SELEÇÃO DE MODELOS DE TEMPOS COM LONGA-DURAÇÃO PARA DADOS DE FINANÇAS

DANIELE CRISTINA TITA GRANZOTTO

Orientador: Prof. Dr. **FRANCISCO LOUZADA NETO**
Coorientadora: Profa. Dra. **GLEICI DA S. C. PERDONÁ**

São Carlos-SP
Julho - 2008

SELEÇÃO DE MODELOS DE TEMPOS COM LONGA-DURAÇÃO PARA DADOS DE FINANÇAS

DANIELE CRISTINA TITA GRANZOTTO

Orientador: Prof. Dr. **FRANCISCO LOUZADA NETO**
Coorientadora: Profa. Dra. **GLEICI DA S. C. PERDONÁ**

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos-SP
Julho - 2008

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

G765sm

Granzotto, Daniele Cristina Tita.
Seleção de modelos de tempos com longa-duração para dados de finanças / Daniele Cristina Tita Granzotto. -- São Carlos : UFSCar, 2008.
88 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Análise de sobrevivência. 2. Finanças. 3. Seleção de modelos. 4. Simulação. I. Título.

CDD: 519.5 (20ª)

RESUMO

Os modelos de análise de sobrevivência com fração de cura incorporam a heterogeneidade de duas populações (susceptíveis e imunes ao evento de interesse) e são conhecidos na literatura como modelos de longa-duração. Com o objetivo de exemplificar a aplicabilidade dos modelos de longa-duração em dados da área de finanças, trabalhou-se com o modelo proposto por Berckson e Gage usando-se para isto os modelos Weibull e log-logístico. Estudou-se a adequabilidade dos modelos e métodos para seleção e verificação de ajuste. Um estudo de simulação foi realizado com o propósito de testar a medida de distância entre curvas como alternativas às métricas usuais e também verificar o comportamento destas métricas em diferentes situações de percentuais de censura e tamanhos de amostras. Neste estudo verificou-se que, uma métrica simples como a medida de distância entre curvas, é capaz de selecionar o modelo mais apropriado aos dados na presença de longa-duração e grandes carteiras de clientes.

Palavras-chave: Sobrevivência, modelos de finanças, longa-duração, critérios de seleção de modelo, simulação.

AGRADECIMENTOS

Para chegar ao programa de mestrado foi necessário trilhar alguns caminhos. Desta forma, em primeiro lugar agradeço ao meu pai por ter me apoiado, acreditado e tornado possível os meus estudos.

Meus agradecimentos ao meu esposo que esteve ao meu lado durante parte da minha graduação e todo o mestrado, sempre me apoiando e ajudando em tudo o que foi possível. Agradeço-o por ter dividido comigo seus conhecimentos e também seu carinho e atenção.

Sempre terei muito a agradecer aos professores e amigos da UEM pois, foi junto com eles, que consegui concluir a primeira fase, a graduação.

Tenho muito a agradecer também, as pessoas que integram o Departamento de Estatística da UFSCar por me acolherem e tornarem possível a conclusão de mais uma etapa. Sempre me recordarei com carinho destas pessoas.

Em especial, tenho meus sinceros agradecimentos aos professores Neto e Gleici que me orientaram neste trabalho. Agradeço-os pela paciência, compreensão, atenção disposta e por sua dedicação, reconhecendo sempre que este trabalho não seria possível sem seus auxílios.

Além dos orientadores, agradeço aos professores que lecionaram as disciplinas durante o mestrado. Agradeço-os por terem compartilhado comigo seus conhecimentos e pela disponibilidade de me ajudar e responder aos meus questionamentos.

Tenho que agradecer também aos amigos que conheci e com eles compartilhei momentos inesquecíveis. Foram eles que, em muitos momentos de dificuldades estiveram ao meu lado.

Finalmente, agradeço a todas as outras pessoas não citadas, pessoas estas, não menos importantes e também agradeço aos softwares SAS e R que tornaram todas as análises possíveis.

Sumário

Capítulo 1

Introdução

A fidelização e retenção de clientes têm papel fundamental nas empresas que hoje atuam em mercados altamente competitivos, principalmente naquelas ligadas à área de finanças: bancos, financiadoras, seguradoras etc.

Para a manutenção ou acréscimo da lucratividade é primordial que estas instituições identifiquem antecipadamente clientes com alto potencial de ruptura de relacionamento, possibilitando ações preventivas que evitem a perda desses clientes.

Neste contexto, técnicas estatísticas usadas em análise de sobrevivência e confiabilidade têm sido bastante aplicadas e mesmo desenvolvidas. Uma característica importante presente em carteiras de clientes é a fração de clientes fidelizados, essa característica exige, além das técnicas usuais, a utilização de modelos com longa-duração.

Outra característica a ser ressaltada, é a presença de grandes bancos de dados na área de finanças, ao contrário do que geralmente ocorre nas aplicações em análise de sobrevivência. Por esta razão, além de desenvolver modelos com fração de fidelizados, as técnicas estatísticas para modelagem devem ser aprimoradas e adequadas para estas carteiras de clientes.

Portanto o objetivo deste trabalho é investigar a utilização de modelos de longa-duraçãoem grandes bases de dados. Para isto, consideraremos uma revisão dos conceitos básicos de análise de sobrevivência, Capítulo 2. Nesse capítulo definimos os conceitos de evento de interesse e tempo de falha. Outro conceito que foi definido, talvez o mais importante é a censura. Entende-se por censura a observação parcial da resposta e, isto pode ser um complicador no momento de analisar os dados, uma vez que, a variável resposta, tempo, não é medida instantaneamente. Basicamente, descrevemos os quatro tipos mais comuns: censura aleatória, censura do tipo I ou à direita, censura do tipo II e censura intervalar.

Um enfoque especial é dado a censura do tipo I ou à direita, uma vez que, este tipo

de censura foi a observada nos dados que foram considerados (aplicação apresentada no capítulo posterior) e também, durante a apresentação dos modelos de longa-duração.

Ainda no segundo capítulo, descrevemos o tempo (variável aleatória positiva) e as funções que podem ser consideradas para descrever esta variável. Em especial, descrevemos a função densidade de probabilidade, a qual tem como resposta, por exemplo, a chance de um cliente vir a abandonar a carteira, a função de sobrevivência que apresenta o tempo até a ocorrência do evento de interesse, neste caso abandono da carteira, e, a função de risco, a qual mede a taxa de falha instantânea do indivíduo abandonar a carteira.

Técnicas não-paramétricas para a obtenção das funções densidade de probabilidade, sobrevivência e risco são apresentadas também no Capítulo 2. Para a obtenção das curvas de sobrevivência observadas, duas técnicas serão apresentadas: o estimador de Kaplan-Meier e o estimador atuarial.

Assim, durante a aplicação e simulação, utilizamos destas técnicas não-paramétricas para visualizar a distribuição dos dados e ajustá-los via modelos paramétricos, também apresentados no Capítulo 2. Estes modelos paramétricos são descritos aqui em detalhes, sendo apresentado, para todos, suas principais características e as formas das suas funções de sobrevivência e risco.

No Capítulo 3, introduzimos o modelo de longa-duração, em especial, o modelo de sobrevivência com longa-duração proposto por Berckson e Gage em 1952 e suas principais características. Nesse capítulo, introduzimos também os conceitos de estimação via máxima verossimilhança para dados com censura à direita e usando-se destes conceitos, apresentamos as verossimilhanças e log-verossimilhanças dos modelos exponencial, Weibull e log-logístico, modelos estes que serão usados para ajustar os dados da aplicação e durante um estudo de simulação realizado. Métodos para seleção de modelos são apresentados. Os critérios considerados neste capítulo são utilizados para ajuste e seleção dos modelos trabalhados na aplicação e simulação. Em especial, utilizamos os critérios AIC e BIC e, alternativamente, propomos uma métrica bastante conhecida, a distância Euclidiana. Apesar de conhecida, a norma Euclidiana foi usada de forma diferenciada. Esta métrica foi empregada não só para medir distâncias mas para selecionar o modelo mais apropriado aos dados, usando para isto, as distâncias das curvas estimadas com relação as curvas observadas.

No Capítulo 4 consideramos um estudo de simulação, baseado no método Bootstrap, onde as diferentes métricas consideradas foram avaliadas em termos de diferentes modelos e diferentes valores de censura.

Finalmente, no Capítulo 5, apresentamos uma conclusão geral do estudo e assim, enfatizamos os principais resultados obtidos.

Capítulo 2

Teoria Introdutória em Análise de Sobrevivência

Nas últimas décadas, as técnicas de análise para dados de sobrevivência e confiabilidade tem sido bastante desenvolvidas e utilizadas. A razão deste crescimento é explicada pelo desenvolvimento e aprimoramento de técnicas estatísticas combinada com computadores cada vez mais velozes e eficazes. Este crescimento pode ser quantificado pelo número crescente de aplicações na área médica, que tem por objetivo estudar a função de risco/sobrevivência (que serão definidas posteriormente), de pacientes submetidos a determinados tratamentos, e pela imensa procura destas técnicas por financiadoras e seguradoras, as quais buscam estudar e prever possíveis comportamentos de seus clientes. Em geral buscam fidelizar os clientes já existentes em suas carteiras.

A metodologia de análise de sobrevivência consiste, em sua essência, numa coleção de procedimentos estatísticos utilizados para analisar dados relacionados ao tempo até a ocorrência de um determinado evento de interesse (morte, cura, contração ou recidiva de uma doença etc), a partir de um tempo inicial pré-estabelecido. Na área financeira, o evento de interesse pode ser o abandono de um cliente, o não pagamento de empréstimos, a ocorrência de um sinistro etc. Neste contexto, a análise de sobrevivência, permite determinar quais variáveis afetam o risco de ocorrência de determinado fenômeno.

A principal característica relacionada a estes dados diz respeito a presença de observações censuradas, que consiste na observação parcial da resposta e se dá, geralmente pelo fato de um cliente abandonar a carteira ou, estes clientes não experimentaram o evento de interesse em estudo.

Neste capítulo, pretende-se introduzir alguns conceitos básicos em análise de sobrevivência, descrevendo algumas particularidades, tais como: evento de interesse, tempo de falha, censura, variáveis explicativas e funções de interesse neste contexto. Também serão

apresentados, neste capítulo, métodos não-paramétricos para estimação das funções de interesse (função densidade de probabilidade, função de risco e de sobrevivência), bem como métodos para análise gráfica do comportamento dos dados (análise descritiva), tais como: estimador Kaplan-Meier, atuarial e TTT-plot (Lee & Wang, 2003).

2.1 Evento de Interesse

O evento de interesse, como o próprio nome já diz, é aquele que se tem interesse em observar no experimento. Na área financeira, o evento de interesse diz respeito ao cliente vir a sinistrar, estudando assim tempo até a ocorrência do sinistro; ou em outros casos, o tempo até o pagamanto de determinado produto ou mesmo o tempo até abandono da carteira etc.

Em algumas situações, a definição de falha é clara, mas em outras pode assumir termos ambíguos. Por exemplo, fabricantes de produtos alimentícios desejam saber informações sobre o tempo de validade (tempo de falha) de seus produtos expostos em balcões frigoríficos de supermercados. O tempo de falha vai do tempo inicial de exposição (chegada ao supermercado) até o produto ficar "inapropriado ao consumo". Este evento deve ser claramente definido antes de se iniciar o estudo. Por exemplo, o produto fica inapropriado para o consumo quando algumas características específicas do produto forem alteradas (Prentice *et al.*, 1978).

2.2 Tempo de Falha

Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e, muito frequentemente, pelas censuras, os quais constituirão a resposta. No caso de estudos financeiros, o tempo de falha pode ser o tempo decorrido até o abandono dos clientes em carteiras de seguradoras ou bancos ou ainda, medir o tempo até a adesão ao programa.

O tempo de falha é o período decorrido até a ocorrência do evento de interesse. Assim sendo, três elementos compõem o tempo de falha:

- **Início do Experimento:** É imprescindível que o tempo de início do estudo seja precisamente definido. Os indivíduos devem ser comparáveis na origem do estudo, com exceção de diferenças medidas pelas covariáveis. Em estudos aleatorizados, a data da aleatorização é a escolha natural para a origem do estudo. A data em que o cliente aderiu ao programa ou, a data do contrato são outras escolhas possíveis.

- **Escala de Medida:** A escala de medida é quase sempre o tempo real (ou do "relógio"), apesar de existirem outras alternativas. Em carteiras de clientes podem surgir outras escalas de medida, como o número de vezes que o cliente inicia um relacionamento com a empresa ou o número de produtos que este cliente vem a adquirir.

- **Evento de Interesse:** como descrito na seção (2.1), na maioria dos casos consiste no abandono por parte do cliente da carteira da instituição.

2.3 Censura

Quando se trata de dados de sobrevivência (que agrupam tanto os tempos de sobrevivência como um conjunto de variáveis observáveis que podem estar relacionadas com estes tempos), um complicador presente está relacionado com o fato da variável de interesse, ou evento de interesse não ser medido instantaneamente e independentemente do tamanho da resposta. Valores grandes da variável tempo necessitam de mais tempo e persistência para serem observados. Em situações extremas, este fato pode comprometer a observação do valor da variável para alguns clientes, uma vez que o evento de interesse pode não ocorrer até o final do estudo.

Assim sendo, a censura nada mais é que a observação parcial da resposta e se dá geralmente pelo abandono da carteira por parte do cliente, antes que este experimente o evento de interesse.

Desta forma, existe a necessidade da introdução de uma variável extra na análise, indicando se o cliente teve seu tempo até a ocorrência do evento de interesse exatamente observado ou não. Esta variável é conhecida na literatura como variável indicadora de censura e assume os valores

$$\delta_i = \begin{cases} 0 \text{ se } t_i \text{ corresponde a um tempo de falha} \\ 1 \text{ se } t_i \text{ corresponde a um tempo censurado,} \end{cases} \tag{2.1}$$

para cada cliente i, $i = 1, \ldots, n$, sendo t_i os tempos observados (censura ou não).

Os tempos, mesmo censurados devem ser usados na análise estatística, pois, mesmo incompletas, as observações censuradas nos fornecem informações sobre o tempo falha de clientes. Assim sendo, a omissão das censuras nos cálculos estatísticos certamente acarretarão em conclusões viciadas (Louzada, Mazucheli & Achcar, 2001).

Existem basicamente quatro tipos de censuras:

- **Censura do tipo I ou à direita:** este tipo de censura ocorre quando, geralmente, o tempo para o fim do estudo é pré-estabelecido, assim, alguns clientes deixam de experimentar o evento de interesse ao fim deste estudo, tendo os seus tempos censurados à direita. Por exemplo, um banco deseja verificar o tempo até que os clientes, de determinada carteira, se tornem inadimplentes. Estuda-se portanto, esta carteira durante um tempo pré-determinado pela instituição e, ao fim, alguns desses deixaram de experimentar o evento de interesse (portanto são não inadimplentes), observando assim, a censura do tipo I.

- **Censura do tipo II:** ao invés do tempo final do estudo ser pré-estabelecido, o estudo será terminado após um determinado número, n, de indivíduos experimentar o evento de interesse, ou seja, após um número n de ocorrências o experimentador finaliza a pesquisa e os clientes que deixaram de experimentar o evento de interesse terão seus tempos censurados. No caso da instituição financeira que tem interesse em prosseguir o estudo até que 10% de sua carteira de clientes se torne inadimplente. Após o número de ocorrência determinado, os clientes sob risco terão seus tempos censurados à direita.

- **Censura aleatória:** diferentemente das outras censuras, este tipo de censura foge ao controle do experimentador. Geralmente ocorre quando o cliente abandona a carteira de clientes ou, abandona determinado experimento sem ter experimentado o evento de interesse. A censura aleatória é um caso mais geral, tendo como caso particular a censura tipo I ou censura à direita, já apresentada.

- **Censura intervalar:** ocorre quando não se conhece o tempo exato em que ocorreu o evento de interesse, mas sim, que este ocorreu em um intervalo de tempo. Sendo assim, sabe-se que o evento (abandono do cliente ou inadimplência, por exemplo), ocorreu entre um tempo antecessor e um sucessor.

A Figura 2.1-a, ilustra a situação em que os tempos de falha são exatamente observado. Na Figura 2.1-b tem-se a presença de observações censuradas à direita (censura do tipo II), Louzada, Mazucheli & Achcar (2001).

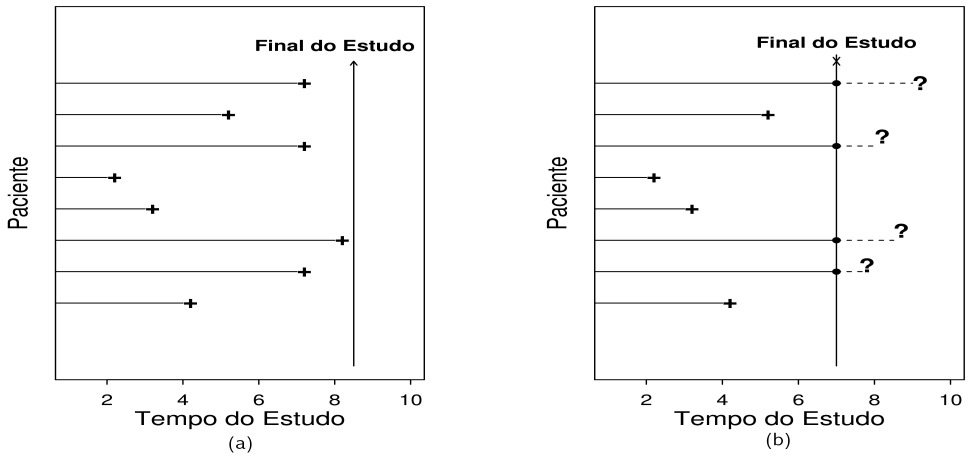


Figura 2.1: (a) Todos os clientes experimentam o evento de interesse antes do final do estudo. (b) Ao final do estudo alguns clientes ainda não haviam experimentado o evento de interesse.

2.4 Presença de Variáveis Explicativas

Além do tempo de sobrevivência e da variável indicadora de censura, também se pode observar nos dados, variáveis que representam tanto a heterogeneidade existentes na população, tais como idade, sexo, dentre outras. Estas variáveis são conhecidas como variáveis explicativas ou covariáveis.

Muitas vezes, o objetivo da análise de sobrevivência está centrado na relação entre o tempo de sobrevivência e algumas variáveis explicativas. A questão é saber se existe efeito destas covariáveis no tempo de sobrevivência bem como, se as interações entre as variáveis explicativas são importantes.

Desta maneira, do ponto de vista estatístico, temos as variáveis tempo de sobrevivência, variável indicadora de censura, e um vetor de variáveis explicativas disponíveis para a análise, Louzada, Mazucheli & Achcar (2001).

Uma característica adicional que também pode ocorrer na análise de sobrevivência é encontrarmos variáveis explicativas que dependem do tempo, ou seja, os valores da

covariável no final do experimento podem não ser os mesmos que no seu início (Colosimo & Giolo, 2006). Por exemplo, pode-se ter um experimento em que o valor do crédito concedido por bancos, financiadoras ou empresas de cartão de crédito, é modificado ao longo do experimento devido a algumas políticas pré-estabelecidas pelo banco.

2.5 Funções de Interesse: Densidade de Probabilidade, Sobrevivência e Risco

Quando se trata de análise de sobrevivência, a variável aleatória contínua e não-negativa tempo (T) é usualmente especificada pelas funções densidade de probabilidade, sobrevivência e risco.

A função densidade de probabilidade é expressa como o limite da probabilidade de um indivíduo vir a experimentar o evento de interesse no intervalo de tempo $[t, t + \Delta t)$ por unidade de tempo e é expressa como sendo

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}, \tag{2.2}$$

onde $f(t) \geq 0$ para todo t. Por se tratar de uma função densidade de probabilidade, tem-se a área abaixo da curva igual a 1 (Lee, 1992).

Já, a função de sobrevivência $S(t)$, é uma das principais funções usadas para descrever a variável aleatória "tempo" e é definida como sendo a probabilidade de um indivíduo não falhar (ou de o evento de interesse não ocorrer) até um determinado tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, denota-se como

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t) \\ &= 1 - \int_0^t f(u)\, du, \end{aligned} \tag{2.3}$$

onde, $f(\cdot)$ é a função densidade de probabilidade (fdp), dada por Lawless (1982).

Alternativamente, (2.3) pode ser escrita na forma

$$S(t) = 1 - F(t), \tag{2.4}$$

onde $F(t)$ é a probabilidade de um indivíduo experimentar o evento de interesse ao tempo t, assim, à partir de (2.4), tem-se $S(t) + F(t) = 1$.

Pelas propriedades da função de sobrevivência e da função densidade acumulada tem-se que

$$\begin{array}{rcl} \lim\limits_{t\to 0} S(t) & = & 1 \quad \text{e} \quad \lim\limits_{t\to\infty} S(t) = 0, \\ \lim\limits_{t\to 0} F(t) & = & 0 \quad \text{e} \quad \lim\limits_{t\to\infty} F(t) = 1. \end{array}$$

Tendo em vista estas propriedades, tem-se que $F(t)$ é uma função monótona crescente e $S(t)$ uma função monótona decrescente, ou não crescente, Lee & Wang (2003).

A função de sobrevivência pode ser obtida também à partir das relações:

$$f(t) = \frac{d}{dt}F(t) \quad \text{ou} \quad f(t) = -\frac{d}{dt}S(t). \tag{2.5}$$

Como o $100(1-p)\,\%$ percentil da variável aleatória T é definido como o valor de t_p tal que $P(T \leq t_p) = p$, tem-se

$$F(t_p) = p \Longrightarrow F^{-1}[F(t_p)] = F^{-1}(p) \tag{2.6}$$

Portanto, $t_p = F^{-1}(p)$, Lee & Wang (2003) e Hosmer & Lemeshow (1999).

A função de risco fornece a taxa instantânea de falha, isto é, dado que o indivíduo não experimentou o evento de interesse até um determinado tempo t. Esta função fornece a probabilidade do indivíduo experimentar determinado evento de interesse no intervalo de tempo $t + \Delta t$ com $\Delta t \to 0$, ou seja,

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \,|T \geq t)}{\Delta t}. \tag{2.7}$$

Devido a sua interpretação, a função de risco (2.7) tem sido preferida por muitos autores para descrever o comportamento do tempo de sobrevivência e sua importância é descrita por Klein & Moeschberger (1997). A função de risco descreve como a probabilidade instantânea de falha (taxa de falha) se modifica com o passar do tempo. Ela é também conhecida como taxa de falha instantânea, força de falha e taxa de falha condicional (Cox & Oakes, 1984).

Além disso, através da função de risco pode-se caracterizar classes especiais de distribuições de tempo de sobrevivência, de acordo com o seu comportamento como função do tempo. A função de risco pode ser constante, crescente, decrescente ou mesmo não monótona.

A probabilidade de uma falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser

expressa em termos da função de sobrevivência como sendo

$$S(t_1) - S(t_2). \tag{2.8}$$

A taxa de falha, ou risco, no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Algebricamente,

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1) S(t_1)}. \tag{2.9}$$

De forma geral, redefinindo o intervalo como sendo $[t, t + \Delta t)$, tem-se à partir da função (2.9)

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{2.10}$$

Além do interesse em estimar as funções especificadas anteriormente (densidade de probabilidade, sobrevivência e risco) , tem-se o interesse em outras quantidade tais como o tempo médio de vida e a vida média residual.

O tempo médio de vida, como o próprio nome sugere, mede o tempo médio até a ocorreência do evento de interesse para um determinado cliente na carteira e é obtido pela área sob a função de sobrevivência. Já a vida média residual, mede o tempo médio para experimentar o evento de interesse que o cliente tem a partir de um tempo t, ou seja, para um cliente com um certo tempo de permanência t, o tempo médio restante de permanência é $vmr(t)$ (Colosimo & Giolo, 2006).

Para o cálculo do tempo médio e vida média residual tem-se as respectivas expressões dadas respectivamente por (2.11) e (2.12).

$$t_m = \int_0^\infty S(t)dt \tag{2.11}$$

$$vmr(t) = \frac{\int_t^\infty (u - t) f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)} \tag{2.12}$$

As funções de sobrevivência, de densidade e de risco são matematicamente equivalentes. Uma vez definida qualquer uma delas, respeitando-se suas propriedades, tem-se as demais por conseqüência.

À partir da função de risco, dada em (2.7), tem-se que

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right] \frac{1}{S(t)} \\ &= \frac{d}{dt} F(t) \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{2.13}$$

onde $f(t)$ é a função densidade da variável aleatória T.

Usando-se da relação apresentada em (2.5) tem-se que (2.13) pode ainda ser escrita como sendo

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log [S(t)]. \tag{2.14}$$

À partir da integração de (2.14) tem-se que

$$S(t) = \exp\left[-\int_0^t h(u)\, du\right]. \tag{2.15}$$

onde $\int_0^t h(u)\, du$ é a função de risco acumulada e é finita para algum tempo $t > 0$.

Desta forma, tem-se duas expressões alternativas para representar a função de densidade da variável aleatória t. São elas,

$$f(t) = h(t) S(t) \quad \text{ou} \quad f(t) = h(t) \exp\left[-\int_0^t h(u)\, du\right]. \tag{2.16}$$

2.6 Métodos Não-Paramétricos

Técnicas estatísticas não-paramétricas são frequentemente utilizadas, em análise de sobrevivência, com o intuito de descrever e caracterizar a distribuição dos dados.

Inicialmente, ao se trabalhar com dados de sobrevivência e confiabilidade, tem-se o interesse em estimar a função densidade de probabilidade, a função de sobrevivência e a função de risco. Desta forma, os métodos não-paramétricos contribuem muito, uma vez que, estas funções podem ser estimadas diretamente a partir dos dados amostrais.

Utilizando-se das técnicas não paramétricas, a função densidade de probabilidade pode ser estimada a partir dos dados amostrais por meio da seguinte expressão

$$\hat{f}(t) = \frac{\begin{array}{c}\text{Número de clientes que experimentaram o evento de interesse na carteira} \\ \text{no intervalo começando em } t\end{array}}{(\text{Número total de clientes}) \times (\text{Amplitude do intervalo})}, \tag{2.17}$$

onde $\hat{f}$ denota o estimador da função densidade f (Louzada, Mazucheli & Achcar, 2001).

A função de sobrevivência pode ser estimada, a partir dos dados, como a proporção de clientes que não experimentaram o evento de interesse mais que um tempo t, e é dada por

$$\hat{S}(t) = \frac{\text{Número de clientes que estão sob risco no tempo} \geq t}{\text{Número total de clientes}} \tag{2.18}$$

Louzada, Mazucheli & Achcar (2001).

Já a função de risco, estimada a partir dos dados amostrais, é dada por (Klein & Kleinbaum, 2005)

$$\hat{h}(t) = \frac{\begin{array}{c}\text{Número de clientes que experimentaram o evento de interesse na carteira} \\ \text{no intervalo começando em } t\end{array}}{(\text{Número de clientes com tempos } > t) \times (\text{Amplitude do intervalo})}. \tag{2.19}$$

2.6.1 Estimador Atuarial

O estimador atuarial tem uma importância histórica pois foi utilizado em informações provenientes de censos demográficos para, essencialmente, estimar características associadas ao tempo de vida dos seres humanos. Este estimador foi proposto por demógrafos e atuários no século passado e usado basicamente em grandes amostras.

Este estimador é conhecido na literatura como estimador atuarial ou tábua de vida e, é obtido através da divisão do tempo total de estudo em k intervalos, geralmente de tamanhos iguais. Então, o estimador atuarial da função de sobrevivência é dado por

$$\hat{S}_{AT}(t) = \frac{n'_1 - d_1}{n'_1}\frac{n'_2 - d_2}{n'_2}\cdots\frac{n'_k - d_k}{n'_k} \tag{2.20}$$

para $t'_k \leq t \leq t'_{k+1}$, $k = 1, \ldots, m$, com $n'_k = n_k - (c_k/2)$, onde n_k representa o número de clientes que não experimentaram o evento de interesse no início do intervalo k, d_k é o número de clientes que experimentaram este evento em estudo e c_k é o número de tempos censurados no intervalo k (Louzada, Mazucheli & Achcar, 2001).

O estimador atuarial da função de risco, no período de tempo compreendido no intervalo k, é dado por

$$\hat{h}_{AT}(t) = \frac{1}{A_u}\left(1 - \frac{n'_k - d_k}{n'_k}\right) \tag{2.21}$$

onde A_u é a amplitude do intervalo u (Louzada, Mazucheli & Achcar, 2001).

É importante ressaltar que os estimadores atuarias são importantes quando se tem uma grande quantidade de dados, e os valores dos tempos de sobrevivência somente são expressos por meio de intervalos. Porém, o estimador de Kaplan-Meier, que será apresentado posteriormente, e de mais fácil implementação e por isso foi utilizado durante o estudo de simulação e aplicação.

2.6.2 Estimador de Kaplan-Meier

O Kaplan-Meier foi proposto por Kaplan & Meier (1958) e é também chamado de estimador produto-limite, proposto inicialmente por Böhmer (1912) (Louzada, Mazucheli & Achcar, 2001).

Este estimador consiste em um método não-paramétrico para a análise de dados de sobrevivência, uma vez que, a presença de observações censuradas é um problema para as técnicas convencionais de análise descritiva que envolvem a média, o desvio padrão e técnicas gráficas como histograma e box-plot, dentre outros (Collet, 1994).

Desse modo, o principal componente da análise descritiva envolvendo dados de sobrevivência é a função de sobrevivência. O procedimento consiste em encontrar uma estimativa para a função de sobrevivência e, a partir dela, estimar estatísticas de interesse tais como, dentre outras, o tempo médio ou mediano e alguns percentis.

O referido estimador considera, na sua construção um número de intervalos de tempo igual ao número de falhas distintas sendo, os limites dos intervalos, os tempos de falha da amostra.

Suponha que existem n clientes no estudo e $k(\leq n)$ eventos de interesse distintos nos tempos $t_1 < t_2 < \ldots < t_k$. Considerando $S(t)$ como uma função discreta com probabilidade maior que zero somente nos tempos em que se experimentou o evento de interesse t_i, $i = 1, \ldots, k$, tem-se que

$$S(t_i) = (1 - q_1)(1 - q_2) \dots (1 - q_i), \tag{2.22}$$

onde q_i é a probabilidade de um indivíduo experimentar o evento de interesse na carteira no intervalo $[t_{i-1}, t_i)$ sabendo que ele não experimentou até t_{i-1} e considerando $t_0 = 0$. Ou seja, pode-se escrever q_i como

$$q_i = P\left(T \epsilon [t_{i-1}, t_i) | T \geq t_{i-1}\right). \tag{2.23}$$

Desta forma, tem-se a expressão geral de $S(t)$ em termos de probabilidades condicionais. O estimador de Kaplan-Meier reduz-se então, a estimar q_i que é dado por,

$$\hat{q}_i = \frac{\text{número de observações do evento em estudo no intervalo } [t_{i-1}, t_i)}{\text{número de observações sob risco em } t_{i-1}},$$

para $i = 1, \dots, k$.

Assim, para a expressão geral do estimador de Kaplan-Meier, considera-se:

- $t_1 \leq t_2 \leq \dots \leq t_k$, os k tempos distintos em que observou-se o evento de interesse ou não;
- d_j o número de clientes que experimentaram o evento de interesse da carteira em t_j, $j = 1, 2, \dots, k$, e
- n_j o número de indivíduos em risco em t_j, ou seja, os indivíduos que não experimentaram o evento de interesse e não censuraram até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier, é então, dado por:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.24}$$

Assim, $\hat{S}(t)$ é uma função escada com degraus nos tempos em que se observou o evento de interesse de tamanho $1/n$ onde n é o número de amostras. Por ser uma adaptação da função de sobrevivência empírica, na ausência de censura reduz-se a:

$$\hat{S}\left(t\right) = \frac{\text{número de indivíduos que não experimentaram o evento até o tempo } t}{\text{número total de indivíduos no estudo}}. \tag{2.25}$$

Assim, o estimador de Kaplan-Meier, tem como suas principais propriedades, o fato de ser não-viciado, fracamente consistente, possuir distribuição assintótica normal e ser estimador de máxima verossimilhança de $S(t)$.

A consistência e normalidade assintótica foram provadas por Breslow e Crowley (1974) e, no artigo original, Kaplan e Meier (1958) mostram que $\hat{S}(t)$ é estimador de máxima verossimilhança de $S(t)$.

A função de verossimilhança é escrita na forma

$$L\left[S\left(\cdot\right)\right] = \prod_{i=0}^{k} \left\{ \left[S\left(t_i\right) - S\left(t_i + 0\right)\right]^{d_i} \prod_{j=1}^{m_i} S\left(t_{ij} + 0\right) \right\}.$$

Naturalmente, o estimador de Kaplan-Meier se reduz à função de sobrevivência empírica (2.25) se não existirem censuras. Este estimador também mantém esta forma em estudos envolvendo os mecanismos de censura do tipo I e II mas não atinge $\hat{S}\left(t\right) = 0$, pois as últimas observações são censuradas.

2.6.3 TTT-Plot (Gráfico Tempo Total em Teste)

Este método, como o próprio nome sugere, consiste em uma técnica gráfica utilizada para auxiliar na seleção de um modelo para a análise paramétrica dos dados de forma simples, verificando assim o ajuste do modelo.

Em muitas aplicações existe informação qualitativa e, muitas vezes, estrutural a respeito do fenômeno em questão, que pode ser utilizada na determinação empírica da forma da função de risco. Informações estruturais estão diretamente vinculadas ao conhecimento do pesquisador sobre o fenômeno, enquanto que informações qualitativas podem ser extraídas por meio de uma análise gráfica (Louzada, Mazucheli & Achcar, 2001).

Neste contexto, um gráfico conhecido como gráfico do tempo total em teste (curva TTT) é de grande utilidade e foi inicialmente proposto por Barlow & Campo (1975).

Para a obtenção desta curva tem-se a expressão (2.26) que segue, a qual, deve ser plotada versus r/n,

$$G\left(r/n\right) = \frac{\left[\left(\sum_{i=1}^{r} T_{(i:n)}\right) + (n-r)T_{(r:n)}\right]}{\sum_{i=1}^{r} T_{(i:n)}} \tag{2.26}$$

Utilizando a expressão anterior, tem-se a Figura 2.2 a qual, apresenta duas curvas TTT-Plot. Estas curvas foram obtidas através de valores gerados de duas distribuições Weibull: uma com os valores gerados de uma Weibull com parâmetros de forma e escala iguais a 2 e outra, que representa valores de uma Weibull com parâmetro de forma igual a 2 e escala igual a 3.

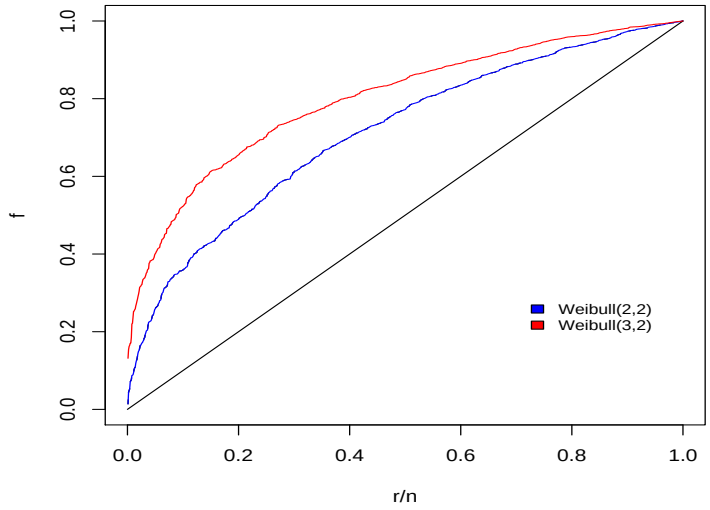


Figura 2.2: Exemplo de curva TTT-Plot.

Assim, a Figura 2.2 indica que a forma da função de risco é monótona crescente. Isto é comprovado pelo fato da mesma ter sido gerada por um modelo Weibull com parâmetro de forma maior que 1.

Por exemplo, se a curva apresentada fosse inicialmente convexa e no final côncava, a função de risco teria forma unimodal, sendo possíveis candidatos para o ajuste, o modelo log-logístico ou log-normal. Se fosse inicialmente convexa e então côncava, indicaria uma função de risco unimodal em forma de "U", sendo assim, as possíveis candidatas para o ajuste são as distribuições com risco duplo.

2.7 Modelos Paramétricos Usualmente Utilizados em Análise de Sobrevivência

Embora existam vários modelos probabilísticos, alguns destes ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas. Em particular, nesta

seção, serão abordadas as principais distribuições de probabilidade utilizadas na modelagem de dados de sobrevivência. Dentre essas distribuições, apresentamos a exponencial, Weibull, log-normal, valor extremo e log-logística quando se considera a existência de fração de fidelizados.

Essas distribuições apresentam algumas particularidades, tal como a distribuição exponencial, que acomoda funções de risco constantes. Enquanto que se a função de risco for monotonicamente crescente ou decrescente em t, tem-se uma distribuição Weibull. As distribuições log-logística e log-normal acomodam funções de risco multimodais e as em forma de "U"também podem ser observadas. Assim, nesta seção apresenta-se uma revisão dos modelos usuais de análise de sobrevivência.

2.7.1 Modelo Exponencial

A distribuição exponencial, é uma das mais simples e importantes distribuição de probabilidade utilizada na modelagem de dados que representam o tempo até a ocorrência de algum evento de interesse. A distribuição exponencial se caracteriza por ser a única distribuição que apresenta uma função de taxa de falha constante, ou seja, a função de risco independe do tempo (Lee & Wang, 2003).

A função de densidade de probabilidade da distribuição exponencial é dada por

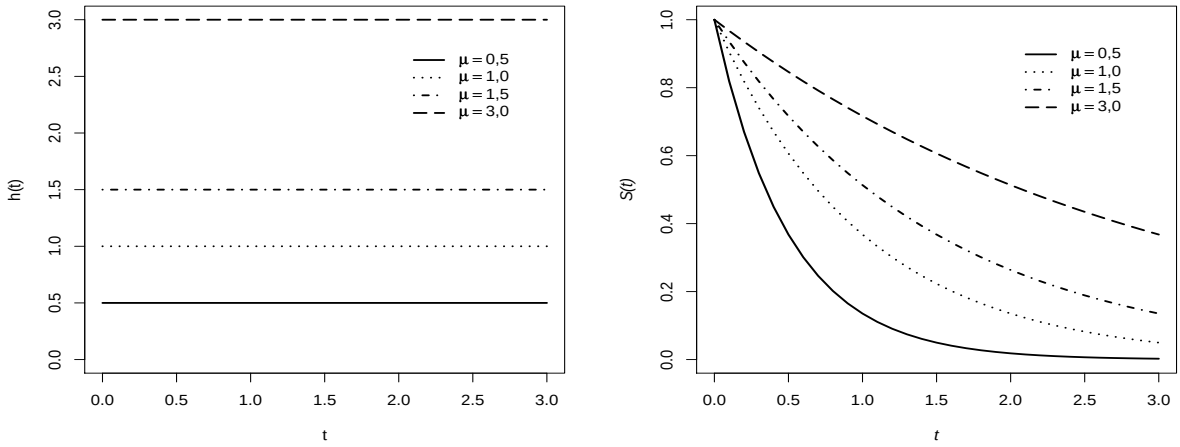
$$f_0(t) = \frac{1}{\mu} \exp\left(-\frac{t}{\mu}\right), \tag{2.27}$$

onde o parâmetro $\mu > 0$ é o tempo médio de vida e naturalmente tem a mesma unidade dos dados.

A partir de (2.27) tem-se que a função de sobrevivência, $S_0(t)$, é dada por

$$S_0(t) = \exp\left(-\frac{t}{\mu}\right). \tag{2.28}$$

A taxa de falha, ou a função de risco, associada à distribuição exponencial é, como já citado anteriormente, constante e igual a $\frac{1}{\mu}$. Isto significa que tanto um cliente antigo quanto um novo cliente, que ainda não experimentaram o evento de interesse, tem a mesma probabilidade de experimentá-lo em um dado intervalo de tempo. Esta importante propriedade da distribuição exponencial é denominada falta de memória. A Figura 2.3 apresenta estas propriedades apresentadas para o risco, bem como a função de sobrevivência para diferentes valores de μ.



(a) (b)

Figura 2.3: (a): Função de risco; (b): Função de sobrevivência.

2.7.2 Modelo Weibull

A distribuição Weibull foi proposta originalmente por W. Weibull em 1951 em estudos relacionados ao tempo de falha devido a fadiga de metais. Esta distribuição muito usada para descrever o tempo de vida de produtos industriais. Além disto, é muito importante na prática, pois apresenta uma grande variedade de formas para a função de risco (é possível a visualização na Figura 2.4-a), e todas com uma propriedade em comum: a sua taxa de falha é monótona, ou seja, ela é crescente para $\beta > 1$, decrescente para $\beta < 1$ ou constante para $\beta = 1$. No caso do parâmetro $\beta = 1$ tem-se a distribuição exponencial que é um caso particular da distribuição Weibull, dada quando a taxa de falha é constante.

A função densidade de probabilidade Weibull é dada por

$$f(t) = \frac{\beta}{\mu}\left(\frac{t}{\mu}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \tag{2.29}$$

onde a variável aleatória T é não negativa e, $\beta > 0$ e $\mu > 0$, são parâmetros de forma e escala, respectivamente (Lawless, 1982).

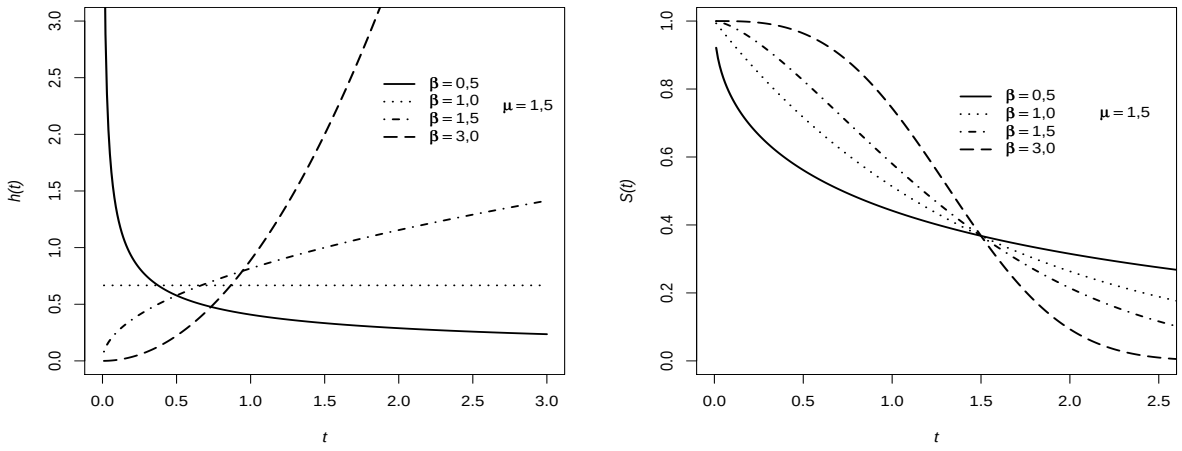
A partir de (2.29) , tem-se a função de sobrevivência,

$$S_0(t) = \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right]. \tag{2.30}$$

Para a visualização do comportamento das curvas de sobrevivência da distribuição Weibull, tem-se a Figura 2.4-b. A função de risco, ou taxa de falha, é dada por

$$h_0(t) = \left(\frac{1}{\mu}\right)^{\beta} \beta\,(t)^{\beta-1}. \tag{2.31}$$

para os valores do parâmetro β já descritos anteriormente.



(a) (b)

Figura 2.4: (a): Função de risco; (b): Função de sobrevivência.

É comum encontrar na literatura a distribuição Weibull escrita sob várias parametrizações, porém, a expressão dada em (2.29), permite uma interpretação direta do parâmetro de escala, $\mu \simeq 63^o$ percentil da distribuição da variável aleatória T. Uma forma alternativa de escrever a função densidade de distribuição Weibull é

$$f(t) = \lambda\beta t^{\beta-1} \exp\left[-\lambda t^{\beta}\right], \tag{2.32}$$

em que $\lambda = \mu^{-\beta}$.

2.7.3 Modelo Valor Extremo

Considerando-se o logaritmo dos tempos da distribuição Weibull, tem-se a distribuição de valor extremo. Portanto para

$$\begin{aligned} Y &\sim \log(T) \\ g^{-1}(y) &= \exp(y), \end{aligned}$$

onde a função densidade de probabilidade é

$$g(y) = f\left[g^{-1}(y)\right] \left| \frac{d}{dt} g^{-1}(y) \right|,$$

desenvolvendo tem-se

$$\begin{aligned} g(y) &= \frac{\beta}{\mu} \left[\frac{\exp(y)}{\mu}\right]^{\beta-1} \exp\left\{-\left[\frac{\exp(y)}{\mu}\right]^{\beta}\right\} \times \exp(y) \\ &= \frac{\beta}{\mu^{\beta}} \exp(\beta y - y) \exp\left[-\frac{\exp(y\beta)}{\mu^{\beta}}\right] \times \exp(y) \\ &= \frac{\beta \exp(\beta y)}{\mu^{\beta}} \exp\left[-\frac{\exp(y\beta)}{\mu^{\beta}}\right]. \end{aligned}$$

Reparametrizando, $\alpha = \log \mu$ e $\sigma = \beta^{-1}$ então,

$$\begin{aligned} \mu &= \exp \alpha \\ \beta &= \frac{1}{\sigma}. \end{aligned}$$

Assim, $g(y)$ é dada por

$$\begin{aligned} g(y) &= \frac{1}{\sigma} \frac{\exp\left(\frac{y}{\sigma}\right)}{\exp(\alpha)^{\frac{1}{\sigma}}} \exp\left[-\frac{\exp\left(\frac{y}{\sigma}\right)}{\exp(\alpha)^{\frac{1}{\sigma}}}\right] \\ &= \frac{1}{\sigma} \exp\left(\frac{y-\alpha}{\sigma}\right) \exp\left[-\exp\left(\frac{y-\alpha}{\sigma}\right)\right] \\ &= \frac{1}{\sigma} \exp\left[\left(\frac{y-\alpha}{\sigma}\right) - \exp\left(\frac{y-\alpha}{\sigma}\right)\right]. \end{aligned}$$

Reescrevendo, tem-se a função densidade de probabilidade dada por

$$f(t) = \frac{1}{\sigma} \exp\left[\left(\frac{t-\mu}{\sigma}\right) - \exp\left(\frac{t-\mu}{\sigma}\right)\right], \tag{2.33}$$

onde a variável aleatória T e o parâmetro μ são definidos para o intervalo $-\infty < \mu, t < \infty$, e σ é estritamente positivo.

A distribuição valor extremo apresenta uma particularidade muito importante para a análise de dados de confiabilidade pois apresenta uma taxa de falha acelerada e também tem uma representação de falha proporcional (Louzada, Mazucheli & Achcar, 2001).

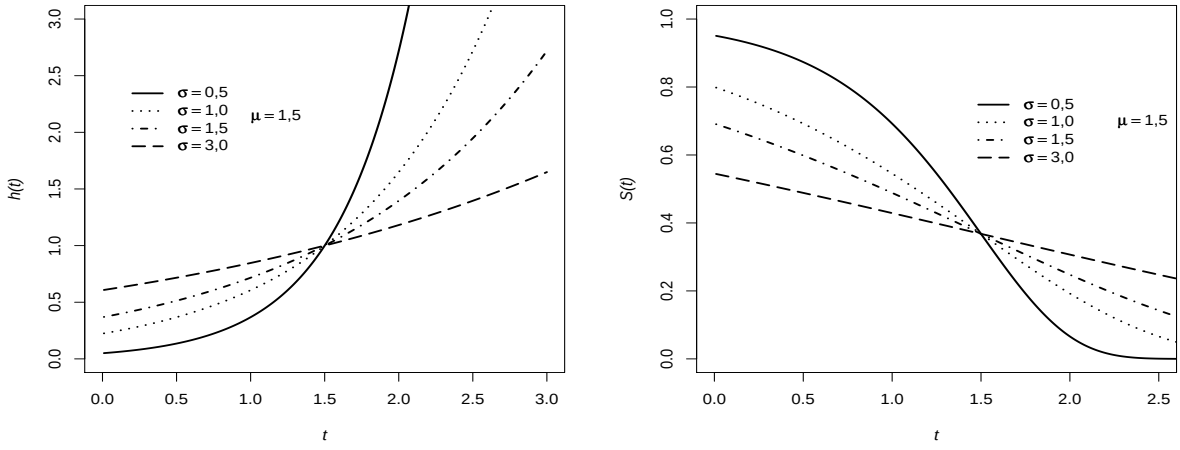
A partir da equação 2.33, tem-se a função de sobrevivência do modelo dada por

$$S_0(t) = \exp\left[-\exp\left(\frac{t-\mu}{\sigma}\right)\right], \tag{2.34}$$

para os mesmos intervalos definidos anteriormente. A função de risco, ou taxa de falha, é definida como sendo

$$h_0(t) = \exp\left(\frac{t-\mu}{\sigma}\right). \tag{2.35}$$

A Figura 2.5 apresenta as funções de risco e sobrevivência do modelo valor extremo.



(a) (b)

Figura 2.5: (a): Função de risco; (b): Função de sobrevivência.

2.7.4 Modelo Log-Normal

Para o modelo log-normal, tem-se a função densidade de probabilidade definida como sendo (Klein & Moeschberger, 1997)

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{[\log(t) - \mu]^2}{2\sigma^2}\right\}, \tag{2.36}$$

onde a variável aleatória $t \geq 0$, $-\infty < \mu < \infty$ e $\sigma > 0$.

Como o próprio nome sugere, existe uma relação entre a distribuição log-normal e a distribuição normal, o que facilita a apresentação e análise de dados provenientes da distribuição log-normal. Ou seja, o logaritmo de uma variável com distribuição normal com parâmetros μ e σ tem uma distribuição log-normal com média μ e desvio padrão σ, ou variância igual a σ^2.

Portanto, dados provenientes de uma distribuição o log-normal podem ser analisados segundo uma distribuição normal, se trabalhar com o logaritmo dos dados ao invés dos valores originais.

A função de sobrevivência da distribuição log-normal é dada por

$$S_0(t) = 1 - \phi\left(\frac{\log(t) - \mu}{\sigma}\right), \tag{2.37}$$

onde ϕ é a função de distribuição acumulada de uma variável com distribuição normal padrão.

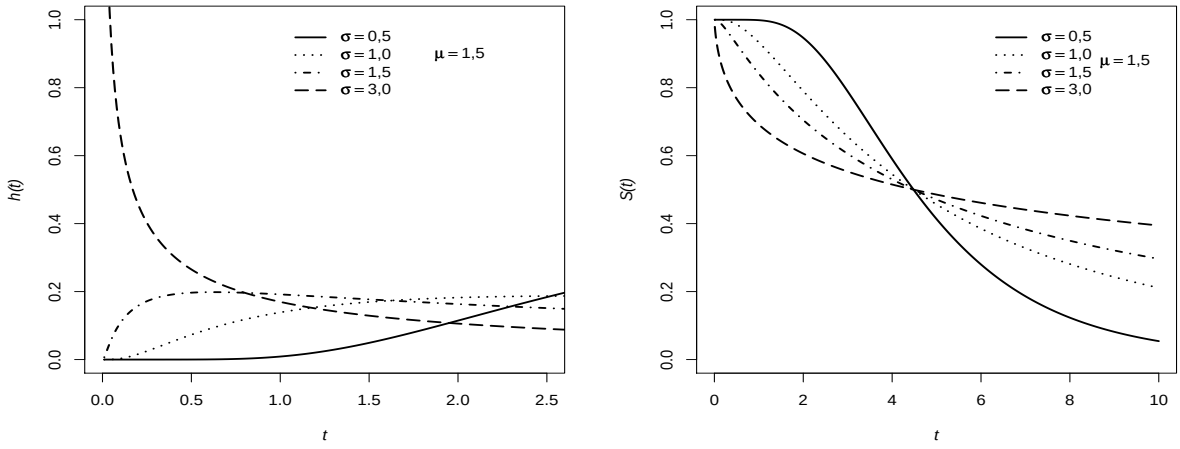
A função de risco é dada por

$$h_0(t) = \frac{f(t)}{S(t)},$$

ou escrita em função da distribuição normal como sendo

$$h_0(t) = \frac{\frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{[\log(t)-\mu]^2}{2\sigma^2}\right\}}{1 - \phi\left(\frac{\log(t)-\mu}{\sigma}\right)}.$$

A função de risco e sobrevivência do modelo lognormal pode sre visualizada através da Figura 2.6.



(a) (b)

Figura 2.6: (a): Função de risco; (b): Função de sobrevivência.

2.7.5 Modelo Log-Logístico

Uma variável aleatória não negativa T, segue uma distribuição log-logística com parâmetros $\mu > 0$ e $\beta > 0$, se o logaritmo, $Y = \log(T)$, tem distribuição logística com densidade dada por (Klein & Moeschberger, 1997)

$$f(y) = \frac{\exp\left[(y-\mu)/\sigma\right]}{\sigma\left\{1+\exp\left[(y-\mu)/\sigma\right]\right\}^2}, \tag{2.38}$$

onde $-\infty < \mu < \infty$ e $\sigma > 0$ são os parâmetros de locação e escala, respectivamente.

Da mesma forma que para o modelo de valor extremo, pode-se considerar a reparametrização da forma $Y = \ln(T) = \mu + \sigma W$, onde $W \sim$Logística$(0,1)$ que é apresentada em detalhes por Louzada, Mazucheli & Achcar (2001). As Figuras 2.7-a e b representam respectivamente as formas das funções de risco e sobrevivência da distribuição logística.

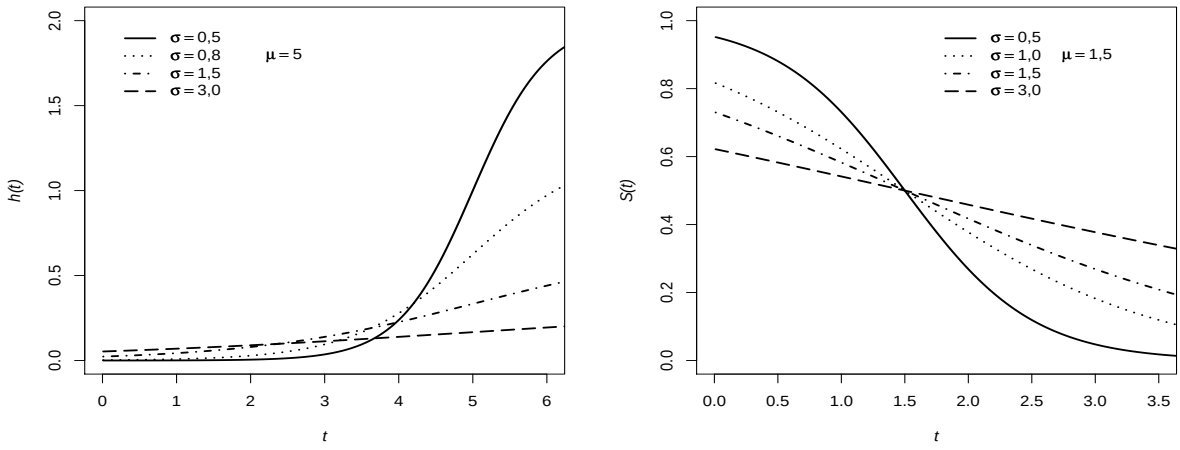
A função de sobrevivência do modelo log-logístico é expressa como sendo

$$S_0(t) = \frac{1}{1+\exp\left[(y-\mu)/\sigma\right]},$$

e a função de risco é dada por

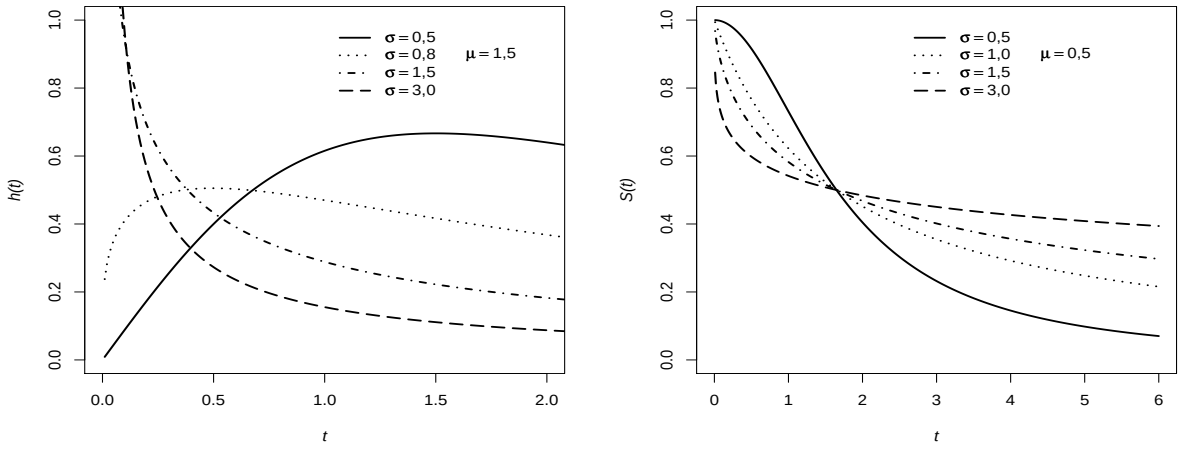
$$h_0(t) = \frac{1}{\sigma\{1 + \exp[(y-\mu)/\sigma]\}}. \tag{2.39}$$

As Figuras 2.8-a e b representam as formas das funções de risco e sobrevivência da distribuição log-logística.



(a) (b)

Figura 2.7: (a): Função de risco; (b): Função de sobrevivência.



(a) (b)

Figura 2.8: (a): Função de risco; (b): Função de sobrevivência.

Capítulo 3

Análise de Sobrevivência Aplicada a Finanças

As instituições financeiras, na entrada do atual século, estão se conscientizando da necessidade de intensificar o foco no *Marketing* de Relacionamento para manter e ampliar seus mercados. Os bancos e outras instituições financeiras estão aprendendo que podem reduzir as taxas de cancelamento voluntário e melhorar a rentabilidade, vendendo mais produtos e serviços para atuais clientes, através de técnicas conhecidas como *Marketing* de Relacionamento; que pode ser definido como um conjunto de estratégias de negócio pelas quais a empresa objetiva construir um relacionamento com seus clientes de maneira personalizada e duradoura. Para isso é essencial que a mesma dedique-se a uma constante melhora nesse relacionamento, para que ambas as partes sejam beneficiadas (Oliveira, 2000).

A gestão do relacionamento do cliente (*Customer Relationship Management*) está totalmente ligada à filosofia do *Marketing* de Relacionamento aliado à utilização constante de informação.

A utilização desse tipo de gestão já era praticada no passado. Um pequeno comércio atendia as pessoas de forma personalizada e quase sempre, o próprio dono do negócio conhecia seus clientes pelo nome, bem como seus hábitos e preferências de compras.

O conceito de gestão de relacionamento de cliente parte da premissa de que é de cinco até dez vezes mais caro obter um novo cliente do que reter os existentes, e que o importante não é ter uma imensa carteira de clientes, mas uma boa base de clientes fiéis (Oliveira, 2000).

Cada vez mais as empresas estão se conscientizando que nada adianta gastar com campanhas publicitárias e de *marketing* se não for possível manter os clientes fiéis a seus produtos e/ou serviços.

Em mercados altamente competitivos, a fidelização e retenção de clientes têm papel fundamental nas empresas. Entende-se por fidelizados, os clientes que apresentam baixo risco de não pagar aos empréstimos feitos ou, clientes que não vão deixar de fazer parte da carteira de determinada instituição (estes clientes serão tratados aqui no contexto de longa-duração) (Oliveira, 2000). Identificar, antecipadamente, clientes com alto potencial de ruptura de relacionamento permitindo ações preventivas, tornou-se imprescindível para a manutenção ou acréscimo da lucratividade das empresas.

Nesta seção, serão apresentados alguns conceitos em modelagem de longa-duração, bem como os modelos paramétricos, métodos de estimação usados em sobrevivência e confiabilidade, enfatizando suas principais propriedades. A metodologia foi aplicada a um estudo da área financeira.

3.1 Modelos de Longa-Duração

Os modelos em análise de sobrevivência com longa-duração, possuem vantagem em relação aos modelos de sobrevivência usuais pois incorporam a heterogeneidade de duas subpopulações (susceptíveis e imunes) e são conhecidos também como modelos com fração de imunes (*cure rate models*).

Dentre esses modelos, o tipo mais comum é o modelo de mistura (*mixture model*) no qual se considera que a população é dividida em duas subpopulações: imunes (ou, fidelizados) e susceptíveis (ou não fidelizados) ao evento de interesse.

Para a análise de dados com longa-duração, isto é, quando é previsto uma porcentagem de não ocorrência do evento de interesse na população, vários modelos foram formulados sendo alguns mais antigos, Lawless (1982) e Berckson & Gage (1952), e outros mais recentes, por exemplo, os apresentados por Chen, Ibrahim & Sinha (1999). Neste contexto, enfocaremos principalmente, o modelo proposto por Berkson & Gage (1952).

No artigo escrito por Berkson & Gage (1952), trabalhou-se o exemplo de pacientes com câncer submetido a um tratamento usual e outro tratamento novo para a época. Desta forma, verificou-se que, durante o estudo, alguns pacientes deixaram de experimentar o evento de interesse, neste caso a morte. Quando isto ocorria, a curva de sobrevivência era afetada de maneira que esta se estabilizava na proporção de pacientes curados.

Para solucionar este problema, os autores propuseram estimar, além dos parâmetros do modelo, um outro parâmetro que indicaria o percentual de pacientes curados.

A partir deste artigo, muitos outros foram escritos considerando o modelo proposto não só para dados relacionados a experimentos na área da saúde mas também para experimentos de finanças, sinistros em seguradoras dentre outras áreas.

Assim, na área de finanças que é o foco deste trabalho, admite-se que os indivíduos podem ser classificados como fidelizados (sem possibilidade de apresentar o evento de interesse) com probabilidade p, ou ser não fidelizados com probabilidade $q = 1-p$. A cada indivíduo associamos uma variável aleatória T, representando o tempo até a ocorrência do evento de interesse ou até a censura.

Desta forma, dada uma função de sobrevivência, $S(t)$, temos que $\lim_{t\to\infty} S(t) = p$, onde p é a proporção de não ocorrência do evento de interesse na população. O modelo proposto por Berkson & Gage (1952) é caracterizado pela função de sobrevivência que é dada por

$$S(t) = P\left(T > t\right) = (1 - p) + p \times S_0\left(t\right) \tag{3.1}$$

em que $S_0\left(\cdot\right)$ é a função de sobrevivência para indivíduos não fidelizados, de tal forma que para $t \to +\infty$, $S_0(t) \to 0$ e assim, $\lim_{t\to\infty} S(t) = (1 - p)(> 0)$. A função $S_0(t)$ é especificada por funções de sobrevivência de modelos paramétricos tais como, as funções de sobrevivência dos modelos Weibull, exponencial, logístico, etc (Maller & Zhou 1996).

Também, as relações entre $S(t)$, $h(t)$ e $f(t)$ já citadas no capítulo anterior, podem ser utilizadas neste contexto. Assim, a função de risco com fração de fidelizados, é dada pela seguinte expressão:

$$h_p(t) = \frac{f_p(t)}{S_p(t)} \tag{3.2}$$

e, a função densidade de probabilidade dada por

$$f_p(t) = (1 - p)f(t). \tag{3.3}$$

Nos modelos de análise de sobrevivência com fração de fidelizados apresentados, os indivíduos fiéis não são identificados a priori, mas a presença dos mesmos pode ser inferida se a quantidade de indivíduos censurados no estudo for suficientemente grande.

3.1.1 Modelo Weibull de Longa-Duração

Para o caso Weibull de longa-duração, usando-se do modelo proposto por Berckson & Gage acima, tem-se a função de sobrevivência dada por

$$S(t) = P(T > t) = (1 - p) + p \times \exp\left[-\left(\frac{1}{\mu}\right)^{\beta}\right], \tag{3.4}$$

onde $\mu > 0$ é o parâmetro de escala da distribuição Weibull, p é o percentual de imunes ao evento de interesse ou, no caso de finanças, pode-se considerar p como sendo o percentual de clientes fidelizados na população em estudo ou ainda, clientes imunes ao evento de interesse em estudo.

Assim, a função de verossimilhança, contemplando a presença de dados censurados a direita, é dada por

$$L(\theta) = \prod_{i=1}^{n} [f_p(t_i;\theta)]^{\delta_i} [S_p(t_i;\theta)]^{1-\delta_i}. \tag{3.5}$$

Para o modelo Weibull de longa-duração tem-se a função de verossimilhança como sendo

$$\begin{aligned} L(p,\mu,\beta|\mathbf{t},\boldsymbol{\delta}) &= \prod_{i=1}^{n} \left[(1-p)\frac{\beta}{\mu}\left(\frac{t_i}{\mu}\right)^{\beta-1} \exp\left[-\left(\frac{1}{\mu}\right)^{\beta}\right]\right]^{\delta_i} \\ &\quad \left[p + (1-p)\exp\left[-\left(\frac{t_i}{\mu}\right)^{\beta}\right]\right]^{1-\delta_i} \end{aligned} \tag{3.6}$$

e o logaritmo desta função de verossimilhança é dado por

$$\begin{aligned} l(p,\mu,\beta|\mathbf{t},\boldsymbol{\delta}) &= \ln[(1-p)\beta]\sum_{i=1}^{n}\delta_i - \beta\ln(\mu)\sum_{i=1}^{n}\delta_i + (\beta-1)\sum_{i=1}^{n}\delta_i \ln t_i - \sum_{i=1}^{n}\delta_i\left(\frac{t_i}{\mu}\right)^{\beta} \\ &\quad + \sum_{i=1}^{n}(1-\delta_i)\log\left[p + (1-p)\exp\left[-\left(-\frac{t_i}{\mu}\right)^{\beta}\right]\right]. \end{aligned} \tag{3.7}$$

A fim de obter as estimativas dos parâmetros da log-verossimilhança, fez-se uma reparametrização onde,

$$\begin{aligned} \mu_0 &= \ln(\mu) \\ \beta_0 &= \ln(\beta) \\ p_0 &= \ln\left(\frac{p}{1-p}\right). \end{aligned}$$

Esta reparametrização foi proposta com o intuito de garantir que os parâmetros estimados, μ e β, sejam sempre positivos e que o parâmetro estimado p esteja no intervalo $[0,1]$ (Maller e Zhou, 1996).

Como a distribuição exponencial é um caso particular da distribuição Weibull quando o parâmetro de forma $\beta = 1$, tem-se a função de sobrevivência de longa-duração para este

caso dada por

$$S(t) = P(T > t) = (1 - p) + p \times \exp\left[-\left(\frac{1}{\mu}\right)\right], \tag{3.8}$$

onde $\mu > 0$ e p é percentual de clientes fidelizados na população em estudo ou imunes ao evento de interesse.

Assim, a função de verossimilhança para o modelo exponencial é dada por

$$L(p, \mu|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left[(1-p)\frac{1}{\mu}\exp\left[-\left(\frac{1}{\mu}\right)\right]\right]^{\delta_i} \left[p + (1-p)\exp\left[-\left(\frac{t_i}{\mu}\right)\right]\right]^{1-\delta_i} \tag{3.9}$$

e o logaritmo da função de verossimilhança é dado por

$$\begin{aligned} l(p, \mu, \beta|\mathbf{t}, \boldsymbol{\delta}) &= \ln[(1-p)]\sum_{i=1}^{n}\delta_i - \ln(\mu)\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}\delta_i\left(\frac{t_i}{\mu}\right) \\ &\quad + \sum_{i=1}^{n}(1-\delta_i)\log\left[p + (1-p)\exp\left[-\left(-\frac{t_i}{\mu}\right)\right]\right]. \end{aligned} \tag{3.10}$$

A fim de obter as estimativas dos parâmetros da log-verossimilhança, fez-se também uma reparametrização onde,

$$\begin{aligned} \mu_0 &= \ln(\mu) \\ p_0 &= \ln\left(\frac{p}{1-p}\right). \end{aligned}$$

3.1.2 Modelo Log-Logístico de Longa-Duração

Para o modelo log-logístico de longa-duração, tem-se a função de sobrevivência dada por

$$S(t) = (1 - p) + \frac{p}{1 + \exp\left[(t - \mu)/\sigma\right]}, \tag{3.11}$$

onde $-\infty < \mu < \infty$, $\sigma > 0$ e $0 \leq p \leq 1$.

Assim, a função de verossimilhança para a distribuição log-logística na presença de tempos com longa-duração é dada por

$$
\begin{aligned}
L(p,\mu,\sigma|\mathbf{t},\boldsymbol{\delta}) &= \prod_{i=1}^{n}\left[f_p(t_i;\theta)\right]^{\delta_i}\left[S_p(t_i;\theta)\right]^{1-\delta_i} \qquad (3.12)\\
&= \prod_{i=1}^{n}\left[(1-p)\frac{\exp\left[(t-\mu)/\sigma\right]}{\sigma\left[1+\exp\left((t-\mu)/\sigma\right)^2\right]}\right]^{\delta_i}\left[(1-p)+\frac{p}{1+\exp\left[(t-\mu)/\sigma\right]}\right]^{1-\delta_i}
\end{aligned}
$$

e o logaritmo da função de verossimilhança é dado por

$$
\begin{aligned}
l(p,\mu,\sigma|\mathbf{t},\boldsymbol{\delta}) &= \log(1-p)\sum_{i=1}^{n}\delta_i+\sum_{i=1}^{n}\delta_i\log\left[\exp\left(\frac{t-\mu}{\sigma}\right)\right] \qquad (3.13)\\
&\quad -\sum_{i=1}^{n}\delta_i\log\left[\sigma\left(1+\exp\left(\frac{t-\mu}{\sigma}\right)\right)^2\right]\\
&\quad +\sum_{i=1}^{n}(1-\delta_i)\log\left[p+\frac{1-p}{1+\exp\left(\frac{t-\mu}{\sigma}\right)}\right].
\end{aligned}
$$

Também, para garantir a positividade do parâmetro σ e para que o parâmetro p esteja no intervalo $[0,1]$, a seguinte reparametrização será considerada:

$$
\begin{aligned}
\mu_0 &= \ln(\mu)\\
\sigma_0 &= \ln(\sigma)\\
p_0 &= \ln\left(\frac{p}{1-p}\right).
\end{aligned}
$$

3.2 Método de estimação dos Parâmetros

Os modelos probabilísticos apresentados no Capítulo 2 são caracterizados por quantidades desconhecidas, denominadas parâmetros, que devem ser estimados a partir de observações amostrais para que seja possível responder as perguntas de interesse.

Existem vários métodos de estimação dos parâmetros. O primeiro a ser citado, e talvez o mais conhecido, é o método de mínimos quadrados ordinários, geralmente apresentado em cursos de estatística dentro do contexto de regressão linear. Por não ser capaz de incorporar censuras no seu processo de estimação, este método é inapropriado para estudos em análise de sobrevivência e confiabilidade, principalmente em análises onde se tem o interesse em estudar uma população imune ao evento de interesse.

O método de máxima verossimilhança, dentro deste contexto, mostra-se mais apropriado para este tipo de dados, uma vez que este, além de incorporar as censuras, é relativamente simples de ser entendido e possui propriedades ótimas para grandes amostras, Bickel & Doksum (1977). Na Seção 3.2.1 apresenta-se o método de máxima verossimilhança para dados censurados em modelos de longa duração.

3.2.1 Método de Máxima Verossimilhança

O método da máxima verossimilhança trata o problema de estimação da seguinte forma: baseado nos resultados obtidos pela amostra, busca qual é a distribuição, entre todas aquelas definidas pelos possíveis valores de seus parâmetros, com maior possibilidade de ter gerado tal amostra. Em outras palavras, se por exemplo a distribuição do tempo de falha é a distribuição Weibull, para cada combinação diferente de μ e β tem-se diferentes distribuições Weibull, e o estimador de máxima verossimilhança escolhe aquele par de μ e β que melhor explique a amostra observada.

A seguir, a idéia do método de máxima verossimilhança é traduzida para conceitos matemáticos a fim de que seja possível obter estimadores para os parâmetros. Suponha, inicialmente, uma amostra de observações $t_1, \ldots, t_n$ de uma certa população de interesse em que todas são não censuradas.

Suponha, ainda, que a população é caracterizada pela sua função de densidade de probabilidade $f(t)$. Por exemplo, se $f(t) = (1/\mu)\exp(-t/\mu)$, significa que as observações vêm de uma distribuição exponencial com parâmetro μ a ser estimado. A função de verossimilhança para um parâmetros genérico θ desta população é então expressa por

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta). \tag{3.14}$$

A dependência de f em θ é preciso agora ser mostrada pois L é função de θ. Nesta expressão, θ pode estar representando um único parâmetro ou um conjunto de parâmetros. No modelo log-normal por exemplo, $\theta = (\mu, \sigma)$. A tradução, em termos matemáticos, para a frase "a distribuição que melhor explique a amostra observada"é encontrar o valor de θ que maximize a função $L(\theta)$. Isto é, o valor de θ que maximize a probabilidade da amostra observada ocorrer.

A função de verossimilhança $L(\theta)$ mostra que a contribuição de cada observação não censurada é a sua função densidade. A contribuição de cada observação censurada não é, contudo, a sua função densidade. Estas observações somente nos informam que o tempo de falha é maior que o tempo de censura observado e, portanto, que a sua contribuição

para $L(\theta)$ é a sua função de sobrevivência $S(t)$. As observações podem então ser divididas em dois conjuntos, as r primeiras são as não censuradas $(1, 2, \ldots, r)$, e as $n - r$ seguintes, são as censuradas $(r + 1, r + 2, \ldots, n)$.

A função de verossimilhança assume assim a seguinte forma

$$L(\theta) = \prod_{i=1}^{r} [f(t_i; \theta)]^{\delta_i} \prod_{j=r+1}^{n} [S(t_j; \theta)]^{1-\delta_j} \tag{3.15}$$

ou equivalentemente,

$$L(\theta) = \prod_{i=1}^{n} [h(t_i; \theta)]^{\delta_i} S(t_i; \theta), \tag{3.16}$$

em que δ_i é a variável indicadora de falha ou censura apresentada na Seção 2.3, Lee & Wang (2003), Lawless (1982). A expressão (3.15), ou (3.16) para a função de verossimilhança é válida para os mecanismos de censura do tipo I e II, censura aleatória e sob a suposição de que o mecanismo de censura é não-informativo (não carrega informações sobre o parâmetro). Por questões computacionais e de simplificação, é sempre conveniente, no entanto, trabalhar com o logaritmo da função de verossimilhança.

Os estimadores de máxima verossimilhança são os valores de θ, vetor de parâmetros, que maximizam $L(\theta)$, função de verossimilhança em questão, ou, de forma equivalente, $\log(L(\theta)) = l(\theta)$. Os θ's são encontrados resolvendo-se o sistema de equações

$$U(\theta) = \frac{\partial l(\theta)}{\partial \theta} = 0. \tag{3.17}$$

Quando se considera modelos de longa-duração, existe um parâmetro no modelo que caracteriza a presença da longa-duração nos dados. Sendo assim, a função de verossimilhança para os modelos com longa-duração é dada por

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f_p(t_i; \theta)]^{\delta_i} [S_p(t_i; \theta)]^{1-\delta_i} \\ &= \prod_{i \in \overline{C}} f_p(t_i; \theta) \prod_{i \in C} S_p(t_i; \theta) \\ &= \prod_{i \in \overline{C}} [(1-p) f_p(t_i; \theta)] \prod_{i \in C} [p + (1-p) S_p(t_i; \theta)], \end{aligned} \tag{3.18}$$

onde, C é o conjunto de censurados e $\overline{C}$ o conjunto de não censurados ou observados.

Os modelo considerados para dados de sobrevivência, em sua maioria, não possuem uma expressão fechada para as equações não lineares obtidas através do método de máxima verossimilhança. Desta forma, para a estimação dos parâmetros, deve-se usar

algum método numérico, como por exemplo o método de Newton Raphson (Lee & Wang, 2003).

3.3 Critérios de Seleção de Modelos

Um problema importante consiste da avaliação e escolha do modelo que melhor represente a situação em estudo.

Apresenta-se aqui, algumas métricas usuais para seleção de modelo e, propõem-se também uma forma alternativa de seleção.

3.3.1 Critérios AIC e BIC

Para avaliar o ajuste do modelo, utilizou-se: AIC, BIC e distância entre curvas (medida através da norma Euclidiana), sendo a distância entre curvas uma alternativa as métricas usuais.

O primeiro método de seleção de modelo que será apresentado aqui, é o critério AIC (*Akaike Information Criterio*) definido por

$$\Delta AIC = -2\ln\left[\frac{\sup_{M_1} f(x|\theta_1, M_1)}{\sup_{M_2} f(x|\theta_2, M_2}\right] - 2(d_2 - d_1), \tag{3.19}$$

onde, d_i, $i = 1, 2$, representa o número de parâmetros de cada modelo. Este critério é baseado em considerações frequentistas de eficiência assintótica, Akaike (1973).

O segundo método em questão, é critério BIC (*Bayesian Information Criterio*). Assim, para o cálculo, tem-se que

$$\Delta BIC = -2\ln\left[\frac{\sup_{M_1} f(x|\theta_1, M_1)}{\sup_{M_2} f(x|\theta_2, M_2}\right] - (d_2 - d_1)\ln n, \tag{3.20}$$

onde, n é o tamanho da amostra em questão e d_i, $i = 1, 2$, é o número de parâmetros de cada modelo (Schwarz, 1978).

Os dois critério apresentados, AIC e BIC, têm como objetivo introduzir a complexidade do modelo no critério de seleção, pois são critérios que "penalizam"a verossimilhança (Paulino, Turkman & Murteira, 2003).

Carlin e Louis (2000) sugerem a seguinte expressão para o valor de BIC,

$$BIC_i = 2E[\ln L(\theta_i|x, M_i)] - d_i \ln n, \tag{3.21}$$

onde, n é a dimensão da amostra e d_i o número de parâmetros do modelo M_i e, o critério de seleção, neste caso em particular, é o modelo que apresenta maior valor de BIC.

Uma forma simplificada para o cálculo destes critérios é dado por

$$AIC = -2l(\boldsymbol{\theta}) + 2d \tag{3.22}$$

$$BIC = -2l(\boldsymbol{\theta}) + d \log n, \tag{3.23}$$

em que $l(\boldsymbol{\theta})$ representa o máximo da função de log-verossimilhança, d é a dimensão do modelo, e n o número de observações (tempos em estudo). Estas expressões são apresentadas pelo *software SAS* e serão usadas durante a aplicação e o estudo de simulação que serão apresentados a seguir.

Contrário ao que foi dito anteriormente, quando optamos por usar as equações simplificadas, o modelo que mais se adequa aos dados em questão, é aquele que apresenta menor valor para AIC e BIC Paulino, Turkman & Murteira (2003). Os critérios acima também foram apresentados por Anderson& Burnham (2004).

3.3.2 Norma Euclidiana Para Seleção do Modelo

Alternativamente, para selecionar o modelo que mais se aproxima dos dados, tem-se a norma Euclidiana, que consiste em medir as distâncias entre pontos da curva empírica dos dados e, pontos da curva plotada a partir de modelos paramétricos.

Assim, optaremos sempre pela curva que tiver mais proximidade, ou seja, menor distância com a curva empírica em análise de sobrevivência, neste caso trabalharemos com as curvas atuariais e Kaplan-Meier anteriormente apresentadas nas Seções 2.6.1 e 2.6.2. Desta forma tem-se

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}, \tag{3.24}$$

onde a representa o valor da curva estimada pela empírica de sobrevivência (curva atuarial ou Kaplan-Meier), b o valor estimado pela sobrevivência do modelo paramétrico em teste e n o número de valores plotados.

3.4 Estudo de Um Caso na Área Financeira

Nesta seção considerou-se dados reais fornecidos por uma instituição financeira brasileira de 65.535 cadastros de clientes, onde o interesse é observar o tempo em que o cliente deixa de pagar determinado empréstimo. Quando o cliente deixa de pagar o empréstimo fornecido pela instituição, o seu tempo é dito observado. Para o caso em que o cliente não paga como foi acordado, o seu tempo é censurado.

Para os dados fornecidos tem-se a presença de 41.787 censuras, ou seja, $63,76\%$ dos clientes têm seus tempos censurados, isto é, são clientes fidelizados. O tempo máximo observado no estudo foi de 201 meses e o mínimo 0 meses. Os tempos iguais a zero foram considerados clientes que não iniciaram um relacionamento com a instituição e, desta forma, foram descartados da análise (em um total de 5 clientes).

A Figura 3.1 apresenta o TTT-Plot dos dados da instituição.

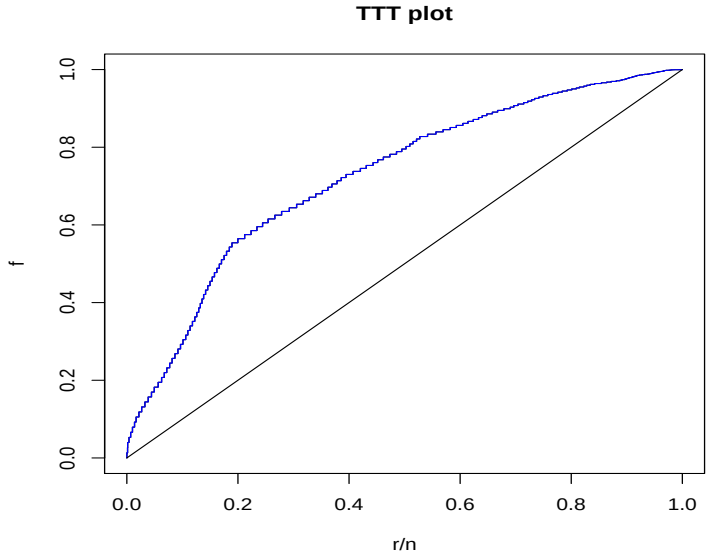


Figura 3.1: TTT-Plot para os tempos até o não pagamento dos empréstimos.

Como apresentado na Seção 2.6.3, a Figura 3.1 indica que a forma da função de risco é monótona crescente. Sendo assim, um possível modelo para ajuste deste dados, seria

o modelo Weibull com parâmetro de forma maior que 1. Outra distribuição candidata para ajustar os tempos até o não pagamento dos empréstimos é a função log-logítica que também apresenta esta forma na função de risco.

Inicialmente, na Figura 3.2 apresenta-se a curva estimada via Kaplan-Meier para os tempos até o não pagamento dos empréstimos. Observa-se nesta figura que um modelo que se adequaria a curva estimada seria um modelo de longa-duração e não modelos usuais em análise de sobrevivência, uma vez que a curva estimada para a sobrevivência não tende a zero como é esperado em situações de susceptíveis ao evento de interesse.

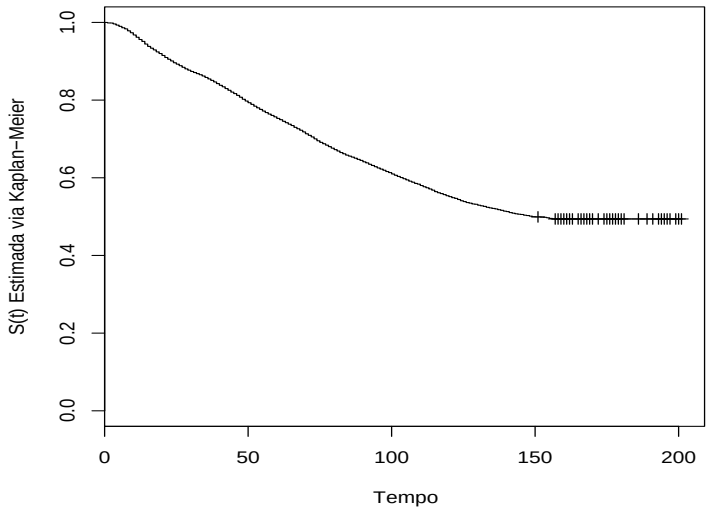


Figura 3.2: Curva estimada via Kaplan-Meier para os tempos até o não pagamento dos empréstimos.

Na tentativa de ajustar um modelo para os tempos até o não pagamento dos empréstimos, usaremos o modelo de Berkson & Gage (1952) apresentado anteriormente.

Para estes dados, será ajustado o modelo Weibull com longa-duração apresentado na Seção 3.1.1. Assim, utilizando-se da função de log-verossimilhança (3.7), através do método de Newton-Raphson, tem-se na Tabela 3.1, as estimativas dos parâmetros do modelo Weibull de longa-duração, os erros padrões das estimativas e os p-valores. Considerou-se para a obtenção das estimativas as seguintes reparametrizações: $\mu = \exp(\mu_0)$, $\beta = \exp(\beta_0)$ e $p = \frac{\exp(\gamma_0)}{(1+\exp(\gamma_0))}$. Os p-valores listados nessa tabela e em todos

as outras apresentadas nessa seção, são fornecidos pelo *software SAS*. O p-valor descrito representa o resultado da hipótese nula, obtido do teste bilateral baseado na estatística t-Student, medindo o quão significativo é o valor estimado para o modelo em questão. Nesse caso tem-se todas as estimativas dos parâmetros significativas ao nível de 5% de significância, representado que os parâmetros são significativos.

Tabela 3.1: Estimativas dos parâmetros (μ, β e p) do modelo Weibull.

Parâmetros	Estimativas	Erro Padrão	P-valor	Reparametrização
μ_0	$4,596283$	$0,020762$	$< 0,0001$	$99,1152$
β_0	$0,264239$	$0,008086$	$< 0,0001$	$1,3024$
γ_0	$0,465302$	$0,040781$	$< 0,0001$	$0,6143$

Através da tabela acima, verifca-se que o 63^o estimado através do modelo Weibull é de aproximadamente 99 meses; a função de risco tem forma crescente uma vez que, o parâmetro de forma estimado pela distribuição é de $1,30$; o p estimado é de aproximadamente 61% indicando o percentual de clientes que pagará o empréstimo de forma conveniente.

A Figura 3.3 apresenta a curva ajustada através do modelo Weibull de longa-duração junto a curva estimada pelo método de Kaplan-Meier.

Com o intuito de verificar os procedimentos de escolha de modelo, apresentados na Seção 3.3, um caso particular do modelo Weibull, o modelo exponencial, será também ajustado aos dados assim como o modelo log-logístico de longa-duração.

Para o modelo exponencial de longa-duração (3.9), e o ajuste do modelo apresentado na Figura 3.4, Considerando-se a reparametrização apresentada também na Seção 3.1.1, tem-se as estimativas para esta aplicação.

Considerando o modelo log-logístico 3.12, reparametrizações para o modelo sendo $\mu =$

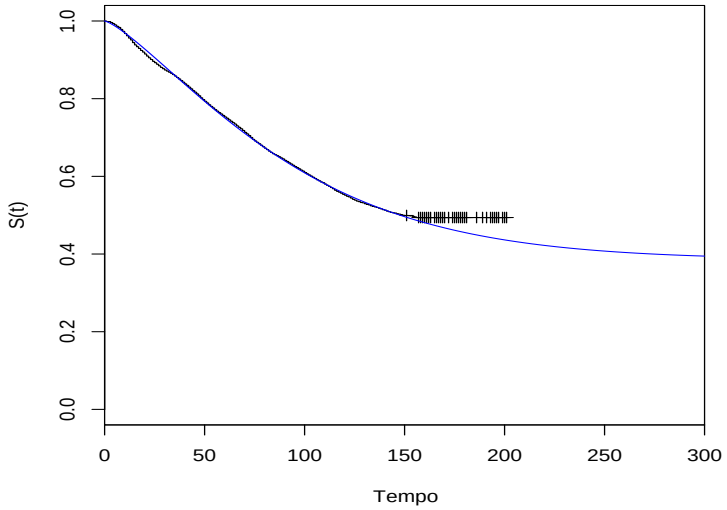


Figura 3.3: Curva estimada via Kaplan-Meier e curva ajustada através do modelo Weibull para os tempos.

Tabela 3.2: Estimativas dos parâmetros (μ e p) do modelo exponencial.

Parâmetros	Estimativas	Erro Padrão	P-valor	Reparametrização
μ_0	$5,337256$	$0,006491$	$< 0,0001$	$207,9413$
γ_0	$13,168575$	$44,161371$	$0,76555$	$0,9999$

$\exp(\mu_0)$, $\sigma = \exp(\sigma_0)$ e $p = \frac{\exp(\gamma_0)}{(1+\exp(\gamma_0))}$, os resultados do ajuste são apresentados na Tabela 3.3.

Tabela 3.3: Estimativas dos parâmetros (μ, σ e p) do modelo log-logístico.

Parâmetros	Estimativas	Erro Padrão	P-valor	Reparametrização
μ_0	$4,062554$	$0,005528$	$< 0,0001$	$58,1226$
σ_0	$3,083985$	$0,006743$	$< 0,0001$	$21,8453$
γ_0	$-0,129251$	$0,011177$	$< 0,0001$	$0,4677$

Na Figura 3.4 observa-se que, as curvas de sobrevivência ajustadas pelos modelos Weibull e exponencial, estão muito próximas da curva estimada via Kaplan-Meier inicialmente, entretanto, o modelo exponencial apresenta dificuldade para captar tempos com longa-duração. Também, é possível verificar que o modelo log-logístico não se ajusta a estes dados.

Da Tabela 3.2, tem-se que, para o modelo Weibull e log-logístico, respectivamente, $p = 0,6143$ e $p = 0,4677$, ou seja, aproximadamente $61,4\%$ e $46,8\%$. Para o modelo exponencial, $p = 0,9999$, ou seja, aproximadamente 1, e com erro padrão muito grande, o que sugere um modelo sem o termo longa-duração.

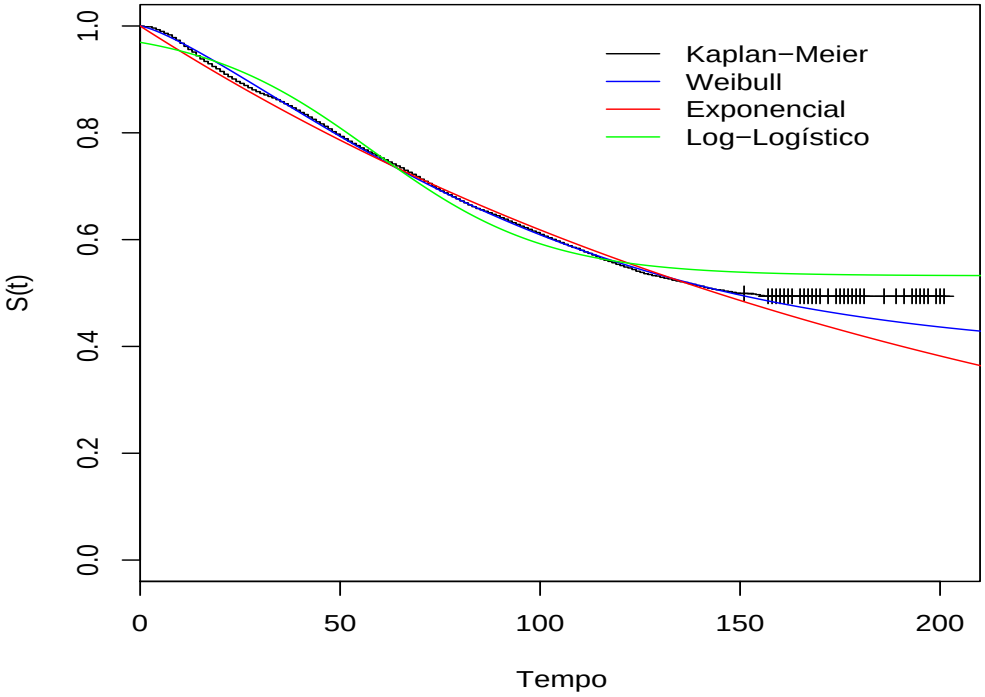


Figura 3.4: Curva estimada via Kaplan-Meier e curvas ajustadas através dos modelos Weibull, exponencial e log-logístico para os tempos.

Portanto, utilizando-se de técnicas gráficas, o melhor modelo ajustado para os dados, é o modelo Weibull. Entretanto, para a escolha do modelo mais apropriado, além do critério subjetivo (verificação gráfica), os métodos de seleção apresentados na Seção 3.3 serão empregados.

Os valores obtidos para os métodos de seleção de modelo, são apresentados na Tabela

3.4 que segue.

Tabela 3.4: Valores de AIC e BIC e norma Euclidiana (NE), para os modelos Weibull, exponencial e log-logístico.

Modelo	AIC	BIC	NE
Exponencial	300.934, 929	300.953, 109	0, 434
Weibull	300.076, 164	300.103, 436	0, 224
Log-Logístico	307.777, 400	307.750, 100	5, 084

Segundo os critérios AIC e BIC, o modelo que mais se adequa aos dados é o modelo Weibull, uma vez que, para este modelo, os valores dos critérios são menores do que para o modelo exponencial e log-logístico. Também, usando a distância entre a curva empírica (Kaplan-Meier) e as curvas ajustadas pelos modelos Weibull, log-logístico e exponencial, a menor distância apresentada é para o modelo Weibull, mostrando que este é mais adequado para os dados.

Considerando apenas os ajustes Weibull e exponencial, uma alternativa para verificar ajuste dos modelos é a formulação de testes de hipóteses usando-se da razão de verossimilhanças. Neste caso, o modelo exponencial é um caso particular do modelo Weibull (quando o parâmetro de forma $\beta = 1$), podendo assim, formular as seguintes hipóteses:

$$\begin{cases} H_0 : \beta = 1 \\ H_1 : \beta \neq 1, \end{cases}$$

e testá-las sob a suposição de que H_0 é verdadeira, pelo seguinte teste

$$w_0 = -2\log\left(\frac{L_1}{L_0}\right) \sim \chi_1 \tag{3.25}$$

ou seja, o teste de razão de verossimilhanças tem uma aproximação qui-quadrado com graus de liberdade igual ao número de parâmetros a ser estimado sob a hipótese H_1 menos o número de parâmetros a ser estimado sob H_0, sendo assim, $3 - 2 = 1$ grau de liberdade.

Desta forma, sob a hipótese nula $-2\log(L_0) = 300.930, 929$ e, sob a hipótese alternativa $-2\log(L_1) = 300.070, 169$. Assim, tem-se $w_0 = 860, 765$. Para a distribuição qui-quadrado com um grau de liberdade, obtém-se $p - valor < 0, 0001$. Sendo assim, existe evidência para rejeitar H_0 ao nível de 5% de significância, ou seja, existe evidência de que o modelo a ser ajustados é o modelo Weibull.

Considerando agora o modelo log-logístico, para construir um teste de hipótese com o intuito de comparar os modelos seria trabalhoso uma vez que os modelos agora não

são encaixados (como no caso do modelo Weibull com relação ao exponencial). Ou seja, neste caso, a medida de distância entre curvas seria uma opção bastante razoável para selecionar o modelo mais apropriado.

Computacionalmente, medir as distâncias é muitos mais simples do que construir um teste de hipóteses. Também, nem sempre é possível utilizar-se de teste de hipóteses simples, como é o caso do teste de razão de verossimilhanças, uma vez que, nem sempre os modelos a serem testados são encaixados. Uma questão importante é sobre o funcionamento do método de medir distância entre curvas (norma Euclidiana), quando trabalhamos com amostras menores.

Assim, com o intuito de responder a esta questão, adota-se a seguinte estratégia de investigação com o intuito de validar os procedimentos aplicados anteriormente.

Usando os dados apresentados na aplicação desta seção, uma amostra aleatória foi retirada proporcional a amostra original, (selecionou-se uma amostra de 1% do tamanho da amostra original, ponderada pelos tempos observados e tempos censurados). Assim, totalizou-se 418 tempos censurados e 217 tempos exatamente observados. Usando-se desta amostra, é possível verificar, se para quantidades menores de observações, os procedimentos descritos são válidos.

Para esta amostra, o modelo Weibull foi ajustado considerando-se a reparametrização feita anteriormente e, as estimativas dos parâmetros são apresentadas na Tabela 3.5.

Tabela 3.5: Estimativas dos parâmetros do modelo Weibull.

Parâmetros	Estimativas	Erro Padrão
μ_0	$4,5044$	$0,1992$
β_0	$0,3094$	$0,0931$
γ_0	$0,2614$	$0,3508$

Para a mesma amostra usada anteriormente, estimou-se os parâmetros usando-se do modelo exponencial, também usando a reparametrização anterior, obtendo assim as estimativas apresentadas na Tabela 3.6 e, para o modelo log-logístico cujas estimativas são apresentadas na Tabela 3.7.

Assim, para estes valores estimados para as distribuições Weibull, log-logística e exponencial, as seguintes curvas foram plotadas e são apresentadas na Figura 3.5.

Tabela 3.6: Estimativas dos parâmetros do modelo exponencial.

Parâmetros	Estimativas	Erro Padrão
μ_0	$5,3592$	$0,0776$
γ_0	$12,4489$	$410,0525$

Tabela 3.7: Estimativas dos parâmetros do modelo Log-Logístico.

Parâmetros	Estimativas	Erro Padrão
μ_0	$4,0384$	$0,0616$
σ_0	$3,0376$	$0,0802$
γ_0	$-0,1930$	$0,1254$

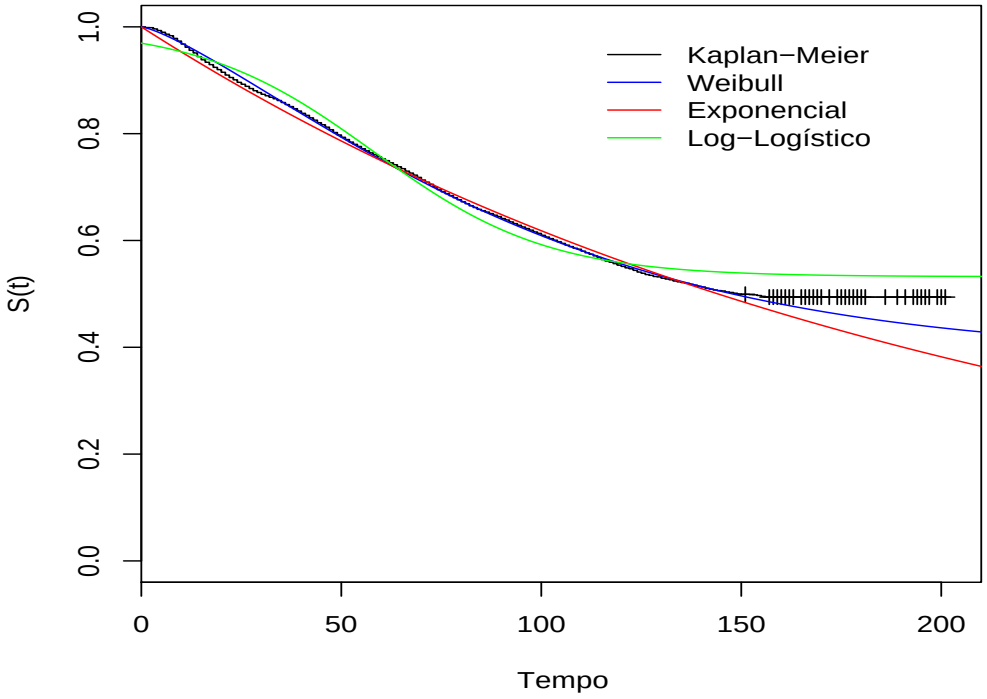


Figura 3.5: Curvas ajustadas para a amostra dos tempos até os pagamentos de empréstimos.

Ainda, como critério de escolha do modelo, os valores de AIC, BIC e norma Euclidiana foram calculados e estão dispostos na Tabela 3.8.

Através dos valores apresentados para AIC e BIC, tem-se que o modelo mais apropriado, segundo estes critérios, é o modelo Weibull o qual apresenta menores valores para

Tabela 3.8: Valores de AIC e BIC e norma Euclidiana para os modelos Weibull e exponencial.

Modelo	AIC	BIC	NE
Exponencial	2.115, 2440	2.124, 1513	0, 4007
Weibull	2.109, 1034	2.122, 4643	0, 1546
Log-logístico	2.146, 429	2.119, 158	3, 8597

estas métricas. Usando-se das distâncias entre a curva empírica e a curva estimada via os modelos Weibull, log-logístico e exponencial, verifica-se que a menor distância é dada pelo modelo Weibull, concordando com os outros critérios de seleção.

Também, realizou-se o mesmo teste de hipótese considerando o modelo Weibull e o modelo exponencial por serem modelos encaixados e, agora, sob a hipótese nula $-2\log(L_0) = 2.111, 2440$ e, sob a hipótese alternativa $-2\log(L_1) = 2.103, 1034$. Assim, tem-se $w_0 = 8.1406$. Usando-se da distribuição qui-quadrado, obtém-se $p - valor = 0, 0043$. Sendo assim, rejeita-se H_0 ao nível de 5% de significância, ou seja, existem evidências de que o modelo a ser ajustado seja o modelo Weibull.

Desta forma, verifica-se que através da análise gráfica, da distância entre curvas e dos critérios de seleção de modelo AIC e BIC, que o modelo Weibull se mostrou mais adequado. Tem-se as mesmas conclusões tanto para grandes amostras, como para amostras pequenas.

Capítulo 4

Avaliação do Procedimento de Estimação dos Parâmetros e os Critérios de Seleção de Modelos

Na Seção 3.4 do Capítulo 3, resultados adequados foram obtidos para uma amostra específica, entretando é necessário verificar o comportamento das estimativas e dos critérios de seleção de modelos para diferentes tamanhos de amostras e porcentagens de censura. Para isso desenvolvemos um estudo de simulação para verificar o desempenho das métricas para amostras geradas de duas distribuições (Weibull e log-logística), com diferentes valores para os parâmetros.

Simulação é um processo que emprega modelos (matemáticos ou estatísticos), com o objetivo de imitar um processo ou operação para descrever o comportamento de um sistema, estimar distribuição de variáveis aleatórias, testar hipóteses estatísticas, comparar diferentes cenários, avaliar comportamento de uma solução analítica etc, (Perin Filho, 1995).

Particularmente trabalhamos com simulação *Bootstrap*, com o intuito de verificar quais métricas são mais adequadas para avaliar o ajuste de modelos na presença de longa-duração nos dados. Também analisaremos para quais tamanhos de amostras e quantidades de censuras as métricas estudadas apresentam melhores resultados.

4.1 Objetivos

Quando os dados compreendem um amostra fidedigna da população e esta é regida por alguma distribuição de probabilidade, a família de estimadores correspondentes a todas

as possíveis amostras tem a mesma distribuição de probabilidade. A estatística clássica investiga estas distribuições dos estimadores, com o propósito de estabelecer propriedades básicas, tais como, confiabilidade e incerteza. Para reamostras, também existem técnicas para avaliar estas mesmas propriedades dos estimadores (Efron e Tibshirani, 1993).

Assim, para o procedimento de estimação, quase sempre é preciso seguir alguns pressupostos. Se um modelo, que pode ser pensado como um conjunto de pressupostos, está incorreto, espera-se que as estimativas para este também sejam incorretas. Sendo assim, um dos objetivos da investigação estatística é obter formas para verificar os pressupostos necessários para uma boa estimativa (robustes estatística, Huber, 1981). Estes conceitos são apresentados por Glymour *et al.* (1997).

A fim de fazer uma investigação sobre os procedimentos de estimação, uma questão importante a ser levantada é qual o modelo mais apropriado aos dados em estudo e, para responder a esta questão, propõem-se um estudo de simulação. Deve-se considerar que se pretendemos trabalhar com grandes bancos de dados pertencentes a empresas ligadas a área de finanças ou seguradoras e que estes, apresentam um alto percentual de censura.

Outro motivo para este estudo, é o fato de que, quando se trabalha com carteiras de clientes para a análise de crédito, financiamento ou dados de seguradoras, existem muitos clientes chamados fidelizados, ou seja, clientes que não estão expostos ao evento de interesse. Assim, ao longo do experimento, estes terão seus tempos de estudo censurados. Quando isto ocorre, existem tempos censurados à direita, o que justificaria a utilização de modelos apropriados, por exemplo, o uso de modelos com longa-duração.

Para comparar modelos e verificar o ajuste, algumas métricas usuais para análise de adequabilidade do modelo tais como, AIC e BIC são utilizadas; mas além dessas métricas usuais propõem-se a utilização da norma Euclidiana como forma alternativa para a análise da adequabilidade do modelo estimado.

A fim de cumprir com os objetivos, utilizaremos o procedimento de reamostragem *Bootstrap* para verificar quais das métricas é a mais adequada, para diferentes tamanhos de amostras e porcentagens de censura.

4.2 Processo de Reamostragem

A idéia de reamostragem surgiu em meados de 1935 e, Bradley Efron em 1979 introduziu a técnica de reamostragem *Bootstrap* como abordagem alternativa ao cálculo de intervalos de confiança, em circunstâncias em que outras técnicas não eram aplicáveis, em particular, no caso em que o tamanho da amostra era pequeno.

A técnica *Bootstrap* é em primeiro lugar, uma maneira de encontrar a distribuição amostral de uma estatística de interesse, pelo menos aproximadamente, a partir de uma amostra disponível. Uma distribuição amostral está baseada em muitas amostras aleatórias da população. Entretanto, ao invés de retirar-se muitas amostras da população, obtém-se reamostras, com reposição, a partir de uma única amostra. Cada reamostra tem o mesmo tamanho da amostra original.

Assim, este processo de reamostragem tenta realizar o que seria desejável na prática, isto é, repetir a experiência de amostragem B vezes. Além disso, o procedimento trata a amostra observada como se esta representasse exatamente toda a população. Sendo assim, para $t = (t_1, t_2, \ldots, t_n)$ uma amostra aleatória contendo n tempos de sobrevivência disponíveis para análise, com $\delta_i = 1$ para os tempos exatamente observados e, $\delta_i = 0$, para os tempos censurados a direita, $i = 1, 2, \ldots, n$. O processo de reamostragem *Bootstrap* consiste em reamostrar B amostras $T^{*(1)}, T^{*(2)}, \ldots, T^{*(B)}$, independentes e identicamente distribuídas, cada uma de tamanho n, onde T é o conjunto de dados disponíveis dado por $T = (t, \delta)$.

Após a obtenção das B amostras *Bootstrap*, pode-se obter, por exemplo, estimativas para os parâmetros de interesse para cada uma dessas amostras, obtendo-se assim, um vetor de parâmetros estimados dado por: $\hat{\theta}^* = (\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)})$. A partir do vetor $\hat{\theta}^*$, é possível fazer inferências a respeito do parâmetro θ, associado a amostra original.

Utilizando-se desses conceitos, neste estudo foram geradas amostras de duas distribuições: Weibull, apresentada na Seção 3.1.1 e log-logística, apresentada na Seção 3.1.2. As amostras foram geradas de tamanhos e com valores de parâmetros pré-estabelecidos. Dentro do contexto de reamostragem *Bootstrap*, as amostras geradas serão conhecidas como "amostras originais".

Tendo em mãos estas "amostras originais "geradas para específicos casos que serão detalhados, o processo de reamostragem *Bootstrap* é empregado e, $B = 1.000$ reamostras é retirada para cada caso em estudo estimando-se os parâmetros do modelo em questão pelo método de máxima verossimilhança (distribuição usada para gerar a amostra original) e também estima-se os parâmetros da outra distribuição que foi considerada. Por exemplo, se a amostra original for gerada da distribuição Weibull, estimamos os parâmetros do modelo Weibull e também do modelo log-logístico.

Desta forma, obtém-se 1.000 estimativas para cada um dos parâmetros da distribuição e uma média destas estimativas é obtida. Posteriormente, calcula-se os valores das métricas (AIC, BIC e norma Euclidiana), para cada uma dessas reamostras, bem como seus valores médios (Apêndice A).

A fim de comparar os resultados, considerou-se a razão destas métricas. Para a obtenção destas razões, foi considerado, para o caso em que amostra é gerada a partir da distribuição Weibull, as estimativas das métricas para este modelo foram divididas pelas estimativas obtidas para o modelo log-logístico. O processo inverso também foi considerado quando fixou-se amostras geradas a partir do modelo log-logístico.

Quanto menor o valor obtido pela métrica, mais adequado é o modelo em questão. Espera-se que o valor da métrica, para o modelo a partir do qual gerou-se os dados, seja menor do que a do outro modelo comparado, desta forma, a razão obtida deverá ser menor do que 1. Quando isto não ocorre, valores maiores do que 1, significa que o critério de seleção (métrica) não foi capaz de identificar o modelo.

Para a realização deste estudo de simulação, alguns casos particulares foram considerados e estes, são apresentados na Seção 4.3.

4.3 Estudo de Simulação

A fim de cumprir com os objetivos propostos, considerou-se alguns casos particulares para as distribuição Weibull e log-logística.

Desta forma, para os dois casos, considerou-se quatro tamanhos de amostras: $n = 100$, 5.000, 15.000 e 30.000. Também para ambos os casos, utilizou-se quatro quantiles de censuras: 10%, 25%, 50% e 75%, sendo estas, censuras à direita, uma vez que tem-se o objetivo de estudar modelos com longa-duração.

Porém, para o caso Weibull, três valores de β (parâmetro de forma), foram usados para gerar a amostra original: $\beta = 0,5$; $1,0$ e $1,5$. Estes valores foram selecionados devido ao comportamento diferenciado da forma da função de risco do modelo. Para $\beta = 0,5$ tem-se a forma do risco decrescente, para $\beta = 1,0$ tem-se a forma do risco constante no tempo e, para $\beta = 1,5$ tem-se a forma do risco crescente com o tempo.

Para o modelo log-logístico, usou-se valores de σ para geração da amostra original com a forma da função de risco unimodal, sendo $\theta = 1,0$; $2,0$ e $3,0$, onde, $\theta = 1/\sigma$.

Assim, para o modelo Weibull e log-logístico, estudou-se 48 casos resultando um total de 96 estudos de simulação.

Para melhor explicar os casos estudados, tem-se a Figura 4.1.

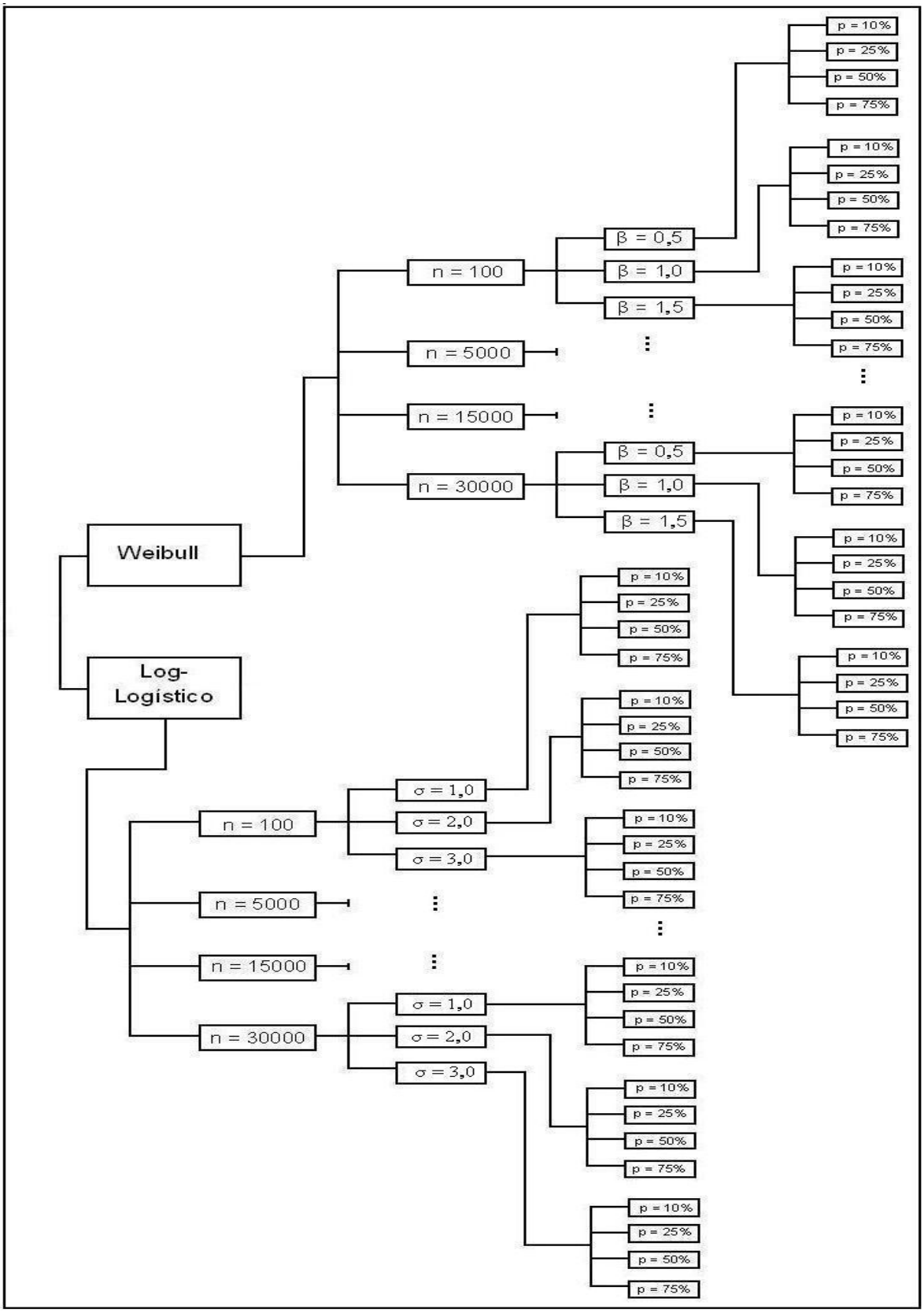


Figura 4.1: Casos estudados.

4.4 Resultados

Com o intuito de responder às questões levantadas nos objetivos, os casos apresentados na seção anterior foram estudados e os resultados foram resumidos em tabelas que estão dispostas no Apêndice A.

Nesta seção, são apresentados alguns box-plots, que resumem estes resultados. Para a construção dos mesmos, utilizando os conceitos apresentados anteriormente, foram geradas amostras das distribuições Weibull e log-logística.

Assim, como explicitado na seção anterior, tendo em mãos estas amostras originais geradas para específicos casos que foram detalhados, faz-se então o processo de reamostragem, gerando 1.000 reamostras para cada caso estudado. Assim, para estas reamostras, estimou-se os parâmetros para os dois modelos em questão, por exemplo, se a amostra original for gerada da distribuição Weibull, estimaremos os parâmetros do modelo Weibull e também do modelo log-logístico.

Desta forma, 1.000 estimativas dos parâmetros foram obtidas para cada distribuição considerando os casos estudados e, no final, uma razão das métricas foi calculada. Sendo assim, para o caso em que a amostra original é Weibull, fez-se a razão das métricas para as estimativas deste mesmo modelo com relação as métricas estimadas pelo modelo log-logístico e vice-versa.

Por exemplo, para a métrica AIC: se a razão de AIC's for maior que 1, significa que o valor de AIC para o modelo estimado a partir da amostra gerada pela mesma distribuição, é maior que o valor de AIC calculado para o outro modelo, o que implica na falha do critério de seleção do modelo.

Desta forma, obtem-se a Figura 4.2 que apresenta as razões de normas Euclidianas para os quatro níveis de censura estudados: 10%, 25%, 50% e 75%, com amostra gerada de uma distribuição Weibull com parâmetro de forma igual a $0,5$, ou seja, para risco decrescente.

Na Figura 4.2-(a), verifica-se que, para 10% de censura, foi possível identificar que o modelo mais apropriado é o modelo Weibull, uma vez que as razões são menores do que 1. Este resultado é o que se espera, uma vez que os dados foram gerados da mesma distribuição.

Nas Figuras 4.2-(b), (c) e (d), tem-se as mesmas conclusões, podendo afirmar assim que, a medida de distância entre curvas para a distribuição Weibull com risco decrescente, é uma boa métrica para selecionar o modelo; isto considerando tamanhos de amostra que variam entre 100 a 30.000 com os percentuais de censura estudados.

Pode-se verificar ainda, através da Figura 4.2 que quanto maior o tamanho da amostra gerada, maior a distância entre a curva estimada do modelo log-logístico com relação à curva empírica. Este resultado é muito importante pois, uma das motivações para a simulação feita é estudar o comportamento das métricas para seleção de modelos, em grandes bancos de dados.

Também, verifica-se que, quanto menor o tamanho da amostra, maior a dispersão das razões de normas Euclidianas observadas.

Uma outra comparação a ser feita é com relação ao percentual de censura que foi gerada a amostra. Tem-se que, para percentuais de censura baixos (neste caso 10% de censura), melhor a métrica identifica como sendo o mais apropriado o modelo de geração dos dados. Ou seja, para percentuais de censura baixos menores são as razões das distâncias calculadas.

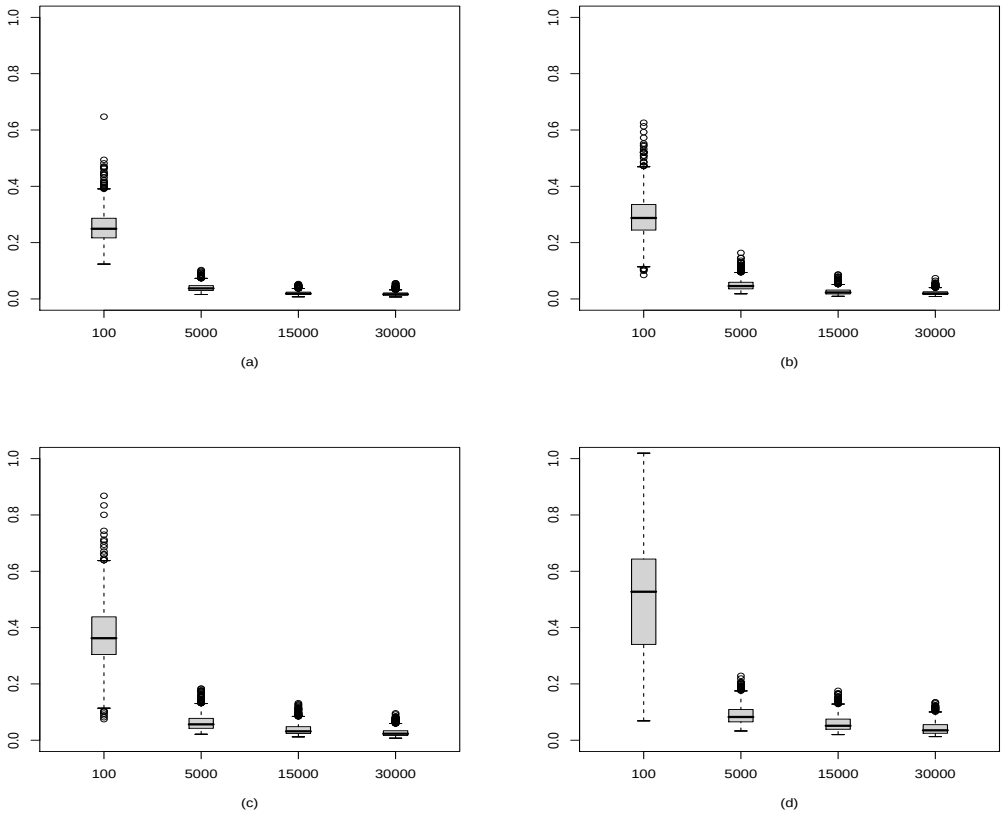


Figura 4.2: Box-Plot das razões de normas Euclidianas do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Ainda para amostras geradas de uma distribuição Weibull com risco decrescente, tem-se as Figuras 4.3 e 4.4 que representam, respectivamente, as razões de AIC's e BIC's.

As mesmas considerações feitas anteriormente, para as razões de normas Euclidianas, são válidas para as razões dos critérios AIC e BIC. Nas figuras a seguir, tem-se que, quanto maior o percentual de censura, mais difícil se torna identificar como apropriado o modelo através do qual se gerou a amostra e, quanto maior o tamanho de amostra, mais identificável se torna o modelo.

Verifica-se assim que as duas métricas são adequadas para seleção de modelos, pois as razões destes critérios (AIC e BIC), são menores do que 1, ou seja, os valores de AIC e BIC para o modelo Weibull são menores do que para o modelo log-logístico.

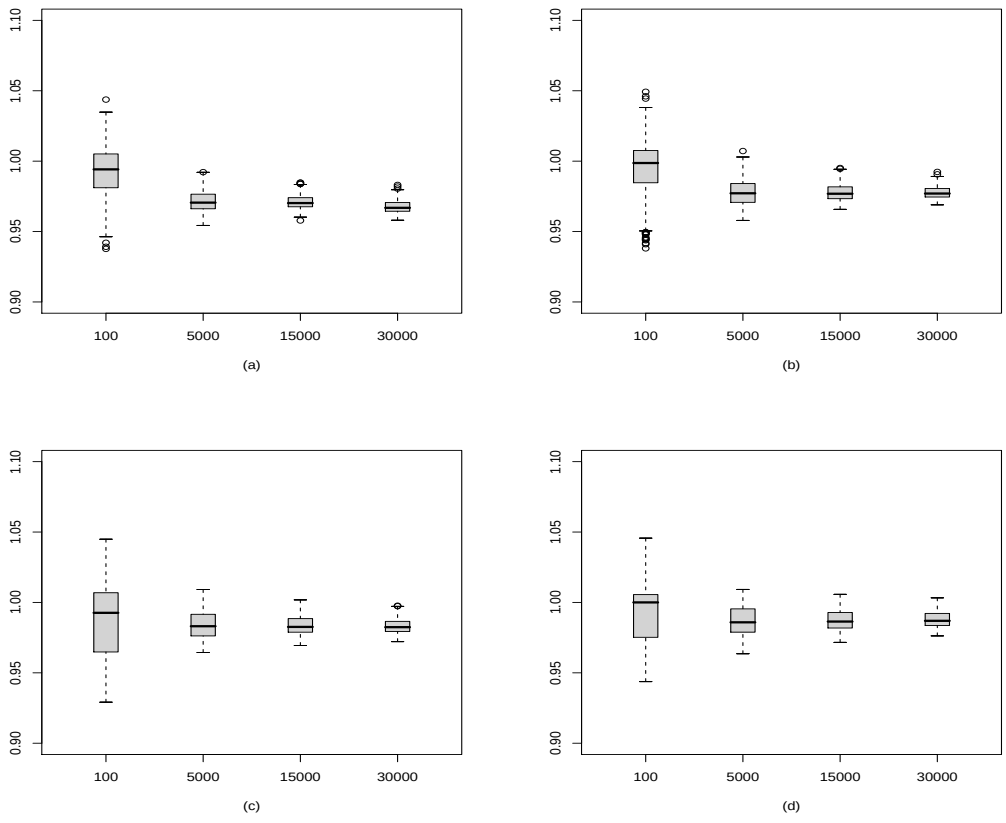


Figura 4.3: Box-Plot das razões de AIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Quando comparados por tamanhos de amostras, quanto maior a amostra mais adequadas as métricas para seleção.

De maneira geral, para amostras geradas de uma distribuição Weibull com parâmetro de forma menor que 1, ou seja, risco decrescente, as métricas estudadas (AIC, BIC e norma Euclidiana), são adequadas para seleção de modelos com termos de longa-duração, mesmo quando os percentuais de censura são altos.

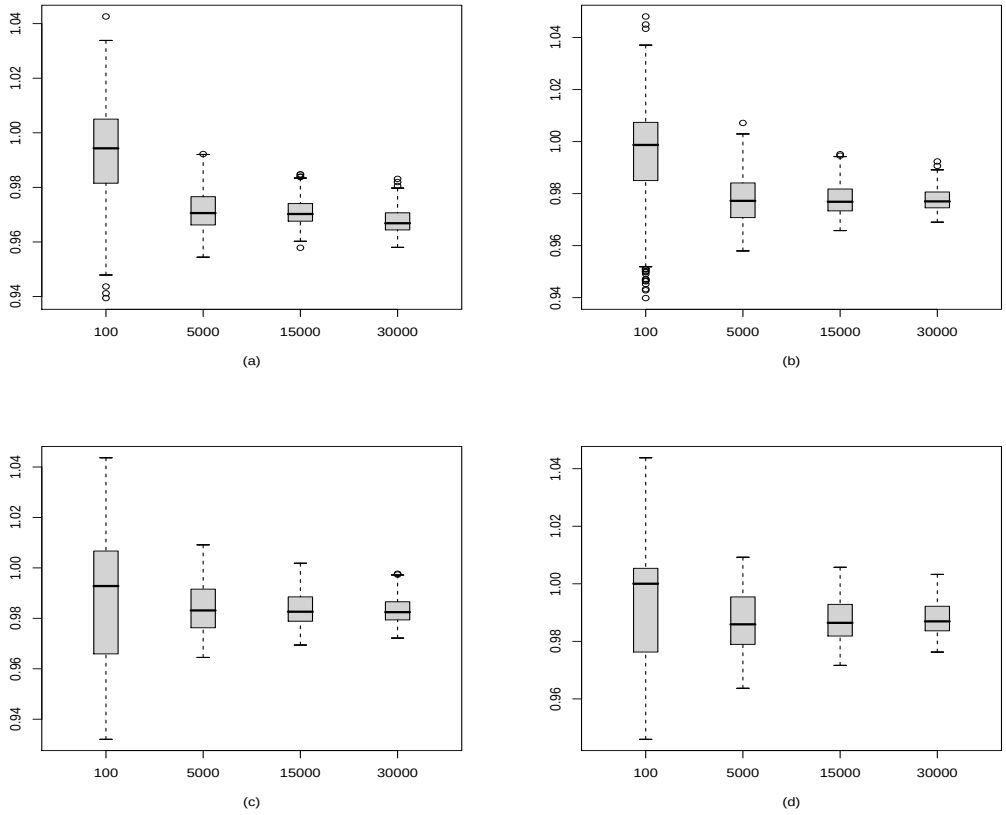


Figura 4.4: Box-Plot das razões de BIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

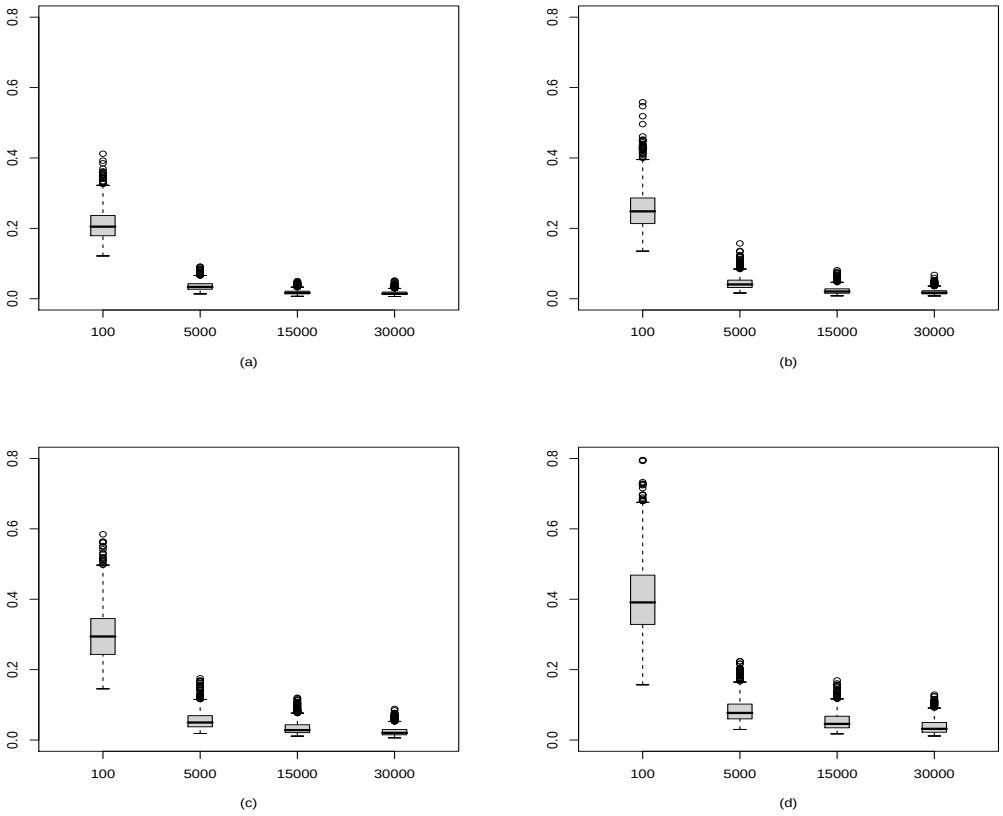


Figura 4.5: Box-Plot das razões de normas Euclidianas do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,0$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

As Figuras 4.5, 4.6 e 4.7 foram construídas para amostras geradas de uma distribuição Weibull com parâmetro de forma igual a 1, ou seja, risco constante (neste caso tem-se a distribuição exponencial, caso particular da distribuição Weibull).

Na Figura 4.5 apresenta-se as razões de normas Euclidianas. Pode-se verificar que, para tamanho de amostra $n = 100$, tem-se uma maior dispersão nos resultados. Esta dispersão diminui conforme aumenta o tamanho da amostra.

Também, para amostras maiores, os valores das razões de normas Euclidianas é menor, ou seja, quanto maior o tamanho da amostra, mais é possível identificar o modelo a partir do qual geramos os dados.

Confrontando a Figura 4.5-(a), amostras geradas com 10% de censura, com a Figura 4.5-(d), amostras geradas com 75% de censura, é possível verificar que, os box-plots da primeira figura estão mais próximos de zero do que para a segunda. Ou seja, quanto

menor o percentual de censura, mais se identifica o modelo utilizado para geração.

No caso da razão de normas, os valores são sempre menores do que 1, demonstrando assim que esta é uma boa métrica pra seleção de modelos, considerando todos os tamanhos de amostras e percentuais de censura estudados.

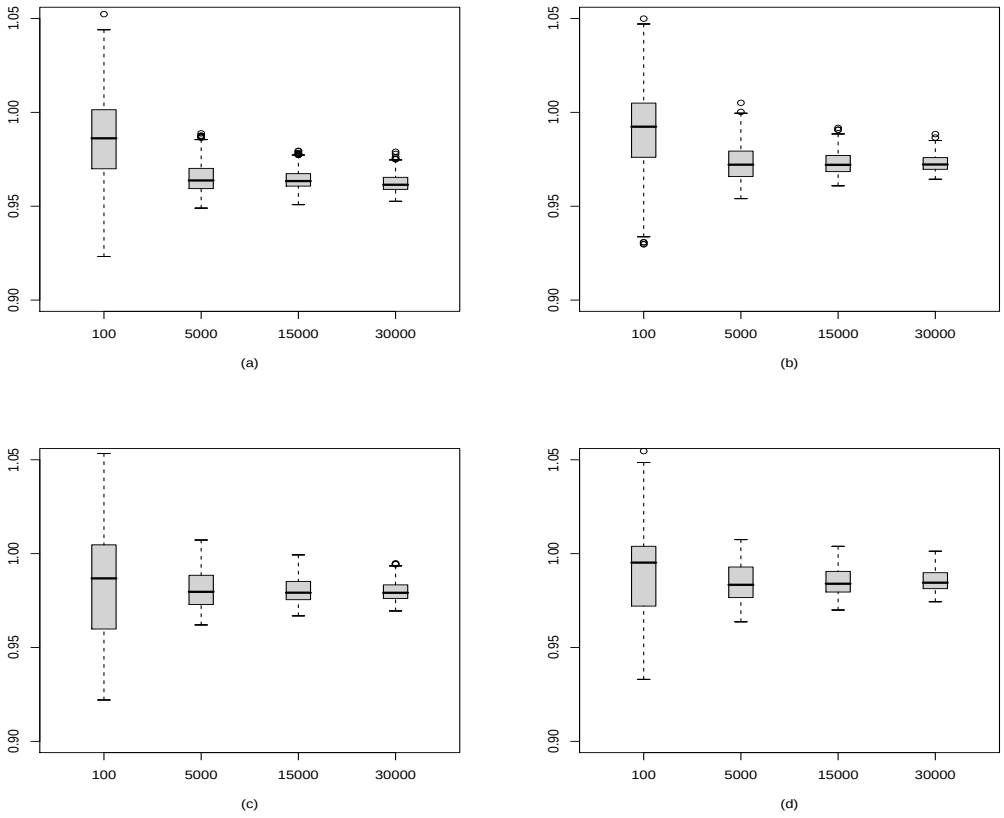


Figura 4.6: Box-Plot das razões de AIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,0$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Assim como para as razões de normas, quanto maior o tamanho de amostra, mais o critério AIC identifica o modelo Weibull como sendo o modelo apropriado aos dados (Figura 4.6).

Com relação aos percentuais de censura estudados, quanto menor o percentual de

censura, mais identificamos o modelo através do qual geramos os dados, não deixando de identificar este modelo para alto percentual de censura (neste caso 75%).

Também verifica-se que, para tamanhos menores de amostras existe uma maior dispersão, confirmando assim que, para amostras menores identificamos o modelo através do qual geramos os dados, porém, os valores de AIC para os dois modelos estão muito próximos. Esta é uma afirmação importante pois, tem-se grande interesse em estudar as métricas de seleção de modelos para amostras grandes uma vez que, em geral, as carteiras de clientes de empresas ligadas a área de finanças apresentam um grande número de observações.

As mesmas conclusões obtidas para o critério AIC são observadas para as razões de BIC's, Figura 4.7. Assim, de forma geral, quando as amostras são geradas de uma distribuição Weibull com risco constante, as métricas estudadas, não só AIC e BIC, mas a medida de distância entre curvas, são apropriadas para seleção de modelos na presença de tempos com termo de longa-duração.

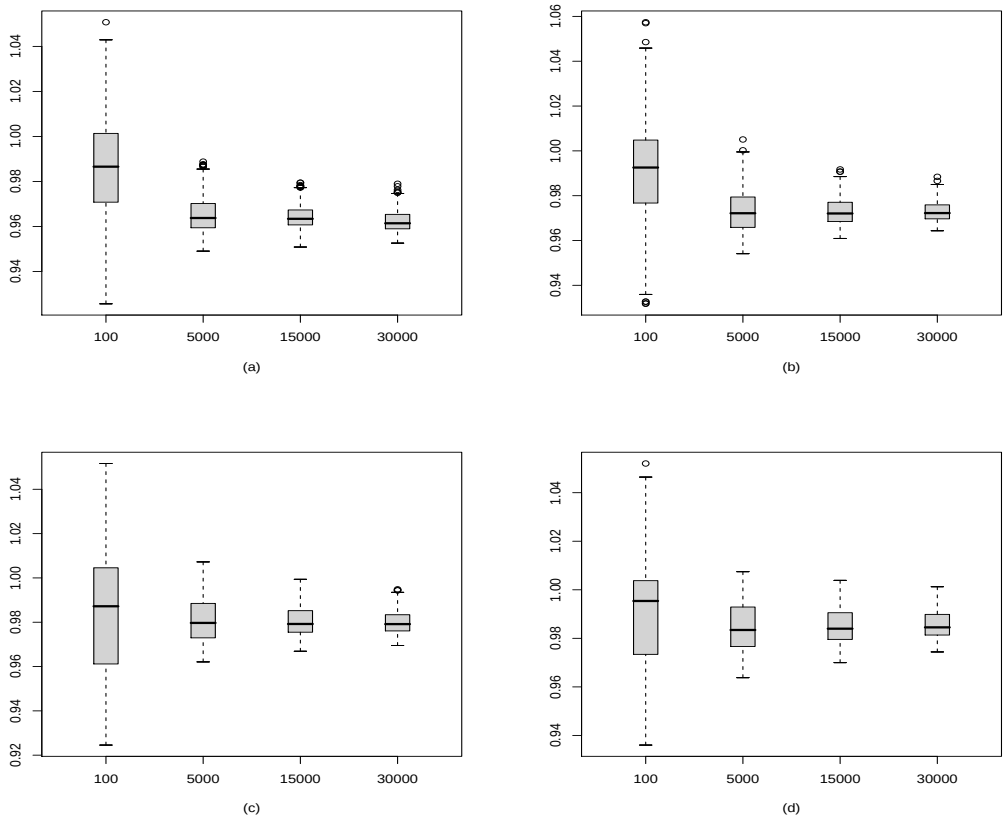


Figura 4.7: Box-Plot das razões de BIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,0$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

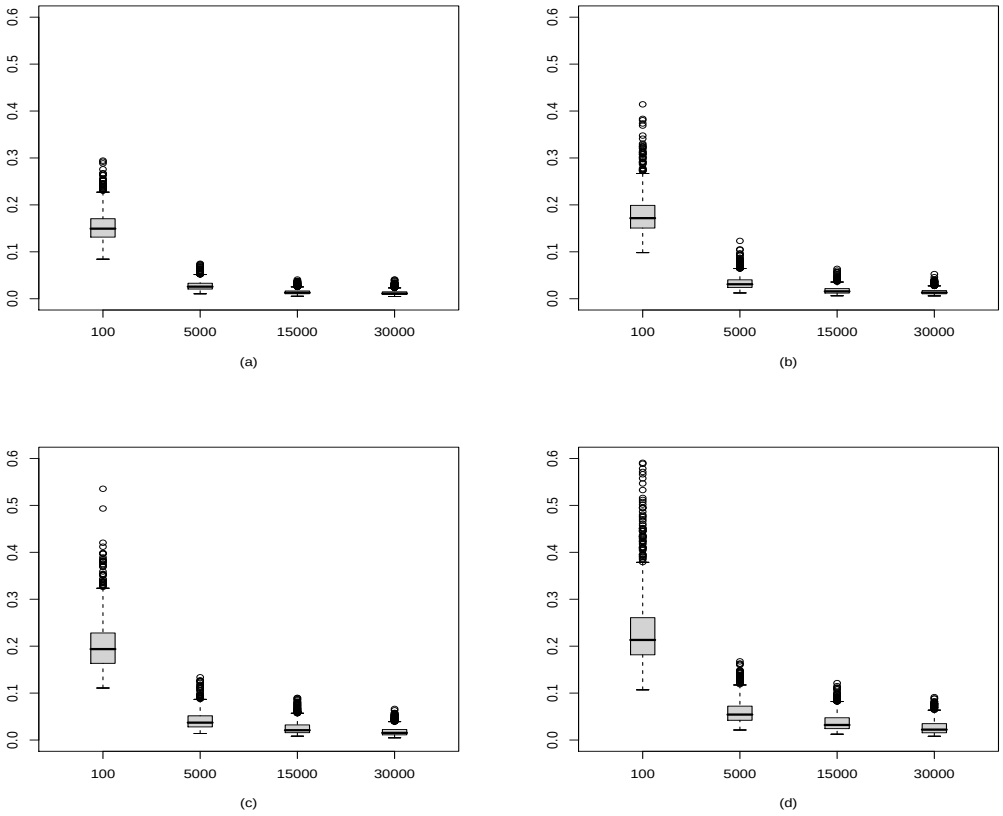


Figura 4.8: Box-Plot das razões de normas Euclidianas do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

A Figura 4.8 apresenta as razões de normas Euclidianas para os quatro níveis de censura estudados: 10%, 25%, 50% e 75%, com amostra gerada de uma distribuição Weibull com parâmetro de forma igual a $1,5$, ou seja, para risco crescente.

Na Figura 4.8-(a), verifica-se que, para 10% de censura, foi possível identificar que o modelo mais apropriado é o modelo Weibull, uma vez que as razões são menores do que 1. Nas Figuras 4.8-(b), (c) e (d), tem-se as mesmas conclusões, podendo afirmar assim que, a medida de distância entre curvas para a distribuição Weibull com risco crescente, é uma boa métrica para selecionar o modelo, isto considerando tamanhos de amostra que variam entre 100 a 30.000 com todos os percentuais de censura estudados.

É possível visualizar ainda, através da Figura 4.8 que, quanto maior o tamanho da amostra gerada, maior a distância entre a curva estimada do modelo log-logístico com relação à curva empírica. Também, verifica-se que, quanto menor o tamanho da amostra,

maior a dispersão das razões de normas Euclidianas observadas.

Uma outra comparação a ser feita é com relação ao percentual de censura que foi gerada a amostra. Tem-se que, para percentual de censura baixo (10% de censura), maior é a distância da curva do modelo que está sendo comparado (log-logístico) com relação à curva empírica, do que para o modelo através do qual geramos os dados (modelo Weibull). Ou seja, para o percentual de censura baixo, menores são as razões das distâncias calculadas.

Ainda para amostras geradas de uma distribuição Weibull com risco crescente, tem-se as Figuras 4.9 e 4.10 que representam, respectivamente, as razões de AIC's e BIC's.

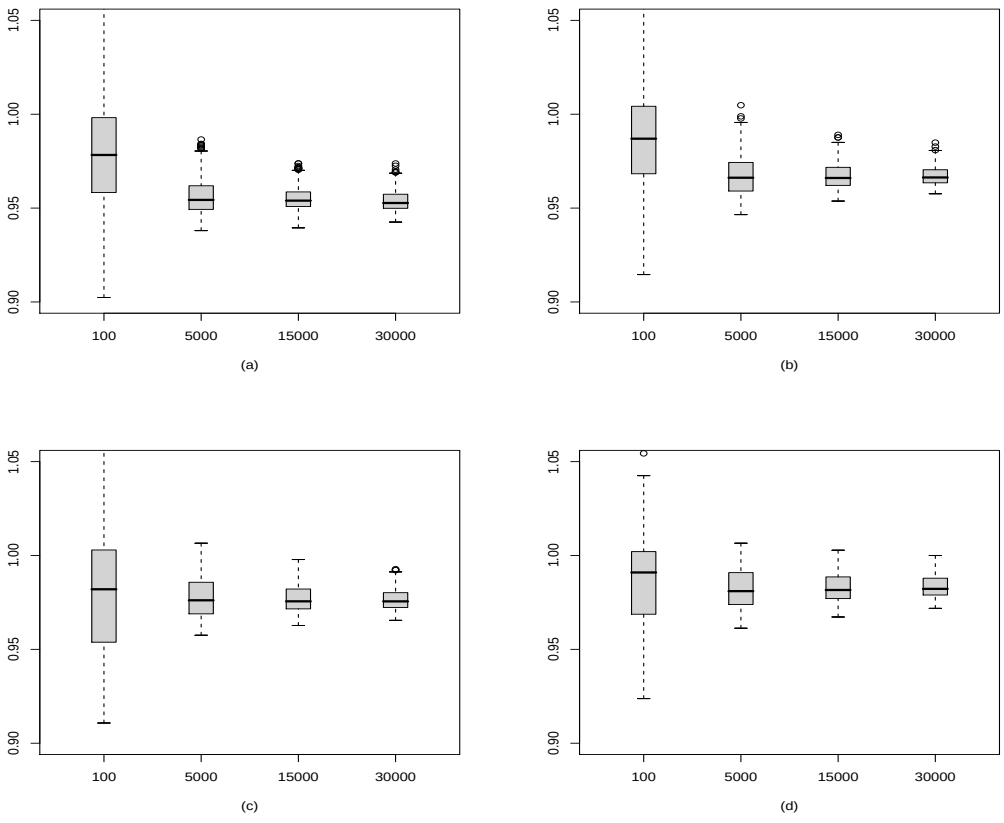


Figura 4.9: Box-Plot das razões de AIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

As mesmas considerações feitas anteriormente, para as razões de normas Euclidianas, são válidas para as razões dos critérios AIC e BIC. Nas figuras a seguir, tem-se que, quanto maior o percentual de censura, mais próximos os valores de AIC calculados para os dois modelos, porém, continua sendo menor o valor deste critério para o modelo Weibull. Quanto maior o tamanho de amostra, mais distante os valores destes critérios e menor a razão desta métrica.

Verifica-se assim que as duas métricas são adequadas para seleção de modelos, pois as razões destes critérios (AIC e BIC), são menores do que 1, ou seja, os valores de AIC e BIC para o modelo Weibull são menores do que para o modelo log-logístico.

De maneira geral, para amostras geradas de uma distribuição Weibull com parâmetro de forma maior que 1, ou seja, risco crescente, as métricas estudadas (AIC, BIC e norma Euclidiana), são adequadas para seleção de modelos com tempos de longa-duração, mesmo quando os percentuais de censura são altos.

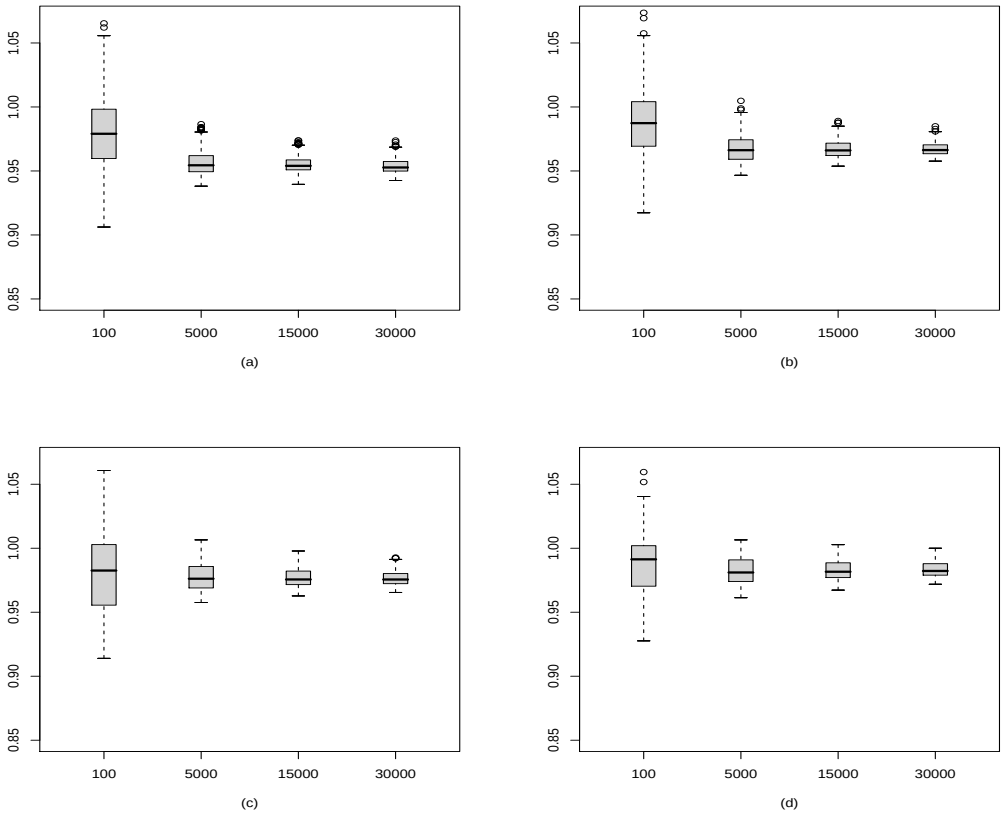


Figura 4.10: Box-Plot das razões de BIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 1,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Para o caso de amostras provenientes da distribuição Weibull, considerando-se todas as formas de sua função de risco e variados tamanhos de amostras (amostras pequenas ou grandes), as métricas aqui estudadas, conseguem identificar qual o modelo mais apropriado (nestes casos, pode-se afirmar que o modelo Weibull é o mais apropriado, uma vez que as amostras foram geradas desta distribuição).

Também, se confrontarmos as Figuras 4.2, 4.5 e 4.8, verifica-se que, para amostras geradas da distribuição Weibull com risco decrescente, primeira figura, as razões de normas Euclidianas são maiores do que para os outros casos. Isto é explicado uma vez que, a distribuição log-logística pode apresentar função de risco descrescente, diferentemente do caso de risco constante ou crescente.

Para amostras geradas da distribuição Weibull conclui-se que, tanto as métricas usuais para seleção de modelos (AIC e BIC), quando a métrica proposta neste trabalho (norma Euclidiana), são boas opções a serem usadas em seleção de modelos com tempos de longa-duração, o que em geral ocorre em carteiras de bancos, seguradoras, etc, quando se tem presente clientes fidelizados. Estas conclusões são válidas para grandes amostras ou pequenas carteiras.

Além deste primeiro estudo de simulação, outro estudo foi feito para o caso onde se tem amostras geradas da distribuição log-logística, confrontando assim, com o modelo Weibull.

A Figura 4.11 apresenta as razões de normas Euclidianas para amostras geradas da distribuição log-logística com parâmetro $1/\sigma = 2$. Nesta figura pode-se verificar que, para todos os percentuais de censura estudados, a norma Euclidiana consegue identificar qual o modelo mais apropriado aos dados. Verificou-se nesta, que as razões são sempre menores do que 1, como era esperado.

Porém, para amostras geradas com percentuais de censura iguais a 10%, os valores obtidos para as razões de normas Euclidianas foram maiores do que para percentuais de censura iguais a 75%. Mas, amostras geradas com percentuais de censura iguais a 25% e 50%, Figura 4.11-(b) e (c) respectivamente, os valores das razões foram maiores do que para amostras geradas com 10% de censura, Figura 4.11-(a).

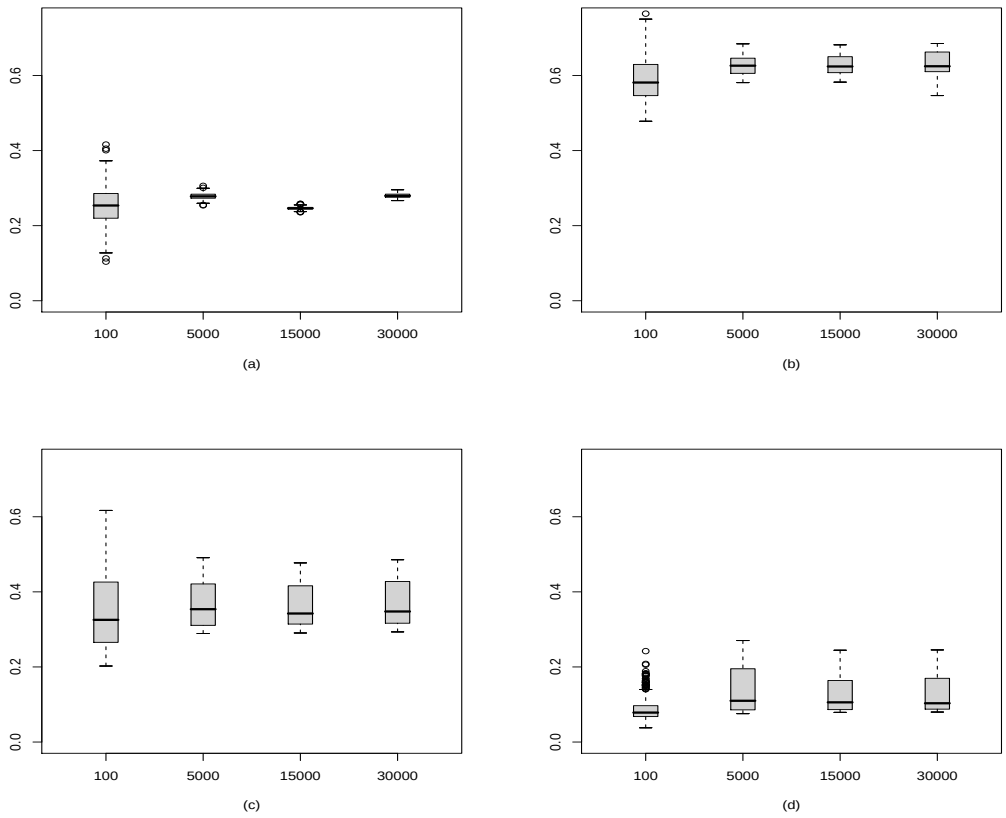


Figura 4.11: Box-Plot das razões de normas Euclidianas do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Agora, se compararmos com relação ao tamanho da amostra, tem-se que, para amostras menores, existe uma dispersão maior nas razões calculadas. Uma excessão observada é com relação ao percentual de censura igual a 75% onde, para amostra igual a 100, a dispersão das razões de normas é menor do que para amostras de tamanhos iguais a 5.000, 15.000 e 30.000, como mostra a Figura 4.11-(d).

Todavia, para o caso de amostras geradas da distribuição log-logística com parâmetro $1/\sigma = 2$, todas as razões de normas Euclidianas são menores do que 1, indicando que a distância entre a curva empírica com relação a curva do log-logístico estimado é menor do que para o modelo Weibull. Ou seja, em todos os casos para este valor de parâmetro, o modelo log-logístico foi identificado como o mais apropriado aos dados.

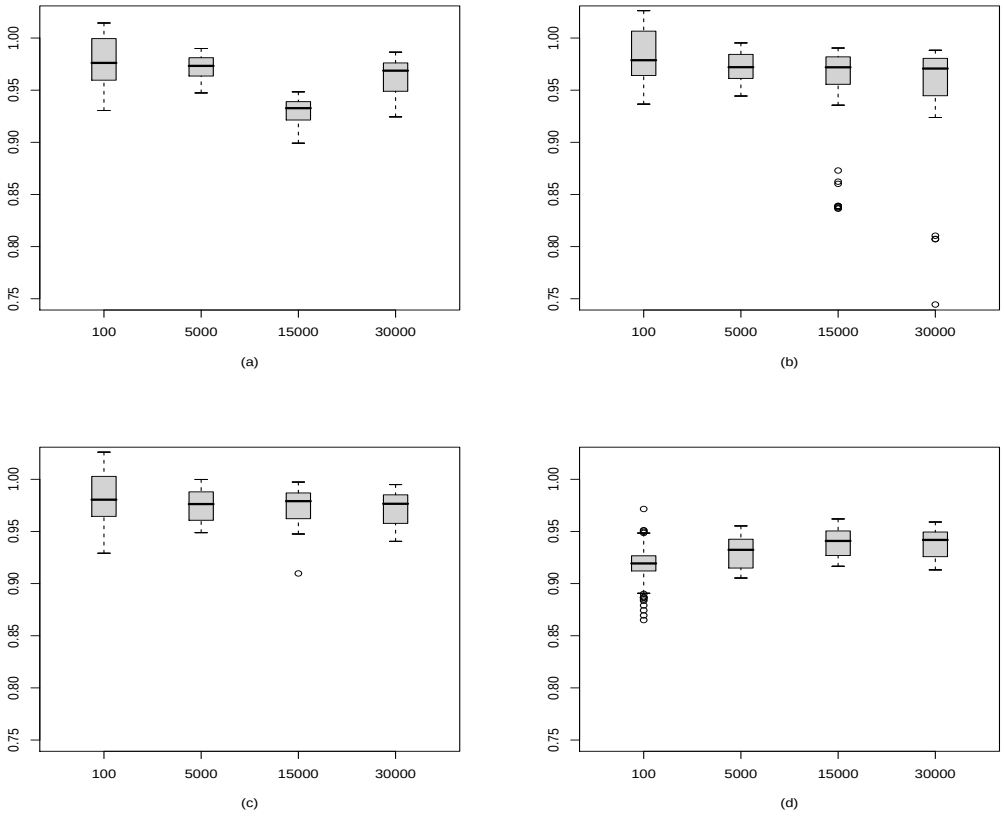


Figura 4.12: Box-Plot das razões de AIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Na Figura 4.12, apresenta-se as razões de AIC's para amostras geradas da distribuição log-logística também com parâmetro $1/\sigma = 2$. Nesta figura, verifica-se que, conforme aumenta o percentual de censura, mais identificável se torna o modelo.

E, quando comparamos com relação ao tamanho de amostra gerada, quanto menor a amostra, maior dispersão nas razões de AIC's. Conforme aumenta o tamanho de amostra, a dispersão das razões diminuem gradativamente.

Com relação a conseguir identificar como mais apropriado o modelo a partir do qual gerou-se a amostra, se considerarmos as medianas das razões calculadas, os valores de AIC's para o modelo log-logístico são sempre menores do que para o modelo Weibull.

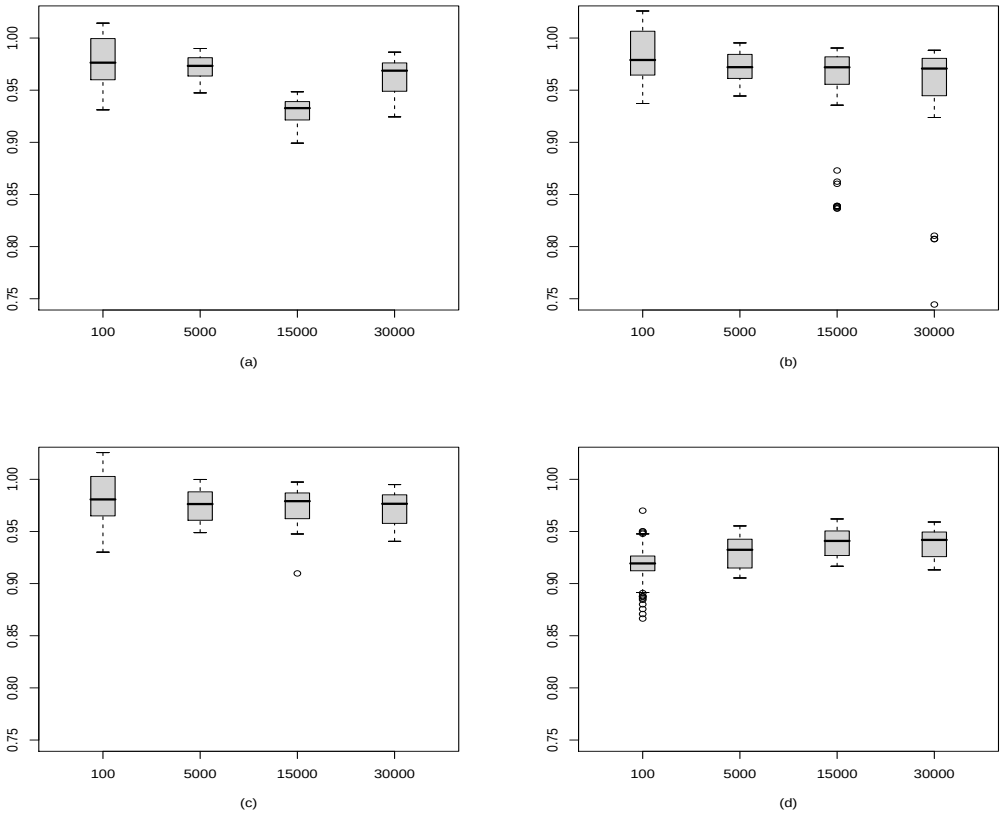


Figura 4.13: Box-Plot das razões de BIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Na Figura 4.13, apresenta-se as razões de BIC's para amostras geradas da distribuição log-logística também com parâmetro $1/\sigma = 2$. Também, assim como para o critério AIC, conforme aumenta o percentual de censura, mais identificável se torna o modelo.

Se compararmos com relação ao tamanho de amostra gerada, quanto menor a amostra, maior dispersão nas razões de BIC's. Conforme aumenta o tamanho de amostra, a dispersão das razões diminuem gradativamente.

Através desta figura, se considerarmos as medianas das razões, identificamos o modelo que gerou os dados como sendo o mais apropriado, igualmente, para qualquer tamanho de amostra, demonstrando assim, de forma geral, que o critério BIC é um bom critério para seleção de modelos com longa-duração, uma vez que todas as medianas das razões desta métrica são menores do que 1.

Para amostras geradas de uma distribuição log-logística com parâmetro $1/\sigma = 3$, tem-se as Figuras 4.14, 4.15 e 4.16 que apresentam, respectivamente, as razões de normas Euclidianas, AIC's e BIC's.

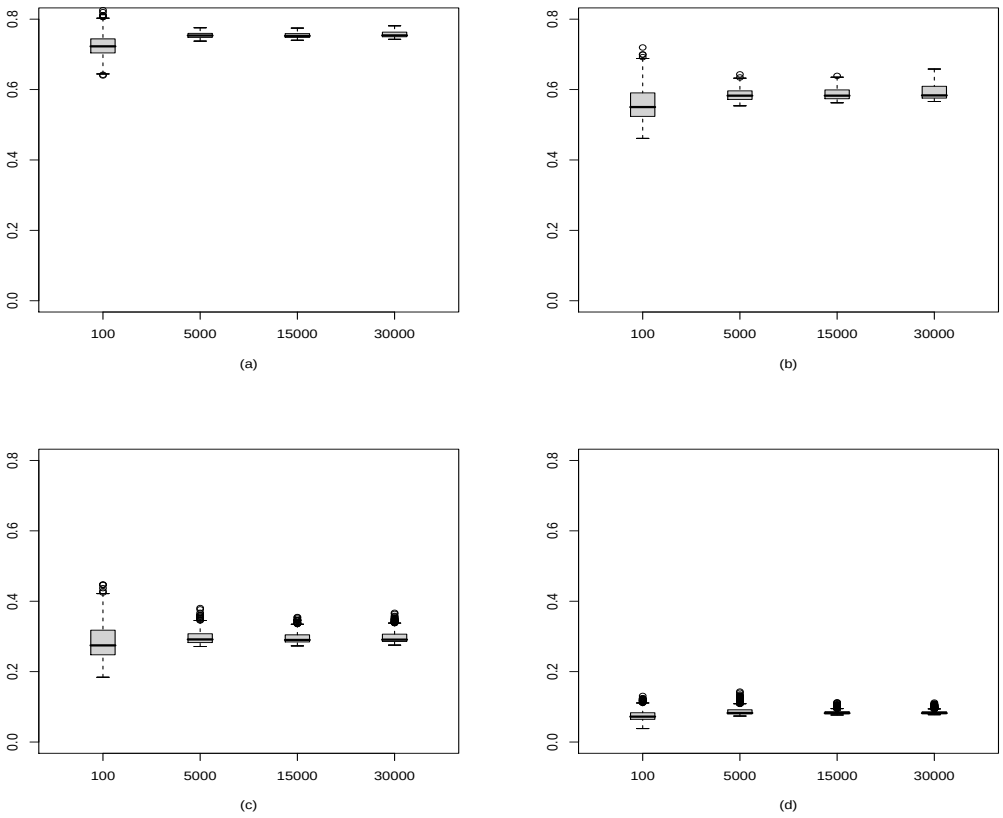


Figura 4.14: Box-Plot das razões de normas Euclidianas do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 3$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Na Figura 4.14 é possível visualizar que, para todos os percentuais de censura estudados, a norma Euclidiana consegue identificar qual o modelo mais apropriado aos dados. Verificou-se nesta, que as razões são sempre menores do que 1, como se espera.

Com relação ao percentual de censura, quanto maior o número de censurados, mais a norma Euclidiana consegue identificar o modelo. Ou seja, quanto maior o número de censura, mais distante a curva Weibull está da curva empírica se compararmos com a distância da curva estimada log-logística com relação à empírica. Na Figura 4.14-(a), que apresenta as razões para amostras geradas com 10% de censura, tem-se que estas razões de normas Euclidianas estão próxima de $0,8$. Quando analisamos a Figura 4.14-(d), para amostras geradas com percentuais de censura iguais a 75%, as razões estão próximas de $0,1$.

Agora, se analisarmos com relação ao tamanho de amostra, quanto maior a amostra, menor a dispersão dos dados. Porém, conforme aumenta o tamanho da amostra, a mediana das razões se mantém, indicando assim que, não importa o tamanho de amostra trabalhado, a identificabilidade do modelo é a mesma.

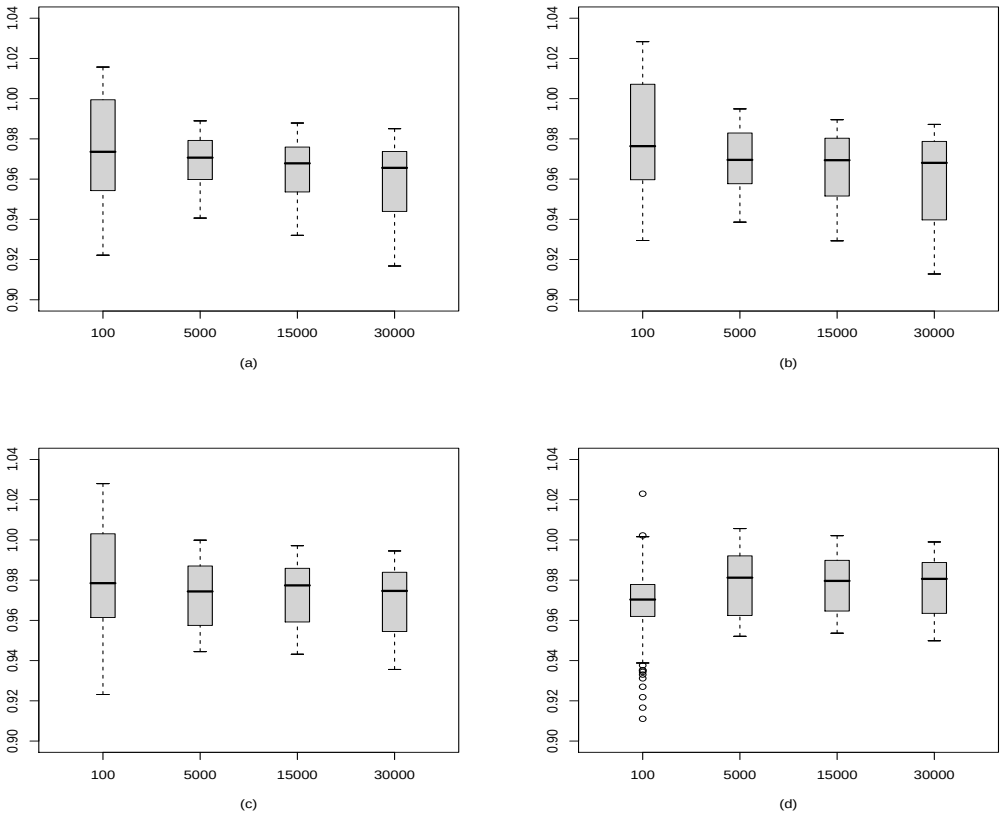


Figura 4.15: Box-Plot das razões de AIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 3$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Para os critérios AIC e BIC, estes também conseguem identificar o modelo log-logístico como sendo o mais apropriado aos dados, não importa o tamanho de amostra, nem o percentual de censura observado.

Para amostra pequena, $n = 100$, as razões destes critérios têm uma maior dispersão, diminuindo conforme aumenta o tamanho de amostra. Porém, as medianas destas razões não se alteram de forma significativa, nem quando aumentamos o tamanho de amostra, nem para os diferentes percentuais de censura.

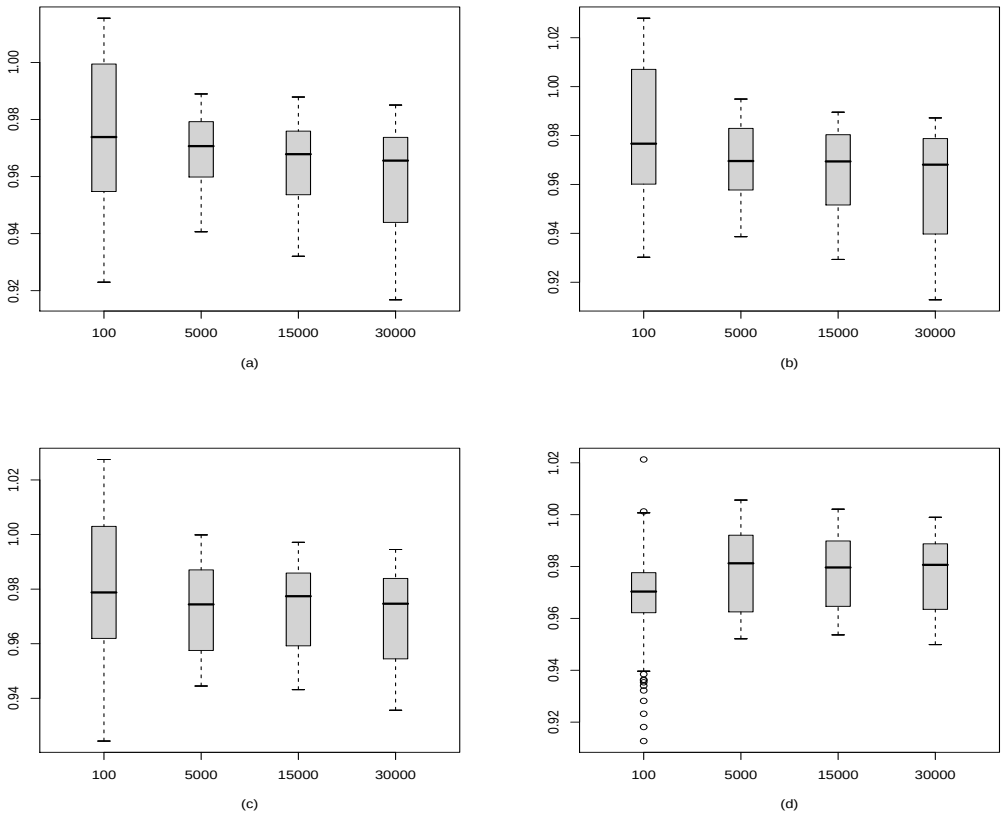


Figura 4.16: Box-Plot das razões de BIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 3$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Uma última análise feita, foi para amostras geradas de uma distribuição log-logística com parâmetro $1/\sigma = 4$. Para estas amostras foram plotadas as Figuras 4.17, 4.18 e 4.19 que apresentam, respectivamente, as razões de normas Euclidianas, AIC's e BIC's.

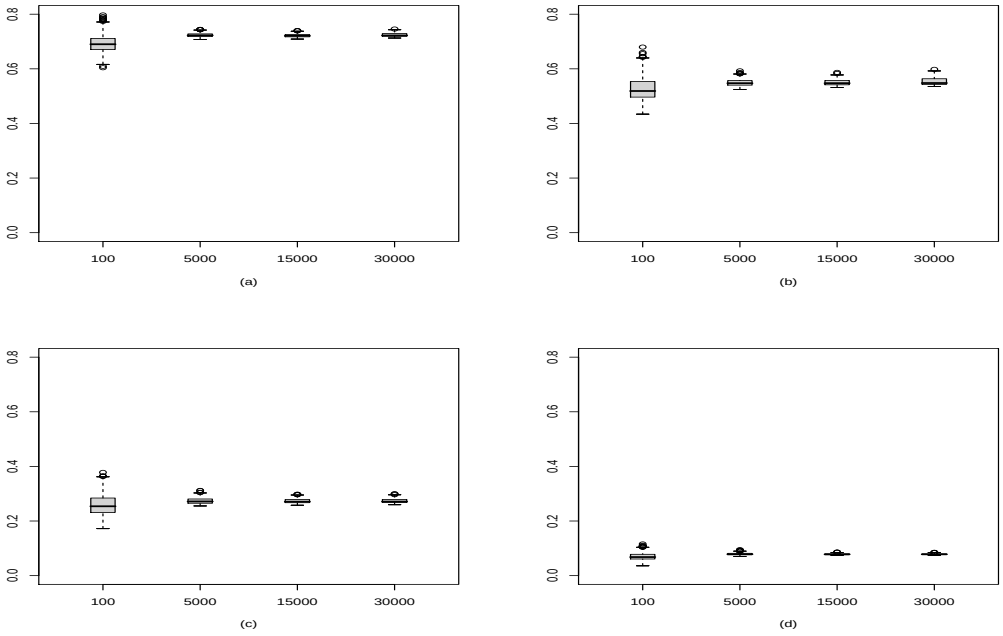


Figura 4.17: Box-Plot das razões de normas Euclidianas do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 4$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

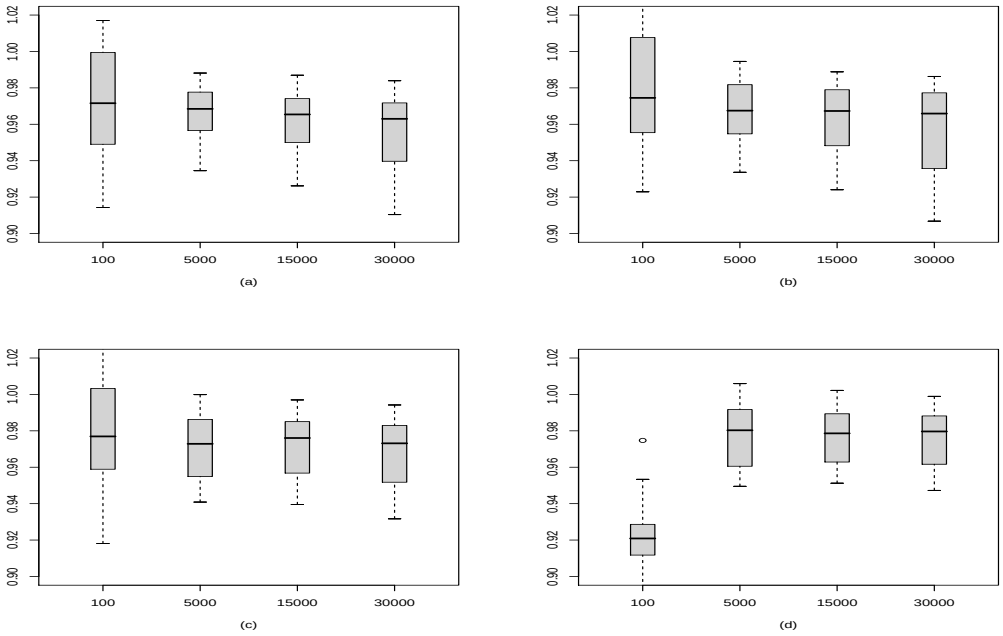


Figura 4.18: Box-Plot das razões de AIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 4$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Os mesmos resultados obtidos para os valores de parâmetros $1/\sigma = 2$ e 3 são verificados aqui. Diferente das tendências dos outros casos, neste caso em particular e para razões de BIC's com amostras geradas com percentual de censura igual a 75%, quanto menor o tamanho da amostra mais idenficável e torna o modelo que gerou os dados.

Nas Figuras 4.18-(d) e 4.19-(d), para amostras de tamanho 100, as razões de métricas são menores, ou seja, o modelo é mais identificável para amostras de tamanho 100 do que para amostras de tamanho 30.000.

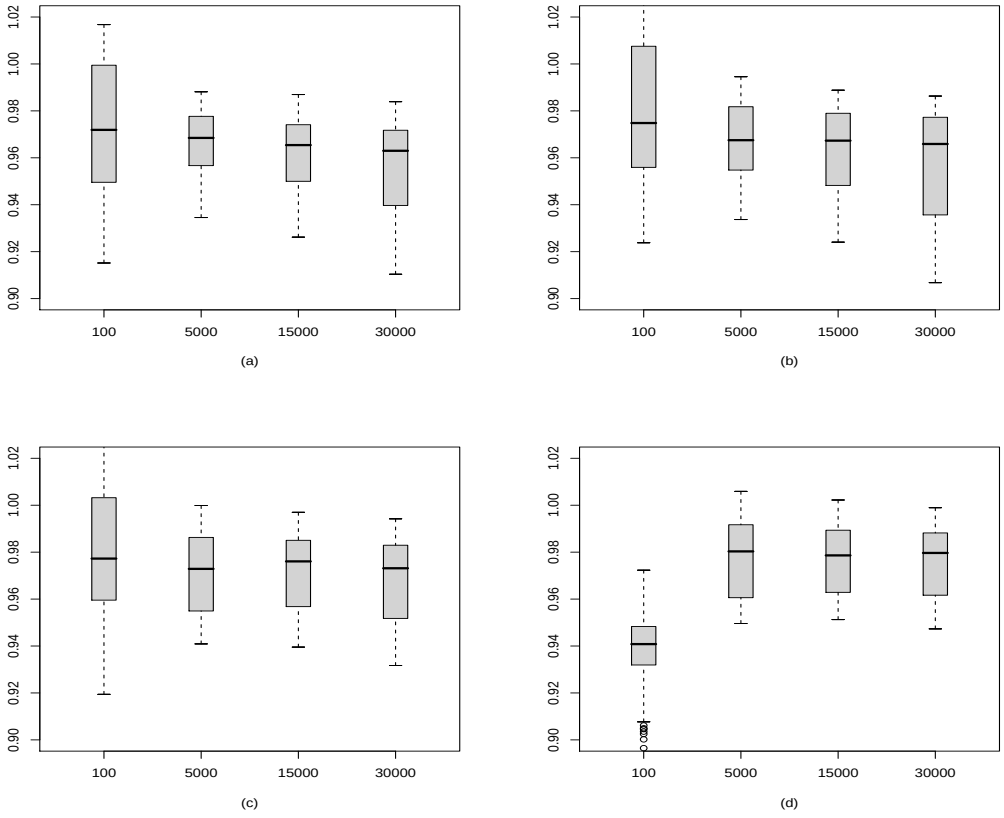


Figura 4.19: Box-Plot das razões de BIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 4$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

De forma geral, tanto para amostras geradas da distribuição Weibull, quanto para

amostras geradas da distribuição log-logística, os critérios estudados demonstraram-se satisfatórios.

Como temos o interesse em trabalhar os conceitos apresentados em carteiras de clientes de instituições financeiras, temos de enfatizar que, o estudo de simulação foi realizado na presença de longa-duração. Não só deve-se considerar a presença de longa-duração nos tempos, mas também, que esta longa-duração foi aqui caracterizada com alto percentual de censura, chegando a 75% de dados parcialmente observados. Esses percentuais de censura altos, em geral, é o que na prática se observa em carteiras de clientes.

Outra particularidade é que, neste estudo de simulação, trabalhou-se com grandes bancos de dados, considerando-se amostras de até 30.000 tempos, o que também se verifica na prática.

Durante a simulação, além da distribuição através da qual geramos os dados, ajustamos para as amostras um outro modelo. Um problema encontrado foi o fato de, mesmo sendo inapropriada determinada distribuição para os dados, no caso de grandes amostras, existe a convergência para as estimativas dos parâmetros.

Neste estudo de simulação, todas as estimativas foram obtidas através do *software* *R*, utilizando-se do pacote "nlm". Assim, como critério de convergência, o pacote utiliza um valor inteiro indicando quando o processo de otimização é concluído. Desta forma, 5 níveis são apresentados, sendo:

1: o ponto encontrado é a provável solução, ou seja, o gradiente da função está próximo de zero;

2: após sucessivas iterações dentro da tolerância, a ultima é provavelmente a solução;

3: o último passo foi reprovado como sendo a solução, ou seja, existe outro ponto melhor que o estimado. Cada estimativa é um mínimo local aproximado da função ou o erro de uma iteração para outra é muito pequeno;

4: o limite de iterações excedeu;

5: não existe ponto aproximado de convergência.

Utilizando-se deste critério apresentado em todo estudo de simulação, as estimativas dos parâmetros dos modelos convergiram com critério 1 em $99,328\%$ das vezes. Também, houve convergência com critério 2, em $0,666\%$ dos casos.

Estes percentuais foram obtidos em 96.000 estimativas obtidas para cada caso estudado, ou seja, um total de 192.000 estimativas calculadas. Em 191.988 das vezes houve convergência dos parâmetros para o modelo adequado ou inadequado. Em apenas 12 tentativas de se estimar os parâmetros dos modelos trabalhados, observou-se o critério de

convergência 4, ou seja, o limite de iterações do software R excedeu e não conseguiu-se estimar os parâmetros da distribuição.

Apesar deste problema, os métodos de seleção de modelos mostraram-se aptos a identificar qual a distribuição dos tempos.

Capítulo 5

Conclusões

Com o objetivo de trabalhar com dados relativos a carteiras de clientes de financiadoras e empresas ligadas a área de finanças, estudamos os conceitos de análise de sobrevivência e confiabilidade e os aplicamos a dados reais. Esta aplicação pôde ser vista no Capítulo 3. Verificamos que na área financeira é comum encontrarmos duas particularidades: grandes bancos de dados e a presença de muitas observações censuradas. Diferente do que ocorre na área médica e/ou biológica, onde a maioria dos estudos contemplam amostras de tamanho pequeno e poucas observações censuradas.

Ao analisar os dados, nos deparamos com problemas para verificar adequabilidade do modelo a ser escolhido. Este problema se deve ao tamanho do banco de dados em questão (65.535 observações). Desta forma, alguns modelos que não são apropriados aos dados foram ajustados e não houve problemas em se estimar os parâmetros destes modelos (na maior parte das vezes houve a convergência das estimativas dos parâmetros dos modelos estimados), devido ao número grande de observações.

Nesta aplicação, usamos das métricas usuais para selecionar o modelo mais adequado aos dados. Assim, verificamos que a medida de distância entre curvas, através da norma Euclidiana e, dos critérios AIC e BIC, foram boas métricas para selecionar modelos quando temos a presença de tempos com longa-duração.

Verificamos que estas métricas são adequadas não apenas para tamanhos de amostras grandes, como para tamanhos de amostras pequenos, uma vez que, selecionamos para validação 1% do tamanho original da amostra com percentuais de censuras proporcionais a esta amostra.

Para verificarmos os resultados obtidos no estudo de um caso particular da área financeira, desenvolvemos um estudo de simulação para duas distribuições usuais na área de finanças: a distribuição log-logística e a distribuição Weibull. Escolhidas por estas apresentarem diferentes formas para a função de risco. A distribuição log-logística apresenta

forma da função de risco crescente, decrescente e unimodal; enquanto que a distribuição Weibull tem formas crescentes, descrescentes e constantes.

Foram empregados três procedimentos para seleção dos modelos: AIC, BIC e uma medida de distância entre curvas (norma Euclidiana).

Nos casos de amostras geradas da distribuição Weibull, com parâmetro de forma decrescente, $\beta = 0,5$, todos os critérios conseguem discriminar os modelos, entretanto, a norma Euclidiana tem desempenho melhor. Nestes casos, as razões de normas estão próximas de zero diferente do que occorre para as métricas usuais, AIC e BIC. Para estas métricas, as razões estão mais próximas de 1.

Para os casos de amostras geradas por uma exponencial (Weibull com $\beta = 1$) ou, para amostras geradas da Weibull com risco crescente ($\beta = 1,5$), temos as mesmas conclusões.

Quando geramos amostras da distribuição log-logística, optou-se por estudar os casos de risco unimodal. Assim, estudamos os casos de $\sigma = 1/2$, $1/3$ e $1/4$. Para estes casos, verificamos que, à medida que diminui o valor de σ, a norma Euclidiana discrimina mais intensamente os modelos. Diferente dos critérios AIC e BIC que, a medida que aumenta o valor de σ, diminui a variabilidade das razões dessas métricas.

De forma geral, os critérios AIC e BIC demonstraram-se adequados para seleção de modelos em todos os tamanhos de amostras e percentuais de censura estudados ($n = 100$, 5000, 15000 e 30000; $p = 10\%$, 25%, 50% e 75%). Também de forma geral, a norma Euclidiana demonstrou-se uma boa alternativa as métricas usuais estudadas.

Finalmente podemos concluir que, para grandes bancos de dados e, na presença de longa-duração, características estas encontradas em carteiras de clientes de financiadoras, bancos ou seguradoras, é possível analisar e selecionar o modelo apropriado aos dados de forma usual, através das métricas conhecidas ou, de forma prática, medindo-se as distâncias entre as curvas estimadas e empírica.

Apêndice A

Tabelas Simulação

Tabela A.1: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 100, geradas para o modelo Weibull com β_0's especificados.

	Weibull			Log-Logístico		
β_0	$0,5$	$1,0$	$1,5$	2	3	4
	10% de censura					
p	$0,0948$	$0,0974$	$0,0982$	$0,0437$	$0,0633$	$0,0736$
μ	$1,4938$	$1,2020$	$1,1280$	$0,7850$	$0,8441$	$0,8796$
β	$0,5732$	$1,1499$	$1,7211$	$0,7519$	$1,5504$	$2,3691$
LV	$-162,9877$	$-130,9809$	$-104,8676$	$-164,0960$	$-132,8634$	$-107,1434$
AIC	$331,9754$	$267,9619$	$215,7352$	$334,1919$	$271,7267$	$220,2869$
BIC	$339,7909$	$275,7774$	$223,5507$	$342,0075$	$279,5422$	$228,1024$
NE	$0,2468$	$0,2467$	$0,2483$	$0,9756$	$1,1786$	$1,6168$
	25% de censura					
p	$0,2435$	$0,2476$	$0,2486$	$0,2009$	$0,2214$	$0,2310$
μ	$1,6674$	$1,2446$	$1,1523$	$0,8645$	$0,8582$	$0,8858$
β	$0,5693$	$1,1441$	$1,7115$	$0,7546$	$1,5616$	$2,3877$
LV	$-168,3720$	$-140,3614$	$-118,2117$	$-168,9735$	$-141,6380$	$-119,8256$
AIC	$342,7441$	$286,7227$	$242,4235$	$343,9469$	$289,2761$	$245,6511$
BIC	$350,5596$	$294,5382$	$250,2390$	$351,7624$	$297,0916$	$253,4666$
NE	$0,2277$	$0,2264$	$0,2275$	$0,8034$	$0,8988$	$1,2823$
	50% de censura					
p	$0,4930$	$0,4976$	$0,4986$	$0,4675$	$0,4816$	$0,4877$
μ	$1,6660$	$1,1587$	$1,0909$	$0,9193$	$0,8165$	$0,8524$
β	$0,6068$	$1,2213$	$1,8316$	$0,7979$	$1,6432$	$2,5059$
LV	$-133,6989$	$-119,1193$	$-105,6502$	$-135,1634$	$-120,9717$	$-107,7333$
AIC	$273,3977$	$244,2387$	$217,3005$	$276,3268$	$247,9435$	$221,4666$
BIC	$281,2132$	$252,0542$	$225,1160$	$284,1423$	$255,7590$	$229,2821$
NE	$0,1665$	$0,1641$	$0,1637$	$0,4921$	$0,5629$	$0,8321$
	75% de censura					
p	$0,7356$	$0,7478$	$0,7489$	$0,7167$	$0,7379$	$0,7430$
μ	$7,1782$	$1,4613$	$1,2578$	$7,1960$	$1,0823$	$0,9999$
β	$0,6699$	$1,3618$	$2,0468$	$0,8711$	$1,8071$	$2,7620$
LV	$-97,2950$	$-84,3031$	$-75,6618$	$-97,9320$	$-85,1572$	$-76,6696$
AIC	$200,5900$	$174,6061$	$157,3236$	$201,8640$	$176,3145$	$159,3392$
BIC	$208,4055$	$182,4216$	$165,1391$	$209,6795$	$184,1300$	$167,1547$
NE	$0,0926$	$0,0893$	$0,0878$	$0,2828$	$0,2341$	$0,3997$

Tabela A.2: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 5.000, geradas para o modelo Weibull com β_0's especificados.

	Weibull			Log-Logístico		
β_0	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,099	0,100	0,100	0,080	0,0874	0,092
μ	0,992	0,996	0,997	0,426	0,642	0,740
β	0,495	0,990	1,485	0,6732	1,3721	2,082
LV	−6600,457	−6121,105	−5177,363	−6792,8653	−6342,585	−5415,086
AIC	13206,915	12248,209	10360,726	13591,731	12691,171	10836,171
BIC	13226,466	12267,761	10380,277	13611,282	12710,722	10855,723
NE	0,296	0,298	0,301	7,362	8,199	10,724
	25% de censura					
p	0,250	0,250	0,250	0,233	0,240	0,243
μ	1,004	1,001	1,001	0,434	0,648	0,745
β	0,499	0,998	1,496	0,680	1,386	2,102
LV	−7005,655	−6568,104	−5768,740	−7163,923	−6750,209	−5964,234
AIC	14017,311	13142,208	11543,479	14333,846	13506,418	11934,467
BIC	14036,863	13161,760	11563,031	14353,397	13525,970	11954,019
NE	0,281	0,281	0,283	5,577	6,276	8,264
	50% de censura					
p	0,500	0,500	0,500	0,489	0,493	0,495
μ	1,010	1,003	1,002	0,435	0,648	0,745
β	0,499	0,998	1,497	0,684	1,392	2,112
LV	−6269,724	−5972,268	−5437,409	−6369,359	−6087,527	−5561,518
AIC	12545,448	11950,535	10880,817	12744,718	12181,053	11129,035
BIC	12564,999	11970,087	10900,369	12764,270	12200,605	11148,587
NE	0,195	0,194	0,195	3,067	3,438	4,624
	75% de censura					
p	0,750	0,750	0,750	0,744	0,746	0,747
μ	1,021	1,005	1,003	0,436	0,646	0,742
β	0,487	0,975	1,462	0,657	1,340	2,035
LV	−4200,100	−4076,491	−3817,384	−4253,439	−4138,627	−3884,660
AIC	8406,201	8158,983	7640,767	8512,877	8283,253	7775,321
BIC	8425,753	8178,534	7660,319	8532,429	8302,805	7794,873
NE	0,102	0,101	0,101	1,133	1,202	1,700

Tabela A.3: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 15.000, geradas para o modelo Weibull com β_0's especificados.

	Weibull			Log-Logístico		
β_0	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,100	0,100	0,100	0,081	0,089	0,093
μ	0,989	0,994	0,996	0,425	0,642	0,740
β	0,503	1,007	1,510	0,687	1,399	2,121
LV	−19812,842	−18257,533	−15385,765	−20405,609	−18935,585	−16111,244
AIC	39631,685	36521,065	30777,529	40817,218	37877,169	32228,489
BIC	39654,532	36543,913	30800,377	40840,065	37900,017	32251,336
NE	0,264	0,266	0,269	12,742	14,356	18,721
	25% de censura					
p	0,250	0,250	0,250	0,234	0,240	0,244
μ	0,986	0,993	0,995	0,423	0,640	0,739
β	0,502	1,004	1,505	0,685	1,396	2,117
LV	−20845,779	−19581,206	−17198,397	−21320,151	−20125,597	−17782,446
AIC	41697,558	39168,412	34402,794	42646,303	40257,194	35570,892
BIC	41720,406	39191,259	34425,642	42669,150	40280,041	35593,739
NE	0,256	0,257	0,258	9,741	10,925	14,355
	50% de censura					
p	0,500	0,500	0,500	0,489	0,493	0,496
μ	1,001	1,000	1,000	0,429	0,645	0,742
β	0,499	0,998	1,497	0,682	1,389	2,107
LV	−18766,403	−17895,472	−16297,623	−19072,710	−18249,379	−16678,646
AIC	37538,807	35796,945	32601,246	38151,419	36504,757	33363,292
BIC	37561,654	35819,792	32624,093	38174,267	36527,604	33386,140
NE	0,207	0,206	0,207	5,318	5,935	7,980
	75% de censura					
p	0,750	0,750	0,750	0,744	0,747	0,748
μ	1,009	1,002	1,001	0,435	0,648	0,745
β	0,500	0,999	1,499	0,681	1,388	2,105
LV	−12629,645	−12189,999	−11389,608	−12786,920	−12371,352	−11584,768
AIC	25265,290	24385,999	22785,216	25579,839	24748,703	23175,535
BIC	25288,138	24408,846	22808,064	25602,686	24771,550	23198,383
NE	0,115	0,115	0,115	1,910	2,115	3,017

Tabela A.4: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 30.000, geradas para o modelo Weibull com β_0's especificados.

	Weibull			Log-Logístico		
β_0	$0,5$	$1,0$	$1,5$	2	3	4
	10% de censura					
p	$0,100$	$0,100$	$0,100$	$0,088$	$0,092$	$0,095$
μ	$0,979$	$0,989$	$0,993$	$0,408$	$0,632$	$0,734$
β	$0,499$	$0,999$	$1,498$	$0,692$	$1,403$	$2,121$
LV	$-39342,542$	$-36474,131$	$-30810,082$	$-40650,996$	$-37895,020$	$-32295,225$
AIC	$78691,084$	$72954,262$	$61626,163$	$81307,992$	$75796,039$	$64596,451$
BIC	$78716,011$	$72979,188$	$61651,090$	$81332,919$	$75820,966$	$64621,378$
NE	$0,341$	$0,343$	$0,345$	$18,013$	$20,348$	$26,419$
	25% de censura					
p	$0,250$	$0,250$	$0,250$	$0,234$	$0,240$	$0,243$
μ	$0,979$	$0,989$	$0,993$	$0,419$	$0,638$	$0,737$
β	$0,500$	$1,001$	$1,502$	$0,683$	$1,391$	$2,110$
LV	$-41522,372$	$-39107,533$	$-34379,123$	$-42470,427$	$-40196,162$	$-35548,068$
AIC	$83050,745$	$78221,065$	$68764,246$	$84946,854$	$80398,323$	$71102,135$
BIC	$83075,671$	$78245,992$	$68789,173$	$84971,781$	$80423,250$	$71127,062$
NE	$0,302$	$0,302$	$0,303$	$13,787$	$15,445$	$20,287$
	50% de censura					
p	$0,500$	$0,500$	$0,500$	$0,490$	$0,494$	$0,496$
μ	$0,977$	$0,988$	$0,992$	$0,419$	$0,638$	$0,737$
β	$0,502$	$1,004$	$1,506$	$0,686$	$1,397$	$2,118$
LV	$-37196,870$	$-35579,921$	$-32425,206$	$-37829,028$	$-36304,422$	$-33202,535$
AIC	$74399,740$	$71165,842$	$64856,412$	$75664,056$	$72614,844$	$66411,069$
BIC	$74424,667$	$71190,769$	$64881,339$	$75688,983$	$72639,771$	$66435,996$
NE	$0,209$	$0,208$	$0,209$	$7,543$	$8,455$	$11,351$
	75% de censura					
p	$0,750$	$0,750$	$0,750$	$0,745$	$0,747$	$0,748$
μ	$1,003$	$1,000$	$1,000$	$0,430$	$0,645$	$0,743$
β	$0,499$	$0,999$	$1,498$	$0,682$	$1,389$	$2,108$
LV	$-25247,494$	$-24371,688$	$-22771,865$	$-25553,023$	$-24725,009$	$-23152,628$
AIC	$50500,989$	$48749,375$	$45549,730$	$51112,046$	$49456,019$	$46311,255$
BIC	$50525,916$	$48774,302$	$45574,657$	$51136,973$	$49480,945$	$46336,182$
NE	$0,116$	$0,115$	$0,115$	$2,711$	$2,987$	$4,269$

Tabela A.5: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 100, geradas para o modelo log-logístico com σ_0's especificados.

	Weibull			Log-Logístico		
$1/\sigma_0$	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,0995	0,0995	0,0995	0,0959	0,0965	0,0968
μ	2,9982	2,9068	2,8612	13,2963	2,6273	2,6498
β	1,2301	1,8350	2,4331	2,1577	3,2391	4,3191
LV	−382,8341	−351,1814	−327,9848	−374,0576	−342,0875	−318,5483
AIC	771,6683	708,3628	661,9697	754,1152	690,1751	643,0965
BIC	779,4838	716,1783	669,7852	761,9307	697,9906	650,9120
NE	4,3511	4,4584	4,5139	1,0977	3,2327	3,1259
	25% de censura					
p	0,2493	0,2495	0,2496	0,2463	0,2470	0,2474
μ	3,0120	2,9166	2,8692	2,5849	2,6284	2,6505
β	1,2434	1,8552	2,4601	2,1068	3,1627	4,2160
LV	−349,0972	−322,8992	−303,6911	−342,9686	−316,5200	−297,0410
AIC	704,1943	651,7984	613,3822	691,9371	639,0400	600,0820
BIC	712,0098	659,6139	621,1977	699,7526	646,8555	607,8975
NE	4,1107	4,3858	4,5187	2,4142	2,4475	2,3767
	50% de censura					
p	0,4986	0,4993	0,4995	0,4969	0,4977	0,4982
μ	3,1192	2,9843	2,9186	2,6760	2,6878	2,6945
β	1,3558	2,0313	2,7030	2,3313	3,5031	4,6739
LV	−267,3236	−247,7889	−233,8675	−262,2931	−242,7169	−228,7418
AIC	540,6473	501,5778	473,7350	530,5863	491,4338	463,4836
BIC	548,4628	509,3933	481,5505	538,4018	499,2493	471,2991
NE	4,1002	4,7551	4,9780	1,3638	1,3437	1,2870
	75% de censura					
p	0,7485	0,7492	0,7495	0,7421	0,7455	0,7470
μ	2,8794	2,8241	2,7985	2,5775	2,6118	2,6345
β	1,6233	2,4345	3,2373	2,4025	3,6493	4,8945
LV	−146,1475	−137,9652	−131,8042	−145,9738	−137,8945	−131,7767
AIC	298,2949	281,9304	269,6083	297,9476	281,7889	269,5534
BIC	306,1104	289,7459	277,4238	305,7632	289,6044	277,3689
NE	5,6903	6,1292	6,2020	0,4734	0,4549	0,4326

Tabela A.6: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 5.000, geradas para o modelo log-logístico com σ_0's especificados.

	Weibull			Log-Logístico		
$1/\sigma_0$	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,100	0,100	0,100	0,100	0,100	0,100
μ	3,184	3,029	2,951	15,232	2,722	2,721
β	1,040	1,559	2,076	1,988	2,984	3,979
LV	−20368,372	−18537,414	−17241,343	−19793,424	−17960,013	−16661,257
AIC	40742,743	37080,828	34488,687	39592,848	35926,026	33328,515
BIC	40762,295	37100,380	34508,238	39612,400	35945,578	33348,066
NE	30,765	31,644	32,125	8,566	23,871	23,239
	25% de censura					
p	0,250	0,250	0,250	0,250	0,250	0,250
μ	3,182	3,027	2,950	2,717	2,718	2,718
β	1,036	1,553	2,069	2,003	3,006	4,009
LV	−18438,089	−16915,004	−15835,745	−17917,410	−16393,232	−15312,737
AIC	36882,177	33836,008	31677,490	35840,821	32792,465	30631,473
BIC	36901,729	33855,560	31697,042	35860,372	32812,016	30651,025
NE	28,583	31,024	32,144	17,902	18,149	17,650
	50% de censura					
p	0,500	0,500	0,500	0,500	0,500	0,500
μ	3,186	3,030	2,952	2,717	2,717	2,717
β	1,040	1,560	2,079	2,007	3,012	4,017
LV	−13894,093	−12878,175	−12158,300	−13538,003	−12521,571	−11801,088
AIC	27794,187	25762,351	24322,600	27082,005	25049,142	23608,176
BIC	27813,739	25781,902	24342,151	27101,557	25068,694	23627,728
NE	27,251	33,357	35,147	9,752	9,878	9,603
	75% de censura					
p	0,750	0,750	0,750	0,750	0,750	0,750
μ	3,213	3,048	2,966	2,721	2,720	2,719
β	1,028	1,541	2,055	1,943	2,917	3,890
LV	−8071,006	−7559,648	−7197,930	−7901,784	−7390,361	−7028,516
AIC	16148,012	15125,295	14401,860	15809,567	14786,722	14063,032
BIC	16167,564	15144,847	14421,411	15829,119	14806,274	14082,583
NE	29,174	40,574	43,132	3,444	3,500	3,413

Tabela A.7: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 15.000, geradas para o modelo log-logístico com σ_0's especificados.

	Weibull			Log-Logístico		
$1/\sigma_0$	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,100	0,100	0,100	0,100	0,100	0,100
μ	3,176	3,023	2,947	15,210	2,721	2,720
β	1,025	1,537	2,048	2,021	3,032	4,043
LV	−61110,319	−55618,650	−51729,576	−59145,651	−53647,770	−49752,292
AIC	122226,638	111243,300	103465,152	118297,303	107301,539	99510,585
BIC	122249,486	111266,147	103487,999	118320,150	107324,387	99533,432
NE	53,246	54,780	55,631	14,743	41,290	40,139
	25% de censura					
p	0,248	0,250	0,250	0,250	0,250	0,250
μ	3,199	3,029	2,951	2,724	2,722	2,721
β	1,011	1,521	2,027	2,017	3,025	4,034
LV	−55545,060	−50893,204	−47643,190	−53734,391	−49138,437	−45884,998
AIC	111096,120	101792,409	95292,380	107474,782	98282,874	91775,996
BIC	111118,967	101815,256	95315,228	107497,629	98305,721	91798,843
NE	49,353	53,582	55,590	30,989	31,434	30,563
	50% de censura					
p	0,500	0,500	0,500	0,500	0,500	0,500
μ	3,185	3,028	2,951	2,720	2,720	2,719
β	1,034	1,551	2,067	2,004	3,007	4,010
LV	−41683,623	−38630,693	−36469,062	−40618,979	−37566,713	−35403,630
AIC	83373,246	77267,385	72944,124	81243,958	75139,425	70813,261
BIC	83396,093	77290,233	72966,972	81266,806	75162,273	70836,108
NE	47,267	57,948	60,935	16,853	17,102	16,638
	75% de censura					
p	0,750	0,750	0,750	0,750	0,750	0,750
μ	3,185	3,030	2,952	2,717	2,718	2,718
β	1,050	1,575	2,100	2,006	3,010	4,014
LV	−24058,116	−22532,516	−21451,464	−23546,733	−22020,961	−20939,597
AIC	48122,232	45071,033	42908,928	47099,466	44047,922	41885,194
BIC	48145,080	45093,880	42931,775	47122,313	44070,769	41908,041
NE	52,866	72,052	75,047	5,968	6,049	5,882

Tabela A.8: Tabela das médias dos parâmetros e métricas, dos resultados obtidos das amostras de tamanho 30.000, geradas para o modelo log-logístico com σ_0's especificados.

	Weibull			Log-Logístico		
$1/\sigma_0$	0,5	1,0	1,5	2	3	4
	10% de censura					
p	0,100	0,100	0,100	0,100	0,100	0,100
μ	3,184	3,029	2,951	15,289	2,724	2,723
β	0,978	1,467	1,955	2,011	3,016	4,022
LV	−123223,241	−112191,079	−104385,597	−118580,316	−107540,267	−99726,828
AIC	246452,482	224388,158	208777,193	237166,631	215086,535	199459,656
BIC	246477,409	224413,085	208802,120	237191,558	215111,462	199484,583
NE	75,065	77,238	78,520	21,005	58,444	56,855
	25% de censura					
p	0,249	0,250	0,250	0,250	0,250	0,250
μ	3,209	3,033	2,954	2,728	2,725	2,723
β	0,968	1,455	1,939	2,012	3,017	4,023
LV	−111829,121	−102545,914	−96035,182	−107598,698	−98389,652	−91873,977
AIC	223664,241	205097,829	192076,363	215203,395	196785,305	183753,953
BIC	223689,168	205122,756	192101,290	215228,322	196810,232	183778,880
NE	69,131	75,096	78,195	43,809	44,467	43,256
	50% de censura					
p	0,500	0,500	0,500	0,500	0,500	0,500
μ	3,189	3,032	2,954	2,728	2,725	2,723
β	1,013	1,519	2,025	2,022	3,033	4,044
LV	−83629,253	−77485,818	−73140,903	−81219,715	−75074,686	−70728,068
AIC	167264,505	154977,635	146287,806	162445,430	150155,372	141462,137
BIC	167289,432	155002,562	146312,733	162470,357	150180,299	141487,064
NE	65,872	81,489	85,988	23,879	24,211	23,535
	75% de censura					
p	0,750	0,750	0,750	0,750	0,750	0,750
μ	3,186	3,030	2,952	2,721	2,720	2,719
β	1,036	1,554	2,072	2,004	3,006	4,009
LV	−48174,209	−45119,765	−42955,812	−47104,003	−44049,271	−41884,871
AIC	96354,418	90245,530	85917,623	94214,006	88104,542	83775,741
BIC	96379,345	90270,457	85942,550	94238,933	88129,469	83800,668
NE	73,974	101,890	106,103	8,431	8,555	8,323

Apêndice B

Programas Usados

Seguem abaixo, alguns dos comandos utilizados para gerar os resultados apresentados no decorrer deste trabalho. Estas linhas foram executadas no *Software R*.

Listagem B.1: Geração da amostra Weibull.

```
1   library(survival)
2   n <- tamanho_amostra
3   namostra <- n_amostra
4   shape <- forma
5   scale <- escala
6   # VALOR DE P
7   p <- 0.1
8   #dados Weibull
9   amostra <- rweibull(n*(1-p),shape,scale)
10  #dados uniformes
11  amostra=sort(amostra)
12  amostra=c(amostra,runif(n*p,max(amostra),
13              3*max(amostra)))
14  #delta
15  delta=c(rep(1,n*(1-p)),rep(0,n*p))
16  # BOOTSTRAP
17  set.seed(1234)
18  dados <- matrix(NA,n,namostra)
19  for(i in 1:namostra)
20  {
21  id <- sample(1:n,n,replace=T)
22  dados[,i] <- amostra[id]
```

```
23      dados[,i] <- sort(dados[,i])
24      }
```

Listagem B.2: Geração da amostra Log-logística.

```
1       library(survival)
2       n <- tamanho_amostra
3       namostra <- n_amostra
4       shape <- 1/sigma
5       scale <- exp(mu)
6       # VALOR DE P
7       p <- 0.1
8       amostra_logis <- rlogis(n*(1-p),0,1)
9       amostra <- exp(scale+shape*amostra_logis)
10      #dados uniformes
11      amostra=sort(amostra);
12      amostra=c(amostra,runif(n*p,max(amostra),
13                3*max(amostra)))
14      #delta
15      delta=c(rep(1,n*(1-p)),rep(0,n*p))
16      chute <- c(p,scale,shape)
17      # BOOTSTRAP
18      set.seed(1234)
19      dados <- matrix(NA,n,namostra)
20      for(i in 1:namostra)
21      {
22      id <- sample(1:n,n,replace=T)
23      dados[,i] <- amostra[id]
24      dados[,i] <- sort(dados[,i])
25      }
```

Listagem B.3: Estimativas Weibull.

```
1       estimativas_weibull <- matrix(0,1000,8)
2       for(m in 1:namostra)
3       {
4       chute <- c(p,scale,shape)
5         #ESTIMANDO OS ÂPARMETROS WEIBULL
6         weibull <- function(theta)
```

```
7         {
8         gama=exp(theta[1])/(1+exp(theta[1]))
9         mu=exp(theta[2])
10        beta1=exp(theta[3])
11              l <- sum(delta)*log(gama*beta1/(mu**beta1))
12              + sum((beta1-1)*delta*log(dados[,m])) -
13              sum(delta*((dados[,m]/mu)**beta1)) + sum((1-delta)*
14              log(1-gama+gama*exp(-(dados[,m]/mu)**beta1)))
15              -l
16          }
17        est_weibull <- nlm(weibull, chute)
18        ekm <- survfit(Surv(dados[,m],delta))
19        gamaw1 = exp(est_weibull$estimate[1])/
20                 (1+exp(est_weibull$estimate[1]))
21        beta = exp(est_weibull$estimate[3])
22        mu = exp(est_weibull$estimate[2])
23        sw <- (1-gamaw1) + gamaw1*(exp(-(ekm$time/mu)**beta))
24        normaw <- sqrt(sum((ekm$surv-sw)**2))
25        log_ver = -weibull(c(est_weibull$estimate[1],
26                    est_weibull$estimate[2],est_weibull$estimate[3]))
27        aicw <- -2*(log_ver)+3*2
28        bicw <- -2*(log_ver)+3*log(n)
29        estimativas_weibull[m,] = c((1-gamaw1),mu,beta,
30                                    est_weibull$code,
31                                    log_ver,normaw,aicw,bicw)
32    }
```

Listagem B.4: Estimativas Log-logístico.

```
1  estimativas_llogis <- matrix(0,1000,8)
2  for(m in 1:namostra)
3  {
4   chute <- c(p,scale,shape)
5   llogis<-function(x)
6   {
7     mu<-exp(x[1]);
8     beta1<-1/(x[2]);
9     gama=exp(x[3])/(1+exp(x[3]))
```

```
10      l1=log(gama*beta1/(mu^beta1))
11      l2=(beta1-1)*log(dados[,m])
12      l3=2*(log(1+(dados[,m]/mu)^beta1))
13      l4=(1-delta)*log(1-gama+gama/(1+(dados[,m]/mu)^beta1))
14      f= -sum(delta*(l1+l2-l3) +(1-delta)*l4)
15      g <-(f)
16      g
17     }
18    est_llogis <- nlm(llogis,chute)
19    ekm = survfit(Surv(dados[,m],delta))
20    gamall1 = (exp(est_llogis$estimate[3])/
21              (1+exp(est_llogis$estimate[3])))
22    mu = exp(est_llogis$estimate[1])
23    beta = 1/est_llogis$estimate[2]
24    sllogis = (1-gamall1)+gamall1/(1+exp((ekm$time-mu)/beta))
25    normallogis = sqrt(sum((ekm$surv-sllogis)**2))
26    aicll = -2*(-llogis(c(est_llogis$estimate[1],
27            est_llogis$estimate[2],est_llogis$estimate[3])))+3*2
28    bicll = -2*(-llogis(c(est_llogis$estimate[1],
29            est_llogis$estimate[2],est_llogis$estimate[3])))+3*log(n)
30    estimativas_llogis[m,] = c(1-gamall1,mu,beta,
31                      est_llogis$code,
32                      -llogis(c(est_llogis$estimate[1],
33                      est_llogis$estimate[2],
34                      est_llogis$estimate[3])),
35                      normallogis,aicll,bicll)
36  }
```

Listagem B.5: Exemplo geração figuras Cap. 3.

```
1       dev.off()
2       pdf(file="C:\\graficos\\weibull-risco.pdf")
3       t<- seq(0.01,3,0.01)
4       Survival      <- function(t,mu,beta)
5       {
6            (beta/mu)*(t/mu)**(beta-1)
7       }
8       mu          <-  1.5
```

```
9       beta      <-  0.5
10      S         <-  Survival(t,mu,beta)
11      plot(t,S,type="l",ylim=c(0,3),xlim=c(0,3),
12           lty=1,font=7,font.axis=3,font.lab=3,
13           lwd=2,ylab="h(t)", xlab="t")
14      beta      <-  1
15      S         <-  Survival(t,mu,beta)
16      lines(t,S,lty=3,lwd=2)
17      beta      <-  1.5
18      S         <-  Survival(t,mu,beta)
19      lines(t,S,lty=4,lwd=2)
20      beta      <-  3.0
21      S         <-  Survival(t,mu,beta)
22      lines(t,S,lty=5,lwd=2)
23      legend(1.8,2.6,col=c("black","black","black","black"),
24                bty="n",lty=c(1,3,4,5),lwd=2,
25                c(expression(beta=="0,5"),
26                expression(beta=="1,0"),
27                expression(beta=="1,5"),
28                expression(beta=="3,0")))
29      legend(2.5,2.4,col="black",bty="n",
30             expression(mu=="1,5"))
31      dev.off()
32      #
33      dev.off()
34      pdf(file="C:\\graficos\\weibull-survival.pdf")
35      t<- seq(0.01,3,0.01)
36      Survival      <- function(t,mu,beta)
37      {
38              exp(-(t/mu)**beta)
39      }
40      mu        <-  1.5
41      beta      <-  0.5
42      S         <-  Survival(t,mu,beta)
43      plot(t,S,type="l",ylim=c(0,1),xlim=c(0,2.5),
44           lty=1,font=7,font.axis=3,font.lab=3,
45           lwd=2,ylab="S(t)", xlab="t")
```

```
46      beta     <- 1
47      S        <- Survival(t,mu,beta)
48      lines(t,S,lty=3,lwd=2)
49      beta     <- 1.5
50      S        <- Survival(t,mu,beta)
51      lines(t,S,lty=4,lwd=2)
52      beta     <- 3.0
53      S        <- Survival(t,mu,beta)
54      lines(t,S,lty=5,lwd=2)
55      legend(1.3,0.83,col=c("black","black","black","black"),
56          bty="n",lty=c(1,3,4,5),lwd=2,
57          c(expression(beta=="0,5"),
58          expression(beta=="1,0"),
59          expression(beta=="1,5"),
60          expression(beta=="3,0")))
61      legend(2,0.78,col="black",bty="n",
62          expression(mu=="1,5"))
63      dev.off()
```

Referências Bibliográficas

Akaike, H. (1973). Information Theory and the Maximum Likelihood Principle, in International Simposium on Information Theory, eds. V. Petrov and F. Csáki, Budapest: Akademiai Kiádo.

Anderson, D. R.; Burnham, K. P. (2004). Multimodel Inference Understanding AIC and BIC in Model Selection. Sociological Methods e Research, Vol. 33, No. 2, 261 - 304.

Barlow, R. E. ; Campo, R. A. (1975), Time on Test Processes and Applications to Failure Data Analysis, Caligornia University Berkeley Operations Research Center.

Berkson, J. and Gage, R. (1952), Survival cure for cancer patients following treatment, Journal of the American Statistical Association 47, 501 - 515.

Breslow, N. E., and Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates under Random Censoring. Annals of Statistics, 2, 437 - 453.

Bickel P. J.; Doksum, K. A., (1977). Mathematical Statistics. INC: Holden-Day.

Böhmer, P. E. (1912). Theorie der unabhangigen. Wahrscheinlichkeiten Rapports Memories et Proces-verbaux de Septieme Congres International d'Actuaires, 2 : 327 - 343.

Carlin, B. P., Lois, T. A. (2000). Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. Chapman and Hall, London.

Chen, M. H., Ibrahim, J. G., Sinha, D. (2001). Bayesian Survival Analysis. Springer Series in Statistics, New York.

Collett, D. (1994). Modelling Survival Data in Medical Research. Chapman and Hall, New York.

Colosimo, E. A., Giolo, S. R. (2006): Análise de Sobrevivência Aplicada. ABE - Projeto Fisher, São Paulo, Brasil.

Cox, D. R. e Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Efron, B.; Tibshirani, R.J. (1993). An Introduction to the Boostrap. New York: Chapman and Hall.

Glymour, C.; Mandingan, D.; Pregibon, D.; Smyth, P. (1997). Statistical Themes and Lessons for Data Mining. Kluwer Academic Publishers.

Hernandez, P. J., Navarro, J. and Wondmagegnehu, E. T. (2005): Bathtub Shaped Failure Rates From Mixtures: A Practical Point of View. IEEE Transactions on Reliability, Vol. 54, n. 2.

Hosmer, D. W., Lemeshow, S. (1999): Applied Survival Analysis, Regression Modeling of Time to Event Data. John Wiley & Sons.

Huber, P.J. (1981). Robust Statistics. New York: Wiley.

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

Kleinbaum, D. G., Klein, M. (2005) Survival Analysis A Self-Learning Text. Springer, Second Edition, USA.

Klein, J. P. e Moeschberger, M. L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag, New York.

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. John Wiley and Sons, New York, New York.

Lee, E. T. (1992). Statistical Methods for Survival Data Analysis, second edition, Wiley, New York.

Lee, E. T., Wang, J. W. (2003): Statistical Methods for Survival Data Analysis. John Wiley & Sons.

Louzada-Neto, F. and Rodrigues, J. (2007): On The Unification of The Long-Term Survival Models. Departamento de Estatística UFSCar, Brasil.

Louzada-Neto, F., Mazucheli, J., Achcar, J. A. (2001): Lifetime Models with Nonconstant Shape Parameters. Departamento de Estatística UFSCar, Brasil.

Louzada-Neto, F., Mazucheli, J., Achcar, J. A. (2001): Uma Introdução à Análise de Sobrevivência e Confiabilidade. Minicurso: III Jornada Regional de Estatística e II Semana da Estatística, Universidade Estadual de Maringá.

Maller, R. and Zhou, X. (1996), Survival Analysis with Long-Term Survivors, London: Wiley Series in Probability and Statistics.

Oliveira, W. (2000), CRM & e-business, Florianópolis: Visual Books, p. 154.

Paulino, C. D.; Turkman, M. A. A.; Murteira, B. Estatística Bayesiana. Lisboa: Fundação Calouste Gulbenkian, 2003.

Prentice, R. L. (1978). Linear rank tests with right censored data. Biometrika, 65 : 167 - 179.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6, 416-464.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo