Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

DIEGO ALVES GOMES

INTRODUÇÃO A ANÁLISE DE SOBREVIVÊNCIA COM APLICAÇÃO A DADOS DE TRANSPLANTADOS DE MEDULA ÓSSEA

Campina Grande
10 de Dezembro de 2014

DIEGO ALVES GOMES

INTRODUÇÃO A ANÁLISE DE SOBREVIVÊNCIA COM APLICAÇÃO A DADOS DE TRANSPLANTADOS DE MEDULA ÓSSEA

Trabalho Acadêmico Orientado apresentado ao curso de Bacharelado em Estatística do Departamento 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:
Tiago Almeida de Oliveira

Campina Grande
10 de Dezembro de 2014

G633i Gomes, Diego Alves.
Introdução a análise de sobrevivência com aplicação a dados de transplantados de medula óssea [manuscrito] / Diego Alves Gomes. - 2014.
46 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Tiago Almeida de Oliveira, Departamento de Estatística".

1. Análise de sobrevivência. 2. Distribuições paramétricas. 3. Censura - Estatística. I. Título.

21. ed. CDD 519.53

DIEGO ALVES GOMES

INTRODUÇÃO A ANÁLISE DE SOBREVIVÊNCIA COM APLICAÇÃO A DADOS DE TRANSPLANTADOS DE MEDULA ÓSSEA

Trabalho Acadêmico Orientado 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.

Aprovado em: 10 / 12 / 2014

Banca Examinadora:

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba - DE/CCT
Orientador

Prof. Dr. João Gil de Luna
Universidade Estadual da Paraíba - DE/CCT
Examinador

Prof. Ms. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba - DE/CCT
Examinador

Dedicatória

Dedico este trabalho a DEUS e a meus pais Vânia Maria e Antônio Gomes, ao meu amado irmão e amigo Diogo Alves e a meus tios, em especial a meu querido tio Carlos pela força nos momentos de duvidas e dificuldades, e a minha bondosa e querida tia Zefinha agradeço a DEUS por ter vocês em minha vida.

Agradecimentos

Agradeço a Deus, em primeiro lugar, pelas inúmeras vitorias que conquistei e ainda conquistarei, pela minha saúde e discernimento para pode seguir com meus objetivos e sonhos.

Ao professor e orientador Tiago Almeida de Oliveira pela paciência, estimulo, confiança e dedicação durante a orientação deste trabalho e do Pibic.

A minha amada e querida família meus pais Vânia e Antônio ao meu grande amigo e irmão Diogo, aos meus tios Carlos e Zefinha pelo apoio nos momentos de dificuldades.

A Universidade Estadual da Paraíba (UEPB) ao Centro de Ciências Tecnologia e em especial ao departamento de estatística e a todos os professores que ajudaram ao longo dos meus estudos.

Aos meus amigos em especial Alcione e Fábio pela turma dos três que formamos, pelos momentos em que passamos juntos de alegrias, felicidades, dificuldades e pelos sábados que passamos estudando e por ter me ajudado nos momentos estressantes que tive na minha graduação, a minha namorada Allana Lívia pela compreensão, paciência, carinho e amor, e aos meus queridos amigos Rosendo, Sidcleide, Sônia e a Leomir, agradeço a DEUS por ter conhecido vocês e a todos que de uma forma ou de outra me ajudaram durante minha graduação.

Resumo

O termo análise de sobrevivência refere-se a situações médicas envolvendo dados censurados. Em geral, os bancos de dados utilizados para este tipo de análise são constituídos de indivíduos suscetíveis e não suscetíveis a um determinado evento que pode ser cura, recidiva ou morte, etc. O termo censura refere-se ao acompanhamento de um indivíduo por um certo período de tempo e o mesmo vir a sair do estudo por motivos que não estão diretamente relacionadas com o objetivo da pesquisa, por exemplo um paciente acompanhado até ser interrompido o estudo, por diversos motivos tais como mudança de cidade, morte, desistência voluntária, entre outros. O objetivo da análise de sobrevivência é estimar, comparar, e interpretar as variáveis como o tempo de sobrevivência e as funções de risco. Para que se possa conhecer o comportamento de determinadas doenças ou fenômenos é necessário o acompanhamento dos indivíduos sobre risco e para tal informação é preciso ter um banco de dados que reúna as condições para se realizar a análise de sobrevivência e assim poder estimar parâmetros de distribuições e estimar funções de sobrevivência e risco. Os métodos estatísticos usados para facilitar a interpretação e melhorar o ajuste dos modelos propostos aos dados são variados, tais como: métodos não-paramétricos, semi-paramétricos, paramétricos. O auxílio de *softwares*, tais como o R são essenciais para a melhor aplicabilidade dos métodos. Neste estudo, aplicou-se os métodos não-paramétricos de Kaplan-Meier, Nelson Aalen, teste log-rank, distribuições paramétricas de weibull, exponencial e log-normal, afim de se estimar a função de sobrevivência para um banco de dados constituído de 137 pacientes (99 AML, 38 ALL), os quais são uma distinção entre os tipos de cancer, leucemia linfoblástica aguda e leucemia linfoblástica (ALL), estes classificados em 3 grupos de doença. Os resultados encontrados sugerem que a distribuição log-normal foi a que melhor se ajustou aos dados em estudo, além de que os indivíduos do grupo 3 leucemia linfoblástica (All) alto risco, foram os indivíduos que tiveram maior probabilidade de recidiva em relação aos outros grupos.

Palavras-Chave: Distribuições paramétricas; Kaplan-Meier; Leucemia Linfoblástica.

Abstract

The term survival analysis relates to medical situations involving censored data. In general, the databases used for this type of analysis consist of susceptible individuals and not susceptible to a particular event that may be healing, recurrence or death, etc. The term censorship refers to the monitoring of an individual for a certain period of time and even come out of the study for reasons other that are not directly related to the purpose of the research, for example a patient together until interrupted the study, for various reasons such as moving to another city, the patient's death, voluntary withdrawal, among others. The goal of survival analysis is to estimate, compare, and interpret the variables as the survival time and risk functions. In order to understand the behavior of certain diseases or phenomena monitoring of individuals about risk and such information is required you must have a database that meets the conditions to perform survival analysis and thus to estimate distributions of parameters and estimate survival and risk functions. Statistical methods used to facilitate interpretation and improve the fit of the models proposed to the data are varied, such as non-parametric methods, semi-parametric, parametric. The aid software such as R, are essential for better applicability of the methods. In this study, we applied the non-parametric methods of Kaplan-Meier, Nelson Aalen, log-rank test, parametric Weibull distributions, exponential and log-normal, in order to estimate the survival function for a database consisting of 137 patients (99 AML, 38 ALL), which is a distinction between the types of cancer, acute leukemia and linfobastica linfobastica leukemia (ALL), can be classified in three disease groups. The results suggest that the log-normal distribution was the best fit to the data under study, and that individuals in the group 3 linfobastica leukemia All High risk were individuals who were more likely to reicidiva in relation to other groups, addition to a lower survival rate.

Key-Words: Parametric Distribution; Kaplan-Meier; Linfoblastic Leukemia.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A análise de sobrevivência é um conjunto de ferramentas estatísticas que auxiliam na análise de dados em um intervalo de tempo pré-determinado, em que a variável resposta é o tempo até que um determinado evento ocorra nos indivíduos que constituem o banco de dados, estes eventos são denominados de falhas. É uma das áreas da estatísticas que mais cresceram nas últimas duas décadas devido ao desenvolvimento e aprimoramento de técnicas estatísticas combinadas com computadores avançados (COLOSIMO; GIOLO, 2006).

O transplante de medula óssea (TMO) é um tratamento padrão para a leucemia aguda. A recuperação após o TMO é um processo complexo. O prognóstico para a recuperação pode depender de fatores de risco conhecidos no momento do transplante, tais como paciente e/ou a idade do doador e o sexo, da etapa da doença inicial, o tempo a partir do diagnóstico até o transplante, etc. O prognóstico final pode mudar à medida que a história pós-transplante do paciente desenvolve com a ocorrência de eventos em momentos aleatórios durante o processo de recuperação, como o desenvolvimento da doença do enxerto versus hospedeiro aguda ou crônica (GVHD), retorno da contagem de plaquetas a níveis normais, retorno de granulócitos para níveis normais, entre outras. Transplante pode ser considerado uma falha quando um paciente de leucemia tem a morte causada pela doença e censura quando tem a volta da doença (recidiva).

Na realização deste trabalho estimou-se a curva de sobrevivência não paramétrica para 3 grupos de pacientes com diferentes tipos de leucemia linfoblástica, procedeu-se o ajuste das distribuições paramétricas aos dados de sobrevivência afim de compará-las entre si e finalmente aplicou-se os modelos de tempo de vida acelerado afim de se encontrar as covariáveis que alteram o tempo de sobrevivência para pacientes que receberam o trasplante de medula óssea.

2 Fundamentação Teórica

2.1 Princípios da análise de sobrevivência

A análise de sobrevivência é uma técnica estatística que cresceu nos últimos 40 anos devido a sua vasta área de aplicação, que pode ser usada na medicina, na área financeira e também ao desenvolvimento das tecnologias da informação. Análise de sobrevivência possui a característica de trabalhar com covariáveis relacionadas com o tempo de sobrevivência, permitindo ao pesquisador melhor compreender quais fatores que influenciam na curva de sobrevivência de um indivíduo.

Para melhor compreender a ação dessas covariáveis saber o tempo de vida é necessário utilizar metodologias especificas da estatística para estimar os parâmetros e funções a partir de métodos paramétricos, métodos semi-paramétricos e métodos não paramétricos, esses métodos também auxiliam na interpretação de dados censurados, ou seja, indivíduos que por diversos motivos o evento de interesse não ocorreu, ou a pesquisa começou com o evento já ocorrido. Strapasson (2007), afirma que 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 censura, 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, *i=1,...,n* são observações representadas por um par (t_i, δ_i).

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

Pode-se ainda ocorrer outros dois tipos de censuras à esquerda e a censura intervalar. Segundo Strapasson (2007), censura à esquerda ocorre quando o evento de interesse já aconteceu, quando o indivíduo foi observado: ou seja, o tempo de vida é menor que o observado.

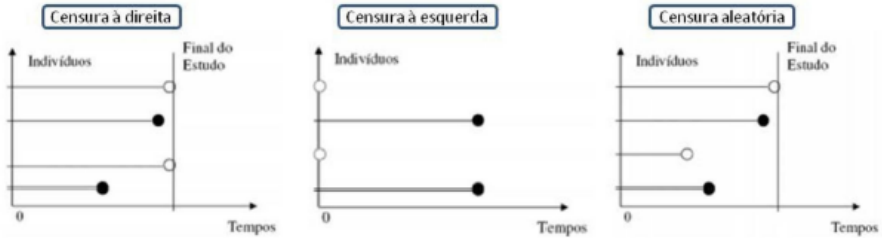

Figura 1: Representação gráfica de censura, em que • representa o tempo de falha e ∘ tempo de censura.

Censura intervalar é quando não se sabe o tempo exato da ocorrência do evento de interesse, sabe-se que ele ocorreu dentro de um intervalo especificado, por exemplo, ocorre quando não se conhece o exato momento da morte, mas sabe-se que ocorreu no intervalo de tempo. Segundo Colosimo e Giolo (2006), o mecanismo de censura aleatória é aquele em que os tempos de censura são variáveis aleatórias mutuamente independentes e ainda independentes dos tempos de vida. A censura do tipo I é um caso particular da censura aleatória, cuja variável aleatória t tem probabilidade maior do que zero, ou seja, t é uma variável aleatória mista com um componente contínuo e outro discreto. Dados censurados são representados por sinal "+".

2.2 Métodos não paramétricos

Na análise de sobrevivência, uma das utilizações dos métodos não-paramétricos é em situações em que os modelos paramétricos (probabilísticos) não estão se adequando aos dados, segundo (Colosimo e Giolo 2006) existem técnicas não-paramétricas para estimar parâmetros em análise de sobrevivência.

Desta forma é possível estimar funções de densidade de probabilidade aos tempos de vida, *f(t)*, como a função de sobrevivência, *S(t)*, e a função de risco , *h(t)*. A função de densidade de probabilidade, $\widehat{f}(t)$, pode ser estimada a partir dos dados amostrais, se não existirem observações censuradas. Caso existam outros meios, usar os dados censurados. A função de sobrevivência, $\widehat{S}(t)$, é estimada a partir dos dados, como a proporção de pacientes que sobreviveram após um certo período de tempo, t, e a função de risco, $\widehat{h}(t)$, é estimada a partir dos dados amostrais quando não existirem observações censuradas.

Os estimadores da probabilidade de sobrevida, $\hat{S}(t)$, utilizados nos teste não paramétricos se resumem a três, que são: o teste de Kaplan-Meier, 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 e o estimador de Nelson-Aalen, que apresenta propriedades similares ao de Kaplan-Meier.

2.2.1 Estimador de Kaplan-Meier

Esse método e usado principalmente para casos onde existem censuras nos dados em estudos, a função de Kaplan-Meier consegue, mesmo com censuras, estimar a função de risco e de sobrevivência. A construção do estimador de Kaplan-Meier considera o número de intervalos iguais ao número de falhas distintas e os limites dos intervalos são os próprios tempos de falhas na amostra e são ordenados do primeiro ao último, podendo existir mais de uma falha ao mesmo tempo, $t_1 \leq t_2 \leq t_3 \leq ... \leq t_k$, segundo Colosimo e Giolo (2006) é expresso por,

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

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

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

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

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

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

As principais propriedades do estimador são: ele é não viciado para amostras grandes, é fracamente consistente, converge assintoticamente para um processo gaussiano e é o estimador de máxima verossimilhança de *S(t).*

2.2.2 Tábua de vida ou atuarial

A Tábua de vida também conhecida na literatura como Atuarial é um instrumento dos mais antigos utilizados na análise de sobrevivência e com o passar dos anos foi substituída pelo estimador de Kaplan-Meier, mais nem por isso deixou de ser usado. Em certos casos, através do método Atuarial, é possível mostrar o tempo de sobrevivência dos elementos de amostra homogêneas em que geralmente usa-se uma amostra de no mínimo 30 elementos

para que se possa organizar os tempos de vida em intervalos, sendo bastante empregada em demografia e epidemiologia.

Segundo Ferreira (2007) para se construir uma tabela de vida primeiramente divide-se o período total de observação em um número conveniente de intervalos e para cada intervalo estima-se a função de sobrevivência. Como o estimador de Kaplan-Meier usa um número maior de intervalos em relação a tábua de vida para estimar as funções de vida e de risco, este último é considerado menos eficaz comparado com o Kaplan-Meier. Por isso é menos utilizado em análise de sobrevivência.

A tábua de vida é dada por Ferreiro (2007):

$$\hat{h}\left(t_{i-1}\right) = \frac{\mathrm{N}^0 \;\; falhas \;\; em\left[t_{i-1}, t_i\right)}{\left(\mathrm{N}^0 \;\; sob \;\; risco \;\; em \;\; t_{i-1}\right) - \left(\mathrm{N}^0 \;\; censuras \;\; em\left[t_{i-1}, t_i\right)\right) \diagup 2}$$

em que $i = 1..., n, t = t_1, ..., t_n$ e $t_0 = 0$ verifica-se na equação acima que observações censuradas no intervalo são tratadas como se estivessem sob risco durante a metade do intervalo considerado. Suponha um estudo iniciado com n indivíduos. Então a probabilidade de falhar até $\mathrm{t_i}$ é $\hat{h}$, ou seja, dos n indivíduos $n\left[\hat{h}\left(t_1\right)\right]$ não chegarão a t_1. Assim, no final do primeiro período $n\left[1-\hat{h}\left(t_1\right)\right]$ indivíduos ainda estarão vivos. Dessa maneira, a função de sobrevivência, que é a probabilidade de sobreviver além de t_1, pode então ser estimada por,

$$\hat{S}\left(t_1\right) = \frac{n\left[1-\hat{h}\left(t_1\right)\right]}{n} = 1-\hat{h}\left(t_1\right)$$

De forma análoga, dos $n\left[1-\hat{h}\left(t_1\right)\right]$ indivíduos que sobreviveram ao final do primeiro período apenas $n\left[1-\hat{h}\left(t_1\right)\right]\left[1-\hat{h}\left(t_2\right)\right]$ chegarão até o final do período. Portanto,

$$\hat{S}\left(t_2\right) = n\left[1-\hat{h}\left(t_1\right)\right]\left[1-\hat{h}\left(t_2\right)\right]$$

Assim, de uma forma geral, para qualquer tempo t o estimador atuarial da função de sobrevivência é definido por:

$$\hat{S}_{tv} = \left(t_i\right)\prod_{j=1}^{i}\left[1-\hat{h}\left(t_{j-1}\right)\right], j \leq i$$

A representação gráfica do estimativa da função de sobrevivência tem forma escada,

com valores constantes para a função em cada intervalo de tempo.

2.2.3 Nelson-Aalen

Esse método, em relação ao estimador de Kaplan-Meier e a Tabela de Vida é mais recente na literatura especializada. Foi proposta inicialmente Nelson (1972), como um estimador para função de risco acumulado $\Lambda(t)$; alguns anos depois Aalen (1978) provou suas propriedades usando processos de contagem. Por isso, esse estimador é conhecido como Nelson-Aalen. De acordo Colosimo e Giolo (2006) esse estimador tem a seguinte forma

$$\tilde{\Lambda}(t) = \sum_{j:t_j<t} \frac{d_j}{n_j}$$

em que, d_j e n_j são definidos com base na função de sobrevivência, tal como, os estimadores de Kaplan-Meier e o de Nelson-Aalen.Isto é,

$$S(t) = \exp\{-\Lambda(t)\}$$

e sua estimativa é dada por:

$$\tilde{S}(t) = \exp\left\{-\tilde{\Lambda}(t)\right\}$$

Bohoris (1994) citado por Colosimo e Giolo (2006), mostrou que $\tilde{S}(t) \geq \hat{S}(t)$ para todo t, ou seja, as estimativas obtidas por meio do estimador de Nelson-Aalen são maiores ou iguais às obtidas por meio do estimador de Kaplan-Meier.

2.2.4 Teste logrank

Muitas vezes é importante determinar se duas curvas de sobrevivência apresentam diferenças significativas entre si. Nestes casos, o teste logrank Mantel (1996), é um dos mais conhecidos e usados na área de sobrevivência.

A estatística do teste é a diferença entre o número observado de falhas em cada grupo e uma quantidade que, para muitos propósitos, pode ser pensada como o correspondente número esperado de falhas sob a hipótese nula. Considere inicialmente o teste de igualdade de duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. Sejam t_1, t_2, ..., t_k os tempos de falha distintos da amostra formada pela combinação das duas amostras individuais. Suponha que no tempo t_j ocorram d_j falhas e que n_j indivíduos estejam sob risco em um tempo

imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i = 1, 2 e j = 1, ..., k. Em cada tempo de falha t_j, os dados podem ser dispostos em forma de tabela de contingência 2 x 2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i.

Tabela 1: Tabela de contingência gerada no tempo t_j

	grupos		
	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Condicionado às ocorrências de falha e censura até o tempo anterior a t_j (fixando as marginais de coluna) e ao número de falhas no tempo t_j (fixando as marginais de linha), a distribuição de d_{2j} é uma hipergeométrica:

$$\frac{\begin{pmatrix} n_{1j} \\ d_{1j} \end{pmatrix} \cdot \begin{pmatrix} n_{2j} \\ d_{2j} \end{pmatrix}}{\begin{pmatrix} n_j \\ d_j \end{pmatrix}}.$$

A média de d_{2j} é $w_{2j} = n_{2j} \times d_j \times n_j - 1$, o que equivale a dizer que, se não houver diferença entre as duas populações no tempo t_j, o número total de falhas d_j pode ser dividido entre as duas amostras de acordo com a razão entre o número de indivíduos sob risco em cada amostra e o número total de indivíduos sob risco. A variância de d_{2j} obtida a partir da distribuição hipergeométrica é dada por

$$(V_j)_2 = d_j \left(\frac{n_{1j} \cdot n_{2j}}{n_j^2}\right) \left(\frac{n_j - d_j}{n_j - 1}\right).$$

Então, a estatística $d_{2j} - w_{2j}$ tem média zero e variância $(V_j)_2$. Se as k primeiras tabelas de contingência forem condicionalmente independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística

$$T = \frac{\left[\sum_{j=1}^{k} (d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} (V_j)_2},$$

em que sob a hipótese nula $H_0 : S_1(t) = S_2(t)$, para todo t no período de acompanhamento, tem uma distribuição qui-quadrado com 1 grau de liberdade para grandes amostras.

2.3 Métodos paramétricos

Os métodos paramétricos são ferramentas estatísticas que são trabalhadas através dos parâmetros estimados das amostras em estudos e geralmente possuem um número de suposições, que satisfeitas dão uma maior confiança na hora da interpretação dos resultados obtidos.

2.3.1 Modelos probabilísticos

Embora existam vários modelos probabilísticos, alguns ocupam maior destaque por sua comprovada adequação a várias situações reais, ou seja, por modelar os tempos de sobrevivência. Os principais modelos probabilísticos utilizados na análise de sobrevivência são o Exponencial, o Weibull e o Log-Normal, pois as variáveis tratam do tempo até a falha sendo positivos. Por outro lado, a Gaussiana (normal) e a binomial são adequadas para variáveis clínicas e industriais.

A distribuição Exponencial é das mais simples e importantes distribuições de probabilidade utilizadas para modelagem de dados que representam o tempo até a ocorrência do evento de interesse, apresentando a função de risco constante. A distribuição Weibull é a generalização da distribuição Exponencial, sendo bastante utilizada no ajuste de dados de confiabilidade em diversas áreas do conhecimento, apresenta função de risco crescente, decrescente ou ainda constante. A distribuição Log-Normal é usada para ajustar dados referentes à confiabilidade, como a distribuição Weibull, sendo que a Weibull e Log-Normal são caracterizados por dois parâmetros e a Exponencial por apenas um.

2.3.2 Distribuição Exponencial

A distribuição Exponencial possui bastante aplicação na modelagem de tempos de vidas (componentes eletrônicos, organismos vivos, etc.) e consequentemente na teoria de confiabilidade. Por volta do ano de 1940 alguns pesquisadores começaram a utilizar a exponencial para esses estudos, seja uma variável aleatória T, não-negativa, absolutamente contínua cujo o tempo de sobrevida $T \geq 0$ tem distribuição Exponencial com parâmetros λ se sua função de densidade de probabilidade é dada da seguinte forma:

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0 \tag{2.1}$$

onde λ é uma constante positiva, que é o parâmetro da distribuição também conhecida como taxa da exponencial. A sua função densidade de probabilidade acumulada, $F(t)$, é calculada a partir de sua função de densidade $f(t)$ por definição, temos que

$$F(t) = P(T \leq t) = \int_0^t f(u)du, \qquad t \geq 0 \tag{2.2}$$

agora substituindo pela (2.1) em (2.2)

$$\int_0^t \lambda e^{-\lambda u}du = \lambda \int_0^t e^{-\lambda u}du = \lambda \left[-\frac{e^{-\lambda u}}{\lambda}\right]_0^t = -\left[e^{-\lambda u}\right]_0^t = -\left[e^{-\lambda t} - e^{-\lambda 0}\right] = -e^{-\lambda t} + 1$$

organizando temos

$$F(t) = 1 - e^{-\lambda t}$$

Com a f.d.p, $f(t)$ e a função de probabilidade acumulada, $F(t)$, podemos calcular o valor da função de sobrevivência $S(t)$, $S(t)$ é definida como a probabilidade de um individuo sobreviver até um certo tempo t, sem que tenha ocorrido o evento. Sendo uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência, definidas por:

$$S(t) = P(T > t) = 1 - F(t) = \int_0^t f(u)du, \qquad t \geq 0 \tag{2.3}$$

Agora substituindo $F(t)$ calculado com base em (2.2) na equação (2.3) temos que

$$1 - F(t) = 1 - \left(1 - e^{-\lambda t}\right) = 1 - 1 + e^{-\lambda t} = e^{-\lambda t} = S(t)$$

.

A função de risco ou taxa de falha descreve a forma com que a taxa de falha muda com o tempo, ou seja, demonstra o risco do individuo falhar no tempo. A função de risco pode ser definida em termos da função de distribuição ,$F(t)$, e da função de densidade de probabilidade, $f(t)$. Porém, podemos substituir $1 - F(t)$ por $S(t)$ da seguinte forma:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}, \qquad t \geq 0 \tag{2.4}$$

substituindo $f(t)$ e $S(t)$ pelos resultados obtidos em (2.1) e (2.3) obtemos

$$h\left(t\right) = \frac{f\left(t\right)}{S\left(t\right)} = \frac{\lambda \cdot e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

2.3.3 Distribuição Weibull

A distribuição Weibull foi proposta originalmente por Wallodi Weibull (WEIBULL, 1939) e discutida por Weibull (1951), por este mesmo autor. Esta distribuição é usada frequentemente para estudos biomédicos e industriais, pois apresenta uma grande variedade de formas devido à sua simplicidade, todas com propriedades básicas: função de taxa de falha é monótona, isto é, crescente, decrescente ou constante. O autor ressalta que essa distribuição é tão importante para análise paramétrica de dados de sobrevivência quanto a distribuição normal é para modelos lineares.

De acordo com Colosimo e Giolo (2006), a sua f.d.p é da forma,

$$f(t) = \frac{\gamma}{\alpha\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \qquad t \geq 0 \tag{2.5}$$

em que $\gamma > 0$ e $\alpha > 0$ são parâmetros de forma e escala. Pode-se observar, que para $\gamma < 1$, tem-se função de taxa de falha decrescente, enquanto $\gamma > 1$ as funções de taxa de falha são crescente, e $\gamma = 1$ a função de taxa de falha é constante. O parâmetro α tem mesma unidade de medida de t, γ não tem unidade.

As funções de risco e de sobrevivência são, respectivamente,

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{2.6}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha\gamma} t^{\gamma-1} \tag{2.7}$$

para $t \geq 0$, α e $\gamma > 0$.

Quando $\gamma = 1$, obtêm-se a distribuição exponencial como caso particular da distribuição Weibull, sendo algumas formas das funções de densidade de sobrevivência e de taxa de falha (risco) da variável T. A partir da função densidade de probabilidade da Weibull, $f(t)$, podemos calcular a sua função de densidade acumulada $F(t)$, para facilitar os cálculos futuros reescrevendo (2.5) temos o $f(t)$ da seguinte maneira

$$f\left(t\right) = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \qquad t \geq 0 \tag{2.8}$$

usando (2.8) em (2.3) para calcular $F(t)$ temos que

$$F\left(t\right)=\int_{0}^{t}\frac{\gamma}{\alpha}\left(\frac{x}{\alpha}\right)^{\gamma-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\gamma}\right\}dx$$

fazendo uma mudança de variável

$$h=\left(\tfrac{x}{\alpha}\right)^{\gamma}\Rightarrow dh=\tfrac{\gamma}{\alpha}\left(\tfrac{x}{\alpha}\right)^{\gamma-1}$$
$$x=0\Rightarrow h=0;x=t\Rightarrow h=\left(\tfrac{t}{\alpha}\right)^{\gamma}$$

temos que

$$F\left(t\right)=\int_{0}^{t}\frac{\gamma}{\alpha}\left(\frac{x}{\alpha}\right)^{\gamma-1}\exp^{-\left(\frac{x}{\alpha}\right)^{\gamma}}=\int_{0}^{\left(\frac{t}{\alpha}\right)^{\gamma}}\exp^{-h}dh=-\exp^{-h}|_{0}^{\left(\frac{t}{\alpha}\right)^{\gamma}}=1-\exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right]$$

calculando a função de sobrevivência $S(t)$ a partir de $1-F(t)$ obtemos

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

e já função de risco $h(t)$ com base em

$$h\left(t\right)=\frac{f\left(t\right)}{1-F\left(t\right)}=\frac{f\left(t\right)}{S\left(t\right)}$$

encontramos o seguinte

$$=\frac{\frac{\gamma}{\alpha^{\gamma}}t^{\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}{\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}=\frac{\gamma}{\alpha^{\gamma}}t^{\gamma-1}$$

O modelo Weibull é um caso particular do modelo de Cox para dados de sobrevivência intervalar. Segundo Strapasson (2007) pode-se ainda mostrar que se $T\sim$ Weibull (α,γ), então, Y= log (T) $\sim$ Gumbel, ou seja, Y tem distribuição do valor extremo, com função de densidade de probabilidade, função de sobrevivência e função da taxa dado respectivamente, por

$$f(y)=\frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right)-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\},\tag{2.9}$$

$$S(y)=\exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}\tag{2.10}$$

e

$$\lambda(y)=\frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}\tag{2.11}$$

em que Y e $\mu \in \mathbb{R}$ e $\sigma > 0$. Se $\mu = 0$ e $\sigma = 1$, tem-se a distribuição do valor extremo padrão. Os parâmetros μ e σ são denominados parâmetros de locação e escala, respectivamente, e relacionam-se com os parâmetros da distribuição de Weibull e do valor extremo apresentando as seguintes relações de igualdade, $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp\{\mu\}$.

A constante de Euler é conhecido pela média e variância, ou seja, $\mu - \upsilon\sigma$ e $(\frac{\pi^2}{6})\sigma^2$ com υ= 0,5772... . O percentil $100_p\%$, t_p, expresso por (COLOSIMO; GIOLO, 2006) é,

$$t_p = \mu + \sigma \log[-\log(1-p)] \tag{2.12}$$

Na análise de dados de sobrevivência, é comum trabalhar com o logaritmo dos tempo de vida dos indivíduos.

2.3.4 Distribuição Log-Normal

Essa distribuição é bastante usada para se estudar tempos de vida de produtos e indivíduos, ela é muito utilizada em estudos de tempos clínicos, como o de leucemia entre outras doenças. A sua f.d.p para uma variável aleatória T , Colosimo e Giolo (2006) e dada por:

$$f(t) = \frac{1}{\sqrt{2\pi t\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)\right\}, t > 0, \tag{2.13}$$

em que μ é a média do logaritmo do tempo de falha, assim como σ é o desvio-padrão. Uma observação importante sobre a log-normal é que sua função de sobrevivência $S(t)$ e a função de risco, $\lambda(t)$, não contém na literatura uma forma analítica explicita podendo então ser apresenta como:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right)$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}$$

2.3.5 Distribuição Gama e Gama Generalizada

A Distribuição Gama é bastante utilizada para descrever dados relacionados com tempo, porém isso é graças aos pioneiros Brown e Flood (1947) que exploraram essa

distribuição para explicar o tempo de vida de copos de vidro em uma cafeteria. Na área medica, o uso dessa distribuição é recente, principalmente quando existem fatores aleatórios. A função de densidade da distribuição Gama com π o parâmetro de forma e δ o de escala, para $\pi, \delta > 0$ é expresso, por

$$f(t) = \frac{1}{\Gamma(\pi)\,\delta^{\pi}} t^{\pi-1} \exp\left\{-\left(\frac{t}{\delta}\right)\right\}, t > 0$$

e sua função de sobrevivência $S(t)$ expressada da seguinte maneira:

$$S(t) = \int_{t}^{\infty} \frac{1}{\Gamma(\pi)\,\delta^{\pi}} v^{\pi-1} \exp\left\{-\left(\frac{v}{\delta}\right)\right\} dv$$

e acrescentando mais um parâmetro γ de forma, essa distribuição se torna mais flexível para mais informações ver Stacy (1962).

2.4 Estimação dos parâmetros

Segundo Colosimo e Giolo (2006), os parâmetros são características dos modelos de probabilidade para estudos de tempo de vida, existindo-se alguns métodos de estimação. O método de máxima verossimilhança é uma opção apropriada para dados censurados, incorporando-se as censuras relativamente simples por possuir propriedades para grandes amostras.

2.4.1 Máxima Verossimilhança

O método de máxima verossimilhança apresenta os procedimentos de estimação para os parâmetros dos modelos de sobrevivência. Por exemplo se a distribuição do tempo de falha é a da Weibull, para cada combinação diferente γ e α, tendo diferentes distribuições Weibull. O estimador de máxima verossimilhança escolhe o par de γ e α que melhor explique a amostra observada.

Considera-se uma amostra de observações aleatórias $t_1, ..., t_n$ de uma variável aleatória T com tempos de sobrevivência e de confiabilidade de uma certa população de interesse com n observações independentes de t_i, em que t_i, $i = 1, \ldots, n$, indica o tempo de falha ou censura, onde todas são não-censuradas. Com um vetor de parâmetros $\boldsymbol{\theta}=(\alpha, \beta, \gamma)$, tem-se a função de verossimilhança para um parâmetro genérico θ da população (COLOSIMO e GIOLO 2006).

O método de máxima verossimilhança é baseado geralmente para modelo em inferência

paramétrica e sua teoria assintótica, onde a função de verossimilhança para o vetor de parâmetros $\boldsymbol{\theta}$ é expressa por,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}) \tag{2.14}$$

sendo que θ pode ser um único parâmetro ou um vetor de parâmetros. A função de verossimilhança $L(\boldsymbol{\theta})$ mostra que a contribuição de cada observação não-censurada, sendo esta a função de densidade. As observações censuradas não representam a função de densidade, estando ligadas a função de sobrevivência $S(t)$. As observações são dividas em dois conjuntos, em que as r primeiras ordenadas são as não-censuradas $(1, 2, ...r)$ e as $(n - r)$ seguintes são as censuradas $(r + 1, r + 2, ..., n)$. Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$, onde temos a dependência de f em $\boldsymbol{\theta}$, em que L é função de $\boldsymbol{\theta}$. Com a presença de censura, tem-se,

$$L(\theta) = \prod_{\mathbf{i=1}}^{\mathbf{r}} \mathbf{f}(\mathbf{t_i}; \theta) \prod_{\mathbf{i=r+1}}^{\mathbf{n}} \mathbf{S}(\mathbf{t_i}; \theta), \tag{2.15}$$

em que o termo relacionado a censura tem a forma $\prod_{i=r+1}^{n} S(c, \theta) = [\mathbf{S}(\mathbf{x}, \theta)^{\mathbf{n-r}}]$, sendo $T = C$

Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$ ou equivalentemente o logaritmo de $L(\theta)$. Eles são encontrados resolvendo o sistema de equações,

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} = 0 \tag{2.16}$$

Segundo Strapasson (2007), as propriedades assintóticas dos estimadores de máxima verossimilhança são necessárias para construção de intervalos de confiança e testes de hipóteses sobre os parâmetros do modelo sob condições de regularidade com média $\boldsymbol{\theta}$, matriz de variância e covariância dada pelo inverso da matriz de Fisher $(I(\theta)^{-\mathbf{1}})$.

2.5 Teste de Hipóteses

Em muitas situações é preciso trabalhar com vetores de parâmetros $\Theta = (\Theta_1, \cdots, \Theta_p)$ com isso surge a necessidade de testa hipóteses relacionadas com esse vetor ou subcon-

juntos desse vetor, para realizar essas tarefa existe alguns testes como o Wald e Razão de Verossimilhanças entre outros existentes.

2.5.1 Teste de Wald

Esse teste é uma generalização do teste t de Student (Wald, 1943), é baseado na distribuição assintótica de $\hat{\Theta}$ e geralmente usado para o caso de hipóteses relacionadas a apenas um parâmetro Θ_i suas hipótese são as seguintes:

$$\begin{cases} H_0 : \Theta = \Theta_0 \\ H_1 : \Theta \neq \Theta_0 \end{cases}$$

sua estatística do teste e dado por:

$$W_{calc} = \left(\hat{\Theta} - \Theta_0\right)' F\left(\Theta_0\right) \left(\hat{\Theta} - \Theta_0\right),$$

em H_0 tem aproximadamente uma distribuição qui-quadrado com α graus de liberdade χ^2_α. Sua interpretação é dada da seguinte maneira para todos os níveis de significância em que os valores de W_{calc} maiores que os valores tabelados de uma qui-quadrado indicam evidencias para rejeição de H_0.

2.5.2 Teste da Razão de Verossimilhanças

O teste da razão de verossimilhanças (TVR) e usado para a comparação dos valores logaritmo da função de verossimilhança maximizada e não possui restrição e sob H_0, assim sendo a comparação $\log L(\hat{\theta})$ e $\log L(\hat{\theta}_0)$. As hipóteses a serem testadas são as seguintes:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

e sua estatística de teste e dada por:

$$TRV = -2 \log \left[\frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \right] = 2[\log L(\hat{\theta}) - \log L(\hat{\theta}_0)], \tag{2.17}$$

E sob H_0 temos que segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade. Para amostras relativamente grandes, rejeita-se H_0 a um nível de $100\alpha\%$ de significância, se $TRV > \chi^2_{p,1-\alpha}$.

2.6 Métodos Gráficos

Os métodos gráficos são usados como evidencias para comparação da função de sobrevivência proposta e estimada por Kaplan-Meier. Inicialmente ajustam-se os modelos estatísticos que propõem explicar o comportamento dos dados, e após a estimação dos parâmetros dos respectivos modelos são estimadas as suas respectivas funções de sobrevivência como por exemplo para os modelos Exponencial, Weibull e log-normal, etc. Para o Conjunto de dados também é estimada a função de sobrevivência $\hat{S}$ pelo estimador de Kaplan-Meier, por fim é comparado a função de sobrevivência estimada pelo Kaplan-Meier contra a função de sobrevivência estimada pelos modelos propostos $\hat{S}$ versus $\hat{S}_{\exp}(t)$ etc. A sua interpretação é dada da seguinte maneira, a curva do modelo que mais se assemelha a curva estimada pelo Kaplan-Meier tem a maior evidência de explicar o conjunto de dados. Uma outra forma de comparação é usando a função de taxa de falha acumulada $\hat{\Lambda}$ proposta por Nelson(1990a), ou seja, colocando em um gráfico a função taxa de falha $\hat{\Lambda}$ versus t, $\hat{\Lambda}$ é estimada pelos modelos probabilísticos, e a relação da taxa de falha acumulada Λ com a função de sobrevivência é dada da seguinte forma:

$$\Lambda(t) = -\log(S(t))$$

.

Uma observação importante é que apenas o uso dos métodos gráficos não assegura evidências suficientes para a escolha de um modelo sendo preciso reunir o maior número de informações possíveis antes de tomar qualquer decisão, o uso dos testes estatísticos ainda é a melhor forma de reunir evidencias a favor de um modelo ou técnica.

2.7 Modelos de regressão para dados de sobrevivência

Na área médica é comum estudos que envolvem a inclusão de covariáveis, e quando essas covariáveis estão relacionadas com o tempo de sobrevivência é preciso usar os modelos de regressão, para poder acomodar de forma correta os efeitos da inclusão das covariáveis, na análise estatística dos dados, os modelos que melhor explicam os efeitos e relações das covariáveis nos dados de sobrevivência, estão divididos em dois grupos, os paramétricos e semi-paramétricos, o primeiro grupo também é denominado como modelos de tempo de vida acelerado, que são mais eficazes porém menos flexíveis que o segundo grupo.

2.7.1 Modelo de Regressão Linear

Quando desejamos compreender a relação entre uma única covariável e a resposta, onde essa resposta é o tempo de ocorrência de um determinado evento de interesse, usamos um modelo estatístico para explicar essa relação, o mais conhecido é o modelo de regressão linear Draper e Smith (1998) nesse modelo de regressão linear a resposta é associada com a covariável por meio de um modelo linear. Esse modelo é representado da seguinte maneira

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{2.18}$$

onde a equação da reta é o componente determinístico do modelo de regressão linear e a variação em torno dessa reta é o componente estocástico do modelo, o componente estocástico geralmente tem uma distribuição normal. A descrição do modelo é a seguinte Y é a resposta, x é a covariável, e os parâmetros são β_0 e β_1 para serem estimados e ϵ é o erro aleatório com distribuição normal. Lembrando que esse modelo de regressão linear é usado para dados sem censura. Entretanto o que mais ocorre em estudos de sobrevivência é a presença de censuras nos dados e isso impede o uso do modelo (2.18). Além disso a distribuição da resposta geralmente tende a ser assimétrica na direção dos maiores tempos de sobrevivência, isso torna incorreto o uso da distribuição normal para o componente estocástico do modelo de regressão. Para contornar esse problema da modelagem estatística em análise de sobrevivência existem duas maneiras, a primeira seria fazer uma transformação nos dados para voltar ao modelo linear normal, a segunda seria usar um componente determinístico não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico os dois métodos se equivalem em determinadas situações.

2.7.2 Modelos de Regressão Exponencial

Para acrescentar covariáveis, utilizando o modelo de regressão exponencial, consideramos o seguinte:

$$V_e = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i \tag{2.19}$$

onde o $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$, ε_j são independentes identicamente distribuídos (iid) variáveis

aleatórias com uma distribuição exponencial dupla valor ou extrema, e esse modelo de regressão exponencial tem a seguinte função de risco, densidade e sobrevivência:

$$\begin{aligned} h(V_e) &= \lambda_i = \exp\left[-\left(\beta_0 + \textstyle\sum_{j=1}^{p} \beta_j x_{ji}\right)\right] \\ f(V_e) &= \lambda_i \exp\left(-\lambda_i t\right) \\ S(V_e) &= \exp\left(-\lambda_i t\right) \end{aligned}$$

Para o modelo (2.19), a estimação dos seus parâmetros e dada pelo método de máxima verossimilhança onde sua função de verossimilhança é a seguinte:

$$L\left(\lambda_{ij}\right) = \sum\nolimits_{i=1}^{w} \prod\nolimits_{j=1}^{n} \left(\lambda_{ij}\right)^{\delta_{ij}} \exp\left(-\lambda_{ij} v_{ij}\right) \tag{2.20}$$

e para a obtenção dos estimadores de máxima verossimilhança, é necessário substituir as funções de densidade e sobrevivência pelos de valores extremos na função de verossimilhança $L\left(\lambda_{ij}\right)$ e depois aplica-se o log na função de verossimilhança encontramos o seguinte resultado $\log L\left(\lambda_{ij}\right) = \sum_{j=1}^{n} \delta_{ij} x_{lij}$, a função de verossimilhança (2.20) e a forma geral e serve para todos os modelos de regressão paramétricos, para informações mais detalhadas sobre estimadores de máxima verossimilhança para, essa parametrização do modelo de regressão exponencial que foi baseada com algumas modificações no modelo de Lee (2003).

2.7.3 Modelos de Regressão Weibull

O modelo de regressão exponencial apresenta alguns problemas para dados de sobrevivência por ser muito simples na sua forma de analisar os dados com covariáveis e censuras, para isso existem outros modelos como o modelo de regressão Weibull que permite acrescentar várias covariáveis, que dependendo do contexto em estudo pode ser igual ou melhor que o modelo de regressão exponencial. Considere o modelo de regressão Weibull da seguinte maneira

$$V_w = \beta_0 + \sum\nolimits_{k=1}^{p} \beta_k x_{ki} + \sigma \varepsilon_i \tag{2.21}$$

para $\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}$ e ε onde V_w tem uma distribuição Weibull, pois $log(V_w)$ tenha uma distribuição de valor extremo com parâmetro σ de escala. Para esse modelo de regressão a função de risco, f.d.p e função de sobrevivência e dada da seguinte forma:

$$h(V_w) = \lambda_i \gamma v^{\gamma-1}$$
$$f(V_w) = \lambda_i \gamma v^{\gamma-1} \exp\left(-\lambda_i v^{\gamma}\right)$$
$$S(V_w) = \exp\left(-\lambda_i v^{\gamma}\right)$$

E a função de verossimilhança para estimação dos seus parâmetros e dada como:

$$L\left(\lambda_{ij}\right) = \sum\nolimits_{i=1}^{w} \prod\nolimits_{j=1}^{n} \left(\lambda_{ij}\right)^{\delta_{ij}} \exp\left(-\lambda_{ij} v_{ij}\right),$$

lembrando que e a função (2.20) e a função de verossimilhança na sua forma geral, onde substituímos os valores das funções de densidade e sobrevivência por aquelas da distribuição de valores extremos e depois tomando o logaritmo de $L\left(\lambda_{ij}\right)$, encontramos os estimadores de máxima verossimilhança dos parâmetros.

2.7.4 Modelos de Regressão lognormal

No modelo de regressão lognormal consideramos o modelo a seguir com V_n como sendo uma variável aleatória com distribuição lognormal.

$$V_n = \beta_0 + \sum\nolimits_{k=1}^{p} \beta_1 x_{ki} + \sigma\varepsilon_i \tag{2.22}$$

e sua f.d.p e função de sobrevivência dada por

$$f\left(V_n\right) = \frac{\exp\left[\frac{-\left(\log v - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\right)\right)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma v} \tag{2.23}$$

$$S\left(V_n\right) = 1 - \left(\left(\sqrt{2\pi}\right)^{-1} \int_{-\infty}^{\varepsilon} e^{-x^2/2} dx\right)\left(\frac{\log v - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\right)}{\sigma}\right) \tag{2.24}$$

Como podemos observar analiticamente não é simples de se resolver, e para isso podemos ter o auxilio de um software, para estimar tais funções e com a função de verossimilhança (2.20) não é diferente. É possível obter vários modelo de regressão a partir de generalizações do modelo (2.21) podemos considerar as distribuições gama e log-logística entre outras para a variável aleatória V_w e para o ε a mais comum é a generalização proposta por Cox (1972).

2.8 Adequação do Modelo de Regressão Paramétrico Ajustado

A escolha adequada do modelo de regressão é uma das partes mais importantes da análise dos dados, pois o modelo de regressão que mais se ajusta aos dados é que terá a interpretação mais próxima da realidade dos dados. Para essa escolha do modelo existe algumas técnicas que ajudam a fazer a escolha mais coerente, para modelos de regressão podemos recorrer a análise dos resíduos; essas técnicas servem para mostrar quais os modelos que estão claramente inapropriados para os dados e não para provar que tal modelo é o correto, pois em muitas situações um ou mais modelos podem demostrar bons ajustes. A análise dos resíduos pode demostrar o ajuste global do modelo de regressão, a forma funcional de uma covariável nos dados e a acurácia do modelo entre outras funções, podemos destacar na literaturas os seguintes resíduos:

- Resíduos de Cox-Snell
- Resíduos Martingal
- Critério de Akaike (AIC)

Existe outras técnicas para a escolha dos modelos, porém nesse estudo vamos focar em alguns resíduos citados acima e o critério de Akaike.

2.8.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell foram propostos por Cox é Snell (1968) e melhorado por Klein é Moeschberger (1997). Eles servem para verificar o ajuste global, considerando o modelo de regressão (2.21) os resíduos de Cox-Snell é dado por:

$$\hat{e}_{cs_i} = \frac{\hat{V}_i - \left(\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{ki}\right)}{\hat{\sigma}}, \qquad i = 1, 2, ..., n \tag{2.25}$$

Os resíduos $\hat{e}_{cs}$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão se o modelo for adequado (Lawless, 1982). Se o modelo for bem adequado aos dados o gráfico $\hat{e}_{cs}$ contra a função de risco acumulada deve ser aproximadamente uma reta com inclinação igual ao valor 1, e para todo i tem um valor exp(1) isto com taxa de distribuição igual a 1.

Um resíduo similar ao Cox-Snell é o resíduo padronizado que são representados baseados no modelo de regressão lognormal para mais informação sobre esse resíduos ver Colossimo e Giolo (2006).

2.8.2 Resíduos Martingal

Esse resíduo é utilizado para verificar o comportamento linear e quadrático entre outras formas de uma certa covariável geralmente contínua acrescentada aos modelos de regressão, ele é definido como

$$\hat{R}_i = \hat{\delta}_i - \hat{e}_{cj_i} \tag{2.26}$$

onde $\hat{\delta}_i$ é uma variável que mostra as falhas e $\hat{e}_{cj_i}$ e uma forma modificada dos resíduos de Cox-Snell, pois nessa situação são interpretados como uma estimativa das falhas em excesso observadas nos dados e não é predito pelo modelo.

2.8.3 Criterio de seleção de Akaike (AIC)

O critério de seleção proposto por Akaike (1974) se baseia pelo método de máxima verossimilhança, e o AIC e definido da seguinte maneira $AIC = -2$ (log EMV) $+2$ (números de parâmetro). Sua interpretação é simples quanto menor for valor do AIC melhor o ajuste do modelo, uma observação importante e que o critério de Akaike não e um teste de hipóteses.

3 Aplicação

3.1 Descrição dos Dados

O banco de dados é constituído de um total de 137 pacientes (99 AML, 38 ALL) os quais são uma distinção entre os tipos de câncer, leucemia linfoblástica aguda e leucemia linfoblástica alto e baixo risco, foram tratados em um de quatro hospitais avaliados: 76 pacientes no Hospital da Universidade Estadual de Ohio (OSU), em Columbus, 21 pacientes de Hahnemann University (HU), na Filadélfia, 23 pacientes no Hospital St. Vincent (SVH) em Sydney, Austrália, e 17 pacientes no Alfred Hospital (AH) em Melbourne.

O estudo consiste de transplantes realizados nessas instituições a partir de 1 de março de 1984, a 30 de Junho de 1989. O máximo de acompanhamento foi de 7 anos. Foram 42 pacientes que recaíram e 41 que morreram enquanto em remissão. Vinte e seis pacientes tiveram um episódio de GVHD aguda (doença do hospedeiro), e 17 pacientes tiveram recidiva ou morreram em remissão, sem suas plaquetas voltarem para níveis normais.

Vários fatores potenciais de risco foram medidos no momento do transplante. Para cada doença, os pacientes foram agrupados em categorias de risco com base no seu estado no momento de transplante. Essas categorias foram os seguintes: ALL (38 pacientes), AML baixo risco primeira remissão (54 pacientes), AML e segunda remissão de alto risco ou não tratada primeira recaída (15 pacientes) ou segunda ou maior de recaída ou nunca em remissão (30 pacientes). Foi estudado o tempo até a morte do pacientes com leucemia linfoblástica. As covariáveis utilizadas foram 10 ao todo e identificadas como:

Z_1: Idade do paciente; Z_2: Idade do doador;

Z_3: Sexo do paciente; Z_4: Sexo do doador;

Z_5: CMV do paciente; Z_6: CMV do doador;

Z_7: Tempo de espera em dias para o transplante de medula óssea; Z_8: FAB;

Z_9: Hospital; Z_{10}: MTX.

3.2 Analise dos dados

A tabela 2 descreve algumas estatísticas básicas calculadas a partir da função de sobrevivência estimada, através do estimador de Kaplan-Meier:

Tabela 2: Tabela com estatísticas básicas, para os dados de leucemia linfoblástica (ALL), grupo 1, Leucemia linfoblástica aguda (AML) baixo e alto risco respectivamente grupo 2 e 3.

Grupos	Nmax	eventos	mediana	Limite inferior	Limite superior
grupos 1	38	23	466	194	NA
grupos 2	54	21	NA	1074	NA
grupos 3	45	33	242	120	456

Pela Tabela 2, percebe-se que o grupo com maior número de indivíduos é o grupo 2 que é constituído de indivíduos com leucemia linfoblástica aguda (AML) de baixo risco, de modo que 21 indivíduos tiveram a recidiva da doença, outro fator importante é que no caso mais grave da doença a porcentagem de recidiva da doença é de 73%, enquanto que nos demais casos a porcentagem não ultrapassa 60%. O gráfico das curvas estimadas sobrevida livre de doença para os três grupos são apresentadas na Figura 2.

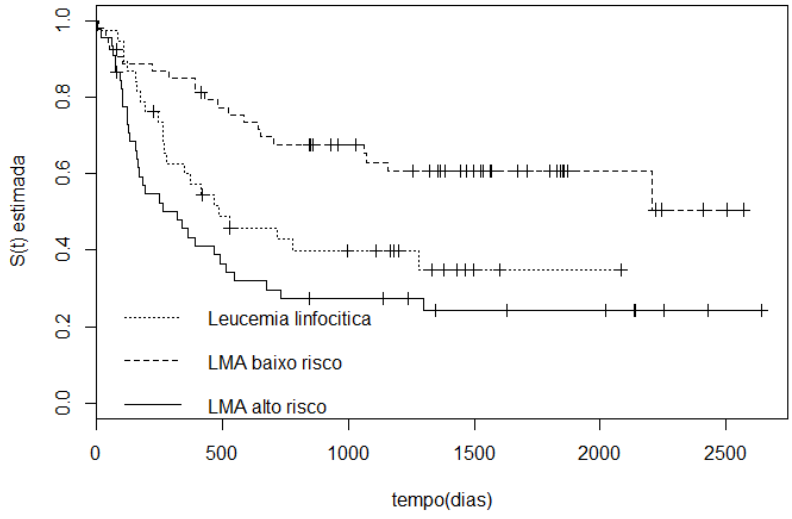

Figura 2: Representação gráfica das curvas de sobrevivência estimadas, a partir do estimador de Kaplan-Meier, para pacientes com leucemia linfocítica, e AML (baixo risco e alto risco de desenvolver esse tipo de leucemia) e as censuras representadas por (+)

Após essa análise previa da Tabela 2, foi usado o software R para estimar as curvas de sobrevivência, e consequentemente observamos o comportamento das 3 curvas de

sobrevivência, para isso foi usado o estimador de Kaplan-Meier.

Na figura 2, observa-se que as curvas terminam em pontos diferentes, porque os maiores tempos de estudo são diferentes para os três grupos (2081 dias para ALL , 2.569 para AML de baixo risco, e 2640 para AML alto risco). No AML baixo risco os pacientes têm o melhor prognóstico e pacientes de alto risco AML tem o prognóstico menos favorável.

A seguir uma comparação entre duas formas distintas de estimação de tempos de sobrevivência e de riscos, fazendo primeiramente as estimativas usando estimador de Kaplan-Meier e logo após o estimador de Nelson-Aalen:

Tabela 3: Estimativas da sobrevivência para o grupo 3 AML Alto risco, usando o estimador de Kaplan-Meier.

Tempo	n° Risco	n° Evento	Sobrevivência	E.P.	L.I.	L.S.
2	45	1	0.978	0.0220	0.936	1.000
16	44	1	0.956	0.0307	0.897	1.000
62	43	1	0.933	0.0372	0.863	1.000
63	42	1	0.911	0.0424	0.832	0.998
73	41	1	0.889	0.0468	0.802	0.986
74	40	1	0.867	0.0507	0.773	0.972
93	38	1	0.844	0.0542	0.744	0.957
97	37	1	0.821	0.0574	0.716	0.942
105	36	2	0.775	0.0626	0.662	0.908
121	34	1	0.753	0.0648	0.636	0.891
122	33	1	0.730	0.0667	0.610	0.873
128	32	1	0.707	0.0684	0.585	0.855
129	31	1	0.684	0.0699	0.560	0.836
153	30	1	0.661	0.0712	0.536	0.817
162	29	1	0.639	0.0723	0.512	0.797
164	28	1	0.616	0.0732	0.488	0.777
168	27	1	0.593	0.0740	0.464	0.757
183	26	1	0.570	0.0746	0.441	0.737
195	25	1	0.547	0.0750	0.418	0.716
248	24	1	0.525	0.0753	0.396	0.695
265	23	1	0.502	0.0754	0.374	0.673
318	22	1	0.479	0.0753	0.352	0.652
341	21	1	0.456	0.0751	0.330	0.630
363	20	1	0.433	0.0747	0.309	0.608
392	19	1	0.411	0.0742	0.288	0.585
469	18	1	0.388	0.0735	0.267	0.562
491	17	1	0.365	0.0726	0.247	0.539
515	16	1	0.342	0.0716	0.227	0.516
547	15	1	0.319	0.0703	0.207	0.492
677	14	1	0.296	0.0689	0.188	0.468
732	13	1	0.274	0.0673	0.169	0.443
1298	9	1	0.243	0.0663	0.143	0.415

L.S. Limite Superior; L.I. Limite Inferior; E.P. Erro Padrão

Tabela 4: Estimativas da sobrevivência para o grupo 3 AML Alto risco, usando o estimador de Nelson-Aalen.

Tempo	n° Risco	n° Eventos	Sobrevivência	E.P.	L.I.	L.S.
2	45	1	0.978	0.0217	0.936	1.000
16	44	1	0.956	0.0304	0.898	1.000
62	43	1	0.934	0.0368	0.865	1.000
63	42	1	0.912	0.0420	0.833	0.998
73	41	1	0.890	0.0464	0.804	0.986
74	40	1	0.868	0.0502	0.775	0.972
93	38	1	0.846	0.0537	0.747	0.958
97	37	1	0.823	0.0568	0.719	0.942
105	36	2	0.779	0.0618	0.666	0.910
121	34	1	0.756	0.0640	0.640	0.892
122	33	1	0.733	0.0660	0.615	0.875
128	32	1	0.711	0.0677	0.590	0.857
129	31	1	0.688	0.0692	0.565	0.838
153	30	1	0.666	0.0705	0.541	0.819
162	29	1	0.643	0.0716	0.517	0.800
164	28	1	0.621	0.0726	0.493	0.781
168	27	1	0.598	0.0734	0.470	0.761
183	26	1	0.575	0.0740	0.447	0.740
195	25	1	0.553	0.0745	0.425	0.720
248	24	1	0.530	0.0748	0.402	0.699
265	23	1	0.508	0.0749	0.380	0.678
318	22	1	0.485	0.0749	0.359	0.657
341	21	1	0.463	0.0747	0.337	0.635
363	20	1	0.440	0.0744	0.316	0.613
392	19	1	0.418	0.0739	0.295	0.591
469	18	1	0.395	0.0733	0.275	0.568
491	17	1	0.372	0.0725	0.254	0.545
515	16	1	0.350	0.0715	0.234	0.522
547	15	1	0.327	0.0704	0.215	0.499
677	14	1	0.305	0.0691	0.195	0.475
732	13	1	0.282	0.0675	0.177	0.451
1298	9	1	0.252	0.0666	0.151	0.423

L.S. Limite Superior; L.I. Limite Inferior; E.P. Erro Padrão

Ainda na Tabela 3 pelo método de Kaplan-Meier, percebe-se que após 2 anos a probabilidade dos indivíduos sobreviverem a doença está em torno de $25,2\%$ e na tabela 4 temos as mesma estimativas, tendo como método de Nelson-Aelen podemos observar nos limites inferiores e superiores uma precisão menor quando utilizamos o estimador de Nelson-Aalen. Mais detalhes sobre esta diferença de estimadores em Bohoris (1994).

Tabela 5: Tabela com valores do teste *logrank*, para comparação das curvas de sobrevivência estimadas para os dados de leucemia linfoblástica, grupo 1 (ALL), Leucemia linfoblástica (AML) grupo 2 baixo risco e 3 alto risco, usando o estimador de Kaplan-Meier.

Grupos	N	Observado	Esperado	$(O-E)^2/E$	$(O-E)^2/V$
grupos=1	38	23	20,3	0,381	0,522
grupos=2	54	21	36,9	6,871	13,483
grupos=3	45	33	19,8	8,717	11,871

Para a Tabela 5, o valor do teste de logrank para a comparação entre os três grupos de pacientes em estudos resultou em uma $\chi^2 = 16,3$ com 2 graus de liberdade e um p-valor igual a $0,000294$, ou seja, indicando que existe uma diferença entre as três curvas de sobrevivência. É interessante fazer comparações uma a uma, isto é, comparar a curva 1 com a 2, curva 1 com 3 e 2 com 3 para que ser possa ter uma melhor interpretação de quais curvas (grupos) diferem entre sí.

Tabela 6: Tabela com valores do teste *logrank*, para comparação das curvas de sobrevivência estimadas para os dados de leucemia linfoblástica, grupo 1 (All), Leucemia linfoblástica (AML) grupo 2 baixo risco, usando o estimador de Kaplan-Meier.

Grupos	N	observado	Esperado	$(O-E)^2/E$	$(O-E)^2/V$
grupos[1:92]=1	38	23	15,1	4,11	6,43
grupos[1:92]=2	54	21	28,9	2,15	6,43

Para a Tabela 6, o valor do teste de logrank para a comparação entre os grupos 1 e 2 de pacientes em estudos resultou em uma $\chi^2 = 6,43$ com 2 graus de liberdade e um p-valor igual a $0,0121$, utilizando o método de Bonferroni que usa o alfa $0,05/3 = 0,017$ para cada um dos testes, assim o nível global de significância é de $0,05$, ou seja, indicando que existe uma diferença entre as curvas de sobrevivência do grupo 1 e grupo 2 esse tipo de comparação e chamada de comparações múltiplas.

Na tabela 7, aplicou-se os métodos paramétricos, afim de verificar se alguma distribuição paramétrica poderia ser utilizada no lugar dos métodos não paramétricos, para isso as distribuições exponencial, weibull e log-normal e estimou-se as suas funções de sobrevivência ($\hat{S}_{te}$, $\hat{S}_{tw}$, $\hat{S}_{tln}$, respectivamente as estimativas das funções de sobrevivência das distribuições exponencial, weibull e log-normal) pelo método de máxima verossimilhança.

Tabela 7: Estimativas da sobrevivência para os grupos de Leucemia linfoblástica usando o estimador de Kaplan-Meier e as distribuições exponencial, weibull e log-normal

indivíduos	tempo	$\hat{S}_t$	$\hat{S}_{te}$	$\hat{S}_{tw}$	$\hat{S}_{tln}$
1	1.00	0.99	1.00	0.99	1.00
2	2.00	0.99	1.00	0.99	1.00
3	10.00	0.98	0.99	0.96	0.98
4	16.00	0.97	0.99	0.95	0.97
5	35.00	0.96	0.98	0.92	0.94
6	48.00	0.96	0.97	0.90	0.92
...	...	...	...	...	...
131	2640	0.3726825	0.1706415	0.2731495	0.3129255

De acordo com a tabela 7 não podemos visualizar de uma forma clara qual distribuição está melhor se adequando aos dados, porém as estimativas da distribuição Weibull e log-normal estão em condições melhores que a distribuição exponencial. Nas figuras 3 e 4 pode-se ter uma melhor distinção.

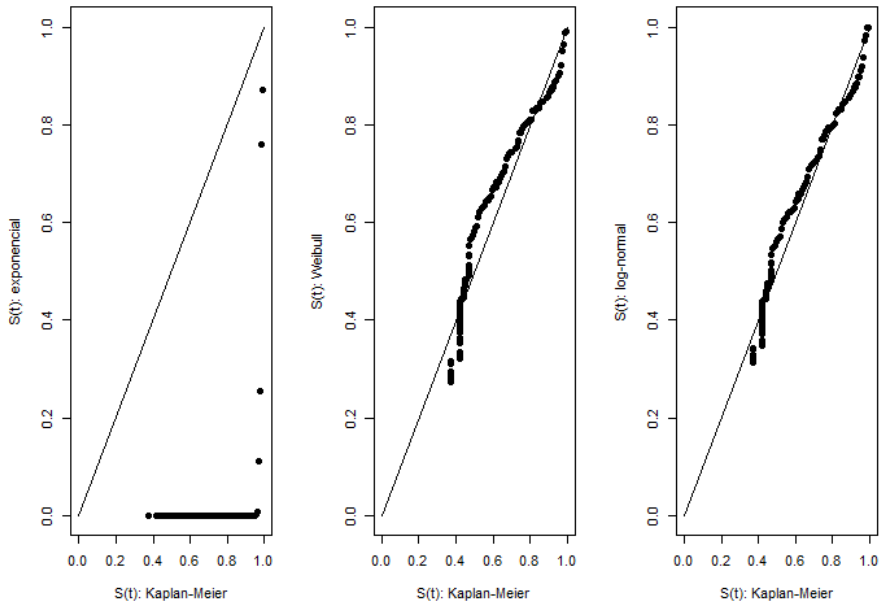

Figura 3: Gráficos das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas das distribuições exponencial, weibull, e log-normal

Na Figura 3 tem-se que o gráfico da esquerda para o modelo exponencial não há uma aproximação dos pontos a uma reta, para o modelo weibull no gráfico do centro da figura, os pontos se acomodam de forma melhor que a exponencial, assim a indícios que a Weibull em relação a exponencial seja melhor. Não obstante, o gráfico da direita também conseguiu ajustar seus pontos com a reta do estimador de kaplan-Meier tão bem quanto a Weibull e não conseguimos ter certeza de qual das duas distribuições Weibull e log-normal

conseguiu melhor explicar os dados. Para melhorar a visualização foi feito a linearização dos modelos como mostra a figura 4.

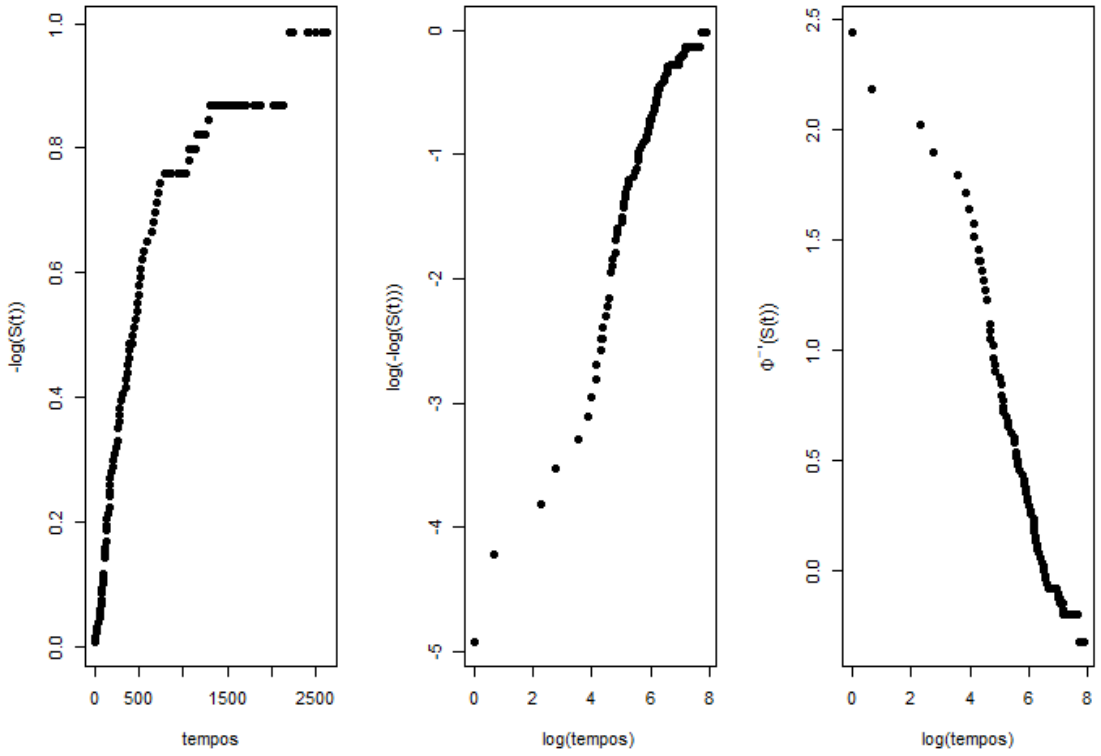

Figura 4: t versus $-log(\hat{S}(t)), log(t)$ versus $log(-log(\hat{S}(t)))$ e $log(t)$ versus $phi^{-1}(\hat{S}(t))$.

Na figura 4 percebemos que realmente a exponencial não explica bem os dados isso ficou bem claro, já os modelos Weibull e a log-normal demostram ser bons candidatos para explicar o comportamento dos dados, mas para ter uma escolha com maior confiança será feito o teste da razão de verossimilhança e sera usado o critério de Akaike.

Tabela 8: Logaritmo da função $L(\theta)$ e resultados dos TRV e AIC

Modelo	$log(L(\theta))$	TRV	Valor P	AIC (menor melhor)
Gama Generalizado	-632,2	–	–	–
Exponencial	-639,8	$2(639,8-632,2)=15,4$	$<0,00045$	1296,12
Weibull	-627,6	$2(627,6-632,2)=-9,2$	1	1275,34
Log-normal	-623,7	$2(623,7-632,2)=-17,0$	1	1268,42

Com base na tabela 8 usando o Teste da Razão de Verossimilhança (TRV) para um $\alpha = 0.01$ de significância para as seguintes hipóteses para H_0, $i)$ o modelo exponencial é adequado, $ii)$ o modelo Weibull é adequado, $iii)$ o modelo da log-normal é adequado. Como podemos observar o p-valor da exponencial e muito pequeno havendo evidências suficientes para rejeitamos a hipóteses: $i)$ com $\alpha = 0.01$ de significância. Para a log-normal

é Weibull tem-se um p-valor alto ou seja não há evidências suficientes para se rejeitar a hipótese ii) e iii) sendo plausível escolher a Weibull ou log-normal usando apenas o p-valor como critério, porém foi usado o critério de Akaike que possui o menor valor para o modelo usando a log-normal, sendo assim foi escolhido o modelo da log-normal para os dados em estudos como uma distribuição adequada para modelar o tempo de sobrevivência, risco e morte dos pacientes, uma observação importante e que se fossemos se basear apenas na Figura 4 e no TVR a Weibull também estaria se adequando aos dados.

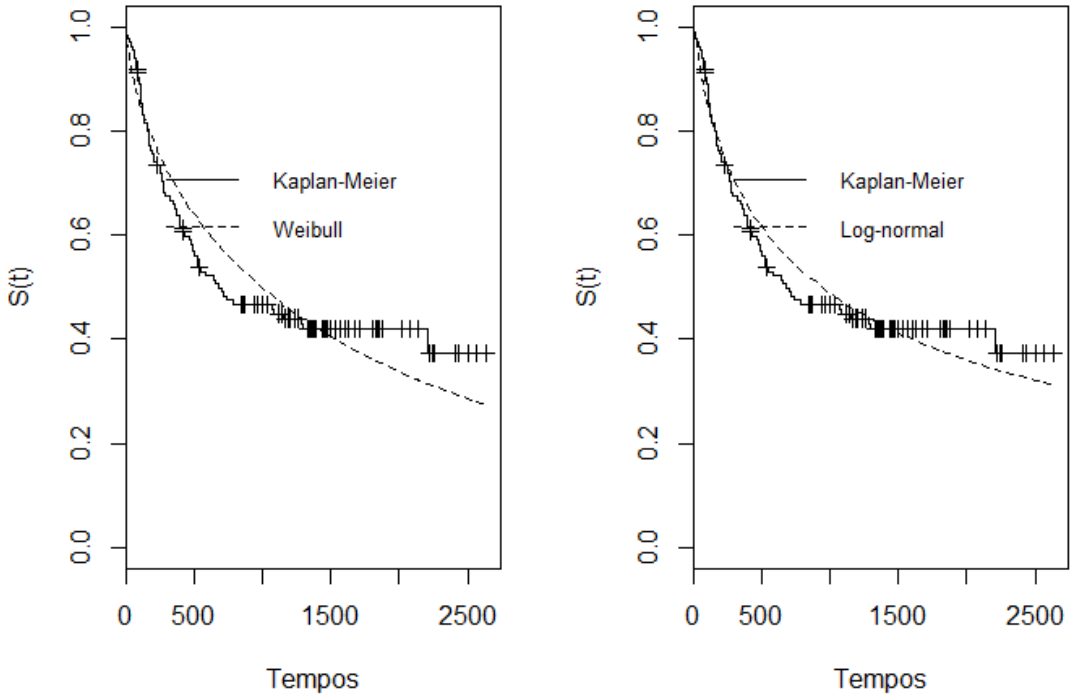

Figura 5: Curva de sobrevivência estimada pelo Kaplan-Meier versus curvas de sobrevivência estimadas usando a Weibull e a Log-normal

Podemos observar na Figura 5 o gráfico da esquerda o modelo da Weibull ajustado em conjunto com a curva de Kaplan-Meier, de forma que o ajuste da weibull tem uma leve diferença em relação a log-normal, já no gráfico da direita o modelo de regressão log-normal. A distribuição log-normal tem uma pequena melhora em relação a Weibull por isso usou-se a log-normal para esse banco de dados Leucemia linfoblástica, conforme os resultados do critério de Akaike.

O método utilizado para fazer a seleção de covariáveis foi proposto por Collett (1994), nesse estudo os passos usados nessa análise estão descrito abaixo:

1. Ajustou-se todos os modelos contendo uma única covariável Z_1 depois com Z_2,Z_3,Z_4, Z_5,Z_6,Z_7,Z_8,Z_9 até Z_{10}. Todas as covariáveis que foram significativas ao nível de 0,10 foi usado como modelo inicial ou seja Z_2+Z_8+Z_9+Z_{10}. Utilizo-se o teste da razão de verossimilhança nesse passo para retirar ou deixar uma covariável.

2. As covariáveis significativas no passo 1 foram ajustadas conjuntamente e foi retirada uma de cada vez e a cada retirada o modelo era ajustado, ou seja, o primeiro ajuste foi Z_2+Z_8+Z_9 sem Z_{10}, depois Z_2+Z_8+$Z_1$0 sem Z_9, e assim sucessivamente. No final dessa etapa restaram as mesma covariáveis do passo anterior Z_2+Z_8+Z_9+Z_{10}.

3. Ajustou-se o modelo do passo anterior Z_2+Z_8+Z_9+Z_{10} e foi acrescentada as covariáveis eliminadas no passo 1 uma a uma para verifica se elas realmente devem ser descartadas do modelo, ou seja Z_2+Z_8+Z_9+Z_{10}+Z_1 depois Z_2+Z_8+Z_9+Z_{10}+Z_3 e assim sucessivamente. No final dessa etapa restou apenas Z_2+Z_8+Z_9+Z_{10}, isto é, realmente as covariáveis deveriam ser retiradas.

O modelo final ajustado pela log-normal continha apenas 4 covariáveis com Z_2+Z_8+ Z_9+Z_{10} das 10 covariáveis estudadas para explicar os dados. As estimativas do parâmetros do modelo de regressão log-normal estão na tabela 9. Onde podemos visualizar os coeficientes na escala logarítmica que foram calculados usando o software R.

Tabela 9: Estimativas dos parâmetros do modelo de regressão log-normal ajustado aos dados de pacientes com leucemia linfoblástica.

Covariável	Estimativa	Erro-Padrão	p-valor
Constante	7.5349	0.5977	< 0.0001
Z_2	-0.0382	0.0169	0.0242
Z_8	-1.1653	0.3679	0.0015
Z_9	0.6558	0.1949	0.0008
Z_{10}	-1.5179	0.4429	0.0006
Parâmetro de forma	0.6441	0.2094	N/A

A figura 6 mostra o resultado dos resíduos na escala logarítmica com uma transformação exponencial nos resíduos, $\hat{v}_i$, ou seja, $\hat{e}_i^* = \exp\{\hat{v}_i\}$. Essa transformação é necessária para poder utilizar o estimador de Kaplan-Meier, pois os resíduos são censurados e produzem tanto valores positivos como negativos, e com essa transformação os resultados seguem uma distribuição conhecida, a log-normal padrão. Os resíduos são positivos o que facilita o uso do estimador de Kaplan-Meier.

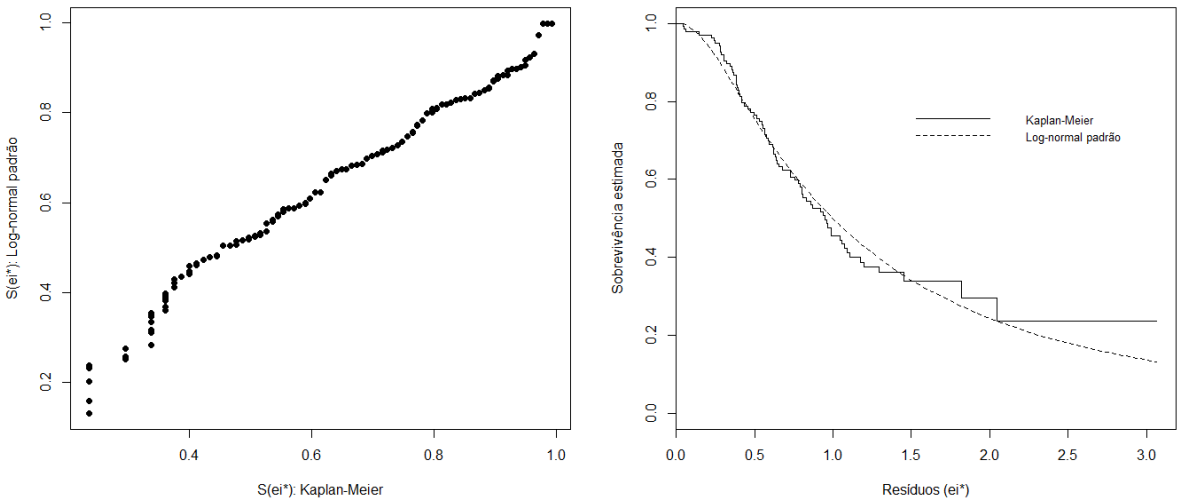

Figura 6: sobrevivências dos resíduos e_i^* estimadas pelo método de Kaplan-Meier e pelo modelo log-normal padrão (gráfico da esquerda) e respectivas curvas de sobrevivência estimadas (gráfico da direita

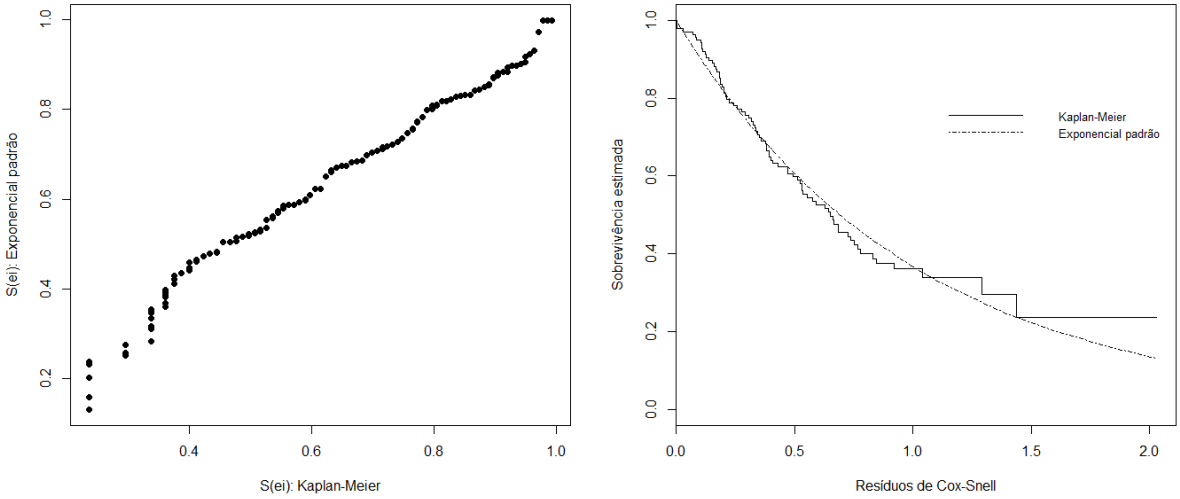

Figura 7: sobrevivências dos resíduos Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo exponencial pdrão (gráfico da esquerda) e respectivas curvas de sobrevivência estimadas (gráfico da direita)

Como podemos observar na figura 7, os resíduos Cox-Snell devem seguir uma distribuição exponencial padrão para o modelo log-normal seja adequado e os resíduos estão seguindo de forma coerente.

4 Conclusão

A partir das técnicas estatísticas usadas nesse trabalho evidenciou-se que análise de sobrevivência é uma importante ferramenta na área da saúde desde que todos os critérios que cada técnica estatística possuir sejam seguidos de forma correta. Essas técnicas ajudam a compreender o comportamento de certas patologias e de seus respectivos tratamentos, essas técnicas ajudaram a entender quais os fatores de risco que existe para o transplante de médula óssea em pacientes que possuíam leucemia linfoblástica aguda (AML) alto e baixo riscos e leucemia linfoblástica (ALL) a técnica do estimador de kaplan-Meier para os tempos de sobrevivência dos três grupos de estudo mostrou que as curvas terminam em pontos diferentes, porque os maiores tempos de estudo são diferentes para os três grupos (2081 dias para ALL , 2.569 para AML de baixo risco, e 2640 para AML alto risco). No AML baixo risco os pacientes têm o melhor prognóstico e pacientes de alto risco AML tem o prognóstico menos favorável.

A distribuição de probabilidade que foi usada para explicar os dados da amostra, foi a log-normal selecionada pelos métodos gráficos, teste da razão de verossimilhança e critério de Akaike, e posteriormente o modelo de regressão log-normal que auxilio na escolha das covariáveis que possuem maior influência, para o aumento do risco de morte do transplante de médula óssea, para a amostra de pacientes com leucemia linfoblástica foram a idade do doador, FAB, Hospital e MTX, com essas técnicas a análise de sobrevivência permite aos especialistas da área da saúde entender os fatores que afetam o tempo de sobrevivência e elaborar estratégias de como conduzir os tratamentos para essa patologia.

Referências

Aalen, O. O. Nonparametric Inference for a Family of Counting Processes. **The Annals of Statistics**, p. 701 - 726, 1978.

Bohoris, G. Comparison of the Cumulative-Hazard and Kaplan-Meier Estimators of the Survivor Function, **IEEE Transactions on Reliability**, v. 43, n. 2, p. 230 - 232, 1994.

Collett, D. **Modelling Suival Data in Medical Research**. London, Chapman and Hall, 1994.

Colosimo, E. A.; Giolo, S. R. ,**Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blucher Ltda, 2006.

Cox, D. R. Regression models and Life Tables. **Journal os the Royal Statistical Society**, p. 187-220, 1972.

Cox, D. R; Snell, E. J. A General Definition of Residuals. **Journal os the Royal Statistical Society**, p. 248-275, 1968.

Draper N. R.; Smith, H. **Applied Regression Analysis**, 3. ed. New York: Jhon Wiley and Sons, 1998.

Ferreira, J. M. Tabela de Vida. in: **Análise de sobrevivência: uma visão de risco comportamental na utilização de cartão de crédito**, 2007. Disponível em ¡http://www.pgbiom.ufrpe.br/dissertacoes/2007/d2007-08.pdf¿. Acesso em: 24 mar. 2014.

Klein, J. P. **Moeschberger, M.L. Survival Analysis: Techniques for Censored and Truncated Data**. 2. ed. New York: Springer, 2003.

Lawless, J. F. **Statistical Models and Methods for lifetime Data.** New York: Jhon Wiley and Sons, 1982.

Lee, E. T; Wang, J. W. **Statistical methods for survival data analysis**. 3. ed. Okalahoma: Wiley, p. 134-135, 2003.

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**. p. 163-170, 1996.

Nelson, W. Theory and Applications of Hazard Plotting for Censored Failure Data. **Technometrics**, p. 945 - 965, 1972.

Nelson, W. **Accelerated Life Testing: Statiscal Models, Data Analysis and Test Plans**. New York: Jhon Wiley and Sons, 1990.

Stacy, E.W. Generalization of the Gamma Distribution. **Ann. Math. Stat**. v:33, p. 1187-1192, 1962.

Wald, A. Test Statistical Hypotheses concerning Several Parameters when the Number of Observations is Large. **Trans. Amer. Math. Soc**. p. 426-482, 1943.

Strapasson, E. **Comparação de modelos com censura intervalos em análise de sobrevivência**. Piracicaba: ESALQ, 2007.

Weibull, W. A Statistical Theory of the Strength of Material. Proc. **Ingeniors Vetenskaps Akademien Handlingar**. n:151, p. 293-297, 1939.

Weibull, W. A statistical distribution function of wide applicability. **Jounal of Applied Mechanics**. Sweden, p. 293-297, 1951.

APÊNDICE A – Rotina da análise feita utilizando o software R (Gratuito)

Toda a rotina foi baseada na do livro de colosimo e giolo (2006), abaixo é apresentada a parte inicial da rotina.

```
require(survival) #caso não tenha o pacote
library(survival)
# usando o estimador de kaplan-Meier

#tempo1= "Tempo da morte"
#censura=  " Recidiva da doença"
#falha= "Morte"

tempo1=c(2081, 1602, ...  ,363 )
length(tempo1)

censura=c(0,0,..,1)
length(censura)

grupos<-c(rep(1,38),rep(2,54),rep(3,45))
length(grupos)

ekm=survfit(Surv(tempo1,censura)~grupos)
summary(ekm)

## Kaplan-Meier gráfico
```

```
plot(ekm, lty=c(3,2,1),xlab="tempo(dias)",ylab="S(t) estimada")
legend(1,0.34, lty=3:1,c("Leucemia linfoblástica" ,"LMA baixo risco",
"LMA alto risco"), lwd=1,bty="n")

## ic do kaplam-Meier
ekm<- survfit(Surv(tempo1,censura)~grupos,conf.type="plain")
summary(ekm)
plot(ekm,conf.int=T,  xlab="Tempo (em meses)", ylab="S(t) estimada", bty="n")
##########################################################
# Nelson-Aalen
#ss<-survfit(coxph(Surv(tempo1,censura)~grupos,method="breslow"))
# caso geral sem fazer diferenças de grupo usando todos os grupos em um só
ss<- survfit(coxph(Surv(tempo1[grupos==3],censura[grupos==3])~1,method =
"breslow"))
summary(ss)
racum=-log(ss$surv)
racum
#########################################################################
####### teste logrank
ekm=survfit(Surv(tempo1,censura)~grupos)
summary(ekm)

plot(ekm,lty=c(1,4,2),xlab="Tempos",ylab="S(t) estimada")
legend(1,0.35,lty=c(1,4,2),c("Leucemia Linforblástica",
"LMA baixo risco","LMA baixo alto"), lwd=1,bty="n",cex=0.8)
survdiff(Surv(tempo1,censura)~grupos,rho=0)
survdiff(Surv(tempo1[1:92],censura[1:92])~grupos[1:92],rho=0)
# grupo1 vs grupo2
survdiff(Surv(tempo1[39:137],censura[39:137])~grupos[39:137],rho=0)
# grupo2 vs grupo3
survdiff(Surv(c(tempo1[1:38],tempo1[55:137]),c(censura[1:38],censura[55:137]))~
c(grupos[1:38],grupos[55:137]),rho=0)
#grupo1 vc grupo3
```