UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Estimação de curvas de sobrevivência para estudos de custo-efetividade

Autora: Letícia Herrmann
Orientadora: Professora Dr. Patrícia Klarmann Ziegelmann

Porto Alegre, 6 de Dezembro de 2011.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Estimação de curvas de sobrevivência para estudos de custo-efetividade

Autora: Letícia Herrmann

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dr. Patrícia Klarmann Ziegelmann
Dr. Rodrigo Antonini Ribeiro

Porto Alegre, 6 de Dezembro de 2011.

Dedico este trabalho aos meus pais, Elaine e Roque,
e ao meu irmão, Samuel,
que sempre me apoiaram e incentivaram.

"Se não puder se destacar pelo talento, vença pelo esforço."

Dave Weinbaum

Agradecimentos

Agradeço ao meu irmão Samuel por ser o grande “culpado” por eu estar onde estou. Obrigada por me fazer descobrir e sonhar em estudar na UFRGS. Te amo!

Agradeço aos meus pais, que sempre me apoiaram, incentivaram e acreditaram em mim, tornando a realização deste sonho possível. Obrigada por terem entendido quando não pude ir pra casa ou quando não pude estar tão presente quanto gostaria. Foi pensando em vocês e no que vocês fizeram por mim que eu me dediquei e tentei sempre dar o meu melhor. Obrigada por tudo, amo muito vocês!

Agradeço aos professores da UFRGS pelas aulas e pelo conhecimento compartilhado. Faço um agradecimento especial à professora Patrícia Ziegelmann, uma ótima professora e orientadora. Obrigada pelas ótimas aulas, por ter me aceitado como tua orientanda, por todas as ajudas e, principalmente, pela paciência comigo. Agradeço também ao médico Rodrigo Ribeiro pelo banco de dados utilizado neste trabalho.

Agradeço ao Andrigo Rodrigues que foi um grande companheiro. Obrigado pelas horas de estudo juntos e pela ajuda especialmente no início do curso. Muito obrigada pelo carinho, pela atenção, pelos colos, por toda paciência que você teve, por sempre me encorajar, apoiar, acreditar em mim e não ter me deixado desanimar.

Agradeço aos colegas de curso e amigos que conquistei. Agradeço em especial a Natália Barbieri pela amizade criada já no dia da matrícula, pelos milhares de e-mails trocados, por ter se tornado uma irmã para mim; a Mariana Bartels pela amizade, listas emprestadas, dúvidas tiradas; ao Mauro Lacerda por tantas ajudas, inclusive em gráficos para este trabalho; a Pricila Maciel pelos nossos estudos juntos; ao Mateus Becker, Japa, por ter o dom de conseguir me acalmar; ao Henrique Helfer, Alemão, por acordar cedo de segunda e quarta este semestre só porque eu pedi pra ele ser meu colega; ao Jonas Pacheco pela parceria e bom humor, deixando qualquer um feliz também.

Agradeço aos colegas de trabalho Eduardo Schneider, Marcel Becker, Fernanda Quadros e Melissa Pivoto que me ensinaram muito. Agradeço em especial a Ana Paula Queiroz Sperotto que é um exemplo de profissional e chefia, e que tem sido uma amiga muito querida com quem tenho aprendido muito.

Agradeço aos meus amigos de Canudos do Vale, de maneira especial às minhas amigas Patrícia Schmidt e Alana Ledur, pela amizade verdadeira desde sempre e pela compreensão nos momentos de ausência.

Resumo

Esta monografia tem como objetivo principal produzir um material prático e claro sobre análise de dados de sobrevivência, que possa vir a auxiliar aos que desejarem utilizar essas técnicas, especialmente quando aplicadas a estudos de custo-efetividade. Entende-se por dados de sobrevivência dados provenientes de estudos longitudinais em que indivíduos são acompanhados até a ocorrência do evento de interesse. A análise de sobrevivência modela o tempo até a ocorrência do evento de interesse e incorpora a informação das censuras, ou seja, utiliza o tempo até a censura daqueles indivíduos que participaram do estudo e não falharam.

Neste trabalho, são introduzidos os principais conceitos e funções da análise de sobrevivência, como também são apresentadas diferentes maneiras de executá-la. Além da descrição dos modelos mais conhecidos da análise de sobrevivência, como os modelos não-paramétricos de Kaplan-Meier e a tábua de vida, são apresentados também os modelos paramétricos e modelos de regressão paramétricos, que são amplamente utilizados em análise de custo-efetividade por permitirem a extrapolação da função de sobrevivência. Para estas técnicas, foi desenvolvido um exemplo passo a passo no software STATA, utilizando um banco de dados de acompanhamento de pacientes do ambulatório de insuficiência cardíaca do Hospital de Clínicas de Porto Alegre, a fim de auxiliar a compreensão das técnicas e servir de guia para aqueles que desejarem conduzir uma análise de sobrevivência.

Abstract

This monograph's main objective is to produce a clear and practical material on the analysis of survival data, which can assist those who wish to use these techniques, especially when applied to studies of cost-effectiveness. Survival data is described as data from longitudinal studies in which individuals are followed until the occurrence of the event of interest. Survival analysis models the time until the occurrence of the event of interest and incorporates information from censoring, in other words, it uses the time until the censoring of those individuals who participated in the study and did not get the event.

This work introduces the main concepts and functions of survival analysis and also presents different ways to perform it. In addition to the description of the most well-known models of survival analysis, as the non-parametric models of Kaplan-Meier and the life table, it is presented the parametric models and parametric regression models that are widely used in cost-effectiveness analysis, as they allow the extrapolation of the survivor function. For these techniques, a step by step example in the STATA software was developed using a database of patients monitored by the ambulatory of heart failure at the Hospital de Clínicas de Porto Alegre, in order to support the understanding of the techniques and serve as a guide for those who wish to conduct a survival analysis.

SUMÁRIO

1. Introdução

Em diversas áreas, e em especial na área médica, é prática frequente trabalhar com estudos longitudinais que investigam o tempo até a ocorrência de determinado evento de interesse. Dados de sobrevivência são provenientes deste tipo de estudo e definem a variável resposta como o tempo até a ocorrência do evento de interesse. Contudo, este tipo de estudo geralmente possui dados censurados, ou seja, observações em que o evento de interesse não ocorreu durante o tempo de duração do estudo. Estas características exigem técnicas estatísticas apropriadas para analisar este tipo de observação.

A análise de sobrevivência é uma metodologia desenvolvida para trabalhar com dados de tempo até a ocorrência de um determinado evento de interesse e que tem como característica principal incorporar a informação proveniente dos dados censurados. Algumas das referências mais importantes da análise de sobrevivência são Hosmer e Lemeshow (1999) e Collett (2003) e, na literatura nacional, Colosimo e Giolo (2006). Estes textos foram utilizados como base deste trabalho. Além desses, Cleves *et al* (2008) é um texto interessante, que concilia a teoria dos modelos de análise de sobrevivência com alguns comandos necessários para realizar as análises no *software* STATA.

Em estudos de análise de sobrevivência, é necessário definir o evento de interesse para que se possam obter as duas informações essenciais para realizar a análise: o tempo até a ocorrência do evento de interesse e a definição da natureza deste evento como falha ou censura.

O estimador Kaplan-Meier é provavelmente a técnica de análise de sobrevivência mais utilizada na análise clínica. Este estimador é utilizado quando se dispõe dos dados do tempo até a ocorrência da falha ou censura para todos os indivíduos do estudo, e seu resultado irá representar a realidade daqueles dados, tendo, portanto, um caráter descritivo. A tábua de vida é uma das técnicas mais antigas utilizadas em análise de sobrevivência. Com as tábuas de vida populacionais é possível estimar a probabilidade de morte para cada idade. Ainda, é possível obter uma tábua de vida para uma população específica a partir de uma tábua de vida para população geral, se a informação da razão das taxas de falha da população geral em relação à específica é conhecida. Os modelos de regressão de Cox também são amplamente utilizados em análise de sobrevivência. Este modelo é semi-paramétrico e, portanto, bastante flexível.

Além disso, é um modelo muito difundido, uma vez que é a técnica de análise de sobrevivência mais abordada na literatura. Seu objetivo é identificar os fatores de risco, ou seja, fatores que influenciam na sobrevida das pessoas.

Os modelos paramétricos e os modelos de regressão paramétricos são menos difundidos que os modelos de análise de sobrevivência citados anteriormente. Estes modelos têm como característica básica a suposição de uma distribuição densidade de probabilidades para o tempo até a ocorrência do evento de interesse. Esta suposição precisa ser verificada e, para que possam ser utilizados, algum modelo paramétrico deve ajustar-se bem aos dados. Uma vantagem da utilização dos modelos paramétricos está no fato de que, em consequência desta suposição, é possível realizar estrapolação dos dados. A extrapolação é utilizada quando se deseja obter uma probabilidade de sobrevivência para algum tempo superior ao período de tempo de duração dos estudos.

A análise de sobrevivência é uma ferramenta importante para estudos de Avaliação de Tecnologias em Saúde (ATS). Estudos de Avaliação de Tecnologias em Saúde objetivam fornecer informações quanto ao possível impacto e consequências de uma nova tecnologia, ou da mudança de uma tecnologia já existente, para os tomadores de decisão. A análise de custo-efetividade considera os custos e a efetividade das intervenções a fim de informar qual das opções representa um maior benefício e a qual custo incremental. A modelagem de dados sobrevivência é importante neste tipo de análise como medida de efetividade, onde é possível medir o impacto de uma nova tecnologia na sobrevivência (Latimer, 2011).

Gray *et al* (2011) abordam os modelos de análise de sobrevivência como ferramenta para a análise de custo-efetividade, tratando da necessidade de extrapolação das curvas de sobrevivência. Este tipo de prática é muito comum em estudos de custo-efetividade. Recentemente, o *National Institute for Health and Clinical Excellence* (NICE) publicou um relatório técnico (Latimer, 2011) tratando da utilização dos modelos de análise de sobrevivência em estudos de avaliação econômica, enfatizando o tema da extrapolação dos dados. Estas bibliografias mostram a importância da utilização de modelos paramétricos na análise de sobrevivência, uma vez que somente com estes modelos é possível fazer extrapolação.

O objetivo principal desta monografia é produzir um material sobre análise de sobrevivência que tenha cunho prático e não seja tão teórico quanto à maioria dos livros de estatística sobre análise de sobrevivência. O foco deste trabalho voltou-se às técnicas de análise de sobrevivência mais utilizadas em estudos de custo-efetividade. Especial

atenção é dada aos modelos paramétricos, visto sua aplicabilidade para extrapolação de curvas de sobrevivência. Ainda, são destacados métodos para auxiliar na escolha do modelo paramétrico a ser utilizado.

No Capítulo 2, são apresentadas as características dos dados de sobrevivência, quais dados são necessários para realizar uma análise de sobrevivência e quais são as funções mais utilizadas neste tipo de análise. No Capítulo 3, são apresentadas as técnicas de Kaplan-Meier e tábua de vida através de exemplos. No Capítulo 4, são apresentas os modelos paramétricos exponencial, Weibull, log-normal e Gompertz, técnicas que auxiliam na escolha do modelo adequado aos dados e um exemplo. No Capítulo 5, são apresentados os modelos de regressão paramétricos, tanto pela abordagem de modelos de taxas de falha proporcionais como pela abordagem de modelos de tempo de vida acelerado, técnicas que auxiliam na escolha do modelo adequado aos dados, técnicas para verificar a qualidade do ajuste e um exemplo para cada abordagem dos modelos de regressão paramétricos. No Capítulo 6, são apresentadas as considerações finais.

2. Características de Dados de Sobrevivência

A análise de sobrevivência é uma técnica estatística adequada para analisar dados de acompanhamento de indivíduos ao longo do tempo até a ocorrência de um determinado evento de interesse. Contudo, há estudos em que o evento de interesse não ocorrerá para todos os indivíduos, resultando em observações incompletas, ditas censuradas.

As técnicas de análise de sobrevivência incorporam a informação da censura. Esta característica é muito importante e a diferencia das demais técnicas estatísticas usuais.

2.1. Dados Necessários para Análise de Sobrevivência

Para realizar uma análise de sobrevivência, é necessário dispor das informações sobre tempo de falha e censura, porque estas duas informações constituem a resposta da análise de sobrevivência. A variável do tempo de falha armazena o tempo até a ocorrência do evento de interesse e a variável de censura indica se determinado indivíduo falhou ou foi censurado. Os conceitos de tempo de falha e censura são explorados na Seção 2.1.1. e na Seção 2.1.2..

2.1.1. Tempo de Falha

O tempo de falha é o tempo transcorrido entre a entrada do indivíduo no estudo e a ocorrência do evento de interesse. Esta variável deve estar bem definida e, para tanto, o tempo inicial do estudo, a escala de medida desse tempo e o evento que caracterizará a falha também precisam estar bem definidos.

O tempo de origem do estudo é utilizado para o cálculo do tempo até a falha. Em estudos clínicos geralmente essa data é definida como o início do estudo. Isso acontece quando, por exemplo, os pacientes começam a tomar diferentes medicações ao mesmo tempo. Porém, há estudos em que o indivíduo não participa do estudo a partir da sua data inicial. Neste caso o tempo de falha é o tempo que esse indivíduo realmente ficou em tratamento, ou seja, o tempo transcorrido desde o momento da entrada do paciente no estudo até a ocorrência da falha. Note que este tempo será diferente do tempo transcorrido a contar da data inicial do estudo até a ocorrência do evento de interesse.

A escala de medida geralmente é o tempo real ou o tempo cronológico. Pode ser medida em dias, horas, semanas, etc.. É importante definir a escala de medida utilizada para garantir que os tempos até a falha ou censura de todos os indivíduos sejam medidos na mesma escala.

O evento que caracteriza a falha deve estar definido desde o início do estudo e de forma muito clara. Como este evento geralmente refere-se a algum evento indesejado, ele é também chamado de falha. É a partir da definição desse evento que é possível calcular o tempo até a falha dos indivíduos estudados. Se a falha estiver bem definida não haverá dúvidas quanto à natureza do evento que ocorreu com o indivíduo, se foi uma falha ou se foi uma censura. Se as definições de falha e censura forem confundidas, a coleta das informações poderá conter erros, comprometendo a qualidade dos resultados da análise de sobrevivência.

A falha pode, ainda, ser definida como a ocorrência de um único evento ou de mais de um evento. Quando dois ou mais eventos representam uma falha, há o chamado risco competitivo. Os modelos que serão apresentados neste trabalho são apropriados para falhas definidas pela ocorrência de um único evento.

2.1.2. Censura

A censura ocorre quando um indivíduo participante do estudo não falha. Isto habitualmente acontece nos estudos de acompanhamentos dos indivíduos ao longo do tempo. Em estudos clínicos, por exemplo, pode-se perder contato com algum paciente caso ele passe a residir em outra cidade, morra em função de fatores externos, ou porque o estudo acabou sem que este paciente tenha falhado.

A informação que a censura fornece é de que determinado paciente não falhou até o tempo em que foi censurado. Ou seja, o tempo até a ocorrência do evento será superior ao tempo registrado até o último acompanhamento do indivíduo. Utilizar os dados censurados nas análises é extremamente importante porque, mesmo que parcial, a censura fornece informação. Se essa informação não for utilizada no cálculo das estatísticas desejadas, pode-se gerar resultados viciados (Colosimo e Giolo, 2006), tornando inadequado o uso de estatísticas como média e desvio padrão, por exemplo.

A variável indicadora de censura, δ_i é definida como:

$$\delta_i = \begin{cases} 1 \; se \; t_i \; é \; um \; tempo \; de \; falha \\ 0 \; se \; t_i \; é \; um \; tempo \; censurado \end{cases}$$

A variável resposta da análise de sobrevivência é formada pelo par de informações do tempo de falha t_i, e pela variável indicadora de censura δ_i. Além disso, quando são coletadas informações adicionais dos pacientes, prática frequente em estudos clínicos, as covariáveis passam a fazer parte da resposta. A resposta será dada, então, por $(t_i, \delta_i, \boldsymbol{x}_i)$, onde $\boldsymbol{x}_i$ representa o vetor de covariáveis.

Há três possíveis mecanismos de censura: tipo I, tipo II e aleatório. A censura tipo I ocorre quando um estudo tem tempo de duração pré-estabelecido e os indivíduos falham em função do término do estudo. A censura tipo II é aquela que ocorre quando se estabelece o fim estudo após a ocorrência de um número pré-determinado de falhas. A censura aleatória ocorre, por exemplo, quando um indivíduo sai do estudo durante sua execução, ou morre de alguma causa diferente da definida como o evento da falha. Segundo Colosimo e Giolo (2006), censura aleatória é o mecanismo de censura mais frequente na prática médica. A Figura 1 apresenta os mecanismos de censura, onde as falhas são representadas pelo símbolo "●" e as censuras pelo símbolo "∘".

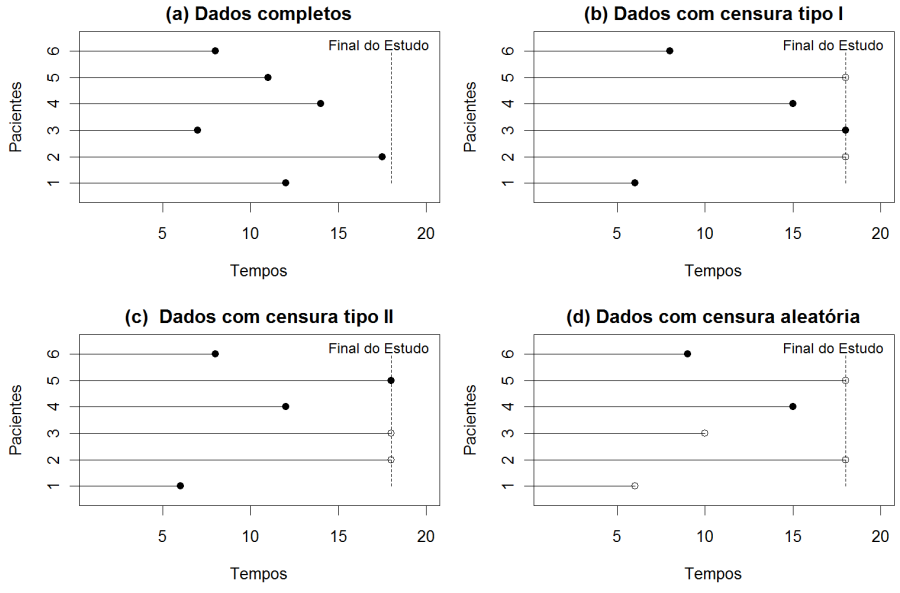

Figura 1: Mecanismos de censura. Figura retirada de Colosimo e Giolo (2006).

Há também três possíveis formas de censura: à direita, à esquerda e intervalar. A censura à direita ocorre quando o tempo de ocorrência do evento está à direita do tempo registrado, como é o caso das censuras apresentadas na Figura 1. A censura à esquerda ocorre quando o tempo registrado é maior que o tempo de falha, ou seja, no momento em que o indivíduo é observado a falha já ocorreu. Uma pessoa que possui HIV e começa a participar de um estudo, mas que não sabe exatamente quando foi infectada, é um exemplo de censura à esquerda. Há casos chamados duplamente censurados, que acontecem quando ocorre censura à esquerda e à direita simultaneamente.

A censura intervalar ocorre quando o paciente é acompanhado periodicamente, e o evento de interesse ocorre dentro de um intervalo de tempo. Imagine uma situação em que uma pessoa retirou um tumor através de um procedimento cirúrgico. Esta pessoa irá consultar-se de dois em dois meses, a fim de saber se o tumor voltou ou não. Se na consulta dois meses após a cirurgia não há tumor, mas na nova consulta quatro meses após a cirurgia há um tumor, sabe-se apenas que a volta do tumor ocorreu entre dois e quatro meses após a realização da cirurgia. Neste caso, a censura é intervalar.

A censura também pode ser classificada com informativa e não informativa. A censura informativa ocorre quando a perda do indivíduo é em decorrência de uma causa associada ao evento de interesse do estudo, como por exemplo, quando um paciente abandona um tratamento em função da piora no seu estado de saúde. A censura não informativa, também chamada de independente, ocorre quando o motivo da perda da informação é independente do desfecho. Segundo Bastos e Rocha (2006), a censura informativa deve ser evitada, porque um indivíduo censurado no instante t não será representativo de todos os indivíduos com as mesmas características e que também sobreviveram até o instante de tempo t.

É preciso tomar cuidado para não confundir censura com truncamento. Um truncamento é feito quando se utiliza algum critério para a seleção dos indivíduos que participarão do estudo. Quando truncamentos são realizados, há a exclusão de indivíduos que não satisfazem os critérios desejados.

Há duas formas de truncamento: à esquerda e à direita. O truncamento à direita ocorre quando um critério de seleção inclui somente indivíduos que já tenham sofrido o evento que vai ser estudado, e o truncamento à esquerda ocorre quando a perda da informação está relacionada aos indivíduos que foram excluídos do estudo porque já tinham experimentado o evento de interesse.

As técnicas de análise de sobrevivência tratadas neste trabalho são adequadas para censura à direita e para censura não informativa. Elas também são utilizadas, na prática, para dados oriundos dos três mecanismos de censura. Colosimo e Giolo (2006) e Collett(2003) apresentam algumas técnicas de análise de sobrevivência para dados com censura intervalar, e Turnbull (1976) apresenta alguns tratamentos especiais para dados truncados e censurados.

2.2. Apresentação dos Dados de Sobrevivência

Dados de sobrevivência são geralmente representados por duas funções, chamadas de função de sobrevivência e função taxa de falha. Para o cálculo destas funções, define-se o tempo de falha ou censura como a variável aleatória T. Note que T é uma variável não negativa e, em geral, do tipo contínua.

A Seção 2.2.1 apresenta a função de sobrevivência com mais detalhes e a Seção 2.2.2. apresenta a função taxa de falha em mais detalhes. A Seção 2.2.3. mostra que a função de sobrevivência e a função taxa de falha são matematicamente relacionadas.

2.2.1. Função de Sobrevivência

A função de sobrevivência é a forma mais intuitiva de apresentar dados de sobrevida. Ela é definida como a probabilidade de uma observação não falhar até o tempo t, ou equivalentemente, de sobreviver ao tempo t. Ou seja,

$$S(t) = P(T \geq t).$$

A função de sobrevivência também pode ser encontrada utilizando a função densidade de probabilidade acumulada $F(t)$. Isto é,

$$S(t) = 1 - F(t).$$

A Figura 2 apresenta curvas de sobrevivência para dois grupos de pacientes submetidos a dois diferentes tratamentos. Analisando o gráfico pode-se observar que a probabilidade de sobrevivência do grupo 1 é superior à do grupo 2 até pouco mais de dois meses. A partir deste momento, o grupo 2 passa a ter probabilidade de

sobrevivência superior ao grupo 1. Ainda, é possível verificar que aproximadamente 20% dos pacientes do grupo 1 ainda estão vivos quando completam quatro meses de tratamento. Para este mesmo período de duração do tratamento, há aproximadamente 30% dos pacientes do grupo 2 ainda vivos.

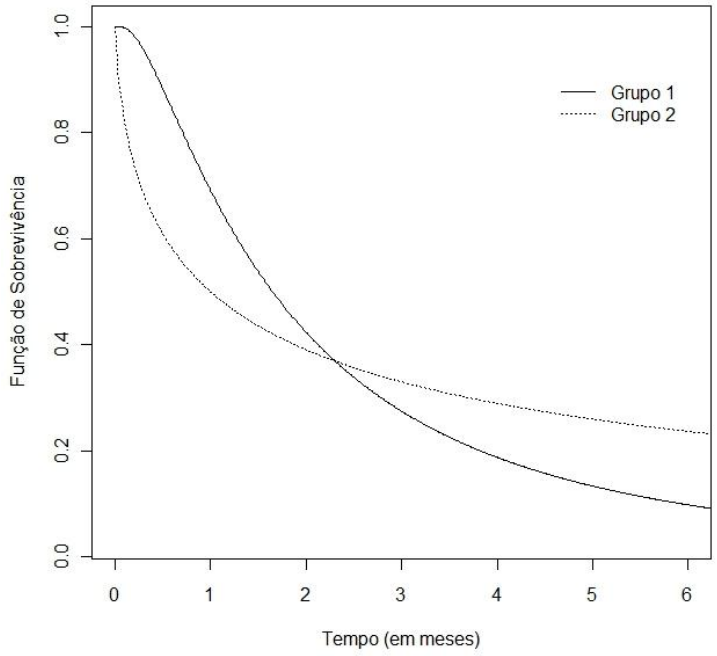

Figura 2: Curvas de sobrevivência para o grupo 1 e para o grupo 2.

2.2.2. Função Taxa de Falha

A função taxa de falha é mais conhecida como função *hazard*. Ela é calculada através da razão entre a probabilidade de uma falha ocorrer em um determinado intervalo de tempo $[t, t + \Delta t)$, dado que não tenha acontecido falha até o instante de tempo t, e o tamanho do intervalo. Ou seja,

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Na expressão acima, assumindo Δt como um valor próximo de zero, o intervalo de tempo será tão pequeno que a função taxa de falha pode ser considerada a taxa de

falha instantânea para um indivíduo que sobreviveu ao tempo t. Então esta função pode então ser reescrita como:

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

A função taxa de falha pode ser interpretada como um indicador natural da propensão à falha após uma unidade de tempo ter transcorrido. Note que a função taxa de falha é sempre positiva e não é limitada, podendo assumir valores maiores que 1.

A função taxa de falha pode ser decrescente, constante ou crescente. Se ela for decrescente significa que a taxa de falha diminui à medida que o tempo passa, se for crescente significa que a taxa de falha aumenta com o transcorrer do tempo e se for constante a taxa de falha será sempre a mesma, independente do tempo. Ainda, há situações em que ela pode começar crescente e, a partir de um instante de tempo t, passar a decrescer, e há situações em que ela pode iniciar decrescente e passar a crescer em algum instante de tempo t.

Outra função que pode ser utilizada em análise de sobrevivência é a função taxa de falha acumulada, $H(t)$. Esta função, porém, possui difícil interpretação e não será detalhada neste trabalho.

2.2.3. Relações entre as funções

Uma interessante característica entre a função de sobrevivência, a função taxa de falha e a função taxa de falha acumulada, é que elas são matematicamente relacionadas. Essa relação pode ser útil nos processos de estimação ou em situações que se conhece uma das funções e se deseja obter outra função.

Há situações em que a estimativa de uma função terá propriedades melhores que a estimativa de outra função e, neste caso, estima-se a função que possui o melhor estimador e encontra-se a outra função através da relação entre elas.

Essas relações são expressas por:

$$1)\ h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)),$$

onde $f(t)$ é a função densidade de T,

$$2)\ H(t) = \int_0^t h(u)du = -\log S(t),$$

$$3)\ S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)du\right\}.$$

2.3. Descrição dos Dados Utilizados nos Exemplos

Este trabalho apresenta diferentes maneiras de conduzir uma análise de sobrevivência, e propõe-se a exemplificar o uso destas técnicas. Estes exemplos irão utilizar o banco de dados aqui caracterizado.

O banco de dados utilizado é um banco de dados reais, do ambulatório de insuficiência cardíaca do Hospital de Clínicas de Porto Alegre (HCPA). O evento de interesse deste estudo foi definido como a morte ocasionada por todos os tipos de causas.

Compõe o banco de dados as variáveis:

- follow: tempo até a morte ou censura dos pacientes do ambulatório de insuficiência cardíaca do HCPA, medido em dias;
- obitot: variável indicadora de falha, onde 1 indica que o paciente falhou e 0 indica que o paciente foi censurado;
- inter: variável de internação, onde 1 indica que o paciente já havia sido internado pelo menos uma vez antes do início do estudo e 0 indica que o paciente nunca havia sido internado antes do início do estudo;
- sexo: variável indicadora do sexo do paciente, onde 1 significa que o paciente é do sexo feminino e 0 indica que o paciente é do sexo masculino;
- id: variável de identificação do paciente.

As variáveis follow e obitot são as duas variáveis essenciais para realizar a análise de sobrevivência, porque juntas formam a variável resposta. As variáveis inter e sexo são covariáveis cuja informação também foi coletada durante o estudo.

A Tabela 1 apresenta estas variáveis para os sete pacientes com menor tempo de falha ou censura. Por exemplo, o paciente 1 (id=1) é do sexo masculino (sexo=0), morreu (obitot=1) depois de 35 dias de acompanhamento (follow=35) e não havia sido

internado em nenhuma ocasião anterior ao início do estudo (inter=0). Já o paciente 113 (id=113), também do sexo masculino (sexo=0), que já havia sido internado pelo menos uma vez antes do início do estudo (inter=1), foi censurado (obitot=0) depois de 101 dias de acompanhamento (follow=101).

Tabela 1: Dados dos sete pacientes com menor tempo de falha ou censura.

id	follow	obitot	inter	sexo
1	35	1	0	0
2	69	1	1	0
3	75	1	0	1
4	80	1	0	0
104	82	0	0	0
5	96	1	0	0
113	101	0	1	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

O estudo é composto por 318 pacientes, onde apenas 96 pacientes tiveram como desfecho a morte ocasionada por qualquer tipo de causa. Destes 318 pacientes, 219 são do sexo masculino e 153 já haviam sido internados pelo menos uma vez.

3. Modelos Não-Paramétricos

As funções de análise de sobrevivência podem ser estimadas segundo três abordagens: utilizando métodos não-paramétricos, semi-paramétricos e paramétricos.

Os modelos não-paramétricos, que serão abordados neste capítulo, são assim definidos porque não há suposição de distribuição de probabilidade ao tempo de sobrevivência. Estes modelos tem caráter basicamente descritivo, e podem ser utilizados para auxiliar na escolha do modelo paramétrico adequado. O estimador de Kaplan-Meier e o método da tábua de vida serão abordados na Seção 3.1. e Seção 3.2., respectivamente. Estes modelos serão apresentados através do desenvolvimento de um exemplo.

O estimador de Nelson-Aalen não será abordado neste trabalho porque ele possui características muito semelhantes as do estimador de Kaplan-Meier. Além do mais, suas estimativas para a função de sobrevivência são iguais ou superiores as obtidas pelo estimador Kaplan-Meier (Colosimo e Giolo, 2006). Detalhes sobre este estimador podem ser encontrados em Colosimo e Giolo (2006) e em Hosmer e Lemeshow (1999).

Ainda, Colosimo e Giolo (2006) destacam que, entre os estimadores não-paramétricos, há superioridade do estimador de Kaplan-Meier, indicando que é preferível a utilização deste estimador.

3.1. Estimador Kaplan-Meier

O estimador Kaplan-Meier também é conhecido como estimador limite-produto. Ele é, provavelmente, a técnica de análise de sobrevivência mais conhecida e utilizada. Este estimador, criado por Kaplan e Meier, é uma adaptação da função de sobrevivência empírica, ou seja, a função de sobrevivência estimada na ausência de censura. Ele pode ser utilizado como análise exploratória dos dados.

O estimador Kaplan-Meier estima uma curva de sobrevivência incorporando a informação da censura. Com o objetivo de melhor caracterizar a incorporação da censura, inicialmente será apresentado um exemplo com dados sem censura. Para este exemplo, é suposto não existir censura nos dados descritos na Seção 2.3., ou seja, é suposto que todos os pacientes do estudo tiveram como desfecho a morte. No caso de

não haver censura, a função de sobrevivência é estimada usando a função de sobrevivência empírica, que é definida por:

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

A Tabela 2 apresenta os valores da função de sobrevivência estimados através da função empírica para os sete primeiros e sete últimos pacientes que morreram e que haviam sido internados pelo menos uma vez. Note que optou-se por estimar curvas de sobrevivência separadamente para o grupo pacientes que haviam sido internados pelo menos uma vez e para o grupo de pacientes que nunca haviam sido internados. A Tabela 2 também apresenta o intervalo de confiança com 95% de confiança para as probabilidades de sobrevivência estimadas (as fórmulas para estimar a variância e o intervalo de confiança podem ser encontradas em Colosimo e Giolo, 2006).

A probabilidade de sobrevivência estimada entre o tempo 0 e o tempo 69, tempo em que ocorre a primeira morte, é igual a $153/153 = 1$, indicando que todos os pacientes sobreviveram até o tempo 69. Note que a probabilidade de sobrevivência estimada altera-se somente quando ocorre uma falha. Por este motivo, o valor da função de sobrevivência é constante para o intervalo de tempo entre uma falha e outra.

No período entre o tempo 69 e o tempo 101 ocorreu apenas uma morte entre os 153 pacientes participantes do estudo. Portanto, a probabilidade de sobrevivência estimada para este período, denotada por $\hat{S}(69)$, é dada por $152/153 = 0{,}9935$. Ou seja, estima-se que a probabilidade de sobreviver ao tempo 69 é de $0{,}9935$ (IC95% de $0{,}9545$ a $0{,}9991$).

Tabela 2: Tempo até a morte, número de pacientes que não falharam até o tempo t, função de sobrevivência e intervalo de confiança dos pacientes que já haviam sido internados pelo menos uma vez.

Tempo até a morte (em dias)	Nº de pacientes que não falharam até o tempo t	$\hat{S}(t_j)$	IC95%	
69	153	0,9935	0,9545	0,9991
101	152	0,9869	0,9487	0,9967
133	151	0,9804	0,9404	0,9936
150	150	0,9739	0,9318	0,9901
158	149	0,9608	0,9148	0,9822
161	147	0,9542	0,9064	0,9779
204	146	0,9477	0,8982	0,9735
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2754	7	0,0392	0,0161	0,0787
2758	6	0,0327	0,0123	0,0700
2767	5	0,0261	0,0086	0,0611
2807	4	0,0196	0,0054	0,0520
2853	3	0,0131	0,0026	0,0425
2927	2	0,0065	0,0006	0,0330
3477	1	0,0000	-	-

Para incorporar a informação de censura, Kaplan e Meier propuseram um estimador, que tem no denominador o número de pessoas sob risco em determinado período. Ou seja, este estimador estará considerado apenas aqueles pacientes que estavam vivos no instante de tempo t.

O estimador da função de sobrevivência de Kaplan-Meier é definido como:

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

onde d_j é o número de falhas no tempo t_j, e n_j é o número de indivíduos sob risco em t_j, ou seja, o número de indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j. Nesta notação, considera-se que os tempos de falha estão ordenados, ou seja, $t_1 < t_2 < \cdots < t_k$, para k tempos distintos.

Este estimador tem como característica ser uma função escada, onde os "degraus" ocorrem nos instantes de tempo t em que ocorrem as falhas: serão tantos intervalos quanto forem o número de falhas distintas. O tamanho dos degraus pode

variar: será $1/n$ se não houver empates e será $n^{\underline{o}}\ de\ empates/n$ se houverem empates. Entende-se por empate duas ou mais falhas ocorridas no mesmo instante de tempo.

A Tabela 3 apresenta os valores da função de sobrevivência estimados por Kaplan-Meier para os sete primeiros e os dez últimos pacientes, que morreram ou foram censurados, e que haviam sido internados pelo menos uma vez. De maneira equivalente ao que ocorre quando se estima a função de sobrevivência na ausência de censura, a probabilidade de sobrevivência estimada até o tempo em que ocorre a primeira falha é igual a 1.

A primeira falha ocorre no tempo 69. A estimativa da probabilidade de sobrevivência neste tempo, $\hat{S}(69)$, é calculada por $1 * \left(1 - \frac{1}{153}\right) = 0{,}9935$, onde o valor 1 é da probabilidade de sobrevivência estimada até o tempo 69 e a fração $\left(1 - \frac{1}{153}\right)$ é calculada levando em conta que houve uma morte quando haviam 153 pacientes sob risco.

O próximo tempo observado é o tempo 101. Como este tempo é de censura, a estimativa da sobrevivência não se altera, ou seja, continua sendo $0{,}9935$. Esta estimativa só ira se alterar no próximo tempo de falha que, neste caso, é o tempo 150.

A estimativa da probabilidade de sobreviver ao tempo 150 é obtida multiplicando a probabilidade de sobrevivência estimada do intervalo anterior ($0{,}9935$) por $\left(1 - \frac{1}{150}\right)$, uma vez que ocorreu uma morte neste intervalo e há 150 pacientes sob risco: os 153 pacientes que iniciaram o estudo menos um paciente que morreu antes do tempo 150 menos dois pacientes que foram censurados antes do tempo 150. Ou seja, $\hat{S}(150) = 1 * \left(1 - \frac{1}{153}\right) * \left(1 - \frac{1}{150}\right) = 0{,}9935 * \left(1 - \frac{1}{150}\right) = 0{,}9868$. Portanto, estima-se que a probabilidade de sobreviver ao tempo 150 é de $0{,}9868$ (IC95% de $0{,}9484$ a $0{,}9967$).

A estimativa da probabilidade de sobrevivência em um tempo situado entre dois tempos de falha será a igual à probabilidade de sobrevivência estimada do menor tempo de falha. Por exemplo, no tempo 2590 a probabilidade de sobrevivência estimada será $0{,}3675$. Pode-se também utilizar interpolação linear para calcular esta estimativa (Colosimo e Giolo, 2006).

Tabela 3: Tempo até a morte ou censura, variável indicadora de falha, função de sobrevivência estimada por Kaplan-Meier e intervalo de confiança para pacientes que já haviam sido internados pelo menos uma vez.

Tempo até a morte ou censura (em dias)	Variável indicadora de falha	$\hat{S}(t_j)$	IC95%	
69	1	0,9935	0,9545	0,9991
101	0	0,9935	0,9545	0,9991
133	0	0,9935	0,9545	0,9991
150	1	0,9868	0,9484	0,9967
158	0	0,9868	0,9484	0,9967
161	1	0,9801	0,9397	0,9935
204	1	0,9734	0,9307	0,9899
.		.	.	.
.		.	.	.
.		.	.	.
2585	1	0,3675	0,2496	0,4859
2667	1	0,3267	0,2018	0,4575
2670	0	0,3267	0,2018	0,4575
2754	0	0,3267	0,2018	0,4575
2758	0	0,3267	0,2018	0,4575
2767	0	0,3267	0,2018	0,4575
2807	0	0,3267	0,2018	0,4575
2853	0	0,3267	0,2018	0,4575
2927	0	0,3267	0,2018	0,4575
3477	0	0,3267	0,2018	0,4575

Ainda observando a Tabela 3, note que o último tempo de falha ocorre com 2667 dias. Portanto, o último degrau da função de sobrevivência ocorre no tempo 2667, e o valor da função de sobrevivência estimado para os tempos seguintes será igual ao valor da função de sobrevivência estimado no tempo 2667. Outra característica da função de sobrevivência estimada por Kaplan-Meier, é que ela não atinge o valor zero caso o último tempo observado seja uma censura.

Note que há diferença entre os valores da função de sobrevivência estimados na ausência de censura e na presença de censura com o mesmo banco de dados (Tabela 2 e Tabela 3). Os valores estimados da função de sobrevivência na ausência de censura decaem mais rapidamente do que os valores da função de sobrevivência quando há censura, uma vez que a censura indica que o indivíduo censurado sobreviveu até pelo menos o tempo t e, portanto, seu tempo de sobrevivência será maior do que seria se este indivíduo tivesse morrido neste mesmo tempo t.

Algumas características do estimador de Kaplan-Meier podem ser observadas graficamente. A Figura 3, que apresenta as funções de sobrevivência estimadas por Kaplan-Meier para os pacientes que haviam sido internados pelo menos uma vez e para os pacientes que nunca foram internados, é uma função escada, que apresenta degraus de tamanhos distintos e que não atinge o valor zero.

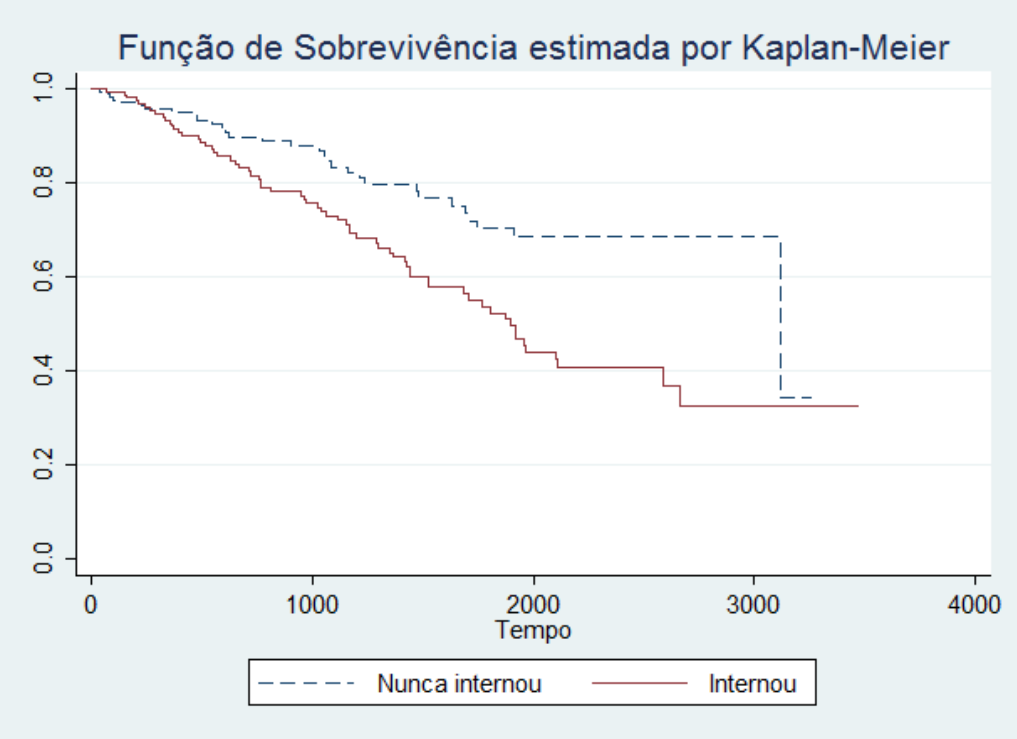

Figura 3: Funções de sobrevivência estimadas por Kaplan-Meier para pacientes que foram internados pelo menos uma vez e que nunca foram internados.

Ainda, a Figura 3 indica que apenas para os primeiros dias os pacientes que haviam sido internados pelo menos uma vez apresentaram sobrevida estimada maior que os pacientes que nunca haviam sido internados, enquanto que para o restante do período do estudo a probabilidade de sobrevivência estimada dos pacientes que nunca haviam sido internados é superior a dos pacientes que já haviam sido internados pelo menos uma vez.

Uma questão natural que surge observando este gráfico é se estas curvas são de fato diferentes. Por se tratar de curvas de sobrevida estimadas, um teste estatístico se faz necessário para verificar se a diferença observada é significativa ou pode ser atribuída ao acaso da amostragem. O teste log-rank é um dos testes mais utilizados para comparar curvas de sobrevivência de diferentes grupos. As hipóteses deste teste são:

$$\begin{cases} H_0: \text{não há } diferença \text{ entre as } curvas \\ H_1: \text{há } diferença \text{ entre as } curvas \end{cases}$$

Portanto, definindo um nível de significância de 5%, um p-valor menor que 0,05 indica que há diferença significativa entre as curvas de sobrevivência dos diferentes grupos.

Para o exemplo dos dados do HCPA, o p-valor do teste de log-rank obtido é de $0{,}0007$. Ou seja, há diferença significativa entre as curvas de sobrevivência dos pacientes que nunca haviam sido internados e dos pacientes que já haviam sido internados pelo menos uma vez ($p - valor = 0{,}0007$).

3.2. Tábua de Vida

O método da tábua de vida, também conhecido como método atuarial, é um método não-paramétrico que pode ser utilizado para estimar curvas de sobrevivência.

Há dois tipos de tábuas de vida: tábuas de vida populacionais e tábuas de vida clínicas. As tábuas de vida populacionais são as tábuas que resumem a experiência de mortalidade de uma população específica por um período específico de tempo, e as tábuas de vida clínicas são as tábuas de vida construídas para pacientes com determinada doença e que foram seguidos por um período de tempo (Lee e Wang, 2003).

Embora os cálculos para estas duas tábuas de vida sejam similares, as informações necessárias para construir cada uma das duas são diferentes. A tábua de vida populacional e a tábua de vida clínica são apresentadas na Seção 3.2.1. e Seção 3.2.3., respectivamente.

3.2.1. Tábua de Vida Populacional

A tábua de vida populacional pode ser dividida em dois tipos de tábuas de vida: as tábuas de vida por coorte e as tábuas de vida atuais. As tábuas de vida por coorte descrevem a sobrevivência ou experiência de mortalidade de um grupo de indivíduos que são acompanhados do nascimento até a morte. Para construir uma tábua de vida por coorte é preciso um tempo de acompanhamento muito longo, o que pode dificultar sua

utilização. As tábuas de vida atuais são tábuas em que se aplica uma taxa específica de mortalidade por idade a uma coorte hipotética de pessoas.

As tábuas de vida podem ainda ser tábuas de vida completas ou tábuas de vida abreviadas. Uma tábua de vida é completa quando cada linha da tábua refere-se a uma idade, ou seja, as informações da tábua de vida serão calculadas para cada idade separadamente. Neste tipo de tábua de vida a idade varia de zero anos até, por exemplo, cem anos e mais. Note que a última idade será uma faixa etária com limite superior aberto. Uma tábua de vida abreviada contém as idades agrupadas em faixas - geralmente a idade está agrupada em faixas quinquenais, com exceção dos menores de 5 anos de idade.

Segundo Bandeira (2009), é necessário fazer algumas suposições na utilização de tábuas de vida atuais:

- A coorte é fechada, ou seja, não há migrações e imigrações;
- As pessoas morrem em cada idade conforme um esquema pré-fixado que não pode ser alterado;
- A raiz da tábua é definida geralmente como 1, 100, 1.000, 10.000 ou 100.000, para facilitar comparações entre diferentes tábuas de vida;
- As mortes são igualmente distribuídas entre um aniversário e outro, ou seja, metade das mortes esperadas entre as idades de 20 e 21 anos ocorre quando a população está com 20 anos e meio. Esta suposição não se aplica aos primeiros anos de vida;
- A tábua de vida deve, preferencialmente, referir-se apenas a um sexo, porque as taxas de mortalidade de homens e mulheres possuem comportamentos diferentes.

Uma tábua de vida populacional, completa ou abreviada, possui as seguintes funções:

- ${}_nq_x$: a probabilidade de morrer entre as idades exatas x e $x+n$;
- ${}_np_x$: a probabilidade de sobreviver entre as idades exatas x e $x+n$;
- l_x: o número de pessoas vivas que tem exatos x anos de vida;
- ${}_nd_x$: o número de pessoas que morrem entre as idades exatas x e $x+n$;
- ${}_na_x$: a proporção média do tempo vivido entre as idades x e $x+n$ por aqueles que morreram neste intervalo;

- ${}_nL_x$: o número de anos vividos pelas pessoas do intervalo de idade exato x e $x + n$;
- T_x: o número total de anos vividos pelas pessoas com idade superior ou igual a idade x;
- e_x: número médio de anos que se espera que uma pessoa viva após ter x anos de idade;

A notação ${}_n$, que antecede cada uma das funções, refere-se ao tamanho do intervalo de idade utilizado. Por exemplo, suponha que se deseje conhecer a probabilidade de morrer entre 20 e 25 anos. A função da tábua de vida necessária para responder esta questão é ${}_5q_{20}$ - inicia nos 20 anos de idade e o intervalo tem amplitude de 5 anos. Quando se utiliza uma tábua de vida completa tem-se $n = 1$, e a notação ${}_n$ pode ser suprimida.

Um exemplo da construção de uma tábua de vida atual será desenvolvido para que seja possível reproduzir alguns cálculos.

Quando se constrói uma tábua de vida atual é necessário utilizar as probabilidades de morte de uma população. Ao escolher a tábua de vida que servirá como base, deve-se tomar o cuidado de escolher uma tábua de vida que tenha as taxas específicas de mortalidade e características semelhantes as da população para a qual se está construindo a tábua de vida com a coorte hipotética.

O primeiro passo, então, é encontrar uma tábua de vida de uma população semelhante a que se está estudando, e adotar suas probabilidades de morte como as probabilidades de morte desta população. O exemplo que será desenvolvido irá criar uma tábua de vida atual para as mulheres gaúchas, e irá utilizar como base as probabilidades de morte da tábua de vida das mulheres do Brasil de 2009 (Fonte: IBGE, Diretoria de Pesquisas (DPE), Coordenação de População e Indicadores Sociais (COPIS)).

A Tabela 4 apresenta as funções da tábua de vida atual para as idades de 0 a 15 para uma coorte hipotética de mulheres gaúchas.

Tabela 4: Tábua de vida de uma coorte hipotética submetida as probabilidades de morte das mulheres do Brasil de 2009

Idade	q_x	p_x	l_x	d_x	L_x	T_x	e_x
0	0,01879	0,98121	100.000	1879	98.309	7.701.156	77,01
1	0,00181	0,99819	98.121	177	98.015	7.602.847	77,48
2	0,00093	0,99907	97.944	91	97.889	7.504.832	76,62
3	0,00061	0,99939	97.853	59	97.817	7.406.943	75,69
4	0,00044	0,99956	97.793	43	97.767	7.309.126	74,74
5	0,00034	0,99966	97.750	34	97.733	7.211.359	73,77
6	0,00028	0,99972	97.716	27	97.703	7.113.626	72,80
7	0,00024	0,99976	97.689	23	97.678	7.015.923	71,82
8	0,00021	0,99979	97.666	20	97.656	6.918.245	70,84
9	0,00019	0,99981	97.646	19	97.636	6.820.590	69,85
10	0,00019	0,99981	97.627	18	97.618	6.722.953	68,86
11	0,00019	0,99981	97.609	19	97.600	6.625.335	67,88
12	0,00022	0,99978	97.590	22	97.580	6.527.736	66,89
13	0,00026	0,99974	97.569	26	97.556	6.430.156	65,90
14	0,00030	0,99970	97.543	29	97.529	6.332.600	64,92
15	0,00034	0,99966	97.514	34	97.497	6.235.072	63,94
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

A primeira coluna da Tabela 4 é referente às idades. Note que esta é uma tábua de vida completa (a tábua de vida foi apresentada de maneira resumida – apresentou-se apenas as idades de 0 a 15). A segunda coluna, q_x, foi copiada da tábua de vida das mulheres brasileiras de 2009 do IBGE. Uma observação importante sobre esta função é que deve-se fixar a probabilidade de morrer do último intervalo, que é aberto, em 1, uma vez que todas pessoas que estão vivas no início desse intervalo irão morrer durante ele (Bandeira, 2009).

A terceira coluna, p_x, é dada pelo complemento de q_x, ou seja, $p_x = 1 - q_x$. Ou, para tábuas de vida abreviadas, ${}_np_x = 1 - {}_nq_x$. Por exemplo, $p_0 = 1 - q_0 = 1 - 0{,}01879 = 0{,}98121$, e $p_1 = 1 - q_1 = 1 - 0{,}00181 = 0{,}99819$.

A quarta coluna, l_x, é calculada com base no número de pessoas vivas na idade exata $x - n$ e na probabilidade de sobreviver a idade exata $x - n$. Ou seja, $l_x = l_{x-n} *_n p_{x-n}$. Lembre-se que para tábuas de vida completas $n = 1$. Por exemplo, $l_1 = l_{1-1} * p_{1-1} = l_0 * p_0 = 100000 * 0{,}98121 = 98121$. Lembre-se também que l_0 é a raiz da tábua, valor pré-fixado.

A quinta coluna, ${}_nd_x$, é calculada subtraindo as pessoas vivas na idade exata x as pessoas vivas na idade $x+n$. Ou seja, ${}_nd_x = l_x - l_{x+n}$. Por exemplo, $d_0 = l_0 - l_1 = 100000 - 98121 = 1879$.

A sexta coluna, ${}_nL_x$, é calculada adicionando aos anos de vida dos vivos os anos de vida vividos por aqueles que morreram. Ou seja, ${}_nL_x = n(l_x - {}_nd_x + {}_na_x *_n d_x)$. Para este cálculo geralmente utiliza-se, para a maioria das idades, ${}_na_x = 0{,}5$, porque supõe-se que a morte ocorre de forma igualmente espaçada entre as idades. Porém, para as idades iniciais essa suposição não é válida, uma vez que em crianças as mortes tendem a ocorrer mais cedo. Segundo Bandeira (2009), a convenção é usar $a_0 = 0{,}3$ nos países com alta mortalidade e $a_0 = 0{,}1$ nos países com baixa mortalidade, e ${}_4a_1 = 0{,}4$.

Para este exemplo, assumiu-se que a mortalidade é baixa e, portanto, utilizou-se $a_0 = 0{,}1$. Desta forma, $L_0 = (l_0 - d_0 + a_0 * d_0) = (100000 - 1879 + 0{,}1 * 1879) = 98309$. Ainda, $a_1 = \cdots = a_4 = 0{,}4$, então $L_1 = (l_1 - d_1 + a_1 * d_1) = (98121 - 177 + 0{,}4 * 177) = 98015$. Para as demais idades tem-se que $a_i = 0{,}5$. Logo, por exemplo, $L_{10} = (l_{10} - d_{10} + a_{10} * d_{10}) = (97627 - 18 + 0{,}5 * 18) = 97618$.

Para o último intervalo o cálculo deve se feito de forma diferenciada, uma vez que a informação de l_{x+n} não existe. Segundo Bandeira (2009), uma das melhores maneiras de estimar L_{x+} é através de uma tábua de vida modelo com níveis de mortalidade parecidos. Neste exemplo, utilizou-se $L_{80} = 520836$ proveniente da tábua de vida das mulheres brasileiras de 2009.

A sétima coluna, ${}_nT_x$, é calculada acumulando a coluna ${}_nL_x$ de baixo para cima. Ou seja, equivalentemente, $T_x = T_{x+n} + {}_nL_x$. Por exemplo, $T_0 = T_1 + L_0 = 7602847 + 98309 = 7701156$.

A oitava e última coluna, e_x, é calculada pela razão entre T_x e l_x, ou seja, $e_x = T_x / l_x$. Por exemplo, $e_0 = T_0 / l_0 = 7701156/100000 = 77{,}01$ e então a expectativa de vida ao nascer é de 77,01 anos.

3.2.2. O uso de Tábuas de Vida Populacionais em Estudos de Custo-Efetividade

Em estudos de custo-efetividade, é comum a situação de não ter dados individuais de pacientes para obter estimativas de curvas de sobrevivência para grupos específicos de pacientes, mas ter informação de estudos anteriores com, por exemplo, a

razão de taxas de falhas entre o grupo de pacientes de interesse e pacientes controle. Com base nesse resultado, pode-se utilizar tábuas de vida para a população em geral (fornecidas pelo IBGE, por exemplo) para criar tábuas de vida específicas para certo grupo de pacientes. Por exemplo, se há uma tábua de vida para a população em geral e se tem uma estimativa da razão de taxa de falhas de pessoas com doença de coração em relação à população em geral e a prevalência desta doença na população, pode-se obter uma tábua de vida específica para pessoas com doenças do coração.

Segundo Gray *et al* (2011), multiplica-se a probabilidade de morte ${}_nq_x$ da tábua da população geral por θ para encontrar a probabilidade de morte da população específica. Para tanto, θ é definido como:

$$\theta = \frac{HR}{p * HR + (1 - p)}$$

onde HR é a razão taxa de falha ou *hazard ratio* da população com a doença em relação a população geral e p é a prevalência da doença na população.

Por exemplo, suponha que se deseja criar uma tábua de vida específica para mulheres gaúchas com câncer de mama. Suponha ainda, que a prevalência de câncer de mama é 10% em mulheres até 40 anos e 20% para mulheres com 40 anos ou mais, e que a razão taxa de falha entre mulheres gaúchas com câncer de mama e mulheres gaúchas controle é 4 para todas as idades. Então, para mulheres com menos de 40 anos, tem-se:

$$\theta_1 = \frac{4}{0{,}10 * 4 + (1 - 0{,}10)} = 3{,}0769,$$

e para mulheres com 40 anos ou mais tem-se:

$$\theta_2 = \frac{4}{0{,}20 * 4 + (1 - 0{,}20)} = 2{,}5.$$

A probabilidade de morte das mulheres gaúchas com menos de 40 anos com câncer de mama é obtida pela multiplicação da probabilidade de morte das mulheres gaúchas com menos de 40 anos por $\theta_1 = 3{,}0769$, ou seja, $q_x * \theta_1$. Trocando θ_1 por θ_2

obtêm-se a probabilidade de morte para as mulheres gaúchas com câncer de mama com 40 anos ou mais.

Por exemplo, da tábua de vida populacional apresentada na Tabela 4, sabe-se que a probabilidade de morte das mulheres gaúchas com 15 anos de idade é $0{,}00034$. Como a idade 15 anos é menor que 40 anos, multiplica-se esta probabilidade por θ_1, ou seja, $0{,}00034 * \theta_1 = 0{,}00034 * 3{,}0769 = 0{,}001046$, para obter a probabilidade de morte das mulheres gaúchas de 15 anos e com câncer de mama. Analogamente, para uma mulher gaúcha de 60 anos, a probabilidade de sobrevivência é $0{,}010021$ (continuação da Tabela 4 que não aparece no texto), e a probabilidade de sobrevivência para mulheres gaúchas de 60 anos com câncer de mama é dada por $0{,}010021 * \theta_2 = 0{,}010021 * 2{,}5 = 0{,}025053$.

Esta modificação no valor das probabilidades de morte causará, consequentemente, alterações nos valores das demais funções da tábua de vida.

3.2.3. Tábua de Vida Clínica

A tábua de vida clínica permite estimar as funções de sobrevivência e taxa de falha utilizando as informações de dados provenientes de estudos longitudinais com censura. Uma tábua de vida clínica possui as seguintes informações:

- $[t_i, t_{i+n})$: intervalos dos tempos do estudo;
- t_{mi}: ponto médio de cada intervalo;
- b_i: amplitude de cada intervalo;
- c_i: número de observações censuradas em cada intervalo;
- d_i: número de mortes no i-ésimo intervalo;
- n'_i: número de pessoas que entram no i-ésimo intervalo – n'_1 é o tamanho total da amostra;
- n_i: número de pessoas sob risco no i-ésimo intervalo;
- $\hat{q}_i$: estimativa da probabilidade condicional de morte no i-ésimo intervalo dado risco de morte;
- $\hat{p}_i$: estimativa da probabilidade condicional de sobrevivência no i-ésimo intervalo;
- $\hat{S}(t_i)$: estimativa da função de sobrevivência no tempo t_i;

- $\hat{h}(t_{mi})$: estimativa da função taxa de falha para o ponto médio do i-ésimo intervalo;

Para encontrar n'_i, $\hat{q}_i$, $\hat{p}_i$, $\hat{S}(t_i)$ e $\hat{h}(t_{mi})$ são utilizadas informações da tábua como n_i, c_i, d_i e b_i. Estas funções serão definidas depois do exemplo para que seja possível reproduzir alguns cálculos.

A Tabela 5 apresenta as estimativas para a função de sobrevivência pelo método da tábua de vida clínica para os pacientes do ambulatório de insuficiência cardíaca do HCPA que haviam sido internados pelo menos uma vez, para os sete primeiros e sete últimos intervalos de tempo. A amplitude dos intervalos de tempo foi escolhida como 100 dias.

Tabela 5: Função de sobrevivência para pacientes que já foram internados pelo menos uma vez pelo método da tábua de vida clínica.

Intervalo de tempo até a morte (em dias)	n'_i	Mortes (d_i)	Censura (c_i)	n_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i+)$
[0,100)	153	1	0	153,0	0,007	0,993	0,993
[100,200)	152	2	4	150,0	0,013	0,987	0,980
[200,300)	146	5	1	145,5	0,034	0,966	0,946
[300,400)	140	6	1	139,5	0,043	0,957	0,905
[400,500)	133	3	6	130,0	0,023	0,977	0,884
[500,600)	124	4	7	120,5	0,033	0,967	0,855
[600,700)	113	3	9	108,5	0,028	0,972	0,831
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
[2600,2700)	9	1	1	8,5	0,118	0,882	0,325
[2700,2800)	7	0	3	5,5	0,000	1,000	0,325
[2800,2900)	4	0	2	3,0	0,000	1,000	0,325
[2900,3000)	2	0	1	1,5	0,000	1,000	0,325
[3400,3500)	1	0	1	0,5	0,000	1,000	0,325
[2900,3000)	2	0	1	1,5	0,000	1,000	0,325
[3400,3500)	1	0	1	0,5	0,000	1,000	0,325

Para calcular o número de pessoas que entram no i-ésimo intervalo, n'_i, para $i \neq 1$, tem-se que: $n'_i = n'_{i-1} - c_{i-1} - d_{i-1}$. Por exemplo, $n'_2 = n'_1 - c_1 - d_1 = 153 - 0 - 1 = 152$.

Esta informação é utilizada para calcular o número de pessoas sob risco no i-ésimo intervalo, n_i, uma vez que $n_i = n'_i - \frac{1}{2}c_i$. Por exemplo, $n_1 = n'_1 - \frac{1}{2}c_1 = 153 - \frac{1}{2}0 = 153$ e $n_2 = n'_2 - \frac{1}{2}c_2 = 152 - \frac{1}{2}4 = 152 - 2 = 150$.

A estimativa da probabilidade de morte no i-ésimo intervalo é obtida por: $\hat{q}_i = \frac{d_i}{n_i}$. Para o último intervalo da tábua de vida esta probabilidade será 1 (Lee e Wang). A probabilidade de sobrevivência estimada, por sua vez, é dada por: $\hat{p}_i = 1 - \hat{q}_i$. Por exemplo, $\hat{q}_1 = \frac{d_1}{n_1} = \frac{1}{153} = 0{,}007$ e $\hat{p}_1 = 1 - \hat{q}_1 = 1 - 0{,}007 = 0{,}993$, e $\hat{q}_2 = \frac{d_2}{n_2} = \frac{2}{150} = 0{,}013$ e $\hat{p}_2 = 1 - \hat{q}_2 = 1 - 0{,}013 = 0{,}987$.

A função de sobrevivência estimada é definida por $\hat{S}(t_i) = \hat{p}_{i-1}\hat{S}(t_{i-1})$. Por definição, $\hat{S}(t_1) = 1$. Então, por exemplo, $\hat{S}(t_2) = \hat{p}_1\hat{S}(t_1) = 0{,}993 * 1 = 0{,}993 = \hat{S}(t_1 +)$ e $\hat{S}(t_3) = \hat{p}_2\hat{S}(t_2) = 0{,}987 * 0{,}993 = 0{,}980 = \hat{S}(t_2 +)$.

A função taxa de falha pode ser estimada por $\hat{h}(t_{mi}) = \frac{2\hat{q}_i}{b_i(1+\hat{p}_i)}$. Por exemplo, $\hat{h}(t_{50}) = \frac{2\hat{q}_1}{b_1(1+\hat{p}_1)} = \frac{2*0{,}007}{100(1+0{,}993)} = 0{,}00007$.

A tábua de vida clínica não é uma técnica de análise de sobrevivência muito utilizada. Pode-se considerar a tábua de vida clínica como um resumo dos resultados estimados por Kaplan-Meier, uma vez que o estimador Kaplan-Meier calcula as estimavas para cada falha e a tábua de vida clínica calcula as estimativas para intervalos de tempo definidos.

Como já foi dito anteriormente, o estimador Kaplan-Meier apresenta superioridade sobre os demais estimadores não-paramétricos. Portanto, sugere-se que, quando forem utilizados modelos não-paramétricos, priorize-se o uso do estimador Kaplan-Meier.

4. Modelos Paramétricos

Os modelos paramétricos são assim definidos porque são baseados na suposição de uma distribuição de probabilidade para o tempo de sobrevivência. Devido a esta suposição, estes modelos são também chamados de modelos probabilísticos.

A grande vantagem da utilização de modelos paramétricos é a possibilidade de fazer extrapolações das curvas de sobrevivência. Em estudos de custo-efetividade, muitas vezes é necessário extrapolar curvas de sobrevida (Latimer, 2011), visto que os dados disponíveis para estimar as curvas de sobrevivência são oriundos de estudos cujo tempo de acompanhamento é, em geral, menor que aquele desejado para os estudos de custo-efetividade. É de extrema importância que a distribuição de probabilidade utilizada seja adequada aos dados, porque a extrapolação será feita com base nesta suposição. Se o modelo paramétrico escolhido não é adequado aos dados, a extrapolação pode gerar estimativas incorretas.

As distribuições de probabilidades mais utilizadas para fazer análise de sobrevivência com modelos paramétricos são: exponencial, Weibull, log-normal e Gompertz. Estes modelos serão abordados com mais detalhes nas Seções 4.1., 4.2., 4.3. e 4.4.. Note que essas funções são assimétricas, diferente da distribuição normal, geralmente utilizada como base para técnicas estatísticas inferenciais. Ainda, a Seção 4.5. apresenta algumas ferramentas que podem ser utilizadas na escolha do modelo paramétrico que melhor se ajusta aos dados.

4.1. Modelo Exponencial

O modelo exponencial é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidades exponencial. Este modelo paramétrico é apontado como o modelo mais simples em termos matemáticos, e é também considerado o modelo paramétrico mais importante. Lee e Wang (2003) comparam a sua importância na análise de sobrevivência à importância de uma distribuição normal nas diversas análises da área da estatística.

A função densidade de probabilidade para a variável aleatória tempo de falha T é dada pela expressão:

$$f(t) = \lambda \exp\{-\lambda t\}, \quad t \geq 0 \text{ e } \lambda > 0.$$

O modelo exponencial apresenta apenas um parâmetro, λ. Este parâmetro representa o inverso do tempo médio de sobrevivência, ou seja, o tempo médio de sobrevivência é obtido por $1/\lambda$.

A função de sobrevivência do modelo exponencial é dada por:

$$S(t) = \exp\{-\lambda t\}, \quad t \geq 0 \text{ e } \lambda > 0,$$

e a função taxa de falha por:

$$h(t) = \lambda, \quad t \geq 0 \text{ e } \lambda > 0.$$

Uma característica marcante do modelo exponencial é que ele possui a função taxa de falha constante ao longo do tempo, ou seja, o risco de falha é sempre o mesmo para qualquer tempo t. O valor da função taxa de falha é igual ao valor do parâmetro da distribuição, como pode ser visto na expressão acima. Essa propriedade é conhecida como falta de memória da distribuição exponencial, e pode ser vista na Figura 4.

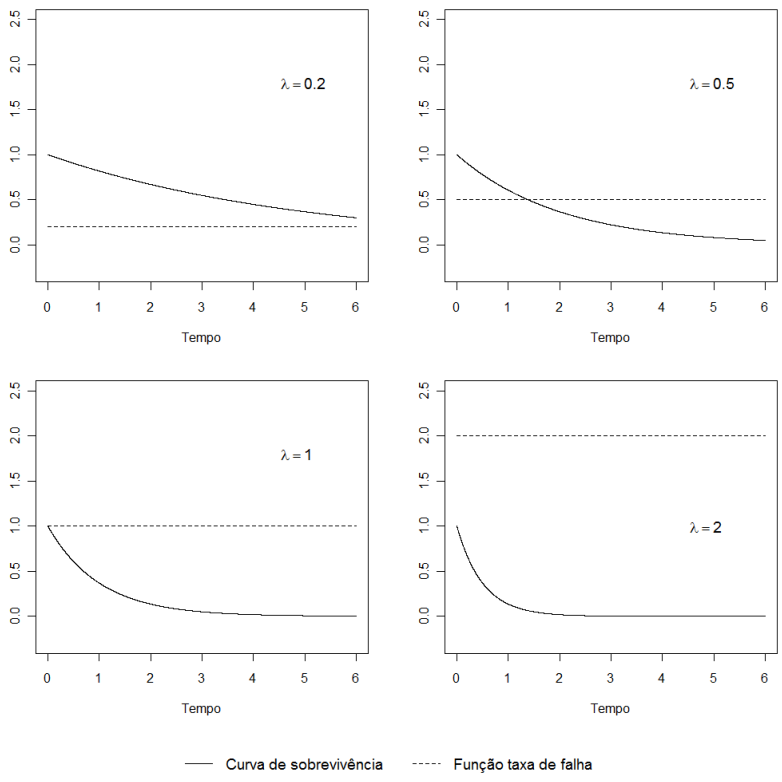

Figura 4: Função sobrevivência e função taxa de falha da do modelo exponencial para diferentes valores de λ.

Na Figura 4 é possível ver a influência dos valores do parâmetro λ na função de sobrevivência e na função taxa de falha. Note que para valores grandes de λ o risco será alto e a sobrevida baixa, e para valores pequenos de λ o risco será baixo e a sobrevida alta (Lee e Wang, 2003).

Por exemplo, considere que um estudo está sendo realizado para investigar o tempo até a morte de pacientes com determinada doença. Se o modelo exponencial for adequado para analisar esses dados, sabe-se, automaticamente, que o risco de morte para pacientes que estão com a doença há pouco ou há muito tempo, dado que estes ainda não tenham morrido, é o mesmo.

4.2. Modelo Weibull

O modelo Weibull é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidades Weibull. Este modelo paramétrico tem se mostrado bastante útil porque ele apresenta uma grande variedade de formas e, por isso, consegue se adaptar a várias situações práticas.

A função densidade de probabilidade para a variável aleatória tempo de falha T é dada pela expressão:

$$f(t) = \lambda p t^{p-1} \exp\{-\lambda t^p\}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

A função de sobrevivência do modelo Weibull é dada por:

$$S(t) = \exp\{-\lambda t^p\}, \quad t \geq 0 \text{ e } p,\ \lambda > 0,$$

e a função taxa de falha por:

$$h(t) = \lambda p t^{p-1}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Note que o modelo exponencial é um caso particular do modelo Weibull, quando $p = 1$.

A função taxa de falha da distribuição Weibull tem como propriedade ser uma função monótona, ou seja, é crescente, decrescente ou constante. Ela será crescente

quando $p > 1$, constante quando $p = 1$ e decrescente quando $p < 1$, como pode ser visto na Figura 5.

Na função de sobrevivência, o parâmetro λ influencia na rapidez com que a curva decresce: valores altos para este parâmetro fazem a curva de sobrevivência decair mais rapidamente do que valores baixos. Note também que à medida que p aumenta o decaimento da curva de sobrevivência ocorre mais rapidamente, indicando que os parâmetros p e λ influenciam conjuntamente a forma da curva de sobrevivência.

Para a função taxa de falha, no caso onde $p \neq 1$, o parâmetro λ influencia na rapidez com que a função taxa de falha cresce ou decresce: o crescimento será mais rápido ($p > 1$) ou o decrescimento será mais lento ($p < 1$), para valores maiores do parâmetro λ. Para $p = 1$, a função taxa de falha é constante com taxa de falha maior para valores maiores de λ.

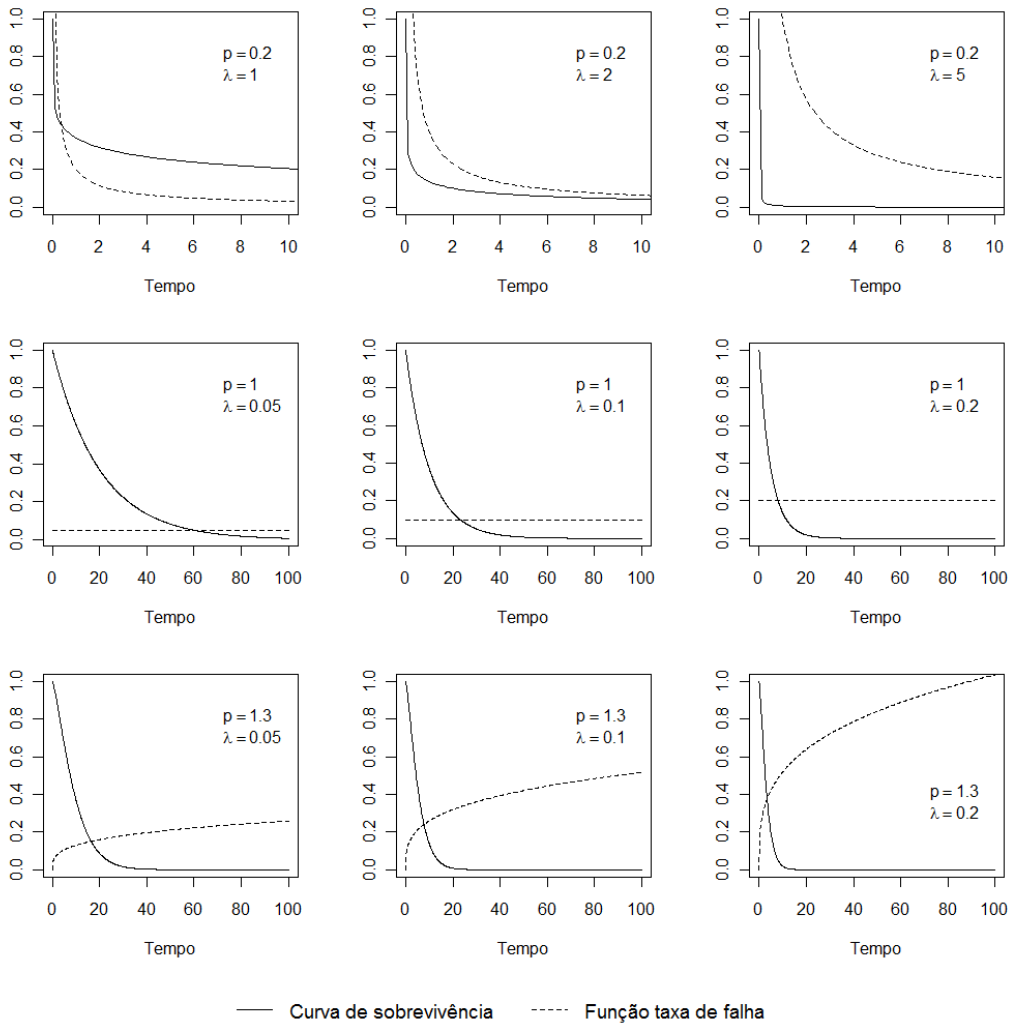

Figura 5: Função sobrevivência e função taxa de falha do modelo Weibull para diferentes valores de p e λ.

Há situações em que é conveniente utilizar o logaritmo do tempo de falha T. Quando o tempo de falha segue distribuição Weibull, o logaritmo do tempo de falha segue uma distribuição de Gambel, também chamada de distribuição do valor extremo. Ou seja, se T segue distribuição Weibull, $Y = \log(T)$ segue distribuição do valor extremo.

A função densidade de probabilidade, a função de sobrevivência e a função taxa de falha de Y são dadas por:

$$f(y) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}, \qquad y, \mu \in \Re, \sigma > 0$$

$$S(y) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}, \qquad y, \mu \in \Re, \sigma > 0$$

e

$$\lambda(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}, \qquad y, \mu \in \Re, \sigma > 0.$$

A relação entre as distribuições Weibull e Gambel também pode ser expressa por uma relação entre os parâmetros dessas duas distribuições. Para tanto, tem-se que $p = 1/\sigma$ e $\lambda = \left(1/\exp\{\mu\}\right)^p$.

4.3. Modelo Log-normal

O modelo log-normal é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidades log-normal. Uma característica interessante desta distribuição é que o logaritmo de uma variável com distribuição log-normal, com parâmetros μ e σ, tem distribuição normal, com média μ e desvio-padrão σ.

A função densidade de probabilidade para a variável aleatória tempo de falha T é dada pela expressão:

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

onde μ e σ representam, respectivamente, a média e o desvio-padrão do logaritmo do tempo de falha.

A função de sobrevivência do modelo log-normal é dada por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right), \qquad t > 0, \mu \in \Re, \sigma > 0,$$

e a função taxa de falha por:

$$h(t) = \frac{f(t)}{S(t)}, \qquad t > 0, \mu \in \Re, \sigma > 0,$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada de uma distribuição normal padrão. Note que essas funções não apresentam forma analítica explícita.

A característica da função taxa de falha do modelo paramétrico log-normal é ser inicialmente crescente e, quando atingir o ponto de máximo passar a decrescer, como pode ser observado na Figura 6.

A influência dos valores dos parâmetros μ e σ podem também ser entendidas a partir da Figura 6. Valores grandes para μ produzem curvas de sobrevivência com sobrevida maior que valores pequenos para μ. Ou seja, quanto maior for o parâmetro μ, maior será o tempo de sobrevida dos pacientes. Consequentemente, quanto menor for o valor de μ, mais acentuado será o pico da função taxa de falha. O parâmetro σ influencia na variabilidade das curvas, ou seja, curvas de sobrevivência que tem valores mais altos para σ terão probabilidade de sobrevivência maior para tempos maiores do que curvas com valores baixos para σ. Uma grande variabilidade acarreta em uma função taxa de falha menor do que seria se a variabilidade fosse pequena.

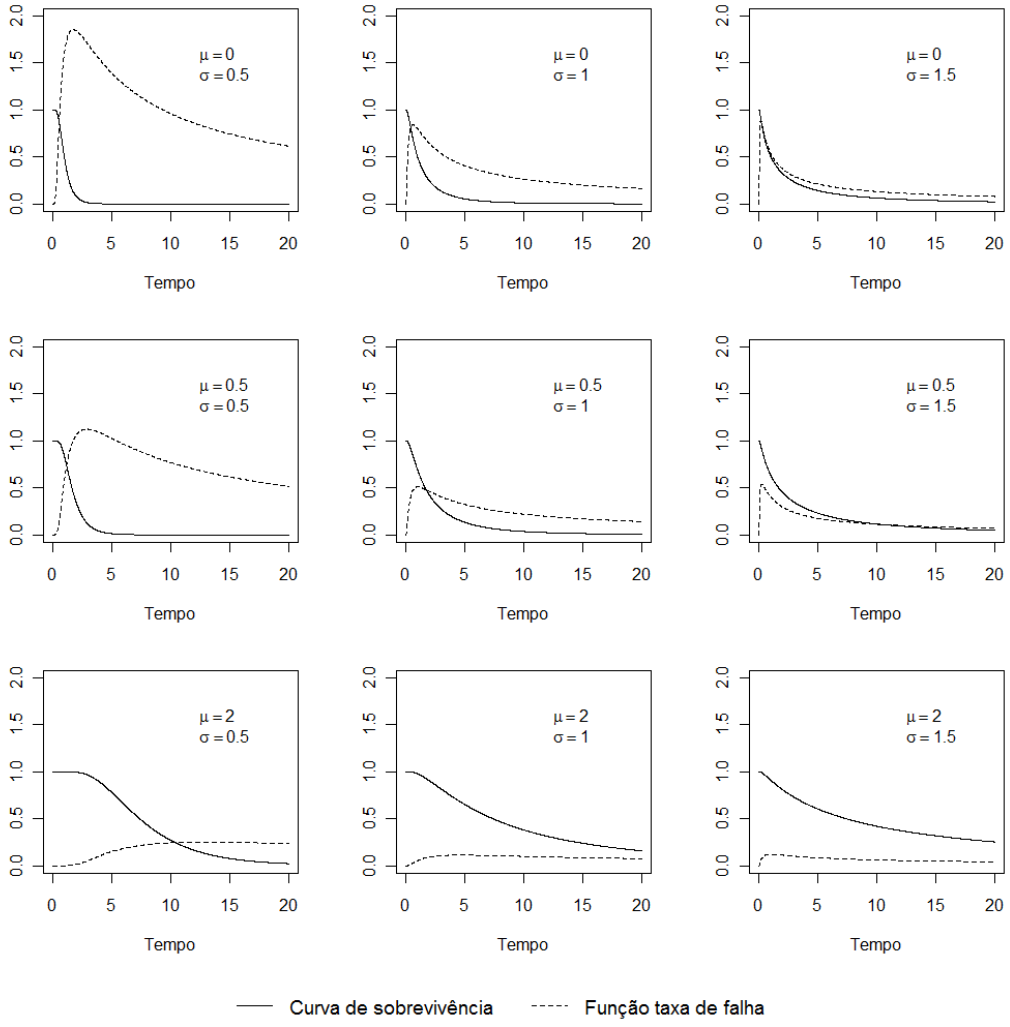

Figura 6: Função sobrevivência e função taxa de falha do modelo log-normal para diferentes valores de μ e σ.

4.4. Modelo Gompertz

O modelo Gompertz é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidades Gompertz. Uma característica interessante do modelo paramétrico Gompertz é que sua função taxa de falha é igual à função taxa de falha do modelo exponencial, quando $\alpha = 0$ (Cleves *et al*, 2008; Wienke, 2009).

A função densidade de probabilidade para a variável aleatória tempo de falha T é dada pela expressão:

$$f(t) = \theta \exp[\alpha t] \exp\left[\frac{\theta}{\alpha}(1 - \exp(\alpha t))\right], \qquad t > 0, \theta > 0, \alpha \in \Re.$$

A função de sobrevivência do modelo Gompertz é dada por:

$$S(t) = \exp\left[\frac{\theta}{\alpha}(1 - \exp(\alpha t))\right], \qquad t > 0, \theta > 0, \alpha \in \Re$$

e função taxa de falha por:

$$h(t) = \theta \exp[\alpha t], \qquad t > 0, \theta > 0, \alpha \in \Re.$$

A função taxa de falha do modelo Gompertz é uma função monótona. Ela é crescente para valores positivos de α e é decrescente para valores negativos de α. Se $\alpha = 0$, a função taxa de falha é constante. Estas propriedades do modelo Gompertz podem ser verificadas na Figura 7.

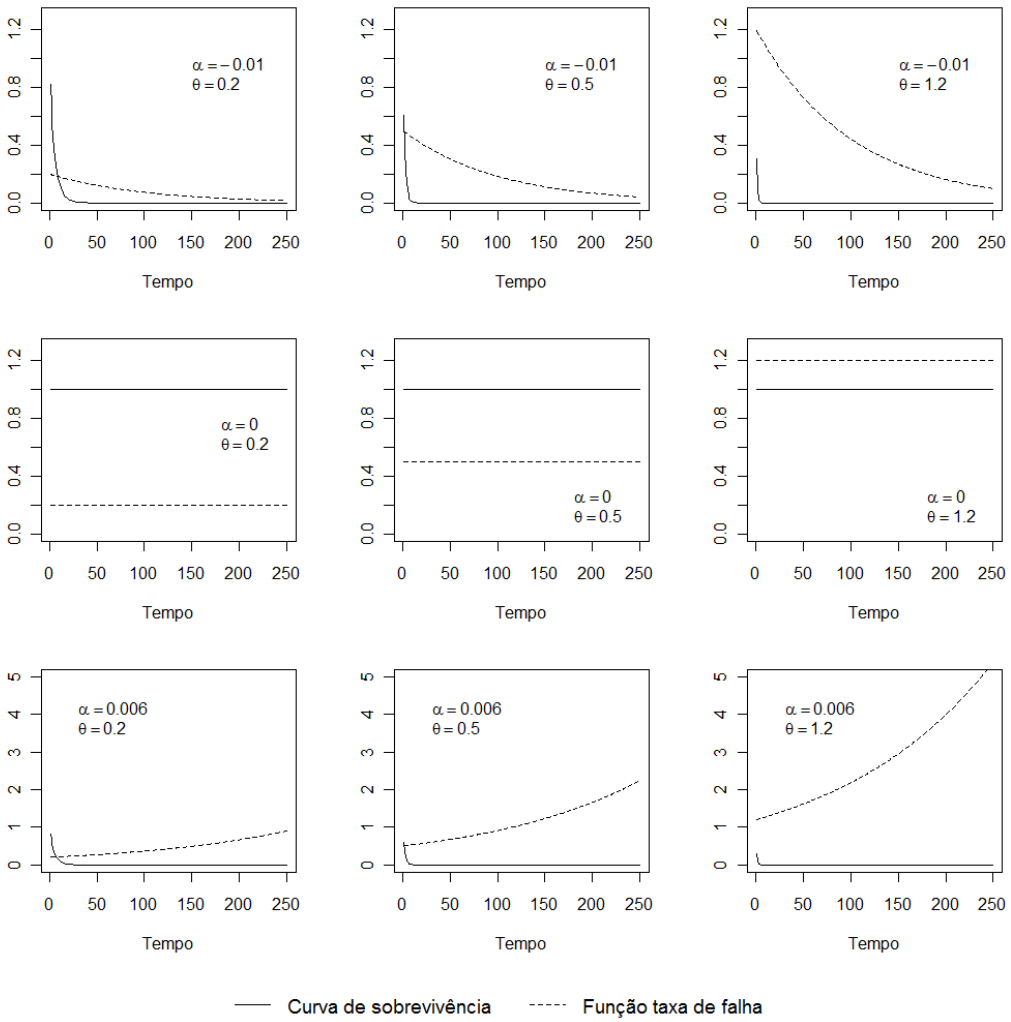

Figura 7: Função sobrevivência e função taxa de falha do modelo Gompertz para diferentes valores de α e θ.

Através da Figura 7 também é possível entender a influência do parâmetro θ sobre a curva de sobrevivência e a função taxa de falha. Quanto maior for o valor de θ maior será o valor da função taxa de falha ($\alpha = 0$), ou maior será o crescimento ($\alpha > 0$) ou decrescimento ($\alpha < 0$) da função taxa de falha. Para a curva de sobrevivência, valores grandes de θ causam um decrescimento acelerado da função em relação a valores pequenos de θ. Note também que a função taxa de falha cresce ou decresce exponencialmente com o tempo (Lee e Wang, 2003).

4.5. Escolha do Modelo Paramétrico

A escolha do modelo paramétrico adequado para analisar um conjunto de dados de sobrevivência é uma etapa muito importante da análise. Se o modelo utilizado não é adequado aos dados os resultados podem não retratar corretamente a realidade, invalidando toda análise realizada.

Existem métodos para auxiliar na escolha do modelo paramétrico que melhor se ajusta aos dados, ou seja, indicam qual distribuição probabilística se adequa melhor as características específicas dos dados. Entre estes métodos estão alguns métodos gráficos, o teste de modelos encaixados e os critérios de informação.

Os métodos gráficos tendem a ser mais fáceis de construir e compreender, porém há subjetividade na decisão. Estes métodos servem como ferramenta para indicar que algum modelo paramétrico não é adequado aos dados, e não para indicar qual é o modelo que melhor se ajusta aos dados. Ou seja, servem para excluir a possibilidade de utilização de um modelo paramétrico proposto nas situações em que os gráficos gerados por este modelo não são adequados.

O teste de modelos encaixados tem a decisão baseada em um nível de significância definido, o que exclui a subjetividade da decisão. Contudo, este teste não pode ser calculado para o modelo paramétrico Gompertz, como será discutido a seguir. Neste caso, não é possível escolher o modelo paramétrico que melhor se ajusta aos dados somente com este teste, porque neste caso não estamos testando o ajuste do modelo Gompertz. Então, este teste também irá servir para indicar que algum modelo pode não ser adequado aos dados, e não para decidir qual modelo deve ser utilizado.

Os critérios de informação obedecem a uma regra de decisão, não havendo, portanto, subjetividade na decisão. Estes critérios indicam o modelo que melhor se ajusta aos dados. Contudo, não se deve utilizar somente este método, porque ele pode

indicar que determinado modelo gera o melhor ajuste aos dados, enquanto que as demais ferramentas indicam que este modelo não é adequado. Neste caso, os critérios de informação levariam a uma decisão equivocada.

4.5.1. Métodos Gráficos

Nesta seção serão apresentados dois métodos gráficos, que são discutidos por Colosimo e Giolo (2006). O primeiro método baseia-se essencialmente na comparação da função de sobrevivência estimada pelo modelo paramétrico proposto com a curva de sobrevivência estimada pelo método de Kaplan-Meier. O segundo método consiste em linearizar a função de sobrevivência do modelo proposto.

Para construir o primeiro método gráfico é necessário ajustar os modelos paramétricos desejados e, então, fazer o gráfico da curva de sobrevivência estimada pelo modelo proposto com a curva de sobrevivência estimada pelo método de Kaplan-Meier. Note que no eixo x estará o tempo t. O que deve ser observado nestes gráficos é a proximidade das funções de sobrevivência estimadas pelos modelos paramétricos e Kaplan-Meier. Os modelos paramétricos cujas curvas de sobrevivência estimadas se aproximarem da curva de Kaplan-Meier apresentam um bom ajuste aos dados.

Uma variação deste método é utilizar a função taxa de falha acumulada estimada ao invés da função de sobrevivência estimada. Novamente, os modelos paramétricos que possuírem função taxa de falha acumulada estimada próxima da função taxa de falha acumulada estimada de Kaplan-Meier apresentam um bom ajuste aos dados.

Outra variação deste método é fazer o gráfico da função de sobrevivência estimada por Kaplan-Meier *versus* a função de sobrevivência estimada pelo modelo paramétrico proposto. Note que a função de sobrevivência estimada por Kaplan-Meier estará no eixo x e a função de sobrevivência estimada pelo modelo proposto no eixo y. Portanto, comparam-se os pontos gerados por este gráfico com a reta $x = y$, onde pontos retas próximas indicam bom ajuste do modelo paramétrico aos dados. Falhas na linearidade indicam que o modelo não é adequado.

O segundo método gráfico baseia-se na ideia de linearização da função de sobrevivência. Desta forma, se o modelo paramétrico proposto for adequado o resultado será uma reta. Violações na linearidade são facilmente identificadas, indicando que o modelo paramétrico proposto não é adequado.

No modelo exponencial $-\log(S(t))$ é função linear de t. Então, se o modelo exponencial for adequado, o gráfico $-\log\left(\hat{S}(t)\right)$ *versus* t, onde $\hat{S}(t)$ é a curva de sobrevivência estimada por Kaplan-Meier, será aproximadamente linear.

Para o modelo de Weibull, a relação entre $\log(-\log(S(t)))$ e $\log(t)$ é linear. Portanto, se o modelo Weibull for adequado, o gráfico $\log\left(-\log\left(\hat{S}(t)\right)\right)$ *versus* $\log(t)$, onde $\hat{S}(t)$ é a curva de sobrevivência estimada por Kaplan-Meier, será aproximadamente linear. Se este gráfico passar pela origem e sua inclinação for 1, há indícios de que o modelo exponencial é o mais adequado.

No modelo log-normal sabe-se que $\Phi^{-1}(S(t))$ é linear a $\log(t)$. Neste caso, o gráfico de $\Phi^{-1}(\hat{S}(t))$ *versus* $\log(t)$, onde $\hat{S}(t)$ é a curva de sobrevivência estimada por Kaplan-Meier, será aproximadamente linear quando o modelo log-normal for adequado.

Para o modelo Gompertz esta técnica não pode ser utilizada porque não há relação linear entre a função de sobrevivência e o tempo t.

Segundo Colosimo e Giolo (2006), há situações em que os métodos gráficos não apontam grandes diferenças entre os modelos paramétricos. Isso pode acontecer se o tamanho de amostra for pequeno ou se há poucas falhas. Neste caso, os modelos tendem a gerar resultados parecidos, podendo apresentar uma pequena diferença nas caudas das distribuições. Os métodos gráficos também podem indicar que nenhum modelo paramétrico é adequado. Neste caso, técnicas não-paramétricas e modelos paramétricos mais flexíveis podem ser uma solução.

4.5.2. Comparando Modelos Encaixados

Os modelos exponencial, Weibull e log-normal são casos especiais de um modelo geral chamado de modelo gama generalizado, ou seja, os modelos exponencial, Weibull e log-normal são modelos encaixados do modelo gama generalizado. Detalhes sobre o modelo gama generalizado podem ser encontrados em Colosimo e Giolo (2006) e Collett (2003).

As hipóteses do teste de modelos encaixados são:

$$\begin{cases} H_0: o\ modelo\ \acute{e}\ adequado \\ H_1: o\ modelo\ n\tilde{a}o\ \acute{e}\ adequado \end{cases}$$

O teste dos modelos encaixados é baseado na verossimilhança do modelo paramétrico proposto e na verossimilhança do modelo gama generalizado, comparando o ajuste aos dados do modelo paramétrico proposto com o do modelo gama generalizado. Se o ajuste dos dois modelos for parecido, então os valores das verossimilhanças serão próximos e o modelo paramétrico proposto será adequado.

O teste é calculado através de duas vezes a diferença entre o logaritmo da verossimilhança do modelo gama generalizado, que chamaremos de A, e o logaritmo da verossimilhança do modelo proposto, que chamaremos de B. Ou seja, $2(A - B)$. Esta estatística tem distribuição qui-quadrado, onde os graus de liberdade são dados pela diferença entre o número de parâmetros da distribuição gama generalizada e do modelo paramétrico proposto.

O modelo proposto não será adequado quando rejeita-se $H_0: o\ modelo$ é $adequado$, ou seja, quando o p-valor for menor que o nível de significância estabelecido, geralmente 5%. Se $p - valor > 0{,}05$, então não há evidências estatísticas de que o modelo proposto não seja adequado.

Note que o modelo Gompertz não é um modelo encaixado do modelo gama generalizado e, portanto, o teste de modelos encaixados não pode ser utilizado para este modelo paramétrico. Portanto, não há como utilizar somente este teste para obter o modelo paramétrico que melhor se ajusta aos dados, uma vez que neste caso não se estaria considerando o ajuste do modelo Gompertz.

4.5.3. Critérios de Informação

Os critérios de informação são outra forma de avaliar o ajuste dos modelos. Com os critérios de informação como o *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC) é possível comparar diferentes modelos com diferentes números de parâmetros.

O AIC e o BIC são critérios que penalizam a verossimilhança para que um modelo mais parcimonioso seja selecionado. Quanto menor for o valor do AIC e do BIC melhor é o ajuste do modelo.

4.6. Exemplo

Nesta seção os dados apresentados na Seção 2.3. são utilizados para desenvolver um exemplo passo a passo, visando auxiliar o entendimento da modelagem paramétrica em análise de sobrevivência. Esta é, provavelmente, uma das melhores maneiras de compreender e fixar as técnicas descritas no Capítulo 4.

O *software* STATA versão 11 é utilizado para fazer as análises. À medida que o exemplo é desenvolvido, alguns comandos e saídas desse *software* são disponibilizados no texto.

4.6.1. Preparação dos Dados no STATA

O primeiro passo, ainda antes de começar as análises, é salvar o banco de dados no formato de banco de dados do STATA, com terminação .dta. O banco de dados deste exemplo é originalmente um arquivo do Excel, ou seja, com terminação .xls. Neste caso, a maneira mais fácil encontrada de fazer este processo é abrir o banco de dados em Excel, selecionar os dados e copiá-los para colar no STATA.

Para colar os dados no STATA, com o *software* aberto, deve-se ir no Menu *Data*, clicar em *Data Editor* e clicar em *Data Editor (Edit)*. Assim irá abrir uma planilha onde deve-se colar os dados, com crtl+v, por exemplo. Quando isto for feito, o *software* abre uma janela onde você deve indicar se o banco de dados que você está colando possui os nomes das variáveis ou não. Para salvar o banco de dados como banco de dados do STATA, deve-se ir no Menu *File*, clicar em *Save As...* e escolher o local que o banco de dados será salvo, bem como o nome que será dado para o arquivo. O nome dado para este banco de dados é "exemplo".

Para abrir este banco de dados no STATA, por programação, pode-se utilizar o comando:

```
use "D:\Monografia\Exemplo Cap. 4\exemplo.dta", clear
```

onde o endereço entre as aspas é o endereço em que foi salvo o banco de dados `exemplo.dta`. O `clear` faz parte do comando para que seja fechado qualquer outro banco que ocasionalmente estiver aberto.

Para trabalhar com análise de sobrevivência, deve-se declarar os dados como dados de análise de sobrevivência. Para isso utiliza-se o comando `stset`:

```
stset  follow, failure(obitot)
```

onde `follow` é a variável de tempo de falha e `obitot` é a variável indicadora de censura, variáveis descritas na Seção 2.3.. Este comando assume, para a variável `obitot`, que o valor 1 indica falha e o valor 0 indica censura. Como resposta, este comando irá apresentar um pequeno resumo do banco de dados, como total de observações, número de observações que não serão utilizadas por causa da restrição imposta, entre outros, como pode ser visto abaixo:

```
          failure event:  obitot != 0 & obitot < .
obs. time interval:  (0, follow]
 exit on or before:  failure

--------------------------------------------------------------------------
      318  total obs.
        0  exclusions
--------------------------------------------------------------------------
      318  obs. remaining, representing
       96  failures in single record/single failure data
   379953  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      3477
```

É importante observar esta saída do *software* para confirmar se os dados foram declarados corretamente. Por exemplo, ela nos fornece a informação de que há 318 observações e que nenhuma observação foi excluída.

4.6.2. Estimativas por Kaplan-Meier

Antes de iniciar a modelagem, lembre-se que estes dados já haviam sido utilizados anteriormente no Capítulo 3, para explicar o método de estimação de Kaplan-Meier. A Figura 3, naquele capítulo, indicava que o comportamento das curvas de sobrevivência de pacientes que nunca foram internados e de pacientes que já haviam sido internados pelo menos uma vez, não é igual.

A programação que cria aquele gráfico no STATA, utilizando o comando `sts`, é:

```
sts graph, by(inter) ylabels(0 .2 0.4 0.6 0.8 1, format(%9.1f)) title
(Função de Sobrevivência estimada por Kaplan-Meier) xtitle(Tempo)
legend(label(1 Nunca internou) label(2 Internou))
```

O comando `graph` indica que deseja-se fazer o gráfico da função de sobrevivência estimada. Um comando muito importante é o comando `by(inter)`, que faz com que as funções de sobrevivência sejam estimadas para cada nível da variável `inter`, ou seja, para pacientes que nunca haviam sido internados e para pacientes que já haviam sido internados pelo menos uma vez. Os demais comandos que vem depois da vírgula são opcionais e são referentes à formatação do gráfico, como título do gráfico (`title`), subtítulo do gráfico (`subtitle`) e nome do eixo x (`xtitle`). O comando menos intuitivo é o das escalas do eixo y (`ylabels`), que primeiro declara os pontos que devem aparecer no gráfico, e depois declara que estes pontos vão estar justificados a direita e irão ter 1 casa decimal (`format(%9.1f)`). Embora não sejam obrigatórios, é interessante utilizar estes comandos opcionais, porque assim o gráfico fica mais completo.

Esta programação cria um gráfico muito próximo ao da Figura 3. Porém, dessa forma as duas linhas do gráfico serão de cores diferentes, porém ambas contínuas, o que pode gerar alguma confusão. Como não se conseguiu alterar a forma das linhas via programação, este ajuste foi feito via Menu. Para fazer esta alteração, deve-se clicar com o botão direito do *mouse* em cima do gráfico e clicar na opção *Start Graph Editor*. Para melhor diferenciar as duas curvas, a curva de sobrevivência estimada por Kaplan-Meier para os pacientes que nunca foram internados será uma linha pontilhada. Com um duplo-clique sobre a linha da função de sobrevivência estimada por Kaplan-Meier para os pacientes que nunca foram internados, abre-se uma janela *Line proprieties*, onde, na opção *Pattern* troca-se de *Solid* para *Dash*, seguido do *ok*.

Ainda no Capítulo 3, a hipótese da diferença entre as funções de sobrevivência dos pacientes que nunca haviam sido internados e dos pacientes que haviam sido internados pelo menos uma vez foi testada pelo teste log-rank.

A programação utilizada para realizar o teste log-rank no STATA é:

```
sts test inter
```

onde a resposta fornecida pelo *software* é:

```
Log-rank test for equality of survivor functions

      |   Events         Events
inter |  observed       expected
------+-------------------------
0     |        31          47.65
1     |        65          48.35
------+-------------------------
Total |        96          96.00

           chi2(1) =      11.58
           Pr>chi2 =     0.0007
```

Ou seja, há diferença significativa entre as curvas de sobrevivência dos pacientes que nunca haviam sido internados e dos pacientes que já haviam sido internados pelo menos uma vez ($p - valor = 0{,}0007$). Esta diferença entre as curvas de sobrevivência indica que os dois grupos de pacientes devem ser analisados separadamente, ou esta variável deve entrar na análise como covariável.

Os modelos paramétricos abordados no Capítulo 4 não consideram a possibilidade da inclusão de covariáveis e, portanto, será feito uma estratificação dos dados. O estudo contou com 318 pacientes, onde 153 já haviam sido internados pelo menos uma vez e 165 nunca haviam sido internados. Como o tamanho da amostra de cada grupo não ficará pequeno, não há problemas em fazer estratificação.

A modelagem deve ser feita para os dois grupos de pacientes separadamente e, portanto, o modelo paramétrico adequado para descrever a sobrevivência de um grupo pode ser diferente do modelo paramétrico adequado para descrever a sobrevivência do outro grupo. Neste trabalho será feita apenas a modelagem passo a passo para os pacientes que já foram internados pelo menos uma vez. Para os pacientes que nunca foram internados, apenas será indicado qual modelo paramétrico é o mais adequado.

Como a modelagem será feita para os dois grupos de pacientes separadamente é preciso, novamente, declarar os dados como dados de sobrevivência, agora separando em grupos. A programação utilizada para fazer isso no STATA é:

```
stset  follow if inter==1, failure(obitot)
```

O comando condicional é utilizado para que apenas os pacientes que já haviam sido internados pelo menos uma vez (`if inter==1`) entrem na análise.

A função de sobrevivência de Kaplan-Meier deve ser estimada porque esta curva é considerada uma análise descritiva da análise de sobrevivência. A programação utilizada para fazer este gráfico no STATA é:

```
sts graph, ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f)) title (Curva de
Sobrevivência estimada por Kaplan-Meier) subtitle (Pacientes que já
foram internados) xtitle(Tempo)
```

A curva de sobrevivência estimada por Kaplan-Meier para os pacientes que já foram internados pelo menos uma vez é apresentada na Figura 8. Note que a partir de pouco mais de 2000 dias as observações censuradas começam a ser frequentes. Para os pacientes que foram internados pelo menos uma vez ainda vivos ao final do estudo, a probabilidade de sobrevivência é aproximadamente 0,35.

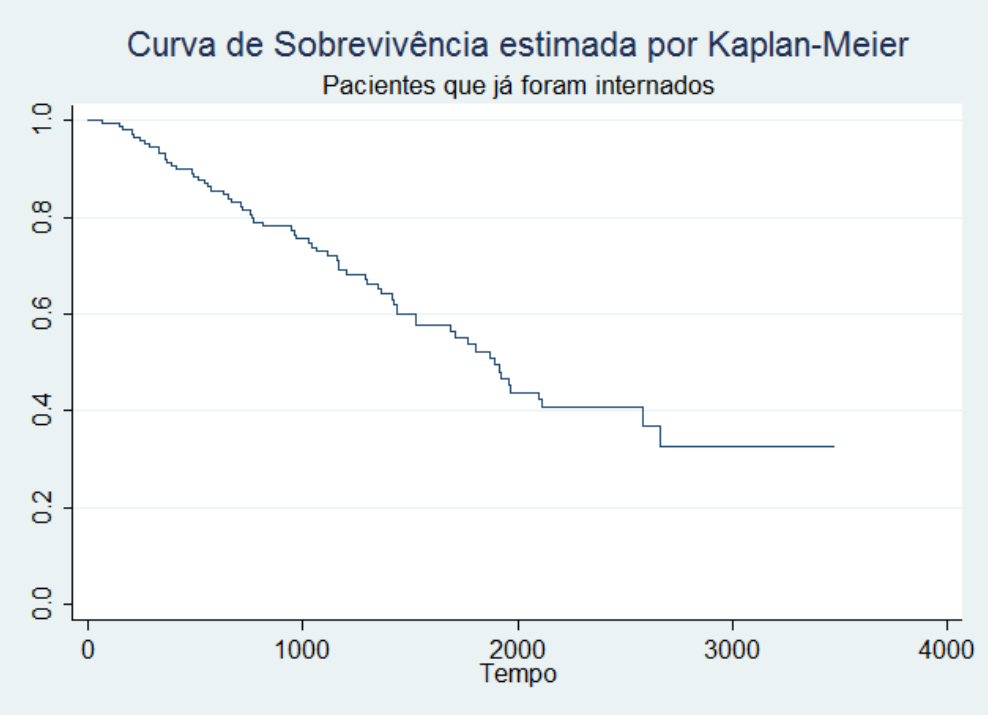

Figura 8: Curva de sobrevivência estimada por Kaplan-Meier para pacientes que foram internados pelo menos uma vez.

4.6.3. Estimativas dos Modelos Paramétricos

O próximo passo é ajustar os modelos paramétricos que serão testados. Isto porque, como foi abordado no Capítulo 4, para escolher entre os modelos paramétricos disponíveis, deve-se investigar qual deles proporciona o melhor ajuste aos dados. Neste

exemplo serão ajustados os quatro modelos paramétricos discutidos neste trabalho. Para tanto, o comando `streg` é utilizado.

Modelo Exponencial

Para ajustar o modelo paramétrico exponencial no STATA utiliza-se o comando:

```
streg, dist(exp) nohr
```

Note que `dist(exp)` é a parte do comando que indica qual modelo paramétrico será utilizado. `nohr` indica que o resultado não é para ser apresentado em termos do *hazard ratio*, ou seja, da razão das taxas de falha. Quando se ajustam os modelos de regressão paramétricos de taxas de falhas proporcionais, a razão taxa de falhas é importante, e por isso o *default* do *software* é apresentar a razão das taxas de falha para cada covariável. Contudo, não há razão de taxas de falha nesse exemplo porque não há covariáveis. Ainda, como o interesse neste caso é obter o parâmetro λ do modelo exponencial, precisa-se do valor da constante que é obtido, no STATA, com o comando `streg` com `nohr`.

O STATA fornece algumas informações quando se executa o comando para ajustar o modelo paramétrico exponencial. Abaixo está apenas a tabela fornecida como resposta ao comando apresentado anteriormente.

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -7.984578   .1240347   -64.37   0.000    -8.227682  -7.741474
------------------------------------------------------------------------------
```

Uma estimativa para o parâmetro λ do modelo paramétrico exponencial é encontrada aplicando a função matemática exponencial no valor da constante estimada pelo STATA, ou seja, $\exp(-7{,}984578)$. Para o STATA fazer esta conta e armazenar o valor do parâmetro estimado em `lambda_e` utiliza-se a programação:

```
scalar lambda_e = exp(_b[_cons])
scalar list lambda_e
  lambda_e =  .00034068
```

O comando `scalar` é utilizado quando se deseja criar um objeto com algum valor e `_b[_cons]` é utilizado para que o valor da constante estimada pelo STATA seja

chamado automaticamente. O comando `list` é utilizado para que o valor do objeto `lambda_e` seja apresentado.

Desta forma, tem-se $\hat{\lambda} = 0{,}00034068$.

Modelo Weibull

Para ajustar o modelo paramétrico Weibull no STATA utiliza-se a programação:

```
streg, dist(weib) nohr
```

onde a tabela da resposta fornecida pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   -10.6618    1.04969   -10.16   0.000    -12.71916   -8.604449
-------------+----------------------------------------------------------------
        /ln_p |   .3094508   .1029522     3.01   0.003     .1076682    .5112335
-------------+----------------------------------------------------------------
           p |   1.362677   .1402906                      1.113678    1.667347
         1/p |   .7338498   .0755515                      .5997553    .8979255
------------------------------------------------------------------------------
```

Utiliza-se estes resultados produzidos pelo STATA para obter as estimativas para os parâmetros p e λ do modelo paramétrico Weibull. O parâmetro p é estimado pelo valor p da tabela, portanto, $\hat{p} = 1{,}362677$. Para obter a estimativa do parâmetro λ deve-se aplicar a função matemática exponencial no valor da constante. Ou seja:

```
scalar lambda_w = exp(_b[_cons])
scalar list lambda_w
  lambda_w =  .00002342
```

Então, $\hat{\lambda} = 0{,}00002342$.

Modelo Log-normal

Para ajustar o modelo paramétrico log-normal no STATA utiliza-se a programação:

```
streg, dist(lnormal)
```

Note que quando se ajusta o modelo paramétrico log-normal não se utiliza o comando `nohr`. Isto porque o modelo de regressão paramétrio log-normal não é um modelo de taxa de falhas proporcionais, mas sim um modelo de tempo de vida

acelerado e, portanto, não faz sentido apresentar a razão das taxas de falha para as covariáveis. A tabela da resposta fornecida pelo STATA é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |       7.559   .1209324    62.51   0.000     7.321977    7.796024
-------------+----------------------------------------------------------------
      /ln_sig |   .1040875   .0923062     1.13   0.259    -.0768294   .2850044
-------------+----------------------------------------------------------------
        sigma |   1.109698    .102432                      .9260479    1.329768
------------------------------------------------------------------------------
```

Utiliza-se os resultados produzidos pelo STATA para obter as estimativas dos parâmetros μ e σ do modelo paramétrico log-normal. O parâmetro μ é estimado pelo valor `_cons` da tabela e o parâmetro σ é estimado pelo valor `sigma` da tabela. Portanto, $\hat{\mu} = 7{,}5590003$ e $\hat{\sigma} = 1{,}109698$.

Modelo Gompertz

Para ajustar o modelo paramétrico Gompertz no STATA utiliza-se a programação:

```
streg, dist(gomp) nohr
```

onde a tabela da resposta fornecida pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -8.335556   .2180755   -38.22   0.000    -8.762976   -7.908136
-------------+----------------------------------------------------------------
      /gamma |   .0003694   .0001724     2.14   0.032     .0000316    .0007073
------------------------------------------------------------------------------
```

Utiliza-se os resultados produzidos pelo STATA para obter as estimativas dos parâmetros α e θ do modelo paramétrico Gompertz. O parâmetro α é estimado pelo valor `gamma` da tabela, portanto, $\hat{\alpha} = 0{,}0003594$. Para obter a estimativa do parâmetro θ deve-se aplicar a função matemática exponencial no valor da constante. Ou seja:

```
scalar teta = exp(_b[_cons])
scalar list teta
      teta =  .00023984
```

Então, $\hat{\theta} = 0{,}00023984$.

4.6.4. Escolha do Modelo Paramétrico Adequado

Uma vez que os quatro modelos paramétricos foram ajustados, utiliza-se as técnicas descritas na Seção 4.5. para verificar a qualidade do ajuste de cada modelo paramétrico, a fim de escolher o modelo paramétrico que melhor se ajusta aos dados.

Métodos Gráficos

A primeira técnica gráfica abordada na Seção 4.5. baseia-se na comparação da curva de sobrevivência estimada por Kaplan-Meier com a curva de sobrevivência estimada pelo modelo paramétrico proposto. Uma maneira de fazer esta comparação é colocando as duas curvas de sobrevivência em um mesmo gráfico. A programação utilizada para fazer este gráfico no STATA, para o modelo paramétrico exponencial é:

```
quietly streg, dist(exp)
predict s_e, surv
label variable s_e "Exponencial"
sts graph, addplot(line s_e _t, sort) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) title(Curvas de Sobrevivência) subtitle(Pacientes que
já foram internados) xtitle(Tempo)
```

Primeiramente é ajustado o modelo paramétrico exponencial. O comando `quietly` é utilizado para suprimir a resposta deste comando, ou seja, o comando `streg` será executado, mas não apresentará resposta. O comando `predict` calcula os valores preditos para o modelo exponencial ajustado, para a função de sobrevivência porque escolheu-se a opção `surv`, e armazena-se estes valores em `s_e`. O *label* da variável `s_e` é definido para que esta informação já apareça corretamente no gráfico que será feito.

Na programação para construir o gráfico, o único comando que ainda não havia sido utilizado é `addplot(line s_e _t, sort)`, que é utilizado para adicionar a linha da função de sobrevivência estimada pelo modelo paramétrico exponencial ao gráfico da função de sobrevivência estimada por Kaplan-Meier.

Esta programação cria um gráfico muito próximo ao da Figura 9, apresentado a seguir. Porém, como não se conseguiu alterar alguns pontos que precisam de ajustes via programação, estes ajustes foram feitos via Menu. Para fazer estas alterações, deve-se clicar com o botão direito do *mouse* em cima do gráfico e clicar na opção *Start Graph Editor*. A primeira alteração é referente às linhas do gráfico, já que as duas serão contínuas. Neste caso a linha da função de sobrevivência estimada pelo modelo paramétrico exponencial será alterada para uma linha pontilhada através de

procedimento já descrito anteriormente. A segunda alteração é referente à legenda. Com um duplo-clique sobre a legenda *Survivor function* abre-se uma janela *Textbox proprietis*, onde, na opção *Text*, deve-se digitar Kaplan-Meier porque esta linha refere-se à função de sobrevivência estimada por Kaplan-Meier, seguido do *ok.*

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Weibull é:

```
quietly streg, dist(weib)
predict s_w, surv
label variable s_w "Weibull"
sts graph, addplot(line s_w _t, sort) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) title(Curvas de Sobrevivência) subtitle(Pacientes que
já foram internados) xtitle(Tempo)
```

Note que o modelo paramétrico Weibull é novamente ajustado porque os comandos subsequentes baseiam-se no último modelo ajustado. Então, antes de fazer os demais gráficos, deve-se ajustar cada modelo paramétrico. Note ainda que as alterações feitas para a linha e legenda para o modelo paramétrico exponencial terão que ser feitas para os gráficos dos demais modelos paramétricos ajustados.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico log-normal é:

```
quietly streg, dist(lnormal)
predict s_ln, surv
label variable s_ln "Log-normal"
sts graph, addplot(line s_ln _t, sort) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) title(Curvas de Sobrevivência) subtitle(Pacientes que
já foram internados) xtitle(Tempo)
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Gompertz é:

```
quietly streg, dist(gomp)
predict s_g, surv
label variable s_g "Gompertz"
sts graph, addplot(line s_g _t, sort) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) title(Curvas de Sobrevivência) subtitle(Pacientes que
já foram internados) xtitle(Tempo)
```

A Figura 9 apresenta a curva de sobrevivência estimada por Kaplan-Meier e a estimada pelo modelo paramétrico exponencial. Note que estas duas curvas são, de maneira geral, próximas, muito embora quando analisadas pontualmente a sobreposição não ocorre perfeitamente.

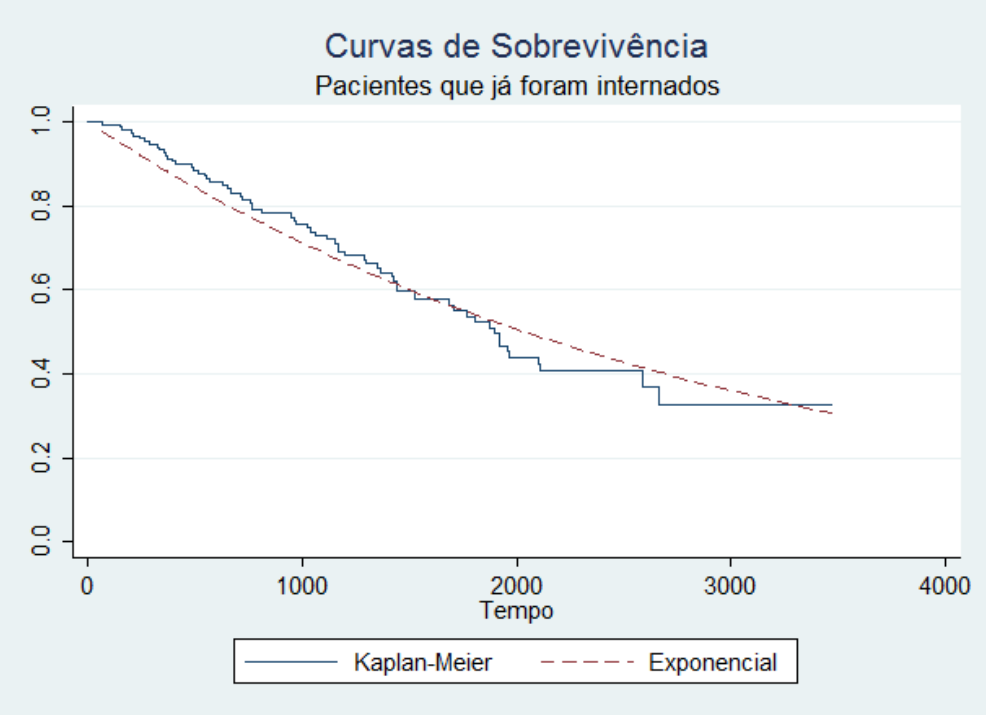

Figura 9: Curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo exponencial.

A Figura 10 apresenta a curva de sobrevivência estimada por Kaplan-Meier e a estimada pelo modelo paramétrico Weibull. Estas curvas sobrepõem-se quase que totalmente nos tempos iniciais, até aproximadamente 2000 dias. Contudo, a partir deste período, as duas curvas começam a se distanciar.

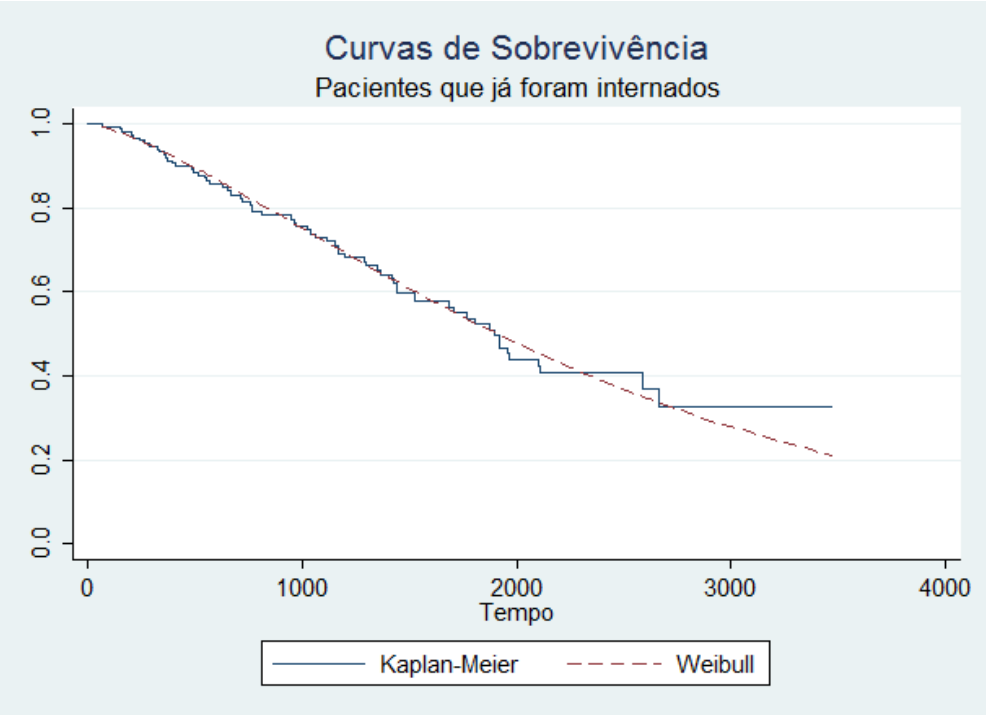

Figura 10: Curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo Weibull.

A Figura 11 apresenta a curva de sobrevivência estimada por Kaplan-Meier e a estimada pelo modelo paramétrico log-normal. Estas curvas, de maneira geral, estão sobrepostas. Note que este modelo paramétrico parece ajustar melhor o tempo final.

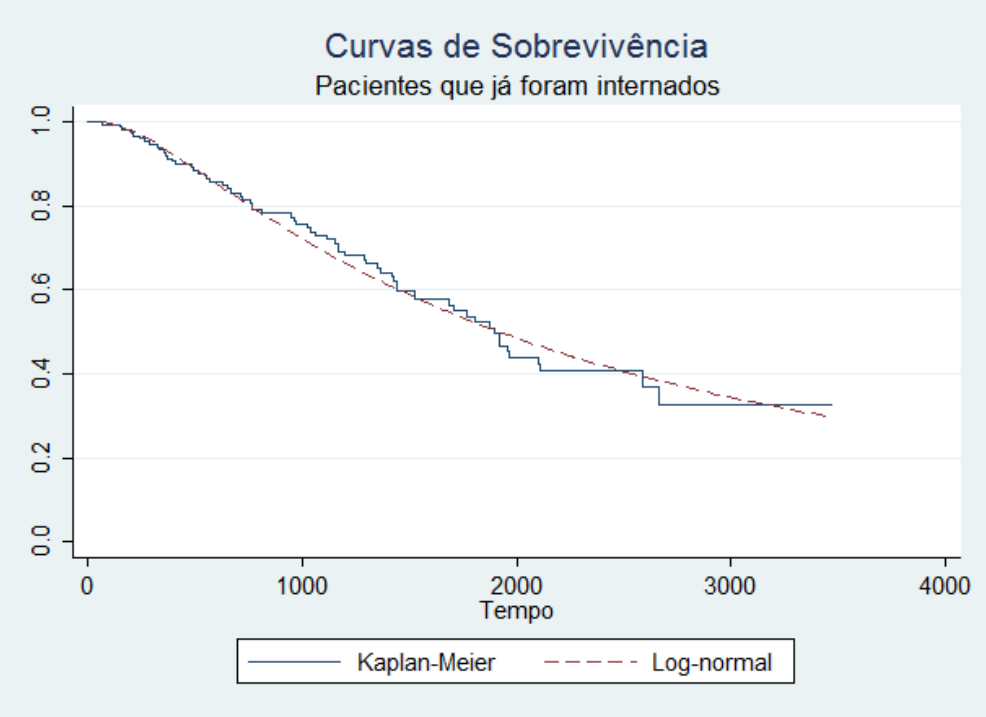

Figura 11: Curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo log-normal.

A Figura 12 apresenta a curva de sobrevivência estimada por Kaplan-Meier e a estimada pelo modelo paramétrico Gompertz. Estas duas curvas são bem próximas, apresentando um distanciamento maior para os últimos períodos de tempo do estudo, muito semelhante com o que ocorreu para o modelo paramétrico Weibull, visto na Figura 10.

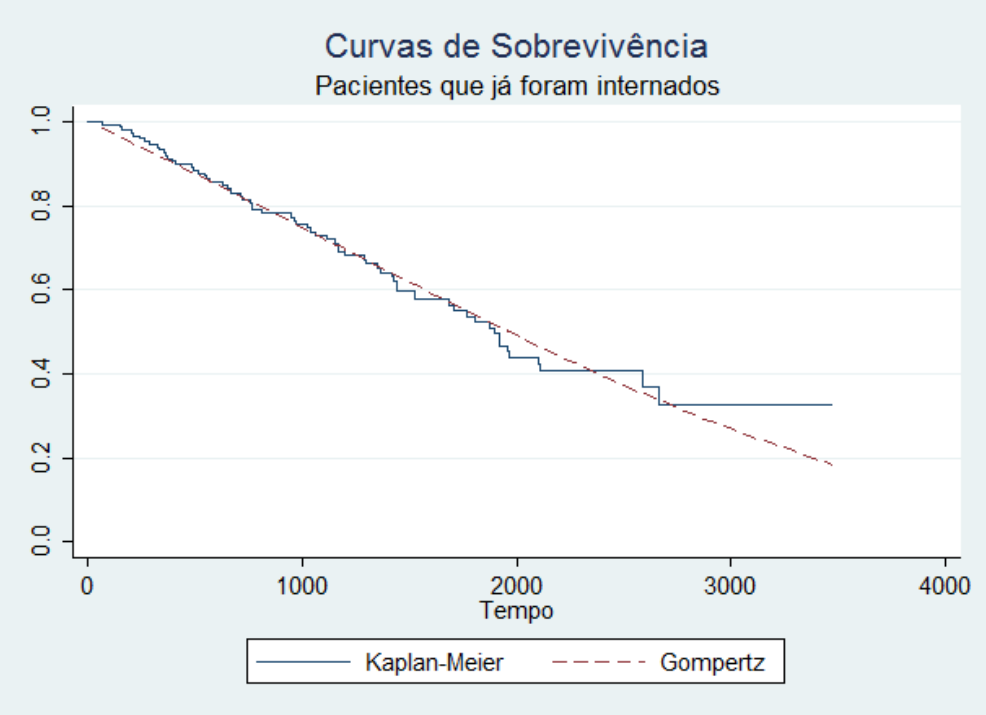

Figura 12: Curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo Gompertz.

Através dos gráficos apresentados observa-se que, de maneira geral, nenhum modelo paramétrico estimou uma curva de sobrevivência com pontos muito distantes da curva estimada por Kaplan-Meier. Pode-se, então, concluir que não houve algum modelo paramétrico que tenha se ajustado tão mal aos dados a ponto desta técnica excluir a possibilidade de sua utilização.

Uma variação deste método gráfico é utilizar a função taxa de falha acumulada onde antes se utilizava a função de sobrevivência. A programação utilizada para criar este gráfico para o modelo paramétrico exponencial no STATA é:

```
generate double H = -ln(km)
label variable H "Kaplan-Meier"
quietly streg, dist(exp)
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada)
subtitle(Pacientes que já foram internados) xtitle(Tempo) ytitle(Taxa
de falha acumulada)
```

O primeiro comando `generate` é utilizado para criar novas variáveis. Neste caso, utiliza-se a relação entre a função de sobrevivência e entre a função taxa de falha acumulada, apresentada na Seção 2.2.3., para criar uma variável chamada `H` com os valores da função taxa de falha acumulada de Kaplan-Meier. O comando `stcurve,`

`cumhaz` é utilizado para construir o gráfico da função taxa de falha acumulada do modelo paramétrico ajustado.

Para este gráfico também foi necessário fazer os ajustes para a linha e para a legenda, como foi feito anteriormente. A única diferença foi que para a legenda, neste caso, foi necessário alterar o nome *Cumulative Hazard* para o nome do modelo paramétrico utilizado.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Weibull é:

```
quietly streg, dist(weib)
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada)
subtitle(Pacientes que já foram internados) xtitle(Tempo) ytitle(Taxa
de falha acumulada)
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico log-normal é:

```
quietly streg, dist(lnorm)
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada)
subtitle(Pacientes que já foram internados) xtitle(Tempo) ytitle(Taxa
de falha acumulada)
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Gompertz é:

```
quietly streg, dist(gomp)
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada)
subtitle(Pacientes que já foram internados) xtitle(Tempo) ytitle(Taxa
de falha acumulada)
```

A Figura 13 apresenta as funções taxas de falha acumulada estimadas por Kaplan-Meier e pelo modelo paramétrico exponencial. As duas curvas são próximas, de uma maneira geral. Contudo, a maioria dos pontos não se sobrepõem, são apenas próximos.

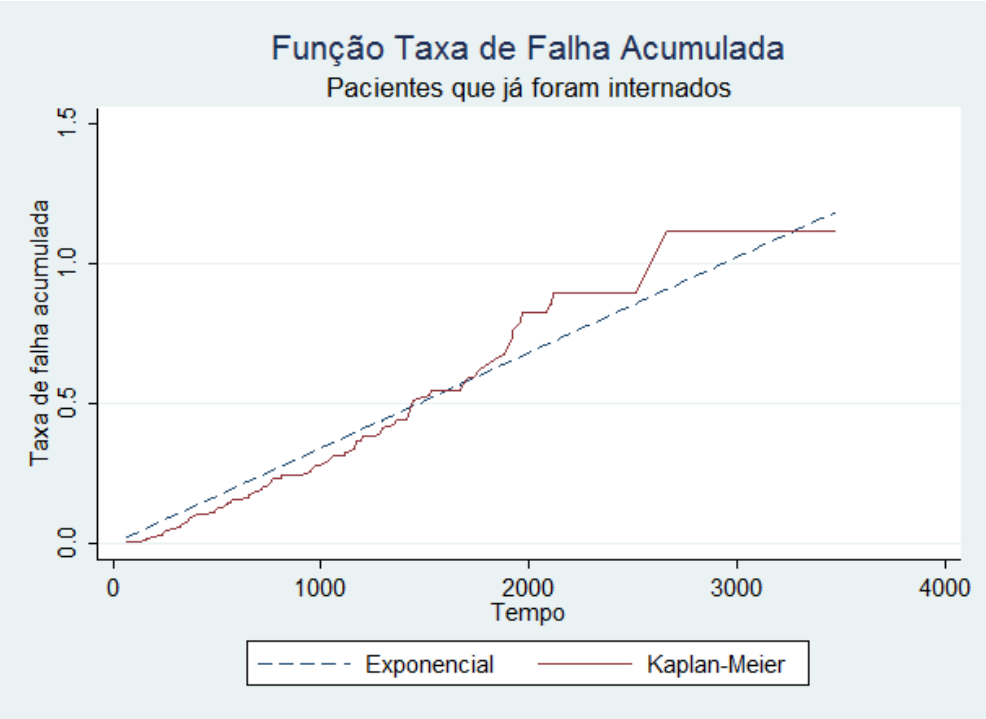

Figura 13: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo exponencial.

A Figura 14 apresenta as funções taxas de falha acumulada estimadas por Kaplan-Meier e pelo modelo paramétrico Weibull. As curvas inicialmente se sobrepõem quase que totalmente, contudo, nos últimos períodos do tempo estas curvas começam a se distanciar.

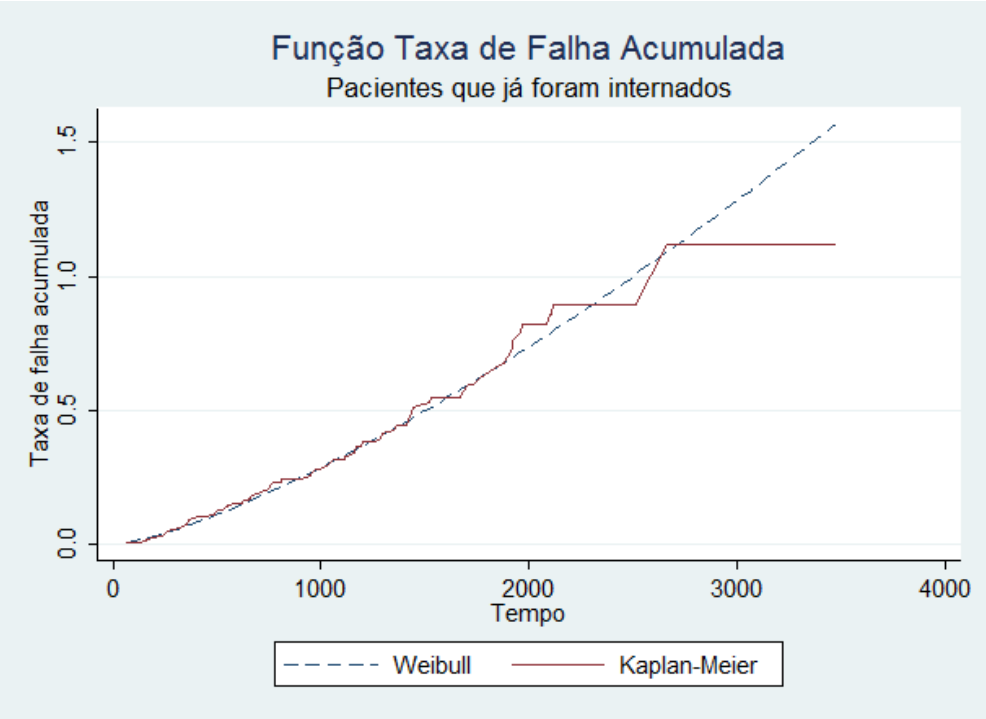

Figura 14: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo Weibull.

A Figura 15 apresenta as funções taxas de falha acumulada estimadas por Kaplan-Meier e pelo modelo paramétrico log-normal. Estas curvas sobrepõem-se quase que totalmente, e a distância dos pontos em que as curvas não estão uma sobre a outra são relativamente pequenas, se comparado aos demais gráficos.

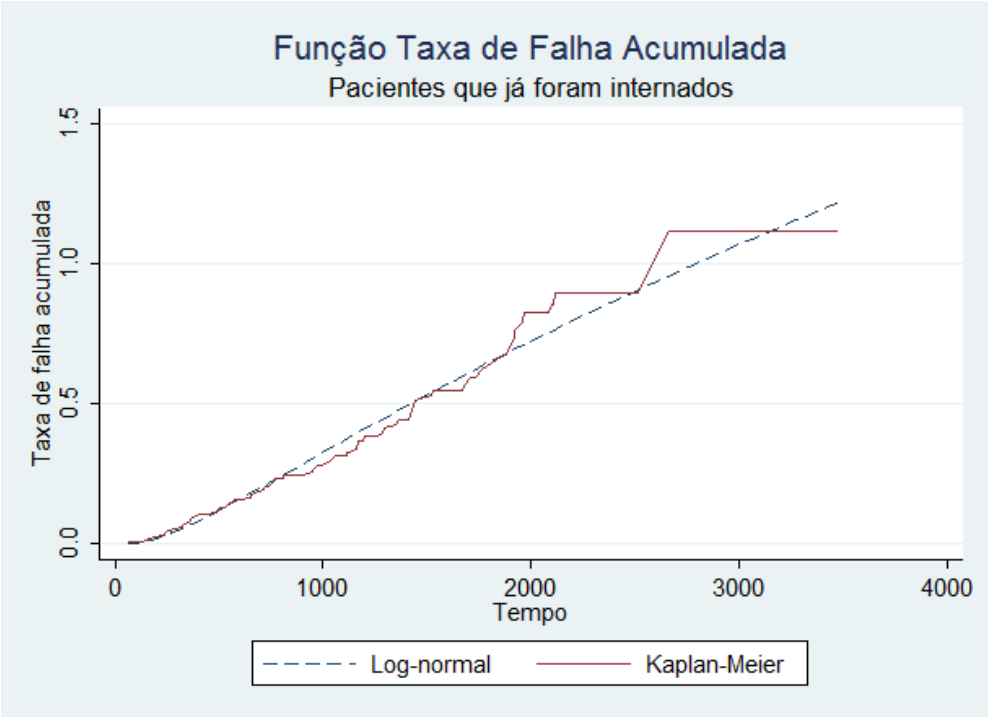

Figura 15: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo log-normal.

A Figura 16 apresenta as funções taxas de falha acumulada estimadas por Kaplan-Meier e pelo modelo paramétrico Gompertz. As duas curvas estão sobrepostas inicialmente e, para os tempos finais, se distanciam. Note que este gráfico é muito parecido com o do modelo paramétrico Weibull.

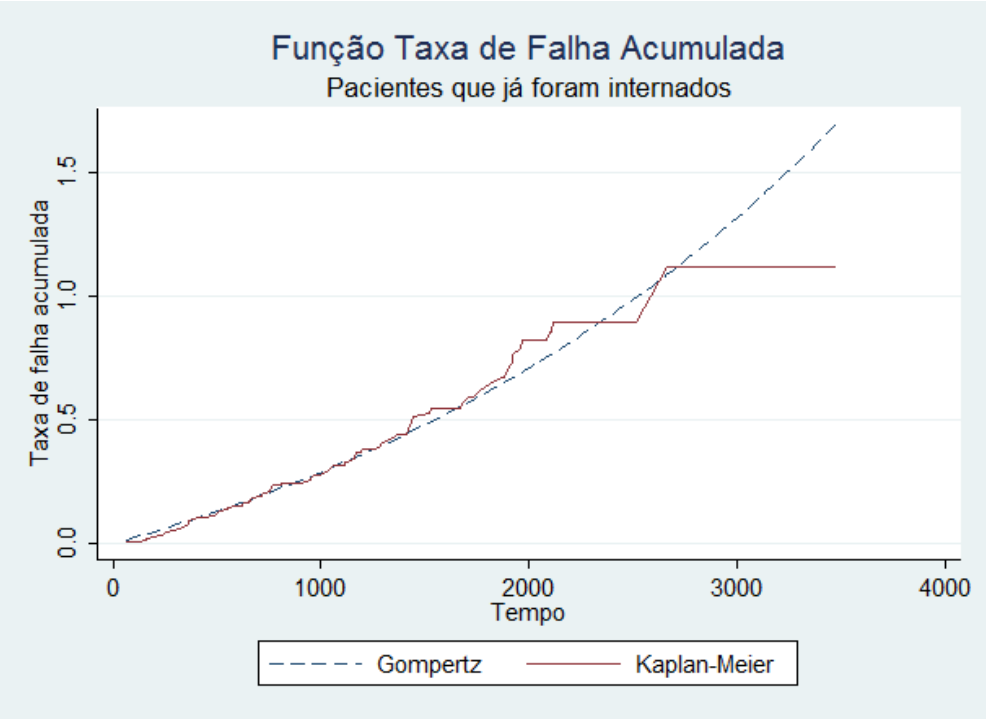

Figura 16: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo Gompertz.

Através dos gráficos apresentados, observa-se que, de maneira geral, nenhum modelo paramétrico estimou uma função taxa de falha acumulada com pontos muito distantes da função taxa de falha acumulada estimada por Kaplan-Meier. Pode-se, então, concluir que não houve algum modelo paramétrico que tenha se ajustado tão mal aos dados a ponto desta técnica excluir a possibilidade de sua utilização.

Outra variação deste método gráfico pode ser feita através do gráfico da função de sobrevivência estimada por Kaplan-Meier *versus* a função de sobrevivência estimada pelo modelo paramétrico proposto. Este gráfico deve gerar uma reta para indicar que o modelo paramétrico proposto ajusta-se bem aos dados.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico exponencial é:

```
sts generate km = s
scatter km s_e, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Exponencial)
```

O comando `scatter` é utilizado para fazer gráficos de dispersão, onde a função de sobrevivência estimada por Kaplan-Meier ficará no eixo x e a estimada pelo modelo paramétrico exponencial no eixo y. O comando `msize` é utilizado para definir o tamanho dos pontos do gráfico de dispersão. Para este gráfico os pontos serão pequenos.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Weibull é:

```
scatter km s_w, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Weibull)
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico log-normal é:

```
scatter km s_ln, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Log-normal)
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Gompertz é:

```
scatter km s_g, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Gompertz)
```

A Figura 17, a Figura 18, a Figura 19 e a Figura 20 apresentam a função de sobrevivência estimada por Kaplan-Meier *versus* a função de sobrevivência estimada pelos modelos paramétricos exponencial, Weibull, log-normal e Gompertz, respectivamente. O gráfico para os quatro modelos paramétricos foi muito semelhante: apesar do comportamento dos primeiros pontos não ser exatamente linear, de modo geral tem-se uma reta.

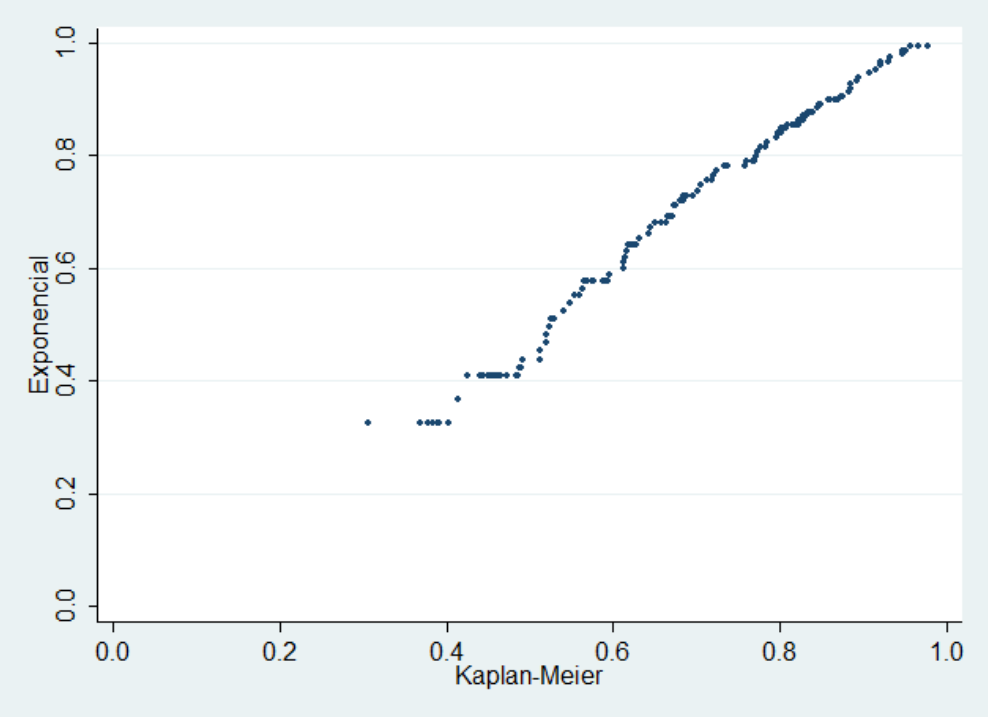

Figura 17: Curva de sobrevivência estimada por Kaplan-Meier *versus* curva de sobrevivência estimada pelo modelo exponencial.

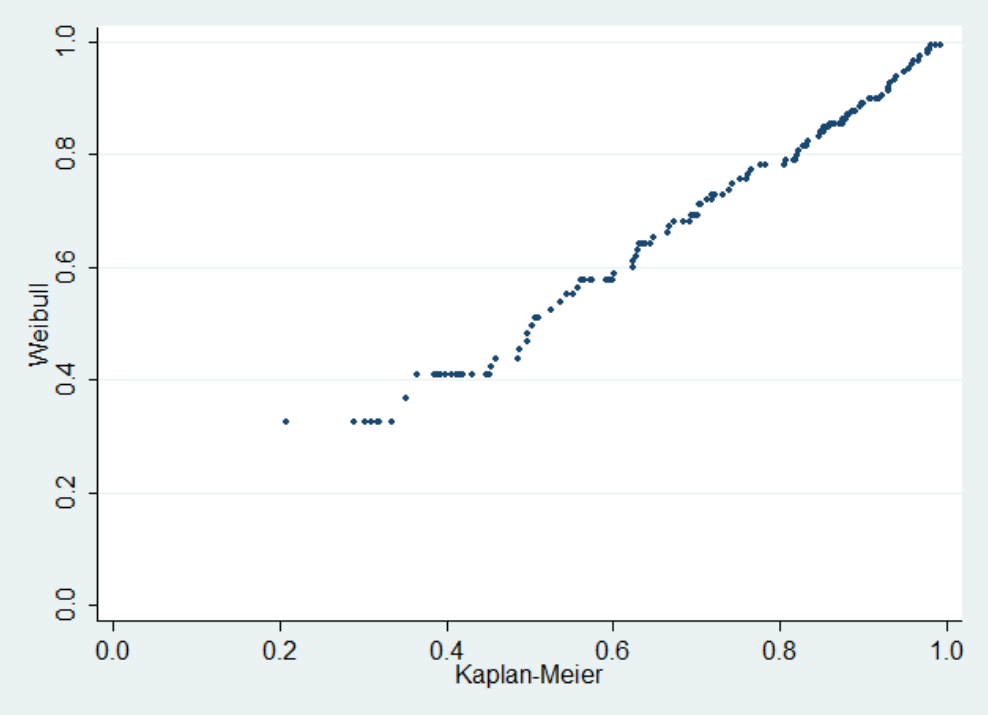

Figura 18: Curva de sobrevivência estimada por Kaplan-Meier *versus* curva de sobrevivência estimada pelo modelo Weibull.

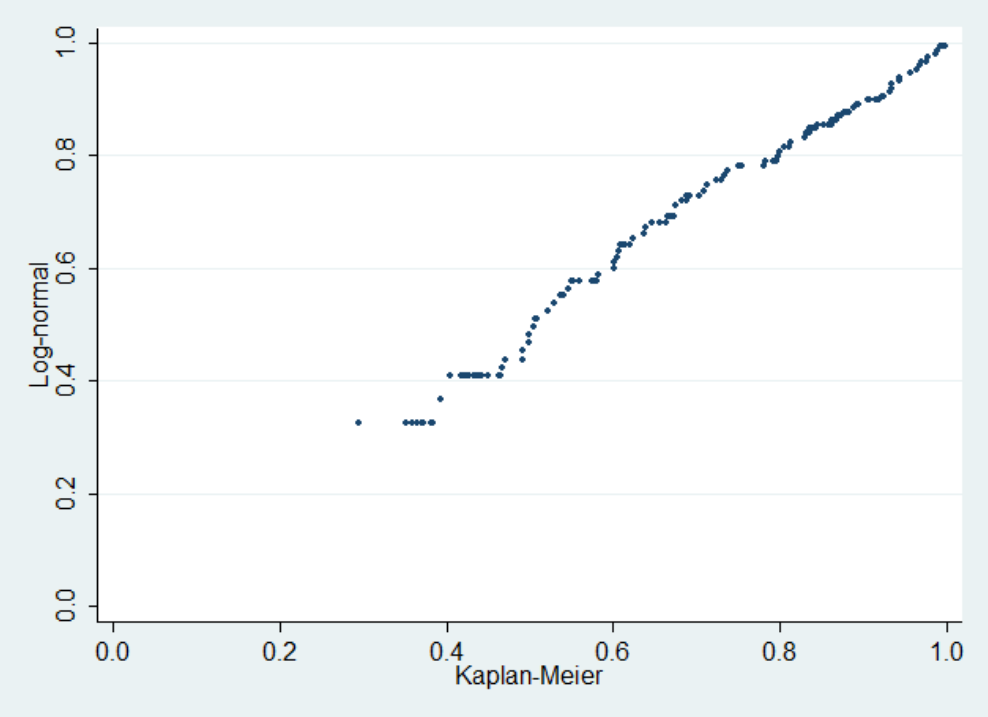

Figura 19: Curva de sobrevivência estimada por Kaplan-Meier *versus* curva de sobrevivência estimada pelo modelo log-normal.

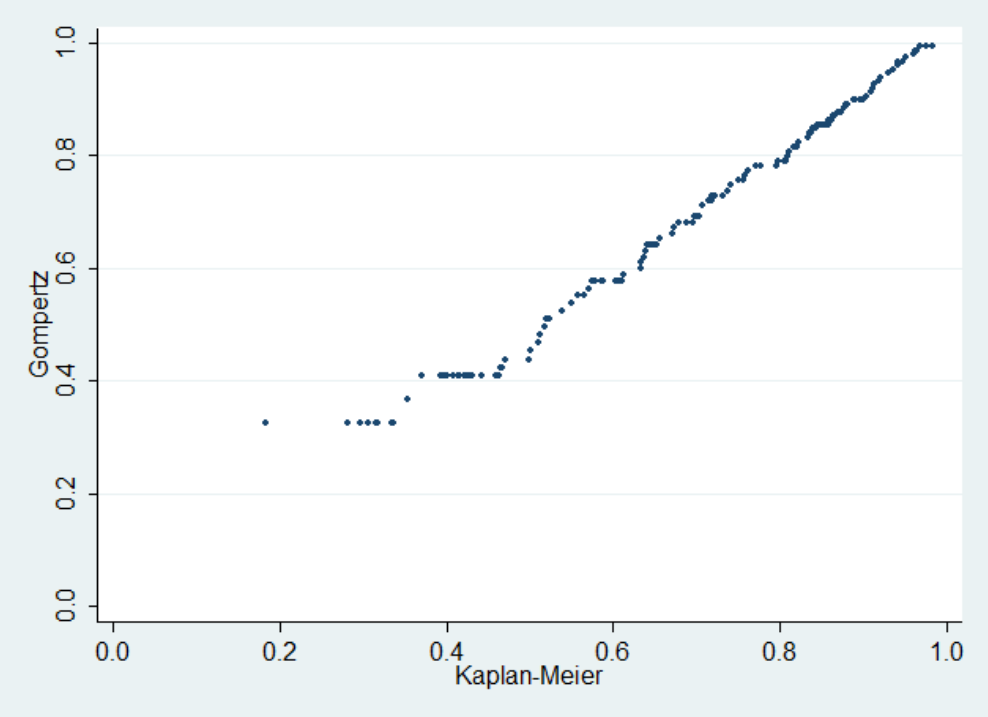

Figura 20: Curva de sobrevivência estimada por Kaplan-Meier *versus* curva de sobrevivência estimada pelo modelo Gompertz.

Através dos gráficos apresentados, observa-se que o comportamento dos quatro modelos paramétricos são muito próximos, já que todos produziram aproximadamente uma reta. Pode-se, então, concluir que não houve algum modelo paramétrico que tenha se ajustado tão mal aos dados a ponto desta técnica excluir a possibilidade de sua utilização.

O segundo método gráfico refere-se à linearização da curva de sobrevivência dos modelos paramétricos. Lembre-se que este método não é aplicado ao modelo Gompertz, uma vez que não é possível obter uma relação linear entre o tempo e a função de sobrevivência para este modelo paramétrico.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico exponencial é:

```
generate lns = -ln(km)
scatter lns _t, msize(small) ylabels(0 0.5 1 1.5, format(%9.1f))
xtitle(t) ytitle("-log(S(t))")
```

Note, no comando acima, que utiliza-se a variável `_t`. Esta variável contêm os diferentes tempos de falha ou censura. Ainda, o comando `ln` é utilizado para calcular o logaritmo natural da variável `km`, que contêm os valores da função de sobrevivência estimados por Kaplan-Meier.

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico Weibull é:

```
generate lnt = ln(_t)
generate lnlns = ln(-ln(km))
scatter lnlns lnt, msize(small) xtitle("ln(t)") ytitle("log(-
log(S(t)))")
```

A programação utilizada para criar este gráfico no STATA para o modelo paramétrico log-normal é:

```
generate norms = invnormal(km)
scatter norms lnt, msize(small) xtitle("ln(t)")
ytitle("invnormal(S(t))")
```

A Figura 21 apresenta o gráfico resultante da linearização do modelo exponencial. Este gráfico é, de maneira geral, uma reta, apesar de apresentar alguns pontos que se distanciam um pouco de uma reta a partir do tempo 2000.

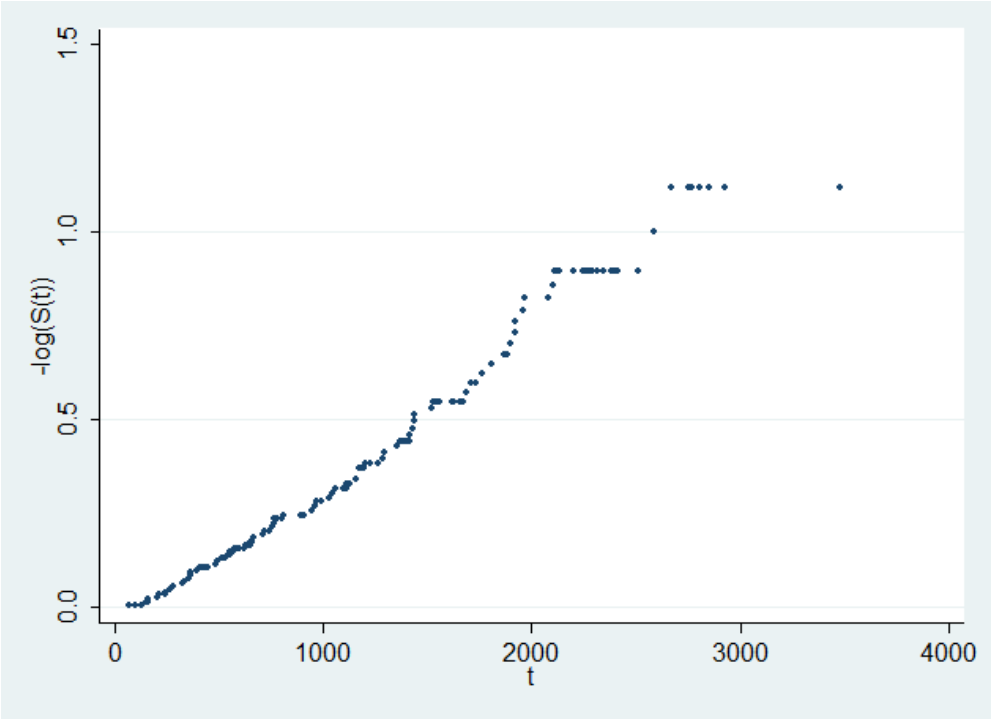

Figura 21: Linearização do modelo exponencial.

A Figura 22 apresenta o gráfico resultante da linearização do modelo Weibull. Os três primeiros pontos da curva distanciam-se um pouco da reta, contudo, o restante da curva é uma reta praticamente perfeita.

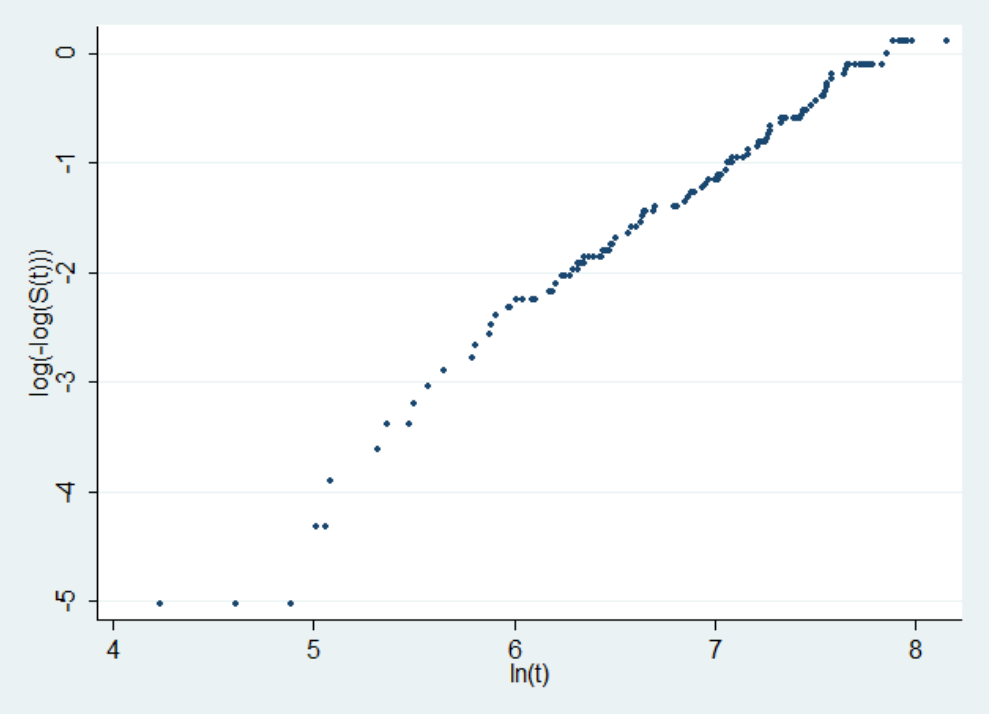

Figura 22: Linearização do modelo Weibull.

A Figura 23 apresenta o gráfico resultante da linearização do modelo log-normal. Os dois primeiros pontos da curva distanciam-se um pouco da reta, contudo, o restante da curva é uma reta praticamente perfeita.

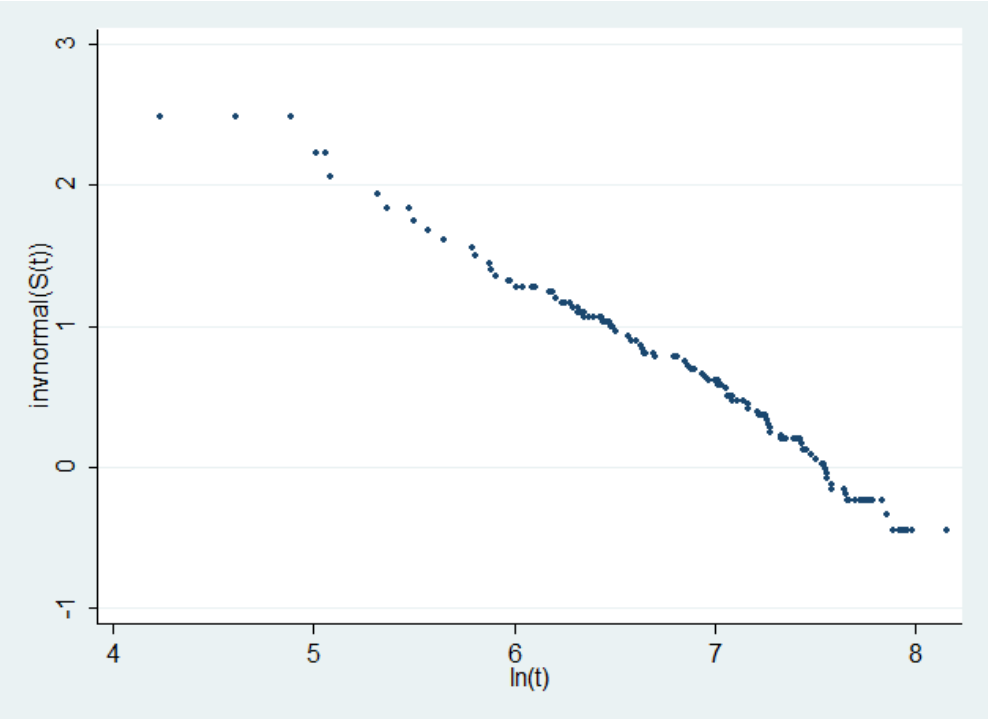

Figura 23: Linearização do modelo log-normal.

Através dos gráficos apresentados, observa-se que nenhum modelo paramétrico apresentou problema na linearização. Pode-se então concluir que não houve algum modelo paramétrico que tenha se ajustado tão mal aos dados a ponto da técnica de linearização excluir a possibilidade da utilização de algum modelo paramétrico.

Através de todos gráficos apresentados, não descartou-se a possibilidade de utilizar nenhum modelo paramétrico. Note que as técnicas gráficas são mais apropriadas para indicar que algum modelo paramétrico não se ajusta bem aos dados do que que algum modelo paramétrico é o mais adequado para fazer a modelagem.

Teste dos Modelos Encaixados

O teste para modelos encaixados apresentado na Seção 4.5.2. será agora executado. Para tanto, ajusta-se o modelo gama generalizado e testa-se, através do log da verossimilhança, o ajuste dos modelos exponencial, Weibull e log-normal. Lembre-se que este teste não é aplicado para o modelo Gompertz, porque este modelo não é um caso particular do modelo gama generalizado.

A programação utilizada para ajustar o modelo paramétrico gama generalizado no STATA é:

```
streg, dist(gamma)
```

onde uma parte da resposta do comando fornecida pelo *software* é:

```
Gamma regression -- accelerated failure-time form

No. of subjects =          153             Number of obs   =       153
No. of failures =           65
Time at risk    =       190797
                                            Wald chi2(0)    =         .
Log likelihood  =   -144.26732              Prob > chi2     =         .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   7.696137   .1689376    45.56   0.000     7.365025    8.027249
-------------+----------------------------------------------------------------
     /ln_sig |  -.0626634   .2177901    -0.29   0.774    -.4895242   .3641974
      /kappa |   .4639342   .4745216     0.98   0.328    -.466111    1.393979
-------------+----------------------------------------------------------------
       sigma |   .9392596   .2045615                      .6129179    1.439358
------------------------------------------------------------------------------
```

Para realizar o teste dos modelos encaixados é necessário armazenar o valor do logaritmo da verossimilhança de cada um dos modelos paramétricos. Quando se ajusta um modelo paramétrico este valor é fornecido como um resultado, chamado de `Log likelihood`. Para o modelo gama generalizado tem-se: `Log likelihood = -144.26732`. Portanto, faz-se:

```
scalar ll_g = -144.26732
```

A programação com o valor do logaritmo da verossimilhança do modelo paramétrico exponencial e que calcula o valor da estatística do teste e p-valor do teste dos modelos encaixados para o modelo paramétrico exponencial no STATA é:

```
scalar ll_e = -148.75858
scalar trv_e =  2*(ll_g - ll_e)
scalar p_e =  1 - chi2(2,trv_e)
```

A estatística do teste tem distribuição qui-quadrado, motivo pelo qual o p-valor é calculado com o comando `chi2`. Este comando calcula a probabilidade acumulada a

esquerda da distribuição qui-quadrado com 2 graus de liberdade para o modelo exponencial, motivo pelo qual utiliza-se `1 - chi2(2,trv_e)` para calcular o p-valor.

A programação com o valor do logaritmo da verossimilhança do modelo paramétrico Weibull e que calcula o valor da estatística do teste e p-valor do teste dos modelos encaixados para o modelo paramétrico Weibull no STATA é:

```
scalar ll_w = -144.81288
scalar trv_w = 2*(ll_g-ll_w)
scalar p_w = 1-chi2(1,trv_w)
```

A programação com o valor do logaritmo da verossimilhança do modelo paramétrico log-normal e que calcula o valor da estatística do teste e p-valor do teste dos modelos encaixados para o modelo paramétrico log-normal no STATA é:

```
scalar ll_ln = -144.73577
scalar trv_ln= 2*(ll_g-ll_ln)
scalar p_ln = 1-chi2(1,trv_ln)
```

A programação utilizada para exibir o p-valor de cada modelo paramétrico acima calculado no STATA é:

```
scalar list p_e p_w p_ln
```

onde a resposta do comando fornecida pelo *software* é:

```
       p_e =  .01120651
       p_w =  .29622319
      p_ln =  .33307636
```

Há evidências estatísticas de que o modelo exponencial não é um modelo adequado aos dados ($p - valor = 0{,}0112$). Para os modelos Weibull e log-normal não há evidências estatísticas de que estes modelos não sejam adequados aos dados ($p - valor = 0{,}2962$ e $p - valor = 0{,}3331$).

Com esse resultado, o modelo paramétrico exponencial não precisaria ser analisado na próxima etapa, quando se analisam os critérios de informação. Contudo, como este trabalho tem o objetivo de auxiliar quem irá reproduzir a modelagem dos modelos paramétricos, os critérios de informação serão também analisados para o modelo paramétrico exponencial.

Critérios de Informação

Os critérios de informação são importantes porque são eles que indicam o modelo paramétrico mais adequado aos dados. Contudo, não deve-se utilizar apenas os critérios de informação para escolher o modelo paramétrico, porque as ferramentas apresentadas anteriormente, que são úteis para indicar modelos paramétricos inadequados, podem indicar que o modelo que possui menores valores para os critérios de informação não são adequados. Ou seja, por mais que sejam os critérios de informação que indicam o modelo adequado, esta ferramenta deve ser utilizada conjuntamente com as demais.

A programação utilizada no STATA para armazenar os valores dos critérios de informação para o modelo paramétrico exponencial é:

```
quietly streg, dist(exp)
estimates store exp
```

A programação utilizada no STATA para armazenar os valores dos critérios de informação para o modelo paramétrico Weibull é:

```
quietly streg, dist(weib)
estimates store weib
```

A programação utilizada no STATA para armazenar os valores dos critérios de informação para o modelo paramétrico log-normal é:

```
quietly streg, dist(lnormal)
estimates store lnorm
```

A programação utilizada no STATA para armazenar os valores dos critérios de informação para o modelo paramétrico Gompertz é:

```
quietly streg, dist(gomp)
estimates store gomp
```

A programação utilizada para exibir o valor dos critérios de informação é:

```
estimates stats _all
```

onde a resposta do comando fornecida pelo *software* é:

```
-----------------------------------------------------------------------------
Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
------+----------------------------------------------------------------------
  exp |    153   -148.7586   -148.7586      1     299.5172    302.5476
 weib |    153   -144.8129   -144.8129      2     293.6258    299.6866
lnorm |    153            .   -144.7358      2     293.4715    299.5324
 gomp |    153            .   -146.5849      2     297.1698    303.2307
-----------------------------------------------------------------------------
         Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Os modelos Weibull e log-normal apresentaram os menores valores para os critérios de informação. Note que os valores do AIC e do BIC foram muito próximos para estas duas distribuições. Como em ambos os critérios o menor valor é para o modelo log-normal, este modelo foi o que melhor se ajustou aos dados.

4.6.5. Conclusões

A escolha do modelo paramétrico adequado para o grupo de pacientes que nunca havia sido internado e para os pacientes que já haviam sido internados pelo menos uma vez foi realizada separadamente, como discutiu-se na Seção 4.6.2..

A modelagem para os pacientes que haviam sido internados pelo menos uma vez foi apresentada passo a passo nos itens anteriores. A modelagem para o grupo de pacientes que nunca haviam sido internados foi feita de maneira análoga, e aqui serão apresentados apenas os resultados. As conclusões para os dois grupos de pacientes são apresentadas a seguir.

Grupo de Pacientes que Haviam Sido Internados Pelo Menos Uma Vez

O modelo paramétrico log-normal é o modelo adequado para explicar o tempo até a morte ocasionada por todas as causas, dos pacientes do ambulatório de insuficiência cardíaca do Hospital de Clínicas de Porto Alegre, que haviam sido internados pelo menos uma vez.

A função de sobrevivência estimada dos pacientes do ambulatório de insuficiência cardíaca do HCPA, que foram internados pelo menos uma vez, é apresentada na Figura 24.

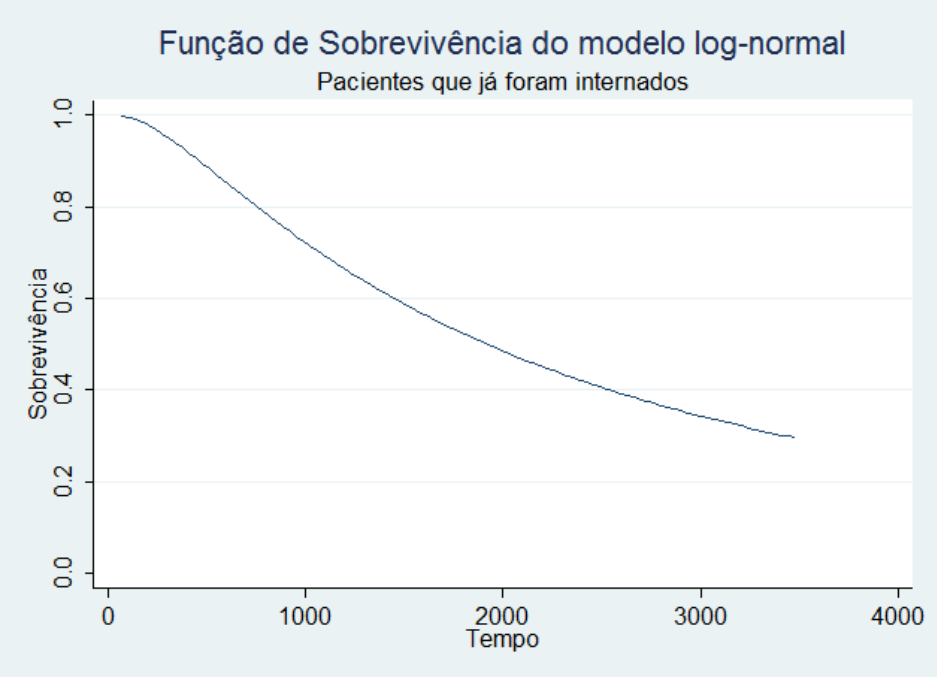

Figura 24: Função de sobrevivência estimada pelo modelo paramétrico log-normal para os pacientes que haviam sido internados pelo menos uma vez.

A programação utilizada para construir o gráfico da função de sobrevivência do modelo paramétrico log-normal no STATA é:

```
streg, dist(lnormal)
stcurve, survival ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
title(Função de Sobrevivência do modelo log-normal) subtitle(Pacientes
que já foram internados) xtitle(Tempo) ytitle(Sobrevivência)
```

Como já havia sido dito, uma das grandes vantagens dos modelos paramétricos é a possibilidade de fazer extrapolação da curva de sobrevivência. Isso quer dizer que é possível estimar a probabilidade de sobreviver a 4000 dias dos pacientes do ambulatório de insuficiência cardíaca do HCPA, dado que eles estão vivos até esse dia.

Como o modelo paramétrico log-normal é o modelo adequado, a função de sobrevivência pode ser estimada pela expressão:

$$\hat{S}(t) = \Phi\left(\frac{-\log(t) + 7{,}5590003}{1{,}109698}\right),$$

onde os valores $\hat{\mu} = 7{,}5590003$ e $\hat{\sigma} = 1{,}109698$ foram estimados anteriormente, no início desta seção.

Para obter a estimativa da probabilidade de sobreviver a 4000 dias substitui-se, na expressão acima, $t = 4000$. Ou seja,

$$\hat{S}(4000) = \Phi\left(\frac{-\log(4000) + 7{,}5590003}{1{,}109698}\right).$$

A programação utilizada para fazer este cálculo no STATA é:

```
scalar mi = _b[_cons]
scalar sigma = 1.109698
scalar prev_4000 = normal((-log(4000)+mi)/sigma)
scalar list prev_4000
```

onde o comando `normal` calcula a função de distribuição acumulada de uma normal padrão. A resposta do comando fornecida pelo *software* é:

```
prev_4000 =  .25386167
```

Logo, estima-se que a probabilidade de sobreviver a 4000 dias daqueles pacientes do ambulatório de insuficiência cardíaca do HCPA, que foram internados pelo menos uma vez e que estão vivos até esse dia, é de 0,254, se o comportamento do tempo de sobrevida destes pacientes continuar nas mesmas condições.

Grupo de Pacientes que Nunca Haviam Sido Internados

A escolha do modelo paramétrico adequado aos dados dos pacientes que nunca haviam sido internados foi feita separadamente, e somente o modelo paramétrico escolhido será apresentado neste trabalho. Contudo, algumas observações sobre esta modelagem serão feitas.

O gráfico da linearização do modelo paramétrico exponencial não gerou uma reta, como pode ser visto na Figura 25. Conclui-se, então, que o modelo paramétrico exponencial não é adequado aos dados e, portanto, não deve ser utilizado.

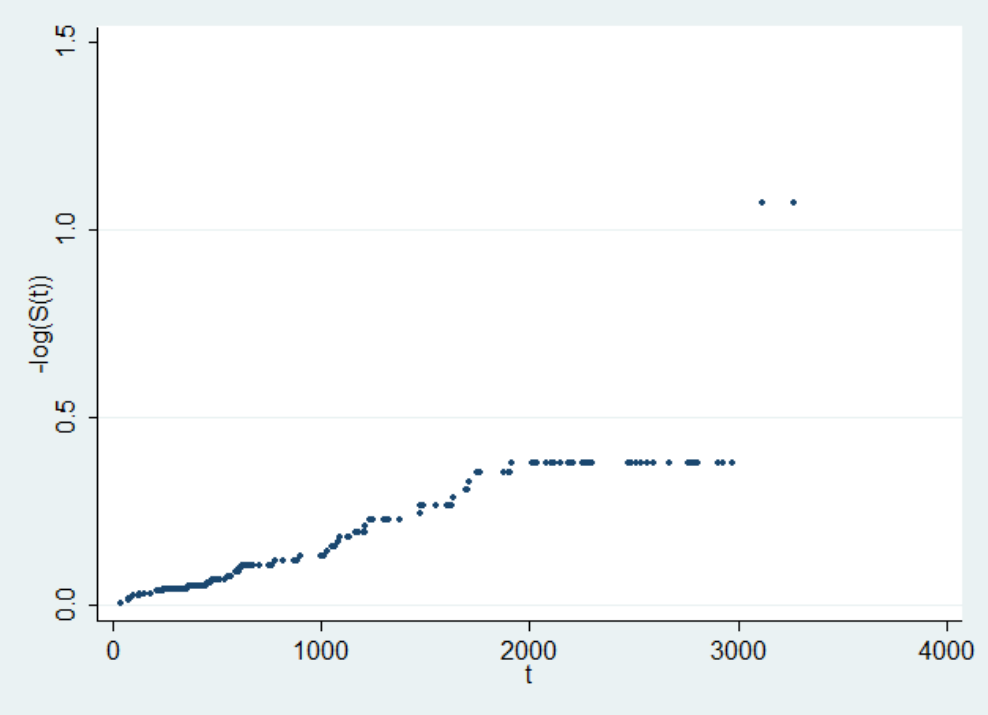

Figura 25: Linearização do modelo exponencial.

Contudo, para os critérios de informação a resposta do comando fornecida pelo STATA é:

```
-----------------------------------------------------------------------------
 Model |    Obs    ll(null)   ll(model)     df       AIC          BIC
-------+---------------------------------------------------------------------
   exp |    165   -102.3923   -102.3923      1     206.7846     209.8906
  weib |    165   -102.2887   -102.2887      2     208.5774     214.7893
 lnorm |    165            .   -102.9852      2     209.9704     216.1822
  gomp |    165            .   -102.3227      2     208.6454     214.8573
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Ou seja, os critérios de informação indicam que o modelo paramétrico exponencial é o modelo que melhor se ajusta aos dados. Como, pela Figura 25, concluímos que o modelo paramétrico exponencial não é adequado aos dados, então o modelo paramétrico Weibull, que apresentou os segundos menores valores para o AIC e para o BIC, e que não apresentou nenhum problema nas análises gráficas e teste dos modelos encaixados, é o modelo que melhor se ajustou aos dados. Este exemplo mostra a importância das análises gráficas, porque se apenas os critérios de informação tivessem sido utilizados para escolher o modelo paramétrico, um modelo inadequado aos dados teria sido selecionado.

Logo, o modelo paramétrico Weibull é o modelo adequado para explicar o tempo até a morte ocasionada por todas as causas, dos pacientes do ambulatório de insuficiência cardíaca do Hospital de Clínicas de Porto Alegre, que nunca haviam sido internados.

A função de sobrevivência dos pacientes do ambulatório de insuficiência cardíaca do HCPA que nunca haviam sido internados é apresentada na Figura 26.

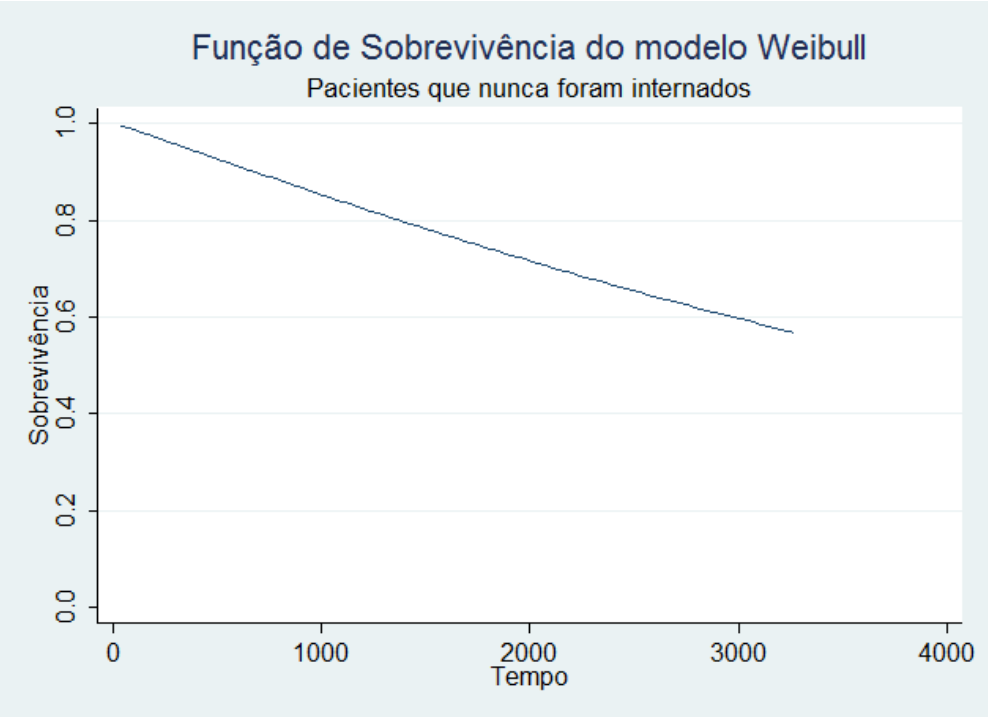

Figura 26: Função de sobrevivência estimada pelo modelo paramétrico Weibull para os pacientes que nunca haviam sido internados.

A programação utilizada para construir o gráfico da função de sobrevivência do modelo paramétrico log-normal no STATA é:

```
streg, dist(weibull)
stcurve, survival ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
title(Função de Sobrevivência do modelo Weibull) subtitle(Pacientes
que nunca foram internados) xtitle(Tempo) ytitle(Sobrevivência)
```

Como o modelo paramétrico Weibull é o modelo adequado, a função de sobrevivência pode ser estimada pela expressão:

$$\hat{S}(t) = \exp\{-0{,}00009702t^{1{,}071498}\},$$

onde $\hat{\lambda} = 0{,}00009702$ e $\hat{p} = 1{,}071498$.

Para obter a estimativa da probabilidade de sobreviver a 4000 dias substitui-se, na expressão acima, $t = 4000$. Ou seja,

$$\hat{S}(4000) = \exp\{-0{,}00009702 * 4000^{1{,}071498}\}$$

A programação utilizada para fazer este cálculo no STATA é:

```
scalar aux = 4000^1.071498
scalar prev2_4000 = exp(-lambda_w*aux)
scalar list prev2_4000
```

onde a resposta do comando fornecida pelo *software* é:

```
prev2_4000 =   .4954949
```

Logo, estima-se que a probabilidade de sobreviver a 4000 dias, daqueles pacientes do ambulatório de insuficiência cardíaca do HCPA, que nunca haviam sido internados e estão vivos até esse dia, é de 0,495, se o comportamento do tempo de sobrevida destes pacientes continuar nas mesmas condições.

5. Modelos de Regressão Paramétricos

Os modelos de regressão paramétricos são, de certa forma, uma extensão dos modelos paramétricos apresentados no Capítulo 4, uma vez que eles também supõem uma distribuição densidade de probabilidades para a variável aleatória tempo de falha T. A diferença entre os modelos paramétricos e os modelos de regressão paramétricos é que os modelos de regressão paramétricos permitem a inclusão de covariáveis na análise, de modo que elas auxiliem na estimação da curva de sobrevivência.

Segundo Colosimo e Giolo (2006), os modelos de regressão paramétricos são mais eficientes que os modelos de regressão semi-paramétricos, porém menos flexíveis. O modelo de Cox, um dos modelos de análise de sobrevivência mais utilizados e difundidos, é um exemplo de modelo de regressão semi-paramétrico. Detalhes sobre o modelo de Cox podem ser encontrados em Hosmer e Lemeshow (1999) e Colosimo e Giolo (2006).

Há duas abordagens para os modelos de regressão paramétricos: os modelos de taxa de falha proporcionais e os modelos de tempo de vida acelerado (Collett, 2003; Cleves *et al*, 2008). Os modelos de taxas de falha proporcionais serão abordados na Seção 5.1. e os modelos de tempo de vida acelerado na Seção 5.2.. Nas seções seguintes, são apresentados métodos para selecionar covariáveis, para escolher o modelo de regressão paramétrico adequado aos dados, para verificar a qualidade do ajuste do modelo de regressão paramétrico selecionado, uma seção referente à interpretação dos coeficientes estimados pelos modelos e uma seção onde será resolvido um exemplo passo a passo.

5.1. Modelos de Taxas de Falhas Proporcionais

Os modelos de taxas de falha proporcionais também são conhecidos como modelos de riscos proporcionais. Estes modelos têm como característica principal a suposição de taxas de falha proporcionais, e por isso são conhecidos como a versão paramétrica do modelo de regressão de Cox (Collett, 2003). Esta suposição impõe que as taxas de falha (ou riscos) sejam proporcionais ao longo do tempo, ou seja, que a razão das taxas de falha seja constante ao longo do tempo.

A expressão geral do risco de morte do indivíduo i no tempo t, para os modelos de regressão de taxas de falhas proporcionais, é dada por:

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) h_0(t), \quad t \geq 0 \text{ e } i = 1, \ldots, n$$

para p covariáveis. A expressão acima é igual a expressão da função taxa de falha do modelo de Cox. A diferença entre estes modelos é que para os modelos paramétricos $h_0(t)$ terá suposição paramétrica, ou seja, será utilizada uma distribuição de probabilidade para $h_0(t)$.

Os modelos exponencial, Weibull e Gompertz são modelos de taxas de falha proporcionais e serão discutidos nas próximas seções.

5.1.1. Modelo de Regressão Exponencial

O modelo de regressão exponencial é adequado quando o tempo de falha é bem descrito através de uma distribuição de probabilidades exponencial. Para o modelo de regressão exponencial, $h_0(t)$ é dada por:

$$h_0(t) = \lambda, \quad t \geq 0 \text{ e } \lambda > 0$$

Consequentemente, a função risco de morte é dada por:

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) \lambda, \quad t \geq 0 \text{ e } \lambda > 0.$$

Note, através da expressão acima, que o risco de morte será constante ao longo do tempo. Neste caso, o efeito das covariáveis será aumentar ou diminuir este risco constante.

A função de sobrevivência do modelo de regressão exponencial é dada por:

$$S(t) = \exp\left(-\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)\, \lambda t\right) \quad t \geq 0 \text{ e } \lambda > 0.$$

5.1.2. Modelo de Regressão Weibull

O modelo de regressão Weibull é adequado quando o tempo de falha é bem descrito através de uma distribuição de probabilidades Weibull. Para o modelo de regressão Weibull, $h_0(t)$ é dada por:

$$h_0(t) = \lambda p t^{p-1}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Consequentemente, a função risco de morte é dada por:

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) \lambda p t^{p-1}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Note, através da expressão acima, que o risco de morte não é constante, mas sim crescente ou decrescente ao longo do tempo (para $p \neq 1$). Neste caso, o efeito das covariáveis é modificar o crescimento ou decrescimento das funções taxas de falha.

A função de sobrevivência do modelo de regressão Weibull é dada por:

$$S(t) = \exp\left(-\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)\, \lambda t^p\right), \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Note que o modelo de regressão exponencial é um caso particular do modelo de regressão Weibull, quando $p = 1$.

5.1.3. Modelo de Regressão Gompertz

O modelo de regressão Gompertz é adequado quando o tempo de falha é bem descrito através de uma distribuição de probabilidades Gompertz. Para o modelo de regressão Gompertz, a função risco de morte é dada por:

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) \theta e^{\alpha t}, \quad \theta > 0, -\infty < \alpha < \infty \text{ e } t \geq 0.$$

De maneira análoga ao modelo de regressão paramétrico Weibull, o risco de morte é crescente ou decrescente ao longo do tempo, e o efeito das covariáveis é modificar este crescimento ou decrescimento das funções taxas de falha.

5.1.4. Adequabilidade da suposição de taxas de falhas proporcionais

Para utilizar os modelos de regressão de taxas de falhas proporcionais, a suposição de taxas de falhas proporcionais deve ser válida para os dados que estão sendo avaliados. A ideia básica é observar o que os dados dizem sobre a razão das taxas de falhas. Se os dados apresentam evidências que esta razão não é constante ao longo do

tempo, então os modelos de taxas de falhas proporcionais não são adequados, visto que eles iriam ajustar taxas de falha proporcionais. Há alguns métodos gráficos que podem ser utilizados para auxiliar na avaliação da validade da suposição e, para o modelo Weibull há, também, um teste estatístico.

O método gráfico baseia-se na comparação visual das funções taxa de falha para os níveis de cada covariável, ou equivalentemente, a comparação das funções taxa de falha acumuladas. Para esta comparação, uma ideia seria fazer o gráfico com as funções taxas de falha acumuladas para os níveis de cada covariável, e verificar se são proporcionais. Outra ideia, menos intuitiva mas que facilita a tomada de decisão, é transformar as funções taxas de falhas acumuladas utilizando a transformação $\ln\left(-\ln\left(\widehat{H}_i\right)\right)$, e fazer o gráfico desta nova função. Estes dois gráficos podem ser visualizados na Figura 27.

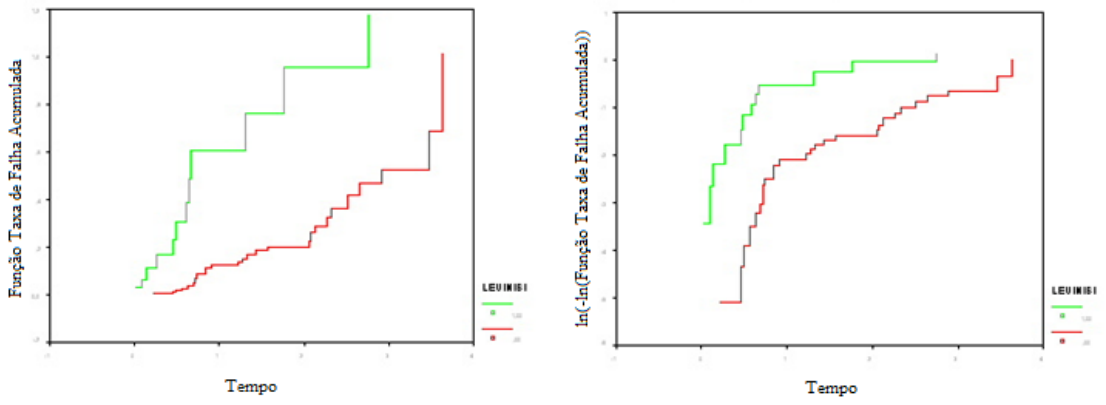

Figura 27: Gráficos para verificação da suposição de taxas de falhas proporcionais.

O primeiro gráfico da Figura 27 é um exemplo de um gráfico onde são utilizadas as funções taxas de falha acumuladas. Neste caso a suposição é válida porque as curvas são proporcionais, ou seja, as funções taxas de falha acumuladas vão se afastando de maneira proporcional com o passar do tempo. O segundo gráfico da Figura 27 é um exemplo de gráfico com a transformação $\ln\left(-\ln\left(\widehat{H}_i\right)\right)$. Neste caso a suposição é valida porque as curvas são paralelas, ou seja, a distância entre as duas curvas é constante ao longo do tempo.

Indica-se que estes gráficos sejam feitos para os níveis de cada covariável. Por exemplo, suponha que as covariáveis sexo (indica o sexo do paciente) e alergia (indica se o paciente é alérgico ou não a determinado medicamento) devem ser incluídas na

análise. Neste caso são necessários três gráficos: um comparando o grupo de homens com o grupo de mulheres, outro comparando o grupo de alérgicos com o grupo de não-alérgicos e um último comparando grupos formados pela combinação destas duas covariáveis, ou seja, homens alérgicos, homens não-alérgicos, mulheres alérgicas e mulheres não-alérgicas. Quando há covariáveis contínuas no modelo, indica-se categorizar a covariável (preferencialmente em poucas categorias) para que a suposição de taxas de falhas proporcionais possa ser verificada através destas categorias. Collet (2003) discute os métodos gráficos com detalhes. Ainda, é importante destacar que esta técnica gráfica deve ser utilizada com cuidado quando os estratos tiverem tamanhos muito pequenos.

Para verificar a validade da suposição para o modelo Weibull, há também um teste estatístico. Segundo Collett (2003), a suposição de taxas de falhas proporcionais para g grupos no modelo de regressão Weibull é equivalente a dizer que o parâmetro p da função taxa de falha basal $h_0(t)$ é o mesmo para cada grupo.

No caso das covariáveis sexo e alérgico, a ideia seria realizar o teste três vezes, uma considerando a covariável sexo, outra considerando a covariável alérgico e o terceiro considerando a combinação das duas. Por exemplo, para a covariável sexo ajusta-se 2 modelos ($g = 2$) de regressão de taxas de falhas proporcionais Weibull, um modelo para os homens e outro para as mulheres. Estes modelos provavelmente terão diferentes estimativas para os parâmetros λ e p, e, consequentemente, terão diferentes valores para o logaritmo da verossimilhança. Calcula-se então a soma dos valores da estatística $-2\log(verossimilhança)$ de cada um dos dois modelos, valor que chamaremos de A. Ajusta-se então o modelo de taxas de falhas proporcionais Weibull com todas as covariáveis (modelo saturado), obtendo-se o valor da estatística $-2\log(verossimilhança)$ deste modelo, que chamaremos de B. A diferença $(B - A)$ é então comparada com uma distribuição qui-quadrado com $(g - 1)$ graus de liberdade, ou seja, com 1 grau de liberdade para o exemplo. A hipótese nula deste teste é que a suposição de taxas de falhas proporcionais é válida. Então, $p - valores$ pequenos para este teste indicam evidências de que a suposição de taxas de falhas não é válida.

5.2. Modelos de Tempo de Vida Acelerado

Os modelos de tempo de vida acelerado são assim denominados porque o efeito das covariáveis é multiplicativo na escala do tempo (Hosmer e Lemeshow, 1999). Isto

quer dizer que o efeito das covariáveis é de acelerar ou desacelerar o tempo de sobrevivência (Colosimo e Giolo, 2006).

A expressão geral da função taxa de falha do indivíduo i no tempo t, dos modelos de tempo de vida acelerado, é dada por:

$$h_i(t) = \exp(-\eta_i) h_0\left(\frac{t}{\exp(\eta_i)}\right), \quad t \geq 0,$$

onde $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$, para p covariáveis.

Frequentemente são utilizados os modelos log-lineares da variável T, ou seja,

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma \varepsilon_i,$$

onde μ e σ são o intercepto e o parâmetro de escala, e ε_i é o componente aleatório que assume uma distribuição de probabilidades para cada modelo de regressão. Os valores estimados de $\alpha_1, \alpha_2, \ldots, \alpha_p$ refletem o efeito que cada covariável tem sobre o tempo de sobrevivência: valores positivos indicam que o aumento do valor da covariável ocasiona aumento no tempo de sobrevivência, e valores negativos indicam que o decréscimo no valor da covariável indica um decréscimo no valor do tempo de sobrevivência.

Os modelos regressão exponencial, Weibull e log-normal podem ser escritos segundo esta abordagem. Segundo Colosimo e Giolo (2006), poucas situações práticas são ajustadas adequadamente pelo modelo de regressão exponencial devido à simplicidade deste modelo. Portanto, somente os modelos Weibull e log-normal serão discutidos nas próximas seções.

5.2.1. Modelo de Regressão Weibull

O modelo de regressão Weibull é adequado quando o tempo de falha é bem descrito através de uma distribuição de probabilidades Weibull. Para o modelo de regressão Weibull, $h_0(t)$ é dada por:

$$h_0(t) = \lambda p t^{p-1}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Consequentemente, a função taxa de falha é dada por:

$$h_i(t) = \exp(-\eta_i)\lambda p(\exp(-\eta_i)t)^{p-1} = \exp(-\eta_i)^p\, \lambda p t^{p-1}, \quad t \geq 0 \text{ e } p,\ \lambda > 0.$$

Neste caso, o tempo de sobrevivência segue uma distribuição Weibull com os parâmetros $\lambda \exp(-p\eta_i)$ e p. Segundo Cleves *et al* (2008), a função de sobrevivência do modelo de regressão paramétrico Weibull é dada por:

$$S(t) = \exp\left\{-\left[\exp\left(-\alpha_0 - \boldsymbol{x}_j\boldsymbol{\alpha}_j\right) t_j\right]^p\right\}, \quad t > 0 \text{ e } p > 0.$$

Como o modelo de regressão de Weibull pode ser modelado segundo as duas abordagens (modelos de taxa de falhas proporcionais e modelos de tempo de vida acelerado), é possível relacionar os resultados das duas modelagens. Segundo Cleves *et al* (2008), tem-se que $\alpha_j = -\beta_j/p$, onde β_j são os coeficientes das covariáveis do modelo de taxa de falhas proporcionais de Weibull e α_j são os coeficientes do modelo de tempo de vida acelerado. Ainda, tem-se que $\sigma = p^{-1}$.

5.2.2. Modelo de Regressão Log-normal

O modelo de regressão log-normal é adequado quando o tempo de falha é bem descrito através de uma distribuição de probabilidades log-normal. Para o modelo de regressão log-normal, $S_0(t)$ é dada por:

$$S_0(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t \geq 0,\ \mu \in \Re \text{ e } \sigma > 0.$$

A função de sobrevivência do modelo de regressão log-normal é dada por:

$$S(t) = S_0(\exp(-\eta_i)\, t), \qquad t \geq 0$$

ou seja,

$$S(t) = 1 - \Phi\left(\frac{\log t - \eta_i - \mu}{\sigma}\right), \qquad t \geq 0,\ \mu \in \Re \text{ e } \sigma > 0.$$

Neste caso, o tempo de sobrevivência segue uma distribuição densidade de probabilidades log-normal, com parâmetros $\mu + \boldsymbol{\alpha}'\boldsymbol{x}_i$ e σ. O modelo log-linear deste modelo paramétrico segue uma distribuição densidade de probabilidades normal, como foi discutido na Seção 4.3 do Capítulo 4.

5.3. Seleção de Covariáveis

Ao ajustar um modelo de regressão, é fundamental que as covariáveis que compõem o modelo sejam relevantes, ou seja, a informação que a covariável contêm é importante para melhorar o ajuste do modelo. O número de covariáveis no modelo dita o número de estimativas necessárias, ou seja, quanto mais covariáveis há no modelo, mais parâmetros precisam ser estimados. Portanto, covariáveis que não são significativas e não são clinicamente importantes devem sair do modelo, visto que quanto mais parâmetros são estimados, piores são as estimativas.

Ajustar todos os modelos possíveis para encontrar o que melhor se ajusta aos dados pode não ser uma boa alternativa quando se trabalha com muitas covariáveis. Isto porque a quantidade de modelos pode ser muito grande, visto que se tem um modelo para cada uma das covariáveis, um modelo para cada par de covariáveis e assim por diante, até o modelo que estiver com todas covariáveis. Se as interações forem consideradas, o número de modelos será ainda maior.

Existem rotinas automáticas para seleção de modelos, como o método *forward*, *backward* ou *stepwise*. Estes métodos são utilizados para auxiliar na seleção das covariáveis que farão parte do modelo. Contudo, estes métodos só irão selecionar covariáveis estatisticamente significativas. Para modelos na área da saúde isto pode ser um problema, visto que podem existir covariáveis que não sejam estatisticamente significativas (talvez porque o tamanho da amostra não tenha sido suficiente), mas que são clinicamente importantes. Estas covariáveis devem, em geral, fazer parte do modelo. Portanto, será apresentada uma maneira simples de fazer a seleção de covariáveis e que permite uma participação maior do pesquisador.

O primeiro passo é fazer os modelos de regressão univariados, ou seja, um modelo de regressão para cada covariável: as covariáveis que possuírem $p - valor \leq 0{,}20$ serão inicialmente selecionadas. O segundo passo é fazer um modelo de regressão com todas as covariáveis previamente selecionadas e com as covariáveis que são clinicamente importantes que não foram significativas. No modelo com todas as

covariáveis, deve-se retirar as covariáveis de maior $p - valor$ e não significativas, uma de cada vez: um modelo novo é feito após cada retirada de covariável. Novamente, as variáveis clinicamente importantes não precisam ser retiradas do modelo e devem fazer parte dessas etapas da modelagem. Repete-se esse processo até que todas as covariáveis possuam $p - valor \leq 0{,}05$, com exceção das covariáveis clinicamente importantes. Quando todas covariáveis forem significativas, tem-se as covariáveis selecionadas para compor o modelo de regressão.

Neste processo, Colosimo e Giolo (2006) indicam que deve-se ajustar um modelo de regressão paramétrico geral, como o modelo gama generalizada, e utilizar seus resultados para fazer a seleção das covariáveis. Para os modelos exponencial, Weibull e log-normal, que são modelos encaixados da gama generalizada, esta forma de modelagem é intuitiva.

5.4. Escolha do Modelo de Regressão Paramétrico

A escolha do modelo de regressão paramétrico que melhor se ajusta aos dados é extremamente importante. Esta escolha pode ser feita de duas maneiras para os modelos de regressão: pode-se modelar somente o tempo de falha, selecionar o modelo paramétrico e verificar a qualidade do ajuste após a inclusão das covariáveis, através dos resíduos descritos na próxima seção, ou modelar diretamente os modelos de regressão e verificar a qualidade do ajuste com os resíduos.

Quando modela-se somente o tempo de falha, as técnicas descritas na Seção 4.5. para escolha do modelo paramétrico podem ser utilizadas.

Quando modela-se diretamente os modelos de regressão, deve-se inicialmente selecionar as covariáveis que farão parte do modelo. Com as covariáveis selecionadas, ajusta-se todos os modelos de regressão paramétricos desejáveis. Alguns métodos gráficos descritos na Seção 4.5.1. podem ser utilizados na escolha do modelo de regressão paramétrico adequado, como por exemplo a comparação das funções taxa de falha acumulada estimada por Kaplan-Meier e pelo modelo de regressão paramétrico proposto e a linearização dos modelos paramétricos. O teste de modelos encaixados também pode ser utilizado, embora este teste não possa ser utilizado para o modelo de regressão Gompertz. Os critérios de informação são a alternativa mais completa porque permitem comparar todos os modelos de regressão paramétrica sem a subjetividade da

análise gráfica: quanto menor for o valor do AIC e do BIC, melhor é o ajuste do modelo aos dados.

5.5. Adequação do Modelo Ajustado

A adequabilidade do modelo de regressão selecionado deve ser verificada. Isto é fundamental, porque os resultados somente serão válidos se o modelo selecionado estiver bem ajustado. Análises gráficas de resíduos são geralmente utilizadas para este fim. O maior potencial desse tipo de análise é apontar falhas nas suposições, embora elas também sejam úteis na detecção de observações atípicas.

Colosimo e Giolo (2006) propõem a análise de quatro resíduos: resíduos de Cox-Snell, resíduos padronizados, resíduos *martingal* e resíduos d*eviance*. Os dois primeiros resíduos são utilizados para analisar o ajuste global do modelo, o terceiro é utilizado para determinar a forma funcional de uma covariável contínua e o quarto é útil quando se examina a acurácia do modelo. Pode-se encontrar em Collett (2003) outras opções de resíduos para serem analisados.

5.5.1. Qualidade Geral do Ajuste

Os resíduos de Cox-Snell e os resíduos padronizados são úteis para avaliar a qualidade geral do ajuste do modelo de regressão paramétrico.

Resíduos de Cox-Snell

Os resíduos de Cox-Snell são definidos por:

$$\hat{e}_i = \widehat{H}(t_i) = -\log\left(\hat{S}(t_i)\right),$$

onde $\widehat{H}(t_i)$ é a estimativa da função taxa de falha acumulada e $\hat{S}(t_i)$ é a estimativa da função de sobrevivência do indivíduo i.

Se o modelo paramétrico proposto ajusta-se bem aos dados, então a distribuição dos resíduos $\hat{e}_i$ deve ser exponencial. Para verificar se os resíduos tem distribuição exponencial, segue-se os seguintes passos:

1. Calcula-se os resíduos de Cox-Snell;

2. Ajusta-se a função de sobrevivência utilizando Kaplan-Meier para os resíduos;
3. Ajusta-se a função de sobrevivência utilizando o modelo exponencial para os resíduos;
4. Faz-se um gráfico com curvas de sobrevivência ajustadas em 2. e 3.. Neste gráfico, curvas próximas indicam que a distribuição dos resíduos $\hat{e}_i$ deve ser exponencial;
5. Faz-se um gráfico com a curva ajustada em 2. *versus* a curva ajustada em 3.. Se este gráfico gerar uma reta, há indícios de que a distribuição dos resíduos $\hat{e}_i$ deve ser exponencial.

Colosimo e Giolo (2006) destacam que para os casos em que o modelo exponencial ou o modelo Weibull são utilizados e o número de censuras for pequeno, é necessário ajustar os resíduos de Cox-Snell para os indivíduos censurados. Ou seja, os resíduos de Cox-Snell para os indivíduos censurados são calculados por:

$$\hat{e}_i = \widehat{H}(t_i) + 1.$$

Uma limitação da análise dos resíduos de Cox-Snell é que a conclusão de que o modelo não está bem ajustado pode ser devido ao fato de se utilizar as estimativas para os parâmetros do modelo exponencial (passo 3) e não o seu verdadeiro valor. Ou seja, o resultado pode ser um falso negativo. Em resumo, esta análise é sensível para tamanhos de amostras pequenos.

Resíduos Padronizados

Os resíduos padronizados são definidos por:

$$\hat{v}_i = \frac{\left(\log(t_i) - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i} - \cdots - \hat{\alpha}_p x_{pi}\right)}{\hat{\sigma}}.$$

Note que se utiliza o logaritmo dos tempos, remetendo a representação dos modelos log-lineares. Desta forma, o modelo de regressão exponencial ou o modelo de regressão Weibull ajusta-se bem aos dados quando os resíduos seguem a distribuição do

valor extremo padrão. Analogamente, o modelo de regressão log-normal ajusta-se bem aos dados se os resíduos seguem distribuição normal padrão.

O modelo de regressão exponencial é considerado adequado quando o gráfico com a curva de sobrevivência dos resíduos estimados por Kaplan-Meier *versus* os valores obtidos utilizando-se a distribuição do valor extremo padrão forem uma reta. A mesma análise gráfica é utilizada para o modelo de regressão Weibull. O modelo de regressão log-normal pode ser considerado adequado quando o gráfico de probabilidade normal dos resíduos padronizados estimados for aproximadamente uma reta.

5.5.2. Forma Funcional das covariáveis

Resíduos Martingal

Os resíduos *martingal* são geralmente utilizados para analisar a forma funcional das covariáveis. Eles são definidos por:

$$\hat{m}_i = \delta_i - \hat{e}_i,$$

onde δ_i é a variável indicadora de censura e $\hat{e}_i$ são os resíduos de Cox-Snell.

Para ter indícios de qual a melhor maneira que cada covariável contínua deve entrar no modelo, pode-se fazer um diagrama de dispersão da covariável analisada com os resíduos *martingal*. Por exemplo, se o diagrama de dispersão da covariável x_1 com os resíduos *martingal* apresentar uma relação linear, não será necessário fazer transformação na covariável x_1. Contudo, se o diagrama de dispersão apresentar a forma de uma parábola, então há indícios de que é necessário acrescentar um termo quadrático da variável x_1 no modelo de regressão.

5.5.3. Acurácia do Modelo

Resíduos Deviance

Os resíduos *deviance* são geralmente utilizados para detectar pontos atípicos. Eles são definidos por:

$$\hat{d}_i = sinal(\hat{m}_i)[-2(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i))]^{1/2},$$

onde $\hat{m}_i$ são os resíduos *martingal* e δ_i é a variável indicadora de falha. Estes resíduos tentam tornar os resíduos *martingal* mais simétricos em torno de zero, o que facilita a detecção de *outliers*.

Se os resíduos *deviance* apresentarem comportamento aleatório, então o modelo é apropriado. Isto pode ser verificado através do gráfico dos resíduos *deviance versus* os tempos. Este gráfico também é uma boa ferramenta para detectar a presença de *outliers*.

5.6. Interpretação

Os resultados dos modelos de regressão paramétricos devem ser cuidadosamente interpretados. Há diferença na interpretação dos coeficientes dos modelos de taxas de falhas proporcionais e dos modelos de tempo de vida acelerado.

Nos modelos de taxa de falha proporcionais, o valor da função matemática exponencial aplicada no valor do coeficiente, ou seja, $\exp(\beta_j)$, é o valor da razão das taxas de falhas. Como consequência da suposição de taxas de falhas proporcionais, tem-se que esta razão de riscos será constante ao longo do tempo.

Se a variável x_l for, por exemplo, a variável que indica o tratamento que o paciente recebeu, onde pacientes pertencentes ao grupo 0 receberam o tratamento padrão e pacientes pertencentes ao grupo 1 receberam o novo tratamento, o risco de morte para os pacientes que receberam o novo tratamento é $\exp\{\hat{\beta}_l\}$ vezes o risco de morte de pacientes que receberam o tratamento padrão, mantendo-se as demais variáveis constantes.

É interessante utilizar intervalo de confiança (IC) como uma medida para completar a informação fornecida pela razão das taxas de falha. O intervalo de confiança para a razão de taxas de falhas é encontrado aplicando a função matemática exponencial no intervalo de confiança do coeficiente $\hat{\beta}_l$.

Se o intervalo de confiança contém o valor 1, então não há evidências estatísticas de que o risco de morte para o grupo 1 e para o grupo 0 sejam diferentes, com determinada confiança. Consequentemente, se o intervalo de confiança não contém o valor 1, então o risco de morte para pacientes do grupo 0 e do grupo 1 são estatisticamente diferentes, com a confiança que foi previamente estabelecida.

Note que se a razão das taxas de falha for 1, então o risco de morte dos pacientes do grupo 1 é uma vez o risco de morte dos pacientes do grupo 0, ou seja, o risco de morte dos pacientes do grupo 1 e do grupo 0 são iguais. Se o valor de $\exp\{\hat{\beta}_l\}$ for maior que 1, o grupo 1 tem risco de morte maior que o grupo 0, e se $\exp\{\hat{\beta}_l\}$ for menor que 1, o grupo 1 tem risco de morte menor que o risco de morte do grupo 0.

Por exemplo, em um modelo de regressão com uma única covariável, onde o grupo 0 recebeu o tratamento padrão e o grupo 1 recebeu o novo tratamento, tenham gerado as seguintes estimativas: $\exp\{\hat{\beta}_1\} = 3$ e $IC95\% = (2{,}5; 4)$. Sabe-se, portanto, que o risco de morte de pacientes que receberam o novo tratamento é 3 vezes o risco dos pacientes que receberam o tratamento padrão, com $IC95\% = (2{,}5; 4)$. A assimetria do intervalo de confiança é dada porque constrói-se o intervalo de confiança para a razão das taxas de falhas aplicando-se a função matemática exponencial no intervalo de confiança para $\hat{\beta}_1$.

Quando a covariável é dicotômica, como foi o caso do exemplo acima, a categoria de referência será sempre o grupo 0. Portanto, a estimativa do coeficiente fornece quantas vezes é o risco do grupo 1 em relação ao risco do grupo 0. Deve-se ter atenção ao definir o grupo de referência para as interpretações sejam no sentido desejado.

Se a covariável for categórica, cria-se $número\ de\ categorias - 1$ variáveis indicadoras. É preciso definir uma categoria como grupo de referência - para esta categoria não é necessário a criação da variável indicadora. Os coeficientes estimados irão expressar a relação entre a categoria que está representada na variável indicadora e a categoria de referência.

Por exemplo, suponha que existem três tratamentos: tratamento padrão, tratamento A e tratamento B. Define-se a categoria de referência como o tratamento padrão. Cria-se então as variáveis indicadoras x_1 e x_2: a variável x_1 recebe 1 se o paciente recebeu o tratamento A e 0 se não recebeu o tratamento A, e a variável x_2 recebe 1 se o paciente recebeu o tratamento B e 0 se não recebeu o tratamento B. Então, o risco de morte de pacientes que receberam o tratamento A é $\exp(\hat{\beta}_1)$ vezes o risco dos que receberam o tratamento padrão e o risco de morte de pacientes que receberam o tratamento B é $\exp(\hat{\beta}_2)$ vezes o risco dos pacientes que receberam o tratamento padrão.

Se a covariável for contínua, a interpretação do coeficiente altera-se. Suponha que idade é a covariável do modelo de regressão, então o aumento de 1 ano de idade aumenta o risco de morte em $\left(1-\exp\left(\hat{\beta}_1\right)\right)100\%$.

Para os modelos de tempo de vida acelerado, $\exp\left(\alpha_j\right)$ representa a razão dos tempos medianos e não da função taxa de falha (Colosimo e Giolo, 2006). A interpretação segue a mesma lógica dos modelos de taxa de falha proporcionais, contudo, quando se fala em razão das taxas de falha nos modelos de taxa de falha proporcionais, nos modelos de tempo de vida acelerado deve-se falar em razão de tempos medianos.

5.7. Exemplo

Nesta seção, os dados apresentados na Seção 2.3 serão utilizados para desenvolver um exemplo utilizando a metodologia dos modelos de regressão paramétricos. As covariáveis `inter` e `sexo` serão utilizadas na modelagem, ou seja, participação do processo de seleção de covariáveis e, caso sejam selecionadas, farão parte dos modelos de regressão.

Quando se trabalha com os modelos de regressão paramétricos pode-se conduzir a análise seguindo os seguintes passos:

1. Seleção de covariáveis;
2. Escolha do modelo de regressão paramétrico
    a. Verificação da suposição de taxas de falhas proporcionais. Se rejeitar a suposição segue para o passo c.;
    b. Escolha entre as distribuições exponencial, Weibull e Gompertz para o modelo de taxas de falhas proporcionais e segue para 3.;
    c. Escolha entre as distribuições Weibull e log-normal para os modelos de tempo de vida acelerado e segue para 3..
3. Adequação do modelo ajustado;
4. Interpretação e extrapolação.

O *software* STATA versão 11 é utilizado para fazer as análises.

5.7.1. Seleção de Covariáveis

A seleção das covariáveis que farão parte do modelo de regressão deve ser a primeira etapa da modelagem, porque este resultado pode ser utilizado tanto para os modelos de taxas de falha proporcionais como para os modelos de tempo de vida acelerado. Como discutido na Seção 5.3., o modelo paramétrico gama generalizado é utilizado para ajustar os modelos e analisar a significância das covariáveis.

A primeira etapa do método de seleção de covariáveis é ajustar os modelos univariados. A programação utilizada para criar o modelo paramétrico gama generalizado com a covariável `inter` é:

```
streg inter, dist(gamma)
```

onde a covariável deve ser declarada antes da vírgula. A resposta fornecida pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       inter |  -.5855586   .1817404    -3.22   0.001    -.9417633    -.229354
       _cons |   8.443597   .1786096     47.27   0.000     8.093529    8.793666
-------------+----------------------------------------------------------------
     /ln_sig |  -.2133276   .2548146    -0.84   0.402     -.712755    .2860998
     /kappa  |   .9824737   .4192358     2.34   0.019     .1607866    1.804161
-------------+----------------------------------------------------------------
       sigma |   .8078914   .2058625                      .4902916    1.331225
------------------------------------------------------------------------------
```

A covariável `inter` tem $p - valor = 0{,}001$, ou seja, há evidências estatísticas de que o coeficiente estimado da covariável `inter` é diferente de zero. Esta covariável possui $p - valor \leq 20$ e, portanto, é selecionada para participar da próxima etapa da seleção de covariáveis.

A programação utilizada para criar o modelo paramétrico gama generalizado com a covariável `sexo` é:

```
streg sexo, dist(gamma)
```

onde a resposta fornecida pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |   .3066395   .1929836     1.59   0.112    -.0716014    .6848803
       _cons |   7.958694   .1349303    58.98   0.000     7.694236    8.223153
-------------+----------------------------------------------------------------
      /ln_sig|  -.0571274   .2267914    -0.25   0.801    -.5016304    .3873757
      /kappa |   .7006693   .3831102     1.83   0.067    -.0502129    1.451551
-------------+----------------------------------------------------------------
       sigma |   .9444738   .2141986                      .6055426     1.47311
------------------------------------------------------------------------------
```

A covariável `sexo` tem $p - valor = 0{,}112$, ou seja, possui $p - valor \leq 20$ e, portanto, é selecionada para participar da próxima etapa da seleção de covariáveis.

Com as regressões univariada realizadas, o próximo passo é fazer um modelo de regressão com as covariáveis selecionadas, `inter` e `sexo`. A programação utilizada para criar o modelo paramétrico gama generalizado com as covariáveis `sexo` e `inter` é:

```
streg sexo inter, dist(gamma)
```

onde a resposta fornecida pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |   .2918057    .189323     1.54   0.123    -.0792607     .662872
       inter |  -.5830034   .1816287    -3.21   0.001    -.9389891   -.2270176
       _cons |   8.337966   .1852239    45.02   0.000     7.974933    8.700998
-------------+----------------------------------------------------------------
      /ln_sig |-.1820484   .2295955    -0.79   0.428    -.6320473    .2679505
      /kappa |   .9173383   .3781489     2.43   0.015       .17618    1.658497
-------------+----------------------------------------------------------------
       sigma |    .833561   .1913818                      .5315026    1.307282
------------------------------------------------------------------------------
```

A covariável `inter` apresentou $p - valor = 0{,}001$ e a covariável `sexo` apresentou $p - valor = 0{,}123$, ou seja, a covariável `sexo` deveria ser retirada do modelo porque apresentou $p - valor \geq 0{,}05$. Contudo, esta covariável é clinicamente importante e, por este motivo, opta-se por deixá-la no modelo.

Com isso, tem-se que as duas covariáveis testadas, `inter` e `sexo`, são selecionadas para fazer parte dos modelos de regressão que serão ajustados.

Uma observação importante que pode ser feita sobre seleção de covariáveis é que, em casos com poucas covariáveis como este exemplo, poderiam ter sido ajustados todos os modelos e comparado os valores do AIC e BIC para escolher o modelo. Os

métodos de seleção de covariáveis são mais interessantes quando há mais covariáveis, contudo ele foi utilizado neste exemplo para que a dinâmica do método fosse entendida.

5.7.2. Escolha do Modelo de Regressão Paramétrico

A escolha do modelo de regressão paramétrico é uma etapa muito importante da modelagem, visto que deseja-se obter o modelo que melhor se ajusta aos dados. Esta escolha está diretamente relacionada com a abordagem que será utilizada, uma vez que os modelos exponencial, Weibull e Gompertz são modelados segundo a abordagem de taxas de falhas proporcionais e os modelos Weibull e log-normal são modelados segundo a abordagem de tempo de vida acelerado.

Os modelos de taxas de falhas proporcionais podem ser mais fáceis de entender, uma vez que estes modelos são semelhantes ao conhecido modelo de Cox. Logo, uma vantagem dessa abordagem é que se interpretam os coeficientes como razão de taxas de falhas proporcionais, também chamado de *hazard ratio*. Por este motivo, inicialmente tenta-se modelar os dados segundo a abordagem de modelos de taxas de falhas proporcionais, onde deve-se verificar a validade da suposição de taxas de falhas proporcionais. Caso a suposição seja válida, essa abordagem é utilizada e a escolha do modelo paramétrico será entre os modelos exponencial, Weibull e Gompertz. Somente se a suposição não for válida ajustam-se os modelos de tempo de vida acelerado, quando deve-se escolher entre os modelos Weibull e log-normal.

Verificação da Suposição de Taxas de Falhas Proporcionais

A verificação da suposição de taxas de falhas proporcionais para as covariáveis que farão parte do modelo de regressão é o primeiro passo quando se deseja ajustar os modelos de taxas de falhas proporcionais. Portanto, a suposição de taxas de falhas proporcionais deve ser testada para homens e mulheres, e pacientes que já foram internados pelo menos uma vez e pacientes que nunca foram internados.

Para testar esta suposição no STATA serão utilizados os comandos do modelo de regressão de Cox. O comando `stphplot` cria um gráfico onde linhas paralelas indicam que a suposição de taxas de falhas proporcionais não é violada. As curvas deste gráfico são referentes as funções de sobrevivência estimadas para os níveis das covariáveis e não a função taxa de falha acumulada, como é mais usual. Ou seja, este gráfico não irá gerar as curvas apresentadas na Seção 5.1.4., mas sim um gráfico

equivalente que pode ser utilizado sem prejuízo para a análise. A programação utilizada para fazer este gráfico no STATA para a covariável `inter` é:

```
stphplot, by(inter) xtitle(Logaritmo do tempo)
ytitle(log(log(probabilidade de sobrevivência))) legend(label(1 Nunca
Internou) label(2 Internou))
```

O comando `stphplot` cria o gráfico que testa a suposição de taxas de falhas proporcionais, neste caso para a variável `inter` porque utilizou-se `by(inter)`. A programação acima gera uma figura muito próxima da Figura 28, contudo aparecem bolas sobre as linhas e as duas linhas são contínuas, opções que serão alteradas via Menu para que o gráfico fique mais claro. Com o gráfico habilitado para edição, dá-se um duplo-clique sobre a linha da função, onde irá abrir uma janela na qual deve-se ir para a aba *Markers* e, na opção *Symbol* altera-se de *Circle* para *Point*. Esta alteração deve ser feita para as duas linhas e, para uma das linhas deve-se alterar o estilo da linha para pontilhado (a maneira com que altera-se o estilo da linha foi explicado no exemplo da Seção 4.6.).

A Figura 28 mostra que as curvas se cruzam, logo, os dados apresentam evidências de que a suposição não é válida. Ou seja, não é indicado utilizar a covariável `inter` nos modelos com abordagem de taxas de falhas proporcionais.

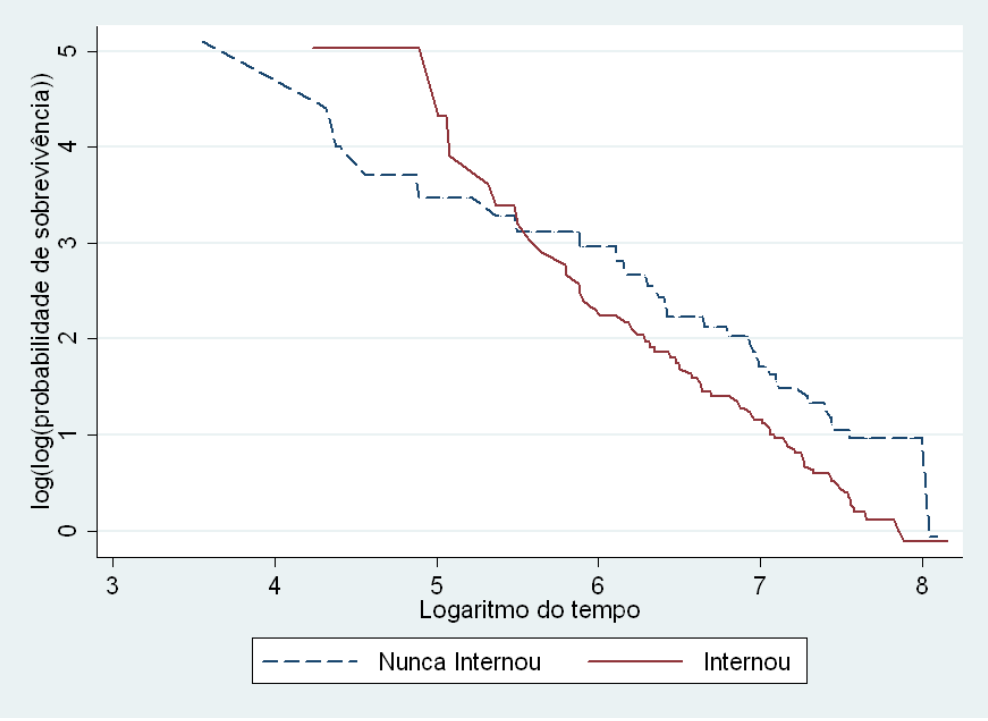

Figura 28: Verificação da suposição de taxas de falha proporcionais para a covariável inter.

A programação utilizada no STATA para testar a suposição de taxas de falhas proporcionais para a covariável `sexo` é:

```
stphplot, by(sexo) xtitle(Logaritmo do tempo)
ytitle(log(log(probabilidade de sobrevivência))) legend(label(1 Homem)
label(2 Mulher))
```

A Figura 29 mostra que as curvas para a covariável sexo não se cruzam, mas praticamente se tocam quando o logaritmo do tempo é aproximadamente igual a 5,2. Além disso, as curvas são muito próximas, o que pode ser explicado pela não significância da covariável, que indica que os grupos não diferem significativamente.

Ainda, como esta covariável não foi significativa e a covariável `inter` não pode ser utilizada pela falha da suposição, construir um modelo de regressão de taxas de falhas proporcionais somente com a covariável `sexo` não parece adequado. Portanto, indica-se não utilizar modelos de taxas de falhas proporcionais.

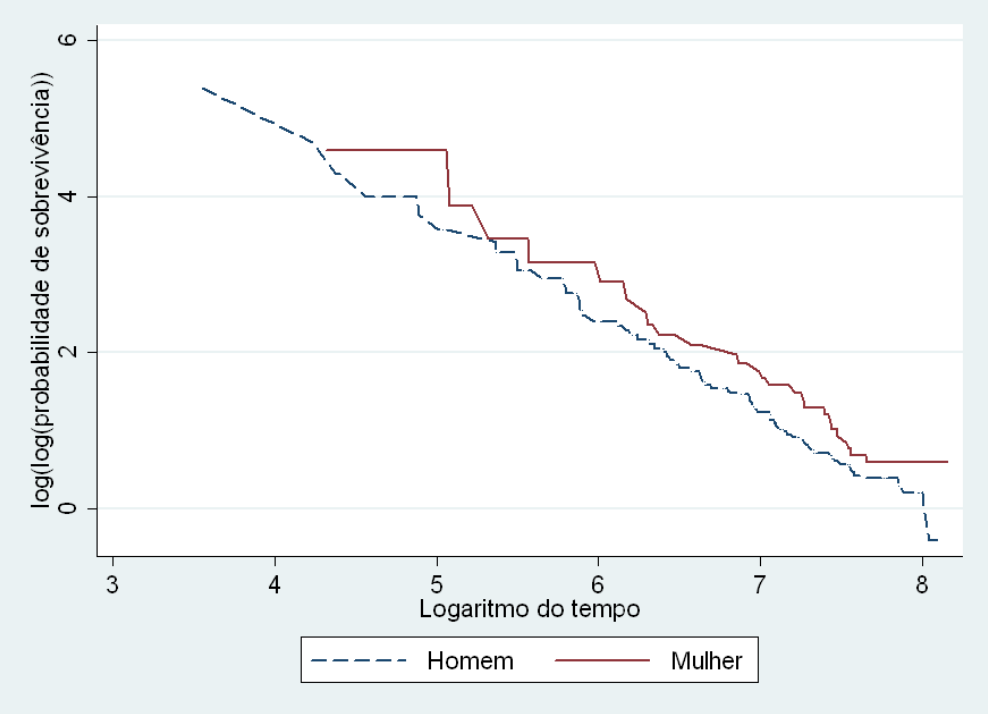

Figura 29: Verificação da suposição de taxas de falha proporcionais para a covariável sexo.

Escolha entre as Distribuições Weibull e Log-Normal para os Modelos de Tempo de Vida Acelerado

Para a escolha do modelo paramétrico de regressão adequado aos dados, segundo abordagem de modelos de tempo de vida acelerado, inicia-se ajustando os modelos Weibull e log-normal.

Ajuste do Modelo Weibull

A programação utilizada para ajustar o modelo de regressão Weibull no STATA, sob a abordagem de tempo de vida acelerado, com as covariáveis `inter` e `sexo` é:

```
streg inter sexo, dist(weib) time
```

Note que ao final deste comando é utilizado o comando `time`. Como o modelo de regressão paramétrico Weibull pode ser ajustado segundo as duas abordagens, é o comando `time` que indica qual abordagem está sendo utilizada. Quando este comando é utilizado, o modelo é ajustado segundo modelos de tempo de vida acelerado, enquanto que se este comando não for utilizado, o modelo será ajustado segundo a abordagem de taxas de falhas proporcionais.

Ajuste do Modelo Log-Normal

A programação utilizada para ajustar o modelo de regressão log-nomal no STATA, sob a abordagem de tempo de vida acelerado, com as covariáveis `inter` e `sexo` é:

```
streg inter sexo, dist(lnormal)
```

Note que o comando `time` não é utilizado para o modelo log-normal porque este modelo só pode ser ajustado segundo a abordagem de modelos de tempo de vida acelerado.

Métodos Gráficos

Alguns métodos gráficos podem ser utilizados para auxiliar na escolha ou rejeição do modelo regressão paramétrico proposto. O primeiro método gráfico compara

as funções de sobrevivência estimadas pelo modelo paramétrico proposto e por Kaplan-Meier, através do gráfico de uma curva *versus* a outra.

A programação utilizada para construir este gráfico no STATA para o modelo de regressão Weibull é:

```
sts generate km = s
quietly streg inter sexo, dist(weib) time
predict s_w, surv
label variable s_w "Weibull"
scatter km s_w, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Weibull)
```

A programação utilizada para construir este gráfico no STATA para o modelo de regressão log-normal é:

```
quietly streg inter sexo, dist(lnormal)
predict s_ln, surv
label variable s_ln "Log-normal"
scatter km s_ln, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier) ytitle(Log-normal)
```

A Figura 30 e a Figura 31 apresentam a função de sobrevivência estimada por Kaplan-Meier *versus* a estimada pelo modelo de regressão paramétrio Weibull e log-normal, respectivamente. Para que o ajuste do modelo esteja adequado, é necessário observar quatro retas de pontos. Para estes dados, apesar de não serem retas perfeitas, pode-se considerar que há quatro retas de pontos para os dois modelos de regressão paramétricos testados.

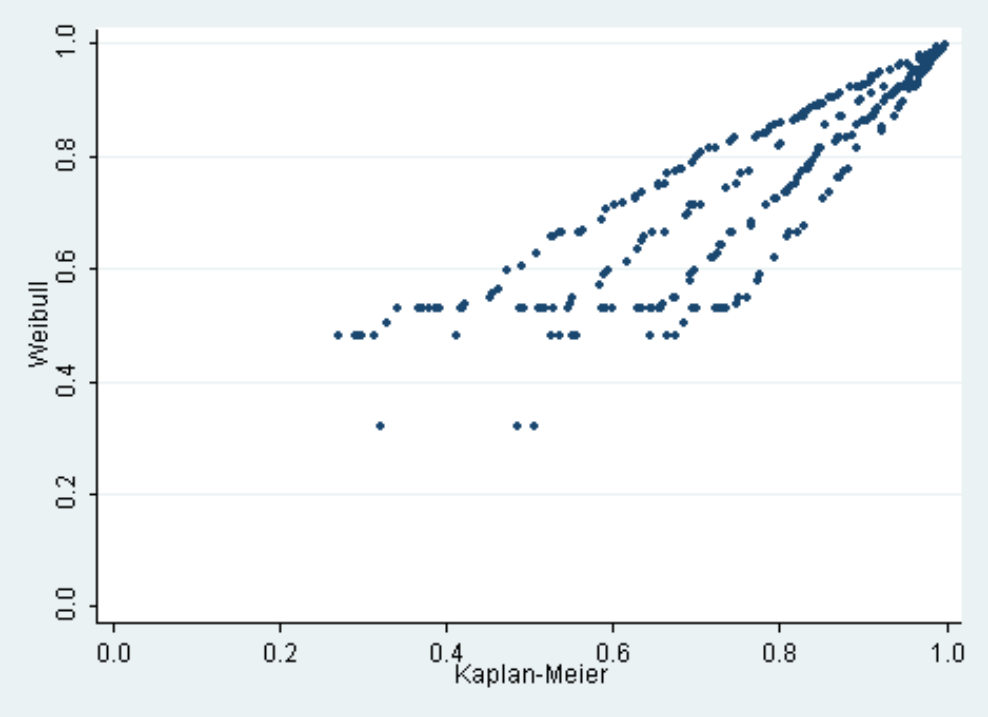

Figura 30: Curva de sobrevivência estimada por Kaplan-Meier *versus* curva de sobrevivência estimada pelo modelo Weibull.

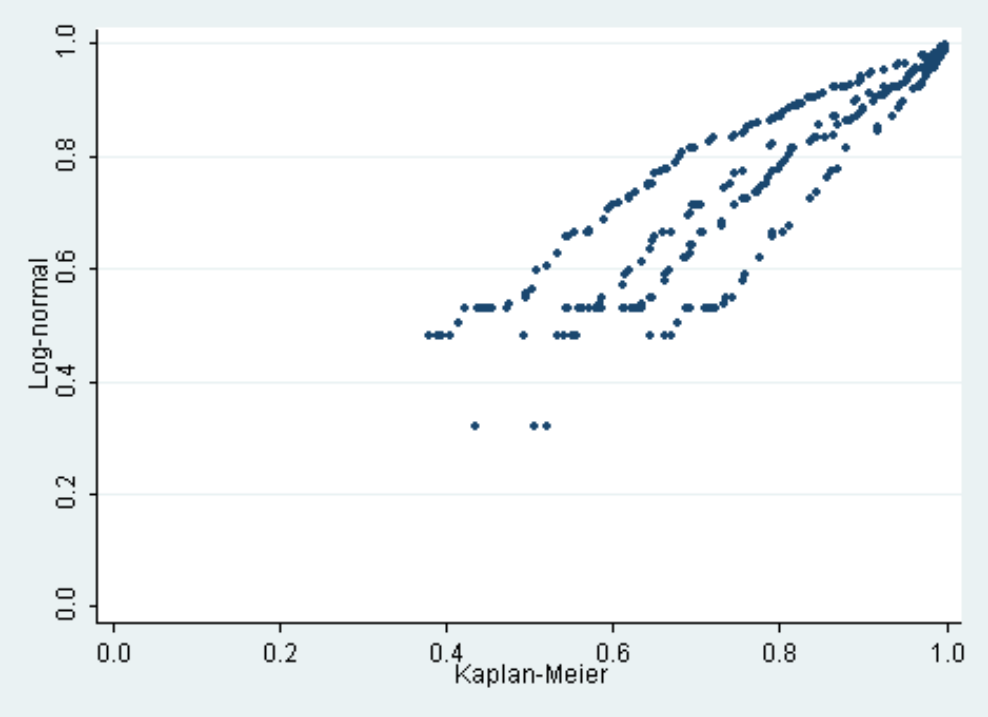

Figura 31: Curva de sobrevivência estimada por Kaplan-Meier versus curva de sobrevivência estimada pelo modelo log-normal.

Pode-se então concluir que nenhum dos dois modelos de regressão paramétricos ajustou-se tão mal aos dados a ponto deste método gráfico excluir a possibilidade de sua utilização.

Uma variação deste método gráfico é comparar a função taxa de falha acumulada estimada pelo método de Kaplan-Meier com a estimada pelo modelo paramétrico proposto, através de um gráfico com estas duas curvas. A programação utilizada para construir este gráfico no STATA para o modelo de regressão Weibull é:

```
generate double H = -ln(km)
label variable H "Kaplan-Meier"
quietly streg inter sexo, dist(weib) time
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada) xtitle(Tempo)
ytitle(Taxa de falha acumulada)
```

A programação utilizada para construir este gráfico no STATA para o modelo de regressão log-normal é:

```
quietly streg inter sexo, dist(lnorm)
stcurve, cumhaz addplot(line H _t, sort) ylabels(0 0.5 1 1.5,
format(%9.1f)) title(Função Taxa de Falha Acumulada) xtitle(Tempo)
ytitle(Taxa de falha acumulada)
```

A Figura 32 apresenta as funções taxa de falha acumuladas estimadas por Kaplan-Meier e pelo modelo de regressão paramétrico Weibull. As curvas são, de maneira geral, sobrepostas. Nos tempos inicias a sobreposição ocorre quase que perfeitamente, enquanto que para os últimos tempos há alguns pontos em que as curvas são um pouco distantes.

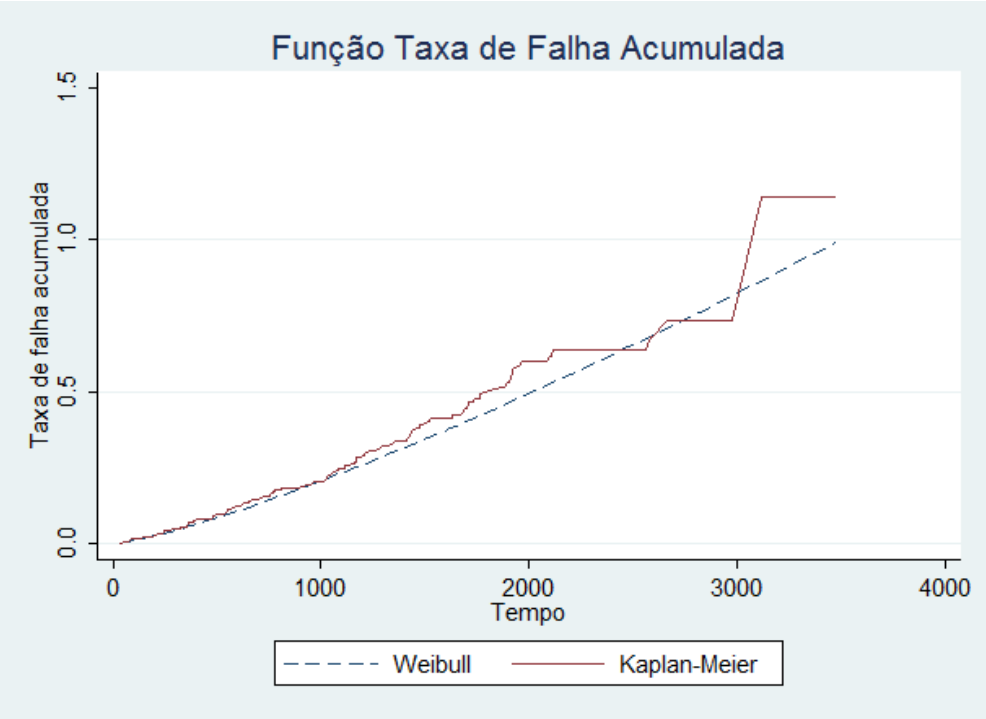

Figura 32: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo Weibull.

A Figura 33 apresenta as funções taxa de falha acumuladas estimadas por Kaplan-Meier e pelo modelo de regressão paramétrico log-normal. As curvas sobrepõem-se para os tempos inicias e para os últimos tempos há pontos distantes.

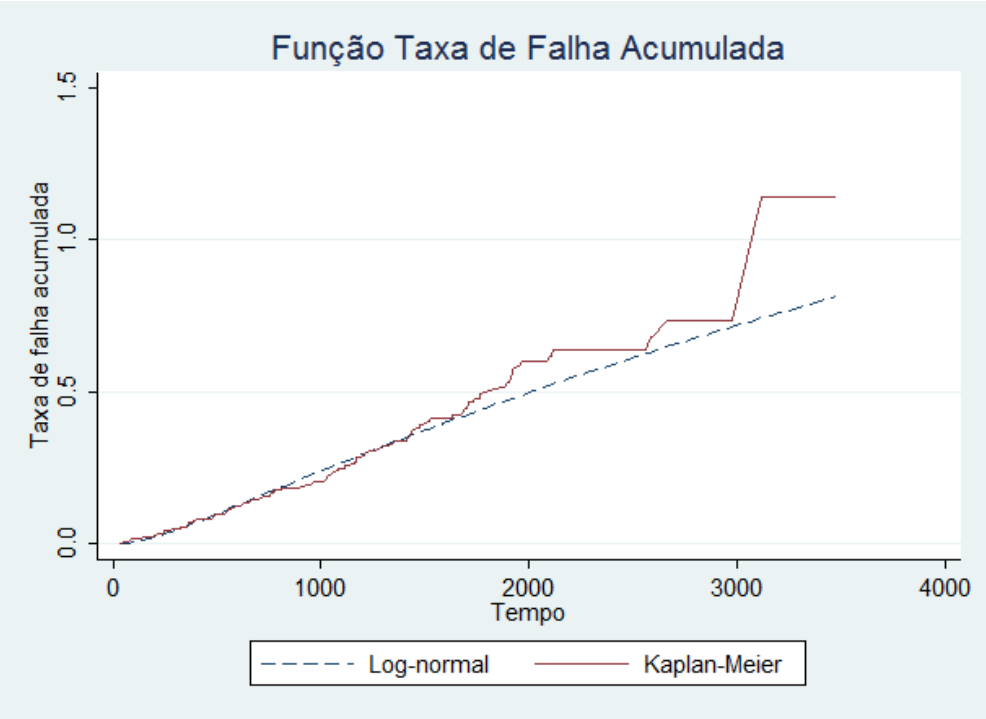

Figura 33: Funções taxa de falha acumulada estimadas por Kaplan-Meier e pelo modelo log-normal.

Através dos gráficos apresentados, observa-se que o comportamento dos dois modelos de regressão paramétricos são semelhantes, uma vez que os dois estimaram a função taxa de falha acumulada muito próxima da estimada por Kaplan-Meier. Pode-se então concluir que não houve algum modelo de regressão paramétrico que tenha se ajustado tão mal aos dados a ponto deste método gráfico excluir a possibilidade de sua utilização.

Outro método gráfico refere-se a linearização da curva de sobrevivência dos modelos de regressão paramétricos.

A programação utilizada para construir este gráfico no STATA para o modelo de regressão Weibull é:

```
generate lnt = ln(_t)
generate lnlns = ln(-ln(km))
scatter lnlns lnt, msize(small) xtitle("ln(t)") ytitle("log(-
log(S(t)))")
```

A programação utilizada para construir este gráfico no STATA para o modelo de regressão log-normal é:

```
generate norms = invnormal(km)
scatter norms lnt, msize(small) xtitle("ln(t)")
ytitle("invnormal(S(t))")
```

A Figura 34 e a Figura 35 apresentam os gráficos resultantes da linearização do modelo de regressão paramétrico Weibull e log-normal, respectivamente. O comportamento do gráfico para estes dois modelos é semelhante, uma vez que a sequência de pontos de cada um destes gráficos pode ser considerada uma reta.

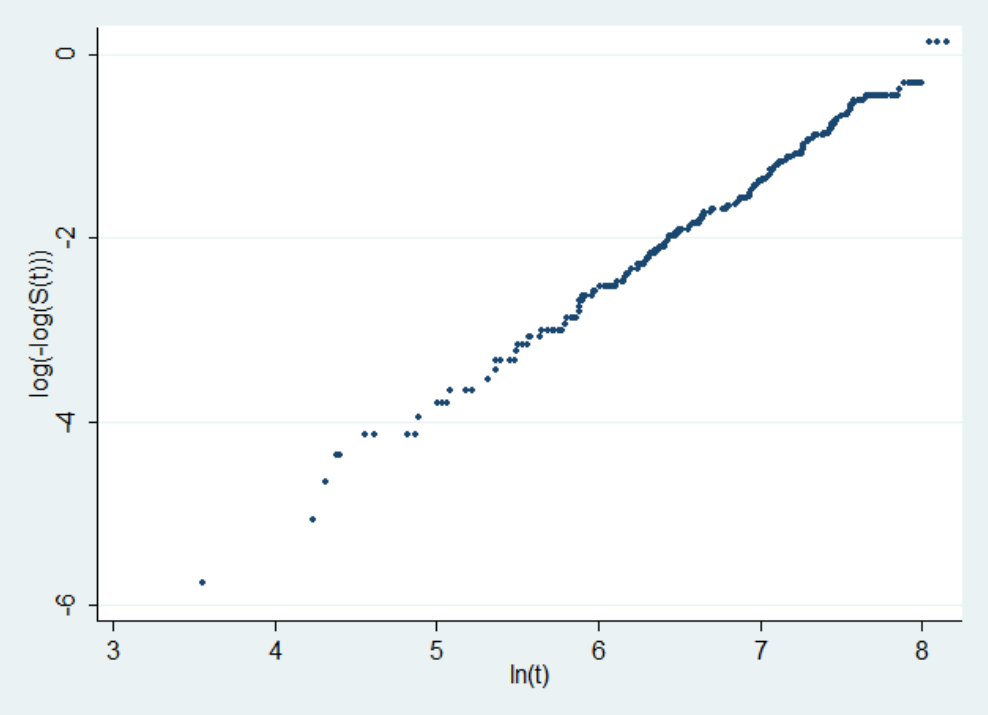

Figura 34: Linearização do modelo Weibull.

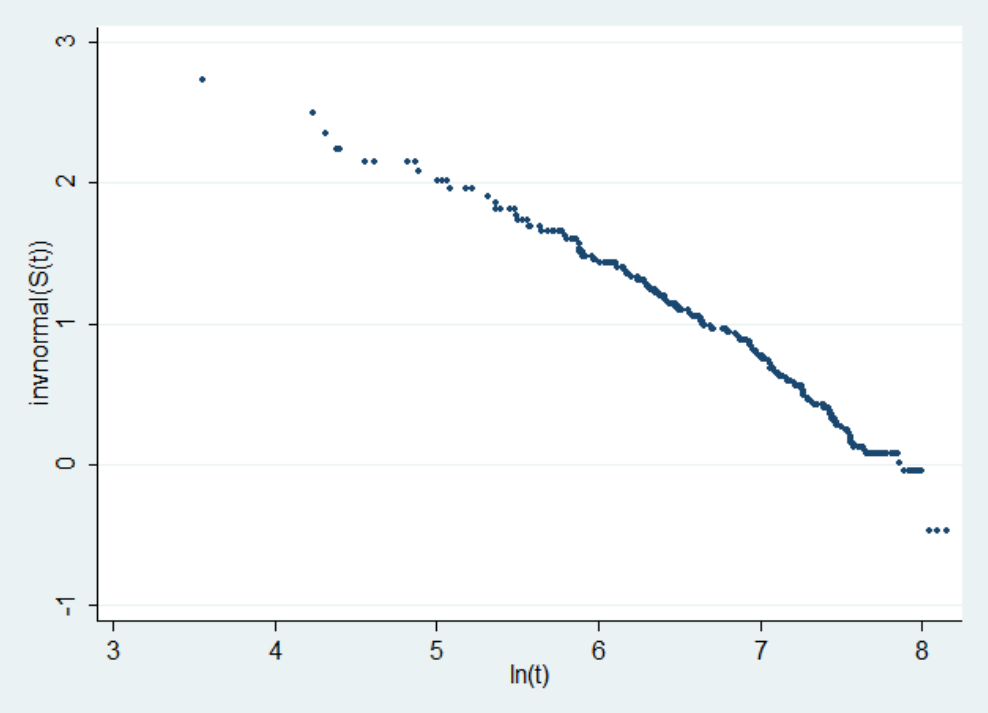

Figura 35: Linearização do modelo log-normal.

Através dos gráficos apresentados, observa-se que nenhum modelo de regressão paramétrico apresentou problemas na linearização. Pode-se então concluir que não

houve algum modelo de regressão paramétrico que tenha se ajustado tão mal aos dados a ponto deste método gráfico excluir a possibilidade de sua utilização.

Teste dos Modelos Encaixados

O teste para modelos encaixados pode ser executado para os modelos de regressão Weibull e log-normal. Para executar este teste é necessário também ajustar o modelo de regressão gama generalizado.

A programação utilizada no STATA para ajustar o modelo de regressão gama generalizado e armazenar o valor do logaritmo da verossimilhança do ajuste é:

```
streg inter sexo, dist(gamma)
scalar ll_g = -246.7197
```

A programação utilizada no STATA para ajustar o modelo de regressão Weibull, armazenar o valor do logaritmo da verossimilhança, calcular a estatística do teste e seu p-valor é:

```
streg inter sexo, dist(weib) time
scalar ll_w = -246.74214
scalar trv_w = 2*(ll_g-ll_w)
scalar p_w = 1-chi2(1,trv_w)
```

A programação utilizada no STATA para ajustar o modelo de regressão log-normal, armazenar o valor do logaritmo da verossimilhança, calcular a estatística do teste e seu p-valor é:

```
streg inter sexo, dist(lnormal)
scalar ll_ln = -250.39912
scalar trv_ln= 2*(ll_g-ll_ln)
scalar p_ln = 1-chi2(1,trv_ln)
```

A programação utilizada para exibir o p-valores dos dois modelos de regressão paramétricos acima calculados no STATA é:

```
escalar list p_w p_ln
```

onde a resposta fornecida pelo *software* é:

```
       p_w =  .83222484
      p_ln =  .00667338
```

Há evidências estatísticas que o modelo de regressão paramétrico log-normal não é um modelo adequado aos dados ($p - valor = 0{,}0067$), e não há evidências estatísticas que o modelo de regressão paramétrico Weibull não seja adequado aos dados ($p - valor = 0{,}8322$).

Note que, neste caso, não é necessário utilizar os critérios AIC e BIC, uma vez que, como somente dois modelos foram ajustados e o teste dos modelos encaixados indicou que um modelo não é adequado, há apenas um modelo de regressão paramétrico adequado aos dados. Ou seja, o modelo de regressão paramétrico que pode ser utilizado é o modelo Weibull.

5.7.3. Adequação do Modelo Ajustado

Com o modelo de regressão paramétrico definido, é preciso testar a adequação do modelo ajustado. Para tanto, serão analisados alguns dos resíduos descritos na Seção 5.5..

O resíduo de Cox-Snell é utilizado para testar a qualidade geral do ajuste. Como havia sido discutido na Seção 5.5.1., quando estes resíduos são bem descritos por uma distribuição de probabilidades exponencial, o modelo está bem ajustado.

A programação utilizada para calcular os resíduos de Cox-Snell no STATA é:

```
streg inter sexo, dist(weib) time
predict res_cs, csnell
```

O comando `predict`, juntamente com o comando `csnell` depois da vírgula, é utilizado para fazer a predição dos resíduos de Cox-Snell, que são armazenados na variável `res_cs`.

Para analisar estes resíduos é necessário ajustar um modelo de sobrevivência para eles, uma vez que se compara a função de sobrevivência estimada pelo modelo exponencial com a função de sobrevivência estimada por Kaplan-Meier. Para isso, os resíduos devem ser declarados como os tempos de falha do modelo de análise de sobrevivência, ou seja:

```
stset res_cs
```

A programação utilizada no STATA para ajustar o modelo de regressão exponencial, estimar sua função de sobrevivência e fazer o gráfico da função de

sobrevivência estimada por Kaplan-Meier com a função de sobrevivência estimada pelo modelo paramétrico exponencial é:

```
streg, dist(exp)
predict sobr_e, surv
label variable sobr_e "Exponencial"
sts graph, addplot(line sobr_e _t, sort) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) xlabels(0 0.5 1 1.5, format(%9.1f)) title(Análise do
Resíduo Cox-Snell) xtitle(Resíduos Cox-Snell)
```

A Figura 36 apresenta a curva de sobrevivência dos resíduos Cox-Snell estimada por Kaplan-Meier e curva de sobrevivência dos resíduos Cox-Snell estimada pelo modelo paramétrico exponencial. Este gráfico mostra que os resíduos Cox-Snell são bem ajustados com uma distribuição densidade de probabilidades exponencial, ou seja, o modelo de regressão paramétrico Weibull ajusta-se bem aos dados.

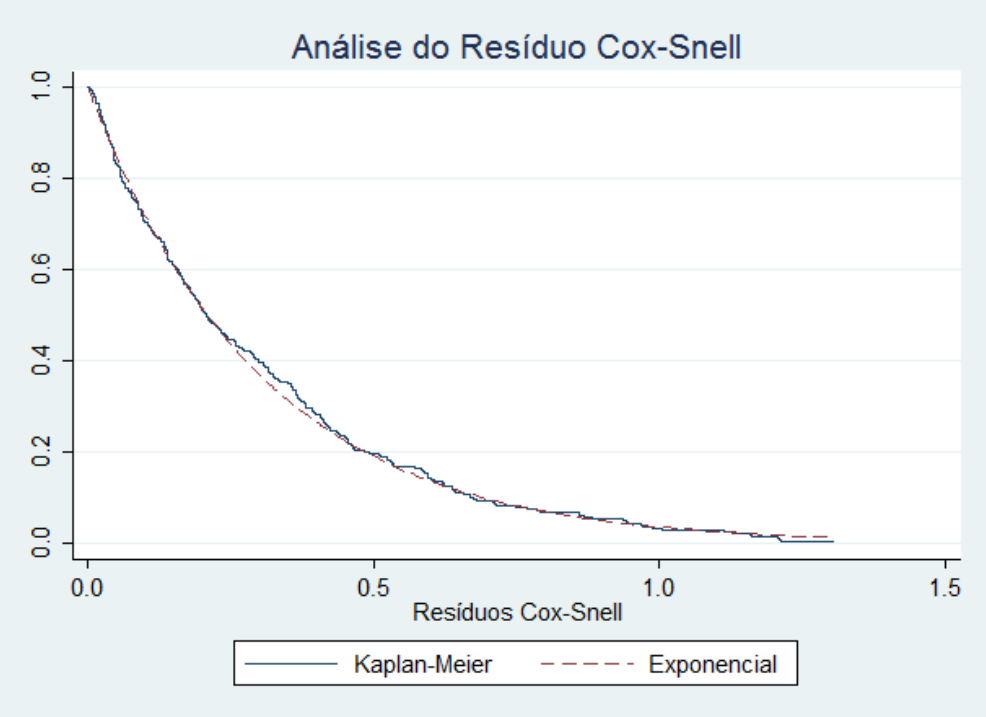

Figura 36: Curvas de sobrevivência dos resíduos Cox-Snell estimadas por Kaplan-Meier e pelo modelo exponencial.

A programação utilizada no STATA para estimar a função de sobrevivência por Kaplan-Meier e fazer o gráfico da função de sobrevivência estimada por Kaplan-Meier *versus* a função de sobrevivência estimada pelo modelo paramétrico exponencial é:

```
sts generate km2 = s
scatter km2 sobr_e, msize(small) xlabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ylabels(0 0.2 0.4 0.6 0.8 1, format(%9.1f))
xtitle(Kaplan-Meier dos Resíduos) ytitle(Exponencial dos Resíduos)
```

A Figura 37 apresenta a curva de sobrevivência dos resíduos Cox-Snell estimada por Kaplan-Meier *versus* a curva de sobrevivência dos resíduos Cox-Snell estimada pelo modelo paramétrico exponencial. Este gráfico gerou uma reta praticamente perfeita, indicando que os resíduos Cox-Snell são bem ajustados com uma distribuição densidade de probabilidades exponencial e, consequentemente, que o modelo de regressão paramétrico Weibull ajusta-se bem aos dados.

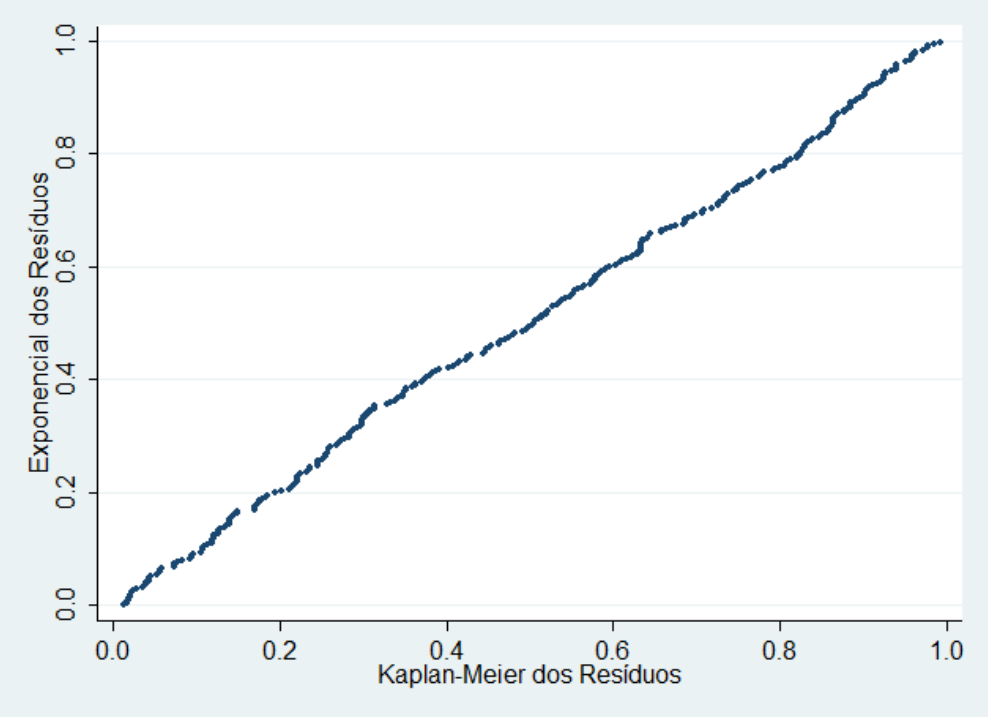

Figura 37: Curva de sobrevivência dos resíduos Cox-Snell estimada por Kaplan-Meier *versus* curva de sobrevivência dos resíduos Cox-Snell estimadas pelo modelo exponencial.

5.7.4. Interpretação e Extrapolação

O modelo de regressão paramétrico Weibull, segundo a abordagem de modelos de tempo de vida acelerado, é o modelo adequado para explicar o tempo até a morte ocasionada por todas as causas, dos pacientes do ambulatório de insuficiência cardíaca do Hospital de Clínicas de Porto Alegre, ajustado pelo sexo do paciente e pelo fato do

paciente ter sido ou não internado pelo menos uma vez. A programação utilizada no STATA para ajustar este modelo de regressão paramétrico Weibull é:

```
streg inter sexo, dist(weib) time
```

As funções de sobrevivência estimadas pelo modelo de regressão paramétrico Weibull para os níveis de cada covariável podem ser desenhadas. A programação utilizada no STATA para fazer este gráfico é:

```
stcurve, surv at1(inter=0 sexo=0) at2(inter=0 sexo=1) at3(inter=1
sexo=0) at4(inter=1 sexo=1) ylabels(0 0.2 0.4 0.6 0.8 1,
format(%9.1f)) ytitle(Sobrevivência) title(Curvas de Sobrevivência
estimadas) subtitle(Modelo de Regressão Weibull) xtitle(Tempo)
legend(label(1 Homens que nunca internaram) label(2 Mulheres que nunca
internaram) label(3 Homens que já internaram) label(4 Mulheres que já
internaram))
```

O comando `at` é o responsável por estimar as curvas para cada covariável, ou seja, utilizando `at1(inter=0 sexo=0)` desenha-se a curva de sobrevivência estimada pelo modelo de regressão paramétrico Weibull para homens (`sexo=0`) que nunca haviam sido internados (`inter=0`), utilizando `at2(inter=0 sexo=1)` desenha-se a curva de sobrevivência estimada pelo modelo de regressão paramétrico Weibull para mulheres (`sexo=1`) que nunca haviam sido internadas (`inter=0`), utilizando `at3(inter=1 sexo=0)` desenha-se a curva de sobrevivência estimada pelo modelo de regressão paramétrico Weibull para homens (`sexo=0`) que haviam sido internados pelo menos uma vez (`inter=1`) e utilizando `at4(inter=1 sexo=1)` desenha-se a curva de sobrevivência estimada pelo modelo de regressão paramétrico Weibull para mulheres (`sexo=1`) que haviam sido internadas pelo menos uma vez (`inter=1`). O estilo das linhas e o tamanho da fonte da legenda foram alterados via Menu, conforme passos já explicados anteriormente.

A Figura 38 apresenta as curvas de sobrevivência estimadas pelo modelo de regressão paramétrico Weibull para os níveis das covariáveis `sexo` e `inter`. Note que o grupo de pacientes do sexo masculino que já haviam sido internados pelo menos uma vez foi o grupo que apresentou menor probabilidades de sobrevivência, enquanto que o grupo de pacientes do sexo feminino que nunca haviam sido internadas foi o grupo que apresentou probabilidades de sobrevivência mais elevadas.

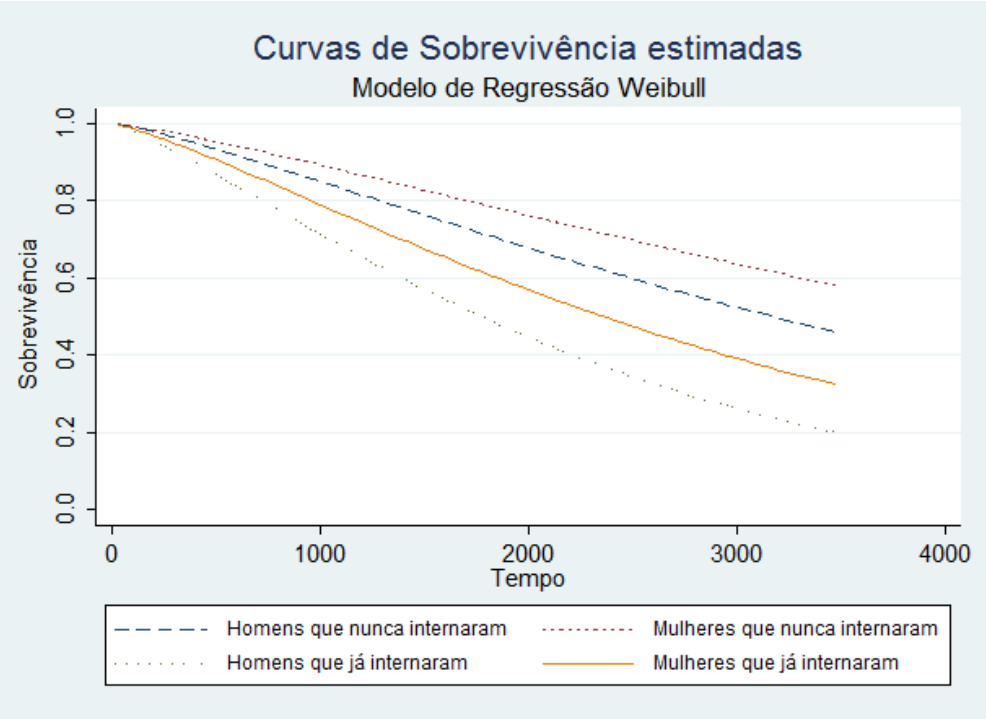

Figura 38: Funções de sobrevivência estimadas pelo modelo de regressão paramétrico Weibull para os níveis das covariáveis.

Ainda, quando se ajusta o modelo de regressão Weibull sob a abordagem de modelos de tempo de vida acelerado, no STATA, a tabela fornecida como resposta pelo *software* é:

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       inter |  -.5807721   .1801663    -3.22   0.001    -.9338916   -.2276526
        sexo |   .2838407   .1834787     1.55   0.122     -.075771    .6434523
       _cons |    8.35216   .1720159    48.55   0.000     8.015014    8.689305
-------------+----------------------------------------------------------------
        /ln_p |   .2288404   .0845645     2.71   0.007      .063097    .3945837
-------------+----------------------------------------------------------------
           p |   1.257141   .1063095                      1.06513    1.483766
         1/p |   .7954555   .0672673                     .6739605    .9388524
------------------------------------------------------------------------------
```

Os resultados acima são importantes porque os coeficientes das covariáveis são utilizados para calcular a razão dos tempos de falha medianos. Este cálculo é feito aplicando a função matemática exponencial no coeficiente. Ainda, aplicando a função matemática exponencial também nos valores estimados para os intervalos de confiança, tem-se a estimativa para os intervalos de confiança da razão dos tempos de falha medianos.

A programação utilizada no STATA para calcular estas estimativas para a covariável `inter` é:

```
scalar rtm_inter = exp(_b[inter])
scalar rtm_li_inter = exp(-0.9338916)
scalar rtm_ls_inter = exp(-0.2276526)
```

Note que para a estimativa pontual não é necessário digitar o valor do coeficiente porque `_b[inter]` contêm este valor. Para o intervalo de confiança é necessário digitar os valores que se encontram na tabela exibida anteriormente.

A programação utilizada no STATA para exibir estas estimativas é:

```
scalar list rtm_li_inter rtm_inter rtm_ls_inter
```

onde a resposta fornecida pelo *software* é:

```
rtm_li_inter =  .39302125
  rtm_inter =  .55946625
rtm_ls_inter =  .79640088
```

A programação utilizada no STATA para calcular estas estimativas para a covariável `sexo` é:

```
scalar rtm_sexo = exp(_b[sexo])
scalar rtm_li_sexo = exp(-0.075771)
scalar rtm_ls_sexo = exp(0.6434523)
```

A programação utilizada no STATA para exibir estas estimativas é:

```
scalar list rtm_li_sexo rtm_sexo rtm_ls_sexo
```

onde a resposta fornecida pelo *software* é:

```
rtm_li_sexo =  .92702847
  rtm_sexo =  1.3282213
rtm_ls_sexo =  1.9030394
```

Isto significa que o tempo mediano até a morte ocasionada por qualquer tipo de causa, para pacientes que já haviam sido internados pelo menos uma vez, é aproximadamente a metade do tempo mediano até a morte ocasionada por qualquer tipo de causa, para pacientes que nunca haviam sido internados. A diferença entre a razão dos tempos medianos de morte destes dois grupos de pacientes é significativa, uma vez que $\text{IC95\%} = (0{,}393\ ,0{,}796)$ não contêm o valor 1.

É interessante destacar que se pode trocar o sentido da interpretação, ou seja, é possível interpretar a razão de tempos medianos dos pacientes que nunca haviam sido internados em relação aos pacientes que já haviam sido internados pelo menos uma vez através do cálculo $1/\alpha_1$, ou seja, $1/0{,}55946625 = 1{,}787418$.

Portanto, pode-se concluir, de maneira equivalente, que o tempo mediano até a morte ocasionada por qualquer tipo de causa, para pacientes que nunca haviam sido internados, é 1,8 vezes o tempo mediano até a morte ocasionada por qualquer tipo de causa, para pacientes que já haviam sido internados pelo menos uma vez.

A estimativa pontual para a razão de tempos medianos da covariável sexo é 1,3. Contudo, esta razão não é significativa, uma vez que IC95% $= (0{,}927\ ,\ 1{,}903)$ contêm o valor 1.

Como utilizou-se um modelo de regressão paramétrico, pode-se também fazer extrapolação, ou seja, estimar a probabilidade de sobrevivência para tempos de sobrevivência superiores ao do estudo. Geralmente, este é um o grande objetivo quando se ajustam modelos paramétricos.

Para estimar a probabilidade de sobreviver a t dias, dos pacientes do ambulatório de insuficiência cardíaca do HCPA, para mulheres que haviam sido internadas pelo menos uma vez, considerando que o modelo de regressão paramétrico Weibull é adequado aos dados, tem-se a expressão:

$$\hat{S}(t) = \exp\left\{-\left[\exp\left(-8.35216 - (-0.5807721 * 1 + 0.2838407 * 1)\right) t\right]^{1.257141}\right\},$$

onde $\alpha_0 = 8.35216$, $\alpha_1 = -0.5807721$, $\alpha_2 = 0.2838407$ e $p = 1.257141$ são retirados da tabela gerada quando se ajusta o modelo de regressão paramétrico Weibull. Note que os coeficientes das covariáveis estão multiplicados por 1, ou seja, $-0.5807721 * 1$ e $0.2838407 * 1$. Isto é feito porque deseja-se estimar a probabilidade de morte para mulheres (`sexo=1`) que haviam sido internadas pelo menos uma vez (`inter=1`).

Para obter a estimativa da probabilidade de sobreviver a 4000 dias, dos pacientes do ambulatório de insuficiência cardíaca do HCPA, para mulheres que já haviam sido internadas pelo menos uma vez, substitui-se, na expressão acima, $t = 4000$. Ou seja,

$$\hat{S}(4000) = \exp\left\{-\left[\exp\left(-8.35216 - (-0.5807721 * 1 + 0.2838407 * 1)\right) * 4000\right]^{1.257141}\right\}$$

A programação utilizada para fazer este cálculo no STATA é:

```
scalar prev_mr_4000 = exp(-(exp(-8.35216 -(-0.5807721*1)-
(0.2838407*1))*4000)^1.257141)
scalar list prev_mr_4000
```

onde a resposta do comando fornecida pelo *software* é:

```
prev_mr_4000 =  .25919577
```

Estima-se que a probabilidade de sobreviver a 4000 dias, daqueles pacientes do ambulatório de insuficiência cardíaca do HCPA que estão vivos até esse dia, que são do sexo feminino e que já haviam sido internados pelo menos uma vez, é 0,259, se o comportamento do tempo de sobrevida desses pacientes continuar sob as mesmas condições.

6. Considerações Finais

O objetivo deste trabalho foi produzir um texto de fácil compreensão e que auxiliasse a reprodução das técnicas de análise de sobrevivência, principalmente as mais utilizadas em estudos de custo-efetividade. O foco principal desta monografia voltou-se aos modelos paramétricos, porque estes modelos permitem a realização de extrapolações.

Os conceitos básicos de tempo de falha e censura foram introduzidos neste trabalho, porque é necessário entender estes conceitos para entender a análise de sobrevivência. Os modelos não-paramétricos de Kaplan-Meier e tábua de vida foram explicados através da resolução de um exemplo, para que a lógica de cada modelo pudesse ser entendida. O modelo de Kaplan-Meier é um importante modelo de análise de sobrevivência, provavelmente um dos mais conhecidos e utilizados. Este modelo é considerado uma análise descritiva dos modelos de análise de sobrevivência, além de servir como ferramenta na escolha do modelo paramétrico adequado aos dados e, por isso, é muito importante.

Os modelos paramétricos e os modelos de regressão paramétricos foram apresentados neste trabalho em um capítulo para cada tópico. Os capítulos iniciaram com a teoria referente a estes modelos e abordaram técnicas utilizadas para selecionar a distribuição densidade de probabilidades adequada para descrever os dados de tempo de sobrevivência. Quando foram abordados os modelos de regressão paramétricos, modelos paramétricos que permitem inclusão de covariáveis na análise, também foram discutidos métodos de seleção de covariáveis e medidas de verificação da qualidade do ajuste. Ainda, foram abordadas as duas maneiras de fazer análise de sobrevivência com modelos de regressão paramétrica: os modelos de taxas de falhas proporcionais e os modelos de tempo de vida acelerada.

Considera-se que a grande contribuição deste trabalho é a exemplificação detalhada da modelagem para os modelos paramétricos e os modelos de regressão paramétricos, tanto na abordagem de modelos de taxas de falhas proporcionais quanto na de modelos de tempo de vida acelerada. Ainda, as programações utilizadas para desenvolver o exemplo, que foi resolvido no *software* STATA, foram fornecidas e houve a preocupação de explicar alguns comandos destas programações visando facilitar a execução da análise.

Havia-se imaginado fazer estas análises também no *software* R, que é um *software* livre e todas as pessoas podem ter acesso, já que o *software* STATA é um *software* pago. Contudo, não foi possível realizar algumas análises no R porque, por exemplo, ele não ajusta o modelo gama generalizado ou, ao calcular a função de sobrevivência para o modelo Gompertz ele não considera a censura. Estas limitações inviabilizaram o uso desse *software*.

Uma sugestão para trabalhos futuros é comparar quando é vantajoso utilizar os modelos paramétricos ou os modelos de regressão paramétricos. Por exemplo, quando utiliza-se os modelos paramétricos estratificando por alguma covariável tem-se a vantagem de poder supor distribuições de probabilidade diferentes para os diferentes grupos, respeitando as especificidades de cada grupo. Por outro lado, pode-se ter estratos de tamanhos pequenos, o que pode comprometer a qualidade da análise. Pode-se, portanto, estudar quais situações tornam adequado o uso do modelo paramétrico ou do modelo de regressão paramétrico.

Referências Bibliográficas

BANDEIRA, M.D. Estatística Demográfica, Porto Alegre: UFRGS, 2009.

BASTOS, J.; ROCHA, C. Análise de Sobrevivência: Conceitos Básicos. *Arq Med*, 2006, vol.20, nº.5/6, p.185-187.

CLEVES, M. A.; GOULD, W.; GUTIERREZ, R.; MARCHENKO, Y. An introduction to survival analysis using Stata. 2ª edição. Editora STATA Press, 2008.

COLLETT, D. Modelling Survival Data in Medical Research. 2ª edição. Editora Chapman & Hall, 2003.

COLOSIMO, E. A.; GIOLO, S. R. Análise de Sobrevivência Aplicada. 1ª edição. São Paulo: Editora Edgard Blücher, 2006.

GRAY, A. M.; CLARKE, P. M.; WOLSTENHOLME, J. L.; WORDSWORTH, S. Applied Methods of Cost-effectiveness Analysis in Health Care. Editora Oxford, 2011.

HOSMER, D. W. JR.; LEMESHOW, S. Applied Survival Analysis: regression modeling of time to event data. Editora John Wiley & Sons, 1999.

LATIMER, N., Survival Analysis for Economic Evaluations Alongside Clinical Trials – Extrapolation with Patient-Level Data, Relatório Técnico do NICE (disponível online, acessado em 20 de outubro de 2011, http://www.nicedsu.org.uk/NICE DSU TSD Survival analysis_finalv2.pdf).

LEE, E. T.; WANG, J. W. Statistical Methods for Survival Data Analysis. 3ª edição. New Jersey: Editora John Wiley & Sons, 2003.

TURNBULL, B. W. Nonparametric Estimation of a Survivorship Function whit doubly Censored Data. J. R. Statist. Soc. B, 38, 290-295, 1976.

WIENKE, A. Frailty Models in Survival Analysis. Editora Chapman & Hall, 2009.

Apêndice 1: Programação para Tábua de Vida Clínica no STATA

A programação utilizada para estimar a função de sobrevivência para pacientes que já foram internados pelo menos um vez pelo método da tábua de vida com o tempo em intervalos de 100 dias é:

```
ltable follow obitot if inter==1, intervals(100)
```

A programação utilizada para fazer o gráfico da função de sobrevivência estimada para os pacientes que já foram internados pelo menos um vez pelo método da tábua de vida com o tempo em intervalos de 100 dias é:

```
ltable follow obitot if inter==1, intervals(100) graph
```

A programação utilizada para estimar a função taxa de falha para pacientes que já foram internados pelo menos um vez pelo método da tábua de vida com o tempo em intervalos de 100 dias é:

```
ltable follow obitot if inter==1, hazard intervals(100)
```