Pontifícia Universidade Católica
do Rio de Janeiro

Carolina Marques Portilho

Estimação da Persistência de Segurados de Planos de Previdência Privada Via Modelos de Sobrevivência

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós–graduação em Engenharia Elétrica da PUC–Rio

Orientador: Prof. Cristiano Augusto Coelho Fernandes

Rio de Janeiro
Abril de 2013

Pontifícia Universidade Católica
do Rio de Janeiro

Carolina Marques Portilho

Estimação da Persistência de Segurados de Planos de Previdência Privada Via Modelos de Sobrevivência

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Cristiano Augusto Coelho Fernandes
Orientador
Departamento de Engenharia Elétrica — PUC–Rio

Prof. Eduardo Fraga L. de Melo
Superintendência de Seguros Privados-Ministério da Fazenda

Prof. Kaizô Iwakami Beltrão
FGV

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 05 de Abril de 2013

Carolina Marques Portilho

Graduou-se em Estatística na ENCE (Escola Nacional de Ciências Estatísticas) em 2010 com a monografia titulada "Modelos Econométricos de Previsão da Inflação".

Ficha Catalográfica

Portilho, Carolina

Estimação da Persistência de Segurados de Planos de Previdência Privada Via Modelos de Sobrevivência / Carolina Marques Portilho; orientador: Cristiano Augusto Coelho Fernandes. – 2013.

61 f.: il. ; 30 cm

1. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2013.

Inclui referências bibliográficas.

1. Engenharia Elétrica – Tese. 2. Previdência privada. 3. Análise de sobrevivência. 4. Estimador de Kaplan-Meier. 5. Modelo de Cox. 6. Função de sobrevivência. 7. Função de risco. 8. PGBL. 9. VGBL. I. Fernandes, Cristiano. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Agradecimentos

Agradeço a minha família ao apoio em toda a minha vida. Aos meus pais, Deuza e Wagner, e a minha irmã, Fernanda, obrigada.

Agradeço ao meu companheiro André por toda a dedicação, amor, carinho e compreensão que sempre teve comigo, especialmente nos momentos delicados que precisei passar para a conclusão do mestrado.

Agradeço ao meu orientador Cristiano pela dedicação na minha orientação. Agradeço também ao professor Álvaro pela oportunidade de ensinamento.

Agradeço aos amigos de mestrado, de graduação e a todos os amigos da vida.

Agradeço à CAPES pela concessão da bolsa de estudos.

Resumo

Portilho, Carolina Marques; Fernandes, Cristiano Augusto Coelho (Orientador). **Estimação da Persistência de Segurados de Planos de Previdência Privada Via Modelos de Sobrevivência**. Rio de Janeiro, 2013. 61p. Dissertação de Mestrado — Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação tem como objetivo propor uma abordagem pouco utilizada na área financeira para prever cancelamentos em planos de previdência privada. Métodos e modelos provenientes da análise de sobrevivência foram utilizados para prever o cancelamento, além de estimar o risco associado de o cliente cancelar. Os modelos propostos, modelo de regressão paramétrico e modelo de riscos proporcionais de Cox, foram estimados utilizando-se uma base de dados de clientes acompanhados durante um período de 6 anos e meio de uma seguradora nacional. Os modelos mostraram-se equivalentes, tendo capacidade de generalização de 58%.

Palavras–chave

Previdência privada; Análise de sobrevivência; Estimador de Kaplan-Meier; Modelo de Cox; Função de sobrevivência; Função de risco; PGBL; VGBL;

Abstract

Portilho, Carolina Marques; Fernandes, Cristiano Augusto Coelho (Advisor). **Estimation of insured persistency pension plans via survival models**. Rio de Janeiro, 2013. 61p. MSc Dissertation — Department of Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This paper aims to propose an approach rarely used in finance to predict cancellations in private pension plans. Methods and models from survival analysis were used to predict the cancellation and estimate the risk associated with the client cancellation. The proposed models, regression model and parametric proportional hazards model of Cox, were estimated using a database of customers followed for a period of 6 and a half years of a national insurer. The models showed equivalence and generalizability of 58%.

Keywords

Pension plans; Survival analysis; Kaplan-Meier estimator; Cox model; Survival function; Hazard function; PGBL; VGBL;

Sumário

1
Introdução

Define-se por previdência privada uma forma de complemento de aposentadoria fundamentado em um acúmulo de reservas pelo indivíduo em um determinado período de tempo, como uma poupança a longo prazo. Para tal, é necessário que haja disciplina e comprometimento por parte do segurado para que o investimento corresponda às expectativas.

Segundo a Federação Nacional de Previdência Privada e Vida (1), o mercado de previdência privada brasileiro cresceu 920% entre os anos 2000 e 2010. Impulsionados por esse crescimento, seguradoras e instituições financeiras preocupam-se com uma importante questão: a persistência em planos de previdência privada. Neste contexto, persistência refere-se ao ato de perseverar nos pagamentos dos prêmios enquanto dura o contrato. Ou seja, são persistentes aqueles investidores que pagam regularmente ou não resgatam o prêmio de uma única vez antes do previsto.

Como apontado por Lian (2), o cancelamento dos contratos de forma prematura resulta em perdas e desperdício de esforços para todos os participantes do negócio: o segurado, o corretor e a companhia de seguros. O segurado perde os prêmios pagos e a proteção do seguro; os gastos com a aquisição do plano podem não ser recuperados pela companhia; e o corretor perde suas comissões por renovação de contrato e pode até comprometer seu emprego.

Ao incorporar um novo cliente em um plano de previdência, a seguradora precisa saber previamente quais são os riscos envolvidos com essa ação. O novo cliente permanecerá com o plano por um período mínimo de forma que se torne rentável à seguradora? O novo cliente é um potencial investidor? A previdência privada é o melhor produto que se pode ser vendido ao cliente com um dado perfil? Para responder tais perguntas, uma das principais ferramentas de controle de risco utilizada pelas empresas é a modelagem estatística.

Quando se trata da longevidade do cliente, é necessário que se analise os cancelamentos dos contratos e o comportamento do cliente. Com isso, a empresa saberá se o risco de perder um cliente é alto ou baixo e quais são as suas causas. Tendo conhecimento do perfil do cliente em relação à longevidade, pode-se tratar melhor questões como o prêmio pago e também criar margens

a fim de separar os clientes mais rentáveis dos outros.

Nessa dissertação, a metodologia proposta para a estimação da persistência é a análise de sobrevivência, técnica amplamente utilizada na área biomédica e pouco explorada nas demais. Na economia, a análise de sobrevivência também é conhecida como análise de duração (*duration analysis*); na engenharia, análise de confiança (*reliability analysis*); em sociologia, análise da história de um evento (*event history analysis*).

A unidade básica na análise de sobrevivência é a variável resposta, que é o tempo até a ocorrência de um evento de interesse. Na medicina, o evento de interesse costuma ser a morte de um paciente ou até mesmo a cura de uma doença; na indústria, a falha de um produto sob teste; em economia, desemprego, promoções ou aposentadoria. Nessa dissertação, o evento de interesse é o cancelamento do plano de previdência pelo cliente. Desta forma, a variável resposta será dada pelo tempo decorrido entre a adesão do plano de previdência pelo cliente até o seu cancelamento prematuro.

Em um estudo como esse, onde o interesse é fazer previsões, a amostra utilizada geralmente é coletada em um período finito pré-determinado, não sendo possível acompanhar todos os clientes até o encerramento de seus planos. Surge assim uma das principais características em dados de sobrevivência: a *censura*. Um dado é dito censurado se até o final do período de observação não for observado o evento de interesse, nesse caso o cancelamento do plano de previdência. Outra característica comum em análise de sobrevivência é a presença de dados truncados. Esses conceitos, tanto a censura como o truncamento dos dados, serão cuidadosamente discutidos no capítulo 2, seção 2.1 e relacionados com os dados em estudo.

Os dados utilizados nessa dissertação são provenientes de uma seguradora nacional focados em dois tipo de planos: PGBL e VGBL. Ambos são planos de previdência complementar no qual o cliente acumula recursos por um prazo contratado e a partir de uma data contratada estipulada pelo cliente, ele recebe a renda de forma única ou mensal. A diferença entre o Plano Gerador de Benefício Livre (PGBL) e o Vida Gerador de Benefício Livre (VGBL) está na tributação. O plano PGBL é indicado àqueles que fazem declaração completa do Imposto de Renda (IR), pois os valores depositados podem ser deduzidos da base de cálculo do IR em até 12% da renda bruta anual. O plano VGBL não permite que os valores depositados sejam deduzidos da base do cálculo do IR, porém apenas valores referentes ao rendimento (ganho de capital) alcançado no plano estão sujeitos à tributação de IR no momento do resgate.

Resumidamente, o evento cancelamento será correlacionado com um conjunto de características do cliente, através dos modelos de sobrevivência,

de forma que os fatores de risco de cancelamento sejam discriminados e quantificados. Assim, é possível, por exemplo, estimar qual a proporção de clientes com um dado perfil vai permanecer ativa no plano até um certo tempo e desses qual é o risco de cancelarem em um certo tempo.

Antes da modelagem, uma análise descritiva própria para dados censurados é usada: as estimativas das curvas de sobrevivência pelo estimador não-paramétrico Kaplan-Meier. Por ela, pode-se obter a probabilidade de clientes sob risco de cancelamento em um dado instante de tempo. Foco será dado a dois pontos importantes na área de seguros: a probabilidade de clientes sobreviverem ao primeiro ano de plano e o tempo mediano, segundo recortes feitos pelas variáveis explicativas.

Dois modelos serão propostos no tratamento do problema. O primeiro é o modelo de regressão paramétrico, cuja variável resposta tem distribuição adequada aos dados, tais como Exponencial, Weibull, Gama, etc. Essas e outras distribuições propostas nessa dissertação são detalhadas no capítulo 2, seção 2.4. O segundo é o modelo de riscos proporcionais de Cox, de natureza semi-paramétrica, detalhado no capítulo 2, seção 2.5. Os modelos ajustados são invariantes no tempo pois as co-variáveis disponíveis permanecem invariantes ao longo do período de observação do cliente.

Os métodos estatísticos da análise de sobrevivência aplicados no estudo da persistência em seguros foram poucos explorados até hoje. Com um banco de dados de onze seguradoras de vida de Singapura, Lian (2) propôs uma abordagem um pouco diferente da que será usada neste trabalho: foi utilizado o estimador da tabela de vida para estimar as funções de sobrevivência e de risco e o modelo paramétrico Weibull para analisar a distribuição do tempo de duração do cliente. Os resultados foram apresentados de maneira superficial e inconclusiva. Não houve testes de significância para as funções de sobrevivência, avaliação do modelo e do poder preditivo.

Gustafsson (3) fez um estudo sobre a duração de clientes na indústria de seguros não-vida, especificamente seguros de automóveis da Dinamarca. Nesse caso, o cliente poderia cancelar o seguro por diversos motivos diferentes, configurando um cenário de riscos competitivos. Para estimar as funções de sobrevivência, foi utilizado o estimador de Kaplan-Meier. O modelo de riscos proporcionais de Cox estratificado, um para cada motivo de cancelamento, foi ajustado com a variável resposta sendo o tempo de duração do cliente até o cancelamento e um conjunto de variáveis regressoras invariantes no tempo relacionadas ao cliente. O modelo ajustado tornou-se uma simplificação excessiva do cancelamento em relação às características do clientes, além de não ter sido feita nenhuma avaliação da qualidade do ajuste do modelo.

No Brasil, Pereira e Torrini (4) aplicaram a técnica de Análise de Sobrevivência para planos de previdência privada. Ajustou-se um modelo de riscos proporcionais para modelagem da duração do cliente e este mostrou-se satisfatório para prever o não cancelamento.

Chun (5) utilizou a técnica *Chain Ladder* e Modelos Lineares Generalizados para modelar a persistência em planos de previdência de dois tipos de planos: benefício definido e contribuição definida. Porém ao final dos ajustes, não houve uma medida comparativa entre os modelos, tornando os resultados incompletos e superficiais.

Saindo do contexto de seguros, um importante artigo tornou-se percursor do uso da análise de sobrevivência em problemas de *credit-scoring*. Stepanova e Thomas (6) propuseram a estimação de um modelo de riscos proporcionais de Cox e de uma modelo de regressão logística para prever quando um cliente detentor de um empréstimo iria se tornar inadimplente ou pagar precocemente sua dívida. Para cada um dos dois eventos (inadimplência e pagamento precoce) foi ajustado um modelo de cada classe (riscos proporcionais de Cox e regressão logística). Depois das estimações dos modelos, avaliou-se o modelo via os resíduos produzidos. Em seguida, utilizando-se uma amostra de validação (*holdout*), calculou-se as curvas ROC e a tabela de classificação para avaliar o poder preditivo do modelo. Os modelos se mostraram adequados para o tratamento do problema. Posteriormente, Abreu (7) usou a mesma metodologia para o cenário brasileiro utilizando uma base de dados de uma instituição financeira sediada no Brasil. O foco do trabalho foi variar diversas amostras de acordo com balanceamento de bons e maus clientes (classificados pela ocorrência ou não de inadimplência). A validação foi feita pela curva ROC e os resultados foram satisfatórios.

Essa dissertação propõe uma abordagem original para o tratamento da persistência em planos de previdência privada, baseado principalmente pelo artigo de Stepanova e Thomas (6), percursor da aplicação de análise de sobrevivência em problemas de gestão financeira. A estimação dos modelos será feita de maneira criteriosa, passando pela adequação através dos resíduos e pela validação dentro e fora da amostra de estimação.

O capítulo 2 apresenta a teoria envolvida nos modelos e métodos usados nessa dissertação. Os dados utilizados são descritos e detalhados no capítulo 3. O capítulo 4 contém os resultados da análise descritiva, estimação dos modelos, adequação e validação deles. Por fim, o capítulo 5 apresenta a conclusão e sugestões e trabalhos futuro.

2
Modelos e métodos

Neste capítulo introduziremos definições básicas e teóricas de conceitos e modelos estatísticos úteis para a estimação da persistência em planos de previdência privada e a determinação dos fatores que influenciam na persistência. Para uma leitura introdutória, veja Colosimo (8).

2.1
Censura e dados truncados

Seja T o tempo medido em alguma unidade de um objeto a partir de um ponto de referência (ponto de origem) até a ocorrência de um evento de interesse (falha) ou até o final do período de estudo. O tamanho desse intervalo é o tempo de sobrevivência e este é tratado como uma variável aleatória contínua não-negativa com função de distribuição $F(t) = P(T \leq t)$ e função de densidade $f(t) = \frac{d}{dt}F(t)$. Nos modelos de sobrevivência, T é a variável resposta, que no nosso contexto é o tempo de duração entre a entrada do participante em um plano de previdência privada até o seu cancelamento.

A principal característica em dados de sobrevivência é a presença de censura. A censura se refere a um dado que não teve o tempo de sobrevivência observado, seja porque o evento não ocorreu no período de estudo ou porque foi retirado da amostra por alguma razão que não a ocorrência do evento. O tempo observado de sobrevivência do indivíduo será o mínimo entre o tempo de falha e o de censura.

Existem diferentes tipos de censura (ver Klein (9)). A censura do tipo I é aquela em que existe um período fixo pré-determinado para o estudo, portanto o número de falhas que podem acontecer é aleatório. Quando o interesse é de se observar um pré-determinado número de falhas, então neste estudo ocorre a censura do tipo II. Se o indivíduo for retirado do estudo antes da ocorrência da falha por alguma razão diferente do evento de interesse, ocorre a censura do tipo aleatória.

De uma forma mais geral, pode-se representar a censura aleatória da seguinte maneira: seja T_i a variável aleatória representando o tempo de falha do *i-ésimo* indivíduo na amostra e C_i uma outra variável aleatória representando

o tempo de censura deste mesmo paciente. Assume-se que essas duas variáveis aleatórias são independentes, pois o evento acontece por razões que não podem ser controladas. Assim, criamos uma variável relacionada com cada indivíduo para indicar quando houve a falha, δ_i, tal que

$$\delta_i = \begin{cases} 1 & \text{, se falha} \\ 0 & \text{, se censura} \end{cases} \tag{2-1}$$

Quando há presença de co-variáveis, os dados de sobrevivência para o *i-ésimo* indivíduo serão representados por (t_i, δ_i, x_i), onde t_i é o tempo observado e x_i é o vetor de variáveis explicativas.

Quando a falha não ocorre durante o período de observação mas em algum momento depois, diz-que que o dado é censurado à direita, pois a ocorrência do evento esta à direita do tempo registrado. Este tipo de censura é o mais recorrente em estudos com dados de sobrevivência. A censura à esquerda ocorre quando o evento de interesse já ocorreu quando o objeto de estudo foi observado, ou seja, o tempo registrado é maior que o tempo de falha. Se é sabido que a falha sucedeu em algum ponto entre dois tempos, o dado é um caso de censura intervalar.

Outra característica presente em alguns dados de análise de sobrevivência é o truncamento. Se o indivíduo não é observado desde o início do estudo, mas a partir de algum ponto u_i tal que $u_i \leq t_i$, dizemos que o tempo observado é truncado à esquerda. Um exemplo para ilustrar o truncamento à direita foi dado por Kalbfleisch (10). Em um estudo para estimar o tempo de vida entre a infecção pelo vírus HIV e o diagnóstico da AIDS em pessoas que contraíram o vírus por transfusão de sangue, o grupo de pessoas selecionadas consistiam naquelas que tiveram o diagnóstico da AIDS antes do dia primeiro de julho de 1986. Como a data de infecção pelo vírus era conhecida, devido à origem dos dados, tem-se então dados truncados à direita.

Nos dados utilizados neste estudo, a data do início do contrato e do seu respectivo cancelamento (se houver) são conhecidas e, assim os dados censurados serão aqueles em que o evento de interesse, cancelamento do plano, não for observado até o final do período de estudo. Ou seja, todas as observações censuradas aconteceram simplesmente porque os indivíduos não foram acompanhados por tempo suficiente. Desta forma, quando for mencionado *dado censurado* ao longo do texto, trata-se da censura à direita. Quando há presença de censura, não é possível ter alguma ideia sobre distribuição do tempo de vida. Neste caso, como há censuras à direita, não se tem conhecimento da cauda superior da distribuição.

A unidade de medida do tempo até o cancelamento do plano de pre-

vidência é o número de dias. Para que a interpretação dos resultados sejam mais simples, a duração foi convertida para anos. Assim, a duração observada do cliente nesse estudo pode ter qualquer valor entre o intervalo $[0; 6, 53)$.

Para que se possa fazer inferências sobre o tempo de falha T, a censura deve ser considerada não-informativa, isto é, assume-se que T e C são variáveis aleatórias independentes. Na prática isto significa que se o dado for censurado, a única afirmação sobre o tempo de vida deste é indivíduo é que seu tempo de falha é maior que o observado.

2.2 Função de risco e probabilidade de sobrevivência

Seja T uma variável aleatória contínua representando o tempo de sobrevivência e x sendo o vetor de co-variáveis. No nosso contexto, probabilidade de um cliente não cancelar seu plano de previdência no tempo t será dada por:

$$S(t) = P(T > t) = 1 - F(t) \tag{2-2}$$

sob as condições $t \geq 0, S(0) = 1$ e $S(\infty) = 0$.

Outra importante função é a função de risco, também conhecida por taxa de risco ou de falha, definida por:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{2-3}$$

Pode-se chegar a seguinte relação, a partir de 2-3:

$$h(t) = \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt} \cdot \frac{1}{S(t)} = \frac{f(t)}{S(t)} \tag{2-4}$$

descrevendo assim a relação entre as três mais importantes funções utilizadas para representar o tempo de sobrevivência.

A taxa de falha é interpretada como o risco de acontecer a falha instantaneamente após t, dado o vetor x de co-variáveis e que o indivíduo sobreviveu à t. Ela é calculada pela a probabilidade do indivíduo falhar no intervalo $[t, t + dt)$ dado um conjunto de co-variáveis x e que este sobreviveu até o tempo t dividida por dt, tornando-se assim uma taxa.

Sabendo que $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$, pode-se chegar a:

$$h(t) = -\frac{d}{dt} log\left[S(t)\right] \Longleftrightarrow S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right) \tag{2-5}$$

onde $H(t) = \int_0^t h(u)du$ é a função de falha acumulada e, sendo assim, a relação (seção 2-5) só é válida para distribuições contínuas em T.

As relações entre as funções taxa de risco, de densidade e de probabilidade de sobrevivência explicitadas em nas equações (2-3) e (2-5) são importantes no desenvolvimento dos modelos, uma vez que ter conhecimento de uma dessas funções implica em ter conhecimento das demais.

2.2.1 Estimador de Kaplan-Meier

Para estimar a função de sobrevivência $S(t)$, utiliza-se o estimador de máxima verossimilhança não-paramétrico de Kaplan-Meier (11) definido em (2-6). Este é o estimador mais utilizado para estimar $S(t)$ em análise de sobrevivência. Segundo Breslow (12), esse estimador é fracamente consistente e possui distribuição assintótica normal, sendo dado por:

$$\widehat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{n_j}) \tag{2-6}$$

onde t_j é o tempo de falha do *j-ésimo* indivíduo tal que $j = 1, 2, ..., k$ e k são os tempos distintos e ordenados de falha, d_j é o número de falhas em t_j e n_j é o número de indivíduos que podem falhar em t_j, ou seja, são indivíduos que sobreviveram e não foram censurados até o instante imediatamente anterior a t_j.

A variância do estimador Kaplan-Meier é estimada através da expressão (2-7), conhecida como fórmula de Greenwood (13):

$$\widehat{Var}(\widehat{S}(t)) = \left[\widehat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2-7}$$

Consequentemente, como $\widehat{S}(t)$ tem distribuição assintótica normal, o intervalo de $100(1-\alpha)\%$ confiança para $S(t)$ será dado por:

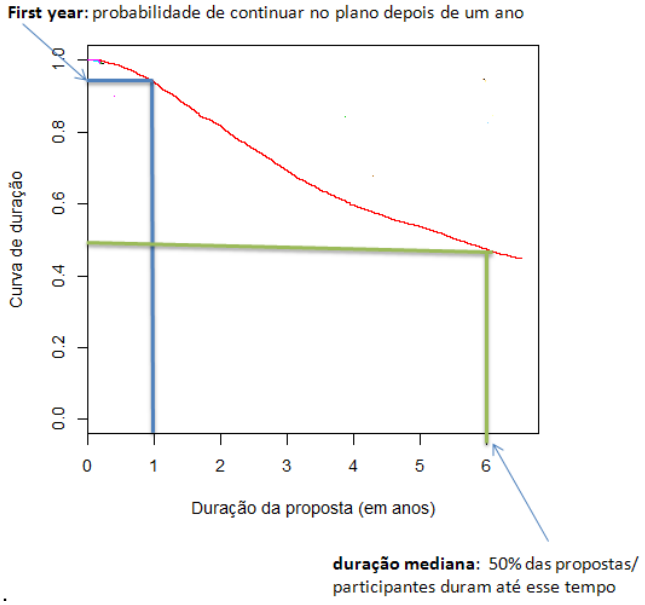
$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{S}(t))} \tag{2-8}$$

onde $\alpha/2$ é o percentil da distribuição Normal com média igual zero e variância igual a um.

A Figura (2.1) apresenta uma forma geral de uma curva de sobrevivência estimada por Kaplan-Meier e dois importantes pontos no contexto de previdência privada: o chamado *first year* e o tempo mediano. O primeiro ano (*first year*) de um segurado em uma instituição com um plano de previdência privada é muitas vezes o período em que se define os prêmios a serem pagos de acordo com a rentabilidade neste ano. Assim sendo, essa questão será tratada analisando-se a estimativa de Kaplan-Meier dada quando $t = 1$. Ou seja, saberemos qual é a probabilidade de um cliente permanecer ativo após o aniversário de um ano do seu plano de previdência. No caso da Figura (2.1), a probabilidade do cliente não cancelar o plano é de aproximadamente 97%. O

tempo mediano dá a informação de que 50% dos clientes duram esse tempo. Na Figura (2.1) o tempo mediano é de seis anos.

Figura 2.1: Exemplo de estimativa da função de sobrevivência por Kaplan-Meier.



Para cada variável explicativa presente no estudo é possível verificar se as r funções de sobrevivência $S_1(t), ..., S_r(t)$ diferem entre si, onde r é o número de níveis da variável explicativa em questão. Dois testes não-paramétricos serão utilizados para este fim nesse trabalho: o teste *logrank* (14) e o teste de Wilcoxon (15), que serão brevemente discutidos a seguir.

Teste logrank

Sejam $t_1 < t_2 < ... < t_k$ os k tempos de falha distintos da amostra. Suponha que no tempo t_j haja d_j falhas e n_j indivíduos estejam sob risco no tempo imediatamente anterior a t_j, onde $j = 1, 2, ...k$. Considere o teste de igualdade de r funções de sobrevivência $S_1(t), ..., S_r(t)$. Cada uma dessas funções será calculada com sua respectiva amostra. Por exemplo, seja uma variável dicotômica indicadora do gênero. Serão duas funções de sobrevivência estimada, uma para o gênero feminino e outra para o masculino.

A distribuição conjunta de $d_{2j}, ..., d_{rj}$, onde d_{ij} é o número de falhas no tempo j e amostra r, é:

$$\frac{\prod_{i=1}^{r} \binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}} \tag{2-9}$$

tal que a média de d_{ij} é

$$w_{ij} = n_{ij} d_j n_j^{-1}, \tag{2-10}$$

a variância igual a

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1} \tag{2-11}$$

e covariância de d_{ij} e d_{lj} igual a

$$(V_j)_{il} = -n_{ij} n_{lj} d_j (n_j - d_j)^{-2} (n_j - 1)^{-1}. \tag{2-12}$$

A estatística de teste sob a hipótese nula de igualdade das r funções de sobrevivência é dada por:

$$T = v' V^{-1} v, \tag{2-13}$$

tal que a estatística $v = \sum_j^k v_j$ é o vetor de dimensão $(r-1) \times 1$ formado pela soma dos elementos:

$$v_j' = (d_{2j} - w_{2j}, ..., d_{rj} - w_{rj}) \tag{2-14}$$

que tem média zero e matriz de variância-covariância V_j de tamanha $r-1$ formada por $(V_j)^{ii}, i = 2, ..., r$ na diagonal principal e $(V_j)^{il}, i, l = 2, ..., r$ nas demais posições.

Sob H_0, a estatística de teste 2-13 tem distribuição χ^2_{r-1}.

O teste de Wilcoxon é obtido através de uma variação do teste *logrank*: para cada amostra é dado um peso n_j, que é o número de indivíduos sob risco em $j, j = 1, ..., k$.

2.3 Modelos

Frequentemente, dados relacionados com estudos de tempo de sobrevivência são provenientes de populações heterogêneas, ou seja, além do tempo observado, há a presença de variáveis explicativas que podem estar relacionadas com este tempo. Em casos como esses, justifica-se o uso de modelos de regressão. A seguir serão apresentados as duas classes de modelos mais usados em análise de sobrevivência: modelo de tempo de vida acelerado e modelo de regressão de Cox, também conhecido por modelo de riscos proporcionais, proposto por Cox (16).

Antes de introduzirmos a metodologia a ser utilizada nesse trabalho, os modelos de sobrevivência, faremos um breve resumo sobre modelos lineares generalizados de forma a relacionarmos com os modelos de sobrevivência.

Os modelos clássicos de regressão linear são definidos por:

$$Y_i = \boldsymbol{X_i'}\boldsymbol{\beta} + \epsilon_i, i = 1, 2, ..., n \tag{2-15}$$

sendo Y_i a variável resposta aleatória do modelo associada ao i-ésimo indivíduo da amostra de tamanho n, $\boldsymbol{X_i'}$ o vetor de variáveis independentes de dimensão $(1 \times p)$, $\boldsymbol{\beta}$ o vetor de parâmetros desconhecidos de dimensão $(p \times 1)$ e o componente aleatório ϵ_i com distribuição normal de média zero e variância σ^2. Consequentemente, a variável resposta tem distribuição condicional normal com média $E(Y_i|X_i) = \mu_i = \boldsymbol{X_i'}\boldsymbol{\beta}$ e variância $Var(Y_i|X_i) = \sigma^2$.

O modelo linear generalizado é uma extensão do modelo clássico de regressão linear na qual a variável resposta pode ter distribuição pertencente à família de distribuições exponencial.

A família exponencial uniparamétrica é caracterizada por uma função de probabilidade ou densidade de forma:

$$f(y, \theta) = h(y)exp\left[\eta(\theta)t(y) - b(\theta)\right] \tag{2-16}$$

tal que as funções $h(y)$, $t(y)$, $\eta(\theta)$ e $b(\theta)$ possuem valores no conjunto dos reais e θ é o vetor de parâmetros da distribuição de Y.

Para que uma distribuição pertença à família de distribuições exponencial, esta deve ser expressa na forma da equação (2-16). Por exemplo, seja uma variável aleatória Y $Poisson(\theta)$ com $\theta > 0$. Sua função de probabilidade pode ser escrita como:

$$f(y; \theta) = \frac{e^{-\theta}\theta^y}{y!} = \frac{1}{y!}exp\left\{ylog(\theta) - \theta\right\} \tag{2-17}$$

onde $h(y) = \frac{1}{y!}$; $\eta(\theta) = log(\theta)$; $t(y) = y$ e $b(\theta) = \theta$. Logo, a distribuição Poisson pertence à família de distribuições exponencial.

Além da distribuição Poisson, outras importantes distribuições também pertencem à família exponencial: binomial, normal, gama e normal inversa são outros exemplos.

Portanto, tendo-se uma variável resposta univariada, um conjunto de variáveis explicativas e uma amostra aleatória de tamanho n, podemos definir as componentes que formam um modelo linear generalizado (para maiores detalhes, ver Gauss (17)):

1. a variável resposta com distribuição pertencente à família de distribuições conforme (2-16);

2. o conjunto das variáveis explicativas que têm uma relação linear no modelo, formando o preditor linear $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$;

3. uma função de ligação adequada para conectar a parte aleatória do modelo (variável resposta) com a parte sistemática (variáveis explicativas).

Definindo de forma generalizada, a função de densidade ou de probabilidade de um conjunto de variáveis aleatórias Y_i, onde $i = 1, ..., n$ é dada por:

$$f(y_i, \theta_i, \phi) = exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\} \tag{2-18}$$

tal que $\phi > 0$ é um parâmetro de dispersão, θ_i é o parâmetro canônico, $b(.)$ e $c(.)$ são funções conhecidas. O componente sistemático é dado por η_i

O valor esperado $E(Y_i)$ e a variância $Var(Y_i)$ podem ser obtidos através da função de verossimilhança utilizando-se as propriedades da função escore (ver Gauss (17) p.9). As expressões do valor esperado e da variância são dadas por $E(Y_i) = \mu_i = b'(\theta_i)$ e $Var(Y_i) = \phi b''(\theta_i)$, respectivamente.

O componente sistemático é dado por $\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r$ ou $\boldsymbol{\eta} = \boldsymbol{X\beta}$, onde $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$ é a matriz de variáveis explicativas do modelo, $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)'$ o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)'$ o preditor linear. A variável aleatória é relacionada com o preditor linear através de uma função de ligação $\eta_i = g(\mu_i)$ tal que $g(.)$ é uma função monótona e diferenciável.

A escolha da distribuição em um modelo linear generalizado é feita levando-se em conta o tipo da variável resposta (discreta ou contínua) e seu intervalo de variação. A função de ligação é escolhido de acordo com o problema a ser tratado. As funções identidade, logarítmica e logística são alguns exemplos de possíveis funções de ligação.

Os modelos lineares generalizados são adequados em situações onde a variável resposta pode ser dicotômica, discreta ou contínua. Portanto, são uma classe de modelo bem flexível.

Um tipo específico de variável resposta é frequentemente encontrado em análise de dados: o tempo de sobrevivência. Pode ser o tempo de sobrevida de um indivíduo acometido por uma doença, o tempo de vida útil uma máquina ou o tempo de sobrevivência até que ocorra qualquer outro evento de interesse, chamado falha. Diante de situações como essas, é natural que nem todos os dados observados venham a "morrer", ou seja, nem todos os dados falharão até o final do tempo de observação do estudo. Esses dados são chamados de censura, visto que o evento de interesse não ocorreu durante o período de observação.

Os modelos de sobrevivência são ideais nesses casos onde a variável resposta é o tempo até a ocorrência do evento de interesse (falha), pois têm

como diferencial o fato de incorporar dados censurados nas análises de dados e estimação dos modelos, além dos dados em que ocorreram o evento de interesse. Nessa dissertação, o tempo de sobrevivência é o tempo de duração de um cliente até o cancelamento do seu plano de previdência privada.

O tempo de sobrevivência pode ser expresso através de três funções: a função de densidade do tempo de sobrevivência, $f(t)$; a função de sobrevivência $S(t) = P(T > t)$ que está diretamente relacionada com a função de distribuição acumulada $F(t)$ de forma que $S(t) = 1 - F(t)$; $h(t)$, que é a função de risco instantâneo de falha em t, dada por $h(t) = F'(t)/[1 - F(t)]$. Essas três funções serão especificadas detalhadamente na seção 2.2.

O modelo proposto por Cox (16) relaciona de forma multiplicativa a função de risco dependente exclusivamente do tempo com o efeito das variáveis explicativas. Ele é dado por:

$$h(t|\boldsymbol{x}) = \lambda(t)exp(\boldsymbol{x'\beta}) \tag{2-19}$$

tal que $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ é o vetor de parâmetros desconhecidos de dimensão $(p \times 1)$ associados ao vetor de variáveis explicativas $\boldsymbol{x'} = (x_1, ..., x_p)$, $\lambda(t)$ é uma função não-negativa dependente do tempo e da distribuição de T e $\boldsymbol{\eta = x'\beta}$ é o preditor linear.

A partir da equação (2-19) pode-se chegar à função de sobrevivência:

$$S(t|\boldsymbol{x}) = exp\left[-H(t)exp(\boldsymbol{x'\beta})\right] \tag{2-20}$$

onde $H(t) = \int_{-\infty}^{t} h(u)du$.

A função de densidade de probabilidade de T é dada por:

$$f(t|\boldsymbol{x}) = H'(t)exp\left[\eta - h(t)exp(\eta)\right]. \tag{2-21}$$

Segundo Cordeiro e Demétrio (17), a distribuição do tempo de sobrevivência T do modelo com função de densidade (2-21) pertence à família exponencial não-linear, mas não à família (2-16).

Em geral, a distribuição da variável resposta tende a ser assimétrica à direita, Em situações como essa, algumas distribuições do tempo de sobrevivência T são mais adequadas como por exemplo exponencial, Weibull, Gama, Gama generalizada, Log-normal e Log-logística.

Quando assume-se que a variável resposta T tem uma distribuição paramétrica pertencente à família exponencial não-linear, o modelo mais utilizado para estimação é o modelo de tempo de vida acelerado na forma log-linear, dado por:

$$Y = log(T) = \boldsymbol{x'\beta} + \sigma\nu \tag{2-22}$$

em que σ é um parâmetro de escala para acomodar adequadamente a distribuição do erro ν de acordo com a distribuição de T. Por exemplo, se no modelo regressão (2-22) T tiver distribuição Weibull, então o $log(\nu)$ terá distribuição do valor extremo com parâmetro de escala σ.

A classe de modelos de tempo de vida acelerado tem esse nome porque as variáveis explicativas têm a função de acelerar ou desacelerar o tempo de sobrevivência.

O método de estimação dos coeficientes do modelo de tempo de vida acelerado é feito por método de máxima verossimilhança, que será descrito na seção 2.4.6.

Outro modelo largamente utilizado na análise de sobrevivência é o modelo de riscos proporcionais de Cox (16). Trata-se de uma versão semi-paramétrica do modelo de riscos proporcionais (2-19) formada por duas partes: a primeira a parte não-paramétrica composta pela função $\lambda(t)$ não-negativa e arbitrária dependente de t e a segunda é a parte paramétrica formada pelo componente paramétrico $exp\{\boldsymbol{x'\beta}\}$.

Na estimação do modelo de riscos proporcionais de Cox, não é preciso atribuir nenhuma distribuição de probabilidade ao tempo de sobrevivência T. O método de estimação utilizado nesse modelo, o de máxima verossimilhança parcial, condiciona a função de verossimilhança à história passada de falhas e censuras eliminando assim da estimação a função $\lambda(t)$. Esse método será explicado com mais detalhes na seção 2.5.1.

2.4 Modelo de tempo de vida acelerado

A expressão geral de um modelo de tempo de vida acelerado com a variável aleatória resposta T é dada por:

$$T = exp\{\boldsymbol{x'\beta}\}exp\{\sigma\nu\} \tag{2-23}$$

onde $\boldsymbol{x} = (1, x_1, ..., x_p)'$ é o vetor com p co-variáveis, $\boldsymbol{\beta'} = (1, \beta_1, ..., \beta_p)$ é o vetor de p parâmetros associados às co-variáveis, σ um parâmetro de escala e ν é o erro aleatório.

Para que se possa fazer inferências estatísticas no contexto de Análise de Sobrevivência, pode-se linearizar o modelo 2-23 tomando-se o logaritmo de T e transformando-o no modelo de regressão linear:

$$Y = log(T) = \boldsymbol{x'\beta} + \sigma\nu \tag{2-24}$$

A distribuição da variável do erro aleatório ν é determinada pela distribuição da variável aleatória resposta T. Devido à forma assimétrica dos

tempos em análise de Sobrevivência, algumas distribuições de T específicas são usualmente utilizadas: Weibull ou exponencial, log-normal, gama, gama generalizada e logística. Respectivamente, a distribuição do erro aleatório será: valor extremo, normal, log-gama, log-gama generalizada[1] e logística.

Pode-se obter as funções de sobrevivência para Y condicional a $\boldsymbol{x}$ e para T condicional a $\boldsymbol{x}$:

$$S(y|\boldsymbol{x}) = exp\left\{-exp\left\{\frac{y-\boldsymbol{x'\beta}}{\sigma}\right\}\right\} \tag{2-25}$$

e

$$S(t|\boldsymbol{x}) = exp\left\{-\left(\frac{t}{exp\left\{\boldsymbol{x'\beta}\right\}}\right)^{1/\sigma}\right\} \tag{2-26}$$

A seguir serão mostradas as funções de densidade das distribuições usadas nesse trabalho.

2.4.1 Modelo de regressão exponencial

A função de densidade de probabilidade para uma variável aleatória que mede o tempo de duração com distribuição exponencial é:

$$f(t) = \frac{1}{\alpha}exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \tag{2-27}$$

em que $t \geqslant 0$ e o parâmetro $\alpha > 0$.

Substituindo $1/\alpha$ por $exp\{\boldsymbol{x'\beta}\}$, temos a função de densidade de um modelo de regressão exponencial, dado por:

$$f(t|\boldsymbol{x}) = exp\{\boldsymbol{x'\beta}\}exp\left\{-\left(\frac{t}{exp\{\boldsymbol{x'\beta}\}}\right)\right\} \tag{2-28}$$

com função de risco:

$$h(t|\boldsymbol{x}) = exp\{\boldsymbol{x'\beta}\} \tag{2-29}$$

e função de sobrevivência:

$$S(t|\boldsymbol{x}) = exp\left\{-\left(\frac{t}{exp\{\boldsymbol{x'\beta}\}}\right)\right\}. \tag{2-30}$$

[1]Neste caso, é acrescentado mais um parâmetro de escala na fórmula 2-24.

2.4.2
Modelo de regressão Weibull

A função de densidade de probabilidade para uma variável aleatória que mede o tempo de duração com distribuição Weibull é:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{2-31}$$

em que $t \geqslant 0$, o parâmetro de escala $\alpha > 0$ e o parâmetro de forma $\gamma > 0$.

Para incorporar o vetor de variáveis explicativas na equação (2-31), basta fazer a reparametrização substituindo o parâmetro de escala α por $exp\{\boldsymbol{x'\beta}\}$. O modelo de regressão terá a seguinte função de densidade:

$$f(t|\boldsymbol{x}) = \frac{\gamma}{[exp\{\boldsymbol{x'\beta}\}]^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{exp\{\boldsymbol{x'\beta}\}}\right)^{\gamma}\right\}. \tag{2-32}$$

A função de risco será dada por:

$$h(t|\boldsymbol{x}) = \gamma\lambda t^{\gamma-1} \tag{2-33}$$

e a função de sobrevivência:

$$S(t|\boldsymbol{x}) = exp\left(-\lambda t^{\gamma}\right) \tag{2-34}$$

onde $\lambda = exp\{\boldsymbol{x'\beta}\}$.

2.4.3
Modelo de regressão Log-normal

A função de densidade de probabilidade para uma variável aleatória que mede o tempo de duração com distribuição log-normal é:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\} \tag{2-35}$$

em que $t > 0$, μ é a média do logaritmo do tempo de falha e σ o desvio padrão.

O modelo de regressão log-normal tem a função de densidade igual a equação (2-35) substituindo μ por $\boldsymbol{x'\beta}$.

Nesse caso, a função de sobrevivência será:

$$S(t|\boldsymbol{x}) = 1 - \Phi\left\{\frac{ln(t)-\mu}{\sigma}\right\} \tag{2-36}$$

em que $\Phi(z)$ é a função de distribuição cumulativa normal padrão no ponto z, onde $z = \dfrac{ln(t)-\mu}{\sigma}$.

2.4.4
Modelo de regressão Log-logística

A função de densidade de probabilidade para uma variável aleatória que mede o tempo de duração com distribuição log-logística é:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \left(1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right)^{-2} \tag{2-37}$$

em que $t > 0$, $\alpha > 0$ é o parâmetro de forma e $\gamma > 0$ é o parâmetro de escala.

Quando substituímos o parâmetro de escala α por $exp\{-\boldsymbol{x'\beta}\}$, temos a função de densidade de um modelo de regressão log-logístico dada por:

$$f(t|\boldsymbol{x}) = \frac{\lambda^{1/\gamma} t^{1/(\gamma-1)}}{\gamma\{1 + (\lambda t)^{1/\gamma}\}^2} \tag{2-38}$$

com função de sobrevivência dada por:

$$S(t|\boldsymbol{x}) = \left\{1 + (\lambda t)^{1/\gamma}\right\}^{-1} \tag{2-39}$$

2.4.5
Modelo de regressão Gama generalizado

A função de densidade de probabilidade para uma variável aleatória que mede o tempo de duração com distribuição gama generalizada é:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{\gamma k-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{2-40}$$

em que $t > 0$, $\Gamma(k)$ é a função gama, $\alpha > 0$ o parâmetro de escala e $\gamma > 0$ e $k > 0$ os parâmetros de forma.

A função de densidade gama generalizada na presença de variáveis explicativas fica como a equação (2-40) substituindo o parâmetro de escala α por $\boldsymbol{x'\beta}$.

2.4.6
Estimação dos parâmetros do modelo de regressão paramétrico

A estimação dos parâmetros de um modelo de regressão paramétrico é feita pelo método de verossimilhança. Como já foi dito, na base de dados utilizada nessa dissertação nem todos os clientes cancelaram seus planos no término do período de observação, sendo assim esses dados censurados.

Se a observação não for censurada, ou seja, o cliente tiver cancelado em algum ponto até o final do período observado, sua contribuição na função de verossimilhança é a própria função de densidade. Caso a observação seja censurada (cliente não cancelou o plano até o final do estudo), então a

sua contribuição para a função de verossimilhança será a sua função de sobrevivência. Nesse caso, o dado censurado apenas informa que o tempo até o cancelamento é maior que o tempo de censura observado.

Suponha que a amostra contenha n clientes e que esse conjunto seja dividido em dois sub-conjuntos: as r primeiras observações são os clientes que cancelaram e as seguintes $n - r$ são os clientes com dado censurado. Sendo assim, a função de verossimilhança será dada por:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \boldsymbol{x_i}; \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i; \boldsymbol{x_i}; \boldsymbol{\theta}) \tag{2-41}$$

onde $f(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})$ é a função de densidade do $i-$ésimo cliente se este não for censurado, com vetor de parâmetros $\boldsymbol{\theta}$ e conjunto de variáveis explicativas $\boldsymbol{x_i}$. Caso o dado seja censurado, $S(t_i; \boldsymbol{x_i}; \theta)$ será a função de sobrevivência deste i-ésimo dado censurado, com vetor de parâmetros $\boldsymbol{\theta}$ e conjunto de variáveis explicativas $\boldsymbol{x_i}$.

Usando-se a relação entre as funções de densidade, de risco e de sobrevivência, pode-se escrever a função de verossimilhança da seguinte forma:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i; \boldsymbol{x_i}; \boldsymbol{\theta}) S(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i; \boldsymbol{x_i}; \boldsymbol{\theta})]^{\delta_i} S(t_i; \boldsymbol{x_i}; \boldsymbol{\theta}) \end{aligned} \tag{2-42}$$

Por conveniência para cálculos, toma-se o logaritmo de (2-42), já que o máximo de uma função monotônica não se modifica com a transformação logarítmica. Então, os estimadores de máxima verossimilhança serão os valores do vetor de parâmetros $\boldsymbol{\theta}$ que maximizam a logverossimilhança $log(L(\theta))$, calculados pela equação dada pela derivada de $log(L(\theta))$ igualada a zero:

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} \tag{2-43}$$

A equação (2-43) será formada pelos parâmetros da distribuição escolhida. Para resolver esse sistema de equações, um método numérico deve ser utilizado, como por exemplo o de Newton-Raphson.

2.4.7
Interpretação dos coeficientes estimados no modelo de regressão paramétrico

A interpretação dos coeficientes estimados em um modelo de regressão paramétrico em análise de sobrevivência é feita através da razão dos tempos medianos estimados proposta por Hosmer (18). A mediana estimada $t_{0,5}$ pode ser calculada através de uma das três funções: de densidade, $f(t)$, de sobrevivência, $S(t)$ ou de risco, $h(t)$. Para qualquer um dos modelos que serão utilizados nessa dissertação, o tempo mediano estimado será dado por:

$$\widehat{t}_{0,5}(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}) = u(\widehat{\theta}) * exp\{\boldsymbol{x'}\boldsymbol{\beta}\} \tag{2-44}$$

em que $u(\widehat{\theta})$ é uma função dos parâmetros de forma da distribuição sendo utilizada.

Por exemplo, se T tem uma distribuição de Weibull com parâmetros $exp\{\beta_0 + \beta_1 x\}$ e γ, onde x é uma variável dicotômica, o tempo mediano será:

$$\widehat{t}_{0,5} = (-ln0,5)^{\widehat{\gamma}} exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x\}. \tag{2-45}$$

A razão dos tempos medianos ($RT_{0,5}$) entre o indivíduo que tem $x = 1$ e $x = 0$ será:

$$RT_{0,5}(x=1) = \frac{\widehat{t}_{0,5}(x=1, \widehat{\boldsymbol{\beta}})}{\widehat{t}_{0,5}(x=0, \widehat{\boldsymbol{\beta}})} = \frac{(-ln0,5)^{\widehat{\gamma}} exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x\}}{(-ln0,5)^{\widehat{\gamma}} exp\{\widehat{\beta}_0\}} = exp\{\widehat{\beta}_1\}. \tag{2-46}$$

Sendo assim, o tempo mediano do indivíduo com $x = 1$ será $exp\{\widehat{\beta}_1\}$ vezes o tempo mediano do indivíduo com $x = 0$. Essa proporcionalidade é garantida para todos os modelos de tempo de vida acelerado.

Se a co-variável for categórica com m níveis, a interpretação pela razão dos tempos medianos também poderá ser feita, de modo que as estimativas serão comparadas com o grupo do m-ésimo nível de referência.

2.5
Modelo de riscos proporcionais de Cox

Considere $\boldsymbol{x} = (1, x_1, ..., x_p)'$ um vetor com p co-variáveis e $\boldsymbol{\beta'} = (1, \beta_1, ..., \beta_p)$ o vetor de p parâmetros associados às co-variáveis. A expressão geral da função de risco do modelo é:

$$h(t|\boldsymbol{x}) = h_0(t) g(\boldsymbol{x'}\boldsymbol{\beta}) \tag{2-47}$$

onde $h(t|\boldsymbol{x})$ é a função de risco no tempo t para um indivíduo com vetor de co-variáveis x, o componente não-paramétrico $h_0(t)$ é uma função do tempo positiva e não-especificada conhecida como função de base ou basal, e o componente paramétrico $g(\boldsymbol{x'\beta})$ é uma função não negativa a ser especificada. Devido a esta construção, o modelo é de natureza semi-paramétrica.

Para garantir que $h(t|\boldsymbol{x})$ seja sempre positiva, a forma mais utilizada na literatura e que também será utilizada neste trabalho é a função na forma multiplicativa $g(\boldsymbol{x'\beta}) = exp\{\boldsymbol{x'\beta}\}$, resultando no modelo com taxa de risco:

$$h(t|\boldsymbol{x}) = h_0(t)\exp\{\boldsymbol{x'\beta}\} \tag{2-48}$$

O modelo de regressão de Cox (2-48) também é conhecido como *modelo de riscos proporcionais*, pois supõe-se que a razão das taxas de falha de dois indivíduos diferentes é constante ao longo do tempo. Ou seja,

$$\frac{h(t|\boldsymbol{x_i})}{h(t|\boldsymbol{x_j})} = \frac{h_0(t)\exp\{\boldsymbol{x_i'\beta}\}}{h_0(t)\exp\{\boldsymbol{x_j'\beta}\}} = \exp\{\boldsymbol{x_i'\beta} - \boldsymbol{x_j'\beta}\} = exp\{(\boldsymbol{x_i'} - \boldsymbol{x_j'})\boldsymbol{\beta}\}. \tag{2-49}$$

Duas importantes generalizações deste modelo devem ser citadas. A primeira generalização,é apropriada para casos onde existem variáveis explicativas dependentes do tempo, ou seja, co-variáveis que variam no decorrer do tempo. O modelo terá a seguinte função de taxa de risco:

$$h[t|\boldsymbol{x(t)}] = h_0(t)\exp\{\boldsymbol{x'(t)\beta}\} \tag{2-50}$$

em que $x(t)$ é o vetor de co-variáveis que podem depender do tempo. Neste estudo, não há a presença de co-variáveis dependentes do tempo e portanto esta generalização não será utilizada.

Na outra generalização, o componente não-paramétrico $h_0(t)$ arbitrário pode variar em subconjuntos específicos formados no banco de dados. Ou seja, se houver interesse em dividir os dados em m estratos, o modelo terá a função de risco dada por

$$h_j(t|\boldsymbol{x_j(t)}) = h_{0_j}(t)\exp\{\boldsymbol{x_j'(t)\beta}\} \tag{2-51}$$

com $j = 1, ..., m$. Fazendo isto, os riscos dos indivíduos nos diferentes subgrupos são afetados igualmente pelas co-variáveis, mas podem se diferenciar na função de base. As funções basais são arbitrárias e não relacionadas. Este modelo é adequado em casos onde há evidências de não-proporcionalidade causada por alguma co-variável.

2.5.1
Estimação dos parâmetros do modelo de riscos proporcionais

Para que sejam feitas inferências sob os parâmetros de um modelo estatístico, o método de máxima verossimilhança, proposto por Cox e Hinkley (19), é frequentemente utilizado. Porém, como o modelo de Cox com função de taxa de risco apresentado em (2-48) é semi-paramétrico, esse método deve ser adaptado. O método conhecido como *máxima verossimilhança parcial*, proposto por Cox (20) é utilizado. Na sua derivação não é necessário fazer nenhuma suposição sobre a função basal. Sua solução é apresentada a seguir.

Seja uma amostra de n objetos com $r \leqslant n$ falhas distintas nos tempos $t_1 < t_2 ... < t_r$. A função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} \right)^{\delta_i} \tag{2-52}$$

em que $i = 1, ..., n$ é o conjunto ordenado dos tempos observados, $R(t_i)$ é o conjunto de indivíduos sob risco no tempo j, δ_i é o indicador de falha, $\mathbf{x}$ é o vetor de co-variáveis e $\boldsymbol{\beta}$ é o vetor de parâmetros a ser estimado.

Tomando $log\,[L(\boldsymbol{\beta})]$ e seu respectivo vetor de derivadas de primeira ordem igualado a zero, teremos um sistema a ser resolvido para obter os valores de $\boldsymbol{\beta}$ que maximizam $L(\boldsymbol{\beta})$:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = U(\boldsymbol{\beta}) = \sum_{i \in n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_i)} x_j exp\{\mathbf{x}_j'\widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} exp\{\mathbf{x}_j'\widehat{\boldsymbol{\beta}}\}} \right] = 0 \tag{2-53}$$

Como assume-se que os tempos observados são contínuos, supõe-se que os tempos observados de falha são distintos, ou seja, não existe empate entre esses tempos. Porém, nem sempre é possível que o tempo medido seja coletado em uma determinada escala de medida de tal forma que não haja empates. Em situações com presença de empates, deve-se usar alguma aproximação para a função de verossimilhança. Neste trabalho, duas aproximações serão utilizadas:

– A função de verossimilhança parcial modificada proposta por Breslow (21):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{exp\{\mathbf{s}_i'\boldsymbol{\beta}\}}{\left(\sum_{j \in R(t_i)} exp\{\mathbf{x}_j'\boldsymbol{\beta}\} \right)^{d_i}} \tag{2-54}$$

em que s_j é a soma dos vetores das p co-variáveis para os indivíduos que falharam no mesmo tempo t_i $(i = 1, ..., r)$ e d_i é o número de falhas em

t_i. Essa aproximação é adequada nos casos em que o número de empates não é grande.

– A função de verossimilhança parcial modificada proposta por Efron (22):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{exp\{\mathbf{s}_i'\boldsymbol{\beta}\}}{\prod_{l=1}^{d_i}\left(\sum_{j\in R(t_i)}\ exp\{\mathbf{x}_j'\boldsymbol{\beta}\} - \frac{l-1}{d_i}\sum_{j\in D_i}\ exp\{\mathbf{x}_j'\boldsymbol{\beta}\}\right)} \tag{2-55}$$

onde $R(t_i)$ é o conjunto de todos os indivíduos não-censurados no instante $t_{(j)}$. Quando o número de empates é grande, este método é mais acurado que o método anterior.

A estimação do vetor de parâmetros $\boldsymbol{\beta}$ utilizando as aproximações apresentadas é feita como em 2-53: aplica-se o logaritmo da função de verossimilhança parcial e em seguida tira-se a primeira derivada, iguala-se a zero e resolve-se o sistema de maneira que a estimativa de $\boldsymbol{\beta}$ maximize $L(\boldsymbol{\beta})$.

Os estimadores de máxima verossimilhança parcial são consistentes e assintoticamente normais sob certas condições de regularidade (Andersen (23)). Assim, pode-se realizar inferências acerca dos parâmetros, que serão apresentadas posteriormente.

Como visto anteriormente, a construção da função de verossimilhança parcial elimina o componente $h_0(t)$. Assim, não é possível estimá-lo através da mesma.

Para estimar $H_0(t)$ na presença de variáveis explicativas, Breslow (21) sugere a seguinte estimativa:

$$\widehat{H}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\in R_j} exp\{\mathbf{x}_l'\widehat{\boldsymbol{\beta}}\}} \tag{2-56}$$

onde d_j é o número de falhas em t_j.

2.5.2 Interpretação dos coeficientes estimados no modelo de riscos proporcionais de Cox

Nesse caso, a interpretação dos coeficientes estimados no modelo de riscos proporcionais de Cox é feita tomando-se a razão das taxas de falha (função de risco). Considere um modelo de riscos proporcionais de Cox de apenas uma variável dicotômica x dado por $exp\{\beta x\}$. Sendo assim, a razão das funções de risco entre um indivíduo que tem a característica indicada por x, $x = 1$ e outro não tem, $x = 0$, será dada por:

$$RR(x=1) = \frac{\widehat{h}(t|x=1,\widehat{\boldsymbol{\beta}})}{\widehat{h}(t|x=0,\widehat{\boldsymbol{\beta}})} = \frac{exp\{\widehat{\beta} * 1\}}{exp\{\widehat{\beta} * 0\}} = exp\{\widehat{\beta}\}. \tag{2-57}$$

Ou seja, o indivíduo com característica indicada por x terá risco de cancelar $exp\{\widehat{\beta}\}$ vezes o risco de um indivíduo que não tem essa característica.

No caso da co-variável ser categórica com m níveis, a interpretação pela razão dos ricos também poderá ser feita, de modo que as estimativas serão comparadas com o grupo do m-ésimo nível de referência.

2.6
Testes relacionados ao vetor de parâmetros estimados nos modelos

Depois da estimação do vetor de parâmetros $\boldsymbol{\beta}$, é necessário realizar testes de hipóteses. Os seguintes testes relacionados ao vetor de parâmetros do modelo podem ser realizados, tanto nos modelos de regressão paramétricos como nos modelos de riscos proporcionais de Cox:

1. Teste da Razão de Verossimilhança

   Estatística de teste sob $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$:

   $$TRV = -2log\left[\frac{L\left(\widehat{\boldsymbol{\beta_0}}\right)}{L\left(\widehat{\boldsymbol{\beta}}\right)}\right] = 2\left[logL\left(\widehat{\boldsymbol{\beta}}\right) - logL\left(\widehat{\boldsymbol{\beta_0}}\right)\right]. \tag{2-58}$$

   tal que $\widehat{\boldsymbol{\beta}}$ é o vetor de parâmetros estimados do modelo irrestrito e $\widehat{\boldsymbol{\beta_0}}$ é o vetor de parâmetros estimados do modelo restrito.

   Sob H_0 TRV segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade. A um nível de significância α, a hipótese nula é rejeitada se $TRV > \chi^2_{p,1-\alpha}$ em grandes amostras.

2. Teste Escore

   Estatística de teste sob $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$:

   $$S = U'(\boldsymbol{\beta_0})\left[I(\boldsymbol{\beta_0})\right]^{-1} U(\boldsymbol{\beta_0}) \tag{2-59}$$

   onde $U(\boldsymbol{\beta_0})$ é a função escore $U(\boldsymbol{\beta}) = \dfrac{\partial logL(\beta)}{\partial \beta} \mid_{\beta=\beta_0}$, $I(\boldsymbol{\beta_0})$ é a matriz de variância-covariância tal que $I(\boldsymbol{\beta_0}) = -E\left(\dfrac{\partial^2 logL(\beta)}{\partial\beta\partial\beta'}\right) \mid_{\beta=\beta_0}$. Sob H_0 S segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade. A hipótese nula é rejeitada se $S > \chi^2_{p,1-\alpha}$ a um nível de significância α.

3. Teste de Wald

   Estatística de teste sob $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$:

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})' I(\boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{2-60}$$

   em que $I(\boldsymbol{\beta_0})$ é a matriz descrita no teste Escore. A um nível de significância α, rejeita-se a hipótese nula se $W > \chi^2_{p,1-\alpha}$.

Na prática, os três testes descritos anteriormente são assintoticamente equivalentes em grandes amostras, sendo o teste de Wald o mais utilizado devido a sua simplicidade.

2.7
Adequação e validação dos modelos ajustados

Na literatura de Análise de Sobrevivência, a adequação do modelo é feita basicamente através de técnicas gráficas utilizando diferentes resíduos propostos ao longo do tempo. A seguir apresentaremos os principais tipos de resíduos discutidos na literatura.

1. Resíduos de Cox-Snell

   Os resíduos de Cox-Snell (24) são definidos por:

$$\widehat{e_i} = \widehat{H}(t_i|\boldsymbol{x}_i) = \widehat{H}_0(t_i) exp\{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\} \tag{2-61}$$

   onde $i = 1, ..., n$ é o i-ésimo indivíduo, $\widehat{\beta}$ é o vetor de parâmetros estimados pelo função de máxima verossimilhança e $\widehat{H}_0(t)$ estimado por 2-56. Se o modelo for adequado, então os resíduos $\widehat{e_i}$ devem se comportar como uma amostra censurada seguindo distribuição exponencial. Para isso, $\widehat{e_i}$ deve ter seu correspondente t_i como uma observação censurada. Analisando graficamente, $\widehat{S}(\widehat{e_i})$ *versus* $log(\widehat{e_i})$ devem ser aproximadamente uma reta com inclinação igual a -1. Também se pode fazer o gráfico de $\widehat{H}(\widehat{e_i})$ *versus* $\widehat{e_i}$ e este seve ser aproximadamente uma reta com inclinação igual a 1.

2. Resíduos de Schoenfeld

   Os resíduos de Schoenfeld (25) são analisados para averiguar se a suposição de riscos proporcionais é violada no modelo de regressão de Cox.

   Esses resíduos são definidos por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} exp\{\boldsymbol{x'_j}\widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} exp\{\boldsymbol{x'_j}\widehat{\boldsymbol{\beta}}\}} \tag{2-62}$$

em que $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})'$ é o vetor de co-variáveis do i-ésimo indivíduo condicionado à falha, $\boldsymbol{r}_i = (r_{1i}, r_{2i}, ..., r_{pi})$ é o vetor de resíduos de Schoenfeld e $q = 1, ..., p$ e p o número de co-variáveis.

O conjunto dos resíduos de Schoenfeld formam uma matriz $d \times p$, onde d é o número de falhas.

A forma padronizada dos resíduos de Schoenfeld para a avaliação da suposição de riscos proporcionais no modelo de regressão de Cox é dada por:

$$\boldsymbol{s^*}_i = \left[I(\widehat{\boldsymbol{\beta}})\right]^{-1} \times \boldsymbol{r}_i \tag{2-63}$$

tal que $I(\widehat{\boldsymbol{\beta}})$ é a matriz de informação observada.

Nesse trabalho, serão calculados os coeficientes de correlação de Pearson dos resíduos de Schoenfeld com o tempo para avalizar se a suposição de riscos proporcionais no modelo de Cox é válida.

Outras duas variações do resíduo de Cox-Snell são utilizadas na literatura de análise de sobrevivência: resíduos martingal e resíduos deviance, propostos por Therneau (26). O primeiro é usado para detectar a melhor forma funcional para uma dada variável e o segundo é usado para detecção de *outliers*. Ambos resíduos são utilizados em gráficos de pontos. Por isso, deve-se ser cauteloso quando usá-los. No caso de grandes amostras, como nesse estudo, a utilização deles é duvidosa.

2.7.1
Tabela de classificação e curva ROC

Como o tamanho da amostra é grande, é possível verificar o desempenho dos modelos através da *validação cruzada* (Kohavi (27)). Essa técnica consiste em avaliar o capacidade de generalização do modelo a partir de um conjunto de dados que não foi utilizado na estimação do modelo. Existem alguns métodos diferentes de validação cruzada, e o que será utilizado nessa dissertação será o método *holdout*. Este método consiste em dividir o banco de dados em duas partes: uma para estimação e outra para validação. Aqui, 70% dos dados (48278) foram utilizados na estimação dos modelos e os 30% restantes (20690) utilizados na validação. O método *holdout* é ideal quando o tamanho da amostra em questão é grande.

Em caso de pequenas amostras, outros métodos são indicados, tais como os métodos *k-fold* e *leave-one-out*. O método *k-fold* consiste em dividir a amostra total em k sub-amostras. Dessas, as *k-1* formam a amostra de estimação dos parâmetros e a amostra restante é utilizada como amostra de validação. Calcula-se a acurácia do modelo por algum métrica escolhida (por exemplo, a média entre a diferença entre o valor predito e o valor real) dentro da amostra de validação. Repete-se esses passos com todas as combinações possíveis entre as k sub-amostras que formam os conjuntos de estimação e validação. A medida final de acurácia do modelo é calculada utilizando-se todas as medidas de erros encontradas. O método *leave-one-out* de validação utiliza apenas uma única observação da amostra como validação e as demais são utilizadas para a estimação dos parâmetros.

A validação cruzada permite avaliar se os escores produzidos pelos modelos estimados estão bem correlacionados com a probabilidade/risco de o cliente cancelar o plano de previdência. O primeiro método estatístico de validação do modelo estimado proposto é o cálculo da tabela de classificação (Hosmer (28)). Para simplificar e explicar como ela é calculada, daremos um exemplo que é o próprio caso desta dissertação. Na amostra de validação desse trabalho 9634 clientes cancelaram o plano. No modelo paramétrico, a estimativa dada é o tempo mediano de cada indivíduo. Então, nesse caso, o escore é o tempo mediano e quanto menor ele for, "pior"será o cliente. Ou seja, os 9634 clientes com menores estimativas do tempo mediano serão classificados como "maus"clientes e os demais classificados como "bons". No modelo de riscos proporcionais de Cox, a estimativa dada pelo modelo é a função de risco estimada de o cliente cancelar o plano dado um conjunto de variáveis explicativas. Assim sendo, os 9634 clientes com maiores riscos estimados serão os clientes classificados como maus e os demais classificados como bons.

Posteriormente, verifica-se quantas classificações foram feitas corretamente e quantos classificadas erradas. Assim, pode-se formar a tabela de classificação na forma da Tabela (2.1):

	Cancelados previstos	Ativos previstos	Total
Cancelados reais	c_c	c_a	c
Ativos reais	a_c	a_a	a
Total	C	A	n

Tabela 2.1: Tabela de classificação.

Em uma situação ideal, c_a e a_c tem valores iguais a zero.

A partir da tabela de classificação pode-se calcular duas importantes medidas *sensibilidade* e *especificidade*que são, respectivamente, a probabilidade de um cliente cancelar (classificado como mau cliente) dado que ele realmente

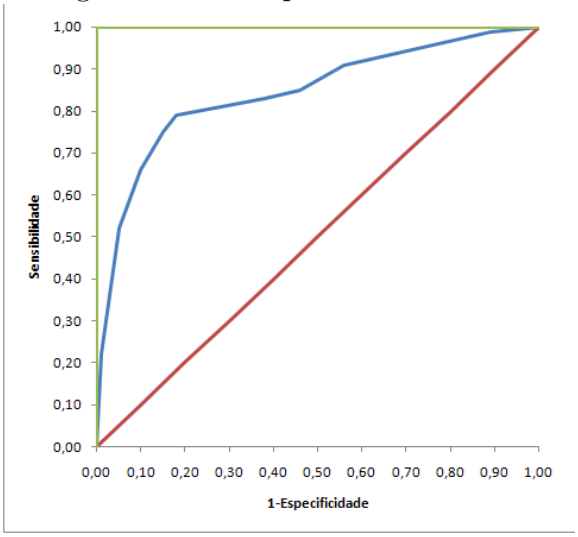
cancelou e a probabilidade de um cliente manter seu plano ativo (classificado como bom cliente) dado que ele realmente permaneceu ativo. Essas probabilidade são calculadas como:

$$\begin{aligned} Sensibilidade &= \frac{c_c}{C} \\ Especificidade &= \frac{a_a}{A}. \end{aligned} \tag{2-64}$$

Outra importante medida que pode ser calculada é a *acurácia* do modelo, dada pela capacidade de acertos total.

A segunda ferramenta proposta nessa dissertação para avaliar a capacidade de generalização dos modelos é a curva ROC (*Receiver Operating Characteristic*). Esse gráfico consiste em diversos pares de valores para diversos pontos de cortes de (1-especificidade, sensibilidade). A Figura (2.2) apresenta três exemplos de curvas ROC que um modelo pode gerar. A reta vermelha indica que o modelo classifica os clientes aleatoriamente. Quanto mais próxima a curva estiver do canto superior esquerdo da área do gráfico, melhor será o modelo. No caso da Figura (2.2), o melhor modelo é aquele com curva verde e o modelo com curva azul é considera bom.

Figura 2.2: Exemplo de curva ROC



3
Dados

Os dados utilizados nesse estudo são provenientes de uma seguradora brasileira. A base de dados original é constituída de registros referentes ao histórico de movimentações financeiras durante um período de seis anos e meio. Essas transações podem ser vistas sob duas perspectivas: proposta ou participante. A proposta é a unidade básica que diz respeito sobre as características do produto vendido, tais como: tipo de plano, forma de pagamento, etc. Já o participante reúne atributos do contratante: idade, sexo, estado civil, etc. As variáveis serão descritas detalhadamente mais à frente.

O banco de dados original precisou ser arduamente trabalhado antes que qualquer análise pudesse ser feita. Entre alguns dos problemas detectados, pode-se citar a presença de múltiplos registros, que foram excluídos. Também havia a presença de registros mal preenchidos, incompletos e incoerentes que precisaram ser excluídos do banco de dados. Alguns participantes tinham registro de movimentação financeira com data antes da própria venda do plano, o que também foi tratado como incoerência e deletado do banco de dados. Depois da limpeza no banco de dados, foi feita a amostragem aleatória de 70 mil participantes utilizada nessa dissertação.

Na amostra em estudo, aproximadamente 72% dos participantes possuem apenas uma proposta durante o período de observação, o que é equivalente a olharmos o banco de dados sob a unidade proposta ou sob a unidade participante. Para este estudo, optou-se por trabalhar com a unidade participante por ser mais interessante comercialmente. Assim, pode-se analisar o tempo de relacionamento entre o cliente e a empresa e quais são os fatores de risco que levam o participante a sair da instituição financeira.

Se o participante teve mais de um produto durante o período de observação, leva-se em conta as características da proposta mais recente. O conjunto de características da proposta mais recente será associado ao cliente. Os atributos referentes ao participantes são invariantes no tempo de observação.

O tempo de duração do participante na seguradora é medido em meses desde a data da venda da proposta até o seu cancelamento. Para participantes com mais de uma proposta, o seu tempo de duração será o intervalo de tempo

entre a sua primeira compra até o cancelamento da proposta mais recente.

O período de observação teve início em 3 de janeiro de 2005 e final em 14 de agosto de 2011. Os contratos podem ter sido iniciados e cancelados em qualquer data nesse intervalo de tempo. O participante pode cancelar sua proposta por iniciativa própria, encerrando assim seu relacionamento com a empresa. A proposta também pode ter sido encerrada por outras razões: quando o participante entra em gozo do benefício a proposta é naturalmente encerrada; quando há falecimento; ou quando a seguradora decide cancelar por razões diversas, como um problema no processo de venda. Além das propostas que não foram encerradas, somente propostas canceladas por decisão do próprio participante foram consideradas, visto que o interesse nesse estudo é analisar o comportamento voluntário do cliente. Então, quando neste texto for mencionado cancelamento, entenda-se por cancelamento por vontade do participante.

Contextualizando os dados à análise de sobrevivência, a variável resposta é o tempo até o cancelamento da proposta, sendo assim o evento de interesse (falha) o cancelamento. A presença de dados censurados se dá por aquelas propostas que não foram canceladas até o final do período de observação. Ou seja, esses dados são censurados à direita.

A Tabela (3.1) descreve as variáveis utilizadas nesse estudo.

Variável	Descrição
duracao	Tempo até o cancelamento (em meses)
duracao_anos	Tempo até o cancelamento (em anos)
cancelado	Variável indicadora: 1, se falha; 0, se censura
estado_civil	Variável categórica: fatores de 1 a 4
faixa_etaria	Variável categórica: fatores de 1 a 7
sexo	Variável indicadora: 1, se masculino; 0 se feminino
vgbl	Variável indicadora: 1, se VGBL; 0 se PGBL
tipo_pagamento	Variável categórica: 1-3
forma_pagamento	Variável indicadora: 1, se parcela única; 0, se mensal
fez_aporte	Variável indicadora: 1, se fez algum aporte; 0 c.c.
faixa_contribuicao	Variável categórica: 1-8

Tabela 3.1: Variáveis presentes no banco de dados da seguradora.

A variável *duracao* é o tempo de duração do participante entre a sua primeira compra até o dia do cancelamento da sua proposta mais recente. Caso a proposta seja censurada, a duração do participante será o tempo entre a sua primeira compra até o último dia de observação do estudo. A variável *duracao_anos* é a duração medida em anos: dividiu-se por 365 a variável *duracao*. Esta variável foi criada para facilitar a interpretação dos resultados.

A variável *estado_civil* indica o estado civil do participante. As categorias que podem ser atribuídas a ela são: 1, se casado; 2 se divorciado; 3 se solteiro

ou 4 se viúvo.

Para a variável *faixa_etaria*, indicadora da faixa etária do participante no dia da sua primeira aquisição da proposta, as possíveis categorias são: 1, se faixa etária é de 0 a 19 anos; 2, se a faixa etária é de 20 a 24 anos; 3, se a faixa etária é de 25 a 29 anos; 4, se a faixa etária é de 30 a 49 anos; 5, se a faixa etária é de 50 a 59 anos; 6, se a faixa etária é de 60 a 64 anos; 7, se a faixa etária é de mais de 65 anos.

O tipo de plano adquirido pelo participante é indicado pelo variável *vgbl*: se o plano for VGBL, a variável será igual a 1. Se o plano for PGBL, a variável será igual a 0.

Quando o participante adquire um plano de previdência privada, ele pode optar por um dos três tipos de pagamento disponíveis, discriminados pela variável *tipo_pagamento*: se o tipo de pagamento for débito em conta corrente, então *tipo_pagamento* será igual a 1; se for em carnê, igual a 2 e se for débito em poupança, igual a 3.

O participante pode optar em fazer o pagamento da contribuição de forma única ou mensal. Essa carecterística é dada pela variável dicotômica *forma_pagamento*: se for igual a 1, então o pagamento é feito por parcela única; se for igual a 0, então o pagamento é mensal.

A variável *fez_aporte* quando igual a 1 indica que, durante o período de observação, o participante já fez algum aporte extraordinário. Ou seja, além da contribuição, ele depositou uma quantia extra no fundo de previdência.

A variável *faixa_contribuicao* tem como referência o pagamento mensal. Se o participante opta por fazer o pagamento de forma única (uma vez ao ano), então o seu valor de contribuição será divido por 12. As faixas de contribuição podem ser: 1, se a contribuição estiver até R$99,00; 2, se a contribuição estiver entre R$100,00 e R$399,00; 3, se a contribuição estiver entre R$400,00 e R$899,00; 4, se a contribuição estiver entre R$900,00 e R$1.599,00; 5, se a contribuição estiver entre R$1.600,00 a R$2.499,00; 6, se a contribuição estiver entre R$2.500,00 a R$5.624,00; 7, se a contribuição estiver entre R$5.625,00 a R$9.999,00; 8, se a contribuição for maior que R$10.000,00.

Foi realizada uma amostragem aleatória simples para a obtenção dos resultados. Setenta mil participantes foram sorteados aleatoriamente, sob as restrições já discutidas. Desse total, 1,5% dos participantes tinham valor de contribuição igual a zero, que foram considerados uma contaminação na amostra, pois indicam uma falsa persistência: a proposta está ativa, mas o participante em si não está ativamente contribuindo. No total são 68.968 observações, sendo 70% (48.278) delas usadas para estimação e 30% (20.690) separadas para validação dos modelos a serem ajustados.

Estado civil	Nº de participantes	%
Casado	36990	53,6%
Solteiro	25946	37,6%
Viúvo	3177	4,6%
Divorciado	2855	4,1%

Tabela 3.2: Distribuição dos participantes por estado civil.

Faixa etária	Nº de participantes	%
0 a 19 anos	3009	4,4%
20 a 24 anos	5760	8,4%
25 a 29 anos	9007	13,1%
30 a 49 anos	30961	44,9%
50 a 59 anos	10092	14,6%
60 a 64 anos	3645	5,3%
65 anos ou mais	6494	9,4%

Tabela 3.3: Distribuição dos participantes por faixa etária.

Sexo	Nº de participantes	%
Masculino	36981	53,6%
Feminino	31987	46,4%

Tabela 3.4: Distribuição dos participantes por sexo.

Uma breve descrição das distribuições dos participantes pelas diversas variáveis da Tabela (3.1) é apresentada a seguir.

A Tabela (3.2) mostra que mais da metade dos segurados da amostra são casados.

A maior parte das pessoas que possuem um plano de previdência privada têm de 25 a 59 anos, que presumidamente é a faixa de indivíduos assalariados, como pode ser visto na Tabela (3.3).

A distribuição de homens e mulheres, Tabela (3.4), é equilibrada, sendo o sexo masculino ligeiramente mais representativo.

Os planos de previdência privada são majoritariamente do tipo VGBL, como visto na Tabela (3.5).

Quase 80% dos segurados optam por pagar suas contribuições por débito em conta corrente, Tabela (3.6).

A distribuição de propostas por forma de pagamento é equilibrada: metade dos segurados pagam suas contribuições de forma mensal e a outra metade de forma única. Esses números encontram-se na Tabela (3.7).

Apenas 25% dos participantes fizeram algum aporte extraordinário durante o período de estudo, apresentado na Tabela (3.8).

A maioria dos participantes (mais de 70%) contribui com no máximo R$899,00 mensais,. A distribuição dos clientes por faixa de contribuição encontra-se na Tabela (3.9).

Tipo do plano	Nº de participantes	%
VGBL	63545	92,1%
PGBL	5423	7,9%

Tabela 3.5: Distribuição dos participantes por tipo de plano.

Tipo de pagamento	Nº de participantes	%
Débito em conta corrente	54797	79,5%
Débito em poupança	10072	14,6%
Carnê	4099	5,9%

Tabela 3.6: Distribuição dos participantes por tipo de pagamento.

Forma de pagamento	Nº de participantes	%
Mensal	34896	50,6%
Única	34072	49,4%

Tabela 3.7: Distribuição dos participantes por forma de pagamento.

Já fez aporte	Nº de participantes	%
Não	51703	75,0%
Sim	17265	25,0%

Tabela 3.8: Distribuição dos participantes por presença de aporte.

Faixa de contribuição	Nº de participantes	%
até R$99,00	13944	20,2%
de R$ 100,00 a R$399,00	22511	32,6%
de R$ 400,00 a R$899,00	13010	18,9%
de R$ 900,00 a R$1.599,00	4954	7,2%
de R$ 1.600,00 a R$2.499,00	4004	5,8%
de R$ 2.500,00 a R$5.624,00	6266	9,1%
de R$ 5.625,00 a R$9.999,00	2147	3,1%
mais de R$10.000,00	2132	3,1%

Tabela 3.9: Distribuição dos participantes por faixa de contribuição.

4
Resultados

Neste capítulo serão apresentados os resultados obtidos pela modelagem estatística dos tempos de duração dos clientes em um plano de previdência privada. Para tal, foram utilizados os *softwares* estatísticos *Stata* e *R* [1].

Antes de fazer a estimação dos modelos, é necessária uma análise descritiva dos dados. Como a amostra é censurada, usaremos a estimativa da função de sobrevivência dada pelo estimador de Kaplan-Meier. Para verificar se há diferença entre as categorias de cada variável, utilizaremos o teste log-rank. Assim, em um primeiro momento, ganharemos conhecimento sobre o comportamento das variáveis e a modelagem poderá ser melhor conduzida.

A seguir serão apresentados as estimativas da função de sobrevivência para cada uma das variáveis explicativas descritas na Tabela (3.1). Duas medidas particularmente importantes para as seguradoras serão explicitadas a diante :a probabilidade de o cliente permanecer durante pelo menos um ano no plano de previdência e o tempo mediano, como dito no capítulo 2 na seção 2.2.1.

A Figura (4.1(a)) apresenta a estimativa da função de sobrevivência para a variável *estado_civil*. Como podemos ver, segurados com estado civil casado ou viúvo apresentam as maiores probabilidade de sobrevivência, ou seja, são os participantes mais persistentes. Depois de um ano, permaneceram ativos aproximadamente 78% dos participantes casados ou viúvos, enquanto 74,8% dos participantes divorciados e 75,8% dos solteiros. Os viúvos têm o maior tempo de duração mediano até o cancelamento, 3,86 anos, seguidos daqueles que são casados (tempo mediano de 3,64 anos). Os menores tempos medianos são dos solteiros e divorciados: 2,90 e 2,97 anos, respectivamente.

A estimativa da função de sobrevivência para a variável *faixa_etaria*, Figura (4.1(b)), mostra que pessoas com mais de 60 anos tem as maiores per-

[1] O *software R* foi utilizado para estimação das curvas de Kaplan-Meier com o pacote *survival* através da função *survfit*. No *Stata* foram estimados os modelos paramétrico e o modelo de riscos proporcionais de Cox utilizando-se as funções *streg* e *stcox*. A análise de resíduos foi feita exportando-se os dados do *Stata* para o *R* e as funções utilizadas foram programadas. A validação dos modelos estimados foi feita no *R*: a curva ROC foi calculada pelas funções *prediction* e performance e para o cálculo da tabela de classificação uma programação foi feita.

sistências, juntamente com participantes de até 19 anos de idade [2] Segurados de 20 a 29 anos têm as menores probabilidade de permanecerem no plano de previdência. Durante o primeiro ano, aproximadamente 19% participantes das faixas etárias de 0 a 19 anos, de 60 a 64 anos e de mais de 65 anos cancelaram seus planos de previdência. Entre as pessoas com idade entre 20 e 24 houve a menor persistência durante o primeiro ano: 29% desses planos foram cancelados. O tempo de permanência mediano até o cancelamento daqueles clientes que têm mais de 65 anos é 5,07 anos. Pessoas com até 19 anos possuem o segundo maior tempo de duração mediano, 4,27 anos. Os jovens com idade entre 20 e 24 anos têm o pior desempenho em relação ao tempo mediano: permanecem por 2,15 anos medianamente.

As curvas de sobrevivência estimadas para a variável *sexo* representadas na Figura (4.1(c)) mostram que não há diferença significativa entre os participantes do sexo masculino e feminino. Seguradas são apenas discretamente mais persistentes que os segurados. A probabilidade de participantes do sexo feminino continuarem com o seguro é 78,1%, enquanto participantes do sexo masculino tem 76,6%. O tempo de duração mediano das mulheres em um plano de previdência é ligeiramente maior, 3,49 anos, enquanto que homens têm o tempo de duração até o cancelamento de aproximadamente 3,21.

Pela Figura (4.1(d)) vê-se que planos PGBL são mais persistentes que os planos VGBL. Durante o primeiro ano, 11,4% dos planos PGBL e 23,5% dos planos VGBL são cancelados. Entre aqueles que têm um plano de previdência do tipo PGBL observados na amostra 57,3% permaneceram ativos no final do estudo, não sendo assim observado o tempo mediano de duração. Já aqueles detentores de um plano VGBL levam 3,21 anos para cancelar o plano, em relação ao tempo mediano.

As maiores persistências, Figura (4.1(h)), por tipo de pagamento são dos participantes que optam por pagar suas contribuições em débito em conta corrente. As probabilidades de sobrevivência dos outros dois tipo de pagamento, carnê e débito em poupança, se misturam no decorrer do tempo. Aproximadamente 78,7% dos participantes que tem o tipo de pagamento débito em conta corrente, 67,4% dos que tem tipo de pagamento em carnê e 73,9% dos que tem tipo de pagamento débito em poupança continuam sob risco depois do primeiro ano. Em relação ao tempo mediano de até o cancelamento, os clientes que pagam suas contribuições em débito em conta corrente têm 3,64 anos de tempo mediano até o cancelamento, 2,23 anos para aqueles que pagam

[2]Vale deixar claro que no banco de dados as variáveis referentes ao participante são associados ao beneficiário do plano. Então, se, por exemplo, um pai resolve criar um plano de previdência privada para seu filho menor de idade, as características anotadas no banco de dados serão da criança/adolescente.

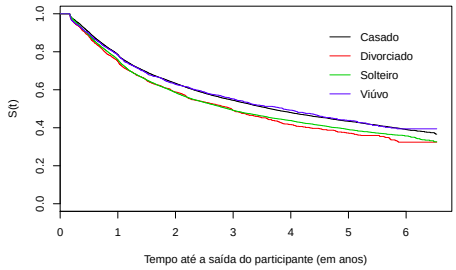
em carnê e 2,45 anos para os que pagam em débito em poupança.

Os planos com pagamentos mensais são sutilmente mais persistentes que planos com pagamento único, como pode ser visto na Figura (4.1(f)). Depois de um ano, 80% das propostas com pagamentos mensais continuam ativas e 74,5% das propostas com pagamentos únicos também permanecem ativas. Clientes que pagam mensalmente têm o tempo mediano até o cancelamento um pouco maior que aqueles que pagam em parcela única: 3,55 anos contra 3,11 anos, respectivamente.

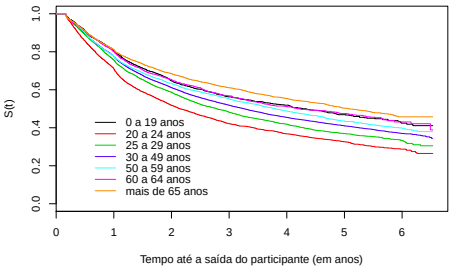
Imagina-se que quem faz um aporte extraordinário, ou seja, uma contribuição extra, tenha um comprometimento maior com o plano de previdência privada. Isso é confirmado através da Figura (4.1(g)): quem fez algum aporte extraordinário durante o período observado tem maior persistência que aquelas que não fizeram. Durante o primeiro ano, 26,4% dos participantes que não fizeram aporte extraordinário cancelaram seus planos e apenas 12,3% dos segurados que fizeram algum aporte extraordinário desistiram de seus planos no mesmo período. O tempo mediano de duração de um cliente que fez um aporte extraordinário (6,16 anos) é mais de duas vezes maior que aqueles que não fizeram (2,57 anos).

Quanto à faixa de contribuição, Figura (4.1(h)), o grupo com menor persistência é aquele que pagam mensalidades entre R$400,00 e R$899,00 e com maior persistência é o grupo que faz as maiores contribuições, acima de R$10.000,00. Dos participantes que contribuem com parcelas entre R$400,00 e R$899,00, 73% continuam com seus planos de previdência depois de um ano. 86,2% das pessoas que contribuem com mais de R$10.000,00 permanecem ativas depois do primeiro ano de plano. Dentre aqueles que contribuem com mais de R$10.000,00 53,5% permaneceram ativos no final do período de estudo, não sendo possível calcular o tempo mediano. De todo forma, este fato implica que os participantes que fazem as maiores contribuições tem os maiores tempo. O menor tempo mediano de duração de um cliente é daqueles que contribuem com mensalidades entre R$400,00 e R$899,00: 2,68 anos.

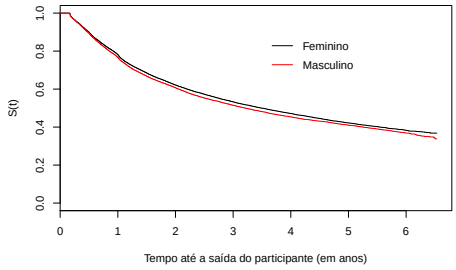
O teste *logrank* e Wilcoxon foram feitos para verificar se as curvas de sobrevivência em cada variável diferente entre si. Os resultados estão nas Tabelas (4.1) e (4.2). Com base nesses resultados, tomando-se um nível de significância de 5%, pode-se rejeitar a hipótese nula para todas as variáveis. Sendo assim, todas as variáveis devem ser incluídas na modelagem estatística.



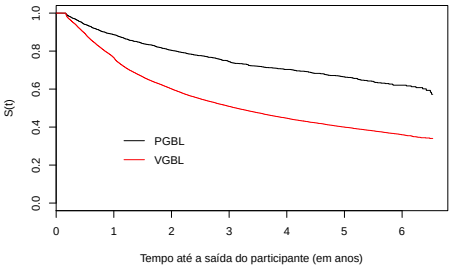
4.1(a): Curva de sobrevivência estimada por Kaplan-Meier por estado civil.



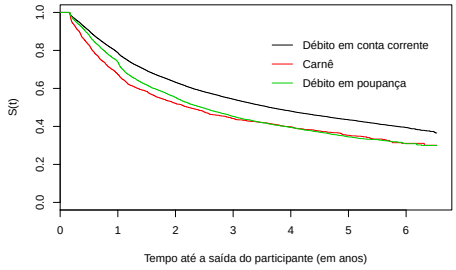
4.1(b): Curva de sobrevivência estimada por Kaplan-Meier por faixa etária.



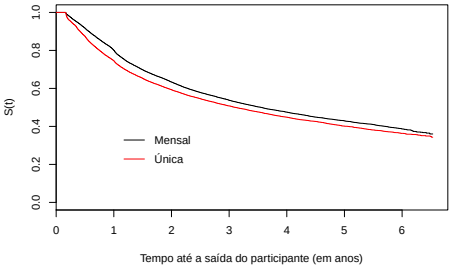
4.1(c): Curva de sobrevivência estimada por Kaplan-Meier por sexo.



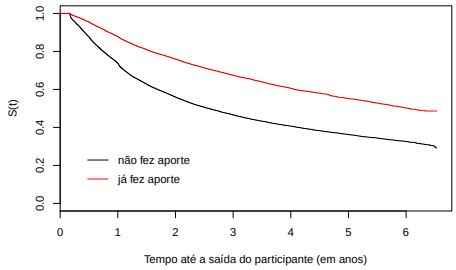
4.1(d): Curva de sobrevivência estimada por Kaplan-Meier por tipo plano.



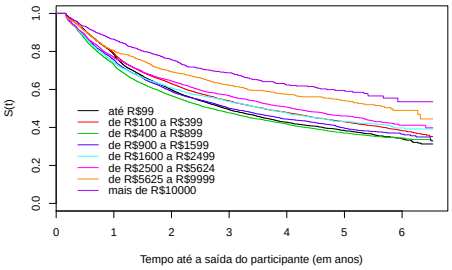
4.1(e): Curva de sobrevivência estimada por Kaplan-Meier por tipo de pagamento.



4.1(f): Curva de sobrevivência estimada por Kaplan-Meier por forma de pagamento.



4.1(g): Curva de sobrevivência estimada por Kaplan-Meier por presença de aporte extraordinário.



4.1(h): Curva de sobrevivência estimada por Kaplan-Meier por faixa de contribuição.

Figura 4.1: Curvas de sobrevivência estimada por Kaplan-Meier por variável explicativa.

Variável	Estatística de teste	Graus de liberdade	p-valor
vgbl	477	1	0,0000
forma_pagamento	138	1	0,0000
tipo_pagamento	346	2	0,0000
sexo	15	1	0,0001
estado_civil	107	3	0,0000
contribuicao	302	7	0,0000
faixa_etaria	342	6	0,0000
fez_aporte	1576	1	0,0000

Tabela 4.1: Resultado do teste *logrank* para as variáveis explicativas.

Variável	Estatística de teste	Graus de liberdade	p-valor
vgbl	502	1	0,0000
forma_pagamento	86,3	1	0,0000
tipo_pagamento	307	2	0,0000
sexo	13,7	1	0,0002
estado_civil	105	3	0,0000
contribuicao	302	7	0,0000
faixa_etaria	363	6	0,0000
fez_aporte	1445	1	0,0000

Tabela 4.2: Resultado do teste Wilcoxon para as variáveis explicativas.

4.1 Estimação do modelo paramétrico

Escolhidas as variáveis candidatas a regressoras, o primeiro passo na especificação do modelo foi fazer a seleção entre essas candidatas daquelas mais relevantes do ponto de vista da significância estatística utilizando o método de seleção de variáveis *stepwise-forward* (com nível de significância igual a 10%).

Para a especificação do modelo paramétrico, nesse primeiro passo optou-se pela distribuição gama generalizada para a escolha das variáveis de efeitos diretos, visto que esta distribuição inclui como casos especiais as distribuições exponencial, Weibull e log-normal. Depois testou-se o efeito das interações duplas manualmente considerando apenas as variáveis seleciondas pelo método *stepwise-forward*. Por fim, com o conjunto final de variáveis de efeitos diretos e interações duplas, ajustou-se modelos que são casos particulares da distribuição gama generalizada: exponencial, log-normal e Weibul, além da distribuição log-logística. A estimação é feita com o modelo de tempo de vida acelerado.

A Tabela (4.3) apresenta o modelo gama generalizada selecionado com apenas efeitos diretos. Nela são apresentados os parâmetros estimados, o desvio padrão das estimativas, o valor da estatística de teste de Wald e seu p-valor.

O modelo irrestrito, com os efeitos diretos contidos na Tabela (4.3) mais todas as combinações possíveis de interações duplas dentre esses efeitos foi ajustado. Obviamente, nem todas as interações duplas foram estatisticamente significantes. Uma vez que a amostra utilizada para estimação dos modelos

Variável	Estimativa	Desvio Padrão	Estatística de teste	p-valor
constante	1,536	0,057	26,91	0,000
estado_civil_2	-0,179	0,036	-4,91	0,000
estado_civil_3	-0,070	0,018	-3,95	0,000
estado_civil_4	-0,085	0,037	-2,29	0,022
sexo_m	-0,113	0,015	-7,60	0,000
faixa_etaria_2	-0,420	0,043	-9,83	0,000
faixa_etaria_3	-0,265	0,041	-6,52	0,000
faixa_etaria_4	-0,152	0,039	-3,93	0,000
faixa_etaria_5	-0,057	0,043	-1,33	0,183
faixa_etaria_6	-0,021	0,050	-0,41	0,680
faixa_etaria_7	0,132	0,047	2,83	0,005
fez_aporte	0,726	0,017	42,63	0,000
vgbl	-0,597	0,031	-19,32	0,000
tipo_pagamento_2	-0,360	0,031	-11,81	0,000
tipo_pagamento_3	-0,148	0,021	-7,17	0,000
forma_pagamento	-0,311	0,021	-15,01	0,000
contribuicao_2	-0,018	0,021	-0,86	0,389
contribuicao_3	-0,110	0,026	-4,23	0,000
contribuicao_4	-0,056	0,034	-1,64	0,101
contribuicao_5	0,015	0,037	0,40	0,691
contribuicao_6	0,074	0,033	2,23	0,026
contribuicao_7	0,175	0,047	3,68	0,000
contribuicao_8	0,341	0,050	6,85	0,000
parâmetro de forma	0,402	0,005	79,650	0,000
parâmetro de escala	-0,817	0,030	-27,040	0,000

Tabela 4.3: Estimativas dos parâmetros do modelo de regressão gama generalizado ajustado para os dados em estudo.

estatísticos é grande, parâmetros estatisticamente significantes são facilmente obtidos. Desta forma, optou-se por um nível de significância restritivo para as interações duplas: somente estimativas de parâmetros de interação com p-valores menores ou iguais a 10^{-3} foram consideradas relevantes e incluídas no modelo final, Tabela (4.4).

A Tabela (4.5) mostra os critérios *AIC* e *BIC* do modelo de efeitos fixos, que está na Tabela (4.3) e do modelo final com as interações duplas, que é o modelo que minimiza os valores de *AIC* e *BIC* e maximiza a log-verossimilhança.

Com as variáveis e interações selecionadas no modelo final, ajustou-se os modelos que são casos particulares da distribuição gama generalizada além da distribuição log-normal. A log-verossimilhança e os critérios *AIC* e *BIC* são apresentados na Tabela (4.6).

Assim sendo, o modelo paramétrico escolhido é o modelo de regressão gama generalizado com os parâmetros estimados na Tabela (4.4).

O passo seguinte foi verificar a adequação do modelo através da análise de resíduos. Os primeiros resíduos analisados são os resíduos de *Cox-Snell*.

Como dito anteriormente, se o modelo estiver bem ajustado, então os

Parâmetro	Variável	Estimativa	D.P.	Est. de teste	p-valor
$\widehat{\beta}_0$	constante	1,227	0,005	78,14	0,000
$\widehat{\beta}_1$	estado_civil_2	-0,167	0,036	-4,62	0,000
$\widehat{\beta}_2$	estado_civil_3	-0,096	0,030	-3,17	0,002
$\widehat{\beta}_3$	estado_civil_4	-0,055	0,037	-1,48	0,139
$\widehat{\beta}_4$	sexo_m	-0,115	0,015	-7,78	0,000
$\widehat{\beta}_5$	faixa_etaria_2	-0,401	0,042	-9,44	0,000
$\widehat{\beta}_6$	faixa_etaria_3	-0,390	0,055	-7,07	0,000
$\widehat{\beta}_7$	faixa_etaria_4	-0,382	0,046	-8,34	0,000
$\widehat{\beta}_8$	faixa_etaria_5	-0,069	0,046	-1,48	0,140
$\widehat{\beta}_9$	faixa_etaria_6	-0,034	0,054	-0,63	0,530
$\widehat{\beta}_{10}$	faixa_etaria_7	0,102	0,051	2,01	0,045
$\widehat{\beta}_{11}$	fez_aporte	1,118	0,072	15,43	0,000
$\widehat{\beta}_{12}$	vgbl	-0,973	0,071	-13,71	0,000
$\widehat{\beta}_{13}$	tipo_pagamento_2	-0,185	0,049	-3,82	0,000
$\widehat{\beta}_{14}$	tipo_pagamento3	-0,135	0,020	-6,6	0,000
$\widehat{\beta}_{15}$	forma_pagamento	0,210	0,042	5,02	0,000
$\widehat{\beta}_{16}$	contribuicao_2	0,018	0,020	0,86	0,388
$\widehat{\beta}_{17}$	contribuicao_3	-0,059	0,026	-2,26	0,024
$\widehat{\beta}_{18}$	contribuicao_4	-0,005	0,034	-0,15	0,880
$\widehat{\beta}_{19}$	contribuicao_5	0,075	0,037	2,02	0,043
$\widehat{\beta}_20$	contribuicao_6	0,136	0,033	4,1	0,000
$\widehat{\beta}_{21}$	contribuicao_7	0,236	0,047	4,99	0,000
$\widehat{\beta}_{22}$	contribuicao_8	0,396	0,050	7,99	0,000
$\widehat{\beta}_{23}$	estado_civil_3 $*$ faixa_etaria_4	0,129	0,035	3,68	0,000
$\widehat{\beta}_{24}$	estado_civil_3 $*$ forma_pagamento	-0,188	0,032	-5,90	0,000
$\widehat{\beta}_{25}$	estado_civil_3 $*$ tipo_pagamento_2	-0,272	0,062	-4,42	0,000
$\widehat{\beta}_{26}$	fez_aporte $*$ forma_pagamento	-0,373	0,034	-10,96	0,000
$\widehat{\beta}_{27}$	fez_aporte $*$ vgbl	0,637	0,074	8,64	0,000
$\widehat{\beta}_{28}$	forma_pagamento $*$ faixa_etaria_3	0,225	0,051	4,37	0,000
$\widehat{\beta}_{29}$	forma_pagamento $*$ faixa_etaria_4	0,280	0,034	8,14	0,000
$\widehat{\beta}_{30}$	vgbl $*$ faixa_etaria_3	-0,210	0,088	-2,38	0,017
$\widehat{\beta}_{31}$	vgbl $*$ forma_pagamento	0,514	0,066	7,76	0,000
$\widehat{\sigma}$	parâmetro de forma	0,394	0,005	78,14	0,000
$\widehat{\kappa}$	parâmetro de escala	-0,824	0,030	-27,6	0,000

Tabela 4.4: Estimativas dos parâmetros do modelo final de regressão gama generalizado ajustado para os dados em estudo.

	Log-verossimilhança	Nº parâmetros estimados	AIC	BIC
Modelo de efeitos diretos	-52962	25	105974	106194
Modelo final	-52756	33	105578	105782

Tabela 4.5: Critérios de escolha do modelo paramétrico gama generalizado.

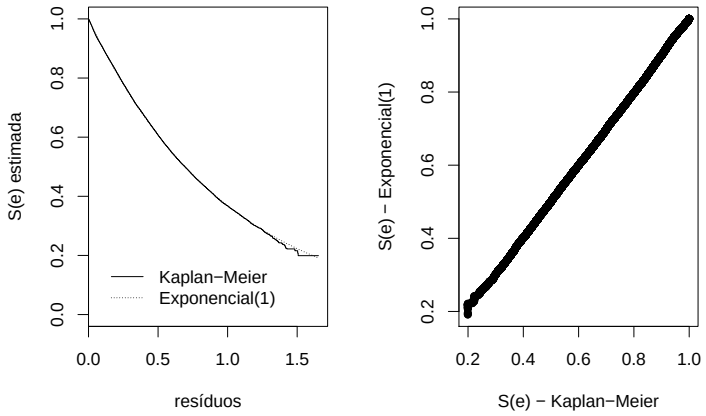
	Log-likelihood	Nº parâmetros estimados	AIC	BIC
Gama generalizada	-52756	33	105578	105778
Log-normal	-53153	32	106370	106684
Log-logística	-53764	32	107592	107940
Weibull	-54842	32	109748	110191
Exponencial	-54855	31	109772	110217

Tabela 4.6: Critérios de seleção do modelo paramétrico.

resíduos *Cox-Snell* serão provenientes de uma amostra aleatória censurada com distribuição exponencial com parâmetro igual a um. Como os resíduos são censurados, o estimador de Kaplan-Meier deve ser utilizado para estimar a função de sobrevivência dos resíduos.

A Figura (4.2) apresenta dois gráficos: o da esquerda apresenta as curvas estimadas por Kaplan-Meier e pela distribuição exponencial padrão e o da direita, o gráfico de pontos das estimativas das sobrevivências pelos dois métodos.

Figura 4.2: Curvas de sobrevivência dos resíduos de Cox-Snell estimadas por Kaplan-Meier e pelo modelo exponencial padrão (esquerda) e respectivas sobrevivências (direita).



Visualmente o modelo parece bem ajustado, visto que o os pontos do gráfico à direita formam uma reta e as curvas de sobrevivência estimadas estão muito próximas.

4.1.1 Interpretação dos coeficientes estimados

Neste trabalho foi realizada a interpretação dos tempos medianos proposta por Hosmer (18) para se ter conhecimento dos efeitos das co-variáveis sob o tempo de duração do participante.

Para cada interpretação, considere que as outras características são as mesmas entre os grupos que estão sendo comparados (*ceteris paribus*).

O tempo mediano das participantes divorciados é aproximadamente 15% menor que o tempo mediano dos casados. A diferença entre os tempos medianos dos participantes viúvos é mínima: 5% menor em relação aos casados. Em relação aos solteiros, é necessário que se considere as interações associadas a esta variável. A razão dos tempos medianos nesse caso será dado por:

$$RT_{0,5}(solteiro = 1) = exp\{-0,096 + 0,129 * faixa\_etaria\_4 \\ -0,188 * forma\_pagamento - 0,272 * tipo\_pagamento\_2\} \quad (4\text{-}1)$$

onde $faixa\_etaria\_4 = 1$ se o cliente tem idade entre 30 e 49 anos, $forma\_pagamento = 1$ se o cliente faz pagamentos em parcela única e $tipo\_pagamento\_2 = 1$ se o cliente paga suas contribuições em carnê. Então, se o participante solteiro tem todas essas características, seu tempo mediano é 30% menor que o tempo dos casados. Se, por exemplo, o cliente solteiro não tiver entre 30 e 49 anos, fizer pagamentos em carnê com parcela única, então ele terá o tempo mediano de duração aproximadamente 35% menor que o grupo de referência, os casados.

A diferença entre os tempo medianos de homens e mulheres é pequena: participantes do sexo feminino têm o tempo mediano 11% maior que participantes do sexo masculino.

Detentores de planos com idade entre 25 e 29 anos têm razão dos tempos medianos de duração até o cancelamento dependente das interações:

$$RT_{0,5}(faixa\_etaria\_3 = 1) = exp\{-0,390 + 0,225 * forma\_pagamento \\ -0,210 * vgbl\} \quad (4\text{-}2)$$

em que $forma\_pagamento = 1$ se o cliente faz pagamento em parcela única e $vgbl = 1$ se o plano for VGBL. Nesse caso, se o participante paga mensalmente e tem um plano VGBL, então seu tempo mediano é 55% menor que o grupo de referência entre as faixas etárias até 19 anos.

Aqueles que tem entre 30 e 49 anos também têm a razão dos tempos medianos dependente de interações, dado por:

$$RT_{0,5}(faixa\_etaria\_4 = 1) = exp\{-0,382 + 0,129 * solteiro + \\ 0,280 * forma\_pagamento\} \quad (4\text{-}3)$$

tal que se o cliente não for solteiro ($solteiro = 0$) e fizer pagamentos em parcela mensal ($forma\_pagamento = 0$), seu tempo mediano será 30% menor que aqueles que tem até 19 anos.

Aqueles que têm um maior comprometimento com o plano de previdência privada e fazem um aporte extraordinário tem o tempo mediano 5,8 vezes maior em relação aos que nunca fizeram esse tipo de transação financeira dado que o plano é VGBL e forma de pagamento mensal (razão dos tempos medianos

dada pela equação a seguir).

$$RT_{0,5}(fez\_aporte = 1) = exp\{1,118 + 0,637 * vgbl \\ -0,376 * forma\_pagamento\} \tag{4-4}$$

Se o cliente tiver um plano de previdência do tipo VGBL a razão dos tempos medianos será dada por:

$$RT_{0,5}(vgbl = 1) = exp\{-0,973 + 0,637 * fez\_aporte - 0,21 * faixa\_etaria\_3 + \\ 0,514 * forma\_pagamento\} \tag{4-5}$$

onde $fez\_aporte = 1$ se o cliente fez um aporte extraordinário durante o período de observação, $faixa\_etaria\_3 = 1$ se tiver entre 25 e 29 anos e $forma\_pagamento = 1$ se o pagamento for feito em parcela única. Se o cliente tiver entre 25 e 29 anos, pagar mensalmente suas contribuições e nunca tiver feito um aporte extraordinário, seu tempo mediano será 70% menor que aqueles detentores de um plano PGBL.

Quem faz as contribuições em parcela única tem razão dos tempos medianos dependente das interações :

$$RT_{0,5}(forma\_pagamento = 1) = exp\{0,21 - 0,188 * solteiro + \\ 0,225 * faixa\_etaria\_3 + 0,280 * faixa\_etaria\_4 + 0,514 * vgbl\}. \tag{4-6}$$

Assim, se o cliente tiver idade entre 30 e 49 anos, plano VGBL e não for solteiro, seu tempo mediano será 2,7 vezes o tempo mediano daqueles que pagam mensalmente.

Levando-se em conta somente o tipo de pagamento, aqueles que pagam em débito em conta corrente têm os maiores tempos mediano. Quem paga em débito em poupança têm o tempo mediano até o cancelamento apenas 17% menor que o grupo de referência (débito em conta corrente). Já os cliente que optam por pagar em carnê têm um decréscimo maior no tempo mediano em relação ao grupo de referência: 27%, se forem solteiros (dado pela equação abaixo).

$$RT_{0,5}(tipo\_pagamento\_2 = 1) = exp\{-0,185 - 0,272 * solteiro\} \tag{4-7}$$

Quanto às faixas de contribuição, o grupo tido como referência é formado

por aqueles que optaram a contribuir com até R$399,00 mensais. Todos que fazem contribuições maiores que R$2.499,00 tem tempos medianos maiores que o grupo de controle: os participantes que estão na faixa de contribuição de R$2.500,00 a R$5.624,00 tem tempos medianos apenas 15% maior; aqueles que contribuem com algum valor entre R$5.625,00 e R$9.999,00 tem um aumento de 27% no tempo mediano e os que contribuem com mais de R$10.000,00 tem o tempo mediano quase dobrado (aumento de 49%). Aqueles que fazem depósitos com valor entre R$400,00 e R$899,00 tem diminuição no tempo mediano de apenas 6% em relação ao grupo de referência.

4.2 Estimação do modelo de riscos proporcionais de Cox

Os passos feitos da estimação do modelo de riscos proporcionais são os mesmos que passos feitos na estimação do modelo paramétrico. Primeiramente, o método de seleção de variáveis *stepwise-forward* com nível de significância igual a 0,1 foi utilizado para a seleção de variáveis relevantes que produzem efeitos diretos.

Selecionadas as variáveis de efeitos diretos, testou-se os efeitos das interações duplas. Neste passo, os *softwares* utilizados apresentam uma limitação: no máximo 12 interações podem ser testadas. Assim, uma solução foi aproveitar o conjunto de interações duplas estatisticamente significativas encontrado no ajuste do modelo paramétrico. O mesmo critério foi usado: somente interações com p-valores menores ou iguais a 10^{-3} foram mantidos no modelo, já que, como dito anteriormente, as interações são facilmente estatisticamente significantes quando a amostra a ser utilizada na modelagem é grande.

As duas aproximações para a função de verossimilhança parcial propostas para o tratamento de empates foram testadas. A aproximação com a melhor performance foi a de Efron. Os resultados estão na Tabela (4.7).

Aproximação	Log-likelihood	Nº de parâmetros estimados	AIC	BIC
Efron	-225716	29	449171	449417
Breslow	-225726	29	449241	449487

Tabela 4.7: Critérios de seleção do modelo de riscos proporcionais.

Sendo assim, o modelo escolhido foi o modelo ajustado pela aproximação de Efron. Os parâmetros estimados estão na Tabela (4.8).

Calculou-se o coeficiente de correlação de Pearson ρ entre os resíduos padronizados de Schoenfeld e o tempo para cada co-variável para avaliar se a suposição de riscos proporcionais no modelo de regressão de Cox é violada. Como essa correlação tem valores próximos de zero em todos os casos, não há

Variável	Estimativa	RR	Erro-padrão	Est. de teste	p-valor
estado_civil_2	0,138	1,148	0,038	4,13	0,000
estado_civil_3	0,047	1,048	0,030	1,65	0,100
estado_civil_4	0,077	1,080	0,038	2,19	0,028
sexo_m	0,108	1,114	0,015	7,78	0,000
faixa_etaria_2	0,418	1,519	0,062	10,33	0,000
faixa_etaria_3	0,290	1,336	0,053	7,27	0,000
faixa_etaria_4	0,354	1,425	0,061	8,24	0,000
faixa_etaria_5	0,074	1,077	0,049	1,64	0,101
faixa_etaria_6	-0,007	0,993	0,053	-0,12	0,901
faixa_etaria_7	-0,152	0,859	0,043	-3,03	0,002
fez_aporte	-1,226	0,293	0,032	-11,13	0,000
vgbl	1,202	3,327	0,358	11,17	0,000
tipo_pagamento_2	0,162	1,176	0,052	3,69	0,000
tipo_pagamento_3	0,151	1,163	0,021	8,33	0,000
forma_pagamento	-0,031	0,970	0,038	-0,79	0,427
contribuicao_2	-0,053	0,948	0,018	-2,80	0,005
contribuicao_3	0,027	1,027	0,024	1,14	0,254
contribuicao_4	-0,040	0,961	0,029	-1,31	0,190
contribuicao_5	-0,152	0,859	0,029	-4,44	0,000
contribuicao_6	-0,229	0,795	0,025	-7,43	0,000
contribuicao_7	-0,353	0,703	0,033	-7,45	0,000
contribuicao_8	-0,547	0,579	0,032	-9,88	0,000
estado_civil_3 $*$ faixa_etaria_4	-0,109	0,897	0,029	-3,32	0,001
estado_civil_3 $*$ tipo_pagamento_2	0,214	1,239	0,068	3,91	0,000
estado_civil_3 $*$ forma_pagamento	0,161	1,174	0,034	5,47	0,000
forma_pagamento $*$ faixa_etaria_4	-0,188	0,829	0,025	-6,24	0,000
fez_aporte $*$ vgbl	-0,742	0,476	0,053	-6,67	0,000
fez_aporte $*$ forma_pagamento	0,227	1,254	0,043	6,57	0,000
vgbl $*$ forma_pagamento	-0,516	0,597	0,048	-6,42	0,000

Tabela 4.8: Resultados do ajuste do modelo de regressão de Cox para os dados em estudo e correspondentes razões de risco.

evidências para a rejeição da suposição de riscos proporcionais. Os resultados estão na Tabela (4.9).

Para verificar se a suposição de riscos proporcionais no modelo de Cox é violada, poderíamos realizar um teste de hipóteses baseado nesses resíduos para saber se os coeficientes de correlação são iguais. Ou seja, se a hipótese de riscos proporcionais é válida. Porém, como a amostra utilizada neste estudo é muito grande, não se deve tomar decisões a partir do p-valor, pois este consequentemente tem valores muito pequenos devido ao tamanho amostral.

4.2.1 Interpretação dos coeficientes estimados

A interpretação dos coeficientes estimados no modelo de riscos proporcionais de Cox é feita usando-se a propriedade de riscos proporcionais.

Para cada interpretação, considere que as outras características são as mesmas entre os grupos que estão sendo comparados (*ceteris paribus*).

Variável	ρ
contribuicao_2	-0,010
contribuicao_3	-0,025
contribuicao_4	-0,011
contribuicao_5	-0,025
contribuicao_6	-0,016
contribuicao_7	-0,013
contribuicao_8	-0,001
estado_civil_d_2	-0,002
estado_civil_s_3	-0,016
estado_civil_v_4	0,002
faixa_etaria_2	-0,008
faixa_etaria_3	0,013
faixa_etaria_4	0,020
faixa_etaria_5	0,005
faixa_etaria_6	0,002
faixa_etaria_7	0,004
fez_aporte	0,010
forma_pagamento	0,008
sexo_m	-0,010
tipo_pagamento_2	-0,036
tipo_pagamento_3	-0,004
vgbl	-0,003
estado_civil_s_3∗faixa_etaria_4	-0,018
estado_civil_s_3∗forma_pagamento	0,029
estado_civil_s_3∗tipo_pagamento_2	0,007
faixa_etaria_3∗forma_pagamento	-0,018
fez_aporte∗forma_pagamento	0,043
fez_aporte∗vgbl	0,000
forma_pagamento∗faixa_etaria_4	-0,024
vgbl∗forma_pagamento	-0,017

Tabela 4.9: Coeficientes de correlação de Pearson (ρ) entre resíduos padrozinados de Schoenfeld e o tempo.

O risco dos clientes divorciados e viúvos não é significante maior que os clientes do grupo de referência (casados): 15% e 8% maiores, respectivamente. Os participantes solteiros têm o risco de cancelar dependente de outras variáveis devido a interações. O risco estimado é dado por:

$$\begin{aligned} RR(solteiro = 1) = exp\{0,047 - 0,109 * faixa\_etaria\_4 \\ +0,214 * tipo\_pagamento\_2 + 0,161 * forma\_pagamento\}. \end{aligned} \tag{4-8}$$

Assim sendo, quando o cliente opta por fazer pagamentos em carnê, em parcela única e não tem entre 30 e 49 anos, seu risco será 52% maior que o risco de pessoas casadas cancelarem o plano de previdência.

Homens tem o risco de cancelamento apenas 12% maior que o risco associado às mulheres. Sendo assim, não há diferenças significativas entre a classe gênero.

Entre as faixas etárias, aqueles clientes com idade entre 20 e 24 anos têm o risco de cancelar 52% maior que o risco do grupo basal (pessoas com até 19 anos). Se o participante tiver entre 30 e 49 anos, seu risco estimado de cancelar o plano será dado por:

$$RR(faixa\_etaria\_4 = 1) = exp\{0,354 - 0,109 * solteiro - 0,188 * forma\_pagamento\} \quad (4\text{-}9)$$

e se o cliente dessa faixa etária não for solteiro e fizer pagamentos em parcela mensal, seu risco será 43% maior que o risco do grupo de base. Aqueles com idade entre 20 e 24 anos tem o risco aumentado também em 43% em relação ao grupo de referência.

O cliente que fez algum aporte extraordinário, tem um plano VGBL e faz pagamentos em parcela mensal têm o risco de cancelar igual a 86% menor que o risco daqueles que nunca fizeram um aporte extraordinário. O risco estimado em relação à ocorrência de um aporte extraordinário é dado pela equação a seguir:

$$RR(fez\_aporte = 1) = exp\{-1,226 - 0,742 * vgbl + 0,227 * forma\_pagamento\} \quad (4\text{-}10)$$

O fato de um cliente que possui um plano VGBL não ter feito um aporte extraordinário está associado a um aumento significativo no risco de cancelamento. O risco estimado para clientes com plano VGBL é dado por:

$$RR(vgbl = 1) = exp\{1,202 - 0,742 * fez_aporte + 0,227 * forma\_pagamento\} \quad (4\text{-}11)$$

que quando $fez\_aporte = 1$ e pagamento mensal ($forma\_pagamento = 0$) será igual a 4,17 vezes o risco daqueles que têm um plano PGBL.

O tipo de pagamento com menor risco estimado de cancelamento é entre aqueles que escolhem pagar em débito em conta corrente. Quando o pagamento é em débito em poupança, o risco estimado de cancelamento é 16% maior que o risco dos que pagam em débito conta corrente. O risco estimado de cancelar o plano daqueles que pagam em carnê (fórmula abaixo) é 46% maior que o risco dos que pagam em débito em conta corrente quando cliente é solteiro.

$$RR(tipo\_pagamento\_2 = 1) = exp\{0,162 + 0,214 * solteiro\} \quad (4\text{-}12)$$

Em relação à forma de pagamento, o risco estimado associado àqueles que pagam em parcela única é dado por:

$$RR(forma\_pagamento = 1) = exp\{-0,031 + 0,161 * solteiro - 0,516 * vgbl - 0,188 * faixa\_etaria\_4 + 0,227 * fez\_aporte\}. \quad (4\text{-}13)$$

Quando não for solteiro, com idade entre 30 e 49 anos, não fez aporte extraordinário e o plano é VGBL, o risco estimado de cancelamento é metade do risco daqueles que pagam mensalmente.

Entre as faixas de contribuição, os dois grupos com os maiores valores mensais tem os menores risco de cancelar: 30% e 43% menores que o risco do grupo de referência (contribuições até R$99,00), respectivamente.

4.3 Comparação e validação dos modelos

Ajustados os modelos e feita a análise de adequação através dos resíduos, o próximo passo é comparar os modelos paramétrico e de Cox e medir o poder de previsão deles. Ou seja, o interesse agora é saber o quanto o escore produzido pelos modelos consegue distinguir os bons e maus clientes e assim poder identificar de forma prévia se o cliente é um risco para a seguradora ou um potencial investidor.

Para tal, duas ferramentas foram utilizados nesse trabalho: tabela de classificação, que gera medidas de capacidade de acerto e a curva ROC, que ilustra visualmente qual o modelo é melhor.

O escore utilizado no modelo de regressão paramétrico para o cálculo da curva ROC é o tempo mediano estimado. Nesse caso, os 9634 clientes com menores tempos medianos foram previstos como cancelados e os demais ativos. No caso do modelo de Cox, o escore utilizado foi a função de risco estimada: os 9634 clientes com maiores riscos são classificados como maus e os demais como bons.

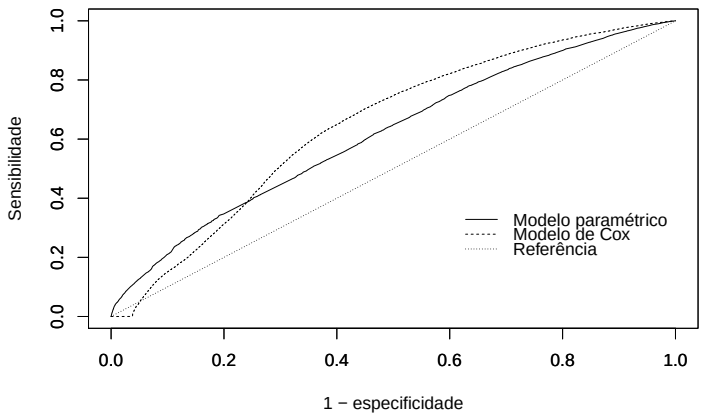
A curva ROC e a tabela de classificação foram ajustados para as amostras de estimação (70% dos dados) e validação do modelo (30% dos dados).

Analisando os resultados através da amostra de estimação pela curva ROC, Figura (4.3), podemos ver que o modelo de riscos proporcionais de Cox melhor comparado ao modelo paramétrico gama generalizado. A capacidade de acerto total é de 62,2% para o modelo de Cox e 57,74% para o modelo paramétrico. A capacidade de acerto dos maus e dos bons clientes é de 58,88% e 65,02% no modelo de Cox e 54,03% e 60,89% no modelo paramétrico, respectivamente. Essas medidas de capacidade preditiva foram calculadas com

base nos números da Tabela (4.12), para o modelo paramétrico e com base nos números da Tabela (4.10), para o modelo de riscos proporcionais de Cox.

Quando avaliados pela amostra de validação, os dois modelos apresentaram poderes preditivos praticamente iguais, como pode ser visto pela curva ROC, Figura (4.4). A capacidade de acertos total, dos maus e dos bons clientes do modelo paramétrico é, respectivamente, 58,19%, 55,11% e 60,88%; no modelo de Cox essas medidas são 58,63%, 55,57% e 61,29%, respectivamente. Essas medidas de capacidade preditiva foram calculadas com base nos números da Tabela (4.13), para o modelo paramétrico e com base nos números da Tabela (4.11), para o modelo de riscos proporcionais de Cox.

Figura 4.3: Curva ROC dentro da amostra.



	Cancelado previsto	Ativo previsto
Cancelado real	13067	9125
Ativo real	9125	16961

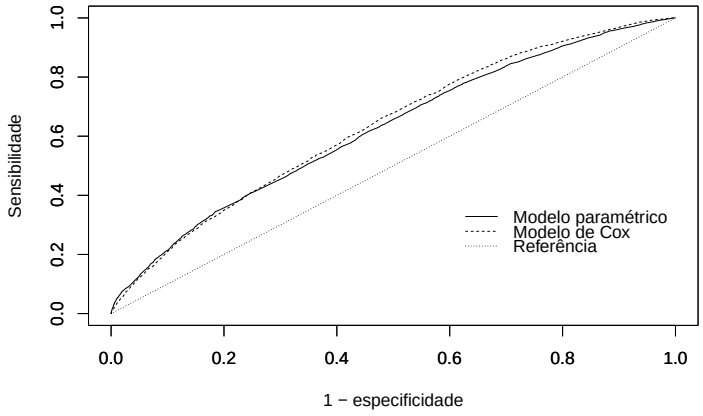
Tabela 4.10: Acertos/erros modelo de riscos proporcionais de Cox dentro da amostra.

	Cancelado previsto	Ativo previsto
Cancelado real	5354	4280
Ativo real	4280	6776

Tabela 4.11: Acertos/erros modelo de riscos proporcionais de Cox fora da amostra.

Outra comparação que pode ser feita é em relação às estimativas dos modelos e do estimador de Kaplan-Meier. Porém esta comparação se torna complexa a medida que cada modelo têm uma combinação parâmetros e interações estimadas e, além disso, o estimador de Kaplan-Meier calcula estimativas para cada uma das co-variáveis sem levar em conta a presença de outras

Figura 4.4: Curva ROC fora da amostra.



	Cancelado previsto	Ativo previsto
Cancelado real	11990	10202
Ativo real	10202	15884

Tabela 4.12: Acertos/erros modelo paramétrico dentro da amostra.

	Cancelado previsto	Ativo previsto
Cancelado real	5309	4325
Ativo real	4325	6731

Tabela 4.13: Acertos/erros modelo paramétrico fora da amostra.

co-variáveis. Usaremos a co-variável estado civil para fazer tal comparação. Entre os divorciados, o tempo mediano estimado por Kaplan-Meier e pelo modelo paramétrico indicaram uma diminuição de aproximadamente 20% em relação aos casados. O risco de clientes divorciados cancelarem o plano de previdência é 20% maior que os clientes casados. No caso dos clientes solteiros, o tempo mediano e o risco de cancelamento estimados pelos modelos paramétrico e de Cox, respectivamente, dependem de interações. Por isso, as estimativas entre os três métodos não são parecidas: segundo Kaplan-Meier, o tempo mediano dos clientes solteiros é 20% menor que o grupo de referência (casados), enquanto que segundo o modelo paramétrico é 40% menor se esses clientes pagarem as contribuições em carnê e de forma única e não tiverem idade entre 30 e 49 anos. Com as mesmas características dos clientes solteiros no modelo paramétrico, segundo o modelo de riscos proporcionais de Cox esses participantes terão risco de cancelamento 30% menor que os casados.

5
Conclusão e trabalhos futuros

O objetivo desse trabalho foi propor um novo tipo de abordagem para a estimação de persistência de clientes em planos de previdência privada, explicando a durabilidade do segurado através de características suas e do produto adquirido, como tipo do plano, sexo, idade, etc. Os dados utilizados são provenientes de uma seguradora nacional.

Iniciou-se o estudo por uma análise descritiva dos dados através do estimador não-paramétrico de Kaplan-Meier. Para cada uma das variáveis explicativas contidas na base de dados, a função de sobrevivência foi estimada para cada uma das sub-categorias. A partir das estimativas, pôde-se verificar quais características estão relacionadas às maiores probabilidades de sobrevivência, especialmente a proporção de clientes que permanecem ativos após o aniversário de um ano do plano e o tempo mediano de cada uma das sub-categorias. Por essas estatísticas foi possível adquirir conhecimento sobre a relação do tempo de duração de um cliente com suas próprias características e, assim, construir um ideia prévia das variáveis relevantes na estimação dos modelos.

Dois modelos de análise de sobrevivência foram propostos: um modelo de regressão com uma distribuição adequada aos dados e um modelo de natureza semi-paramétrica, o modelo de riscos proporcionais de Cox. Na estimação do modelo de regressão paramétrico, partiu-se de um modelo com distribuição gama generalizada, visto que esta tem como casos particulares outras distribuições também adequadas para o tipo de dados em questão. A seleção de variáveis foi feita pelo método *stepwise-forward* com um nível de significância de 10%. Depois, testou-se as possíveis interações dupla. Com o conjunto de variáveis explicativas e interações duplas escolhidas, ajustou-se modelos com as outras distribuições: exponencial, Weibull, log-logística e log-normal. Escolheu-se o modelo de regressão gama generalizado, que minimizou os critérios AIC e BIC de seleção de modelo. O modelo de riscos proporcionais de Cox foi ajustado da mesma forma que o modelo de regressão paramétrico: método *stepwise* para seleção de variáveis. O mesmo conjunto de interações selecionadas no modelo de regressão foi utilizado. Para o tratamento

de empates na estimação, foram testadas as aproximações da função de verossimilhança parcial propostas por Breslow e Efron e a segunda mostrou uma melhor performance.

Para avaliar a adequação do modelo de regressão gama generalizado, ajustou-se os resíduos de Cox-Snell, que se mostraram adequados para o modelo ajustado. Os resíduos padronizados de Schoenfeld foram utilizados para verificar se a suposição de riscos proporcionais foi violada, o que não aconteceu.

O método *holdout* de validação cruzada foi utilizado para verificar a capacidade preditiva não só entre os dados da estimação, mas também para uma amostra de validação que não fora utilizada na estimação. O desempenho dentro da amostra do modelo de riscos proporcionais de Cox foi melhor que o modelo de regressão gama generalizada, tendo acurácia igual a 62,2% e 57,74%, respectivamente. Na amostra de validação, o desempenho dos dois modelos segundo a acurácia foi praticamente igual: 58,19% no modelo de regressão gama generalizado e 58,63% no modelo de riscos proporcionais de Cox.

Os modelos de sobrevivência permitem que a informação daqueles em que o evento de interesse (no caso o cancelamento do plano pelo cliente) não ocorreu seja absorvida. Além disso, leva-se em consideração não só a ocorrência do cancelamento, mas também a relação que o evento tem com o tempo.

Outro ponto interessante é a previsão do tempo de duração fornecida pelos modelos paramétricos de acordo com o escore do cliente. Com isso, pode-se direcionar de maneira otimizada um melhor fundo de investimento de modo que a receita seja maximizada para o banco. As previsões do tempo de duração do cliente podem direcionar a seguradora a fazer campanhas e dar incentivos àqueles que tem os menores tempos.

Se o interesse for utilizar os modelos para estimação na prática, mais atenção deve ser dada ao banco de dados. Uma base mais estruturada e com informações mais completas aprimoraria o modelo e consequentemente suas conclusões. Visto que planos de previdência completar são investimentos feitos a longo prazo, uma base de dados com maior tempo de observação e variáveis explicativas que acompanhem o cliente (ou seja, variantes no tempo) provavelmente trariam mais acurácia às estimações. Por exemplo, se pudéssemos acompanhar os resgates ao longo do tempo do cliente, poderíamos incorporar uma variável exógena, como a taxa de retorno, de modo que através dessa variável o modelo conseguisse captar a relação entre a rentabilidade de outros produtos com a evasão nos planos de previdência privada.

Referências Bibliográficas

[1] FENAPREVI. **Planos de Caráter Previdenciário - Dados Estatísticos**. Technical report, Federação Nacional de Previdência Privada e Vida, dezembro 2012. 1

[2] LIAN, K.; YUAN, W. ; LOI, S. Survival analysis of terminated life insurance policies. **Singapore International Insurance and Actuarial Journal**, v.2, p. 101–119, 1998. 1

[3] GUSTAFSSON, E. **Customer duration in non-life insurance industry**. Suécia, abril 2009. Dissertação de Mestrado - Dept of Mathematics – Stockholm University. 1

[4] TORRINI, F.; PEREIRA, N. **Análise de Sobrevivência Aplicada à Previdência Privada: Um Ajuste para o Tempo de Relacionamento dos Clientes**. Monografia final de curso, Escola Nacional de Ciências Estatísticas, Rio de Janeiro, dezembro 2010. 1

[5] CHUN, R. **Análise de Persistência de Participantes em Planos de Previdência**. Rio de Janeiro, abril 2007. Dissertação de Mestrado - Pontifícia Universidade Católica do Rio de Janeiro. 1

[6] STEPANOVA, M.; THOMAS, L. Survival Analysis Methods for Personal Loan Data. **Operations Research**, v.50, n.2, p. 277–289, March-April 2002. 1

[7] DE ABREU, H. J. **Aplicação da Análise de Sobrevivência em um problema de Credit Scoring e comparação com a Regressão Logística**. outubro 2004. Dissertação de Mestrado - Universidade Federal de São Carlos. 1

[8] COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. ABE - Projeto Fisher. 1ª. ed., São Paulo: Edgard Blücher, 2006. 2

[9] KLEIN, J. P.; MOESCHBERGER, M. L. **Survival Analysis - Techniques for Censored and Truncated Data**. New York: Springer, 1997. 2.1

[10] KALBFLEISCH, J. D.; LAWLESS, J. F. Inference based on retrospective ascertainment: an analysis of the data on transfusion related AIDS. **Journal of American Statistical Association**, v.84, p. 360–372, 1989. 2.1

[11] KAPLAN, E.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, p. 457–81, 1958. 2.2.1

[12] BRESLOW, N.; CROWLEY, J. A Large Sample Study of the Life Table an Product Limit Estimates under Random Censorship. **Annals os Statistics**, v.2, p. 437–453, 1974. 2.2.1

[13] KALBFLEISCH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data**. 2ª. ed., New Jersey: John Wiley and Sons, 2002. 2.2.1

[14] MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**, v.50, p. 163–170, 1966. 2.2.1

[15] GEHAN, E. A. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. **Biometrika**, v.52, p. 203–223, 1965. 2.2.1

[16] COX, D. R. Regression models and life tables. **Journal of the Royal Statistical Society Series B**, v.34, p. 187–220, 1972. 2.3, 2.3, 2.3

[17] CORDEIRO, G. M.; DEMÉTRIO, C. G. **Modelos Lineares Generalizados e Extensões**, 2009. 2.3, 2.3, 2.3

[18] HOSMER, D. W.; LEMESHOW, S. **Applied Survival Analysis**. John Wiley and Sons, 1999. 2.4.7, 4.1.1

[19] COX, D.; HINKLEY, D. **Theoretical Statistcs**. London: Chapman and Hall, 1974. 2.5.1

[20] COX, D. Partial Likelihood. **Biometrika**, v.62, p. 269–76, 1975. 2.5.1

[21] BRESLOW, N. Contribuição à Discussão do artigo de D.R. Cox. **Journal of the Royal Statistical Society B**, v.34, p. 216–217, 1972. 2.5.1, 2.5.1

[22] EFRON, B. The Efficiency of Cox's Likelihood Function for Censored Data. **Journal of the American Statistical Association**, v.72, p. 557–565, 1977. 2.5.1

[23] ANDERSEN, P.; GILL, R. Cox's Regression Model for Couting Processes: A Large Sample Study. **Annals os Statistics**, v.10, p. 1100–1200, 1982. 2.5.1

[24] COX, D.; SNELL, E. A general definition of residuals (with discussion). **Journal of the Royal Statistical Society Series B**, v.30, p. 248–275, 1968. 1

[25] SCHOENFELD, D. Partial Residuals for the Proportional Hazard Regression Model. **Biometrika**, v.69, p. 239–241, 1982. 2

[26] M. T. THERNEAU, P. G.; FLEMING, T. Martingale-based residuals for survival models. **Biometrika**, v.77, p. 147–160, 1990. 2.7

[27] KOHAVI, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 1995. 2.7.1

[28] HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. John Wiley and Sons, 2000. 2.7.1