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

Inferência para modelos de fração de cura zeros inflacionados aplicados a dados de risco de crédito

Matheus Henrique Felix

Trabalho de Conclusão de Curso

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

Inferência para modelos de fração de cura zeros inflacionados aplicados a dados de risco de crédito

Matheus Henrique Felix
Orientadora: Profª. Drª Vera Tomazella

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

São Carlos
26 de Novembro de 2021

Matheus Henrique Felix

Inferência para modelo de longa duração zeros inflacionados aplicados a dados de risco de crédito

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Matheus Henrique Felix e aprovado pela banca examinadora.
São Carlos, 28 de outubro de 2021.

Banca Examinadora

- Profª. Drª Vera Tomazella
- Dr. Oilson Alberto Gonzatto Junior
- Prof. Dr. Márcio Luis Lanfredi Viola

Agradecimentos

Durante a graduação pude crescer como pessoa e como profissional. Esta caminhada foi única porque estive rodeado de pessoas que me tornaram uma pessoa melhor. Gostaria de agradecer a minha mãe Maria, meu pai Adão, meus irmãos Lucas e Adan, minha irmã Sabrina, e a Giovanna Nesterick, Julia Beltramini, Graziela Valero, Leticia Cristina, Natalia Tomazella e ao Matheus Ramalho por todo o apoio que me deram durante este processo. Amo ter minha vida.

Expresso minha mais sincera gratidão a professora Vera Lucia Damasceno Tomazella pela atenção, respeito, assessoria e confiança. Você além de uma ótima orientadora é uma amiga!

Resumo

Dada a grande demanda para realização de concessões de crédito e bens de serviço, há necessidade de poder controlar o risco envolvido no processo. Essa medida visa administrar os possíveis eventos indesejados, por exemplo, a inadimplência, a fim de viabilizar a geração de lucro ou controlar os prejuízos para que não sejam superiores aos que a instituição financeira poderia suportar. Em tempos de crise, torna-se necessário cada vez mais o uso de ferramentas que possam auxiliar a tomada de decisão de forma mais confíavel. Assim, diversas técnicas estatísticas são utilizadas para construir modelos que possam expressar cenários de risco, entre elas existe a análise de sobrevivência, a qual tem por objetivo, por exemplo, prever situações como o tempo até indivíduos inadimplentes voltarem a seus status iniciais de adimplentes (recuperação de crédito). Com a aplicação destas técnicas, instituições financeiras podem basear-se nos resultados a fim de fornecer um valor de crédito ideal, para que não gere prejuízos para a mesma, assim como estimativas para a retomada das operações de créditos. Neste contexto, este trabalho tem por objetivo estudar o modelo de sobrevivência com fração de cura inflacionado de zero. Nesta abordagem é possível incorporar três classes de indivíduos: indíviduos com tempo igual a zero, não suscetíveis e suscetíveis ao evento de interesse. A metodologia proposta é aplicada a uma base de dados de uma empresa financeira.

Palavras-chave: *análise de sobrevivência, proporção de cura, modelos de longa duração, inflação de zeros, mercado financeiro, risco de credito, inadimplente.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Existem algumas problemáticas quando o objetivo é definir o que é risco dado a diversificação de campos em que este conceito se encontra. Entretanto, à medida em que observa-se estes acontecimentos, nota-se que eles estão relacionados a cenários de incerteza, cujos possíveis eventos de desfecho não podem ser assegurados, sendo associado a probabilidade ou possibilidades de ocorrência em um horizonte de tempo especificado. A fundamentação do que é risco não é vista por aquilo que está acontecendo, mas para o que acontecerá (Adam *et al.*, 2000).

Douglas e Wildavsky (1983) afirmam que risco é uma construção social em que, regularmente, delinea-se como algo não previsível, dado que, em sua maioria, não conseguimos ter realmente certeza de que os esforços realizados para evitar acidentes ou efeitos indesejados realmente serão satisfatórios e/ou seguros.

Em uma instituição financeira, quando ocorrem eventos de concessão de recursos, diz que tal começa a dispor-se do chamado risco de crédito. Este tipo de risco pode ser definido como a possibilidade de não ocorrência das obrigações contratuais advindas de seu devedor (Jorion, 2007), o qual passa a descumprir o contrato realizado com seu credor no momento da realização do acordo. A inadimplência é caracterizada quando obrigações contratuais não são cumpridas, as quais sempre ocorrem em operações de crédito, contudo existe a presença de perdas inesperadas derivadas das avaliações realizadas pelo agente contratante, assim melhor definindo o risco de crédito (Chaia, 2003).

A análise de crédito surge com a necessidade de saber para quais indivíduos conceder crédito para evitar eventos que possam gerar perdas a instituição. Este tipo de análise, embora leve em consideração a experiência do analista, apresenta algumas desvantagens. A primeira refere-se ao fato que a aprovação da solicitação de crédito está suscetível ao

analista que a faz, sendo assim, há uma grande possibilidade de determinações diferentes entre os mesmos, havendo aprovações para algumas análises e para outras não. A segunda é que o risco envolvido nas operações não é mensurado (Sicsú, 2010). Com isto, modelos de *credit scoring* se mostram proficientes por inúmeras situações, como consistências na tomada de decisões, capacidade de monitoramento, maior agilidade nos processos, entre outros. Tendo em vista a volatilidade dos cenários econômicos, a monitoração e revisão de modelos de riscos sempre devem ser realizadas. Silva (2000) assegura que:

> "as condições na análise de crédito são os fatores externos e macroeconômicos. Estes fatores externos, muitas vezes imprevisíveis, não são controláveis pela empresa. Mudanças na política econômica do governo podem afetar positivamente ou negativamente uma empresa. Toda empresa está envolvida em um sistema onde diversas forças e fatores exercem influência sobre ela. Exemplos disto são, as conjunturas nacionais e internacionais, o governo, o meio ambiente, a concorrência etc. O ramo de atividade também é um fator que influi na existência da empresa. Alguns ramos de atividade funcionam em uma cadeia e só atendem a um outro ramo, se este ramo entra em crise, com certeza a crise irá lhe afetar".

Uma crise ecônomica ocasiona eventos que causam impactos dentro de um país ou instituição, entre eles pode-se considerar o aumento do endividamento dos indivíduos devido ao aumento da inflação, altos indíces de desemprego e diminuição da conceção de crédito (Toledo, 2011). Neste contexto, a utilização de técnicas capazes de contribuir para a reabilitação do crédito é fundamental para entender como funciona o ciclo dentro deste setor, assim, podendo deduzir e entender como são dados os períodos de recuperação necessários para este mercado. Toledo (2011) afirma que a recuperação dentro de um sistema financeiro é lenta, necessitando assim de previsões que possam explicar bem a volatilidade dos processos de recuperação. Deste modo, a utilização de modelos matemáticos, em especial a Análise de Sobrevivência é um ferramenta útil para inferir e auxiliar nestas operações.

Análise de sobrevivência é um conglomerado de técnicas estatísticas frequentemente utilizada na área de saúde, ciências biológicas e engenharia. O tempo de ocorrência de um determinado evento é caracterizado como sendo a variável resposta, porém, para determinados indivíduos a não ocorrência do evento é possível, resultando nas censuras. Collett (2015) afirma que, por este motivo, as técnicas usuais da inferência clássica se tornam inviabilizada, dado que para estes modelos é assumido a ocorrência de todos os tempos de falha. Portanto, a utilização de modelos de análise de sobrevivência é visto com a capacidade de incorporação da informação dos dados censurados.

Uma grande vantagem vista com a utilização da análise de sobrevivência é que ela está atrelada com o espaço temporal de ocorrência de determinado acontecimento, possibilitando a previsão do evento de interesse. Para o caso em estudo neste trabalho é a ocorrência ou não da inandimplência de um certo cliente (DINIZ e LOUZADA, 2012). Louzada-Neto *et al.* (2001) afirmam que a utilização de censuras no modelo oferece um campo mais amplo de estudo, visto que, a desconsideração de censuras podem implicar em soluções viesadas, o que eventualmente pode acontecer em modelos estatísticos tradicionais.

Em algumas situações, existem certos grupos de indivíduos que podem ser considerados "curados", ou seja, não estão passíveis a ocorrência do evento de interesse (Fernandes, 2013). Diz-se, então, que esses indivíduos são "imunes" ao evento de interesse e o conjunto de dados de sobrevivência aos quais eles pertencem possui uma fração de cura. No contexto do mercado financeiro, o intuito é prever o tempo de recuperação de clientes. Recuperado é aquele cliente que retorna ao *status* de adimplência. Em casos em que a inadimplência não tenha acontecido ou para casos em que ocorreu a antecipação do prazo de empréstimo, a conclusão de bom ou mal pagador não é possível ser efetuada ao final do supervisionamento.

O uso de modelos de fração de cura, aplicados no mercado financeiro, é uma boa ferramenta para estimação de eventos, sejam eles, o prazo de retorno até o status de adimplência ou a realização/atraso de uma das parcelas do empréstimo (Toledo, 2011). Além disto, ainda dentro deste contexto, indíviduos os quais nunca efetuaram o ato da compra de determinado produto pode ser considerado como imune, tendo em vista que elas não estão suscetíveis a ocorrência do evento, assim como apresentado por Farewell (1986) e Meeker (1987).

Dentro da teoria de modelos de fração de cura, houveram muitos autores que contribuíram sendo Boag (1949) um dos vanguardista. Indo na mesma linha de pensamento, foi proposto, posteriormente, o modelo de mistura padrão (Berkson e Gage, 1952). Posteriormente, modelos mais sofisticados foram propostos por autores como Tsodikov *et al.* (1996), Ibrahim *et al.* (2014) dentre outros presentes na literatura. Também com uma expansão, com a teoria unificada por Rodrigues *et al.* (2009). Por fim, Granzotto *et al.* (2008) realizaram, a partir de dados financeiros, um estudo para inferir o tempo até a ruptura do cliente com a instituição, aplicando a metodologia proposta por Berkson e Gage (1952), modelando o tempo com modelos Weibull e log-logístico.

Em alguns estudos, indivíduos são suscetíveis a falhas precoces, o que resulta em um tempo de sobrevivência igual a zero ou muito próximo de zero. Aqui, iremos nos referir como inflacionado de zero. Na prática, é comum os pesquisadores excluírem essas unidades do estudo. No entanto, retirar essa informação pode levar a conclusões errôneas, como taxas de sobrevida superestimadas. Contudo, Martin *et al.* (2005) afirmam que a ocorrência de valores iguais a zero podem ser decorrentes de dois grupo, sendo "Contagem de Zeros Verdadeira"e "Contagem de Zeros Falsos", sendo a primeira resultante de zeros reais que decorrem da baixas probabilidades de ocorrência deste tipo evento, enquanto a segunda está associado ao erro humano, sendo ele, tanto decorrente de erro amostral, quanto a zeros falsos. A aplicação com zeros inflacionados na área de sobrevivência ainda é pouco explorada na literatura, mesmo não sendo tão incomum esta presença em dados com censura. Nas finanças um exemplo de aplicação deste tipo de modelagem póde ser vista em (de Oliveira *et al.*, 2017).

Respaldado pelo anseio de aplicar modelos de sobrevivência com a presença de excesso de zeros, ao cenário de risco de crédito, será utilizado neste trabalho métodos inferenciais frequentistas, considerando distribuições de probabilidades Gompertz e Weibull para os tempos de recuperação de crédito de clientes. Neste âmbito, necessita-se averiguar se um cliente em estado de inadimplência tem probabilidades maiores de se recuperar de imediato, contudo, também há a possibilidade de adquirir a informação se a adimplência ocorrerá dentro de um contexto de análise de sobrevivência, assim analisando a probabilidade de ser um cliente adimplente.

1.1 Objetivo

1.1.1 Objetivo Geral

O objetivo deste Trabalho de Conclusão de Curso é utilizar a metodologia de análise de sobrevivência que possibilita a inserção de tempos iguais a zero, em cenários em que observa-se risco de crédito. Será investigado a fração de indivíduos que recuperam o status de adimplência ou recuperados e de consumo de crédito.

1.1.2 Objetivos Específicos

1. Estudar a metodologia de análise de sobrevivência de longa duração e sua aplicabilidade;

2. Considerar o problema em três cenários "falhas no tempo zero (falhas iniciais)", "suscetíveis"ou "não suscetíveis"ao evento de interesse.

3. Considerar um estudo de simulação

4. Aplicar a metodologia existente para dados reais voltado ao mercado financeiro

O conjunto de dados a ser utilizado neste trabalho foi fornecido por uma empresa de crédito, contendo informações sobre os indivíduos (pessoas e empresas) como informações cadastrais, compromissos de crédito e hábitos de pagamento. Os indivíduos acompanhados neste estudo referem-se apenas as pessoas físicas que contraíram dívidas no segmento financeiro, varejo e telecomunicações no período da crise econômica, as quais foram acompanhadas em um tempo pré-determinado.

1.2 Organização do Trabalho

Este trabalho está disposto da seguinte maneira. No Capítulo 2, será apresentado uma revisão da literatura com conceitos básicos da análise de sobrevivência, até metodologias de modelos de longa duração. No Capítulo 3, será introduzido modelos de longa duração inflacionados no zero, assim como um estudo de simulação. No Capítulo 4, será apresentada uma aplicação a dados financeiros baseada na metodologia proposta anteriormente. E por fim, o Capítulo 5 apresentará conclusões dos resultados obtidos com a aplicação da metodologia exposta nos capítulos anteriores.

Capítulo 2

Revisão da Literatura

Neste capítulo, foi feita uma breve revisão dos conceitos básicos que sustentam a análise de sobrevivência, tais como o conceito de censura, funções básicas importantes e outras definições que por ventura possam se tornar importante para o desenvolvimento do trabalho. Informações e conceitos expressos nesta seção serão obtidas de Colosimo e Giolo (2006a) a menos que seja dito o contrário.

2.1 Conceitos básicos de Análise de Sobrevivência

O termo análise de sobrevivência é comumente utilizado para expressar a análise que se utiliza de tempos de origem até a ocorrência de algum evento específico (Collett, 2015), também conhecido como tempo de falha, podendo ser o tempo até algum problema em um equipamento eletrônico, bem como a cura ou reaparecimento de alguma doença. A inferência realizada nestes tipos de modelos pode ser feita de maneira frequentista ou bayesiana, as quais serão brevemente abordadas posteriormente.

Um elemento importante para a área de sobrevivência é a presença de observações incompletas ou parciais, denominadas censuras. Quando há a existência de observações como estas, uma opção viável, ao invés de eliminá-las, é utilizar técnicas de sobrevivência, já que a exclusão delas pode acarretar em resultados viesados.

2.1.1 Censura

Dada a motivação para o uso de censuras, agora o interesse é classificar os tipos existentes, lembrando-se que a introdução da análise de sobrevivência deu-se com o objetivo

de identificar o tempo de sobrevivência em pacientes durante um tratamento de determinada doença. Visto isto, serão definidos a seguir os tipos mais comuns de censura, encontrados na literatura.

Censura do Tipo I

Quando há um tempo pré-estabelecido para o término do estudo, pode ocorrer a presença de indivíduos que não vivenciam o evento de interesse até o fim do estudo, consequentemente, terão seus tempos censurados. Em um contexto financeiro, a ocorrência deste tipo de censura pode ser encontrada, por exemplo, quando uma agência bancária almeja verificar o tempo até um indivíduo se tornar inadimplente, pré-fixando um período de tempo. Com este tipo, nota-se que ao final alguns indivíduos não apresentaram o evento de interesse (se tornar inadimplentes), deste modo, obtendo a censura do tipo I.

Censura do Tipo II

No caso que o estudo é finalizado após a ocorrência de um número pré-estabelecido do evento de interesse, caracteriza-se a censura do tipo II. Desta forma, após este número pré-estabelecido de ocorrências do evento, aqueles indivíduos que deixaram de apresentar o evento de interesse terá seu tempo censurado. Este tipo de experimento tem uma vantagem, visto que, para certos indivíduos, a ocorrência do evento de interesse (ou seja, ocorrência da falha) pode demorar muito tempo, portanto, em muitos casos, reduzindo tempo e dinheiro investidos.

Censura Aleatória

Diferente das abordadas anteriormente, pode existir algumas censuras que fogem da alçada do pesquisador, ou seja, mesmo tomando muitas medidas, ocorrem de maneiras excepcionais. Frequentemente observada em experimentos, como o abandono do indíviduo em uma pesquisa, ou a morte súbita do paciente por algum motivo diferente do evento de interesse, a qual é a mais observada na área médica.

Além destas citadas acima, existem outras definições para tempos de censura, como censuras a direita, esquerda, entre outras (ver Colosimo e Giolo (2006a) e Lawless (2011)).

De forma a definir matematicamente a censura, seja i o índice associado ao $i-ésimo$ indivíduo suscetível a ocorrência do evento, $i = 1, ..., n.$. Pode-se definir a censura a partir

de uma função indicadora δ_i, com t_i sendo considerado o tempo de falha. Deste modo, tem-se que.

$$\delta_i = \begin{cases} 1 & , \text{ se } t_i \text{ for o tempo até a falha;} \\ 0 & , \text{ se } t_i \text{ for o tempo até a censura.} \end{cases}$$

2.1.2 Funções de Interesse

Dentro de análise de sobrevivência, usualmente, denota-se T como sendo uma variável aleatória não-negativa, $T \geq 0$, que representa o tempo de falha. Geralmente é apresentada na forma de sua função de sobrevivência ou função taxa de falha (ou risco), as quais são amplamente utilizadas dentro desta área.

Inicialmente, considerando que o tempo de falha é usualmente dado por uma variável aleatória contínua, define-se a função de densidade de probabilidade da ocorrência da falha, como sendo o limite desta probabilidade, no intervalo de tempo $[t, t + \Delta t)$ por unidade, assim como expressa abaixo.

$$f_T(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t}. \tag{2.1}$$

Por consequência da Equação (2.1), a sua função de distribuição acumulada, é dada por:

$$F_T(t) = \mathbb{P}(T \leq t) = \int_0^t f_T(s)ds. \tag{2.2}$$

Visto isto, a Equação (2.2) nos retorna qual é a probabilidade de que um indivíduo esteja vivo antes de um tempo t, porém, em análise de sobrevivência o objetivo é verificar qual é a probabilidade de um indivíduo sobreviver após um tempo t, o que pode ser visto como o cálculo complementar desta probabilidade.

$$S_T(t) = \mathbb{P}(T \geq t) = \int_t^\infty f_T(s)ds = 1 - F_T(t), \tag{2.3}$$

tendo as seguintes propriedades:

1. $S_T(t)$ é não decrescente;
2. $S_T(0) = 1$;
3. $\lim_{t \longrightarrow \infty} S_T(t) = 0$.

Sabendo que o indivíduo sobreviveu após um tempo t, pode-se ter interesse na taxa de falha instântanea, isto é, obter o risco da ocorrência de falha do indivíduo no intervalo $[t, t + \Delta t)$, com $\Delta t \to 0$, o qual é dado por:

$$h_T(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f_T(t)}{S_T(t)}. \tag{2.4}$$

A função de risco pode apresentar comportamentos variados dependendo a distribuição de T, como formas crescentes, decrescentes, constantes, ou até mesmo a conhecida como "curva de banheira", a qual é uma característica das curvas do risco de morte dos seres humanos.

Por fim, outra função conveniente é a função de risco acumulado, que assim como o nome diz, calcula o acúmulo dos riscos do indivíduo, assim como definido na Equação (2.4).

$$H_T(t) = \int_0^t h_T(s) ds. \tag{2.5}$$

Em decorrência das equações vistas anteriormente, é possível obter algumas relações entre elas que podem ser interessantes dependendo o contexto que está sendo trabalhado na prática.

$$h_T(t) = -\frac{d}{dt}\log(S_T(t)), \quad H_T(t) = -\log(S_T(t)), \quad S_T(t) = \exp(-H_T(t)).$$

O interessante destas relações é que, por exemplo, conhecendo-se $H_T(t)$, as outras funções podem ser obtidas por consequência.

2.1.3 Estimador de Kaplan-Meier

A literatura apresenta algumas classes de estimadores para a função de sobrevivência com a presença de censura. Entre elas existe a classe de estimadores não paramétricos, os quais não associam que a variável T siga uma distribuição em específico. Por exemplo, existe o estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), assim como o estimador de Nelson-Aalen proposto por Nelson (1972) e, posteriormente, tendo suas propriedades estudadas por Aalen (1978). Há um terceiro estimador, chamado tabela de vida, sendo uma das técnicas mais antigas para a estimação do tempo de falha.

Também conhecido como estimador produto limite, o estimador de Kaplan-Meier,

pode ser visto como uma adaptação do estimador empirico da função sobrevivência, quando há a existência de censura. Este estimador, na sua construção, leva em consideração inúmeros intervalos, assim como os distintos tempos de falhas observados.

Deste modo, o estimador de Kaplan-Meier é dado pela seguinte expressão:

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

em que $t_1, t_2, ..., t_r$ são considerados os r tempos de falhas ordenados em ordem crescente, d_j é dito o número de falhas no tempo t_j, $j = 1, ..., r$, e n_j visto como o número de indivíduos expostos ao risco em t_j, $j = 1, ..., r$. Um adendo para este estimador é que a curva resultante possui um comportamento em forma de escada, em que cada degrau corresponde a intervalos de tempo resultante entre duas falhas distintas.

Em certas situações, tem-se o interesse em estimar curvas de sobrevivências separando-as por algum grupo de interesse. Para isto, a utilização do estimador de Kaplan-Meier estratificado é uma opção. Porém, é razoavel considerar esta divisão entre grupos distintos quando é notada uma diferença significativa entre as curvas de sobrevivência. Com isto, nasce a necessidade de avaliar esta diferença de forma inferencial, o que pode ser feito utilizando o teste não paramétrico Log-rank.

Proposto por Mantel e Haenszel (1959), com o intuito de verificar diferenças entre dois grupos distintos, este teste foi estendido por Aalen (1978) e Gill (1980) com o objetivo de verificar a presença de diferenças entre curvas de sobrevivência. A estatística teste é fundamentada nas diferenças entre o número observado e esperado das falhas dentro de cada grupo. Assim, definindo matematicamente sua hipótese nula, tem-se $H_0 : S_1(t) = S_2(t)$, com a respectiva estatística do teste.

$$T = \frac{\left[\sum_{j=1}^{r} (d_{2j} - w_{2j}) \right]^2}{\sum_{j=1}^{r} (V_j)_2}, \tag{2.7}$$

Sob a hipótese nula, T possui distribuição assintótica qui-quadrado com 1 grau de liberdade.

Sendo representado na Equação (2.7), d_{2j} é visto como o número de falhas observadas no grupo 2 no tempo t_j, w_{2j} é visto como o número esperado de falhas no grupo 2 no tempo t_j, sendo $w_{2j} = n_{2j} d j n_j^{-1}$, com d_j e n_j considerados a quantidade total de falhas observadas e número total de indivíduos sob o risco no tempo t_j, respectivamente, já n_{2j} denotado como o número de indivíduos sob risco do grupo 2 no instante de tempo t_j. Já

$(V_j)_2$ é vista como:

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

o que pode ser descrito como a variância de d_{2j}.

2.1.4 Estimação por Máxima Verossimilhança

Em muitos casos estamos interessados em estimar a função de sobrevivência, com a suposição de que o tempo de falha, definido como a variável aleatória T, segue uma distribuição específica. Nestes casos, ao supor que os dados seguem alguma distribuição, existem os parâmetros associados a ela, os quais são desconhecidos. Portanto, o intuito desta técnica é obter, entre todos os valores possíveis de seu espaço paramétrico, aquele que retorna maior possibilidade de ter gerado a amostra (Colosimo e Giolo, 2006a).

Em um contexto em que não há a presença de censuras, conforme visto em Bolfarine e Sandoval (2001). Dado $(T_1, T_2, ..., T_n)$ uma amostra aleatória de tamanho n, proveniente da variável aleatória T, com função densidade dada por $f_{T_i}(t_i|\theta)$, sendo θ o vetor de parâmetros do modelo, a função de verossimilhança é:

$$L(\theta; \mathbf{t}) = \prod_{j=1}^{n} f_{T_i}(t_i|\theta).$$

Assim, o estimador de máxima verossimilhança para o vetor de parâmetros $\boldsymbol{\theta}$ é aquele que maximiza a função de verossimilhança $L(\theta; \mathbf{t})$.

Porém, na análise de sobrevivência trabalha-se com dados aonde existem a presença de censura. Visto que a sua utilização pode trazer informação adicional ao seu modelo, já que devido ao fato dele ter sido censurado, sabe-se que o tempo de falha daquele indivíduo é maior do que o tempo censurado. Com isto, a sua incorporação na função de verossimilhança é dado com o uso da função sobrevivência.

$$L(\theta; \mathbf{t}) = \prod_{j=1}^{n} \left[f_{T_i}(t_i|\theta)\right]^{\delta_i} \left[S_{T_i}(t_i|\theta)\right]^{1-\delta_i} = \prod_{j=1}^{n} \left[h_{T_i}(t_i|\theta)\right]^{\delta_i} S_{T_i}(t_i|\theta), \tag{2.8}$$

em que, δ_i é a função indicadora que representa a presença de censura. Um adendo é que a Expressão (2.8) é válida, apenas quando as censuras encontradas forem independentes e do tipo I, II, aleatória ou sob a suposição de mecanismo não informativo.

Porém, é apropriado e conveniente utilizar-se o logarítmo da função de verossimilhança,

$log(L(\theta;\mathbf{t}))$, para obter as estimativas do seu vetor de parâmetros, visto que os valores do vetor θ que maximizam a sua função de verossimilhança coincidem com os valores que maximizam sua função de log-verossimilhança, os quais são encontrados resolvendo o seguinte sistema de equações.

$$U(\theta) = \frac{\partial \log(L(\theta;\mathbf{t}))}{\partial \theta} = \frac{\partial l(\theta;\mathbf{t})}{\partial \theta}.$$

Em sua maioria, estas equações não resultam em formas algébricas fechadas, por consequência, não conseguindo-se obter resultados analíticos dada a complexidade das equações. Com isto, faz-se necessário a utilização e implementação de algorítmos númericos e interativos para a sua estimação.

2.2 Modelos Probabilísticos

Nesta seção, são apresentados alguns modelos probabilísticos que são bastante utilizados na área de análise de sobrevivência para modelar o tempo de falha. Mesmo com a grande gama de distribuições que são utilizadas para este intuito, pode-se destacar algumas, visto suas posições de destaque em número de publicações sobre o assunto. Com isto, serão introduzidos os modelos Exponencial, Weibull e Gompertz.

2.2.1 Distribuição Exponencial

A distribuição exponencial pode ser vista como a mais simplória, matematicamente dizendo, sendo ela amplamente usada em problemas em que o tempo corresponde ao tempo de falha de determinado produto, tempo de vida de óleos isolantes, entre outros (Colosimo e Giolo, 2006a).

Se T é uma variável aleatória com distribuição exponencial, com parâmetro $\lambda > 0$, sua função densidade de probabilidade é dada por.

$$f_T(t) = \lambda \mathbf{exp}\{-\lambda t\}\mathbb{I}_{\{t \geq 0\}} \tag{2.9}$$

Consequentemente, a função sobrevivência e taxa de falha são obtidas por meio da Expressão (2.9).

$$S_T(t) = \mathbb{P}(T \geq t) = \mathbf{exp}\{-\lambda t\}, \qquad h_T(t) = \frac{f_T(t)}{S_T(t)} = \lambda.$$

Por consequência da propriedade de falta de memória desta distribuição, nota-se que a função taxa de falha é constante. Isto significa que, independentemente do tempo em que o objeto ou indivíduo está no estudo, o risco de falhar em um intervalo futuro é o mesmo.

O interessante em se utilizar uma distribuição de probabilidade associada aos dados é que a forma das suas funções estão suscetíveis a alteração do valor de seus parâmetros, assim como mostra a Figura 2.1.

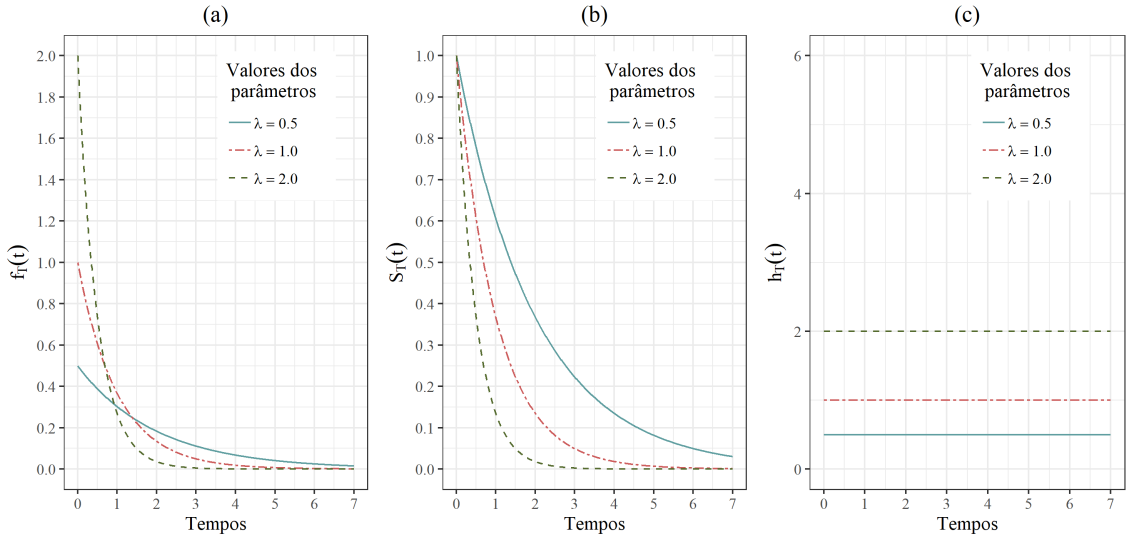

Figura 2.1: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Exponencial.

2.2.2 Distribuição Weibull

Inicialmente proposta por Weibull (1939), a distribuição weibull teve sua aplicabilidade também abordada, posteriormente, pelo mesmo autor (Weibull, 1951). Desde então, pode ser vista sua utilização em variadas áreas como modelos biométricos, áreas industriais, laborátorio e análise de confiabilidade (Martz e Waller, 1982). Sua grande fama e utilização pode ser vista por uma característica muito interessante, que são as várias formas de suas funções se sobrevivência, taxa de falha e densidade, especialmente a função taxa de falha apresentando um comportamento monótono, podenso ser crescente, decrescente ou constante.

Seja T uma variável aleatória com distribuição weibull, com parâmetros dados por $\lambda > 0$ e $\gamma > 0$, sendo γ um parâmetro de forma e λ um parâmetro de escala. Assim, sua função densidade de probabilidade é dada por.

$$f_T(t) = \gamma\lambda^\gamma t^{\gamma-1}\mathbf{exp}\{-(\lambda t)^\gamma\}\mathbb{I}_{\{t \geq 0\}}. \tag{2.10}$$

A sua função de sobrevivência é obtida por meio de sua função densidade de probabilidade, que é:

$$S_T(t) = \int_t^\infty \gamma\lambda^\gamma t^{\gamma-1}\mathbf{exp}\{-(\lambda t)^\gamma\}dt = \mathbf{exp}\{-(\lambda t)^\gamma\},$$

Utilizando as equações anteriores se obtêm a função taxa de falha da weibull,

$$h_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\gamma\lambda^\gamma t^{\gamma-1}\mathbf{exp}\{-(\lambda t)^\gamma\}}{\mathbf{exp}\{-(\lambda t)^\gamma\}} = \gamma\lambda^\gamma t^{\gamma-1}.$$

Um peculiariedade desta distribuição é que, para específicos valores do parâmetro de forma γ, ontêm-se distribuições diferentes já relatadas na literatura. Por exemplo, quando $\gamma = 1$, T segue uma distribuição Exponencial; $\gamma = 2$, uma distribuição Rayleigh, além de duas situações para quando $\gamma = 2.5$ e $\gamma = 3.6$, sendo aproximações das distribuições Log-Normal e Normal, respectivamente.

Assim, como dito anteriormente, o interessante em se utilizar uma forma paramétrica a seus dados é que a forma das suas funções está suscetível à alteração do valor de seus parâmetros, assim como mostra a Figura 2.2.

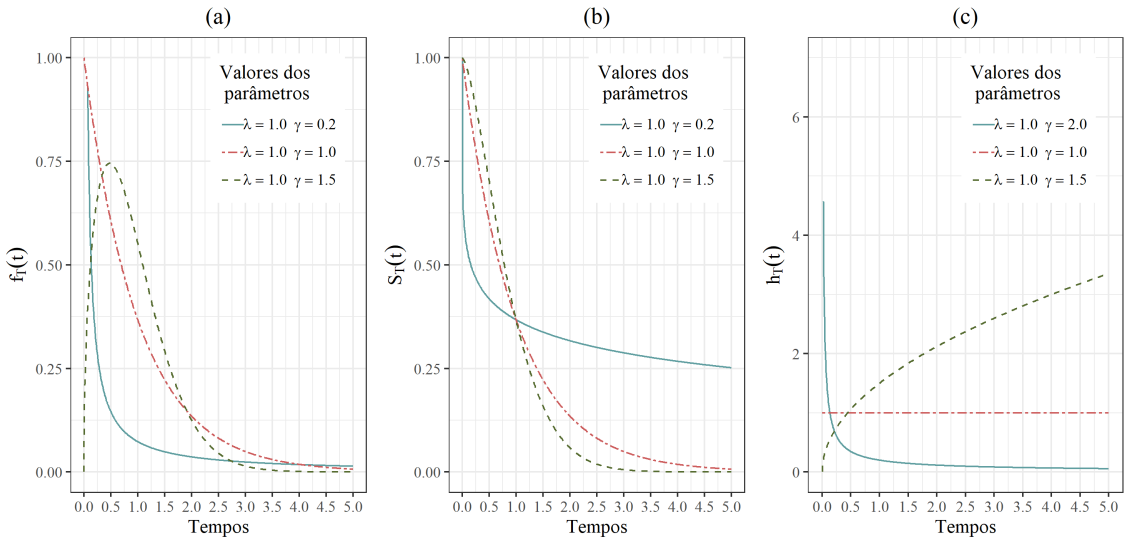

Figura 2.2: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Weibull.

2.2.3 Distribuição Gompertz

Desenvolvida por Benjamin Gompertz, visando sua aplicabilidade em estudos demográficos (Gompertz, 1825), esta distribuição é amplamente utilizada em áreas de atuária, demografia e outros estudos aonde é possivel a aplicação de análise de sobrevivência (Gieser *et al.*, 1998). Esta distribuição, em sua formulação, contêm dois parâmetros, λ e γ, sendo γ um parâmetro de escala e λ um parâmetro de forma.

Seja T uma variável aleatória com distribuição gompertz, com parâmetros dados por $\lambda > 0$ e $\gamma > 0$. Sua função densidade de probabilidade é dada por:

$$f_T(t) = \lambda e^{\gamma t}\mathbf{exp}\left\{ -\left(\frac{\lambda}{\gamma}\right)(\mathbf{exp}(\gamma t) - 1)\right\}\mathbb{I}_{\{t \geq 0\}}. \tag{2.11}$$

Consequentemente, obtêm-se a função sobrevivência e taxa de falha por meio da Equação (2.11).

$$S_T(t) = \mathbf{exp}\left\{ -\left(\frac{\lambda}{\gamma}\right)(\mathbf{exp}(\gamma t) - 1)\right\}, \qquad h_T(t) = \lambda\mathbf{exp}(\gamma t).$$

Assim, como mencionado anteriormente, o interessante em se utilizar um modelo paramétrico a seus dados é que a forma das suas funções estão suscetíveis a alteração do valor de seus parâmetros, assim como mostrado pela Figura 2.3.

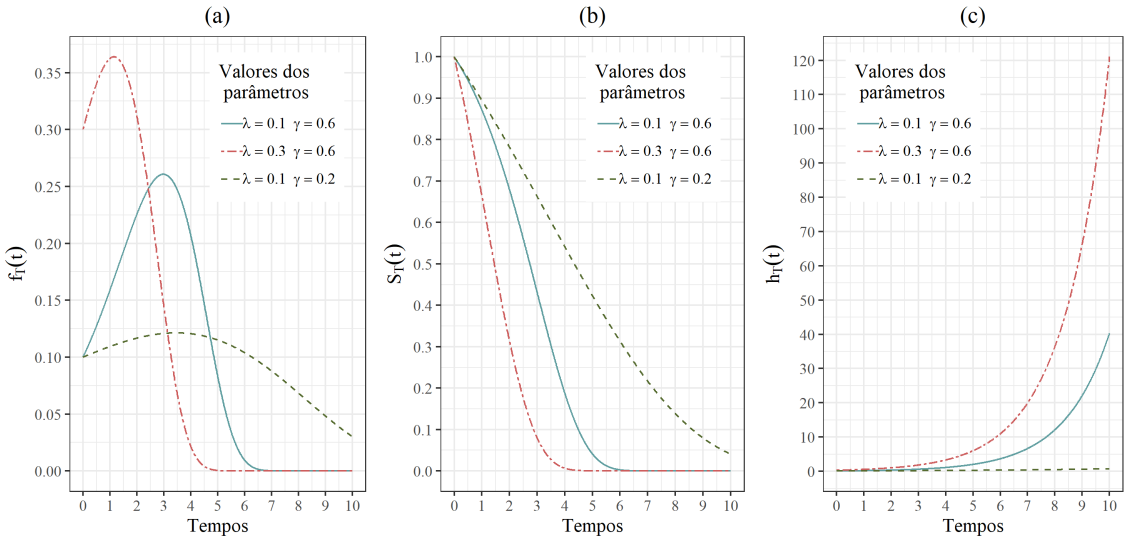

Figura 2.3: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Gompertz.

2.3 Modelos de Longa duração

Dentro da análise de sobrevivência, há certas situações que nem todos os indivíduos inclusos no estudo apresentam o evento de interesse mesmo após longos períodos de tempos, contrariando, assim, a atribuição, usualmente utilizada, de que todos estão suscetíveis a apresentar o evento. Comumente a proporção de indivíduos não suscetíveis ao evento de interesse é intitulada: curadas/imunes, fidelizados (Ibrahim *et al.*, 2014), (Granzotto *et al.*, 2008). Com isto nasce a motivação e necessidade para se utilizar modelos de sobrevivência de longa duração.

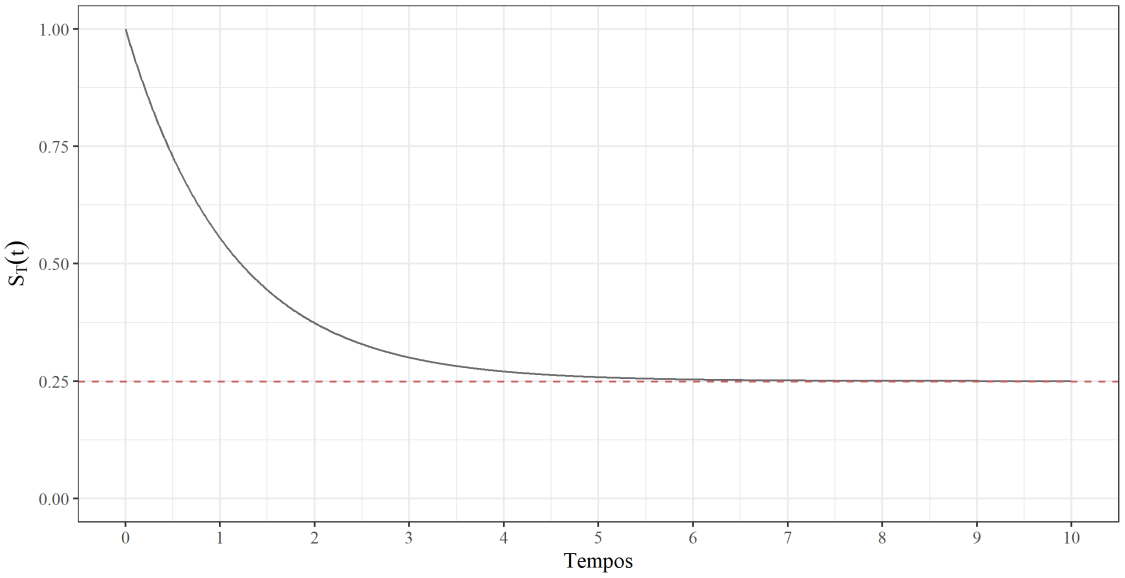

Figura 2.4: Função de sobrevivência associada aos modelos de longa duração.

A Figura 2.4 mostra exatamente o comportamento usual que motiva a utilização de modelos de longa duração, visto que, ao invés da curva de sobrevivência estabilizar no ponto $S_T(t) = 0$ à medida que o tempo aumenta, ela acaba estabilizando no ponto $S_T(t) = 0.25$. Deste modo então, este ponto limitante que a função de sobrevivência estabiliza é exatamente a proporção de indivíduos que são considerados curados/imunes.

A modelagem de longa duração, dentro da análise de sobrevivência, tem grande importância, e com isto, surge na literatura vários artigos com diferentes métodos para se ajustar a situações distintas. Inclusive, é possível acomodar tempos inflacionados de zeros em modelos de longa duração, porém, em algumas modelagens mais comuns, como o modelo de mistura padrão, introduzido por Berkson e Gage (1952), não é possível.

2.3.1 Modelo de Mistura Padrão

Proposto por Berkson e Gage (1952), o modelo de mistura padrão é um dos modelos mais simples e comuns dentro da classe de modelos de longa duração. Esta metodologia é baseada em escrever a função de sobrevivência em duas partes, uma delas considerando os indivíduos curados (ou adimplentes para o contexto desse trabalho), e a outra se tratando da parte da população de indivíduos considerados não curados. Uma peculiariedade é que a função de sobrevivência associada à população total é dita imprópria, o que poderá ser entendido melhor posteriormente.

Assim, é possível deduzir o modelo de mistura padrão considerando uma variável $M_i \sim \text{Bernoulli}(\theta)$, para os indivíduos que estão ou não sob risco. Logo,

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

sendo $\mathbb{P}(M_i = 0) = 1 - \theta$ e $\mathbb{P}(M_i = 1) = \theta$.

Deste modo, pode-se verificar que há a existência de duas subpopulações, porém nota-se que apenas a função de sobrevivência relacionada à proporção de não curados é dita própria, já que, para o caso dos indivíduos curados, os tempos de falha são infinitos, tendo em vista que o evento de interesse não irá acontecer, resultando assim em uma função de sobrevivência imprópria para esta subpopulação. Considerando T uma variável aleatória não negativa e contínua, representando o tempo de falha, obtêm-se

$$\mathbb{P}(T > t \ | M_i = 1) = S_T(t) \text{ e } \mathbb{P}(T > t \ | M_i = 0) = 1.$$

Com o resultado anterior, é possível então decompor a função de sobrevivência da população geral através do teorema da probabilidade total, particionando-a entre as duas subpopulações. Assim:

$$\begin{aligned} S_{pop}(t) &= \mathbb{P}(T > t) = \mathbb{P}(T > t \ | M_i = 0)\mathbb{P}(M_i = 0) + \mathbb{P}(T > t \ | M_i = 1)\mathbb{P}(M_i = 1) \\ &= (1 - \theta) + \theta S_T(t), \quad t \geq 0, \end{aligned}$$

em que $S_T(\cdot)$ é a função de sobrevivência própria associada à subpopulação de indivíduos considerados não curados. Como dito anteriormente, a função de sobrevivência associada à população total, $S_{pop}(\cdot)$, é imprópria possuindo as seguintes propriedades:

- Se $\theta = 1$, então $S_{pop}(t) = S_T(t)$;

- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é uma função decrescente;
- $\lim_{t\to\infty} S_{pop}(t) = 1 - \theta$.

Em especial, a última propriedade caracteriza o fato da função de sobrevivência populacional ser imprópria, uma vez que o limite da sobrevivência não vai a zero à medida que o tempo tende ao infinito, sendo o valor $1 - \theta$ justamente a proporção de indivíduos curados.

Assim como visto na Subseção 2.1.2, obtendo uma das funções de interesse, é possível obter as demais. Desta forma, a função densidade populacional é dada por:

$$f_{pop}(t) = -\frac{d\big[S_{pop}(t)\big]}{dt} = \theta f_T(t),$$

sendo $f_T(\cdot)$ a função densidade própria relacionada a subpopulação de indivíduos considerados não curados, ou seja, aqueles que estão sob risco. Com isto, a função de risco populacional é

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{\theta f_T(t)}{1 - \theta + \theta S_T(t)}.$$

E a função de risco própria, é:

$$h_T(t) = \frac{S_{pop}(t) h_{pop}(t)}{(1-\theta) S_T(t)} = \left[\frac{S_{pop}(t)}{S_{pop}(t) - (1-\theta)}\right] h_{pop}(t)$$

Não é muito difícil perceber que $\{S_{pop}(t)/[S_{pop}(t) - (1-\theta)]\} > 1$, logo têm-se que $h_{pop}(t) < h_T(t)$, ou seja, a função taxa de risco populacional é limitada pela função de risco relacionada a população de não curados. Verifica-se também que $h_T(t)$ não irá ter a propriedade de risco proporcional, tendo em vista que $\{S_{pop}(t)/[S_{pop}(t) - (1-\theta)]\}$ sempre estará dependendo de t. Além disso,

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

Nota-se que, quanto mais o tempo aumenta, o risco converge a zero. Com isto é possivel verificar mais um fato que ocorre devido à estabilização da curva de sobrevivência populacional em um valor diferente de zero, ou seja, a sua fração de cura, o que é um

indicativo de que uma parcela da sua população não obteve seu evento de interesse e, possivelmente, “curados”.

2.3.2 Modelos Unificados de Fração de Cura

Outro modelo de longa duração visto na literatura é o modelo unificado de fração de cura, estudado por Chen *et al.* (1999) e Rodrigues *et al.* (2009). De forma geral, a ideia básica do modelo unificado de fração de cura está baseada na noção de ocorrência do evento de interesse em um processo em dois estágios:

• Estágio de iniciação: Seja uma variável aleatória, N, a qual representa o número de riscos ou causas que competem a ocorrência de determinado evento de interesse. Há o desconhecimento da ocorrência do evento, com N sendo não observada, seguindo uma distribuição de probabilidade p_n representando a distribuição da variável e com as caudas dadas por:

$$p_n = \mathbb{P}(N = n) \quad \text{e} \quad q_n = \mathbb{P}(N > n),$$

com $n = 0, 1, 2, ...$

• Estágio de maturação. Estabelecido um valor para a variável de causas competitivas, $[N = n]$, sejam Z_k, com $k = 1, 2, ..., n$, variáveis aleatórias não negativas representando o tempo até a ocorrência do evento de interesse atrelado à k-ésima causa independentes entre si, com função acumulada dada por $F_Z(z) = 1 - S_Z(z)$ e independentes de N. Com o objetivo de inserção de indivíduos não suscetíveis ao evento, define-se o tempo até a ocorrência como:

$$T = \min\{Z_0, Z_1, Z_2, ..., Z_N\},$$

em que $P[Z_0 = \infty] = 1$, tendo então a possibilidade de uma parcela da população p_0 não apresentar o evento de interesse, sendo T uma variável aleatória observável a qual em muitos casos poderá ser censura e Z_j e N são variáveis não observáveis, ou seja, variáveis latentes.

De acordo com Feller (2008). Seja uma sequência de números reais $\{a_n\}$. Se

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

converge para valores de s contidas no intervalo [0, 1], por consequência $A(s)$ pode ser definida como uma função geradora de sequência $\{a_n\}$.

Se $a = a_n = p_n = p$, discorre que $A_p(s) = E[s^N]$. É possível então observar que a função geradora de probabilidades pode ser descrita com a função geradora de momentos da variável aleatória latente N no ponto $\log(s)$, ou seja, $A_p(s) = E[\exp\{\log(s)N\}]$, a qual converge. Então $A(s)$ é definida como a função geradora da sequência $\{a_n\}$.

Logo é possível definir a função de sobrevivência populacional com distribuição T da seguinte forma:

$$\begin{aligned} S_{pop}(t) &= \mathbb{P}(N=0) + \mathbb{P}(Z_1 > t, Z_2 > t, ..., Z_N > t, N \geq 1), \\ &= \mathbb{P}(N=0) + \sum_{n=1}^{\infty} \mathbb{P}(N=n)\mathbb{P}(Z_1 > t, Z_2 > t, ..., Z_N > t), \\ &= p_0 + \sum_{n=1}^{\infty} p_n S_T(t)^n, \\ &= A(S_T(t)), \end{aligned}$$

em que $A(\cdot)$ é a função geradora da sequência $\{p_n\}$. Ou seja, pode-se definir a função de sobrevivência populacional relacionada à variável aleatória T, como um modelo de longa duração em dois estágios, sendo então um agrupamento entre a função geradora de probabilidades e a função de sobrevivência. Porém, nota-se que a função de sobrevivência do modelo de longa duração em dois estágio, $S_{pop}(t)$, é definida como não própria.

Sendo assim, a função de sobrevivência populacional segue as seguintes características:

- Se $p_0 = 1$, então $S_{pop}(t) = S_T(t)$;
- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é uma função decrescente;
- $\lim_{t\to\infty} S_{pop}(t) = p_0$.

Assim como visto na Subseção 2.1.2, obtendo uma das funções de interesse, é possível obter as demais. Desta forma, a função densidade populacional e de risco populacionais são dadas respectivamente por:

$$f_{pop}(t) = f_T(t)\frac{d\big[A(s)\big]}{ds}\Big|_{s=S_T(t)}$$

e

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = f_T(t)\frac{\frac{d[A(s)]}{ds}\Big|_{s=S_T(t)}}{S_{pop}(t)}$$

Em especial, pode-se dar destaque a algumas distribuições que são utilizadas, comumente, para o número de causas competitivas, variável aleatória N, sendo estas dadas por Bernoulli, Poisson, Binomial Negativa, Binomial e Geométrica, abrangendo assim uma maior variedade para diferentes dispersões. Podendo então, obter mais expressões relacionadas a elas com mais detalhes a seguir.

Modelo Bernoulli

Seja o número de causas competitivas latentes para o desfecho do evento de interesse, N, dado por uma distribuição Bernoulli, com parâmetro θ. Então sua função de probabilidade como:

$$P[N = n] = \theta^n(1-\theta)^{1-n},\ n = 0, 1.\text{ e } 0 < \theta < 1. \tag{2.12}$$

Logo, sua função geradora de probabilidades para N é dada por:

$$A(s) = 1 - \theta + \theta s,\ 0 \leq s \leq 1. \tag{2.13}$$

Consequentemente, a função de sobrevivência populacional é descrita por:

$$S_{pop}(t) = A(S(t)) = 1 - \theta + \theta S(t), \tag{2.14}$$

tendo assim, o modelo de mistura padrão proposto por (Berkson e Gage, 1952), descrito na seção anterior. O modelo de mistura padrão é denominado assim, pois ele é visto como uma mistura de distribuições paramétricas, tendo uma, definindo a parcela da população considerada suscetíveis ao evento de interesse e outra para os curados/imunes. Podendo então obter a proporção de curados por:

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = 1 - \theta. \tag{2.15}$$

As funções densidade e risco são, respectivamente:

$$f_{pop}(t) = \theta f(t) \tag{2.16}$$

e

$$h_{pop}(t) = \frac{\theta f(t)}{1 - \theta + \theta S(t)}. \tag{2.17}$$

Modelo Poisson

Seja o número de causas competitivas latentes para o desfecho do evento de interesse, N, dado por uma distribuição Poisson, com parâmetro θ. A função geradora de probabilidades da Poisson é dada por $A(s) = \exp[\theta(1 - s)]$. Então sua função de sobrevivência é:

$$S_{pop}(t) = A(S(t)) = \exp\left[-\theta F(t)\right]. \tag{2.18}$$

Consequentemente, as funções densidade e risco são, respectivamente:

$$f_{pop}(t) = -\frac{dS_{pop}(t)}{dt} = \theta f(t) \exp\left[-\theta F(t)\right] \tag{2.19}$$

e

$$h_{pop}(t) = \theta f(t). \tag{2.20}$$

Dessa forma, de (2.18) obtem a fração de cura dada por $p_0 = \lim_{t\to\infty} S_{pop}(t) = \exp(-\theta)$.

É possível verificar, por meio da Tabela 2.1, as funções de sobrevivência populacional, densidades impróprias e as frações de curas para diferentes distribuições associadas ao número de causas latentes N.

Tabela 2.1: Função de sobrevivência $S_{pop}(t)$, função de densidade $f_{pop}(t)$ e fração de cura para diferentes distribuições do número de causas latentes, N. Sendo θ^* outra parametrização, com $\theta^* = \theta/(1+\theta)$.

Distribuição	$S_{pop}(t)$	$f_{pop}(t)$	p_0
Bernoulli(θ)	$1-\theta+\theta S(t)$	$\theta f(t)$	$1-\theta$
Binomial(K, θ^*)	$(1-\theta^*+\theta^* S(t))^K$	$K\theta^* f(t)(1-\theta^*+\theta^* S(t))^{K-1}$	$(1-\theta^*)^K$
Poisson(θ)	$\exp(-\theta F(t))$	$\theta f(t)\exp(-\theta F(t))$	$e^{-\theta}$
Geométrica(θ)	$\{1+\theta F(t)\}^{-1}$	$\theta f(t)\{1+\theta F(t)\}^{-2}$	$1/(1+\theta)$
Binomial Negativa(η, θ)	$\{1+\eta\theta F(t)\}^{-1/\eta}$	$\theta f(t)\{1+\eta\theta F(t)\}^{-1-1/\eta}$	$(1+\eta\theta)^{-1/\eta}$

2.4 Considerações finais

Este capítulo buscou apresentar conceitos básicos da análise de sobrevivência, como a função de sobrevivência, definição de censura, além de distribuições que são comumente utilizadas para as modelagens. Objetivando a metodologia que será proposta para este trabalho, também foi abordado conceitos de modelos de longa duração, o qual em sua definição fazem a incorporação de indivíduos que não são suscetíveis ao eventos de interesse, sendo eles considerados curados.

Capítulo 3

Modelos de Fração de cura inflacionados de zero

Modelos inflacionados de zeros têm se tornado cada vez mais utilizado, dado a necessidade da incorporação de uma componente responsável por expressar o excesso de zeros no modelo. Sendo usado nas mais diversas áreas de estudo, por exemplo, o número de defeitos por lote de produção; na área médica quando há interesse de verificar a quantidade de casos de determinada doença; ou na espacial com a contagem de colisões de meteoritos em um teste de satélite durante sua órbita (Shanker e Hagos, 2016).

Nesta classe de modelo é considerado a mistura de duas distribuições, sendo uma responsável por tratar o excesso de zeros, e a segunda, responsável pelo restante, ou seja, valores não inflacionados de zeros, incluindo o zero (Martin *et al.*, 2005). Em análise de sobrevivência, na literatura existe pouca informação sobre a utilização de modelos que incorporam a inflação de zeros, observada em alguns tipos de experimentos. Diante desta necessidade, recentes trabalhos propuseram uma extensão do modelo proposto por Berkson e Gage (1952), que diferentemente da modelagem inicial, permite incluir os tempos iguais a zero, sendo inicialmente proposta em um contexto de dados financeiros, por de Oliveira *et al.* (2017), tendo sua função de sobrevivência para qualquer tempo, dada por:

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

sendo $S_0^*(t)$ é a função de sobrevivência associada à proporção de indíviduos suscetíveis à falha, $(1-p_0-p_1)$, com p_0 representando a proporção de tempos inflacionados de zero, e p_1

a proporção de indivíduos curados/imunes, ou seja, a fração de cura. Consequentemente, verifica-se as seguintes propriedades:

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

Implicando que, para casos em que a inflação de zero não existe, $p_0 = 0$. Este modelo se resume ao modelo de mistura padrão proposto por Berkson e Gage (1952). Observa-se na Figura 3.1 exatamente o comportamento da função de sobrevivência expressa por este modelo.

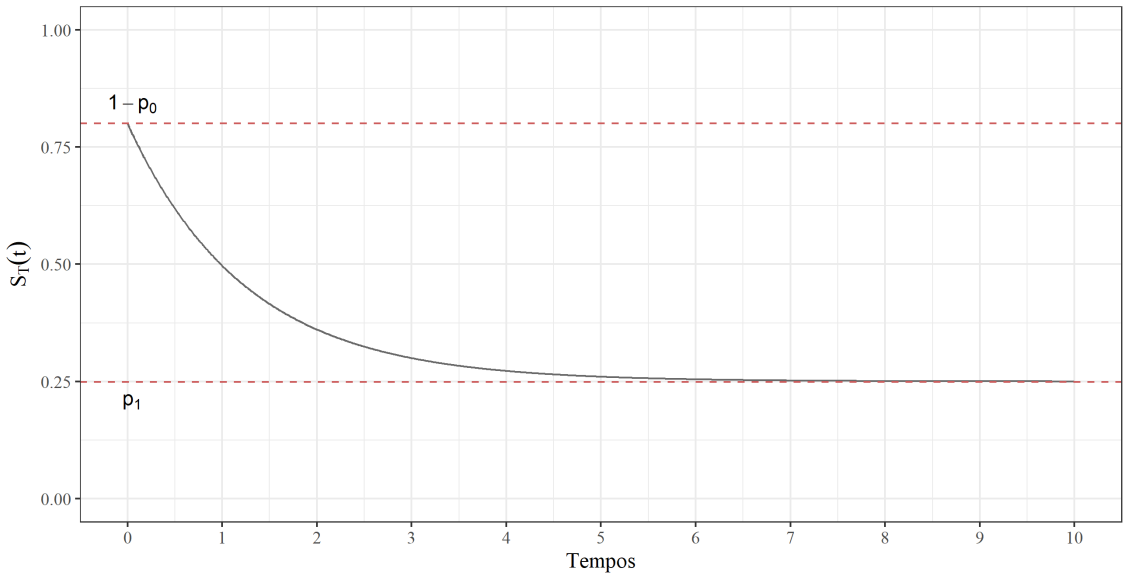

Figura 3.1: Função de sobrevivência associada ao modelo de fração de cura inflacionado de zeros.

A Figura 3.1 expressa os dois limitantes para a curva de sobrevivência deste modelo, ou seja, há o decaimento da curva iniciando em um ponto menor que um, $1 - p_0$, representando a inflação de zeros; e uma estabilização da curva em um ponto maior que zero, p_1, representando a fração de cura. Modelos que atendem simultaneamente a estes dois critérios são, comumente, denominados de modelos de sobrevivência zero-inflacionados (ou zero-ajustados) com fração de cura.

Assim, supondo que o número N de causas competitivas segue uma distribuição de probabilidade Poisson com parâmetro θ, estende-se a formulação e obtenção da curva de sobrevivência para os indivíduos suscetíveis a falha, $S_0^*(t)$, utilizando como base a

expressão para o modelo Poisson visto na Tabela (2.1). Deste modo, a função de sobrevivência é obtida através de uma modificação da função proposta por Ibrahim *et al.* (2014).

$$S_0^*(t) = \mathbb{P}(T > t|N \geq 1) = \frac{\exp\{-\theta F_T(t)\} - \exp\{-\theta\}}{1 - \exp\{-\theta\}}.$$

Verifica-se que $S_0^*(t)$ é própria já que as seguinte propriedades são satisfeitas $S_0^*(0) = 1$ e $S_0^*(\infty) = 0$. A sua função densidade é dada por:

$$f_0^*(t) = -\frac{d}{dt}\left[S_0^*(t)\right] = \left(\frac{\exp\{-\theta F_T(t)\}}{1 - \exp\{-\theta\}}\right)\theta f_T(t), \quad t \geq 0.$$

Também não é muito dificil perceber que $S_0^*(t)$ pode ser derivada por meio da função de sobrevivência sugerida na Tabela 2.1, quando se supõe que as causas, N, segue uma distribuição de poisson com parâmetro θ, visto que para o caso da tabela, $S_{pop}(t) = \mathbb{P}(T > t|N \geq 0)$. Deste modo, então, sendo possível verificar que $p_1 = \exp\{-\theta\}$, com p_1 denotando a fração de cura do modelo. Porém, é visto em de Oliveira *et al.* (2017) que para acomodar o excesso de zeros é sugerido, $p_0 = \exp\{-k\}$, sendo $k > 0$, ou seja o modelo proposto é dado por:

$$S_{pop}(t) = \exp\{-\theta\} + (1 - \exp\{-k\} - \exp\{-\theta\})S_0^*(t), \quad t \geq 0, \;\; \theta, k > 0. \tag{3.1}$$

Entretando, para garantir de que p_0, p_1 e $(1 - p_0 - p_1) \in (0, 1)$, conforme visto em Pereira *et al.* (2013), Hosmer Jr *et al.* (2013) e de Oliveira *et al.* (2017), uma proposta de vinculo entre os dois vetores de covariáveis e os parâmetros associados à inflação de zeros e fração de cura, é dada por $p_{0i} = \exp\{-k_i\}$ e $p_{1i} = \exp\{-\theta_i\}$, no qual:

$$k_i = -\log\left\{\frac{\exp(x_{1i}^t\beta_1)}{1 + \exp(x_{1i}^t\beta_1) + \exp(x_{2i}^t\beta_2)}\right\} \quad \text{e} \quad \theta_i = -\log\left\{\frac{\exp(x_{2i}^t\beta_2)}{1 + \exp(x_{1i}^t\beta_1) + \exp(x_{2i}^t\beta_2)}\right\}, \tag{3.2}$$

sendo β_1 e β_2 vetores de coeficientes de regressão desconhecidos, os quais precisam ser estimados, os quais podem ser interpretados como a influências das variáveis na influência para a inflação de zeros e para a fração de cura.

Um ponto importante a ser dito é que, embora o modelo unificado de fração de cura tenha sido proposto em um contexto biológico, há uma grande utilização desta modelagem em outras áreas do conhecimento. Em alguns contextos, é utilizado a variável N como o

número de causas que competem para a ocorrência de um determinado evento de interesse, por exemplo, a inadimplência ou adimplência de um indivíduo. Então, admitindo estender a metodologia para outros contextos, assim como realizado por Barriga *et al.* (2015), o objetivo é estudar o tempo até a inadimplência em uma carteira de crédito, tendo o número de causas competitivas, N, como uma distribuição geométrica, e função acumulada dada pela Weibull Inversa.

3.1 Modelo de taxa de cura inflacionado de zero Gompertz (MTCIZ-Gompertz)

Considerando a modelagem para o caso geral, nesta seção é apresentada uma das proposta para este trabalho que é associar a distribuição gompertz, apresentada na Subseção 2.2.3, a função de sobrevivência relacionada aos indivíduos suscetíveis a falha.

Seja T uma variável aleatória com distribuição Gompertz com parâmetros $\lambda > 0$ e $\gamma > 0$. Sua função de distribuição acumulada é dada por:

$$F_T(t) = 1 - \mathbf{exp}\left\{ -\left(\frac{\lambda}{\gamma}\right)(\mathbf{exp}(\gamma t) - 1)\right\}, \tag{3.3}$$

em que $\lambda > 0, \gamma > 0$ e $t > 0$, λ parâmetro de escala e γ parâmetro de forma.

Não é muito difícil constatar que a sua função de risco acumulado $H_T(t)$ é dada por:

$$H_T(t) = \frac{\lambda}{\gamma}\left(\exp\{\gamma t\} - 1\right). \tag{3.4}$$

Deste modo, é possível definir a função de sobrevivência populacional para o modelo de taxa de cura inflacionado de zero Gompertz, que é:

$$S_{pop}(t) = \exp\{-\theta\} + (1 - \exp\{-k\} - \exp\{-\theta\})S_0^*(t), \quad t \geq 0, \;\; \theta, k > 0, \tag{3.5}$$

em que $p_0 = \exp(-k)$ e $p_1 = \exp(-\theta)$.

Considerando a função de densidade para a distribuição Gompertz, definida na Subseção 2.11, e a função de distribuição acumulada (3.3), pode-se definir a função de sobrevivência associada a população suscetível a falha por meio de:

$$\begin{aligned} S_0^*(t) &= \frac{\exp\{-\theta F_T(t)\} - \exp\{-\theta\}}{1 - \exp\{-\theta\}} \\ &= \frac{\exp\Big\{ -\theta\big(1 - \mathbf{exp}\big\{ -(\tfrac{\lambda}{\gamma})(\mathbf{exp}(\gamma t) - 1)\big\}\big)\Big\} - \exp\big\{ -\theta\big\}}{1 - \exp\big\{ -\theta\big\}}, \end{aligned}$$

assim como a função densidade associada à população suscetível a falha,

$$\begin{aligned} f_0^*(t) &= \left(\frac{\exp\{-\theta F_T(t)\}}{1 - \exp\{-\theta\}}\right)\theta f_T(t) \\ &= \left(\frac{\exp\Big\{ -\theta(1 - \mathbf{exp}\big\{ -(\tfrac{\lambda}{\gamma})(\mathbf{exp}(\gamma t) - 1)\big\}\Big\}}{1 - \exp\big\{ -\theta\big\}}\right)\theta\lambda e^{\gamma t}\mathbf{exp}\Big\{ -\Big(\frac{\lambda}{\gamma}\Big)(\mathbf{exp}(\gamma t) - 1)\Big\}. \end{aligned}$$

3.2 Modelo de taxa de cura inflacionado de zero Weibull (MTCIZ-Weibull)

Uma segunda proposta de aplicação para este trabalho é associar a distribuição Weibull, apresentada na Subseção 2.2.2, à função de sobrevivência relacionada aos indivíduos suscetíveis a falha.

Seja então T uma variável aleatória com distribuição Weibull com parâmetros $\lambda > 0$ e $\gamma > 0$. Sua função acumulada é dada por:

$$F_T(t) = 1 - \mathbf{exp}\{-(\lambda t)^\gamma\}, \tag{3.6}$$

onde $\lambda > 0, \gamma > 0$ e $t > 0$, λ parâmetro de escala e γ parâmetro de forma.

Consequentemente, a função de risco acumulado $H_T(t)$ é definida por:

$$H_T(t) = (\lambda t)^\gamma. \tag{3.7}$$

Deste modo, é possível obter a função de sobrevivência populacional para o modelo de taxa de cura zero inflacionado Weibull por meio da seguinte equação:

$$S_{pop}(t) = \exp\{-\theta\} + (1 - \exp\{-k\} - \exp\{-\theta\})S_0^*(t), \quad t \geq 0, \;\; \theta, k > 0, \tag{3.8}$$

em que $p_0 = \exp(-k)$ e $p_1 = \exp(-\theta)$.

Considerando a função de densidade para a distribuição Weibull, definida na Subseção 2.10, e a função de distribuição acumulada (3.6), pode-se definir a função de sobrevivência associada à população suscetível a falha por meio de:

$$\begin{aligned} S_0^*(t) &= \frac{\exp\{-\theta F_T(t)\} - \exp\{-\theta\}}{1 - \exp\{-\theta\}} \\ &= \frac{\exp\Big\{ -\theta(1 - \mathbf{exp}\{-(\lambda t)^\gamma\})\Big\} - \exp\big\{ -\theta\big\}}{1 - \exp\big\{ -\theta\big\}}, \end{aligned}$$

assim como a função densidade associada a população suscetível a falha,

$$\begin{aligned} f_0^*(t) &= \left(\frac{\exp\{-\theta F_T(t)\}}{1 - \exp\{-\theta\}}\right)\theta f_T(t) \\ &= \left(\frac{\exp\Big\{ -\theta(1 - \mathbf{exp}\{-(\lambda t)^\gamma\})\Big\}}{1 - \exp\big\{ -\theta\big\}}\right)\theta\gamma\lambda^\gamma t^{\gamma-1}\mathbf{exp}\{-(\lambda t)^\gamma\}. \end{aligned}$$

3.3 Inferência

O interesse agora é realizar a estimação dos parâmetros associados ao modelo. Para isto, então, é utilizado o método de máxima verossimilhança. Em de Oliveira *et al.* (2017), os autores dividem a contribuição dos indivíduos em três diferentes subgrupos: (I) indivíduos que realizam o evento de interesse logo no início, ou seja, com tempo igual a zero; (II) indivíduos não suscetíveis ao evento, ou seja, os que são considerados curados/imunes e, por fim, (III) aqueles que estão suscetíveis ao evento de falha. Assim, atribui-se os seguintes valores para cada um dos subgrupos:

$$\begin{cases} p_{0i}, & \text{se} \quad t_i = 0; \\ (1 - p_{0i} - p_{1i}) f_0^*(t_i), & \text{se} \quad t_i \text{ é falha}; \\ p_{1i} - (1 - p_{0i} - p_{1i}) S_0^*(t_i), & \text{se} \quad t_i \text{ é censura}. \end{cases}$$

Considere, então, os dados disposto da seguinte forma forma, $\mathbb{D} = (t_i, \delta_i, \mathbf{X})$, representando os tempos observados, a indicadora de censura e a matrix de covariáveis respectivamente. Leve em conta também o vetor de parâmetros associados ao modelo, $\zeta = (\theta, \beta_\theta, \beta_k)$, representando para este caso os parâmetros associados à distribuição do tempo de falhas, e os coeficientes do modelo de regressão referente a obtenção de p_0 e p_1. Assim, a função de verossimilhança para o modelo é dada pela seguinte expressão:

$$L(\zeta;\mathbb{D}) = \prod_{t_i=0}\{p_{0i}\}\prod_{t_i>0}\left\{\left[(1-p_{0i}-p_{1i})f_0^*(t)\right]^{\delta_i}\left[p_{1i}+(1-p_{0i}-p_{1i})S_0^*(t)\right]^{1-\delta_i}\right\}.$$

Assim, na maioria dos casos, a complexidade da função de verossimilhança obtida pelo modelo é bem grande, fazendo, assim, necessário a maximizar a função de verossimilhança de maneira numérica. Neste Trabalho em questão, será realizada a utilização da função *Optim* do pacote *stats4* do *software R*, utilizando assim o método "BFGS"para a maximização da função de verossimilhança.

Conforme Migon *et al.* (2014) e Ospina e Ferrari (2012), para amostras suficientemente grandes, as inferências acerca dos parâmetros são baseadas em propriedade de normalidade assintótica. Seja, então, $\widehat{\zeta_i}$ a estimativa de verossimilhança para o i-ésimo parâmetro do modelo, então $\widehat{\zeta_i} - \zeta_i$ possui distribuição assintótica a qual é a normal p-variada, com média zero, e matriz de variância e covariâncias $I^{-1}(\hat{\zeta})$, sendo assim a inversa da matriz de informação de Fisher estimada. Logo, obtém-se intervalos de confiança assintóticos, utilizando-se $100(1-\alpha)\%$ como o nível de confiança.

$$IC(\zeta_i; 100(1-\alpha)\%) = \widehat{\zeta_i} \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{\zeta_i})},$$

em que $z_{\alpha/2}$ é o percentil de $\alpha/2$ referente à distribuição normal padrão, e $\text{Var}(\widehat{\zeta_i})$ representa os valores da diagonal da matriz $I^{-1}(\hat{\zeta})$, respectivo ao seu parâmetro.

De acordo com Ceccotti (2015) em pequenas amostras pode ocorrer de que algumas propriedades do estimador de máxima verossimilhança não sejam satisfeitas. Consequentemente, podendo ocorrer problemas, tais como valores de intervalos que não sejam contemplados pelo espaço paramétrico, comumente conhecido como problema de fronteira.

Para resolver o problema de fronteira, é de costume usar transformações nos parâmetros. Construindo-se um intervalo para este novo parâmetro transfomado, por exemplo, usando o método delta e, em seguida, reajustando o intervalo para retornar à escala original (Ceccotti, 2015). De todo modo, neste trabalho este problema não será tratado.

Além disto, testes como o de Wald, Escore e Razão de Verossimilhança são muito utilizados para testar hipóteses relacionadas aos parâmetros (Colosimo e Giolo, 2006b). Utiliza-se de critérios como o de Akaike, proposto por Akaike (1974), e o critério bayesiano por Schwarz *et al.* (1978), os quais são frequentemente utilizados nas mais variadas áreas. Em ambos, o melhor modelo é aquele associados aos menores valores para seus critérios.

3.4 Simulação

Nesta seção conduzimos um estudo de simulação para investigar a consistência e eficiência dos EMVs que podem ser obtidos através das equações (3.3) com base em diferentes tamanhos amostrais. Para tanto utizamos 3 cenários e três critérios: o viés, e erro quadrático médio (EQM) e a probabilidade de cobertura (PC), os quais são dados, respectivamente, por:

$$\text{Vies}\left(\widehat{\zeta}_w\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\zeta}_w^{(m)} - \zeta_w\right);$$

$$\text{EQM}\left(\widehat{\zeta}_w\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\zeta}_w^{(m)} - \zeta_w\right)^2$$

e

$$\text{PC}\left(\widehat{\zeta}_w\right) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}\left(\zeta_w \in \widehat{\zeta}_w \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{\zeta}_w)}\right)$$

para $w = 1, \ldots, \kappa$, em que M é o numero de replicações Monte Carlo e $\zeta = (\zeta_1, \ldots, \zeta_\kappa) = \left(\theta, \beta_{11}, \ldots, \beta_{pn_p}, \alpha_{11}, \ldots, \alpha_{pn_p}\right)$ representa o vetor de parâmetros. Entretanto, $\widehat{\zeta}_w^{(m)}$ denota o EMV de ζ_w obtida da amostra m, para $m = 1, \ldots, M$. Por meio de 1000 amostras bootstrap.

Por esta abordagem, espera-se que bons estimadores tenham viés e EMQ próximos de zero. Por sua vez, os intervalos de confiança razoáveis, que são obtidos aqui usando a normalidade assintótica dos EMVs, devem ter amplitude pequena e com probabilidade de cobertura próxima ao valor nominal de 95%. Neste trabalho, todos os cálculos e simulações foram realizados no software R (R Core Team, 2019).

Para verificar o impacto de covariáveis no modelo, é considerado uma variável binária, X, a qual assume dois valores 0 e 1 de uma distribuição Bernoulli com parâmetro $0,5$. Por exemplo, poderíamos considerar esta variável como sendo individuos que pertecem ou não a um determinado grupo, atribuindo assim valores de 1 e 0, respectivamente.

A distribuição Weibull com parâmetros α_1 e α_2 foi usada para modelar o tempo de falha. Assim, considerando presença da covariável teremos oitos parâmetros regressores (β_{10}, β_{11}, β_{20}, β_{21}, β_{30}, β_{31}, β_{40} e β_{41}), os quais são visto nas seguintes expressão:

$$k_{1i} \;=\; -\log\left(\frac{\exp\{\beta_{10} + x_i\beta_{11}\}}{1 + \exp\{\beta_{10} + x_i\beta_{11}\} + \exp\{\beta_{20} + x_i\beta_{21}\}}\right),$$

$$\theta_{1i} = -\log\left(\frac{\exp\{\beta_{20} + x_i\beta_{21}\}}{1 + \exp\{\beta_{10} + x_i\beta_{11}\} + \exp\{\beta_{20} + x_i\beta_{21}\}}\right),$$

$$\alpha_{1i} = \exp\{\beta_{30} + x_i\beta_{31}\},$$

$$\alpha_{2i} = \exp\{\beta_{40} + x_i\beta_{41}\}.$$

Os cenários escolhidos para a simulação do modelo foram estabelecidos utilizando os valores para os parâmetros de regressão, dados pela Tabela 3.1:

Tabela 3.1: Valores dos parâmetros para diferentes cenários.

Parâmetros	Cenários		
	1	2	3
β_{10}	-1.75	-0.25	-0.50
β_{11}	0.50	1.00	0.50
β_{20}	-0.75	0.50	-0.75
β_{21}	0.50	-1.00	0.75
β_{30}	0.50	0.50	-0.75
β_{31}	0.50	1.50	1.00
β_{40}	1.50	-0.75	1.25
β_{41}	2.00	3.00	1.00

Com os valores da Tabela 3.1, calculou-se os valores das proporções de cura e zeros correspondentes com os valores fixos de (β_{10}, β_{11}, β_{20}, β_{21}, β_{30}, β_{31}, β_{40} e β_{41}) e aos valores das covariáveis como sendo 0 ou 1.

Tabela 3.2: Valores para as proporções de zeros e cura para diferentes cenários.

Proporções de zeros ou cura	X	Cenários		
		1	2	3
p_0	0	28,69%	22,72%	22,72%
	1	37,74%	16,29%	33,33%
p_1	0	10.55%	48,10%	29,17%
	1	13,87%	56,85%	33,33%

Algo interessante a se destacar, é que ao realizar a comparação entre os cenários, o

primeiro cenário é considerado com baixa ocorrência de individuos que não apresentaram o evento ao final do estudo, enquanto 3 e 2 são considerados como média e alta incidência, respectivamente.

O seguinte algoritmo será utilizado para a geração de observações para o modelo:

1. Estabeleça valores para os parâmetros do modelo β_{10}, β_{11}, β_{20}, β_{21}, β_{30}, β_{31}, β_{40} e β_{41};
2. Gere x_i distribuido como uma Bernoulli com parâmetro 0,5 e compute p_{0i}, p_{1i}, α_{1i} e α_{2i};
3. Gera u_i distribuido uniformemente no intervalo (0,1);
4. Caso $u_i \leq p_{0i}$, $s_i = 0$;
5. Caso $u_i > p_{0i}$, $s_i = \infty$;
6. Caso $p_{0i} < u_i \leq p_{1i}$, gere v_i distribuido uniformemente no intervalo $(p_{0i},\ 1 - p_{1i})$ e escolha s_i como sendo a raiz de $F(s_i) - v_i = 0$;
7. Simule w_i distribuido como uma uniforme no intervalo $(0, \max(s_i)$, levando em conta apenas os valores finitos para s_i;
8. Calcule $t_i = \min(s_i, w_i)$, caso $t_i < w_i$, use $\delta_i = 1$, caso contrário, $\delta_i = 0$;
9. Refaça n vezes os passos de 2 a 8. Repare que a censura é distruida uniformemente com um alcance limitado, pois deste modo é possível deixar as taxas de censuras em uma quantia razoável. (Rocha *et al.*, 2017, p. 12).

No estudo realizado obteve-se as EMV's dos parâmetros e seus erros padrão para cada cenário. Essas estimativas foram usadas para calcular o viés, o erro quadrático médio (EQMR) e a probabilidade de cobertura (PC) para cada tamanho de amostra e cenário estabelecido. Para os três cenários, foram considerados diferentes tamanhos de amostra (n): $100, 250, 500, 750$ e 1.000.

As Figuras 3.2- 3.4 mostram os resultados do estudo de simulação considerando, em que cenário 1 (■), cenário 2 (▲) e cenário 3 (●).

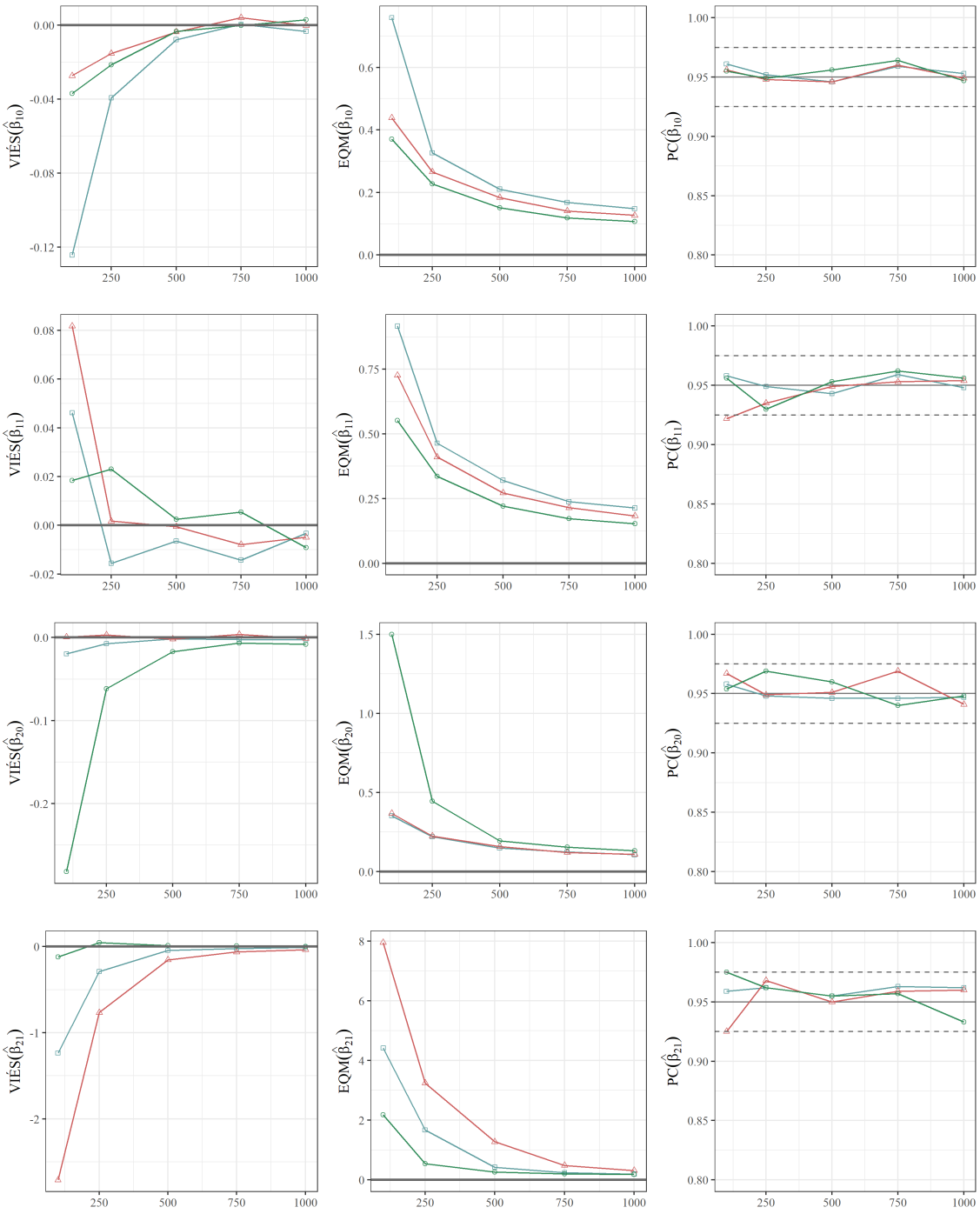

Figura 3.2: Viés, raiz quadrada do erro quadrático médio e probabilidade de cobertura (CP) do estimador de máxima verossimilhança de ($\hat{\beta}_{10}$, $\hat{\beta}_{11}$, $\hat{\beta}_{20}$, $\hat{\beta}_{21}$) do modelo de taxa de cura inflacionado de zero Weibull utilizando dados simulados sob os três cenários de parâmetros e sob diferentes tamanhos de amostrais (n).

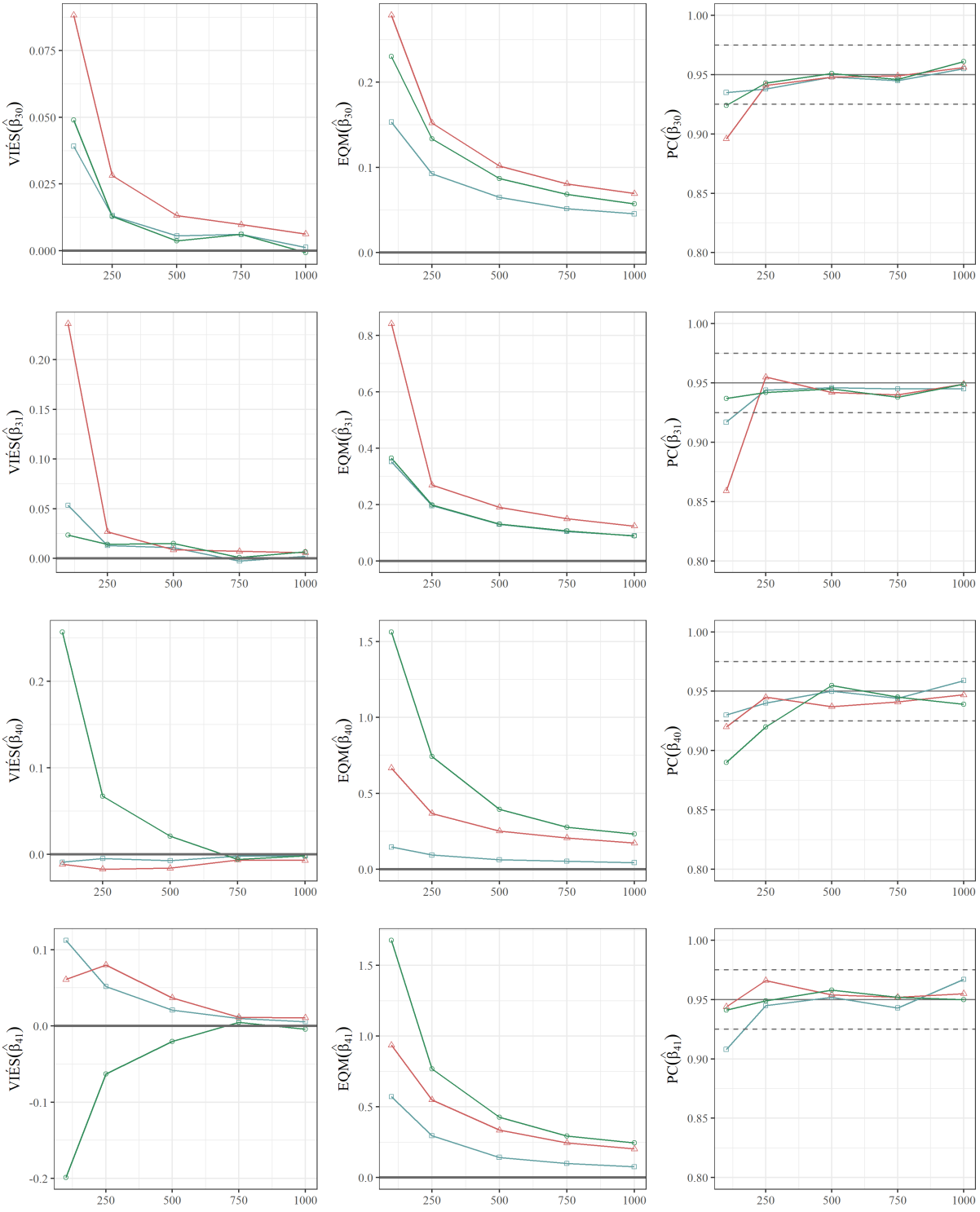

Figura 3.3: Viés, raiz quadrada do erro quadrático médio e probabilidade de cobertura (CP) do estimador de máxima verossimilhança de ($\hat{\beta}_{30}$, $\hat{\beta}_{31}$, $\hat{\beta}_{40}$, $\hat{\beta}_{41}$) do modelo de taxa de cura inflacionado de zero Weibull utilizando dados simulados sob os três cenários de parâmetros e sob diferentes tamanhos de amostrais (n).

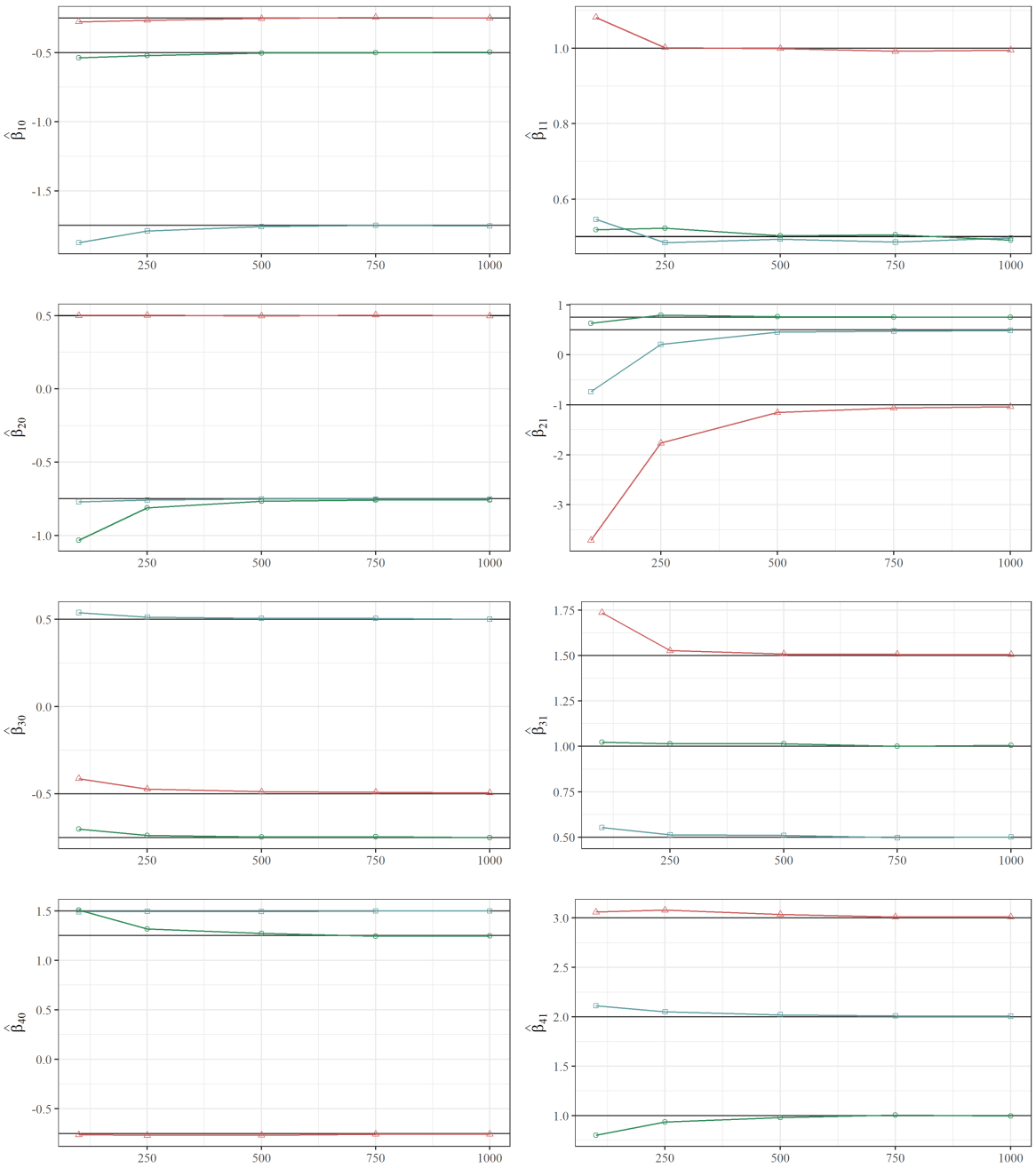

Figura 3.4: Média das estimativas de máxima verossimilhança de todos os parâmetros.

De acordo com os resultados mostrados nas Figuras 3.2, 3.3 e 3.4, podemos concluir o seguinte: É possível verificar, que em média, o estimador de máxima verossimilhança obtêm valores bem próximos aos valores reais em todos os cenários, ou seja, pertos dos valores prefixados para a geração dos cenários. Entretanto, para o parâmetro β_{21} fez-se necessário uma amostra maior, de pelo menos 500 observações, para que se fosse obtida a convergência, em contrapartida tem-se que os parâmetros β_{31} e β_{41} obtiveram viéses com convergência mais devagar para os cenários cenário 2 (▲) e cenário 3 (●), respectivamente.

Pelas Figura 3.2 e Figura 3.3 é possível verificar que há uma tendência de diminuição nos valores dos víeses e erros quadráticos médios à medida que aumenta-se o tamanho amostral. Também é possível salientar, sem especificar um parâmetro em específico, que a probabilidade de cobertura obtém o verdadeiro valor, isto é, considera-se que está próximo do valor esperado de 95%.

Observa-se também que o cenário 1 (■) tem a menor presença de excesso de zeros e de curados, as médias do EMQ, Viés e CP obtiveram melhores resultados aos parâmetros voltados aos coeficientes de regressão que ajudam a estimar α_1 e α_2, quando comparado aos outros dois cenários sendo dado pelo maior número de observações com o tempo com desfecho do evento de interesse. Em compensação, os outros dois cenários, os quais têm maiores frações de curas e de zeros, obtiveram melhores métricas aos parâmetros de regressão que são utilizados para realizar a estimação de $p_0 = \exp\{-k\}$ e $p_1 = \exp\{-\theta\}$, sendo um efeito do maior número de observaçõs contidas no excesso de zeros ou censura.

3.5 Consideração finais

Este capítulo apresentou a metodologia principal deste Trabalho, os modelos de fração de cura inflacionados de zero, sendo eles os modelos MTCIZ-Weibull e MTCIZ-Gompertz. Para estimação dos parâmetros, foi considerada uma abordagem clássica utilizando método de Máxima Verossimilhança e com o intuito de verificar as propriedades frequentistas foi feito um estudo de simulação.

Capítulo 4

Aplicação a dados financeiros

O conjunto de dados, utilizado neste trabalho, foi concedido por uma instituição financeira brasileira, a qual realiza serviços voltados ao mercado de crédito. Tais dados reúnem informações que envolvem características voltadas aos hábitos e costumes de indivíduos em torno de compromissos envolvendo solicitações de crédito.

Foi considerado o período após a recessão brasileira, iniciada em meados de 2014, acentuando, assim, a situação da crise financeira no país. Foi considerado uma amostra aleatória de 9.645 CPF's ativos. Os indivíduos, que englobam esta base de dados, têm como característica principal a aquisição de dívidas, ou seja, a composição é feita por clientes com dívidas vencidas e não quitadas no período entre julho/2015 a dezembro/2015. Uma característica do perfil dos endividados é que cerca de 65% destas dívidas vieram de origem de instituições financeiras ou bancos, como, por exemplo, empréstimos bancários, cheque especial ou até mesmo cartão de crédito, contrapondo, assim, com os 35% advindos de outros ramos da economia, como empresas de varejo, indústrias, entre outros meios de prestação de serviço.

Em ambas as áreas, o processo de realização de cobrança é feita de maneira tradicional. Neste tipo de cobrança, comumente, o processo é efetuado por meio de cobranças telefônicas, cartas de cobrança ou por meio de ligações extrajudiciais. Algo acometido pelo cenário da crise ecônomica é a lentidão do processo de restituição do *status* dos clientes ao de adimplente, o que implica na utilização de modelos estatísticos de modo estimar o prazo para a ocorrência destes eventos.

O tempo de falha considerado para este estudo é o tempo de espera entre a data de aquisição da dívida até a finalização do estudo, sendo ele de 24 meses. Para verificar diferenças entre os comportamentos dos clientes para diferentes cenários, será estudado

o cenário com a utilização de duas covariáveis, as quais demonstraram efetividade ao estudo.

- **Informação de consulta:** Indicativo de consulta de empresas (de qualquer segmento) aos relatórios de créditos do cliente nos últimos 180 dias, tendo assim indícios de solicitação constante de crédito ao mercado;
- **Tipo de dívida:** Caracteriza a origem da dívida do cliente durante o período de crise, seja ela de origem financeira (bancos) ou outros segmentos.

Deste modo, é possível verificar, por meio da Tabela 4.1, a composição das covariáveis com relação as suas categorias.

Tabela 4.1: Quantidade por covariável.

Covariável	Descrição	Categoria	n	%
x_1	Informação de consulta	0 - Sem consulta	295	03,06%
		1 - Com consulta	9350	96,90%
x_2	Tipo de dívida	0 - Banco	5103	52,90%
		1 - Outros segmentos	4542	47,10%

Contudo, também é possível verificar qual é a distribuição dos subgrupos de clientes. Tendo assim o objetivo de verificar se, para diferentes características, é possível notar comportamentos diferentes em relação ao tempo de recuperação do *status* de adimplência para os diferentes subgrupos de clientes, sendo eles:

- **(i) Cliente com evento no tempo zero:** Clientes que realizaram o evento de interesse logo ao início, ou seja, indica se houve a regularização da dívida no tempo zero, tornando-se adimplentes imediatamente;
- **(ii) Cliente suscetível ao evento:** Clientes suscetíveis ao evento, ou seja, para o contexto são os indivíduos que regularizam a dívida no período de 24 meses, revertendo seu status novamente para ao de adimplência;
- **(iii) Cliente não suscetível ao evento:** Clientes não suscetíveis ao evento, que pela teoria são considerados curados/imunes, ou seja, para o contexto são os indivíduos que permaneceram com o status de inadimplentes mesmo após o período observado de 24 meses, continuando com suas dívidas em aberto.

Tabela 4.2: Subgrupos de clientes

Número de Clientes	Subgrupo		
	(I) Cliente com evento no tempo zero	(II) Cliente suscetível ao evento	(III) Cliente não suscetível ao evento
9.645	2.292	5.268	2.085
100%	24%	55%	22%

A Tabela 4.2 mostra que há uma grande concentração de eventos no tempo zero, aproximadente 24% das observações, caracterizando assim o excesso de zeros. Além disso, uma proporção de 22% dos clientes não apresentaram o evento de interesse resultando na parcela de indivíduos que, pela teoria, são considerados imunes, visto que, mesmo considerando um grande intervalo de tempo, não apresentaram o evento de interesse.

A Figura 4.1 mostra a distribuição dos tempos de regularização das dívidas para os dados observados.

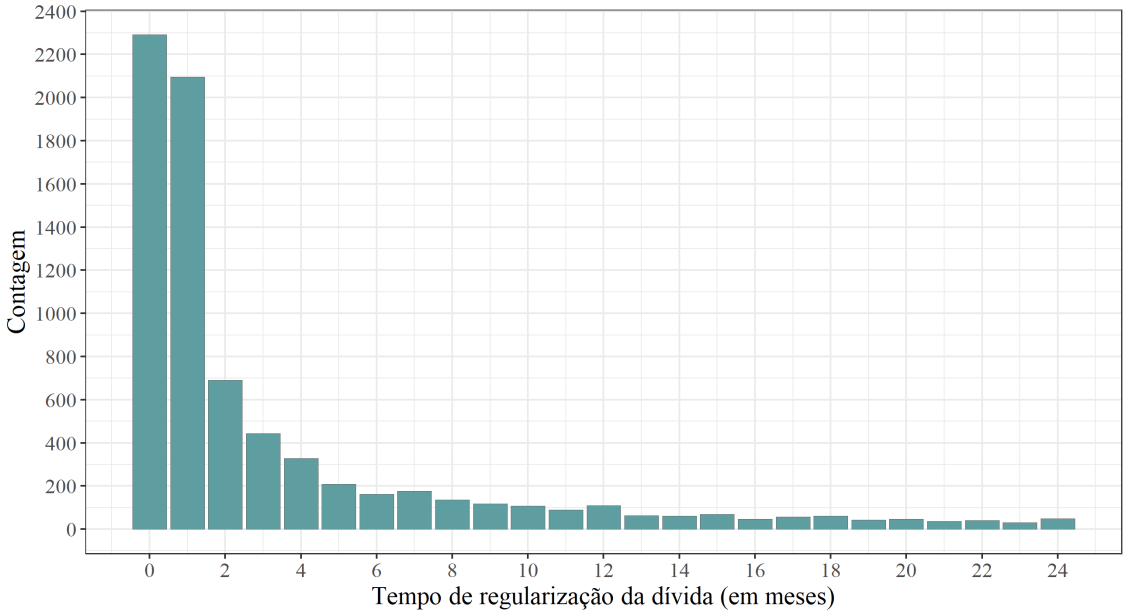

Figura 4.1: Gráfico de barras referente ao tempo de regularização da dívida (em meses).

É possível visualizar, pela Figura 4.1, a inflação de zeros para este banco de dados, o que é uma característica interessante visto que o estudo está sendo realizado em um cenário de crise econômica, caracterizando que uma grande parcela dos clientes endividados preferem realizar a quitação de suas dívidas logo no inicio, o que pode ser dado pelo interesse de normalização de seu *status* para realizações de outros feitos pessoais, os quais a inadimplência poderia impedir.

A Figura 4.2 mostra a curva de Kaplan-Meier estimada para os tempos de regularização da dívida. É possível verificar um grande número de censuras à direita, visto que, logo ao início do estudo, um grande número de clientes que adquiriram dívidas já haviam as quitado, notando assim a inflação de zeros. Outro ponto importante é a não estabilização da curva para o ponto $\widehat{S}_T(t) = 0$, observando visualmente a presença de fração de cura para os dados.

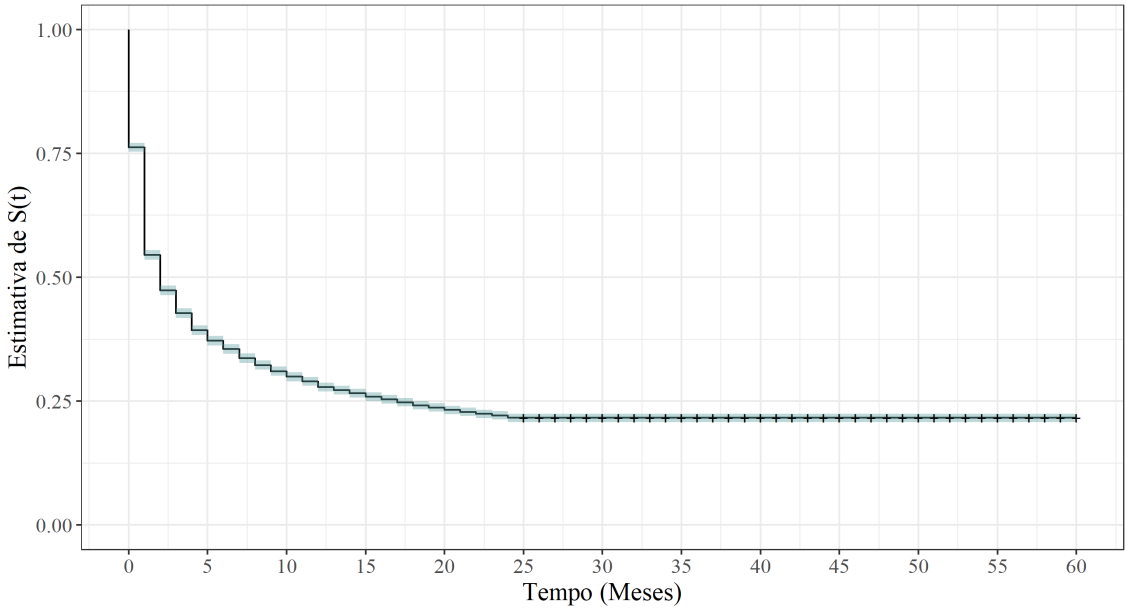

Figura 4.2: Curva de Kaplan-Meier para os tempos de regulaziação da dívida.

É importante salientar que, por padrão, os gráficos das estimativas de Kaplan-Meier, produzidos pelo pacote *ggplot2* do *Software R*, apresentam uma reta vertical no ponto 0 do eixo x, dando a impressão de que a curva começa no ponto 1. Porém, é importante reforçar que a curva só possui massa de probabilidade a partir do ponto que ela fica na horizontal, ou seja, para este gráfico em específico, a curva começa aproximadamente no ponto 0.75, verificando a presença de inflação de zeros.

Outro ponto interessante é a construção das curvas de Kaplan-Meier de maneira estratificada por covariáveis, visto na Figura 4.3, em que é possível verificar diferenças das curvas para diferentes entre as categorias dentro da covariável, caracterizando uma diferença entre os tempos de sobrevida, com destaque para o tipo de dívida, em que clientes com dívidas de origem financeira (bancos) costumam realizar a quitação de suas dívidas em uma maior proporção quando comparado às dívidas de origem provenientes de outros seguimentos, com uma grande diferença na estabilização das curvas, assim como características diferentes na inflação de zeros.

Para as curvas de sobrevivência estimadas será realizado o teste não parametrico “Log-Rank”, apresentado na Subseção 2.1.3, possibilitando verificar se a diferença observada entre as curvas são, estatisticamente, significantes.

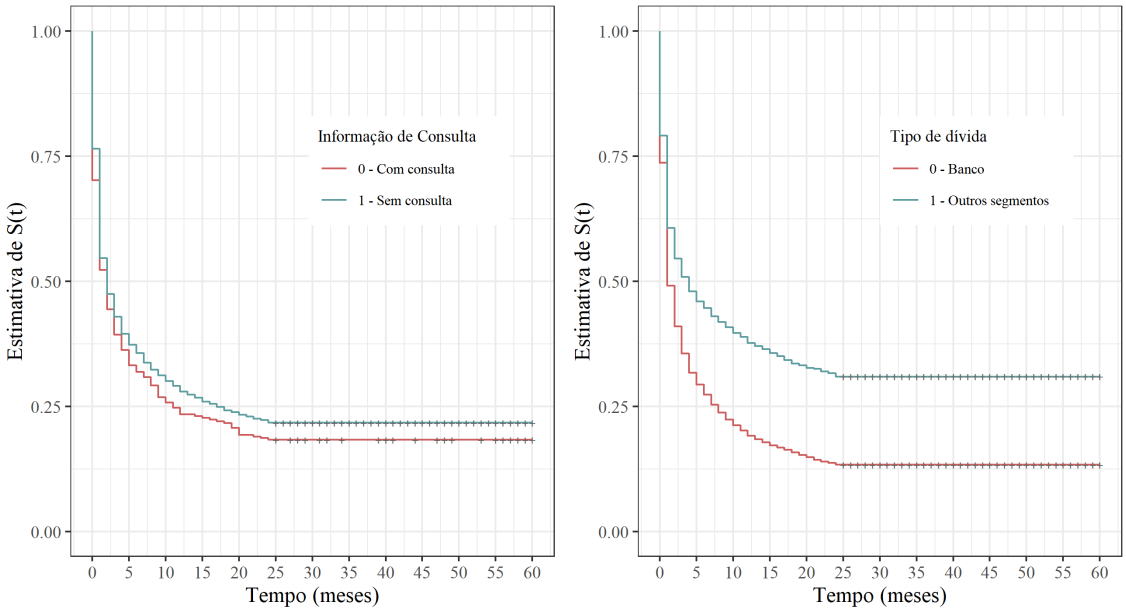

Figura 4.3: Curvas de Kaplan-Meier estimada considerando as covariáveis: Consulta aos relatórios de credito e Seguimento da divida adquirida.

Com base nas curvas observadas na Figura 4.3, o teste de “Log-Rank”nos retorna que o p-valor associado à variável “Informação de Consulta”é de 0.07, enquanto que, para a variável “Tipo de Dívida”, o p-valor resultante é $< 2 \cdot 10^{-16}$. Portanto, ao nível de significância de 0.10, conclui-se que que há diferença entre as curvas, resultando que clientes em situação de crise tendem a priorizar mais dívidas provenientes de instituições financeiras. Além disto, clientes sem consulta aos relatórios de crédito acabam priorizando mais o pagamento de dívida do que os que tiveram consulta.

4.1 Ajuste dos Modelos MTCIZ-Gompertz e MTCIZ-Weibull

Nesta seção a metodologia proposta, considerando como função de risco de base os modelos Gompertz e Weibull (ver Seções 3.1 e 3.2, respectivamente), foi aplicada ao conjunto de dados. Inicialmente, foram ajustados os modelos sem a presença de covariáveis e, posteriormente, a utilização das covariáveis será feita partindo da aplicação de cada uma

separadamente. Por fim, será analizado o modelo com todas as covariáveis conjuntamente.

Os resultados das análises do ajuste dos modelos são mostrados em forma de tabelas com estimativas dos parâmetros, assim como, erro padrão e intervalos de confiança, sendo também realizado testes para saber a significância dos parâmetros por estes intervalos com um posterior suporte de análise gráfica dos ajustes realizados. Para o ajuste dos modelos considerando as duas covariáveis conjuntamente, são mostrados somente as tabelas com as estimativas. Por fim, a escolha do melhor modelo será feita utilizando algumas métricas tais como AIC e BIC.

De modo a simplificar as interpretações, considere que os β_{1i} estão associados à influência das variáveis em relação a inflação de zeros, β_{2i} associado à influência das covariáveis na fração de cura.

Além disso, $p_{0i} = \exp\{-k_i\}$ é relacionado às proporções de zeros e $p_{1i} = \exp\{-\theta_i\}$ às proporções de cura, em que k_i e θ_i são obtidos de acordo com as expressões dadas pelas expressões 3.2. Ainda tem-se que λ e γ são os parâmetros da distribuição Gompertz ou Weibull. 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 covariável

A Tabela 4.3 mostra os resultados das estimativas dos parâmetros, erro padrão e seus intervalos de confiança de 95% obtidos no ajuste dos modelos MTCIZ-Weibull e MTCIZ-Gompertz dados nas Seções 3.1 e 3.2, respectivamente, sem a presença de covariáveis.

Tabela 4.3: Estimativas de máxima verossmilhança (EMV), erro-padrão (EP), Intervalo de confiança - IC(95%) para os modelos MTCIZ-Weibull e MTCIZ-Gompertz

	MTCIZ-Gompertz				MTCIZ-Weibull			
Parâmetros	EMV	EP	IC(95%)		EMV	EP	IC(95%)	
			LI	LS			LI	LS
γ	0.0031	0.0034	-0.0035	0.0098	1.0969	0.0110	1.0754	1.1186
λ	0.1340	0.0030	0.1278	0.1397	0.1379	0.0022	0.1337	0.1422
θ	1.4356	0.0250	-0.8828	-0.7847	1.4316	0.0250	-0.8822	-0.7841
k	1.5407	0.0259	-0.9896	-0.8881	1.5537	0.0259	-1.0060	-0.9046
p_0	0.2380	0.0043	0.2294	0.2464	0.2389	0.0043	0.2304	0.2474
p_1	0.2142	0.0041	0.2061	0.2224	0.2114	0.0041	0.2033	0.2195

A Tabela 4.3 mostra que as estimativa dos parâmetros associados a proporção de cura p_0 e proporção de zeros p_1 foram bem próximas para ambos os modelos, assim como para os erros padrões, apenas diferenciando as estimativas relacionadas aos parâmetros da distribuição (Weibull ou Gompertz). Algo interessante de se notar é que, praticamente, todos os parâmetros são significativos, ao nível de confiança de 5%, visto que a maioria das regiões de confiança não englobam o valor zero.

A Figura 4.4 mostra o ajuste dos modelos Weibull e Gompertz sem a presença de covariáveis. É possível verificar que o modelo Gompertz (a) obteve um ajuste bem similar quando comparado ao modelo Weibull (b), visto que as curvas estimadas para ambos os modelos estão bem próximas da curva estimada de Kaplan-Meier. O interessante de se observar é que, para a região central, aparentemente o modelo Gompertz possui um melhor ajuste, porém, quando foca-se mais na região das caudas, nota-se que o modelo Weibull se ajusta melhor, o que é uma característica da própria distribuição, a qual contêm caudas mais pesadas.

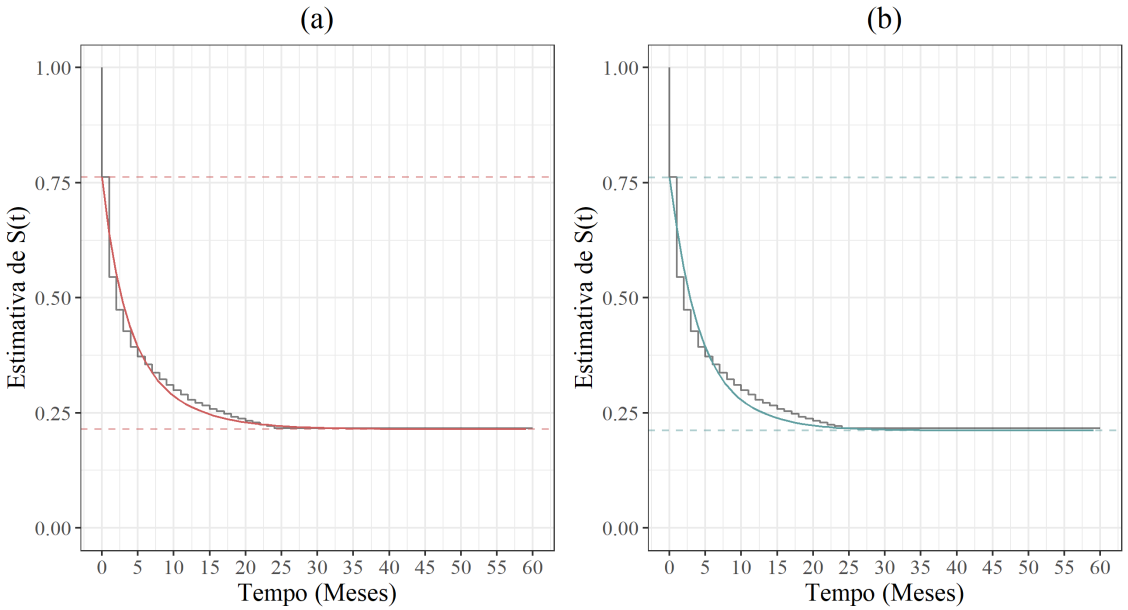

Figura 4.4: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo MTCIZ-Gompertz (a), MTCIZ-Weibull (b), sem a presença de covariável.

Como os dois modelos MTCIZ-Weibull e MTCIZ-Gompertz tiveram um bom ajuste, a escolha do melhor modelo sem a presença de covariável poderá ser dada por qualquer um dos dois, dependendo do interesse do pesquisador. De qualquer modo, isto poderá ser visto ao final deste capítulo.

Considerando a função de sobrevivencia dada pela Equação 3.1, é possível obter uma

relação com a função de risco acumulado populacional, $\hat{H}_{pop}(t) = -log(\hat{S}_{pop}(t))$.A Figura (4.5) mostra a curva estimada da função de risco acumulada

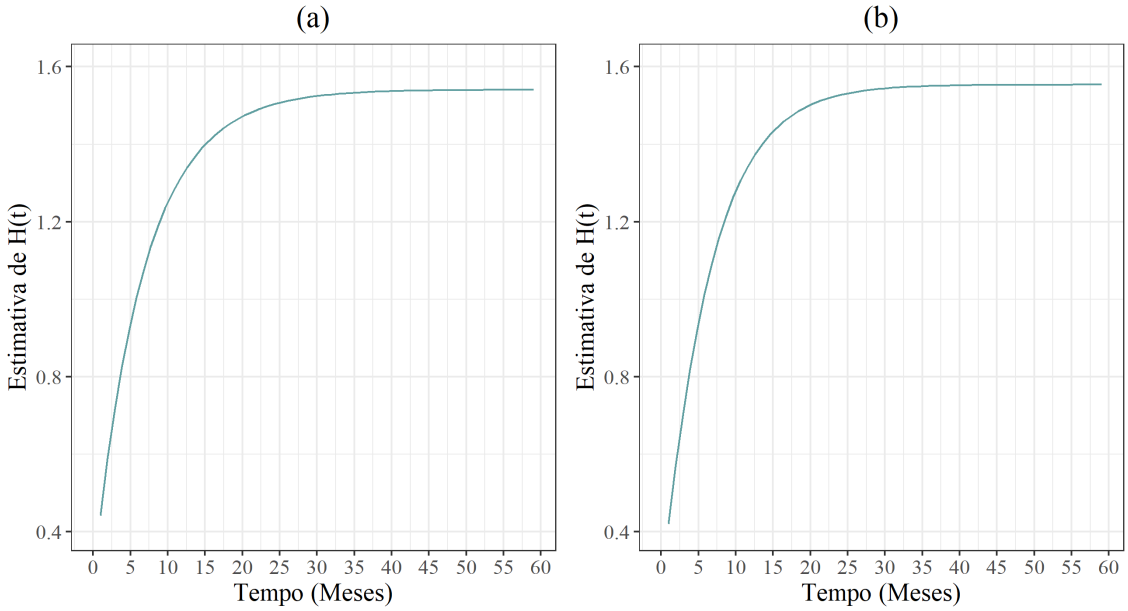

Figura 4.5: Estimativa da função de risco acumulada pelo MTCIZ-Gompertz (a), MTCIZ-Weibull (b), sem a presença de covariável.

Pela Figura 4.5 verifica-se que há um risco maior de um indivíduo, que adquiriu alguma dívida, quitar sua dívida com maior chance até o trigésimo mês, visto que a curva de risco acumulada estimada se estabiliza logo após este ponto, tendo que o risco de um indívuo quitar sua dívida em até 35 meses ou em até 60 meses quase a mesma medida.

4.1.2 Ajuste dos modelos na presença das covariáveis (Separadamente)

A Tabela 4.4 mostra a estimativa dos parâmetros, erro padrão e seus intervalos de confiança de 95% obtidos pelo ajuste do MTCIZ-Gompertz para cada covariável separadamente.

Tabela 4.4: Estimativa de máxima verossimilhança (EMV), erro-padrão (EP) e Intervalo de confiança - IC(95%) MTCIZ-Gompertz para as covariáveis x_1 e x_2

	Covariável x_1				Covariável x_2			
Parâmetros	EMV	EP	IC(95%)		EMV	EP	IC(95%)	
			LI	LS			LI	LS
γ	0.0031	0.0034	-0.0036	0.0098	0.0059	0.0059	-0.0007	0.0125
λ	0.1337	0.0030	0.1278	0.1397	0.1275	0.1275	0.1218	0.1333
$\beta_{10_{(intercepto)}}$	-0.5556	0.1339	-0.8180	-0.2933	-0.8313	-0.8313	-0.8954	-0.7672
$\beta_{11_{(X_i=1)}}$	-0.2880	0.1363	-0.5551	-0.0210	-0.0058	-0.0059	-0.1054	0.0937
$\beta_{20_{(intercepto)}}$	-0.9907	0.1457	-1.2764	-0.7050	-1.4845	-1.4846	-1.5629	-1.4062
$\beta_{21_{(X_i=1)}}$	0.0534	0.1475	-0.2358	0.3426	1.0098	1.0098	0.9102	1.1093
p_{00}	0.2950	0.0890	0.1206	0.4694	0.2620	0.0234	0.2162	0.3078
p_{01}	0.2361	0.0186	0.1996	0.2726	0.2107	0.0288	0.1543	0.2670
p_{10}	0.1909	0.1111	-0.0268	0.4086	0.1363	0.0334	0.0708	0.2019
p_{11}	0.2150	0.0197	0.1764	0.2535	0.3027	0.0221	0.2594	0.3460

A Tabela 4.4 mostra que a covariável "informação de consulta"tem maior inflação de zeros $p_{00} = 0.2950$ para clientes que não obtiveram consulta, enquanto a menor proporção de zeros é para o modelo "Tipo de débito"com $p_{01} = 0.2107$ para quando a dívida é referente a outros seguimentos. A maior proporção de cura é dada para o modelo com a covariável "Tipo de débito", com $p_{11} = 0.3027$ para clientes com dívidas de outros segmentos, já a menor proporção de cura é vista também para o mesmo modelo, com $p_{10} = 0.1363$ só que agora para quando a dívida é do segmento financeiro (bancos).

Observa-se que quase todos os parâmetros envolvendo os modelos foram significativos, dado que em sua maioria as regiões de confiança estabelecidas não contemplam o valor zero. Em especial, nota-se que a proporção de cura p_{10} foi não significativa.

A Figura 4.6 mostra a curva de sobrevivência estimada pelo modelo MTCIZ-Gompertz para as covariáveis x_1 =Informação de Consulta e x_2 = Tipo de dívida. Nota-se um bom ajuste do modelo proposto para ambas as covariáveis, visto o comportamento similar das curvas com as estimativas de kaplan meier.

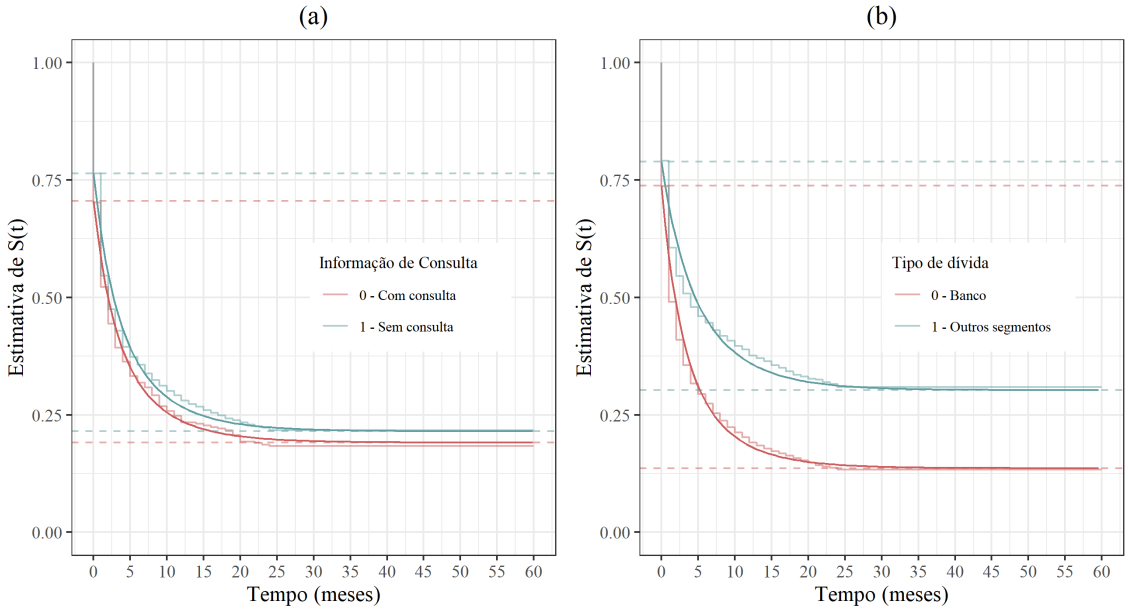

Figura 4.6: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo MTCIZ-Gompertz por covariável, Consulta aos relatórios de credito (a), Segmento da divida adquirida (b).

A Figura 4.7 mostra a curva estimada da função de risco acumulada, $\hat{H}_{pop}(t) = -log(\hat{S}_{pop}(t))$ do modelo MTCIZ-Gompertz (Seção 3.1).

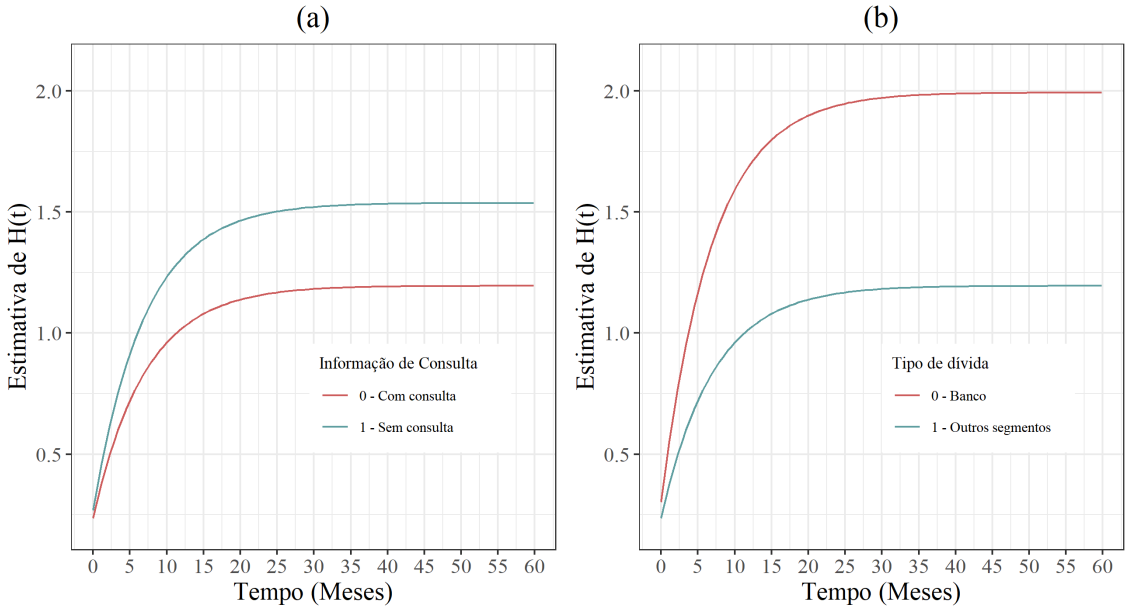

Figura 4.7: Estimativa da função de risco acumulado pelo MTCIZ-Gompertz por covariável, Consulta aos relatórios de credito (a), Segmento da divida adquirida (b).

Observa-se, pela figura 4.7, que o risco de um individuo vir em até determinado instante

de tempo quitar sua dívida, é maior para para clientes com o tipo de dívida advinda de bancos e sem histórico de consulta aos relatórios de crédito daquele cliente. O que possivelmente implicaria que um cliente proveniente desta combinação é o que mais terá chance de quitar suas dividas.

Tabela 4.5: Estimativa de máxima verossimilhança (EMV), erro-padrão (EP) e Intervalo de confiança - IC(95%) do MTCIZ-Weibull para x_1 e x_2

	Covariável x_1				Covariável x_2			
Parâmetros	EMV	EP	IC(95%)		EMV	EP	IC(95%)	
			LI	LS			LI	LS
γ	1.0970	0.0110	1.0754	1.1185	1.1061	0.0111	1.0843	1.1279
λ	0.1380	0.0021	0.1337	0.1422	0.1340	0.0021	0.1298	0.1381
$\beta_{10_{(intercepto)}}$	-0.5540	0.1337	-0.8162	-0.2917	-0.8310	0.0327	-0.8949	-0.7667
$\beta_{11_{(X_i=1)}}$	-0.2889	0.1461	-0.5558	-0.0219	-0.0056	0.0510	-0.1052	0.0939
$\beta_{20_{(intercepto)}}$	-1.0097	0.1465	-1.2969	-0.7225	-1.5113	0.0400	-1.5898	-1.4328
$\beta_{21_{(X_i=1)}}$	0.0560	0.1483	-0.2348	0.3467	1.0283	0.0510	0.9283	1.1284
p_{00}	0.2964	0.0888	0.1223	0.4705	0.2630	0.0234	0.2173	0.3088
p_{01}	0.2371	0.0186	0.2007	0.2735	0.2113	0.0287	0.1550	0.2676
p_{10}	0.1879	0.1122	-0.0320	0.4077	0.1332	0.0336	0.0673	0.1991
p_{11}	0.2122	0.0197	0.1735	0.2508	0.3009	0.0221	0.2575	0.3443

Pela Tabela 4.5 é possível ter as mesmas conclusões do modelo utilizando a distribuição gompertz em relação às maiores e menores proporções estimadas, o que já era esperado. O interessante é que se comparado os erros padrões destes modelos com os anteriores, vistos na Tabela 4.4, há uma grande proximidade entre os valores, o que acaba levando também as mesmas conclusões de significância de parâmetros dadas anteriormente.

Também é possível observar que a estimativa do parâmetro λ da distribuição Weibull obteve valores bem similares para ambos os modelos, assim como o parâmetro de forma γ, fazendo com que as estimativas de seus erros padrões praticamente se coincidam.

A Figura 4.8 mostra a curva de sobrevivência estimada pelo modelo MTCIZ-Weibull para cada covariável.

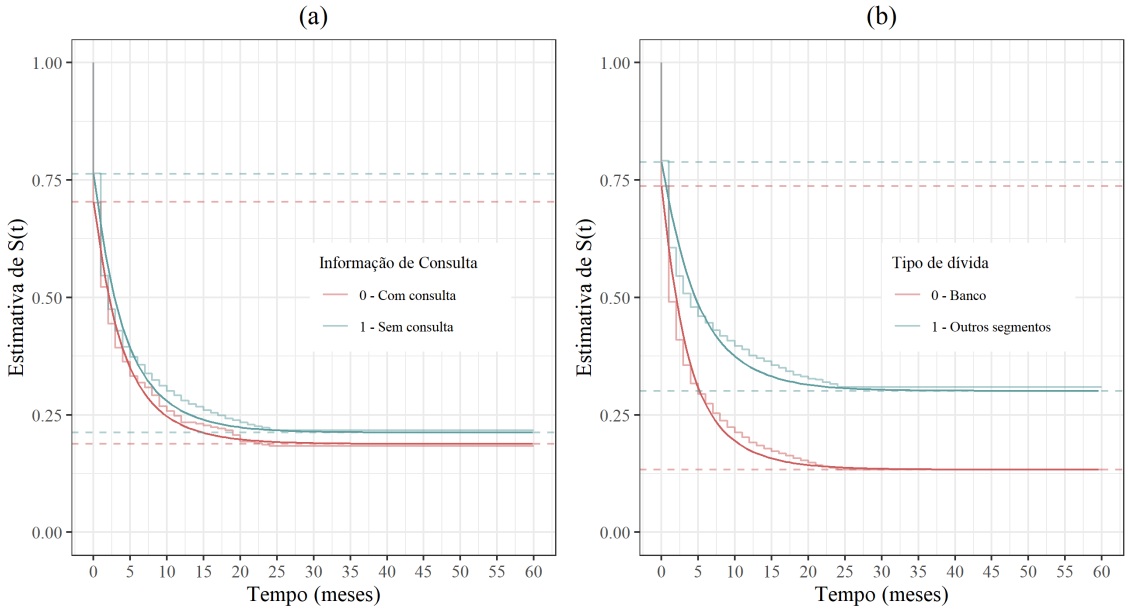

Figura 4.8: Estimativa de Kaplan-Meier e a curva de sobrevivência estimada pelo modelo MTCIZ-Weibull por covariável, Consulta aos relatórios de credito (a), Segmento da divida adquirida (b).

A Figura 4.8 mostra um ajuste bom para ambas as covariáveis envolvendo o modelo MTCIZ-Weibull, porém, quando comparado ao MTCIZ-Gompertz, apresentado anteriormente, é possível verificar um distanciamento maior para a regiões centrais de decaimento das curvas em relação à estimativa de Kaplan-Meier. De qualquer modo, isto poderá ser visto por meio dos critérios de escolha para os modelos que será dado posteriormente.

Em decorrência da estimativa para as curvas de sobrevivência do MTCIZ-Weibull visto anteriormente, é possível obter a relação com a função de risco acumulada por cada classe da covariável, tendo suas seguintes estimativas dadas pela Figura 4.9.

A Figura 4.9 mostra a função de risco acumulada do MTCIZ-Weibull (Seção 3.2), a qual pode ser interpretada da mesma forma que a Figura 4.7. Em que nota-se que o risco de um individuo vir em até determinado instante de tempo quitar sua dívida, é maior para para clientes com o tipo de dívida advinda de bancos e sem histórico de consulta aos relatórios de crédito daquele cliente. O que possivelmente implicaria que um cliente proveniente desta combinação é o que mais terá chance de quitar suas dividas.

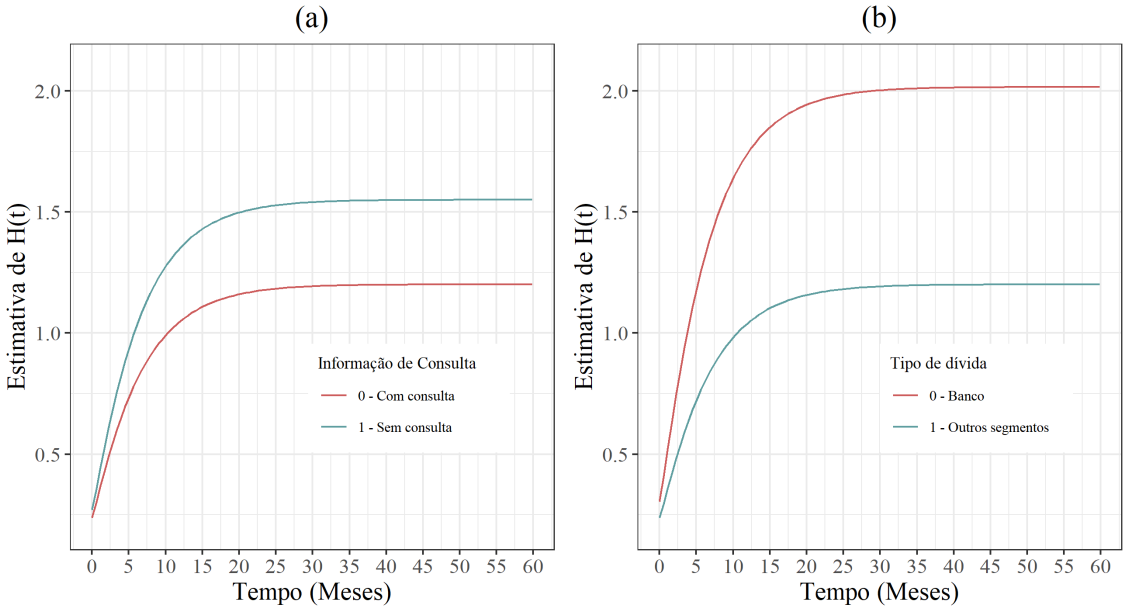

Figura 4.9: Estimativa da função de risco acumulado pelo modelo MTCIZ-Weibull por covariável, Consulta aos relatórios de credito (a), Segmento da divida adquirida (b).

4.1.3 Ajuste dos modelos na presença das covariáveis (Conjuntamente)

Nesta subseção será considerado o ajuste dos modelos para as duas variáveis conjuntamente. Inicialmente foi ajustado o modelo MTCIZ-Gompertz com a presença de ambas as covariáveis mostrado pelas Tabelas 4.6 e 4.7.

Tabela 4.6: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), Intervalo de confiança - IC(95%) do MTCIZ-Gompertz com x_1 e x_2.

Parâmetros	EMV	Erro-Padrão	IC(95%)	
			LI	LS
γ	0.0060	0.0034	-0.0006	0.0126
λ	0.1274	0.0029	0.1217	0.1332
$\beta_{10_{(intercepto)}}$	-0.5527	0.1353	-0.8178	-0.2876
$\beta_{11_{(X_1=1)}}$	-0.2882	0.1361	-0.5550	-0.0214
$\beta_{12_{(X_2=1)}}$	-0.0045	0.0508	-0.1040	0.0950
$\beta_{20_{(intercepto)}}$	-1.5317	0.1513	-1.8282	-1.2352
$\beta_{21_{(X_1=1)}}$	0.0475	0.1502	-0.2468	0.3418
$\beta_{22_{(X_2=1)}}$	1.0086	0.0508	0.9090	1.1082

Observa-se, na Tabela 4.6, as estimativas dos parâmetros associado ao modelo MTCIZ-Gompertz, sendo estes os parâmetros associados à distribuição Gompertz λ e γ, e as estimativas dos parâmetros de regressão β ligados às proporções, apontando também que a maioria dos parâmetros foram significativos, dado o mesmo critério visto nos modelos anteriores considerando os intervalos de confiança.

Pela Tabela 4.7, nota-se que a maior proporção de indivíduos que regularizam a sua dívida logo no instante zero está associado aos clientes que obtiveram consulta aos relatórios de créditos e dívidas advindas do ramo financeiro (bancos) com uma proporção de $p_{000} = 0.3212$, contrastando com a menor proporção de zeros, que é dada para clientes que não tiveram consulta e dívidas com origem de outros segmentos com apenas $p_{001} = 0.2094$.

Algo interessante a observar é que os clientes que não tiveram consulta e dívidas com origem de outros segmentos também são os que mais concentram indivíduos que não quitam suas dívidas, mesmo após um período de 24 meses, com uma proporção de $p_{111} = 0.3030$. Já clientes que obtiveram consulta e adquiriram dívidas do ramo financeiro (bancos) são os que menos ficam com dívidas pendentes, com apenas $p_{110} = 0.1367$.

Tabela 4.7: Estimativa das proporções de cura para o modelo MTCIZ-Gompertz com x_1 e x_2.

Proporção de cura e zero	x_1	x_2	Estimativa	E.P.	IC(95%)	
					LI	LS
p_0	0	0	0.3212	0.0876	0.1495	0.4929
		1	0.2645	0.0954	0.0775	0.4515
	1	0	0.2601	0.0237	0.2137	0.3066
		1	0.2094	0.0290	0.1525	0.2662
p_1	0	0	0.1207	0.1258	-0.1259	0.3673
		1	0.2737	0.1027	0.0724	0.4750
	1	0	0.1367	0.0336	0.0708	0.2026
		1	0.3030	0.0485	0.2080	0.3981

Pode ser observado quais são os padrões de clientes que mais tendem ou não a pagar suas dívidas ao final do período, de qualquer modo, quase todos as proporções do modelo se mostraram significativas, com exceção da proporção p_{100}.

Para o ajuste do modelo MTCIZ-Weibull com a presença de ambas as covariáveis

conjuntamente, as estimativas são mostradas nas Tabelas 4.8 e 4.9.

Tabela 4.8: Estimativas de máxima verossimilhança (EMV), erro-padrão (EP), Intervalo de confiança - IC(95%) do MTCIZ-Weibull com x_1 e x_2.

Parâmetros	EMV	Erro-Padrão	IC(95%)	
			LI	LS
γ	1.1061	0.0111	1.0843	1.1279
λ	0.1340	0.0021	0.1298	0.1381
$\beta_{10_{(intercepto)}}$	-0.5519	0.1353	-0.8171	-0.2866
$\beta_{11_{(X_1=1)}}$	-0.2888	0.1362	-0.5558	-0.0219
$\beta_{12_{(X_2=1)}}$	-0.0051	0.0508	-0.1047	0.0945
$\beta_{20_{(intercepto)}}$	-1.5529	0.1518	-1.8505	-1.2554
$\beta_{21_{(X_1=1)}}$	0.0428	0.1507	-0.2526	0.3382
$\beta_{22_{(X_2=1)}}$	1.0283	0.0510	0.9283	1.1284

Assim, como no modelo MTCIZ-Gompertz, nota-se, pela Tabela 4.8, as estimativas dos parâmetros associadas ao modelo Weibull, sendo estes os parâmetros associados à distribuição como λ e γ, e estimativas dos parâmetros de regressão β ligados as proporções, notando que a maioria dos parâmetros foram significativos, dado o mesmo critério visto nos modelos anteriores.

Tabela 4.9: Estimativa das proporções para o modelo MTCIZ-Weibull com x_1 e x_2.

Proporções de cura e zero	X_1	X_2	Estimativa	E.P.	IC(95%)	
					LI	LS
p_0	0	0	0.3222	0.0876	0.1505	0.4938
		1	0.2647	0.0955	0.0774	0.4519
	1	0	0.2611	0.0237	0.2147	0.3075
		1	0.2097	0.0290	0.1528	0.2666
p_1	0	0	0.1184	0.1266	-0.1298	0.3666
		1	0.2734	0.1032	0.0712	0.4756
	1	0	0.1337	0.0338	0.0674	0.2000
		1	0.3018	0.0488	0.2062	0.3974

Observa-se, pela Tabela 4.9, que as conclusões dadas para o modelo MTCIZ-Weibull são iguas as dadas para o modelo MTCIZ-Gompertz, visto os valores bem próximos ao

encontrados na Tabela 4.7, tanto para as estimativas dos parâmetros quanto para os erros padrões, o que, consequentemente, resulta em intervalos de confiança com a mesma interpratibilidade.

4.2 Critérios de seleção

Nesta seção são apresentados os critérios de akaike (AIC) e critério bayesiano (BIC) para a escolha do modelo, assim como mostrado pela Tabela 4.10.

Tabela 4.10: Critérios de seleção para os modelos ajustados.

	Critério	Covariável			
		sem cov	x_1	x_2	x_1 e x_2
MTCIZ-Gompertz	AIC	46319.38	46317.77	45867.06	45865.70
	BIC	46348.07	46360.82	45910.10	45923.09
MTCIZ-Weibull	AIC	46240.42	46238.83	45775.97	45774.62
	BIC	46269.11	46281.87	45819.01	45832.01

A Tabela 4.10 mostra que o modelo MTCIZ-Weibull apresenta critérios melhores quando comparado com o respectivo modelo MTCIZ-Gompertz, podendo contradizer um pouco o que foi dito em relação às curvas ajustadas, o que talvez pode ter sido ocasionado por um melhor ajuste do modelo MTCIZ-Weibull para as caudas das distribuição. Contudo, a diferença entre os critérios não são tão distantes uns dos outros. Outro ponto observado é que os modelos ajustado sem nenhuma covariável ou apenas utilizando a variável x_1 foram os que obtiveram os menores desempenhos, sendo muito próximos entre si.

Em resumo observa-se que os modelos contendo apenas a variável x_2 ou contendo x_1 e x_2 foram os modelos com melhores desempenhos, sendo o primeiro mais eficiente pelo critério bayesiano e o segundo pelo critério de Akaike, implicando que a escolha de qualquer um deles poderia ser realizada. Outro ponto importante é que a escolha entre o modelo MTCIZ-Weibull ou modelo MTCIZ-Gompertz é indiferente, visto que ambos demonstraram bastante eficazes, ou seja, qualquer um dos dois poderia ser escolhido, visto a insignificância das diferença entre os critérios dentro de cada uma das classes de covariáveis.

4.3 Considerações finais

Neste capítulo foi possível realizar uma aplicação a um conjunto de dados reais, o qual verificou-se que atendia as características principais desta modelagem, sendo elas a fração de cura (Clientes inadimplentes) e zeros (Clientes que regularizaram as suas dívidas no início do estudo). Foi possível constatar que os modelos MTCIZ-Weibull e MTCIZ-Gompertz se ajustaram bem, porém o modelo MTCIZ-Weibull foi selecionado como o mais adequado na presença da variável "Tipo de dívida".

Além disto, uma ténica que poderia ser aplicada é a análise de resíduos, a qual por meio dela poderia ser visto se existe ou não a adequabilidade do modelo proposto. Porém, neste trabalho não foi realizado, visto que a sua implementação depende de esforços computacionais que ainda não estão disponíveis.

Capítulo 5

Conclusão

Neste trabalho foi estudado a modelagem estatística baseado em análise de sobrevivência denominada como "modelo de fração de cura inflacionado de zero", tendo como uma das suas principais peculiariedades a incorporação de indivíduos que não estão suscetíveis ao evento de interesse mesmo após considerar um longo período de tempo e também a agregação de uma parcela de indivíduos que apresentam a falha no tempo zero, sendo esta segunda ainda pouco explorada na literatura.

A fim de demonstrar a efetividade das propriedades frequentistas, foi realizado um estudo de simulação considerando a adição de uma covariável binária, a qual foi incorporada em todas as partes do modelo, sendo elas, os parâmetros de forma e escala associados à distribuição de Weibull e às proporções de cura e zeros. Neste estudo observou-se que com o aumento do tamanho da amostra, as estimativas dos parâmetros, em média, são muito próximos aos valores reais, com diminuição dos viéses e erros quadrático médios. Além disso, também foi possível observar, sem especificar um parâmetro em específico, que a probabilidade de cobertura obteve o verdadeiro valor, isto é, considerou-se que estava próximo do valor esperado de 95%.

Foi possível mostrar a aplicabilidade do modelo por meio de um conjunto de dados reais de clientes que adquiriram dívidas, entre os meses de julho e dezembro de 2015. Utilizando-se a modelagem com a agregação de diferentes covariáveis, notou-se que clientes que foram consultados por empresas aos relátorios de busca e com dívidas vindas de outros segmentos, foram os que obtiveram a pior perfomance dentro de uma instituição, visto que resultaram na menor proporção de indivíduos que quitaram suas dívidas logo ao inicio do estudo e tiveram a maior proporção de indivíduos que ao final do estudo continuaram com o seu status de inadimplência. Em contrapartida, observou-se que clientes com dívidas de

origem bancária e sem consulta de crédito são os clientes com menores riscos de perdas monetárias para as instituições, considerando que são os que mais costumam pagar suas dívidas logo no inicio do período e os que mais costumam voltar ao status de adimplência pois é observado sua menor proporção de cura entre as classe de clientes estudadas.

A utilização dos modelos de taxa de cura inflacionados de zeros Weibull e Gompertz, os quais denominamos "MTCIZ-Weibull"e "MTCIZ-Gompertz"respectivamente, se mostraram adequados para o conjunto de dados analizado. Neste contexto, os modelos contendo apenas a variável "Tipo de dívida"ou "Informação de consulta"e "Tipo de dívida"obtiveram melhores desempenhos com base em medidas como AIC ou BIC. É importante salientar que o real desempenho do modelo a ser utilizado poderá ser realmente mensurado a partir do início do uso em empresas do ramo, tendo a possibilidade de explorá-lo mais ainda com a agregação de outras variáveis, dado que a modelagem permite a utilização de quantas variáveis forem necessárias.

Com este estudo foi possível fornecer informações sobre o comportamento de clientes dentro do mercado financeiro, possibilitando a tomada de decisões, assim como medidas necessárias para evitar certas situações com base em tempos estimados. Considerando propostas futuras, é possível além da realização de estimação via métodos frequentistas, realizar a estimação através de métodos bayesianos, podendo assim comparar com os resultados já obtidos neste relatório.

Por fim, para trabalhos futuros seria interessante a incorporação da análise de resíduos, pois com ela seria possível verificar se o modelo está ou não adequado. Neste Trabalho não foi realizado a incorporação da análise de resíduos por motivos de dificuldade de realizar a programação necessária.

Referências Bibliográficas

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

Adam, B., Beck, U. e Van Loon, J. (2000). *The risk society and beyond: critical issues for social theory*. Sage.

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

Barriga, G. D., Cancho, V. G. e Louzada, F. (2015). A non-default rate regression model for credit scoring. *Applied Stochastic Models in Business and Industry*, **31**(6), 846–861.

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.

Ceccotti, T. B. (2015). Intervalos de confiança baseados em deviance para os hiper-parâmetros em modelos estruturais.

Chaia, A. J. (2003). *Modelos de gestão do risco de crédito e sua aplicabilidade ao mercado brasileiro.*. Ph.D. thesis, Universidade de São Paulo.

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.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Colosimo, E. A. e Giolo, S. R. (2006a). *Análise de sobrevivência aplicada*. Editora Blucher.

Colosimo, E. A. e Giolo, S. R. (2006b). *Análise de Sobrevivência Aplicada*. Number 15 in Série do livro. Edgard Blucher, São Paulo. ISBN XX-XXX-XXXX-X. Bibliografia: p. 131–132.

de Oliveira, M. R., 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.

DINIZ, C. e LOUZADA, F. (2012). Modelagem estatística para risco de crédito. *ABE, São Paulo-SP*.

Douglas, M. e Wildavsky, A. (1983). *Risk and culture: An essay on the selection of technological and environmental dangers*. Univ of California Press.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Feller, W. (2008). *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons.

Fernandes, L. M. (2013). Inferencia bayesiana em modelos discretos com fracao de cura.

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.

Gill, A. E. (1980). Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society*, **106**(449), 447–462.

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.

Hosmer Jr, D. W., Lemeshow, S. e Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Ibrahim, J. G., Chen, M.-H. e Sinha, D. (2014). B ayesian 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.

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

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

Louzada-Neto, F., Mazucheli, J. e Achcar, J. A. (2001). *Uma introdução à análise de sobrevivência e confiabilidade*. Sociedad Chilena de Estadística.

Mantel, N. e Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, **22**(4), 719–748.

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.

Martz, H. F. e Waller, R. (1982). Bayesian reliability analysis. *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, 1982, 704*.

Meeker, W. Q. (1987). Limited failure population life tests: application to integrated circuit reliability. *Technometrics*, **29**(1), 51–65.

Migon, H. S., Gamerman, D. e Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**(4), 945–966.

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**, 90–99.

Ospina, R. e Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, **56**(6), 1609–1623.

Pereira, G. H., Botter, D. A. e Sandoval, M. C. (2013). A regression model for special proportions. *Statistical Modelling*, **13**(2), 125–151.

Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F. e Eudes, A. (2017). New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, **26**(4), 1737–1755.

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. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Shanker, R. e Hagos, F. (2016). On poisson–sujatha distribution and its applications to model count data from biological sciences. *Biometrics & Biostatistics International Journal*, **3**(4), 1–7.

Sicsú, A. L. (2010). *Credit Scoring: desenvolvimento, implantação, acompanhamento*. Blucher.

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

Toledo, J. S. B. (2011). Modelos de taxa de cura inflacionado de zero aplicado a dados de risco de crédito.

Tsodikov, A. D., Yakovlev, A. Y. e Asselain, B. (1996). *Stochastic models of tumor latency and their biostatistical applications*, volume 1. World Scientific.

Weibull, W. (1939). A statistical theory of the strength of materials, 1939. *Generalstabens Litografiska Anstalts Förlag*.

Weibull, W. (1951). A statistical distribution function of wide applicability. journal of applied mechanics 18: 293-297. *Statistical and Computational Analysis*, **291**.