Análise de dados de sobrevivência com presença de riscos competitivos e fração de cura

Tamie Beatriz Medeiros Komino

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

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

São Paulo, 12 de setembro de 2016

Análise de dados de sobrevivência com presença de riscos competitivos e fração de cura

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Tamie Beatriz Medeiros Komino
e aprovada pela Comissão Julgadora.

Análise de dados de sobrevivência com presença de riscos competitivos e fração de cura

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Tamie Beatriz Medeiros Komino
e aprovada pela Comissão Julgadora.

Comissão Julgadora:

- Profª. Drª. Gisela Tunes da Silva (Presidente) - IME-USP
- Prof. Dr. Antonio Carlos Pedroso de Lima - IME-USP
- Prof. Dr. Rinaldo Artes - INSPER

Porque nada é impossível...
e nada é ao acaso.

Agradecimentos

Agradeço a todas circunstâncias da vida que me proporcionaram chegar até aqui e a todos aqueles que, de uma forma intencional ou mesmo despretenciosa, facilitaram essa jornada.

Em especial, gostaria de agradecer imensamente a minha orientadora Gisela Tunes da Silva pela sua infinita paciência e pela fé em mim. Muito obrigada por todo seu admirável profissionalismo em todos os momentos desses anos, por ser uma excelente professora e orientadora, pela incansável compreensão e principalmente, por tornar esse sonho possível. Com certeza, serei eternamente grata e sua fã.

Também agradeço ao meu marido, Erick Skorupa Parolin, que foi minha constante fonte de motivação e auto-confiança, por todas as broncas disciplinares que eu precisava ouvir, por cuidar de mim e da minha saúde sempre, por todo o suporte nas pequenas e grandes coisas ao longo desses anos e por fazer meus dias mais leves e alegres. Com você ao meu lado, tudo foi mais fácil e fez mais sentido.

Agradeço ao meu irmão Caio Sadao Medeiros Komino, que sugeriu que eu prestasse estatística mesmo quando eu não tinha a menor idéia do quanto eu ia me apaixonar pelo curso, e à Ana Luisa Montagnoli Lorenzeti, que tanto me ajudou com sua alegria e materiais impecáveis na graduação. Agradeço meus pais, Sandra Medeiros Komino e Eduardo Sadao Komino, por me fazerem ser quem sou. Agradeço a minha irmã Ayumi Cristina Medeiros Komino pelas caronas, conversas e amizade.

Agradeço a todos os professores que me incentivaram a fazer o curso de mestrado, em especial aos queridos Ademir José Petenate e Mauro Sérgio de Freitas Marques. E por fim, agradeço ao Renato Fadel Fava, por me mostrar que era possível realizar esse sonho sem precisar abrir mão das demais coisas importantes da vida, e à Jacqueline David, a grande propulsora dessa oportunidade maravilhosa que vivi nos últimos anos. Muito obrigada!

Resumo

A análise de riscos competitivos é um ramo da análise de sobrevivência dedicado ao caso particular em que mais de um tipo de evento é observado e que a ocorrência de um deles impede a ocorrência dos demais. A fração de cura é outro ramo que também requer metodologias apropriadas, e que acontece se uma porção da população não for suscetível aos eventos estudados. Embora ambos os ramos tenham sido explorados na literatura por muitos anos, a presença simultânea dessas duas particularidades em um conjunto de dados é um tema novo e ainda pouco desenvolvido.

Nesse trabalho pode ser encontrada uma comparação entre três métodos de estimação direta da função de incidência acumulada da causa específica. Dois deles baseiam-se na distribuição Gompertz e foram propostos para modelagem de dados com presença simultânea de riscos competitivos e fração de cura. O terceiro método baseia-se no uso de pseudo-valores, que foi inicialmente proposto para uso em riscos competitivos, mas como atende às premissas de dados com fração de cura, foi explorado nesse trabalho.

O estudo que motivou a comparação dessas técnicas reside na análise de cancelamento passivo e ativo de cartões de crédito, em que claramente um tipo de cancelamento impede o outro e nem todas as contas serão canceladas. O principal interesse é estimar a proporção de cartões ativos após estabilização da incidência de cancelamentos,para fornecer dados importantes para análises de rentabilidade e prejuízos. Os métodos foram comparados pelo uso de simulações e as particularidades de suas implementações detalhadas nas aplicações ao conjunto de dados que motivou a pesquisa. O uso de pseudo-valores, apesar da presença da fração de cura e de seu custo de processamento, se mostrou mais flexível e muito eficiente. Os métodos paramétricos apresentaram resultados menos ajustados ao conjunto de dados estudado, mas podem ser também uma boa alternativa para dados com riscos competitivos e fração de cura.

Palavras chave: Análise de sobrevivência; fração de cura; riscos competitivos; distribuição Gompertz; pseudo-valores; função de incidência acumulada.

Abstract

Competing Risks Analysis is a Survival Analysis field dedicated to a particular case where more than one event type is observed and its occurrence prevents occurrences from other types. Cure rate is another field that also requires proper methodologies, and that happens when a population portion is not susceptible to studied events. Although both of them have been explored in the literature for many years, the simultaneous presence of these two characteristics in a data set is a new theme and not well developed.

In this work it can be found a comparison of three methods for direct estimation of the cumulative incidence function. Two of them are based on Gompertz distribution and have been proposed for modeling data with simultaneous presence of competing risks and cure rate. Third method is based on using pseudo-values, which was originally proposed for use in competing risks, but as it suits cure rate premises, it was explored in this study.

The study that led to the comparison of these techniques lies in the analysis of credit cards passive and active attrition, where clearly an attrition type prevents the occurrence from another type and not all accounts will be canceled by the end of study. The main interest is to estimate active cards proportion after attrition incidence steady state in order to provide important data for profit and loss analysis. Methods were compared by simulations and their implementation particularities were detailed in applications to the data set that motivated this research. Pseudo-values method, in spite of cure rate presence and processing cost, proved to be more flexible and very efficient. Parametric methods presented results that were less adjusted to data set studied, but may also be a good alternative for data with presence of competing risks and cure rate.

Keywords: Survival analysis; Cure Rate; Competing Risks; Gompertz Distribution; Pseudo-Values; Cumulative Incidence Function.

Índice

Capítulo 1

Introdução

A análise de sobrevivência é uma área da estatística que engloba um conjunto de metodologias apropriadas para análise de dados cujo principal interesse reside no estudo do tempo até a ocorrência de um evento, também denominado por falha. Se esse evento for classificado em diferentes tipos ou causas e a ocorrência de uma determinada causa impede a ocorrência das demais, o conjunto de dados é dito com presença de *riscos competitivos*, já que os riscos de cada uma das causas competem entre si. Esse ramo da análise de sobrevivência, já muito explorado, surgiu do fato simples de que a existência de uma causa alternativa a que se pretende estudar afeta a incidência da causa de interesse não somente impedindo sua observação, como seria no caso de censura, mas impedindo a sua ocorrência e assim, requer uma metodologia específica.

Outro tema de importância crescente na análise de sobrevivência é a chamada *fração de cura*. A fração de cura ocorre se existir uma parcela dos indivíduos na população de um estudo para a qual o evento não irá acontecer, isto é, se existir essa parcela da população imune (ou ainda não suscetível) ao evento de interesse independente do tempo de seguimento desse estudo. É importante ressaltar que independente da observação da falha, a própria natureza dela poderá ditar a presença de cura nos dados, mas a não observação de falhas não garante a existência de fração de cura nos dados.

A presença da fração de cura rompe uma das premissas de muitos modelos em análise de sobrevivência, que é a ocorrência da falha para todos os indivíduos da população estudada. Com essa premissa rompida, a modelagem de dados requer técnicas que incorporem a fração de cura e até mesmo, sejam capazes de estimá-la. Os indivíduos não suscetíveis são chamados na literatura por "curados", "imunes" ou "sobreviventes de longo prazo" (em inglês, *long-term survivors*), ainda que o estudo não esteja relacionado a doenças ou morte. Se houver censura nos dados, isto é, indivíduos para os quais não é possível observar a falha, fica claro que a presença da cura estará confundida na parcela das observações censuradas, e que então, são necessárias técnicas apropriadas para esse tipo de dados a fim de se obter uma modelagem adequada e, principalmente, realizar inferências sobre essa fração de curados.

Apesar de ambos os temas já estarem cobertos por muitas propostas na literatura, pouco foi desenvolvido e proposto para a modelagem de dados com presença simultânea de riscos

competitivos e fração de cura. O estudo que motivou essa pesquisa reside na análise dados com essa presença simultânea, e de forma geral, consiste em fornecer dados importantes para a análise de rentabilidade de cartões de crédito. Nele, os contratos de cartões iniciados em um determinado mês são acompanhados até o final do estudo ou até que seja observado um dos dois tipos de cancelamento, um ativo e outro passivo. O cancelamento ativo ocorre quando o cliente cancela o cartão por desinteresse no produto e o passivo quando a instituição financeira cancela a conta de cartão de crédito por atraso de pagamento de um débito superior a um determinado prazo. Como era importante para análise de prejuízos dessa carteira que fosse estudada a incidência de cancelamentos passivos considerando a incidência de cancelamento ativos, os dados foram classificados com presença de riscos competitivos. No entanto, como é possível que uma cartão não seja cancelado por um desses dois tipos, os dados contêm fração de cura. Nesse contexto, a fração de cura é a proporção de cartões ativos após a estabilização das curvas de incidência acumulada dos dois tipos de cancelamento, e é uma métrica muito importante e de interesse para a análise de rentabilidade dessa carteira.

Nesse trabalho, motivado pelo estudo de dados de cartões de crédito, a ideia central é comparar as metodologias paramétricas propostas por Jeong e Fine (2006) e Jeong e Fine (2007) para dados com riscos competitivos e fração de cura baseadas na distribuição Gompertz ao uso de pseudo-valores, que originalmente foi proposto por Klein e Andersen (2005) somente para dados com riscos competitivos. Os métodos estão explicados nesse texto, particularidades de suas implementações são descritas nas aplicações aos dados de cartão de crédito e eles são comparados também por simulações, com suas vantagens e desvantagens detalhadas ao longo do texto.

1.1 Revisão Bibliográfica

O primeiro registro sobre a ideia central da análise de sobrevivência com riscos competitivos encontra-se em um livro de memórias lido em 1760 por Daniel Bernoulli e publicado posteriormente (Bernoulli, 1766). Motivado por uma controvérsia sobre os méritos de inoculação de varíola, Bernoulli propôs uma tabela de vida caso a varíola fosse eliminada, pois se questionara como seriam os efeitos na taxa de mortalidade se uma das causas de morte fosse modificada.

Estudando essa mesma ideia de comparação e relação entre diferentes causas de um evento de interesse (na literatura de análise de sobrevivência denominado por falha), Cox (1959) apresentou e discutiu modelos que poderiam ser utilizados para analisar dados nos quais as falhas eram classificadas entre apenas duas causas distintas. Nesse trabalho, ele sugeriu uma abordagem paramétrica com o uso das distribuições exponencial e Weibull. Também notou que em dados com presença de riscos competitivos apenas o tempo da falha e sua causa são observáveis, e que o tempo representa o menor dentre potenciais tempos de falha não observáveis cujas distribuições marginais são não identificáveis. Uma discussão mais profunda sobre a questão da não identificabilidade de modelos de riscos competitivos pode ser encontrada em Tsiatis (1975).

Apesar das discussões promovidas por Bernoulli e Cox, foi Gail (1975) quem formalizou a notação de riscos competitivos e apresentou as premissas para a análise desses dados. Em Prentice et al. (1978), os autores apresentaram um condensado do já estudado na época, introduzindo o conceito de função da taxa de falha da causa específica, e desenvolveram uma modelagem baseada nessa função, mostrando que era a quantidade essencial a ser estimada. Este material é mais facilmente encontrado no livro desses autores, Kalbfleisch e Prentice (2002). Uma abordagem de modelagem alternativa para a função da taxa de falha da causa específica foi apresentada por Larson e Dinse (1985), que utilizaram um modelo paramétrico de mistura no contexto de riscos competitivos, pelo qual modela-se a função taxa de falha condicional à causa e os componentes da mistura referem-se às proporções das causas.

Somente no final da década de 1980, Gray (1988) propôs um estimador não paramétrico da função de incidência acumulada e desenvolveu também testes não paramétricos para comparar as funções de incidência acumulada entre os diferentes grupos, assim como um método para analisar a relação entre covariáveis com as diferentes causas de falha. Com essa contribuição, os estudos passaram a envolver a modelagem da função de incidência acumulada e não somente a taxa de falha da causa específica. Benichou e Gail (1990) discutiram a modelagem da função de incidência acumulada através da parametrização completa da função taxa de falha da causa específica utilizando as distribuições exponencial e exponencial por partes. Na mesma época, Korn e Dorey (1992) demonstraram como o uso do estimador de Kaplan-Meier (Kaplan e Meier, 1958) pode superestimar a proporção de sobreviventes na presença de riscos competitivos reforçando a importância da função de incidência acumulada e, em seguida, Pepe e Mori (1993) compararam os métodos não paramétricos no contexto de riscos competitivos.

Mais tarde, Fine e Gray (1999) propuseram um modelo semi-paramétrico que ajusta diretamente a função de incidência acumulada e é muito flexível e semelhante ao modelo de regressão

de David (1972). Devido a essa semelhança e por ser capaz de fornecer interpretações simples e diretas, o modelo de Fine e Gray (1999) se popularizou. Mais recentemente, outros pesquisadores propuseram extensões para este modelo que acomodam variáveis dependentes no tempo (Scheike et al., 2008; Beyersmann e Schumacher, 2008), assim como outras extensões que incorporam um termo de fragilidade (Scheike et al., 2010; Dixon et al., 2011; Katsahian et al., 2006; Katsahian e Boudreau, 2011).

Bryant e Dignam (2004) também propuseram outro método semi-paramétrico para estimar a função de incidência acumulada que se mostrou melhor quando comparado aos métodos não paramétricos estudados. Também seguindo uma abordagem semi-paramétrica, foram propostas extensões do modelo de mistura que permitem a inclusão de covariáveis cujas componentes são formadas pelas sobrevivências marginais modeladas por Cox e pelas probabilidades de cada causa obtidas através de um modelo logístico (Chang et al., 2007; Lu e Peng, 2008).

Seguindo uma abordagem paramétrica, Jeong e Fine (2007) também propuseram a modelagem direta da função de incidência acumulada capaz de incorporar covariáveis e compararam com a modelagem usual da taxa de falha da causa específica. O modelo proposto por eles é capaz ainda de incorporar o que é chamado em análise de sobrevivência por fração de cura.

A fração de cura é um tema mais recente na análise de sobrevivência, mas o primeiro modelo encontrado na literatura para dados com presença de cura é o proposto por Boag (1949) e posteriormente desenvolvido por Berkson e Gage (1952), em que uma parcela da população é considerada imune ao evento observado enquanto assume-se que a parcela remanescente irá sofrer o evento. A função de sobrevivência para todos os indivíduos é então escrita como uma mistura dessas duas porções, a imune e a suscetível ao evento, que possuirá sua própria função de sobrevivência.

Haybittle (1965) ajustou uma proposta paramétrica utilizando a distribuição Gompertz, que para determinados valores assume sobrevivência imprópria, e assim parte dos indivíduos apresentam tempos de sobrevivência infinito. Seu modelo, diferente das duas propostas mencionadas anteriormente, não requer duas estimações e permite o teste da presença de cura nos dados.

Outra proposta paramétrica que combina o uso da função logística para modelar a proporção de curados e o uso da distribuição Weibull para modelar a função da taxa de falha da parcela suscetível pode ser encontrada em Farewell (1982) e um desenvolvimento semi-paramétrico, combinando o modelo logístico com o modelo de Cox em Kuk e Chen (1992). Em Taylor (1995), o autor propõe algo semelhante a Kuk e Chen, porém fazendo o uso do estimador de Kaplan-Meier para modelar a sobrevivência dos indivíduos suscetíveis só que restrito ao caso sem presença de covariáveis. Os artigos de Peng e Dear (2000) e Sy e Taylor (2000) também seguem a mesma abordagem que Kuk e Chen (1992), mas baseiam-se no algoritmo EM ao invés da aproximação por Monte Carlo, investigando os modelos semi-paramétricos e comparando-os com os paramétricos. O trabalho de Li et al. (2001) discute sobre a identificabilidade em modelos de cura seguindo essa abordagem. Mais adiante, Peng (2003) propõe uma abordagem semiparamétrica para o ajuste de modelos de cura baseado no algoritmo EM com o passo M alternando entre o modelo logístico e o de riscos proporcionais usando funções conhecidas e

disponíveis na maioria dos softwares estatísticos, tornando essa abordagem mais atrativa.

No trabalho de Chen et al. (1999) encontra-se uma discussão sobre alguns pontos desfavoráveis do uso de modelos de mistura, como a não possibilidade de construir modelos que atendam a propriedade de riscos proporcionais com a presença de covariáveis e como a definição de uma porção imune pode ser, em muitos contextos, questionável e não representativa do processo (biológico ou não) que gera a falha. Os autores sugerem o modelo de tempo de promoção (em inglês, *Bounded Cumulative Hazard (BCH)*), desenvolvido por Yakovlev et al. (1993), como uma alternativa viável para os modelos de mistura pois além de se derivar de uma motivação biológica, permite riscos proporcionais e ainda mantém uma relação matemática com o modelo de mistura. Uma discussão que aborda vantagens e desvantagens oferecidas por esses dois tipos de modelos também pode ser encontrada em Ibrahim et al. (2005a), livro que consolida técnicas bayesianas na análise de sobrevivência estudadas até a época.

Além dessa alternativa, em Aalen (1992) encontra-se outro modelo baseado na distribuição de Poisson composta (em inglês, *Compound Poisson Distribution*), que é gerada pela soma de variáveis aleatórias que seguem a distribuição Gama e que é capaz de incorporar a fração de cura pois permite a existência de uma parcela da população com suscetibilidade igual a zero. No trabalho de Rodrigues et al. (2011) encontra-se o desenvolvimento de uma classe de modelos baseada também na distribuição de Poisson composta.

Ainda com o interesse de garantir as propriedades de riscos proporcionais, no trabalho de Tsodikov (1998), o modelo de riscos proporcionais é modificado para permitir a existência da fração de cura e são sugeridos neste trabalho algoritmos de estimação para o modelo proposto. O autor discute ainda a estimação da fração de cura e mostra que no caso sem covariáveis, o estimador se resume ao estimador de Kaplan-Meier no final do estudo.

Em Yin e Ibrahim (2005b), os autores propuseram uma classe de modelos construídos a partir da transformação de Box-Cox aplicada na função da taxa de falha e uma estrutura aditiva, capaz de acomodar a fração de cura, se existir, e cuja o modelo de tempo de promoção é um caso particular. Em um trabalho seguinte, Yin e Ibrahim (2005a) propuseram a transformação aplicada à função de sobrevivência e discutiram formas de estimação para essa classe de modelos. Mais recentemente, Rodrigues et al. (2009) propuseram uma outra classe de modelos paramétricos oriunda de um conceito de riscos competitivos combinado ao uso da distribuição de Conway-Maxwell Poisson (Conway e Maxwell, 1962) que tem como casos particulares algumas distribuições conhecidas como a Poisson, Bernoulli e a Geométrica.

Alguns livros que abordam brevemente os chamados modelos de cura são Ibrahim et al. (2005b) e Lawless (2011). O livro de Maller e Zhou (1996) sintetiza muito bem os tópicos abordados até a época de sua publicação e traz uma revisão completa até a data, inclusive com uma proposta de estimador não paramétrico para a fração de cura baseado diretamente no estimador de Kaplan-Meier. Em Klein et al. (2016), é possível encontrar um conteúdo rico e mais moderno, porém bem resumido em um capítulo que discute as pesquisas mais recentes na área além de apresentar um consolidado das técnicas mais exploradas.

Apesar do trabalho extenso em riscos competitivos e em modelos para dados com presença de cura, a combinação desses dois fatores por enquanto foi pouco explorada por pesquisadores.

Um dos primeiros trabalhos encontrados foi o de Greenhouse e Wolfe (1984) em que os autores propõem modelos com fração de cura em dados com riscos competitivos, porém assumindo independência entre os tempos de falha das diferentes causas, o que torna essa abordagem ingênua e pouco útil uma vez que não é razoável em muitos casos assumir essa independência.

Outros trabalhos seguindo a abordagem bayesiana são o de Chao (1998), em que o autor não assume independência entre os riscos mas se limita predizer a proporção de cura em dados que possuem a característica de riscos competitivos e fração de cura, e o de Basu e Tiwari (2010), trabalho que desenvolve um modelo que combina o modelo de mistura e riscos competitivos e usa amostragem de cadeias de Markov para análise bayesiana do modelo, discutindo modelagem e problemas computacionais de sua estrutura.

Em Maller e Zhou (2002), os autores propõem um modelo de mistura paramétrico para riscos competitivos que permite a existência de indivíduos imunes e demonstram a existência, unicidade, consistência e normalidade assintótica dos estimadores de máxima verossimilhança quando os parâmetros atendem às restrições impostas. No entanto, os autores não oferecem a possibilidade de incorporar covariáveis no modelo nem para a modelagem dos tempos de falha e nem para a fração de cura, visando simplificar o desenvolvimento e a obtenção de resultados teóricos.

Os trabalhos mais recentes encontrados, e que são base para esse texto, focam na modelagem direta da função de incidência acumulada, como o Jeong e Fine (2006), em que essas funções são modeladas parametricamente pela distribuição Gompertz mas sem a possibilidade de incorporar covariáveis. Em Jeong e Fine (2007), os mesmos autores propõem uma mudança na estrutura do modelo, ainda baseado na distribuição Gompertz, para generalizá-lo e contemplar a incorporação de covariáveis explicativas. Os dois trabalhos se valem de que a distribuição Gompertz pode ser imprópria dependendo dos seus parâmetros, porém em ambos não existe a possibilidade de modelar a fração de cura diretamente através de covariáveis. A fração de cura é consequência da distribuição imprópria.

No caso do trabalho de Klein e Andersen (2005), os autores sugerem a utilização de pseudo-valores para a estimação da função de incidência acumulada sem qualquer imposição de modelagem entre os diferentes riscos. Aproveitando o raciocínio dos modelos anteriomente citados, esse modelo pode ser uma alternativa para dados com presença de cura, em que a estimação da fração de cura é indireta pela estimação das funções de incidência acumulada.

O trabalho mais recente encontrado é o de Choi et al. (2015) que é uma sequência de Choi e Huang (2014). Nesse primeiro trabalho, os autores propõem um algoritmo de estimação para o modelo semi-parámetrico de mistura proposto para o uso em riscos competitivos e sugerem que é possível combiná-lo ao método de estimação proposto por Peng e Dear (2000) para que seu modelo incorpore a fração de cura, porém não desenvolvem esse raciocínio. No trabalho posterior, os autores completam esse desenvolvimento, unificando riscos competitivos e fração de cura em um modelo de mistura que combina regressão logística multinomial para definir a probabilidade de cada causa específica a uma classe de modelos semiparamétricos para estimação do tempo de ocorrência desses eventos. Essa alternativa é interessante pois permite a modelagem direta da fração de cura por covariáveis e consequentemente, estimação

dos efeitos dessa covariáveis diretamente nessa proporção, mas não foi explorada nesse texto.

1.2 Objetivos e Organização do trabalho

O principal objetivo desse trabalho é comparar e testar as metodologias aqui apresentadas para análise de dados com presença de riscos competitivos e fração de cura por meio de aplicações e simulações. Nesse capítulo introdutório, o leitor pode encontrar a revisão bibliográfica realizada para riscos competitivos, fração de cura e ambos os ramos de análise de sobrevivência juntos.

No Capítulo 2, são apresentados alguns conceitos básicos de riscos competitivos, descritas algumas quantidades essenciais e exemplificada a análise descritiva em riscos competitivos, assim como é definida a notação que será seguida ao longo do texto. Nele também, são apresentados alguns conceitos básicos de fração de cura e listados seus modelos mais conhecidos.

No Capítulo 3, encontra-se a descrição das modelagens da função de incidência acumulada comparadas nesse texto, a notação adotada, assim como os métodos de estimação e inferência sugeridos pelos autores das propostas. Os modelos comparados são: o paramétrico de Jeong e Fine (2006), baseado na distribuição Gompertz, mas sem o uso de covariáveis; o paramétrico de Jeong e Fine (2007), também baseado na distribuição Gompertz, mas com o uso de covariáveis e adotada uma estrutura de transformação para a função de incidência acumulada; e pelo uso da técnica baseada nos pseudo-valores proposta por Klein e Andersen (2005), que apesar de não ser proposto inicialmente para dados com fração de cura, pela flexibilidade do modelo foi explorado nesse texto.

A seguir, são encontradas as simulações realizadas no Capítulo 4 e aplicações aos dados que motivaram o estudo no Capítulo 5, seções em que comentários sobre as especificidades das implementações de cada abordagem também foram descritas. A discussão final e comparações entre os métodos pode ser encontrada no último capítulo desse texto.

Capítulo 2

Conceitos Básicos de Riscos Competitivos e Fração de Cura

Conforme mencionado na introdução, os modelos de sobrevivência usuais assumem que todos os indivíduos apresentarão o evento e assim, se o seguimento do estudo for suficientemente longo e sem quaisquer impedimentos para a observação dele, todas as falhas serão observadas. Se nem todos os indivíduos apresentam falha até o fim de um estudo, é comum associar a ausência desta ao impedimento de sua observação, que no contexto de análise de sobrevivência é chamada de "censura", assumindo que a falha ocorreu ou ocorrerá mas que somente não foi possível observá-la durante a coleta de dados. No entanto, existem situações em que os eventos de interesse podem, de fato, não ocorrer para todos os indivíduos da população estudada e os dados são ditos com presença de "cura" ou com "fração de cura". Os indivíduos que não apresentam a falha podem ser casos de censura ou indivíduos curados, isto é, não suscetíveis ao evento de interesse. A modelagem para esse tipo de dados deve então ser adequada a essa característica do evento estudado, considerando a possibilidade da existência de uma parcela da população não suscetível à falha, a fim de proporcionar estimativas e inferências adequadas para essa situação.

Outra situação que difere da modelagem usual acontece quando os indivíduos em um estudo estão sujeitos *simultaneamente* a mais de um tipo de falha, ou ainda, quando existam outros eventos terminais que estejam impedindo a ocorrência do evento de interesse. Nesses casos, é necessário entender como esses demais eventos se comportam e se relacionam com o estudado para inferir informações sobre ele. Um conjunto de dados que possui mais de um tipo de falha e no qual se deseja estudá-los separadamente é referido por um conjunto de dados com "riscos competitivos", pois os tipos diferentes de falha são considerados riscos que *competem entre si*.

Muitas abordagens foram propostas na literatura para modelagem de dados com fração de cura e também muitas propostas podem ser encontradas para riscos competitivos. Como o objetivo desse trabalho é estudar propostas de modelagem para dados em que ambas essas situações estão presentes, as seções que seguem abordarão conceitos básicos de fração de cura

e de riscos competitivos, a fim de estabelecer uma notação, nomenclatura para as funções e quantidades e resumir principais resultados utilizados nesse texto.

2.1 Fração de Cura

Quando os dados de um estudo apresentam a fração de cura, além de estudar o tempo até a ocorrência de falha para os indívíduos suscetíveis, pode ser interessante estudar essa proporção de indivíduos não suscetíveis, bem como os fatores que interferem nessa proporção. Em especial, a caractéristica da possibilidade de presença de fração de cura deve ser levada em consideração para uma modelagem adequada pois a premissa de que a falha ocorrerá para todos os indivíduos foi rompida.

As principais questões que envolvem o estudo da fração de cura são:

- Estimar a fração de cura para a população e para seus subgrupos;
- Estudar os fatores que interferem nessa proporção de não suscetíveis.

Duas discussões muito importantes em análises de dados de sobrevivência são sobre a real existência da fração de cura e sobre a escolha do seguimento do estudo, isto é, o tempo máximo que os indivíduos são observados. Apesar da clara necessidade de que o seguimento do estudo seja "suficientemente" longo, em contextos práticos, a coleta de dados para um estudo geralmente é interrompida quando julga-se que o tempo de observação é suficiente para obter estimativas e informações úteis. Porém em muitos contextos, por exemplo naqueles que o evento tem incidência muito baixa por um longo período inicial, pode ser dificíl estabelecer se o tempo de observação é adequado. Uma característica importante para um conjunto de dados em que se pretende estudar a fração de cura é então o seguimento do estudo. Uma discussão sobre o seguimento suficiente pode ser em encontrada em Maller e Zhou (1996).

Em contextos que a própria natureza do evento permite a possibilidade de uma parcela da população não suscetível, fica mais claro assumir a fração de cura. Um exemplo para ilustrar esse tipo de situação é o estudo do tempo desde o casamento até a separação por divórcio de um casal. Nesse caso, fica clara a possibilidade de casais que não irão se divorciar, independente da escolha do seguimento, e portanto, a necessidade de contemplar essa parcela da população para realizar uma modelagem adequada.

Em contextos nos quais a fração de cura está presente, a censura à direita está confundida com a fração de cura, pois todo indíviduo curado aparece como censurado no estudo, mas nem todos os censurados são indivíduos curados. Dessa forma, uma característica que pode indiciar a presença de cura é o alto percentual de censurados no final do estudo, e portanto, na análise descritiva dos dados a curva estimada de sobrevivência se estabiliza acima de zero até o final do seguimento do estudo. De qualquer maneira, é importante ressaltar que ainda que seja observada a estabilização da curva de sobrevivência, trata-se de um indício, mas não de uma evidência da fração de curados. Em contextos práticos, a análise descritiva pode ser também um meio de avaliar a suficiência do seguimento do estudo para a melhor estimação da fração de cura, mas novamente não para garanti-la.

Em geral, muitos modelos de regressão assumem a existência da fração de cura conceitualmente, porque o pesquisador entende como possível a existência de uma porção da população que se mantém imune ao evento. Somente conceitualmente é possível justificar a existência de fração de cura em um estudo, e para ele propor modelagens que acomodem sua presença. No entanto, em alguns contextos, apesar de sua existência conceitualmente comprovada, os dados podem não evidenciar a presença de cura e a modelagem pode ser simplificada. Nesses casos, é importante notar que a definição correta do seguimento de um estudo é crucial para que inferências e resultados adequados sejam obtidos.

Em alguns casos, a existência de cura pode não ser justificada conceitualmente, mas pode haver interesse em questioná-la. Alguns modelos se propõem a testar a existência de fração de cura, o que deve ser entendido como um teste acerca da possível estabilização da incidência do evento a níveis praticamente nulos. Por exemplo, em um teste de uma nova vacina preventiva para uma determinada doença, pode haver interesse em testar se a vacina de fato evita a determinada doença para uma parte da população, e assim, um modelo para o tempo até o diagnóstico dessa doença após o recebimento da vacina, que comporta a presença de cura, é interessante para proporcionar o teste da existência de uma parcela da população que a vacina de fato evitou a doença. Uma discussão sobre o seguimento e o detalhamento sobre os testes para a presença de cura pode ser encontrada em Maller e Zhou (1996).

Em linhas gerais, dependendo do modelo adotado, o teste para a existência de fração de cura poderá ser realizado e além disso, será possível estudar quais fatores interferem na proporção de curados. Na seção a seguir, encontra-se um resumo com a estrutura dos principais modelos de cura.

2.1.1 Principais Modelos de Fração de Cura

O **Modelo de Mistura** é o mais antigo com fração de cura e foi proposto por Boag (1949) e posteriormente desenvolvido por Berkson e Gage (1952). Nele supõe-se que se a população de um estudo pode ser dividida em duas parcelas: uma suscetível ao evento e outra não suscetível. Seja Y o indicador de suscetibilidade, isto é, $Y = 1$ se o indivíduo é suscetível ao evento de interesse e $Y = 0$, caso contrário. Define-se $P(Y = 1) = \pi$. Se T é o tempo até a ocorrência do evento, então:

$$\begin{aligned} S(t) = P(T > t) = & \quad P(Y = 1)P(T > t|Y = 1) + P(Y = 0)P(T > t|Y = 0) \\ = & \quad \pi P(T > t|Y = 1) + (1 - \pi) \times 1 \\ = & \quad \pi S_{nc}(t) + (1 - \pi). \end{aligned}$$

Mesmo sendo bastante antiga, essa proposta continua sendo utilizada dado o apelo da facilidade de interpretação e de separação de efeitos sobre a proporção de curados e sobre a sobrevivência dos não curados, S_{nc}. Assim, modela-se π em função de determinadas covariáveis, por exemplo, pelo modelo logístico e modela-se S_{nc} de inúmeras maneiras distintas (modelos paramétricos, não paramétricos, riscos proporcionais, tempo de falha acelerado, etc.). Algumas

propostas que exploram combinações diferentes de métodos distintos podem ser encontradas em Berkson e Gage (1952), Farewell (1982), Kuk e Chen (1992), Taylor (1995), Peng e Dear (2000), Sy e Taylor (2000) e Peng (2003). O modelo porém, além de não ter a propriedade de riscos proporcionais, foi criticado pois em muitos contextos não é razoável assumir que parte da população é completamente imune ao evento estudado considerando o mecanismo que gera esse evento, como por exemplo em estudos de pacientes com câncer, em que se estuda o evento reincidência após um determinado tratamento pois nenhum paciente é "imune"à reincidência do câncer, apenas seu câncer não é detectável.

Como uma alternativa ao modelo de mistura, a abordagem desenvolvida por Yakovlev et al. (1993) foi posteriormente proposta por Chen et al. (1999) no contexto de fração de cura e é denominada por **Modelo de Tempo de Promoção**. Essa abordagem foi motivada pelo contexto biológico acima citado, e diferente do modelo de mistura, tem a propriedade de riscos proporcionais. Suponha que se pretende estudar o tempo até um indivíduo apresentar câncer detectável. Seja M o número de células carcinogênicas ativas em um indivíduo e D_j o tempo até a j-ésima célula carcinogência apresentar um câncer detectável. Assume-se que $M \sim Poisson(\theta)$ e que $D_1, D_2, \ldots, D_M$ são variáveis aleatórias independentes com função de distribuição acumulada $F_D(t|\boldsymbol{\lambda}) = 1 - S_D(t|\boldsymbol{\lambda})$, independentes de M. O tempo T até o aparecimento do câncer detectável é dado por $T = \min(D_0, D_1, D_2, \ldots, D_M)$, em que D_0 é tal que $P(D_0 = \infty) = 1$ e trata-se da forma como o modelo acomoda a fração de cura. Seguindo essas premissas:

$$
\begin{aligned}
S(t|\boldsymbol{\lambda}, \theta) &= P(T > t|\boldsymbol{\lambda}, \theta) \\
&= \textstyle\sum_{m=0}^{\infty} P(T > t|M = m, \boldsymbol{\lambda}) P(M = m|\theta) \\
&= P(T > t|M = 0, \boldsymbol{\lambda}) P(M = 0|\theta) + \textstyle\sum_{m=1}^{\infty} P(T > t|M = m, \boldsymbol{\lambda}) P(M = m|\theta) \\
&= P(M = 0|\theta) + \textstyle\sum_{m=1}^{\infty} S_Y(t|\boldsymbol{\lambda})^m P(M = m|\theta) \\
&= \textstyle\sum_{m=0}^{\infty} S_Y(t|\boldsymbol{\lambda})^m P(M = m|\theta) \\
&= e^{-\theta} \textstyle\sum_{m=0}^{\infty} \frac{[S_Y(t|\boldsymbol{\lambda})\theta]^m}{m!} \\
&= e^{-\theta} e^{\theta S_Y(t|\boldsymbol{\lambda})} \\
&= e^{-\theta F_Y(t|\boldsymbol{\lambda})}.
\end{aligned}
$$

Esse modelo garante a propriedade de riscos proporcionais e, adotando-se diferentes distribuições para F_Y, algumas propostas foram exploradas em Chen et al. (1999), Yakovlev et al. (1993) e Ibrahim et al. (2005a). Algumas extensões também foram propostas derivadas desse modelo em Rodrigues et al. (2009) e Rodrigues et al. (2011), como é o caso dos **Modelos Destrutivos de Poisson**.

Além dessas duas principais estruturas aqui descritas, outras estruturas mais simples foram exploradas utilizando-se de distribuições paramétricas impróprias, como por exemplo a **Distribuição Gompertz** em Haybittle (1965). Independente da estrutura adotada, o importante

é destacar que, em dados com presença de fração de cura, a estrutura de modelagem deve ser modificada para acomodar uma parcela da população que não sofrerá a falha.

2.2 Riscos Competitivos

Na presença de riscos competitivos, é importante destacar que mesmo que o pesquisador queira estudar apenas um específico tipo de falha, os demais tipos não devem ser ignorados ou tratados diretamente como censura (à direita) (Klein e Moeschberger, 2003), pois podem levar a conclusões incoerentes. Em geral, as motivações para a análise de sobrevivência com presença de riscos competitivos podem ser classificadas em uma das três situações a seguir:

- Estimação da relação entre covariáveis e a taxa de incidência de cada um dos possíveis tipos de falha;
- Estudo da interrelação entre os tipos de falha sob determinadas condições;
- Estimação das taxas de falha de determinadas causas de falha dado a remoção de alguns ou todas os demais causas.

Portanto, dependendo do cenário de interesse do pesquisador, isto é, com a presença dos demais riscos ou dada a remoção completa ou parcial deles, será necessário estudar as quantidades adequadas que resumam o conhecimento sobre a ocorrência do evento. Será comum se referir a essas probabilidades de interesse por:

- **Probabilidade Bruta:** probabilidade da falha considerando a presença dos demais fatores de risco.
- **Probabilidade Líquida:** probabilidade de falha removendo-se os demais riscos.
- **Probabilidade Parcial:** probabilidade de falha na situação em que somente parte dos demais riscos são removidos.

Para determinar quais delas serão as quantidades adequadas, o pesquisador deve ter claro qual é o cenário que pretende estudar, isto é, o da realidade em que os demais riscos competitivos existem e interferem na ocorrência daqueles de interesse ou, o cenário hipotético em que os riscos, ou parte deles, podem ser removidos. Para ilustrar, suponha que um pesquisador pretende estudar como é a mortalidade de uma determinada doença respiratória. No entanto, a mesma população estudada sofre com uma epidemia de uma outra doença fulminante. No cenário no qual a epidemia já está erradicada, que pode ser o interesse do pesquisador, fica claro que a mortalidade pela doença respiratória estudada poderá ser diferente da observada no cenário real, uma vez que os indivíduos que faleceram pela epidemia da outra doença poderiam falecer por doenças respiratórias, mas isso não foi observado. Se o interesse do pesquisador for no cenário de epidemia erradicada, o pesquisador deve buscar formas de estimar

a probabilidade líquida ou parcial, mas caso seu interesse seja no cenário real, deve buscar estimar as probabilidades brutas.

Nas seções a seguir, encontra-se um resumo sobre as principais funções de interesse, a notação adotada nesse texto e as formas de análise descritiva apropriadas para dados com riscos competitivos. Para o estudo mais detalhado e completo sobre o tema o leitor deve recorrer a Kalbfleisch e Prentice (2002), Klein et al. (2016), Crowder (2001), Pintilie (2006), Beyersmann et al. (2011a) e Crowder (2012).

2.2.1 Conceitos e Funções Básicas em Riscos Competitivos

Formulação Geral

Suponha que os indivíduos de um estudo estejam sujeitos a K tipos de falha e que, para cada um deles, somente é possível observar se a falha aconteceu, quando ocorreu e qual o seu tipo. Seja T uma variável contínua que representa o tempo de falha observado, $J \in \{1, 2, \ldots, K\}$ o indicador da causa da falha e $\mathbf{X}(t) = \{\mathbf{x}(u) : 0 \leq u \leq t\}$ o vetor de covariáveis. A **função da taxa de falha global** representa a taxa de ocorrência no instante t do evento de interesse - independente de seu tipo - dado o vetor de covariáveis, sendo expressa por:

$$\lambda(t, \mathbf{X}(t)) = \lim_{h \to 0} \frac{1}{h} P\left(t \leq T < t + h | T \geq t, \mathbf{X}(t)\right). \tag{2.1}$$

Para considerar qual é o tipo da falha, é necessário definir a **taxa de falha da causa específica**, que representa a taxa instantânea para as falhas do tipo j no tempo t dado $\mathbf{X}(t)$ e a presença de todos os demais riscos. Em outras palavras, representa a taxa de falha para os indivíduos que ainda não falharam por nenhum tipo de falha falharem pelo tipo j no instante t dado as covariáveis associadas. Sua definição é:

$$\lambda_j(t, \mathbf{X}(t)) = \lim_{h \to 0} \frac{1}{h} P\left(t \leq T < t + h\ ,\ J = j | T \geq t, \mathbf{X}(t)\right),\ j = 1, 2, \ldots, K. \tag{2.2}$$

Dessa forma, se somente um dos K tipos de falha pode ocorrer, claramente $\lambda(t, \mathbf{X}(t)) = \sum_{j=1}^{K} \lambda_j(t, \mathbf{X}(t))$.

Considerando o caso em que $\mathbf{X}(t)$ é vetor de covariáveis fixas, a **função de sobrevivência global**, que representa a probabilidade de um indivíduo não falhar por qualquer que seja o tipo até um instante t dado suas covariáveis, será obtida por:

$$S(t; \mathbf{X}) = P(T > t | \mathbf{X}) = \exp\left[-\int_0^t \lambda(u, \mathbf{X}) du\right]. \tag{2.3}$$

Define-se a **função de densidade** como:

$$f(t; \mathbf{X}) = \lim_{h \to 0} \frac{1}{h} P\left(t \leq T < t + h\ | \mathbf{X}\right) = \sum_{j=1}^{K} f_j(t; \mathbf{X}), \tag{2.4}$$

em que

$$f_j(t;\mathbf{X}) = \lim_{h\to 0}\frac{1}{h}P\left(t \leq T < t+h\ ,\ J=j|\mathbf{X}\right) = \lambda_j(t,\mathbf{X})S(t;\mathbf{X}),\ \ j=1,2,\ldots,K, \tag{2.5}$$

que representa **função de subdensidade para o tipo *j***. Essa função é chamada de subdensidade porque não se trata de uma densidade, mas sim de uma partição pelos diferentes tipos da densidade de T dado $\mathbf{X}$.

Uma das funções mais utilizadas em análise de riscos competitivos é a **função de incidência acumulada para o tipo *j* (FIA)**, também chamada de **função de subdistribuição da causa específica *j***:

$$CI_j(t;\mathbf{X}) = P(T\leq t, J=j;\mathbf{X}) = \int_0^t f_j(u,\mathbf{X})du = \int_0^t \lambda_j(u,\mathbf{X})S(u;\mathbf{X})du,\ t>0,\ j=1,2,\ldots,K. \tag{2.6}$$

Essa função expressa a probabilidade bruta do tipo j, isto é, a probabilidade de falha na presença dos demais riscos. Isso fica claro em sua fórmula, pois essa quantidade depende não somente do tipo de falha específico, mas de todas as taxas que os demais riscos ocorrem, já que seu cálculo envolve a função de sobrevivência global. Novamente, isso fortalece o argumento de quão importante é considerar os riscos que competem com um risco no qual se deseja estudar. Não é possível, porém, chamar essa quantidade de função de distribuição pois:

$$p_j = P(J=j) = \lim_{t\to\infty} CI_j(t;\mathbf{X}),\ j=1,2,\ldots,K.$$

Usualmente, assume-se que $\sum_{j=1}^{K} p_j = 1$, isto é, todos os indivíduos irão falhar em algum momento e somente um dos K tipos de falha pode ocorrer. É importante ressaltar que no caso de dados com a presença de fração de cura essa soma não será igual a um.

Em resumo, para estudar e realizar inferências sobre dados com presença de riscos competitivos as principais quantidades a se conhecer são:

- Função taxa de falha global;
- Função taxa de falha da causa específica;
- Função de sobrevivência global;
- Função de densidade;
- Função de subdensidade da causa específica;
- Função de incidência acumulada da causa específica.

Formulação por Tempos Latentes

A abordagem por tempos latentes é um caso particular para dados com presença de riscos competitivos em que supõe-se que os indivíduos de um estudo estão sujeitos a K tipos de falha e a existência de potenciais tempos de falha que não são observáveis $L_j, j=1,2,\ldots,K$,

chamados na literatura de tempos latentes. O que se observa para cada indivíduo é o valor mínimo $T = \min(L_1, L_2, \ldots, L_K)$ e $J \in \{1, 2, \ldots, K\}$, o indicador da causa da falha, além do vetor de covariáveis $\mathbf{X}(t) = \{\mathbf{x}(u) : 0 \leq u \leq t\}$. Portanto, dada a existência dos tempos latentes, torna-se possível obter a sobrevivência conjunta destes e se a estrutura de dependência entre os tempos latentes for conhecida, derivar e inferir as funções e quantidades de interesse.

Seja a função de sobrevivência conjunta dos potenciais tempos de falha dada por $S(t_1, t_2, \ldots, t_K) = P(L_1 > t_1, \ldots, L_K > t_K)$. A função da taxa de falha da causa específica pode ser encontrada por:

$$\lambda_j(t) = \frac{-\partial S(t_1, \ldots, t_K)/\partial t_j|_{t_1=\ldots=t_K=t}}{S(t, \ldots, t)}.$$

Uma dificuldade e desvantagem desse abordagem está em especificar a relação de dependência entre as tempos latentes, o que leva ao chamado **Dilema da Identificabilidade**. Em geral, assume-se alguma estrutura de dependência a partir de um conhecimento prévio sobre a natureza dos dados, mas como os potenciais tempos de falha não são observáveis em aplicações práticas, não é possível estudar a estrutura de dependência entre eles nem testar se qualquer premissa adotada para tal estrutura está adequada. Em muitos trabalhos, os tempos latentes são considerados independentes, o que motiva uma das principais críticas a essa abordagem: dificilmente os dados justificam tal independência e, se essa última for assumida, os demais riscos quando comparados a um específico podem ser encarados como censuras, o que simplifica o problema, mas utiliza uma premissa muito restritiva e não verificável. Mais ainda, se o interesse é estudar um risco específico isolando-o dos demais que competem com ele, assumir uma estrutura entre eles é interferir diretamente no que se deseja conhecer.

Assim, se os tempos latentes forem independentes com funções de sobrevivência $S_j(t)$ para $j = 1, 2, \ldots, K$, é possível notar que a função da taxa de falha da j-ésima causa específica coincide com a função da taxa de falha do j-ésimo tempo latente, pois:

$$\lambda_j(t) = \frac{-\partial \prod_{i=1}^{K} S_i(t_i)/\partial t_j|_{t_1=\ldots=t_K=t}}{\prod_{i=1}^{K} S_i(t)} = \frac{-\partial S_j(t_j)/\partial t_j|_{t_j=t}}{S_j(t)},$$

o que certamente simplifica o problema mas baseia-se em uma premissa muito restritiva.

Talvez um dos motivos pelos quais ainda exista uma grande atração por essa abordagem seja porque para muitos ela é mais intuitiva e permite de forma simples a resposta de questões sobre as probabilidades líquidas e parciais. Mesmo assim, alguns autores como Prentice et al. (1978) e Kalbfleisch e Prentice (2002) argumentam contra a existência dos tempos latentes e contra a possibilidade de tentar extrapolar qualquer inferência para algo que não foi observado nos dados. Ao contrário, Aalen (1975) argumenta que os modelos poderiam ir além dos dados e que tais extrapolações poderiam gerar discussões ou especulações úteis. Além desses autores, outros livros comparam a abordagem por tempos latentes com a geral, sobretudo apontando as desvantagens da primeira, como é o caso de Klein et al. (2016), Pintilie (2006) e Beyersmann et al. (2011b).

2.2.2 Análise Descritiva para Riscos Competitivos

Como explicado nas seções anteriores, a análise de dados com riscos competitivos difere da análise de sobrevivência usual e essa seção apresenta três técnicas para análise descritiva nesse contexto. Usualmente, dada a presença de mais de um tipo de falha, estimam-se as curvas de incidência acumulada (subdistribuição), ao invés da curva de sobrevivência, ou a possível partição da função de sobrevivência global entre as causas de falha.

O complementar do estimador de Kaplan-Meier é um estimador comumente utilizado, mas se trata de uma abordagem ingênua. Denotando-se por $t_1 < t_2 < \ldots < t_N$ os instantes distintos de falha e de censuras ordenados, esse estimador é da forma:

$$\widetilde{CI_j(t)} = 1 - \prod_{i:t_i \leq t} \left(\frac{Y_i - r_{ij}}{Y_i} \right),$$

em que

Y_i é o número de indivíduos em risco em t_i, isto é, indivíduos que nesse instante ainda não falharam por nenhuma causa, e

r_{ij} é o número de falhas do tipo j em t_i.

Em sua construção, deve-se adotar que as ocorrências dos demais riscos sejam tratadas como censuras. Este pode ser interpretado como um estimador da probabilidade líquida, isto é, da ocorrência de um determinado tipo de falha até um instante t num cenário hipotético em que os demais riscos são completamente removidos. O uso desse estimador pode ser completamente inadequado dependendo do interesse do estudo, porque:

- Raramente o interesse de um estudo reside no entendimento do cenário hipotético onde os riscos são removidos;
- Em muitas situações o cenário hipotético é inviável;
- Assumir os demais riscos como censuras no estimador Kaplan-Meier requer independência entre os riscos, que não pode ser verificada e que, em muitos contextos, não é razoável.

Assim, esse estimador poderá ser utilizado quando desejar-se inferências sobre o cenário em que o risco de interesse esteja isolado e somente no caso restritivo em que os riscos são independentes, lembrando que essa deve ser uma premissa justificável conceitualmente já que não poderá ser verificada.

Um segundo método, não paramétrico, para estimar a j-ésima função de incidência acumulada é construído denotando-se por:

- $t_1 < t_2 < \ldots < t_N$ os instantes distintos de falha e de censura ordenados (de qualquer um dos K tipos de falha);
- Y_i o número de indivíduos em risco em t_i, isto é, o número de indivíduos que ainda não falharam por qualquer causa ou que não foram censurados até imediatamente antes do instante t_i;

- r_i o número de falhas do tipo de interesse, j, em t_i;
- d_i o número de falhas dos demais tipos, que não j, em t_i.

O estimador é definido por (Kalbfleisch e Prentice, 2002):

$$\widehat{CI_j(t)} = \begin{cases} 0, & t < t_1 \\ \sum_{t_i \leq t} \left\{ \prod_{l=1}^{i-1} 1 - \frac{(d_l + r_l)}{Y_l} \right\} \frac{r_i}{Y_i} = \sum_{t_i \leq t} \hat{S}(t_i^-) \frac{r_i}{Y_i}, & t \geq t_1, \end{cases} \tag{2.7}$$

em que $\hat{S}(t_i^-)$ é o estimador de Kaplan-Meier para a falha independente de sua causa avaliado imediatamente antes do instante t_i. A j-ésima função de incidência acumulada (FIA) representa a probabilidade bruta do evento do tipo j ocorrer antes do instante t e antes dos demais riscos competitivos. Por essa quantidade ser na maioria das vezes o objeto de interesse em um estudo, o uso desse estimador é mais recomendado. A derivação para obtenção do estimador pode ser encontrado em (Pintilie, 2006), assim como para estimar sua variância e obter os resultados assintóticos. Uma forma de estimar sua variância é:

$$\widehat{V}\left[\widehat{CI_j(t)}\right] = \sum_{t_i \leq t} \hat{S}(t_i)^2 \left\{ \left[\widehat{CI_j(t)} - \widehat{CI_j(t_i)}\right]^2 \frac{(r_i + d_i)}{Y_i^2} + \left[1 - 2(\widehat{CI_j(t)} - \widehat{CI_j(t_i)})\right] \frac{r_i}{Y_i^2} \right\}.$$

A partir de alguma dessas possíveis estimativas da variância, os intervalos de confiança pontuais $(1-\alpha)100\%$ para a FIA podem ser obtidos por meio de $\left[\widehat{CI_j(t)} \pm Z_{1-\alpha/2} \sqrt{V(\widehat{CI_j(t)})}\right]$.

A terceira forma de sintetizar os dados é chamada de **função de probabilidade condicional para o risco** j e é definida por:

$$CP_j(t) = \frac{CI_j(t)}{1 - CI_{j^C}(t)} = \frac{P(T \leq t, J = j)}{(P(T \leq t, J = j) + P(T > t, J = j) + P(T > t, J = j^C))}.$$

Essa função é um estimador da probabilidade condicional da falha ocorrer pelo tipo j até o instante t dado que a falha não ocorreu por nenhuma das outras causas até esse momento. Essa quantidade incorpora a informação sobre a causa específica e sua relação com as demais causas, e pode ser útil pois em alguns casos a função de incidência acumulada apresenta valores baixos porque um dos demais riscos competitivos apresenta valores muito altos. Essa relação considera inclusive a probabilidade do evento ocorrer após o instante t, seja pela causa especificada ou por seus riscos competitivos.

Para construir um estimador dessa quantidade de um particular risco j, sejam respectivamente $\widehat{CI_j(t)}$ e $\widehat{CI_{j^C}(t)}$ as estimativas da funções de incidência acumulada para ele e para todos os demais riscos agrupados como um só. O estimador será definido por:

$$\widehat{CP_j(t)} = \frac{\widehat{CI_j(t)}}{1 - \widehat{CI_{j^C}(t)}}.$$

Nota-se que, por sua forma, esse estimador muda de valor toda vez que uma falha ocorre, seja do tipo específico ou não, e ainda que a probabilidade condicional sempre será maior que

a função de incidência acumulada.

Pepe e Mori (1993) mostraram que o estimador $\widehat{CP_j(t)}$ é assintoticamente normal com variância estimada por:

$$\widehat{V}\left[\widehat{CP_j(t)}\right] = \frac{\widehat{S}(t^-)^2}{\left(1-\widehat{CI_{j^C}(t)}\right)^4}\sum_{t_i\leq t}\frac{\left(1-\widehat{CI_{j^C}(t_i)}\right)^2 r_i + \widehat{CI_j(t_i)}^2 d_i}{Y_i(Y_i-1)},$$

em que $t_1 < t_2 < \ldots < t_N$ são os instantes distintos de falha e de censura ordenados, $\widehat{S}(t^-)$ é o estimador de Kaplan-Meier avalidado no instante imediatamente antes de t, $\widehat{CI_j(t)}$ e $\widehat{CI_{j^C}(t)}$ são as estimativas da funções de incidência acumulada para ele e para todos os demais riscos agrupados como um só, r_i e d_i representam respectivamente o número de falhas da causa específica e o das demais causas em t_i e Y_i representa o número de indivíduos em risco em t_i.

Os três métodos aqui apresentados podem ser utilizados para sintetizar os dados e os gráficos sugeridos são dessas estimativas contra o eixo do tempo. É importante apresentar as curvas para todos os riscos já que a incidência de um afeta a do outro, e geralmente é melhor quando todos estão num mesmo gráfico para proporcionar mais facilmente a comparação entre eles. Claramente, dependendo do principal interesse da pesquisa, poderá ser útil apresentar gráficos de cada um dos riscos separadamente e neles, comparar diferentes subgrupos da população.

Outro gráfico geralmente apresentado no contexto de riscos competitivos é o da partição do complementar do estimador de Kaplan-Meier. Seja r_{jl} o número de falhas do tipo j no instante l e o estimador de Kaplan-Meier obtido por:

$$\widehat{S}(t_i) = \prod_{l=1}^{i}\left(\frac{Y_l - R_l}{Y_l}\right) \text{ em que } R_l = \sum_{j=1}^{K} r_{jl}.$$

Note que $\widehat{S}(t_i) = \widehat{S}(t_{i-1})\left(\frac{Y_i-R_i}{Y_i}\right)$, e portanto:

$$\begin{aligned}
\widehat{S}(t_i) + \textstyle\sum_{j=1}^{K}\widehat{CI_j(t_i)} &= \widehat{S}(t_i) + \textstyle\sum_{j=1}^{K}\sum_{l=1}^{i}\widehat{S}(t_{l-1})\frac{r_{jl}}{Y_l} \\
&= \widehat{S}(t_{i-1})\left(\frac{Y_i - R_i}{Y_i}\right) + \widehat{S}(t_{i-1})\left(\frac{R_i}{Y_i}\right) + \textstyle\sum_{l=1}^{i-1}\widehat{S}(t_{l-1})\frac{R_l}{Y_l} \\
&= \widehat{S}(t_{i-2})\left(\frac{Y_{i-1} - R_{i-1}}{Y_{i-1}}\right) + \widehat{S}(t_{i-2})\left(\frac{R_{i-1}}{Y_{i-1}}\right) + \textstyle\sum_{l=1}^{i-2}\widehat{S}(t_{l-1})\frac{R_l}{Y_l} \\
&\vdots \\
&= \widehat{S}(t_0)\left(\frac{Y_1 - R_1}{Y_1}\right) + \widehat{S}(t_0)\left(\frac{R_1}{Y_1}\right) = 1
\end{aligned}$$

Dessa forma, assim como é conhecida a relação das funções de sobrevivência e de incidência

acumulada da causa específica por:

$$S(t;\mathbf{X}) = 1 - P(T \leq t|\mathbf{X}) = 1 - P(T \leq t, J = j;\mathbf{X}) = 1 - \sum_{j=1}^{K} CI_j(t), \tag{2.8}$$

o desenvolvimento acima garante que o somatório de todas funções de incidência acumulada estimadas não parametricamente, como descrito no texto, é o complementar do estimador de Kaplan-Meier:

$$\widehat{S}(t) = 1 - \sum_{j=1}^{K} \widehat{CI_j(t)} \text{ ou } \sum_{j=1}^{K} \widehat{CI_j(t)} = 1 - \widehat{S}(t). \tag{2.9}$$

Assim, a área total composta por todas as funções de incidência acumulada representa a área complementar ao estimador de Kaplan-Meier para a falha, sem distinção de causa. O gráfico pode ser construído da seguinte forma:

1. Obtenha $CI_j(t_i)$ para $j = 1, 2, \ldots, K$ e para todos os distintos tempos de falha (t_i com $i = 1, 2, \ldots, t_N$);
2. Obtenha $C_1(t_i) = CI_1(t_i) \ \ \forall t_i$;
3. Para $k = 2, \ldots, K$, obtenha $C_k(t_i) = \sum_{j=1}^{k} CI_j(t_i) \forall t_i$;
4. Apresente os valores obtidos para $C_k(t_i)$ com $k = 1, 2, \ldots, K$ no eixo vertical e os tempos no eixo horizontal.

Exemplo com Dados Fictícios

Para ilustrar os cálculos e os gráficos obtidos na análise descritiva de um conjunto de dados com riscos competitivos, serão utilizados os dados fictícios extraídos de Pintilie (2006). Suponha que um estudo sobre queda na terceira idade conduzido em um instituto de reabilitação incluiu dez pacientes e que o evento de interesse era queda com lesão grave que resultava em uma reabilitação tardia. Dois outros eventos competem com o evento de interesse: o falecimento do indivíduo ou sua alta do instituto. Ambos podem ser considerados eventos terminais que impedem a ocorrência do evento de interesse e portanto deve ser considerado um modelo de riscos competitivos a fim de obter estimativas adequadas às restrições e características dos dados. Somente o número de dias até o primeiro evento e seu tipo foram gravados, sendo codificado 0 para censura, 1 para a queda com lesão grave, 2 para o falecimento do paciente e 3 para a alta deste.

A Tabela 2.1 apresenta os dados, as funções de incidência acumulada estimadas para cada um dos três tipos de evento e a estimativa da probabilidade condicional para o evento de interesse, incluindo os cálculos necessários para obter as estimativas.

Além dos dados tabelados, se o conjunto de dados fosse maior, seria mais fácil e elegante apresentar os dados através de análises descritivas tais como a Figura 2.1, que é uma das formas sugeridas nesse texto para apresentar dados com riscos competitivos. Nele as funções

Tabela 2.1: Cálculo das funções de incidência acumulada e probabilidade condicional para a causa específica

ID	t	J	$\hat{S}(t)$	$\widehat{CI_1(t)}$	$\widehat{CI_2(t)}$	$\widehat{CI_3(t)}$	$\widehat{CI_{1C}(t)}$	$\widehat{CP_1(t)}$
-	0	-	1,0000	0,0000	0,0000	0,0000	0,0000	0/ (1-0) = 0,0000
04	24	2	0,9000	$0 + 1 \times (0/10) = 0,0000$	0,1000	0,0000	0,1000	0/ (1-0,1) = 0,0000
08	37	0	0,9000	$0 + 0,9 \times (0/9) = 0,0000$	0,1000	0,0000	0,1000	0/ (1-0,1) = 0,0000
10	49	1	0,7875	$0 + 0,9 \times (1/8) = 0,1125$	0,1000	0,0000	0,1000	0,1125/ (1-0,1) = 0,1250
03	61	2	0,6750	$0,1125 + 0,7875 \times (0/7) = 0,1125$	0,2125	0,0000	0,2125	0,1125/ (1-0,2125) = 0,1429
06	65	3	0,5625	$0,1125 + 0,675 \times (0/6) = 0,1125$	0,2125	0,1125	0,3250	0,1125/ (1-0,325) = 0,1667
02 09	80	1 3	0,3375	$0,1125 + 0,5625 \times (1/5) = 0,2250$	0,2125	0,2250	0,4375	0,225/ (1-0,4375) = 0,4000
07	106	2	0,2250	$0,225 + 0,3375 \times (0/3) = 0,2250$	0,3250	0,2250	0,5500	0,225/ (1-0,55) = 0,5000
05	113	0	0,2250	$0,225 + 0,225 \times (0/2) = 0,2250$	0,3250	0,2250	0,5500	0,225/ (1-0,55) = 0,5000
01	128	1	0,0000	$0,225 + 0,225 \times (1/1) = 0,4500$	0,3250	0,2250	0,5500	0,45/ (1-0,55) = 1,0000

de incidência acumulada para cada um dos tipos de falha estão sobrepostas permitindo fácil comparação entre elas para cada instante do tempo.

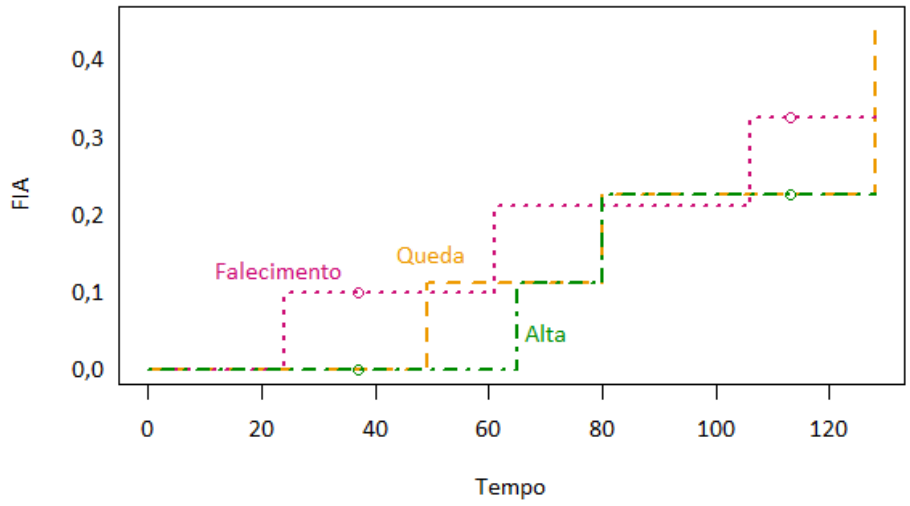


Figura 2.1: Função de incidência acumulada de cada um dos riscos competitivos

Outra forma sugerida aqui está apresentada na Figura 2.2, em que cada uma das funções de incidência acumulada são apresentadas como áreas e formam uma das partição do complementar de Kaplan-Meier (se não houvesse distinção de causa da falha). Esse gráfico pode ser de interesse quando os dados apresentam fração de cura pois permite fácil visualização da área remanescente, que representa a proporção de indivíduos que não sofreram nenhum evento até cada instante. Em especial, quando há presença de fração de cura, a proporção de indivíduos livres de falhas estabilizará como uma constante após um determinado instante de tempo.

Note que o complementar do estimador de Kaplan-Meier particionado nesse gráfico considera todos os tipos de falha como um só e é diferente do primeiro estimador mencionado nesse texto, que visa estimar uma curva para cada causa específica.

Nos gráficos da Figura 2.3, é possível comparar as curvas estimadas pelo complementar do

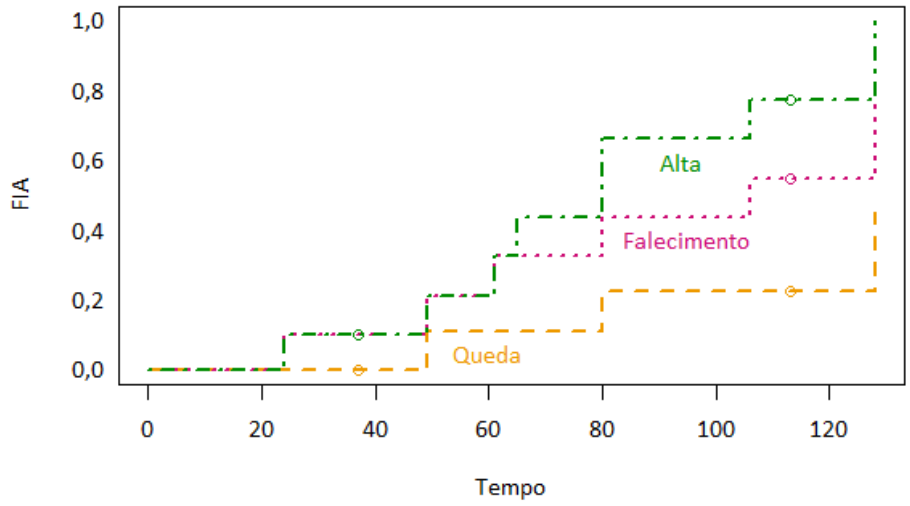


Figura 2.2: Gráfico da partição do Complementar do Kaplan-Meier

Kaplan-Meier para cada causa específica, em que a ocorrência dos demais riscos são considerados censuras aleatórias, com a função de incidência acumulada da mesma causa específica: o complementar do KM sempre apresenta curva superior à função de incidência acumulada pois ignora o fato de que os indivíduos que já falharam não poderão falhar pelo causa específica de interesse. Dessa forma, fica claro que o uso do complementar do Kaplan-Meier para análise descritiva de dados com presença de riscos competitivos é uma abordagem ingênua e pode levar a interpretações incorretas pois desconsidera a característica fundamental de riscos competitivos de que os riscos competem entre si. A função de incidência acumulada da causa específica fornece a estimativa mais correta para a probabilidade acumulada de um determinado evento de interesse na presença de riscos competitivos sem a necessidade de premissas sobre a dependência entre os riscos.

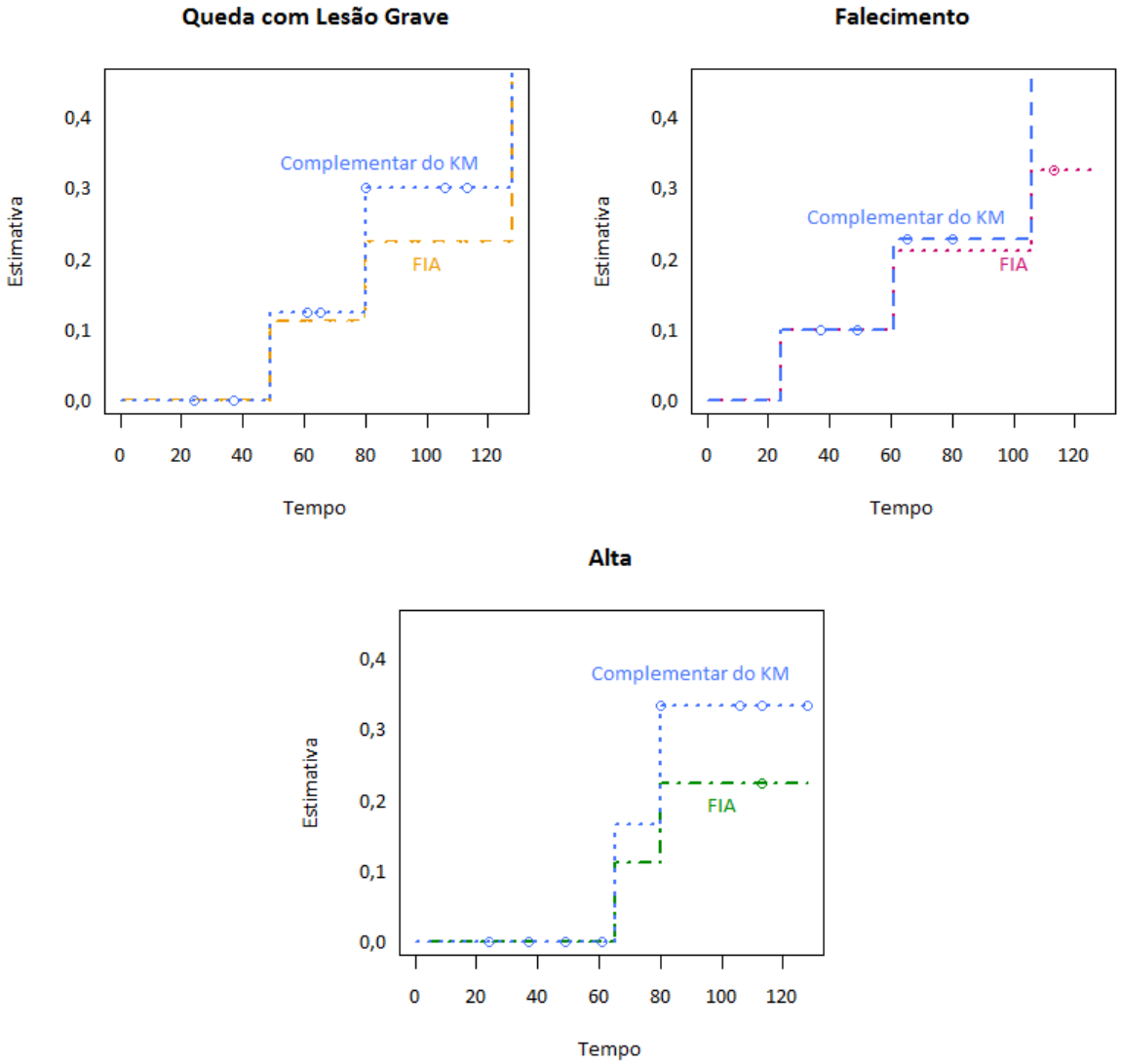


Figura 2.3: Comparação entre as curvas estimadas pelo complementar do Kaplan-Meier e pela FIA de cada causa específica

Capítulo 3

Riscos Competitivos com Fração de Cura

No capítulo anterior estão listadas todas as quantidades necessárias para o estudo de dados com presença de riscos competitivos. Nota-se que é suficiente conhecer todas as funções de incidência acumulada da causa específica ou todas as funções taxa de falha da causa específica pois a partir delas é possível obter as demais quantidades de interesse. Assim, muitas abordagens para modelagem de dados com presença de riscos competitivos envolvem modelar as **funções taxa de falha da causa específica** e estabelecer uma relação entre as causas como é o caso nos trabalhos de Benichou e Gail (1990) e Larson e Dinse (1985). Alternativamente, é possível modelar a **função de incidência acumulada da causa específica**, ou ainda explorar uma transformação dessa função, assim como a proposta de Fine e Gray (1999) em que os autores impuseram uma estrutura de riscos proporcionais aplicando uma transformação sobre a função de incidência acumulada da causa específica.

Em análise de sobrevivência sem a presença de riscos competitivos, existe uma relação direta entre a função da taxa de falha e a função de sobrevivência (ou ainda a função distribuição de probabilidade acumulada), e portanto, o efeito das covariáveis sobre a função taxa de falha pode ser obtido algebricamente pelos efeitos dessas covariáveis sobre a função de sobrevivência, e vice-versa. No caso de riscos competitivos, não é possível permutar a função de incidência acumulada da causa específica com a função da taxa de falha da mesma causa, pois a função de incidência acumulada de um risco específico depende da taxa de falha de todos os riscos simultaneamente:

$$CI_j(t;\mathbf{X}) = P(T \leq t, J = j;\mathbf{X}) = \int_0^t \lambda_j(u,\mathbf{X})S(u;\mathbf{X})du = \int_0^t \lambda_j(u,\mathbf{X}) \exp\left(-\int_0^u \sum_{l=1}^{K} \lambda_l(v,\mathbf{X})dv\right)du, \tag{3.1}$$

em que $t > 0,\ j = 1, 2, \ldots, K$. Levando isso em consideração, a escolha da técnica de regressão deve permitir a inferência direta sobre os efeitos das covariáveis na função desejada, seja ela a função de incidência acumulada ou da taxa de falha da causa específica, dependendo do contexto do estudo.

No trabalho de Jeong e Fine (2006), os autores propõem uma abordagem paramétrica pela modelagem direta da função de incidência acumulada fazendo uso da distribuição Gompertz

(Gompertz, 1825) e em um segundo trabalho, Jeong e Fine (2007) acrescentam a possibilidade de introduzir covariáveis no modelo e usar uma estrutura de modelo transformada assim como em Fine e Gray (1999) e Fine (2001). Nas seções que seguem, esses trabalhos estão descritos com maior detalhe pois são capazes de acomodar um conjunto de dados com presença de riscos competitivos e fração de cura, fornecendo inclusive estimativas para a fração de cura.

Outra proposta detalhada nesse capítulo é a de Klein e Andersen (2005), que trata de uma abordagem baseada no uso de pseudo-valores originalmente proposta para análise de dados com presença de riscos competitivos somente. Denotando $p_j = P(J = j) = \lim_{t\to\infty} CI_j(t; \mathbf{X})$, $j = 1, 2, \ldots, K$, como esse modelo não impõe qualquer restrição tal que $\sum_{j=1}^{K} p_j = 1$, o modelo pode incorporar a fração de cura. Dessa forma, essa metodologia também será explorada nesse texto como uma alternativa para dados com presença de fração de cura e riscos competitivos, sendo que alguns resultados utilizados nos trabalhos de Jeong e Fine (2007) serão também aplicados ao uso de pseudo-valores para assim obter estimativas para a fração de cura.

3.1 Modelagem Paramétrica da Função de Incidência Acumulada

Nesta seção está detalhada a metodologia proposta por Jeong e Fine (2006) que se baseia no uso da distribuição Gompertz, já muito explorada para a modelagem de dados com fração de cura pois dependendo de seus parâmetros pode assumir forma imprópria. Sua distribuição acumulada pode ser escrita como:

$$F(t; \rho, \tau) = 1 - \exp\left(\frac{\tau}{\rho}(1 - \exp(\rho t))\right), \text{ em que } -\infty < \rho < \infty \text{ e } 0 < \tau < \infty. \tag{3.2}$$

Logo, sua função taxa de falha e função taxa de falha acumulada são dadas por:

$$\lambda(t; \rho, \tau) = \frac{1}{1 - F(t; \rho, \tau)} \frac{\delta F(t; \rho, \tau)}{\delta t} = \tau \exp(\rho t)$$

e

$$\Lambda(t; \rho, \tau) = \int_0^t \lambda(u; \rho, \tau) du = \int_0^t \tau \exp(\rho u) du = \frac{\tau}{\rho}(\exp(\rho t) - 1). \tag{3.3}$$

Nota-se que a distribuição Gompertz é imprópria sempre que ρ for negativo, pois:

$$\lim_{t\to\infty} F(t; \rho, \tau) = \begin{cases} 1 & \text{, se } \rho \geq 0 \\ 1 - \exp(\tau/\rho) & \text{, se } \rho < 0. \end{cases} \tag{3.4}$$

Além disso, se ρ for negativo o limite quando $t \to \infty$ assume valores entre 0 e 1. Portanto, a distribuição Gompertz é capaz de acomodar propriedades da função de incidência acumulada e proporcionar uma forma de estimar a fração de cura, se ela existir.

3.1.1 Modelo

Seja K o número de riscos competitivos, CI_j a j-ésima função de incidência acumulada e $F^G(.;\rho,\tau)$ a função distribuição acumulada da Gompertz como especificada em (3.2) com parâmetros ρ e τ. O modelo proposto é da forma:

$$CI_j(t) = F^G(t;\rho_j,\tau_j) = 1 - \exp\left(\frac{\tau_j}{\rho_j}(1-\exp(\rho_j t))\right), \tag{3.5}$$

para $t > 0$, $j = 1,2,\ldots,K$, $-\infty < \rho_j < \infty$ e $0 < \tau_j < \infty$.

Uma vantagem dessa formulação é ser parsimoniosa em relação às abordagens paramétricas mencionadas anteriormente, pois dispensa uma forma de relacionar as K funções. Nela, somente $2K$ parâmetros deverão ser estimados. A desvantagem é que esse modelo não permite a comparação entre grupos pois não prevê o uso de covariáveis. Se o interesse do estudo não for a comparação entre grupos, mas apenas a estimação da fração de cura da população como um todo essa é uma abordagem simplista que acomoda presença de riscos competitivos e fração de cura simultaneamente.

Além de atender a necessidade da estimação da fração de cura em dados com a presença de riscos competitivos, uma das vantagens desse modelo é que ele permite o teste da presença da fração de cura. Em alguns contextos esse teste é muito útil pois pode por si só endereçar questões do estudo. Em outros pode ser incoerente, uma vez que assume-se conceitualmente a presença de cura, ainda que o seguimento da amostra seja insuficiente para garanti-la. Portanto, apesar de limitado, esse modelo oferece uma ferramenta interessante dependendo do interesse do estudo.

3.1.2 Estimação dos Parâmetros

Seja uma amostra composta por N indivíduos, denote os tempos de falha ou censura observados para cada indivíduo por T_i para $i = 1,2,\ldots,N$ e defina os indicadores da causa específica como:

$$\delta_{ji} = \begin{cases} 1 & \text{se o } i\text{-ésimo indivíduo falhou pela } j\text{-ésima causa} \\ 0 & \text{caso contrário,} \end{cases}$$

$j = 1,2,\ldots,K$. Defina, para $j = 1,2,\ldots,K$, as funções de subdensidade como:

$$f_j^G(t;\rho_j,\tau_j) = \frac{\delta F_j^G(t;\rho_j,\tau_j)}{\delta t} = \tau_j \exp(\rho_j t)\exp\left(\frac{\tau_j}{\rho_j}(1-\exp(\rho_j t))\right).$$

Sabendo que:

$$S(t;\boldsymbol{\psi}) = 1 - \sum_{j=1}^{K} F_j^G(t;\rho_j,\tau_j),$$

a função de verossimilhança e a função de log-verossilhança para esse modelo respectivamente serão:

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{N}\left[\left(\prod_{j=1}^{K} f_j^G(t_i;\rho_j,\tau_j)^{\delta_{ji}}\right) S(t_i;\boldsymbol{\psi})^{1-\sum_{j=1}^{K}\delta_{ji}}\right] \tag{3.6}$$

e

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{N}\left[\left(\sum_{j=1}^{K}\delta_{ji}\log(f_j^G(t_i;\rho_j,\tau_j))\right) + \left(1-\sum_{j=1}^{K}\delta_{ji}\right)\log(S(t_i;\boldsymbol{\psi}))\right], \tag{3.7}$$

em que $\boldsymbol{\psi} = (\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \ldots, \boldsymbol{\psi_K})$ e $\boldsymbol{\psi_j} = (\rho_j, \tau_j)$.

Utilizando algum método numérico de maximização de funções aplicado na função de log-verossilhança, como sugerido no artigo, obtém-se os estimadores para os parâmetros desconhecidos. Portanto os $2K$ parâmetros serão estimados simultaneamente e todas quantidades de interesse (FIA, funções taxa de falha das causas, função de sobrevivência, fração de cura, etc.) serão derivadas desses parâmetros. Como consequência, uma fragilidade do modelo é que se houver erro na especificação da distribuição para um dos riscos, todas as estimativas para os demais parâmetros serão afetadas, diferente do que acontece se forem modeladas as funções taxa de falha da causa específica, pois a verossimilhança pode ser fatorada em K parcelas que podem ser estimadas separadamente. Para ilustrar, suponha que foi adotada um técnica de modelagem da função taxa de falha da causa específica, então o modelo é da forma:

$$\lambda_j(t) = h^A(t;\boldsymbol{\varphi_j}), \text{ para j = 1, 2, . . . , K,}$$

em que $h^A(.;\varphi_j)$ é uma função especificada que atende às propriedadas da função da taxa de falha da causa específica e $\boldsymbol{\varphi_j}$ é o vetor de parâmetros associados ao j-ésimo risco. Como é possível fatorar a função de sobrevivência por:

$$S(t;\boldsymbol{\varphi}) = \exp\left(-\int_0^t \sum_{j=1}^{K}\lambda_j(u;\boldsymbol{\varphi_j})du\right) = \prod_{j=1}^{K}\exp\left(-\int_0^t h^A(u;\boldsymbol{\varphi_j})du\right) = \prod_{j=1}^{K} S_j(t;\boldsymbol{\varphi_j}),$$

e utilizando a subdensidade da forma expressa em (2.5), note que a função de verossimilhança pode ser fatorada em K parcelas que envolvem somente os parâmetros de cada um dos K riscos:

$$\begin{aligned}
L(\boldsymbol{\varphi}) &= \prod_{i=1}^{N}\left[\left(\prod_{j=1}^{K} f_j(t_i;\boldsymbol{\varphi_j})^{\delta_{ji}}\right) S(t_i;\boldsymbol{\varphi})^{1-\sum_{j=1}^{K}\delta_{ji}}\right] \\
&= \prod_{i=1}^{N}\left[\left(\prod_{j=1}^{K}\lambda_j(t_i;\boldsymbol{\varphi_j})S(t_i;\boldsymbol{\varphi_j})^{\delta_{ji}}\right) S(t_i;\boldsymbol{\varphi})^{1-\sum_{j=1}^{K}\delta_{ji}}\right] \\
&= \prod_{i=1}^{N}\left[\left(\prod_{j=1}^{K} h^A(t_i;\boldsymbol{\varphi_j})^{\delta_{ji}}\right)\prod_{j=1}^{K} S_j(t_i;\boldsymbol{\varphi_j})\right] \\
&= \prod_{i=1}^{N}\left[\prod_{j=1}^{K}\left(h^A(t_i;\boldsymbol{\varphi_j})^{\delta_{ji}} S_j(t_i;\boldsymbol{\varphi_j})\right)\right].
\end{aligned}$$

Para evitar o erro na especificação da distribuição, uma sugestão dos autores é testar diferentes distribuições para riscos competitivos diferentes, de acordo com o que for observado na análise descritiva dos dados.

Outra crítica importante a esse modelo é que para um vasto subespaço de $\boldsymbol{\Psi}$, o espaço paramétrico definido para o modelo em (3.5), os valores obtidos para a função de sobrevivência a partir de um determinado tempo são negativos. Claramente, como nenhuma imposição ao espaço paramétrico ou a sua estrutura é feita nesse modelo, ao utilizar métodos numéricos é possível esbarrar em tais valores absurdos e não calculáveis quando, por exemplo, pretende-se obter o logaritmo da função de sobrevivência em pontos que essa última for negativa. Dependendo do software e pacote utilizado, o algoritmo de maximização pode ser interrompido por erro ou ainda alertar que não é possível avaliar alguns subespaços de $\boldsymbol{\Psi}$. Os pacotes porém irão avaliar os parâmetros somente para os valores de tempo observados na amostra, e portanto, modelos que assumem valores absurdos para grandes valores de tempo que não estiverem presentes na amostra poderão ser sugeridos. Dessa forma, se houver interesse em estimativas para grandes valores de tempo esse modelo proposto nem sempre proporcionará estimativas interpretáveis e nem mesmo razoáveis. Essa deficiência do modelo é grave pois um dos apelos desse modelo é para o uso em dados com fração de cura, que requerem boas, razoáveis e interpretáveis estimativas para o final das curvas de incidência acumulada das causas específicas e sobrevivência. Os autores argumentam em seu trabalho que a estimação seria muito mais complexa se fossem consideradas as restrições para definir o espaço paramétrico de tal forma que esses problemas não acontecessem, mas não propõem uma solução nem indicam como obtê-la.

É importante notar também que ao implementar esse modelo, é necessário atentar-se para possíveis erros numéricos oriundos da forma da distribuição Gompertz. Note que:

$$\lim_{\rho \to 0} F^G(t, \boldsymbol{\psi}) = 1 - \exp(-t\tau), \tag{3.8}$$

e que ao especificar a função em um software, caso esses limites sejam desconsiderados, os valores obtidos serão incorretos.

Além das críticas listadas, esse modelo é simples e não incorpora covariáveis, limitando seu uso para comparar subpopulações. Mas se todas curvas de incidência acumulada na amostra estudada apresentarem platô, isto é, apresentam estabilização em um valor a partir de um determinado instante de tempo, a grande deficiência do modelo em apresentar função de sobrevivência negativa não acontecerá e o modelo é uma alternativa razoável para modelagem de dados com presença de riscos competitivos e fração de cura. Os conjuntos de dados em que os platôs não estiverem claros para todos os riscos competitivos, seja por uma alta proporção de censura, seguimento insuficiente ou por uma presença de fração de cura muito baixa, irão apresentar dificuldade na estimação se for adotado esse modelo e resultarão em estimativas não razoáveis.

3.1.3 Inferência

Além das estimativas para os parâmetros são necessárias as estimativas de suas variâncias para realizar inferências. Também a partir delas, é possível derivar as estimativas para as

variâncias das funções de incidência acumulada das diferentes causa se assim obter intervalos de confiança desejados.

Para calcular as estimativas das variâncias o procedimento é encontrar a inversa da matriz de informação observada. Determina-se a matriz de informação $I(\boldsymbol{\psi})$ a partir da função de log-verossimilhança $l(\boldsymbol{\psi})$ em que $\boldsymbol{\psi} = (\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \ldots, \boldsymbol{\psi_K})$ com $\boldsymbol{\psi_j} = (\rho_j, \tau_j)$ por:

$$I(\boldsymbol{\psi}) = (-1)\begin{bmatrix} \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_1}\boldsymbol{\psi_1}} & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_1}\boldsymbol{\psi_2}} & \cdots & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_1}\boldsymbol{\psi_K}} \\ \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_2}\boldsymbol{\psi_1}} & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_2}\boldsymbol{\psi_2}} & \cdots & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_2}\boldsymbol{\psi_K}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_K}\boldsymbol{\psi_1}} & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_K}\boldsymbol{\psi_2}} & \cdots & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_K}\boldsymbol{\psi_K}} \end{bmatrix} \text{ em que } \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \boldsymbol{\psi_i}\boldsymbol{\psi_j}} = \begin{bmatrix} \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \rho_i \rho_j} & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \rho_i \tau_j} \\ \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \tau_i \rho_j} & \frac{\delta^2 l(\boldsymbol{\psi})}{\delta \tau_i \tau_j} \end{bmatrix}. \tag{3.9}$$

Seguindo esses cálculos, a diagonal da inversa da matriz $I(\boldsymbol{\psi})$ será o vetor das variâncias estimadas para o vetor $\boldsymbol{\psi}$. Pelo princípio da invariânca, podem ser obtidas as estimativas de máxima verossimilhança para as funções de incidência acumulada da causa específica:

$$\widehat{CI_j(t)} = F^G(t; \hat{\rho}_j, \hat{\tau}_j) = 1 - \exp(\hat{\tau}_j(1 - \exp(\hat{\rho}_j t))/\hat{\rho}_j), \tag{3.10}$$

e conforme sugerido pelos autores, assumindo válidas as condições usuais de regularidade para os estimadores de máxima verossimilhança, podem ser obtidas as estimativas para as variâncias de $CI_j(t)$, $j = 1, 2, \ldots, K$ pelo Método Delta multivariado por:

$$\hat{V}\left[\widehat{CI_j(t)}\right] = \hat{V}\left[F^G(t; \hat{\rho}_j, \hat{\tau}_j)\right] = \left(\frac{\delta F_j^G(t; \boldsymbol{\psi_j})}{\delta \boldsymbol{\psi_j}}\right)\Bigg|_{\boldsymbol{\psi_j}=\hat{\boldsymbol{\psi}}_j} \hat{V}(\hat{\boldsymbol{\psi}}_j) \left(\frac{\delta F_j^G(t; \boldsymbol{\psi_j})}{\delta \boldsymbol{\psi_j}}\right)'\Bigg|_{\boldsymbol{\psi_j}=\hat{\boldsymbol{\psi}}_j} \text{ para } j = 1, 2, \ldots, K, \tag{3.11}$$

em que $\delta F_j^G(t; \boldsymbol{\psi_j})/\delta \boldsymbol{\psi_j}$ é um vetor 1×2 das primeiras derivadas da função de incidência acumulada da j-ésima causa em relação ao vetor $\boldsymbol{\psi_j} = (\rho_j, \tau_j)$ e $\hat{V}(\hat{\boldsymbol{\psi}}_j)$ é a submatriz da inversa da matriz de informação correspondente ao vetor $\boldsymbol{\psi_j}$ avaliada em $\boldsymbol{\psi_j} = \hat{\boldsymbol{\psi}}_j$.

Assim, uma possível forma de se obter um intervalo de confiança pontual de α% para a j-ésima FIA avaliada no tempo t é por:

$$F_j^G(t; \hat{\rho}_j, \hat{\tau}_j) \pm z_{\alpha/2} \hat{V}\left[F^G(t; \hat{\rho}_j, \hat{\tau}_j)\right], \tag{3.12}$$

em que $z_{\alpha/2}$ é tal que $P(Z > z_{\alpha/2}) = \alpha/2$ sendo $Z \sim \text{Normal}(0, 1)$. Em especial, se houver interesse no estudo da fração de cura, esse intervalo pode ser avaliado para $\widehat{CI_j(t)}$ em $t \to \infty$ (como pode ser observado em (3.4)).

Para os autores, a fração de cura existe para o evento estudado e os demais riscos competitivos são fatores importantes a se considerar para ajustar corretamente o modelo, de tal forma que se um indivíduo falha por um dos riscos pode ser considerado um curado para outro risco competitivo a ele. A cura se relaciona ao tipo do evento e não ao evento como um todo. Esse

raciocínio contraria a intuição de que a fração de cura corresponde à porção de indivíduos que não sofrem *qualquer evento* pois interpreta que se o indivíduo estiver curado de um determinado tipo de falha de interesse, nada impede que esse apresente uma falha por um segundo motivo. Num contexto de dados sem riscos competitivos mas com fração de cura, todos os curados são censurados mas nem todas as censuras são de indivíduos curados. De forma análoga, o modelo proposto sugere que todos os curados para o tipo de falha estudado são censurados ou apresentam uma falha de outro tipo, mas nem todos aqueles que foram censurado ou falharam por demais tipos são curados. Dessa forma, a fração de cura para o j-ésimo risco é:

$$Q_j = \lim_{t\to\infty} \; 1 - F_j^G(t;\boldsymbol{\psi}_j). \tag{3.13}$$

Assim como derivado em (3.4), o limite da distribuição Gompertz para ρ_j positivo é 1, e portanto, não existe cura para o risco estudado indexado por j. Dessa forma, é possível testar a presença da cura testando se ρ_j é negativo por $H_0 : \rho_j \geq 0$ vs $H_1 : \rho_j < 0$.

Além da fração de cura proposta pelos autores que segue o raciocínio descrito acima, pode ser estimada uma fração de cura mais intuitiva, que corresponde a indivíduos que não apresentam falhas (qualquer que seja o tipo). Vale lembrar que se pelo menos uma das curvas de incidência acumulada não apresentar platô, a estimativa sugerida a seguir poderá apresentar valores negativos, isto é, não razoáveis ou interpretáveis. No caso em que todas as curvas apresentam platô, e então os valores estimados para os parâmetros ρ_j forem negativos, a *cura global* é:

$$Q = \lim_{t\to\infty} \; 1 - \sum_{j=1}^{K} F_j^G(t;\boldsymbol{\psi}_j), \tag{3.14}$$

e assim como na cura do j-ésimo risco, pode ser estimada tomando-se $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$.

3.2 Modelagem Paramétrica da Função de Incidência Acumulada com Covariáveis

Uma forma de modelar dados em análise de sobrevivência é adotar uma classe de modelos obtidos pela transformação da função de sobrevivência, ou no caso de riscos competitivos, pela transformação da função de distribuição acumulada. Assim como descrito anteriormente, no trabalho de Fine e Gray (1999), essa estrutura de transformação foi adotada para a função de incidência acumulada pela vantagem da facilidade de interpretação direta sobre o efeito das covariáveis nessa função. Assumindo funções conhecidas e crescentes $g_j(.), j = 1,\ldots,K$, tal que:

$$g_j(CI_j(t;\mathbf{Z})) = u_j(t) + \mathbf{Z}^T\boldsymbol{\beta}_j, \;\; j = 1,\ldots,K, \tag{3.15}$$

em que $u_j(t)$ é uma função inversível e monótona crescente, $\boldsymbol{\beta}_j$ é o vetor $P \times 1$ dos parâmetros associados às covariáveis e $\mathbf{Z}$ o vetor $P \times 1$ das covariáveis fixas, essa transformação é conveniente pois permite que para dois indivíduos distintos com vetores de covariáveis $\mathbf{Z_1}$ e $\mathbf{Z_2}$ a

diferença entre as funções transformadas seja um deslocamento vertical, isto é:

$$g_j(CI_j(t;\mathbf{Z_1})) - g_j(CI_j(t;\mathbf{Z_2})) = (\mathbf{Z_1} - \mathbf{Z_2})^T\boldsymbol{\beta}_j, \forall t.$$

Fine e Gray (1999) optaram pelo uso da transformação chamada de complemento log-log, em que $g(u) = \log(-\log(1-u))$, por ser a mesma utilizada para o conhecido modelo de riscos proporcionais (sem a presença de riscos competitivos). Dessa forma,

$$g_j(CI_j(t;\mathbf{Z})) = \log(-\log(1 - CI_j(t;\mathbf{Z}))) = \log(-\log(1 - P(T \leq t, J = j;\mathbf{Z})))$$

$$\Rightarrow \log(-\log(P((T > t) \cup (T \leq t \cap J \neq j);\mathbf{Z}))) = u_j(t) + \mathbf{Z}^T\boldsymbol{\beta}_j.$$

Seguindo esse desenvolvimento é possível encontrar a função taxa de falha modificada de Fine e Gray (1999), por:

$$\begin{aligned} -\log(P((T > t) \cup (T \leq t \cap J \neq j);\mathbf{Z})) &= \exp(u_j(t))\exp(\mathbf{Z}^T\boldsymbol{\beta}_j) \\ \Rightarrow \frac{\delta}{\delta t}\left(-\log(P((T > t) \cup (T \leq t \cap J \neq j);\mathbf{Z}))\right) &= \frac{\delta}{\delta t}\Big(\exp(u_j(t))\Big)\exp(\mathbf{Z}^T\boldsymbol{\beta}_j) \\ \Rightarrow \lambda_j^*(t) &= \lambda_{j_0}(t)\exp(\mathbf{Z}^T\boldsymbol{\beta}_j). \end{aligned}$$

É importante notar que essa não é uma função taxa de falha da causa específica usual pois seu grupo em risco também inclui indivíduos que já falharam por demais causas até o instante de interesse, t. Portanto, a interpretação da função de incidência acumulada da j-ésima causa g-transformada é contra-intuitiva se analisada por sua correspondente função taxa de falha da j-ésima causa mas, assim como argumentam os autores, a interpretação dos coeficientes é direta para a função de incidência acumulada e não depende da estrutura probabilística da função taxa de falha da causa específica. A interpretação direta é uma vantagem para estudos em que se deseja testar efeitos de covaráveis sobre a função de incidência acumulada e não sobre a taxa de falha da causa específica, pois como estudado por Gray (1988) e Pepe (1991) o efeito das covariáveis sobre essas funções pode ser muito diferente. Para a estimação desses coeficientes, Fine e Gray (1999) construíram a verossimilhança parcial comportando os indivíduos desse grupo de risco modificado.

Com motivações semelhantes quanto a facilidade de interpretação e ainda propondo uma extensão desse trabalho, Jeong e Fine (2007) utilizaram uma classe de transformações paramétricas da forma:

$$g_j(u;\alpha_j) = \log\left(\frac{(1-u)^{-\alpha_j} - 1}{\alpha_j}\right), -\infty < \alpha_j < \infty. \tag{3.16}$$

Portanto, a proposta pode ser considerada mais genérica e flexível uma vez que cada um dos riscos competitivos assume uma função de transformação paramétrica diferente. Além disso, essa tranformação assume dois importantes casos particulares. Se $\alpha_j \to 0$, a função de ligação se reduz ao complemento log-log:

$$\begin{aligned}\lim_{\alpha_j \to 0} g_j(u;\alpha_j) &= \log\left[\lim_{\alpha_j \to 0}\left(\frac{(1-u)^{-\alpha_j}-1}{\alpha_j}\right)\right] \\ &= \log\left[\lim_{\alpha_j \to 0}\left(\frac{-(1-u)^{-\alpha_j}\log(1-u)}{1}\right)\right] \\ &= \log\left(-\log(1-u)\right).\end{aligned} \tag{3.17}$$

Apesar dos autores chamarem o modelo de riscos proporcionais se $\alpha_j \to 0$, vale ressaltar que utilizando a função de ligação complemento log-log, o modelo assume forma semelhante ao de Fine e Gray (1999), e na realidade baseia-se em uma função taxa de falha diferente da usual, $\lambda_j^*(t)$. Rigorosamente, não é correto chamar esse modelo de riscos proporcionais, pois em riscos competitivos a função de incidência acumulada de uma causa específica não se relaciona diretamente à função da taxa de falha da mesma causa específica ou à função da taxa de falha global:

$$-\log(1-CI_j(t)) \neq \int_0^t \lambda_j(u)du \text{ e } -\log(1-CI_j(t)) \neq \int_0^t \lambda(u)du,$$

e, dessa forma, ao considerar a função de incidência acumulada g-transformada pelo complemento log-log o modelo não atende a propriedade de riscos proporcionais, diferente do que acontece em análise de sobrevivência sem riscos competitivos se o complemento log-log for aplicado à função distribuição acumulada.

Se $\alpha_j = 1$, a função de ligação será o logito:

$$\begin{aligned}g_j(u;\alpha_j) = g_j(u;1) &= \log\left((1-u)^{-1}-1\right) \\ &= \log\left(\frac{u}{1-u}\right),\end{aligned} \tag{3.18}$$

e nesse caso, podemos dizer que o modelo proposto é de chances proporcionais se, ao interpretar os efeitos das covariáveis, o conceito de "chance" na presença de riscos competitivos for a relação $\dfrac{CI_j(t)}{1-CI_j(t)}$.

Além da transformação aplicada à função de incidência acumulada, para que o modelo seja completamente especificado, é preciso estabeler a forma de $u_j(t)$. A escolha dos autores também pretende estender o trabalho anterior, Jeong e Fine (2006), e utilizar a distribuição Gompertz como base para essa função, e assim, sofisticar a proposta anterior incorporando covariáveis.

3.2.1 Modelo

Seja K o número de riscos competitivos, $\mathbf{Z}$ o vetor $P \times 1$ de covariáveis independentes no tempo, $\boldsymbol{\beta_j}$ o vetor $P \times 1$ de parâmetros associados às covariáveis para o j-ésimo risco, CI_j a j-ésima função de incidência acumulada e $g_j(.;\alpha_j)$ a função de ligação proposta por Jeong e Fine assim como em (3.16) para a estrutura transformada como descrito em (3.15). Assim, para o

j-ésimo evento:

$$g_j(CI_j(t;\mathbf{Z});\alpha_j) = \log\left(\frac{(1-CI_j(t;\mathbf{Z}))^{-\alpha_j}-1}{\alpha_j}\right) = u_j(t) + \mathbf{Z}^T\beta_\mathbf{j},\ j = 1,\ldots,K. \tag{3.19}$$

Desenvolvendo a equação acima, obtem-se:

$$\begin{aligned}
\log\left(\frac{(1-CI_j(t;\mathbf{Z}))^{-\alpha_j}-1}{\alpha_j}\right) &= u_j(t) + \mathbf{Z}^T\boldsymbol{\beta_j} &\Leftrightarrow\\
(1-CI_j(t;\mathbf{Z}))^{-\alpha_j} &= 1 + \alpha_j \exp\left(u_j(t) + \mathbf{Z}^T\boldsymbol{\beta}_j\right) &\Leftrightarrow\\
CI_j(t;\mathbf{Z}) &= 1 - \left(1 + \alpha_j \exp\left(u_j(t)\right)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}_j\right)\right)^{\frac{-1}{\alpha_j}}.
\end{aligned}$$

Nota-se que se $u_j(t)$ deve ser uma função inversível e monotonamente crescente, o mesmo se aplicará para $\exp(u_j(t))$. Portanto, a sugestão dos autores é utilizar a função taxa de falha acumulada da Gompertz para substituir a expressão $\exp(u_j(t))$ na equação derivada acima. A função taxa de falha acumulada da Gompertz foi obtida em (3.3) e, tomando $\exp(u_j(t)) = \tau_j(\exp(\rho_j t)-1)/\rho_j$, o modelo fica especificado por:

$$CI_j(t,\boldsymbol{\psi}_j;\mathbf{Z}) = 1 - \left(1 + \alpha_j\frac{\tau_j}{\rho_j}(\exp(\rho_j t)-1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}_j\right)\right)^{\frac{-1}{\alpha_j}} \tag{3.20}$$

para $j = 1,2,\ldots,K, -\infty < \alpha_j < \infty, -\infty < \rho_j < \infty$ e $0 < \tau_j < \infty$, em que $\boldsymbol{\psi}_j = (\alpha_j, \boldsymbol{\beta}_j^T, \rho_j, \tau_j)$.

Os autores argumentam que a escolha da distribuição Gompertz é estratégica para que as funções de incidência acumulada componham uma função de sobrevivência imprópria (dependendo dos parâmetros) e portanto, acomode quando necessário a presença de cura nos dados. Para verificar se a função de sobrevivência pode ser imprópria, os limites da função (3.20) para $t \to \infty$ foram estudados e os resultados estão sumarizados a seguir, com a remoção do índice j apenas para simplificar a notação:

$$\begin{aligned}
&\text{(I) para } \rho \geq 0 \text{ e } \alpha < 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = \infty\\
&\text{(II) para } \rho \geq 0 \text{ e } \alpha \geq 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1\\
&\text{(III) para } \rho < 0 \text{ e } \alpha < 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1 - \left(1 - \alpha\frac{\tau}{\rho}\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)^{\frac{-1}{\alpha}}\\
&\text{(IV) para } \rho < 0 \text{ e } \alpha > 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1 - \left(1 - \alpha\frac{\tau}{\rho}\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)^{\frac{-1}{\alpha}}\\
&\text{(V) para } \rho < 0 \text{ e } \alpha \to 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1 - \exp\left(\frac{\tau}{\rho}\exp(\mathbf{Z}^T\boldsymbol{\beta})\right)\\
&\text{(VI) para } \rho \to 0 \text{ e } \alpha < 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = \infty\\
&\text{(VII) para } \rho \to 0 \text{ e } \alpha \geq 0: && \lim_{t\to\infty} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1.
\end{aligned} \tag{3.21}$$

Nota-se que para que a função (3.20) seja imprópria, não é suficiente que ρ_j seja negativo.

Além disso, essa função pode assumir valores superiores a 1 em (I), (III), (IV) e (VI) para algum subespaço de $\boldsymbol{\Psi}_j$ o que resultará, independente dos demais riscos, em uma função de sobrevivência com valores negativos. Mesmo que todas as funções de incidência acumulada sejam impróprias, o modelo também não estabelece nenhuma restrição para os parâmetros tal que comporte as características básicas das funções principais em riscos competitivos, nem que garanta que a função de sobrevivência resultante seja imprópria. Portanto, uma crítica para esse modelo é que o domínio dos parâmetros não está bem definido, e assim, as principais quantidades em riscos competitivos segundo esse modelo podem assumir valores não razoáveis e interpretáveis.

Esse segundo trabalho de Jeong e Fine estende o trabalho mencionado na seção anterior em dois sentidos: incorporando o uso de covariáveis no modelo e generalizando a função de ligação para uma forma que tem como casos particulares modelos já explorados e interessantes do ponto de vista de interpretação de parâmetros como o de chances proporcionais e o de "riscos" proporcionais de Fine e Gray (1999). A proposta continua sendo de acomodar simultaneamente riscos competitivos e a possibilidade de fração de cura.

Como o trabalho anterior, esse modelo pode ser interessante pois permite o teste da presença de cura, e como trata-se de uma abordagem mais completa pois incorpora o uso de covariáveis explicativas, permite comparações entre subgrupos de um estudo. Além disso, o modelo é mais flexível pois a estrutura de transformação utilizada e as covariáveis poderão ser diferentes para cada tipo de evento. Outro fator interessante é a possibilidade de testar a qualidade do ajuste para modelos de chances proporcionais, que seguindo a sugestão dos autores, pode ser realizado testando se $H_0 : \alpha_j = 1$ contra $H_a : \alpha_j \neq 1$.

3.2.2 Estimação dos Parâmetros

Seja uma amostra composta por N indivíduos, os tempos de falha ou censura observados para cada indivíduo denotados por T_i para $i = 1, 2, \ldots, N$, K tipos de eventos competitivos, Z_i o vetor $P \times 1$ de covariáveis fixas e os indicadores da causa específica definidos como:

$$\delta_{ji} = \begin{cases} 1 & \text{se o } i\text{-ésimo indivíduo falhou pela } j\text{-ésima causa} \\ 0 & \text{caso contrário.} \end{cases}$$

A função de verossimilhança para esse modelo será:

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{N} \left[\left(\prod_{j=1}^{K} f_j(t_i, \boldsymbol{\psi}_j; \mathbf{z}_i)^{\delta_{ji}} \right) S(t_i, \boldsymbol{\psi}_j; \mathbf{z}_i)^{1-\sum_{j=1}^{K} \delta_{ji}} \right], \tag{3.22}$$

em que $\boldsymbol{\psi} = (\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \ldots, \boldsymbol{\psi_K})$ e $\boldsymbol{\psi_j} = (\alpha_j, \boldsymbol{\beta_j^T}, \rho_j, \tau_j)$,

$$f_j(t, \boldsymbol{\psi}_j; \mathbf{z_i}) = \frac{\delta CI_j(t, \boldsymbol{\psi_j}; \mathbf{z_i})}{\delta t}, \text{ para } j = 1, 2, \ldots, K \text{ e}$$

$$S(t, \boldsymbol{\psi}_j; \mathbf{z_i}) = 1 - \sum_{j=1}^{K} CI_j(t, \boldsymbol{\psi}_j; \mathbf{z_i}).$$

Utilizando algum método numérico de maximização de funções aplicado à função de log-verossilhança, obtém-se os estimadores de máxima verossimilhança para os parâmetros desconhecidos. Seguindo a mesma notação, a função de log-verossilhança é:

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{N} \left[\left(\sum_{j=1}^{K} \delta_{ji} \log(f_j(t_i, \boldsymbol{\psi}_j; \mathbf{z}_i)) \right) + \left(1 - \sum_{j=1}^{K} \delta_{ji} \right) \log(S(t_i, \boldsymbol{\psi}_j; \mathbf{z_i})) \right]. \tag{3.23}$$

O número de parâmetros estimados simultaneamente é $K \times (P + 3)$ e as quantidades de interesse (FIA, funções taxa de falha das causas, função de sobrevivência, fração de cura, etc.) serão todas derivadas desses parâmetros. Novamente, embora esse modelo seja mais flexível que o anterior assumindo diferentes transformações para riscos diferentes, uma fragilidade do modelo é que caso ocorra erro na especificação da distribuição para um dos riscos, todas as estimativas para os parâmetros dos demais riscos serão afetadas. A sugestão dos autores é testar outras distribuições para os riscos competitivos diferentes da Gompertz, de acordo com o que for observado na análise descritiva dos dados e nos resultados de cada modelo, a fim de tentar evitar esse erro de especificação da distribuição.

A crítica mais forte a esse modelo é semelhante ao trabalho anterior: para um vasto subespaço de $\boldsymbol{\Psi}$, o espaço paramétrico definido para o modelo, as principais funções em análise de sobrevivência com riscos competitivos assumem valores não razoáveis, como por exemplo a função de sobrevivência a partir de um determinado tempo pode ser negativa e as funções de incidência acumulada podem assumir valores superiores a 1. As mesmas dificuldades descritas na seção anterior serão encontradas nesse modelo se pelo menos uma das K curvas de incidência acumulada do tipo específico não apresentar platô, pois a curva de sobrevivência assume valores negativos a partir de um determinado ponto e a estimação de parâmetros por métodos numéricos será impraticável ou resultará em valores não interpretáveis. No próprio artigo, Jeong e Fine (2007) aplicam sua proposta a um conjunto de dados e as estimativas obtidas para os parâmetros geram uma curva de sobrevivência que assume valores negativos, só que para instantes de tempo muito maiores do que os presentes em seu conjunto de dados.

Ainda que a função de sobrevivência negativa não aconteça nos instantes de tempo presentes nos dados, a deficiência do modelo é classificada aqui como forte, pois, num contexto de presença de fração de cura, o interesse está em obter estimativas razoáveis e interpretáveis para essa fração, que segundo a proposta são obtidas para $t \to \infty$. A fim de contornar essa deficiência na implementação em um pacote estatístico, é necessário atentar-se para os valores que não podem ser calculados, por exemplo o logaritmo da função de sobrevivência em pontos que essa é negativa, e às propriedades básicas que as funções de incidência acumulada e de sobrevivência precisam atender para que a log-verossimilhança seja calculável e resulte em valores razoáveis. Outro ponto de atenção ao implementar o método são as possíveis complicações numéricas resultantes da forma proposta para as funções de incidência acumulada, e para evitá-las, é

importante especificar a forma de (3.20) para os casos em que os parâmetros assumem valores que geram tais complicações numéricas, como por exemplo, $\alpha \to 0$ e $\rho \to 0$.

Assim como no trabalho anterior, vale ressaltar que se todas curvas de incidência acumulada na amostra estudada apresentarem platô, a grande deficiência do modelo em apresentar função de sobrevivência negativa não acontece e o modelo é uma alternativa razoável para modelagem de dados com presença de riscos competitivos e fração de cura. Para os conjuntos de dados em que os platôs não estiverem claros para todos os riscos competitivos, esse modelo não é recomendado por essa deficiência aqui descrita.

3.2.3 Inferência

Seguindo um raciocínio semelhante ao explicado para o trabalho anterior - Jeong e Fine (2006) -, podem ser obtidas as estimativas para as variâncias a fim de realizar inferências. Também a partir delas, é possível derivar as estimativas para as variâncias das funções de incidência acumulada das diferentes causas e obter intervalos de confiança assim como já detalhado no texto na Seção 3.1.3.

Novamente, seguindo o raciocínio dos autores, a fração de cura está vinculada a cada risco separadamente e ao estudar um tipo de evento, os demais riscos que competem com ele devem ser levados em consideração para que a estimação dos parâmetros seja adequada. A fração de cura é formada por indivíduos que não apresentam falhas do tipo de interesse mesmo com um seguimento suficientemente grande, sejam eles indivíduos que não apresentam falhas de qualquer tipo ou aqueles que falharam pelos riscos competitivos. Dessa forma, o grupo em risco que inclui indivíduos que já falharam por demais causas resultante da estrutura transformada inspirada no modelo de Fine e Gray (1999) é justificado por esse raciocínio. Assim, para estimar a fração de cura $Q_{j|\mathbf{Z}}$ calcula-se em $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$:

$$\begin{aligned} Q_{j|\mathbf{Z}} &= \lim_{t\to\infty} \; 1 - CI_j(t;\mathbf{Z}) \\ &= \lim_{t\to\infty} \; \left(1 + \alpha_j \frac{\tau_j}{\rho_j}(\exp(\rho_j t) - 1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}_j\right)\right)^{\frac{-1}{\alpha_j}}. \end{aligned} \tag{3.24}$$

Mas se o interesse do pesquisador for estimar a *fração de cura global*, isto é, não atrelada somente ao risco de interesse mas a todos os eventos estudados, pode ser calculada pela expressão a seguir avaliada em $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$:

$$\begin{aligned} Q_{\cdot|\mathbf{Z}} &= \lim_{t\to\infty} \; 1 - \textstyle\sum_{j=1}^{K} CI_j(t;\mathbf{Z}) \\ &= \lim_{t\to\infty} \; 1 - \textstyle\sum_{j=1}^{K} \left\{1 - \left(1 + \alpha_j \frac{\tau_j}{\rho_j}(\exp(\rho_j t) - 1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}_j\right)\right)^{\frac{-1}{\alpha_j}}\right\}. \end{aligned} \tag{3.25}$$

Note que a partir dos resultados apresentados em (3.21), fração de cura Q_j nem sempre está bem definida tal que assuma valores no intervalo $[0,1]$. Sua existência depende que ρ_j seja negativo e também de sua relação com os demais parâmetros, pois nos casos (III) e (IV) pode assumir valores superiores a 1. A fração de cura global também não é garantida se todos os ρ_j, $j = 1,\ldots,K$ forem negativos e existirá somente se todos as funções de incidência acumulada estimadas forem impróprias.

3.3 Modelagem da Função de Incidência Acumulada via Pseudo-Valores

Nessa seção será detalhada outra técnica para modelar diretamente a função de incidência acumulada apresentada por Klein e Andersen (2005), que se baseia no uso de pseudo-valores obtidos pelo método *Jackknife* a partir da função de incidência acumulada.

O *Jackknife*, também chamado de *"leave-one-out"*, consiste em um método de reamostragem utilizado em diferentes contextos para calcular N estimativas, em um conjunto de dados com N indivíduos, pela eliminação sequencial de um indivíduo da amostra. Dessa forma, as N estimativas são obtidas por N amostras de tamanho $N-1$ que diferem do conjunto de dados total apenas pela remoção de um indivíduo por vez.

Assim como descrito em seções anteriores, a modelagem direta da função de incidência acumulada é vantajosa se o interesse reside em realizar inferências sobre os efeitos de covariáveis de forma direta na função de incidência acumulada. Uma vantagem específica dessa técnica é que, após calculados os pseudo-valores, o ajuste se resume a um modelo de regressão linear generalizado, que pode assumir diferentes funções de ligação, e que também já foi muito explorado na literatura, como em McCullagh e Nelder (1989) e Agresti e Kateri (2013), somada à aplicação de técnicas de estimação propostas e descritas em Liang e Zeger (1986), Diggle (2002) e Hardin (2005).

Esse modelo foi inicialmente proposto por Klein e Andersen (2005) para o uso em dados com presença de riscos competitivos somente. Nesse texto, o modelo é sugerido como uma proposta válida para o caso de presença simultânea de riscos competitivos e fração de cura, pois nesse modelo não existe qualquer imposição tal que a função de sobrevivência seja própria, isto é, ele não assume que todos os indivíduos irão falhar por algum tipo de evento. Dessa forma, porque é capaz de comportar a fração de cura e utilizando o raciocínio semelhante ao de Jeong e Fine (2007), que estima a fração de cura pelo complementar da função de incidência acumulada estimada, o modelo com base no uso de pseudo-valores é uma alternativa viável para análise de dados com presença de cura e riscos competitivos.

3.3.1 Modelo

Seja um conjunto de dados com K riscos competitivos composto por N indivíduos, em que para o i-ésimo indivíduo foi observado:

- T_i: o tempo da falha ou da censura;
- δ_i: o indicador de falha que assume valor 1 para a falha e 0 para a censura;
- ϵ_i: o indicador da causa da falha que assume valores entre 1 e K referenciando a causa da falha ou 0 no caso de censura;
- $\mathbf{Z_i}$: o vetor de covariáveis fixas.

Define-se, conforme descrito no Capítulo 2, o estimador não paramétrico para a função de incidência acumulada como:

$$\widehat{CI_j(t)} = \begin{cases} 0, & t < t_1 \\ \sum_{t_i \leq t} \left\{ \prod_{l=1}^{i-1} 1 - \frac{(d_l + r_l)}{Y_l} \right\} \frac{r_i}{Y_i} = \sum_{t_i \leq t} \hat{S}(t_i^-) \frac{r_i}{Y_i}, & t \geq t_1, \end{cases} \tag{3.26}$$

para $j = 1, 2, \ldots, K$, em que:

- $t_1 < t_2 < \ldots < t_N$ são os instantes distintos de falha e de censura ordenados (de qualquer um dos K tipos de falha);
- Y_i é o número de indivíduos em risco em t_i, isto é, o número de indivíduos que ainda não falharam por qualquer causa ou que não foram censurados até imediatamente antes do instante t_i.
- r_i é o número de falhas do tipo de interesse, j, em t_i.
- d_i é o número de falhas dos demais tipos, que não j, em t_i.

Esse método consiste na regressão em instantes de tempo pré-determinados, e somente para eles, as estimativas serão obtidas. Assim, dada a escolha de uma grade de M pontos $\tau_1, \tau_2, \ldots, \tau_M$, os pseudo-valores são obtidos por:

$$\theta_{jih} = N\hat{CI}_j(\tau_h) - (N-1)\hat{CI}_j^{(i)}(\tau_h), \tag{3.27}$$

para $j = 1, \ldots, K$, $i = 1, \ldots, N$, $h = 1, \ldots, M$ em que $\hat{CI}_j^{(i)}(.)$ representa o estimador da função de incidência acumulada calculado a partir da amostra na qual foi removido o i-ésimo indivíduo.

O número de pseudo-valores obtidos será $N \times M \times K$ se a grade de pontos escolhida for de tamanho M. Apesar da escolha $M = 1$ ser suficiente para que os parâmetros do modelo sejam estimados, é mais eficiente escolher diversos pontos para esse ajuste. No trabalho de Klein e Andersen (2005), os autores simulam qual é o impacto nas estimativas e suas variâncias com o uso de uma grade contendo 5, 10 e 20 pontos e concluem que pouca diferença foi observada no conjunto de dados utilizado por eles. Mas é importante ressaltar que, em um conjunto de dados diferente, essa conclusão deveria ser ao menos testada. Uma opção alternativa é modificar o método de estimação e ajustar o modelo baseando-se em todos os pontos da amostra, como descrito no trabalho de Scheike e Zhang (2007). Nesse trabalho será adotada a metodologia proposta por Klein e Andersen (2005) que se vale do uso de uma grade de pontos finita para o cálculo dos pseudo-valores e consequente estimação dos efeitos de covariáveis sobre a função de incidência acumulada, apesar da existência de uma proposta mais completa, que é muito exaustiva para as simulações que compararão os métodos nesse trabalho.

Outro ponto importante a se destacar é que, se o conjunto de dados não tiver nenhuma censura, adotando a mesma notação de (3.26):

$$\hat{S}(t_i^-) = \prod_{l=1}^{i-1} 1 - \frac{(d_l + r_l)}{Y_l} = \prod_{l=1}^{i-1} \frac{(Y_l - d_l - r_l)}{Y_l} = \prod_{l=1}^{i-1} \frac{Y_{l+1}}{Y_l} = \frac{Y_i}{Y_1},$$

e denotando Y_1 e $Y_1^{(i)}$ respectivamente como o número de indivíduos em risco no caso da amostra completa e no caso da amostra após a remoção do i-ésimo indivíduo, observa-se que para um dado j e τ_h:

$$\hat{CI}_j(\tau_h) = \sum_{t_l<t} \hat{S}(t_l^-)\frac{r_l}{Y_l} = \sum_{t_l<t} \frac{Y_l}{Y_1}\frac{r_l}{Y_l} = \frac{1}{N}\sum_{t_l<t} r_l \text{ e}$$

$$\hat{CI}_j^{(i)}(\tau_h) = \sum_{t_l<t} \hat{S}^{(i)}(t_l^-)\frac{r_l^{(i)}}{Y_l^{(i)}} = \sum_{t_l<t} \frac{Y_l^{(i)}}{Y_1^{(i)}}\frac{r_l^{(i)}}{Y_l^{(i)}} = \frac{1}{(N-1)}\sum_{t_l<t} r_l^{(i)}.$$

Portanto substituindo os valores obtidos acima na equação de cálculo dos pseudo-valores, conclui-se que:

$$\theta_{jih} = N\frac{1}{N}\sum_{t_l<t} r_l - (N-1)\frac{1}{(N-1)}\sum_{t_l<t} r_l^{(i)} = \sum_{t_l<t} r_l - \sum_{t_l<t} r_l^{(i)} = \mathbb{1}(T_i \leq \tau_h, J = j), \tag{3.28}$$

para $j = 1, 2, \ldots, K$, $i = 1, \ldots, N$ e $h = 1, \ldots, M$. Em palavras, quando não há censura nos dados, o pseudo-valor θ_{jih} se resume ao indicador da ocorrência da falha do tipo j para o i-ésimo indivíduo até o instante no tempo τ_h e portanto, os pseudo-valores para indivíduos distintos são independentes. Assim, no caso sem censura de dados é trivial concluir que:

$$E(\theta_{jih}) = E\left(\mathbb{1}(T_i \leq \tau_h, J = j)\right) = P\left(T_i \leq \tau_h, J = j\right) = CI_j(\tau_h), \tag{3.29}$$

para $j = 1, 2, \ldots, K$, $i = 1, \ldots, N$ e $h = 1, \ldots, M$.

No caso de presença de censura nos dados, os pseudo-valores serão valores próximos a esses indicadores: indivíduos que apresentaram a falha do tipo j até o tempo τ_h terão pseudo-valores próximos a 1; indivíduos que não apresentaram falha ou apresentaram outros tipos de falha que não j até o tempo τ_h terão pseudo-valores próximos ou menores que 0; enquanto os indivíduos que foram censurados até o tempo τ_h apresentarão valores entre 0 e 1 (Andersen e Perme, 2009). No trabalho de Graw et al. (2009), encontra-se um desenvolvimento teórico que demonstra que os pseudo-valores obtidos por (3.27), são estimadores assintoticamente não viesados para a função de incidência acumulada, condicionalmente às covariáveis, assumindo que o mecanismo de censura é independente dos tempos de evento e das covariáveis. Em resumo, considerando as premissas:

- (P1) A censura é estocasticamente independente dos tempos e tipos de eventos e das covariáveis;
- (P2) O resultado é válido para todo t, tal que $t < \phi$, em que $P(C_i > \phi) > 0$ e C_i é o tempo de censura,

os autores demonstram as seguintes propriedades para os pseudo-valores:

- $\theta_{jih}, i = 1, \ldots, N$, podem ser aproximados por variáveis aleatórias independentes e identicamente distribuídas;

- $E\left(\theta_{jih}\right) = CI_j(\tau_h) + o_P(1)$, para $i = 1, \ldots, N$, $j = 1, \ldots, K$ e $h = 1, \ldots, M$;
- $E\left(\theta_{jih}|\mathbf{Z_i}\right) = CI_j(\tau_h|\mathbf{Z_i}) + o_P(1)$, para $i = 1, \ldots, N$, $j = 1, \ldots, K$ e $h = 1, \ldots, M$.

O leitor interessado na demonstração completa deve recorrer ao texto original (Graw et al., 2009), pois nesse texto são mencionadas apenas principais propriedades dos pseudo-valores que justificam seu uso como estimadores da função de incidência acumulada, e que garantem a consistência e normalidade assintótica desses estimadores obtidos mesmo com presença de censura nos dados. No caso de ausência de censura nos dados, a demonstração decorre de (3.28) e de (3.29) e assim, se $\mathbf{Z_i}$ for o vetor de covariáveis para o i-ésimo indivíduo, utilizando as mesmas premissas (P1) e (P2), temos:

- $\theta_{jih} = \mathbb{1}(T_i \leq \tau_h, J = j), i = 1, \ldots, N$ e portanto, são variáveis aleatórias independentes e identicamente distribuídas;
- $E\left(\theta_{jih}\right) = CI_j(\tau_h)$, para $i = 1, \ldots, N$, $j = 1, \ldots, K$ e $h = 1, \ldots, M$;
- $E\left(\theta_{jih}|\mathbf{Z_i}\right) = CI_j(\tau_h|\mathbf{Z_i})$, para $i = 1, \ldots, N$, $j = 1, \ldots, K$ e $h = 1, \ldots, M$.

Essa demonstração é essencial para garatir o uso apropriado das equações de estimação generalizadas (GEE) combinadas ao uso dos pseudo-valores como estimadores da função de incidência acumulada no modelo proposto por Klein e Andersen (2005). As equações de estimação generalizadas compõem um método de estimação de parâmetros sugerido por Liang e Zeger (1986) para uso em dados longitudinais. Na proposta de Klein e Andersen (2005), os pseudo-valores são interpretados como dados longitudinais, isto é, dados obtidos através de leituras repetidas para um mesmo indivíduo em instantes diferentes de tempo, para então serem estimados os parâmetros associados aos efeitos das covariáveis.

Para a definição do modelo, é escolhida uma função de ligação $g(.)$ dentre essas possibilidades aplicáveis:

- Logito: $g(x) = \log(x/(1-x))$;
- Complemento do log-log em x: $g(x) = -\log(-\log(x))$;
- Complemento do log-log em $1-x$: $g(x) = -\log(-\log(1-x))$,

e assume-se o modelo linear generalizado da forma:

$$g(\theta^*_{jih}) = \alpha_{jh} + \boldsymbol{\gamma}_j^T \mathbf{Z}_i = \boldsymbol{\beta}_{jh}^T \mathbf{Z}_{ih}, \tag{3.30}$$

para $j = 1, 2, \ldots, K$, $i = 1, \ldots, N$ e $h = 1, \ldots, M$, em que:

- $E(\theta_{jih}) = \theta^*_{jih}$;
- α_{jh} representa o termo da combinação linear associado ao tempo τ_h e ao risco j;
- $\boldsymbol{\gamma}_j$ é o vetor $p \times 1$ dos parâmetros associados ao efeitos das covariáveis no risco j;

- $\mathbf{Z}_i$ é o vetor $p \times 1$ das covariáveis observadas para o i-ésimo indivíduo;
- $\boldsymbol{\beta}_{jh}$ é o vetor $(p+1) \times 1$ da forma $\boldsymbol{\beta}_{jh}^T = \left[\alpha_{jh} \;\; \boldsymbol{\gamma}_j^T\right]$;
- $\mathbf{Z}_{ih}$ é o vetor $(p+1) \times 1$ da forma $\mathbf{Z}_{ih}^T = \left[1 \;\; \mathbf{Z}_i^T\right]$.

Esse modelo tem uma característica principal que o difere dos demais: apenas os pontos escolhidos para a grade $\tau_1, \tau_2, \ldots, \tau_M$ terão estimativas tanto para o efeito das covariáveis quanto para as funções de incidência acumulada, o que limita algumas inferências mas pode ser suficiente para atender a necessidade de um pesquisador, dependendo do seu interesse. Se o interesse for especial na fração de cura, como a escolha da grade deve ser de pontos presentes na amostra, fica claro aqui o quanto é importante nesse modelo que o seguimento do estudo seja longo o suficiente para que a escolha do último ponto da grade seja representativa da real fração de cura. Assim como nos modelos anteriores, esse modelo não permite a incorporação de covariáveis explicativas diretamente para a estimação da cura. Uma diferença é que, apesar de não existir especificação ou imposição de relação entre os diferentes riscos, esse modelo não leva aos mesmos problemas das abordagens anteriores pois como todos os pontos da curva a serem estimados são pontos existentes na amostra, o modelo não sofre com extrapolações em tempos não presentes na amostra que poderiam geram estimativas de sobrevivência negativas e não razoáveis.

3.3.2 Estimação dos Parâmetros

Conforme previamente descrito, os pseudo-valores são considerados dados longitudinais e para a estimação dos parâmetros de interesse é utilizado o método das equações de estimação generalizadas (GEE). O GEE é considerado um método de estimação e não um modelo, pois nele não existe qualquer associação de distribuição para a variável resposta. Essa ausência de distribuição e consequente ausência de função verossimilhança limita o uso de testes e inferências baseados em métodos que requerem a verossimilhança. Apesar da limitação, o método não requer muito esforço computacional quando comparado com a maximização da verossimilhança e trata-se de uma alternativa que requer somente a especificação dos primeiros dois momentos.

Para utilizar o método, define-se $\boldsymbol{\theta}_{ji} = \left(\theta_{ji1}, \ldots, \theta_{jiM}\right)^T$, $c_{ji} = \left(CI_j(\tau_1|\mathbf{Z}_i), \ldots, CI_j(\tau_M|\mathbf{Z}_i)\right)^T$. A função de ligação inversa é dada por:

$$\boldsymbol{\theta}_{ji}^* = g^{-1}(\mathbf{Z}_i^T \boldsymbol{\beta}_j), \tag{3.31}$$

em que $\mathbf{Z}_i$ é matriz $(M+p) \times M$ e $\boldsymbol{\beta}_j$ é vetor $(M+p) \times 1$ da forma:

$$\mathbf{Z}_i^T = \begin{bmatrix} 1 & 0 & \ldots & 0 & z_{i1} & \ldots & z_{ip} \\ 0 & 1 & \ldots & 0 & z_{i1} & \ldots & z_{ip} \\ \vdots & \vdots & \ddots & \vdots & z_{i1} & \ldots & z_{ip} \\ 0 & 0 & \ldots & 1 & z_{i1} & \ldots & z_{ip} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\beta}_j = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_M \\ \boldsymbol{\gamma} \end{bmatrix}.$$

Denotando $\partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j) = \frac{\partial g^{-1}(\mathbf{Z}_i^T\boldsymbol{\beta}_j)}{\partial\boldsymbol{\beta}_j}$ e a matriz de covariância de trabalho por $V_i^{-1}(\boldsymbol{\beta}_j)$, as estimativas dos parâmetros associados aos efeitos das covariáveis serão encontradas pelas soluções das equações a seguir:

$$U(\boldsymbol{\beta}_j) = \sum_i \partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j)V_i^{-1}(\boldsymbol{\beta}_j)(\boldsymbol{\theta}_{ji} - c_{ji}) = \mathbf{0}. \tag{3.32}$$

Portanto, para a obtenção dessas estimativas, além da parte sistemática definida em (3.30) é necessário estabelecer a escolha da matriz de covariância de trabalho. Algumas opções são a matriz identidade, a matriz de covariância auto-regressiva, que supõe correlação maior entre observações mais próximas no tempo e menor conforme a distância de tempo aumenta, ou mesma a covariância "exata", em que todos os elementos da matriz são formados por parâmetros a se estimar. No entanto, no trabalho de Klein e Andersen (2005) os autores mostram que diferenças irrelevantes são derivadas dessa escolha e sugerem que o uso da matriz identidade para a matriz de covariância de trabalho é a alternativa simples e que não afeta a inferência. Outra sugestão oferecida pelos autores para reduzir a instabilidade numérica quando se pretende trabalhar com a matriz exata é escolher inicialmente a matriz identidade a fim de obter a matriz de covariância exata e então utilizá-la num segundo passo de forma fixa nas equações de estimação generalizadas.

Como as estimativas para a função de incidência acumulada nesse método são obtidas por:

$$\hat{CI}^{PS}_{j|\mathbf{Z}}(\tau_h) = g^{-1}(\hat{\alpha}_{jh} + \hat{\boldsymbol{\gamma}}_j^T\mathbf{Z}) = g^{-1}(\hat{\boldsymbol{\beta}}_{jh}^T\mathbf{Z}_h), \tag{3.33}$$

seguindo o raciocínio semelhante ao de (3.13) e (3.14), as frações de cura por risco e global podem ser estimadas por:

$$\hat{Q}_{j|\mathbf{Z}} = 1 - g^{-1}(\hat{\alpha}_{jM} + \hat{\boldsymbol{\gamma}}_j^T\mathbf{Z}) \quad \text{e} \quad \hat{Q}_{\cdot|\mathbf{Z}} = 1 - \textstyle\sum_{j=1}^K \hat{CI}^{PS}_{j|\mathbf{Z}}(\tau_M). \tag{3.34}$$

3.3.3 Inferência

Seguindo o desenvolvimento de Liang e Zeger (1986) e denotando por $\hat{\boldsymbol{\beta}}_j$ a solução de (3.32), apesar de não envolver o uso de verossimilhança, assim como descrito anteriormente, no trabalho de Graw et al. (2009) os autores demonstram propriedades para os pseudo-valores que garantem as propriedades desejadas para o uso das equações de estimação generalizadas e assim, $\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ é assintoticamente normal com média zero e covariância que pode ser estimada por:

$$\hat{\Sigma} = I(\hat{\boldsymbol{\beta}}_j)^{-1}\hat{\text{var}}(U(\hat{\boldsymbol{\beta}}_j))I(\hat{\boldsymbol{\beta}}_j)^{-1} \tag{3.35}$$

em que

$$I(\boldsymbol{\beta}_j) = \sum_i \partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j)V_i^{-1}(\boldsymbol{\beta}_j)\partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j)^T \text{ e}$$

$$\hat{\text{var}}(U(\hat{\boldsymbol{\beta}}_j)) = \sum_i \left(\partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j)V_i^{-1}(\boldsymbol{\beta}_j)(\hat{\boldsymbol{\theta}}_i - c_i)\right)\left(\partial\boldsymbol{\mu}_i(\boldsymbol{\beta}_j)V_i^{-1}(\boldsymbol{\beta}_j)(\hat{\boldsymbol{\theta}}_i - c_i)\right)^T.$$

Assim como em (3.32), a matriz de covariância de trabalho deve ser pré-especificada para

a obtenção das estimativas das variâncias, mas os autores mostram que pouca diferença é observada no conjunto de dados utilizado em sua aplicação. A fim de testar tal afirmação para o conjunto de dados que motivou essa dissertação, na seção de aplicação o leitor pode encontrar a comparação entre os valores obtidos para as estimativas e variâncias estimadas utilizando a matriz exata e a matriz identidade como possíveis escolhas para a matriz de covariância de trabalho.

Capítulo 4

Simulação

A fim de avaliar e comparar a performance dos modelos descritos no capítulo anterior, neste trabalho foram gerados diversos conjuntos de dados simulados sob diferentes cenários de porcentagens de cura, censura, número de observações e distribuições para que fossem ajustados os métodos descritos e assim, fosse possível a comparação da qualidade das estimativas. Nesse capítulo, o leitor poderá encontrar uma breve descrição da metodologia utilizada para geração desses conjuntos de dados, a lista dos cenários testados, os principais resultados desse estudo por simulação para cada modelo e uma comparação entre eles.

Os dados foram gerados através do método sugerido em Beyersmann et al. (2011b), que se baseia em propriedades das principais funções em análise de sobrevivência com riscos competitivos e dispensa a necessidade de especificação de dependência entre os diferentes riscos. Nesse trabalho, foram consideradas duas distribuições para a geração de dados, a Weibull e a Gompertz. A primeira foi escolhida por permitir uma geração tal que não houvesse preferência natural por um dos modelos, e porque ao mesmo tempo, é flexível para assumir funções de incidência acumulada com formas bem diferentes das geradas pela distribuição Gompertz e semelhantes às encontradas no conjunto de dados utilizado no Capítulo 5. A segunda foi escolhida para testar o desempenho dos modelos baseados nessa distribuição sob o caso de especificação correta da distribuição.

Sabendo que a função da taxa de falha global é o somatório de todas as funções taxa de falha das causas específicas, o método de geração de um conjunto de dados com N observações para a distribuição Weibull segue os seguintes passos:

1. Especificar as funções da taxa de falha da causa específica, $\lambda_j(t), j = 1, \ldots, K$, para cada uma das K causas;

2. Gerar os tempos de falha $t_i, i = 1, \ldots, N$ de acordo com a função da taxa de falha global acumulada, obtida por $\int_0^t \lambda(u)du = \sum_{j=1}^{K} \int_0^t \lambda_j(u)du$;

3. Definir a causa da falha através de um experimento multinomial, em que a i-ésima observação tem probabilidade de ser do tipo j igual a $\dfrac{\lambda_j(t)}{\sum_{j=1}^{K} \lambda_j(t)}$. Note que esse passo é

resultado de $P(J = j|T = t) = \dfrac{P(J = j, T = t)}{\sum_{j=1}^{K} P(J = j, T = t)} = \dfrac{\lambda_j(t)}{\sum_{j=1}^{K} \lambda_j(t)}$;

4. Gerar o indicador de suscetibilidade $y_i, i = 1, \ldots, N$, por experimento Bernoulli com $P(Y_i = 1) = \pi$;

5. Gerar os tempos de censura $c_i, i = 1, \ldots, N$, da forma usual, e censurar os dados se $c_i < t_i$ ou se $y_i = 1$;

No caso da distribuição Gompertz, a presença de cura é consequência natural da impropriedade da distribuição, e portanto, foi suficiente escolher parâmetros tais que houvesse presença de cura nos dados. Assim, a geração dos dados segue o algoritmo:

1. Especificar as funções de incidência acumulada $CI_j(t)$, $j = 1, \ldots, K$, para cada uma das K causas;

2. Gerar um valor aleatório u_i distribuído pela Uniforme[0,1];

3. Se o valor u_i pertencer à imagem da função de probabilidade acumulada $F(t) = \sum_{j=1}^{K} CI_j(t)$, o tempo será o valor t_i tal que $F(t_i) = u_i$; Se o valor u_i não pertencer à imagem dessa função, o tempo assume o maior valor possível pré-determinado;

4. Definir a causa da falha através de um experimento multinomial, em que a i-ésima observação tem probabilidade de ser do tipo j igual a $\dfrac{\lambda_j(t_i)}{\sum_{j=1}^{K} \lambda_j(t_i)}$. Note que esse passo é resultado de $P(J = j|T = t) = \dfrac{P(J = j, T = t)}{\sum_{j=1}^{K} P(J = j, T = t)} = \dfrac{\lambda_j(t)}{\sum_{j=1}^{K} \lambda_j(t)}$;

5. Gerar os tempos de censura $c_i, i = 1, \ldots, N$, da forma usual, e censurar os dados se $c_i < t_i$.

Em ambos os casos, além da censura aleatória, os dados foram gerados com um valor máximo de tempo observado pré-determinado, a fim de espelhar os dados da aplicação.

Foram simulados 2000 conjuntos de dados para cada um dos 48 cenários diferentes, que além de considerarem as duas distribuições escolhidas, foram gerados com dois diferentes números de observações, com quatro opções de censura e três proporções diferentes de cura. Ao longo do texto, estrato será utilizado para denotar a porção da população gerada pelas possíveis combinações entre níveis das covariáveis. Assim, todos os dados simulados continham dois riscos competitivos e no caso da Weibull, seis estratos formados por uma variável binária e outra categórica com três níveis. No caso da Gompertz, dois estratos formados por uma variável binária. Como não era necessário e nem possível comparar dois cenários gerados pelas diferentes distribuições, já que as proporções de curas eram diferentes, não foi necessário que o número de estratos fosse igual. Na Gompertz, a inclusão de muitos estratos não mudava a forma das curvas e portanto, foi suprimida uma das covariáveis. É importante porém, ao realizar as leituras, lembrar que no caso da Gompertz o número de observações em cada estrato é superior ao caso da Weibull. A seguir, na Figura 4.1 encontra-se a lista completa dos cenários simulados.

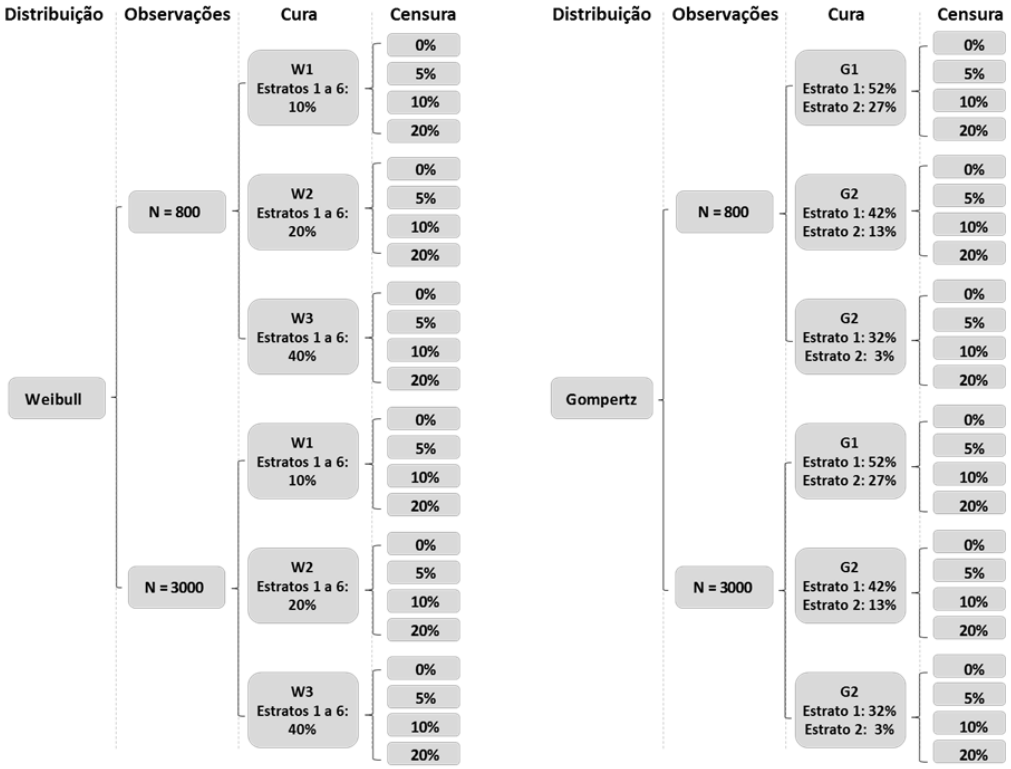


Figura 4.1: Ilustração dos 48 cenários simulados.

As funções de incidência acumulada teóricas simuladas estão apresentadas nas Figuras 4.2 e 4.3. Note que a forma das curvas simuladas pela Weibull propositalmente difere das distribuições Gompertz quando assume sua forma imprópria pois apresentam um ponto de inflexão no tempos iniciais, ou incidência acentuada em um ponto. Essa característica foi escolhida estrategicamente para espelhar uma propriedade do conjunto de dados utilizado no capítulo de aplicações, que motivou esse estudo, e para avaliar os métodos paramétricos sob o erro de espeficicação da distribuição também.

Outro ponto importante é notar que, como os dados foram simulados, é possível distinguir a observação que é curada da observação censurada e portanto, ao longo desse capítulo, por proporção de censura entende-se somente a proporção de indivíduos que não são curados mas foram censurados e por proporção de cura entende-se somente as proporções de indivíduos curados. Assim, a proporção de cura a de censura não de sobrepõem, isto é, se a proporção de censura for 20% e a de cura for 30%, os 50% indivíduos remanescentes apresentarão falhas. As proporções de curas e censuras foram simuladas tais que somadas não ultrapassassem 60%, que é a proporção encontrada no conjunto de dados utilizado na aplicação. Essa simulação poderia ser tão exaustiva quanto se desejasse, testando números de observações menores, outras distribuições, outras proporções de cura e censura, mas a princípio as escolhas acima explicadas buscavam responder as seguintes perguntas:

- A qualidade das estimativas para cada método é muito vulnerável ao número de in-

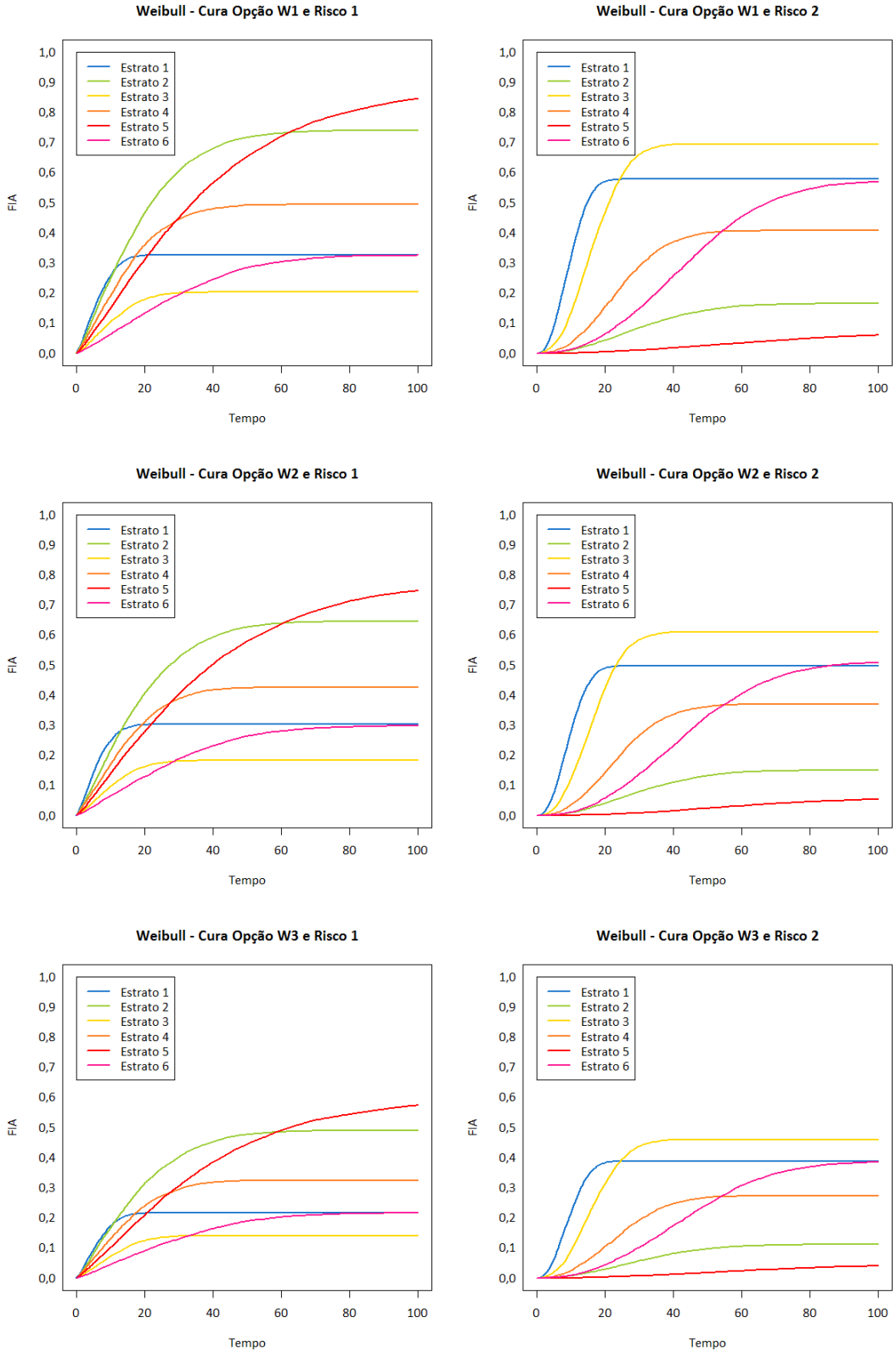


Figura 4.2: Funções de Incidência Acumulada simuladas pela distribuição Weibull.

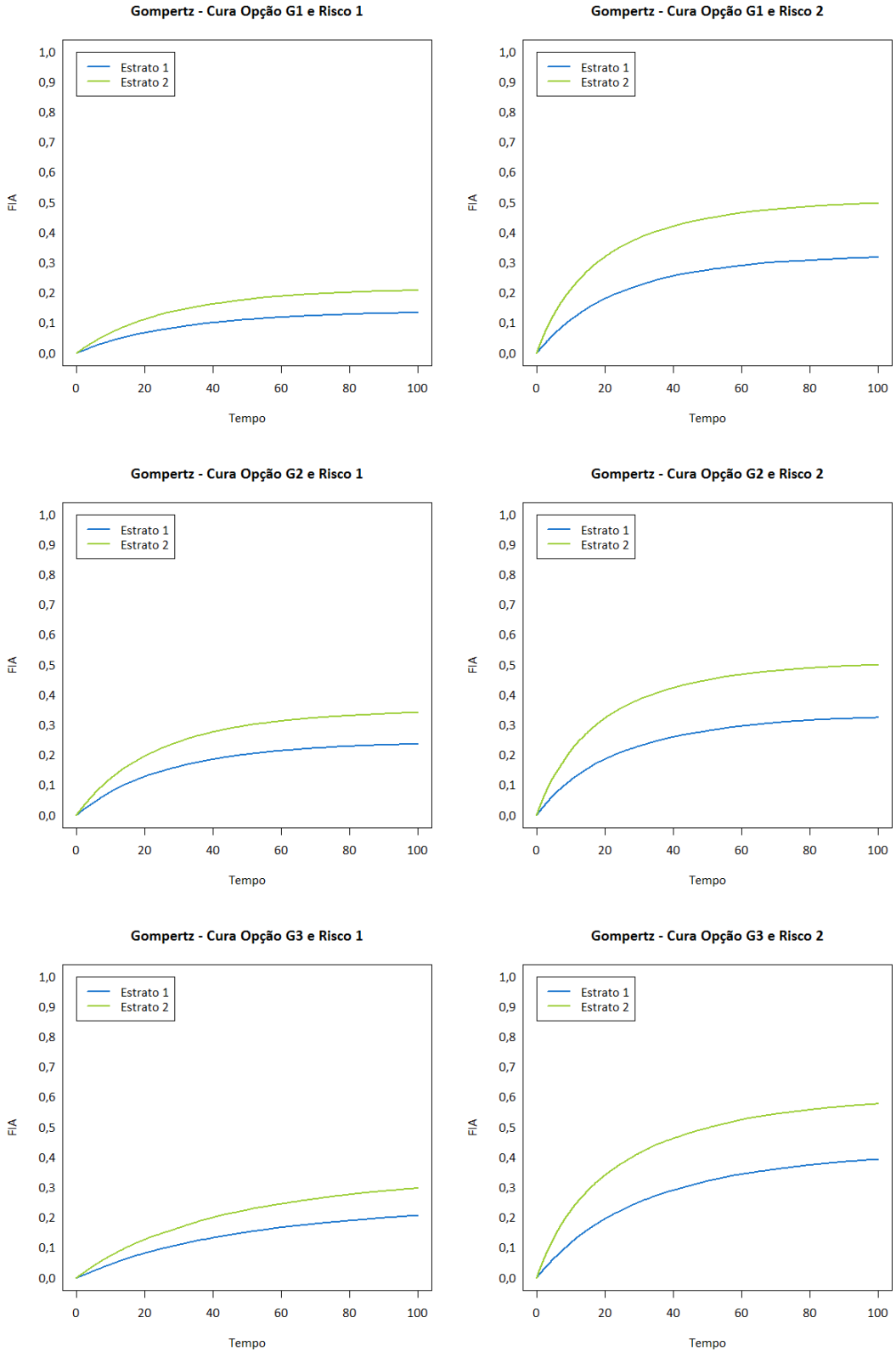


Figura 4.3: Funções de Incidência Acumulada simuladas pela distribuição Gompertz.

divíduos na amostra?

- A qualidade das estimativas para cada método é muito vulnerável às proporções de censura?
- A qualidade das estimativas para cada método é muito vulnerável às proporções de cura?
- O método paramétrico é muito vulnerável ao erro de especificação da distribuição?
- O método com uso de pseudo-valores é uma alternativa boa para dados com presença de cura?
- Qual método apresenta as *melhores estimativas?*

Nas seções a seguir encontram-se os resultados resumidos para cada um dos modelos testados e na última seção, uma comparação entre eles.

4.1 Método Paramétrico sem Covariáveis

O primeiro ponto a se destacar sobre esse modelo é que como ele não contempla a presença de covariáveis, as estimativas foram comparadas às funções de incidência acumulada não estratificadas. Quando avaliado o viés médio das estimativas para cada cenário, não foram observadas diferenças significantes para diferentes proporções de censura, e pouca diferença entre cenários de cura diferentes, como está apresentado na Figura 4.4. Nos dados gerados pela distribuição Gompertz, o viés é muito próximo de zero. Já nos dados simulados pela Weibull, nota-se que o modelo superestima os menores tempos e subestima os maiores, pois não consegue capturar a forma da curva que contém um ponto de inflexão no seu início.

Outro ponto importante é que a média das estimativas da variância é sempre muito superior à variância observada para as estimativas, além de ser crescente conforme aumenta o valor do tempo. Portanto, se o tamanho da amostra for pequeno, o modelo oferece inferências pouco úteis para esses pontos no tempo com valores altos, o que pode ser entendido como uma limitação do modelo quando o interesse reside no estudo da fração de cura, já que ela é obtida pelo complementar do somatório das funções de incidência acumulada em um ponto no tempo suficientemente grande. A Tabela 4.1 mostra o primeiro quartil, a mediana, o terceiro quartil e a média das variâncias estimadas para as 2000 estimativas das funções de incidência acumulada no tempo igual a 100, para um mesmo cenário de cura (W1) e para as diferentes proporções de censura e tamanho de amostra. A partir dos valores obtidos para $N = 800$, podemos dizer que as variâncias estimadas como sugeridas no trabalho de Jeong e Fine (2006) são pouco úteis por serem muito altas, além de muito superiores se comparadas à variância observada para as estimativas nessa simulação. O tamanho da amostra precisa ser maior para que essas estimativas possam fornecer informações úteis. Apesar da escolha de um dos cenários para ilustrar essa argumentação, esse padrão se repete em todos os riscos, em ambos casos simulados pela Weibull e pela Gompertz e em todas as proporções de cura testadas.

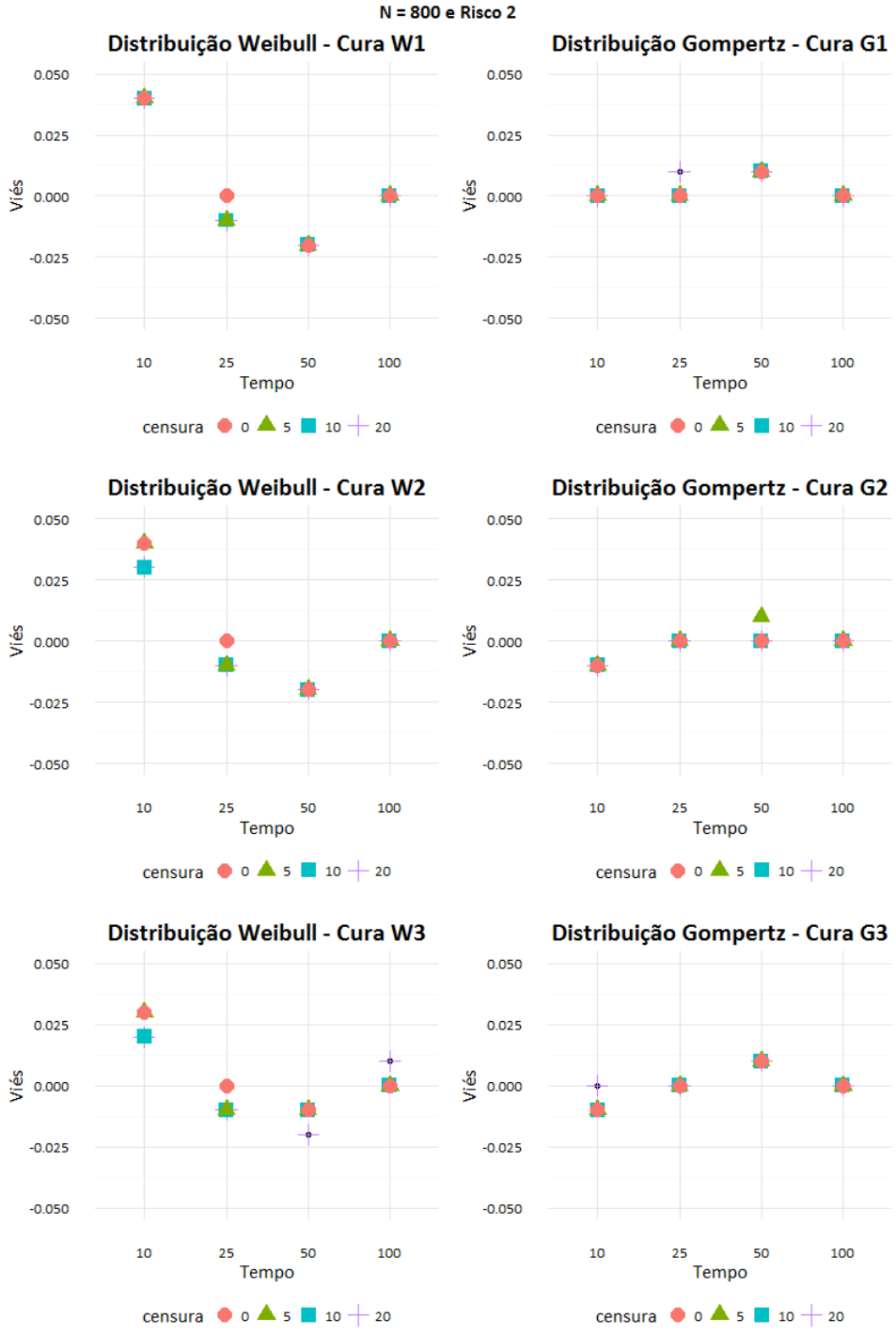


Figura 4.4: Comparação entre o viés médio das estimativas da FIA em diferentes pontos, proporções de cura e censura.

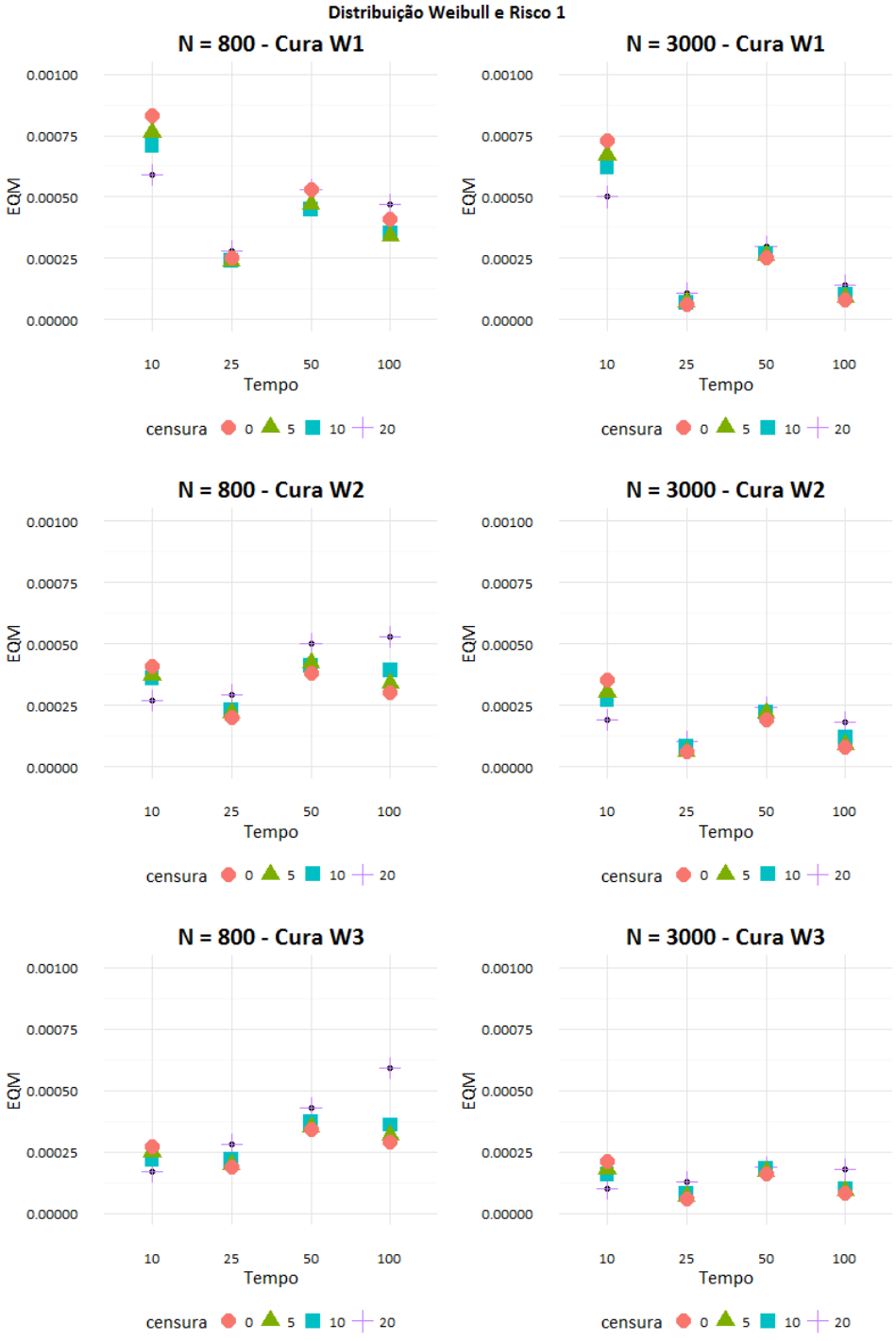


Figura 4.5: Comparação entre os EQMs para diferentes pontos da FIA, com diferentes proporções de cura e censura.

Tabela 4.1: Comparação entre estimativas das variâncias e variância observada para a FIA nos cenários de cura W1, risco 1 e distribuição utilizada Weibull.

	N = 800				N = 3000			
	Censura				Censura			
	0%	5%	10%	20%	0%	5%	10%	20%
Variância Observada	0,0004	0,0003	0,0003	0,0004	0,0001	0,0001	0,0001	0,0001
Média das Variâncias Estimadas	0,5024	0,5336	0,5664	0,6489	0,1343	0,1422	0,1501	0,1726
Primeiro Quartil das Variâncias Estimadas	0,4681	0,4961	0,5253	0,597	0,1299	0,1374	0,1448	0,1653
Mediana das Variâncias Estimadas	0,5011	0,5305	0,5625	0,6468	0,1339	0,142	0,1501	0,1723
Terceiro Quartil das Variâncias Estimadas	0,5343	0,5671	0,6034	0,697	0,1386	0,1467	0,1551	0,1793

Além dos pontos citados, o erro quadrático médio (EQM) é relativamente muito menor com o aumento do número de observações de 800 para 3000 e é influenciado pela proporção de censura, mas não fica evidente uma deterioração pela presença de uma proporção de cura maior. Na Figura 4.5 essa diferença está ilustrada para os dados simulados pela Weibull e para o risco 1, sendo que as proporções de curados são para $W1 = 10\%$, $W2 = 20\%$ e $W3 = 40\%$. Note que a redução do erro quadrático médio é maior conforme aumenta o tempo, reforçando a necessidade de uma amostra suficientemente grande para que inferências úteis sejam realizadas sobre os tempos maiores.

4.2 Método Paramétrico com Covariáveis

O modelo proposto por Jeong e Fine (2007) foi avaliado sob o erro de especificação do modelo e sob a especificação correta. Assim como no modelo anterior, para os cenários gerados pela distribuição Gompertz, o viés é, para todos os cenários e pontos, praticamente igual a zero, como pode ser visto nas Figura 4.6 em que foram escolhidos dois estratos para ilustrar tal argumentação. Dessa forma, a análise do erro quadrático médio é o mesmo que avaliar a variância das estimativas obtidas pelo modelo.

Sob a especificação correta da distribuição, observa-se que a média das variâncias estimadas para as funções de incidência acumulada é muito semelhante à variância das estimativas obtidas, mostrando que a forma de estimação da variância sugerida é bastante precisa, e pouco influenciada pela proporção de cura, censura ou tamanho da amostra. Na Tabela 4.2 encontra-se a comparação lado a lado entre a variância das estimativas (VE) e a média das variâncias estimadas (MV) para os cenários simulados pela Gompertz com Cura G1, risco 1 e estrato 1. Apesar de somente um risco e escolha de parâmetros G1 ser apresentada aqui, esse padrão se repete para todos os cenários simulados pela Gompertz. Note que, também como o modelo anterior, tanto a variância estimada como a observada aumentam conforme o valor do tempo aumenta, mas que diferente do modelo anterior, assumem valores baixos já com 800 indivíduos e portanto úteis para realizar inferências, mesmo em valores de tempo maiores.

Como parte dos dados foi gerado justamente seguindo esse modelo paramétrico, uma análise adicional foi realizada avaliando as estimativas de cada parâmetro. Na Tabela 4.3 é

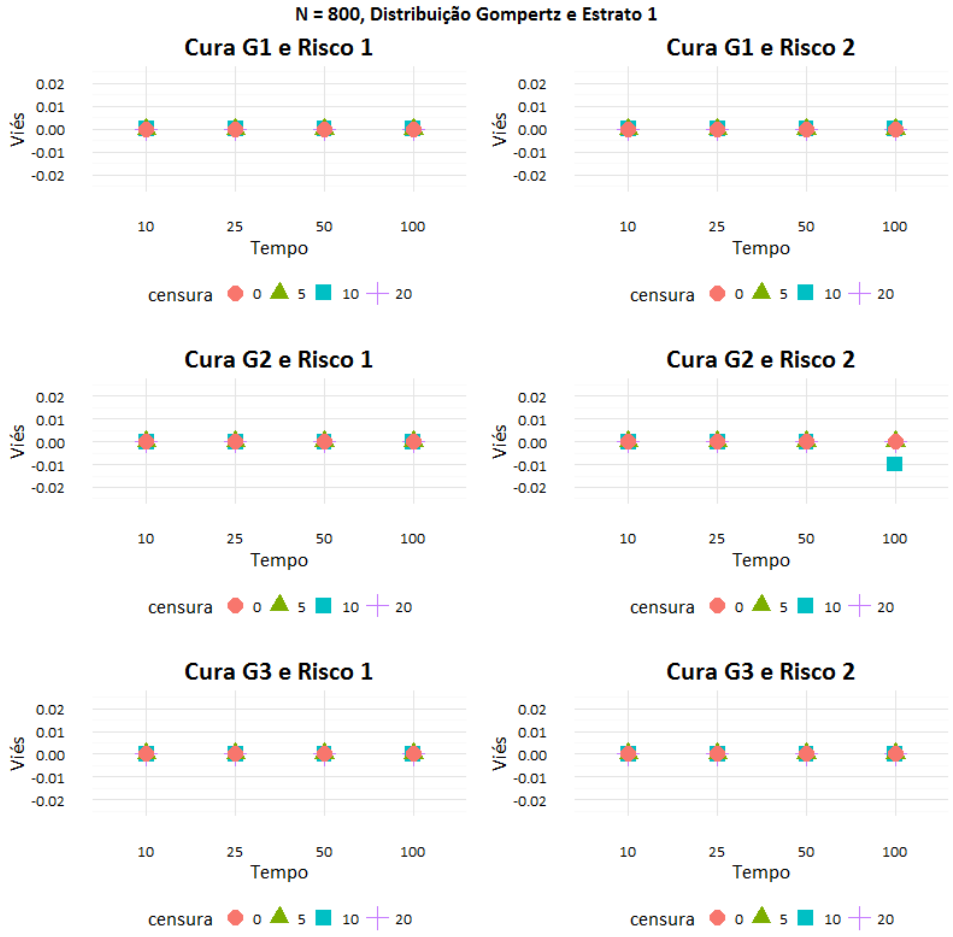


Figura 4.6: Comparação entre o viés médio das estimativas da FIA observado para o estrato 1, riscos 1 e 2 dos cenários gerados pela Gompertz com N = 800.

Tabela 4.2: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA no risco 1, estrato 1, nos cenários de cura G1 e distribuição utilizada Gompertz.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00005	0,00005	0,00005	0,00005	0,00006	0,00006	0,00007	0,00007
	25	0,00014	0,00014	0,00016	0,00015	0,00016	0,00016	0,00020	0,00019
	50	0,00024	0,00023	0,00026	0,00024	0,00026	0,00027	0,00034	0,00033
	100	0,00031	0,00029	0,00033	0,00032	0,00035	0,00037	0,00056	0,00058
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00001	0,00001	0,00001	0,00001	0,00001	0,00001	0,00002	0,00002
	25	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00005	0,00005
	50	0,00006	0,00006	0,00006	0,00006	0,00008	0,00007	0,00009	0,00009
	100	0,00008	0,00008	0,00008	0,00008	0,00011	0,00010	0,00015	0,00015

possível ver que as estimativas médias são exatamente iguais aos valores reais dos parâmetros independente da censura e para os três conjuntos de parâmetros testados, com exceção dos parâmetros α que são sempre muito diferentes dos reais. Através da análise da estatística C90, que representa a porcentagem de amostras que o intervalo de confiança 90% cobre o valor verdadeiro, nota-se que com exceção do parâmetro α, todos os demais parâmetros são cobertos pelos intervalos de confiança em aproximadamente 90% dos casos, como desejado. A variância das estimativas é também sempre muito semelhante à média das variâncias estimadas, e aumenta com maiores proporções de censura assim como esperado, desconsiderando os parâmetros α. Para esses parâmetros, as estimativas apresentam uma variabilidade muito grande e o viés médio é alto. Além disso, a média das variâncias estimadas é muito maior que a variância das estimativas para todos os cenários, e quando analisados os primeiros e terceiros quartis para as variâncias estimadas, fica claro que a variância estimada não é precisa. Por consequência, a cobertura dos intervalos de confiança 90% não apresentam valores próximos a 90%. Apesar da imprecisão observada para os parâmetros α, vale reforçar que as estimativas das funções de incidência acumulada apresentam viés muito próximo de zero e variabilidade baixa, mostrando a pouca influência desse parâmetro no modelo. Para evitar essa complicação, é possível seguir uma alternativa sugerida pelos autores que se trata do uso do modelo com esse parâmetro pré-fixado, e posterior definição do melhor modelo pelo AIC. Nos exercícios de simulação desse trabalho, porém, não foi avaliado quais seriam os resultados se os parâmetros α fossem pré-fixados no ajuste.

Sob o erro de especificação da distribuição, nota-se que esse modelo não consegue capturar a forma das funções de incidência acumulada simuladas pela Weibull, que contém um ponto de inflexão nos tempos iniciais. Para as curvas que assumem essa forma, o modelo superestima a FIA nos pontos iniciais e subestima nos pontos maiores no tempo. Para as curvas que possuem uma incidência alta em algum ponto inicial e logo estabilizam, como é o caso do risco 1 e estrato

Tabela 4.3: Comparação entre estimativas médias (E), variâncias das estimativas (VE), média das variâncias estimadas (MV) e cobertura do parâmetro real pelo intervalo de confiança 90%(C90) nos cenários gerados pela distribuição Gompertz, para ambos os riscos 1 e 2.

			Censura															
			0%				20%				0%				20%			
	Parâmetro	Valor Real	E	VE	MV	C90	E	VE	MV	C90	E	VE	MV	C90	E	VE	MV	C90
Cura G1																		
R1	α_1	0,0	-2,5	22,658	682,138	0,66	-3,1	35,331	495,845	0,58	-2,2	17,724	154,465	0,68	-2,7	24,235	1901,160	0,67
	η_1	-5,3	-5,3	0,037	0,038	0,91	-5,3	0,060	0,060	0,91	-5,3	0,009	0,009	0,90	-5,3	0,013	0,013	0,91
	ρ_1	0,0	0,0	$< 0,001$	$< 0,001$	0,90	0,0	$< 0,001$	$< 0,001$	0,91	0,0	$< 0,001$	$< 0,001$	0,86	0,0	$< 0,001$	$< 0,001$	0,88
	β_1	0,5	0,6	0,054	0,061	0,92	0,6	0,088	0,098	0,91	0,5	0,014	0,015	0,91	0,5	0,019	0,022	0,92
/	α_2	0,0	-0,8	7,064	76,147	0,87	-1,3	13,876	210,025	0,82	-0,1	0,539	1,399	0,93	-0,5	3,837	19,268	0,89
R2	η_2	-4,2	-4,2	0,015	0,015	0,91	-4,2	0,021	0,022	0,91	-4,2	0,004	0,004	0,90	-4,2	0,005	0,005	0,90
	ρ_2	0,0	0,0	$< 0,001$	$< 0,001$	0,86	0,0	$< 0,001$	$< 0,001$	0,87	0,0	$< 0,001$	$< 0,001$	0,91	0,0	$< 0,001$	$< 0,001$	0,88
	β_2	0,8	0,8	0,030	0,030	0,88	0,8	0,039	0,041	0,89	0,8	0,008	0,008	0,90	0,8	0,011	0,011	0,89
Cura G2																		
R1	α_1	0,0	-2,1	18,233	841,148	0,72	-2,8	25,956	1215,326	0,69	-0,8	5,684	40,262	0,86	-1,8	15,456	100,525	0,78
	η_1	-4,6	-4,6	0,021	0,022	0,92	-4,6	0,028	0,029	0,91	-4,6	0,005	0,005	0,91	-4,6	0,007	0,007	0,90
	ρ_1	0,0	0,0	$< 0,001$	$< 0,001$	0,84	0,0	$< 0,001$	$< 0,001$	0,86	0,0	$< 0,001$	$< 0,001$	0,87	0,0	$< 0,001$	$< 0,001$	0,83
	β_1	0,5	0,5	0,032	0,036	0,92	0,5	0,041	0,046	0,91	0,5	0,009	0,009	0,90	0,5	0,012	0,012	0,90
R2	α_2	0,0	-1,2	12,241	29,599	0,82	-2,1	22,104	107,894	0,76	-0,2	1,424	1,250	0,91	-0,4	3,808	50,294	0,89
	η_2	-4,2	-4,2	0,015	0,015	0,90	-4,2	0,019	0,020	0,91	-4,2	0,004	0,004	0,88	-4,2	0,005	0,005	0,90
	ρ_2	0,0	0,0	$< 0,001$	$< 0,001$	0,85	0,0	$< 0,001$	$< 0,001$	0,82	0,0	$< 0,001$	$< 0,001$	0,90	0,0	$< 0,001$	$< 0,001$	0,90
	β_2	0,8	0,8	0,030	0,029	0,88	0,8	0,037	0,036	0,88	0,8	0,009	0,008	0,89	0,8	0,010	0,010	0,90
Cura G3																		
R1	α_1	0,0	-1,9	13,771	130,526	0,86	-2,6	27,350	289,760	0,64	-0,7	5,215	47,405	0,87	-3,4	30,815	353,608	0,34
	η_1	-5,3	-5,3	0,023	0,026	0,90	-5,3	0,039	0,040	0,92	-5,3	0,007	0,007	0,91	-5,3	0,005	0,009	0,98
	ρ_1	0,0	0,0	$< 0,001$	$< 0,001$	0,87	0,0	$< 0,001$	$< 0,001$	0,89	0,0	$< 0,001$	$< 0,001$	0,87	0,0	$< 0,001$	$< 0,001$	0,85
	β_1	0,5	0,5	0,034	0,049	0,94	0,6	0,061	0,072	0,92	0,5	0,011	0,015	0,93	0,6	0,013	0,017	0,98
R2	α_2	0,0	-0,7	6,378	53,471	0,87	-0,8	8,409	38,367	0,86	-0,1	0,349	0,425	0,93	-0,2	0,389	3,436	0,98
	η_2	-4,2	-4,2	0,014	0,014	0,89	-4,2	0,017	0,017	0,89	-4,2	0,004	0,004	0,90	-4,2	0,004	0,004	0,90
	ρ_2	0,0	0,0	$< 0,001$	$< 0,001$	0,86	0,0	$< 0,001$	$< 0,001$	0,88	0,0	$< 0,001$	$< 0,001$	0,90	0,0	$< 0,001$	$< 0,001$	0,89
	β_2	0,8	0,8	0,024	0,028	0,92	0,8	0,032	0,034	0,89	0,8	0,007	0,008	0,90	0,8	0,004	0,009	0,97

1, o modelo subestima os pontos iniciais e superestima os pontos finais tentando acomodar a forma da distribução Gompertz. O viés médio para as 2000 amostras geradas para cada cenário alcançou valores superiores a 0,15, mostrando que o modelo é vulnerável ao erro de especificação, como pode ser visto nas Figuras 4.7 e 4.8 para alguns estratos selecionados. Reforçando que esse é um problema de especificação da distribuição, o viés médio é menor nos cenários que a proporção de cura é maior, pois as curvas simuladas começam a ser mais semelhantes à distribuição Gompertz.

Outro ponto importante é que, sob o erro de especificação, a variância observada para as estimativas é, em muitos cenários, maior que a média das variâncias estimadas. Esse padrão se repete para a maioria dos estratos e ambos os riscos, e pode ser avaliado na Tabela 4.4 para os cenários de cura W1 para o risco 1 e estrato 5. Portanto a sugestão de cálculo da variância é vulnerável ao erro de especificação da distribuição, e essa é uma fragilidade do modelo, pois em contextos práticos não é possível determinar se houve erro de especificação e a distribuição Gompertz não é tão flexível ou conhecida por ser uma distribuição que facilmente acomoda dados em análise de sobrevivência.

Novamente, a variância observada para as estimativas da função de incidência acumulada e as variâncias estimadas aumentam conforme aumenta o valor do tempo, mas sempre com valores que fornecem informação útil para realizar inferências. O erro quadrático médio, no caso do erro de especificação, não se altera com a mudança do número de indivíduos na amostra, pois enquanto a variância observada diminui, mas é muito baixa, o viés independe

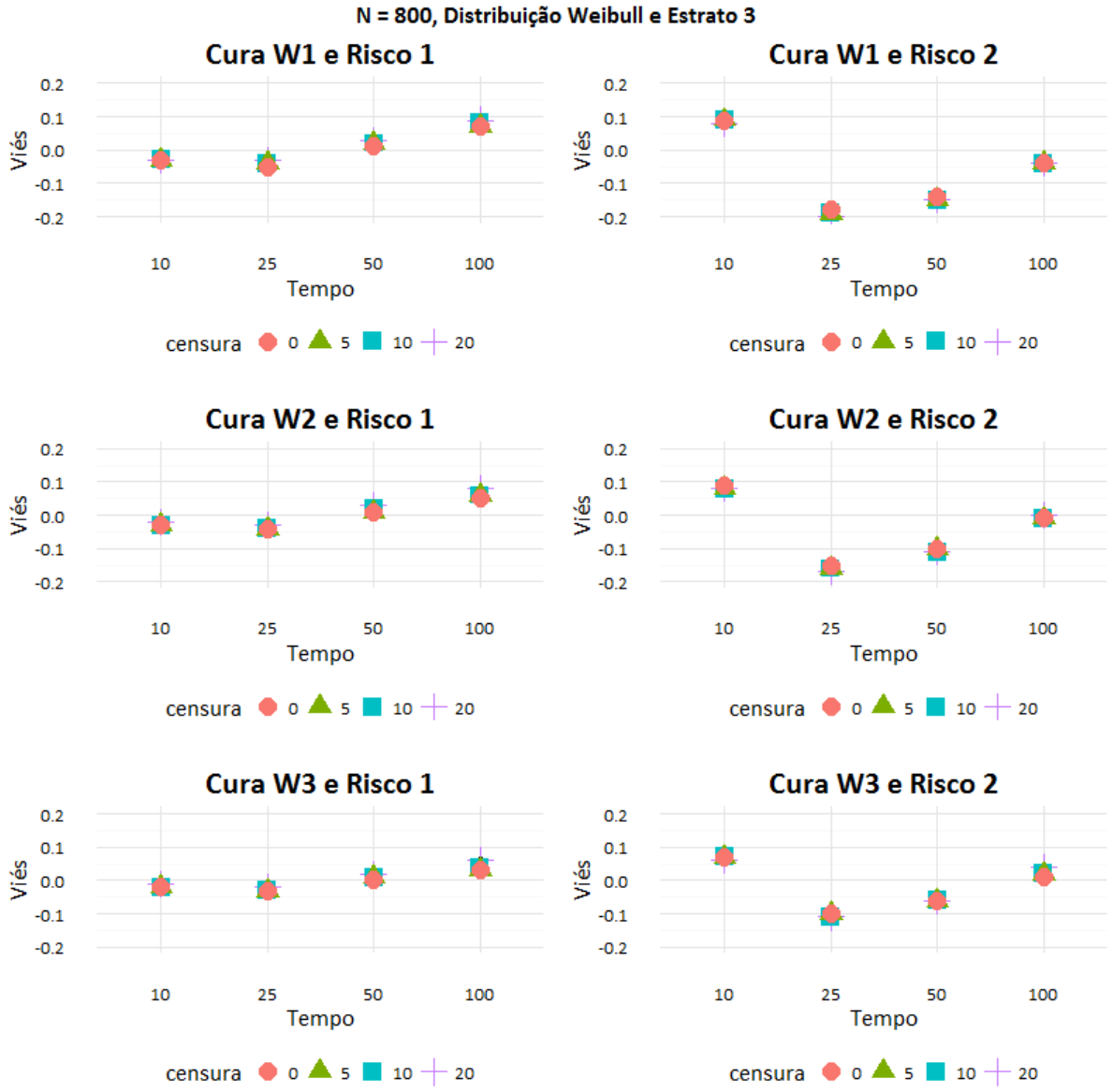


Figura 4.7: Comparação entre o viés médio das estimativas da FIA observado para o estrato 3, riscos 1 e 2 dos cenários gerados pela Weibull com N = 800.

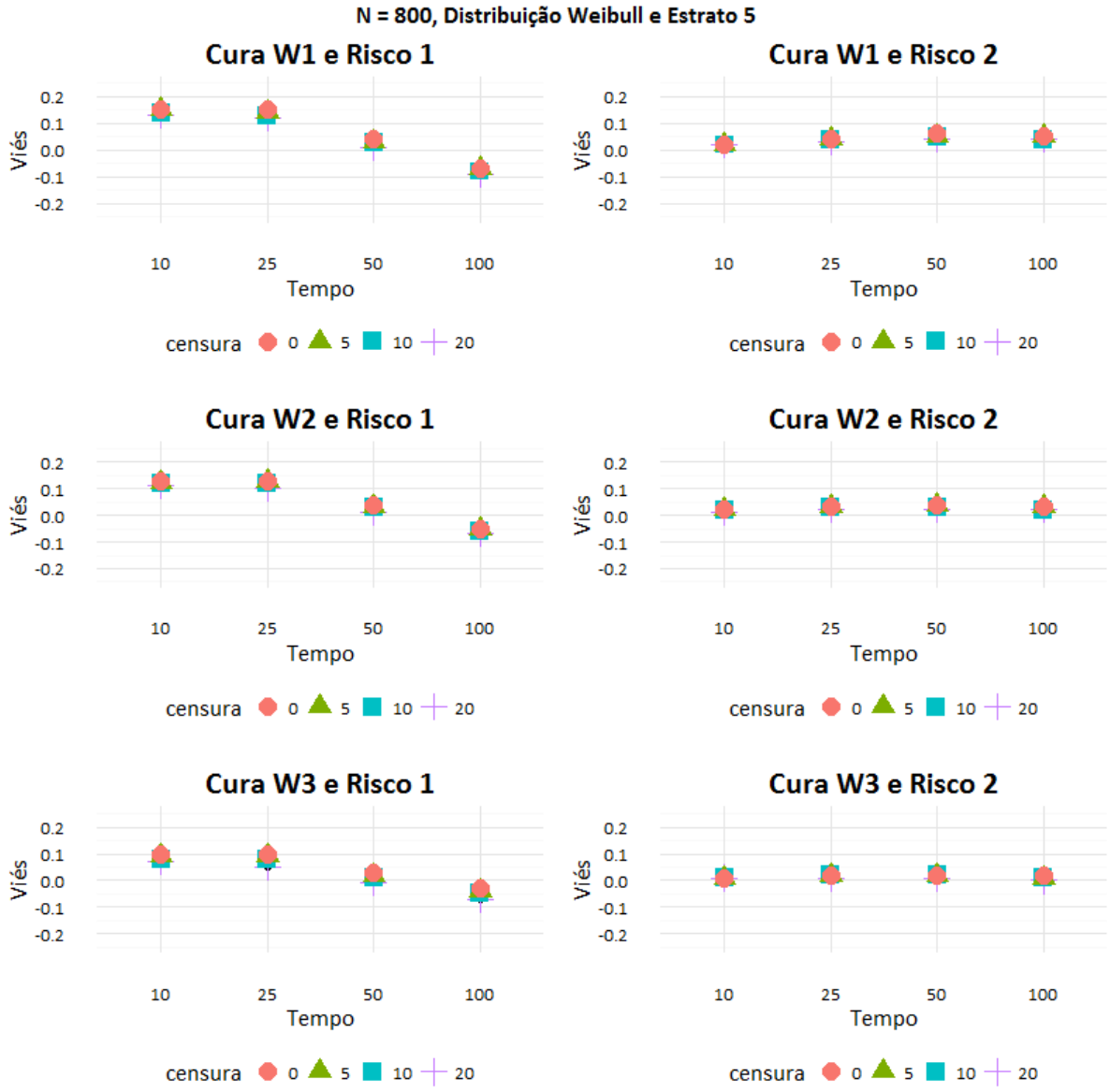


Figura 4.8: Comparação entre o viés médio das estimativas da FIA observado para o estrato 5, riscos 1 e 2 dos cenários gerados pela Weibull com N = 800.

Tabela 4.4: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA no risco 1, estrato 5, nos cenários de cura W1 e distribuição utilizada Weibull.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00059	0,00025	0,00056	0,00031	0,00057	0,00029	0,00077	0,00030
	25	0,00188	0,00080	0,00180	0,00101	0,00184	0,00093	0,00246	0,00094
	50	0,00318	0,00137	0,00307	0,00177	0,00312	0,00161	0,00425	0,00158
	100	0,00400	0,00180	0,00388	0,00237	0,00394	0,00217	0,00564	0,00211
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00017	0,00007	0,00020	0,00007	0,00020	0,00025	0,00019	0,00011
	25	0,00057	0,00022	0,00066	0,00023	0,00066	0,00090	0,00060	0,00035
	50	0,00097	0,00038	0,00113	0,00041	0,00114	0,00171	0,00104	0,00063
	100	0,00120	0,00051	0,00141	0,00054	0,00143	0,00241	0,00134	0,00087

do número de observações, e é ele a parcela que mais contribui para compor o EQM.

4.3 Pseudo-Valores

4.3.1 Função de Ligação Logito

O modelo que se baseia no uso de pseudo-valores foi ajustado com a função de ligação logito e complemento log-log. No primeiro caso, o modelo se mostrou muito eficiente nos cenários simulados pela Gompertz, com viés muito próximo a zero em todos os cenários, independente das proporções de censura ou cura testadas, como está ilustrado para o risco 1 e estrato selecionados nas Figuras 4.9 e 4.10. Para os cenários simulados pela Weibull, o modelo apresentou viés médio baixo, mas não tão próximo de zero como no caso anterior, e nenhum padrão que relaciona as proporções de cura ou censura foi identificado. Em alguns cenários, o viés médio alcançou valores superiores a 0,15 ou inferiores a -0.15 revelando uma imprecisão para esse modelo.

A variabilidade das estimativas cresce, conforme esperado, quando as proporções de cura ou censura aumentam e se o número de indivíduos é menor. No entanto, em todos os cenários simulados, a variabilidade foi muito pequena. Como no trabalho de Klein e Andersen (2005) não foi proposto nenhuma forma de estimar a variância das estimativas das funções de incidência acumulada, uma alternativa testada nesse exercício de simulação foi calcular a estimativa conforme proposto nos trabalhos paramétricos aqui apresentados e comparar as estimativas obtidas com a variabilidade observada. Nas Tabelas 4.5 e 4.6, encontra-se essa comparação para alguns cenários selecionados. Note que os valores não são tão parecidos, porém a média das estimativas para a variância sempre é maior que a variância observada, o que poderia indicar que essa é um forma de estimar a variância que pelo menos não oferece estimativas subestimadas.

Tabela 4.5: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA no risco 1, estrato 1, nos cenários de cura G1 e distribuição utilizada Gompertz.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00006	0,00006	0,00007	0,00007	0,00007	0,00007	0,00008	0,00008
	25	0,00015	0,00017	0,00017	0,00018	0,00016	0,00020	0,00022	0,00024
	50	0,00025	0,00030	0,00028	0,00032	0,00027	0,00035	0,00039	0,00046
	100	0,00032	0,00040	0,00036	0,00043	0,00035	0,00049	0,00055	0,00071
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002
	25	0,00004	0,00005	0,00004	0,00005	0,00005	0,00005	0,00006	0,00007
	50	0,00007	0,00008	0,00007	0,00009	0,00008	0,00009	0,00010	0,00012
	100	0,00009	0,00011	0,00009	0,00012	0,00011	0,00013	0,00015	0,00019

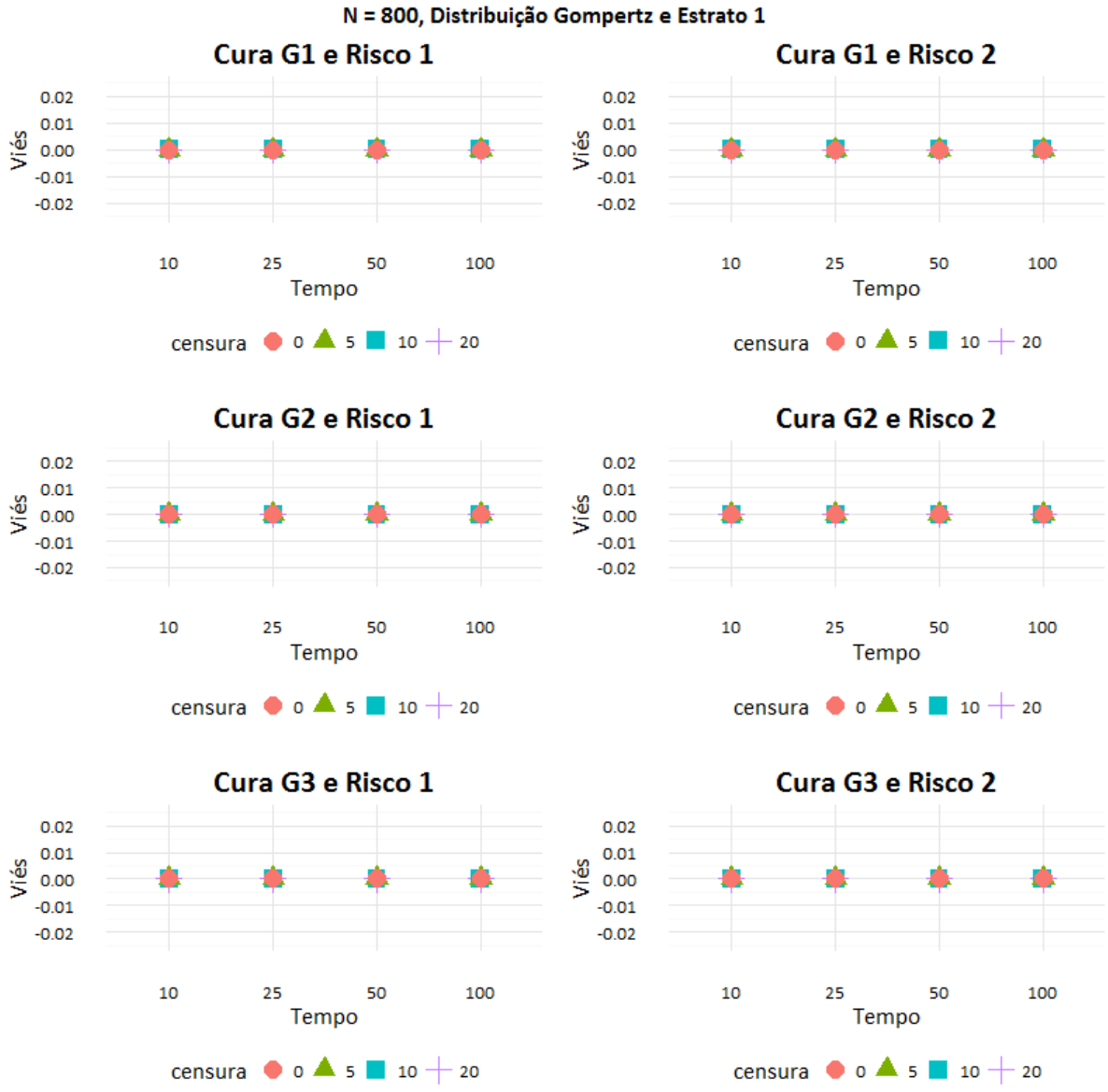


Figura 4.9: Comparação entre o viés médio das estimativas da FIA observado para o estrato 1, riscos 1 e 2 dos cenários gerados pela Gompertz com N = 800.

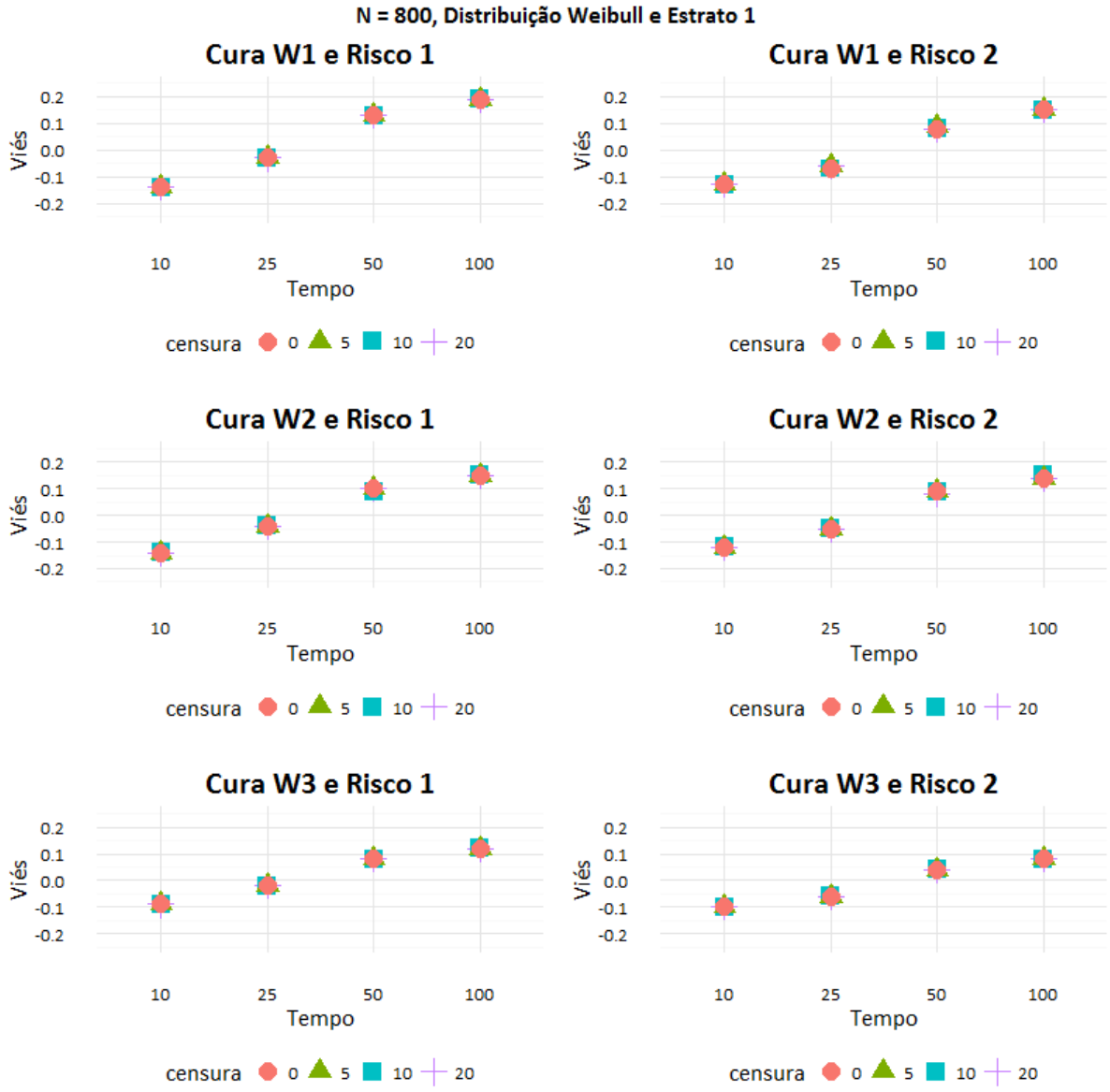


Figura 4.10: Comparação entre o viés médio das estimativas da FIA observado para o estrato 1, riscos 1 e 2 dos cenários gerados pela Weibull com N = 800.

Tabela 4.6: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA, no risco 1, estrato 1 nos cenários de cura W1 e distribuição utilizada Weibull.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00066	0,00059	0,00074	0,00060	0,00073	0,00063	0,00075	0,00067
	25	0,00189	0,00266	0,00216	0,00273	0,00208	0,00285	0,00228	0,00308
	50	0,00224	0,00407	0,00253	0,00418	0,00242	0,00434	0,00267	0,00469
	100	0,00238	0,00427	0,00267	0,00439	0,00256	0,00454	0,00283	0,00496
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00017	0,00015	0,00019	0,00016	0,00018	0,00016	0,00020	0,00017
	25	0,00049	0,00072	0,00054	0,00074	0,00051	0,00076	0,00061	0,00082
	50	0,00056	0,00111	0,00064	0,00115	0,00061	0,00118	0,00069	0,00128
	100	0,00060	0,00118	0,00069	0,00121	0,00066	0,00125	0,00075	0,00137

4.3.2 Função de ligação Complemento Log-Log

As conclusões para a simulação do modelo que se baseia no uso dos pseudo-valores com a função de ligação complemento log-log são bem semelhantes às apresentadas na seção anterior, em que o modelo se mostrou eficiente nos cenários simulados pela Gompertz, com viés muito próximo a zero em todos os cenários, independente das proporções de censura ou cura testadas, como está ilustrado para o risco 1 e estrato selecionados nas Figuras 4.11 e 4.12. Novamente, apesar de um viés médio baixo, nos cenários simulados pela Weibull, não foi possível identificar um padrão que relaciona as proporções de cura ou censura com o viés médio. Em alguns cenários, o viés médio alcançou valores superiores a 0,15 ou inferiores a -0.15 revelando uma imprecisão para esse modelo.

A variabilidade também se comporta da mesma maneira, crescendo quando as proporções de cura ou censura aumentam e se o número de indivíduos é menor, mas sempre apresentando valores baixos. Note que os valores não são tão parecidos, mas desta vez, a média das variância estimadas não é sempre maior que a variância observada para as estimativas da função de incidência acumulada.

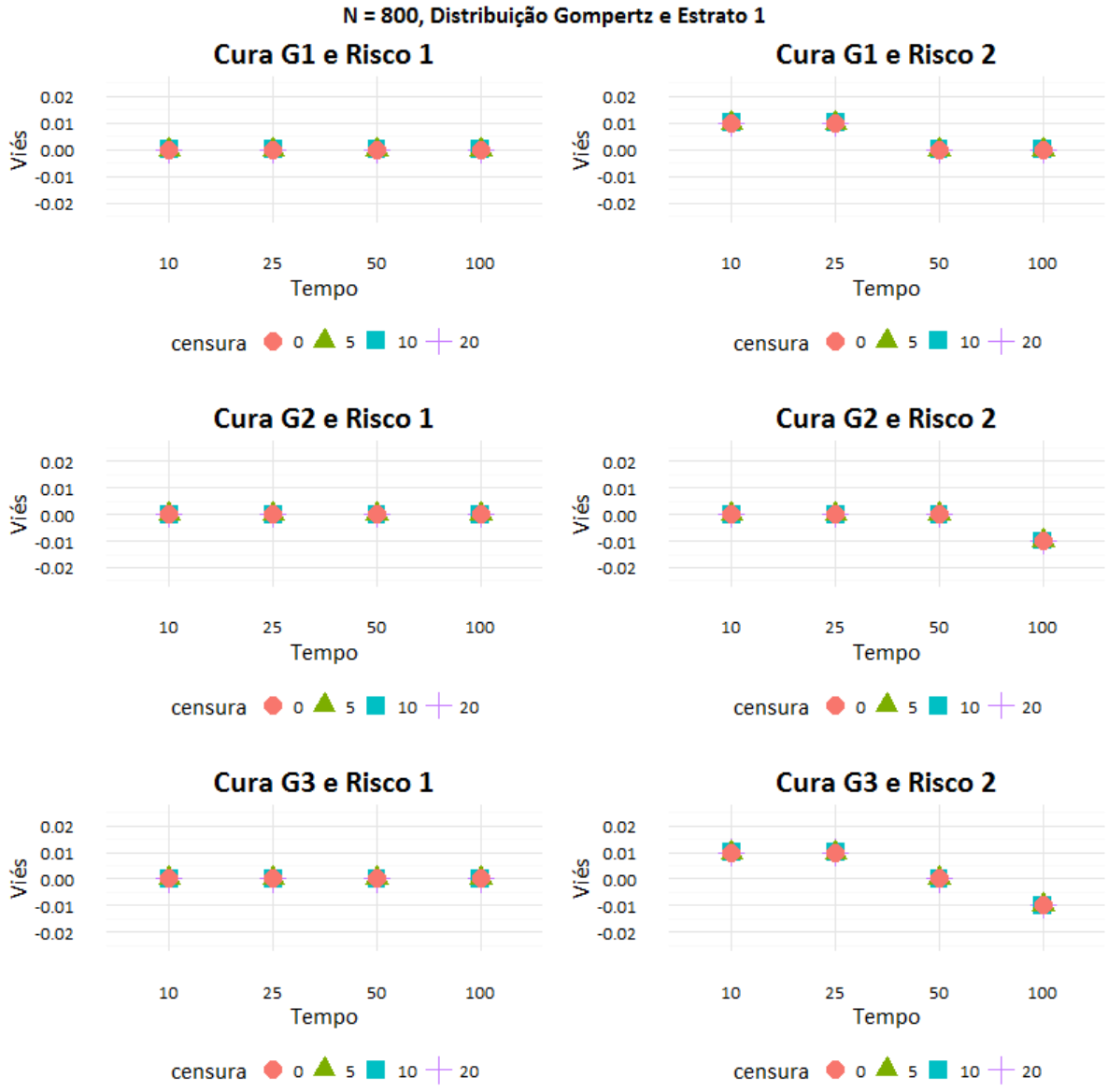


Figura 4.11: Comparação entre o viés médio das estimativas da FIA observado para o estrato 1, riscos 1 e 2 dos cenários gerados pela Gompertz com N = 800.

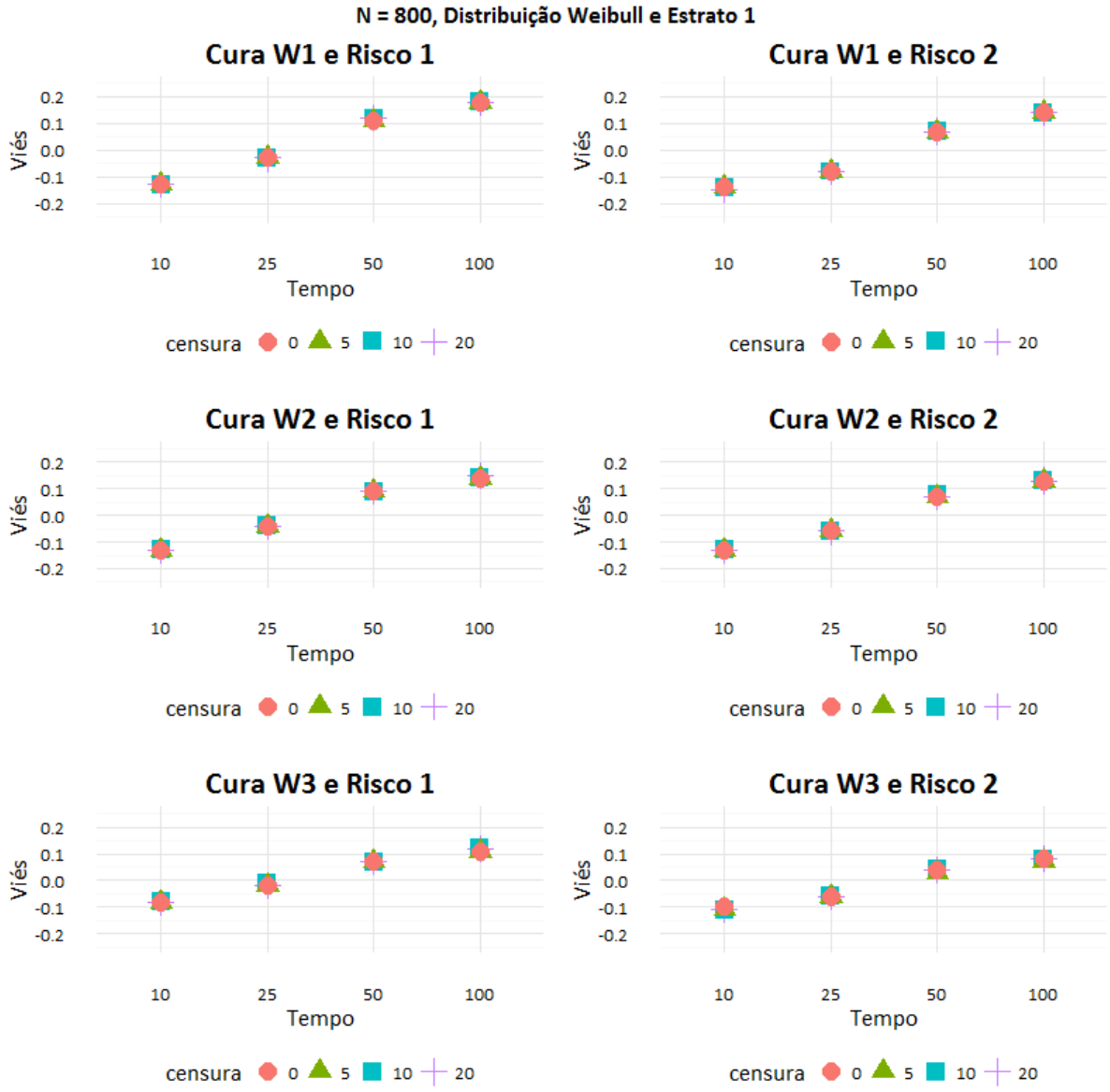


Figura 4.12: Comparação entre o viés médio das estimativas da FIA observado para o estrato 1, riscos 1 e 2 dos cenários gerados pela Weibull com N = 800.

Tabela 4.7: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA no risco 1, estrato 1, nos cenários de cura G1 e distribuição utilizada Gompertz.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00006	0,00006	0,00007	0,00006	0,00006	0,00006	0,00008	0,00008
	25	0,00015	0,00014	0,00016	0,00015	0,00016	0,00016	0,00021	0,00020
	50	0,00025	0,00023	0,00028	0,00025	0,00027	0,00028	0,00039	0,00036
	100	0,00033	0,00031	0,00036	0,00033	0,00036	0,00037	0,00056	0,00054
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002	0,00002
	25	0,00004	0,00004	0,00004	0,00004	0,00005	0,00004	0,00005	0,00006
	50	0,00007	0,00006	0,00007	0,00007	0,00008	0,00008	0,00010	0,00010
	100	0,00009	0,00008	0,00009	0,00009	0,00011	0,00001	0,00015	0,00015

Tabela 4.8: Comparação entre média das variâncias estimadas (MV) e variância das estimativas (VE) para a FIA, no risco 1, estrato 1, nos cenários de cura W1 e distribuição utilizada Weibull.

		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
800	10	0,00049	0,00039	0,00054	0,00040	0,00054	0,00041	0,00055	0,00044
	25	0,00144	0,00126	0,00164	0,00130	0,00159	0,00135	0,00175	0,00146
	50	0,00201	0,00192	0,00227	0,00198	0,00218	0,00206	0,00241	0,00224
	100	0,00227	0,00222	0,00254	0,00229	0,00245	0,00237	0,00270	0,00259
		Censura							
		0%		5%		10%		20%	
N	Tempo	VE	MV	VE	MV	VE	MV	VE	MV
3000	10	0,00012	0,00010	0,00014	0,00011	0,00013	0,00011	0,00015	0,00012
	25	0,00037	0,00034	0,00041	0,00035	0,00039	0,00036	0,00046	0,00039
	50	0,00050	0,00052	0,00058	0,00054	0,00055	0,00055	0,00062	0,00060
	100	0,00057	0,00060	0,00065	0,00062	0,00063	0,00064	0,00071	0,00070

4.4 Comparação entre Modelos

A ideia inicial desse exercício de simulação era justamente comparar o método paramétrico na situação em que esse deveria ser o melhor modelo, isto é, quando os dados fossem gerados pela Gompertz, e também ser capaz de compará-lo à alternativa testada sob o erro de especificação da distribuição. Após a geração dos dados e leitura dos resultados dos ajustes, fica claro que é importante separar a leitura em três situações:

1. Dados gerados pela Gompertz;
2. Dados em que as funções de incidência acumulada assumem formas bem diferentes das possíveis formas adotadas pela Gompertz;
3. Dados em que, apesar de não serem gerados pela distribuição Gompertz, assumem formas semelhantes.

Para realizar comparações justas, ao longo dessa seção foram comparados os modelos que permitem a inclusão de covariáveis, e o leitor deve adotar a notação "J7"para o modelo proposto em Jeong e Fine (2007), "PC"e "PL"para o modelo descrito em Klein e Andersen (2005), proposto aqui para o uso em dados com fração de cura, com o uso das funções de ligação complemento log-log e logito, respectivamente.

Assim, para os **dados gerados pela Gompertz**, ao contrário do que era esperado, o viés e o erro quadrático médio (EQM) para ambos os modelos são muito semelhantes, e não existe uma preferência clara por um modelo. Para exemplificar, os resultados para o estrato 2 estão apresentados nas Figuras 4.13 e 4.14. Não foi observado um padrão claro que relacionasse proporção de censura ou de cura impactando viés ou EQM, mas isso pode ser consequência da escolha dos tamanhos amostrais testados, que visavam valores semelhantes aos encontrados nos dados da aplicação. O único padrão claro é que conforme o valor do tempo aumenta, a variância das estimativas aumenta, mas sempre adotando um valor baixo mostrando a precisão dos modelos com o número médio de observações por estrato de 400 indivíduos.

Já na situação em que os dados **assumem formas para a FIA bem distintas daquelas geradas pela Gompertz**, o modelo paramétrico nem sempre apresenta resultados tão bons quanto o modelo que se baseia nos pseudo-valores. De forma geral, para os cenários W2 e W3, que apresentam proporções de cura superiores a 20%, as diferenças entre os modelos observadas foram amenizadas, uma vez que conforme a proporção de cura aumenta, as FIAs atingem valores mais baixos, e consequentemente, os ajustes apresentam menor viés e valores mais baixos para o EQM. Vale reforçar então que, para fins de simplicação nesse texto somente, os resultados serão apresentados para os cenários W1, mesmo que resultados semelhantes, mas em escala menor, foram observados para os cenários com proporção de cura maior. Outra escolha para simplificação foi apresentar somente os gráficos comparando os cenários com número de observações igual a 800, já que neles as diferenças são acentuadas.

Alguns exemplos foram selecionados para ilustrar as observações seguindo o raciocínio descrito acima, e para cada um deles, além de apresentados os EQMs observados, foram

Figura 4.13: Comparação entre o viés médio das estimativas da FIA observado para o estrato 2, risco 1 nos cenários gerados pela Gompertz com N = 800.

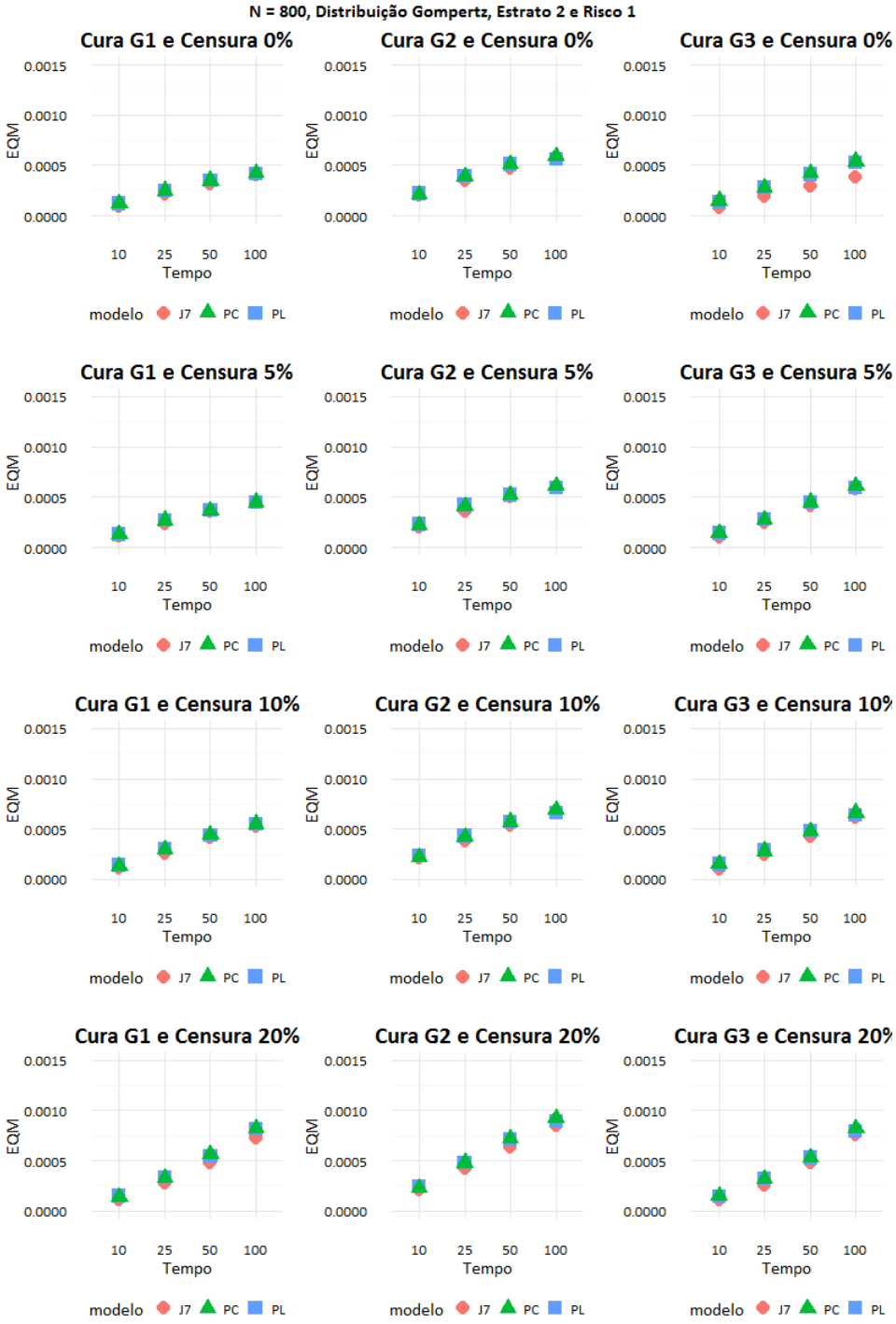


Figura 4.14: Comparação entre os valores de EQM para FIA no estrato 2, risco 1 nos cenários gerados pela Gompertz com N = 800.

selecionados aleatoriamente um conjunto de dados de 800 observações para ilustrar as FIAs estimadas por cada um dos métodos comparadas às curvas teóricas. Na Figura 4.15, é possível notar que para $t = 25$ o EQM é muito superior para o modelo J7, mas que se comparados os tempos finais, esse modelo na realidade apresenta melhores valores de EQM. Na comparação entre as curvas estimadas e a curva teórica em todos os pontos, é possível ver que o modelo J7 subestima a curva real praticamente para todo o período do estudo, mas nos maiores tempos é o que mais se aproxima da curva real. No caso de ambos os ajustes baseados nos pseudo-valores, as estimativas se aproximam mais da curva teórica ao longo de todo o estudo, porém o modelo superestima os tempos maiores e não consegue capturar a incidência acentuada nos tempos iniciais muito bem.

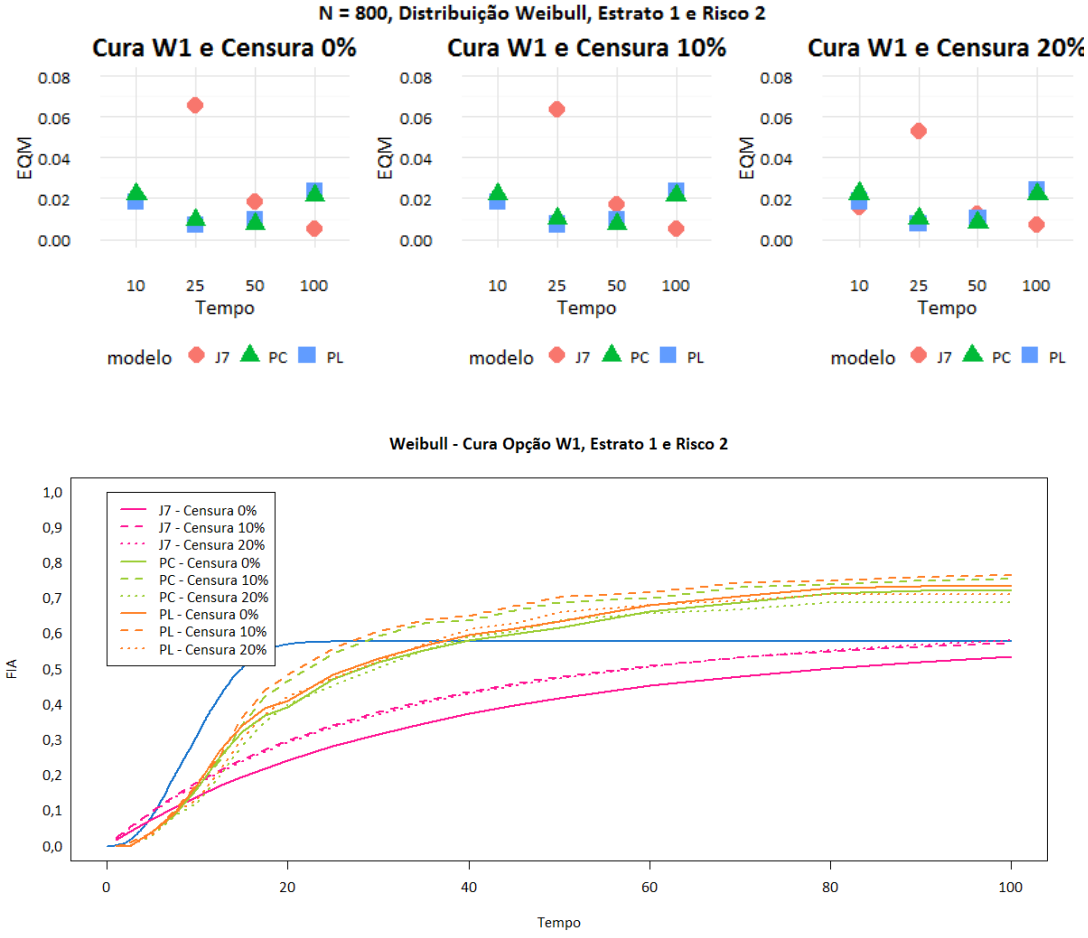


Figura 4.15: Comparação entre os valores de EQM para a FIA no estrato 1, risco 2 nos cenários gerados pela Gompertz com N = 800.

Outro exemplo é o risco 2 no estrato 3, apresentado na Figura 4.16, em que foi observado um comportamento bem semelhante ao descrito acima, mas nesse caso, o modelo baseado nos pseudo-valores se ajusta bem melhor nos tempos iniciais, conseguindo capturar a incidência acentuada inicial, mas ainda superestimando os tempos maiores. Por outro lado, o modelo J7 que estima muito bem a FIA nos tempos maiores, está subestimando a curva por quase todo o

período considerado e portanto, pelo EQM, a preferência seria pelos modelos PC ou PL.

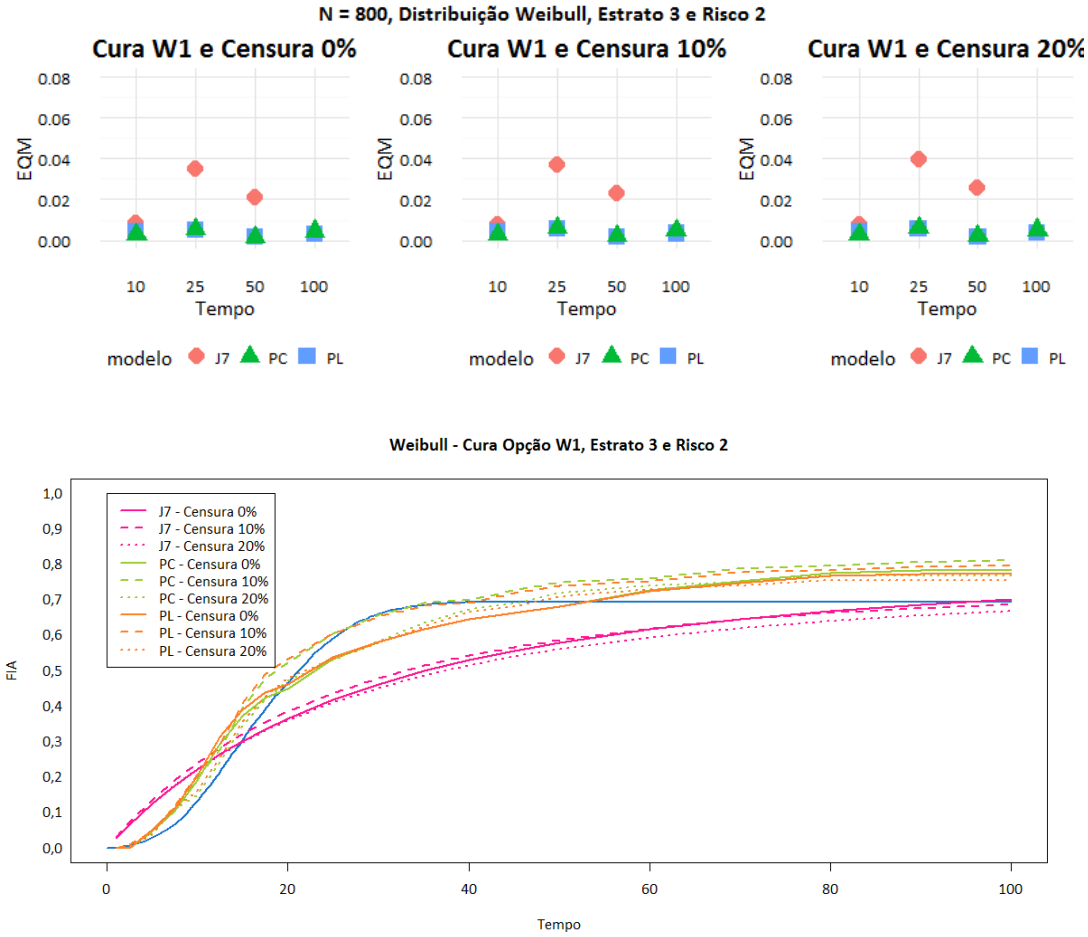


Figura 4.16: Comparação entre os valores de EQM para a FIA no estrato 3, risco 2 nos cenários gerados pela Gompertz com N = 800.

Claramente, outras formas para as funções de incidência acumuladas podem ser testadas, mas nesse exercício limitou-se a comparar os modelos sobre curvas que se assemelham às curvas dos dados da aplicação. Algumas curvas teóricas porém, **apesar de serem geradas pela Weibull, não apresentam formas muito distintas da Gompertz** e os resultados de todos os modelos para elas é muito semelhante, não ficando clara uma preferência por um modelo, como pode ser visto nas Figuras 4.17 e 4.18.

O risco 1 no estrato 1 apresenta resultados interessantes, pois diferente do esperado, o modelo parámetrico se mostra como a melhor alternativa mesmo que os dados sejam gerados pela Weibull, isso porque os ajustes baseados nos pseudo-valores superestimam a curva teórica para quase todo o período, enquanto o modelo J7, apesar de subestimar no início do período de observação, se ajusta melhor conforme o tempo vai aumentando. Na comparação das curvas, essa leitura fica clara e, para casos como esse, cabe a discussão do que é prioridade para o pesquisador a fim de escolher o modelo que melhor se ajustará aos dados de interesse.

De forma geral, o exercício de simulação permitiu a avaliação de que não era tão supe-

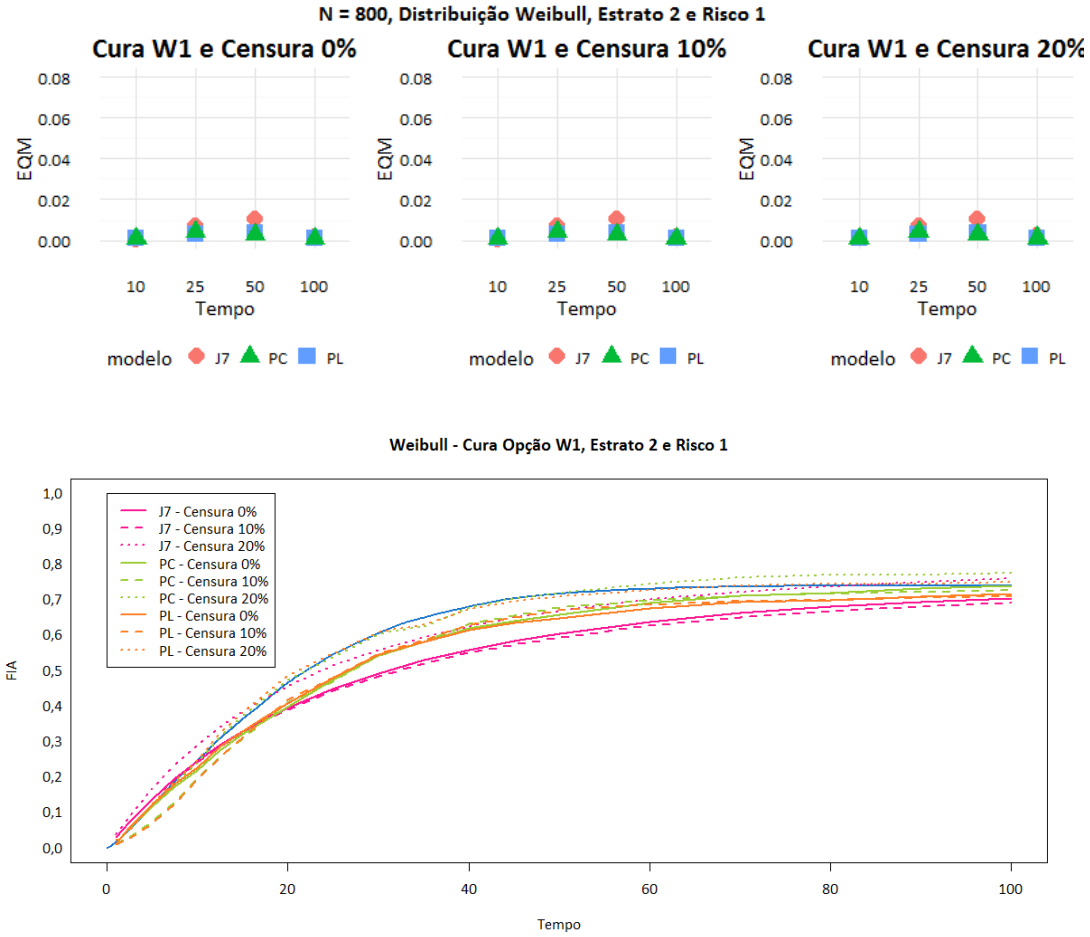


Figura 4.17: Comparação entre os valores de EQM para a FIA no estrato 2, risco 1 nos cenários gerados pela Gompertz com N = 800.

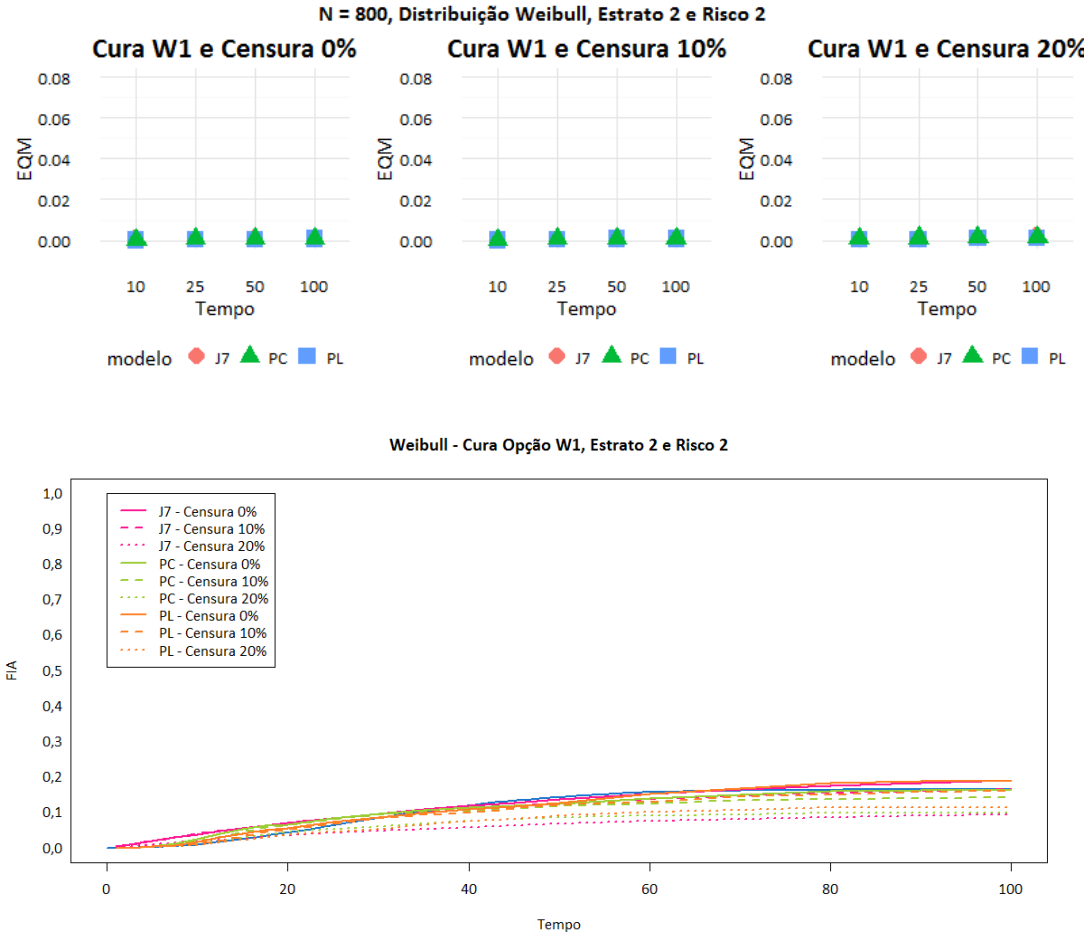


Figura 4.18: Comparação entre os valores de EQM para a FIA no estrato 2, risco 2 nos cenários gerados pela Gompertz com N = 800.

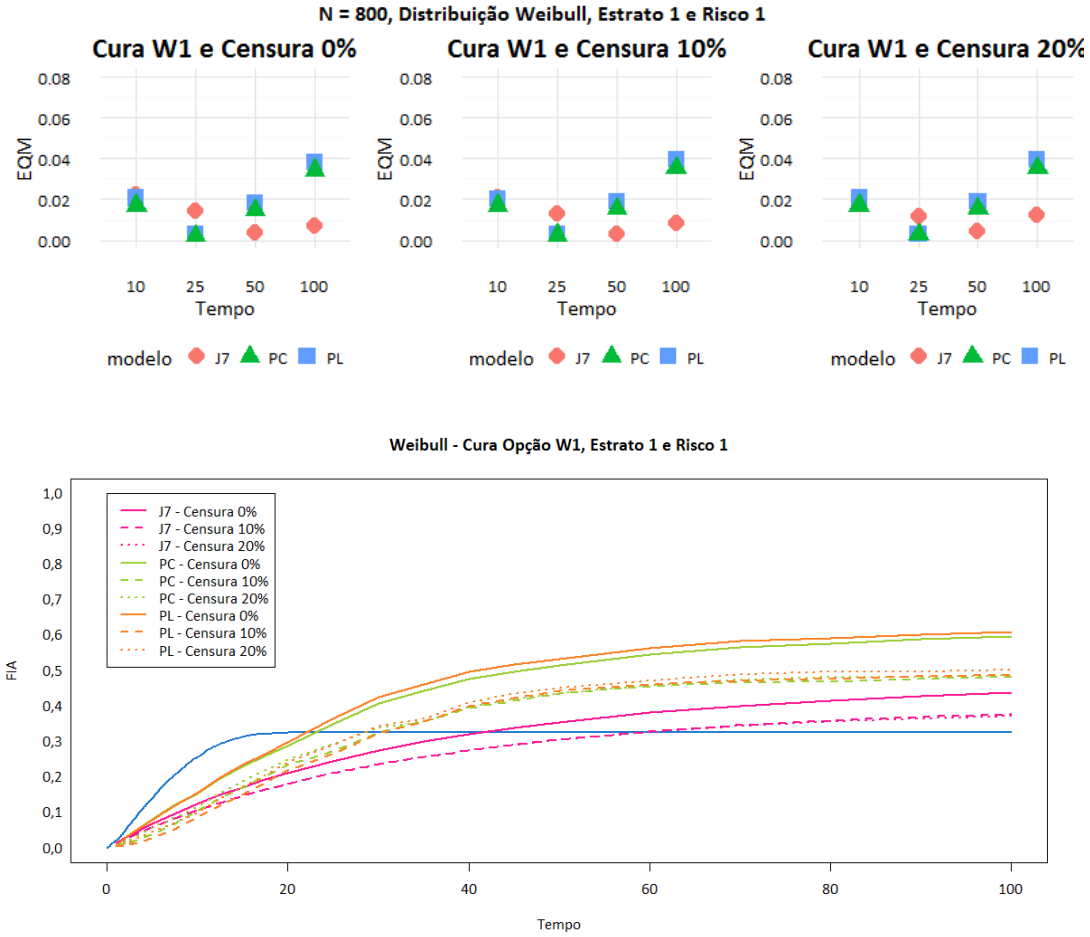


Figura 4.19: Comparação entre os valores de EQM para a FIA no estrato 1, risco 1 nos cenários gerados pela Gompertz com N = 800.

rior a preferência esperada pelos modelos paramétricos na situação de dados gerados pela distribuição em que se baseiam, mas que também esse modelo pode se ajustar bem a dados em que, apesar de alguma diferença, as funções de incidência acumulada se assemelham - não necessariamente em todos os pontos - às possiveis formas geradas pela Gompertz. O modelo que se baseia no uso de pseudo-valores é uma alternativa boa, e que se ajusta bem a dados com riscos competitivos e fração de cura, independente da forma da curva. Um passo importante é, portanto, realizar uma análise descritiva dos dados para conhecer as formas das funções de incidência acumulada que se deseja estimar, e a partir do que for observado testar os modelos que apresentam melhores resultados.

Capítulo 5

Aplicação a Dados de Cartão de Crédito

O estudo que motivou explorar métodos que combinassem a presença de riscos competitivos e fração de cura está presente em análises de crédito, em que é necessário prever o nível de perdas para então precificar o produto, delinear planos de recompensa e definir elegibilidade e concessão de linhas crédito adequadas para cada possível cliente. Ao estimar as perdas financeiras, é comum trabalhar sobre a previsão da proporção de contratos cancelados *involuntariamente*, isto é, contratos cancelados pela instituição financeira após finalizado um prazo determinado para que cliente retome o pagamento de sua dívida depois de ter atrasado algum pagamento. Entretanto, essa abordagem ignora que o nível de cancelamento *voluntário* (realizado pelo cliente por desinteresse no produto) pode afetar muito as estimativas buscadas. Além disso, quando se trata de um produto como cartão de crédito, que não tem uma data final pré-estabelecida, além das duas situações descritas acima (cancelamento voluntário e cancelamento involuntário), existe a possibilidade de nenhuma das duas acontecerem, gerando a possível presença de sobreviventes de longo prazo. Nesse caso, a fração de cura é melhor entendida como sobrevivência de longo prazo, uma vez que todos os contratos de cartões terão uma data final, por exemplo, com o falecimento de seu titular.

O conjunto de dados utilizado para ilustrar a aplicação das abordagens descritas no capítulo anterior é composto por 13114 contratos de cartão de crédito iniciados em agosto de 2011. O objetivo principal é ser capaz de prever qual a porcentagem de contratos que estará ativa após um longo prazo, caso ela exista. Como objetivos secundários, é de interesse identificar características disponíveis no momento da contratação do cartão de crédito relacionadas a diferentes níveis de perda (cancelamento involuntário) e fornecer informações importantes sobre o tempo de seguimento suficiente para calcular apropriadamente lucros e prejuízos. Seguindo esse raciocínio, foram coletadas as possíveis covariáveis disponíveis no momento da contratação (informações presentes na proposta) e todos os contratos foram acompanhados pelos 56 meses seguintes para serem classificados entre cancelamento involuntário, cancelamento voluntário ou censura. Como não há perda de informação ao longo dos 56 meses, os dados apresentam censura do tipo I, isto é, todos os indivíduos que não apresentaram falha (cancelamento) até o final do seguimento do estudo são considerados censurados. Parte dessa

população censurada é na realidade a porção de curados (cartões que não serão cancelados por nenhum dos dois tipos descritos), uma vez que existe a possibilidade de nenhum desses tipos de cancelamento ocorrer. A seguir, todas as covariáveis disponíveis estão listadas com suas descrições e codificações entre parênteses:

- **Idade:** faixa de idade do aplicante no momento da proposta, tomando valores (1) "menor que 20 anos", (2) "de 20 a menor que 30 anos", (3) "de 30 a menor que 50 anos", (4) "de 50 a menor que 70 anos" ou (5) "maior ou igual a 70 anos";

- **Faixa de Escore de Crédito:** faixa do escore de crédito utilizado pela instituição financeira para estimar o risco do aplicante, tomando valores (1) "Sem escore", (2) "Alto Risco", (3) "Médio Risco", (4) "Baixo Risco" ou (5) "Baixíssimo Risco";

- **Gênero:** gênero do aplicante no momento da proposta, tomando valores ('F') "Feminino" ou ('M') "Masculino";

- **Canal:** canal de aquisição do cartão de crédito, tomando valores (1) "Tele-Marketing Ativo", em que a instituição financeira aborda o possível aplicante via telefone, (2) "Agência de Varejo", em que o aplicante requer o cartão por meio da agência, (3) "Mala Direta", em que instituição financeira aborda o possível aplicante via carta ou (4) "Internet / Tele-Marketing Passivo", em que o aplicante requer o cartão via internet ou telefone;

- **Produto:** classificação do tipo de cartão requerido na proposta pelo aplicante, assumindo valores (1) "Black", (2) "Gold", (3) "Internacional" ou (4) "Platinum". As quatro diferentes categorias oferecem características variadas e a decisão fica a critério do aplicante uma vez que tenha renda mínima estabelecida para cada produto;

- **Região:** estado do país onde a proposta foi realizada, tomando valores (1) "SP", (2) "RJ", (3) "PR, RS ou SC", (4) "MG, BA, DF, GO, ES, MT ou MS" ou (5) "Demais estados".

5.1 Análise Descritiva

Dentre todos os 13114 contratos, 4160 cartões foram cancelados pelos clientes, 1962 cancelados pela instituição financeira e 6992 permanecem ativos após os 56 meses, e então são considerados censurados. Os gráficos na Figura 5.1 revelam o número e a proporção de cancelamentos voluntários, involuntários e cartões ativos ao final dos 56 meses de observação por categoria de cada covariável. Através deles, também é possível observar a distribuição das categorias de cada covariável e que a proporção de cada tipo de cancelamento ou cartões ativos é bem diferente entre as categorias para as covariáveis "Faixa de Escore de Crédito", "Canal" e "Produto".

Avaliando as funções de incidência acumulada na Figura 5.2 observa-se que ambas as curvas parecem formar um platô a partir do mês 50, indiciando a presença de cura nesses dados. Note que as curvas são apresentadas sobrepostas e como partição do complementar

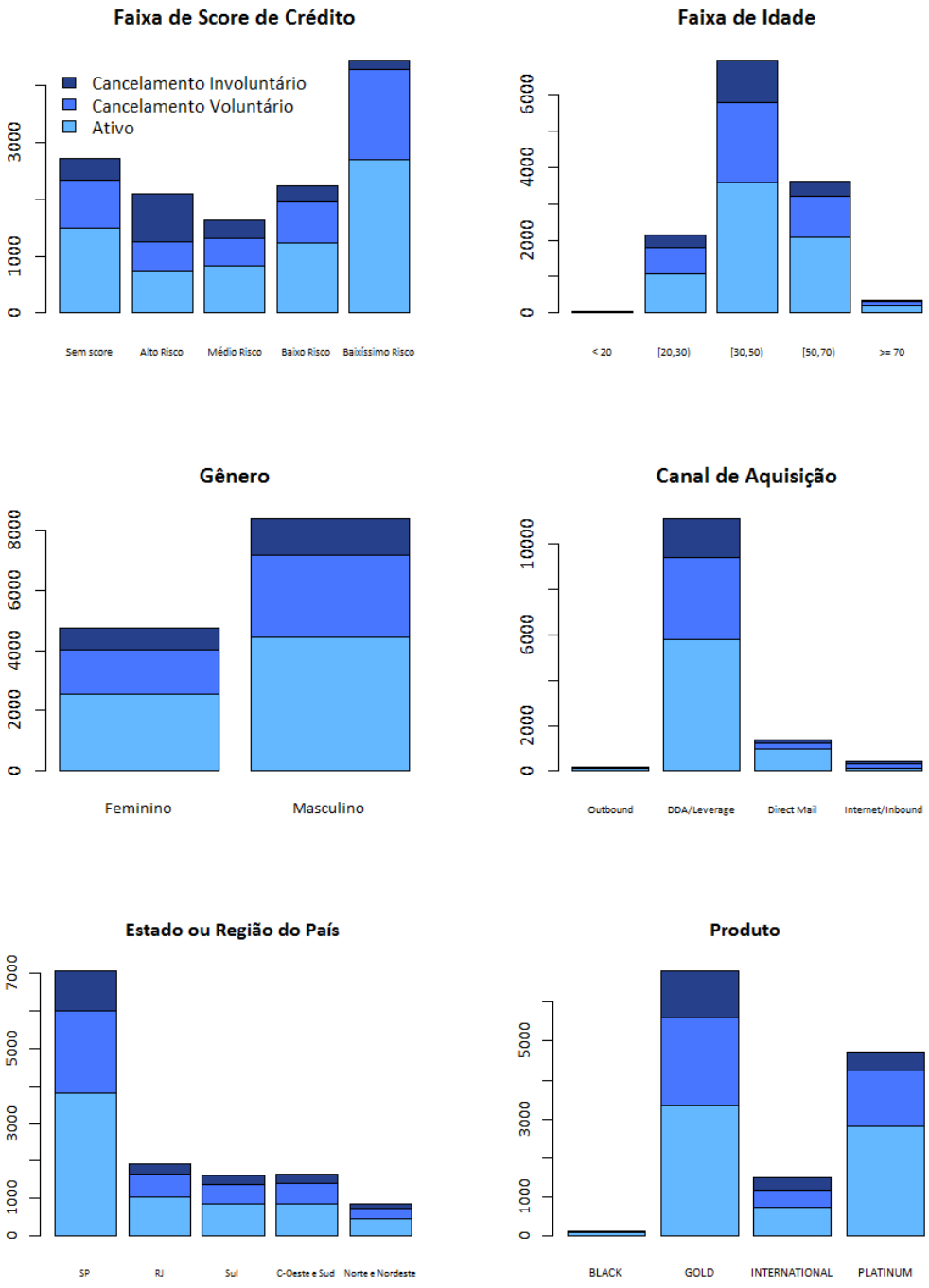


Figura 5.1: Proporção de Tipos de Cancelamento por Categorias das Covariáveis: Escore de Crédito, Idade, Gênero e Canal de Aquisição

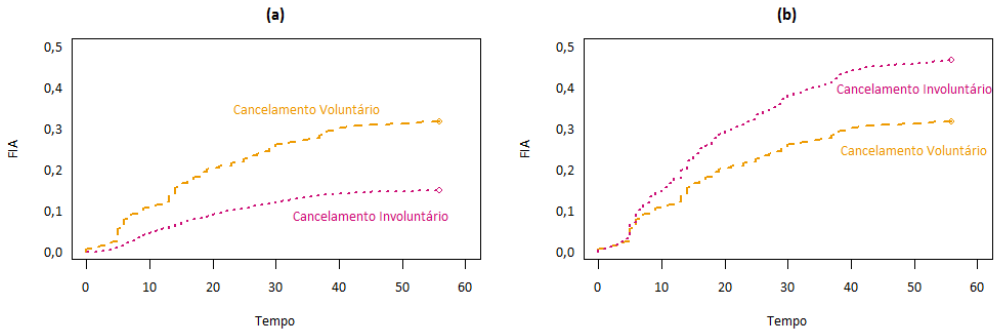


Figura 5.2: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário: em (a) as curvas são apresentadas sobrepostas e em (b) estão representadas pela área (acumuladas)

da curva de sobrevivência, assim como em 2.2. Em qualquer apresentação, fica claro que no quinquagésimo sexto mês, ainda existe uma porção considerável de contratos que não foram cancelados. Por fim, o aspecto importante é a avaliação das curvas de incidência acumulada por categoria de cada covariável.

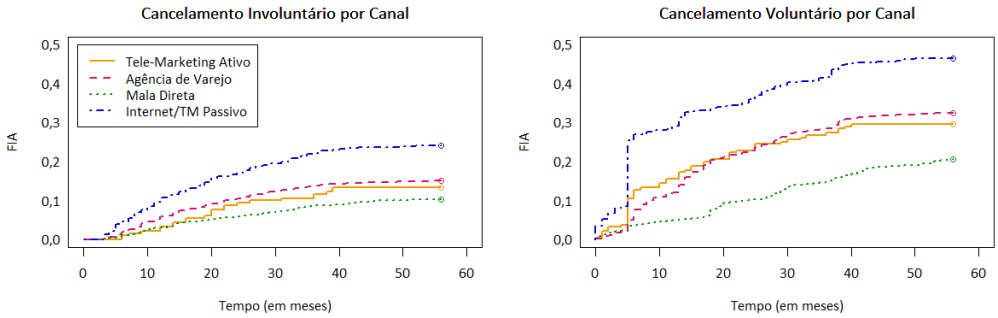


Figura 5.3: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Canal de Aquisição

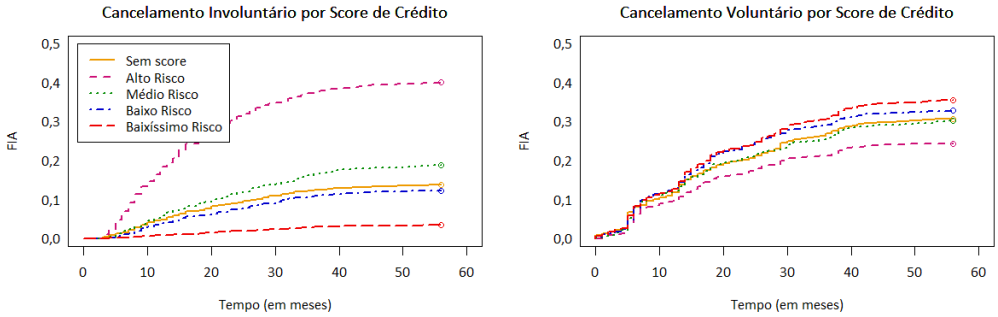


Figura 5.4: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Faixa de Escore de Crédito

Conforme esperado, o canal "Internet/Tele-Marketing Passivo" é o que apresenta maiores

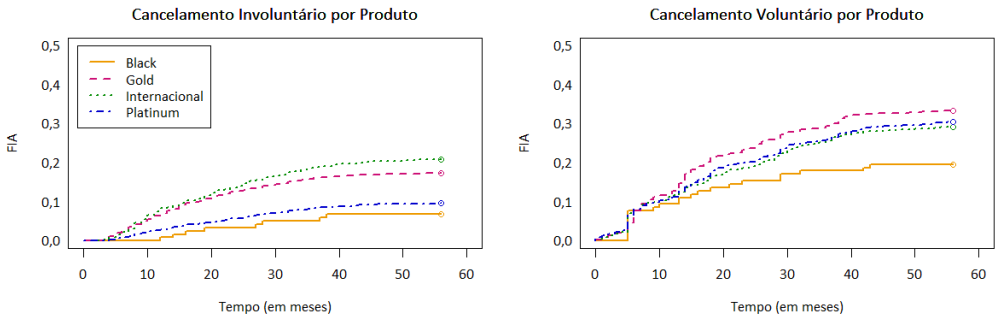


Figura 5.5: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Tipo de Produto

valores para a função de incidência acumulada pois acredita-se existir uma seleção adversa dado que por esse canal os aplicantes que buscaram o crédito. Também o mesmo canal apresenta a maior probabilidade de cancelamento voluntário. Note que no sexto mês o salto nas curvas de cancelamento voluntário é reflexo do procedimento automático do banco que considera cancelado um contrato de cartão de crédito se, depois de seis meses do momento da aquisição, o cartão ainda estiver bloqueado para seu uso.

As faixas do "Escore de Crédito", que é uma ferramenta criada justamente com o propósito de estimar o risco, conforme esperado apresentam curvas de incidência acumuladas para o cancelamento involuntário bem distintas, em especial entre os extremos "Alto Risco" e "Baixíssimo Risco". No caso das curvas de cancelamento voluntário, o interessante é notar que a ordenação de maior para menor incidência é reversa quando comparada ao cancelamento involuntário, o que é justificado pelo raciocínio de que os clientes de menor risco (com menor probabilidade de contratos cancelados por atraso) são também os clientes que menos precisam do crédito e por sua vez, cancelam seus contratos por desinteresse no produto. Quando avaliadas as curvas para os diferentes tipos de produto, não há clara distinção nas curvas do cancelamento voluntário, com excessão da categoria "Black" que tem volume muito baixo (são 117 casos de "Black" sem considerar o tipo do cancelamento) mas existe uma ordenação esperada entre os produtos para as curvas de cancelamento involuntário, pois as restrições de renda para a aquisição de cada produto são diferentes. Quanto mais restrito, menor é o risco observado nessas curvas.

Apesar de usualmente apresentarem relação muito fraca com os tipos cancelamento, também foram avaliadas as covariáveis demográficas "Gênero", "Região" e "Idade" por estarem disponíveis, mas que não apresentam claras diferenças nas funções de incidência acumulada estimadas para ambos os tipos de cancelamento. No caso da covariável faixa de idade, não há clara diferença entre as curvas de ambos os tipos de cancelamento com excessão da faixa de idade "menor que 20 anos", que pode estar descolando das demais por essa faixa ser constituída apenas por 50 casos.

Nas seções a seguir, os métodos apresentados nos capítulos anteriores são aplicados a esse conjunto de dados e no final desse capítulo encontram-se as comparações e conclusões. Em

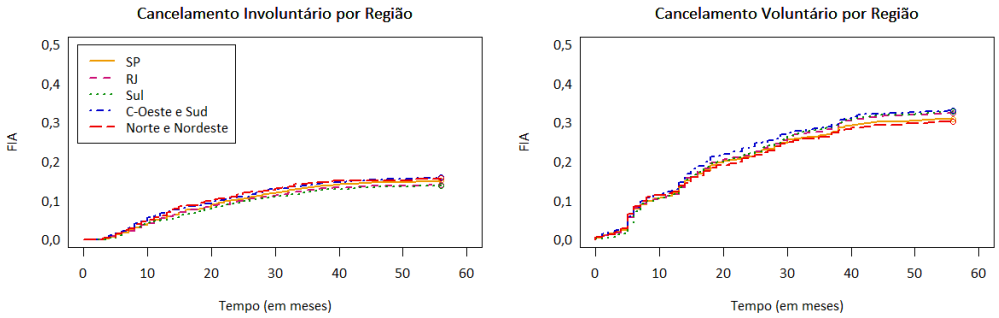


Figura 5.6: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Estado ou Região

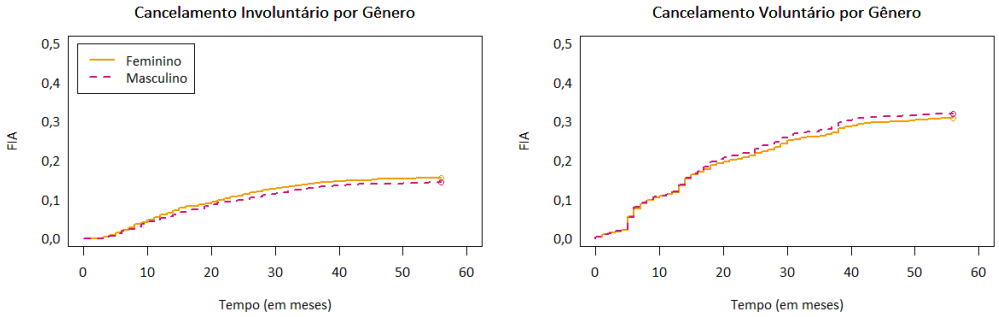


Figura 5.7: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Gênero

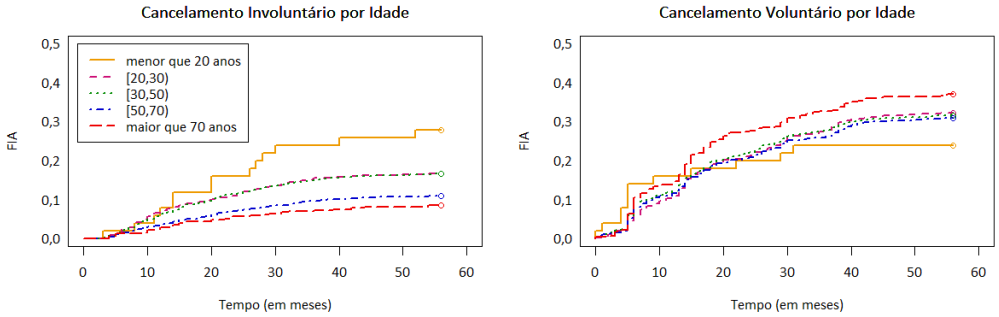


Figura 5.8: Funções de Incidência Acumulada para Cancelamento Voluntário e Involuntário por Faixa de Idade

todos os métodos os riscos serão tratados como:

- Risco 1: Cancelamento Involuntário;
- Risco 2: Cancelamento Voluntário.

5.2 Modelagem Paramétrica

O primeiro método pelo qual os dados foram ajustados é aquele que se baseia no uso da distribuição Gompertz descrito na Seção 3.1. Essa é a proposta mais simples que combina presença de riscos competitivos e fração de cura, mas que não prevê o uso de covariáveis para o ajuste. A ausência de covariáveis nesse contexto é claramente um ponto negativo pois não é possível relacionar as funções de incidência acumulada com as diferentes características de um aplicante. Apesar disso, o método foi aplicado a esses dados porque pode oferecer alguma informação sobre qual é a proporção de cartões ativos (fração de cura) após uma janela de tempo longa para fornecer informações úteis para o cálculo de lucros e prejuízos.

Tabela 5.1: Estimativas para os parâmetros da distribuição Gompertz para os modelos de cancelamento involuntário e voluntário.

	$\hat{\eta}$	DP	Valor-P	$\hat{\rho}$	DP	Valor-P
Cancelamento Involuntário	-5.0578	0.0368	< 0.0001	-0.0332	0.0015	< 0.0001
Cancelamento Voluntário	-4.1991	0.0248	< 0.0001	-0.0334	0.0011	< 0.0001

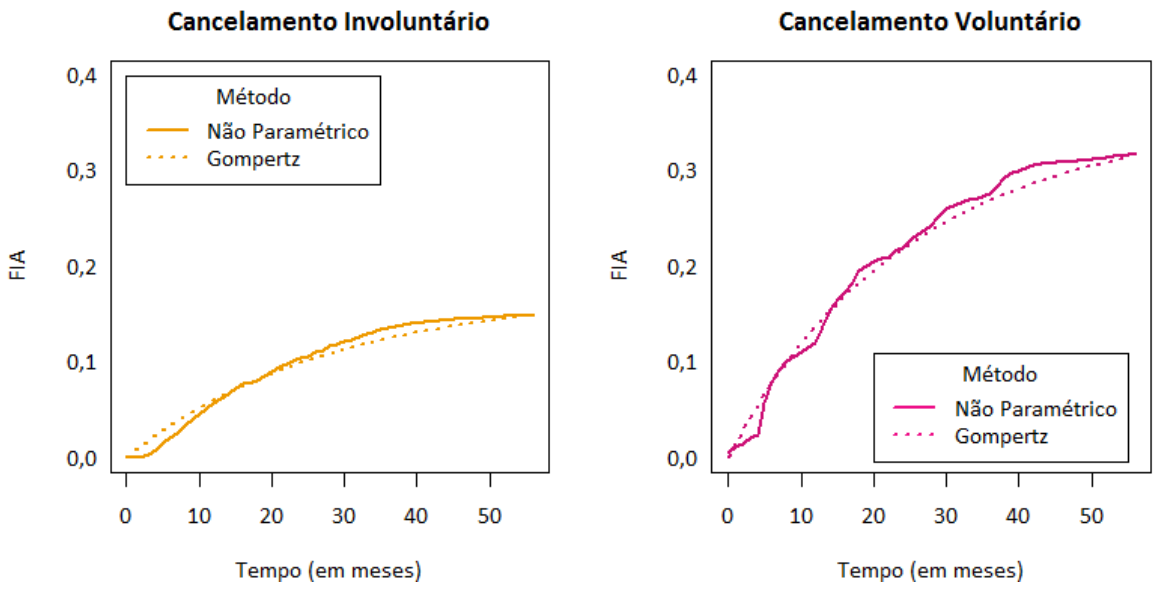


Figura 5.9: Comparação entre métodos paramétrico baseado na distribuição Gompertz e não-paramétrico para obtenção das estimativas das funções de incidência acumulada do cancelamento involuntário e voluntário.

Ambas as funções de incidência acumulada para cancelamento involuntário e voluntário foram ajustadas segundo o modelo (3.5), com a seguinte reparametrização a fim de evitar complicações numéricas ao maximizar a verossimilhança:

$$CI_j(t) = F^G(t; \rho_j, \eta_j) = 1 - \exp\left(\frac{\exp(\eta_j)}{\rho_j}(1 - \exp(\rho_j t))\right), \tag{5.1}$$

com $t > 0$, $j = 1, 2$, $-\infty < \rho_j < \infty$ e $-\infty < \eta_j < \infty$. Na Tabela 5.1 encontram-se os valores estimados para os parâmetros desse modelo e as demais informações sobre essas estimativas. Note que as estimativas para ρ_1 e ρ_2 são valores negativos, assim como esperado pois as curvas de incidência acumulada nesse conjunto de dados se assemelham a alguma distribuição imprópria, que no caso da Gompertz acontece quando esses valores são negativos. Seguindo também a sugestão dos autores, é possível testar a presença de cura para cada risco (assim como detalhado em (3.13)), testando se cada um dos parâmetros ρ_1 e ρ_2 são negativos. O teste é da forma $H_0 : \rho_j \geq 0$ contra $H_a : \rho_j < 0$ e para realizá-lo utiliza-se a estatística:

$$T_{\rho_j} = \frac{\hat{\rho}_j}{\hat{DP}(\hat{\rho}_j)}, j = 1, 2, \tag{5.2}$$

em que $\hat{DP}(\hat{\rho}_j)$ é a estimativa do desvio padrão para a estimativa de ρ_j. Para ambos os tipos de cancelamento, o teste apresenta valor-p inferior a $0,0001$ e podemos então dizer que os dados possuem fração de cura. Assim, as frações de cura estimadas segundo o modelo são $82,6\%$ e $63,8\%$ para o cancelamento involuntário e voluntário, respectivamente. A fração de cura global estimada é $46,4\%$, e nesse caso trata-se da estimativa da proporção de cartões ativos extrapolando para um instante em que ambas as curvas de incidência acumulada já estão estabilizadas. Vale ressaltar que essa informação seria mais rica se fosse possível relacionar diferentes estratos, porém trata-se de uma informação útil para avaliar uma carteira de cartões de créditos como um todo. Note que por ser um modelo paramétrico, essas estimativas refletem a projeção das estimativas obtidas com base nos primeiros 56 meses contidos nos dados e, antes de realizar essa extrapolação, é importante considerar se esse seguimento é longo suficiente para tanto. Na Figura 5.9, podemos comparar as curvas estimadas pelo método de Jeong e Fine (2006) com as estimativas não-paramétricas. Apesar das curvas de incidência acumulada estimadas não-parametricamente não possuírem a forma da distribuição Gompertz, ambos os métodos fornecem estimativas muito semelhantes. Note que no início da janela de observação até o sexto mês, para ambos os tipos de cancelamento, o método baseado na Gompertz superestima as funções de incidência acumulada comparado ao método não paramétrico, como um reflexo dessa diferença de forma, e que também, no cancelamento voluntário esse método paramétrico não é capaz de capturar uma incidência acentuada que ocorre por volta do quadragésimo mês. Sem levar em conta essas diferenças sutis, podemos dizer que o método é uma forma simples e barata do ponto de vista de processamento para a obtenção das estimativas das funções de incidência acumulada e que é capaz de fornecer estimativas para a fração de cura, além de permitir o teste de sua presença.

5.3 Modelagem Paramétrica com Covariáveis

O segundo método explorado nesse texto é na verdade um aprimoramento do modelo anterior, que incorpora o uso de covariáveis e pelo qual uma transformação é aplicada à função de incidência acumulada. Esse método paramétrico poderia ser uma alternativa simples para dados com presença de riscos competitivos e fração de cura, mas dada algumas dificuldades em sua implementação e vulnerabilidade da maximização da log-verossimilhança dados diferentes pontos iniciais, trata-se de uma alternativa exaustiva para a aplicação.

Ao implementá-lo é importante notar que existe um subespaço no espaço paramétrico definido por Jeong e Fine (2007) em que na realidade as funções principais atingem valores não razoáveis, como por exemplo a função de sobrevivência e subdensidade podem atingir valores negativos. Ao maximizar a função de log-verossimilhança um método pode esbarrar nesse subespaço indesejado e ser interrompido ao tentar calcular o logaritmo de um número negativo. Além disso, como a implementação requer uma maximização numérica, foi importante o estudo analítico dos limites atingidos pelas funções quando os parâmetros assumem valores específicos. Relembre em (3.23) que a função de log-verossimilhança depende basicamente da função de subdensidade e função de incidência acumulada:

$$\begin{aligned} l(\boldsymbol{\psi}) &= \sum_{i=1}^{N}\left[\left(\sum_{j=1}^{K}\delta_{ji}\log(f_j(t_i,\boldsymbol{\psi}_j;\mathbf{z}_i))\right)+\left(1-\sum_{j=1}^{K}\delta_{ji}\right)\log(S(t_i,\boldsymbol{\psi}_j;\mathbf{z_i}))\right] \\ &= \sum_{i=1}^{N}\left[\left(\sum_{j=1}^{K}\delta_{ji}\log(f_j(t_i,\boldsymbol{\psi}_j;\mathbf{z}_i))\right)+\left(1-\sum_{j=1}^{K}\delta_{ji}\right)\log(1-\sum_{j=1}^{K}CI_j(t_i,\boldsymbol{\psi}_j;\mathbf{z_i}))\right]. \end{aligned} \tag{5.3}$$

Portanto, foram estudados os limites dessas funções para valores dos parâmetros que pudessem gerar uma instabilidade numérica na implementação. Para simplicidade de notação, os índices i e j relacionados ao indivíduo e ao risco foram suprimidos nos resultados abaixo e a reparametrização $\tau = \exp(\eta)$ foi adotada por ser conveniente para o uso dos métodos de maximização. Dessa forma, a função de incidência acumulada e os limites utilizados foram:

$$CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1-\left(1+\alpha\frac{\exp(\eta)}{\rho}(\exp(\rho t)-1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)^{\frac{-1}{\alpha}}$$

$$\text{(I)}\lim_{\alpha\to 0} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1-\exp\left(-\frac{\exp(\eta)}{\rho}(\exp(\rho t)-1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)$$

$$\text{(II)}\lim_{\rho\to 0} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1-\left(1+\alpha\exp(\eta)t\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)^{\frac{-1}{\alpha}}$$

$$\text{(III)}\lim_{\substack{\rho\to 0\\ \alpha\to 0}} CI(t,\boldsymbol{\psi};\mathbf{Z}) = 1-\exp\left(-\exp(\eta)t\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)$$

A função de log-verossimilhança e os limites obtidos para ela foram:

$$\log f(t,\boldsymbol{\psi};\mathbf{Z}) = \left(\frac{-1}{\alpha}-1\right)\log\left(1+\alpha\frac{\exp(\eta)}{\rho}(\exp(\rho t)-1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)\right)+\mathbf{Z}^T\boldsymbol{\beta}+\eta+\rho t$$

$$\text{(I)}\lim_{\alpha\to 0} f(t,\boldsymbol{\psi};\mathbf{Z}) = \mathbf{Z}^T\boldsymbol{\beta}+\eta+\rho t-\frac{\exp(\eta)}{\rho}(\exp(\rho t)-1)\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)$$

$$\text{(II)} \lim_{\rho \to 0} f(t, \boldsymbol{\psi}; \mathbf{Z}) = \left(\frac{-1}{\alpha} - 1\right) \log\left(1 + \alpha \exp(\eta) t \exp\left(\mathbf{Z}^T \boldsymbol{\beta}\right)\right) + \mathbf{Z}^T \boldsymbol{\beta} + \eta$$

$$\text{(III)} \lim_{\substack{\rho \to 0 \\ \alpha \to 0}} f(t, \boldsymbol{\psi}; \mathbf{Z}) = -\exp(\eta) t \exp\left(\mathbf{Z}^T \boldsymbol{\beta}\right) + \mathbf{Z}^T \boldsymbol{\beta} + \eta$$

Outro ponto importante a ressaltar sobre esse método é que como as estimativas para todos os parâmetros de todos os riscos são obtidas simultaneamente, a remoção de variáveis pouco significantes em um risco interfere nas estimativas e significâncias das categorias de outras covariáveis, até mesmo em outro risco competitivo a ele. Os autores de Jeong e Fine (2007) sugerem que tanto é possível estimar α como fixá-lo, em 0 ou 1, para que a função de transformação aplicada a função de incidência acumulada sejam respectivamente as funções complemento log-log e logito. A fim de testar mais de uma combinação, foram ajustados os modelos com α_1 e α_2 livres, $\alpha_1 = \alpha_2 = 0$ e $\alpha_1 = \alpha_2 = 1$. As estimativas obtidas estão apresentadas na Tabela 5.2, em que é possível notar que existe pouca diferença entre as estimativas e que inclusive, pela estimativa de α_1 no modelo com ele livre, não é descartada a hipótese de $\alpha_1 = 0$. Apesar da pouca diferença, note que os modelos finais contém variáveis e categorias agrupadas de forma diferente, e que o método mostra alta significância para todas as categorias em cada um dos modelos dado o número de observações, mas nas Figuras 5.10, 5.11 e 5.12 é possível ver que as estimativas não são muito próximas às curvas de incidência acumulada estimadas pelo método não paramétrico, especialmente quando o número de falhas nos estratos é menor (essas quantidades podem ser encontradas na Tabela 5.4).

Tabela 5.2: Estimativas para os modelos com α_1 e α_2 fixado em 1, 0 e livres.

Cancelamento Involuntário	Opções de Modelo								
	$\alpha_1 = \alpha_2 = 1$			$\alpha_1 = \alpha_2 = 0$			α_1 e α_2 livres		
Parâmetros / Efeitos	Estimativa	DP	Valor-P	Estimativa	DP	Valor-P	Estimativa	DP	Valor-P
α_1	1,000	-	-	0,000	-	-	0,063	0,296	0,8306
η_1	-5,105	0,077	$< 0,0001$	-5,176	0,044	$< 0,0001$	-5,288	0,055	$< 0,0001$
ρ_1	-0,024	0,002	$< 0,0001$	-0,030	0,002	$< 0,0001$	-0,030	0,003	$< 0,0001$
Escore de Crédito									
Sem Escore ou Baixo	-	-	-	-	-	-	-	-	-
Alto	1,488	0,061	$< 0,0001$	1,188	0,047	$< 0,0001$	1,278	0,077	$< 0,0001$
Médio	0,398	0,072	$< 0,0001$	-	-	-	0,351	0,066	$< 0,0001$
Baixíssimo	-1,395	0,090	$< 0,0001$	-1,446	0,085	$< 0,0001$	-1,335	0,089	$< 0,0001$
Canal									
Tele-Marketing Ativo / Mala-Direta	-	-	-	-	-	-	-	-	-
Agência de Varejo	-0,264	0,072	0,0003	-	-	-	-	-	-
Internet / Tele-Marketing Passivo	-	-	-	0,650	0,076	$< 0,0001$	0,825	0,107	$< 0,0001$
Cancelamento Voluntário									
α_2	1,000	-	-	0,000	-	-	-8,318	0,233	$< 0,0001$
η_2	-3,626	0,079	$< 0,0001$	-3,862	0,067	$< 0,0001$	-4,436	0,018	$< 0,0001$
ρ_2	-0,024	0,001	$< 0,0001$	-0,033	0,001	$< 0,0001$	-0,103	0,002	$< 0,0001$
Escore de Crédito									
Sem Escore, Baixo ou Médio	-	-	-	-	-	-	-	-	-
Alto	-0,397	0,056	$< 0,0001$	-0,363	0,050	$< 0,0001$	-0,043	0,008	$< 0,0001$
Baixíssimo	0,189	0,040	$< 0,0001$	0,169	0,033	$< 0,0001$	0,014	0,003	$< 0,0001$
Canal									
Tele-Marketing Ativo e Passivo / Internet	-	-	-	-	-	-	-	-	-
Agência de Varejo	-0,529	0,078	$< 0,0001$	-0,314	0,064	$< 0,0001$	-	-	-
Mala-Direta	-1,280	0,098	$< 0,0001$	-0,930	0,085	$< 0,0001$	-	-	-

Para escolher um desses modelos, uma alternativa é comparar os valores do AIC para cada

Figura 5.10: Comparação entre as estimativas das FIAs pelo método não paramétrico (linhas cheias) e pelo método paramétrico (linhas tracejadas) para o modelo com α_1 e α_2 livres.

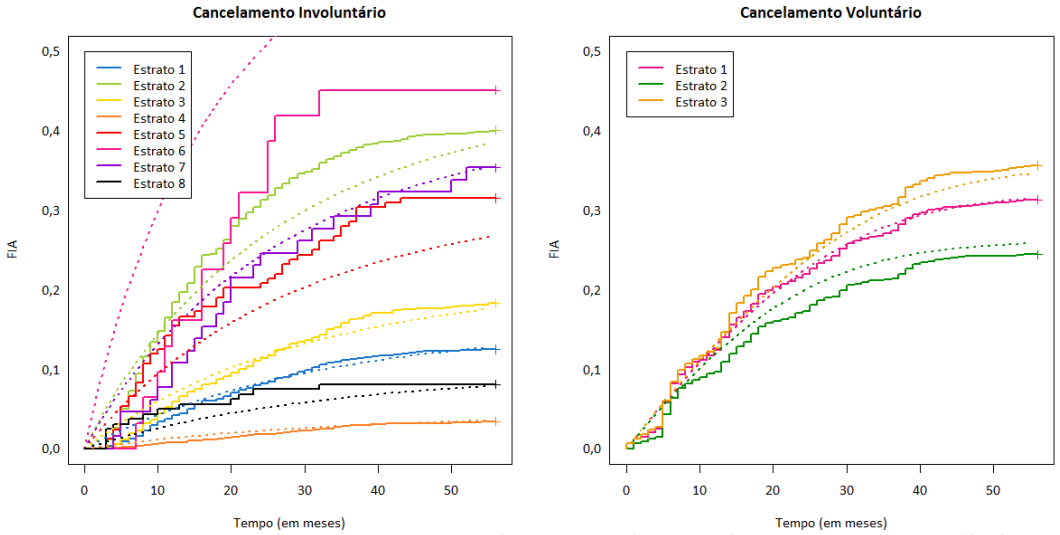


Figura 5.11: Comparação entre as estimativas das FIAs pelo método não paramétrico (linhas cheias) e pelo método paramétrico (linhas tracejadas) para o modelo com $\alpha_1 = \alpha_2 = 0$.

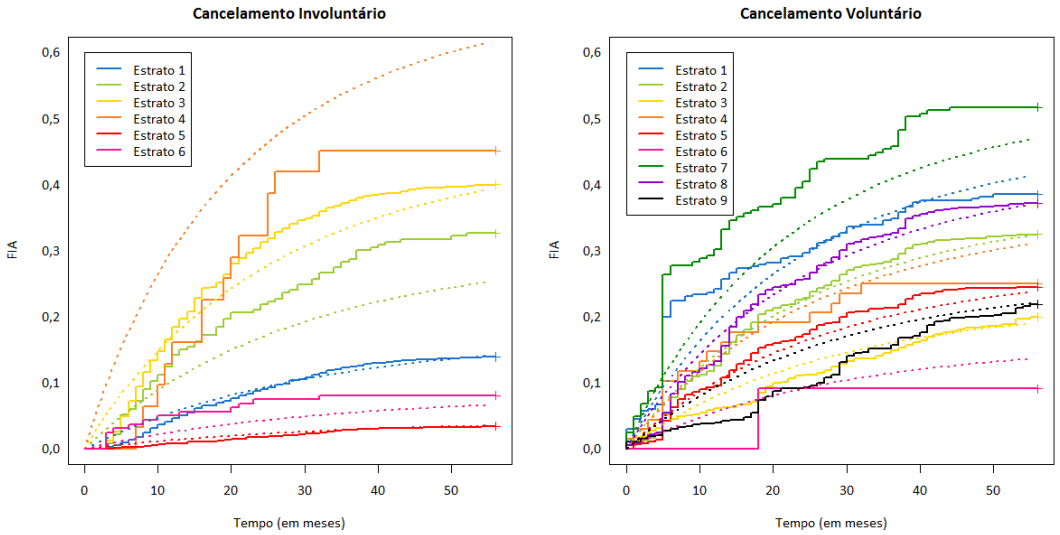
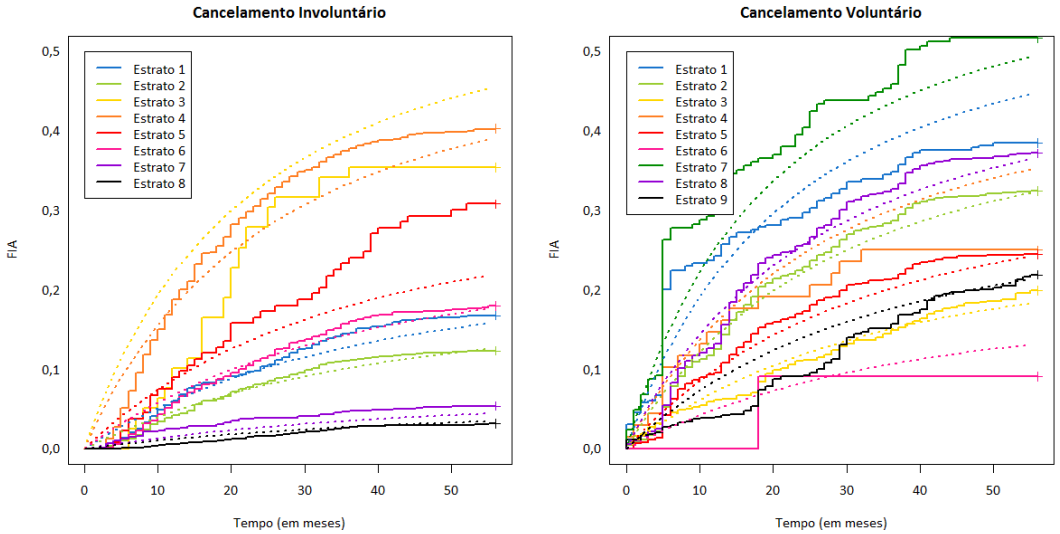


Figura 5.12: Comparação entre as estimativas das FIAs pelo método não paramétrico (linhas cheias) e pelo método paramétrico (linhas tracejadas) para o modelo com $\alpha_1 = \alpha_2 = 1$.



um dos modelos, que são 71180,56, 71248,27 e 71313,73 para os modelos com α_1 e α_2 livres, com $\alpha_1 = \alpha_2 = 0$ e com $\alpha_1 = \alpha_2 = 1$, respectivamente. Com base nisso, podemos dizer que a inclusão de dois parâmetros não foi suficiente para penalizar o modelo que já era esperado como a melhor alternativa. A vantagem dos modelos alternativos testados seria a facilidade de interpretação, como já explicado anteriormente.

Em todos os modelos é possível testar se ρ_j é negativo, visando *testar a presença de cura do j-ésimo risco*. O teste é da forma $H_0 : \rho_j \geq 0$ contra $H_a : \rho_j < 0$ e para realizá-lo utiliza-se a estatística:

$$T_{\rho_j} = \frac{\hat{\rho}_j}{\hat{DP}(\hat{\rho}_j)}, j = 1, 2, \tag{5.4}$$

em que $\hat{DP}(\hat{\rho}_j)$ é a estimativa do desvio padrão para a estimativa de ρ_j. Na Tabela 5.2, encontram-se os valores-p obtidos para o teste bilateral e portanto, os valores-p são menores que 0,0001 para o teste unilateral. Conclui-se que para ambos os riscos nos três modelos apresentados anteriormente existe a presença da fração de cura. Dessa forma, como todos os riscos em todos os modelos apresentam valores negativos, é possível estimá-la para cada combinação de estrato e risco. Além de calculadas as estimativas para $t \to \infty$, estão dispostas na Tabela 5.3 as estimativas para $t = 56$ que é o último ponto observado no conjunto de dados para o modelo escolhido. Nota-se que as estimativas obtidas para a fração de cura pela forma sugerida pelos autores, isto é, em $t \to \infty$, podem ser bem diferentes das obtidas em $t = 56$ para alguns estratos. Novamente, vale destacar que essa extrapolação deve ser realizada com cautela. Nesse contexto, seria razoável assumir a extrapolação sob a premissa de que nenhum evento inesperado irá acontecer após o mês 56.

Tabela 5.3: Fração de cura estimada em $t = 56$ e $t \to \infty$ para os estratos do modelo com α_1 e α_2 livres.

Modelo: α_1 e α_2 livres

Cancelamento Involuntário				Cancelamento Voluntário			
Estrato/Descrição		Fração de Cura (%)		Estrato/Descrição		Fração de Cura (%)	
		$t = 56$	$t \to \infty$			$t = 56$	$t \to \infty$
1	Escore: Sem Escore ou Baixo Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	87,15	84,38	1	Escore: Sem Escore, Baixo ou Médio	68,43	67,82
2	Escore: Alto Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	61,35	54,81	2	Escore: Alto	74,13	73,81
3	Escore: Médio Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	82,28	78,62	3	Escore: Baixíssimo	65,29	64,40
4	Escore: Baixíssimo Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	96,43	95,61	4			
5	Escore: Sem Escore ou Baixo Canal: Tele-Marketing Passivo ou Internet	73,18	68,05	5			
6	Escore: Alto Canal: Tele-Marketing Passivo ou Internet	33,50	26,17	6			
7	Escore: Médio Canal: Tele-Marketing Passivo ou Internet	64,30	58,05	7			
8	Escore: Baixíssimo Canal: Tele-Marketing Passivo ou Internet	92,06	90,28	8			

Tabela 5.4: Descrição e número de falhas para cada estrato dos riscos 1 e 2 apresentados nas Figuras 5.10, 5.11 e 5.12.

Modelo: α_1 e α_2 livres

Cancelamento Involuntário			Cancelamento Voluntário		
Estrato	Descrição	Falhas	Estrato	Descrição	Falhas
1	Escore: Sem Escore ou Baixo Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	599	1	Escore: Sem Escore, Baixo ou Médio	2065
2	Escore: Alto Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	825	2	Escore: Alto	511
3	Escore: Médio Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	287	3	Escore: Baixíssimo	1584
4	Escore: Baixíssimo Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	148	4		
5	Escore: Sem Escore ou Baixo Canal: Tele-Marketing Passivo ou Internet	53	5		
6	Escore: Alto Canal: Tele-Marketing Passivo ou Internet	14	6		
7	Escore: Médio Canal: Tele-Marketing Passivo ou Internet	23	7		
8	Escore: Baixíssimo Canal: Tele-Marketing Passivo ou Internet	13	8		

Modelo: $\alpha_1 = \alpha_2 = 0$

Cancelamento Involuntário			Cancelamento Voluntário		
Estrato	Descrição	Falhas	Estrato	Descrição	Falhas
1	Escore: Sem Escore, Baixo ou Médio Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	886	1	Escore: Sem Escore, Baixo ou Médio Canal: Tele-Marketing Ativo, Passivo ou Internet	127
2	Escore: Sem Escore, Baixo ou Médio Canal: Tele-Marketing Passivo ou Internet	76	2	Escore: Sem Escore, Baixo ou Médio Canal: Mala Direta	1792
3	Escore: Alto Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	825	3	Escore: Sem Escore, Baixo ou Médio Canal: Agência de Varejo	146
4	Escore: Alto Canal: Tele-Marketing Passivo ou Internet	14	4	Escore: Alto Canal: Tele-Marketing Ativo, Passivo ou Internet	17
5	Escore: Baixíssimo Canal: Tele-Marketing Ativo, Mala Direta ou Agência de Varejo	148	5	Escore: Alto Canal: Mala Direta	493
6	Escore: Baixíssimo Canal: Tele-Marketing Passivo ou Internet	13	6	Escore: Alto Canal: Agência de Varejo	11
			7	Escore: Baixíssimo Canal: Tele-Marketing Ativo, Passivo ou Internet	106
			8	Escore: Baixíssimo Canal: Mala Direta	1336
			9	Escore: Baixíssimo Canal: Agência de Varejo	142

Modelo: $\alpha_1 = \alpha_2 = 1$

Cancelamento Involuntário			Cancelamento Voluntário		
Estrato	Descrição	Falhas	Estrato	Descrição	Falhas
1	Escore: Sem Escore ou Baixo Canal: Tele-Marketing Ativo e Passivo, Internet ou Mala Direta	156	1	Escore: Sem Escore, Baixo ou Médio Canal: Tele-Marketing Ativo, Passivo ou Internet	127
2	Escore: Sem Escore ou Baixo Canal: Agência de Varejo	496	2	Escore: Sem Escore, Baixo ou Médio Canal: Agência de Varejo	1792
3	Escore: Alto Canal: Tele-Marketing Ativo e Passivo / Internet / Mala Direta	28	3	Escore: Sem Escore, Baixo ou Médio Canal: Mala-Direta	146
4	Escore: Alto e Canal: Agência de Varejo Canal: Agência de Varejo	811	4	Escore: Alto e Canal: Tele-Marketing Ativo e Passivo / Internet Canal: Tele-Marketing Ativo e Passivo / Internet	17
5	Escore: Médio Canal: Tele-Marketing Ativo e Passivo / Internet / Mala Direta	41	5	Escore: Alto Canal: Agência de Varejo	493
6	Escore: Médio Canal: Agência de Varejo	269	6	Escore: Alto Canal: Mala-Direta	11
7	Escore: Baixíssimo Canal: Tele-Marketing Ativo e Passivo / Internet / Mala Direta	46	7	Escore: Baixíssimo Canal: Tele-Marketing Ativo e Passivo / Internet	106
8	Escore: Baixíssimo Canal: Agência de Varejo	115	8	Escore: Baixíssimo Canal: Agência de Varejo	1336
			9	Escore: Baixíssimo Canal: Mala-Direta	142

5.4 Pseudo-Valores

O método de modelagem direta das funções de incidência acumulada que se baseia no uso de pseudo-valores também foi aplicado ao conjunto de dados descrito no início do capítulo em que se pretende estimar probabilidades de cancelamento após a abertura de um contrato de cartão de crédito. Assim como descrito no capítulo anterior, é fundamental para a aplicação desse método a escolha de uma grade de pontos no tempo para os quais os valores das funções de incidência acumulada dos dois tipos de cancelamento serão estimados.

No trabalho de Klein e Andersen (2005), os autores argumentam que a escolha da quantidade de pontos, desde que superior a 5, não altera as inferências obtidas, assim como não deve alterar por grandes diferenças absolutas as estimativas obtidas para os efeitos das covariáveis. Sugerem ainda que a escolha seja de 5 a 10 pontos e que devem ser extraídos de tal maneira que entre os pontos existam um número igual de eventos. Com o objetivo de testar e exemplificar essa sugestão, foram escolhidos quatro diferentes grades de pontos seguindo o mesmo raciocínio proposto no artigo, com 5, 10, 25 e 50 pontos.

Seguindo o trabalho posterior desses autores em Klein et al. (2008), mais focado na implementação do método, o pacote "pseudo" do R foi utilizado para calcular os pseudo-valores e a partir deles o pacote "geepack" foi utilizado para ajustar os modelos, da forma descrita em (3.30), com as seguintes funções de ligação:

- Complemento Log-Log, em que $g(x) = \log(-\log(1-x))$;
- Logito, em que $g(x) = \log(\frac{x}{1-x})$.

Para cada função de ligação e para cada tipo de cancelamento, um ajuste foi realizado utilizando a matriz de covariância de trabalho como a matriz identidade, e então iterativamente foram removidas as covariáveis não significativas e agrupadas as categorias para enfim resultarem nos modelos finais que têm as estimativas dos efeitos das covariáveis apresentadas nas Tabelas 5.5 e 5.6.

Fica claro que os ajustes de cada tipo de cancelamento resultaram em diferentes covariáveis e agrupamentos, mas que as duas funções de ligação testadas para cada um deles não alteram esse resultado. Outro ponto importante é que as covariáveis para um mesmo modelo final, bem como o agrupamento de suas categorias, são iguais e com significâncias bem semelhantes independente da escolha da grade, apenas com alterações segunda casa decimal de suas estimativas. Podemos concluir que a escolha da quantidade de pontos influencia muito pouco nas estimativas e que, pelo menos nesses dados, 25 pontos parecem ser a melhor escolha que balanceia tempo de processamento e inferências semelhantes ao caso extremo de 50 pontos aqui testado. Vale lembrar também que as estimativas das funções de incidência acumulada só estarão disponíveis para os pontos presentes na grade, o que talvez justificasse uma escolha de pontos mais ampla. Em todas as grades testadas, as estimativas para os parâmetros associados ao tempo foram significativos a 1%.

Tabela 5.5: Estimativas baseadas nas grades de 5, 10, 25 e 50 pontos e matriz de covariância de trabalho identidade para o modelo final do **cancelamento involuntário**.

Função de ligação: Logito												
	Grade de 5 pontos			Grade de 10 pontos			Grade de 25 pontos			Grade de 50 pontos		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal												
Tele-Marketing Ativo / Agência de Varejo / Mala Direta	-	-	-	-	-	-	-	-	-	-	-	-
Internet / Tele-Marketing Passivo	0,814	0,139	$< 0,0001$	0,801	0,141	$< 0,0001$	0,796	0,143	$< 0,0001$	0,788	0,144	$< 0,0001$
Escore de Crédito												
Sem Escore	-	-	-	-	-	-	-	-	-	-	-	-
Alto Risco	1,458	0,072	$< 0,0001$	1,448	0,073	$< 0,0001$	1,452	0,074	$< 0,0001$	1,451	0,074	$< 0,0001$
Médio Risco	0,284	0,087	0,0011	0,266	0,089	0,0027	0,257	0,090	0,0043	0,252	0,091	0,0054
Baixo Risco	-0,176	0,089	0,0477	-0,187	0,090	0,0385	-0,185	0,092	0,0434	-0,187	0,092	0,0421
Baixíssimo Risco	-1,549	0,108	$< 0,0001$	-1,558	0,110	$< 0,0001$	-1,572	0,113	$< 0,0001$	-1,580	0,114	$< 0,0001$
Gênero												
Feminino	-	-	-	-	-	-	-	-	-	-	-	-
Masculino	-0,179	0,061	0,0035	-0,183	0,062	0,0032	-0,18	0,063	0,0045	-0,184	0,064	0,0038
Função de ligação: Complemento Log-Log												
	Grade de 5 pontos			Grade de 10 pontos			Grade de 25 pontos			Grade de 50 pontos		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal												
Tele-Marketing Ativo / Agência de Varejo / Mala Direta	-	-	-	-	-	-	-	-	-	-	-	-
Internet / Tele-Marketing Passivo	0,666	0,124	$< 0,0001$	0,656	0,127	$< 0,0001$	0,653	0,129	$< 0,0001$	0,646	0,13	$< 0,0001$
Escore de Crédito												
Sem Escore	-	-	-	-	-	-	-	-	-	-	-	-
Alto Risco	1,286	0,065	$< 0,0001$	1,283	0,066	$< 0,0001$	1,293	0,067	$< 0,0001$	1,293	0,067	$< 0,0001$
Médio Risco	0,268	0,081	0,0009	0,252	0,082	0,0022	0,244	0,084	0,0036	0,24	0,085	0,0046
Baixo Risco	-0,157	0,084	0,0598	-0,168	0,085	0,0476	-0,168	0,087	0,0527	-0,170	0,087	0,0509
Baixíssimo Risco	-1,491	0,103	$< 0,0001$	-1,503	0,106	$< 0,0001$	-1,520	0,109	$< 0,0001$	-1,529	0,109	$< 0,0001$
Gênero												
Feminino	-	-	-	-	-	-	-	-	-	-	-	-
Masculino	-0,154	0,054	0,0042	-0,159	0,055	0,0037	-0,156	0,056	0,005	-0,160	0,056	0,0043

Para a escolha da função de ligação mais adequada para o conjunto de dados, os autores sugerem que sejam analisadas contra o eixo do tempo as seguintes diferenças:

$$D_{ja_i}(t) = g(\hat{CI}_{ja_i}(t)) - g(\hat{CI}_{ja_{ref}}(t)),$$

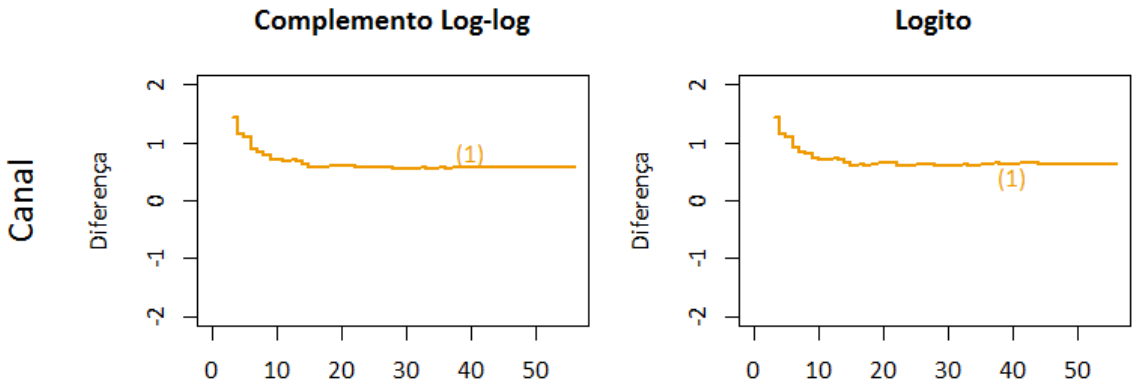
em que $g(.)$ é a função de ligação testada, $\hat{CI}_{ja_i}(t)$ é a função de incidência acumulada estimada não-parametricamente para o j-ésimo risco no i-ésimo estrato definido pela covariável a e a_{ref} representa uma categoria dessa mesma covariável que foi escolhida para ser referência. Se a função de ligação testada for correta, as curvas graficadas devem ser aproximadamente linhas horizontais dada a forma do modelo. Portanto, foram avaliadas as covariáveis "Escore de Crédito", "Canal", "Gênero", "Produto" e "Região" com as categorias agrupadas como nos modelos finais. A representação gráfica das diferenças obtidas no tempo considerando as funções de ligação complemento log-log e logito para o cancelamento involuntário e voluntário são apresentadas nas Figuras 5.13, 5.14 e 5.14. Note que a partir delas não é clara uma preferência por uma das funções de ligação, mas que ambas transformações para os dois tipos de cancelamento mostram linhas horizontais e portanto, a escolha é adequada e indiferente. Dessa forma, foram escolhidos os modelos com a função de ligação logito pela facilidade de interpretação dos parâmetros, além de uma sutil preferência por essa função no caso da covariável "Escore de Crédito" no cancelamento involuntário.

O último item testado foi como seriam alteradas as estimativas e inferências ao ajustar esses mesmos modelos finais com a matriz de covariância de trabalho exata. Como o ajuste com a matriz de covariância de trabalho exata é extremamente caro do ponto de vista de consumo de tempo, vale ressaltar que nesse trabalho não foi testado se levariam a modelos diferentes caso

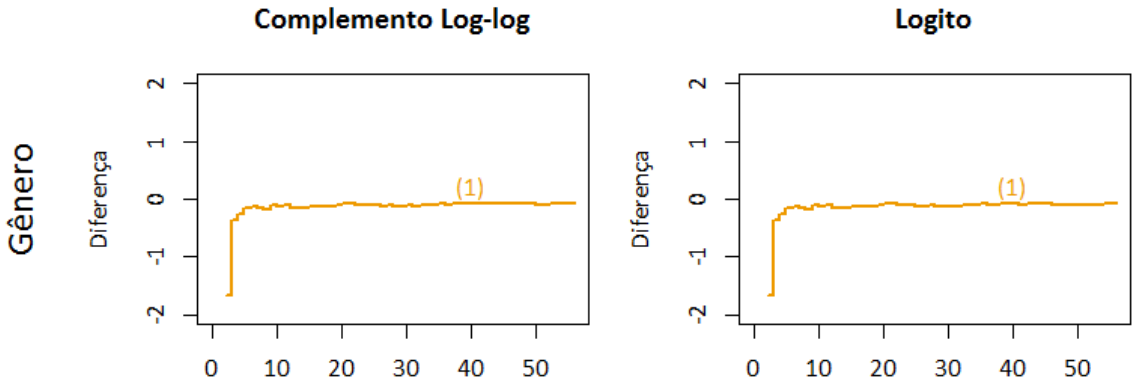
Tabela 5.6: Estimativas baseadas nas grades de 5, 10, 25 e 50 pontos e matriz de covariância de trabalho identidade para o modelo final do **cancelamento voluntário**.

Função de ligação: Logito												
	Grade de 5 pontos			Grade de 10 pontos			Grade de 25 pontos			Grade de 50 pontos		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal												
Tele-Marketing Ativo / Agência de Varejo	-	-	-	-	-	-	-	-	-	-	-	-
Mala-Direta	-0,948	0,072	$< 0,0001$	-0,983	0,074	$< 0,0001$	-1,019	0,077	$< 0,0001$	-1,036	0,078	$< 0,0001$
Internet / Tele-Marketing Passivo	0,870	0,106	$< 0,0001$	0,892	0,108	$< 0,0001$	0,910	0,109	$< 0,0001$	0,909	0,109	$< 0,0001$
Escore de Crédito												
Sem Escore / Baixo Risco	-	-	-	-	-	-	-	-	-	-	-	-
Alto Risco	-0,392	0,061	$< 0,0001$	-0,395	0,062	$< 0,0001$	-0,388	0,063	$< 0,0001$	-0,388	0,063	$< 0,0001$
Médio Risco	-0,138	0,064	0,0300	-0,153	0,065	0,0182	-0,146	0,065	0,0255	-0,150	0,066	0,0226
Baixíssimo Risco	0,148	0,045	0,0009	0,134	0,045	0,0031	0,132	0,046	0,0040	0,131	0,046	0,0044
Gênero												
Feminino	-	-	-	-	-	-	-	-	-	-	-	-
Masculino	0,087	0,040	0,03120	0,090	0,041	0,0280	0,091	0,041	0,0277	0,091	0,041	0,0285
Produto												
Black / Platinum / Gold	-	-	-	-	-	-	-	-	-	-	-	-
Internacional	-0,292	0,067	$< 0,0001$	-0,282	0,068	$< 0,0001$	-0,292	0,070	$< 0,0001$	-0,293	0,070	$< 0,0001$
Região / Estado												
Demais estados*	-	-	-	-	-	-	-	-	-	-	-	-
MG, BA, DF, GO, ES, MT ou MS	0,169	0,058	0,0036	0,171	0,059	0,0036	0,173	0,059	0,0035	0,175	0,060	0,0034
Função de ligação: Complemento Log-Log												
	Grade de 5 pontos			Grade de 10 pontos			Grade de 25 pontos			Grade de 50 pontos		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal												
Tele-Marketing Ativo / Agência de Varejo	-	-	-	-	-	-	-	-	-	-	-	-
Mala-Direta	-0,840	0,065	$< 0,0001$	-0,878	0,069	$< 0,0001$	-0,916	0,071	$< 0,0001$	-0,933	0,072	$< 0,0001$
Internet / Tele-Marketing Passivo	0,692	0,084	$< 0,0001$	0,712	0,085	$< 0,0001$	0,729	0,086	$< 0,0001$	0,730	0,087	$< 0,0001$
Escore de Crédito												
Sem Escore / Baixo Risco	-	-	-	-	-	-	-	-	-	-	-	-
Alto Risco	-0,342	0,054	$< 0,0001$	-0,347	0,055	$< 0,0001$	-0,343	0,056	$< 0,0001$	-0,343	0,056	$< 0,0001$
Médio Risco	-0,120	0,055	0,0292	-0,133	0,056	0,0177	-0,128	0,057	0,0244	-0,132	0,057	0,0217
Baixíssimo Risco	0,128	0,038	0,0007	0,117	0,039	0,0025	0,116	0,039	0,0032	0,115	0,040	0,0036
Gênero												
Feminino	-	-	-	-	-	-	-	-	-	-	-	-
Masculino	0,073	0,035	0,0350	0,076	0,035	0,0310	0,077	0,036	0,0312	0,077	0,036	0,0320
Produto												
Black / Platinum / Gold	-	-	-	-	-	-	-	-	-	-	-	-
Internacional	-0,240	0,058	$< 0,0001$	-0,234	0,059	0,0001	-0,243	0,061	0,0001	-0,244	0,061	0,0001
Região / Estado												
Demais estados*	-	-	-	-	-	-	-	-	-	-	-	-
MG, BA, DF, GO, ES, MT ou MS	0,143	0,049	0,0037	0,146	0,050	0,0037	0,149	0,051	0,0035	0,150	0,051	0,0033

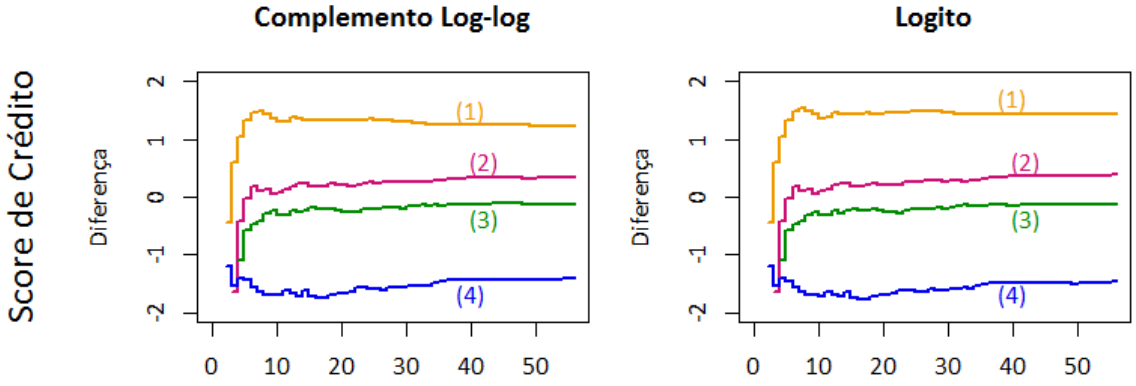
A categoria "Demais estados" para a variável "Região" equivale aos estados que não estão na lista "MG, BA, DF, GO, ES, MT ou MS".



Referência: "Tele-Marketing Ativo", "Agência de Varejo" e"Mala Direta"; (1) "Internet / Tele-Marketing Passivo".

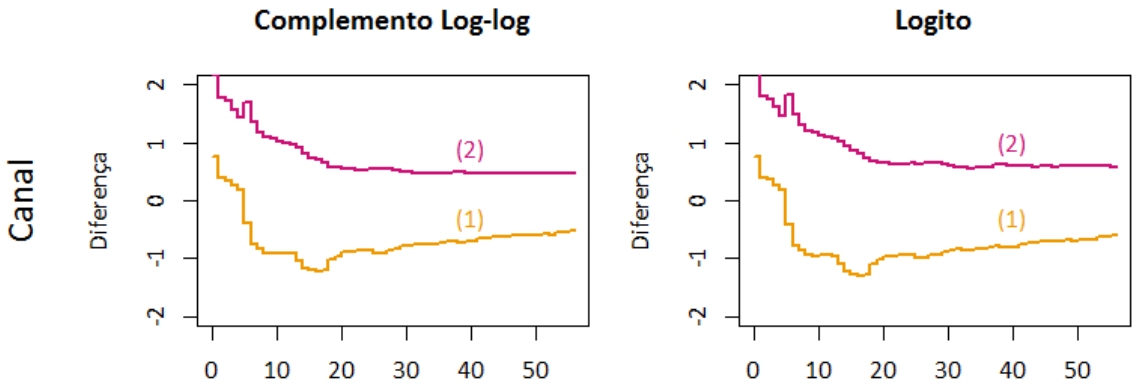


Referência: "Feminino"; (1) "Masculino".

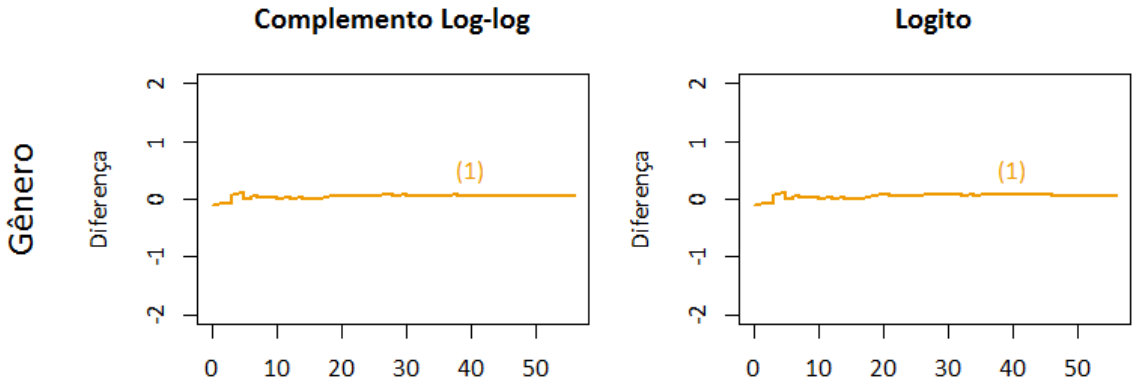


Referência: "Sem escore"; (1) "Alto Risco"; (2) "Médio Risco"; (3) "Baixo Risco"; (4) "Baixíssimo Risco".

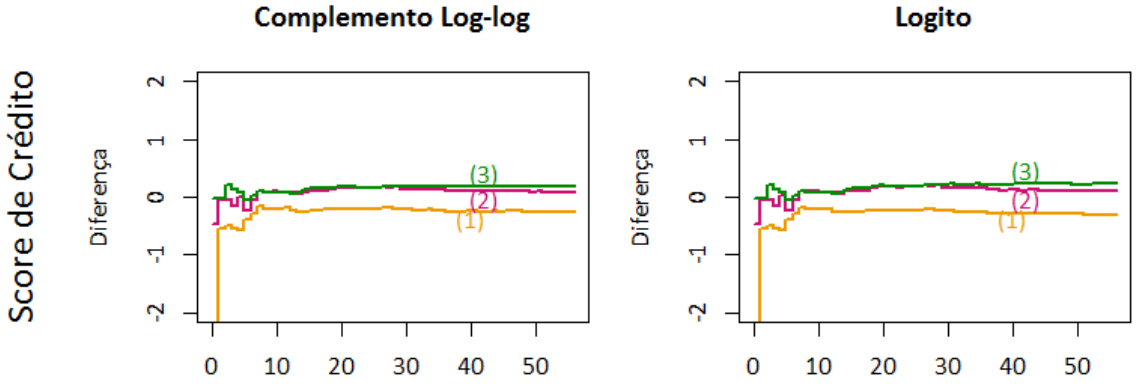
Figura 5.13: Diferenças entre logito e complementar log-log das FIAs para escolha da função de ligação do modelo de **cancelamento involuntário**.



Referência: "Tele-Marketing Ativo" e "Agência de Varejo"; (1) "Mala Direta"; (2) "Internet / Tele-Marketing Passivo".



Referência: "Feminino"; (1) "Masculino".



Referência: "Sem escore" e "Baixo Risco"; (1) "Alto Risco"; (2) "Médio Risco"; (3) "Baixíssimo Risco".

Figura 5.14: Diferenças entre logito e complementar log-log das FIAs para escolha da função de ligação do modelo de **cancelamento voluntário**.

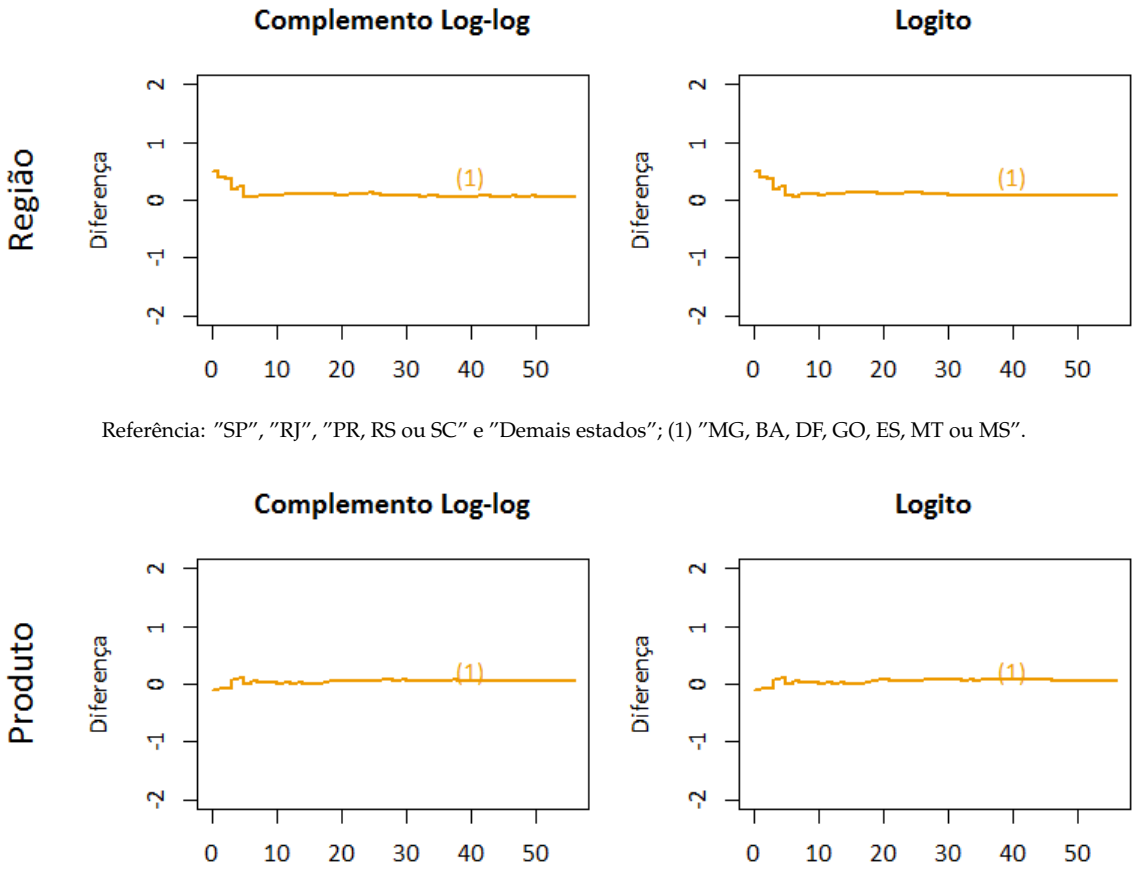


Figura 5.14: Diferenças entre logito e complementar log-log das FIAs para escolha da função de ligação do modelo de **cancelamento voluntário**.

fossem aplicados desde a primeira iteração antes de todo o processo de seleção de variáveis. Nas Tabelas 5.7 e 5.8 é possível comparar as estimativas para os mesmos modelos dado a escolha da função de ligação logito, considerando os mesmos 25 pontos de grade, sendo a única diferença a escolha da matriz de covariância de trabalho entre identidade e exata. A partir delas confirma-se a argumentação dos autores que a mudança da matriz de covariância de trabalho praticamente não altera as estimativas, nem seus desvios padrão para justificar a utilização de uma matriz de covariância de trabalho diferente da matriz identidade, proposta por eles como uma simplificação de implementação do método.

Tabela 5.7: Estimativas baseadas na grade de 25 pontos e matriz de covariância de trabalho identidade ou exata para o modelo escolhido de **cancelamento involuntário**.

Função de ligação: Logito	Matriz Identidade			Matriz Exata		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal						
Tele-Marketing Ativo / Agência de Varejo / Mala Direta	-	-	-	-	-	-
Internet / Tele-Marketing Passivo	0,801	0,141	$< 0,0001$	0,781	0,138	$< 0,0001$
Escore de Crédito						
Sem Escore	-	-	-	-	-	-
Alto Risco	1,448	0,073	$< 0,0001$	1,449	0,072	$< 0,0001$
Médio Risco	0,266	0,089	0,0027	0,286	0,087	0,0010
Baixo Risco	-0,187	0,090	0,0385	-0,173	0,088	0,0489
Baixíssimo Risco	-1,558	0,110	$< 0,0001$	-1,537	0,105	$< 0,0001$
Gênero						
Feminino	-	-	-	-	-	-
Masculino	-0,183	0,062	0,0032	-0,179	0,061	0,0034

Tabela 5.8: Estimativas baseadas na grade de 25 pontos e matriz de covariância de trabalho identidade ou exata para o modelo escolhido de **cancelamento voluntário**.

Função de ligação: Logito	Matriz Identidade			Matriz Exata		
	$\hat{\beta}$	DP	Valor-P	$\hat{\beta}$	DP	Valor-P
Canal						
Tele-Marketing Ativo / Agência de Varejo	-	-	-	-	-	-
Mala-Direta	-0,983	0,074	$< 0,0001$	-0,865	0,070	$< 0,0001$
Internet / Tele-Marketing Passivo	0,892	0,108	$< 0,0001$	0,923	0,108	$< 0,0001$
Escore de Crédito						
Sem Escore / Baixo Risco	-	-	-	-	-	-
Alto Risco	-0,395	0,062	$< 0,0001$	-0,415	0,060	$< 0,0001$
Médio Risco	-0,153	0,065	0,0182	-0,152	0,063	0,0153
Baixíssimo Risco	0,134	0,045	0,0031	0,148	0,044	0,0008
Gênero						
Feminino	-	-	-	-	-	-
Masculino	0,090	0,041	0,028	0,092	0,040	0,0202
Produto						
Black / Platinum / Gold	-	-	-	-	-	-
Internacional	-0,282	0,068	$< 0,0001$	-0,259	0,067	0,0001
Região / Estado						
Demais estados*	0,171	0,059	0,0036	0,173	0,057	0,0025
MG, BA, DF, GO, ES, MT ou MS	0,171	0,059	0,0036	0,173	0,057	0,0025

A categoria "Demais estados" para a variável "Região" equivale aos estados que não estão na lista "MG, BA, DF, GO, ES, MT ou MS".

Considerando como modelos finais aqueles com a função de ligação logito e matriz de covariância de trabalho identidade, as funções de incidência acumulada estimadas pelo método

com base nos pseudo-valores obtidos para a grade de 25 pontos também foram comparadas com as estimativas não paramétricas, resultando em curvas muito semelhantes para estratos com volume considerável. A seguir, nas Figuras 5.15 e 5.15 alguns estratos foram selecionados para ilustrar a comparação e é possível notar o quanto as curvas são parecidas em patamar e forma. Nota-se que, na Figura 5.15, no cancelamento voluntário o ajuste é capaz de se adaptar bem ao início das curvas, inclusive no salto do sexto mês que ocorre dado o processo de cancelamento automático após os clientes não desbloquearem seus cartões nos seis meses iniciais. Enfim, podemos interpretar o efeito das covariáveis nas funções de incidência acumulada, pois se $\hat{\beta}$ é a estimativa de um parâmetro relacionado a uma categoria de uma cováriável de interesse, $\exp(\hat{\beta})$ é a razão de chances entre essa categoria e a referência dessa covariável num mesmo ponto no tempo quando todas as demais covariáveis permanecerem iguais.

A chance de cancelamento involuntário, que está diretamente atrelada à inadimplência, é 4,25 vezes maior quando um aplicante é classificado como "Alto Risco" no "Escore de Crédito" contra outro aplicante classificado como "Sem escore", e também 20,2 vezes maior contra um aplicante classificado como "Baixíssimo Risco". Também nesse modelo, as aplicantes do gênero feminino tem chance 1,2 maior de cancelarem seus cartões por inadimplência do que os homens e, como esperado, o canal mais arriscado é o "Internet / Tele-marketing Passivo", pelo qual os aplicantes que buscam a instituição financeira.

O modelo de cancelamento voluntário indica que a mesma covariável que tanto discrimina probabilidade de cancelamento involuntário também é útil para discriminar a probabilidade de cancelamento por desinteresse no produto de forma inversa, no qual a chance é estimada 1,7 vezes maior por um aplicante considerado como "Baixíssimo Risco" contra um de "Alto Risco". Essa informação é útil e deve ser considerada em uma análise de rentabilidade para balancear o risco do produto e seu retorno financeiro. Provavelmente por ser um produto mais barato, o tipo de cartão "Internacional" tem chance de cancelamento 0,75 vezes a chance de um cartão "Black", "Platinum" ou "Gold" mas que pode indicar que os planos de recompensa dessas últimas categorias podem não ser suficientes para justificar seus preços. Diferente do esperado, a chance de cancelamento por desinteresse no produto no canal "Internet / Tele-Marketing Passivo" é a maior, 2,48 vezes a da referência "Agência de Varejo" e "Tele-marketing Ativo", o que levanta um questionamento sobre o quão bem divulgados por esse canal são as características do produto, já que os próprios aplicantes que buscaram o produto passam a se desinteressar por mantê-lo.

Como o interesse reside principalmente na proporção de cartões ativos para cada estrato, essas quantidades foram estimadas a partir, como proposto em (3.34), e na Tabela 5.9 seguem os valores obtidos apenas para alguns estratos selecionados, dado que são mais de 86 estratos considerando todas as combinações de estratos do modelo do cancelamento involuntário e voluntário. É interessante notar que para as faixas de menor risco as proporções de ativos são maiores para mulheres que homens, enquanto nas demais faixas mais arriscadas as proporções de ativos são maiores para homens. Essa inversão ocorre devido a maior probabilidade de cancelamento voluntário pelos homens do que mulheres e pelo comportamento inverso no cancelamento involuntário. As proporções de cartões ativos, bem como o tempo para que

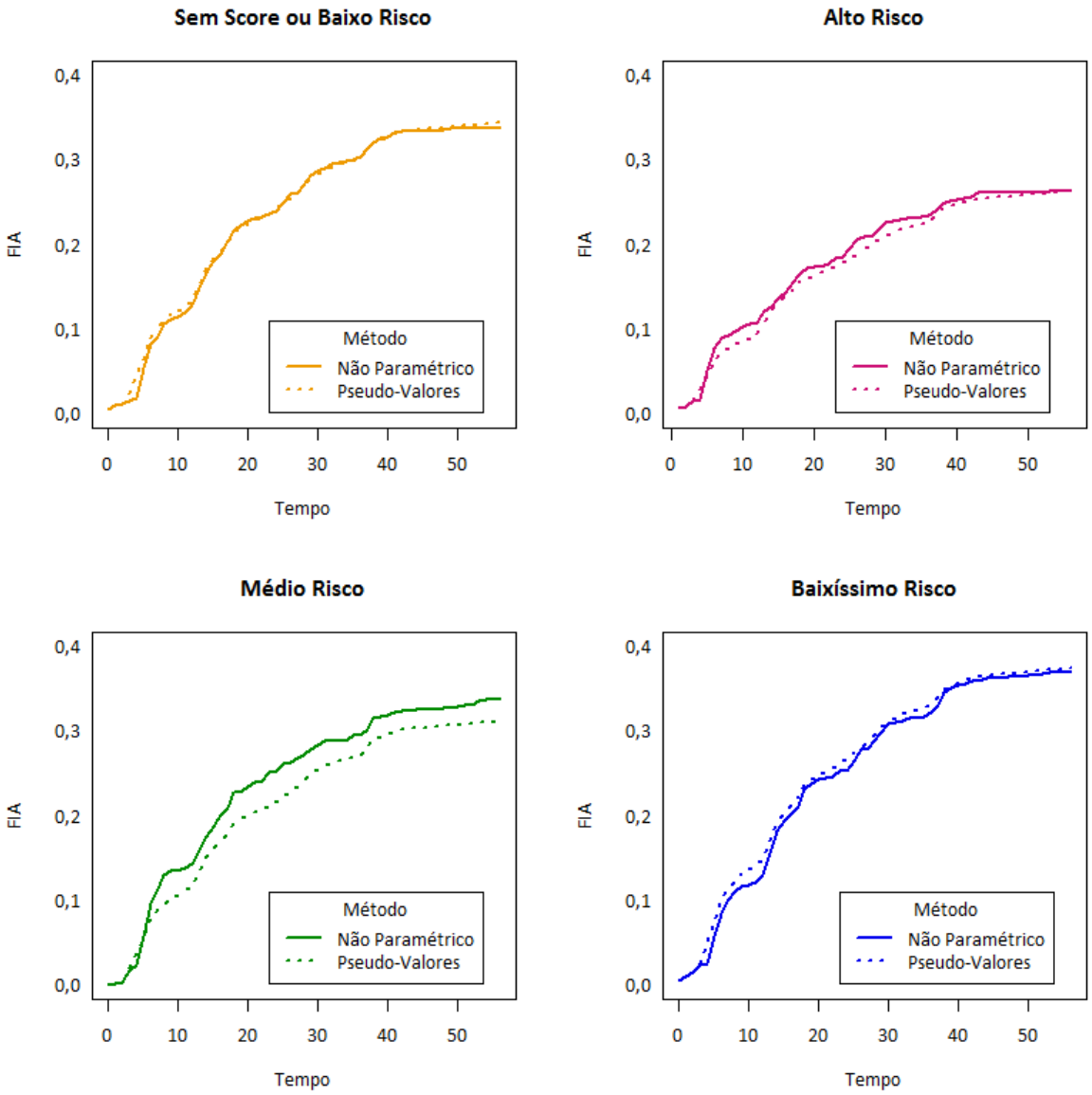


Figura 5.15: Comparação entre estimativas das FIAs pelo método não paramétrico e pelos pseudo-valores. Os estratos selecionados foram aqueles com canal "Tele-Marketing Ativo ou Agência de Varejo", gênero "Masculino", produto "Black, Platinum ou Gold" e região "SP, RJ, PR, RS, SC" e "Demais estados", com as diferentes faixas de "Escore de Crédito". O risco é o **cancelamento voluntário**.

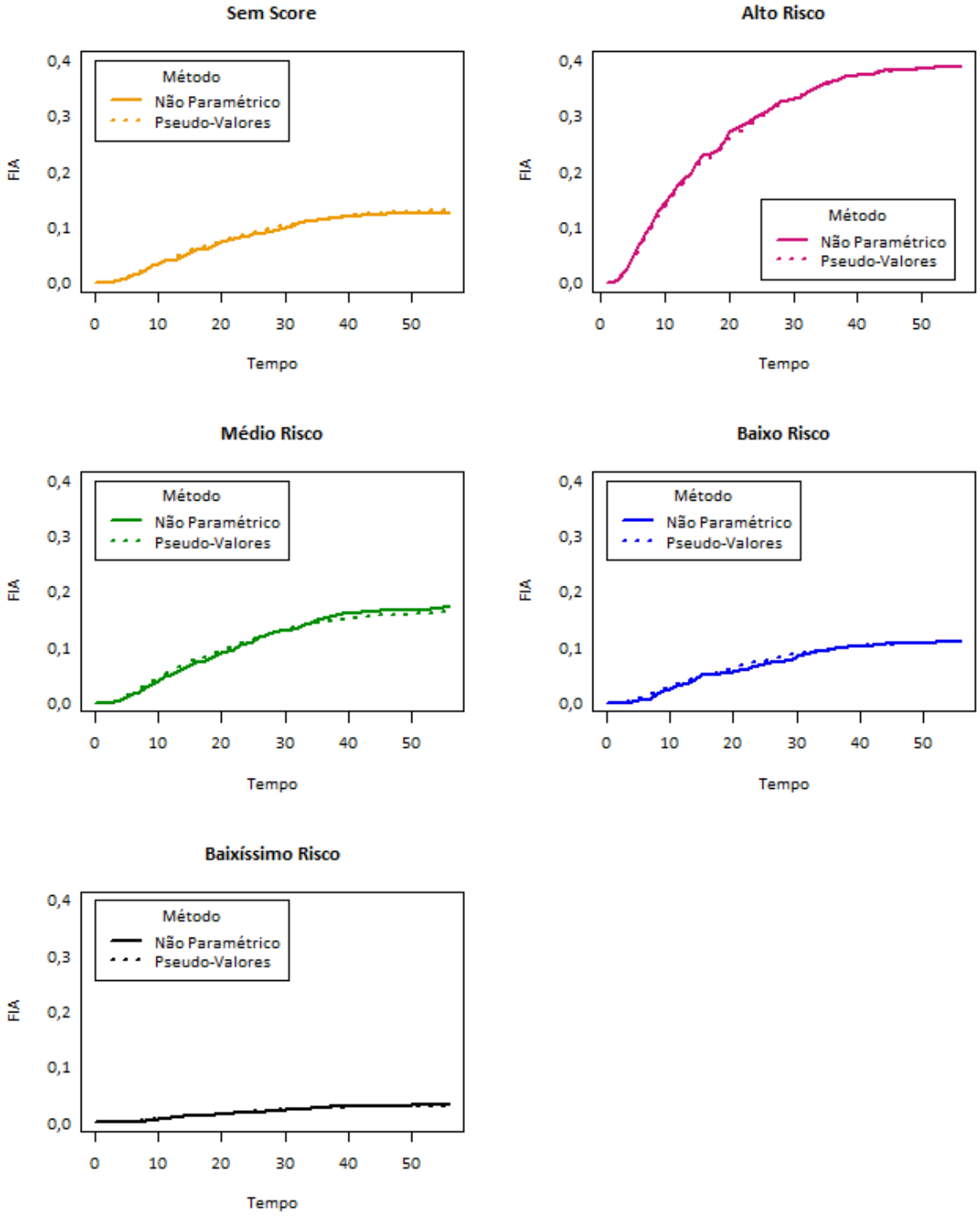


Figura 5.15: Comparação entre estimativas das FIAs pelo método não paramétrico e pelos pseudo-valores. Os estratos selecionados foram aqueles com canal "Tele-Marketing Ativo, Agência de Varejo ou Mala Direta", gênero "Masculino", com as diferentes faixas de "Escore de Crédito". O risco é o **cancelamento involuntário**.

essas proporções sejam atingidas formam importante subsídio para análises de rentabilidade do produto, projeção de perdas e planejamento de investimentos.

Tabela 5.9: Estimativas para proporções de cartões ativos após 56 meses para estratos selecionados.

	Gênero	
Escore de Crédito	Feminino	Masculino
Alto Risco	31,97%	34,60%
Médio Risco	51,79%	52,47%
Baixo Risco	54.55%	54,42%
Baixíssimo Risco	61,00%	59,47%
Sem escore	52,30%	52,46%

As categorias selecionadas foram agrupamento de "Tele-Marketing Ativo" e "Agência de Varejo" para a covariável "Canal", o agrupamento das categorias "Black", "Platinum" e "Gold" para a covariável "Produto" e todas as categorias da variável "Região" que não a "MG, BA, DF, GO, ES, MT ou MS".

Baseando-se nos resultados obtidos, algumas vantagens desse método são a qualidade do ajuste que captura a forma das curvas de incidência acumulada detalhadamente, a flexibilidade do modelo que permite a escolha de diferentes funções de ligação e a facilidade de interpretação dos efeitos das covariáveis dependendo dessa escolha. Por outro lado, é importante notar que a geração dos pseudo-valores é extremamente cara do ponto de vista de processamento, espaço em memória e tempo consumido e, depedendo do tamanho da amostra em um estudo, esse método pode até se tornar impraticável.

5.5 Comparação

Para comparar ambos os métodos utilizados nas seções anteriores, foram ajustados os modelos que contém somente a covariável "Escore de Crédito" com os níveis "Sem escore ou Baixo Risco"', "Alto Risco", "Médio Risco" ou "Baixíssimo Risco". A covariável "Escore de Crédito" é especialmente interessante pois a relação entre seus níveis e a incidência de cancelamentos é reversa quando comparados cancelamento involuntário e voluntário. Enquanto para o cancelamento involuntário, quanto maior o risco de crédito maior será a incidência de cancelamentos, para o cancelamento voluntário, quanto maior o risco de crédito menor será a incidência de cancelamentos. Essa característica reforça a importância do uso de técnicas apropriadas para riscos competitivos nesses dados. Outra simplificação nessa seção é o uso dos mesmos níveis para a covariável escolhida, a fim de facilitar o cálculo das estimativas das frações de cura global por estrato.

Nas Figuras 5.16 e 5.17, encontram-se as funções de incidência acumulada estimadas pelos métodos não paramétrico, paramétrico e pelo uso dos pseudo-valores. Apesar de não existir grande diferença entre as estimativas obtidas a partir de cada um dos métodos, o uso de pseudo-valores mostra-se mais adequado para esse conjunto de dados especialmente no cancelamento involuntário, que demora cerca de seis meses para ter uma incidência acentuada e portanto, difere bastante das possíveis formas da função de incidência acumulada derivada da

modelagem com a distribuição Gompertz.

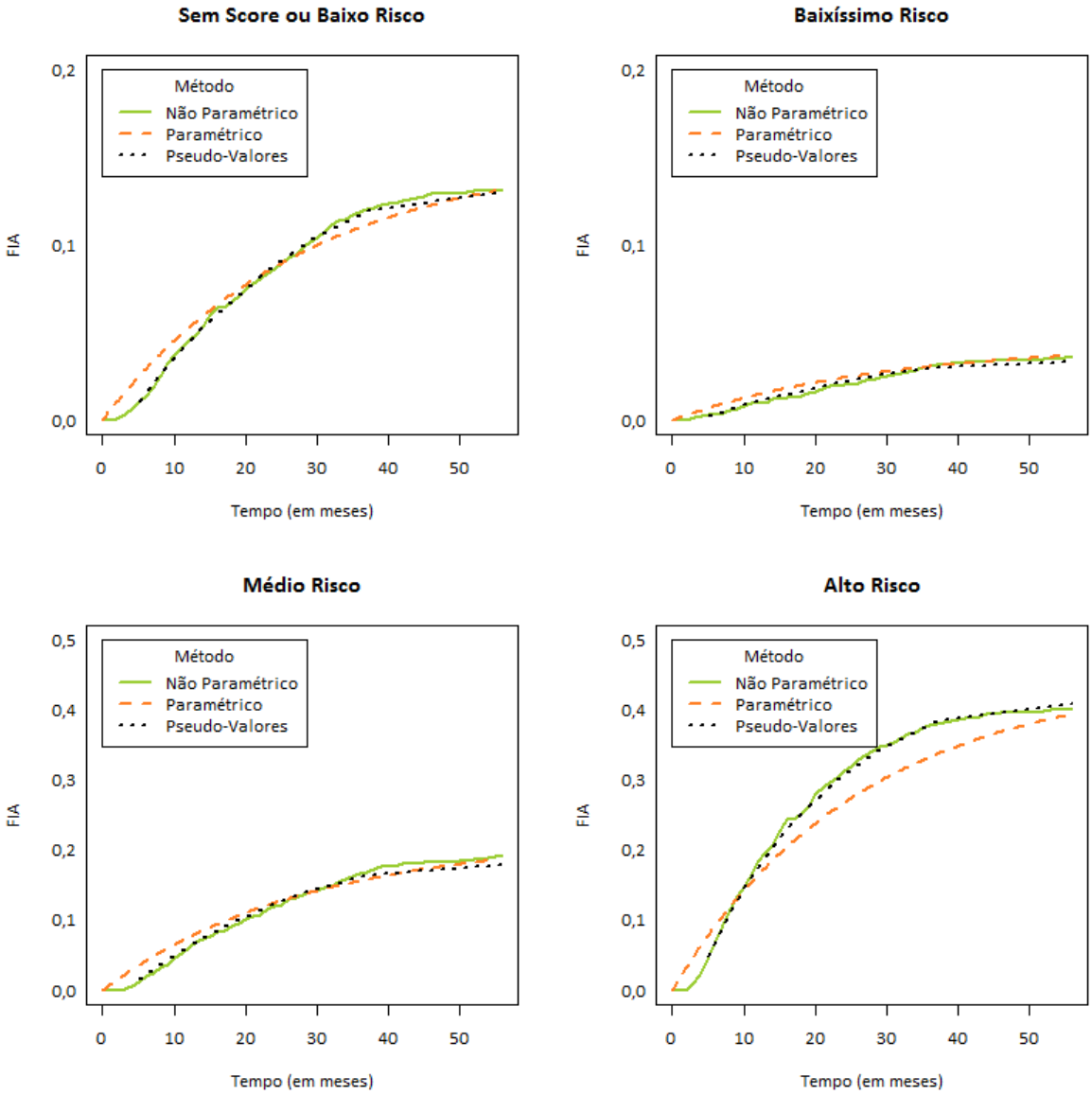


Figura 5.16: Comparação entre estimativas das FIAs pelo método não paramétrico, paramétrico e uso de pseudo-valores. O risco é o **cancelamento involuntário**.

Finalmente, as frações de cura por risco e global estão apresentadas da Tabela X. Novamente, para o método paramétrico foram apresentadas as estimativas da fração de cura em $t = 56$ e $t \to \infty$. Nota-se que não há muita diferença entre as estimativas obtidas pelo uso de pseudo-valores e pelo método paramétrico em $t = 56$, mas as estimativas em $t \to \infty$ para os níveis "Alto Risco" e "Médio Risco" são bem menores relativamente que as avaliadas em $t = 56$. Assumindo a premissa que nenhum evento inesperado ocorrerá após o quinquagésimo sexto mês, a extrapolação para outros pontos no tempo, por exemplo $t = 60$ ou $t = 72$, é muitas vezes interessante no contexto de análises de perdas e rentabilidade e essa técnica é portanto, uma

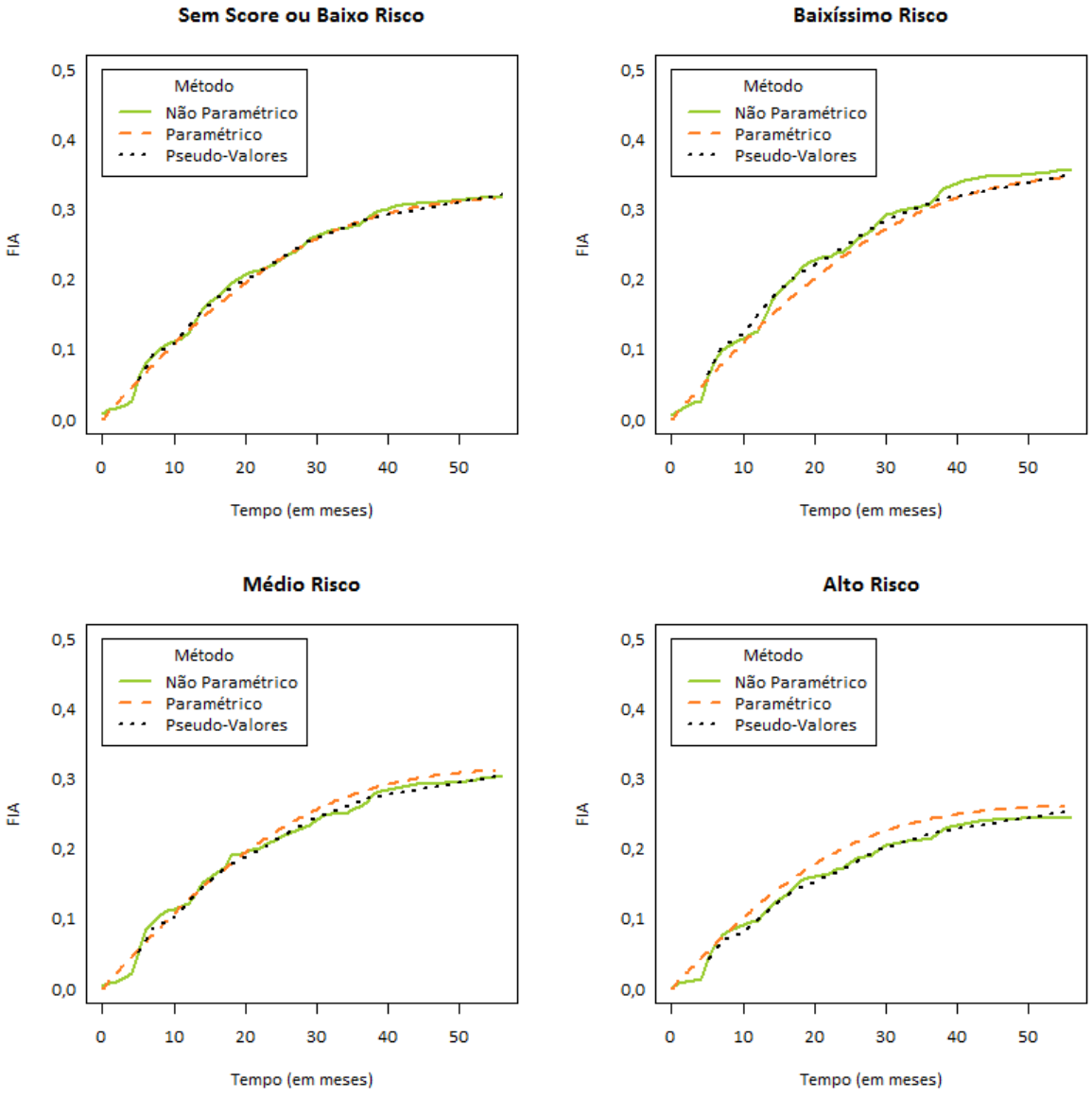


Figura 5.17: Comparação entre estimativas das FIAs pelo método não paramétrico, paramétrico e uso de pseudo-valores. O risco é o **cancelamento voluntário**.

alternativa para obter tais estimativas. Assim, apesar do uso de pseudo-valores mostrar-se mais adequado nesse conjunto de dados, os métodos paramétricos baseados na distribuição Gompertz também podem ser úteis.

Tabela 5.10: Estimativas para proporções de cartões ativos pelo uso de pseudo-valores e método paramétrico em $t = 56$ e $t \to \infty$.

	Fração de Cura								
	Cancelamento Involuntário			Cancelamento Voluntário			Global		
Escore de Crédito	Pseudo-Valores	Paramétrico		Pseudo-Valores	Paramétrico		Pseudo-Valores	Paramétrico	
		$t = 56$	$t \to \infty$		$t = 56$	$t \to \infty$		$t = 56$	$t \to \infty$
Sem Escore ou Baixo Risco	86,95%	86,91%	84,62%	67,92%	68,44%	67,81%	54,87%	55,35%	52,43%
Alto Risco	59,14%	60,84%	54,61%	74,63%	73,90%	73,57%	33,77%	34,74%	28,18%
Médio Risco	82,03%	81,43%	78,24%	69,58%	68,78%	68,18%	51,61%	50,21%	46,42%
Baixíssimo Risco	96,65%	96,34%	95,69%	65,12%	65,48%	64,56%	61,78%	61,82%	60,25%

Capítulo 6

Discussão

Durante o trabalho desenvolvido e apresentado nesse texto, buscou-se conduzir uma revisão bibliográfica cuidadosa para tentar cobrir todos os métodos propostos para dados com presença simultânea de riscos competitivos e fração de cura até então. Dessa forma, é possível dizer que poucas propostas, até onde se sabe, foram desenvolvidas para dados que contêm ambas as características e a contribuição desse texto é, além de consolidar mais de uma alternativa para tal, compará-las a fim de entender em quais condições um método pode superar o outro.

Os métodos paramétricos, testados nesse texto, apresentam fragilidades numéricas na implementação que podem ser contornadas, mas que acabam limitando sua aplicabilidade. A maximização da função de log-verossimilhança sugerida pelos autores é muito vulnerável à escolha dos parâmetros iniciais nas funções de otimização, o que torna a implementação e aplicação do método mais exaustiva. Apesar disso, os modelos são alternativas interessantes pois, em geral, fornecem boas estimativas para valores de tempo altos em que observa-se a estabilização das curvas de incidência acumulada, além de permitir o teste sobre a existência dessa estabilização para um determinado risco competitivo. Um desenvolvimento teórico sobre a não identificabilidade desses modelos ainda é necessário, como é pontuado pelos próprios autores. Além disso, os métodos ainda são carentes de um análise da qualidade do ajuste formal e bem fundamentada teoricamente.

O uso de pseudo-valores para a estimação da função de incidência acumulada *na presença de fração de cura* se mostrou um método eficiente e muito flexível, que se ajusta a formas de curvas de incidência acumulada bem específicas, como é o caso do cancelamento voluntário na aplicação de dados de cartão de crédito em que é possível notar um comportamento sazonal (de seis em seis meses) de incidência acentuada. A escolha da grade de pontos e a escolha da matriz de covariância de trabalho parecem influenciar muito pouco nas estimativas, e portanto recomenda-se que a grade seja composta pelo menos por 10 pontos equidistantes na escala de eventos e que a matriz seja a identidade para agilizar e simplificar a implementação do método. Como esse modelo fornece estimativas somente para pontos no tempo que estiverem contidos nos dados observados, se o interesse for por uma estimativa projetada em um tempo

maior que o seguimento dos dados, o método não atenderá as expectativas. Embora essa seja uma limitação do método quando comparado aos modelos paramétricos aqui testados, não se trata de um ponto negativo uma vez que essa extrapolação para pontos no tempo não observados nos dados é uma prática criticável por muitos autores. Uma segunda limitação do método, principalmente se comparado aos métodos paramétricos aqui testados, é o tempo de processamento e memória exigidos. É possível que em determinados casos, com amostras grandes - os dados de cartão de crédito com 13.114 indivíduos já são um exemplo -, o método se torne até impraticável dependendo da máquina disponível, pois a geração de pseudo-valores requer um grande esforço computacional e pode demandar muito espaço em disco. Para esse método também, até onde se sabe, não existe uma técnica de avaliação da qualidade do ajuste bem fundamentada e difundida.

Nesse trabalho, os modelos foram comparados empiricamente via simulações e, dentre as alternativas testadas, se for praticável a geração de pseudo-valores, recomenda-se o uso dessa técnica para a modelagem de dados com fração de cura e riscos competitivos. Essa técnica se mostrou mais flexível e em geral, acomodou bem a diferentes curvas de incidência acumulada testadas. Se pela análise descritiva a forma das funções de incidência acumulada for semelhante a forma da Gompertz, espera-se que a escolha entre as técnicas seja indiferente, sendo que os métodos paramétricos oferecem vantagens como a extrapolação para pontos não presentes nos dados e, para a maioria dos casos testados, oferecem estimativas menos viesadas nos valores de tempo maiores quando comparado à técnica baseada nos pseudo-valores. Enquanto não forem estabelecidas técnicas para a comparação dos ajustes, a prioridade do pesquisador por tempos maiores ou menores e a análise descritiva dos dados pode, então, fornecer uma ideia inicial de qual técnica melhor atenderá o estudo.

Além de existirem poucos modelos propostos para essa combinação de fração de cura e riscos competitivos, ainda são pontos a desenvolver como estabelecer uma estatística que ajude a decidir sobre qual o melhor modelo, métodos de diagnóstico para a qualidade do ajuste realizado, como comparar modelos paramétricos e não paramétricos, e até mesmo, como testar a suficiência da extensão do seguimento ou desenvolver técnicas para estabelecê-la nesse contexto. Assim, como próximos passos para esse trabalho, o modelo e a técnica de estimação propostas em Choi et al. (2015) podem ser explorados e comparados aos modelos aqui apresentados, e outro ponto muito interessante, é pesquisar e desenvolver métodos para avaliar a qualidade dos ajustes nesses modelos.

Apêndice A

Principais Resultados da Simulação

Nessa seção o leitor pode encontrar a comparação entre os EQMs obtidos para cada um dos métodos descritos nesse trabalho para todos os cenários com N = 800.

A.1 Cenários Gerados pela Distribuição Gompertz

As figuras apresentadas a seguir estão ordenadas por estrato, e em cada estrato, por risco. Em cada um delas, é possível comparar os valores de EQM obtidos para cada um dos modelos, para os cenários com N = 800, gerados pela distribuição Gompertz, para os quatro níveis de censura e três níveis de cura testados.

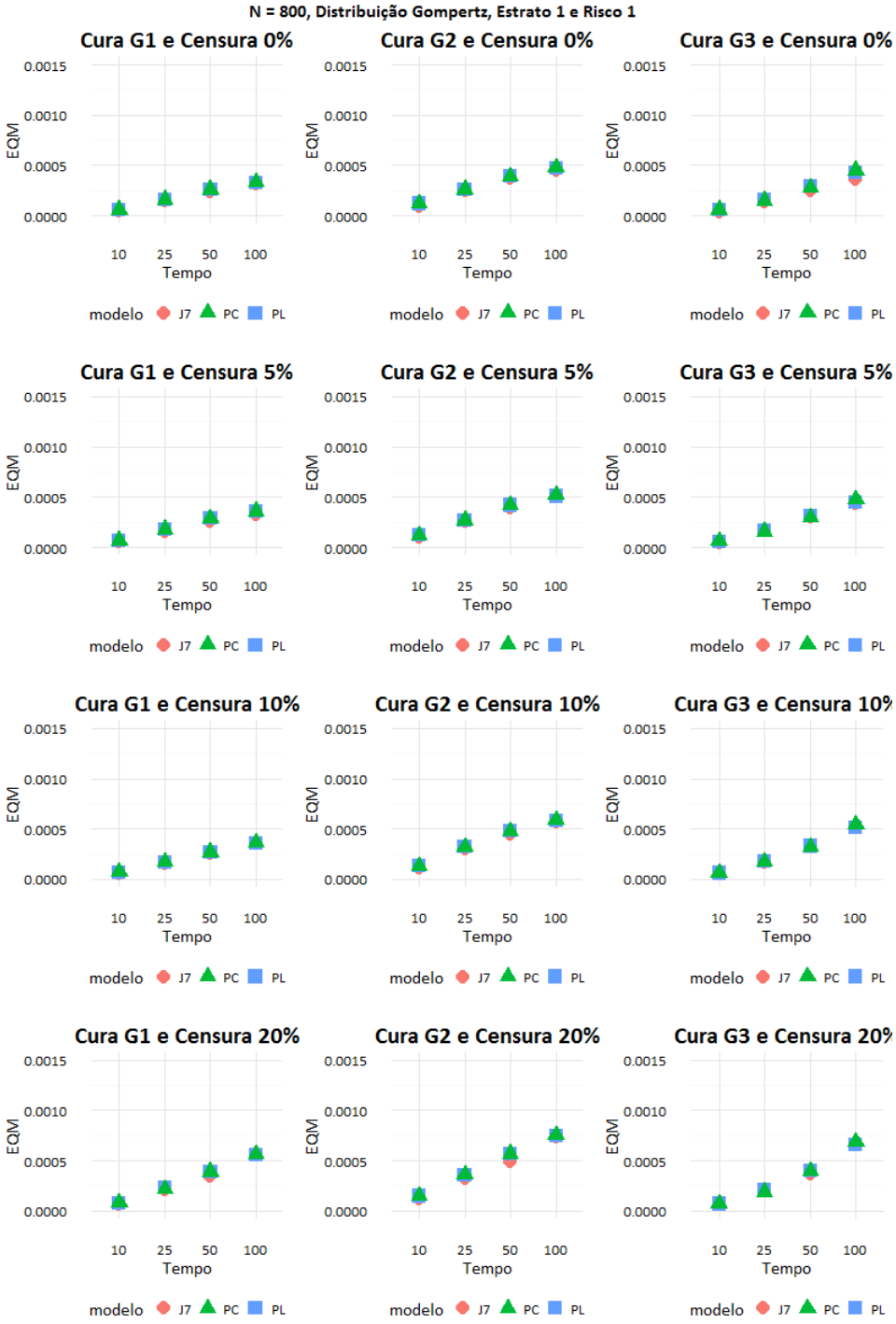


Figura A.1: Comparação entre os valores de EQM para o estrato 1, risco 1 nos cenários gerados pela Gompertz com N = 800.

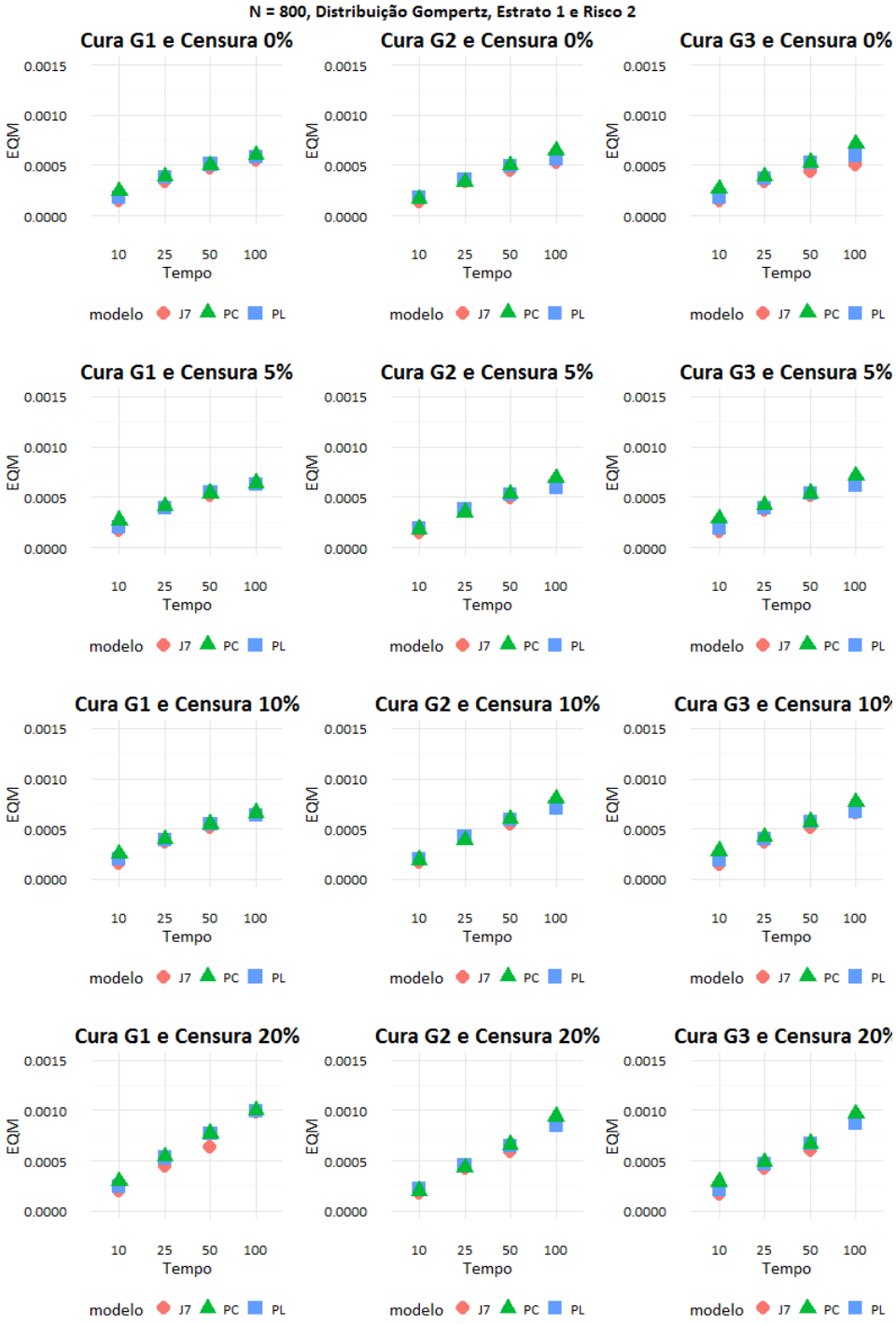


Figura A.2: Comparação entre os valores de EQM para o estrato 1, risco 2 nos cenários gerados pela Gompertz com N = 800.

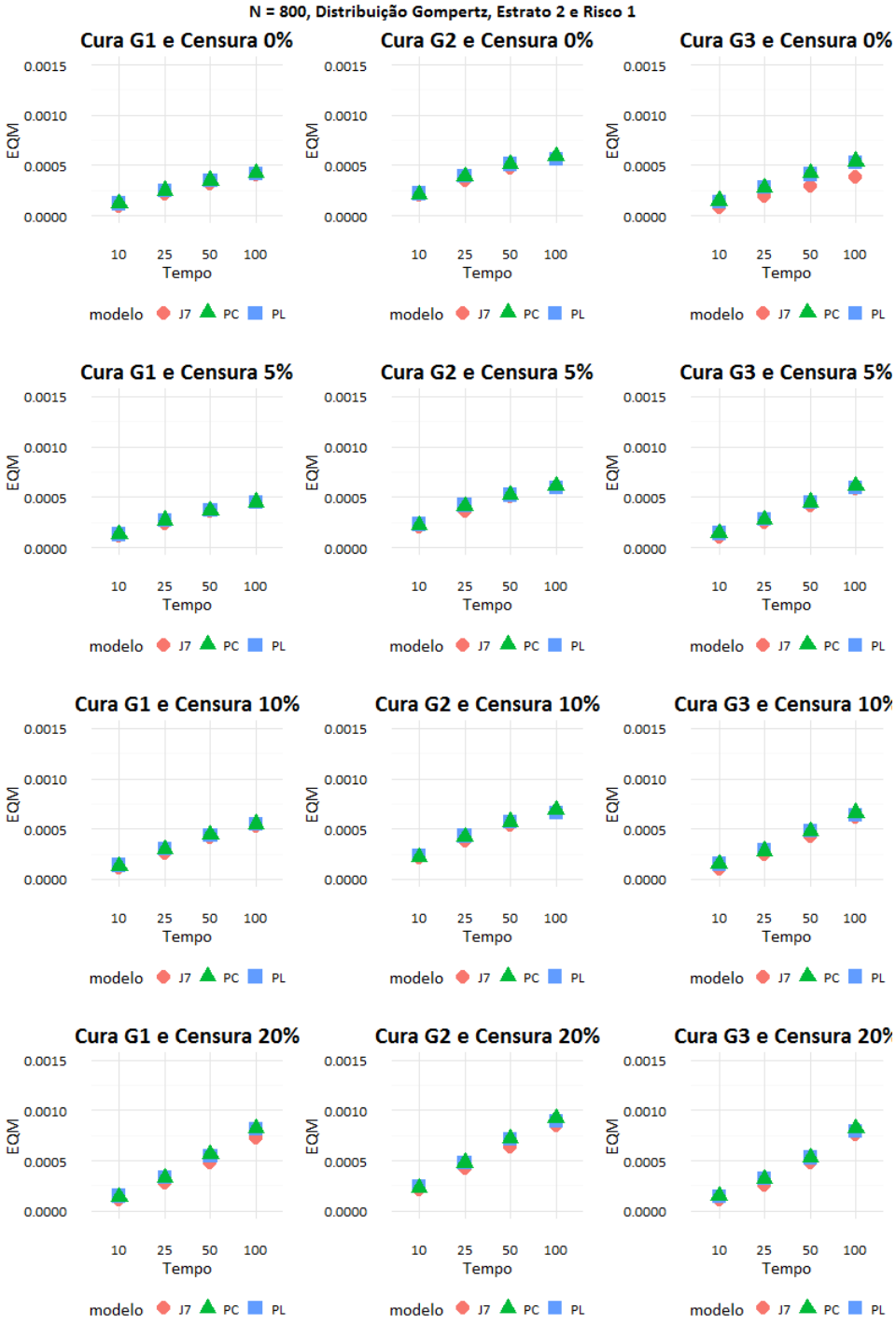


Figura A.3: Comparação entre os valores de EQM para o estrato 2, risco 1 nos cenários gerados pela Gompertz com N = 800.

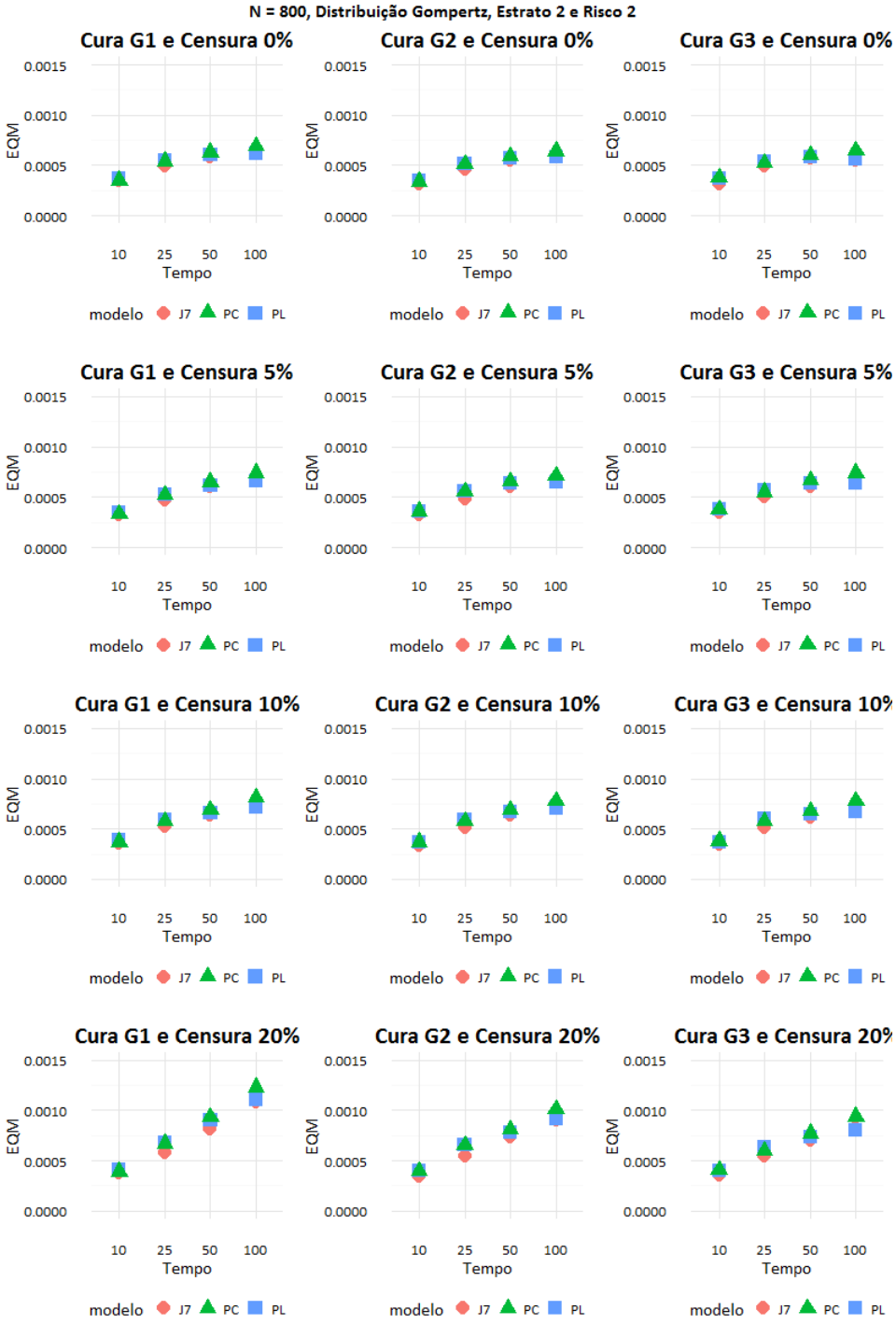


Figura A.4: Comparação entre os valores de EQM para o estrato 2, risco 2 nos cenários gerados pela Gompertz com N = 800.

A.2 Cenários Gerados pela Distribuição Weibull

As figuras apresentadas a seguir estão ordenadas por estrato, e em cada estrato, por risco. Em cada um delas, é possível comparar os valores de EQM obtidos para cada um dos modelos, para os cenários com N = 800, gerados pela distribuição Weibull, para os quatro níveis de censura e três níveis de cura testados.

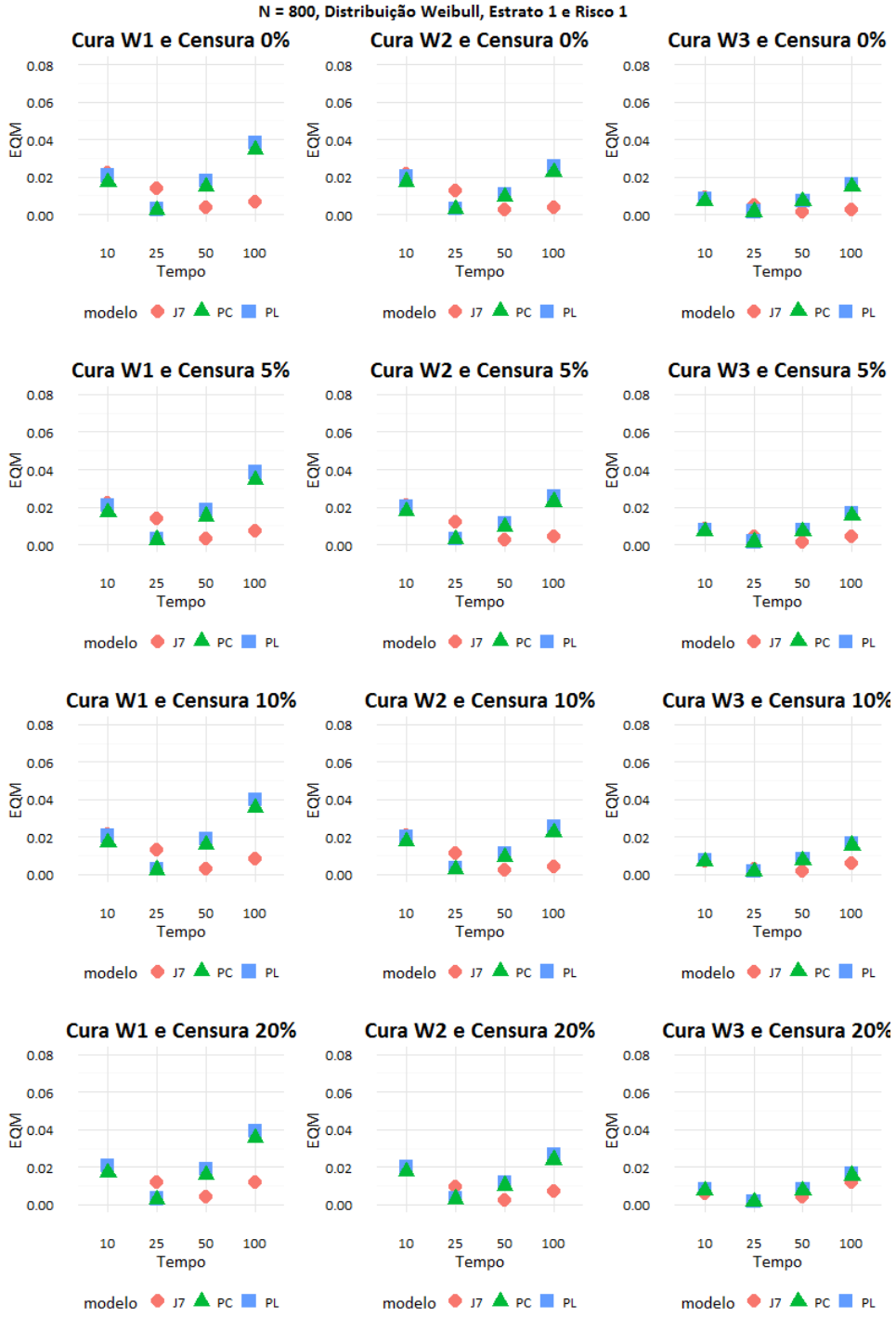


Figura A.5: Comparação entre os valores de EQM para o estrato 1, risco 1 nos cenários gerados pela Weibull com N = 800.

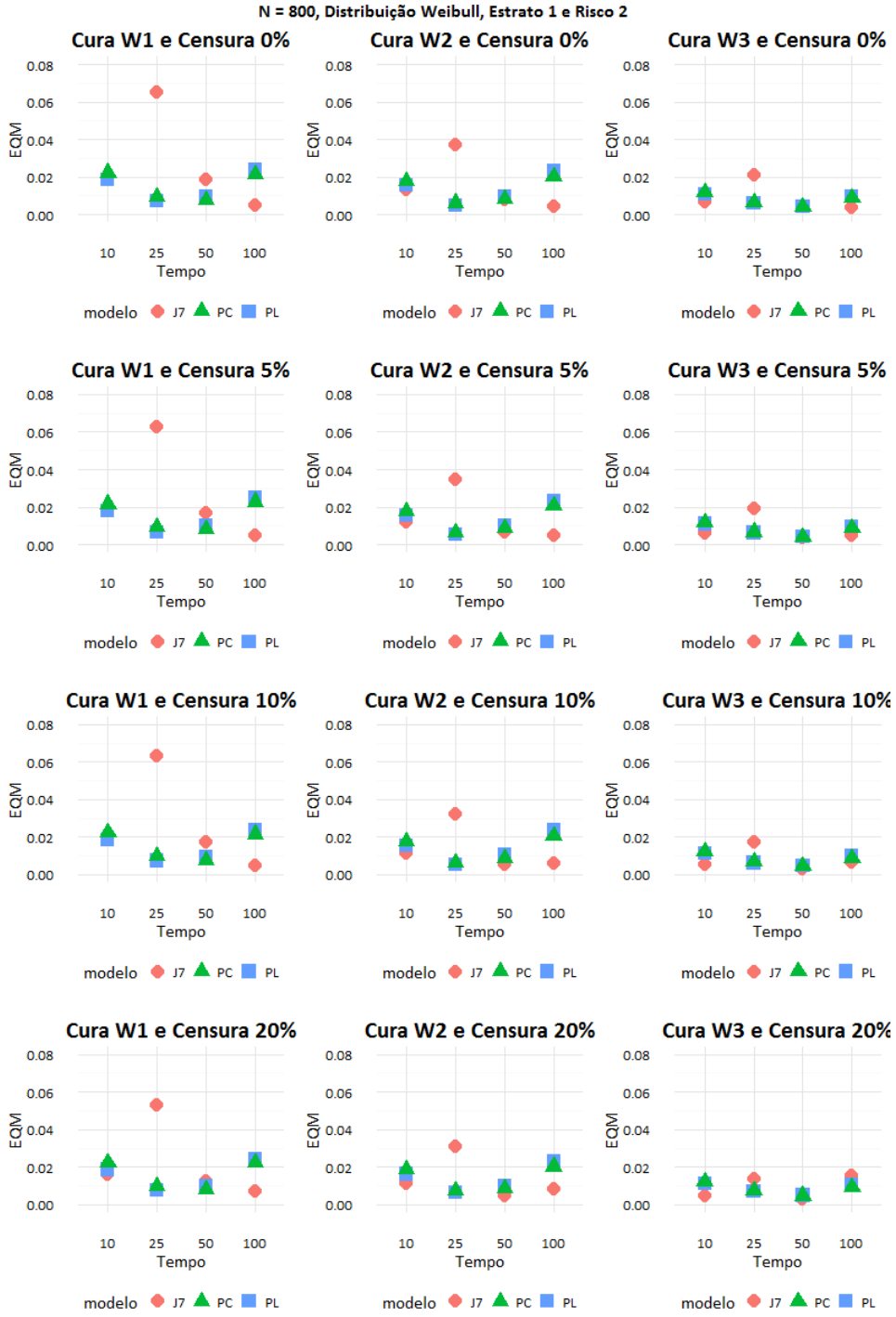


Figura A.6: Comparação entre os valores de EQM para o estrato 1, risco 2 nos cenários gerados pela Weibull com N = 800.

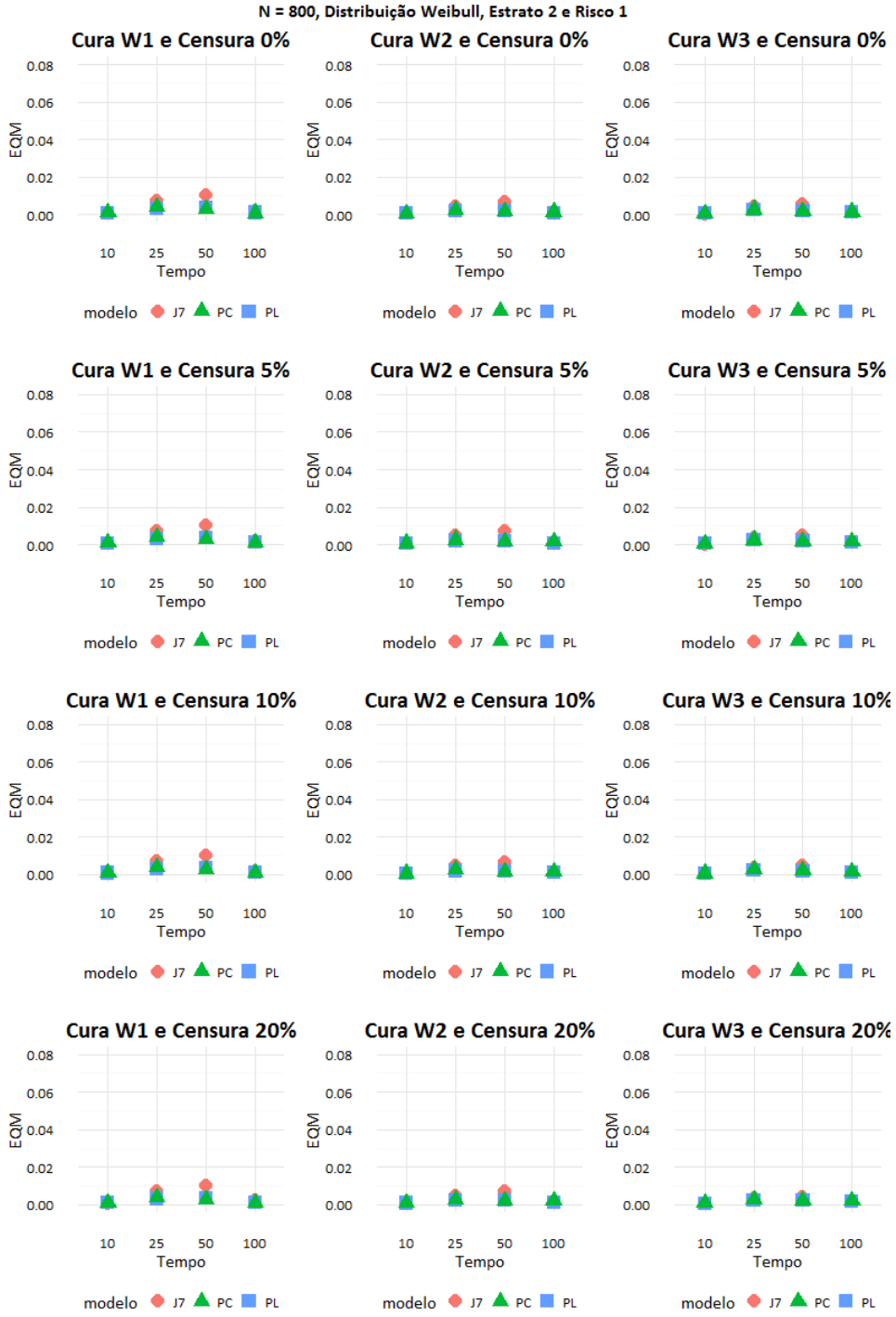


Figura A.7: Comparação entre os valores de EQM para o estrato 2, risco 1 nos cenários gerados pela Weibull com N = 800.

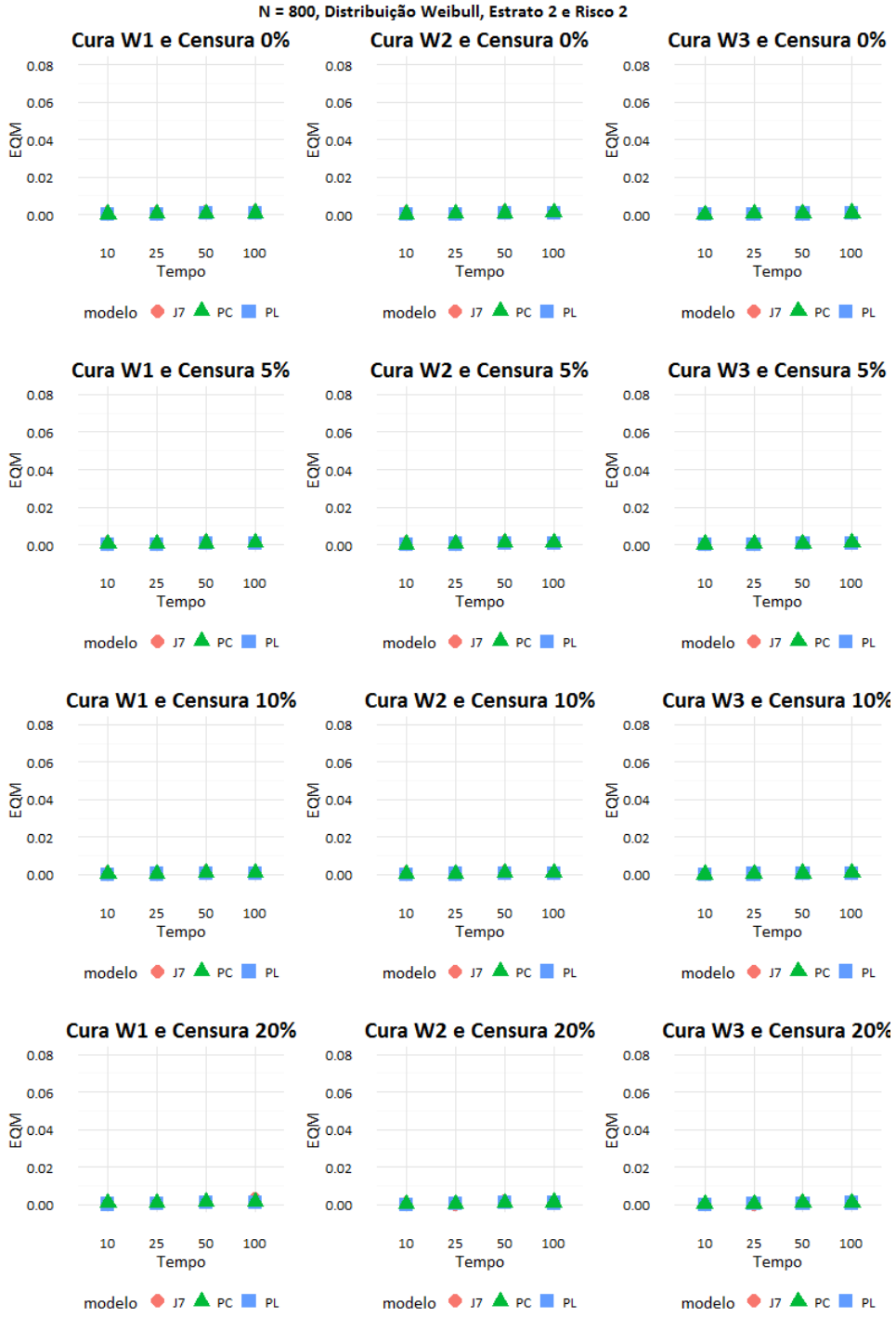


Figura A.8: Comparação entre os valores de EQM para o estrato 2, risco 2 nos cenários gerados pela Weibull com N = 800.

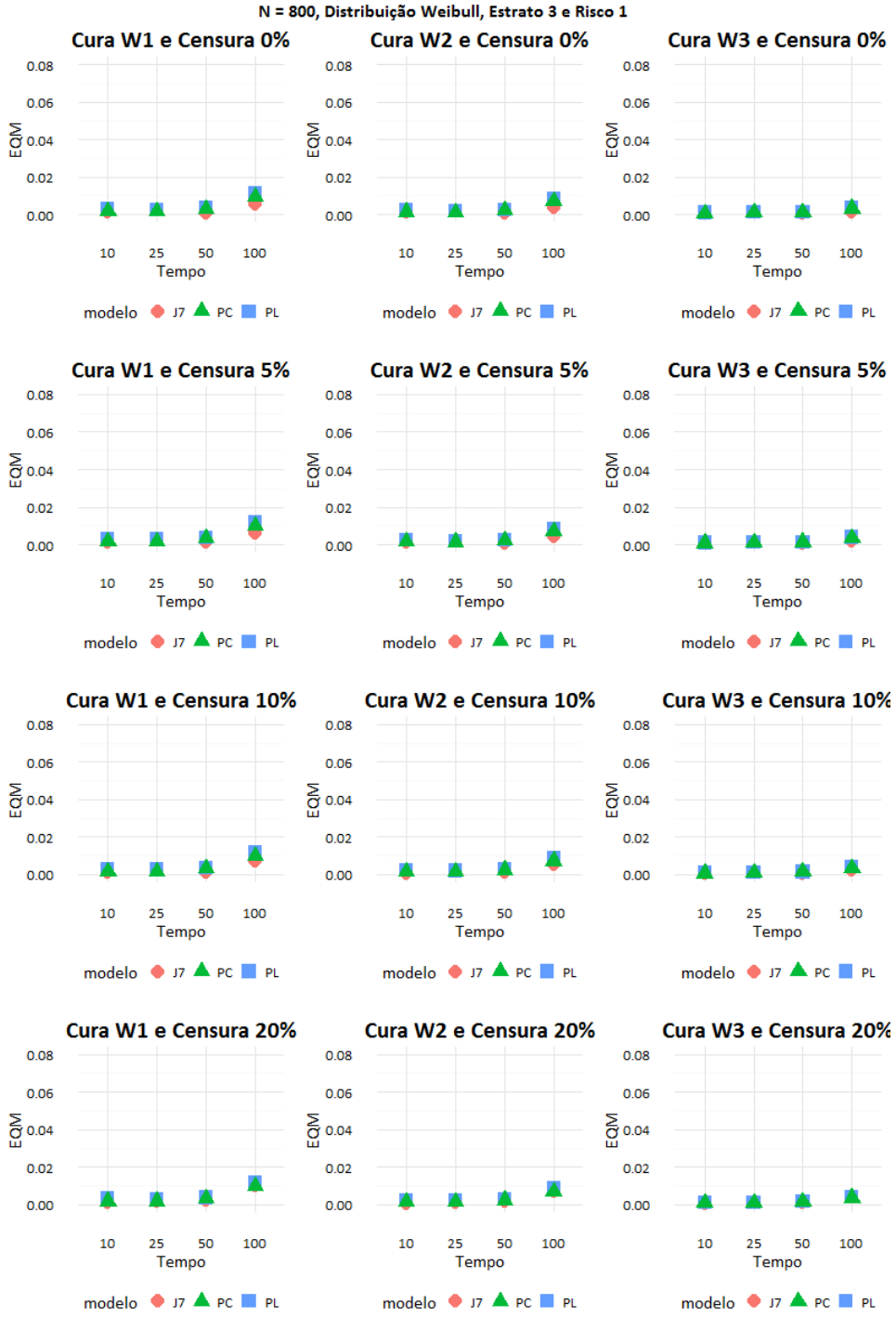


Figura A.9: Comparação entre os valores de EQM para o estrato 3, risco 1 nos cenários gerados pela Weibull com N = 800.

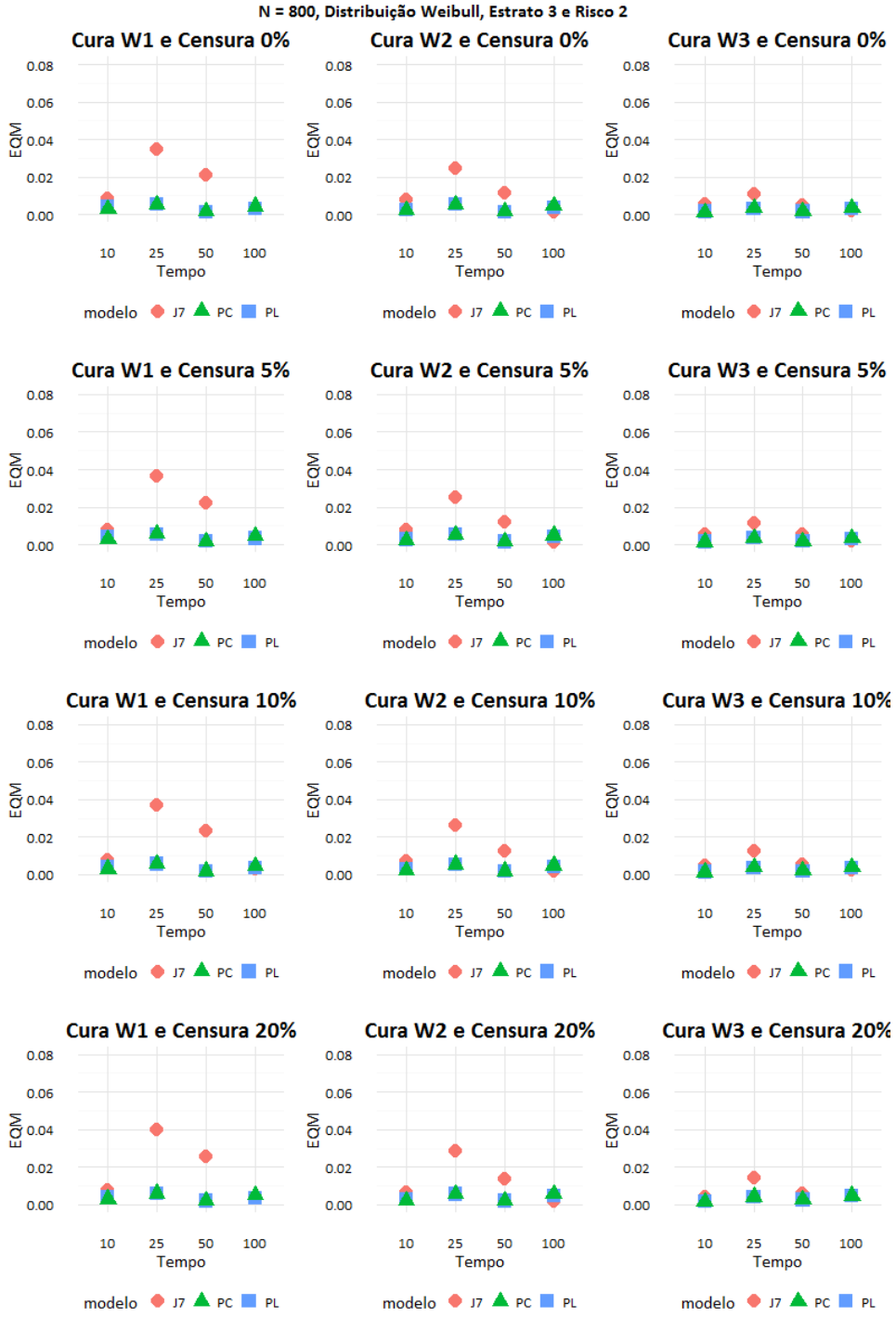


Figura A.10: Comparação entre os valores de EQM para o estrato 3, risco 2 nos cenários gerados pela Weibull com N = 800.

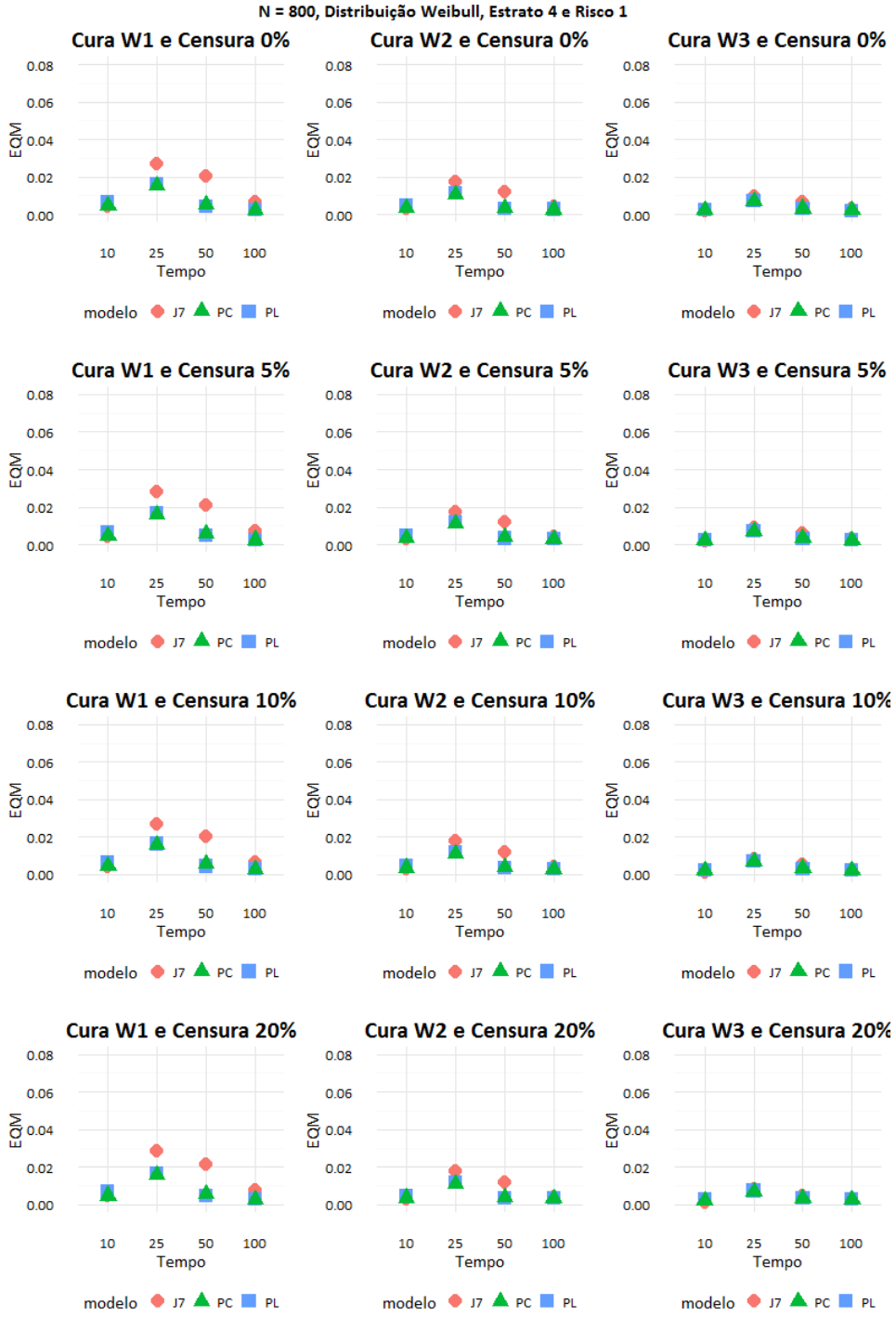


Figura A.11: Comparação entre os valores de EQM para o estrato 4, risco 1 nos cenários gerados pela Weibull com N = 800.

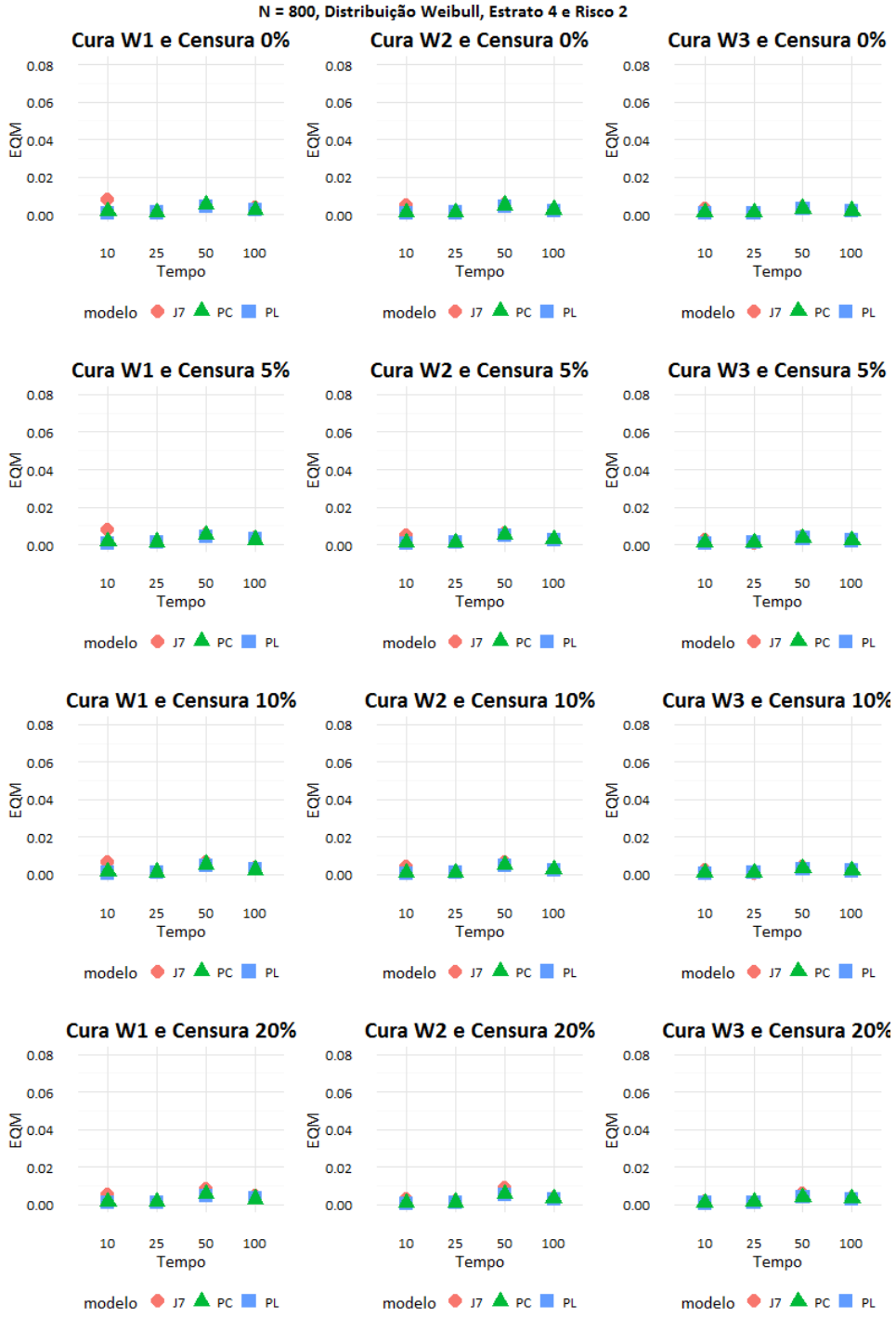


Figura A.12: Comparação entre os valores de EQM para o estrato 4, risco 2 nos cenários gerados pela Weibull com N = 800.

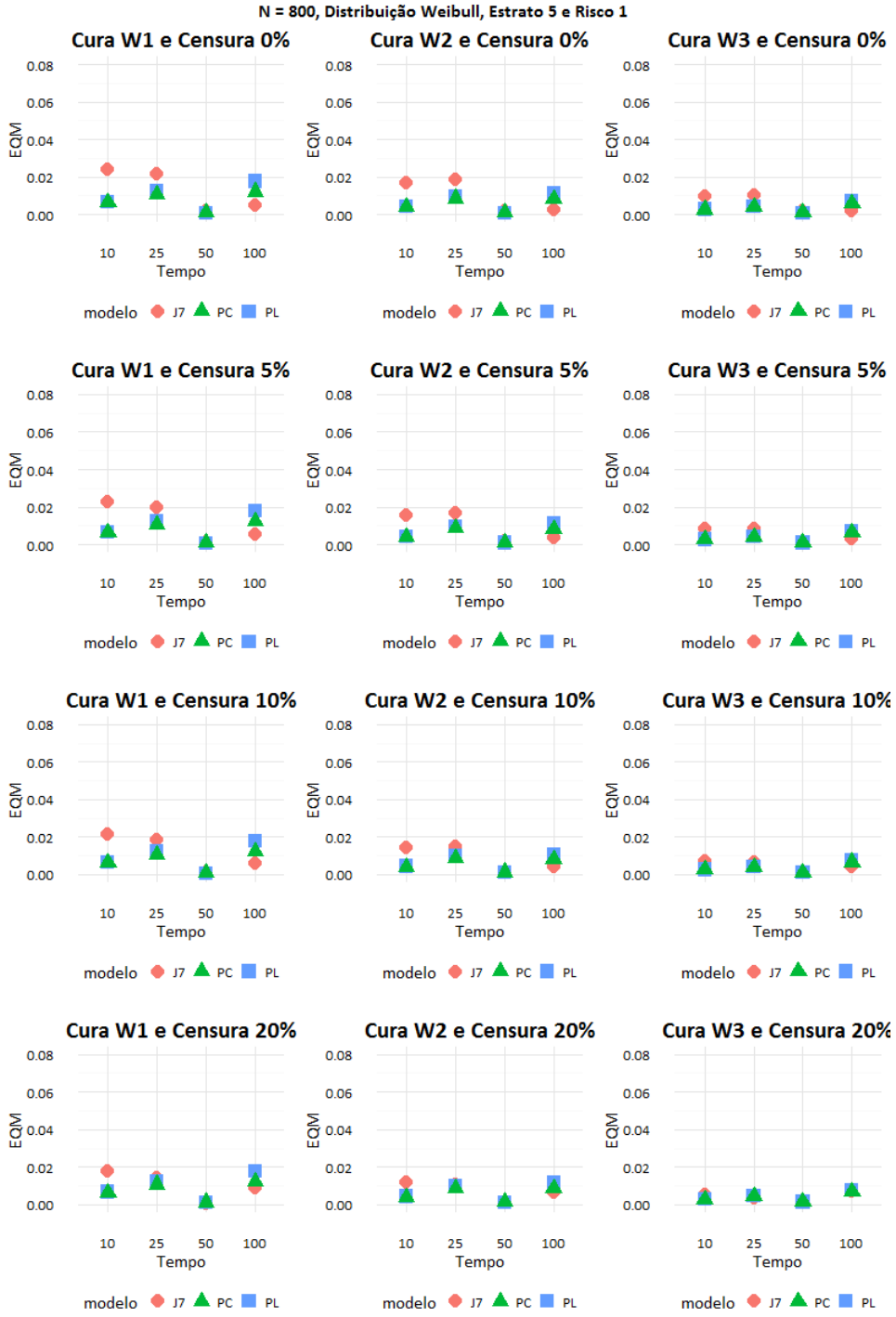


Figura A.13: Comparação entre os valores de EQM para o estrato 5, risco 1 nos cenários gerados pela Weibull com N = 800.

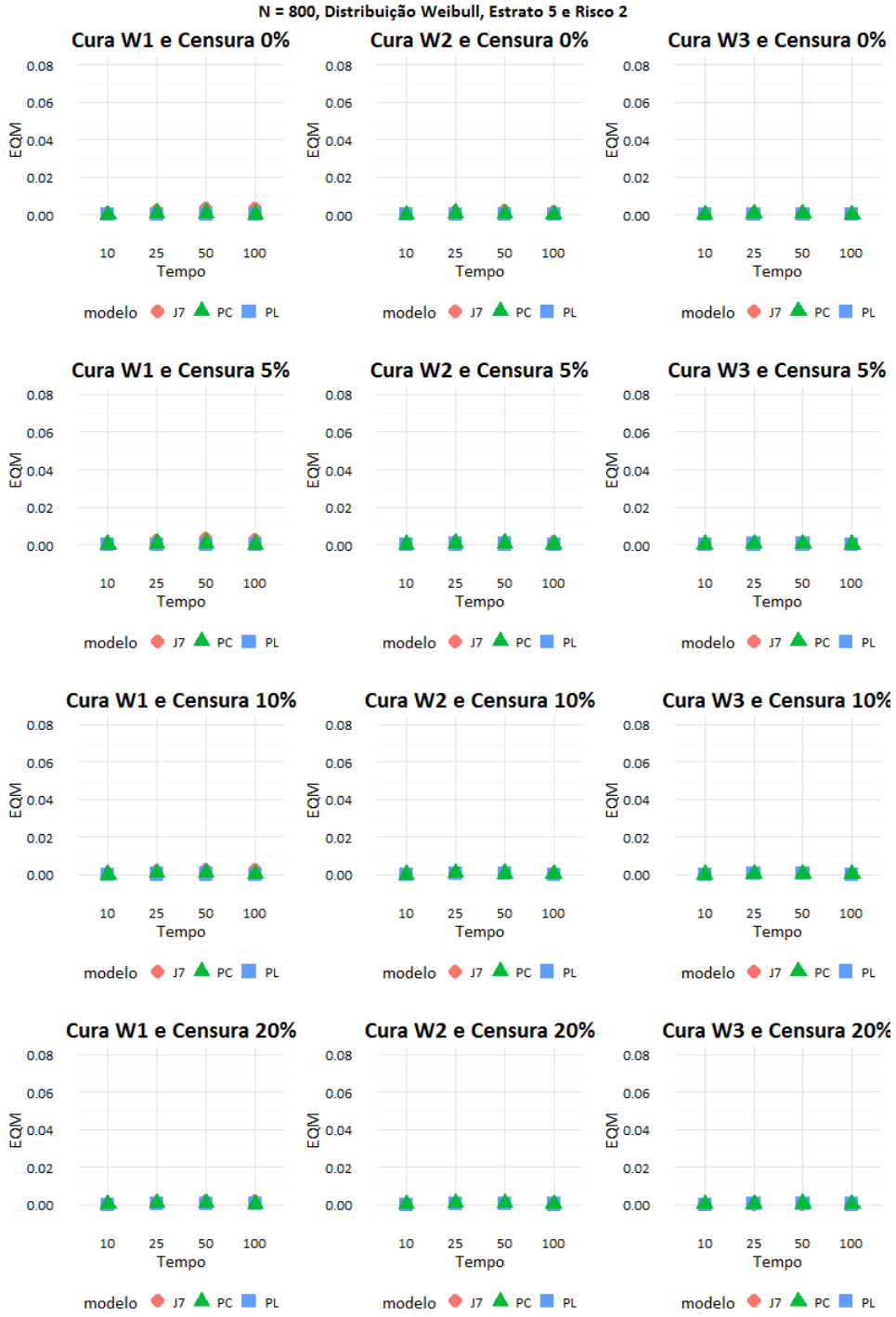


Figura A.14: Comparação entre os valores de EQM para o estrato 5, risco 2 nos cenários gerados pela Weibull com N = 800.

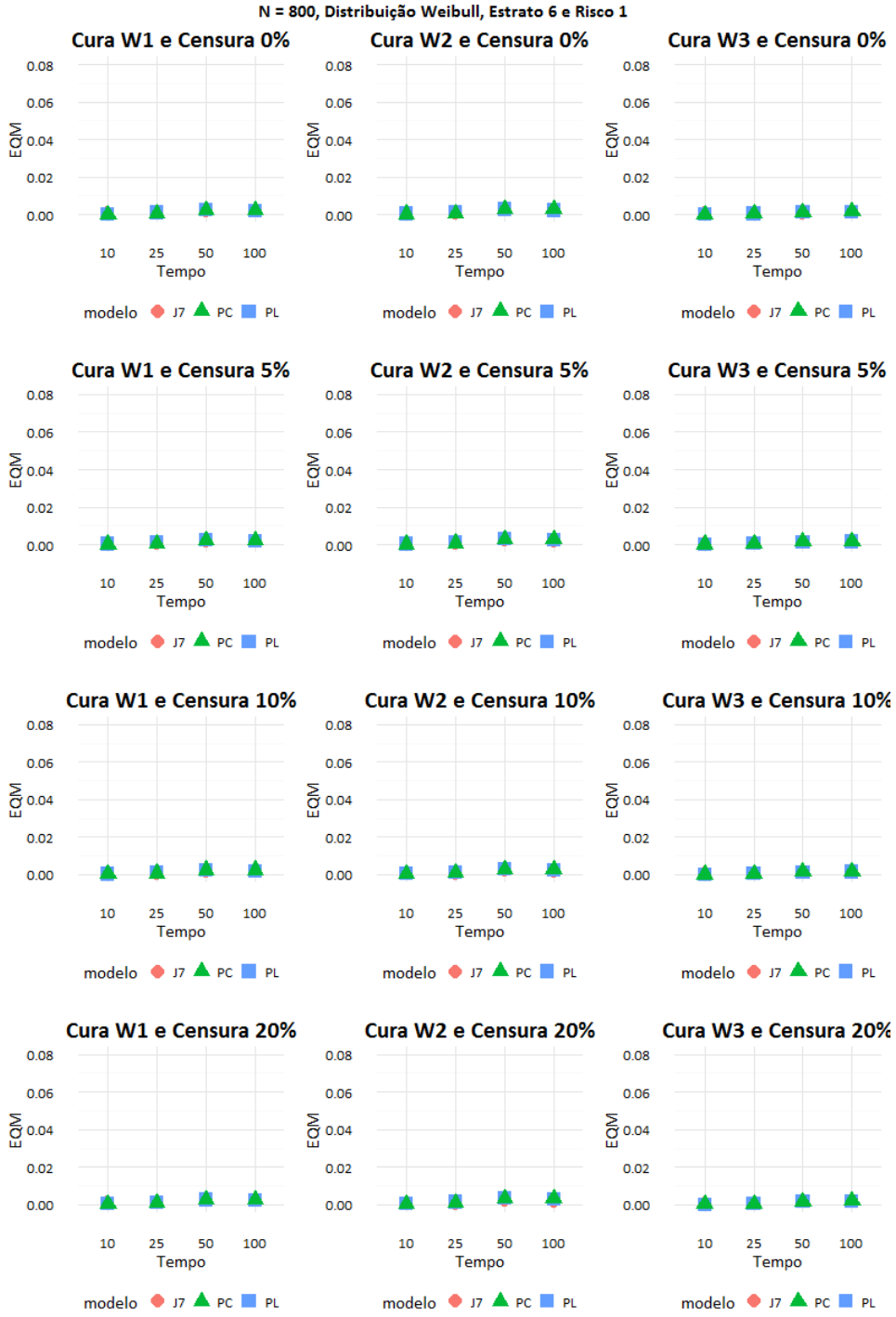


Figura A.15: Comparação entre os valores de EQM para o estrato 6, risco 1 nos cenários gerados pela Weibull com N = 800.

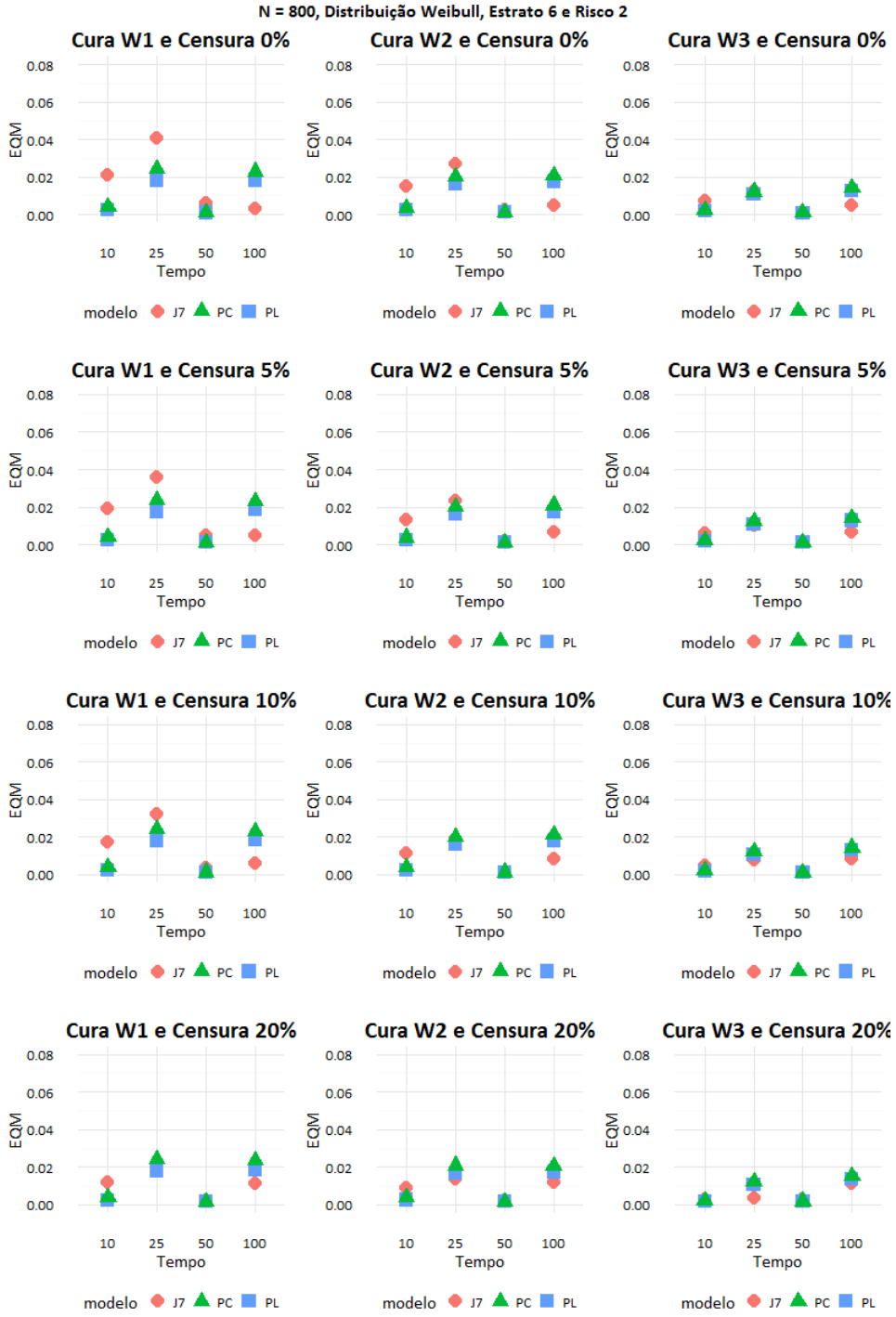


Figura A.16: Comparação entre os valores de EQM para o estrato 6, risco 2 nos cenários gerados pela Weibull com N = 800.

Referências Bibliográficas

Aalen, O. (1975). Statistical inference for a family of counting processes. proquest llc, *Ann Arbor, MI* .

Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution, *The Annals of Applied Probability* pp. 951–972.

Agresti, A. e Kateri, M. (2013). *Categorical data analysis*, Springer.

Andersen, P. K. e Perme, M. P. (2009). Pseudo-observations in survival analysis, *Statistical methods in medical research* .

Basu, S. e Tiwari, R. C. (2010). Breast cancer survival, competing risks and mixture cure model: a bayesian analysis, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(2): 307–329.

Benichou, J. e Gail, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies, *Biometrics* pp. 813–826.

Berkson, J. e Gage, R. P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**(259): 501–515.

Bernoulli, D. (1766). An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it, *Mem. Math. Phys. Acad. Royal Science* **1**.

Beyersmann, J., Allignol, A. e Schumacher, M. (2011a). *Competing risks and multistate models with R*, Springer Science & Business Media.

Beyersmann, J., Allignol, A. e Schumacher, M. (2011b). *Competing risks and multistate models with R*, Springer Science & Business Media.

Beyersmann, J. e Schumacher, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks, *Biostatistics* **9**(4): 765–776.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society. Series B (Methodological)* **11**(1): 15–53.

Bryant, J. e Dignam, J. J. (2004). Semiparametric models for cumulative incidence functions, *Biometrics* **60**(1): 182–190.

Chang, I., Hsiung, C. A., WEN, C.-C., WU, Y.-J., YANG, C.-C. et al. (2007). Non-parametric maximum-likelihood estimation in a semiparametric mixture model for competing-risks data, *Scandinavian Journal of Statistics* **34**(4): 870–895.

Chao, E. C. (1998). Gibbs sampling for long-term survival data with competing risks, *Biometrics* **54**: 350–366.

Chen, M.-H., Ibrahim, J. G. e Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association* **94**(447): 909–919.

Choi, S. e Huang, X. (2014). Maximum likelihood estimation of semiparametric mixture component models for competing risks data, *Biometrics* **70**(3): 588–598.

Choi, S., Huang, X. e Cormier, J. N. (2015). Efficient semiparametric mixture inferences on cure rate models for competing risks, *Canadian Journal of Statistics* **43**(3): 420–435.

Conway, R. W. e Maxwell, W. L. (1962). A queuing model with state dependent service rates, *Journal of Industrial Engineering* **12**(2): 132–136.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure, *Journal of the Royal Statistical Society. Series B (Methodological)* **21**: 411–421.

Crowder, M. J. (2001). *Classical competing risks*, CRC Press.

Crowder, M. J. (2012). *Multivariate survival analysis and competing risks*, CRC Press.

David, C. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)* **34**: 187–220.

Diggle, P. (2002). *Analysis of longitudinal data*, Oxford University Press.

Dixon, S. N., Darlington, G. A. e Desmond, A. F. (2011). A competing risks model for correlated data based on the subdistribution hazard, *Lifetime data analysis* **17**(4): 473–495.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**: 1041–1046.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities, *Biostatistics* **2**(1): 85–97.

Fine, J. P. e Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American statistical association* **94**(446): 496–509.

Gail, M. (1975). A review and critique of some models used in competing risk analysis, *Biometrics* **31**: 209–222.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, *Philosophical transactions of the Royal Society of London* pp. 513–583.

Graw, F., Gerds, T. A. e Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models, *Lifetime Data Analysis* **15**(2): 241–255.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of statistics* **16**: 1141–1154.

Greenhouse, J. B. e Wolfe, R. A. (1984). A competing risks derivation of a mixture model for the analysis of survival data, *Communications in Statistics-Theory and Methods* **13**(25): 3133–3154.

Hardin, J. W. (2005). *Generalized estimating equations (GEE)*, Wiley Online Library.

Haybittle, J. (1965). A two-parameter model for the survival curve of treated cancer patients, *Journal of the American Statistical Association* **60**(309): 16–26.

Ibrahim, J. G., Chen, M.-H. e Sinha, D. (2005a). *Bayesian survival analysis*, Wiley Online Library.

Ibrahim, J. G., Chen, M.-H. e Sinha, D. (2005b). *Bayesian survival analysis*, Wiley Online Library.

Jeong, J.-H. e Fine, J. (2006). Direct parametric inference for the cumulative incidence function, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55**(2): 187–200.

Jeong, J.-H. e Fine, J. P. (2007). Parametric regression on cumulative incidence function, *Biostatistics* **8**(2): 184–196.

Kalbfleisch, J. e Prentice, R. (2002). *The Statistical Analysis of Failure Time Data.*, Wiley-Interscience, Hoboken, NJ.

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American statistical association* **53**(282): 457–481.

Katsahian, S. e Boudreau, C. (2011). Estimating and testing for center effects in competing risks, *Statistics in medicine* **30**(13): 1608–1617.

Katsahian, S., Resche-Rigon, M., Chevret, S. e Porcher, R. (2006). Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution, *Statistics in medicine* **25**(24): 4267–4278.

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

Klein, J. P. e Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics* **61**(1): 223–229.

Klein, J. P., Gerster, M., Andersen, P. K., Tarima, S. e Perme, M. P. (2008). Sas and r functions to compute pseudo-values for censored data regression, *Computer methods and programs in biomedicine* **89**(3): 289–300.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. e Scheike, T. H. (2016). *Handbook of survival analysis*, CRC Press.

Korn, E. L. e Dorey, F. J. (1992). Applications of crude incidence curves, *Statistics in medicine* **11**(6): 813–829.

Kuk, A. Y. e Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika* **79**(3): 531–541.

Larson, M. G. e Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data, *Applied statistics* pp. 201–211.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, Vol. 362, John Wiley & Sons.

Li, C.-S., Taylor, J. M. e Sy, J. P. (2001). Identifiability of cure models, *Statistics & Probability Letters* **54**(4): 389–395.

Liang, K.-Y. e Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.

Lu, W. e Peng, L. (2008). Semiparametric analysis of mixture regression models with competing risks data, *Lifetime data analysis* **14**(3): 231–252.

Maller, R. A. e Zhou, X. (1996). *Survival analysis with long-term survivors*, Wiley New York.

Maller, R. A. e Zhou, X. (2002). Analysis of parametric models for competing risks, *Statistica Sinica* **12**: 725–750.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models*, Vol. 37, CRC press.

Peng, Y. (2003). Fitting semiparametric cure models, *Computational statistics & data analysis* **41**(3): 481–490.

Peng, Y. e Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation, *Biometrics* **56**(1): 237–243.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies, *Journal of the American Statistical Association* **86**(415): 770–778.

Pepe, M. S. e Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data?, *Statistics in medicine* **12**(8): 737–751.

Pintilie, M. (2006). *Competing risks: a practical perspective*, Vol. 58, John Wiley & Sons.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. e Breslow, N. (1978). The analysis of failure times in the presence of competing risks, *Biometrics* **34**: 541–554.

Rodrigues, J., de Castro, M., Balakrishnan, N. e Cancho, V. G. (2011). Destructive weighted Poisson cure rate models, *Lifetime data analysis* **17**(3): 333–346.

Rodrigues, J., de Castro, M., Cancho, V. G. e Balakrishnan, N. (2009). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data, *Journal of Statistical Planning and Inference* **139**(10): 3605–3611.

Scheike, T. H., Sun, Y., Zhang, M.-J. e Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data, *Biometrika* **97**(1): 133–145.

Scheike, T. H. e Zhang, M.-J. (2007). Direct modelling of regression effects for transition probabilities in multistate models, *Scandinavian Journal of Statistics* **34**(1): 17–32.

Scheike, T. H., Zhang, M.-J. e Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression, *Biometrika* **95**(1): 205–220.

Sy, J. P. e Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model, *Biometrics* **56**(1): 227–236.

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics* **51**: 899–907.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**(1): 20–22.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors, *Biometrics* **54**: 1508–1516.

Yakovlev, A. Y., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A. e Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, *Biometrie et analyse de donnees spatio-temporelles* **12**: 66–82.

Yin, G. e Ibrahim, J. G. (2005a). Cure rate models: a unified approach, *Canadian Journal of Statistics* **33**(4): 559–570.

Yin, G. e Ibrahim, J. G. (2005b). A general class of bayesian survival models with zero and nonzero cure fractions, *Biometrics* **61**(2): 403–412.