UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO
CURSO DE CIÊNCIAS ATUARIAIS

THALES DA SILVA SIQUEIRA

INTRODUÇÃO AO MODELO DE COX

FORTALEZA

2017

THALES DA SILVA SIQUEIRA

INTRODUÇÃO AO MODELO DE COX

Monografia apresentada ao Curso de Ciências Atuariais da Faculdade de Economia, Administração, Atuária e Contabilidade da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Orientadora: Prof.ª Luciana Moura Reinaldo

FORTALEZA

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca da Faculdade de Economia, Administração, Atuária e Contabilidade

S632i Siqueira, Thales da Silva.
Introdução ao Modelo de Cox / Thales da Silva Siqueira. – 2017.
35 f.: il.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Ciências Atuariais, Fortaleza, 2017.
Orientação: Profa. Luciana Moura Reinaldo.

1. Modelo de Cox. 2. Análise de Sobrevivência. 3. Modelo Semi-Paramétrico. I. Título.

CDD 368.01

THALES DA SILVA SIQUEIRA

INTRODUÇÃO AO MODELO DE COX

Monografia apresentada ao Curso de Ciências Atuariais da Faculdade de Economia, Administração, Atuária e Contabilidade da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Aprovada em: ___/___/_____.

BANCA EXAMINADORA

_____
Profª. Luciana Moura Reinaldo (Orientador)
Universidade Federal do Ceará (UFC)

_____
Profª. Me. Alana Katielli Azevedo de Macedo
Universidade Federal do Ceará (UFC)

_____
Prof. Me. Romulo Alves Soares
Universidade Estadual do Ceará (UFC)

AGRADECIMENTOS

Agradeço a Deus, pois sem ele eu não teria forças para essa longa jornada.
Aos meus pais, pelo carinho e apoio para que eu chegasse até esta etapa de minha vida.
Aos amigos e colegas, pelo incentivo e pelo apoio constante.
Agradeço também a todos os professores que me acompanharam durante a graduação, em especial à Prof.ª Luciana, responsável pela realização deste trabalho.

RESUMO

Um dos modelos mais aplicados na análise de sobrevivência é o modelo de regressão de Cox. Suas técnicas podem ser utilizadas na medicina, na engenharia, na análise de seguros e até mesmo para analisar o risco de um cliente se tornar inadimplente, ou seja, este modelo se diferencia dos demais por sua ampla aplicação em diversas áreas e pelo fato de ser um modelo semi-paramétrico, o que lhe permite fazer uso de covariáveis associadas aos indivíduos presente no estudo, como por exemplo, idade, sexo, doenças pré-existentes, renda, grau de escolaridade, local de residência, entre outras, para que desta forma possa modelar o efeito dessas covariáveis sobre o tempo de sobrevivência do indivíduo, que é o objetivo principal da análise de sobrevivência. O objetivo desta monografia foi realizar uma introdução sobre o modelo semi-paramétrico de regressão de Cox, incluindo os métodos de adequação e ajuste aos dados de sobrevivência, proporcionando um breve estudo sobre a análise de sobrevivência. A teoria estudada foi ilustrada com uma aplicação do modelo à um conjunto de dados de pacientes com câncer avançado de pulmão, onde foi analisado o tempo até a morte, que nesse estudo foi considerado o evento de interesse. A aplicação e os resultados encontrados foram satisfatórios, pois através da análise da significância das covariáveis e da análise dos resíduos, foi encontrado o melhor modelo. Utilizou-se o pacote *survival* no *software* R para ajuste do modelo.

Palavras-chave: Analise de sobrevivência. Modelo semi-paramétrico. Modelo de Cox.

ABSTRACT

One of the most applied models in survival analysis is the Cox regression model. Its techniques can be used in medicine, engineering, insurance analysis and even to analyze the risk of a customer becoming a defaulter, that is, this Model is different from the others because of its wide application in several areas and because it is a semi-parametric model, which allows it to make use of covariates associated with the individuals present in the study, such as, for example, age, sex, preexisting diseases , Income, educational level, place of residence, among others, so that it can model the effect of these covariates on the survival time of the individual, which is the main objective of the survival analysis. The objective of this monograph is to make an introduction about the semi-parametric Cox regression model, including the methods of adaptation and adjustment to survival data, providing a brief study on the survival analysis. The theory studied is illustrated with an application of the model to a set of data from patients with advanced lung cancer, where the time to death is analyzed, which in this study is the event of interest. The survival package was used in software R for model adjustment.

Keywords: Survival analysis. Semi-parametric model. Cox model.

LISTAS DE ILUSTRAÇÕES

LISTAS DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise de sobrevivência está entre os ramos da Estatística que mais cresceram nos últimos anos. Esse crescimento deve-se ao aprimoramento e desenvolvimento de técnicas, que associadas ao desenvolvimento computacional permitem que ela seja aplicada em diversas áreas. Suas técnicas podem ser utilizadas na medicina, para prever o tempo de sobrevida de pacientes (SCHNEIDER; D'ORSI, 2009), na engenharia podem auxiliar no estudo do tempo de vida útil de peças industrial (DANTAS, 2008), até mesmo na área de seguros as técnicas de análise de sobrevivência podem ser utilizadas para modelar o risco de sinistros entre os segurados (PORTILHO, 2013), dentre outras diversas áreas de conhecimento que permitem sua utilização.

Os modelos com dados em sobrevivência são baseados nos artigos seminais de Kaplan e Meier (1948) e de Cox (1972), estes autores contribuíram para esse reconhecimento e crescimento da utilização da análise de sobrevivência. Duas características importantes da análise de sobrevivência são: o tempo de falha e a censura. O tempo de falha é na análise de sobrevivência a variável resposta, que consiste no tempo até a ocorrência do evento de interesse. Já as censuras são observações parciais ou incompletas da variável resposta, que ocorrem quando por algum motivo o acompanhamento do indivíduo no experimento foi interrompido.

Com a análise de sobrevivência é possível observar ou explicar como fatores relacionados ao indivíduo presente no estudo, como por exemplo, idade, sexo, doenças pré-existentes, renda, grau de escolaridade, local de residência, entre outros diversos fatores, que são chamados de covariáveis, podem estar relacionados com a variável resposta. Para que seja possível realizar estudos sobre esta relação, a análise de sobrevivência faz uso de modelos de regressão.

Existem diversos modelos de regressão para verificar se as covariáveis influenciam no tempo de sobrevivência ou de censura, por exemplo, o modelo de regressão paramétrico Exponencial e o modelo semi-paramétrico de Cox. Nesse contexto, o objetivo desse estudo consiste em apresentar o modelo de regressão de Cox, no qual permite modelar o efeito das covariáveis sobre o tempo de sobrevivência.

Essa monografia tem como finalidade introduzir o modelo de Cox, onde inicialmente pretende-se constituir sua fundamentação teórica e a descrição breve da Análise de Sobrevivência, através das técnicas paramétricas, com os conceitos do estimador de

Kaplan-Meier e a utilização de técnicas semi-paramétricas, com ênfase no modelo de regressão de Cox, apresentando sua origem, fundamentos e técnicas de ajuste de precisão. A teoria estudada é ilustrada com uma aplicação do modelo de Cox a um conjunto de dados reais que estuda o tempo até a morte de pacientes com câncer avançado de pulmão, disponível no pacote survival do software R (R CORE TEAM, 2015).

2 CONCEITOS BÁSICOS

A análise de sobrevivência consiste em um conjunto de modelos e métodos utilizados para inferir estatisticamente sobre dados de sobrevivência. Por natureza, sua resposta é longitudinal, ou seja, podem existir ao longo do tempo em apenas uma unidade amostral muitas observações. Em estudos de análise de sobrevivência a variável resposta é o tempo de falha, que pode ser o tempo em que o indivíduo é acompanhado durante um estudo, até que ocorra o evento de interesse, que geralmente é um evento indesejável, como por exemplo, o tempo até a morte de um determinado paciente, a recidiva de uma doença ou até mesmo o tempo de vida útil de uma peça industrial, como demonstrado por Bustamante-Teixeira, Faerstein e Latorre (2002).

A principal característica de dados de sobrevivência é a presença de censura (COLOSIMO; GIOLO, 2010), que é uma observação parcial em relação ao paciente analisado, no caso de uma aplicação em medicina. Essa censura é utilizada em casos onde o acompanhamento é interrompido por algum motivo, antes do final do estudo. Isto quer dizer que o tempo de observação é sempre inferior ao tempo de falha. O fato de existir a censura é que torna essencial a utilização da análise de sobrevivência, pois em estudos onde não há censura, as técnicas de estatísticas clássicas como a análise de regressão, poderiam ser perfeitamente utilizadas.

2.1 Tempo de Falha

O tempo de falha é definido pelo tempo inicial do período de estudo, a escala de medida e o evento de interesse. Os três elementos devem ser perfeitamente definidos para que se obtenha sucesso no estudo, como já demonstrado por Dantas *et al.* (2010) e Botelho, Silva e Cruz (2009)

De acordo com Colosimo e Giolo (2010), para definir o tempo de falha, é preciso ajustar que o tempo de início deve ser aplicado de forma que os indivíduos sejam comparáveis no início do estudo, levando em consideração as exceções recorrentes das variações provenientes das covariâncias aplicada; definir a escala de medida do tempo e qual será o evento de interesse (falha) comum em estudos de analise sobrevivência, como por exemplo a morte ou a recidiva de uma doença.

2.2 Censura

Em estudos clínicos, é comum que nem todos os pacientes cheguem ao momento de falha, o que torna as informações incompletas (censuras). Essas censuras podem ocorrer por diversos motivos, como a não ocorrência da falha no paciente antes do final do estudo, ou até mesmo a perda do acompanhamento do paciente, entre outros motivos, como já demonstrado por Lee (1992).

Porém, mesmo com dados censurados contendo informações incompletas ou parciais sobre um paciente, todos os dados registrados devem ser considerados, visto que o tempo até a ocorrência do evento (falha), para todos os pacientes, é superior ao tempo registrado até o ultimo acompanhamento. Ressaltando que os dados, mesmo que censurados, fornecem informações importantes sobre o tempo de vida do paciente e a não utilização de tais dados podem fornecer resultados viciados.

Como já descrito por Bastos *et al.* (2006), é possível dizer que uma variável aleatória é censurada quando não há meios de observar o seu valor exato, mas é possível obter um limite inferior para este valor (censura à direita), ou limite superior (censura à esquerda), ou até mesmo ambos (censura intervalar). De acordo com Colosimo e Giolo (2010) as censuras são descritas conforme as informações apresentadas a seguir:

Dentre os 3 tipos de censuras citadas anteriormente, a censura à direita é a mais comum, pois ocorre quando o tempo até a ocorrência da falha está à direita do tempo de registro, isto é, ela acontece quando o evento de interesse não ocorre durante o período de estudo. Por exemplo, o indivíduo não morrer de câncer até o final do período do estudo.

A censura à direita ainda pode ser classificada em 3 tipos: A censura à direita de tipo 1 geralmente ocorre em casos onde os estudos possuem um tempo de encerramento pré-estabelecido, o que pode acarretar que a falha ou tempo até a ocorrência do evento de interesse, não ocorra durante o período do estudo; já a censura à direita de tipo 2 ocorre quando é pré-estabelecido o encerramento dos estudos, após a ocorrência de uma certa quantidade de eventos e por fim, a censura à direita de tipo 3 ocorre quando o indivíduo interrompe o experimento, por algum motivo antes do encerramento do estudo e antes da ocorrência da falha. A censura à esquerda acontece quando o tempo registrado é maior que o tempo de falha. Esse tipo de censura geralmente ocorre quando o evento já ocorreu no momento em que o indivíduo foi observado (COLOSIMO; GIOLO, 2010).

Quando os indivíduos são acompanhados periodicamente, o evento só pode ser confirmado no momento da consulta, portanto caso ocorra o evento, a informação que deve

ser considerada é que o evento ocorreu entre uma consulta e outra, não podendo ser especificado o momento exato da falha. Este tipo de censura é determinada como intervalar, pois o tempo de falha T pertence a um intervalo $(L,U]$. Lindsey *et al.*(1998) dizem que tempos exatos de falhas são caso especiais de dados de sobrevivência intervalar com $L = U$, para censuras à direita $U = \infty$ e para censuras à esquerda $L = 0$.

2.3 Funções básicas do tempo de sobrevivência

Nesta seção definem-se as funções básicas da análise de sobrevivência. De acordo Colosimo e Giolo (2010), a variável aleatória não-negativa *T*, geralmente contínua, que representa o tempo de falha, é referenciada na análise de sobrevivência por sua função de sobrevivência ou pela função de taxa de falha. Essas funções que são de extrema importância para análise de dados de sobrevivência, assim como a função de risco acumulada, são apresentadas a seguir.

2.3.1 Função de sobrevivência

Conforme definida por Carvalho *et al.* (2011), a probabilidade de um evento não ocorrer até um certo tempo t, ou seja, a função de sobrevivência é a probabilidade de sobreviver ao tempo t . Essa função é definida por:

$$S(t) = P(T > t), t \geq 0 \quad (1)$$

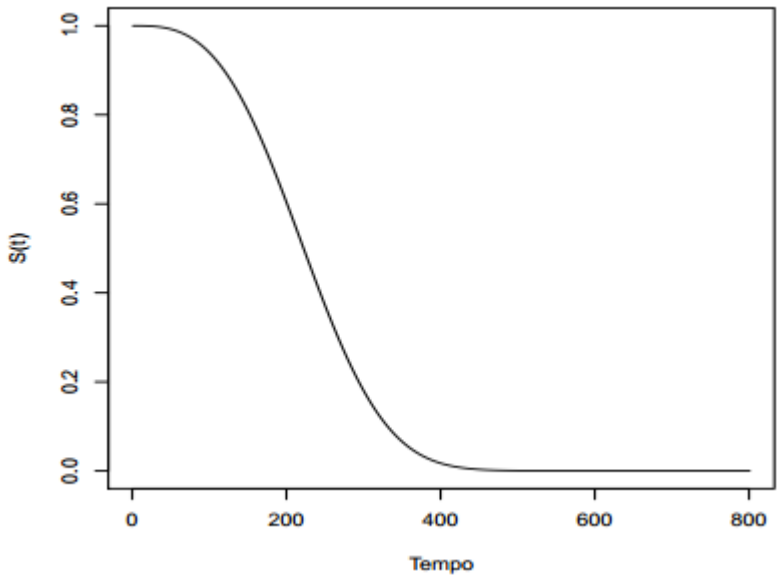
Por consequência, a função distribuição acumulada é definida pela probabilidade de não sobrevivência ao tempo t, que pode ser descrita pela seguinte função:

$$F(t) = 1 - S(t). \quad (2)$$

Entre as definições da função de sobrevivência, é evidenciado que é uma função não crescente no tempo, que a probabilidade de sobrevivência no tempo infinito é zero e que a probabilidade de sobrevivência de pelo menos superar o tempo zero é 1.

Conforme descrito por Carvalho *et al.* (2011), a função de sobrevivência pode ser descrita por um gráfico de $S(t)$ *versus* (t) que é conhecido como curva de sobrevivência. A Figura 1 ilustra a forma típica da função de sobrevivência. O gráfico pode apresentar uma curva íngreme, que representa um tempo de sobrevivência curto ou uma baixa razão de sobrevivência, como também o gráfico pode apresentar uma curva gradual ou plana, que significa que existe uma longa sobrevivência, o que evidencia a alta taxa de sobrevivência.

Figura 1 – Forma típica da função de sobrevivência



2.3.2 Função da Taxa de Risco ou Falha

Em termos da função de sobrevivência e de acordo com Alves *et al*. (2010) pode-se expressar a probabilidade da falha ocorrer em um determinado intervalo de tempo $[t_1, t_2)$, como:

$$S(t_1) - S(t_2). \tag{3}$$

Dado que a falha não ocorreu antes do intervalo t_1, definimos a taxa de falha durante o período $[t_1, t_2)$ como a probabilidade de ocorrer a falha durante esse intervalo, dividida pelo comprimento do intervalo, Assim, a função que expressa a taxa de falha no intervalo $[t_1, t_2)$ é definida por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{4}$$

Redefinindo alguns termos das funções apresentadas interiormente, com o intuito de descrever a forma que a taxa instantânea de falha muda com o tempo, temos a seguinte função:

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t\, S(t)}. \tag{5}$$

Em que o intervalo $[t_1, t_2)$, foi redefinido como $[t, t + \Delta t)$.

Assumindo Δt bem pequena, $\lambda(t)$ representa a taxa de falha instantânea no tempo t, condicional à sobrevivência até o tempo t. Essa função é muito utilizada para descrever a distribuição de sobrevivência de um paciente. Sendo assim, a taxa de Falha de T é definida por:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \quad (6)$$

2.3.3 Função de risco acumulada

A função de taxa de risco acumulada, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo presente no estudo e pode ser definida por:

$$\Lambda(t) = \int_0^t \lambda(u)du \quad (7)$$

De acordo com Colosimo e Giolo (2010), essa função não possui uma interpretação direta, mas pode ser de grande utilidade na avaliação da função de taxa de falha $\lambda(t)$. Isso ocorre essencialmente na estimação não-paramétrica, onde $\Lambda(t)$ apresenta um estimador com ótimas propriedades, ao contrário de $\lambda(t)$.

2.4 Estimação da função de sobrevivência

A análise de sobrevivência em casos relacionados a medicina, geralmente buscam um comparativo entre tratamentos específicos, identificação de fatores de prognóstico para uma determinada doença, dentre outros vários exemplos e a resposta para todas essas analises são feitas a partir de um conjunto de dados de sobrevivência. Portanto, sendo a resposta encontrada a partir de um conjunto de dados de sobrevivência, conforme sugerido por Colosimo e Giolo (2010), é essencial para o início de qualquer análise estatística a descrição dos dados.

Contudo, os dados censurados podem-se tornar um grande problema quando utilizadas técnicas convencionas de analise descritiva, que utilizam média e desvio padrão, além de técnicas gráficas como histograma, *Box-plot*, entre outros. De acordo com Carvalho *et al.* (2011), nos casos de análise descritiva envolvendo dados de tempo de vida, o principal componente é a função de sobrevivência. Portanto, como passo inicial deve-se encontrar um estimador para a função de sobrevivência, em seguida estimar as estatísticas desejadas, que geralmente são o tempo médio, percentis de falha em tempos fixos de acompanhamento.

2.4.1. Tabela de vida

De acordo com Bowers *et al.* (1997) a tabela de vida é uma forma de resumir a vida de indivíduos de uma determinada população estudada, onde a probabilidade de sobrevivência de cada um dependerá das covariáveis associadas a cada indivíduo, como: sexo,

idade, renda, entre outros. Como também já demonstrado por César (2005), sua construção consiste em dividir o eixo de tempo em um determinado número de intervalos. Supondo que a divisão do eixo do tempo seja realizada em s intervalos, dados pelos pontos de corte, $t_1, t_2, \dots, t_s$.

Isto é $I_j = [t_{j-1}, t_j)$, para $j = 1, \dots, s$, em que $t_0 = 0$ *e* $t_s = +\infty$.

O estimador da tabela de vida pode ser expresso da seguinte forma:

$$\hat{S}(t) = \prod_{i=1}^{j} (1 - \widehat{q_{i-1}}), \qquad t \in I_j, \tag{8}$$

Onde a estimativa para q_j na tabela de vida é dada por:

$$\widehat{q_j} = \frac{nº\ de\ falhas\ no\ intervalo\ [t_{j-1}, t_j)}{\left[nº\ sob\ risco\ em\ t_{j-1}\right] - \left[\frac{1}{2}\ x\ o\ nº\ de\ censuras\ em\ [t_{j-1}, t_j)\right]} \tag{9}$$

2.4.2 Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier, também chamado de estimador limite-produto, foi proposto inicialmente por Böhmer (1912) e após alguns anos teve suas propriedades estudadas por Kaplan e Meier (1958), para estimar a função de sobrevivência. O estimado de Kaplan-Meier pode ser definido pela função:

$$\hat{S}(t) = \frac{nº\ de\ observações\ que\ não\ falharam\ até\ o\ tempo\ t}{nº\ total\ de\ observações\ no\ estudo} \tag{10}$$

Em que, $S(t)$ é uma função escada, com degraus nos tempos de falha de tamanho $1/n$, em que n é o tamanho da amostra analisada. Em caso de empates em um determinado tempo t, o tamanho do degrau fica multiplicado pela quantidade de empates. Em sua construção, o estimador de Kaplan-Meier, leva em consideração o número de falhas para determinar os intervalos de tempo, ou seja, um intervalo de tempo para cada falha. De acordo com Colosimo e Giolo (2010), ao considerar $S(t)$ como uma função discreta com saltos, com probabilidade maior que zero somente nos tempos de falha $t_j, j = 1, \dots, k$, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2) \dots (1 - q_j), \tag{11}$$

Em que q_j é a probabilidade de morte de um indivíduo no intervalo $[t_{j-1}, t_j)$, sabendo que ele sobreviveu até t_{j-1} e considerando $t_0 = 0$. Dessa forma pode-se descrever q_j como:

$$q_j = P\left(T \in \left[t_{j-1}, t_j\right) \middle| T \geq t_{j-1}\right) \tag{12}$$

A partir das informações apresentadas anteriormente, deve-se definir a expressão geral de $S(t)$ em termos de probabilidades condicionais, de forma que o estimador de Kaplan-Meier se reduza a estimar q_j, dado por:

$$\hat{q}_j = \frac{nº\ de\ falhas\ em\ t_{j-1}}{nº\ de\ observações\ sob\ risco\ em\ t_{j-1}} \tag{13}$$

Para $j = 1, \ldots, k+1$, em que $t_{k+1} = \infty$.

Levando em consideração as seguintes definições:

- $t_1 < t_2 \ldots < t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em $t_j, j = 1, \ldots, k$, e
- n_j o número de indivíduos em risco no tempo t_j, ou seja , os indivíduos não censurados ou que não falharam até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right) \tag{14}$$

De acordo com Colosimo e Giolo (2010), são definidas como uma das principais propriedades do Estimador Kaplan-Meyer:

- Quando utilizados em amostras grandes é não viciado.
- É fracamente consistente.
- Converge assintoticamente para um processo gaussiano
- É estimador de máxima verossimilhança de $S(t)$.

> A consistência e normalidade assintótica de $\hat{S}(t)$ foram provadas, sob certas condições de regularidade, por Breslow e Crowley (1974) e Meier (1975), e no artigo original, Kaplan e Meier (1958) mostram que $\hat{S}(t)$ é o estimador de máxima verossimilhança de $S(t)$ (COLOSIMO; GIOLO, 2010, P.40).

Como já descrito por Aranha (2009), é necessário avaliar a precisão do estimador de Kaplan-Meyer, assim como de outros estimadores não citados neste trabalho, para que possamos construir intervalos de confiança e realizar testes de hipóteses para $S(t)$, pois os estimadores estão sujeitos a variações que devem ser descritas em termos de estimações

intervalares. A estimação da variância assintótica do estimador de Kaplan-Meier pode é estimada pela formula de *Green-Wood*, expressa da seguinte forma:

$$\widehat{Var}\left(\hat{S}(t)\right) = [\hat{S}(t)]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}. \quad (15)$$

Como S(t), para *t* fixo, possui uma distribuição assintótica Normal, segue que um intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é definido por:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))}, \quad (16)$$

Em que, $\alpha/2$ denota o $\alpha/2$-percentil da distribuição Normal padrão.

No próximo capítulo descreve-se o modelo no qual é possível estimar os efeitos das covariáveis sem a necessidade de supor distribuição de probabilidade do tempo de sobrevivência.

3 MODELO DE REGRESSÃO DE COX

Conforme Colosimo e Giolo (2010), o modelo de Regressão de Cox (1972) permite a análise de dados provenientes de tempo de vida, em que a variável resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis. E ainda, a suposição básica para o uso do modelo de regressão de Cox, que é: as taxas de falha devem ser proporcionais ou equivalentes, assim como também as taxas de falha acumuladas devem ser proporcionais.

Suponha que ao compararmos pacientes em um novo experimento, com o intuito de analisar o tempo de falha após receber um novo tratamento, sejam selecionados aleatoriamente dois grupos, um que receberá o tratamento padrão (grupo 0) e o outro grupo, novo tratamento (grupo 1). Ao representarmos a taxa de falha do grupo 0, por $\lambda_0(t)$ e a do grupo 1, por $\lambda_1(t)$ pode-se definir as proporcionalidades entre as funções como:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = K, \tag{17}$$

Em que K é definida como a razão das taxas de falhas dos grupos 0 e 1, constante para todo o período do estudo (tempo t). Sendo x a variável indicadora de grupo, em que:

$$x = \begin{cases} 0 & se\ grupo\ 0, \\ 1 & se\ grupo\ 1, \end{cases} \tag{18}$$

e $K = \exp\{\beta x\}$, então:

$$\lambda(t|x) = \lambda_0(t)\exp\{\beta x\}, \tag{19}$$

ou seja,

$$\lambda(t|x) = \begin{cases} \lambda_1(t) = \lambda_0(t)\exp\{\beta\}, & se\ x = 1 \\ \lambda_0(t), & se\ x = 0 \end{cases} \tag{20}$$

Genericamente considerando p covariáveis, de modo que x seja um vetor com os componentes $x = (x_1, \dots, x_p)'$, A expressão geral do modelo de regressão de Cox considera:

$$\lambda(t|x) = \lambda_0(t)g(x'\beta), \tag{21}$$

Na qual $g(x'\beta)$ é uma função não-negativa que deve ser especificada, tal que $g(0) = 1$. Este modelo é definido pelo o produto de um componente paramétrico e outro não-paramétrico($\lambda_0(t)$). Usualmente o modelo não-paramétrico é denominado como função da taxa de falha base, pois $\lambda(t|x) = \lambda_0(t)$, quando $x = 0$; além de não ser especificado e ser uma função não-negativa do tempo. O componente paramétrico é utilizado com frequência na seguinte forma multiplicativa:

$$g(x'\beta) = \exp\{x'\beta\} = \exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}, \quad (22)$$

Sendo β é o vetor de parâmetros associados às covariáveis, garantindo que $\lambda(t|x)$ seja sempre não-negativa.

Contudo, de acordo com Colosimo e Giolo (2010) é necessário realizar a estimação dos efeitos das covariáveis, que no modelo de regressão de Cox é realizado através do método de máxima verossimilhança parcial, que é descrito na próxima seção.

3.1 Método de Máxima Verossimilhança Parcial

Os efeitos das covariáveis sobre a função de taxa de falha são medidos pelos coeficientes $\beta's$, que são características do modelo de regressão de Cox. Essas quantidades devem ser estimadas por meio de observações amostrais, de forma que o modelo fique determinado, o que faz necessário um método de estimação para que possam ser realizadas inferências acerca dos parâmetros do modelo.

Cox propôs em seu artigo original e formalizou em um artigo subsequente (COX, 1975) o método necessário para realização da estimação, denominando de método de máxima verossimilhança parcial, visto que não seria apropriado utilizar o usual método de máxima verossimilhança (COX; HINKLEY, 1974), devido a presença do componente não-paramétrico $\lambda_0(t)$.

Conforme demonstrado por Gouvêa (2009), a máxima verossimilhança parcial considera o seguinte argumento condicional: a probabilidade condicional da i-esima observação vir a falhar no tempo t_i, levando em consideração que as observações sob risco em t_i são conhecidas, é:

$$\begin{aligned} &P(indivíduo\ falhar\ em\ t_i | uma\ falha\ em\ t_i\ e\ história\ até\ t_i) = \\ &= \frac{P(indivíduo\ falhar\ em\ t_i | sobreviveu\ a\ t_i\ e\ história\ até\ t_i)}{P(uma\ falha\ em\ t_i |\ história\ até\ t_i)} = \\ &= \frac{\lambda_i(t|x_i)}{\sum_{j\in R(t_i)} \lambda_j(t|x_j)} = \frac{\lambda_0(t)\exp\{x'_i\beta\}}{\sum_{j\in R(t_i)} \lambda_0(t)\{x'_j\beta\}} = \frac{\exp\{x'_i\beta\}}{\sum_{j\in R(t_i)} exp\{x'_j\beta\}}, \end{aligned} \quad (23)$$

Em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_j. Sendo assim, pode-se observar que condicional ao histórico de censuras e falhas até o tempo t_i, componente não-parametrico $\lambda_0(t)$ não aparece na expressão citada acima. Assim a forma utilizada da função de verossimilhança que nos permite inferir sobre os parâmetros do modelo, é formada pelo produto dos termos apresentados na expressão acima, associados aos tempos distintos de falha, ou seja,

$$L(\beta) = \prod_{i=1}^{k} \frac{exp\{x'_j\beta\}}{\sum_{j\in R(t_i)} exp\{x'_j\beta\}} = \prod_{i=1}^{n} \left(\frac{exp\{x'_j\beta\}}{\sum_{j\in R(t_i)} exp\{x'_j\beta\}} \right)^{\delta_i}, \quad (24)$$

Em que δ_i é o indicador de falha. β os valores que maximizam a função de verossimilhança parcial, $L(\beta)$, são obtidos resolvendo-se o sistema de equações definido por $U(\beta) = 0$, onde $U(\beta)$ é o vetor escore de derivadas de primeira ordem da função $l(\beta) = \log\big(L(\beta)\big)$, ou seja:

$$U(\beta) = \sum_{i=1}^{n} \delta_1 \left[x_i \frac{\sum_{j\in R(t_i)} x_j exp\{x'_j\hat{\beta}\}}{\sum_{j\in R(t_i)} exp\{x'_j\beta\}} \right] = 0 \quad (25)$$

Dentre suas definições, a função de verossimilhança parcial assume que os tempos de sobrevivência são contínuos, e por consequência não pressupõe a possibilidade de empates nos valores observados, o que na prática podem ocorrer nos tempos de falha ou censura, devido à escala de medida. "Quando ocorrem empates entre falhas e censuras, usa-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha" (COLOSIMO; GIOLO, 2010, P.161)

As observações empatadas devem ser incorporadas, quando presentes, tendo que ser utilizada a função de verossimilhança parcial modificada para realização da incorporação das observações. A aproximação para a equação citada acima, proposta por Breslow (1972) e Peto (1972) considera s_i o vetor formado pelo somatório das correspondentes covariáveis p para indivíduos que falham no mesmo tempo $t_i (i = 1, \dots, k)$ e d_j a quantidade de falhas ocorridas no mesmo tempo. A aproximação se dá pela seguinte função:

$$L(\beta) = \prod_{i=1}^{k} \frac{exp\{s'_j\beta\}}{\left[\sum_{j\in R(t_i)} exp\{x'_j\beta\}\right]} \quad (26)$$

De acordo com Colosimo e Giolo (2010), existem diversos outros modelos de aproximação para a função de verossimilhança, além do modelo abordado acima proposto por Breslow (1972) e Peto (1972), como os propostos por Efron (1977), Farewell e Prentice (1980).

As propriedades assintóticas dos estimadores de máxima verossimilhança parcial são necessárias para a realização da construção dos intervalos de confiança, assim como para realizar os testes de hipóteses sobre os coeficientes do modelo, sendo possível utilizar as estatísticas de Wald, da Razão de Verossimilhança e Escore, para se fazer inferências sobre os parâmetros do modelo de Cox utilizando-se a função de verossimilhança parcial.

3.2 Adequação do Modelo de Cox

Conforme informado anteriormente, devido à presença de componentes não-paramétricos, o modelo de regressão de Cox é bastante flexível, mas ainda sim ele não se ajusta a qualquer tipo de situação clínica, sendo necessária a utilização de técnicas que permitam avaliar sua adequação.

Existem diversos métodos que permitem avaliar a adequação do modelo, que em geral baseiam-se em avaliar por meio dos resíduos, a distribuição dos erros. Conforme afirmado por Klein e Moeschberger (2003), "Estas técnicas devem ser utilizadas como um meio de rejeitar modelos claramente inapropriados e não para 'provar' que um particular modelo paramétrico está correto" (COLOSIMO; GIOLO, 2010, P. 123).

Está afirmação é importante, pois em diversos tipos de aplicações, mais de um modelo pode fornecer ajustes razoáveis, assim como estimativas similares das quantidades de interesse, isso válido para modelos paramétricos e não-paramétricos. A seguir serão apresentados alguns métodos utilizados para avaliar a qualidade geral do ajuste do modelo, a proporcionalidade dos riscos, entre outros.

3.2.1 Avaliação da qualidade geral do modelo de ajuste

Os resíduos de Cox-Snell (1968), são utilizados no modelo de Cox, com o mesmo propósito que são utilizados em modelos paramétricos, avaliar a qualidade geral do ajuste, como já demonstrado por Santos e Nakano (2015). No modelo de Cox, os resíduos de Cox-Snell são definidos como:

$$\hat{e}_i = \hat{\Lambda}_0(t_i) \exp\left\{\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\right\}, \qquad i = 1, \dots, n, \tag{27}$$

Sendo $\hat{\Lambda}_0(t_i)$, estimada através de uma função escada com saltos nos tempos distintos de falha, proposta por Breslow (1972), onde d_j, é o número de falhas em t_j:

$$\hat{\Lambda}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\epsilon R_j} exp\{x'_l\hat{\beta}\}}, \tag{28}$$

Dessa forma, aplicando os resíduos de Cox-Snell, caso o modelo estiver ajustado, os $\hat{e}_i's$ devem ser considerados como uma amostra censurada de uma distribuição exponencial padrão, sendo o gráfico entre $\hat{e}_i$ versus $\hat{\Delta}(t_i)$, como exemplo, deve ser aproximado a uma reta. Porém, caso o gráfico demonstre o contrário, ou seja, de forma não linear, isso demonstra que

o modelo não está corretamente ajustado. Porém essa análise gráfica não permite identificar qual o problema está ocorrendo.

Assim, de acordo com Colosimo e Giolo (2010), não são recomendados gráficos que envolvam resíduos, para que possa ser realizada uma avaliação sobre a suposição das taxas de falhas proporcionais. Contundo, existem técnicas gráficas, que permitem avaliar a suposição de taxas de falhas proporcionais no modelo de Cox:

- Método gráfico descrito.

Esta técnica gráfica consiste em separar os dados em *m* extratos diferentes, levando em consideração alguma covariável, como por exemplo, o sexo. Em seguida deve ser realizada a estimação de $\hat{\Delta}_{0_j}(t)$, utilizando a expressão citada acima.Caso as curvas dos logaritmos t versus $\hat{\Delta}_{0_j}(t)$, apresentarem diferenças aproximadamente constantes no tempo, significa que a suposição é válida, Brito e Neto (2008). Caso as curvas não sejam paralelas, significam desvios da suposição de taxa de falha proporcional.O ideal é que seja construído um gráfico para cada covariável, para que possa ser indicado com facilidade qual a covariável que apresenta a violação da suposição.

- Método com coeficiente dependente do tempo.

Esse método consiste em analisar os resíduos de Schoenfeld (1982). Para defini-los no modelo de Cox, deve-se considerar que se o *i*-ésimo individuo com vetor de covariáveis $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ é observado falhar, deve-se considerar para este individuo um vetor de resíduos de Schoenfeld $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \dots, r_{ip})$, onde em cada componente r_{iq}, para $q = 1, \dots, p$, é definido a partir de:

$$r_{iq} = x_{iq} - \frac{\sum_{j \epsilon R(t_i)} x_{jq} exp\{x'_j \hat{\beta}\}}{\sum_{j \epsilon R(t_i)} exp\{x'_j \hat{\beta}\}}. \quad (29)$$

Os resíduos não são definidos pelas censuras e sim para cada falha. Ao considerarmos que para cada uma das *p* covariáveis, existe para o individuo *i*, um resíduo de Schoenfeld correspondente e sabendo que os resíduos são definidos em cada falha, pode-se construir um através do conjunto de resíduos de Schoenfeld uma matriz com *d* linhas e *p* colunas, sendo *d* o número de falhas.

Cada linha da matriz corresponde a um tempo de falha e cada coluna uma das *p* covariáveis utilizadas no modelo. Condicional a uma falha no conjunto de risco $R(t_i)$, o valor

esperado da covariável para esta falha é obtido por $\frac{\sum_{j\epsilon R(t_i)} x_{jq} exp\{x'_j\hat{\beta}\}}{\sum_{j\epsilon R(t_i)} exp\{x'_j\hat{\beta}\}}$, dessa forma a interpretação de r_{iq} como um resíduo é correta.

De acordo com Colosimo e Giolo (2010) existe uma forma padronizada dos resíduos de Schoenfeld (*scaled Schoenfeld residual*) que permite que a estrutura de correlação dos resíduos seja considerada. Essa forma padronizada é definida como:

$$S_i^* = \left[I(\widehat{\beta})\right]^{-1} \times r_i, \tag{30}$$

Na qual $I(\widehat{\beta})$ representa a matriz de informação observada (THERNEAU; GRAMBSCH, 2000). De acordo com Colosimo e Giolo (2010), esse modelo padronizado é baseado em um resultado importante apresentado por Grambsch e Therneau (1994) que considera o modelo expresso pela seguinte função:

$$\lambda(t|X) = \lambda_0(t) exp\{x'\beta(t)\}, \tag{31}$$

Ressalta-se que há a restrição de que $\beta(t) = \beta$, como uma forma alternativa de representar o modelo de Cox, onde essa restrição implica diretamente na proporcionalidade das taxas de falha. "Quando $\beta(t)$ não for constante, o impacto de uma ou mais covariáveis na taxa de falha pode variar como o tempo" (COLOSIMO; GIOLO, 2010, P.168). Assim, caso o gráfico de $\beta_q(t)$ versus *t* seja uma linha horizontal, deve-se supor que a as taxas de falhas proporcionais é válida.

Contudo, as técnicas gráficas apresentadas nos apresentam conclusões subjetivas, pois os resultados dependem exclusivamente da interpretação de quem os analisa, o que torna a utilização de medidas estatísticas e testas de hipóteses, métodos de grande valia para a avaliação do modelo de ajuste.

Dentre os diversos testes, são citados em Colosimo e Giolo (2010), como o coeficiente de correlação de Pearson (*p)* e testes de hipótese geral de proporcionalidade.

4 APLICAÇÃO

Neste capítulo utilizou-se o modelo de Cox para avaliar fatores associados a qualidade de vida de pacientes com câncer avançado de pulmão, e o método de Kaplan-Meier para estimar a função de sobrevivência. Nestes pacientes, a qualidade de vida é afetada por vários fatores, por exemplo, o nível da gravidade da doença. Conforme Franceschini *et al.* (2008), a qualidade de vida inicial é uma medida importante para avaliação do prognóstico e da sobrevida dos pacientes.

Os pacientes foram submetidos a um sistema de pontuação de avaliação da qualidade de vida dos pacientes conhecido por *Medical OutComes Study-36-item Short-Form Health Survey* (SF-36). Segundo Francheschini (2008), trata-se de um questionário genérico, multidimensional, composto de oito subitens, que estão associados à capacidade funcional dos pacientes, as limitações quanto ao tipo e a qualidade de trabalho e das atividades da vida diária, a presença de dor, o estado geral de saúde do paciente (como se sente), itens que consideram o nível de energia e de fadiga avaliando a vitalidade dos mesmos, aspectos sociais e psicológicos. Estes tipos de avaliações têm se mostrado atualmente com frequência como importantes fatores prognósticos para a sobrevivência em portadores de câncer de pulmão (LOPRINZINI et al, 1994).

Nesta aplicação considera-se um banco de dados que estuda o tempo até a morte de pacientes com câncer avançado de pulmão, disponível no pacote survival do software R (R CORE TEAM, 2015).

4.1 Descrição dos dados

A base de dados é composta de 228 pacientes, nos quais 72,37% foram ao óbito (falha) e 27,63% sobreviveram até o final do estudo (censura).

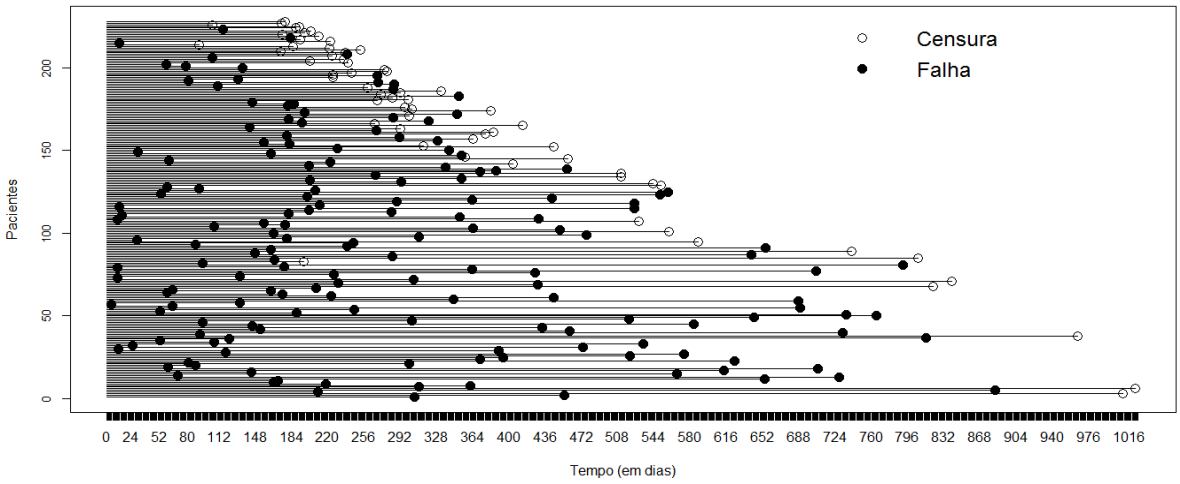
A Tabela 1 mostra um resumo das principais medidas descritivas para as trajetórias de tempo dos indivíduos envolvidos no estudo.

Tabela 1 – Medidas descritivas para os tempos de observação dos pacientes com câncer

Medidas descritivas	Tempos (em dias)		
	Geral	Censurados	Falhas
Mínimo	5,0	92,0	5,0
1º Quartil	166,8	221,5	135,0
Mediana	255,5	284,0	226,0
Média	305,3	363,5	283,0
3º Quartil	396,5	428,5	387,0
Máximo	1022,0	1022,0	883,0

A Figura 2 mostra a representação gráfica dos tempos de observação individuais dos pacientes com câncer, incluindo a informação sobre a ocorrência de censura e falha.

Figura 2 – Ilustração dos tempos de falha e censura dos pacientes com câncer



Nesta aplicação, define-se a variável aleatória não-negativa T como o tempo (em dias) até o óbito do paciente com câncer de pulmão. As variáveis disponíveis na base de dados são descritas no Quadro 1 a seguir.

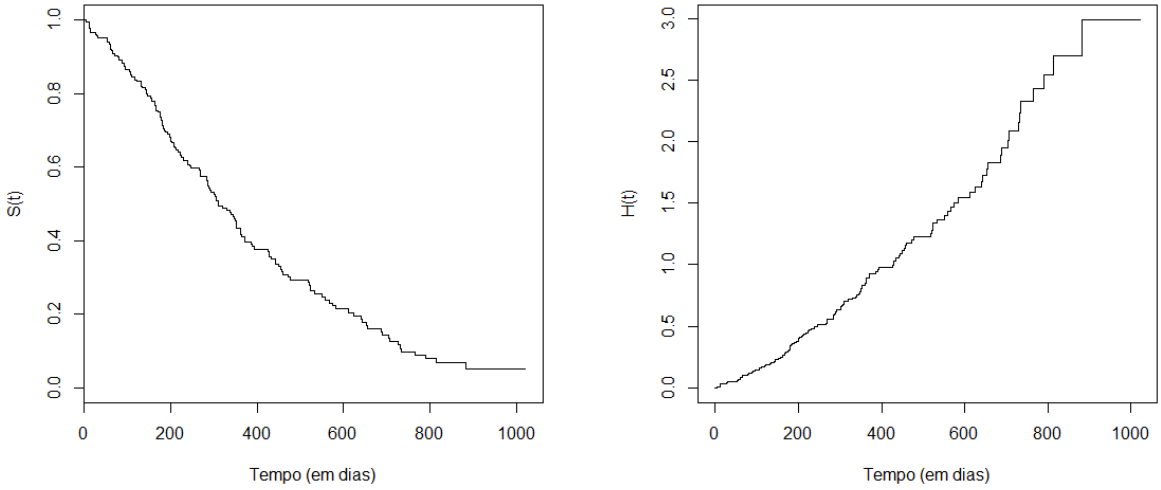
Quadro 1 – Descrição das variáveis utilizadas no estudo

Variáveis	**Descrição**
inst	código da instituição
time	dias até o óbito
status	indica o óbito com câncer de pulmão em estagio grave ou censura
age	idade em anos
sex	Sexo
ph.ecog	pontuação de desempenho que varia entre 0 (boa) e 5(morto) pelo critério do ECOG, uma variável categórica ordinal
ph.karno	pontuação de desempenho que varia entre 0 (mau) e 100(bom) avaliado por médico, pelo critério de Karnofsky, uma variável categórica ordinal
pat.karno	pontuação de desempenho como avaliado por paciente, pelo critério de Karnofsky, uma variável categórica ordinal
meal.cal	calorias consumidas durante as refeições
wt.loss	a perda de peso em seis meses

Inicialmente realizou-se uma análise descritiva para compreender o comportamento dos dados. A Figura 3 mostra a curva estimada de sobrevivência e a função de risco acumulado, estimados para os dados dos pacientes de câncer de pulmão. A função de sobrevivência caracteriza-se por começar com $S(0) = 1$, ou seja, a probabilidade de um

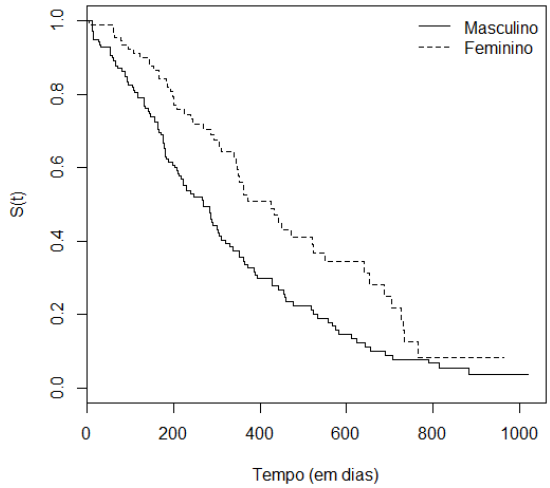
indivíduo sobreviver por mais de 0 dias é 1. Com o aumento do tempo, nesta aplicação, a função $S(t)$ decresce, assim, $t \to \infty, S(\infty) = 0$.

Figura 3 – Gráficos da função de sobrevivência e da função de risco acumulado



O comportamento das funções de sobrevivência estimadas por meio do estimador de Kaplan-Meier para a variável sexo são mostradas na Figura 4. A comparação entre as curvas (Figura 2) sugere que existe diferença entre a sobrevivência de homens e mulheres. Observa-se que a curva dos homens parece estar menor do que para as mulheres durante o período observado, esse padrão indica que uma sobrevivência menor entre os homens.

‘’ Figura 4 – Sobrevivência estimada por Kaplan-Meier por sexo



Segundo Carvalho *et al.* (2011), para comparar as curvas de sobrevivência mais formalmente, deve-se recorrer a testes de hipótese. Neste estudo, utilizou-se o teste de

Mantel-Haenzel, ou log-rank, que compara os valores observados e esperados em cada estrato sob a hipótese de que o risco é mesmo em todos os grupos. Portanto, pela estatística de *log-rank* pode-se concluir pela existência de diferenças significativas entre os sexos (χ^2=10,3, p-valor= 0,00131), confirmando o que foi verificado na Figura 4. A seguir, utiliza-se o modelo de Cox para avaliar os fatores que estão associados a qualidade de vida de pacientes com câncer avançado de pulmão através das covariáveis (Quadro 1).

4.2 Modelo de Cox

Os dados foram ajustados ao modelo de Cox considerando todas as variáveis estudadas. As estatísticas desse modelo inicial estão a seguir:

Tabela 2 – Estimativa do modelo inicial que contém todas as covariáveis em estudo.

Cováriavel	Paramêtro	Estimativas	Exp (β)	z	Valor-p
age	β_1	1.06	1.01	0.92	0.3591
sex	β_2	-5.51	5.76	-2.74	0.0061
ph.ecog	β_3	7.34	2.08	3.29	0.0010
ph.karno	β_4	2.25	1.02	2.00	0.0457
pat.karno	β_5	-1.24	9.88	-1.54	0.1232
wt.loss	β_6	-1.43	9.86	-1.84	0.0652
meal.cal	β_7	3.33	1.00	0.13	0.8979

Observa-se pelas estatísticas na Tabela 2 que este modelo não possui riscos proporcionais, além de confirmar que nem todas as covariáveis são significativas. Com o intuito de chegar ao modelo final, foram retiradas uma variável por vez, considerando como critério para retirada, a que apresentava o menor Valor-p para a suposição de riscos proporcionais. Foi observado que a cada retirada de uma variável as estatísticas das outras foram modificadas.

Após a análise de diversos modelos, formados por várias combinações de variáveis, o modelo que o que melhor representa os dados e que sustenta todas as hipóteses que são necessárias para a validação do modelo, é formado pelas variáveis: *sex, ph.ecog e ph.karno.*

Tabela 3 – Estimativa do modelo final, contendo apenas 3 covariáveis.

Cováriavel	Paramêtro	Estimativas	Exp (β)	z	Valor-p
sex	β_1	-0.56882	0.56619	-3.37	0.00075
ph.ecog	β_2	0.64036	1.89716	3.59	0.00032
ph.karno	β_3	0.01105	1.01112	1.16	0.24629

A partir das estatísticas na Tabela 3, assume-se que o modelo final possui riscos proporcionais e que todas as variáveis são significantes.

4.3 Análise dos Resíduos

A proporcionalidade dos riscos foi avaliada com base na representação gráfica dos resíduos de Schoenfeld (Figura 5). Através destes gráficos é razoável presumir a existência de proporcionalidade das funções de risco, visto que não é encontrada uma tendência marcada os resíduos Schoenfeld em função do tempo. O teste estatístico que reforça esta interpretação, permite testar a existência de correlação entre os resíduos de Schoenfeld em função do tempo.

A Tabela 4 apresenta os valores-p relativos a este teste, confirmando a existência de proporcionalidade dos riscos para todas as covariáveis.

Figura 5 – Resíduos padronizados de Shoenfeld *versus* os tempos para as covariáveis

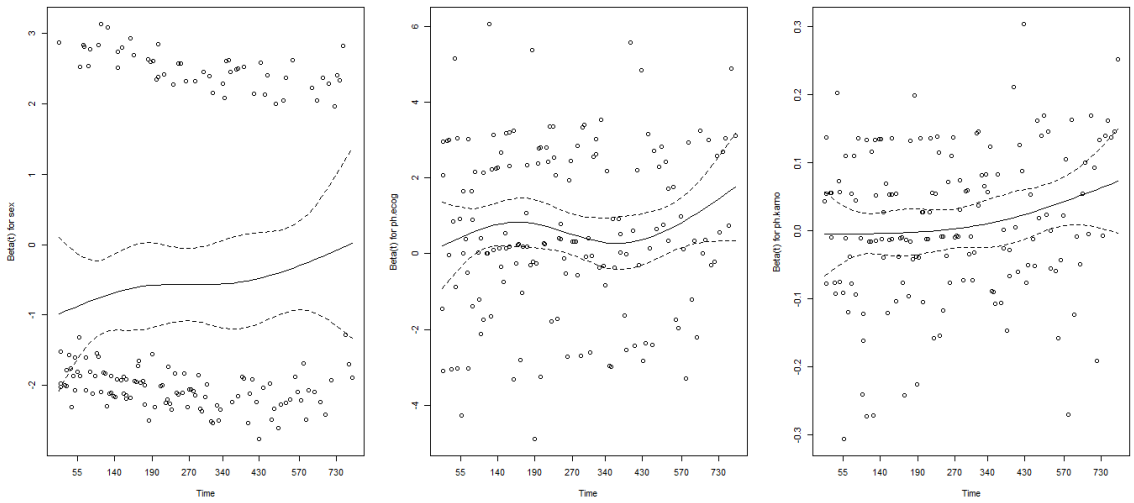


Tabela 4 – Estatísticas a respeito do Risco Proporcional das variáveis selecionadas

Cováriavel	**Paramêtro**	ρ	X^2	**Valor-p**
sex	β_1	0.0934	1.403	0.2363
ph.ecog	β_2	0.0554	0.444	0.5053
ph.karno	β_3	0.1762	4.065	0.0438
Global	-	-	**7.867**	**0.0488**

5 CONSIDERAÇOES FINAIS

Neste trabalho foram apresentados brevemente os principais conceitos sobre analise de sobrevivência, assim como a modelo de regressão de Cox. Este modelo, que é tema central desse trabalho, se destaca de outros modelos regressão pelo fato de permitir a utilização de covariáveis, o que permite que sua ampla utilização em diversas áreas.

Foram também apresentadas técnicas paramétricas como o estimado de Kaplan-Meier, que é de extrema importância para análise de sobrevivência e para a aplicação do modelo de regressão de Cox, pois nos permite visualizar de forma clara o comportamento dos dados, com ênfase na curva estimada de sobrevivência e na função de risco acumulada.

O modelo de Cox foi aplicado a um conjunto de dados de pacientes com câncer de pulmão em estado avançado, como forma de comprovar sua eficácia ao nos permitir inferir o quanto as covariáveis podem estar relacionadas ao evento de interesse de um estudo, no caso, a morte do paciente. A aplicação e os resultados encontrados foram satisfatórios, pois através da análise da significância das covariáveis e posteriormente a análise dos resíduos, foi encontrado o melhor modelo.

Como pesquisa futura seria interessante a utilização de outros modelos também ligados a análise de sobrevivência que permitam inferir também sobre pacientes que nunca irão chegar ao evento de interesse, que o caso da fração de cura, uma técnica relativamente nova, mas de grande importância para análise de sobrevivência, assim como o modelo de regressão de Cox.

REFERÊNCIAS

ALVES, Bruno Cardoso et al. **Modelos heterogéneos de sobrevivência: uma aplicação ao risco de crédito**. 2010. Tese de Doutorado. Disponível em < https://repositorio.iscte-iul.pt/bitstream/10071/4402/1/Alves%2c%20Modelos%20Heterog%C3%A9neos%20de%20Sobreviv%C3%AAncia.pdf> Acesso em 20 de Maio de 2017.

ARANHA, Guiomar Terezinha Carvalho et al. Identification of a statistical method as a quality tool: patient's length of stay in the operating room. **Brazilian Journal of Cardiovascular Surgery**, v. 24, n. 3, p. 382-390, 2009. Disponível em < http://www.scielo.br/scielo.php?pid=S0102-76382009000400019&script=sci_arttext&tlng=pt> Acesso em 13 de Junho de 2017.

BASTOS, Joana; ROCHA, Cristina. Análise de sobrevivência: Conceitos Básicos. **Arquivos de Medicina**, v. 20, n. 5-6, p. 185-187, 2006.Disponível em < http://www.scielo.mec.pt/pdf/am/v20n5-6/v20n5-6a07.pdf> Acesso em 26 de Abril de 2017.

BELLOC, Nedra B.; BRESLOW, Lester. Relationship of physical health status and health practices. **Preventive medicine**, v. 1, n. 3, p. 409-421, 1972 Disponível em < http://www.sciencedirect.com/science/article/pii/009174357290014X> Acesso em 04 de Junho de 2017.

BÖHMER, P. E. Theorie der unabhängigen Wahrscheinlichkeiten. In: **Rapports Memoires et Proces verbaux de Septieme Congres International dActuaires Amsterdam**. 1912. p. 327-343.

BOTELHO, Francisco; SILVA, Carlos; CRUZ, Francisco. Epidemiologia explicada–análise de sobrevivência. **Acta Urológica**, v. 26, n. 4, p. 33-38, 2009.Disponível em < http://www.apurologia.pt/acta/4-2009/epidem-explic.pdf> Acesso em 18 de Junho de 2017.

BRESLOW, Norman et al. A large sample study of the life table and product limit estimates under random censorship. **The Annals of Statistics**, v. 2, n. 3, p. 437-453, 1974. Disponível em < http://projecteuclid.org/download/pdf_1/euclid.aos/1176342705> Acesso em 25 de Junho de 2017.

BUSTAMANTE-TEIXEIRA, Maria Teresa; FAERSTEIN, Eduardo; DO ROSÁRIO LATORRE, Maria. Técnicas de análise de sobrevida Survival analysis techniques. **Cad. Saúde Pública**, v. 18, n. 3, p. 579-594, 2002. Disponível em < https://www.researchgate.net/profile/Eduardo_Faerstein/publication/26359790_Tecnicas_de_analise_de_sobrevida/links/0deec52048fa362ebb000000/Tecnicas-de-analise-de-sobrevida.pdf> Acesso em 24 de Junho de 2017.

CARVALHO, Marilia Sá et al. **Análise de Sobrevivência: teoria e aplicações em saúde**. SciELO-Editora FIOCRUZ, 2011.

CÉSAR, Kelly Araújo. Análise estatística de sobrevivência: um estudo com pacientes com câncer de mama. 2005. Disponível em < https://repositorio.ucb.br/jspui/bitstream/10869/1713/1/Kelly%20Araujo%20Cesar.pdf> Acesso em 14 de junho de 2017.

COLOSIMO, E. A.; GIOLO, S. R. Modelo de regressão de Cox. **Colosimo EA, Giolo SR. Análise de sobrevivência aplicada. São Paulo: Edgard Blücher**, p. 155-200, 2006.

COX, D. R. et al. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society**, v. 34, p. 187-220, 1972.

DANTAS, Maria Aldilene et al. Modelo de regressão Weibull para estudar dados de falha de equipamentos de sub-superfície em poços petrolíferos. **Production**, v. 20, n. 1, p. 127-134, 2010. Disponível em <http://www.redalyc.org/pdf/3967/396742038011.pdf> Acesso em 23 de Junho de 2017.

ESTEVÃO FREIRE, Elisabeth E. C. Monteiro e Julio C. R. Cyrino, Instituto de Macromoléculas, Universidade Federal do Rio de Janeiro, ex.? 68.525,21945-900 - Rio de Janeiro, RJ - Polímeros: Ciência e Tecnologia - Jul/Set-94 – Disponível em <http://revistapolimeros.org.br/files/v4n3/v4n3a02.pdf > Acesso em 25 de junho de 2017.

DE OLIVEIRA SANTOS, Rayany; NAKANO, Eduardo Yoshio. Análise do tempo de permanência de trabalhadores no mercado de trabalho do Distrito Federal via modelo de riscos proporcionais de Cox e Log-normal. **Rev. Bras. Biom**, v. 33, n. 4, p. 570-584, 2015. Disponível em < http://jaguar.fcav.unesp.br/RME/fasciculos/v33/v33_n4/A10_Santos_Nakano.pdf> Acesso em 30 de Junho de 2017.

EFRON, Bradley. The efficiency of Cox's likelihood function for censored data. **Journal of the American statistical Association**, v. 72, n. 359, p. 557-565, 1977. Disponível em < http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1977.10480613> Acesso em 01 Julho de 2017.

FAREWEL, V. T.; PRENTICE, Ross L. The approximation of partial likelihood with emphasis on case-control studies. **Biometrika**, v. 67, n. 2, p. 273-278, 1980. Disponível em < https://academic.oup.com/biomet/article-abstract/67/2/273/314682> Acesso em 26 de Maio de 2017.

FOWLER, Robert G.; DEGNEN, Gerald E.; COX, Edward C. Mutational specificity of a conditional Escherichia coli mutator, mutD5. **Molecular and General Genetics MGG**, v. 133, n. 3, p. 179-191, 1974. Disponível em < https://link.springer.com/article/10.1007/BF00267667> Acesso em 27 de Junho de 2017.

FRANCESCHINI, J. SANTOS, A. A., MOUALLEN, I. E., JAMLIK, S., UEHARA, C., FERNANDES, A. L. G., SANTORO, I. L. **Avaliação da qualidade de vida em pacientes com câncer de pulmão através da aplicação do questionário Medical Outcomes Study-36-item short-Form Health Survey**. Jornal Brasileiro Pneumologia, vol. 34, n. 6, 2008.

FUGL-MEYER, Axel R. et al. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. **Scandinavian journal of rehabilitation medicine**, v. 7, n. 1, p. 13-31, 1975. Disponível em < http://europepmc.org/abstract/med/1135616> Acesso em 20 de Junho de 2017.

GOUVÊA, Graziela Dutra Rocha; DE OLIVEIRA, Fernando Luiz Pereira; VIVANCO, Mário Javier Ferrua. Análise de eventos competitivos: uma aplicação aos dados de hemodiálise da cidade de Lavras-MG. **Rev. Bras. Biom., São Paulo**, v. 27, n. 3, p. 491-500, 2009. Disponível em < http://jaguar.fcav.unesp.br/RME/fasciculos/v27/v27_n3/A10_Graziela.pdf> Acesso em 28 de Maio de 2017.

KAPLAN, Edward L.; MEIER, Paul. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, v. 53, n. 282, p. 457-481, 1958.

KLEIN, John P.; MOESCHBERGER, Melvin L. **Survival analysis: techniques for censored and truncated data**. Springer Science & Business Media, 2005.

LEE, Elisa T.; WANG, John. **Statistical methods for survival data analysis**. John Wiley & Sons, 2003. Disponível em < https://books.google.com.br/books?hl=pt-BR&lr=&id=3QiBBonpRW0C&oi=fnd&pg=PR7&dq=+Statistical+Methods+for+Survival+Data+Analysis&ots=KbSFATg2ss&sig=a9HN1A3V7OLAT7fzqiZKQiYq5uE#v=onepage&q=Statistical%20Methods%20for%20Survival%20Data%20Analysis&f=false> Acesso em 30 de Junho de 2017.

LINDSEY, Jane C.; RYAN, Louise M. Methods for interval-censored data. **Statistics in medicine**, v. 17, n. 2, p. 219-238, 1998. Disponível em < http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19980130)17:2%3C219::AID-SIM735%3E3.0.CO;2-O/full> Acesso em 24 de Junho de 2017.

LOPRINZI CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7, 1994.

PETO, Richard. Rank tests of maximal power against Lehmann-type alternatives. **Biometrika**, v. 59, n. 2, p. 472-475, 1972. Disponível em < https://academic.oup.com/biomet/article-abstract/59/2/472/325639> Acesso em 01 de Julho de 2017.

PORTILHO, Carolina Marques. **Estimaçao da Persistência de Segurados de Planos de Previdência Privada Via Modelos de Sobrevivência**. 2013. Tese de Doutorado. PUC-Rio. Disponível em < http://www.dbd.puc-rio.br/pergamum/tesesabertas/1021478_2013_completo.pdf> Acesso em 27 de Maio de 2017.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

SCHOENFELD, David. Partial residuals for the proportional hazards regression model. **Biometrika**, v. 69, n. 1, p. 239-241, 1982.

SCHNEIDER, Ione Jayce Ceola; D'ORSI, Eleonora. Sobrevida em cinco anos e fatores prognósticos em mulheres com câncer de mama em Santa Catarina, Brasil. **Cadernos de Saúde Pública**, v. 25, n. 6, p. 1285-1296, 2009. Disponível em

http://www.scielosp.org/scielo.php?pid=S0102-311X2009000600011&script=sci_abstract&tlng=pt > Acesso em 19 de Junho de 2017.

SILVA BRITO, Giovani Antonio; ASSAF NETO, Alexandre. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade & Finanças-USP**, v. 19, n. 46, 2008. Disponível em < http://www.producao.usp.br/bitstream/handle/BDPI/6154/art_BRITO_Modelo_de_classificacao_de_risco_de_credito_2008.pdf?sequence=1> Acesso em 05 de Junho de 2017.

THERNEAU, Terry M (2015). _A Package for Survival Analysis in S_. version 2.38, <URL: http://CRAN.R-project.org/package=survival>.

THERNEAU, Terry M.; GRAMBSCH, Patricia M. The Cox model. In: **Modeling survival data: extending the Cox model**. Springer New York, 2000. p. 39-77. Disponível em < https://link.springer.com/chapter/10.1007/978-1-4757-3294-8_3>