UNIVERSIDADE ESTADUAL DA PARAÍBA - UEPB

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

CURSO DE BACHARELADO EM ESTATÍSTICA

RELATORIO: ANÁLISE DE SOBREVIVÊNCIA – COMPARAÇÃO ENTRE O MODELO DE REGRESSÃO EXPONENCIAL E WEIBULL PARA PACIENTES COM CÂNCER DE PULMÃO

Lucas Manoel Batista de Albuquerque
Cleanderson Romualdo Fidelis

Análise de Sobrevivência

2023

1 INTRODUÇÃO

O câncer de pulmão é uma das principais causas de morte por câncer em todo o mundo, afetando significativamente a qualidade de vida e a sobrevida dos pacientes. A análise de sobrevida é uma ferramenta estatística amplamente utilizada para estudar o tempo de sobrevida de pacientes com câncer, estimar a taxa de sobrevida e identificar fatores prognósticos. Neste relatório, utilizaremos dados provenientes do *North Central Cancer Treatment Group* (NCCTG) para realizar uma análise de sobrevivência abrangente do câncer de pulmão. Nosso objetivo principal é investigar a sobrevida dos pacientes e explorar a influência de diferentes fatores no tempo de sobrevida.

Para tanto, empregaremos o estimador de Kaplan-Meier, método não paramétrico amplamente utilizado para estimar funções de sobrevivência em análises de sobrevivência. Este estimador é particularmente adequado para lidar com dados de tempo de sobrevivência censurados, onde o evento de interesse (como morte) não foi observado para todos os indivíduos do estudo.

Além disso, exploraremos o uso de modelos probabilísticos para descrever a distribuição dos tempos de sobrevivência. Neste relatório, consideraremos os modelos Exponencial e Weibull. Esses modelos fornecem uma abordagem paramétrica para analisar a sobrevivência, permitindo que parâmetros sejam estimados e informações sobre a forma da função de sobrevivência sejam obtidas.

Comparando os resultados dos modelos probabilísticos através de gráficos e do teste da razão de verossimilhança, poderemos determinar qual modelo é o mais adequado para os dados de câncer de pulmão. Esta informação é fundamental para entender a distribuição dos tempos de sobrevivência, ajudando a prever a sobrevivência em pacientes com câncer de pulmão e identificando fatores de risco associados.

2 METODOLOGIA

2.1 ESTIMADOR DE KAPLAN MEIER

O estimador de Kaplan-Meier é uma técnica estatística não-paramétrica que é comumente utilizada para estimar a função de sobrevivência em estudos de sobrevivência. Esse estimador leva em conta dados censurados, ou seja, dados que não apresentam informações completas sobre a duração do evento de interesse. O estimador de Kaplan-Meier leva em conta esses dados censurados e é capaz de produzir uma curva de sobrevivência que mostra a proporção de indivíduos sobreviventes em diferentes momentos de tempo.

Construção do Estimador de Kaplan-Meier:

- Ordenar os tempos distintos de falha:

$$t_1 < t_2 < \cdots < t_k$$

- Utilizando a seguinte notação:

  - d_i: número de falhas no tempo t_i;
  - n_i: número de observações sob risco (não falhou e não foi censurado) até o tempo t_i (exclusivo);

O estimador de Kaplan-Meier é:

$$\hat{S}(t) = \prod_{\substack{i \\ t_j < t}} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{\substack{i \\ t_j < t}} \left(1 - \frac{d_i}{n_i} \right)$$

É definida como a probabilidade de uma observação não falhar até certo instante t, ou seja, a probabilidade de esta sobreviver ao tempo t.

O estimador de Kaplan-Meier será usado com o objetivo principal de ser uma base de comparação para a qualidade dos ajustes dos modelos probabilísticos.

2.2 TESTE LOG-RANK E LOG-RANK GENERALIZADO

O teste log-rank e log-rank generalizado é um teste estatístico utilizado na análise de sobrevivência para comparar as curvas de sobrevida entre dois ou mais grupos independentes, respectivamente.

Ambos os testes calculam uma estatística de teste que segue aproximadamente uma distribuição qui-quadrado, permitindo determinar se as diferenças observadas entre as curvas de sobrevida são estatisticamente significativas. Para o teste log-rank essa distribuição qui-quadrado possui 1 grau de liberdade e a do teste log-rank generalizado possui $r - 1$ graus de liberdade, onde r é a quantidade de curvas de sobrevivência a serem comparados.

A hipótese nula do teste log-rank é de que não há diferenças entre as curvas de sobrevida dos dois grupos comparados, ou seja, as taxas de eventos são iguais em ambos os grupos ao longo do tempo. A hipótese alternativa, por sua vez, é de que existe pelo menos uma diferença significativa nas taxas de eventos entre os grupos. Tais hipóteses são representadas como:

$$\begin{aligned} H_0 &: S_i(t) = S_j(t) \\ & \qquad \forall\, i \neq j \\ H_1 &: S_i(t) \neq S_j(t) \end{aligned} \tag{1}$$

Para todo t no período de acompanhamento.

2.3 MODELOS PROBABILÍSTICOS

Serão comparados dois modelos probabilísticos de regressão através do teste da razão de Verossimilhança. Tais modelos são: o modelo de regressão Exponencial e o modelo de regressão Weibull.

2.3.1 DISTRIBUIÇÃO EXPONENCIAL

A função de densidade de probabilidade para a variável aleatória tempo de falha T com distribuição exponencial é dada por:

$$f(t) = \frac{1}{\alpha} exp\left\{-\frac{t}{\alpha}\right\} \quad t \geq 0 \tag{1}$$

em que o parâmetro $\alpha \geq 0$ é o tempo médio de vida. O parâmetro α tem a mesma unidade do tempo de falha t. Isto é, se t é medido em horas, α também será medido em horas.

Ainda, a função de sobrevivência $S(t)$ e para o modelo de regressão exponencial é dada por

$$S(t|x) = exp\left\{-\left(\frac{\mathrm{t}}{\exp(\mathbf{x\beta})}\right)\right\} \tag{2}$$

sendo $\boldsymbol{\beta}$ as estimativas dos efeitos das covariáveis e $\boldsymbol{x}$ o valor das covariáveis

2.3.2 DISTRIBUIÇÃO DE WEIBULL

Para uma variável aleatória T com distribuição de Weibull tem-se a função de densidade de probabilidade dada por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, t \geq 0 \tag{3}$$

em que γ, o parâmetro de forma, e α, o de escala, são ambos positivos. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade.

Para esta distribuição, a função de sobrevivência é

$$S(t|x) = exp\left\{-\left(\frac{t}{exp(\boldsymbol{x\beta})}\right)^{\gamma}\right\} \tag{4}$$

sendo $\boldsymbol{\beta}$ as estimativas dos efeitos das covariáveis e $\boldsymbol{x}$ o valor das covariáveis. É possível notar que quando γ é igual a 1, temos a distribuição exponencial. Assim, o teste da razão de Verossimilhança não só irá ser útil para testar a significância das covariáveis no modelo como também irá nos indicar se o parâmetro γ é significativamente diferente de 1.

2.4 MÉTODO DA LINEARIZAÇÃO PARA ESCOLHA DO MODELO

Este método na linearização da função de sobrevivência tendo como ideia básica a construção de gráficos que sejam aproximadamente lineares caso o modelo proposto seja apropriado. Violações da linearidade podem ser rapidamente verificadas visualmente COLOSIMO (2006)

A seguir são apresentados exemplos de linearização para os modelos exponencial e Weibull.

- Linearização do modelo exponencial

Para o modelo exponencial, a função de sobrevivência sem nenhuma covariável é dada por:

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

Assim,

$$-log[S(t)] = \frac{t}{\alpha} = \left(\frac{1}{\alpha}\right)t$$

o que mostra que $-log[S(t)]$ é uma função linear de t. Logo, o gráfico de $-log\left[\hat{S}(t)\right]$ versus t deve ser aproximadamente linear, passando pela origem, se o modelo exponencial for apropriado. $\hat{S}(t)$ é o estimador de Kaplan-Meier.

- Linearização do modelo de Weibull

A função de sobrevivência para o modelo Weibull de parâmetros (α, γ) sem o uso de covariáveis é dada por:

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \quad t \geq 0$$

Desse modo,

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

$$log\left[-log[S(t)]\right] = -\gamma log(\alpha) + \gamma log(t)$$

o que mostra que $log\left[-log[S(t)]\right]$ é uma função linear de $log(t)$. Portanto, o gráfico de $log\left[-log\left[\hat{S}(t)\right]\right]$ *versus* $log(t)$, sendo $\hat{S}(t)$ o estimador de Kaplan-Meier, deve ser aproximadamente linear se o modelo Weibull for apropriado.

2.5 TESTE DA RAZÃO DE VEROSSIMILHANÇA

Uma forma de discriminar os modelos é através de testes de hipóteses, com a vantagem de obter uma conclusão direta, não envolvendo qualquer componente subjetivo na sua interpretação.

As hipóteses a serem testadas são:

H_0: O modelo de interesse é adequado

contra uma hipótese alternativa de que o modelo não é adequado.

O teste da razão de Verossimilhança é realizado usando um modelo generalizado tal que os modelos de interesse são casos particulares. O teste é realizado a partir dos seguintes dois ajustes: (1) modelo generalizado é obtenção do valor do logaritmo de sua função de verossimilhança ($log\left(L\left(\widehat{\theta_G}\right)\right)$); (2) modelo de interesse e obtenção do valor do logaritmo de sua função de verossimilhança ($log\left(L\left(\widehat{\theta_M}\right)\right)$). A partir destes valores é possível calcular a estatística da razão de verossimilhanças, isto é,

$$TRV = 2\left[log\left(L\left(\widehat{\theta_G}\right)\right) - log\left(L\left(\widehat{\theta_M}\right)\right)\right] \quad (5)$$

que, sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros $\left(\widehat{\theta_G}\right)$ e $\left(\widehat{\theta_M}\right)$ dos modelos sendo comparados.

Vimos anteriormente que o modelo exponencial de regressão é um caso particular do modelo de regressão Weibull quando o parâmetro γ é igual a 1, ou seja, o teste da razão de Verossimilhança será útil para determinar qual modelo usar.

2.6 ANÁLISE DOS RESÍDUOS

A avaliação da adequação do modelo ajustado através da análise dos resíduos é parte essencial da análise dos dados, pois nos ajudam como um meio de rejeitar modelos claramente inapropriados.

Os resíduos de Cox-Snell auxiliam a examinar o ajuste global do modelo final. Esses resíduos são quantidades calculadas por

$$\widehat{e_i} = \widehat{\Lambda}(t_i|\boldsymbol{x}_i) \quad (6)$$

em que $\widehat{\Lambda}(.)$ é a função de risco acumulada obtida do modelo ajustado. Para os modelos de regressão exponencial e Weibull, os resíduos de Cox-Snell são dados, respectivamente, por

Exponencial: $\widehat{e_i} = \left[t_i exp\left(-\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}\right)\right]$

Weibull: $\widehat{e_i} = \left[t_i exp\left(-\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}\right\}\right]^{1/\hat{\sigma}=\hat{\gamma}}$

Os resíduos $\widehat{e_i}$, que são estimativas dos erros que vem de uma população homogênea, devem seguir uma distribuição exponencial padrão.

Para considerar o modelo como adequado, o gráfico $\widehat{e_i}$ *versus* $-log\left(\hat{S}(\widehat{e_i})\right)$ deve ser aproximadamente uma reta com inclinação 1 quando o modelo exponencial for adequado. Aqui, $\hat{S}(\widehat{e_i})$ é a função de sobrevivência dos $\widehat{e_i}$'s obtida pelo estimador de Kaplan-Meier. O gráfico da curva de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, também auxiliam a verificar a qualidade do modelo ajustado. Quanto mais próximo elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

3 SOBRE OS DADOS

O Banco de Dados de Câncer de Pulmão NCCTG é composto por algumas informações clínicas de 166 pacientes com câncer de pulmão, incluindo tempo de sobrevivência do paciente em dias, se houve ou não censura para o paciente (0 para censura e 1 para morte), sexo do paciente (1 para homens e 0 para mulheres), Escala de performance ECOG realizada pelo médico (0 para pacientes capazes de realizar todas suas atividades sem restrição ou com restrição a atividades físicas rigorosas, 1 para pacientes incapazes de realizar qualquer atividade de trabalho). Esses dados são coletados de forma sistemática e rigorosa em ensaios clínicos multicêntricos conduzidos pelo NCCTG, garantindo a confiabilidade e a consistência das informações.

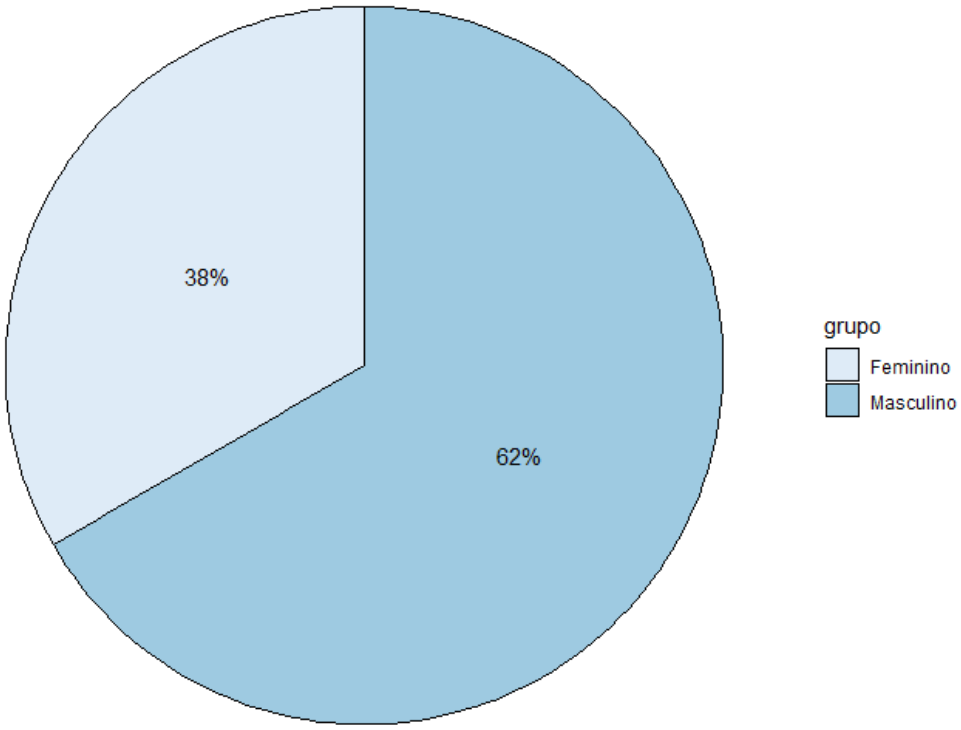
Na tabela 1 vemos uma amostra dos dados de 5 pacientes para termos uma noção de como os dados se comportam em tabelas. As colunas estão dispostas respectivamente ao texto do parágrafo anterior:

Tabela 1 – Amostra dos dados de câncer de pulmão do NCCTG

Tempo	Censura	Sexo	ECOG
455	1	1	0
210	1	1	0
1022	0	1	0
310	1	0	1
361	1	0	1

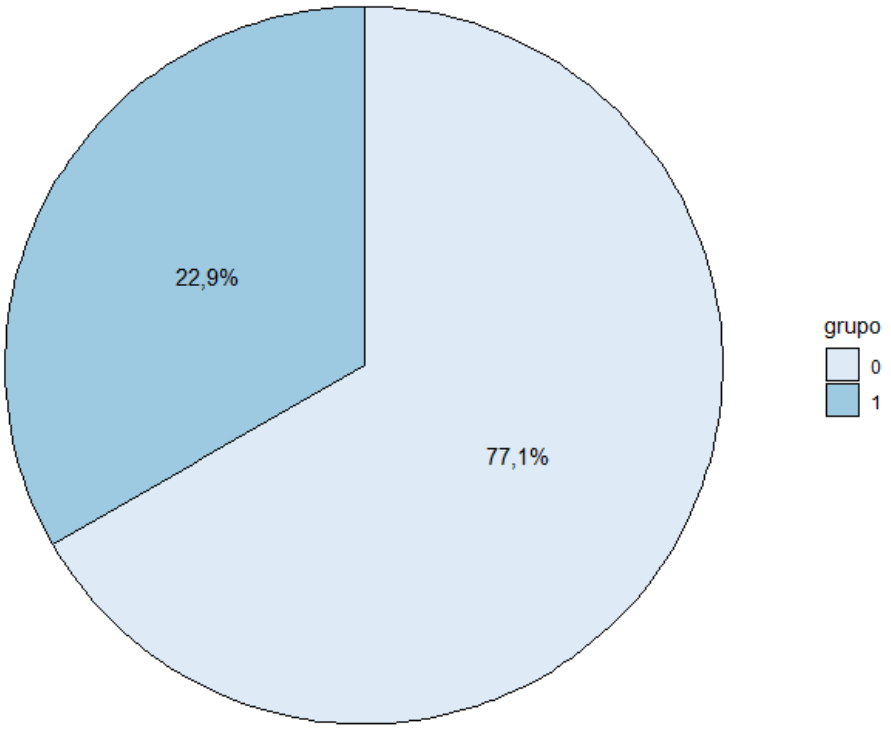
Considerando a Figura 1 e a Figura 2, podemos observar através do gráfico de setores a porcentagem de indivíduos do sexo masculino e feminino presentes na amostra, assim como a porcentagem de tipos de diagnóstico para a escala de performance:

Figura 1 - Gráfico de setores para a variável sexo dos pacientes



Fonte: Autoria própria

Figura 2 - Gráfico de setores para o tipo de diagnóstico ECOG



Fonte: Autoria própria

4 OBJETIVOS

A ideia central é observar se as variáveis sexo e ECOG influenciam no tempo de sobrevida de pacientes com câncer de pulmão.

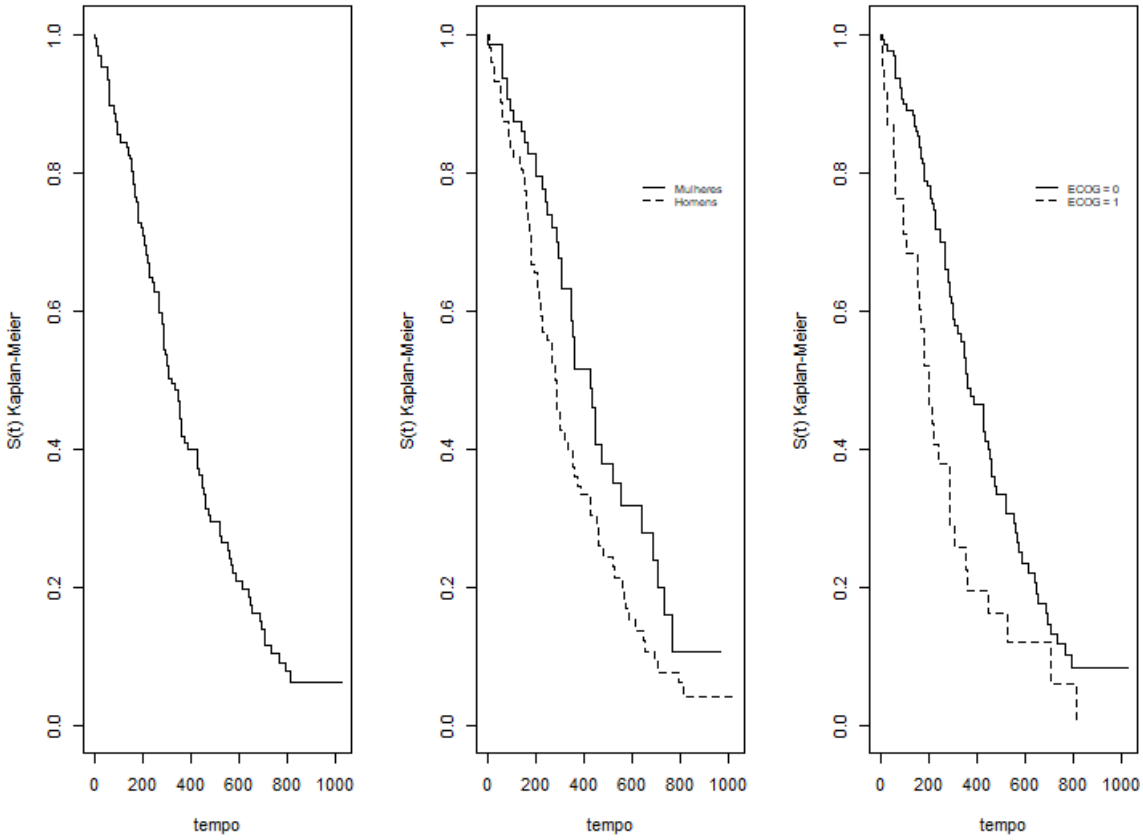
Além disso, temos como objetivo:

- Realizar curvas de sobrevivência usando o estimador de Kaplan-Meier Usando como covariáveis o sexo do paciente e escala de performance.
- Observar se algum modelo probabilístico se comporta bem aos dados e indicar qual o mais adequado

5 RESULTADOS E DISCUSSÕES

Na Figura 3 está expresso as curvas de sobrevivência estimadas pelo estimador não paramétrico de Kaplan-Meier para os dados sem levar em consideração nenhuma covariável, depois levando em consideração apenas o sexo e, por fim, levando em consideração apenas a performance ECOG, respectivamente.

Figura 3 - Gráficos das estimativas da curva de Kaplan-Meier não levando em consideracao nenhuma covariável, levando em consideração o sexo, levando em consideração o diagnóstico ECOG, respectivamente.

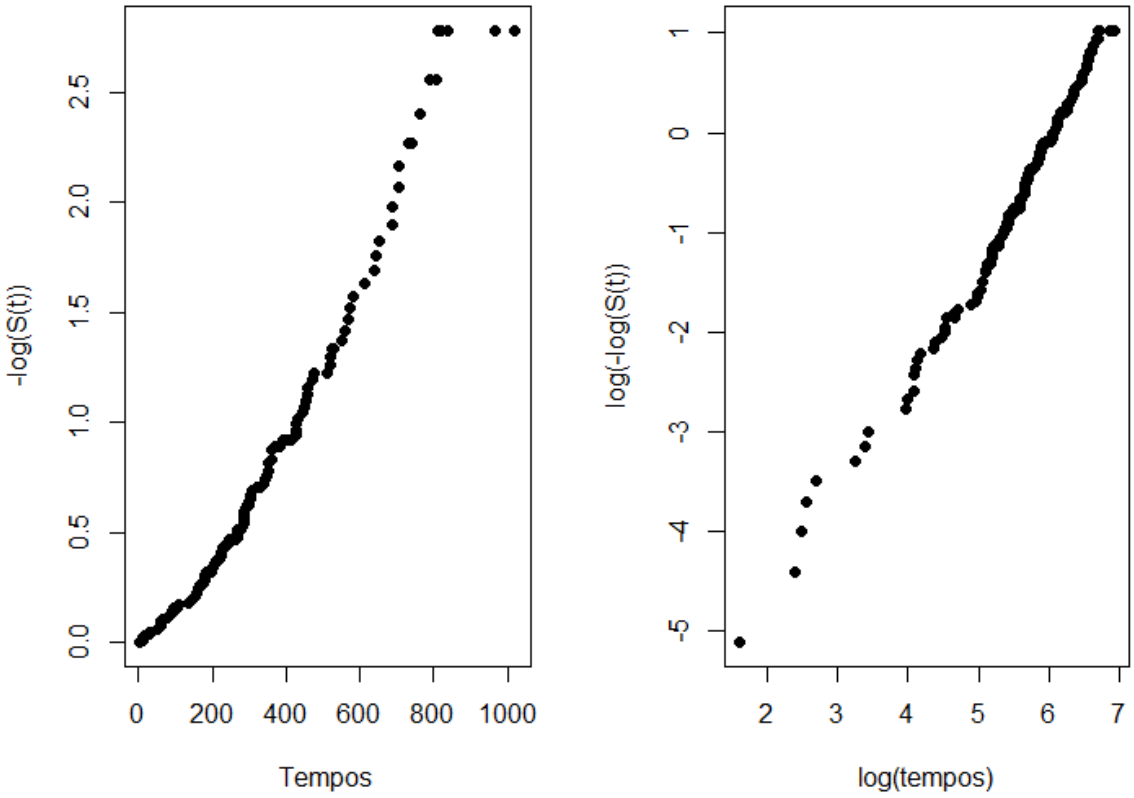


Fonte: Autoria própria

Realizando o teste log-rank para as curvas estimadas para o sexo dos pacientes e para o diagnóstico ECOG, rejeitou-se a hipótese nula de que as curvas são iguais a um nível de 5% de significância, sendo esse um bom indicativo de que tais covariáveis serão úteis na aplicação dos nossos modelos de regressão probabilísticos.

Antes de realizar a aplicação dos dados nos modelos de regressão, vamos avaliar qual modelo parece ser mais adequado através dos gráficos das linearizações, ignorando as covariáveis de sexo e ECOG. Na Figura 3 temos os gráficos das linearizações para a distribuição exponencial e Weibull, respectivamente.

Figura 4 - Gráficos das linearizações para a distribuição exponencial e Weibull, respectivamente



Fonte: Autoria própria

Pela Figura 3, vemos que o gráfico que segue mais linear é o que favorece a aplicação do modelo Weibull, ou seja, o segundo gráfico. Ajustando os modelos de regressão Exponencial e Weibull usando as covariáveis de sexo e ECOG, obtemos as estimativas que estão apresentadas na Tabela 2. É possível observar que o valor de γ para o modelo de Weibull está distante de 1. Usando o teste da razão de verossimilhança e testando as hipóteses $H_0: \gamma = 1$ contra $H_1: \gamma \neq 1$ obteve-se *p-valor* $< 0{,}00004$ e, portanto, não rejeitamos o modelo de regressão Weibull em favor do exponencial.

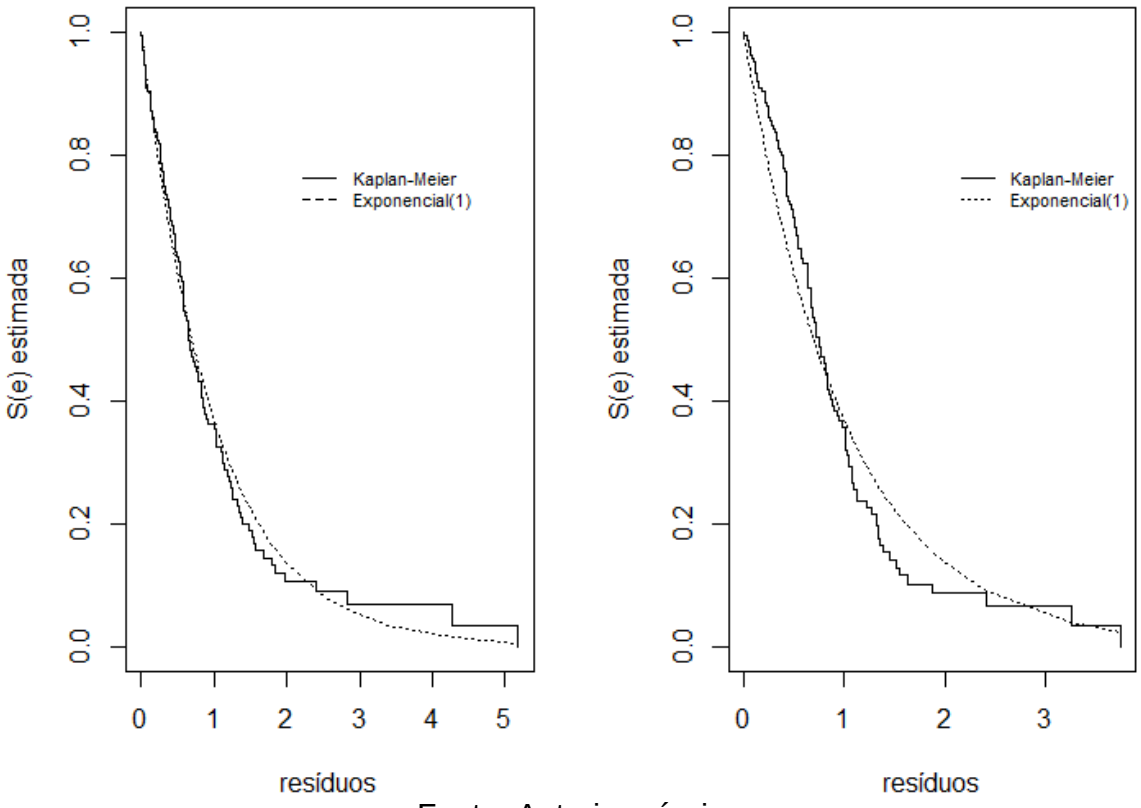
Tabela 2 – Estimativa dos modelos de regressão exponencial e Weibull

Regressão exponencial	Regressão Weibull
$\widehat{\beta_0} = 6{,}51$	$\widehat{\beta_0} = 6{,}39$
$\widehat{\beta_1} = -0{,}47$	$\widehat{\beta_1} = -0{,}35$
$\widehat{\beta_2} = -0{,}66$	$\widehat{\beta_2} = -0{,}53$
$\hat{\gamma} = 1$	$\hat{\gamma} = 1{,}353$

Testou-se também, a um nível de 5% de significância, para ambos os modelos de regressão, $H_0: \beta_1 = 0$ e $H_0: \beta_2 = 0$ contra $H_1: \beta_1 \neq 0$ e $H_1: \beta_2 \neq 0$, respectivamente. Com todas as hipóteses nulas foram rejeitadas, concluiu-se que, para ambos os modelos, as covariáveis de sexo e ECOG influenciam no tempo de sobrevida de pacientes com câncer de pulmão.

Até agora, de acordo com o teste da razão de verossimilhança e os gráficos das linearizações, podemos afirmar que a regressão de Weibull possui uma vantagem para modelar o tempo até a morte de pacientes com câncer. Para concluir qual modelo é mais adequado, vamos examinar os resíduos de Cox-Snell de ambos, realizando os gráficos de sobrevivência desses resíduos através do estimador Kaplan-Meier e comparando-os com uma distribuição exponencial padrão. Quanto mais próximos as estimativas estiverem, melhor será o modelo. Pela Figura 4 vemos que a regressão Weibull apresenta uma melhor performance quando comparado com o modelo de regressão exponencial. Dessa forma, usaremos o modelo de regressão Weibull para estimar nossas curvas de sobrevivência para cada caso das nossas covariáveis.

Figura 5 - Análise dos resíduos de Cox-Snell do modelo de regressão de Weibull e Exponencial ajustados, respectivamente

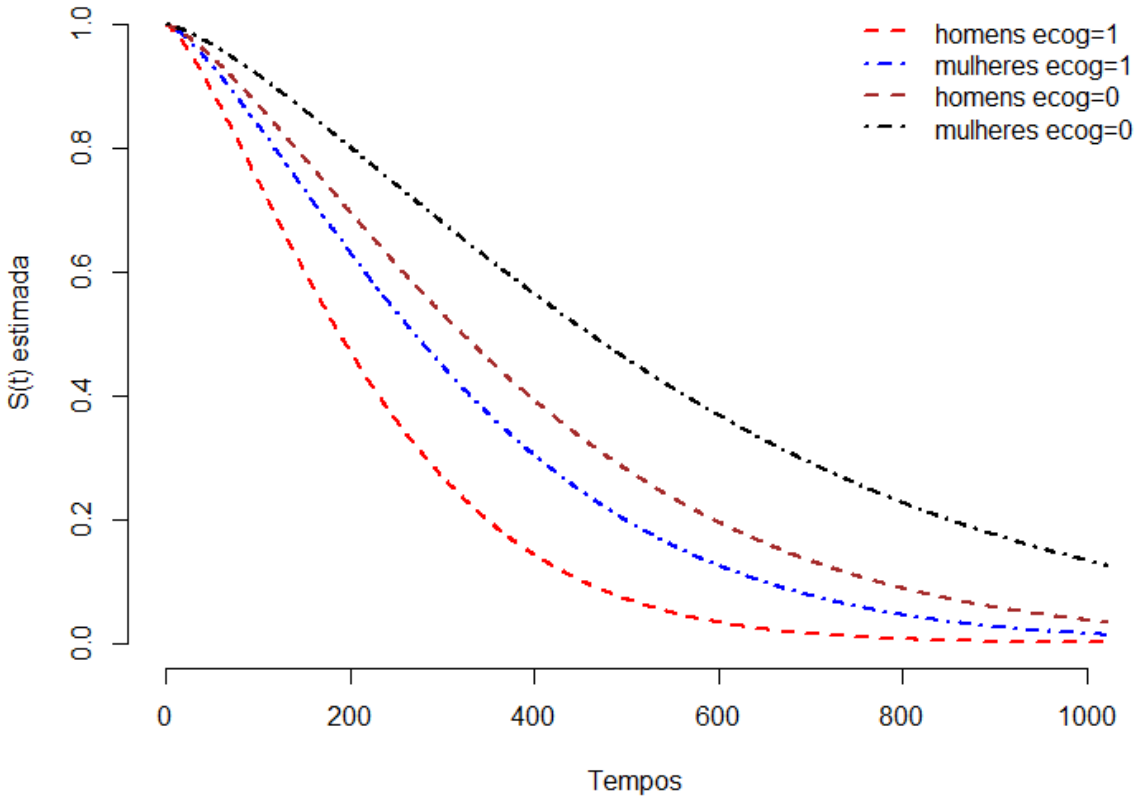


Fonte: Autoria própria

Na Figura 5 estão as curvas de sobrevivência do modelo de regressão Weibull levando em consideração os valores das covariáveis sexo e ECOG. Também temos, pela equação 7, a função de sobrevivência estimada para o modelo de Weibull em questão.

$$S(t|x) = exp\left\{-\left(\frac{t}{exp(6{,}39 - 0{,}35 * Sexo_i - 0{,}53 * ECOG_i)}\right)^{1{,}37}\right\} \qquad (7)$$

Figura 6 - Curvas de sobrevivência estimadas pelo modelo regressão Weibull para cada valor das covariáveis sexo e ECOG



Fonte: Autoria própria

Pela função de sobrevivência do modelo de regressão Weibull presentes na equação 7 e na Figura 5, vemos que homens com diagnóstico ECOG igual a 1 têm probabilidades de sobrevivência menor, seguidos de mulheres com diagnostico ECOG igual a 1, homens com diagnóstico ECOG igual a 0 e mulheres com diagnóstico ECOG igual a 0.

Dessa forma, foi possível mostrar que o tempo de sobrevida de pacientes com câncer de pulmão sofre influência do sexo do paciente e que o diagnóstico ECOG é útil para indicar se um paciente necessita de mais cuidados especiais.

6 REFERÊNCIAS

COLOSIMO E. A; GIOLO S. R. ANÁLISE DE SOBREVIVÊNCIA APLICADA. Blucher, 1 janeiro de 2006.

CARVALHO M. S. et al. Análise de sobrevivência: teoria e aplicações em saúde. 2º edição. Rio de Janeiro: Editora Fiocruz, 2011.