Universidade de Brasília
Departamento de Estatística

Modelagem de risco de crédito via modelo de chances de riscos proporcionais

Filipe Oliveira do Vale Ribeiro

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

Brasília
2025

Filipe Oliveira do Vale Ribeiro

Modelagem de risco de crédito via modelo de chances de riscos proporcionais

Orientador(a): Prof.º Eduardo Yoshio Nakano

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

Brasília
2025

Filipe Oliveira do Vale Ribeiro

Credit scoring modeling using Proportional Odds Hazard Model

Advisor: Prof.º Eduardo Yoshio Nakano

Dissertation presented to the Department of Statistics of the University of Brasília as part of the requirements for obtaining the degree of Master in Statistics.

Brasília
2025

Agradecimentos

Gostaria de agradecer ao Professor Doutor Eduardo Yoshio Nakano pela sua orientação e a todos os docentes que me formaram desde as etapas mais simples da escola aos meus atuais formadores na universidade.

Um muito obrigado também aos meus colegas de curso pelos dias e noites de estudos e pelos diversos momentos de alegria na Universidade de Brasília. E um agradecimento muito especial à minha família por terem me dado todas as condições para ter chegado aonde cheguei.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Este trabalho investiga a gestão do risco de inadimplência na concessão de crédito, uma preocupação central para instituições financeiras. O objetivo principal foi aplicar o Modelo de Chances de Riscos Proporcionais (MCRP) para a modelagem do risco de crédito, desenvolvendo um escore capaz de classificar clientes com base em sua probabilidade de inadimplência. O MCRP foi escolhido por sua flexibilidade em lidar com a natureza discreta do tempo até a inadimplência e por sua capacidade de extrair informações detalhadas sobre o comportamento do cliente. A metodologia foi validada com dados simulados, demonstrando que o escore de risco proposto é robusto, apresenta desempenho consistente e superior a técnicas tradicionais como o modelo logístico. Conclui-se que o escore é uma contribuição eficaz para a literatura de modelagem de risco de crédito, oferecendo uma ferramenta prática para a tomada de decisões na concessão de crédito.

Palavras-chave: Escore de Crédito, Inadimplência, Modelo de Chances de Riscos Proporcionais, Análise de Sobrevivência, Tempo Discreto, Escore de Risco.

Abstract

This study addresses the critical issue of default risk in credit granting, a central concern for financial institutions. The main objective was to apply the Proportional Odds Hazard Model (POHM) for credit risk modeling, developing a score capable of classifying clients based on their probability of default. The POHM was chosen for its flexibility in handling the discrete nature of time-to-default and its ability to extract detailed information about customer behavior. The methodology was validated with simulated data, demonstrating that the proposed risk score is robust, exhibits consistent performance, and surpasses traditional techniques like the logistic model. It is concluded that the score is an effective contribution to the credit risk modeling literature, offering a practical tool for credit granting decision-making.

Keywords: Credit Scoring, Default, Proportional Odds Hazard Model, Survival Analysis, Discrete time-to-event, Risk Score.

Sumário

1 Introdução 8

2 Conceitos Básicos em Análise de Sobrevivência 9

2.1 Introdução 9

2.2 Representação do tempo de sobrevivência 10

2.2.1 Representação para tempo contínuo 10

2.2.2 Representação para tempo discreto 13

2.3 Tempos discretos 15

2.3.1 Discretização de distribuições contínuas 16

2.3.2 Distribuição Weibull Discreta 17

2.3.3 Distribuição log-logística discreta 18

2.4 Obtenção de estimadores 19

2.4.1 Estimador de Kaplan-Meier de S(t) 19

2.4.2 Estimadores de Máxima verossimilhança 20

2.5 Modelo de riscos proporcionais de Cox 21

2.5.1 Formulação do modelo 22

2.5.2 Estimação dos parâmetros 24

2.5.3 Avaliação do modelo (verificação do ajuste) 25

3 Modelo de chances de riscos proporcionais (MCRP) 27

3.1 Formulação do modelo 27

3.2 MCRP Weibull discreto 29

3.3 Verificação da suposição de chances de riscos proporcionais 30

4 Modelagem de risco de crédito 32

4.1 Introdução 32

4.2 Escore de risco 35

4.3 Classificação dos clientes pelo escore de risco 35

4.4 Avaliação da acurácia do modelo 36

5 Ilustração da metodologia proposta 38

5.1 Banco de dados 38

5.2 Análise descritiva 40
5.3 Ajuste do MCRP 41
5.4 Obtenção do escore de risco e classificação dos indivíduos segundo a metodologia proposta 43
5.5 Comparativo de Desempenho Preditivo 46
6 Considerações finais 48
Referências 49
Apêndice 52

1 Introdução

A concessão de crédito desempenha um papel crucial nas economias desenvolvidas. A alocação eficiente desse capital em empreendimentos rentáveis impulsiona o crescimento econômico, fomentando a criação de produtos e serviços que atendem às demandas do mercado.

Contudo, credores se preocupam com o risco envolvido nessa operação, visto que os clientes podem não conseguir honrar o pagamento de seus empréstimos, por diversos motivos. Sendo assim, é de interesse dessas instituições financeiras avaliar o risco associado ao cliente, antes de fazer a concessão do crédito, para evitar casos de inadimplência e, assim, tornar o negócio lucrativo. Sob esse cenário, surgiram os Modelos de *Credit Scoring*, como ferramenta capaz de quantificar o risco de crédito envolvido em uma operação.

Este trabalho tem como objetivo principal a modelagem do risco de crédito utilizando o modelo de chances de riscos proporcionais (MCRP), proposto por Vieira et al. (2023). Esse modelo permite uma análise mais detalhada das variáveis que influenciam a inadimplência, considerando tanto aspectos temporais quanto características específicas dos clientes. A aplicação dessa metodologia proporciona uma avaliação mais precisa do risco, auxiliando instituições financeiras na tomada de decisões estratégicas quanto à concessão de crédito. Mais especificamente, será proposto um escore que medirá o risco de inadimplência de um cliente. Este escore será calculado a partir das estimativas do MCRP e utilizado para classificar os clientes em mal ou bons pagadores.

Além disso, o uso do MCRP oferece vantagens significativas em termos de flexibilidade e interpretação dos resultados. Ao contrário de modelos tradicionais, que frequentemente assumem uma distribuição contínua para o tempo até a inadimplência, o modelo de chances de riscos proporcionais não impõe essa restrição, permitindo uma análise mais realista e adaptada às características dos dados disponíveis. A metodologia proposta será ilustrada em um conjunto de dados artificiais que imitam dados presentes na literatura.

A estrutura deste estudo é organizada da seguinte maneira: primeiramente, apresenta-se uma revisão da literatura sobre modelagem de risco de crédito, juntamente com os fundamentos teóricos do modelo de Cox. Em seguida, detalha-se a metodologia empregada, incluindo a descrição dos dados utilizados e as técnicas de análise aplicadas. Posteriormente, os resultados obtidos são apresentados e discutidos, com ênfase nas implicações práticas para a gestão de risco em instituições financeiras. Finalmente, são apresentadas as considerações finais e sugestões para pesquisas futuras.

2 Conceitos Básicos em Análise de Sobrevivência

2.1 Introdução

A Análise de Sobrevivência é uma classe de métodos estatísticos utilizada para análise de dados no qual o foco de interesse é o tempo até ocorrência de determinado evento de interesse. Essa técnica é muito utilizada na área médica para estudos sobre morte, mas ganhou aplicações em vários outros campos do conhecimento, como sociologia na análise histórica de eventos, engenharia com análise do tempo de vida de equipamentos e na economia com análise de inadimplência.

As técnicas de Análise de Sobrevivência têm ganhado destaque em diversas áreas, pois, além de identificar a ocorrência ou não de um evento de interesse, estimam o momento em que tal evento ocorre, permitindo situá-lo temporalmente.

De acordo com Colosimo e Giolo (2006), uma característica crucial nesse conjunto de técnicas é a presença de censura, que se refere à possibilidade de informações incompletas nos dados, ou seja, quando uma observação não é acompanhada até a ocorrência do evento de interesse. Um exemplo na área econômica seria quando o acompanhamento dos depósitos de crédito de um cliente de um banco termina para análise, e não ocorre o evento de inadimplência. Isso implica que toda a informação disponível se resume ao conhecimento de que o evento não aconteceu durante o período observado, deixando em aberto o momento em que ele poderia ocorrer após o término do acompanhamento.

Considerando a censura, é preciso incluir uma variável que identifique se o tempo da ocorrência do evento foi observado ou não. Essa variável é conhecida na literatura como variável indicadora de censura ou falha. É importante destacar que, mesmo nos casos em que ocorre censura, todos os resultados do estudo devem ser considerados na análise estatística, pois a exclusão da censura nos cálculos estatísticos pode levar a estimativas distorcidas.

A classificação dos tipos de censura pode ser dividida em três categorias principais: censura à direita, censura à esquerda e censura intervalar. Na censura à direita, o tempo registrado (ou censurado) é menor do que o tempo que teria sido observado se não houvesse interrupção. Por outro lado, na censura à esquerda, o evento de interesse ocorre antes mesmo do tempo ser registrado, ou seja, o tempo registrado é maior do que o tempo de falha.

Por fim, na censura intervalar, não é possível identificar o tempo exato em que ocorreu a falha; apenas conhecemos um intervalo de tempo em que o evento de interesse ocorreu.

Assim, é necessário incluir uma variável dicotômica na análise para indicar se o tempo de sobrevida de um indivíduo foi observado ou não. Essa variável, chamada de variável indicadora de censura ou simplesmente censura, é definida como igual a um se o tempo de sobrevida é observado e igual a zero se o tempo de sobrevida é censurado.

2.2 Representação do tempo de sobrevivência

Na Análise de Sobrevivência, busca-se estimar o comportamento da variável aleatória tempo de sobrevivência, $T > 0$, o que pode ser aprimorado com o uso de variáveis explicativas. O comportamento da variável resposta pode ser expresso por meio de várias funções equivalentes. Essas funções, como a função de sobrevivência ou a função de risco (ou taxa de falha), são utilizadas para descrever diferentes aspectos do tempo de sobrevivência, que pode ser discreto ou contínuo. Se uma dessas funções é especificada, as outras podem ser derivadas.

2.2.1 Representação para tempo contínuo

Função densidade de probabilidades

Considere T como uma variável aleatória não negativa e contínua. A Função Densidade de Probabilidade (FDP) de T, denotada por $f(t)$, é uma função que satisfaz as seguintes condições (MEYER, 1983):

1. $f(t) \geq 0$ para todo $t \geq 0$
2. $\int_0^\infty f(t)\, dt = 1$
3. $P(a \leq T \leq b) = \int_a^b f(t)\, dt \ , \forall\ 0 \leq a \leq b.$

Essa função pode ser interpretada como o limite da probabilidade de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t+\Delta t)$, dividida pela duração do intervalo e pode ser expressa por:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < \Delta t)}{\Delta t},\ t \geq 0. \tag{2.2.1}$$

Função de sobrevivência

Denotada por $S(t)$, a função de sobrevivência representa a probabilidade de um indivíduo não apresentar o evento de interesse até um dado instante t, ou seja, a probabilidade de que o indivíduo sobreviva além desse tempo t. Esta função é uma das principais funções probabilísticas utilizadas para descrever o tempo de sobrevivência e é definida por:

$$S(t) = P[T > t] = \int_t^{\infty} f(u)\, du\ ,\ t \geq 0. \tag{2.2.2}$$

A função de sobrevivência (2.2.2) é uma função não crescente e absolutamente contínua, tal que $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = 0$

Função de risco

A Função de Risco, também denominada função taxa de falha e denotada por $h(t)$, representa o risco instantâneo de um indivíduo apresentar o evento de interesse em um dado instante t. Para uma variável aleatória contínua, esta função é definida como o limite da razão da probabilidade condicional de um indivíduo experienciar o evento de interesse no intervalo de tempo $[t, t + \Delta t)$, dado que não tenha experienciado o evento antes de t, pelo intervalo de tempo Δt. A função $h(t)$ é expressa por:

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},\ t \geq 0. \tag{2.2.3}$$

No caso de variáveis aleatórias contínuas, a função de risco $h(t)$ assume valores reais positivos e não possui um limite superior.

Função de risco acumulado

Outra função importante derivada da função $h(t)$ é a Função de Risco Acumulada, ou Taxa de Falha Acumulada, representada por $H(t)$. Embora $H(t)$ não tenha uma interpretação direta, ela é útil em procedimentos de estimação não-paramétricos e na escolha do modelo mais adequado para ajustar um conjunto de dados específico. A função $H(t)$ fornece o risco acumulado até o tempo t e, no caso de uma variável aleatória contínua, é definida por:

$$H(t) = \int_0^t h(u)\, du\ ,\ t \geq 0. \tag{2.2.4}$$

Principais relações entre as funções $f(t)$, $S(t)$, $h(t)$ e $H(t)$

Apresentam-se, a seguir, algumas relações matemáticas importantes entre as funções de densidade, sobrevivência e risco. Tais relações podem ser utilizadas para determinar uma das funções, dado o conhecimento de outra. Observe que:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(\{t \leq T < t + \Delta t\} \cap \{T \geq t\})}{\Delta t P(T \geq t)} \\ &= \frac{\lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}}{P(T > t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{2.2.5}$$

A função densidade de probabilidades, $f(t)$, é definida como a derivada da Função de Distribuição Acumulada, $F(t)$, isto é,

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

Visto que $F(t) = 1–S(t)$, tem-se que

$$f(t) = \frac{d}{dt}[1 - S(t)] = -\frac{d}{dt} S(t) = -S'(t). \tag{2.2.6}$$

Uma vez que $\frac{d}{du} log(u) = \frac{u'}{u}$

substituindo (2.2.5) em (2.2.6) obtém-se que

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} log S(t). \tag{2.2.7}$$

Consequentemente, integrando ambos os termos (2.2.8) resulta em

$$log S(t) = -\int_0^t h(u)\, du = -H(t),$$

o que implica em

$$S(t) = exp\left\{ -\int_0^t h(u)\, du \right\} = exp\left\{ -H(t) \right\}. \tag{2.2.8}$$

De (2.2.6) tem- se que

$$f(t) = h(t) S(t).$$

Desta forma, ao substituir (2.2.8) em (2.2.6) resulta em

$$f(t) = h(t) exp\left\{ -\int_0^t h(u)\, du \right\}. \tag{2.2.9}$$

2.2.2 Representação para tempo discreto

Função de probabilidade

Considere T como uma variável aleatória discreta que assume valores inteiros não negativos, ou seja, $t = 0, 1, 2, \ldots$ A função de probabilidade, ou distribuição de probabilidade de T, é uma função que associa a cada possível valor da variável aleatória sua respectiva probabilidade. Essa função, denotada por $p(t) = P(T = t)$, deve satisfazer as seguintes condições (JAMES, 2015):

1. $0 \leq p(t) \leq 1$ para todo $t = 0, 1, 2, \ldots$ e
2. $\sum_{t=0}^{\infty} p(t) = 1$.

Função de sobrevivência

Definida como a probabilidade de um indivíduo não apresentar o evento de interesse até um dado instante t, a função de sobrevivência, no caso em que T é uma variável aleatória discreta, é expressa por:

$$S(t) = P[T > t] = \sum_{k=t+1}^{\infty} p(k) = \sum_{k=t+1}^{\infty} P(T = k) \ , t = 0, 1, 2, \ldots \ . \tag{2.2.10}$$

A função de sobrevivência é uma função definida em todos os reais não negativos.

Função de risco

No caso de uma variável aleatória discreta, a Função de Risco (ou função taxa de falha) é definida como a probabilidade condicional de um indivíduo apresentar o evento de interesse no instante t, dado que não o tenha apresentado antes de t. Assim, temos:

$$h(t) = P(T = t | T \geq t), \ t = 0, 1, 2, \ldots \ . \tag{2.2.11}$$

Observe que, para os valores de t que são negativos ou não inteiros, a função de risco (2.2.11) é igual a zero. Além disso, por se tratar de uma probabilidade condicional, a função de risco para variáveis aleatórias discretas é limitada ao intervalo $[0, 1]$.

Função de risco acumulado

A função de risco acumulado, $H(t)$, representa o risco acumulado do indivíduo até o tempo t. No caso de uma variável aleatória discreta, essa função é definida por

$$H(t) = \sum_{k=0}^{t} h(k),\ t = 0, 1, 2, \dots\ . \tag{2.2.12}$$

A função de risco acumulado (2.2.12) não tem interpretação direta e é uma função que assume valores reais positivos, não sendo limitada superiormente.

Principais relações entre as funções $f(t)$, $S(t)$, $h(t)$ e $H(t)$

Descrevem-se, a seguir, algumas relações matemáticas importantes entre as funções de densidade, sobrevivência e risco de uma variável aleatória discreta.

Dadas as equações (2.2.10) e (2.2.11), observa-se que

$$\begin{aligned} h(t) = P(T = t | T \geq t) &= \frac{P(T = t \cap T \geq t)}{P(T \geq T)} = \frac{P(T = t)}{P(T = t) + P(T > t)} \\ &= \frac{p(t)}{p(t) + S(t)},\ t = 0, 1, 2, \dots, \end{aligned} \tag{2.2.13}$$

que resulta em

$$p(t) = \frac{h(t)}{1 - h(t)} S(t),\ t = 0, 1, 2, \dots\ . \tag{2.2.14}$$

Além disso, a distribuição de probabilidades pode ser expressa em termos da função de sobrevivência por meior da seguinte expressão:

$$p(t) = \begin{cases} 1 - S(0), & \text{se } t = 0. \\ S(t-1) - S(t), & \text{se } t = 1, 2, \dots\ . \end{cases} \tag{2.2.15}$$

Veja ainda que, para $t = 1, 2, \dots$ tem-se que

$$S(t) = \frac{S(0)}{1} \frac{S(1)}{S(0)} \frac{S(2)}{S(1)} \cdots \frac{S(t-1)}{S(t-2)} \frac{S(t)}{S(t-1)} = S(0) \prod_{k=1}^{t} \frac{S(k)}{S(k-1)}\ . \tag{2.2.16}$$

Visto que $S(0) = 1 - p(0)$ e $h(0) = p(0)$, a função de sobrevivência pode ser obtida a partir da função de risco por meio da seguinte expressão:

$$
\begin{aligned}
S(t) &= [1-h(0)]\prod_{k=1}^{t}\frac{S(k)}{p(k)+S(k)} = [1-h(0)]\prod_{k=1}^{t}\left[1-\frac{p(k)}{p(k)+S(k)}\right] \\
&= [1-h(0)]\prod_{k=1}^{t}[1-h(k)] \\
&= \prod_{k=0}^{t}[1-h(k)],\ t=0,1,2,\ldots,
\end{aligned}
\tag{2.2.17}
$$

A partir das equações (2.2.14) e (2.2.17), é possível expressar a distribuição de probabilidades em termos da função de risco por meio da seguinte expressão:

$$
p(t) = \frac{h(t)}{1-h(t)}\prod_{k=0}^{t}[1-h(k)],\ t=0,1,2,\ldots\ . \tag{2.2.18}
$$

2.3 Tempos discretos

O objetivo de diversas análises estatísticas, especialmente na análise de sobrevivência, é modelar o tempo até a ocorrência do evento de interesse. Conforme Berger e Schmid (2018), é comum assumir que o tempo de sobrevivência é uma variável aleatória medida em uma escala contínua, e existe uma vasta literatura sobre o tema. No entanto, na prática, as medições de tempo costumam ser discretas. Em algumas situações, a hora exata do evento pode não ser conhecida, apenas o intervalo durante o qual o evento ocorreu.

De acordo com Tutz e Schmid (2016), o tempo até a ocorrência do evento de interesse pode ser discreto devido a:

1. Medições intrinsecamente/genuinamente discretas;
2. Dados agrupados.

Os dados agrupados representam eventos que ocorrem em intervalos de tempo específicos, e a variável resposta se refere a um desses intervalos, que podem ter tamanhos iguais ou diferentes. Exemplos desse tipo de análise incluem estudos como o tempo, em meses, até a morte de homens diagnosticados com Síndrome da Imunodeficiência Adquirida (AIDS) (BRUNELLO; NAKANO, 2015) e o tempo, também em meses, de pacientes com câncer de cabeça e pescoço (CARDIAL; FACHINI-GOMES; NAKANO, 2020).

Em relação aos tempos genuinamente discretos, cujas medições representam números naturais, destacam-se diversas aplicações, tais como: o tempo até a gravidez, que em estudos clínicos é geralmente medido pelo número de ciclos menstruais (BERGER; SCHMID, 2018); o tempo até a evasão universitária, medido em semestres (VALLEJOS; STEEL, 2016); e o tempo até a degeneração macular relacionada à idade entre idosos, monitorado por visitas anuais de estudo (BERGER et al., 2019).

De acordo com Tutz e Schmid (2016), os métodos estatísticos desenvolvidos para tempos discretos oferecem várias vantagens:

1. **Interpretação facilitada**: A consideração de modelos para tempos discretos permite formular os riscos como probabilidades condicionais, o que facilita a interpretação em comparação com as funções de risco contínuas.

2. **Adequação aos dados discretos**: Na prática, diversos tempos de evento são intrinsecamente discretos ou observados em uma escala discreta. Portanto, a utilização de modelos para tempos discretos mostra-se mais adequada do que a aproximação dos dados observados por meio de um modelo de sobrevivência contínuo.

3. **Ausência de problemas com empates**: Diferentemente dos modelos de sobrevivência para tempo contínuo, os modelos para eventos discretos não apresentam problemas com empates.

4. **Possibilidade de incorporação em MLG**: Modelos para tempos discretos podem ser integrados na estrutura de um Modelo Linear Generalizado (MLG), permitindo que a estimativa seja obtida usando softwares padrão para a estimativa de MLG.

5. **Aplicação em modelos avançados**: A incorporação na estrutura de MLG permite utilizar a metodologia também para modelos avançados, como aqueles que incluem parâmetros específicos em modelos de fragilidade.

2.3.1 Discretização de distribuições contínuas

Em muitas situações práticas, variáveis que têm uma natureza contínua podem ser registradas de forma discreta. Por exemplo, o tempo até a ocorrência de inadimplência de crédito pode ser medido em dias ou meses, mesmo que o conceito subjacente seja contínuo. Nesse contexto, torna-se útil e apropriado modelar tais variáveis por meio de distribuições discretas, derivadas de modelos contínuos, preservando uma ou mais propriedades características da distribuição original, tais como sua função de densidade de probabilidade, função geradora de momentos, função da taxa de risco, entre outras. Chakraborty (2015) fornece uma revisão abrangente sobre métodos e técnicas para criar versões discretas de

distribuições de probabilidade contínuas. De acordo com Jayakumar e Babu (2018), entre as várias abordagens para discretizar distribuições contínuas, destacam-se as seguintes:

- Discretizar a função de distribuição acumulada contínua;
- Discretizar a função densidade de probabilidade contínua;
- Discretizar a função de risco contínua;
- Obter distribuição discreta de tempo de vida a partir da taxa de falha alternativa.

A primeira metodologia mencionada, que será a única abordada em detalhes neste trabalho, será explicada a seguir. Considere X como uma variável aleatória contínua não negativa com função de distribuição acumulada $F_X(x)$. A variável aleatória discreta T pode ser obtida através de $T = \lfloor X \rfloor$, em que $\lfloor X \rfloor$ representa a "parte inteira de X", ou seja, o maior inteiro que é menor ou igual a X. Dessa forma, a distribuição de probabilidades da variável aleatória discreta T pode ser expressa como (NAKANO; CARRASCO, 2006):

$$P(T = t) = P(t \leq X < t + 1) = F_X(t + 1) - F_X(t),\ t = 0, 1, 2, \dots\ . \tag{2.3.1}$$

Nesse contexto, diversas publicações sobre análise de sobrevivência utilizam essa metodologia para determinar análogos discretos de distribuições contínuas, resultando em novas distribuições de probabilidade discretas. Além disso, esses análogos discretos são aplicados em diferentes áreas. Exemplos incluem Nakagawa e Osaki (1975), que desenvolveram a distribuição Weibull discreta (WD); Jayakumar e Babu (2018), que introduziram a distribuição Weibull Geométrica discreta; Vieira et al. (2023), que trabalhou com a distribuição Log-logística discreta; Cardial, Fachini-Gomes e Nakano (2020), que estudaram a distribuição Weibull discreta exponenciada (WDE); Sarhan (2017), que apresentou a distribuição banheira de dois parâmetros discreta (DTPBT); Chakraborty (2015), que propôs a distribuição Gumbel discreta; e Brunello e Nakano (2015), que também investigaram a distribuição WD em um contexto bayesiano. A seguir será apresentada a distribuição Weibull discreta e a distribuição log-logística.

2.3.2 Distribuição Weibull Discreta

A distribuição Weibull é amplamente reconhecida e utilizada na modelagem de dados de sobrevivência contínuos devido à sua grande flexibilidade. Essa versatilidade se deve à presença de dois parâmetros principais: o parâmetro de escala e o parâmetro de

forma. Esses parâmetros permitem uma variedade de formas para a distribuição, além de uma função taxa de falha que pode ser crescente, decrescente ou constante, caracterizando-se por sua monotonicidade.

Neste trabalho, será abordada a versão discreta, introduzida por Nakagawa e Osaki (1975). Uma variável aleatória T segue uma distribuição Weibull discreta (WD) com parâmetros $\eta > 0$ e $q \in (0,1)$, denotado por $T \sim \text{WD}(q, \eta)$, se sua função de probabilidade é dada por:

$$p(t) = q^{t^\eta} - q^{(t+1)^\eta},\ t = 0, 1, 2, \dots\ . \tag{2.3.2}$$

A função de sobrevivência da distribuição Weibull discreta (WD) e a função de risco são expressas, respectivamente, por:

$$S(t) = q^{(t+1)^\eta}\ \ , \tag{2.3.3}$$

e

$$h(t) = \left(\frac{q^{t^\eta} - q^{(t+1)^\eta}}{q^{t^\eta}}\right)\ ,\ t = 0, 1, 2, \dots\ . \tag{2.3.4}$$

A função de risco da distribuição Weibull discreta assume diferentes formas dependendo do valor de η. Ela é estritamente crescente se $\eta > 1$, constante se $\eta = 1$ (neste caso, a distribuição Weibull discreta se reduz a uma distribuição geométrica), e estritamente decrescente se $\eta < 1$.

2.3.3 Distribuição log-logística discreta

A distribuição log-logística discreta é uma versão discretizada da distribuição log-logística contínua. A distribuição log-logística discreta é útil em diversas áreas, incluindo biomedicina, engenharia de confiabilidade e ciências sociais. Sua principal diferença em relação à distribuição WD é que a mesma acomoda funções de risco unimodais.

Segundo Santos (2017), considerando $\alpha > 0$ como o parâmetro de escala e $\gamma > 0$ como o parâmetro de forma da distribuição log-logística contínua, a função de probabilidade da distribuição log-logística discreta pode ser derivada a partir da Equação (2.3.1) e é dada por:

$$p(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^\gamma} - \frac{1}{1 + \left(\frac{t+1}{\alpha}\right)^\gamma},\ t = 0, 1, 2, \dots\ . \tag{2.3.5}$$

Dessa forma, as funções de sobrevivência e de risco são dadas, respectivamente, por

$$S(t) = \frac{1}{1 + \left(\frac{t+1}{\alpha}\right)^{\gamma}} \quad , \tag{2.3.6}$$

e

$$h(t) = \frac{\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}}}{\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}}}, \ t = 0, 1, 2, \ldots \ . \tag{2.3.7}$$

2.4 Obtenção de estimadores

Na análise de sobrevivência, a obtenção de estimadores desempenha um papel fundamental na modelagem e na interpretação dos dados relacionados ao tempo até a ocorrência de um evento de interesse. Do ponto de vista prático, o interesse inicial reside na estimativa da função densidade de probabilidade $f(t)$, da função de sobrevivência $S(t)$ e da função de risco $h(t)$. Essas funções podem ser estimadas diretamente a partir dos dados amostrais utilizando procedimentos não paramétricos.

2.4.1 Estimador de Kaplan-Meier de S(t)

O estimador de Kaplan-Meier é amplamente empregado como uma ferramenta não paramétrica para estimar a função de sobrevivência em presença de censuras. Ele é particularmente eficaz para uma análise preliminar dos dados, uma vez que as técnicas convencionais de cálculo de medidas resumo tendem a falhar nesse cenário (COLOSIMO; GIOLO, 2006).

Considere os tempos distintos de falha $t_1, t_2, \ldots, t_k$, onde $t_1 < t_2 < \ldots < t_k$. Existem n indivíduos com seus respectivos tempos de sobrevivência, e entre esses, k são tempos distintos que não apresentam censura. Dessa forma, temos que $k \leq n$, e cada tempo t_j (para $j = 1, \ldots, k$) pode ser observado mais de uma vez. Este estimador é também conhecido na literatura como estimador limite-produto e é definido como (KAPLAN; MEIER, 1958)

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right), \tag{2.4.1}$$

sendo d_j o número de falhas no tempo t_j e n_j é o número de indivíduos que não experimentaram do evento de interesse, e que não foram censurados até o tempo imediatamente anterior a t_j (COLOSIMO; GIOLO, 2006).

Seja $w > 0$ e $t_1 \leq w < t_2$, onde w é um tempo qualquer e t_1 e t_2 são tempos de falha observados. As estimativas de $S(t)$ respeitam a seguinte relação:

$$\hat{S}(w) = P(T > w) = P(T > t_1) = \hat{S}(t_1).$$

Esse é um dos motivos pelos quais a representação gráfica da função de sobrevivência estimada pelo método Kaplan-Meier assume a forma de uma escada.

2.4.2 Estimadores de Máxima verossimilhança

Atualmente, este é o método de estimação amplamente utilizado na inferência frequentista. Além disso, possui uma base teórica robusta, desenvolvida para uma ampla gama de situações como citado em Dani, Francisco e Migon (2014).

Seja $t_1, t_2, \ldots, t_n$ uma amostra aleatória observada de uma variável aleatória discreta T. Aqui, $p(t; \boldsymbol{\phi})$ representa a função de probabilidade da população e $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)^T$ é o vetor de parâmetros da função de probabilidade. A função de verossimilhança para $\boldsymbol{\phi}$, na ausência de censuras, é dada por:

$$L(\boldsymbol{\phi}; \mathbf{t}) = \prod_{i=1}^{n} p(t_i; \boldsymbol{\phi}). \tag{2.4.2}$$

No entanto, na presença de censura (à direita), os dados censurados devem ser distinguidos daqueles que sofreram o evento de interesse, frequentemente referidos como dados não censurados.

Dessa maneira, as observações podem ser reordenadas e divididas em dois grupos: os primeiros k elementos são os não censurados $(1, 2, \ldots, r)$, cuja contribuição para a função de verossimilhança é dada por $p(t_i; \boldsymbol{\phi})$, e os $k - r$ elementos restantes são os censurados $(k+1, k+2, \ldots, n)$, cuja contribuição para a função de verossimilhança é representada pela função de sobrevivência $S(t_i; \boldsymbol{\phi})$. Neste caso, a função de verossimilhança é expressa da seguinte forma:

$$L(\boldsymbol{\phi}; \boldsymbol{t}) \propto \prod_{i=1}^{k} p(t_i; \boldsymbol{\phi}) \prod_{i=k+1}^{n} S(t_i; \boldsymbol{\phi}) \ , \tag{2.4.3}$$

que é equivalente a:

$$L(\boldsymbol{\phi}; \boldsymbol{t}) \propto \prod_{i=1}^{n} [p(t_i; \boldsymbol{\phi})]^{\delta_i} [S(t_i; \boldsymbol{\phi})]^{1-\delta_i} \ , \tag{2.4.4}$$

em que δ_i é a variável indicadora de falha, que assume valor 1 se o tempo t_i for de falha e 0 se for censura à direita. Em (2.4.4), $p(\cdot; \boldsymbol{\phi})$ e $S(\cdot; \boldsymbol{\phi})$ são, respectivamente, a distribuição de probabilidade e função de sobrevivência do modelo considerado, $i = 1, 2, \ldots, n$.

Ao aplicar o logaritmo na função de verossimilhança na equação (2.4.4), tem-se:

$$\ell(\boldsymbol{\phi}; \boldsymbol{t}, \boldsymbol{\delta}) = c + \sum_{i=1}^{n} [\delta_i \log [p(t_i; \boldsymbol{\phi})] + (1 - \delta_i) \log [S(t_i; \boldsymbol{\phi})]] \ , \tag{2.4.5}$$

em que c é uma constante que não depende de $\boldsymbol{\phi}$.

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\phi}$ que maximizam $L(\boldsymbol{\phi}; \boldsymbol{t}, \boldsymbol{\delta})$ ou equivalentemente $\ell(\boldsymbol{\phi}; \boldsymbol{t}, \boldsymbol{\delta})$, normalmente representado por $\hat{\boldsymbol{\phi}}$, e são obtidos resolvendo o sistema de equações:

$$\frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{t}, \boldsymbol{\delta})}{\partial \boldsymbol{\phi}} = 0 \ . \tag{2.4.6}$$

2.5 Modelo de riscos proporcionais de Cox

De acordo com Colosimo e Giolo (2006), o modelo de regressão de Cox é uma ferramenta poderosa para a análise de dados provenientes de estudos de tempo de vida, onde a variável resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis. Este modelo é amplamente utilizado em estudos de sobrevivência devido à sua versatilidade. Fundamentado na suposição de que os riscos são proporcionais, a regressão de Cox não requer a especificação de uma distribuição de probabilidade para os tempos de sobrevivência, o que o torna um modelo robusto e flexível.

Existem várias razões que tornam a regressão de Cox atraente, como a capacidade de lidar com covariáveis dependentes do tempo, realizar análises estratificadas para controle de variáveis com ruído e funcionar tanto para medidas de tempo discretas quanto contínuas. No seu artigo original, Cox (1972) introduziu dois conceitos inovadores: o modelo de riscos proporcionais (posteriormente generalizado para riscos não proporcionais) e um novo método de estimação denominado máxima verossimilhança parcial.

Considerando $\mathbf{x}$ um vetor de covariáveis com p componentes, o modelo de regressão de Cox é dado por:

$$h(t) = h_o(t) g(\boldsymbol{x'\beta}) \ , \tag{2.5.1}$$

em que $g(\cdot)$ é uma função não negativa tal que $g(\mathbf{0}) = 1$ e $h_0(t)$ é a função de risco base (função de risco quando todas as covariáveis são iguais a zero).Dessa forma, o modelo de Cox é definido como o produto de dois componentes: um paramétrico e outro não paramétrico, razão pela qual também é chamado de modelo semiparamétrico. O componente não paramétrico é geralmente denominado função de risco base ou basal. Isso ocorre porque $h(t) = h_0(t)$ quando $\mathbf{x} = 0$, ou seja, $h_0(t)$ pode ser considerada a taxa de falha

de um indivíduo cujas covariáveis possuem valor zero. A função base não é especificada, mas deve ser uma função não negativa ao longo do tempo.

Por outro lado, a parte paramétrica é uma função positiva e contínua das covariáveis. Embora existam outras formas na literatura para essa componente, ela é comumente escrita na forma exponencial, devido à sua propriedade de ser sempre positiva, da seguinte maneira:

$$g(\boldsymbol{x'\beta}) = \exp(\boldsymbol{x'\beta}) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p) \ , \tag{2.5.2}$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos.

Devido à sua linearidade no modelo, é comum referir a soma $\beta_1 x_1 + \ldots + \beta_p x_p$ como preditor linear ou escore, que, na forma matricial, é denotado por $\boldsymbol{x'\beta}$.

É importante observar que a constante β_0 (intercepto), presente nos modelos paramétricos, não aparece na função $g(\boldsymbol{x'\beta})$. Isso se deve à presença do componente não paramétrico no modelo, que absorve esse termo constante. A expressão do modelo em (2.5.2) implica que a razão das taxas de falha ou de risco entre dois indivíduos, l e m, é constante ao longo do tempo, sendo uma função apenas das covariáveis, como pode ser observado em (2.5.3).

$$\frac{h_l(t)}{h_m(t)} = \frac{h_0(t)\exp(\boldsymbol{x_l'\beta})}{h_0(t)\exp(\boldsymbol{x_m'\beta})} = \exp(\boldsymbol{x_l'\beta} - \boldsymbol{x_m'\beta}). \tag{2.5.3}$$

Devido a essa razão, o Modelo de Cox também é conhecido como Modelo de Riscos Proporcionais. Apesar do modelo de Cox ser muito flexível devido ao componente não paramétrico, a suposição fundamental de taxas de falha proporcionais não pode ser violada para a correta utilização do Modelo de Cox. Para avaliar a proporcionalidade dos riscos, podem ser empregadas técnicas gráficas e testes estatísticos.

2.5.1 Formulação do modelo

Dado um conjunto de observações de sobrevivência, o objetivo comum é estimar modelos preditivos nos quais o risco do evento depende de covariáveis. Uma maneira de determinar tal modelo é estimando os coeficientes $\boldsymbol{\beta' s}$ que mensuram os efeitos dos atributos sobre a função taxa de falha no Modelo de Cox.

Assim, é necessário um método de estimação que permita a construção de inferências sobre os parâmetros do modelo. O método da máxima verossimilhança, frequentemente utilizado, não pode ser empregado devido à presença do componente não paramétrico. Por isso, Cox propôs um novo método de estimação: a máxima verossimilhança parcial, através do qual é possível estimar os coeficientes das covariáveis sem a

necessidade de especificar a função base $h_0(t)$.

Uma forma simples de entender esse método, segundo Colosimo e Giolo (2006), considera o seguinte argumento condicional: a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i, conhecendo quais indivíduos estão sob risco em t_i é:

$$P(\text{indivíduo falhar em } t_i|\text{uma falha em } t_i \text{ e história até } t_i) =$$

$$\frac{P(\text{indivíduo falhar em } t_i|\text{sobreviveu a } t_i \text{ e história até } t_i)}{P(\text{uma falhar em } t_i|\text{história até } t_i)} =$$

$$\frac{h_i(t|\boldsymbol{x_i})}{\sum_{j\in R(t_i)} h_j(t|\boldsymbol{x_j})} = \frac{h_0(t)\exp(\boldsymbol{x_i'\beta})}{\sum_{j\in R(t_i)} h_0(t)\exp(\boldsymbol{x_j'\beta})} = \frac{\exp(\boldsymbol{x_i'\beta})}{\sum_{j\in R(t_i)} \exp(\boldsymbol{x_j'\beta})},$$

em que $R(t_i)$ representa o conjunto dos índices das observações sob risco em t_i. Cox propôs utilizar o registro histórico passado de falhas e censuras na forma de probabilidade condicional para eliminar o termo não paramétrico da função de verossimilhança.Assim, a função de máxima verossimilhança parcial é dada por:

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

em que δ_i é o indicador de falha, n é o tamanho da amostra, $k < n$ o número de falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$.

Essa função obtida para o modelo de riscos proporcionais não é uma verossimilhança verdadeira, porque não utiliza os verdadeiros tempos de sobrevivência dos clientes censurados e não censurados. Por isso, ela é chamada de verossimilhança parcial.

O logaritmo dessa função de verossimilhança é dado por:

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x_j'\beta} - \log \sum_{j\in R(t_i)} \exp(\boldsymbol{x_j'\beta}) \right). \tag{2.5.4}$$

As estimativas de verossimilhança dos parâmetros $\boldsymbol{\beta}'$s são obtidos maximizando-se (2.5.4), ou seja, resolvendo o sistema de equações definido $U(\boldsymbol{\beta}) = 0$, em que U é o vetor escore formado pelas primeiras derivadas de $\ell(\boldsymbol{\beta})$.

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(x_i - \frac{\sum_{j\in R(t_i)} x_j \exp(\boldsymbol{x_j'\hat{\beta}})}{\sum_{j\in R(t_i)} \exp(\boldsymbol{x_j'\hat{\beta}})} \right) = 0\ . \tag{2.5.5}$$

Assim, o termo "regressão de Cox" refere-se à combinação do modelo de riscos proporcionais com o método de estimação da máxima verossimilhança parcial.

Este método de estimação possui duas das três propriedades padrão dos estimadores de máxima verossimilhança: resultados consistentes e assintoticamente normais. Isso significa que, em grandes amostras, os estimadores são aproximadamente não viesadas e a sua distribuição amostral é aproximadamente normal.

Tanto o modelo de riscos proporcionais quanto a função de verossimilhança parcial pressupõem que os tempos de sobrevivência são contínuos, o que, em teoria, impede a ocorrência de empates nos valores observados. No entanto, na prática, empates podem ocorrer devido a escalas de medição, processos de coleta de dados, arredondamentos e aproximações, bem como a ocorrência de múltiplos eventos no mesmo instante de tempo. Também é possível haver empates entre observações censuradas e entre falhas e censuras. Portanto, são necessárias adequações à função de verossimilhança para lidar com esses empates.

Com as estimativas dos coeficientes $\boldsymbol{\beta}$ e seus erros-padrão, é possível construir um intervalo de confiança de 100(1 - α)% para um determinado $\boldsymbol{\beta_p}$, utilizando os percentis da distribuição normal padrão. Se o intervalo de confiança não incluir o valor zero, pode-se afirmar que há evidências suficientes para considerar que o coeficiente $\boldsymbol{\beta_p}$ é significativamente diferente de zero.

2.5.2 Estimação dos parâmetros

Dado um Modelo de Cox com o vetor $\mathbf{x}$ de covariáveis de dimensão p e as respectivas estimativas dos coeficientes, então a função taxa de falha para o j-ésimo indivíduo é dada por:

$$\hat{h}_j(t) = \hat{h}_0(t) \exp(\boldsymbol{x_j'}\hat{\boldsymbol{\beta}}) \ , \tag{2.5.6}$$

em que $\hat{h}_0(t)$ é a estimativa da função base.

Outras funções relacionadas a $h_0(t)$ são importantes, especialmente em análises gráficas para avaliar a adequação do modelo ajustado. No entanto, como $h_0(t)$ não é especificado de forma paramétrica, outras técnicas são utilizadas para sua estimativa. A função de risco acumulada base pode ser estimada de maneira simples, segundo Breslow (1975), onde uma função em degraus com saltos nos tempos distintos de falha é empregada da seguinte maneira:

$$\hat{H}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l \in R_j} \exp(\boldsymbol{x_j'}\hat{\boldsymbol{\beta}})} \ , \tag{2.5.7}$$

em que d_j é o número de falhas em t_j.

Consequentemente, é possível estimar as funções de sobrevivência $S_0(t)$ e $S(t)$ da seguinte forma

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\} \tag{2.5.8}$$

e

$$\hat{S}(t|\boldsymbol{x_j}) = \hat{S}_0(t)^{\exp(\boldsymbol{x_j'}\hat{\boldsymbol{\beta}})}. \tag{2.5.9}$$

2.5.3 Avaliação do modelo (verificação do ajuste)

Apesar de o Modelo de Cox ser flexível, é necessário avaliar a adequação dos dados à aplicação da metodologia. Uma maneira de verificar se o modelo escolhido é o mais apropriado consiste em examinar o comportamento dos resíduos entre os valores preditos e observados. Isso permite analisar a suposição de riscos proporcionais e identificar dados discrepantes na amostra.

Para verificar a suposição de riscos proporcionais no modelo de Cox, é comum usar um gráfico específico. Inicialmente, divide-se os dados em m estratos, geralmente com base em uma covariável, como sexo. Em seguida, estima-se $\hat{h}_{0j}(t)$ para cada estrato. Se a suposição de riscos proporcionais for válida, as curvas do logaritmo de $\hat{h}_{0j}(t)$ versus t, ou $log(t)$, devem ser aproximadamente paralelas. Curvas não paralelas indicam desvios dessa suposição. É recomendável construir esse gráfico para cada covariável do estudo. Para covariáveis contínuas, sugere-se agrupá-las em poucas categorias. Essa técnica gráfica tem a vantagem de indicar qual covariável está violando a suposição, se for o caso.

Uma proposta adicional que vem sendo usada para verificar a suposição de riscos proporcionais no modelo de Cox é a de analisar os resíduos de Schoenfeld (1982). Considere que o indivíduo i experimentou o evento de interesse, sendo observado o tempo de falha e o vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. O resíduo de Schoenfeld é definido como $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{ip})'$, onde cada componente r_{iq}, para $q = 1, 2, \ldots, p$, é dado por:

$$r_{iq} = x_{i1} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp(\boldsymbol{x_j'}\hat{\boldsymbol{\beta}})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x_j'}\hat{\boldsymbol{\beta}})} \ . \tag{2.5.10}$$

Esse resíduo não é definido para censuras, apenas para tempos de falha. Entretanto, essa medida, definida dessa forma, é pouco utilizada, pois não considera a estrutura de correlação entre os resíduos. Para contornar essa limitação, foram desenvolvidos os Resíduos Padronizados de Schoenfeld, que são frequentemente utilizados. Nesse caso,

é necessário usar a matriz de informação observada, $I(\beta)$, como um fator multiplicativo aplicado ao resíduo simples, conforme a seguinte fórmula:

$$s_i^* = [I(\beta)]^{-1} r_i \ . \tag{2.5.11}$$

Consequentemente, se a suposição de riscos proporcionais for válida, o gráfico de $\beta_q(t)$ versus t deve ser uma reta horizontal, já que uma inclinação zero indica a proporcionalidade dos riscos.

3 Modelo de chances de riscos proporcionais (MCRP)

No capítulo anterior, foi apresentada a ferramenta do modelo de Cox para descrever o tempo de sobrevivência considerando a influência das covariáveis em uma variável resposta contínua T. No entanto, não é possível formular um modelo de riscos proporcionais (totalmente) paramétrico quando a resposta é discreta. Como alternativa discreta ao modelo de riscos proporcionais, este trabalho adotará o modelo de chances de riscos proporcionais (MCRP), proposto por Vieira et al. (2023).

3.1 Formulação do modelo

Como discutido em (2.2.11), quando T é uma variável aleatória discreta, a função de risco é uma probabilidade, ou seja, $0 < h(t) < 1$. Nesse contexto, não é possível adotar a estrutura de riscos proporcionais para incluir covariáveis no modelo de regressão. No entanto, é possível utilizar a razão de chances (odds) de $h(t)$, ou seja:

$$odds\{h(t)\} = \frac{h(t)}{1-h(t)} \ . \tag{3.1.1}$$

Como $odds\{h(t)\} > 0$, o modelo proposto é denominado como modelo de chances de riscos proporcionais (MCRP) e considera que as covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ agem multiplicativamente (proporcionalmente) na chance do risco. Isto é,

$$\frac{h(t|\boldsymbol{x})}{1-h(t|\boldsymbol{x})} = g(\boldsymbol{x}'\boldsymbol{\beta})\frac{h_0(t)}{1-h_0(t)}, \tag{3.1.2}$$

em que $h_0(t)$ é a função de risco base, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ é o vetor de coeficientes associado ao vetor de covariáveis $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ e $g(.)$ é uma função de ligação que satisfaz as seguintes condições:

1. $g(a) > 0, \ \forall a \in \mathbb{R}$;
2. $g(0) = 1$.

Note que o intercepto β_0 não aparece no preditor linear $\boldsymbol{x}'\boldsymbol{\beta}$. Isto porque a função de risco base, $h_0(t)$, absorve este termo constante (VIEIRA et al., 2023).

Além disso, a propriedade de chances de riscos proporcionais permite interpretar os coeficientes estimados. Considerando a função de ligação exponencial ($g(\cdot) = \exp(\cdot)$), a razão de chances do risco entre dois indivíduos (r e s) que possuem os mesmos valores para todas as covariáveis, exceto a m-ésima, é expressa por

$$\frac{odds\{h(t|\boldsymbol{x_r}\})}{odds\{h(t|\boldsymbol{x_s}\})} = \frac{\exp\{\beta_m x_{rm}\}}{\exp\{\beta_m x_{sm}\}} = \exp\{\beta_m(x_{rm} - x_{sm})\} \ , \tag{3.1.3}$$

que não depende de t.

Observe que a Equação (3.1.3) representa uma razão de chances (odds ratio) de riscos. Por exemplo, se x_m for a covariável dicotômica sexo, com $x_{rm} = 1$ (masculino) e $x_{sm} = 0$ (feminino), então a chance de falha (odds do risco) para indivíduos do sexo masculino é $\exp(\beta_m)$ vezes a chance de falha para indivíduos do sexo feminino, mantendo as demais covariáveis constantes.

A partir da equação (3.1.2), observa-se que a função de risco de um indivíduo com covariáveis $\boldsymbol{x}$ é dada por

$$h(t|\boldsymbol{x}) = \frac{g(\boldsymbol{x'\beta})h_0(t)}{1 - h_0(t) + g(\boldsymbol{x'\beta})h_0(t)} \ . \tag{3.1.4}$$

Segundo a equação (2.2.17) e (3.1.4), a função de sobrevivência na presença de covariáveis pode ser escrita como

$$S(t|\boldsymbol{x}) = \prod_{u=0}^{t}[1 - h(u|\boldsymbol{x})] = \prod_{u=0}^{t}\left[\frac{1 - h_0(u)}{1 - h_0(t) + g(\boldsymbol{x'\beta})h_0(u)}\right] \ t = 0, 1, 2, \ldots . \tag{3.1.5}$$

Portanto, utilizando as equações (2.2.15) e (2.2.17), a função de probabilidade é expressa por:

$$p(t|\boldsymbol{x}) = \begin{cases} h(0|\boldsymbol{x}), & \text{se } t = 0 \ . \\ h(t|\boldsymbol{x})S(t-1|\boldsymbol{x}), & \text{se } t = 1, 2, \ldots \ . \end{cases} \tag{3.1.6}$$

Para ajustar o MCRP, é necessário estimar os parâmetros da distribuição base e da componente de regressão. Esse processo é realizado maximizando a função de verossimilhança, dada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\left\{\left[h(0|\boldsymbol{x})^{\mathbb{1}_{\{0\}}(t_i)}\left[h(t_i|\boldsymbol{x})S(t_i-1|\boldsymbol{x})\right]^{1-\mathbb{1}_{\{0\}}(t_i)}\right]^{\delta_i}\left[S(0|\boldsymbol{x})^{\mathbb{1}_{\{0\}}(t_i)}S(t_i|\boldsymbol{x})^{1-\mathbb{1}_{\{0\}}(t_i)}\right]^{1-\delta_i}\right\}, \tag{3.1.7}$$

em que

$$\mathbb{1}_{\{0\}}(t_i) = \begin{cases} 1, & \text{se } t_i = 0 \\ 0, & \text{se } t_i \neq 0, \end{cases}$$

δ_i é o indicador de falha do i-ésimo indivíduo e $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\beta}')$ é o vetor de parâmetros do modelo, sendo $\boldsymbol{\phi}$ o vetor de parâmetros da distribuição base do tempo de sobrevivência e $\boldsymbol{\beta}$ o vetor de coeficientes de regressão.

3.2 MCRP Weibull discreto

Considerando T como uma variável aleatória com distribuição Weibull discreta, a partir das equações (2.3.4) e (3.1.4) e usando a função de ligação $g(\boldsymbol{x'\beta}) = \exp(\boldsymbol{x'\beta})$, foi derivada a seguinte fórmula para a função de risco no MCRP Weibull discreto.

$$h(t|\boldsymbol{x}) = \frac{e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}{q^{(t+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}, t = 0, 1, 2, \dots \ . \tag{3.2.1}$$

Substituindo $h_0(t)$ e $g(\boldsymbol{x'\beta})$ em (3.1.5), a função de sobrevivência pode ser rescrita como

$$S(t|\boldsymbol{x}) = \prod_{u=0}^{t} \left[\frac{q^{(u+1)^\eta}}{q^{(u+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{u^\eta} - q^{(u+1)^\eta})} \right], \ t = 0, 1, 2, \dots \ . \tag{3.2.2}$$

A partir de (3.1.6), (3.2.1) e (3.2.2), tem-se que a função de probabilidade é dada por

$$p(t|\boldsymbol{x}) = \begin{cases} \frac{e^{x'\beta}(1-q)}{q+e^{x'\beta}(1-q)}, \text{ se } t = 0 \ . \\ \\ \frac{e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}{q^{(t+1)^\eta}} \prod_{u=0}^{t-1} \left[\frac{q^{(u+1)^\eta}}{q^{(u+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{u^\eta} - q^{(u+1)^\eta})} \right], \text{ se } t = 1, 2, \dots \ , \end{cases} \tag{3.2.3}$$

que resulta em

$$p(t|\boldsymbol{x}) = \frac{e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}{q^{(t+1)^\eta}} \prod_{u=0}^{t} \left[\frac{q^{(u+1)^\eta}}{q^{(u+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{u^\eta} - q^{(u+1)^\eta})} \right], t = 0, 1, 2, \dots \ . \tag{3.2.4}$$

Sabe-se que a distribuição geométrica é um caso particular da distribuição Weibull discreta. Portanto, ao substituir $\eta = 1$ nas equações (3.2.2), (3.2.3) e (3.2.4), obtemos, respectivamente, as funções de risco, de sobrevivência e de probabilidade para o MCRP geométrico. Com base na definição dessas funções, é possível determinar a função de verossimilhança a partir da Equação (3.1.7) e, assim, estimar os parâmetros do modelo.

3.3 Verificação da suposição de chances de riscos proporcionais

Segundo Vieira et al. (2023), o MCRP (3.1.2) pressupõe que as chances (*odds*) de risco para dois indivíduos são proporcionais. Considerando, por exemplo, uma covariável z dicotômica que assume os valores 0 e 1, o modelo supõe que

$$\frac{h(t|z=1)}{1-h(t|z=1)} = C\frac{h(t|z=0)}{1-h(t|z=0)} \ , \tag{3.3.1}$$

em que $h(.)$ é a função de risco e C é uma constante que não depende do tempo t.

Seja $o_l(.)$ a função chance (*odds*) de risco de um indivíduo com covariável $z = l, l = 0, 1$,

$$o_l(t) = \frac{h(t|z=l)}{1-h(t|z=l)}, \ l = 0, 1 \ , \tag{3.3.2}$$

e $O_l(.)$ sua respectiva função chance de risco acumulada. Isto é,

$$O_l(t) = \sum_{u=0}^{t} o_l(u) = \sum_{u=0}^{t} \frac{h(u|u=l)}{1-h(u|z=l)}, \ l = 0, 1 \ . \tag{3.3.3}$$

Note que sob a suposição das chances dos riscos serem proporcionais, tem-se a partir das Equações (3.3.1), (3.3.2) e (3.3.3) que

$$O_1(t) = CO_0(t) \ . \tag{3.3.4}$$

Aplicando o logaritmo em ambos os lados da Equação (3.3.4), obtém-se a seguinte expressão:

$$log[O_1(t)] = log[C] + log[O_0(t)] \ . \tag{3.3.5}$$

Portanto, a relação entre $\log[O_1(t)]$ e $\log[O_0(t)]$ é uma linha reta com coeficiente angular igual a 1 e intercepto igual a $\log[C]$.

Assim, a suposição de chances de riscos proporcionais pode ser verificada graficamente ajustando uma reta de regressão simples com coeficiente angular igual a 1 ($b = 1$ fixo). O procedimento consiste em construir um gráfico cujos pontos são dados pelas coordenadas $(\log[O_0(t)], \log[O_1(t)])$ e o comportamento esperado é que os pontos do gráfico estejam próximos da reta de regressão ajustada.

Alternativamente, pode-se criar gráficos de t ou $\log(t)$ versus $\log[O_l(t)]$, $l = 0$, 1. Curvas paralelas, que mantêm uma distância vertical constante, indicam chances de

riscos proporcionais.

Ademais, um teste de hipóteses pode ser adotado para checar se as chances de riscos são proporcionais. Assim, se $t_{(j)}$, com $j = 1, 2, \ldots, J$, é o j-ésimo tempo distinto observado (censurado ou não), a verificação de proporcionalidade das chances de riscos pode ser feita testando-se a hipótese de que o coeficiente angular de uma reta de regressão simples é diferente de 1, isto é, a hipótese de interesse é descrita por:

$$H_0 : b_1 = 1 \quad vs. \quad H_1 : b_1 \neq 1. \tag{3.3.6}$$

A estatística do teste da hipótese (3.3.6) é dada por:

$$B = \frac{\widehat{b_1} - 1}{\sqrt{\frac{\sum_{j=1}^{J}(z_j - \bar{z})^2}{(J-2)\sum_{j=1}^{J}(y_j - \bar{y})^2}}}, \tag{3.3.7}$$

em que $\widehat{b_1} = \dfrac{J\sum_{j=1}^{J} z_j y_j - \sum_{j=1}^{J} z_j \sum_{j=1}^{J} y_j}{J\sum_{j=1}^{J} z_j^2 - \left(\sum_{j=1}^{J} z_j\right)^2}$, $\bar{z} = \dfrac{\sum_{j=1}^{J} z_j}{J}$ e $\bar{y} = \dfrac{\sum_{j=1}^{J} y_j}{J}$ com $z_j = \log\left[O_0(t_j)\right]$ e $y_j = \log\left[O_1(t_j)\right]$. Assumindo a normalidade de $\log\left[O_1(t)\right]$, B segue uma distribuição t de Student com $J - 2$ graus de liberdade.

Os procedimentos descritos para verificação da suposição de chances de riscos proporcionais podem ser facilmente estendido para covariáveis categóricas com três ou mais níveis, comparando todos os níveis dois-a-dois. Para covariáveis númericas, os mesmos procedimentos podem ser adotados, bastando categorizar a covariável e comparando seus pares de níveis.

4 Modelagem de risco de crédito

4.1 Introdução

Ao longo da história do desenvolvimento econômico e social das sociedades, o crédito tem sido um dos principais fatores a serem considerados, pois, permite que agentes sociais de diferentes setores alcancem expansão econômica.

Sob essa perspectiva financeira:

> O crédito corresponde a um valor monetário disponibilizado ao tomador de recursos financeiros, na forma de empréstimo ou financiamento, por um período previamente acordado, com a promessa de pagamento futuro, ao qual se acrescenta uma remuneração, denominada juros. Consequentemente, o risco é inerente ao processo de concessão de crédito, uma vez que existem incertezas quanto ao futuro das quantias emprestadas (MACHADO, 2015).

Segundo Santos (2011), "o risco é definido pela incerteza de retorno de um investimento frente à possibilidade de um evento futuro, incerto e independente da vontade do investidor, cuja ocorrência pode causar prejuízos". Nesse contexto, o risco de crédito está associado a fatores internos e externos ao credor que podem dificultar a recuperação do montante emprestado. Para o Banco Central do Brasil, conforme o Art. 2º da Resolução 3.721/2009, o risco de crédito é definido como a possibilidade de perdas associadas ao não cumprimento das obrigações financeiras pelo tomador ou contraparte nos termos pactuados, à desvalorização de contratos de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação.

Dentro da esfera do risco, diversos aspectos devem ser analisados, como:

- Risco do Cliente - associado aos C's do Crédito:
    1. Capacidade - habilidade em pagar. Relaciona-se aos meios financeiros para honrar os compromissos assumidos;
    2. Colateral - garantia;
    3. Caráter - confiabilidade e "vontade"de pagar;
    4. Condição - condições ambientais externas, internas e indicadores econômicos;
    5. Capital - reservas e patrimônio.

- Risco da Operação - envolve características do produto, prazo, formas de pagamento, garantia e preço;
- Risco de Carteira - relacionado ao conjunto de clientes e tipos de negócios;
- Risco de Administração de Crédito - compreende o acompanhamento do crédito concedido.

Nesse cenário, surgiram os Modelos de *Credit Scoring*, como ferramentas capazes de quantificar o risco de crédito envolvido em uma operação de forma automatizada, padronizada e objetiva.

Os Modelos de *Credit Scoring* utilizam algoritmos matemáticos e técnicas estatísticas para calcular a probabilidade de que determinado evento ocorra. Aplicando fórmulas, o sistema atribui uma pontuação específica para cada característica do proponente/cliente, com o objetivo de prever um resultado.

Historicamente, os modelos de *Credit Scoring (CS)* foram iniciados pelos estudos de Durand (1941) na área de financiamento ao consumidor após a Grande Depressão nos EUA. Este projeto foi pioneiro na utilização da Estatística como uma ferramenta para análise de risco de crédito. Na pesquisa, foi utilizada a Análise Discriminante desenvolvida por Fisher (1936) para identificar bons e maus empréstimos. Nesse contexto, a pesquisa de Durand pode ser considerada o ponto de partida para futuros estudos focados no desenvolvimento de metodologias de suporte à concessão de crédito.

No início dos anos 1950, Bill Fair e Earl Isaac criaram a primeira companhia de consultoria em métodos de *scoring*, utilizando dados históricos para melhorar as decisões de negócios. Posteriormente, em 1958, venderam o primeiro sistema de *Credit Scoring* para a área de cartões de crédito, fato considerado um marco importante na história dos modelos de *scoring*. Entretanto, o sucesso da companhia e seu foco comercial não resultaram em um desenvolvimento significativo da literatura sobre o tema, uma vez que o conhecimento se tornou valioso e foi pouco divulgado pelas empresas.

Apesar dos Modelos de *CS* representarem uma melhoria na análise de risco de crédito, diversos fatores dificultavam seu crescimento, como a relutância dos executivos, limitações tecnológicas, obstáculos no desenvolvimento e implementação dos modelos e, segundo Myers e Forgy (1963), a falta de estatísticos para promover a área de crédito e transformar a ideia em uma ferramenta operacional bem-sucedida e útil. Diante disso, apesar da expansão do crédito nos EUA, poucos estudos sobre *Credit Scoring* foram produzidos até os anos 1960.

A partir de 1960, outras pesquisas relevantes foram publicadas, como:

- Desenvolvimento de Sistemas Numéricos de Avaliação de Crédito (MYERS; FORGY,

1963). Os autores verificaram a eficácia das fórmulas preditivas de *scoring* e introduziram o conceito de amostra *hold-out*, diferente daquela utilizada para a modelagem, importante para testar a capacidade preditiva do modelo em novas amostras.

- Conceitos e Utilização de Técnicas de *CS* (WEINGARTNER, 1966). O autor ressaltou a importância de testar os escores de crédito antes de usá-los e sugeriu validar a fórmula aplicando-a a clientes inadimplentes para verificar se os escores são baixos.
- Índices Financeiros, Análise Discriminante e Previsão de Falência de Empresas (ALTMAN, 1968). Introduziu modelos de *scoring* para empresas.
- Um Modelo de *CS* para Empréstimos Comerciais (ORGLER, 1970). Propôs um modelo para avaliar periodicamente a qualidade dos empréstimos concedidos.

A partir dos anos 1970, com o crescimento econômico e a consequente demanda por crédito, muitas instituições financeiras nos EUA cresceram de forma insustentável, sem conseguir manter a lucratividade. Ao mesmo tempo, a reconstrução da Europa pós-guerra contribuiu para que os Modelos de *CS* fossem reconhecidos como uma indústria. Desde o início dos anos 1990, os Modelos de *CS* tornaram-se o principal mecanismo para avaliação de risco na concessão de diversos tipos de empréstimos, com decisões sendo tomadas sem intervenção humana.

Com a divulgação do Acordo de Basileia II em 2004, os Modelos de *CS* tornaram-se ainda mais importantes, destacando a necessidade de técnicas que permitissem às instituições e supervisores avaliar corretamente os diversos riscos enfrentados pelos bancos. Muitas organizações desenvolveram melhores modelos ou modificaram os já existentes para conformidade com as novas regras e melhores práticas de mercado, dado que os reguladores impuseram regras mais rigorosas sobre o desenvolvimento, implementação e validação dos modelos internos utilizados para estimar o capital a ser provisionado.

Com o contínuo desenvolvimento e crescimento dos mercados financeiros, o crédito tornou-se ainda mais crucial para a economia. A globalização e a sofisticação dos meios de comunicação, como a internet, fazem com que os consumidores busquem ofertas de crédito mais atrativas. Por isso, as instituições financeiras procuram desenvolver ferramentas eficientes para avaliar e controlar os riscos.

Os Modelos de *CS*, inicialmente utilizados apenas para decidir a concessão de crédito, hoje são parte integral de todo o ciclo do crédito, presentes em cada etapa da gestão estratégica de riscos.

4.2 Escore de risco

A mensuração do risco de crédito é o processo de quantificar a credibilidade de um solicitante de crédito, utilizando variáveis explicativas para classificar os solicitantes como "bons" ou "maus" pagadores. O objetivo dessa classificação prévia é prever comportamentos que possam indicar padrões de inadimplência, evitando maiores prejuízos e a perda de bons clientes para a instituição financeira.

A proposta desse trabalho é formular o escore de risco a partir dos resultados do MCRP, sendo a sua grandeza associada ao valor estimado do preditor linear do modelo. No MCRP, quanto maior for o valor do preditor linear, maior é a função de risco do cliente. Isso implica, no contexto desse modelo, maior probabilidade do cliente inadimplir (ou inadimplir em um tempo mais recente). Desta forma, o preditor linear $\boldsymbol{x'\beta}$ pode ser considerado como um escore de risco para o MCRP, isto é,

$$ER = \boldsymbol{x'\beta} = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p. \tag{4.2.1}$$

Note que o escore de risco definido em (4.2.1) assume valores reais. Alternativamente, as transformações $\exp\{\cdot\}$ ou $\exp\{-\exp\{\cdot\}\}$ podem ser incorporadas para obter escores de risco positivos ou limitados no intervalo 0–1, respectivamente.

4.3 Classificação dos clientes pelo escore de risco

A classificação dos clientes é uma das etapas mais cruciais para as instituições financeiras. Essa classificação orienta as posturas e estratégias em relação às concessões de crédito. O objetivo principal é maximizar os lucros e minimizar os riscos, garantindo uma gestão eficaz das concessões de crédito.

Com base nos pontos mencionados, os critérios de classificação dos clientes pelo escore de risco têm os seguintes objetivos:

1. Minimizar o erro ao classificar maus solicitantes como de risco baixo, evitando assim concessões de crédito com alto risco.

2. Minimizar o erro ao classificar bons solicitantes como de risco alto, evitando a perda de clientes valiosos.

3. Maximizar a precisão total na classificação dos clientes, garantindo que a avaliação seja justa e precisa.

4.4 Avaliação da acurácia do modelo

Para exemplificar os termos de acurácia, considere o seguinte caso : Seja X o *status* de um cliente (1 = mal pagador, 2 = bom pagador) e Y o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco; e negativo, quando o modelo classifica o cliente como de baixo risco).

Validação cruzada :

Essa técnica envolve a divisão dos dados em duas amostras, geralmente de tamanhos iguais: uma amostra de estimação e uma amostra de validação. A subamostra de estimação é utilizada para calcular os parâmetros do modelo, enquanto a subamostra de validação serve para validar esses parâmetros e verificar o poder preditivo do modelo. Esse processo permite avaliar quantitativamente a capacidade de previsão do modelo em relação a novas observações.

Total de acertos:

Corresponde ao número de classificações corretas do modelo para a variável resposta em relação à variável explicativa.

Sensibilidade:

Corresponde à probabilidade de o modelo alocar o indivíduo i na categoria K, dado que ele realmente pertence a essa categoria. Considerando o exemplo citado anteriormente, a sensibilidade será definida como a probabilidade de o diagnóstico do modelo acertar a classificação de risco como alto ($Y = +1$) para um mal pagador (AGRESTI, 2019), ou seja,

$$\text{Sensibilidade} = P(Y = +1|X = 1)$$

Especificidade:

Corresponde à probabilidade de o modelo não alocar o indivíduo i na categoria K, dado que ele realmente não pertence a essa categoria. Considerando o exemplo citado anteriormente, a especificidade será definida como a probabilidade de o diagnóstico do modelo classificar um bom pagador como de baixo risco ($Y = -1$) (AGRESTI, 2019), ou seja,

$$\text{Especificidade} = P(Y = -1 \mid X = 2)$$

Falso Positivo:

Um falso positivo ocorre quando o modelo classifica incorretamente uma ob-

servação como pertencente à categoria de sucesso, quando na verdade ela pertence à categoria de fracasso. Considerando o exemplo citado anteriormente, um falso positivo seria quando o modelo de risco classifica erroneamente um bom pagador como de alto risco.

Falso Negativo:

Um falso negativo ocorre quando o modelo classifica incorretamente uma observação como pertencente à categoria de fracasso, quando na verdade ela pertence à categoria de sucesso. Considerando o exemplo citado anteriormente, um falso negativo seria quando o modelo de risco classifica erroneamente um mal pagador como de baixo risco.

Matriz de confusão:

A matriz de confusão é uma tabela que compara os valores reais com os valores preditos pelo modelo, relatando o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Considerando o exemplo citado anteriormente, a matriz de confusão seria construída da seguinte forma:

Tabela 1: Matriz de confusão

	Valores reais (X)	
Diagnóstico do modelo de risco (Y)	Mal pagador	Bom pagador
Alto risco	Verdadeiro positivo	Falso positivo
Baixo risco	Falso negativo	Verdadeiro negativo

O número total de acertos é dado pela soma da diagonal principal da matriz de confusão.

5 Ilustração da metodologia proposta

Este capítulo apresenta a aplicação da metodologia desenvolvida neste trabalho para a classificação de clientes, detalhando as etapas necessárias para a construção e avaliação do modelo.

Inicialmente, o banco de dados foi dividido em dois subconjuntos: treino e teste. O conjunto de treino, correspondente a 70% do total de dados, é utilizado para o desenvolvimento e ajuste do modelo. Por sua vez, o conjunto de teste, que representa os 30% restantes, é empregado após a criação do modelo, permitindo simular previsões em cenários reais e avaliar seu desempenho e capacidade de generalização.

O desenvolvimento do modelo MCRP segue as seguintes etapas principais:

1. **Divisão dos dados**: Separação dos dados em conjuntos de treino e teste.
2. **Construção do modelo**: Aplicação do modelo de chances de riscos proporcionais (MCRP) no conjunto de treino.
3. **Avaliação no conjunto de teste**: Aplicação do modelo ao conjunto de teste para verificar seu desempenho em dados não utilizados no treinamento.
4. **Análise de métricas**: Avaliação das métricas de desempenho, como acurácia, sensibilidade, especificidade para determinar a qualidade do modelo.

Essa abordagem busca garantir a robustez do modelo e sua capacidade de fornecer classificações confiáveis em situações práticas.

5.1 Banco de dados

Para a seleção da base de dados utilizada no presente trabalho, realizou-se uma ampla revisão bibliográfica com o objetivo de identificar bases de dados reais que atendessem aos objetivos propostos. Contudo, em virtude do elevado valor comercial associado aos dados de risco de crédito e à inclusão de informações sensíveis dos clientes, que poderiam infringir os princípios estabelecidos pela Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018), constatou-se a inexistência de bases públicas disponíveis nas publicações revisadas.

Diante dessa limitação, optou-se pela realização de uma análise detalhada da literatura existente, com o objetivo de identificar as variáveis mais frequentemente empregadas na distinção entre solicitantes de baixo e alto risco no contexto da análise de risco de crédito. Com base nessa investigação, selecionaram-se as seguintes variáveis explicativas:

- **Limite de Credito:** Refere-se ao limite de credito concedido a cada cliente. Foi utilizado a transformação log da variável para melhor modelagem dos dados.
- **Sexo:** Refere-se ao sexo de cada cliente. Foi utilizado dois níveis numéricos, com "0"representando o gênero feminino e "1"o gênero masculino.
- **Classificação Social por renda:** Indica a classificação social da renda, de acordo com os critérios da Fundação Getulio Vargas (2014). A variável foi codificada em 5 níveis: da Classe E à Classe A, com a Classe B servindo como base de comparação.
- **Estado civil:** Representa o estado civil de cada cliente, codificado da seguinte forma: "0"para casados(as), "1"para solteiro(a) e "2"para Viúvo/Separado.
- **Idade:** Variável que indica a idade de cada cliente, em anos.

A definição dos parâmetros para cada variável baseou-se no trabalho de Bogoni e Pavan (2014) que analisou como cada característica pode influenciar o risco. A idade foi considerada uma variável diretamente proporcional ao risco, com a hipótese de que clientes mais velhos podem apresentar maior probabilidade de inadimplência. Na variável sexo, homens foram associados a um maior risco em relação às mulheres, possivelmente devido a padrões comportamentais identificados em estudos prévios. Quanto à classificação por renda, níveis mais altos foram correlacionados a um maior risco, indicando que maior renda pode estar associada a maior acesso a concessões de crédito e, consequentemente, maior exposição ao risco de inadimplência. Por fim, no estado civil, indivíduos casados foram considerados de maior risco em comparação com solteiros ou viúvos/separados, possivelmente devido ao impacto de maiores responsabilidades financeiras na administração familiar.

A partir dessas variáveis, criou-se uma base de dados simulada com 1000 observações, incluindo também a variável tempo (em meses) até a ocorrência da inadimplência. Em relação à censura, observou-se que as taxas de censura geralmente se mostram elevadas, resultando em um número reduzido de casos de inadimplência registrados. Tal cenário é esperado, considerando a lucratividade característica do setor de empréstimos.

Nesse trabalho foi considerada uma taxa de censura aproximada de 50%, visando a uma aproximação aos dados reais encontrados na literatura em trabalhos como o de Dirick e Baesens (2017). Essa estratégia possibilitou que a base simulada representasse com maior fidelidade as características observadas em cenários reais de análise de risco de crédito.

As variáveis explicativas foram geradas da seguinte forma: o limite de crédito a partir de uma distribuição lognormal, o sexo por meio de uma distribuição de Bernoulli, a classificação social por uma distribuição multinomial com 5 categorias, o estado civil

por uma distribuição multinomial com 3 categorias e a idade por uma distribuição de Poisson. Os parâmetros dessas distribuições foram escolhidos de modo que seus valores (sinais) estivessem de acordo com as observações do trabalho de Bogoni e Pavan (2014). Os tempos de sobrevivência foram gerados utilizando o método da transformação inversa, partindo inicialmente de um MCRP Weibull contínuo. Os tempos de censura foram gerados combinando os mecanismos de Tipo 1 (considerando financiamentos com duração de 36 meses) e aleatório (através de uma distribuição exponencial independente do tempo de sobrevivência), com o parâmetro da distribuição exponencial ajustado para aproximar a taxa de censura desejada. Por fim, os tempos de sobrevivência resultantes foram discretizados, tomando-se a parte inteira de seus valores.

A base de dados criada subsidiou a aplicação da metodologia proposta, baseada na técnica MCRP. Essa abordagem permitiu o desenvolvimento de um modelo robusto e alinhado aos objetivos do estudo, garantindo, simultaneamente, a conformidade com os princípios éticos e de proteção à privacidade dos dados.

5.2 Análise descritiva

Nessa seção será apresentado a análise descritiva do banco de dados gerado. Será utilizado tabelas e gráficos para demonstração.

As Tabelas 2 e 3 representam uma análise descritiva das variáveis numéricas e categóricas.

Tabela 2: Medidas resumo de cada variável Numérica

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
Limite de Credito	5,704	8,006	10,332	8,010	0,69
Idade	23	40	60	39,87	6,20

Tabela 3: Medidas resumo de cada variável categórica

Variável	Nível	Frequência Absoluta	Frequência Relativa(%)
Sexo	Homem: x = 1	498	49,8
	Mulher : x = 0	502	50,2
Classe Social	Classe A : x >R$ 11.262	202	20,2
	Classe B : R$ 8.641<x <R$ 11.261	195	19,5
	Classe C : R$ 2.005 <x <R$ 8.640	191	19,1
	Classe D : R$ 1.255 <x <R$ 2.004	194	19,4
	Classe E : R$ 0 <x <R$ 1.254	218	21,8
Estado Civil	Casado : x = 0	335	33,5
	Solteiro : x = 1	354	35,4
	Viúvo/Separado : x = 2	311	31,1

A partir da Tabela 2 observa-se que os dados numéricos não apresentam uma

dispersão elevada em torno da média/mediana, exibindo um comportamento centrado nas medidas centrais, além de apresentarem valores baixos para o desvio padrão.

A Tabela 3 apresenta os dados categóricos e suas frequências absolutas e relativas. Entre as principais informações, observa-se que o comportamento dentro das classes de todas as variáveis exibe similaridade, apresentando uma proporção homogênea.

A base de dados utilizada neste trabalho consistiu em $n = 1000$ observações, das quais 469 (46,9%) foram censuradas. Todas as censuras foram do tipo à direita, indicando que o evento de inadimplência não ocorreu dentro do período do estudo.

A Figura 1 apresenta a estimativa do tempo em meses até a inadimplência, sem a consideração das variáveis explicativas.

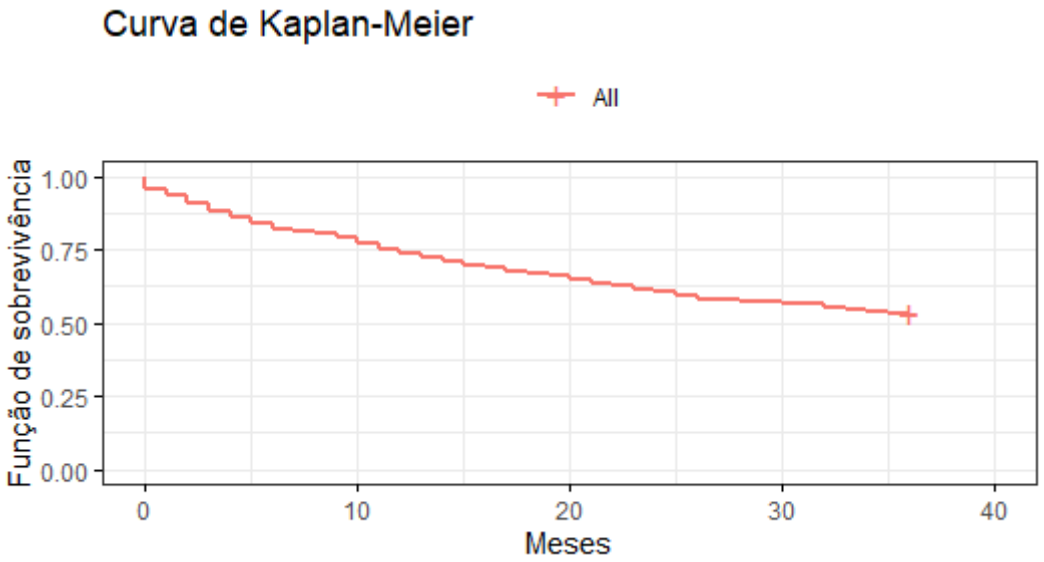

Figura 1: Estimativa de Kaplan-Meier da função de sobrevivência sem a presença das variáveis explicativas

5.3 Ajuste do MCRP

Agora, nesta seção do trabalho, será feito ajuste do MCRP. Como dito previamente, o banco de dados é dividido em "treino" e "teste", onde a tabela de treino é usada para estimação dos parâmetros e a tabela de teste é usada para avaliar o desempenho do modelo. A divisão do banco de dados (em 70% treino e 30% teste) foi realizada por meio de sorteio aleatório das observações.

Outro aspecto relevante consiste na transformação de variáveis categóricas em variáveis *dummies*, procedimento que envolve a criação de variáveis binárias 0, 1 para cada categoria, com exceção da categoria de referência. Por exemplo, no conjunto referente à Classe Social, com cinco classes (A a E), ao utilizar a Classe B como referência, um

indivíduo pertencente à Classe D seria representado com o valor 1 apenas na variável Classe D e com o valor 0 nas demais (Classes A, C e E). Tal procedimento possibilita a interpretação dos coeficientes de regressão como efeitos diferenciais em relação à categoria de referência (AGRESTI, 2013).

Considerando T como uma variável aleatória com distribuição Weibull discreta e usando a função de ligação $g(\boldsymbol{x'\beta}) = \exp(\boldsymbol{x'\beta})$, derivou-se a seguinte fórmula para a função de risco no MCRP de Weibull discreto:

$$h(t|\boldsymbol{x}) = \frac{e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}{q^{(t+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{t^\eta} - q^{(t+1)^\eta})}, t = 0, 1, 2, \dots \ . \tag{5.3.1}$$

A função de sobrevivência pode ser reescrita como

$$S(t|\boldsymbol{x}) = \prod_{u=0}^{t} \left[\frac{q^{(u+1)^\eta}}{q^{(u+1)^\eta} + e^{\boldsymbol{x'\beta}}(q^{u^\eta} - q^{(u+1)^\eta})} \right] \ t = 0, 1, 2, \dots \ . \tag{5.3.2}$$

As estimativas do MCRP Weibull discreto para os dados do banco treino são apresentadas na Tabela 4:

Tabela 4: Estimativas dos parâmetros do modelo MCRP Weibull discreto para os dados do banco treino

Variável	Estimativa	LI IC 95%	LS IC 95%
η	0,8260	0,7459	0,9062
q	0,4332	0,0413	0,8250
Limite de Crédito	-0,2231	-0,3814	-0,0647
Sexo Feminino	*		
Sexo Masculino	0,8800	0,6464	1,1135
Classe Social A	0,2309	-0,0752	0,5371
Classe Social B	*		
Classe Social C	-0,3068	-0,6501	0,0365
Classe Social D	- 0,7849	-1,1504	-0,4194
Classe Social E	- 0,7928	-1,1706	-0,4149
Estado Civil Casado	*		
Estado Civil Solteiro;	- 0,7183	-1,0093	-0,4273
Estado Civil Viúvo/Separado	0,0211	-0,2782	0,2359
Idade	- 0,0420	-0,0608	-0,0232

Nota : As variáveis com * são as de nível de referência

Com a tabela 4 pode-se observar os coeficientes β para cada variável, juntamente com seus intervalos de confiança. Valores positivos indicam um aumento nas chances de inadimplência, e valores negativos, uma redução.

Como os dados foram gerados a partir de um MCRP Weibull contínuo, a verificação da suposição de chances de riscos proporcionais torna-se desnecessária.

5.4 Obtenção do escore de risco e classificação dos indivíduos segundo a metodologia proposta

O Escore de Risco apresentado em (4.2.1) foi calculado para cada observação da amostra, com base nas estimativas dos parâmetros do modelo apresentadas na Tabela 4.

Dessa forma, o escore de risco de um cliente é estimado por:

$$\begin{aligned} ER = \boldsymbol{x'\beta} = & (-0,2231 \times \text{Limite de Crédito} + 0,8800 \times \text{Sexo Masculino} \\ & + 0,2309 \times \text{Classe Social A} - 0,3068 \times \text{Classe Social C} \\ & - 0,7849 \times \text{Classe Social D} - 0,7928 \times \text{Classe Social E} \\ & - 0,7183 \times \text{Estado Civil Solteiro} - 0,0211 \times \text{Estado Civil Viúvo/Separado} \\ & - 0,0420 \times \text{Idade }) \end{aligned} \tag{5.4.1}$$

Por exemplo, o escore de risco de um cliente, com limite de crédito $R\$1200$, do sexo feminino, da classe social C, do estado civil Viúvo/Separado e com 36 anos de idade é estimado por :

$$\begin{aligned} ER = x'\beta = & -\log(1200) \times 0,2230 \ + 0,8800 \times 0 \ + 0,2309 \times 0 \ - 0,3068 \times 1 \\ & - 0,7849 \times 0 \ - 0,7928 \times 0 \\ & - 0,7183 \times 0 \ - 0,0211 \times 1 \\ & - 0,0420 \times 36. = 0,0326 \end{aligned} \tag{5.4.2}$$

O escore de risco (5.4.1) foi calculado para cada indivíduo da amostra de teste. A Figura 2 apresenta a densidade da distribuição dos escores obtidos, de acordo com as duas categorias da variável resposta (Inadimplência/Adimplência).

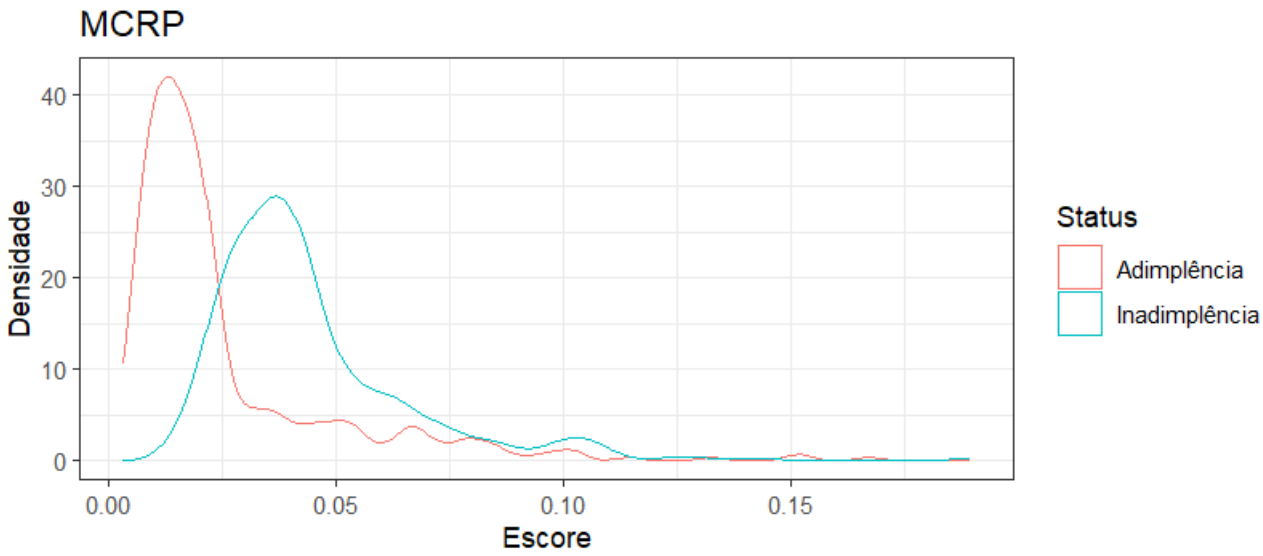

Figura 2: Gráfico da densidade dos escores de risco

A partir da Figura 2 , observa-se que a categoria de menor risco apresentou um pico entre os escores mais baixos, evidenciando uma diferenciação em relação à categoria de maior risco.

O principal objetivo da determinação do escore de risco consiste em minimizar: o erro de categorizar clientes com alto risco como pertencentes a uma categoria de baixo risco e o erro de categorizar clientes com baixo risco como pertencentes a uma categoria de alto risco.

A Figura 3 apresenta o percentual dos erros associados a cada nível de corte no banco de treino.

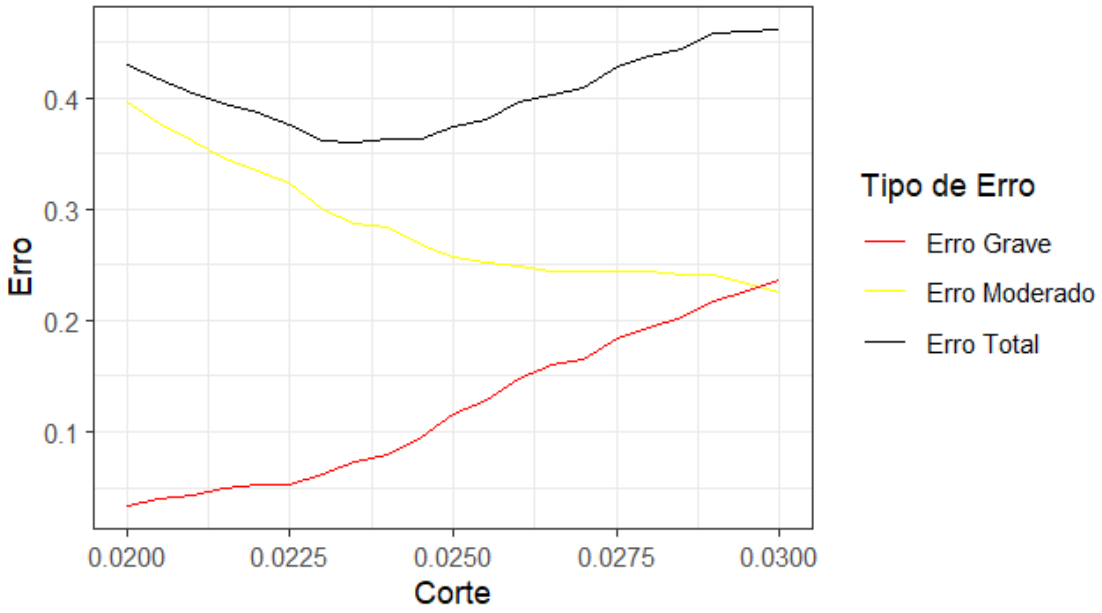

Figura 3: Gráfico do erros associados ao corte

A partir da análise da Figura 3, percebe-se que o erro moderado e o erro total apresentam uma diminuição à medida que o nível de corte aumenta, com seus mínimos próximos a $0,03$ e $0,0237$, respectivamente. Em contrapartida, o erro grave exibe um comportamento crescente à medida que o nível de corte aumenta, apresentando seu mínimo próximo ao ponto de corte $0,02$.

Na escolha do escore, consideraram-se os dois tipos de erro na classificação dos clientes. O erro do Tipo I, considerado o mais grave, consiste em classificar clientes com alto risco como clientes com baixo risco, acarretando, assim, uma perda de capital nas operações. O erro do Tipo II, por sua vez, consiste na classificação de clientes com baixo risco como clientes com alto risco, o que acarreta a perda de potenciais clientes. Buscou-se minimizar ambos os cenários, com prioridade para o erro do Tipo I.

Considerando os pontos mencionados anteriormente, determinou-se o ponto de corte em $0,0235$. Ou seja, clientes com escore de risco inferior a $0,0235$ foram classificados como de baixo risco, e clientes com $ER \geq 0,0235$ foram classificados como de alto risco. Tais pontos de corte foram definidos com base na Figura 2, visando ao controle dos erros do Tipo I e do Tipo II (probabilidade de erro do Tipo I inferior a 10% e probabilidade de erro do Tipo II inferior a 36%), além de buscar um nível de erro total próximo ao mínimo possível.

Assim, um indivíduo com ER = 0,0326 (5.4.2) é classificado como de alto risco de inadimplência. Ademais, o valor do Escore de Risco foi calculado para todos os indivíduos dos bancos treino e teste e seus resultados são apresentados nas Tabelas 5 e 6, respectivamente.

Tabela 5: Classificação - banco treino

	Predito		
Observado	Adimplência	Inadimplência	Total
Adimplência	267	107	374
Inadimplência	24	302	326
% de acerto	91,75%*	73,84%**	81,29%***

Nota : * Valor preditivo positivo ** valor predivivo negativo *** Acerto total

O modelo de MCRP apresentou um bom desempenho, com $81,29\%$ de acerto global, alcançando $91,75\%$ na categorização dos solicitantes de menor risco e uma porcentagem próxima do acerto global para os solicitantes de maior risco. Conforme mencionado anteriormente, o principal objetivo consistia na obtenção de um escore que classificasse de maneira adequada as categorias de maior risco para as instituições financeiras.

Tabela 6: Classificação - banco teste

	Predito		
Observado	Adimplência	Inadimplência	Total
Adimplência	120	37	157
Inadimplência	12	131	143
% de acerto	90,90%*	77,98%**	83,67%***

Nota : * Valor preditivo positivo ** valor predivivo negativo *** Acerto total

A análise da Tabela 6 revela que os dados se mostram próximos aos do banco de treino, com $83,67\%$ de acerto global e com acertos nas categorias bastante similares. Tal característica demonstra que o modelo apresentou bom desempenho em relação a novos dados, com resultados satisfatórios.

5.5 Comparativo de Desempenho Preditivo

Nesta seção, comparamos o desempenho preditivo do modelo proposto com os modelos de regressão logística e de riscos proporcionais de Cox, ambos amplamente utilizados na elaboração de escores de risco.

Os resultados monstram que o MCRP apresentou desempenho superior ou semelhante aos modelos comparados. Especificamente, o modelo logístico alcançou 74% de acertos no banco de treino e 73% no banco de teste, resultados inferiores aos do MCRP. Além da acurácia, o MCRP se destaca por ser mais informativo que o modelo logístico, fornecendo dados sobre a curva de risco e a sobrevivência ao longo do tempo, úteis para

o desenvolvimento de estratégias de negócio. Além de que o modelo logístico não leva em consideração as observações censuradas como o MCRP.

Por sua vez, o modelo de Cox obteve 82% de acertos no banco de treino e 85% no banco de teste, resultados próximos aos do MCRP. Contudo, o MCRP apresenta uma vantagem conceitual significativa: diferentemente do modelo de riscos proporcionais de Cox, que pressupõe uma distribuição contínua para a variável tempo, o MCRP considera sua distribuição discreta. Essa característica é particularmente relevante em contextos onde o tempo de evento é naturalmente discreto ou medido em intervalos, conferindo ao MCRP uma maior adequação para tais dados.

6 Considerações finais

Diante do aquecimento da economia e do aumento da oferta de crédito, as instituições financeiras constataram a necessidade de estruturar processos objetivos e eficientes para a gestão dos riscos em todas as etapas do ciclo de crédito. Contudo, a expansão do crédito acarreta também o aumento da inadimplência, aspecto que motivou o desenvolvimento do presente trabalho.

A proposta do trabalho consistiu no desenvolvimento de um escore de risco e na avaliação de sua capacidade preditiva para a classificação de solicitantes em categorias de risco.

A escolha do modelo de chances de risco proporcionais justificou-se pela característica da informação do tempo, que apresenta uma distribuição discreta ao longo da mensuração dos meses. A vantagem da técnica MCRP reside no ganho de informação proveniente das técnicas de análise de sobrevivência, tais como a função de sobrevivência e a função de risco ao longo do tempo. Dessa forma, além da classificação final dos clientes, o estudo apresenta informações sobre o comportamento ao longo do tempo.

O trabalho empregou um banco de dados simulado, elaborado com base em estudos da literatura. Tal banco de dados possibilitou o desenvolvimento do escore de risco e a classificação de clientes entre as categorias de Inadimplentes e Adimplentes.

Os resultados observados demonstraram que o escore de risco proposto neste trabalho se mostrou adequado para a classificação de clientes, apresentando desempenhos consistentes tanto na amostra de treino quanto na amostra de teste, além de exibir resultados superiores aos de técnicas comuns na literatura e na indústria, como o modelo logístico.

O objetivo de propor um escore de risco consistente foi alcançado. As constatações apresentadas neste trabalho representam uma contribuição para a análise de dados discretos e ampliam a literatura sobre o tema.

Como tópicos para estudos futuros, a possível aplicação da metodologia proposta em bancos de dados maiores e reais, com a inclusão de mais variáveis explicativas significativas podem ser realizados. Ademais, outros modelos de regressão para dados discretos de sobrevivência com estrutura proporcional também podem ser considerados, como por exemplo modelo de chances de sobrevivência proporcionais (CARDIAL; COBRE; NAKANO, 2025) ou o modelo log sobrevivências proporcionais (CHANDIONA; CARDIAL; NAKANO, 2025).

Referências

AGRESTI, A. *Categorical data analysis*. 3ed.. ed. [S.l.]: Wiley, 2013. 42

AGRESTI, A. *An Introduction to Categorical Data Analysis*. 3rd. ed. [S.l.]: John Wiley & Sons, 2019. (Wiley Series in Probability and Statistics). 36

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 1968. 34

BERGER, M.; SCHMID, M. Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, v. 18, n. 3-4, p. 322–345, 2018. Disponível em: ⟨https://doi.org/10.1177/1471082X17748084⟩. 15, 16

BERGER, M. et al. A classification tree approach for the modeling of competing risks in discrete time. *Advances in Data Analysis and Classification*, v. 13, n. 4, p. 965–990, 2019. 16

BOGONI, N. M.; PAVAN, R. Análise de inadimplência de crédito e microcrédito em uma cooperativa de crédito localizada na região norte do estado do rio grande do sul (rs): utilização do modelo econométrico de logit. *convibra*, convibra, 2014. 39, 40

BRESLOW, N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, [Wiley, International Statistical Institute (ISI)], v. 43, n. 1, p. 45–57, 1975. ISSN 03067734, 17515823. Disponível em: ⟨http://www.jstor.org/stable/1402659⟩. 24

BRUNELLO, G. H. V.; NAKANO, E. Y. Inferência bayesiana no modelo weibull discreto em dados com presença de censura. *Trends in Computational and Applied Mathematics*, v. 16, n. 2, p. 97–110, 2015. Disponível em: ⟨https://doi.org/10.5540/tema.2015.016.02.0097⟩. 15, 17

CARDIAL, M. R. P.; COBRE, J.; NAKANO, E. Y. A discrete weibull proportional odds survival model. *Journal of Applied Statistics*, v. 52, n. 2, p. 429–447, 2025. Disponível em: ⟨https://doi.org/10.1080/02664763.2024.2373929⟩. 48

CARDIAL, M. R. P.; FACHINI-GOMES, J. B.; NAKANO, E. Y. Exponentiated discrete weibull distribution for censored data. *Brazilian Journal of Biometrics*, v. 38, n. 1, p. 35–56, Mar. 2020. Disponível em: ⟨http://www.biometria.ufla.br/index.php/BBJ/article/view/425⟩. 15, 17

CHAKRABORTY, S. Generating discrete analogues of continuous probability distributions-a survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2015. Disponível em: ⟨https://doi.org/10.1186/s40488-015-0028-6⟩. 16, 17

CHANDIONA, T. C. E. F.; CARDIAL, M. R. P.; NAKANO, E. Y. Proportional log survival model for discrete time-to-event data. *Mathematics*, v. 13, n. 5, p. 800, 2025. Disponível em: ⟨https://doi.org/10.3390/math13050800⟩. 48

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. [S.l.]: Edgard Blucher, 2006. 9, 19, 21, 23

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 34, n. 2, p. 187–202, 1972. Disponível em: ⟨https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x⟩. 21

DANI, G.; FRANCISCO, L.-N.; MIGON, H. d. S. *Statistical inference : an integrated approach.* Second edition. [S.l.]: Chapman & Hall/CRC, 2014. (Chapman & Hall/CRC texts in statistical science series). ISBN 9781439878828,143987882X. 20

DIRICK, G. C. L.; BAESENS, B. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, Taylor & Francis, v. 68, n. 6, p. 652–665, 2017. Disponível em: ⟨https://doi.org/10.1057/s41274-016-0128-9⟩. 39

DURAND, D. *Risk Elements in Consumer Instalment Financing.* [S.l.]: National Bureau of Economic Research, 1941. 33

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936. 33

Fundação Getulio Vargas. *Qual a faixa de renda familiar das classes?* 2014. Acesso em: 24 jan. 2025. Disponível em: ⟨https://cps.fgv.br/qual-faixa-de-renda-familiar-das-classes⟩. 39

JAMES, B. R. *Probabilidade um curso intermediário.* 4. ed. [S.l.]: IMPA, 2015. (Projeto Euclides). 13

JAYAKUMAR, K.; BABU, M. G. Discrete weibull geometric distribution and its properties. *Communications in Statistics - Theory and Methods*, Taylor & Francis, v. 47, n. 7, p. 1767–1783, 2018. Disponível em: ⟨https://doi.org/10.1080/03610926.2017.1327074⟩. 17

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, American Statistical Association, Taylor & Francis, Ltd., v. 53, n. 282, p. 457–481, 1958. ISSN 01621459. Disponível em: ⟨http://www.jstor.org/stable/2281868⟩. 19

MACHADO, A. R. *Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionais de Cox.* Dissertação (Mestrado) — Universidade de Brasília, 2015. 32

MEYER, P. *Probabilidade: Aplicações à Estatística.* [S.l.]: LTC - GRUPO GEN, 1983. 10

MYERS, J. H.; FORGY, E. W. The development of numerical credit evaluation systems. *Journal of the American Statistical Association*, Taylor & Francis, 1963. Disponível em: ⟨https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500889⟩. 33, 34

NAKAGAWA, T.; OSAKI, S. The discrete weibull distribution. *IEEE Transactions on Reliability*, R-24, n. 5, p. 300–301, 1975. 17, 18

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Computational and Applied Mathematics*, v. 7, n. 1, p. 91–100, 2006. Disponível em: ⟨https://tcam.sbmac.org.br/tema/article/view/276⟩. 17

ORGLER, Y. E. A credit scoring model for commercial loans. *Journal of Money, Credit Banking (Ohio State University Press) 2, 435-445*, 1970. 34

SANTOS, D. F. dos. *Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência*. Dissertação (Mestrado) — Universidade de Brasília, 2017. 18

SANTOS, W. de Almeida Aguiar Yamamoto; Edson Aparecida de Araújo Querido Oliveira ; Vilma da S. O gerenciamento de risco de crédito em um banco de varejo: um estudo do segmento pessoas físicas. *XV Encontro Latino Americano de Iniciação Científica e XI Encontro Latino Americano de Pós-Graduação*, 2011. 32

SARHAN, A. M. A two-parameter discrete distribution with a bathtub hazard shape. *Communications for Statistical Applications and Methods*, The Korean Statistical Society, and Korean International Statistical Society, v. 24, n. 1, p. 15–27, 2017. 17

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. *Biometrika*, v. 69, n. 1, p. 239–241, 04 1982. ISSN 0006-3444. Disponível em: ⟨https://doi.org/10.1093/biomet/69.1.239⟩. 25

TUTZ, G.; SCHMID, M. *Modeling Discrete Time-to-Event Data*. 1. ed. [S.l.]: Springer International Publishing, 2016. (Springer Series in Statistics). ISBN 978-3-319-28156-8, 978-3-319-28158-2. 15, 16

VALLEJOS, C. A.; STEEL, M. F. J. Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, v. 180, n. 2, p. 613–631, 07 2016. ISSN 0964-1998. Disponível em: ⟨https://doi.org/10.1111/rssa.12211⟩. 16

VIEIRA, M. G. F. et al. Proportional odds hazard model for discrete time-to-event data. *Axioms*, v. 12, n. 12, p. 1102, 2023. Disponível em: ⟨https://doi.org/10.3390/axioms12121102⟩. 8, 17, 27, 30

WEINGARTNER, H. *Concepts and Utilization of Credit-Scoring Techniques*. [S.l.]: Banking 58,51-54, 1966. 34

Código para Estimativas MRCP e do ER

```
    ########################
##### COVARIAVEIS #####
########################
# x1. Limite de credito (continua)
# x2. Sexo (categorica 2 niveis)
# x3. Classe Social (categorica 5 niveis: x31, x32, x33 e x34)
# x4. Estado Civil (categorica 3 niveis: x41 e x42)
# x5. Idade (discreta)


##################################
##### definicao das funcoes #####
##################################

## risco base - WEIBULL DISCRETA
h0<-function(t,q,eta){
  (q^(t^eta)-q^((t+1)^eta))/q^(t^eta)
}

## funcao de ligacao
link.theta<-function(beta1,beta2,beta31,beta32,beta33,beta34,
beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5){
  g <- exp(beta1 *x1     +
           beta2 *x2     +
           beta31*x31    +
           beta32*x32    +
           beta33*x33    +
           beta34*x34    +
           beta41*x41    +
           beta42*x42    +
           beta5 *x5)
  g
}

## risco na presenca de covariaveis
h.x<-function(t,q,eta,beta1,beta2,beta31,beta32,beta33,beta34,
beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5){
  g <- link.theta(beta1,beta2,beta31,beta32,beta33,beta34,beta41,
```

```
  beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5)
  g*h0(t,q,eta) / (1-h0(t,q,eta) + g*h0(t,q,eta))
}

### sobrevivencia na presenca de covariaveis
s.x<-function(t,q,eta,beta1,beta2,beta31,beta32,beta33,beta34,
beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5){
  sobrev<-1
  for (u in 0:t){
    sobrev<-sobrev * (1-h.x(u,q,eta,beta1,beta2,beta31,beta32,
    beta33,beta34,beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,
    x41,x42,x5))
  }
  return(sobrev)
}

### distribuicao de probabilidades na presenca de covariaveis
p.x<-function(t,q,eta,beta1,beta2,beta31,beta32,beta33,beta34,
beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5){
  if (t==0) return(h.x(0,q,eta,beta1,beta2,beta31,beta32,beta33,
  beta34,beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5))
  else return(h.x(t,q,eta,beta1,beta2,beta31,beta32,beta33,beta34,
  beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5)
  *s.x(t-1,q,eta,beta1,beta2,beta31,beta32,beta33,
  beta34,beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5))
}

### log verossimilhanca
log_vero_MCRP <- function(parametros,t,x1,x2,x31,x32,x33,x34,x41,
x42,x5,censura) {
  q     <- parametros[1]
  eta   <- parametros[2]
  beta1 <- parametros[3]
  beta2 <- parametros[4]
  beta31<- parametros[5]
  beta32<- parametros[6]
  beta33<- parametros[7]
  beta34<- parametros[8]
  beta41<- parametros[9]
```

```
  beta42<- parametros[10]
  beta5 <- parametros[11]

  theta <- link.theta(beta1,beta2,beta31,beta32,beta33,beta34,
  beta41,beta42,beta5,x1,x2,x31,x32,x33,x34,x41,x42,x5)
  if ((q>0)&&(q<1)&&(eta>0)) {
    lvero<-0
    for (i in 1:length(t)){
      lvero<-lvero + ( censura[i] *log(p.x(t[i],q,eta,beta1,
      beta2,beta31,beta32,beta33,beta34,beta41,beta42,beta5,
      x1[i],x2[i],x31[i],x32[i],x33[i],x34[i],x41[i],x42[i],x5[i]))
                  + (1-censura[i])*log(s.x(t[i],q,eta,beta1,
beta2,beta31,beta32,beta33,beta34,beta41,
                beta42,beta5,x1[i],x2[i],x31[i],x32[i],x33[i],
                x34[i],x41[i],x42[i],x5[i])) )
    }
    return(-1*lvero)
  }
  else return(-Inf)
}


### obtencao das estimativas dos parametros do MCRP
chute.inicial
est <- nlm(log_vero_MCRP,chute.inicial,
            t=tempo,censura=censura,
            x1=x1,x2=x2,x31=x31,x32=x32,x33=x33,
            x34=x34,x41=x41,x42=x42,x5=x5,
            iterlim = 500,hessian=T)
est
est_beta<-est$estimate[3:11]

ES<- exp(est_beta[1]*x1 +
         est_beta[2]*x2 +
         est_beta[3]*x31 +
         est_beta[4]*x32 +
         est_beta[5]*x33 +
         est_beta[6]*x34 +
         est_beta[7]*x41 +
```

```
        est_beta[8]*x42 +
        est_beta[9]*x5)

### figura com os ER
plot(density(ES[censura==0]))
points(density(ES[censura==1]),col=2,type="l")
```