Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

SurvControl: Um Painel de Controle Baseado em Modelos de Sobrevivência

Gabriel Grandemagne dos Santos

18 de abril de 2023

Gabriel Grandemagne dos Santos

SurvControl: Um Painel de Controle Baseado em Modelos de Sobrevivência

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Profa. Dra. Silvana Schneider

Porto Alegre
Abril de 2023

Gabriel Grandemagne dos Santos

SurvControl: Um Painel de Controle Baseado em Modelos de Sobrevivência

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Profa. Dra. Silvana Schneider, UFRGS
Doutor(a) pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Banca Examinadora:

Profa. Dra. Silvana Schneider, UFRGS
Doutora pela Universidade Federal de Minas Gerais – Belo Horizonte, MG

Prof. Dr. Rodrigo Citton Padilha dos Reis, UFRGS
Doutor pela Universidade Federal de Minas Gerais – Belo Horizonte, MG

Porto Alegre
Abril de 2023

Resumo

A Análise de Sobrevivência engloba diversos métodos utilizados para analisar o tempo até a ocorrência do evento de interesse, conhecido como taxa de falha, como o estimador de Kaplan-Meier, modelo de regressão de Cox, modelos de fragilidade e suas extensões, capazes de acomodar a censura informativa e dependente. Todos os diferentes métodos citados podem ser utilizados para prever o tempo de sobrevida dos indivíduos. Nosso trabalho tem como intuito aplicar esses métodos estatísticos nos dados oncológicos RHC fornecidos pela Fundação Oncocentro de São Paulo (FOSP, 2022) e disponibilzar uma ferramenta computacional para análise da sobrevivência, como forma de Painel de Controle ou de monitoramento dos pacientes presentes na base. Nosso aplicativo (`https://ggrandemagne.shinyapps.io/SurvControl2/`) tem como objetivo fornecer uma maneira intuitiva do usuário escolher e entender os efeitos das covariáveis, como por exemplo, idade, sexo, grau de escolaridade, entre outras, sobre o tempo de sobrevivência de diferentes grupos ou indivíduos, assim como associar se há dependência entre o tempo de falha e de censura nos dados. Com isto em mente, também iremos apresentar uma breve análise comparativa entre pacientes de câncer de mama e ovário, onde avaliamos que há efeito da associação entre o tempo de falha e censura, sendo assim, é vantajoso aplicarmos os modelos que levam em consideraçao o efeito dessa relação.

Palavras-Chave: Análise de Sobrevivência, Fragilidade, Dados Oncológicos, Painel de Controle.

Abstract

Survival analysis encompasses various methods used to analyze the time until the occurrence of an event of interest, known as failure time, such as the Kaplan-Meier estimator, Cox regression model, frailty models, and their extensions, capable of accommodating informative and dependent censoring. All the different methods mentioned can be used to predict the survival time. Our work aims to apply these statistical methods to the RHC oncology data, provided by the São Paulo Oncocenter Foundation (FOSP, 2022) and provide a computational tool for survival analysis, as a form of Control Panel for cancer patients. Our application (https://ggrandemagne.shinyapps.io/SurvControl2/) aims to provide an intuitive way for the user to choose and understand the effects of covariates, such as age, sex, educational level, among others, on the survival time of different groups or individuals, as well as to associate whether there is dependence between the failure time and censoring time in the data. With this in mind, we will also present a brief comparative analysis between breast and ovarian cancer patients, where we evaluated that there is an effect of the association between failure and censoring time, thus it is advantageous to apply models that take into account the effect of this association.

Keywords: Survival Analysis, Frailty, Oncology Data, Control Panel.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Análise de Sobrevivência é um conjunto de procedimentos estatísticos aplicados na análise de dados, onde o objetivo é estimar o tempo até a ocorrência de determinado evento de interesse (Carvalho et al., 2011). Esse tempo até o evento de interesse é denominado tempo de falha, tempo de vida, ou tempo de sobrevivência. Por exemplo, tempo até o óbito de pacientes, ou tempo até a falha de uma máquina (chamado de análise de confiabilidade).

Uma das principais características dos dados de sobrevivência é a presença de observações incompletas nomeadas de censura, que é a observação parcial da resposta. A censura pode ocorrer quando por alguma razão o acompanhamento de algum indivíduo no estudo for interrompido, seja porque ele mudou de cidade, o estudo terminou para a análise dos dados ou, o paciente veio a óbito por causas diferentes das estudadas (Colosimo e Giolo, 2006).

De acordo com Colosimo e Giolo (2006), a censura pode ser classificada da seguinte forma:

1. Censura à direita, é a situação mais comum para estudos que envolvem dados de sobrevivência, ocorre quando o tempo de ocorrência do evento de interesse se encontra à direita do tempo observado, podemos ter 3 tipos, são eles:

    - Censura do tipo I: Ocorre nos estudos em que, ao serem finalizados, constam em seu término alguns indivíduos que ainda não apresentaram o evento de interesse.
    - Censura do tipo II: Resultam de estudos que são finalizados após a ocorrência de um número pré-definido de eventos de interesse em um número pré-estabelecido de indivíduos.
    - Censura Aleatória: É o mais comum na prática, ocorre quando o indivíduo for retirado do estudo sem ter ocorrido o evento de interesse ou se o indivíduo falhar por outras causas, sem correlação com o evento de interesse.

2. Censura à esquerda, ocorre quando o tempo registrado é maior que o tempo de falha, isto é, o evento já ocorreu quando o indivíduo foi observado.

3. Censura intervalar, é um tipo de censura mais geral que acontece comumente em estudos em que os pacientes são acompanhados em visitas periódicas, logo, sabemos apenas que o evento de interesse ocorreu entre um intervalo de tempo

Também podemos classificar a censura entre dependente ou independente, onde a primeira está relacionada com a ocorrência do evento de interesse e a segunda ocorre independentemente do mesmo (Kalbfleisch e Prentice, 2011). Além disso, os dados censurados são uma parte fundamental para a modelagem da sobrevida, pois apesar de serem incompletas, as observações censuradas fornecem informações importantes sobre o tempo de falha (Schneider et al., 2019).

> "A relação entre a censura e o tempo de falha pode ser diferenciada de acordo com as causas geradoras da censura. Por exemplo, em estudos clínicos sobre câncer de mama, algumas causas de observações incompletas podem ser: a) término do estudo; b) óbito devido a causas externas; c) remoção do estudo se houver evidências clínicas da ineficiência do tratamento; d) abandono do estudo devido aos efeitos colaterais da terapia que a paciente está recebendo. Em a) e b) a censura ocorre independentemente da falha, ou seja, ela não traz informação alguma sobre o tempo da falha, nesses casos dizemos que a censura é não informativa. Em c) e d) as observações censuradas parecem predizer algo sobre o tempo de falha, ou seja, elas estão relacionadas com a falha, nesses casos dizemos que a censura é informativa" (Schneider, 2017).

Dentre os métodos utilizados na análise de sobrevivência, existem diversos modelos de regressão para verificar se as covariáveis influenciam no tempo de sobrevivência ou de censura. Por exemplo, o modelo semi-paramétrico de Cox (Cox, 1972), no qual permite modelar os efeitos das covariáveis sobre o tempo até o evento de interesse.

Segundo Carvalho et al. (2011), a capacidade do modelo de Cox é limitada quando falta covariáveis importantes, como *status* socioeconômico, o que reflete na variabilidade das observações. Isso quer dizer que os indivíduos apresentam grande heterogeneidade ou diferentes fragilidades não atribuíveis a qualquer característica medida. Esse problema pode ser tratado com a inclusão de um efeito aleatório para cada indivíduo, ou grupo de indivíduos, que torna a estimativa dos efeitos das covariáveis mais eficientes. Logo, a inclusão da fragilidade em um estudo é a inclusão de um efeito aleatório na função taxa de falha para descrever alguma possível associação entre indivíduos, ou grupos de indivíduos (Colosimo e Giolo, 2006).

> "Segundo Vaupel et al. (1979) a fragilidade é definida como uma variável aleatória multiplicativa na taxa de mortalidade. Podemos dizer que isso representa a associação entre os tempos de falha, onde esta associação pode ser encontrada de forma intrapessoal, quando possuímos vários tempos observados para um mesmo indivíduo (eventos recorrentes, como ataque do miocárdio), ou tempos de diferentes indivíduos provindo da mesma família (correlação genética), ou até pacientes provindos do mesmo bloco hospitalar" (Schneider, 2017).

Neste trabalho, iremos utilizar o Software R (R Core Team, 2022) para aplicar as técnicas de análise de sobrevivência, como o modelo de regressão de Cox (Cox, 1972) e a curva de Kaplan-Meier (Kaplan e Meier, 1958), utilizando o pacote *survival* (Therneau, 2022), com visualização feita através do pacote *survminer* (Kassambara et al., 2021) e modelos para acomodar censura dependente, utilizando o pacote *DepCens* (Schneider e Grandemagne dos Santos, 2022), pois também apresentaremos uma situação em que há evidências de dependência entre o tempo de falha e censura para pacientes de diferentes tipos de câncer (mama e ovário), que pode ocasionar em uma mudança na curva de sobrevivência.

Neste trabalho, os modelos empregados para censura dependente são baseados em extensões dos modelos de fragilidade, capazes de acomodar a dependências entre tempos de falha e de censura, utilizando as distribuições marginais Weibull e Exponencial por partes. Para isto utilizamos, o pacote *DepDens*, que contém os modelos propostos em Schneider et al. (2019).

Utilizando principalmente os pacotes mencionados acima, temos como objetivo desenvolver um aplicativo chamado Painel de Controle, em formato *shinydashboard* (Chang e Borges Ribeiro, 2021). Este aplicativo, (`https://ggrandemagne.shinyapps.io/SurvControl2/`), proporciona ao usuário uma fácil visualização e entendimento do tempo de sobrevivência e do risco estimado, para pacientes com diferentes tipos de cânceres. Para isso, construímos um algoritmo em R, no formato *shiny* (Chang et al., 2022), que utiliza os pacotes de análise de sobrevivência e os aplica na base de dados da Fundação Oncocentro de São Paulo (FOSP) (FOSP, 2022). Esse Painel de Controle possibilita uma aplicação de diversos filtros, gráficos e saídas, podendo ser avaliado, por exemplo, se há estrutura de dependência entre o tempo de falha e censura, e como isso afeta na estimação do tempo de sobrevivência dos pacientes.

Os dados que utilizamos são disponibilizados pela Secretaria de Saúde do estado de São Paulo e atualizados pela FOSP (FOSP, 2022), trimestralmente, desde 2000. A Fundação Oncocentro de São Paulo é um órgão de apoio da Secretaria de Saúde, para assessorar a política de saúde em câncer no Estado de São Paulo. O banco que utilizamos para a versão corrente do aplicativo contém os dados a partir de

2014 para indivíduos com câncer de ovário, mama ou próstata, totalizando 67.369 participantes.

Logo, iremos utilizar o *software R*, em conjunto com o banco de dados disponibilizados pela Fundação Oncocentro de São Paulo para aplicar a curva de Kaplan-Meier (Kaplan e Meier, 1958), modelo de Cox (Cox, 1972) e modelos para censura dependente (Schneider et al., 2019) com o intuito de avaliar o tempo de sobrevivência em algumas doenças (câncer de mama, útero, ovário, próstata, entre outros). Iremos proporcionar ao usuário uma ferramenta em forma de *Shiny Dashboard*, onde vamos apresentar um Painel de Controle para pacientes com esses tipos de cânceres. Esta *Dashboard* irá proporcionar uma visualização da curva de sobrevivência, relação com a taxa de falha e influência de covariáveis, através dos métodos e pacotes citados anteriormente. Além disso, neste trabalho iremos avaliar os dados de pacientes que possuem câncer de mama e/ou ovário, com o intuito de melhor compreender seu comportamento na população, utilizando nossa *Dashboard* como ferramenta analítica.

O Capítulo 2 apresenta uma introdução metodológica aos estimadores e modelos aplicados neste trabalho. O Capítulo 3 introduz a implementação computacional, que apresenta os algoritmos utilizados nas estimações e ajustes utilizados, assim como a criação do painel de controle. O Capítulo 4 apresenta a aplicação computacional, mostrando como os dados foram selecionados e manipulados para sua aplicação em *shinydashboard*, assim como uma breve análise da dependência entre o tempo de falha e censura para pacientes de câncer de mama e/ou ovário. Por último, no Capítulo 5 apresentamos as conclusões e discussões finais.

2 Metodologia

2.1 Definições das funções de sobrevivência

Nesta seção iremos apresentar algumas funções utilizadas na análise de sobrevivência. As definições desta seção foram retiradas de Carvalho et al. (2011).

Suponha T uma variável contínua e positiva, com função densidade de probabilidade $f(t)$, representando o tempo até a ocorrência de um evento de interesse. Em análise de sobrevivência essa função pode ser representada como a probabilidade de um indivíduo sofrer um evento em um intervalo imediato de tempo, em que ϵ é o incremento de tempo infinitamente pequeno (Carvalho et al., 2011),

$$f(t) = \lim_{\epsilon \to 0^+} \frac{P(t \leq T < t + \epsilon)}{\epsilon}.$$

Segundo Carvalho et al. (2011), a função de sobrevivência, $S(t)$, é a probabilidade de um indivíduo não falhar ou sobreviver a um certo tempo t, estabelecido no estudo. Essa função é dada por

$$S(t) = P(T > t).$$

A função de distribuição acumulada, $F(t)$, da variável aleatória T é a probabilidade de um evento ocorrer até o tempo t, tal que

$$F(t) = P(T \leq t).$$

Logo podemos perceber que a função de sobrevivência, $S(t)$, é simplesmente o complementar da função de distribuição acumulada vista anteriormente, $F(t)$, dado que

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

A função taxa de falha, $h(t)$, é definida como a taxa instantânea de um indivíduo falhar entre o tempo t e $t + \epsilon$, dado que o indivíduo sobreviveu até o tempo t. A

função taxa de falha, $h(t)$, é definida como (Carvalho et al., 2011)

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

2.2 Estimador de Kaplan-Meier

As definições e equações desta Seção foram extraídas do livro de Colosimo e Giolo (2006). O estimador não-paramétrico de Kaplan-Meier, também conhecido como estimador limite-produto, foi proposto por Kaplan e Meier (1958) para ser aplicado na estimação da função de sobrevivência. O estimador de Kaplan-Meier é obtido pela seguinte função:

$$\hat{S}(t) = \frac{\text{nº de obsevações sem falha até o tempo t}}{\text{nº total de obsevações}}.$$

Aqui, a função de sobrevivência é uma função do tipo escada, com saltos de tamanho $1/n$, onde n é o tamanho amostral. O estimador de Kaplan-Meier considera o número de falhas distintas para determinar o número de intervalos no eixo do tempo, ou seja, um intervalo de tempo para cada falha distinta observada. Segundo Colosimo e Giolo (2006), temos que, ao considerar a função de sobrevivência uma função discreta com saltos, com probabilidade maior que zero somente nos tempos de falha $t_j, j = 1, \ldots, k$, temos que

$$S(t_j) = (1 - q_1)(1 - q_2) \ldots (1 - q_j),$$

onde q_j representa a probabilidade de óbito de um indivíduo ocorrer no intervalo $[t_{j-1}, t_j)$, sendo que o mesmo indivíduo sobreviveu até t_{j-1} e considerando $t_0 = 0$. Assim podemos descrever q_j como

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}).$$

Com as informações apresentadas anteriormente, definimos a expressão geral da função de sobrevivência $S(t)$ em termos de probabilidades condicionais, para $j = 1, \ldots, k + 1$, onde $t_{k+1} = \infty$ temos que o estimador de q_j é dado por

$$\hat{q}_j = \frac{\text{nº de falhas em } t_{j-1}}{\text{nº total de obsevações sob risco em } t_{j-1}}.$$

Para definir o estimador de Kaplan-Meier devemos levar em consideração as seguintes definições:

- $t_1 < t_2 < \cdots < t_k$, k tempos de falha distintos e ordenados;

- d_j o número de falhas em t_j, $j = 1, \ldots, k$;
- n_j o número de indivíduos em risco no tempo t_j.

Então, podemos obter o estimador de Kaplan-Meier pela seguinte equação:

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

De acordo com Colosimo e Giolo (2006), o estimador Kaplan-Meier é o estimador de máxima verossimilhança de $S(t)$, também é não-viciado quando utilizados em amostras grandes, é fracamente consistente e converge assintoticamente para um processo Gaussiano.

2.3 Modelo de Regressão de Cox

De acordo a Colosimo e Giolo (2006), o modelo de regressão proposto por Cox (1972) permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando por covariáveis. As definições e equações desta Seção foram extraídas do livro de Colosimo e Giolo (2006).

A suposição básica para a aplicação do modelo de regressão de Cox é que as taxas de falha sejam proporcionais. Suponhamos que temos como objetivo comparar o tempo de falha entre dois diferentes grupos, em que os pacientes são selecionados aleatoriamente para receber o tratamento padrão (grupo 0) ou o novo tratamento (grupo 1). Ao representar a função de taxa de falha do grupo 0 por $h_0(t)$ e a do segundo grupo por $h_1(t)$, assumindo proporcionalidade entre essas funções temos que

$$\frac{h_1(t)}{h_0(t)} = K,$$

em que K é definida pela razão das taxas de falha dos grupos, constante para todo o tempo de acompanhamento de estudo.

Considere X uma variável indicadora de grupo, em que:

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

e seja $K = \exp\{\beta x\}$. Logo, a função taxa de falha pode ser obtida por

$$h(t|x) = h_0(t) \exp\{\beta x\},$$

ou seja,

$$h(t|x) = \begin{cases} h_1(t) = h_0(t)\exp\{\beta\}, & \text{se } x = 1 \\ h_0(t), & \text{se } x = 0 \end{cases}$$

A expressão acima define o modelo de Cox (1972) para uma única covariável. Genericamente, considerando p covariáveis, de modo que X seja um vetor com as covariáveis contendo seus respectivos componentes $X = (X_1, \ldots, X_p)'$. A expressão geral do modelo de regressão de Cox resulta em

$$h(t|x) = h_0(t)g(x'\beta),$$

em que $g(x'\beta)$ é uma função não-negativa que deve ser especificada, tal que $g(0) = 1$.

Portanto, o modelo é definido pelo produto de duas funções, uma paramétrica e outra não-paramétrica. A função $h_0(t)$ é denominado por taxa de falha basal, pois a função $h(t|x) = h_0(t)$ caso $x = 0$. O componente paramétrico é utilizado na seguinte forma multiplicativa:

$$g(x'\beta) = \exp\{x'\beta\} = \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\},$$

em que β denota o vetor de parâmetros associado às covariáveis, garantindo que $h(t|x)$ seja sempre não-negativa. Este modelo é também denomidado por modelo de taxas proporcionais, pois tem a propriedade de garantir que a razão entre as taxas de falha de dois indivíduos diferentes sejam constantes no tempo t, desde que o efeito das covariáveis também sejam constantes ao longo do tempo, ou seja, a razão das funções de taxa de falha para indivíduos i e j, $i \neq j$ não depende do tempo. Esta razão é dada por

$$\frac{h(t|x_i)}{h(t|x_j)} = \frac{h_0(t)\exp\{x_i'\beta\}}{h_0(t)\exp\{x_j'\beta\}} = \exp\{x_i'\beta - x_j'\beta\}.$$

Para definirmos a função de verossimilhança, precisamos primeiramente apresentar que a função taxa de falha pode ser relacionada com a função de sobrevivência da seguinte forma (Carvalho et al., 2011)

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

A partir dessa propriedade, notamos que a função taxa de falha, $h(t)$, e função de sobrevivência, $S(t)$, são inversamentre proporcionais (Carvalho et al., 2011). A função taxa de falha acumulada, $H(t)$, é utilizada quando queremos avaliar melhor a função taxa de falha (Colosimo e Giolo, 2006). Ela pode ser calculada ao integrar

a função taxa de falha, dada por

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

Agora, em relação as funções de taxa de falha, taxa de falha acumulada e função de sobrevivência, podemos relacioná-lás da seguinte maneira (Colosimo e Giolo, 2006)

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

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

$$S(t) = \exp\{-H(t)\} = \exp\Big\{-\int_0^t h(u)du\Big\}.$$

Segundo Colosimo e Giolo (2006), vemos que a função de verossimilhança geral é dada por

$$L(\theta) = \prod_{i=1}^{n}[h(t_i)S(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i},$$

$$L(\theta) = \prod_{i=1}^{n}[h(t_i)^{\delta_i}S(t_i)].$$

Levando em consideração as relações pontuadas acima entre a função de sobrevivência $S(t)$ e função taxa de falha $h(t)$, Colosimo e Giolo (2006) aponta que a função de verossimilhança com inclusão de covariáveis para o modelo de regressão de Cox é dada por

$$L(\theta) = \prod_{n=1}^{n}(h_0(t_i)\exp\{x_i'\beta\})^{\delta_i}[S_0(t_i)]^{\exp\{x_i'\beta\}}.$$

No ajuste do modelo de regressão de Cox nós utilizamos o método de máxima verossimilhança para a estimação de parâmetros. Esse método implica em estimar o parâmetro θ maximizando a equação $L(\theta)$, que é equivalente a maximizar o logaritmo da função de verossimilhança (Colosimo e Giolo, 2006). Computacionalmente iremos utilizar a função *coxph()* do pacote *survival* (Therneau, 2022) no software R (R Core Team, 2022).

2.4 Modelos com Fragilidade

Para os métodos estatísticos apresentados anteriormente a suposição considerada foi de que os tempos de sobrevivência de diferentes indivíduos são independentes. Esta suposição é válida para muitos estudos, porém também pode ser inadequada

para outros. É lógico supor alguma relação entre o tempo de sobrevivência de indivíduos quando, por exemplo, observamos em estudos médicos indivíduos provindo da mesma família (indivíduos de uma mesma família tendem a ter características semelhantes), ou ala hospitalar, o que caracteriza situações em que a suposição de independência entre os tempos de sobrevivência pode não ser válida.

A fragilidade foi proposta por Vaupel et al. (1979), que ao analisar uma tábua de vida publicada na época, percebeu que ignorava o fato de que alguns indivíduos específicos eram mais propensos a ter um tempo de vida menor que os demais. Ele ainda observou que estes indivíduos mais "frágeis" morriam antes, logo, concluiu que a estimativa de taxa de risco seria viesada. Para resolver esse problema de heterogeneidade, ele propôs a inclusão de um fator aleatório, chamado por fragilidade, na taxa de falha, fator do qual modificaria o risco de forma multiplicativa.

Segundo Hanagal (2011), os modelos de fragilidade podem ser aplicados a dados de sobrevivência multivariados para captar a correlação entre os tempos dos indivíduos de um mesmo grupo ou para captar a correlação presente nos eventos recorrentes para um mesmo indivíduo (Schneider, 2017). O modelo de fragilidade aplicado a estes casos é chamado por *modelo de fragilidade compartilhada*, pois o efeito aleatório é compartilhado entre os indivíduos de um mesmo grupo (Colosimo e Giolo, 2006).

O modelo de fragilidade é formulado por meio da extensão do modelo proposto por Cox (1972), incluindo um efeito aleatório, denominado fragilidade. As definições apresentadas a seguir foram extraídas de Colosimo e Giolo (2006).

A função taxa de falha para o i-ésimo indivíduo no k-ésimo grupo, condicional à fragilidade z_k e a um vetor de dimensão p de covariáveis, que podem ser em nível de grupo e em nível de indivíduo, $x_{i,k}$, é expressada por

$$h_{i,k}(t|z_k, x_{i,k}) = h_0(t) z_k \exp(\beta x_{i,k}),$$

em que $t_{i,k}$ denota o tempo de sobrevivência para o i-ésimo sujeito no k-ésimo grupo, sendo $i = 1, \ldots, n_k$, onde n_k representa o número de indivíduos no k-ésimo grupo, e $k = 1, \ldots, m$, onde m representa o número total de grupos, β é o vetor com dimensão $p \times 1$ de coeficientes de regressão desconhecidos, $h_0()$ denota a função taxa de falha basal e z_k é a fragilidade associada ao k-ésimo grupo.

Podemos reescrever o modelo apresentado acima como

$$h_{i,k}(t|w_k, x_{i,k}) = h_0(t_{i,k}) \exp(\beta' x_{i,k} + w_k),$$

em que $z_k = \exp(w_k)$.

Geralmente, assume-se que os efeitos aleatórios w_k possuem média zero e variância desconhecida. Os modelos de fragilidade podem ser aplicados para casos de

sobrevivência univariada, quando cada indivíduo tem sua própria fragilidade, e para casos multivariados, em que a fragilidade corresponde ao efeito dos grupos (Colosimo e Giolo, 2006). A função de verossimilhança, com a inclusão da fragilidade é dada por

$$L(\beta, \theta | Z_i, \ldots, Z_n) = \prod_{n=1}^{n} \left[Z_i h_0(t_i; \theta) e^{\beta' x_i} \right]^{\delta_i} \exp\{-Z_i H_0(t_i; 0) e^{\beta' x_i}\}.$$

2.5 Modelos para Censura Dependente

As equações e definições apresentadas nesta Seção foram extraídas de Schneider (2017). Portanto, informações adicionais podem ser pesquisadas neste trabalho.

2.5.1 Construção da função de verossimilhança

Seja $Y = min(T, C, A)$, onde Y é uma variável aleatória não-negativa representando o tempo observável, sendo T a variável aleatória que denota o tempo até a falha, C a variável aleatória que denota o tempo até a censura dependente (informativa) e A a variável aleatória que denota o tempo até a censura administratriva. Também consideramos $\delta^{(T)}$ a variável aleatória indicadora de falha e $\delta^{(C)}$ a variável aleatória indicadora de censura dependente. Logo, temos que

$$\delta^{(T)} = \begin{cases} 1, & \text{se } T \leq min(C, A), \\ 0, & \text{caso contrário} \end{cases} \quad \text{e} \quad \delta^{(C)} = \begin{cases} 1, & \text{se } T \leq min(T, A), \\ 0, & \text{caso contrário} \end{cases}$$

Caso haja censura administrativa, temos que $\delta^{(T)} = 0$ e $\delta^{(C)} = 0$.

Se assumirmos que todas as censuras são independentes, podemos obter a função de verossimilhança pelo simples produto entre as funções dos tempos de falha e tempos de censura. Portanto, a função de verossimilhança é tal que

$$L(.) \propto \prod_{i=1}^{n} [h^{(T)}(y_i)]^{\delta_i^{(T)}} S^{(T)}(y_i) \times [h^{(C)}(y_i)]^{\delta_i^{(C)}} S^{(C)}(y_i).$$

Consequentemente, quando a censura é independente, temos que a função de verossimilhança pode ser obtida pela sequinte equação:

$$L(\theta^{(T)}) \propto \prod_{i=1}^{n} [h^{(T)}(y_i | \theta^{(T)})]^{\delta_i^{(T)}} S^{(T)}(y_i | \theta^{(T)}).$$

Se a censura é dependente, o cálculo da probabilidade conjunta $P(T, C)$ precisa ser levado em consideração na função de verossimilhança. Podemos notar que o tempo até a censura administrativa (variável A) é independente das variáveis T e

C (variável aleatória tempo até a falha e tempo até censura dependente, respectivamente). Logo, temos que a função de verossimilhança é dada por

$$\begin{aligned} L(\theta|D) &= \prod_{i=1}^{n} \Big[P(T \in (y_i, y_i + \Delta y_i], C > y_i|\theta)\}\Big]^{\delta_i^{(T)}} \\ &\times \Big[P(C \in (y_i, y_i + \Delta y_i], T > y_i|\theta)\}\Big]^{\delta_i^{(C)}} \\ &\times \Big[P(T > y_i, C > y_i)\Big]^{(1-\delta_i^{(T)})(1-\delta_i^{(C)})}, \end{aligned}$$

em que $\theta = (\theta^{(T)}, \theta^{(C)})$ e $D = (\delta^{(T)}, \delta^{(C)}, x^{(T)}, x^{(C)})$ é o conjunto de dados observados.

Condicional na fragilidade W, temos a função de verossimilhança condicional para o i-ésimo indivíduo e no k-ésimo grupo é dada por

$$\begin{aligned} L_{ik}(d_{ik}; \theta, w_k) &\propto \Big[h_0^{(T)}(y_{ik}) \exp(x_{ik}^{(T)}\beta^{(T)} + w_k)\Big]^{\delta_{ik}^{(T)}} \\ &\times \exp\Big\{ - H_0^{(T)}(y_{ik}) \exp(x_{ik}^{(T)}\beta^{(T)} + w_k)\Big\} \\ &\times \Big[h_0^{(C)}(y_{ik}) \exp(x_{ik}^{(C)}\beta^{(C)} + \alpha w_k)\Big]^{\delta_{ik}^{(C)}} \\ &\times \exp\Big\{ - H_0^{(C)}(y_{ik}) \exp(x_{ik}^{(C)}\beta^{(C)} + \alpha w_k)\Big\}, \end{aligned} \tag{2.1}$$

em que $d_{ik} = (\delta_{ik}^{(T)}, \delta_{ik}^{(C)}, x_{ik}^{(T)}, x_{ik}^{(C)})$, a variável aleatória transformada $W = \log(Z)$ tem distribuição Normal com média zero, Normal$(0, \tau)$, em que $\tau = \frac{1}{\sigma^2}$. Para modelar a distribuição do efeito aleatório $Z \sim Log{-}Normal(0, \tau)$, onde τ é chamado por parâmetro de precisão, com função densidade dada por

$$f(z) = \frac{\sqrt{\tau}}{\sqrt{2\pi z}} \exp\left\{\frac{-\tau(\log(z))^2}{2}\right\}.$$

Podemos perceber que as função taxa de falha $h_0^{(T)}(y)$ e $h_0^{(C)}(y)$ e as funções taxa de falha acumuladas $H_0^{(T)}(y)$ e $H_0^{(C)}(y)$ estão definidas de forma genérica na equação 2.1. Portanto, basta especificar a distribuição de probabilidade dos tempos e o modelo estará definido.

2.5.2 Funções de taxa de falha para o Modelo Weibull

Se os tempos de falha e censura, T e C, possuem distribuição Weibull, com parâmetro denotados por $\kappa^{(T)}$ e $\kappa^{(C)}$, respectivamente, e escala denotados por $\gamma^{(T)} e \gamma^{(C)}$. As funções densidade são dada por

$$f_T(y) = \kappa^{(T)} y^{\kappa^{(T)}-1} \gamma^{(T)} \exp(-\gamma^{(T)} y^{\kappa^{(T)}}),$$

$$f_C(y) = \kappa^{(C)} y^{\kappa^{(C)}-1} \gamma^{(C)} \exp(-\gamma^{(C)} y^{\kappa^{(C)}}).$$

Consequentemente, as funções de taxa de falha de base são dadas por

$$h_0^{(T)}(y) = \kappa^{(T)} y^{\kappa^{(T)}-1} \gamma^{(T)}$$
$$h_0^{(C)}(y) = \kappa^{(C)} y^{\kappa^{(C)}-1} \gamma^{(C)}.$$

A distribuição Weibull possui apenas dois parâmetros e é capaz de acomodar funções de taxa de falha estritamente crescentes, decrescentes e constantes.

2.5.3 Funções de taxa de falha para o Modelo Exponencial por Partes

Se os tempos de falha e censura, T e C, possuem distribuição Exponencial por Partes, a função densidade de probabilidade é dada por

$$f(t|\lambda, p) = \lambda_j \exp\left\{ -\sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \right\}$$

em que λ_j é a taxa de falha no j-ésimo intervalo.

Seja $\rho^{(T)}$ a grade de tempos que particiona o eixo dos tempos de falha em b intervalos, sendo a partição finita do eixo do tempo de falha T dada por $0 < s_1 < s_2 < \cdots < s_b < \infty$, com $s_b > t$, em que t é o tempo de falha observado, com b intervalos. Analogamente, seja $\rho^{(C)}$ a grade de tempos que particiona o eixo do tempo até a censura em d intervalos, sendo a partição finita do eixo do tempo de censura C dada por $0 < c_1 < c_2 < \cdots < c_d < \infty$, com $c_d > c$, em que c é o tempo de censura observado, com d intervalos.

Consequentemente, as funções de taxa de falha de base para os tempos de falha e censura são dadas, respectivamente, por

$$h_0^{(T)}(y) = \lambda_j^{(T)},\ y \in I_j^{(T)} = (s_{j-1}, s_j],\ j = 1, \ldots, b$$
$$h_0^{(C)}(y) = \lambda_l^{(C)},\ y \in I_l^{(C)} = (c_{j-1}, c_j],\ l = 1, \ldots, d.$$

A distribuição Exponencial por Partes é bastante flexível para ajustar qualquer função taxa de falha, pois apresenta uma característica não-paramétrica a medida que o número de intervalos aumenta.

3 Implementação computacional

O *software* de programação R (R Core Team, 2022) é uma linguagem aberta (*open source*) utilizada por muitos estatísticos. Nela, podemos realizar problemas simples como obter a média de uma coluna de variáveis, através de funções como *mean*(), e problemas mais complexos como a criação de um app em Shiny. Através da linguagem R é possível montar uma interface gráfica chamada Shiny, na qual é apresentado resultados relativos às análises utilizadas de acordo com a demanda do usuário. Isso ajuda na comunicação dos dados, de forma mais visual e intuitiva para que a população em geral possa compreender cada item.

Nesta Seção iremos apresentar a forma como operacionalizamos a implementação do Shiny, através dos pacotes *shiny* (Chang et al., 2022) e *shinydashboards* (Chang e Borges Ribeiro, 2021), utilizando o software R. Também iremos demonstrar como aplicamos os métodos da Análise de Sobrevivência, vistos anteriormente, com a utilização de pacotes como *survival*, para regressão de Cox, *survminer* para os gráficos estimados por Kaplan-Meier e *DepCens* para modelos de censura dependente.

3.1 Shiny

Shiny é uma interface gráfica e pacote desenvolvido por Chang et al. (2022) do software R. Sua utilidade está em sua reatividade, pois no aplicativo Shiny temos a possibilidade de atualizamos as variáveis de acordo com o que buscamos visualizar, por exemplo, podemos determinar um tipo específico de câncer. Portanto, no aplicativo (SurvControl, `https://ggrandemagne.shinyapps.io/SurvControl2/`) será mostrado apenas o *output* filtrado nessas condições.

No Shiny, dividimos as funções em dois componentes, uma nomeada por interface de usuário (UI) e uma função de servidor (*server*), que são passados como argumentos para a função *shinyApp*, criando assim um objeto Shiny a partir das informações fornecidas nos algoritmos referentes a interface do usuário e servidor (Lisa, 2022). Podemos dividir a UI em diferentes partes, uma delas é a *header*, que cria um espaço de cabeçalho onde é possível colocar objetos como nomes e mensagens. Outra parte

é a *sidebar*, que cria um espaço na lateral esquerda do programa e funciona como uma divisão de páginas ou projetos, como um navegador da internet com diversas abas abertas. O mecanismo mais importante da interface de usuário é chamado de *body*, é onde preparamos toda a interface para receber as funções do *server* através da entrada do usuário (Chang et al., 2022). Algumas entradas (*inputs*) são:

- textInput: Adiciona uma caixa de texto.
- DateInput: Adiciona um input de data.
- SliderInput: Adiciona uma escolha de número.
- RadioButtons: Adiciona uma escolha de botões.
- CheckBox: Adiciona uma escolha de caixas.

O pacote *shinydashboard*, desenvolvido por Chang e Borges Ribeiro (2021), nos disponibiliza uma série de funções pré-programadas que são projetadas para facilitar a criação dessa interface em forma de *dashboard* HTML, através de funções como:

- dashboardHeader(): utilizada para customizar o cabeçalho do aplicativo.
- dashboardSidebar(): que serve como um menu de navegação para o usuário, assim como um local para inserir os inputs reativos globais.
- dashboardBody(): similarmente a função body do pacote base Shiny, esta funciona como uma interface receptora das entradas do usuário.
- dashboardPage(): utilizada como um canvas de cada seção a ser apresentada no painel.

O pacote também incluí diversos componentes de interface, como caixas de informações, valores e paineis que podem ser utilizados para criar conteúdos dinâmicos e interativos. Para informações mais detalhadas de como utilizar o *shinydashboard* visite o site <https://rstudio.github.io/shinydashboard/>.

3.2 Pacotes estatísticos

Podemos utilizar diversos outros pacotes para que o aplicativo seja realizado. Os pacotes mais comuns utilizados para a visualização de dados é o *ggplot2* (Wickham, 2016), que gera gráficos de fácil visualização e interatividade. Outros pacotes, como o *dplyr* (Wickham et al., 2022a), foram utilizados para manusear bancos de dados internamente no aplicativo.

O pacote *tidyverse* (Wickham et al., 2019) é um compilado de *libraries* no R, pacotes como *dplyr* e *ggplot2* são encontrados nele. Aqui, podemos facilmente realizar a proposta de filtragem do banco na UI, além de alterações para que o banco se encaixe na forma desejada pela função.

Em relação aos pacotes utilizados como ferramenta para dados de sobrevivência temos principalmente o pacote *survival* (Therneau, 2022), onde podemos fazer análises descritivas, modelos de tempo de vida acelerado, modelo de Cox, modelos paramétricos e outros. As principais funções do pacote *survival* que utilizamos foram

- Surv(): Cria um objeto de classe 'survival'.
- survfit(): Ajusta uma curva de sobrevivência utilizando a fórmula fornecida pelo usuário.
- coxph(): Ajusta um modelo de regressão de Cox para risco proporcionais
- cox.zph(): Testa as presunções de riscos proporcionais de um modelo de regressão de Cox.

Através do pacote *survminer* (Kassambara et al., 2021) nós aplicamos a função *ggsurvplot*, que utiliza operações do pacote *ggplot2* para apresentar o gráfico da curva de sobrevivência ou taxa de falha acumulada, utilizando os estimadores de Kaplan-Meier através dos valores disponibilizados pela rotina do pacote *survival*.

Desenvolvemos o pacote *DepCens* utilizando o pacote de desenvolvimento *devtools* (Wickham et al., 2022b), com o intuito de disponibilizar novas funções para criar e implementar um novo modelo para análises de dados de sobrevivência com censura dependente. Posteriormente, nós desenvolvemos outras funções em R para apresentar um sumário, critérios de informação e os gráficos para as distribuições Weibull e Exponencial por partes, com o objetivo de facilitar a análise de dados de sobrevivência com censura dependente.

Esses modelos disponíveis no pacote *DepCens* são baseados em extenções dos modelos de fragilidade, capazes de acomodar a dependencia entre o tempo de falha e censura, utilizando as distribuições marginais Weibull ou Exponencial por partes. As funções utilizadas do pacote são

- dependent.censoring(): Ajusta os dados de sobrevivência com censura dependente, também pode ser utilizado para incluir informações sobre a censura informativa.
- summary_dc(): Disponibiliza um sumário contendo as informações disponibilizadas pelo pacote, como p-valor, critérios de informação e afins.

- plot_dc(): Cria um gráfico para visualizar a função de sobrevivência do modelo ajustado.

Ainda sobre o pacote *DepCens*, de nossa autoria, podemos comentar que o mesmo se encontra atualmente na lista de pacotes oficiais da linguágem R, ou seja, o mesmo pode ser baixado através da função *install.packages("DepCens")*.

Os códigos do Painel de Controle apresentado neste trabalho podem ser acessados através do caminho https://github.com/GabrielGrandemagne/SurvControl. Assim como, o App está hospedado em *shinyapps.io* através do link https://ggrandemagne.shinyapps.io/SurvControl2/.

4 Aplicação computacional

Neste Capítulo descrevemos como aplicamos computacionalmente os modelos listados no Capítulo Metodologia, assim como as ferramentas utilizadas no desenvolvimento dos pacotes estatísticos e uma breve introdução ao *Shiny Dashboards*.

Nosso objetivo é a construção de uma dashboard utilizando o pacote *Shiny*, do software R, que utilize técnicas de Análise de Sobrevivência no banco disponibilizado pela FOSP. Este aplicativo tem como fim servir de painel de controle para a área administrativa hospitalar de diferentes especialidades. Mais especificamente, temos como objetivo:

- Desenvolver recursos gráficos e disponibilizá-los com os resultados da pesquisa, através dos pacotes de visualização como o survival, survminer e ggplot2.
- Manipular os dados conforme necessário, para viabilizar a utilização dos pacotes de análise/visualização de dados.
- Utilizar modelos de regressão com censura informativa para dados de sobrevivência multivariados, com o intuito de prever a sobrevida dos pacientes da base de dados.

Os dados utilizados neste trabalho são referentes ao banco de dados RHC (Registro Hospitalar de Câncer) da FOSP, disponível publicamente por Oncocentro (2021), que foram selecionados porque incluem informações referentes a data de diagnóstico e última data informada do paciente sendo então possível fazer o delineamento necessário para a análise de sobrevivência, contendo apenas as observações registradas de casos de câncer de mama, ovário e próstata a partir de 2014 até 2022. Este banco de dados fornecido pela FOSP é atualizado periodicamente, a cada três meses, sendo possível assim adaptar o painel de controle futuramente como um painel de monitoramento, atualizando os dados conforme eles são disponibilizados. A FOSP (2022) destaca alguns pontos sobre seus dados, são eles (Oncocentro, 2021):

- O início do Registro Hospitalar de Câncer (RHC) é em 01/01/2000, cadastrando novos casos diagnostigados a partir desta data até o presente momento;

- Segundo a FOSP (2022), os dados coletados pelo RHC não podem ser utilizados para cálculo de incidência, uma vez que retratam apenas o perfil de atendimento de uma determinada instituição;
- Todos os indivíduos presentes na base de dados não são identificados.

4.1 Manipulação

O banco de dados utilizado neste trabalho contém uma partição do banco de dados do RHC (Registro Hospitalar de Câncer) fornecidos pela FOSP (2022), contendo observações de indivíduos do estudo a partir de 2014 até 2022, com 3 tipos de câncer, sendo eles câncer de mama, ovário e próstata, o que totaliza em uma base de dados contendo 17.647 indivíduos.

Além disso, utilizando o *software R* (R Core Team, 2022), com o intuito de obter uma base de dados adaptada ao *Shiny Dashboard* para ajuste dos devidos modelos de sobrevivência, o banco de dados passou por diversos filtros que iremos apresentar a seguir. Entre eles, foi realizado uma seleção de variáveis, que são as seguintes:

Tabela 4.1: Covariáveis selecionadas

Lista das variáveis selecionadas para o estudo	
Variável	Descrição
ESCOLARI	Código para escolaridade do paciente
SEXO	Sexo do paciente
IDADE	Idade do paciente
CATEATEND	Categoria de atendimento ao diagnóstico
CLINICA	Código da clínica
DTDIAG	Data do diagnóstico
DTTRAT	Data de início do tratamento
TOPOGRUP	Grupo da topografia
T	Classificação TNM - T
N	Classificação TNM - N
P	Classificação TNM - P
TRATAMENTO	Código de combinação dos tratamentos realizados
DTULTINFO	Data da última informação do paciente
ULTINFO	Última informação sobre o paciente
ANODIAG	Ano de diagnóstico
PERDASEG	Perda de seguimento

Para a criação da variável indicadora de censura, utilizamos as informações disponibilizadas pela variável *ULTINFO*. Portanto, para a criação das variáveis $\delta^{(C)}$

e $\delta^{(T)}$ são utilizadas as indicadoras: $\delta^{(T)} = 1$, se o paciente teve óbito por câncer; $\delta^{(C)} = 1$, se o paciente teve óbito outras causas.

Para a variável do tempo observável, utilizamos a diferença entre a data da última informação do paciente ($DTULTINFO$) e a data de início do tratamento ($DTTRAT$). Esse tempo foi dividido por 360, para que o tempo seja considerado anos, ao invés de dias.

Em relação às covariáveis do modelo, nós primeiramente criamos a variável referente ao estágio do câncer, seguindo o estadiamento clínico (cTNM) 8ª Edição AJCC (American Joint Comission on Cancer) - 2017 (Gomes et al., 2017), onde a partir das variáveis T, N e M nós obtivemos a informação do estágio do paciente, sendo ele de estágio 0 até o máximo 4.

ESTADIAMENTO CLÍNICO (cTNM) | 8ª EDIÇÃO AJCC – 2017

ESTADIO	T	N	M
0	Tis	N0	M0
IA	T1a	N0	M0
IB	T1b ou T2a	N0	M0
IIA	T2b ou T3a	N0	M0
IIB	T3b ou T4a	N0	M0
IIC	T4b	N0	M0
III	Qualquer T	≥ N1	M0
IV	Qualquer T	Qualquer N	M1

Figura 4.1: Estadiamento clínico (cTNM)

Também criamos uma variável derivada de estágio, chamada *ESTAGIOt*, com seu valor associado 1 = Estágio 4, 0 = caso contrário, para avaliar casos específicos (cancêr de ovário) onde está presente na base apenas indivíduos no estágio 4 ou 0, como apresentamos na Tabela 4.5.

Nós também optamos por agregar alguns níveis da variável referente a escolaridade do paciente, com apenas 4 níveis, sendo eles: até ensino fundamental incompleto, ensino médio completo, ensino superior incompleto e não informado.

Além disso, modificamos a variável relacionada à Categoria de Atendimento (CATEATEND), sendo categoria de referência 0 = Particular ou Convênio (não-SUS) e 1 = SUS.

Com relação a variável de tratamento, nós optamos por utilizá-la de forma di-

cotômica com valor de referência 1 = Quimioterapia e 0 = Outro(s) para fins de simplificação e aplicabilidade dos modelos disponíveis através do *DepCens*.

Então, em nossas análises utilizamos as seguintes informações:

1. tempo: tempo observável, é a diferença entre a data da última informação do paciente e a data de início do tratamento,
2. delta_t: variável indicadora de falha (1 = Óbito por câncer, 0 = caso contrário),
3. delta_c: variável indicadora de censura (1 = Óbito por outra causa, 0 = caso contrário),
4. IDADE: idade do paciente,
5. ESTAGIO: variável indicadora de estágio (0 à 4),
6. ESTAGIOt: variável indicadora de estágio 4 (1 = estágio 4, 0 = caso contrário),
7. CATEATEND: variável indicadora de tratamento SUS (1 = SUS, 0 = caso contrário),
8. ESCOLARI: escolaridade do paciente (1 = Ensino Fundamental, 2 = Ensino Médio, 3 = Ensino Superior, 4 = Não informado),
9. SEXO: sexo do paciente (0 = Feminino, 1 = Masculino),
10. TRATAMENTO: variável indicadora de tratamento quimioterapia (1 = Utilizou quimioterapia no tratamento, 0 = caso contrário).

Também vale ressaltar que utilizamos a variável correspondente ao código da clínica (*CLINICA*) como variável de *cluster* para os modelos utilizados no pacote *DepCens*. Dada as formatações de variáveis descritas acima, nosso banco de dados (contendo apenas os três tipos de câncer) é resumido pela seguinte tabela descritiva:

Tabela 4.2: Tabela descritiva para o banco completo.

	Vivo	Óbito por câncer	Óbito por outras causas
	(N=13289)	(N=2797)	(N=1561)
IDADE			
Média (DP)	59.5 (12.2)	60.8 (14.1)	67.9 (12.3)
Mediana [q25, q75]	61 [51, 68]	61 [51, 72]	69 [61, 77]
SEXO			
Masculino	5786 (43.5%)	875 (31.3%)	830 (53.2%)
Feminino	7503 (56.5%)	1922 (68.7%)	731 (46.8%)
ESCOLARI			
Fundamental Incompleto	3258 (24.5%)	985 (35.2%)	684 (43.8%)
Ensino Médio	4420 (33.3%)	917 (32.8%)	502 (32.2%)
Ensino Superior	1825 (13.7%)	176 (6.3%)	63 (4.0%)
Ignorada	3786 (28.5%)	719 (25.7%)	312 (20.0%)
CATEATEND			
Convênio	2621 (19.7%)	142 (5.1%)	111 (7.1%)
SUS	10038 (75.5%)	2645 (94.6%)	1437 (92.1%)
Particular	630 (4.7%)	10 (0.4%)	13 (0.8%)
TRATAMENTO			
Cirurgia	3690 (27.8%)	177 (6.3%)	250 (16.0%)
Radioterapia	779 (5.9%)	68 (2.4%)	143 (9.2%)
Quimioterapia	361 (2.7%)	411 (14.7%)	101 (6.5%)
Cirurgia + Radio	568 (4.3%)	32 (1.1%)	63 (4.0%)
Cirurgia + Quimio	1089 (8.2%)	377 (13.5%)	100 (6.4%)
Radio + Quimio	86 (0.6%)	93 (3.3%)	19 (1.2%)
Cirurgia + Radio + Quimio	683 (5.1%)	183 (6.5%)	62 (4.0%)
Cirurgia + Radio + Quimio + Hormonio	1608 (12.1%)	244 (8.7%)	68 (4.4%)
Outras	4425 (33.3%)	1212 (43.3%)	755 (48.4%)
ESTAGIO			
0	7529 (56.7%)	726 (26.0%)	775 (49.6%)
1	1396 (10.5%)	57 (2.0%)	149 (9.5%)
2	1379 (10.4%)	106 (3.8%)	168 (10.8%)
3	2598 (19.6%)	748 (26.7%)	294 (18.8%)
4	387 (2.9%)	1160 (41.5%)	175 (11.2%)

4.2 Análises do Painel de Controle

Nesta Seção iremos pontuar as abas disponíveis até o momento na dashboard, onde as diversas abas são referentes às diferentes análises e pacotes aplicados no banco de dados.

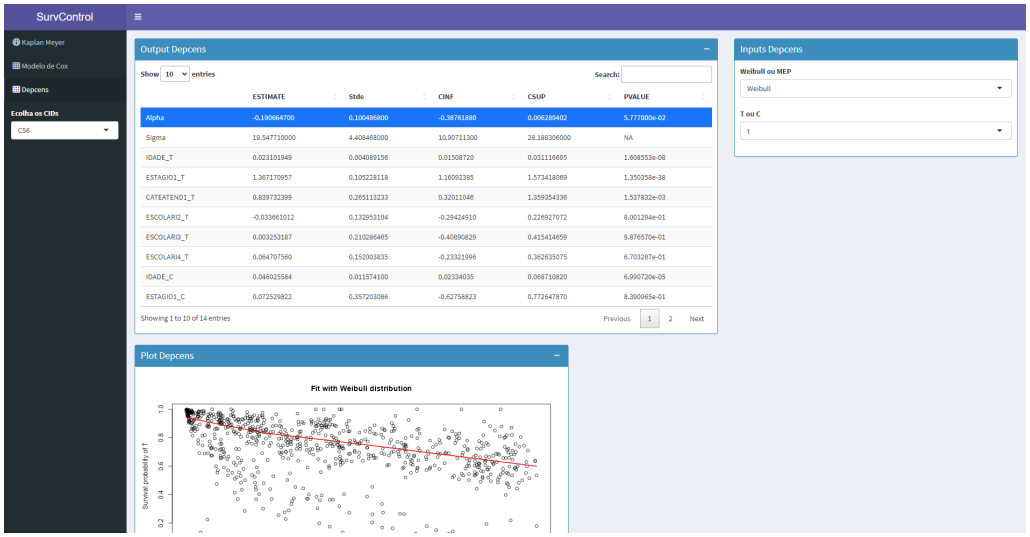

Figura 4.2: Aplicativo SurvControl

A Figura 4.2 mostra como as abas estão situadas no aplicativo, cada uma com rotinas e funções diferentes correspondente as análises da aba.

Primeiramente, para fins de avaliativos, utilizamos diferentes métodos com o intuito de comparar graficamente as curvas de sobrevivência. Utilizamos primeiramente o estimador de Kaplan-Meier, onde fora operacionalizado a criação do gráfico da curva de sobrevivência e de risco acumulativo utilizando a função *ggsurvplot* do pacote *survminer* para que seja possível a visualização do comportamento da curva de sobrevivência estratificando pelos níveis das demais covariáveis.

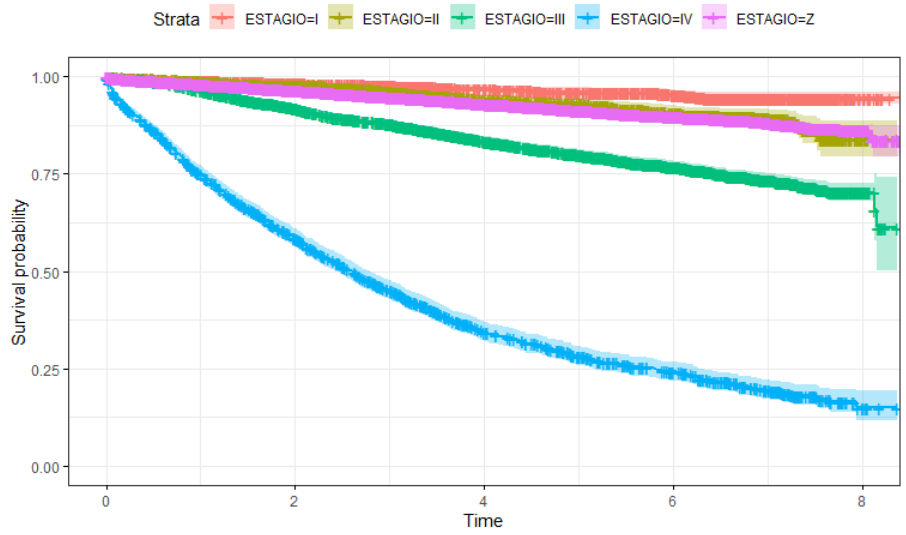

Figura 4.3: Função de sobrevivência utilizando o estimador de Kaplan-Meier

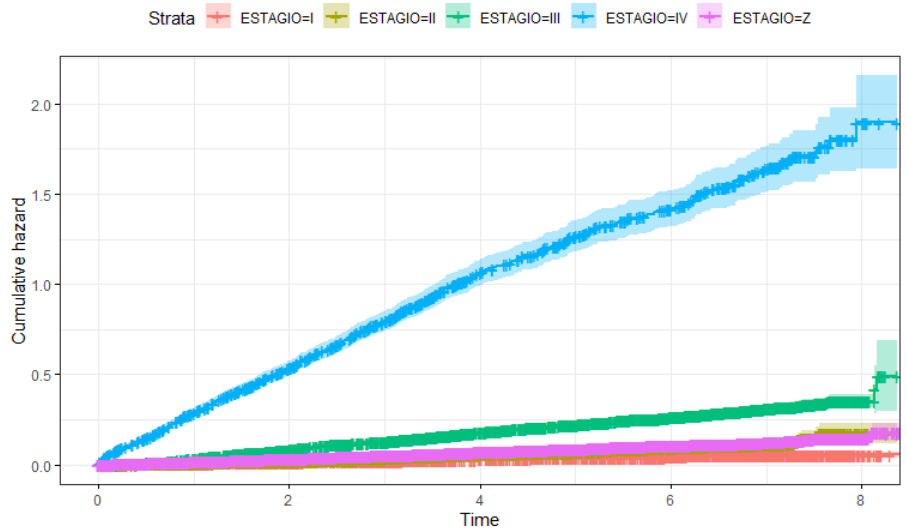

Figura 4.4: Função taxa de falha acumulada utilizando o estimador de Kaplan-Meier

No aplicativo *SurvControl* nós criamos uma variável reativa de *input* onde o usuário do painel pode escolher qual a variável estratificadora a ser apresentada nos gráficos Kaplan-Meier. A variável estratificadora da Figura (4.3) e Figura (4.4) é o Estágio, que indica o avanço da doença, logo faz sentido os pacientes em estágios mais avançados terem uma sobrevida menor que os demais.

A segunda aba do Painel de Controle contém a aplicação do modelo de regressão de Cox. O usuário pode escolher o tipo de câncer a ser apresentado, que usaremos em conjunto para a análise na próxima Seção. Vale ressaltar que também adicionamos o teste de Schoenfeld para que o usuário seja capaz de verificar se o pressuposto de riscos proporcionais estão bem atendindos. Um exemplo de *output* que podemos ver no Painel de Controle na segunda aba é referente a Tabela 4.3.

Tabela 4.3: Modelo de regressão de Cox para o banco geral

Termo	Estimativa	Erro Padrão	Estatística	P-Valor
IDADE	0.0131	0.0017	7.7327	0.00
ESTAGIO1	-0.6998	0.1382	-5.0648	0.00
ESTAGIO2	-0.0436	0.1064	-0.4097	0.68
ESTAGIO3	0.6806	0.0553	12.2990	0.00
ESTAGIO4	2.5718	0.0482	53.3385	0.00
CATEATEND1	0.6621	0.0855	7.7422	0.00
ESCOLARI2	-0.1345	0.0473	-2.8453	0.00
ESCOLARI3	-0.3071	0.0846	-3.6312	0.00
ESCOLARI4	0.0106	0.0498	0.2122	0.83
SEXO1	0.1555	0.0521	2.9854	0.00
TRATAMENTO1	0.5168	0.0471	10.9722	0.00

A terceira e última aba do nosso Painel de Controle contém os *outputs* provindos do pacote *DepCens*, considerando as 1.561 observações censuradas por óbito devido à outros fatores como censura dependente. O modelo tem a capacidade de acomodar a dependência entre o tempo de falha e censura, utilizando a variável que mencionamos anteriormente $\delta^{(C)}$. Um exemplo de resultado a ser encontrado no painel é a Tabela 4.4, onde as colunas *CINF* e *CSUP* apresentam os limites superiores e inferiores, respectivamente, do intervalo de confiança.

Tabela 4.4: Tabela DepCens Weibull para o banco C50

IND	Nome	Estimativa	Erro Padrão	CINF	CSUP	P-Valor
	α	0.51	0.27	-0.02	1.05	0.06
	σ^2	8.11	2.68	2.85	13.37	
$\delta^{(T)}$	IDADE	0.01	0.00	0.00	0.01	0.00
	ESTAGIOI	-1.46	0.41	-2.27	-0.65	0.00
	ESTAGIOII	1.10	0.20	0.72	1.49	0.00
	ESTAGIOIII	1.17	0.08	1.02	1.32	0.00
	ESTAGIOIV	3.08	0.08	2.92	3.23	0.00
	CATEATEND1	0.62	0.11	0.41	0.84	0.00
	ESCOLARI2	-0.14	0.06	-0.27	-0.02	0.02
	ESCOLARI3	-0.43	0.11	-0.64	-0.22	0.00
	ESCOLARI4	0.03	0.07	-0.10	0.17	0.62
	SEXO1	-0.49	0.30	-1.08	0.11	0.11
	TRATAMENTO1	0.41	0.06	0.30	0.52	0.00
$\delta^{(C)}$	IDADE	0.05	0.00	0.04	0.06	0.00
	ESTAGIOI	-0.23	0.20	-0.62	0.16	0.24
	ESTAGIOII	0.23	0.27	-0.29	0.76	0.38
	ESTAGIOIII	0.28	0.09	0.11	0.45	0.00
	ESTAGIOIV	1.06	0.13	0.81	1.31	0.00
	CATEATEND1	0.16	0.13	-0.09	0.41	0.20
	ESCOLARI2	0.01	0.09	-0.17	0.19	0.94
	ESCOLARI3	-0.48	0.18	-0.83	-0.12	0.01
	ESCOLARI4	-0.22	0.11	-0.43	-0.00	0.05
	SEXO1	-0.70	0.36	-1.40	-0.00	0.05
	TRATAMENTO1	-0.14	0.09	-0.31	0.03	0.10

Como o α é positivo, podemos dizer que há evidências de dependência entre o tempo de falha e censura dependente, para os dados de sobrevivência contendo indivíduos com diagnóstico de câncer de mama (CID C50). Portanto, neste caso a associação entre o tempo até o óbito por cancer e o tempo até o óbito por outros fatores é positiva.

4.3 Análise câncer de ovário

Nesta Seção iremos discutir brevemente sobre o câncer de ovário (CID C56), assim como utilizar os métodos disponibilizados no Painel de Controle para fazer uma breve análise. O câncer é uma doença genética não transmissível, podendo ser causada ou agravada por fatores intrínsecos como idade ou sexo, e extrínsecos como tabagismo ou obesidade (Albrecht, 2011). Segundo OMS (2022), no ano de 2006 o câncer foi a segunda maior causa de óbito em países desenvolvidos e a terceira em

países em desenvolvimento. Segundo Instituto Nacional do Câncer (INCA, 2022), em 2022 no Brasil foram registrados 7.310 novos casos de câncer de ovário, o que consiste em 3% dos novos casos de câncer entre as mulheres e 3.921 óbitos.

Para esta análise, utilizamos o banco de dados fornecido pela FOSP (2022) com observações registradas a partir de 2014, contendo apenas indivíduos com caso positivo de câncer de ovário. Com isso o banco ficou com 774 indivíduos, com 323 falhas (óbitos por câncer de ovário), o que representa aproximadamente 41% dos dados. A seguir a Tabela 4.5 apresenta uma breve análise descritiva dos dados.

Tabela 4.5: Tabela descritiva para indivíduos com câncer de ovário.

	Vivo	Óbito por câncer	Óbito por outras causas
	(N=395)	(N=323)	(N=56)
IDADE			
Mean (SD)	50.5 (15.7)	59.2 (13.4)	61.4 (15.2)
Median [q25, q75]	52.0 [42.5, 61.0]	60.0 [51.0, 68.0]	63.5 [49.5, 75.3]
ESCOLARI			
Fundamental Incompleto	102 (25.8%)	108 (33.4%)	22 (39.3%)
Ensino Médio	141 (35.7%)	104 (32.2%)	18 (32.1%)
Ensino Superior	49 (12.4%)	30 (9.3%)	5 (8.9%)
Ignorada	103 (26.1%)	81 (25.1%)	11 (19.6%)
CATEATEND			
Convênio	86 (21.8%)	16 (5.0%)	4 (7.1%)
SUS	298 (75.4%)	307 (95.0%)	52 (92.9%)
Particular	11 (2.8%)	0 (0%)	0 (0%)
TRATAMENTO			
Cirurgia	148 (37.5%)	56 (17.3%)	16 (28.6%)
Radioterapia	1 (0.3%)	1 (0.3%)	0 (0%)
Quimioterapia	27 (6.8%)	77 (23.8%)	10 (17.9%)
Cirurgia + Radio	3 (0.8%)	0 (0%)	2 (3.6%)
Cirurgia + Quimio	192 (48.6%)	170 (52.6%)	20 (35.7%)
Radioterapia + Quimio	0 (0%)	3 (0.9%)	2 (3.6%)
Cirurgia + Radio + Quimio	4 (1.0%)	5 (1.5%)	1 (1.8%)
Cirurgia + Radio + Quimio + Hormonio	1 (0.3%)	0 (0%)	0 (0%)
Outras	19 (4.8%)	11 (3.4%)	5 (8.9%)
Nenhum	0 (0%)	0 (0%)	0 (0%)
ESTAGIO			
0	371 (93.9%)	179 (55.4%)	46 (82.1%)
4	24 (6.1%)	144 (44.6%)	10 (17.9%)

Em relação à questão computacional, como mencionado anteriormente, nós tivemos que modificar a variável de *ESTAGIO* para uma variável dicotômica com estágio 0 como categoria de referência, pois como este banco contém apenas indivíduos com câncer de ovário, como podemos ver na Tabela 4.5, só há indivíduos situados no estágio 0 ou 4 de câncer de ovário registrados na base de dados.

A Tabela 4.6 é o resultado da aplicação do modelo de regressão de Cox através da função *coxph()* do pacote *survival* (Therneau, 2022).

Tabela 4.6: Modelo de regressão de Cox para C56

Termo	Estimativa	Erro Padrão	Estatística	P-Valor
IDADE	0.0225	0.0047	4.7551	0.00
ESTAGIOt1	1.3087	0.1201	10.8922	0.00
CATEATEND1	0.8988	0.2636	3.4094	0.00
ESCOLARI2	-0.0107	0.1412	-0.0759	0.94
ESCOLARI3	0.0436	0.2107	0.2068	0.84
ESCOLARI4	0.1462	0.1491	0.9806	0.33
TRATAMENTO1	0.2108	0.1423	1.4814	0.14

Através da função *cox.zph()* nós percebemos que a suposição de taxas de falha proporcionais não é bem atendida (p-valor global do teste de Schoenfeld < 0.05), o que é um indicativo de que devemos utilizar algum outro modelo capaz de acomodar esses dados.

Logo, a Tabela 4.7 corresponde ao resultado do ajuste com o modelo de censura dependente, utilizando a distribuição marginal exponencial por partes, que aplicamos através do pacote *DepCens* para avaliar se há alguma dependência entre o tempo de falha e de censura. O tempo de falha é o tempo devido ao óbito pelo cancer, tempo de censura dependente é o tempo até o óbito por outras causas. Aqui também vale destacar que utilizamos a variável correspondendo ao TRATAMENTO (como pode ser visto na Tabela 4.5) em sua forma dicotômica, com categoria de referência 1 = Quimioterapia, 0 = caso contrário, com a variável situada desta maneira podemos perceber que, a partir da Tabela 4.7, os indivíduos que fizeram o tratamento por quimioterapia possuem um risco de $\exp(0.19) \approx 1.21$ vezes o risco dos indivíduos que fizeram outros tipos de tratamento.

Tabela 4.7: Tabela DepCens MEP para C56.

IND	Nome	Estimativa	Erro Padrão	CINF	CSUP	P-Valor
	α	-0.20	0.09	-0.37	-0.0236	0.03
	σ^2	18.96	3.22	0.00	56.13	
$\delta^{(T)}$	IDADE	0.02	0.00	0.01	0.03	0.00
	ESTAGIOt1	1.35	0.12	1.10	1.59	0.00
	CATEATEND1	0.88	0.27	0.35	1.41	0.00
	ESCOLARI2	-0.02	0.15	-0.31	0.26	0.87
	ESCOLARI3	0.02	0.22	-0.41	0.45	0.93
	ESCOLARI4	0.05	0.15	-0.25	0.35	0.75
	TRATAMENTO1	0.19	0.14	-0.09	0.47	0.19
$\delta^{(C)}$	IDADE	0.05	0.01	0.02	0.07	0.00
	ESTAGIOt1	0.20	0.37	-0.52	0.92	0.58
	CATEATEND1	0.76	0.54	-0.29	1.82	0.16
	ESCOLARI2	-0.11	0.33	-0.75	0.53	0.74
	ESCOLARI3	-0.03	0.50	-1.02	0.95	0.95
	ESCOLARI4	-0.13	0.38	-0.87	0.61	0.73
	TRATAMENTO1	-0.39	0.28	-0.95	0.17	0.17

Na Tabela 4.7, observando o α significativo à 5% podemos dizer que há uma estrutura de dependência entre o tempo de falha (tempo até óbito pelo câncer de ovário) e censura dependente (tempo até óbito por outras causas), para os dados contendo apenas indivíduos com câncer de ovário. Também podemos apontar que, por exemplo, uma pessoa no estágio 4 tem um risco de ≈ 3.86 vezes o risco dos indivíduos no estágio 0, logo este também é considerado um fator de risco.

Ainda, com o auxílio do pacote *DepCens*, podemos mostrar a curva de sobrevivência em relação ao tempo de acompanhamento, dadas por:

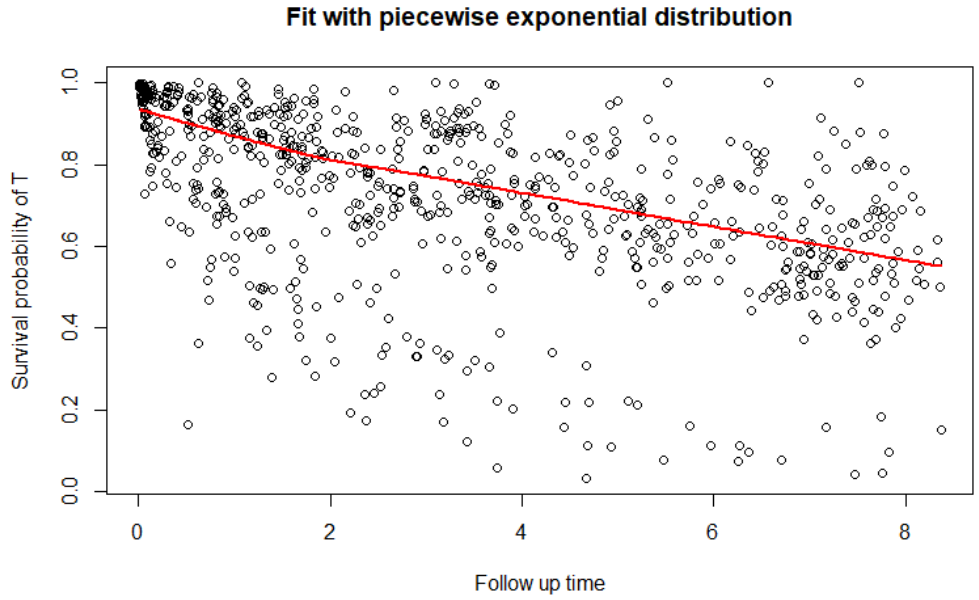

Figura 4.5: Função de sobrevivência utilizando a distribuição Exponencial por partes para T (tempo até óbito pelo câncer de ovário) (C56).

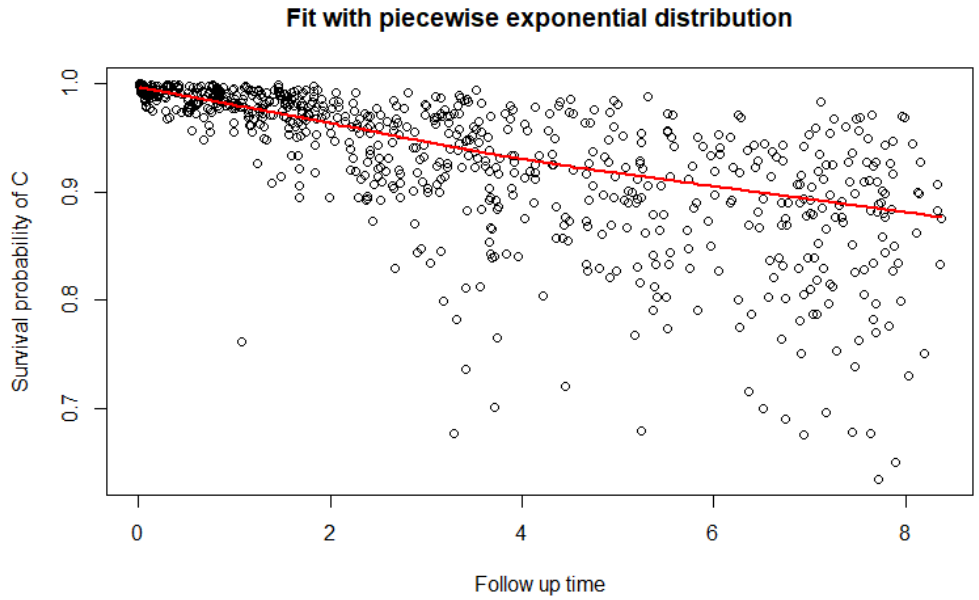

Figura 4.6: Função de sobrevivência utilizando a distribuição Exponencial por partes para C (tempo até óbito por outras causas) (C56).

4.4 Análise câncer de mama

Entre as mulheres, a neoplastia de mama é a mais incidente. Segundo Instituto Nacional do Câncer (INCA, 2022), em 2022 no Brasil foram registrados 73.610 casos novos de neoplastia da mama, o que consiste em 30.1% dos novos casos de câncer

entre as mulheres, e 17.825 óbitos, com o risco estimado de desenvolver a doença no estado do Rio Grande do Sul de 37 casos a cada 100 mil mulheres, porém esse risco varia entre regiões, o que torna evidente a necessidade de estratégias governamentais que priorizem a detecção precoce da doença em determinados estados (Albrecht, 2011).

Para esta análise utilizamos o banco de dados fornecido pela FOSP (2022) com observações registradas a partir de 2014 até 2022, contendo apenas indivíduos com caso positivo de câncer de mama. Com isso o banco ficou com 9429 indivíduos, com 1610 falhas por câncer de mama (óbitos por câncer), o que representa 17% dos dados. A seguir a Tabela 4.8 apresenta uma breve análise descritiva dos dados.

Tabela 4.8: Tabela descritiva para indivíduos com câncer de mama.

	Vivo	Óbito por câncer	Óbito por outras causas
	(N=7136)	(N=1610)	(N=683)
IDADE			
Mean (SD)	55.4 (12.6)	56.3 (14.2)	64.4 (14.8)
Median [q25, q75]	55.0 [46.0, 64.0]	55.5 [46.0, 65.0]	65.0 [54.0, 76.0]
SEXO			
Masculino	28 (0.4%)	11 (0.7%)	8 (1.2%)
Feminino	7108 (99.6%)	1599 (99.3%)	675 (98.8%)
ESCOLARI			
Fundamental Incompleto	1713 (24.0%)	523 (32.5%)	271 (39.7%)
Ensino Médio	2632 (36.9%)	584 (36.3%)	248 (36.3%)
Ensino Superior	987 (13.8%)	119 (7.4%)	38 (5.6%)
Ignorada	1804 (25.3%)	384 (23.9%)	126 (18.4%)
CATEATEND			
Convênio	1548 (21.7%)	97 (6.0%)	66 (9.7%)
SUS	5421 (76.0%)	1509 (93.7%)	609 (89.2%)
Particular	167 (2.3%)	4 (0.2%)	8 (1.2%)
TRATAMENTO			
Cirurgia	761 (10.7%)	42 (2.6%)	66 (9.7%)
Radioterapia	157 (2.2%)	25 (1.6%)	29 (4.2%)
Quimioterapia	299 (4.2%)	315 (19.6%)	80 (11.7%)
Cirurgia + Radio	286 (4.0%)	14 (0.9%)	32 (4.7%)
Cirurgia + Quimio	883 (12.4%)	196 (12.2%)	72 (10.5%)
Radioterapia + Quimio	51 (0.7%)	80 (5.0%)	9 (1.3%)
Cirurgia + Radio + Quimio	668 (9.4%)	171 (10.6%)	59 (8.6%)
Cirurgia + Radio + Quimio + Hormonio	1604 (22.5%)	230 (14.3%)	65 (9.5%)
Outras	2427 (34.0%)	537 (33.4%)	271 (39.7%)
ESTAGIO			
0	3885 (54.4%)	250 (15.5%)	292 (42.8%)
1	531 (7.4%)	6 (0.4%)	28 (4.1%)
2	86 (1.2%)	29 (1.8%)	15 (2.2%)
3	2471 (34.6%)	706 (43.9%)	266 (38.9%)
4	163 (2.3%)	619 (38.4%)	82 (12.0%)

Aplicamos o modelo de Cox através da função *coxph()* do pacote *survival* (Therneau, 2022) para o banco contendo indivíduos apenas com câncer de mama, obtivemos o resultado que apresentamos na Tabela 4.9.

Tabela 4.9: Modelo de regressão de Cox para C50

Termo	Estimativa	Erro Padrão	Estatística	PVALUE
IDADE	0.0073	0.0020	3.5603	0.00
ESTAGIO1	-1.4784	0.4134	-3.5764	0.00
ESTAGIO2	1.1423	0.1969	5.8031	0.00
ESTAGIO3	1.1791	0.0751	15.6904	0.00
ESTAGIO4	3.1361	0.0761	41.1847	0.00
CATEATEND1	0.5543	0.1051	5.2720	0.00
ESCOLARI2	-0.0995	0.0627	-1.5879	0.11
ESCOLARI3	-0.3610	0.1047	-3.4483	0.00
ESCOLARI4	0.0240	0.0686	0.3501	0.72
SEXO1	-0.4995	0.3039	-1.6436	0.10
TRATAMENTO1	0.4016	0.0556	7.2288	0.00

Desta vez, ao análisar a suposição de riscos proporcinais, percebemos que ela está bem atendida para a maioria das covariáveis, porém o teste global de Schoenfeld ainda rejeita a hipótese nula de riscos proporcionais (p-valor global do teste de Schoenfeld < 0.05).

Os resultados dos modelos de censura dependende ajustados pelo pacote *DepCens* para indivíduos com câncer de mama são apresentados na Tabela 4.10. Aqui, também vale destacar que utilizamos a variável correspondendo ao TRATAMENTO (como pode ser visto na Tabela 4.8) em sua forma dicotômica, com categoria de referência 1 = Quimioterapia, 0 = caso contrário, com a variável situada desta maneira podemos perceber que, a partir da Tabela 4.10, os indivíduos que fizeram o tratamento por quimioterapia possuem um risco de $\exp(0.42) \approx 1.52$ vezes o risco dos demais indivíduos que não fizeram tratamento por quimioterapia.

Tabela 4.10: Tabela DepCens MEP para C50.

IND	Nome	Estimativa	Erro Padrão	CINF	CSUP	P-Valor
	α	0.62	0.28	0.071	1.168	0.03
	σ^2	7.61	4.43	0.00	22.52	
$\delta^{(T)}$	IDADE	0.01	0.00	0.00	0.01	0.00
	ESTAGIOI	-1.46	0.41	-2.27	-0.65	0.00
	ESTAGIOII	1.10	0.20	0.72	1.49	0.00
	ESTAGIOIII	1.17	0.08	1.02	1.32	0.00
	ESTAGIOIV	3.08	0.08	2.92	3.23	0.00
	CATEATEND1	0.61	0.11	0.39	0.83	0.00
	ESCOLARI2	-0.14	0.06	-0.27	-0.02	0.03
	ESCOLARI3	-0.43	0.11	-0.64	-0.22	0.00
	ESCOLARI4	0.03	0.07	-0.11	0.16	0.69
	SEXO1	-0.51	0.31	-1.11	0.09	0.10
	TRATAMENTO1	0.42	0.06	0.30	0.53	0.00
$\delta^{(C)}$	IDADE	0.05	0.00	0.04	0.06	0.00
	ESTAGIOI	-0.24	0.20	-0.63	0.15	0.24
	ESTAGIOII	0.24	0.27	-0.29	0.76	0.38
	ESTAGIOIII	0.28	0.09	0.11	0.45	0.00
	ESTAGIOIV	1.05	0.13	0.80	1.30	0.00
	CATEATEND1	0.17	0.13	-0.08	0.42	0.18
	ESCOLARI2	0.02	0.09	-0.16	0.19	0.87
	ESCOLARI3	-0.47	0.18	-0.82	-0.12	0.01
	ESCOLARI4	-0.21	0.11	-0.43	0.01	0.06
	SEXO1	-0.60	0.37	-1.32	0.12	0.10
	TRATAMENTO1	-0.14	0.09	-0.31	0.03	0.10

Na Tabela 4.10, observando o α significativo podemos dizer que há uma estrutura de dependência entre o tempo de falha e censura para os dados contendo apenas indivíduos com câncer de mama, onde o tempo de falha é o tempo até óbito por câncer de mama e o tempo de censura dependente é o tempo até óbito por outras causas. Também podemos apontar que, por exemplo, um indivíduo que fez seu tratamento pelo SUS (valor de referência para a variável *CATEATEND*) tem risco relativo ≈ 1.8 vezes o risco dos demais indivíduos que fizeram tratamento através do convênio ou particular.

Ainda com o auxílio do pacote *DepCens*, podemos plotar a curva de sobrevivência em relação ao tempo, dadas por:

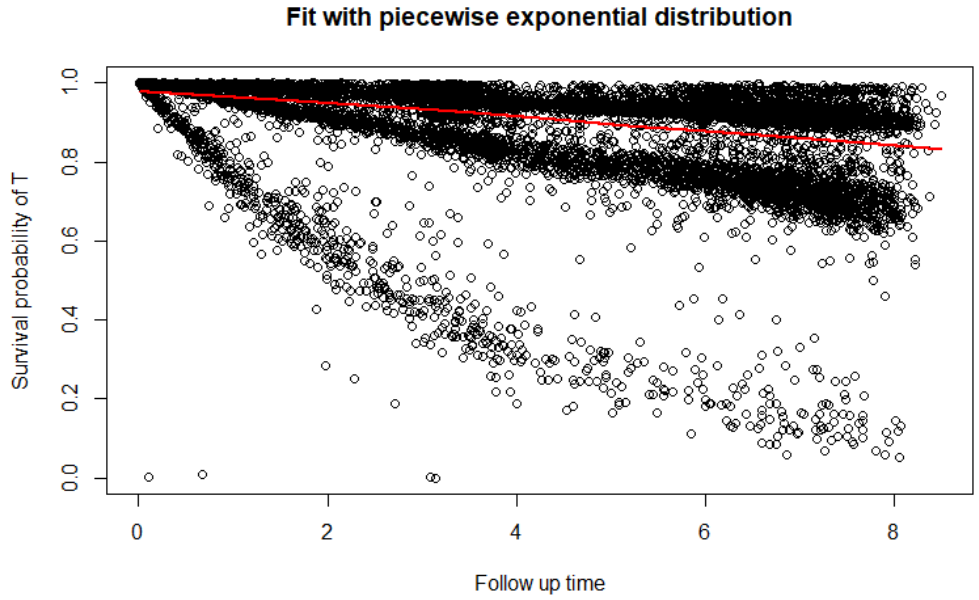

Figura 4.7: Função de sobrevivência utilizando a distribuição Exponencial por partes para T (tempo até óbito pelo câncer de mama) (C50).

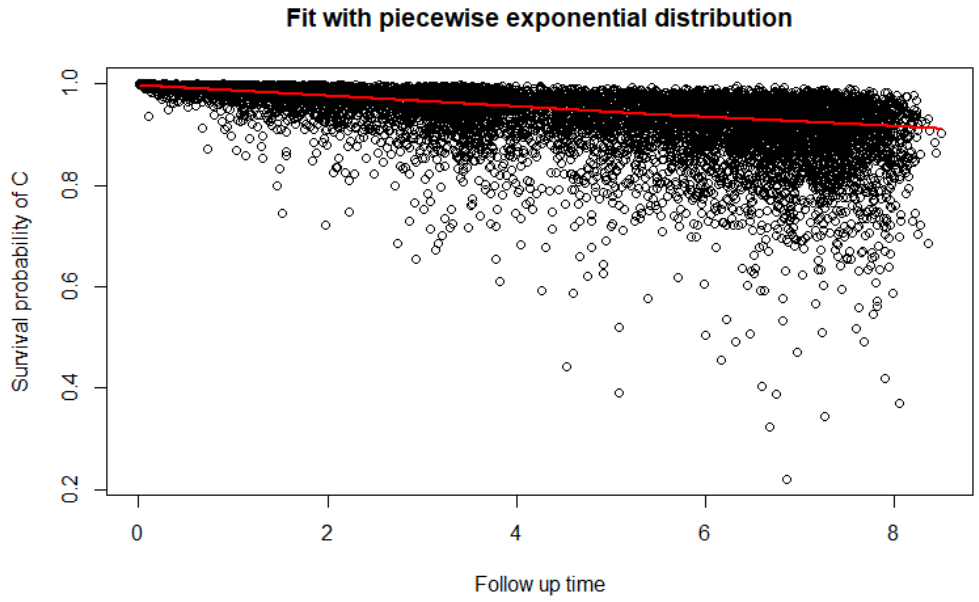

Figura 4.8: Função de sobrevivência utilizando a distribuição Exponencial por partes para C (tempo até óbito por outras causas) (C50).

Para complementar nossa análise, em nosso aplicativo podemos selecionar ambos os dados de câncer de ovário e câncer de mama em conjunto e gerar outputs a partir do pacote *DepCens*, assim como foi apresentado na seção de análise para cada tipo de cancêr. Então, para fins comparativos iremos agregamos pacientes com câncer de mama e câncer de ovário na mesma base, utilizando o pacote em R *DepCens* que

é capaz de utilizar duas variáveis indicadoras de óbitos para cada tipo de câncer. A Tabela 4.11 contém apenas os indivíduos diagnosticados com câncer de ovário ou câncer de mama onde agora nossa variável indicadora de falha foi estratificada por causa do óbito, se foi óbito devido a câncer de mama ou de ovário. O *output* para essa rotina é dada por:

Tabela 4.11: Tabela DepCens Weibull para C50 e C56.

Óbito	Nome	Estimativa	Erro Padrão	CINF	CSUP	P-Valor
	α	-6.42	1.06	-8.49	-4.34	0.00
	σ^2	1.66	1.97	0.00	5.52	
C50	IDADE	0.01	0.00	0.00	0.01	0.00
	ESTAGIOt1	2.38	0.07	2.25	2.51	0.00
	CATEATEND1	0.73	0.11	0.52	0.94	0.00
	ESCOLARI2	-0.12	0.07	-0.26	0.02	0.08
	ESCOLARI3	-0.42	0.13	-0.67	-0.16	0.00
	ESCOLARI4	0.08	0.07	-0.06	0.21	0.26
	TRATAMENTO1	0.55	0.05	0.44	0.65	0.00
C56	IDADE	0.02	0.00	0.01	0.03	0.00
	ESTAGIOt1	1.91	0.12	1.68	2.15	0.00
	CATEATEND1	0.90	0.27	0.36	1.43	0.00
	ESCOLARI2	-0.14	0.14	-0.42	0.14	0.33
	ESCOLARI3	0.13	0.21	-0.29	0.55	0.54
	ESCOLARI4	0.01	0.16	-0.30	0.32	0.94
	TRATAMENTO1	1.07	0.14	0.79	1.35	0.00

Na Tabela 4.11, observando o α significativo podemos dizer que há uma estrutura de dependência entre o tempo de falha por câncer de ovário e tempo de falha por câncer de mama.

Também podemos apontar que, por exemplo, um indivíduo que fez seu tratamento pelo SUS (variável CATEATEND1) tem risco ≈ 1.8 vezes o risco dos demais indivíduos que fizeram tratamento através do convênio ou particular para câncer de mama (CID C50), já para pacientes diagnosticados com câncer de ovário esse risco sobre para ≈ 2.91 vezes o risco dos indivíduos que fizeram tratamento por convênio.

5 Conclusão

Este trabalho teve como objetivo estudar, aplicar e desenvolver um aplicativo que utiliza o modelo de Cox (Cox, 1972), estimadores de Kaplan-Meier (Kaplan e Meier, 1958) e o modelo para censura dependente (Schneider et al., 2019) em uma partição dos dados de sobrevivência provindos da FOSP (2022). Para desenvolver o aplicativo utilizamos o software R (R Core Team, 2022) em conjunto com os pacotes *shiny* e *shinydashboard*, assim com os pacotes gráficos *ggplot2* e de análise de sobrevivência *survival* e *DepCens*.

Neste trabalho apresentamos a implementação dessas técnicas de modelagem para fazer uma breve análise da sobrevida de indivíduos com câncer de mama e de ovário. Para isso, utilizamos o modelo de Cox (Cox, 1972) e o modelo de censura dependente para análise de sobrevivência, proposto por Schneider et al. (2019), método que nos permite avaliar se há estrutura de dependência entre os tempos de falha e censura. Através deste método foi observado que o parâmetro da dependência α é significativo para quase todos os casos apresentados nesse trabalho. Por meio das Tabelas 4.7, 4.10 e 4.11 podemos apontar que há indícios de alguma estrutura de dependência entre o tempo até o óbito por câncer e tempo até o óbito por outros fatores, para estes dados, o que ocasiona em uma mudança na estimativa da curva de sobrevivência e taxa de falha para estes dados de sobrevivência.

De acordo com as estimativas dos modelos de censura dependente apresentados nas Tabelas anteriormente vistas, em geral, indivíduos atendidos pelo SUS possuem maior risco de falhar que os indivíduos atendidos por convênio ou privado. Quanto aos diferentes tipos de tratamentos, pacientes que foram submetidos à quimioterapia apresentaram maior risco de falhar que os demais pacientes.

A partir da Tabela 4.7 podemos perceber que, para indivíduos com câncer de ovário na base de dados da FOSP, há estrutura de dependência aparente entre o tempo de falha e censura. Também podemos apontar que, com base na amostra do banco RHC de 2014 até 2022 disponibilado pela FOSP (FOSP, 2022), foi encontrado que uma pessoa passar por tratamento pelo SUS por câncer de ovário tem um risco de óbito aproximadamente 2.41 vezes o risco de ter óbito se o indivíduos fez o

tratamento particular ou por convênio. Também podemos apontar que a variável indicadora de estágio (*ESTAGIOt*) é considerada um fator de risco, o que indica que indivíduos no estágio 4 da doença possuem um risco de aproximadamente 3.85 vezes o risco dos demais indivíduos.

A partir da Tabela 4.10 podemos novamente apontar que o parâmetro de dependência é significativo. Portanto, a relação entre o tempo até o óbito pelo câncer de mama e o tempo até o óbito por outros fatores é significativa. Também podemos mostrar que, através do ajuste do modelo para censura dependente, o indivíduo do estágio 1 da doença é aparentemente um fator protetor com relação ao estágio 0, o que não é intuitivo porém é um comportamento perceptível ao analisar a curva de sobrevivência utilizando os estimadores de Kaplan-Meier, estratificados por estágio apresentada na Figura 4.3. Pois, podemos perceber claramente que a curva relacionada ao estágio 0 (cor rosa) está em alguns momentos abaixo da curva relacionada ao estágio 1.

Com relação aos dados conjuntos de câncer de mama e câncer de ovário, podemos mencionar que, a partir da Tabela 4.11, temos resultados similares aos demais, com um α significativo podemos apontar que há indícios de dependência entre o tempo de falha por câncer de mama e o tempo de falha por câncer de ovário. Também podemos mencionar que as variáveis *IDADE*, *TRATAMENTO*, *ESTAGIOt* e *CATEATEND* são fatores de risco para ambos indivíduos com câncer de mama ou câncer de ovário.

Há diversos artigos na literatura que ressaltam a existência algum tipo de correlação genética entre câncer de mama e câncer de ovário. O trabalho de apresentado por Suszynska et al. (2019) apresenta evidências da existência de uma associação significativa com o aumento de risco para indivíduos com câncer de mama e ovário para diversos genes. Também podemos mencionar o trabalho de pesquisadores como Lerda et al. (2019), que ao analisar dados hereditários de câncer de ovário e mama, concluíram que indíviduos com histórico familiar desses tipos de câncer, possuem uma procentagem elevada de genes BRCA.

Nosso Painel de Controle contém atualmente apenas os métodos, ajustes e banco de dados vistos ou comentados neste trabalho. Porém, temos a intenção de estender este Painel de Controle, como implementar outras análises, distribuições e gráficos. Também pretendemos operacionalizar o aplicativo futuramente em formato *react*, para que seja possível fazer a implementação direta com a base de dados completa da FOSP, atualizando-a periodicamente ou quando necessário.

Referências Bibliográficas

Albrecht, C. A. M. (2011). Análise de sobrevida de pacientes com câncer de mama atendidas no hospital santa rita de cássia, na cidade de vitória, espírito santo.

Carvalho, M., Andreozzi, V., Codeço, C., Campos, D., Barbosa, M., e Shimakura, S. (2011). *Análise de sobrevivência: teoria e aplicações em saúde*. SciELO - Editora FIOCRUZ.

Chang, W. e Borges Ribeiro, B. (2021). *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.2.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., e Borges, B. (2022). *shiny: Web Application Framework for R*. R package version 1.7.4.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*, volume 1. São Paulo: Blücher.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

FOSP (2022). *Fundação Oncocentro de São Paulo: Banco de dados do rch.* https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/.

Gomes, E., Landman, G., Belfort, F., e Schmerling, R. (2017). Estadiamento do melanoma pela ajcc. *Melanoma. 76 (2017)*, pages 1–7.

Hanagal, D. D. (2011). Modeling survival data using frailty models.

INCA (2022). *Instituto Nacional de Câncer (INCA), avaliado em 03/23*.

Kalbfleisch, J. e Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kassambara, A., Kosinski, M., e Biecek, P. (2021). *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9.

Lerda, D., Pellicioni, P., Biaggi, M., Labrador, J., Illescas, E., Bella, S., Llugdar, J., e Cortes, M. (2019). Consejo genético y detección de vías moleculares en pacientes con cáncer hereditario. *Methodo Investigación Aplicada a las Ciencias Biológicas*, 4(3):71–80.

Lisa, D. (2022). *Building Web Apps with R Shiny*. https://github.com/debruine/shinyintro/.

OMS (2022). *Organização Mundial de Saúde*. Cancer control: knowledge into action. WHO guide for effective programmes. Geneva, WHO press, 2006.

Oncocentro, F. (2021). Banco de dados do rhc. disponível em: https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/. acesso em: 05 abr. 2023.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schneider, S. (2017). Modelos para dados de sobrevivÊncia multivariados com censura informativa.

Schneider, S., Demarqui, F. N., Colosimo, E. A., e Mayrink, V. D. (2019). An approach to model clustered survival data with dependent censoring. *Biometrical Journal*, 62(1).

Schneider, S. e Grandemagne dos Santos, G. (2022). *depcens: Dependent censoring regression models for survival multivariate data*. R package version 3.4-0.

Suszynska, M., Klonowska, K., Jasinska, A. J., e Kozlowski, P. (2019). Large-scale meta-analysis of mutations identified in panels of breast/ovarian cancer-related genes — providing evidence of cancer predisposition genes. *Gynecologic Oncology*, 153(2):452–462.

Therneau, T. M. (2022). *survival: A Package for Survival Analysis in R*. R package version 3.4-0.

Vaupel, J. W., Manton, K. G., e Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., e Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., François, R., Henry, L., e Müller, K. (2022a). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.

Wickham, H., Hester, J., Chang, W., e Bryan, J. (2022b). *devtools: Tools to Make Developing R Packages Easier*. R package version 2.4.5.