Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de regressão defeituosos zero ajustados aplicados a dados de risco de crédito

Crystiane Fernanda de Souza

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de regressão defeituosos zero ajustados aplicados a dados de risco de crédito

Crystiane Fernanda de Souza
Orientador(a): Profª Drª Vera Tomazella

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Setembro de 2023

Crystiane Fernanda de Souza

Modelos de regressão defeituosos zero ajustados aplicados a dados de risco de crédito

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Crystiane Fernanda de Souza e aprovado pela banca examinadora.

Aprovado em 18 de agosto de 2023.

Banca Examinadora:

- Profª. Drª. Vera Tomazella
- Prof. Dr. Jeremias da Silva Leão
- Prof. Dr. Carlos Alberto Ribeiro Diniz

Agradecimentos

Aos meus pais, Carmen Aparecida Ara de Souza e Valdeci Felisberto de Souza, por todo suporte que me deram para meu avanço educacional e profissional, por todo apoio, compreensão, paciência e incentivo.

Ao meu irmão, Anderson Luiz Ara Souza, por todo suporte, por todos os ensinamentos e por ter me incentivado a seguir nas Ciências Estatísticas.

Ao meu noivo, Fernando Luiz Nais Junior, pelo imenso apoio, pela paciência, carinho e compreensão durante todo esse período.

A minha orientadora, Vera Lucia Damasceno Tomazella, minha gratidão, pela amizade e por todo o conhecimento transferido.

E aos meus colegas de graduação, que se tornaram amigos para a vida, Isadora Almeida, Juliano Cesar, Matthews Martis e Moyses Tenório, que também me acompanharam por todo esse processo.

Se você pode sonhar, você pode conseguir.

(Zig Ziglar)

Resumo

Com o aumento do consumo de bens, serviços e concessão de crédito, torna-se necessário controlar o risco do processo. Essa medida visa evitar uma possível inadimplência maior do que a suportada pelas instituições financeiras e, ao mesmo tempo, possibilita a geração de lucros. Várias técnicas estatísticas podem ser utilizadas para a construção de modelos que apresentem o panorama de risco, sendo uma delas é a análise de sobrevivência. A aplicação dessa técnica no mercado financeiro busca estudar, por exemplo, o tempo que um indivíduo leva para recuperar um crédito após a finalização de uma crise financeira em seu país. A utilização desse tipo de dado pode embasar a previsão do valor de crédito ideal a ser provisionado nos possíveis cenários de crise e inferir em que prazo poderá ocorrer a retomada das operações de crédito. Neste contexto, este trabalho tem por objetivo estudar dois modelos de regressão defeituosos para modelagem de dados de sobrevivência zero ajustado no cenário de risco de crédito. Essa abordagem, permite que três tipos de unidades sejam acomodadas, como os clientes com tempos de sobrevida "zero", ou seja, falhas precoces, os clientes suscetíveis e não suscetíveis ao evento de interesse. A metodologia estudada será aplicada a uma base de dados fornecida por uma instituição líder em serviços e informações para crédito no Brasil.

Palavras-chave: *Análise de sobrevivência, Dados financeiros, Risco de crédito, Distribuição defeituosa, Modelo de longa duração, Fração de cura, Zero ajustado.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O risco pode ser definido como a volatilidade de eventos inesperados, como a representação do valor de ativos, patrimônios ou ganhos (Jorion, 2007). Nesse contexto, nas instituições financeiras uma operação de concessão de crédito é caracterizada como risco de crédito. Este tipo de risco é inerente a qualquer operação financeira, sendo definido como a possibilidade de inadimplência das obrigações contratuais por parte do devedor, que deixa de honrar o acordo estabelecido com o credor no momento da contratação. Assim, é de extrema importância que as instituições financeiras adotem medidas e procedimentos adequados para gerenciar o risco de crédito, garantindo a saúde financeira da instituição e a confiança do mercado e dos clientes.

A análise de crédito desempenha um papel crucial nas empresas, uma vez que é fundamental avaliar a capacidade financeira e a relação do indivíduo com o mercado para determinar a viabilidade da concessão de crédito. A análise leva em consideração a renda, histórico de crédito, entre outros fatores, a fim de evitar prejuízos financeiros para a instituição. Nesse sentido, a utilização de modelos de *credit scoring* mostra-se benéfica por permitir a consistência de decisões na análise de crédito, criar a automatização na concessão, aumento no valor de análises, capacidade de monitorar e administrar o risco de uma carteira de crédito, dentre outros. Além disso, de acordo com da Silva (2000), ao considerar a dinâmica que os cenários econômicos estão vinculados e a forma que afeta diretamente o risco de inadimplência, a decisão de concessão de crédito, baseada em modelos de riscos devem ser monitoradas e revisadas quando necessário.

A ocorrência de uma crise financeira no país é caracterizada por uma redução do nível de produção no país, resultando em uma série de impactos. Entre esses impactos, é possível destacar o endividamento da população, causado por diversos fatores, como o

aumento da inflação, o alto índice de desemprego e a restrição do acesso ao crédito. Em tal cenário, a retomada do sistema financeiro pode ser um processo lento e incerto, carecendo de previsões assertivas acerca do momento ideal de recuperação. Dessa maneira, torna-se essencial empregar modelos estatísticos, como a Análise de Sobrevivência, que é uma ferramenta que pode fornecer um importante suporte em tais circunstâncias.

A análise de sobrevivência é composta por um conjunto de técnicas e métodos estatísticos utilizados para estudar o tempo decorrido até a ocorrência de um evento interesse. O termo de análise de sobrevivência é habitualmente utilizado na área médica, onde pode-se caracterizar o tempo de falha como: a morte, a cura, o aparecimento de uma doença, o efeito colateral de um medicamento, entre outros. Entretanto, além do campo médico, a análise de sobrevivência pode ser aplicada em outras áreas, como o mercado financeiro.

A modelagem de fração de cura, também conhecida como modelagem de longa duração estuda casos em que, supostamente, existem observações não suscetíveis ao evento de interesse. Boag (1949) foi um dos pioneiros dentro da modelagem de longa duração. Posteriormente, outros modelos foram propostos, como o modelo de mistura padrão por Berkson e Gage (1952), o modelo unificado de fração de cura por Rodrigues *et al.* (2009), dentre outros. Neste tipo de modelagem, existem indivíduos que não são suscetíveis a ocorrência do evento de interesse, podendo ser considerados como indivíduos curados/imunes ao evento de interesse e o conjunto de dados de sobrevivência a que pertencem possui uma fração de cura. No mercado financeiro, o intuito é prever o tempo de recuperação de clientes, em que o recuperado é aquele cliente que retorna ao status de adimplência. O uso dos modelos de longa duração, no mercado financeiro, são considerados uma boa ferramenta para a estudar o tempo até a ocorrência do evento de interesse, como, o prazo de retorno até o status da adimplência ou a realização/atraso de uma parcela de empréstimo (Toledo *et al.*, 2022).

Dessa forma, aplicado no mercado financeiro, utiliza-se a análise de sobrevivência de longa duração com o intuito de estimar o tempo de um evento, como por exemplo, o prazo decorrido desde a aquisição de um empréstimo até o atraso de uma das parcelas, ou mesmo, como estudado por Granzotto *et al.* (2008),o início do relacionamento de um cliente com a instituição até a ruptura desse relacionamento. Para essa análise, foi aplicado o modelo proposto por Berkson e Gage (1952), utilizando-se os modelos Weibull e Log-logístico para modelar o tempo.

Contudo, em alguns estudos existem indivíduos que são suscetíveis a falhas precoces, os quais resultam em um tempo de sobrevivência igual ou próximo de zero. Nesse caso, este cenário será referido como ajustados para zero. Sendo assim, no contexto de fração de cura, os modelos defeituosos oferecem a estratégia para modelar dados de sobrevivência ajustados para zero. Embora alguns artigos já tenham utilizado a ideia de modelos defeituosos, Balka *et al.* (2011), Rocha *et al.* (2017), Scudilio *et al.* (2019) e Calsavara *et al.* (2019b) popularizaram recentemente o termo "defeituoso". Na literatura, existem diversas distribuições de probabilidade que possuem a forma defeituosa.

Nesse contexto, motivado pela necessidade de aplicar modelos de sobrevivência com a presença de excesso de zeros, em um cenário de risco de crédito, será utilizado neste trabalho uma abordagem proposta por Calsavara *et al.* (2019a) para acomodar uma proporção de falhas no tempo zero ou tempo de vida ajustado para zero, em que são considerados modelos defeituosos Gompertz e Gaussiana-Inversa, para estimar a função de sobrevivência com a possibilidade de taxa de cura e uma proporção de tempo de vida ajustado para zero.

1.1 Objetivos

O principal objetivo do Trabalho de Conclusão de Curso foi considerar uma abordagem proposta por Calsavara *et al.* (2019a), denominado "Modelos de regressão defeituosos zero ajustados" para análise de dados de risco de crédito no mercado financeiro. Essa abordagem, permite que três tipos de unidades sejam acomodadas, como os clientes com tempos de sobrevida "zero", ou seja, falhas precoces, os clientes suscetíveis e não suscetíveis ao evento de interesse. Para estimar a função de sobrevivência com a possibilidade de fração de cura e uma proporção de tempo de vida ajustado para zero consideramos os modelos defeituosos Gompertz e Gaussiana-Inversa.

O conjunto de dados utilizado na aplicação foi analisado por Toledo *et al.* (2022). Os dados foram concedidos por uma instituição financeira brasileira, a qual realiza serviços voltados ao mercado de crédito, contendo informações que envolvem características voltadas aos hábitos e costumes de indivíduos em torno de compromissos envolvendo solicitações de crédito.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte maneira. No Capítulo 2 será apresentado uma revisão da literatura com os conceitos básicos da análise de sobrevivência, metodologias de modelagem de longa duração e a modelagem defeituosa. No Capítulo 3 será introduzido o modelo defeituoso ajustados a zero. No Capítulo 4 será apresentado uma aplicação a dados financeiros baseado na metodologia proposta ao longo do trabalho. Por fim, o Capítulo 5 apresentará as conclusões obtidas ao longo do trabalho.

Capítulo 2

Revisão da Literatura

Neste capítulo será realizado uma revisão dos conceitos básicos da teoria de análise de sobrevivência, tais como o tipo de censura, funções básicas de sobrevivência e outras definições importantes ao decorrer do trabalho.

2.1 Conceitos básicos de Análise de Sobrevivência

A análise de sobrevivência ou confiabilidade, consiste em um conjunto de técnicas e métodos estatísticos, onde são utilizados para estudar o tempo até a ocorrência de um determinado evento de interesse, em geral denominado tempo de sobrevivência, vida ou falha. Além disso, essa área tem uma ampla literatura, com vários artigos e autores como Andersen *et al.* (1993); Hougaard (2000); Kalbfleisch e Prentice (2002); Aalen *et al.* (2008), dentre outros, que publicaram livros sobre o assunto com diversos tipos de abordagens.

A principal característica relacionada a estes dados, diz respeito à presença de observações incompletas. Comumente denominado por censura, esta ocorre por vários motivos, dentre elas, o abandono do tratamento, a saída do estudo por outros motivos além do estudado ou a não ocorrência da falha até o término do experimento. É importante observar, que mesmo sendo observações parciais, essas trazem alguma informação sobre o tempo até a falha e não devem ser omitidas na análise do problema.

2.1.1 Censura

Existem alguns tipos mais comuns de censura, encontrados na literatura que serão citados a seguir:

- **Censura do tipo I:** O estudo será terminado após um período pré-estabelecido de tempo, ou seja, quem não falha no tempo fixado é considerado como censura.
- **Censura do tipo II:** O estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos.
- **Censura do tipo aleatória:** Ocorre quando o indivíduo é retirado do estudo por algum motivo diferente da estudada, sem ter ocorrido a falha.

De acordo com Colosimo e Giolo (2006), os dados de sobrevivência para o i-ésimo indivíduo sob estudo são representados, em geral, pelo par, (t_i, δ_i), onde t_i é o tempo de falha ou censura e δ_i é a variável indicadora de falha ou censura, que é representada da seguinte forma,

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

2.1.2 Funções de Interesse

A função de densidade de probabilidade de T, pode ser interpretada como a probabilidade de um indivíduo experimentar um evento em um determinado intervalo de tempo,

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{2.1}$$

em que Δt é o incremento de tempo infinitamente pequeno.

Sua função de distribuição acumulada é dada por,

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{2.2}$$

A função de sobrevivência de um indivíduo é a probabilidade do indivíduo sobreviver por um período superior a t, ou seja, é a probabilidade de um indivíduo não falhar até um tempo t. É definida por,

$$S(t) = P(T \geq t) \ = \int_t^\infty f(u)du \ = 1 - F(t) \tag{2.3}$$

A função de sobrevivência segue as seguintes propriedades:

- $S(t)$ é não crescente;

- $S(0) = 1$;
- $\lim_{t\to\infty} S(t) = 0$.

A função de risco (ou função de taxa de falha) é a taxa instantânea de falha no tempo t, dado que o indivíduo estava vivo até o tempo t. Isto é,

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.4}$$

Nesse contexto, a função de risco pode apresentar diversos comportamentos, por exemplo, decrescente, constante, crescente e também a formas não monótonas, comumente conhecido por "curva da banheira", que representa a função de risco de morte dos seres humanos.

A função de risco acumulado $H(t)$ é importante em análises gráficas para verificar a adequação de modelos estatísticos e é definida por,

$$H(t) = \int_0^t h(u)du = -\log S(t). \tag{2.5}$$

Existem algumas relações matemáticas importantes entre as equações definidas anteriormente, sendo:

$$h(t) = \frac{f(t)}{S(t)} = \frac{d}{dt}(\log S(t)),$$

$$H(t) = -\log(S(t)),$$

$$S(t) = \exp\{-H(t)\}.$$

2.1.3 Estimador de Kaplan-Meier

O estimador Kaplan-Meier é o mais utilizado nos estudos clínicos e ganha cada vez mais espaço em estudos de confiabilidade. Esse estimador não-paramétrico foi proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, e também é conhecido por estimador de limite-produto. Suponha que existem n itens sob testes e $k(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \leq ... \leq t_k$. Ocasionalmente, pode ocorrer mais de uma falha simultaneamente, o que é chamado de empate. Dessa forma, se:

- d_i denota o número de falhas em t_i.

- n_i denota o número de itens sob risco em t_i, ou seja, denota o número de itens que não falhou e não foi censurado até o momento imediatamente anterior em t_i.

Sendo assim, o estimador de Kaplan-Meier de $S(t)$ é definido por:

$$\hat{S}(t) = \prod_{t_i:t_i \leq k} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{t_i:t_i \leq k} \left(1 - \frac{d_i}{n_i} \right). \tag{2.6}$$

A Expressão 2.6 é uma função escada nos tempos observados de falha.

2.1.4 Estimação por Máxima Verossimilhança

Na análise de sobrevivência, em diversos casos o interesse é estimar a função de sobrevivência, em que a suposição do tempo até o evento de interesse, definido como a variável aleatória T, segue uma determinada distribuição. Nesse contexto, ao supor que os dados seguem alguma distribuição, os parâmetros dessa distribuição serão desconhecidos. Logo, a ideia é obter o melhor conjunto de parâmetros da distribuição dos dados (Colosimo e Giolo, 2006).

No método de máxima verossimilhança existe a possibilidade da incorporação de censuras, além da ótima propriedade para amostras suficientemente grandes. Para dados sem censuras, conforme visto em Bolfarine e Sandoval (2001) a função de verossimilhança para uma amostra aleatória de tamanho n é definida por:

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

em que T é a variável aleatória que representa o tempo até o evento de interesse com função de densidade de probabilidade $f(t; \theta)$ e θ é o vetor de parâmetros.

Contudo, os dados censurados trazem informações importantes, uma vez que quando temos uma censura é possível verificar que o tempo até o evento de interesse do indivíduo é maior do que aquele em que foi censurado. Dessa forma, a contribuição para $L(\theta)$ é dada pela função de sobrevivência $S(t)$, assim, pode-se dividir as observações das amostras aleatórias em dois casos, as censuradas e não censuradas, sendo a função de verossimilhança com dados censurados dada por:

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i; \theta)^{\delta_i}][S(t_i; \theta)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i; \theta)]^{\delta_i} S(t_i; \theta), \tag{2.7}$$

em que δ_i será a variável indicadora de censura. Além disso, a Expressão 2.7 é válida para as censuras do tipo I, II e aleatória.

Os estimadores de máxima verossimilhança, basicamente são os valores de θ que maximizam os valores de $L(\theta)$, ou seja, $l(\theta) = \log|L(\theta)|$. Para encontrar esses estimadores, é necessário resolver o seguinte sistema de equações, sendo:

$$U(\theta) = \frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta} = 0.$$

Em geral, esse sistema não resulta em formas algébricas fechadas, devido a sua complexidade. Logo, é necessário o uso de métodos numéricos para realizar a estimação.

2.2 Modelos de Longa Duração

No contexto de análise de sobrevivência usual, em alguns casos, o evento de interesse é identificar o óbito de um paciente, o diagnóstico de uma doença, a cura de um indivíduo, o tempo de vida de um componente eletrônico, entre outros.

Contudo, existem determinadas situações em que uma parte da população não irá apresentar a ocorrência do evento de interesse, independentemente de serem acompanhados em um extenso período de tempo. Tais indivíduos são imunes ou curados ao evento de interesse (Ibrahim *et al.*, 2014). Nesse sentido, um indivíduo é considerado imune/curado, quando não sofre o evento de interesse no tempo de observação definido, ou seja, sempre que sua observação é censurada. Logo, têm-se a motivação do uso de modelos de sobrevivência de longa duração.

Os modelos de longa duração ou fração de cura, são de extrema importância na análise de sobrevivência, uma vez que surge na literatura diversos métodos para ajustar tais modelos, sendo Berkson e Gage (1952), Chen *et al.* (1999), Lawless (2011), entre outros.

Na Figura 2.1 têm-se o comportamento que motiva o uso de modelos de sobrevivência de longa duração, uma vez que, ao invés da curva estabilizar em $S_t = 0$, temos que a medida que o tempo aumenta, a curva estabiliza em $S_t = 0.25$. Dessa forma, o ponto que a curva estabiliza, será a proporção de indivíduos que será considerado imunes ou curados.

A seguir, serão apresentados os modelos de longa duração mais conhecidos na literatura, sendo esses o modelo de mistura padrão e o modelo unificado de fração de cura.

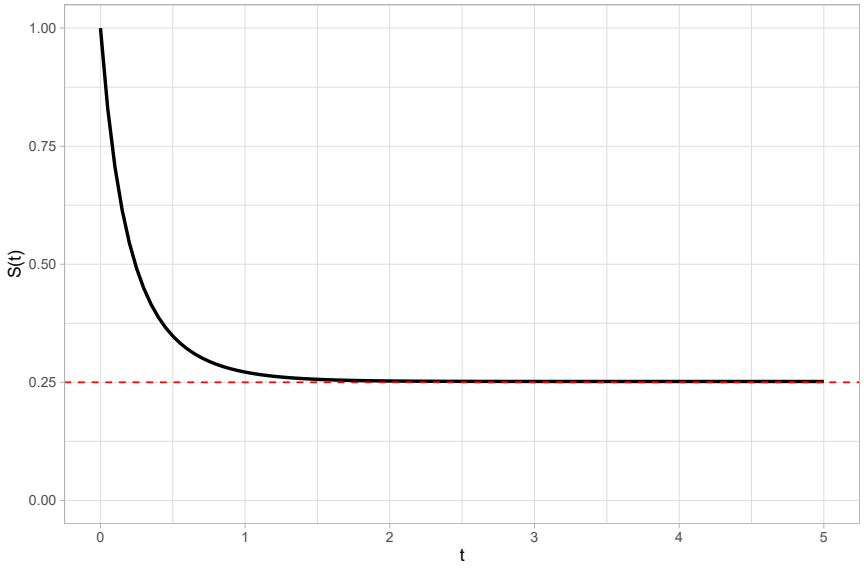

Figura 2.1: Exemplo de função de sobrevivência associada aos modelos de fração de cura.

2.2.1 Modelo de Mistura Padrão

Proposto por Berkson e Gage (1952), o modelo de mistura padrão é um dos modelos mais famosos dentro da modelagem de longa duração. Este constitui-se de uma mistura de distribuições paramétricas, em que uma função de sobrevivência imprópria é considerada para a população total e uma função de sobrevivência própria é considerada por parte da população formada pelos não curados.

O modelo de mistura padrão é derivado considerando uma variável de Bernoulli, não observável M_i aos indivíduos curados e não curados na amostra. Em que,

$$M_i = \begin{cases} 0, \text{ se o indivíduo } i \text{ não está em risco;} \\ 1, \text{ se o indivíduo } i \text{ está em risco,} \end{cases}$$

sendo $P(M_i = 0) = p$ e $P(M_i = 1) = 1 - p$.

Como existem duas subpopulações na amostra, sendo curados e não curados, as funções de sobrevivência para os indivíduos não curados, é própria, enquanto para os indivíduos considerados imunes/curados é esperado que a função de sobrevivência seja imprópria, uma vez que seus tempos de vida são considerados infinitos.

Nesse contexto, seja T uma variável aleatória não negativa e contínua, representando o tempo de vida, logo,

$$P(T > t|M_i = 1) = S(t) \text{ e } P(T > t|M_i = 0) = 1. \tag{2.8}$$

A probabilidade do tempo de vida ser maior do que um determinado tempo t, independentemente ao grupo que pertença, é dada por:

$$\begin{aligned} S_{pop}(t) &= P(T > t) \\ &= P(T > t|M_i = 0)P(M_i = 0) + P(T > t|M_i = 1)P(M_i = 1) \\ &= p + (1-p)S(t),\ t \geq 0. \end{aligned}$$

Dessa forma, a função de sobrevivência populacional será definida por:

$$S_{pop}(t) = p + (1-p)S(t), \tag{2.9}$$

em que $S(.)$ representa a função de sobrevivência própria associada aos indivíduos de risco.

A função acima (2.9) possui as respectivas propriedades:

- Se $p = 0$, então $S_{pop} = S(t)$;
- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é decrescente;
- $\lim_{t \longrightarrow \infty} S_{pop}(t) = p$.

A partir da última propriedade, é possível verificar que a função de sobrevivência populacional é imprópria, uma vez que a curva de sobrevivência estabiliza em p, sendo a probabilidade de cura da população.

Como visto na Subseção 2.1.2, ao obter uma das funções de interesse é possível obter as restantes. Nesse sentido, a função de densidade populacional (imprópria) é dada por,

$$f_{pop}(t) = -\frac{d[S_{pop}(t)]}{dt} = (1-p)f(t)$$

e a função de risco populacional é definida por,

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{(1-p_)f(t)}{p + (1-p)S(t)}.$$

A partir da função anterior, a função de risco própria é dada por.

$$h(t) = \frac{S_{pop}(t)h_{pop}(t)}{(1-p)S(t)} = \left[\frac{S_{pop}(t)}{S_{pop}(t) - p}\right] h_{pop}(t).$$

É possível perceber que $\{S_{pop}(t)/[S_{pop}(t)-p]\} > 1$ tem-se que $h_{pop}(t) < h(t)$, ou seja, a função de risco populacional é limitada pela função de risco base. Decorre-se que $h(t)$ também não terá a propriedade de risco proporcional, uma vez que $\{S_{pop}(t)/[S_{pop}(t)-p]\}$ sempre estará dependendo de t. Ademais,

$$\lim_{t\to\infty} h_{pop}(t) = \lim_{t\to\infty} \frac{(1-p)f(t)}{S_{pop}(t)} = \left(\frac{1-p}{p}\right) \lim_{t\to\infty} f(t) = 0.$$

Nesse sentido, quanto mais o tempo aumentar, o risco da população será convergido para zero, o que retrata a ocorrência da curva de sobrevivência populacional estabilizar em um determinado valor, ou seja, sua fração de cura, que indicará que uma parcela dos indivíduos não obtiveram o evento de interesse e possivelmente foram curados no experimento.

Função de Verossimilhança

Ao considerar que n indivíduos foram observados, o conjunto de dados observados será composto pelos vetores $t = (t_1, ..., t_n)'$ e $\delta = (\delta_1, ..., \delta_n)'$. Dessa forma, sejam $\mathbf{D} = (t, \delta)$ o conjunto de dados observado e $\boldsymbol{\vartheta}$ o vetor de parâmetros a ser estimado.

Logo, a função de verossimilhança para o modelo de mistura padrão será dada por,

$$\begin{aligned} L(\boldsymbol{\vartheta}; \mathbf{D}) &\propto \prod_{i=1}^{n} [f_{pop}(t_i|\boldsymbol{\vartheta})]^{\delta_i} [S_{pop}(t_i|\boldsymbol{\vartheta})]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} [(1-p)f(t_i;\theta)]^{\delta_i} [p + (1-p)S(t_i;\theta)]^{1-\delta_i}. \end{aligned}$$

Modelo de Mistura Padrão Exponencial

Seja T uma variável aleatória não negativa e contínua, que possui distribuição exponencial com parâmetro $\lambda > 0$, se sua função de densidade de probabilidade for escrita da seguinte forma,

$$f(t) = \lambda e^{-\lambda t}, \ t \geq 0 \text{ e } \lambda > 0.$$

A função de sobrevivência é dada por,

$$S(t) = e^{-\lambda t},$$

e a função de risco é definida por,

$$h(t) = \lambda.$$

Nesse contexto, a função de densidade populacional será de,

$$f_{pop}(t) = (1-p)\lambda e^{-\lambda t},$$

e a função de risco populacional é,

$$h_{pop}(t) = \frac{(1-p)\lambda e^{-\lambda t}}{p + (1-p)e^{-\lambda t}}.$$

2.2.2 Modelos Unificados de Fração de Cura

Um outro tipo de modelo de longa duração comumente conhecido na literatura, é o modelo unificado de fração de cura, estudado por Chen *et al.* (1999) e Rodrigues *et al.* (2009). O modelo unificado de fração de cura, se baseia na ocorrência do evento de interesse em um determinado processo divididos em dois estágios:

- **Primeiro estágio:** Seja M, uma variável aleatória, no qual representa o número de riscos ou causas que competem para à ocorrência de um determinado evento de interesse. A variável não é observada, seguindo uma distribuição de probabilidade p_m,

  $$p_m = P[M = m],$$

  em que $m = 0, 1, 2, ...$

- **Segundo estágio:** Dado que $M = m$, e sejam Z_i, com $i = 1, 2, \cdots$, variáveis aleatórias contínuas e não negativas representando o tempo até a ocorrência do evento de interesse atrelado a i-ésima causa independentes entre si, com função acumulada dada por $F_Z(z) - 1 - S_Z(z)$ e independentes de M, o tempo para a ocorrência do evento de interesse é dado por,

  $$T = \min\{Z_0, Z_1, Z_2, \cdots, Z_M\},$$

em que $P[Z_0 = \infty] = 1$, fazendo com que obtenha-se uma parcela da população p_0 de não apresentar o evento de interesse, sendo T uma variável aleatória observável ou censurada em diversos casos e as variáveis Z_i e M sendo latentes, ou seja, não observáveis.

De acordo com a definição de Feller (1991), seja $\{a_m\}$ uma definição de números reais. Se

$$A(s) = a_0 + a_1 s + a_2 s^2 + \cdots ,$$

converge para valores de s em um intervalo de $[0, 1]$, logo, $A_a(s)$ é definida como função geradora da sequência $\{a_m\}$.

Nesse sentido, é possível definir a função de sobrevivência populacional com distribuição T, da seguinte forma:

$$\begin{aligned} S_{pop}(t) &= P(M = 0) + P(Z_1 > t, Z_2 > t, ..., Z_M > t, M \geq 1) \\ &= P(M = 0) + \sum_{m=1}^{\infty} P(M = m) P(Z_1 > t, Z_2 > t, ..., Z_M > t) \\ &= p_0 + \sum_{m=1}^{\infty} p_m [S(t)]^m \\ &= A[S(t)], \end{aligned}$$

em que $A(.)$ é a função geradora da sequência $\{p_m\}$ que converge no intervalo $0 \leq S(t) \leq 1$.

A função de densidade de probabilidade associada a função de sobrevivência populacional é definida por,

$$f_{pop}(t) = f(t) \left. \frac{\mathrm{d}[A(s)]}{\mathrm{d}s} \right|_{s=S(t)},$$

e a função de risco é dada por,

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = f(t) \frac{\left. \frac{\mathrm{d}[A(s)]}{\mathrm{d}s} \right|_{s=S(t)}}{S_{pop}(t)}.$$

Em Feller (1991), pode-se destacar algumas distribuições que são utilizadas, comumente, para o número de risco ou causas que competem para a ocorrência do evento de interesse, sendo algumas a Bernoulli, Poisson, e Geométrica, apresentadas a seguir.

- $M \sim Bernoulli(\theta) : p_m = (1-\theta)^{1-m}$, com $0 < \theta < 1$ e $m = 0, 1$. Logo,

$$A(s) = (1-\theta)\theta s.$$

- $M \sim Poisson(\theta) : p_m = \frac{e^{-\theta}\theta^m}{m!}$, com $\theta > 0$ e $m = 0, 1, 2, \ldots$. Logo,

$$A(s) = \exp -\theta(1-s).$$

- $M \sim Geometrica(\theta) : p_m = (1-\theta)^m\theta$, com $0 < \theta < 1$ e $m = 0, 1, 2, \ldots$ .. Logo,

$$A(s) = \frac{\theta}{1-(1-\theta)^s}.$$

2.2.3 Exemplo de aplicação

Para melhor entendimento dos modelos de mistura padrão, será feita uma aplicação em um conjunto de dados reais. Os dados dessa aplicação foram retirados de Kersey *et al.* (1987) e são referentes a um estudo de recorrência de leucemia em pacientes que foram submetidos a um certo tipo de transplante. O conjunto de dados possui 44 observações, com 9 censuras, sendo 20.45% da base de dados. O tempo máximo de observação foi de aproximadamente 5 anos. Na Figura 2.2, têm-se a curva de sobrevivência estimada pelo estimador de Kaplan e Meier (1958).

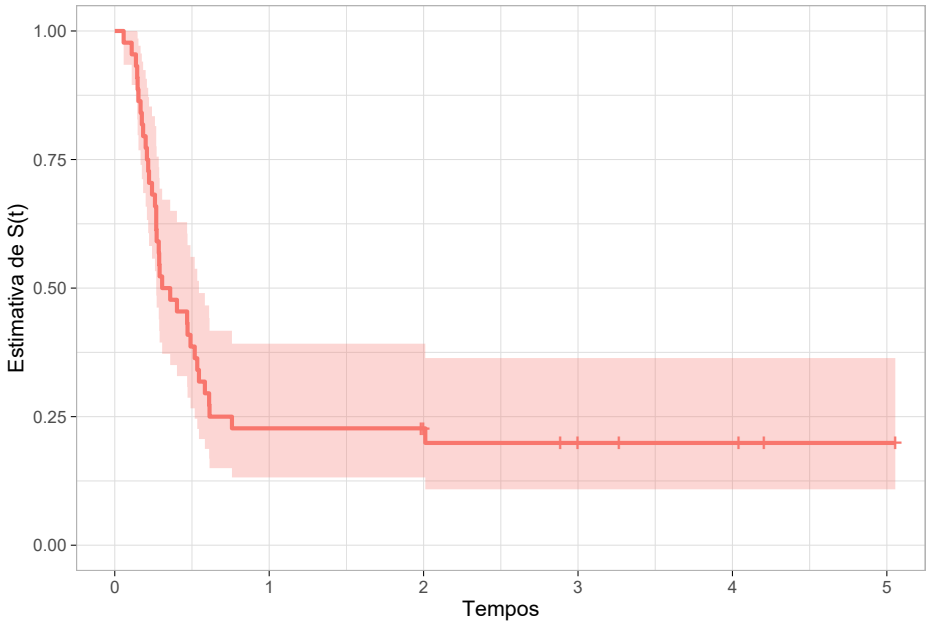

Figura 2.2: Curva de sobrevivência estimada através de Kaplan-Meier.

Dessa forma, é possível observar através da Figura 2.2 estabiliza em aproximadamente $S(t) = 0.20$, logo, têm-se que 20% das pacientes são imunes ao evento de interesse, que é a recorrência de leucemia. Nesse contexto, será ajustado o modelo de mistura padrão Exponencial. A função de sobrevivência populacional será dada por,

$$\begin{aligned} S_{pop}(t) &= p + (1-p)S(t) \\ &= p_0 + (1-p)e^{-\lambda t}, \end{aligned}$$

e a função de densidade populacional será de,

$$\begin{aligned} f_{pop}(t) &= (1-p)f(t) \\ &= (1-p)\lambda e^{-\lambda t}. \end{aligned}$$

Então, a função de verossimilhança para o modelo de mistura padrão exponencial será definida por,

$$\begin{aligned} L(\lambda|\boldsymbol{t}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} [f_{pop}(t_i|\lambda)]^{\delta_i} [S_{pop}(t_i|\lambda)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [(1-p)\lambda e^{-\lambda t_i}]^{\delta_i} [p_0 + (1-p)e^{-\lambda t_i}]^{1-\delta_i}, \end{aligned}$$

e consequentemente, a função de log-verossimilhança, será definida por,

$$l(\lambda|\boldsymbol{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \left(\delta_i \log((1-p)\lambda e^{-\lambda t_i}) + (1-\delta_i) \log(p + (1-p)e^{-\lambda t_i}) \right). \tag{2.10}$$

Os parâmetros são estimados via maximização direta de 2.10. Sendo assim, a Tabela 2.1 possui os parâmetros estimados e o intervalo de confiança do modelo de mistura padrão Exponencial.

Tabela 2.1: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP) e Intervalo de Confiança - IC(95%) para o modelo de mistura padrão Exponencial.

Parâmetros	EMV	EP	Intervalo de Confiança	
			LI	LS
λ	2.680	0.462	1.867	3.684
p	0.204	0.06	0.103	0.338

A partir da Tabela 2.1 é possível verificar que $\hat{p} = 0.204$. Dessa forma, cerca de 20.4%

indivíduos são curados ao evento de interesse, ou seja, não terão a recorrência de leucemia. Ademais, é possível verificar que $\hat{p}$ coincide com o que foi visto na Figura 2.2. Contudo, para verificar se o modelo de mistura padrão Exponencial está bem ajustado aos dados, é necessário curva do estimador de Kaplan e Meier (1958) e a curva do modelo ajustado, que está representado na Figura 2.3.

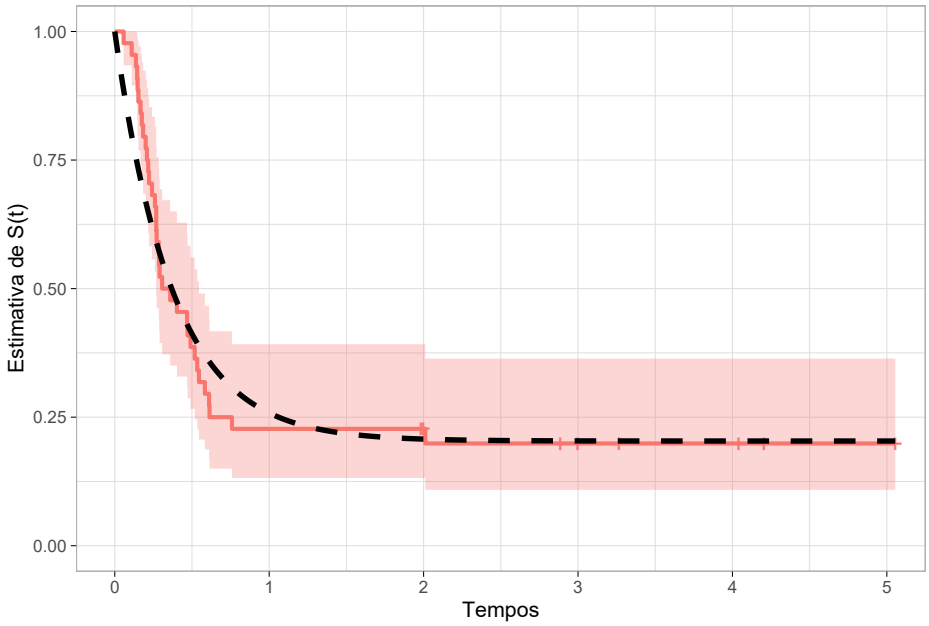

Figura 2.3: Comparação da curva de Kaplan-Meier com a curva do modelo ajustado.

Na Figura 2.3 nota-se que o modelo de mistura padrão Exponencial se ajustou bem aos dados, uma vez que a curva do modelo ajustado está bem próxima da curva do estimador de Kaplan e Meier (1958). Dessa forma, é razoável a utilização do modelo de mistura padrão Exponencial para os dados de recorrência de leucemia em pacientes submetidos a um determinado tipo de transplante.

2.3 Modelos Defeituosos

Nesta seção será apresentado alguns modelos probabilísticos defeituosos encontrados na literatura que são utilizados na modelagem de dados de sobrevivência. De acordo com o tema do trabalho, será considerado os modelos Gompertz e Gaussiana-Inversa.

Uma distribuição de probabilidade é chamada de defeituosa, se a integral da função de densidade de probabilidade não resulta em 1, mas sim, um valor $p \in (0, 1)$, quando o valor dos parâmetros é mudado.

A função de distribuição acumulada é defeituosa, quando não se aproxima mais de 1,

e sim de p. A Figura 2.4 ilustra a função de distribuição acumulada de uma distribuição defeituosa.

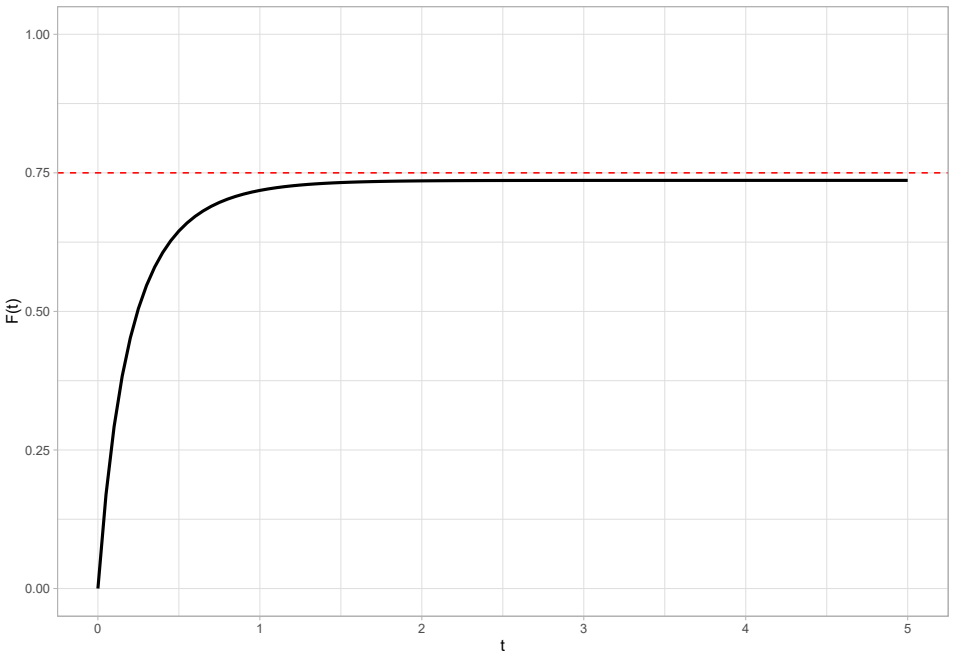

Figura 2.4: Função de distribuição acumulada defeituosa.

Nas seções 2.3.1 e 2.3.2, será apresentado as distribuições de probabilidade Gompertz e Gaussiana-Inversa em suas formas defeituosas, respectivamente.

2.3.1 Distribuição Gompertz Defeituosa

A distribuição Gompertz foi proposta por Benjamin Gompertz, visando a aplicabilidade em estudos demográficos (Gompertz, 1825), além disso, a distribuição também é frequentemente utilizada em modelagem de dados de sobrevivência em diversas áreas do conhecimento (Gieser *et al.*, 1998). A função de densidade de probabilidade, é dada por,

$$f(t;\theta) = be^{at}e^{-\frac{b}{a}(e^{at}-1)}. \tag{2.11}$$

Nessa distribuição, a é o parâmetro de escala e b é o parâmetro de forma. Nesse sentido, considerando T uma variável aleatória com distribuição Gompertz, com parâmetros dados por $a > 0$, $b > 0$ e $\theta = (a, b)^T$, a função de sobrevivência é dada por,

$$S(t;\theta) = P(T \geq t) = e^{-\frac{b}{a}(e^{at}-1)}, \tag{2.12}$$

e a função de risco é definida por,

$$h(t;\theta) = be^{at}. \tag{2.13}$$

Na Figura 2.5 é possível verificar que as funções da distribuição Gompertz, são suscetíveis a alteração do valor de seus parâmetros.

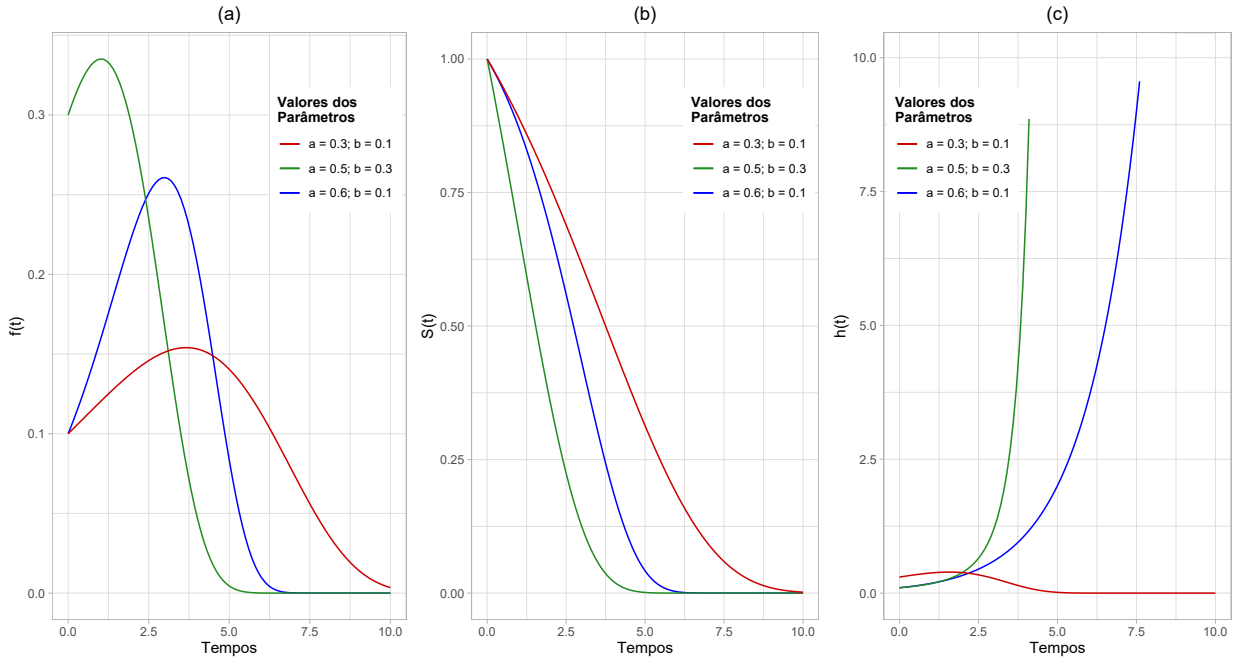

Figura 2.5: Funções de densidade de probabilidade (a), de sobrevivência (b) e taxa de falha (c) para a distribuição Gompertz.

A distribuição Gompertz defeituosa, é a distribuição Gompertz que permite que o parâmetro de escala possua valores negativos ($a < 0$). Nesse contexto, quando o parâmetro a é negativo, a distribuição de Gompertz torna-se uma distribuição imprópria, o que é muito útil para modelar dados de sobrevivência na presença de uma fração sobrevivente.

A fração de cura para a população é calculada como o limite da função de sobrevivência ($S(t)$) quando $a < 0$, e é definida por,

$$p = \lim_{t\to\infty} S(t;\theta) = \lim_{t\to\infty} e^{-(b/a)(e^{at}-1)} = e^{b/a} \in (0,1). \tag{2.14}$$

Na Figura 2.6 têm-se o comportamento das funções de sobrevivência e risco quando a distribuição é defeituosa.

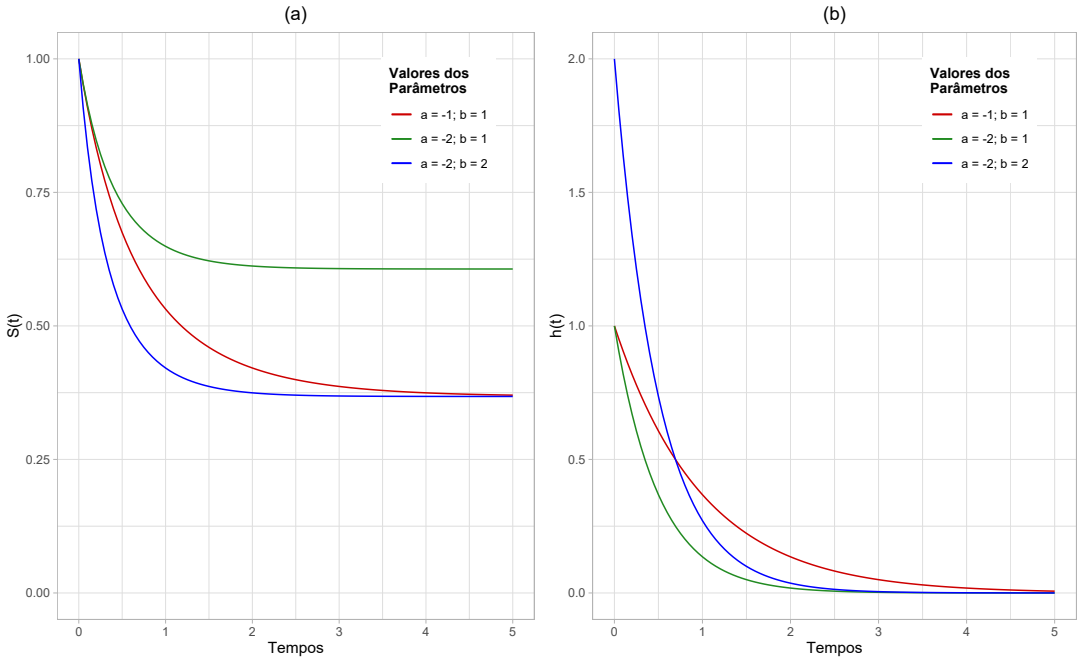

Figura 2.6: Funções de sobrevivência (a) e taxa de falha (b) para a distribuição Gompertz defeituosa.

2.3.2 Distribuição Gaussiana-Inversa Defeituosa

A distribuição Gaussiana-Inversa defeituosa possui função de densidade de probabilidade que é dada por,

$$f(t;\theta) = \frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\}, \tag{2.15}$$

em que $a > 0, b > 0$, $t > 0$ e $\theta = (a,b)^T$. A função de sobrevivência da distribuição, é dada por,

$$S(t;\theta) = 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right], \tag{2.16}$$

em que $\Phi(.)$ denota a função de distribuição acumulada da normal padrão.

A função de risco da distribuição será definida por,

$$h(t;\theta) = \frac{\frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1 - \left\{\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right\}} \tag{2.17}$$

Na Figura 2.7 têm-se as ilustrações de vários cenários para a função de densidade de probabilidade, função de sobrevivência e taxa de falha (função de risco) para a distribuição Gaussiana-Inversa.

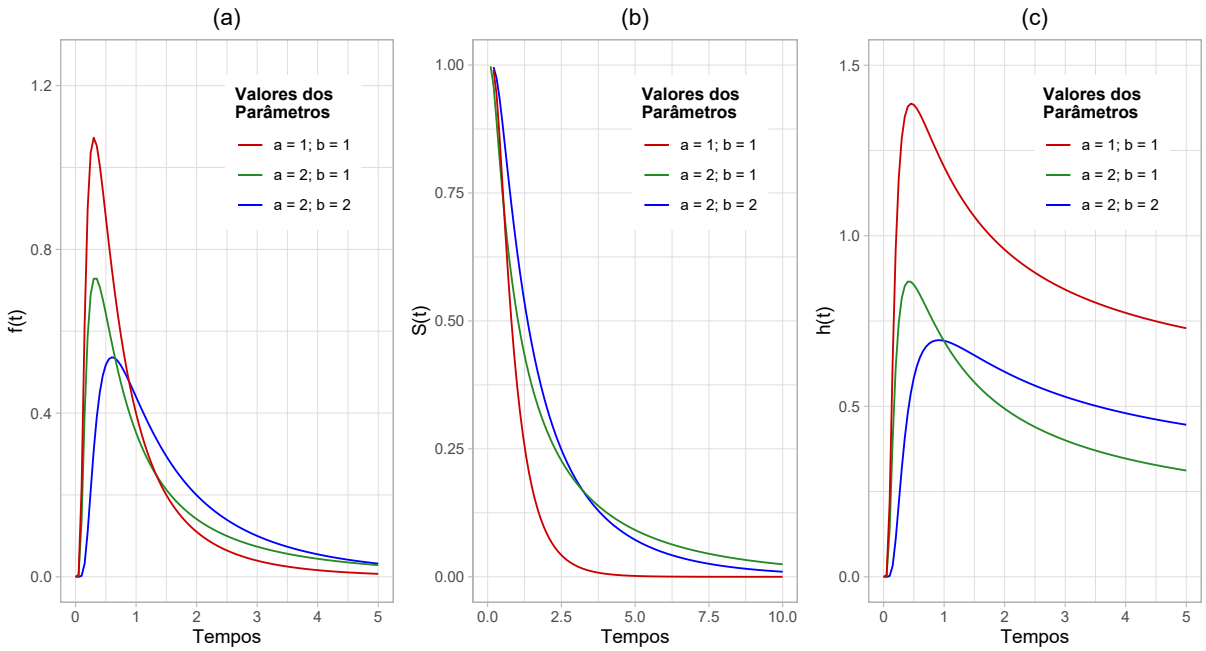

Figura 2.7: Funções de densidade de probabilidade (a), de sobrevivência (b) e taxa de falha (c) para a distribuição Gaussiana-Inversa.

A distribuição Gaussiana-Inversa permite valores negativos para a. Quando $a < 0$ temos uma distribuição defeituosa e a fração de cura correspondente é dada por,

$$p = \lim_{t\to\infty} S(t;\theta) = \lim_{t\to\infty} \left\{ \Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right\} = (1 - e^{2a/b}) \in (0,1). \tag{2.18}$$

Na Figura 2.8 têm-se o comportamento das funções de sobrevivência e função de risco da distribuição Gaussiana-Inversa, quando esta é defeituosa.

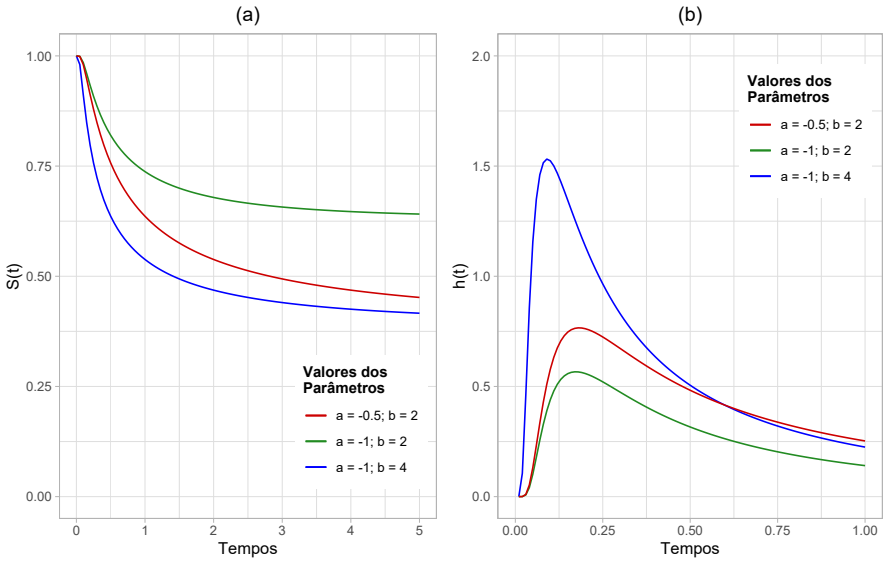

Figura 2.8: Funções de sobrevivência (a) e taxa de falha (b) para a distribuição Gaussiana-Inversa defeituosa.

Se o parâmetro a estimado for negativo ($a < 0$), então a fração de cura para os modelos defeituosos Gompertz e Gaussiana-Inversa, pode ser obtida por 2.14 e 2.18, respectivamente. Caso contrário, se o parâmetro a estimado for positivo ($a > 0$), então não haverá fração de cura, de acordo com os modelos defeituosos, e 2.12 e 2.16 são funções de sobrevivência usuais.

2.3.3 Exemplo de aplicação

Para esta aplicação, o conjunto de dados utilizado foi retirado de Carvalho *et al.* (2011). Este conjunto trata-se de dados provenientes de coortes hospitalares de pacientes portadores de HIV. Desta coorte, foram obtidos uma amostra de 193 indivíduos que foram diagnosticados como portadores de AIDS durante o período de acompanhamento em dias.

Neste caso, iremos supor que os dados podem ser modelados a partir da distribuição Gompertz. Com isso, o método de máxima verossimilhança é o mais indicado para encontrar as estimativas dos parâmetros. A função de verossimilhança da distribuição Gompertz é dada por,

$$\begin{aligned} L(a,b|\boldsymbol{t},\boldsymbol{\delta}) &= \prod_{i=1}^{n} f(t_i|a,b)^{\delta_i} S(t_i|a,b)^{1-\delta_i} \\ &= \prod_{i=1}^{n} (be^{at_i - \frac{b}{a}(e^{at_i}-1)})^{\delta_1} (e^{-\frac{b}{a}(e^{at_i}-1)})^{1-\delta_i}, \end{aligned}$$

e consequentemente, a função de log-verossimilhança será definida por,

$$\begin{aligned} l(a,b|\boldsymbol{t},\boldsymbol{\delta}) &= \sum_{i=1}^{n} \delta_i \log(be^{at_i - \frac{b}{a}(e^{at_i}-1)}) + \sum_{i=1}^{n} (1-\delta_i) \ln\left(e^{-\frac{b}{a}(e^{at_i}-1)}\right) \\ &= \sum_{i=1}^{n} \delta_i \log(b) + \sum_{i=1}^{n} \delta_i \left(at_i - \frac{b}{a}(e^{at_i}-1)\right) + \sum_{i=1}^{n} -(1-\delta_i)\frac{b}{a}(e^{\gamma t_i}-1) \\ &= \log(b)\sum_{i=1}^{n} \delta_i + a\sum_{i=1}^{n} \delta_i t_i - \frac{b}{a}\sum_{i=1}^{n} \delta_i (e^{at_i}-1) - \frac{b}{a}\sum_{i=1}^{n} (e^{at_i}-1) + \frac{b}{a}\sum_{i=1}^{n} \delta_i (e^{at_i}-1). \end{aligned}$$

Dessa forma, temos que a função de log-verossimilhança da distribuição Gompertz é dada por,

$$l(a,b|\boldsymbol{t},\boldsymbol{\delta}) = \log(b)\sum_{i=1}^{n} \delta_i + a\sum_{i=1}^{n} \delta_i t_i - \frac{b}{a}\sum_{i=1}^{n} (e^{at_i}-1).$$

Para ajustar o modelo, foi utilizado o pacote *flexsurv* do software R Core Team (2013),

e com isso, as estimativas do modelo e do intervalo de confiança, com nível de confiança de 95%, serão,

Tabela 2.2: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP) e Intervalo de Confiança - IC(95%) para o modelo Gompertz.

Parâmetros	Estimativas	Intervalo de Confiança	
		LI	LS
a	−0.000668	−0.000884	−0.000452
b	0.000775	0.000581	0.001033

Nesse sentido, através da Tabela 2.2 nota-se que o parâmetro a assumiu um valor negativo, logo, a distribuição Gompertz torna-se uma distribuição imprópria, ou seja, defeituosa. Com isso, é possível possível obter a fração de cura da população, que será calculada como o limite da função de sobrevivência, sendo,

$$p = \lim_{t\to\infty} S(t;\theta) = e^{b/a} = e^{0.000775/-0.000668} = 0.3134.$$

Logo, pode-se dizer que aproximadamente 31.34% dos indivíduos diagnosticados com AIDS, são imunes/curados ao evento de interesse. Por fim, na Figura 2.9 nota-se que o modelo Gompertz se ajustou bem aos dados, uma vez que a curva de sobrevivência do modelo ajustado é bem próxima a curva do estimador de Kaplan e Meier (1958). Logo, para este conjunto de dados é razoável a utilização do modelo Gompertz, independente do modelo ser defeituoso.

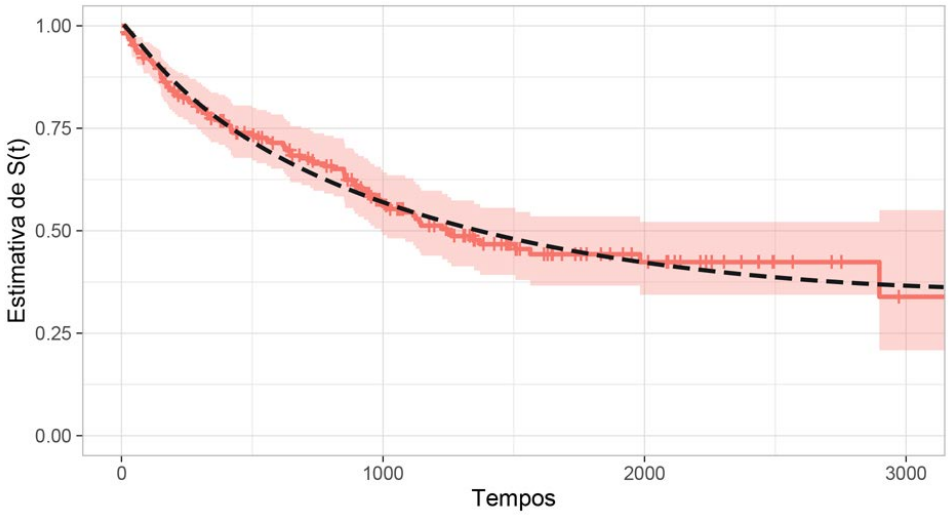

Figura 2.9: Comparação da curva de Kaplan-Meier com a curva do modelo Gompertz ajustado.

2.4 Critério de seleção de modelos

Quando nos deparamos com problemas que envolvem incerteza, torna-se fundamental a utilização de critérios que permitam comparar e selecionar adequadamente modelos paramétricos durante a análise de dados.

Akaike (1974) propõe um método baseado na medida de Informação de Kullback-Leibler. Seja k o número de parâmetros a serem estimados, n o número de observações de t e $\hat{\theta}$ uma estimativa de θ, o critério de informação de Akaike (AIC) é obtido por:

$$AIC = -2\log(L(\hat{\theta}; t)) + 2k$$

Dado um conjunto de modelos candidatos para t, ajustados os dados, o mais adequado será aquele que fornecer o menor AIC. Além de selecionar um ótimo ajuste, o critério penaliza a adição de parâmetros, não ocorrendo o *overfitting*, ou seja, não ocorrendo a seleção de um modelo complexo e com muitos parâmetros que tenham um pobre desempenho preditivo.

Outro critério muito utilizado é o Critério de Informação Bayesiana (BIC), desenvolvido por Schwarz (1978). A estrutura desta métrica é semelhante ao AIC. Nesse caso, o BIC é obtido através de resultados assintóticos e da suposição de que os dados pertencem à família exponencial, em que, a equação é dada por:

$$BIC = -2\log(L(\hat{\theta}; t)) + k\log(n).$$

Este critério também penaliza a adição de parâmetros, logo, dos modelos candidatos, o mais adequado será aquele que possuir o menor BIC.

2.5 Considerações Finais

Este capítulo buscou apresentar os conceitos básicos de análise de sobrevivência, como a censura, funções de interesse, estimador de Kaplan-Meier e a estimação por máxima verossimilhança. Considerando a metodologia que será aplicada neste trabalho, foi estudado a modelagem de longa duração, no qual faz parte a abordagem de indivíduos não suscetíveis ao evento de interesse. Ademais, foi considerado a metodologia de distribuição defeituosa e duas distribuições comumente utilizadas na literatura.

Capítulo 3

Modelos Fração de Cura Defeituosos Zero Ajustados

Como foi visto no Capítulo 2, na análise de sobrevivência existem situações em que algumas unidades do estudo não são suscetíveis ao evento de interesse, sendo denominados por indivíduos imunes ou curados. A classe de modelos que utiliza essa a abordagem é a de modelagem de longa duração, ou também comumente conhecida como modelagem de fração de cura.

Outra abordagem da modelagem de fração de cura é através dos riscos competitivos, apresentados por Borges *et al.* (2012), Cancho *et al.* (2013), Chen *et al.* (1999), entre outros. Esses modelos obtiveram êxito ao serem aplicados nos quais as unidades não são suscetíveis a falhas no tempo zero ou tempos de sobrevivência ajustados para zero, ou seja, a falhas precoces.

Entretanto, em alguns estudos os indivíduos são suscetíveis a falhas precoces, resultando em um tempo de sobrevida igual a zero ou muito próximo de zero. Em um contexto prático, muitos pesquisadores tomam a decisão de remover essa informação, consequentemente levando a conclusões errôneas, com taxas de sobrevivência superestimadas, além de ignorar as características da unidade que resultam na falha precoce do evento. Em outras áreas na Estatística, existem diversas abordagens que foram propostas para lidar com o excesso de zeros em um determinado modelo.

Na literatura, a abordagem mais adequada para lidar de uma forma adequada com a questão do excesso de zeros é a partir da classe de distribuições conhecida como modelos inflacionados de zeros.

3.1 Modelo de Fração de Cura Zero Ajustado

Nesse tipo de modelagem, utiliza-se uma mistura de duas distribuições com dois processos subjacentes, um que trata do excesso de zeros e outro que trata da parte diferente de zero (Martin *et al.*, 2005). Nesse sentido, considerando a circunstância de dados financeiros, foi proposto uma extensão do modelo de mistura padrão de Berkson e Gage (1952) que permite a adição de uma proporção de tempos iguais a zero por Ribeiro de Oliveira Jr *et al.* (2017), em que a função de sobrevivência de todos os tempos possíveis, é dada por:

$$S_{pop}(t) = p_1 + (1 - p_0 - p_1)S_0^*(t), \quad t \geq 0, \tag{3.1}$$

em que $S_0^*(t)$ é a função de sobrevivência associada à proporção de indivíduos suscetíveis a falha, p_0 é a proporção de tempos de sobrevivência inflacionados com zero e p_1 é proporção de indivíduos imunes ou curados na população. Note que o modelo dado pela equação 3.1 tem as seguintes propriedades.

- $\lim_{t\to\infty} S_{pop}(t) = p_1 > 0$,
- $S_{pop}(0) = 1 - p_0 < 1$.

Além do mais, se $p_0 = 0$, ou seja, sem a inflação de zeros, será obtido o modelo de mistura padrão dado por Berkson e Gage (1952).

Na Figura 3.1 tem-se o comportamento da função de sobrevivência deste modelo e podemos observar que existem duas características nos dados as quais devem ser contempladas adequadamente nos modelos de sobrevivência: a proporção de zeros e a fração de cura.

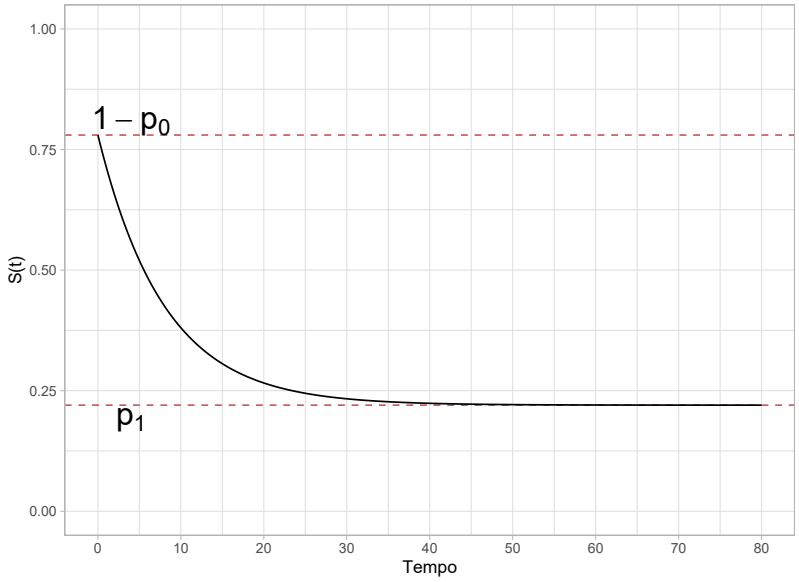

Figura 3.1: Exemplo de função de Sobrevivência do modelo de taxa de cura inflacionado no zero.

3.2 Modelo Defeituoso Zero Ajustado

Como já foi visto, no contexto de fração de cura, os modelos defeituosos oferecem a estratégia para modelar dados de sobrevivência ajustados para zero. Nesse sentido, ao invés de estimar a fração de cura p_1 diretamente, como no modelo de mistura padrão, o modelo defeituoso é uma alternativa para modelar dados de tempo de vida de longa duração.

Nesse contexto, para acomodar os tempos de vida ajustados a zero em modelos defeituosos, Calsavara *et al.* (2019a) propôs uma nova função de sobrevivência como segue:

$$S_{pop}(t;\boldsymbol{\theta}^*) = (1-p_0)S(t;\boldsymbol{\theta}), \quad t>0, \tag{3.2}$$

em que $S(\cdot;\boldsymbol{\theta})$ é uma função de sobrevivência própria ou imprópria, $0 \leq p_0 \leq 1$ denota a proporção de zero-ajustados e $\boldsymbol{\theta}^* = \left(p_0, \boldsymbol{\theta}^\top\right)^\top$ é um vetor de parâmetros.

É importante ressaltar que se $S(\cdot;\boldsymbol{\theta})$ é uma função de sobrevivência própria, ou seja, $\lim_{t\to\infty} S(\cdot;\boldsymbol{\theta}) = 0$, o modelo (3.2) torna-se um modelo padrão de sobrevivência ajustado a zero. Caso contrário, se a função de sobrevivência $S(\cdot;\boldsymbol{\theta})$ for imprópria, então o modelo proposto satisfaz,

$$S_{pop}(t;\boldsymbol{\theta}^*) = (1-p_0) \leq 1,$$

e o limite da função de sobrevivência é

$$p_1 = \lim_{t\to\infty} S_p\left(t;\boldsymbol{\theta}^*\right) = (1-p_0)\lim_{t\to\infty} S(t;\boldsymbol{\theta}) = (1-p_0)\,p \in (0,1),$$

em que p é a fração de cura da distribuição imprópria/defeituosa.

As funções de distribuição acumulada e densidade de probabilidade associadas são, respectivamente,

$$F_{pop}(t;\boldsymbol{\theta^*}) = p_0 + (1-p_0)F(t;\theta), \quad t>0,$$

e

$$f_{pop}(t;\boldsymbol{\theta^*}) = \begin{cases} p_0, & \text{se } t=0, \\ (1-p_0)f(t;\theta), & \text{se } t>0. \end{cases}$$

Note que, se $p_0 = 0$, o modelo padrão defeituoso é obtido como um caso especial.

3.2.1 Modelo Defeituoso Gompertz Zero Ajustado

Com base na equação 3.2 com a função de sobrevivência em 2.12, a função de sobrevivência do modelo defeituoso Gompertz zero ajustado, será dado por:

$$S_{pop}(t;\boldsymbol{\theta^*}) = (1-p_0)\exp\left\{-\frac{b}{a}(e^{at}-1)\right\},$$

em que $\boldsymbol{\theta^*} = (p_0, a, b)^\top$ é um vetor de parâmetros, onde $0 \leq p_0 \leq 1$, $a \in \mathcal{R}$ e $b > 0$.

A função de densidade de probabilidade correspondente é definida por,

$$f_{pop}(t;\boldsymbol{\theta^*}) = (1-p_0)b \times \exp\left\{at - \frac{b}{a}(e^{at}-1)\right\}.$$

Como visto na distribuição Gompertz defeituosa (2.3.1), a distribuição Gompertz defeituosa zero ajustada, também permite valores negativos para o parâmetro a. Nesse caso, a fração de cura correspondente quando $a<0$, é dada por,

$$p_1 = \lim_{t\to\infty} S_p(t;\boldsymbol{\theta^*}) = (1-p_0)\lim_{t\to\infty} e^{(-b/a)(e^{at}-1)} = (1-p_0)e^{b/a} = (1-p_0)p \in (0,1). \quad (3.3)$$

A partir de 3.3 a distribuição Gompertz defeituosa zero ajustada, mostra que a fração de cura diminui a medida que b aumenta.

3.2.2 Modelo Defeituoso Gaussiana-Inversa Zero Ajustado

Novamente, com base na equação 3.2 com a função de sobrevivência em 2.16, a função de sobrevivência do modelo defeituoso Gaussiana-Inversa zero ajustado, é dado por:

$$S_{pop}(t;\boldsymbol{\theta}^*) = (1-p_0)\left[1-\left\{\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right\}\right],$$

em que $\boldsymbol{\theta}^* = (p_0, a, b)^\top$ é um vetor de parâmetros, onde $0 \leq p_0 \leq 1$, $a \in \mathcal{R}$ e $b > 0$.

A função de densidade de probabilidade correspondente, é dada por,

$$f_{pop}(t;\boldsymbol{\theta}^*) = \frac{1-p_0}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}(1-at)^2\right\}.$$

Seguindo o mesmo conceito do modelo defeituoso Gompertz zero ajustado, o modelo defeituoso Gaussiana-Inversa zero ajustado permite $a < 0$, e sua fração de cura é,

$$p_1 = \lim_{t\to\infty} S_p(t;\boldsymbol{\theta}^*) = (1-p_0)\lim_{t\to\infty}\left[1-\left\{\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right\}\right]$$

$$= (1-p_0)(1-e^{2a/b}) = (1-p_0)p \in (0,1). \tag{3.4}$$

A partir de 3.4 a distribuição Gaussiana-Inversa defeituosa zero ajustada, mostra que a fração de cura diminui a medida que b aumenta.

Nesse sentido, se o parâmetro a estimado for negativo ($a < 0$), então a fração de cura para os modelos defeituosos Gompertz e Gaussiana-Inversa ajustados em zero, pode ser obtidas, respectivamente, de 3.3 e 3.4. Caso contrário, se o parâmetro estimado do modelo for positivo, não haverá fração de cura, de acordo com os modelos defeituosos zero ajustados.

A vantagem do modelo proposto por Calsavara *et al.* (2019a) está na capacidade de acomodar uma proporção do tempo de vida ajustados para zero, bem como a possibilidade de fração de cura na população.

3.3 Inferência

Nessa seção, será descrito o procedimento de inferência, baseado na máxima verossimilhança e também na teoria assintótica de grandes amostras. Seja $T \geq 0$ uma variável

aleatória que representa o tempo até a ocorrência do evento de interesse. Considere o tempo da variável indicadora δ_i^*, isto é, $\delta_i^* = 0$ se $T = 0$ (tempo de sobrevivência ajustado para zero) e $\delta_i^* = 1$ se $T > 0$, $i = 1, ..., n$. Ademais, seja δ_i a variável indicadora de censura, em que $\delta_i = 0$ se os dados são censurados e $\delta_i = 1$, caso contrário. As variáveis explicativas, serão incorporadas ao modelo com um conjunto de vetores de duas variáveis, $\mathbf{x_1} \in \mathbb{R}^{s+1}$ e $\mathbf{x_2} \in \mathbb{R}^{q+1}$, de tal modo que $\mathbf{x}^\top = (\mathbf{x_1}^\top, \mathbf{x_2}^\top) \in \mathbb{R}^w$ é um vetor de covariável com dimensão w, em que $w = s + q + 2$. De acordo com Calsavara *et al.* (2019a), foram consideradas as funções de ligação logito e log, sendo:

$$\ln\left(\frac{p_{0\mathbf{x_{1i}}}}{1 - p_{0\mathbf{x_{1i}}}}\right) = \mathbf{x_{1i}}^\top \boldsymbol{\beta_0} \quad \text{e} \quad \ln b(\mathbf{x_{2i}}) = \mathbf{x_{2i}}^\top \boldsymbol{\beta_1},$$

em que $x_{1i}^\top = (1, x_{1_{i1}}, ..., x_{1_{is}})$ e $x_{2i}^\top = (1, x_{2_{iq}}, ..., x_{1_{is}})$ são os conjuntos de covariáveis e $\beta_0^\top = (\beta_{00}, \beta_{01}, ..., \beta_{0s})$ e $\beta_1^\top = (\beta_{10}, \beta_{11}, ..., \beta_{1q})$ e seus coeficientes de regressão, respectivamente. Desta forma, a função de ligação dependerá das covariáveis e pode ser expressa da seguinte maneira:

$$p_{0_{\mathbf{x_{1i}}}} = \frac{\exp\{\mathbf{x_{1i}}^\top \boldsymbol{\beta_0}\}}{1 + \exp\{\mathbf{x_{1i}}^\top \boldsymbol{\beta_0}\}} \quad \text{e} \quad b(\mathbf{x_{2i}}) = \exp\{\mathbf{x_{2i}}^\top \boldsymbol{\beta_1}\}.$$

Na prática, os vetores de covariáveis podem ser os mesmos, ou seja, $x = x_1 = x_2$. Ademais, as funções de ligação logito e log serão utilizadas para manter a amplitude dos valores de p_0 e b, respectivamente.

Outras funções de ligações podem ser utilizadas para a proporção de falhas precoces, como a função de ligação probito e complementar log-log. Considerando os parâmetros de escala e forma, é possível utilizar o quadrado inverso e as funções de ligação recíproca, respectivamente. No entanto, neste trabalho não será abordado a sensibilidade das funções de ligação.

Nesse sentido, no conjunto de dados a ser observado, temos $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta}, \boldsymbol{\delta^*}, \mathbb{X})$, em que $\mathbf{t} = (t_1, ..., t_n)^\top$ serão os tempos de vida observados, $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)^\top$ e $\boldsymbol{\delta^*} = (\delta_1^*, ..., \delta_n^*)^\top$, são respectivamente, os indicadores de censura e tempo de censura, e $\mathbb{X}$ é a matriz contendo as informações das covariáveis. Consideramos que T_i's são variáveis aleatórias dependentes e identicamente distribuídas com a função de sobrevivência especificada por $S_p(\cdot; \boldsymbol{\vartheta})$, em que $\boldsymbol{\vartheta} = (a, \beta_0, \beta_1)^\top$ é um vetor de parâmetros desconhecidos. Assumimos que T é independente do tempo de censura. Logo, a função de verossimilhança de $\boldsymbol{\vartheta}$ sob censura

não-informativa é expressa como,

$$L(\boldsymbol{\vartheta};\mathbf{D}) \propto \prod_{i=1}^{n}(p_{0\mathbf{x_{1i}}})^{1-\delta_i^*}\{f_p(t_i;\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}})^{\delta_i}S_p(t_i;\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}})^{1-\delta_i}\}^{\delta_i^*}. \tag{3.5}$$

A correspondente log-verossimilhança, é dada por,

$$\begin{aligned} l(\boldsymbol{\vartheta}) &= \log L(\boldsymbol{\vartheta};\mathbf{D}) \\ &\propto \sum_{i=1}^{n}(1-\delta^*)\log(p_{0\mathbf{x_{1i}}}) + \sum_{i=1}^{n}\delta_i^*\delta_i \log f_p(\mathbf{t_i};\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}}) + \sum_{i=1}^{n}(1-\delta_i)\delta_i^* \log S_p(\mathbf{t_i};\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}}) \end{aligned}$$

A função de log-verossimilhança anterior, pode ser reescrita da seguinte forma,

$$\begin{aligned} l(\boldsymbol{\vartheta}) \propto &\sum_{i=1}^{n}(1-\delta^*)\log(p_{0\mathbf{x_{1i}}}) + \sum_{i=1}^{n}\delta_i^*\delta_i\log(1-p_{0\mathbf{x_{1i}}}) \\ &+ \sum_{i=1}^{n}\delta_i^*\delta_i \log f(\mathbf{t_i};\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}}) + \sum_{i=1}^{n}(1-\delta_i)\delta_i^* \log S(\mathbf{t_i};\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}}), \end{aligned}$$

em que $f(\cdot;\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}})$ e $S(\cdot;\boldsymbol{\vartheta},\mathbf{x_{1i}},\mathbf{x_{2i}})$ são respectivamente, a densidade de probabilidade e a função de sobrevivência associadas com a distribuição defeituosa. A prova completa da função de verossimilhança pode ser encontrada em Calsavara *et al.* (2019a).

As estimativas de máxima verossimilhança dos parâmetros são obtidas maximizando numericamente a função de log-verossimilhança. Existem diversos métodos para essa maximização numérica, entretanto, foi utilizado a rotina do *optim* no software estatístico R para essa maximização.

Dessa forma, as propriedades assintóticas das estimativas de máxima verossimilhança são necessárias para a construção de intervalos de confiança e testes de hipóteses sobre os parâmetros do modelo. Sob certas condições, $\hat{\boldsymbol{\vartheta}}$ possui distribuição normal multivariada assintótica com média $\boldsymbol{\vartheta}$ e variância $\boldsymbol{\Sigma}(\hat{\boldsymbol{\vartheta}})$, sendo estimada por,

$$\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\vartheta}}) = \left\{ - \left. \frac{\mathrm{d}l(\boldsymbol{\vartheta})}{\mathrm{d}\boldsymbol{\vartheta} d\boldsymbol{\vartheta}^\top} \right|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} = \hat{\boldsymbol{\vartheta}} \right\}^{-1}$$

Portanto, um intervalo de confiança aproximado $100(1-\alpha)\%$ para ϑ_i é $(\hat{\vartheta}_i \pm z_{\alpha/2}\sqrt{\Sigma^{ii}})$, em que Σ^{ii} denota o i-ésimo elemento da diagonal da inversa de $\boldsymbol{\Sigma}$ avaliado em $\hat{\boldsymbol{\vartheta}}$ e z_α denota o percentil $100(1-\alpha)$ da variável aleatória da normal padrão.

Os resultados da normalidade assintótica das estimativas de máxima verossimilhança

são válidas sob certas condições. Em Calsavara *et al.* (2019a) foi realizado um estudo de simulação para verificar se as assíntotas usuais das estimativas de máxima verossimilhança são válidas, uma vez que simulações têm sido usadas em muitos trabalhos para verificar o comportamento assintótico de estimativas de máxima verossimilhança, especialmente quando uma investigação analítica não é trivial.

3.4 Considerações Finais

Neste Capítulo, foi apresentado a metodologia principal a ser usada neste trabalho, os modelos defeituosos zero ajustados, sendo esses o modelo defeituoso Gompertz zero ajustado e o modelo defeituoso Gaussiana-Inversa zero ajustado.

Capítulo 4

Aplicação a Dados Financeiros

Neste trabalho, o conjunto de dados para essa aplicação foi concedido por uma instituição financeira, na qual realiza serviços voltados para crédito. Esses dados foram analisados por Toledo *et al.* (2022) considerando o modelo proposto por Ribeiro de Oliveira Jr *et al.* (2017) dado na Equação 3.1.

O período considerado foi após a recessão econômica brasileira, com início em meados de 2014, na qual houve um aumento da crise financeira no país. Para esta aplicação, será considerado uma amostra aleatória de 9.645 CPF's. A característica principal dos indivíduos que englobam este conjunto de dados, é a aquisição de dívidas, ou seja, há clientes com dívidas vencidas e não quitadas no período de julho/2015 à dezembro/2015.

O processo de realização de cobrança das dívidas em aberto, é feita de maneira tradicional. Este tipo de processo, pode ser efetuado por meio de cobranças telefônicas, cartas de cobrança ou ligações extrajudiciais. Devido ao cenário da crise econômica, há a lentidão do processo de restituição do *status* do clientes de inadimplente para adimplente, sendo necessário a utilização de modelos estatísticos para estimar o prazo para a ocorrência destes eventos.

O tempo de falha neste estudo é o tempo de entre a data de aquisição da dívida até a finalização do estudo, sendo um período de 24 meses. Nesse contexto, para identificar diferenças nos comportamentos dos clientes para diferentes cenários, será estudo a situação com o uso de duas covariáveis.

- **Informação de Consulta:** Indicativo de consultas de empresas aos relatórios de créditos dos clientes nos últimos 180 dias.

- **Tipo de Dívida:** Caracterização da origem da dívida do cliente durante o período

de crise econômica, sendo por origem financeira (bancos) ou outros segmentos.

Através da Tabela 4.1 é possível verificar a composição das variáveis em relação com as suas categorias.

Tabela 4.1: Quantidade por covariável.

Covariável	**Descrição**	**Categoria**	**n**	**%**
X_1	Informação de Consulta	0: Sem Consulta	295	3.06%
		1: Com Consulta	9350	96.90%
X_2	Tipo de Dívida	0: Banco	5103	52.90%
		1: Outros Segmentos	4542	47.10%

Ademais, através do conjunto de dados é possível averiguar qual a distribuição dos subgrupos de clientes, uma vez que é possível notar comportamentos distintos em relação ao tempo de recuperação do *status* de adimplência para os diferentes subgrupos de clientes, sendo,

1. **Cliente com evento no tempo zero:** Clientes que realizaram o evento de interesse logo no início do estudo, ou seja, regularizaram a quitação da dívida no tempo zero, tornando-se adimplente novamente;

2. **Cliente suscetível ao evento:** Clientes que são suscetíveis ao evento de interesse, ou seja, os clientes que regularizaram a dívida dentro do período de 24 meses, assim, tornando-se adimplente novamente;

3. **Cliente não suscetível ao evento:** Clientes não suscetíveis ao evento de interesse, que de acordo com a teoria são considerados imunes/curados, ou seja, são os clientes que continuaram com suas dívidas em abertos após o período de 24 meses, assim, permanecendo com o *status* de inadimplência.

Na Tabela 4.2 há a quantidade de cada subgrupo presente no conjunto de dados.

Tabela 4.2: Subgrupos de Clientes no Conjunto de Dados.

Subgrupos	**Quantidade de Clientes**	**% de Clientes**
(I) Cliente com evento no tempo zero	2292	23.76%
(II) Cliente suscetível ao evento	5268	54.62%
(III) Cliente não suscetível ao evento	2085	21.62%
Total	9645	100%

Deste modo, a Tabela 4.2 mostra que há uma concentração de eventos no tempo zero, sendo cerca de 23.76% das observações, identificando o excesso de zeros. Ademais, cerca de 21.62% dos clientes não apresentaram o evento de interesse, que pela teoria, são considerados imunes. Por fim, cerca de 54.62% dos clientes apresentaram o evento de interesse, ou seja, quitaram suas dívidas dentro do período de 24 meses.

A Figura 4.1 representa a distribuição dos tempos de regularização das dívidas para o conjunto de dados observado.

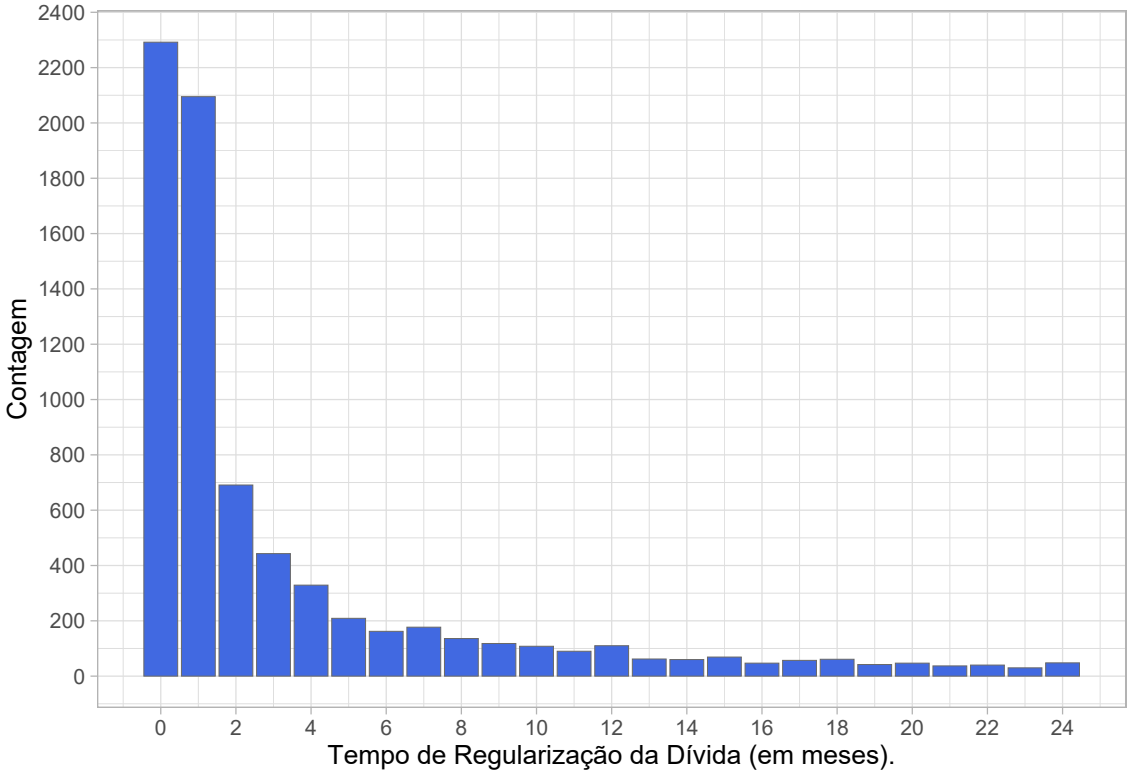

Figura 4.1: Gráfico referente ao Tempo de Regularização da Dívida (em meses).

Nesse sentido, é possível notar na Figura 4.1, a inflação de zeros para este conjunto de dados. Essa característica é interessante, uma vez que o estudo está sendo realizado em um cenário de crise econômica e grande parte dos clientes endividados realizaram a quitação de dívidas logo no início do estudo. Isto pode ser dado, devido ao interesse dos clientes na normalização de seu *status* para realizações de outros ações, na qual a inadimplência poderia impedir.

Na Figura 4.2 têm-se a curva de Kaplan-Meier estimada para os tempos de regularização da dívida dos clientes. É possível verificar uma grande quantidade de censuras à direita, ou seja, uma grande quantidade de clientes que não quitaram suas dívidas dentro do período de 24 meses. Além disso, é possível identificar que não há a estabilização da

curva de sobrevivência no ponto em que $\hat{S}(t) = 0$, observando-se então a indicação da presença de fração de cura para esses dados.

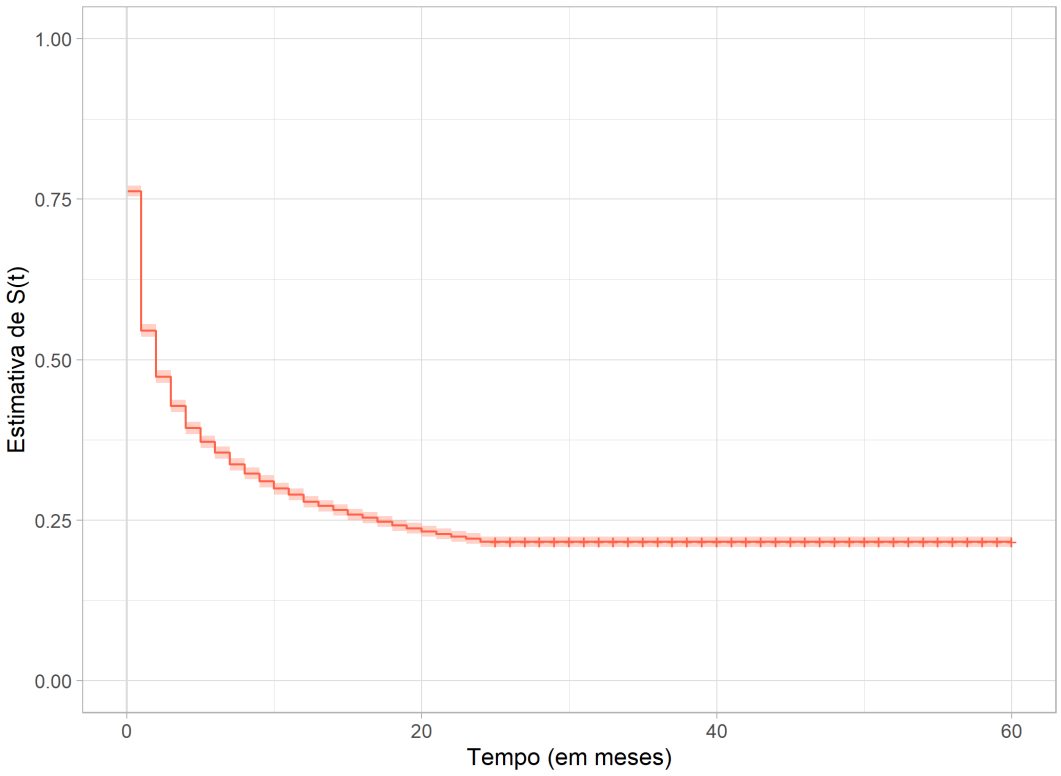

Figura 4.2: Gráfico de Kaplan-Meier para os Tempos de Regularização da Dívida (em meses).

Além disso, é importante salientar que a curva de sobrevivência estimada na Figura 4.2 começa aproximadamente no ponto 0.75, devido a presença de inflação de zeros visto na Figura 4.1.

Na construção das curvas de Kaplan-Meier, um outro ponto relevante são as curvas de maneira estratificada por covariáveis, como é possível verificar na Figura 4.3, em que existem diferenças das curvas para diferentes categorias dentro da covariável, assim, representando uma diferença nos tempos de sobrevida.

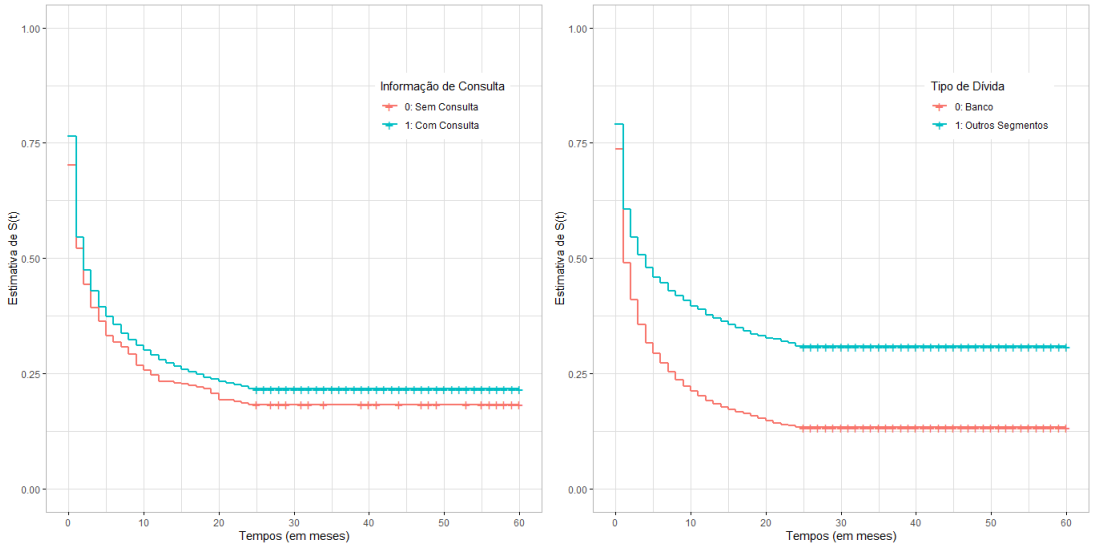

Figura 4.3: Gráfico de Kaplan-Meier considerando as covariáveis: 1) Informação de Consulta aos Relatórios de Crédito; 2) Segmento de Dívida Adquirida.

Dessa forma, na Figura 4.3 é possível destacar a covariável Tipo de Dívida, em que os clientes que possuem dívidas em bancos, realizam a quitação de suas dívidas em uma maior proporção quando comparado as dívidas que são provenientes de outros segmentos. Em relação a covariável Informação de Consulta, é possível verificar que clientes que não possuem consultas em seus relatórios de crédito tendem a priorizar mais o pagamento da dívida em relação daqueles que aqueles que possuem consulta.

4.1 Ajuste dos Modelos Defeituosos Zero Ajustados Gompertz e Gaussiana-Inversa

Nesta seção, a metodologia proposta na Seção 3 foi aplicada ao conjunto de dados. Primeiramente, foram ajustados os modelo sem a presença de covariáveis e, posteriormente, os modelos com a presença das covariáveis separadamente. Por fim, será realizado o ajuste do modelo com todas as covariáveis conjuntamente.

Os resultados dos ajustes dos modelos serão mostrados em formas de tabelas com as estimativas dos parâmetros, erro padrão e intervalos de confiança. Ademais, será utilizado um suporte de análise gráfica dos ajustes realizados. Por fim, para realizar a escolha do melhor modelo, será utilizado duas métricas para mensurar sua qualidade, sendo o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC).

Com o objetivo de simplificar as interpretações, considere que os coeficientes β_{0i} estão

associados à influência das covariáveis na inflação de zeros, enquanto os coeficientes β_{1i} estão relacionados à influência das mesmas covariáveis no parâmetro b das distribuições Gompertz e Gaussiana-Inversa. Além disso, temos as seguintes relações, a proporção de zeros, p_{0i}, é dada por,

$$p_{0i} = \frac{\exp\{\beta_{00} + x_{1i}\beta_{01} + x_{2i}\beta_{02}\}}{1 + \exp\{\beta_{00} + x_{1i}\beta_{01} + x_{2i}\beta_{02}\}},$$

e o parâmetro b está relacionado com o parâmetro b das distribuições Gompertz e Gaussiana-Inversa e é expresso como,

$$b(x) = \exp\{\beta_{1i} + x_{1i}\beta_{11} + x_{2i}\beta_{12}\}.$$

O parâmetro a também é um parâmetro presente nas distribuições Gompertz e Gaussiana-Inversa, e a proporção de cura p_{1i} é obtida através das equações 3.3 e 3.4 dos modelos defeituosos zero-ajustados Gompertz e Gaussiana-Inversa, respectivamente. As estimativas para os erros-padrões para as proporções estimadas foram determinadas através do método delta (Oliveira *et al.*, 1997).

4.1.1 Ajuste do modelo sem a presença de covariáveis

A Tabela 4.3 apresenta os resultados das estimativas de parâmetros, erro padrão e intervalos de confiança de 95% obtidos no ajuste do modelos defeituosos zeros ajustados Gompertz e Gaussiana-Inversa, respectivamente, sem a presença de covariáveis.

Tabela 4.3: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), intervalo de confiança para os Modelos Defeituosos Zero Ajustados Gompertz e Gaussiana-Inversa sem covariáveis.

Parâmetros	Modelos Defeituosos Zero Ajustados							
	Gompertz				Gaussiana-Inversa			
	EMV	EP	LI	LS	EMV	EP	LI	LS
a	-0.147	0.002	-0.152	-0.142	-0.065	0.003	-0.071	-0.059
b	0.187	0.007	0.173	0.200	0.454	0.007	0.440	0.467
$\beta_{00_{(intercepto)}}$	-1.166	0.024	-1.213	-1.119	-1.166	0.024	-1.213	-1.119
$\beta_{10_{(intercepto)}}$	-1.679	0.018	-1.715	-1.643	-0.790	0.018	-0.826	-0.754
p_0	0.238	0.004	0.230	0.245	0.238	0.004	0.230	0.245
p_1	0.213	0.004	0.206	0.221	0.190	0.007	0.177	0.204

A partir da Tabela 4.3 é possível verificar que as estimativas dos parâmetros associados a proporção de zeros (p_0), é bem próxima para ambos os modelos. Em relação ao parâmetro da proporção de cura (p_1), nota-se, que no modelo Gompertz a proporção de cura encontrada é superior ao modelo Gaussiana-Inversa. Um ponto relevante é que, no geral, todos os parâmetros são significativos, ao nível de significância de 5%, uma vez que as regiões de confiança, não englobam o valor zero.

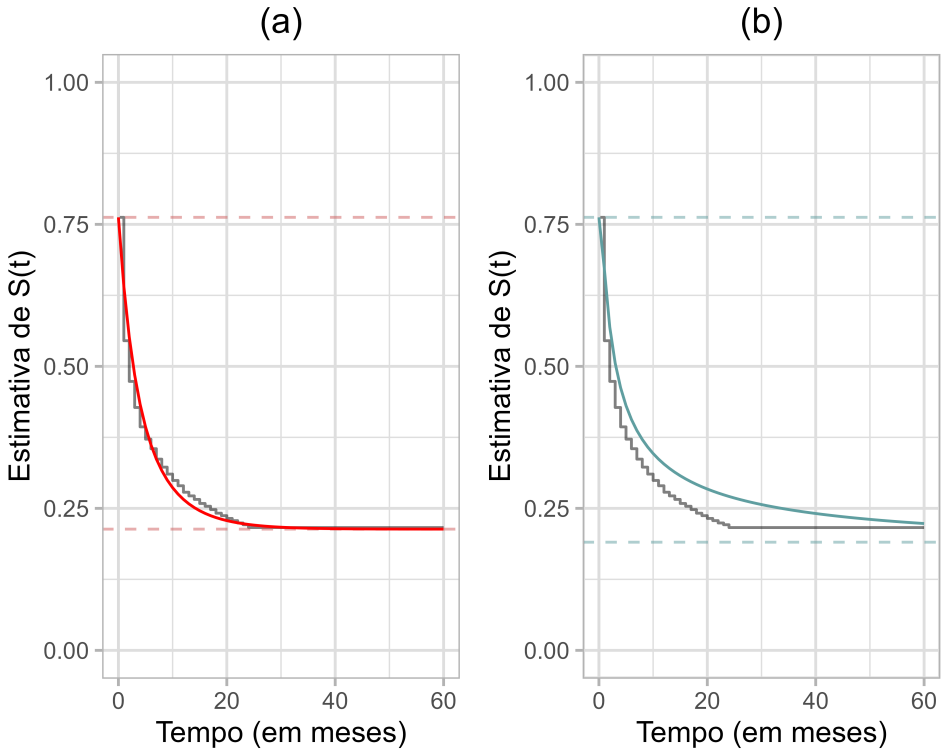

Figura 4.4: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), sem a presença de covariável.

A Figura 4.4 mostra o ajuste dos modelos defeituosos zero ajustados Gompertz (a) e Gaussiana-Inversa (b), respectivamente, sem a presença de covariáveis. Nesse sentido, é possível verificar que o modelo Gompertz (a), obteve um ajuste melhor quando comparado ao modelo Gaussiana-Inversa (b), uma vez que a curva de sobrevivência estimada pelo modelo Gompertz está muito próxima da curva estimada de Kaplan-Meier.

Ao analisar as funções de sobrevivência representadas pelas Equações 3.2.1 e 3.2.2, é possível estabelecer uma relação com a função de risco acumulado populacional, em que $\hat{H}_{pop} = -\log(\hat{S}_{pop}(t))$. Na Figura 4.5, têm-se as curvas estimadas da função de risco acumulada para cada um dos modelos.

Nesse sentido, ao analisar a Figura 4.5, na qual o modelo Gompertz demonstrou o

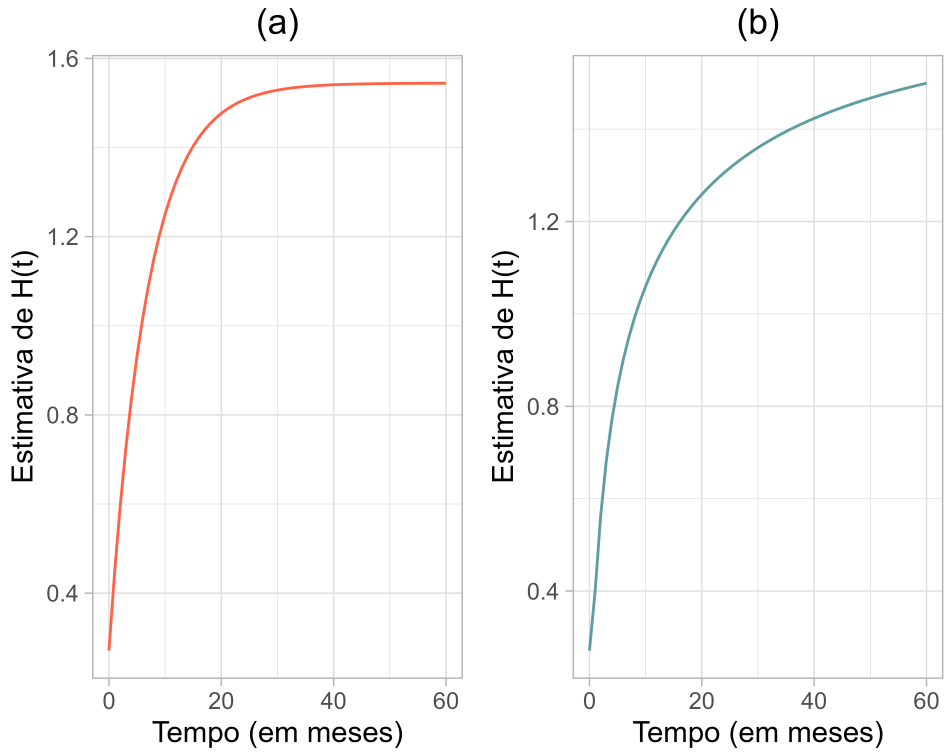

Figura 4.5: Estimativa da função de risco acumulada ($H(t)$) pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), sem a presença de covariável.

melhor ajuste aos dados, podemos observar que há um maior risco do indivíduo, que adquiriu uma dívida, quitar sua dívida com maior chance até o 30º mês, visto que a curva acumulada estimada se estabiliza logo após esse ponto. Entretanto, vale ressaltar que, ao quitar a dívida em até 35 meses ou em até 60 meses, têm-se quase o mesmo risco.

4.1.2 Ajuste do modelo com a presença separadamente das covariáveis

A Tabela 4.4 mostra a estimativa dos parâmetros, erro-padrão e seus intervalos de confiança de 95% para a variável de "Informação de Consulta" para cada um dos modelos propostos.

Novamente, na Tabela 4.4 verifica-se que as estimativas dos parâmetros associados a proporção de zeros (p_0), são bem próximas para ambos os modelos. Nesse sentido, pode-se afirmar que a covariável "Informação de Consulta" possui uma maior inflação de zeros $p_{00} = 0.298$ para clientes que não obtiveram nenhuma consulta de empresas ao seu relatório de crédito.

Tabela 4.4: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), intervalo de confiança para os Modelos Defeituosos Zero Ajustados Gompertz e Gaussiana-Inversa para a covariável x_1.

Parâmetros	Modelos Defeituosos Zero Ajustados							
	Gompertz				Gaussiana-Inversa			
	EMV	EP	LI	LS	EMV	EP	LI	LS
a	-0.147	0.002	-0.152	-0.142	-0.065	0.003	-0.071	-0.059
b	0.186	0.054	0.171	0.292	0.454	0.008	0.438	0.469
$\beta_{00_{(intercepto)}}$	-0.855	0.127	-1.105	-0.606	-0.856	0.127	-1.105	-0.607
$\beta_{01_{(X_1=1)}}$	-0.321	0.130	-0.575	-0.067	-0.320	0.130	-0.574	-0.066
$\beta_{10(intercepto)}$	-1.641	0.082	-1.802	-1.481	-0.800	0.094	-0.985	-0.615
$\beta_{11_{(X_1=1)}}$	-0.038	0.082	-0.199	0.122	0.010	0.095	-0.176	0.196
p_{00}	0.298	0.027	0.245	0.351	0.298	0.027	0.245	0.351
p_{01}	0.236	0.004	0.228	0.244	0.236	0.004	0.228	0.244
p_{10}	0.187	0.021	0.146	0.228	0.177	0.017	0.143	0.210
p_{11}	0.214	0.004	0.206	0.222	0.191	0.007	0.177	0.205

Em relação a proporção de cura, a maior proporção é dada pela modelagem utilizando a distribuição Gompertz $p_{11} = 0.214$ para clientes que obtiveram consultas de empresas ao seu relatório de crédito, já a menor proporção de cura é dada utilizando a distribuição Gaussiana-Inversa $p_{10} = 0.177$, para clientes que não obtiveram nenhuma consulta.

Além disso, é possível constatar que a maioria dos parâmetros associados aos modelos são significativos, uma vez que a grande maioria das regiões de confiança estabelecidas não inclui o valor zero.

A Figura 4.6 mostra a curva de sobrevivência estimada pelos modelos defeituosos zero ajustados Gompertz (a) e Gaussiana-Inversa (b), respectivamente, com a presença da covariável Informação de Consulta aos relatórios de Crédito. Nota-se que, para a covariável de Informação de Consulta, o modelo Gompertz apresenta um ajuste mais adequado aos dados em comparação ao modelo Gaussiana-Inversa. O destaque para o modelo Gompertz se deve ao fato de que sua curva de sobrevivência estimada se aproximar significativamente da curva de sobrevida estimada de Kaplan-Meier.

Na Figura 4.7, têm-se as curvas estimadas da função de risco acumulada populacional, $\hat{H}_{pop} = -\log(\hat{S}_{pop}(t))$, para cada um dos modelos.

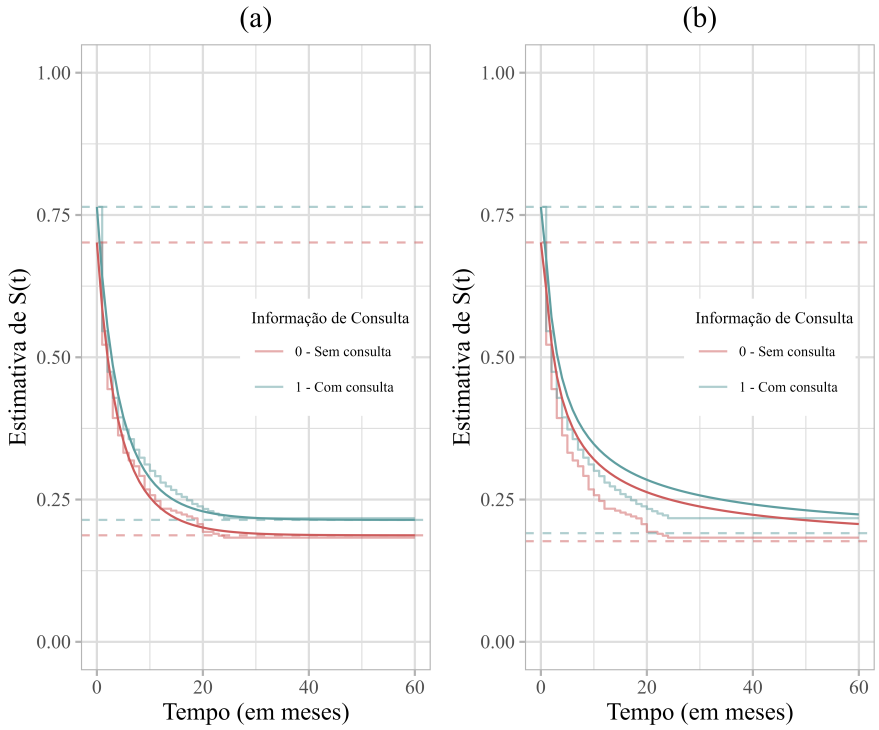

Figura 4.6: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), com a presença da covariável de Informação de Consulta aos relatórios de Crédito.

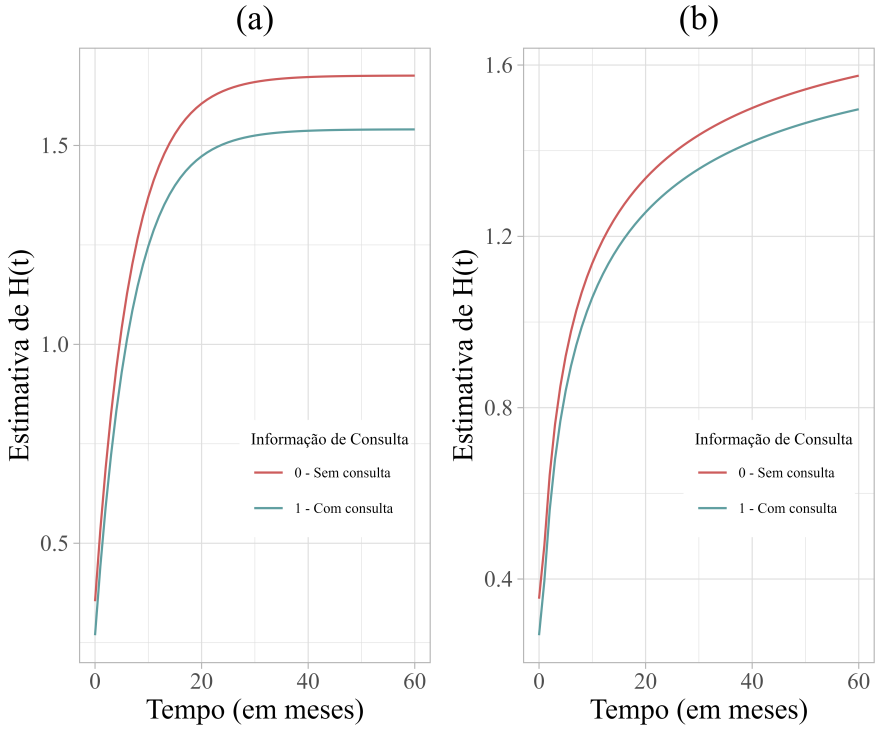

Figura 4.7: Estimativa da função de risco acumulada ($H(t)$) pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), com a presença da covariável de Informação de Consulta aos relatórios de Crédito.

É possível observar, na Figura 4.7, que o risco de um indivíduo em um determinado instante de tempo quitar a sua dívida, é maior para os clientes que não obtiveram nenhuma consulta aos seus relatórios de crédito.

A Tabela 4.5 mostra a estimativa dos parâmetros, erro-padrão e seus intervalos de confiança de 95% para a variável de "Tipo de Dívida" para cada um dos modelos propostos.

Tabela 4.5: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), intervalo de confiança para os Modelos Defeituosos Zero Ajustados Gompertz e Gaussiana-Inversa para a covariável x_2.

Parâmetros	Modelos Defeituosos Zero Ajustados							
	Gompertz				Gaussiana-Inversa			
	EMV	EP	LI	LS	EMV	EP	LI	LS
a	-0.142	0.002	-0.146	-0.137	-0.066	0.003	-0.071	-0.060
b	0.136	0.010	0.116	0.155	0.367	0.010	0.347	0.387
$\beta_{00_{(intercepto)}}$	-1.030	0.032	-1.092	-0.967	-1.030	0.032	-1.092	-0.967
$\beta_{02_{(X_2=1)}}$	-0.302	0.048	-0.397	-0.207	-0.302	0.048	-0.397	-0.207
$\beta_{10_{(intercepto)}}$	-1.426	0.021	-1.467	-1.481	-0.606	0.024	-0.653	-0.558
$\beta_{12_{(X_2=1)}}$	-0.571	0.028	-0.626	-0.516	-0.397	0.031	-0.458	-0.335
p_{00}	0.263	0.006	0.251	0.275	0.263	0.006	0.251	0.275
p_{01}	0.209	0.006	0.197	0.221	0.209	0.006	0.197	0.221
p_{10}	0.135	0.005	0.125	0.145	0.158	0.006	0.146	0.170
p_{11}	0.303	0.007	0.290	0.317	0.238	0.008	0.223	0.254

Em relação a Tabela 4.5, verifica-se também que as estimativas dos parâmetros associados a proporção de zeros (p_0) são bem próximas para ambos os modelos. Nesse caso, a variável Tipo de Dívida, possui uma maior inflação de zeros $p_{00} = 0.263$ para clientes com dívidas em bancos, enquanto a menor inflação de zeros $p_{01} = 0.209$ para clientes que possui dívidas de outros segmentos. Além disso, em ambos os modelos a maior proporção de cura é dada para clientes que possuem dívidas em outros segmentos, enquanto a menor proporção de cura é dada para clientes com dívidas em bancos.

Também é possível observar que todos os parâmetros dos modelos ajustados considerando a covariável Tipo de Dívida, são significativos, visto que as regiões de confiança estabelecidas em ambos os modelos, não contemplam o valor zero.

A Figura 4.8 representa a curva de sobrevivência estimada pelos modelos defeituosos zero ajustados Gompertz (a) e Gaussiana-Inversa (b), respectivamente, para a covariável

Tipo de Dívida.

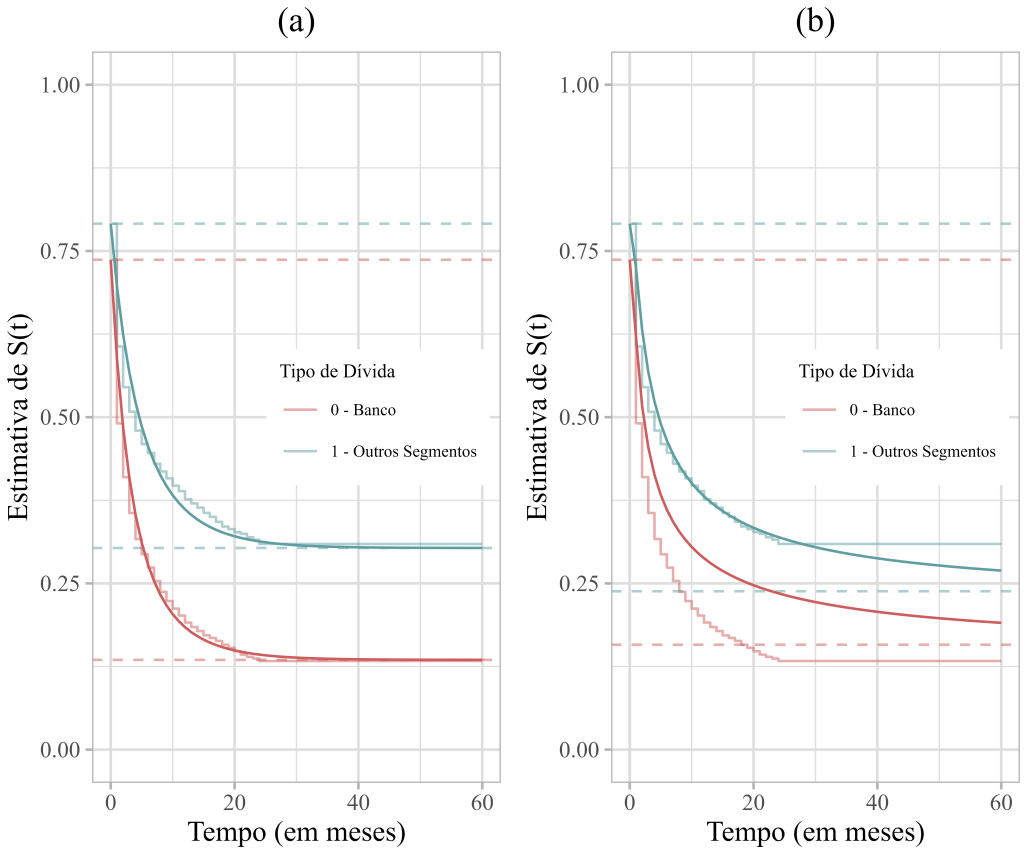

Figura 4.8: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), com a presença da covariável de Tipo de Dívida.

Novamente, para a covariável Tipo de Dívida, é possível verificar que o melhor ajuste ocorre no modelo Gompertz em comparação ao modelo Gaussiana-Inversa. A partir da Figura 4.8, , nota-se que as curvas de sobrevivência estimadas para dívidas em bancos e outros segmentos, a partir do modelo Gaussiana-Inversa, estão muito distantes das curvas de sobrevivência estimadas por Kaplan-Meier.

A Figura 4.9, mostra as curvas estimadas da função de risco acumulada populacional, $\hat{H}_{pop} = -\log(\hat{S}_{pop}(t))$, para cada um dos modelos, considerando a covariável Tipo de Dívida. Nesse sentido, a partir da Figura 4.9, nota-se que o risco de um indivíduo em um determinado instante de tempo quitar sua dívida, é maior para os clientes que possuem dívidas no segmento financeiro, ou seja, em bancos. Em contrapartida, o risco de um indivíduo quitar a dívida em um determinado instante de tempo, é menor para clientes que possuem dívidas em outros segmentos.

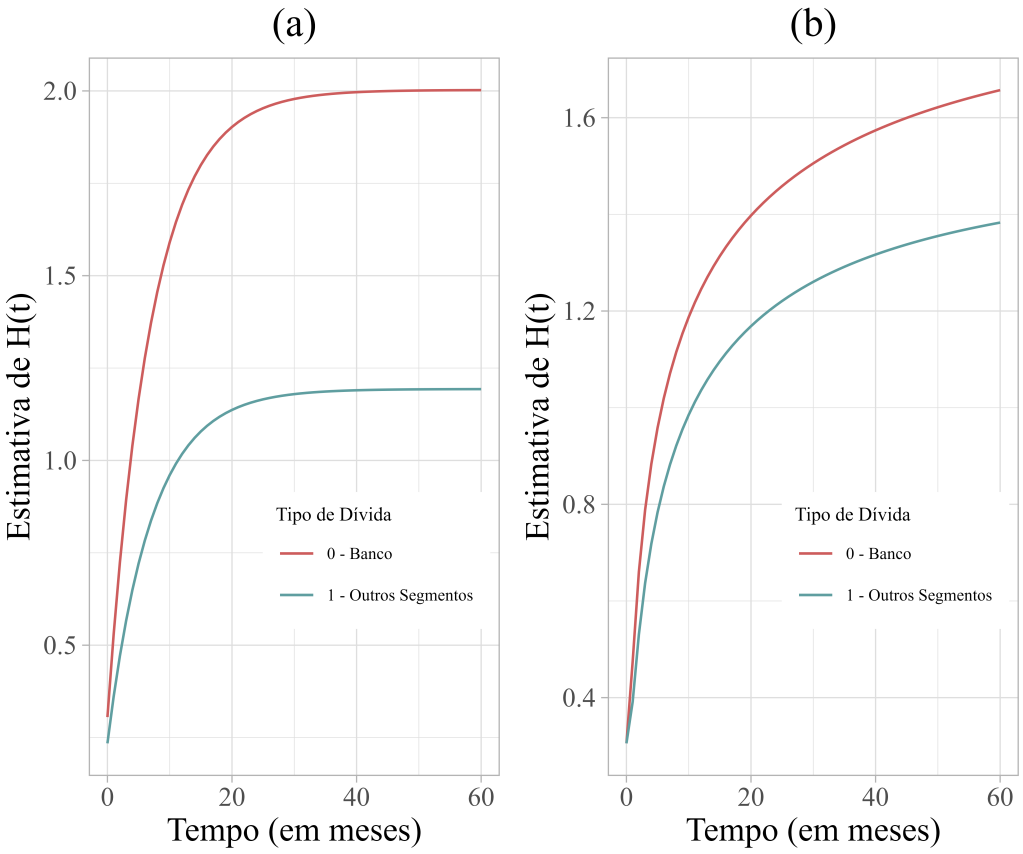

Figura 4.9: Estimativa da função de risco acumulada ($H(t)$) pelo Modelo Defeituoso Zero Ajustado Gompertz (a) e Modelo Defeituoso Zero Ajustado Gaussiana-Inversa (b), com a presença da covariável de Tipo de Dívida.

4.1.3 Ajuste do modelo com a presença das covariáveis de forma conjunta

Nesse contexto, após realizar os ajustes dos modelos considerando as covariáveis separadamente, será considerado o ajuste dos modelos para as duas variáveis conjuntamente. Na Tabela 4.6 têm-se o ajuste dos modelos defeituosos zero ajustados Gompertz e Gaussiana-Inversa, considerando ambas covariáveis.

Ao observar a Tabela 4.6, nota-se que os parâmetros associados as distribuições Gompertz e Gaussiana-Inversa são significativos, além disso, verifica-se que a maioria das estimativas dos parâmetros de regressão β ligados ao parâmetro b e a proporção de zero foram significativos, ao considerar o mesmo critério das regiões de confiança, como visto nos modelos anteriores.

Tabela 4.6: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), intervalo de confiança para os Modelos Defeituosos Zero Ajustados Gompertz e Gaussiana-Inversa para as covariáveis x_1 e x_2 de forma conjunta.

Parâmetros	Modelos Defeituosos Zero Ajustados							
	Gompertz				Gaussiana-Inversa			
	EMV	EP	LI	LS	EMV	EP	LI	LS
a	-0.142	0.002	-0.146	-0.137	-0.066	0.003	-0.071	-0.060
b	0.136	0.010	0.188	0.227	0.367	0.010	0.348	0.387
$\beta_{00_{(intercepto)}}$	-0.723	0.129	-0.976	-0.470	-0.721	0.129	-0.974	-0.468
$\beta_{01_{(X_1=1)}}$	-0.317	0.130	-0.572	-0.063	-0.318	0.130	-0.573	-0.064
$\beta_{02_{(X_2=1)}}$	-0.301	0.048	-0.396	-0.206	-0.303	0.048	-0.398	-0.208
$\beta_{10_{(intercepto)}}$	-1.397	0.082	-1.558	-1.235	-0.618	0.095	-0.805	-0.431
$\beta_{11_{(X_1=1)}}$	-0.030	0.082	-0.191	0.131	0.011	0.095	-0.175	0.197
$\beta_{12_{(X_2=1)}}$	-0.571	0.028	-0.626	-0.516	-0.395	0.031	-0.456	-0.334

Na Tabela 4.7, têm-se as estimativas das proporções de zero e proporções de cura para o modelo defeituoso zero ajustado Gompertz para as covariáveis Informação de Consulta e Tipo de Dívida.

Tabela 4.7: Estimativas das proporções de zeros e de cura para o Modelo Defeituoso Zero Ajustado Gompertz para as covariáveis x_1 e x_2 de forma conjunta.

Proporção de Zeros e Cura	x_1	x_2	Estimativa	Erro Padrão	I.C. 95%	
					LI	LS
p_0	0	0	0.3267	0.028	0.2718	0.3816
		1	0.2642	0.025	0.2152	0.3132
	1	0	0.2611	0.006	0.2494	0.2729
		1	0.2073	0.006	0.1955	0.2190
p_1	0	0	0.1173	0.018	0.0820	0.1526
		1	0.2742	0.024	0.2271	0.3212
	1	0	0.1356	0.005	0.1258	0.1454
		1	0.3041	0.007	0.2904	0.3178

Nesse contexto, é possível observar através da Tabela 4.7 que a maior proporção de indivíduos que regularizam sua dívida no tempo zero, está associado aos clientes que não obtiveram alguma consulta aos seus relatórios de crédito por empresas e possuem dívidas do segmento financeiro, ou seja, em bancos com uma proporção de $p_{000} = 0.3267$. Em contrapartida, a menor proporção de indivíduos que regularizam sua dívida no tempo zero, está associado aos clientes que obtiveram alguma consulta aos seus relatórios de

crédito e que possuem dívidas advindas de outros segmentos.

Além disso, é notável que os clientes que tiveram seus relatórios de crédito consultados por alguma empresa e possuem dívidas originadas de outros segmentos são aqueles que apresentam a maior concentração de indivíduos que não quitaram suas dívidas dentro do período de 24 meses, uma vez que a proporção de cura é dada por $p_{111} = 0.3041$. No que diz respeito aos clientes cujos relatórios de crédito não foram consultados por nenhuma empresa e que possuem dívidas provenientes de bancos, é importante notar que eles apresentam a menor quantidade de dívidas pendentes, já que a proporção de cura é $p_{100} = 0.1173$.

Em Toledo *et al.* (2022), foi utilizado o mesmo conjunto de dados, porém com mais covariáveis. Foi apresentado um resumo das proporções de zeros e de cura das covariáveis "Informação de Consulta" e "Tipo de Dívida" que foram selecionadas como o melhor modelo, utilizando a metodologia do modelo de fração de cura zero ajustado proposto por Ribeiro de Oliveira Jr *et al.* (2017), em que a função de sobrevivência é dada pela equação 3.1.

Nesse contexto, ao comparar os resultados obtidos por Toledo *et al.* (2022) e os resultados na Tabela 4.7, podemos constatar que as estimativas das proporções de zeros e cura são bastante semelhantes, o que evidencia a eficácia do modelo. É relevante destacar que a metodologia empregada neste estudo possui a vantagem de ter a necessidade de estimar apenas os parâmetros do modelo defeituoso e a proporção de zeros (p_0), enquanto na outra metodologia, era necessário estimar a proporção de zeros (p_0), a proporção de cura (p_1) e os parâmetros dos modelos de sobrevivência base.

Por fim, através da Tabela 4.7, foi observado quais são os padrões de clientes que tendem a pagar ou não suas dívidas dentro do período de 24 meses, utilizando o modelo defeituoso zero ajustado Gompertz. Ademais, todas as proporções de zeros e cura são significativas.

Já na Tabela 4.8, têm-se as estimativas das proporções de zero e proporções de cura para o modelo defeituoso zero ajustado Gaussiana-Inversa para as covariáveis Informação de Consulta e Tipo de Dívida.

Através da Tabela 4.8, as conclusões das proporções de zeros e proporções cura dadas para o modelo defeituoso zero ajustado Gompertz é semelhante as conclusões dadas para o modelo defeituoso zero ajustado Gaussiana-Inversa. Isso ocorre devido à proximidade dos valores encontrados em relação aos valores apresentados na Tabela 4.7. Além disso,

Tabela 4.8: Estimativas das proporções de zeros e de cura para o Modelo Defeituoso Zero Ajustado Gaussiana-Inversa para as covariáveis x_1 e x_2 de forma conjunta.

Proporção de Zeros e Cura	x_1	x_2	Estimativa	Erro Padrão	I.C. 95%	
					LI	LS
p_0	0	0	0.3271	0.028	0.2722	0.3820
		1	0.2641	0.025	0.2151	0.3131
	1	0	0.2613	0.006	0.2495	0.2730
		1	0.2070	0.006	0.1953	0.2188
p_1	0	0	0.1455	0.014	0.1180	0.1729
		1	0.2232	0.020	0.1840	0.2624
	1	0	0.1581	0.006	0.1464	0.1699
		1	0.2384	0.008	0.2227	0.2540

nesse caso todas as proporções de zeros e cura também são significativas.

4.2 Critérios de Seleção de Modelos

Nesta Seção, serão apresentados o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), para a escolha do modelo mais adequado, representado pela Tabela 4.9.

Tabela 4.9: Critérios de seleção para os modelos ajustados.

Modelo Defeituoso Zero Ajustado	Critério	Covariáveis		
		x_1	x_2	x_1 e x_2
Gompertz	AIC	-46305.09	-45851.15	-45841.30
	BIC	-46223.35	-45769.41	-45726.86
Gaussiana-Inversa	AIC	-45249.05	-45055.32	-45045.59
	BIC	-45167.31	-44973.57	-44931.15

A partir da Tabela 4.9, verifica-se que o modelo defeituoso zero ajustado Gompertz, considerando a covariável Informação de Consulta, exibe os menores valores de AIC e BIC. Dessa forma, é possível verificar que realmente o critério faz sentido, pois através da Figura 4.10, identifica-se que as curvas de sobrevivência estimadas estão muito próximas das curvas estimadas de Kaplan-Meier.

Outro ponto observado, é que o modelo defeituoso zero ajustado Gaussiana-Inversa, considerando as covariáveis separadamente e conjuntamente, foram os que obtiveram a menor performance. Portanto, pode-se concluir que o modelo mais adequado a ser sele-

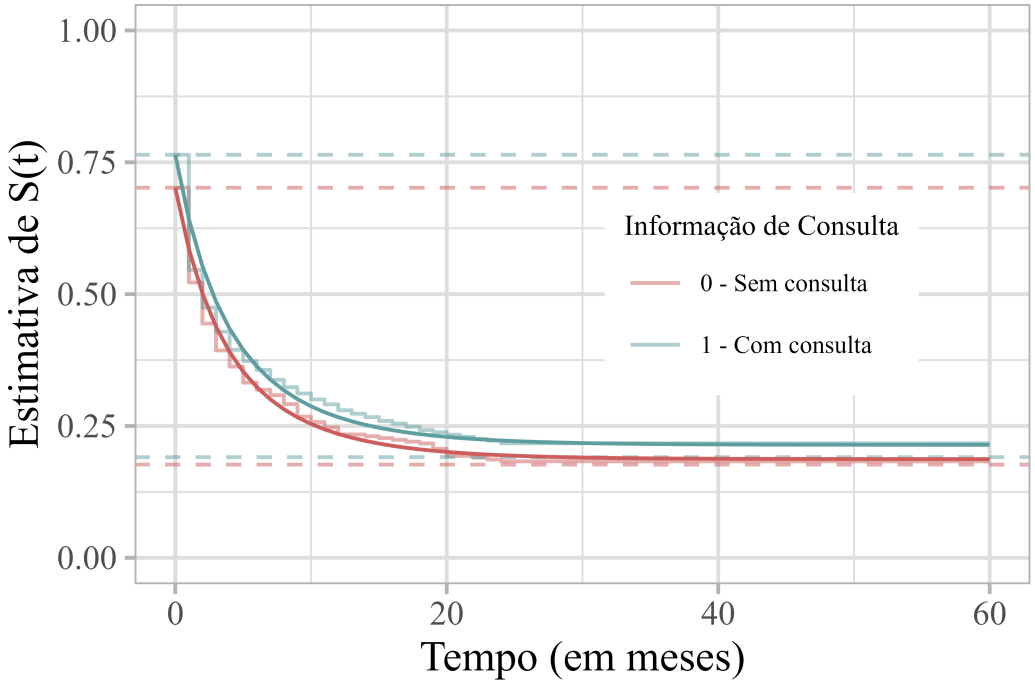

Figura 4.10: Modelo Defeituoso Zero Ajustado Gompertz com a covariável de Informação de Consulta.

cionado é o modelo defeituoso zero ajustado Gompertz, na presença da covariável "Informação de Consulta".

4.3 Considerações Finais

Neste Capítulo foi possível realizar uma aplicação a um conjunto de dados reais, em que foi visto que o conjunto atende as principais características da modelagem, sendo a fração de cura, representando os clientes inadimplentes que não quitaram suas dívidas dentro do período de 24 meses, e a proporção de zeros, representando os clientes que regularizaram suas dívidas no início do estudo, ou seja, no tempo zero.

Capítulo 5

Conclusão

Neste trabalho, foi estudado modelos estatísticos de sobrevivência denominado Modelos de regressão defeituosos zero ajustado. Esses modelos têm duas características principais que os diferenciam de modelos usuais de sobrevivência: a incorporação de uma parcela dos indivíduos que não apresentam o evento de interesse, mesmo após um longo tempo de acompanhamento, e também a possibilidade de que uma proporção dos tempos em estudo seja iguais a zero.

Para ilustrar a metodologia aqui apresentada, analisamos os dados de sobrevivência de um banco de dados real de clientes que adquiriram dívidas, entre os meses de julho e dezembro de 2015, que foi disponibilizado pela Serasa Experian, uma instituição líder em informações e serviços de crédito no Brasil. O modelo possibilitou estimar as proporções de três grupos de clientes em um conjunto de dados específico: um grupo em que o tempo é igual a zero (clientes que liquidaram suas dívidas no tempo zero, recuperando imediatamente sua capacidade de pagamento); outro grupo com clientes suscetíveis ao evento de interesse (clientes que quitaram suas dívidas ao longo do tempo, recuperando, em seguida, sua capacidade de pagamento); e um grupo de clientes não suscetíveis ao evento (clientes que não liquidaram suas dívidas).

Os resultados mostraram que o modelo defeituoso zero ajustado Gompertz apresentou um melhor desempenho. Os critérios avaliados para seleção do melhor modelo foi medida pelo AIC e BIC. No entanto, é importante ressaltar que o desempenho real dos modelos apresentado aqui poderá ser avaliado considerando seu uso diário pelas empresas, aplicando uma maior variedade de dados e covariáveis disponíveis, uma vez que o modelo permite o uso de quantas covariáveis forem necessárias, sejam elas contínuas ou categóricas.

Além disso, foi visto que a metodologia apresentada aqui é semelhante a metodologia dos modelos de taxa de cura zero ajustado estudado por Toledo *et al.* (2022). No entanto, os modelos defeituosos zero ajustados possuem uma vantagem significativa, pois requerem a estimação de um parâmetro a menos, os parâmetros do modelo defeituoso e o da proporção de zeros, ou seja, a proporção de clientes que quitaram suas dívidas no início do estudo.

No desfecho deste estudo, constatou-se que é possível adquirir conhecimento adicional, o que leva à conclusão de que podemos utilizar a técnica de análise de sobrevivência para estimar e escolher um modelo eficiente em carteiras de clientes com acesso a crédito, como as de grandes bancos ou varejistas. Por fim, uma alternativa adicional que pode ser empregada é a análise de resíduos, possibilitando a avaliação da adequação do modelo proposto.

Referências Bibliográficas

Aalen, O. O., Borgan e Gjessing, H. K. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Andersen, P., Gill, R. e Keiding, N. (1993). *Statistical models based on counting process*. Springer Series in Statistics. Springer, second edition.

Balka, J., Desmond, A. F. e McNicholas, P. D. (2011). Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, **38**(1), 127–144.

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

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

Bolfarine, H. e Sandoval, M. C. (2001). *Introdução à inferência estatística*, volume 2. SBM.

Borges, P., Rodrigues, J. e Balakrishnan, N. (2012). Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. *Computational Statistics & Data Analysis*, **56**(6), 1703–1713.

Calsavara, V. F., Rodrigues, A. S., Rocha, R., Louzada, F., Tomazella, V., Souza, A. C., Costa, R. A. e Francisco, R. P. (2019a). Zero-adjusted defective regression models for modeling lifetime data. *Journal of Applied Statistics*, **46**(13), 2434–2459.

Calsavara, V. F., Rodrigues, A. S., Rocha, R., Tomazella, V. e Louzada, F. (2019b). Defective regression models for cure rate modeling with interval-censored data. *Biometrical Journal*, **61**(4), 841–859.

Cancho, V. G., Bandyopadhyay, D., Louzada, F. e Yiqi, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme. *Statistical methodology*, **13**, 48–68.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, S. E. (2011). *Análise de sobrevivência: teoria e aplicações em saúde*. SciELO-Editora FIOCRUZ.

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

Colosimo, E. e Giolo, S. (2006). *Análise de sobrevivência aplicada.*. ABE - Projeto Fisher.

da Silva, J. P. (2000). *Gestão e análise de risco de crédito* .. Editora Atlas SA.

Feller, W. (1991). *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons.

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J. e Pullen, J. (1998). Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, **17**(8), 831–839.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, (115), 513–583.

Granzotto, D. C. T. *et al.* (2008). Seleção de modelos de tempos com longa-duração para dados de finanças.

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

Ibrahim, J., Chen, M. e Sinha, D. (2014). Bayesian survival analysis. wiley statsref: Statistics reference online.

Jorion, P. (2007). *Value at risk: the new benchmark for managing financial risk*. The McGraw-Hill Companies, Inc.

Kalbfleisch, J. e Prentice, R. (2002). *The statistical analysis of failure time data*. Series in Probability and Statistics. John Wiley & Sons, second edition.

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

Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B. *et al.* (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, **317**(8), 461–467.

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

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. e Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, **8**(11), 1235–1246.

Oliveira, N. F. d., Santana, V. S. e Lopes, A. A. (1997). Razões de proporções e uso do método delta para intervalos de confiança em regressão logística. *Revista de Saúde Pública*, **31**(1), 90–99.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro de Oliveira Jr, M., Moreira, F. e Louzada, F. (2017). The zero-inflated promotion cure rate model applied to financial data on time-to-default. *Cogent Economics & Finance*, **5**(1), 1395950.

Rocha, R., Nadarajah, S., Tomazella, V. e Louzada, F. (2017). A new class of defective models based on the marshall–olkin family of distributions for cure rate modeling. *Computational Statistics & Data Analysis*, **107**, 48–63.

Rodrigues, J., Cancho, V. G., de Castro, M. e Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**(6), 753–759.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, páginas 461–464.

Scudilio, J., Calsavara, V. F., Rocha, R., Louzada, F., Tomazella, V. e Rodrigues, A. S. (2019). Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*, **46**(3), 484–507.

Toledo, J. S., Tomazella, V. L. D., Lima, C. M. M. e Felix, M. H. (2022). Gompertz zero-inflated cure rate regression models applied to credit risk data. *Applied Stochastic Models in Business and Industry*.