ANÁLISE DE SOBREVIVÊNCIA APLICADA

Enrico Antônio Colosimo
Departamento de Estatística, UFMG

Suely Ruiz Giolo
Departamento de Estatística, UFPR

Prefácio

Sumário

Capítulo 1

Conceitos Básicos e Exemplos

1.1 Introdução

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado. A razão deste crescimento é o desenvolvimento e aprimoramento de técnicas estatísticas combinado com computadores cada vez mais velozes. Uma evidência quantitativa deste sucesso é o número de aplicações de análise de sobrevivência em medicina. Baillar III e Mosteller (1992, Capítulo 3) verificaram que o uso de métodos de análise de sobrevivência cresceu de 11% em 1979 para 32% em 1989 nos artigos do conceituado periódico *The New England Journal of Medicine*. Esta foi a área da estatística, segundo os autores, que mais se destacou no período avaliado. Outro indicador deste crescimento é apresentar os dois artigos mais citados em toda a literatura estatística no período de 1987 a 1989 (Stigler, 1994). Os artigos são o do estimador de Kaplan-Meier para a função de sobrevivência (Kaplan e Meier, 1958) e o do modelo de Cox (Cox, 1972).

Em análise de sobrevivência a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é denominado **tempo de falha**, podendo ser o tempo até a morte do paciente, bem como até a cura ou recidiva de uma doença. Em estudos de câncer, é usual o registro das datas correspondentes ao diagnóstico da doença, à remissão (após o tratamento, o paciente fica livre dos sintomas da doença), à recorrência da doença (recidiva) e à morte do paciente. O tempo de falha pode ser, por exemplo, do diagnóstico até a morte ou da remissão até a recidiva.

A principal característica de dados de sobrevivência é a presença de **censura**, que é a observação parcial da resposta. Isto se refere a situações em que, por al-

guma razão, o acompanhamento do paciente foi interrompido; seja porque o paciente mudou de cidade, o estudo terminou para a análise dos dados, ou o paciente morreu de causa diferente da estudada. Isto significa que toda informação referente à resposta se resume ao conhecimento de que o tempo de falha é superior àquele observado. Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e planejamento de experimentos, poderiam ser utilizadas na análise deste tipo de dados, provavelmente usando uma transformação para a resposta. Suponha, por exemplo, que o interesse seja comparar o tempo médio de vida de três grupos de pacientes. Se não houver censuras, pode-se usar as técnicas usuais de análise de variância para fazer tal comparação. No entanto, se houver censuras, que é o mais provável, tais técnicas não podem ser utilizadas pois elas necessitam de todos os tempos de falha. Desta forma, faz-se necessário o uso dos métodos de análise de sobrevivência, que possibilitam incorporar na análise estatística a informação contida nos dados censurados.

O termo análise de sobrevivência refere-se basicamente a situações médicas envolvendo dados censurados. Entretanto, condições similares ocorrem em outras áreas em que se usam as mesmas técnicas de análise de dados. Em engenharia, são comuns os estudos em que produtos ou componentes são colocados sob teste para se estimar características relacionadas aos seus tempos de vida, tais como o tempo médio ou a probabilidade de um certo produto durar mais do que 5 anos. Exemplos podem ser encontrados em Nelson (1990a), Meeker e Escobar (1998) e Freitas e Colosimo (1997). Os engenheiros denominam esta área de confiabilidade. O mesmo ocorre em ciências sociais, em que várias situações de interesse têm como resposta o tempo entre eventos (Allison, 1984; Elandt-Johnson e Johnson, 1980). Criminalistas estudam o tempo entre a liberação de presos e a ocorrência de crimes; estudiosos do trabalho se concentram em mudanças de empregos, desempregos, promoções e aposentadorias; demógrafos com nascimentos, mortes, casamentos, divórcios e migrações. O crescimento observado no número de aplicações em medicina também pode ser observado nestas outras áreas.

Este texto foi motivado por ilustrações essencialmente na área clínica. Desta forma, os exemplos e colocações são conduzidos para esta área. No entanto, enfatiza-se que as técnicas estatísticas são de ampla utilização em outras áreas do conhecimento como mencionado anteriormente.

1.2 Objetivo e Planejamento dos Estudos

Os estudos clínicos são investigações científicas realizados com o objetivo de provar ou não uma determinada hipótese de interesse. Estas investigações são conduzidas coletando dados e analisando-os através de métodos estatísticos. Em geral, estes estudos podem ser divididos nas seguintes três etapas:

1. Formulação da hipótese de interesse.

2. Planejamento e coleta dos dados.

3. Análise estatística dos dados para validar ou não a hipótese formulada.

Estas etapas são comuns em qualquer estudo envolvendo análise estatística de dados. No presente texto, o interesse envolve situações em que a variável resposta é o tempo até a ocorrência de um evento de interesse como descrito na Seção 1.1.

A primeira etapa de um estudo clínico é gerada pela curiosidade científica do pesquisador. Identificar fatores de risco ou prognóstico para uma doença é um objetivo que aparece freqüentemente em estudos clínicos. A comparação de drogas ou diferentes opções terapêuticas é outro objetivo usualmente encontrado neste tipo de estudo.

Os textos técnicos estatísticos concentram todo o esforço na terceira etapa, ou seja, na análise estatística dos dados, mesmo admitindo a importância de um adequado planejamento do estudo. Este texto não é diferente dos demais. No entanto, o restante desta seção é dedicado a uma breve descrição desta segunda etapa.

Em análise de sobrevivência a resposta é por natureza longitudinal. O delineamento destes estudos pode ser observacional ou experimental assim como ele pode ser retrospectivo ou prospectivo. As quatro formas básicas de estudos clínicos são: descritivo, caso-controle, coorte e clínico aleatorizado. Os três primeiros são observacionais e o quarto é experimental pois existe a intervenção do pesquisador ao alocar, de forma aleatória, tratamento ao paciente.

O estudo envolvendo somente uma amostra, usualmente de doentes, é descritivo pois não existe um grupo de comparação. Nestes estudos o objetivo é frequentemente a identificação de fatores de prognóstico para a doença em estudo. Os outros tipos de estudo são comparativos. Isto significa que o objetivo do estudo é a comparação de dois ou mais grupos.

O estudo caso-controle é usualmente retrospectivo. Um grupo de doentes (casos) e outro de não-doentes (controles) são comparados através da exposição a um ou

mais fatores de interesse. Este estudo é simples, de baixo custo e rápido pois a informação já está disponível. No entanto, ele sofre de algumas limitações por estar sujeito a alguns tipos de vícios. Estes vícios estão relacionados a informação disponível sobre a história da exposição assim como a incerteza sobre a escolha do grupo controle. Uma discussão mais profunda sobre este tipo de estudo foge o escobo deste livro. No entanto, devido a sua grande utilização existem várias bibliografias sobre o assunto, entre elas, Breslow e Day (1980) e Rothman e Greenland (1998).

As limitações do estudo caso-controle podem ser vencidas pelos estudos prospectivos conhecidos por coorte. Um grupo exposto e outro não-exposto ao fator de interesse são acompanhados por um período de tempo registrando a ocorrência da doença ou evento de interesse. A vantagem deste estudo sobre o caso-controle é poder avaliar a comparabilidade dos grupos no início do estudo e identificar as variáveis de interesse a serem medidas. Por outro lado, é um estudo longo e mais caro pois os indivíduos são acompanhados por um período de tempo muitas vezes superior a um ano. Também não é um estudo indicado para doenças consideradas raras. Uma referência importante sobre estes estudos é Breslow e Day (1987).

A forma mais consagrada de pesquisa clínica é o estudo clínico aleatorizado. Este estudo é importante por ser experimental. Isto significa que existe a intervenção direta do pesquisador ao alocar, de forma aleatória, tratamento ao paciente. Este procedimento garante a comparabilidade dos grupos. Este estudo é bastante analisado na literatura e pode-se citar os seguintes livros, entre outros, Pocock (?) e Friedman, Furbecg e DeMets (?).

Na Seção 1.5 é apresentado alguns exemplos reais que são analisados ao longo do livro. Entre estes exemplos existem estudos descritivos, de coorte e clínico aleatorizado.

Os estudos industriais são usualmente de campo ou realizados na própria empresa simulando situações de campo. No entanto, existem alguns estudos industriais planejados com o objetivo de reduzir o tempo de vida das unidades sob teste e desta forma obter dados amostrais mais rápidos. Estes estudos são chamados de testes de vida acelerados. Os itens amostrais são estressados para falhar mais rápido e através de modelos de regressão obtém-se as estimativas para as quantidades de interesse nas condições de uso utilizando extrapolações. Uma discussão mais profunda sobre estes testes pode ser encontrada em Nelson (1990) e Freitas e Colosimo (1997).

Uma extensão destes testes é o de degradação que pode ser ou não acelerado. Nestes testes uma variável numérica associada ao tempo de falha é registrada ao longo do período de acompanhamento. A partir destes valores é possível obter as

estimativas de interesse mesmo em situações em que nenhuma falha foi registrada. Estes testes estão ganhando espaço na literatura técnica de engenharia. Mais informações sobre estes testes podem ser encontradas em Meeker e Escobar (1998) e Oliveira e Colosimo (2004).

1.3 Caracterizando Dados de Sobrevivência

Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e, muito freqüentemente, pelas censuras. Estes dois componentes constituem a resposta. Em estudos clínicos, um conjunto de covariáveis é também, geralmente, medido em cada paciente. Os seguintes três elementos constituem o tempo de falha: o tempo inicial, a escala de medida e o evento de interesse (falha). Estes elementos devem ser claramente definidos e, juntamente com a censura, são discutidos em detalhes a seguir.

1.3.1 Tempo de Falha

O tempo de início do estudo deve ser precisamente definido. Os indivíduos devem ser comparáveis na origem do estudo, com exceção de diferenças medidas pelas covariáveis. Em um estudo clínico aleatorizado, a data da aleatorização é a escolha natural para a origem do estudo. A data do diagnóstico ou do início do tratamento de doenças também são outras escolhas possíveis.

A escala de medida é quase sempre o tempo real ou "de relógio", apesar de existirem outras alternativas. Em testes de engenharia podem surgir outras escalas de medida, como o número de ciclos, a quilometragem de um carro ou qualquer outra medida de carga.

O terceiro elemento é o evento de interesse. Estes eventos são, na maioria dos casos, indesejáveis e, como já mencionado, chamados de falha. É importante, em estudos de sobrevivência, definir de forma clara e precisa o que vem a ser a falha. Em algumas situações, a definição de falha já é clara, tais como morte ou recidiva, mas em outras pode assumir termos ambíguos. Por exemplo, fabricantes de produtos alimentícios desejam saber informações sobre o tempo de vida 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 iniciar o estudo. Por exemplo, o produto fica inapropriado para o consumo quando atingir mais que

uma determinada concentração de microorganismos por mm^2 de área do produto.

O evento de interesse (falha) pode ainda ocorrer devido a uma única causa ou devido a duas ou mais. Situações em que causas de falha competem entre si são denominadas na literatura de *riscos competitivos* (Prentice et al., 1978). Apenas as que consideram uma única causa de falha são abordadas neste texto.

1.3.2 Censura e Dados Truncados

Os estudos clínicos que envolvem uma resposta temporal são freqüentemente prospectivos e de longa duração. Mesmo sendo longos, os estudos clínicos de sobrevivência usualmente terminam antes que todos os indivíduos no estudo venham a falhar. Uma característica decorrente destes estudos é, então, a presença de observações incompletas ou parciais. Estas observações, denominadas censuras, podem ocorrem por uma variedade de razões, dentre elas, a perda de acompanhamento do paciente no decorrer do estudo e a não ocorrência do evento de interesse até o término do experimento. Note que toda informação obtida sobre estes indivíduos é que o seu tempo até o evento é superior ao tempo registrado até o último acompanhamento.

Ressalta-se o fato de que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística. Duas razões justificam tal procedimento: (i) mesmo sendo incompletas, as observações censuradas nos fornecem informações sobre o tempo de vida de pacientes; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode acarretar em conclusões viciadas.

Alguns mecanismos de censura são diferenciados em estudos clínicos. Censura do tipo I é aquela em que o estudo será terminado após um período pré-estabelecido de tempo. Censura do tipo II é aquela em que o estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos. Um terceiro mecanismo de censura, o do tipo aleatório, é o que mais ocorre na prática médica. Isto acontece quando um paciente é retirado no decorrer do estudo sem ter ocorrido a falha. Isto também ocorre, por exemplo, se o paciente morrer por uma razão diferente da estudada.

Uma representação simples do mecanismo de censura aleatória é feita usando duas variáveis aleatórias. Seja T uma variável aleatória representando o tempo de falha de um paciente e seja C uma variável aleatória independente de T, representando o tempo de censura associado a este paciente. Os dados observados são

$$t = \min(T, C)$$

e

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

Suponha que os pares (T_i, C_i), $i = 1, \ldots, n$, formam uma amostra aleatória de n pacientes. Observe que se todos $C_i = C$, uma constante fixa sob o controle do pesquisador, tem-se a censura do tipo I. Ou seja, a censura do tipo I é um caso particular da aleatória.

A Figura 1.1 ilustra estes mecanismos de censura. Os exemplos apresentados na Seções 1.5.1, 1.5.2 e 1.5.5 são caracterizados pela censura do tipo aleatória.

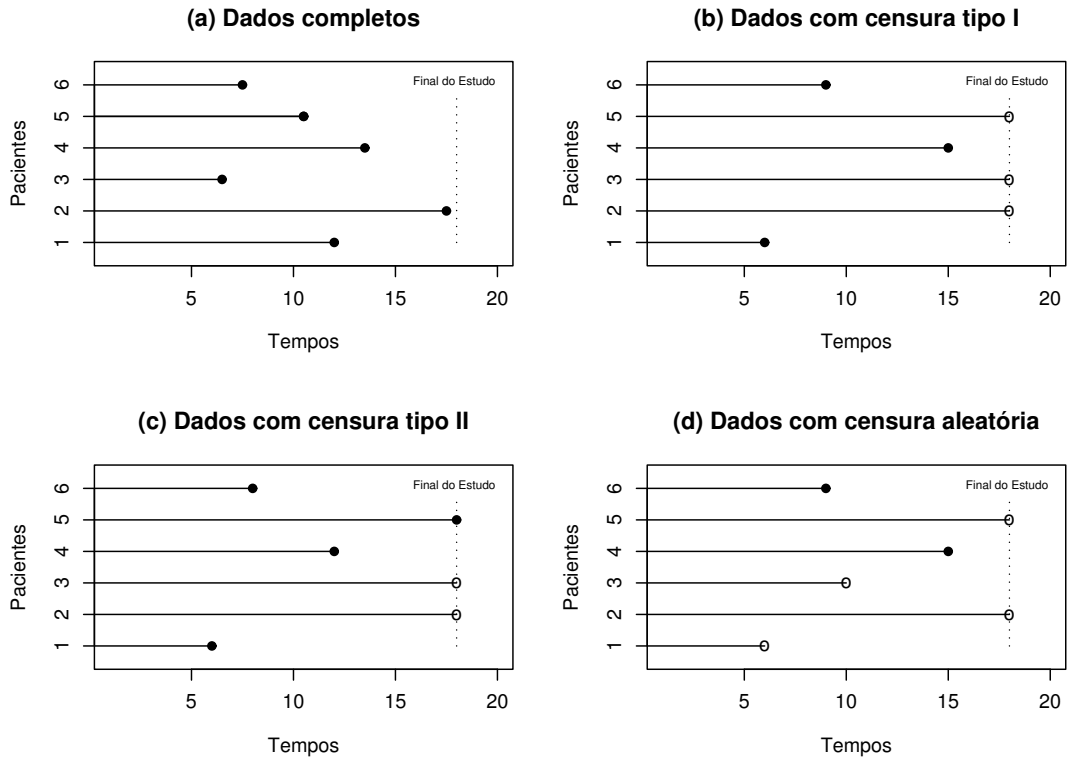

Figura 1.1: Ilustração de alguns mecanismos de censura em que • falha e ∘ censura. (a) todos os pacientes experimentaram o evento antes do final do estudo, (b) alguns pacientes não experimentaram o evento até o final do estudo, (c) o estudo foi finalizado após a ocorrência de um número pré-estabelecido de falhas e (d) o acompanhamento de alguns pacientes foi interrompido por alguma razão e alguns pacientes não experimentaram o evento até o final do estudo.

Os mecanismos de censura apresentados na Figura 1.1 são conhecidos por censura à direita, pois o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Esta é a situação freqüentemente encontrada em estudos envol-

vendo dados de sobrevivência. Os conjuntos de dados apresentados na Seção 1.5, com exceção daqueles nas Seções 1.5.6 e 1.5.7, são exemplos de censura à direita. No entanto, outras duas formas de censura podem ocorrer: censura à esquerda e intervalar.

A censura à esquerda ocorre quando o tempo registrado é maior que o tempo de falha. Isto é, o evento de interesse já aconteceu quando o indivíduo foi observado. Um estudo para determinar a idade em que as crianças aprendem a ler em uma determinada comunidade pode ilustrar a situação de censura à esquerda. Quando os pesquisadores começaram a pesquisa algumas crianças já sabiam ler e não lembravam com que idade isto tinha acontecido, caracterizando desta forma observações censuradas à esquerda. Neste mesmo estudo, pode ocorrer simultaneamente censura à direita para crianças que não sabiam ler quando os dados foram coletados. Os tempos de vida neste caso são chamados de duplamente censurados (Turnbull, 1974).

A intervalar é um tipo mais geral de censura que acontece, por exemplo, em estudos em que os pacientes são acompanhados em visitas periódicas e é conhecido somente que o evento de interesse ocorreu em um certo intervalo de tempo. Pelo fato de o tempo de falha T_i não ser conhecido exatamente, mas sim pertencer a um intervalo, isto é $T_i \in (L_i, U_i]$, tais dados são denominados *dados de sobrevivência intervalar*. A não ocorrência do evento é por sua vez denominada, nestes estudos, de *censura intervalar*. Lindsey et al. (1998) observam que tempos exatos de falha bem como censuras à direita e à esquerda, são casos especiais de dados de sobrevivência intervalar com $L_i = U_i$ para tempos exatos de falha, $U_i = \infty$ para censuras à direita e $L_i = 0$ para censuras à esquerda.

Uma outra característica de alguns estudos de sobrevivência é o truncamento que é muitas vezes confundido com censura. Truncamento é uma condição que exclui certos indivíduos do estudo. Em estudos com truncamento, somente indivíduos que experimentam algum evento são incluídos no estudo. Em estudos de AIDS, a data da infecção é uma origem de tempo bastante utilizada e o evento de interesse pode ser o desenvolvimento da AIDS. Neste caso, o número de pacientes infectados é desconhecido. Então, indivíduos já infectados e que ainda não desenvolveram a doença são desconhecidos para o pesquisador e não serão incluídos na amostra. Estas observações são chamadas de truncadas à direita. Outros exemplos de truncamento podem ser encontrados em Nelson (1990b), Kalbfleisch e Lawless (1992) e Klein e Moeschberger (1997).

A presença de censuras traz problemas para a análise estatística. A censura do tipo II é, em princípio, mais tratável que os outros tipos. Métodos exatos de

inferência estatística existem para a censura do tipo II, mas para situações bem simples que raramente acontecem em estudos clínicos. Na prática, faz-se uso de resultados assintóticos para realizar a análise estatística dos dados de sobrevivência, que não exigem o reconhecimento do mecanismo de censura. Isto significa que as mesmas técnicas estatísticas são utilizadas na análise de dados oriundos dos três mecanismos de censura.

Neste texto, atenção estará voltada aos dados de sobrevivência com censura à direita, que é a situação encontrada com mais freqüência em estudos tanto em medicina quanto em engenharia e ciências sociais. No entanto, comentários serão feitos ao longo do texto sobre o tratamento de censura à esquerda e truncamento. Um tratamento geral para dados censurados e truncados pode ser encontrado em Turnbull (1976). No caso particular de dados de sobrevivência com censura intervalar, algumas técnicas especializadas de análise serão apresentadas nos Capítulos 7 e 8. Desta forma, quando for simplesmente mencionada a palavra *censura* entenda-se, neste texto, censura à direita.

1.4 Representação dos Dados de Sobrevivência

Os dados de sobrevivência para o indivíduo i $(i = 1, \cdots, n)$ sob estudo, são representados, em geral, pelo par $(t_i,\, \delta_i)$ sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

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

Na presença de covariáveis medidas no i-ésimo indivíduo tais como, dentre outras, $\mathbf{x}_i$ = (sexo, idade, tratamento recebido), os dados ficam representados por $(t_i, \delta_i, \mathbf{x}_i)$. No caso especial de dados de sobrevivência intervalar tem-se, ainda, a representação $(\ell_i, u_i, \delta_i, \mathbf{x}_i)$ em que ℓ_i e u_i são, respectivamente, os limites inferior e superior do intervalo observado para o i-ésimo indivíduo.

1.5 Exemplos de Dados de Sobrevivência

Existem vários exemplos de aplicação dos modelos de análise de sobrevivência. Na área médica, eles são muito utilizados na identificação de fatores de prognóstico para uma doença, bem como na comparação de tratamentos. Em oncologia qualquer nova terapêutica ou droga para o combate ao câncer requer um estudo, em que a resposta

de interesse é geralmente o tempo de sobrevida dos pacientes, que é chamada de sobrevida global pelos oncologistas. Estudos epidemiológicos da AIDS são outros exemplos, em que as técnicas de análise de sobrevivência vêm sendo usadas com freqüência. Jacobson et al. (1993) mostram um estudo típico nesta área.

A seguir são descritos brevemente alguns dos exemplos, cujos dados encontram-se apresentados no Apêndice, que serão utilizados para ilustrar as técnicas estatísticas descritas no restante do texto.

1.5.1 Dados de Hepatite

Um estudo clínico aleatorizado foi realizado para investigar o efeito da terapia com esteróide no tratamento de hepatite viral aguda (Gregory et al., 1975). Vinte e nove pacientes com esta doença foram aleatorizados para receber um placebo ou o tratamento com esteróide. Cada paciente foi acompanhado por 16 semanas ou até a morte (evento de interesse) ou até a perda de acompanhamento. Os tempos de sobrevivência observados, em semanas, para os dois grupos encontram-se apresentados na Tabela 1.1 (+ indica censura).

Tabela 1.1: Tempos de sobrevivência observados no estudo de hepatite.

Grupo	Tempo de sobrevivência em semanas
Controle	1+, 2+, 3, 3, 3+, 5+, 5+, 16+, 16+, 16+, 16+, 16+, 16+, 16+, 16+
Esteróide	1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+

Este exemplo é utilizado no Capítulo 2 para ilustrar as técnicas não-paramétricas para dados de sobrevivência.

1.5.2 Dados de Leucemia Pediátrica

Este conjunto de dados foi obtido a partir de um estudo de crianças com leucemia, desenvolvido pelo Grupo Cooperativo Mineiro para Tratamento de Leucemias Agudas. Este é um estudo descritivo. A leucemia aguda é a neoplasia de maior incidência na população com menos de 15 anos de idade. Calcula-se que, nesta faixa etária, a incidência anual gire em torno de 5 a 6 casos novos por 100 mil crianças, sendo a grande maioria dos casos de Leucemia Linfoblástica Aguda (LLA).

Apesar do progresso alcançado no tratamento, em particular, da leucemia linfoblástica, as leucemias agudas continuam sendo a causa mais comum de morte por

neoplasia. O objetivo do tratamento médico de uma criança com LLA é obter longos períodos de sobrevida livre da doença, o que, muitas vezes, significa sua "cura". Os avanços terapêuticos obtidos nos últimos 25 anos têm sido grandes na LLA. Na década de 60, menos de 1% das crianças com LLA sobreviviam mais de 5 anos após o diagnóstico. Atualmente, com a intensificação da quimioterapia para os grupos com prognóstico mais desfavorável, 60 a 70% do total de crianças, com diagnóstico de LLA, são sobreviventes de longo prazo e encontram-se provavelmente "curadas". Nos grupos de melhor prognóstico, as proporções de "cura"já se situam no patamar de 90%.

Com o objetivo de entender melhor quais fatores afetam o tempo de sobrevivência de uma criança brasileira com LLA, um grupo de 128 crianças, com idade inferior a 15 anos, foi acompanhado no período de 1988 a 1992, em alguns hospitais de Belo Horizonte. A variável resposta de interesse é o tempo a partir da remissão (ausência da doença) até a recidiva ou morte (a que ocorrer primeiro). Das 128 crianças, 120 entraram em remissão e são elas que formam o conjunto de dados em estudo. Os fatores registrados para cada criança e que compõem o banco de dados são os seguintes: idade, peso, estatura, contagem de leucócitos, porcentagem de linfoblastos, porcentagem de vacúolos, fator de risco e indicador de sucesso da remissão. Informações adicionais sobre este estudo podem ser encontradas em Viana et al. (1994) e Colosimo et al. (1992).

1.5.3 Dados de Sinusite em Pacientes Infectados pelo HIV

O estudo da epidemia da AIDS é uma área de intensa pesquisa e vários trabalhos já estão registrados na literatura. A maioria deles tem como foco principal de atenção o tempo de vida de pacientes. Um estudo desenvolvido pela Profa. Denise Gonçalves do Departamento de Otorrinolaringologia da UFMG teve como interesse a ocorrência de manifestações otorrinolaringológicas em pacientes HIV positivos. O objetivo mais específico e que é explorado neste texto é verificar a hipótese de que a infecção pelo HIV aumenta a incidência de sinusite e que estas aumentam em freqüência com a progressão da imunodeficiência.

Neste estudo foram utilizadas informações provenientes de 91 pacientes HIV positivo e 21 HIV negativo, somando assim 112 pacientes estudados. Estes pacientes foram acompanhados no período compreendido entre março de 1993 até fevereiro de 1995. A classificação do paciente quanto à infecção pelo HIV seguiu os critérios do CDC (*Centers of Disease Control*, 1987). Os pacientes foram classificados como: HIV soronegativo (não possuem o HIV) ou HIV soropositivo assintomático (possuem

o vírus mas não desenvolveram o quadro clínico de AIDS) ou com ARC, *AIDS Related Complex*, (apresentam baixa imunidade e outros indicadores clínicos que antecedem o quadro clínico de AIDS) ou com AIDS (já desenvolveram infecções oportunistas que definem AIDS, segundo os critérios do CDC de 1987). Esta é a principal covariável a ser considerada no estudo. Ela é dependente do tempo, pois os pacientes mudam de classificação ao longo do estudo. Esta característica requer técnicas especializadas que são apresentadas no Capítulo 6. Outras covariáveis neste estudo, como contagem de CD4 e CD8, também são dependentes do tempo. No entanto, elas somente foram medidas no início do estudo.

A cada consulta, a classificação do paciente foi reavaliada. Cada paciente foi acompanhado através de consultas trimestrais. A freqüência mediana de consultas foi 4. A resposta de interesse foi o tempo, contado a partir da primeira consulta, até a ocorrência de sinusite. O objetivo foi identificar fatores de risco para cada uma destas manifestações. Os possíveis fatores de risco, incluídos no estudo, estão listados na Tabela 1.2.

Tabela 1.2: Covariáveis medidas no estudo de ocorrência de sinusite.

Idade do Paciente	Foi medida em anos
Sexo do Paciente	1 - Masculino
	2 - Feminino
Grupos de Risco	1 - Paciente HIV Soronegativo
	2 - Paciente HIV Soropositivo Assintomático
	3 - Paciente com ARC
	4 - Paciente com AIDS
CD4	Contagem de CD4
CD8	Contagem de CD8
Atividade Sexual	1 - Homossexual
	2 - Bissexual
	3 - Heterossexual
Uso de Droga	1 - Sim
Injetável	2 - Não
Uso de Cocaína	1 - Sim
por Aspiração	2 - Não

Para as covariáveis CD4 e CD8 foram registrados 41 valores perdidos, assim como nas covariáveis Atividade Sexual, Uso de Droga e Uso de Cocaína, em que

também foram registrados 23 valores perdidos.

Mais informações sobre este estudo podem ser encontradas em Gonçalves (1995) e Colosimo e Vieira (1996).

1.5.4 Dados de Aleitamento Materno

As Organizações Internacionais de Saúde recomendam o leite materno como a única fonte de alimentação para crianças entre 4 e 6 meses de idade. Identificar fatores determinantes do aleitamento materno em diferentes populações é, portanto, fundamental para alcançar tal recomendação.

Os Profs. Eugênio Goulart e Cláudia Lindgren do Departamento de Pediatria da UFMG realizaram um estudo no Centro de Saúde São Marcos, localizado em Belo Horizonte, com o objetivo principal de conhecer a prática do aleitamento materno de mães que utilizam este centro, assim como os possíveis fatores de risco ou de proteção para o desmame precoce. Um inquérito epidemiológico composto por questões demográficas e comportamentais foi aplicado a 150 mães de crianças menores de 2 anos de idade. A variável resposta de interesse foi o tempo máximo de aleitamento materno, ou seja, o tempo contado a partir do nascimento até o desmame completo da criança.

Uma análise estatística utilizando modelos paramétricos e semi-paramétricos é realizada nos Capítulos 4 e 5 para estes dados. Desta forma, pode-se comparar os resultados obtidos usando ambos os modelos.

1.5.5 Dados Experimentais utilizando Camundongos

Um estudo laboratorial foi realizado para investigar o efeito protetor do fungo *Saccharomycs boulardii* em ratos debilitados imunologicamente. O estudo utilizou 93 ratos provenientes do mesmo biotério. Inicialmente o sistema imunológico dos ratos eram debilitados quimicamente e a seguir, 4 tratamentos (controle e o fungo nas dosagens: 10mg, 1mg e 0.1mg) foram alocados aleatoriamente a cada animal. A resposta de interesse é o tempo de vida, medido em dias, após a aplicação do tratamento. O objetivo do estudo é comparar os tratamentos controlando pelo peso inicial do rato. Uma característica desses dados é a presença de empates. Existiam 61 tempos de censura e 13 tempos de falha distintos entre as 32 mortes observadas durante o período do estudo. A possibilidade de ajustar um modelo de regressão discreto para um conjunto de dados com vários empates é discutido no Capítulo 8.

1.5.6 Dados de Câncer de Mama

Um estudo retrospectivo foi realizado com 94 mulheres com diagnóstico precoce de câncer de mama com o objetivo de pesquisar duas terapias: (a) somente radioterapia e (b) radioterapia em conjunto com quimioterapia. Um total de 46 delas recebeu a primeira terapia e as demais receberam a segunda. As pacientes foram acompanhadas a cada 4-6 meses sendo, em cada visita, registrados uma medida da retração da mama (nenhuma, moderada ou severa) bem como o evento de interesse que, neste estudo, foi o tempo até o aparecimento de uma retração moderada ou severa da mama. Como as pacientes foram visitadas em alguns tempos aleatórios, o tempo exato de ocorrência da primeira retração da mama é conhecido somente ocorrer entre duas das visitas realizadas. Por outro lado, o que se sabe a respeito das pacientes que não apresentaram retração da mama até a última visita é que o evento não ocorreu até aquele momento e que, caso venha a ocorrer, será a partir daquele momento em diante. Este exemplo é analisado no Capítulo 7 em que é abordado a análise de dados de sobrevivência intervalar. Informações adicionais sobre este estudo podem ser encontradas em Klein e Moeschberger (1997).

1.5.7 Dados de Tempo de Vida de Mangueiras

Um ensaio, em delineamento em blocos ao acaso, foi conduzido no Departamento de Horticultura da ESALQ/USP no período de 1971 a 1992. O objetivo foi verificar a resistência das mangueiras a uma praga denominada seca da mangueira, que mata a planta. O interesse concreto era identificar novas mangueiras obtidas a partir de enxertos, resistentes à seca da mangueira. Um experimento fatorial completamente aleatorizado foi realizado com 6 copas enxertadas sobre 7 porta-enxertos (fatorial 6×7). Todas as 42 combinações foram replicadas em 5 diferentes blocos, totalizando 210 unidades experimentais. O estudo iniciou-se em 1971 e a resposta de interesse era o tempo de vida das mangueiras. O experimento foi visitado 12 vezes durante o período do estudo e registrada a condição de cada unidade experimental (viva ou morta). Os dados provenientes desse estudo são de natureza intervalar, ou seja, o evento de interesse (morte da mangueira) acontece entre duas visitas consecutivas e o tempo exato da morte é desconhecido. Este exemplo é analisado no Capítulo 8 que é dedicado a dados grupados. Mais informações sobre este estudo podem ser encontradas em Chalita et al. (1999).

1.6 Especificando o Tempo de Sobrevivência

A variável aleatória não-negativa T, que representa o tempo de falha, é usualmente especificada em análise de sobrevivência pela sua função de sobrevivência ou pela função de taxa de falha (ou risco). Estas duas funções, e funções relacionadas, que são extensivamente usadas na análise de dados de sobrevivência são apresentadas a seguir.

1.6.1 Função de Sobrevivência

Esta é uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência. A função de sobrevivência é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como

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

Em conseqüência, a função de distribuição acumulada, isto é, $F(t) = 1 - S(t)$, é definida como a probabilidade de uma observação não sobreviver ao tempo t.

Na Figura 1.2 pode ser observada a forma típica de duas funções de sobrevivência. Estas curvas, supostas representarem as funções de sobrevivência de dois grupos diferentes de pacientes, o grupo 1 tratado com a droga A e o grupo 2 com a droga B, fornecem informações importantes. Note, por exemplo, que o tempo de vida dos pacientes do grupo 1 é superior ao dos pacientes do grupo 2. Para os pacientes do grupo 1, o tempo para o qual cerca de 50% (tempo mediano) deles estarão mortos é de 20 anos, enquanto que para os pacientes do grupo 2, este tempo é menor (10 anos). Outra informação importante e possível de ser retirada desta figura é o percentual de pacientes que ainda estarão vivos até um determinado tempo de interesse. Por exemplo, para os pacientes do grupo 1, cerca de 90% deles ainda estarão vivos após 10 anos do início do estudo enquanto que para os do grupo 2, apenas 50%.

1.6.2 Função de Taxa de Falha ou de Risco

A probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como

$$S(t_1) - S(t_2).$$

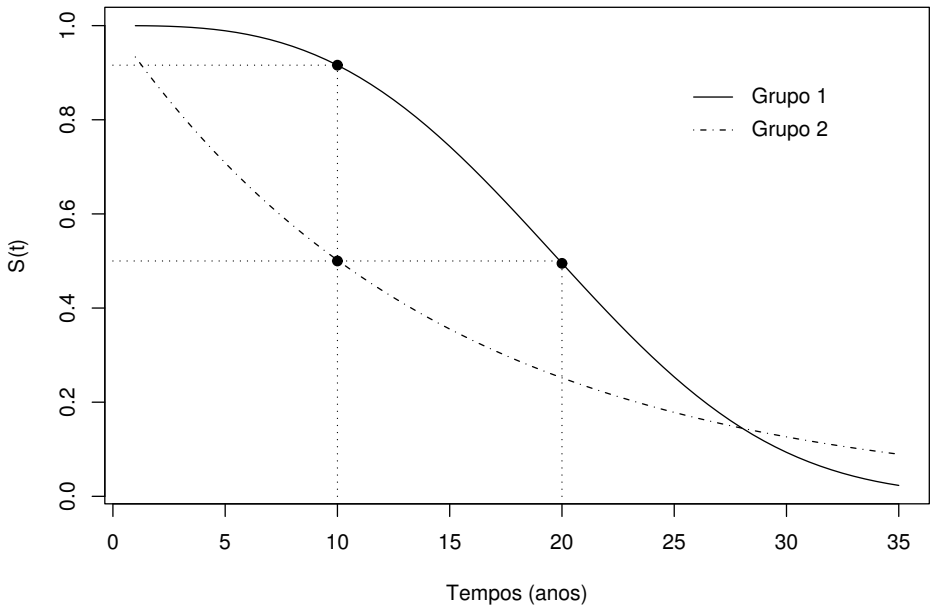

Figura 1.2: Funções de sobrevivência para dois grupos de pacientes.

A taxa de falha 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. Assim, a taxa de falha no intervalo $[t_1, t_2)$ é expressa por

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)\, S(t_1)}. \tag{1.1}$$

De uma forma geral, redefinindo o intervalo como $[t, t+\Delta t)$, a expressão (1.1) assume a seguinte forma

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

Assumindo Δt bem pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. Observe que as taxas de falha são números positivos, mas sem limite superior. A função de taxa de falha $\lambda(t)$ é bastante útil para descrever a distribuição do tempo de vida de pacientes. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo.

A função de taxa de falha de T é então definida como

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t}. \tag{1.2}$$

A Figura 1.3 mostra três funções de taxa de falha. A função crescente indica que a taxa de falha do paciente aumenta com o transcorrer do tempo. Este comportamento

mostra um efeito gradual de envelhecimento. A função constante indica que a taxa de falha não se altera com o passar do tempo. A função decrescente mostra que a taxa de falha diminui à medida que o tempo passa.

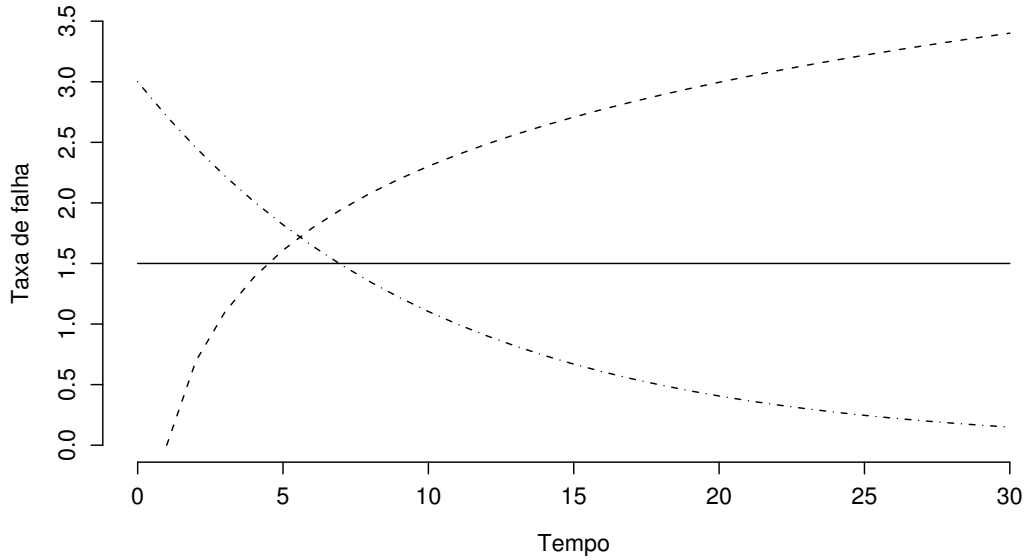

Figura 1.3: Funções de taxa de falha - - crescente, — constante e -·-· decrescente.

Sabe-se, ainda, que a taxa de falha para o tempo de vida de seres humanos é uma combinação das curvas apresentadas na Figura 1.3 em diferentes períodos de tempo. Ela é conhecida como "curva da banheira"e tem uma taxa de falha decrescente no período inicial, representando a mortalidade infantil, constante na faixa intermediária e crescente na porção final. Uma representação desta curva é mostrada na Figura 1.4.

1.6.3 Função de Taxa de Falha Acumulada

Outra função útil em análise de dados de sobrevivência é a função de risco acumulada que, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo. Esta função é definida por:

$$\Lambda(t) = \int_0^t \lambda(u)\, du.$$

A função de taxa de falha acumulada, $\Lambda(t)$, não tem uma interpretação direta mas pode ser útil na avaliação da função de maior interesse que é a de taxa de falha, $\lambda(t)$. Isto acontece essencialmente na estimação não-paramétrica em que $\Lambda(t)$ apresenta um estimador com propriedades ótimas e $\lambda(t)$ é difícil de ser estimada.

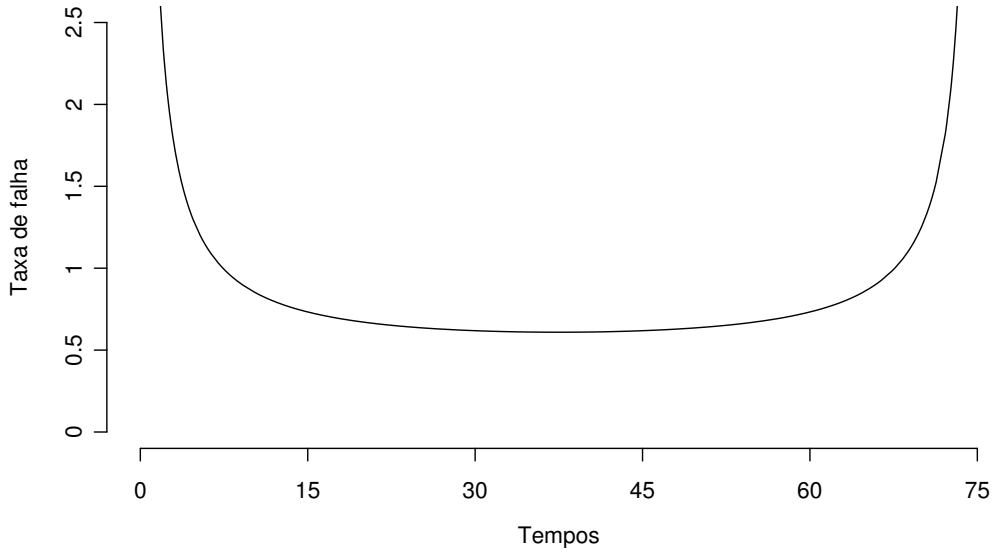

Figura 1.4: Representação da função de taxa de falha conhecida como *curva da banheira*.

1.6.4 Tempo Médio e Vida Média Residual

Outras duas quantidades de interesse em análise de sobrevivência são: o tempo médio de vida e a vida média residual. A primeira é obtida pela área sob a função de sobrevivência. Isto é,

$$t_m = \int_0^\infty S(t)\, dt.$$

Já a vida média residual é definida condicional a um certo tempo de vida t. Ou seja, para indivíduos com idade t esta quantidade mede o tempo médio restante de vida e é, então, a área sob a curva de sobrevivência à direita do tempo t dividida por $S(t)$. Isto é,

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

sendo $f(\cdot)$ a função de densidade de T.

1.6.5 Relações entre as Funções

Em termos das funções definidas anteriormente, tem-se algumas relações matemáticas importantes entre elas, a saber:

$$\lambda(t) \;=\; \frac{f(t)}{S(t)} = -\frac{d}{dt}\Big(\log S(t)\Big),$$

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\log S(t)$$

e

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u)\, du\right\}.$$

Tais relações mostram que o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais, isto é, de $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$.

Outras relações envolvendo estas funções são as seguintes:

$$S(t) = \frac{\text{vmr}(0)}{\text{vmr}(t)}\ \exp\left\{-\int_0^t \frac{du}{\text{vmr}(u)}\right\}$$

e

$$\lambda(t) = \left(\frac{d\,\text{vmr}(t)}{dt} + 1\right)/\text{vmr}(t).$$

1.7 Exercícios

1. Um grande número de indivíduos foi acompanhado para estudar o aparecimento de um certo sintoma. Os indivíduos foram incluídos ao longo do estudo e foi considerado como resposta de interesse a idade em que este sintoma apareceu pela primeira vez. Para os seis indivíduos selecionados e descritos a seguir, identifique o tipo de censura apresentado.

   (a) O primeiro indivíduo entrou no estudo com 25 anos já apresentando o sintoma.

   (b) Outros dois indivíduos entraram no estudo com 20 e 28 anos e não apresentaram o sintoma até o encerramento do estudo.

   (c) Outros dois indivíduos entraram com 35 e 40 anos e apresentaram o sintoma no segundo e no sexto exames respectivamente, após terem entrado no estudo. Os exames foram realizados a cada dois anos.

   (d) O último indivíduo selecionado entrou no estudo com 36 anos e mudou da cidade depois de 4 anos sem ter apresentado o sintoma.

2. Mostre que $\lambda(t) = \dfrac{f(t)}{S(t)} = -\dfrac{d}{dt}\Big(\log S(t)\Big)$.

3. Mostre que $\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$. (Dica: utilize o exercício 2).

4. Mostre que $S(t) = \dfrac{\text{vmr}(0)}{\text{vmr}(t)} \exp\left\{-\displaystyle\int_0^t \frac{du}{\text{vmr}(u)}\right\}$.
   $\Big($Dica: utilize uma integral por partes sabendo que $f(u)du = -\frac{d}{du}S(u)\Big)$.

5. Suponha que a taxa de falha da variável aleatória tempo de falha T seja expressa pela função linear $\lambda(t) = \beta_0 + \beta_1 t$, com β_0 e $\beta_1 > 0$. Obtenha $S(t)$ e $f(t)$.

6. Suponha que a vida média residual de T seja dada por $\text{vmr}(t) = t + 10$. Obtenha $E(T)$, $\lambda(t)$ e $S(t)$.

Capítulo 2

Técnicas Não-Paramétricas em Análise de Sobrevivência

2.1 Introdução

Os objetivos de uma análise estatística envolvendo dados de sobrevivência geralmente estão relacionados, em medicina, à identificação de fatores de prognóstico para uma certa doença ou à comparação de tratamentos em um estudo clínico enquanto controlado por outros fatores. Vários exemplos podem ser encontrados na literatura médica. No estudo da leucemia pediátrica, por exemplo, apresentado na Seção 1.5.2, leucometria (contagem de células brancas) ao diagnóstico e idade são conhecidos fatores de prognóstico para o tempo de vida de crianças com leucemia.

Por mais complexo que seja o estudo, as respostas às perguntas de interesse são dadas a partir de um conjunto de dados de sobrevivência, e o passo inicial de qualquer análise estatística consiste em uma descrição dos dados. A presença de observações censuradas é, contudo, um problema para as técnicas convencionais de análise descritiva, envolvendo média, desvio-padrão e técnicas gráficas como histograma, box-plot, entre outras. Os problemas gerados por observações censuradas podem ser ilustrados numa situação bem simples em que se tenha interesse na construção de um histograma. Se a amostra não contiver observações censuradas, a construção do histograma consiste na divisão do eixo do tempo em um certo número de intervalos e, em seguida, conta-se o número de ocorrências de falhas em cada intervalo. Entretanto, quando existem censuras, não é possível construir um histograma, pois não se conhece a freqüência exata associada a cada intervalo.

Entretanto, algumas técnicas usuais podem ser utilizadas com o devido cuidado.

Por exemplo, em uma análise descritiva inicial dos dados é comum o exame do gráfico de cada covariável contínua com a resposta. Este gráfico nos possibilita avaliar através da nuvem de pontos uma possível relação linear entre elas ou a adequação de um modelo proposto. A presença de observações censuradas gera dificuldades na interpretação deste gráfico, mas com um certo cuidado continua gerando informações descritivas sobre a relação entre as variáveis. A Figura 2.1 apresenta um gráfico envolvendo o tempo entre a remissão e a recidiva (tempo de sobrevivência em anos) e a raiz quadrada da leucometria ao diagnóstico (contagem de células brancas iniciais) para os dados de leucemia pediátrica apresentados na Seção 1.5.2. A transformação raiz quadrada é usual em covariáveis como esta que apresentam uma escala de medida muito ampla. Cox e Snell (1981, p. 148) apresentam uma transformação similar em uma situação envolvendo esta mesma covariável. Os símbolos diferentes na Figura 2.1 são utilizados para diferenciar falha e censura.

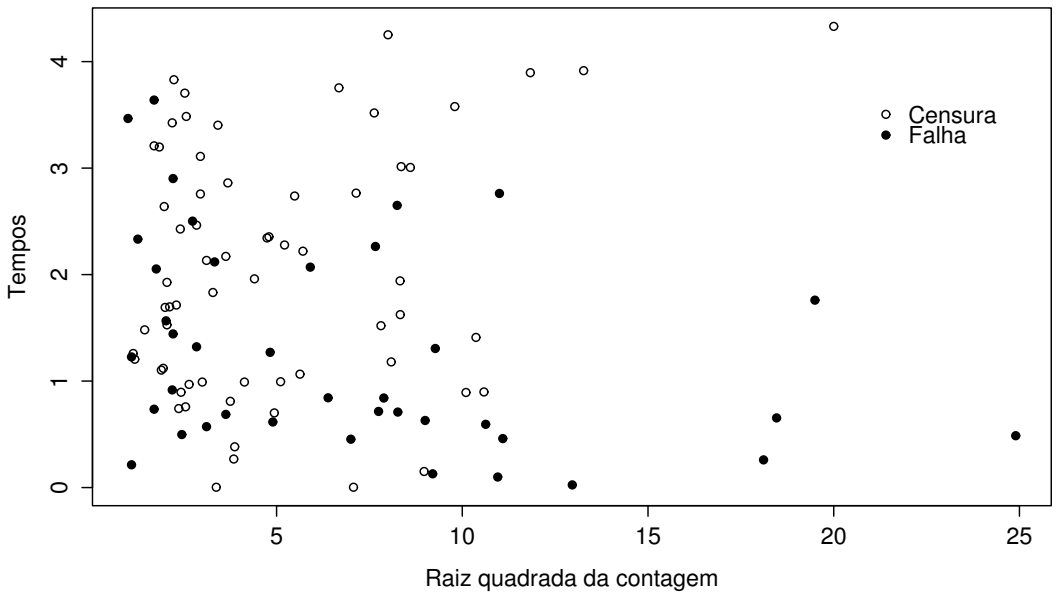

Figura 2.1: Gráfico de dispersão do tempo de sobrevivência versus a raiz quadrada da contagem inicial de leucócitos para os dados de leucemia pediátrica.

A natureza da associação entre a leucometria e o tempo de sobrevivência pode ser visualizada no gráfico apresentado na Figura 2.1. A nuvem de pontos (falhas) é densa para tempos de sobrevivência curtos e os pontos vão lentamente diminuindo para os tempos maiores, limitada pelo valor de $4,33$ anos que foi o paciente com o maior tempo de acompanhamento. A forma do gráfico é controlada pela associação entre a leucometria e o tempo de sobrevivência e pela informação de que a distribuição desta

última tende a ser assimétrica à direita. A leucometria tem uma associação negativa com o tempo de sobrevivência. Ou seja, os tempos são menores para os valores mais altos de leucometria. Se todos os sujeitos são acompanhados pelo mesmo período de tempo, tem-se então mais observações censuradas entre os pacientes com contagem baixa do que entre aqueles com contagem alta. Entretanto, se a entrada dos pacientes for uniforme durante o período de estudo e independente da leucometria, espera-se uma proporção igual de observações censuradas em todos os valores de leucometria. O exemplo mostrado na Figura 2.1 vem de um estudo deste tipo e a figura indica que as observações censuradas e as não-censuradas estão misturadas para todos os valores de leucometria.

Nos textos básicos de estatística, uma análise descritiva consiste essencialmente em encontrar medidas de tendência central e variabilidade. Como a presença de censuras invalida este tipo de tratamento aos dados de sobrevivência, o principal componente da análise descritiva envolvendo dados de tempo de vida é a função de sobrevivência. Nesta situação, o procedimento inicial é encontrar uma estimativa para esta função de sobrevivência e então, a partir dela, estimar as estatísticas de interesse que usualmente são o tempo médio ou mediano, alguns percentis ou certas frações de falhas em tempos fixos de acompanhamento. Nas Seções 2.2 a 2.4 serão apresentados alguns estimadores não-paramétricos para a função de sobrevivência, dentre eles o conhecido estimador de Kaplan-Meier. Algumas quantidades de interesse, como a mediana e a média, são obtidas a partir desta função e estão apresentadas na Seção 2.5. A Seção 2.6 finaliza o capítulo apresentado os testes não-paramétricos para a comparação de duas ou mais funções de sobrevivência.

2.2 Estimação na Ausência de Censura

Nesta seção será tratado a estimação da função de sobrevivência em uma situação sem censura. A função de taxa de falha também será estimada, mas existem dificuldades em obter tal estimativa na presença de censura. Na realidade, existem dificuldades em estimar esta função em termos não-paramétricos mesmo na ausência de censuras. A Figura 2.2 apresenta um histograma que mostra a distribuição do tempo de falha associado a um certo grupo de indivíduos. Este histograma foi obtido a partir de uma amostra de 54 observações não-censuradas.

Uma estimativa para a taxa de falha no período compreendido entre 400 e 500

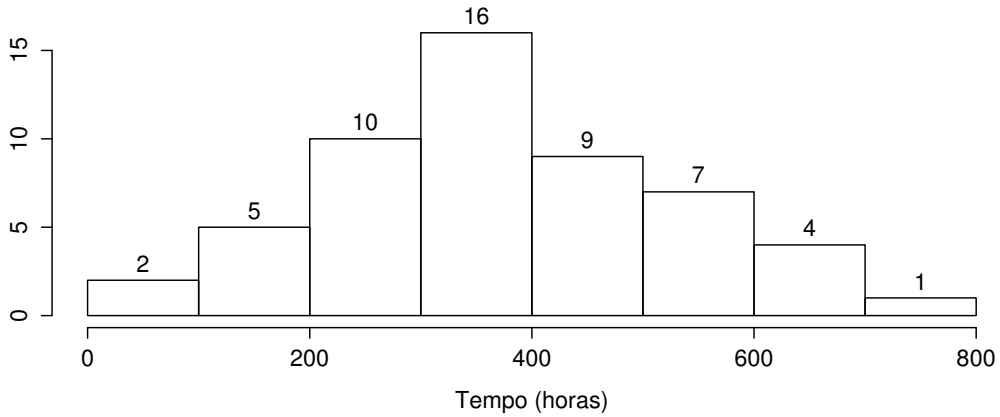

Figura 2.2: Histograma mostrando a distribuição do tempo de falha associado a um certo grupo de indivíduos.

horas é dada por

$$\widehat{\lambda}([400,500)) = \frac{\text{número de falhas no período } [400,500)}{\text{número que não falharam até } t = 400} = \frac{9}{21} = 0,429. \qquad (2.1)$$

Em palavras, a taxa de falha é de $42,9\%$ durante o período de 100 horas, compreendido entre 400 e 500 horas a partir do início do estudo. Isto significa que, entre 100 indivíduos que sobreviveram até 400 horas, espera-se que 57 sobrevivam mais 100 horas. A taxa de falha pode também ser expressa como $42,9\%/100$ horas ou $0,429\%/$hora. Usando o mesmo tipo de cálculo para os outros intervalos de tempo, obtêm-se os resultados mostrados na Tabela 2.1. Desta tabela, pode-se notar que a taxa de falha é do tipo crescente.

A função de sobrevivência no tempo $t = 400$ horas é estimada por:

$$\begin{aligned} \widehat{S}(400) &= \frac{\text{no. de indivíduos que não falharam até o tempo } t = 400}{\text{número de indivíduos no estudo}} \\ &= \frac{21}{54} = 0,389. \end{aligned}$$

Em palavras, este número significa que 39% destes indivíduos sobrevivem mais do que 400 horas. Repetindo o mesmo tipo de cálculo para cada tempo de falha, obtiveram-se os resultados mostrados na Tabela 2.1. A partir destes valores pode-se obter informações importantes sobre o tempo de vida dos indivíduos em estudo.

A forma utilizada para calcular as estimativas para as taxas de falhas foi bastante intuitiva. Uma forma alternativa é usar a expressão dada em (1.1) que apresenta a taxa de falha em termos da função de sobrevivência. Assim tem-se

Tabela 2.1: Estimativas das funções de sobrevivência e de taxa de falha para os dados do histograma mostrado na Figura 2.2.

Intervalo	Taxa de Falha (%/hora)	Sobrevivência (%)
0-100	0,037	100,0
100-200	0,096	96,3
200-300	0,213	87,0
300-400	0,432	68,5
400-500	0,429	38,9
500-600	0,583	22,2
600-700	0,800	09,3
700-800	-	01,9

$$\widehat{\lambda}([400,500)) = \frac{\widehat{S}(400) - \widehat{S}(500)}{(500-400)\,\widehat{S}(400)} = \frac{0,389 - 0,222}{(100)\,0,389} = 0,0043/\text{hora}$$

ou 0,43%/hora.

É importante observar que as taxas de falha estimadas neste exemplo e apresentadas na Tabela 2.1 foram para os intervalos definidos na Figura 2.2. Desta forma, estas taxas não são instantâneas como prescrito na definição (1.2) de $\lambda(t)$. A partir do banco de dados com os valores reais, existem propostas de estimadores para estimar $\lambda(t)$ (Klein e Moeschberger, 1997).

2.3 O Estimador de Kaplan-Meier

No exemplo da Seção 2.2, foram estimadas as funções de sobrevivência e de taxa de falha para um estudo em que todas as observações falharam, ou seja, não existiram censuras. Na prática, entretanto, o conjunto de dados amostrais de tempos de falha apresenta censuras, o que requer técnicas estatísticas especializadas para acomodar a informação contida nestas observações. A observação censurada informa que o tempo até a falha é maior do que aquele que foi registrado.

Nesta seção será apresentado o conhecido estimador de Kaplan-Meier para a função de sobrevivência que é, sem dúvida, o mais utilizado em estudos clínicos e vem ganhando cada vez mais espaço em estudos de confiabilidade. O estimador

conhecido por Nelson-Aalen, proposto por Nelson (1972) e suas propriedades estudadas por Aalen (1978) será apresentado na seção 2.4.1. Este estimador e o de Kaplan-Meier apresentam essencialmente as mesmas características. Um terceiro estimador, o estimador da tabela de vida ou atuarial, por ser uma das mais antigas técnicas estatísticas utilizadas para estimar características associadas à distribuição dos tempos de falha, também será apresentado na Seção 2.4.2.

O estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é também chamado de estimador limite-produto. Ele é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{\text{no. de observações que não falharam até o tempo } t}{\text{no. total de observações no estudo}}. \tag{2.2}$$

$\widehat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se exitirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

O estimador de Kaplan-Meier, na sua construção, considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra. A seguir é apresentado a idéia intuitiva deste estimador para depois mostrar a sua expressão geral, assim como foi proposto por seus autores.

Considere os tempos de sobrevivência do grupo esteróide dos dados de hepatite apresentados na Seção 1.5.1 e reproduzidos na Tabela 2.2. O procedimento para obter a estimativa de Kaplan-Meier envolve uma sequência de passos, em que o próximo depende do anterior. Isto significa, por exemplo, que

$$S(5) = P(T \geq 5) = P(T \geq 1)P(T \geq 5 | \, T \geq 1).$$

Desta forma, para o indivíduo sobreviver a 5 semanas, ele vai precisar sobreviver, em um primeiro passo, à primeira semana e depois sobreviver à quinta semana, sabendo que ele sobreviveu à primeira. Os tempos de 1 e 5 semanas foram tomados por serem os dois primeiros tempos distintos de falha nos dados do grupo esteróide. Os passos são gerados a partir de intervalos definidos pela ordenação dos tempos de falha de forma que cada um deles começa em um tempo observado e termina no próximo tempo. A Tabela 2.2 apresenta os tempos ordenados indicando que se tem 6 intervalos, iniciando com $[0, 1)$, até o sexto intervalo que é $[10, 16)$. O limite superior deste último intervalo é definido como sendo 16 por ser este o maior tempo de acompanhamento do estudo.

Tabela 2.2: Estimativas de Kaplan-Meier para o grupo esteróide.

t_j	Intervalos	d_j	n_j	$\widehat{S}(t_j)$
0	[0,1)	0	14	1,000
1	[1;5)	3	14	0,786
5	[5;7)	1	9	0,698
7	[7;8)	1	8	0,611
8	[8;10)	1	7	0,524
10	[10,16)	1	6	0,437

Todos os indivíduos estavam vivos em $t = 0$ e se mantêm até a primeira morte que ocorre em $t = 1$ semana. Então a estimativa de $S(t)$ deve ser 1 neste intervalo compreendido entre 0 e 1 semana. No valor correspondente a 1 semana, a estimativa deve cair devido a três mortes que ocorrem neste tempo. No segundo intervalo, $[1, 5)$, existem então 14 indivíduos que estavam vivos (sob risco) antes de $t = 1$ e 3 morrem. Desta forma, a estimativa da probabilidade condicional de morte neste intervalo é $3/14$ e a probabilidade de sobreviver é $1 - 3/14$. Isto pode ser escrito como,

$$\widehat{S}(1) = \widehat{P}(T \geq 0)\widehat{P}(T \geq 1 | \, T \geq 0) = (1)(11/14) = 0,786.$$

Assim sucessivamente, para qualquer t, $S(t)$ pode ser escrito em termos de probabilidades condicionais. Suponha que existem n pacientes no estudo e $k(\leq n)$ falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$. Considerando $S(t)$ como uma função discreta com probabilidade maior que zero somente nos tempos de falha $t_j, j = 1, \ldots, k$, tem-se que

$$S(t_j) = (1 - q_1)(1 - q_2) \ldots (1 - q_j), \tag{2.3}$$

em que q_j é a probabilidade de um indivíduo morrer no intervalo $[t_{j-1}, t_j)$ sabendo que ele não morreu até t_{j-1} e considerando $t_0 = 0$. Ou seja, pode-se escrever q_j como

$$q_j = P(T \in [t_{j-1}, t_j) | \, T \geq t_{j-1}). \tag{2.4}$$

Desta forma pode-se escrever a expressão geral de $S(t)$ em termos de probabilidades condicionais. O estimador de Kaplan-Meier se reduz então, a estimar q_j que adaptado da expressão (2.2) é dado por,

$$\widehat{q}_j = \frac{\text{no. de falhas em } t_j}{\text{no. de observações sob risco em } t_{j-1}}, \tag{2.5}$$

para $j = 1, \ldots, k$.

A expressão geral do estimador de Kaplan-Meier pode então ser apresentada após estas considerações preliminares. Formalmente, considere:

- $t_1 < t_2 \cdots < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em t_j, $j = 1, \cdots, k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é, então, definido como:

$$\hat{S}(t) = \prod_{j\,:\,t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j\,:\,t_j < t} \left(1 - \frac{d_j}{n_j} \right). \tag{2.6}$$

Uma justificativa simples para a expressão (2.6) do estimador de Kaplan-Meier vem da decomposição de $S(t)$ em termos dos q_j's apresentada em (2.3). O estimador de Kaplan-Meier é obtido a partir de (2.6) se os q_j's forem estimados por d_j/n_j que foi expresso em palavras em (2.5). No artigo original, Kaplan e Meier justificam a expressão (2.6) mostrando que ela é o estimador de máxima verossimilhança de $S(t)$. Os principais passos desta prova são indicados a seguir. Suponha, como feito anteriormente, que d_j observações falham no tempo t_j, para $j = 1, \ldots, k$, e m_j observações são censuradas no intervalo $[t_j, t_{j+1})$, nos tempos $t_{j1}, \ldots, t_{jm_j}$. A probabilidade de falha no tempo t_j é então

$$S(t_j) - S(t_j + 0),$$

com $S(t_j + 0) = \lim_{\Delta t \to 0+} S(t_j + \Delta t)$, $j = 1, \ldots, k$. Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em $t_{j\ell}$, $\ell = 1, \ldots, m_j$, é

$$P(T > t_{j\ell}) = S(t_{j\ell} + 0).$$

A função de verossimilhança pode então ser escrita como

$$L(S(\cdot)) = \prod_{j=1}^{k} \left\{ \Big[S(t_j) - S(t_j + 0) \Big]^{d_j} \prod_{\ell=1}^{m_j} S(t_{j\ell} + 0) \right\}.$$

Pode-se mostrar que $S(t)$ que maximiza $L(S(\cdot))$ é exatamente a expressão (2.6). Esta definição do estimador de máxima verossimilhança é uma generalização do conceito usual utilizado em modelos paramétricos em que se tem tantos parâmetros quanto falhas distintas. Entretanto, o resultado de problemas como este, em que

muitos parâmetros estão envolvidos, deve ser tratado com cuidado. Detalhes desta prova são encontrados em Kalbfleisch e Prentice (1980).

Naturalmente, o estimador de Kaplan-Meier se reduz à função de sobrevivência empírica (2.2) 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 $\widehat{S}(t) = 0$, pois as últimas observações são censuradas.

A Tabela 2.2 mostra os cálculos das estimativas de Kaplan-Meier para a função de sobrevivência do grupo esteróide dos dados de hepatite. Assim, a estimativa de $S(5)$ usando a expressão (2.6) fica

$$\widehat{S}(5) = (1 - 3/14)(1 - 1/9) = 0,698.$$

Observe que $\widehat{S}(6)$ é também igual a $0,698$, pois $\widehat{S}(t)$ é uma função escada com saltos somente nos tempos de falha.

A partir dos valores obtidos na Tabela 2.2, o mais prático é fazer um gráfico destes resultados, através do qual é possível responder a possíveis perguntas de interesse. O gráfico é construído mantendo o valor de $\widehat{S}(t)$ constante entre os tempos de falhas. A forma gráfica do estimador de Kaplan-Meier é apresentada na Figura 2.3. Neste gráfico, também é mostrada a estimativa de Kaplan-Meier para o grupo controle que é de simples cálculo, pois só apresenta um tempo distinto de falha. Observe que este gráfico não atinge o valor $\widehat{S}(t) = 0$. Como foi dito, isto sempre acontecerá quando o maior tempo observado na amostra corresponder a uma censura.

As estimativas para o grupo esteróide apresentadas na Tabela 2.2, bem como as estimativas para o grupo controle e suas respectivas representações gráficas apresentadas na Figura 2.3, podem ser obtidas no pacote estatístico R por meio dos comandos apresentados a seguir.

```
> require(survival)
> tempos<- c(1,2,3,3,3,5,5,16,16,16,16,16,16,16,16,1,1,1,1,4,5,7,8,10,10,12,16,16,16)
> cens<-c(0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1,1,0,0,0,0,0)
> grupos<-c(rep(1,15),rep(2,14))
> ekm<- survfit(Surv(tempos,cens)~grupos)
> summary(ekm)
> plot(ekm, lty=c(2,1), xlab="Tempo (semanas)",ylab="S(t) estimada")
> legend(1,0.3,lty=c(2,1),c("Controle","Esteróide"),lwd=1, bty="n")
```

Quanto as principais propriedades do estimador de Kaplan-Meier, estas consistem basicamente em:

i) é não-viciado para amostras grandes,

ii) é fracamente consistente,

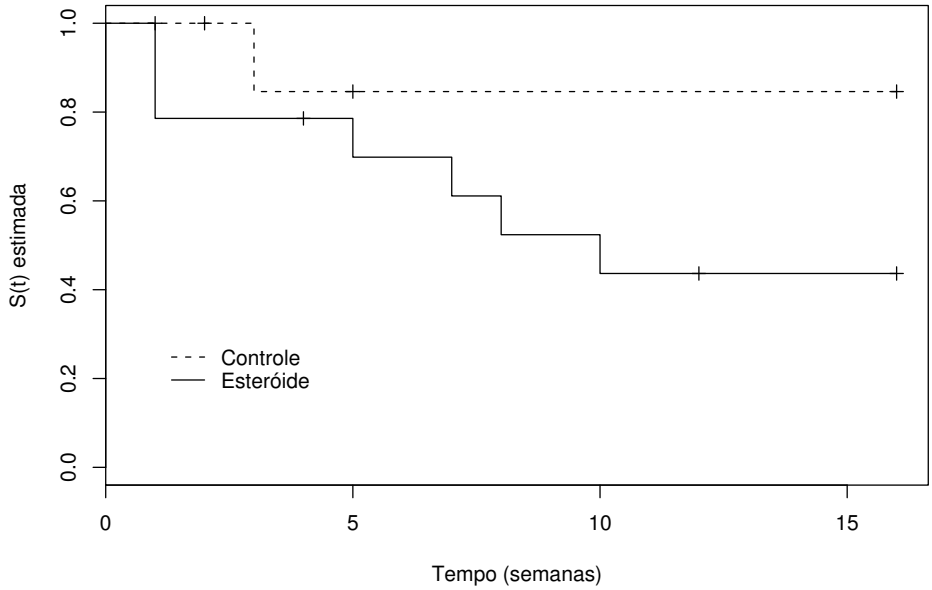

Figura 2.3: Estimativas de Kaplan-Meier para os grupos correspondentes aos dados de hepatite apresentados na Seção 1.5.1.

iii) converge assintoticamente para um processo gaussiano e

iv) é estimador de máxima verossimilhança de $S(t)$.

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

No entanto, para construir intervalos de confiança e testar hipóteses para $S(t)$, faz-se necessário avaliar a precisão do estimador de Kaplan-Meier. Este estimador, assim como outros, está sujeito à variações que devem ser descritas em termos de estimações intervalares. A expressão para a variância assintótica do estimador de Kaplan-Meier é dada por:

$$\widehat{Var}(\widehat{S}(t)) = \left[\widehat{S}(t)\right]^2 \sum_{j:\, t_j < t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.7}$$

Esta expressão é conhecida como fórmula de Greenwood, e pode ser obtida a partir da expressão (2.6). Isto é feito em Kalbfleisch e Prentice (1980, p. 12-14). A estimativa da variância de $\widehat{S}(6)$, para o exemplo considerado, é então, dada por

$$\widehat{Var}(\widehat{S}(6)) = (0,698)^2 \left[\frac{3}{14.11} + \frac{1}{9.8}\right] = 0,0163.$$

Como $\widehat{S}(t)$, para t fixo, tem distribuição assintótica Normal, segue que, um intervalo de 100(1- α)% de confiança para $S(t)$ é dado por:

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

em que $\alpha/2$ denota o $\alpha/2$-percentil superior da distribuição Normal padrão. O intervalo de 95% de confiança para $S(6)$ é $0,698 \pm \sqrt{0,0163}$ ou seja, $(0,45; 0,95)$.

Entretanto, para valores extremos de t, este intervalo de confiança pode apresentar limite inferior negativo ou limite superior maior do que 1. Nesses casos, o problema é resolvido utlizando uma transformação para $S(t)$ como, por exemplo, $\widehat{U}(t) = \log[-\log(\widehat{S}(t))]$ que têm variância assintótica estimada por

$$\widehat{Var}(\widehat{U}(t)) = \frac{\displaystyle\sum_{j\,:\,t_j<t} \frac{d_j}{n_j(n_j-d_j)}}{\left[\displaystyle\sum_{j\,:\,t_j<t} \log\Big(\frac{n_j-d_j}{n_j}\Big)\right]^2}.$$

Assim, um intervalo aproximado de 100(1- α)% de confiança para $S(t)$ é dado por:

$$\widehat{S}(t)^{\exp\left\{\pm\, z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{U}(t))}\right\}} \tag{2.8}$$

que assume valores no intervalo [0,1] e resulta no intervalo $(0,38; 0,88)$ de 95% de confiança para $S(6)$ no exemplo dos dados de esteróide.

2.4 Outros Estimadores Não-Paramétricos

Como foi dito anteriormente, o estimador de Kaplan-Meier é sem dúvida o mais utilizado para estimar $S(t)$ em análise de sobrevivência. Existem muitos pacotes estatísticos que calculam este estimador e ele também é apresentado em vários textos de estatística básica. Entretanto, outros dois estimadores de $S(t)$ têm importância na literatura mais especializada desta área. Eles são: o estimador de Nelson-Aalen e o estimador da tabela de vida. O primeiro, o de Nelson-Aalen, é mais recente que o de Kaplan-Meier e apresenta, aparentemente, propriedades similares ao deste último. O segundo 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 utilizado basicamente em grandes amostras.

2.4.1 Estimador de Nelson-Aalen

Este estimador, como mencionado anteriormente, é mais recente do que o de Kaplan-Meier e baseia-se na função de sobrevivência expressa por:

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

em que $\Lambda(t)$ é a função de risco acumulada definida na Seção 1.6.3.

Um estimador para $\Lambda(t)$ foi inicialmente proposto por Nelson (1972) e retomado por Aalen (1978) que provou suas propriedades assintóticas usando processos de contagem. Este estimador é denominado na literatura por Nelson-Aalen e tem a seguinte forma,

$$\tilde{\Lambda}(t) = \sum_{j:\, t_j < t} \left(\frac{d_j}{n_j}\right), \tag{2.9}$$

em que d_j e n_j são definidos como no estimador de Kaplan-Meier. A variância deste estimador, proposta por Aalen (1978b), é dada por

$$\widehat{V}ar(\tilde{\Lambda}(t)) \;=\; \sum_{t_j < t} \left(\frac{d_j}{n_j^2}\right). \tag{2.10}$$

Um estimador alternativo para a variância de $\tilde{\Lambda}(t)$ proposto por Klein (1991) é

$$\widehat{V}ar(\tilde{\Lambda}(t)) \;=\; \sum_{t_j < t} \frac{(n_j - d_j)\, d_j}{n_j^3},$$

mas, por apresentar menor vício, o estimador (2.10) é preferível a este último.

Desse modo, e com base no estimador de Nelson-Aalen, um estimador para a função de sobrevivência é expresso por

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

A variância deste estimador, devido a Aalen e Johansen (1978), pode ser obtida por

$$\widehat{V}ar(\tilde{S}(t)) \;=\; \Big[\tilde{S}(t)\Big]^2 \sum_{t_j < t} \left(\frac{d_j}{n_j^2}\right)$$

ou, alternativamente, substituindo-se $\widehat{S}(t)$ por $\tilde{S}(t)$ na expressão (2.7).

O estimador $\tilde{S}(t)$ e o de Kaplan-Meier apresentam na maioria das vezes estimativas muito próximas para $S(t)$. Bohoris (1994) mostrou que $\tilde{S}(t) \geq \widehat{S}(t)$ para todo t. Ou seja, o estimador de Nelson-Aalen é maior ou igual que o estimador de

Kaplan-Meier. A Tabela 2.3 apresenta as estimativas de Nelson-Aalen para o grupo esteróide dos dados de hepatite. Observe que estas estimativas são bem próximas das de Kaplan-Meier mostradas na Tabela 2.2, mesmo neste caso em que a amostra é relativamente pequena.

Tabela 2.3: Estimativas de Nelson-Aalen para o grupo esteróide.

t_j	n_j	d_j	$\tilde{\Lambda}(t_j)$	$\tilde{S}(t_j)$	e.p.$(\tilde{S}(t_j))$	I.C.$(\tilde{S}(t_j))_{95\%}$
0	14	0	0	1		
1	14	3	0,214	0,807	0,0999	(0,633; 1,000)
5	9	1	0,325	0,722	0,1201	(0,521; 1,000)
7	8	1	0,450	0,637	0,1326	(0,424; 0,958)
8	7	1	0,593	0,553	0,1394	(0,337; 0,906)
10	6	1	0,760	0,468	0,1414	(0,259; 0,846)

O estimador de Kaplan-Meier tem a vantagem de estar disponível em vários pacotes estatísticos, o que não acontece em geral com o de Nelson-Aalen. No pacote estatístico R, por exemplo, as estimativas de Nelson-Aalen para o grupo esteróide dos dados de hepatite, apresentadas na Tabela 2.3, podem ser obtidas por meio dos comandos:

```
> require(survival)
> tempos<- c(1,2,3,3,3,5,5,16,16,16,16,16,16,16,16,1,1,1,1,4,5,7,8,10,10,12,16,16,16)
> cens<-c(0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1,1,0,0,0,0,0)
> grupos<-c(rep(1,15),rep(2,14))
> ss<-(survfit(coxph(Surv(tempos[grupos==2],cens[grupos==2])~1,method = "breslow")))
> summary(ss)
> racum<- -log(ss$surv)
> racum
```

Cabe aqui uma observação final sobre a função $\Lambda(t)$. Ela não tem interpretação probabilística mas tem utilidade na seleção de modelos. O gráfico da estimativa desta função, em papéis especiais, é utilizado para verificar a adequação de modelos paramétricos. Este ponto é discutido com mais detalhes no Capítulo 4.

2.4.2 Estimador da Tabela de Vida ou Atuarial

A construção de uma tabela de vida consiste em dividir o eixo do tempo em um certo número de intervalos. Suponha que o eixo do tempo foi dividido em $s+1$ intervalos, definidos pelos seguintes pontos de corte, $t_1, t_2, \ldots, t_s$. Isto é, $I_j = [t_{j-1}, t_j);\ j = 1, \ldots, s$, em que $t_0 = 0$ e $t_s = +\infty$. O estimador da tabela

de vida apresenta a forma (2.3) do estimador de Kaplan-Meier mas utiliza um estimador ligeiramente diferente para q_j uma vez que, neste caso, tem-se para d_j e n_j que:

- d_j = nº de falhas no intervalo $[t_{j-1}, t_j)$ e
- $n_j = \Big[n^o$ de indivíduos sob risco em $\mathrm{t}_{j-1}\Big]$ - $\Big[\frac{1}{2}\times($ nº de censuras em $[t_{j-1}, t_j))\Big]$.

Assim, a estimativa para q_j na tabela de vida é dada por

$$\widehat{q_j} = \frac{\text{nº de falhas no intervalo } [t_{j-1}, t_j)}{\Big[n^o \text{ sob risco em } \mathrm{t}_{j-1}\Big] - \Big[\frac{1}{2}\times(\text{ nº de censuras em } [t_{j-1}, t_j))\Big]}. \tag{2.11}$$

A explicação para o segundo termo do denominador da expressão (2.11) é que observações para as quais a censura ocorreu no intervalo $[t_{j-1}, t_j)$ são tratadas como se estivessem sob risco durante a metade do intervalo considerado.

Utilizando a expressão (2.3), o estimador da tabela de vida fica expresso por:

$$\widehat{S}(t_{I_j}) = \begin{cases} 1 & j = 1 \\ \widehat{S}(t_{I_{(j-1)}})\,(1 - \widehat{q}_{j-1}) & j = 2, \ldots, s. \end{cases}$$

cuja representação gráfica para a função de sobrevivência é uma função escada, com valor constante em cada intervalo de tempo.

Suponha o exemplo da hepatite com os dados do grupo esteróide divididos em 4 intervalos: $[0, 5)$, $[5, 10)$, $[10, 15)$ e $[15, \infty)$. A estimativa de q_2 correspondente ao intervalo $[5, 10)$ é

$$\widehat{q}_2 = \frac{3}{9} = 0,33.$$

Isto significa que a taxa de morte entre a 5^a e a 10^a semana da terapia com esteróide é de 33,3%. O cálculo pode ser estendido da mesma forma para os outros intervalos e estes valores são mostrados na Tabela 2.4.

A estimativa para a função de sobrevivência no tempo $t = 10$ semanas é

$$\widehat{S}(10) = (0,769)(1 - 0,333) = 0,513.$$

Isto significa que um paciente no grupo esteróide tem uma probabilidade de $51,3\%$ de sobreviver a 10 semanas de tratamento. Na Tabela 2.4 estão também apresentados os valores estimados da função de sobrevivência para os outros intervalos de tempo.

A variância assintótica para $\widehat{S}(t_{I_j})$ é, neste caso, obtida por

$$Var(\widehat{S}(t_{I_j})) \cong \Big[\widehat{S}(t_{I_j})\Big]^2 \sum_{\ell=1}^{j-1} \frac{\widehat{q}_\ell}{n_\ell\,(1 - \widehat{q}_\ell)}, \qquad j = 2, \ldots, s.$$

Tabela 2.4: Estimativas da tabela de vida para o grupo esteróide.

Intervalo I_j	No. sob Risco	No. de Falhas	No. de Censuras	$\widehat{q_j}$ (%)	$\widehat{S}(t_{I_j})$ (%)
$[0, 5)$	14	3	2	23,1	100,0
$[5, 10)$	9	3	0	33,3	76,9
$[10, 15)$	6	1	2	20,0	51,3
$[15, \infty)$	3	0	3	0,0	41,0

2.4.3 Comparação dos Estimadores de $S(t)$

A grande diferença entre os estimadores de $S(t)$ está no número de intervalos utilizados para a construção de cada um deles. O estimador de Kaplan-Meier e o de Nelson-Aalen são sempre baseados em um número de intervalos igual ao número de tempos de falha distintos, enquanto que na tabela de vida, os tempos de falha são agrupados em intervalos de forma arbitrária. Isto faz com que a estimativa obtida pelo estimador de Kaplan-Meier seja baseada freqüentemente em um número de intervalos maior que a estimativa obtida através da tabela de vida.

No exemplo discutido nesta seção, o eixo do tempo foi dividido em cinco intervalos de tempo, correspondendo a cada falha distinta, para o estimador de Kaplan-Meier, enquanto que no estimador da tabela de vida foram utilizados somente quatro intervalos de tempo. É natural esperar que, quanto maior o número de intervalos, melhor seja a aproximação para a verdadeira distribuição do tempo de falha. Pode-se então perguntar: porque não usar cinco ou mais intervalos para o cálculo do estimador da tabela de vida? Isto poderia ser feito; no entanto, observa-se que na prática isto não acontece devido às suas origens. A justificativa reside no fato deste estimador ter sido proposto por demógrafos e atuários no século passado e usado sempre em grandes amostras (por exemplo, proveniente de censos demográficos). A divisão em um número arbitrário e grande de intervalos é justificada por ser a amostra muito grande, o que não acontece em resultados provenientes de estudos clínicos ou ensaios de confiabilidade.

O uso da tabela de vida, considerando um número igual ou maior de intervalos que o do estimador de Kaplan-Meier, gera estimativas *exatamente* iguais às estimativas de Kaplan-Meier se o mecanismo de censura for do tipo I ou II. Entretanto, se o mecanismo de censura for do tipo aleatório, as estimativas serão próximas mas não necessariamente coincidentes.

Nesta última situação, alguns autores estudaram as propriedades assintóticas dos dois estimadores. Estes estudos mostraram a superioridade do estimador de Kaplan-Meier. Ele é um estimador não-viciado para a função de sobrevivência em grandes amostras, enquanto o estimador da tabela de vida não o é, com um vício que fica pequeno à medida que o comprimento dos intervalos diminuem. Com amostras de pequeno ou médio porte, existe alguma evidência empírica também da superioridade do estimador de Kaplan-Meier. Desta forma, o mais indicado é então usar o estimador de Kaplan-Meier ou eventualmente o de Nelson-Aalen, ao invés daquele da tabela de vida, quando o interesse se concentrar em informações provenientes da função de sobrevivência.

2.5 Estimação de Quantidades Básicas

A utilização direta da curva de Kaplan-Meier nos informa a probabilidade estimada de sobrevivência para um determinado tempo. Um exemplo é a probabilidade do paciente sobreviver a 12 semanas de tratamento. A estimativa de Kaplan-Meier para este valor é diretamente obtida da Figura 2.3 e é igual a 44%. Se o valor do tempo de interesse estiver ao longo de um degrau da curva de Kaplan-Meier pode-se também utilizar uma interpolação linear. Por exemplo, como havia sido observado na Seção 2.3, a probabilidade estimada de um paciente do grupo esteróide sobreviver a 6 semanas obtida diretamente da curva de Kaplan-Meier é de $0,698$. No entanto, se a interpolação linear for utilizada obtém-se

$$\frac{7-5}{0,611-0,698} = \frac{6-5}{\widehat{S}(6)-0,698}$$

cuja solução é a estimativa de 0,655. Esta última estimativa deve ser preferida (Colosimo et al., 2002).

A partir da curva de Kaplan-Meier também é possível obter estimativas de percentis. Uma informação muito útil é o tempo mediano de vida. Como a curva de Kaplan-Meier é uma função escada, a estimativa mais adequada para a mediana é novamente obtida através de uma interpolação linear. Isto é

$$\frac{10-8}{0,437-0,524} = \frac{\text{MED}-8}{0,50-0,524}$$

cuja solução é a estimativa de 8,55 semanas. Esta forma de estimar estes valores é equivalente a conectar por retas as estimativas de Kaplan-Meier ao invés de se

utilizar $\widehat{S}(t)$ na forma de escada. Esta forma usualmente gera uma melhor representação da distribuição contínua do tempo de falha (Colosimo et al., 2002). De forma análoga, pode-se obter estimativas de outros percentis da distribuição do tempo de vida dos pacientes.

A variância assintótica do estimador de percentis ($\widehat{t_p}$) é expressa por:

$$Var(\widehat{t_p}) = \frac{Var\Big(\widehat{S}(\widehat{t_p})\Big)}{\Big[f(\widehat{t_p})\Big]^2}.$$

A dificuldade em se obter uma estimativa para $f(\widehat{t_p})$ inviabiliza a utilização desta expressão. Brookmeier e Crowley (1985) propõem um estimador alternativo para a mediana invertendo a região de rejeição de um teste não-paramétrico.

Outra quantidade que pode ser de interesse é o tempo médio de vida do paciente. Esta quantidade, no entanto, nem sempre é estimada adequadamente utilizando estimadores não-paramétricos em estudos incluindo censuras. O tempo médio de vida pode ser mostrado, por argumentos probabilísticos, ser dado pela área (integral) sob a função de sobrevivência. Uma estimativa para o tempo médio é então obtida calculando a área sob a curva da estimativa de Kaplan-Meier. Como esta curva é uma função escada, esta integral é simplesmente a soma de áreas de retângulos, isto é,

$$\widehat{t_m} = t_1 + \sum_{j=1}^{k-1} \widehat{S}(t_j)\Big(t_{j+1} - t_j\Big)$$

em que $t_1 < \cdots < t_k$ são os k tempos distintos e ordenados de falha.

Entretanto, um problema surge se o maior tempo observado for uma censura. Isto acontece com frequência em estudos clínicos como é o caso dos dados de hepatite. Neste caso a curva de Kaplan-Meier não atinge o valor zero e o valor do tempo médio de vida fica subestimado. Nestes casos, tal estimativa deve ser interpretada com bastante cuidado ou talvez até mesmo evitada. Uma alternativa é utilizar a mediana ao invés do tempo médio de vida. Ambas são medidas de tendência central, representando um valor típico da distribuição do tempo de vida da população sob estudo. A mediana, no entanto, pode ser extraída facilmente da função de sobrevivência como foi estimada anteriormente para os pacientes do grupo esteróide. A mediana deve somente ser evitada se o número de censuras for maior que o de falhas. Neste caso, ela não é estimável utilizando a curva de Kaplan-Meier. Uma outra forma de estimar o tempo médio de vida será apresentada no Capítulo 3 utilizando modelos paramétricos para os dados de sobrevivência.

Kaplan e Meier (1958) mostraram que a variância assintótica de $\widehat{t}_m$ pode ser estimada por:

$$\widehat{Var}(\widehat{t_m}) = \frac{r}{r-1}\left[\sum_{j=1}^{r-1}\frac{(A_j)^2}{n_j(n_j-d_j)},\right]$$

com $A_j = \widehat{S}(t_j)(t_{j+1}-t_j) + \cdots + \widehat{S}(t_{r-1})(t_r - t_{r-1})$ e r é o número de observações não censuradas, isto é, o número de falhas. Observe que r é igual ao número de falhas e não ao número de falhas *distintas*.

Outra quantidade possivelmente de interesse é o tempo médio restante daqueles pacientes que se encontram livres do evento em um determinado tempo t. Como visto anteriormente, este tempo é estimado pela área sob a curva de sobrevivência à direita de t dividido por $S(t)$, isto é,

$$\text{vmr} = \frac{\text{área sob a curva } S(t) \text{ à direita de } t}{S(t)}.$$

Este estimador apresenta as mesmas limitações de $\widehat{t}_m$.

2.5.1 Exemplo sobre reincidência de tumor sólido

A título de ilustração considere este outro exemplo em que se deseja avaliar os tempos de reincidência de 10 pacientes com tumor sólido (Lee, 1980). Dos 10 pacientes, seis deles apresentaram reincidência em 3; 6,5; 6,5; 10; 12 e 15 meses de seus respectivos ingressos no estudo; um deles perdeu-se contato após 8,4 meses de acompanhamento e três deles permaneceram sem reincidência após 4; 5,7 e 10 meses de acompanhamento. Os esquemas que ilustram hipoteticamente o acompanhamento dos pacientes deste estudo são apresentados na Figura 2.4. Do esquema (a), apresentado nesta figura, observa-se que o experimento foi planejado para durar 18 meses e teve início com três pacientes. Após ter decorrido um mês do início do experimento ocorreu o ingresso do quarto paciente e assim sucessivamente até o décimo paciente que ingressou após decorridos 14 meses de andamento do experimento. O esquema apresentado em (b) mostra, por sua vez, quanto tempo cada paciente permaneceu no estudo. Note que o uso do referencial "zero"neste último esquema, possibilita que o tempo até a ocorrência da falha ou da censura de cada paciente sob estudo, seja observado de maneira mais fácil e direta do que no esquema (a).

Para os dados deste exemplo, as estimativas da função de sobrevivência $S(t)$ e seus respectivos intervalos a 95% de confiança, obtidos utilizando-se o estimador de Kaplan-Meier com o auxílio do pacote estatístico R e os comandos:

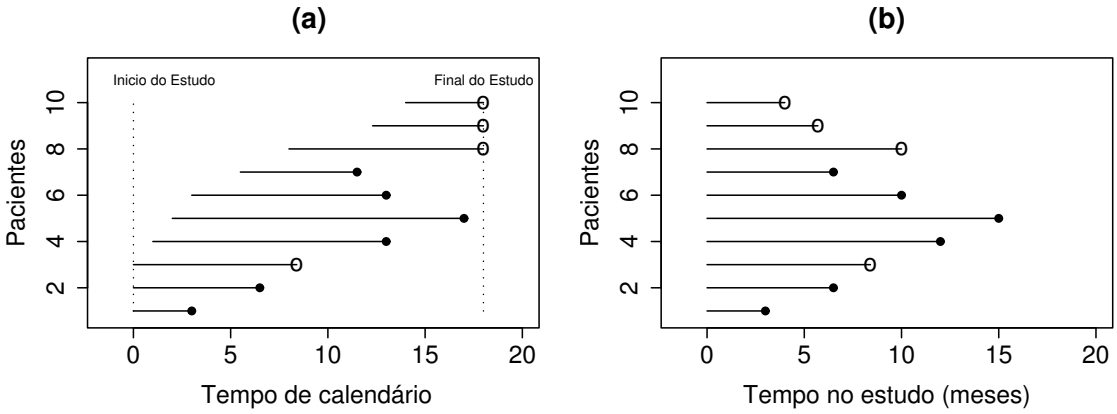

Figura 2.4: (a) esquema ilustrativo dos ingressos no estudo dos pacientes com tumor sólido e seus respectivos períodos de permanência no mesmo; (b) esquema ilustrativo dos tempos até a ocorrência de falha (•) ou censura (o) dos pacientes deste mesmo estudo.

```
> require(survival)
> tempos<- c(3,4,5.7,6.5,6.5,8.4,10,10,12,15)
> cens<- c(1,0,0,1,1,0,1,0,1,1)
> ekm<- survfit(Surv(tempos,cens))
> summary(ekm)
```

encontram-se apresentados na Tabela 2.5.

Tabela 2.5: Estimativas obtidas por meio do estimador de Kaplan-Meier.

Tempos	Intervalos	n_j	d_j	$\left(1-\frac{d_j}{n_j}\right)$	$\hat{S}(t)$	e.p.$(\hat{S}(t))$	$I.C.(\hat{S}(t))_{95\%}$
3	[3; 6,5)	10	1	9/10	0,900	0,900	(0,0949; 1)
6,5	[6,5; 10)	7	2	5/7	0,643	0,643	(0,1679; 1)
10	[10; 12)	4	1	3/4	0,482	0,482	(0,1877; 1)
12	[12; 15)	2	1	1/2	0,241	0,241	(0,1946; 1)
15	[15; ∞)	1	1	0	0,000	0,000	-

A partir da Tabela 2.5, segue que:

$$\hat{S}(t) = \begin{cases} 1 \text{ se t} < 3 \\ 0{,}9 \text{ se } 3 \leq \text{t} < 6{,}5 \\ 0{,}643 \text{ se } 6{,}5 \leq \text{t} < 10 \\ 0{,}482 \text{ se } 10 \leq \text{t} < 12 \\ 0{,}241 \text{ se } 12 \leq \text{t} < 15 \\ 0 \text{ se t} \geq 15 \end{cases}$$

cuja representação gráfica, com os respectivos intervalos a 95% de confiança para todo t tal que $0 \leq t \leq 15$, obtida por meio dos comandos

```
> plot(ekm,conf.int=T,  xlab="Tempo (em meses)", ylab="S(t) estimada", bty="n")
```

é mostrada na Figura 2.5.

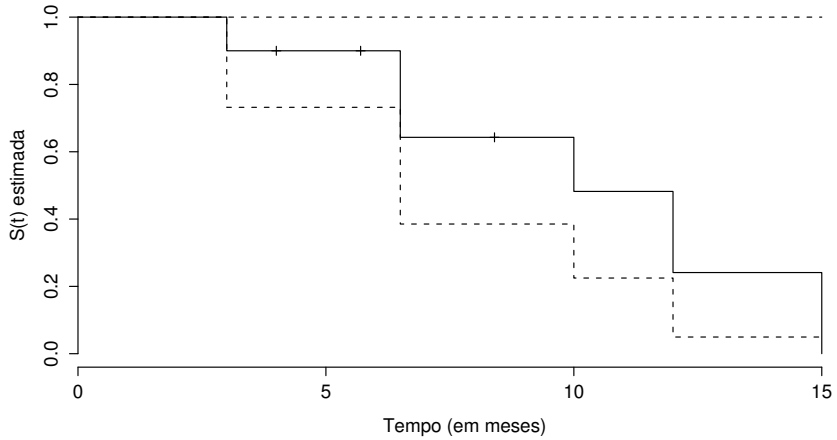

Figura 2.5: Sobrevivência estimada por Kaplan-Meier para os dados de tumor sólido.

Para o tempo mediano, obtido por meio de uma interpolação linear, tem-se que:

$$\frac{10 - 6,5}{0,482 - 0,643} = \frac{\text{MED} - 6,5}{0,50 - 0,643}$$

cuja solução é 9,61 meses. Isto significa uma estimativa do tempo em que 50% dos pacientes permanece vivos é de 9,6 meses. Tem-se, ainda, para os pacientes deste exemplo, um tempo médio de vida estimado de $\widehat{t_m} = 10{,}088$ meses. Tal estimativa pode ser obtida utilizando-se o pacote estatístico R por meio dos comandos:

```
> t<- tempos[cens==1]
> tj<-c(0,as.numeric(levels(as.factor(t))))
> surv<-c(1,as.numeric(levels(as.factor(ekm$surv))))
> surv<-sort(surv, decreasing=T)
> k<-length(tj)-1
> prod<-matrix(0,k,1)
>   for(j in 1:k){
>     prod[j]<-(tj[j+1]-tj[j])*surv[j]
>   }
> tm<-sum(prod)
> tm
```

Observe, neste exemplo, que o tempo médio apresenta-se bem estimado uma vez que o maior tempo observado trata-se de uma falha. O mesmo não seria verdade, como discutido anteriormente, se o referido tempo correspondesse a uma censura. A variância estimada de $\widehat{t}_m$ foi também obtida e esta resultou em:

$$\begin{aligned}\widehat{Var}(\widehat{t_m}) &= \frac{6}{5}\left[\frac{(A_1)^2}{9\times 10}+\frac{(A_2)^2}{6\times 7}+\frac{(A_3)^2}{5\times 6}+\frac{(A_4)^2}{3\times 4}+\frac{(A_5)^2}{1\times 2}\right]\\ &= 2,33\end{aligned}$$

sendo,

$$\begin{aligned}A_1 &= 0,9(6,5-3)+0,643(10-6,5)+0,482(12-10)+0,241(15-12)=7,088\\ A_2 &= A_3 = 0,643(10-6,5)+0,482(12-10)+0,241(15-12)=3,938\\ A_4 &= 0,482(12-10)+0,241(15-12)=1,687\\ A_5 &= 0,241(15-12)=0,723.\end{aligned}$$

Para pacientes que sobreviverem até, por exemplo, o tempo $t = 10$ meses, estima-se também que os mesmos tenham um tempo médio de vida restante de:

$$\widehat{\text{vmr}} = \frac{\text{área sob a curva } \widehat{S}(t) \text{ à direita de } t{=}10}{\widehat{S}(10)} = 3,5 \text{ meses.}$$

Note, a partir da Tabela 2.5 e também da Figura 2.5, que os intervalos de confiança obtidos para $\widehat{S}(t)$ são relativamente amplos. Isto se deve em particular ao tamanho amostral relativamente pequeno ($n = 10$).

2.6 Comparação de Curvas de Sobrevivência

O estudo clínico controlado, apresentado na Seção 1.5.1, foi realizado para investigar o efeito da terapia com esteróide no tratamento de hepatite viral aguda. Isto significa que o objetivo principal do estudo é comparar o grupo tratado com esteróide e o controle. Um procedimento natural usaria os resultados assintóticos de $\widehat{S}(t)$, apresentados na seção anterior, para testar a igualdade de funções de sobrevivência em determinado tempo t. Esta forma, no entanto, não faria uso eficiente dos dados disponíveis, pois não se estaria usando todo o período do estudo. Estatísticas mais comumente usadas podem ser vistas como generalizações para dados censurados, de conhecidos testes não-paramétricos. O teste de logrank (Mantel, 1966) é o mais usado em análise de sobrevivência. Gehan (1965) propôs uma generalização para

a estatística de Wilcoxon. Outras generalizações foram propostas por Peto e Peto (1972) e Prentice (1978), entre outros. Latta (1981) fez uso de simulações de Monte Carlo para comparar vários testes não-paramétricos.

Neste texto ênfase será dada ao teste logrank. Este teste é muito utilizado em análise de sobrevivência, e é particularmente apropriado quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante. Isto é, as populações têm a propriedade de riscos proporcionais. A estatística deste teste é a diferença entre o número observado de falhas em cada grupo e uma quantidade que, para muitos propósitos, pode ser pensada como o correspondente número esperado de falhas sob a hipótese nula. A expressão do teste logrank é obtida de forma similar do conhecido teste de Mantel-Hanzel (1959), para combinar tabelas de contingência. O teste logrank tem também a mesma expressão do teste escore para o modelo de regressão de Cox que será apresentado no Capítulo 5. Outros testes também serão apresentados nesta seção.

Considere, inicialmente, o teste de igualdade de duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. Sejam $t_1 < t_2 < \ldots < t_k$ os tempos de falha distintos da amostra formada pela combinação das duas amostras individuais. Suponha que no tempo t_j acontecem d_j falhas e n_j indivíduos estão sob risco em um tempo imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i; $i = 1, 2$ e $j = 1, \ldots, k$. Em cada tempo de falha t_j, os dados podem ser dispostos em forma de uma tabela de contingência 2 x 2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i. Isto é mostrado na Tabela 2.6.

Tabela 2.6: Tabela de contingência gerada no tempo t_j.

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

Condicional à experiência de falha e censura até o tempo t_j (fixando as marginais de coluna) e ao número de falhas no tempo t_j (fixando as marginais de linha), a distribuição de d_{2j} é então uma hipergeométrica

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

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

$$(V_j)_2 = n_{2j}(n_j - n_{2j})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}.$$

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

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}. \tag{2.12}$$

que tem uma distribuição qui-quadrado com 1 grau de liberdade para grandes amostras.

O objetivo principal do estudo controlado dos dados de hepatite é comparar a terapia com esteróide e o grupo controle. As curvas de Kaplan-Meier para os dois grupos apresentadas na Figura 2.3 indicam que possivelmente a terapia com esteróide não é um tratamento adequado para pacientes com hepatite viral aguda. No entanto, é necessário uma evidência quantitativa deste fato através de um teste de significância. O valor do teste logrank para a comparação entre estes dois grupos resultou em

$$T = 3,67,$$

o que implica em um valor $p = 0,055$, indicando uma diferença entre as duas curvas de sobrevivência. O valor deste teste e seu correspondente valor p podem ser obtidos no pacote estatístico R por meio dos comandos:

```
> require(survival)
> tempos<- c(1,2,3,3,3,5,5,16,16,16,16,16,16,16,16,1,1,1,1,4,5,7,8,10,10,12,16,16,16)
> cens<-c(0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1,1,0,0,0,0,0)
> grupos<-c(rep(1,15),rep(2,14))
> survdiff(Surv(tempos,cens)~grupos,rho=0)
```

A generalização do teste logrank para a igualdade de $r > 2$ funções de sobrevivência $S_1(t), \ldots, S_r(t)$ não é complicada. Considere a mesma notação anterior, com o índice i agora variando entre 1 e r. Desta forma, os dados podem ser arranjados em forma de uma tabela de contingência 2 x r com d_{ij} falhas e $n_{ij} - d_{ij}$

sobreviventes na coluna i. Ou seja, a Tabela 2.6 passaria a ter r colunas ao invés de simplesmente duas.

Condicional à experiência de falha e censura até o tempo t_j e ao número de falhas no tempo t_j, a distribuição conjunta de $d_{1j}, d_{2j}, \ldots, d_{rj}$ é então uma hipergeométrica multivariada, isto é,

$$\frac{\prod_{i=1}^{r} \begin{pmatrix} n_{ij} \\ d_{ij} \end{pmatrix}}{\begin{pmatrix} n_j \\ d_j \end{pmatrix}}.$$

A média de d_{ij} é $w_{ij} = n_{ij} d_j n_j^{-1}$, a variância de d_{ij} e a covariância de d_{ij} e d_{lj} são, respectivamente,

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij} n_{lj} d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}.$$

Então a estatística $v_j^{'} = (d_{2j} - w_{2j}, \ldots, d_{rj} - w_{rj})$ tem média zero e matriz de variância-covariância V_j. Pode-se então formar a estatística v, somando sobre todos os tempos distintos de falha, isto é,

$$v = \sum_{j}^{k} v_j$$

com v um vetor de dimensão $r \times 1$ cujos elementos são as diferenças entre os totais observados e esperados de falha.

Considerando novamente a suposição de que as k tabelas de contingência são independentes, a variância da estatística v será $V = V_1 + \ldots + V_k$. Um teste aproximado para a igualdade das r funções de sobrevivência pode ser baseado na estatística

$$T = v^{'} V^{-1} v, \tag{2.13}$$

que tem uma distribuição qui-quadrado com $r-1$ graus de liberdade para amostras grandes. Os graus de liberdade são $r-1$ e não r, pois os elementos de v somam zero.

2.6.1 Outros Testes

Outros testes não-paramétricos foram propostos para comparar funções de sobrevivência. No caso particular da comparação de duas funções de sobrevivência a

seguinte forma geral inclui os testes mais importantes na literatura e generaliza a estatística T em (2.12),

$$S = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2}, \tag{2.14}$$

com u_j os pesos que especificam os testes. Sob a hipótese nula de que as funções de sobrevivência são iguais, a estatística S tem distribuição qui-quadrado com 1 grau de liberdade para amostras grandes. O teste logrank (2.12) é obtido tomando-se $u_j = 1, \quad j = 1, \ldots, k$. Outro teste bastante utilizado na prática é o de Wilcoxon obtido quando se toma $u_j = n_j$. Este teste foi adaptado para dados censurados a partir do conhecido teste não-paramétrico de Wilcoxon (Gehan, 1965, Breslow, 1970). O teste de Tarone e Ware (1977) propõe peso $u_j = \sqrt{n_j}$, que fica entre os pesos do logrank e do Wilcoxon. A Tabela 2.7 apresenta os resultados dos três testes para os dados de hepatite.

Tabela 2.7: Testes não-paramétricos para comparação das curvas de sobrevivência dos grupos esteróide e controle dos dados de hepatite.

Teste	Estatística	valor-p
Logrank	3,67	0,055
Wilcoxon	3,19	0,074
Tarone-Ware	3,43	0,064

A escolha do peso na expressão (2.14) direciona o tipo de diferença a ser detectado nas funções de sobrevivência. O teste de Wilcoxon que utiliza peso igual ao número de indivíduos sob risco, coloca mais pesos na porção inicial do eixo do tempo. No início do estudo todos indivíduos estão sob risco e saindo do estado "sob risco" à medida que falham ou são censurados. O teste logrank, por outro lado, coloca mesmo peso para todo o eixo do tempo, o que reforça o enfoque nos tempos maiores quando comparado ao teste de Wilcoxon. O teste de Tarone-Ware se localiza em uma situação intermediária.

Peto e Peto (1972) e Prentice (1978) sugerem utilizar uma função do peso que depende diretamente da experiência passada de sobrevivência observada das duas amostras combinadas. A função do peso é uma modificação do estimador de Kaplan-Meier e é definido de tal forma que seu valor é conhecido antes da falha ocorrer. O

estimador modificado da função de sobrevivência é

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

e os pesos utilizados são

$$u_j = \tilde{S}(t_{j-1}) \frac{n_j}{n_j + 1}.$$

Este estimador é conhecido por Peto-Prentice. Outra classe de pesos para a expressão (2.14) foi proposta por Harrington-Fleming (1982) como

$$u_j = \left[\widehat{S}(t_{j-1}) \right]^{\rho}.$$

Se $\rho = 0$, obtém-se $u_j = 1$ e tem-se então o teste logrank. Entretanto, se $\rho = 1$, então o peso é o Kaplan-Meier no tempo de falha anterior que é um teste similar ao de Peto-Prentice.

A principal vantagem dos testes de Peto-Prentice e Harrington-Fleming é que a ponderação é feita relativa à experiência de sobrevivência anterior. Isto não acontece com os testes de Wilcoxon e logrank. O teste de Wilcoxon, em particular, pondera pelo número de indivíduos sob risco que depende da experiência de sobrevivência assim como da de censura. Se o padrão de censura é nitidamente diferente nos dois grupos, então o teste pode rejeitar ou não rejeitar, não somente com base nas diferenças das sobrevivências entre os grupos mas também devido ao padrão de censura.

2.7 Exercícios

1. Mostre que a partir da transformação $U(t) = \log[-\log S(t)]$ obtém-se o intervalo de 95% de confiança para $S(t)$ mostrado em (2.8).

2. Os dados mostrados abaixo representam o tempo até a ruptura de um tipo de isolante elétrico sujeito a uma tensão de estresse de 35 Kvolts. O teste consistiu em deixar 25 destes isolantes funcionando até que 15 deles falhassem (censura do tipo II) obtendo-se os seguintes resultados (em minutos):

0,19	0,78	0,96	1,31	2,78	3,16	4,67	4,85
6,50	7,35	8,27	12,07	32,52	33,91	36,71	

   A partir destes dados amostrais, deseja-se obter as seguintes informações:

(a) uma estimativa para o tempo mediano de vida deste tipo de isolante elétrico funcionando a 35 Kvolts,

(b) uma estimativa (pontual e intervalar) para a fração de defeituosos esperada nos dois primeiros minutos de funcionamento,

(c) uma estimativa (pontual) para o tempo médio de vida destes isoladores funcionando a 35 Kvolts (limitado em 40 minutos) e,

(d) o tempo necessário para 20% dos isolantes estarem fora de operação.

3. Os dados da Tabela 2.8 referem-se aos tempos de sobrevivência (em dias) de pacientes com câncer submetidos à radioterapia (+ indica censura). Para esses

Tabela 2.8: Tempos de sobrevivência de pacientes submetidos à radioterapia.

7, 34, 42, 63, 64, 74^+, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157, 160, 160, 165, 173, 176, 185^+, 218, 225, 241, 248, 273, 277, 279^+, 297, 319^+, 405, 417, 420, 440, 523, 523^+, 583, 594, 1101, 1116^+, 1146, 1226^+, 1349^+, 1412^+, 1417

Fonte: Louzada Neto et al. (2002)

dados obtenha:

(a) a função de sobrevivência estimada pelos métodos de Kaplan-Meier e Nelson-Aalen. Apresente-as em tabelas e graficamente;

(b) os tempos mediano e médio;

(c) as probabilidades de um paciente com câncer sobreviver a: i) 42 dias, ii) 100 dias, iii) 300 dias e iv) 1000 dias;

(d) o tempo médio de vida restante dos pacientes que sobreviverem 1000 dias;

(e) interprete as estimativas obtidas nos itens 1.2) a 1.4).

(f) para quais tempos tem-se: i) $\hat{S}(t) = 0,80$, ii) $\hat{S}(t) = 0,30$ e $\hat{S}(t) = 0,10$? Interprete.

4. Os dados apresentados na Tabela 2.9 representam o tempo (em dias) até a morte de pacientes com câncer de ovário tratados na Mayo Clinic (Fleming et al., 1980). O símbolo "+"indica censura.

(a) Construa estimativas de Kaplan-Meier para as funções de sobrevivência de ambos os grupos e apresente-as no mesmo gráfico.

Tabela 2.9: Tempos (em dias) dos pacientes no estudo de câncer de ovário.

Amostra	Tempo de sobrevivência em dias
1. Tumor Grande	28, 89, 175, 195, 309, 377+, 393+, 421+, 447+, 462, 709+, 744+, 770+, 1106+, 1206+
2. Tumor Pequeno	34, 88, 137, 199, 280, 291, 299+, 300+, 309, 351, 358, 369, 369, 370, 375, 382, 392, 429+, 451, 1119+

(b) Repita a letra (a) usando as estimativas de Nelson-Aalen.

(c) Usando os intervalos de confiança assintóticos das estimativas de Kaplan-Meier, teste a hipótese de igualdade das funções de sobrevivência dos dois grupos em $t = 6$ meses e 18 meses.

(d) Teste a hipótese de igualdade das funções de sobrevivência dos dois grupos usando dois testes diferentes.

(e) Apresente informações sobre o *software* (se algum) utilizado para realizar os cálculos deste problema.

5. Um estudo de sobrevivência foi realizado para comparar dois métodos para a realização de transplante de medula em pacientes com leucemia. A resposta de interesse era o tempo contado a partir do transplante até a morte do paciente.

(a) Os seguintes resultados foram obtidos:

$$\sum_{j=1}^{k}(d_{2j} - w_{2j}) = 3,964 \qquad \text{e} \qquad \sum_{j=1}^{k}(V_j)_2 = 6,211.$$

Estabeleça as hipótese, obtenha o teste logrank e conclua. Use o nível de significância de 5% (3,84).

(b) Neste estudo os pesquisadores não têm interesse em detectar diferenças entre os métodos nos tempos iniciais devido a toxicidade dos medicamentos. Você usaria o teste logrank ou Wilcoxon nesta situação? Justifique sua resposta.

Capítulo 3

Modelos Probabilísticos em Análise de Sobrevivência

3.1 Introdução

O objetivo deste capítulo é apresentar a análise estatística de dados de sobrevivência envolvendo distribuições de probabilidade. Elas são chamadas de modelos probabilísticos para o tempo de falha. Estas distribuições são bastante utilizadas, principalmente para produtos industriais, por se mostrarem adequadas para descrever estes tempos de vida. Os modelos paramétricos vêm sendo utilizados com mais freqüência na área industrial do que na médica. A principal razão deste fato é que os estudos envolvendo componentes e equipamentos industriais podem ser planejados e conseqüentemente as fontes de perturbação (heterogeneidade) podem ser controladas. Nestas condições a busca por um modelo paramétrico adequado fica facilitada e a análise estatística dos dados fica mais precisa.

Existem diversos livros de probabilidade que fazem uma apresentação exaustiva dos modelos paramétricos e que podem ser usados pelo leitor em busca de mais informações. Entre eles pode-se citar Johnson e Kotz (1970). Os principais modelos probabilísticos utilizados em análise de sobrevivência são apresentados na Seção 3.2. O método de máxima verossimilhança para a estimação dos parâmetros dos modelos é introduzido na Seção 3.3. Nesta seção são também apresentadas as propriedades destes estimadores, que possibilitam a construção de intervalos de confiança para os parâmetros do modelo ou para uma função deles. A Seção 3.4 apresenta técnicas gráficas e o teste da razão de verossimilhança para discriminar entre modelos probabilísticos. Exemplos são analisados na Seção 3.5.

3.2 Principais Distribuições em Sobrevivência

Algumas distribuições de probabilidade são certamente familiares para o leitor, como é o caso da normal (gaussiana) e da binomial. Elas descrevem de forma adequada certas variáveis clínicas e industriais. Por outro lado, quando se trata de descrever a variável "tempo até a falha", outras distribuições se mostram mais adequadas.

Embora exista uma série de modelos probabilísticos utilizados em análise de dados de sobrevivência, alguns destes modelos ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas. Entre estes modelos é possível citar o exponencial, o de Weibull e o log-normal.

O leitor deve se ater às características de cada uma das distribuições. É importante entender que cada distribuição de probabilidade pode gerar estimadores diferentes para a mesma quantidade desconhecida. Desta forma, a utilização de um modelo inadequado acarreta em erros grosseiros nas estimativas destas quantidades. A escolha de um modelo probabilístico adequado para descrever o tempo de falha deve, então, ser feita com bastante cuidado. Este tópico é abordado na Seção 3.4. Algumas das principais distribuições de probabilidade usadas em análise de sobrevivência são apresentadas a seguir.

3.2.1 Distribuição Exponencial

Em termos matemáticos, a distribuição exponencial é um dos modelos probabilísticos mais simples usados para descrever o tempo de falha. Esta distribuição apresenta um único parâmetro e é a única que se caracteriza por um ter uma função de taxa de falha (ou de risco) constante. Ela tem sido extensivamente usada como um modelo para o tempo de vida de certos produtos e materiais e tem descrito adequadamente o tempo de vida de óleos isolantes e dielétricos, entre outros. Cox e Snell (1981) utilizaram o modelo exponencial para descrever o tempo de vida de pacientes adultos com leucemia.

A função de densidade de probabilidade para a variável aleatória tempo de falha T com distribuição exponencial é dada por:

$$f(t) = \frac{1}{\alpha} \exp\Big\{ -\Big(\frac{t}{\alpha}\Big)\Big\}, \qquad t \geq 0 \tag{3.1}$$

em que o parâmetro $\alpha \geq 0$ é o tempo médio de vida. O parâmetro α tem a mesma unidade do tempo de falha t. Isto é, se t é medido em horas, α também será medido em horas.

Ainda, as funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$ são dadas, respectivamente, por

$$S(t) = \exp\Big\{-\Big(\frac{t}{\alpha}\Big)\Big\} \tag{3.2}$$

e

$$\lambda(t) = \frac{1}{\alpha} \quad \text{para } t \geq 0. \tag{3.3}$$

A forma típica destas três funções para diferentes valores de α pode ser observada na Figura 3.1.

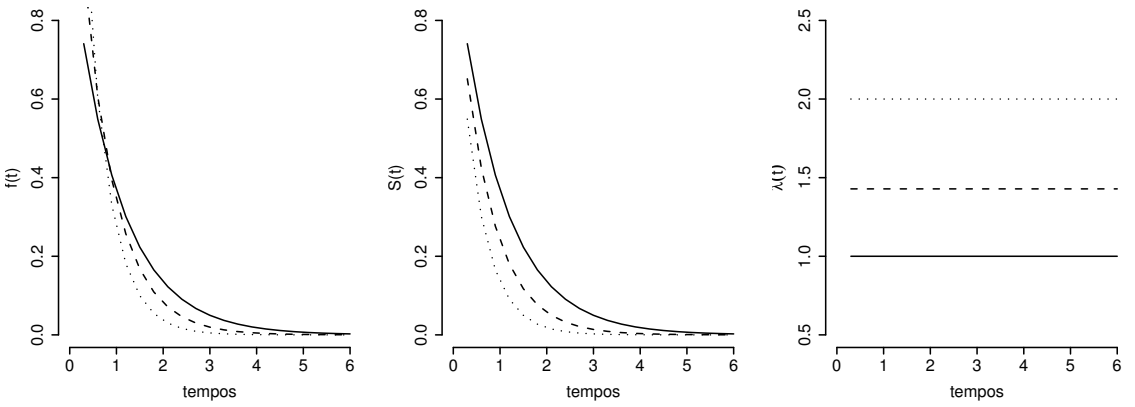

Figura 3.1: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição exponencial para α = 1,0 (–), 0,7 (- -) e 0,5 ($\cdots$).

Como dito anteriormente, somente a distribuição exponencial tem uma taxa de falha constante. Isto significa que tanto uma unidade velha quanto uma nova, que ainda não falharam, têm o mesmo risco de falhar em um intervalo futuro. Esta propriedade é chamada de falta de memória da distribuição exponencial.

Outras características de interesse são a média, a variância e os percentis. A média da distribuição exponencial é α, e a variância α^2. O percentil $100p\%$ corresponde ao tempo em que $100p\%$ dos produtos ou indivíduos falharam. Os percentis são importantes para obtenção, por exemplo, de informações a respeito de falhas prematuras. Eles podem ser obtidos a partir da função de densidade ou da função de sobrevivência. Para o caso da distribuição exponencial, o percentil $100p\%$, t_p, pode ser obtido por

$$t_p = -\alpha \log(1-p).$$

Conhecido então o valor de α, o percentil correspondente a mediana, por exemplo, é facilmente obtido por $t_{0,5} = -\alpha \log(1 - 0,5)$. A média da distribuição exponencial corresponde ao $t_{0,63}$, ou seja, o percentil 63%.

Alguns livros de confiabilidade (Meeker e Escobar, 1998, Ebeling, 1997) apresentam o modelo exponencial com dois parâmetros. Neste modelo, um parâmetro de locação t_0 é incluído para representar um período inicial de tempo em que a falha nunca ocorre. Este parâmetro é conhecido como tempo de garantia. A função de densidade desta nova variável T é obtida substituindo-se t por $t - t_0$ na expressão (3.1) e o suporte de T fica definido a partir de t_0. É difícil, contudo, em situações práticas, assumir com certeza que ocorra este período inicial sem falhas. Observe que esta afirmação é determinística.

3.2.2 Distribuição de Weibull

A distribuição de Weibull foi proposta originalmente por W. Weibull (1954) em estudos relacionados ao tempo de falha devido a fadiga de metais e, desde então, vem sendo freqüentemente usada em estudos biomédicos e industriais. A sua popularidade em aplicações práticas deve-se ao fato dela apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de taxa de falha é monótona. Isto é, ou ela é crescente ou decrescente ou constante.

Para uma variável aleatória T com distribuição de Weibull tem-se a função de densidade de probabilidade dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma}\, t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\}, \quad t \geq 0 \tag{3.4}$$

em que γ, o parâmetro de forma, e α, o de escala, são ambos positivos. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade.

Para esta distribuição, as funções de sobrevivência e de risco são, respectivamente,

$$S(t) = \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\} \tag{3.5}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma}\, t^{\gamma-1} \tag{3.6}$$

para t, α e $\gamma \geq 0$. Observe que quando $\gamma = 1$ tem-se a distribuição exponencial e, sendo assim, a distribuição exponencial é um caso particular da distribuição de

Weibull. Algumas formas das funções de densidade, de sobrevivência e de taxa de falha (risco) de uma variável T com distribuição de Weibull são mostradas na Figura 3.2.

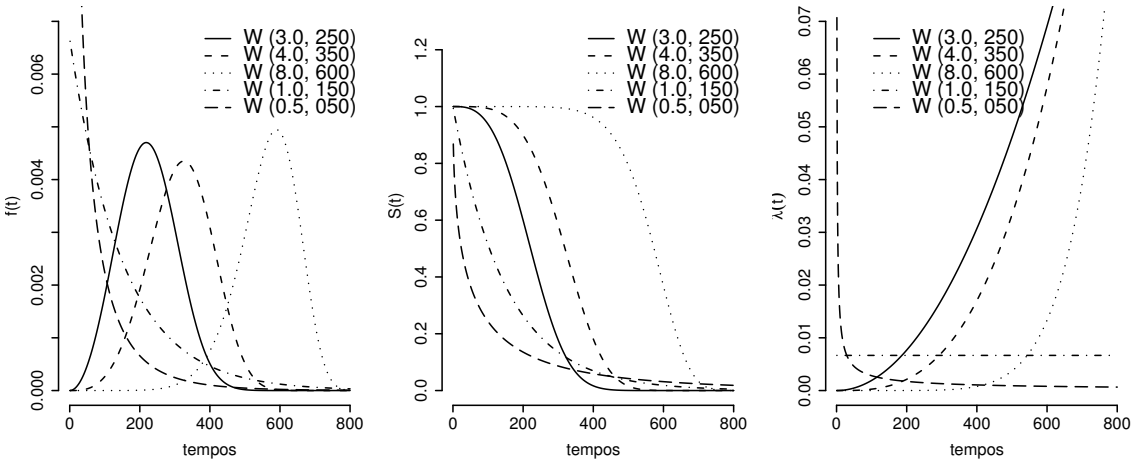

Figura 3.2: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull para alguns valores dos parâmetros (γ, α).

Observe, a partir da Figura 3.2, que a função de taxa de falha $\lambda(t)$ é estritamente crescente para $\gamma > 1$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 1$. Para $\gamma = 1$ tem-se a função de taxa de falha da distribuição exponencial que, como mencionado, é um caso particular da de Weibull.

As expressões para a média e a variância da Weibull incluem o uso da função gama, isto é,

$$\begin{aligned} E[T] &= \alpha\,\Gamma[1 + (1/\gamma)], \\ Var[T] &= \alpha^2\Big[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2\Big], \end{aligned}$$

sendo $\Gamma(r) = (r-1)!$ para r inteiro. A tabela da função gama deve ser utilizada para valores não inteiros de r (Abramowitz e Stegun, 1965). Os percentis são dados por

$$t_p = \alpha\Big[-\log(1-p)\Big]^{1/\gamma}.$$

É importante neste ponto, introduzir uma distribuição que é bastante relacionada à de Weibull. Ela é chamada de distribuição do valor extremo ou de Gambel e surge quando se toma o logaritmo de uma variável com a distribuição de Weibull. Isto é, se a variável T tem uma distribuição de Weibull com $f(t)$ dada por (3.4), então a

variável $Y = \log(T)$ tem uma distribuição do valor extremo com a seguinte função de densidade

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

em que y e $\mu \in \Re$ e $\sigma > 0$. Se $\mu = 0$ e $\sigma = 1$ tem-se a, assim denominada, distribuição do valor extremo padrão. Os parâmetros μ e σ são chamados de parâmetros de locação e escala, respectivamente. Os parâmetros das distribuições de Weibull e do valor extremo apresentam as seguintes relações de igualdade: $\gamma = 1/\sigma$ e $\alpha = \exp\{\mu\}$.

As funções de sobrevivência e de taxa de falha da variável Y são dadas, respectivamente, por

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

e

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

A média e a variância são, respectivamente, $\mu - \nu\sigma$ e $(\pi^2/6)\sigma^2$, com $\nu = 0,5772...$ a conhecida constante de Euler. O percentil $100p\%$ é dado por

$$t_p = \mu + \sigma \log[-\log(1-p)].$$

Na análise de dados de tempo de vida é muitas vezes conveniente trabalhar com o logaritmo dos tempos de vida observados. Este fato é explorado nos modelos de regressão a ser discutido no Capítulo 4. Desta forma, se os dados tiverem uma distribuição de Weibull, a distribuição do valor extremo aparece naturalmente na modelagem.

3.2.3 Distribuição Log-normal

Assim como a distribuição de Weibull, a distribuição log-normal é muito utilizada para caracterizar tempos de vida de produtos e indivíduos. Isto inclui, fadiga de metal, semicondutores, diodos e isolação elétrica. Ela também é bastante utilizada para descrever situações clínicas, como o tempo de vida de pacientes com leucemia.

A função de densidade de uma variável aleatória T com distribuição log-normal é dada por

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\} \qquad t \geq 0 \tag{3.7}$$

em que μ é a média do logaritmo do tempo de falha assim como σ é o desvio-padrão.

Existe uma relação entre as distribuições log-normal e normal similar à relação existente entre as distribuições de Weibull e do valor extremo. Esta relação facilita a apresentação e análise de dados provenientes da distribuição log-normal. Como o nome sugere, o logaritmo de uma variável com distribuição log-normal com parâmetros μ e σ tem uma distribuição normal com média μ e desvio-padrão σ. Esta relação significa que dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, desde de que, é claro, se considere o logaritmo dos dados ao invés dos valores originais.

As funções de sobrevivência e de taxa de falha de uma variável log-normal não apresentam uma forma analítica explícita e são, desse modo, expressas, respectivamente, por

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right) \qquad \text{e} \qquad \lambda(t) = \frac{f(t)}{S(t)}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

A Figura 3.3 apresenta a forma de algumas funções de densidade, de sobrevivência e de taxa de falha da disribuição log-normal para alguns valores de μ e σ.

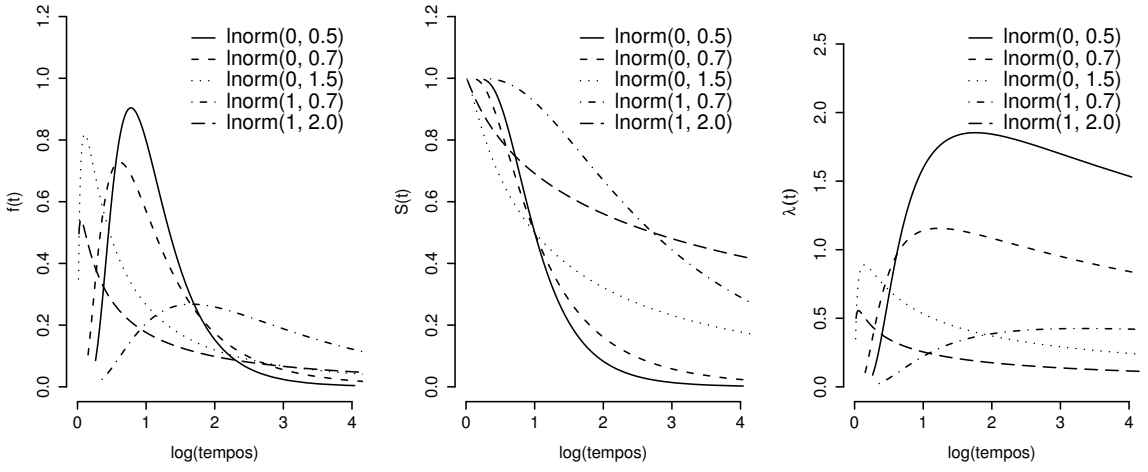

Figura 3.3: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição log-normal para alguns valores dos parâmetros (μ, σ).

Observe que as funções de taxa de falha não são monótonas como as da distribuição de Weibull. Elas crescem, atingem um valor máximo e depois decrescem. Os percentis para a distribuição log-normal podem ser obtidos a partir da tabela da

normal padrão, usando-se a seguinte expressão

$$t_p = \exp\{z_p\,\sigma + \mu\}$$

com z_p o $100p\%$ percentil da distribuição normal padrão. A média e a variância da distribuição log-normal são dadas, respectivamente, por $E[T] = \exp\{\mu + \sigma^2/2\}$ e $\mathrm{Var[T]} = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1)$.

3.2.4 Distribuições Gama e Gama Generalizada

A distribuição gama, que também inclui a exponencial como um caso especial, foi usada por Brown e Flood (1947) para descrever o tempo de vida de copos de vidro circulando em uma cafeteria e também por Birnbaum e Saunders (1958) para descrever o tempo de vida de materiais. Desde então, esta distribuição tem sido usada em problemas de confiabilidade pois a mesma se ajusta adequadamente a uma variedade deles. Em problemas da área médica sua utilização é mais recente. É por exemplo, a distribuição assumida com maior freqüência nos modelos de fragilidade tratados no Capítulo 9.

A função de densidade da distribuição gama, que é caracterizada por dois parâmetros, k e α, em que $k > 0$ é chamado parâmetro de forma e $\alpha > 0$ de escala, é expressa por

$$f(t) = \frac{1}{\Gamma(k)\;\alpha^k}\;t^{k-1}\exp\Big\{-\Big(\frac{t}{\alpha}\Big)\Big\} \qquad t > 0. \tag{3.8}$$

Para $k > 1$, esta função apresenta um único pico em $t = (k-1)/\alpha$. A respectiva função de sobrevivência desta distribuição é dada por

$$S(t) = \int_t^\infty \frac{1}{\Gamma(k)\;\alpha^k}\;u^{k-1}\exp\Big\{-\Big(\frac{u}{\alpha}\Big)\Big\}\,du. \tag{3.9}$$

A função de taxa de falha, obtida da relação $\lambda(t) = f(t)/S(t)$, apresenta um padrão crescente ou decrescente convergindo, no entanto, para um valor constante quando t cresce de 0 a infinito.

Representações gráficas das funções de densidade e de taxa de falha da distribuição gama, para alguns valores de k e α, podem ser observadas na Figura 3.4. Note, a partir desta figura, que para $k > 1$, a taxa de falha cresce monotonicamente de 0 até α quando t cresce de 0 a infinito. Já para $0 < k < 1$, a taxa de falha decresce monotonicamente de infinito até α quando t cresce de 0 a infinito. Observe, ainda, que para $k = 1$ tem-se a distribuição exponencial como um caso especial da gama e sendo assim, a taxa de falha é, neste caso, constante.

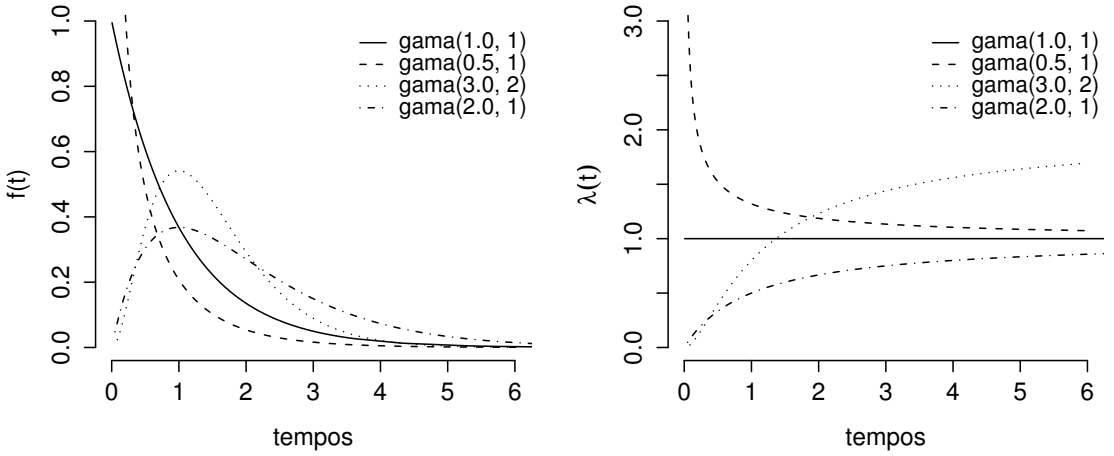

Figura 3.4: Forma típica das funções de densidade de probabilidade e de taxa de falha da distribuição gama para alguns valores dos parâmetros (k, α).

A média e variância da distribuição gama são dadas, respectivamente, por $k\,\alpha$ e $k\,\alpha^2$. A distribuição gama com o parâmetro k restrito a valores inteiros $(1, 2, \ldots)$ é conhecida como distribuição Erlangian (Lee, 1980).

Outra distribuição que merece destaque em análise de sobrevivência é a distribuição gama generalizada. Esta distribuição foi introduzida por Stacy(1962) e é caracterizada por três parâmetros, γ, k e α, todos positivos. Sua função de densidade é dada por

$$f(t) = \frac{\gamma}{\Gamma(k)\ \alpha^{\gamma k}}\ t^{\gamma k-1} \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)^{\gamma} \Big\} \qquad t > 0$$

em que $\Gamma(k)$ é a função gama, isto é, $\Gamma(k) = \int_0^\infty x^{k-1} \exp\{-x\} dx$. Para esta distribuição tem-se um parâmetro de escala, α, e dois de forma, γ e k, o que a torna bastante flexível.

Note, a partir da função de densidade da distribuição gama generalizada, que:

i) para $\gamma = k = 1$ tem-se $T \sim \text{Exp}(\alpha)$,

ii) para $k = 1$ tem-se $T \sim \text{Weibull}(\gamma, \alpha)$,

iii) e para $\gamma = 1$ tem-se $T \sim \text{Gama}(k, \alpha)$.

Pode-se ainda mostrar (Lawless, 1980) que a distribuição log-normal aparece como um caso limite da distribuição gama generalizada quando $k \to \infty$.

Do que foi exposto, tem-se que a distribuição gama generalizada inclui, como casos especiais, as distribuições exponencial, de Weibull, gama e log-normal. Esta propriedade da gama generalizada faz com que a mesma seja de grande utilidade, por exemplo, na decisão entre modelos probabilísticos alternativos, como será visto na Seção 3.4.2.

3.2.5 Outros Modelos Probabilísticos

Existem outras distribuições de probabilidade apropriadas para modelar o tempo de falha de produtos, materiais e situações clínicas. Dentre elas, podem ser citadas as distribuições logística, log-logística, log-gama, Rayleigh, normal inversa e Gompertz.

Diversos textos apresentam a popular função de taxa de falha do tipo "banheira"que descreve o comportamento das taxas de falhas de certos produtos industriais e, principalmente, do tempo de vida dos seres humanos. Para esta função, representada graficamente na Figura 3.5, distinguem-se três regiões distintas:

1ª) *Período de Falhas Prematuras ou Mortalidade Infantil*: é caracterizado por uma taxa de falha alta que decresce rapidamente com o tempo. Neste período, uma pequena porcentagem da população apresenta falhas devido a defeitos grosseiros de fabricação ou itens que sofreram solicitações (estresses) extraordinárias antes do uso. As falhas prematuras são usualmente removidas por um pré-envelhecimento conhecido por "burn-in"(Jensen e Petersen, 1982). Esta porção da curva é também conhecida por fase de mortalidade infantil.

2ª) *Período de Vida Útil*: este período se caracteriza por uma taxa de falha aproximadamente constante. As falhas ocorrem de forma ocasional, decorrentes de solicitações normais de uso, diferentes combinações de condições de uso, acidentes causados pelo uso incorreto e manutenção inadequada e até debilidades inerentes ao projeto. Este período é caracterizado, nos seres humanos, pela fase intermediária da vida, ou seja, após os primeiros meses de vida até o início do envelhecimento.

3ª) *Período de Desgaste*: apresenta uma taxa de falha crescente devido ao processo natural de envelhecimento ou desgaste do produto. Estas falhas podem ser evitadas por um programa adequado de manutenção preventiva. Nos seres humanos, este período tem início na fase "de envelhecimento" (em geral, na assim denominada, terceira idade).

Distribuições teóricas com função de taxa de falha na forma da apresentada na Figura 3.5 encontram-se apresentadas na literatura. Entretanto, elas são bastante complexas e conseqüentemente difíceis de serem tratadas (ver Nelson, 1990a, p.27).

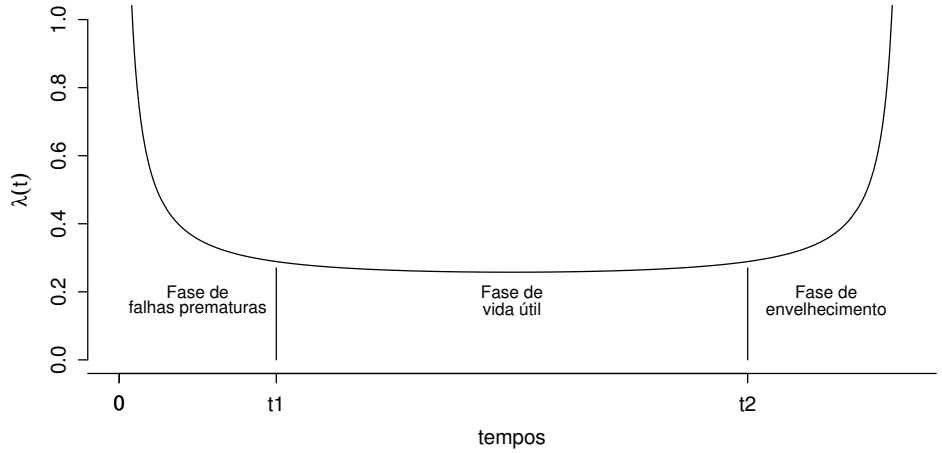

Figura 3.5: Função de taxa de falha do tipo "banheira" e suas três regiões distintas.

Ênfase será dada, neste texto, às distribuições exponencial, de Weibull e log-normal uma vez que, em um contexto prático, elas acomodam grande parte das situações reais. A distribuição gama generalizada, por ser útil na comparação de modelos probabilísticos, e a distribuição gama, por desempenhar um importante papel nos modelos de fragilidade, serão utilizadas, respectivamente, nos Capítulos 4 e 9.

3.3 Estimação dos Parâmetros dos Modelos

Os modelos probabilísticos apresentados na seção anterior são caracterizados por quantidades desconhecidas, denominadas parâmetros. O modelo gama generalizado é caracterizado por três parâmetros, os modelos de Weibull, log-normal e gama por dois parâmetros e, o exponencial, por apenas um. Estas quantidades conferem uma forma geral aos modelos probabilísticos. Entretanto, em cada estudo envolvendo tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais para que o modelo fique determinado e assim seja possível responder às perguntas de interesse.

Existem alguns métodos de estimação conhecidos na literatura estatística. Talvez o mais conhecido seja o método de mínimos quadrados, geralmente apresentado em cursos básicos de estatística dentro do contexto de regressão linear. No entanto, este método é inapropriado para estudos de tempo de vida. A principal razão é a sua incapacidade de incorporar censuras no seu processo de estimação. O método de máxima verossimilhança surge como uma opção apropriada para este tipo de dados. Ele incorpora as censuras, é relativamente simples de ser entendido e possui

propriedades ótimas para grandes amostras. Na Seção 3.3.1 é feita a apresentação do método de máxima verossimilhança para dados censurados.

3.3.1 O Método de Máxima Verossimilhança

O método de máxima verossimilhança trata o problema de estimação da seguinte forma: baseado nos resultados obtidos pela amostra, 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 de Weibull, para cada combinação diferente de γ e α tem-se diferentes distribuições de Weibull. O estimador de máxima verossimilhança escolhe aquele par de γ e α 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 $f(t)$. Por exemplo, se $f(t) = (1/\alpha)\exp(-t/\alpha)$, significa que as observações vêm de uma distribuição exponencial com parâmetro α a ser estimado. A função de verossimilhança para um parâmetro genérico θ desta população é então expressa por

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

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, encontrar 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 de densidade. A contribuição de cada observação censurada não é, contudo, a sua função de 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 $(1, 2, \cdots, r)$, e as $n - r$ seguintes, são as censuradas $(r+1, r+2, \cdots, n)$.

A função de verossimilhança assume assim a seguinte forma

$$L(\theta) = \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta), \tag{3.10}$$

ou eqüivalentemente,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Big[f(t_i;\theta)\Big]^{\delta_i} \Big[S(t_i;\theta)\Big]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \Big[\lambda(t_i;\theta)\Big]^{\delta_i} S(t_i;\theta) \end{aligned} \tag{3.11}$$

em que δ_i é a variável indicadora de falha ou censura apresentada na Seção 1.4. A expressão (3.10), ou (3.11), para a função de verossimilhança é válida para os mecanismos de censura do tipo I e II e sob a suposição de que o mecanismo de censura é não-informativo (não carrega informações sobre os parâmetros). Vale também para o mecanismo de censura do tipo aleatório. Esta suposição é freqüentemente válida em estudos clínicos e industriais. É sempre conveniente, no entanto, trabalhar com o logaritmo da função de verossimilhança. Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$ ou eqüivalentemente $\log(L(\theta))$. Eles são encontrados resolvendo-se o sistema de equações

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

3.3.2 Exemplos de Aplicações

Os cálculos a serem realizados para obter os estimadores de máxima verossimilhança são ilustrados a seguir para as distribuições exponencial e de Weibull. No caso da distribuição de Weibull, não existem expressões fechadas para os estimadores de γ e α, e sendo assim, optou-se por apresentar os passos seguidos pelo método numérico. Neste caso, as estimativas para um conjunto de dados de tempo de vida devem ser obtidas por meio de um pacote estatístico.

Suponha para as situações ilustradas a seguir, uma amostra de n itens em que $r \leq n$ são falhas e os demais, $n - r$, são censuras.

3.3.2.1 Distribuição Exponencial

A função de verossimilhança para a distribuição exponencial, obtida a partir das expressões (3.1) e (3.2) da Seção 3.2.1, é dada por

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} \left[\frac{1}{\alpha} \exp\left\{ -\left(\frac{t_i}{\alpha} \right) \right\} \right]^{\delta_i} \left[\exp\left\{ -\left(\frac{t_i}{\alpha} \right) \right\} \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\alpha} \right]^{\delta_i} \exp\left\{ -\left(\frac{t_i}{\alpha} \right) \right\}. \end{aligned}$$

Tomando-se o logaritmo de $L(\alpha)$ segue que,

$$\log(L(\alpha)) = \sum_{i=1}^{n} \delta_i \log(1/\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} t_i = -\sum_{i=1}^{n} \delta_i \log(\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} t_i$$

e assim,

$$\frac{\partial \log(L(\alpha))}{\partial \alpha} = -\frac{1}{\alpha} \sum_{i=1}^{n} \delta_i + \frac{1}{\alpha^2} \sum_{i=1}^{n} t_i.$$

Igualando-se a última equação a zero e avaliando-a em $\alpha = \hat{\alpha}$, obtém-se o estimador de máxima verossimilhança de α dado por:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} \delta_i} = \frac{\sum_{i=1}^{n} t_i}{r}.$$

O termo $\sum_{i=1}^{n} t_i$ é denominado *tempo total sob teste*. Observe que se todas as observações fossem não-censuradas, $\widehat{\alpha} = \overline{t}$, a média amostral.

3.3.2.1 Distribuição de Weibull

A função de verossimilhança para uma amostra de dados de tempos de vida provenientes de uma distribuição de Weibull é obtida a partir das expressões (3.4) e (3.5) da Seção 3.2.2 e dada por

$$\begin{aligned} L(\gamma, \alpha) &= \prod_{i=1}^{n} \left[\frac{\gamma}{\alpha^{\gamma}} \, t_i^{\gamma-1} \exp\left\{ -\left(\frac{t_i}{\alpha} \right)^{\gamma} \right\} \right]^{\delta_i} \left[\exp\left\{ -\left(\frac{t_i}{\alpha} \right)^{\gamma} \right\} \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{\gamma}{\alpha^{\gamma}} \, t_i^{\gamma-1} \right]^{\delta_i} \exp\left\{ -\left(\frac{t_i}{\alpha} \right)^{\gamma} \right\}. \end{aligned}$$

Assim, o respectivo logaritmo desta função é

$$\begin{aligned}\log(L(\gamma,\alpha)) &= \log\left[\prod_{i=1}^{n}\left[\frac{\gamma}{\alpha^{\gamma}}\,t_i^{\gamma-1}\right]^{\delta_i}\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\gamma}\right\}\right]\\ &= \sum_{i=1}^{n}\delta_i\log(\gamma)-\sum_{i=1}^{n}\delta_i\gamma\log(\alpha)+(\gamma-1)\sum_{i=1}^{n}\delta_i\log(t_i)-\alpha^{-\gamma}\sum_{i=1}^{n}t_i^{\gamma}\\ &= r\log(\gamma)-r\,\gamma\log(\alpha)+(\gamma-1)\sum_{i=1}^{n}\delta_i\log(t_i)-\alpha^{-\gamma}\sum_{i=1}^{n}t_i^{\gamma}.\end{aligned}$$

De forma alternativa, fazendo-se $y_i = \log(t_i)$ e utilizando-se a distribuição do valor extremo tem-se

$$\log(L(\mu,\alpha)) = -r\log(\sigma)+\sum_{i=1}^{n}\delta_i\Big(\frac{y_i}{\sigma}\Big)-\frac{r\mu}{\sigma}-\sum_{i=1}^{n}\exp\left\{\frac{(y_i-\mu)}{\sigma}\right\}$$

que é mais simples do que o logaritmo da função de verossimilhança obtida para a distribuição de Weibull. Derivando $\log(L(\mu,\sigma))$ em relação aos parâmetros μ e σ e igualando as expressões resultantes a zero, obtém-se o seguinte sistema de equações:

$$\begin{aligned}\frac{\partial\log L(\mu,\sigma)}{\partial\mu} &= \frac{1}{\sigma}\left[-r+\sum_{i=1}^{n}\exp\left\{\frac{(y_i-\mu)}{\sigma}\right\}\right]=0\\ \frac{\partial\log L(\mu,\sigma)}{\partial\sigma} &= \frac{1}{\sigma^2}\left[-r\sigma-\sum_{i=1}^{n}\delta_i y_i+r\mu+\sum_{i=1}^{n}\exp\left\{\frac{(y_i-\mu)}{\sigma}\right\}(y_i-\mu)\right]=0.\end{aligned}$$

Os estimadores de máxima verossimilhança são os valores de μ e σ que satisfazem às equações acima. A solução deste sistema de equações para um conjunto de dados particular deve ser obtida por meio de um método numérico como, por exemplo, o de Newton-Raphson. Este método utiliza a matriz de derivadas segundas ($\mathcal{F}$) do logaritmo da função de verossimilhança e a sua expressão

$$\widehat{\theta}_{(k+1)} = \widehat{\theta}_{(k)} - \Big[\mathcal{F}(\widehat{\theta}_{(k)})\Big]^{-1}U(\widehat{\theta}_{(k)})$$

é baseada numa expansão de $U(\widehat{\theta}_{(k)})$ em série de Taylor em torno de $\widehat{\theta}_{(k)}$. Partindo de um valor inicial $\widehat{\theta}_{(0)}$, em que é usual tomar $\widehat{\theta}_{(0)} = 0$, vai-se atualizando este valor a cada passo. Em geral, obtém-se convergência em poucos passos, com erro relativo menor que, por exemplo, 0,001 entre dois passos consecutivos.

Observe que $\mathcal{F}$ para o modelo exponencial é um único número e igual a

$$\begin{aligned}\mathcal{F}(\alpha) &= \frac{\partial^2\log L(\alpha)}{\partial\alpha^2}\\ &= \frac{r}{\alpha^2}-\frac{2\sum_{i=1}^{n}t_i}{\alpha^3}\end{aligned}$$

e $\mathcal{F}(\gamma, \alpha)$, para o modelo de Weibull é uma matriz simétrica 2×2 consistindo dos seguintes elementos:

$$\begin{aligned}
\mathcal{F}_{11}(\gamma, \alpha) &= \frac{\partial^2 \log L(\gamma, \alpha)}{\partial \gamma^2}, \\
\mathcal{F}_{22}(\gamma, \alpha) &= \frac{\partial^2 \log L(\gamma, \alpha)}{\partial \alpha^2}, \\
\mathcal{F}_{12}(\gamma, \alpha) &= \mathcal{F}_{21}(\gamma, \alpha) = \frac{\partial^2 \log L(\gamma, \alpha)}{\partial \gamma \partial \alpha}.
\end{aligned}$$

Informações adicionais sobre o método iterativo de Newton-Raphson podem ser encontradas no Apêndice D deste texto.

3.3.3 Precisão das Estimativas e Intervalos de Confiança

O método de máxima verossimilhança foi utilizado para obter estimadores para os parâmetros do modelo. Estes valores são chamados de estimadores pontuais. Este método também permite a construção de intervalos de confiança para os parâmetros. Isto é feito a partir das propriedades para grandes amostras destes estimadores. As justificativas matemáticas destas propriedades são bastantes complexas e neste texto serão simplesmente apresentadas as propriedades importantes que são suficientes para os objetivos propostos. As provas destas propriedades e maiores informações podem ser encontradas em Cox e Hinkley (1974) e Cordeiro (1992).

A propriedade ou resultado mais importante diz respeito à precisão do estimador de máxima verossimilhança e estabelece que

$$Var(\widehat{\theta}) \approx -\Big[E(\mathcal{F}(\theta))\Big]^{-1}.$$

Ou seja, que a matriz de variância-covariância dos estimadores de máxima verossimilhança é aproximadamente menos a inversa da esperança da matriz de derivadas segundas de $\log L(\theta)$. Em situações em que a esperança é impossível ou difícil de ser calculada, usa-se simplesmente $-\Big[\mathcal{F}(\theta)\Big]^{-1}$. Os elementos da diagonal principal destas matrizes são as variâncias dos estimadores e os outros elementos suas respectivas covariâncias. Geralmente a $Var(\widehat{\theta})$ depende de θ. Uma estimativa para $Var(\widehat{\theta})$ é então obtida substituindo θ por $\widehat{\theta}$.

Na construção de intervalos de confiança é necessário ter uma estimativa para o erro-padrão de $\widehat{\theta}$, isto é, para $\sqrt{Var(\widehat{\theta})}$. No caso especial em que θ é um escalar, um intervalo aproximado de $(1-\alpha)100\%$ de confiança para θ é dado por

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{\theta})}.$$

Por exemplo, um intervalo de 95% de confiança para o parâmetro α do modelo exponencial é dado por

$$\widehat{\alpha} \pm 1,96 \times \sqrt{\frac{\widehat{\alpha}^2}{r}}$$

pois $E\left[\frac{r}{\alpha^2} - \frac{2\sum_{i=1}^{n} t_i}{\alpha^3}\right] = -\frac{r}{\alpha^2}$. No caso em que θ é um vetor de parâmetros, um intervalo de confiança pode ser construído para cada um deles separadamente. Basta obter uma estimativa para o seu erro-padrão a partir da matriz de variância-covariância $Var(\widehat{\theta})$.

Suponha que $\theta = (\gamma, \alpha)$, como no modelo de Weibull. Algumas vezes o interesse é estimar uma função dos parâmetros $\phi = g(\gamma, \alpha)$. Por exemplo, a mediana da Weibull, $t_{0,5} = \alpha[-\log(1-0,5)]^{1/\gamma}$. O estimador de máxima verossimilhança para ϕ é $\widehat{\phi} = g(\widehat{\gamma}, \widehat{\alpha})$. Ou seja, para estimar $\phi = g(\gamma, \alpha)$ basta substituir γ e α por seus respectivos estimadores de máxima verossimilhança. Esta é outra propriedade importante do estimador de máxima verossimilhança. Se além de estimar ϕ, existir interesse em construir um intervalo de confiança, é necessário obter uma estimativa para o erro padrão de ϕ. Isto é feito usando o *método delta* que é descrito a seguir.

Considere inicialmente que θ é um escalar e que há interesse em avaliar a $Var(g(\widehat{\theta}))$. Expandindo $g(\widehat{\theta})$ em torno de $E[\widehat{\theta}] \doteq \theta$ e ignorando os termos superiores ao de primeira ordem tem-se

$$g(\widehat{\theta}) \doteq g(\theta) + (\widehat{\theta} - \theta)\left(\frac{dg(\theta)}{d\theta}\right)$$

e, portanto,

$$Var(g(\widehat{\theta})) \doteq Var(\widehat{\theta})\left(\frac{dg(\theta)}{d\theta}\right)^2.$$

A versão multivariada do *método delta* é necessária para as distribuições que envolvem mais de um parâmetro. Suponha como anteriormente que $\theta = (\gamma, \alpha)$, e que há interesse em $\phi = g(\gamma, \alpha)$. Procedendo de forma similar segue que

$$Var(\widehat{\phi}) \doteq Var(\widehat{\alpha})\left(\frac{\partial\phi}{\partial\alpha}\right)^2 + 2\,Cov(\widehat{\alpha}, \widehat{\gamma})\left(\frac{\partial\phi}{\partial\alpha}\right)\left(\frac{\partial\phi}{\partial\gamma}\right) + Var(\widehat{\gamma})\left(\frac{\partial\phi}{\partial\gamma}\right)^2.$$

3.4 Escolha do Modelo Probabilístico

A escolha do modelo a ser utilizado é um tópico extremamente importante na análise paramétrica de dados de tempo de vida. O método de máxima verossimilhança só

pode ser aplicado após ter sido definido um modelo probabilístico adequado para os dados. Por exemplo, somente após ter definido que o modelo log-normal se ajusta bem aos dados é que o método de máxima verossimilhança pode ser usado para estimar μ e σ. Entretanto, se o modelo log-normal for usado inadequadamente para um certo conjunto de dados, toda a análise estatística fica comprometida e conseqüentemente as respostas às perguntas de interesse ficam distorcidas.

Mas porque usar o modelo log-normal e não o de Weibull? Algumas vezes existem evidências provenientes de testes realizados no passado de que um certo modelo se ajusta bem aos dados. No entanto, em muitas situações este tipo de informação não está disponível. A solução para estas situações é basicamente empírica.

Sabe-se que as distribuições apresentadas na Seção 3.2 são típicas para dados de tempos de vida. A proposta empírica consiste em ajustar os modelos probabilísticos apresentados (exponencial, de Weibull etc.) e, com base na comparação entre valores estimados e observados, decidir qual deles "melhor"explica os dados amostrais. A forma mais simples e eficiente de selecionar o "melhor"modelo a ser usado para um conjunto de dados é através de técnicas gráficas. Entretanto, testes de hipóteses com modelos encaixados também podem ser utilizados para esta finalidade.

A seguir são apresentados dois métodos gráficos e o teste da razão de verossimilhanças para a discriminação de modelos .

3.4.1 Métodos Gráficos

3.4.1.1 Método 1

O primeiro método gráfico a ser apresentado consiste na comparação da função de sobrevivência do modelo proposto com o estimador de Kaplan-Meier. Neste procedimento ajustam-se os modelos propostos ao conjunto de dados (por exemplo, os modelos log-normal e de Weibull) e, a partir das estimativas dos parâmetros de cada modelo, estimam-se suas respectivas funções de sobrevivência, representadas aqui para os modelos log-normal e de Weibull por $\widehat{S}_{ln}(t)$ e $\widehat{S}_{w}(t)$, respectivamente. Para o conjunto de dados, obtém-se também a estimativa de Kaplan-Meier para a função de sobrevivência ($\widehat{S}(t)$).

Finalmente, comparam-se graficamente as funções de sobrevivência estimadas para cada modelo proposto com $\widehat{S}(t)$. O modelo (ou os modelos) adequado será aquele em que sua curva de sobrevivência se aproximar daquela do estimador de Kaplan-Meier. Na prática isto é feito através dos gráficos $\widehat{S}(t)$ *versus* $\widehat{S}_{w}(t)$ e $\widehat{S}(t)$ *versus* $\widehat{S}_{ln}(t)$. Assim, o "melhor"modelo será aquele cujos pontos da função de

sobrevivência estimada estiverem mais próximos dos valores obtidos pelo estimador de Kaplan-Meier. Em outras palavras, o melhor modelo será aquele cujos pontos no gráfico estiverem mais próximos da reta $y = x$, com $x = \widehat{S}(t)$ e $y = \widehat{S}_w(t)$ ou $y = \widehat{S}_{ln}(t)$.

Uma outra forma de comparação é colocar no mesmo gráfico as curvas $\widehat{S}(t)$ *versus* t e $\widehat{S}_w(t)$ *versus* t, por exemplo. Alguns autores, por exemplo Nelson (1990a), sugerem o uso da função de taxa de falha acumulada $\Lambda(t)$, que foi apresentada na Seção 1.6.3. Isto também é feito colocando no mesmo gráfico as curvas $\widehat{\Lambda}(t)$ *versus* t e $\widehat{\Lambda}_w(t)$ *versus* t, por exemplo.

A função de taxa de falha acumulada $\Lambda(t)$ é relacionada com a função de sobrevivência por meio da expressão

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

e sendo assim, uma estimativa para $\Lambda(t)$ é obtida substituindo-se $S(t)$ por sua correspondente estimativa na expressão acima. Exemplificando, nos casos dos modelos de Weibull e log-normal tem-se, respectivamente,

$$\widehat{\Lambda}(t) = -\log(\widehat{S}_w(t)) = \left(\frac{t}{\widehat{\alpha}}\right)^{\widehat{\gamma}}$$

e

$$\widehat{\Lambda}(t) = -\log\left(\Phi\left[-(\log(t) - \widehat{\mu})/\widehat{\sigma}\right]\right).$$

Essencialmente, gráficos envolvendo a função de sobrevivência ou a função de taxa de falha acumulada são úteis para discriminar modelos. A idéia é comparar estas funções com o estimador de Kaplan-Meier e selecionar o modelo cuja curva melhor se aproximar da curva deste último.

3.4.1.2 Método 2

Este método consiste na linearização da função de sobrevivência tendo como idéia básica a construção de gráficos que sejam aproximadamente lineares caso o modelo proposto seja apropriado. Violações da linearidade podem ser rapidamente verificadas visualmente.

O gráfico utilizado é o de uma transformação que lineariza a função de sobrevivência do modelo proposto. Isto gera como resultado final, uma reta se o modelo

proposto for adequado. A seguir são apresentados exemplos de linearização para os modelos exponencial, de Weibull e log-normal.

a) Linearização no modelo exponencial

Para o modelo exponencial, a função de sobrevivência, apresentada na Seção 3.2.1, é dada por:

$$S(t) \;=\; \exp\Big\{-\Big(\frac{t}{\alpha}\Big)\Big\}$$

Assim,

$$-\log\big[S(t)\big] = \frac{t}{\alpha} = \left(\frac{1}{\alpha}\right) t$$

o que mostra que $-\log[S(t)]$ é uma função linear de t. Logo, o gráfico de $-\log[\hat{S}(t)]$ *versus* t deve ser aproximadamente linear, passando pela origem, se o modelo exponencial for apropriado. $\widehat{S}(t)$ é o estimador de Kaplan-Meier.

b) Linearização no modelo de Weibull

A função de sobrevivência para o modelo Weibull de parâmetros $(\gamma,\, \alpha)$ é, como visto anteriormente, dada por:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \qquad t \geq 0.$$

Desse modo,

$$\begin{aligned} -\log\big[S(t)\big] \;&=\; \left(\frac{t}{\alpha}\right)^{\gamma} \\ \log\big[-\log[S(t)]\big] \;&=\; -\gamma\log(\alpha) + \gamma\log(t) \end{aligned}$$

o que mostra que $\log\big[-\log[S(t)]\big]$ é uma função linear de $\log(t)$. Portanto, o gráfico de $\log\big[-\log[\hat{S}(t)]\big]$ *versus* $\log(t)$, sendo $\widehat{S}(t)$ o estimador de Kaplan-Meier, deve ser aproximadamente linear se o modelo Weibull for apropriado. Se além de linear, o gráfico passar pela origem e tiver inclinação igual a 1, é uma indicação a favor do modelo exponencial.

c) Linearização no modelo log-normal

Similarmente, a função de sobrevivência para o modelo log-normal, isto é,

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

pode ser linearizada, e apresenta a seguinte forma

$$\Phi^{-1}\Big(S(t)\Big) = \frac{-\log t + \mu}{\sigma}$$

em que $\Phi^{-1}(\cdot)$ são os percentis da distribuição Normal padrão. Isto significa que o gráfico de $\Phi^{-1}(\hat{S}(t))$ *versus* $\log(t)$ deve ser aproximadamente linear, com intercepto μ/σ e inclinação $-1/\sigma$, se o modelo log-normal for apropriado.

Observe que é possível, a partir desses gráficos, obter estimativas grosseiras para os parâmetros dos modelos. Por exemplo, se o modelo Weibull for adequado pode-se traçar uma reta no gráfico de $\log[-\log(\widehat{S}(t))]$ *versus* $\log(t)$. A inclinação desta reta é uma estimativa para γ e o intercepto para $\gamma\log(\alpha)$. De forma análoga, obtém-se estimativas para μ e σ no modelo log-normal e para α no modelo exponencial. Entretanto, a forma mais indicada para se obter estimativas para os parâmetros, após selecionar o modelo, é utilizar o método de máxima verossimilhança.

Mesmo sendo estes modelos típicos para dados de tempo de vida, podem ocorrer situações em que nenhum deles seja adequado. Estas situações exigem modelos paramétricos mais flexíveis, envolvendo mais que dois parâmetros como, por exemplo, o modelo gama generalizado, ou simplesmente uma análise estatística toda baseada em técnicas não-paramétricas, como aquelas apresentadas no Capítulo 2. Dentre os pacotes estatísticos disponíveis no mercado, não muitos contudo são capazes de ajustar um modelo com mais de dois parâmetros. O *SAS*, por exemplo, é um dos pacotes aptos a ajustar a distribuição gama generalizada.

Existem ainda outras situações em que os gráficos apresentados não discriminam os modelos mas indicam que eles são igualmente bons. A principal razão deste fato se deve aos tamanhos de amostra pequenos ou eqüivalentemente, um número pequeno de falhas. Na prática isto significa que as conclusões serão similares ao se usar um ou outro modelo, podendo apresentar alguma diferença na cauda das distribuições.

3.4.2 Teste da Razão de Verossimilhança

Como foi dito anteriormente, as técnicas gráficas são extremamente úteis na seleção de modelos. Entretanto, as conclusões a partir delas podem diferir para diferentes analistas. Ou seja, existem nas técnicas gráficas um componente subjetivo na sua interpretação. Outra forma de discriminar modelos é através de testes de hipóteses. Neste caso, a conclusão é direta e portanto, não envolve qualquer componente subjetivo na sua interpretação.

As hipóteses a serem testadas são:

$$H_0 : \text{O modelo de interesse é adequado}$$

versus uma hipótese alternativa vaga, de que o modelo não é adequado.

Este teste é usualmente realizado utilizando a estatística da razão de verossimilhanças em modelos encaixados (Cox e Hinkley, 1977). Isto significa que deve ser identificado um modelo generalizado tal que os modelos de interesse são casos particulares. O teste é realizado a partir dos seguintes dois ajustes: (1) modelo generalizado e obtenção do valor do logaritmo de sua função de verossimilhança ($\log L(\widehat{\theta}_G)$); (2) modelo de interesse e obtenção do valor do logaritmo de sua função de verossimilhança ($\log L(\widehat{\theta}_M)$). A partir destes valores é possível calcular a estatística da razão de verossimilhanças, isto é,

$$\text{TRV} = -2\log\left[\frac{L(\widehat{\theta}_M)}{L(\widehat{\theta}_G)}\right] = 2\big[\log L(\widehat{\theta}_G) - \log L(\widehat{\theta}_M)\big]$$

que, sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros ($\widehat{\theta}_G$ e $\widehat{\theta}_M$) dos modelos sendo comparados.

No contexto de análise de sobrevivência, este teste é usualmente realizado utilizando a distribuição gama generalizada que apresenta os modelos exponencial, de Weibull, log-normal e gama, como modelos encaixados uma vez que todos eles, como visto anteriormente, são casos especiais da gama generalizada.

3.5 Exemplos

As técnicas estatísticas apresentadas neste capítulo são aplicadas nesta seção a dois conjuntos de dados. O primeiro refere-se ao tempo de reincidência de um grupo de pacientes com câncer de bexiga submetidos a um procedimento cirúrgico feito por laser e, o segundo, ao tempo até os primeiros sinais de alterações no estado de saúde de um grupo de pacientes submetidos à quimioterapia após cirurgia de intestino.

3.5.1 Exemplo 1

Neste exemplo são considerados os tempos de reincidência, em meses, de um grupo de 20 pacientes com câncer de bexiga que foram submetidos a um procedimento cirúrgico feito por laser. Os tempos obtidos foram: 3, 5, 6, 7, 8, 9, 10, 10^+, 12, 15, 15^+, 18, 19, 20, 22, 25, 28, 30, 40, 45^+ em que o símbolo + indica censura.

Para este exemplo, as expressões das estimativas das funções de sobrevivência para os modelos exponencial, de Weibull e log-normal são, respectivamente,

$$\begin{aligned}\widehat{S}(t) &= \exp\{-t/20,41\}\\ \widehat{S}(t) &= \exp\{-(t/21,34)^{1,54}\}\\ \widehat{S}(t) &= \Phi[-(\log(t)-2,72)/0,76].\end{aligned}$$

Os valores que aparecem nas expressões apresentadas são as estimativas de máxima verossimilhança dos parâmetros de cada um dos modelos. Estas estimativas podem ser obtidas no pacote estatístico R por meio dos comandos

```
> require(survival)
> tempos<-c(3,5,6,7,8,9,10,10,12,15,15,18,19,20,22,25,28,30,40,45)
> cens<-c(1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,1,1,1,0)
> ajust1<-survreg(Surv(tempos,cens)~1,dist='exponential')
> ajust1
> alpha<-exp(ajust1$coefficients[1])
> alpha
> ajust2<-survreg(Surv(tempos,cens)~1,dist='weibull')
> ajust2
> alpha<-exp(ajust2$coefficients[1])
> gama<-1/ajust2$scale
> cbind(gama, alpha)
> ajust3<-survreg(Surv(tempos,cens)~1,dist='lognorm')
> ajust3
```

O valor destas funções em, por exemplo, $t = 10$ meses pode, então, ser calculado e fornecem, respectivamente,

$$\begin{aligned}\widehat{S}(t) &= \exp\{-10/20,41\} = 0,612\\ \widehat{S}(t) &= \exp\{-(10/21,34)^{1,54}\} = 0,732\\ \widehat{S}(t) &= \Phi[-(\log(10)-2,72)/0,76] = 0,708.\end{aligned}$$

Observe que as estimativas obtidas pelos modelos de Weibull e log-normal são bem próximas. O mesmo não é observado para o modelo exponencial, que apresenta um valor estimado ligeiramente diferente dos obtidos para os outros dois modelos.

A Tabela 3.1 mostra as estimativas das funções de sobrevivência para os tempos de reincidência usando os modelos exponencial, de Weibull e log-normal e também o Kaplan-Meier. Os comandos utilizados no R para obtenção das estimativas foram:

```
> time<-ekm$time
> st<-ekm$surv
> ste<- exp(-time/20.41)
> stw<- exp(-(time/21.34)^1.54)
> stln<- pnorm((-log(time)+ 2.72)/0.76)
> cbind(time,st,ste,stw,stln)
```

Tabela 3.1: Estimativas da sobrevivência para os tempos de reincidência usando o estimador de Kaplan-Meier e os modelos exponencial, de Weibull e log-normal.

Tempos	Kaplan-Meier	Exponencial	Weibull	Log-normal
3	0,950	0,863	0,952	0,983
5	0,900	0,782	0,898	0,928
6	0,850	0,745	0,867	0,889
7	0,800	0,709	0,835	0,845
8	0,750	0,675	0,801	0,800
9	0,700	0,643	0,767	0,754
10	0,650	0,612	0,732	0,708
12	0,595	0,555	0,662	0,621
15	0,541	0,479	0,559	0,506
18	0,481	0,413	0,463	0,411
19	0,421	0,394	0,433	0,383
20	0,361	0,375	0,404	0,358
22	0,300	0,340	0,350	0,312
25	0,240	0,293	0,279	0,255
28	0,180	0,253	0,218	0,210
30	0,120	0,229	0,184	0,185
40	0,060	0,140	0,071	0,101
45	0,060	0,110	0,042	0,076

Para a escolha de um dos modelos, utilizou-se inicialmente do primeiro método gráfico apresentado na Seção 3.4.1. Foram então construídos os gráficos das estimativas das sobrevivências obtidas pelo método de Kaplan-Meier *versus* as estimativas das sobrevivências obtidas a partir dos modelos exponencial, de Weibull e log-normal, respectivamente. Estes gráficos encontram-se apresentados na Figura 3.6 e foram obtidos com o auxílio do *R* por meio dos comandos

```
> par(mfrow=c(1,3))
> plot(ste,st,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)), ylab = "S(t): Kaplan-Meier",
> xlab="S(t): exponencial")
> lines(c(0,1), c(0,1), type="l", lty=1)
> plot(stw,st,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)), ylab = "S(t): Kaplan-Meier",
> xlab="S(t): Weibull")
> lines(c(0,1), c(0,1), type="l", lty=1)
> plot(stln,st,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)), ylab = "S(t): Kaplan-Meier",
> xlab="S(t): log-normal")
> lines(c(0,1), c(0,1), type="l", lty=1)
```

A partir dos gráficos apresentados na Figura 3.6, é possível observar que o modelo exponencial parece não ser o mais adequado para estes dados pois a curva se apresenta um tanto afastada da reta $y = x$. Por outro lado, os modelos de Weibull e log-normal acompanham mais de perto a reta $y = x$, indicando ser um desses modelos, possivelmente, adequado para os dados sob estudo.

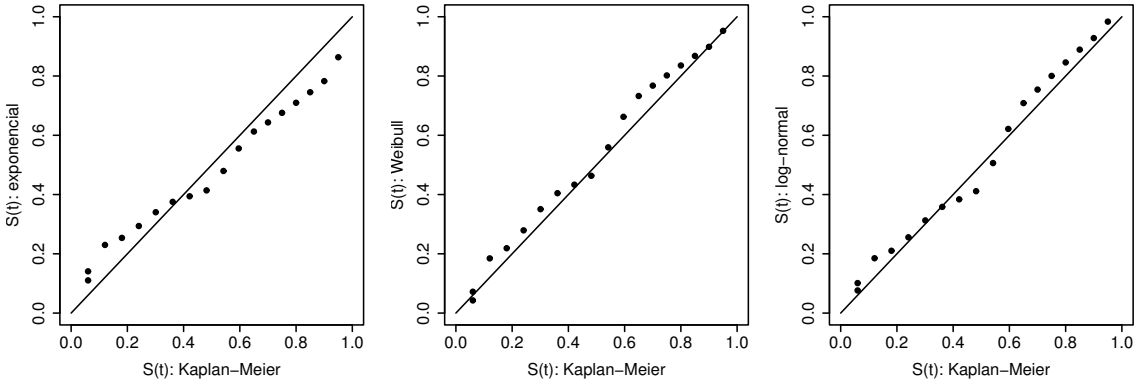

Figura 3.6: Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as sobrevivências estimadas pelos modelos exponencial, de Weibull e log-normal.

Na tentativa de confirmar os resultados obtidos pelo método 1, foram construídos os gráficos linearizados (método 2) para os modelos exponencial, de Weibull e log-normal. Eles estão mostrados na Figura 3.7 e foram obtidos no R com o auxílio dos comandos:

```
> par(mfrow=c(1,3))
> invst<-qnorm(st)
> plot(time, -log(st),pch=16,xlab="tempos",ylab="-log(S(t))")
> plot(log(time),log(-log(st)),pch=16,xlab="log(tempos)",ylab="log(-log(S(t)))")
> plot(log(time),invst, pch=16,xlab="log(tempos)", ylab=expression(Phi^-1 * (S(t))))
```

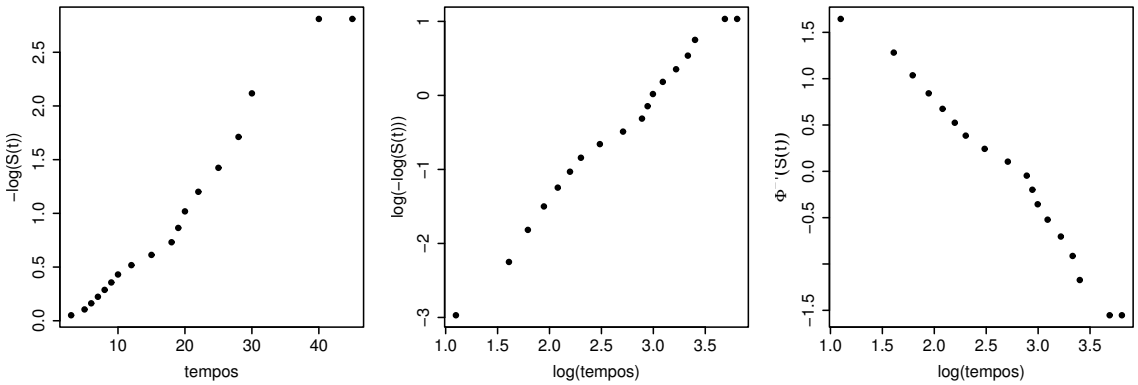

Figura 3.7: Gráficos de t *vs* $-\log(\widehat{S}(t))$, $\log(t)$ *vs* $\log(-\log(\widehat{S}(t)))$ e $\log(t)$ *vs* $\Phi^{-1}(\widehat{S}(t))$.

Os gráficos para os modelos de Weibull e log-normal apresentados na Figura 3.7 não mostram afastamentos marcantes de uma reta. Já para o modelo exponencial observa-se um certo desvio da reta. Esses gráficos confirmam os resultados observados quando do uso do método 1 e indicam os modelos de Weibull e log-normal a serem usados na análise dos dados. Os dois modelos indicados pelos procedimentos gráficos devem apresentar, como comentado na Seção 3.4.1, resultados similares e igualmente bons.

Os testes da razão de verossimilhança para as hipóteses de que: i) o modelo exponencial é adequado, ii) o modelo de Weibull é adequado e iii) o modelo log-normal é adequado, foram realizados utilizando-se o modelo gama generalizado. Os valores do logaritmo da função de verossimilhança para os quatro modelos e os testes da razão de verossimilhança (TRV) resultaram nos valores apresentados na Tabela 3.2.

Tabela 3.2: Logaritmo da função de verossimilhança e resultados dos TRV.

Modelo	$\log(L(\theta))$		TRV	valor p
Gama Generalizado	$\log(L(\gamma, k, \alpha))$	= -65,69*	-	-
Exponencial	$\log(L(\alpha))$	= -68,27	2(68,27 - 65,69) = 5,16	0,075
Weibull	$\log(L(\gamma, \alpha))$	= -66,13	2(66,13 - 65,69) = 0,88	0,348
Log-normal	$\log(L(\mu, \sigma))$	= -65,74	2(65,74 - 65,69) = 0,10	0,752

* valor obtido com o auxílio do pacote estatístico *SAS*.

Os resultados apresentados na Tabela 3.2, em que os valores do logaritmo das funções de verossimilhança foram obtidos com o auxílio do pacote estatístico *SAS* [1] para a distribuição gama generalizada (*R* ainda não disponibiliza a gama generalizada no procedimento *survival*), e do pacote *R*, com os comandos

```
> ajust1$loglik[2]
> ajust2$loglik[2]
> ajust3$loglik[2]
```

para as demais distribuições, indicam a adequação dos modelos de Weibull e log-normal para a análise dos dados desse exemplo, confirmando as conclusões apresentadas quando da utilização das técnicas gráficas. As curvas de sobrevivência estimadas por meio do ajuste de ambos os modelos *versus* a curva de sobrevivência estimada por Kaplan-Meier podem ser observadas na Figura 3.8. Note, a partir

[1] Cabe observar que o *SAS* fornece por *default* o logaritmo da função de verossimilhança construída para o log(tempos). O *R*, por sua vez, fornece o logaritmo da função de verossimilhança construída para os tempos.

desta figura, que ambos os modelos apresentam ajustes satisfatórios. Os comandos utilizados no R para obtenção desta figura foram:

```
> par(mfrow=c(1,2))
> plot(ekm, conf.int=F, xlab="Tempos", ylab="S(t)")
> lines(c(0,time),c(1,stw), lty=2)
> legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Weibull"),bty="n",cex=0.8)
> plot(ekm, conf.int=F, xlab="Tempos", ylab="S(t)")
> lines(c(0,time),c(1,stln), lty=2)
> legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Log-normal"),bty="n",cex=0.8)
```

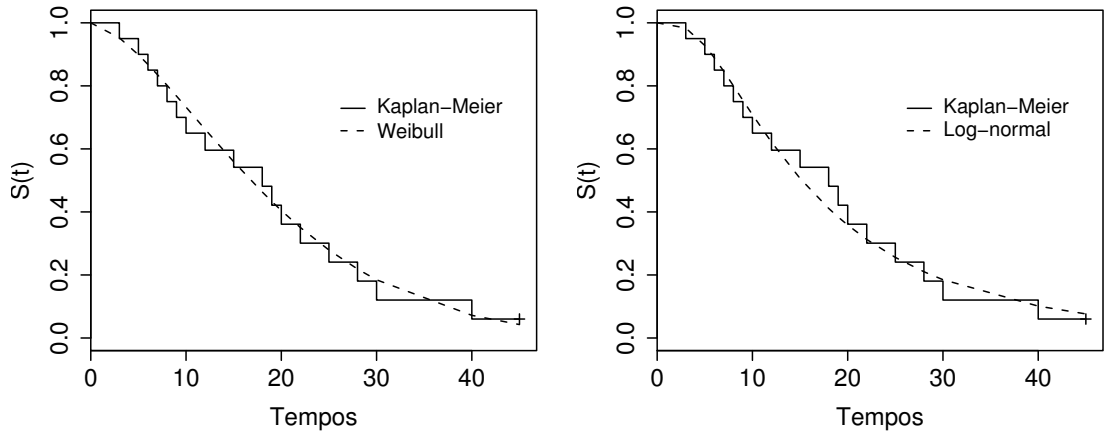

Figura 3.8: Curvas de sobrevivência estimadas pelos modelos de Weibull e log-normal *versus* a curva de sobrevivência estimada por Kaplan-Meier.

Estimativas para o tempo médio, com base nas distribuições de Weibull e log-normal, são calculadas a partir das expressões da média apresentadas nas Seções 3.2.2 e 3.2.3. Desta forma, tem-se, respectivamente, para o modelo de Weibull e log-normal, as estimativas:

$$\begin{aligned} \widehat{E}(T) &= 21,34[\Gamma(1+(1/1,54))] = 19,206 \text{ meses} \\ \widehat{E}(T) &= \exp\left\{2,72+(0,76^2/2)\right\} = 20,263 \text{ meses.} \end{aligned}$$

Intervalos de confiança para $E[T]$ podem ser obtidos após obtenção de estimativas para a $Var(\widehat{E}(T))$. Isto é feito a partir do método delta apresentado na Seção 3.3.3. Para o modelo log-normal, por exemplo, tem-se $\widehat{V}ar(\hat{\mu}) = 0{,}031$, $\widehat{V}ar(\hat{\sigma}) = 0{,}0176$ e $\widehat{C}ov(\hat{\mu},\hat{\sigma}) = 0{,}00207$, de modo que

$$\begin{aligned}\widehat{Var}(\widehat{E}(T)) &\doteq \widehat{V}ar(\hat{\mu})\left[\exp\left\{\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right\}\right]^2 + \widehat{V}ar(\hat{\sigma})\left[\hat{\sigma}\exp\left\{\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right\}\right]^2 \\ &+ 2\,\widehat{C}ov(\hat{\mu},\hat{\sigma})\left[\exp\left\{\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right\}\right]\left[\hat{\sigma}\exp\left\{\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right\}\right] \\ &= (0,031)(20,263)^2 + (0,0176)((0,76)*(20,263))^2 \\ &+ 2(0,00207)(0,76)(20,263)^2 = 18,2.\end{aligned}$$

Tem-se, assim, um intervalo de 95% de confiança para $E[T]$ de $(11,90; 28,62)$ meses. Ainda, uma estimativa para o tempo mediano, obtida a partir da expressão dos percentis, é

$$\widehat{t}_{0,5} = \exp(z_{0,5}0,76 + 2,72) = 15,18 \text{ meses.}$$

O estimador de Kaplan-Meier, usando interpolação linear, fornece um valor de 17,05 meses como uma estimativa para o tempo mediano bem como uma estimativa para o tempo médio de reicindência, embora subestimada pois a última unidade foi censurada, de 18,43 meses. Observe para, por exemplo, $t = 20$ meses, que uma estimativa para $S(t)$ usando o modelo log-normal é de 35,5%. Essa mesma estimativa usando o estimador de Kaplan-Meier é de 36,1%. Estes valores são bastante próximos e significa que um paciente tem uma probabilidade de cerca de 36% de estar livre de reincidência após 20 meses da realização do procedimento cirúrgico.

3.5.2 Exemplo 2

No estudo analisado neste exemplo, são apresentados na Tabela 3.3 os tempos, em dias, até a ocorrência dos primeiros sinais de alterações indesejadas no estado geral de saúde de 45 pacientes de ambos os sexos que receberam tratamento quimioterápico após terem sido submetidos à cirurgia de intestino. Foi registrado um total de 250 dias desde a entrada do primeiro paciente até o término do estudo.

Tabela 3.3: Tempos até a ocorrência dos primeiros sinais de alterações pós-cirúrgicas de pacientes que receberam tratamento quimioterápico após cirurgia de intestino.

$7, 8, 10, 12, 13, 14^+, 19, 23, 25^+, 26, 27, 31, 31^+, 49, 59^+, 64^+, 87, 89, 107, 117, 119, 130, 148,$ $153, 156, 159, 191, 222, 200^+, 203^+, 210^+, 220^+, 220^+, 228^+, 230^+, 233^+, 235^+, 240^+, 240^+,$ $240^+, 241^+, 245^+, 247^+, 248^+, 250^+$

(+ indica censura)

Na tentativa de escolher entre os modelos exponencial, de Weibull e log-normal, utilizou-se o método gráfico 2. Os gráficos das linearizações correspondentes aos três modelos encontram-se apresentados na Figura 3.9 e indicam que o modelo log-normal é o que parece apresentar desvios menos acentuados de uma reta sendo, desse modo, o mais adequado, dentre os três modelos analisados, para a análise deste conjunto de dados.

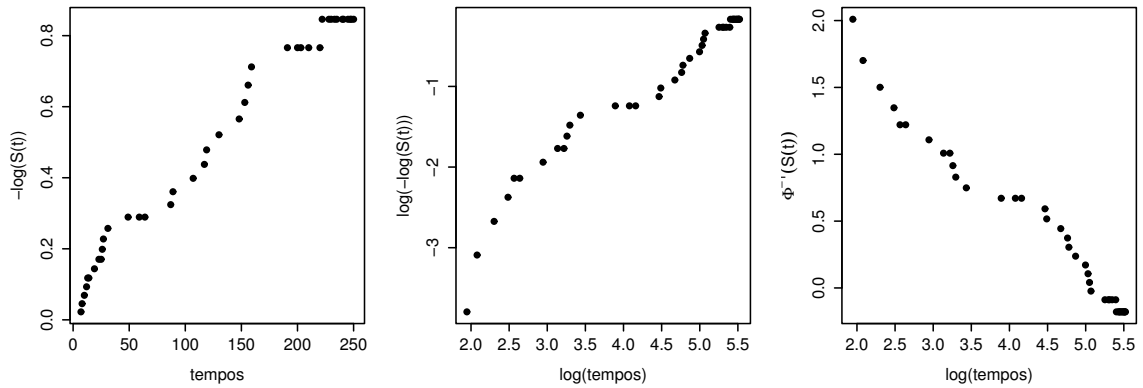

Figura 3.9: Gráficos de t *vs* $-\log(\widehat{S}(t))$, $\log(t)$ *vs* $\log(-\log(\widehat{S}(t)))$ e $\log(t)$ *vs* $\Phi^{-1}(\widehat{S}(t))$.

Os resultados dos testes da razão de verossimilhança (TRV) apresentados na Tabela 3.4, confirmam a indicação do modelo log-normal, obtida no procedimento gráfico, como o mais adequado para a análise desses dados. Note, contudo, que os modelos exponencial e de Weibull não foram totalmente descartados.

Tabela 3.4: Logaritmo da função de verossimilhança e resultados dos TRV.

Modelo	$\log(L(\theta))$		TRV	valor p
Gama Generalizada	$\log(L(\gamma, k, \alpha))$	= 149,66*	-	-
Exponencial	$\log(L(\alpha))$	= 151,07	2(151,07 - 149,66) = 2,82	0,24
Weibull	$\log(L(\gamma, \alpha))$	= 150,55	2(150,55 - 149,66) = 1,78	0,18
Log-normal	$\log(L(\mu, \sigma))$	= 149,81	2(149,81 - 149,66) = 0,30	0,58

* valor obtido com o auxílio do pacote estatístico *SAS*.

A partir da Figura 3.10, que mostra as curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo log-normal, e considerando a existência de uma quantidade considerável de censuras observadas neste exemplo (em torno de 50%), pode-se notar que o modelo indicado apresenta-se razoável para a análise dos dados deste estudo. Assim, uma estimativa para o tempo médio, encontrada a partir da expressão da média do modelo log-normal, é de

$$\widehat{E}(T) = \exp\left\{5,181 + (1,724^2/2)\right\} \;=\; 786 \text{ dias.}$$

A estimativa para o tempo mediano é obtida a partir da expressão dos percentis e fornece um valor de

$$\widehat{t}_{0,5} = \exp\left\{z_{0,5}1,724 + 5,181\right\} = \; 178 \text{ dias.}$$

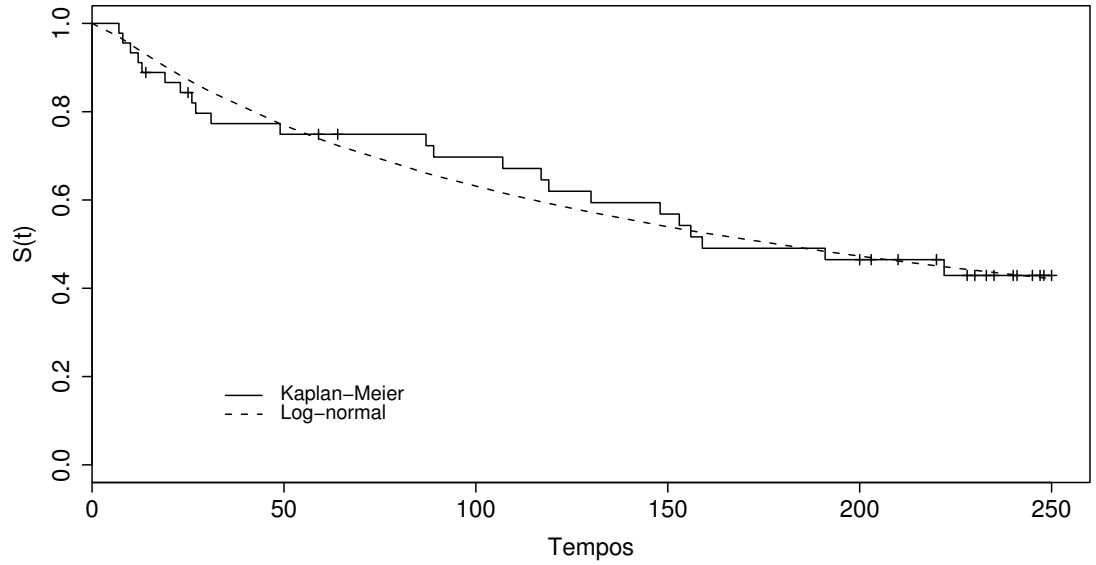

Figura 3.10: Curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo log-normal para os dados dos pacientes submetidos à cirurgia de intestino e quimioterapia.

O estimador de Kaplan-Meier fornece um valor de 158 dias como uma estimativa do tempo mediano e não permite a obtenção de uma estimativa adequada para o tempo médio de vida, pois os últimos pacientes apresentaram tempos de censura.

Uma estimativa do percentual de pacientes sem nenhum sinal de alterações indesejadas no seu estado de saúde em, por exemplo, $t = 200$ dias pode então ser obtida usando a expressão do modelo log-normal, isto é,

$$\widehat{S}(t) \;\; = \;\; \Phi[-(\log(t) - 5,181)/1,724]$$

que fornece o valor de 47,3%. Esta mesma estimativa obtida pelo estimador de Kaplan-Meier fornece o valor de 46,5%. Assim, um paciente que é submetido à quimioterapia após cirurgia do intestino, apresenta uma probabilidade de cerca de 47% de estar sem alterações indesejáveis em seu estado de saúde após 200 dias da cirurgia e início da quimioterapia.

3.6 Exercícios

1. O tempo em dias para o desenvolvimento de tumor em ratos expostos a uma substância cancerígena segue uma distribuição de Weibull

$$S(t) = \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)^{\gamma}\Big\},$$

com $\alpha = 100$ e $\gamma = 2$.

   (a) Qual é a probabilidade de um rato sobreviver sem tumor aos primeiros 30 dias? E aos primeiros 45 dias?

   (b) Qual é o tempo médio até o aparecimento do tumor?

   (c) Qual é o tempo mediano até o aparecimento do tumor?

   (d) Ache a taxa de falha de aparecimento de tumor aos 30, 45 e 60 dias. Interprete estes valores.

2. Deseja-se comparar duas populações de tempos de vida. Uma amostra de tamanho n ($r \leq n$ falhas) foi obtida da população 1 que tem distribuição exponencial com média α. Uma amostra de tamanho m ($s \leq m$ falhas) foi obtida da população 2 que tem distribuição exponencial com média $\alpha + \Delta$.

   (a) Estabeleça as hipóteses que se deseja testar.

   (b) Apresente a função de verossimilhança para $\theta = (\alpha, \Delta)^{'}$.

   (c) Apresente o vetor escore ($U(\theta)$) e a matriz de informação observada ($\mathcal{F}(\theta)$).

   (d) Obtenha as expressões dos testes de Wald e da razão de verossimilhanças para as hipóteses apresentadas em (a).

3. Os dados mostrados a seguir representam o tempo até a ruptura de um tipo de isolante elétrico sujeito a uma tensão de estresse de 35 Kvolts. O teste consistiu em deixar 25 destes isolantes funcionando até que 15 deles falhassem (censura do tipo II) obtendo os seguintes resultados (em minutos):

0,19	0,78	0,96	1,31	2,78	3,16	4,67	4,85
6,50	7,35	8,27	12,07	32,52	33,91	36,71	

   Este exercício foi proposto no Capítulo 2 para ser resolvido utilizando métodos não-paramétricos. O que se deseja aqui é que o exercício seja repetido utilizando modelos paramétricos. Inicialmente deve-se identificar um modelo

paramétrico para explicar estes dados e em seguida responder novamente às mesmas perguntas. Isto é, a partir destes dados amostrais, deseja-se obter as seguintes informações:

(a) Uma estimativa para o tempo mediano de vida deste tipo de isolante elétrico funcionando a 35 Kvolts.

(b) Uma estimativa (pontual e intervalar) para a fração de defeituosos esperada nos dois primeiros minutos de funcionamento?

(c) Uma estimativa (pontual e intervalar) para o tempo médio de vida destes isoladores funcionando a 35 Kvolts.

(d) O tempo necessário para 20% dos isolantes estarem fora de operação.

4. O fabricante de um tipo de isolador elétrico quer conhecer o comportamento de seu produto funcionando na temperatura de 200^{o}C. Um teste de vida foi realizado nestas condições usando 60 isoladores elétricos. O teste terminou quando 45 deles haviam falhado (censura do tipo II). As 15 unidades que não haviam falhado ao final do teste foram desta forma, censuradas no tempo $t = 2729$ horas. O fabricante tem interesse em estimar o tempo médio e mediano de vida do isolador e o percentual de falhas após 500 horas de uso. Os tempos (em horas) obtidos encontram-se apresentados na Tabela 3.5.

Tabela 3.5: Tempos (horas) dos isolantes elétricos funcionando a 200^{o}C.

151	164	336	365	403	454	455	473	538	577	592	628	632	647	675	727	785
801	811	816	867	893	930	937	976	1008	1040	1051	1060	1183	1329	1334		
1379	1380	1633	1769	1827	1831	1849	2016	2282	2415	2430	2686	2729				
2729^+	2729^+	2729^+	2729^+	2729^+	2729^+	2729^+	2729^+	2729^+	2729^+							
2729^+	2729^+	2729^+	2729^+	2729^+												

Responda às questões de interesse do fabricante usando o modelo exponencial, de Weibull ou log-normal, aquele que se apresentar mais apropriado para descrever os dados.

Capítulo 4

Modelos de Regressão em Análise de Sobrevivência

4.1 Introdução

Os estudos na área médica muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Por exemplo, a contagem de CD4 e CD8 ao diagnóstico são duas covariáveis que a literatura médica mostra serem importantes fatores de prognóstico para o tempo até a ocorrência de AIDS em pacientes infectados pelo HIV. Certamente, estas covariáveis devem ser incluídas na análise estatística dos dados. As técnicas não-paramétricas apresentadas no Capítulo 2 não permitem a inclusão de covariáveis na análise. Estas técnicas são importantes para descrever os dados de sobrevivência pela sua simplicidade e facilidade de aplicação, pois não envolvem nenhuma estrutura paramétrica. No entanto, este fato inviabiliza uma análise mais elaborada incluindo covariáveis.

Uma forma simples de fazer isto é dividir os dados em estratos de acordo com estas covariáveis e usar as técnicas não-paramétricas apresentadas no Capítulo 2. A simplicidade dos cálculos e a facilidade de entendimento são as grandes vantagens da análise estratificada. No entanto, ela apresenta sérias limitações. A mais importante é que uma análise envolvendo várias covariáveis gera um número muito grande de estratos que podem conter poucas, ou talvez nenhuma, observações. Isto faz com que as comparações fiquem impossíveis de serem realizadas.

A forma mais eficiente de acomodar o efeito destas covariáveis é utilizar um modelo de regressão apropriado para dados censurados. Em análise de sobrevivência existem duas classes de modelos propostas na literatura: os modelos paramétricos

e os semi-paramétricos. Os modelos paramétricos, também chamados de modelos de tempo de vida acelerado, são mais eficientes, porém menos flexíveis do que os modelos semi-paramétricos. A segunda classe de modelos, também chamada simplesmente de modelo de regressão de Cox, tem sido bastante utilizada em estudos clínicos. Além da flexibilidade, este modelo permite incorporar facilmente covariáveis dependentes do tempo, que ocorrem com freqüência em várias áreas de aplicação. O modelo de regressão de Cox é tratado em detalhes no Capítulo 5.

4.2 Modelo Linear para Dados de Sobrevivência

Considere uma situação simples de modelagem envolvendo uma única covariável em que o objetivo seja explorar a relação entre a covariável e a resposta, que é o tempo até a ocorrência de um evento de interesse. Um gráfico de dispersão entre esta covariável e a resposta pode auxiliar na detecção de uma possível associação entre elas. Outras análises descritivas para explorar esta relação podem também ser realizadas utilizando as técnicas apresentadas no Capítulo 2. Por exemplo, a covariável pode ser estratificada e um estimador de Kaplan-Meier pode ser construído para cada estrato. Como foi dito anteriormente, esta análise é limitada e nesta seção será explorado a utilização de um modelo estatístico para explicar esta relação.

O modelo de regressão linear (Draper e Smith, 1998) é o mais conhecido em estatística e será tomado como ponto de partida. Neste modelo a resposta é associada com as variáveis explicativas ou covariáveis através de um modelo linear. No caso de uma única covariável, o gráfico desta versus a resposta deve mostrar evidências de uma relação linear, caso o modelo seja aceitável para esta situação. Ou seja, a nuvem de pontos deste gráfico deve dar indicações de que uma reta é uma boa aproximação para a relação entre as variáveis. A equação da reta é o componente sistemático do modelo de regressão e a variação em torno desta reta é representada pelo componente estocástico. No caso do modelo linear, este último componente segue uma distribuição normal. A representação deste modelo é a seguinte

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{4.1}$$

em que Y é a resposta, x é a covariável, β_0 e β_1 são os parâmetros a serem estimados e ϵ é o erro aleatório com distribuição normal.

Retornando à situação de interesse, em que se tem uma resposta envolvendo o tempo até a ocorrência de um evento e a presença de censura, o que se deseja é utilizar um modelo de regressão para estudar a relação entre as variáveis. No

entanto, o tipo de resposta e o comportamento das variáveis não permitem, em geral, a utilização direta do modelo (4.1). Junte-se a isto o fato de que a distribuição da resposta tende também, em geral, a ser assimétrica na direção dos maiores tempos de sobrevivência, o que torna inapropriado o uso da distribuição normal para o componente estocástico do modelo.

Existem duas formas de enfrentar o problema da modelagem estatística em análise de sobrevivência. São elas:

1. transformar a resposta para tentar retornar ao modelo linear-normal ou,

2. utilizar um componente sistemático não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico.

Na verdade as duas formas são eqüivalentes. Utilizar um modelo linear para a transformação logarítmica da resposta é eqüivalente a usar o seguinte componente sistemático

$$\exp\{\beta_0 + \beta_1 x\} \tag{4.2}$$

e distribuição log-normal para o erro. Existem, no entanto, outras distribuições assimétricas possíveis para o erro, que não possibilitam o retorno para o modelo linear. Nas Seções 4.2.1 a 4.2.3 são descritos alguns modelos paramétricos usuais que apresentam distribuições assimétricas para o erro.

4.2.1 Modelo de Regressão Exponencial

A utilização da distribuição exponencial para o erro e um componente sistemático da forma (4.2) é certamente o modelo de regressão mais simples e historicamente mais utilizado na literatura de análise de sobrevivência. Este modelo, envolvendo uma única covariável, será utilizado para introduzir a modelagem de uma situação simples em análise de sobrevivência.

A combinação de um componente sistemático e uma distribuição exponencial com média unitária ($f(\epsilon) = \exp\{-\epsilon\}$) para o erro gera o seguinte modelo:

$$T = \exp\{\beta_0 + \beta_1 x\}\ \epsilon \tag{4.3}$$

que é o modelo de regressão exponencial. Este modelo admite uma relação não-linear entre T e x no seu componente sistemático e erro com distribuição assimétrica. Na linguagem de modelos lineares generalizados (McCullagh e Nelder, 1989), tem-se uma função de ligação logarítmica e a resposta com distribuição exponencial.

Observe que o modelo (4.3) é linearizável se for considerado o logaritmo de T. Assim, obtém-se

$$Y = \log(T) = \beta_0 + \beta_1 x + \nu, \tag{4.4}$$

com $\nu = \log(\epsilon)$. O modelo (4.4) é semelhante ao modelo linear, com exceção da distribuição dos erros que não é normal. O erro ν segue uma distribuição do valor extremo padrão ($f(\nu) = \exp\{\nu - \exp\{\nu\}\}$. Esta distribuição é bastante utilizada em análise de sobrevivência, pois caracteriza de forma adequada a distribuição do logaritmo de certos tempos de vida. Mais informações sobre esta distribuição podem ser encontradas em Lawless (1982).

Note de (4.4) e (4.3), que x atua linearmente em Y e, então, multiplicativamente em T. Ainda, a função de sobrevivência para Y condicional a x é, para este modelo, expressa por

$$S(y \mid x) = \exp\left\{ - \exp\Big\{y - (\beta_0 + \beta_1 x)\Big\}\right\}$$

e, para $T = \exp\{Y\}$ dado x, por:

$$S(t \mid x) \;\; = \;\; \exp\left\{ - \left(\frac{t}{\exp\{\beta_0 + \beta_1 x\}} \right)\right\}. \tag{4.5}$$

O passo seguinte, após a especificação do modelo, é a estimação dos seus parâmetros. No caso particular do modelo (4.4) é necessário estimar e fazer inferência sobre os parâmetros β_0 e β_1. No modelo linear utiliza-se o método de mínimos quadrados para esta finalidade, pois ele tem propriedades desejáveis na presença de erros com distribuição normal (Seber, 1977). Na ausência de normalidade dos erros e, principalmente, na presença de censuras este método se torna inadequado. O método da máxima verossimilhança, discutido no Capítulo 3, se apresenta então como uma opção apropriada.

A construção da função de verossimilhança, como visto anteriormente, é dividida em duas partes separadas, correspondentes às falhas e censuras. No caso de falhas, que correspondem, como visto na Seção 1.4, a dados representados por $(t_i; 1; x_i)$, sabe-se que a falha para o indivíduo i ocorreu no tempo t_i. Desta forma, a contribuição deste indivíduo com covariável x_i para a função de verossimilhança, é a "probabilidade"de que o mesmo tenha a recidiva ou morte no tempo t_i. Isto é dado pela sua função de densidade. No caso de censuras, que correspondem a dados representados por $(t_i; 0; x_i)$, sabe-se que o tempo de falha é superior a t_i. Então, a contribuição para a verossimilhança é a probabilidade do indivíduo com covariável

x_i sobreviver ao tempo t_i. Isto é dado por sua função de sobrevivência. Tratando os dados como independentes, a função de verossimilhança para o modelo linear na forma (4.4) pode então ser escrita para toda a amostra como

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big[f(y_i, \boldsymbol{\beta} \mid x_i) \Big]^{\delta_i} \Big[S(y_i, \boldsymbol{\beta} \mid x_i) \Big]^{(1-\delta_i)} \tag{4.6}$$

em que $y_i = \log(t_i)$, ou, ainda, para modelos na forma (4.3), por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big[f(t_i, \boldsymbol{\beta} \mid x_i) \Big]^{\delta_i} \Big[S(t_i, \boldsymbol{\beta} \mid x_i) \Big]^{(1-\delta_i)}. \tag{4.7}$$

Para obtenção dos estimadores de máxima verossimilhança, é necessário substituir as funções de densidade e sobrevivência por aquelas da distribuição do valor extremo em (4.6) ou da exponencial em (4.7). Fazendo isto em (4.6) e tomando-se o logaritmo de $L(\boldsymbol{\beta})$ tem-se

$$\begin{aligned} l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Big[\delta_i \Big((y_i - \beta_0 - \beta_1 x_i) - \exp\{y_i - \beta_0 - \beta_1 x_i\} \Big) \\ &\quad -(1-\delta_i) \exp\{y_i - \beta_0 - \beta_1 x_i\} \Big] \\ &= \sum_{i=1}^{n} \Big[\delta_i (y_i - \beta_0 - \beta_1 x_i) - \exp\{y_i - \beta_0 - \beta_1 x_i\} \Big]. \end{aligned} \tag{4.8}$$

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\beta}$ que maximizam a função $l(\boldsymbol{\beta})$ mostrada em (4.8). Para isso, é necessário encontrar as derivadas de (4.8) em função de β_0 e β_1, igualar as expressões obtidas a zero e resolver o sistema de equações resultante. Como as equações são não-lineares em β_0 e β_1 e não têm solução analítica, devem ser resolvidas numericamente o que, usualmente, envolve a utilização de um pacote estatístico.

4.2.2 Modelo de Regressão Weibull

O modelo de regressão exponencial apresentado é simples e interessante de ser manuseado para a introdução de modelagem com dados de sobrevivência. No entanto, devido à sua simplicidade, somente poucas situações na prática são adequadamente ajustadas por este modelo. De acordo com Nelson (1990) somente 10% de produtos industriais têm tempo de vida com distribuição exponencial. Uma forma de generalizar o modelo (4.4) é incluir um parâmetro extra de escala em sua

formulação. Isto é eqüivalente a passar de uma distribuição normal padrão para o erro, em modelos lineares, para uma normal com variância σ^2. O modelo linear (4.4) passa então, a ter a forma $Y = \log(T) = \beta_0 + \beta_1 x + \sigma\nu$ ou, considerando a presença de p covariáveis, tem-se:

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = X\boldsymbol{\beta} + \sigma\nu. \tag{4.9}$$

Este modelo é conhecido como modelo de regressão Weibull pois T deve ter uma distribuição de Weibull para que $\log(T)$ tenha uma distribuição do valor extremo com parâmetro de escala σ. Sendo assim, a função de sobrevivência para Y condicional a $\mathbf{x}$ é expressa por

$$S(y \mid \mathbf{x}) = \exp\left\{ -\exp\left\{ \frac{y - \mathbf{x}\boldsymbol{\beta}}{\sigma} \right\} \right\}$$

e, para T condicional a $\mathbf{x}$, por:

$$S(t \mid \mathbf{x}) = \exp\left\{ -\left(\frac{t}{\exp\{\mathbf{x}\boldsymbol{\beta}\}} \right)^{1/\sigma} \right\}.$$

4.2.3 Modelo de Tempo de Vida Acelerado

O modelo (4.9) pode ser estendido considerando outras distribuições para ν ou T. Distribuições adequadas para T são, por exemplo, a log-normal, gama e log-logística, entre outras. De forma correspondente, a distribuição para ν é normal, log-gama e logística. O modelo na forma (4.9) é bastante utilizado na prática e conhecido como **modelo de tempo de vida acelerado**. Isto porque a função das covariáveis é acelerar ou desacelerar o tempo de vida. Este fato pode ser melhor entendido se for considerado a escala original

$$T = \exp\{X\boldsymbol{\beta}\} \exp\{\sigma\nu\}. \tag{4.10}$$

A generalização deste modelo pode ser obtida em termos paramétricos se for acrescentado mais um parâmetro de forma. A gama generalizada é um exemplo de tal modelo. No entanto, a parte inferencial e seu correspondente aspecto computacional se tornam complexos. A generalização mais utilizada é no entanto, a proposta por Cox (1972) que sugere um modelo semi-paramétrico em que alguns modelos na forma (4.9) aparecem como casos particulares. Devido à importância deste modelo na análise de dados de sobrevivência, o Capítulo 5 será dedicado para a sua apresentação e discussão.

4.3 Adequação do Modelo Ajustado

Uma avaliação da adequação do modelo ajustado é parte fundamental da análise dos dados. No modelo de regressão linear usual, uma análise gráfica dos resíduos é usada para esta finalidade. Nos modelos de regressão apresentados neste capítulo, a definição de resíduos não é tão clara e, desse modo, diversos resíduos têm sido propostos na literatura para acessar o ajuste do modelo (Lawless, 1980, Klein e Moeschberger, 1997, Therneau e Grambsch, 2000).

Técnicas gráficas, que fazem uso dos diferentes resíduos propostos são, em particular, bastante utilizadas para examinar diferentes aspectos do modelo. Um desses aspectos é o de avaliar, por meio dos resíduos, a distribuição dos erros. Estas técnicas, no entanto, como bem observado por Klein e Moeschberger (1997), devem ser utilizadas como um meio de rejeitar modelos claramente inapropriados e não para "provar" que um particular modelo paramétrico está correto, mesmo porque, em muitas aplicações, dois ou mais modelos paramétricos podem fornecer ajustes razoáveis bem como estimativas similares das quantidades de interesse.

Nas seções que se seguem os seguintes resíduos serão descritos: i) os *resíduos de Cox-Snell* (1968) e os *resíduos padronizados*, úteis para examinar o ajuste global do modelo final, ii) os *resíduos martingale*, úteis para determinar a forma funcional (linear, quadrática etc.) de uma covariável, em geral contínua, sendo incluída no modelo de regressão e iii) os *resíduos deviance* que auxiliam a examinar a acurácia do modelo para cada indivíduo sob estudo.

4.3.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell (1968) auxiliam, como dito anteriormente, a examinar o ajuste global do modelo final. Esses resíduos são quantidades calculadas por

$$\hat{e}_i = \widehat{\Lambda}(t_i \mid \mathbf{x}_i) \tag{4.11}$$

em que $\widehat{\Lambda}(\cdot)$ é a função de risco acumulada obtida do modelo ajustado. Para os modelos de regressão exponencial, Weibull e log-normal, os resíduos de Cox-Snell são dados, respectivamente, por

$$\begin{aligned} \text{Exponencial: } \hat{e}_i &= \left[t_i \exp\{-\mathbf{x}_i\widehat{\boldsymbol{\beta}}\}\right] \\ \text{Weibull: } \hat{e}_i &= \left[t_i \exp\{-\mathbf{x}_i\widehat{\boldsymbol{\beta}}\}\right]^{1/\hat{\sigma}=\hat{\gamma}} \end{aligned}$$

e

$$\text{log-normal:}\ \ \hat{e}_i = -\log\left[1 - \Phi\left(\frac{\log(t_i) - \mathbf{x}_i\widehat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\right].$$

Os resíduos $\hat{e}_i$, que são estimativas dos erros que vêm de uma população homogênea, devem seguir uma distribuição exponencial padrão (Lawless, 1980). Desse modo, o gráfico $\hat{e}_i$ *versus* $\widehat{\Lambda}(\hat{e}_i)$, com $\widehat{\Lambda}(\hat{e}_i)$ o risco acumulado dos $\hat{e}_i$'s obtido pelo estimador de Nelson-Aalen, deve ser uma reta com inclinação 1 caso o modelo se ajuste bem aos dados.

Além do gráfico $\hat{e}_i$ *versus* $\widehat{\Lambda}(\hat{e}_i)$, pode-se também fazer uso das técnicas gráficas apresentadas na Seção 3.4. Assim, o gráfico $\hat{e}_i$ *versus* $-\log(\widehat{S}(\hat{e}_i))$ deve ser aproximadamente uma reta com inclinação 1 quando o modelo exponencial for adequado. Aqui, $\widehat{S}(\hat{e}_i)$ é a função de sobrevivência dos $\hat{e}_i$'s obtida pelo estimador de Kaplan-Meier ou de Nelson-Aalen. O gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier (ou Nelson-Aalen) e pelo modelo exponencial padrão, também auxiliam a verificar a qualidade do modelo ajustado. Quanto mais próximas elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

De acordo com Lawless (1980), quando existirem poucas observações censuradas e os modelos exponencial ou de Weibull estiverem sendo usados, é conveniente ajustar os resíduos censurados e tratá-los como se fossem não censurados. Assim, para todo t_i correspondente a um tempo censurado, tem-se, para estas situações, os correspondentes resíduos redefinidos por

$$\hat{e}_i = \left[t_i \exp(-\mathbf{x}_i\,\hat{\boldsymbol{\beta}})\right]^{1/\widehat{\sigma}} + 1 = \widehat{\Lambda}(t_i \mid \mathbf{x}_i) + 1.$$

Embora os resíduos de Cox-Snell sejam úteis para examinar o ajuste global do modelo, eles não indicam o tipo de falha, detectado a partir do modelo, quando o gráfico de $\hat{e}_i$ *versus* $\hat{\Lambda}(e_i)$ apresentar-se não linear (Crowley e Storer, 1983). Outros tipos de resíduos como, por exemplo, os resíduos martingale, podem ser úteis nessas situações.

Klein e Moeschberger (1997) observam, ainda, que os resíduos de Cox-Snell deveriam ser usados com cuidado pois a distribuição exponencial dos mesmos mantém-se somente quando os verdadeiros valores dos parâmetros são usados em (4.11). Quando as estimativas dessas quantidades são usadas para o cálculo dos resíduos, como é feito aqui, falhas quanto à distribuição exponencial podem ocorrer devido, parcialmente, à incerteza no processo de estimação dos parâmetros $\boldsymbol{\beta}$. Essa incerteza é maior na cauda direita da distribuição e para amostras pequenas.

4.3.2 Resíduos Padronizados

Examinar o ajuste do modelo por meio dos resíduos de Cox-Snell é eqüivalente a fazer uso dos assim denominados *resíduos padronizados* baseados na representação dos modelos log-lineares apresentados em (4.4) e (4.9). Neste caso, e por analogia aos resíduos usados no modelo de regressão normal usual, os resíduos padronizados são quantidades calculadas por

$$\widehat{\nu_i} = \frac{(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}} \tag{4.12}$$

com $y_i = \log(t_i)$.

Assim se, por exemplo, o modelo de regressão exponencial for adequado, esses resíduos deveriam ser uma amostra censurada da distribuição valor extremo padrão. De modo análogo, se o modelo log-normal for apropriado, os mesmos deveriam ser uma amostra censurada da distribuição normal padrão.

Note que os resíduos $\hat{\nu}_i$ são estimativas dos erros que vêm de uma população homogênea. Desta forma, o gráfico das sobrevivências destes resíduos obtida pelo estimador de Kaplan-Meier (ou de Nelson-Aalen) *versus* as sobrevivências obtidas pelo modelo valor extremo padrão deve ser aproximadamente uma reta para que o modelo de regressão exponencial seja considerado adequado. O mesmo vale para o modelo de regressão Weibull.

Similarmente, o modelo de regressão log-normal será considerado adequado se o gráfico de probabilidades normal dos resíduos $\hat{\nu}_i$ for aproximadamente uma reta. Eqüivalentemente, o gráfico das sobrevivências dos resíduos $\hat{e}_i^* = \exp\{\hat{\nu}_i\}$, obtida pelo estimador de Kaplan-Meier (ou Nelson-Aalen), *versus* as sobrevivências destes resíduos obtidas pelo modelo log-normal padrão, deve ser aproximadamente uma reta para que o modelo de regressão log-normal apresente ajuste satisfatório. O gráfico das curvas de sobrevivência dos e_i^*'s, obtidas por Kaplan-Meier (ou Nelson-Aalen) e pelo modelo log-normal padrão, também auxiliam a verificar a qualidade do modelo ajustado. Quanto mais próximas elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

4.3.3 Resíduos Martingale

Para os modelos de regressão paramétricos apresentados neste capítulo, os resíduos martingale são definidos por

$$\hat{m}_i = \delta_i - \hat{e}_i. \tag{4.13}$$

em que δ_i é a variável indicadora de censura e $\hat{e}_i$ os resíduos de Cox-Snell. Esses resíduos, que na realidade são uma ligeira modificação dos resíduos de Cox-Snell, são vistos como uma estimativa do número de falhas em excesso observada nos dados mas não predito pelo modelo. Os mesmos são usados, em geral, para examinar a melhor forma funcional (linear, quadrática etc.) para uma dada covariável em um modelo de regressão assumido para os dados sob estudo.

Para, por exemplo, uma covariável contínua X_1, os pares $(x_{1i}, \hat{m}_i)$, $i = 1, \cdots n$, são representados graficamente. Uma curva suavizada do diagrama de dispersão resultante é tipicamente usado. Se esta curva suavizada for linear, nenhuma transformação em X_1 é necessária. Se esta curva, contudo, apresentar um mudança em um determinado valor de X_1, uma versão discretizada da covariável é indicada. Outros comportamentos desta curva podem indicar, por exemplo, a inclusão de um termo quadrático da covariável no modelo ou sugerir alguma transformação da mesma.

4.3.4 Resíduos Deviance

Os resíduos deviance nos modelos de regressão paramétricos são definidos por

$$\hat{d}_i = \text{sinal}(\hat{m}_i)\Big[-2\Big(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i)\Big)\Big]^{1/2}. \tag{4.14}$$

Esses resíduos, que são uma tentativa de tornar os resíduos martingale mais simétricos em torno de zero, facilitam, em geral, a detecção de pontos atípicos (*outliers*). Se o modelo for apropriado, estes resíduos deveriam apresentar um comportamento aleatório em torno de zero. Gráficos dos resíduos martingale, ou deviance, *versus* o tempo fornecem assim, uma forma de verificar a adequação do modelo ajustado bem como auxiliam na detecção de observações atípicas.

4.4 Aplicações

4.4.1 Sobrevida de Pacientes com Leucemia Aguda

Considere os tempos de sobrevida, em semanas, de 17 pacientes com leucemia aguda (Lawless, 1982) apresentados na Tabela 4.1. Para esses pacientes a contagem de glóbulos brancos (WBC) foi registrada na data do diagnóstico e encontram-se, com seus correspondentes logaritmos na base 10, apresentados também na Tabela 4.1.

Observe, neste estudo, que a covariável WBC é contínua e, a menos que a mesma seja estratificada em no máximo dois estratos uma vez que o tamanho amostral

Tabela 4.1: Tempos de sobrevida de pacientes com leucemia aguda.

Tempos	WBC	$\log_{10}$(WBC)	Tempos	WBC	$\log_{10}$(WBC)
65	2300	3,36	143	7000	3,85
156	750	2,88	56	9400	3,97
100	4300	3,63	26	32000	4,51
134	2600	3,41	22	35000	4,54
16	6000	3,78	1	100000	5,00
108	10000	4,02	1	100000	5,00
121	10000	4,00	5	52000	4,72
4	17000	4,23	65	100000	5,00
39	5400	3,73			

é relativamente pequeno, fica inviável a obtenção da curva de sobrevivência pelo método de Kaplan-Meier. Analisar os dados por meio de um modelo de regressão que considere a covariável WBC, ou o $\log_{10}$(WBC), parece ser, portanto, uma alternativa viável. Para isso, e como uma ferramenta auxiliar no processo de escolha do modelo de regressão adequado, foi inicialmente ignorado a covariável WBC e construído os gráficos das linearizações, discutidos na Seção 3.4, dos modelos exponencial, de Weibull e log-normal. Estes gráficos encontram-se apresentados na Figura 4.1 e foram obtidos no *R* por meio dos comandos

```
> temp<-c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65)
> cens<-rep(1,17)
> lwbc<-c(3.36,2.88,3.63,3.41,3.78,4.02,4.00,4.23,3.73,3.85,3.97,
          4.51,4.54,5.00,5.00,4.72,5.00)
> dados<-cbind(temp,cens,lwbc)
> require(survival)
> dados<-as.data.frame(dados)
> i<-order(dados$temp)
> dados<-dados[i,]
> ekm<- survfit(Surv(dados$temp,dados$cens))
> summary(ekm)
> st<-ekm$surv
> temp<-ekm$time
> invst<-qnorm(st)
> par(mfrow=c(1,3))
> plot(temp, -log(st),pch=16,xlab="Tempos",ylab="-log(S(t))")
> plot(log(temp),log(-log(st)),pch=16,xlab="log(tempos)",ylab="log(-log(S(t))")
> plot(log(temp),invst,pch=16,xlab="log(tempos)",ylab=expression(Phi^-1 * (S(t))))
```

As distribuições exponencial e de Weibull, pelo que foi discutido na Seção 3.4 e o que pode ser observado na Figura 4.1, apresentaram-se aparentemente como as melhores candidatas, dentre as consideradas, para a análise dos dados deste estudo. Considerando-se, então, os modelos de regressão exponencial e Weibull com

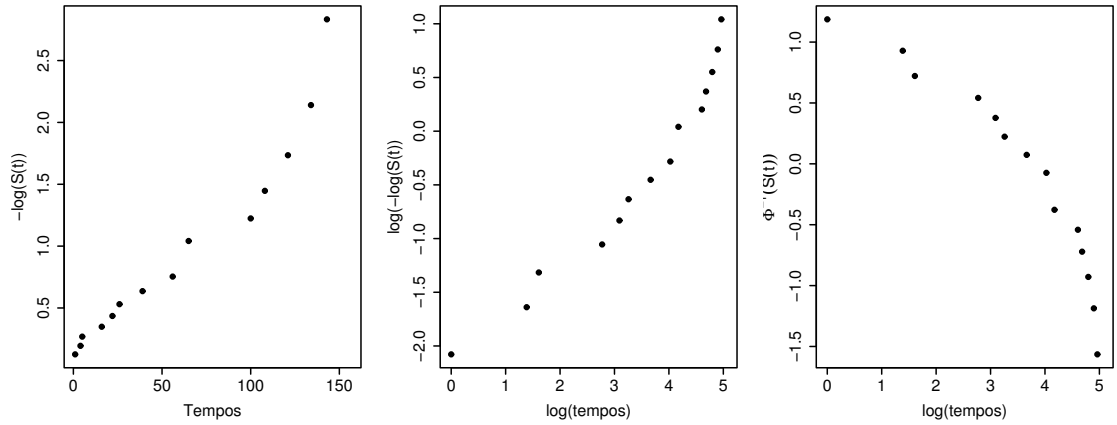

Figura 4.1: Gráficos $t \times -\log(\hat{S}(t))$, $\log t \times \log(-\log(\hat{S}(t))$ e $\log t \times \Phi^{-1}(\hat{S}(t))$

a covariável $X_1 = \log(\text{WBC})$, foram obtidas, no R, as estimativas dos parâmetros apresentadas na Tabela 4.2.

```
> ajust1<-survreg(Surv(dados$temp, dados$cens)~dados$lwbc, dist='exponential')
> ajust1
> ajust1$loglik
> ajust2<-survreg(Surv(dados$temp, dados$cens)~dados$lwbc, dist='weibull')
> ajust2
> ajust2$loglik
> gama<-1/ajust2$scale
> gama
```

Tabela 4.2: Estimativas para os dados de leucemia aguda.

Regressão exponencial	Regressão Weibull
$\hat{\beta}_0 =$ 8,4775	$\hat{\beta}_0 =$ 8,4408
$\hat{\beta}_1 =$ -1,1093	$\hat{\beta}_1 =$ -1,0982
$\gamma = 1$ (fixo)	$\hat{\gamma} =$ 1,0218

Observa-se a partir da Tabela 4.2 que a estimativa do parâmetro $\gamma = 1/\sigma$ encontra-se muito próxima de 1. Testando-se, assim, as hipóteses H_0: $\gamma = 1$ *versus* H_A: $\gamma \neq 1$ obteve-se $TRV = 2(83,87705 - 83,87136) = 0,0113$ ($p = 0,915$, g.l. $= 1$) e, portanto, rejeita-se o modelo de regressão Weibull em favor do exponencial.

Utilizando o modelo de regressão exponencial, testou-se então a hipótese nula H_0: $\beta_1 = 0$ por meio do teste da razão de verossimilhança que resultou em $TRV = 2(87,28983$ - $83,87705) = 6,83$ ($p = 0,009$, g.l. $= 1$). Concluiu-se, portanto, pela rejeição da

hipótese H_0 e sendo, assim, é possível dizer que a variação observada nos tempos de sobrevida dos pacientes relaciona-se à contagem de glóbulos brancos.

A função de sobrevivência estimada obtida pelo modelo de regressão exponencial ajustado para os dados desse exemplo é, portanto, expressa por:

$$\widehat{S}(t \mid x_1) = \exp\left\{ -\left(\frac{t}{\exp\{8,4775 - 1,1093\, x_1\}} \right) \right\} \qquad t \geq 0$$

em que x_1 = logaritmo, na base 10, da contagem de glóbulos brancos.

Note, ainda, que $\hat{\beta}_1$ é negativo o que implica que quanto maior o valor de x_1 menor a probabilidade de sobrevivência estimada. Este fato pode ser claramente observado na Figura 4.2 em que as curvas de sobrevivência estimadas para dois pacientes, um com $x_1 = 4,0$ e outro com $x_1 = 3,0$, são apresentadas.

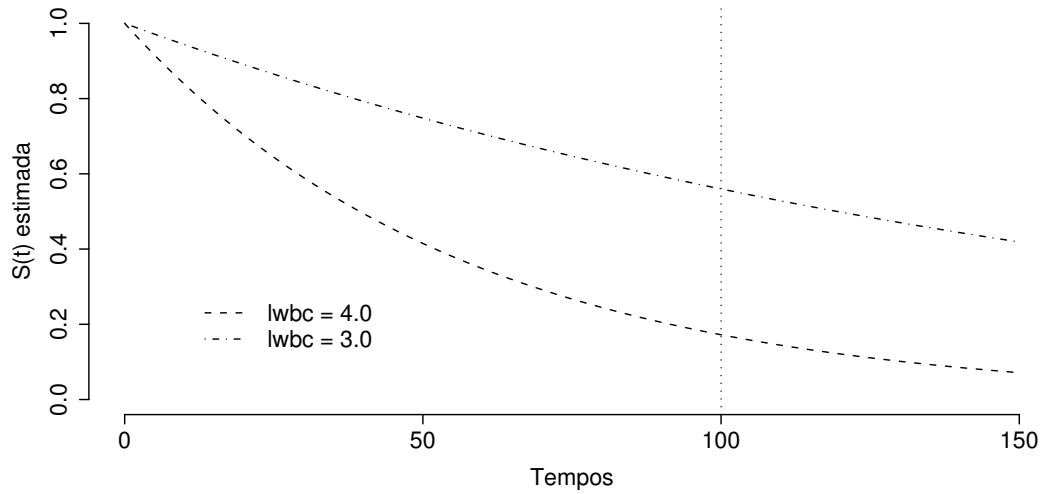

Figura 4.2: Curvas de sobrevivência estimadas pelo modelo de regressão exponencial para os dados de leucemia aguda.

A partir da Figura 4.2 pode-se observar que $\widehat{S}(100 \mid x_1 = 4) = 0,172$, o que significa que em torno de 17% dos pacientes que apresentarem, no diagnóstico, logaritmo da contagem de glóbulos brancos igual a 4,0, estarão vivos no tempo $t = 100$ semanas (linha vertical apresentada no gráfico). Por outro lado, estima-se, para pacientes que no diagnóstico apresentarem logaritmo da contagem de glóbulos brancos igual a 3,0, que em torno de 56% deles estarão vivos na 100ª semana, visto que $\widehat{S}(100 \mid x_1 = 3) = 0,559$.

Para avaliar o ajuste do modelo de regressão exponencial aos dados deste estudo, pode-se examinar os resíduos de Cox-Snell, definidos aqui por:

$$\hat{e}_i = \left[t_i \exp(-\hat{\beta}_0 - \hat{\beta}_1 x_1) \right]$$

e que, neste modelo, são considerados, como visto anteriormente, uma amostra aleatória proveniente da distribuição exponencial padrão (Exp(1)). Assim, as estimativas das sobrevivências dos resíduos obtidas por Kaplan-Meier ($\hat{S}(\hat{e}_i)_{KM}$) e pelo modelo exponencial padrão ($\hat{S}(\hat{e}_i)_{Exp}$), deverão estar próximas bem como, o gráfico dos pares de pontos ($\hat{S}(\hat{e}_i)_{KM}$, $\hat{S}(\hat{e}_i)_{Exp}$), deverão ser aproximadamente uma reta para que o modelo ajustado possa ser considerado satisfatório.

A Figura 4.3 apresenta ambos os gráficos citados e pode-se observar, destes gráficos, que o modelo exponencial padrão é aceitável para os resíduos.

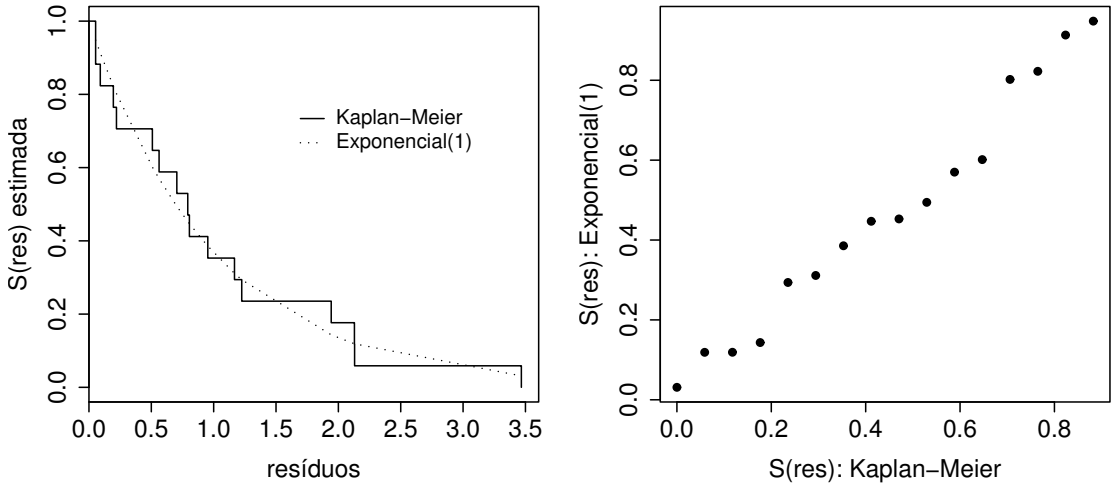

Figura 4.3: Análise dos resíduos de Cox-Snell do modelo de regressão exponencial ajustado para os dados de leucemia aguda.

Dos resultados apresentados, verificou-se, portanto, que o modelo de regressão exponencial apresentou ajuste satisfatório aos dados dos tempos de sobrevida dos pacientes de leucemia aguda analisados nesse exemplo. De maneira geral, pode-se ainda concluir, do modelo ajustado e seus resultados, que o tempo de sobrevida estimado dos pacientes diminui à medida que são observadas, no diagnóstico, contagens crescentes de glóbulos brancos sendo, as probabilidades de sobrevivência, em qualquer tempo $t \geq 0$ e para um dado valor de x_1, estimadas por:

$$\widehat{S}(t \mid x_1) = \exp\left\{ -\left(\frac{t}{\exp\{8,4775 - 1,1093\ x_1\}} \right) \right\} \qquad t \geq 0.$$

Os comandos utilizados no R para obtenção das Figuras 4.2 e 4.3 encontram-se apresentados no Apêndice B deste texto.

4.4.2 Grupos de Pacientes com Leucemia Aguda

Considere, neste estudo, os mesmos tempos de sobrevida, em semanas, dos 17 pacientes com leucemia aguda apresentados na Seção 4.4.1, pacientes estes que apresentaram Ag positivo, bem como outro grupo de 16 pacientes, também com leucemia aguda, mas com Ag negativo (Louzada-Neto et al., 2002). Para todos os pacientes, a covariável contagem de glóbulos brancos (WBC) foi registrada na data do diagnóstico. A WBC e seus respectivos logaritmos na base 10, para cada um dos dois grupos (Ag+ e Ag-), encontram-se apresentados na Tabela 4.3.

Tabela 4.3: Sobrevida de dois grupos de pacientes com leucemia aguda.

	Ag+			Ag-	
Tempos	WBC	$\log_{10}$(WBC)	Tempos	WBC	$\log_{10}$(WBC)
65	2300	3,36	56	4400	3,64
156	750	2,88	65	3000	3,48
100	4300	3,63	17	4000	3,60
134	2600	3,41	7	1500	3,18
16	6000	3,78	16	9000	3,95
108	10000	4,02	22	5300	3,72
121	10000	4,00	3	10000	4,00
4	17000	4,23	4	19000	4,28
39	5400	3,73	2	27000	4,43
143	7000	3,85	3	28000	4,45
56	9400	3,97	8	31000	4,49
26	32000	4,51	4	26000	4,41
22	35000	4,54	3	21000	4,32
1	100000	5,00	30	79000	4,90
1	100000	5,00	4	100000	5,00
5	52000	4,72	43	100000	5,00
65	100000	5,00			

Observe a existência de duas covariáveis de interesse neste estudo: X_1 = logaritmo da contagem de glóbulos brancos e X_2 = grupos (Ag+ ou Ag-). Para esta última será considerado:

$$X_2 = \begin{cases} 0 & \text{se grupo Ag+} \\ 1 & \text{se grupo Ag-} \end{cases}$$

Para o grupo Ag+, analisado na Seção 4.4.1, foi escolhido o modelo de regressão exponencial. Procedendo então uma investigação inicial de modo análogo ao que foi feito para o grupo Ag+, foram obtidos no R (ver comandos no Apêndice B) os gráficos das linearizações dos modelos exponencial, de Weibull e log-normal, apresentados na Figura 4.4, para ambos os grupos (Ag+ e Ag-). A partir destes gráficos, é possível notar que as linearizações para o grupo Ag- são muito similares às obti-

das para o grupo Ag+. Há, portanto, indicações favoráveis ao modelo de regressão exponencial também para o grupo Ag-.

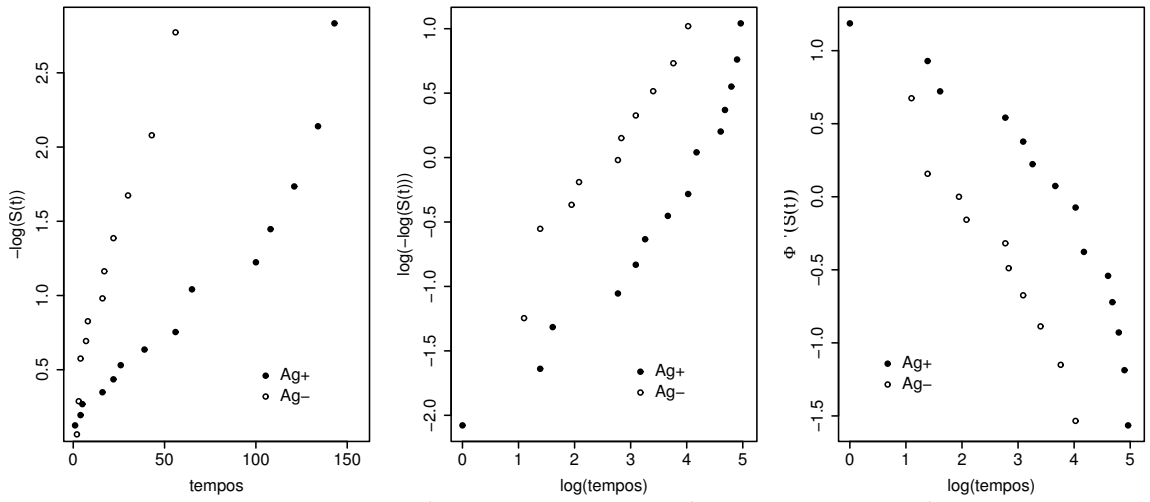

Figura 4.4: Gráficos de $t \times -\log(\hat{S}(t))$, $\log t \times \log(-\log(\hat{S}(t))$ e $\log t \times -\Phi^{-1}(\hat{S}(t))$ para os grupos de pacientes Ag+ e Ag- com leucemia aguda.

Considerando então o modelo de regressão exponencial e as covariáveis $X_1 = \log_{10}(\text{WBC})$ e $X_2 =$ grupos, foram obtidos os resultados das estimativas dos parâmetros e os valores dos logaritmos das funções de verossimilhança, apresentados na Tabela 4.4, para 4 modelos possíveis, um deles com a interação entre X_1 e X_2.

Tabela 4.4: Estimativas dos parâmetros e logaritmo das funções de verossimilhança dos modelos de regressão exponencial ajustados para os dados de leucemia aguda.

Modelo	Covariáveis no modelo	Estimativas	Log verossimilhança
1	nenhuma	$\hat{\beta}_0 = 3{,}71$	$l_1 =$ -155,5
2	X_1	$\hat{\beta}_0 = 7{,}37$	
		$\hat{\beta}_1 =$ -0,92	$l_2 =$ -150,3
3	X_1 e X_2	$\hat{\beta}_0 = 6{,}83$	
		$\hat{\beta}_1 =$ -0,70	
		$\hat{\beta}_2 =$ -1,02	$l_3 =$ -146,5
4	X_1, X_2 e $X_1 * X_2$	$\hat{\beta}_0 = 8{,}47$	
		$\hat{\beta}_1 =$ -1,11	
		$\hat{\beta}_2 =$ -4,14	
		$\hat{\beta}_3 = 0{,}755$	$l_4 =$ -145,7

Para testar a significância da interação $X_1 * X_2$, foi usado o teste da razão de verossimilhança que resultou em $TRV = 2[146{,}5 - 145{,}7] = 1{,}6$ (valor $p = 0{,}2059$, g.l. $= 1$). Deste resultado pode-se concluir não haver evidências estatísticas de que

a interação entre X_1 e X_2 seja significativa. Desse modo, foram testados os efeitos das covariáveis X_1 e X_2 cujos resultados, apresentados na Tabela 4.5, mostram evidências estatísticas de efeito da covariável X_1 bem como evidências de efeito da covariável X_2 na presença de X_1, com valores p, obtidos da distribuição $\chi^2_{(1)}$, de 0,0013 e 0,0058, respectivamente.

Tabela 4.5: Resultados obtidos para os testes da razão de verossimilhanças.

Efeito	Hipótese nula	Estatística de teste (TRV)	valor p
interação: X_1*X_2	H_0: $\beta_3 = 0$	2(146,5 - 145,7)= 1,6	0,2059
de X_2	H_0: $\beta_2 = 0$	2(150,3 - 146,5)= 7,6	0,0058
de X_1	H_0: $\beta_1 = 0$	2(155,5 - 150,3)= 10,4	0,0013

Considerando-se, portanto, o modelo de regressão exponencial com as covariáveis X_1 e X_2, em que as estimativas de seus respectivos parâmetros encontram-se mostradas na Tabela 4.4, tem-se que a sobrevivência estimada, por este modelo, para um paciente com leucemia aguda é obtida por:

$$\widehat{S}(t \mid x_1, x_2) = \exp\left\{ -\left(\frac{t}{\exp\{6,83 - 0,70\ x_1 - 1,02\ x_2\}} \right) \right\} \qquad t \geq 0$$

em que x_1 é o logaritmo, na base 10, da contagem de glóbulos brancos observada para este paciente e x_2 indica se o paciente pertence ao grupo Ag+ ou Ag-. Para pacientes do grupo Ag+ tem-se $x_2 = 0$. Em caso contrário, $x_2 = 1$.

A análise dos resíduos de Cox-Snell desse modelo, análogo ao que foi feito no estudo anterior, é apresentada na Figura 4.5. Desta figura, observa-se que o modelo de regressão exponencial apresenta ajuste razoável aos dados dos tempos de sobrevida desses dois grupos de pacientes com leucemia aguda.

Note, para o modelo ajustado, que $\hat{\beta}_1$ é negativo o que implica que quanto maior o valor de x_1 menor a probabilidade de sobrevivência estimada. Observe, ainda, que $\hat{\beta}_2$ também é negativo o que implica que pacientes no grupo Ag- ($x_2 = 1$) apresentarão probabilidade de sobrevivência estimada menor do que a dos pacientes do grupo Ag+ ($x_2 = 0$) . Este fato pode ser claramente observado na Figura 4.6 em que as curvas de sobrevivência estimadas para dois pacientes pertencentes ao grupo Ag+ e dois outros pacientes pertencentes ao grupo Ag-, um com $x_1 = 4,0$ e outro com $x_1 = 3,0$, são apresentadas.

Os correspondentes riscos estimados dos pacientes considerados na Figura 4.6 encontram-se apresentados na Figura 4.7. Estes são constantes ao longo do tempo,

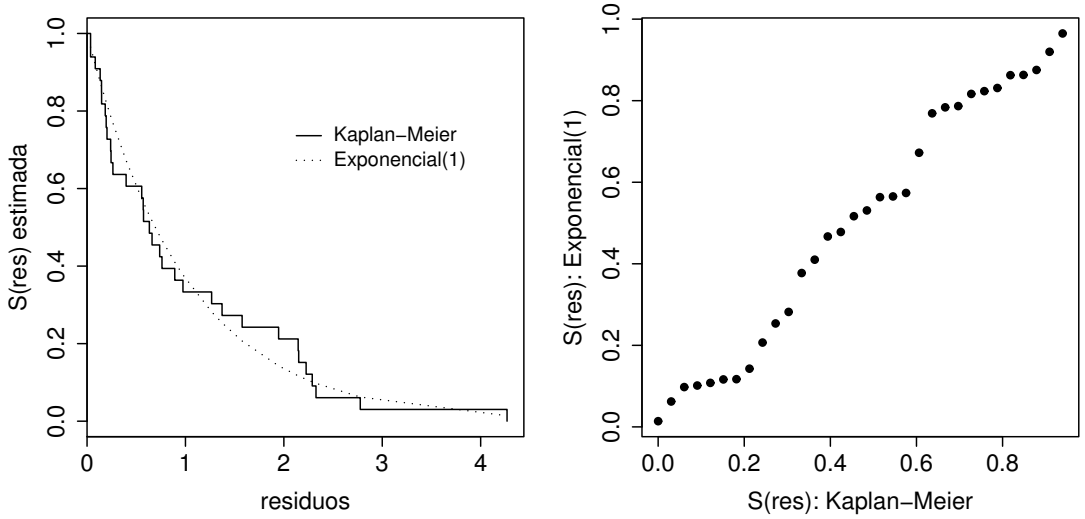

Figura 4.5: Análise dos resíduos para os dados de leucemia com as covariáveis X_1 e X_2.

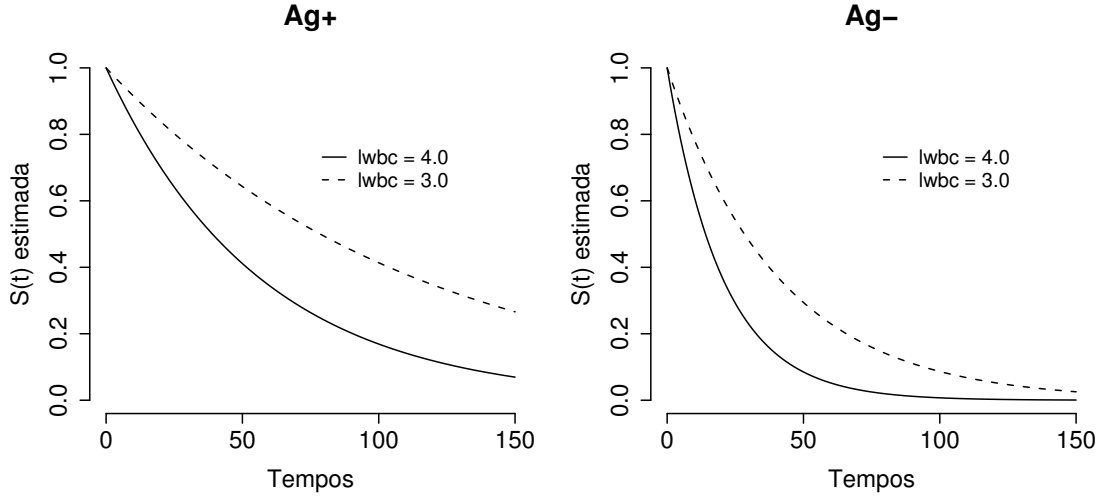

Figura 4.6: Curvas de sobrevivência estimadas pelo modelo de regressão exponencial para dois pacientes do grupo Ag+ e dois pacientes do grupo Ag- com leucemia aguda e diferentes contagens de glóbulos brancos no diagnóstico.

o que é uma característica do modelo exponencial, podendo-se notar que a taxa instântanea de falha do paciente com $x_1 = 4,0$, em relação ao paciente com $x_1 = 3,0$, é maior tanto no grupo Ag+ quanto no grupo Ag-. Portanto, quanto maior a contagem de glóbulos brancos no diagnóstico, maior o risco de falha. Comparativamente, os pacientes no grupo Ag- apresentam risco de falha estimado maior do que o referido risco estimado para os pacientes no grupo Ag+.

No Apêndice C o leitor encontra os comandos usados no R para obtenção dos resultados e figuras apresentadas para este estudo.

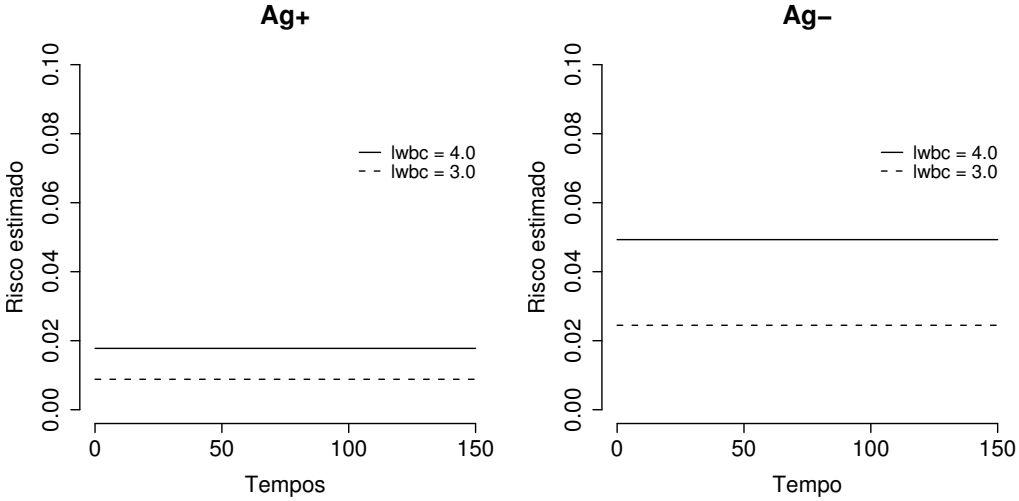

Figura 4.7: Riscos estimados pelo modelo de regressão exponencial para para dois pacientes do grupo Ag+ e dois pacientes do grupo Ag- com leucemia aguda e diferentes contagens de glóbulos brancos no diagnóstico.

4.4.3 Análise dos Dados de Aleitamento Materno

4.4.4.1 Descrição do estudo e das variáveis

As Organizações Internacionais de Saúde recomendam o leite materno como a única fonte de alimentação para crianças entre 4 e 6 meses de idade. Identificar fatores determinantes do aleitamento materno em diferentes populações é, portanto, fundamental para alcançar tal recomendação. Um artigo publicado na revista *American Institute of Nutrition* intitulado "Exclusive Breast-Feeding Duration is Associated with Attitudinal, Socioeconomic and Biocultural Determinants in Three Latin American Countries"(Pérez-Escamilla et al., 1985) apresenta um estudo realizado em Honduras, México e Brasil nos anos de 1992 e 1993 cujo principal objetivo era identificar determinantes do aleitamento exclusivamente materno em populações urbanas de baixa renda. Os resultados desse estudo mostram que a condição sócio-econômica (Honduras e México) e o peso ao nascimento da criança (Brasil e Honduras) estão associados com o aleitamento exclusivamente materno. Além disso, as mulheres que têm acesso à maternidades que promovem programas de aleitamento são mais bem sucedidas. Nessa mesma linha de pesquisa, um outro estudo foi realizado pelos Profs. Eugênio Goulart e Cláudia Lindgren do Departamento de Pediatria da UFMG. Este estudo foi realizado no Centro de Saúde São Marcos, localizado em Belo Horizonte, que é um ambulatório municipal que atende essencialmente a população de baixa renda. Esse estudo tem como objetivos prin-

cipais conhecer a prática do aleitamento materno de mães que utilizam este centro, assim como os possíveis fatores de risco ou de proteção para o desmame precoce. Um inquérito epidemiológico composto por questões demográficas e comportamentais foi aplicado a 150 mães de crianças menores de 2 anos de idade. A variável resposta de interesse foi o tempo máximo de aleitamento materno, ou seja, o tempo contado a partir do nascimento até o desmame completo da criança.

No estudo foram registradas 11 covariáveis e a variável resposta. Algumas crianças não foram acompanhadas até o desmame e, portanto, registra-se a presença de censuras. O banco de dados, que se encontra no Apêndice, é composto por 13 variáveis: as 11 covariáveis (fatores) e a variável resposta que é representada pelo tempo de acompanhamento e uma variável indicadora de ocorrência do desmame. A Tabela 4.6 apresenta uma descrição das 11 covariáveis estudadas.

Tabela 4.6: Descrição das covariáveis utilizadas no estudo sobre aleitamento materno.

Código	Descrição	Categorias
V1	Experiência anterior de amamentação	0 se sim e 1 se não
V2	Número de filhos vivos	0 se dois ou menos e 1 se mais de dois
V3	Conceito materno sobre o tempo ideal de amamentação	0 se $>$ 6 meses e 1 se $\leq$ 6 meses
V4	Dificuldades para amamentar nos primeiros dias pós-parto	0 se não e 1 se sim
V5	Tipo de serviço em que realizou o pré-natal	0 se serviço público e 1 se privado/convênios
V6	Recebeu exclusivamente leite materno na maternidade	0 se sim e 1 se não
V7	A criança tem contato com o pai	0 se sim e 1 se não
V8	Renda per capita (em SM/mês)	0 se $\geq$ 1 SM e 0 se $<$ 1 SM
V9	Peso ao nascimento	0 se $\geq$ 2,5kg e 1 se $<$ 2,5kg
V10	Tempo de separação mãe-filho pós-parto	0 se $\leq$ 6 horas e 1 se $>$ 6 horas
V11	Permanência no berçário	0 se não e 1 se sim

Na análise estatística destes dados será utilizado, nesta seção, as técnicas de análise de sobrevivência apresentadas neste capítulo e nos anteriores. Nos Capítulos 5 e 6, esses dados serão também modelados por meio do modelo de regressão de Cox.

4.4.4.2 Análise Descritiva e Exploratória

A primeira etapa de qualquer análise estatística de dados consiste de uma análise

descritiva das variáveis em estudo. Em análise de sobrevivência esta etapa consiste em utilizar os métodos não-paramétricos apresentados no Capítulo 2.

Todas as covariáveis são dicotômicas e portanto é possível construir as estimativas de Kaplan-Meier para comparar as duas categorias. Isto foi feito para as 11 covariáveis assim como foi testado a hipótese de igualdade das duas curvas utilizando os testes de Wilcoxon e logrank. Os 11 gráficos não serão apresentados aqui mas, a título de ilustração, a Figura 4.8, obtida no R por meio dos comandos

```
> desmame<-read.table("c:/desmame.txt",h=T)
> attach(desmame)
> require(survival)
> ekm<- survfit(Surv(tempo,cens)~V4)
> summary(ekm)
> survdiff(Surv(tempo,cens)~V4,rho=0)
> plot(ekm,lty=c(1,4),xlab="Tempo até o desmame (meses)",ylab="Sobrevivência")
> legend(15,0.9,lty=c(1,4),c("mais de seis meses","seis meses ou menos"),bty="n",cex=0.8)
> text(18.5,0.93,c("Tempo ideal de amamentação"),bty="n", cex=0.85)
```

apresenta as curvas de Kaplan-Meier para a covariável dificuldades para amamentar nos primeiros dias pós-parto (V4).

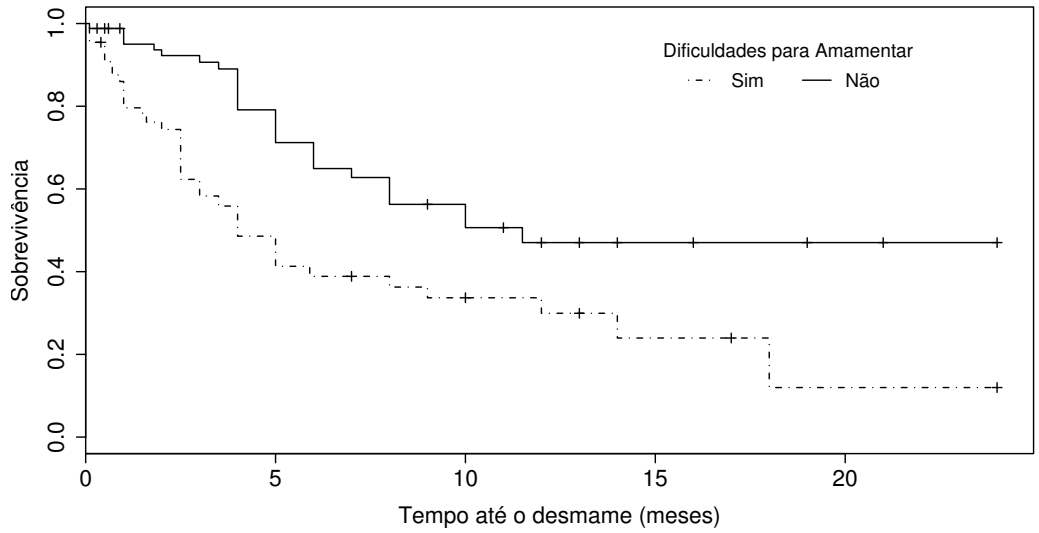

Figura 4.8: Curvas de sobrevivência estimadas pelo método de Kaplan-Meier para a covariável dificuldades para amamentar nos primeiros dias pós-parto (V4).

A Figura 4.8 indica que as mães que não tiveram dificuldades para amamentar nos primeiros dias pós-parto apresentam um tempo até o desmame maior que aquelas que tiveram dificuldades. A Tabela 4.7 mostra os valores dos testes logrank e de Wilcoxon.

Tabela 4.7: Testes logrank e de Wilcoxon utilizados para testar a igualdade das curvas de sobrevivência obtidas para as covariáveis consideradas no estudo de aleitamento.

	Testes (valor p)	
Covariável	logrank	Wilcoxon
V1: Experiência anterior de amamentação	$3,95$ $(0,047)$	$6,73$ $(0,010)$
V2: Número de filhos vivos	$2,60$ $(0,107)$	$2,02$ $(0,155)$
V3: Tempo ideal de amamentação	$6,15$ $(0,013)$	$8,54$ $(0,004)$
V4: Apresentou dificuldades para amamentar	$12,26$ $(0,001)$	$15,45$ $(< 0,001)$
V5: Tipo de serviço do pré-natal	$1,38$ $(0,241)$	$1,09$ $(0,296)$
V6: Recebeu exclusivamente leite materno	$7,47$ $(0,006)$	$6,31$ $(0,012)$
V7: Contato com o pai	$1,84$ $(0,175)$	$0,90$ $(0,344)$
V8: Renda per capita	$2,11$ $(0,146)$	$2,60$ $(0,107)$
V9: Peso ao nascimento	$1,87$ $(0,171)$	$2,59$ $(0,108)$
V10: Tempo de separação mãe-filho	$2,60$ $(0,107)$	$0,97$ $(0,325)$
V11: Permanência no berçário	$2,93$ $(0,087)$	$0,90$ $(0,343)$

A próxima etapa da análise consiste em modelar separadamente cada uma das covariáveis com a resposta. Esta etapa tem por objetivo selecionar quais variáveis explicativas (covariáveis) prosseguirão na análise. O critério utilizado nesse trabalho foi o de permanecer com aquelas que apresentaram valores p inferiores a 0,25 em pelo menos um dos testes de comparação das curvas de sobrevivência. Esta proposta em escolher um nível relativamente modesto de significância é baseado em recomendações para regressão linear dada por Bendel e Afifi (1977), para análise discriminante dada por Costanza e Afifi (1979) e para mudanças nos coeficientes do modelo dada por Mickey e Greenland (1989).

Com base nos resultados apresentados na Tabela 4.7, verifica-se que todas as covariáveis passaram por esse critério, e portanto deverão ser incluídas na etapa de modelagem estatística.

As técnicas utilizadas até o momento são importantes para descrever os dados de sobrevivência pela sua simplicidade e facilidade de aplicação, pois não envolvem nenhuma estrutura paramétrica. No entanto, elas não permitem a inclusão conjunta das covariáveis na análise. A forma mais eficiente de acomodar o efeito das covariáveis é utilizar um modelo de regressão apropriado para dados censurados. Entretanto, antes de realizar o ajuste destes modelos será discutido um passo importante na análise estatística que é o de seleção de variáveis.

4.4.4.3 Estratégia para a Seleção de Covariáveis

Onze covariáveis potencialmente importantes para descrever o comportamento da resposta foram selecionadas para serem incluídas no modelo. Existem, portanto, 2048 possíveis modelos formados pela combinação de todas estas covariáveis. É certamente impraticável ajustar todos estes possíveis modelos a fim de ser selecionado o que melhor explique a resposta. Nessas situações, rotinas automáticas para seleção de covariáveis podem ser utilizadas, tais como os métodos *forward*, *backward* ou *stepwise*. Estes métodos estão implementados e, portanto, disponíveis em pacotes estatísticos. Entretanto, tais rotinas possuem algumas desvantagens. Tipicamente, elas tendem a identificar um particular conjunto de covariáveis, ao invés de possíveis conjuntos igualmente bons para explicar a resposta. Esse fato impossibilita que dois ou mais conjuntos de covariáveis igualmente bons sejam apresentados para o pesquisador, para a escolha do mais relevante em sua área de aplicação. Isto significa que estes métodos são automáticos e fazem com que o pacote estatístico escolha o modelo. Na realidade, o que se defende aqui é que o estatístico juntamente com o pesquisador tenham uma postura pro-ativa neste processo. Isto implica, por exemplo, que covariáveis importantes em termos clínicos devem ser incluídas independente de significância estatística, assim como a importância clínica deve ser considerada em cada passo de inclusão ou exclusão no processo de seleção de covariáveis.

Frente à estas limitações das rotinas automáticas optou-se por utilizar métodos que envolvem a interferência mais de perto do analista. O leitor interessado em mais informações sobre os métodos *stepwise* podem consultar Draper e Smith (1998). Na verdade, a filosofia do método é essencialmente a mesma para qualquer classe de modelos. Neste estudo optou-se por utilizar uma estratégia de seleção de modelos derivada da proposta de Collett (1994). Os passos utilizados no processo de seleção são apresentados a seguir.

1. Ajustar todos os modelos contendo uma única covariável. Incluir todas as covariáveis que forem significativas ao nível de $0,10$. É aconselhável utilizar o teste da razão de verossimilhanças neste passo.

2. As covariáveis significativas no passo 1 são então ajustadas conjuntamente. Na presença de certas covariáveis, outras podem deixar de ser significativas. Conseqüentemente, ajusta-se modelos reduzidos, excluindo uma única covariável. Verifica-se as covariáveis que provocam um aumento estatisticamente significativo na estatística da razão de verossimilhanças. Somente aquelas que atingiram a significância per-

manecem no modelo.

3. Ajusta-se um novo modelo com as covariáveis retidas no passo 2. Neste passo as covariáveis excluídas no passo 2 retornam ao modelo para confirmar que elas não são estatisticamente significativas.

4. As eventuais covariáveis significativas no passo 3 são incluídas ao modelo juntamente com aquelas do passo 2. Neste passo retorna-se com as covariáveis excluídas no passo 1 para confirmar que elas não são estatisticamente significativas.

5. Ajusta-se um modelo incluindo as covariáveis significativas no passo 4. Neste passo é testado se alguma delas pode ser retirada do modelo.

6. Utilizando as covariáveis que sobreviveram ao passo 5 ajusta-se o modelo final para os efeitos principais. Para completar a modelagem deve-se verificar a possibilidade de inclusão de termos de interação. Testa-se cada uma das interações duas a duas possíveis entre as covariáveis incluídas no modelo. O modelo final fica determinado pelos efeitos principais identificados no passo 5 e os termos de interação significativos identificados neste passo.

Ao ser utilizado este procedimento de seleção, deve-se incluir as informações clínicas no processo de decisão e evitar ser muito rigoroso ao testar cada nível individual de significância. Para decidir se um termo deve ser incluído, o nível de significância não deve ser muito baixo, sendo recomendado um valor próximo de $0,10$. Variações deste método de seleção de variáveis podem ser encontrados na literatura. Hosmer e Lemeshow (1998) discutem estes métodos com bastante elegância.

4.4.4.4 Ajuste de um modelo de regressão paramétrico

Nesta seção será utilizado métodos paramétricos para modelar o tempo até o desmame em função das covariáveis medidas. A utilização destes métodos requer a especificação de uma distribuição de probabilidade para a variável resposta. Nessa situação, o passo mais importante da modelagem é encontrar uma distribuição de probabilidade adequada para os dados em estudo. Somente após encontrar esta distribuição é que será possível estimar e testar as quantidades de interesse.

Para determinar qual distribuição de probabilidade melhor se ajusta aos dados, partiu-se da distribuição gama generalizada. Esta distribuição, como discutido na Seção 3.2.4, assume uma variedade imensa de formas pois tem dois parâmetros de forma além do parâmetro de escala. Além disso, as distribuições comumente utilizadas para modelagem de dados de sobrevivência, como a Weibull e log-normal,

são casos especiais dessa distribuição o que a torna útil na discriminação dos modelos mencionados. Adicionalmente, essa mesma distribuição, quando plausível, pode ser utilizada para descrever o estudo mas deve-se evitar este fato pela dificuldade de interpretação dos parâmetros em um modelo tão complexo.

Os passos da implementação da estratégia de seleção de covariáveis, descritos anteriormente, considerando o modelo gama generalizado, estão apresentados na Tabela 4.8 e foram obtidos no pacote estatístico *SAS*. De forma a acompanhar os passos do processo, não foi utilizado os nomes originais das covariáveis mas seus respectivos códigos identificadores, apresentados na Tabela 4.6. Em cada passo do processo de seleção de covariáveis, a estatística de teste, apresentada na Tabela 4.8, foi obtida utilizando o teste da razão de verossimilhanças com uma distribuição qui-quadrado de referência com graus de liberdade igual ao número de termos excluídos (diferença entre o número de parâmetros dos dois modelos a serem comparados).

Um ponto deve ser destacado no processo de seleção das covariáveis apresentado na Tabela 4.8. Foi observado um efeito de multicolinearidade entre as covariáveis V1 e V8. Além disso, constatou-se que os modelos que continham apenas V1 ou apenas V8 não apresentavam muita discrepância nos valores da estatística teste. Isso indica que os modelos são similares. Dessa forma, a decisão sobre qual das covariáveis deveria permanecer no modelo foi baseada em evidências clínicas. Assim, os pesquisadores decidiram por manter a covariável V1 (experiência anterior de amamentação). O modelo final ficou composto pelas covariáveis: experiência anterior de amamentação (V1), conceito materno sobre o tempo ideal de amamentação (V3), dificuldades de amamentação nos primeiros dias pós-parto (V4) e recebimento exclusivo de leite materno na maternidade (V6). Nenhum termo de interação foi significativo.

Uma vez escolhido o conjunto de variáveis prognósticas, o interesse se concentra agora em investigar a utilização de modelos mais simples, casos especiais da gama generalizada, mas não menos adequados aos dados. O teste da razão de verossimilhanças, utilizado para selecionar os modelos, apresentou os seguintes resultados:

i) adequação do modelo de regressão Weibull: $TRV = 5{,}347$ $(p = 0,0207)$

ii) adequação do modelo de regressão log-normal: $TRV = 0{,}218$ $(p = 0,6406)$.

A partir destes resultados é possível concluir que o modelo de regressão log-normal é adequado para ajustar os tempos até o desmame. Desse modo, todas as análises seguintes são baseadas nesse modelo. Vale salientar que as covariáveis selecionadas para o modelo de regressão gama generalizado são as mesmas utilizadas para o modelo de regressão log-normal.

Tabela 4.8: Seleção de covariáveis considerando o modelo gama generalizado.

Passos	Modelo	$-2 \log L$	Estatística	valor p
Passo 1	Nulo	335,540	–	–
	V1	330,235	5,305	0,0212
	V2	332,715	2,825	0,0933
	V3	329,746	5,794	0,0161
	V4	322,692	12,848	0,0003
	V5	333,756	1,784	0,1816
	V6	328,524	7,016	0,0080
	V7	333,291	2,249	0,1337
	V8	332,561	2,979	0,0843
	V9	332,592	2,948	0,0859
	V10	333,599	1,941	0,1635
	V11	333,449	2,091	0,1481
Passo 2	V1+V2+V3+V4+V6+V8+V9	304,038	–	–
	V2+V3+V4+V6+V8+V9	305,287	1,248	0,2639
	V1+V3+V4+V6+V8+V9	304,165	0,126	0,7226
	V1+V2+V4+V6+V8+V9	307,398	3,360	0,0667
	V1+V2+V3+V6+V8+V9	312,484	9,446	0,0021
	V1+V2+V3+V4+V8+V9	309,478	5,440	0,0201
	V1+V2+V3+V4+V6+V9	307,512	3,474	0,0623
	V1+V2+V3+V4+V6+V8	305,346	1.308	0,2527
Passo 3	V3+V4+V6+V8	307,485	–	–
	V3+V4+V6+V8+V1	305,529	1,956	0,1619
	V3+V4+V6+V8+V2	306,357	1,128	0,2882
	V3+V4+V6+V8+V9	306,382	1,103	0,2936
Passo 4	V3+V4+V6+V8	307,485	–	–
	V3+V4+V6+V8+V5	307,485	0,000	1,0000
	V3+V4+V6+V8+V7	305,725	1,759	0,1847
	V3+V4+V6+V8+V10	307,231	0,253	0,6149
	V3+V4+V6+V8+V11	307,322	0,163	0,6864
Passo 5	V3+V4+V6+V8	307,485	–	–
	V4+V6+V8	311.306	3,821	0,0506
	V3+V6+V8	320,594	13,109	0,0003
	V3+V4+V8	312,582	5,097	0,0239
	V3+V4+V6	312,999	5,514	0,0188
Passo 6	V3+V4+V6+V8	307,485	–	–
	V3+V4+V6+V8+V3*V4	306,777	0,708	0,4004
	V3+V4+V6+V8+V3*V6	305,678	1,807	0,1789
	V3+V4+V6+V8+V3*V8	307,206	0,279	0,5973
	V3+V4+V6+V8+V4*V6	306,735	0,750	0,3864
	V3+V4+V6+V8+V4*V8	306,740	0,745	0,3883
	V3+V4+V6+V8+V6*V8	307,200	0,285	0,5941
Etapa Final*	V3+V4+V6+V8	307,485		
	V1+V3+V4+V6	309,544		
Modelo Final	V1+V3+V4+V6	309,544		

* Escolha baseada em evidências clínicas e discussões realizadas com o pesquisador

As estimativas dos parâmetros do modelo de regressão log-normal foram obtidas no R utilizando-se o método de máxima verossimilhança

```
> ajust1<-survreg(Surv(tempo,cens)~V1+V3+V4+V6, dist='lognorm')
> ajust1
> summary(ajust1)
```

e encontram-se apresentadas na Tabela 4.9. Os coeficiente estimados, estão expressos na escala logaritímica do tempo, isto é, para $Y = \log(T) = \mathbf{X}\boldsymbol{\beta} + \sigma\nu$.

Tabela 4.9: Estimativas dos parâmetros do modelo de regressão log-normal.

Covariável	Estimativa	Erro-Padrão	Valor p
Constante	3,293	0,304	$< 0,0001$
V1: Experiência anterior de amamentação	-0,572	0,301	0,057
V3: Conceito sobre tempo de amamentação	-0,631	0,290	0,029
V4: Dificuldades amamentação pós-parto	-0,824	0,279	0,014
V6: Recebimento exclusivo de leite materno	-0,680	0,275	0,017
Parâmetro de forma	1,439	0,129	0,001

4.4.4.5 Adequação do Modelo

Antes de proceder a interpretação das estimativas dos parâmetros do modelo ajustado, é desejável utilizar os resíduos para confirmar a adequação do modelo log-normal. Os métodos gráficos são bastante utilizados para este fim como discutido na Seção 4.3.

Se o modelo log-normal para o tempo T estiver bem ajustado para estes dados, a distribuição dos resíduos na escala logaritímica ($\hat{\nu}_i$) deve estar bastante próxima da normal padrão. Como os resíduos são censurados, o estimador de Kaplan-Meier deve ser utilizado para estimar a função acumulada dos resíduos. No entanto, os resíduos $\hat{\nu}_i$ apresentam tanto valores positivos quanto negativos, e isso causa um pequeno problema para o cálculo do Kaplan-Meier em pacotes estatísticos já que estes esperam valores de uma variável estritamente positiva. Por esse motivo, aplicou-se a transformação exponencial nos resíduos $\hat{\nu}_i$, isto é, $\hat{e}_i^* = \exp\{\hat{\nu}_i\}$ que, não somente resolve o problema de estimação da função de sobrevivência, mas produz resíduos de uma distribuição conhecida, a log-normal padrão. O gráfico das sobrevivências dos resíduos estimadas por Kaplan-Meier e pelo modelo log-normal padrão bem como o gráfico de suas respectivas curvas de sobrevivência estimadas encontram-se apresentados na Figura 4.9. A partir dessa figura pode-se acreditar que o modelo de regressão log-normal se encontra bem ajustado aos dados sob análise.

A Figura 4.9 foi obtida no R utilizando-se os comandos a seguir.

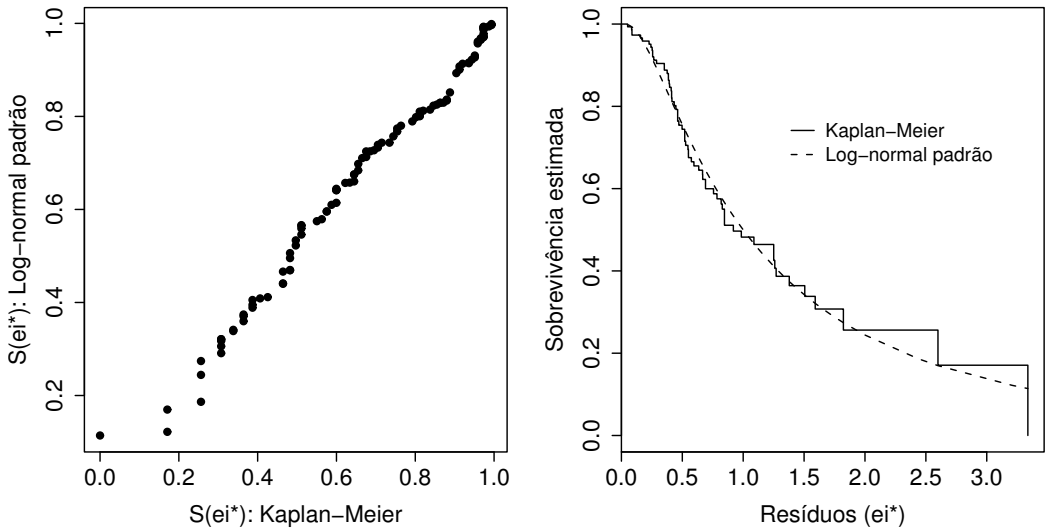

Figura 4.9: Sobrevivências dos resíduos e_i^* estimadas pelo método de Kaplan-Meier e pelo modelo log-normal padrão (gráfico à esquerda) e respectivas curvas de sobrevivência estimadas (gráfico à direita).

```
> xb<-ajust1$coefficients[1]+ajust1$coefficients[2]*V1+ajust1$coefficients[3]*V3+
      ajust1$coefficients[4]*V4+ ajust1$coefficients[5]*V6
> sigma<-ajust1$scale
> res<-(log(tempo)-(xb))/sigma                    # residuos padronizados
> resid<-exp(res)                                 # exponencial dos residuos padronizados
> ekm<- survfit(Surv(resid,cens)~1)
> resid<-ekm$time
> sln<-pnorm(-log(resid))
> par(mfrow=c(1,2))
> plot(ekm$surv,sln, xlab="S(ei*): Kaplan-Meier",ylab="S(ei*): Log-normal padrão",pch=16)
> plot(ekm,conf.int=F,mark.time=F,xlab="Resíduos (ei*)",ylab="Sobrevivência estimada",pch=16)
> lines(resid,sln,lty=2)
> legend(1.3,0.8,lty=c(1,2),c("Kaplan-Meier","Log-normal padrão"),cex=0.8,bty="n")
```

Eqüivalentemente, os resíduos de Cox-Snell deveriam seguir a distribuição exponencial padrão para que o modelo de regressão log-normal possa ser considerado adequado. A partir dos gráficos apresentados na Figura 4.10 pode-se também observar, da análise desses resíduos, que o modelo se encontra bem ajustado.

```
> ei<- -log(1-pnorm(res))                          # residuos de Cox-Snell
> ekm1<-survfit(Surv(ei,cens)~1)
> t<-ekm1$time
> st<-ekm1$surv
> sexp<-exp(-t)
> par(mfrow=c(1,2))
> plot(st,sexp,xlab="S(ei): Kaplan-Meier",ylab="S(ei): Exponencial padrão",pch=16)
> plot(ekm1,conf.int=F,mark.time=F, xlab="Resíduos de Cox-Snell", ylab="Sobrevivência estimada")
> lines(t,sexp,lty=4)
> legend(1.0,0.8,lty=c(1,4),c("Kaplan-Meier","Exponencial padrão"),cex=0.8,bty="n")
```

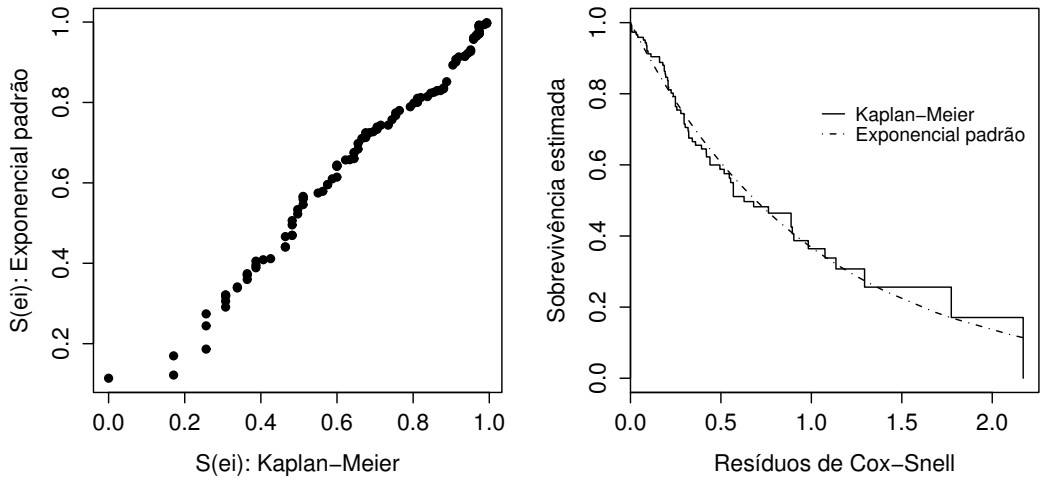

Figura 4.10: Sobrevivências dos resíduos de Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo exponencial padrão (gráfico à esquerda) e respectivas curvas de sobrevivência estimadas (gráfico à direita).

4.4.4.6 Interpretação dos coeficientes estimados

Tomando-se o exponencial dos coeficientes estimados apresentados na Tabela 4.9, obtém-se a razão dos tempos medianos de sobrevivência (Hosmer e Lemeshow, 1998). Ou seja, para uma covariável codificada (0 e 1), como são as do estudo em questão, esta razão compara o tempo mediano de sobrevivência do grupo 1 em relação ao do grupo 0. Desse modo, as interpretações dos resultados obtidos são as seguintes:

i) o tempo mediano até o desmame de mães que não tiveram experiência anterior de amamentação é aproximadamente a metade daquele das mães que já tiveram esta experiência;

ii) as mães que acreditam que o tempo ideal de amamentação é superior a seis meses têm um tempo mediano até o desmame de aproximadamente duas vezes maior do que o das mães que pensam ser esse tempo inferior ou igual a seis meses;

iii) o tempo mediano até o desmame das mães que não apresentaram dificuldades de amamentar nos primeiros dias após o parto é 2,3 vezes maior do que o tempo das que sofreram essas dificuldades e, ainda,

iv) as crianças que receberam exclusivamente leite materno na maternidade têm um tempo mediano de amamentação duas vezes maior do que o tempo daquelas

que receberam outro tipo de alimentação juntamente com o leite materno.

4.5 Exercícios

1. Os dados apresentados na Tabela 4.10 referem-se aos tempos de sobrevida, em meses, de dois grupos de pacientes com a mesma doença que foram submetidos a um de dois tratamentos alternativos (A ou B). (+ indica censura)

Tabela 4.10: Tempos de sobrevida de pacientes submetidos aos tratamentos A ou B.

Tratamentos	Tempos de sobrevida
A	1 2 2 2 2^+ 6 8 8 9 9^+ 13 13^+ 16 17 22^+ 25^* 29 34 36 43^+ 45^+
B	1 2 5 7 7^+ 11^+ 12 19 22 30 35^+ 39 42 46 55

Considerando a covariável X = tratamento recebido em que

$$X = \begin{cases} 0 & \text{se tratamento A} \\ 1 & \text{se tratamento B,} \end{cases}$$

1. Ajuste os modelos de regressão exponencial, Weibull e log-normal e verifique qual é mais adequado a esses dados.

2. Utilizando o modelo escolhido no item anterior, use o teste da razão de verossimilhanças para testar se os tratamentos A e B diferem.

3. Para o modelo final, apresente graficamente a análise dos resíduos e a(s) curva(s) de sobrevivência estimada(s).

4. Utilizando o modelo ajustado, obtenha e interprete a sobrevivência estimada em $t = 40$ meses.

Capítulo 5

Modelo Semi-Paramétrico de Riscos Proporcionais de Cox

5.1 Introdução

A modelagem em análise de sobrevivência, utilizada para avaliar o poder de explicação das covariáveis, foi tratada, em um contexto paramétrico, no Capítulo 4. Naquele capítulo foram apresentados os modelos de regressão exponencial e Weibull e estes foram então generalizados para o modelo de tempo de vida acelerado. Outro modelo, no entanto, é utilizado com freqüência na análise de dados de sobrevivência, o modelo de regressão de Cox. Este modelo é o mais utilizado em estudos clínicos por sua versatilidade e é tema deste capítulo.

O modelo de regressão de Cox (Cox, 1972) abriu uma nova fase na modelagem de dados clínicos. Uma evidência quantitativa deste fato aparece em Stigler (1994). O autor usa citações feitas a periódicos indexados de todas as áreas entre os anos de 1987 e 1989, para quantificar a importância de algumas publicações na literatura estatística. O artigo de Cox (1972), em que o modelo é apresentado, foi neste período o segundo artigo mais citado na literatura estatística, somente ultrapassado pelo artigo de Kaplan-Meier (1958). Isto significa, em números, uma média de 600 citações por ano, o que representa aproximadamente 25% das citações anuais ao *Journal of the Royal Statistical Society B*, a revista que publicou o artigo.

O objetivo deste capítulo é apresentar este importante modelo para a análise de dados de sobrevivência. Inicialmente o modelo é introduzido de forma simples e intuitiva e, em seguida, apresentado em sua forma geral. Vários aspectos relacionados ao modelo são também apresentados. Sua aplicação é ilustrada por meio da análise

de três conjuntos de dados, sendo um deles o da leucemia pediátrica descrito na Seção 1.5.2.

5.2 O Modelo de Cox

O modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando por covariáveis. No caso especial em que a única covariável é um indicador de grupos, o modelo de Cox assume a sua forma mais simples. Este caso é apresentado a seguir para introduzir a forma do modelo de Cox.

Suponha um estudo controlado que consiste na comparação dos tempos de falha de dois grupos. Os pacientes foram selecionados aleatoriamente para receber o tratamento padrão (grupo 0) ou o novo tratamento (grupo 1). A função de taxa de falha do primeiro grupo será representada por $\lambda_0(t)$ e a do segundo grupo $\lambda_1(t)$. Assumindo proporcionalidade entre estas funções, tem-se que

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K$$

em que K é a razão das taxas de falhas ou risco relativo, constante para todo tempo t de acompanhamento do estudo. Se X é a variável indicadora de grupo, em que

$$X = \begin{cases} 0 & \text{se grupo 0} \\ 1 & \text{se grupo 1} \end{cases}$$

e $K = \exp\{\beta x\}$, então

$$\lambda(t) = \lambda_0(t)\exp(\beta x) \tag{5.1}$$

ou seja,

$$\lambda(t) = \begin{cases} \lambda_1(t) = \lambda_0(t)\exp(\beta) & \text{se } x = 1 \\ \lambda_0(t) & \text{se } x = 0. \end{cases}$$

A expressão (5.1) é o modelo de Cox para uma única covariável.

De uma forma genérica, considere p covariáveis de modo que $\mathbf{x}$ é um vetor com os componentes $\mathbf{x} = (x_1, \ldots, x_p)$. A expressão geral do modelo de regressão de Cox considera

$$\lambda(t) = \lambda_0(t)\ g(\mathbf{x}'\boldsymbol{\beta}) \tag{5.2}$$

em que g é uma função que deve ser especificada, tal que $g(0) = 1$. Este modelo é composto pelo produto de dois componentes, um não-paramétrico e outro

paramétrico. O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. Ele é usualmente chamado de função de base, pois $\lambda(t) = \lambda_0(t)$ quando $\mathbf{x} = \mathbf{0}$. O componente paramétrico é freqüentemente usado na seguinte forma multiplicativa

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta}) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p) \tag{5.3}$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis. Esta forma garante que $\lambda(t)$ será sempre positiva. Outras formas para a função $g(\mathbf{x}'\boldsymbol{\beta})$ foram propostas na literatura (Storer et al., 1983). Entretanto, a forma multiplicativa é a mais utilizada e adotada neste texto. Observe que a constante β_0, presente nos modelos paramétricos, não aparece no componente mostrado em (5.3). Isto ocorre devido à presença do componente não-paramétrico no modelo que absorve este termo constante.

Este modelo é também chamado de modelo de riscos proporcionais, pois a razão das taxas de falha de dois diferentes indivíduos é constante no tempo. Isto é, a razão das funções de taxa de falha para dois indivíduos diferentes, i e j, é

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\lambda_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \exp\left\{\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_j'\boldsymbol{\beta}\right\},$$

que não depende do tempo. Por exemplo, se um indivíduo no início do estudo tem um risco de morte igual a duas vezes o risco de um segundo indivíduo, então esta razão de riscos será a mesma para todo o período de acompanhamento.

A suposição básica para o uso do modelo de regressão de Cox é, portanto, que as taxas de falha sejam proporcionais. A Figura 5.1 apresenta uma situação em que o uso deste modelo é inadequado. Esta figura mostra as curvas das taxas de falha para dois grupos. O grupo 2 tem uma alta taxa de mortalidade no início do acompanhamento. Esta taxa decresce rapidamente, ficando menor que a taxa do grupo 1 no restante do tempo. Neste caso, as taxas de falha não são proporcionais e portanto, violam a suposição básica do modelo. As curvas seriam proporcionais se elas mantivessem uma diferença constante ao longo do período de acompanhamento.

O modelo de regressão de Cox é utilizado extensivamente em estudos médicos. A principal razão desta popularidade é a presença do componente não-paramétrico, que torna o modelo bastante flexível. Um exemplo da flexibilidade deste modelo é possuir alguns conhecidos modelos paramétricos como casos particulares (Kalbleisch e Prentice, 1980). O modelo de regressão Weibull apresentado no Capítulo 4 é, por exemplo, um caso particular do modelo de Cox. Na Seção 5.8 este assunto é abordado com mais detalhes.

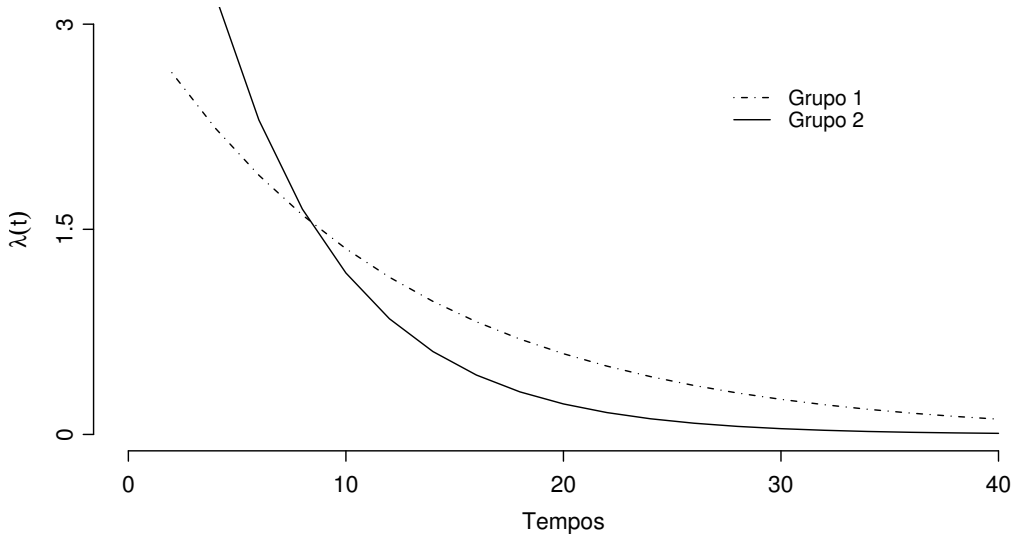

Figura 5.1: Exemplo de duas curvas de taxa de falha que não são proporcionais.

5.3 Ajustando o Modelo de Cox

O modelo de regressão de Cox é caracterizado pelos coeficientes β's, que medem os efeitos das covariáveis sobre a função de taxa de falha. Estas quantidades devem ser estimadas a partir das observações amostrais para que o modelo fique determinado.

Um método de estimação é necessário para se fazer inferência no modelo. O método de máxima verossimilhança é bastante conhecido (Cox e Hinkley, 1974) e freqüentemente utilizado para este propósito. No entanto, a presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança, torna este método inapropriado. Uma solução razoável consiste em condicionar a verossimilhança para eliminar esta função de perturbação. Foi exatamente isto que Cox propôs no seu artigo original e formalizou em um artigo subseqüente (Cox, 1975), denominando de método de máxima verossimilhança parcial.

5.3.1 Método da Máxima Verossimilhança Parcial

Usando a mesma notação dos capítulos anteriores para escrever a função de verossimilhança parcial, considere que em uma amostra de n indivíduos existam $k \leq n$ falhas distintas nos tempos $t_1 \leq t_2 \ldots \leq t_k$. Uma forma simples de entender a verossimilhança parcial considera o seguinte argumento condicional: a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i conhecendo quais observações

estão sob risco em t_i é

$$\frac{\lambda_i(t)}{\sum_{j\in R(t_i)}\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\lambda_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} \tag{5.4}$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i. Observe que condicional à história de falhas e censuras até o tempo t_i, o componente não-paramétrico $\lambda_0(t)$ desaparece de (5.4).

A função de verossimilhança a ser utilizada para fazer inferências no modelo é, então, formada pelo produto de todos os termos representados por (5.4) associados aos tempos distintos de falha, isto é,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k}\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n}\left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}\right)^{\delta_i}. \tag{5.5}$$

com δ_i o indicador de falha. Os valores de $\boldsymbol{\beta}$ que maximizam $L(\boldsymbol{\beta})$, a função de verossimilhança parcial, são obtidos resolvendo-se o sistema de equações definido por $\mathrm{U}(\boldsymbol{\beta}) = 0$, em que $\mathrm{U}(\boldsymbol{\beta})$ é o vetor escore de primeiras derivadas da função $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$. Isto é,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\delta_i\left[x_i - \frac{\sum_{j\in R(t_i)}x_j\exp\{\mathbf{x}_j'\widehat{\boldsymbol{\beta}}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\widehat{\boldsymbol{\beta}}\}}\right] = 0. \tag{5.6}$$

A função de verossimilhança parcial (5.5) assume que os tempos de sobrevivência são contínuos e, conseqüentemente, não pressupõe a possibilidade de empates nos valores observados. Na prática, podem ocorrer empates nos tempos de falhas ou censuras devido a escala de medida. Por exemplo, o tempo não é necessariamente registrado em horas, podendo em alguns estudos ser medido em dias, meses ou até mesmo em anos, dependendo da dificuldade em se obter a medida. Da mesma forma, podem ocorrer empates nas falhas ou censuras. Quando ocorrem empates entre falhas e censuras, usa-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha.

A função de verossimilhança parcial (5.5) deve ser modificada para incorporar as observações empatadas quando estas estão presentes. A aproximação para (5.5) proposta por Breslow (1972) e Peto (1972) é simples e freqüentemente usada nos pacotes estatísticos comerciais. Considere $\mathbf{s}_i$ o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i; $i = 1, \ldots, k$ e d_i o número de falhas neste mesmo tempo. A aproximação mencionada anteriormente considera a seguinte função de verossimilhança parcial

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\mathbf{s}_i'\boldsymbol{\beta})}{\left[\sum_{j \in R(t_i)} \exp(\mathbf{x}_j'\boldsymbol{\beta})\right]^{d_i}}. \tag{5.7}$$

Esta aproximação é adequada quando o número de observações empatadas em qualquer tempo não é grande. Naturalmente, a expressão (5.7) se reduz a (5.5) quando não houver empates. Outras aproximações para empates foram propostas por Efron (1977), Farewell e Prentice (1980), entre outros. Quando o número de empates em qualquer tempo é grande, o modelo de regressão de Cox para dados grupados deve ser usado (Lawless, 1982, Prentice e Gloeckler, 1978). Outras aproximações para a verossimilhança parcial na presença de empates são mostradas no Capítulo 8. Neste capítulo é também apresentado os modelos de regressão discretos que são indicados quando o número de empates é grande.

As propriedades assintóticas dos estimadores de máxima verossimilhança parcial são necessárias para a construção de intervalos de confiança e testar hipóteses sobre os coeficientes do modelo. Vários autores estudaram estas propriedades (Cox, 1975, Tsiatis, 1981), mas foram Andersen e Gill (1982) que apresentaram as provas mais gerais das propriedades destes estimadores. Eles usaram a relação entre os tempos de falhas e martingales para mostrar que estes estimadores são consistentes e assintoticamente normais sob certas condições de regularidade. Desta forma, é possível utilizar as conhecidas estatísticas de Wald e da razão de verossimilhança para fazer inferências no modelo de regressão de Cox.

5.4 Interpretação dos Coeficientes

A partir da expressão (5.2) do modelo de Cox, pode-se observar que o efeito das covariáveis é de acelerar ou desacelerar a função de risco. No entanto, a propriedade de riscos proporcionais do modelo deve ser usada para interpretar os coeficientes estimados. Tomando a razão das taxas de falhas de dois indivíduos (i e j) que têm os mesmos valores para as covariáveis com exceção da l-ésima, tem-se

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\left\{\beta_l(x_{il} - x_{jl})\right\}$$

que pode ser interpretado como a razão de riscos ou o risco relativo instantâneo no tempo t. Entretanto, como esta razão é constante para todo o acompanhamento, pode-se suprimir a palavra instantânea da interpretação. Por exemplo, suponha que x_l seja uma covariável dicotômica indicando pacientes hipertensos. O risco de

morte entre os hipertensos é $\exp(\beta_l)$ vezes o risco de pacientes com pressão normal, mantida fixas as outras covariáveis.

5.5 Estimando Funções Relacionadas a $\lambda_0(t)$

Os coeficientes de regressão $\boldsymbol{\beta}$ são as quantidades de maior interesse na modelagem estatística de dados. Entretanto, funções relacionadas com $\lambda_0(t)$ são também importantes no modelo de Cox. Funções relacionadas a $\lambda_0(t)$ referem-se basicamente à função de taxa de falha de base acumulada

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du$$

e a correspondente função de sobrevivência

$$S_0(t) = \exp(-\Lambda_0(t)).$$

A maior importância destas funções diz respeito ao uso delas em técnicas gráficas para avaliar a adequação do modelo ajustado. Isto será visto na próxima seção. A função de sobrevivência $S(t) = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}$ é também útil quando se deseja concluir a análise em termos de percentis associados a grupos de indivíduos.

Se $\lambda_0(t)$ fosse especificado parametricamente, poderia ser estimado usando a função de verossimilhança. Entretanto, na verossimilhança parcial, o argumento condicional elimina completamente esta função da verossimilhança. Desta forma, os estimadores para estas quantidades serão de natureza não-paramétrica.

Uma estimativa simples para $\Lambda_0(t)$ proposta por Breslow (1972), é uma função escada com saltos nos tempos distintos de falha e expressa por

$$\widehat{\Lambda}_0(t_i) = \sum_{j\,:\,t_j \le t} \frac{d_j}{\sum_{l\in R_j} \exp\{\mathbf{x}_l'\widehat{\boldsymbol{\beta}}\}} \tag{5.8}$$

com d_j o número de falhas em t_j. Conseqüentemente as funções de sobrevivência $S_0(t)$ e $S(t)$ podem ser estimadas a partir de (5.8) por, respectivamente,

$$\widehat{S}_0(t) = \exp\big\{-\widehat{\Lambda}_0(t)\big\}$$

e

$$\widehat{S}(t) = [\widehat{S}_0(t)]^{\exp\{\mathbf{x}'\widehat{\boldsymbol{\beta}}\}}.$$

Tanto $\widehat{S}_0(t)$ quanto $\widehat{S}(t)$ são funções escada decrescentes com o tempo. Note que na ausência de covariáveis, a expressão (5.8) reduz-se a

$$\widehat{\Lambda}_0(t_i) = \sum_{j\,:\,t_j \leq t} \left(\frac{d_j}{n_j} \right)$$

que é o estimador de Nelson-Aalen descrito no Capítulo 2. Por este fato, o estimador apresentado em (5.8) é também referenciado na literatura como estimador de Nelson-Aalen-Breslow.

5.6 Adequação do Modelo de Cox

O modelo de regressão de Cox é bastante flexível devido à presença do componente não-paramétrico. Mesmo assim, ele não se ajusta a qualquer situação clínica e como qualquer outro modelo estatístico, requer o uso de técnicas para avaliar a sua adequação. Em particular, como mencionado na Seção 5.2, ele tem uma suposição básica que é a de riscos proporcionais. A violação desta suposição pode acarretar sérios vícios na estimação dos coeficientes do modelo (Struthers e Kalbfleisch, 1986).

Alguns métodos para avaliar a adequação deste modelo são propostas na literatura. Nesta seção, alguns desses métodos serão apresentados.

5.6.1 Verificando a Suposição de Riscos Proporcionais

5.6.1.1 Métodos Gráficos

Para verificar a suposição de riscos proporcionais no modelo de Cox, um gráfico simples e bastante usado é proposto para esta finalidade. A obtenção deste gráfico consiste, inicialmente, em dividir os dados em m estratos, usualmente de acordo com alguma covariável. Por exemplo, dividir os dados em dois estratos de acordo com a covariável sexo. Em seguida, deve-se estimar $\widehat{\Lambda}_{0_j}(t)$ para cada estrato usando a expressão (5.8). Se a suposição for válida, as curvas do logaritmo de $\widehat{\Lambda}_{0_j}(t)$ *versus* t, ou $\log(t)$, devem apresentar diferenças aproximadamente constantes no tempo. Curvas não paralelas significam desvios da suposição de riscos proporcionais, como por exemplo, a situação mostrada na Figura 5.1. É razoável construir este gráfico para cada covariável incluída no estudo. Se a covariável for de natureza contínua, uma sugestão é agrupá-la em um pequeno número de categorias. Uma vantagem desta técnica gráfica é a de indicar a covariável que está gerando a violação da

suposição, caso isto ocorra.

Uma proposta adicional que vem sendo usada para verificar a suposição de riscos proporcionais no modelo de Cox é a de analisar os resíduos de Schoenfeld (1982). Para definir tais resíduos, considere que existam $k \leq n$ tempos distintos de falha $t_1 < t_2 < \cdots < t_k$. Se o indivíduo i com vetor de covariáveis $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})$ é observado falhar, tem-se para este indivíduo, um vetor de resíduos Schoenfeld $\mathbf{r}_i = (r_{1i}, r_{2i}, \cdots, r_{pi})$ em que cada componente r_{qi} $(q = 1, \cdots, p)$ é definido por

$$r_{qi} = x_{qi} - \frac{\sum_{j \in R(t_i)} x_{qj} \ \exp\{\mathbf{x}_j' \widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \widehat{\boldsymbol{\beta}}\}}. \tag{5.9}$$

Note que para cada uma das p covariáveis consideradas no modelo, tem-se, para o indivíduo i, um correspondente resíduo Schoenfeld. Como os resíduos são definidos em cada falha, o conjunto de resíduos Schoenfeld é, desse modo, uma matriz com k linhas e p colunas. Cada linha corresponde a um tempo distinto de falha e cada coluna a uma das p covariáveis consideradas no modelo. A $k-$ésima linha desta matriz é obtida por (5.9). Condicional a uma falha no conjunto de risco $R(t_i)$, o valor esperado da covariável para esta falha é expresso pelo termo $\frac{\sum_{j \in R(t_i)} x_{qj} \ \exp\{\mathbf{x}_j' \widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \widehat{\boldsymbol{\beta}}\}}$ apresentado em (5.9) e, assim, a interpretação de $\mathbf{r}_{qi}$ como um resíduo é apropriada. Como usual para resíduos, $\sum_i \mathbf{r}_i = \mathbf{0}$. Para, em um certo sentido, levar em conta uma estrutura de correlação dos resíduos, uma forma escalonada dos resíduos de Schoenfeld (*scaled Schoenfeld residuals*) é freqüentmente usada e esta é definida por

$$r_i^* = d \times i(\beta)^{-1} \times r_i$$

em que $i(\beta)$ é a matriz de informação e d é o número observado de falhas.

Se o modelo de riscos proporcionais for apropriado, os gráficos dos resíduos r_i^* *versus* t_i, para cada uma das p covariáveis, não deveriam exibir tendências ao longo do tempo t. A Figura 5.3 ilustra tais gráficos em uma situação em que duas covariáveis (X_1 e X_2) são consideradas. O gráfico à esquerda, mostrado nesta figura, não apresenta nenhuma tendência acentuada ao longo do tempo. O mesmo não se pode concluir para o gráfico à direita. A suposição de riscos proporcionais parece, portanto, não ser apropriada e há evidências de que a covariável X_2 esteja gerando a violação desta suposição.

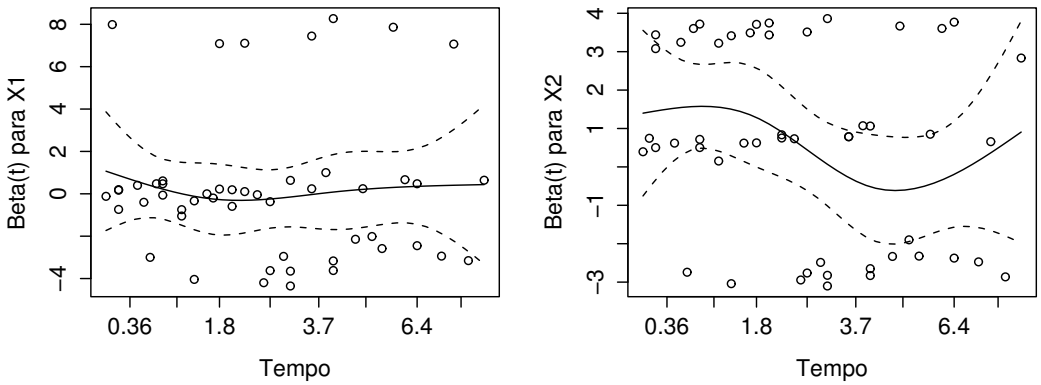

Figura 5.2: Resíduos escalonados de Schoenfeld *versus* tempo para as covariáveis X_1 (gráfico à esquerda) e X_2 (gráfico à direita).

5.6.1.2 Testes

Cox (1979) propôs um teste para examinar a suposição de riscos proporcionais. O teste consiste em acrescentar ao modelo uma covariável dependente do tempo. Covariáveis dependentes do tempo generalizam o modelo de Cox apresentado em (5.2) e serão abordadas no Capítulo 6.

Para apresentação do teste mencionado, considere um estudo clínico controlado em que cada paciente foi alocado de forma aleatória a dois grupos, um deles correspondendo ao tratamento padrão e o outro a um novo tratamento. Uma situação como esta foi apresentada na Seção 5.2. O interesse é verificar se a razão das taxas de falhas é a mesma em qualquer tempo t. Como visto, o modelo de Cox para esta situação é dado por

$$\lambda(t) = \lambda_0(t) \exp\{\beta_1 x_1\}$$

em que x_1 é a covariável indicadora de tratamento, isto é, $x_1 = 0$ se tratamento padrão e $x_1 = 1$ se tratamento novo. Como discutido anteriormente, a razão das taxas de falhas em qualquer tempo de um tratamento em relação ao outro é $\exp\{\beta_1\}$, se o modelo for adequado para os dados.

Uma outra covariável $x_2 = t$ pode ser adicionada ao modelo e, assim,

$$\lambda(t) = \lambda_0(t) \exp\{\beta_1 x_1 + \beta_2 t\}$$

de modo que a razão das taxas de falhas é agora

$$\exp\{\beta_1 + \beta_2 t\}$$

e, portanto, não é mais constante no tempo e nem o modelo é mais de riscos proporcionais. Em particular, se $\beta_2 < 0$, a razão das taxas de falhas decresce com o

tempo. Isto significa que o risco de falha usando o novo tratamento, relativo ao padrão, diminui com o tempo. Por outro lado, se $\beta_2 > 0$, o risco de falha do novo tratamento em relação ao padrão, aumenta com o tempo. No caso em que $\beta_2 = 0$, esse risco é constante e igual a $\exp\{\beta_1\}$, mostrando que esta hipótese corresponde à suposição de riscos proporcionais. Esta situação é ilustrada na Figura 5.4.

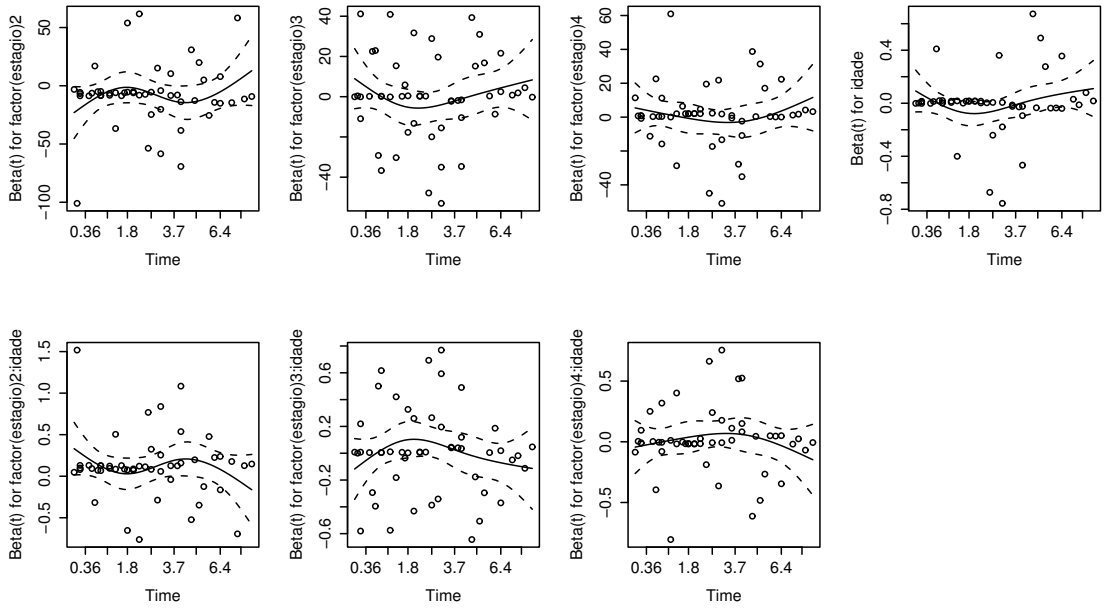

Figura 5.3: Gráfico do risco relativo $\exp\{\beta_1 + \beta_2\, t\}$ *versus* t para diferentes valores de β_2.

Modelos incluindo covariáveis dependentes do tempo, como x_2, não podem ser ajustados da mesma maneira como aqueles que incluem somente covariáveis que não mudam com o tempo. A razão disto é que estas covariáveis assumem diferentes valores em diferentes tempos complicando o cálculo do denominador da função de verossimilhança parcial apresentada em (5.5). O ajuste desses modelos é abordado no Capítulo 6.

Outros testes de adequação foram propostos para o modelo de Cox. Entretanto eles tem sérias limitações no seu uso. Alguns testes (Schoenfeld, 1980; Andersen, 1982) consideram uma partição arbitrária do eixo do tempo para a sua aplicação. Um grave problema é que diferentes partições geram testes diferentes. Outros testes, como o de Wei (1984) não necessita desta partição, entretanto, só pode ser usado no modelo com uma única covariável. Todas estas limitações associadas aos testes de adequação indicam que as técnicas gráficas envolvendo resíduos definidos adequadamente, deverão ser as ferramentas a serem utilizadas para esta finalidade.

5.6.2 Verificando outros Aspectos do Modelo de Cox

Além da suposição de riscos proporcionais, há interesse em examinar outros aspectos do modelo de Cox. Dentre eles: i) verificar o ajuste global do modelo ajustado; ii) verificar a melhor forma funcional para explicar a influência de uma dada covariável na sobrevivência, na presença das demais covariáveis; iii) verificar a presença de potenciais indivíduos atípicos (*outliers*) que, talvez, devessem ser excluídos da análise e iv) examinar a influência que cada indivíduo exerce em vários aspectos do modelo ajustado.

Para examinar os aspectos mencionados, técnicas de diagnóstico também se encontram disponíveis para o modelo de regressão de Cox. Estas baseiam-se, essencialmente, nos mesmos tipos de resíduos definidos para os modelos paramétricos apresentados no Capítulo 4. Alguns deles tais como, os resíduos de Cox-Snell, Martingale e deviance são apresentados a seguir.

5.6.2.1 Resíduos de Cox-Snell

Para o modelo de Cox, os resíduos de Cox e Snell (1968) são definidos por:

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

com $\hat{\Lambda}_0(t_i)$ estimado por (5.8). Se o modelo estiver bem ajustado, os $\hat{e}_i$'s podem ser olhados como uma amostra censurada de uma distribuição exponencial padrão e, então, o gráfico de, por exemplo, $\hat{\Lambda}(\hat{e}_i)$ versus $\hat{e}_i$ deveria ser aproximadamente uma reta. Assim como nos modelos paramétricos, os resíduos de Cox-Snell são úteis para examinar o ajuste global do modelo de Cox. Os mesmos comentários feitos na Seção 4.3.1 para esses resíduos quanto aos cuidados e desvantagem de sua utilização, são também válidos para o modelo de Cox.

O uso de gráficos envolvendo estes resíduos para verificar a suposição de riscos proporcionais não são recomendados pelas razões apresentadas a seguir. A idéia por trás da definição desses resíduos vem do fato de que a variável tempo de falha T sendo contínua e não-censurada, implica que $S_T(t)$, a função de sobrevivência de T, tem uma distribuição uniforme em $(0, 1)$. Conseqüentemente, $\Lambda_T(t) = -\log S_T(t)$ tem uma distribuição exponencial padrão. Usando a definição geral de resíduos (Cox e Snell, 1968), $\hat{e}_i = \widehat{\Lambda}_i(t_i)$ pode ser definido como o resíduo para a i-ésima observação da amostra, com tempo de falha observado em t_i. No caso de observação censurada, $\hat{e}_i = \widehat{\Lambda}_i(t_i)$ é considerada como proveniente de uma distribuição exponencial padrão censurada. Supostamente, espera-se que o gráfico destes resíduos em um

papel de probabilidade exponencial padrão seja uma reta se o modelo ajustado for adequado. Por outro lado, afastamentos da reta indicariam que o modelo é inapropriado para a situação estudada. Entretanto, Crowley e Storer (1983) mostraram, usando simulações de Monte Carlo, que não existe consistência entre afastamentos da exponencial padrão e a adequação do modelo ajustado. Em particular, ajustando o modelo sem nenhuma covariável resulta exatamente nas estatísticas de ordem de uma amostra proveniente de uma exponencial padrão. Isto mostra que estes resíduos não são apropriados para verificar a adequação do modelo ajustado, em particular a suposição de riscos proporcionais.

Outras definições de resíduos tais como os resíduos generalizados de Barlow e Prentice (1988) e os de Schoenfeld (1982), este último apresentado anteriormente, têm um grande potencial para desempenhar bem esta função.

5.6.2.2 Resíduos Martingale

Como visto no Capítulo 4, os resíduos martingale resultam de uma modificação dos resíduos de Cox-Snell. Assim, quando os dados apresentam censuras à direita e todas as covariáveis são fixadas no início do estudo, ou seja, não forem dependentes do tempo, os resíduos martingale para o modelo de Cox são definidos por:

$$\hat{m}_i = \delta_i - \hat{\Lambda}_0(t_i) \exp\left\{ \sum_{k=1}^{p} x_{ip}\hat{\beta}_k \right\} = \delta_i - \hat{e}_i, \qquad i = 1, \cdots, n$$

Esses resíduos podem ser usados para verificar a adequação do modelo mas, na prática, são usados para verificar a forma funcional das covariáveis, isto é, se estas deveriam ser usadas no modelo como $\log(x_i)$, x_i^2, e assim por diante, em vez de x_i.

5.6.2.2 Resíduos Deviance

Os resíduos deviance no modelo de Cox são definidos por

$$\hat{d}_i = \text{sinal}(\hat{m}_i)\Big[-2\Big(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i)\Big)\Big]^{1/2}. \tag{5.10}$$

Esses resíduos facilitam, em geral, a detecção de pontos atípicos (*outliers*) e deveriam apresentar um comportamento aleatório em torno de zero caso o modelo seja apropriado.

5.7 Aplicações

5.7.1 Análise de um Estudo sobre Câncer de Laringe

Neste exemplo os dados considerados referem-se a um estudo, descrito em Klein e Moeschberger (1997), realizado com 90 pacientes do sexo masculino diagnosticados no período de 1970-1978 com câncer de laringe e que foram acompanhados até 01/01/1983. Para cada paciente foram registrados, no diagnóstico, a idade (em anos) e o estágio da doença (I = tumor primário, II = envolvimento de nódulos, III = metástases e IV = combinações dos 3 estágios anteriores) bem como seus respectivos tempos de falha ou censura (em meses). Os estágios encontram-se ordenados pelo grau de seriedade da doença (menos sério para mais sério).

Utilizando-se o modelo de Cox para a análise desses dados, foram ajustados diversos modelos seqüências cujos resultados, obtidos no R por meio dos comandos abaixo, encontram-se apresentados na Tabela 5.1.

```
> laringe<-read.table("c:/Temp/laringe.txt", h=T)
> attach(laringe)
> require(survival)
> fit2<-coxph(Surv(tempos,cens)~factor(estagio), data=laringe,
        x = T, method="breslow")
> summary(fit2)
> fit2$loglik
> fit3<- coxph(Surv(tempos,cens)~factor(estagio)+ idade, data=laringe,
         x = T, method="breslow")
> summary(fit3)
> fit3$loglik
> fit4<-coxph(Surv(tempos,cens) ~ factor(estagio) + idade + factor(estagio)*idade,
        data=laringe, x = T, method="breslow")
> summary(fit4)
> fit4$loglik
```

A partir da Tabela 5.1 tem-se, para o teste da razão de verossimilhança parcial associado à interação entre estágio e idade, o resultado $TRV = 6{,}2038$ ($p = 0{,}1021$, g.l. = 3) indicando que esta interação seria não significativa. No entanto, análise dos resultados dos testes locais dessa interação, apresentados na Tabela 5.2, mostram evidências de que pelo menos um dos β's associados à referida interação difere significativamente de zero, no caso β_5 com p-valor = 0,022.

Em conseqüência dos resultados encontrados decidiu-se pela realização dos diagnósticos dos resíduos de Schoenfeld dos modelos de Cox com, e sem, a presença da interação para, então, proceder a escolha de um desses dois modelos.

Tabela 5.1: Estimativas obtidas para os modelos de Cox ajustados aos dados de laringe.

Modelo	Covariáveis no modelo	Estimativas	Log verossimilhança parcial
1	nenhuma	-	l_1 = -197,2129
2	X_1: estágio (II)	$\hat{\beta}_1$ = 0,0658	
	(III)	$\hat{\beta}_2$ = 0,6121	
	(IV)	$\hat{\beta}_3$ = 1,7228	l_2 = -189,0812
3	X_1: estágio (II)	$\hat{\beta}_1$ = 0,1386	
	(III)	$\hat{\beta}_2$ = 0,6383	
	(IV)	$\hat{\beta}_3$ = 1,6931	
	X_2: idade	$\hat{\beta}_4$ = 0,0189	l_3 = -188,1794
4	X_1: estágio (II)	$\hat{\beta}_1$ = -7,9461	
	(III)	$\hat{\beta}_2$ = -0,1225	
	(IV)	$\hat{\beta}_3$ = 0,8470	
	X_2: idade	$\hat{\beta}_4$ = -0,0026	
	$X_1 * X_2$ (II* id)	$\hat{\beta}_5$ = 0,1203	
	(III* id)	$\hat{\beta}_6$ = 0,0114	
	(IV* id)	$\hat{\beta}_7$ = 0,0137	l_4 = -185,0775

Tabela 5.2: Testes locais dos parâmetros associados à interação.

parâmetro	estimativa	erro padrão	z	p-valor
β_5 = idade:est2	0,1203	0,0523	2,2990	**0,022**
β_6 = idade:est3	0,0114	0,0374	0,3031	0,760
β_7 = idade:est4	0,0137	0,0360	0,3802	0,700

Desse modo, e utilizando-se dos resíduos escalonados de Schoenfeld do modelo de Cox com a interação, foram obtidos

```
> residuals.coxph(fit4,type="scaledsch")
> cox.zph(fit4)
> par(mfrow=c(2,4))
> plot(cox.zph(fit4))
```

os resultados apresentados na Tabela 5.3 e Figura 5.4. Dos resultados apresentados nesta tabela pode-se observar que os valores obtidos para ρ são próximos de zero e que a hipótese nula de não tendência no tempo, para cada termo no modelo, isto é, H_0: $\rho = 0$ não é rejeitada para nenhum deles (todos os valores $p > 0{,}40$). Observando-se os gráficos apresentados na Figura 5.4 pode-se, visualmente, confirmar este fato. A suposição de riscos proporcionais não é, desse modo, rejeitada para esse modelo, uma vez que tendências ao longo do tempo não são evidentes. O

Tabela 5.3: Testes para tendências no modelo de Cox com a interação.

	rho (ρ)	chisq	p
estágio II	0,1126	0,69477	0,405
estágio III	0,0278	0,05705	0,811
estágio IV	0,0112	0,00736	0,932
idade	0,0894	0,64048	0,424
estágio II * idade	-0,1105	0,67773	0,410
estágio III * idade	-0,0598	0,26537	0,606
estágio IV * idade	-0,0276	0,04479	0,832

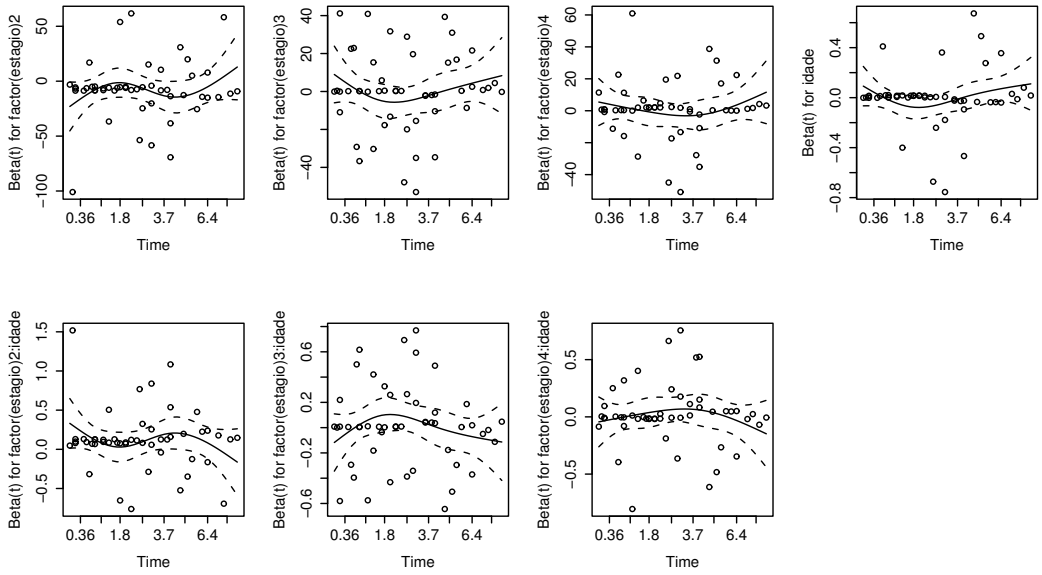

Figura 5.4: Resíduos escalonados de Schoenfeld do modelo de Cox com a interação.

modelo de Cox com a presença da interação apresenta-se, desse modo, como uma opção satisfatória para a análise dos dados desse exemplo.

De modo análogo, foram obtidos, para o modelo de Cox sem a presença da interação, os resultados apresentados na Tabela 5.4 e Figura 5.5.

```
> residuals.coxph(fit3,type="scaledsch")
> cox.zph(fit3)
> par(mfrow=c(1,4))
> plot(cox.zph(fit3))
```

Note desta tabela que, embora não significativo ao nível de 5% de significância, o estágio III é marginalmente significativo ($p = 0{,}0736$) sugerindo uma possível falha da suposição de riscos proporcionais para este nível da covariável.

Tabela 5.4: Testes para tendências no modelo de Cox sem a interação.

	rho (ρ)	chisq	p
estágio II	-0,0163	0,0140	0,9057
estágio III	-0,2591	3,2005	0,0736
estágio IV	-0,1100	0,5336	0,4651
idade	0,1138	0,8584	0,3542

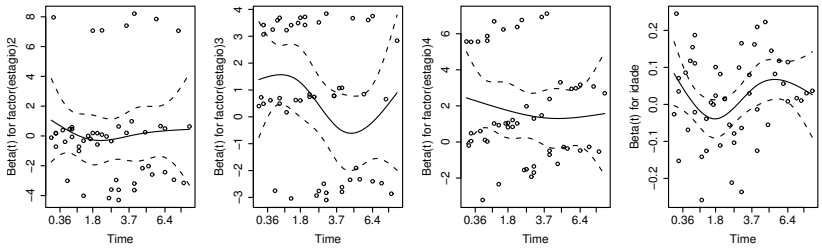

Figura 5.5: Resíduos escalonados de Schoenfeld do modelo de Cox sem interação.

Para amostras grandes, o que não é o caso do estudo aqui analisado, cabe ressaltar que para os resultados dos testes de tendências, atenção maior deve ser dada aos valores de ρ uma vez que, em sendo o tamanho amostral grande, significância estatística será indicada mesmo para valores de ρ muito próximos de zero.

Comparando-se então os resultados de ambos os diagnósticos apresentados, decidiu-se pelo uso do modelo de Cox com a presença da interação.

5.7.1.1 Funções de sobrevivência e risco estimadas

Para o modelo escolhido, as funções de sobrevivência e de risco estimadas são expressas por, respectivamente,

$$\widehat{S}(t \mid \mathbf{x}) = \begin{cases} \left[\widehat{S}_0(t)\right]^{\exp\{\hat{\beta}_4\, x_2\}} & \text{se estágio I} \\ \left[\widehat{S}_0(t)\right]^{\exp\{\hat{\beta}_1+(\hat{\beta}_4+\hat{\beta}_5)\, x_2\}} & \text{se estágio II} \\ \left[\widehat{S}_0(t)\right]^{\exp\{\hat{\beta}_2+(\hat{\beta}_4+\hat{\beta}_6)\, x_2\}} & \text{se estágio III} \\ \left[\widehat{S}_0(t)\right]^{\exp\{\hat{\beta}_3+(\hat{\beta}_4+\hat{\beta}_7)\, x_2\}} & \text{se estágio IV} \end{cases}$$

e

$$\widehat{\alpha}(t \mid \mathbf{x}) = \begin{cases} \widehat{\alpha}_0(t)\ \exp\{\hat{\beta}_4\, x_2\} & \text{se estágio I} \\ \widehat{\alpha}_0(t)\ \exp\{\hat{\beta}_1 + (\hat{\beta}_4+\hat{\beta}_5)\, x_2\} & \text{se estágio II} \\ \widehat{\alpha}_0(t)\ \exp\{\hat{\beta}_2 + (\hat{\beta}_4+\hat{\beta}_6)\, x_2\} & \text{se estágio III} \\ \widehat{\alpha}_0(t)\ \exp\{\hat{\beta}_3 + (\hat{\beta}_4+\hat{\beta}_7)\, x_2\} & \text{se estágio IV} \end{cases}$$

em que x_2 = idade.

As estimativas $\widehat{S}_0(t)$ e $\widehat{\alpha}_0(t)$ são, como pode ser observado nas expressões das funções de sobrevivência e risco, necessárias para obtenção de suas respectivas estimativas. Estas estimativas bem como as estimativas $\widehat{\Lambda}_0(t)$ encontram-se apresentadas na Tabela 5.5 e foram obtidas no R por:

```
> ss<-survfit(fit4)
> round(ss$surv,digits=5)       # S(t|x) para x = xbar (default R) #
> b<-fit4$coefficients
> b<-as.vector(b)
> x<- fit4$x
> xbar<-as.matrix(apply(x,2,mean))
> embx<-exp(-sum(b*xbar))
> s0<-(ss$surv)^embx
> H0<- -log(s0)
> x1<-as.matrix(H0)
> n<-nrow(x1)
> a0<-rep(0,n)
> for(i in 1:n){a0[i]<-H0[i+1] - H0[i]}
> alpha0<-c(H0[1],a0[1:(n-1)])
> alpha0<-c(H0[1],a0[1:(n-1)])
> round(cbind(ss$time,s0,alpha0,H0),digits=5)
```

Tabela 5.5: Estimativas de $S_0(t)$, $\alpha_0(t)$ e $\Lambda_0(t)$ para os dados de laringe.

	Tempo	$\widehat{S}_0(t)$	$\widehat{\alpha}_0(t)$	$\widehat{\Lambda}_0(t)$		Tempo	$\widehat{S}_0(t)$	$\widehat{\alpha}_0(t)$	$\widehat{\Lambda}_0(t)$
1	0,1	0,99377	0,00625	0,00625	18	3,2	0,77965	0,02581	0,24891
2	0,2	0,98739	0,00644	0,01269	19	3,3	0,76923	0,01345	0,26236
3	0,3	0,96737	0,02049	0,03318	20	3,5	0,73805	0,04138	0,30374
4	0,4	0,96039	0,00723	0,04041	21	3,6	0,71695	0,02901	0,33274
5	0,5	0,95319	0,00753	0,04794	22	3,8	0,70498	0,01684	0,34959
6	0,6	0,94596	0,00761	0,05555	23	4,0	0,66650	0,05613	0,40572
7	0,7	0,93875	0,00766	0,06321	24	4,3	0,65242	0,02135	0,42707
8	0,8	0,91713	0,02329	0,08650	25	5,0	0,63406	0,02854	0,45561
9	1,0	0,90154	0,01715	0,10365	26	5,3	0,61308	0,03364	0,48925
10	1,3	0,88552	0,01792	0,12158	27	6,0	0,59024	0,03798	0,52723
11	1,5	0,87745	0,00916	0,13073	28	6,2	0,56680	0,04051	0,56775
12	1,6	0,86907	0,00959	0,14033	29	6,3	0,54126	0,04611	0,61386
13	1,8	0,85231	0,01948	0,15980	30	6,4	0,48988	0,09974	0,71360
14	1,9	0,83549	0,01993	0,17974	31	6,5	0,46291	0,05661	0,77022
15	2,0	0,81848	0,02056	0,20030	32	7,0	0,43005	0,07363	0,84384
16	2,3	0,80945	0,01109	0,21140	33	7,4	0,39625	0,08187	0,92571
17	2,4	0,80004	0,01170	0,22309	34	7,8	0,35937	0,09769	1,02341

Na Figura 5.6 encontram-se representadas as curvas de sobrevivência estimadas para pacientes com idades de 50 e 65 anos, em cada um dos 4 estágios da doença. Desta figura pode-se observar que as curvas de sobrevivência estimadas para os estágios I, III e IV não apresentam diferenças muito acentuadas quando comparadas para as idades de 50 e 65 anos. No estágio II, contudo, observa-se um decréscimo expressivo desta curva quando comparada para pacientes de 50 e 65 anos de idade. Este fato justifica, assim, a presença da interação entre o estágio e idade no modelo, em especial entre idade e o estágio II.

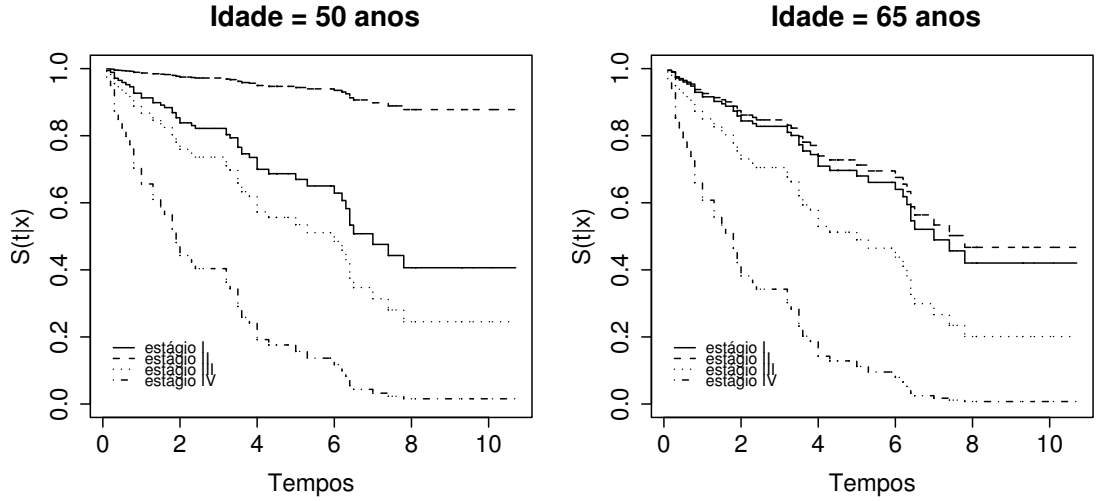

Figura 5.6: Sobrevivências estimadas pelo modelo de Cox para os dados de laringe.

Na Figura 5.7 encontram-se representadas as correspondentes curvas dos riscos estimados para pacientes com idades de 50 e 65 anos, em cada um dos 4 estágios da doença.

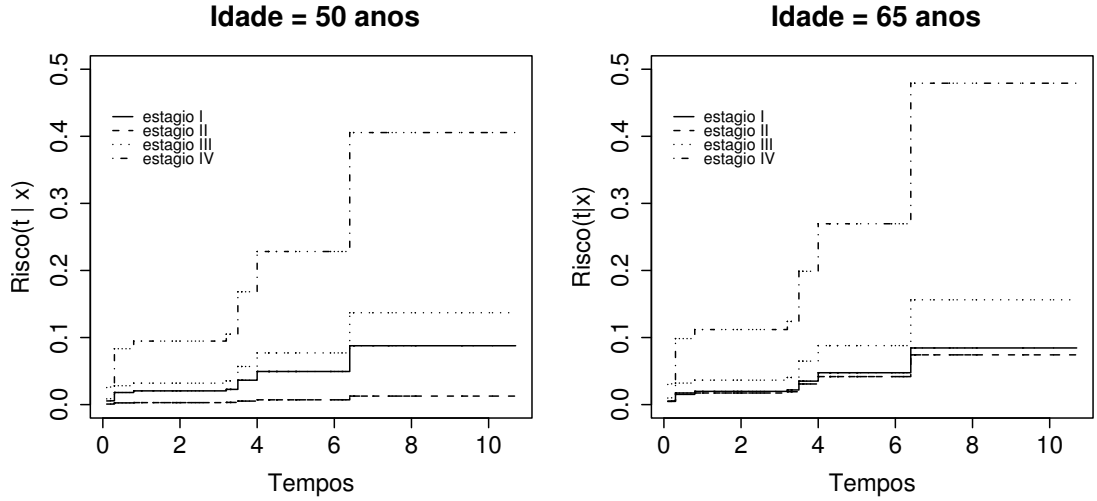

Figura 5.7: Riscos estimados pelo modelo de Cox para os dados de laringe.

Assim, por exemplo, para os pacientes i e l em que ambos encontram-se no estágio II da doença, mas um deles apresenta idade de 65 anos e o outro de 50 anos, tem-se que a razão de riscos entre eles é de:

$$\begin{aligned}\frac{\alpha(t \mid \mathbf{x}_i)}{\alpha(t \mid \mathbf{x}_l)} &= \frac{\exp\left\{\hat{\beta}_1 + (\hat{\beta}_4 + \hat{\beta}_5) * 65\right\}}{\exp\left\{\hat{\beta}_1 + (\hat{\beta}_4 + \hat{\beta}_5) * 50\right\}} \\ &= \exp\left\{(\hat{\beta}_4 + \hat{\beta}_5) * (65 - 50)\right\} = 5,844\end{aligned}$$

o que significa que o risco de falha de pacientes com 65 anos de idade e no estágio II da doença, é de aproximadamente 6 vezes maior do que o risco de falha de pacientes com 50 anos e no mesmo estágio da doença.

Por outro lado, tem-se, por exemplo, para os pacientes j e k em que ambos têm 50 anos de idade mas um deles encontra-se no estágio IV da doença e outro no estágio III, que a razão de riscos entre eles é de:

$$\frac{\alpha(t \mid \mathbf{x}_j)}{\alpha(t \mid \mathbf{x}_k)} = \frac{\exp\left\{\hat{\beta}_3 + (\hat{\beta}_4 + \hat{\beta}_7) * 50\right\}}{\exp\left\{\hat{\beta}_2 + (\hat{\beta}_4 + \hat{\beta}_6) * 50\right\}} = 2,9611.$$

Desse modo, tem-se que o risco de falha de pacientes com 50 anos de idade e no estágio IV da doença é de aproximadamente 3 vezes maior do que o risco de falha de pacientes também com 50 anos de idade mas que se encontram no estágio III da doença.

Razões de risco para todas as demais comparações de interesse podem ser obtidas e discutidas de forma análoga a apresentada. No apêndice B o leitor encontra os comandos utilizados no R para obtenção das Figuras 5.6 e 5.7.

5.7.2 Análise dos Dados de Aleitamento Materno

No Capítulo 4, após análise descritiva e exploratória das variáveis, métodos paramétricos foram utilizados para modelar o tempo máximo de aleitamento materno em função das covariáveis registradas no estudo. Dentre os modelos analisados, o modelo log-normal foi encontrado ser o mais adequado para ajustar os tempos até o desmame. Fazendo uso da estratégia de seleção de covariáveis, descrita na Seção 4.4.4.3, permaneceram no modelo final as covariáveis: experiência anterior de amamentação (V1), conceito materno sobre o tempo ideal de amamentação (V3), dificuldades de amamentação nos primeiros dias pós-parto (V4) e recebimento exclusivo de leite materno na maternidade (V6).

De forma alternativa, a modelagem do tempo até o desmame pode ser feita com base no modelo semi-paramétrico de Cox apresentado neste capítulo. Con-

siderando então este modelo, os passos da implementação da estratégia de seleção das covariáveis encontram-se apresentados na Tabela 5.6.

Tabela 5.6: Seleção de covariáveis usando o modelo de regressão de Cox.

Passos	Modelo	$-2 \log L$	Estatística	valor p
Passo 1	Nulo	560,628	–	–
	V1	556,958	3,670	0,0554
	V2	557,922	2,706	0,1000
	V3	554,920	5,708	0,0169
	V4	549,455	11,173	0,0008
	V5	559,402	1,226	0,2682
	V6	554,008	6,620	0,0101
	V7	558,420	2,208	0,1373
	V8	558,617	2,011	0,1562
	V9	558,597	2,031	0,1541
	V10	558,137	2,491	0,1145
	V11	557,872	2,756	0,0969
Passo 2	V1+V2+V3+V4+V6+V11	536,196	–	–
	V2+V3+V4+V6+V11	538,771	2,575	0,2358
	V1+V3+V4+V6+V11	536,196	0,000	1,0000
	V1+V2+V4+V6+V11	541,104	4,908	0,0267
	V1+V2+V3+V6+V11	543,629	7,433	0,0064
	V1+V2+V3+V4+V11	540,242	4,046	0,0443
	V1+V2+V3+V4+V6	536,346	0,150	0,6985
Passo 3	V3+V4+V6	539,433	–	–
	V3+V4+V6+V1	536,347	3,086	0,0790
	V3+V4+V6+V2	538,823	0,610	0,4348
	V3+V4+V6+V11	539,359	0,074	0,7856
Passo 4	V3+V4+V6+V1	536,347	–	–
	V3+V4+V6+V1+V5	536,076	0,271	0,6027
	V3+V4+V6+V1+V7	534,108	2,239	0,1346
	V3+V4+V6+V1+V8	533,257	3,090	0,0788
	V3+V4+V6+V1+V9	535,012	1,335	0,2479
	V3+V4+V6+V1+V10	536,268	0,079	0,7787
Passo 5	V1+V3+V4+V6+V8	533,257	–	–
	V3+V4+V6+V8	534,492	1,235	0,2497
	V1+V4+V6+V8	538,540	5,283	0,0215
	V1+V3+V6+V8	542,136	8,879	0,0029
	V1+V3+V4+V8	538,172	4,915	0,0266
	V1+V3+V4+V6	536,347	3,090	0,0788
Passo 6	V1+V3+V4+V6	536,347	–	–
	V1+V3+V4+V6+V1*V3	535,922	0,425	0,5145
	V1+V3+V4+V6+V1*V4	536,123	0,224	0,6360
	V1+V3+V4+V6+V1*V6	536,005	0,342	0,5587
	V1+V3+V4+V6+V3*V4	535,136	1,211	0,2711
	V1+V3+V4+V6+V3*V6	534,673	1,674	0,1957
	V1+V3+V4+V6+V4*V6	535,873	0,474	0,4912
Modelo Final	V1+V3+V4+V6	536,347		

Após o processo de seleção, o modelo de Cox resultante, incluiu o mesmo conjunto de covariáveis identificadas pelo modelo paramétrico (V1, V3, V4 e V6). Este

fato mostra que tais covariáveis são realmente importantes para descrever o comportamento do tempo até o desmame. Os comandos usados no R para obtenção dos resultados apresentados na Tabela 5.6 para, por exemplo, o modelo final foram:

```
> require(survival)
> desmame<-read.table("c:/Temp/desmame.txt",h=T)
> fit<-coxph(Surv(tempo,cens)~V1+V3+V4+V6,data=desmame,x = T,method="breslow")
> summary(fit)
> fit$loglik
```

5.7.2.1 Adequação do Modelo

Como discutido anteriormente, a suposição de riscos proporcionais deve ser atendida para que o modelo de Cox possa ser considerado adequado aos dados desse estudo. Dois métodos gráficos foram apresentados para essa finalidade, um deles envolvendo o logaritmo da função de risco acumulada e outro os resíduos de Schoenfeld. Em ambos os métodos, um gráfico deve ser construído para cada covariável incluída no modelo final.

Na Figura 5.8 encontram-se os gráficos envolvendo o logaritmo da função de risco acumulada para as covariáveis V1, V3, V4 e V6. Como pode ser observado desta figura, as curvas não indicam violação da suposição de riscos proporcionais. Embora as mesmas não sejam perfeitamente paralelas ao longo do eixo do tempo, não existem, em termos descritivos, afastamentos marcantes desta característica. A situação extrema de violação é caracterizada por curvas que se cruzam.

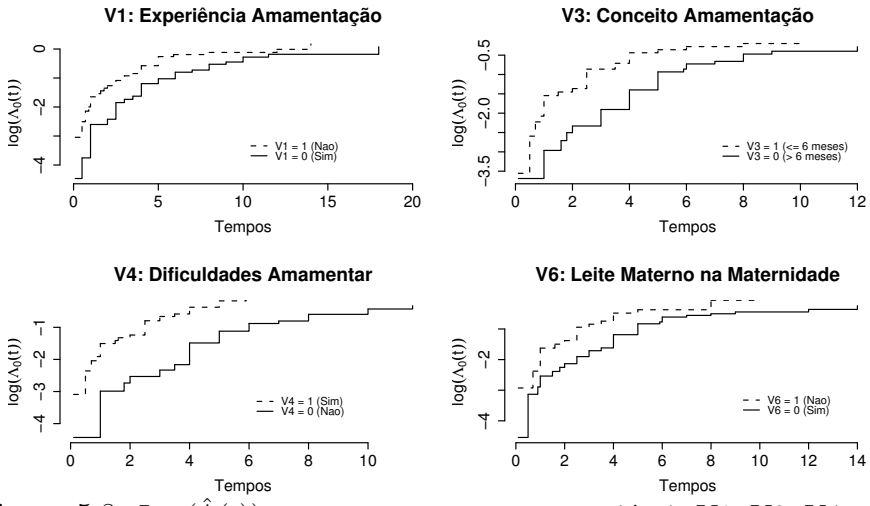

Figura 5.8: $\text{Log}(\hat{\Lambda}(t))$ *versus* tempo para as covariáveis V1, V3, V4 e V6.

No apêndice B o leitor encontra os comandos utilizados no R para obtenção da Figura 5.8. Os resíduos escalonados de Schoenfeld encontram-se por sua vez

apresentados na Figura 5.9. Desta figura pode-se observar a ausência de tendências acentuadas para qualquer uma das covariáveis presentes no modelo. Desse modo, a análise desses resíduos mostram também não haver evidências de violações da suposição de riscos proporcionais.

```
> residuals.coxph(fit,type="scaledsch")
> cox.zph(fit)
> par(mfrow=c(2,2))
> plot(cox.zph(fit))
```

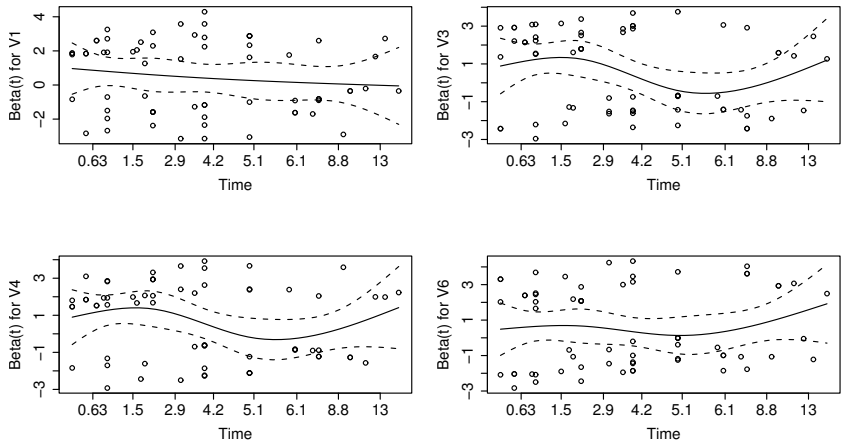

Figura 5.9: Suposição de riscos proporcionais para as covariáveis V1, V3, V4 e V6 fazendo uso dos resíduos escalonados de Schoenfeld.

5.7.2.2 Resultados Finais e Interpretação

Os resultados obtidos do ajuste do modelo de riscos proporcionais de Cox com as covariáveis selecionadas, isto é, V1, V3, V4 e V6, encontram-se apresentados na Tabela 5.7.

Tabela 5.7: Resultado do ajuste do modelo de regressão de Cox para os dados de aleitamento materno.

Covariável	Estimativa	Erro-Padrão	Valor-p	RR	IC(RR, 95%)
V1: Exper. Amam.	0,471	0,268	0,079	1,601	(0,94; 2,71)
V3: Conc. Amam.	0,579	0,262	0,027	1,785	(1,07; 2,99)
V4: Dific. Amam.	0,716	0,264	0,007	2,046	(1,22; 3,43)
V6: Leite Excl.	0,578	0,264	0,028	1,783	(1,06; 2,99)

As seguintes interpretações podem ser obtidas a partir da Tabela 5.7:

i) O risco de desmame precoce em mães que não tiveram experiência anterior de amamentação é 1,6 vezes o risco das mães que tiveram essa experiência. Além disso, podemos afirmar com 95% de confiança que esse risco varia entre $0,95$ e $2,71$.

ii) O risco de desmame precoce em mães que acreditam que o tempo ideal de amamentação é menor ou igual a 6 meses é aproximadamente 1,8 vezes o risco das mães que acreditam que o tempo ideal de amamentação é superior a 6 meses. Além disso, podemos afirmar com 95% de confiança que esse risco varia entre $1,07$ e $2,99$.

iii) O risco de desmame precoce em mães que apresentaram dificuldades de amamentar nos primeiros dias pós-parto é aproximadamente 2 vezes o risco das mães que não apresentaram essas dificuldades. Além disso, podemos afirmar com 95% de confiança que esse risco é superior a $1,22$.

iv) O risco de desmame precoce em crianças que não receberam exclusivamente leite materno na maternidade é 1,8 vezes o risco de desmame precoce em crianças que receberam exclusivamente o leite materno. Além disso, podemos afirmar com 95% de confiança que esse risco varia entre $1,06$ e $2,99$.

5.7.2.3 Considerações Finais

Os modelos de Cox e paramétrico log-normal foram utilizados na análise dos dados de aleitamento materno com o intuito de identificar as covariáveis associadas ao tempo até o desmame. Dentre aquelas registradas no estudo, os dois modelos foram consistentes nos resultados, identificando o mesmo conjunto de fatores explicativos. Assim, o resultado alcançado para os ajustes confirmaram aquilo que se esperava, que independente da estrutura de modelagem, as varivéis que melhor explicam a resposta (tempo até o desmame) são as mesmas. Além disso, destaca-se o fato de que as estimativas para os parâmetros das covariáveis de cada modelo apontam na mesma direção. Isto significa que apesar dos sinais dos coeficientes estimados serem contrários eles apontam na mesma direção em termos de interpretação. Isto ocorre devido a estrutura distinta de cada modelo. No caso do modelo de Cox a função de risco é modelada e no caso do modelo log-normal é a própria resposta. No primeiro modelo, um coeficiente positivo indica um aumento da taxa de falha e por conseqüência uma redução do tempo até a falha. Esta é a razão dos coeficientes com sinais contrários.

Frente a dois modelos com estruturas diferentes a interpretação dos coeficientes é realizada de acordo com a forma do modelo. Isto pôde ser observado ao longo da análise quando interpretação no modelo de Cox foi feita em termos de razão de taxas de falha e no modelo paramétrico em termos de razão de tempos medianos de falha. Desta forma, não pode-se comparar a ordem de grandeza dos coeficientes estimados.

Os modelos paramétricos, se bem ajustados, devem produzir resultados mais precisos do que os do modelo de Cox. Isto acontece devido ao caráter semi-paramétrico do modelo de Cox. Ou seja, a estimação utilizando o método de máxima verossimilhança parcial exclui parte da informação da amostra pois baseia-se nos postos das observações. Isto foi mostrado por Cox (1975) quando na construção da verossimilhança parcial. Ele indica que partindo da verossimilhança usual parte desta última é descartada para formar a parcial. Isto pode ser constatado na comparação dos ajustes dos dois modelos apresentados nas Tabelas 4.9 e 5.7. Isto não pode ser feito simplesmente comparando as estimativas dos erros padrões pois os coeficientes estimados são diferentes como já dito. No entanto, está comparação pode ser realizada comparando as estatísticas de teste que estão na mesma unidade. O que se pode constatar fazendo isto é que o modelo paramétrico log-normal realmente apresenta valores maiores para estas estatísticas confirmando a maior precisão destes modelos. Entretanto, a diferença é bastante pequena indicando que a perda de precisão do modelo de Cox é mínima e certamente o ganho dele em termos de flexibilidade compensa largamente esta perda.

5.7.3 Análise dos Dados de Leucemia Pediátrica

Nesta seção, os dados de leucemia em crianças, descrito na Seção 1.5.2, é analisado por meio do modelo de riscos proporcionais de Cox. As covariáveis consideradas nesta análise foram medidas na data do diagnóstico e encontram-se apresentadas na Tabela 5.8. Desta tabela pode-se notar que todas as covariáveis foram dicotomizadas sendo a categoria inferior representada por 0 e a superior por 1. Esta categorização é arbitrária mas deve ser explicitada a fim de que seja feito a interpretação dos resultados. Isto significa, que é possível utilizar qualquer representação destas variáveis categóricas. Os resultados irão registrar esta configuração mas as conclusões serão exatamente as mesmas.

Nesta análise estão incluídas 103 crianças com leucemia. Dezessete crianças foram excluídas por apresentarem valores omissos em pelo menos uma das covariáveis listadas na Tabela 5.8. Este conjunto de dados, com as covariáveis não

Tabela 5.8: Descrição das covariáveis utilizadas no estudo sobre leucemia pediátrica.

Código	Descrição	Categorias
LEUINI	Número de leucócitos no sangue periférico	0 se $\leq$ 75000 leucócitos/mm^3
		1 se $>$ 75000 leucócitos/mm^3
IDADE	Idade em meses	0 se $\leq$ 96 meses
		1 se $>$ 96 meses
ZPESO	Peso padronizado pela idade e sexo	0 se $\leq$ -2 e 1 se $>$ -2
ZEST	Altura padronizada pela idade e sexo	0 se $\leq$ -2 e 1 se $>$ -2
PAS	Porcentagem de linfoblastos medulares que reagiram positivamente ao ácido periódico de Schiff	0 se $\leq$ 5% e 1 se $>$ 5%
VAC	Porcentagem de vacúolos no citoplasma dos linfoblastos	0 se $\leq$ 15% e 1 se $>$ 15%
RISK	Fator de risco obtido a partir de uma fórmula que é função dos tamanhos do fígado e do baço e do número de blastos	0 se $\leq$ 1,7% e 1 se $>$ 1,7%
R6	Remissão na sexta semana de tratamento	0 se não e 1 se sim

dicotomizadas, é apresentado no Apêndice A. Nas análises, contudo, as covariáveis encontram-se dicotomizadas.

Assumindo que o modelo de Cox é adequado para estes dados, foram obtidos, no *R*, os resultados apresentados na Tabela 5.9. A segunda coluna desta tabela corresponde às estimativas de máxima verossimilhança parcial. Os valores-p apresentados na última coluna da Tabela 5.9 correspondem ao teste de Wald.

```
> leuc<-read.table("c:/Dados/leucemia.txt", h=T)    ## lendo no R os dados do Apendice A1
> attach(leuc)
> idadec<-ifelse(idade>96,1,0)
> leuinic<-ifelse(leuini>75,1,0)
> zpesoc<-ifelse(zpeso>-2,1,0)
> zestc<-ifelse(zest>-2,1,0)
> pasc<-ifelse(pas>0.05,1,0)
> vacc<-ifelse(vac>15,1,0)
> pasc<-ifelse(pas>5,1,0)
> riskc<-ifelse(risk>1.7,1,0)
> r6c<-r6
> leucc<-as.data.frame(cbind(leuinic,tempos,cens,idadec,zpesoc,zestc,pasc,vacc,riskc,r6c))
> detach(leuc)
> attach(leucc)
> require(survival)
> fit<-coxph(Surv(tempos,cens)~leuinic+idadec+zpesoc+zestc+pasc+vacc+riskc+r6c,
             data=leucc, x = T, method="breslow")
> summary(fit)
```

Uma análise preliminar desta tabela indica que possivelmente as covariáveis RISK, R6 e ZEST não são importantes para explicar o tempo até a recidiva ou

Tabela 5.9: Modelo de Cox para os dados de leucemia com as oito covariáveis.

Covariável	Coeficiente	Erro-Padrão	valor-p
LEUINI	0,979	0,424	0,021
IDADE	0,743	0,375	0,048
ZPESO	$-1,369$	0,788	0,082
ZEST	$-0,811$	0,759	0,256
PAS	$-1,041$	0,496	0,036
VAC	1,316	0,450	0,003
RISK	0,0005	0,476	1,000
R6	$-0,573$	0,521	0,270

morte de crianças com leucemia, na presença das demais. A Tabela 5.10 mostra os valores de menos 2 vezes o logaritmo da máxima verossimilhança parcial ($\mathcal{L}$) para alguns modelos. Para o modelo 3, por exemplo, foi utilizado no R os comandos:

```
> fit3<-coxph(Surv(tempos,cens)~leuinic+idadec+zpesoc+pasc+vacc,data=leucc,x = T,method="breslow")
> summary(fit3)
> -2*fit3$loglik[2]
```

As covariáveis IDADE e leucometria inicial (LEUINI) foram mantidas em todos os modelos pois sabe-se a partir da literatura médica que elas são importantes fatores de prognóstico.

Tabela 5.10: Valores de $\mathcal{L}$ = - 2(log-verossimilhança) obtidos para alguns modelos.

MODELO	$\mathcal{L}$
1- IDADE + LEUINI + ZPESO + ZEST + PAS + VAC + RISK + R6	280,45
2- IDADE + LEUINI + ZPESO + ZEST + PAS + VAC	281,60
3- IDADE + LEUINI + ZPESO + PAS + VAC	282,64
4- IDADE + LEUINI + ZEST + PAS + VAC	285,30
5- IDADE + LEUINI + ZPESO + PAS	291,11
6- IDADE + LEUINI + ZPESO + VAC	291,71
7- IDADE + LEUINI + ZPESO	297,47

O teste da razão de verossimilhança parcial será utilizado para comparar alguns modelos a partir dos valores apresentados na Tabela 5.10. O teste da importância conjunta das covariáveis RISK, R6 e ZEST é feito comparando os modelos 1 e 3 através da estatística da razão de verossimilhança (TRV) parcial

$$TRV = 282,64 - 280,45 = 2,19$$

que, sob a hipótese nula, tem aproximadamente uma distribuição qui-quadrado com 3 graus de liberdade. O que gera um valor-p igual a $0,53$. Este valor mostra que estas covariáveis perdem o seu valor prognóstico na presença das outras covariáveis.

Sabe-se que o peso e a altura das crianças são importantes para explicar a resposta, mas são fortemente associados. A partir do modelo que inclui ambas (modelo 2), pode-se testar a possibilidade de exclusão de cada uma delas na presença das demais (modelos 3 e 4). Os seguintes valores foram obtidos:

$$TRV = 282,64 - 281,60 = 1,04 \quad (p = 0,31) \qquad \text{(excluir ZEST)},$$

$$TRV = 285,30 - 281,60 = 3,70 \quad (p = 0,054) \quad \text{(excluir ZPESO)}.$$

Estes testes praticamente confirmam a afirmação estabelecida acima. Ou seja, na presença de altura, o peso perde sua importância e vice-versa. No entanto, este efeito é muito mais acentuado para a exclusão da altura. O modelo 7 inclui IDADE, LEUINI e ZPESO. Os modelos 6 e 7 são usados para testar a inclusão de VAC (TRV = 5,76, p = 0,016) e os modelos 5 e 7 a inclusão de PAS (TRV = 6,36, p = 0,012). A inclusão de VAC e PAS simultaneamente é testada utilizando os modelos 3 e 7 (TRV = 14,83, p = 0,002). Desta forma o modelo 3 é o final, cujas estimativas estão apresentadas na Tabela 5.11.

Tabela 5.11: Modelo de Cox final para os dados de leucemia pediátrica.

Covariável	Coeficiente	Erro-Padrão	valor-p	razão de riscos
LEUINI	1,109	0,394	0,005	$e^{1,109} = 3,03$
IDADE	0,711	0,371	0,055	$e^{0,711} = 2,04$
ZPESO	$-2,055$	0,496	<0,001	$e^{-2,055} = 0,13$
PAS	$-1,225$	0,456	0,007	$e^{-1,225} = 0,29$
VAC	1,324	0,414	0,001	$e^{1,324} = 3,76$

A Tabela 5.11 mostra que valores mais altos da leucometria inicial, da idade e da porcentagem de vacúolos aumenta o risco de recidiva ou morte entre crianças com leucemia. O inverso acontece com as covariáveis PAS e ZPESO. A interpretação, por exemplo, do coeficiente estimado associado à idade é que o risco de recidiva ou morte entre crianças com mais de 96 meses (8 anos) é cerca de 2 vezes o risco daquelas com menos de 8 anos, mantidas as outras covariáveis fixas.

5.7.3.1 Diagnóstico do Modelo Ajustado

Para verificar a suposição de riscos proporcionais no modelo de Cox ajustado para os dados de leucemia pediátrica, os métodos gráficos descritos na Seção 5.6.1 foram utilizados. A Figura 5.10 mostra as curvas do logaritmo de $\hat{\Lambda}_{0j}(t)$ *versus* os tempos para cada covariável mantida no modelo final ajustado. A partir desta figura, cujos comandos usados no R para sua obtenção encontram-se apresentados no apêndice B, pode-se observar que as curvas não se cruzam para nenhuma das covariáveis e, embora existam alguns desvios quanto ao paralelismo das curvas, em especial para as covariáveis PAS e VAC, não há evidências de que estes desvios possam sugerir uma séria violação da suposição de riscos proporcionais.

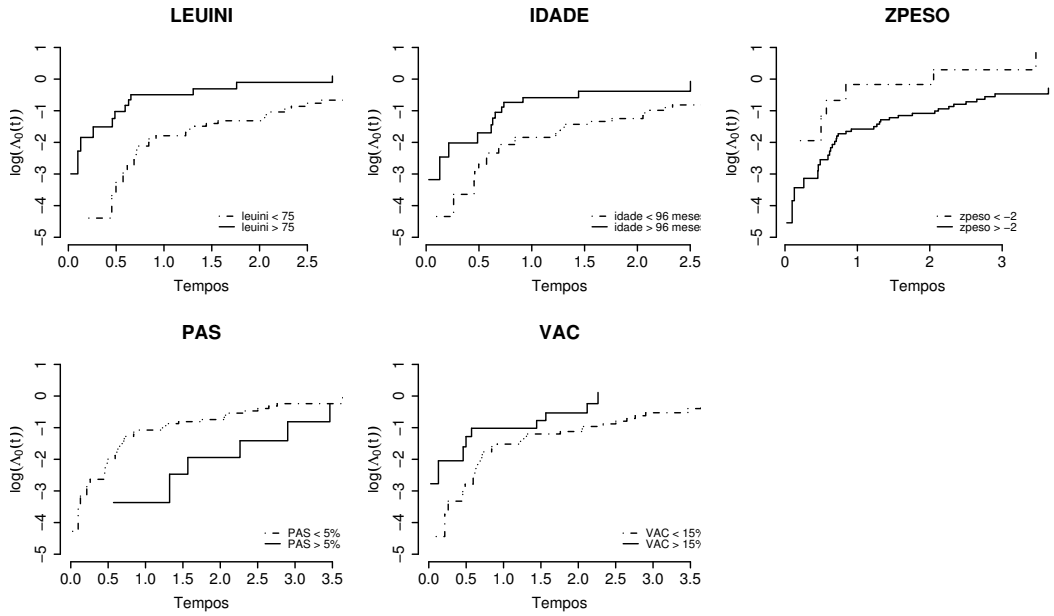

Figura 5.10: Log($\hat{\Lambda}(t)$) *versus* tempos para as covariáveis leuini, idade, zpeso, pas e vac.

A Figura 5.11 obtida no R por:

```
> residuals.coxph(fit3,type="scaledsch")
> cox.zph(fit3)
> par(mfrow=c(2,3))
> plot(cox.zph(fit3))
```

apresenta, ainda, para estas mesmas covariáveis, os resíduos escalonados de Schoenfeld *versus* os tempos. Desta figura, tendências ao longo do tempo, embora não muito acentuadas, podem ser observadas para as covariáveis LEUINI, PAS e VAC. Tais tendências sugerem uma possível violação da suposição de riscos proporcionais

bem como que as covariáveis citadas estariam gerando esta violação. Como visto, contudo, na Figura 5.10, situações extremas dessa violação, que são caracterizadas por curvas que se cruzam, não foram observadas para nenhuma dessas covariáveis. A análise das Figuras 5.10 e 5.11 sugere, desse modo, não haver evidências de séria violação da suposição de riscos proporcionais.

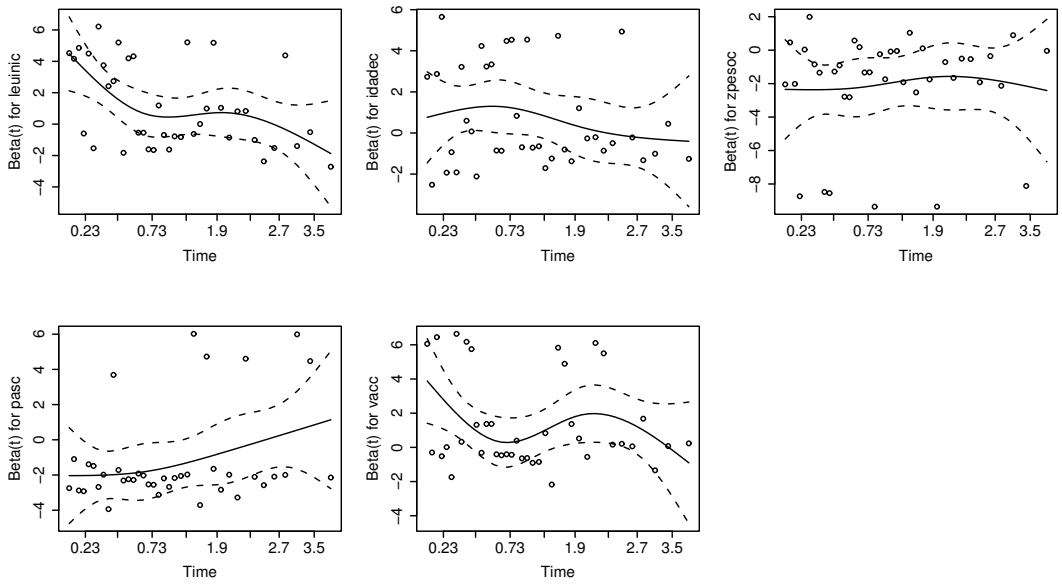

Figura 5.11: Suposição de riscos proporcionais para as covariáveis leuini, idade, zpeso, pas e vac fazendo uso dos resíduos escalonados de Schoenfeld.

A Figura 5.12, que pode ser obtida no *R* por

```
> par(mfrow=c(1,2))
> mart<-residuals.coxph(fit3,type="martingale")
> dev<-residuals.coxph(fit3,type="deviance")
> plot(mart,ylab="res",pch=20)
> title("Resíduos Martingale")
> plot(dev,ylab="res",pch=20)
> title("Resíduos Deviance")
```

apresenta, adicionalmente, os gráficos dos resíduos martingale e deviance do modelo ajustado. Tais gráficos não sugerem a existência de pontos que possam ser considerados atípicos (*outliers*), com uma possível exceção ao resíduo martingale de -3,15. Para este resíduo martingale tem-se, contudo, um correspondente resíduo deviance de -2,51 o qual é um valor aceitável dentro da variação observada para estes resíduos. O comportamento aleatório dos resíduos deviance em torno de zero, observado no gráfico à direita da Figura 5.12, fornece ainda, indicativos favoráveis à adequação do modelo ajustado aos dados deste estudo.

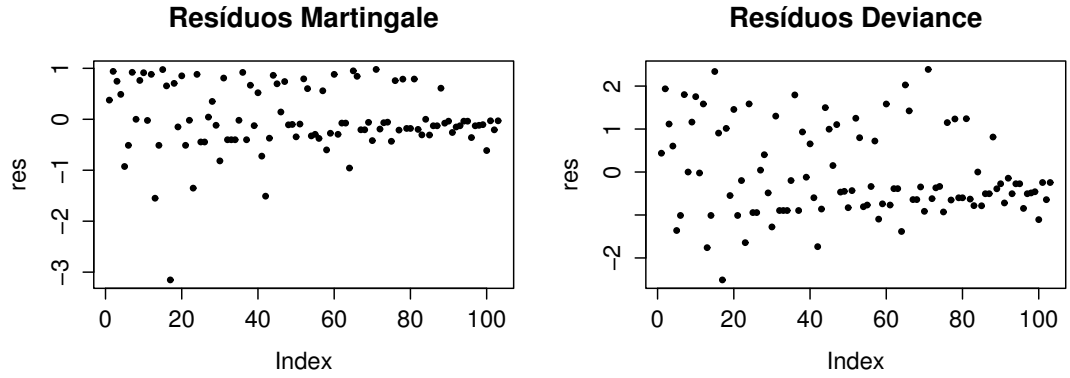

Figura 5.12: Resíduos Martingale e Deviance do modelo de Cox final ajustado.

5.8 Comentários sobre o Modelo de Cox

O modelo de regressão Cox é, como dito anteriormente, extensivamente utilizado em estudos médicos devido essencialmente a presença do componente não-paramétrico, o que o torna bastante flexível. Este modelo apresenta ainda, alguns modelos paramétricos como casos particulares (Kalbfleisch e Prentice, 1980). O modelo de Weibull é por exemplo um desses casos quando se toma $\lambda_0(t) = \frac{\gamma}{\alpha^\gamma}\, t^{\gamma-1}$ na expressão dada em (5.2). Como o modelo exponencial é um caso especial do de Weibull, segue que o mesmo também é um modelo de riscos proporcionais.

Kalbfleish e Prentice (1980) mostraram que o modelo Weibull de parâmetros (γ, α) é o único modelo que pertence tanto a classe de modelos log-lineares quanto a classe de modelos de riscos proporcionais. O modelo exponencial, como já citado, inclui-se nesse resultado por ser um caso especial.

A família de modelos de riscos proporcionais é essencialmente distinta da família de modelos log-lineares apresentada no Capítulo 4. Os modelos Weibull e exponencial são os únicos modelos log-lineares que são também de riscos proporcionais.

5.9 Exercícios

1. Os seguintes dados representam o tempo (em dias) até a morte de pacientes com câncer de ovário tratados na Mayo Clinic (Fleming et al, 1980). O "+"indica censura.

   Amostra 1 (tumor grande): 28, 89, 175, 195, 309, 377+, 393+, 421+, 447+, 462, 709+, 744+, 770+, 1106+, 1206+

Amostra 2 (tumor pequeno): 34, 88, 137, 199, 280, 291, 299+, 300+, 309, 351, 358, 369, 369, 370, 375, 382, 392, 429+, 451, 1119+.

(a) Escreva a forma do Modelo de Cox para estes dados.

(b) Escreva a forma da função de verossimilhança parcial.

(c) Ajuste o modelo de Cox e construa um intervalo de confiança para o parâmetro do modelo.

(d) Teste a hipótese de igualdade dos dois grupos. Caso exista diferença entre os grupos, interprete o coeficiente estimado.

(e) Sabendo que o teste logrank **coincide** com o teste escore associado ao modelo de Cox, use este teste para testar a hipótese estabelecida em (d).

2. Um estudo foi realizado para comparar dois tratamentos pós-cirúrgicos de câncer de ovário. O estudo envolveu o acompanhamento de 26 mulheres após a cirurgia de remoção do tumor. A resposta foi o tempo, em dias, do início do tratamento (aleatorização) até a morte do paciente. As seguintes covariáveis foram registradas: tratamento, idade, resíduo: se o resíduo da doença foi completamente (2) ou parcialmente (1) removido e status: é a condição do doente no início do estudo, boa (1) ou ruim (2). Os dados encontram-se apresentados na Tabela 5.12.

Tabela 5.12: Conjunto de dados referente ao Exercício 2

Paciente	tempo	ind. falha	tratamento	idade	resíduo	status
1	156	1	1	66	2	2
2	1040	0	1	38	2	2
3	59	1	1	72	2	1
4	421	0	2	53	2	1
5	329	1	1	43	2	1
6	769	0	2	59	2	2
7	365	1	2	64	2	1
8	770	0	2	57	2	1
9	1227	0	2	59	1	2
10	268	1	1	74	2	2
11	475	1	2	59	2	2
12	1129	0	2	53	1	1
13	464	1	2	56	2	2
14	1206	0	2	44	2	1
15	638	1	1	56	1	2
16	563	1	2	55	1	2
17	1106	0	1	44	1	1
18	431	1	1	50	2	1
19	855	0	1	43	1	2
20	803	0	1	39	1	1
21	115	1	1	74	2	1
22	744	0	2	50	1	1
23	477	0	1	64	2	1
24	448	0	1	56	1	2
25	353	1	2	63	1	2
26	377	0	2	58	1	1

1. Ajuste o modelo de Cox para estes dados e apresente o seu melhor ajuste.

2. Use uma técnica de adequação de modelo para verificar a suposição de riscos proporcionais.

3. Caso a suposição seja válida, use o modelo ajustado em (a) para verificar se existe diferença entre os tratamentos.

4. Qual a probabilidade de uma paciente com 45 anos, resíduo = 1 e status = 2, sobreviver aos primeiros dois anos após o uso do tratamento 2?

Capítulo 6

Extensões do Modelo de Riscos Proporcionais e o Modelo de Riscos Aditivos de Aalen

6.1 Introdução

Algumas situações práticas envolvendo medidas longitudinais não são ajustadas adequadamente usando o modelo de Cox na sua forma original como apresentado no Capítulo 5. Existem covariáveis que são monitoradas durante o estudo, e seus valores podem mudar ao longo desse período. Por exemplo, pacientes podem mudar de grupo durante o tratamento ou a dose de quimioterapia aplicada em pacientes com câncer pode sofrer alterações durante o curso do tratamento. Se estes valores forem incorporados na análise estatística, resultados mais precisos podem ser obtidos comparados àqueles em que utilizam somente as mesmas medidas registradas no início do estudo. Em outros exemplos, a não inclusão destes valores pode acarretar em sérios vícios. Este tipo de covariável é chamada de dependente do tempo e o modelo de Cox pode ser estendido para incorporar as informações longitudinais registradas para esta variável.

Em outras situações a suposição de riscos proporcionais é violada e o modelo de Cox não é adequado. Modelos alternativos existem para enfrentar esta situação. Um deles é uma extensão do próprio modelo de Cox chamado de modelo de riscos proporcionais estratificado. Neste caso supõe-se que os riscos proporcionais valem em cada estrato mas não valem entre estratos. Um outro modelo alternativo é o aditivo de Aalen. Neste caso, o efeito das covariáveis é aditivo na função de risco ao

invés de multiplicativo. Este tipo de modelagem gera vantagens e desvantagens em situações reais. A grande vantagem do modelo de riscos aditivos de Aalen é a de possibilitar o monitoramento do efeito da covariável ao longo do acompanhamento enquanto que a desvantagem é a de ser permitido valores estimados negativos para a função de riscos.

O objetivo deste capítulo é apresentar generalizações do modelo de Cox, úteis em situações práticas assim como o modelo aditivo de Aalen. A modelagem envolvendo covariáveis dependentes do tempo é apresentada na Seção 6.2. A seguir as outras duas generalizações do modelo de Cox são apresentadas: o modelo estratificado na Seção 6.3 e o modelo aditivo de Aalen na Seção 6.4. O capítulo finaliza na Seção 6.5 aplicando estes modelos em uma situação real envolvendo a ocorrência de sinusite em pacientes infectados com o HIV. Esta aplicação foi apresentada na Seção 1.2.

6.2 Covariáveis Dependentes do Tempo

As covariáveis no modelo de Cox consideradas no Capítulo 5, foram medidas no início do estudo ou na origem do tempo. Entretanto, existem covariáveis que são monitoradas durante o estudo e seus valores podem mudar ao longo do período de acompanhamento. Um estudo bastante analisado na literatura é o do programa de transplante de coração de Stanford (Crowley e Hu, 1977). Neste estudo os pacientes eram aceitos no programa quando se tornavam candidatos a um transplante de coração. Quando surgia um doador, os médicos escolhiam, de acordo com alguns critérios, o candidato que iria receber o coração. Alguns pacientes morreram sem receber o transplante. A forma de alocação estava fortemente viciada na direção daqueles pacientes com maior tempo de sobrevivência pois somente estes pacientes viveram o suficiente para receber o coração. O uso de uma covariável assumindo o valor zero para aqueles esperando o transplante e um para aqueles com coração novo, serve para minimizar esse vício e ela muda de valor assim que o transplante é realizado e é, portanto, dependente do tempo.

O estudo da ocorrência de sinusite em pacientes infectados pelo HIV que foi apresentado na Seção 1.2 é outro exemplo com uma covariável dependente do tempo. A classificação do paciente (soropositivo assintomático, ARC e AIDS) pode mudar ao longo do estudo. Ou seja, alguns pacientes que iniciaram o estudo com a classificação soropositivo assintomático evoluiram para AIDS no final do estudo passando por ARC. Este estudo será analisado na Seção 6.4 utilizando o modelo de Cox com covariáveis dependentes do tempo apresentado a seguir.

As covariáveis que alteram seu valor ao longo do período de acompanhamento são conhecidas como covariáveis dependentes do tempo e podem ser incorporadas ao modelo de regressão de Cox, generalizando-o como

$$\lambda(t) = \lambda_0(t) \exp\left\{\mathbf{x}'(t)\boldsymbol{\beta}\right\}. \tag{6.1}$$

Definido desta forma, o modelo (6.1) não é mais de riscos proporcionais pois a razão das funções de risco no tempo t para dois indivíduos i e j fica sendo

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\left\{\mathbf{x}_i'(t)\boldsymbol{\beta} - \mathbf{x}_j'(t)\boldsymbol{\beta}\right\}$$

que é dependente do tempo. A interpretação dos coeficientes $\boldsymbol{\beta}$ do modelo deve considerar o tempo t. Cada coeficiente β_l, $l = 1, \ldots, p$, pode ser interpretado como o logaritmo da razão de riscos cujo valor da l-ésima covariável no tempo t difere de uma unidade, quando as outras covariáveis assumem o mesmo valor neste tempo.

O ajuste do modelo de Cox é obtido estendendo a função de log-verossimilhança parcial. Isto é feito usando

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i(t_i) - \frac{\sum_{j \in R(t_i)} x_j(t_i) \exp\left\{\mathbf{x}_j'(t_i)\widehat{\boldsymbol{\beta}}\right\}}{\sum_{j \in R(t_i)} \exp\left\{\mathbf{x}_j'(t_i)\widehat{\boldsymbol{\beta}}\right\}} \right] = 0$$

que é uma extensão da expressão (5.6) considerando covariáveis dependentes do tempo. Propriedades assintóticas dos estimadores de máxima verossimilhança parcial, para que se possa construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo, foram obtidas por Andersen e Gill (1982). Eles apresentaram provas bastante gerais das propriedades para o modelo de Cox incluindo covariáveis dependentes do tempo. Eles usaram a relação entre os tempos de falhas e martingais, como foi mencionado no Capítulo 5, para mostrar que estes estimadores são consistentes e assintoticamente normais sob certas condições de regularidade. Desta forma, pode-se usar as conhecidas estatísticas de Wald e da razão de verossimilhança para fazer inferências no modelo de regressão de Cox.

6.3 O Modelo de Cox Estratificado

Na Seção 5.6 foram apresentadas técnicas estatísticas para avaliar a adequação do modelo de Cox. Essencialmente, estas técnicas avaliam a suposição de riscos proporcionais. O modelo (5.2) não pode ser usado se esta suposição for violada. Nestes casos, uma solução para o problema é estratificar os dados de modo que a suposição

seja válida em cada estrato. Por exemplo, os riscos podem não ser proporcionais entre homens e mulheres mas esta suposição pode valer no estrato formado somente por homens e naquele formado somente por mulheres.

A análise estratificada consiste em dividir os dados de sobrevivência em m estratos, de acordo com uma indicação de violação da suposição. O modelo de riscos proporcionais (5.2) é então expresso como

$$\lambda_{ij}(t) = \lambda_{0_i}(t) \exp\left\{\boldsymbol{\beta}' X_{ij}\right\} \tag{6.2}$$

para $i = 1, \ldots, m$ e $j = 1, \ldots, n_i$, em que n_i é o número de observações no i-ésimo estrato. As funções de base λ_{0_1},...,λ_{0_m}, são arbitrárias e completamente não relacionadas.

A estratificação não cria nenhuma complicação na estimação do vetor $\boldsymbol{\beta}$. Uma verossimilhança parcial (5.5) é construída para cada estrato e a estimação dos β's é baseada na soma dos logaritmos das verossimilhanças parciais (Kalbfleisch e Prentice, 1980, p.87-88).

As propriedades assintóticas destes estimadores são obtidas a partir dos estimadores do modelo não estratificado (Colosimo, 1996). O modelo estratificado deve somente ser usado caso realmente necessário, ou seja, na presença de violação da suposição de riscos proporcionais. O uso desnecessário da estratificação acarreta em uma perda de eficiência das estimativas obtidas. Informações adicionais sobre o modelo estratificado de Cox podem ser encontradas em Colosimo (1991).

6.4 Modelo Aditivo de Aalen

O modelo de riscos proporcionais de Cox apresenta as vantagens de ter uma simples interpretação dos resultados, ser facilmente estendido para incorporar covariáveis dependentes do tempo e de estar disponível em vários pacotes estatísticos. Entretanto Aalen (1989) citou algumas limitações deste modelo. A primeira delas é que as suposições do modelo podem não valer, as vezes o modelo de Cox é usado na literatura sem que suas propriedades sejam verificadas. Isto ocorre com freqüência na literatura médica. Além disto, também não é claro se satisfazendo as propriedades usuais de proporcionalidade garantem a adequação do modelo de Cox. Em segundo lugar, mudanças ao longo do tempo na influência das covariáveis não são facilmente descobertas e o modelo de Cox não é adaptado para uma descrição detalhada de efeitos de covariáveis ao longo do tempo. Por último, a suposição de proporciona-

lidade do risco é vulnerável à mudanças no número de covariáveis modeladas. Se as covariáveis são retiradas de um modelo ou medidas com um diferente nível de precisão, a proporcionalidade é geralmente afetada. Portanto verifica-se uma falta de consistência do modelo de Cox a este respeito.

Estas limitações conduziram a propostas de modelos alternativos ao de Cox para modelar a função de risco. Uma alternativa foi sugerida originalmente por Aalen (1980) que é um modelo de risco aditivo para análise de regressão de dados censurados. Este modelo aditivo de Aalen fornece uma alternativa útil ao modelo de riscos proporcionais de Cox pois permite que ambos, os parâmetros e os vetores de covariáveis, variem com o tempo. Já que efeitos temporais não são assumidos serem proporcionais para cada covariável, o modelo de Aalen é capaz de fornecer informações detalhadas a respeito da influência temporal de cada covariável. Os modelos de Cox e Aalen diferem fundamentalmente, o de Cox tem uma função básica não-paramétrica mas o efeito das covariáveis é modelado parametricamente. Por outro lado, o modelo de Aalen é completamente não-paramétrico no sentido de que funções são ajustadas e não parâmetros. Ou seja, na estimação dos parâmetros o modelo de Aalen usa apenas informação local o que faz este modelo bastante flexível. Os estimadores propostos por Aalen generalizam o tão conhecido estimador de Nelson-Aalen que é o estimador natural no caso de população homogênea. Aplicações foram apresentadas por Mau (1986, 1988) e Andersen e Vaeth (1989) e resultados teóricos foram realizados por McKeague (1986), McKeague e Utikal (1988) e Huffer e McKeague (1987) indicando que o modelo pode ser útil e é sem dúvida razoável para explorar vantagens da linearidade analogamente a teoria clássica de modelo linear.

Em um estudo típico, um número de indivíduos são observados ao longo do tempo para verificar a ocorrência de um determinado evento. O acontecimento deste evento é assumido independente entre os indivíduos. Como no modelo de risco multiplicativo, tem-se um tempo até a ocorrência do evento T_i para cada indivíduo, cuja distribuição depende de um vetor dado por $\mathbf{x}_i(t) = (1, x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))'$ em que $x_{ij}(t)$, com $j = 1, \ldots, p$, são os valores das covariáveis que podem variar no tempo. Seja n o número de indivíduos, p o número de covariáveis na análise e $\lambda_i(t)$ a função de risco para o tempo de sobrevivência t_i de um indivíduo i.

O modelo de risco aditivo de Aalen é dado por

$$\lambda_i(t) = \alpha_0(t) + \sum_{j=1}^{p} \alpha_j(t) x_{ij}(t).$$

Considerando a forma matricial

$$\lambda(t) = \alpha(t)Y(t),$$

em que $\alpha(t) = (\alpha_0(t), \alpha_1(t), \ldots, \alpha_p(t))'$ é um vetor de funções do tempo desconhecidas, cujo primeiro elemento $\alpha_0(t)$ é interpretado como uma função de parâmetro básica, enquanto que $\alpha_i(t)$, $i = 1, \ldots, p$, chamados aqui funções de regressão medem a influência das respectivas covariáveis.

O modelo aditivo de Aalen pode ser obtido a partir de uma expansão em Taylor do modelo de Cox ou de uma forma mais geral da função de risco. Ou seja, expandindo em série de Taylor em torno de $X = 0$ e ignorando os termos superiores ao de primeira ordem.

A matriz $Y(t)$ de ordem $n \times (p+1)$ é construída da seguinte maneira: se o evento considerado ainda não ocorreu para o i-ésimo indivíduo e ele não é censurado então a i-ésima linha de $Y(t)$ é o vetor $\mathbf{x}_i(t) = (1, x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))'$. Caso contrário, se o indivíduo não está sob risco no tempo t, então a linha correspondente de $Y(t)$ contém apenas zeros.

Este modelo é considerado não-paramétrico pois nenhuma forma paramétrica particular é assumida para as funções de regressão. Como visto, estas funções podem variar arbitrariamente com o tempo, revelando mudanças na influência das covariáveis. Esta é uma das vantagens do modelo acima bem como a não exigência de tamanho de amostra extremamente grande.

O modelo de riscos proporcionais assume que os efeitos das covariáveis agem multiplicativamente na função de risco. Os coeficientes estimados da estrutura de regressão são constantes desconhecidas cujos valores não mudam com o tempo. No modelo de Aalen assume-se que as covariáveis agem de maneira aditiva na função de risco e os coeficientes de riscos desconhecidos podem ser funções do tempo, ou seja, o efeito das covariáveis pode variar durante o estudo. Dessa forma os estimadores dos parâmetros são baseados nas técnicas de mínimos quadrados. A derivação desses estimadores é similar a derivação do estimador de Nelson-Aalen da função de risco acumulada apresentado na Seção 2.4.1.

A aproximação para estimação depende das suposições sobre a forma funcional das funções de regressão que neste caso são não-paramétricas. A estimação direta das funções de regressão é difícil na prática sendo mais fácil a estimação da função de regressão acumulada. Isto ocorre pelo mesmo motivo que é mais fácil estimar a função de distribuição acumulada do que a função de densidade de probabilidade. Considera-se então a estimação do vetor coluna $A(t)$ com elementos $A_j(t)$ dados por

$$A_j(t) = \int_0^t \alpha_j(s)ds$$

Sejam $T_1 < T_2 < \ldots$ os tempos de falhas ordenados. Aalen considerou um estimador razoável de $A(t)$, denominado estimador de mínimos quadrados de Aalen, que é dado por.

$$A^*(t) = \sum_{T_k \leq t} Z(T_k) I_k, \tag{6.3}$$

em que I_k é um vetor de zeros que assume o valor 1 para o indivíduo cujo evento ocorre no tempo T_k. Enquanto que $Z(t)$ é a inversa generalizada de $Y(t)$. Em princípio, $Z(t)$ pode ser qualquer inversa generalizada de $Y(t)$. Uma escolha sinples pode ser baseada no princípio de mínimos quadarados local, ou seja

$$Z(t) = [Y(t)'Y(t)]^{-1}Y(t)'.$$

Esta inversa usada comumente em modelos de regressão, em geral, pode não ser ótima. Uma escolha ótima dependerá do conhecimento dos verdadeiros valores dos parâmetros. Huffer e McKeague (1987) sugeriram o uso de uma outra inversa definindo assim o estimador de mínimos quadrados ponderados. Neste trabalho será usada a inversa de mínimos quadrados.

É importante notar que o estimador de $A(t)$ é definido apenas sobre um intervalo de tempo onde $Y(t)$ tem posto completo, ou seja, a estimação pára quando $Y(t)$ perde o posto completo, que é uma consequência do princípio não paramétrico. Os componentes de $A^*(t)$ convergem assintoticamente, sob condições apropriadas, para um processo gaussiano (Aalen 1989). Então um estimador da matriz de covariância de $A^*(t)$ é dado por

$$\Omega^*(t) = \sum_{T_k \leq t} Z(T_k) I_k^D Z(T_k)', \qquad i = 1, \ldots, k$$

em que I_k^D é uma matriz diagonal com I_k como diagonal.

As funções de regressão acumuladas são obtidas em cada tempo de falha distinto pela estimação da contribuição instantânea das covariáveis para o risco. $A_j^*(t)$ pode ser considerada como uma função empírica descrevendo a influência da j-ésima covariável. A inclinação do gráfico da função de regressão acumulada contra o tempo fornece informação sobre a influência de cada covariável, sendo possível verificar

se uma covariável particular tem um efeito constante ou varia com o tempo ao longo do período de estudo. Por exemplo, se $\alpha_j(t)$ é constante, então o gráfico deve aproximar-se de uma linha reta. Inclinações positivas ocorrem durante períodos em que aumentos dos valores das covariáveis são associados com aumentos na função de risco. Por outro lado, inclinações negativas ocorrem em períodos quando crescimentos nos valores das covariáveis estão associados com decréscimos na função de risco. As funções de regressão acumuladas têm inclinações aproximadamente iguais a zero em períodos em que as covariáveis não influenciam a função risco. Ramlau-Hansen (1983) mostrou que também é possível estimar a função mais diretamente utilizando métodos de estimação da densidade de probabilidade.

Não é difícil verificar, como consequência dos resultados obtidos anteriormente, que pode-se estimar o risco acumulado e a função de sobrevivência correspondentes dados os valores das covariáveis. Seja $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)'$ o conjunto de valores das covariáveis fixados no tempo zero. O estimador do risco acumulado $H^*(t)$ é dado por

$$H^*(t) = A^*(t)'X.$$

A partir da relação apresentada no Capítulo 2 entre a função de sobrevivência e a função de risco acumulada, esta é estimada então por

$$S^*(t) = \exp(-H^*(t)). \tag{6.4}$$

Alternativamente, baseada no estimador de Kaplan-Meier, a função de sobrevivência pode ser estimada como

$$S^{**}(t) = \prod_{T_k \leq t} [1 - (Z(T_k)I_k)'x].$$

A função de sobrevivência estimada não é necessariamente monótona sobre todo o período de observação. Ela pode aumentar para alguns valores de t e de acordo com a equação (6.4) decrescer para algum t.

É freqüentemente de interesse testar se uma covariável específica tem algum efeito na função de risco total. Para o modelo aditivo de Aalen isto corresponde a testar a hipótese nula de que não existe efeito da covariável sob a função de risco. A hipótese nula para algum $j \geq 1$ é estabelecida como

$$H_j : \alpha_j(t) = 0, \qquad t \in [0, T]$$

É importante lembrar que no contexto não-paramétrico a hipótese nula acima pode apenas ser testada sobre intervalos de tempo onde $Y(t)$ tem posto completo. Dentro da estrutura do modelo, Aalen (1980, 1989) desenvolveu para todo tempo de falha uma estatística de teste para H_j dada pelo j−ésimo elemento U_j do vetor

$$U = \sum_{T_k} K(T_k)Z(T_k)I_k, \tag{6.5}$$

em que $K(t)$, uma função peso não negativa, é uma matriz diagonal $(p+1)\times(p+1)$.

A estatística de teste da Equação (6.5) surge como uma combinação ponderada da soma do estimador de $A_j(t)$ apresentado na equação (6.3). Os elementos diagonais de $K(t)$ são funções pesos e suas escolhas podem depender das alternativas para a hipótese nula de interesse.

Uma escolha ótima da função peso necessitará do conhecimento das verdadeiras variâncias dos estimadores, entretanto isto dependerá de funções de parâmetros desconhecidas. Aalen considerou duas escolhas para a função peso. A primeira possibilidade é considerar cada função peso igual ao número de pacientes que permanecem no conjunto de risco em algum tempo dado. Neste caso a matriz $K(t)$ é substituída por um escalar $K_1(T_k)$ dado por

$$K_1(T_k) = \sum_{i=1}^{n} K_{1i}(t),$$

em que $K_{1i} = 1$, se o i-ésimo indivíduo está sob risco no tempo t e $K_{1i} = 0$ em caso contrário.

Uma segunda escolha é tomar $K_2(t) = \{\text{diag}[(\text{Y(t)}'\text{Y(t)})^{-1}]\}^{-1}$, em que $K_2(t)$ é dada como a inversa de uma matriz diagonal tendo a mesma diagonal principal da matriz $(Y(t)'Y(t))^{-1}$. Este peso é escolhido por analogia ao problema da regressão de mínimos quadrados em que as variâncias dos estimadores são proporcionais aos elementos diagonais da matriz $(Y'Y)^{-1}$ sendo Y o desenho da matriz. Estudos preliminares parecem indicar que a escolha da segunda opção pode ser mais poderosa em algumas situações. Neste trabalho foi utilizada esta última opção como função peso.

Um estimador da matriz de covariância de U dado pela Equação (6.5) é

$$V = \sum_{T_k} K(T_k)Z(T_k)I_k^D Z(T_k)'K(T_k)'.$$

Suponha que se queira testar simultâneamente todos H_j para j em algum subconjunto A de $\{1, \ldots, p\}$ consistindo de s elementos. Seja U_A definido como o subvetor correspondente de U e V_A a submatriz correspondente de V, isto é, V_A é a matriz de covariâncias estimadas de U_A. A estatística de teste normalizada $U_A' V_A^{-1} U_A$ é assintoticamente distribuída como uma qui-quadrado com s graus de liberdade quando H_j vale para todo j em A. Se o interesse é testar apenas uma das hipóteses H_j, então é usada a estatística de teste $U_j V_{jj}^{-1/2}$. Esta estatística tem uma distribuição assintótica normal padrão sob a hipótese nula.

Através da escolha de diferentes pesos, Lee e Weissfeld (1998) obtiveram quatro novas estatísticas de testes para o modelo de riscos aditivos. A primeira função peso contém $K_1(t)$ como caso especial e é dada por uma função quadrada, contínua e integrável em $[0, 1]$. A segunda função peso derivada é uma combinação da primeira função peso proposta e de $K_2(t)$ e a terceira é baseada na estimativa de Kaplan-Meier. Por último a quarta função peso proposta combina esta última função e a função peso $K_2(t)$. Estas estatísticas foram comparadas com as duas estatísticas propostas por Aalen usando simulação de Monte Carlo.

6.5 Análise dos dados de pacientes infectados pelo HIV

6.5.1 Descrição dos dados

Este estudo foi brevemente descrito na Seção 1.2. Nesta seção é apresentado mais informações sobre o estudo e o mesmo é analisado utilizando o modelo de regressão de Cox para covariáveis dependentes do tempo (6.1) e o de riscos aditivos de Aalen.

Neste estudo foram utilizadas informações provenientes de 91 pacientes HIV positivo e 21 HIV negativo, somando assim 112 pacientes estudados. Estes pacientes foram acompanhados no período entre março de 1993 a fevereiro de 1995. Somente foram considerados os pacientes que tiveram entrada até julho de 1994. Todos os pacientes incluídos no estudo foram encaminhados ao Centro de Treinamento e Referência em Doenças Infecto-parasitárias (CTR-DIP) da cidade de Belo Horizonte-MG, por pertencerem a grupos de comportamento de risco para adquirir o HIV ou por terem um exame HIV positivo. Após a primeira consulta clínica, os pacientes foram encaminhados ao Serviço de Otorrinolaringologia da Universidade Federal de Minas Gerais.

As doenças otorrinolaringológicas (ORL) avaliadas foram definidas com base nos estudos de prevalência destas manifestações na literatura em pacientes infectados pelo HIV. Nesta seção encontram-se apresentados os resultados para a infecção *sinusite*. A classificação do paciente quanto à infecção pelo HIV seguiu os critérios do CDC (*Centers of Disease Control*, 1989). Os pacientes foram classificados como: HIV soronegativo, HIV soropositivo assintomático, com ARC (*AIDS Related Complex*) e com AIDS. Na covariável Grupo de Risco pacientes HIV soronegativo são aqueles que não possuem o HIV. Pacientes HIV soropositivo assintomáticos são aqueles que possuem o vírus mas não desenvolveram o quadro clínico de AIDS e que apresentam um perfil imunológico estável. Pacientes com ARC são aqueles que apresentam baixa imunidade e outros indicadores clínicos que antecedem o quadro clínico de AIDS. Pacientes com AIDS são aqueles que já desenvolveram infecções oportunistas que definem esta doença, segundo os critérios do CDC de 1989. Esta covariável depende do tempo pois os pacientes mudam de classificação ao longo do estudo. Outras covariáveis neste estudo, como contagem de CD4, também são dependentes do tempo. No entanto, elas somente foram medidas no início do estudo.

A cada consulta, a classificação do paciente foi reavaliada. Cada paciente foi acompanhado através de consultas trimestrais. A frequência mediana foi de 4 consultas. A resposta de interesse foi o tempo, contado a partir da primeira consulta, até a ocorrência das manifestações ORL. O objetivo foi identificar fatores de risco para cada uma destas manifestações. Os possíveis fatores de risco foram listados na Tabela 1.1 e as covariáveis importantes que foram identificadas após utilização das técnicas descritas no Capítulo 2 estão repetidas na Tabela 6.1.

Para a covariável CD4 foram registrados 41 valores perdidos, assim como nas covariáveis Atividade Sexual e Uso de Cocaína em que também foram registrados 23 valores perdidos.

6.5.2 Modelagem Estatística

Os resultados do ajuste do modelo de Cox incluíndo a covariável Grupo de Risco, que depende do tempo, são apresentados na Tabela 6.2. Esta tabela também apresenta as estimativas para outras covariáveis listadas na Tabela 6.1. Pode-se observar que com exceção da covariável grupo que é dependente do tempo, as demais parecem ser não significativas. Removendo-se estas covariáveis gradativamente chegou-se no modelo final. A Tabela 6.3 apresenta as estimativas do modelo final para a ocorrência de sinusite. A idade aparece neste modelo apesar de ser não significativa no modelo apresentado na Tabela 6.2. Idade e CD4 apresentaram-se altamente associadas e ao

Tabela 6.1: Covariáveis medidas no estudo de ocorrência de sinusite.

Idade do Paciente	medida em anos
Sexo do Paciente	1 - Masculino
	2 - Feminino
Grupos de Risco	1 - Paciente HIV Soronegativo
	2 - Paciente HIV Soropositivo Assintomático
	3 - Paciente com ARC
	4 - Paciente com AIDS
CD4	Contagem de CD4
CD8	Contagem de CD8
Atividade Sexual	1 - Homossexual
	2 - Bissexual
	3 - Heterossexual
Uso de Droga	1 - Sim
Injetável	2 - Não
Uso de Cocaína	1 - Sim
por Aspiração	2 - Não

ser retirado CD4 do modelo, idade passou a ser significativa.

Pode-se observar que Idade e Grupos de Risco foram identificados como fatores de risco para a ocorrência desta infecção. Foi verificado que a cada aumento de 10 anos na idade do paciente, o risco de desenvolver Sinusite diminui em $\exp(0,77) = 2,16$ vezes, o que indica que pacientes mais jovens estão mais sujeitos a esta infecção. Notou-se também que o risco de pacientes HIV soropositivo assintomáticos não difere significativamente do grupo HIV soronegativo. Entretanto, no grupo com ARC o risco de se desenvolver Sinusite é 9,7 vezes o risco do grupo HIV soronegativo. Para o grupo com AIDS o risco de desenvolver sinusite é 14,2 vezes o risco em relação ao grupo HIV soronegativo. Por outro lado, a precisão das estimativas associadas a estes dois últimos riscos relativos é bastante reduzida como pode ser observado pela grande amplitude de seus respectivos intervalos de confiança.

Usando o modelo de regressão de Cox foi possível incluir esta covariável dependente do tempo na análise dos dados. Os resultados obtidos a partir da análise estatística deste estudo são importantes para explicar a incidência de manifestações ORL em pacientes HIV positivos. A análise feita nesta seção é somente parte do estudo. Mais informações sobre este estudo e a interpretação clínica dos achados na

Tabela 6.2: Estimativas do modelo de Cox para as covariáveis listadas na Tabela 6.1.

Covariável	Coeficiente Estimado	Valor-p
Idade	-0,025	0,404
Sexo	-0,492	0,413
HIV soropososito assintomático	-1,80	0,002
com ARC	0,270	0,750
com AIDS	2,564	<0,001
Atividade Bissexual	0,558	0,468
Heterossexual	-0,1347	0,825
Aspira	1,045	0,577
CD4	-0,0027	0,085

Tabela 6.3: Estimativas do modelo de Cox final ajustado para os dados de sinusite.

Covariável	Coeficiente de Regressão	Erro Padrão	Valor-p	Risco Relativo Estimado(I.C. 95%)
Idade	-0,077	0,031	0,014	0,926 (0,871; 0,984)
HIV soropos. assint.	-0,755	1,000	0,451	0,470 (0,066; 3,338)
com ARC	2,274	0,837	0,007	9,717 (1,884; 50,124)
com AIDS	2,651	0,790	<0,001	14,168 (3,012; 66,646)

análise dos dados podem ser encontradas em Gonçalves (1995).

Os resultados do ajuste do modelo aditivo de Aalen para os dados de sinusite encontram-se apresentados na Tabela 6.4. Nota-se que a variável "sexo"não é significativa, o que significa que ela será retirada do modelo.

Tal como no modelo de Cox, as covariáveis idade do paciente e os grupos de riscos foram consideradas como fatores influentes na ocorrência da sinusite (ver Tabela 6.4). Assim como no modelo de Cox a covariável "x_2" que indica o grupo HIV soropositivo assintomático permaneceu no modelo por representar um dos grupos de classificação quanto a infecção pelo HIV. O risco de desenvolver sinusite em pacientes HIV soropositivo assintomáticos não diferiu significativamente (valor $p = 0,498$) do grupo HIV soronegativo. Pacientes que fazem parte do grupo com AIDS têm um risco maior de desenvolver a sinusite do que os pacientes dos demais grupos de classificação. Comparando-se com o grupo HIV soronegativo este risco é de aproximadamente 1,6 vezes. Verifica-se também, por exemplo, que um aumento

Tabela 6.4: Resultados do Ajuste inicial do Modelo Aditivo de Aalen para os dados de Sinusite em Pacientes com Aids.

Covariável	Coeficiente	Erro Padrão	p-valor	I.C. (95%)
constante	1,051	0,415	0,005	(0,238; 1,865)
idade	$-0,033$	0,015	0,013	(−0,062;−0,004)
sexo	0,063	0,195	0,889	$(-0,319; 0,445)$
HIV assintomático	0,004	0,140	0,463	$(-0,270; 0,278)$
ARC	0,917	0,371	0,020	(0,189; 1,644)
AIDS	1,566	0,545	0,000	(0,497; 2,635)

de 20 anos na idade do paciente diminue em $0,66$ vezes o risco de ocorrência da sinusite, o que confirma que quanto maior a idade do paciente menor o risco de desenvolvimento desta doença.

Tabela 6.5: Resultados do Ajuste Final do Modelo Aditivo de Aalen para os dados de Sinusite em Pacientes com Aids.

Covariável	Coeficiente	Erro Padrão	p-valor	I.C (95%)
constante	1,038	0,405	0,004	(0,244;1,831)
idade	$-0,031$	0,013	0,011	$(-0,057;-0,005)$
HIV assintomático	0,004	0,136	0,498	$(-0,263;0,271)$
ARC	0,833	0.336	0,020	(0,175;1,491)
AIDS	1,544	0.536	0,000	(0,493;2,595)

Através da análise gráfica das funções de regressão acumuladas contra o tempo, apresentadas na Figura 6.1, pode-se observar o comportamento do efeito de cada covariável significativa no modelo de Aalen. A função de regressão acumulada para a idade tem uma inclinação consistentemente negativa e seu efeito no risco da ocorrência da sinusite diminue razoalvemente com o tempo. Isto indica que crescimentos nos valores da idade, neste período, estão associados com decréscimos na função de risco. A covariável que indica o grupo com ARC parece ter uma influência clara e crescente por cerca de 10 meses com uma influência menor que parece desaparecer depois desse período.

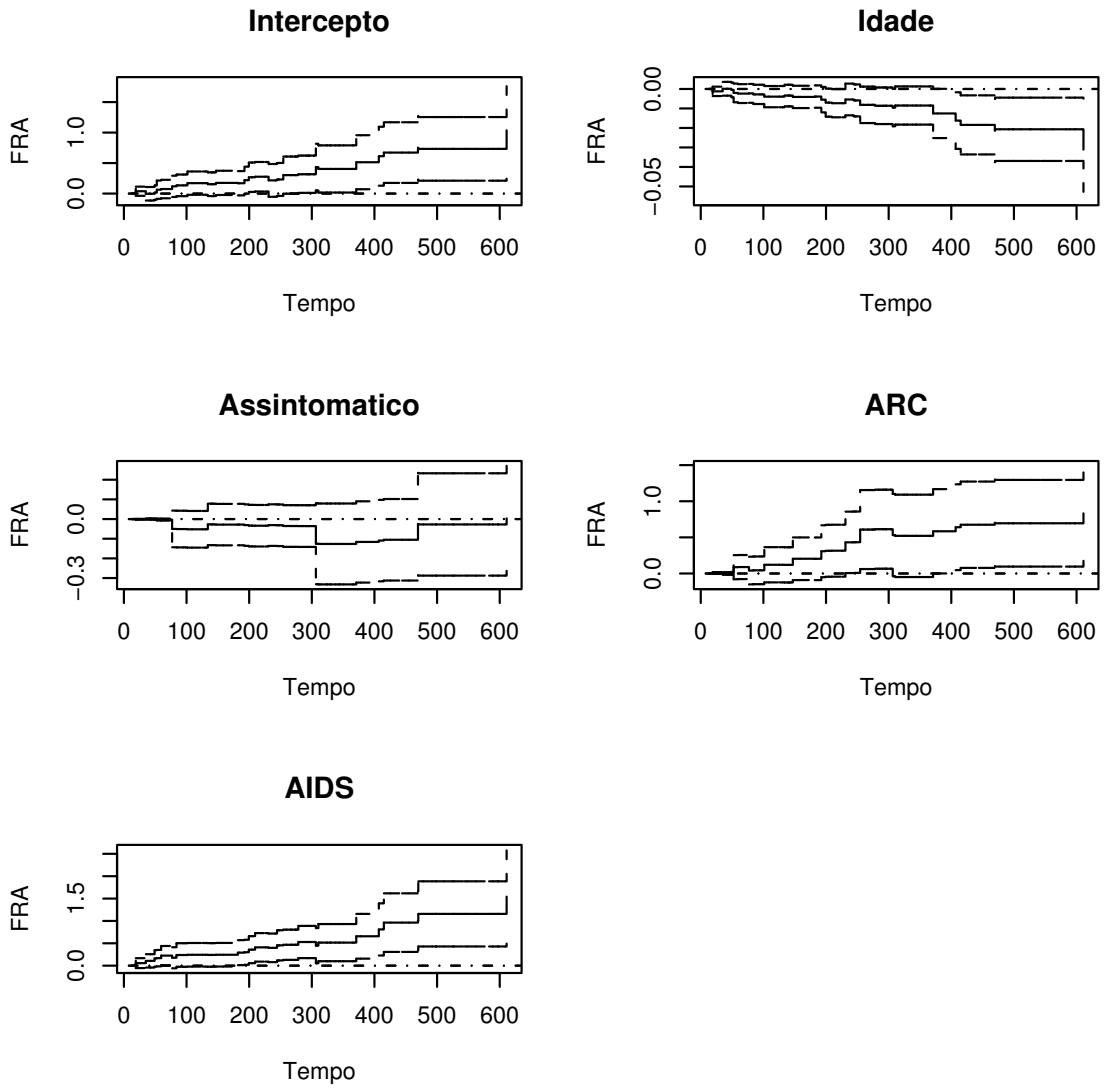

Figura 6.1: Estimativas das Funções de Regressão Acumuladas (FRA) com Intervalo de 95% de Confiança para os dados de Sinusite em Pacientes infectados pelo HIV.

APÊNDICE A

Dados Utilizados no Texto

A.1 Dados de Leucemia Pediátrica
A.2 Dados de Sinusite em Pacientes Infectados pelo HIV
A.3 Dados de Aleitamento Materno
A.4 Dados Experimentais Utilizando Camundongos
A.5 Dados de Câncer de Mama
A.6 Dados de Tempo de Vida de Mangueiras
A.7 Dados de Câncer de Laringe

A.1 Conjunto de dados utilizados no estudo sobre leucemia pediátrica.

leuini	tempos	cens	idade	zpeso	zest	pas	vac	risk	r6
380	1.76	1	60.52	-0.97	-0.48	0.1	5.7	1.58	1
328	0.26	1	68.04	0.36	1.44	0.6	1.5	1.64	0
84.7	0.129	1	159.93	-1.84	-2.17	0.6	20.4	1.26	1
2.9	3.639	1	92.91	-1.06	-0.69	0.7	1.5	0.96	1
400	4.331	0	156.98	-0.84	-0.82	13.7	1	1.32	1
64	4.252	0	69.62	-0.2	-0.19	2.3	2	1.4	1
13.2	0.687	1	79.08	0.02	-2.29	0.3	2	1.52	1
50	0.003	0	112.43	-1.86	-2.42	0	2.7	1.72	1
34.9	2.07	1	47.97	0.15	0.49	4.5	0	1.8	1
68.3	0.709	1	37.91	-0.21	1.27	0	2.1	1.2	1
1	3.466	1	95.21	-2.08	0	43.7	6	0.54	1
24	0.616	1	146.37	-0.49	0.13	0.1	0	1.24	1
140	3.896	0	56.77	-0.07	-1.8	1.2	1.6	1.79	1
5	3.83	0	32.33	-0.32	0.23	0.7	15	0.85	1
49	0.454	1	29.57	0.27	1.11	0	0	2.26	1
68	2.65	1	79.74	-0.66	-1.15	0.1	1	1.3	1
176	3.915	0	160.13	-0.33	-0.98	0.7	4.5	1.06	1
1.6	2.333	1	65.25	2.78	-0.47	0	5	0.6	1
44.6	3.754	0	57.79	0.43	0.19	6.2	13	1.6	1
23.3	1.27	1	69.88	0.57	-0.74	0.7	6.8	1.4	1
6.4	3.704	0	41.59	1.78	1.04	0.2	9.6	0.82	1
15	0.383	0	60.06	-0.51	-0.75	0.5	1.5	1.54	0
96	3.578	0	85.09	-0.74	-1.1	0.3	1.6	1.8	0
4.9	2.902	1	87.06	0.27	0.38	7.8	14	0.72	1
58.2	3.518	0	36.86	-0.17	0.64	0.3	1	1.14	1
6.6	3.485	0	35.94	-0.88	-0.23	0.9	12.8	1.45	1
11.1	2.119	1	86.57	-1.43	-0.33	3.7	24.5	1.16	1
7.5	2.502	1	176.56	-0.84	0.52	0.5	4.3	1.06	0
4.8	3.425	0	70.28	-0.79	-0.36	11.2	1.5	1.3	1
11.7	3.403	0	130.14	0.04	-0.05	0.3	5.3	1.22	1
60	0.715	1	100.34	-0.08	-0.72	0.2	6	1.6	0
3.4	3.198	0	24.41	0.94	2.2	0	5.6	0.9	1
8.7	3.11	0	70.44	-0.31	-1.1	1.2	8.5	0.95	1
2.9	3.209	0	49.45	-0.21	1.6	0.4	12.2	0.58	1
14.8	0.268	0	31.97	0.52	-0.26	0.5	1	1.58	1
168	0.025	1	107.99	0.2	1.38	1.8	16.2	1.36	0
69.8	3.014	0	90.61	-1.91	0.26	0	4.8	1.66	1
123	0.46	1	8.51	-1.44	-0.65	0	15.7	1.6	1
121	2.762	1	38.44	-0.15	0.09	0.4	2.3	1.52	1
86	1.306	1	55.06	0.06	-2.72	0.3	5	1.56	1
3.1	2.053	1	51.52	-3.66	-2.1	0	1.1	0.1	1
74	3.006	0	60.32	3.42	0.63	0	44.1	2	1
13.6	2.861	0	72.48	0.25	-1.09	0	0	1.16	1
1.2	1.227	1	57.86	0.24	1.33	0.3	0.1	0.64	1
58.7	2.264	1	36.96	1.48	0.62	21.9	87.7	1.6	1
62.2	0.841	1	82.89	-2.26	-2.77	0.3	0	1.14	1
4.8	0.917	1	124.81	0.11	0.35	0.3	1	1.08	1
51	2.765	0	61.7	-0.46	0.16	35.8	9	1.28	1
30.1	2.738	0	77.73	-1.1	-0.77	6	7.2	1.18	1
8.7	2.757	0	94.29	-1.43	-0.04	0.1	2.5	1.26	1
3.9	2.639	0	90.22	1.07	2.98	51.7	11.7	0.58	1
2.9	0.736	1	99.48	-1.49	-0.98	1.8	0.7	1.42	1
81	0.63	1	132.4	-1.5	-1.85	0	1	2.1	1
8.1	2.464	0	46.16	-1.44	-0.38	39.5	50.3	0.7	1
5.8	2.428	0	42.12	1.04	0.45	4.5	0.2	1.25	1

A.1 Continuação.

leuini	tempos	cens	idade	zpeso	zest	pas	vac	risk	r6
4.9	1.443	1	105.53	0.14	-0.19	1	55.7	2.1	1
340.8	0.654	1	132.47	-0.56	-0.67	0	3.6	1.72	1
23	2.355	0	153.13	1.59	0.64	0.9	1.8	1.38	1
27.2	2.278	0	84.34	-0.51	0.23	0	1.5	1.3	1
40.8	0.843	1	16.56	-1.63	-0.34	0	0.1	1.74	1
22.5	2.344	0	48.07	0.01	-0.41	1	2.1	1.68	1
13.2	2.171	0	54.93	0.07	-1.14	29.9	0	1.5	1
9.7	2.133	0	18.96	0.87	0.68	27.4	11.9	1.65	1
32.6	2.22	0	29.21	-0.66	0.05	3.3	15.9	1.2	1
8.1	1.322	1	39.69	0.12	0.63	57.4	5.6	1.26	1
113	0.594	1	60.85	0.43	0.71	0	2	2.28	1
19.4	1.96	0	94.46	1.26	-0.95	0	5	1.78	1
4.2	1.927	0	43.93	-0.56	-0.09	4.7	0	0.95	1
10.8	1.832	0	21.98	-0.7	-0.22	49.4	11.5	1	1
69.3	1.941	0	133.13	-0.71	-1.01	0	12	1.8	1
120	0.099	1	90.25	-0.73	-1.43	0.3	0	1.21	1
5.3	1.714	0	33.25	0.53	1.4	0	0	0.92	1
80.5	0.151	0	137.46	-1.21	-0.01	1.8	1.6	1.3	1
4.5	1.697	0	79.67	0.28	0.1	22	0.1	0.89	1
4	1.692	0	115.25	-0.48	0.45	45.7	39.5	0.62	1
1.2	0.214	1	169.07	-2.32	-1.95	3.3	2	0.88	1
69.4	1.624	0	52.96	-0.93	-1.08	37.1	17.9	1.52	1
4.1	1.566	1	75.17	-1.02	0.08	8.4	19.7	1.5	1
4.2	1.528	0	48.99	-1.56	-0.36	0.3	1	0.76	1
61	1.52	0	62	0.71	-0.99	0.4	0	1.06	1
620	0.487	1	115.22	2.06	1.51	0.6	1	2.7	0
2.1	1.481	0	81.64	0.04	0.48	83.1	64.4	0.78	1
107.5	1.41	0	105	-0.38	-0.15	40.5	5	1.4	1
11.4	0.003	0	63.08	-1.65	-0.34	0.5	1.5	1.28	1
1.3	1.259	0	98.3	-1.03	-0.55	21.3	68.7	1.1	1
1.4	1.205	0	49.68	-1.23	-2.55	0.7	0.3	0.78	1
65.4	1.18	0	79.11	0.31	1.01	0.4	10.1	1.4	1
9.7	0.572	1	66.76	-2.46	-3.05	71.4	19.7	1.12	1
3.8	1.12	0	97.18	-0.33	-0.16	5.7	4	1.7	1
3.6	1.103	0	20.47	-0.93	-0.42	52.3	8.2	1.42	1
31.7	1.065	0	141.54	-1.55	-0.59	4.2	6.5	1.12	1
6	0.498	1	23.69	-2.72	-2.21	1.5	40.5	0.92	1
9	0.991	0	52.27	-0.91	-0.35	1.5	4.9	1.2	1
17.1	0.991	0	74.55	-1.86	-1.18	7.9	3.1	1	1
26.1	0.994	0	86.7	-0.16	-0.34	6.6	5.7	0.88	1
112	0.898	0	57.43	-0.12	-0.99	3	1.7	1.7	1
7	0.969	0	37.91	-1.79	-1.61	0.9	1.1	1.6	1
5.9	0.895	0	90.09	-1.06	-0.96	0.2	2	0.85	1
102	0.893	0	56.54	0.35	-0.35	53	14.2	1.24	1
24.4	0.701	0	72.18	-2.68	-3.7	2.9	3.2	1.46	0
14.1	0.81	0	21.59	-0.82	-0.19	13.3	12.7	1.2	1
5.6	0.742	0	122.58	0	0.34	0.7	2.5	0.72	1
6.5	0.758	0	88.25	-0.97	-0.11	6.3	1.7	0.75	1

leuini em 1000 leucócitos/mm^3; tempos = resposta em anos; cens = 1 se falha e 0 se censura; indica censura; idade em meses; zpeso = peso padronizado pela idade e sexo; zest = altura padronizada pela idade e sexo; pas em %; Vac em %, risk = fator de risco em % e r6 = 1 se sucesso.

A.3 Conjunto de dados utilizados no estudo sobre aleitamento materno.

id	tempo	cens	V3	V2	V7	V11	V4	V1	V6	V10	V8	V9	V5
1	6	1	0	0	0	1	0	0	0	1	1	1	0
5	8	1	0	0	0	1	1	1	1	1	1	1	1
6	0.1	1	1	0	0	0	1	1	0	1	0	0	1
8	5	1	0	1	0	1	1	0	0	0	0	0	0
9	3	1	0	0	0	1	1	0	0	1	0	0	0
15	5	1	1	0	0	0	1	1	0	0	0	0	1
18	7	1	0	0	0	1	0	0	0	1	0	0	0
22	2	1	0	0	0	1	1	1	0	1	0	0	0
24	3	1	0	0	0	1	0	1	1	1	1	0	0
27	4	1	0	1	0	1	0	0	0	1	0	1	0
30	4	1	1	0	0	0	0	0	0	1	1	0	1
34	1	1	1	0	0	1	1	0	1	1	1	0	0
36	5	1	0	0	0	1	1	1	0	1	0	0	1
37	2.5	1	1	0	0	0	1	1	0	0	1	0	0
44	4	1	1	0	0	0	0	0	0	0	1	0	0
49	6	1	0	0	0	0	0	0	0	0	1	0	0
51	9	1	0	0	0	0	1	0	0	0	0	0	0
57	10	1	1	0	0	1	0	0	1	1	0	0	0
60	1	1	0	1	0	0	1	0	1	0	1	0	0
61	1	1	1	1	0	0	0	0	0	1	1	0	0
62	0.1	1	0	0	0	1	1	1	1	0	1	0	1
65	14	1	1	0	0	1	1	1	0	1	1	0	1
68	8	1	0	1	0	0	0	0	1	1	1	0	0
69	1.8	1	0	0	0	1	0	1	0	1	1	0	1
71	4	1	0	0	0	0	1	0	1	1	0	0	0
72	0.1	1	1	0	0	1	0	0	1	0	1	0	1
76	0.7	1	1	0	0	1	1	1	1	1	0	0	0
77	5	1	0	0	0	0	0	1	0	1	1	0	0
78	1	1	1	0	0	1	1	1	1	1	1	0	0
79	4	1	0	1	0	0	1	0	0	0	0	0	0
80	0.5	1	1	0	0	0	1	1	0	0	1	0	0
81	1	1	1	0	0	1	1	1	0	1	1	0	0
82	0.5	1	1	1	0	1	1	0	0	1	1	0	0
83	0.5	1	1	0	1	1	1	1	0	1	1	0	0
86	4	1	0	0	0	1	0	1	1	0	1	0	0
87	1	1	1	0	0	0	0	0	1	1	0	0	0
91	2.5	1	1	0	0	0	1	0	1	1	0	0	0
93	11.5	1	1	0	0	1	0	0	1	0	0	1	0
95	10	1	1	1	0	1	0	0	1	1	1	0	0
96	18	1	1	0	0	1	1	0	1	1	0	0	1
104	8	1	0	0	0	1	0	0	1	1	1	0	1
106	2	1	1	1	0	0	0	0	1	1	0	0	0
107	0.7	1	1	0	0	0	1	1	1	0	1	0	1
108	2.5	1	1	1	0	0	1	0	0	0	1	0	0
111	12	1	0	0	0	1	1	1	0	1	1	0	0
114	6	1	1	1	0	0	0	0	0	0	1	0	0
116	3.5	1	1	0	0	1	1	1	1	0	0	0	1
117	8	1	1	0	0	0	0	0	0	1	0	0	0
118	4	1	1	0	0	1	0	1	1	0	0	0	0
122	5	1	0	0	0	0	0	1	0	0	1	0	0
129	4	1	0	0	0	0	0	1	0	0	1	0	0
131	2.5	1	1	0	0	1	1	0	1	1	0	0	1
132	4	1	1	0	0	1	1	1	0	1	1	0	1
133	0.9	1	1	0	0	0	1	1	0	0	0	0	0
135	3.5	1	1	0	0	0	0	0	0	0	0	0	0
136	5	1	0	0	0	0	0	1	0	0	0	0	0
139	3	1	0	0	0	1	1	1	0	1	0	1	0
140	0.1	1	0	0	0	0	1	1	1	0	1	0	0
143	2.5	1	1	0	0	0	1	0	1	0	0	0	1
144	1	1	0	0	0	0	0	1	1	0	0	0	0
146	2.5	1	1	0	0	1	1	1	1	1	1	0	0
148	1.6	1	0	0	0	1	1	1	1	1	1	1	0
149	1.5	1	1	0	0	0	1	1	0	1	1	0	0
152	5	1	0	0	0	1	0	0	1	1	0	0	0
154	5.9	1	0	0	1	0	1	1	0	0	1	0	0
2	10	0	0	0	1	0	0	1	0	0	0	1	0
3	17	0	1	0	0	0	1	0	0	0	0	0	0
4	0.5	0	1	0	0	1	1	0	1	1	1	0	0
7	11	0	0	1	0	0	0	0	0	0	1	0	0
10	2	0	1	0	0	1	1	1	1	1	0	0	0
11	2	0	1	1	0	0	0	0	0	0	1	0	0
12	2	0	1	0	0	1	1	1	1	1	1	0	0
13	1	0	0	0	0	0	1	0	0	0	0	0	0
16	1	0	0	0	0	0	1	1	1	1	0	0	0

A.3 Continuação.

id	tempo	cens	V3	V2	V7	V11	V4	V1	V6	V10	V8	V9	V5
17	21	0	1	1	0	1	0	0	0	1	0	0	0
19	0.5	0	0	0	0	1	0	1	0	1	0	0	0
20	2	0	0	1	0	0	0	0	0	1	1	0	0
21	8	0	0	1	0	0	0	0	0	0	1	0	0
23	2	0	1	0	0	1	1	1	1	1	0	1	1
25	12	0	0	0	0	0	1	1	0	1	0	0	0
26	4	0	0	0	0	0	0	0	0	0	0	0	0
28	24	0	1	0	0	0	0	0	0	0	0	0	0
29	8	0	0	0	0	0	0	0	0	0	1	0	0
31	24	0	0	0	1	0	1	1	0	0	0	1	0
32	19	0	0	1	0	0	0	0	0	0	1	0	0
33	4	0	0	0	0	0	0	1	0	0	1	0	0
35	5	0	0	0	0	1	0	1	0	1	1	1	0
38	3.5	0	1	0	0	0	0	0	0	0	0	0	0
39	1	0	0	0	0	1	0	0	1	1	0	0	0
40	0.9	0	0	0	0	0	0	0	0	0	0	0	0
41	0.4	0	1	1	1	0	1	0	0	1	0	0	0
42	1	0	0	0	0	0	0	1	0	0	0	0	0
43	4	0	1	0	0	1	0	0	1	1	0	0	0
45	2	0	1	1	0	1	0	0	1	0	1	0	0
46	12	0	0	0	0	0	0	0	0	0	1	0	0
47	1	0	1	0	0	1	1	1	1	1	1	0	1
48	11	0	0	0	0	1	0	0	1	0	0	0	1
50	3	0	1	0	0	0	1	1	0	1	1	0	0
52	6	0	0	0	0	0	0	0	0	0	0	0	0
53	9	0	0	0	0	0	0	1	0	0	0	0	0
54	0.9	0	0	0	1	0	0	0	1	1	1	0	0
55	16	0	0	0	0	1	0	0	0	1	0	1	0
56	4	0	1	0	0	1	0	0	1	1	1	0	0
58	1	0	0	1	0	1	0	0	1	0	0	0	1
59	3	0	0	0	0	1	1	0	1	1	1	0	0
63	4	0	0	0	1	1	0	1	1	1	1	0	0
64	2	0	0	1	0	0	0	0	0	0	1	0	0
66	10	0	1	0	0	0	0	1	0	0	0	0	0
67	2	0	1	1	0	1	0	0	0	1	1	0	0
70	0.3	0	1	0	0	0	0	0	0	0	0	0	0
73	8	0	0	1	0	0	0	0	0	0	1	0	1
74	0.6	0	1	0	0	0	0	0	0	0	0	0	1
75	2	0	1	0	0	1	1	1	1	1	0	0	0
84	2	0	0	0	0	0	0	0	0	0	0	1	0
85	3	0	0	1	0	1	1	0	1	1	0	0	0
89	8	0	0	1	0	0	0	0	0	0	1	0	1
90	9	0	0	0	0	0	1	0	0	1	0	0	0
92	16	0	1	1	0	1	0	0	1	1	0	1	0
94	4	0	1	0	0	0	0	0	0	1	1	0	1
97	2	0	0	0	0	0	0	0	0	0	1	0	0
98	2	0	1	1	0	0	0	0	0	1	0	0	0
99	1	0	1	0	0	0	0	1	0	0	1	0	0
100	10	0	0	0	0	1	1	1	0	1	1	0	1
101	7	0	0	0	0	1	1	1	1	1	1	0	0
102	1	0	0	1	0	0	0	0	0	1	0	1	0
103	1	0	1	0	0	1	0	0	1	1	0	0	0
105	2	0	0	1	0	0	0	0	0	0	1	0	0
109	9	0	0	1	1	0	1	0	0	1	0	0	0
110	16	0	1	0	0	0	0	0	0	0	1	0	0
112	10	0	0	0	0	0	1	1	0	1	0	0	0
113	4	0	0	1	0	0	0	0	0	0	1	1	0
115	1	0	1	1	0	1	1	0	1	1	0	0	0
120	2	0	1	0	0	1	0	1	0	0	0	1	0
121	5	0	1	0	0	0	0	0	0	0	0	0	0
123	2	0	0	0	0	1	1	1	1	1	0	0	0
124	13	0	1	0	0	0	0	1	0	0	1	0	0
125	3	0	0	1	0	1	1	0	0	1	0	1	0
126	3	0	0	1	0	1	1	0	0	1	1	1	0
127	1.6	0	1	0	0	1	1	0	0	1	1	0	1
128	16	0	1	0	1	0	0	1	0	0	1	0	1
130	12	0	0	0	0	0	0	1	0	0	0	1	0
134	4	0	0	0	0	1	0	0	1	1	0	0	0
138	14	0	0	0	0	0	0	1	0	0	0	0	0
141	13	0	0	0	0	0	1	1	0	0	1	0	1
142	14	0	0	0	0	0	0	1	1	1	1	0	0
145	14	0	0	1	0	1	0	0	0	0	0	0	0
147	17	0	1	0	0	1	1	0	1	1	0	0	0
150	12	0	0	1	0	0	1	0	0	0	0	0	0
151	0.1	0	1	0	0	1	1	1	1	1	0	0	0
153	9	0	1	1	0	1	0	0	1	1	1	0	0

id = identificação da mãe; tempo = tempo de aleitamento materno (meses);
cens = indicadora de censura (1 = falha e 0 = censura); V1 a V11 descritas no texto.

A.7 Dados utilizados no estudo sobre câncer de laringe.

id	tempos	cens	idade	estagio	id	tempos	cens	idade	estagio
1	0.6	1	77	1	46	4.3	0	64	2
2	1.3	1	53	1	47	5.0	0	66	2
3	2.4	1	45	1	48	7.5	0	50	2
4	3.2	1	58	1	49	7.6	0	53	2
5	3.3	1	76	1	50	9.3	0	61	2
6	3.5	1	43	1	51	0.3	1	49	3
7	3.5	1	60	1	52	0.3	1	71	3
8	4.0	1	52	1	53	0.5	1	57	3
9	4.0	1	63	1	54	0.7	1	79	3
10	4.3	1	86	1	55	0.8	1	82	3
11	5.3	1	81	1	56	1.0	1	49	3
12	6.0	1	75	1	57	1.3	1	60	3
13	6.4	1	77	1	58	1.6	1	64	3
14	6.5	1	67	1	59	1.8	1	74	3
15	7.4	1	68	1	60	1.9	1	53	3
16	2.5	0	57	1	61	1.9	1	72	3
17	3.2	0	51	1	62	3.2	1	54	3
18	3.3	0	63	1	63	3.5	1	81	3
19	4.5	0	48	1	64	5.0	1	59	3
20	4.5	0	68	1	65	6.3	1	70	3
21	5.5	0	70	1	66	6.4	1	65	3
22	5.9	0	47	1	67	7.8	1	68	3
23	5.9	0	58	1	68	3.7	0	52	3
24	6.1	0	77	1	69	4.5	0	66	3
25	6.2	0	64	1	70	4.8	0	54	3
26	6.5	0	79	1	71	4.8	0	63	3
27	6.7	0	61	1	72	5.0	0	49	3
28	7.0	0	66	1	73	5.1	0	69	3
29	7.4	0	73	1	74	6.5	0	65	3
30	8.1	0	56	1	75	8.0	0	78	3
31	8.1	0	73	1	76	9.3	0	69	3
32	9.6	0	58	1	77	10.1	0	51	3
33	10.7	0	68	1	78	0.1	1	65	4
34	0.2	1	86	2	79	0.3	1	71	4
35	1.8	1	64	2	80	0.4	1	76	4
36	2.0	1	63	2	81	0.8	1	65	4
37	3.6	1	70	2	82	0.8	1	78	4
38	4.0	1	81	2	83	1.0	1	41	4
39	6.2	1	74	2	84	1.5	1	68	4
40	7.0	1	62	2	85	2.0	1	69	4
41	2.2	0	71	2	86	2.3	1	62	4
42	2.6	0	67	2	87	3.6	1	71	4
43	3.3	0	51	2	88	3.8	1	84	4
44	3.6	0	72	2	89	2.9	0	74	4
45	4.3	0	47	2	90	4.3	0	48	4

id = identificação do paciente; tempos = tempo até a morte (meses);
cens = indicadora de censura (1 = falha e 0 = censura); estágio = estágio da doença.

APÊNDICE B

Comandos Utilizados no Pacote Estatístico *R*

B.1 Obtenção da Figura 4.2
B.2 Obtenção da Figura 4.3
B.3 Obtenção da Figura 4.4
B.4 Modelos Ajustados na Seção 4.4.2
B.5 Obtenção da Figura 4.5
B.6 Obtenção da Figura 4.6
B.7 Obtenção da Figura 4.7
B.8 Obtenção da Figura 5.6
B.9 Obtenção da Figura 5.7
B.10 Obtenção da Figura 5.8
B.11 Obtenção da Figura 5.10

B.1 Obtenção da Figura 4.2

```
> ajust1<-survreg(Surv(dados$temp, dados$cens)~dados$lwbc, dist='exponential')
> ajust1
> x1<-4.0
> temp1<-0:150
> ax1<-exp(ajust1$coefficients[1]+ajust1$coefficients[2]*x1)
> ste1<-exp(-(temp1/ax1))
> x1<-3.0
> temp2<-0:150
> ax2<-exp(ajust1$coefficients[1]+ajust1$coefficients[2]*x1)
> ste2<-exp(-(temp2/ax2))
> par(mfrow=c(1,1))
> plot(temp1,temp1*0,pch=" ",ylim=range(c(0,1)), xlim=range(c(0,150)),
> xlab="Tempos",ylab="S(t) estimada",bty="n")
> lines(temp1,ste1,lty=2)
> lines(temp2,ste2,lty=4)
> abline(v=100,type="l",lty=3)
> legend(10,0.3,lty=c(2,4),c("lwbc = 4.0","lwbc = 3.0"),lwd=1, bty="n")
```

B.2 Obtenção da Figura 4.3

```
> temp<-c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65)
> t<-sort(temp)
> x<-dados$lwbc
> bo<- ajust1$coefficients[1]
> b1<- ajust1$coefficients[2]
> res<- t*exp(-bo-b1*x)
> ekm <- survfit(Surv(res,dados$cens)~1,type=c("kaplan-meier"))
> summary(ekm)
> par(mfrow=c(1,2))
> plot(ekm, conf.int=F,lty=c(1,1),xlab="resíduos",ylab="S(e) estimada")
> res<-sort(res)
> exp1<-exp(-res)
> lines(res,exp1,lty=3)
> legend(2,0.8,lty=c(1,3),c("Kaplan-Meier","Exponencial(1)"),lwd=1,bty="n",cex=0.7)
> st<-ekm$surv
> t<-ekm$time
> sexp1<-exp(-t)
> plot(st,sexp1,xlab="S(e) - Kaplan-Meier", ylab= "S(e) - Exponencial(1)",pch=16)
```

B.3 Obtenção da Figura 4.4

```
> temp<-c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65,
          56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)
> cens<-c(rep(1,17),rep(1,16))

> lwbc<-c(3.36,2.88,3.63,3.41,3.78,4.02,4.00,4.23,3.73,3.85,3.97,4.51,
          4.54,5.00,5.00,4.72,5.00,3.64,3.48,3.6,3.18,3.95,3.72, 4.0,
          4.28,4.43,4.45,4.49,4.41,4.32,4.90,5.0,5.0)
> grupo<-c(rep(0,17),rep(1,16))
> require(survival)
> ekm1<-survfit(Surv(temp,cens)~grupo)
> summary(ekm1)
```

```
> st1<-ekm1[1]$surv
> time1<-ekm1[1]$time
> invst1<-qnorm(st1)
> st2<-ekm1[2]$surv
> time2<-ekm1[2]$time
> invst2<-qnorm(st2)
> par(mfrow=c(1,3))
> plot(time1, -log(st1),pch=16,xlab="tempos",ylab="-log(S(t))")
> points(time2, -log(st2))
> legend(100,0.6,pch=c(16,1),c("Ag+","Ag-"),bty="n")
> plot(log(time1),log(-log(st1)),pch=16,xlab="log(tempos)",ylab="log(-log(S(t)))")
> points(log(time2),log(-log(st2)))
> legend(3,-1.5,pch=c(16,1),c("Ag+","Ag-"),bty="n")
> plot(log(time1),invst1,pch=16,xlab="log(tempos)",ylab=expression(Phi^-1 * (S(t))))
> points(log(time2),invst2)
> legend(0.5,-1,pch=c(16,1),c("Ag+","Ag-"),bty="n")
```

B.4 Modelos Ajustados na Seção 4.4.2

```
> dados<-as.data.frame(cbind(temp,cens,lwbc,grupo))
> attach(dados)
> require(survival)
> ajust1<-survreg(Surv(temp,cens)~1,dist='exponential')
> ajust1
> ajust2<-survreg(Surv(temp,cens)~lwbc,dist='exponential')
> ajust2
> ajust3<-survreg(Surv(temp,cens)~lwbc+grupo,dist='exponential')
> ajust3
> ajust4<-survreg(Surv(temp,cens)~lwbc+grupo+lwbc*grupo,dist='exponential')
> ajust4
```

B.5 Obtenção da Figura 4.5

```
> t<-temp
> x1<-lwbc
> x2<-grupo
> bo<-6.83
> b1<--0.7
> b2<--1.02
> res<- t*exp(-bo-b1*x1-b2*x2)
> ekm <- survfit(Surv(res,dados$cens)~1,type=c("kaplan-meier"))
> par(mfrow=c(1,2))
> plot(ekm, conf.int=F,lty=c(1,1),xlab="residuos",ylab="S(res) estimada")
> res<-sort(res)
> exp1<-exp(-res)
> lines(res,exp1,lty=3)
> legend(2,0.8,lty=c(1,3),c("Kaplan-Meier","Exponencial(1)"),lwd=1, bty="n", cex=0.8)
> st<-ekm$surv
> t<-ekm$time
> sexp1<-exp(-t)
> plot(st,sexp1,xlab="S(res): Kaplan-Meier", ylab= "S(res):Exponencial(1)",pch=16)
```

B.6 Obtenção da Figura 4.6

```
> x1<-4.0
> x2<-0.0
> temp1<-0:150
> ax1<-exp(6.83-0.70*x1-1.02*x2)
> ste1<-exp(-(temp1/ax1))
> x1<-3.0
> x2<-0.0
> temp2<-0:150
> ax2<-exp(6.83-0.70*x1-1.02*x2)
> ste2<-exp(-(temp2/ax2))
> par(mfrow=c(1,2))
> plot(temp1,temp1*0,pch=" ",ylim=range(c(0,1)), xlim=range(c(0,150)),
      xlab="Tempos",ylab="S(t) estimada",bty="n")
> lines(temp1,ste1,lty=1)
> lines(temp2,ste2,lty=2)
> legend(75,0.8,lty=c(1,2),c("lwbc = 4.0","lwbc = 3.0"),lwd=1, bty="n",cex=0.8)
> title("Ag+")
> x1<-4.0
> x2<-1.0
> temp1<-0:150
> ax1<-exp(6.83-0.70*x1-1.02*x2)
> ste1<-exp(-(temp1/ax1))
> x1<-3.0
> x2<-1.0
> temp2<-0:150
> ax2<-exp(6.83-0.70*x1-1.02*x2)
> ste2<-exp(-(temp2/ax2))
> plot(temp1,temp1*0,pch=" ",ylim=range(c(0,1)), xlim=range(c(0,150)),
> xlab="Tempos",ylab="S(t) estimada",bty="n")
> lines(temp1,ste1,lty=1)
> lines(temp2,ste2,lty=2)
> legend(75,0.8, lty=c(1,2),c("lwbc = 4.0","lwbc = 3.0"),lwd=1, bty="n",cex=0.8)
> title("Ag-")
```

B.7 Obtenção da Figura 4.7

```
> x1<-4.0
> x2<-0.0
> temp1<-0:150
> risco1<-1/(exp(6.83-0.70*x1-1.02*x2))
> risco1<-rep(risco1,151)
> x1<-3.0
> x2<-0.0
> temp2<-0:150
> risco2<-1/(exp(6.83-0.70*x1-1.02*x2))
> risco2<-rep(risco2,151)
> plot(temp1,temp1*0,pch=" ",ylim=range(c(0,0.1)), xlim=range(c(0,150)),
> xlab="Tempos",ylab="Risco estimado",bty="n")
> lines(temp1,risco1,lty=1)
> lines(temp2,risco2,lty=2)
> legend(100,0.08,lty=c(1,2),c("lwbc = 4.0","lwbc = 3.0"),lwd=1, bty="n",cex=0.8)
> title("Ag+")
```

```
> x1<-4.0
> x2<-1.0
> temp1<-0:150
> risco1<-1/(exp(6.83-0.70*x1-1.02*x2))
> risco1<-rep(risco1,151)
> x1<-3.0
> x2<-1.0
> temp2<-0:150
> risco2<-1/(exp(6.83-0.70*x1-1.02*x2))
> risco2<-rep(risco2,151)
> plot(temp1,temp1*0,pch=" ",ylim=range(c(0,0.1)), xlim=range(c(0,150)),
> xlab="Tempo",ylab="Risco estimado",bty="n")
> lines(temp1,risco1,lty=1)
> lines(temp2,risco2,lty=2)
> legend(100,0.08,lty=c(1,2),c("lwbc = 4.0","lwbc = 3.0"),lwd=1, bty="n",cex=0.8)
> title("Ag-")
```

B.8 Obtenção da Figura 5.6

```
> tt<-sort(tempos)
> aux1<-as.matrix(tt)
> n<-nrow(aux1)
> aux2<-as.matrix(cbind(ss$time,s0))
> S00<-rep(max(aux2[,2]),n)
> s0
> for(i in 1:n){
     if(tt[i]> min(aux2[,1])){
        i1<- aux2[,1]<= tt[i]
        S00[i]<-min(aux2[i1,2])}}
> ts0<-cbind(tt,S00)
> ts0
> b<-fit4$coefficients
> id<-50
> st1<- S00^(exp(b[4]*id)) # S(t|x) para estágio I e idade = 50 anos #
> st2<- S00^( exp(b[1]+( ( b[4]+b[5] )*id) ) ) # S(t|x) para estágio II e idade = 50 anos #
> st3<- S00^( exp(b[2]+( ( b[4]+b[6] )*id) ) ) # S(t|x) para estágio III e idade = 50 anos #
> st4<- S00^( exp(b[3]+( ( b[4]+b[7] )*id) ) ) # S(t|x) para estágio IV e idade = 50 anos #
> id<- 65
> st11<- S00^(exp(b[4]*id)) # S(t|x) para estágio I e idade = 65 anos #
> st21<- S00^( exp(b[1]+( ( b[4]+b[5] )*id) ) ) # S(t|x) para estágio II e idade = 65 anos #
> st31<- S00^( exp(b[2]+( ( b[4]+b[6] )*id) ) ) # S(t|x) para estágio III e idade = 65 anos #
> st41<- S00^( exp(b[3]+( ( b[4]+b[7] )*id) ) ) # S(t|x) para estágio IV e idade = 65 anos #
> par(mfrow=c(1,2))
> plot(tt,st1,type="s",ylim=range(c(0,1)),xlab="Tempos",ylab="S(t|x)",lty=1)
> lines(tt,st2,type="s",lty=2)
> lines(tt,st3,type="s",lty=3)
> lines(tt,st4,type="s",lty=4)
> legend(0,0.2,lty=c(1,2,3,4),c("estágio I","estágio II","estágio III","estágio IV"),lwd=1,bty="n",cex=0.7)
> title(" Idade = 50 anos")
> plot(tt,st11,type="s",ylim=range(c(0,1)),xlab="Tempos",ylab="S(t|x)",lty=1)
> lines(tt,st21,type="s",lty=2)
> lines(tt,st31,type="s",lty=3)
> lines(tt,st41,type="s",lty=4)
> legend(0,0.2,lty=c(1,2,3,4),c("estágio I","estágio II","estágio III","estágio IV"),lwd=1,bty="n",cex=0.7)
> title(" Idade = 65 anos")
```

B.9 Obtenção da Figura 5.7

```
tt<-sort(tempos)
aux1<-as.matrix(tt)
n<-nrow(aux1)
aux2<-as.matrix(cbind(ss$time,alpha0))
alpha00<-rep(min(aux2[,2]),n)
  for(i in 1:n){
   if(tt[i]> min(aux2[,1])){
      i1<- aux2[,1]<= tt[i]
      alpha00[i]<-max(aux2[i1,2])}
}
talpha0<-cbind(tt,alpha00)
b<-fit4$coefficients
id<-50
rt1<- alpha00*(exp(b[4]*id)) # risco  para estagio I e idade = 50 anos #
rt2<- alpha00*( exp(b[1]+( ( b[4]+b[5] )*id) ) ) # risco para estagio II e idade = 50 anos #
rt3<- alpha00*( exp(b[2]+( ( b[4]+b[6] )*id) ) ) # risco para estagio III e idade = 50 anos #
rt4<- alpha00*( exp(b[3]+( ( b[4]+b[7] )*id) ) ) # risco para estagio IV e idade = 50 anos #
id<-65
rt11<- alpha00*(exp(b[4]*id)) # risco para estagio I e idade = 65 anos #
rt21<- alpha00*(exp(b[1]+( ( b[4]+b[5] )*id) ) ) # risco para estagio II e idade = 65 anos #
rt31<- alpha00*(exp(b[2]+( ( b[4]+b[6] )*id) ) ) # risco para estagio III e idade = 65 anos #
rt41<- alpha00*(exp(b[3]+( ( b[4]+b[7] )*id) ) ) # risco para estagio IV e idade = 65 anos #
par(mfrow=c(1,2))
plot(tt,rt1,type="s",ylim=range(c(0,0.5)),xlab="Tempos", ylab="Risco(t | x)", lty=1)
lines(tt,rt2,type="s",lty=2)
lines(tt,rt3,type="s",lty=3)
lines(tt,rt4,type="s",lty=4)
legend(0,0.45, lty=c(1,2,3,4),c("estagio I","estagio II","estagio III","estagio IV"),lwd=1,bty="n",cex=0.7)
title(" Idade = 50 anos")
plot(tt,rt11,type="s",ylim=range(c(0,0.5)),xlab="Tempos", ylab="Risco(t|x)", lty=1)
lines(tt,rt21,type="s",lty=2)
lines(tt,rt31,type="s",lty=3)
lines(tt,rt41,type="s",lty=4)
legend(0,0.45, lty=c(1,2,3,4),c("estagio I","estagio II","estagio III","estagio IV"),lwd=1,bty="n",cex=0.7)
title(" Idade = 65 anos")
```

B.10 Obtenção da Figura 5.8

```
desmame<-read.table("c:/Temp/desmame.txt",h=T)
attach(desmame)
require(survival)
par(mfrow=c(2,2))
fit1<-coxph(Surv(tempo[V1==0],cens[V1==0])~1, data=desmame, x = T, method="breslow")
summary(fit1)
fit1$loglik
ss<- survfit(fit1)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
plot(ss$time,log(H0), xlim=range(c(0,20)),xlab="Tempos",ylab=expression(log(Lambda[0]*(t))),bty="n",type="s")
fit2<-coxph(Surv(tempo[V1==1],cens[V1==1]) ~ 1, data=desmame, x = T, method="breslow")
ss<- survfit(fit2)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
```

```
lines(ss$time,log(H0),type="s",lty=2)
legend(10,-3,lty=c(2,1),c("V1 = 1 (Nao)","V1 = 0 (Sim)"),lwd=1,bty="n",cex=0.7)
title("V1: Experiência Amamentação")
```

Obs: análogo para as demais covariáveis.

B.11 Obtenção da Figura 5.10

```
par(mfrow=c(2,3))
fit<-coxph(Surv(tempos[leuinic==1],cens[leuinic==1]) ~ 1, data=leucc, x = T, method="breslow")
ss<- survfit(fit)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
plot(ss$time,log(H0), xlab="Tempos",ylim=range(c(-5,1)),ylab = expression(log(Lambda[0]* (t))),bty="n",type="s")
fit<-coxph(Surv(tempos[leuinic==0],cens[leuinic==0]) ~ 1, data=leucc, x = T, method="breslow")
ss<- survfit(fit)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
lines(ss$time,log(H0),type="s",lty=4)
legend(1.5,-4,lty=c(4,1),c("leuini < 75","leuini > 75 "),lwd=1,bty="n",cex=0.8)
title("LEUINI")
```

Obs: análogo para as demais covariáveis.

APÊNDICE C

Comandos Utilizados no *Software SAS*

C.1 Ajuste de Diversos Modelos Gama Generalizados - Tabela 4.10

```
data desmame;
input id tempo cens V3 V2 V7 V11 V4 V1 V6 V10 V8 V9 V5;
V13=V1*V3;
V14=V1*V4;
V16=V1*V6;
V34=V3*V4;
V36=V3*V6;
V38=V3*V8;
V46=V4*V6;
V48=V4*V8;
V68=V6*V8;
cards;
1   6   1   0   0   0   1   0   0   0   1   1   1   0
5   8   1   0   0   0   1   1   1   1   1   1   1   1
...
153 9   0   1   1   0   1   0   0   1   1   1   0   0
;
proc lifereg;
model tempo*cens(0)= /distribution=gamma;
run;
proc lifereg;
model tempo*cens(0)=V1 /distribution=gamma;
run;
proc lifereg;
model tempo*cens(0)=V2 /distribution=gamma;
run;

proc lifereg;
model tempo*cens(0)= V1 V2 V3 V4 V6 V8 V9 /distribution=gamma;
run;
proc lifereg;
model tempo*cens(0)= V2 V3 V4 V6 V8 V9 /distribution=gamma;
run;

proc lifereg;
model tempo*cens(0))= V3 V4 V6 V8/distribution=gamma;
run;
proc lifereg;
model tempo*cens(0)= V3 V4 V6 V8 V34/distribution=gamma;
run;
proc lifereg;
model tempo*cens(0)= V3 V4 V6 V8 V68/distribution=gamma;
run;
```

APÊNDICE D

D.1 Método Iterativo de Newton-Raphson

O método iterativo de Newton-Raphson é um método numérico, usado para resolver um sistema de equações não-lineares, baseado na expansão de $U(\hat{\theta}_{(k)})$ em série de Taylor. Relembrando que a expansão em série de Taylor de 1^a ordem de uma função $f(x)$ em torno de x_0 é expressa por $f(x) = f(x_0) + f'(x_0)(x - x_0)$ segue, por analogia, e para $k = 1$, que:

$$U(\hat{\theta}_{(1)}) = U(\theta_{(0)}) + U'(\theta_{(0)})(\hat{\theta}_{(1)} - \theta_{(0)}).$$

Tomando-se, então, um valor inicial $\hat{\theta}_{(0)}$ para θ_0 e igualando-se a expressão obtida a zero obtém-se:

$$\begin{aligned} U(\hat{\theta}_{(0)}) \quad &+ \quad U'(\hat{\theta}_{(0)})\,(\hat{\theta}_{(1)}) - \; U'(\hat{\theta}_{(0)})\,\hat{\theta}_{(0)} = 0 \\ \hat{\theta}_{(1)} \quad &= \quad \hat{\theta}_{(0)} - \left[U'(\hat{\theta}_{(0)})\right]^{-1} U(\hat{\theta}_{(0)}) \\ \hat{\theta}_{(1)} \quad &= \quad \hat{\theta}_{(0)} - \left[\mathcal{F}(\hat{\theta}_{(0)})\right]^{-1} U(\hat{\theta}_{(0)}). \end{aligned}$$

em que $U'(\hat{\theta}_{(0)}) = \dfrac{\partial^2 \log(\theta)}{\partial^2 \theta} \mid_{\theta=\hat{\theta}_{(0)}} = \mathcal{F}(\hat{\theta}_{(0)})$ e $[\mathcal{F}(\hat{\theta}_{(0)})]^{-1}$ é a inversa da matriz $\mathcal{F}(\hat{\theta}_{(0)})$.

Repetindo esse procedimento para $k = 2$, e tomando-se $\hat{\theta}_{(1)}$ obtido no passo anterior, obtém-se:

$$\hat{\theta}_{(2)} = \hat{\theta}_{(1)} - \left[\mathcal{F}(\hat{\theta}_{(1)})\right]^{-1} U(\hat{\theta}_{(1)}).$$

No $(k+1)$-ésimo passo, a expressão para o método iterativo de Newton-Raphson será, portanto,

$$\hat{\theta}_{(k+1)} = \hat{\theta}_{(k)} + \left[\mathcal{F}(\hat{\theta}_{(k)})\right]^{-1} U(\hat{\theta}_{(k)}).$$

Um critério de parada (convergência) definido para esse procedimento iterativo é, por exemplo,

$$\frac{\mid \hat{\theta}_{(k)} \mid}{\mid \hat{\theta}_{(k+1)} \mid} < \epsilon$$

em que ϵ é um valor tão pequeno quanto desejável (por exemplo, $\epsilon = 10^{-8}$).

Algumas observações importantes sobre este método iterativo são:

1. se a função de verossimilhança for unimodal, o método apresenta-se bastante eficiente com convergência sendo obtida em poucos passos;
2. se a função de verossimilhança for multimodal, o método não é muito eficiente pois pode-se obter um máximo local em vez do máximo global;
3. se a função de verossimilhança apresentar um "platô", esse método, assim como tantos outros, apresentará problemas de convergência;
3. o método é muito sensível ao valor inicial, $\theta_{(0)}$, devendo este ser próximo de θ para que convergência possa ser obtida.