Universidade de Brasília
Departamento de Estatística

Um escore de risco via modelo de regressão de Cox com covariáveis dependentes no tempo

Camila Carneiro Brito Bomfim

Monografia apresentada para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Camila Carneiro Brito Bomfim

Um escore de risco via modelo de regressão de Cox com covariáveis dependentes no tempo

Orientador(a): Prof. Eduardo Yoshio Nakano

Monografia apresentada para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Resumo

O objetivo deste trabalho é propor um escore de risco baseado no tempo até o atraso do pagamento de um tomador, utilizando o modelo de regressão de Cox com covariáveis dependentes no tempo. O banco de dados utilizado é uma simulação de dados básicos de clientes de uma instituição financeira e o modelo verifica a influência de cada variável no tempo do tomador se tornar inadimplente. As análises foram realizadas por meio do *software R*.

Palavras-chave: Análise de Sobrevivência, Modelo de Regressão, Dados Censurados, Riscos Proporcionais de Cox, Behavior Score, Covariáveis Dependentes no Tempo.

Abstract

The aim of this work is to propose a risk score based on the time until default of payment using Time-varying covariates in Cox regression models. The database used is a simulation of basic customer data from a financial institution and the model verifies the weight of each covariates on the time the customer becomes non-performing. All analysis were done using *software R*.

Keywords: Survival Analysis, Regression Model, Censored Data, Cox Proportional Hazards, Behavior Score, Time-varying covariates.

Sumário

1 Introdução

A Análise de Sobrevivência é uma área da estatística que vem ganhando cada vez mais destaque com o avanço da tecnologia, envolvendo principalmente a ciência de dados. Essa área consiste no uso de técnicas e modelos estatísticos para estipular o tempo de duração esperado até a ocorrência de um ou mais eventos de interesse. De forma geral, a unidade em estudo pode ser desde um objeto até um animal, mas usualmente os indivíduos são os mais análisados nessa área. Diante disso, a composição de um estudo de análise de sobrevivência inclui a unidade de estudo, as covariáveis que serão analisadas e o tempo que representa o objeto principal do estudo, sendo esse tempo chamado de tempo de sobrevivência, tempo de falha ou tempo de vida.

As covariáveis referidas anteriormente, geralmente são medidas em um único momento ao longo do tempo e um dos modelos de regressão mais populares na análise de dados de sobrevivência é o modelo de riscos proporcionais de Cox (1972). No entanto, alguns problemas podem apresentar situações em que as covariáveis são medidas periodicamente ou se alteram ao longo do tempo.

No contexto de modelagem de risco de crédito, geralmente é obtido um escore que avalia o risco de ocorrência de perdas associadas ao não cumprimento pelo tomador de suas obrigações financeiras. Neste tipo de problema, essa perda é percebida principalmente devido à falta (ou atraso) de pagamento das prestações. Esses escores de risco (credit scoring) são geralmente obtidos por modelos de regressão logística (quando o desfecho é a ocorrência do atraso) ou modelos de regressão em análise de sobrevivência (quando o desfecho é o tempo até o atraso) considerando as covariáveis fixas. No entanto, considerar fixas as covariáveis pode ser uma limitação na modelagem, principalmente nos modelos de Behaviour Scoring, que são aplicados em clientes ativos visando a reavaliação/renovação do contrato. De fato, o atraso do pagamento pode ser devido à uma mudança de situação durante a vigência do contrato, como a redução da renda, por exemplo.

Neste contexto, o objetivo desse trabalho é propor um escore de risco baseado no tempo até o atraso do pagamento, utilizando o modelo de regressão de Cox com covariáveis dependentes no tempo. De início, será realizada uma revisão do modelo de Cox (para covariáveis fixas), seguida por uma estruturação e explicação do modelo de regressão de Cox com covariáveis dependentes no tempo, para posteriormente propor um escore de risco com base no modelo. Por fim, a metodologia proposta neste trabalho será ilustrada com um conjunto de dados simulados de uma instituição financeira.

2 Conceitos Básicos em Análise de Sobrevivência

A análise de sobrevivência é uma técnica muito utilizada na área da saúde, devido ao fato de envolver covariáveis que podem estar relacionadas com o tempo de sobrevivência de um depeterminado objeto de estudo. Entretanto, além da área da saúde, a análise de sobrevivência vem crescendo bastante, devido ao número de aplicações também na área financeira.

Em análise de sobrevivência, a variável resposta é constituída de dois componentes: o tempo de falha e as censuras, ambos caracterizam os dados de sobrevivência. Segundo Colosimo e Giolo (2006), o tempo de falha é constituído por três elementos (tempo inicial, escala de medida e evento de interesse), estes devem ser claramente definidos. O tempo inicial é quando começa a ser realizado o estudo, utilizado para comparação dos indivíduos na origem do estudo. A escala de medida é o "tempo" que será contabilizado, por exemplo, tempo real, meses, semanas, dias, números de ciclos, medidas de carga e muitas outras. O evento de interesse é a própria falha, na maioria dos casos indesejável, em análises de sobrevivência deve ser definido de forma clara e precisa. Após definir os três elementos, determina-se a variável tempo de falha. Além disso, outro fator importante que deve ser ressaltado são as denominadas censuras, que de acordo com Nakano e Carrasco (2006), são observações parciais da resposta, mas são informações úteis e importantes para a análise.

2.1 Censura

As censuras ocorrem, pois os estudos de análise de sobrevivência envolvem uma resposta temporal. Diante disso, existe a possibilidade de que um estudo termine antes que ocorra o evento de interesse para todos os casos da amostra, tornando algumas informações incompletas ou parciais. Exitem três meios de censura mais conhecidos. A censura do tipo I, ocorre quando o estudo finaliza após um tempo estabelecido. As observações que não apresentarem o evento de interrese durante esse período são ditas censuras. Outro tipo de censura, a do tipo II é aquela onde se estabelece um número de indivíduos para que ocorra o evento de interesse e após isso se encerra o estudo. Por fim, o terceiro tipo de censura é a do tipo aleatória, que ocorre se a observação for retirada no decorrer do estudo sem que o evento de interesse tenha ocorrido.

As censuras apresentadas anteriormente são conhecidas por censura à direita, pois o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Sendo essa, a censura mais encontrada em estudos que envolvem dados de sobrevivência. Entretanto, outras duas formas de censura podem ocorrer: censura à esquerda e intervalar.

2.2 Representações do Tempo de Sobrevivência

Em análise de sobrevivência, o tempo de falha também é especificado pela função de sobrevivência e pela função de taxa de falha. Além dessas, a função densidade de probabilidade também se enquadra entre as três funções básicas utilizadas na análise de sobrevivência.

Colosimo e Giolo (2006) definem a função de sobrevivência como sendo a probabilidade de uma observação sobreviver até o tempo "t". Em termos probabilísticos, para uma variável aleatória contínua positiva, define-se a função de sobrevivência como:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)\, du, t > 0, \tag{2.2.1}$$

em que $f(.)$ é a função de densidade da variável T.

Hoel, Port e Stone (1978) definem a função de distribuição entre o intervalo [0,1] para todo valor de t e F uma função não decrescente de t. Então, como $S(t) = 1 - F(t)$, afirma-se que a função de sobrevivência é não crescente e também definida no intervalo [0,1].

Uma importante função para especificar o tempo de falha é a função taxa de falha que segundo Colosimo e Giolo (2006), é a probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$. Essa função pode ser expressa em termos da função de sobrevivência como:

$$S(t_1) - S(t_2). \tag{2.2.2}$$

Essa taxa de falha é definida como a probabilidade de que a falha tenha ocorrido nesse intervalo e não tenha ocorrido antes do instante t_1, dividida pelo comprimento do intervalo. Sendo assim expressa por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{2.2.3}$$

De forma geral, pode ser expressa por:

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{2.2.4}$$

Assumindo também um Δt muito pequeno, $h(t)$ representa a taxa de falha instantânea. Logo, a taxa de falha de T é definida como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2.5}$$

Outra função importante é a taxa de risco acumulada, que fornece a taxa de falha acumulada do indivíduo, sendo definida por:

$$H(t) = \int_0^t h(u)\, du. \tag{2.2.6}$$

Colosimo e Giolo (2006) evidenciam algumas relações matemáticas importantes, são elas:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(log(S(t)), \tag{2.2.7}$$

$$H(t) = \int_0^t h(u)\, du = -logS(t) \tag{2.2.8}$$

e

$$S(t) = exp\{-H(t)\} = exp\{-\int_0^t h(u)\, du\}. \tag{2.2.9}$$

Dada essas relações, é possível verificar que o conhecimento de uma função, implica no conhecimento das demais.

A presença da censura não permite que a análise descritiva usual seja realizada. Uma alternativa para esse problema é utilizar o estimador de Kaplan-Meier, para estimar a função de sobrevivência e a função de taxa de falha. Com isso, é possível realizar o estudo descritivo dos dados de análise de sobrevivência.

2.3 Kaplan-Meier

O estimador de Kaplan-Meier é uma técnica estatística não paramétrica para a estimação da função de sobrevivência. Nessa técnica a função de sobrevivência $S(t)$ é caracterizada como uma função escada com degraus nos tempos observados de falha. A partir do Estimador de Kaplan-Meier é possível comparar os tempos de falha segundo variáveis qualitativas.

Para obter a estimativa de Kaplan-Meier, obtem-se uma sequência de passos, em que o próximo passo depende do anterior. Ou seja, para encontrar a estimativa para um tempo t_j, devemos considerar a probabilidade de sobrevivência em, $j = 1, 2, ...k$. De forma matemática, é possível definir por:

$$S(t_j) = P(T \geq t_j) = P(T \geq t_{j-1}, T \geq t_j) \tag{2.3.1}$$

$$S(t_j) = P(T \geq t_{j-1})P(T \geq t_j | T \geq t_{j-1}). \tag{2.3.2}$$

Sabendo que $S(t)$ é uma função com degraus, ou seja, com probabilidade maior do que zero apenas nos tempos de falha t_j, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2)...(1 - q_j), \tag{2.3.3}$$

em que q_j é a probabilidade de uma observação falhar no intervalo $[t_{j-1}, t_j)$ sabendo que sobreviveu até o t_{j-1}.

Para o estimador de Kaplan-Meier, estima-se q_j como:

$$\hat{q_j} = \frac{\text{número de falhas em } t_j}{\text{número de observações sob risco em } t_{j-1}} \tag{2.3.4}$$

Diante disso, o estimador de Kaplan-Meier, então definido como:

$$\hat{S(t)} = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j}), \tag{2.3.5}$$

onde:

- $t_1 < t_2 < ... < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em $t_j, j = 1, ..., k$,
- n_j número de observações sob risco em t_j.

Pode-se utilizar outros dois estimadores não-paramétricos, são eles: Nelson-Aalen, proposto por Nelson (1972) e estudado por Aalen (1978) e a Tabela de Vida ou Atuarial, que constrói uma tabela de vida e divide o eixo do tempo em vários intervalos. Os resultados desses dois estimadores são bem próximos dos estimados pelo Kaplan-Meier.

3 Modelo de riscos proporcionais de Cox

Esse modelo é popular por não assumir distribuição de probabilidade para o tempo de sobrevivência. O modelo considera as covariáveis de interesse para modelar dados de sobrevivência, por meio da função de risco.

A forma geral do modelo é expressa por

$$h(t) = h_0(t)g(x'\beta), \tag{3.0.1}$$

em que t é o tempo e x' é o vetor das p covariáveis explicativas do modelo. Nesse modelo observa-se um componente não-paramétrico, $h_0(t)$, que não é especificado e é uma função não-negativa do tempo. E o componente paramétrico, que frequentemente é usado da seguinte forma:

$$g(x'\beta) = exp(x'\beta) = exp(\beta_1 x_1 + ... + \beta_p x_p) \tag{3.0.2}$$

em que β é o vetor de parâmetros associados às p covariáveis explicativas e $h_i(t)$ é a função taxa de falha para o i-ésimo indivíduo no tempo t.

A suposição básica para o uso do modelo de riscos proporcionais de Cox é que as taxas de falha sejam proporcionais, ou seja, a razão das taxas de falha de dois diferentes indivíduos, i e j, é constante no tempo. Essa razão pode ser expressa por:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)exp\{x_i'\beta\}}{h_0(t)exp\{x_j'\beta\}} = exp\{x_i'\beta - x_j'\beta\}, \tag{3.0.3}$$

que não depende do tempo.

3.1 Ajuste do Modelo de Cox

Para a estimação dos parâmetros desconhecidos, o modelo de Cox utiliza a função de verossimilhança parcial com as censuras, através do conjunto de risco.

Segundo Colosimo e Giolo (2006), mantendo a notação utilizada anteriormente, considerando uma amostra de n indivíduos, a verossimilhança parcial considera o seguinte argumento condicional: a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i conhecendo quais observações estão sob risco em t_i é:

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

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

$$= \frac{h_i(t|x_i)}{\sum_{j\in R(t_i)} h_j(t|x_j)} = \frac{h_0(t)exp(x_i'\beta)}{\sum_{j\in R(t_i)} h_0(t)exp(x_i'\beta)} = \frac{exp(x_i'\beta)}{\sum_{j\in R(t_i)} exp(x_i'\beta)}. \tag{3.1.1}$$

Cox propôs a utilização do registro histórico passado de falhas e censuras em forma de probabilidade condicional para eliminar o termo não paramétrico da função de verossimilhança.

Diante disso, a função de verossimilhança a ser utilizada é expressa por:

$$L(\beta) = \prod_{i=1}^{k} \left(\frac{exp(x_i'\beta)}{\sum_{j\in R(t_i)} exp(x_j'\beta)} \right)^{\delta_i}. \tag{3.1.2}$$

com δ_i o indicador de falha. Caso ocorra empates nos valores observados, a função de verossimilhança parcial deve ser modificada para incorporar essas observações empatadas. Breslow (1975) propõe uma aproximação, considerando s_i o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i e d_i o número de falhas neste mesmo tempo. A aproximação considera a seguinte função de verossimilhança parcial

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(s_i'\beta)}{[\sum_{j\in R(t_i)} exp(x_j'\beta)]^{d_i}}. \tag{3.1.3}$$

Com as estimativas dos $\beta's$ e os erros-padrão, é possível estimar um intervalo de $100(1-\alpha)\%$ de confiança para determinado β_i a partir do percentil da distribuição Normal Padrão. Caso o intervalo calculado não inclua o valor zero, então pode-se dizer que há evidências para afirmar que o coeficiente β_i é diferente de zero.

O Modelo de Cox contém a suposição de riscos proporcionais, ou seja, as funções de risco não se cruzam. Tal suposição pode ser verificada por meio de métodos gráficos, onde se o modelo de riscos proporcionais for apropriado, os gráficos dos resíduos versus o tempo, para cada uma das p covariáveis, não deveriam exibir tendências ao longo do tempo t. Além disso, pode-se aplicar também testes como método de verificação.

Para verificar a qualidade do ajuste global do modelo, podem ser utilizados os resíduos de Cox-Snell, proposto por Cox e Snell (1968). Esses resíduos são definidos por:

$$\hat{e_i} = \hat{H}_0(t_i)exp\left\{ \sum_{k=1}^{p} x_{ip}\hat{\beta}_k \right\}, i = 1, 2, ..., n. \tag{3.1.4}$$

com $\hat{H}_0(t_i)$ estimado por Breslow (1975).Considerando adequado o ajuste do modelo, têm distribuição exponencial padrão, ou seja, o gráfico dos resíduos $\hat{e}_i$ versus a função de risco acumulada deve ser aproximadamente uma reta.

3.2 Interpretação dos Coeficientes

Segundo Giolo (1994), a interpretação dos parâmetros do modelo medem o efeito das covariáveis sobre a taxa de falha, ou seja, esse efeito pode ser acelerar ou desacelerar a função de risco.

De acordo com Machado (2015), a interpretação do coeficiente de uma covariável no modelo de Cox pode ser dado como o logaritmo da razão de risco do evento de dois indivíduos com atributos diferentes para uma mesma variável.

Assumindo que essa diferença é de uma unidade então pode-se assumir que o risco de se observar o evento de interesse para o indivíduo com maior valor da covariável, é $exp(\hat{\beta})$ vezes o risco para o outro indivíduo. Generalizando para quando essa diferença é de y unidades, tem-se que a taxa de falha é $exp(y\hat{\beta})$ vezes o risco do indivíduo com menor valor da covariável.

Caso a variável seja categórica, então assume-se que determinado grupo é referência e compara-se com os demais.

3.3 Avaliação do Modelo

O modelo de Cox não se ajusta a qualquer situação, e assim como acontece em qualquer outro modelo estatístico, demanda o uso de técnicas para avaliar a sua adequação.

Uma das técnicas mais utilizadas atualmente é a análise de resíduos de Schoenfeld. Esse resíduo pode ser utilizado tanto por meio de teste de hipótese quanto por técnicas gráficas.

Ao considerar o i-ésimo indivíduo, correspondente a um evento, com covariáveis $x_i = (x_{i1}, ..., x_{ip})'$, o vetor de resíduos de Schoenfeld $r_i = (r_{i1}, ..., r_{ip})$ é definida para cada componente r_{iq}, $q = 1, ..., p$, por:

$$r_{iq} = x_{iq} \frac{\sum_{j \in R(t_i)} x_{jq} exp\{x_j' \hat{\beta}\}}{\sum_{j \in R(t_i)} exp\{x_j' \hat{\beta}\}} \tag{3.3.1}$$

Os resíduos padronizados de Schoenfeld são dados por:

$$s_i^* = [I(\hat{\beta})]^{-1} \times r_i \tag{3.3.2}$$

em que $I(\hat{\beta})$ é a matriz de informação observada.

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

4 Modelo de Cox com covariáveis dependentes no Tempo

4.1 Covariáveis Dependentes no Tempo

Covariáveis dependentes no tempo ocorre quando uma covariável muda no decorrer do tempo de estudo. Essas variáveis podem ser analisadas utilzando o modelo de riscos proporcionais de Cox para estimar se afetam o tempo de sobrevivência. Para trabalhar com esse modelo é necessário organizar o banco em forma de contagem. Em situações em que a suposição de riscos proporcionais do modelo de regressão de Cox não se sustenta, dizemos que o efeito da covariável varia com o tempo. O coeficiente de variação de tempo pode ser descrito com uma função de tempo paramétrica.

Estudos que consideram esse tipo de covariável podem fornecer melhores resultados e caso opte por não incluir esses valores pode gerar sérios vícios. Diante disso, estas covariáveis têm muita aplicação em análise de sobrevivência, já que podem ser úteis para modelar o efeito de indivíduos que mudam de grupo durante um estudo, quando para acomodar medidas que variam com o tempo.

Segundo Kalbfleisch e Prentice (2011), tais covariáveis podem ser consideradas dentro de duas amplas classificações referidas como covariáveis internas e covariáveis externas. Covariáveis internas são aquelas que são medidas durante o estudo e apenas podem ser medidas enquanto o objeto de estudo sobrevive.

Em contrapartida, covariáveis externas são aquelas que não necessariamente requer a sobrevivência do objeto de estudo para existir. Um tipo de variável externa é aquela que muda de tal forma que seus valores serão conhecidos caso se avance em um tempo futuro.

4.2 Ajuste do Modelo

Os diferentes tipo de covariáveis dependentes do tempo apresentador anteriormente podem ser incorporador ao modelo de Cox. Para formularização do modelo, primeiro deve-se considerar a fórmula geral do modelo de Cox

$$h(t|x) = h_0(t)exp(\beta' x), \tag{4.2.1}$$

em que $h_0(t)$ é a função de risco da linha base e $\beta' = (\beta_1, \beta_2, ..., \beta_p)$ é o vetor de coeficientes de regressão. Nessa forma proposta, $x = (x_1, x_2, ..., x_p)$, é um vetor de covariáveis fixas

no tempo.

Tendo em vista que o foco desse trabalho é utilizar covariáveis dependentes no tempo, segundo Therneau e Grambsch (2000), uma abordagem é estender o modelo de riscos proporcionais de Cox para permitir as covariáveis dependentes no tempo.

$$h(t|Z(t)) = h_0(t)exp(\theta' x + \gamma' X g(t)), \tag{4.2.2}$$

em que θ' são os coeficientes das covariáveis fixas e γ' são os coeficientes de covariáveis variáveis no tempo. Suponha que $Z(t)$ represente a covariável, então:

$$Z(t) = [x_1, x_2 ... x_{q_1}, X_1 g(t), X_2 g(t) ..., X_{q_2} g(t)] \tag{4.2.3}$$

Pode-se optar por generalizar o modelo como:

$$h(t|Z(t)) = h_0(t)exp(\beta' Z(t)), \tag{4.2.4}$$

em que $\beta = (\theta_1, ..., \theta_{q_1}, \gamma_1, ..., \gamma_{q_2})$. Aqui, $p = q_1 + q_2$ é o número de coeficientes do modelo.

É importante verificar que definindo dessa forma, o modelo dado não é mais de risco proporcional. Os valores das covariáveis $Z(t)$ dependem do tempo t. E a razão de risco no tempo t para dois indivíduos i e j é dada por:

$$\hat{HR} = \frac{h_i(t|Z(t))}{h_j(t|Z(t))} exp(Z_i(t)\beta - Z_j(t)\beta), \tag{4.2.5}$$

é também dependente do tempo.

Ao estender a função escore parcial para:

$$U(\beta) = \sum_{i=1}^{n} \delta_i [Z(t_i)\beta - log \sum_{j \in R(t_i)} exp(Z(t_j)\beta], \tag{4.2.6}$$

obtêm-se as estimativas dos parâmentros do modelo de Cox com covariáveis dependentes do tempo. Essa função é uma extensão da equação $U(\beta) = 0$, considerando covariáveis dependentes do tempo.

Segundo Andersen e Gill (1982), para construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo, são necessárias propriedades assintótica dos estimadores de máxima verossimilhança parcial. Indepêndente da dependência do tempo das covariáveis, a verossimilhança parcial preserva as mesmas propriedades do método clássico de máxima verossimilhança.

Vale ressaltar que a taxa de risco acumulada e a função de sobrevivência podem ser definidas e estimadas apenas se o processo for constante ($Z(t) = Z$).

4.3 Interpretação dos Coeficientes

Assim como no modelo de riscos proporcionais de Cox com covariáveis fixas, no modelo agora exposto, a interpretação dos coeficientes se mantém quase a mesma se diferenciando apenas pela inclusão do tempo.

Sendo assim, a interpretação dos coeficiente do modelo deve considerar o tempo t. Os coeficientes $\beta_l, l = 1, ..., p$ podem ser interpretados como o logaritmo da razão de risco para dois indivíduos cujo valor da l-ésima covariável no tempo t difere de uma unidade quando as outras covariáveis assumem o mesmo valor neste tempo.

4.4 Avaliação do Modelo

Ao contrário do exposto na Seção 3.3, no modelo de Cox com covariáveis dependentes no tempo, a suposição de riscos proporcionais é violada. Ou seja, para que esse modelo seja adequado, é necessário que uma covariável que foi medida no início do estudo não tenha efeito constante no resultado ao longo do acompanhamento.

Uma ferramenta de diagnóstico é considerada parte fundamental a fim de avaliar o ajuste do modelo e garantir que as suposições acerca do mesmo sejam plausíveis aos dados disponíiveis. Os resíduos de Cox-Snell são úteis para verificar o ajuste global de um modelo final. Os resíduos de Cox-Snell são um tipo de resíduos padronizados usados em análises de sobrevivência. Um resíduo é a diferença entre um ponto de dados observado e um valor predito. Um resíduo de Cox-Snell considera os parâmetros da distribuição e estimados a partir do modelo de regressão de vida útil.

Os resíduos de Cox-Snell são iguais ao negativo do logaritmo natural da probabilidade de sobrevivência para cada observação e a partir do gráfico dos resíduos é possível obter um diagnóstico visual que permite identificar observações extremas que precisam de investigação adicional e avaliar se a distribuição do tempo de falha ou do log do tempo de falha é adequada.

5 Risco de Crédito

O mercado financeiro move o país e essa grande moviventação atual teve início com a Revolução industrial, onde o crédito surgiu e possibilitou a fomentação da economia do país, se tornando então, uma das operações mais importantes.

O crédito nada mais é, que um contrato entre um consumidor e uma instuição financeira. Contrato esse em que a instuição financeira oferece ao consumidor um valor que deve ser devolvido em um determinado período acrescido de juros.

Diante disso, temos que toda operação de crédito está diretamente relacionada a "confiança" apresentada entre o consumidor e a instituição financeira. Confiança essa que tem como principal foco analisar o tomador, bem como seu histórico, presente e possíveis tendências futuras, para assim reduzir dos riscos para a empresa.

Essa análise geral deve ser bem feita, para evitar o endividamento do cliente ou grandes prejuízos a empresa. Ademais, essa análise deve ser padronizada, evitando assim que para um mesmo tomador, várias respostas sejam apresentadas. Sendo assim, para uma boa análise e previsão de risco de crédito, utiliza-se modelos de *credit scoring*, esses modelos garantem consistência nas decisões, automatização na concessão e melhores análises.

Os modelos de *credit scoring* se dividem em dois grupos. O primeiro, denomidado *application scoring*, ferramenta utilizada para captação da proposta e análise do risco de crédito com foco no perfil de risco do proponente baseada em aspectos sociais, financeiros e demográficos para decisão de aceitação e definição de limites e condições, ou seja, verifica a probabilidade do proponente não pagar seu compromisso antes de completar um período prefixado, esse modelo tem como objetivo principal conceder crédito para novos clientes. Já o segundo grupo, *behavior scoring*, tem como finalidade administarar os créditos de clientes já ativos. Tomaremos o segundo grupo como base para o desenvolvimento do trabalho em questão.

5.1 Behavior Score

O *Behavior Score* é aplicado para a manutenção de crédito, ou seja, para clientes que já tem um contrato com a instuição, onde são utilizados modelos para assessorar na concessão de crédito e na análise de risco.

A modelagem de *behavior score* auxilia principalmente na recalibragem de limites concedidos anteriormente ao tomador, garantindo assim uma melhor classificação e tornando o sistema financeiro de crédito mais eficiente. Além disso, contribui para ren-

tabilização, manutenção e retenção uma vez que para clientes bons há a possibilidade de ofertar novos produtos.

Sendo assim, tomando como base o ponto escolhido, primeiro analisa-se o intervalo anterior a esse ponto, denominado período histórico. Essa análise serve como um "rascunho" e como suporte para a classificação do tomador. Já o intervalo posterior ao ponto é chamado de período de performance. Conforme Sicsú (2010), sendo esse o período em que é possível verificar o atual risco do tomador, dependendo do critério definido no começo do estudo.

Esse modelo de *behavior score* pode ser formulado por diversas técnicas, entre elas encontra-se a análise de sobrevivência que além de apresentar uma resposta temporal, também é capaz de considerar dados censurados, o que garante menor perda de informação.

5.2 Escore de Risco

Tendo em vista o modelo de *behavior score*, um modelo de pontuação de risco será obtido usando os resultados de uma análise de sobrevivência, mais especificamente, um modelo de regressão de Cox com covariáveis dependentes no tempo.

Conforme definido no Capítulo 3 deste trabalho, o modelo de riscos proporcionais de Cox considera as covariáveis de interesse para modelar dados de sobrevivência, por meio da função de risco. Considerando a função de ligação logarítmica, tem-se que quanto maior o preditor linear, maior o risco do cliente atrasar o seu pagamento. Desta forma, levando em consideração as informações apresentadas e tendo como foco principal as variações que algumas covariáveis podem manifestar com o decorrer do tempo, o presente trabalho propõe um escore de risco para o cliente i, $i = 1, 2, ..., n$, como sendo:

$$ER_i = exp\left\{ \sum_{q=1}^{p} \sum_{j=1}^{J_i} \frac{t_{(i,j)} - t_{(i,j-1)}}{max_j\{t_{(i,j)}\}} \beta_q Z_q(t_{ij}) \right\} \tag{5.2.1}$$

em que $t_{(i,j)}$ é o maior tempo do i-ésimo cliente quando a sua q-ésima covariável $Z_q(t_i)$ assume o seu j-ésimo valor, $j = 1, 2, ..., J_i$, com $t_{(i,0)} = 0$ e β_q é o q-ésimo coeficiente associado à covariável $Z_q(t_i)$, $q = 1, 2, ..., p$.

Em, (5.2.1), $\frac{t_{(i,j)} - t_{(i,j-1)}}{max_j\{t_{(i,j)}\}}$ mede a contribuição da covariável $Z_q(t_i)$ em cada intervalo de tempo. Note que a covariável $Z_q(t_i)$ pode ou não depender do tempo t_i. De fato, quando $Z_q(t_i)$ não depende do tempo, tem-se que $J_i = 1$, resultando em $\frac{t_{(i,j)} - t_{(i,j-1)}}{max_j\{t_{(i,j)}\}} = 1$.

O escore ER definido em (5.2.1) assume valores positivos e, quanto maior o seu valor, maior o risco de default do cliente (risco do mesmo atrasar o pagamento).

5.3 Métricas de Performace e Classificação dos clientes por meio do escore de risco

A validação do modelo desenvolvido é uma das partes mais importantes do processo, é a etapa onde se pode comprovar se o modelo irá gerar bons resultados para o público em questão. Segundo Congalton (1991), um dos métodos usados para essa validação é chamado de matriz de confusão.

Esta matriz permite determinar quantos acertos e erros o modelo obteve para a classificação. Em geral, para criação de um modelo, utiliza-se uma base de dados dividida em duas partes. A primeira, denominada "treino", corresponde a 70% dos dados e tem como objetivo utilizar a classificação como informação para o modelo. Já a segunda parte, denominada "teste", é manipulado o restante dos dados, 30%, para que o próprio algoritmo realize a classificação.

A matriz de confusão é uma ferramenta padrão para avaliar modelos. Ela contém todas as respostas do modelo e é utilizada durante a fase de treinamento, sendo possível comparar com os resultados reais. Portanto, a matriz é dividida em quatro quadrantes: verdadeiros positivos; falsos negativos; falsos positivos; verdadeiros negativos.

		Predito	
		Alto Risco	**Baixo Risco**
Real	**Alto Risco**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Baixo Risco	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 1: Matriz de Confusão

A matriz de confusão se divide em dois grupos: Real e Previsto. Sendo o valor previsto o resultado do modelo, esse valor é comparado na tabela com o real.

Caso o modelo erre na classificação, a matriz apresenta dois tipos de erro, o tipo 1 que é o Falso Positivo e o tipo 2 que é o Falso Negativo. No contexto apresentado nesse trabalho, a interpretação seria:

- Erro 1: Um cliente classificado como alto risco ser, na realidade, de baixo riso. Ou seja, a instituição deixou de conceder crédito para um bom tomador.

- Erro 2: Um cliente classificado como baixo risco ser, na realidade, de alto risco. Ou seja, a instituição concede crédito para um mau pagador.

Ademais, existem outras métricas que podem medir a performance do modelo de risco, como: Total de acertos, Precisão, Sensibilidade e Especificidade. Essas métricas são especificados abaixo.

- **Total de Acertos**: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente.

$$\text{Total de Acertos } = \frac{VP + VN}{VP + VN + FP + FN} \tag{5.3.1}$$

- **Precisão**: dentre todas as classificações de alto risco que o modelo fez, quantas estão corretas.

$$\text{Precisão } = \frac{VP}{VP + FP} \tag{5.3.2}$$

- **Sensibilidade**: dentre todas as situações de alto risco na amostra, quantas estão corretas.

$$\text{Sensibilidade } = \frac{VP}{VP + FN} \tag{5.3.3}$$

- **Especificidade**: dentre todas as situações de baixo risco na amostra, quantas estão corretas.

$$\text{Especificidade } = \frac{VN}{VN + FP} \tag{5.3.4}$$

Para cada estudo, existem situções diferentes em que deve-se decidir minimizar um dos erros ou os dois. No contexto deste trabalho, optou-se por minimizar o erro 2, pois acredita-se que no cenário de uma grande instituição financeira o erro mais grave é o de conceder crédito para um mau pagador, correndo assim, um maior risco de inadiplência. Portanto, o erro 2 deve ser controlado utilizando o cenário:

$$\text{P(Erro 2)} = \frac{FN}{FN + VN} \tag{5.3.5}$$

Diante disso, a classificação de um cliente pode ser dada por:

- $ER > k$: o cliente é considerado de alto risco;
- $ER \leq k$: o cliente é considerado de baixo risco.

Em que o valor de k, é definido de forma a controlar o erro 2, isto é, um valor tal que:

$$\mathrm{P}(\text{Erro 2}) < \alpha, \tag{5.3.6}$$

sendo α a máxima probabilidade do erro 2 admitida pela instituição.

6 Ilustração em um conjunto de dados simulados

Diante de um cenário de uma grande instituição financeira, ao realizar o cadastro de um novo cliente, informações são coletadas e armazenadas em um banco de dados para que possam ser analisadas.

Essa análise, no momento do cadastro (*application scoring*), garante que o servidor possa tomar a decisão de liberação de cartões, limites, empréstimos entre outros serviços. Para que essa decisão seja tomada, a instituição possui modelos para calcular uma pontuação, também chamado de escore. Esse escore os ajuda a tomar a decisão se esse cliente é ou não é um bom credor, garantindo assim créditos para novos clientes.

Entretanto, para os clientes que já estão vinculados à instituição, as informações cadastrais são atualizadas periodicamente, assim como o escore é reprocessado (*behavior scoring*), mas esse processo de atualização do escore não leva em consideração todo o histórico anterior ao momento em que é processado o modelo.

Uma solução para a manutenção de crédito de clientes já ativos é a utilização de um escore de risco via modelo de regressão de Cox com covariáveis dependentes no tempo. Posto isto, o estudo seria capaz de observar um histórico dos tomadores, bem como suas mudanças cadastrais no decorrer do tempo e gerar um modelo mais adequado com a realidade.

6.1 Dados

Para a aplicação da metodologia proposta, foi utilizado um banco de dados simulados com um histórico de 10 mil tomadores no período de 60 meses. O pacote do *software R* utilizado para gerar os dados foi o "*sim.survdata*", onde apenas a covariável renda foi criada por meio de um *looping* para que ela se alterasse até 3 vezes por indivíduo, além disso, o percentual de censura definido foi de 4%. As variáveis utilizadas no estudo foram:

- **id** - Identificação do cliente;
- **sexo** - Sexo do cliente (variável fixa). 1: Masculino e 0: Feminino;
- **escolaridade** - Escolaridade do cliente (variável fixa). 1: "Não Possui Ensino Superior" e 0:"Possui Ensino Superior";
- **idade** - Idade do cliente (variável fixa medida em anos);
- **log_renda** - Log da renda do cliente (variável dependente no tempo).
- **tstart** - Tempo inicial da análise, medida em mês.

- **tstop** - Tempo final da análise, medida em mês.
- **endpt** - Determina se o tomador apresentou ou não a falha (inadimplência) em cada corte de tempo. 1: "Inadimplente" e 0: "Não Inadimplente".

Para a implementação do modelo em estudo no *software* ***R***, o banco de dados precisa estar estruturado diferente do que é de costume. Em geral, para cada linha do banco, tem-se a informação de um cliente. Porém, para o modelo com covariáveis dependentes no tempo, um mesmo cliente pode aparecer em mais de uma linha.

Essa mudança se dá, pois para cada mudança na covariável que depende no tempo, uma nova linha é criada contendo as informações do tomador durante o período em que observou-se essa mudança. A Figura 2 apresenta a visualização das 20 primeiras linhas do banco de dados em estudo.

	id	sexo	esc_nao_superior	idade	tstart	tstop	endpt	renda	log_renda
1	1	1	1	34	0	4	0	6607.000	8.795885
2	1	1	1	34	4	6	0	5980.000	8.696176
3	1	1	1	34	6	7	0	6866.000	8.834337
4	1	1	1	34	7	31	1	6607.000	8.795885
5	2	1	1	65	0	6	0	5931.000	8.687948
6	2	1	1	65	6	10	0	7108.000	8.868976
7	2	1	1	65	10	11	0	5199.000	8.556222
8	2	1	1	65	11	18	1	5980.000	8.696176
9	3	1	1	47	0	4	0	5291.000	8.573763
10	3	1	1	47	4	5	0	6625.000	8.798606
11	3	1	1	47	5	20	0	7573.400	8.932397
12	4	1	1	68	0	5	0	5639.000	8.637462
13	4	1	1	68	5	11	0	5193.000	8.555067
14	4	1	1	68	11	13	0	5052.135	8.527566
15	5	1	0	47	0	4	0	5680.000	8.644707
16	5	1	0	47	4	7	0	5502.000	8.612867
17	5	1	0	47	7	8	0	8670.000	9.067624
18	5	1	0	47	8	13	0	6068.757	8.710909
19	6	1	1	39	0	4	0	4365.000	8.381373
20	6	1	1	39	4	10	0	6030.000	8.704502

Figura 2: Visualização das 20 primeiras linhas do banco de dados

Note pela Figura 2 que cada mudança da covariável de um cliente é representada por uma nova linha do banco de dados. O cliente "id=1" por exemplo, apresentou 4 valores distintos da variável renda ao longo do estudo e por isso está representado em 4 linhas diferentes. Já o cliente "id=4" só apresentou 3 valores distintos para variável renda, logo, ele aparece em apenas 3 linhas, mantendo os valores das variáveis que não foram alteradas.

A base de dados simulada foi dividida em duas partes (treino e teste), em seguida, uma breve análise descritiva foi feita para cada variável na base de treino e seus resultados

serão disponibilizados em seguida.

6.2 Análise Descritiva

De acordo com as Tabelas 1 e 2, pode-se tirar algumas conclusões básicas a respeito do banco de dados em análise (banco treino). É possível verificar que 60,00% dos clientes em estudo são do sexo masculino e 74,64% não apresenta nível superior, informações que podem ser essenciais para identificar o perfil de um possível tomador inadimplente.

Além disso, para o estudo, apenas clientes maiores de idade (acima de 18 anos) foram analisados. Na Figura 3, observa-se que tanto o sexo quando a inadimplência apresentam idade média muito próximas e algumas idades que se diferenciam drasticamente de todas as outras. Na Figura 3 é possível notar também, que a média das rendas assemelham-se para ambos os sexos e para os dois tipos de clientes, aqueles que apresentaram inadimplência e aqueles que não apresentaram.

É importante ressaltar que durante o período de 60 meses de estudo, grande parte dos clientes apresentaram atualizações cadastrais envolvendo a renda, essa mudança pode ser dada devido à uma troca emprego, uma promoção ou até mesmo à uma demissão.

Ademais, no banco de treino foram observados 2.796 (39,94%) clientes que falharam, ou seja, se tornaram inadimplente em algum momento do estudo.

Tabela 1: Tabela de frequência da variável sexo (banco treino)

Sexo	Frequência Absoluta	Frequência Relativa	Frequência Relativa %
0	2.801	0,4000	340,00%
1	4.199	0,6000	60,00%

Tabela 2: Tabela de frequência da variável escolaridade (banco treino)

Escolaridade Não Superior	Frequência Absoluta	Frequência Relativa	Frequência Relativa %
0	1.775	0,2536	25,36%
1	5.225	0,7464	74,64%

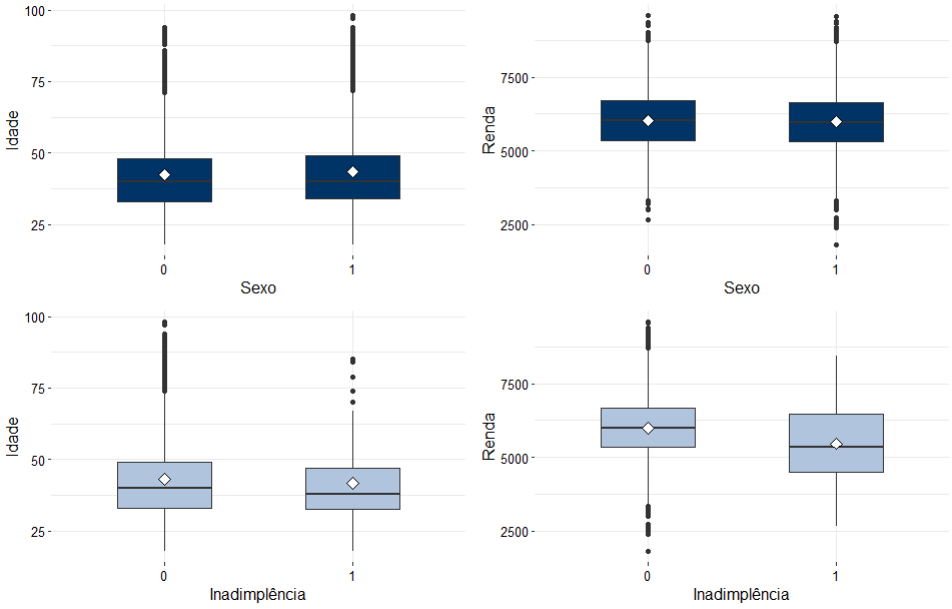

Figura 3: Boxplot das covariáveis Idade e Renda por Sexo e Inadimplência

Posteriormente, é realizada a análise descritiva dos dados, como os dados de sobrevivência possuem observações parciais da resposta, as censuras, será utilizado técnicas não paramétricas para descrever o banco. Na Figura 4 foi realizada a estimação da função de sobrevivência pelo método de Kaplan-Meier. Ao observar o gráfico da função de sobrevivência de todos os tempos, sem considerar variáveis explicativas é possível verificar o comportamento decrescente, que já era esperado, onde a queda de probabilidade diminui de forma gradativa até tempo final de 60 meses, se aproximando de zero. Entretanto, como há presença de censuras nos últimos valores, o gráfico apresenta uma pequena elevação no final, o que faz com que a função de sobrevivência não decaia totalmente para o valor zero. Após a análise dos gráficos para a covariável sexo (Figura 5) e escolaridade (Figura 6), com o objetivo de verificar se realmente é possível identificar diferenças entre as curvas de sobrevivência para algumas covariáveis, constatou-se que a suposição de risco proporcionais não é razoável, pois as curvas de sobrevivência se cruzam em todos os casos.

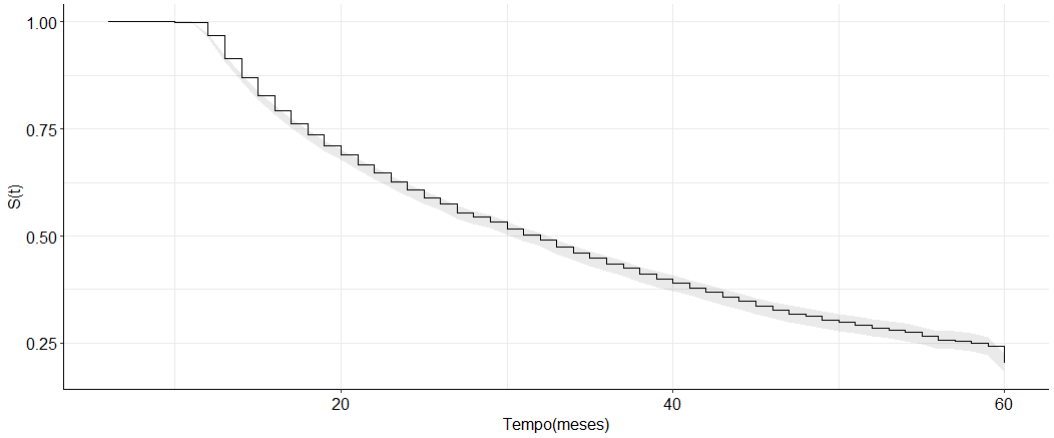

Figura 4: Gráfico da função de sobrevivência - Kaplan-Meier

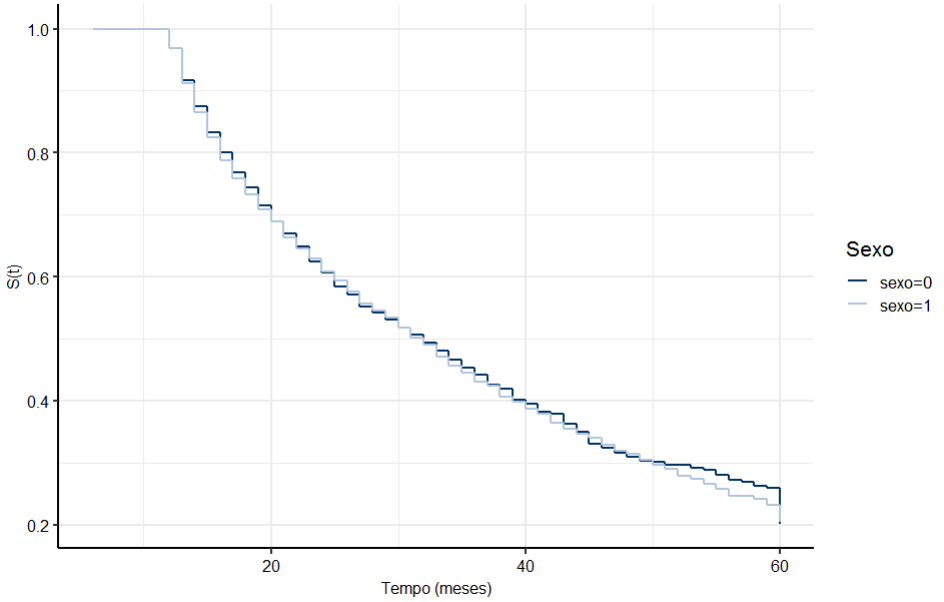

Figura 5: Gráfico da função de sobrevivência da variável sexo

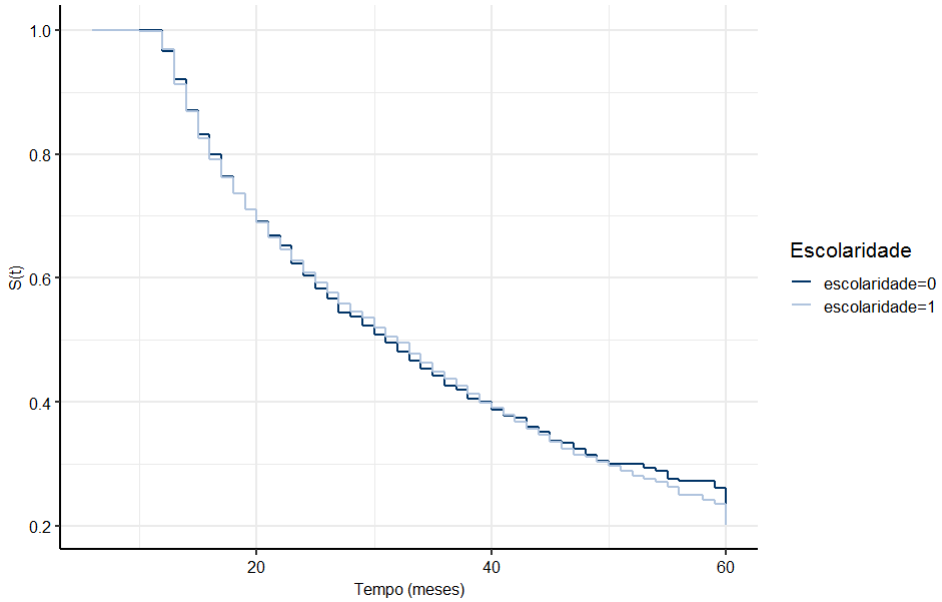

Figura 6: Gráfico da função de sobrevivência da variável escolaridade

6.3 Modelo

O modelo de Cox com covariáveis dependentes no tempo gerado a partir de todas as variáveis observadas no banco de dados de treino resultou na tabela a seguir:

Tabela 3: Estimativa dos parâmetros do modelo de Cox com covariáveis dependentes no tempo (banco treino)

Variável	Coeficiente	Exp(coef)	P-valor
escolaridade	-0,0009947	0,9990058	0,982
sexo	-0,0102163	0,9898357	0,793
idade	-0,0003051	0,9996949	0,821
log_renda	-0,8918520	0,4098959	<2e-16

Nota: Ensino Superior e Feminino são os níveis de referência das covariáveis escolaridade e sexo, respectivamente.

Em situações em que há um grande número de covariáveis, um método de seleção de covariáveis pode ser utilizado (como por exemplo do tipo Stepwise). No entando, visto que o objetivo neste trabalho é ilustrar a metodologia proposta, optou-se por não seguir esses passos e utilizar todas as covariáveis, por mais que não sejam significativas, para dar continuidade a construção do escore de risco.

Como o modelo em questão não necessita da validação de riscos proporcionais. Pode-se analisá-lo, a partir da Tabela 3 que sugere a seguinte conclusão: A taxa de falha,

ou seja, a chance de um tomador se tornar inadimplente é aproximadamente 0,41 vezes para cada unidade do log da renda a mais, em um determinado tempo, mantida fixas as outras covariáveis. Para as demais covariáveis, como elas não apresentaram significância, optou-se por não realizar a interpretação dos coeficientes.

6.4 Escore de Risco

Conforme estabelecido na Seção 5.2, utilizou-se o resultado do modelo de Cox com covariáveis dependentes no tempo, encontrado anteriormente, para criação de um modelo de pontuação de risco. Para validar o modelo desenvolvido e determinar um bom ponto de corte para classificação dos tomadores em bom ou mau pagador, optou-se por observar o comportamento dos erros do tipo 1, tipo 2 e também a acurácia. Sendo que, no cenário do trabalho em questão, a melhor opção é minimizar o erro 2. Tendo em vista que, para uma grande instituição financeira, o erro mais grave seria o de conceder crédito para um mau pagador.

A Figura 7 apresenta os valores do Total de Acertos do modelo e Erros 1 e 2, de acordo com o ponto de corte, k, adotado.

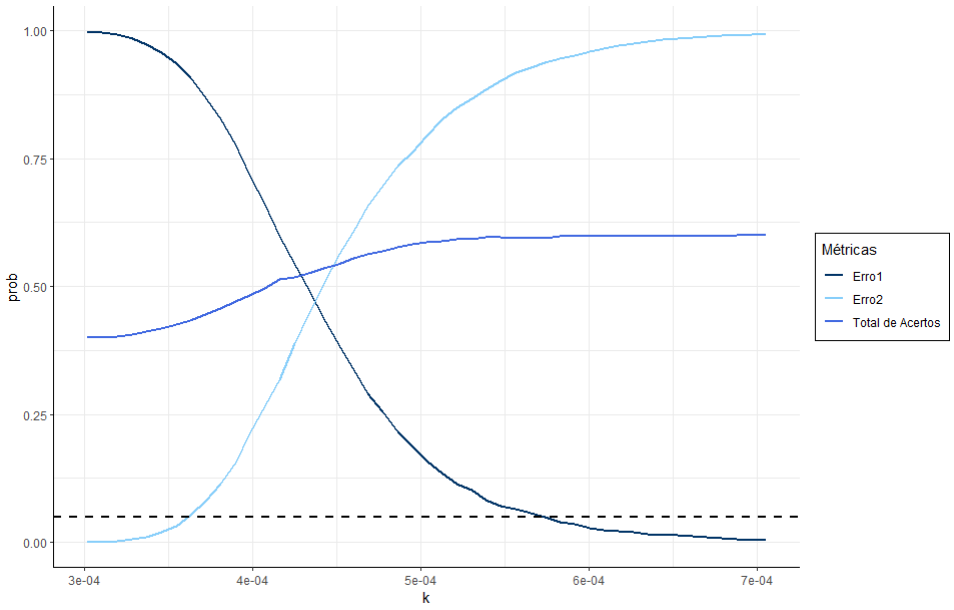

Figura 7: Gráfico de comportamento do erro 1, erro 2 e acurácia (banco treino)

Assumindo um $\alpha = 0,05$, o valor de k observado do corte da reta pontilhada com a curva do erro2, foi de $0,0003545296$. Esse valor foi definido de forma a controlar o erro 2, sendo $0,05$ a probabilidade máxima de se conceder crédito para um mau tomador. Assim,

se $ER < 0,0003545296$ o cliente é classificado como baixo risco e se $ER \leq 0,0003545296$, o cliente é classificado como alto risco. A Figura 8 apresenta a classificação dos 10 primeiros clientes do banco treino segundo esse ponto de corte definido.

	id	falha	tempo_final	ER	estimador	classificacao_estimada
1	1	1	31	0.0003852640	1	1
2	2	1	18	0.0004050044	1	1
3	3	0	20	0.0003626790	1	1
4	4	0	13	0.0004590723	1	1
5	5	0	13	0.0004182106	1	1
6	6	0	17	0.0004039621	1	1
7	7	1	14	0.0004841759	1	1
8	8	0	22	0.0005327638	1	1
9	9	0	23	0.0004412377	1	1
10	10	0	12	0.0004199856	1	1
11	11	0	12	0.0003330644	0	0

Figura 8: Classificação dos 10 primeiro cliente do banco treino segundo o corte k definido.

De forma ilustrativa, ao selecionar o cliente de $id = 1$, tem-se as seguintes informações:

id	sexo	esc	idade	tstart	tstop	endpt	log_renda
1	1	1	34	0	4	0	8,795885
1	1	1	34	4	6	0	8,696176
1	1	1	34	6	7	0	8,834337
1	1	1	34	7	31	1	8,795885

Para calcular o escore de risco desse cliente, deve-se calcular para cada intervalo observado e em seguida somar a contribuição de cada intervalo e aplicar o exponencial sobre esse valor, chegando assim no ER_1.

Para o primeiro intervalo, temos:

$$\frac{4-0}{31}(-0,0009947302\times 1+-0,01021629\times 1+-0,0003051258\times 34+-0,891852\times 8,795885)$$

$$= 0,12903226 \times -7,8662123 = -1,0149952, \tag{6.4.1}$$

em que 4 é o maior tempo do cliente $id = 1$ no primeiro intervalo, 0 é o maior tempo do cliente 1 no intervalo anterior (como não existe intervalo anterior ao primeiro, defini-se o maior tempo como 0) e 31 é o máximo de tempo observado pelo cliente 1 durante todos

os intervalos. Os valores apresentados dentro dos parênteses são os coeficientes estimados para cada covariável multiplicado pelo valor que essa covariável assume no intervalo.

Seguindo os mesmos passos para o segundo, terceiro e quarto intervalo, obtêm-se:

Para o 2^{o} intervalo:

$$= 0,06451613 \times -7,777287 = -0,5017604. \tag{6.4.2}$$

Para o 3^{o} intervalo:

$$= 0,03225806 \times -7,900506 = -0,2548550. \tag{6.4.3}$$

Para o 4^{o} intervalo:

$$= 0,77419355 \times -7,866213 = -6,0899711. \tag{6.4.4}$$

Logo, a soma da contribuição de cada intervalo é:

$$-1,0149952 - 0,5017604 - 0,2548550 - 6,0899711 = -7,861582. \tag{6.4.5}$$

Aplicando o exponencial ao valor obtido anteriormente:

$$ER_1 = 0,000385264. \tag{6.4.6}$$

Conforme o critério de classificação estabelecido na Seção 5.3, como $ER_1 > k$, em que $k = 0,0003545296$, o cliente $id = 1$ é classificado como alto risco.

Tendo esse exemplo como referência, após todos os cliente passarem pelo modelo apresentado, eles receberão um escore de risco que os classificarão como alto ou baixo risco. Sendo assim, a instituição financeira será capaz de minimizar as perdas com clientes que geram prejuízos ao se tornarem inadimplentes.

Para verificar a performance do modelo classificar corretamente um novo cliente, foi implementado o modelo de Cox e posteriormente o escore de risco na base de teste. Após cada cliente receber uma pontuação, o valor de corte $k = 0,0003545296$, dividiu os tomadores entre alto e baixo risco e a proporção de acertos foi apresentada na Figura 9. Vale ressaltar que, uma desvantagem apresentada quando se define apenas um erro para controlar é que como o objetivo principal da classificação foi de minimizar o erro de conceder crédito para maus pagadores, a quantidade de clientes classificados como alto

risco, que são na realidade de baixo risco aumentou.

		Predito	
		Baixo Risco	Alto Risco
Real	Baixo Risco	90	1.735
	Alto Risco	59	1.116

Figura 9: Tabela da proporção de acertos e erros do modelo na amostra de teste.

A partir da Figura 6 é possível verificar que apenas 59 tomadores (5,02%) foram classificados com baixo risco, quando na verdade eram de alto risco (Erro 2 = 5, 02%), ou seja, uma proporção muito baixa dentre os 3.000 clientes que passaram pela classificação de teste. Em contra partida, o número de pagadores classificados erroneamente como alto risco foi alto (95%), indicando que o modelo apresentado é mais criterioso e tem uma margem de erro muito maior para deixar de conceder crédito para um bom tomador.

7 Conclusões Finais

O presente trabalho teve por objetivo estudar o modelo de Cox com covariáveis dependentes no tempo, aplicado ao mercado financeiro, em que a variável resposta é o tempo até um tomador se tornar inadimplente. O foco do trabalho foi apresentar uma proposta de escore de risco com base nesse modelo, para garantir uma pontuação para cada cliente que já possua um histórico na instituição, sendo aplicado ao behavior score.

Diante disso, a partir dessa pontuação foi gerado um ponto de corte capaz de classificar os clientes em bons e maus pagadores, garantindo uma minimização do erro de conceder crédito para maus pagadores e diminuindo então a probabilidade de possíveis prejuízos gerados por um grande número de tomadores inadimplentes.

Os resultados obtidos mostraram que o escore de risco proposto é útil para a ordenação e classificação de clientes em situação onde as suas características (covariáveis) se modificam ao longo do tempo. A metodologia se mostrou eficaz em controlar o erro desejado (em geral, o mais grave do ponto de vista da instituição financeira) no entanto, a sua acurácia depende da qualidade do modelo de regressão ajustado.

Vale ressaltar que, por motivos de dificuldade de obtenção e liberação pela instituição financeira, a metodologia proposta neste trabalho foi ilutrada por meio de dados simulados. Diante disso, propõe-se para estudos futuros a aplicação da metodologia em banco de dados reais.

Referências

Andersen, P. K.; Gill, R. D. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, JSTOR, p. 1100–1120, 1982.

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.

Colosimo, E.; Giolo, S. *Análise de Sobrevivência Aplicada*. [S.l.]: Edgar Blücher, São Paulo, 2006. 1-87 p. ABE - Projeto Fisher.

Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, Elsevier, v. 37, n. 1, p. 35–46, 1991.

Cox, D. Regression models and life tables. *Journal of Royal Statistical Society.*, v. 39, p. 1–38, 1972.

Cox, D.; Snell, E. A general definition of residuals. *Journal of the Royal Statistical Society*, v. 30, n. 2, p. 248–275, 1968.

Giolo, S. *Modelos de analise de sobrevivencia para experimentos dose-resposta*. Dissertação (Mestrado) — Universidade Estadual de Campinas, Instituto de Matematica, Estatistica e Ciencia da Computação, Campinas,SP, 1994.

Hoel, P. G.; Port, S.; Stone, C. J. Introdução a teoria da probabilidade. Editora lnterciencia Ltda., Rio de Janeiro, 1978.

Kalbfleisch, J.; Prentice, R. *The Statistical Analysis of Failure Time Data*. Wiley, 2011. (Wiley Series in Probability and Statistics). ISBN 9781118031230. Disponível em: ⟨https://books.google.com.br/books?id=BR4Kq-a1MIMC⟩.

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

Nakano, E.; Carrasco, C. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. tema - tend. mat. apl. comput. v. 7, n. 1, p. 91–100, 2006.

Sicsú, A. *Credit scoring: desenvolvimento, implantação,acompanhamento.* [S.l.]: Blücher, São Paulo, 2010.

Therneau, T.; Grambsch, P. The cox model. In: *Modeling survival data: extending the Cox model*. [S.l.]: Springer, 2000. p. 39–77.