Universidade de Brasília
Departamento de Estatística

Modelos de Regressão Discretos para Dados Grupados:

Uma Aplicação em Avaliação de Risco em Produto de Crédito Parcelado.

Tatiana Santos Rocha

Brasília
2013

Tatiana Santos Rocha

Modelos de Regressão Discretos para Dados Grupados:

Uma Aplicação em Avaliação de Risco em Produto de Crédito Parcelado

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Juliana Betini Fachini
Co-orientador: Prof. Dr. Afrânio Márcio Corrêa Vieria

Brasília
2013

DEDICATÓRIA

À minha querida mãe,
Nívia Eulália *por dedicar a vida aos*
filhos lutando junto a eles em busca
dos sonhos de cada um.
Ao meu irmão,

Fábio Augusto*, pela torcida e amizade.*

EPÍGRAFE

"Grande é a tarefa que nos espera...
Para todos os seres humanos, constitui quase um dever pensar que o que já se tiver realizado é sempre pouco em comparação com o que resta por fazer"

João XXIII

AGRADECIMENTOS

A Deus pelas oportunidades da minha vida e por me mostrar os melhores caminhos para aproveitá-las da melhor forma. A Ele agradeço pelo suporte concedido nesses quatro anos de desafios, além da força e saúde.

A querida Profa. Dra. Juliana pela orientação sem igual, amizade, conselhos e por acreditar na minha capacidade. Agradeço toda dedicação e ensinamentos que foram de extrema importância para a realização deste trabalho.

Ao co-orientador, Prof. Dr. Afrânio, pela colaboração e disposição em partilhar sua experiência na área de risco de crédito.

A minha amada mãe Nívia pelos valores ensinados e por toda estrutura fornecida pra que eu chegasse até aqui. Obrigada pelo apoio, incentivo e por acreditar em mim de maneira incondicional.

Ao irmão Fábio pelo apoio, amizade e torcida.

Ao meu pai pelos ensinamentos, contribuições e torcida.

A toda minha família pela preocupação, energia positiva e, por acompanhar cada etapa, mesmo a quilômetros de distância.

Ao Marcelo pela amizade, compreensão e apoio emocional. Obrigada por todo carinho e por acreditar na minha vitória.

As amigas de curso Bruna e Lívia pela parceria durante esses anos e por todos os desafios e horas de estudo que enfrentamos juntas.

Aos colegas da instituição pelo apoio e confiança. Em especial o Fabiano pela amizade, preocupação e incentivo; obrigada pela confiança, oportunidades e por tudo que me ensinou. A Luciane pela oportunidade e apoio e Aline pela atenção, conselhos e pela troca de conhecimentos.

Aos professores do departamento de estatística da Universidade de Brasília por todos os ensinamentos.

A todos que contribuíram de alguma forma com a realização deste trabalho.

SUMÁRIO

RESUMO

Modelos de Regressão Discretos para Dados Grupados: Uma Aplicação em Avaliação de Risco em Produto de Crédito Parcelado

Com a popularização e crescimento do sistema de concessão de crédito no mercado brasileiro, é crescente a necessidade em mensurar o risco dessas operações para que eventos como a inadimplência sejam prevenidos. A análise de crédito é um processo decisório bastante complexo, envolvendo experiência anterior, conhecimento sobre o que está sendo decidido, método para tomar a decisão e utilização de instrumentos e técnicas específicas. Dentre as diversas metodologias estatísticas que dão suporte a esse procedimento sugerir-se-á a análise de sobrevivência como metodologia alternativa para o desenvolvimento de modelos de risco de crédito. Tal técnica se refere a um conjunto de metodologias estatísticas que estudam dados relacionados ao tempo decorrido até a ocorrência de um evento de interesse. A partir de dados disponibilizados a respeito de empréstimos concedidos por uma instituição financeira brasileira, modelos de regressão discretos serão desenvolvidos. Dessa forma, o objetivo do presente trabalho é apresentar à instituição financeira uma metodologia alternativa em busca de melhorar a qualidade dos modelos atuais, além de acrescentar informações não conhecidas a respeito dos dados, como o tempo até o cliente se tornar inadimplente. Tendo em vista a característica dos dados, serão ajustados modelos de regressão discretos sob a ótica de dados grupados para dados provenientes de uma linha de crédito parcelado com prazo de contratação de dezoito meses.

Palavras-chave: Crédito; Risco de crédito; Análise de sobrevivência; Modelos discretos; Dados grupados.

1 INTRODUÇÃO

Juntamente com a expansão da concessão de créditos financeiros no mercado brasileiro, é crescente a necessidade em mensurar o risco dessas operações, bem como o limite de crédito a ser concedido aos clientes. Uma justificativa relevante é que a concessão de crédito ganhou força na rentabilidade das empresas do setor financeiro, tornando-se uma das principais fontes de receita. Para dar suporte a esses procedimentos, algumas metodologias como regressão logística, análise discriminante, redes neurais, entre outros, são encontradas na literatura.

No presente trabalho sugerir-se-á a análise de sobrevivência como metodologia alternativa para o desenvolvimento de modelos de risco de crédito, tendo em vista que atualmente, a metodologia mais utilizada pelas instituições financeiras é a regressão logística. Alguns métodos de análise de sobrevivência são antigos, mas segundo Lawless (2003), sua rápida expansão no que diz respeito à metodologia, teoria e campo de aplicação se deu por volta de 1970.

A análise de sobrevivência se refere a um conjunto de metodologias estatísticas que estudam dados relacionados ao tempo decorrido até a ocorrência de um evento de interesse. Dessa forma, acredita-se que a análise de sobrevivência pode propor melhoria aos modelos atuais, uma vez que possibilita a estimação da probabilidade de um cliente não se tornar inadimplente em determinado tempo. Isso significa que são utilizadas informações do tempo no estudo, não apenas se o evento ocorreu ou não.

A variável resposta deste estudo está associada com a inadimplência do cliente. O cliente foi considerado inadimplente quando apresentou atraso do pagamento de uma parcela mais de 60 dias. Dessa forma, o evento de interesse que será considerado neste estudo é a inadimplência do cliente. Sendo assim, a variável resposta é definida como o tempo até o cliente se tornar inadimplente e também é composta pela variável indicadora de censura que dirá se o tempo associado a cada cliente é um tempo de falha (quando o evento de interesse aconteceu) ou tempo de censura (quando por algum motivo não foi observado o evento de interesse).

Esta proposta é de interesse de instituições financeiras no que diz respeito ao envolvimento do risco de crédito. A qualidade de um modelo de risco de crédito é de extrema importância, uma vez que quanto melhor desenvolvido, maior o poder de discriminação e, consequentemente, maior a confiabilidade em conceder crédito a um cliente bom e em não conceder crédito a um cliente mau pagador.

A partir de dados disponibilizados a respeito de empréstimos concedidos por uma instituição financeira brasileira, modelos discretos de sobrevivência serão desenvolvidos devido à característica discreta dos dados. Todas as unidades amostrais serão avaliadas nos

mesmos intervalos de tempo, o que acarreta em um grande número de empates e indica a utilização de técnicas para dados grupados.

2 OBJETIVOS

O intuito deste trabalho é apresentar à instituição financeira uma metodologia alternativa em busca de melhorar a qualidade dos modelos atuais, além de acrescentar informações não conhecidas a respeito dos dados, como o tempo até o cliente se tornar inadimplente.

Dessa forma, objetiva-se ajustar modelos de regressão discretos sob a ótica de dados grupados para modelar dados da instituição financeira cujo produto é uma linha de crédito pacelado com prazo de contratação de dezoito meses.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Crédito

Silva (2008) define crédito como a entrega de um valor presente mediante uma promessa de pagamento. Santos (2003) conceitua crédito como a modalidade de financiamento destinada a possibilitar a realização de transações comerciais entre empresas e seus clientes. Segundo Schrickel (1994), crédito é todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte de seu patrimônio a um terceiro, com a expectativa de que esta parcela volte a sua posse integralmente, após decorrido o tempo estipulado.

Em um banco, que tem como principal atividade o intermédio financeiro, o crédito consiste em colocar à disposição do cliente certo valor sob a forma de empréstimo ou financiamento, mediante promessa de pagamento em uma data futura (Silva, 2008). Em outras palavras, o banco disponibiliza determinado valor em troca de recebê-lo futuramente acrescido de uma taxa pré-definida, a qual é denominada como juro.

Em vista desses conceitos é necessário analisar o perfil do tomador a fim de decidir se a capacidade financeira do mesmo é favorável ou não para arcar com a dívida dentro dos prazos estabelecidos, uma vez que caso isso não ocorra a instituição sofre impacto direto no que diz respeito à perdas financeiras ocasionadas pela inadimplência, por exemplo. Assim, existem critérios que classificam o cliente como bom, mau, ou intermediário.

Em linhas gerais, aquele cliente que apresentam probabilidades maiores de perdas em alguma operação de crédito é qualificado como "mau", usualmente são aqueles que atrasam a parcela sessenta dias ou mais. Já aqueles clientes que não apresentam atrasos nas parcelas, são classificados como "bons". Existe ainda, a possibilidade de classificar clientes como intermediários, no caso em que o tempo de atraso da parcela está entre trinta e sessenta dias.

A análise de crédito envolve a habilidade de fazer uma decisão de crédito, dentro de um cenário de incertezas e constantes mutações e informações incompletas. Esta habilidade depende da capacidade de analisar logicamente situações, não raro, complexas, e chegar a uma conclusão clara, prática e factível de ser implementada (Schrickel, 1994).

Neste contexto é de extrema importância por parte das instituições financeiras, fazer a análise do crédito. Conforme Sicsú (2010) seja o crédito solicitado, seja oferecido pelo credor, sempre existe a possibilidade de perda. Essa probabilidade pode ser considerada como o risco de crédito, pelo qual se baseia a decisão de maneira mais confiável.

3.2 Risco de Crédito e Modelos de *Credit Scoring*

O risco de crédito, ou *credit scoring*, consiste na probabilidade de perda, ou seja, trata-se da probabilidade de conceder crédito a um cliente e o mesmo não honrar com sua dívida. A estimativa dessa probabilidade é obtida a partir de informações do solicitante do crédito, bem como da operação.

Outro ponto de vista é dado por Silva (1993), que defende que o risco de crédito serve para caracterizar os diversos fatores que poderão contribuir para que aquele que concedeu o crédito não receba do devedor na época acordada.

As avaliações do risco são feitas de forma quantitativa, ou subjetiva. Porém, este último método não quantifica o risco de crédito e, portanto, não é tão preciso. Medir o risco de maneira quantitativa é vantajoso no que diz respeito à decisões consistentes, adequadas e mais eficientes, devido ao subsídio computacional.

Dessa forma, são utilizadas ferramentas que estimam essa probabilidade de perda. Isto é, através das fórmulas de cálculo denominadas como modelos de *credit scoring*, obtém-se escores que quantificam o risco (probabilidade de inadimplência), de forma que os gestores passam a ter o subsídio para tomar a decisão de conceder ou não o crédito ao solicitante. A Figura 1 esquematiza o processo, que consiste na obtenção de resultados, a partir de características dos clientes, as quais servirão como base para a avaliação de crédito.

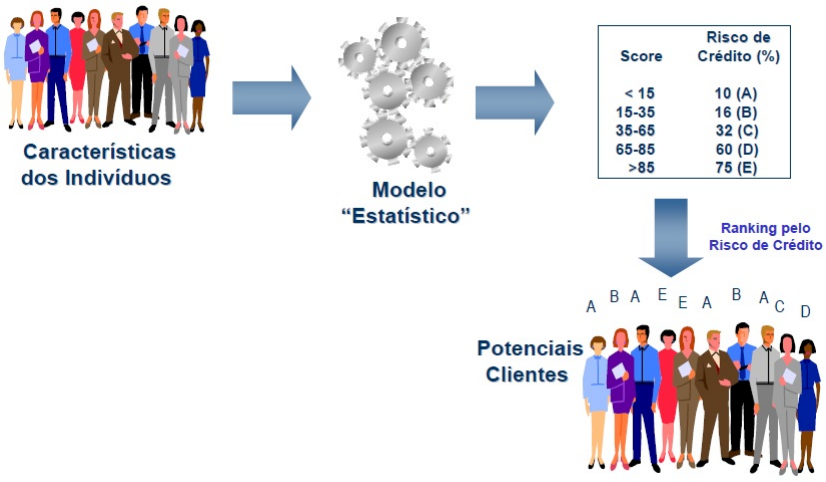

Figura 1 – Ilustração de um Modelo de *Credit Scoring* (Fonte: Louzada-Neto, 2005)

Conforme Zerbini (2000), o modelo de *credit scoring* é uma ferramenta valiosa para decisões de aprovação ou não de pedidos de crédito, obedecendo à hipótese de que o público alvo da carteira de crédito, após a implementação do modelo, se mantenha o mesmo

que no passado recente sobre o qual todo o procedimento estatístico se baseia.

Assim, a ideia principal dos modelos de *credit scoring* é identificar fatores que influenciam na adimplência ou na inadimplência dos clientes, de forma que é possível classificá-los em grupos (por ratings) e, consequentemente, decidir se é viável ou não conceder crédito ao solicitante. Esses modelos podem ser aplicados tanto na análise de crédito de pessoas físicas quanto de pessoas jurídicas.

O roteiro para desenvolvimento, implantação e gestão de um modelo de escoragem segundo Sicsú (2008), consiste, em suma, nas seguintes etapas:

- Planejamento e definições: estabelecer objetivos, o tipo de operação de crédito, mercado-alvo, definir bom e mau cliente, datas e períodos que serão analisados;
- Identificação das variáveis potenciais: identificar variáveis previsoras que tem potencial para discriminar bons e maus clientes;
- Planejamento e seleção da amostra: consiste na coleta de dados de clientes bons e maus que tomaram crédito no passado;
- Análise e tratamento dos dados: verificar se os dados foram coletados corretamente, analisar as características de cada variável individualmente (análise univariada) e analisar a relação entre as variáveis (análise bivariada);
- Cálculo da fórmula de escoragem: aplicar a fórmula do modelo de acordo com a metodologia utilizada;
- Análise e validação da fórmula: avaliação da fórmula baseada em critérios estatísticos e
- Ajuste final do modelo: aperfeiçoar o modelo a partir da análise e validação da fórmula.

O processo de concessão e gestão de crédito envolve não só a necessidade do modelo pelo qual é obtido o escore para tomada de decisão. Conforme Sicsú (2010) também é importante que se tenha uma política de crédito bem definida, um sistema de informações gerenciais com dados do cliente, operação, formas de pagamento, políticas de cobrança, entre outros.

4 METODOLOGIA

Os dados utilizados no estudo são de uma linha de crédito de uma instituição financeira com prazo de contratação de dezoito meses. Devido à inviabilidade de obter informações contínuas dos dados, um intervalo fixo de 30 dias será atribuído. Neste caso, ocorre a situação de dados grupados, uma vez que todas as unidades amostrais serão avaliadas nos mesmos intervalos de tempo.

Conforme Colosimo e Giolo (2006) esse tipo de dados é muitas vezes identificado por um número excessivo de empates. Segundo Hashimoto (2008), a importância dos estudos de dados de sobrevivência grupados se deve a compreensão da natureza dos dados e a forma adequada de tratar o tempo de vida quando há presença de censura e empates.

O ponto de partida da análise de sobrevivência para obter as primeiras informações a respeito dos dados é através da análise exploratória. Neste caso, é desenvolvida por meio de técnicas não-paramétricas, as quais fornecem estimativas para a função de sobrevivência. Desta forma será considerado o estimador de Kaplan Meier, proposto por Kaplan e Meier (1958), pelo fato de ser um estimador de máxima verossimilhança, não viciado e fracamente consistente.

Uma vez que a resposta será dada pelo tempo até a ocorrência de um evento de interesse e por covariáveis, um possível modelo para analisar esses dados é o modelo de Cox. Esse modelo permite estudar o efeito das covariáveis em relação à função taxa de falha, que descreve a distribuição do tempo decorrido até os clientes se tornarem inadimplentes. Como o conjunto de dados possui características de empates, modificações na função de verossimilhança parcial devem ser consideradas para estimar os parâmetros do modelo.

Pelo fato de os dados apresentarem característica de dados grupados, outra metodologia indicada para tratá-los é aplicar modelos de regressão discretos. Nesses modelos a estrutura de regressão é especificada em termos da probabilidade de um indivíduo sobreviver a certo tempo condicional a sua sobrevivência ao tempo anterior. A partir de algumas funções de ligação para modelar a estrutura de regressão, diferentes modelos serão definidos para analisar dados grupados.

Os softwares estatísticos SAS 9.3 e R 2.15.2 serão utilizados como suporte durante todo o desenvolvimento e análise estatística dos dados. O SAS será utilizado devido a parceria acadêmica entre SAS Institute Brasil e o Departamento de Estatística da Universidade de Brasília

4.1 Notação e conceitos básicos em Análise de Sobrevivência

A análise de sobrevivência se refere a um conjunto de metodologias estatísticas que buscam estudar dados relacionados ao tempo decorrido até a ocorrência de um evento de interesse a partir de um tempo inicial, pré-definido. Esse período é designado como tempo de falha e é constituído pelo tempo inicial, que deve ser precisamente definido; a escala de medida, geralmente o tempo real; e o evento de interesse, definido previamente.

No caso da não ocorrência do evento de evento de interesse, ou seja, quando não há falha, os dados referentes são definidos como censurados e resultam em observações parciais ou incompletas. Essas observações devem ser consideradas, visto a capacidade que elas têm em fornecer informações sobre tempo de vida de indivíduos e de evitar que conclusões viciadas sejam obtidas na análise.

Existem mecanismos que diferenciam os tipos de censura quanto ao tempo registrado e ao tempo de falha. A censura à direita é verificada quando o evento de interesse não ocorre até o momento final em que se observa o indivíduo, o tempo de falha está à direita do tempo registrado. A censura à esquerda ocorre quando o evento de interesse acontece em uma data desconhecida e anterior ao início do acompanhamento do indivíduo, o tempo de falha está à esquerda do tempo registrado. Por fim, a censura intervalar é observada quando os dados de sobrevivência são registrados em intervalos de tempo, neste caso os tempos de vida são chamados de duplamente censurados, visto que dados de sobreviência intervalar generalizam qualquer situação em que combinações de tempos de falha e censuras à direita e à esquerda possam ocorrer em um estudo. Um caso particular de censura intervalar são os dados grupados.

A censura à direita é classificada em três formas. Censura do tipo I é observada quando nem todos os indivíduos chegaram a falhar até o final do estudo, o qual é previamente especificado. A Censura do tipo II ocorre quando um número pré-estabelecido de falhas é observado e, assim, informações dos demais indivíduos que participaram do experimento deixam de ser observadas. Já a censura aleatória é a mais comum, visto que ocorre quando se perde informação do indivíduo por motivos não controláveis.

Os dados de sobrevivência são compostos por tempos de falha e de censura, os quais constituem a resposta. Contudo, segundo Colosimo e Giolo (2006), a presença de censuras traz problemas para a análise estatística e, na prática, resultados assintóticos são utilizados para analisar esses dados.

A representação dos dados de sobrevivência é dada pelo par(t_i, δ_i), onde t_i é o tempo de falha ou censura e δ_i é a variável indicadora de falha ou censura. Dessa forma:

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

Assim, duas colunas representam a variável aleatória resposta no banco de dados. No caso em que há presença de covariáveis $\mathbf{x}_i$ referentes ao i-ésimo indivíduo, os dados passam a ser representados como $(t_i, \delta_i, \mathbf{x}_i)$. Ou ainda por $(l_i, u_i, \delta_i, \mathbf{x}_i)$ no caso de sobrevivência intervalar, onde l_i e u_i são os limites inferior e superior do i-ésimo intervalo, respectivamente.

Para estudar dados de sobrevivência, mais precisamente para especificar os tempos de sobrevivência, ou seja, a varável aleatória não-negativa T, três funções são muito utilizadas. São elas a função de sobrevivência $S(t)$, a função densidade de probabilidade $f(t)$ e a função risco $h(t)$.

A função densidade de probabilidade $f(t)$ é definida como o limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $[t, t+\Delta t)$ por unidade de Δt (comprimento do intervalo), ou simplesmente por unidade de tempo. É expressa por (LEE, 1992):

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq T + \Delta t)}{\Delta t}, \tag{1}$$

em que $f(t) \geq 0$ para todo t e a área abaixo da curva de $f(t)$ é igual a 1.

A função de sobrevivência é uma das principais funções probabilísticas usadas para descrever dados de sobrevivência. Ela é monotonicamente decrescente e representa a probabilidade de o indivíduo não falhar, ou seja, do indivíduo sobreviver ao tempo t, conforme segue:

$$S(t) = P(T \geq t) = \int_0^\infty f(x)dx, \tag{2}$$

onde T é uma variável aleatória que representa o tempo, o qual assume valores não-negativos e os valores de $S(t)$ variam entre 0 e 1, uma vez que se trata de uma probabilidade.

A função de sobrevivência é própria quando todos os indivíduos são suscetíveis ao evento de interesse. Já quando não tende a zero à medida que o tempo tende a infinito é dita imprópria e, neste caso, indica que existe uma proporção de indivíduos curados. Dessa forma, a função de sobrevivência juntamente com suas propriedades são muito importantes na identificação de dados com a presença de indivíduos curados.

Como consequência desta definição tem-se a função de distribuição acumulada que é o contrário da função de sobrevivência, ou seja, é dada pela probabilidade de um indivíduo não sobreviver ao tempo t, isto é:

$$F(t) = 1 - S(t). \tag{3}$$

A função taxa de falha ou risco descreve como a taxa de falha se modifica com o passar do tempo. É o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t)$, dado que este indivíduo sobreviveu até o tempo t, dividido pelo comprimento do intervalo e é representada por (LAWLESS, 2003) como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq T + \Delta t | T \geq t)}{\Delta t}. \tag{4}$$

Ou ainda, em termos da função densidade de probabilidade e da função de sobrevivência, isto é:

$$h(t) = \frac{f(t)}{S(t)}. \tag{5}$$

Segundo Colosimo e Giolo (2006) a modelagem da função taxa de falha é um importante método para dados de sobrevivência, visto que é mais informativa do que a função de sobrevivência. Tal afirmação pode ser explicada, pelo fato de que diferentes funções de taxa de falha podem diferir significantemente entre si, enquanto diferentes funções de sobrevivência podem ter formas parecidas. Com isso, a função risco é muitas vezes utilizada para descrever o comportamento dos tempos de sobrevivência.

Uma outra função utilizada para representar o tempo de sobrevivência é a função taxa de falha acumulada que fornece o risco acumulado do indivíduo e pode ser usada para obter $h(t)$ na estimação não-paramétrica. É obtida por meio da função risco:

$$H(t) = \int_0^t h(u)du, \tag{6}$$

e em função da $S(t)$:

$$H(t) = -\log(S(t)). \tag{7}$$

4.1.1 Estimador de Kaplan-Meier

O ponto de partida da análise de sobrevivência para obter as primeiras informações a respeito dos dados é a análise exploratória. Neste caso, é desenvolvida por meio de técnicas não-paramétricas devido à dificuldade em encontrar medidas de tendência central e variabilidade quando há observações censuradas. Essas técnicas fornecem estimativas para a função de sobrevivência e, a partir delas, é possível estimar as estatísticas de interesse, como tempo médio, mediano e percentis.

Os três principais métodos que estimam a função de sobrevivência na presença de censura são: Nelson Aalen, Tabela de Vida ou Atuarial e Kaplan-Meier. O estimador de Kaplan-Meier é um dos mais utilizados para analisar dados de sobrevivência. Segundo Stigler (1994) o artigo do estimador para a função de sobrevivência esteve entre os dois mais citados em toda literatura estatística no período de 1987 a 1989.

O método de Kaplan-Meier é também conhecido como estimador limite-produto e segundo Colosimo e Giolo (2006) é dado por uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº total de observações no estudo}}, \tag{8}$$

sendo $\widehat{S}(t)$ uma função escada a qual tem o tamanho do degrau multiplicado pelo número de empates, caso ocorra em certo tempo t.

O método de Kaplan-Meier considera, na sua construção, o número de intervalos de tempo igual ao número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

Para qualquer tempo t, a função de sobrevivência pode ser escrita em termos de probabilidades condicionais. Suponha um estudo com n indivíduos, neste caso clientes, e k falhas distintas nos tempos $t_1 < t_2 < ... < t_k$. Considerando $S(t)$ uma função com probabilidade maior que zero somente nos tempos de falha $t_j, j = 1, ..., k$ tem-se:

$$S(t) = (1 - q_1) - (1 - q_2)...(1 - q_j), \tag{9}$$

sendo que q_j é a probabilidade de um cliente se tornar inadimplente no intervalo $[t_{j-1}, t_j)$ sabendo que ele não falhou até t_{j-1} e considerando $t_0 = 0$, isto é:

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}). \tag{10}$$

Assim, o estimador de Kaplan-Meier se adapta à expressão (8) de forma que S(t) é escrita em termos de probabilidade condicional:

$$\widehat{q_j} = \frac{\text{nº de falhas em } t_{j-1}}{\text{nº de observações sobre risco em } t_{j-1}}, \tag{11}$$

para $j = 1, ..., k + 1, t_{k+1} = \infty$.

Colosimo e Giolo (2006) definem a fórmula geral do estimador de Kaplan-Meier como:

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right), \tag{12}$$

em que:

- $t_1 < t_2 < ... < t_k$ são os tempos distintos e ordenados de falha;
- d_j é o número de falhas em $d_j, j = 1, ..., k$, e
- n_j é o número de indivíduos que não falharam nem foram censurados até o instante imediatamente anterior a t_j.

Definida a expressão geral do estimador de Kaplan-Meier é importante ressaltar suas propriedades. Ele é não viciado para amostras grandes, é fracamente consistente, converge assintoticamente para um processo gaussiano e é um estimador de máxima verossimilhança.

Comparando com outros estimadores, o método de Kaplan-Meier é benéfico no que diz respeito às suas propriedades e, portanto, é o mais indicado para este estudo. Em relação ao estimador de Nelson-Aalen, o de Kaplan-Meier é preferido para estimar a função de sobrevivência, pois Nelson-Aalen não é um estimador de máxima verossimilhança. Já o estimador da tabela de vida tem certo vício e possui menos intervalos, o que não é interessante para os dados em questão, pois quanto mais intervalos, mais próximo da densidade de probabilidade da variável resposta é possível chegar.

Para estudos que envolvem dados de sobrevivência intervalar, Thurnbull (1976) propôs um outro estimador limite-produto, o qual não tem forma analítica fechada e baseia-se em um procedimento iterativo (Colosimo e Giolo, 2006).

Tomando $0 = \tau_0 < \tau_1 < ... < \tau_m$ como uma sequência de tempos dos pontos contidos em L_i e U_i, $(i = 1, ..., n)$ e definindo um peso α_{ij} de forma que seja igual a 1 caso o intervalo $(\tau_{j-1}, \tau_j]$ esteja contido em $(L_i, U_i]$ e zero caso contrário. É necessário adotar um valor inicial para $S(\tau_j)$ e, então, o algoritmo de Turnbull é obtido definindo, primeiramente, a probabilidade de um evento ocorrer no tempo τ_j , isto é:

$$p_j = S(\tau_{j-1}) - S(\tau_j). \tag{13}$$

Em seguida, é necessário estimar o número de eventos ocorridos no tempo τ_j:

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k}. \tag{14}$$

O próximo passo consiste em obter o número estimado de indivíduos em risco em τ_j :

$$Y_j = \sum_{k=j}^{m} d_k. \tag{15}$$

Por fim, deve-se atualizar o estimador com base nos resultados obtidos em (13) e (14). Repete-se o procedimento iterativo até que a estimativa atualizada de $S(.)$ esteja próxima da anterior para todo τ_j.

Colosimo e Giolo (2006) defendem que este estimador é mais adequado para dados grupados, porém a diferença entre as estimativas obtidas para a função de sobrevivência pelos métodos de Kaplan-Meier e Turnbull é ínfima, conforme observado em Ramos (2013). Dessa forma, ambos estimadores podem ser utilizados neste estudo.

4.2 Modelo de Riscos Proporcionais de Cox

Quando é de interesse para o estudo conhecer a relação entre o tempo de sobrevivência e uma ou mais covariáveis é necessário aplicar metodologias capazes de explorar tal interesse. Um dos principais objetivos é modelar a função de risco e determinar potencias covariáveis que influenciam na sua forma. Outra finalidade é mensurar a função de sobrevivência de cada indivíduo, bem como o risco individual de cada um falhar, que no presente estudo, equivale ao risco de cada cliente se tornar inadimplente.

O modelo de riscos proporcionais de Cox, ou modelo de regressão de Cox analisa dados provenientes de estudos de tempo de sobrevivência em que a variável resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis. A suposição do modelo é a proporcionalidade dos riscos para diferentes níveis de covariáveis, neste caso perfis de clientes.

Neste modelo não é necessário fazer qualquer suposição sobre a distribuição do tempo de sobrevivência. Outros benefícios do modelo são que além de ser um caso particular do modelo Weibull, é flexível devido à presença de um componente não-paramétrico. Desta forma o modelo é dito semi-paramétrico visto que não assume distribuição para a função de risco basal $\lambda_0(t)$. Considerando p covariáveis de modo que $\mathbf{x}$ seja um vetor, o modelo é definido de forma geral como segue:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}'\boldsymbol{\beta}) \tag{16}$$

sendo g uma função especificada, tal que o componente paramétrico $g(0) = 1$ e o componente não-paramétrico $\lambda_0(t)$ deve ser não-negativo no tempo. Segundo Colosimo e Giolo (2006), o componente paramétrico é comumente utilizado na seguinte forma multiplicativa:

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\boldsymbol{\beta}_1 x_1 + ... + \boldsymbol{\beta}_p x_p\} \tag{17}$$

onde $\boldsymbol{\beta}$ é o vetor de coeficientes das covariáveis. Esse formato garante que $\lambda(t|\mathbf{x})$ seja sempre não-negativo. Note que a constante $\boldsymbol{\beta}_0$ presente em outros modelos, não aparece na equação (17). Este termo constante é absorvido devido à presença do componente não paramétrico no modelo.

Sabendo que $S_0(t)$ é a função de sobrevivência base e considerando o modelo de riscos proporcionais de Cox para o tempo de sobrevivência T, a função de sobrevivência desse modelo é dada por:

$$S(t|\mathbf{x}) = \exp\{-\int_0^t \lambda(u|\mathbf{x})du\} = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}. \tag{18}$$

A razão entre as taxas de falha para indivíduos i e j é dada por:

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_j)} = \frac{\lambda_0(t)\exp\mathbf{x}'\boldsymbol{\beta}}{\lambda_0(t)\exp\mathbf{x}_\mathbf{j}'\boldsymbol{\beta}} = \exp\{\mathbf{x}_\mathbf{i}'\boldsymbol{\beta} - \mathbf{x}_\mathbf{j}'\boldsymbol{\beta}\} \tag{19}$$

que não depende do tempo e, portanto, é também denominado como modelo de taxas de falha proporcionais. Daí a suposição inicial do modelo de Cox quanto aos riscos proporcionais.

Em situações em que esse modelo é adequado, é importante verificar a suposição de proporcionalidade, uma vez que sua violação leva a estimativas viciadas dos coeficientes do modelo (Struthers e Kalbfleisch, 1986). Para esta finalidade pode-se fazer a análise gráfica das taxas de falha, de forma que curvas razoavelmente paralelas indicam que são proporcionais, e curvas que se entrelaçam indicam que não há proporcionalidade. Segundo Colosimo e Giolo (2006) a vantagem dessa técnica é que ela indica a covariável que gera a suposição de proporcionalidade, caso isto ocorra. A desvantagem é que a conclusão é subjetiva, pois depende da interpretação dos gráficos. Outra maneira de averiguar se os riscos são proporcionais é através do método com coeficiente dependente do tempo, o qual analisa os resíduos de Schoenfeld (1982). Para maiores detalhes a respeito do método, consultar Colosimo e Giolo (2006).

O ajuste de modelo de Cox envolve a estimação dos coeficientes $\boldsymbol{\beta}'s$ que medem a influência das covariáveis em relação à função de risco. Para fazer essas inferências necessita-se de um método de estimação, como o método de máxima verossimilhança, muito utilizado para este fim.

Contudo esse método é apropriado em caso de modelos paramétricos e sabendo que o modelo de Cox possui uma componente não-paramétrica, seria necessário fazer alterações na função de verossimilhança. A partir desta motivação, Cox (1972) propôs em seu artigo o método de máxima verossimilhança parcial que considera uma probabilidade condicional à história de falhas e de censuras até o tempo i de forma que o elemento não-paramétrico desaparece, tornando possível então obter a função de verossimilhança, isto é:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x_i'}\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x_j'}\boldsymbol{\beta}\}} \right)^{\delta_i}, \tag{20}$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i e δ_i é o indicador de falha.

O modelo de riscos proporcionais e, consequentemente, a função de máxima verossimilhança assumem que os tempos de sobrevivência são contínuos e, sob essa perspectiva, não é possível que ocorra empates entre esses tempos. Porém na prática, é comum se utilizar escalas de tempo discretos, como semanas, meses, ou anos e, assim, observa-se com frequência a ocorrência de tempos de sobrevivência iguais para dois ou mais indivíduos da amostra.

Existem modificações da verossimilhança parcial do modelo de Cox para casos em que o número de empates não é grande. Breslow (1971) e Peto (1972) sugeriram a seguinte

aproximação para a função de verossimilhança parcial:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} = \left(\frac{\exp\{\mathbf{s_i'}\boldsymbol{\beta}\}}{\left[\sum_{j \in R(t_i)} \exp\{\mathbf{x_j'}\boldsymbol{\beta}\}\right]^{d_i}} \right) \tag{21}$$

sendo p o vetor da soma das p variáveis correspondentes aos indivíduos que empataram no tempo $(i = 1, ..., k)$ e d_i o número de falhas neste tempo. Na literatura é possível encontrar outras aproximações para corrigir o problema de muitos empates, como as que foram propostas por Efron (1977), Farewell e Prentice(1980).

O presente estudo trata de produtos de crédito parcelados, os quais são analisados periodicamente e têm como consequência a ocorrência de um grande número de empates na análise. Nesse caso a natureza de tempos de falha discretos é definida e o mais indicado é trabalhar com modelos de regressão discretos. Com base nessa afirmação, apenas a modelagem discreta será considerada na aplicação deste trabalho.

4.3 Dados Grupados

Em análises estatísticas existem situações em que é conveniente estudar os dados separando-os em intervalos, seja devido ao grande número de observações, seja por algum interesse específico. No caso da análise de sobrevivência, quando não se conhece o tempo exato de falha e sim o intervalo em que a mesma ocorreu, define-se esta situação como censura intervalar.

Um caso particular da censura intervalar são os dados grupados, os quais ocorrem quando todas as observações são avaliadas em um mesmo momento, ocasionando portanto, um grande número de falhas no mesmo intervalo. Este resultado é devido à amplitude da unidade de medida, que pode ser dias ou meses, por exemplo. Desse modo uma das principais características aqui é o grande número de empates, o que mostra a natureza discreta dos tempos de falha.

Para tratar esse tipo de dados, alguns autores sugerem ignorar a característica de censura intervalar dos dados considerando o tempo como contínuo de forma que é possível utilizar os métodos tradicionais de análise de sobrevivência. Em contrapartida Rücker e Messerer (1988), Odell; Anderson e D'Agostinho (1992), Dorey e Lindsey e Ryan (1998) advertem que tomar tempos de falhas intervalares como tempos exatos de falha pode conduzir a inferências inválidas.

Colosimo e Giolo (2006) abordam sobre as possibilidades de métodos que tratam esse tipo de dados como discretos no que diz respeito à modelos de regressão discretos e aproximações para a função de verossimilhança parcial no contexto do modelo de riscos proporcionais.

Sob esse cenário, considerar-se-á no presente estudo modelos de regressão dis-

cretos para tratar os dados, uma vez que serão grupados visto que os indivíduos do estudo serão avaliados nos mesmos intervalos de tempos.

4.3.1 Modelos de Regressão Discretos

Segundo Lawless (2003), modelos discretos podem ser utilizados em duas situações. A primeira quando os tempos de vida são propriamente discretos. A segunda quando os tempos contínuos apenas podem ser observados em determinados intervalos, ocasionando em dados grupados de forma equivalente ao caso do presente estudo.

Em termos da estrutura de regressão desses conjuntos de dados, a mesma é especificada em função da probabilidade de um indivíduo sobreviver a um certo intervalo dado que ele havia sobrevivido na visita anterior, conforme Colosimo e Giolo (2006).

Dessa forma, assumindo que as censuras ocorrem no final do intervalo e considerando que os tempos de vida são grupados em k intervalos $I_i = a_{i-1}, a_i$, *i=1,2,..,k* sendo $0 = a_0 < a_1 < ... < a_k = \infty$, de forma que R_i representa o conjunto do observações sob risco no tempo a_{i-1} e δ_{li}, uma variável indicadora de falha e censura tal que $\delta_{li} = 1$ se ocorreu falha do l-ésimo indivíduo no I_i-ésimo intervalo e $\delta_{li} = 0$, caso contrário. Seja T o tempo de vida de n indivíduos e $\mathbf{x}_l$ o vetor de covariáveis regressoras, a probabilidade do l-ésimo indivíduo falhar até a_i dado que ele não falhou até a_{i-1} é dada por:

$$p_i(\mathbf{x}_l) = P[T_l < a_i \mid T_l \geq a_{i-1}, \mathbf{x}_l], \tag{22}$$

vale ressaltar que em termos da função de sobrevivência $S(.)$ a expressão acima é escrita como:

$$p_i(\mathbf{x}_l) = P[a_{i-1} \leq T_l < a_i \mid T_l \geq a_{i-1}, \mathbf{x}_l] = 1 - \frac{S_0(a_i)}{S_0(a_{i-1})} \tag{23}$$

Assim, a contribuição de uma observação não censurada no intervalo I_i para a função de verossimilhança é (Strapasson,2007):

$$\begin{aligned} P[a_{i-1} \leq T_l < a_i | x_l] &= S(a_{i-1}|x_l) - S(a_i|x_l) \\ &= [\{1 - p_1(\mathbf{x}_l)\}...\{1 - p_{i-1}(\mathbf{x}_l)\}]p_i(\mathbf{x}_l) \end{aligned} \tag{24}$$

e a contribuição de uma observação censurada em a_i para a função e verossimilhança é:

$$\begin{aligned} P[T_l \geq a_i | x_l] &= S(a_i|x_l) \\ &= [\{1 - p_1(\mathbf{x}_l)\}...\{1 - p_i(\mathbf{x}_l)\}]. \end{aligned} \tag{25}$$

Portanto, com base nas equações (24) e (25) a função de verossimilhança é escrita, conforme Colosimo e Giolo (2006), da seguinte forma:

$$\prod_{i=1}^{k} \prod_{l \epsilon R_i} \{p_i(\mathbf{x}_l)\}^{\delta_{li}} \{1 - p_i(\mathbf{x}_l)\}^{1-\delta_i}, \tag{26}$$

a qual vem de uma função de máxima verossimilhança da distribuição Bernoulli, em que δ_i é a variável resposta e $p_i(\mathbf{x}_l)$ é a probabilidade de sucesso. A equação (26) pode ser modelada usando diferentes funções de ligação na probabilidade $p(x)$.

4.3.2 Modelos de Riscos Proporcionais

Segundo Hashimoto (2008) é possível modelar a estrutura de regressão dada em termos de $p_i(\mathbf{x}_l)$ através de diferentes funções de ligação, que de mandeira geral é representada por:

$$p_i(\mathbf{x}_l) = g(\eta_{li}) \tag{27}$$

para $l = 1, ..., n$ e $i = 1, ..., k$ e $\eta_{li} = \gamma_i + \mathbf{x}_l^T\boldsymbol{\beta}$, sendo $g(.)$ uma função estritamente monótona e duplamente diferenciável que relaciona as variáveis independentes com um preditor linear e $\boldsymbol{\beta}$ o vetor de parâmetros associados a cada covariável.

Algumas funções de ligação são utilizadas frequentemente em estudos que envolvem análise de sobrevivência para dados grupados, como a função complemento log-log, logito e probito. O benefício em utilizá-las está relacionado com a interpretação simplória dos parâmetros fornecidos pelas mesmas.

Para cada ligação citada, a função $g(.)$ assume as seguintes formas:

- Ligação complemento log-log

$$g(\eta_{li}) = 1 - \exp[-\exp(\eta_{li})]. \tag{28}$$

Ao utilizar essa função de ligação obtem-se o modelo de riscos proporcionais para dados grupados. Esse fato pode ser observado ao assumir o modelo de riscos proporcionais de Cox para o tempo de vida T. A função de sobrevivência para esse modelo, como definida em (18), tem a seguinte forma:

$$S(t|\mathbf{x}) = \exp\{-\int_0^t \lambda(u|\mathbf{x})du\} = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}. \tag{29}$$

Assim, $p_i(\mathbf{x}_l)$ é definida como:

$$p_i(\mathbf{x}_l) = 1 - \left[\frac{S_0(a_i)}{S_0(a_{i-1})}\right]^{\exp\{\mathbf{x}_\mathbf{l}'\boldsymbol{\beta}\}}. \tag{30}$$

e pode ser reescrita por:

$$p_i(\mathbf{x}_l) = 1 - \gamma_i^{\exp\{\mathbf{x}_\mathbf{l}'\boldsymbol{\beta}\}}. \tag{31}$$

O modelo (31) pode ser linearizado ao utlizar a transformação complemento log-log. Isto é,

$$\log[-\log\{1 - p_i(\mathbf{x}_l)\}] = \gamma_i^* + \mathbf{x}_\mathbf{l}'\boldsymbol{\beta} = \eta_{li}, \tag{32}$$

em que $\gamma_i^* = \log(-\log\gamma_i)$ é o efeito do i-ésimo intervalo e η_{li} é o preditor linear.

- Ligação logito

$$g(\eta_{li}) = \frac{[\exp(\eta_{li})]}{1+\exp(\eta_{li})}. \tag{33}$$

Ao utilizar essa função de ligação obtem-se o modelo logístico para dados grupados. Esse fato pode ser observado ao assumir o seguinte modelo para o tempo de vida T:

$$p_i(\mathbf{x}_l) = 1 - (1+\gamma_i \exp\{\mathbf{x_l}'\boldsymbol{\beta}\})^{-1}. \tag{34}$$

em que $\gamma_i = p_i(0)/1 - p_i(0)$, para $i = 1, ..., k$. O modelo (34) pode ser linearizado utilizando-se a transfrmação *logito*, de forma que:

$$\log \frac{p_i(\mathbf{x}_l)}{1-p_i(\mathbf{x}_l)} = \gamma_i^* + \mathbf{x_l}'\boldsymbol{\beta} = \eta_{li}. \tag{35}$$

- Ligação probito

$$g(\eta_{li}) = \Phi(\eta_{li}), \tag{36}$$

em que $\Phi(n_{li})$ é função distribuição acumulada da distribuição normal padrão.

O estimador de máxima verossimilhança é obtido em termos das funções de ligação, basta modelá-lo substituindo a probabilidade $p_i(\mathbf{x}_l)$ pela forma da ligação de interesse na função de máxima verossimilhança definida em (26).

6 CONCLUSÃO

O processo de avaliação de crédito antigamente era feito de maneira muito lenta e subjetiva, sendo influenciado pela particularidade de cada analista. Com a popularização e crescimento desse mercado, foi necessário buscar meios de padronizar o sistema de avaliação e melhorar sua qualidade de forma que se tornasse mais prático e preciso, levando em consideração a rentabilidade e a importância do controle da inadimplência.

A partir do seu desenvolvimento, o processo de avaliação de risco eliminou qualquer subjetividade envolvida, melhorando a qualidade e acurácia dos modelos. Esse resultado foi possível graças ao suporte de técnicas estatísticas como regressão logística, árvores de decisão e análise discriminante.

Atualmente é de interesse por parte das instituições melhorar suas análises buscando metodologias cada vez mais adequadas para diminuir a inadimplência e melhorar a rentabilidade. Sob essa perspectiva, é comum encontrar estudos relacionados à sugestões de novos métodos de modelagem de *Credit Scoring* que chamam a atenção das instituições.

Este estudo utilizou uma técnica em presente ascensão na área de risco de crédito: a análise de sobrevivência. Esse método fornece, como resultante do modelo, a probabilidade de ocorrência de um evento associada a cada instante ao longo do horizonte de previsão. No caso dos modelos aqui realizados, esse evento foi associado à inadimplência. Outro ponto levado em consideração na modelagem foi a característica discreta dos dados pelo o grande número de tempos de falha empatados, o que define a ótica de dados grupados. Sob essa perspectiva, três modelos de regressão foram desenvolvidos de acordo com cada função de ligação: complemento log-log, logito e probito.

Antes dos modelos serem ajustados, alguns cuidados iniciais foram tomados em relação aos dados com o intuito de otimizar os resultados e evitar erros. Depois de tratar os dados e analisá-los exploratoriamente, ajustou-se os modelos e os resultados foram concordantes para todos, levando a conclusões iguais e estimativas aproximadas.

A variável renda não foi significativa, mas poderia ser mantida nos modelos, o que acarretaria em perda na qualidade dos mesmos. Foi observado na análise preliminar que renda e a classificação dos clientes como bons ou maus eram independentes. Essa falta de associação pode ter sido a causa da variável não ter sido significante na modelagem.

O presente trabalho foi muito voltado para a exploração dos dados e ao estudo do comportamento dos clientes em relação ao tempo e às covariáveis. A técnica de análise de sobrevivência e as suas interpretações foram bastante exploradas com o intuito de mostrar as ferramentas da análise de sobrevivência e os benefícios que a mesma pode trazer para os modelos da instituição.

Não foi possível mensurar a qualidade do ajuste dos modelos com base nas

estatísticas KS, AUROC, GINI, pois não existe um forma fechada para obtenção desses valores na metodologia utilizada de dados grupados. Isso porque a probabilidade de falha de cada indivíduo é calculada em cada intervalo, ou seja, o mesmo cliente tem diferentes probabilidades de inadimplência, de acordo com cada intervalo. Contudo, testou-se a qualidade do ajuste dos modelos utilizando o teste de Hosmer e Lemeshow e o resultado mostrou que os ajustes dos modelos foram adequados.

A grande diferença da análise de sobrevivência e a regressão logística, técnica mais utilizada nas intituições financeiras atualmente, é a capacidade que ela tem de estimar o tempo de sobrevivência, não só a probabilidade de inadimplência. Essa informação do tempo pode ser muito útil nos modelos de *Credit Scoring*. O presente trabalho teve como objetivo levar essa ideia do uso da análise de sobrevivência para a instituição financeira, com o intuito de propor a aplicabilidade de uma metodologia nova e eficaz.

A característica dos dados da linha de crédito utilizada permite ainda que metodologias mais específicas sejam aplicadas para o ajuste dos modelos. O indicativo de fração de cura é forte, segundo os gráficos das funções de sobrevivência estimados pelo método de Kaplan Meier. Assim, propõe-se para trabalhos futuros o desenvolvimento de modelos baseados na técnica de fração de cura.

APÊNDICE

Apêndice A: Programação

```
/* Dados grupados - Obtém intervalos */
data intervs;
      retain interv1-interv10 0;
      array dd[10] interv1-interv10;
      set base.base;
      if tempo = 10 then do interv=1 to 10;
         y=0; dd[interv]=1;
         output;
         dd[interv]=0;
      end;
      else do interv=1 to tempo;
         if interv=tempo then y=1;
         else y=0;
         dd[interv]=1;
         output;
         dd[interv]=0;
      end;
  run;

/* Modelo Função Complemento log-log */;
 proc logistic data=intervs descending outest=est1;
 class idade estcivil grauinst valor / param=reference ref=first;
      model y=  interv1-interv10 idade estcivil grauinst valor /lackfit noint
                                    link=cloglog technique=newton;
output out=resultados predicted=prob;
run;

/* Modelo Função Logit */;
 proc logistic data=intervs descending outest=est1;
 class idade estcivil grauinst valor / param=reference ref=first;
      model y=  interv1-interv10 idade estcivil grauinst valor /lackfit noint
                                    link=logit technique=newton;
output out=resultados predicted=prob;
run;

/* Modelo Função Probito */;
 proc logistic data=intervs descending outest=est1;
 class idade estcivil grauinst valor / param=reference ref=first;
      model y=  interv1-interv10 idade estcivil grauinst valor /lackfit noint
                                    link=probit technique=newton;
output out=resultados predicted=prob;
run;
```

REFERÊNCIAS

ALLISON, P. D. **Survival Analysis Using the SAS System: a Practical Guide**. Cary, NC, USA: SAS institute Inc., 1995.

BUSSAB, W. O, MORETTIN, P. A. **Estatística básica**. São Paulo: Saraiva, 2005, 5ª edição.

CARAVALHO, M. S. ANDEREOZZI, V.L.; CODEÇO, C. T; CAMPOS, D.P; BARBOSA, M. S.; SHIMAKURA, S. E. **Análise de sobrevivência: Teoria e aplicações em saúde**. FIOCRUZ, RIO DE JANEIRO, 2011.

COLOSIMO, Enrico Antonio; GIOLO, Suely Ruiz. **Análise de sobrevivência aplicada**. São Paulo: E. Blucher, 2006. Xv, 369p.

COX, D. R. **Regression Models and Life Tables (with discussion)**. Journal Royal Statistical Society, 1972

DINIZ, C; Louzada, F. **Modelagem estatística para risco de crédito. In: Simpósio Nacional de probabilidade e estatística**. 20, 2012, João Pessoa – PB. São Paulo: ABE – Associação Brasileira de Estatística. 2012

EFRON, B.**The Efficiency os Cox's Likelihood Function for Censored Data**. Journal of the American Statistical Association. 1977.

FACHINI, J. B. **Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência**. 2011. p. 64-65. Tese (Doutorado em Ciências) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba 2011.

FAREWELL, V. T e PRENTICE, R. L.**The aApproximation of Partial Likelihood with Emphasis on Case-Control Studies**. Biometrika. 1980.

HASHIMOTO, E. M. **Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados**. 2008. p. 121. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba 2008.

GRAMBSCH, P. M.; THERNEAU, T. M. **Proportional Hazards Tests and Diagnostics based on Weighted Residuals**. Biometrika, 1994.

KAAPLAN, E.L.; MEIER, P. **Nonparametric estimation from incomplete observations**. Journal of the American Statistical Association, 53, 457-81.

KALBFLEISH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data**. 2nd ed. New York: John Wiley, 2002. 439 p.

HOSMER, David W.; LEMESHOW, Stanley; MAY, Susanne. **Applied survival analysis: regression modeling of time-to-event data** . 2nd ed. Hoboken, N.J.: Wiley-Interscience, c2008. xiii, 392 p.

LAWLESS, J. F. **Statistical Methods and Models for Lifetime Data**. John Wiley Sons, New York, 2003.

LINDSEY, J. C; RYAN, L. M. **Tutorial in biostatiscs methods for interval-censored data**. Statiscs in Medicine, Chichestes, 1998

LEE, E. T. **Statistical Methods for Survival Data Analysis**.Lifetime Learning Publications, New York, 1992

LOUZADA-NETO, F. **Análise de sobrevivência aplicada ao *Credit Scoring***. Seminário internacional de Credit Scoring Serasa. 2005.

MACHADO, A. R. **Modelos estatísticos para avaliação de risco em produtos de crédito parcelados**. 2010. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade de Brasília, Brasília 2010.

ODELL, P.M.; ANDERSON, K.M.; D'AGUSTINHO,R.B.**Maximum likelihood estimation for interval-censored data using a weigbull-based failure time model**. Biometrics, Washington, DC. 1992.

EFRON, B.**The Efficiency os Cox's Likelihood Function for Censored Data**. Journal of the American Statistical Association. 1977

PEREIRA, C. G. **Análise de crédito bancário: um sistema especialista com técnicas difusas para os limites da agência**. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolls 1995.

RAMOS, A. L. **Análise de Sobrevivência com dados grupados: Uma Aplicação na ocorrência de Hipotensão em Idosos**. 2010. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade de Brasília, Brasília 2013.

RÜCKER, G.;MESSERER, D. **Remission duration: an example of interval-censored observations**. Statistics in Medicine, Chichester. 1988

SANTOS, José Odálio. **Análise de Crédito – Empresas e Pessoas Físicas**. SãoPaulo. 2ª edição, 2003 – Editora Atlas

SICSÚ, A. L. **Crédit Scoring: desenvolvimento, implantação e acompanhamento**. São Paulo: Blucher, 2010.

SCHRICKEL, W.K. **Análise de Crédito: Concessão e gerência de empréstimos**. São Paulo, Atlas, 1994.

SCHOFELD, D. A. **Partial Residuals for the Proportional Hazard Regression Model**. Biometrika, 1982.

SILVA, J. P. **Análise de decisão de crédito**. São Paulo: Atlas, 1993

SILVA, J. P. **Gestão e Análise de Risco de Crédito**. São Paulo: Atlas, 2008.

STIGLER, S. M. **Citation Patterns in Journals of Statistics and Probability**.

Statiscal Science. 1994)

STRAPASSON, E.**Comparação de Modelos com Censura Intervalar em Análise de Sobrevivência**. 2008. p. 121. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba 2007.

STRUTHERS, C. A; KALBFLEISCH, J. D.**Misspecified Proportional Hazards Models**. Biometrika. 1986)

TURNBULL, B. W.**Nonparametric Estimation of a Surviorship Function with doubly Censored Data**. J.R. *Statist. Soc.* B, 38, 290-295. 1976)

WIENKE, A; LICHTENSTEIN, P; YASHIN, A. I. **A Bivariate Fraity Model with a Cure Fraction for Modeling Familial Correlations in Deseases**. Max Planck Institute for Demographic Research, Rostock, Germany, p. 1178-1179, 2003.

ZERBINI, M. B. A. A. **Três Ensaios Sobre Crédito**. Tese (Doutorado em Administração) - FEA-USP, São Paulo, 2000.