UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

ANDRÉ LUÍS SOBRAL DE CARVALHO

ESTUDO DO TEMPO DE INADIMPLÊNCIA ATRAVÉS DA ANÁLISE DE SOBREVIVÊNCIA E REGRESSÃO DE COX

São Cristóvão – SE

2021

ANDRÉ LUÍS SOBRAL DE CARVALHO

ESTUDO DO TEMPO DE INADIMPLÊNCIA ATRAVÉS DA ANÁLISE DE SOBREVIVÊNCIA E REGRESSÃO DE COX

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Orientador (a): Allan Robert da Silva

São Cristóvão – SE

2021

ANDRÉ LUÍS SOBRAL DE CARVALHO

ESTUDO DO TEMPO DE INADIMPLÊNCIA ATRAVÉS DA ANÁLISE DE SOBREVIVÊNCIA E REGRESSÃO DE COX

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Apresentado em____/____/____.

Banca Examinadora:

_____

Prof. Allan Robert da Silva

Orientador

_____

Prof. Carlos Raphael Araújo Daniel

1° Examinador

_____

Prof. Luiz Henrique Gama Dore de Araujo

2° Examinador

AGRADECIMENTOS

Agradeço primeiramente a Deus que sempre esteve comigo durante as dificuldades nesta trajetória de graduação, me dando forças para essa conquista.

Dedico essa conquista em especial aos meus Pais, Carlos Eduardo de Carvalho e Joana d'Arc Franco Sobral. O cuidado e dedicação dos senhores deram em alguns momentos, a esperança para seguir.

Ao Prof º. Allan Robert da Silva pela orientação e oportunidade a mim concedida. Aos Profªs Carlos Raphael e Luiz Henrique. É um prazer tê-los na banca examinadora.

A toda minha família por sempre acreditar em mim, em especial a minha Avó, Maria Isabel Franco Sobral, inspiração dada durante a graduação. Sem vocês, tudo isso não seria possível.

A minha esposa, Eliane Santos Silva Sobral e meu filho Kauã Matheus. Com vocês, tenho me sentido mais vivo de verdade. Obrigado pelo carinho e atenção de sempre nesta caminhada.

Aos meus Amigos: Fillipe Marinho, Sibelle Sá, Mércia Valeria, Marcos Oliveira, Mayara Oliveira. Conviver com vocês foi muito gratificante.

Por fim, a todos que direta ou indiretamente fizeram parte do meu trajeto. Muito Obrigado.

“O sucesso nada mais é que ir de fracasso em fracasso sem que se perca o entusiasmo”.

Winston Churchill

RESUMO

A relevância da análise de crédito pelas entidades financeiras visa reduzir o risco de inadimplência. Para isso, diversas técnicas estatísticas podem ser utilizadas tal como a análise de sobrevivência a fim de estudar o tempo até a ocorrência do pagamento (evento de interesse). Avaliar o tempo de sobrevivência do adimplente e identificar fatores de risco tais como sexo, estado civil, tipo de residência, dependentes e faixa etária. Foram analisados 7751 clientes atrasados em até 180 dias de uma carteira de crédito de uma empresa brasileira. Análise descritiva por meio do número de eventos, número de pagamentos em atraso, média de tempo de pagamento em atraso, mediana do tempo de pagamento em atraso para os cruzamentos; Kaplan-Meier para construção da função de sobrevida; Teste de Log-Rank e de Gehan, para comparar duas ou mais curvas de sobrevida. Foram estimadas as razões de risco bruta e ajustada através da regressão de Cox e a função de risco. Foi utilizado o software R Core Team 2020 e nível de significância adotado foi de 5%. Os tempos medianos para o sexo masculino foram de 18 dias, ($p<0,024$), solteiros, 18 dias ($p<0,001$), morar em casa alugada, 18 dias ($p<0,001$), possui dependentes 18 dias ($p<0,001$) e com menos de 20 anos, 31 dias ($p<0,001$). Em 30 dias, não pagaram 40% dos que moram em casa alugada, 35% dos que possuem dependentes, 50% dos com menos de 20 anos e 37% dos solteiros. Clientes do sexo masculino, solteiros, que residem em casa alugada, com menos de 20 anos são propensos a atrasos maiores.

Palavras-chave: Sobrevivência; Kaplan Meier; Cox; Crédito, Adimplência

ABSTRACT

The relevance of credit analysis by financial entities aims to reduce the risk of default. For this, several statistical techniques can be used, such as survival analysis in order to study the time until the payment occurs (event of interest). Assess the time of survival of the non-performing and identify risk factors such as sex, marital status, type of residence, dependents and age group. 7751 customers overdue within 180 days of a credit portfolio of a Brazilian company were analyzed. Descriptive analysis through the number of events, number of arrears, average arrears, median arrears for crossings; Kaplan-Meier to construct the survival function; Log-Rank and Gehan test, to compare two or more survival curves. Gross and adjusted risk ratios were estimated using Cox regression and the risk function. The R Core Team 2020 software was used and the level of significance adopted was 5%. The median times for males were 18 days, ($p < 0.024$), single, 18 days ($p < 0.001$), living in a rented house, 18 days ($p < 0.001$), having dependents 18 days ($p < 0.001$) and less than 20 years old, 31 days ($p < 0.001$). In 30 days, they did not pay 40% of those who live in a rented house, 35% of those who have dependents, 50% of those under 20 years of age and 37% of singles. Single, male customers who live in a rented home under the age of 20 are prone to longer delays.

Keywords: Survival; Kaplan Meier; Cox; Credit, Default

LISTA DE TABELAS E ILUSTRAÇÕES

SUMÁRIO

1) INTRODUÇÃO

As organizações que trabalham com concessões de credito sempre lidam com o risco, ou seja, um certo grau de incerteza que envolve a operação de credito, que pode ser mensurado pelo setor responsável pela concessão de credito, além de calcular o custo agregado de incerteza em relação ao possível recebimento do credito concedido (XAVIER, 2011).

Entretanto com a atual conjuntura econômica exige das empresas que oferecem concessões de credito um maior gerenciamento e controle sobre o risco, para tentar minimizar as perdas financeiras, segundo Corrêa e Vellasco (2008) antigamente as empresas recorriam a especialistas da área que analisavam o risco somente em critérios julgamentais, entretanto com o passar do tempo, iniciaram a utilização de modelos quantitativos para auxiliar na decisão, esses modelos são denominados de credit scoring¸ que analisa o risco do credito através de características e dados históricos do proponente juntamente com técnicas estatísticas, gerando assim um score, sendo esse valor analisado pela empresas para formar um ranking sobre o risco, e consequentemente decidir sobre o empréstimo do credito ou não.

Segundo Araújo e Carmona (2007) as técnicas quantitativas de credit scorings já estão consolidada nas instituições financeiras mais tradicionais, ou seja, estão sendo utilizada amplamente nas instituições financeira, devido a sua confiabilidade. Nesse mesmo âmbito os autores Scarpel e Milioni (2002) relatam que os estudos sobre modelos que utilizam métodos quantitativos são objetos contínuos, sempre tendo como objetivo a melhoria na tomada de decisão de credito. Entre as variadas técnicas o método da regressão logística se tornou padrão para análise de credito e previsão de inadimplência (Thomas 2002).

Todavia esse estudo tem como objetivo utilizar a técnica de análise de sobrevivência para gerar o modelo de credit scoring. Essa técnica já é utilizada em outras aplicações financeiras como a compra de produtos financeiros (TANG et al 2007) e o desenvolvimento de pontuações genéricas para cartões de varejos (ANDREEVA, 2006). E existe um interesse na utilização da análise de sobrevivência como técnica para gerar o modelo de credit scorings (CROOK ET AL 2007).

1.1) SITUAÇÃO PROBLEMA

A necessidade por parte do mercado financeiro de um sistema que identifique os perfis dos clientes tomadores de crédito para realizar a classificação em grupos de risco para o respectivo provisionamento e também o grande potencial de discriminação dos clientes antes

de inadimplir.

1.2) OBJETIVOS

Classificar os clientes inadimplentes a partir dos dados cadastrais.

2) FUNDAMENTAÇÃO TEÓRICA

2.1) ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado. A razão deste crescimento é o desenvolvimento e aproximação de técnicas estatísticas combinadas com computadores cada vez mais velozes (COLOSIMO E GIOLO, 2006).

Dados de sobrevivência são provenientes de estudos longitudinais que visam o tempo até a ocorrência de determinado evento de interesse, ou seja, o tempo de falha que é a variável resposta. A característica principal desses dados é a presença de censura, que é a observação parcial da resposta, isso significa que o evento de interesse ocorreu antes ou depois do tempo de duração do estudo. Para analisar dados com esta característica é necessário usar métodos estatísticos apropriados que possibilitem incorporar no estudo as informações contidas nas censuras, como o estimador de Kaplan-Meier proposto por Kaplan e Meier em 1958, muito utilizado nas áreas de medicina e engenharia.

Os componentes que constituem a resposta de certo conjunto de dados de sobrevivência são: tempo de falha e censura. O tempo de falha é constituído pelo tempo inicial, escala de medida e o evento de interesse (falha). O tempo de início do estudo deve ser precisamente definido. A escala de medida na maioria das vezes é o tempo real ou “de relógio”. E o evento de interesse muitas das vezes é algo indesejável, chamado de falha. Em estudos desse tipo, é de extrema importância definir de forma clara o que vem a ser a falha, antes de dar início ao estudo. Com isso o tempo de falha vai do tempo inicial até o evento de interesse, que pode ocorrer devido a uma ou mais causas.

Já a censura, refere-se à perda de informação do indivíduo, seja por falta de acompanhamento do paciente no decorrer do estudo ou pela não ocorrência do evento de interesse até o término do experimento. Ressalta-se o fato de que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística. Duas razões justificam tal procedimento: (i) mesmo sendo incompletas, as

observações censuradas fornecem informações sobre o tempo de vida de pacientes; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas (COLOSIMO E GIOLO; 2006).

Os mecanismos de censura são diferenciados em três tipos, Colosimo e Giolo (2006) apresentam quais são: censura tipo I, censura tipo II e censura aleatória. Censura do tipo I ocorre naqueles estudos que ao serem finalizados após um período pré-estabelecido de tempo registram, em seu término, alguns indivíduos que ainda não apresentaram o evento de interesse. Censuras do tipo II resultam de estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos. O terceiro mecanismo é o que mais ocorre na prática médica. Isso acontece quando um paciente é retirado no decorrer do estudo sem ter ocorrido a falha, ou também, se o paciente morre por uma razão diferente da estudada.

2.2) ESTIMADOR DE KAPLAN-MEIER

Na literatura mais especializada em análise de sobrevivência, destacam-se três estimadores de S(t), são eles: o estimador de Kaplan-Meier, o de Nelson-Aalen e o da Tabela de Vida. O de Nelson-Aalen é mais recente que o de Kaplan-Meier, e ambos apresentam basicamente as mesmas características, pois consideram o número de intervalos de tempo o mesmo que o número de falhas distintas. O terceiro tem um reconhecimento histórico, foi proposto no final do século XIX por demógrafos e atuários, aplicado em informações originárias de censos demográficos para medir características associadas ao tempo de vida dos seres humanos. Este estimador, utilizado em grandes amostras tem seus tempos de falhas agrupados em intervalos de forma arbitrária.

O presente trabalho limita-se a tratar somente do estimador de Kaplan-Meier, que é o mais utilizado em análise de sobrevivência e que vem ganhando espaço em estudos de confiabilidade. Proposto por Kaplan e Meier (1958) para estimar a função de sobrevida quando se tem observações com censuras, denominado também de estimador produto-limite.

Deste modo, o estimador de Kaplan-Meier é definido, como:

$$\hat{S}(t) = \prod_{j:t_{(j)} \le t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_{(j)} \le t} \left(1 - \frac{d_i}{n_i} \right) \qquad (1)$$

em que $t(1),t(2),\ldots,t(k)$ representa os k tempos de falhas distintos e ordenados,dj é o número de falhas em $t(j)$, $j = 1,\ldots,k$, e nj é o número de indivíduos sob risco em $t(j)$, ou os

indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a $t(j)$. Em seu artigo original Kaplan e Meier justificam a equação (I) mostrando que ela é o estimador de máxima verossimilhança da função de sobrevivência S(t).

De acordo com Colosimo e Giolo (2006), as propriedades básicas do estimador de Kaplan-Meier são as seguintes: é não viciado para amostras grandes, é fracamente consistente, converge assintoticamente para um processo gaussiano e é estimador de máxima verossimilhança de S(t).

Os autores Breslow e Crowley (1974) e Meier (1975) provaram a consistência e normalidade assintótica do estimador não-paramétrico de S(t), e Kaplan e Meier (1958) mostraram que $S\hat{}(t)$ é o estimador de máxima verossimilhança de S(t).

Lima (2010) utilizou o estimador de Kaplan-Meier para analisar a sobrevida em lojas de shopping centers de forma a identificar um percentual de KPCS (Custo total de ocupação dividido pelas vendas), que sinalize risco de descontinuidade para o lojista, o que resultou no desenvolvimento do primeiro modelo de sobrevida para o mercado de shopping centers.

Com o objetivo de ajudar uma instituição financeira a identificar os fatores que possam indicar previamente eventuais perdas, Gomes e Roslindo (2008) apresentaram técnicas de análise de sobrevivência. Carvalho (2011) utilizou-se do estimador de Kaplan-Meier para relacionar o conceito de fração de cura no contexto de análises financeiras, algo ainda pouco explorado, em que a fração de cura teria como interpretação a quantidade de tomadores de empréstimos que quitaram sua dívida antes de findo o prazo e/ou que quitaram todo o plano sem entrar em inadimplência.

3) RELEVÂNCIA DO TEMA DE PESQUISA PARA O NEGÓCIO

O tema desta pesquisa é relevante para a instituição por ter potencial para originar uma solução de crédito que possibilitará as empresas conhecerem os clientes aos quais concederam o crédito com relação ao tempo de sobrevivência na carteira até a ocorrência da inadiplência.

4) METODOLOGIA

O banco de dados utilizado neste trabalho consiste de uma carteira de clientes do ano de 2015 de uma Instituição Financeira, num período de seis meses, com 7.751 registros de Clientes distribuidos nas variáveis abaixo.

× Cód_Identificador – corresponde ao cliente;

- Gênero – corresponde ao sexo do Cliente (Masculino/Feminino);
- Faixa Etaria – corresponde a faixa etária do Cliente;
- Possui_Dependente – Se o cliente tem cartões adicionais;
- UF_Residência – corresponde ao Estado de Origem do Cliente;
- Estado Civil – corresponde ao estado cívil do Cliente;
- Tipo de Residência – corresponde ao Tipo de residência do Cliente (Alugada/Própia);
- DataCrédito – corresponde a data de cadastro na empresa;
- DataNascimento – corresponde a data de nascimento do cliente;
- Limite – corresponde ao valor do limite concedido;
- Saldo – corresponde ao valor do Saldo Total nos últimos seis meses e no atual;
- MesAtraso – refere-se à presença de atraso ou não, contado em meses;
- DiasAtraso – refere-se à presença de atraso ou não, contado em dias;
- FaixasAtraso – refere-se à presença de atraso ou não, Classificados em Faixas;
- Acordo – registra se houve a presença de acordo ou não;

Os dados deverão ser oriundos de empresas do ramo financeiro que trabalhem com fornecimento de crédito a pessoas físicas e/ou jurídicas e como pode ser verificado, contém informações históricas de crédito de seus clientes, assim como algumas informações cadastrais. Além disso, para o fim ao qual a base se destina, é interessante que clientes que nunca realizaram movimentação, estejam em atraso a mais de 180 dias ou tenham solicitado o cancelamento do crédito não sejam incluídos. De posse dessa base de dados, será feita a eliminação de atributos irrelevantes para a análise de risco, utilizando o Software R.

4.1) Teste log-rank

Considere a hipótese nula de que não existe diferença entre as sobrevivências dos dois grupos. Uma forma de testar a validade desta hipótese é considerar uma medida da diferença entre o número observado de indivíduos que falham nos dois grupos em cada tempo de falha e o número esperado sob a hipótese nula.

H_0: não há diferença entre as curvas;

H1: há diferença entre as curvas;

Para avaliar se essa diferença é estatisticamente significativa, ou seja, se é provável que se deva a efeitos aleatórios, utiliza-se tal como na análise clássica o valor de p ou intervalo de

confiança. Para o cálculo deste p, o teste de significância mais utilizado é o Teste log-rank (também conhecido por estatística do logrank de Mantel, estatística do logrank de Cox-Mantel). Este teste compara o número de eventos observados em cada grupo com o número de eventos que seria esperado com base no número de eventos dos dois grupos combinados, ou seja, não importa a que grupo pertence o indivíduo.

Um teste do qui-quadrado aproximado é usado para testar a significância de uma expressão matemática envolvendo o número de eventos esperados e observados.

Características do Teste log-rank:

Compara duas ou mais curvas de sobrevida (H_0 → as curvas são iguais);

Ordena os tempos de "falhas" dos indivíduos em dois (ou mais) grupos e atribuem postos;

O número esperado de falhas é calculado para cada intervalo para cada grupo;

Calcula um X^2 entre as falhas esperadas versus falhas observadas;

Assume intervalo de tempo pequeno (Ex.: 1 dia ou 1 "falha").

A estatística do teste é dada por:

$$estatística \quad do \quad \log-rank = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)} \sim \chi^2_{(G-1)} \qquad (2)$$

G = 2 grupos

Onde:

$$(O_2 - E_2) = \sum_{j=1}^{k} (m_{2j} - e_{2j}) \qquad (3)$$

K= número de tempos de falha diferentes.

G=número de grupos diferentes.

O Teste de hipóteses de log-rank é o teste mais indicado para os estudos que assumem - riscos proporcionais, ou seja, é apropriado quando a razão das funções de risco dos grupos a

serem comparados é aproximadamente constante. É ainda bastante potente quando as funções de risco não forem proporcionais e não se cruzarem, porém, mostra-se inadequado para funções de risco que se cruzam.

A curva de sobrevivência pode ser usada para comparar o número de eventos observados em cada grupo com o número de eventos esperados ordenando os tempos de "falhas" dos indivíduos nos dois grupos e fornece a possibilidade de determinar quantidades relevantes, por exemplo, mediana, percentis e outras medidas, comparando a distribuição da ocorrência dos eventos observados em cada grupo com a distribuição que seria esperada se a incidência fosse igual em todos os grupos. Caso a distribuição for equivalente à distribuição esperada, então a curva de sobrevivência dos produtos pertencentes ao estrato é equivalente à curva de sobrevivência dos produtos em geral (a covariável não tem efeito na confiabilidade). Este teste produz gráficos e várias estatísticas que são úteis em termos de confiabilidade e análise de sobrevivência.

Os modelos de Credit Scoring são as principais ferramentas de suporte à concessão de crédito, nesse ramo, Chagas (2013) utiliza-se do teste log-rank para modelar o risco de crédito, entender melhor a relação da inadimplência com potenciais covariáveis e ter maior precisão na hora de decidir qual a classificação de risco de cada cliente.

Silva (2012) investiga se a probabilidade de default é afetada pelas condições gerais da economia ao longo do tempo, utilizando a técnica de Análise de Sobrevivência, em que os resultados sugerem que o nível de desemprego mensal tem maior relevância na explicação de inadimplência.

4.2) Teste de Gehan

O teste de Gehan Bastos (Joana e Cristina Rocha, 2007) é uma generalização do teste de Mann-Whitney-Wilcoxon adaptado para dados censurados a partir do conhecido teste não paramétrico de Wilcoxon (Gehan, 1965, Braslow, 1970), baseia-se numa estatística semelhante à utilizada no teste log-rank e nos permite testar a hipótese nula de igualdade das funções de sobrevivência.

A estatística de teste é representada por:

$$U_G = \sum_{j=1}^{r} n_j (D_{1j} - e_{1j}), \qquad (4)$$

Onde n é o número de itens em teste no tempo t, D1j são variáveis aleatórias independentes entre si, r é o número de instantes de morte distintos na amostra conjunta e o valor médio desta estatística é zero sendo E(D1j) = e1j.

Sendo (d1j - e1j) a diferença entre o número total de falhas observadas no grupo 1 e o equivalente número esperado, ponderada pelo número de indivíduos em risco nj, observa-se que será atribuído menor peso às diferenças (d1j - e1j) correspondentes aos instantes em que o número de indivíduos em risco é relativamente pequeno, isto é, para instantes perto do limite de observação.

A variância de UG é dada por:

$$var(U_G) = \sum_{j=1}^{r} n_j^2 v_{1j} = V_G \qquad (5)$$

Consequentemente, sob a validade de H_0 (Hipótese Nula), a estatística segue aproximadamente a distribuição assintótica qui-quadrado com um grau de liberdade.

Bastos (Joana e Cristina Rocha, 2007) constatou que o teste de Gehan é menos sensível do que o teste log-rank às diferenças entre o número observado e o número esperado de falhas que se verifiquem na cauda direita da distribuição do tempo de vida. Em seu trabalho Bastos conclui que o teste log-rank é mais potente na detecção de afastamentos da hipótese de igualdade das distribuições que sejam do tipo risco proporcionais. Entretanto, quando funções se cruzam, como observado no presente estudo, o teste log-rank pode não conseguir detectar diferenças significativas entre as curvas do estimador de Kaplan-Meier, sendo mais indicada a aplicação do teste de Gehan.

A literatura disponível sugere vários ramos de aplicação do Teste de Gehan, de relevante importância para o ramo da saúde e economia, como por exemplo, no artigo Evolução de Idosos em Síndrome Coronariana Aguda Sem Supradesnível do Segmento ST(SCASS de ST) Submetidos a dois diferentes protocolos, publicado em Agosto/2007 na Revista SOCERJ, a autora Elizabeth Mesquita de Sousa aplicou o teste de Wilcoxon (Gehan) para comparar as curvas de sobrevida, buscando avaliar a evolução tardia de dois diferentes protocolos de tratamento em idosos com SCASS de ST.

Já na tese apresentada ao Programa de Pós-Graduação em Economia (IPE/USP, 2013), Modelos de Duração Aplicados à Sobrevivência das Empresas Paulistas entre 2003 e 2007, o autor André Luís Pavão, utiliza-se do Teste de Gehan para analisar as principais causas para a mortalidade das empresas paulistas criadas entre 2003 e 2007. Entretanto, notou-se que a

aplicação do Teste de Gehan para a análise de risco de crédito é pouco utilizada.

4.3) Modelo de Regressão de Cox

O modelo de regressão de Cox, também conhecido como modelo de riscos proporcionais, foi construído com base em Kaplan & Meier (Kaplan e Meier, 1958). A grande vantagem desse modelo de regressão em relação aos demais é não exigir que os dados tenham uma distribuição específica. O modelo é composto por uma parte paramétrica e outra não paramétrica, por isso é classificado como um modelo semi paramétrico (COX, 1972).

A forma geral do modelo é dada por

$$h_i(t) = h_0(t)exp(\boldsymbol{\beta}_i \boldsymbol{x}_i), \qquad (6)$$

onde t é o tempo, x é o vetor das p covariáveis explicativas do modelo (x1, x2, ..., xp), β é o vetor de coeficientes associados às p covariáveis explicativas (β1, β2, ..., βp) e hi(t) é a função taxa de falha para o i-ésimo indivíduo no tempo t. A parte não paramétrica do modelo, h0(t), é o risco basal. A parte paramétrica do modelo, exp(βixi), é uma função dos valores dos vetores x e β para o i-ésimo indivíduo. 10

É importante salientar que o modelo é conhecido como de riscos proporcionais, porque a razão entre a função taxa de falha de dois indivíduos independe do tempo, como pode ser visto a seguir,

$$\frac{h_0(t)exp(\boldsymbol{\beta}\boldsymbol{x}_1)}{h_0(t)exp(\boldsymbol{\beta}\boldsymbol{x}_2)} = exp(\boldsymbol{\beta}\boldsymbol{x}_1 - \boldsymbol{\beta}\boldsymbol{x}_2) = exp(\boldsymbol{\beta}(\boldsymbol{x}_1 - \boldsymbol{x}_2)). \qquad (7)$$

Para ajustar o modelo de riscos proporcionais dado pela função (6) a um conjunto de dados, é preciso que os componentes da função sejam estimados. O componente paramétrico e o componente não paramétrico podem ser estimados separadamente, no entanto a estimação de h0(t) depende da estimação do componente linear da função, exp(βixi).

Para estimar o componente linear, podem-se utilizar métodos como o de máxima verossimilhança, o de verossimilhança parcial, o de verossimilhança aproximada e o de Cox para dados agrupados. O tempo é contínuo, porém, para a coleta dos dados, ele é discretizado em dias, semanas, meses ou outro intervalo, de acordo com o objetivo do estudo. Dessa forma, podem ocorrer eventos de interesse em um mesmo tempo, os chamados empates. O método de verossimilhança parcial é robusto e útil para a estimação quando existem dados censurados, porém, como se assume que a função risco taxa de falha é contínua, não seria possível a existência de empates (COX, 1975).

Para tratar-se empates, utiliza-se uma aproximação do método de verossimilhança parcial. O método mais simples e mais amplamente utilizado foi proposto por Breslow (BRESLOW,1974). A aproximação utiliza a informação do número de empates ocorridos em cada ponto de tempo e a soma dos coeficientes de cada covariável, para os indivíduos que tiveram o evento de interesse empatado com o de outros, de forma a ponderar os coeficientes quando ocorrerem empates. Para encontrar os estimadores dos coeficientes (β), é então utilizado algum método numérico, em geral o de Newton-Raphson, a fim de maximizar a função do componente linear (COLLETT, 2003).

A estimação da parte não paramétrica será baseada no método de máxima verossimilhança. Supondo que foram observados r tempos de sobrevida distintos para os indivíduos que apresentaram o evento de interesse, ordenam- 11 se esses tempos ascendentemente e, considerando que existem dj empates no tempo tj, a estimativa do risco basal é dada por

$$\hat{h}_0(t_j) = 1 - \hat{\xi}, \qquad (8)$$

tal que $\hat{\xi}$ é a solução da seguinte equação,

$$\sum_{l \in D(t_j)} \frac{exp(\hat{\boldsymbol{\beta}} \boldsymbol{x}_l)}{1 - \hat{\xi}^{\exp(\hat{\boldsymbol{\beta}} \boldsymbol{x}_l)}} = \sum_{l \in R(t_j)} exp(\hat{\boldsymbol{\beta}} \boldsymbol{x}_l), \qquad (9)$$

sendo D(tj) o conjunto de todos os indivíduos que tiveram empates no j-ésimo tempo, tj, e R(tj) o conjunto de todos os indivíduos que estão sob risco de incorrerem no evento de interesse no j-ésimo tempo. Utilizando as estimativas das partes paramétrica e não paramétrica do modelo de riscos proporcionais, chega-se à estimativa da função risco para o i-ésimo indivíduo,

$$\hat{h}_i(t) = \hat{h}_0(t) exp(\hat{\boldsymbol{\beta}}_i \boldsymbol{x}_i). \qquad (10)$$

A partir dela, é possível chegar às estimativas da função densidade de probabilidade e da função de sobrevivência (KALBFLEISCH, 1973).

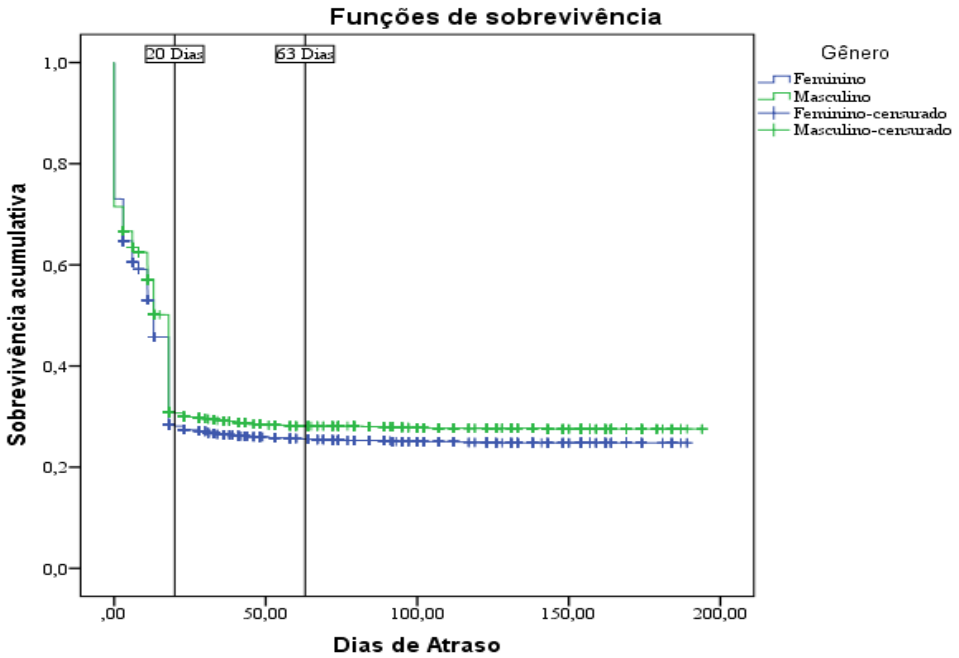
5) RESULTADOS

Os tempos medianos para o sexo masculino foram de 18 dias, (p<0,024), solteiros, 18 dias (p<0,001), morar em casa alugada, 18 dias (p<0,001), possui dependentes 18 dias (p<0,001) e com menos de 20 anos, 31 dias (p<0,001).

Tabela 1: Teste de Long-Rank e Gehan aplicado as variáveis observadas.

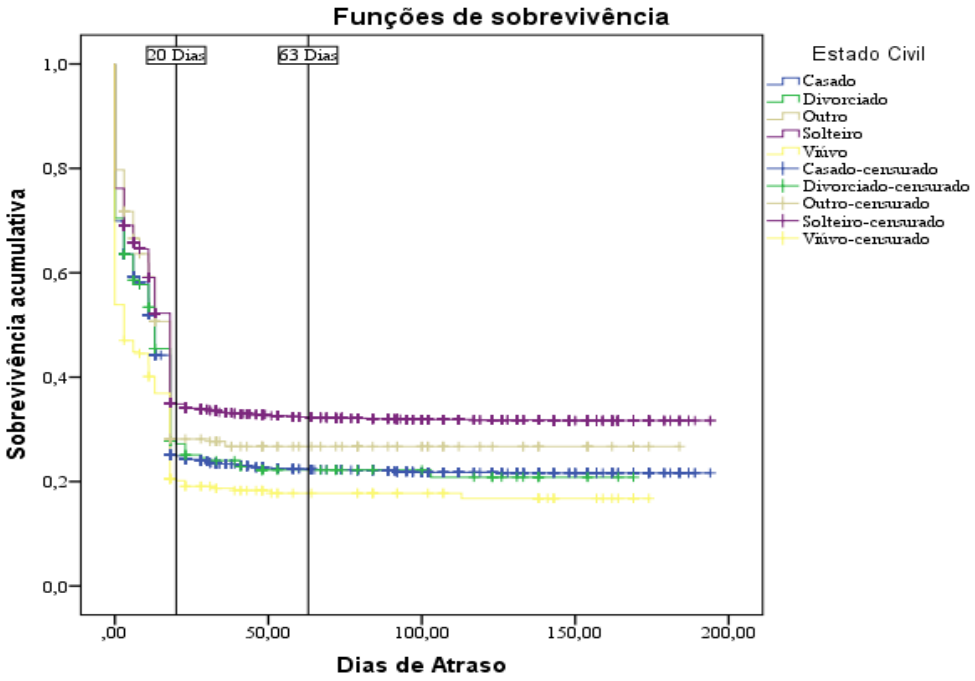
	N total	N de Eventos	Média	Mediana	Log-Rank	Gehan
Gênero						
Feminino	4315	3082	53,9	13	0,001	0,024
Masculino	3439	2394	60,4	18		
Estado Civil						
Casado	3521	2623	49,3	13	<0,001	<0,001
Divorciado	247	182	43,6	13		
Outro	276	193	56,1	18		
Solteiro	3389	2217	68,1	18		
Viúvo	321	261	35,4	3		
Tipo de Residência						
Alugada	579	348	73,3	18	<0,001	<0,001
Família	719	492	62,0	18		
Outros	120	91	43,3	13		
Própria	6336	4545	55,2	13		
Possui Dependentes						
Não	2954	2280	44,9	11	<0,001	<0,001
Sim	4800	3196	65,2	18		
Faixa Etária						
<= 20,00	830	420	95,0	31	<0,001	<0,001
21,00 - 25,00	973	611	73,8	18		
26,00 - 30,00	925	626	64,0	18		
31,00 - 40,00	1922	1377	55,0	13		
41,00 - 50,00	1448	1071	47,3	13		
51,00+	1656	1371	32,8	6		

Figura 1. Curva de sobrevivência do tempo de permanência na inadimplência por gênero.



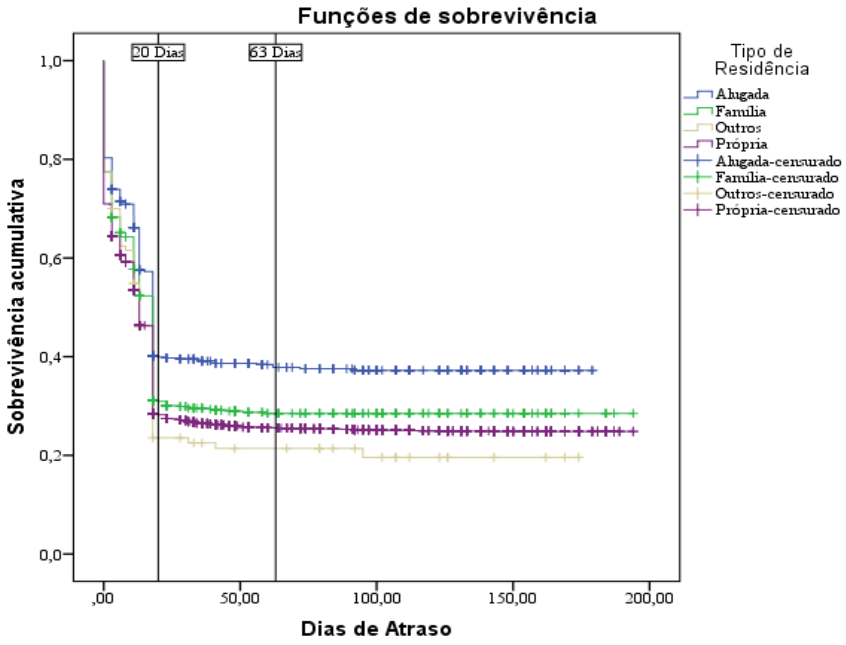
Fonte: Autoria própria

Figura 2. Curva de sobrevivência do tempo de inadimplência por estado civil



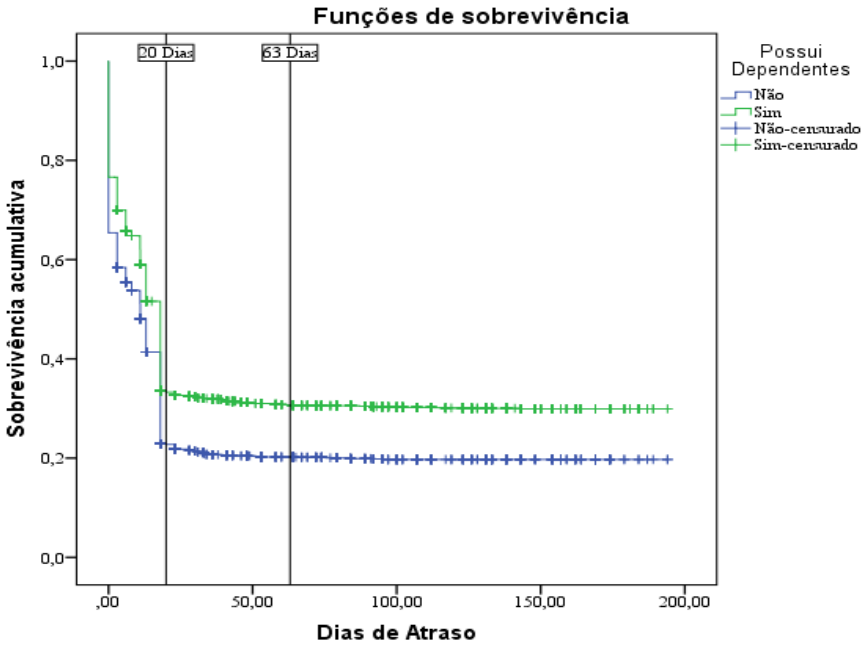
Fonte: Autoria própria

Figura 3. Curva de sobrevivência acumulada do tempo de permanência na inadimplência por tipo de residência.



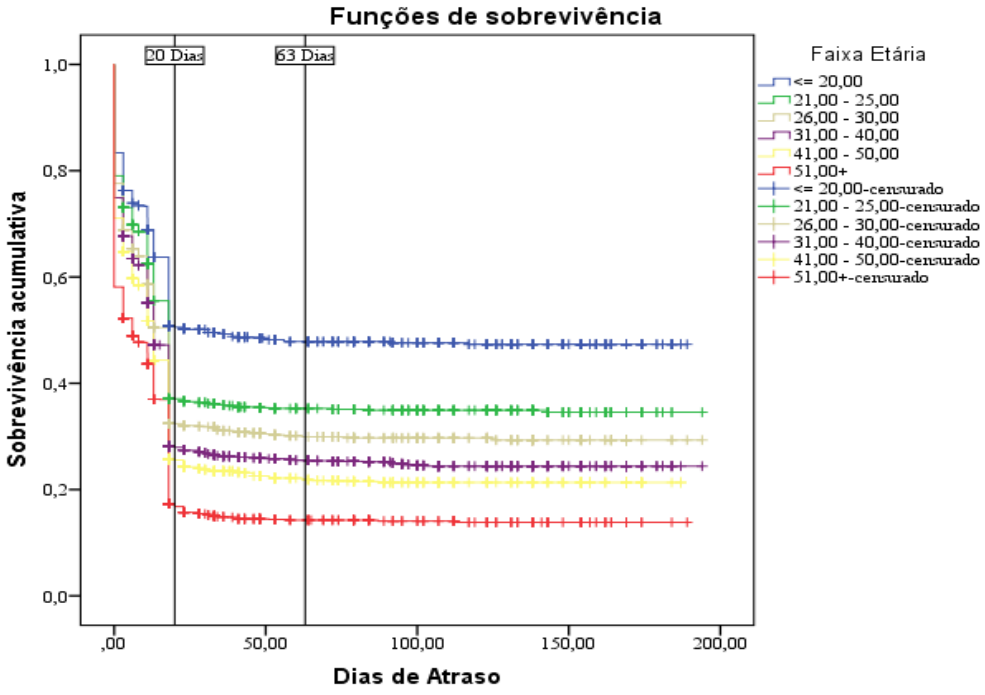
Fonte: Autoria Própia

Figura 4. Curva de sobrevivência de inadimplência por cartões dependentes.



Fonte: Autoria própria

Figura 4. Curva de sobrevivência de inadimplência por Faixa Etária.



Fonte: Autoria Própia

Tabela 2: Regressão de Cox Univariada e Multivariada.

	RR Bruta (IC95%)	p	RR Ajustada (IC95%)	p
Gênero				
Feminino	1,08 (1,02-1,14)	0,005	1,07 (1,01-1,13)	0,02
Masculino	Referência		Referência	
Estado Civil				
Casado	0,82 (0,72-0,93)	0,003	1,01 (0,89-1,16)	0,83
Divorciado	0,81 (0,67-0,98)	0,03	0,88 (0,73-1,07)	0,19
Outro	0,71 (0,59-0,86)	<0,001	0,91 (0,75-1,1)	0,32
Solteiro	0,66 (0,58-0,75)	<0,001	0,87 (0,76-0,99)	0,04
Viúvo	Referência		Referência	
Tipo de Residência				
Alugada	0,73 (0,66-0,82)	<0,001	0,81 (0,73-0,91)	<0,001
Família	0,90 (0,82-0,99)	0,02	1 (0,91-1,11)	0,93
Outros	1,04 (0,85-1,29)	0,68	1,04 (0,85-1,28)	0,70
Própria	Referência		Referência	
Possui Dependentes				
Não	1,31 (1,25-1,39)	<0,001	1,29 (1,22-1,37)	<0,001
Sim	Referência		Referência	
Faixa Etária				
<= 20,00	Referência		Referência	
21,00 - 25,00	1,34 (1,19-1,52)	<0,001	1,29 (1,14-1,46)	<0,001
26,00 - 30,00	1,53 (1,35-1,73)	<0,001	1,41 (1,25-1,6)	<0,001
31,00 - 40,00	1,69 (1,51-1,88)	<0,001	1,54 (1,37-1,72)	<0,001
41,00 - 50,00	1,83 (1,63-2,05)	<0,001	1,69 (1,5-1,9)	<0,001
51,00+	2,28 (2,04-2,54)	<0,001	2,06 (1,83-2,32)	<0,001

RR – Razão de Risco; IC95% – Intervalo de Confiança para 95%.

6) CONCLUSÃO

Em 30 dias, não pagaram 40% dos que moram em casa alugada, 35% dos que possuem dependentes, 50% dos com menos de 20 anos e 37% dos solteiros sendo assim podemos concluir que: Clientes do sexo masculino, solteiros, que residem em casa alugada, com menos de 20 anos são propensos a atrasos maiores, representando maior risco de inadimplência, ou seja, menor chance de honrar com seus compromissos financeiros.

7) REFERÊNCIAS

Andreeva G. European generic scoring models using survival analysis. J Opl Res Soc 57(10): 1180-1187 , 2006

ARAÚJO, E. A.; CARMONA, C. U. M. Desenvolvimento de Modelos Credit Scoring com abordagem de regressão logística para a gestão da inadimplência de uma instituição de microcrédito. Contabilidade Vista & Revista, v. 18, n. 3, p. 107-131, 2007.

BASTOS, Joana; ROCHA, Cristina. Análise de Sobrevivência Métodos Não Paramétricos. Arq Med, Porto , v. 21, n. 3-4, p. 111-114, 2007 . Disponível em <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0871-34132007000300007&lng=pt&nrm=iso>. acessos em 10 jan. 2021.

BRESLOW, N. Covariance analysis of censored survival data. Biometrics. v. 30, p. 89-100, 1974.

CARVALHO, Thiago Morais de. **Análise de Sobrevivência Aplicada ao Risco de Crédito:** Ajuste de Modelos Paramétricos Contínuos a Dados de Tempo Discretos. Disponível em: http://bdm.unb.br/bitstream/10483/4050/6/2011_ThiagoMoraisdeCarvalho.pdf, 2011 Acesso em: 01 fev. 2021.

CHAGAS, Caio Martins. **Risco de Crédito:** Credit Scoring e Aplicações em Análise de Sobrevivência, 2013.

COLLETT, D. Modelling survival data in medical research. 2ª ed. Boca Raton: Chapman &

Hall/CRC, 2003.

CORRÊA, M. F; VELLASCO, M. Análise de risco de crédito em correspondentes bancários através de redes neurais. Revista Inteligência Computacional Aplicada, v. 1, n. 1, p. 23-37, 2008.

COX, DAVID R. Regression models and life-tables. Journal of the Royal Statistical Society, Series B (Methodological), p. 187-220, 1972.

COX, DAVID R. Partial likelihood. Biometrika, v. 62(2), p. 269-276, 1975.

Gehan, E.A. A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52**, 203–233, 1965.

GOMES, G. C. ROSLINDO, J. J. **Análise de Sobrevivência como ferramenta auxiliar na originação e manutenção do ciclo de crédito**. Disponível em: <http://200.17.213.49/lib/exe/fetch.php/disciplinas:ce229:tcc2008_gean_jessica.pdf>,2008 Acesso em: 11 fev. 2020.

KALBFLEISCH, J. D., PRENTICE, R. L. Marginal likelihoods based on Cox's regression and life model. Biometrika, v. 60, p. 267-278, 1973.

KAPLAN, EDWARD L., MEIER P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, v. 53, p. 457-481, 1958.

LIMA, Marcio Targa de. **Inadimplência de Lojistas no Setor de Shopping Centers:** Um

estudo baseado na Análise de Sobrevivência. Disponível em:http://tede.mackenzie.com.br/tde_arquivos/10/TDE-2010-06-18T101123Z-930/Publico/Marcio%20Targa%20de%20Lima.pdf, 2010. Acesso em: 11 fev. 2020.

PAVÃO, André Luís. **Modelos de duração aplicados à Sobrevivência das empresas paulistas entre 2003 e 2007.** Disponível em: https://teses.usp.br/teses/disponiveis/12/12140/tde-24072013-154206/publico/AndreLuisPavao.pdf, 2013 Acesso em: 13 fev. 2020.

SCARPEL E MILIONI: **UTILIZAÇÃO CONJUNTA DE MODELAGEM ECONOMÉTRICA E OTIMIZAÇÃO EM DECISÕES DE CONCESSÃO DE CRÉDITO** http://www.scielo.br/pdf/pope/v22n1/a04v22n1, 2002. Acesso em: 13 fev. 2020.

SILVA, Sandra Almeida. **Estudo de Risco de Crédito em Operações de Cartão de Crédito usando variáveis Macroeconômicas e Técnicas de Análise de Sobrevivência**. Disponível em:http://www.mackenzie.br/fileadmin/PUBLIC/UP_MACKENZIE/servicos_educacionais/stricto_sensu/Administracao_Empresas/Teses_e_Dissertacoes/Sandra_Almeida_Silva.pdf, 2012 Acesso em: 13 fev. 2020.

SOUSA, Elizabeth Mesquita de. **Evolução de Idosos em Síndrome Coronariana Aguda** Sem Supradesnível do Segmento ST (SCASS de ST) Submetidos a Dois Diferentes Protocolos. Disponível em: http://sociedades.cardiol.br/socerj/revista/2007_04/a2007_v20_n04_art01.pdf, 2007. Acesso em: 10 fev. 2020.

S. R. Giolo and E. A. Colosimo. **Analise de Sobrevivência Aplicada**. Edgard Blucher, São Paulo, SP, first edition, 2006.

Tang L, Thomas LC, Thomas S, Bozzetto J-F. It's the economy stupid: modelling financial product purchases. International Journal of Bank Marketing. Vol.25, issue 1, pp.22-38, 2007.

Thomas LC, Edelman DB and Crook JN. Credit Scoring and its Applications. SIAM Monographs on Mathematical Modeling and Computation. SIAM: Philadelphia, USA, 2002.

XAVIER, C. G. Risco na análise de crédito. Florianópolis: UFSC, 2011.