Análise de dados com riscos semicompetitivos

Elizabeth González Patiño

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística
Orientador: Profa. Dra. Gisela Tunes da Silva

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, outubro de 2012

Análise de dados com riscos semicompetitivos

Esta é a versão original da dissertação elaborada pela candidata Elizabeth González Patiño, tal como submetida à Comissão Julgadora.

Análise de dados com riscos semicompetitivos

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 16/08/2012. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dra. Gisela Tunes da Silva (orientadora) - IME-USP
- Prof. Dr. Antonio Carlos Pedroso de Lima - IME-USP
- Prof. Dra. Silvia Emiko Shimakura - UFPR

A minha mãe,
sua fortaleza e tenacidade
são meu exemplo.

Agradecimentos

Em especial, agradeço a minha orientadora, professora Gisela Tunes, pelo tempo, dedicação e paciência, sem seus bons conselhos não teria sido possível. A meus pais e meus irmãos Juan Da e Adri pela força embora a distância. A Juan Fer, pela compreensão, companhia e apoio incondicional; "Gracias niño". A todas as pessoas do IME que de uma ou outra forma contribuíram neste processo, em especial a Aleja, Fran, Diana e Gio pelo carinho e boas experiências. Também agradeço a Eliardo, Jayme A, Renata, Cata Rua, Nubia e Diego. A Alejandra Restrepo, Susana Ramirez e à Dra. Maria Isabel Múnera do Laboratorio Médico Echavarria de Colombia, pelos dados que motivaram o trabalho.

Agradeço ao professor Antonio Carlos Pedroso pela oportunidade.

Resumo

Em análise de sobrevivência, usualmente o interesse está em estudar o tempo até a ocorrência de um evento. Quando as observações estão sujeitas a mais de um tipo de evento (por exemplo, diferentes causas de óbito) e a ocorrência de um evento impede a ocorrência dos demais, tem-se uma estrutura de riscos competitivos. Em algumas situações, no entanto, o interesse está em estudar dois eventos, sendo que um deles (evento terminal) impede a ocorrência do outro (evento intermediário), mas não vice-versa. Essa estrutura é conhecida como *riscos semicompetitivos* e foi definida por Fine et al.(2001). Neste trabalho são consideradas duas abordagens para análise de dados com essa estrutura. Uma delas é baseada na construção da função de sobrevivência bivariada por meio de cópulas da família Arquimediana e estimadores para funções de sobrevivência são obtidos. A segunda abordagem é baseada em um processo de três estados, conhecido como *processo doença-morte*, que pode ser especificado pelas funções de intensidade de transição ou funções de risco. Neste caso, considera-se a inclusão de covariáveis e a possível dependência entre os dois tempos observados é incorporada por meio de uma fragilidade compartilhada. Estas metodologias são aplicadas a dois conjuntos de dados reais: um de 137 pacientes com leucemia, observados no máximo sete anos após transplante de medula óssea, e outro de 1253 pacientes com doença renal crônica submetidos a diálise, que foram observados entre os anos 2009-2011.

Palavras chave: Análise de sobrevivência, riscos semicompetitivos, cópula família Arquimediana, processo doença-morte, fragilidade compartilhada.

Abstract

In survival analysis, usually the interest is to study the time until the occurrence of an event. When observations are subject to more than one type of event (e.g, different causes of death) and the occurrence of an event prevents the occurrence of the other, there is a competing risks structure. In some situations, nevertheless, the main interest is to study two events, one of which (terminal event) prevents the occurrence of the other (nonterminal event) but not vice versa. This structure is known as *semicompeting risks*, defined initially by Fine et al. (2001). In this work, we consider two approaches for analyzing data with this structure. One approach is based on the bivariate survival function through Archimedean copulas and estimators for the survival functions are obtained. The second approach is based on a process with three states, known as *Illness-Death process*, which can be specified by the transition intensity functions or risk functions. In this case, the inclusion of covariates and a possible dependence between the two times is taken into account by a shared frailty. These methodologies are applied to two data sets: the first one is a study with 137 patients with leukemia that received an allogeneic marrow transplant, with maximum follow up of 7 years; the second is a dataset of 1253 patients with chronic kidney disease on dialysis treatment, followed from 2009 until 2011.

Keywords: survial analysis, semicompeting risks, Archimedean copula, illness-death process, shared frailty model.

Índice

1 Introdução **1**

1.1 Motivação do estudo 2

1.2 Censuras e riscos semicompetitivos 3

1.3 Dependência em Análise de Sobrevivência 4

1.4 Revisão bibliográfica 6

1.5 Organização do trabalho 9

2 Estudo da função de sobrevivência por meio de cópulas **11**

2.1 Notação 12

2.2 Cópulas 13

2.3 Estrutura de dependência 16

2.4 Estimação 22

2.4.1 Parâmetro de dependência da cópula α 23

2.4.2 Função de Sobrevivência Marginal $S_X(x)$ 27

2.4.3 Função de Sobrevivência bivariada 30

2.4.4 Função de sobrevivência condicional quando não há recaída 31

2.4.5 Função de sobrevivência condicional quando há recaída 32

3 Processo Doença - Morte **35**
3.1 Notação . . . 38
3.2 Modelo . . . 38
3.2.1 Modelo restrito, sem covariáveis . . . 40
3.2.2 Modelo geral, com covariáveis . . . 45
3.3 Estimação . . . 48
3.3.1 Modelo restrito, sem covariáveis . . . 50
3.3.2 Modelo geral, com covariáveis . . . 52
3.4 Propriedades dos estimadores . . . 58

4 Aplicação **61**
4.1 Transplante de Medula Óssea para Leucemia . . . 61
4.1.1 Análise descritiva por meio de cópulas . . . 63
4.1.2 Processo doença-morte . . . 74
4.2 Doença Renal Crônica . . . 78
4.2.1 Análise descritiva por meio de cópulas . . . 82
4.2.2 Processo doença-morte . . . 92

5 Discussão **97**

A Apêndice **99**
A.1 τ de Kendall para cópula de Frank . . . 99

CAPÍTULO 1

Introdução

Em análise de sobrevivência, o principal interesse é o estudo do tempo até a ocorrência de um evento que pode ser, por exemplo, o óbito. No entanto, existem muitas situações em que o óbito pode ocorrer devido a diferentes causas e deseja-se estudar separadamente cada tipo de causa. Neste contexto, surgem os modelos para dados com riscos competitivos, nos quais assume-se que o evento de interesse pode ocorrer por diversas causas ou que existem alguns eventos de interesse tais que um impede a ocorrência dos demais.

Para a análise de dados com estrutura de riscos competitivos, uma possível formulação é supor que existem tempos latentes correspondentes a cada tipo de falha e observa-se o mínimo entre esses tempos. Muitos modelos supõem independência entre os tempos latentes, o que nem sempre é razoável. Mais recentemente, surgiram modelos baseados na função de incidência acumulada (Klein e Moeschberger, 2003), que são baseados em uma formulação que não necessita da suposição de independência entre os tempos.

Dados com estrutura de riscos competitivos são muito comuns na área médica, em que o óbito pode ser devido ao tratamento (quimioterapia) ou devido à própria doença. Recentemente, os tratamentos para vários tipos de câncer evoluíram bastante, o que ocasionou uma diminuição na mortalidade de pacientes e um aumento na proporção de pacientes com recaída da doença. Desse modo, surge o interesse no estudo dos dois eventos: óbito e recaída. Alguns consideram a

estrutura de riscos competitivos para a análise do tempo até recaída e o tempo até óbito *livre de recaída* (ou seja, óbito antes da recaída). Observe que, neste caso, a estrutura de riscos competitivos só é válida se considerado o tempo até ocorrência do primeiro evento. Contudo se o interesse está no estudo do tempo até recaída e óbito (tendo ou não recaída), então a estrutura dos dados não é mais de riscos competitivos, mas sim de riscos *semicompetitivos*.

Mais especificamente, a estrutura de riscos semicompetitivos é caracterizada quando o interesse é o estudo da ocorrência de dois eventos, sendo um deles não terminal e outro terminal. Considera-se que o evento terminal impede a ocorrência do outro evento, porém, este último não impede a ocorrência do evento terminal.

Este tipo de estrutura é definida inicialmente por Fine et al. (2001) e apresenta outras aplicações na área médica além de estudos de câncer. Por exemplo, esta estrutura pode ser observada no estudo da insuficiência renal quando o paciente entra em estágios avançados da doença e, conseqüentemente, requer tratamento por meio de diálise. Nesta situação, o paciente também pode ter estes dois eventos: progressão da doença (i.e., o paciente apresenta piora no seu quadro clínico) e/ou óbito. Os pacientes são observados desde o início das diálises até a progressão da doença, que é determinada segundo avaliações mensais nos níveis de fósforo inorgânico, cálcio, nitrogênio uréico ou potássio no sangue, e óbito. É claro que, no decorrer do período de observação, o paciente pode não apresentar nenhum dos eventos até o fim do estudo, gerando assim dados censurados. Essa situação foi a principal motivação deste trabalho e está detalhadamente descrita a seguir.

1.1 Motivação do estudo

Atualmente, a doença renal crônica é considerada uma patologia de alto custo, além de ser uma doença com incidência relativamente alta. Especificamente na Colômbia, segundo o informe da CRC (2011), "Riesgo de progresión en Insuficiencia Renal Crónica", a cada ano a incidência da doença renal crônica terminal aumenta no país em aproximadamente 4000 pacientes. Estima-se que, para cada paciente em diálise, existem 18 pessoas com algum grau de possibilidade de patologia renal. No contexto mundial, é uma condição que afeta a mais de 2,1 milhões de pessoas. Deste modo, há grande interesse em estudos e pesquisas associadas a essa doença.

Um conjunto de dados de pacientes com esta patologia foi disponibilizado por Laboratório Echavarria e Fresenius Medical Care da Colômbia. Os dados correspondem ao acompanhamento

entre os anos 2009 e 2011 de 1253 pacientes, de várias cidades colombianas, com doença renal crônica em estágio 4 ou 5 (dano renal com diminuição severa da função renal ou falha renal) e que iniciaram tratamento de diálise.

Os dados contemplam informação de tempos de falha (óbito), exames mensais que indicam o estado do paciente e variáveis como sexo, idade, cidade.

É de interesse então, estudar o tempo de vida desses pacientes. Neste caso, técnicas já bem estabelecidas de análise de sobrevivência podem ser aplicadas. Porém, considerando que a doença renal crônica é uma patologia que gera um forte impacto econômico no Sistema de Seguridade Social em Saúde, também é importante o estudo da "progressão"ou "piora"do quadro clínico.

Desta forma, pode-se perceber que os dados apresentam uma estrutura de riscos semicompetitivos e uma análise estatística adequada deve considerar tal estrutura. No primeiro contato com o conjunto de dados, já pode ser percebido que a análise descritiva não seria trivial, pois curvas de Kaplan-Meier não são adequadas para a estimação da função de sobrevivência do evento não terminal (progressão da doença). Assim, essa foi a motivação para a busca de abordagens apropriadas para dados com riscos semicompetitivos. Buscou-se, dessa forma, estudar uma metodologia para uma análise descritiva apropriada dos dados e também para uma análise inferencial, via modelo de regressão.

1.2 Censuras e riscos semicompetitivos

Em estudos envolvendo o tempo de sobrevivência, uma característica comum nos dados é a existência de dados censurados. No contexto usual de sobrevivência, um dado é censurado se não é conhecido o instante exato de falha, mas sabe-se que a falha ocorreu dentro de um intervalo ou após um certo instante. Há vários tipos de censura, como censura à direita, à esquerda e intervalar.

Por exemplo, é comum a perda de contato de um paciente por mudança de plano de saúde, cidade, etc. Nesta situação, o paciente estava vivo na data da última consulta, por exemplo, e não há informação sobre a data do óbito; sabe-se apenas que o tempo de sobrevida foi maior do que um determinado valor. Neste caso, o dado foi censurado à direita, pois a informação disponível é que o instante do óbito é posterior ao último retorno do paciente. A censura à direita pode ser dividida também em alguns tipos: censura tipo I, censura tipo II e censura aleatória. Em linhas gerais, na censura tipo I, os pacientes são censurados se até o fim do estudo (que é fixo) não apresentaram o evento de interesse. Na censura tipo II, não há um tempo limite para o fim do estudo e os

pacientes são observados até que uma certa quantidade pré-determinada de falhas seja observada. Este tipo de censura ocorre principalmente em estudos de tempo de vida de equipamentos. Por fim, os dados estão sujeitos a censura aleatória quando os pacientes estão sujeitos a algum evento aleatório que os força a serem retirados do estudo antes de apresentarem o evento de interesse. Neste caso, assume-se que existe uma variável aleatória C que represente o tempo de censura e a observação é censurada se o tempo de censura C observado é menor do que o tempo de falha.

A censura à direita é o tipo de censura mais comum em estudos da área médica, porém, censura à esquerda e censura intervalar também podem ser observadas. No caso em que não é possível conhecer o tempo de falha de um paciente, mas sabe-se que é menor que um instante de censura C, então trata-se de censura à esquerda. Por fim, quando se sabe que o tempo de falha ocorreu num dado intervalo, mas não é conhecido o instante exato, então tem-se censura intervalar. Para uma discussão mais aprofundada sobre os tipos de censura, Klein e Moeschberger (2003) pode ser consultado.

No caso de riscos semicompetitivos, os dados também estão sujeitos a censura. Será considerada neste trabalho apenas censura aleatória à direita. Assim, o dado será considerado censurado se:

(i) o paciente não apresentou nenhum evento (recorrência ou óbito) antes da perda de acompanhamento, ou seja, a informação disponível é que a recaída pode ter acontecido após o tempo de censura e o óbito ocorreu após esse instante;

(ii) o paciente apresentou a recaída, porém, não foi observado o tempo do evento terminal (óbito).

É importante observar que, no caso em que o óbito ocorre antes da recaída, o evento intermediário (recaída) já não pode ser observado, mas não significa que seja censurado, simplesmente sua ocorrência foi impedida pelo óbito. O inverso, no entanto, não ocorre, pois ainda pode ser observado o óbito após a recorrência.

1.3 Dependência em Análise de Sobrevivência

Conforme já discutido, em dados com riscos semicompetitivos, o interesse está na análise do tempo até o evento intermediario e o tempo até o evento terminal. Assim, tem-se um par de variáveis aleatórias observadas no mesmo paciente, e é razoável considerar a existência de uma estrutura

de dependência entre essas duas variáveis. Portanto, torna-se conveniente uma discussão sobre situações que podem gerar dependência em dados envolvendo o tempo até a ocorrência de um evento.

Em análise de sobrevivência, existem muitos mecanismos que geram dependência. Hougaard (2000) os classifica em três principais categorias:

i. **Eventos comuns:** quando vários eventos acontecem simultaneamente, i.e, dados paralelos, por exemplo, acidentes ou desastres que levam a morte de muitas pessoas ao mesmo tempo.

ii. **Riscos comuns:** quando os indivíduos objeto de estudo são dependentes pela existência de alguns fatores de risco comuns que, geralmente, são não observáveis. Neste caso, são incluídos efeitos aleatórios no modelo com a função de englobar os fatores comuns não observáveis. O ponto chave é a independência condicional quando os riscos comuns são conhecidos. Aqui são usados modelos de risco latente, sendo o modelo de fragilidade o mais comum. A dependência gerada no modelo de riscos semicompetitivos faz parte deste tipo de dependência, pois se um indivíduo apresenta algum dos eventos de interesse, recaída ou óbito, isto não muda a sobrevida de cada um dos outros pacientes, mas sim o conhecimento da sobrevivência deles.

iii. **Evento-Relativo:** o evento atual pode mudar o risco de eventos futuros, por exemplo, quando uma pessoa adquire um vírus, o risco que outras sejam infectadas aumenta.

Além da natureza da dependência, há outra consideração importante: a duração da dependência. Podem ser consideradas as seguintes situações:

- Dependência instantânea: dois ou mais eventos ocorrem ao mesmo tempo;
- Dependência de curto prazo: a dependência é mais pronunciada imediatamente após outros indivíduos no grupo experimentarem o evento;
- Dependência de longo prazo: um evento implica que o risco entre os membros do outro grupo incrementa para sempre.

Intuitivamente, espera-se que, na maior parte das aplicações, a dependência em dados com estrutura de riscos semicompetitivos possivelmente se enquadre em dependências de curto ou longo prazo, esta última devido a que em um maior tempo de observação a proporção de óbito

poderia aumentar em relação aos pacientes em que é observada a recaída, aumentando então a força da dependência entre os dois eventos.

Na quantificação da dependência entre duas variáveis X e Y podem ser usadas medidas bivariadas, sendo que a mais conhecida é o coeficiente tradicional de correlação ρ de Pearson. No entanto, esta medida não é preservada nas cópulas (Hougaard, 2000), isto é, dois pares de variáveis correlacionadas com a mesma cópula podem ter diferentes correlações. Além disso, não é uma medida apropriada dado que as distribuições marginais não são normais e a estrutura de dependência é não linear no contexto de variáveis aleatórias que medem tempo de falha.

Dessa forma, existem outras medidas mais gerais, como o coeficiente de concordância τ de Kendall (Hougaard, 2000) que é constante para uma mesma cópula, isto é, quaisquer variáveis correlacionadas da mesma cópula apresentam o mesmo τ correspondente a essa cópula.

Também podem ser usados o coeficiente de correlação de Spearman (ρ_S) e concordância mediana. Todos os coeficientes mencionados não dependem da unidade de medida.

1.4 Revisão bibliográfica

Atualmente, existem na literatura alguns trabalhos que abordam a análise de dados com riscos semicompetitivos. As principais dificuldades na análise de dados com esse tipo de estrutura são, além da existência de dados censurados, (*i*) a hipótese de que se tem um par de variáveis aleatórias com a restrição que uma é menor que a outra (a recaída necessariamente ocorre antes do óbito) ou uma delas não é observada e (*ii*) a incorporação da provável dependência entre o tempo até a observação do evento terminal e o tempo até a observação do evento intermediário. Segundo Hsieh et al. (2008), se a relação entre os dois eventos intermediário (não terminal) e terminal não é completamente especificada, então a distribuição marginal do tempo até a observação do evento não terminal é não identificável devido à possível censura dependente.

Assim, partindo da idea de os eventos (intermediário e terminal) estarem associados, Fine et al. (2001) propõem incorporar a dependência entre os tempos observados, definidos como: X tempo até o evento intermediário (recaída) e Y tempo até o evento terminal (óbito), através de um modelo de fragilidade gama no quadrante superior ($X \leq Y$), gerando assim uma função de sobrevivência bivariada, que no caso sem covariáveis, corresponde à cópula de Clayton (1978), pertencente à família de cópulas Arquimedianas. Devido à dependência entre os eventos, não podem ser usados os procedimentos habituais para a estimação das funções de sobrevivência marginais, as quais

podem ser usadas para a estimação da função de sobrevivência bivariada por meio de cópulas. Deste modo, Fine et al. (2001) propõem também um estimador não paramétrico para as funções de sobrevivência marginais e derivam suas propriedades assintóticas.

Considerando que a cópula de Clayton pertence à família de cópulas Arquimediana, Lakhal e Abdous (2008) mostram que o desenvolvimento de Fine et al. (2001) pode ser estendido para qualquer cópula desta família dado que elas fornecem uma forma conveniente de expressar distribuições conjuntas de duas ou mais variáveis aleatórias, principalmente quando elas podem ser correlacionadas. Os autores encontram, a partir do estudo de Oakes (1989) que trata modelos de sobrevivência bivariados induzidos por fragilidades, uma equação de estimação para o parâmetro de dependência da cópula que é importante para obter a estimação da função de sobrevivência conjunta e das marginais, principalmente $S_X(x)$ que corresponde à função de sobrevivência do tempo até o evento não terminal, X.

É importante ressaltar que o estimador Kaplan-Meier não é apropriado para estimar a função de sobrevivência marginal $S_X(x)$ do evento não terminal. Assim, Lakhal e Abdous (2008) propõem usar o estimador Cópula-Gráfico introduzido por Zheng e Klein (1982), que estudam funções de sobrevivência marginal em riscos competitivos assumindo uma cópula conhecida. Este estimador foi estudado posteriormente por Rivest e Wells (2001) que obtiveram suas propriedades assintóticas usando teoria de processos de contagem. Essa abordagem é extremamente interessante por possibilitar a construção de gráficos, semelhantes a gráficos com curvas de Kaplan-Meier, com estimativas de funções de sobrevivência de interesse (marginais e condicionais) obtidas incorporando de forma conveniente a dependência entre os dois tempos observados.

Uma outra abordagem, ainda mais recente, é apresentada por Xu et al. (2010), em que o estudo dos eventos intermediário (recaída) e terminal (óbito) é feito com base em um processo de três estados, conhecido na literatura como *Processo Doença-Morte*, estudado inicialmente por Fix e Neyman (1951). No processo doença-morte, o estado inicial é representado pelo momento em que o paciente passa a integrar o estudo; depois, quando o evento intermediário é observado, o paciente faz uma transição ao segundo estado e depois, se for o caso, passa ao terceiro estado, chamado também estado absorvente, que corresponde ao evento terminal. No entanto, o paciente pode passar diretamente a este terceiro. Como a dependência entre os eventos intermediário e terminal é uma das características principais dos riscos semicompetitivos, os autores a incluem no modelo por meio de uma fragilidade compartilhada, com distribuição gama.

Uma vantagem deste processo é que, além da inclusão da fragilidade nas funções de intensidade

de transição, também podem ser incorporadas covariáveis às funções de risco condicionadas à fragilidade, através de um modelo de regressão semelhante ao modelo de riscos proporcionais de Cox (1972). Nesta proposta, a estimação dos parâmetros é feita por meio de máxima verossimilhança. Xu et al. (2010) apresentam um algoritmo para a aproximação numérica, e discutem as propriedades assintóticas, seguindo a abordagem baseada nos resultados de Murphy (1995) referentes à teoria assintótica para modelos de fragilidade.

Considerando a incorporação de covariáveis no problema de riscos semicompetitivos, há ainda algumas outras referências recentes. Uma delas é o trabalho de Peng e Fine (2007). Os autores apresentam um estudo de modelo de regressão de dados com riscos semicompetitivos formulando os efeitos das covariáveis na função de sobrevivência marginal do evento não terminal, que influencia na sobrevivência conjunta, dado que é modelada segundo o esquema de cópula proposto por Fine et al. (2001). Os estimadores propostos são não paramétricos, derivados de equações de estimação não lineares. Os autores demonstram que estes estimadores são uniformemente consistentes e convergem fracamente a um processo Gaussiano. Peng e Fine (2007) também apresentam o desenvolvimento de testes não paramétricos para o efeito das covariáveis e os parâmetros das cópulas usadas. É importante ressaltar que a diferença em relação ao trabalho de Xu et al. (2010) que estudam riscos semicompetitivos segundo o processo doença-morte, é que neste caso é possível observar o efeito das covariáveis em três situações (ou estados): recaída, óbito e óbito após recaída.

Hsieh et al. (2008) também propõem um modelo de regressão para dados deste tipo, com enfoque no modelo para o evento não terminal. Para tratar o problema de não identificabilidade deste evento, os autores usam cópulas da família Arquimediana (Clayton e Frank), e propõem procedimentos para escolher a cópula que permite um melhor ajuste. A metodologia é desenvolvida para covariáveis discretas.

Uma proposta ainda mais atual para análise de dados com riscos semicompetitivos é a proposta de Chen (2012), que apresenta modelos de regressão marginais para ambos os eventos, não terminal (intermediário) e terminal, e assume o modelo de cópula para a distribuição conjunta. O autor propõe estimação dos parâmetros por máxima verossimilhança não paramétrica, apresenta desenvolvimentos teóricos diretos com processos de contagem e propõe uma implementação computacional. O trabalho de Chen (2012) é mais geral, comparado com as outras metodologias, dado que a formulação dos modelos é feita em termos de processos de contagem. Esta abordagem não é estudada detalhadamente neste trabalho devido sua recente publicação.

Até agora só foi tratada a estrutura de riscos semicompetitivos assumindo que o esquema de censura é a direita. No entanto outras situações são encontradas em estudos médicos. Por exemplo, Porta (2010) apresenta um trabalho de doutorado motivado pelo maior estudo de câncer de bexiga na Espanha (SBC/EPICURO) com esquema de censura intervalar. No estudo, a autora considera modelos de riscos competitivos e modelos multiníveis. No entanto, com estas metodologias não foi possível estudar a distribuição marginal da progressão da doença devido ao problema de identificabilidade. Deste modo, Porta (2010) amplia a metodologia semiparamétrica de Fine et al. (2001), assumindo a cópula de Clayton (1978), e modifica o problema de riscos semicompetitivos quando os dados estão censurados num intervalo. A autora desenvolve um algoritmo iterativo para estimar conjuntamente tanto o parâmetro da cópula quanto a função de sobrevivência marginal do evento não terminal.

Li e Peng (2011) consideram o problema de riscos semicompetitivos quando os dados estão sujeitos a truncagem à esquerda do evento terminal, ou seja, quando não é observado o evento terminal em parte da população. Segundo Li e Peng (2011), isto pode complicar a análise de regressão sob o evento não terminal, como nos modelos estudados por Peng e Fine (2007) e Hsieh et al. (2008), de modo que para esta nova situação os autores propõem um método de regressão quantílica, que apresenta a boa característica de que as interpretações são simples. Os autores apresentam também procedimentos de estimação e inferência, além de propriedades assintóticas dos estimadores (consistência e convergência).

De maneira geral, os trabalhos que tratam da análise de dados com riscos semicompetitivos são recentes e a literatura ainda é um pouco escassa. Há, por conseguinte, muitos problemas que não foram abordados nesse contexto (como avaliação de qualidade do ajuste, entre outros) que precisam ainda ser devidamente trabalhados.

1.5 Organização do trabalho

Neste capítulo, foram apresentados alguns conceitos preliminares e descrita a estrutura dos riscos semicompetitivos que será utilizada ao longo do trabalho. O Capítulo 2 é dedicado ao estudo e estimação da função de sobrevivência conjunta dos eventos intermediário e terminal por meio de cópulas, que pode ser utilizada na análise descritiva dos dados. A inclusão de covariáveis em um modelo de regressão é discutida no Capítulo 3, em que um modelo de fragilidade compartilhada é utilizado para modelar as intensidades de transição em um processo doença-morte.

No Capítulo 4 é apresentada uma aplicação dos modelos estudados a dois conjuntos de dados reais, um contendo pacientes com leucemia, observados após transplante de medula óssea e, outro, composto por pacientes com doença renal crônica submetidos a tratamento de diálise. Por fim, uma discussão da metodologia e as conclusões são apresentadas no Capítulo 5.

CAPÍTULO 2

Estudo da função de sobrevivência por meio de cópulas

A primeira etapa de qualquer análise estatística é uma análise descritiva completa dos dados. No caso de dados com estrutura de riscos semicompetitivos, a primeira ideia seria a utilização do estimador de Kaplan-Meier para a análise descritiva do tempo até os eventos de interesse. Embora o estimador de Kaplan-Meier possa ser utilizado para a estimação da função de sobrevivência do evento terminal, para a estimação da função de sobrevivência do evento não terminal, este estimador não é apropriado, dada a possível dependência entre os eventos de interesse. Se for considerado que o óbito é também censura para o evento não terminal, a suposição de censura independente não é verificada e, portanto, não se pode utilizar técnicas de análise estatística que se baseiam nessa suposição. Nesse contexto, foram propostos vários estimadores: estimador *não paramétrico* de Fine et al. (2001), estimador *auto consistente* de Jiang et al. (2005) e o estimador *Cópula-Gráfico* de Zheng e Klein (1982), que foi adaptado por Rivest e Wells (2001) para censura dependente e por Lakhal e Abdous (2008) para dados com estrutura de riscos semicompetitivos.

O propósito deste capítulo é a estimação das funções de sobrevivência marginais que usualmente são de interesse para uma análise descritiva de um conjunto de dados de tempos de vida que satisfazem a estrutura de riscos semicompetitivos. Mais especificamente, será considerada a abordagem de Lakhal e Abdous (2008), que utiliza cópulas Arquimedianas para modelar a possível dependência entre os tempos até os eventos de interesse. Para isso, inicialmente será

apresentada a caracterização dos tempos até os eventos por meio de cópulas para, posteriormente, ser considerada a estimação de quantidades de interesse tais como:

- A probabilidade de sobrevida de um paciente, dado que no instante x não teve recaída nem apresentou o óbito, isto é, $P(Y > y|X > x, Y > x)$;
- A probabilidade de sobrevida de um paciente dado que no tempo x teve recaída e continua vivo no instante $t > x$, ou seja, $P(Y > y|X = x, Y > t)$.

Vale ressaltar que, neste capítulo, não será considerada a inclusão de covariáveis, a ser abordada no Capítulo 3.

2.1 Notação

Será denotado por Y o tempo até ocorrência do evento terminal (óbito) e por X o tempo até a ocorrência do evento intermediário (não terminal, recaída). A variável aleatória associada ao tempo de censura será denotada por B. Neste trabalho será considerada a situação em que a censura B, à direita, é independente de (X, Y). Será denotada por $\pi(x, y)$ a função de sobrevivência conjunta de X e Y, ou seja,

$$\pi(x, y) = S_{XY}(x, y) = P(X > x, Y > y), \qquad 0 \leq x \leq y, \tag{2.1}$$

e por $S_X(x)$ e $S_Y(y)$ as funções de sobrevivência marginais de X e Y, respectivamente.

São definidas ainda as variáveis $Z = \min(X, Y)$ e $\delta_x = I_{\{X<Y\}}$, com $I_{\{X<Y\}}$ uma função indicadora, ou seja, $I_{\{X<Y\}} = 1$ se $X < Y$ e $I_{\{X<Y\}} = 0$, caso contrário. Observe que δ_x será igual a zero se o evento terminal for observado antes do intermediário e, dessa forma, o evento intermediário não será observado. As seguintes quantidades são observáveis:

- $R = \min(Y, B)$, em que B é independente de (X, Y);
- $S = \min(Z, B)$;
- $\delta_y = I_{\{Y<B\}}$;
- $\delta_z = I_{\{Z<B\}}$;
- $\delta_{xz} = \delta_x \delta_z$.

Observe que, devido à censura, δ_x nem sempre é conhecido. No entanto, se δ_x não é conhecido, $\delta_z = 0$ e δ_{xz} será igual a zero.

Com estas definições, o conjunto de dados consiste de n réplicas independentes

$\Delta = \{(R_i, S_i, \delta_{y_i}, \delta_{z_i}, \delta_{xz_i}), i = 1, \cdots, n\}$ das variáveis observáveis $(R, S, \delta_y, \delta_z, \delta_{xz})$.

Por exemplo, para o i-ésimo indivíduo que apresentou os dois eventos (intermediário e terminal) tem-se, $\Delta = \{(R_i = x, S_i = y, \delta_{y_i} = 1, \delta_{z_i} = 1, \delta_{xz_i} = 1)$. Quando no i-ésimo indivíduo é observado só o evento terminal, então $\Delta = \{(R_i = y, S_i = y, \delta_{y_i} = 1, \delta_{z_i} = 1, \delta_{xz_i} = 0)$. Quando o i-ésimo indivíduo só apresentou o evento intermediário (recaída) tem-se $\Delta = \{(R_i = b, S_i = x, \delta_{y_i} = 0, \delta_{z_i} = 1, \delta_{xz_i} = 1)$ e finalmente, se nenhum dos eventos foi observado no i-ésimo indivíduo então $\Delta = \{(R_i = b, S_i = b, \delta_{y_i} = 0, \delta_{z_i} = 0, \delta_{xz_i} = 0)$.

2.2 Cópulas

As cópulas fornecem uma forma conveniente de expressar distribuições conjuntas de duas ou mais variáveis aleatórias. No caso de riscos semicompetitivos, trata-se de duas variáveis não negativas que representam os tempos até a observação dos eventos intermediário (não terminal) X e terminal Y. A cópula pode ser definida como uma função que *acopla* as funções de distribuição marginais $F(x)$ e $G(y)$ das variáveis X e Y, respectivamente, numa função de distribuição bivariada $H(x, y)$ por meio da relação $H(x, y) = C(F(x), G(y))$. Um resultado que estabelece a relação entre cópulas e funções bivariadas é o teorema de Sklar, de grande importância na teoria de cópulas e para as aplicações em estatística. A seguir enuncia-se o teorema, e detalhes da prova podem ser consultados em Nelsen (2006).

Teorema 2.1 *(Teorema de Sklar.) Seja H uma função de distribuição bidimensional com funções de distribuição marginais F e G. Então existe uma cópula C tal que $H(x, y) = C\{F(x), G(y)\}$. Reciprocamente, para qualquer funções de distribuição F e G e qualquer cópula C, a função H definida acima é uma função de distribuição bidimensional com marginais F e G. Além disso, se F e G são contínuas, C é única.*

Assim, o teorema de Sklar mostra que as cópulas são uma boa alternativa, na construção de distribuições multivariadas a partir das distribuições marginais, principalmente quando tem-se uma dependência entre as variáveis.

Sejam (X, Y) variáveis aleatórias contínuas com funções de sobrevivência e funções de densidade de probabilidade marginais dadas por (S_X, S_Y) e (f_X, f_Y), respectivamente. Considerando

o contexto de riscos semicompetitivos, em que X denota o evento intermediário e Y é o evento terminal, a correspondente função de sobrevivência conjunta baseada em uma cópula C_α pode ser expressa como

$$\begin{aligned} S_{XY}(x,y) &= P(X > x; Y > y) \\ &= C_\alpha\{S_X(x), S_Y(y)\}, \qquad 0 \leq x \leq y, \end{aligned} \tag{2.2}$$

e a função de densidade conjunta como

$$f_{XY}(x,y) = C_\alpha\{S_X(x), S_Y(y)\} f_X(x) f_Y(y). \tag{2.3}$$

O parâmetro α nas funções de sobrevivência e densidade conjuntas (dadas em 2.2 e 2.3) é um parâmetro da cópula que está associado à dependência entre as variáveis que representam os eventos de interesse (intermediário e terminal).

Neste trabalho serão consideradas as cópulas da família Arquimediana, indexadas pelo parâmetro α, que têm a seguinte forma

$$C_\alpha(u,v) = \phi_\alpha^{-1}\{\phi_\alpha(u) + \phi_\alpha(v)\} \qquad 0 \leq u \leq 1, 0 \leq v \leq 1, \tag{2.4}$$

em que $\phi_\alpha(\cdot)$ é conhecida como função geradora da cópula. A função $\phi_\alpha(\cdot)$ é contínua, estritamente decrescente ($\phi_\alpha^{'}(\cdot) < 0$) e convexa ($\phi_\alpha^{''}(\cdot) \geq 0$), definida de [0,1] a [0, ∞] tal que $\phi_\alpha(1) = 0$ (Nelsen, 2006). Suas derivadas de primeira e segunda ordem, assim como sua função inversa, serão muito importantes para a obtenção de resultados descritos mais adiante.

No trabalho realizado por Lakhal e Abdous (2008), a metodologia é desenvolvida para cópulas da família Arquimediana, porém, os autores consideram em detalhes apenas as cópulas de Clayton e Frank. Além dessas duas cópulas, existem varias cópulas que pertencem a esta família, sendo que Nelsen (2006) cita vinte e duas delas. Neste trabalho serão consideradas, além das cópulas de Clayton e Frank, mais duas outras cópulas, uma delas dada em Nelsen (2006) (cópula 4.2.2, p. 116), escolhida pela forma simples da função geradora, e a cópula de Gumbel, bastante referenciada na literatura e usada por Zheng e Klein (1982).

A seguir serão descritas com detalhes as quatro cópulas consideradas neste trabalho denotadas por conveniência como família A, família B, família C e família D .

Família A

Esta família é conhecida como cópula de Clayton e foi proposta inicialmente por Kimeldorf e Sampson (1975). Posteriormente ficou conhecida como cópula de Clayton com o trabalho de Clayton (1978), foi definida também em Joe (1997) e Alsina et al. (2006). Esta cópula é dada por

$$C_\alpha(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \qquad \alpha \in [-1,\infty)/\{0\}, \tag{2.5}$$

com função geradora $\phi_\alpha(k) = \dfrac{1}{\alpha}(k^{-\alpha} + 1)$ e derivadas de primeira e segunda ordem iguais a $\phi'_\alpha(k) = -k^{-\alpha-1}$ e $\phi''_\alpha(k) = (\alpha+1)(k^{-\alpha-2})$. Sua função geradora inversa é dada por $\phi_\alpha^{-1}(k) = \left(\dfrac{1}{\alpha k + 1}\right)^{1/\alpha}$.

Família B

Esta cópula está definida em Nelsen (2006) (cópula 4.2.2, pg. 116), adaptada de Alsina et al. (2006), e é dada por

$$C_\alpha(u,v) = 1 - [(1-u)^\alpha + (1-v)^\alpha]^{1/\alpha}, \qquad \alpha \in [1,\infty). \tag{2.6}$$

A função geradora é $\phi_\alpha(k) = (1-k)^\alpha$ e suas derivadas de primeira e segunda ordem são $\phi'_\alpha(k) = -\alpha(1-k)^{\alpha-1}$ e $\phi''_\alpha(k) = \alpha(\alpha-1)(1-k)^{\alpha-2}$, respectivamente. Sua função geradora inversa é $\phi_\alpha^{-1}(k) = 1 - k^\alpha$.

Família C

Esta é a cópula de Frank, definida inicialmente por Frank (1979), dada por

$$C_\alpha(u,v) = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\right), \qquad \alpha \in (-\infty,\infty)/\{0\}, \tag{2.7}$$

com função geradora $\phi_\alpha(k) = -\ln\left(\dfrac{e^{-\alpha k}-1}{e^{-\alpha}-1}\right)$ e derivadas de primeira e segunda ordem dadas por $\phi'_\alpha(k) = -\dfrac{\alpha e^{-\alpha k}}{1-e^{-\alpha k}}$ e $\phi''_\alpha(k) = \dfrac{\alpha^2 e^{-\alpha k}}{(1-e^{-\alpha k})^2}$. Sua função geradora inversa é

$$\begin{aligned}\phi_\alpha^{-1}(k) &= -\frac{1}{\alpha}\ln(1 + \exp\{\ln(e^{-\alpha}-1) - k\}) \\ &= -\frac{1}{\alpha}\ln(1 + (e^{-\alpha}-1)e^{-k}).\end{aligned}$$

Família D

Por fim, será considerada a cópula de Gumbel (1960) que é definida por

$$C_\alpha(u,v) = \exp\{-[(-\ln u)^\alpha + (-\ln v)^\alpha]^{1/\alpha}\}, \qquad \alpha \in [1,\infty), \tag{2.8}$$

com função geradora $\phi_\alpha(k) = (-\ln k)^\alpha$ e suas derivadas de primeira e segunda ordem dadas por $\phi'_\alpha(k) = -\dfrac{\alpha}{k}(-\ln k)^{\alpha-1}$ e $\phi''_\alpha(k) = \dfrac{\alpha}{k^2}(-\ln k)^{\alpha-1}\left(\dfrac{\alpha-1}{(-\ln k)}+1\right)$. A função geradora inversa é $\phi_\alpha^{-1}(v) = \exp\{-v^{1/\alpha}\}$.

2.3 Estrutura de dependência

O parâmetro α nas cópulas, como dito anteriormente, está associado à dependência entre os tempos de falha, ou seja, a dependência na função dada em (2.2). Mesmo sem ter interpretação própria, é útil para medir a capacidade do evento intermediário (não terminal) predizer a ocorrência do evento terminal (Lakhal e Abdous, 2008).

Dado que α é um parâmetro referente à intensidade da dependência entre os tempos de falha, é interessante estudar a relação dele com alguma medida de dependência, como o coeficiente de concordância τ de Kendall, que permite quantificar mais claramente a dependência entre os dois eventos de interesse intermediário (não terminal) e terminal.

O coeficiente τ de Kendall é uma medida de dependência global, que mede a diferença entre a probabilidade de um par ser concordante e um par ser discordante. Um par de variáveis é concordante se grandes ou pequenos valores de uma variável tendem a estar associados com grandes ou pequenos valores da outra variável. Em outra palavras, dados os pares observados (x_1, y_1) e (x_2, y_2) de um vetor (X, Y) de variáveis aleatórias contínuas, eles são concordantes se $x_i > x_j$ e $y_i > y_j$ ou se $x_i < x_j$ e $y_i < y_j$, e eles são discordantes se $x_i > x_j$ e $y_i < y_j$ ou se $x_i < x_j$ e $y_i > y_j$. Pode-se ainda verificar que os pares são concordantes se $(x_i - x_j)(y_i - y_j) > 0$ e discordantes se $(x_i - x_j)(y_i - y_j) < 0$. O coeficiente τ de Kendall é definido com base nessa idéia, conforme pode ser observado na definição 2.1.

Definição 2.1 *Sejam (X_1, Y_1) e (X_2, Y_2) duas réplicas independentes de um par (X, Y) de variáveis aleatórias*

quaisquer. O coeficiente τ de Kendall é definido por:

$$\begin{aligned}\tau &= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \\ &= E\{\mathrm{sgn}((\mathrm{X}_1 - \mathrm{X}_2)(\mathrm{Y}_1 - \mathrm{Y}_2))\}.\end{aligned}$$

em que $\mathrm{sgn(x)} = \begin{cases} -1 & \text{se } x < 0 \\ 0 & \text{se } x = 0 \\ 1 & \text{se } x > 0 \end{cases}.$

Com base nesta definição, assumindo X e Y contínuas, vale o resultado

$$\begin{aligned}\tau &= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \\ &= 4E\{S_{XY}(x, y)\} - 1 \\ &= 4\int\int S_{XY}(x, y) f_{XY}(x, y) dxdy - 1,\end{aligned}$$

em que $f_{XY}(x, y)$ denota a função densidade de probabilidade conjunta de X e Y.

Se $S_{XY}(x, y)$ é representada usando a cópula de sobrevivência bivariada, neste caso da família Arquimediana, que é gerada pela função $\phi_\alpha(\cdot)$ como em (2.4), então partindo da definição anterior, tem-se que [1]

$$\begin{aligned}\tau &= 4\int\int S_{XY}(x, y) f_{XY}(x, y) dxdy - 1 \\ \tau &= 4\int_0^1 \frac{\phi_\alpha(u)}{\phi'_\alpha(u)} du + 1. \qquad (2.9)\end{aligned}$$

Assim, o coeficiente de concordância τ de Kendall para as quatro cópulas consideradas neste trabalho é calculado a seguir.

[1] Detalhes da prova podem ser encontrados em Nelsen (2006), página 163.

Família A

Para a Cópula de Clayton, tem-se que o coeficiente τ de Kendall é dado por:

$$\begin{aligned}\tau &= 4\int_0^1 \frac{\frac{1}{\alpha}(u^{-\alpha}-1)}{-\frac{1}{u^{\alpha+1}}}du + 1 \\ &= -\frac{4}{\alpha}\int_0^1 u - u^{\alpha+1}du + 1 \\ &= \frac{\alpha}{\alpha+2}.\end{aligned}$$

Neste caso, pode-se verificar que o caso limite $\alpha \to 0$ corresponde ao caso de independência (Nelsen, 2006).

Família B

Neste caso, chega-se a

$$\begin{aligned}\tau &= 4\int_0^1 -\frac{(1-u)^{\alpha}}{\alpha(1-u)^{\alpha-1}}du + 1 \\ &= -\frac{4}{\alpha}\int_0^1 (1-u)du + 1 \\ &= \frac{\alpha-2}{\alpha}.\end{aligned}$$

Embora $\tau = 0$ corresponde ao caso $\alpha = 2$, esta situação não corresponde ao caso de independência, o que é uma desvantagem desta cópula.

Família C

Para a Cópula de Frank, tem-se que

$$\begin{aligned}\tau &= 4\int_0^1 \frac{\ln(e^{-\alpha}-1)-\ln(e^{-\alpha u}-1)}{\alpha e^{-\alpha u}(e^{-\alpha u}-1)^{-1}}du + 1 \\ &= 1 - \frac{4}{\alpha^2}\int_0^{\alpha}\left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha}.\end{aligned}$$

Detalhes do calculo de τ encontram-se no apêndice. Para a cópula de Frank, o caso de independência ocorre quando $\alpha \to 0$.

Família D

Por fim, para a cópula de Gumbel, tem-se

$$\begin{aligned}\tau &= 4\int_0^1 -\frac{u(-\ln u)^{\alpha}}{\alpha(\ln u)^{\alpha-1}}du + 1\\ &= -\frac{4}{\alpha}\int_0^1 u(-\ln u)du + 1.\end{aligned}$$

Integrando por partes, tem-se, $\int_0^1 u(-\ln u)du = \dfrac{1}{4}$, logo

$$\tau = -\frac{4}{\alpha}\left(\frac{1}{4}\right) + 1 = \frac{\alpha - 1}{\alpha}.$$

Na cópula de Gumbel, ocorre independência quando $\alpha \to 1$.

Esses resultados mostram claramente que o parâmetro α está associado com a dependência entre os tempos de falha para os eventos intermediário X e terminal Y. Deste modo, torna-se evidente que é de fundamental importância a estimação do parâmetro de dependência α, uma vez que a medida de concordância τ é função deste.

Para estimação de α no contexto de riscos semicompetitivos, é útil ainda considerar uma versão do τ de Kendall condicional (Oakes, 1989) que está associada à função *Cross-Ratio* ou razão cruzada. Para definir o coeficiente de τ de Kendall condicional e a razão cruzada, considere dois pares (X_i, Y_i) e (X_j, Y_j) independentes e identicamente distribuídos, com distribuição dada segundo a função (2.1), para $x, y \in [0, \infty)$. Denotando por $\tilde{X}_{ij} = \min(X_i, X_j)$ e $\tilde{Y}_{ij} = \min(Y_i, Y_j)$, a função *Cross-Ratio*, referida por Oakes (1989), é dada conforme a Definição 2.2.

Definição 2.2 *A função* Cross-ratio *ou razão cruzada é definida como*

$$\begin{aligned}\theta^*_{\alpha}(x, y) &= \frac{P\{(X_i - X_j)(Y_i - Y_j) > 0 | \tilde{X}_{ij} = x; \tilde{Y}_{ij} = y\}}{P\{(X_i - X_j)(Y_i - Y_j) < 0 | \tilde{X}_{ij} = x; \tilde{Y}_{ij} = y\}}\\ &= \frac{S_{XY}(x, y)D_1D_2S(x, y)}{D_1S_{XY}(x, y)D_2S_{XY}(x, y)},\end{aligned} \tag{2.10}$$

em que os operadores D_1 e D_2 representam as derivadas parciais negativas, isto é, $D_1 = -\partial/(\partial x)$ e $D_2 = -\partial/(\partial y)$.

Considerando a segunda expressão da definição 2.2, observe que $\theta^*_\alpha(x, y)$ depende dos tempos até a ocorrência do evento intermediário (não terminal, recaída) X e do evento terminal (óbito) Y, por meio de $S_{XY}(x, y)$. Sendo $S_{XY}(x, y)$ uma subclasse das distribuições Arquimedianas com função geradora $\phi_\alpha(k)$ (Oakes, 1989), então $S_{XY}(x, y)$ definida em (2.1) pode ser escrita segundo (2.4) como:

$$S_{XY}(x, y) = \phi_\alpha^{-1}[\phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}].$$

Dessa forma, segue que $\theta^*_\alpha(x, y)$ pode ser expressa como

$$\begin{aligned}
\theta^*_\alpha(x, y) &= \frac{\phi_\alpha^{-1}[\phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}]D_1D_2\phi_\alpha^{-1}[\phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}]}{D_1\phi_\alpha^{-1}[\phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}]D_2\phi_\alpha^{-1}[\phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}]} \\
&= \frac{\phi_\alpha^{-1}(u)\phi_\alpha^{-1''}(u)}{[\phi_\alpha^{-1'}(u)]^2}, \qquad (2.11)
\end{aligned}$$

em que $u = \phi_\alpha\{S_X(x)\} + \phi_\alpha\{S_Y(y)\}$. Note que a expressão da razão cruzada em (2.11) está escrita em termos da função inversa de $\phi_\alpha(\cdot)$. Denotando, $\phi_\alpha^{-1}(u) = v$ e $\phi_\alpha(v) = u$, a primeira derivada fica dada por:

$$\begin{aligned}
\frac{d}{dv}\phi_\alpha^{-1}(\phi_\alpha(v)) &= \frac{d}{dv}v \\
\phi_\alpha^{-1'}(\phi_\alpha(v))\phi_\alpha^{'}(v) &= 1 \\
\phi_\alpha^{-1'}(u) &= \frac{1}{\phi_\alpha^{'}(v)}.
\end{aligned}$$

A segunda derivada fica dada por:

$$\begin{aligned}
\frac{d^2}{dv^2}\phi_\alpha^{-1'}(\phi_\alpha(v)) = (\phi_\alpha^{-1'}(\phi_\alpha(v)))^{'} &= \left(\frac{1}{\phi_\alpha^{'}(v)}\right)^{'} \\
\phi_\alpha^{-1''}(\phi_\alpha(v))\phi_\alpha^{'}(v) &= -\frac{1}{(\phi_\alpha^{'}(v))^2}\phi_\alpha^{''}(v) \\
\phi_\alpha^{-1''}(u) &= -\frac{\phi_\alpha^{''}(v)}{(\phi_\alpha^{'}(v))^3}.
\end{aligned}$$

Logo, reescrevendo (2.11) em termos das derivadas da função geradora inversa, tem-se

$$\theta^*_\alpha(x,y) = \frac{-v\frac{\phi''_\alpha(v)}{(\phi'_\alpha(v))^3}}{\left(\frac{1}{\phi'_\alpha(v)}\right)^2} = -\frac{v\phi''_\alpha(v)}{\phi'_\alpha(v)} = \theta_\alpha(v), \tag{2.12}$$

com $v = S_{XY}(x,y)$.

Na análise de riscos semicompetitivos, a região de estudo faz sentido quando $x \leq y$, assim a equação anterior só é valida nessa região, sempre que $S_{XY}(x,y)$ puder ser escrita em termos de uma cópula Arquimediana. Para as quatro famílias Arquimedianas descritas na Seção 2.2, a Tabela 2.1 apresenta as respectivas funções de razão cruzada.

Tabela 2.1: *Função de razão cruzada*

Cópula	Razão cruzada
Família A - Cópula de Clayton	$\theta_\alpha(k) = \alpha + 1$
Família B	$\theta_\alpha(k) = \frac{k(\alpha-1)}{1-k}$
Família C - Cópula de Frank	$\theta_\alpha(k) = \frac{\alpha k}{1-e^{-\alpha k}}$
Família D - Cópula de Gumbel	$\theta_\alpha(k) = 1 - \frac{(\alpha-1)}{\ln k}$

A existência de censura à direita, no entanto, torna necessária uma versão modificada da função de razão cruzada. Quando há censura, dois pares (X_i, Y_i) e (X_j, Y_j) só podem ser comparados quando $\tilde{X}_{ij} \leq \tilde{Y}_{ij} \leq \tilde{B}_{ij}$, em que $\tilde{X}_{ij} = \min(X_i, X_j)$, $\tilde{Y}_{ij} = \min(Y_i, Y_j)$ e $\tilde{B}_{ij} = \min(B_i, B_j)$. O evento $\tilde{X}_{ij} \leq \tilde{Y}_{ij} \leq \tilde{B}_{ij}$ é denotado por A_{ij}. Assim, uma pequena modificação na razão cruzada, dada na Definição 2.2, fornece uma expressão do coeficiente τ de Kendall condicional, apresentado na Definição 2.3.

Definição 2.3 *O τ de Kendall condicional é definido como*

$$\tau_\alpha = E\{\mathrm{sgn}[(\mathrm{X_i} - \mathrm{X_j})(\mathrm{Y_i} - \mathrm{Y_j})]|\mathrm{A_{ij}}\}.$$

O τ de Kendall condicional está relacionado à função de razão cruzada. Para encontrar essa relação, observe que

$$E\{\mathrm{sgn}[(X_i - X_j)(Y_i - Y_j)]|\tilde{X}_{ij} = x, \tilde{Y}_{ij} = y\} = \frac{\theta(x, y) - 1}{\theta(x, y) + 1}. \tag{2.13}$$

Além disso, usando propriedades de esperança condicional, vem que

$$\begin{aligned} \tau_\alpha &= E\{\mathrm{sgn}[(X_i - X_j)(Y_i - Y_j)]|A_{ij}\} \\ &= E\left[E\{\mathrm{sgn}[(X_i - X_j)(Y_i - Y_j)]|A_{ij} \cap (\tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)\}|A_{ij}\right]. \end{aligned}$$

Conhecidos os valores $\tilde{X}_{ij} = x$, e $\tilde{Y}_{ij} = y$, como se trabalha na região $x \leq y$, então A_{ij} necessariamente ocorre (i.e., os pares são comparáveis) e $A_{ij} \cap (\tilde{X}_{ij} = x, \tilde{Y}_{ij} = y) = (\tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)$, de forma que

$$\tau_\alpha = E[E\{sgn[(X_i - X_j)(Y_i - Y_j)]|\tilde{X}_{ij} = x; \tilde{Y}_{ij} = y\}|A_{ij}].$$

Por (2.13), chega-se a

$$\tau_\alpha = E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})\} - 1}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})\} + 1}\middle| A_{ij}\right\}. \tag{2.14}$$

Deste modo o τ_α de Kendall condicional fica expresso como um valor esperado condicionado na área de interesse em que $\tilde{X}_{ij} \leq \tilde{Y}_{ij} \leq \tilde{B}_{ij}$, e fornece a base para a estimação do parâmetro α, que é descrita a seguir.

2.4 Estimação

Como já mencionado, tem-se interesse em estimar as funções de sobrevivência marginais, a conjunta e algumas condicionais nas quais o parâmetro α sempre está envolvido. Dessa forma, torna-se fundamental estimar inicialmente este parâmetro. Neste trabalho, será considerado o método de estimação proposto por Lakhal e Abdous (2008). Após a estimação de α, são descritos métodos de estimação das funções de sobrevivência de interesse.

2.4.1 Parâmetro de dependência da cópula α

Para estimar α a ideia é obter uma equação de estimação. Para motivar essa equação, os autores partem da expressão do τ_α de Kendall condicional dada em (2.14) e chegam a

$$\begin{aligned}\tau_\alpha &= E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}-\frac{1}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\middle|A_{ij}\right\}\\ &= E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\middle|A_{ij}\right\}-E\left\{\frac{1}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}|A_{ij}\right\}. \qquad (2.15)\end{aligned}$$

Dada à definição da razão cruzada em termos de probabilidades de concordância e discordância condicionada (segundo a Definição 2.2), então o primeiro termo de (2.15) fica dado por

$$E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\middle|A_{ij}\right\}=E\left\{\frac{\frac{P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}{P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}}{\frac{P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}+P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}{P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}}\middle|A_{ij}\right\}.$$

Sendo que $P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}+P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}=1$, tem-se

$$\begin{aligned}E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\middle|A_{ij}\right\} &= E\left\{\frac{\frac{P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}{P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}}{\frac{1}{P\{(X_i-X_j)(Y_i-Y_j)<0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}}}\middle|A_{ij}\right\}\\ &= E\left\{P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}|A_{ij}\right\},\end{aligned}$$

e, aplicando a propriedade $E(X)=E[E(X|Y)]$ à segunda igualdade e considerando que condicionado a A_{ij}, o evento $(\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y)$ necessariamente ocorre, chega-se a

$$\begin{aligned}E\left\{\frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\middle|A_{ij}\right\} &= E\left[E\left\{P\{(X_i-X_j)(Y_i-Y_j)>0|A_{ij}\cap(\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y)\}\right\}|A_{ij}\right]\\ &= E\left[E\left\{P\{(X_i-X_j)(Y_i-Y_j)>0|\tilde{X}_{ij}=x;\tilde{Y}_{ij}=y\}|A_{ij}\right\}\right]\\ &= E\left[I_{\{P\{(X_i-X_j)(Y_i-Y_j)>0\}}|A_{ij}\right].\end{aligned}$$

Deste modo, pode-se concluir que

$$E\left\{I_{\{P\{(X_i-X_j)(Y_i-Y_j)>0\}} - \frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}|A_{ij}\right\} = 0.$$

Retomando a notação descrita na Seção 2.1, sejam $\tilde{S}_{ij} = \min(S_i, S_j)$ e $\tilde{R}_{ij} = \min(R_i, R_j)$, em que $S_i = \min(\min(X_i, Y_i), B_i)$ e $R_i = \min(Y_i, B_i)$. A equação anterior motiva a seguinte equação de estimação:

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})1_{\{A_{ij}\}} \times \left\{I_{\{(X_i-X_j)(Y_i-Y_j)>0\}} - \frac{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\right\} = 0, \tag{2.16}$$

em que $w(\cdot,\cdot)$ é uma função de pesos aleatórios associada ao par (i, j), que converge para uma função determinística. Fine et al. (2001) sugerem o uso dos seguintes pesos:

$$w_{a,b}^{-1}(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{S_i\geq\min(u,a),R_i\geq\min(b,v)\}},$$

com a e b constantes positivas.

Observe que a equação de estimação (2.16) depende da função de sobrevivência bivariada $S_{XY}(x, y)$, que é desconhecida. Os autores solucionam este problema sugerindo substituir $S_{XY}(x, y)$ pelo estimador não paramétrico $\tilde{S}_{XY}(\cdot,\cdot)$ baseado na técnica IPCW (Inverse Probability of Censoring Weight) proposto por Lin e Ying (1993), dado por

$$\tilde{S}_{XY}(x,y) = \frac{\sum_{i=1}^{n} I_{\{S_i>x,R_i>y\}}}{\hat{G}(y)} \qquad 0 \leq x \leq y; \tag{2.17}$$

em que $\hat{G}(\cdot)$ é o estimador Kaplan-Meier (K-M) da função de sobrevivência da censura B.

Assim, a partir da expressão (2.16) é obtida uma equação para a estimação do parâmetro α da cópula. No caso da cópula da família A (Cópula de Clayton), pode-se encontrar uma expressão analítica para $\hat{\alpha}$ dado que a função de razão cruzada (*Cross-Ratio*) é constante. No entanto, para as demais cópulas, é preciso utilizar um método numérico para a obtenção da estimativa de α.

Para facilitar a obtenção das equação de estimação das cópulas estudadas, considere $D_{ij} = 1_{\{A_{ij}\}}$ e $\Delta_{ij} = I_{\{(X_i-X_j)(Y_i-Y_j)>0\}}$. Então, a partir de (2.16) e das funções de razão cruzada (*Cross-Ratio*)

obtidas para cada cópula, a seguir é descrita a obtenção das equações de estimação para cada cópula considerada.

- **Família A** - Cópula de Clayton:

  A partir da função *Cross-Ratio* desta cópula dada na Tabela 2.1, tem-se

$$\begin{aligned}
\frac{1}{\binom{n}{2}} \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \times \left\{ \Delta_{ij} - \frac{\alpha+1}{\alpha+2} \right\} &= 0 \\
\alpha \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \Delta_{ij} + 2 \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \Delta_{ij} - \alpha \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} - \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} &= 0.
\end{aligned}$$

  Portanto, é claro que

$$-\alpha \left[\sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} - \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \Delta_{ij} \right] = -\sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} (2\Delta_{ij} - 1).$$

  Logo, chega-se à expressão

$$\hat{\alpha} = \frac{\sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} (2\Delta_{ij} - 1)}{\sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} (1 - \Delta_{ij})},$$

  que é o mesmo estimador proposto por Fine et al. (2001).

- **Família B**:

  A partir da função *Cross-Ratio* desta cópula dada na Tabela 2.1, tem-se

$$\frac{1}{\binom{n}{2}} \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \times \left\{ \Delta_{ij} - \frac{\frac{\tilde{S}_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})(\alpha-1)}{1-\tilde{S}_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})}}{\frac{\tilde{S}_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})(\alpha-1)}{1-\tilde{S}_{XY}(\tilde{X}_{ij}, \tilde{Y}_{ij})} + 1} \right\} = 0$$

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})(\alpha-1)}{\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})(\alpha-1)+1-\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})}\right\} = 0$$

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})(\alpha-1)}{\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})(\alpha-2)+1}\right\} = 0.$$

Esta última expressão é a equação de estimação para α na cópula da família B.

- **Família C** - Cópula de Frank:

  A partir da função *Cross-Ratio* desta cópula apresentada na Tabela 2.1, chega-se a

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{\frac{\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})}{1-\exp(-\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))}}{\frac{\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})}{1-\exp(-\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))}+1}\right\} = 0$$

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij})}{\alpha\tilde{S}(\tilde{X}_{ij},\tilde{Y}_{ij})+1-\exp(-\alpha\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))}\right\} = 0$$

  que corresponde à equação de estimação para α na cópula de Frank (família C).

- **Família D** - Cópula de Gumbel:

  Por fim, a partir da função *Cross-Ratio* desta cópula na Tabela 2.1, tem-se

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{1-\frac{(\alpha-1)}{\ln(\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))}}{2-\frac{(\alpha-1)}{\ln(\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))}}\right\} = 0$$

$$\frac{1}{\binom{n}{2}}\sum_{i<j} w(\tilde{S}_{ij},\tilde{R}_{ij})D_{ij}\times\left\{\Delta_{ij}-\frac{\ln(\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))-(\alpha-1)}{2\ln(\tilde{S}_{XY}(\tilde{X}_{ij},\tilde{Y}_{ij}))-(\alpha-1)}\right\} = 0.$$

  Esta expressão corresponde a equação de estimação para a cópula de Gumbel (família D).

Os autores sugerem o uso do jackknife para se obter uma estimativa do erro padrão do estimador.

2.4.2 Função de Sobrevivência Marginal $S_X(x)$

Dada a possível existência de uma estrutura de dependência entre os eventos de intermediário e terminal (recaída e óbito), os estimadores habituais para a função de sobrevivência, como Kaplan-Meier, não podem ser utilizados para a função de sobrevivência marginal $S_X(x)$. Porém, alguns autores propõem alternativas, discutidas a seguir.

- Estimador não paramétrico (Fine et al., 2001)

  A função de sobrevivência marginal é definida como

  $$\hat{S}_X^F(t) = \phi_{\hat{a}}^{-1}\{\phi_{\hat{a}}[\hat{S}_Z(t)] - \phi_{\hat{a}}[\hat{S}_Y(t)]\},$$

  em que $\hat{S}_Z$ e $\hat{S}_Y$ correspondem aos estimadores de Kaplan-Meier das funções de sobrevivência de $Z = \min(Y, X)$ e Y, respectivamente. A maior dificuldade deste estimador está na estimação da cauda de S_X quando há alta dependência com Y.

- Estimador autoconsistente (Jiang et al., 2005)

  O estimador autoconsistente para $S_X(x)$ é obtido por meio de uma equação de estimação. Cada indivíduo contribui para essa equação de estimação de $S_X(x)$ com a probabilidade de que ele não tenha recaída no instante x, dada a história até x. Contudo, a contribuição das observações censuradas envolvem termos que incluem $S_X(x)$, de forma que não é possível a obtenção de uma expressão analítica e métodos numéricos devem ser empregados.

  Jiang et al. (2005) mostram que a solução é uma função de passo crescente que salta nos valores observados do evento intermediário (X).

- Estimador *Cópula-Gráfico*

  Este estimador foi introduzido por Zheng e Klein (1982) para estimar a função de sobrevivência sob censura dependente. Os autores assumiram uma cópula conhecida para a distribuição conjunta do tempo de falha e censura, e derivaram uma equação de estimação para as funções de sobrevivência marginais. Mais tarde, Rivest e Wells (2001) adaptaram o estimador para cópulas da família Arquimediana para dados com estrutura de riscos semicompetitivos e estudaram sua distribuição assintótica usando teoria de processos de contagem.

Segundo Lakhal e Abdous (2008), dentre estas três propostas de estimação para $S_X(x)$, o estimador Cópula-Gráfico apresentou o melhor desempenho e, portanto, é o estimador usado no desenvolvimento deste trabalho e é apresentado a seguir com maior detalhamento.

Considere

$$\hat{S}_Z(t) = \phi_{\hat{a}}^{-1}[\phi_{\hat{a}}\{\hat{S}_X(t)\} + \phi_{\hat{a}}\{\hat{S}_Y(t)\}].$$

Se t é o tempo de falha observado para o evento intermediário (X), então

$$\begin{aligned}
\phi_{\hat{a}}\{\hat{S}_Z(t)\} &= \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\} + \phi_{\hat{a}}\{\hat{S}_Y(t)\} \\
\phi_{\hat{a}}\{\hat{S}_Z(t)\} - \phi_{\hat{a}}\{\hat{S}_Z(t^-)\} &= \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(t^-)\} + \phi_{\hat{a}}\{\hat{S}_Y(t)\} - \phi_{\hat{a}}\{\hat{S}_Y(t^-)\},
\end{aligned}$$

sendo $\hat{S}_X^{CG}(t)$ o estimador Cópula-Gráfico. Como X não é censurado por Y no tempo t e como $Z = \min(X, Y)$, então este tempo t é também um tempo de falha observado para Z. E é um ponto de descontinuidade para $\hat{S}_Z(.)$. De outro lado, t não pode ser tempo de falha para Y por continuidade e daí $\hat{S}_Y(.)$ não salta no tempo t.

Logo, pode-se concluir que

$$\phi_{\hat{a}}\{\hat{S}_Z(t)\} - \phi_{\hat{a}}\{\hat{S}_Z(t^-)\} = \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(t^-)\}.$$

Somando estes termos sob todos os tempos de falha observados antes de t, tem-se

$$\begin{aligned}
\sum_{S_i \le t; \delta_{xzi}=1} \phi_{\hat{a}}\{\hat{S}_Z(S_i)\} - \phi_{\hat{a}}\{\hat{S}_Z(S_i^-)\} &= \sum_{S_i \le t; \delta_{xzi}=1} \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_i)\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_i^-)\} \\
&= \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_{k_t})\} + \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_{k_t})\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_{k_t-1})\} + \\
&\quad \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_{k_t-1})\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_{k_t-2})\} + \cdots + \\
&\quad \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_2)\} - \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_1)\} + \phi_{\hat{a}}\{\hat{S}_X^{CG}(S_1)\} \\
&= \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\},
\end{aligned}$$

em que k_t é o número de falhas de x ocorridas em (0,t].

Deste modo, a seguinte expressão é obtida,

$$\sum_{S_i \le t; \delta_{xzi}=1} \phi_{\hat{a}}\{\hat{S}_Z(S_i)\} - \phi_{\hat{a}}\{\hat{S}_Z(S_i^-)\} = \phi_{\hat{a}}\{\hat{S}_X^{CG}(t)\}.$$

Assim, o estimador Cópula-Gráfico geral é dado por

$$\hat{S}_X^{CG}(t) = \phi_{\hat{a}}^{-1}\left[\sum_{S_i \leq t; \delta_{xzi}=1} \phi_{\hat{a}}\{\hat{S}_Z(S_i)\} - \phi_{\hat{a}}\{\hat{S}_Z(S_i^-)\}\right]. \tag{2.18}$$

Este estimador foi calculado para as cópulas em estudo e depende da função geradora inversa definida na Seção 2.2, como segue:

- **Família A** - Cópula de Clayton:

$$\begin{aligned}\hat{S}_X^{CG}(t) &= \left(\hat{\alpha}\left\{\sum_{S_i \leq t; \delta_{xzi}=1} \frac{1}{\hat{\alpha}}\left[\{\hat{S}_Z(S_i)^{-\hat{\alpha}}+1\} - \{\hat{S}_Z(S_i^-)^{-\hat{\alpha}}+1\}\right]\right\}+1\right)^{-1/\hat{\alpha}} \\ &= \left(\left\{\sum_{S_i \leq t; \delta_{xzi}=1} \{\hat{S}_Z(S_i)^{-\hat{\alpha}}\} - \{\hat{S}_Z(S_i^-)^{-\hat{\alpha}}\}\right\}+1\right)^{-1/\hat{\alpha}}. \end{aligned} \tag{2.19}$$

- **Família B:**

$$\hat{S}_X^{CG}(t) = 1 - \left\{\sum_{S_i \leq t; \delta_{xzi}=1} [\{1-\hat{S}_Z(S_i)\}^{\hat{\alpha}} - \{1-\hat{S}_Z(S_i^-)\}^{\hat{\alpha}}]\right\}^{1/\hat{\alpha}}. \tag{2.20}$$

- **Família C** - Cópula de Frank:

$$\begin{aligned}\hat{S}_X^{CG}(t) &= -\frac{1}{\hat{\alpha}}\ln\left[1+(e^{-\hat{\alpha}}-1)\exp\left\{\sum_{S_i \leq t; \delta_{xzi}=1} \ln\left(\frac{e^{-\hat{\alpha}\{\hat{S}_Z(S_i)\}}-1}{e^{-\hat{\alpha}}-1}\right) - \ln\left(\frac{e^{-\hat{\alpha}\{\hat{S}_Z(S_i^-)\}}-1}{e^{-\hat{\alpha}}-1}\right)\right\}\right] \\ &= -\frac{1}{\hat{\alpha}}\ln\left[1+(e^{-\hat{\alpha}}-1)\exp\left\{\sum_{S_i \leq t; \delta_{xzi}=1} \ln\left(\frac{e^{-\hat{\alpha}\{\hat{S}_Z(S_i)\}}-1}{e^{-\hat{\alpha}\{\hat{S}_Z(S_i^-)\}}-1}\right)\right\}\right]. \end{aligned} \tag{2.21}$$

- **Família D:** Cópula de Gumbel,

$$\hat{S}_X^{CG}(t) = \exp\left\{-\left[\sum_{S_i \leq t; \delta_{xzi}=1} (-\ln\{\hat{S}_Z(S_i)\})^{\hat{\alpha}} - (-\ln\{\hat{S}_Z(S_i^-)\})^{\hat{\alpha}}\right]^{1/\hat{\alpha}}\right\}. \tag{2.22}$$

Depois de se obter um estimador de $S_X(x)$, é possível construir uma função de sobrevivência conjunta semiparamétrica (devido ao parâmetro de dependência da cópula) como em (2.2) dos

eventos intermediário e terminal.

2.4.3 Função de Sobrevivência bivariada

Na estimação da função de sobrevivência bivariada, há duas alternativas: um estimador não paramétrico definido por Lin e Ying (1993), dado na expressão (2.17), que é usado na equação de estimação para α, e o estimador paramétrico que depende do estimador das funções de sobrevivência marginais dos eventos intermediário, X, e terminal, Y.

Estimador não paramétrico

Com as variáveis anteriormente definidas na Seção 2.1, note que é possível obter as funções de sobrevivência estimadas Kaplan-Meier do evento terminal, Y, e de $Z = \min(X, Y)$, baseadas respectivamente em (R_i, δ_{yi}) e (S_i, δ_{zi}) dado que elas estão sujeitas a censura independente. No caso da censura C, também é possível obter seu estimador K-M da função de sobrevivência, denotado por $\hat{G}(.)$, baseado em $(R_i, 1 - \delta_{yi})$.

Assim, com estas definições, Lin e Ying (1993) obtiveram o estimador de $S(x, y)$ na região $x \leq y$, como visto na estimação de α, seção 2.4.1, dado na expressão (2.17).

Estimador semiparamétrico

O estimador semiparamétrico é baseado na função de sobrevivência bivariada dada pela cópula Arquimediana em (2.4), sendo que é preciso conhecer as estimativas das funções de sobrevivência marginais dos eventos intermediário (recaída) e terminal (óbito), dadas por $\hat{S}_X^{CG}(x)$ e $\hat{S}_Y^{KM}(x)$, respectivamente. Assim a função de sobrevivência conjunta pode ser estimada como

$$\hat{S}_{XY}(x, y) = \hat{C}_\alpha\{\hat{S}_X(x), \hat{S}_Y(y)\}, \qquad 0 \leq x \leq y. \tag{2.23}$$

Para as cópulas em estudo, o estimador para a função de sobrevivência conjunta bivariada é:

- **Família A** - Cópula de Clayton:

$$\hat{S}_{XY}(x, y) = (\hat{S}_X^{CG}(x)^{-\hat{\alpha}} + \hat{S}_Y^{KM}(y)^{-\hat{\alpha}} - 1)^{-1/\hat{\alpha}}. \tag{2.24}$$

- **Família B:**

$$\hat{S}_{XY}(x, y) = 1 - [(1 - \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (1 - \hat{S}_Y^{KM}(y))^{\hat{\alpha}}]^{1/\hat{\alpha}}. \tag{2.25}$$

- **Família C** - Cópula de Frank:

$$\hat{S}_{XY}(x,y) = -\frac{1}{\hat{\alpha}} \ln\left(1 + \frac{(e^{-\hat{\alpha}\hat{S}_X^{CG}(x)} - 1)(e^{-\hat{\alpha}\hat{S}_Y^{KM}(y)} - 1)}{e^{-\hat{\alpha}} - 1}\right). \tag{2.26}$$

- **Família D** - Cópula de Gumbel:

$$\hat{S}_{XY}(x,y) = \exp\{-[(-\ln \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (-\ln \hat{S}_Y^{KM}(y))^{\hat{\alpha}}]^{1/\hat{\alpha}}\}. \tag{2.27}$$

Além de estimar a função de sobrevivência conjunta, pode haver interesse em determinar outras quantidades, como as descritas nas seções seguintes.

2.4.4 Função de sobrevivência condicional quando não há recaída

Um estimador proposto por Lakhal e Abdous (2008) para a probabilidade de sobrevivência de um paciente com uma certa doença e que não teve recaída no tempo x é

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{\tilde{S}(x,y)}{\tilde{S}(x,x)}.$$

O estimador desta função pode ser obtido a partir do estimador não paramétrico (2.17). Porém, também pode ser usado um estimador semiparamétrico:

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{C_\alpha\{\hat{S}_X(x), \hat{S}_Y(y)\}}{C_\alpha\{\hat{S}_X(x), \hat{S}_Y(x)\}}. \tag{2.28}$$

Para as cópulas em estudo, estimadores para a probabilidade condicional (2.28) são descritos a seguir.

- **Família A** - Cópula de Clayton:

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{(\hat{S}_X^{CG}(x)^{-\hat{\alpha}} + \hat{S}_Y^{KM}(y)^{-\hat{\alpha}} - 1)^{-1/\hat{\alpha}}}{(\hat{S}_X^{CG}(x)^{-\hat{\alpha}} + \hat{S}_Y^{KM}(x)^{-\hat{\alpha}} - 1)^{-1/\hat{\alpha}}}.$$

- **Família B:**

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{1 - [(1 - \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (1 - \hat{S}_Y^{KM}(y))^{\hat{\alpha}}]^{1/\hat{\alpha}}}{1 - [(1 - \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (1 - \hat{S}_Y^{KM}(x))^{\hat{\alpha}}]^{1/\hat{\alpha}}}.$$

- **Família C** - Cópula de Frank:

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{\ln\left(1 + \frac{(e^{-\hat{\alpha}\hat{S}_X^{CG}(x)} - 1)(e^{-\hat{\alpha}\hat{S}_Y^{KM}(y)} - 1)}{e^{-\hat{\alpha}} - 1}\right)}{\ln\left(1 + \frac{(e^{-\hat{\alpha}\hat{S}_X^{CG}(x)} - 1)(e^{-\hat{\alpha}\hat{S}_Y^{KM}(x)} - 1)}{e^{-\hat{\alpha}} - 1}\right)}.$$

- **Família D** - Cópula de Gumbel:

$$\tilde{P}(Y > y|X > x; Y > x) = \frac{\exp\{-[(-\ln \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (-\ln \hat{S}_Y^{KM}(y))^{\hat{\alpha}}]^{1/\hat{\alpha}}\}}{\exp\{-[(-\ln \hat{S}_X^{CG}(x))^{\hat{\alpha}} + (-\ln \hat{S}_Y^{KM}(x))^{\hat{\alpha}}]^{1/\hat{\alpha}}\}}.$$

2.4.5 Função de sobrevivência condicional quando há recaída

Também pode ser de interesse determinar a probabilidade de sobrevivência de um paciente com uma certa doença que teve recaída em x, mas segue vivo no tempo $t > x$. Segundo Lakhal e Abdous (2008), um estimador é

$$\tilde{P}(Y > y|X = x; Y > t) = \frac{\phi'_\alpha(C_\alpha\{\hat{S}_X(x), \hat{S}_Y(t)\})}{\phi'_\alpha(C_\alpha\{\hat{S}_X(x), \hat{S}_Y(y)\})}.$$

Para as cópulas em estudo este estimador é apresentado a seguir.

- **Família A** - Cópula de Clayton:

$$\tilde{P}(Y > y|X = x; Y > t) = \frac{C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\}^{-\hat{\alpha}-1}}{C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\}^{-\hat{\alpha}-1}},$$

  com $C_{\hat{\alpha}}\{\hat{S}_X^{CG}(u), \hat{S}_Y^{KM}(v)\}$ sendo a função de sobrevivência bivariada segundo a cópula de Clayton dada pela expressão (2.24).

- **Família B:**

$$\tilde{P}(Y > y|X = x; Y > t) = \frac{(1 - C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\})^{\hat{\alpha}-1}}{1 - (C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\})^{\hat{\alpha}-1}},$$

  com $C_{\hat{\alpha}}\{\hat{S}_X^{CG}(u), \hat{S}_Y^{KM}(v)\}$ sendo a função de sobrevivência bivariada segundo a cópula da

família B dada pela expressão (2.25).

- **Família C** - Cópula de Frank:

$$\tilde{P}(Y > y|X > x; Y > t) = \frac{\hat{\alpha}(\exp\{-\hat{\alpha}C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\}\} - 1)\exp\{-\hat{\alpha}C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\}\}}{\hat{\alpha}(\exp\{-\hat{\alpha}C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\}\} - 1)\exp\{-\hat{\alpha}C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\}\}},$$

  com $C_{\hat{\alpha}}\{\hat{S}_X^{CG}(u), \hat{S}_Y^{KM}(v)\}$ sendo a função de sobrevivência bivariada segundo a cópula de Frank dada pela expressão (2.26).

- **Família D** - Cópula de Gumbel:

$$\tilde{P}(Y > y|X > x; Y > t) = \frac{C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\}(-\ln C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\})^{\hat{\alpha}-1}}{C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(t)\}(-\ln C_{\hat{\alpha}}\{\hat{S}_X^{CG}(x), \hat{S}_Y^{KM}(y)\})^{\hat{\alpha}-1}},$$

  com $C_{\hat{\alpha}}\{\hat{S}_X^{CG}(u), \hat{S}_Y^{KM}(v)\}$ sendo a função de sobrevivência bivariada segundo a cópula de Gumbel dada pela expressão (2.27).

Em suma, neste capítulo foram descritos estimadores para a função de sobrevivência marginal de X, função de sobrevivência bivariada de (X, Y) e funções de sobrevivência condicionais de Y (dado que houve ou não recaída). Apenas relembrando, a função de sobrevivência marginal de Y pode ser estimada via Kaplan-Meier.

Assim, todas as funções de sobrevivência de interesse podem ser estimadas adequadamente considerando-se a estrutura de riscos semicompetitivos dos dados, e a análise descritiva de dados com essas características pode ser realizada de forma apropriada.

CAPÍTULO 3

Processo Doença - Morte

Em estudos aplicados de análise de sobrevivência, em geral é de interesse incorporar o possível efeito de covariáveis no risco de indivíduos apresentarem o evento considerado. A abordagem apresentada no capítulo anterior, baseada em cópulas, apesar de bastante útil em uma análise descritiva dos dados em estudo, não permite a inclusão de covariáveis. Neste capítulo, será considerada a inclusão de covariáveis em um modelo para dados com estrutura de riscos semi-competitivos.

Mais precisamente, será considerada a abordagem proposta por Xu et al. (2010). Os autores defendem a ideia de que dados com estrutura de riscos semicompetitivos podem ser descritos por um processo de três estados, que nada mais é do que o processo de doença-morte já bastante discutido na literatura (Fix e Neyman, 1951). O estado ocupado pelo indivíduo no tempo $t = 0$ é chamado *em estudo* (estado 0); o paciente permanece neste estado até que o evento intermediário ou o óbito ocorra. Quando o evento intermediário (não terminal) ocorre, o estado é chamado *recaída* (estado 1) e o *óbito* é o estado final ou absorvente (estado 2). Note que no processo não há transição do estado intermediário ao estado inicial.

Segundo a formulação deste processo, apresentada na Figura 3.1, é claro que existem várias aplicações reais que se enquadram nesta estrutura. Por exemplo, considere um estudo de tratamento de câncer, em que o estado inicial é *não resposta ao tratamento*, o estado 1 corresponde

a *resposta ao tratamento* e o estado absorvente é *recaída ou óbito*. Outros exemplos são as duas aplicações consideradas neste trabalho, que serão discutidas em detalhes no Capítulo 4.

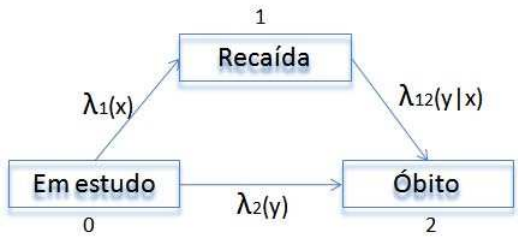

Figura 3.1: *Modelo Doença-Morte para dados com estrutura de risco semi competitivo.*

O modelo Doença-Morte fica completamente especificado pelas funções de intensidade de transição, que descrevem o risco instantâneo de transição entre os três estados, denotadas na Figura 3.1 por $\lambda_1(x)$, $\lambda_2(y)$ e $\lambda_{12}(y|x)$. Uma possível abordagem para analisar processos de doença-morte é focar o interesse nas probabilidades de transição. Se a probabilidade para a transição do estado 1 ao estado 2 não depende da duração no estado 1, então trata-se de um processo de Markov com matriz de intensidade de transição que pode ser representada por

$$\boldsymbol{\lambda}(t) = \begin{bmatrix} -(\lambda_1(t) + \lambda_2(t)) & \lambda_1(t) & \lambda_2(t) \\ 0 & -\lambda_{12}(y|x) & \lambda_{12}(t|x) \\ 0 & 0 & 0 \end{bmatrix}.$$

Desta matriz $\boldsymbol{\lambda}(t)$ podem ser obtidas as probabilidades de transição, sendo que, para um indivíduo no estado 0 (em estudo) no tempo s, denota-se por $P_{01}(s,t)$ a probabilidade de haver transição ao estado 1 (recaída) até um instante t posterior a s. Se no tempo t o indivíduo fica no estado 0, tem-se que $P_{00}(s,t)$ é a probabilidade de não haver transição de estados em (s,t). Similarmente, para um indivíduo com recaída (estado 1) no tempo s denota-se por $P_{11}(s,t)$ a probabilidade de ficar nesse

estado no tempo t. Estas probabilidades são dadas por

$$\begin{aligned} P_{00}(s,t) &= \exp\left\{-\int_s^t [\lambda_1(u)+\lambda_2(u)]du\right\}, \\ P_{11}(s,t) &= \exp\left\{-\int_s^t \lambda_{12}(u)du\right\}, \\ P_{01}(s,t) &= \int_s^t P_{00}(s,u)\lambda_1(u)P_{11}(u,t)du. \end{aligned}$$

Embora esta abordagem em termos de probabilidades de transição do processo de Markov seja de interesse[1], uma outra abordagem de interesse é assumir um modelo que incorpore covariáveis nas intensidades de transição, de maneira semelhante ao modelo semiparamético de Cox (1972), que é a abordagem considerada neste trabalho.

Para incorporar uma possível dependência dos tempos de transição entre os estados, Xu et al. (2010) consideram a inclusão de um termo de fragilidade compartilhada. A variável de fragilidade é um efeito aleatório incorporado às funções de transição de forma multiplicativa, com uma distribuição de probabilidade conhecida, em geral com valor esperado igual a 1 e variância desconhecida. A forma mais simples do processo Doença-Morte é o modelo restrito sem covariáveis, em que não há inclusão de covariáveis e assume-se que a transição de qualquer estado (estado inícial 0 ou intermediário 1) ao estado absorvente (estado 2), óbito, é a mesma, ou seja, $\lambda_2(t) = \lambda_{12}(t)$. É interessante observar que este caso simples, utilizando a distribuição gama para o efeito aleatório, é equivalente a formulação utilizando a cópula de Clayton o qual é mostrado mais adiante.

Conforme já mencionado, a inclusão de covariáveis é também feita de forma multiplicativa nas intensidades de transição, de forma semelhante ao modelo semiparamétrico de Cox. A estimação dos parâmetros do modelo é feita por máxima verossimilhança, sendo que os estimadores possuem boas propriedades assintóticas, discutidas por Xu et al. (2010). A seguir é apresentada em detalhes essa metodologia. Na Seção 3.2 discute-se com detalhes a formulação do modelo, tanto no caso sem quanto no caso com covariáveis. A construção da verossimilhança e estimação dos parâmetros são consideradas na Seção 3.3 e a discussão das proporiedades dos estimadores é apresentada na Seção 3.4.

[1]Detalhes são encontrados em Aalen et al. (2008)

3.1 Notação

A notação utilizada neste capítulo é semelhante à utilizada no capítulo anterior, com algumas pequenas alterações. Assim, conforme já definido anteriormente, será denotado por Y o tempo até a a ocorrência do evento terminal (óbito) e por X o tempo até a ocorrência do evento intermediário (não terminal, recaída). São consideradas, além da censura denotada por B e independente de (X, Y), as seguintes variáveis observáveis:

- $R = \min(Y, B)$;
- $S = \min(X, R) = \min(X, \min(Y, B))$;
- $\delta_x = I_{\{X \leq R\}} = I_{\{X \leq \min(Y,B)\}}$;
- $\delta_y = I_{\{Y \leq B\}}$; e
- Vetor de p covariáveis: $\boldsymbol{z} = [z_1, z_2, \cdots, z_p]^\top$.

Deste modo, o conjunto de dados consiste de n réplicas independentes $\Delta = \{(R_i, S_i, \delta_{x_i}, \delta_{y_i}, \boldsymbol{z}_i), i = 1, \cdots, n\}$ das variáveis observáveis $(R, S, \delta_x, \delta_y, \boldsymbol{z})$.

3.2 Modelo

Segundo o trabalho de Xu et al. (2010), uma possível abordagem da estrutura de dados com riscos semicompetitvos sob processo doença-morte é por meio das funções de intensidade de transição, cuja interpretação intuitiva é a taxa instantânea de risco. A Definição 3.1 apresenta as definições apropriadas para o processo de doença-morte das intensidades de transição. É importante observar na Definição 3.1 que as taxas $\lambda_1(x)$ e $\lambda_2(y)$ não são iguais às funções de taxa de falha (ou risco) usualmente definidas, elas correspondem à função de risco da causa específica utilizada no contexto de riscos competitivos (Klein e Moeschberger, 2003).

Definição 3.1 *As taxas de risco ou funções de transição entre os estados 0 (em estudo), 1 (recaída) e 2*

(óbito) são definidas por

$$\begin{aligned}
\lambda_1(x) &= \lim_{\Delta\to 0} \frac{P(x \le X < x+\Delta | X \ge x, Y \ge x)}{\Delta}, \\
\lambda_2(y) &= \lim_{\Delta\to 0} \frac{P(y \le Y < y+\Delta | X \ge y, Y \ge y)}{\Delta} \; e \\
\lambda_{12}(y|x) &= \lim_{\Delta\to 0} \frac{P(y \le Y < y+\Delta | X = x, Y \ge y)}{\Delta},
\end{aligned}$$

com $0 < x < y$.

Assim como no modelo de cópulas, é importante observar que a área de interesse é o quadrante superior tal que $X \le Y$, conforme pode ser visualizado na Figura 3.2, pois o tempo até chegar ao estado intermediário é necessariamente menor que o tempo até o estado absorvente, óbito. Será considerado que o modelo de probabilidade é absolutamente contínuo, com função de densidade conjunta no quadrante superior $x \le y$. Porém, não é trivial obter uma função densidade de probabilidade absolutamente contínua $f^*_{XY}(x,y)$ no quadrante $x \le y$ incluindo a possibilidade do evento associado ao tempo X não ser observado, com a propriedade $\int_0^\infty \int_x^\infty f^*_{XY}(x,y)dydx = 1$. A alternativa encontrada é considerar uma função densidade de probabilidade absolutamente contínua $f_{XY}(x,y)$ com $\int_0^\infty \int_0^\infty f_{XY}(x,y)dydx = 1$. Como o interesse está na região $x \le y$, será utilizada a densidade $f_{XY}(x,y)$ nessa região apenas. Observe que, neste caso, tem-se $\int_0^\infty \int_x^\infty f_{XY}(x,y)dydx \le 1$. Como existe a possibilidade de transição do estado *em estudo* para o óbito diretamente, X pode ser considerada como igual a infinito, ou seja, será definido $X = \infty$ com função de densidade $f_\infty(y)$, $y > 0$. Essa densidade $f_\infty(y)$ será obtida de tal forma que $\int_0^\infty \int_x^\infty f_{XY}(x,y)dydx + \int_0^\infty f_\infty(y)dy = 1$. Nesta formulação, não está sendo associada qualquer probabilidade positiva ao quadrante $\infty > x > y$, o que é razoável dado o processo considerado.

A estimação dos parâmetros de interesse do processo doença-morte é feita por máxima verossimilhança. Portanto é preciso obter tanto a função de densidade $f_{X,Y}(x,y)$ como a função de sobrevivência $S_{XY}(x,y)$ conjuntas dos tempos até o estado intermediário, recaída (X), e o estado final, óbito (Y). Deste modo, nas próximas seções são obtidas as funções de densidade de probabilidade e de sobrevivência conjunta do modelo de fragilidade compartilhada, considerando-se ambos casos com a inclusão ou não de covariáveis no modelo. Será tratado também o caso particular em que se assume que as duas funções de transição ao estado final são iguais, que será denominado caso restrito. Apesar do interesse principal estar no caso com covariáveis, o caso sem covariável é útil para facilitar a compreensão da situação mais geral.

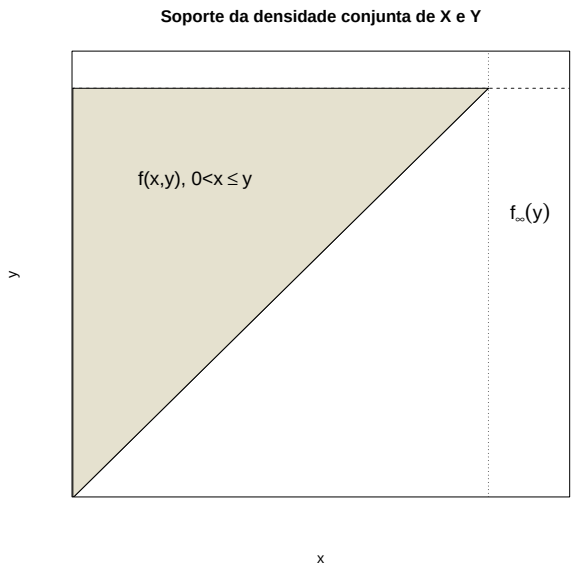

Figura 3.2: *Suporte da Função de densidade conjunta de X e Y.*

3.2.1 Modelo restrito, sem covariáveis

Note na Figura 3.1 que as taxas $\lambda_{12}(y|x)$ e $\lambda_2(y)$ correspondem ao risco de ocorrência do evento terminal dado que o evento intermediário ocorreu ou não, respectivamente. Veja que $\lambda_{12}(y|x) = \lambda_2(y)$ significa que a ocorrência do evento intermediário X não afeta a ocorrência do risco do evento terminal (Y) e caracteriza um modelo Markoviano. A razão $\dfrac{\lambda_{12}(y|x)}{\lambda_2(y)}$ também pode ser usada para caracterizar a dependência entre X e Y: se esta razão é 1, a ocorrência de X é independente da ocorrência de Y. Quando isto acontece, trata-se do modelo restrito.

No entanto, pode ser incluída uma estrutura de dependência entre os estados intermediário e terminal nas funções de intensidade de transição dadas pela definição (3.1). Para tal propósito é usado um efeito aleatório ou fragilidade compartilhada denotada pela variável aleatória $\gamma > 0$. No modelo de fragilidade compartilhada, assume-se que os tempos considerados são variáveis independentes dada a variável de fragilidade.

Será assumido que a variável de fragilidade γ tem distribuição gama com média 1 e variância $\theta > 0$, cuja função de densidade é dada por $f_\Gamma(\gamma) = \dfrac{\theta^{-\frac{1}{\theta}}}{\Gamma(1/\theta)}\gamma^{1/\theta-1}e^{-\gamma/\theta}, \quad \gamma \geq 0.$

Assim, as funções de risco dadas na Definição 3.1 condicionadas à variável de fragilidade compartilhada ficam dadas por:

$$\begin{aligned} \lambda_1(x|\gamma) &= \gamma\lambda_{01}(x), \quad x > 0, \\ \lambda_2(y|\gamma) &= \gamma\lambda_{02}(y), \quad y > 0, \\ \lambda_{12}(y|\gamma, x) &= \gamma\lambda_{03}(y), \quad 0 < x < y. \end{aligned} \tag{3.1}$$

Condicionalmente a γ, esse é um modelo de doença-morte Markoviano usual, o que ocorre também se γ é constante e igual a 1. No caso geral, com $\lambda_{02}(y)$ e $\lambda_{03}(y)$ arbitrários, a dependência entre X e Y é descrita pela razão $\dfrac{\lambda_{03}(y)}{\lambda_{02}(y)}$ e pela fragilidade comum γ. O caso particular em que $\lambda_{03}(y) = \lambda_{02}(y)$ é interessante, pois a dependência entre os eventos intermediário e terminal é completamente explicada pela variável aleatória γ, e deste modo tem-se o modelo restrito em que

$$\lambda_{12}(y|\gamma, x) = \gamma\lambda_{02}(y), \quad 0 < x < y. \tag{3.2}$$

Segundo Xu et al. (2010), a equação (3.2) especifica um tipo de independência condicional entre os eventos intermediário e terminal. Com as funções de risco assim definidas, pode-se obter a distribuição marginal dos tempos envolvidos, que será posteriormente utilizada para a construção da função de verossimilhança. Assim, a seguir são apresentados os passos para a obtenção da função densidade de probabilidade conjunta de X e Y (integrando-se a densidade conjunta de X, Y e γ em relação à última variável), bem como a correspondente função de sobrevivência. De maneira geral, as ideias seguidas para a obtenção das funções de sobrevivência e densidade de probabilidade marginais são as mesmas usadas para qualquer modelo de fragilidade, porém, com pequenos detalhes para incorporar a informação de que a densidade é definida apenas no quadrante $0 < x < y$, com uma densidade correspondente ao caso $X = \infty$, conforme discutido anteriormente.

Nesta situação, a noção de transformada de Laplace de uma variável aleatória Z torna-se extremamente útil. A transformada de Laplace de uma variável aleatória Z é definida como $\mathbb{L}(s) = E[\exp\{-sZ\}]$.

No caso em que $\gamma \sim \Gamma(1/\theta; 1/\theta)$, a transformada de Laplace para a variável γ é dada por

$$\begin{aligned}
\mathbb{L}(s) &= E[\exp\{-s\gamma\}] \\
&= \int_0^\infty \exp\{-s\gamma\} f_\Gamma(\gamma) d\gamma \\
&= (1+\theta s)^{-\frac{1}{\theta}}.
\end{aligned}$$

Na modelagem de fragilidade, a função de sobrevivência bivariada marginal pode ser obtida por (Wienke, 2011):

$$\begin{aligned}
S_{XY}(x,y) &= E[S_{XY}(x,y|\gamma)] \\
&= E[S_X(x|\gamma) S_Y(y|\gamma)] \\
&= E\left[\exp\left\{-\int_0^t \lambda_1(x|\gamma)dx\right\} \exp\left\{-\int_0^t \lambda_2(y|\gamma)dy\right\}\right] \\
&= E[\exp\{-\gamma(\Lambda_{01}(x)+\Lambda_{02}(y))\}] \\
&= \mathbb{L}(\Lambda_{01}(x)+\Lambda_{02}(y)).
\end{aligned}$$

Conclui-se então que

$$S_{XY}(x,y) = [1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{-1/\theta}. \tag{3.3}$$

Para as funções de sobrevivência marginais de X e Y tem-se também que

$$S_X(x) = E[S_X(x|\gamma)] = E[\exp\{-\gamma(\Lambda_{01}(x))\}] = \mathbb{L}(\Lambda_{01}(x)) = \phi_\theta(\Lambda_{01}(x))$$

e

$$S_Y(y) = E[S_Y(y|\gamma)] = E[\exp\{-\gamma(\Lambda_{02}(y))\}] = \mathbb{L}(\Lambda_{02}(y)) = \phi_\theta(\Lambda_{02}(y)),$$

sendo $\mathbb{L}(\cdot) = \phi_\theta(\cdot)$. Observe que $\Lambda_{01}(x) = \phi_\theta^{-1}(S_X(x))$ e $\Lambda_{02}(y) = \phi_\theta^{-1}(S_Y(y))$, com ϕ_θ^{-1} função inversa de ϕ_θ. Então a função de sobrevivência marginal incondicional é dada por:

$$S_{XY}(x,y) = \phi_\theta^{-1}\{\phi_\theta(S_X(x)) + \phi_\theta(S_Y(y))\}.$$

Esse resultado é interessante, pois esta função corresponde à família Arquimediana de cópulas com função geradora $\phi_\theta(\cdot)$ e parâmetro de dependência da cópula θ.

Considerando a distribuição da fragilidade $\Gamma(1/\theta; 1/\theta)$, as funções de sobrevivência marginais

são $S_X(x) = \mathbb{L}(\Lambda_{01}(x)) = (1 + \theta(\Lambda_{01}(x)))^{-1/\theta}$ e $S_Y(y) = \mathbb{L}(\Lambda_{02}(y)) = (1 + \theta(\Lambda_{02}(y)))^{-1/\theta}$. Portanto, a função de sobrevivência dada pela equação (3.3) fica expressa como

$$S_{XY}(x, y) = [(S_X(x))^{-\theta} + (S_Y(y))^{-\theta} - 1]^{-1/\theta}, \tag{3.4}$$

que corresponde à cópula de Clayton discutida no capítulo anterior.

A partir das funções de risco condicionais dada a fragilidade nas equações (3.1) e (3.2), da distribuição da fragilidade e considerando a definição 3.1, o modelo marginal das funções de risco, segundo Wienke (2011) é dado por

$$\begin{aligned} \lambda_1(x) = \lambda_1(x|Y \geq x) &= -\frac{\partial}{\partial x} \ln(S_{XY}(x, y)) \mid_{Y=x} \\ &= -\frac{\partial}{\partial x}\left(-\frac{1}{\theta}\right) \ln[1 + \theta\Lambda_{01}(x) + \theta\Lambda_{02}(y)] \mid_{Y=x} \\ &= \lambda_{01}(x)[1 + \theta\Lambda_{01}(x) + \theta\Lambda_{02}(x)]^{-1}, \qquad x > 0. \end{aligned} \tag{3.5}$$

Analogamente, para $\lambda_2(y)$ obtém-se

$$\begin{aligned} \lambda_2(y) = \lambda_2(y|X \geq y) &= -\frac{\partial}{\partial y} \ln(S_{XY}(x, y)) \mid_{X=y} \\ &= \lambda_{02}(y)[1 + \theta\Lambda_{01}(y) + \theta\Lambda_{02}(y)]^{-1}, \qquad y > 0, \end{aligned} \tag{3.6}$$

e, para $\lambda_{12}(y|x)$ (Wienke, 2011), tem-se

$$\begin{aligned} \lambda_{12}(y|x) = \lambda_{12}(y|X = x) &= -\frac{\partial}{\partial y} \ln\left[-\frac{\partial}{\partial x}(S_{XY}(x, y))\right] \\ &= -\frac{\partial}{\partial y} \ln\left[\left(\frac{1}{\theta}\right)[1 + \theta\Lambda_{01}(y) + \theta\Lambda_{02}(y)]^{-(1/\theta)-1}\theta\lambda_{01}(x)\right] \\ &= (1 + \theta)\lambda_{02}(y)[1 + \theta\Lambda_{01}(y) + \theta\Lambda_{02}(y)]^{-1}, \qquad 0 < x < y. \end{aligned} \tag{3.7}$$

Das funções de risco (3.6) e (3.7), considerando-se o modelo com a restrição $\lambda_2(y) = \lambda_{12}(y)$, pode-se obter a relação

$$\lambda_2(y|X \geq y) = (1 + \theta)\lambda_{12}(y|X = x), \tag{3.8}$$

que define de forma única a função de sobrevivência bivariada, como segue:

$$
\begin{aligned}
\lambda_2(y|X \geq y) &= (1+\theta)\lambda_{12}(y|X=x) \\
-\frac{\partial}{\partial Y}\ln\left(-\frac{\partial}{\partial X}S_{XY}(x,y)\right) &= (1+\theta) - \frac{\partial}{\partial Y}\ln(S_{XY}(x,y)) \\
\int_0^y \frac{\partial}{\partial T}\ln\left(-\frac{\partial}{\partial X}S_{XY}(x,t)\right)dt &= (1+\theta)\int_0^y \frac{\partial}{\partial T}\ln\left(S_{XY}(x,t)\right)dt \\
\ln\left(-\frac{\partial}{\partial X}S_{XY}(x,t)\right) - \ln\left(-\frac{\partial}{\partial X}S_X(x)\right) &= (1+\theta)[\ln(S_{XY}(x,y)) - \ln(S_X(x))] \\
\ln\left(-\frac{\partial}{\partial X}S_{XY}(x,t)\right) - \ln(S_{XY}(x,y))^{(1+\theta)} &= \ln\left(-\frac{\partial}{\partial X}S_X(x)\right) - \ln(S_X(x))^{(1+\theta)} \\
\frac{\frac{\partial}{\partial X}S_{XY}(x,y)}{S_{XY}(x,y)^{1+\theta}} &= \frac{\frac{\partial}{\partial X}S_X(x)}{S_X(x)^{1+\theta}} \\
\int_0^x \frac{\partial}{\partial T}S_{XY}(t,y)^{-\theta}dt &= \int_0^x \frac{\partial}{\partial T}S_X(t)^{-\theta}dt \\
S_{XY}(x,y)^{-\theta} - S_Y(y)^{-\theta} &= S_X(x)^{-\theta} - 1 \\
S_{XY}(x,y) &= [(S_X(x))^{-\theta} + (S_Y(y))^{-\theta} - 1]^{-1/\theta}.
\end{aligned}
$$

Com isto, segundo Wienke (2011), pode-se concluir que no modelo restrito existem duas formas de se derivar a função de sobrevivência bivariada (3.4): uma a partir da relação (3.8), referente à proporcionalidade dos riscos condicionais $\lambda_2(y)$ e $\lambda_{12}(y|x)$, e outra usando a transformada de Laplace da variável de fragilidade.

A partir da função de sobrevivência bivariada é obtida sua correspondente função de densidade de probabilidade, considerando que $f_{XY}(x,y) = \frac{\partial}{\partial X \partial Y}F_{XY}(x,y)$ e $F_{XY}(x,y) = S_{XY}(x,y) - S_{XY}(x,0) - S_{XY}(0,y) + 1 = S_{XY}(x,y) - S_X(x) - S_Y(y) + 1$. Com isso, conclui-se que

$$F_{XY}(x,y) = [1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{-1/\theta} - [1+\theta\Lambda_{01}(x)]^{-1/\theta} - [1+\theta\Lambda_{02}(y)]^{-1/\theta} + 1$$

e

$$\frac{\partial}{\partial X \partial Y}F_{XY}(x,y) = \frac{\partial}{\partial X}\left([1+\theta\Lambda_{02}(y)]^{(-1/\theta)-1}\lambda_{02}(y) - [1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{(-1/\theta)-1}\lambda_{02}(y)\right).$$

Portanto, a função de densidade de probabilidade fica dada por:

$$f_{XY}(x,y) = (1+\theta)\lambda_{01}(x)\lambda_{02}(y)[1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{(-1/\theta)-2}, \qquad 0 < x < y. \tag{3.9}$$

Entretanto, em uma estrutura de risco semicompetitivo em que as observações estão restritas na região $0 \leq x \leq y$, ocorre que

$$\int_0^\infty \int_0^y f_{XY}(x,y)dxdy \leq 1.$$

No entanto, $\int_0^\infty \int_0^\infty f_{XY}(x,y)dxdy = 1$, de forma que pode-se escrever

$$\int_0^\infty \int_0^y f_{XY}(x,y)dxdy + \int_0^\infty \int_y^\infty f_{XY}(x,y)dxdy = 1.$$

Desse modo, $\int_y^\infty f_{XY}(x,y)dx$ pode ser entendida como a função de densidade de (x,y) no caso em que $x = \infty$, que fica dada por

$$\begin{aligned} f_\infty(y) &= \int_y^\infty (1+\theta)[1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{(-1/\theta)-2}\lambda_{01}(x)\lambda_{02}(y)dx \\ &= \lambda_{02}(y)[1+\theta\Lambda_{01}(y)+\theta\Lambda_{02}(y)]^{(-1/\theta)-1}, \qquad y > 0. \end{aligned} \tag{3.10}$$

Como mencionado na seção (3.2) veja que a função (3.10) corresponde a $f_\infty(y)$, $y > 0$, que representa o caso em que é observado o evento final (óbito) mas não o intermediário (recaída). Esta é a situação em que Xu et al. (2010) definem $X_i = \infty$.

Finalmente, com as funções de sobrevivência (3.3) e densidade de probabilidade (3.9) e (3.10) definidas, pode-se construir a função de verossimilhança para estimação dos parâmetros no modelo restrito sem covariáveis, o que será discutido posteriormente.

3.2.2 Modelo geral, com covariáveis

Depois de analisar o modelo restrito em que $\lambda_{02}(y) = \lambda_{03}(y)$, considere agora a situação na qual são incluídas covariáveis z e na qual $\lambda_{02}(y) \neq \lambda_{03}(y)$ no modelo descrito pelas funções de intensidade de transição condicionadas na fragilidade com equações (3.1).

Para a incorporação de covariáveis, podem ser considerados vários modelos de regressão, porém, neste trabalho será especificado o modelo similar a riscos proporcionais (Cox, 1972), de

modo que a modelagem descrita em (3.1) fica dada por:

$$\begin{aligned}
\lambda_1(x|\gamma, z) &= \gamma\lambda_{01}(x)\exp\{\boldsymbol{\beta}_1^\top z\}, \quad x > 0, \\
\lambda_2(y|\gamma, z) &= \gamma\lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\}, \quad y > 0, \\
\lambda_{12}(y|\gamma, x, z) &= \gamma\lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\}, \quad 0 < x < y.
\end{aligned} \tag{3.11}$$

em que $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ e $\boldsymbol{\beta}_3$ são os parâmetros desconhecidos e z é o vetor de covariáveis. Como no modelo restrito, será assumido que γ tem distribuição $\Gamma(1/\theta; 1/\theta)$ e é independente de z.

O modelo (3.11), comparado com o modelo sem covariáveis (3.1), é caracterizado pelos coeficientes $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ e $\boldsymbol{\beta}_3$, que contém informação sobre os efeitos das covariáveis nas funções de intensidade condicionadas à fragilidade. Assim como no caso anterior, a estimação dos parâmetros será feita via máxima verossimilhança e, dessa forma, é necessário obter a função de sobrevivência conjunta de X e Y, bem como a correspondente função densidade de probabilidade. Os passos para a obtenção dessas funções são descritos a seguir.

A partir da relação $f(x) = \lambda(x)S(x) = \lambda(x)\exp\{-\Lambda(x)\}$ e das equações (3.11), tem-se

$$f_{X|\Gamma}(x|\gamma, z) = \gamma\lambda_{01}(x)\exp\{\boldsymbol{\beta}_1^\top z - \gamma\Lambda_{01}(x)\exp\{\boldsymbol{\beta}_1^\top z\}\}.$$

Além disso, tem-se que

$$f_{Y|X,\Gamma}(y|x, \gamma, z) = \lambda^*(y|x, \gamma, z)S^*(y|x, \gamma, z),$$

em que $\lambda^*(y|x, \gamma, z)$ é a correspondente função de risco e $S^*(y|x, \gamma, z)$ é a função de sobrevivência.

Considerando as intensidades de transição dadas em (3.11), observe que, para uma dado x fixado, a intensidade de ocorrência de Y é diferente se $x \leq y$ ou $x > y$ (caso que será associado com a transição direta para o óbito, sem ocorrência do evento intermediário). Assim, pode-se verificar que

$$\lambda^*(y|x, \gamma, z) = \begin{cases} \gamma\lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\}, & x \leq y, \\ \gamma\lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\}, & \text{caso contrário.} \end{cases}$$

Para $S^*(y|x,\gamma,z)$, tem-se

$$S^*(y|x,\gamma,z) = \begin{cases} \exp\left\{-\left(\int_0^x \gamma\lambda_{02}(t)\exp\{\boldsymbol{\beta}_2^\top z\}dt + \int_x^y \gamma\lambda_{03}(t)\exp\{\boldsymbol{\beta}_3^\top z\}dt\right)\right\}, & x \le y, \\ \exp\left\{-\int_0^y \gamma\lambda_{02}(t)\exp\{\boldsymbol{\beta}_2^\top z\}dt\right\}, & \text{caso contrário.} \end{cases}$$

$$= \begin{cases} \exp\left\{-\gamma(\Lambda_{02}(x)\exp\{\boldsymbol{\beta}_2^\top z\} + \Lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\} - \Lambda_{03}(x)\exp\{\boldsymbol{\beta}_3^\top z\})\right\}, & x \le y, \\ \exp\left\{-\gamma\Lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\}\right\}, & \text{caso contrário.} \end{cases}$$

Desse modo, a seguinte expressão pode ser obtida

$$f_{Y|X,\Gamma}(y|x,\gamma,z) = \begin{cases} \gamma\lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z - \gamma(\Lambda_{02}(x)\exp\{\boldsymbol{\beta}_2^\top z\} + \Lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\} - \Lambda_{03}(x)\exp\{\boldsymbol{\beta}_3^\top z\})\} & x \le y, \\ \gamma\lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z - \gamma\Lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\} & \text{caso cont.} \end{cases}$$

Portanto, fixando-se as covariáveis z, com $f_{X|\Gamma}(x|\gamma,z)$ e $f_{Y|X,\Gamma}(y|x,\gamma,z)$ assim definidas e com a função de distribuição da fragilidade γ, obtém-se a função de densidade conjunta de x, y e γ como:

$$f_{X,Y,\Gamma}(x,y,\gamma|z) = f_{Y|X,\Gamma}(y|x,\gamma,z)f_{X|\Gamma}(x|\gamma,z)f_\Gamma(\gamma).$$

Para $x \le y$, vem que

$$f_{X,Y,\Gamma}(x,y,\gamma|z) = \frac{\gamma^{(1/\theta)+1}}{\theta^{1/\theta}\Gamma(1/\theta)}\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{03}(y)e^{\boldsymbol{\beta}_3^\top z}\exp\{-\gamma(1/\theta + \Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z} + \Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top z} + (\Lambda_{03}(y) - \Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top z})\}.$$

Para $x > y$, que será associado com a situação em que o evento intermediário não é observado, chega-se a

$$f_{X,Y,\Gamma}(x,y,\gamma|z) = \frac{\gamma^{(1/\theta)+1}}{\theta^{1/\theta}\Gamma(1/\theta)}\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z}\exp\{-\gamma(1/\theta + \Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z} + \Lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z})\}.$$

Finalmente, depois de alguns cálculos, a função de densidade bivariada no quadrante positivo é dada por $f_{X,Y}(x,y|z) = \int_0^\infty f_{X,Y,\Gamma}(x,y,\gamma|z)d\gamma$

$$= \begin{cases} (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{03}(y)e^{\boldsymbol{\beta}_3^\top z}[1 + \theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z} + \theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top z} + \theta(\Lambda_{03}(y) - \Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top z}]^{-(1/\theta)-2} & x \le y, \\ (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z}[1 + \theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z} + \theta\Lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z}]^{-(1/\theta)-2} & x > y. \end{cases} \tag{3.12}$$

Como no modelo sem covariáveis, neste caso também é necessária a obtenção da função densidade de probabilidade quando só o óbito é observado, ou seja, quando $x = \infty$, isto é

$$\begin{aligned} f_\infty(y|\boldsymbol{z}) &= \int_y^\infty (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}}\lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}}+\theta\Lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}}]^{-(1/\theta)-2}dx \\ &= \lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}}[1+\theta\Lambda_{01}(y)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}}+\theta\Lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}}]^{-(1/\theta)-1}. \end{aligned} \tag{3.13}$$

Assim, a partir de função $f_\infty(y|\boldsymbol{z})$ e de $f_{X,Y}(x,y|\boldsymbol{z})$ na região $x \leq y$ pode-se construir a função de verossimilhança correspondente ao modelo (3.11).

Paralelo ao modelo (3.11), há outra modelagem alternativa semelhante como a descrita em (3.5), (3.6) e (3.7), porém as variáveis podem ser incluídas de forma multiplicativa nas funções de intensidade marginais, ou seja,

$$\begin{aligned} \lambda_1(x|\boldsymbol{z}) &= \lambda_{01}(x)[1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(x)]^{-1}\exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z}\}, \qquad x>0, \\ \lambda_2(y|\boldsymbol{z}) &= \lambda_{02}(y)[1+\theta\Lambda_{01}(y)+\theta\Lambda_{02}(y)]^{-1}\exp\{\boldsymbol{\beta}_2^\top \boldsymbol{z}\}, \qquad y>0, \\ \lambda_{12}(y|x,\boldsymbol{z}) &= (1+\theta)\lambda_{03}(y)[1+\theta\Lambda_{01}(y)+\theta\Lambda_{02}(x)+\theta(\Lambda_{03}(y)-\lambda_{03}(x))]^{-1}\exp\{\boldsymbol{\beta}_3^\top \boldsymbol{z}\}, \quad 0<x<y. \end{aligned}$$

Embora esta metodologia represente uma alternativa para a incorporação de covariáveis no processo doença-morte, o foco neste estudo é a primeira abordagem apresentada. Para a estimação dos parâmetros das funções de intensidade marginal condicionadas é utilizado o método de máxima verossimilhança não paramétrica (as funções de risco basais não são especificadas), discutido nas próximas seções.

3.3 Estimação

Segundo a modelagem definida do processo Doença-Morte, nesta seção tem-se interesse em estimar os parâmetros das funções de intensidade de transição para as três distintas transições dos estados 0 (em estudo), 1 (recaída) e 2 (óbito), tanto para o modelo restrito quanto para o geral. Note que em ambos os modelos é preciso estimar o parâmetro referente à fragilidade (θ) e as funções de risco basais. No caso geral, precisa-se também estimar os coeficientes da regressão.

Para este propósito será considerado o método de estimação de máxima verossimilhança não paramétrica, proposto por Xu et al. (2010), sendo que para a construção da função de verossimilhança é fundamental analisar todas as possíveis situações apresentadas pelos da-

dos observados. O conjunto de dados em estudo consiste de n réplicas independentes $\Delta = \{(R_i, S_i, \delta_{x_i}, \delta_{y_i}, \boldsymbol{z}_i)\}$, $i = 1, \cdots, n$, em que, conforme definido anteriormente, $R = \min(Y, B)$, $S = \min(X, R)$, $\delta_x = I_{\{X \leq R\}}$, $\delta_y = I_{\{Y \leq B\}}$ e $\boldsymbol{z} = [z_1, z_2, \cdots, z_p]^\top$ representa o conjunto de p covariáveis. Relembre ainda que X_i e Y_i são os tempos até a recaída e óbito respectivamente para o i-ésimo indivíduo, e B_i representa o tempo de censura.

Segundo a estrutura dos riscos semicompetitivos, quatro situações podem ser observadas:

- *Caso 1:* São observados os dois eventos de interesse: o tempo até a recaída e o tempo até o óbito (transição do estado 0 ao estado 2 passando pelo estado 1), de modo que nesta situação os dados observados são

$$(S_i = x_i, R_i = y_i, \delta_{x_i} = 1, \delta_{y_i} = 1, \boldsymbol{z}_i).$$

- *Caso 2:* É observado o tempo até o óbito livre de recaída (transição direta do estado 0 ao estado 2). Assim, os dados observados neste caso são:

$$(S_i = y_i, R_i = y_i, \delta_{x_i} = 0, \delta_{y_i} = 1, \boldsymbol{z}_i).$$

- *Caso 3:* É observado somente o tempo até a recaída (transição do estado 0 ao estado 1), mas o tempo até o óbito é censurado. Assim, observa-se:

$$(S_i = x_i, R_i = b_i, \delta_{x_i} = 1, \delta_{y_i} = 0, \boldsymbol{z}_i).$$

- *Caso 4:* A observação é censurada, ou seja, o indivíduo não apresentou nenhum dos eventos de interesse no recorrer do período de observação (não houve transição entre os estados do processo, o indivíduo ficou no estado 0), de forma que observa-se

$$(S_i = b_i, R_i = b_i, \delta_{x_i} = 0, \delta_{yi} = 0, \boldsymbol{z}_i).$$

A seguir, é descrita a obtenção da verosimilhança $L(\boldsymbol{\eta})$, em que $\boldsymbol{\eta}$ denota o vetor de todos os parâmetros a serem estimados. A contribuição da verossimilhança é obtida para cada uma das quatro possíveis situações de falhas/censura descritas anteriormente, em ambos os modelos, restrito e geral.

3.3.1 Modelo restrito, sem covariáveis

Observe que, tanto no modelo restrito como no modelo geral, a função densidade de probabilidade (ou sobrevivência) depende das funções de intensidade de transição basais (e também das funções de intensidade de transição basais acumuladas). Assim como no modelo semiparamétrico de Cox, deseja-se deixar tais funções arbitrárias e não especificar uma forma paramétrica para as mesmas funções. Dessa forma, para a estimação dos parâmetros de interesse será utilizada a mesma ideia proposta por Breslow (1972), que consiste em aproximar a função de intensidade de transição acumulada por uma função escada que dá saltos nos instantes em que as correspondentes transições foram observadas.

Dessa forma, no modelo restrito, em que $\lambda_{02}(y) = \lambda_{03}(y)$, os parâmetros a serem estimados correspondem ao vetor de dimensão $(m+f+1)$, $\boldsymbol{\eta} = (\theta, \boldsymbol{\lambda}_{0R}, \boldsymbol{\lambda}_{0D})$, sendo que $\boldsymbol{\lambda}_{0R} = (\lambda_{011}, \lambda_{012}, \cdots, \lambda_{01m})^\top$ e $\boldsymbol{\lambda}_{0D} = (\lambda_{021}, \lambda_{022}, \cdots, \lambda_{02f})^\top$ são os vetores das funções de risco basais correspondentes aos tempos ordenados de recaída, $t_{r1}, t_{r2}, \cdots, t_{rm}$, e óbito, $t_{d1}, t_{d2}, \cdots, t_{df}$, respectivamente, sendo que $\lambda_{01j} = d\Lambda_{01}(t_{rj})$, $j = 1, \cdots, m$ e $\lambda_{02j} = d\Lambda_{02}(t_{dj})$, $j = 1, \cdots, f$.

Para determinar a função de verossimilhança $L(\boldsymbol{\eta})$, é analisada a contribuição de cada observação para cada um dos casos descritos anteriormente.

- *Caso 1:*

  Como os dois eventos de interesse foram observados, a contribuição para a verossimilhança $L(\boldsymbol{\eta})$ é dada pela função de densidade (3.9), ou seja,

$$f_{XY}(x, y) = (1 + \theta)\lambda_{01}(x)\lambda_{02}(y)[1 + \theta\Lambda_{01}(x) + \theta\Lambda_{02}(y)]^{(-1/\theta)-2}, \qquad 0 < x < y.$$

- *Caso 2:*

  Na situação em que é observado somente o óbito, veja que no processo doença-morte há uma transição direta do estado 0 ao estado 2, de modo que a contribuição desta observação é dada pela função de densidade quando $x = \infty$ como em (3.10), isto é,

$$f_{XY}(x, y) = \lambda_{02}(y)[1 + \theta\Lambda_{01}(y) + \theta\Lambda_{02}(y)]^{(-1/\theta)-1}.$$

- *Caso 3:*

  Se ao fim do estudo somente foi observada a recaída, isso implica que houve censura para

o evento óbito, então tem-se que $S = x$ e $R = b$. Intuitivamente, pode-se pensar que o valor observado para a variável R é maior que um valor y, então, assumindo $R > y$ e utilizando uma notação informal, com abuso de notação para denotar probabilidade associada a variável contínua, a contribuição desta observação seria dada por $P(S = x, R > y)$. Como o tempo até recaída x é sempre menor que um possível tempo de óbito y, então esta probabilidade pode ser interpretada como

$$\begin{aligned} P(X = x, Y > y) &= P(Y > y|X = x)P(X = x), \\ &= S_{Y|X=x}(y)f_X(x), \end{aligned}$$

de modo que a contribuição desta observação é:

$$\begin{aligned} S_{Y|X=x}(y|x)f_X(x) &= \int_y^\infty \frac{f_{XY}(x,t)}{f_X(x)} dt f_X(x) \\ &= \int_y^\infty (1+\theta)\lambda_{01}(x)\lambda_{02}(t)[1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(t)]^{(-1/\theta)-2}dt \\ &= \lambda_{01}(x)[1+\theta\Lambda_{01}(x)+\theta\Lambda_{02}(y)]^{(-1/\theta)-1}. \end{aligned}$$

- *Caso 4:*

  Considerando-se que a observação é censurada, sabe-se que os tempos de recaída e óbito são maiores que o tempo de censura observado. Portanto, a contribuição deste tipo de observação é $P(X > x, Y > y)$, interpretado como a função de sobrevivência dada em (3.3), $S_{XY}(x, y) = [1 + \theta\Lambda_{01}(x) + \theta\Lambda_{02}(y)]^{-1/\theta}$, sendo que os valores observados para X e Y são, $S = R = b$.

Finalmente, a verossimilhança para n observações é construída com os quatro componentes descritos, incorporados na função segundo as variáveis indicadoras, δ_{x_i} e δ_{y_i}. A função de verossimilhança fica, portanto, dada por

$$L(\eta) = \prod_{i=1}^{n}(1+\theta)^{\delta_{x_i}\delta_{y_i}}\lambda_{01}(S_i)^{\delta_{x_i}}\lambda_{02}(R_i)^{\delta_{y_i}}\{1+\theta[\Lambda_{01}(S_i)+\Lambda_{02}(R_i)]\}^{-(1/\theta)-\delta_{x_i}-\delta_{y_i}}. \tag{3.14}$$

Assim, a função log verossimilhança de $L(\boldsymbol{\eta})$, da qual é obtido o vetor escore $U(\boldsymbol{\eta})$, é dada por

$$\sum_{i=1}^{n} \delta_{x_i}\delta_{y_i} \log\{1+\theta\} + \delta_{x_i} \log\{\lambda_{01}(S_i)\} + \delta_{y_i} \log\{\lambda_{02}(R_i)\} - [(1/\theta) + \delta_{x_i} + \delta_{y_i}] \log\{1 + \theta[\Lambda_{01}(S_i) + \Lambda_{02}(R_i)]\}.$$

Seguindo o procedimento usual, da função de verossimilhança $L(\boldsymbol{\eta})$ é obtido o vetor escore $U(\boldsymbol{\eta})$ que corresponde às derivadas da log verossimilhança e os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\eta}$ que maximizam $\log(L(\boldsymbol{\eta}))$, que são encontrados resolvendo-se o sistema de equações $U(\boldsymbol{\eta}) = \mathbf{0}$.

As respectivas componentes do vetor escore correspondem as derivadas parciais com respeito a $\theta, \boldsymbol{\lambda}_{0R}$ e $\boldsymbol{\lambda}_{0D}$, ou seja, $U(\boldsymbol{\eta})$ é o gradiente de $\log(L(\boldsymbol{\eta}))$.

O vetor escore é apresentado para o modelo geral, sendo que para o modelo restrito o cálculo é similar e portanto, omitido.

Xu et al. (2010) mostram através de simulações o bom desempenho da estimação por máxima verossimilhança. No entanto, eles apresentam também outros dois métodos diretos para a estimação do parâmetro associado a dependência θ, que para o caso do modelo restrito corresponde ao parâmetro de dependência na cópula de Clayton (Fine et al., 2001).

Um dos métodos diretos corresponde ao método proposto por Fine et al. (2001), estudado no capítulo anterior, que é baseado na função de razão cruzada (*Cross-Ratio*), em que o parâmetro de dependência (no caso, θ) é solução direta da equação:

$$\frac{1}{\binom{n}{2}} \sum_{i<j} w(\tilde{S}_{ij}, \tilde{R}_{ij}) D_{ij} \times \left\{ \Delta_{ij} - \frac{\theta+1}{\theta+2} \right\} = 0.$$

Já o outro método usa uma equação de estimação baseada em processos de contagem e processos de risco dos eventos intermediário e terminal. Neste trabalho, como o método de máxima verossimilhança se mostrou superior aos demais (Xu et al., 2010), optou-se por não discutir em detalhes os demais métodos propostos. Para uma discussão mais aprofundada desses métodos, pode-se consultar Day et al. (1997).

3.3.2 Modelo geral, com covariáveis

No modelo geral, assim como feito no caso do modelo restrito, as funções de risco acumuladas basais serão aproximadas por funções do tipo escada, com saltos nos instantes de falha. Assim,

os vetores de risco basais correspondentes aos diferentes tempos de recaída $\boldsymbol{\lambda}_{0R}$, óbito $\boldsymbol{\lambda}_{0D}$ e óbito após recaída $\boldsymbol{\lambda}_{0RD}$ têm dimensões que dependem do número de instantes distintos em que algum evento de interesse ocorreu. Deste modo, denote os distintos tempos ordenados de recaída por $t_{r1}, t_{r2}, \cdots, t_{rm}$, os distintos tempos ordenados de óbito sem recaída por $t_{d1}, t_{d2}, \cdots, t_{df}$ e os distintos tempos ordenados de óbito após recaída por $t_{rd1}, t_{rd2}, \cdots, t_{rdg}$. As funções de risco basal para cada um deles, com essa aproximação, ficam dadas por $\lambda_{01j} = d\Lambda_{01}(t_{rj}), j = 1, \cdots, m$, $\lambda_{02j} = d\Lambda_{02}(t_{dj}), j = 1, \cdots, f$ e $\lambda_{03j} = d\Lambda_{03}(t_{dj}), j = 1, \cdots, g$. Assim tem-se que $\boldsymbol{\lambda}_{0R} = (\lambda_{011}, \lambda_{012}, \cdots, \lambda_{01m})^\top$, $\boldsymbol{\lambda}_{0D} = (\lambda_{021}, \lambda_{022}, \cdots, \lambda_{02f})^\top$ e $\boldsymbol{\lambda}_{0RD} = (\lambda_{031}, \lambda_{032}, \cdots, \lambda_{03g})^\top$.

Dada a possibilidade de que um ou vários pacientes apresentem o mesmo tempo observado, defina então os empates para os tempos de recaída, óbito e óbito seguido de recaída como: $d_{t_{rj}}$ número de recaídas ao tempo t_{rj}, $d_{t_{dj}}$ número de óbitos ao tempo t_{dj} e $d_{t_{rdj}}$ número de óbitos seguidos de recaídas ao tempo t_{rdj}.

Com estas configurações o vetor de parâmetros a serem estimados é

$$\boldsymbol{\eta} = (\theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\lambda}_{0R}, \boldsymbol{\lambda}_{0D}, \boldsymbol{\lambda}_{0RD}) = (\theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, d\boldsymbol{\Lambda}_{01}, d\boldsymbol{\Lambda}_{02}, d\boldsymbol{\Lambda}_{03}).$$

Analogamente ao modelo restrito, a construção da verossimilhança $L(\boldsymbol{\eta})$ é feita com as contribuições dos diferentes casos. Lembre que o caso 1 corresponde à observação de ambos os eventos de interesse, o caso 2 se refere à observação do tempo até o óbito sem recaída, no caso 3 se observa apenas o evento intermediário (recaída) e, por fim, no caso 4, nenhum dos eventos é observado.

Uma observação do caso 1 contribui com a função de densidade bivariada $f_{XY}(x, y)$ quando $x < y$, como em (3.12). A contribuição à $L(\boldsymbol{\eta})$ de uma observação do caso 2 (evento terminal, em que $x = \infty$), é $f_\infty(y|\boldsymbol{z})$ dada pela expressão (3.13)

Já para o caso 3, observação da recaída, é preciso calcular $\int_y^\infty f_{XY}(x, t)dt$ (como no modelo sem covariáveis), com $f_{XY}(x, t)$ dada em (3.12), assim

$$\begin{aligned}\int_y^\infty f_{XY}(x,t)dt &= \int_y^\infty (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}}\lambda_{03}(t)e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}} + \theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}} + \theta(\Lambda_{03}(t) - \Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}}]^{-(1/\theta)-2}dt \\ &= \lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}} + \theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}} + \theta(\Lambda_{03}(y) - \Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}}]^{-(1/\theta)-1}, \qquad x \le y,\end{aligned}$$

em que y representa, neste caso, o tempo de censura c (dos dados observados $R = b$).

Por último, para o caso 4, quando nenhum dos eventos é observado, a contribuição é a probabilidade que representa a função de sobrevivência avaliada nos valores de $S = R = b$ do i-ésimo

indivíduo, isto é,

$$P(X > b, Y > b) = S_{X,Y}(b,b) = \int_b^\infty \int_b^x f_{XY}(x,y)dydx + \int_b^\infty \int_x^\infty f_{XY}(x,y)dydx$$

com $f_{XY}(x,y)$ sendo a função de densidade bivariada em todo o quadrante positivo dada em (3.12). Logo, pode-se desenvolver a expressão e chegar a

$$S_{X,Y}(b,b) = \int_b^\infty \int_b^x (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{02}(t)e^{\boldsymbol{\beta}_2^\top z}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(t)e^{\boldsymbol{\beta}_2^\top z}]^{-(1/\theta)-2}dtdx+$$

$$\int_b^\infty \int_x^\infty (1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{03}(t)e^{\boldsymbol{\beta}_3^\top z}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top z}+\theta(\Lambda_{03}(t)-\Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top z}]^{-(1/\theta)-2}dtdx$$

$$= \int_b^\infty \lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(b)e^{\boldsymbol{\beta}_2^\top z}]^{-(1/\theta)-1}dx.$$

Portanto, finalmente chega-se a

$$P(X > b, Y > b) = S_{X,Y}(b,b) = [1+\theta\Lambda_{01}(b)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(b)e^{\boldsymbol{\beta}_2^\top z}]^{-1/\theta}.$$

Em resumo, tem-se que as componentes de cada caso para a construção da função de verossimilhança $L(\boldsymbol{\eta})$ são:

- *Caso 1:*

$$(1+\theta)\lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}\lambda_{03}(y)e^{\boldsymbol{\beta}_3^\top z}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top z}+\theta(\Lambda_{03}(y)-\Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top z}]^{-(1/\theta)-2},$$

- *Caso 2:*

$$\lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z}[1+\theta\Lambda_{01}(y)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(y)e^{\boldsymbol{\beta}_2^\top z}]^{-(1/\theta)-1},$$

- *Caso 3:*

$$\lambda_{01}(x)e^{\boldsymbol{\beta}_1^T z}[1+\theta\Lambda_{01}(x)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(x)e^{\boldsymbol{\beta}_2^\top z}+\theta(\Lambda_{03}(b)-\Lambda_{03}(x))e^{\boldsymbol{\beta}_3^\top z}]^{-(1/\theta)-1} \text{ e}$$

- *Caso 4:*

$$[1+\theta\Lambda_{01}(b)e^{\boldsymbol{\beta}_1^\top z}+\theta\Lambda_{02}(b)e^{\boldsymbol{\beta}_2^\top z}]^{-1/\theta}.$$

A função de verossimilhança para uma amostra de n observações pode ser escrita englobando os quatro casos descritos utilizando-se as variáveis indicadoras δ_{x_i} e δ_{y_i}:

$$L(\boldsymbol{\eta}) = \prod_{i=1}^{n} \lambda_{01}(S_i)^{\delta_{x_i}} \lambda_{02}(R_i)^{\delta_{y_i}(1-\delta_{x_i})} \lambda_{03}(R_i)^{\delta_{x_i}\delta_{yi}} \times \exp[\delta_{x_i}\boldsymbol{\beta}_1^\top \boldsymbol{z}_i + \delta_{y_i}(1-\delta_{x_i})\boldsymbol{\beta}_2^\top \boldsymbol{z}_i + \delta_{x_i}\delta_{y_i}\boldsymbol{\beta}_3^\top \boldsymbol{z}_i]$$
$$\times (1+\theta)^{\delta_{x_i}\delta_{y_i}} \{1 + \theta[\Lambda_{01}(S_i)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}_i} + \Lambda_{02}(S_i)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}_i} + (\Lambda_{03}(R_i) - \Lambda_{03}(S_i))e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}_i}]\}^{-(1/\theta)-\delta_{x_i}-\delta_{y_i}}.$$

O termo $\Lambda_{03}(R_i) - \Lambda_{03}(S_i)$ não está presente para as observações que se enquadram nos casos 2 e 4, pois, pela notação adotada, $R_i = S_i$ nesses casos.

O logaritmo da verossimilhança fica dado por

$$\begin{aligned}\log(L(\boldsymbol{\eta})) = \sum_{i=1}^{n} \quad & \delta_{x_i}\log\{\lambda_{01}(S_i)\} + \delta_{y_i}(1-\delta_{x_i})\log\{\lambda_{02}(R_i)\} + \delta_{x_i}\delta_{y_i}\log\{\lambda_{03}(R_i)\} + \delta_{x_i}\boldsymbol{\beta}_1^\top \boldsymbol{z}_i + \\ & \delta_{y_i}(1-\delta_{x_i})\boldsymbol{\beta}_2^\top \boldsymbol{z}_i + \delta_{x_i}\delta_{y_i}\boldsymbol{\beta}_3^\top \boldsymbol{z}_i + \delta_{x_i}\delta_{y_i}\log\{(1+\theta)\} - ((1/\theta) + \delta_{x_i} - \delta_{y_i}) \\ & \log\{1 + \theta[\Lambda_{01}(S_i)e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}_i} + \Lambda_{02}(S_i)e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}_i} + (\Lambda_{03}(R_i) - \Lambda_{03}(S_i))e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}_i}]\}. \qquad (3.15)\end{aligned}$$

Finalmente é preciso obter o gradiente de $\log(L(\boldsymbol{\eta}))$, que corresponde ao vetor escore $U(\boldsymbol{\eta})$, dado que os estimadores de máxima verossimilhança são solução do sistema $U(\boldsymbol{\eta}) = \mathbf{0}$.

Vetor Escore

Para se obter as componentes do vetor escore das derivadas parciais referentes a $\boldsymbol{\lambda}_{0R}$, $\boldsymbol{\lambda}_{0D}$ e $\boldsymbol{\lambda}_{0RD}$, basta verificar que $\Lambda_{01}(k)$, $\Lambda_{02}(k)$ e $\Lambda_{03}(k)$ representam as funções de risco acumuladas até k dos tempos observados na amostra até a recaída, até o óbito sem recaída e até o óbito após recaída, respectivamente. Deste modo, pode-se escrever

$$\begin{aligned}\Lambda_{01}(k) &= \lambda_{011}I_{\{t_{r1}<k\}} + \lambda_{012}I_{\{t_{r2}<k\}} + \cdots + \lambda_{01m}I_{\{t_{rm}<k\}} \\ &= \sum_{j=1}^{m} \lambda_{01j}I_{\{t_{rj}<k\}}.\end{aligned}$$

Analogamente, pode-se escrever $\Lambda_{02}(k) = \sum_{j=1}^{f} \lambda_{02j}I_{\{t_{dj}<k\}}$ e $\Lambda_{03}(k) = \sum_{j=1}^{g} \lambda_{03j}I_{\{t_{rdj}<k\}}$.

A função $\log(L(\boldsymbol{\eta}))$ dada em (3.15) pode ser escrita como

$$\log(L(\boldsymbol{\eta})) = \sum_{i=1}^{n} \delta_{x_i} \log\{\lambda_{01}(S_i)\} + \delta_{y_i}(1-\delta_{x_i}) \log\{\lambda_{02}(R_i)\} + \delta_{x_i}\delta_{y_i} \log\{\lambda_{03}(R_i)\} + \delta_{x_i}\boldsymbol{\beta}_1^\top \boldsymbol{z}_i +$$
$$\delta_{y_i}(1-\delta_{x_i})\boldsymbol{\beta}_2^\top \boldsymbol{z}_i + \delta_{x_i}\delta_{y_i}\boldsymbol{\beta}_3^\top \boldsymbol{z}_i + \delta_{x_i}\delta_{y_i} \log\{(1+\theta)\} - ((1/\theta) + \delta_{x_i} - \delta_{y_i})$$
$$\log\left\{1 + \theta\left[\sum_{j=1}^{m} \lambda_{01_j} I_{\{t_{r_j} < S_i\}} e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}_i} + \sum_{j=1}^{f} \lambda_{02_j} I_{\{t_{d_j} < S_i\}} e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}_i} + \sum_{j=1}^{g} \lambda_{03_j} I_{\{R_i \le t_{rd_j} < S_i\}} e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}_i}\right]\right\}.$$

Para simplificar as expressões das derivadas parcias de cada componente do vetor escore $U(\boldsymbol{\eta})$, defina $A_i = A_{1i} + A_{2i} + A_{3i}$ em que

$A_{1i} = \Lambda_{01}(S_i) \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z}_i\} = \sum_{j=1}^{m} \lambda_{01j} I_{\{t_{rj} < S_i\}} e^{\boldsymbol{\beta}_1^\top \boldsymbol{z}_i}$,

$A_{2i} = \Lambda_{02}(S_i) \exp\{\boldsymbol{\beta}_2^\top \boldsymbol{z}_i\} = \sum_{j=1}^{f} \lambda_{02j} I_{\{t_{dj} < S_i\}} e^{\boldsymbol{\beta}_2^\top \boldsymbol{z}_i}$ e

$A_{3i} = (\Lambda_{03}(R_i) - \Lambda_{03}(S_i)) \exp\{\boldsymbol{\beta}_3^\top \boldsymbol{z}_i\} = \sum_{j=1}^{g} \lambda_{03j} I_{\{R_i \le t_{rdj} < S_i\}} e^{\boldsymbol{\beta}_3^\top \boldsymbol{z}_i}$.

Deste modo, as componentes do vetor escore $U(\boldsymbol{\eta})$ ficam dadas por:

$$\begin{aligned}
U_1 &= \frac{\partial}{\partial \theta} \log(L(\boldsymbol{\eta})) \\
&= \sum_{i=1}^{n} \left\{ \frac{\delta_{x_i}\delta_{y_i}}{1+\theta} + \frac{n}{\theta^2} \log\{1+\theta A_i\} - \left(\frac{A_i}{1+\theta A_i} \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i} \right) \right) \right\}, \\
U_2 &= \frac{\partial}{\partial \beta_1} \log(L(\boldsymbol{\eta})) \\
&= \sum_{i=1}^{n} \left\{ \delta_{x_i} \boldsymbol{z}_i - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i} \right) \frac{\theta A_{1i} \boldsymbol{z}_i}{1+\theta A_i} \right\}, \\
U_3 &= \frac{\partial}{\partial \beta_2} \log(L(\boldsymbol{\eta})) \\
&= \sum_{i=1}^{n} \left\{ (1-\delta_{x_i})\delta_{y_i} \boldsymbol{z}_i - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i} \right) \frac{\theta A_{2i} \boldsymbol{z}_i}{1+\theta A_i} \right\} \text{ e} \\
U_4 &= \frac{\partial}{\partial \beta_3} \log(L(\boldsymbol{\eta})) \\
&= \sum_{i=1}^{n} \left\{ \delta_{x_i}\delta_{y_i} \boldsymbol{z}_i - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i} \right) \frac{\theta A_{3i} \boldsymbol{z}_i}{1+\theta A_i} \right\}.
\end{aligned}$$

Finalmente, os j-ésimos elementos das componentes de U_5, U_6 e U_7 são

$$
\begin{aligned}
U_{5j} &= \sum_{i=1}^{n} \frac{d_{t_{r_j}}}{\lambda_{01_j}} - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i}\right) \frac{I_{\{t_{r_j} < S_i\}} \theta \exp(\boldsymbol{\beta}_1^\top \boldsymbol{z}_i)}{1 + \theta A_i}, \qquad j = 1 \cdots m, \\
U_{6j} &= \sum_{i=1}^{n} \frac{d_{t_{d_j}}}{\lambda_{02j}} - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i}\right) \frac{I_{\{t_{d_j} < S_i\}} \theta \exp(\boldsymbol{\beta}_2^\top \boldsymbol{z}_i)}{1 + \theta A_i}, \qquad j = 1 \cdots f, \\
U_{7j} &= \sum_{i=1}^{n} \frac{d_{t_{r_j}}}{\lambda_{03_j}} - \left(\frac{1}{\theta} + \delta_{x_i} + \delta_{y_i}\right) \frac{I_{\{S_i < t_{r_j} \leq R_i\}} \theta \exp(\boldsymbol{\beta}_3^\top \boldsymbol{z}_i)}{1 + \theta A_i}, \qquad j = 1 \cdots g.
\end{aligned}
$$

Para a solução destas equações, $U(\boldsymbol{\eta}) = \mathbf{0}$, é usada a aproximação numérica proposta por Xu et al. (2010), que, segundo os autores, converge rapidamente. O algoritmo pode ser resumido como:

1. Considere estimativas inicias para os parâmetros: $\theta^{(0)}$, $\boldsymbol{\beta}_1^{(0)}$, $\boldsymbol{\beta}_2^{(0)}$, $\boldsymbol{\beta}_3^{(0)}$, $\Lambda_{01}^{(0)}(k)$, $\Lambda_{02}^{(0)}(k)$, e $\Lambda_{03}^{(0)}(k)$. Em geral, utiliza-se $\theta^{(0)} = 1$, $\boldsymbol{\beta}_1^{(0)} = \boldsymbol{\beta}_2^{(0)} = \boldsymbol{\beta}_3^{(0)} = \mathbf{0}$ e para $\Lambda_{0j}^{(0)}(k)$, $j = 1, 2, 3$ são usadas as estimativas do tipo Nelson-Aalen das respectivas funções de risco acumuladas.

2. Dadas $\Lambda_{0j}^{(0)}(k)$, $j = 1, 2, 3$, atualizar as estimativas de θ, e $\boldsymbol{\beta}_j$ para $\theta^{(1)}$, $\boldsymbol{\beta}_j^{(1)}$, $j = 1, 2, 3$, a partir de U_1, U_2, U_3, U_4.

3. Dados $\theta^{(1)}$, $\boldsymbol{\beta}_j^{(1)}$, $j = 1, 2, 3$, obtidos no passo anterior, resolver as equações $U_5 = 0$, $U_6 = 0$ e, $U_7 = 0$ para agora atualizar as estimativas das funções de risco acumuladas $\Lambda_{0j}^{(1)}(k)$, $j = 1, 2, 3$.

4. Atualizar os valores inicias em (1) e repetir os passos (2) a (4) até atingir convergência para as estimativas.

Uma vez obtidas as estimativas dos parâmetros, é preciso calcular o erro padrão e, para isso, é calculada a matriz de informação observada, que contém as derivadas parciais de $U(\boldsymbol{\eta})$ em relação ao vetor de parâmetros $\boldsymbol{\eta}$.

Defina $\hat{I}(\hat{\boldsymbol{\eta}}) = -\dfrac{\partial U(\hat{\boldsymbol{\eta}})}{\partial \hat{\boldsymbol{\eta}}^\top}$ como a matriz de informação observada para $\hat{\boldsymbol{\eta}}$. Ela é formada pelas derivadas segundas do logaritmo da função de verossimilhança avaliadas no ponto de máximo, ou seja nos estimadores de máxima verossimilhança $\hat{\boldsymbol{\eta}}$. Xu et al. (2010) detalham os cálculos das derivadas $\dfrac{\partial U(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^\top}$.

Neste trabalho, para a obtenção das estimativas dos parâmetros, assim como seus erros padrão, foi desenvolvido e implementado em R o algoritmo proposto por Xu et al. (2010). Os códigos ficarão disponíveis em: www.ime.usp.br/ tunes/Rsemicompetitivos.

3.4 Propriedades dos estimadores

Xu et al. (2010) apresentam propriedades assintóticas dos estimadores propostos. Além disso, incluem um resultado sobre a identificabilidade do modelo proposto. Para estabelecer as propriedades assintóticas dos estimadores Xu et al. (2010), especificam as seguintes condições:

(C1) A variância do parâmetro θ_0 encontra-se em um intervalo conhecido $[0, K]$, $K < \infty$. Os parâmetros de regressão $(\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{03})$ encontram-se ao interior de um conjunto compacto. As funções de risco acumuladas basais Λ_{01}, Λ_{02} e Λ_{03} são estritamente crescentes e continuamente diferenciáveis em $[0, \tau]$ para um valor dado de $\tau < \infty$.

(C2) As observações $(S_i, \delta_{x_i}, R_i, \delta_{y_i})$, $i = 1, \cdots, n$ são independentes e identicamente distribuídas. Condicionalmente em $\boldsymbol{z}$, a censura à direita C é independente de X_i e Y_i.

(C3) A covariável $\boldsymbol{z}$ é limitada com probabilidade um, e se $\boldsymbol{b}$ é um vetor de constantes tal que $\boldsymbol{b}^\top \boldsymbol{z} = \boldsymbol{0}$ com probabilidade um, então $\boldsymbol{b} = \boldsymbol{0}$.

Assumindo $\boldsymbol{\alpha}^\top = (\theta, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)$ e $\Lambda_0 = (\Lambda_{01}, \Lambda_{02}, \Lambda_{03})$, é válido o seguinte teorema (3.1), enunciado e provado por Xu et al. (2010).

Teorema 3.1 *(Identificabilidade e consistência)*
Sob as condições (C1)-(C3),

(i) Os parâmetros $\boldsymbol{\alpha}$ e Λ_0 são identificáveis.

(ii) Com probabilidade um,

$$|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0| + \sup_{t \in [0,\tau)} |\hat{\Lambda}_0 - \Lambda_0| \to 0.$$

Isto é, $\hat{\boldsymbol{\alpha}}$ e $\hat{\Lambda}_0$ são fortemente consistentes.

(iii) $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \hat{\Lambda}_0 - \Lambda_0)$ converge fracamente a um processo gaussiano de média zero no espaço métrico $R^{3p+1} \times \{l^\infty[0, \tau]\}^3$, em que $l^\infty[0, \tau]$ é o espaço linear consistente de todas as funções limitadas em $[0, \tau]$

e está equipado com a norma supremo. Além disso, $\hat{\boldsymbol{\alpha}}$ é assintoticamente eficiente; equivalentemente, sua variância assintótica atinge o limite de eficiência semiparamétrica para $\boldsymbol{\alpha}$.

A prova do teorema (3.1) é apresentada no material suplementar de Xu et al. (2010), que segue dos resultados de Murphy (1995) referentes a teoria assintótica para modelos de fragilidade, e não será apresentada aqui.

CAPÍTULO 4

Aplicação

Este capítulo tem como objetivo principal apresentar a análise de dois conjuntos de dados reais com estrutura de riscos semicompetitivos.

O primeiro conjunto de dados é um estudo bastante utilizado na literatura, correspondente a 137 pacientes com leucemia que receberam transplante de medula óssea (Klein e Moeschberger, 2003). No entanto, conforme mencionado na introdução do trabalho, o principal objetivo está na aplicação dos modelos discutidos a um conjunto de dados reais de 1253 pacientes colombianos com doença renal crônica em um estágio avançado e que, conseqüentemente, requerem um tratamento de hemodiálise ou diálise peritoneal enquanto aguardam um possível transplante de rim. A doença renal crônica atualmente é considerada como uma patologia de alto custo, além de ter relativamente alta incidência. O conjunto de dados em questão foi disponibilizado por Laboratórios Echavarria e Fresenius Medical Care da Colômbia.

4.1 Transplante de Medula Óssea para Leucemia

O conjunto de dados faz parte de um estudo em que 137 pacientes com leucemia receberam transplante de medula óssea, que é um tratamento padrão para a leucemia aguda, foi publicado inicialmente por Copelan et al. (1991) e está disponível em Klein e Moeschberger (2003). A

recuperação após transplante é um processo complexo que depende de um conjunto de fatores de risco conhecidos no momento do transplante, como doador, idade, sexo, fase inicial da doença, tempo desde o diagnóstico para o transplante, etc.

O prognóstico final do paciente pode mudar pós-transplante pois durante o processo de recuperação podem ocorrer eventos tais como a doença de enxerto contra hospedeiro, o regresso da contagem de plaquetas aos níveis normais, o retorno de granulócitos a níveis normais, ou o desenvolvimento de infecções. Além disso, após o transplante pode haver a reincidência da leucemia, que é considerada uma falha no tratamento. Assim, o transplante pode ser considerado uma falha quando um paciente com leucemia apresenta recidiva ou quando ele morre.

Muitas vezes, dados de transplante de medula óssea são analisados com metodologias para riscos competitivos, considerando-se os eventos "recidiva"e "óbito livre da doença"como riscos competitivos. Seria interessante, no entanto, fazer uma análise desses dados considerando-se a estrutura de riscos semicompetitivos.

Deste modo, foram consideradas algumas covariáveis para fazer a aplicação dos modelos de cópulas como uma análise descritiva inicial seguida de uma análise inferencial segundo o processo doença morte, como descrito a seguir.

O transplante de medula óssea foi feito em pacientes com leucemia mielopática aguda (AML) (com alto ou baixo risco) ou leucemia linfoblástica aguda (ALL) (99 AML, 38 ALL), tratados em 4 hospitais: 76 em Ohio State Universiy (Columbus), 21 em Hahnemann University (Philadelphia), 23 em Vincent's Hospital (Sydney) e 17 em Alfred Hospital (Melbourne), entre 1º de março de 1984 e 30 de junho de 1989. O tempo de observação foi registrado em dias e os pacientes tiveram um tempo máximo de seguimento de 7 anos. Além disso, foi registrado o tempo de espera entre o diagnóstico e o transplante (entre 0,8 e 87,2 meses), com média de 19,7 meses.

Durante o período de recuperação, dos 137 pacientes transplantados, 41 morreram, 42 apresentaram recaída (40 morreram após a recaída) e 54 não apresentaram nenhum dos eventos. Antes de observar recaída ou óbito, 43 pacientes apresentaram outros eventos durante a recuperação pós-transplante (26 tiveram a doença de enxerto contra hospedeiro e 17 tiveram regresso da contagem de plaquetas aos níveis normais).

Os pacientes foram agrupados em três categorias segundo o estado da doença no momento do transplante:

- ALL (Leucemia Linfoblástica Aguda): 38 pacientes;
- AML (Leucemia Mielopática Aguda) baixo risco: 54 pacientes;

- AML (Leucemia Mielopática Aguda) alto risco: 45 pacientes.

Outras variáveis observadas no momento do transplante que serão utilizadas na análise estão listadas a seguir:

- a idade do paciente, classificada em três categorias: de 7 a 22 anos (42 pacientes), de 23 a 33 anos (55 pacientes) e de 34 a 52 anos (40 pacientes);
- uma combinação do sexo paciente-doador como segue na Tabela 4.1; e

Tabela 4.1: *Sexo do paciente e do doador*

	Paciente - Doador	nº de casos
1	Mas - Mas	55
2	Mas - Fem	33
3	Fem - Mas	25
4	Fem - Fem	24

- o status imune citomegalovírus (CMV) do paciente.

4.1.1 Análise descritiva por meio de cópulas

Para uma análise conjunta do tempo até recorrência da doença e sobrevida destes pacientes, foram usadas as cópulas estudadas no capítulo 2. Inicialmente, foi estimado o parâmetro de dependência e seu correspondente coeficiente de concordância com os erros padrão. O erro padrão (se) foi estimado para ambos os parâmetros com o estimador de variância Jackknife. Os resultados são resumidos na Tabela 4.2 e nos gráficos da Figura 4.1, nos quais pode-se observar as raízes das funções de estimação de α para as cópulas da família B, família C (Frank) e família D (Gumbel), como descrito no Capítulo 2.

Pode-se observar que, embora a cópula da família B apresente uma estimativa menor do coeficiente de concordância τ de Kendal, em geral as quatro cópulas sugerem uma forte dependência positiva entre o tempo até a recaída e o tempo até o óbito.

Uma vez calculado o estimador do parâmetro de dependência, $\hat{\alpha}$, foram obtidas as funções de sobrevivência marginais da recaída por meio do estimador cópula gráfico para as quatro cópulas

Tabela 4.2: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$*

Cópula	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)
Família A - Cópula de Clayton	7,885 (2,367)	0,798 (0,045)
Família B	6,837 (1,544)	0,707 (0,063)
Família C - Cópula de Frank	14,859 (3,851)	0,761 (0,051)
Família D - Cópula de Gumbel	4,651 (1,102)	0,785 (0,048)

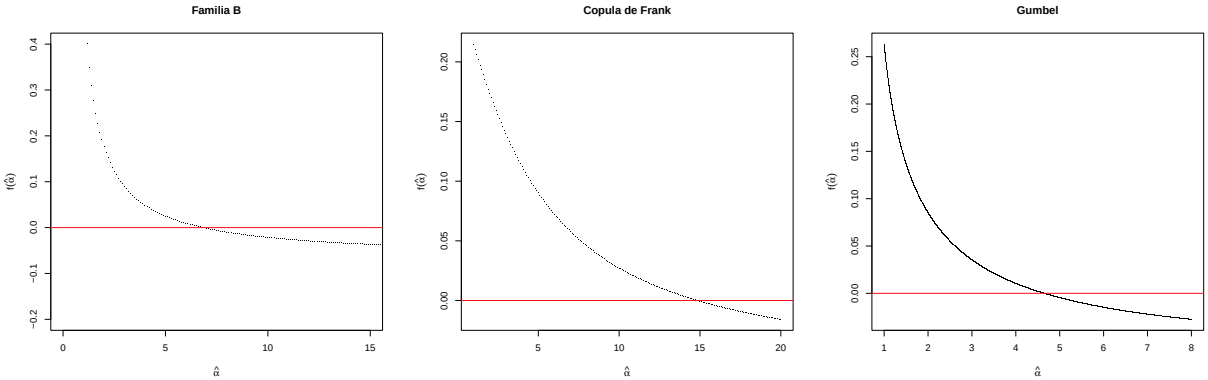

Figura 4.1: *Gráficos da equação de estimação de α, para família B, cópula de Frank e cópula de Gumbel.*

estudadas, apresentados na Figura 4.2. Note que as curvas estimadas pelas diferentes cópulas não apresentam diferenças visualmente significativas.

O primeiro gráfico da Figura 4.3 mostra as funções de sobrevivência marginal da recaída estimadas pelas quatro cópulas sobrepostas. Pode-se verificar que, por exemplo, a probabilidade de apresentar recaída após 500 dias do transplante (com qualquer das quatro cópulas) é aproximadamente 0,5, e visualmente é quase a mesma probabilidade de apresentar óbito após 500 dias do transplante. O segundo gráfico da Figura 4.3 corresponde a função de sobrevivência do óbito, estimada por Kaplan-Meier.

Para fazer uma análise descritiva segundo as covariáveis, foram obtidos os estimadores de sobrevivência marginal para a recaída e para o óbito. Os estimadores Cópula-Gráfico foram obtidos utilizando-se as cópulas de Clayton e Gumbel para as curvas de sobrevivência marginal, além

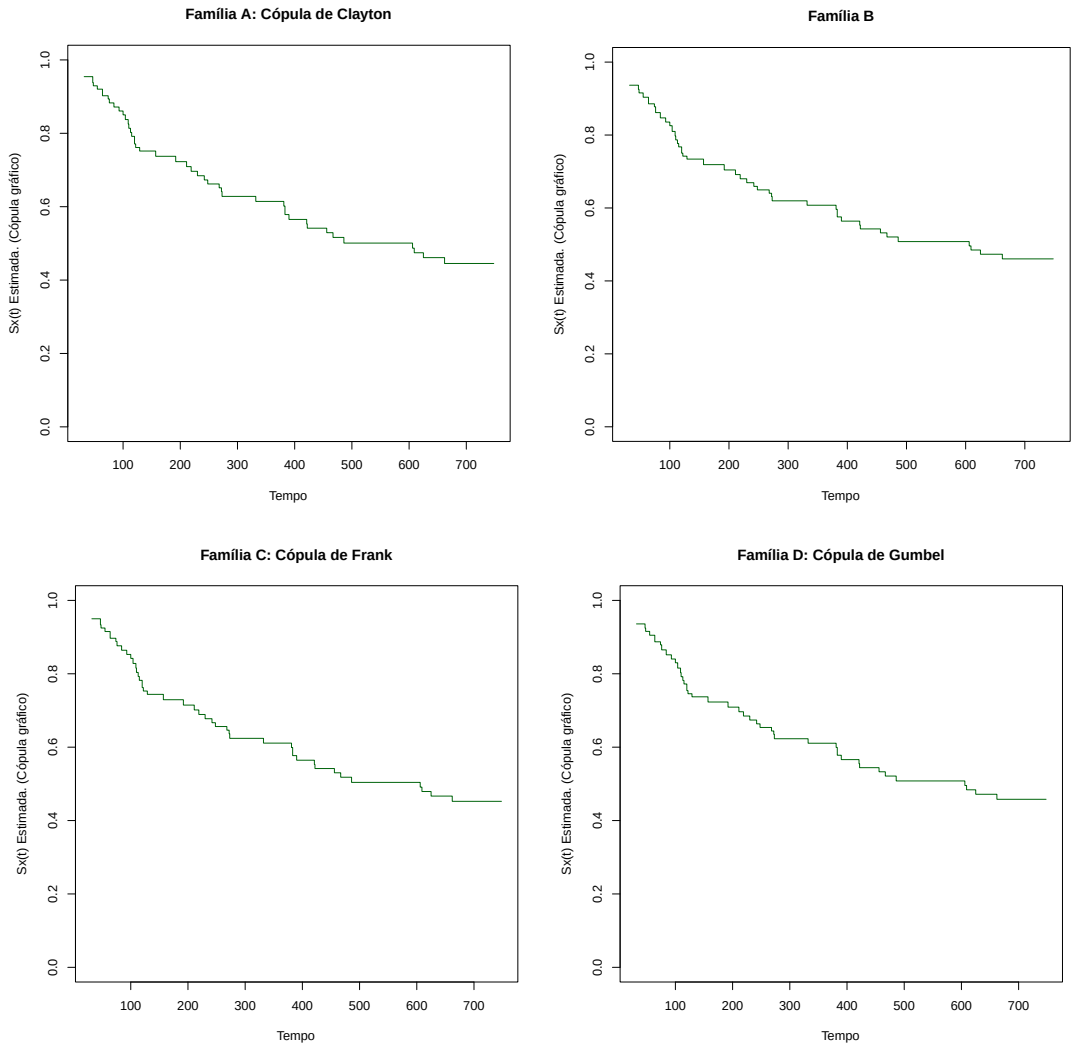

Figura 4.2: *Função de sobrevivência marginal da recaída (obtida pelo estimador Cópula Gráfico).*

dos parâmetros de dependência α e e correlação τ, dos pacientes que apresentaram recaída, e estimadores Kaplan-Meier para as curvas de sobrevivência marginal dos pacientes que apresentaram óbito. Foram utilizadas apenas duas cópulas porque os resultados entre as quatro cópulas foram muito parecidos.

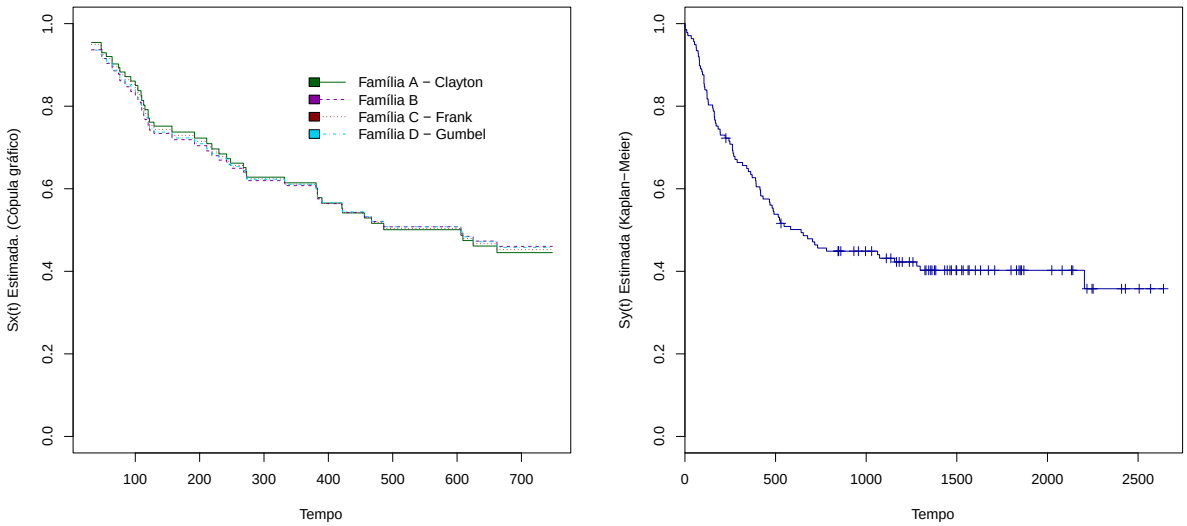

Figura 4.3: *Funções de sobrevivência marginais estimadas de recaída (à esquerda) e óbito (à direita), obtidas pelo estimador Cópula-Gráfico e Kaplan-Meier, respectivamente.*

Deste modo, da Tabela 4.3 pode-se observar que a menor dependência entre os eventos recaída e óbito por grupo de doença é para pacientes com leucemia mielopática aguda (AML) de baixo risco. Comparando as curvas de sobrevivência marginais pelo grupo de doença, como mostra a Figura 4.4, tem-se que os pacientes com leucemia mielopática aguda (AML) de alto risco apresentam a menor probabilidade de sobrevida tanto para a recaída quando para o óbito. Visualmente as curvas estimadas para os dois eventos, recaída e óbito, apresentam diferenças ao longo do tempo, principalmente a curva dos pacientes com leucemia mielopática aguda (AML) de baixo risco.

Com relação a idade, observe que a maior dependência entre recaída e óbito corresponde aos pacientes com idades entre 23 e 33 anos e a menor dependência para os mais jovens, segundo o estimador de τ de ambas as cópulas de Clayton e Gumbel, como mostra a Tabela 4.4. Segundo os gráficos apresentados na Figura 4.5, as pessoas mais idosas apresentaram uma probabilidade menor de sobrevida comparada com as pessoas menos idosas. Com relação às curvas de recaída, observa-se que pacientes no grupo etário de maior idade também apresentam maior incidência de recaída.

Tabela 4.3: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ para o grupo de doença*

Grupo de Doença	Cópula de Clayton $\hat{\alpha}$ (se)	Cópula de Clayton $\hat{\tau}$ (se)	Cópula de Gumbel $\hat{\alpha}$ (se)	Cópula de Gumbel $\hat{\tau}$ (se)
ALL	6,758 (4,827)	0,772 (0,097)	4,454 (2,420)	0,776 (0,103)
AML baixo risco	6,168 (5,000)	0,755 (0,116)	3,263 (1,857)	0,693 (0,136)
AML alto risco	6,230 (4,082)	0,757 (0,093)	5,068 (2,739)	0,803 (0,087)

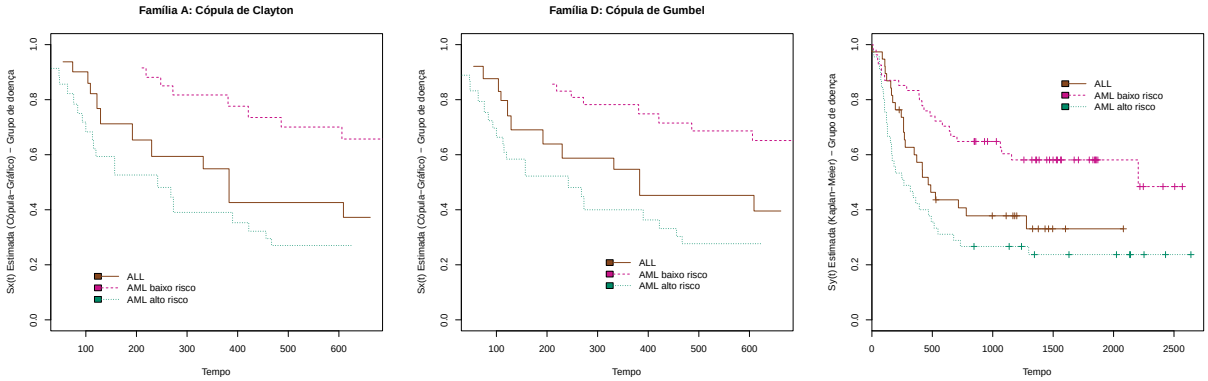

Figura 4.4: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o grupo de doença, obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

Para a configuração do sexo entre paciente e doador, nos gráficos da Figura 4.6 pode-se observar que a menor probabilidade de sobrevida tanto para a recaída quanto para o óbito corresponde a homens que receberam o transplante de mulheres, porém, a probabilidade de sobrevivência é maior quando o doador é homem. Em geral, quando o sexo de receptor e o doador são iguais, a probabilidade de sobrevivência é maior quando comparada a curva de sobrevivência de pacientes com sexos receptor-doador diferentes. Isto é visualizado nas curvas estimadas dos dois eventos, recaída e óbito. Com relação à dependência entre os eventos recaída e óbito pelo sexo do paciente e doador, note pela Tabela 4.5 que, em geral, existe alta dependência, sendo a menor quando o paciente e o doador são de sexo feminino.

Na Tabela 4.6, são resumidos os valores estimados dos parâmetros de dependência α e correlação τ das cópulas Clayton e Gumbel para o status imune do citomegalovírus do paciente. Note que também há dependência forte, sendo que, para pacientes com CMV positivo

Tabela 4.4: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ para a idade*

Idade	Cópula de Clayton		Cópula de Gumbel	
	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)
7 - 22 anos	4,573 (4,581)	0,696 (0,155)	3,309 (2,214)	0,698 (0,155)
23 - 33 anos	13,901 (5,430)	0,874 (0,040)	6,834 (2,425)	0,854 (0,049)
34 - 52 anos	6,525 (9,776)	0,765 (1,400)	4,167 (4,608)	0,760 (0,146)

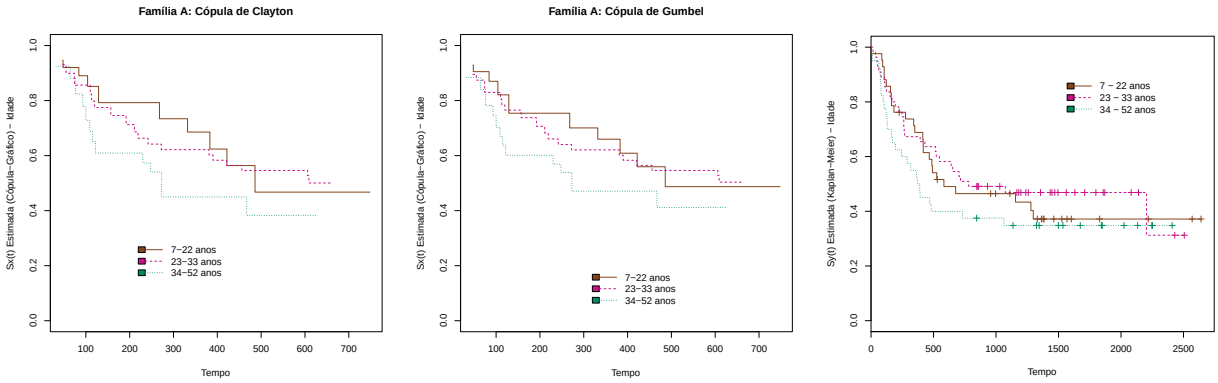

Figura 4.5: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo a faixa etária do paciente, obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

a dependência entre recaída e óbito é ainda maior, acima de 80%. As estimativas das funções de sobrevivência marginais são apresentadas nos gráficos da Figura 4.7. Pode-se observar que pacientes com o status do vírus positivo têm uma probabilidade de sobrevivência menor do que pacientes com status negativo.

Para a covariável hospital, as curvas estimadas são apresentadas nos gráficos das Figuras 4.8 e 4.9. Numa primeira situação, são considerados os quatro hospitais (Figura 4.8) e as curvas de sobrevivência marginais estimadas parecem ser bem diferentes. Note que a probabilidade de sobrevida maior corresponde a pacientes atendidos no hospital americano Hahnemann e a menor, a pacientes atendidos no hospital australiano Alfred, para as curvas feitas tanto para o evento recaída quanto para o óbito.

Como em alguns hospitais o número de pacientes é bastante reduzido, foi considerada também

Tabela 4.5: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ para a combinação do sexo paciente - doador*

Configuração do sexo Pac-Doa	Cópula de Clayton $\hat{\alpha}$ (se)	Cópula de Clayton $\hat{\tau}$ (se)	Cópula de Gumbel $\hat{\alpha}$ (se)	Cópula de Gumbel $\hat{\tau}$ (se)
Pac M - Doa M	9,679 (6,925)	0,829 (0,074)	5,912 (4,151)	0,831 (0,099)
Pac M - Doa F	8,253 (4,273)	0,805 (0,075)	5,704 (2,890)	0,825 (0,084)
Pac F - Doa M	9,642 (7,276)	0,828 (0,092)	7,232 (5,209)	0,862 (0,086)
Pac F - Doa F	4,279 (4,849)	0,681 (0,181)	3,615 (2,855)	0,723 (0,164)

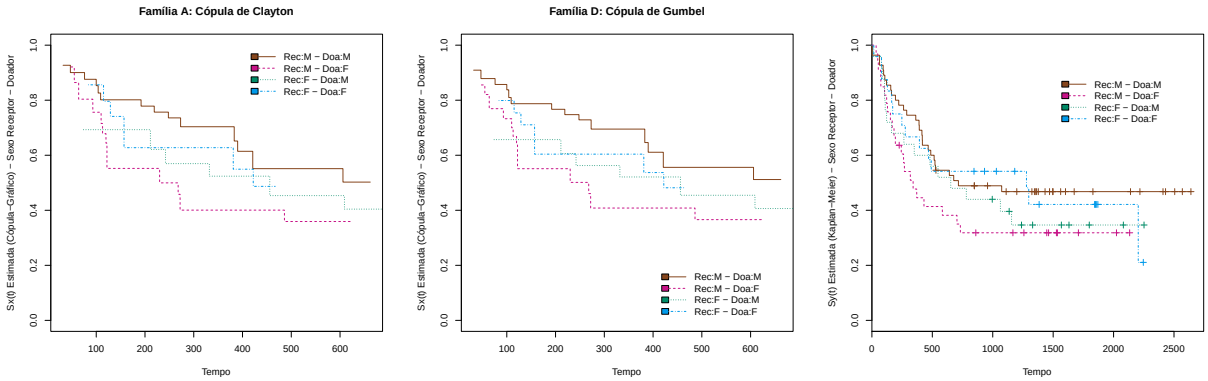

Figura 4.6: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o sexo do paciente e do doador, obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

a covariável hospital com duas categorias:

- Hospitais americanos: Ohio State University e Hahnemann University com 97 pacientes atendidos.
- Hospitais australianos: Alfred e St. Vincent com 40 pacientes atendidos.

Com esta recategorização, foram obtidas as estimativas do parâmetro de dependência para as cópulas de Clayton e Gumbel, apresentadas na Tabela 4.7, e as curvas de sobrevivência marginais estimadas, dispostas na Figura 4.9. Observe que a proporção de pessoas atendidas nos hospitais australianos foi inferior (29%) à dos americanos (71%) e a dependência entre recaída e óbito dos pacientes atendidos nos hospitais de ambos países são muitos próximas, sendo um pouco maior nos hospitais australianos.

Tabela 4.6: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ para o status imune citomegalovírus*

CMV	Cópula de Clayton		Cópula de Gumbel	
	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)
CMV Negativo	6,567 (4,431)	0,767 (0,081)	4,117 (1,651)	0,757 (0,088)
CMV Positivo	8,730 (4,356)	0,814 (0,060)	5,249 (2,060)	0,809 (0,062)

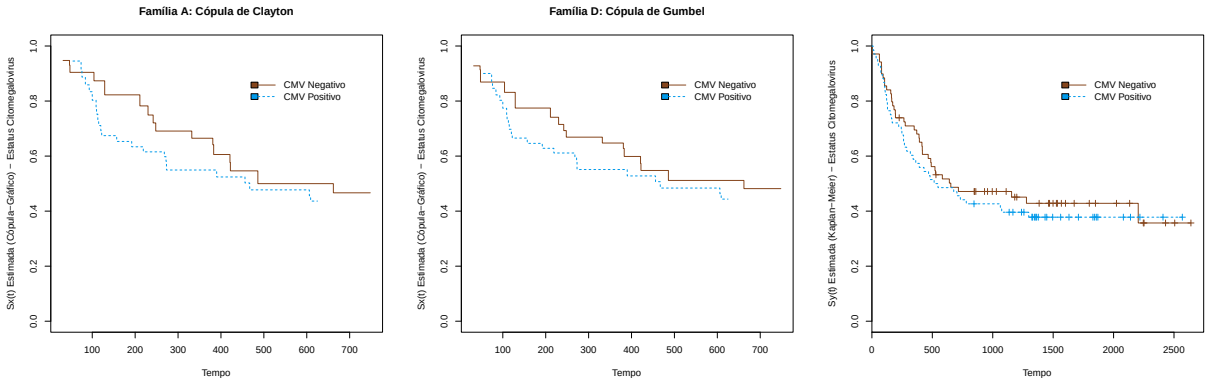

Figura 4.7: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o status imune do citomegalovírus do paciente, obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

Além disso, as conclusões são diferentes quando comparadas as curvas estimadas com o evento recaída e óbito. Observando as curvas correspondentes à recaída, a probabilidade de recaída é maior para pessoas atendidas nos hospitais australianos. Já no caso do óbito, a probabilidade de sobrevida maior corresponde a pessoas atendidas em hospitais americanos.

Além das curvas de sobrevivência marginais, existem outras quantidades de interesse que podem ser obtidas a partir de função de sobrevivência bivarida, tais como a probabilidade de sobrevivência condicional quando não há recaída num tempo x ou quando há recaída em x, mas o paciente segue vivo em t.

Considere um paciente que não teve recaída no tempo x. Duas formas de estimação (não paramétrica e semiparamétrica) da função de sobrevivência condicional são apresentadas na Seção 2.4.4. Suponha por exemplo que não houve recaída após $x = 565$ dias de observação, de modo que a probabilidade a ser calculada é: $\tilde{P}(Y > y|X > 565; Y > 565)$. Esta probabilidade pode ser

Tabela 4.7: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ para o hospital*

Idade	Cópula de Clayton		Cópula de Gumbel	
	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)
Hosp Americanos	6,351 (2,740)	0,760 (1,566)	4,609 (0,068)	0,783 (0,066)
Hosp Australianos	6,929 (3,688)	0,776 (2,307)	5,166 (0,080)	0,806 (0,075)

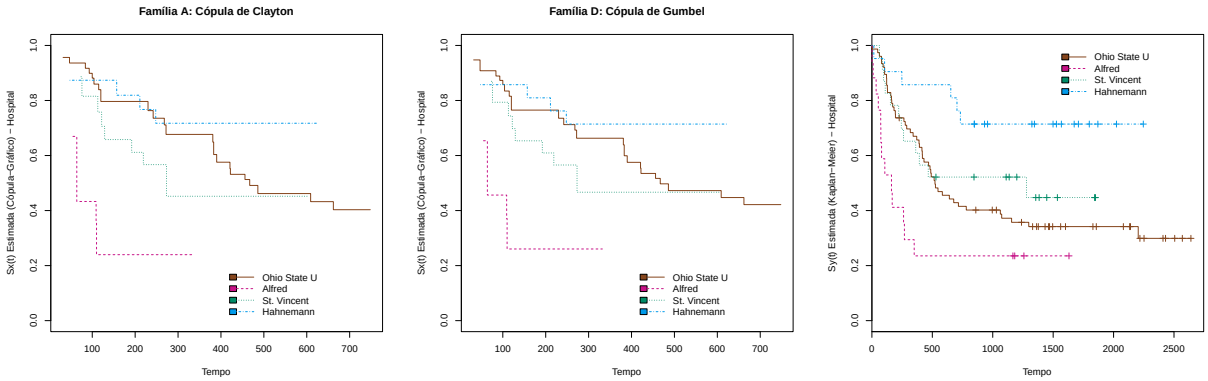

Figura 4.8: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o hospital, obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

estimada tanto pelo estimador não paramétrico quanto pelo semiparamétrico com base na cópula usada na estimação da função de sobrevivência bivariada. Os resultados são apresentados nos gráficos da Figura 4.10. Nesses gráficos, pode-se observar principalmente que o estimador não paramétrico não necessariamente é não crescente em y. Isso ocorre porque a estimativa da função de sobrevivência bivariada não paramétrica é baseada na função de sobrevivência estimada da censura. Por causa disso, ela pode crescer em um tempo de censura observado $R_i > x$ para um dado instante em que $S_i < x$ (Xu et al., 2010). No entanto, as estimativas semiparamétricas construídas por meio das cópulas são sempre não crescentes.

Também foi calculada a probabilidade de sobrevivência de um paciente que não teve recaída no tempo $x = 410$, isto é, $\tilde{P}(Y > y|X = 410; Y > 410)$. Neste caso, conforme pode ser observado na Figura 4.11, o estimador não paramétrico é não crescente. Também as probabilidades de sobrevivência condicional são maiores quando obtidas pelo estimador semiparamétrico usando as cópulas de Frank e Gumbel.

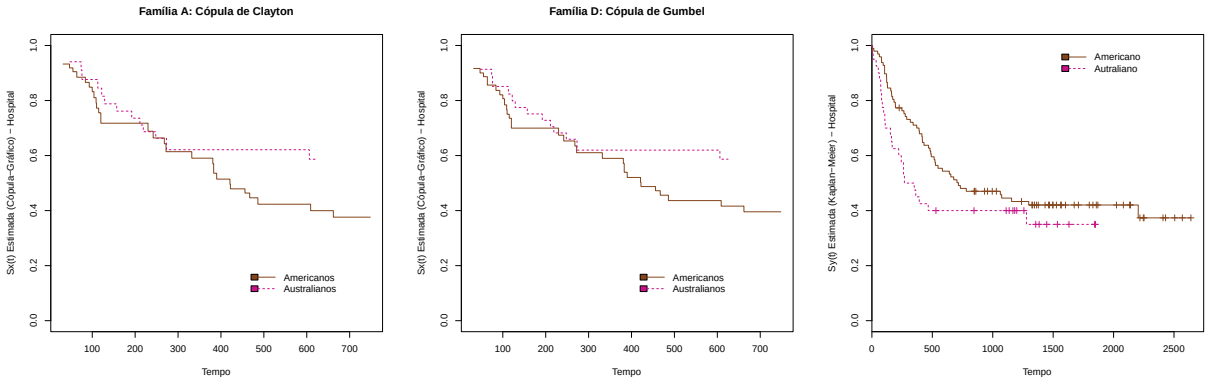

Figura 4.9: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o hospital (recateogrizado), obtidas pelos estimadores Cópula-Gráfico, cópulas de Clayton (à esquerda) e Gumbel (centro) e Kaplan-Meier (à direita).*

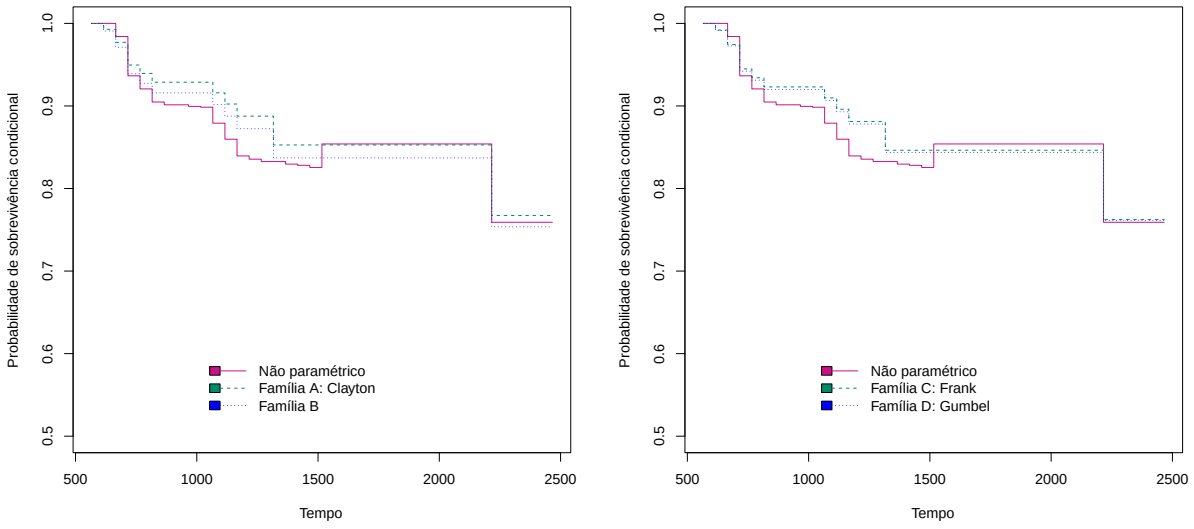

Figura 4.10: *Função de sobrevivência condicional quando não há recaída no tempo $x = 565$. Estimadores semi-paramétricos (cópulas Frank e Gumbel à direita, cópulas Clayton e Família B à esquerda) e não paramétrico.*

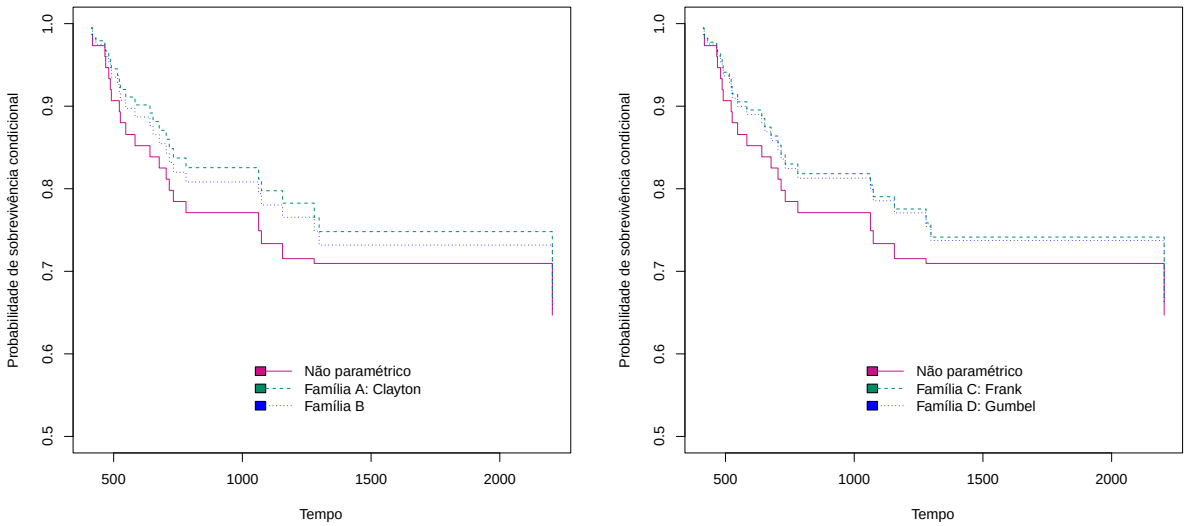

Figura 4.11: *Função de sobrevivência condicional quando não há recaída no tempo $x = 410$. Estimadores semi-paramétricos (cópulas Frank e Gumbel à direita, cópulas Clayton e Família B à esquerda) e não paramétrico.*

Finalmente, como outra aplicação da função de sobrevivência bivariada, foi calculada a probabilidade de sobrevivência de um paciente que teve recaída ao tempo x, mas continua vivo em $t > x$. Para isso foi considerado um paciente com recaída em $x = 486$ que continuava vivo em $t = 500$, ou seja, foi calculada a probabilidade $\tilde{P}(Y > y|X = 486; Y > 500)$ com as quatro cópulas. Os resultados são apresentados na Figura 4.12.

Esta curva é a que as estimativas baseadas nas quatro cópulas apresentam a maior diferença. Pode-se perceber que curva estimada com base na cópula da família B é superior às curvas estimadas usando as outras cópulas, que também apresentam diferenças.

Segundo os resultados do estimador Cópula-Gráfico para a função de sobrevivência marginal do evento intermediário recaída, as cópulas estudadas parecem ter um comportamento similar. No entanto, a cópula da família B apresenta resultados um pouco diferentes quando comparada com as cópulas de Clayton, Frank e Gumbel. Em geral, pode se dizer que os resultados são próximos independentemente da cópula Arquimediana escolhida.

Para analisar os dados de transplante de medula óssea por uma perspectiva inferencial, é

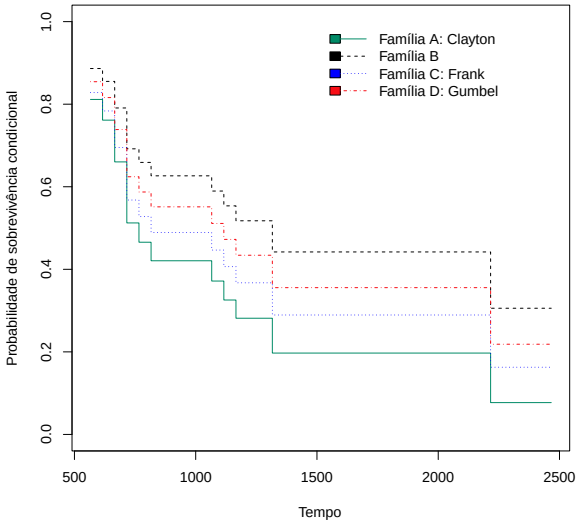

Figura 4.12: *Função de sobrevivência condicional de um paciente que apresentou recaída em $x = 486$.*

utilizado o processo doença-morte como apresentado a seguir.

4.1.2 Processo doença-morte

No processo doença-morte são modeladas as funções de intensidade de transição entre os estados que representam os eventos de interesse, recaída e óbito. No caso geral, incluindo covariáveis, o modelo condicional a ser estimado pode ser expresso por:

$$\begin{aligned}
\lambda_1(x|\gamma, z) &= \gamma\lambda_{01}(x)\exp\{\boldsymbol{\beta}_1^\top z\}, \quad x > 0, \\
\lambda_2(y|\gamma, z) &= \gamma\lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\}, \quad y > 0, \\
\lambda_{12}(y|\gamma, x, z) &= \gamma\lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\}, \quad 0 < x < y,
\end{aligned} \tag{4.1}$$

em que $\gamma \sim \Gamma(1/\theta, 1/\theta)$ é a variável de fragilidade, $\lambda_{01}(x)$, $\lambda_{02}(y)$ e $\lambda_{03}(y)$ representam as funções de risco basais desconhecidas de recaída, óbito e óbito após recaída respectivamente, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ e $\boldsymbol{\beta}_3$ são os vetores dos coeficientes de regressão para a recaída, o óbito e o óbito após recaída, respectivamente,

e z é a matriz de covariáveis.

Tabela 4.8: *Descrição das covariáveis a serem incorporadas no processo doença-morte*

Variável	Categorias
Grupo de doença	1 Leucemia Linfoblástica aguda - ALL 2 Leucemia mielopática aguda, baixo risco - AML baixo risco 3 Leucemia mielopática aguda, alto risco - AML alto risco
Idade do paciente	1 De 7 a 22 anos 2 De 23 a 33 anos 3 De 34 a 52 anos
Sexo receptor-doador	1 Receptor: M, Doador: M 2 Receptor: M, Doador: F 3 Receptor: F, Doador: M 4 Receptor: F, Doador: F
Status imune citomegalovírus	0 Status Imune citomegalovírus - CMV Negativo 1 Status Imune citomegalovírus - CMV Positivo
Hospital	0 Hosp americanos (The Ohio State University, Hahnemann) 1 Hosp australianos (Alfred, St. Vincent)

Na Tabela 4.8 são descritas as covariáveis a serem incluídas no modelo (4.1). Considerando que a maioria das covariáveis tem mais de duas categorias, foram criadas variáveis binárias como segue:

- Grupo de doença

$$z_{1i} = \begin{cases} 1 = \text{Leucemia mielopática aguda baixo risco (AML baixo risco)}, \\ 0 = \text{Caso contrário.} \end{cases}$$

$$z_{2i} = \begin{cases} 1 = \text{Leucemia mielopática aguda alto risco (AML alto risco)}, \\ 0 = \text{Caso contrário.} \end{cases}$$

- Idade do paciente

$$z_{3i} = \begin{cases} 1 = \text{De 23 a 33 anos}, \\ 0 = \text{Caso contrário.} \end{cases}$$

$$z_{4i} = \begin{cases} 1 = \text{Maior de 33 anos,} \\ 0 = \text{Caso contrário.} \end{cases}$$

- Sexo do paciente e do doador

$$z_{5i} = \begin{cases} 1 = \text{Receptor M - Doador F,} \\ 0 = \text{Caso contrário.} \end{cases}$$

$$z_{6i} = \begin{cases} 1 = \text{Paciente F - Doador M,} \\ 0 = \text{Caso contrário.} \end{cases}$$

$$z_{7i} = \begin{cases} 1 = \text{Paciente F - Doador F,} \\ 0 = \text{Caso contrário.} \end{cases}$$

- Status imune citomegalovírus (CMV)

$$z_{8i} = \begin{cases} 1 = \text{CMV Positivo,} \\ 0 = \text{CMV Negativo.} \end{cases}$$

- Hospital

$$z_{9i} = \begin{cases} 1 = \text{Hospital autraliano,} \\ 0 = \text{Hospital americano.} \end{cases}$$

com $i = 1 \cdots 137$.

Conforme discutido no Capítulo 3, as estimativas do modelo (4.1) são obtidas por máxima verossimilhança, considerando-se a distribuição marginal dos tempos e os erros padrão (se) estimados a partir da inversa da matriz de informação observada. Para testar os coeficientes de regressão foi utilizado o teste assintótico de Wald. Os resultados do ajuste completo são resumidos na Tabela 4.9.

Dos resultados obtidos, pode-se verificar que o efeito do grupo de doença foi significativo para o tempo até recaída. Para o tempo de óbito após recaída, só o grupo dos pacientes com AML (Leucemia mielopática aguda) de alto risco foi significativo. Note que o coeficiente de regressão estimado para pacientes com AML (Leucemia mielopática aguda) de baixo risco é negativo, em direção oposta ao coeficiente de regressão estimado para pacientes com AML (Leucemia mielopática aguda) de alto risco que é positivo.

Tabela 4.9: *Estimativas para o modelo geral*

	Tempo até recaída			Tempo até óbito			Tempo até óbito após recaída		
Covariáveis	Estimativa	EP	valor P	Estimativa	EP	P valor	Estimativa	EP	valor P
z_1 (AML baixo vs ALL)	-1,160	0,519	0,026	-0,157	0,458	0,732	-0,605	0,638	0,343
z_2 (AML alto vs ALL)	0,990	0,505	0,050	0,636	0,519	0,221	1,540	0,508	0,002
z_3 (23-33a vs 2-22a)	0,685	0,511	0,180	-0,179	0,512	0,726	0,637	0,636	0,316
z_4 (34-52a vs 2-22a)	0,509	0,591	0,389	0,490	0,550	0,373	0,593	0,636	0,351
z_5 (P:M-D:F vs P:M-D:M)	0,569	0,489	0,245	0,376	0,511	0,462	0,402	0,504	0,425
z_6 (P:F-D:M vs P:M-D:M)	0,071	0,555	0,898	0,510	0,488	0,296	0,284	0,661	0,668
z_7 (P:F-D:F vs P:M-D:M)	0,081	0,572	0,887	0,089	0,565	0,875	-0,285	0,700	0,684
z_8 (CMV)	-0,141	0,485	0,771	-0,379	0,462	0,411	-0,264	0,627	0,673
z_9 (Hospital)	0,837	0,470	0,075	0,835	0,472	0,077	0,518	0,555	0,351
θ	1,138	0,350	0,001						

A covariável Hospital pode ser considerada significativa com um nível de significancia de 10%, referentes ao tempo até recaída e ao tempo até o óbito sem recaída.

Na maioria dos casos o risco de recaída ou óbito é menor para o grupo de comparação. Por exemplo, em relação à idade, o menor risco é para os mais jovens tanto para recaída como para óbito com ou sem recaída. Igualmente ocorre para o hospital, em que o menor risco corresponde aos hospitais americanos, que são o grupo de comparação.

Com relação ao sexo, o menor risco, tanto de recaída quanto o óbito (sem recaída), é para o caso em que paciente e doador são de sexo masculino (grupo de comparação); já para o óbito após recaída, o menor risco é no caso em que paciente e doador são do sexo feminino.

Segundo os resultados da Tabela 4.9, o modelo geral (4.1) condicional à variável de fragilidade fica dado por:

$$\begin{aligned}
\hat{\lambda}_1(x|\hat{\gamma},\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{01}(x)\exp\{-1,16z_1+0,99z_2+0,68z_3+0,51z_4+0,57z_5+0,07z_6+0,08z_7-0,14z_8+0,84z_9\},\\
\hat{\lambda}_2(y|\hat{\gamma},\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{02}(x)\exp\{-0,16z_1+0,64z_2-0,18z_3+0,49z_4+0,38z_5+0,51z_6+0,09z_7-0,38z_8+0,84z_9\} \text{ e}\\
\hat{\lambda}_3(y|\hat{\gamma},x\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{03}(x)\exp\{-0,60z_1+1,54z_2+0,64z_3+0,59z_4+0,40z_5+0,28z_6-0,28z_7-0,26z_8+0,55z_9\}
\end{aligned}$$

Uma possibilidade de interpretação em termos de riscos proporcionais dada a fragilidade (fixada para o mesmo paciente) seria uma situação hipotética. Por exemplo, uma possível interpretação para o coeficiente associado a z_1 em $\hat{\lambda}_1(x|\gamma,\boldsymbol{z})$ seria: o risco de recaída de um paciente do grupo AML de baixo risco seria 0,31 (= $e^{-1,16}$) vezes o risco desse mesmo paciente se ele estivesse no

grupo de doença ALL, porém, o risco de recaída é 2,69 (= $e^{0,99}$) vezes o risco dele caso ele estivesse no grupo AML de alto risco.

Com relação ao parâmetro de fragilidade, a estimativa obtida para o modelo geral gerou uma variância da fragilidade de 1,138 (0,350), indicando associação entre recaída e óbito, porém, uma parte da dependência é capturada pelas funções de risco basais.

O nível descritivo (valor P) foi obtido através do teste assintótico de Wald assumindo que, sob H_0, a estatística do teste tem distribuição chi-quadrado com um grau de liberdade. Embora esta estatística é usada pelos autores Xu et al. (2010), a estatística de Wald pode não ser correta, dado que o valor do parâmetro sob a hipóetese nula encontra-se na fronteira do espaço paramétrico. Assim, apesar de apresentado no trabalho, o valor P deste teste deve ser interpretado com bastante cautela.

4.2 Doença Renal Crônica

Finalmente serão analisados agora os dados do estudo sobre a doença renal crônica, descritos inicialmente na seção 1.1.

Na Colômbia, a incidência da doença renal crônica está aumentando consideravelmente, e sendo uma doença de alto custo, gera grande impacto no sistema de saúde colombiano. Dessa forma, pesquisas associadas a esta doença são de interesse para instituições gestoras de saúde.

Segundo García et al. (2005) e CRC (2011), quando há alteração funcional ou estrutural dos rins, com ou sem diminuição da taxa de filtração glomerular (TFG), manifestada por marcadores de dano renal (anomalias na composição do sangue e da urina) por mais de três meses, considera-se que um paciente apresenta doença renal crônica. A taxa de filtração glomerular (TFG) é estimada com equações que contém variáveis de creatinina, idade, sexo e tamanho corporal. As pessoas diagnosticadas com esta patologia são classificadas em diferentes estágios da doença.

Na literatura médica, existem várias referências para determinar os estágios da doença renal crônica. Na Colômbia, estes estágios são medidos em referência à taxa de filtração glomerular (TFG). Em geral, quanto menor valor da TFG, maior o estágio da doença, sendo que pacientes nos estágios 5 ou 4 seguem tratamento de diálise e precisam de transplante de rim. Na Tabela 4.10, é resumida a descrição dos estágios desta doença.

A doença renal crônica é uma patologia que afeta pessoas de ambos sexos em todas as idades. No acompanhamento dos 1253 pacientes realizado por Fresenius Medical Care de Colombia e

Tabela 4.10: *Estágios da doença renal crônica*

Estágio	Descrição	(TFG) (mL/min/1,73 m^2)
1	Dano renal com filtração normal ou alta	Maior a 90
2	Dano renal com leve diminuição da função renal	60 a 89
3	Dano renal com moderada diminuição da função renal	59 a 30
4	Dano renal com severa diminuição da função renal	15 a 29
5	Falha renal	menor a 15 ou diálise

Laboratorio Médico Echavarria, a maior proporção de pessoas é de sexo masculino (61,2%). Quanto à idade, a maior porcentagem foi de pessoas entre 51 e 65 anos (34%) seguida de maiores de 66 anos (30,6%). Já os mais jovens (menores de 30 anos) representaram 10% da amostra. A idade mínima observada foi de 1 ano e a máxima de 97 anos.

Em relação à cidade, 280 pacientes foram de Manizales, 131 de Rionegro, 138 de Tunja, 370 de Monteria e 334 de Sincelejo, sendo as duas últimas pertencentes à região atlântica Colombiana.

Os dados fornecidos por Laboratorios Echavarria e Fresenius Medical Care de Colômbia correspondem a pacientes com doença renal nos estágios maiores da doença e que portanto requerem tratamento de diálises. A Tabela 4.11 apresenta uma descrição das variáveis registradas no início do seguimento dos pacientes.

Para o estudo da "progressão"ou "piora"do quadro clínico, são avaliadas as concentrações de fósforo inorgânico, cálcio, potássio ou nitrogênio uréico (BUN) no sangue. De maneira geral, a progressão é caracterizada pelos níveis destes exames fora dos establecidos como referência.

Os níveis de referência para os exames feitos nos pacientes com doença renal que iniciaram o tratamento com diálises são: Cálcio de 8,7 a 10,4 mg/dl; Fósforo, de 3,4 a 5,6 mg/dl; Nitrogênio uréico, de 9 a 23 mg/dl; e Potássio, de 4 a 6 mg/dl.

Um paciente quando classificado em estágio 4 ou 5 da doença deve seguir tratamento de diálise, que pode ser diálise peritoneal ou hemodiálise. Dentro do seguimento destes pacientes, são medidos mensalmente os níveis de fósforo inorgânico, cálcio, potássio e nitrogênio uréico (BUN), os quais permitem definir quando acontece uma piora no quadro da doença.

Na estrutura de dados com riscos semicompetitivos há duas variáveis resposta de interesse, o tempo até a recaída e o tempo até o óbito. Nos dados de doença renal, as variáveis resposta foram obtidas a partir das datas em que foram realizados os exames de fósforo inorgânico, cálcio, potássio e nitrogênio uréico (BUN) e as datas de óbitos, quando ocorrem.

Tabela 4.11: *Descrição das covariáveis medidas no estudo da doença renal*

Variável	Categorias
Tratamento	0 Diálises peritoneal 1 Hemodiálises
Sexo	0 Masculino 1 Feminino
Idade do paciente	1 De 1 a 30 anos 2 De 31 a 50 anos 3 De 51 a 65 anos 4 De 66 a 97 anos
Cidade	1 Manizales 2 Monteria 3 Rionegro 4 Sincelejo 5 Tunja

Segundo Steddon e Sharples (2009), existe uma forte associação entre hiperfosfatemia e mortalidade de pacientes em ambos os tipos de tratamento, hemodiálise e diálise peritoneal. Da mesma forma, todas as causas de mortalidade de pacientes com doença renal crônica estão associadas a altos níveis de cálcio. Em geral, altos níveis da concentração dos compostos mencionados podem ser considerados para a medição da piora da doença.

Segundo pesquisadores envolvidos no estudo, diferentes critérios podem ser usados para a definição de um estado pior ou progressão da doença. Deste modo, foram estudadas as variáveis resposta considerando como "recaída"ou "progessão"altos níveis de fósforo inorgânico, cálcio, fósforo e cálcio e todos as concentrações citadas conjuntamente. Foi considerado também que pacientes podem entrar no estudo já no estado de progressão da doença. Deste modo, o número de pacientes que apresentaram os eventos de interesse, recaída e óbito, nestas configurações são apresentados na Tabela 4.12.

Dada a quantidade de observações censuradas (Tabela 4.12), neste trabalho foi considerada a progressão da doença apenas quando o valor do exame de fósforo inorgânico no sangue era superior ao nível de referência, isto é, valores acima de 5,6 mg/dl.

Quando a alteração no nível de fósforo ocorre no primeiro exame, significa que o paciente

Tabela 4.12: *Número de pacientes que apresentaram os eventos de recaída e óbito, considerando como recaída altos níveis de fósforo, cálcio, fósforo e cálcio e todos os exames (fósforo, cálcio, potássio e BUN)*

Evento	Fósforo	Cálcio	Fósforo-Cálcio	Todos
Recaída	511	270	79	23
Óbito sem recaída	121	190	221	233
Óbito após recaída	118	49	18	6
Censurados	503	744	935	991

entrou no estudo já com progressão da doença. No entanto, segundo os pesquisadores, nesta situação a recaída é melhor considerada nos resultados de exames posteriores. Deste modo uma reclassificação é considerada e os pacientes que apresentaram os eventos de interesse são resumidos na Tabela 4.13, em que o número total de recaídas observadas foi 365. É importante ressaltar que esse é um estudo ainda em andamento e, portanto, espera-se que em mais algum tempo de seguimento se tenha disponível mais informação sobre recaída e óbito dos pacientes e a análise deverá ser refeita.

Tabela 4.13: *Número de pacientes que apresentaram os eventos de recaída e óbito, considerando como recaída altos níveis de fósforo*

Evento	Fósforo $> 5,6 mg/dl$
Recaída	303
Óbito sem recaída	177
Óbito após recaída	62
Censurados	711

Na análise descritiva destes dados, é utilizada a modelagem por meio de cópulas como descrito no Capítulo 2.

4.2.1 Análise descritiva por meio de cópulas

Numa primeira etapa da análise descritiva, foi obtida a distribuição do número de pacientes com doença renal crônica que apresentaram os eventos de interesse, no período de observação, para cada categoria das covariáveis consideradas apresentada na Tabela 4.14. Na análise, foi considerado que um paciente apresenta recaída quando seu nível de fósforo no sangue é superior a 5,6 mg/dl.

Tabela 4.14: *Distribuição do número de pacientes com doença renal crônica que apresentaram os eventos de interesse*

Variável	Total de recaídas	Recaída sem óbito	Óbito sem recaída	Óbito após recaída
Tratamento				
0 Diálises peritoneal	72	60	58	12
1 Hemodiálises	293	243	119	50
Sexo				
0 Masculino	149	121	73	28
1 Feminino	216	182	104	34
Idade do paciente				
1 De 1 a 30 anos	45	38	10	7
2 De 31 a 50 anos	106	95	25	11
3 De 51 a 65 anos	140	114	60	26
4 De 66 a 97 anos	74	56	82	18
Cidade				
1 Manizales	102	92	34	10
2 Monteria	92	69	66	23
3 Rionegro	49	45	9	4
4 Sincelejo	73	56	49	17
5 Tunja	49	41	19	8

Uma forma de identificar intuitivamente a força da dependência entre os eventos recaída e óbito é por meio da proporção de morte entre os paciente que apresentaram recaída (Xu et al., 2010). Em geral, nestes dados, esta porcentagem é baixa (62/365=17%) e, portanto, espera-se que não seja muito grande a dependência entre os eventos de interesse.

Segundo os resultados da Tabela 4.14, em relação às covariáveis, a porcentagem de óbito em relação aos pacientes que apresentaram recaída também é baixa. Considerando tratamento e sexo, esta proporção fica entre 17% e 18%. Por idade, esta proporção aumenta nas maiores faixas etárias:

de 51 a 65 anos é 18,6% e para pacientes maiores de 66 anos é 24,3%. Com relação à cidade, os resultados são bem diferentes: para pacientes tratados em Manizales, a porcentagem de óbito em relação aos pacientes que apresentaram recaída é 12,7%; em Monteria 27,2%; em Rionegro 8,2%; em Sincelejo 23,3% e, em Tunja, 16,3%.

Para uma análise descritiva de cada variável em relação aos eventos de interesse foram construídas as estimativas Cópula-Gráfico, para análise de sobrevivência marginal para a progressão, e Kaplan Meier, para o óbito.

Para o estimador Cópula-Gráfico, inicialmente é obtida a estimativa do parâmetro de dependência α das cópulas anteriormente descritas e seu correspondente coeficiente de concordância τ. Os resultados são apresentados na Tabela 4.15 e nos gráficos da Figura 4.13, que correspondem as raízes das funções de estimação de α para cada cópula, como descrito no Capítulo 2.

Tabela 4.15: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$*

Cópula	$\hat{\alpha}$(se)	$\hat{\tau}$ (se)
Família A - Cópula de Clayton	0,586 (0,142)	0,227 (0,042)
Família B	1,559 (0,022)	-0,282 (0,018)
Família C - Cópula de Frank	1,387 (0,295)	0,151 (0,031)
Família D - Cópula de Gumbel	1,141 (0,038)	0,124 (0,029)

No estudo dos pacientes com doença renal, a cópula de Clayton detectou uma leve dependência positiva entre a recaída e o óbito, já as cópulas Frank e Gumbel detectaram uma dependência menor ainda, porém, estes resultados parecem consistentes com a proporção de morte entre os pacientes que apresentaram recaída (17%). No entanto, a dependência detectada com a cópula da família B foi em sentido oposto à dependência segundo as outras cópulas.

Após a estimação do parâmetro de dependência α, o estimador Cópula-Gráfico é calculado para a função de sobrevivência marginal do tempo até recaída. As curvas estimadas são similares para as cópulas estudadas, segundo as Figuras 4.14, e ao serem sobrepostas como mostrado no primeiro gráfico da Figura 4.15, pode ser concluído que não há grandes diferenças, principalmente entre as estimativas das cópulas de Clayton, Frank e Gumbel. Para a recaída, a probabilidade de sobrevivência decresce lentamente ao longo do tempo.

A curva de sobrevivência marginal estimada para o óbito (segundo gráfico da Figura 4.15)

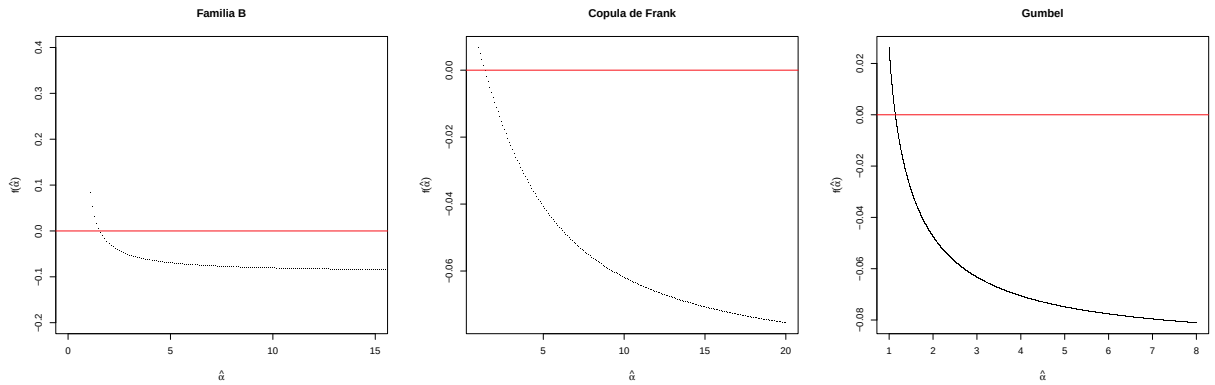

Figura 4.13: *Gráficos da equação de estimação de α, para família B, cópula de Frank e cópula de Gumbel.*

apresenta um comportamento similar à curva estimada da recaída, com uma falha que ressalta no final do período de observação. É importante mencionar que a alta proporção de censura, como é visualizado no gráfico à direita da Figura 4.15, é devida principalmente ao pouco tempo de seguimento dos pacientes com doença renal (dois anos).

No estudo dos pacientes com doença renal, as curvas de sobrevivência marginal parecem ter um comportamento similar quando analisadas em relação às covariáveis.

Para a análise descritiva das funções de sobrevivência marginais segundo as covariáveis foi usada a cópula de Clayton para o estimador Cópula-Gráfico da sobrevivência marginal da recaída, e Kaplan-Meier para o óbito. Os resultados segundo as outras cópulas não são apresentados aqui dado que são semelhantes (principalmente das cópulas Frank e Gumbel).

A Tabela 4.16 apresenta as estimativas obtidas dos parâmetros de dependência α e concordância τ da cópula de Clayton segundo as covariáveis. Pode-se observar, a partir o τ de Kendall estimado, que a dependência detectada entre recaída e óbito foi maior, na maioria dos casos, que a dependência geral (Tabela 4.15, cópula de Clayton, $\hat{\tau} = 0,23$), oscilando entre 0,22 e 0,29.

A dependência entre recaída e óbito foi levemente maior para os pacientes em tratamento de hemodiálise que em diálise peritoneal. Em relação ao sexo, a diferença da dependência detectada também foi pequena, 0,23 para mulheres e 0,22 para homens. Para a idade, a maior dependência dos eventos de interesse foi para os menos idosos (0,29) e a menor para os mais idosos (0,22). Pode-se verificar que há maiores diferenças quando comparadas as cidades, por exemplo para os pacientes de Tunja poderia se dizer que não há dependência entre recaída e óbito pois $\hat{\tau} = 0,08$.

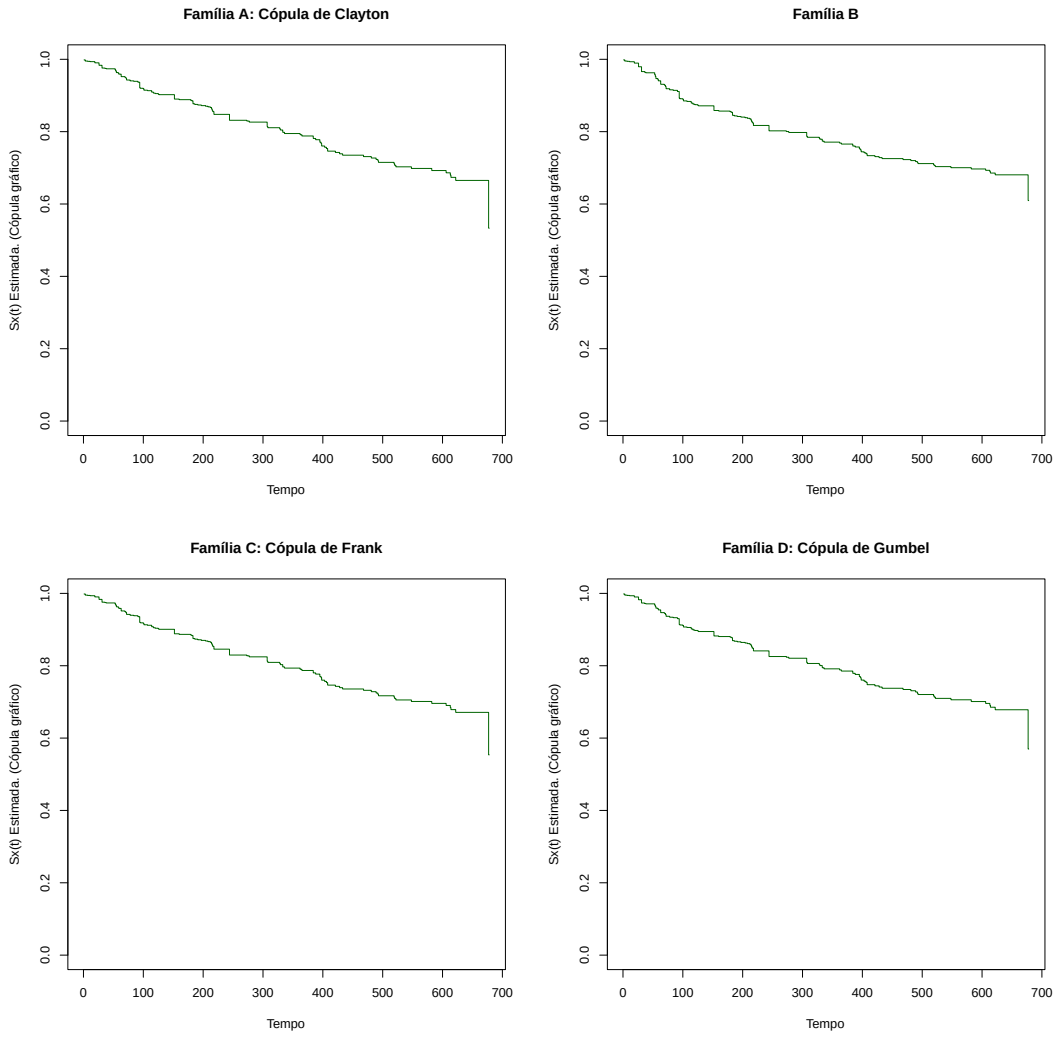

Figura 4.14: *Função de sobrevivência marginal da recaída (obtida pelo estimador Cópula-Gráfico).*

No entanto, para os pacientes das outras cidades, encontrou-se uma leve dependência, sendo a maior para Sincelejo com $\hat{\tau} = 0,27$.

Na Figura 4.16 são comparadas as funções de sobrevivência marginal de recaída e óbito em relação ao tipo de diálise. Quando analisada a recaída, a probabilidade de sobrevida é maior para pacientes em tratamento de hemodiálise do que para pacientes em tratamento de diálise

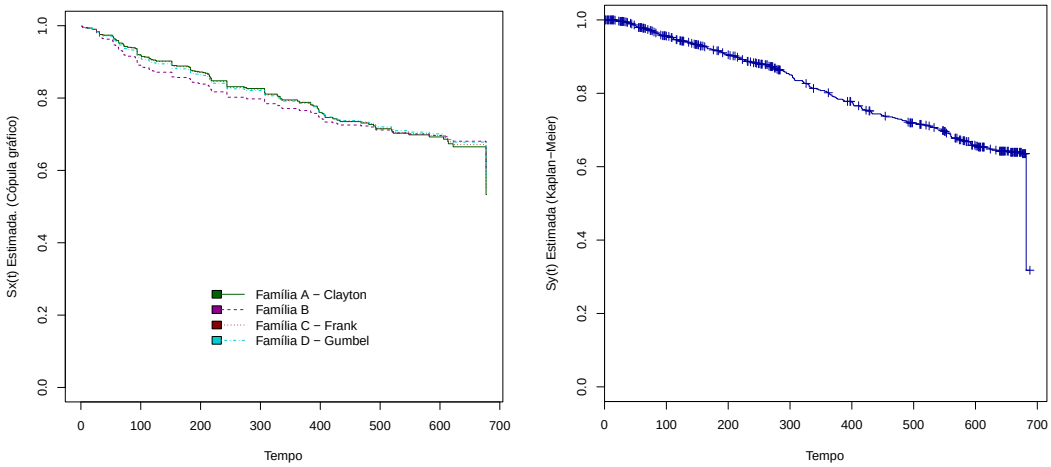

Figura 4.15: *Funções de sobrevivência marginais estimadas de recaída (à esquerda) e óbito (à direita), obtidas pelo estimador Cópula-Gráfico e Kaplan-Meier, respectivamente.*

peritoneal e, visualmente as curvas parecem ser paralelas. Quando analisada a sobrevivência segundo o óbito, as curvas estimadas (Kaplan-Meier) cruzam-se quase até a metade do tempo, depois elas apresentam diferenças, sendo menor a probabilidade de sobrevida dos pacientes em tratamento de diálise peritoneal, que decresce mais rápido.

Na Figura 4.16 são comparadas as funções de sobrevivência marginal de recaída e óbito em relação ao tipo de diálise. Quando analisada a recaída, a probabilidade de sobrevida é maior para pacientes em tratamento de hemodiálise, do que pacientes em tratamento de diálise peritoneal, visualmente as curvas parecem ser paralelas. Quando analisada a sobrevivência segundo o óbito, as curvas estimadas (Kaplan-Meier) cruzam-se quase até a metade do tempo, depois elas apresentam diferenças, sendo menor a probabilidade de sobrevida dos pacientes em tratamento de diálise peritoneal a qual decresce mais rápido.

Na Figura 4.17 são apresentadas as curvas de sobrevivência estimadas em relação ao sexo. Com relação a recaída, as curvas estimadas vão separando-se no recorrer do período, sendo menor a curva de sobrevivência do sexo feminino. Já no caso do óbito, as curvas estimadas não apresentam diferenças visualmente grandes indicando que as probabilidades de sobrevivência são quase iguais, em geral, até metade do período de estudo. No final do período há claramente a

Tabela 4.16: *Parâmetros de dependência e concordância estimados, $\hat{\alpha}$ e $\hat{\tau}$ segundo as covariáveis para a cópula de Clayton*

Covariável	Cópula de Clayton	
	$\hat{\alpha}$ (se)	$\hat{\tau}$ (se)
Tratamento		
0 Diálises peritoneal	0,619 (0,306)	0,236 (0,088)
1 Hemodiálises	0,644 (0,173)	0,244 (0,049)
Sexo		
0 Feminino	0,608 (0,221)	0,233 (0,065)
1 Masculino	0,569 (0,187)	0,221 (0,057)
Idade do paciente		
1 De 1 a 30 anos	0,806 (0,496)	0,287 (0,125)
2 De 31 a 50 anos	0,690 (0,333)	0,256 (0,092)
3 De 51 a 65 anos	0,649 (0,241)	0,245 (0,068)
4 De 66 a 97 anos	0,577 (0,318)	0,224 (0,095)
Cidade		
1 Manizales	0,326 (0,261)	0,287 (0,125)
2 Monteria	0,502 (0,298)	0,200 (0,096)
3 Rionegro	0,646 (0,831)	0,244 (0,247)
4 Sincelejo	0,741 (0,546)	0,270 (0,134)
5 Tunja	0,186 (0,326)	0,085 (0,137)

observação de óbito de um paciente masculino.

Na Figura 4.18 são comparadas as curvas de sobrevivência marginais por idade em relação ao tempo de recaída e óbito. Note que conclusões opostas são obtidas para pessoas maiores de 64 anos, sendo que quando apresentam recaída a probabilidade de sobrevivência é a maior dos grupos etários. Mas quando o evento observado é óbito, a curva estimada de sobrevivência é a menor. Já para os mais jovens (menores de 30 anos), se observa uma situação contrária, ou seja, sobrevivência menor quando observado o tempo até recaída e maior quando observado o tempo até o óbito. Também pode-se verificar que as curvas de sobrevivência marginais da idade referentes à recaída são mais paralelas entre suas categorias ao longo do tempo de estudo que as curvas em relação ao óbito. Do gráfico à direita, pode-se observar que a faixa de idade do paciente que teve óbito ao final do período é de 51 e 65 anos.

Finalmente, ao comparar a sobrevivência marginal pela cidade do paciente, segundo a Figura

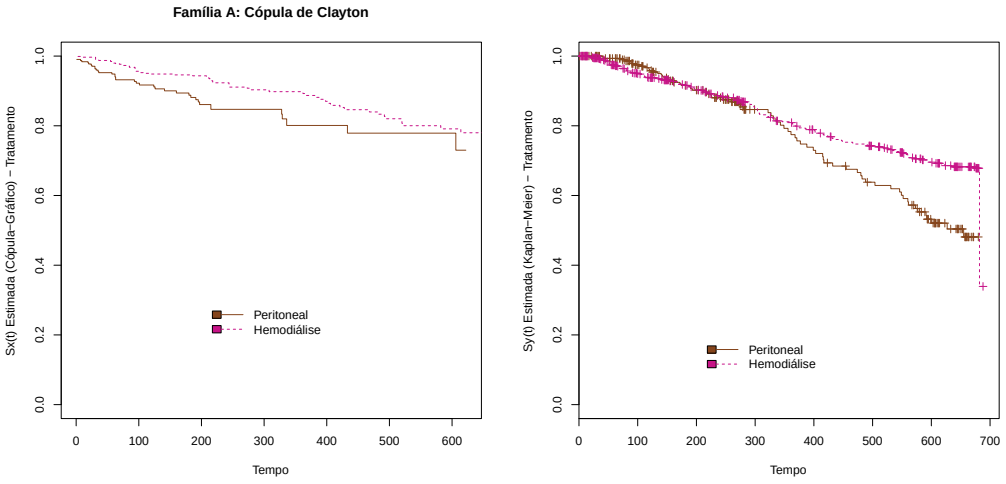

Figura 4.16: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o tratamento, obtidas pelos estimadores Cópula-Gráfico (à esquerda) e Kaplan-Meier (à direita).*

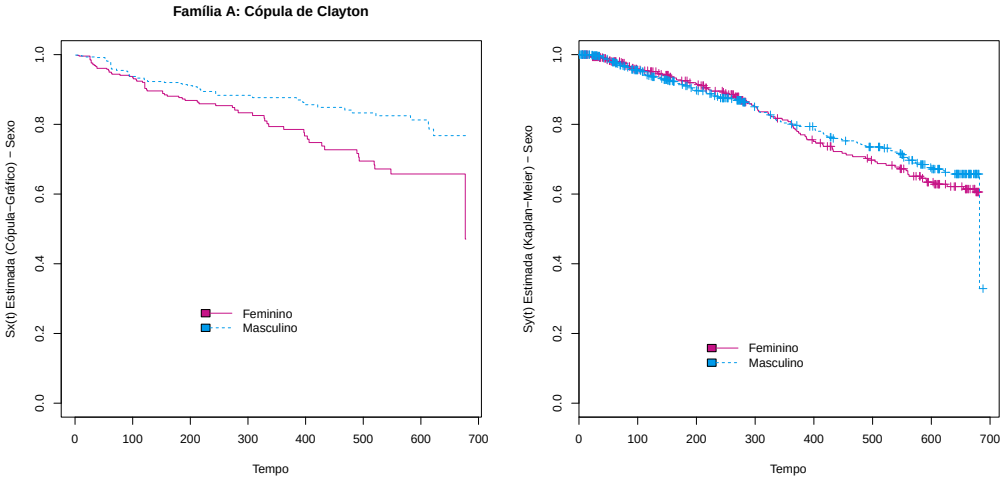

Figura 4.17: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo o sexo, obtidas pelos estimadores Cópula-Gráfico (à esquerda) e Kaplan-Meier (à direita).*

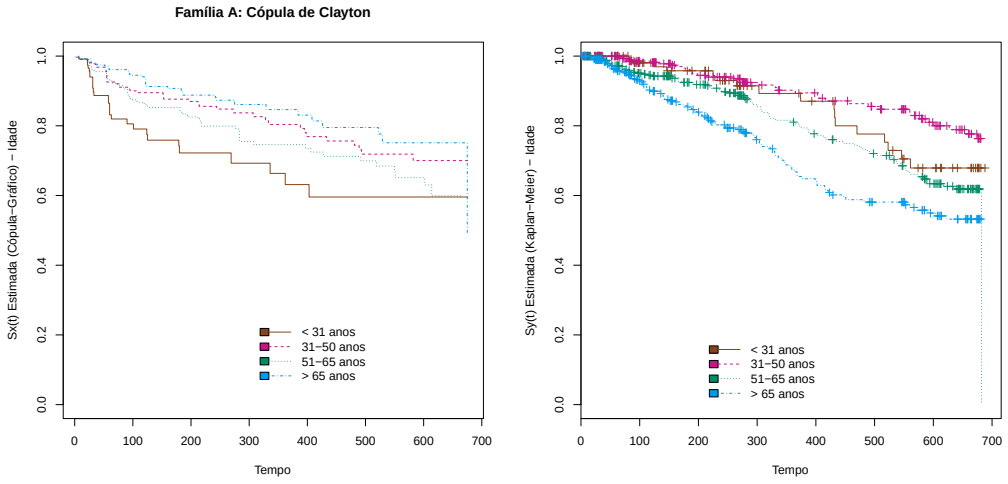

Figura 4.18: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo a idade, obtidas pelos estimadores Cópula-Gráfico (à esquerda) e Kaplan-Meier (à direita).*

4.19, observa-se que em relação ao tempo de recaída as curvas de sobrevivência estimadas dos pacientes de Manizales e Tunja apresentam um comportamento paralelo durante todo o período. Quando comparadas com as curvas estimadas dos pacientes de outras cidades (Monteria, Rionegro e Sincelejo), elas parecem muito diferentes. Pode ser visualizado também que a menor probabilidade de sobrevida é dos pacientes de Rionegro, já as maiores probabilidades correspondem aos pacientes de Monteria e Manizales, cujas curvas estimadas se cruzam em torno de 350 ou 400 dias.

Em relação ao tempo de óbito, segundo a Figura 4.19, pode-se dizer que existe uma nítida diferença entre as curvas estimadas. As cidades Sincelejo e Rionegro apresentam um comportamento similar ao longo do tempo, porém, as curvas são distantes das curvas de sobrevivência correspondentes aos pacientes das cidades Manizales, Monteria e Tunja. A probabilidade de sobrevivência maior corresponde aos pacientes de Manizales e a menor, aos pacientes de Rionegro e Sincelejo, ainda tendo curvas que se cruzam entre si ao longo do tempo.

Sendo de interesse determinar a sobrevivência dos pacientes com doença renal após 245 dias sem apresentar recaída da doença, são obtidos os gráficos da Figura 4.20. As estimativas semiparamétricas foram calculadas como a probabilidade de sobrevivência condicional quando não há recaída no tempo $x = 245$, construídas a partir das funções bivariadas das cópulas de Clayton,

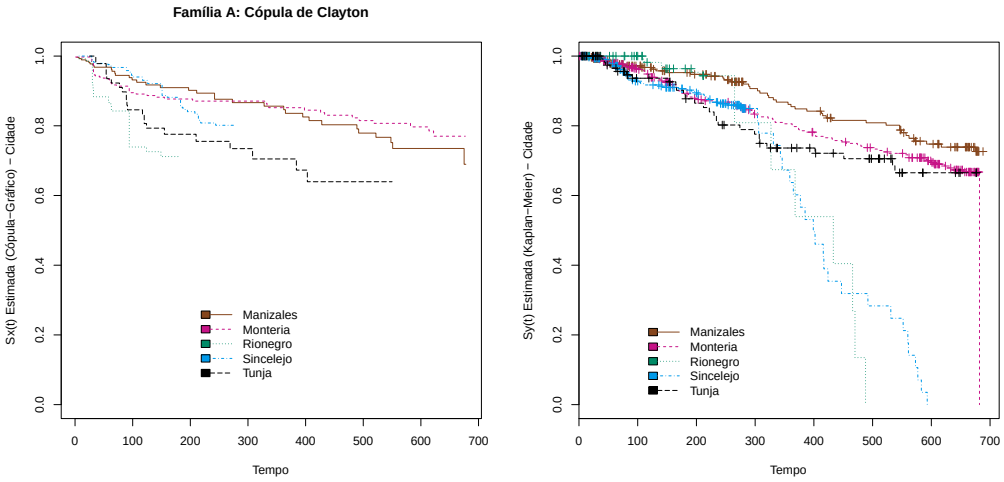

Figura 4.19: *Funções de sobrevivência marginais estimadas de recaída e óbito, segundo a cidade, obtidas pelos estimadores Cópula-Gráfico (à esquerda) e Kaplan-Meier (à direita).*

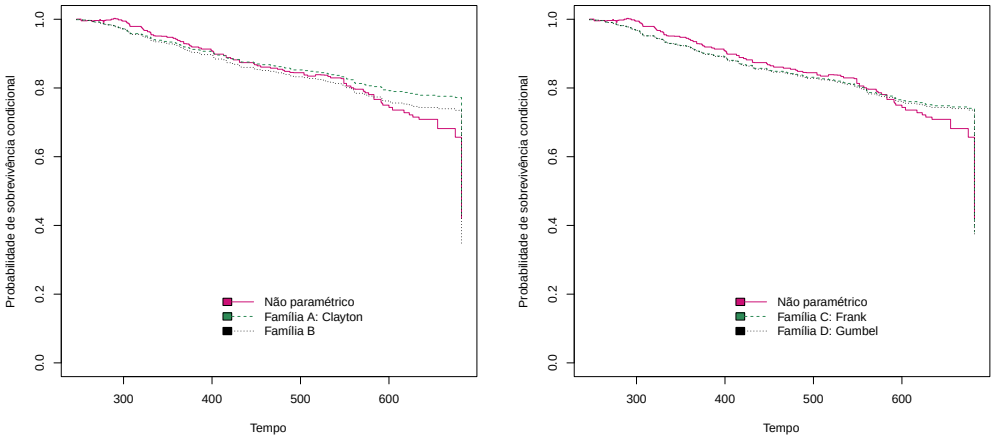

Figura 4.20: *Função de sobrevivência condicional quando não há recaída em 245 dias após inicio do tratamento de diálise (cópulas Clayton e família B à esquerda, cópulas Frank e Gumbel à direita).*

Família B (gráfico à esquerda da Figura 4.20), Frank e Gumbel (gráfico à direita da Figura 4.20) na região observável ($x < y$). Pode-se observar que a cópula da família B apresentou estimativas um pouco acima das demais cópulas, porém, visualmente as diferenças parecem não ser muito grandes. Comparando a estimativa não paramétrica com as semiparamétricas (segundo as cópulas estudadas), a estimativa não paramétrica fornece valores maiores até aproximadamente 500 dias e, depois desse instante, a estimativa fica abaixo das obtidas com estimadores semiparamétricos.

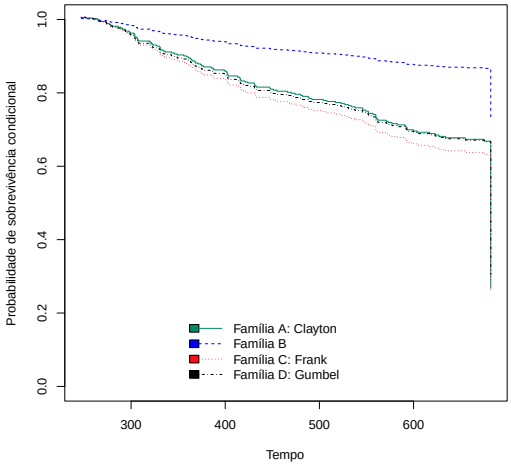

Figura 4.21: *Função de sobrevivência condicional de um paciente que apresentou recaída aos 214 dias após o tratamento, e continua vivo 50 dias depois.*

Também foi calculada a probabilidade de sobrevivência de um paciente que apresentou recaída e continua vivo num tempo previamente fixado, como outra aplicação da função de sobrevivência bivariada. Na Figura 4.21 é apresentada esta estimativa quando a recaída ocorreu após 214 dias de iniciado o tratamento e em que 50 dias depois o paciente estava vivo.

Na estimativa desta probabilidade condicional, vale ressaltar que a cópula da família B fornece estimativas acima das obtidas com as cópulas de Clayton, Frank e Gumbel, o que parece concordar com o coeficiente τ de Kendal estimado para esta cópula (-0,28), que foi oposto aos obtidos para as outras cópulas estudadas. As estimativas mais próximas são as correspondentes as cópulas de Clayton e Gumbel.

Depois de uma visão geral das probabilidades de sobrevivência, a seguir, é feita a análise

inferencial dos dados de doença renal por meio do processo doença-morte.

4.2.2 Processo doença-morte

O modelo a ser estimado é análogo ao modelo da aplicação anterior, que é dado por:

$$\begin{aligned}
\lambda_1(x|\gamma, z) &= \gamma\lambda_{01}(x)\exp\{\boldsymbol{\beta}_1^\top z\}, \quad x > 0, \\
\lambda_2(y|\gamma, z) &= \gamma\lambda_{02}(y)\exp\{\boldsymbol{\beta}_2^\top z\}, \quad y > 0, \\
\lambda_{12}(y|\gamma, x, z) &= \gamma\lambda_{03}(y)\exp\{\boldsymbol{\beta}_3^\top z\}, \quad 0 < x < y,
\end{aligned} \tag{4.2}$$

em que $\gamma \sim \Gamma(1/\theta, 1/\theta)$ é a variável de fragilidade, $\lambda_{01}(x)$, $\lambda_{02}(y)$ y $\lambda_{03}(y)$ representam as funções de risco basais desconhecidas de recaída, óbito e óbito após recaída respectivamente, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ e $\boldsymbol{\beta}_3$ são os vetores dos coeficientes de regressão para a recaída, o óbito e o óbito após recaída, respectivamente, e z é a matriz de covariáveis, composta pelas variáveis binárias:

- Tratamento

$$z_{1i} = \begin{cases} 1 = \text{Hemodiálises}, \\ 0 = \text{Diálises peritoneal}. \end{cases}$$

- Sexo

$$z_{2i} = \begin{cases} 1 = \text{Masculino}, \\ 0 = \text{Feminino}. \end{cases}$$

- Idade do paciente

$$z_{3i} = \begin{cases} 1 = \text{De 31 a 50 anos}, \\ 0 = \text{Caso contrario}. \end{cases}$$

$$z_{4i} = \begin{cases} 1 = \text{De 51 a 65 anos}, \\ 0 = \text{Caso contrario}. \end{cases}$$

$$z_{5i} = \begin{cases} 1 = \text{De 66 a 97 anos}, \\ 0 = \text{Caso contrario}. \end{cases}$$

- Cidade

$$z_{6i} = \begin{cases} 1 = \text{Monteria}, \\ 0 = \text{Caso contrario}. \end{cases}$$

$$z_{7i} = \begin{cases} 1 = \text{Rionegro}, \\ 0 = \text{Caso contrario}. \end{cases}$$

$$z_{8i} = \begin{cases} 1 = \text{Sincelejo}, \\ 0 = \text{Caso contrario}. \end{cases}$$

$$z_{9i} = \begin{cases} 1 = \text{Tunja}, \\ 0 = \text{Caso contrario}. \end{cases}$$

com $i = 1 \cdots 137$.

Os resultados do ajuste do modelo geral (4.2), como descrito no Capítulo 3, são apresentados na Tabela 4.17. Das quatro covariáveis usadas, o sexo não resultou ser significativo em nenhuma das situações, assim como a faixa etária de 31 a 50 anos. No entanto, o grupo da maior faixa etária e a cidade Sincelejo apresentaram um efeito significativo para as três situações: recaída, óbito sem recaída e óbito após recaída.

Tabela 4.17: *Estimativas modelo geral doença-renal*

	Tempo até recaída			Tempo até óbito sem recaída			Tempo até óbito após recaída		
Covariáveis	Estimativa	EP	P valor	Estimativa	EP	P valor	Estimativa	EP	P valor
z_1 (Tratamento)	0,432	0,133	0,001	-0,373	0,166	0,025	0,043	0,316	0,891
z_2 (Sexo)	0,019	0,109	0,862	-0,157	0,155	0,311	-0,176	0,261	0,501
z_3 (31-50a vs 1-30a)	0,070	0,160	0,664	-0,140	0,278	0,614	-0,165	0,425	0,698
z_4 (51-65a vs 1-30a)	0,046	0,150	0,760	0,531	0,235	0,024	0,303	0,381	0,425
z_5 (66-97a vs 1-30a)	-0,515	0,173	0,003	0,978	0,231	0,000	0,829	0,383	0,031
z_6 (Monte vs Maniz)	-0,464	0,143	0,001	0,275	0,205	0,179	0,721	0,333	0,030
z_7 (Rione vs Maniz)	0,484	0,174	0,006	0,332	0,375	0,376	0,928	0,558	0,096
z_8 (Since vs Maniz)	-0,324	0,157	0,039	0,851	0,225	0,000	1,570	0,357	0,000
z_9 (Tunja vs Maniz)	0,202	0,181	0,263	0,651	0,300	0,030	0,533	0,451	0,237
θ	0,048	0,217	0,825						

O tratamento foi significativo para o tempo até recaída e o tempo até o óbito. Em relação a cidade, Monteria e Rionegro apresentaram efeito significativo para a recaída e o óbito após recaída.

Dos resultados obtidos, pode-se verificar que o coeficiente de regressão estimado para óbito sem recaída de pacientes com tratamento de hemodiálise é negativo e em direção oposta aos outros dois, o que parece consistente com o fato de que houve uma proporção menor de pacientes com óbito sem recaída (67,2%) do que com recaída (80,2%) e com óbito após recaída (80,6%) no grupo de tratamento com hemodiálise.

Também para o grupo etário mais idoso (maiores de 67 anos) o coeficiente de regressão estimado para recaída foi negativo e para óbito sem recaída e óbito após recaída foi positivo, indicando que, nos dados de doença renal, houve menor proporção de pacientes idosos com recaída (18,5%) que com óbito (46,3%) e óbito após recaída (29%). Esse mesmo padrão acontece quando considera-se a variável cidade, em que para a categoria significativa (Sincelejo), o coeficiente de regressão estimado para recaída foi negativo e positivo para óbito e óbito após recaída.

Se o risco for determinado segundo as estimativas dos coeficientes de regressão (Wienke, 2011), então o risco de recaída é maior para pacientes em tratamento de hemodiálise. Quanto ao sexo, os homens têm maior risco de recaída. Comparando os grupos de idade com os mais jovens, o maior risco corresponde às idades entre 31 e 65 anos. Em relação à cidade, comparando com Manizales, as cidades que apresentaram maior risco foram Rionegro e Tunja.

No caso do risco de óbito sem recaída da doença renal, o mesmo é menor para tratamento com hemodiálise, para sexo masculino, para a faixa etária de 31 a 50 anos e para a cidade Manizales, o que concorda com a análise descritiva feita a partir das curvas de sobrevivência estimadas para o óbito.

Quando os eventos recaída e óbito são observados, os maiores riscos em geral correspondem a todas as categorias das covariáveis comparadas ao grupo de referência, exceto para sexo, em que as mulheres (grupo de referência) têm maior risco que os homens, e para o grupo de idade de 31 a 50 anos, que tem menor risco quando comparado com os mais jovens.

Com os resultados apresentados na Tabela 4.17 o ajuste do modelo geral (4.2) fica dado por

$$\begin{aligned}
\hat{\lambda}_1(x|\hat{\gamma},\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{01}(x)\exp\{0,43z_1+0,02z_2+0,07z_3+0,05z_4-0,51z_5-0,46z_6+0,48z_7-0,32z_8+0,20z_9\},\\
\hat{\lambda}_2(y|\hat{\gamma},\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{02}(x)\exp\{-0,37z_1-0,16z_2-0,14z_3+0,53z_4+0,98z_5+0,27z_6+0,33z_7+0,85z_8+0,65z_9\} \text{ e}\\
\hat{\lambda}_3(y|\hat{\gamma},x\boldsymbol{z}) &= \hat{\gamma}\hat{\lambda}_{03}(x)\exp\{0,04z_1-0,18z_2-0,16z_3+0,30z_4+0,83z_5+0,72z_6+0,93z_7+1,57z_8+0,53z_9\}.
\end{aligned}$$

De maneira geral, os resultados obtidos estão de acordo com o esperado pelos pesquisadores, pois sabe-se que embora o tipo de tratamento (hemodiálise ou diálise peritoneal) depende do paciente,

a hemodiálise resulta ser uma boa alternativa em muitas situações. Com relação às cidades, sabe-se que Manizales tem uma maior trajetória no tratamento de pacientes renais, no entanto, em Rionegro a unidade renal é mais recente e Sincelejo possui condições menos favoráveis.

Conforme já mencionado anteriormente, esse é um estudo ainda em andamento e espera-se que futuramente se tenha mais tempo de seguimento e mais informações sobre óbito e recaída dos pacientes. Dessa maneira, essa é uma análise preliminar dos dados, porém, já é bastante informativa e útil para os pesquisadores.

Com relação ao parâmetro de fragilidade, a estimativa obtida do modelo geral gerou uma variância da fragilidade pequena, 0,048 (0,217), indicando uma associação fraca entre recaída e óbito, o que é coerente com a baixa proporção de óbito entre os pacientes que apresentaram recaída (17%). É também um resultado esperado considerando que o tempo de observação dos pacientes com esta doença foi curto.

Neste caso, o nível descritivo (P valor) também foi obtido através do teste assintótico de Wald assumindo que, sob H_0, a estatística do teste tem distribuição chi-quadrado com um grau de liberdade. Conforme já discutido, esse resultado pode não ser válido e, portanto, deve-se ter cautela ao avaliar este resultado. Essa questão da distribuição assintótica do teste de Wald para o parâmetro da variância da fragilidade neste modelo de doença-morte é um tópico que merece ser estudado com mais cuidado em estudos futuros.

CAPÍTULO 5

Discussão

Atualmente, têm-se tornado cada vez mais comum na área médica estudos que envolvem o tempo até a ocorrência de mais de um evento. Dependendo da estrutura do problema, tais estudos podem ser enquadrados na estrutura de riscos competitivos. Entretanto, com frequência cada vez maior, surgem dados com estrutura de riscos semicompetitivos, quando um dos eventos impede a ocorrência do outro, porém, o inverso não ocorre.

A primeira dificuldade que surge na análise de dados com estrutura de riscos semicompetitivos é a obtenção de estimadores da função de sobrevivência para sua análise descritiva. Dada a possível dependência entre ambos eventos, o estimador Kaplan-Meier não é apropiado para estimar a função de sobrevivência marginal quando observado o evento intermediário. A proposta de Lakhal e Abdous (2008) de utilizar o estimador Cópula-Gráfico nesta situação parece ter bom resultado, e sua expressão com forma fechada facilita os cálculos. O uso de cópulas da família Arquimediana se mostrou bastante conveniente para estimar a função de sobrevivência bivariada quando há dependência entre os eventos intermediário (recaída) e terminal (óbito). De maneira geral, para fazer a análise descritiva inicial de dados, a abordagem de cópulas é uma boa alternativa. A seleção de uma cópula, em particular da família de cópulas Arquimedianas, pode influenciar as estimativas. Embora as diferenças entre as cópulas observadas nas aplicações não sejam significativas, seria interessante avaliar de alguma forma qual é a cópula mais conveniente

para uma certa situação, porém, não há ainda nenhuma metodologia proposta para isso.

A abordagem por meio de cópulas, no entanto, não permite a inclusão de covariáveis e, em problemas aplicados, é de interesse analisar os dados por meio de modelos de regressão. Uma alternativa para incorporar covariáveis no modelo é a proposta de Xu et al. (2010), que considera a modelagem das funções de intensidade de transição de um processo de doença-morte com inclusão de fragilidade compartilhada para incorporar a dependência entre os tempos observados. Os autores propuseram estimação por máxima verossimilhança (marginal) e o processo de estimação nas aplicações foi rápido, com convergência em poucas iterações. A obtenção de estimativa da variância dos estimadores também é fácil de ser obtida, dado que é possível calcular analiticamente as derivadas do vetor escore. Neste caso, as funções de taxa de falha basais foram estimadas aproximando-as a uma função escada que depende do tamanho da amostra, porém, seria interessante impor uma forma paramétrica para estas funções e avaliar o melhor ajuste.

Xu et al. (2010) utilizam o teste de Wald para testar o parâmetro associado com a variância do efeito aleatório, entretanto esse teste precisa ser melhor avaliado pois o valor do parâmetro na hipótese nula está na fronteira do espaço paramétrico. O processo doença-morte com inclusão de fragilidade ainda é pouco explorado na literatura, e seria bom estender tal metodologia para outras distribuições da variável de fragilidade. Além disso, poderia-se estudar a inclusão de covariáveis dependentes do tempo.

APÊNDICE A

Apêndice

A.1 τ de Kendall para cópula de Frank

O τ de Kendall para cópula de Frank é dado por

$$
\begin{aligned}
\tau &= 4\int_0^1 \frac{\ln(e^{-\alpha}-1)-\ln(e^{-\alpha u}-1)}{\alpha e^{-\alpha u}(e^{-\alpha u}-1)^{-1}}du + 1 \\
&= 1-\frac{4}{\alpha}\int_0^1 \frac{\ln(e^{-\alpha u}-1)-\ln(e^{-\alpha}-1)}{e^{-\alpha u}(e^{\alpha u}-1)^{-1}}du \\
&= 1-\frac{4}{\alpha}\int_0^1 e^{\alpha u}(e^{-\alpha u}-1)\ln\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)du \\
&= 1-\frac{4}{\alpha}\int_0^1 (1-e^{\alpha u})\ln\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)du,
\end{aligned}
$$

Fazendo $t = \alpha u$, $dt = \alpha du$,chega-se a

$$\begin{aligned}\tau &= 1 - \frac{4}{\alpha^2}\int_0^\alpha (1-e^t)\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)dt \\ &= 1 - \frac{4}{\alpha^2}\left[\int_0^\alpha \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)dt - \int_0^\alpha e^t \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)dt\right]. \qquad \text{(A.1)}\end{aligned}$$

Resolvendo por partes a primeira integral de (A.1), tem-se $u = \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)$, $dv = dt$, $du = \frac{e^{-t}}{1-e^{-t}}dt$, $v = t$,

$$\int_0^\alpha \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)dt = t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)\Bigg|_0^\alpha - \int_0^\alpha \frac{t}{e^{-t}-1}dt.$$

Resolvendo por partes a segunda integral de (A.1), chega-se a $u = \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)$, $dv = e^t dt$, $du = -\frac{e^{-t}}{1-e^{-t}}dt$, $v = e^t$,

$$\begin{aligned}\int_0^\alpha e^t \ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)dt &= e^t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)\Bigg|_0^\alpha + \int_0^\alpha \frac{e^t e^{-t}}{e^{-t}-1}dt \\ &= e^t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right)\Bigg|_0^\alpha + \int_0^\alpha \frac{e^t}{1-e^t}dt \\ &= \left[e^t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right) - \ln(1-e^t)\right]\Bigg|_0^\alpha.\end{aligned}$$

Juntando as duas integrais e voltando à expressão (A.1), tem-se finalmente que

$$\tau = 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}\left[t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right) - e^t\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right) + \ln(1+e^t)\right]\Bigg|_0^\alpha,$$

e, sendo que $\ln\left(\frac{e^{-t}-1}{e^{-\alpha}-1}\right) = \ln\left(\frac{1-e^t}{e^t(e^{-\alpha}-1)}\right)$, então

$$\begin{aligned}
\tau &= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}\left[t\ln\left(\frac{1-e^t}{e^t(e^{-\alpha}-1)}\right) - e^t\ln\left(\frac{1-e^t}{e^t(e^{-\alpha}-1)}\right) + \ln(1+e^t)\right]\Bigg|_0^\alpha \\
&= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}[(t-e^t+1)\ln(1-e^t) + (e^t-t)\ln(e^t(e^{-\alpha}-1))]|_0^\alpha.
\end{aligned}$$

Avaliando nos limites α e 0, encontra-se uma forma indeterminada, obtendo

$$\tau = 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}[(\alpha-e^\alpha+1)\ln(1-e^\alpha) + (e^\alpha-\alpha)\ln(1-e^\alpha) - FI - \ln(e^{-\alpha}-1)].$$

Resolvendo o limite da forma indeterminada pelo teorema de L'hôpital, tem-se

$$\begin{aligned}
FI = \lim_{t\to 0}(t-e^t+1)\ln(1-e^t) &= \lim_{t\to 0}\frac{\ln(1-e^t)}{\frac{1}{t-e^t+1}} = \lim_{t\to 0}\frac{-\frac{e^t}{1-e^t}}{-\frac{1-e^t}{(t-e^t+1)^2}} = \lim_{t\to 0}\frac{e^t(t-e^t+1)^2}{(1-e^t)^2} \\
&= \lim_{t\to 0}\frac{e^t 2(t+1-e^t)(1-e^t) + e^t(t+1-e^t)^2}{-2(1-e^t)e^t} \\
&= 0.
\end{aligned}$$

Portanto, o coeficiente τ é dado por

$$\begin{aligned}
\tau &= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}[(\alpha-e^\alpha+1)\ln(1-e^\alpha) + (e^\alpha-\alpha)\ln(1-e^\alpha) - 0 - \ln(e^{-\alpha}-1)] \\
&= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}[\ln(1-e^\alpha)(\alpha-e^\alpha+1+e^\alpha-\alpha) - \ln(e^{-\alpha}-1)] \\
&= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}\left[\ln\left(\frac{1-e^\alpha}{e^{-\alpha}-1}\right)\right] \\
&= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha^2}\left[\ln\left(\frac{e^\alpha(1-e^\alpha)}{1-e^\alpha}\right)\right] \\
&= 1 - \frac{4}{\alpha^2}\int_0^\alpha \left(\frac{t}{e^{-t}-1}\right)dt - \frac{4}{\alpha},
\end{aligned}$$

em que $\int_0^\alpha \frac{t}{e^{-t}-1}dt$ corresponde com a função de Debye de ordem dois.

Referências Bibliográficas

Aalen, O., Borgan, O. e Gjessing, H. (2008). *Survival and Event History Analysis*, Springer, New York.

Alsina, C., Frank, M. e Schweizer, B. (2006). *Associative Functions Triangular Norms and Copulas*, World Scientific Publishing, Singapure.

Breslow, N. (1972). Contribuição para a discusssão do artigo de d. r. cox, *Journal of the Royal Statistical Society, Ser.* **B 34**: 216–217.

Chen, Y. (2012). Maximum likelihood analysis of semicompeting risks data with semiparametric regression models, *Lifetime Data Analysis* **18**: 36–57.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**: 141–151.

Copelan, E., Biggs, J., Thompson, J., Crilley, P., Szer, J., Klein, J. P., Kapoor, N., Avalos, B., Cunningham., Atkinson, K., Downs, K., Harmon, G. S., Darly, M. B., Brodsky, I., Bulova, S. e Tutschka, P. (1991). Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with bu/cy., *Blodd* **78**: 838–843.

Cox, D. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Ser.* **B 34 Nº 2**: 187–220.

CRC (2011). *Clínica Renal Colombiana(CRC). Riesgo de progresión en Insuficiencia Renal Crónica [em linha].*, http://www.clinicarenal.org/index.php?option=com_content&view=article&id=109&lang=en, Bogotá.

Day, R., Bryant, J. e Lefkopoulou, M. (1997). Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic factors, *Biometrika* **84**: 45–56.

Fine, J., Jiang, H. e Chappell, R. (2001). On semicompeting risk., *Biometrika* **88**: 907–919.

Fix, E. e Neyman, J. (1951). A simple stochastic model for recovery, relapse, death and loss of patients, *Human Biology* **23**: 205–241.

Frank, M. (1979). On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, *Aequationes Math* **19**: 194–226.

García, D. L., Ordoñez, I. E., Muñoz, J. E., Rodríguez, K., Robayo, A., Rodríguez, R., Ruiz, M., Sanabria, M., Gómez, R. e Barreto, F. (2005). *Guía para el manejo de la enfermedad renal crónica -ERC- Basada en la Evidencia. Colombia*, Fundación para la investigación y Desarrollo de la Salud y la Seguridad Social, Bogotá.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusiers dimensions, *Publ Inst Statist Univ Paris* **9**: 171–173.

Hougaard, F. (2000). *Analysis of multivariate survival data*, Springer.

Hsieh, J., Wang, W. e Ding, A. (2008). Regression analysis based on semicompeting risk data., *Journal of the Royal Statistical Society, Ser.* **B 70**: 3–20.

Jiang, H., Fine, J., Kosorok, R. e Chappell, R. (2005). Pseudo self-consistent estimation of a copula model with informative censoring., *Scandinavian Journal of Statistics* **32**: 1–20.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Kimeldorf, G. e Sampson, A. (1975). Uniform representations of bivariate distributions, *Communications in Statistics* **4**: 617–627.

Klein, J. e Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edn, Springer.

Lakhal, L.and Rivest, L. e Abdous, B. (2008). Estimating survival and association in a semicompeting risk model, *Biometrics* **64**: 180–188.

Li, R. e Peng, L. (2011). Quantile regression for left-truncated semicompeting risks data, *Biometrics* **67**: 701–710.

Lin, D. Y. e Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring, *Biometrika* **80**: 573–582.

Murphy, S. (1995). Asymptotic theory for the frailty model., *Annals of Statistics* **23**: 182–198.

Nelsen, R. (2006). *An Introduction to Copulas*, 2nd edn, Springer.

Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**: 487–492.

Peng, L. e Fine, J. (2007). Regression modeling of semicompeting risk data, *Biometrics* **63**: 96–108.

Porta, N. (2010). *Interval-Censored semi-competing risk data novel approach for modeling bladder cancer*, PhD thesis, Universidad Politécnica de Cataluña, Espanha.

Rivest, L.-P. e Wells, M. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *Journal of Multivariate analysis* **79**: 138–155.

Steddon, S. e Sharples, E. (2009). *Clinical Practice Guideline. CKD- Mineral and Bone disorders (CKD-MBD) [em linha].*, http://www.renal.org/guidelines, London.

Wienke, A. (2011). *Frailty Models in Survival Analysis*, CRC Press, Güstrow.

Xu, J., Kalbfleisch, J. e Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data, *Biometrics* **66**: 716–725.

Zheng, M. e Klein, J. (1982). Estimates of marginal survival for dependent competing risk based on an assumed copula, *Biometrika* **82**: 127–138.