UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

LUCAS CARDOSO PEREIRA

ANÁLISE DE SOBREVIVÊNCIA APLICADA A DADOS DE CÂNCER DE MAMA EM UM HOSPITAL DE REFERÊNCIA DE CAMPINA GRANDE

CAMPINA GRANDE - PB

2019

LUCAS CARDOSO PEREIRA

ANÁLISE DE SOBREVIVÊNCIA APLICADA A DADOS DE CÂNCER DE MAMA EM UM HOSPITAL DE REFERÊNCIA DE CAMPINA GRANDE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. TIAGO ALMEIDA DE OLIVEIRA

CAMPINA GRANDE - PB

2019

P436a Pereira, Lucas Cardoso.

Análise de sobrevivência aplicada a dados de câncer de mama em um Hospital de Referência de Campina Grande [manuscrito] / Lucas Cardoso Pereira. - 2019.

25 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2019.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira , Coordenação do Curso de Estatística - CCT."

1. Análise de sobrevivência. 2. Kaplan Meier. 3. Câncer de mama. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

LUCAS CARDOSO PEREIRA

ANÁLISE DE SOBREVIVÊNCIA APLICADA A DADOS DE CÂNCER DE MAMA EM UM HOSPITAL DE REFERÊNCIA DE CAMPINA GRANDE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 09 de dezembro de 2019.

BANCA EXAMINADORA

Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba

Ednário Barbosa de Medonça
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Agradecimentos

A Deus por ter me dado saúde e força para superar as dificuldades.

Aos meus pais, Marcos Pereira e Ivanilda Cardoso, e minha irmã, Alessandra Cardoso, pelo amor, incentivo e apoio incondicional.

A minha noiva, Ana Beatriz Aires Pereira, por trazer paz quando preciso, por me apoiar em minhas escolhas e por me incentivar a ser alguém melhor.

Ao professor Tiago Almeida de Oliveira, pela orientação neste trabalho, por me incentivar em relação aos estudos e ter me ajudado no que esteve ao seu alcance.

A todos os professores do departamento de Estatística, pelo incentivo aos alunos do curso, pelo empenho e comprometimento com o trabalho.

Aos colegas e amigos que pude fazer durante minha carreira acadêmica e passar muitos momentos bons, Antônio Leopoldo, Débora Cordeiro, Hiago Andrade, Iago Renan, Pedro Augusto e Sostenes Silva, obrigado pelo apoio, pela ajuda em diversas situações e pelo companheirismo.

Enfim, a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

"Deus é mais glorificado em nós quando estamos mais satisfeitos n'Ele."
(John Piper)

Resumo

Análise de sobrevivência é um ramo da estatística em que é utilizada quando se pretende analisar um fenômeno em relação a um período de tempo, isto é, ao tempo transcorrido entre um evento inicial, no qual um sujeito ou um objeto entra em um estado particular e um evento final, que modifica este estado. Assim, neste trabalho, foi feito um estudo de coorte retrospectivo, onde foi aplicado técnicas de analise de sobrevivência a dados referentes a mulheres que tiveram câncer de mama e fizeram o tratamento no Hospital da FAP (Fundação Assistencial da Paraíba) em Campina Grande, entre os anos de 2005 e 2015. Uma das técnicas utilizadas foi o calculo de curvas de sobrevivência estimadas por Kaplan Meier onde os resultados encontrados foram que para as mulheres que apresentaram o receptor estrogênio, receptor de progesterona e gene C-erb-B2 negativo tiveram tempo menor de sobrevivência em relação as mulheres que tinham estas características positivas. Outra técnica utilizada, foi o modelo de regressão de Cox, onde foi possível concluir que a cada acréscimo de uma unidade de radioterapia, o risco do paciente vir a óbito diminui 3,2%, já para cada acréscimo de uma unidade de hormonoterapia, o risco do paciente vir a óbito diminui 4,2% e o risco do paciente vir a óbito dado que ele apresente o gene C-erb-B2 positivo é 78,5% menor em relação aqueles que apresentaram este gene negativo.

Palavras-chave: Análise de sobrevivência, Kaplan Meier, câncer de mama.

Abstract

Survival analysis is a branch of statistics that is used when analyzing a phenomenon in relation to a period of time, that is, the time is between an initial event, in which a subject or an object enters a particular state and a final event, which modifies this state. Thus, in this study, a retrospective cohort study was performed, where survival analysis techniques were applied to data referring to women who had breast cancer and were treated at the Hospital da FAP (Paraíba Assistencial Foundation) in Campina Grande, Brazil, between the years 2005 and 2015. One of the techniques used was the calculation of survival curves estimated by Kaplan Meier where the results found were that women who presented estrogen receptor, progesterone receptor and negative C-erb-B2 gene had shorter survival time than women. that had these positive characteristics. Another technique used was the Cox regression model, where it was possible to conclude that with each addition of a radiotherapy unit, the patient's risk of death decreases 3.2%, while for each addition of a hormone therapy unit, the The patient's risk of dying decreases by 4.2% and the patient's risk of dying given that he has the positive C-erb-B2 gene is 78.5% lower than those who presented this negative gene.

Key-words: Survival Analysis, Kaplan Meier, Breast Cancer.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O câncer de mama vem se tornando um importante problema de saúde pública no Brasil e no mundo, devido à crescente incidência, morbidade e mortalidade, assim como ao alto custo do tratamento. Menarca precoce, nuliparidade, primeira gestação acima dos 30 anos de idade, uso de anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal são alguns dos vários fatores que já foram estabelecidos como desencadeadores no desenvolvimento do câncer de mama feminino. Segundo Educação (2017), no Brasil, em 2013, ocorreram 14.388 óbitos decorrentes do câncer de mama, sendo 14.207 em mulheres. Para o ano de 2016, em seu relatório mais recente, o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) apontou que são esperados 57.960 casos novos de câncer de mama, com taxa de incidência estimada de 56,20 casos a cada 100 mil mulheres.

A analise de sobrevivência ou sobrevida é uma área da estatística que tem como objetivo a análise do tempo até a ocorrência de um determinado evento de interesse, o qual é definido como falha ou desfecho, de modo que as observações são acompanhadas ao longo de períodos de tempo. Uma particularidade é a possibilidade da presença da censura, que nada mais é, que a observação parcial da resposta de interesse, ou seja, quando o indivíduo não sofre o evento durante o período do estudo. É justamente nas observações censuradas que a técnica análise de sobrevivência se diferencia de outras, como a regressão logística.O uso de co-variáveis afeta o tempo de vida dos indivíduos, daí surge a necessidade de fazer uso da analise de regressão. Na análise de sobrevivência pode-se coletar variáveis que representem a variabilidade existente na população, tais como idade, sexo, entre outras. Nestes casos pode-se adotar a priori duas abordagens os modelos paramétricos e os semi-parametríticos (COLOSIMO; GIOLO, 2006).

A pesquisa teve o objetivo de utilizar métodos de analise de sobrevivência semi-paramétrica para investigar a relação entre as co-variáveis e o tempo até a ocorrência do evento de interesse e determinar curvas de probabilidade de sobrevida dos pacientes (mulheres) que tiveram câncer. O estudo retrospectivo foi feito com autorização do comitê de ética aos prontuários de pacientes que tiveram câncer de mama e fizeram o tratamento no Hospital da FAP (Fundação Assistencial da Paraíba). As informações foram coletadas direto dos prontuários, onde os mesmos foram escolhidos de forma aleatória afim de obtermos uma amostra probabilística.

2 Materiais e Métodos

2.1 Amostragem

Segundo Levy e Lemeshow (2013), uma amostra probabilística caracteriza-se por ter a propriedade de que todo elemento da população tem uma probabilidade conhecida e diferente de zero de ser selecionada para compor a amostra, desta forma gerando estimadores não tendenciosos para os parâmetros populacionais. Por outro lado, amostragem não probabilística, baseia-se num plano amostral que não tem essa característica, desta maneira seu uso não permite avaliar a confiança ou a validade dos estimadores obtidos.

No método de amostragem aleatória simples, onde a premissa é de que cada componente da população estudada tenha a mesma chance de ser escolhido para compor a amostra. A técnica que garante essa igual probabilidade é a seleção aleatória de indivíduos, por exemplo, através de sorteio. O calculo para o tamanho da amostra é feito através equação:

$$n = \frac{N.(Z_{\alpha/2})^2.p.q}{(Z_{\alpha/2})^2.p.q + N.E^2} \quad . \tag{2.1}$$

em que n é o número de indivíduos da amostra, N estimativa do tamanho da população, p proporção populacional de indivíduos que pertencem a categoria que se está interessado em estudar, q proporção populacional de indivíduos que não pertencem a categoria que se está interessado em estudar, $Z_{\alpha/2}$ valor do erro α e E é a margem de erro ou o erro máximo de estimativa.

Segundo Levine, Berenson e Stephan (2000), Quando p e q, parâmetros populacionais, não são conhecidos, e o $\hat{p}$ e $\hat{q}$, parâmetros amostrais também não são conhecidos, recomenda substitui-los por 0,5, assim a formula para o calcula do tamanho da amostra fica:

$$n = \frac{N.(Z_{\alpha/2})^2.0,25}{(Z_{\alpha/2})^2.0,25 + N.E^2} \quad , \tag{2.2}$$

2.2 Análise de sobrevivência

Análise de sobrevivência é um ramo da estatística que analisa tempos até ao acontecimento de determinado evento: tempo que um individuo sobrevive a um determinado tratamento, tempo até o desenvolvimento de uma doença ou simplesmente tempo até a

morte. Segundo Colosimo e Giolo (2006), em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Esses eventos na maioria dos casos, indesejáveis e usualmente chamados de falhas. Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falhas, cuja característica importante é a presença de censura, que representa a observação parcial da resposta.

2.2.1 Censura

Dados censurados são dados coletados ao longo de um tempo pré-determinado, em que a informação obtida naqueles elementos ou indivíduos em que o evento não ocorre é apenas parcial. Por exemplo, em ensaios clínicos ao longo de um tempo, em que o resultado de interesse é o tempo até a reincidência de uma doença, no caso dos pacientes que ao termino do tempo pré-determinado não houve a reincidência da doença, ele censurou a direita. Além da censura a direita, pode-se ainda ocorrer outros tipos de censura como: censura a esquerda e a censura intervalar. Segundo Strapasson (2007), censura à esquerda ocorre quando o evento de interesse já aconteceu, ao se observar o indivíduo, ou seja, o tempo de vida é menor que o observado. Censura intervalar é quando sabe-se que o evento de interesse ocorreu em um intervalo te tempo, mas não se sabe o momento exato.

De acordo com Colosimo e Giolo (2006), o mecanismo de censura aleatória acontece quando no decorrer do estudo, em que o acompanhamento de alguns pacientes é interrompido por razão qualquer, diferente do evento de interesse, ou caso chegue o final do estudo e o evento não ocorreu. A censura do tipo I é um caso particular da aleatória onde nos estudos finalizados após o período pré-estabelecido de tempo foi verificado pacientes que ainda não falharam, já no mecanismo de censura do tipo II, o estudo é finalizado após a ocorrência de um número pré-estabelecido de falhas

Segundo Strapasson (2007), para análise de sobrevivência é necessário que as observações sejam representadas por um vetor (t_i, δ_i, x_i) em que, t_i é o tempo observado de falha ou censura e δ_i uma variável indicadora de falha, em que $\delta_i = 1$, o tempo observado corresponde a uma falha ou $\delta_i = 0$, corresponde a uma censura. Para cada indivíduo observado tem-se uma covariável x_i, onde *i=1,...,n* são observações representadas pelo par (t_i, δ_i).

$$\delta_i = \begin{cases} 1, & \text{quando } T \leq C, \\ 0, & \text{quando } T > C. \end{cases}$$

Dados censurados são representados por sinal “+".

2.2.2 Tempo

Tomando T uma variável aleatória, absolutamente contínua e não-negativa, em que o tempo de sobrevida $T \geq 0$ é expresso por meio de várias funções matematicamente equivalentes, onde, se uma delas é especificada, as outras podem ser derivadas. Segundo Colosimo e Giolo (2006) estas funções são usadas para descrever diferentes aspectos apresentado pelo conjunto de dados e utilizada para caracterizar o comportamento de dados de tempo de sobrevivência, em que t representa o tempo de falha especificada em análise de sobrevivência, cuja distribuição pode ser caracterizada por qualquer umas das seguintes funções: A função densidade de probabilidade, *f(t)*, a função de sobrevivência, *S(t)*, função de risco, *h(t)*.

A função de densidade de probabilidade é caracterizada pelo evento de interesse ao observar um individuo no intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo, que é definida como limite da probabilidade, (LOUZADA; DINIZ, 2012). A f.d.p. é dada por,

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{2.3}$$

em que $f(t) \geq 0$ para todo t e tem a área abaixo da curva igual a 1. Assim, $f(t)$ pode ser interpretada como a probabilidade de um indivíduo sofrer um evento em um intervalo instantâneo de tempo.

Enquanto a função de sobrevivência, denotada por *S(t)*, é definida como a probabilidade de um individuo sobreviver até um certo tempo t sem a ocorrência do evento. Sendo uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência, definida por

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = 1 - \int_0^t f(u)du, \tag{2.4}$$

sendo que $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$ e $F(t) = \int_0^t f(u)$ representa a função de distribuição acumulada.

A função de risco, ou taxa instantâneo de falha, descreve a forma com que a taxa de falha muda com o tempo, ou seja, demostra o risco do individuo falhar no tempo. É definida como o risco instantâneo de um indivíduo sofrer o evento entre o tempo t e $t + \Delta t$, dado que ele sobreviveu ao tempo t, uma definição formal é apresentada por Louzada e Diniz (2012),

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.5}$$

A função taxa de falha pode ser definida, em termos da função de distribuição *F(t)* e da função de densidade de probabilidade *f(t)*, da seguinte forma:

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.6}$$

A taxa instantânea de falha, por unidade de tempo, é fornecida pela função de risco, isto é, pode-se caracterizar classes especiais de distribuições de tempo de sobrevivência de acordo com o comportamento em relação ao tempo, conhecida como força de mortalidade ou taxa de mortalidade condicional. Estas funções são utilizadas na prática com o objetivo de descrever os aspectos apresentados pelo conjunto de dados. A função densidade de probabilidade é definida como a derivada da função densidade de probabilidade acumulada,

$$f(t) = \frac{\partial F(t)}{\partial t}. \tag{2.7}$$

Como $F(t) = 1 - S(t)$ pode-se escrever

$$f(t) = \frac{\partial[1 - S(t)]}{\partial t} = -S'(t) \ , \tag{2.8}$$

substituindo (2.6) em (2.4) obtêm-se,

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{\partial[logS(t)]}{\partial t}. \tag{2.9}$$

Dessa forma, tem-se,

$$\log S(t) = -\int_0^t h(u)du, \tag{2.10}$$

ou seja,

$$S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{2.11}$$

Uma outra função importante é a função risco acumulada, definida como,

$$H(t) = \int_0^t h(u)du. \tag{2.12}$$

Substituindo-se (2.10) em (2.9) tem-se que,

$$S(t) = \exp[-H(t)]. \tag{2.13}$$

Como, $\lim_{t\to\infty} S(t) = 0$ então

$$\lim_{t\to\infty} H(t) = \infty. \tag{2.14}$$

Além disso, de (2.4) e seleção de variáveis

$$f(t) = h(t)S(t). \tag{2.15}$$

Substituindo-se (2.11) em (2.12) tem-se

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right). \tag{2.16}$$

A expressão (2.13) é muito importante quando desenvolve-se os procedimentos de estimação somente sobre a função de risco.

2.3 Técnicas não paramétricas

A função de sobrevivência pode ser estimada considerando-se modelos paramétricos e técnicas não-paramétricas. As técnicas não-paramétricas podem ser utilizadas para verificar se o modelo paramétrico está bem ajustado. Segundo Colosimo e Giolo (2006) existem técnicas não-paramétricas para estimar parâmetros em análise de sobrevivência, obtendo a opção de ajustar os dados utilizando-se os modelos paramétricos probabilístico para tempo de falha.

Os estimadores de probabilidade de sobrevida, $\hat{S}(t)$, utilizados nos testes não-paramétricos se resumem em três: a tabela de vida ou acturial, que é uma das mais antigas técnicas estatística para estimar o tempo de falha, sendo utilizada apenas em grandes amostras, o teste de Kaplan-Meier e o estimador de Nelson-Aalen que apresenta propriedades parecidas com o Kaplan-Meier.

2.3.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier permite realizar testes de hipóteses que não requerem pressupostos sobre a distribuição dos dados, é usado para analisar dados medidos apenas numa escala ordinal podendo ocorrer para dados categorizados que são medidos em escala nominal. É adequado para amostras provenientes de diversas populações, sendo usado se todas as observações falharam, ou seja, não existiram censuras. As observações censuradas informam que o tempo até a falha é maior do que aquele que foi registrado, assim o estimador não-paramétrico de Kaplan-Meier considera a ocorrência de falhas distintas em intervalos de tempo, onde os tempos de sobrevivência são ordenados, isto é, $t_1 \leq t_2 \leq t_3 \leq ... \leq t_k$, podendo ocorrer mais de uma falha no mesmo tempo, expressado por Colosimo e Giolo (2006) por,

i) $t_1 \leq t_2 \leq t_3 \leq ... \leq t_k$ tempos distintos e ordenados de falha;

ii) d_j: número de falhas até o tempo t_j, *j= 1, 2,...,k*;

iii) n_j:número de itens sob risco, ou seja, os indivíduos não falharam e não censurados até t_j.

Segundo CESAR (2005) e Colosimo e Giolo (2006) o estimador $\widehat{R}(t)$ de Kaplan-Meier, é definido por,

$$\widehat{R}(t) = \left(\frac{n_1 - d_1}{n_1}\right) \cdot \left(\frac{n_2 - d_2}{n_2}\right) \times \cdots \times \left(\frac{n_{t_0} - d_{t_0}}{n_{t_o}}\right) = \prod_{j,tj<t} \frac{n_j - d_j}{n_j}$$

em que t_o é o maior tempo de falha menor que t.

As principais propriedades do estimador são:

i) Não viciado para amostras grandes;

ii) Fracamente consistente;

iii) Converge assintoticamente para um processo gaussiano;

iv) Estimador de máxima verossimilhança de *S(t)*.

2.4 Modelo de Cox

Entre os métodos para se aplicar modelo a variável tempo de vida se encontram os modelos de regressão que usam diversas variáveis explicativas, chamadas co-variáveis, e indicam o efeito destas variáveis sobre o tempo de sobrevivência. Entre os modelos de regressão tem-se os modelos de riscos proporcionais.

O modelo apresentado por Cox (1972) é o mais utilizado em estudo clínicos, devido a sua versatilidade. A versatilidade deve-se ao fato de que a estrutura deste modelo conta com um componente não-paramétrico e outro paramétrico, justificando sua denominação de modelo semi-paramétrico, e ele é dado por:

$$\lambda(t) = \lambda_0(t)g(x'\beta), \tag{2.17}$$

sendo g uma função não-negativa que deve ser especificada de modo que $g(0) = 1$. O termo $\lambda_0(t)$ é uma função não negativa do tempo, representando o componente não paramétrico do modelo, que não é especificado. Este componente é usualmente denominado função de base ou basal. O componente paramétrico é frequentemente expresso por:

$$g(x'\beta) = exp\{x'\beta\} \tag{2.18}$$

em que, β é o vetor de parâmetros associados às p co-variáveis. O modelo de cox tem a suposição de riscos proporcionais, que é a taxa de falha de dois indivíduos diferentes ao longo do tempo é constante ao longo do tempo. A função de sobrevivência dado os vetores de co-variáveis é dada por:

$$S(t|x) = [S_0(t)]^{exp(X\beta)}, \tag{2.19}$$

com função de sobrevivência de base definida como:

$$S_0(t) = exp- \int_0^t \lambda_0(y)dy = exp\{-H_0(t)\} \tag{2.20}$$

Devido ao componente não paramétrico, Cox (1975) formalizou o método de máxima verossimilhança parcial, que elimina este componente não paramétrico do modelo.

2.4.1 Método da Máxima Verossimilhança Parcial

A função de verossimilhança a ser utilizada para se fazer inferências acerca dos parâmetros do modelo é, formada pelo produto de todos os termos, descritos na equação abaixo, associados aos tempos de falha, isto é,

$$L(\beta) = \prod_{i=1}^{k} \frac{exp\{x_i^{'}\beta\}}{\sum_{j\in R(t_i)} exp\{x_j^{'}\beta\}} = \prod_{i=1}^{n} \Big(\frac{exp\{x_i^{'}\beta\}}{\sum_{j\in R(t_i)} exp\{x_j^{'}\beta\}} \Big)^{\delta_i}, \tag{2.21}$$

em que, δ_i é o indicador de falha. $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i. Os valores de β que maximizam a função de verossimilhança parcial, $L(\beta)$, são obtidos resolvendo-se o sistema de equações definido por $U(\beta) = 0$, em que $U(\beta)$ é o vetor escore de derivadas de primeira ordem da função $l(\beta) = log(L(\beta))$, isto é,

$$U(\beta) = \sum_{i=1}^{n} \delta_i \Big[x_i - \frac{\sum_{j\in R(t_i)} x_j exp\{x_j^{'}\hat{\beta}\}}{\sum_{j\in R(t_i)} exp\{x_j^{'}\hat{\beta}\}} \Big]. \tag{2.22}$$

2.5 Seleção de Modelos

2.5.1 Critério de Informação de Akaike

A escolha do modelo apropriado, do ponto de vista estatístico, é um tópico extremamente importante na análise de dados. O Critério de Informação de Akaike (AIC) busca ajustar o modelo mais parcimonioso possível, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que busque explicar tão bem ou até mesmo melhor que o modelo saturado o comportamento da variável resposta.

Segundo Moore (2016), um das melhores formas de se avaliar os modelos estatísticos é através do cálculo de AIC, que consiste em avaliar a verossimilhança do modelo, penalizado pelo número de parâmetros. O objetivo é encontrar o modelo tal que a quantidade abaixo seja minima.

$$AIC = -2l(\hat{\beta}) + 2k, \tag{2.23}$$

em que, $l(\hat{\beta})$ é a verossimilhança do modelo e k é o número de parâmetros.

Para Klein e Moeschberger (2005), a inclusão de variáveis no modelo causa uma diminuição no valor de AIC, entretanto, em algum ponto, o critério passa a aumentar, indicando que a inclusão de determinadas variáveis é desnecessário e não contribuíra para as estimativas dos parâmetros.

2.5.2 Seleção de variáveis

Segundo Silva et al. (2006), um dos modelos de seleção de variáveis mais utilizados é o setpwise (seleção passo a passo), que nada mais é que uma modificação do método forward (seleção à frente).

No forward começa com a escolha da variável independente que melhor explica a variável dependente comparando-se, pelo teste da razão de verossimilhança, o modelo ajustado apenas com o intercepto e cada modelo univariado. Logo, a primeira variável escolhida é a que apresenta menor p-valor. O próximo passo é escolher uma segunda variável que produza o maior aumento na razão de verossimilhança quando adicionada ao modelo. Novamente aplica-se o teste da razão de verossimilhança para verificar se a contribuição desta nova variável é significativa. O processo continua até que nenhuma variável acrescida no modelo cause aumento significativo na razão de verossimilhança. Uma característica importante desse procedimento é que, uma vez que a variável foi selecionada e incluída por ser significativa, ela não deve mais ser excluída

Para o método de seleção stepwise, o procedimento começa com o método forward, mas depois que a segunda variável entra no modelo, o teste da razão de verossimilhança é realizado para verificar se a primeira variável permanece no modelo. Caso permaneça, uma terceira variável é selecionada da mesma forma que no procedimento do forward. Se uma terceira variável entra no modelo, testa-se para verificar se as duas primeiras continuam no modelo. Pode acontecer que uma delas ou as duas sejam eliminadas. Tenta-se então a inclusão de uma nova variável. Caso entre, tenta-se a eliminação das que já estão no modelo. O procedimento acaba quando não se consegue nem adicionar, nem eliminar variáveis.

3 Aplicação

Os dados são referentes a pacientes, mulheres, do Hospital da FAP (Fundação Assistencial da Paraíba), que tiveram câncer de mama. Com auxilio da coordenadora do setor de arquivo do hospital, foi sugerido trabalhar com os pacientes a partir do ano de 2005, por os prontuários estarem melhores organizados a partir deste ano. Desta maneira, o estudo teve como inicio os pacientes que deram entrada a partir do ano de 2005 até o ano de 2015, outra informação importante foi conseguida com a coordenadora do setor, que é o número médio de pacientes por ano que fizeram tratamento de câncer de mama no hospital da FAP que foi de aproximadamente 200 mulheres e 2 homens por ano. Assim levando em consideração o erro α à 5% e a margem de erro de 7,5%, temos que o tamanho minimo da amostra era de 158 observações, neste estudo foi utilizado 161 observações. As variáveis coletadas foram, a data da primeira consulta, a data da última consulta, a data da morte do paciente, a idade, o número de hormonoterapia, quimioterapia e radioterapia, o local ou lado da mama em que se encontrava o tumor, o receptores hormonais(receptor de estrogênio e progesterona), a proteína k1-67 e p53, e o gene C-erb-B2.

Os pacientes foram categorizados pelo tempo até o óbito, assim, o tempo do óbito foi calculado a partir da data em que o mesmo deu entrada no hospital até a data de sua morte e o tempo da censura foram os dias entre a data da entrada no hospital e a data da ultima consulta, para aqueles pacientes que não vieram a obter o evento em interesse até o fim da do estudo, no caso o ano de 2015, estes censuraram. Desta forma o banco de dados pode ser dividido em dois grupos, o grupo 1 é formado por aqueles pacientes que não censuram, seja eles terem sido curados, desistido de continuar o tratamento ou por algum outro motivo não terem continuado o tratamento, e o outro é formado pelos pacientes que vieram a óbito devido ao câncer de mama.

As Tabelas 1 e 2 apresentam a média, o desvio padrão, mínimo, máximo, 1º, 2º e 3º quartil das variáveis quantitativas. A partir da Tabela 1, pode-se concluir que a idade média dos pacientes que censuraram era de aproximadamente 61 anos, o tempo médio de atendimento até a censura era de 1200 dias, metade dos pacientes tomaram mais de 52 hormonoterapia, o número médio de quimioterapia foi de aproximadamente 6 por paciente e metade dos pacientes fizeram mais que 28 radioterapia.

Tabela 1 – Estatísticas descritivas das variáveis quantitativas refentes ao grupo censura

Variáveis	**Média**	**D.P.**	**Mín.**	**1º Q.**	**2º Q.**	**3º Q.**	**Máx.**
Idade	60,62	11,55	37,00	51,00	63,00	68,00	84,00
Tempo	1119,69	884,92	34,00	230,50	1093,00	1843,00	2987,00
nº de hormoterapia	34,87	29,41	0,00	0,00	52,50	61,00	75,00
nº de quimioterapia	6,21	14,35	0,00	0,00	0,00	6,50	67,00
nº de radio terapia	27,46	10,71	0,00	25,00	28,00	30,00	54,00

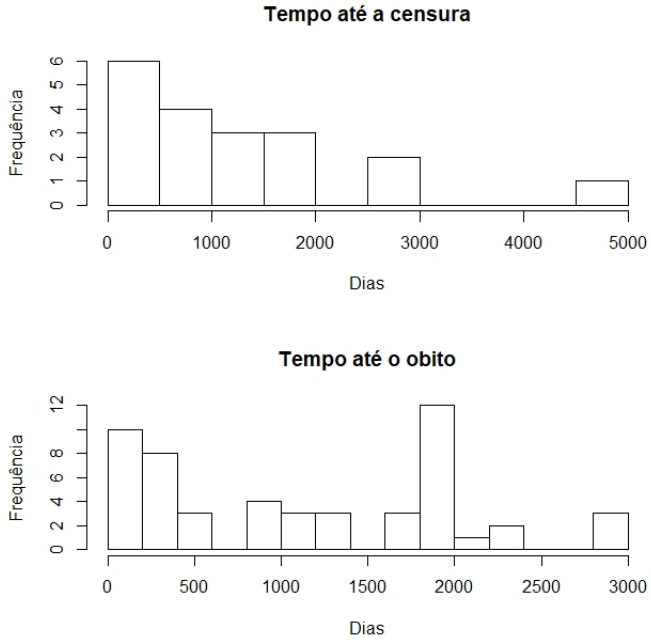
A partir da Tabela 2, tem-se que a idade média dos pacientes que morreram era de aproximadamente 58 anos, o tempo médio de atendimento até o paciente vir a óbito foi de 1250 dias, sendo que metade dos pacientes morreu com 888 dias após ter dado entrada no hospital, metade dos pacientes tomou mais de 5 hormonoterapia, o número médio de quimioterapia foi de aproximadamente 19 por paciente e metade dos pacientes foi submetido a mais que 25 radioterapia.

Tabela 2 – Estatísticas descritivas das variáveis quantitativas refentes ao grupo óbito

Variáveis	**Média**	**D.P.**	**Mín.**	**1º Q.**	**2º Q.**	**3º Q.**	**Máx.**
Idade	57,89	11,91	39,00	49,00	60,00	64,00	84,00
Tempo	1249,26	1160,34	49,00	445,50	888,00	1718,50	4753,00
nº de hormoterapia	20,37	29,86	0,00	1,00	5,00	34,00	100,00
nº de quimioterapia	19,05	17,50	0,00	4,50	18,00	27,50	70,00
nº de radio terapia	19,58	16,27	0,00	0,00	25,00	30,00	49,00

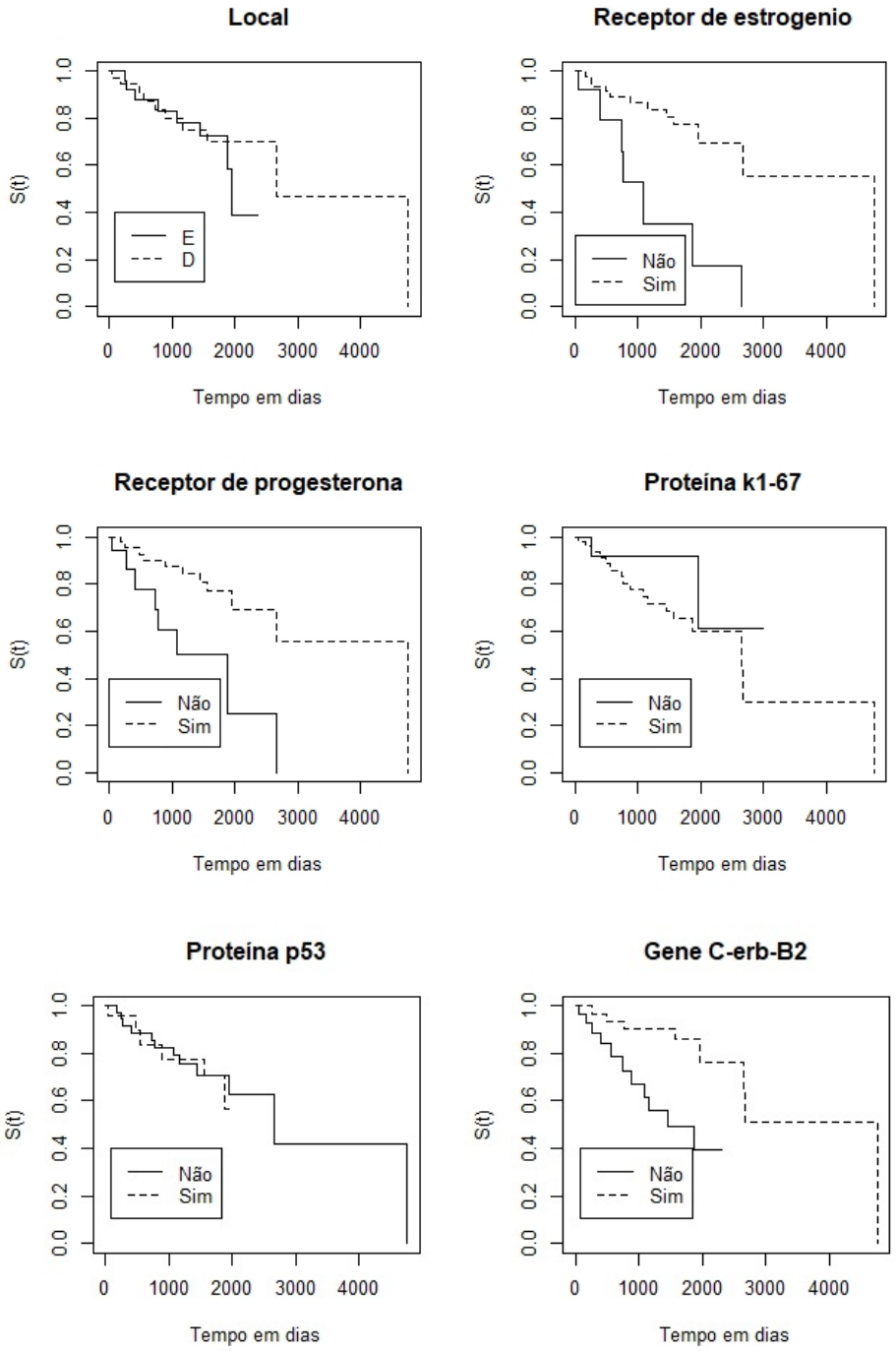
A Figura 1 apresenta o histograma do tempo até a censura e o tempo até o óbito. Desta forma é possível notar que os paciente censuram mais frequentemente nos primeiros 2000 dias, já nos primeiros 500 dias há uma grande frequência de óbito.

Figura 1 – Tempo de atendimento



A Figura 2 apresenta as curvas de sobrevivência para as variáveis local ou lado da mama onde se encontra o tumor, receptor de progesterona e estrogênio, gene C-erb-B2, proteína P53 e k1-67. Pode-se destacar através do teste de log-rank que as variáveis receptor de estrogênio (valor-p $< 0{,}001$), de progesterona (valor-p $= 0{,}003$) e gene C-erb-B2 (valor-p $= 0{,}003$) apresentaram diferença significativa em relação as categorias, sendo que para as mulheres que apresentaram o receptor estrogênio, receptor de progesterona e gene C-erb-B2 negativo tiveram tempo menor de sobrevivência em relação as mulheres que tiveram essas características positivas.

Figura 2 – Curvas de Kaplan-Meier.



Para a construção do modelo de Cox foram utilizadas as variáveis local, idade, numero de radioterapia, de quimioterapia e de hormonoterapia, receptor estrogênio e progesterona, Gene C-erb-B2, proteína k1-67 e p53. Os coeficientes do modelo inicial estão na Tabela 3 e o AIC do modelo foi de 110,03.

Tabela 3 – Modelo de Cox inicial.

Variáveis	**Coef.**	**Exp(coef.)**	**D.P.(coef.)**	**Valor-p**
Local (E)	0,132	1,141	0,579	0,820
Idade	0,012	1,012	0,026	0,657
Nº de radio	-0,038	0,963	0,021	0,063
Nº de quimio	0,013	1,013	0,020	0,511
Nº de hormono	-0,044	0,957	0,014	0,002
Recptor estrogenio (positivo)	1,423	4,149	1,584	0,369
Recptor de progesterona(positivo)	-1,747	0,174	1,486	0,240
Gene C-erb-B2 (positivo)	-2,097	0,123	0,791	0,008
Proteína k1-67 (positivo)	-0,750	0,472	1,267	0,554
Proteína p53 (positivo)	-0,019	0,981	0,617	0,976

Após isso, em busca de um melhor modelo, foi utilizado o critério de seleção de variáveis stepwise, este novo modelo apresentou AIC igual a 98,63 que foi relativamente menor que o do modelo inicial e todas as co-variáveis foram significativas.

Assim, a partir dos resultados apresentados na Tabela 4, pode-se destacar que a cada acréscimo de uma unidade de radioterapia, o risco do paciente vir a óbito diminui 3,2%, já para cada acréscimo de uma unidade de hormonoterapia, o risco do paciente vir a óbito diminui 4,2% e o risco do paciente vir a óbito dado que ele apresente o gene C-erb-B2 positivo é 78,5% menor que aqueles pacientes que apresentaram este gene negativo.

Tabela 4 – Modelo de Cox final

Variáveis	**Coef.**	**Exp(coef.)**	**D.P.(coef.)**	**Valor-p**
Nº de radio	-0,033	0,968	0,018	0,071
Nº de hormono	-0,042	0,958	0,011	0,000
Gene C-erb-B2 (positivo)	-1,539	0,215	0,590	0,009

Segundo Eisenberg e Koifman (2001), o gene C-erbB-2 tem sido extensamente estudado em carcinomas de mama desde que, demonstraram uma associação entre a sua amplificação e um mau prognóstico (SLAMON et al., 1987), assim corroborando com o que modelo de Cox enfatizou que o gene C-erb-B2 é uma variável que indica influência na morte por câncer de mama.

4 Conclusão

Podemos concluir que através das curvas de sobrevivência estimadas por Kaplan e Meier, pode-se destacar, através do teste de log-rank, que as variáveis receptor de estrogênio (valor-p $<$ 0,001), de progesterona (valor-p = 0,003) e gene C-erb-B2 (valor-p = 0,003) apresentaram diferença significativa em relação as categorias, sendo que para as mulheres que apresentaram o receptor estrogênio, receptor de progesterona e gene C-erb-B2 negativo tiveram tempo menor de sobrevivência em relação as mulheres que tiveram essas características positivas. Quanto ao modelo de Cox final, as variáveis que compuseram o modelo foram: o numero de radioterapia, o numero de hormonoterapia e o gene C-erb-B2. Ainda sobre o modelo de Cox, podemos concluir que a cada acréscimo de uma unidade de radioterapia, o risco do paciente vir a óbito diminui 3,2%, já para cada acréscimo de uma unidade de hormonoterapia, o risco do paciente vir a óbito diminui 4,2% e o risco do paciente vir a óbito dado que ele apresente o gene C-erb-B2 positivo é 78,5% menor em relação aqueles que apresentaram este gene negativo.

Referências

CESAR, K. A. Análise estatística de sobrevivência: um estudo com pacientes com câncer de mama. 2005. 12 f. *Monografia (Graduação)–Universidade Católica de Brasília, Brasília*, 2005. Citado na página 15.

COLOSIMO, E.; GIOLO, S. Análise de sobrevivência aplicada. 1ª edição. *São Paulo: Editora Edgard Blücher*, 2006. Citado 4 vezes nas páginas 10, 12, 13 e 15.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado na página 16.

COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. Citado na página 16.

EDUCAÇÃO, I. C. de. *ABC do câncer: abordagens básicas para o controle do câncer*. [S.l.]: Rio de Janeiro: INCA, 2017. Citado na página 10.

EISENBERG, A. L. A.; KOIFMAN, S. Câncer de mama: marcadores tumorais (revisão de literatura). *Rev Bras Cancerol*, v. 47, n. 4, p. 377–88, 2001. Citado na página 23.

KLEIN, J.; MOESCHBERGER, L. Mm (2005) survival analysis: Techniques for censored and truncated data. *New York, NY*, 2005. Citado na página 17.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. Estatística: teoria e aplicações. *Rio de janeiro: LTC*, v. 811, 2000. Citado na página 11.

LEVY, P. S.; LEMESHOW, S. *Sampling of populations: methods and applications*. [S.l.]: John Wiley & Sons, 2013. Citado na página 11.

LOUZADA, F.; DINIZ, C. Modelagem estatística para risco de crédito. *2º Edição, São Paulo: ABE-Associação Brasileira de Estatística*, 2012. Citado na página 13.

MOORE, D. F. *Applied survival analysis using R*. [S.l.]: Springer, 2016. Citado na página 17.

SILVA, C. A. M. et al. Exploração de métodos de seleção de variáveis pela técnica de regressão logística para análise de dados epidemiológicos. [sn], 2006. Citado na página 17.

SLAMON, D. J.; CLARK, G. M.; WONG, S. G.; LEVIN, W. J.; ULLRICH, A.; MCGUIRE, W. L. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *science*, American Association for the Advancement of Science, v. 235, n. 4785, p. 177–182, 1987. Citado na página 23.

STRAPASSON, E. *Comparaçao de Modelos com Censura Intervalar em Análise de Sobrevivência. 2007*. Tese (Doutorado) — Tese de Doutorado-Escola Superior de Agricultura Luiz de Queiroz . . . , 2007. Citado na página 12.