Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

Análise de Sobrevivência com Fração de Cura no contexto de seguros de vida inteira: uma perspectiva atuarial

Milena Doarte da Rocha
Yasmin Santana da Silva

Rio de Janeiro
2025

CIP - Catalogação na Publicação

R672a

Rocha, Milena Doarte da
Análise de sobrevivência com fração de cura no contexto de seguros de vida inteira: uma perspectiva atuarial / Milena Doarte da Rocha ; Yasmin Santana da Silva. --Rio de Janeiro, 2025.
60 f.

Orientador: Viviana das Graças Ribeiro Lobo.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciências atuariais, 2025.

1. Análise de sobrevivência. 2. Fração de cura. 3. Seguros de vida inteira. 4. Inferência bayesiana. I. Silva, Yasmin Santana da. II. Lobo, Viviana das Graças Ribeiro , orient. III. Título.

Fabio Osmar de Oliveira Maciel – CRB-7 6284

Análise de Sobrevivência com Fração de Cura no contexto de seguros de vida inteira: uma perspectiva atuarial

Milena Doarte da Rocha
Yasmin Santana da Silva

Trabalho de conclusão de curso em Ciências Atuariais do Departamento de Métodos Estatísticos do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Bacharel em Ciências Atuariais.

Aprovado por:

Profª. Drª. Viviana das Graças Ribeiro Lobo
IM - UFRJ - Orientadora.

Profª. PhD. Thaís Cristina Oliveira da Fonseca
IM - UFRJ.

Prof. Dr. João Batista de Morais Pereira
IM - UFRJ.

Rio de Janeiro
2025

Agradecimentos

Milena Doarte da Rocha

Primeiramente, agradeço aos meus pais, Marlene e Marco Antônio, por tudo o que sempre fizeram por mim, pelo amor, pelo apoio e por estarem ao meu lado em cada passo desta caminhada. Sou imensamente grata.

Também sou grata ao meu irmão Marcos e à minha cunhada Bárbara pelo apoio constante, e aos meus sogros, Iara e José Nilson, pelo acolhimento e incentivo ao longo desta jornada.

Ao meu namorado Marcus, obrigada por ser meu porto seguro, pelo carinho, pela paciência e por estar sempre ao meu lado. Sou muito grata por me incentivar com tanto amor a seguir em frente, mesmo quando o caminho é difícil.

À professora Viviana Lobo, minha orientadora, agradeço pela confiança, paciência e disponibilidade em cada etapa do processo. Sua orientação foi fundamental para a realização deste projeto.

Agradeço à minha amiga Yasmin pela amizade, pelo companheirismo e por tornar cada momento mais leve e especial.

Aos familiares e amigos que me acompanharam ao longo desta trajetória, agradeço profundamente pela amizade, pelas palavras de incentivo, pelos gestos de apoio e por todas as demonstrações de afeto que fizeram toda a diferença nesta caminhada.

Sou grata aos professores Thaís Fonseca e João Batista por aceitarem compor a banca.

Ao LabMA, obrigada por ter sido um espaço tão importante de aprendizado, troca e crescimento.

Agradeço aos professores da UFRJ pelo conhecimento compartilhado ao longo do curso, que foi essencial para minha formação.

Por fim, deixo meu sincero agradecimento a todos que, de alguma forma, contribuíram para esta etapa tão importante da minha vida.

Yasmin Santana da Silva

A conclusão deste projeto e da minha graduação só foi possível graças à base sólida que recebi da minha família. Aos meus pais, Leda Paula e Ronaldo Francisco, agradeço por todo o amor, dedicação e pelos valores que sempre cultivaram em mim. Desde os primeiros anos de escola até os desafios da universidade, esses valores me sustentaram.

Sou imensamente grata à minha orientadora, professora Viviana das Graças Ribeiro Lobo, por seu apoio, paciência e confiança desde o início deste trabalho, disponibilizando-se sempre a esclarecer os principais conceitos estatísticos e atuariais. Sua orientação foi essencial para o desenvolvimento deste estudo.

Um agradecimento aos professores João Batista e Thaís Oliveira, que concordaram em participar da banca do nosso projeto.

Aos amigos que caminharam comigo nesta trajetória, especialmente Milena Doarte da Rocha, obrigada pelo companheirismo, apoio e incentivo durante os períodos mais desafiadores da graduação. Cada noite de estudo, trabalho e desabafo fortaleceu minha jornada.

Agradeço também ao LabMA, onde tive a oportunidade de aplicar na prática os conhecimentos adquiridos ao longo da graduação. Da mesma forma, sou grata à Mongeral Aegon pela continuidade desse aprendizado no ambiente corporativo. Agradeço ainda aos meus líderes, Carlos Américo e Nelson Emiliano, por todo o conhecimento compartilhado e pela confiança depositada em meu trabalho.

Por fim, gostaria de agradecer a Deus. Por sua graça e misericórdia, tive forças para seguir até aqui. Tudo o que conquistei é, antes de tudo, resultado da Sua vontade. A Ele dedico este projeto com profunda gratidão.

Resumo

Este projeto tem como objetivo introduzir e aplicar modelos de sobrevivência com fração de cura no contexto atuarial, com foco específico em seguros de vida inteira. A metodologia adotada baseia-se em modelos paramétricos sob abordagem Bayesiana, com implementação via amostragem de Monte Carlo Hamiltoniano utilizando a linguagem `Stan`. Na análise de dados são utilizadas duas bases, sendo uma simulada, desenvolvida para representar um cenário ideal com uma fração de curados claramente definida, e outra referente a seguros de vida inteira, que incorpora características observadas no mercado atuarial. Os modelos ajustados incluem versões com e sem fração de cura, além de variações com diferentes especificações para covariáveis associadas à cura. Os resultados são avaliados por meio de critérios de informação e pela análise das curvas ajustadas. Como aplicação prática, busca-se compreender o comportamento da persistência de contratos e identificar segmentos com maior propensão à manutenção do vínculo. O projeto contribui para o aprimoramento da modelagem de curvas de persistência em um cenário de seguros.

Palavras-chave: análise de sobrevivência; fração de cura; seguros de vida inteira; inferência Bayesiana; atuária.

Abstract

This project aims to introduce and apply cure rate survival models in the actuarial context, with a specific focus on whole life insurance. The adopted methodology is based on parametric models under a Bayesian approach, implemented via Hamiltonian Monte Carlo sampling using the `Stan` programming language. The data analysis uses two datasets: one simulated, developed to represent an ideal scenario with a clearly defined cured fraction, and another based on whole life insurance data, incorporating characteristics observed in the actuarial market. The fitted models include versions with and without a cure fraction, as well as variations with different specifications for covariates associated with cure. The results are evaluated using information criteria and by analyzing the fitted survival curves. As a practical application, the goal is to understand contract persistence behavior and to identify segments with a higher propensity to maintain the contractual relationship. This project contributes to the enhancement of persistence curve modeling in a differentiated setting and proposes future directions within an insurance context.

Keywords: survival analysis; cure fraction; whole life insurance; Bayesian inference; actuarial science.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

A análise de sobrevivência é uma vertente estatística amplamente utilizada para estudar o tempo até a ocorrência de um evento de interesse, como a morte, a falência ou a recuperação de uma doença. Embora tradicionalmente aplicada em contextos médicos, epidemiológicos e de confiabilidade, como discutido por Kleinbaum e Klein (2011) e Lawless (2022), essa metodologia também desempenha um papel fundamental nas ciências atuariais, conforme apresentado por Richards (2011), especialmente no segmento de seguros de vida.

No contexto atuarial, exemplos relevantes incluem o tempo até a entrada em invalidez, o tempo até a aposentadoria ou, ainda, o tempo até o cancelamento de uma apólice de seguro. Em particular, o tempo até o cancelamento de uma apólice é uma variável de interesse central, pois influencia diretamente o cálculo de prêmios, a constituição de provisões técnicas, a mensuração da persistência e diversas outras métricas essenciais à gestão atuarial.

É importante destacar que, sobretudo do ponto de vista atuarial, é fundamental incluir no estudo de persistência tanto os indivíduos cujas apólices permanecem ativas quanto aqueles que deixaram o portfólio em decorrência de um sinistro. Esses casos são tratados como dados censurados, uma característica marcante nesse tipo de análise, conforme apresentado por Lobo, Fonseca e Alves (2024). A presença de observações incompletas, nesse contexto, costuma ser significativamente elevada, em contraste com muitos estudos médicos, nos quais os eventos de interesse tendem a ocorrer em prazos mais curtos, resultando em menor proporção de censuras.

Partindo dos modelos tradicionais desenvolvidos no âmbito da análise de sobrevivência, conforme discutido por Colosimo e Giolo (2006) e Carvalho (2011), os modelos de mistura com fração de cura têm ganhado destaque por oferecerem uma estrutura mais flexível e realista para lidar com situações em que existe uma parcela da população que nunca experimentará o evento de interesse, mesmo após longos períodos de acompanhamento. Essa parcela é denominada fração de curados.

Originalmente adotado no contexto biomédico, como em estudos clínicos envolvendo pacientes oncológicos, onde uma fração pode ser considerada efetivamente curada e, portanto, não mais suscetível à recaída assim como descrito por Maller e Zhou (1996), esses modelos vêm sendo progressivamente adaptados para outros domínios, dada sua capacidade de capturar estruturas populacionais com diferentes riscos e comportamentos ao longo do tempo.

Sob a ótica atuarial, especialmente no contexto de seguros de vida inteira, a noção de "cura" adquire uma nova interpretação. Embora o segurado esteja teoricamente co-

berto até o falecimento, conforme definido por Bowers et al. (1997), na prática é comum que o contrato seja encerrado antes disso, por meio do resgate da reserva constituída. Dessa forma, ao reformular a análise com foco nesses eventos como desfechos de interesse, a fração de cura pode ser reinterpretada como representando o grupo de segurados que permanece com suas apólices ativas e não realiza o resgate ao longo de todo o período de observação.

No contexto da persistência em seguros de vida, a "cura" pode ser entendida como uma característica latente de certos segurados que, devido ao comportamento financeiro, perfil demográfico, planejamento de longo prazo ou mesmo vínculo com o contrato, apresentam uma tendência estrutural a manter a apólice até seu vencimento natural. Esses indivíduos podem incluir, por exemplo, segurados com baixa propensão ao consumo imediato, maior compreensão do funcionamento do produto ou que atribuem valor simbólico e afetivo à cobertura contratada.

Adaptar os modelos com fração de cura, também abordados por Amico e Keilegom (2018) a esse contexto diferenciado oferece uma perspectiva promissora para a modelagem atuarial, em especial no que se refere à análise mais realista da persistência de contratos, à identificação de grupos com maior ou menor propensão ao cancelamento e ao ajuste mais preciso de curvas de sobrevivência e de risco. Tal abordagem permite uma compreensão mais refinada do comportamento dos segurados ao longo do tempo, contribuindo para o desenvolvimento de produtos mais adequados, políticas de precificação eficazes e estratégias de retenção direcionadas.

Além disso, a incorporação de uma estrutura de mistura com fração de cura aos modelos de sobrevivência traz ganhos importantes em termos de acurácia das projeções atuariais, diluindo a suposição tradicional de que todos os segurados eventualmente experimentarão o evento de interesse. Em contextos marcados por forte heterogeneidade, tais como diferenças entre perfis de segurados, regiões geográficas ou características contratuais, esses modelos se mostram particularmente vantajosos.

A Figura 1 constitui um ponto de partida essencial para a compreensão da relevância do modelo com fração de cura no contexto da análise de sobrevivência. A curva em vermelho representa a função de sobrevivência tradicional, a qual assume que todos os indivíduos da amostra estão inevitavelmente sujeitos ao evento de interesse — por exemplo, o cancelamento de um contrato. Sob essa perspectiva, à medida que o tempo avança, a probabilidade de sobrevivência tende a zero, já que todos os indivíduos eventualmente experimentarão o evento. A curva em azul, por sua vez, representa um modelo mais flexível, que incorpora a hipótese de que uma fração da população está permanentemente imune ao risco. Essa parcela, denominada "curada", jamais enfrentará o evento de interesse, independentemente do tempo de observação. No contexto de contratos de seguros, por exemplo, isso pode representar clientes altamente fiéis, que nunca cancelarão

a apólice. Esse comportamento é refletido na presença de um platô não nulo na função de sobrevivência — indicando que a curva não decai até zero, mesmo em horizontes de tempo longos, sendo amplamente discutido por Cruz, Fuentes e Padilla (2022).

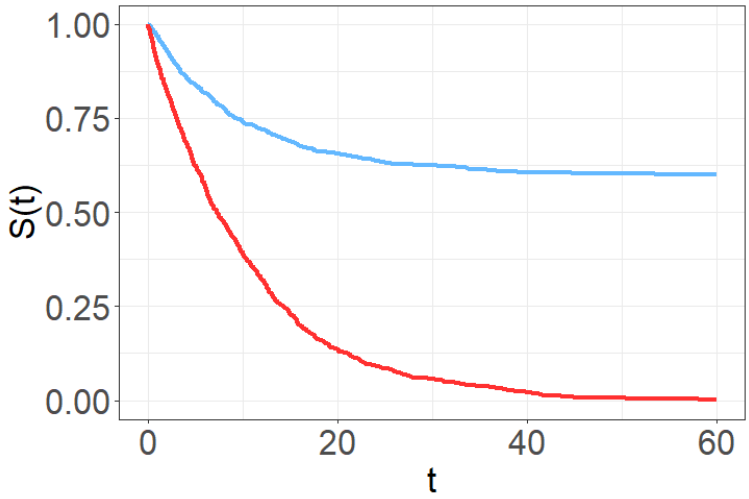


Figura 1: Ilustração da função de sobrevivência empírica usual (vermelho) e com fração de cura (azul).

O objetivo principal deste projeto é explorar e aplicar modelos de análise de sobrevivência com fração de cura no contexto de seguros de vida inteira, a partir de uma perspectiva atuarial. Busca-se compreender como a introdução desse conceito pode aprimorar a modelagem da persistência contratual, facilitar a segmentação de grupos com diferentes níveis de risco e contribuir para uma gestão mais eficiente e estratégica do portfólio. Ao fazer isso, pretende-se ampliar o escopo tradicional da análise de sobrevivência com fração de cura, conforme demonstrado por Martinez et al. (2013), evidenciando sua relevância e aplicabilidade no universo atuarial.

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os principais conceitos da análise de sobrevivência, incluindo as funções de sobrevivência e risco, além de abordar os dados censurados e o estimadores não paramétricos para as funções de sobrevivência e risco. O Capítulo 3 discute o modelo de mistura com fração de cura, detalhando sua estrutura, bem como as funções associadas à modelagem da proporção de indivíduos considerados curados. O Capítulo 4 trata da análise paramétrica, explorando a abordagem Bayesiana para estimação dos parâmetros de interesse, e apresenta os critérios utilizados para seleção dos modelos. O Capítulo 5 descreve as bases de dados utilizadas, a análise exploratória, a modelagem e a comparação dos modelos ajustados, destacando as covariáveis significativas. Por fim, o Capítulo 6 reúne as considerações finais e propõe possíveis extensões deste projeto.

2 Análise de Sobrevivência

Neste capítulo, além da definição formal de análise de sobrevivência, serão apresentados os principais conceitos teóricos que fundamentam esse tipo de estudo, como a função de sobrevivência, a função de risco e a função densidade de probabilidade, bem como as relações entre elas. Também será discutida a definição de dados censurados, com destaque para os diferentes tipos de censura que podem ocorrer. Por fim, será introduzido o estimador não paramétrico de Kaplan-Meier, amplamente utilizado no contexto da análise de sobrevivência para estimar a função de sobrevivência de forma empírica.

A análise de sobrevivência, amplamente discutida por Colosimo e Giolo (2006) e Kleinbaum e Klein (2011) , se dedica ao estudo do tempo até a ocorrência de um evento de interesse, geralmente denominado tempo de sobrevivência e representado por uma variável aleatória contínua T. No contexto do presente trabalho, essa variável representa o tempo até o resgate da reserva para apólices de seguro de vida inteira. Por se tratar de uma variável aleatória, esse tempo pode ser modelado por meio de distribuições probabilísticas, sendo possível analisar tanto a probabilidade de ocorrência do evento ao longo do tempo quanto o comportamento do risco associado.

2.1 Função de Sobrevivência

Dado T (contínuo e não-negativo) e, considerando uma função densidade de probabilidade $f(t)$, temos que a função de distribuição acumulada (FDA) é dada por:

$$F(t) = \int_0^t f(u)\, du = P(T \leq t), \quad 0 < t < \infty. \tag{1}$$

A função de sobrevivência $S(t)$ fornece a probabilidade de um indivíduo sobreviver além de um tempo específico t. Em outras palavras, $S(t)$ representa a probabilidade de a variável aleatória T exceder o tempo t, sendo definida matematicamente como:

$$S(t) = 1 - F(t) = P(T > t), \quad 0 < t < \infty, \tag{2}$$

sendo $S(t)$ uma função monótona decrescente, com as seguintes propriedades: $S(0) = 1$; $S(t) \in [0, 1]$; e $\lim_{t\to\infty} S(t) = 0$.

A Figura 2 apresenta uma função de sobrevivência estimada a partir de dados reais de pacientes com câncer. A curva inicia em 1, indicando probabilidade total de sobrevivência no tempo zero, e decresce ao longo do tempo, à medida que o evento ocorre.

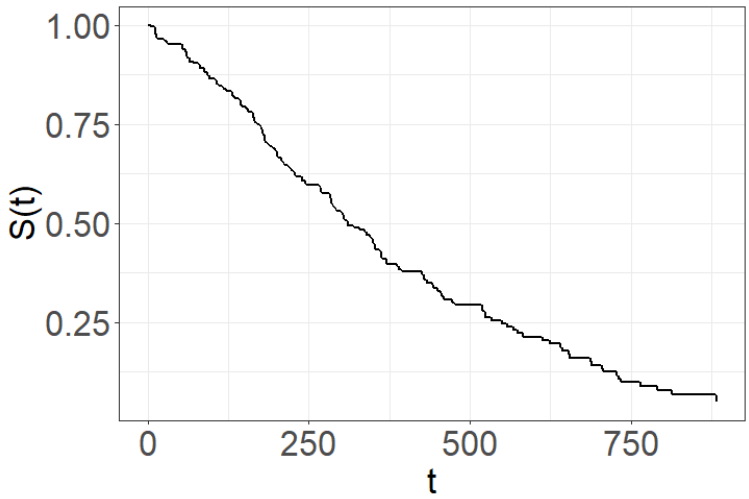


Figura 2: Ilustração da função de sobrevivência com dados de câncer disponíveis no pacote `survival` do R.

Embora todas as curvas de sobrevivência apresentem essas características, a taxa de declínio pode variar dependendo do risco de ocorrência do evento em t. Em certas situações, pode ser mais informativo analisar o comportamento do risco ao longo do tempo, ao invés da própria curva de sobrevivência. Para isso, introduz-se a função de risco.

2.2 Função de Risco

A função de risco representa a taxa instantânea, por unidade de tempo, de ocorrência do evento, dado que o indivíduo ainda não experimentou o evento até o instante t. A função de risco é expressa matematicamente como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}, \quad 0 < t < \infty, \tag{3}$$

tal que $h(t) \geq 0$ e $\displaystyle\int_0^\infty h(t)\, dt = \infty$.

A Figura 3 mostra a estimativa da função de risco com base nos mesmos dados de pacientes com câncer mencionados anteriormente. Essa função detalha como o risco de ocorrência do evento varia ao longo do tempo: curvas crescentes indicam aumento do risco com o tempo, curvas constantes refletem risco uniforme, e curvas decrescentes sugerem redução do risco à medida que o indivíduo continua sem experimentar o evento.

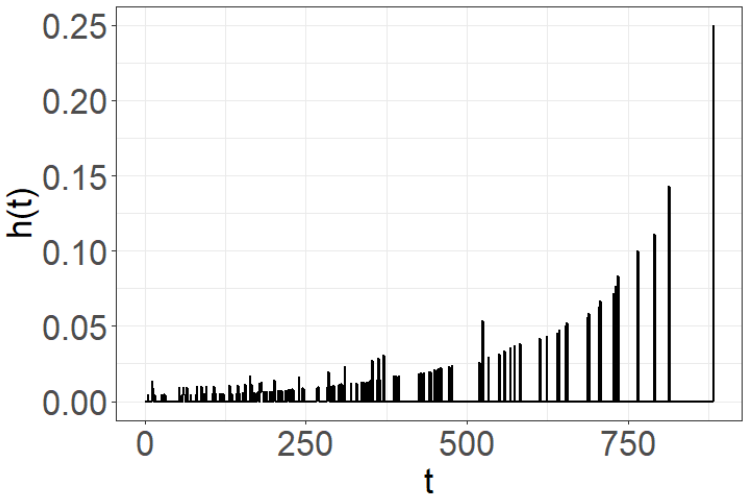


Figura 3: Ilustração da função de risco com dados de câncer disponíveis no pacote `survival` do R.

2.3 Função Densidade de Probabilidade

No contexto da análise de sobrevivência, a função de densidade de probabilidade descreve a distribuição dos tempos até a ocorrência do evento de interesse. Dessa forma, a função densidade de probabilidade $f(t)$ será definida matematicamente como:

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - S(t)) = -\frac{d}{dt}S(t), \quad 0 < t < \infty, \tag{4}$$

e possui as seguintes propriedades: $f(t) \geq 0 \ \forall t \geq 0$; e $\int_0^\infty f(t)\,dt = 1$.

A escolha da função de densidade adequada é um componente fundamental na modelagem em análise de sobrevivência, pois ela define a distribuição dos tempos até o evento de interesse.

2.4 Observações censuradas

De acordo com Colosimo e Giolo (2006) e Carvalho (2011), nos estudos de sobrevivência, é comum que nem todos os indivíduos sofram o evento de interesse antes do término da pesquisa, resultando em dados censurados. A censura ocorre quando o tempo exato do evento não é conhecido, mas sabe-se que ele é maior ou menor que um certo valor registrado.

Os principais tipos de censura em dados de tempo de sobrevivência são: censura à direita, que ocorre quando o evento de interesse ainda não aconteceu até o final do período de observação; censura à esquerda, que se verifica quando o evento já havia ocorrido antes do início da observação; e censura intervalar, que se dá quando se sabe apenas que o evento

ocorreu dentro de um determinado intervalo de tempo, mas não exatamente quando.

A Figura 4 ilustra a diferença entre dados completos e os três principais mecanismos de censura. Partindo de um conjunto de dados completos (Figura 4a), temos a censura do tipo I, quando o estudo termina em um tempo fixado (Figura 4b); a censura do tipo II, quando se observa um número pré-determinado de eventos (Figura 4c); e a censura aleatória, quando indivíduos deixam o estudo por motivos diversos (Figura 4d). Neste trabalho, adotamos a censura aleatória, pois a saída dos indivíduos ocorre por razões não controladas.

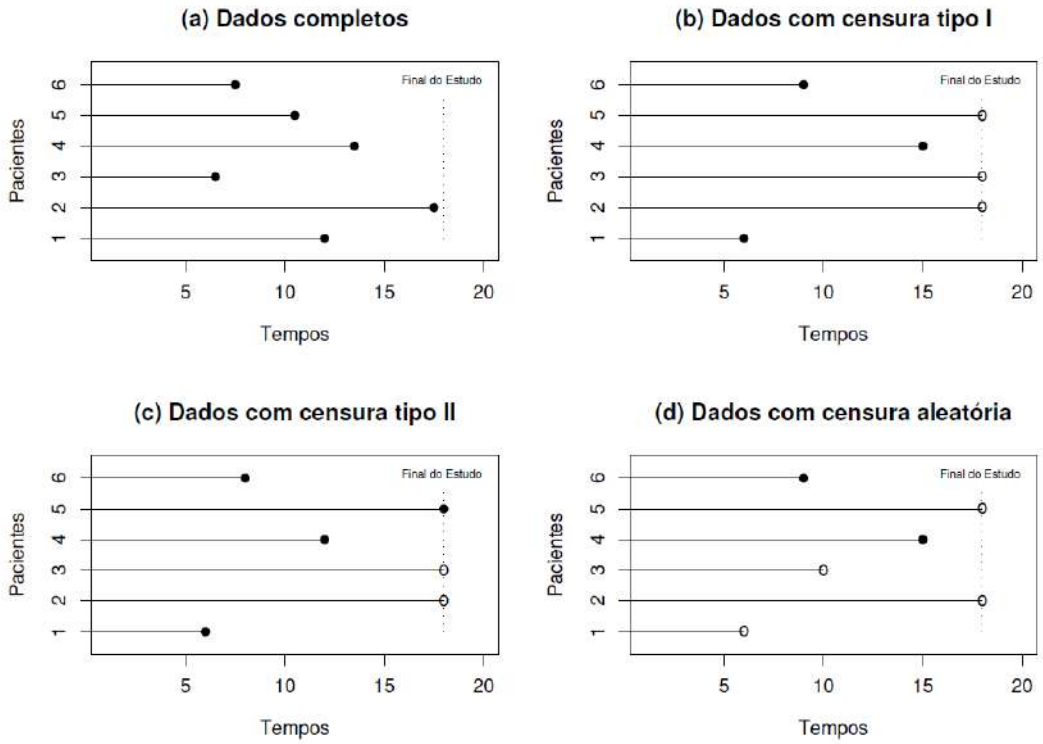


Figura 4: Exemplo de tempos de sobrevivência com diferentes tipos de censura. Fonte: Colosimo e Giolo (2006).

Seja T o tempo até o evento de interesse e C o tempo até a censura. Então, o tempo observado para o elemento i, considerando $i = 1, \ldots, n$ é dado por:

$$t_i = min(T_i, C_i). \tag{5}$$

Dessa forma, os dados observados no contexto da análise de sobrevivência são compostos por $(t_i, I(T_i < C_i))$, isto é, o tempo de acompanhamento e o status de censura. Define-se δ_i como a função indicadora, ou seja:

$$\delta_i = \begin{cases} 1, & \text{se } t_i < c_i \\ 0, & \text{c.c.} \end{cases}$$

Ainda segundo Colosimo e Giolo (2006), mesmo com censura, todas as observações devem ser incluídas na análise estatística, pois contribuem com informações valiosas sobre o tempo de sobrevivência. Métodos como o estimador de Kaplan-Meier e modelagens paramétricas ou semi-paramétricas são amplamente utilizados para tratar esses dados.

2.5 Estimador de Kaplan-Meier e estimador para a função de risco

O estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), é um método não paramétrico para estimar a função de sobrevivência com presença de censura. Ele é amplamente utilizado em estudos de análise de sobrevivência devido à sua capacidade de lidar com tempos de falha observados e censurados.

Supondo que sejam observados p tempos distintos de falha, denotados por $t_1, t_2, \ldots, t_p$, nos quais n_j representa o número de indivíduos em risco imediatamente antes de cada tempo t_j, e d_j corresponde ao número de falhas observadas nesse instante. O estimador de Kaplan-Meier para a função de sobrevivência é definido como:

$$\widehat{S(t)} = \prod_{j|t_j<t} \left(1 - \frac{d_j}{n_j}\right), \tag{6}$$

em que t_j são os tempos distintos de falha, d_j é o número de falhas em t_j e n_j é o número de indivíduos sob risco em t_j.

A motivação para esse estimador decorre da decomposição da função de sobrevivência em termos de probabilidades condicionais sucessivas. Ou seja, a probabilidade de um indivíduo sobreviver além do tempo t, pode ser expressa como o produto das probabilidades de sobreviver a cada tempo de falha anterior, dado que se sobreviveu até aquele ponto:

$$P(T > t) = \prod_{j\,|\,t_j<t} P(T > t_j \mid T \geq t_j). \tag{7}$$

Cada termo condicional $P(T > t_j \mid T \geq t_j)$ é estimado pela proporção de indivíduos que não falharam em t_j, ou seja, $1 - \frac{d_j}{n_j}$. Assim, a função de sobrevivência estimada é construída como um produtório cumulativo dessas probabilidades, resultando em uma estimativa empírica que acomoda naturalmente a presença de censura à direita.

A curva estimada pelo método de Kaplan-Meier é do tipo escada, monótona e não-crescente. Entre as principais propriedades desse estimador, destacam-se: o fato de ser não-viesado para amostras grandes, sua forte consistência, a convergência assintótica para um processo gaussiano, e o fato de corresponder ao estimador de máxima verossimilhança

da função de sobrevivência $S(t)$.

Além do estimador mencionado anteriormente, será considerado também, como alternativa para estimar a função de risco no tempo t, o estimador baseado na proporção de falhas observadas nesse instante. Esse estimador é definido da seguinte forma:

$$\widehat{h(t)} = \frac{d_t}{n_t}, \tag{8}$$

em que d_t representa o número de falhas observadas no tempo t e n_t indica o número de indivíduos em risco imediatamente antes desse tempo.

É importante notar que, embora tanto o estimador de Kaplan-Meier quanto o da função de risco sejam definidos para qualquer instante de tempo t, eles são funções discretas. Isso significa que suas estimativas se modificam ou assumem valores não nulos apenas nos tempos em que há ao menos um evento observado. Nos intervalos entre esses eventos, seus valores permanecem constantes.

3 Modelo de Mistura com Fração de Cura

Os modelos tradicionais de sobrevivência partem da suposição de que todos os indivíduos eventualmente serão submetidos ao evento de interesse em algum momento do tempo e, como resultado, esses modelos podem superestimar o risco a longo prazo.

No entanto, em diversas aplicações práticas, especialmente no contexto atuarial, que é o foco desse estudo, observa-se que uma parcela dos indivíduos não experimentará o evento de interesse, mesmo após longos períodos de observação. Partindo desse pressuposto, segurados com apólices de seguro de vida inteira podem nunca resgatar a reserva acumulada ao longo do tempo, caracterizando assim uma fração de indivíduos potencialmente imunes ao evento de interesse.

Para lidar com esse cenário, os modelos com fração de cura, inicialmente propostos por Boag (1949) e Berkson e Gage (1952) e descritos em detalhes por Maller e Zhou (1996), foram desenvolvidos com o objetivo de identificar e modelar separadamente os indivíduos suscetíveis ao evento e aqueles que são considerados imunes a este.

Neste capítulo, são abordados os principais aspectos teóricos dos modelos de mistura com fração de cura, incluindo a estrutura, a formulação da função de sobrevivência, da função de risco e da função de densidade de probabilidade adaptadas à presença de indivíduos curados.

3.1 Estrutura do modelo

Na formulação do modelo de mistura com fração de cura, assume-se que a população é composta por dois subgrupos distintos: os suscetíveis, que estão em risco e podem vir a sofrer o evento de interesse; e os curados (ou não suscetíveis), que jamais experimentarão tal evento.

Para representar essa heterogeneidade, introduz-se uma variável latente U_i associada a cada indivíduo i, a qual indica sua suscetibilidade ao evento:

$$U_i = \begin{cases} 1, & \text{se o indivíduo é suscetível} \\ 0, & \text{se o indivíduo é imune.} \end{cases}$$

Como se trata de uma variável latente, o valor de U_i não é conhecido diretamente. Apenas quando o evento é observado durante o período de estudo ($\delta_i = 1$) pode-se afirmar com certeza que $U_i = 1$, ou seja, que o indivíduo pertence ao grupo dos suscetíveis.

Por outro lado, nos casos censurados ($\delta_i = 0$), não é possível distinguir se o indivíduo é suscetível ($U_i = 1$) ou curado ($U_i = 0$), uma vez que a ausência do evento

pode ser decorrente tanto da censura quanto da "cura". Essa incerteza é incorporada ao modelo por meio da abordagem de mistura, permitindo estimar a probabilidade de cura com base nas covariáveis observadas.

3.2 Função de Sobrevivência para o modelo de mistura

A partir da estrutura do modelo de mistura, para a variável aleatória T não-negativa, que representa o tempo até o evento, segue que a probabilidade de um indivíduo sobreviver ao tempo t, isto é, a função de sobrevivência considerando tanto os curados quanto os suscetíveis, pode ser expressa da seguinte forma:

$$S_{\text{total}}(t) = P(T > t) = P(U = 0)P(T > t \mid U = 0) + P(U = 1)P(T > t \mid U = 1) \quad (9)$$

$$= p + (1 - p)S(t), \quad (10)$$

em que $p \in [0, 1]$ representa a fração de curados e $S(t)$ representa a função de sobrevivência dos suscetíveis, ou seja, a função de sobrevivência usual mencionada na Seção 2.1.

Entre as principais propriedades da função de sobrevivência que contempla curados e suscetíveis, estão:

1. se $p = 0$, então $S_{\text{total}}(t) = S(t)$;
2. $S_{\text{total}}(0) = 1$;
3. $S_{\text{total}}(t)$ é decrescente;
4. $\lim_{t\to\infty} S_{\text{total}}(t) = p$; e
5. se $p = 1$, então $S_{\text{total}}(t) = 1 \ \forall t$.

Nota-se que, diferentemente do modelo usual de sobrevivência — que assume que todos os indivíduos eventualmente estarão em risco e, portanto, que $\lim_{t\to\infty} S(t) = 0$ — o modelo com fração de cura permite que uma parcela da população nunca experimente o evento de interesse. Essa característica é refletida pelo limite não nulo da função de sobrevivência total o que representa uma das principais diferenças em relação aos modelos tradicionais.

3.3 Função de Risco para o modelo de mistura

A função de risco instantâneo, no contexto do modelo de mistura com fração de curados, passa a incorporar a heterogeneidade da população resultante da mistura entre

curados e suscetíveis. Nesse cenário, a função de risco reflete a combinação ponderada do comportamento de ambos os grupos. Partindo da relação estabelecida na equação (2.2), essa função será expressa por:

$$h_{\text{total}}(t) = \frac{f_{\text{total}}(t)}{S_{\text{total}}(t)} = \frac{(1-p)f(t)}{p + (1-p)S(t)}, \tag{11}$$

em que p é a proporção de curados, $f(t)$ é a função densidade de probabilidade para os suscetíveis e $S(t)$ é a função de sobrevivência para os suscetíveis.

Destacam-se, ainda, as seguintes propriedades da função de risco para suscetíveis e não suscetíveis:

1. se $p = 0$, então $h_{\text{total}}(t) = \frac{f(t)}{S(t)}$;
2. se $p = 1$, então $h_{\text{total}}(t) = 0 \ \forall t$; e
3. $h_{\text{total}}(t) \leq h(t) \ \forall t$.

Diferente dos modelos tradicionais, onde o risco tende a aumentar ou estabilizar com o tempo, nos modelos com fração de cura o risco pode diminuir a medida que o tempo passa. Esse comportamento decorre do fato de que os mais vulneráveis tendem a experimentar o evento mais cedo e, ao longo do tempo, resta no grupo de risco uma proporção maior de curados e indivíduos mais resistentes, fazendo com que o risco reduza a longo prazo.

3.4 Função Densidade de Probabilidade para o modelo de mistura

Considerando que a a população de estudo é composta por dois subgrupos (curados e suscetíveis), a função densidade de probabilidade para curados e suscetíveis, partindo da equação (2.3), será definida por:

$$f_{\text{total}}(t) = -\frac{d}{dt}S_{\text{total}}(t) = -\frac{d}{dt}(p + (1-p)S(t)) = -(1-p)\frac{d}{dt}S(t) = (1-p)f(t), \tag{12}$$

em que p representa a proporção de curados e $f(t)$ representa a função densidade de probabilidade para os suscetíveis.

A função de densidade de probabilidade para o modelo de mistura com fração de cura $f_{\text{total}}(t)$ possui as seguintes propriedades:

1. se $p = 0$, então $f_{\text{total}}(t) = f(t)$;

2. se $p = 1$, então $f_{\text{total}}(t) = 0$;

3. $f_{\text{total}}(t) \leq f(t)\ \forall t$; e

4. $\int_0^\infty f_{\text{total}}(t)\,dt = \int_0^\infty (1-p) f(t)\,dt = (1-p)\int_0^\infty f(t)\,dt = (1-p)$.

A função de densidade de probabilidade total resulta da ponderação da densidade dos indivíduos suscetíveis pela proporção de indivíduos não curados na população. Dessa forma, a escolha da função densidade para os suscetíveis exerce influência direta sobre o comportamento da função de densidade da população como um todo.

4 Análise Paramétrica

A análise paramétrica no contexto de análise de sobrevivência assume que o tempo até o evento de interesse segue uma distribuição de probabilidade completamente especificada por um ou mais parâmetros desconhecidos. Portanto, o objetivo é estimar tais parâmetros com base nos dados observados, que podem incluir tanto observações completas (eventos ocorridos) quanto censuradas (eventos não observados até o fim do tempo de acompanhamento).

A escolha da distribuição apropriada, tais como Exponencial, Weibull, Log-normal, como é feito por Colosimo e Giolo (2006) e analogamente por Carvalho (2011) é essencial, pois influencia diretamente as estimativas de risco e sobrevivência, bem como a inferência sobre o comportamento do tempo até o evento. No presente estudo, conforme mencionado anteriormente, o foco está na modelagem do tempo até o resgate da reserva por parte dos segurados, o que permite uma melhor compreensão da persistência e do efeito do cancelamento.

Neste capítulo, será apresentada a formulação da função de verossimilhança tanto para o modelo usual quanto para o modelo com fração de cura. Em seguida, será discutida a abordagem Bayesiana para a estimação dos parâmetros de interesse. Por fim, serão introduzidos os principais critérios utilizados para seleção e comparação de modelos, os quais servirão de base para a análise e interpretação dos resultados nos capítulos subsequentes.

4.1 Função de Verossimilhança

Nos modelos de sobrevivência tradicionais, assume-se que todos os indivíduos são suscetíveis ao evento. Considere uma amostra aleatória com n indivíduos, na qual se observam, para cada indivíduo $i = 1, 2, \ldots, n$, o tempo $t_i = \min(T_i, C_i)$, sendo T_i o tempo até o evento e C_i o tempo de censura, além da variável indicadora de censura δ_i. Adicionalmente, assume-se que cada indivíduo está associado a um vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$.

Sob a suposição de independência entre os indivíduos e entre os tempos T_i e C_i, a função de verossimilhança do modelo paramétrico com covariáveis é expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})\right]^{1-\delta_i}, \tag{13}$$

em que $f(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ representa a função densidade de probabilidade condicional ao vetor de covariáveis, $S(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ é a função de sobrevivência correspondente e $\boldsymbol{\theta}$ é o vetor de

parâmetros da distribuição assumida.

Modelos com Fração de Cura

Nos modelos com fração de cura, assume-se que uma proporção da população é imune ao evento de interesse, ou seja, jamais o experimentará, enquanto a parcela restante permanece suscetível. Para formalizar essa estrutura, introduz-se uma variável latente U_i, indicadora de cura, tal como descrito na Seção 3.1, em que $U_i = 0$ denota um indivíduo curado e $U_i = 1$, um indivíduo suscetível.

A probabilidade de um indivíduo pertencer ao grupo de curados pode depender de características individuais observáveis. Para isso, associa-se a cada indivíduo i um vetor de covariáveis específicas da componente de cura, denotado por $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})^\top$. Com base nesse vetor, modela-se a probabilidade de cura $p(\mathbf{z}_i)$ por meio de uma função logística, conforme apresentado por Cruz, Fuentes e Padilla (2022):

$$p(\mathbf{z}_i) = \Pr(U_i = 0 \mid \mathbf{z}_i) = \frac{\exp(\mathbf{z}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{z}_i^\top \boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\beta})}, \tag{14}$$

em que $\boldsymbol{\beta}$ representa o vetor de parâmetros da regressão associado à componente de cura.

É importante destacar que os vetores de covariáveis $\mathbf{x}_i$ e $\mathbf{z}_i$ podem ser distintos, refletindo diferentes aspectos da análise de sobrevivência. No contexto deste projeto, que considera o tempo até o resgate da reserva de apólices de seguros de vida inteira, $\mathbf{x}_i$ está associado ao tempo até o resgate por parte dos segurados suscetíveis, enquanto $\mathbf{z}_i$ influencia a probabilidade de o segurado jamais realizar esse resgate — ou seja, de pertencer à fração considerada "curada". Essa fração representa os indivíduos que mantêm suas apólices até o final da vida e, portanto, não resgatam a reserva acumulada.

Dessa forma, a contribuição individual para a função de verossimilhança depende tanto do status de censura quanto da suscetibilidade do indivíduo ao resgate. Essa contribuição pode ser expressa por:

$$L_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = \begin{cases} p(\mathbf{z}_i), & \text{se } U_i = 0 \text{ (imune)} \\ (1 - p(\mathbf{z}_i)) f(t_i \mid \boldsymbol{\theta}), & \text{se } U_i = 1 \text{ e } \delta_i = 1 \text{ (suscetível e evento observado)} \\ (1 - p(\mathbf{z}_i)) S(t_i \mid \boldsymbol{\theta}), & \text{se } U_i = 1 \text{ e } \delta_i = 0 \text{ (suscetível e censurado).} \end{cases} \tag{15}$$

Assim, a função de verossimilhança para o modelo com fração de cura é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[p(\mathbf{z}_i)^{(1-U_i)} (1 - p(\mathbf{z}_i))^{U_i} f(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})^{U_i \delta_i} S(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})^{U_i (1-\delta_i)} \right], \tag{16}$$

em que $p(\mathbf{z}_i)$ é a probabilidade de cura do indivíduo i, conforme a equação (14), $f(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ e $S(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ são, respectivamente, a função densidade e a função de sobrevivência dos suscetíveis.

Vale destacar que, ao se assumir que todos os indivíduos são suscetíveis (ou seja, $U_i = 1$ para todo i), o modelo com fração de cura se reduz ao modelo tradicional apresentado anteriormente na equação (13).

4.2 Distribuição Log-Normal

A distribuição Log-Normal, conforme discutida por Lawless (2022), foi adotada para modelar os tempos até o evento entre os indivíduos suscetíveis, por sua capacidade de representar adequadamente distribuições assimétricas à direita — uma característica comum em dados de sobrevivência — como também destacado por Lobo, Fonseca e Alves (2024). Esta distribuição é particularmente útil quando os logaritmos dos tempos de sobrevivência seguem uma distribuição Normal.

Para essa distribuição, as funções densidade de probabilidade e de sobrevivência são definidas respectivamente por:

$$f(t_i; \mu, \sigma^2) = \frac{1}{t_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln t_i - \mu)^2}{2\sigma^2} \right), \quad t_i > 0, \quad \mu \in \mathbb{R}, \quad \sigma^2 > 0; \tag{17}$$

$$S(t_i; \mu, \sigma^2) = 1 - \Phi\left(\frac{\ln t_i - \mu}{\sigma} \right), \tag{18}$$

em que μ e σ são parâmetros de locação e escala, respectivamente, e $\Phi(\cdot)$ denota a função de distribuição normal padrão.

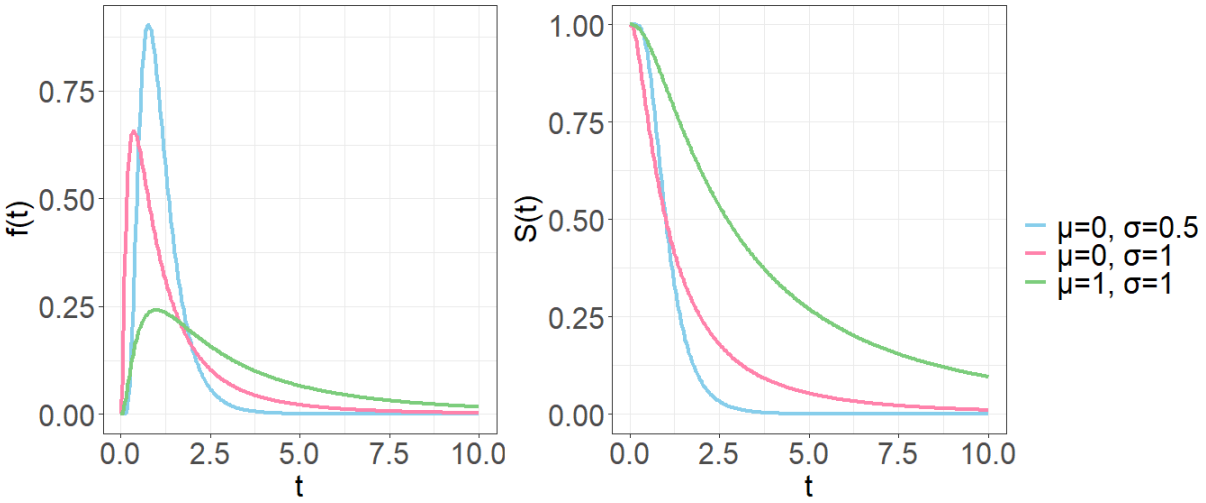


Figura 5: Gráfico das funções densidade de probabilidade (esquerda) e de sobrevivência (direita) variando os parâmetros de locação e escala.

Como mostrado nos painéis da Figura 5, diferentes combinações de μ e σ resultam em formas distintas das curvas de densidade e de sobrevivência. Isso evidencia a flexibilidade da distribuição Log-Normal em se adaptar a diferentes padrões de dados observados em estudos de sobrevivência.

4.3 Abordagem Bayesiana

Na abordagem Bayesiana, conforme discutido por Cruz, Fuentes e Padilla (2022), o interesse está em inferir a respeito da distribuição a posteriori dos parâmetros do modelo considerando a análise de sobrevivência com fração de cura, incorporando incertezas a partir de distribuições a priori.

Assuma que cada indivíduo $i = 1, \ldots, n$ está associado a um vetor de covariáveis $\mathbf{z}_i \in \mathbb{R}^q$ relacionado à fração de cura e a um vetor de covariáveis $\mathbf{x}_i \in \mathbb{R}^p$ relacionado ao tempo de sobrevivência dos suscetíveis.

A probabilidade de cura é modelada por meio de uma função logística, conforme apresentado na equação (14), sendo $\boldsymbol{\beta} \in \mathbb{R}^q$ o vetor de parâmetros da regressão.

Para os indivíduos suscetíveis ($U_i = 1$), modela-se o logaritmo do tempo até o evento, $Y_i = \log(T_i)$, assumindo que $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, onde a média μ_i é uma função das covariáveis e a variância σ^2 é constante. Essa abordagem implica que o tempo original T_i segue uma distribuição Log-Normal. A média do logaritmo do tempo é especificada como:

$$\mu_i = \mathbf{x}_i^\top \boldsymbol{\alpha}, \tag{19}$$

em que $\boldsymbol{\alpha} \in \mathbb{R}^p$ é o vetor de parâmetros associados à média do logaritmo dos tempos de sobrevivência.

Essa formulação permite capturar a heterogeneidade entre os indivíduos tanto em relação à propensão ao resgate da reserva (via $\mathbf{x}_i$), quanto à probabilidade de jamais realizar tal resgate (via $\mathbf{z}_i$).

Função de Verossimilhança Completa

Considerando a variável latente $U_i \in \{0, 1\}$, a contribuição individual à função de verossimilhança é dada por:

$$L_i(\boldsymbol{\theta}, \boldsymbol{\beta}, U_i) = p(\mathbf{z}_i)^{(1-U_i)}(1 - p(\mathbf{z}_i))^{U_i} f(t_i \mid \mu_i, \sigma^2)^{U_i \delta_i} S(t_i \mid \mu_i, \sigma^2)^{U_i(1-\delta_i)}, \tag{20}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \sigma^2)$, e as funções f e S são, respectivamente, a densidade e função de sobrevivência da distribuição Log-Normal.

A função de verossimilhança conjunta é, portanto:

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{U}) = \prod_{i=1}^{n} \left[p(\mathbf{z}_i)^{(1-U_i)}(1 - p(\mathbf{z}_i))^{U_i} f(t_i \mid \mu_i, \sigma^2)^{U_i \delta_i} S(t_i \mid \mu_i, \sigma^2)^{U_i(1-\delta_i)} \right]. \tag{21}$$

Distribuições a Priori

São assumidas distribuições a priori Normais multivariadas para os vetores de parâmetros de regressão $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$. Para o parâmetro de variância σ^2 da distribuição Log-Normal (referente ao logaritmo dos tempos de sobrevivência dos indivíduos suscetíveis), adota-se uma distribuição Inversa Gama.

A escolha de distribuições Normais para os coeficientes de regressão é comum em contextos bayesianos, pois oferece uma estrutura flexível, matematicamente conveniente e permite fácil incorporação de informação sobre a média e (co)variância dos parâmetros.

$$\boldsymbol{\beta} \sim \mathcal{N}_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) \quad \Rightarrow \quad \pi(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}_\beta|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right), \quad \boldsymbol{\beta} \in \mathbb{R}^q;$$

$$\boldsymbol{\alpha} \sim \mathcal{N}_p(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_\alpha) \quad \Rightarrow \quad \pi(\boldsymbol{\alpha}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_\alpha|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\top \boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) \right), \quad \boldsymbol{\alpha} \in \mathbb{R}^p;$$

em que $\boldsymbol{\beta}_0$ e $\boldsymbol{\alpha}_0$ são vetores de médias e $\boldsymbol{\Sigma}_\beta$ e $\boldsymbol{\Sigma}_\alpha$ são matrizes de covariância pré-definidas.

Para o parâmetro de variância σ^2, a priori é dada por:

$$\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma) \quad \Rightarrow \quad \pi(\sigma^2) = \frac{(b_\sigma)^{a_\sigma}}{\Gamma(a_\sigma)} (\sigma^2)^{-(a_\sigma+1)} \exp\left(-\frac{b_\sigma}{\sigma^2} \right), \quad \sigma^2 > 0;$$

em que a_σ e b_σ são os hiperparâmetros de forma e escala, respectivamente, pré-definidos para refletir o conhecimento prévio ou para serem fracamente informativos.

Distribuição a Posteriori

A distribuição a posteriori é proporcional ao produto entre a verossimilhança e as distribuições a priori:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{U} \mid \mathbf{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{U}) \times \pi(\boldsymbol{\beta}) \times \pi(\boldsymbol{\alpha}) \times \pi(\sigma^2). \tag{22}$$

Substituindo as expressões correspondentes, obtemos:

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{U} \mid \mathbf{t}, \boldsymbol{\delta}) \propto & \prod_{i=1}^{n} \left[p(\mathbf{z}_i)^{(1-U_i)} (1 - p(\mathbf{z}_i))^{U_i} f(t_i \mid \mu_i, \sigma^2)^{U_i \delta_i} S(t_i \mid \mu_i, \sigma^2)^{U_i (1-\delta_i)} \right] \\ & \times \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right) \\ & \times \exp\left(-\frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\top \boldsymbol{\Sigma}_\alpha^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) \right) \\ & \times (\sigma^2)^{-(a_\sigma + 1)} \exp\left(-\frac{b_\sigma}{\sigma^2} \right), \end{aligned} \tag{23}$$

em que $\mathbf{t} = (t_1, \ldots, t_n)$ representa os tempos de sobrevivência observados e $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ indica o status de censura.

Observa-se que a distribuição a posteriori, derivada na equação (23), não possui uma forma fechada, o que impossibilita a obtenção de uma solução analítica direta. Para contornar essa limitação, será utilizado o método de amostragem Monte Carlo Hamiltoniano (HMC), uma técnica de MCMC detalhada em Brooks et al. (2011), que será implementado no `Stan`, conforme descrito em Gelman et al. (2013), com o objetivo de obter amostras representativas da distribuição a posteriori dos parâmetros, mesmo em cenários complexos.

Embora as distribuições condicionais completas possam ser obtidas, conforme detalhado no Apêndice da Seção A, optou-se por utilizar o `Stan` devido à sua robustez, eficiência computacional e facilidade de implementação, evitando a necessidade de desenvolver um algoritmo de Gibbs sampling específico para este modelo.

4.4 Critérios de seleção de modelos

A seleção do modelo estatístico mais adequado é uma etapa fundamental na análise de dados, especialmente quando há múltiplas especificações plausíveis. Para comparar modelos, utilizam-se critérios de informação que equilibram a qualidade do ajuste e a complexidade da estrutura, penalizando modelos excessivamente parametrizados. Neste trabalho, foram considerados dois critérios amplamente utilizados no contexto bayesiano: o WAIC (Watanabe-Akaike Information Criterion; Watanabe (2010)) e o LOO-CV

(Leave-One-Out Cross-Validation; Stone (1974)). Uma abordagem prática sobre o uso e a implementação de ambos os critérios é apresentada por Gelman et al. (2013).

O WAIC é um critério inteiramente bayesiano que considera a distribuição posterior completa. Sua formulação é dada por:

$$\text{WAIC} = -2\left(\text{lppd} - p_{\text{WAIC}}\right), \tag{24}$$

em que lppd denota o logaritmo da média posterior preditiva e p_{WAIC} representa a penalização pela complexidade do modelo. Essa métrica é mais apropriada para modelos complexos ou com alta variabilidade nos parâmetros.

O LOO-CV é um método de validação cruzada que avalia a capacidade preditiva do modelo ao excluir uma observação por vez. É expresso por:

$$\text{LOO-CV} = -2\sum_{i=1}^{n} \log\left(p(y_i \mid y_{-i})\right), \tag{25}$$

em que y_i é a i-ésima observação, y_{-i} representa o conjunto de dados com essa observação removida e $p(y_i \mid y_{-i})$ é a densidade preditiva condicional da observação excluída. Pode ser considerado uma alternativa robusta ao WAIC, especialmente útil quando há observações influentes ou distribuições posteriores complexas.

De forma geral, menores valores de WAIC ou LOO-CV indicam modelos com melhor desempenho preditivo e menor complexidade efetiva. Neste estudo, tais critérios foram empregados para comparar diferentes especificações do modelo com fração de cura, auxiliando na seleção da estrutura mais adequada aos dados observados. Para o cálculo desses critérios, foi utilizado o pacote `loo` (Vehtari, Gelman e Gabry (2017)) do R, que permite estimar essas métricas a partir de amostras da posteriori geradas via métodos bayesianos.

5 Aplicações

Neste capítulo, são apresentados os resultados de duas aplicações do modelo de sobrevivência com fração de cura: um estudo simulado, desenvolvido com o objetivo de avaliar o comportamento do modelo proposto na Seção 4 em um cenário controlado, e uma aplicação com dados reais de apólices de seguro de vida inteira. No caso do estudo simulado, o foco recai sobre a validação da metodologia, por meio da definição de um conjunto simplificado de covariáveis e da análise dos parâmetros estimados.

Já, na aplicação com dados reais, o processo analítico é mais detalhado. Inicialmente, são descritas as variáveis de interesse, seguidas por uma análise exploratória que visa compreender a estrutura dos dados e identificar padrões relevantes. Em seguida, são apresentados os modelos ajustados, com suas especificações e métodos de estimação. Por fim, os modelos são comparados com base em critérios de seleção previamente definidos na Subseção 4.4, permitindo avaliar o desempenho relativo de cada configuração e a contribuição das covariáveis para a explicação do fenômeno estudado.

5.1 Estudo simulado

Neste estudo, foi gerada uma base de dados simulada com 5.000 indivíduos, representando um cenário típico de seguros de vida, no qual parte dos segurados é suscetível à ocorrência de um evento (como resgate ou sinistro), enquanto outra parte nunca apresentará o evento, sendo considerada curada. A estrutura dos dados segue um modelo de mistura com fração de cura, no qual os tempos até o evento, entre os suscetíveis, são assumidos como provenientes de uma distribuição Log-Normal. Adicionalmente, foi incorporada uma censura à direita em aproximadamente 60% das observações, refletindo uma alta proporção de dados censurados — característica comum em estudos atuariais de seguros de vida, onde muitos indivíduos permanecem ativos até o fim do período de observação. A construção da base simulada permite verificar se o modelo é capaz de recuperar os efeitos esperados das covariáveis sobre a fração de cura e o tempo até o evento.

Os parâmetros utilizados na simulação incluem um intercepto e os efeitos da covariável *forma_cobranca* sobre o tempo até o evento (α_1 e α_2), bem como um intercepto e os efeitos da covariável *categoria_cliente* sobre a fração de cura (β_1 e β_2), além do parâmetro de escala da distribuição Log-Normal (σ).

O intercepto, em ambos os casos, corresponde à categoria de referência de cada covariável. Os valores dos parâmetros foram definidos de modo a refletir padrões esperados no comportamento dos segurados e são apresentados na Tabela 3 (Valor Real), correspondendo aos coeficientes utilizados na geração dos dados simulados.

Para cada indivíduo, foram simuladas duas covariáveis categóricas com distribuição uniforme entre seus níveis. A primeira, denominada *categoria_cliente*, representa a categoria em que o cliente está inserido e assume dois valores: Cliente A e Cliente B. Essa covariável influencia diretamente a probabilidade de cura, sendo utilizada na parte logística do modelo. Indivíduos da categoria Cliente B são associados a uma maior chance de pertencer ao grupo curado, enquanto aqueles classificados como Cliente A apresentam menor probabilidade de cura.

A segunda covariável, *forma_cobranca*, representa o método de pagamento adotado pelo segurado, categorizada como Dinheiro ou Cartão. Essa covariável influencia exclusivamente o tempo até a ocorrência do evento entre os indivíduos suscetíveis, sendo incorporada ao componente da distribuição Log-Normal responsável por modelar esses tempos.

A análise foi conduzida a partir de três modelagens distintas, nas quais a covariável *forma_cobranca* foi utilizada para modelar o tempo até o evento entre os indivíduos suscetíveis. Essa escolha visa manter a parte relacionada à sobrevivência constante, de modo a isolar e evidenciar os efeitos provocados por diferentes formas de especificação da fração de cura. Com isso, torna-se possível avaliar de forma mais precisa a contribuição dessa componente para o ajuste global do modelo.

A principal diferença entre as modelagens está na forma como a fração de cura é tratada. Na Modelagem 1, utiliza-se um modelo tradicional de sobrevivência, que não contempla a possibilidade de indivíduos curados. Na Modelagem 2, introduz-se a fração de cura de forma homogênea, considerando apenas um intercepto — ou seja, todos os indivíduos possuem a mesma probabilidade de estarem curados. Por fim, na Modelagem 3, a fração de cura passa a ser modelada em função da covariável *categoria_cliente*, incorporando ao modelo o efeito dessa característica na chance de o indivíduo pertencer ao grupo de curados.

As curvas de sobrevivência e de risco para as Modelagens 1, 2 e 3 são apresentadas nas Figuras 6 a 8. Os parâmetros estimados do modelo para cada uma dessas modelagens estão detalhados nas Tabelas 1 a 3.

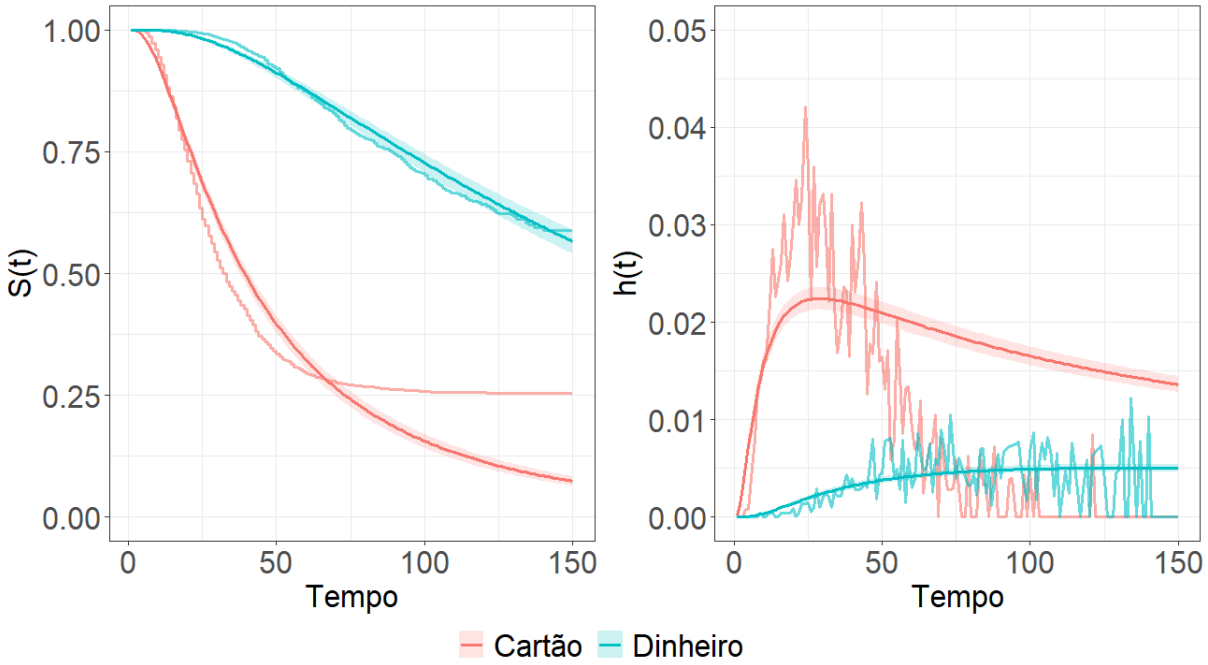


Figura 6: Modelagem 1 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 1: Modelagem 1 - Ajuste do modelo tradicional.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	3,67	(3,63 ; 3,71)
Dinheiro (α_2)	1,50	(1,43 ; 1,57)
σ	0,93	(0,90 ; 0,96)

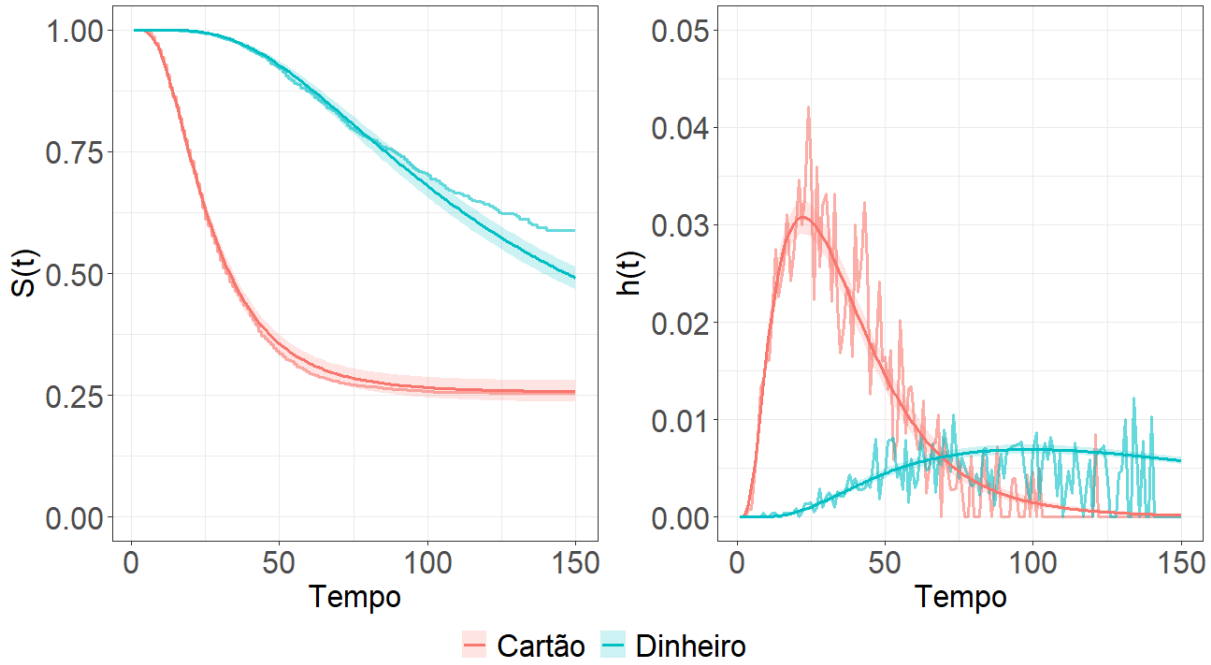


Figura 7: Modelagem 2 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 2: Modelagem 2 - Ajuste do modelo de fração de cura apenas com o intercepto.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	3,23	(3,19 ; 3,26)
Dinheiro (α_2)	1,49	(1,43 ; 1,54)
Intercepto (β_1)	-1,07	(-1,18 ; -0,95)
σ	0,62	(0,60 ; 0,64)
Grupo	**Fração de Cura**	**IC 95%**
Cartão	0,256	(0,235 ; 0,279)
Dinheiro	0,256	(0,235 ; 0,279)

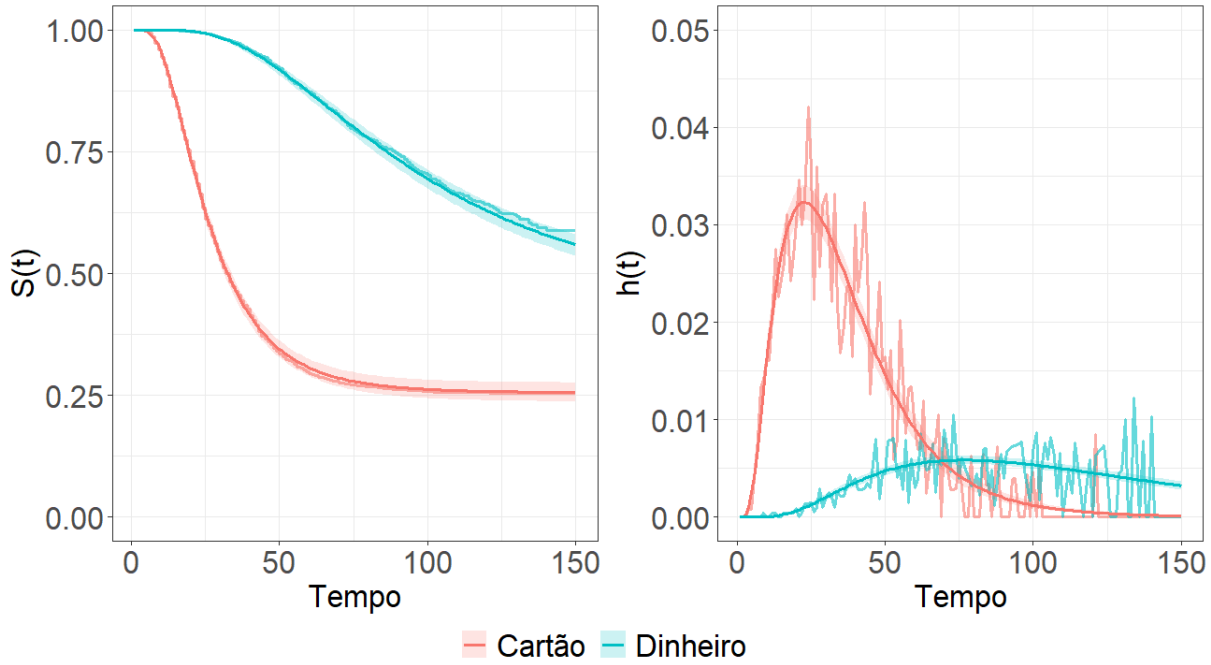


Figura 8: Modelagem 3 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 3: Modelagem 3 - Ajuste do modelo de fração de cura com uma covariável.

Coeficientes	**Valor Real**	**Média**	**IC 95%**
Intercepto (α_1)	3,20	3,22	(3,19 ; 3,25)
Dinheiro (α_2)	1,30	1,32	(1,25 ; 1,38)
Intercepto (β_1)	-1,50	-1,40	(-1,54 ; -1,27)
Cliente B (β_2)	1,50	1,40	(1,21 ; 1,60)
σ	0,60	0,59	(0,57 ; 0,61)
Grupo	**Fração de Cura**	**IC 95%**	
Cartão	0,255	(0,236 ; 0,276)	
Dinheiro	0,442	(0,407 ; 0,475)	

Nas três modelagens, observou-se uma melhora progressiva na adequação do modelo à estrutura dos dados conforme a fração de cura foi incorporada e, posteriormente, associada à covariável *categoria_cliente*. Na Modelagem 1 (Figura 6), sem considerar indivíduos curados, os tempos até o evento foram superestimados, indicando ajuste inadequado. Na Modelagem 2 (Figura 7), a inclusão de uma fração de cura homogênea melhorou significativamente o ajuste mesmo não capturando a heterogeneidade entre os grupos. Na Modelagem 3 (Figura 8), ao modelar a fração de cura em função da *categoria_cliente*, o modelo refletiu o efeito esperado ao introduzir a covariável, conforme especificado na simulação, lidando com uma fração de curados diferente para cada forma

de cobrança.

A Tabela 1 apresenta os parâmetros estimados na Modelagem 1, que assume um modelo de sobrevivência sem fração de cura. Nesse caso, o parâmetro associado à variável Dinheiro é positivo (1,50), indicando que indivíduos pertencentes a esse grupo tendem a apresentar maior tempo até o evento, ou seja, maior persistência ou menor risco. O modelo considera que todos os indivíduos estão suscetíveis ao evento, sem admitir a possibilidade de cura.

Na Modelagem 2, ilustrada na Tabela 2, introduz-se uma fração de cura constante, assumindo que uma proporção fixa da população nunca experimenta o evento. Observa-se que o efeito de Dinheiro no tempo até o evento permanece positivo (1,49), reforçando a associação com maior persistência. A fração de cura estimada é de aproximadamente 25,6% para ambos os grupos, o que sugere que cerca de um quarto da população são considerados curados.

A Tabela 3 exibe os resultados da Modelagem 3, em que a fração de cura é modelada de forma diferenciada por grupo. O efeito da variável Dinheiro sobre o tempo até o evento é novamente positivo (1,30), mantendo o padrão observado nas modelagens anteriores. Além disso, na Modelagem 3, conseguimos recuperar os parâmetros utilizados para a construção da base simulada, indicando que o modelo foi capaz de identificar adequadamente os efeitos planejados, tanto no tempo até o evento quanto na probabilidade de cura entre os grupos.

A Tabela 4 apresenta os valores dos critérios de informação WAIC e LOO-CV obtidos para cada uma das modelagens consideradas no estudo simulado. Ambos os critérios de informação apresentados nessa tabela confirmam a melhora progressiva no ajuste dos modelos à medida que a fração de cura é incorporada e parametrizada em função da covariável *categoria_cliente*. A Modelagem 3, que considera a heterogeneidade na probabilidade de cura, obteve os menores valores de WAIC e LOO-CV, sugerindo melhor qualidade de ajuste e maior capacidade preditiva em comparação com as Modelagens 1 e 2.

Tabela 4: Comparação do ajuste dos modelos considerando os critérios de informação WAIC e LOO-CV, respectivamente.

Modelagem	**WAIC**	**LOO-CV**
1	21.125,4	21.125,4
2	20.601,7	20.601,7
3	20.439,5	20.439,5

5.2 Análise referente a seguros de vida inteira

A base de dados utilizada neste estudo ilustra a aplicação prática de um modelo com fração de cura, que considera tanto o tempo até o resgate da reserva constituída quanto a proporção de indivíduos que nunca experimentarão esse evento.

Foram utilizados os dados do portfólio `uslapseagent`, do pacote `CASdatasets` (Dutang, Charpentier e Gallic (2024)) do software R, que contém informações detalhadas sobre 29.317 apólices de seguro de vida inteira (whole life), todas comercializadas exclusivamente por agentes entre janeiro de 1995 e dezembro de 2008. Cada linha da base representa uma apólice individual e está associada a um conjunto de 14 variáveis, que abrangem características do segurado, do contrato e do contexto econômico vigente ao longo da vida da apólice.

A presença de variáveis na base é fundamental nesse contexto, pois possibilita uma análise mais aprofundada dos fatores associados tanto à susceptibilidade quanto à "cura". Com isso, é possível obter uma compreensão mais detalhada dos mecanismos envolvidos e construir modelos que reflitam de forma mais fiel a heterogeneidade presente na população analisada.

As variáveis disponíveis incluem características do segurado, como gênero, faixa etária de subscrição e local de residência, que ajudam a capturar diferenças comportamentais e socioeconômicas associadas à decisão de manter ou resgatar a reserva da apólice. Também estão presentes informações contratuais relevantes, como a presença de cobertura adicional por morte acidental e a frequência de pagamento, que podem influenciar a atratividade e o comprometimento financeiro assumido com o contrato. Variáveis relacionadas ao risco do segurado, como o estado de risco (fumante ou não fumante) também podem contribuir para contextualizar o ambiente em que a decisão de resgate é tomada.

A variável *duration* representa o tempo decorrido entre a emissão da apólice e o momento do resgate, medido em trimestres. Essa variável é fundamental para o estudo da persistência, pois permite acompanhar a evolução de cada contrato ao longo do tempo. No presente trabalho, ela constitui a variável de interesse, representando o tempo até o cancelamento da apólice, ocorrido no momento de resgate da reserva.

Em particular, para modelos com fração de cura, que assumem a existência de uma parcela da população que nunca experimentará o evento de interesse, a análise cuidadosa dessas variáveis permite distinguir perfis de segurados com alta probabilidade de permanência vitalícia.

A Tabela 5 apresenta um glossário das variáveis disponíveis na base `uslapseagent`.

Tabela 5: Glossário das variáveis.

Variável	**Descrição**
issue.date	Data de emissão da apólice
duration	Duração do contrato em trimestres
acc.death.rider	Indicador da presença de cobertura adicional por morte acidental
gender	Gênero do segurado
premium.frequency	Frequência de pagamento do prêmio
risk.state	Estado de risco do segurado
underwriting.age	Faixa etária na subscrição
living.place	Local de residência do segurado
annual.premium	Valor do prêmio anual, padronizado
DJIA	Última variação trimestral observada do índice Dow Jones (padronizada)
termination.cause	Tipo de término da apólice
surrender	Indicador binário para resgate voluntário por parte do segurado
death	Indicador binário para falecimento do segurado
other	Indicador binário para outras causas de término, como fim contratual
allcause	Indicador binário que consolida todos os tipos de término

O presente estudo utilizará exclusivamente as variáveis categóricas para descrever o processo de persistência considerando o modelo com fração de cura. Essa escolha é motivada tanto por razões metodológicas quanto práticas. Do ponto de vista metodológico, variáveis categóricas são especialmente adequadas para capturar efeitos de grupo, ou seja, diferenças estruturais entre subpopulações que compartilham determinadas características observáveis. Já, do ponto de vista prático, o uso dessas variáveis facilita a interpretação dos resultados, pois permite associar diretamente o comportamento de persistência a perfis identificáveis de segurados.

A Tabela 6 apresenta um resumo das variáveis categóricas, destacando o suporte em que cada uma delas está definida.

Tabela 6: Resumo das variáveis categóricas.

Variável	**Tipo**	**Suporte**
gender	Categórica	{1, 2}
premium.frequency	Categórica	{1, 2, 3}
risk.state	Categórica	{1, 2}
underwriting.age	Categórica	{1, 2, 3}
living.place	Categórica	{1, 2, 3}
acc.death.rider	Categórica	{1, 2}

Será realizada, inicialmente, uma análise exploratória dos dados. Em um primeiro

momento, será examinada a proporção de censuras e resgates, por meio da comparação entre os contratos encerrados por resgate e aqueles que permaneceram ativos até o final do período de observação.

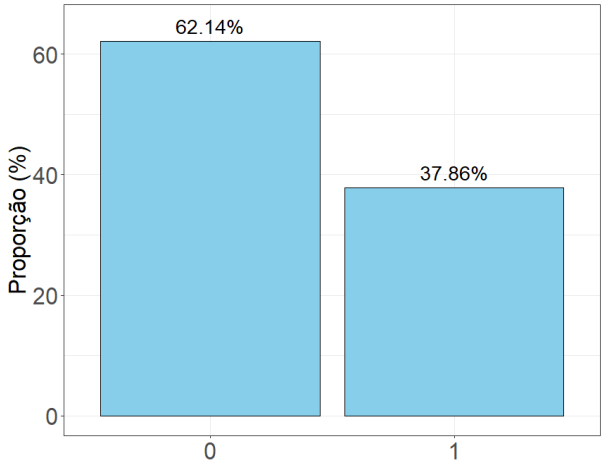
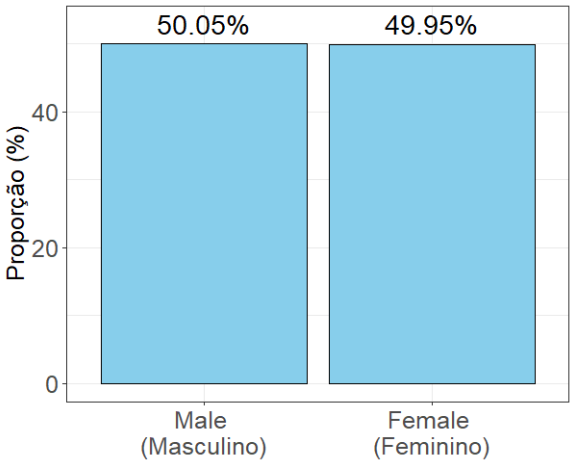


Figura 9: Proporção de resgates (*surrender*) e censuras (contratos ativos ou sinistros).

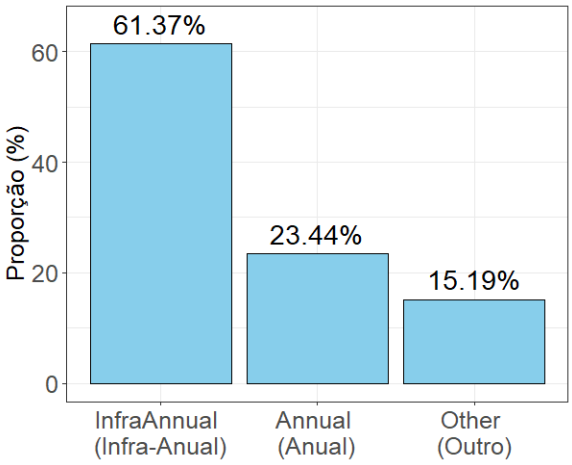
A Figura 9 apresenta a distribuição das observações entre censura (0) e resgate (1) na base de dados analisada. Observa-se que 62,14% das apólices permaneceram vigentes até o final do período de observação, sem que o evento de interesse (resgate) ocorresse, sendo, portanto, classificadas como censuradas. Por outro lado, apenas 37,86% das apólices foram efetivamente encerradas por resgate.

A seguir, considerando a decisão de empregar exclusivamente variáveis categóricas na modelagem com fração de cura, mencionada anteriormente, será realizada uma análise exploratória voltada à identificação de padrões de comportamento entre os diferentes grupos de segurados. O objetivo é compreender como as categorias dessas variáveis se associam à decisão de manter ou resgatar a reserva constituída.

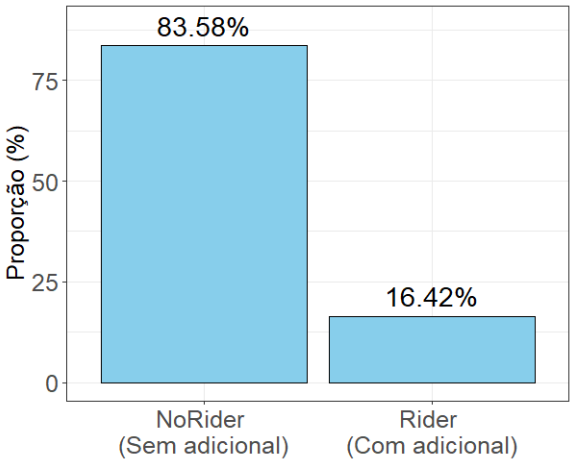
Inicialmente, serão apresentadas as proporções das observações entre as categorias disponíveis, o que ajuda a identificar possíveis desbalanceamentos e a entender melhor o perfil dos segurados que compõem a carteira.



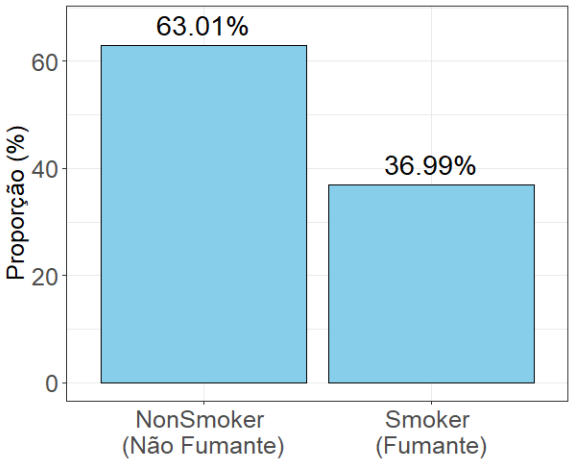
(a) *gender*



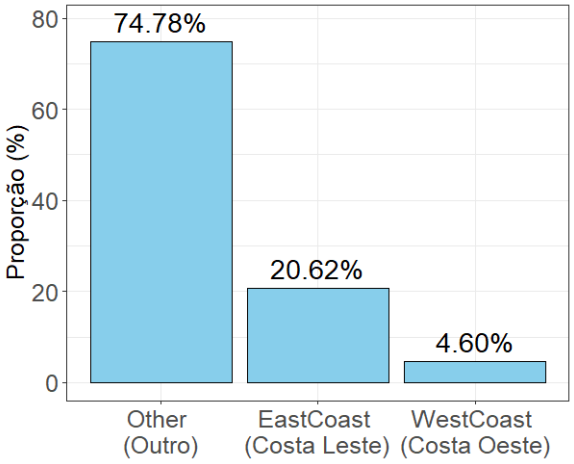
(b) *premium.frequency*



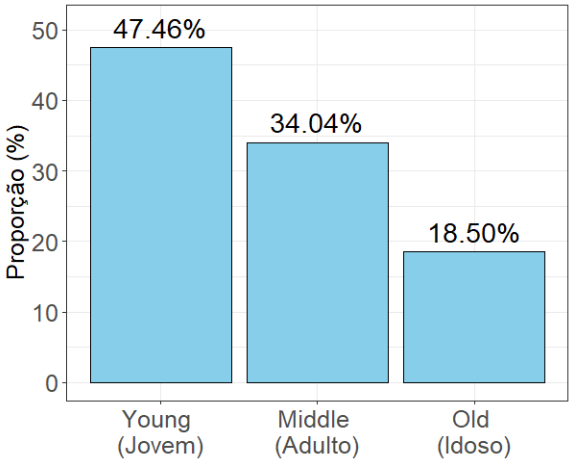
(c) *acc.death.rider*



(d) *risk.state*



(e) *living.place*



(f) *underwriting.age*

Figura 10: Proporções por categoria das variáveis categóricas.

A Figura 10a indica que a distribuição por gênero é equilibrada, indicando ausência

de viés relevante nesse aspecto. Quanto à frequência de pagamento dos prêmios (Figura 10b), observa-se predominância do pagamento infra-anual, seguido do anual e outras modalidades, o que sugere diversidade nos perfis financeiros e de planejamento dos clientes.

Em relação às coberturas adicionais (Figura 10c), a maioria não possui rider vinculado à apólice, o que pode refletir tanto custo adicional quanto baixa percepção de necessidade. Já no que diz respeito ao estado de risco associado ao tabagismo (Figura 10d), grande parte dos segurados é não fumante.

A distribuição geográfica dos clientes (Figura 10e) indica que a maioria reside fora das regiões costeiras, com uma parcela concentrada na costa leste e outra na costa oeste, o que pode refletir a abrangência da atuação comercial da seguradora. Por fim, a análise etária no momento da subscrição (Figura 10f) mostra uma predominância de segurados jovens, seguida por adultos e, em menor proporção, idosos.

Por fim, será analisado como a persistência se comporta ao longo do tempo em cada grupo, utilizando o estimador de Kaplan-Meier para a função de sobrevivência e o estimador empírico para a função de risco tal como apresentado na subseção 2.5. Esse tipo de análise pode indicar a presença de subgrupos com maior tendência a manter o contrato até o fim da vida — justamente o que se espera capturar com o modelo de fração de cura.

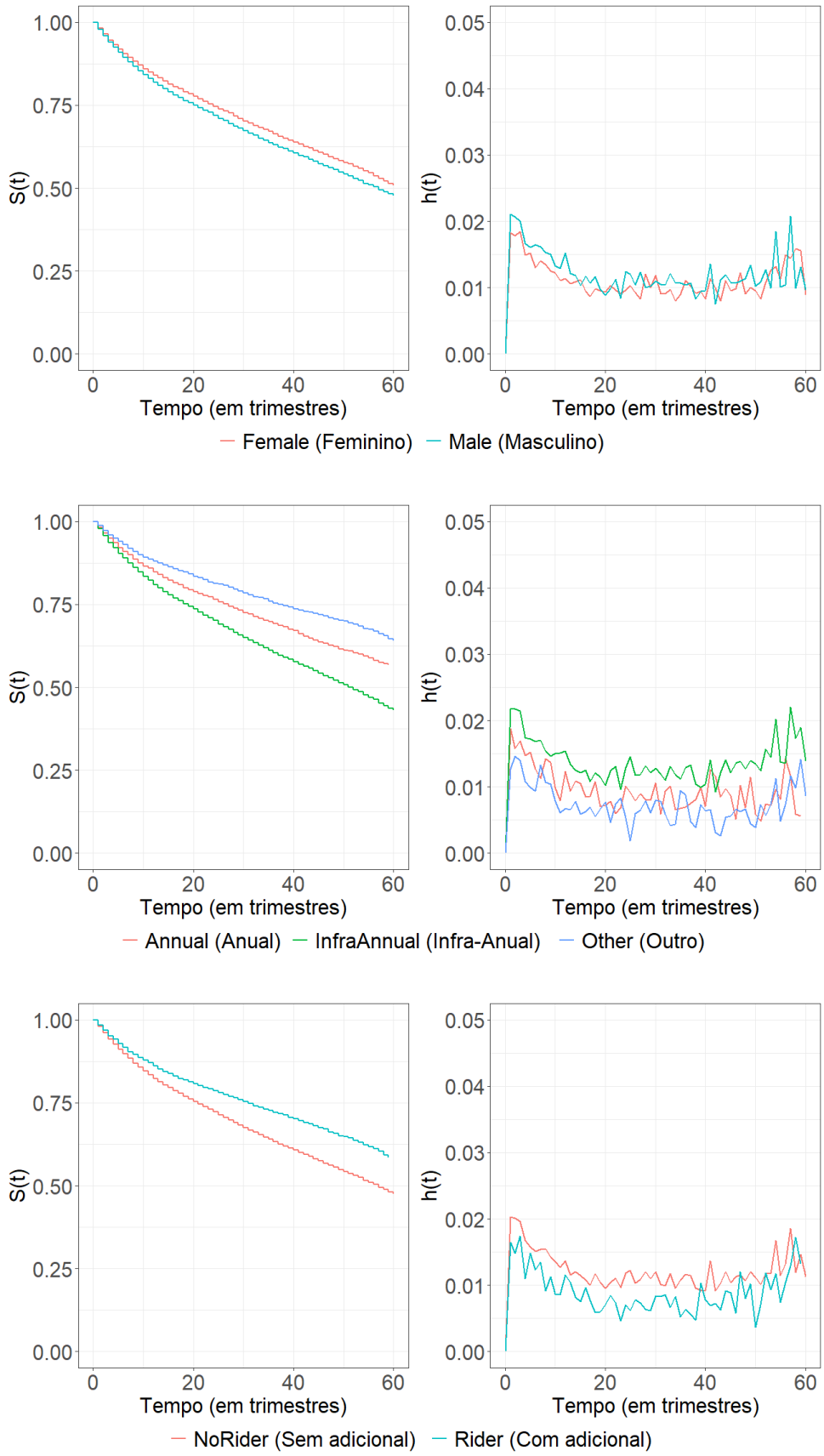


Figura 11: Funções de sobrevivência (esquerda) e de risco (direita) para as variáveis categóricas *gender*, *premium.frequency* e *acc.death.rider*, respectivamente.

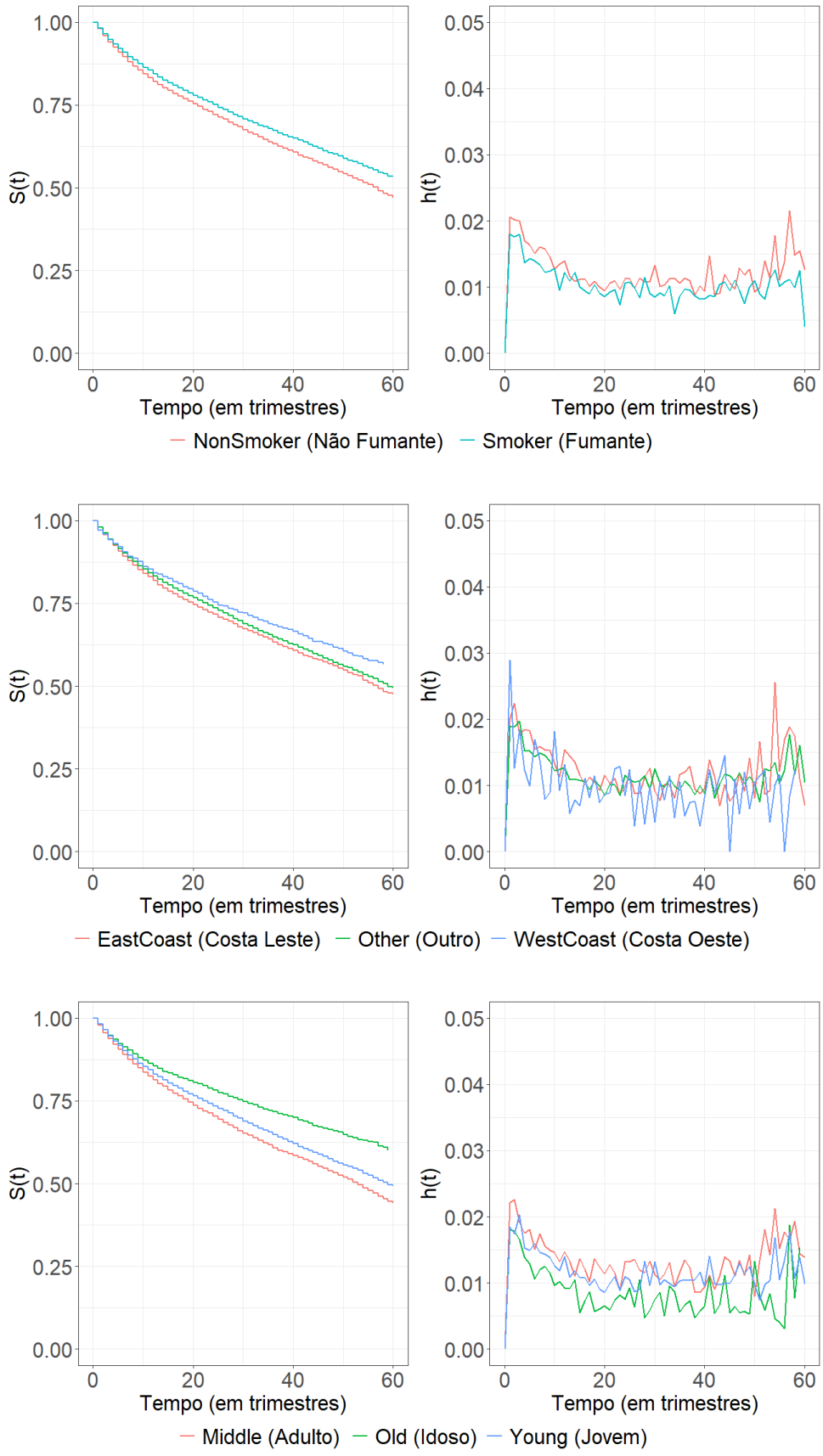


Figura 12: Funções de sobrevivência (esquerda) e de risco (direita) para as variáveis categóricas *risk.state*, *living.place* e *underwriting.age*, respectivamente.

A análise das funções de sobrevivência e de risco associadas às variáveis categóricas revela que, para a variável *premium.frequency*, os clientes que pagam de forma anual apresentam maior persistência ao longo do tempo em comparação àqueles que realizam pagamentos com frequência infra-anual, por exemplo (Figura 11).

De modo semelhante, a variável *risk.state* também apresenta distinções importantes: os não fumantes tendem a cancelar seus contratos com maior rapidez do que os fumantes, conforme evidenciado pela curva de sobrevivência (Figura 12).

Além disso, a variável *underwriting.age* demonstra forte influência na permanência dos segurados. Indivíduos mais jovens tendem a ter menor persistência, enquanto os mais velhos permanecem por períodos mais longos, possivelmente refletindo diferenças nos perfis de compromisso ou na percepção de necessidade do seguro ao longo das faixas etárias (Figura 12).

No que diz respeito à modelagem, em um contexto de modelos com fração de cura, é essencial distinguir entre as covariáveis que afetam a probabilidade de nunca resgatar a reserva de uma apólice de seguro de vida inteira (ou seja, pertencer ao grupo curado) e aquelas que influenciam o tempo até o resgate entre os indivíduos suscetíveis. A seguir, são apresentadas as covariáveis escolhidas para cada parte do modelo, bem como suas respectivas justificativas.

Para a modelagem da fração de cura, foram escolhidas as covariáveis relacionadas à frequência de pagamento (*premium.frequency*) e à idade de subscrição (*underwriting.age*). A covariável *premium.frequency*, que representa a frequência de pagamento dos prêmios, mostra-se relevante para explicar a proporção de "curados". Segurados que optam por pagar em intervalos maiores, como anualmente, tendem a apresentar maior planejamento e estabilidade financeira. Isso sugere um compromisso mais forte com o longo prazo e, portanto, uma menor chance de resgate antecipado, sendo mais prováveis de pertencer ao grupo imune.

A inclusão da covariável *underwriting.age*, que representa a idade do segurado no momento da subscrição, se justifica pelas diferenças de comportamento que tendem a existir entre faixas etárias. Indivíduos mais jovens, por exemplo, podem apresentar maior propensão ao resgate antecipado, seja por instabilidade financeira, mudanças de prioridades ou menor percepção da importância de um seguro de longo prazo. Por outro lado, segurados mais idosos podem demonstrar maior valorização da proteção contratada, optando por manter a apólice ativa por mais tempo e, consequentemente, apresentando maior probabilidade de pertencer ao grupo curado.

Para a modelagem do tempo até o resgate entre os suscetíveis, optou-se por considerar as covariáveis *risk.state* e *acc.death.rider*. A covariável *risk.state*, que indica se o segurado é fumante ou não, mostra-se relevante nesse contexto, pois o comportamento as-

sociado ao tabagismo pode refletir diferentes perfis de risco e distintas atitudes em relação à manutenção do contrato de seguro.

A covariável *acc.death.rider*, que indica a existência de uma cobertura adicional por morte acidental, pode também impactar o tempo até o evento. Por tornar a apólice mais cara devido à proteção adicional, essa cobertura pode levar alguns segurados a reconsiderarem sua permanência no contrato, especialmente em situações de pressão financeira. Dessa forma, a presença dessa cláusula pode estar associada a uma antecipação do resgate entre os indivíduos suscetíveis.

Essa distinção entre os fatores que influenciam a chance de pertencer ao grupo curado e os que impactam o tempo até o resgate permite ao modelo capturar com mais precisão a heterogeneidade no comportamento dos segurados ao longo do tempo.

A escolha de manter essas covariáveis constantes tem como objetivo isolar os efeitos provocados por diferentes especificações na fração de cura, possibilitando uma avaliação mais clara do impacto dessa componente no ajuste do modelo.

A principal distinção entre as modelagens está na forma como a fração de cura é modelada. Na Modelagem 1, considera-se um modelo padrão de sobrevivência, sem a presença da fração de cura, assumindo que todos os indivíduos eventualmente resgatarão a reserva. Na Modelagem 2, introduz-se uma fração de cura modelada apenas com o intercepto, ou seja, todos os indivíduos possuem a mesma probabilidade de pertencer ao grupo de curados, independentemente de suas características individuais.

Na Modelagem 3, incorpora-se a covariável *premium.frequency* na modelagem da fração de cura, permitindo capturar variações nessa probabilidade associadas à frequência de pagamento do prêmio. Por fim, a Modelagem 4 estende essa especificação ao incluir também a covariável *underwriting.age*, considerando que a idade de subscrição pode influenciar a probabilidade de pertencer ao grupo de curados.

As curvas de sobrevivência e de risco para as Modelagens 1 a 4 são apresentadas nas Figuras 13 a 16. Os parâmetros estimados do modelo para cada uma dessas modelagens estão detalhados nas Tabelas 7 a 10.

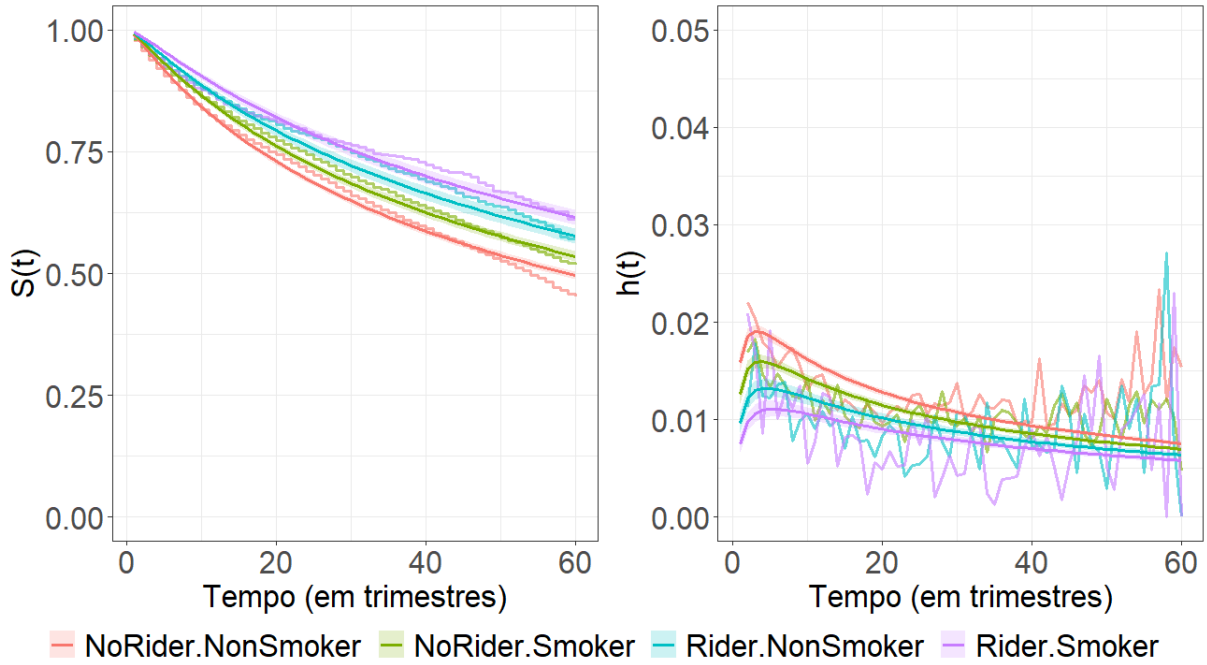


Figura 13: Modelagem 1 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 7: Modelagem 1 - Ajuste do modelo tradicional.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	4,08	(4,04 ; 4,11)
Rider (α_2)	0,36	(0,29 ; 0,43)
Smoker (α_3)	0,17	(0,13 ; 0,22)
σ	1,77	(1,74 ; 1,79)

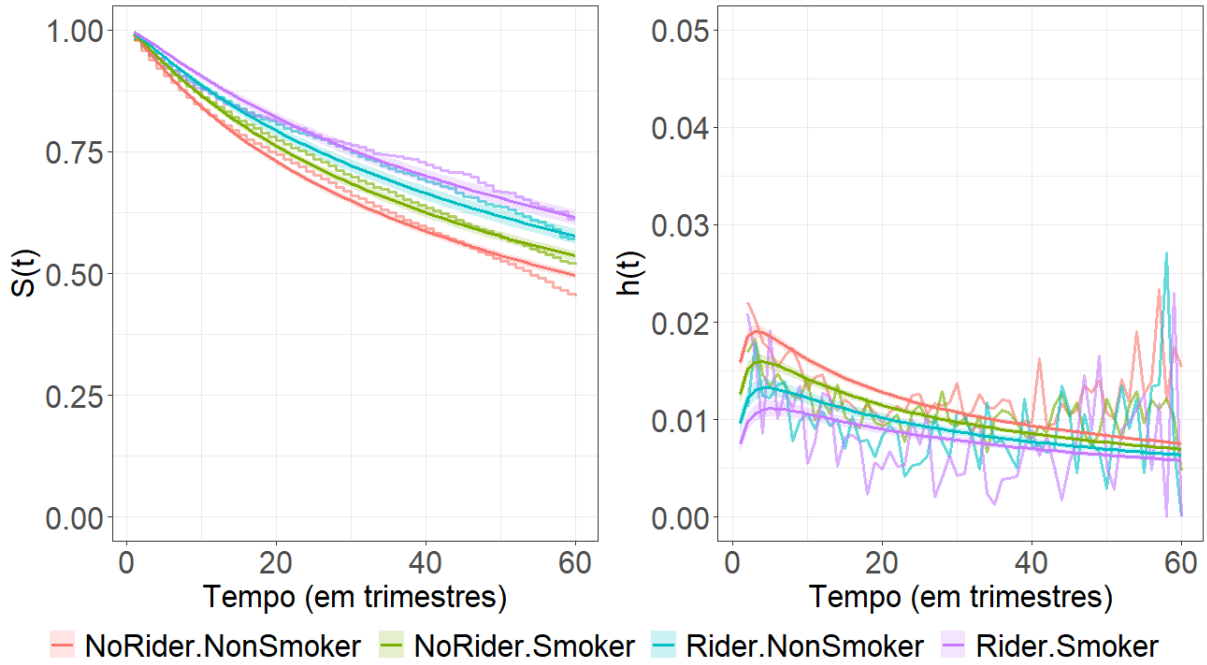


Figura 14: Modelagem 2 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 8: Modelagem 2 - Ajuste do modelo de fração de cura apenas com o intercepto.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	4,08	(4,04 ; 4,11)
Rider (α_2)	0,36	(0,29 ; 0,43)
Smoker (α_3)	0,17	(0,12 ; 0,23)
Intercepto (β_1)	-8,05	(-13,47 ; -4,88)
σ	1,77	(1,74 ; 1,79)
Grupo	**Fração de Cura**	**IC 95%**
NoRider.NonSmoker	0,001	(0,000 ; 0,008)
Rider.NonSmoker	0,001	(0,000 ; 0,008)
NoRider.Smoker	0,001	(0,000 ; 0,008)
Rider.Smoker	0,001	(0,000 ; 0,008)

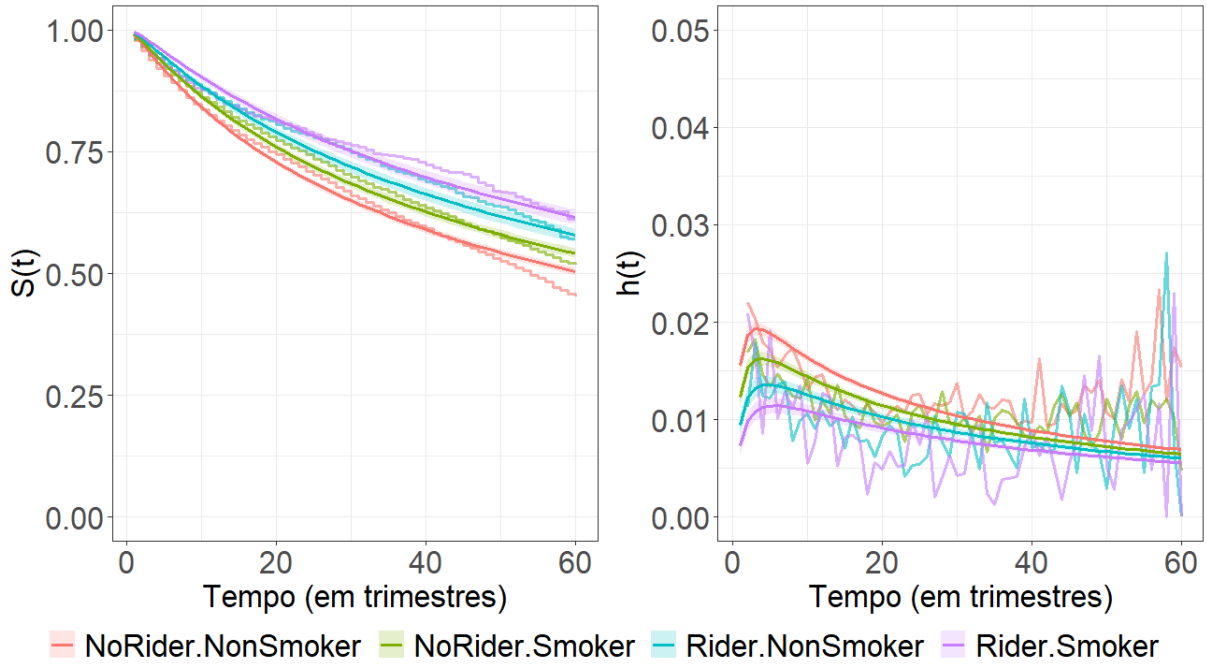


Figura 15: Modelagem 3 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 9: Modelagem 3 - Ajuste do modelo de fração de cura com uma covariável.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	3,87	(3,83 ; 3,91)
Rider (α_2)	0,35	(0,28 ; 0,42)
Smoker (α_3)	0,17	(0,11 ; 0,22)
Intercepto (β_1)	-6,77	(-9,80 ; -4,78)
Annual (β_2)	5,36	(3,40 ; 8,38)
Other (β_3)	6,22	(4,25 ; 9,21)
σ	1,69	(1,67 ; 1,72)
Grupo	**Fração de Cura**	**IC 95%**
NoRider.NonSmoker	0,102	(0,092 ; 0,112)
Rider.NonSmoker	0,103	(0,094 ; 0,113)
NoRider.Smoker	0,106	(0,096 ; 0,116)
Rider.Smoker	0,109	(0,099 ; 0,120)

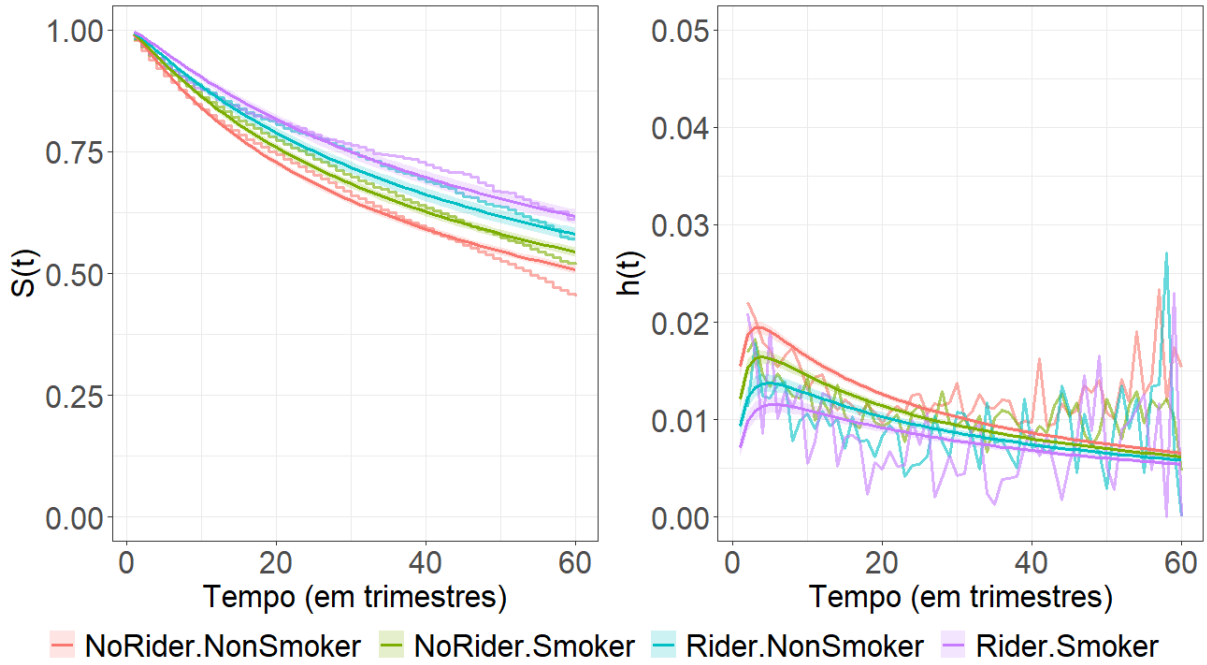


Figura 16: Modelagem 4 - Comparação das curvas de persistência (esquerda) e de risco (direita), mostrando a empírica em degraus (Kaplan-Meier) e a ajustada em linha contínua (média a posteriori), com seu intervalo de credibilidade de 95% (área sombreada).

Tabela 10: Modelagem 4 - Ajuste do modelo de fração de cura com duas covariáveis.

Coeficientes	**Média**	**IC 95%**
Intercepto (α_1)	3,77	(3,71 ; 3,83)
Rider (α_2)	0,35	(0,28 ; 0,42)
Smoker (α_3)	0,17	(0,12 ; 0,22)
Intercepto (β_1)	-2,86	(-3,43 ; -2,44)
Annual (β_2)	1,72	(1,35 ; 2,24)
Other (β_3)	2,47	(2,08 ; 3,00)
Middle (β_4)	-0,70	(-0,97 ; -0,45)
Old (β_5)	0,77	(0,55 ; 0,99)
σ	1,66	(1,63 ; 1,69)
Grupo	**Fração de Cura**	**IC 95%**
NoRider.NonSmoker	0,148	(0,125 ; 0,170)
Rider.NonSmoker	0,150	(0,127 ; 0,173)
NoRider.Smoker	0,151	(0,129 ; 0,174)
Rider.Smoker	0,156	(0,134 ; 0,179)

A Figura 17 apresenta as densidades posteriores estimadas para as frações de cura em cada grupo de segurados, considerando a Modelagem 4, que incorpora múltiplas covariáveis no componente de cura.

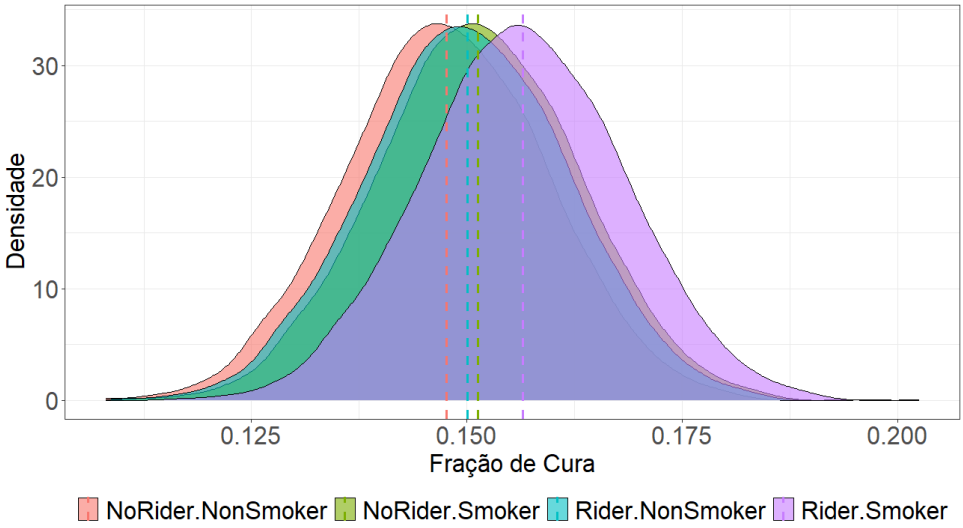


Figura 17: Modelagem 4 - Densidade posterior da fração de cura por grupo.

Nas quatro modelagens aplicados à base referente às apólices de seguro de vida inteira, observa-se uma evolução na complexidade do modelo e na forma como ele tenta capturar tanto o tempo até o evento quanto a presença de indivíduos que nunca resgatarão a reserva constituída. Diferentemente do estudo simulado, em que a melhoria do ajuste era visualmente perceptível nos gráficos gerados para cada modelagem, nesta aplicação a distinção entre os modelos nem sempre se manifesta de forma evidente nos gráficos, podendo estar relacionado possivelmente à natureza dos dados ou à ausência de variáveis preditoras fortes para a cura. Ainda assim, incorporar progressivamente a fração de cura e as variáveis explicativas relevantes permite uma modelagem mais adequada à possível heterogeneidade da população segurada.

A Tabela 7 apresenta os parâmetros estimados na Modelagem 1, que assume um modelo de sobrevivência usual, em que todos os indivíduos são considerados suscetíveis ao evento. Os coeficientes positivos associados às variáveis *Rider* (0,36) e *Smoker* (0,17) indicam um aumento no tempo esperado até o evento para esses grupos, ou seja, menor risco de ocorrência, mesmo na ausência de uma fração de cura.

Na Modelagem 2, apresentada na Tabela 8, é incorporada uma fração de cura constante. Os coeficientes de *Rider* e *Smoker* se mantêm inalterados, sugerindo robustez nos efeitos sobre o tempo até o evento entre os suscetíveis. No entanto, a fração de cura estimada para todos os grupos é extremamente baixa (cerca de 0,1%), o que indica que, embora o modelo permita essa estrutura, ele não identifica presença significativa de indivíduos curados nessa população.

A Tabela 9 apresenta os resultados da Modelagem 3, na qual a fração de cura é modelada em função da variável *premium.frequency*. O tempo até o evento segue afetado positivamente pelas variáveis *Rider* (0,35) e *Smoker* (0,17), enquanto os coeficientes *Annual* (5,36) e *Other* (6,22) no componente da cura indicam que esses métodos de pagamento estão fortemente associados a uma maior probabilidade de cura. De fato, as frações de cura estimadas variam entre 10,2% e 10,9%, apresentando aumento significativo em relação a Modelagem 2, o que reforça a importância de incluir covariáveis explicativas no modelo de cura.

Por fim, na Modelagem 4, representada pela Tabela 10, a modelagem da fração de cura é expandida com a inclusão adicional da covariável *underwriting.age*. Novamente, os efeitos de *Rider* e *Smoker* no tempo até o evento permanecem consistentes. No componente de cura, observa-se que as categorias *Annual* e *Other* continuam associadas a maiores chances de cura, e que a categoria *Old* (coeficiente positivo) está relacionada a maior probabilidade de nunca experimentar o evento, enquanto *Middle* apresenta um efeito negativo. As frações de cura estimadas variam entre 14,8% e 15,6%. Complementando esses valores pontuais, a Figura 17 oferece uma representação visual da incerteza associada e da magnitude das diferenças entre os grupos. O gráfico evidencia uma separação clara entre as distribuições posteriores das frações de cura, destacando que o perfil *Rider.Smoker* apresenta a maior fração média estimada.

A Tabela 11 apresenta os valores dos critérios de informação WAIC e LOO-CV obtidos para cada uma das modelagens aplicadas à base de dados de apólices de seguro.

Tabela 11: Comparação do ajuste dos modelos considerando os critérios de informação WAIC e LOO-CV, respectivamente.

Modelagem	**WAIC**	**LOO-CV**
1	113.884,2	113.884,2
2	113.884,9	113.884,9
3	113.461,7	113.461,8
4	113.382,8	113.382,8

Observa-se que as Modelagens 1 e 2 apresentam valores praticamente idênticos para ambos os critérios de informação, sugerindo que a introdução de uma fração de cura constante na Modelagem 2 não trouxe ganhos significativos em termos de ajuste. Por outro lado, as Modelagens 3 e 4, que incorporam covariáveis explicativas na modelagem da fração de cura, apresentaram reduções consideráveis nesses critérios, com destaque para a Modelagem 4, que obteve os menores valores. Isso indica que a inclusão de variáveis como *premium.frequency* e *underwriting.age* contribuiu para capturar melhor a heterogeneidade da população segurada, resultando em maior qualidade de ajuste e melhor capacidade preditiva em relação as modelagens anteriores.

6 Conclusão

Este trabalho teve como objetivo investigar a aplicabilidade de modelos de análise de sobrevivência com fração de cura com foco no contexto atuarial de seguros de vida, considerando a possibilidade de que uma parcela dos segurados jamais venha a experienciar o evento de interesse, que corresponde ao resgate da reserva de uma apólice. A introdução da fração de cura permite representar de forma mais realista esse comportamento, especialmente em produtos com características de longo prazo.

Ao longo do estudo, foram avaliadas diferentes modelagens, considerando tanto dados reais quanto simulados, com o intuito de explorar a estrutura e a interpretação dos parâmetros envolvidos. Concluímos que, para justificar a utilização de modelos com fração de cura, é fundamental que a curva de sobrevivência empírica apresente uma tendência clara de estabilização ao longo do tempo, formando um platô. Tal comportamento sugere que uma fração não desprezível da população está efetivamente "curada", isto é, não está sujeita ao evento de interesse mesmo em longos períodos de observação.

Contudo, observa-se que os dados reais utilizados apresentam uma limitação importante nesse sentido, isto é, mesmo após um longo período de dados observados e com alta persistência dos contratos no longo prazo, não é possível identificar claramente o comportamento assintótico das curvas de sobrevivência. Em razão disso, optamos por simular uma base de dados com características controladas, incluindo explicitamente uma fração de curados, o que permitiu avaliar o desempenho dos modelos propostos em um ambiente idealizado. Essa simulação também serviu como uma validação da metodologia, permitindo verificar se os modelos utilizados eram capazes de recuperar os parâmetros verdadeiros sob diferentes especificações.

Foram aplicadas diferentes versões de modelagem, incluindo a especificação sem fração de cura, a versão com fração constante e a modelagem com fração parametrizada por covariáveis. A comparação entre as modelagens foi conduzida com o uso dos critérios de informação WAIC e LOO-CV, cujos resultados reforçaram a superioridade dos modelos que incorporam a fração de cura de forma estruturada e dependente de covariáveis explicativas. Tais modelos não apenas apresentaram melhor ajuste, como também maior capacidade preditiva, especialmente no cenário simulado.

Além disso, foi dado destaque à interpretação dos coeficientes associados tanto ao tempo até o evento quanto à probabilidade de cura, permitindo extrair informações relevantes sobre o comportamento dos grupos analisados. A modelagem com fração de cura mostrou-se uma ferramenta alternativa para estudos atuariais, ao possibilitar a diferenciação entre indivíduos suscetíveis e não suscetíveis ao evento, o que é de particular interesse para a precificação, subscrição e avaliação de risco em carteiras de seguros de vida.

Como perspectivas para trabalhos futuros, destacamos três direções principais. A primeira é a extensão dos modelos para lidar com múltiplos desfechos, o que refletiria com maior fidelidade as possíveis razões de encerramento de contratos em contextos reais. A segunda é o desenvolvimento de um pacote completo em R voltado à modelagem de sobrevivência com fração de cura, incluindo funções para estimação, diagnóstico, comparação de modelos e visualizações. Nesse pacote, pretende-se realizar a implementação própria dos algoritmos de inferência, aproveitando o fato de que as distribuições condicionais completas do modelo são conhecidas, o que permitiria evitar o uso de ferramentas como o `Stan` como uma "caixa preta" e, assim, oferecer maior controle e transparência sobre os procedimentos computacionais adotados. Por fim, gostaríamos de propor o uso de modelos baseados em mistura de distribuições Log-Normal para representar os tempos de sobrevivência entre os suscetíveis, conforme abordado por Lobo, Fonseca e Alves (2024), permitindo uma modelagem mais flexível da heterogeneidade observada nesse subgrupo da população.

Os códigos utilizados nas análises e simulações deste projeto estão disponíveis publicamente no repositório GitHub, com o objetivo de garantir a reprodutibilidade dos resultados e incentivar a aplicação da abordagem por outros estudantes, pesquisadores e profissionais da área.

A Apêndice

A.1 Distribuições condicionais completas

As distribuições condicionais completas dos parâmetros e das variáveis latentes, são:

Para $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta} \mid \cdot) \propto \prod_{i=1}^{n} \left[p(\mathbf{z}_i)^{1-U_i}(1-p(\mathbf{z}_i))^{U_i}\right] \times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right), \quad (26)$$

Para $\boldsymbol{\alpha}$:

$$\pi(\boldsymbol{\alpha} \mid \cdot) \propto \prod_{i=1}^{n} \left[f(t_i \mid \mu_i, \sigma^2)^{U_i\delta_i} S(t_i \mid \mu_i, \sigma^2)^{U_i(1-\delta_i)}\right] \times \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\right), \quad (27)$$

Para σ^2:

$$\pi(\sigma^2 \mid \cdot) \propto \prod_{i=1}^{n} \left[f(t_i \mid \mu_i, \sigma^2)^{U_i\delta_i} S(t_i \mid \mu_i, \sigma^2)^{U_i(1-\delta_i)}\right] \times (\sigma^2)^{-(a_\sigma+1)} \exp\left(-\frac{b_\sigma}{\sigma^2}\right), \quad (28)$$

Para U_i, $i = 1, \ldots, n$:

$$\mathbb{P}(U_i = 1 \mid \mathbf{z}_i, t_i, \delta_i) = \frac{(1-p(\mathbf{z}_i)) \times f(t_i \mid \mu_i, \sigma^2)^{\delta_i} \times S(t_i \mid \mu_i, \sigma^2)^{1-\delta_i}}{p(\mathbf{z}_i) + (1-p(\mathbf{z}_i)) \times f(t_i \mid \mu_i, \sigma^2)^{\delta_i} \times S(t_i \mid \mu_i, \sigma^2)^{1-\delta_i}}, \quad (29)$$

$$\mathbb{P}(U_i = 0 \mid \mathbf{z}_i, t_i, \delta_i) = 1 - \mathbb{P}(U_i = 1 \mid \mathbf{z}_i, t_i, \delta_i). \quad (30)$$

A.2 Trajetória das cadeias MCMC

A.2.1 Estudo simulado

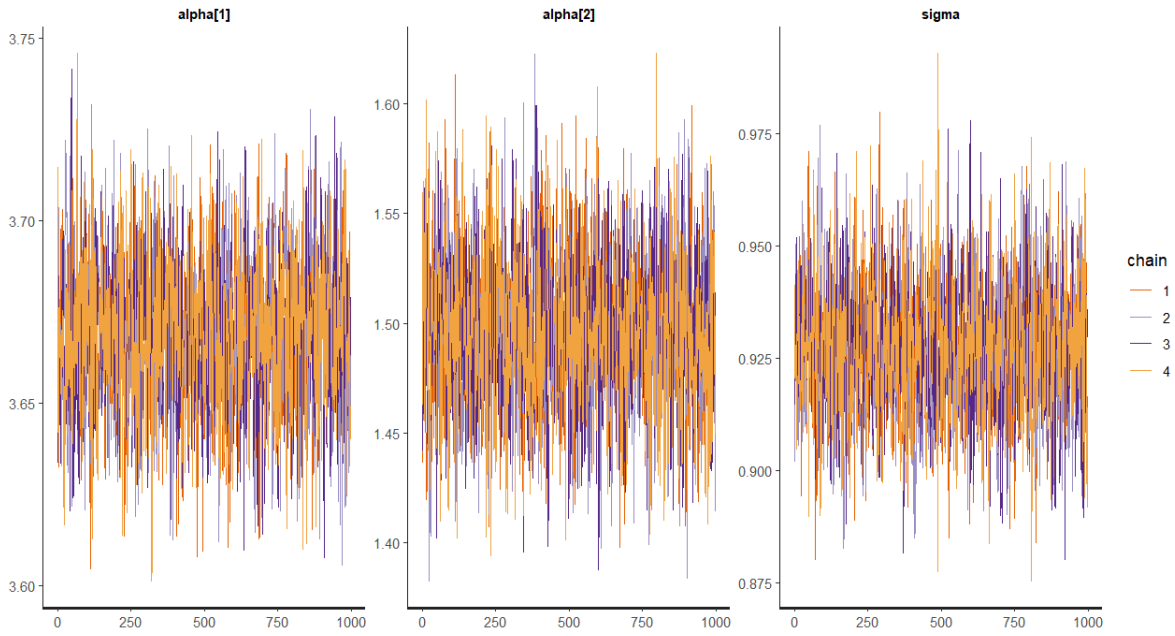


Figura 18: Modelagem 1 - Comportamento das cadeias.

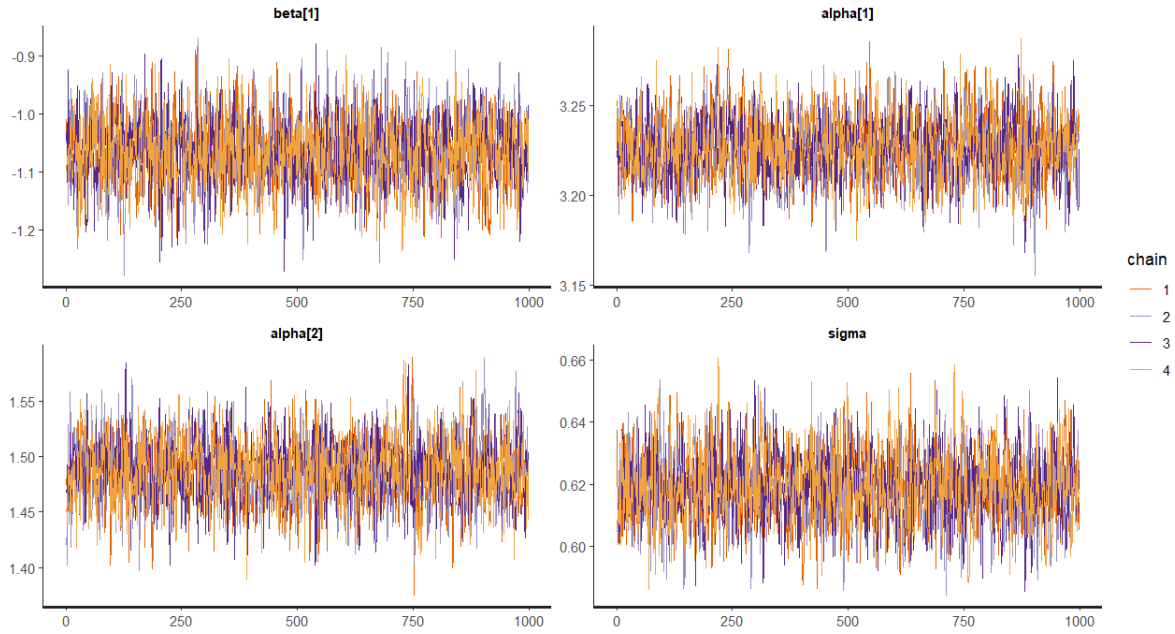


Figura 19: Modelagem 2 - Comportamento das cadeias.

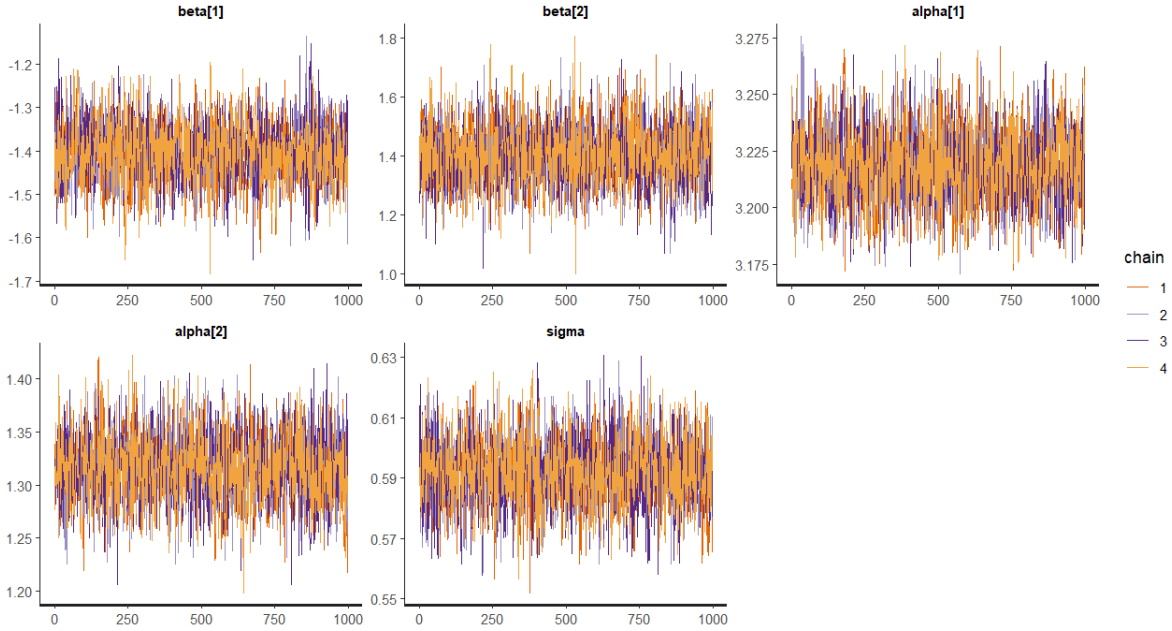


Figura 20: Modelagem 3 - Comportamento das cadeias.

A.2.2 Análise referente a seguros de vida inteira

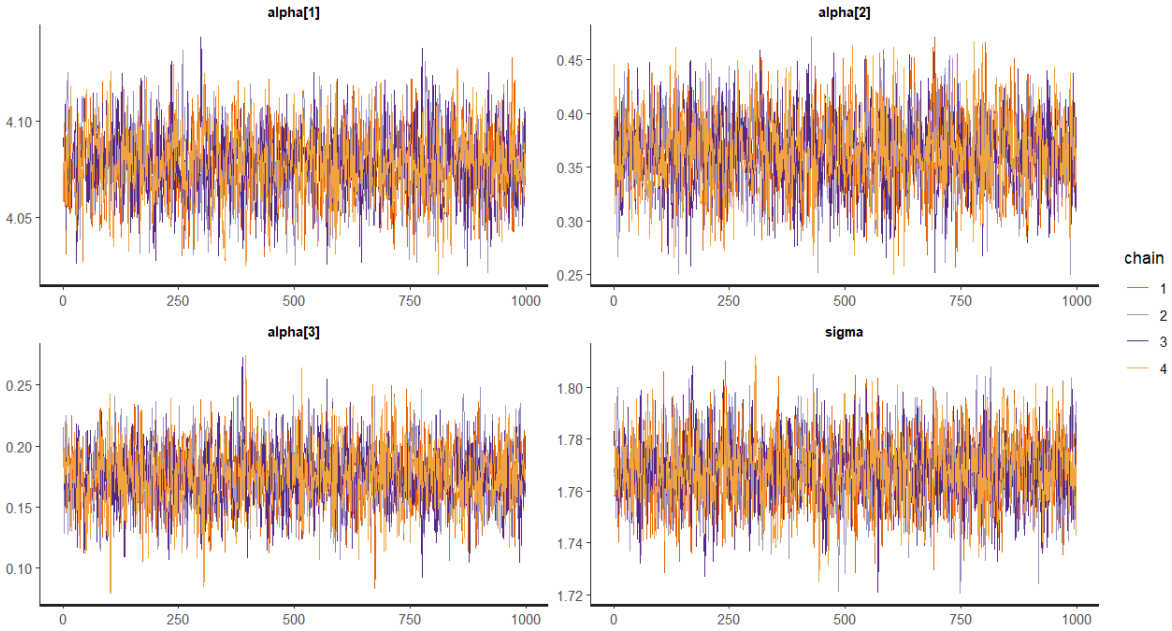


Figura 21: Modelagem 1 - Comportamento das cadeias.

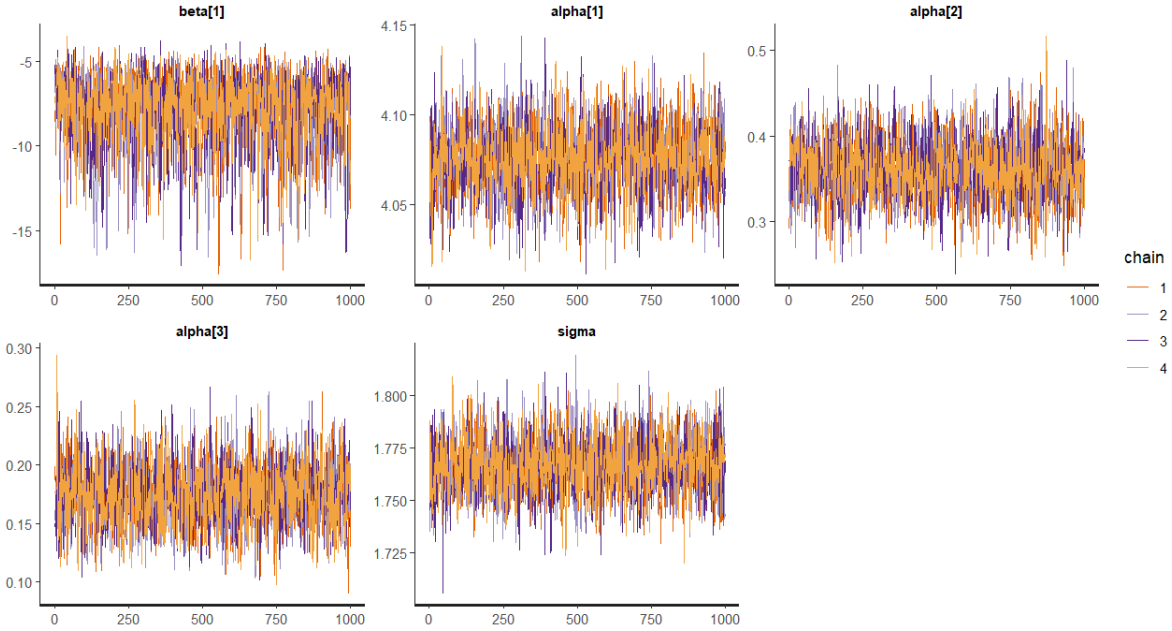


Figura 22: Modelagem 2 - Comportamento das cadeias.

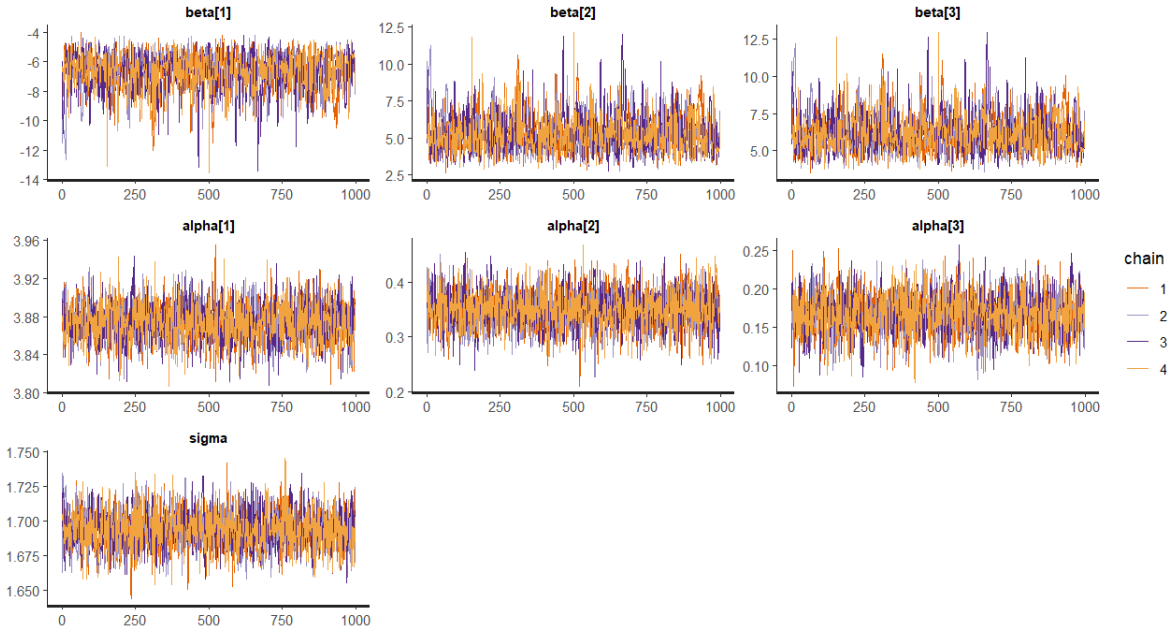


Figura 23: Modelagem 3 - Comportamento das cadeias.

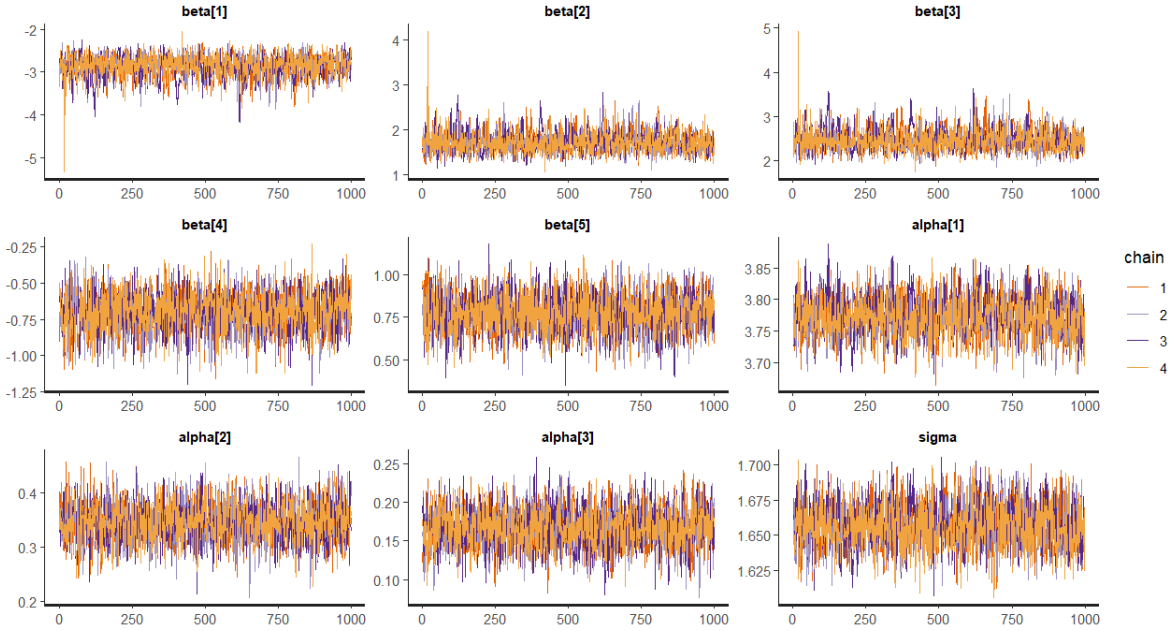


Figura 24: Modelagem 4 - Comportamento das cadeias.

Referências

AMICO, M.; KEILEGOM, I. V. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, v. 5, p. 311–342, 2018. Disponível em: ⟨https://doi.org/10.1146/annurev-statistics-031017-100101⟩.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v. 47, n. 259, p. 501–515, 1952. Disponível em: ⟨https://doi.org/10.1080/01621459.1952.10501187⟩.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 11, n. 1, p. 15–44, 1949. Disponível em: ⟨https://doi.org/10.1111/j.2517-6161.1949.tb00020.x⟩.

BOWERS, N. L. et al. *Actuarial Mathematics*. 2. ed. [S.l.]: Society of Actuaries, 1997. 753 p. (The Society of Actuaries Textbook Series). ISBN 9780938959465.

BROOKS, S. et al. *Handbook of Markov Chain Monte Carlo*. [S.l.]: Chapman and Hall/CRC, 2011.

CARVALHO, S. M. *Análise de Sobrevivência: Teoria e aplicações em saúde*. [S.l.]: FIOCRUZ, 2011.

COLOSIMO, E.; GIOLO, S. *Análise de Sobrevivência Aplicada*. [S.l.]: Blucher, 2006.

CRUZ, R. de la; FUENTES, C.; PADILLA, O. A Bayesian mixture cure rate model for estimating short-term and long-term recidivism. *Entropy*, v. 25, n. 1, p. 56, 2022. Disponível em: ⟨https://doi.org/10.3390/e25010056⟩.

DUTANG, C.; CHARPENTIER, A.; GALLIC, E. *Insurance dataset*. Recherche Data Gouv, 2024. Disponível em: ⟨https://doi.org/10.57745/P0KHAG⟩.

GELMAN, A. et al. *Bayesian Data Analysis*. 3. ed. [S.l.]: Chapman and Hall/CRC, 2013. 675 p. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781439840955.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457–481, 1958. Disponível em: ⟨https://doi.org/10.2307/2281868⟩.

KLEINBAUM, G. D.; KLEIN, M. *Survival Analysis: A Self-Learning Text*. [S.l.]: Springer, 2011.

LAWLESS, F. J. *Statistical Models and Methods for Lifetime Data*. [S.l.]: Wiley, 2022.

LOBO, V. G.; FONSECA, T. C.; ALVES, M. B. Lapse risk modeling in insurance: a Bayesian mixture approach. *Annals of Actuarial Science*, v. 18, p. 126–151, 2024.

MALLER, R. A.; ZHOU, X. *Survival Analysis with Long-Term Survivors*. [S.l.]: Wiley, 1996. 278 p. (Wiley Series in Probability and Statistics). ISBN 9780471962014.

MARTINEZ, E. Z. et al. Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer Methods and Programs in Biomedicine*, v. 112, n. 3, p. 343–355, 2013. Disponível em: ⟨https://doi.org/10.1016/j.cmpb.2013.07.021⟩.

RICHARDS, S. J. *Survival Models for Actuarial Work*. [S.l.], 2011. Disponível em: ⟨https://www.longevitas.co.uk/sites/default/files/Survival%20Models%20for%20Actuarial%20Work%20%281%29.pdf⟩.

STONE, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, v. 36, n. 2, p. 111–147, 1974.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, v. 27, n. 5, p. 1413–1432, 2017. Disponível em: ⟨https://doi.org/10.1007/s11222-016-9696-4⟩.

WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, v. 11, n. 116, p. 3571–3594, 2010. Disponível em: ⟨http://jmlr.org/papers/v11/watanabe10a.html⟩.