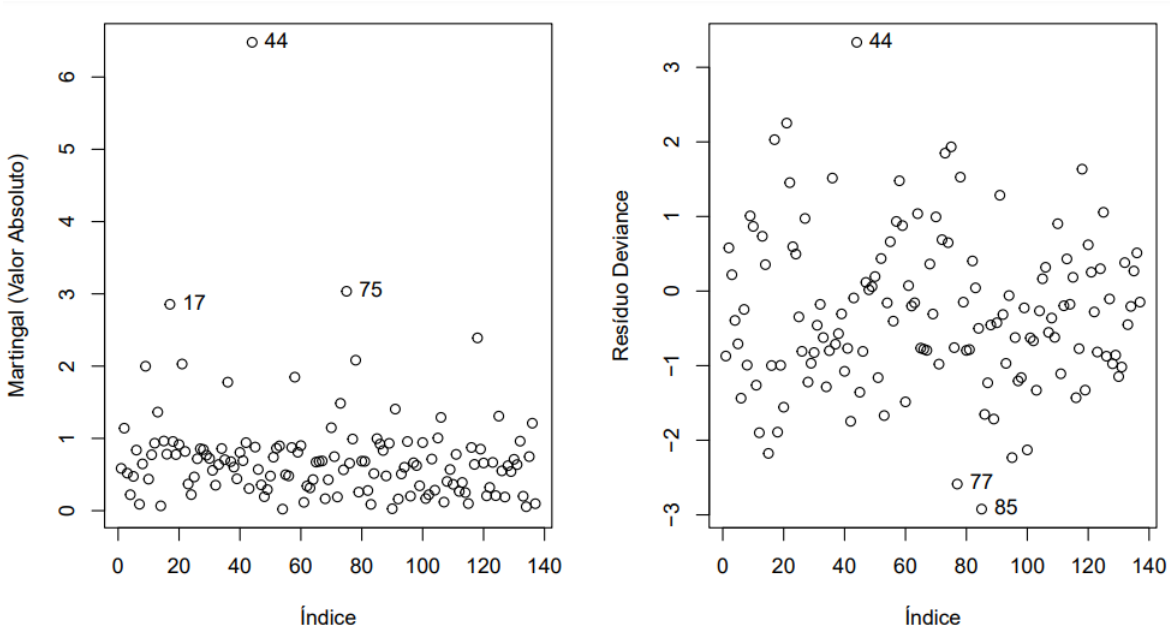
Para verificar a qualidade do ajuste global do modelo nos baseamos inicialmente nos resíduos de Cox-Snell. Na figura abaixo, podemos verificar no gráfico à esquerda que os resíduos de Cox-Snell aparentam seguir a distribuição exponencial padrão, visto que existe uma tendência linear significativa quando comparamos as sobrevivências estimadas por Kaplan-Meier e pela exponencial padrão. No gráfico à direita temos as as curvas de sobrevivência dos resíduos estimadas por Kaplan-Meier e a exponencial padrão. É perceptível que o ajuste do modelo pode ser considerado excelente, ou seja, não há indicíos para rejeição do modelo de regressão selecionado.

Figura 15: Sobrevivências dos resíduos de Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo exponencial padrão (gráfico à esquerda) e respectivas curvas de sobrevivência estimadas (gráfico à direita).



Em seguida verificamos os resíduos de Martingal e Deviance. É possível observar que existe um padrão aleatório dos resíduos em torno do zero quando comparamos os resíduos e o índice (ordem) das observações, o que indica um bom ajuste do modelo em relação aos dados. Apesar disso, algumas observações aparentaram ser pontos de influência, sobretudo, a observação de número 44 que apresenta resíduos de Martingal e Deviance claramente discrepantes.

Figura 16: Residuos de Martingal e Deviance em relação ao índice das observações



Como obtivemos um bom ajuste do modelo quando observamos os resíduos de Cox-Snell, a existência de pontos influentes não parece afetar significativamente os resultados encontrados. Ao retirarmos tais observações e realizar um novo ajuste, não observou-se diferença considerável entre os coeficientes de regressão estimados, os sinais dos coeficientes e a interpretação do que é ou não significativo. Além disso, a análise dos resíduos de Cox-Snell também não foi alterada. Portanto, não se faz necessário repensar o modelo por causa de possíveis medidas influentes, já que os resultados são mantidos com ou sem a inclusão de tais observações. Além disso, cabe ressaltar que a relação funcional encontrada aparenta ser linear, ou seja, não há indícios da necessidade de transformação nas covariáveis sob investigação.

Portanto, os resultados encontrados pela modelagem dos dados são similares as hipóteses levantadas na análise descritiva. A diferença está apenas na variável de tratamento, uma vez que esta não apresenta significância estatística quando observada individualmente, entretanto, esse cenário muda quando analisamos tal variável na presença do tipo de tumor. No próximo tópico finalizamos os resultados com a interpretação do modelo final obtido.

3.3.3 Interpretação dos resultados

Para a correta interpretação dos dados, precisamos considerar que no modelo Weibull precisamos calcular a exponencial de cada um dos coeficientes para encontrarmos a razão dos tempos medianos de sobrevivência. Além da estimativa pontual, vamos considerar os intervalos de confiança. Ademais, temos a presença de variáveis categóricas e um fator de interação, o que nos levará a submodelos que serão interpretados separadamente. Recapitulando o modelo final escolhido, temos:

Tabela 16: Modelo Final com *Célula*, *Score*, *Tratamento* e Interação

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	3,352	0,347	9,64	<0,001
Pequena	-0,306	0,298	-1,02	0,305
Adeno	-1,091	0,392	-2,79	0,005
Grande	0,043	0,347	0,13	0,900
Score	0,028	0,004	6,26	<0,001
Trat. Teste	0,317	0,331	0,96	0,338
Teste*Pequena	-1,020	0,432	-2,36	0,018
Teste*Adeno	-0,076	0,513	-0,15	0,882
Teste*Grande	-0,756	0,481	-1,57	0,116

O preditor linear do modelo pode ser escrito da seguinte maneira: $3,35+0,03X_1-0,31X_{21}-1,09X_{22}+0,04X_{23}+0,32X_3-1,02X_{21}X_3-0,08X_{22}X_3-0,76X_{23}X_3$

- X_1 representa a variável *score*;
- X_{21}, X_{22} e X_{23} representam variáveis binárias para as células do tipo pequena, adeno e grande. Quando todas são iguais a zero temos o submodelo da célula escamosa;
- X_3 representa a variável tratamento. Quando o valor é igual a 1 temos o tratamento teste, quando for igual a 0 o modelo refere-se ao tratamento padrão.

O intercepto não nos retorna uma interpretação válida se considerado de forma isolado, uma vez que o nível da variável score nunca será igual a zero. Por outro lado, temos que $\exp^{0,028} = 1,03$, ou seja, para cada aumento de uma unidade no score do paciente o tempo mediano de sobrevivência aumenta $1,03$ vezes, mantidos tratamento e tipo de tumor contantes. O intervalo com 95% de confiança para essa odds varia entre $1,020$ e $1,036$. A consequência é que a variável *score* desacelera o decaimento da curva de sobrevivência, ou seja, quanto maior o score maior é o tempo e a probabilidade de sobrevivência do paciente. Essa interpretação faz bastante sentido, uma vez que scores baixos indicam indivíduos hospitalizados. Por

sua vez, scores mais altos indicam indivíduos, geralmente, em situação levemente melhor sem necessidade de internação imediata, embora ainda em condição de câncer inoperável.

A cada 10 unidades aumentadas no score a chance de sobrevivência, medida pela razão entre os tempos medianos de sobrevivência, é 10,3 vezes maior. Na comparação entre pacientes totalmente hospitalizados e aqueles que podem cuidar de si mesmos, essa chance pode ser 82,4 vezes maior para pacientes independentes em relação aos internados, por exemplo. Para um indivíduo com tratamento padrão, tumor de célula escamosa e score igual a 20, o tempo médio de sobrevivência é de aproximadamente 50 dias. Quando o score é igual a 50 o tempo médio de sobrevivência é de 116 dias e para score igual a 90 o tempo médio de sobrevivência é de 355 dias.

Com relação aos demais coeficientes do modelo, iniciamos a avaliação a partir dos coeficientes considerados significantes na regressão em questão. O tempo mediano de sobrevivência dos pacientes com câncer adeno é quase um terço se comparado com os pacientes com tumores do tipo escamoso ($\exp^{-1,091} = 0,33$), mantido o restante constante. Com 95% de confiança essa redução no tempo mediano varia entre $0,15$ e $0,72$. O tempo mediano de sobrevivência dos pacientes com câncer pequeno e submetido ao tratamento teste é quase um terço se comparado com os pacientes com tumores do tipo escamoso e tratamento padrão ($\exp^{-1,02} = 0,36$), mantido o restante constante. Com 95% de confiança essa redução no tempo mediano varia entre $0,15$ e $0,84$. Essa análise é importante, entretanto, vamos considerar alguns submodelos possíveis para que a análise seja melhor detalhada e o nível de informações que temos em mãos seja ampliado. Nesse sentido, as comparações podem ser realizadas não só em relação aos níveis de referência, mas em relação a todas as combinações possíveis.

- Célula escamosa e tratamento padrão: $3,35 + 0,03X_1$
- Célula pequena e tratamento padrão: $3,04 + 0,03X_1$
- Célula adeno e tratamento padrão: $2,26 + 0,03X_1$
- Célula grande e tratamento padrão: $3,39 + 0,03X_1$
- Célula escamosa e tratamento teste: $3,67 + 0,03X_1$
- Célula pequena e tratamento teste: $2,36 + 0,03X_1$
- Célula adeno e tratamento teste: $2,51 + 0,03X_1$
- Célula grande e tratamento teste: $3,32 + 0,03X_1$

No caso do tratamento padrão, vemos que existe diferença significativa entre as células adeno e as demais, sobretudo, em relação às células escamosa e grande. Para

os 4 tipos de tumor e um score de 50, por exemplo, o tempo médio de sobrevivência é de 128, 94, 43 e 133 dias, considerando as células escamosa, pequena, adeno e grande, respectivamente.

No caso do tratamento teste, vemos que existe diferença significativa entre as células pequena e as demais, sobretudo, em relação às células escamosa e grande. Para os 4 tipos de tumor e um score de 50, por exemplo, o tempo médio de sobrevivência é de 176, 48, 55 e 124 dias, considerando as células escamosa, pequena, adeno e grande, respectivamente. Resumidamente, podemos estabelecer as seguintes conclusões sobre os resultados:

- Existem diferenças significativas entre tratamentos quando comparamos os tipos de tumores. O tratamento teste afeta negativamente o processo quando a célula cancerígena é do tipo pequena. Existe uma redução considerável no tempo médio de sobrevivência nesse caso, o que acarreta na redução da probabilidade de sobrevivência dos pacientes. Embora os efeitos não tenham sidos considerados significativos, o tratamento teste apresentou melhores resultados do que o tratamento padrão em células escamosas e do tipo adeno.

- Existem diferenças significativas entre os tipos de células. As células pequena e adeno afetam negativamente quando o tratamento é padrão ou de teste. Existe uma redução considerável no tempo médio de sobrevivência nesses casos se comparado com as células grande e escamosa, o que acarreta na redução da probabilidade de sobrevivência dos pacientes.

- Não foi identificado diferença entre as células escamosa e grande, independentemente do tratamento utilizado.

- A célula grande apresenta resultados semelhantes tanto no tratamento padrão quanto no tratamento de teste.

4 Considerações Finais

Neste estudo houve o aprofundamento no entendimento destes dados e em como o tratamento e outras variáveis da base de dados influenciam no resultado final que é o tempo de sobrevivência do paciente com câncer de pulmão. Como foi observado nas seções anteriores, dois modelos eram válidos e se ajustavam bem aos dados, entretanto, como o objetivo do estudo é comparar efeito entre os tratamentos e os pacientes estão com câncer em estágio avançado e inoperável, a inclusão da variável tratamento faz-se necessária para o entendimento do impacto da quimioterapia teste para esta doença. Portanto, o segundo modelo foi escolhido.

As variáveis que são significativas para o resultado do estudo e que compõem o modelo final (que foi selecionado após uma série de processos) são: Score, Célula, Tipo de Tratamento e a interação entre Célula e Tratamento.

Temos que $\exp^{0,028} = 1,03$, ou seja, para cada aumento de uma unidade no score do paciente o tempo mediano de sobrevivência aumenta $1,03$ vezes, mantidos tratamento e tipo de tumor contantes. Também podemos interpretar esse valor como uma razão de chance de sobrevivência. O tempo mediano de sobrevivência dos pacientes com câncer adeno é quase um terço se comparado com os pacientes com tumores do tipo escamoso ($\exp^{-1,091} = 0,33$), mantido o restante constante. Por sua vez, o tempo mediano de sobrevivência dos pacientes com câncer pequeno e submetido ao tratamento teste é quase um terço se comparado com os pacientes com tumores do tipo escamoso e tratamento padrão ($\exp^{-1,02} = 0,36$), mantido o restante constante.

Existem diferenças significativas entre tratamentos quando comparamos os tipos de tumores. O tratamento teste afeta negativamente o processo quando a célula cancerígena é do tipo pequena. Existem diferenças significativas entre os tipos de células. As células pequena e adeno afetam negativamente quando o tratamento é padrão ou de teste, respectivamente. Existe uma redução considerável no tempo médio de sobrevivência nesses casos se comparado com as células grande e escamosa. Não foi identificado diferença entre as células escamosa e grande, independentemente do tratamento utilizado.

Obteve-se as seguintes interpretações do modelo para os tipos de tratamento:

Tratamento padrão: Para os 4 tipos de tumor e um score de 50, por exemplo, o tempo médio de sobrevivência é de 128, 94, 43 e 133 dias, considerando as células escamosa, pequena, adeno e grande, respectivamente.

Tratamento Teste: Para os 4 tipos de tumor e um score de 50, por exemplo, o tempo médio de sobrevivência é de 176, 48, 55 e 124 dias, considerando as células escamosa, pequena, adeno e grande, respectivamente.

Referências

CARVALHO, M. S. et al. *Análise de sobrevivência: teoria e aplicações em saúde*. [S.l.]: SciELO-Editora FIOCRUZ, 2011.

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

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2008.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival analysis : techniques for censored and truncated data*. [S.l.]: New York: Springer, 2003.

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

MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. 6° edição. ed. [S.l.]: Saraiva, 2010.

PRENTICE, R. L. et al. The analysis of failure times in the presence of competing risks. *Biometrics*, JSTOR, p. 541–554, 1978.

SAIKIA, R.; BARMAN, M. P. A review on accelerated failure time models. *International Journal of Statistics and Systems*, v. 12, n. 2, p. 311–322, 2017.

5 Anexo

Link para download do código em R e dados do estudo: Arquivos.