Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Modelagem do risco de crédito via modelo de riscos competitivos

Marcia Araujo Maia

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Brasília
2018

Marcia Araujo Maia

Modelo de riscos competitivos aplicado à modelagem de dados com fração de cura não latente.

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. **Eduardo Yoshio Nakano**

Coorientadora: Prof. Dr. **Juliana Betini Fachini Gomes**

Brasília
2018

Agradecimentos

Agradeço a Deus pela dádiva de realizar os meus sonhos.

Agradeço à Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES) pela apoio financeiro para o desenvolvimento desse trabalho.

Agradeço ao professor Eduardo Yoshio Nakano e à professora Juliana Betini Fachini Gomes por orientar-me, por toda a prestatividade, pela paciência que tiveram e por compartilharem seus conhecimentos.

Agradeço a minha família, principalmente a minha mãe, por todo amor e suporte que me deram para que eu concluísse essa fase.

Agradeço ao André por todo amor e carinho.

Agradeço ao meu chefe Carlos Augusto Pacheco por possibilitar que eu fizesse o mestrado. Obrigada por entender quando eu tinha aula em horários aleatórios e complicados e mesmo assim você fez tudo para que eu pudesse seguir em frente.

Agradeço aos meus amigos do bonde da estatística (Ge, Andressa, Mateus, Cláudia, Mariana, Adolfo, Pablo, Agda e Rodrigo), que mesmo longe me alegravam muito quando vinham para Brasília nas férias ou nos feriados, obrigada por todas as distrações e pelas contínuas trocas de conhecimento.

Agradeço aos amigos de estudo do mestrado (Alisson, Ale, Eric, Ge, Bruno, Kessys, Alex, Maria Gabriela) por todos os momentos de união nos estudos. Sem vocês teria sido bem mais difícil.

Um obrigada especial à Ge pela amizade, por me deixar mudar para sua casa por semanas para estudarmos e por toda a parceria.

Um obrigada especial também para o Adolfo que foi mais que um amigo, principalmente por nos ajudar a resolver as questões que não conseguíamos.

Agradeço ao pessoal da secretaria da estatística, ao pessoal da portaria por serem tão gentis e prestativos.

Agradeço a todos meus amigos que estiveram comigo desde o início da graduação, Rosi, Nahari, Bianca Souza, Maria Júlia, Poliana, Vanessa Rezende, Cadu, Pedro Oliveira, Pedro Rangel, Ana Carolina e Luiza Tuler por todos os momentos de estudo, de brincadeira e alegria.

Resumo

O segmento de empréstimo é um dos investimentos que têm significativa rentabilidade para vários setores da economia, inclusive para Entidades Fechadas de Previdência Complementar (EFPC). Dessa forma, foi construída uma modelagem que auxilie na predição de riscos de inadimplência. O objetivo deste trabalho foi propor escore de risco de inadimplência baseado na modelagem de dados de sobrevivência na presença de eventos competitivos. A metodologia proposta foi aplicada a um conjunto de dados reais sobre clientes que realizaram um empréstimo em um fundo de pensão. Os resultados mostraram que a classificação dos clientes pelo escore de risco baseado no modelo log-logístico se mostrou útil para identificar os clientes "bons pagadores".

Palavras-chave: Análise de Sobrevivência, Dados Censurados, Inadimplência, Riscos Competitivos, Parâmetro de Força-Stress, Escore de Risco, Risco de Crédito.

Abstract

The loan segment is an investment that provides significant profitability for several sectors of the economy, including Pension Funds. Considering that, a model was built to predict the default risk. The main objective of this study was to present a default risk score based on the survival data model, considering the presence of competitive events. The proposed methodology was applied to a group of real data of clients whom have loaned money from Pension Funds. The results showed that the classification of clients by the risk score based on the log-logistic model proved useful to identify a "good-payer"customer.

Key Words: Survival Analysis, Censured Data, Default, Competitive Risks, Stress-Force Parameter, Risk Score, Credit Risk.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O Sistema Financeiro Nacional, segundo Ribeiro (2015), é composto por dois subsistemas: normativo e de intermediação. Uma das instituições financeiras que faz parte do subsistema de intermediação são as entidades fechadas de previdência complementar (EFPC).

A Resolução CMN Nº 4.661 de 25 de maio de 2018 orienta que os investimentos dos recursos dos planos administrados pela entidade fechada de previdência complementar (EFPC) devem ser classificados nos seguintes segmentos de aplicação: renda fixa; renda variável; investimentos estruturados; investimentos no exterior; imóveis; e operações com participantes (que abrange empréstimos e financiamento habitacional).

O segmento de aplicação em empréstimos é uma importante fonte de geração de recursos. No entanto, existem riscos de que o valor concedido não seja devolvido, gerando inadimplência. Essa, por sua vez, afeta a rentabilidade das instituições.

Existem diversos motivos que levam uma pessoa a ficar inadimplente, como o desemprego e até mesmo a extrapolação do consumo além de sua renda. Estudar essas variáveis que afetam o não pagamento da dívida é relevante para ajudar o credor a se antecipar e se proteger dos prejuízos. Nesse sentido, a estatística é uma ciência bastante utilizada para obter maiores informações sobre os dados.

Modelar, por exemplo, o tempo que o indivíduo leva até ficar inadimplente e observar as covariáveis que o fazem atingir o evento de interesse é importante para traçar o perfil de seus clientes. A técnica de Análise de Sobrevivência pode propiciar a modelagem adequada para essa situação.

Neste contexto, o objetivo deste trabalho foi ajustar um modelo para estimar o tempo até que o indivíduo se torne inadimplente. Mais especificamente, o objetivo principal foi ajustar um modelo de riscos competitivos considerando duas causas de interesse: inadimplência (principal evento de interesse) e quitação da dívida (cura). Este trabalho também propôs um escore que mede o risco de inadimplência de um cliente. Esse escore é calculado a partir das estimativas do modelo ajustado e pode ser utilizado para classificar os clientes em bons ou mal pagadores. A metodologia proposta foi aplicada a um conjunto de dados reais sobre participantes que realizaram um empréstimo na Funda-

ção dos Economiários Federais (FUNCEF). A análise considerou os modelos de regressão Weibull, log-normal e log-logística como possíveis modelos para representar o tempo até a inadimplência e quitação.

Este trabalho está organizado da seguinte forma: no Capítulo 2 será apresentado a revisão bibliográfica; no Capítulo 3 conterá uma análise descritiva do problema e dos dados; o Capítulo 4 irá apresentar o modelo de regressão ajustado para os dados e o escore de risco de crédito; o Capítulo 5 será um espaço para as considerações finais e; por fim, será disponibilizado a bibliografia.

Capítulo 2

Revisão Bibliográfica

2.1 Conceitos básicos de Análise de Sobrevivência

A Análise de Sobrevivência é uma técnica estatística que estuda dados relacionados ao tempo decorrido até a ocorrência do evento de interesse, o qual precisa ser claramente definido. Esse tempo é a variável resposta do estudo.

Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e, frenquentemente, pelas censuras (Colosimo e Giolo, 2006). O tempo até a ocorrência de um evento de interesse é denominado tempo de falha (ou tempo de sobrevivência). E, segundo Garcia (2013) a presença de observações incompletas ou parciais são denominadas censura.

Para analisar o tempo de falha é necessário:

- fixar o tempo de início de estudo;
- definir a escala de medida a ser utilizada;
- e estabelecer o evento de interesse, que frequentemente é indesejável e conhecido como falha.

A presença de observações incompletas ou parciais, por sua vez, são denominadas censura (Garcia, 2013).

A censura ocorre quando há perda de informação decorrente de não se ter observado o momento exato da ocorrência do desfecho, resultando em observações parciais ou incompletas. Ou seja, o evento de interesse não pode ser observado por algum motivo. As observações censuradas devem ser incluídas na análise dos dados, pois mesmo sendo incompletas, essas observações fornecem informações sobre o tempo de vida de objetos e indivíduos e a omissão das censuras pode acarretar em conclusões viciadas na análise estatística. Dessa forma, existe a necessidade da introdução de uma variável extra na análise, que indica se o valor do tempo de sobrevivência de um determinado indivíduo foi ou não observado. Essa variável, geralmente representada por δ, é conhecida como variável indicadora de censura e é expressa por:

$$\delta = \begin{cases} 1, \text{ se o indivíduo falhou} \\ 0, \text{ se o indivíduo foi censurado} \end{cases}$$

Em cada estudo a censura pode ocorrer por motivos diferentes. Por isso, existem formas distintas de censura. Elas podem ser censura à direita, censura à esquerda e censura intervalar.

Censura à direita

A censura à direita é caracterizada pelo tempo de ocorrência do evento estar à direita do tempo registrado, ou seja, do tempo de início do estudo. Essa, por sua vez, abrange alguns mecanismos de censura:

1. Censura tipo I: o estudo será terminado após um período de tempo fixo, t_f, e nesse ínterim de tempo uma ou mais observações em estudo não falham. O tempo de início t_0 e o tempo final do estudo t_f devem ser determinado antes do início do estudo;

2. Censura tipo II: o estudo terminará após ocorrer um número fixo k $(k \leq n)$ de falhas. Esse número deve ser determinado antes do início do estudo;

3. Censura aleatória: são todos os casos em que as observações não experimentam o evento de interesse por motivos não controláveis.

A Figura 2.1 ilustra os mecanismos de censura à direita. A falha é representada por "•"e a censura é representada "∘".

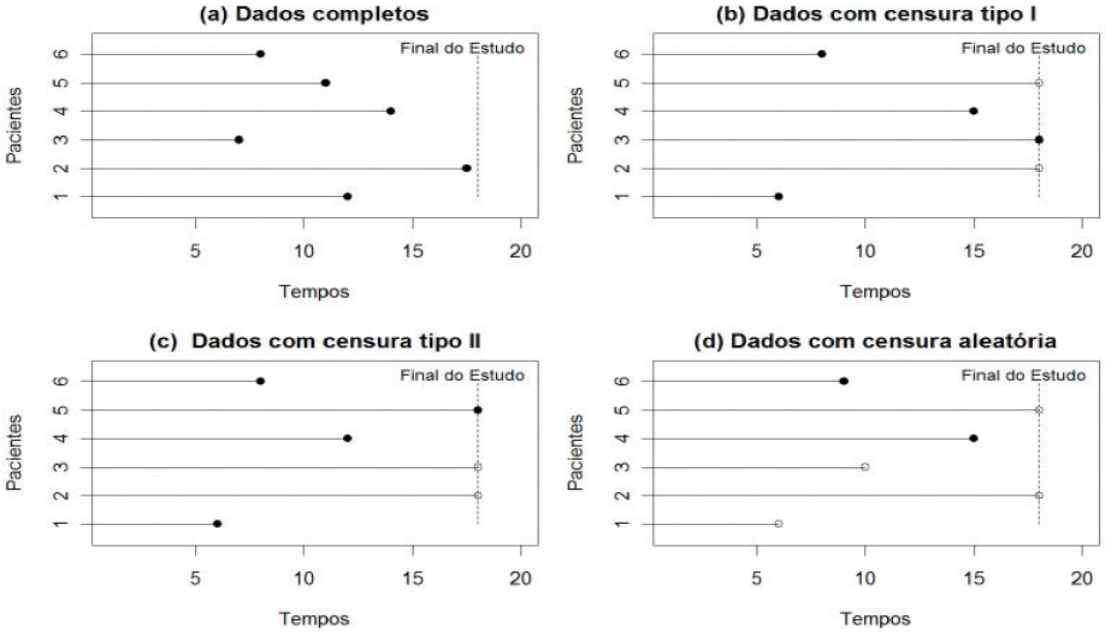

Figura 2.1: Tipos de mecanismos de censura à direita. (Colosimo e Giolo, 2006).

A análise de dados deste trabalho irá considerar a censura à direita.

Censura à esquerda

A censura à esquerda acontece quando o evento de interesse já ocorreu quando o indivíduo começou a ser observado, ou seja, o evento de interesse ocorreu antes do estudo começar.

Censura intervalar

A censura intervalar ocorre em estudos nos quais os indivíduos são acompanhados periodicamente. O evento de interesse ocorre em um intervalo de tempo. Assim, o tempo de falha não é conhecido precisamente, mas pertence a um determinado intervalo.

2.1.1 Descrição do tempo de sobrevivência

Os dados de sobrevivência são usualmente representados pela variável aleatória não negativa T, geralmente contínua. Essa variável representa o tempo de sobrevivência do indivíduo. Algumas maneiras de especificá-la são por meio da função densidade de probabilidade, $f(t)$, função de sobrevivência, $S(t)$; função de risco, $h(t)$; e por meio de relações matemáticas existentes entre essas três funções. Trabalhos que tratam de dados de sobrevivência discretos podem ser vistos em Nakano e Carrasco (2006), Carrasco et al. (2012) e Brunello e Nakano (2015).

A variável indicadora de falha ou censura, δ_i, e o tempo de falha, t_i, representam os dados de sobrevivência para o indivíduo i (i = *1,...,n*). Se houver um vetor de covariáveis x_i, com i= *1,...,n* , os dados de sobrevivência são representados por (t_i, δ_i, x_i).

Função Densidade de Probabilidade

A função densidade de probabilidade é definida como o limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo [T_i, T_i+Δ_t] (Colosimo e Giolo, 2006). Essa função é expressa por:

$$f(t) = \lim_{\Delta_t \to 0} \Big(\frac{P(T_i \leq T \leq T_i + \Delta_t)}{\Delta_i} \Big),$$

$f(t) \geq 0 \; \forall \; t$, e a área abaixo de $f(t)$ é igual a 1.

Função de Distribuição

A função de distribuição também é conhecida como função de distribuição acumulada, pois acumula as probabilidades dos valores inferiores ou iguais a t. Essa função

comporta muitas informações sobre a variável aleatória estudada (Magalhães, 2006). Sua expressão matemática é dada por:

$$F(t) = P(T \in (-\infty, t]) = P(T \leq t),$$

no qual, $t \in \mathbb{R}$.

$F(t)$ possui as seguintes propriedades:

1. $\lim_{t \to -\infty} F(t) = 0$;
2. $\lim_{t \to \infty} F(t) = 1$;
3. $F(t)$ é contínua à direita;
4. $F(t)$ é não decrescente, ou seja, $F(t) \leq F(y)$, sempre que $t \leq y$, $\forall\ t, y \in \mathbb{R}$.

Função de Sobrevivência

A função de sobrevivência, $S(t)$, é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, é a probabilidade de uma observação sobreviver ao tempo t (Colosimo e Giolo, 2006). Em termos probabilísticos essa função é expressa por:

$$S(t) = P(T \geq t),$$

na qual, $S(t)$ é uma função monótona decrescente e contínua. Essa função também pode ser definida em termos da função de distribuição acumulada por: $S(t) = 1 - F(t)$.

Função de Risco

A função de risco, também chamada de Função de Taxa de Falha, é caracterizada pelo limite da probabilidade da falha ocorrer no intervalo de tempo $[t,t + \Delta_t)$, condicional à sobrevivência até o tempo t, dividida pelo comprimento do intervalo de tempo Δ_t (Colosimo e Giolo, 2006). Essa expressão assume a seguinte forma:

$$h(t) = \lim_{\Delta_t \to 0} \Big(\frac{P(t \leq T < t + \Delta_t | T \geq t)}{\Delta_t} \Big),$$

A função $h(t)$ pode assumir a forma crescente, constante ou decrescente. Além disso, pode assumir a forma unimodal ou a forma de curva da banheira.

2.1.2 Função de Risco Acumulada

A função de risco acumulada determina a taxa de falha que é acumulada até o tempo t, e pode ser usada para obter $h(t)$ na estimação não paramétrica (Colosimo e Giolo, 2006). Sua expressão é dada por:

$$H(t) = \int_0^t h(u)du.$$

2.1.3 Relações entre as funções

Segundo Colosimo e Giolo (2006), para uma variável aleatória T contínua e não negativa, podem ser definidas relações relevantes entre elas a partir das funções definidas anteriormente. Algumas relações importantes são:

$$f(t) = \frac{d}{dt}F(t),$$

$$F(t) = 1 - S(t),$$

$$f(t) = \frac{d}{dt}(1 - S(t)) = -\frac{d}{dt}S(t), \tag{2.1}$$

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.2}$$

$$h(t) = \frac{-\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\log[S(t)], \tag{2.3}$$

$$\log[S(t)] = -\int_0^t h(u)du \qquad \text{e} \tag{2.4}$$

$$S(t) = \exp[-H(t)]. \tag{2.5}$$

2.1.4 Estimação não paramétrica da função de sobrevivência

Observações censuradas são um problema para as técnicas convencionais de análise estatística descritiva, que consiste principalmente em encontrar medidas de tendência central e variabilidade, pois a interpretação dos resultados se tornam mais difícil. Sendo assim, o principal componente da análise descritiva, envolvendo dados de tempo de vida, é a função de sobrevivência $S(t)$ (Colosimo e Giolo, 2006). Pode-se estimá-la a partir do estimador não paramétrico de Kaplan-Meier (Kaplan e Meier, 1958).

Estimador de Kaplan-Meier

O estimador não paramétrico de Kaplan-Meier (Kaplan e Meier, 1958), também chamado de estimador limite-produto, é uma adaptação da função de sobrevivência (Colosimo e Giolo, 2006).

O estimador de Kaplan-Meier é definido como:

$$\hat{S} = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \Big(1 - \frac{d_i}{n_i}\Big), \tag{2.6}$$

no qual, $t_1 <...< t_k$ são os k valores distintos dos tempos de falha; d_i é o número de falhas no instante t_i, i= 1,...,k; e n_i é o número de indivíduos sob risco (sobreviventes e não censurados) em até t_i.

As propriedades principais desse estimador são:

- não viciado para grandes amostras;
- fracamente consistente;
- converge assintoticamente para um processo gaussiano;
- estimador de máxima verossimilhança de $S(t)$.

A variância assintótica do estimador de Kaplan-Meier é dada por:

$$\widehat{Var\hat{S}(t)} = [\hat{S}(t)]^2 \sum_{i:t_i<t} \frac{d_i}{n_i(n_i - d_i)}. \tag{2.7}$$

Levando em consideração que $\hat{S}(t)$ tem distribuição assintótica normal, para t fixo, um intervalo aproximado de $100(1-\alpha)\%$ de confiança para $\hat{S}(t)$ é dado por:

$$\hat{S}(t) \pm Z_{\alpha/2} \sqrt{\widehat{Var(\hat{S}(t))}}, \tag{2.8}$$

no qual, $Z_{\alpha/2}$ indica o $\alpha/2$ quantil da distribuição Normal padrão.

2.1.5 Determinação Empírica da Função de Risco

A forma da função de risco será identificada por meio do método gráfico baseado no tempo total em teste (TTT) (Barlow e Campo, 1975). O intuito é encontrar o modelo adequado para os tempos de vida. Contudo, esse método só será eficiente se não houver censuras, ou se existir poucas, pois a curva do gráfico não as leva em consideração. Nesse contexto, seu uso deve ser feito com cautela, principalmente quando a proporção de observações censuradas é alta.

A expressão empírica do TTT (Aarset, 1985) é a seguinte:

$$G(r/n) = \frac{(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}}{\sum_{i=1}^{r} T_{i:n}} \tag{2.9}$$

no qual, r = 1,...,n e $T_{i:n}$, com i= 1,...,n, representa as estatísticas de ordem dos tempos.

A Figura 2.2 mostra várias formas que a curva TTT pode assumir.

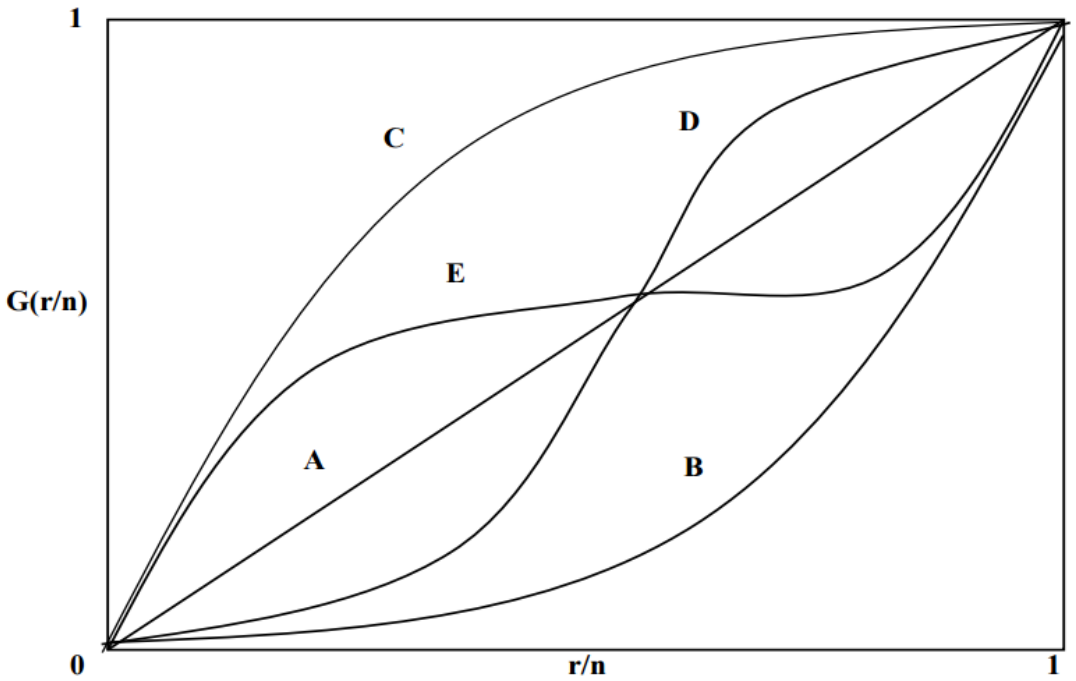

Figura 2.2: Formas da curva do Tempo Total em Teste - TTT

Se a curva TTT assume a forma de uma reta, como no modelo A, é adequado trabalhar com uma função de risco constante. Nesse caso, a distribuição exponencial é indicada, pois sua função de risco é constante para todo o tempo observado.

Quando a curva é convexa, modelo B, ou côncava, modelo C, a função de risco é monotonamente decrescente ou crescente, respectivamente. Assim, a distribuição a ser utilizada será a Weibull, pois apresenta função de risco decrescente quando o parâmetro de forma γ é menor que 1 e crescente quando γ é maior que 1.

Caso a curva apresente uma forma convexa e depois côncava, como na curva D, trata-se de uma função de risco que toma forma de banheira. As distribuições referentes à essa função de risco indicadas para modelagem dos dados são: Weibull exponencializada, Weibull modificada, Weibull aditiva, Burr XII Aditiva, beta Weibull generalizada, entre outras (Fachine-Gomes, 2015).

Se a curva for côncava e em seguida convexa, a função de risco possui característica unimodal. Dessa forma é adequado utilizar distribuições como a log-normal e a log-logística.

2.1.6 Principais modelos paramétricos em Análise de Sobrevivência

Os modelos paramétricos são utilizados para modelar o tempo de sobrevivência até a ocorrência do evento de interesse de forma mais fidedigna (Costa, 2013). Alguns modelos recebem notoriedade por serem adequados à várias situações. Alguns desses modelos que serão utilizados neste trabalho são as distribuições Weibull, exponencial, log-normal e log-logística.

Distribuição Weibull

A distribuição Weibull é popular por apresentar uma grande variedade de formas. A propriedade básica de todas essas formas é que a função de risco é monótona, ou seja, ela é crescente, decrescente ou constante (Colosimo e Giolo, 2006).

A variável aleatória tempo de falha, T, descrito pela distribuição Weibull, tem a seguinte função de densidade de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)^{\gamma} \Big\}, \tag{2.10}$$

no qual, $t \geq 0$, o parâmetro de escala, $\alpha > 0$, tem a mesma unidade de medida de t e o parâmetro de forma,$\gamma > 0$, não tem unidade (Colosimo e Giolo, 2006).

A função de sobrevivência da distribuição Weibull é dada por:

$$S(t) = \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)^{\gamma} \Big\}, t \geq 0. \tag{2.11}$$

E a função de risco é dada por:

$$h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}, t \geq 0. \tag{2.12}$$

A função de risco, $h(t)$, é estritamente crescente quando $\gamma > 1$, estritamente decrescente quado $\gamma < 1$ e constante quando $\gamma = 1$.

Segundo Santos (2017), para verificar a influência das variáveis explicativas no tempo de sobrevivência é necessário utilizar modelos de regressão. Seja $\mathbf{x}^T = (1, x_1, ..., x_p)$ um vetor de covariáveis dos indivíduos em estudo, utiliza-se então uma função de ligação $g(\cdot)$ que conecta a variável resposta às variáveis explicativas. Para um conjunto de p covariáveis, o vetor de parâmetros $\boldsymbol{\theta}$ que será estimado utilizando o vetor $\mathbf{x}$, passa a ser definido como:

$$\boldsymbol{\theta} = q(\boldsymbol{\eta}), \tag{2.13}$$

em que $\boldsymbol{\eta} = \mathbf{x}^T\boldsymbol{\beta}$ é o preditor linear e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor dos coeficientes de regressão.

Para construir o modelo de regressão Weibull, será utilizada a função de ligação logarítmica, e as covariáveis serão inseridas no modelo pela seguinte relação:

$$\alpha = \exp(\mathbf{x}^T\boldsymbol{\beta}). \tag{2.14}$$

Dessa forma, a função densidade de probabilidade e a função de sobrevivência do modelo de regressão Weibull são definidas, respectivamente, por:

$$f(t|x) = \frac{\gamma}{\exp(\mathbf{x}^T\boldsymbol{\beta})^\gamma} t^{\gamma-1} \exp\Big\{ - \Big(\frac{t}{\exp(\mathbf{x}^T\boldsymbol{\beta})}\Big)^\gamma \Big\} \tag{2.15}$$

e

$$S(t) = \exp\Big\{ - \Big(\frac{t}{\exp(\mathbf{x}^T\boldsymbol{\beta})}\Big)^\gamma \Big\},\ t \geq 0. \tag{2.16}$$

Distribuição Exponencial

A distribuição exponencial é um caso particular da distribuição Weibull. Esse modelo é um dos mais simples usados para descrever o tempo de falha. Apresenta apenas um parâmetro e o único que tem função de risco constante (propriedade de falta de memória). Por esse motivo, é indicado que essa distribuição seja usada quando o tempo é curto (Colosimo e Giolo, 2006).

Então, a variável aleatória tempo de falha T, descrito pela distribuição exponencial, tem a seguinte função de densidade de probabilidade:

$$f(t) = \frac{1}{\alpha} \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)\Big\}, \tag{2.17}$$

no qual, $t \geq 0$ e o parâmetro $\alpha > 0$ é o tempo médio de vida e tem a mesma unidade do tempo de falha T.

Além disso, a função de sobrevivência $S(t)$ do modelo em questão é dada por:

$$S(t) = \exp\Big\{ - \Big(\frac{t}{\alpha}\Big)\Big\},\ t \geq 0. \tag{2.18}$$

E a função de risco é dada por:

$$h(t) = \frac{1}{\alpha},\ t \geq 0. \tag{2.19}$$

O modelo de regressão exponencial, assim como foi apresentado acima para o modelo de regressão Weibull, é construído a partir das covariáveis que são alocadas no parâmetro de escala α, que será apresentado a seguir.

$$\alpha = \exp(\mathbf{x}^T\boldsymbol{\beta}). \tag{2.20}$$

Sendo assim, a função densidade de probabilidade e a função de sobrevivência do modelo de regressão exponencial são definidas, respectivamente, por:

$$f(t) = \frac{1}{\exp(\mathbf{x}^T\boldsymbol{\beta})} \exp\Big\{ - \Big(\frac{t}{\exp(\mathbf{x}^T\boldsymbol{\beta})}\Big)\Big\} \tag{2.21}$$

e

$$S(t) = \exp\Big\{ - \Big(\frac{t}{\exp(\mathbf{x}^T\boldsymbol{\beta})}\Big)\Big\},\ t > 0. \tag{2.22}$$

Distribuição log-normal

A distribuição log-normal, assim como a distribuição Weibull, é útil para caracterizar tempos de vida de indivíduos e produtos. A variável aleatória tempo de falha T, com distribuição log-normal, tem a seguinte função densidade (Colosimo e Giolo, 2006):

$$f(t) = \frac{1}{t\sigma\ \sqrt{2\pi}} \exp\Big\{ - \frac{1}{2}\Big(\frac{\log(t) - \mu}{\sigma}\Big)^2\Big\},\ t \geq 0. \tag{2.23}$$

Aqui, $-\infty \leq \mu \leq \infty$ é a média do logaritmo do tempo de falha T, e $\sigma > 0$ é o desvio padrão.

Os dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, considerando o logaritmo dos dados em vez dos valores originais, isto é, $\log(T) \sim \text{Normal}(\mu, \sigma^2)$.

Uma variável log-normal não apresenta uma forma analítica explícita para as funções de sobrevivência e de função de risco, sendo representadas, respectivamente, da seguinte maneira:

$$S(t) = \Phi\Big(\frac{-\log(t) + \mu}{\sigma}\Big) \tag{2.24}$$

e

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.25}$$

no qual, $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

O modelo de regressão log-normal é construído a partir das covariáveis que influenciam no tempo de sobrevida dos indivíduos. Dessa forma, as covariáveis são alocadas no parâmetro μ, conforme apresentado a seguir.

$$\mu = \mathbf{x}^T\boldsymbol{\beta}. \tag{2.26}$$

Portanto, a função densidade de probabilidade e a função de sobrevivência do modelo de

regressão log-normal são definidas, respectivamente, por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2}\Big(\frac{\log(t) - (\mathbf{x}^T\boldsymbol{\beta})}{\sigma}\Big)^2\Big\},\ t \geq 0 \tag{2.27}$$

e

$$S(t) = \Phi\Big(\frac{-\log(t) + (\mathbf{x}^T\boldsymbol{\beta})}{\sigma}\Big),\ t > 0. \tag{2.28}$$

Distribuição log-logística

A variável aleatória tempo de falha T cuja distribuição é dada pelo modelo log-logístico, com parâmetros α e γ, tem a seguinte função de densidade (Colosimo e Giolo, 2006):

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}(1 + (t/\alpha))^{-2},\ \ t > 0. \tag{2.29}$$

Aqui, α>0 e γ>0 são os parâmetros de escala e forma da distribuição, respectivamente.

A função de sobrevivência $S(t)$ e a função de risco $h(t)$ são expressas da seguinte maneira:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma},\ \ t > 0 \tag{2.30}$$

e

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^\gamma]},\ \ t > 0. \tag{2.31}$$

O modelo de regressão log-logístico também é construído a partir das covariáveis, que são importantes para o tempo de sobrevida dos indivíduos. Elas são alocadas no parâmetro de escala α (que representa a mediana da distribuição), conforme apresentado a seguir:

$$\alpha = \exp(\mathbf{x}^T\boldsymbol{\beta}). \tag{2.32}$$

Sendo assim, a função densidade de probabilidade e a função de sobrevivência do modelo de regressão log-logística são definidas, respectivamente, por:

$$f(t) = \frac{\gamma}{\exp(\mathbf{x}^T\boldsymbol{\beta})^\gamma} t^{\gamma-1}(1 + (t/\exp(\mathbf{x}^T\boldsymbol{\beta})))^{-2},\ \ t > 0 \tag{2.33}$$

e

$$S(t) = \exp\Big\{ -\Big(\frac{t}{\exp(\mathbf{x}^T\boldsymbol{\beta})}\Big)^\gamma\Big\}, t > 0. \tag{2.34}$$

2.1.7 Estimação dos parâmetros dos modelos

Os estudos que envolvem tempos de falha são modelados por distribuições. Os parâmetros dessas distribuições devem ser estimados por meio das observações amostrais.

Assim, é possível responder às perguntas de interesse.

Um dos métodos mais conhecidos de estimação, dentro do contexto de regressão linear, é o método de mínimos quadrados. Contudo, usá-lo na análise de estudos de tempo de vida é inapropriado, pois não é possível incorporar censura em seu processo.

Outro método relevante de estimação na literatura é o de máxima verossimilhança. Esse, por sua vez, é mais apropriado para estudos em Análise de Sobrevivência, pois é possível incorporar censuras a ele. Além disso, é relativamente trivial de ser compreendido e possui ótimas propriedades para grandes amostras.

Método de Máxima Verossimilhança

O estimador de máxima verossimilhança, dado uma distribuição, escolhe valores para os parâmetros que melhor explique a amostra observada. Dessa forma, o objetivo é encontrar o valor do vetor de parâmetros $\boldsymbol{\theta}$ que maximiza a função de verossimilhança, $L(\boldsymbol{\theta})$, a qual é expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}), \tag{2.35}$$

em que t_1,...,t_n representam os tempos observados de uma amostra de tamanho n de uma população. E, tem-se que $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$, ou seja $\boldsymbol{\theta}$ pode ser um parâmetro ou um conjunto de parâmetros.

A função de verossimilhança mostra que cada observação não censurada contribui com sua função densidade $f(t)$. Por outro lado, cada informação censurada à direita contribui informando que o tempo de falha é maior que o tempo de censura observado, ou seja, contribuem com sua função de sobrevivência $S(t)$.

Dessa forma, a função de verossimilhança considerando a censura à direita, que é a censura considerada neste trabalho, é dada pela seguinte expressão (Colosimo e Giolo, 2006):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [f(t_i; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{\theta})]^{1-\delta_i}, \tag{2.36}$$

no qual, $\boldsymbol{\theta}$ é o vetor de parâmetros a ser estimado e δ_i é a variável indicadora de falha ou censura.

Contudo, é sempre conveniente trabalhar com o logaritmo da função de verossimilhança. Segundo Colosimo e Giolo (2006), os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou equivalentemente o logaritmo de $L(\boldsymbol{\theta})$, isto é, $\log(L(\boldsymbol{\theta}))$. Esses estimadores são encontrados resolvendo-se o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0, \tag{2.37}$$

no qual, $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$ é dado por:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log[f(t_i;\boldsymbol{\theta})] + (1-\delta_i)\log[S(t_i;\boldsymbol{\theta})]. \tag{2.38}$$

Teste da Razão de Verossimilhanças

Geralmente, há interesse em testar hipóteses relacionadas ao vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$, ou relacionada a um subconjunto desse vetor. Um teste bastante utilizado para isso é o teste da razão de verossimilhanças. A estatística do teste é dada pela razão dos valores do logaritmo da função de verossimilhança maximizada restringido aos valores definido na hipótese nula, H_0 e pelo logaritmo da função de verossimilhança maximizada sem restrição. De uma maneira mais matemática, temos a representação das hipóteses em teste e a expressão da estatística do teste, respectivamente:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0} \qquad vs \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0}, \tag{2.39}$$

$$TRV = -2\log\Big[\frac{L(\boldsymbol{\theta_0})}{L(\boldsymbol{\theta})}\Big]. \tag{2.40}$$

A estatística TRV segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade, χ^2_p. Então, H_0 é rejeitada, ao nível de $100\alpha\%$ de significância se $TRV > \chi^2_{p,1-\alpha}$.

2.1.8 Adequação do Modelo Ajustado

É importante avaliar se o modelo proposto está bem ajustado aos dados. Isso pode ser feito por meio de técnicas gráficas. Essas técnicas avaliam a distribuição dos erros e sua principal utilidade é rejeitar modelos inapropriados, ou seja, o interesse não é aprovar um modelo particular, até porque, em muitos casos, existem mais de um modelo que podem ser utilizados.

Em Análise de Sobrevivência, os resíduos de Cox-Snell, martingal e deviance são os mais utilizados.

Resíduos de Cox-Snell

Segundo Colosimo e Giolo (2006), os resíduos de Cox-Snell (1968), são úteis para examinar o ajuste global do modelo. Esses resíduos são determinados por:

$$\hat{e_i} = \hat{H}(t_i), \tag{2.41}$$

no qual, $\hat{H}(t_i)$ é a função de risco acumulado obtido do modelo ajustado.

Os resíduos $\hat{e_i}$ vêm de uma população homogênea e, se o modelo for adequado, devem seguir uma distribuição exponencial padrão (Lawless, 2003). Para que o modelo em estudo seja adequado, o gráfico de $\hat{e_i}$ *versus* $\hat{H}(\hat{e_i})$ deve ser aproximadamente uma reta. O gráfico das curvas de sobrevivência dos resíduos, $\hat{S}(\hat{e_i})$, e pelo modelo exponencial padrão, $\exp(-\hat{e_i})$, também auxiliam na verificação da qualidade do modelo ajustado, quanto mais próximas, o ajuste do modelo aos dados é melhor.

2.2 Modelo de riscos competitivos

Em determinadas situações há interesse em modelar o tempo de vida de um indivíduo que está suscetível a vários eventos mutuamente exclusivos, decorrentes de um mesmo fator de risco. Assim, a falha desse indivíduo pode ocorrer por m causas distintas. E, quando há várias causas de falha competindo para que ocorra o evento de interesse, na literatura essas causas são definidas como riscos competitivos.

Os dados, na presença de riscos competitivos são representados pela terna (y, c_i, δ), em que,

- $Y = \min(T_1, ..., T_m)$, supondo que T_i é o tempo até a falha devido à causa c_i, $i =$ 1, 2, ...,m. Ou seja, Y é o tempo até a primeira falha;
- c_i é a causa de falha i, $i = 1, 2, ..., m$ e;
- δ representa a censura. Sejam $L_1, L_2, ..., L_m$ os tempos de censuras associadas às m causas de morte, então

$$\delta = \begin{cases} 1, \text{ se o } Y \text{ é observado;} \\ 0, \text{ se o } L = min(L_1, ..., L_m) \text{ é observado.} \end{cases} \tag{2.42}$$

Uma das abordagens que geralmente são utilizadas na análise de riscos competitivos, e que será apresentada no decorrer deste estudo, é a do tempo de falha latente (Nakano, 2003).

2.2.1 Abordagem latente de riscos competitivos

Em uma abordagem de riscos competitivos com tempos de falha latentes, os indivíduos estão sujeitos a m causas de morte e o vetor de variáveis aleatórias $T = (T_1, T_2, ..., T_m)$ representa os tempos de falha devido a cada uma das m causas c_i, $i = 1, 2, ..., m$. Dessa forma, observa-se o mínimo entre os tempos de falha ($Y = min(T_1, ..., T_m)$) e sua respectiva causa de falha (Y,C_i).

A função de sobrevivência conjunta é dada por:

$$S(t_1, t_2, ..., t_m) = P(T_1 \geq t_1, T_2 \geq t_2, ..., T_m \geq t_m), \tag{2.43}$$

em que, $S(0, 0, ..., 0) = 1$ e $S(\infty, \infty, ..., \infty) = 0$. $S(\cdot)$ é contínua à direita e não decrescente em cada argumento (Nakano, 2003).

E a função de sobrevivência total é dada por:

$$S_Y(t) = P[Y \geq t] = P[min(T_1, T_2, ..., T_m) \geq t] = P[T_1 \geq t, T_2 \geq t, ..., T_m \geq t]. \tag{2.44}$$

Essa formulação se aplica se os tempos de morte T_1, T_2, ..., T_m são estatisticamente independentes ou não. A suposição básica de riscos competitivos é que a eliminação da causa c_i não afeta $S(\cdot)$ (Nakano, 2003).

A função marginal de $S(\cdot)$ é denotada por S_i e ocorre apenas na presença da causa c_i. Segundo Nakano (2003), uma dificuldade presente na análise de riscos competitivos deve-se ao fato de $S(\cdot)$ não poder ser estimada empiricamente, visto que T não é observável. Assim, a análise depende de suposições não verificáveis sobre a estrutura de $S(\cdot)$.

A função de subdensidade da causa i é dada por:

$$f_i(t) = \Big(- \frac{\partial S(t_1, t_2, ..., t_m)}{\partial t_i} \Big)_{t_1 = t_2 = ... = t_m = t}. \tag{2.45}$$

A partir das funções de sobrevivência $S_i(t)$, $S_Y(t)$ e $S(\cdot)$, pode-se observar relações que estão associadas a vários riscos. Algumas relações importantes serão mostradas a seguir.

$$h(t) = \frac{\Big(- \frac{\partial S_Y(t)}{\partial t} \Big)}{S_Y(t)}. \tag{2.46}$$

Essa é a função de risco total que expressa a probabilidade de falha por alguma causa c_i em $(t, t+\Delta t)$, dado que sobreviveu de todos os c_i até t (P(falha por alguma causa c_i em $(t, t + \Delta t)|$ sobreviveu de todos os c_i até t)).

$$h_i(t) = \frac{\Big(- \frac{\partial S_i(t)}{\partial t} \Big)}{S_i(t)} = \frac{\Big(- \frac{\partial S(t_i, t_j)}{\partial t_i} \Big)_{t_i = t, t_j = 0}}{S(t, 0)}, i = 1, ..., m; j = 1, ..., m; i \neq j. \tag{2.47}$$

A função de risco marginal é definida pela probabilidade de falhar pela causa c_i em $(t, t + \Delta t)$, quando c_i é o único risco atuando em $t + \Delta t$, dado que sobreviveu de c_i até t (P(falha pela causa c_i em $(t, t + \Delta t)$ quando c_i é o único risco atuando em $t + \Delta t|$

sobreviveu de todos os c_i até t)).

$$g_i(t) = \frac{-(\frac{\partial S_i(t_i)}{\partial t_i})_{t_i=t}}{S_i(t)} = \frac{\Big(-\frac{\partial S_i(t_1,t_2,...,t_m)}{\partial t_i}\Big)_{t_1=t_2=...=t_m=t}}{S(t,t,...,t)}. \tag{2.48}$$

Essa é a probabilidade de falha pela causa c_i em (t, t + Δt), quando todos os riscos estão atuando em (t, t + Δt), dado que sobreviveu de todos os c_i até t (P(falha pela causa c_i em (t, t + Δt) quando c_i é o único risco atuando em (t, t + Δt)| sobreviveu de todos os c_i até t)).

$$G_i(t) = \exp\Big[\int_0^t g_i(u)du\Big]. \tag{2.49}$$

Note que $G_i(t)$ não é uma função de sobrevivência marginal genuína. Definida a partir de $g_i(t)$, $G_i(t)$ é denotada como uma subfunção de sobrevivência, ou seja a probabilidade de um indivíduo sobreviver à causa c_i na presença de todas as outras causas (Nakano, 2003).

A função de sobrevivência total também pode ser dada por:

$$S_Y(t) = \exp\Big[-\int_0^t h(u)du\Big] = \exp\Big[-\int_0^t \sum_{i=1}^m g_i(u)du\Big] = \prod_{i=1}^m \exp\Big[-\int_0^t g_i(u)du\Big]. \tag{2.50}$$

A partir da equação 2.58, obtém-se:

$$S_Y(t) = \prod_{i=1}^m G_i(t). \tag{2.51}$$

Se $g_i(t) = h_i(t)$, tem-se que $h(t) = \sum_{i=1}^m g_i(t) = \sum_{i=1}^m h_i(t)$. Consequentemente, a função de sobrevivência total pode ser dada por:

$$S_Y(t) = \prod_{i=1}^m G_i(t) = \prod_{i=1}^m S_i(t). \tag{2.52}$$

Quando os $T_i's$ são independentes, tem-se que $g_i(t) = h_i(t)$. Porém, a recíproca não é verdadeira. Além disso, se os $T_i's$ são independentes $S(t_1,t_2,...,t_m) = S(t_1)S(t_2)...S(t_m)$, que implica em $S_Y(t) = S(t,t,...,t) = S_1(t)S_2(t)...S_m(t)$. Contudo, a recíproca não é verdadeira.

Os $h_i(t)$ são denotados como função risco "net", também chamados de função de risco específico. Os $g_i(t)$, definidos como função risco "crude", são os riscos de morte devido à causa c_i, quando todas as causas estão atuando sobre o indivíduo (Nakano, 2003).

A distribuição de probabilidade de Y na presença de todas as causas de falha

é dada por:

$$Q_i(t) = P[Y \leq t, c_i = i] = \int_0^t f_i(u)du. \tag{2.53}$$

E,

$$f_i(t) = g_i(t)S_Y(t). \tag{2.54}$$

Logo, a função densidade de Y condicionado à causa de falha c_i é dado por:

$$f_Y(t|c_i) = \frac{1}{\pi_i} g_i(t)S_Y(t), \tag{2.55}$$

onde, $\pi_i = P[c_i = i] = \int_0^t g_i(u)S_Y(u)du$ é a probabilidade de falha pela causa c_i (Nakano, 2003).

Identificabilidade em Riscos Competitivos

Há, em geral, o interesse em identificar as distribuições de sobrevivência conjunta e marginais de $(T_1, T_2, ..., T_m)$ dada a distribuição de (Y, c_i) (a distribuição do mínimo identificado) (Nakano, 2003). Porém, nem sempre a distribuição conjunta pode ser definida a partir das funções marginais de $(T_1, T_2, ..., T_m)$. Esse fato é conhecido por problema de não identificabilidade, observado por Cox em 1959. Para o caso de duas causas de falha, é estudado por Tsiatis em 1975 (Assane, 2013). No entanto, Berman (1963) mostrou que se as causas de falhas atuam de forma independente, a distribuição de (Y,c_i) determina unicamente a distribuição conjunta.

Função de Verossimilhança

A função de verossimilhança em riscos competitivos é construída a partir da função de distribuição $f_i(t_{ij})$ e dos indivíduos censurados, por meio de $S_Y(t_{wj})$.

$$L = \Big\{ \prod_{i=1}^{m} \prod_{j=1}^{D_i} f_i(t_{i_j}) \Big\} \Big\{ \prod_{j=1}^{W} S_Y(t_{wj}) \Big\}. \tag{2.56}$$

$$L = \Big\{ \prod_{i=1}^{m} \prod_{j=1}^{D_i} g_i(t_{i_j})S_Y(t_{ij}) \Big\} \Big\{ \prod_{j=1}^{W} S_Y(t_{wj}) \Big\}. \tag{2.57}$$

$$L = \Big\{ \prod_{i=1}^{m} \prod_{j=1}^{D_i} \Big[g_i(t_{i_j}) \prod_{i=1}^{m} G_m(t_{i_j}) \Big] \Big\} \Big\{ \prod_{j=1}^{W} \prod_{i=1}^{m} G_m(t_{i_j}) \Big\}, \tag{2.58}$$

no qual, D_i é o número de indivíduos que morreram pela causa c_i, W é o número de indivíduos censurados, t_{i_j} é o tempo até a falha dada a causa c_{i_j}, w_j é o j-ésimo indivíduo censurado e t_{w_j} é o tempo de censura do j-ésimo indivíduo.

A função de verossimilhança pode ser reescrita como:

$$L = \prod_{i=1}^{m} \prod_{j=1}^{D_i} g_i(t_{i_j}) \prod_{l=1}^{n} G_i(t_l). \tag{2.59}$$

A expressão acima pode ser reescrita como:

$$L = \prod_{i=1}^{m} \Big\{ \prod_{j=1}^{D_i} \Big[g_i(t_{i_j}) \Big]^{\delta_{ij}} \Big\} \prod_{l=1}^{n} G_i(t_l) = \prod_{i=1}^{m} \prod_{j=1}^{n} \Big[g_i(t_{i_j}) \Big]^{\delta_{ij}} G_i(t_j). \tag{2.60}$$

A função de verossimilhança pode ser separada em componentes separados para m causas de falhas sem a suposição de independência entre as causas de falha. Portanto podemos considerar a maximização para cada causa separadamente, desde que não exista parâmetros comuns para diferentes causas (Nakano, 2003).

Modelos de Regressão em Riscos Competitivos

A observação de covariáveis também é importante em riscos competitivos por influenciar o tempo de sobrevida do indivíduo. Visto isso, a seguir serão apresentados os modelos de regressão com abordagem em riscos competitivos, seguindo a ideia desenvolvida na Seção 2.1.6, para as distribuições Weibull, exponencial, log-normal e log-logística.

Distribuição Weibull

Considere duas variáveis aleatórias independentes T_1 e T_2, em que $T_i \sim$ Weibull (α_i, γ_i). A função de sobrevivência conjunta da distribuição Weibull, para a abordagem em riscos competitivos, segundo Nakano (2003), é dada por:

$$S(t_1, t_2) = S(t_2)S(t_2) = \exp\Big\{ -\Big(\frac{t_1}{\alpha_1}\Big)^{\gamma_1} + \Big(\frac{t_2}{\alpha_2}\Big)^{\gamma_2} \Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.61}$$

O modelo de regressão Weibull, na abordagem de riscos competitivos é construído alocando as covariáveis nos parâmetros de escala α_1 e α_2, que será apresentado a

seguir

$$\alpha_1 = \exp(\mathbf{x}_1^T\boldsymbol{\beta}_1) \quad \text{e} \quad \alpha_2 = \exp(\mathbf{x}_2^T\boldsymbol{\beta}_2). \tag{2.62}$$

Sendo assim, segundo Nakano (2003), a função de sobrevivência conjunta para o modelo de regressão Weibull, para a abordagem em riscos competitivos, é dada por:

$$S(t_1,t_2) = \exp\Big\{-\Big(\frac{t_1}{\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)}\Big)^{\gamma_1} + \Big(\frac{t_2}{\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)}\Big)^{\gamma_2}\Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.63}$$

E a função de sobrevivência total é escrita como

$$S_Y(t) = \exp\Big\{-\Big(\frac{t_1}{\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)}\Big)^{\gamma_1} + \Big(\frac{t_2}{\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)}\Big)^{\gamma_2}\Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.64}$$

Distribuição Exponencial

Considere duas variáveis aleatórias independentes T_1 e T_2, em que $T_i \sim$ exponencial(α_i, γ_i). A função de sobrevivência conjunta da distribuição exponencial, para a abordagem em riscos competitivos, segundo Nakano (2003), é dada por:

$$S(t_1,t_2) = S(t_2)S(t_2) = \exp\Big\{-\Big(\frac{t_1}{\alpha_1}\Big) + \Big(\frac{t_2}{\alpha_2}\Big)\Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.65}$$

O modelo de regressão exponencial, na abordagem de riscos competitivos, assim como foi apresentado acima para a distribuição Weibull, é construído utilizando covariáveis nos parâmetros de escala α_1 e α_2, que será apresentado a seguir

$$\alpha_1 = \exp(\mathbf{x}_1^T\boldsymbol{\beta}_1) \quad \text{e} \quad \alpha_2 = \exp(\mathbf{x}_2^T\boldsymbol{\beta}_2). \tag{2.66}$$

Sendo assim, segundo Nakano (2003), a função de sobrevivência conjunta para o modelo de regressão exponencial, para a abordagem em riscos competitivos, é dada por:

$$S(t_1,t_2) = S(t_1)S(t_2) = \exp\Big\{-\Big(\frac{t_1}{\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)}\Big) + \Big(\frac{t_2}{\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)}\Big)\Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.67}$$

E a função de sobrevivência total é escrita como:

$$S_Y(t) = \exp\Big\{-\Big(\frac{t_1}{\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)}\Big) + \Big(\frac{t_2}{\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)}\Big)\Big\}, \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.68}$$

Distribuição log-normal

Considere duas variáveis aleatórias independentes T_1 e T_2, em que $T_i \sim$ log-normal(μ_i, σ_i). A função de sobrevivência conjunta da distribuição log-normal, para a abordagem em riscos competitivos, segundo Nakano (2003), é dada por:

$$S(t_1,t_2) = S(t_1)S(t_2) = \Phi\Big(\frac{-\log(t_1)+\mu_1}{\sigma_1}\Big)\Phi\Big(\frac{-\log(t_2)+\mu_2}{\sigma_2}\Big), \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0 \tag{2.69}$$

no qual, $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

O modelo de regressão log-normal é obtido a partir das covariáveis, pois influenciam no tempo de sobrevida dos indivíduos. Assim, as covariáveis são colocadas nos parâmetros de escala μ_1 e μ_2, que será apresentado a seguir

$$\mu_1 = \mathbf{x}_1^T\boldsymbol{\beta}_1 \quad \text{e} \quad \mu_2 = \mathbf{x}_2^T\boldsymbol{\beta}_2. \tag{2.70}$$

Sendo assim, segundo Nakano (2003), a função de sobrevivência conjunta do modelo de regressão log-normal, para a abordagem em riscos competitivos, é dada por:

$$S(t_1, t_2) = S(t_1)S(t_2) \tag{2.71}$$

$$= \Phi\Big(\frac{-\log(t) + (\mathbf{x}_1^T\boldsymbol{\beta}_1)}{\sigma_1}\Big)\Phi\Big(\frac{-\log(t) + (\mathbf{x}_2^T\boldsymbol{\beta}_2)}{\sigma_2}\Big), \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0.$$

E a função de sobrevivência total é escrita como

$$S_Y(t) = \Phi\Big(\frac{-\log(t) + (\mathbf{x}_1^T\boldsymbol{\beta}_1)}{\sigma_1}\Big)\Phi\Big(\frac{-\log(t) + (\mathbf{x}_2^T\boldsymbol{\beta}_2)}{\sigma_2}\Big), \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.72}$$

Distribuição log-logística

Considere duas variáveis aleatórias independentes T_1 e T_2, em que $T_i \sim$ log-logística (α_i, γ_i). A função de sobrevivência conjunta da distribuição log-logística, para a abordagem em riscos competitivos, segundo Nakano (2003), é dada por:

$$S(t) = \Big(\frac{1}{1 + (t_1/\alpha_1)^\gamma}\Big)\Big(\frac{1}{1 + (t_2/\alpha_2)^\gamma}\Big), \quad t_1 \geq 0 \quad \text{e} \quad t_2 \geq 0. \tag{2.73}$$

O modelo de regressão log-logística também é construído a partir de covariáveis, que são alocadas nos parâmetros de escala α_1 e α_2, conforme será apresentado a seguir:

$$\alpha_1 = \exp(\mathbf{x}_1^T\boldsymbol{\beta}_1) \quad \text{e} \quad \alpha_2 = \exp(\mathbf{x}_2^T\boldsymbol{\beta}_2). \tag{2.74}$$

Sendo assim, segundo Nakano (2003), a função de sobrevivência conjunta do modelo de regressão log-logística, para a abordagem em riscos competitivos, é dada por:

$$S(t_1, t_2) = S(t_1)S(t_2) = \Big(\frac{1}{1 + (t_1/(\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)))^\gamma}\Big)\Big(\frac{1}{1 + (t_2/(\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)))^\gamma}\Big), \tag{2.75}$$

em que $t_1 \geq 0$ e $t_2 \geq 0$. E a função de sobrevivência total é escrita como

$$S_Y(t) = \Big(\frac{1}{1 + (t_1/(\exp(\mathbf{x}_1^T\boldsymbol{\beta}_1)))^\gamma}\Big)\Big(\frac{1}{1 + (t_2/(\exp(\mathbf{x}_2^T\boldsymbol{\beta}_2)))^\gamma}\Big), \tag{2.76}$$

em que $t_1 \geq 0$ e $t_2 \geq 0$.

Existem duas técnicas para estimação de parâmetros bastante conhecidas: mé-

todo de mínimos quadrados e a estimação por máxima verossimilhança. Assim como na Seção 2.1.7, aqui o método mais adequado será a estimação por máximo verossimilhança, pois além de conter ótimas propriedades para grandes amostras, permite que seja trabalhado dados com censura.

Capítulo 3

Aplicação

3.1 Descrição do problema

As instituições financeiras, incluindo fundos de pensão tem como importante fonte de geração de recursos o segmento de empréstimos. No entanto, o risco de inadimplência é uma preocupação para os credores. Contudo, esse risco pode ser minimizado realizando-se previsões sobre o comportamento futuro dos clientes por meio da análise de dados. Dessa forma, a motivação desse estudo foi obter um modelo estatístico para que se possa reduzir os riscos de um participante ficar inadimplente, seja por meio de reduzir o número de parcelas do empréstimo, redefinindo o valor a ser concedido, ou analisando outras variáveis que sejam estatisticamente significativas.

Os dados utilizados nesse estudo foram fornecidos pela Fundação dos Economiários Federais (FUNCEF). Os dados foram coletados, mensalmente, de fevereiro de 2008 a dezembro de 2016. Haviam sido disponibilizadas 59.053 observações. No entanto, 24.408 observações eram referentes a adiantamento de 13º cuja devolução é feita em uma única prestação. Dessa forma, ao analisar a variável quantidade de prestações a frequência de prestações igual a 1 ficava inflada, fazendo com que as conclusões do estudo ficassem tendenciosas. Sendo assim, foi tomada a decisão de excluir as observações referentes aos adiantamentos de 13º. Portanto, existem 34.645 observações e 5 covariáveis, sendo elas: modalidade (fixo ou variável), quantidade de prestações, valor concedido, sexo (feminino ou masculino), idade. Além disso, serão considerados dois eventos de interesse nesse estudo: tempo até ocorrer a inadimplência e tempo até a quitação.

3.2 Análise descritiva

Inicialmente, será feita uma análise exploratória do conjunto de dados que será utilizado na modelagem do tempo até a ocorrência da inadimplência e na modelagem do tempo até a quitação dos contratos de empréstimos, pois a análise descritiva possibilita conhecer melhor os dados.

Tabela 3.1: Quantidade de participantes que contrataram empréstimos de acordo com as modalidades fixo e variável.

Modalidade	Quantidade	Percentual (%)
fixo	20.113	58,07
variável	14.521	41,93
Total	34.634	100

Existem duas modalidades de empréstimo, a fixa e a variável. Na modalidade fixa, o empréstimo é realizado de acordo com a Tabela Price. E, na modalidade variável, a taxa de juros é realizada de acordo com a Tabela Sac (Sistema de amortização Constante) acrescida do Índice Nacional de Preços ao Consumidor (INPC). Dado que a taxa de juros da modalidade variável é acrescida do INPC e a modalidade fixa tem a taxa de juros composta somente pela taxa de juros da Tabela Price, a taxa de juro da Tabela Price é maior. A vantagem da modalidade fixa em relação à variável é que a prestação da primeira será constante, ou seja, será sempre o mesmo valor. E, a vantagem da modalidade variável em relação à fixa é que o valor concedido será superior pelo fato da taxa ser inferior, logo a margem de concessão será maior.

Tabela 3.2: Quantidade de participantes que contrataram empréstimos de acordo com o sexo (masculino e feminino).

Sexo	Quantidade	Percentual (%)
Masculino	17.932	51,776
Feminino	16.702	48.224
Total	34.634	100

Tabela 3.3: Estatística descritiva para as variáveis: número de prestações, valor concedido e idade.

Estatística Descritiva	Número de Prestações	Valor Concedido	Idade
máximo	72	150.192,00	101
mínimo	2	73,00	8
média	63,88	25.013,84	46
mediana	72	22.805,00	46,09
desvio padrão	16,25	16.726,76	13,05

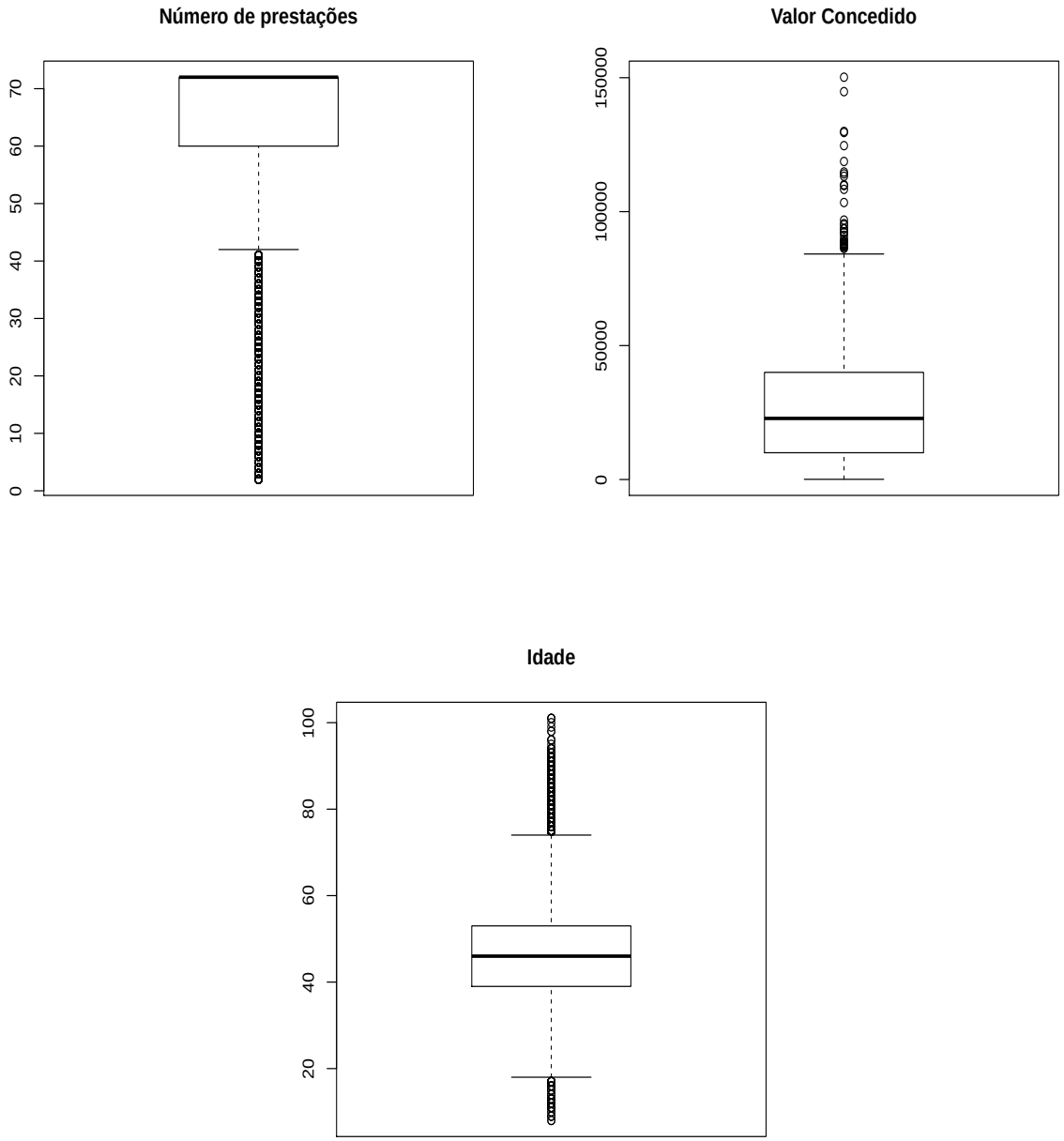

Figura 3.1: Boxplot das variáveis descritivas apresentadas na Tabela 3.3

A quantidade de prestações influencia nas taxas de juros. Dessa forma, foi observado que as prestações 12, 24, 36, 48, 60 e 72 possui maior quantidade de contratos que as outras quantidades, visto que essas quantidades de prestações são o limite para cada taxa de juros. E, a idade dos indivíduos está variando de 8 a 101 anos. Isso ocorre porque há pensionistas menores de idade, portanto existem responsáveis que respondem por eles. E, quando se trata de participantes aposentados, observamos casos de idades avançadas, como participantes com mais de 100 anos. Além disso, observamos nos gráficos

boxplot, na Figura 3.1, que é atípico os contratos terem mais que 42 prestações, o valor concedido ser maior que R$ 85.000 e o contrato ser contraído por pessoas com menos de 18 anos ou mais de 74 anos.

Tabela 3.4: Quantidade de participantes que se encontraram na condição de regular (pagaram o empréstimo regularmente), que ficaram inadimplente, ou quitaram o valor antes do prazo acordado

Evento	Quantidade	Percentual (%)
Quitação	27.033	78,0533
Regular	4.457	12,8689
Inadimplência	3.144	9,0778
Total	34.634	100

O evento inadimplência refere-se às pessoas que atrasaram o pagamento das prestações. O interesse é somente no primeiro atraso. E a quitação diz respeito às pessoas que pagaram todo o valor do empréstimo antecipado, encerrando, assim, o contrato antes do prazo acordado. Já o evento regular diz respeito aos indivíduos que pagaram as parcelas regularmente, conforme acordado no contrato de empréstimo. O tempo até a inadimplência e o tempo até a quitação são os eventos de interesse nesse estudo cujos valores (em dias) serão definidas por T_1 e T_2, respectivamente. Note que ao considerar a variável T_1 os indivíduos regulares e aqueles que quitaram o empréstimo são considerados censurados. Assim, a variável T_1 apresentou 31.490 (90,92%) observações censuradas. Analogamente, os indivíduos regulares e os inadimplentes são tratados como censura ao considerar a variável T_2. Neste caso, a variável T_2 apresentou 7.601 (21.95%) observações censuradas.

A função de sobrevivência dos dados, que é a probabilidade acumulada dos clientes continuar pagando o empréstimo regularmente, ou seja, no primeiro caso, é a probabilidade dos clientes não inadimplirem e no segundo caso é a probabilidade dos clientes não quitarem, se comporta como na Figura 3.2. Note que ambas funções de sobrevivência não atingem o valor zero. Isso ocorre devido ao grande número de censuras observadas.

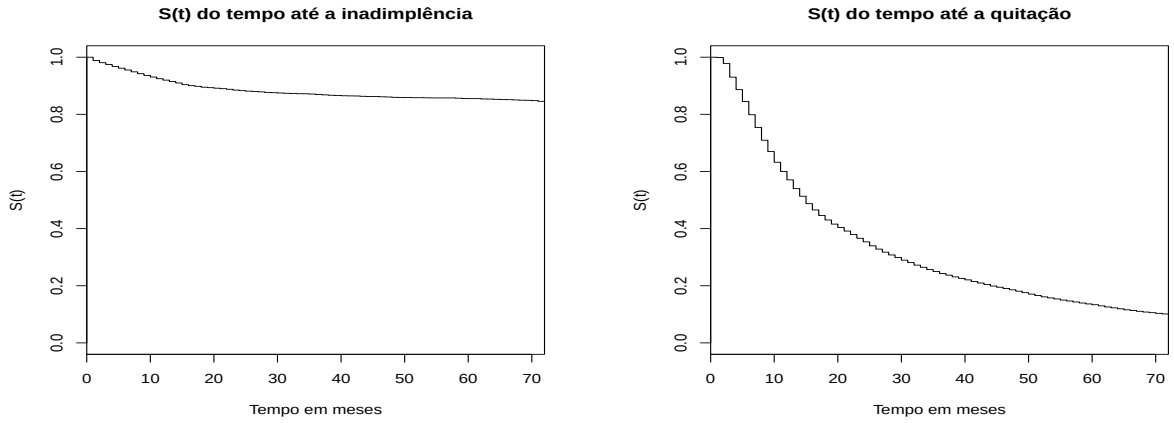

Figura 3.2: Função de sobrevivência estimada do tempo até a inadimplência e função de sobrevivência estimada do tempo até a quitação

Capítulo 4

Resultados

4.1 Caracterização do problema dentro de um contexto de riscos competitivos

A aplicação apresentada no Capítulo 3 pode ser abordada dentro de um contexto de risco competitivos. Note que a variável Y indicando o tempo até a falha (ou seja, tempo até que ocorra o evento de interesse) está suscetível a dois eventos de interesse mutuamente exclusivos: i) primeira inadimplência e ii) quitação da dívida. Esses dois eventos competem entre si para ocasionar a falha.

Definindo como T_1 o tempo até a primeira inadimplência e T_2 como o tempo até a quitação, o que realmente é observado é a variável T, que representa o tempo até a falha, e é descrita por $T = \min(T_1, T_2)$.

Considerando a abordagem latente de riscos competitivos apresentado na Seção 2.2.1, a variável resposta do indivíduo j da amostra pode ser representada pela terna (T_j, δ_{1j} , δ_{2j}), em que T_j é o tempo de falha observado e δ_{ij} é o indicador de censura, i=1,2 e j=1,2,....,n. Aqui, $\delta_{1j} = 1$ indica que o indivíduo j inadimpliu no tempo t e $\delta_{2j} = 1$ indica que o indivíduo j quitou a dívida no tempo t. Uma observação censurada (quando o indivíduo pagou regularmente o empréstimo) é representada por $\delta_{1j} = \delta_{2j} = 0$.

Neste trabalho será considerado T_1 e T_2 independentes. Essa suposição é interessante pois torna o modelo identificável (Seção 2.2.1) e permite que as estimativas dos parâmetros dos modelos associados a T_1 e T_2 possam ser obtidas separadamente por meio da função de verossimilhança (2.73).

4.2 Ajuste do modelo paramétrico sem covariáveis

Como visto na seção anterior, os modelos do tempo até a ocorrência dos eventos de inadimplência e quitação podem ser ajustados separadamente. Desta forma, a primeira etapa busca modelar os dados com base em algum modelo paramétrico, desconsiderando as covariáveis.

Visto que existem várias formas que podem ser assumidas pelos gráficos das funções de risco das variáveis tempo até a quitação e tempo até a inadimplência, foram plotados os gráficos do tempo total em teste (curva TTT) para os tempos até a quitação e para os tempos até a inadimplência, como pode ser observado na Figura 4.1.

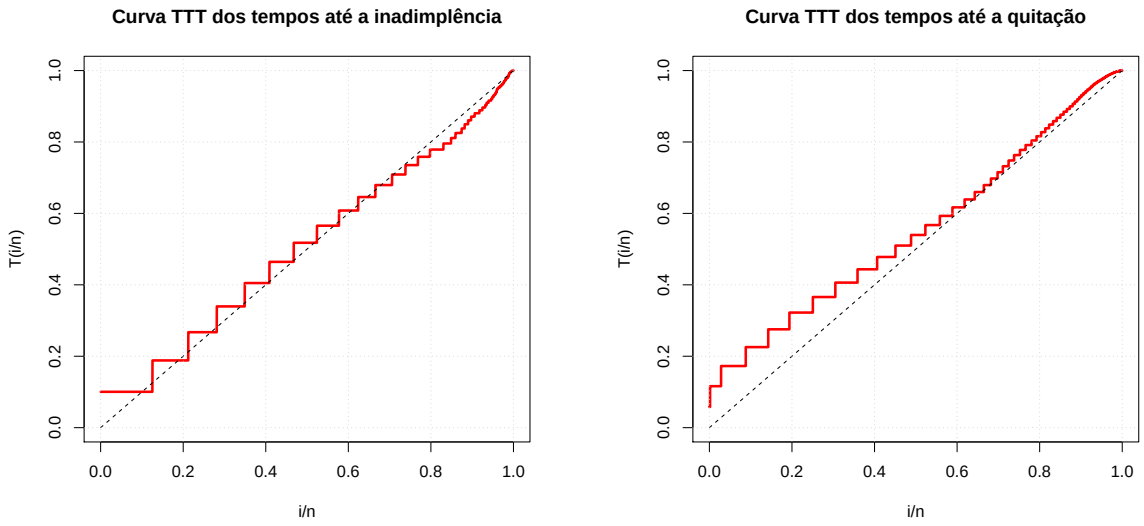

Figura 4.1: Curva TTT dos tempos até a inadimplência (painel da esquerda) e dos tempos até a quitação (painel da direita)

A curva dos tempos até a inadimplência não assume uma forma explícita. Porém, o gráfico indica que o modelo que se ajusta bem aos dados pode ter função de risco crescente ou constante. Já a curva dos tempos até a quitação tem uma forma unimodal. Com base nessas informações, a distribuição Weibull é uma possível candidata ao modelo. No entanto, é importante lembrar que as censuras não são consideradas na construção da curva TTT, o que pode induzir a erros na interpretação e de escolha do modelo nos casos em que o número de censuras é grande. Por esta razão, este trabalho também considerará como distribuições candidatas ao modelo as distribuições log-normal e log-logística, que têm como principal característica representar dados com função de risco unimodal.

Inicialmente, o interesse é analisar a função de sobrevivência dos dados, comparando o resultado do estimador de Kaplan-Meier com três distribuições de probabilidade: Weibull, log-normal e log-logística. Por meio dessa avaliação será possível escolher a distribuição que melhor se adeque aos dados. E, a verificação do ajuste global do modelo será realizada utilizando os resíduos de Cox-Snell. Antes de fazer uma análise mais completa, no momento não serão levados em consideração as covariáveis.

Os eventos foram classificados em inadimplência e censura. Portanto, para comparar os quatro modelos propostos e escolher o mais adequado foram construídos dois

gráficos, um com a função de sobrevivência estimada por Kaplan-Meier e a função de sobrevivência dos modelos Weibull, log-normal e log-logística e o outro com os resíduos de Cox-Snell. Ambos os gráficos foram construídos para o tempo até a inadimplência e para o tempo até a quitação.

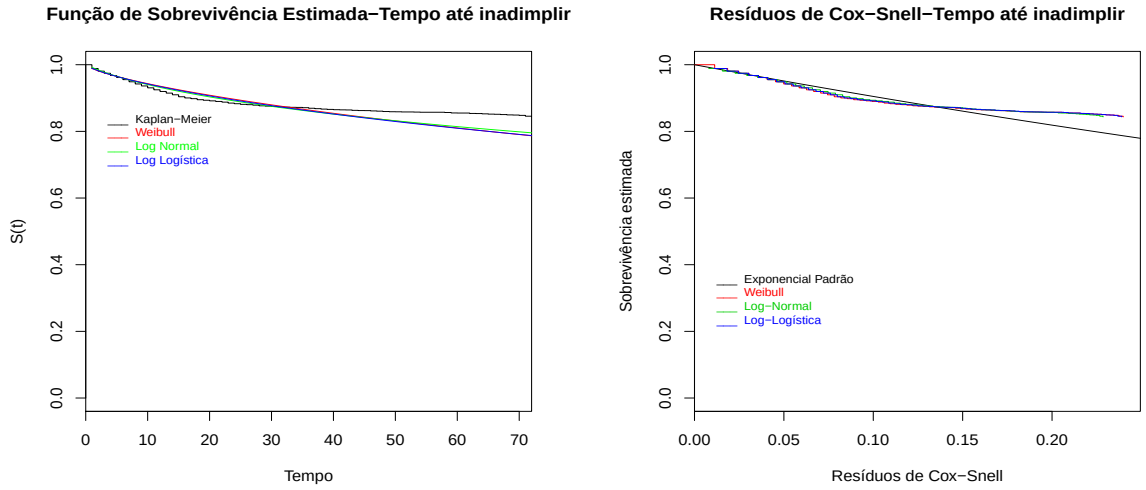

Figura 4.2: Função de sobrevivêcia estimada por Kaplan-Meier e pelos modelos paramétricos Weibull, log-normal e log-logística para o tempo até a inadimplência (Painel da esquerda) e função de sobrevivência dos resíduos de Cox-Snell desses três modelos (Painel da direita).

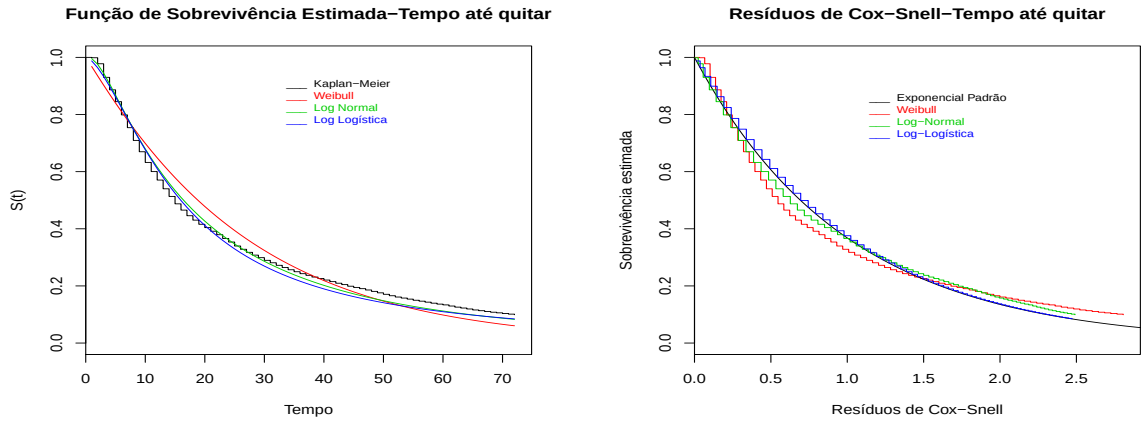

Figura 4.3: Função de sobrevivêcia estimada por Kaplan-Meier e pelos modelos paramétricos Weibull, log-Normal e log-logística para o tempo até a quitação (Painel da esquerda) e função de sobrevivência dos resíduos de Cox-Snell desses três modelos (Painel da direita).

Observando o gráfico da Função de sobrevivência estimada apresentada pela

Figura 4.2 as três distribuições apresentaram um ajuste parecido e ficaram próximas da curva de Kaplan-Meier. Com relação ao gráfico de resíduos de Cox-Snell, as três distribuições também ficaram próximas da distribuição exponencial padrão, porém essa proximidade foi maior na distribuição log-logística. Por esse motivo, optamos por escolher a distribuição log-logística para descrever o tempo até a inadimplência. Um comportamento similar foi observado para o evento quitação. Os três modelos considerados apresentaram bons ajustes aos dados e, novamente, optamos pela distribuição log-logística para representar o tempo até a quitação, pois a mesma aparenta ter apresentado um melhor ajuste (Figura 4.3).

4.2.1 Parâmetro de Força-Estresse

Uma quantidade de interesse na área de modelagem de risco é a probabilidade do cliente inadimplir antes de quitar a dívida, $R = P(T_1 < T_2)$, em que T_1 é o tempo até a primeira inadimplência e T_2 é o tempo até a quitação do contrato. Essa probabilidade é conhecida na literatura de confiabilidade como parâmetro de força-estresse e pode ser considerado como um escore de risco de inadimplência.

Em muitas situações, o cálculo do escore de risco $R = P(T_1 < T_2)$ não pode ser obtido analiticamente, no entanto, o seu valor numérico pode ser facilmente obtido por meio de técnicas de simulação estocástica. O algoritmo abaixo descreve os passos para calcular o valor de R:

1. **Passo 1.** Gerar uma amostra de tamanho M da variável aleatória $T1$;
2. **Passo 2.** Gerar uma amostra de tamanho M da variável aleatória $T2$;
3. **Passo 3.** O valor de R será a frequência relativa do número de vezes (dentre os M valores gerados de cada variável aleatória) em que $T1$ é menor do que $T2$.

Tabela 4.1: Probabilidade de inadimplência (R), estimada pelos modelos de Weibull, log-normal e log-logístico sem covariáves

	Weibull	log-normal	log-logística	empírico*
R	0,09828	0,10149	0,10119	0,090778

* percentual de mal pagadores observado na amostra.

Os valores da probabilidade de inadimplência estimados pelos três modelos foram similares e ficou muito próximo da probabilidade de inadimplência empírica observada na amostra. A probabilidade empírica foi aqui definida como a razão entre o número de observações não censuradas e o tamanho da amostra. Nota-se que as estimativas do parâmetro R foram ligeiramente maiores do que o percentual empírico. De fato, esse comportamento já era esperado, pois o percentual empírico ignora as censuras nos dados e, portanto, subestima a verdadeira proporção de inadimplentes na população.

4.3 Ajuste do modelo paramétrico com covariáveis

Após a análise dos dados pelos modelos sem covariáveis, agora o estudo seguirá com a construção do modelo com covariáveis. As estimativas dos parâmetros do modelo de regressão com covariáveis encontram-se nas Tabelas 4.2, 4.3 e 4.4.

Tabela 4.2: Resultado do ajuste do modelo de regressão Weibull com todas as covariáveis para os dados do tempo até a inadimplência e para os dados do tempo até a quitação do contrato .

	Inadimplência		Quitação	
Covariável	Coeficiente (Erro Padrão)	valor-p	Coeficiente (Erro Padrão)	valor-p
intercepto	β_0 = 7,45 (0,280)	<0,01	β_0 = 3,704 (0,0600)	0,00
modalidade fixo	β_1 = 0		β_1 = 0	
modalidade variável	β_1 = -0,43 (0,051)	<0,01	β_1 = 0,005 (0,0100)	0,67
sexo masculino	β_2 = 0		β_2 = 0	
sexo feminino	β_2 = -0,03 (0,051)	0,48	β_2 = -0,076 (0,0100)	<0,01
prestações	β_3 = -0,02 (0,002)	<0,01	β_3 = -0,003 (0,0005)	<0,01
log(valor)	β_4 = -0,40 (0,028)	0,15	β_4 = 0,010 (0,0060)	0,02
idade	β_5 = 0,020 (0,002)	<0,01	β_5 = -0,006 (0,0005)	<0,01
γ	0,709	<0,01	1,043	<0,01

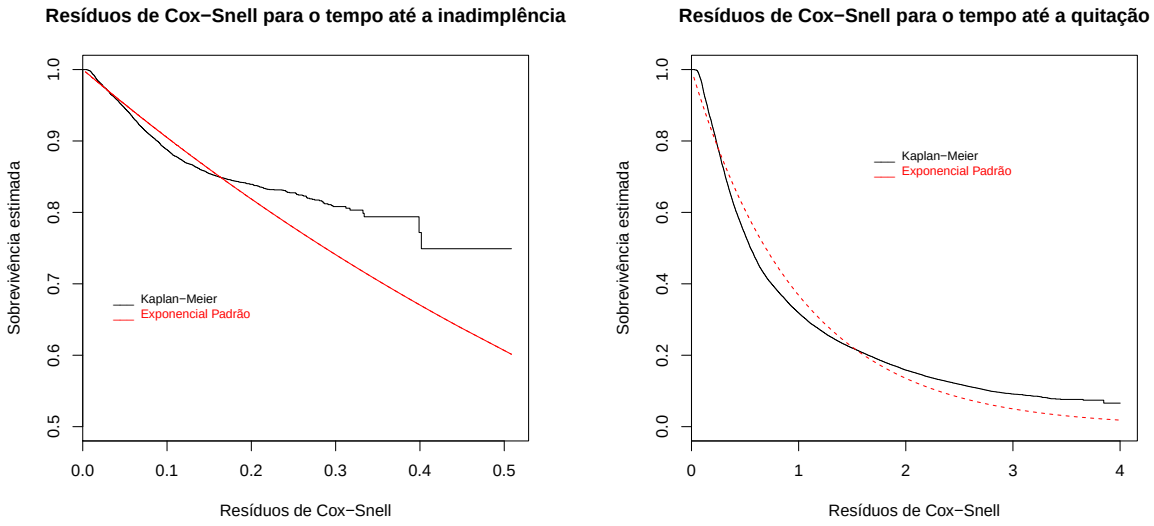

Figura 4.4: Função de sobrevivência dos resíduos de Cox-Snell do modelo Weibull para o tempo até a inadimplência (Painel da esquerda) e Função de sobrevivência dos resíduos de Cox-Snell do modelo Weibull para o tempo até a quitação (Painel da direita).

Tabela 4.3: Resultado do ajuste do modelo de regressão log-normal com todas as covariáveis para os dados do tempo até a inadimplência e para os dados do tempo até a quitação do contrato .

	Inadimplência		Quitação	
Covariável	Coeficiente (Erro Padrão)	valor-p	Coeficiente (Erro Padrão)	valor-p
intercepto	$\beta_0 = 7{,}482$ (0,270)	<0,01	$\beta_0 = 2{,}47$ (0,0569)	0,00
modalidade fixo	$\beta_1 = 0$		$\beta_1 = 0$	
modalidade variável	β_1 = -0,419 (0,052)	<0,01	$\beta_1 = 0{,}025$ (0,0122)	0,0398
sexo masculino	$\beta_2 = 0$		$\beta_2 = 0$	
sexo feminino	β_2 = -0,074 (0,052)	0,15	β_2 = -0,096 (0,0119)	<0,01
prestações	β_3 = -0,020 (0,002)	<0,01	β_3 = -0,003 (0,0004)	<0,01
log(valor)	β_4 = -0,039 (0,028)	0,17	$\beta_4 = 0{,}100$ (0,0065)	<0,01
idade	$\beta_5 = 0{,}022$ (0,002)	<0,01	β_5 = -0,008 (0,0004)	<0,01
σ	2,7	0,00	1,06	<0,01

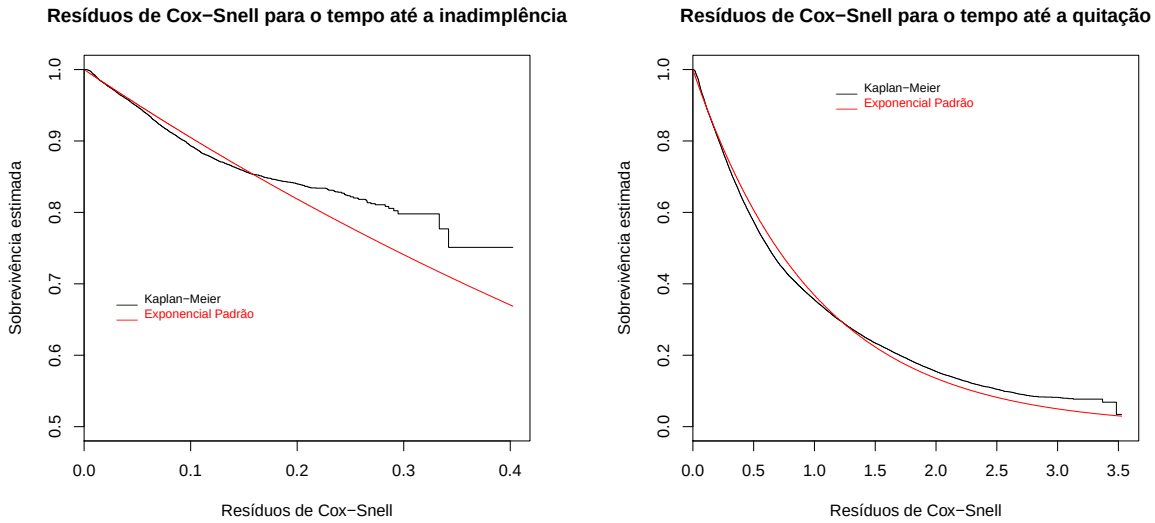

Figura 4.5: Função de sobrevivência dos resíduos de Cox-Snell do modelo log-normal para o tempo até a inadimplência (Painel da esquerda) e função de sobrevivência dos resíduos de Cox-Snell do modelo log-normal para o tempo até a quitação (Painel da direita).

Tabela 4.4: Resultado do ajuste do modelo de regressão log-logística com todas as covariáveis para os dados do tempo até a inadimplência e para os dados do tempo até a quitação do contrato.

	Inadimplência		Quitação	
Covariável	Coeficiente (Erro Padrão)	valor-p	Coeficiente (Erro Padrão)	valor-p
intercepto	β_0 = 7,187 (0,279)	<0,01	β_0 = 2,468 (0,0569)	0,00
modalidade fixo	β_1 = 0		β_1 = 0	
modalidade variável	β_1 = -0,433 (0,052)	<0,01	β_1 = 0,021 (0,0125)	0,09
sexo masculino	β_2 = 0		β_2 = 0	
sexo feminino	β_2 = -0,045 (0,051)	0,38	β_2 = -0,107 (0,0122)	<0,01
prestações	β_3 = -0,023 (0,002)	<0,01	β_3 = -0,005 (0,0004)	<0,01
log(valor)	β_4 = -0,044 (0,029)	0,13	β_4 = 0,115 (0,0068)	<0,01
idade	β_5 = 0,023 (0,002)	<0,01	β_5 = -0,010 (0,0004)	<0,01
γ	0,746	<0,01	1,604	<0,01

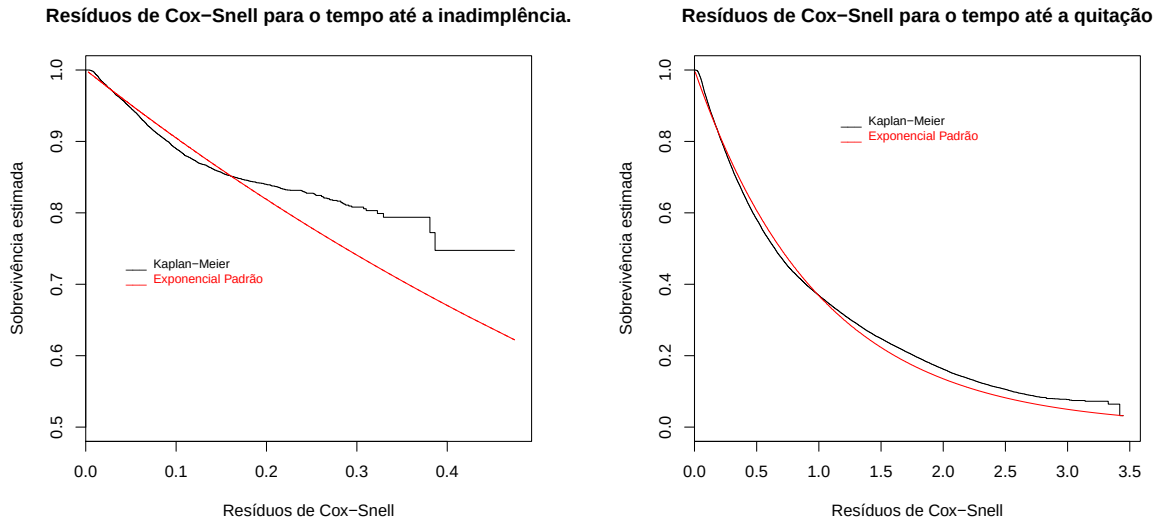

Figura 4.6: Função de sobrevivência dos resíduos de Cox-Snell do modelo log-logística para o tempo até a inadimplência (Painel da esquerda) e função de sobrevivência dos resíduos de Cox-Snell do modelo log-logística para o tempo até a quitação (Painel da direita).

Observando o comportamento do resíduo de Cox-Snell dos três modelos regressão considerados (Figuras 4.4, 4.5 e 4.6), podemos considerar que o modelo log-logístico foi o que melhor se ajustou aos dados.

Assim, tome como exemplo um cliente que tenha contraído o empréstimo na

modalidade variável, que seja do sexo masculino, dividiu o pagamento em 72 prestações, que o valor tenha sido de R$50.000 e que tenha 59 anos de idade (este é o perfil de um cliente inadimplente observado na amostra). Segundo o modelo de regressão log-logístico, a distribuição do tempo até a inadimplência (T_1) desse cliente é dada por $T_1 \sim$ log-logística (α_1, γ_1) e a distribuição do tempo até a quitação (T_2) é dada por $T_2 \sim$ log-logística (α_2, γ_2), em que (Tabela 4.4)

$$\hat{\alpha}_1 = \exp(7,187 - 0,433 + 0 - 72 * 0,023 - \log(50000) * 0,044 + 59 * 0,023) = 395,02,$$

$$\hat{\alpha}_2 = \exp(2,467 + 0,021 + 0 - 72 * 0,005 + \log(50000) * 0,115 - 59 * 0,010) = 16,16,$$

$\hat{\gamma}_1 = 0,746$ e $\hat{\gamma}_2 = 1,604$.

4.3.1 Escore de Risco e classificação dos clientes

Ao ajustar os modelos de regressão para as variáveis T_1: tempo até a primeira inadimplência e T_2: tempo até a quitação do contrato, é possível calcular o parâmetro de força estresse, R=$P(T_1$<$T_2)$.

Como visto na Seção 4.2.1, o valor de R pode ser visto como um escore de risco de inadimplência desse cliente. Assim, ajustado o modelo, é possível calcular o escore de risco R para um cliente qualquer que tem um conjunto de covariáveis x e, com base nesse escore, verificar se o mesmo é classificado como um indivíduo de baixo ou alto risco (de inadimplência).

O ponto de corte para a classificação de um cliente a partir de seu escore de risco pode ser definido como a proporção empírica de inadimplentes observado na amostra.

Neste trabalho foi observado que 9,078% clientes inadimpliram (Tabela 4.1). Desta forma, uma regra de classificação de um cliente pode ser dada por:

Regra de Classificação de um cliente segundo escore de risco R
Classificar o cliente como baixo risco se $R <$ 0,09078
ou
Classificar o cliente como alto risco se $R \geq$ 0,09078

Tomando como exemplo o cliente que tenha contraído o empréstimo na modalidade variável, que seja do sexo masculino, dividiu o pagamento em 72 prestações, que o valor tenha sido de R$50.000 e que tenha 59 anos de idade (este é o perfil de um cliente inadimplente observado na amostra), vimos na seção anterior que $T_1 \sim$ log-logística(395,02;0,746) e $T_2 \sim$ log-logística(16,16; 1,604). Desta forma, este cliente apresenta um escore de risco igual a $R = P(T_1 < T_2) = 0,106 > 0,09078$ e , portanto, esse cliente será classificado como um cliente de alto risco de inadimplência.

Para que seja possível visualizar os dados de previsão de inadimplência comparado aos dados reais coletados para esse estudo é necessário exibir esses valores em uma

matriz denominada de matriz de classificação ou matriz de confundimento (Tabela 4.5):

Tabela 4.5: Matriz de classificação para os valores reais e preditivos de inadimplência.

	Predito pelo modelo	
real	Baixo risco de inadimplência	Alto risco de inadimplência
adimplente	verdadeiro negativo (VF)	falso positivo (FP)
inadimplente	falso negativo (FN)	verdadeiro positivo (VP)

A partir dos resultados da Tabela 4.5 é possível obter uma série de métricas de desempenho do modelo ajustado, tais como sensibilidade, especificidade, valores preditivos positivos, valores preditivos negativos e capacidade total de acertos. Essas métricas são definidas abaixo:

Sensibilidade (S): proporção de maus pagadores, classificados corretamente pelo modelo. Ou seja, é a probabilidade de um indivíduo ser classificado como mau pagador dado que realmente é mau pagador (Silva, 2008). A sensibilidade é definida por:

$$S = \frac{VP}{VP + FN}. \tag{4.1}$$

Especificidade (E): proporção de bons pagadores, classificados corretamente pelo modelo. Ou seja, é a probabilidade de um indivíduo ser classificado como bom pagador dado que realmente é bom pagador (Silva, 2008). A especificidade é definida por:

$$E = \frac{VF}{VF + FP}. \tag{4.2}$$

Valores Preditivos Positivos (VPP):proporção de maus pagadores dado que o modelo assim os identificou (Silva, 2008). O valor preditivo positivo é definido por

$$VPP = \frac{VP}{VP + FP}. \tag{4.3}$$

Valores Preditivo Negativo (VPN): proporção de bons pagadores dado que o modelo assim os identificou (Silva, 2008). O valor preditivo negativo é definido por:

$$VPN = \frac{VF}{VF + FN}. \tag{4.4}$$

Total de Acertos (TA): proporção de acertos de um modelo definido por Silva, 2008:

$$TA = \frac{VP + VF}{VP + FP + VF + FN}. \tag{4.5}$$

Cada um dos indivíduos da amostra teve o seu escore de risco (R) calculado segundo o modelo de regressão log-logístico. Esses indivíduos foram classificados como

baixo risco de inadimplência se R<0,09078 ou como alto risco de inadimplência se R $\geq$ 0,09078. O resultado da classificação desses clientes é apresentada na Tabela 4.6.

Tabela 4.6: Matriz de classificação para os valores reais e preditivos de inadimplência.

	Predito pelo modelo	
real	Baixo risco de inadimplência	Alto risco de inadimplência
adimplente	13.256 (VF)	18.234 (FP)
inadimplente	881 (FN)	2.263 (VP)

De acordo com a Tabela 4.5, a regra de classificação baseada no escore de risco do modelo log-logístico apresentou os seguintes resultados: sensibilidade = 0,72; especificidade = 0,42; valores preditivos positivos = 0,11; valores preditivos negativos = 0,94 e; proporção de acertos =0,45.

Note que, apesar da proporção de acertos do modelo ser baixa (TA=45%), é importante destacar que o valor preditivo negativo é alto (VPN=94%). Isso significa que o modelo é útil para identificar os bons clientes. Note que, a probabilidade de que um cliente classificado como baixo risco vir a inadimplir é de apenas 6%.

Ademais, essas métricas foram obtidas adotando-se a proporção de inadimplentes empírica como ponto de corte de classificação. De fato, outros pontos de corte podem ser definidos visando o controle de alguma das métricas definidas.

Capítulo 5

Considerações Finais

Neste trabalho, a inadimplência de clientes que realizam um empréstimo foi modelada por meio de um modelo de regressão considerando duas causas de falha: 1. Primeira Inadimplência do cliente e; 2. Quitação da dívida. Neste trabalho foi definido como T_1 o tempo até a primeira inadimplência e T_2 o tempo até a quitação da dívida. Apesar de apenas um tempo (o mínimo deles) ser observado, os parâmetros das distribuições podem ser estimados por meio da abordagem latente em riscos competitivos.

Este trabalho também propôs um escore de risco de inadimplência, definido como a probabilidade de um cliente inadimplir antes de quitar a dívida. Esse escore, denotado por R, é definido como $R = P(T_1 < T_2)$, que também é conhecido na literatura como parâmetro de Força-Stress.

A metodologia proposta foi aplicada a um conjunto de dados reais sobre participantes da FUNCEF que realizaram um empréstimo no fundo de pensão. A análise considerou os modelos de regressão Weibull, log-normal e log-logística como possíveis modelos para representar o tempo até a inadimplência e quitação. Desses três modelos, o que apresentou o melhor ajuste foi o log-logístico e o mesmo foi adotado para o calculo do escore de risco das observações da amostra.

Os resultados mostraram que a classificação dos clientes pelo escore de risco baseado no modelo log-logístico apresenta uma baixa taxa de Falsos Negativos. Essa é uma característica interessante para a instituição financeira, pois o modelo é útil para identificar os "bons pagadores".

Referências Bibliográficas

[1] Aarset, M. V. (1987). **How to identify bathtub hazard rate**. IEEE Transactions Reliability, v.36, p.106-108.

[2] Assane, C. C. (2013). **Análise de dados de sobrevivência na presença de risco competitivos**. Dissertação (Mestrado em Engenharia de Produção). Universidade Federal do Rio de Janeiro.

[3] Berman, S. M. (1963). **Note on extreme values competing risks and semi-Markov process.** Ann.Math.Statist., v.34, p.1104 - 1106.

[4] Brunello, G. H. V.; Nakano, E. Y. (2015) **Inferência bayesiana no modelo Weibull discreto em dados com presença de censuras.** TEMA - Tend. Mat. Apl. Comput., V.16, n.2, p.97-110.

[5] Carrasco, C. G.; Tutia, M. H.; Nakano, E. Y. (2012). **Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros.** TEMA - Tend. Mat. Apl. Comput., v.13, n.3, p.247-255.

[6] Colosimo, E. A.; Giolo, S. R. (2006). **Análise de Sobrevivência Aplicada**, (1 ed.). EDGARD BLUCHER.

[7] Conover, W. J. (1999). **Practical nonparametric statistics**. 3. ed. New York: J Wiley, 584p.

[8] Conselho Monetário Nacional. **Dispõe sobre as diretrizes de aplicação dos recursos garantidores dos planos administrados pelas entidades fechadas de previdência complementar**. Resolução n. 4.661, de 25 de maio de 2018. Lex: DOU, 29/5/2018, Seção 1, p. 22-24.

[9] Costa, N.S.S. (2013). **Modelo de riscos múltiplos com fração de cura**. Monografia(Graduação em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília, p. 11.

[10] COX, D. R. (1959). **The analysis of exponentially distributed lifetimes with two types of failure**. Journal of the Royal Statistical Society, v.21, n. 2, p.411-421.

[11] Fachini-Gomes, J. B. (2015) **Análise de Sobrevivência**; Notas de Aula. Departamento de Estatística, Universidade de Brasília, Brasília.

[12] Fachini-Gomes, J. B. (2011). **Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência**. Tese (Doutorado em Ciências).Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

[13] Garcia, P. N. A. (2013). **Aplicação de técnicas de Análise de Sobrevivência em pacientes submetidos à intervenção coronária curtânea**. Monografia(Graduação em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília.

[14] Giordani, N. E. (2013). **Apicação do modelo de risco competitivos em pacientes diagnosticados com câncer no ano de 2006 no Hospital de Clínicas de Porto Alegre**. Monografia(Graduação em Estatística) - Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[15] Giordani, N. E. (2015). **Riscos competitivos: Uma aplicação na sobrevida de pacientes com câncer**. Dissertação (Mestrado em Epidemiologia). Faculdade de Mediciona, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[16] Kaplan, E. L.; Meier, P. (1958). **Nonparametric estimation from incomplete observations.** J. Am. Stat. Assoc., v.53, n.282, p.457-481.

[17] Klein, J. P; Moeschberger, M. L. (2003). **Survival analysis: Techniques for censored and truncated data**. 2.ed. New York: Springer-Verlag, 536p.

[18] Lawless, J. F. (2003). **Statistical models and methods for lifetime data**. 2.ed. New Jersey: John Wiley and Sons, 664p.

[19] Lopes, C. M. C.(2008). **Modelos de sobrevivência com fração de cura e efeitos aleatórios**. Tese (Doutorado em Ciências).Instituto de Matemática e Estatística , Universidade de São Paulo, São Paulo.

[20] Magalhães, M. N. (2006). **Probabilidade e Variáveis Aleatórias**, (2 ed.). EDUSP.

[21] Maia, M. A. (2015). **Análise do tempo até a re-hospitalização de pacientes com esquizofrenia via técnicas de análise de sobrevivência**. Monografia(Graduação em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília.

[22] Nakano, E. Y. (2003). **Uma extensão do modelo weibull bivariado de Ryu: Uma aplicação bayeiana em riscos competitivos**. Dissertação (Mestrado em Estatística). Centro de Ciências Exatas e de Tecnologia. Universidade Federal de São Carlos, São Carlos.

[23] Nakano, E. Y.; Carrasco, C. G. (2006). **Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência.** TEMA - Tend. Mat. Apl. Comput., v.7, n.1, p.91-100.

[24] Nakano, E. Y.; Cunha, J. F. (2012). **Análise do efeito da camuflagem no tempo de segregação em regiões texturizadas utilizando o modelo de riscos proporcionais de Cox.** Semina: Ciências Exatas e Tecnológicas, v.33, n.2, p.141-148.

[25] Nakano, E. Y.; Rodrigues, J. (2006). **Uma extensão do modelo Wibull Bivariado de Ryu: Uma aplicação bayesiana em riscos competitivos.** Revista Mat. Est.São Paulo, v.24, n.4, p.99-115.

[26] Oliveira, M. L. (2014). **Análise de dados de transplante de medula óssea: Proposta do modelo de regressão Kumaraswamy-Weibull com fração de cura**. Monografia(Graduação em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília.

[27] R Core Team (2013). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject.org/.

[28] Ribeiro, R. M. L. (2015). **Aspectos de regulação econômica dos fundos de investimentos: Análise dos ordenamentos jurídicos brasileiro e americano**. Monografia(Graduação em Direito) - Setor de Ciências Jurídicas, Universidade Federal do Paraná, Curitiba.

[29] Santana, T. V. F. (2010). **As distribuições Kumaraswamy-log-logística e Kumaraswamy-logística**. Dissertação (Mestrado em Ciências).Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

[30] Santos, D. F. (2017). **Modelo de Regressão Log-Logístico discreto com fração de cura para dados de sobrevivência.**. Dissertação (Mestrado em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília.

[31] Santos, R. O; Nakano, E. Y. (2015). **Análise do tempo de permanência de trabalhadores no mercado de trabalho do Distrito Federal via modelo de riscos proporcionais de Cox e Log-normal.** Rev. Bras. Biom., v.33, n.4, p.570-584.

[32] Santos, R. O. (2014). **Análise do tempo de permanência do trabalhador formal no mercado de trabalho no Distrito Federal**. Monografia(Graduação em Estatística) - Instituto de ciências Exatas, Universidade de Brasília, Brasília.

[33] Silva, P. H. F. (2008). **Medidas do Valor Preditivo de Modelos de Classificação Aplicados a Dados de Crédito**. Projeto de Pesquisa de Iniciação Científica) - Centro de Estudo de Riscos - Departamento de Estatística, Universidade Federal de São Carlos, São Carlos.

[34] Tsiatis, A. (1975) **A nonidentifiability aspect of the problem of competing riskis.** Proc.Nat.Acad.Sci,U.S.A., v.72, n.1, p.20-22.