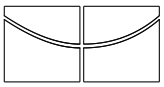


Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionais de Cox

por

Aline Rodrigues Machado

Orientador: Prof.º Eduardo Yoshio Nakano

Brasília

2015

Aline Rodrigues Machado

Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionais de Cox

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof.º Eduardo Yoshio Nakano

Universidade de Brasília

Brasília, 2015

Agradecimentos

Agradeço a Deus por ter me dado força e coragem para não desanimar, mesmo diante a tantos desafios e dificuldades.

A minha família pelo apoio e incentivo de sempre.

A todos os professores da Universidade de Brasília que contribuíram para a minha formação desde a graduação, em especial ao Eduardo Nakano, por toda a compreensão, ajuda e disposição durante a realização deste trabalho.

Aos colegas de trabalho, que às vezes mesmo sem saber colaboraram com o desenvolvimento dessa dissertação.

Agradeço também aos colegas do mestrado, em especial Rafaela, Paulo, Tiago, Augusto e Adélio.

Resumo

Em virtude da profissionalização da política de recuperação de créditos, o objetivo da dissertação foi o desenvolvimento de modelos de *Collection Scoring* para a determinação de melhores estratégias de atuação de cobrança para cada perfil de cliente. Os dados foram obtidos por meio do banco de dados de uma grande instituição financeira de atuação em nível nacional sendo que o universo de aplicação do modelo foi clientes com atraso superior a 90 dias. Foram ajustados um modelo de Regressão Logística e um de Riscos Proporcionais de Cox. Os resultados para os dois modelos foram semelhantes, mas a aplicação com o modelo de Cox permite a estimação considerando também o tempo até a recuperação.

Palavras-Chave: atraso, Basileia, ciclo de crédito, cobrança, descumprimento, inadimplência, modelo de *Credit Scoring*, probabilidade, recuperação, régua de cobrança, risco de crédito.

Abstract

Because of the professionalization of credit and collection policy, the aim of this work was the development of Collection Scoring Models to determine best collection action strategies for each customer profile. Data were obtained from the database of a large financial institution acting at the national level and the universe of application of the model was clients that not pay for more than 90 days. We adjusted a Logistic Regression and a Cox Proportional Hazard Model. Both adjusted models presented similar results, but the Cox model procedures also considers the time until the recovery event.

Key words: delay, Basileia, credit cycle, collection, failure, default, Credit Scoring Model, probability, recovery, collection rule, credit risk.

Sumário

Capítulo 1

Introdução

O mercado de crédito tem papel fundamental na economia de um país, uma vez que sustenta as atividades financeiras, como projetos econômicos, investimentos e aquisição de bens de consumo, influenciando diretamente no PIB nacional. No Brasil, após a estabilização da economia - iniciada com a implantação do Plano Real - o controle da inflação e a maior geração de empregos, a indústria do crédito vem apresentando altas taxas de crescimento, influenciadas pela capacidade de pagamento dos tomadores, tornando o ramo de concessão de crédito atrativo aos interesses das instituições financeiras devido à rentabilidade esperada sobre o capital emprestado. Por outro lado, a expansão do crédito também provoca maior exposição das instituições ao risco de inadimplência, ou seja, de não receberem - ou receberem de forma parcial - o capital previamente emprestado.

Nesse contexto, para garantirem bons resultados financeiros, as empresas necessitam de métodos que auxiliem na gestão estratégica sobre os riscos envolvidos na contratação de crédito, desde a proposta de concessão, até os processos de cobrança, já que a rentabilidade está associada ao número de empréstimos concedidos e ao percentual de clientes que honram os compromissos acordados.

Segundo Thomas *et al.* (2002), até o início do século XX, todas as decisões relativas à concessão de crédito eram baseadas exclusivamente no julgamento subjetivo dos analistas. Somente a partir da publicação, em 1936, da técnica de Análise Linear de Discriminante, desenvolvida por Fisher, é que a estatística começou a ser pensada para identificar bons e maus pagadores. Assim, os primeiros modelos de *Credit Scoring* foram desenvolvidos por Durand (1941), com o objetivo de ordenar os proponentes quanto à probabilidade de pagar o capital emprestado. Diante da maior agilidade na decisão, menor custo, maior objetividade e até mesmo melhor poder preditivo, os modelos de *Credit Scoring* foram aos poucos se popularizando e atualmente são largamente utilizados (Hand e Henley, 1997). Desse modo, Regressão Logística, Análise Discriminante, Análise de Sobrevivência, Árvores de Decisão, Redes Neurais, Cadeias de Markov, Algoritmos Genéticos, Modelos Lineares Generalizados, Análise de Agrupamento e Inferência Bayesiana vêm sendo utilizadas como ferramentas para auxiliar na concessão, acompanhamento, cobrança, retenção e prospecção de clientes.

Nessas circunstâncias, as qualidades encontradas em se reunir técnicas estatísticas com a experiência dos analistas de crédito contribuíram para o desenvolvimento de diversos tipos de modelos de *Credit Scoring*, com objetivos específicos de acordo com o tipo de risco e o estágio no ciclo de crédito, entre os quais, pode-se destacar: modelos de *Prospect Scoring*, *Marketing Propensity Scoring*, *Application Scoring*, *Fraud Scoring*, *Behaviour Scoring*, *Customer Scoring*, *Attrition Scoring*, *Collection Scoring* e *Profit Scoring*.

Ultimamente, os modelos de *Credit Scoring* ganharam mais importância com o Novo Acordo de Basileia - Basileia II - que determina a utilização de técnicas que permitam aos bancos e órgãos supervisores avaliarem os riscos aos quais as instituições estão sujeitas. Portanto, as empresas estão despendendo esforços e estudos para o desenvolvimento de novas técnicas que auxiliem os sistemas de *scoring* e segundo Colosimo e Giolo (2006), uma das mais recentes é a Análise de Sobrevivência, cujo objetivo consiste no tempo até a ocorrência de determinado evento de interesse.

Trabalhos propostos por Narain (1992), Banasik et al (1999), Thomas e Stepanova (2002), Abreu (2004), Andreeva (2006), Tomazela (2007) e Machado (2010) foram relevantes e iniciais no contexto do desenvolvimento de modelos de *Credit Scoring* utilizando Análise de Sobrevivência. Nesses casos, o foco fundamentou-se no acompanhamento do tempo até a ocorrência da inadimplência. Este trabalho, entretanto, tem o objetivo inovador de utilizar a Análise de Sobrevivência para desenvolvimento de um modelo de *Collection Scoring*, visando classificar o risco do cliente inadimplente em termos de pagamentos futuros. Serão utilizados dados de uma grande instituição financeira de atuação nacional para estimar o tempo necessário para normalização dos valores devidos.

O *Collection Scoring* é um modelo de escore baseado em dados de clientes inadimplentes que busca, de forma eficaz, a regularização de créditos em atraso de clientes que não puderam, por motivos diversos, honrar com os compromissos assumidos e corresponder à expectativa de pagamento do capital tomado por meio do financiamento, tornando-se inadimplentes.

Capítulo 2

Contextualização

Este capítulo apresenta uma breve descrição sobre crédito, risco e a utilização de ferramentas para mensuração do risco de crédito nas diversas etapas do ciclo financeiro.

2.1 Modelos de *Credit Scoring*

No decorrer de toda a história de desenvolvimento econômico e social das sociedades, o crédito é um dos fatores mais importantes a serem considerados, pois é por meio dele que as empresas ampliam os seus negócios, gerando emprego e renda, o que impulsiona o consumo e estimula a demanda, permitindo a produção e expansão econômica. Para as pessoas, o crédito representa uma ampliação dos recursos financeiros, para pagamento de dívidas e financiamentos, cumprindo com sua função social, possibilitando aquisição de bens, como automóveis e moradias.

Sob essa perspectiva financeira, o crédito corresponde a um valor monetário disponibilizado ao tomador de recursos financeiros, em forma de empréstimo ou financiamento, por um período previamente pactuado, com a promessa de pagamento futuro, ao qual é acrescido uma remuneração, denominada juros. Nesse contexto, o

risco é inerente ao processo de concessão de crédito, uma vez que existem incertezas quanto ao futuro das quantias emprestadas.

Segundo Yamamoto, Oliveira e Santos (2011), "o risco é definido pela incerteza de retorno de um investimento perante a possibilidade de um evento possível, futuro e incerto, autônomo à vontade do investidor e cuja ocorrência poderá causar prejuízos". Nesse sentido, o risco de crédito está ligado a fatores internos e externos ao concessor que podem prejudicar a recuperação do montante emprestado. Para o Banco Central do Brasil, conforme Art. 2º da Resolução 3.721/2009, risco de crédito é definido como a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação.

Dessa forma, o risco de crédito pode ser analisado sob vários aspectos, dentre eles:

- Risco do Cliente - associado aos C's do Crédito;

    1. Capacidade - habilidade em pagar. Diz respeito aos meios financeiros para honrar com os compromissos assumidos;
    2. Colateral - garantia;
    3. Caráter - confiabilidade e "vontade" de pagar;
    4. Condição - condições ambientais externas, internas e indicadores econômicos;

5. Capital - reservas e patrimônio.

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

Consequentemente, a avaliação do risco é essencial para o sucesso do negócio de concessão de crédito, uma vez que previne perdas e minimiza os riscos. Sob esse cenário, surgiram os Modelos de *Credit Scoring*, como ferramenta capaz de quantificar o risco de crédito envolvido em uma operação. Além disso, esses modelos capacitam os usuários a tomar decisões de forma rápida, automática, padronizada e objetiva, se adequando às milhares de escolhas que devem ser feitas todos os dias.

Os Modelos de *Credit Scoring* (CS) utilizam-se de algorítmos matemáticos e técnicas estatísticas para calcular a probabilidade de que determinado evento aconteça. Aplicando fórmulas, o sistema atribui pontuação específica para cada característica do proponente/cliente para prever um resultado.

Historicamente, os modelos de *Credit Scoring* foram iniciados pelos estudos de Durand (1941) na área de financiamento ao consumidor após a Grande Depressão nos EUA. O projeto foi precursor na utilização da Estatística como ferramenta para análise de risco de crédito. Nesse trabalho, foi utilizada a Análise de Discriminante desenvolvida por Fisher (1936) para identificar bons e maus empréstimos. Nesse contexto, a pesquisa de Durand pode ser considerada como o ponto de partida para

futuros estudos que considerem o desenvolvimento de metodologias de suporte à concessão de crédito.

No início dos anos 1950, Bill Fair e Earl Isaac criaram a primeira Companhia para consultoria em métodos de *scoring* por acreditarem que dados históricos podem melhorar as decisões negociais se utilizados com inteligência. Assim, em 1958 foi vendido o primeiro Sistema de *Credit Scoring* para a área de Cartão de Crédito. Esse fato é considerado o segundo passo importante para a história dos modelos de *scoring*. Por outro lado, o sucesso da Companhia e a sua finalidade comercial não implicaram em desenvolvimento de literatura sobre o tema, uma vez que o conhecimento tornou-se valioso e pouco exibido.

Apesar de Modelos de *Credit Scoring* representarem uma melhoria em relação às análises julgamentais de risco de crédito, houve dificuldades que impediam o seu crescimento, como relutância dos executivos, limitações tecnológicas para aplicação das metodologias, obstáculos no desenvolvimento e implementação dos modelos e, segundo Myers e Forgy (1963), a falta de estatísticos para propagar-se na área de crédito e fazer o trabalho de transformar essa ideia em uma ferramenta operacional bem sucedida e útil. Diante do exposto, apesar do crédito continuar em expansão nos Estados Unidos, poucos estudos sobre *Credit Scoring* foram produzidos até os anos 1960.

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

- Desenvolvimento de Sistemas Numéricos de Avaliação de Crédito, Myers e Forgy (1963). Eles se empenharam em verificar a eficácia das fórmulas predi-

tivas de *scoring* e dessa forma introduziram o conceito de amostra *hold-out*, ou seja, diferente daquela utilizada para a modelagem. Esse foi um fato importante, uma vez que existe chance de um modelo distinguir bons e maus na base original, mas não ser preditivo em outras amostras;

- Conceitos e Utilização de Técnicas de *Credit Scoring*, Weingartner (1966). O autor ressaltou a importância de testes antes da utilização dos escores de crédito e sugeriu uma nova técnica de validação: aplicar a fórmula a clientes inadimplentes para verificar se os escores são baixos;

- Índices Financeiros, Análise de Discriminante e Previsão de Falência de Empresas, Altman (1968). Introdução dos Modelos de *Scoring* para empresas;

- Um Modelo de *Credit Scoring* para Empréstimos Comerciais, Orgler (1970). Propôs um modelo para avaliar periodicamente a qualidade dos empréstimos já concedidos.

A partir de 1970, com aumento da demanda ao crédito ao consumidor, muitas instituições financeiras nos EUA cresceram de forma insustentável, uma vez que avançaram além de suas capacidades de formar e manter uma equipe adequada e experiente de avaliadores de crédito. Aliada a tal fato, a reconstrução da Europa pós Guerra contribuíram para que Modelos de *Credit Scoring* fossem reconhecidos como uma indústria. Desde o início dos anos 1990, os Modelos de CS tornaram-se o método dominante para a avaliação de risco na concessão de vários tipos de empréstimos, sendo as decisões tomadas sem intervenção ou envolvimento de quem está procedendo a avaliação.

A partir da divulgação do Acordo de Basileia II ocorrida em 2004, os Modelos de *Credit Scoring* tornaram-se ainda mais importantes, uma vez que o documento destacou a utilização de técnicas que permitam às instituições e supervisores avaliar corretamente os vários riscos que os bancos enfrentam. Muitas organizações desenvolveram melhores modelos ou modificaram os já existentes para estar em conformidade com as novas regras e com as melhores práticas de mercado, dado que os reguladores forçaram regras mais rigorosas sobre o desenvolvimento, implementação e validação dos modelos internos utilizados para estimar capital a ser provisionado.

Com o contínuo desenvolvimento e crescimento dos mercados financeiros, o crédito tornou-se ainda mais importante na economia. Com a globalização e a sofisticação dos canais, como internet e postos de auto-atendimento, os consumidores tendem a procurar e escolher as ofertas de crédito mais atrativas, sem limitações de tempo e lugar. Por isso as instituições buscam desenvolver eficientes ferramentas para avaliar e controlar os riscos de crédito.

Instituições financeiras estão lidando com grandes volumes de clientes e pequena margem de lucro no nível individual de transação, dessa forma, se esforçam para ter vantagem competitiva, expondo-se a riscos de maneira estratégica. Isso significa que é necessário minimizar todos os riscos, para que perdas possam ser evitadas. Atualmente, isso é alcançado por meio de uma gestão estratégica dos riscos, obtida em grande parte a partir de resultados de modelos estatísticos.

Modelos de *Credit Scoring*, inicialmente utilizados apenas para decisão de conceder ou não determinado valor ou limite, hoje fazem parte de todo o ciclo do crédito,

estando presente em cada etapa da gestão estratégica de riscos. A Figura 2.1 ilustra o ciclo de crédito, apresentando cada uma de suas etapas.

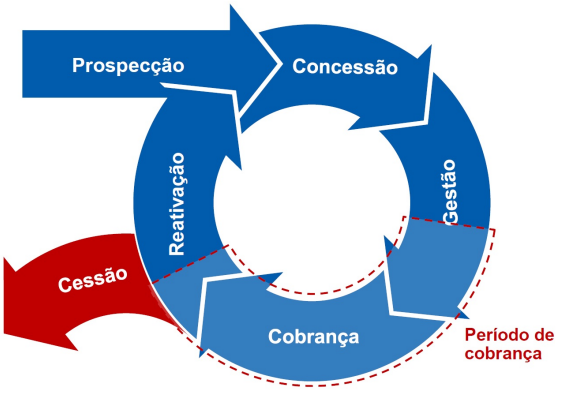


Figura 2.1: O ciclo de crédito (Fonte: SERASA *Experian*, 2011)

Dessa forma, os Modelos de *Credit Scoring* são amplamente aceitos e utilizados pelas Instituições do Mercado Financeiro em todas as etapas do ciclo de crédito, transformando-se em:

1. Modelos de *Prospect Scoring*: avaliar o perfil de risco do proponente com foco no produto, a fim de identificar consumidores mais propensos à concessão, de modo a minimizar a insatisfação dos consumidores com propostas que não dizem respeito aos perfis correspondentes, oferecer crédito à pessoa certa e reduzir custos operacionais - não são utilizados recursos com todos, apenas com aqueles que o modelo mostra ser vantajoso;

2. Modelos de *Marketing Propensity Scoring* - verifica a probabilidade de um proponente/cliente comprar um produto após uma camanha publicitária, com o objetivo de maximizar o retorno envolvido nas campanhas (ROI) - não é necessário investir em campanhas para todos, mas apenas para público específico

determinado pelo modelo e de maximizar a taxa de conversão - aumentar a relação entre o número de ofertas e de contratações;

3. Modelos de *Application Scoring*: ferramenta utilizada para captação da proposta e análise do risco de crédito com foco no perfil de risco do proponente baseada em aspectos sociais, financeiros e demográficos para decisão de aceitação e definição de limites e condições, ou seja, verifica a probabilidade do proponente não pagar seu compromisso antes de completar um período prefixado;

4. Modelos de *Fraud Scoring*: utilizado na validação dos dados do cliente para a prevenção de fraudes. O escore ordena os clientes/proponentes de acordo com a probabilidade da aplicação ser fraudulenta, colaborando com o aumento dos lucros e melhor atendimento ao cliente, por meio da identificação de possíveis fraudes logo no início do relacionamento;

5. Modelos de *Behaviour Scoring*: verifica a probabilidade do cliente apresentar atrasos durante o relacionamento creditício. Método empregado para monitorar e reavaliar o risco de crédito para apoio a processos de cobrança preventiva, renovação e definição de valores de provisão. Além disso, contribui para rentabilização, manutenção e retenção uma vez que para clientes bons há a possibilidade de ofertar novos produtos, aumentando *Cross Selling* e *Up Selling*. Ou seja, o escore de BS quantifica o comportamento dos clientes permitindo melhor gestão da carteira e do cliente, uma vez que proporciona melhor entendimento sobre o consumidor e suas necessidades;

6. Modelos de *Customer Scoring* - sumariza, em uma única medida, o risco do cliente em cada um dos produtos de crédito adquiridos;

7. Modelos de *Attrition Scoring* - verifica a probabilidade do cliente cancelar o produto;

8. Modelos de *Collection Scoring*: avaliar a probabilidade de clientes em atraso regularizar o pagamento em determinado período de tempo com o propósito de ajustar a abordagem e intensidade do processo de cobrança a fim de maximizar a recuperação, reduzir custos, evitar desgastes desnecessários com cliente e automatizar fluxos;

9. Modelos de *Profit Scoring* - verifica a probabilidade de os clientes serem rentáveis para a instituição financeira.

Como demonstrado, no mundo competitivo de hoje, modelos de *Credit Scoring* não são mais uma simples ferramenta para obter a probabilidade de um proponente ser "mau" pagador e, a partir disso, decidir pela concessão ou não. No ambiente atual, o crédito é muito mais complexo, deve haver respeito às determinações legais e normativas, o que exige controle de todos os riscos, além de buscar estar sempre à frente dos concorrentes. O objetivo geral dos Modelos de *Credit Scoring* é ser capaz de atrair "bons" clientes, preserevá-los e gerir de forma adequada toda a carteira, para que o negócio seja rentável. Ou seja, o grande desafio da gestão estratégica do risco de crédito é encontrar o equilíbrio entre risco e retorno a partir de todas as etapas do fluxo de crédito.

Sob esse prisma, apesar da estratégia de *Credit Scoring* ser amplamente utilizada no nível da concessão, os benefícios são muito mais significativos se implementados em todo o ciclo, uma vez que as facilidades de *scoring* possibilitam ganhos consideráveis ao negócio, como agilidade, rapidez e automação.

2.2 Modelos de *Collection Scoring*

A expansão da oferta de crédito pelas instituições financeiras traz consigo o aumento da inadimplência. Assim, a preocupação das empresas com a recuperação de crédito é crescente. Os investimentos são cada vez maiores para aperfeiçoar os processos de cobrança, diminuir custos e riscos e aumentar os resultados de recuperação.

Nessa perspectiva, a cobrança torna-se uma etapa do ciclo operacional e financeiro muito importante para o melhoramento do fluxo de caixa e da redução de perdas, maximizando os resultados financeiros das empresas. Nesse sentido, um conjunto de ações interligadas, como conhecimento do nível de relacionamento, gestão integrada do risco, posicionamento competitivo sustentável, relacionamento preventivo e utilização de modelos preditivos são fundamentais.

Dessa forma, a cobrança vem evoluindo conforme os seguintes estágios:

1. Manual: nessa fase, os processos de cobrança eram manuais, sendo distribuídos em filiais, havendo relação pessoal entre cobrador e devedor, fazendo com que as decisões fossem julgamentais, além de ausência de metas ou orçamentos operacionais;

2. Centralizado: nesse estágio foi introduzido o papel de *staff* de cobrança especializado e casos categorizados por estágio de inadimplência. Mas o relacionamento ainda era pessoal;

3. Automatizado: nessa etapa houve o estabelecimento de processos padronizados e automatizados, abolindo a relação entre cobrador e devedor, mas com ênfase na recuperação de curto prazo. Nesse estágio foi introduzida a utilização de escores como ferramenta para definição de prioridades e ações. A remuneração por resultados foi um avanço;

4. Estratégico: evolução da cobrança para foco no cliente, com especialização de recursos e estratégias otimizadas para definir as melhores ações a serem tomadas. Um aspecto relevante foi o desenvolvimento de métricas para indicar o efeito da cobrança frente aos indicadores de curto e longo prazo. Desenvolvimento de Modelos Estatísticos diferentes para estágios de inadimplência distintos, além de ferramentas de otimização e monitoramento avançadas.

A inadimplência pode ser causada por vários motivos, dentre os quais alteração do ambiente, perda de emprego, atraso de salários, descontrole financeiro, priorização de outras dívidas, não recebimento do boleto, problemas na concessão, falta de controle no processo de cobrança e acompanhamento ineficaz ou inexistente. Entretanto, independentemente da justificativa para o não cumprimento dos termos previstos em contrato, o cliente inadimplente não deixa de ser cliente e, portanto, de impactar nos resultados da instituição. Dessa forma, é necessário o desenvolvimento de uma gestão de relacionamento diferenciada para esses clientes, uma vez que parte deles

deixam essa condição de "mau" pagador em mais ou menos tempo. Portanto, o perfil de cliente inadimplente é diferente para cada tipo de comportamento diante do atraso. Acompanhar essa diferença de perfis permite decidir regras de atuação de cobrança mais específicas, adequadas, personalizadas, baratas e eficientes.

Nesse contexto, um avanço que garante mais eficácia no processo de cobrança diz respeito a segmentação dos diferentes perfis de clientes inadimplentes em relação à propensão ao pagamento e isso pode ser obtido por meio da aplicação da tecnologia de *scoring*. Isto é, desenvolver um sistema que indique a probabilidade de como um cliente inadimplente pagará no futuro e fornecer aos gestores uma visão baseada no risco, permitindo melhores decisões sobre a inadimplência, balanceando custos, receitas e tratamentos.

Diante desse cenário, para uma administração profissional e estratégica dos inadimplentes, surgiram os modelos de *Collection Scoring* como instrumento para classificar o risco do cliente em termos de pagamentos futuros. A literatura sobre o assunto, entretanto ainda é pobre. Nos Estados Unidos, os registros indicam a utilização desse tipo de modelo a partir dos anos 80. Nas conferências internacionais, os modelos de *Collection Scoring* são tratados de forma superficial dentro do tema de gestão da cobrança, com foco na aplicação dos conceitos envolvidos, e não na teoria. No Brasil, a literatura disponível é ainda mais restrita, com pouco conhecimento público a respeito da utilização e desenvolvimento de tais modelos.

Os Modelos de *Collection Scoring* consideram o histórico de inadimplência dos clientes, bem como os custos envolvidos nos atos de cobrança e o tempo necessário

para recuperação com a finalidade de identificar a probabilidade de pagamento dos clientes que já se tornaram inadimplentes, permitindo conhecer o risco individual de cada tomador, por meio das características do seu relacionamento com a instituição e seu comportamento de atraso.

Tanto para o desenvolvimento como para a aplicação de um Modelo de *Collection Scoring*, são necessários dois estágios fundamentais: segmentação e agrupamento. De forma semelhante a qualquer Modelo de *Credit Scoring*, uma etapa fundamental para construção de *Collection Scoring* é segmentar o banco de clientes com base nos atrasos observados. A segunda fase consiste em definir os estágios de inadimplência: recente, média ou tardia. Vale ressaltar que essas definições dependem do produto de crédito e da instituição. A partir disso são definidos os conceitos de bom e mau.

Uma vez desenvolvido, o modelo é aplicado. Nesse estágio, a estratégia de cobrança é definida baseada no escore. A pontuação calculada combinada a outras variáveis, como saldo devedor, quantidade de contas em atraso e número de dias em atraso que guiam a execução e a severidade de cada ação a ser tomada para cada cliente. Por exemplo, se o escore indica baixa probabilidade de recuperação, é necessário maior controle e estímulos para o pagamento, isso implica em aplicação de ações mais rígidas e de forma rápida, para que seja possível antecipar o recebimento do que é permitido cobrar. Por outro lado, se o escore indica alta probabilidade de recuperação, o procedimento indicado é evitar atritos desnecessários com o cliente, para que permaneça ativo e fidelizado. Caso o escore indique probabilidade de recuperação moderada, busca-se a harmonia entre a severidade e a brandura,

esforçando-se para a manutenção do cliente na carteira, analisando as causas da inadimplência e ofertando condições de renegociação.

Em vista do exposto, existem muitas vantagens em implementar os princípios de *scoring* às práticas de cobrança, tanto em relação ao negócio, como ao risco e ao atendimento à regulamentação. Alguns benefícios serão descritos a seguir.

Prevenção e controle da inadimplência, com ofertas para clientes adimplentes, mas com alta probabilidade de tornar-se inadimplente, aumentando o fluxo de caixa por meio da identificação prévia de futuros devedores, minimizando atrasos e subsequente perda.

Outro benefício diz respeito à gestão de políticas de crédito, autorizações e bloqueios, além do gerenciamento das estratégias de cobrança por meio de diversos canais como:

- Cobrança Interna: tratamento dos primeiros ciclos de atraso, buscando a recuperação do cliente e sua retenção;
- Cobrança Externa: maximizar a eficiência por meio da especialização da cobrança nas agências prestadoras de serviços;
- Cessão da carteira: quando todos os esforços foram aplicados e o custo de continuar mantendo o crédito em atraso é maior do que a receita e os benefícios da venda.

Entre as principais utilidades do escore no processo de cobrança está a redução de custos, por meio de automação, diminuição dos procedimentos manuais e centralização do controle, além de redução de ações desnecessárias, uma vez que grupos

diferentes necessitam de estratégias diferentes. As soluções são personalizadas, de acordo com tendências históricas e padrões de comportamentos, e não genéricas.

Por exemplo, existem clientes conhecidos como *self cure*, ou seja, são clientes inadimplentes que regularizam sua situação de atraso sem a necessidade de lembretes ou ações de cobrança. Para esse perfil não é necessário gastar esforços com cobrança, o pagamento é espontâneo, o que evita desgastes e constrangimentos com clientes, aumentando a satisfação do relacionamento com a instituição, indicando uma estratégia de ganha-ganha.

Os Modelos de *Collection* também podem ser utilizados para monitorar e administrar a carteira porque além de identificar clientes em risco, são capazes de reconhecer consumidores curados, gerando aumento de receita e de relacionamento, reduzindo a rotatividade e o cancelamento. E isso é muito importante para as instituições, uma vez que manter clientes é mais barato do que obter novos. Nesse sentido, o relacionamento de longo prazo ganha dimensão importante devido ao benefício mútuo.

Segundo Menck e Moriguchi (2009), do lado da empresa, o benefício do relacionamento está na economia encontrada em não ter que atrair e convencer clientes para novas transações. Do lado do cliente, o benefício do relacionamento está em não ter que pesquisar, conhecer, avaliar e incorrer em riscos a cada transação com um produto novo para ele.

Desse modo, o fato do cliente tornar-se inadimplente não significa que o ciclo de crédito e a gestão do risco terminaram. Pelo contrário, entende-se que a gestão

da cobrança tornou-se um diferencial competitivo, que proporciona a reabilitação e consequente redução de *Churn* (cancelamento).

Outra grande vantagem dos Modelos de *Collection Scoring* refere-se à contribuição dada pela probabilidade de recuperar o montante devido para a estruturação de réguas de cobrança. Isso proporciona melhor alocação de recursos, aumentando a taxa de recuperação, com regras eficientes de priorização e aplicação de ações mais efetivas.

As réguas de cobrança são estruturas fixas de ações a serem aplicadas aos clientes inadimplentes, com o objetivo de definir estratégias de cobrança diferenciadas a partir do perfil dos devedores, como focar em clientes inadimplentes com alta probabilidade de inativação, intensificar a cobrança dos devedores com baixa probabilidade de pagamento e definir o momento correto de enviar a cobrança aos escritórios terceirizados.

A Figura 2.2 apresenta um esquema de ciclo de cobrança com exemplos de régua de cobrança que combinam escore do Modelo de *Collection* com atraso e saldo devedor.

Como é possível perceber, os custos com cobrança aumentam ao caminhar pela régua, podendo chegar a um período de cobrança não rentável, nesse caso, é indicado estabelecer descontos negociais ou venda da carteira, para melhores chances de sucesso.

Adicionalmente, Modelos de *Collection Scoring* estão em conformidade com os princípios de Basileia, por serem ferramentas de gestão de risco, e atendem à Re-

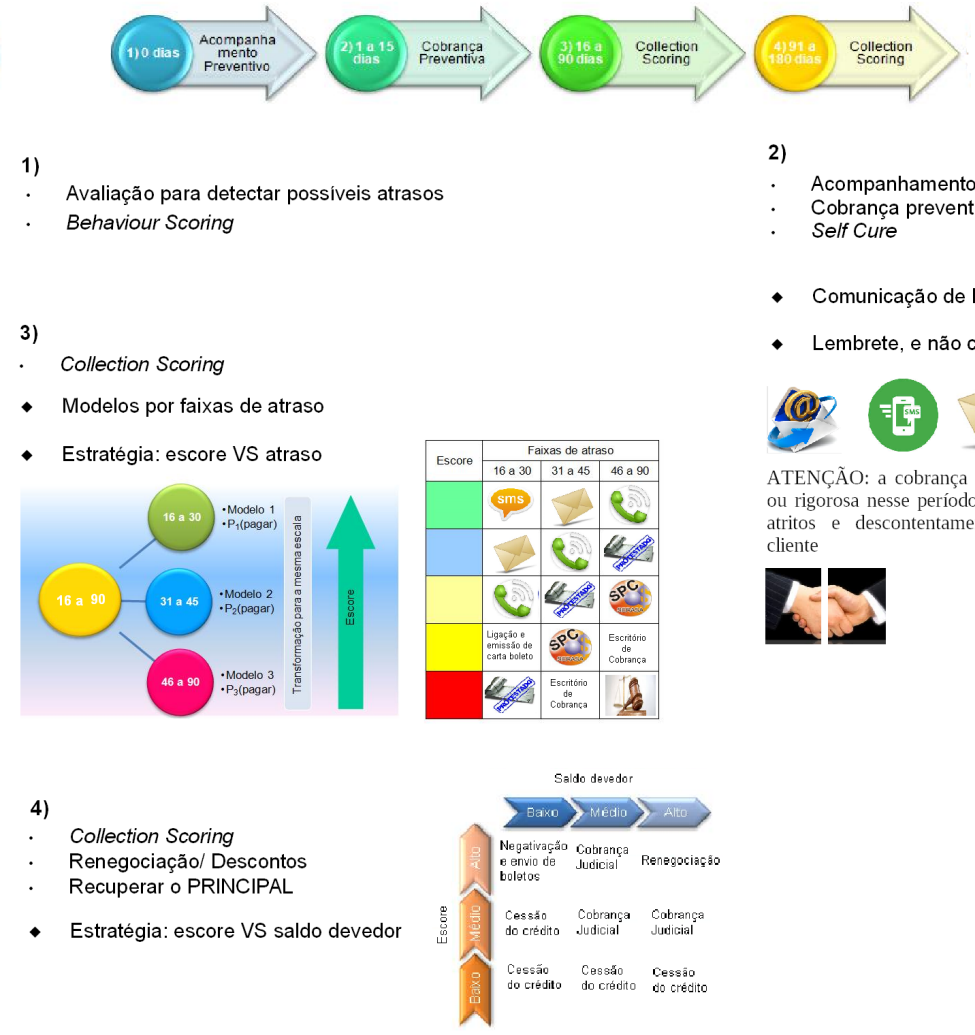


Figura 2.2: Estratégias de Cobrança

solução 2.682/99, uma vez que podem estabelecer critérios para classificação das operações, diminuindo os valores de provisão para créditos de liquidação duvidosa (PDD). Isso é possível porque os escores de cobrança colaboram para alcançar a receita máxima dos bons e reduzir o impacto dos maus nas contas de provisão, diminuindo a exposição da instituição aos riscos e contribuindo para as estratégias de priorização conforme valores esperados de recuperação.

Assim, questões como efetividade da cobrança internamente ou por agências, quais as dívidas enviar à cobrança, quando enviar os inadimplentes às agências de

cobrança, priorização de ações, qual a régua de cobrança mais eficiente, para quais clientes oferecer acordo, quando fazer cessão do empréstimo inadimplente e como medir a eficiência dos processos de cobrança são questões do negócio que podem ser auxiliadas com a implantação de um gerenciamento estratégico da carteira de inadimplentes que ofereça ferramentas de negociação que proporcionam o registro de informações de transações históricas e de relacionamento com o cliente, a efetivação de acordos com valores ótimos que maximizam os resultados e a melhor compreensão da dinâmica de rentabilidade por grupo de produto/cliente e canal/segmento.

Além disso, Modelos de *Collection Scoring* são constituídos sob a forma de um sistema científico e matemático, possibilitando uma avaliação consistente e impessoal, de forma que as decisões sejam as mesmas para grupos semelhantes de clientes. A previsão dos pagamentos é baseada em dados históricos, o que fornece mais acurácia ao sistema, uma vez que, conforme SERASA, o maior indicador de tendência de como um cliente irá lidar com suas finanças futuras é vista pela forma como eles lidaram em situações financeiras anteriores.

Entretanto, assim como qualquer tipo de modelo de *Credit Scoring*, os *Collections* apresentam alguns desafios que devem ser tratados com atenção. O principal é que eles se baseiam em dados históricos, o que nem sempre está atualizado ou disponível, podendo também apresentar séries curtas devido à recentidade dos contratos. Outro desafio consiste na degradação ao longo do tempo, em que o escore original já não discrimina com precisão os indivíduos. Isso pode ser provocado por uma série de fatores, incluindo alterações na população tomadora de financiamentos, nas con-

dições econômicas, e no caso de modelos de cobrança, no perfil dos inadimplentes devido à própria aplicação do modelo. Os modelos também sofrem influências de choques políticos, econômicos, fiscais e jurídicos que podem afetar a capacidade de pagamento dos clientes. Assim, é essencial que eles sejam revistos regularmente.

É importante destacar que Modelos de *Collection Scoring* são ferramentas fundamentais para gestão da carteira de inadimplentes. Mas esses modelos não funcionam sozinhos, é necessária uma caracterização estratégica de todo o processo de cobrança, que envolve: segmentações baseadas em várias dimensões, como, por exemplo, escore, dias em atraso e saldo devedor, definição dos segmentos baseados em metodologias de otimização e utilização de métodos desafiantes para refinamento dos atuais.

Capítulo 3

Análise de Sobrevivência

A Análise de Sobrevivência consiste em uma classe de métodos estatísticos para análise de dados cujo objeto de interesse é o tempo até a ocorrência de determinado evento de interesse. Originalmente, essa é uma técnica que surgiu na área médica para estudos sobre morte, mas que ganhou aplicações em diversos setores, como sociologia (análise histórica de eventos), engenharia (análise de confiança, análise de tempo de falha, tempo de vida de equipamentos) e economia (análise de duração e transição).

Segundo Allison (1995), um evento é uma mudança qualitativa que pode estar situada no tempo, ou seja, é a transição de um estado para outro. Devido às suas origens no campo da saúde, o evento de interesse geralmente estava ligado à morte, dessa forma, associado também à utilização na engenharia, o evento de interesse é denominado falha.

O conjunto de técnicas em Análise de Sobrevivência vem ganhando notoriedade em diversos campos porque além da informação de quem experimenta ou não o evento de interesse, ela é capaz de estimar quando a mudança ocorre. Ou seja, é possível situar o evento no tempo.

Assim, a observação dos indivíduos começa em um ponto bem definido do tempo e é acompanhada por algum período, sendo os tempos nos quais os eventos de interesse ocorrem registrados.

Neste trabalho, a Análise de Sobrevivência será utilizada para estimar um modelo de *Collection Scoring*. Desse modo, para aqueles clientes inadimplentes, deseja-se prever não apenas se eles regularizarão suas dívidas em determinado período, mas qual o tempo necessário para que esse evento ocorra.

É natural supor que pessoas que pagam o valor em atraso uma semana após o descumprimento têm, em média, uma propensão maior de regularizar a situação de inadimplência do que aquelas que não pagam até 60 dias, por exemplo. E ignorar essa informação pode reduzir a precisão das estimativas.

Análise de Sobrevivência tem duas características que dificultam a avaliação com métodos estatísticos convencionais: presença de censura e de variáveis explicativas que podem variar em qualquer ponto durante o período de observação.

Segundo Colosimo e Giolo (2006), a principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta, ou seja, existe alguma informação sobre o tempo de sobrevivência, mas não o conhece-se exatamente. Contextualizando, isto diz respeito a ocasiões em que o acompanhamento do inadimplente é interrompido, por exemplo porque o tempo de observação terminou para a análise de dados, significando que toda a informação sobre a resposta se resume ao conhecimento de que o tempo de falha - nesse caso, de recuperação - é superior ao período observado. Entretanto, essa informação, apesar de incompleta,

é útil e importante para a modelagem, uma vez que sua omissão pode acarretar em estimativas viciadas.

Em estudos de Análise de Sobrevivência, podem existir diversas formas de censura, e isso ocorre devido a vários fatores. Os tipos mais comuns são:

- Censura à direita: ocorre quando tudo que se sabe sobre o tempo do evento é que ele é maior que algum valor c, ou seja, a observação termina antes que o indivíduo experimente o evento de interesse;

    1. Censura do tipo I: o tempo do final do estudo é fixo e determinado pelo pesquisador, além do que todas as observações têm o mesmo tempo de censura;

    2. Censura do tipo II: ocorre quando a coleta de informações termina depois que um número específico de eventos ocorre. Esse tipo de estudo não é comum em ciências sociais;

    3. Censura aleatória: nesse caso, o evento de interesse não pode ser observado por razões fora do controle do pesquisador. Também pode ser produzida quando existe um tempo único de término do estudo, mas os tempos de entrada variam aleatoriamente entre os participantes.

- Censura à esquerda: nesse caso, a única informação sobre o tempo é que ele ocorreu antes de determinado valor. É comum em estudos nos quais a observação dos indivíduos começa após uma amostra já ter experimentado o evento de interesse;

- Censura intervalar: combina censura à direita e à esquerda. Uma observação de uma variável T apresenta censura intervalar se tudo que se sabe é que $a < T < b$, para algum valor de a e b.

Nem todos os dados de sobrevivência são censurados, bem como censuras podem acontecer em outras aplicações, mas como em Análise de Sobrevivência é muito comum, um tratamento especial é necessário.

A primeira particularidade diz respeito a representação dos dados de sobrevivência para um indivíduo i, que, em geral, é dada pelo par (t_i, δ_i), sendo

- t_i: tempo de observação do indivíduo i

- δ_i: uma variável indicadora de ocorrência do evento ou censura, isto é,

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de ocorrência do evento de interesse} \\ 0, & \text{se } t_i \text{ é um tempo censurado} \end{cases}$$

Assim, a variável resposta em Análise de Sobrevivência é representada por duas colunas no banco de dados, uma variável que contém o tempo no qual o evento ocorreu ou, em caso de censura, o último tempo em que o caso foi observado, ambos medidos desde o tempo de origem. Uma segunda variável é necessária quando há casos censurados ou se você deseja identificar diferentes tipos de eventos, ou seja, trata-se de uma variável indicadora. Quando há apenas um tipo de evento, é comum uma variável binária, em que 1 significa casos não censurado e 0 representa censura.

Segundo Diniz e Louzada (2013), na Análise de Sobrevivência, o comportamento da variável aleatória tempo de sobrevivência, $T \geq 0$, pode ser expresso por meio de várias funções equivalentes, tais que, se uma delas é especificada, as outras podem

ser derivadas. Essas funções são utilizadas para descrever diferentes aspectos do tempo de sobrevivência, que pode ser discreto ou contínuo. Estudos que tratam a variável tempo de sobrevivência como uma variável discreta podem ser vistos em Nakano e Carrasco (2006), Carrasco et al (2012) e Brunello e Nakano (2015).

No caso em que T é uma variável aleatória contínua, a função densidade de probabilidade de T pode ser interpretada como o limite da probabilidade de observar o evento de interesse em um intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo e é denotada por $f(t)$ da seguinte maneira:

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

A função de sobrevivência é definida como a probabilidade do indivíduo não falhar (ou não experimentar o evento de interesse) até o tempo t, ou seja, é a probabilidade dele sobreviver ao tempo t. Logo, pode-se descrevê-la como

$$S(t) = P(T > t) = 1 - F(t),$$

em que $F(t) = \int_0^t (u)du$ é a função de distribuição acumulada, $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$. Conforme Collet (1994), a função de sobrevivência é monótona decrescente, isto é, $S(u) \geq S(v)$ para $u < v$.

A função de risco, ou taxa de falha, ou *hazard function* no intervalo $[t1, t2)$ é definida como a probabilidade de que a falha ocora nesse intervalo, dado que não ocorreu antes, dividida pelo comprimento do intervalo. Logo, pode-se expressá-la como

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

Definindo intervalo genérico $[t, t + \Delta t)$ e assumindo Δt pequeno, temos a expressão da taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t, dada por

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

Logo, de maneira geral, a função taxa de falha (ou de risco) de T pode ser escrita como

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

estabelecendo assim o relacionamento entre a função densidade de probabilidade, a função de sobrevivência e a função de risco, todas matematicamente equivalentes para descrever o comportamento da variável aleatória tempo de sobrevivência.

Devido a sua interpretação, a função de risco é preferida por muitos autores, uma vez que ela descreve como a probabilidade instantânea de falha se modifica com o passar do tempo.

Outra função importante é a de risco acumulada, $H(t)$, expressa por

$$H(t) = \int_0^t h(u) du.$$

Em virtude da equivalência entre as expressões matemáticas para as distribuições de T, pode-se escrever

$$f(t) = \frac{\partial F(t)}{\partial t} = \frac{\partial [1 - S(t)]}{\partial t} = -S'(t).$$

Dessa forma, a função risco pode ser dada por

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{\partial [\log S(t)]}{\partial t}.$$

Assim

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

Equivalentemente

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp[-H(t)].$$

E por fim, para completar a relação entre as três funções utilizadas para descrever os dados em análise de sobrevivência, pode-se expressar

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right).$$

Assim como qualquer técnica de regressão, os modelos em Análise de Sobrevivência buscam identificar a relação e a influência das covariáveis com os tempos de sobrevivência. No contexto deste trabalho, para o gerenciamento estratégico da carteira de inadimplentes, o objetivo é estimar o efeito das covariáveis sobre o tempo de sobrevivência, ou melhor, sobre o tempo de falha - que no caso refere-se à regularização do saldo devedor.

Segundo Colosimo e Giolo (2006), existem duas categorias de modelos de regressão em Análise de Sobrevivência, os paramétricos e os semiparamétricos. A primeira classe tem associada uma distribuição de probabilidade à variável aleatória T, que geralmente é tratada por meio de um elemento determinístico não linear nos parâmetros e uma distribuição assimétrica para o comportamento estocástico. Já os semiparamétricos são também conhecidos como Modelo de Cox e caracterizam-se pela flexibilidade, uma vez que não é necessário fazer qualquer suposição sobre a variável aleatória, e a facilidade em incorporar covariáveis dependentes do tempo.

Neste caso, o efeito das covariáveis é estimado por meio da proporcionalidade dos riscos ao longo do tempo de acompanhamento.

Neste trabalho, será desenvolvido um *Collection Scoring* por meio do modelo de Riscos Proporcionais de Cox, como um ensaio para a utilização da Análise de Sobrevivência na construção de modelos de cobrança. O modelo de Cox foi escolhido devido a sua flexibilidade e simplicidade. Nesta proposta de modelagem não foram utilizadas covariáveis dependenes no tempo. As variáveis explicativas foram observadas no momento em que o cliente atingiu atraso superior a 90 dias (e em alguns casos, observou-se os valores defasados, ou seja, um, dois, três ou quatro meses antes do descumprimento).

3.1 Estimador de Kaplan-Meier

Conforme descrito anteriormente, a presença de censuras em dados de sobrevivência dificultam análises com métodos estatísticos convencionais. Para o estudo descritivo, medidas de tendência central e de variabilidade podem provocar interpretações erradas sobre os dados, invalidando esse tipo de diagnóstico.

Desse modo, a função de sobrevivência é o principal instrumento para análise preliminar dos dados, para estimar quantidades de modelos de regressão (como tempos médios ou medianos, percentis ou frações de falhas em tempos fixos) e para avaliar o ajuste de modelos.

Uma das técnicas mais utilizadas para estimar a função de sobrevivência consiste no estimador de Kaplan-Meier. Também conhecido como estimador produto-limite, esse método tem sido utilizado por muitos anos, desde 1958, quando Kaplan e Meier

mostraram que ele é um estimador de máxima verossimilhança não paramétrico de $S(t)$. Isso deu ao método uma sólida justificativa teórica.

Quando todos os dados são não censurados, o estimador de Kaplan-Meier é definido por meio de um ajuste na função de sobrevivência empírica, sendo calculado de maneira simples e intuitiva. Na ausência de censuras, a função de sobrevivência empírica é definida como a proporção:

$$\hat{S}(t) = \frac{\sharp\text{de observações que não falharam até o tempo t}}{\sharp\text{total de observações no estudo}}, \tag{3.1}$$

em que $\hat{S}(t)$ é uma função escada, e se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

Assim, Kaplan e Meir modificaram a proporção (3.1) na presença de censuras para a construção do estimador. Esse estimador considera a quantidade de intervalos igual ao número de falhas distintos, e os limites dos intervalos de tempo são os instantes de falha da amostra.

A expressão geral do estimador de Kaplan-Meier é construída com base nas seguintes suposições:

- sejam $t_1 < t_2 < ... < t_k$ os k tempos distintos e ordenados de falha
- defina d_j como sendo o número de falhas em t_j e
- n_j o número de clientes sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Assim, o estimador de Kaplan-Meier é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{3.2}$$

Em seu artigo original, Kaplan e Meier demonstram que a expressão (3.2) é o estimador de máxima verossimilhança de $S(t)$. Dessa forma, a função de verossimilhança pode ser escrita por:

$$L(S(.)) = \prod_{j=0}^{k} \left\{ [S(t_j) - S(t_{j+})]^{d_j} \prod_{l=1}^{m_j} S(t_{jl+}) \right\}.$$

Os passos para a demonstração dessa fórmula são dados assumindo:

- suponha que d_j indivíduos falharam no tempo t_j, para $j = 1, ..., k$;
- seja m_j a quantidade de censuras no intervalo $[t_j, t_{j+1})$, nos tempos $t_{j1}, ..., t_{jm_j}$;
- então, a probabilidade de falha no tempo t_j é $S(t_j) - S(t_{j+})$, em que $S(t_{j+}) = \lim_{\Delta t \to 0+} (S(t_j + \Delta t))$;
- defina a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_{jl}, para $l = 1, ..., m_j$ como $P(T > t_{jl}) = S(t_{jl+})$.

Apesar de intuitivo, o estimador de Kaplan-Meier é robusto, uma vez que é não viciado para grandes amostras, fracamente consistente e converge assintoticamente para a normal.

As propriedades de consistência e normalidade assintótica foram provadas por Breslow e Crowley (1974) e Meier (1975). A variância assintótica do estimador de Kaplan-Meier é calculada pela fórmula de Greenwood, a partir das propriedades do estimador de máxima verossimilhança, assim:

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

Dessa maneira, como para t fixo, $\hat{S}(t)$ é assintoticamente normal, então um intervalo aproximado de $100(1-\alpha)\%$ de confiança para S(t) é dado por:

$$\hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{S}(t))},$$

em que $\alpha/2$ representa o percentil da Distribuição Normal Padrão.

Para valores extremos de t, pode haver problemas com os intervalos de confiança, porque eles podem não estar entre 0 e 1. Mas uma solução usual foi dada por Collett (1994), em que ele recomenda o cálculo do intervalo para a transformação $\log(-\log \hat{S}(t))$ e posteriormente a conversão dos limites para a métrica original.

Uma vez estimada a curva de sobrevivência, sua utilização é direta para o cálculo da probabilidade de sobreviver a um determinado tempo. Também é possível obter percentis, tempo médio e mediano de vida, além de comparar grupos.

A estimativa do tempo médio, entretanto, deve ser analisada com cuidado, quando há muitas censuras, porque nesse caso, ele é subestimado. Nesses casos, a mediana é preferida.

Dado que a função de sobrevivência fornece um completo conhecimento sobre a experiência de sobrevivência de cada grupo, uma aproximação natural para comparar amostras é testar $H_0 : S_1(t) = S_2(t)$. Para tal finalidade, poderiam ser utilizados os testes de Log-rank e Wilcoxon.

3.2 Modelo de Cox

Conforme Colosimo e Giolo (2006), o modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o

tempo até a ocorrência de um evento de interesse, ajustado por covariáveis.

Esse modelo é utilizado com frequência em estudos de sobrevivência devido a sua versatilidade. Fundamentado na suposição de que os riscos sejam proporcionais, a regressão de Cox não requer a escolha de uma distribuição de probabilidade para os tempos de sobrevivência, por isso é considerado um modelo robusto.

Outras razões tornam a regressão de Cox atrativa, como a possibilidade de trabalhar com covariáveis dependentes no tempo, análises estratificadas para controle de variáveis com ruídos, além de funcionar para medidas de tempo discreta e contínua.

Em seu artigo original, Cox (1972) propôs dois conceitos inovadores, o primeiro referente a um modelo de riscos proporcionais (o que depois foi generalizado para riscos não proporcionais), e o novo método de estimação, denominado de máxima verossimilhança parcial.

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

$$h(t) = h_0(t)g(\boldsymbol{x'\beta}), \tag{3.3}$$

onde g é uma função não negativa, em que $g(0)$ é igual a 1.

Dessa maneira, o Modelo de Cox é definido como o produto entre dois fatores, um paramétrico e outro não paramétrico, por isso ele também é denominado de Modelo Semiparamétrico.

O componente não paramétrico é usualmente chamado de função de risco base ou basal, pois $h(t) = h_0(t)$, quando $\boldsymbol{x} = \mathbf{0}$, ou seja, $h_0(t)$ pode ser considerada como a taxa de falha de um indivíduo para o qual todas as covariáveis têm valor zero. A

função base é não especificada, mas com a exigência de ser uma função não negativa no tempo.

Por outro lado, a parte paramétrica é uma função positiva e contínua das covariáveis. Apesar de existirem outras formas na literatura para essa componente, ela é comumente escrita na forma exponencial, pelo fato de ser sempre positiva, da seguinte maneira:

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

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

Por ser um componente linear do modelo, convencionou-se escrever o somatório $\beta_1 x_1 + ... + \beta_p x_p$ como preditor linear ou escore, que na forma matricial, é dado por $\eta = \boldsymbol{x'\beta}$.

É importante notar que a constante β_0, presente nos modelos paramétricos, não aparece na função $g(\boldsymbol{x'\beta})$. Isso ocorre devido à presença do componente não paramétrico no modelo que absorve esse termo constante.

A expressão do modelo em (3.3) implica que a razão das taxas de falha ou de risco entre dois indivíduos é constante ao longo do tempo, sendo uma função apenas das covariáveis, conforme pode ser visto em (3.4).

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\boldsymbol{x_i'\beta})}{h_0(t)\exp(\boldsymbol{x_j'\beta})} = \exp(\boldsymbol{x_i'\beta - x_j'\beta}). \tag{3.4}$$

Devido a essa razão, o Modelo de Cox também é conhecido como Modelo de Riscos Proporcionais.

Apesar de bastante flexível, devido ao componente paramétrico, a suposição básica de taxas de falha proporcionais não pode ser violada para a correta utilização

do Modelo de Cox.

Para a avaliação da proporcionalidade dos riscos, podem ser empregadas técnicas gráficas e testes estatísticos.

3.3 Ajuste do Modelo de Cox

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

Desse modo, é necessário um método de estimação que permita a construção de inferências sobre os parâmetros do modelo. O método da máxima verossimilhança frequentemente utilizado não pode ser empregado, uma vez que se torna inapropriado devido à presença do componente não paramétrico. Assim, Cox propôs um novo método de estimação: a máxima verossimilhança parcial, a partir do qual é possível estimar os coeficientes das covariáveis sem ter que especificar a função base $h_0(t)$.

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

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

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

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

Em que $R(t_i)$ representa o conjunto dos índices das observações sob risco em t_i. Ou seja, Cox propôs a utilização do registro histórico passado de falhas e censuras em forma de probabilidade condicional para eliminar o termo não paramétrico da função de verossimilhança.

Assim, a função de máxima verossimilhança parcial é dada por:

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

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

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

O logaritmo dessa função de verossimilhança é dado por:

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{\beta' x_i} - \log \sum_{j \in R(t_i)} \exp(\boldsymbol{x_j'}\boldsymbol{\beta}) \right). \tag{3.5}$$

As estimativas de verossimilhança dos parâmetros $\beta's$ são obtidos maximizando-se (3.5), ou seja, resolvendo o sistema de equações definido $U(\boldsymbol{\beta}) = \mathbf{0}$, em que U é o vetor escore formado pelas primeiras derivadas de $L(\boldsymbol{\beta})$.

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

Assim, o termo regressão de Cox refere-se a combinação do modelo e do método de estimação.

Esse método de estimação possui duas das três propriedades padrões das estimativas de máxima verossimilhança, as quais pode-se citar: resultados consistentes e assintoticamente normais, ou seja, em grandes amostras, as estimativas são aproximadamente não viesadas e sua distribuição amostral é aproximadamente normal.

Tanto o modelo de riscos proporcionais, como a função de verossimilhança parcial assumem que os tempos de sobrevivência são contínuos. Nesse contexto, empates nos valores observados não seriam possíveis. Porém, dados empatados podem acontecer na prática devido a escalas de medidas, ao processo de coleta dos dados, arredondamentos e aproximações e a ocorrência de mais de um evento em um mesmo instante de tempo. Também é possível haver empates entre observações censuradas, e entre falhas e censuras. Isto posto, são necessárias adequações à função de verossimilhança.

A função de verossimilhança exata, quando há empates, foi proposta por Kalbfleisch e Prentice (1980), mas isso exige um esforço computacional grandioso. Dessa forma, foram apresentadas diversas modificações para o tratamento apropriado dos dados empatados, entre elas, as propostas por: Breslow (1972) e Peto (1972), Efron (1977), Farewell e Prentice (1980).

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

3.4 Interpretação dos Coeficientes

O coeficiente de uma covariável no modelo semiparamétrico de Cox pode ser interpretado como o logaritmo da razão de risco do evento de dois indivíduos com atributos diferentes para uma variável específica.

Para exemplificar, considere um modelo com apenas uma variável contínua a. As funções de risco para dois indivíduos i e j, respectivamente são:

$$h_i(t) = \exp(\hat{\beta}a_i)h_0(t) \quad \text{e} \quad h_j(t) = \exp(\hat{\beta}a_j)h_0(t).$$

Considere $a_i = x + 1$ e $a_j = x$. Então, a razão das taxas de falha para os indivíduos i e j pode ser escrita por:

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp[\hat{\beta}(x+1)]h_0(t)}{\exp[\hat{\beta}(x)]h_0(t)} = \exp(\hat{\beta}).$$

Dessa forma, podemos assumir que o risco de se observar o evento de interesse para o indivíduo que assume o valor da variável $a = x + 1$ é $\exp(\hat{\beta})$ vezes o risco para aqueles com $a = x$.

Essa é a demonstração quando a variável explicativa é acrescida de uma unidade. Generalizando para quando x é acrescido de y unidades, tem-se que a taxa de falha é $\exp(y\hat{\beta})$ vezes maior.

Para variáveis categóricas classificadas em m níveis, assume-se que determinado grupo é referência e compara-se os demais com aquele.

3.5 Estimação da Função de Risco e Sobrevivência

Dado um Modelo de Cox com o vetor $\boldsymbol{x}$ de covariáveis de dimensão p e as res-

pectivas estimativas dos coeficientes, então a função taxa de falha para o i-ésimo indivíduo é dada por:

$$\hat{h}_i(t) = \hat{h}_0(t) \exp(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}),$$

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

Outras funções relacionadas a $h_0(t)$ são importantes, principalmente em análises gráficas para avaliar a adequação do modelo ajustado. Mas como $h_0(t)$ não é especificado parametricamente, outras técnicas são utilizadas para estimação.

A função de risco acumulada base pode ser estimada de forma simples, conforme proposta de Breslow (1972), em que uma função escada com saltos nos tempos distintos de falha é empregada da seguinte maneira:

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

em que d_j é o número de falhas em t_j.

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

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\} \quad \text{e} \quad \hat{S}(t|x) = [\hat{S}_0(t)]^{(\exp(\boldsymbol{x'}\hat{\boldsymbol{\beta}}))}$$

3.6 Avaliação do Modelo de Cox

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

Existem diversas técnicas gráficas e testes estatísticos disponíveis na literatura, neste trabalho será considerada a análise descritiva a partir dos Resíduos Padronizados de Schoenfeld. Os testes estatísticos não serão aplicados devido ao tamanho da amostra, que por ser grande, leva à rejeição da hipótese nula em qualquer caso.

Considere que o indivíduo i experimentou o evento de interesse, sendo observado o tempo de falha e o vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$. Então, o resíduo de Schoenfeld é definido como $\boldsymbol{r}_i = (r_{i1}, r_{i2}, ..., r_{ip})$, onde cada componente r_{iq}, para $q = 1, 2, , p$ é dado por:

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

Esse resíduo não é definido para censuras, apenas para tempos de falha. Entretanto, essa medida definida dessa forma é pouco utilizada, uma vez que não considera a estrutura de correlação entre os resíduos.

Assim, foi desenvolvido um ajuste frequentemente utilizado denominado Resíduos Padronizados de Schoenfeld. Nesse caso é necessário utilizar a matriz de informação observada como efeito multiplicativo ao resíduo simples, da seguinte maneira:

$$\boldsymbol{s}_i^* = [I(\hat{\boldsymbol{\beta}})]^{-1} \boldsymbol{r}_i.$$

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

Capítulo 4

Regressão Logística

A Regressão Logística é um caso particular de Modelo Linear Generalizado (MLG) desenvolvido por volta de 1960 com o objetivo de realizar predições e estudar a relação entre uma variável aleatória binária (variável dependente) e um conjunto de variáveis independentes.

Apesar da Regressão Logística ter surgido e se desenvolvido na área médica, a sua aplicação se expandiu rapidamente por muitos campos, devido à facilidade e capacidade em explicar a ocorrência de determinados eventos. Além disso, o número de suposições necessárias para a aplicação da técnica é pequeno.

Sob esse aspecto, as restrições necessárias para aplicabilidade da Regressão Logística são:

- valor esperado igual a zero para os resíduos;
- erros não correlacionados;
- variáveis independentes e erros não correlacionados;
- ausência de multicolinearidade perfeita entre as variáveis explicativas.

O objetivo da Regressão Logística é gerar uma função matemática cuja resposta permita estabelecer a probabilidade de uma observação pertencer a um grupo previamente determinado, em razão do comportamento de um conjunto das variáveis independentes. Em modelos de *Credit Scoring*, essa probabilidade representa a chance do tomador de crédito se tornar adimplente ou inadimplente, a depender daquilo que é definido como evento de interesse. Já em modelos de *Collection Scoring*, essa probabilidade representa a chance do cliente regularizar os valores de dívidas em atraso.

Agresti (1990) define MLG como um modelo linear para transformação da esperança de uma variável aleatória cuja distribuição pertence à família exponencial.

Segundo McCullagh e Nelder (1989), um MLG é composto por três elementos fundamentais: um componente aleatório, um componente determinístico ou sistemático e uma função de ligação.

O componente aleatório consiste na variável dependente Y que se deseja modelar, da qual se coletam n observações independentes e cuja distribuição de probabilidades deve pertencer à família exponencial. Mais detalhes sobre a família exponencial podem ser encontrados em Casella e Berguer (2010).

O componente determinístico é definido por um vetor $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ que consiste em uma combinação linear da forma $\boldsymbol{\eta} = \boldsymbol{X\beta}$, onde $\boldsymbol{X}$ é uma matriz de ordem nxp de variáveis independentes (preditoras) e $\boldsymbol{\beta}$ é o vetor p-dimensional de parâmetros desconhecidos do modelo.

A função de ligação $g(.)$ é uma função diferenciável e monótona que associa os

valores esperados das observações (componente aleatório) com as variáveis independentes (componente sistemático). Suponha que uma variável aleatória binária Y_i segue uma distribuição de Bernoulli e assume os seguintes valores:

$$Y_i = \begin{cases} 1, & \text{se o cliente regulariza os créditos em atraso} \\ 0, & \text{se o cliente não regulariza os créditos em atraso} \end{cases}$$

Seja $\boldsymbol{x_i} = (1, x_1, x_2, ..., x_p)^{'}$ o vetor de características do cliente i e $\pi(\boldsymbol{x_i})$ a proporção de clientes que se recuperam em função do perfil dos clientes, a esperança e variância de Y_i são dadas por:

$$E(Y_i) = \pi_i \text{ e } Var(Y_i) = \pi_i(1 - \pi_i).$$

Dado que a distribuição Bernoulli pertence à família exponencial, aplicando MLG utilizando a função *logit* como função de ligação temos:

$$g(E(Y_i)) = g(\pi(\boldsymbol{x_i})) = \ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_0 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i} = \boldsymbol{x_i^{'}\beta}.$$

Podendo também ser escrito da forma:

$$E(Y_i) = \pi_i(\boldsymbol{X}) = \frac{\exp(\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i})}{1 + \exp(\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i})} \quad \text{onde} \quad 0 \leq \pi_i(\boldsymbol{X}) \leq 1.$$

E sua estimativa será

$$\hat{\pi_i} = \frac{\exp(\boldsymbol{x_i^{'}\hat{\beta}})}{1 + \exp(\boldsymbol{x_i^{'}\hat{\beta}})},$$

em que $\pi(\boldsymbol{x_i})$ pode ser interpretado como a probabilidade do i-ésimo cliente voltar à condição de adimplente.

4.1 Estimação dos Coeficientes

Os valores de $\boldsymbol{X}$ são conhecidos e os parâmetros são as únicas quantias desconhecidas que necessitam ser estimadas.

Em modelos de Regressão Logística, a estimativa dos parâmetros é realizada através do método da máxima verossimilhança (Hosmer Lemeshow, 2000).

Sabendo que os dados são oriundos de uma distribuição Bernoulli e uma vez que as observações do conjunto de dados são independentes, a Função de Verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}.$$

Pelo princípio do método da máxima verossimilhança, os valores estimados de $\boldsymbol{\beta}$ são aqueles que maximizam $L(\boldsymbol{\beta})$. Para obtenção desses valores, calcula-se a derivada dessa função em relação a cada um dos parâmetros e procura-se pelo ponto crítico onde a derivada é igual a zero.

Aplicando a transformação monotônica logaritmo natural (ln) à função de verossimilhança, em virtude da propriedade de que o logaritmo de um produto é igual à soma dos logaritmos dos fatores, se obtém:

$$\ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i \ln[\pi(\boldsymbol{x_i})] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x_i})].$$

Essa transformação é realizada para simplificar matematicamente o cálculo das derivadas, tendo em vista que os resultados da maximização das funções $L(\boldsymbol{\beta})$ e $\ln[L(\boldsymbol{\beta})]$ são exatamente os mesmos (Casella e Berger, 2010).

Dessa forma, diferenciando $\ln[L(\boldsymbol{\beta})]$ e igualando a zero obtém-se as equações de verossimilhança, que são expressões não lineares nos parâmetros e portanto, podem ser solucionadas via métodos numéricos iterativos, como por exemplo o método Newton-Raphson.

Os valores encontrados para $\boldsymbol{\beta}$ são chamados Estimadores de Máxima Verossimilhança (EMV) e indicam a importância de cada variável independente para a ocorrência do evento de interesse. (Sicsú, 2010).

Os estimadores possuem diversas características, dentre elas está o conceito de eficiência, que é relacionado com a variância do mesmo, onde infere-se como estimador mais eficiente o de menor variância. Um estimador é dito consistente quando o mesmo converge, em probabilidade, para o seu valor populacional quando o tamanho da amostra n tende para infinito, e não viesado quando a esperança do estimador é o seu valor populacional, ou seja, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Quando n tende a infinito (comportamento assintótico), visto na prática como n suficientemente grande, o estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ possui distribuição aproximadamente Normal com média $\boldsymbol{\beta}$ e variância tendendo para o limite inferior da desigualdade de Cramer-Rao. Além disso, ele é consistente já que $Var(\hat{\boldsymbol{\beta}}) \to \mathbf{0}$ quando $n \to \infty$ (Ehlers, 2009).

A significância dos estimadores pode ser testada através do Teste da Razão de Verossimilhança, que tem o intuito de comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável que se deseja testar. Outro teste que pode ser realizado é o Teste de Wald.

4.2 Interpretação dos Coeficientes

Na Regressão Logística, os coeficientes das variáveis independentes podem ter diversas interpretações, uma vez que eles influenciam o *logit* (logaritmo natural da razão de chance), a razão de chance e as probabilidades.

No *logit*, a estimativa do parâmetro indica a alteração na variável dependente por unidade de variação da variável independente. Ou seja, caso uma variável x_1 tenha coeficiente 10 e todas as outras permaneçam constante, então o acréscimo de uma unidade em x_1 implica no acréscimo de 10 no *logit*. Apesar de simples, em temos práticos, essa interpretação não apresenta nenhum significado intuitivo e nem melhora a qualidade da informação disponível.

Para a interpretação do coeficiente sob a razão de chance, basta fazer e^{β_i} para identificar o impacto dessa variável. Assim, o efeito dos coeficientes sobre a razão de chance é de natureza multiplicativa. Desse modo, um coeficiente igual a 0 significa que o efeito da variável na resposta é nulo, uma vez que a razão de chance fica multiplicada por 1. Isso posto, conclui-se que, coeficientes positivos contribuem para elevar a razão de chance e a probabilidade, e coeficientes negativos reduzem esses valores.

É importante destacar que a relação estabelecida entre as variáveis explicativas e a variável dependente no modelo logístico não é linear.

4.3 Avaliação do Modelo Logístico

Um dos principais mecanismos para avaliar um modelo de Regressão Logística é o *Log Likelihood Value*. Esse indicador tem o objetivo de verificar a capacidade de estimação geral do modelo. Quanto mais próximo de zero, melhor o grau de adequação do modelo.

Mas o *Log Likelihood Value* sozinho oferece pouca informação sobre a qualidade do modelo. Assim, outras medidas são importantes.

O teste de Hosmer e Lemeshow é outra estratégia que pode contribuir para a avaliação do modelo Logístico. Nesse caso, busca-se comparar os valores preditos com os observados. Dessa forma, caso haja diferenças significativas entre eles, conclui-se que o modelo não é capaz de produzir estimativas confiáveis.

Capítulo 5

Ajuste dos Modelos de *Collection Scoring*

5.1 Público Alvo

Modelos de cobrança têm por objetivo identificar a propensão a pagamento de clientes que já se tornaram inadimplentes, então o público alvo de um modelo de *Collection Scoring* é aquele composto por tomadores de crédito que não cumpriram com suas obrigações de pagamentos com as instituições credoras.

Uma base de dados reais será utilizada para ilustração e comparação de metodologias para desenvolvimento de um modelo de *Collection Scoring*. O conjunto de dados considerados para este estudo pertence a uma grande instituição financeira brasileira que atua em diversos segmentos do mercado de crédito. Por questões de sigilo da informação, não serão descritas características das operações envolvidas, tampouco as variáveis analisadas.

Foram observadas variáveis relacionadas ao perfil sociodemográfico do cliente, bem como aquelas de comportamento. Modelos de cobrança são construídos com a inclusão de variáveis que descrevem o relacionamento entre cliente e instituição,

adicionando poder de previsão ao modelo. Foi criada a variável status para indicar se o indivíduo experimentou o evento de interesse ou é censura.

Conforme exposto anteriormente, não é indicado desenvolver um modelo de *Collection Scoring* com toda a base de inadimplentes. A orientação é empregar a segmentação para criar modelos distintos para cada fase da inadimplência. A técnica de segregar a priori os clientes inadimplentes em faixas de atraso é um procedimento que melhora a eficiência dos modelos de cobrança, uma vez que as razões que provocam a chegada dos clientes em etapas crescentes de atraso e a regularização da dívida em cada uma delas é diferente. Ou seja, a utilização da segmentação permite que o modelo seja construído com base em populações que tenham relações homogêneas de previsão, já que apresentam comportamento semelhante. Assim, vários modelos de *Collection Scoring* podem ser construídos isoladamente e posteriormente os escores são alinhados e transpostos para uma mesma escala.

Dessa forma, os clientes inadimplentes foram segmentados conforme o estágio do atraso. Para este trabalho, será considerada a inadimplência tardia, definida em atrasos superiores a 90 dias.

Uma vez identificada a população de interesse, é necessário determinar as safras para a construção do modelo. A quantidade de meses escolhida deve ser suficiente para assegurar que o modelo tenha estabilidade temporal, não descalibrando facilmente, mas também não pode ser excessivamente longa, para que não interfira na acurácia do modelo em períodos mais atuais.

Portanto, para a construção do modelo de *Collection Scoring* a ser proposto neste

trabalho, foram considerados os clientes que descumpriram entre as safras de julho de 2010 a junho de 2011, considerando um ano de estudo, para evitar problemas com sazonalidade.

5.2 Variável Resposta

Independentemente do tipo de modelo de *Credit Scoring* a ser desenvolvido, o objetivo principal é distinguir entre bons e maus clientes, seja do ponto de vista de contratar ou não após uma campanha publicitária, ou apresentar determinado nível de atraso durante um período previamente fixado, ou até mesmo em regularizar créditos vencidos.

Entretanto, a regra utilizada para definição de bons e maus clientes não é rígida, mas depende da visão da empresa em cada tipo de modelo e do propósito que se pretende alcançar com a construção do mesmo. No caso de *Collections*, por exemplo, a performance de cada inadimplente depende daquilo que o departamento de cobrança da instituição entende como bom ou ruim em cada faixa de atraso.

Para o desenvolvimento do modelo de *Collection Scoring* que será apresentado neste trabalho considerou-se bom o cliente que recuperou no mínimo 80% do valor devido no momento do descumprimento.

Com a finalidade de verificar o percentual de recuperação de cada cliente, as quantias recebidas foram ajustadas ao valor no momento do descumprimento pela taxa de juros do contrato.

Inicialmente pensou-se em utilizar 24 meses para verificação da recuperação, entretanto não existiria informação de todas as operações para esse período completo.

Dessa forma, não haveria igualdade de condições na análise de todos os contratos. Outro problema é que não teria disponível safras para validação *out of time*.

Isto posto, cogitou-se considerar 12 meses para apuração dos valores pagos. Contudo, por se tratar de operações comerciais, avaliar um ano após o cliente atingir 90 dias de atraso seria muito tempo. Com as regras estabelecidas pela Resolução 2.682/1999, caso não houvesse pagamentos nesse período, o contrato do cliente seria lançado em prejuízo antes mesmo dos 12 meses.

Diante do exposto, estabeleceu-se observar o comportamento referente a pagamentos dos contratos inadimplentes por 9 meses. Desse modo, todos os contratos estariam em igualdade de condições e o período de análise de performance seria o limite para lançamento em prejuízo.

A Figura 5.1 ilustra os conceitos apresentados.

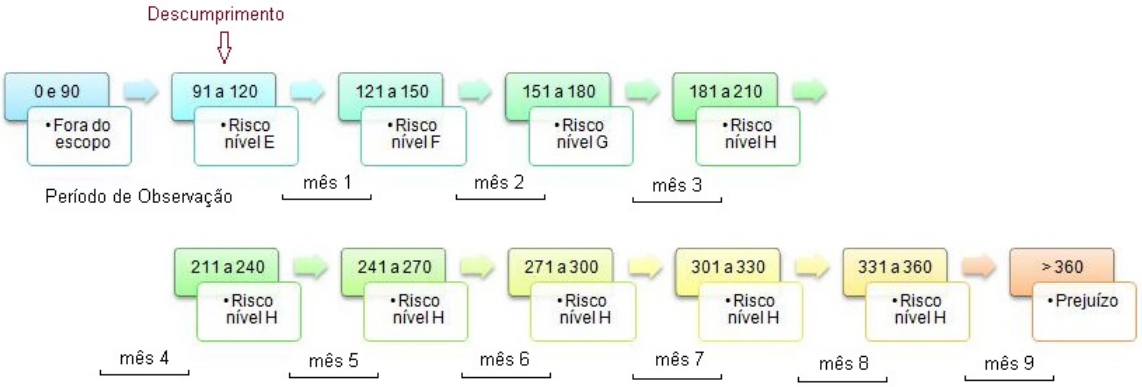


Figura 5.1: Período de Observação do Comportamento de pagamento para o Desenvolvimento do Estudo

Como este trabalho propõe o ajuste de um modelo Logístico e outro de Cox, a variável resposta em cada caso é definida como:

- Regressão Logística: classificação do cliente;

1. Bom: o cliente inadimplente recuperou pelo menos 80% do valor da dívida total no momento do descumprimento durante 9 meses de observação;

2. Mau: o cliente não recuperou;

- Modelo de Riscos Proporcionais de Cox: tempo para recuperação;

1. Tempo, em dias, necessário para a recuperação de pelo menos 80% do valor da dívida total no momento do descumprimento durante 9 meses de observação.

As Figuras 5.2 e 5.3 ilustram a variável resposta em cada caso. Para exemplificar, foram considerados três cenários: o primeiro considera vários pagamentos durante o período de observação, os demais, apenas um pagamento.

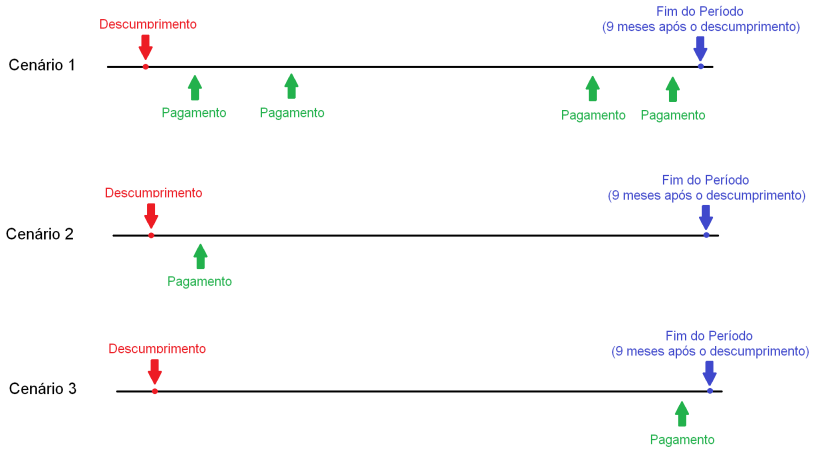


Figura 5.2: Variável Resposta - Regressão Logística

Para a Regressão Logística, não há distinção entre os cenários. Supondo que nos três casos os pagamentos foram suficientes para a recuperação de no mínimo 80% do valor devido, então a variável resposta é a mesma. Já para o Modelo de Cox, a variável resposta é distinta para os três cenários. Em temos de risco e de estratégias

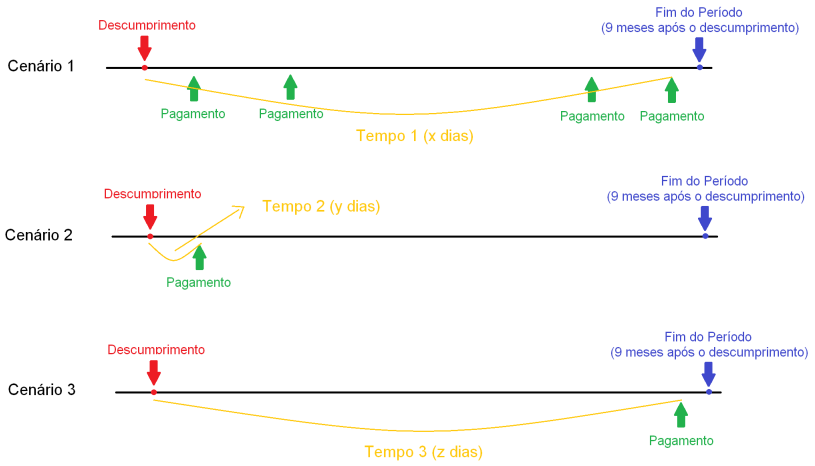


Figura 5.3: Variável Resposta - Modelo de Cox

de cobrança, o cenário 2 é bem diferente dos demais, e a Análise de Sobrevivência é capaz de captar essa diferença na medida em que considera o tempo necessário para a recuperação.

5.3 Análise Exploratória dos Dados

A análise exploratória dos dados consiste no diagnóstico univariado de cada variável explicativa e da variável resposta, com a finalidade de verificar as distribuições, detectar inconsistências, *outliers* e erros de preenchimento.

Outro objetivo dessa análise fundamenta-se na criação de novas variáveis. Sabe-se que a combinação de atributos, transformação de escala e relativização de variáveis absolutas são técnicas que produzem maior robustez aos dados.

Uma variável de valor remanescente, por exemplo, pode explicar bem o desempenho do cliente, mas a inflação, diferenças de classes sociais, condições macroeconômicas e termos do contrato tornam essa característica vulnerável. É diferente comparar dois clientes que devem R$500,00, se um deles tem salário de R$1.000,00

e o outro de R$10.000,00. Dessa forma, a importância de criar variáveis que sejam mais robustas, estáveis ao longo do tempo e com melhor poder explicativo.

Nesta etapa, foram analisadas mais de 100 variáveis, a fim de filtrar aquelas candidatas a serem utilizadas no modelo. A regra inicial utilizada para descartar aquelas não significativas foi baseada no preenchimento. Variáveis com 97% de *missing* foram eliminadas da análise.

O mais indicado para o desenvolvimento de *Collections Scoring* é que a base de clientes utilizada para a regressão não contenha indivíduos que sofreram alguma ação de cobrança, mas nem sempre isso é possível, ainda mais com relação a inadimplência tardia. Uma área de modelagem não consegue sustentar frente a alta administração das instituições que parte dos clientes devedores não devem ser cobrados, para que se observe um comportamento de pagamento espontâneo e assim, seja desenvolvido um modelo com melhor performance.

Portanto, como é difícil observar clientes que atingem 90 dias de atraso e permanecem por mais aproximadamente 9 meses sem nenhuma ação de cobrança, destaca-se a importância de desprezar variáveis correlacionadas com a utilização de outros modelos dentro da estratégia de cobrança, para que isso não influencie o seu escore, provocando alterações de ações dentro da régua de cobrança, tampouco atritos desnecessários com o cliente.

5.4 Análise Bivariada e Categorização das Variáveis

A etapa da análise bivariada é utilizada para verificar a relação entre as variá-

veis independentes e a variável resposta, com o objetivo de identificar associações que possam discriminar e influenciar a variável dependente, bem como comportamentos imprevisíveis e incomuns. Além disso, é nessa etapa que as variáveis são categorizadas de acordo com a entropia das classes em relação à variável resposta. Assim, categorias com características semelhantes são agrupadas no mesmo intervalo, mantendo uma tendência monótona crescente ou decrescente com a variável dependente

Nesta fase, também ocorre o tratamento dos valores *missings* e *outliers*, dado que são vinculados a alguma classe, uma vez que não é indicada a criação de categorias com baixo preenchimento.

Segundo Anderson (2007), a arte dos Modelos de *Credit Scoring* está na escolha sensata das categorias. Dessa forma, é importante utilizar técnicas robustas que dividam a variável de forma que o risco seja homogêneo dentro de cada categoria e heterogêneo entre elas. Assim, utilizou-se a medida de risco relativo para categorizar as variáveis.

As variáveis quantitativas foram ordenadas e divididas em decis. Em cada decil, verifica-se a frequência de bons e maus e o poder de discriminação por meio do risco relativo. Assim, as variáveis são analisadas com relação à variável resposta, sendo reagrupadas quando necessário (classes adjacentes com risco semelhante ou inversão do risco relativo).

Para as variáveis qualitativas, a metodologia para categorização consiste em analisar as frequências dos atributos e calcular o risco relativo de cada característica.

Categorias com risco semelhantes são agrupadas. Caso os atributos não possuíssem o mínimo de 5% de representatividade, qualidades conceitualmente semelhantes eram agrupadas e o risco relativo recalculado.

Nesta etapa procura-se criar variáveis categorizadas que sejam plausíveis, ou seja, devem fazer sentido para o negócio, considerando a característica da informação, e possuir um número significativo de registros nesse caso estabelecido em representatividade de 5%.

Os dados *missings* foram avaliados separadamente, uma vez que a interpretação do dado faltante pode ser diferente para cada variável, dessa forma o tratamento foi manual e individual a depender do aspecto de cada variável.

A Figura 5.4 exemplifica o processo de categorização das variáveis.

Decil	Bom	Mau	RR	Representatividade
a-.-l.	47	564	0,45	0,24%
b-0-l1	3.821	57.057	0,36	23,65%
d-1-l2	2.776	24.635	0,60	10,65%
e-2-l3	2.886	20.922	0,74	9,25%
f-3-l4	2.799	16.895	0,89	7,65%
g-4-l6	5.318	26.994	1,06	12,55%
h-6-l8	4.680	19.128	1,31	9,25%
i-8-l10	4.121	14.291	1,55	7,15%
j-10-l15	6.780	21.383	1,70	10,94%
k-15-l36	7.248	15.078	2,58	8,67%

Categoria	Bom	Mau	RR	Representatividade
1	3.821	57.057	0,36	23,65%
2	2.776	24.635	0,60	10,65%
3	5.685	37.817	0,81	16,90%
4	5365	27558	1,04	12,79%
5	4.680	19.128	1,31	9,25%
6	10.901	35.674	1,64	18,09%
7	7.248	15.078	2,58	8,67%

Figura 5.4: Categorização da Var031

No anexo I são apresentas as curvas estimadas por Kaplan-Meier. Os gráficos demonstram que as categorias apresentam curvas diferentes, indicando que o tempo de sobrevivência é distinto entre as classes criadas para cada variável.

5.5 Amostragem

Os dados utilizados no desenvolvimento das fórmulas foram selecionados de modo

que representem o universo de clientes a serem avaliados pelo respectivo modelo. Como o mesmo cliente pode possuir mais de um contrato inadimplente no período, para a modelagem retirou-se a duplicidade de CPF, adotando as seguintes regras como critério de desempate:

1. regra 1: deixar o contrato com data de descumprimento mais antiga;

2. regra 2: em caso do cliente ter mais de um contrato descumprido na mesma data, considerar aquele não recuperado;

3. regra 3: persistindo o empate, ou seja, mais de um contrato não recuperado com mesma data de descumprimento, considerar aquele com maior exposição no momento de inadimplência.

A base de clientes descumpridos nas safras de desenvolvimento foi segregada em amostra de desenvolvimento (70%) e amostra de validação (30%). A divisão do banco de dados foi realizado por meio do processo de amostragem aleatória simples sem reposição.

Também foi considerada para avaliação do modelo final a amostra *out of time*. Uma amostra *out of time* significa um banco de dados com período diferente daquele utilizado para a estimação dos parâmetros do modelo.

A quantidade de clientes observados está descrito na Tabela 5.1.

5.6 Modelagem

Para a modelagem, foram criadas variáveis *dummies* para representar as categorias das variáveis. O modelo de Regressão Logística e o Modelo de Riscos Propor-

Tabela 5.1: Quantidade de clientes observados nas bases de dados

Amostra	**Recup**	**Não Recup**	**Total**	**% de Recup**
Desenvolvimento	216940	40468	257408	15,72%
Modelagem (70%)	151809	28377	180186	15,75%
Validação (30%)	65131	12091	77222	15,66%
Out of Time	156724	25343	182067	13,92%

cionais de Cox foram ajustados considerando todas as variáveis *dummies* método utilizado para a escolha das significativas foi o *Stepwise*, com nível de significância 0,05 de entrar e 0,15 de permanecer no modelo.

Após a estimação de cada um dos modelos, foi verificada presença de correlação entre as variáveis finais. A presença de correlação foi avalida pelo VIF (*Variance Inflation Factor*). Valores de VIF superiores a 10 indicam presença de associação.

Desse modo, após a primeira regressão, verificou-se o VIF de cada *dummy* e aquela com maior valor, dentre as que apresentaram VIF superiores a 10, foi retirada.

Assim, outra vez o modelo foi estimado, utilizando-se o *Stepwise*. Novamente o VIF foi verificado, e aquela *dummy* com maior VIF (desde que superior a 10) foi retirada. Esse procedimento foi realizado até que não houvesse mais nenhuma *dummy* correlacionada em cada um dos modelos - Logístico e de Cox.

As estimativas dos parâmetros para o modelo de Cox e de Regressão Logística são apresentados na Tabela 5.2.

Os modelos iniciais contaram com 70 variáveis totalizando 216 *dummies*. A quantidade de *dummies* para cada variável corresponde ao número de categorias criadas. A categoria de referência foi escolhida de modo automático pelo *stepwise*. A quantidade de categorias para cada variável pode ser vista no Anexo I.

Como é possível notar, os dois modelos apontam resultados coerentes, não sendo

Tabela 5.2: Estimativas dos Parâmetros dos Modelos de Regressão de Cox e Logístico

Dummy	**Cox**	**Logística**	**Dummy**	**Cox**	**Logística**
VAR01_d1	-0,093		VAR33_d6		0,182
VAR01_d3	-0,050	-0,073	VAR37_d2	-0,188	
VAR02_d2	-0,078		VAR37_d5		0,334
VAR02_d7	0,202	0,247	VAR38_d4	0,166	
VAR07_d3	-0,068		VAR38_d5		-0,214
VAR07_d6		0,090	VAR39_d3	-0,158	-0,328
VAR08_d5	0,402		VAR39_d5		-0,153
VAR10_d1	-0,136	-0,124	VAR39_d7	0,107	
VAR11_d1	-0,137		VAR40_d3		-0,150
VAR12_d1	-0,317	-0,419	VAR40_d5		-0,123
VAR12_d2	-0,144	-0,164	VAR40_d6	0,124	
VAR12_d4	0,138	0,167	VAR40_d7	0,214	0,126
VAR12_d5	0,292	0,368	VAR41_d6	0,132	0,145
VAR12_d6	0,524	0,658	VAR41_d7	0,177	0,236
VAR12_d7	0,794	1,033	VAR42_d4	-0,097	
VAR13_d2	-0,061	-0,072	VAR42_d7		0,149
VAR13_d4	-0,073		VAR43_d2	-0,456	-0,323
VAR13_d6	-0,128		VAR43_d3	-0,161	
VAR15_d4	0,049		VAR43_d5		0,202
VAR15_d6	0,178		VAR43_d7	0,235	0,527
VAR18_d4	-0,069	-0,091	VAR44_d5	-0,082	-0,089
VAR21_d2	0,228		VAR45_d6	0,037	
VAR22_d1	-0,621	-0,719	VAR47_d5	-0,053	-0,059
VAR22_d3	-0,413	-0,484	VAR48_d2		-0,274
VAR22_d4	-0,215	-0,239	VAR48_d5		-0,124
VAR22_d7	0,142	0,225	VAR48_d7	0,230	0,211
VAR22_d8	0,321	0,499	VAR49_d4	0,325	
VAR25_d5	-0,089		VAR49_d5		-0,328
VAR26_d4	-0,086		VAR50_d4	0,097	0,113
VAR27_d3		-0,312	VAR51_d6	0,231	0,269
VAR27_d5	0,272		VAR53_d3	-0,178	-0,115
VAR27_d7	0,392	0,166	VAR54_d4	-0,106	-0,128
VAR28_d7	0,217	0,256	VAR56_d4	0,108	0,141
VAR29_d1	-0,232	-0,425	VAR58_d3	-0,130	-0,169
VAR29_d4		-0,143	VAR59_d6	0,050	0,071
VAR29_d5	0,086		VAR60_d6	0,056	0,085
VAR30_d1		-0,138	VAR62_d4	-0,068	-0,097
VAR31_d2	0,074		VAR62_d6	0,140	0,129
VAR31_d4	-0,044		VAR65_d5	-0,187	-0,215
VAR31_d7	-0,071	-0,122	VAR70_d2	-0,071	
VAR33_d4	-0,144		VAR70_d6		0,110

observadas inversões de sentido em nenhuma covariável, uma vez que negativos em Cox também são negativos na Logística e os positivos em um modelo também são positivos em outro. Isso acarreta na semelhança de interpretação do coeficiente para os dois modelos.

A *dummy* 3 para a variável 1, por exemplo, indica que os clientes que possuem essa característica

- na análise do modelo de riscos proporcionais, tem um risco menor de recuperação, logo é necessário um tempo maior para se recuperar do que um indivíduo que não tem essa covariável
- para a Regressão Logística, o coeficiente negativo indica que a probabilidade de recuperação é menor.

5.7 Validação do Modelo de Cox

Conforme mencionado, para grandes amostras, não é conveniente utilizar testes estatísticos para validar a adequabilidade do modelo. Sobre esse prisma, foram construídos gráficos que indicam se certa variável pode ser inserida no modelo.

A primeira análise sobre o modelo de Cox consiste nos gráficos com as curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo de Riscos Proporcionais.

Como pode ser visto no Anexo I a aproximação entre as curvas é grande nas variáveis finais do modelo de Cox, o que indica não haver violação da suposição de proporcionalidade dos riscos.

Os resíduos padronizados de Schoenfeld, apresentados no Anexo II, também in-

dicam não haver grandes violações da suposição de riscos constantes.

5.8 Avaliação

Os modelos de *Collection Scoring* têm como finalidade discriminar os devedores entre aqueles que conseguem recuperar os valores em atraso e aqueles que continuam na situação de inadimplência. Dessa forma, como qualquer modelo de *Credit Scoring* as questões envolvidas são:

- qual a qualidade de discriminação do modelo;
- qual modelo é melhor.

A resposta a esses questionamentos está relacionada com a capacidade do escore em identificar os clientes bons e maus.

Existem várias medidas que permitem mensurar e comparar o desempenho dos modelos, entre elas a estatística de Kolmogorov-Smirnov (KS), o indicador AUROC e a taxa geral de acerto.

Sob esse prisma, para avaliar a qualidade dos modelos ajustados, foi calculado o escore a partir das estimativas dos parâmetros em cada um dos modelos.

O escore para o Modelo Logístico é dado pelo preditor linear, uma vez que está intimamente ligado à probabilidade, já que quanto maior, maior é a probabilidade do cliente recuperar a dívida em atraso. Então $\boldsymbol{x'\beta}$ pode ser visto como o escore de bom pagador.

Na Regressão de Cox, quanto maior o valor de $\boldsymbol{x'\beta}$, maior o risco de recuperação, da mesma forma $\boldsymbol{x'\beta}$ pode ser visto como um escore de bom pagador.

5.8.1 Estatística de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é utilizado em Estatística Não Paramétrica para testar igualdade entre funções de distribuição. Em *Credit Scoring*, ele é utilizado para comparar a distribuição do escore entre os clientes bons e maus.

Em modelos com boa capacidade de discriminação, espera-se que os clientes bons estejam concentrados nos escores mais altos e os clientes maus nos escores baixos. Assim, calculando a frequência acumulada de bons e maus por classes de escore, define-se a estatística de KS como:

$$KS = max|F_m(e) - F_b(e)|,$$

dessa forma, percebe-se que quanto maior o valor de KS, melhor performance tem o modelo.

A Figura 5.5 apresenta um exemplo teórico da estatística KS:

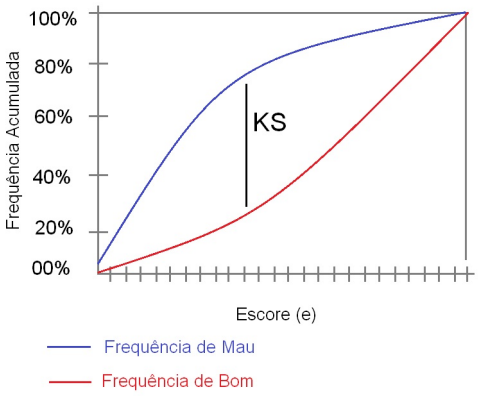


Figura 5.5: Funções de distribuições empíricas para cálculo da estatística KS

5.8.2 AUROC

Para construir a curva ROC (*Receiver Operating Characteristic*) é necessário o entendimento sobre dois conceitos:

- sensibilidade: probabilidade de um indivíduo ser classificado como mau, dado que realmente é mau;

- especificidade: probabilidade de um indivíduo ser classificado com bom, dado que realmente é bom.

Então, variando os pontos de corte ao longo dos escores, calcula-se os valores de sensibilidade e especificidade e dessa forma a curva ROC é construída. O indicador AUROC representa a área sob a curva ROC, conforme Figura 5.6.

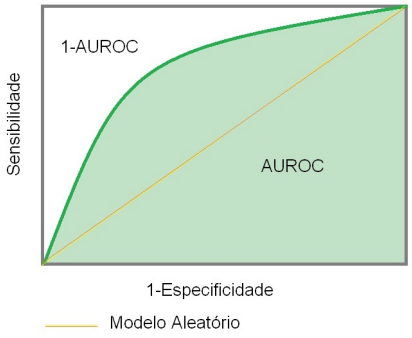


Figura 5.6: Exemplo de curva ROC e indicador AUROC

A Tabela 5.3 apresenta os valores de KS e AUROC para ambos os modelos e nas diversas amostras estudadas.

Tabela 5.3: Nível de Discriminação dos Modelos

Amostra	**Logístico**		**Cox**	
	AUROC	**KS**	**AUROC**	**KS**
Validação (30%)	0,7812	41,91	0,7803	41,46
Modelagem (70%)	0,7808	41,59	0,7804	41,61
Out of Time	0,7452	35,84	0,7456	36,07

Os gráficos observados para os valores de KS e a Curva ROC para cada amostra e tipo de modelo podem ser vistos nas Figuras 5.7 e 5.8.

5.8.3 Taxa de Acerto dos Modelos

Conforme discutido, o interesse principal de qualquer modelo de *Credit Scoring* é a classificação dos indivíduos. Nos *Collections*, o objetivo é prever bom e mau segundo o pagamento dos créditos em atraso.

Uma maneira de analisar a capacidade de acertos de classificação de modelos desse tipo é comparar a classificação obtida pelo modelo com a verdadeira condição dos clientes. Para tal, utiliza-se uma matriz de confusão, que corresponde a uma tabela cruzada entre o resultado previsto por meio do modelo e a suposição de um ponto de corte para a classificação e a condição real observada de cada indivíduo. Nesse caso, a diagonal principal corresponde aos clientes que foram classificados corretamente e os valores fora da diagonal representam os erros de classificação.

A análise da matriz de confusão permite a observação de várias medidas que indicam a capacidade de acerto dos modelos, as quais pode-se citar:

1. capacidade de acerto dos maus: também conhecida como especificidade. Representa a proporção de maus preditos em relação aos maus observados;

2. capacidade de acerto dos bons: também conhecida como sensibilidade. Representa a proporção de bons preditos em relação ao bons observados;

3. capacidade de acerto total: proporção de acertos em relação ao total de clientes. Também conhecida como Acurácia de um modelo. Pode ser interpretada como uma média ponderada da sensibilidade e da especificidade em relação ao total

de observações que apresentam ou não a característica de interesse;

4. valor predidivo positivo: proporção dos bons preditos classificados corretamente em relação ao total previsto de bom;

5. valor preditivo negativo: proporção de maus preditos classificados corretamente em relação ao total previsto de mau;

As Tabelas 5.4 a 5.9 mostram as matrizes de confusão para os modelos. Vale ressaltar que o ponto de corte para os modelos logísticos e de Cox foram escolhidos conforme a especificidade e sensibilidade. O ponto em que esses valores são iguais foi definido como ponto de corte para a classificação dos clientes.

Valores abaixo do ponto de corte classificam o cliente como mau, ou seja, não recuperam a dívida. Por outro lado, clientes com escore acima do ponto de corte indicam que o cliente recuperou os valores devidos e portanto é tido como bom.

É importante ressaltar que nenhuma medida deve ser analisada isoladamente para a avaliação e decisão sobre um modelo, pois todas sofrem influências de características da população. Dessa forma, analisando os indicadores de nível de discriminação entre os dois modelos ajustados, observa-se que o desempenho foi muito semelhante para todas as amostras.

Assim, fatores relacionados ao negócio é que serão fundamentais para a escolha da melhor metodologia. A Regressão Logística é mais difundida, mais aceita, mais utilizada, permite implantação com menor dificuldade. Entretanto, a Análise de Sobrevivência, por estimar o tempo em que a recuperação ocorre, permite a construção de estratégias de cobrança mais específicas, adequadas e efetivas.

Tabela 5.4: Matriz de confusão - Logístico (70%)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	106711	8144	114855
Bom	45098	20233	65331
Total	**151809**	**28377**	**180186**

Taxa de Acerto Geral

Predito	Observado		Total
	Mau	Bom	
Mau	59%	5%	64%
Bom	25%	11%	36%
Total	**84%**	**16%**	**100%**

% em relação aos valores preditos

Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	69%	31%	100%
Total	**84%**	**16%**	**100%**

% em relação aos valores observados

Predito	Observado		Total
	Mau	Bom	
Mau	70%	29%	64%
Bom	30%	71%	36%
Total	**100%**	**100%**	**100%**

Tabela 5.5: Matriz de confusão - Cox (70%)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	107133	8252	115385
Bom	44676	20125	64801
Total	**151809**	**28377**	**180186**

Taxa de Acerto Geral

Predito	Observado		Total
	Mau	Bom	
Mau	59%	5%	64%
Bom	25%	11%	36%
Total	**84%**	**16%**	**100%**

% em relação aos valores preditos

Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	69%	31%	100%
Total	**84%**	**16%**	**100%**

% em relação aos valores observados

Predito	Observado		Total
	Mau	Bom	
Mau	71%	29%	64%
Bom	29%	71%	36%
Total	**100%**	**100%**	**100%**

Uma sugestão para novos trabalhos considerando o desenvolvimento de Modelos de *Collection Scoring* consiste na utilização de técnicas relacionadas a Modelos de Sistemas Reparáveis, uma vez que cada intervenção de cobrança pode alterar (aumentar) o risco de recuperação.

Tabela 5.6: Matriz de confusão - Logístico (30%)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	45841	3467	49308
Bom	19290	8624	27914
Total	**65131**	**12091**	**77222**

Taxa de Acerto Geral

Predito	Observado		Total
	Mau	Bom	
Mau	59%	4%	64%
Bom	25%	11%	36%
Total	**84%**	**16%**	**100%**

% em relação aos valores preditos

Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	69%	31%	100%
Total	**84%**	**16%**	**100%**

% em relação aos valores observados

Predito	Observado		Total
	Mau	Bom	
Mau	70%	29%	64%
Bom	30%	71%	36%
Total	**100%**	**100%**	**100%**

Tabela 5.7: Matriz de confusão - Cox (30%)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	45995	3531	49526
Bom	19136	8560	27696
Total	**65131**	**12091**	**77222**

Taxa de Acerto Geral

Predito	Observado		Total
	Mau	Bom	
Mau	60%	5%	64%
Bom	25%	11%	36%
Total	**84%**	**16%**	**100%**

% em relação aos valores preditos

Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	69%	31%	100%
Total	**84%**	**16%**	**100%**

% em relação aos valores observados

Predito	Observado		Total
	Mau	Bom	
Mau	71%	29%	64%
Bom	29%	71%	36%
Total	**100%**	**100%**	**100%**

Tabela 5.8: Matriz de confusão - Logístico (*Out of time*)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	108784	8549	117333
Bom	47940	16794	64734
Total	**156724**	**25343**	**182067**

Taxa de Acerto Geral

Predito	Observado		Total
	Mau	Bom	
Mau	60%	5%	64%
Bom	26%	9%	36%
Total	**86%**	**14%**	**100%**

% em relação aos valores preditos

Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	74%	26%	100%
Total	**86%**	**14%**	**100%**

% em relação aos valores observados

Predito	Observado		Total
	Mau	Bom	
Mau	69%	34%	64%
Bom	31%	66%	36%
Total	**100%**	**100%**	**100%**

Tabela 5.9: Matriz de confusão - Cox (*Out of time*)

Valores Predito e Observado

Predito	Observado		Total
	Mau	Bom	
Mau	107877	8429	116306
Bom	48847	16914	65761
Total	**156724**	**25343**	**182067**

Taxa de Acerto Geral

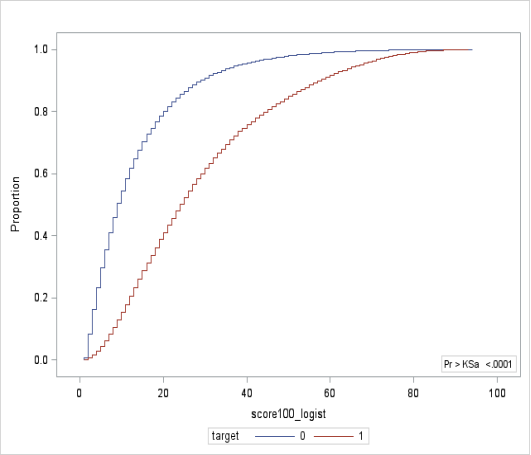
Predito	Observado		Total
	Mau	Bom	
Mau	59%	5%	64%
Bom	27%	9%	36%
Total	**86%**	**14%**	**100%**

% em relação aos valores preditos

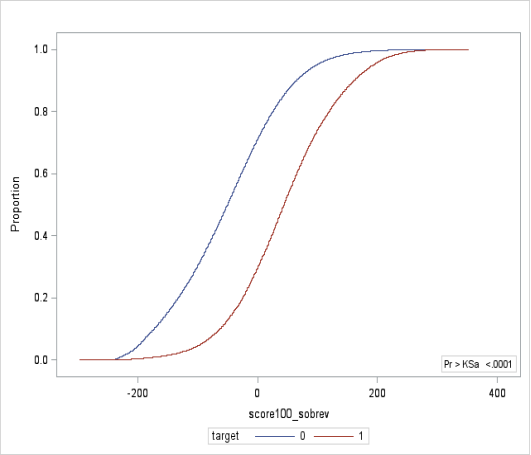
Predito	Observado		Total
	Mau	Bom	
Mau	93%	7%	100%
Bom	74%	26%	100%
Total	**86%**	**14%**	**100%**

% em relação aos valores observados

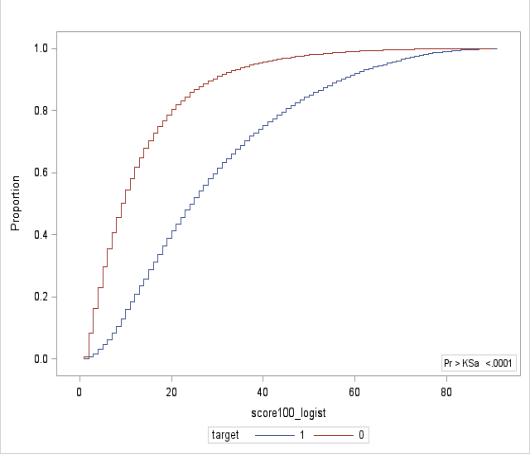
Predito	Observado		Total
	Mau	Bom	
Mau	69%	33%	64%
Bom	31%	67%	36%
Total	**100%**	**100%**	**100%**



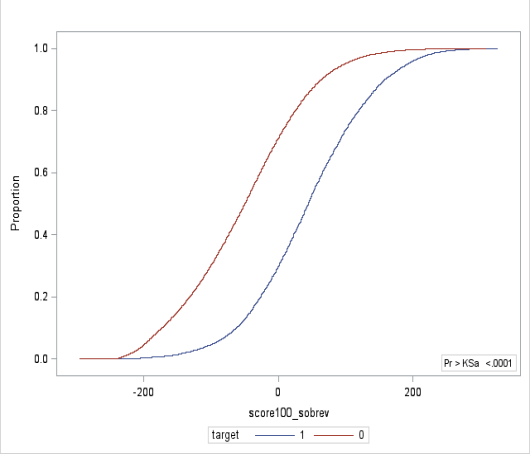
(a) Logística (70)



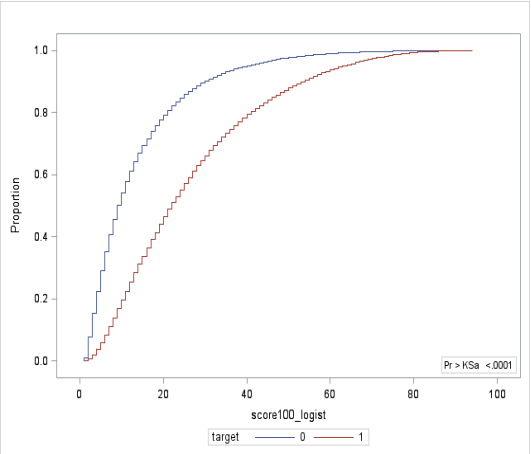
(b) Cox (70)



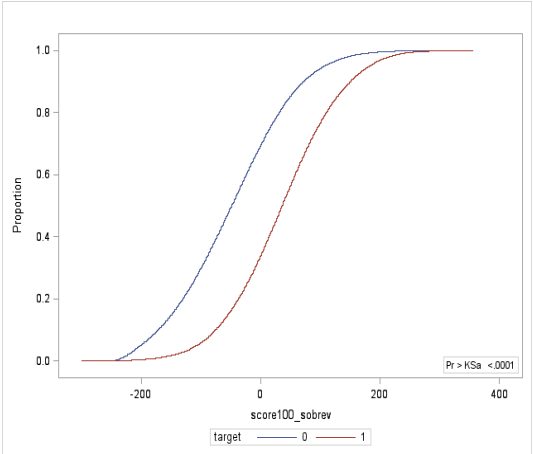
(c) Logística (30)



(d) Cox (30)

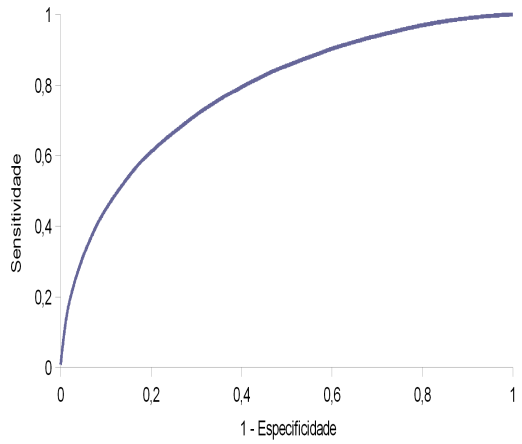


(e) Logística (*Out of time*)

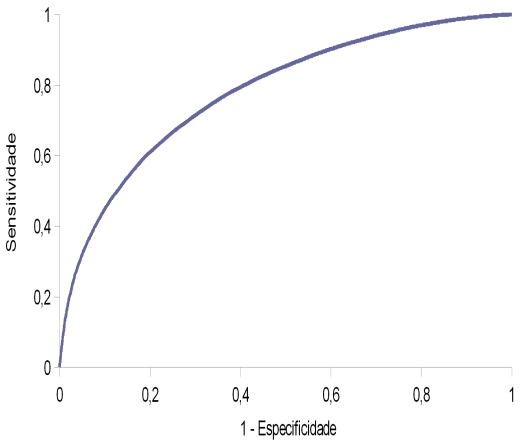


(f) Cox (*Out of time*)

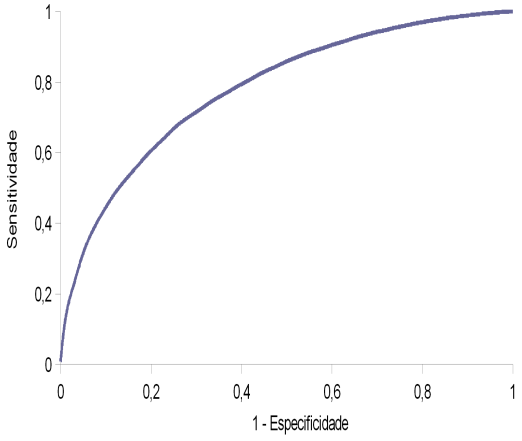
Figura 5.7: Gráficos de KS para as amostras em estudo



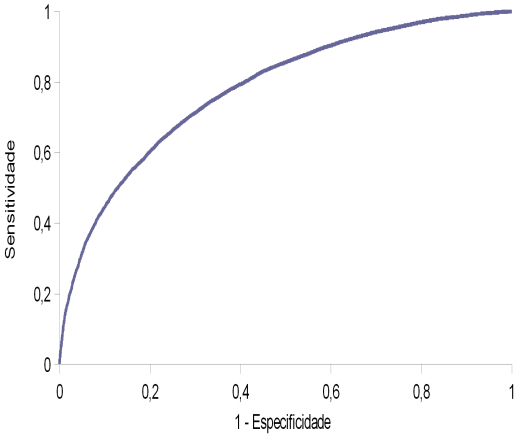
(a) Logística (70)



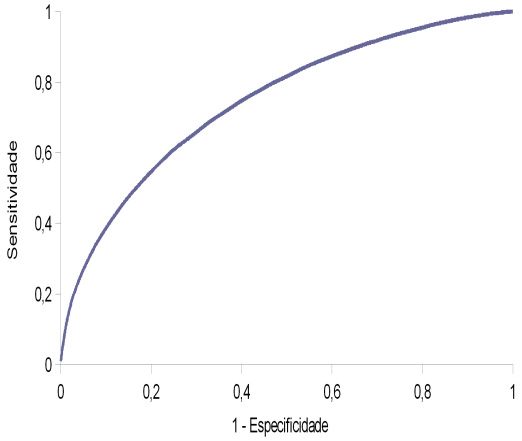
(b) Cox (70)



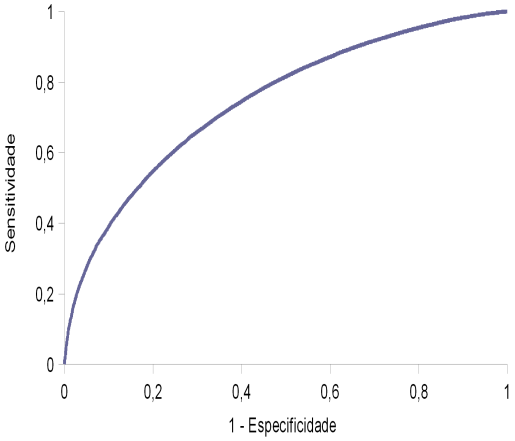
(c) Logística (30)



(d) Cox (30)



(e) Logística (*Out of time*)



(f) Cox (*Out of time*)

Figura 5.8: Curva ROC para as amostras em estudo

Capítulo 6

Conclusão

Com o aquecimento da economia e a oferta de crédito, as instituições financeiras viram a necessidade de estruturar processos objetivos e eficientes para a gestão do risco em todas as etapas do ciclo de crédito, desde a concessão até a cobrança e retenção do cliente. No Brasil, a popularização do crédito se iniciou principalmente com o Plano Real e o controle da inflação e se extente atualmente pelos diversos programas financeiros lançados pelas Instituições. Em 2013, foi lançado por dois dos maiores bancos nacionais programas de incentivo ao fornecimento de capital - Programa Bom Pra Todos (Banco do Brasil) e Caixa Melhor Crédito (Caixa Econômica Federal) - tornando a administração e mitigação do risco importantíssimas para a rentabilidade do negócio e a sustentação da economia. Entretanto, com a expansão do crédito, vem também o aumento da inadimplência, objeto de preocupação deste trabalho.

Nesse sentido, devido a necessidade de formulação de instrumentos eficientes para a mensuração dos riscos envolvidos no processo, os modelos estatísticos passaram a ser ferramentas importantes para auxiliar os gestores a tomarem decisões mais acertadas. Assim, buscar aprimoramento desses modelos tornou-se diferencial

competitivo das instituições financeiras.

Existem diversos trabalhos e estudos sobre modelagem para concessão e gestão do risco de crédito. Entretanto, há pouca literatura disponível referente a modelos de cobrança, e na economia desafiante de hoje, as empresas devem se preocupar não somente com a avaliação do risco na aquisisão do capital, mas também com a cobrança das receitas de clientes. Desse modo, buscou-se com este trabalho contribuir para os conhecimentos relativos aos modelos de *Collection Scoring*.

Para as instituições financieras, o desenvolvimento de um modelo de *Collection Scoring* proporciona melhor gerenciamento do negócio - uma vez que produz melhoria na gestão de cobrança, aumento da efetividade e retorno financeiro, redução dos custos com cobranças e ações impróprias, gestão de campanhas, ofertas, descontos, parcelamentos e acordos mais eficientes - do risco, com ações preventivas, gestão de políticas de crédito e mensuração do risco - e de atendimento à regulamentação, já que atende ao estabelecido pela Resolução CMN 2.682/1999 no sentido de que cria critérios de classificação e regras de gestão para as operações de crédito, além de auxiliar na constituição da provisão para créditos de devedores duvidosos.

Este trabalho apresentou os conceitos que sustentam o desenvolvimento de um modelo de *Collection Scoring* com os dados de clientes inadimplentes de uma instituição financeira de grande porte com atuação nacional por meio da técnica estatística de Análise de Sobrevivência.

As principais características da Análise de Sobrevivência correspondem a sua capacidade de extrair informações de dados censurados, isto é, daqueles clientes

que ao final do estudo não experimentaram o evento de interesse, além de levar em consideração os tempos para a ocorrência da "falha". Foi também desenvolvido um modelo com Regressão Logística para efeitos de comparação.

O que percebemos é que uma técnica não se destaca muito sobre a outra em termos de indicadores de discriminação, mas o ganho que se tem na construção de estratégias de cobrança dadas pela Análise de Sobrevivência são maiores, uma vez que tem-se a estimativa do tempo em que a regularização da dívida pode ocorrer.

Vale ressaltar que modelos de *Collection Scoring* possuem algumas limitações, como alteração de perfil do público alvo, são influenciados por ações de cobrança, na economia e por fatores operacionais. Dessa forma, um desafio é construir um modelo robusto o suficiente para que não descalibre logo que a política de cobrança gerada pelo seu próprio escore comece a ser utilizada. Assim modelos de *scoring* necessitam ser recalibrados regularmente.

Apesar da bibliografia especializada apresentar a elaboração de Modelos de *Collection Scoring* com visão de produto, é sabido que o pagamento de dívidas em atraso depende de todos os compromissos assumidos pelo cliente. Dessa forma, conclui-se que o desenvolvimento desse tipo de modelo observando o cliente sem segregá-lo por suas operações é mais adequado, por considerar a visão do endividamento do cliente. Entretanto, como neste trabalho o objetivo principal foi verificar a aplicação das técnicas estatísticas ao problema e divulgar o mínimo necessário sobre a realidade da Instituição que cedeu os dados, o cliente não foi avaliado de forma completa, no sentido de considerar toda a sua dívida com a instituição, mas apenas

o comprometimento em atraso.

Referências Bibliográficas

1. Abreu, H.J. Aplicação da Análise de Sobrevivência em um Problema de Credit Scoring e Comparação com a Regressão Logística. 2004. Dissertação (Mestrado em Estatística) - UFSCar, 2004.

2. Agresti, A. Categorical data analysis. New York: John Wiley, 1990.

3. Almeida, M.P. Estimação Bayesiana em Modelos de Sobrevivência: uma Aplicação em Credit Scoring. 2008. 67f. Dissertação (Mestrado em Estatística) - Instituto de Ciencias Exatas e Naturais, Universidade Federal do Pará. 2008.

4. Altman, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23. 589-609. 1968.

5. Anderson, R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford: Oxford University Press, 2007.

6. Andreeva, G. European Generic Scoring Models Using Survival Analysis. J. Oper. Res. Soc. 2006.

7. Applied Business Intelligence Group. Collections Score - Powered by Predictive Metrics. 2001. Disponível em: http://www.appliedbigroup.com/content/Collections%20Strategy%20Overview.pdf. Acesso em: 09jun2014.

8. Banasik, J., Crook, J. N., Thomas, L. C. Not if but when will borrowers default. In: Journal of the Operational Research Society, 1999.

9. Banco Central do Brasil. Dispõe sobre a implementação de estrutura de gerenciamento do risco de crédito. Resolução n 3.721, de 30 de abril de 2009.

10. Banco Central do Brasil. Dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa. Resolução n 2.682, de 21 de dezembro de 1999.

11. Barth, N. L. Inadimplência: construção de modelos de previsão. São Paulo: Nobel, 2004.

12. Brunello, G. H. V; Nakano, E. Y. (2015). Inferência bayesiana no modelo weibull discreto em dado com presença de censura. TEMA - Tend. Mat. Apl. Comput., no prelo.

13. Bumacob, V., Ashta, A. The Conceptual Framework of Credit Scoring from its Origins to Microfinance. Disponível em: http://www.rug.nl/research/globalisation-studies-groningen/research/conferencesandseminars/conferences/eumicrofinconf2011/papers/new.10c.bumacov.format.doc. Acesso em: 09jun2014.

14. Bessis, J. Risk Management in Banking. Chichester: John Wiley Sons, 1998.

15. Caouette, J., Altmano, E. Narayanan, P. Gestão do Risco de Crédito: o Próximo Grande Desafio Financeiro. Rio de Janeiro: Qualitymark, 1999.

16. Carrasco, C. G.; Tutia, M. H; Nakano, E. Y. (2012). Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. TEMA - Tend. Mat. Apl. Comput., v.13, n.3, p. 247-255.

17. Casella, G.; Berger, R. L. Inferência Estatística. 2ª Edição. Cengage Learning, 2010.

18. Collet, D. Modeling Survival DAta in Medical Research. London: Chapman and Hall, 1994.

19. Colosimo, E.A., Giolo, S.R. Análise de Sobrevivência Aplicada. 1ed. São Paulo: Edgard Blücher, 2006. 370p. (ISBN 85-212-0384-5).

20. Cox, D.R. Regression Models and life-tables (with discussion). J. Royal Statistics Society Series B. N 74, 1972.

21. Deloitte. Maturidade da Competência de Cobrança. 2011. Disponível em: http://www.abbc.org.br/arquivos/2011_05_02_collection_score_abbc.pdf. Acesso em: 09jun2014.

22. Diniz, C., Louzada, F. Métodos Estatísticos para Análise de Dados de Crédito. *6th Brazilian Conference on Statistical Modeling in Insurance and Finance*, Maresias - SP, 2013.

23. Durand, D. Risk Elements in Consumer Instalment Financing. National Bureau of Economic Research. 1941.

24. Efron, B. The Efficiency of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association, 72, 557-565. 1977.

25. Ehlers, R. S.; Inferência Estatística - Notas de Aula. Departamento de Matemática Aplicada e Estatística da USP. Disponível em <http://www.icmc.usp.br/ ehlers/inf/inf.pdf>. Acesso em 27/04/2015.

26. Experian Decision Analytics. Experian Collections Strategies - Guide One: Maximizing the Performance of Strategies. 2010. Disponível em: http://www.experian.ie/assets/decision-analytics/brochures/maximising-the-performance-of-strategies.pdf. Acesso em: 09jun2014.

27. Experian Decision Analytics. Experian Collections Strategies - Guide Two: Extending the depth of collections capabilities. 2010. Disponível em: http://www.experian.ie/assets/decision-analytics/brochures/extending-depth-of-collections-strategies.pdf. Acesso em: 09jun2014.

28. Experian Decision Analytics. The Value of Implementing Scoring in the Collection Process - A Decision Analytics briefing paper from Experian. 2006. Disponível em: http://www.experian.co.uk/assets/decision-analytics/briefing-papers-global/ExperianDA_BP_ImplementingScoring.pdf. Acesso em: 09jun2014.

29. Farewell, V. T. e Prentice, R. L. The Approximation of Partial Likelihood with Emphasis on Case-Control Studies. Biometrika, 67, 273-279. 1980.

30. Gujarati, D. N. Econometria Básica. 3ed. São Paulo: Makron Books, 2000.

31. Hand, D. J., Henley, W. E. Statistical Classification Methods in Consumer Credit Scoring: a Review. In: J. R. Statist. Soc. A (1997).

32. Hosmer, D. W., Lemeshow, S. Applied Survival Analysis - Regression Modeling of Time to Event Data. 1ed. Estados Unidos da América: John Wiley Sons, Inc, 1999. .

33. Hosmer, D. W., Lemeshow, S. Applied Logistic Regression. 2nd ed. New York: John Wiley Sons, 2000.

34. Intrabase. Desenvolvimento de Modelos Estatísticos de Collection. Disponível em: http://www.intrabase.com.br/imgs/case_collection.pdf. Acesso em: 09jun2014.

35. Kalbfleisch, J. D., Prentice, R. L. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York, 1st edition. 1980.

36. Keramati, A., Yousefi, N. A Proposed Classification of Data Mining Techniques in Credit Scoring. In: International Conference on Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, January 22 24, 2011.

37. Kutner, M.H, Nachtsheim, C.J., Neter, J. Li, W. Applied Linear Statistical Models. 5ed. Nova York: McGraw-Hill Companies, Inc, 2005. 1396p. (ISBN 007-112221-4).

38. Lawrence, D.B. Risco e Recompensa: O Negócio de Crédito ao Consumidor. Tradução de Debbie McKey. Nova York: Individual Bank, Citicorp, 1984. 198p.

39. Liu, Y. New Issues in Credit Application. Institut für Wirtschaftsinformatik, Arbeitsbericht 16/2001.

40. McCULLAGH, P.; NELDER, J.A. Generalized Linear Models. 2.ed. Londres: Chapman Hall, 1989. 532p.

41. Machado, A. R. Modelos Estatísticos para Avaliação de Risco em Produtos de Crédito Parcelados. Monografia (Graduação em Estatística), Universidade de Brasília, 2010.

42. Maia, A. H. N., Teixeira, L. A. J. Uso de Análise de Sobrevivência para Estudos Fenológicos em Fruteiras. XX Congresso Brasileiro de Fruticultura.

43. Manfio, F. O Risco Nosso de Cada Dia. Barueri, São Paulo: Estação das Letras Editora, 2007. 220p. (ISBN 978-85-60166-03-9).

44. Medeiros, K. M., Brito, F. I., Araujo, A. O. Gestão de Crédito e Cobrança: Análise dos Resultados da Terceirização em uma Financeira. Disponível em: https://www.congressousp.fipecafi.org%2Fweb%2Fartigos82008%2F115.pdf &ei=hiuWU6-nLcPNsQTBtYHYDw&usg=AFQjCNFGF7X3zicskTJccPnfby1 rLdBAxQ. Acesso em: 09jun2014.

45. Menck, A. C., Moriguchi, S. N. Marketing. Brasília: UAB, 2009.

46. Myers, J. H. and Forgy, E. W. Development of Numerical Credit Evaluation Systems. Journal of American Statistical Association 50, 797-806, 1963.

47. Nakano, E.Y., Carrasco, C.G. Uma Avaliação do Uso de um Modelo Contínuo na Análise de Dados Discretos de Sobrevivência. TEMA Tend. Mat. Apl. Comput., 7, No.1 (2006), 91-100.

48. Narain, B. Survival Analysis and Credit Granting Decision. In: L.C. Thomas, J.N. Crook, D. B. Edelman, Eds. Credit Scoring and Credit Control, OUP, OXFORD, U.K., 1992.

49. Neto, F. N., Pereira, B. de B. Modelos em Análise de Sobrevivência. Cadernos Saúde Coletiva, Rio de janeiro, 8(1): 8-26, 2000.

50. Orgler, Y. E. A Credit Scoring Model for Commercial Loans. Journal of Money, Credit Banking (Ohio State University Press) 2, 435-445. 1970.

51. Pereira, G. H. de A. Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais. 2004. 96f. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo. São Paulo, 2004.

52. Peto, R. Contribuição à discussão do artigo de D. R. Cox. Journal of the Royal Statistical Society B, 34, 205-207. 1972

53. Régis, D. E., Artes, R. Modelo Multi-Estado de Markov em Cartões de Crédito. *Insper Working Paper*, WPE: 137/2008.

54. Sabato, G. Credit Risk Scoring Models. 2010. Disponível em: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1546347. Acesso em: 09jun2014.

55. Sandroni, P. Dicionário de Economia do Século XXI. São Paulo: Record, 2005.

56. Sarmento, a. Experimentação e Avaliação de Modelos para um Problema de Atribuição de Crédito. 2005. 99f. Dissertação (Mestrado em Análise de Dados e Sistemas de Apoio à Decisão) - Facudade de Economia, Universidade do Porto. 2005.

57. Saunders, A. Medindo o Risco de Crédito: Novas Abordagens para o Value at Risk e Outros Paradigmas. Rio de Janeiro: Qualitumark, 2000.

58. Sawant, A. A. Chawan, P.M. Study of Data Mining Techiniques Used for Financial Data Analysis. *International Journal of Engineering Science and Innovative Technology*, Volume 2, Issue 3, Maio 2013.

59. Sicsú, A. L. Credit Scoring: desenvolvimento, implantação, acompanhamento. São Paulo:Blucher, 2010.

60. Silva, J. P da. Gestão e Análise de Risco de Crédito. São Paulo: Atlas, 1997.

61. Silva, G. D. M., Silva, R. A., Franco, G. C. Modelo de Análise de Sobrevivência para Avaliar os Efeitos do Consumo de Etanol na Memória Espacial de Ratos.

62. Souza, R. B. O Modelo de Collection Scoring como Ferramenta para a Gestão Estratégica do Risco de Crédito. 2000. 75f. Dissertação (Mestrado em Administração) - FGV/EAESP. 2000.

63. Strapasson, E. Comparação de Modelos com Censura Intervalar em Análise de Sobrevivência. 2007. Tese de Doutorado - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, SP.

64. Thomas, L.C. Novos Avanços das Metodologias de Credit Scoring. Tecnologia de Crédito, Ano VI, Número 35, SERASA. 2001.

65. Thomas, L.C., Edelman, D. B., Crook, J. N. Credit Scoring and Its Applications. Siam: Philadelphia, 2002.

66. Thomas, L.C., Stepanova, M. Survival Analysis Methods for Personal Loan Data. Operations Research, 2002.

67. Tomazela, S. M. O. Avaliação de Desempenho de Modelos de Credit Score Ajustados por Análise de Sobrevivência. 2007. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo. São Paulo, 2007.

68. Trevisani, A. T., *et al.* Qualidade de Dados - Desafio Crítico para o Sucesso do Business Intelligence. Itají: XVIII Congresso Latino Americano de Estratégia, 2004.

69. Vieira, A. M. C. Gestão do Risco de Crédito - Automação do Processo de Crédito e Credit Scoring. Aymoré Financiamentos Gerenciamento de Risco, 2004.

70. Vilas Novas, R. S. R., Martinez, J. M. Modelos de Collection Scoring. IMECC - UNICAMP. Disponível em: http://vigo.ime.unicamp.br/Projeto/2013-1/ms777/ms777_renan.pdf. Acesso em: 09jun2014.

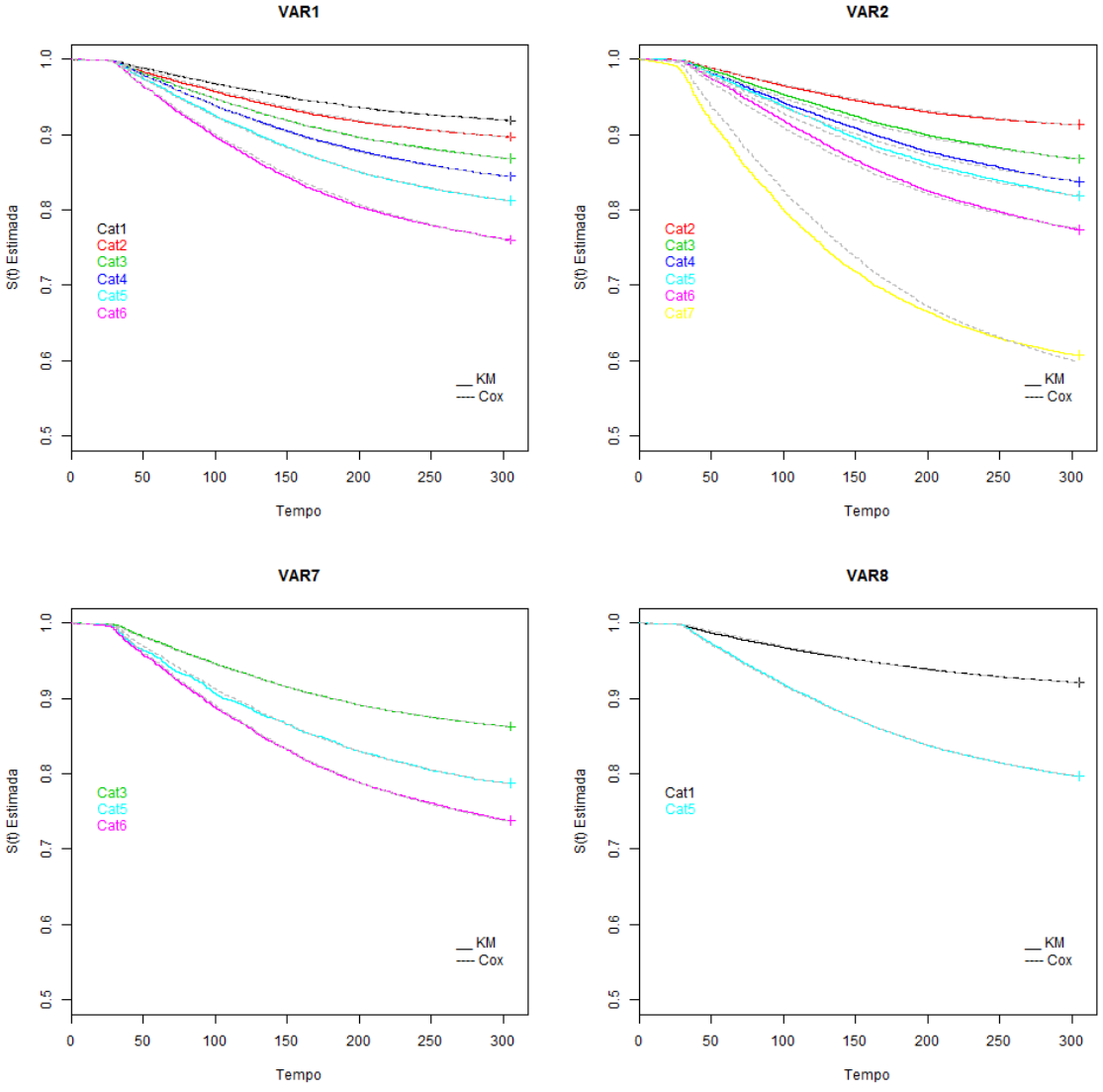
71. Yamamoto, W. A. A., Oliveira, E. A. A. Q., Santos, V. S. O Gerenciamento de Risco de Crédito em um Banco de Varejo: um Estudo do Segmento Pessoas Físicas. In: XV Encontro Latino Americano de Iniciação Científica e XI Encontro Latino Americano de Pós-Graduação - Universidade do Vale do Paraíba. 2011.

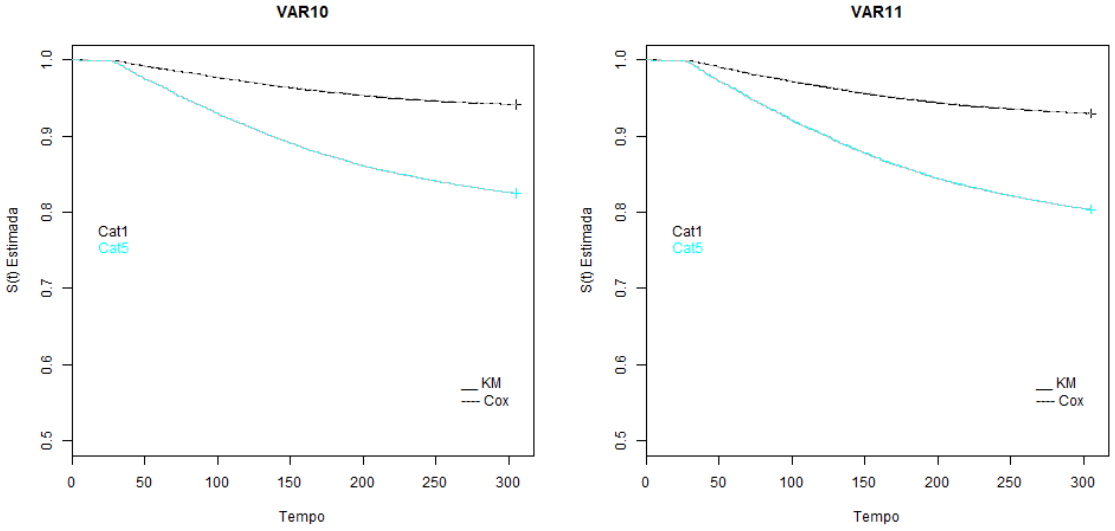
72. Weingartner, H. M. Concepts and Utilization of Credit-Scoring Techniques. Banking 58, 51-54, 1966.

73. Análise de sobrevivência - Instituto de Estudos em Saúde. Disponível em: http://www.iesc.ufrj.br/cursos/bioestatistica/Capitulo5Bioestatisticaanalisedeso.ppt. Acesso em: 28out2009.

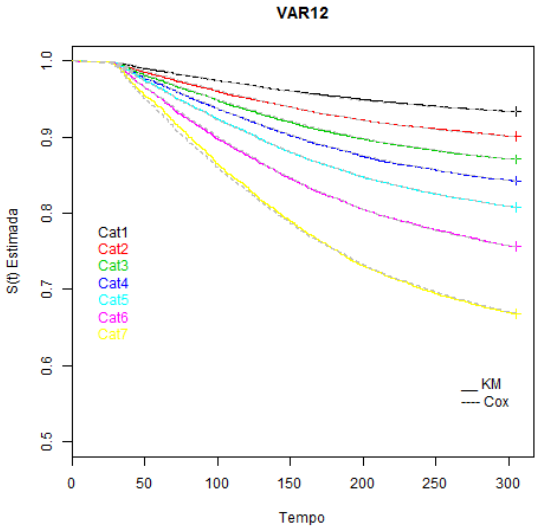
Apêndice A

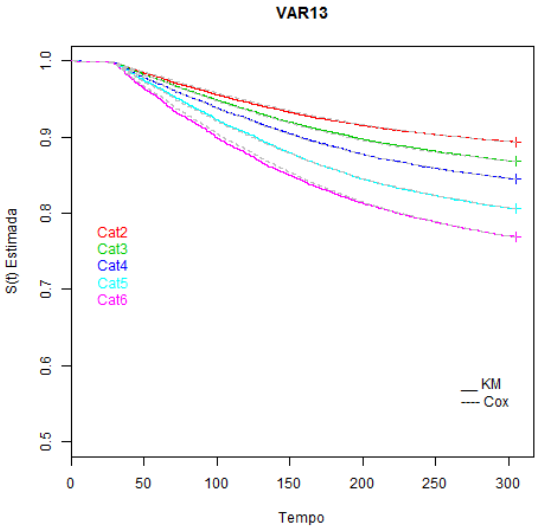
Anexo I - Funções de Sobrevivência Estimadas - Kaplan-Meier e Cox

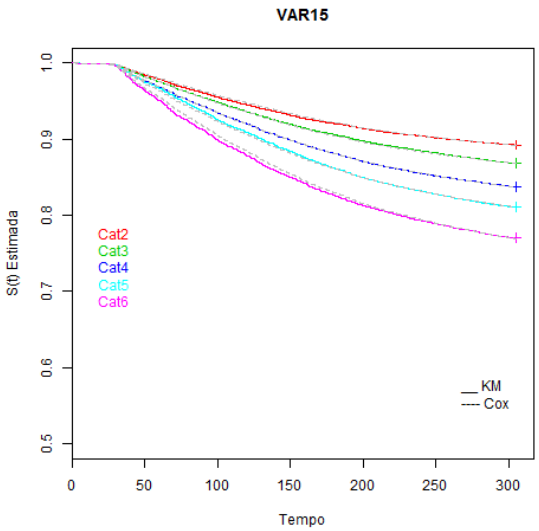


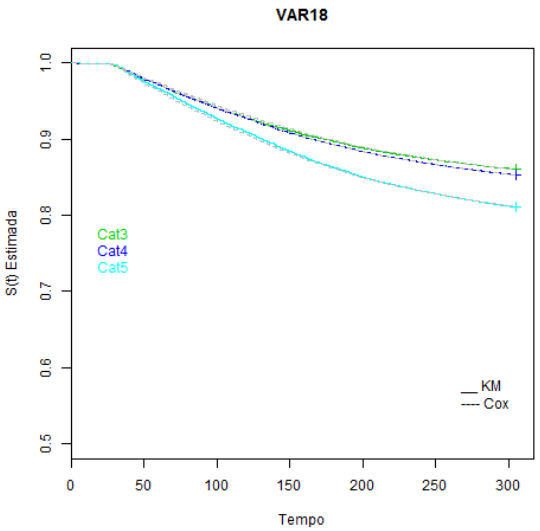

VAR10
S(t) Estimada
Tempo
Cat1
Cat5
KM
Cox

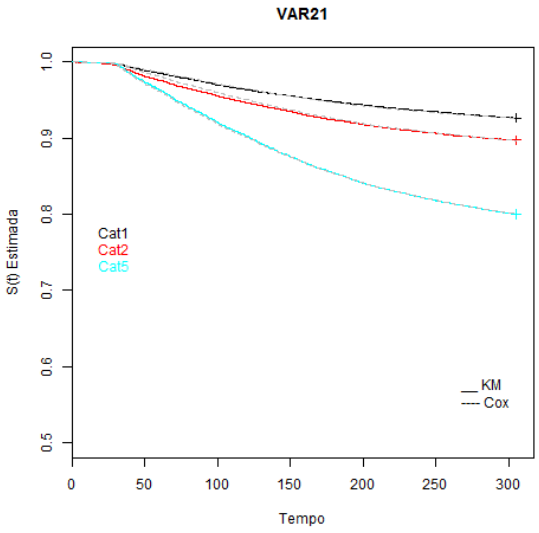
VAR11
S(t) Estimada
Tempo
Cat1
Cat5
KM
Cox

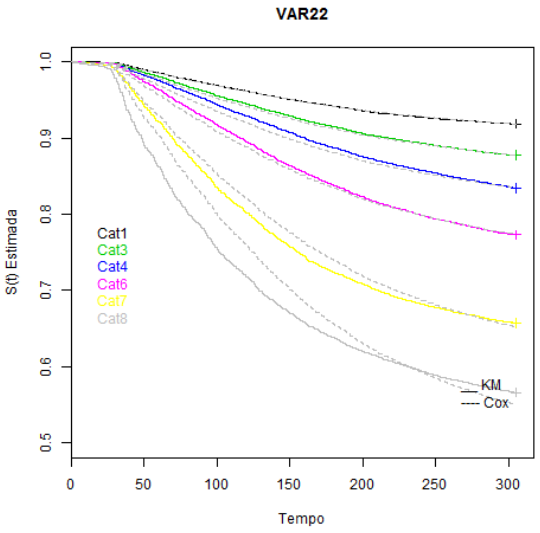

VAR12
S(t) Estimada
Tempo
Cat1
Cat2
Cat3
Cat4
Cat5
Cat6
Cat7
KM
Cox

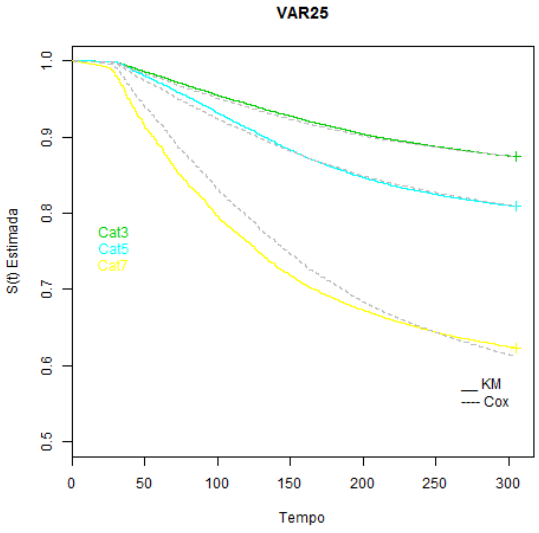

VAR13
S(t) Estimada
Tempo
Cat2
Cat3
Cat4
Cat5
Cat6
KM
Cox

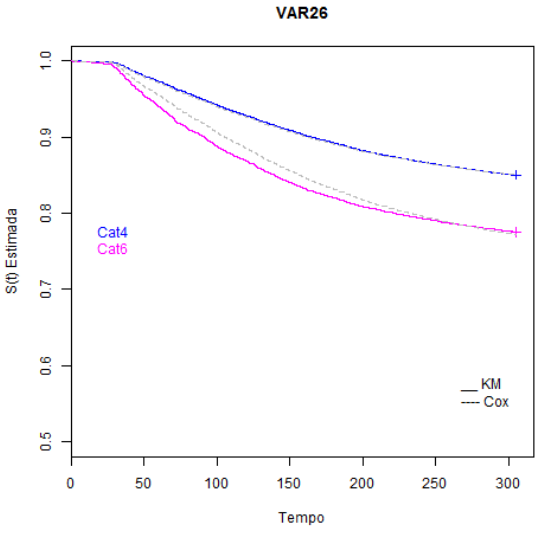

VAR15
S(t) Estimada
Tempo
Cat2
Cat3
Cat4
Cat5
Cat6
KM
Cox

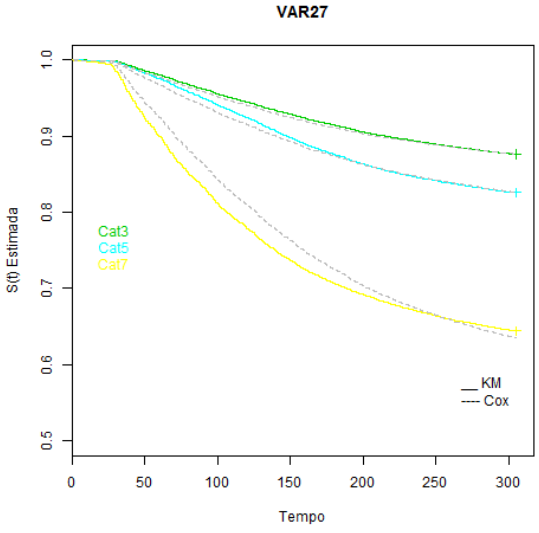

VAR18
S(t) Estimada
Tempo
Cat3
Cat4
Cat5
KM
Cox

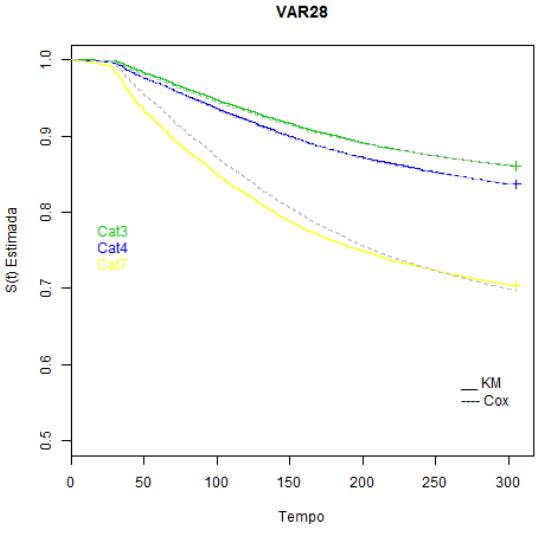

VAR21
S(t) Estimada
Tempo
Cat1
Cat2
Cat5
KM
Cox

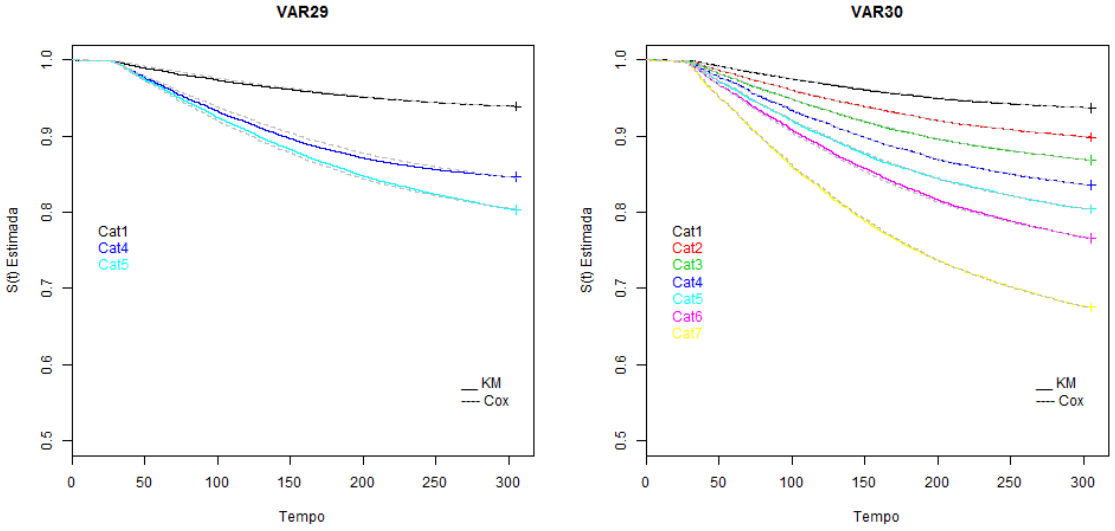

VAR22
S(t) Estimada
Tempo
Cat1
Cat3
Cat4
Cat6
Cat7
Cat8
KM
Cox


VAR25
S(t) Estimada
Tempo
Cat3
Cat5
Cat7
KM
Cox

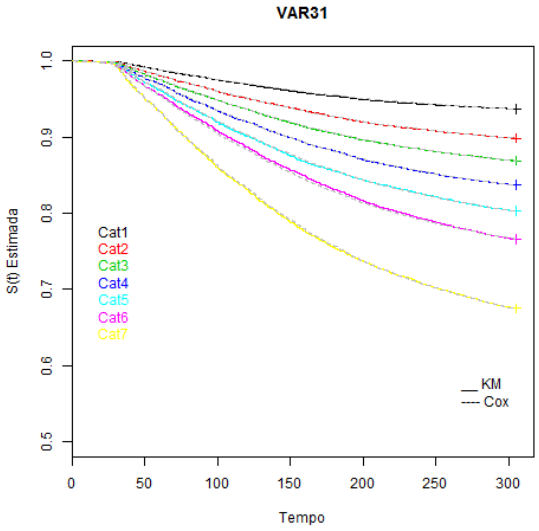

VAR26
S(t) Estimada
Tempo
Cat4
Cat6
KM
Cox

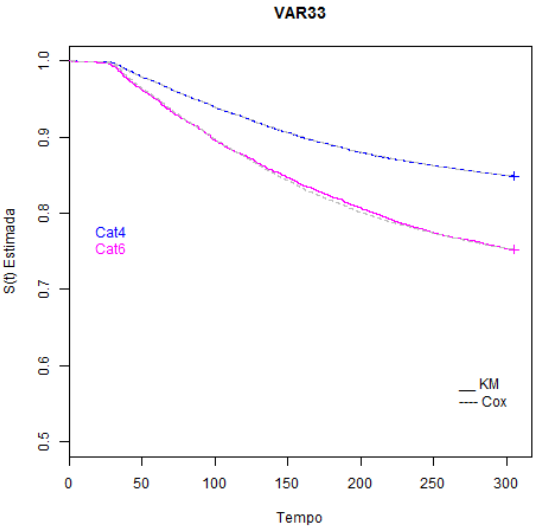

VAR27
S(t) Estimada
Tempo
Cat3
Cat5
Cat7
KM
Cox

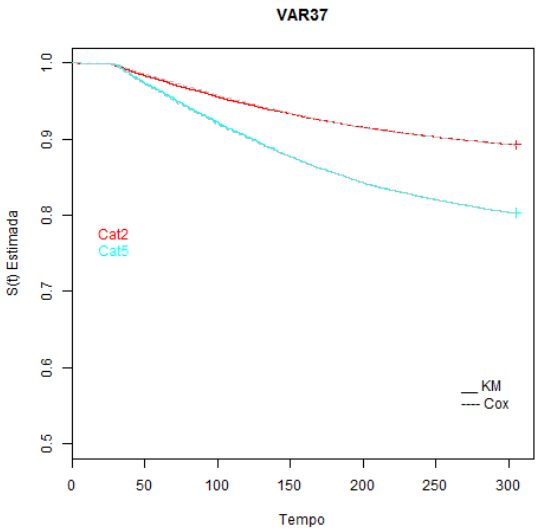

VAR28
S(t) Estimada
Tempo
Cat3
Cat4
Cat7
KM
Cox

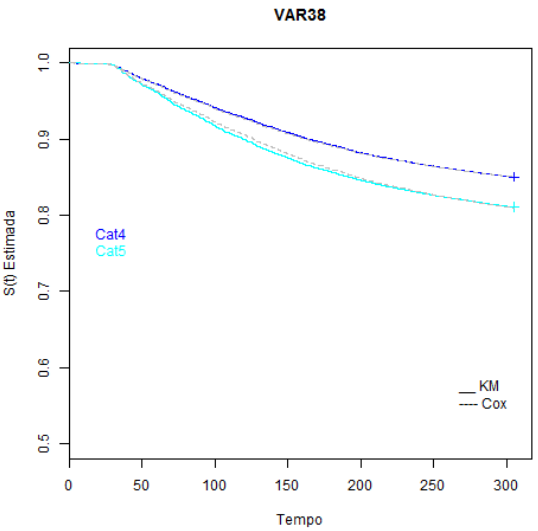

VAR29
S(t) Estimada
Tempo
Cat1
Cat4
Cat5
KM
Cox

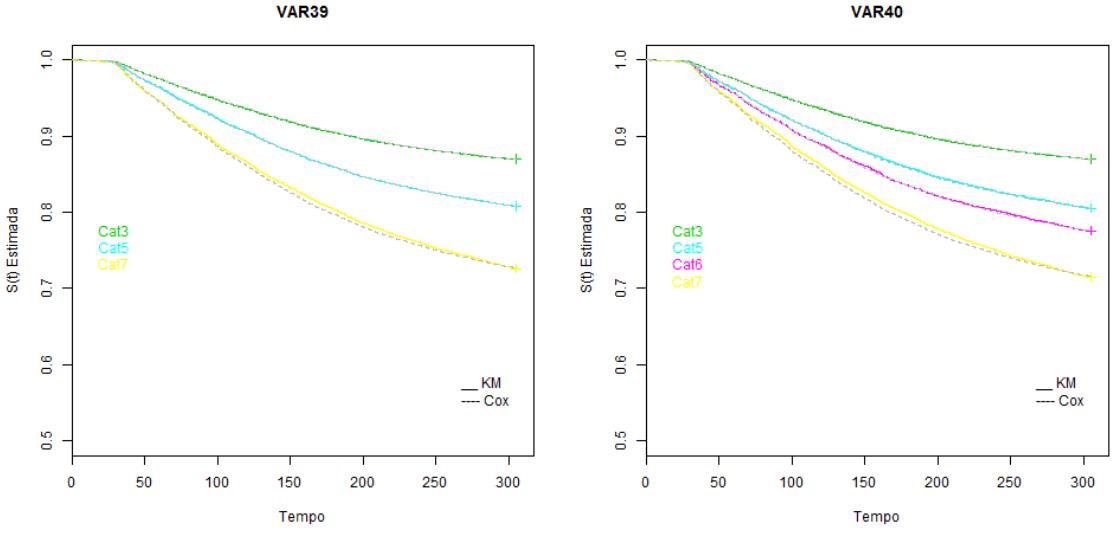
VAR30
S(t) Estimada
Tempo
Cat1
Cat2
Cat3
Cat4
Cat5
Cat6
Cat7
KM
Cox


VAR31
S(t) Estimada
Tempo
Cat1
Cat2
Cat3
Cat4
Cat5
Cat6
Cat7
KM
Cox

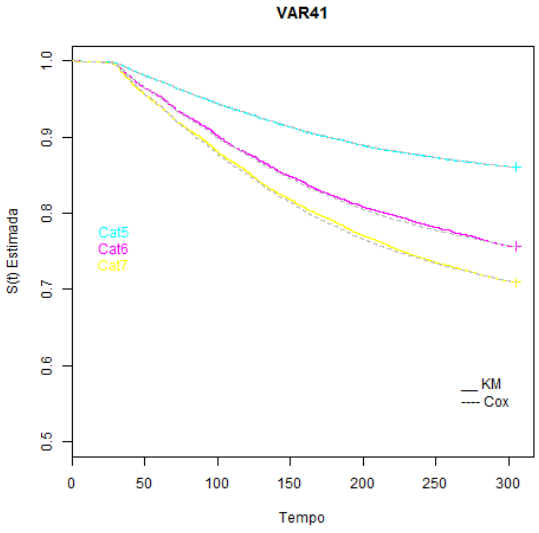

VAR33
S(t) Estimada
Tempo
Cat4
Cat6
KM
Cox

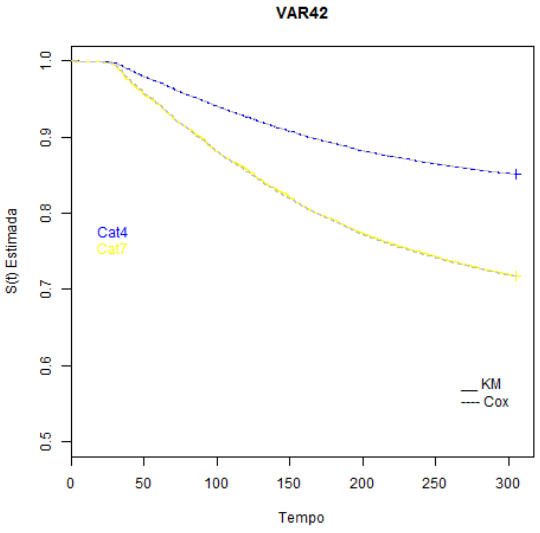

VAR37
S(t) Estimada
Tempo
Cat2
Cat5
KM
Cox

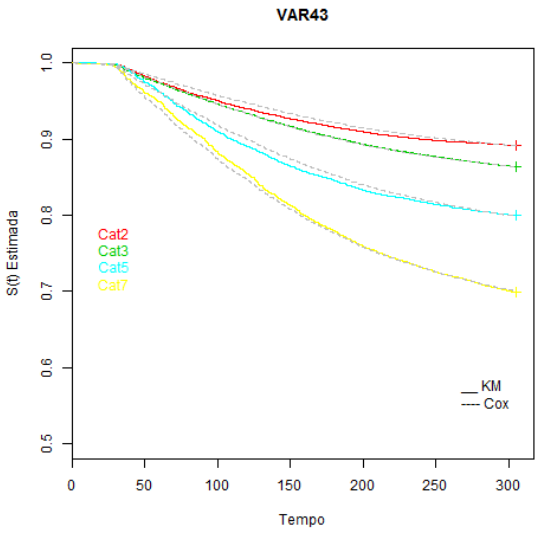

VAR38
S(t) Estimada
Tempo
Cat4
Cat5
KM
Cox

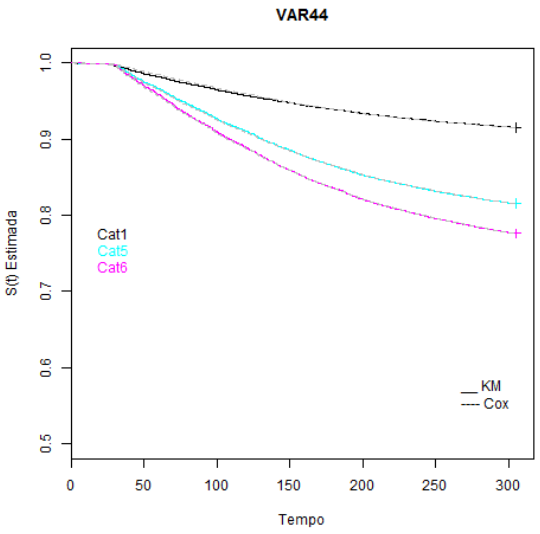

VAR39
S(t) Estimada
Tempo
Cat3
Cat5
Cat7
KM
Cox

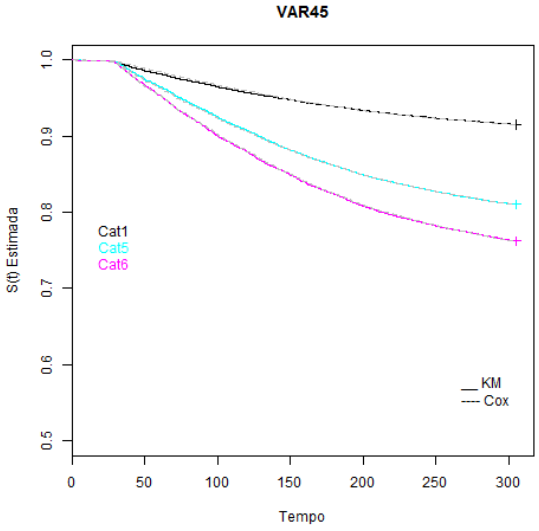
VAR40
S(t) Estimada
Tempo
Cat3
Cat5
Cat6
Cat7
KM
Cox

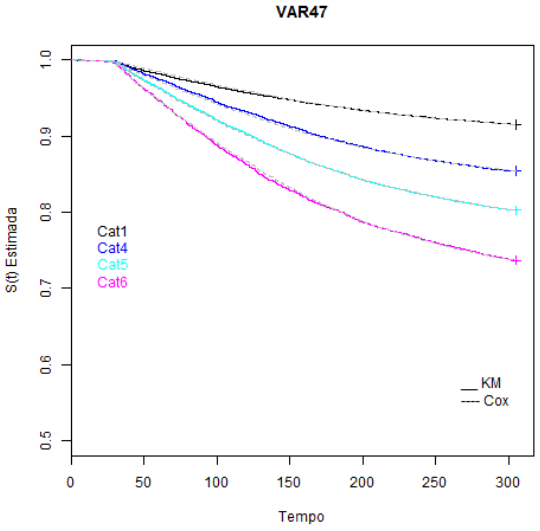

VAR41
S(t) Estimada
Tempo
Cat5
Cat6
Cat7
KM
Cox

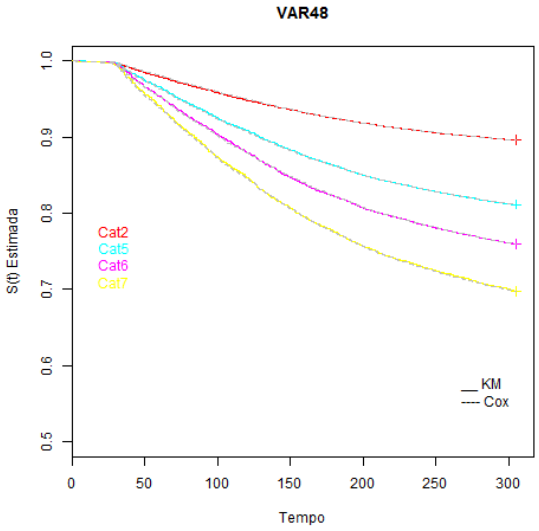

VAR42
S(t) Estimada
Tempo
Cat4
Cat7
KM
Cox

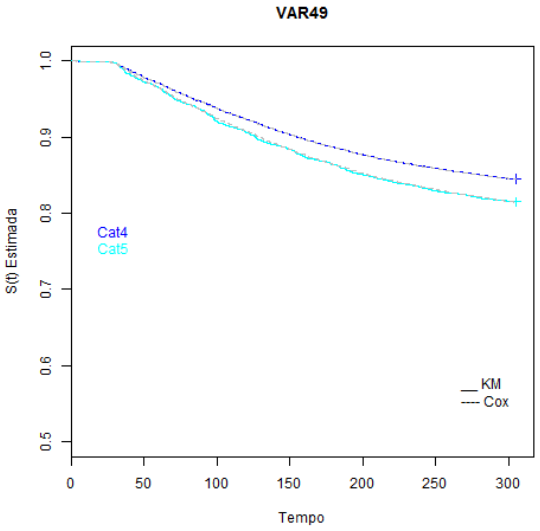

VAR43
S(t) Estimada
Tempo
Cat2
Cat3
Cat5
Cat7
KM
Cox

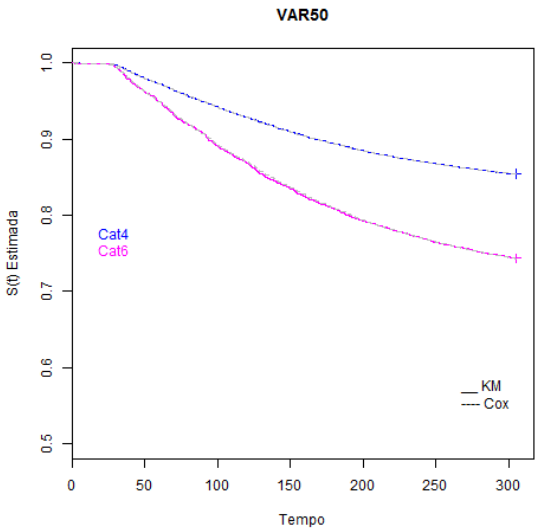

VAR44
S(t) Estimada
Tempo
Cat1
Cat5
Cat6
KM
Cox

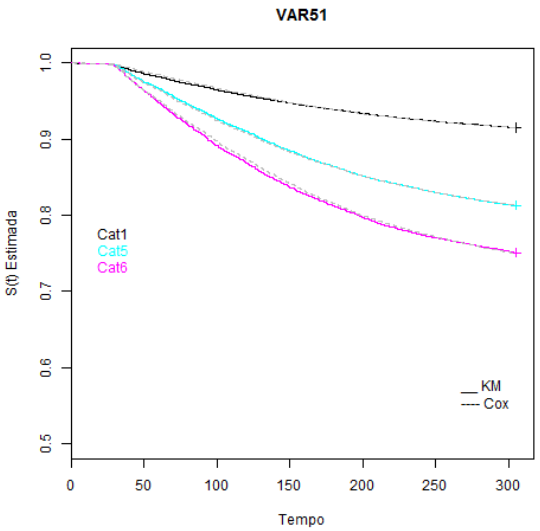

VAR45
S(t) Estimada
Tempo
Cat1
Cat5
Cat6
KM
Cox

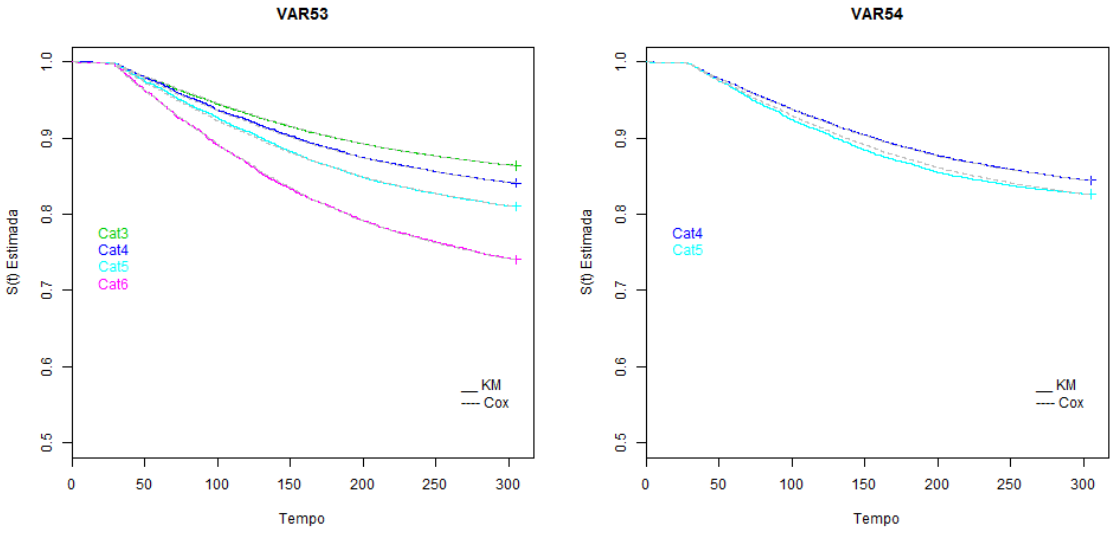

VAR47
S(t) Estimada
Tempo
Cat1
Cat4
Cat5
Cat6
KM
Cox


VAR48
S(t) Estimada
Tempo
Cat2
Cat5
Cat6
Cat7
KM
Cox

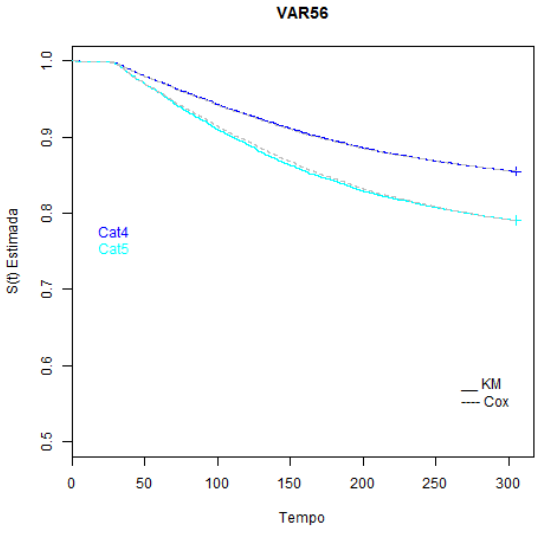

VAR49
S(t) Estimada
Tempo
Cat4
Cat5
KM
Cox

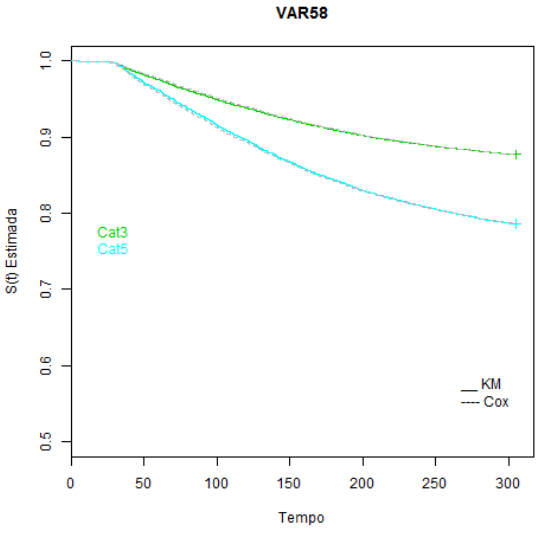

VAR50
S(t) Estimada
Tempo
Cat4
Cat6
KM
Cox

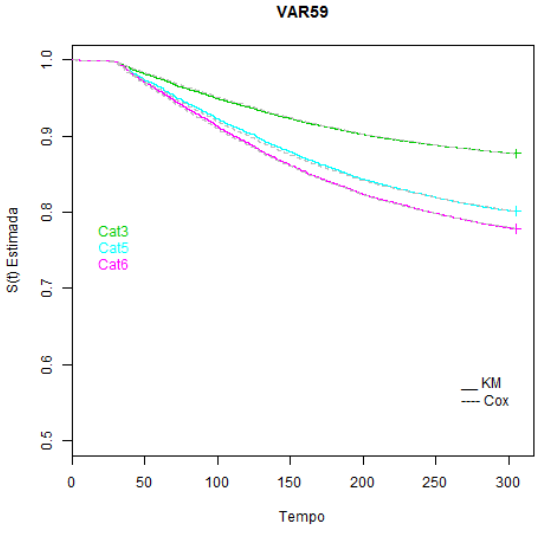

VAR51
S(t) Estimada
Tempo
Cat1
Cat5
Cat6
KM
Cox

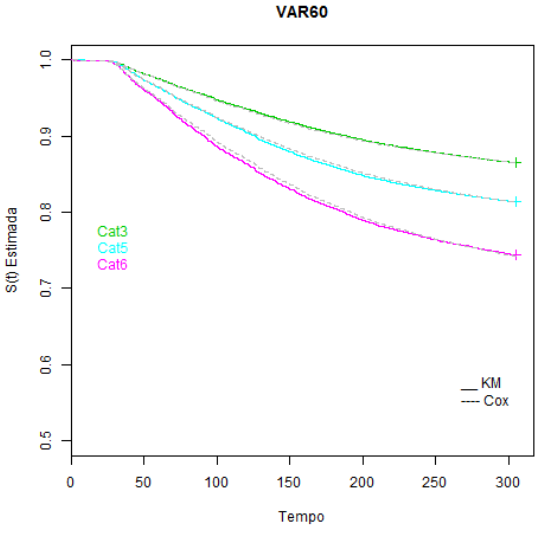

VAR53
S(t) Estimada
Tempo
Cat3
Cat4
Cat5
Cat6
KM
Cox

VAR54
S(t) Estimada
Tempo
Cat4
Cat5
KM
Cox


VAR56
S(t) Estimada
Tempo
Cat4
Cat5
KM
Cox


VAR58
S(t) Estimada
Tempo
Cat3
Cat5
KM
Cox


VAR59
S(t) Estimada
Tempo
Cat3
Cat5
Cat6
KM
Cox


VAR60
S(t) Estimada
Tempo
Cat3
Cat5
Cat6
KM
Cox

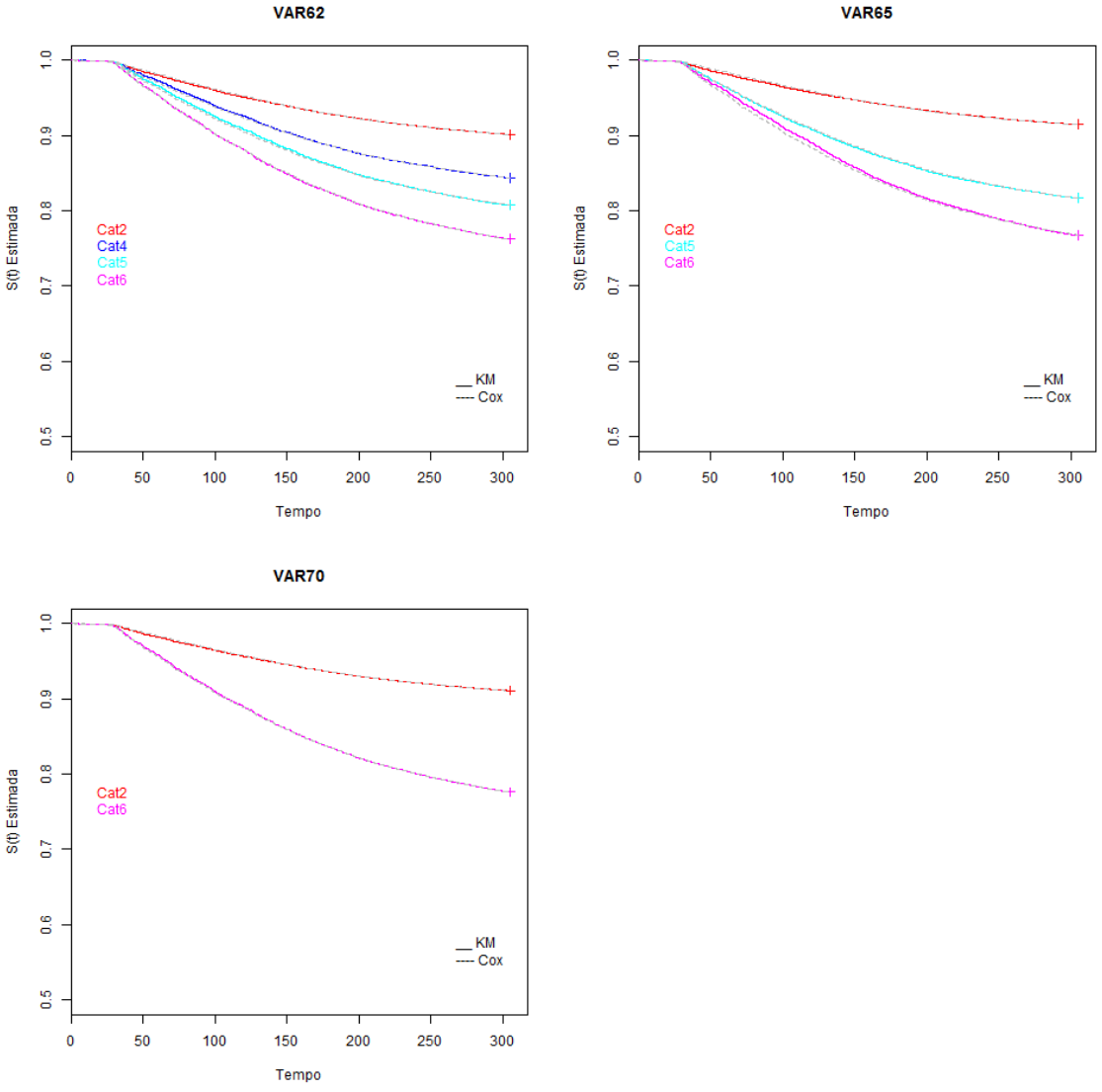
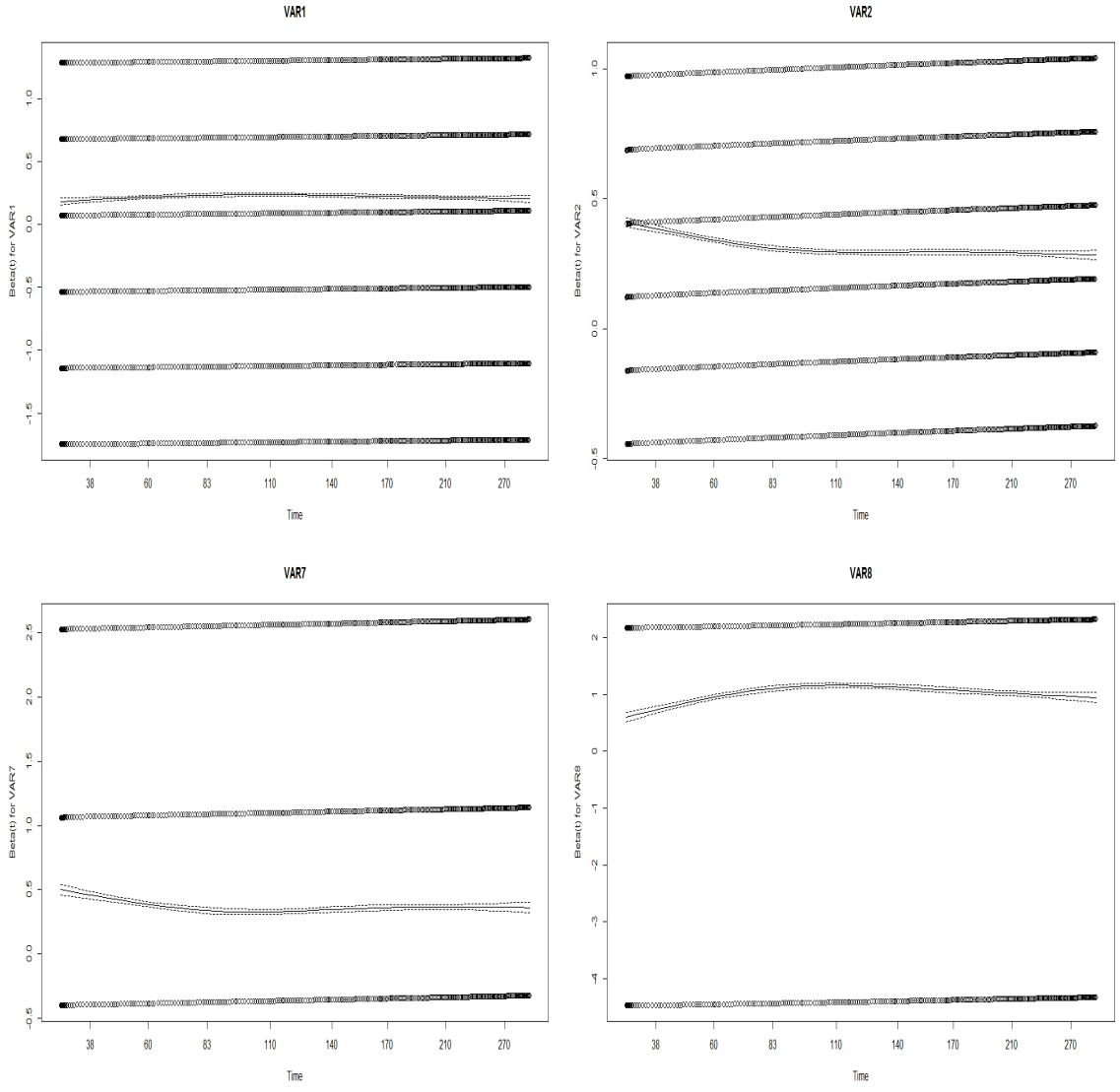
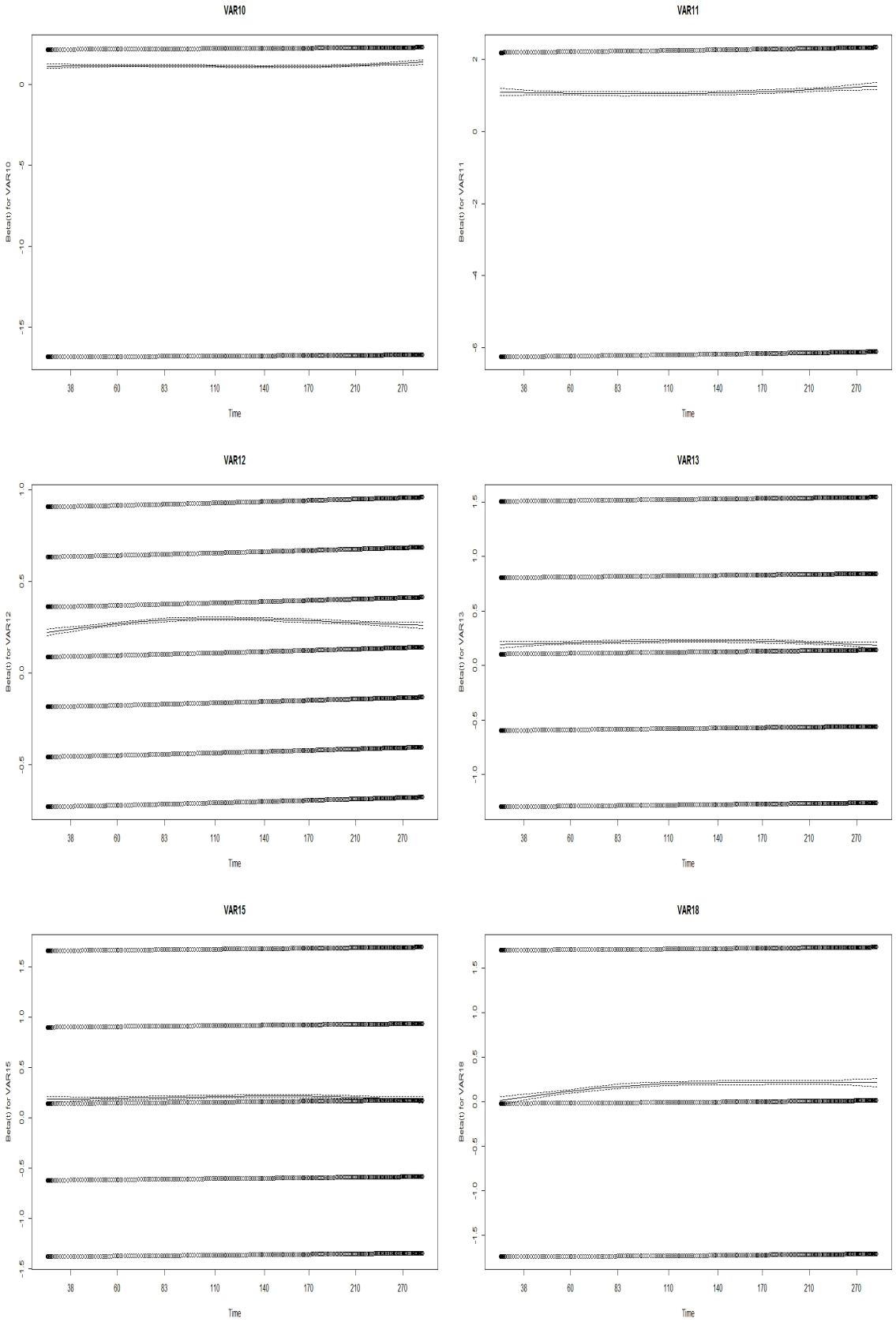


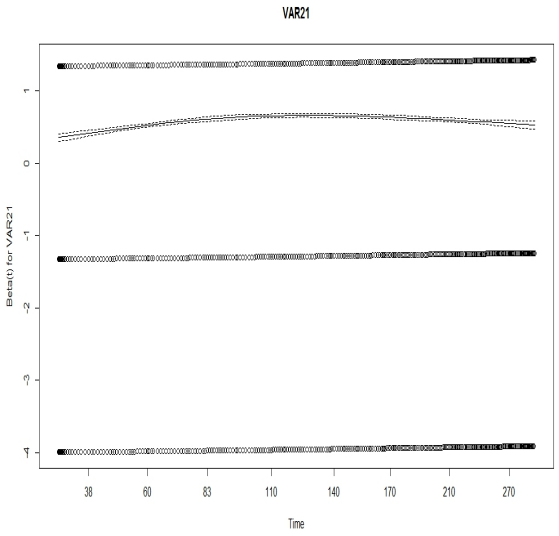
Figura A.1: Curvas de sobrevivência para as covariáveis que entraram nos modelos de Cox e Logístico estimadas por Kaplan-Meier e Cox

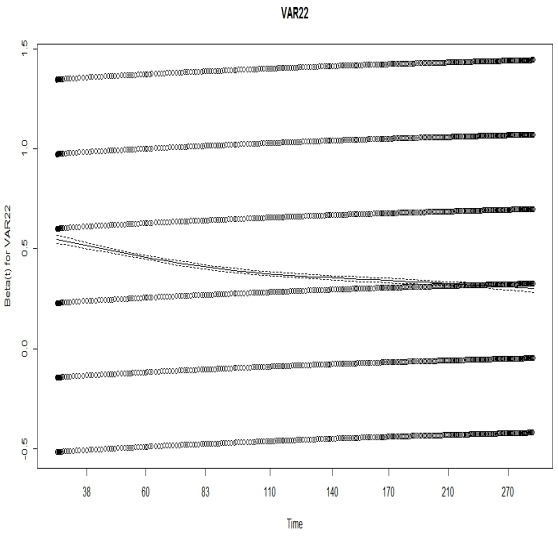
Apêndice B

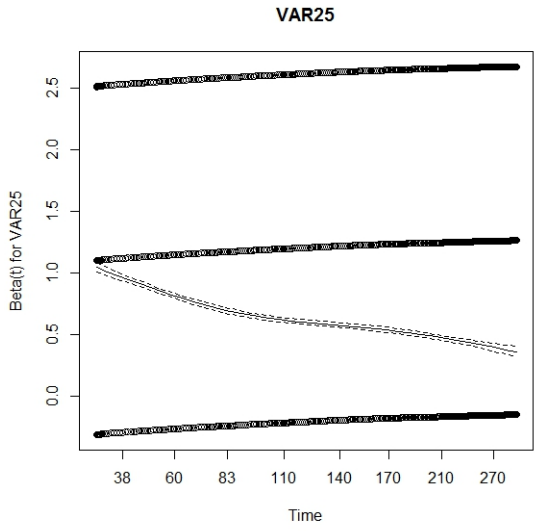
Anexo II - Resíduos Padronizados de Schoenfeld

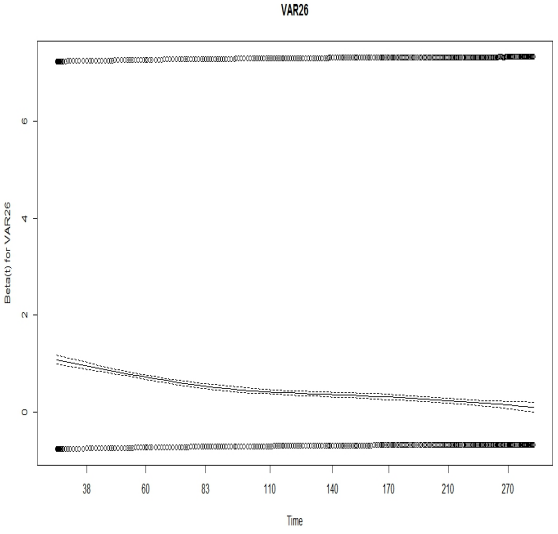


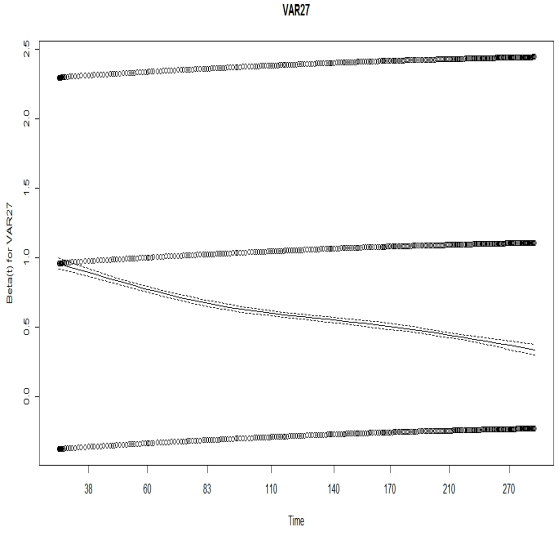

VAR10
Beta(t) for VAR10
0
-5
-10
-15
38
60
83
110
140
170
210
270
Time
VAR11
Beta(t) for VAR11
2
0
-2
-4
-6
38
60
83
110
140
170
210
270
Time
VAR12
Beta(t) for VAR12
1.0
0.5
0.0
-0.5
38
60
83
110
140
170
210
270
Time
VAR13
Beta(t) for VAR13
1.5
1.0
0.5
0.0
-0.5
-1.0
38
60
83
110
140
170
210
270
Time
VAR15
Beta(t) for VAR15
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
38
60
83
110
140
170
210
270
Time
VAR18
Beta(t) for VAR18
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
38
60
83
110
140
170
210
270
Time

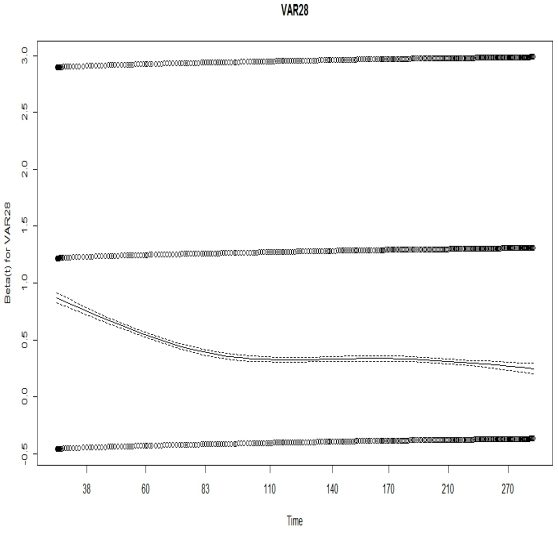

VAR21
Beta(t) for VAR21
Time
38 60 83 110 140 170 210 270

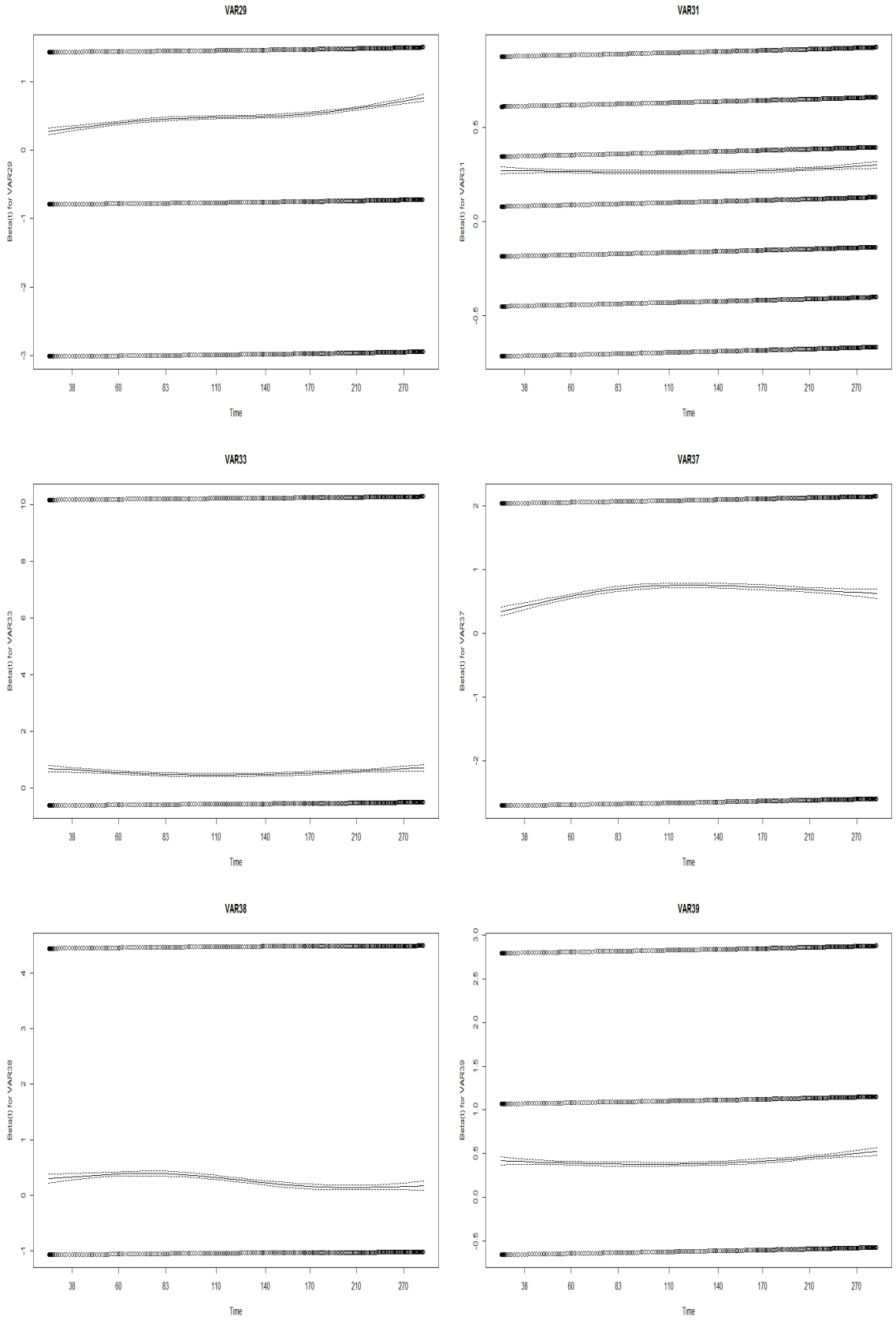

VAR22
Beta(t) for VAR22
Time
38 60 83 110 140 170 210 270

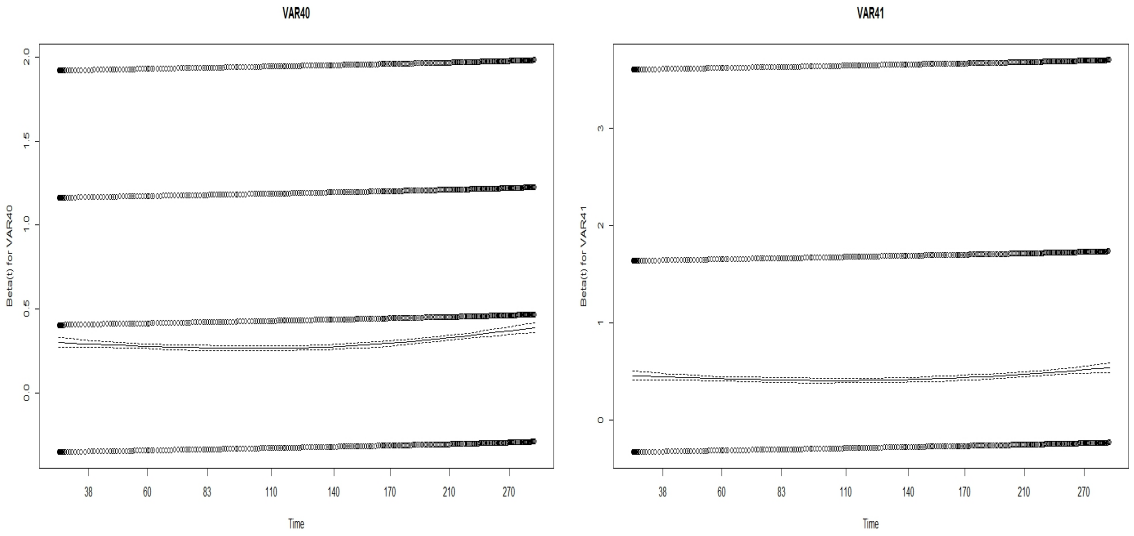

VAR25
Beta(t) for VAR25
Time
38 60 83 110 140 170 210 270


VAR26
Beta(t) for VAR26
Time
38 60 83 110 140 170 210 270

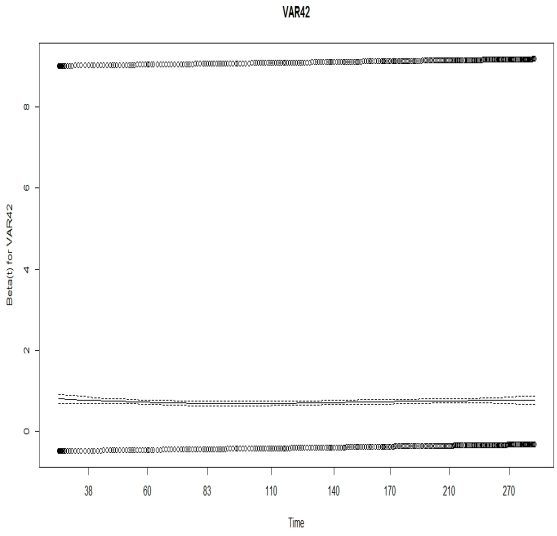

VAR27
Beta(t) for VAR27
Time
38 60 83 110 140 170 210 270

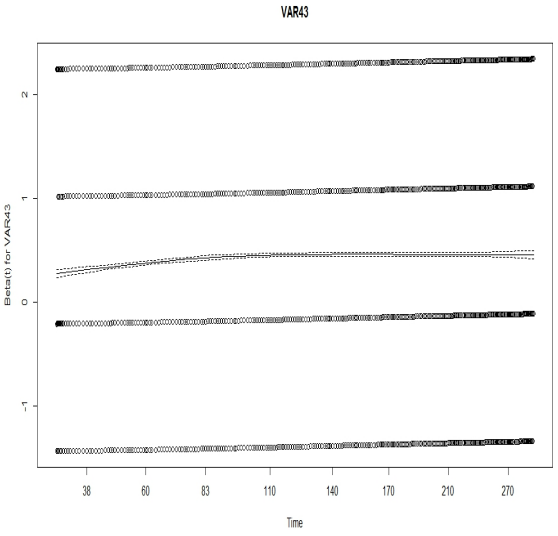

VAR28
Beta(t) for VAR28
Time
38 60 83 110 140 170 210 270

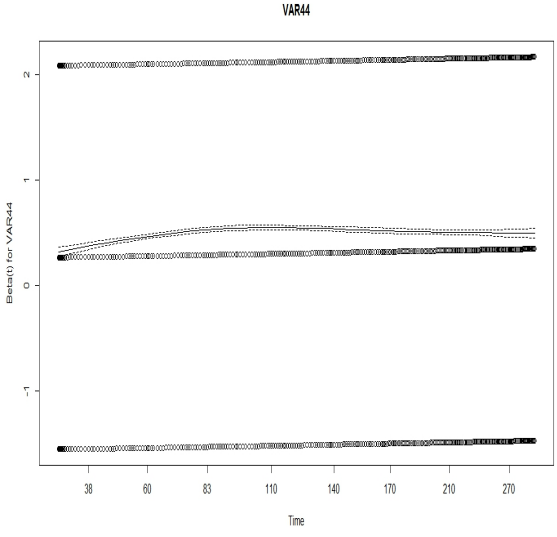

VAR29
Beta(t) for VAR29
Time
VAR31
Beta(t) for VAR31
Time
VAR33
Beta(t) for VAR33
Time
VAR37
Beta(t) for VAR37
Time
VAR38
Beta(t) for VAR38
Time
VAR39
Beta(t) for VAR39
Time

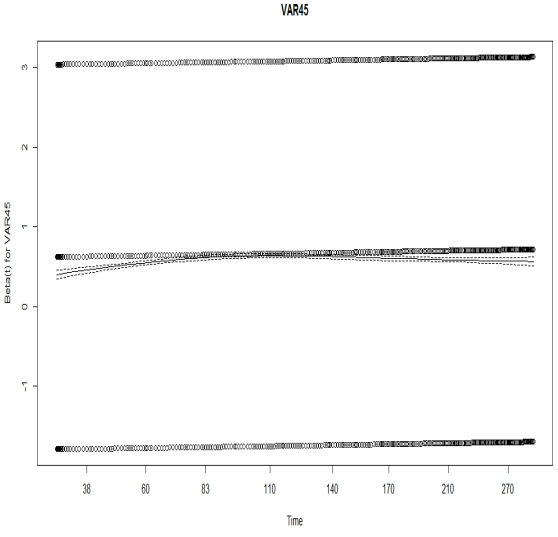

VAR40
Beta(t) for VAR40
Time

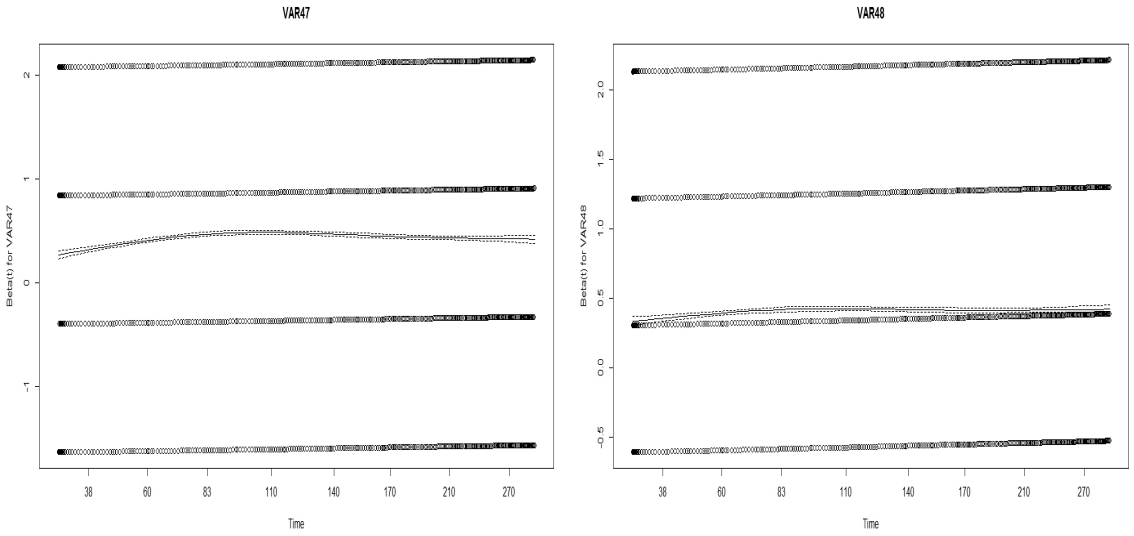
VAR41
Beta(t) for VAR41
Time


VAR42
Beta(t) for VAR42
Time

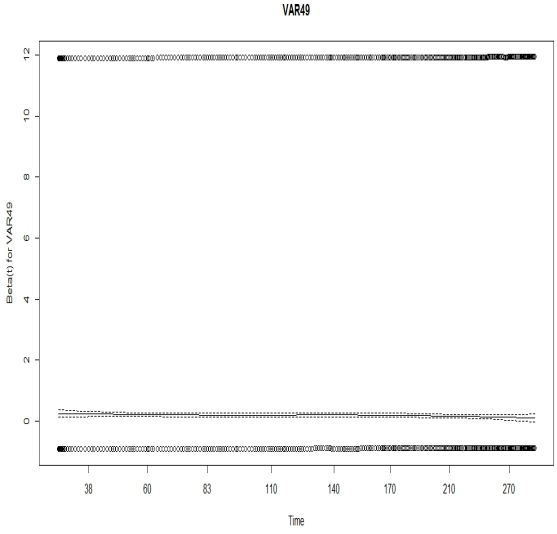

VAR43
Beta(t) for VAR43
Time

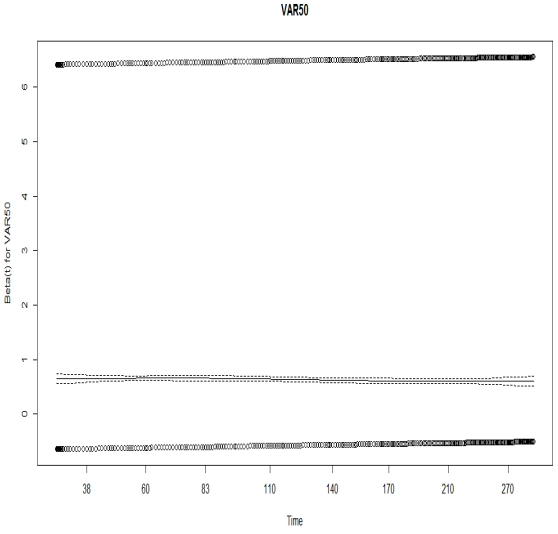

VAR44
Beta(t) for VAR44
Time

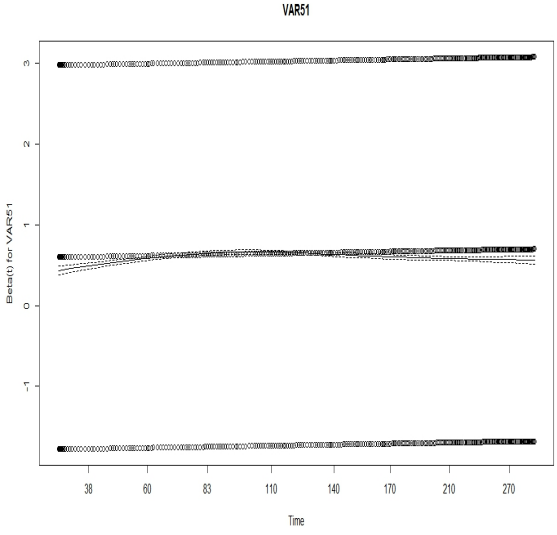

VAR45
Beta(t) for VAR45
Time

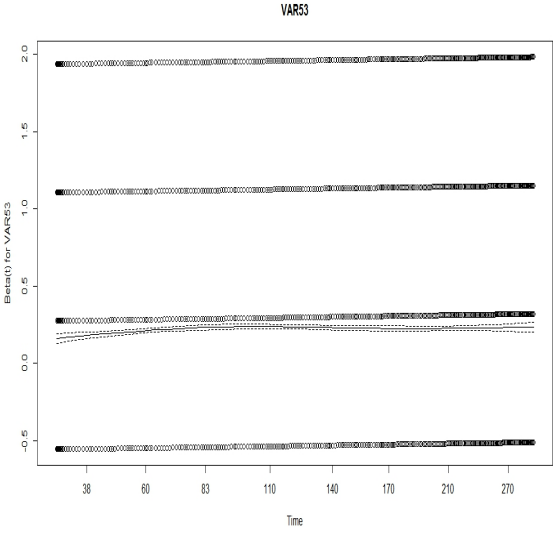

VAR47
Beta(t) for VAR47
Time
38
60
83
110
140
170
210
270

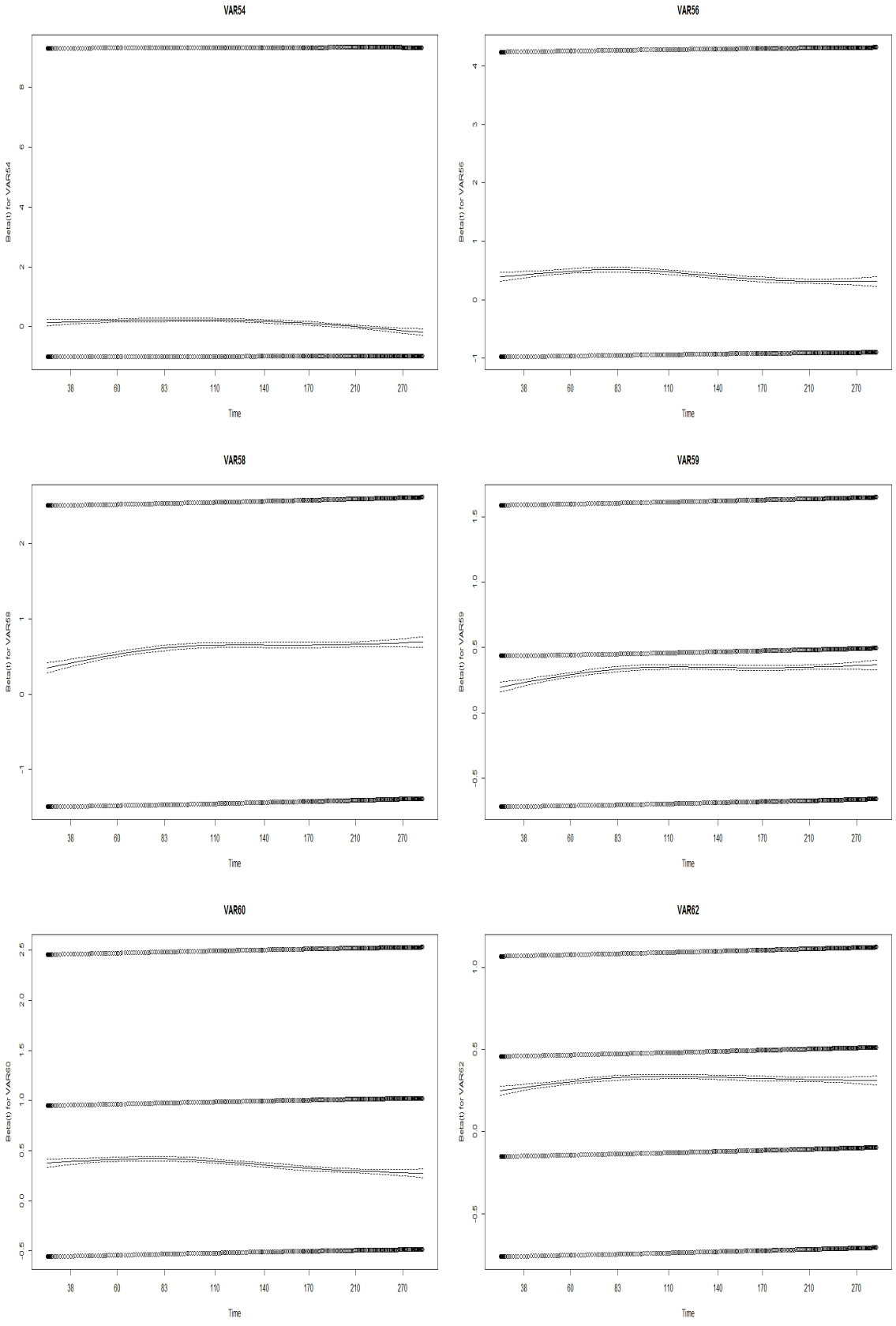
VAR48
Beta(t) for VAR48
Time
38
60
83
110
140
170
210
270


VAR49
Beta(t) for VAR49
Time
38
60
83
110
140
170
210
270


VAR50
Beta(t) for VAR50
Time
38
60
83
110
140
170
210
270


VAR51
Beta(t) for VAR51
Time
38
60
83
110
140
170
210
270


VAR53
Beta(t) for VAR53
Time
38
60
83
110
140
170
210
270


VAR54
Beta(t) for VAR54
Time
VAR56
Beta(t) for VAR56
Time
VAR58
Beta(t) for VAR58
Time
VAR59
Beta(t) for VAR59
Time
VAR60
Beta(t) for VAR60
Time
VAR62
Beta(t) for VAR62
Time

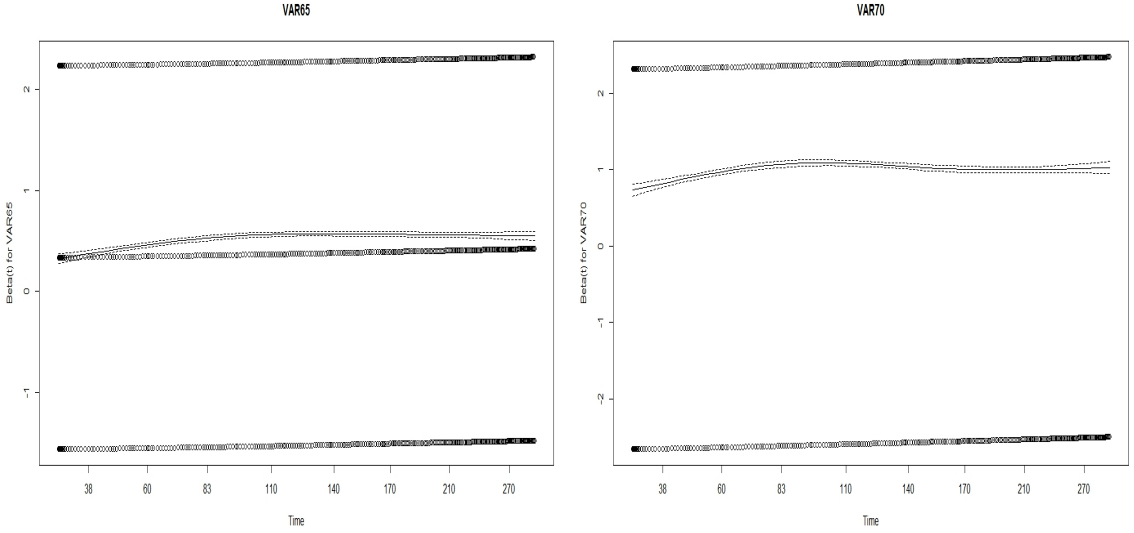


Figura B.1: Resíduos Padronizados de Schoenfeld das covariáveis que entraram no modelo de Riscos Proporcionais