UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

Marco Pollo Almeida

ESTIMAÇÃO BAYESIANA EM MODELOS DE SOBREVIVÊNCIA: UMA APLICAÇÃO EM CREDIT SCORING

Orientadora: Profa. *Dra.* **Maria Regina Madruga Tavares**

Belém
2008

Marco Pollo Almeida

ESTIMAÇÃO BAYESIANA EM MODELOS DE SOBREVIVÊNCIA: UMA APLICAÇÃO EM CREDIT SCORING

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará para obtenção do Título de Mestre em Estatística.

Orientadora: **Profa.** *Dra.* **Maria Regina Madruga Tavares**

Belém
2008

Marco Pollo Almeida

ESTIMAÇÃO BAYESIANA EM MODELOS DE SOBREVIVÊNCIA: UMA APLICAÇÃO EM CREDIT SCORING

Esta dissertação foi julgada e aprovada para a obtenção do título de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará.

Belém, 30 de Junho de 2008

Prof. *Dr.* Mauro de Lima Santos
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística)

Banca Examinadora

Profa. *Dra.* Maria Regina Madruga Tavares
Universidade Federal do Pará
Orientadora

Prof. *Dra.* Marinalva Cardoso Maciel,
Universidade Federal do Pará
Examinadora

Profa. *Dra.* Terezinha Oliveira Morais da Silveira
Universidade Federal do Pará
Examinadora

DEDICATÓRIA

Dedico esse trabalho a meus pais
e meus amados filho e esposa.

Agradecimentos

A Deus por ter permitido que eu chegasse até aqui.

À professora Dra. Regina Madruga por acreditar no tema deste trabalho, pela orientação e por compartilhar seu conhecimento.

À minha esposa pelo seu apoio, compreensão e companheirismo incondicionais.

À minha família pelo apoio, incentivo e compreensão pelo tempo dedicado aos estudos.

Ao meu filho pela candura que vitaliza minha presente existência.

Aos professores e funcionários do Departamento de Estatística e do PPGME da Universidade Federal do Pará pela ajuda, apoio e oportunidade recebidos.

Aos amigos da EMATER-PARÁ pelo apoio e oportunidade concedidos para realização do mestrado.

Em especial aos amigos José Gracildo, Fabio Hipólito, Gustavo Miglio, Pedro Silvestre, Elmer Almeida e Edney Fernandes pela amizade e apoio.

A todos aqueles que indiretamente participaram da realização deste trabalho.

"Our world, our life, our destiny, are dominated by uncertainty; this is perhaps the only statement we may assert without uncertainty."

de Finetti

Resumo

Almeida, Marco Pollo. Estimação Bayesiana Em Modelos De Sobrevivência: Uma Aplicação Em Credit Scoring. 2008. Dissertação (Mestrado em Matemática e Estatística)-PPGME / UFPA, Belém - PA, Brasil.

Nesta dissertação, apresentamos uma análise clássica e uma abordagem bayesiana para obter inferências dos parâmetros de interesse, considerando a distribuição de Weibull para os tempos de sobrevivência. Assumimos a ocorrência de empates dos tempos de sobrevivência e a presença de covariáveis relacionadas com os tempos de sobrevivência com censura à direita. Como aplicação, utilizamos um conjunto de dados reais de uma instituição financeira com objetivo de construir escores contínuos para um sistema de *Credit Scoring* via análise de sobrevivência. No tratamento temporal de *Credit Scoring* via análise de sobrevivência, pretende-se desenvolver escores contínuos voltados à política de concessão de crédito. A partir dos escores estimados, a instituição poderá classificar os clientes com relação ao seu desempenho no pagamento do crédito concedido e propiciar uma avaliação contínua do risco de crédito em quaisquer dos diferentes tempos de relacionamento. O produto de crédito considerado foi o "empréstimo à pessoa física"e as covariáveis selecionadas, via método *stepwise*, para o modelo final ajustado foram sexo, número de parcelas e idade. Os resultados finais da aplicação aos dados reais foram comparados nas abordagens clássica e bayesiana, e observou-se resultados muito similares quando se comparou os modelos baseando-se na medida de desempenho curva ROC.

Palavra-Chave: *Credit Scoring*, Análise de Sobrevivência, Inferência Bayesiana.

Abstract

Almeida, Marco Pollo. Bayesian Estimation in Survival Models: An Application In Credit Scoring. 2008. Thesis (Master in Matematic and Statistic) PPGME/UFPA, Belém - PA, Brazil.

In this paper, we show a classical analysis and a bayesian approach to obtain inferences of the parameters of interest, considering the Weibull distribution to the time of survival. We assumed the occurrence of tied survival times and the presence of covariates related to the right censored survival time. As application, we used a real data set of a financial institution with the objective to build continuous scoring to a credit scoring system via survival analysis. On temporal treatment of credit scoring via survival analysis, we intend to develop continuous scoring related to the credit concession policies. By the estimated scores, the institution could classify the clients according to their performance on the payment of the given credit and favor a continuous assessment of the credit risk in any of the different time relationship. The product of the credit considered was the "personal loan to the customer"and the covariates chosen, via stepwise method, to the final model adjusted were gender, number of parcels and age. The final results of the application to the real data were compared into classical and Bayesian approaches, and it was observed results too similar when it was compared the models based on performance measure: the ROC curve.

Key Words: Credit Scoring, Survival Analysis, Bayesian Inference.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O mercado financeiro é o meio pelo qual o governo, as pessoas e as empresas se interligam no sentido de que cada um desses agentes possa prover suas diferentes necessidades: quem tem muito dinheiro empresta àqueles que têm falta de dinheiro através do mercado financeiro. O mercado financeiro se divide essencialmente em duas partes: o mercado de crédito e o mercado de capitais. Esta pesquisa tem interesse no mercado de crédito, cujo foco principal é a área de concessão de crédito via instituições financeiras, que podem ser os bancos (empresas), ao emprestar para pessoa física (cliente do banco, pessoas e operadoras de cartões de crédito).

As concessões de crédito têm papel fundamental na economia de um país, devido à sua influência direta nos investimentos produtivos, que alavancam o PIB nacional (Pereira, 2004). Esse mercado vem mostrando, nos últimos anos, altas taxas de crescimento o que torna esse ramo rentável e, portanto, atrativo aos interesses das instituições financeiras. Santos e Fama (2007) deixam claro em suas argumentações que as concessões de crédito apresentaram significativo aumento devido à relativa estabilização da economia, êxito no controle da inflação e maior geração de empregos - fatores que interferem diretamente na capacidade de pagamento dos tomadores de crédito. Concomitantemente, ocorre uma exposição maior das instituições financeiras ao risco de inadimplência, isto é, o do não-recebimento, parcial ou total, do pagamento do crédito tomado. Os bancos, em busca de alta rentabilidade (Thomas e Stepanova, 2002), necessitam de métodos que auxiliem na decisão de conceder ou não capital (crédito) a um tomador de crédito (aquele que pede emprestado).

Segundo Anderson (2007), em seu retrospecto sobre a história do crédito, a forma pela qual os créditos eram concedidos aos proponentes, até o início do século XX, era baseado unicamente em julgamentos subjetivos dos analistas de crédito. Em 1936, o estatístico inglês *Sir Ronald Aylmer Fisher* publicou um artigo sobre o uso de uma técnica estatística chamada "Análise Linear Discriminante" usada para classificar diferentes espécies de íris (Fisher, 1936). Durand (1941) mostrou que a mesma técnica poderia ser usada para discriminar bons e maus tomadores de crédito. De acordo com Santos e Fama (2007), atualmente utilizam-se os julgamentos subjetivos em conjunto com os objetivos, sendo este último as técnicas estatísticas. Essas metodologias estatísticas podem ser vistas como as precursoras das que se têm atualmente, tais como, regressão logística, cadeias de *Markov*, redes neurais, algoritmos genéticos etc

Nesse contexto, observa-se a dinâmica do processo de concessão de crédito, o que conduz à busca por uma ferramenta eficaz e rápida, adaptada para absorver a crescente demanda de concessão de créditos sem comprometer a rentabilidade. No trabalho de Pereira (2004) são relatados uma grande variedade de modelos, que vêm sendo construídos com fins específicos por produto e tipo de risco, e estimam a probabilidade de:

(i) Um indivíduo que está solicitando crédito deixar de cumprir seu compromisso antes de completar um período pré-fixado, após a abertura de uma conta ou aquisição de um produto (*application scoring*);

(ii) Clientes que já possuem um determinado produto terem problemas de crédito nos n meses seguintes *behavioural scoring*;

(iii) O risco de crédito do cliente em cada um dos produtos em uma única medida (*Customer Scoring*);

(iv) Dos clientes darem lucro à instituição e ordená-los, basedo nesta probabilidade (*profit scoring*);

(v) Fraudar a instituição (*fraud scoring*);

(vi) Cancelar a conta ou um produto (*attrition scoring*);

(vii) Comprar um produto após uma campanha publicitária (*marketing propensity scoring*);

(viii) E do cliente pagar um empréstimo que já está em atraso (*collection scoring*).

A metodologia de *Credit Scoring* está inserida neste cenário e é amplamente utilizada, devido à sua versatilidade ao reunir a ciência estatística e a subjetividade da experiência dos analistas de crédito. O objetivo é identificar os bons e os maus clientes, baseando-se em suas características (ou variáveis) cadastrais já conhecidas e registradas em alguma base de dados da instituição financeira na qual o cliente está vinculado. Diferenciados os bons dos maus clientes, pode-se relacioná-los ou ordená-los, significativamente, segundo o risco do não pagamento da dívida.

Os modelos de *Credit Scoring* (ou escoragem de crédito) são peculiares no cenário das instituições financeiras, e vêm sendo utilizados como uma das principais técnicas de suporte à concessão de créditos. Resumidamente, *Credit Scoring* é o uso de modelos estatísticos para transformar dados relevantes em medidas numéricas que guiem as decisões de concessão de créditos.

As instituições financeiras estão em um processo de descoberta e desenvolvimento de novas técnicas para auxiliar os sistemas de *Credit Scoring*, e uma das mais recentes é a Análise de Sobrevivência (Colosimo e Giolo, 2006). Na metodologia da análise de sobrevivência, o foco reside no acompanhamento do tempo até a ocorrência de um determinado evento de interesse. No contexto de *Credit Scoring* tal evento seria a inadimplência. Narain (1992) teve a iniciativa de desenvolver um trabalho seguindo esta linha, onde propôs a estimação do tempo que um indivíduo levava para tornar-se inadimplente. A relevância deste paradigma está caracterizada pela necessidade de se saber "quando" o indivíduo se tornará inadimplente, ao invés da idéia tradicional de se conhecer, ao final de um período de acompanhamento do histórico de pagamento, se o mesmo é ou não inadimplente. Banasik *et al.* (1999) faz comparações entre a regressão logística e a análise de sobrevivência, e também demonstra como riscos concorrentes podem ser usados no contexto de escoragem de crédito. Thomas e Stepanova (2002) deram continuidade aos estudos de riscos concorrentes e apresentaram comparações entre as duas abordagens já citadas. Dentre os modelos de sobrevivência vale destacar nesta área, por sua versatilidade, o modelo de regressão de *Cox* (Cox, 1972), também conhecido como modelo de riscos proporcionais ou semiparamétrico de *Cox*, o qual será apresentado em detalhes no capítulo seguinte.

Basicamente, nas propostas de Abreu (2004), Andreeva (2006) e Tomazela (2007), buscou-se avaliar o desempenho dos modelos de sobrevivência ajustados aos sistemas de *Credit Scoring*, com ênfase dada ao modelo semiparamétrico de *Cox*, e compará-los aos modelos de regressão logística. Da mesma forma são implementadas variadas técnicas de modelos paramétricos, semiparamétricos, tempo de vida acelerados, covariáveis dependentes do tempo, entre outros. Vale ressaltar que a literatura sobre *Credit Scoring* com o uso de técnicas de regressão logística é farta, sendo este o procedimento mais utilizado em auxilio à concessão de créditos.

Nesta pesquisa, os dados de escoragem de crédito serão tratados sob a ótica da análise de sobrevivência, pelos motivos já expostos. Além disso, no que tange à parte inferencial, os resultados serão obtidos sob o paradigma da inferência bayesiana. Portanto, o interesse é mostrar a viabilidade de se trabalhar nesta ótica, cujas vantagens são muito significativas, principalmente quando na presença de determinados esquemas de censura complexa e da inviabilidade de usar argumentos assintóticos, como ocorrem na tradicional inferência clássica (Ibrahim *et. al.*, 2001).

As propostas bayesianas no contexto da análise de sobrevivência podem ser encontradas nos trabalhos de Sinha e Dey (1997) e, posteriormente, em Ibrahim *et al.* (2001). No primeiro, investigou-se as potencialidades dos métodos bayesianos semiparamétricos em dados de sobrevivência univariados, na presença de covariáveis fixas e em covariáveis dependentes do tempo, metodologias para tratar dados de tempo de eventos múltiplos e dados de sobrevivência multivariados. O segundo é uma obra de vários autores renomados na área, que mostram parte das técnicas de análise de sobrevivência investigadas no âmbito bayesiano.

1.1 - Justificativa e Importância do Trabalho

A justificativa para a realização deste trabalho deve-se a importância dada às políticas das instituições financeiras em relação à concessão de crédito e, em conseqüência, à avaliação do risco. Portanto, há um forte ensejo em auxiliar essas instituições na concessão e análise do crédito, onde a demanda por instrumentos metodológicos de auxílio à gestão e análise de risco, é grande. Daí a crescente busca por métodos estatísticos de desempenho otimizado.

Além disso, é importante mencionar que a realização deste trabalho se justifica também pelas contribuições na área de inferência bayesiana, observando sua relevância em muitos aspectos. Sob este ponto de vista, Gelman *et al.* (1995) afirma que o uso de métodos bayesianos é uma alternativa vantajosa em relação aos métodos clássicos, pois é livre de certos paradoxos ou violação de princípios, que estão associados com a estatística clássica.

1.2 - Objetivos

1.2.1 - Objetivo Geral

Usar uma metodologia alternativa, análise de sobrevivência, que exponha a necessidade de se obter escores contínuos de apoio à concessão de crédito, no tratamento de um modelo de *Credit Scoring* sob a perspectiva da inferência estatística bayesiana.

1.2.2 - Objetivos Específicos

Os objetivos específicos desta dissertação são:

1. Abordar alguns aspectos estatísticos de um sistema de *Credit Scoring*. Apresentar conceitos associados aos modelos de análise de sobrevivência, paramétricos e semi-paramétricos;

2. Utilizar a metodologia bayesiana para encontrar estimativas das quantidades de interesse na análise de sobrevivência, aplicada ao modelo de *Credit Scoring*: as funções de sobrevivência e de risco;

3. Ajustar um modelo de sobrevivência a dados reais de *Credit Scoring*, apresentando a fórmula de escoragem de acordo com a base de dados obtida;

4. Realizar a validação e verificação da performance do modelo.

1.3 - As Limitações do Trabalho

O presente trabalho encontrou limitações na obtenção da base de dados e na disponibilização de trabalhos na área de análise de sobrevivência via inferência bayesiana aplicada a sistemas de credit scoring. O desenvolvimento de sistemas de credit scoring compreende

etapas de desenvolvimento necessariamente ordenadas. Uma dessas etapas é de fundamental importância nesse contexto. Trata-se da obtenção e tratamento da base de dados. Elas, na maioria das vezes, não são facilmente disponibilizadas por motivos de política de privacidade e sigilo das instituições financeiras. E, em alguns casos, podem ser inadequadas para a construção de sistemas de credit scoring. Estes dois pontos principais em relação à base de dados, se ignorados, podem comprometer completamente o desenvolvimento do sistema.

1.4 - Estrutura do Trabalho

Esta dissertação encontra-se dividida em cinco capítulos, a saber:

- Capítulo 1: apresenta a introdução, importância do trabalho, objetivo e as limitações do trabalho;
- Capítulo 2: Uma descrição geral sobre *Credit Scoring*;
- Capítulo 3: Uma descrição geral sobre Análise de Sobrevivência;
- Capítulo 4: Apresenta a metodologia utilizada em Inferência Bayesiana em Modelos de Sobrevivência;
- Capítulo 5: Apresenta os resultados obtidos em uma aplicação a dados reais;
- Capítulo 6: Apresenta a conclusão e propostas para trabalhos futuros.

Nos próximos Capítulos se fará uma revisão da literatura sobre *Credit Scoring*, Análise de Sobrevivência e Inferência Bayesiana em Modelos de Sobrevivência, buscando dar referências e definições básicas sobre os assuntos pertinentes à dissertação.

Capítulo 2

Credit Scoring

2.1 - Estimando a Probabilidade de Inadimplência

Credit Scoring (Escoragem de Crédito) é, em essência, o uso de modelos estatísticos para transformar dados relevantes, de um tomador de crédito, em medidas numéricas que guiem as decisões de concessão de crédito. Tal método funciona, de modo a reconhecer os diferentes grupos que compõem uma população, que no contexto de escoragem de crédito, trata-se da população de clientes que solicitaram crédito. Existem muitas ferramentas desenvolvidas para este fim, entre elas encontra-se a análise de sobrevivência, como metodologia alternativa, disputando espaço com as mais tradicionalmente utilizadas (pesquisa operacional, regressão logística, inteligência artificial, árvore de decisão, algoritmos genéticos, redes neurais, cadeias de *Markov*, métodos de proximidade etc). Basicamente essas metodologias têm interesse em estimar a probabilidade do evento (inadimplência) ocorrer. O objetivo dessas abordagens é prever quem se tornará inadimplente, de modo que é interessante notar que tantas abordagens diferenciadas possam ser aplicadas ao mesmo problema. Isso se deve ao fato de que os sistemas de *Credit Scoring*, historicamente, sempre se basearam numa abordagem pragmática à questão da concessão de crédito. Os analistas de risco estão voltados para o seguinte lema: se funcionar, use!

Tradicionalmente, toma-se uma amostra de mil a centenas de milhares de clientes que solicitam crédito, de modo a serem acompanhados em um período de observação e para cada cliente da amostra obtêm-se suas variáveis cadastrais (gênero, idade, tempo de emprego, número de filhos, etc). O período de observação, geralmente, é de 12 ou 18 ou 24 meses. Em seguida, define-se quem são os "bons" e os "maus" clientes, considerando que um mau cliente é um tomador de crédito que tenha faltado com pagamento por um período de tempo determinado pela política de crédito-risco das instituições financeiras (Tomazela, 2007), de tal maneira que algumas determinam a falta de pagamento por três

meses consecutivos, outras, 30 dias após o vencimento e, algumas, um dia após a data do vencimento.

Uma questão fundamental neste tipo de problema é o horizonte de tempo adequado para a previsão do sistema de *Credit Scoring* (que é o intervalo de tempo entre a solicitação do crédito e a classificação como bom ou mau cliente). Um período menor que 12 meses subestima a porcentagem de maus clientes e não refletirá as características que permitem prever as inadimplências. E um horizonte com período maior que 24 meses deixa o sistema de *Credit Scoring* suscetível a deslocamentos da população e, com isso, a amostra pode ser diferente daquela em relação à qual o sistema de *Credit Scoring* será utilizado. Estes modelos, essencialmente, são de corte horizontal, ou seja, modelos que ligam dois *flashes* de uma pessoa em diferentes momentos para produzir modelos estáveis, quando analisados longitudinalmente ao longo do tempo.

Outro ponto importante é a respeito da proporção de bons e maus clientes existentes na amostra. Na amostra de desenvolvimento deve haver um número igual de bons e maus clientes, por exemplo, 2000 clientes bons e 2000 clientes maus (Anderson, 2007). Com base nestas características, o sistema de *Credit Scoring* torna-se agora um problema de classificação no qual as variáveis cadastrais são obtidas na base de dados de uma instituição financeira ou de um *Bureau* de referência de crédito (no caso do Brasil temos o Serasa, SPC, etc), de modo que a saída do modelo seja a divisão entre bons e maus clientes. Suponha dividir um dado conjunto C de respostas em dois subconjuntos - um deles contendo as respostas dos clientes que se revelam maus, e o outro por aqueles que se revelam bons. A regra que se aplica aos solicitantes de crédito seria, então, aceitar os que pertencem ao subconjunto dos bons clientes, e rejeitar aqueles que estejam no subconjunto dos maus clientes. Deve-se salientar o fato de que o sistema não será capaz de classificar corretamente todos os elementos da amostra. De qualquer maneira, seria impossível obter uma classificação perfeita uma vez que, em alguns casos, um mesmo subconjunto de respostas é dada por maus e bons clientes. O desejável é um sistema que minimize os erros de classificação. A seguir algumas considerações sobre as metodologias, tradicionais mais conhecidas.

Análise discriminante (Wichern e Johnson, 1988) trata-se de uma técnica que é usada para determinar o número de elementos de um grupo, onde existem dois ou mais grupos conhecidos. A análise discriminante, basicamente, funciona pelo uso de algumas ferramentas de classificação e procura minimizar a distância entre os membros dentro de um grupo, e maximizar as diferenças entre os membros de grupos diferentes. O uso da análise discriminante em *Credit Scoring* usualmente admite o caso simples de dois grupos apenas. Segundo Anderson (2007), análise discriminante recebe muitas críticas contra seu uso pois, facilmente, viola pressupostos básicos do modelo, e é utilizada apenas em variáveis quantitativas. A regressão logística (Hosmer e Lemeshow, 2000) é a técnica estatística mais utilizada no mercado para construção de sistemas de *Credit Scoring* (Anderson, 2007; Thomas e Stepanova, 2002). Sabe-se da regressão logística que a variável resposta é discreta e, na maioria das vezes, é binária. Com essa característica, a regressão logística pode ser utilizada para descrever a relação entre a ocorrência ou não de um evento de interesse e um conjunto de variáveis explanatórias. No contexto de *Credit Scoring*, a variável resposta apresenta-se como o desempenho, em termos de pagamento, dos indivíduos durante um período de tempo determinado e um conjunto de variáveis cadastrais, como já foi citado. Essa metodologia é aplicada em uma amostra adotando-se um horizonte de previsão e considera-se como variável resposta a ocorrência de falta de pagamento dentro desse período, não importando o mês, ou seja, não sendo levado em consideração o momento exato da ocorrência da inadimplência (Tomazela, 2007; Abreu, 2004).

Uma característica da construção de um sistema de *Credit Scoring*, sem levar em conta a técnica usada, é o fato de que a maioria das variáveis cadastrais não são respostas numéricas, mas, categóricas (possui cartão de crédito? possui casa própria? qual o estado civil, etc). Existem muitos métodos estatísticos de classificação, quando se trata de dados categóricos (Tomazela, 2007; Abreu, 2004). As formas de tratamento usadas com as variáveis categóricas também são aplicadas as variáveis quantitativas, como por exemplo, a idade, número de filhos, percentual de comprometimento da renda, entre outras. Segundo Anderson (2007), a arte do *Credit Scoring* está na escolha sensata de categorias. Este objetivo é alcançado por meio de técnicas estatísticas que dividem a variável de forma que o risco de inadimplência seja homogêneo dentro de cada categoria e heterogêneo entre as categorias. Diante disto, uma questão intrigante é: qual o melhor método? Segundo Thomas

e Stepanova (2002) e Anderson (2007), cada um defende a supremacia de sua abordagem e paralelamente, as comparações feitas no âmbito acadêmico são freqüentemente limitadas pelo fato de que a maioria dos dados relevantes, como as pesquisas desenvolvidas por *bureaus* de crédito ou bancos, são muito sigilosas ou dispendiosas para serem repassados. Assim, seus resultados são de natureza meramente indicativa, no entanto, a variação dos erros de classificação dos diferentes métodos é pequena. Diante desta necessidade de se conhecer o quão bom é um modelo, é que surgem as medidas de avaliação de desempenho. Estas são utilizadas para avaliar os modelos ajustados, ou seja, quanto o score produzido pelo modelo consegue distinguir os eventos (bons ou maus pagadores). Uma das idéias envolvidas em medir o desempenho dos modelos está em saber o quão bem eles classificam os clientes. Uma medida de separação muito utilizada nesse contexto é a estatística de *Kolmogorov-Smirnov* (*KS*), assim como a curva *ROC* (*Receiver Operating Characteristic*), ambas são utilizadas nos trabalhos de Abreu (2004), Thomas e Stepanova (2002) e Tomazela (2007). Outras medidas de avaliação de desempenho são propostas no trabalho de Tomazela (2007).

2.2 - Uma Visão Geral Sobre a Comparação de Técnicas de Modelagem Preditiva em *Credit Scoring* : Análise de Sobrevivência *versus* Regressão Logística.

Segundo Anderson (2007) sistemas de escoragem de crédito modelados via regressão logística ainda são evidentemente unânimes como principal procedimento entre as instituições financeiras e seus respectivos analistas de crédito. Muitas propostas alternativas são implementadas em universidades no mundo todo, e sempre trazem comparações em termos de competitividade entre os modelos. Narain (1992) foi um dos primeiros a propor que o uso da análise de sobrevivência poderia ser aplicada à escoragem de crédito como uma alternativa à metodologia usual de regressão logística. A partir disso, alguns anos depois surgiram os trabalhos de Banasik et al. (1999), Thomas e Stepanova (2002), Abreu (2004) e Tomazela (2007) comparando as abordagens de análise de sobrevida e regressão logística; assim como a construção de novos procedimentos.

A presente pesquisa, como afirmado anteriormente, busca implementar uma modelagem via análise de sobrevivência em um modelo de *Credit Scoring*. Portanto, faz-se necessário esclarecer algumas idéias sobre as possíveis vantagens deste modelo. Sob este ponto de vista, Banasik *et al.* (1999) e Louzada-Neto (2005) mostram abordagens significativas sobre a necessidade de se obter scores contínuos de apoio à concessão de crédito. De acordo com os autores, atualmente, os gestores de crédito investigam a propensão à inadimplência de um cliente baseando-se em uma representação discreta do risco de crédito do cliente, ou seja, para o que chamamos de modelo pontual (regressão logística) significa que esperamos até o final do período de desempenho (geralmente fixado em 12 ou 18 meses - intervalo de tempo iniciado na solicitação de crédito até a observação da ocorrência ou não do evento inadimplência) para indicar se o desempenho do cliente foi bom ou mau por meio de uma variável "*flag*" ("bom" ou "mau"), isto é, a classificação dicotômica como bom ou mau pagador. Deste modo, simplifica-se a resposta tal que o tempo (o momento exato do evento) é ignorado.

O detalhe crucial é que apesar dos "pontos de contato" com a instituição de crédito serem discretos (pontuais), o relacionamento do cliente com a instituição financeira desde a sua entrada na base de dados é contínuo.

Logo, é intuitivo pensar em adaptar a técnica de modelagem a uma resposta temporal do cliente à concessão, direcionando os procedimentos estatísticos a uma visão contínua do relacionamento longitudinal do tempo cliente-instituição financeira. A isto se chama modelagem temporal de *Credit Scoring*, e a mesma condiz com os procedimentos de análise de sobrevivência.

Ainda, levando-se em conta a investigação à propensão do cliente à inadimplência, temos agora outro cenário ao considerarmos a modelagem temporal, onde a resposta a essa pergunta é mais sofisticada, a qual nos indica para "quando" o cliente terá propensão à inadimplência, ao invés de simplesmente indicar se ele a terá.

A vantagem da proposta de se utilizar a análise de sobrevida (modelagem temporal) está em permitir a utilização de um score contínuo, ao contrário das técnicas usuais que direcionam para obtenção de um score discreto (pontual), que precisa ser obtido várias vezes, por meio de diversas modelagens, se quisermos ter uma visão longitudinal do com-

portamento do crédito. Segundo Tomazela (2007) e Anderson (2007) na modelagem por Análise de Sobrevivência não há necessidade de fixação de um período único de acompanhamento. Sem o uso da premissa clássica, de período de observação, nesta modelagem obtém-se a probabilidade para todos os instantes até o período máximo observado na amostra. Com isso há uma diminuição da quantidade de modelos, já que não seria mais necessário desenvolver um modelo de previsão de inadimplência para 6 e outros para 12, 18 ou 24 meses, conforme nos modelos ajustados por regressão logística.

Capítulo 3

Análise de Sobrevivência

3.1 - Introdução

A análise de sobrevivência surge como instrumento analítico de modelagem de dados, quando o objeto de interesse é o tempo (Carvalho *et al.* 2005). O tempo é a variável resposta a ser acompanhada, de modo a se estabelecer um tempo de origem bem definido, até a ocorrência de um evento de interesse. Esse evento quando ocorre é chamado de falha.

É de fundamental importância, em dados de sobrevivência, deixar claro e de forma precisa a definição de falha, assim como o início e o fim do acompanhamento (Colosimo e Giolo, 2006). Em estudos na área de sobrevivência nem todos os indivíduos da pesquisa vem a falhar quando chega o fim do estudo. Neste caso, as informações sobre o tempo de sobrevivência desses indivíduos estão incompletas e são chamadas de informações censuradas ou apenas censura. A censura consiste de observações (tempo de sobrevivência) incompletas ou parciais. No caso desta pesquisa serão contempladas as informações com censuras à direita. Uma observação é dita ser censurada à direita em c, se o valor exato da observação não é conhecido, mas sabe-se apenas que este é maior ou igual à c (Hosmer e Lemeshow, 1999).

Formalizando estas idéias deve-se considerar T uma variável aleatória não-negativa e contínua, que representa o tempo de sobrevivência de indivíduos de alguma população. A função densidade de probabilidade de T, denotada por $f(t)$, e sua respectiva função de distribuição é dada por

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{3.1}$$

A probabilidade de um indivíduo sobreviver até um tempo t é dada pela função de sobrevivência:

$$S(t) = 1 - F(t) = P(T > t). \tag{3.2}$$

De acordo com Collet (1994), nota-se que $S(t)$ é uma função decrescente monótona com $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$. A função de risco (*hazard function*), $h(t)$, é a taxa de falha (risco) instantânea no tempo t, e é definida por

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{3.3}$$

Em particular, $h(t)\Delta t$ é a probabilidade de falha aproximada em $(t, t + \Delta t)$, dado que sobrevive até o tempo t. As funções $f(t)$, $F(t)$, $S(t)$ e $h(t)$ dão especificações matematicamente equivalentes das distribuições de T. Pode-se obter expressões para $S(t)$ e $f(t)$ em termos de $h(t)$. Visto que $f(t) = -\dfrac{d}{dt}\, S(t)$ e baseando-se em (3.3), tem-se que

$$h(t) = -\frac{d}{dt} \lg(S(t)) \tag{3.4}$$

Agora, integrando ambos os lados de (3.4) e tomando o exponencial, obtém-se o seguinte

$$S(t) = \exp\left(-\int_0^t h(u)d(u)\right). \tag{3.5}$$

O risco acumulado (*cumulative hazard*), $H(t)$, é definido como

$$H(t) = \int_0^t h(u)d(u),$$

o qual se relaciona com a função de sobrevivência por $S(t) = \exp(-H(t))$. Sabendo-se que $S(\infty) = 0$, segue que $H(\infty) = \lim_{t\to\infty} H(t) = \infty$. Deste modo, a função de risco $h(t)$ tem as seguintes propriedades:

1. $h(t) \geq 0$;
2. $\displaystyle\int_0^\infty h(t)d(t) = \infty$

E obtém-se de (3.4) e (3.5) que

$$f(t) = h(t) \exp\left(-\int_0^t h(u)d(u)\right). \tag{3.6}$$

Geralmente, em estudos com dados de sobrevida, são registradas algumas variáveis (também chamadas de covariáveis) que geram efeitos sobre o tempo de sobrevivência. Com isso, o objetivo principal é estimar o efeito destas covariáveis (variáveis independentes ou preditoras) sobre a variável resposta (ou variável dependente), representada pela variável T. Esta estrutura corresponde à metodologia de um modelo de regressão. No entanto, esta modelagem deve ser adequada para comportar dados censurados (Collet, 1994).

De acordo com Colosimo e Giolo (2006), existem duas classes de modelos na literatura que tratam este tipo de questão: os modelos paramétricos e semiparamétricos. No caso dos modelos paramétricos associa-se uma distribuição de probabilidade à variável aleatória T. Contudo, sendo esta variável uma quantidade contínua e não-negativa, excluímos a possibilidade de modelar o tempo de sobrevivência ao modelo normal, pois o mesmo assume tanto valores positivos quanto negativos. Além disso, geralmente o tempo de sobrevida apresenta-se assimétrico na direção dos maiores tempos de sobrevivência. Isto ocorre porque a maior parte dos tempos observados têm valores pequenos e apresentam, também, poucos indivíduos que registram tempos muito longos.

Uma das alternativas para tratar esses problemas de modelagem da variável resposta, em análise de sobrevida, é utilizar um componente determinístico não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico (Colosimo e Giolo, 2006).

A segunda classe de modelos é conhecida na literatura como modelo de regressão de *Cox* (Cox, 1972). Nesta classe, o fundamento para estimar o efeito das covariáveis é a proporcionalidade dos riscos ao longo de todo o tempo de acompanhamento. Trata-se de um modelo flexível, o que significa dizer que não é necessário fazer qualquer suposição a respeito do tempo de sobrevivência. E, além disso, permite incorporar facilmente covariáveis dependentes do tempo (os valores mudam ao longo do tempo).

3.2 - Modelos Paramétricos

Seguindo as considerações da seção anterior, quanto a associar um modelo probabilístico para o tempo de sobrevivência, as propostas que aparecem com mais freqüência são os modelos exponencial, Weibull e log-normal.

As etapas para construção da modelagem de um modelo paramétrico perpassam pelas seguintes etapas (Colosimo e Giolo, 2006):

1ª **- Descrição do Estudo e das Variáveis:** Neste ponto desenvolve-se a abordagem do problema em questão e levantam-se as perguntas sobre o problema que fomenta o surgimento das variáveis potencialmente importantes;

2ª **- Análise Exploratória dos Dados:** Etapa primordial para qualquer uma análise estatística e que consiste, basicamente, da estatística descritiva das variáveis levantadas;

3ª **- Seleção de Covariáveis:** Dentre as variáveis potencialmente importantes tenta-se ajustar um modelo parcimonioso, que contemple tanto as técnicas implementadas pelos softwares estatísticos, método *Stepwise* (Hosmer e Lemeshow, 1999), quanto pela experiência do estatístico e do pesquisador.

4ª **- Ajuste de um Modelo de Regressão Paramétrico:** Requer a especificação de uma distribuição de probabilidade para a variável resposta.

5ª **- Adequação do Modelo Ajustado:** Análise dos resíduos e suas respectivas particularidades quanto ao ajuste global do modelo, determinação da forma funcional (linear, quadrática etc) de uma covariável e a acurácia do modelo para cada indivíduo sob estudo. Alguns autores sugerem que esta etapa anteceda a interpretação das estimativas dos parâmetros do modelo ajustado.

3.2.1 - Modelo Exponencial

O modelo Exponencial (Colosimo e Giolo, 2006) é o modelo paramétrico de regressão mais simples e historicamente mais utilizado em análise de sobrevida, no entanto, devido a sua simplicidade, poucas situações na prática são adequadamente ajustadas por este modelo. Suponha n tempos de sobrevivência independentes e identicamente distribuídos (*i.i.d.*) representados por $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$, onde cada um admita uma distribuição exponencial com parâmetro (λ). Para esta distribuição usaremos a notação para os indicadores de censura como $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_n)'$, onde $\nu_i = 0$ se y_i for censurado à direita e $\nu_i = 1$ se y_i for um tempo de falha. Sejam

$$f(y_i \mid \lambda) = \lambda \exp(-\lambda y_i)$$

a função densidade de probabilidade de y_i e

$$S(y_i \mid \lambda) = \exp(-\lambda y_i)$$

a função de sobrevivência e $D(n, \boldsymbol{y}, \boldsymbol{\nu})$ os dados observados. Escreve-se a função de verossimilhança dos dados da seguinte forma

$$\begin{aligned} L(D \mid \lambda) &= \prod_{i=1}^{n} f(y_i \mid \lambda)^{\nu_i} S(y_i \mid \lambda)^{1-\nu_i} \\ &= \lambda^d \exp\left(-\lambda \sum_{i=1}^{n} y_i\right), \end{aligned} \tag{3.7}$$

onde: $d = \sum_{i=1}^{n} \nu_i$.

3.2.2 - Modelo Weibull

Na distribuição de *Weibull* supõe-se que o risco não varia linearmente com o tempo. É a distribuição mais utilizada atualmente para modelar tempos de sobrevivência. Ao admitir que o tempo de sobrevivência T tenha uma função densidade de probabilidade (f.d.p.) dada pela Equação (3.8), a seguir

$$f(t) = \begin{cases} \alpha\gamma t^{\alpha-1}\exp(-\gamma t)^{\alpha}, & \text{para } t > 0; \quad \alpha > 0 \text{ e } \gamma > 0 \\ \\ 0, & \text{caso contrário.} \end{cases} \tag{3.8}$$

tem-se a distribuição *Weibull* com parâmetros α e γ, denotada por $T \sim Weibull\,(\alpha;\gamma)$, onde α é o parâmetro que determina a forma da função de risco, e γ, é conhecido como parâmetro de escala da distribuição. Para a distribuição de *Weibull*, $h(t)$ é monotonicamente crescente quando $\alpha > 1$ e monotonicamente decrescente quando $0 < \alpha < 1$.

Agora, suponha que tenhamos n tempos de sobrevida $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$, (*i.i.d.*) de acordo com uma distribuição de *Weibull* como definido em (3.8). Em alguns casos é mais conveniente escrever o modelo acima em termos de uma reparametrização do tipo $\lambda = \log(\gamma)$, levando ao seguinte

$$f(y \mid \alpha;\lambda) = \alpha y^{\alpha-1}\exp(\lambda - \exp(\lambda)y^{\alpha}). \tag{3.9}$$

Com a reparametrização acima, muda-se a notação da distribuição para $Y \sim Weibull$ $(\alpha;\lambda)$. Seja a função de sobrevivência dada por

$$S(y \mid \alpha;\lambda) = \exp(-\exp(\lambda)y^{\alpha}).$$

Desse modo, pode-se escrever a função de verossimilhança da seguinte forma

$$\begin{aligned} L(D \mid \alpha;\lambda) &= \prod_{i=1}^{n} f(y_i \mid \alpha;\lambda)^{\nu_i} S(y_i \mid \alpha;\lambda)^{1-\nu_i} \\ &= \alpha^{d}\exp\left[d\lambda + \sum_{i=1}^{n}\left(\nu_i(\alpha-1)\log(y_i) - \exp(\lambda)y_i^{\alpha}\right)\right], \end{aligned} \tag{3.10}$$

onde d é o mesmo definido em (3.7).

3.2.3 - Modelo Log-Normal

O modelo log-normal é modelo muito usado em análise de sobrevivência para caracterizar tempos de vidas de indivíduos e produtos. Neste modelo admite-se que os logaritmos dos tempos de sobrevida são normalmente distribuídos. Se y_i tem distribuição log-normal com parâmetros (μ, σ^2) , denotado por LN(μ, σ^2), tem-se então

$$f(y_i \mid \mu, \sigma) = (2\pi)^{-\frac{1}{2}}(y_i\sigma)^{-1}exp\left\{-\frac{1}{2\sigma^2}(log(y_i) - \mu)^2\right\} \tag{3.11}$$

Onde μ é a média do logaritmo do tempo de falha, assim como σ é o desvio-padrão. Segundo Colosimo e Giolo (2006), há uma relação entre as distribuições log-normal e normal do mesmo modo à relação existente entre as distribuições de Weibull e do valor extremo. Tal relação torna fácil a apresentação e a análise de dados oriundos da distribuição log-normal. O logaritmo de uma variável com distribuição log-normal de parâmetros μ e σ tem distribuição normal com média μ e desvio-padrão σ. Isso significa que dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, tal que, se considere o logaritmo dos dados em vez dos valores originais.

A função de sobrevivência é dada por

$$S(y_i \mid \mu, \sigma) = 1 - \Phi\left(\frac{log(y_i) - \mu}{\sigma}\right) \tag{3.12}$$

Donde se obtém a função de verossimilhança de (μ, σ)

$$\begin{aligned} L(\mu, \sigma \mid D) \quad &= \prod_{i=1}^{n} f(y_i \mid \mu, \sigma)^{\nu_i} S(y_i \mid \mu, \sigma)^{1-\nu_i} \\ &= (2\pi\sigma^2)^{-\frac{d}{2}} exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\nu_i(log(y_i) - \mu)^2\right\} \\ &\times \prod_{i=1}^{n} y_i^{-\nu_i}\left(1 - \Phi\left(\frac{log(y_i) - \mu}{\sigma}\right)\right)^{1-\nu_i} \end{aligned} \tag{3.13}$$

3.3 - Modelo Semiparamétrico de Riscos Proporcionais

A função de risco depende, na maioria dos casos, tanto do tempo quanto de um conjunto de variáveis (ou covariáveis). O modelo de riscos proporcionais separa estas componentes especificando que o risco para um indivíduo, no instante t, cujo vetor de covariáveis é $\boldsymbol{x}$, será dado por

$$h(t \mid \boldsymbol{x}) = h_0(t) \exp\{G(\boldsymbol{x}, \boldsymbol{\beta})\} \tag{3.14}$$

onde $G(\boldsymbol{x}, \boldsymbol{\beta})$ é uma função positiva e contínua das covariáveis e é escrita na forma exponencial pelo fato de que deve ser positiva. $\boldsymbol{\beta}$ é um vetor de coeficientes de regressão a estimar e está associado às covariáveis. $h_0(t)$ é a função de risco base (ou risco basal) de um indivíduo que possui $G(\boldsymbol{x}, \boldsymbol{\beta}) = 1$ quando $\boldsymbol{x} = \mathbf{0}$.

Convencionou-se admitir que o efeito sobre as variáveis seja multiplicativo, o que implica na seguinte função de risco

$$\begin{aligned} h(t \mid x) &= h_0(t) \exp(\boldsymbol{x}^{'} \boldsymbol{\beta}) \\ &= h_0(t) \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}, \end{aligned} \tag{3.15}$$

onde $\eta = \boldsymbol{x}^{'} \boldsymbol{\beta}$ é chamada de preditor linear. O modelo em (3.15) implica que a razão de riscos para dois indivíduos depende da diferença entre seus preditores lineares em qualquer instante do tempo, ou seja, a razão de riscos de dois indivíduos diferentes é constante ao longo do tempo e é uma função das covariáveis, ou seja, não depende do tempo. Sejam dois indivíduos i e j e suas respectivas funções de risco $h_i(t \mid x_i)$ e $h_j(t \mid x_j)$, então

$$\begin{aligned} \frac{h_i(t \mid x_i)}{h_j(t \mid x_j)} &= \frac{h_0(t) \exp(x_i^{'} \beta)}{h_0(t) \exp(x_j^{'} \beta)} \\ &= \exp\{(x_i^{'} \beta) - (x_j^{'} \beta)\}. \end{aligned} \tag{3.16}$$

Para ajustar o modelo de regressão de *Cox*, parte-se do pressuposto de proporcionalidade descrito em (3.16), de modo que se consegue estimar o efeito das covariáveis sem a necessidade de fazer qualquer suposição da distribuição dos tempos de sobrevivência. Este modelo é dito semiparamétrico porque não se supõe nenhuma distribuição para a função de risco base $h_0(t)$, somente se admite que as covariáveis têm uma ação multiplicativa em relação ao risco e esta é a parte paramétrica do modelo.

Quanto à adequação do modelo de regressão de *Cox*, deve-se considerar como suposição inviolável os riscos proporcionais, de tal forma a atender as seguintes etapas:

1. **Avaliação da Qualidade Geral do Ajuste do Modelo:** Do mesmo modo como nos modelos paramétricos, costuma-se utilizar a análise dos resíduos;

2. **Avaliação da Proporcionalidade dos Riscos:** Nesse contexto surgem técnicas gráficas e testes estatísticos (método com coeficiente dependente do tempo, método com covariável dependente do tempo, etc.).

3.3.1 - Ajuste de um Modelo Semiparamétricos de Riscos Proporcionais

Para ajustar um modelo semiparamétrico, deve-se estimar os coeficientes β' s através do método de máxima verossimilhança, pois, estes coeficientes medem o efeito que as covariáveis exercem sobre a função de risco. Desse modo, constrói-se a função de verossimilhança para um modelo de riscos proporcionais com dados censurados à direita como segue. Suponha que há n sujeitos sob estudo (acompanhamento), e que associado ao *i-ésimo* indivíduo está o tempo de sobrevivência t_i e um tempo de censura c_i. Considera-se que os t_i' s são independentes e identicamente distribuídos (*i.i.d.*) com densidade $f(t)$ e função de sobrevivência $S(t)$. O tempo exato de sobrevivência t_i de um indivíduo será observado se $t_i \leq c_i$. Dessa forma, os dados podem ser representados pelos n pares de variáveis aleatórias $(y_i; \nu_i)$, onde

$$y_i = \min(t_i; c_i) \tag{3.17}$$

e

$$\nu_i = \begin{cases} 1, \text{ se } t_i \leq c_i \\ \\ 0, \text{ se } t_i > c_i. \end{cases} \tag{3.18}$$

então a função de verossimilhança para $(\boldsymbol{\beta}; h_0(.))$ de um conjunto de dados censurados à direita de n indivíduos é

$$\begin{aligned} L(D \mid \boldsymbol{\beta}; h_0(t)) &\propto \prod_{i=1}^{n} [h_0(y_i)\exp(\eta_i)]^{\nu_i} \left(S_0(y_i)^{\exp(\eta_i)}\right) \\ &= \prod_{i=1}^{n} [h_0(y_i)\exp(\eta_i)]^{\nu_i} \exp\left\{ -\sum_{i=1}^{n} \exp(\eta_i) H_0(y_i) \right\}, \end{aligned} \tag{3.19}$$

onde $D = (n, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\nu})$; $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{'}$; $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_n)^{'}$ e $\eta_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}$ é o preditor linear para o indivíduo i, $\boldsymbol{x}_i$ é o vetor de covariáveis para o indivíduo i cuja dimensão é $p \times 1$, $\boldsymbol{X}$ é a matriz $n \times p$ de covariáveis com a *i-ésima* linha $\boldsymbol{x}_i^{'}$, e $S_0(t)$ é a função de sobrevida basal, a qual está relacionada com $h_0(.)$ através da relação

$$\begin{aligned} S_0(t) &= \exp\left(-\int_0^t h_0(u)d(u)\right) \\ &= \exp\left(-H_0(t)\right). \end{aligned}$$

Contudo, no modelo de regressão de *Cox* a presença das duas componentes, paramétrica e semiparamétrica, tornam este método de estimação inapropriado. Em 1975, *Cox* (Cox, 1975) formalizou em seu artigo um método conhecido por *método de máxima verossimilhança parcial*, que resolveria tal problema. Ao considerar n indivíduos de um conjunto de dados de sobrevida, considera-se m tempos de eventos distintos e $n - m$ tempos de sobrevivência censurados à direita. Uma forma de facilitar é admitir que somente um indivíduo morre a cada tempo, ou seja, não há empates nos dados. Denota-se os tempos ordenados e distintos de sobrevivência por $(y_{(1)}, y_{(2)}, \cdots, y_{(m)})$, tal que $y_{(j)}$ seja o *j-ésimo* tempo ordenado de sobrevida. O conjunto de indivíduos que estão em risco no tempo $y_{(j)}$ serão denotados por $\mathcal{R}_j$, tal que $\mathcal{R}_j$ seja o conjunto de indivíduos que não estão em risco e não-censurados apenas no instante antes de $y_{(j)}$. A quantidade $\mathcal{R}_j$ é chamada de conjunto de risco. A *verossimilhança parcial de Cox* para $\boldsymbol{\beta}$ é dada por

$$VP(D \mid \boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp(\boldsymbol{x}_{(j)}^{'}\boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}_j} \exp(\boldsymbol{x}_{\ell}^{'}\boldsymbol{\beta})} \tag{3.20}$$

onde $\boldsymbol{x}_{(j)}$ denota o vetor $p \times 1$ de covariáveis do indivíduo que tem um evento no tempo ordenado de sobrevivência $y_{(j)}$. A somatória no denominador de (3.20) é a soma dos valores de $\exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})$ sobre todos os indivíduos que estão em risco no tempo $y_{(j)}$. Nota-se que o produtório é tomado sobre os indivíduos para quem os tempos dos eventos foram registrados. Indivíduos para os quais os tempos de sobrevida estão censurados, não contribuem para o numerador de (3.20), mas eles entram no somatório dos conjuntos de risco nos tempos dos eventos que ocorrem antes de uma observação censurada. Além

disso, esta verossimilhança depende somente do *ranking* dos tempos dos eventos, sabendo-se que isto determina o conjunto de risco em cada tempo de evento. Conseqüentemente, inferências sobre $\boldsymbol{\beta}$ dependem somente da ordem do posto dos tempos de sobrevivência. A função de verossimilhança em (3.20) ainda pode ser escrita de outra forma

$$VP(D \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}_i} \exp(\boldsymbol{x}_\ell' \boldsymbol{\beta})} \right)^{\nu_i} \tag{3.21}$$

onde ν_i é o mesmo definido em (3.18). As *estimativas de máxima verossimilhança parcial* de $\boldsymbol{\beta}$ podem ser obtidas pela maximização de (3.20) em relação à $\boldsymbol{\beta}$. Estes resultados podem ser obtidos através de métodos numéricos.

Um ponto a ser ressaltado é a questão dos empates. Trata-se da possibilidade de ocorrência de mais de um evento num dado instante. Apesar de no modelo de riscos proporcionais os dados de sobrevivência admitirem que a função de risco seja contínua, e que, sob esta suposição, não é possível ocorrer tempos de sobrevida empatados, na prática tem-se tempos de sobrevivência registrados em dias, meses ou anos, o que conduz ao surgimento de dados empatados (Collet, 1994; Colosimo e Giolo, 2006). Segundo Colosimo e Giolo (2006), existem na literatura duas propostas para tratar de dados empatados. Uma delas é fazer uso de aproximações para a função de verossimilhança parcial no contexto do modelo de riscos proporcionais, e a outra é utilizar modelos de regressão discretos.

Capítulo 4

Inferência Bayesiana em Modelos de Sobrevivência

4.1 - Aspectos Básicos de Inferência Bayesiana

Segundo Paulino *et al.* (2003) e Ibrahim *et al.* (2001), o paradigma bayesiano está baseado na especificação de um modelo de probabilidade para dados que foram observados, $\boldsymbol{D}$, dado um vetor de parâmetros desconhecidos $\boldsymbol{\theta}$, o que nos conduz à obtenção de uma função de verossimilhança $L(D \mid \boldsymbol{\theta})$. Sob essa ótica, $\boldsymbol{\theta}$ é aleatório e possui uma *distribuição a priori* $\pi(\boldsymbol{\theta})$. E, para obter inferências sobre $\boldsymbol{\theta}$, deve-se ter como base o que chama-se de *distribuição a posteriori*, que é obtida através do *Teorema de Bayes*. Portanto, a distribuição *à posteriori* é dada por

$$\pi(\boldsymbol{\theta} \mid D) = \frac{L(D \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} L(D \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{4.1}$$

onde Θ denota o *espaço paramétrico* de $\boldsymbol{\theta}$. De (4.1), diz-se que $\pi(\boldsymbol{\theta} \mid D)$ é *proporcional* à verossimilhança multiplicada pela priori, uma vez que o denominador em (4.1) não depende de $\boldsymbol{\theta}$, ou seja,

$$\pi(\boldsymbol{\theta} \mid D) \propto L(\boldsymbol{\theta} \mid D)\pi(\boldsymbol{\theta}),$$

e nota-se que há nesta sentença uma contribuição importante dos dados observados, como forma de atualização da informação a priori, através da $L(\boldsymbol{\theta} \mid D)$, e uma contribuição da informação a priori quantificada através de $\pi(\boldsymbol{\theta})$. Chama-se de *constante normalizadora* de $\pi(\boldsymbol{\theta})$ ou *distribuição marginal dos dados* ou, ainda, *distribuição preditiva à priori* à quantidade

$$m(D) = \int_{\Theta} L(D \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\theta.$$

Esta quantidade geralmente não tem uma solução analítica simples e, em decorrência disso, afeta $\pi(\boldsymbol{\theta} \mid D)$. Esta dificuldade nos reporta ao uso de métodos numéricos computa-

cionais de obtenção de amostragem da posteriori, assim como métodos de estimação para $m(D)$. Um dos métodos mais populares para se obter amostras de $\pi(\boldsymbol{\theta} \mid D)$ é chamado de *amostrador de Gibbs* (Gelman *et al.*, 1995). O *amostrador de Gibbs* é um poderoso algoritmo de simulação que permite obter amostras de $\pi(\boldsymbol{\theta} \mid D)$ sem se ter conhecimento da distribuição marginal.

Ainda há um aspecto importante a ser relatado sobre o paradigma bayesiano. Trata-se da predição, que é um objetivo importante em problemas que envolvem modelos de regressão. A *distribuição preditiva à posteriori* de um vetor $\boldsymbol{z}$ de observações futuras dado que se tem os dados observados D é definido como

$$\pi(\boldsymbol{z} \mid D) = \int_{\Theta} f(\boldsymbol{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{z} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{4.2}$$

onde $f(\boldsymbol{z} \mid \boldsymbol{\theta})$ denota a densidade amostral de $\boldsymbol{z}$ e $\pi(\boldsymbol{\theta} \mid D)$ é a *distribuição à posteriori* de $\boldsymbol{\theta}$. Nota-se que a Equação (4.2) é a esperança posterior de $f(\boldsymbol{z} \mid \boldsymbol{\theta})$, e dessa forma obter amostragem de (4.2) é feito de maneira simples quando se usa o *amostrador de Gibbs* para $\pi(\boldsymbol{\theta} \mid D)$. Considera-se esta uma boa característica do paradigma bayesiano visto que a Equação (4.2) mostra que predições e distribuições preditivas são fáceis de se obter quando amostras de $\pi(\boldsymbol{\theta} \mid D)$ estejam disponíveis.

Na obra de Ibrahim *et al.* (2001) e no artigo de Sinha e Dey (1997), os autores expõem que nesse contexto, enumera-se algumas vantagens do paradigma bayesiano em relação ao paradigma frequentista (base da abordagem clássica). Na literatura sobre análise de sobrevivência se reconhece a grande dificuldade de ajustar tais modelos, especificamente na presença de esquemas de censura complexa. Com o surgimento do *amostrador de Gibbs* (Gelfand e Smith, 1990; Casella e George, 1992) e outras técnicas de algoritmos de amostragem da linha Monte Carlo Cadeia de *Markov* (MCMC) (Gelfand e Smith, 1990; Casella e George, 1992), o ajuste de modelos de sobrevivência complexos é razoavelmente simples e conta com a disponibilidade do software *Bayesian Inference Using Gibbs Sampler* - WinBugs (http://www.mrc-bsu.cam.ac.uk/bugs/), de fácil implementação dessas técnicas. Além disso, a amostragem MCMC permite obter inferências exatas para quaisquer tamanhos de amostras sem se valer de recursos assintóticos. No paradigma frequentista, a estimativa da variância, por exemplo, requer argumentos assintóticos, os quais podem ser muito complicados para se obter e, em alguns casos, simplesmente não há

solução. Então, existe sempre a questão de que o tamanho da amostra seja grande o suficiente para que a aproximação assintótica seja válida. No paradigma bayesiano acontece exatamente o contrário, a estimativa da variância, assim como qualquer outra medida *à posteriori* surge como um produto derivado do amostrador de *Gibbs* e, por isso, são triviais de se obter depois que as amostras da *distribuição à posteriori* estejam disponíveis.

4.2 - Análise de Sobrevivência sob o Enfoque Bayesiano

A inferência bayesiana ganha, cada vez mais, aceitação como metodologia alternativa por causa do êxito na solução de problemas complexos em diferentes áreas de interesse. Os modelos paramétricos são importantes em análise de sobrevivência, especialmente quando são utilizados os procedimentos bayesianos. No caso semiparamétrico os autores Sinha e Dey (1997) discutem as propostas existentes sobre as técnicas bayesianas para tratar a componente não-paramétrica do modelo de regressão de *Cox*. Os autores apresentam uma priori muito comum utilizada para modelar o componente determinístico do modelo semiparamétrico de Cox. Essa priori é popularmente conhecida na literatura como processo gama. Dias (2002) em sua tese de doutorado trata de forma abrangente de modelos de riscos proporcionais sob o paradigma bayesiano, focando em modelos cuja função de risco basal é especificada por um processo de incremento gama independente.

4.3 - Modelos Paramétricos

Baseando-se na Seção 3.2, aqui definimos algumas condições para construção da abordagem bayesiana dos modelos paramétricos vistos na referida Seção.

4.3.1 - Modelo Exponencial

De acordo com a Subseção 3.2.1, *a priori conjugada* para λ é *a priori gama*. Então seja uma distribuição Gama denotada por $Gama(\alpha_0, \lambda_0)$ e densidade dada por

$$\pi(\lambda \mid \alpha_0, \lambda_0) \propto \lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda).$$

Toma-se uma priori $Gama(\alpha_0, \lambda_0)$ para λ, logo a distribuição posterior de λ é dada por

$$\begin{aligned}
\pi(\lambda \mid D) &\propto L(D \mid \lambda)\pi(\lambda \mid \alpha_0, \lambda_0) \\
&\propto \left[\lambda^{\sum_{i=1}^{n} \nu_i} \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)\right] \left(\lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda)\right) \qquad (4.3) \\
&= \lambda^{\alpha_0 + d - 1} \exp\left[-\lambda\left(\lambda_0 + \sum_{i=1}^{n} y_i\right)\right].
\end{aligned}$$

E reconhece-se o núcleo (*kernel*) da distribuição posterior em (4.3) como uma distribuição $Gama\left(\alpha_0 + d; \lambda_0 + \sum_{i=1}^{n} y_i\right)$. Sua média e variância são respectivamente

$$E(\lambda \mid D) = \frac{\alpha_0 + d}{\lambda_0 + \sum_{i=1}^{n} y_i} \qquad (4.4)$$

e

$$Var(\lambda \mid D) = \frac{\alpha_0 + d}{\left(\lambda_0 + \sum_{i=1}^{n} y_i\right)^2}. \qquad (4.5)$$

A distribuição preditiva posterior de um tempo de falha y_f futuro é dado por

$$\begin{aligned}
\pi(y_f \mid D) &= \int_0^\infty \pi(y_f \mid \lambda)\pi(\lambda \mid D) d\lambda \\
&\propto \int_0^\infty \lambda^{\alpha_0 + d + 1 - 1} \exp\left[-\lambda\left(y_f + \lambda_0 + \sum_{i=1}^{n} y_i\right)\right] d\lambda \\
&= \Gamma(\alpha_0 + d + 1)\left(\lambda_0 + \sum_{i=1}^{n} y_i + y_f\right)^{-(d + \alpha_0 + 1)} \qquad (4.6) \\
&\propto \left(\lambda_0 + \sum_{i=1}^{n} y_i + y_f\right)^{-(d + \alpha_0 + 1)}.
\end{aligned}$$

Para construir um modelo de regressão, deve-se introduzir as covariáveis em λ, e escreve-se $\lambda_i = \varphi(\boldsymbol{x}_i^{'}\boldsymbol{\beta})$, onde $\boldsymbol{x}_i$ é o vetor de covariáveis $p \times 1$, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de coeficientes de regressão, e $\varphi(.)$ é uma função conhecida.

Uma forma para φ é tomar

$$\varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \text{ ou } \varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) = (\boldsymbol{x}_i'\boldsymbol{\beta})^{-1}.$$

E obtém-se, desse modo, a função de verossimilhança

$$\begin{aligned} L(\boldsymbol{\beta} \mid D) &= \prod_{i=1}^{n} f(y_i \mid \lambda_i)^{\nu_i} S(y_i \mid \lambda_i)^{(1-\nu_i)} \\ &= \prod_{i=1}^{n} \left[\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \exp(-y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))\right]^{\nu_i} \left[\exp(-y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))\right]^{(1-\nu_i)} \\ &= \exp\left[\sum_{i=1}^{n} (\boldsymbol{\nu}_i \boldsymbol{x}_i'\boldsymbol{\beta})\right] \exp\left[-\sum_{i=1}^{n} (y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))\right]. \end{aligned} \tag{4.7}$$

Define-se a partir da Equação (4.7), $D = (n; \boldsymbol{y}; \boldsymbol{X}; \boldsymbol{v})$, onde $\boldsymbol{X}$ é uma matriz $n \times p$ das covariáveis com a *i-ésima* linha $\boldsymbol{x}_i'$.

As distribuições *a priori* para $\boldsymbol{\beta}$ mais utilizadas são *prioris* uniforme impróprias, isto é, $\pi(\boldsymbol{\beta}) \propto 1$, e a *priori* normal. No cenário da regressão, não estão disponíveis formas analíticas fechadas para as distribuições *à posteriori* de $\boldsymbol{\beta}$, e para isso, utiliza-se os métodos MCMC.

Suponha que se especifique uma *priori* normal *p-dimensional* para $\boldsymbol{\beta}$, denotada por $N_p(\boldsymbol{\mu}_0; \Sigma_0)$, onde $\boldsymbol{\mu}_0$ denota a média *à priori* e Σ_0 denota a matriz de covariância *à priori*. Então a distribuição *à posteriori* de $\boldsymbol{\beta}$ é dada por

$$\pi(\boldsymbol{\beta} \mid D) \propto L(\boldsymbol{\beta} \mid D)\pi(\boldsymbol{\beta} \mid \boldsymbol{\mu}_0; \Sigma_0), \tag{4.8}$$

onde $\pi(\boldsymbol{\beta} \mid \boldsymbol{\mu}_0; \Sigma_0)$ é a densidade normal multivariada com média $\boldsymbol{\mu}_0$ e matriz de covariância Σ_0. A posteriori (4.8) não tem uma forma fechada na maioria das vezes, e os métodos MCMC são necessários para obtenção de amostras da distribuição *à posteriori* de $\boldsymbol{\beta}$.

4.3.2 - Modelo de Weibull

De acordo com a Subseção 3.2.2, quando se supõe α conhecido, a *priori conjugada* para $\exp(\lambda)$ é a *priori gama*. Nenhuma *priori conjugada* é conhecida quando (α, λ) são ambos parâmetros desconhecidos. Diante disso, geralmente, especifica-se uma *priori conjunta* tomando α e λ independentes, estabelecendo que α tenha uma *priori* $Gama(\alpha_0, \kappa_0)$ e λ tenha uma *priori* $N_p(\mu_0, \sigma_0^2)$. Logo, a distribuição *à posteriori conjunta* para (α, λ) é

$$\begin{aligned}
\pi(\alpha, \lambda \mid D) &\propto L(\alpha, \lambda \mid D)\pi(\alpha \mid \alpha_0, \kappa_0)\pi(\lambda \mid \mu_0, \sigma_0^2) \\
&\propto \left[\prod_{i=1}^{n} f(y_i \mid \alpha, \lambda)^{v_i} S(y_i \mid \alpha, \lambda)^{(1-\nu_i)}\right] \pi(\alpha \mid \alpha_0, \kappa_0)\pi(\lambda \mid \mu_0, \sigma_0^2) \\
&= \alpha^{\alpha_0+d-1} \exp\left\{ d\lambda + \sum_{i=1}^{n} \left(\nu_i(\alpha-1)\log(y_i) - \exp(\lambda) y_i^{\alpha}\right) \right. \\
&- \left. (\kappa_0\alpha) - \left(\frac{1}{2\sigma_0^2}\right)(\lambda - \mu_0)^2 \right\}. \qquad (4.9)
\end{aligned}$$

A distribuição posterior conjunta de (α, λ) não tem uma forma fechada, mas pode ser mostrado que as distribuições posteriores condicionais $(\alpha \mid \lambda, D)$ e $(\lambda \mid \alpha, D)$ são *log-côncavas* e, por esse motivo, torna-se simples o amostrador de *Gibbs* para esse modelo. Analogamente ao modelo exponencial, a construção do modelo de regressão paramétrico de *Weibull*, introduz as covariáveis através de λ e escreve-se $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. As distribuições *à priori* para $\boldsymbol{\beta}$ incluem a *priori* imprópria uniforme, $\pi(\boldsymbol{\beta}) \propto 1$, e a *priori* normal. Admitindo-se uma *priori* para $\boldsymbol{\beta}$ do tipo $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ e uma *priori gama* para α, obtém-se a seguinte distribuição *posterior conjunta*

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \alpha \mid D) &\propto \alpha^{\alpha_0+d-1} \exp\left\{ \sum_{i=1}^{n} (\nu_i \boldsymbol{x}_i'\boldsymbol{\beta} + \nu_i(\alpha-1)\log(y_i) \right. \\
&- y_i^{\alpha}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})) - \kappa_0\alpha \\
&- \left. \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\left(\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right) \right\}. \qquad (4.10)
\end{aligned}$$

Geralmente, não há distribuições *à posteriori* para $\boldsymbol{\beta}$ de forma fechada, e novamente se recorre aos métodos MCMC. Para o modelo de regressão de *Weibull* pode-se mostrar que as distribuições condicionais posteriores de $(\alpha \mid \beta, D)$ e $(\beta \mid \alpha, D)$ são *log-côncavas*, o que torna a implementação do amostrador de *Gibbs* mais simples.

4.3.3 - Modelo Log-Normal

De acordo com a seção 3.2, aqui se apresentam as características do modelo log-normal sob a ótica bayesiana. Seja $\tau = \frac{1}{\sigma^2}$. Não há nenhuma priori conjugada e conjunta quando (μ, τ) são ambos desconhecidos. Neste caso, a especificação de uma priori conjunta é tomar $\mu \mid \tau \sim N(\mu_0, \frac{1}{\tau\tau_0})$ e $\tau \sim Gama(\frac{\alpha_0}{2}, \frac{\lambda_0}{2})$. Sob esta formulação a priori conjugada e conjunta para (μ, τ) é dada por

$$\begin{aligned} \pi(\mu, \tau \mid D) \quad & \propto L(\mu, \sigma \mid D)\pi(\mu, \tau \mid \mu_0, \tau_0, \alpha_0, \lambda_0) \\ & \propto \tau^{\frac{\alpha_0+d}{2}-1} exp\left\{ -\frac{\tau}{2}\left[\sum_{i=1}^{n} \nu_i (log(y_i) - \mu)^2 + \tau_0(\mu - \mu_0)^2 + \lambda_0 \right] \right\} \quad (4.11) \\ & \times \prod_{i=1}^{n} y_i^{-\nu_i} (1 - \Phi(\tau^{\frac{1}{2}}(log(y_i) - \mu)))^{1-\nu_i} \end{aligned}$$

A distribuição a posteriori conjunta de (μ, τ) não tem forma analiticamente tratável. Ao contrário dos modelos exponencial e Weibull, gerar amostras de $\pi(\mu, \tau \mid D)$ não é simples. Pode-se mostrar que a distribuição a posterior condicional de $[\mu \mid \tau, D]$ é log-côncava. No entanto, a distribuição a posterior condicional de $[\tau \mid \mu, D]$ em geral não é log-côncava. Portanto, nesse caso é necessário utilizar o algoritmo de Metropolis-Hastings (Gelfand e Smith, 1990; Casella e George, 1992) para obter amostras de $[\tau \mid \mu, D]$.

Do mesmo modo que os modelos anteriores, para se construir o modelo de regressão, introduzimos covariáveis através de μ e se usa $\mu_i = x_i'\beta$. Usualmente prioris dos tipos uniforme imprópria e normal são usadas para β. Considerando que $\beta \mid \tau \sim N_p(\mu_0, \tau^{-1}\sum_0)$, a posteriori conjunta para (β, τ) é dada por

$$\pi(\beta,\tau \mid D) \quad \propto \tau^{\frac{\alpha_0+d}{2}-1} exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n}\nu_i(log(y_i)-x_i'\beta)^2+(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)+\lambda_0\right]\right\}$$
$$\times\prod_{i=1}^{n} y_i^{\nu_i}\left(1-\Phi\left(\tau^{\frac{1}{2}}(log(y_i-x_i'\beta))\right)\right)^{1-\nu_i} \tag{4.12}$$

A distribuição a priori analiticamente não tratável na equação acima precisa utilizar métodos MCMC ou integração numérica. Para o modelo de regressão log-normal pode-se mostrar que a distribuição condicional a posteriori de $[\beta \mid \alpha, D]$ é log-côncava. Para o caso da obtenção de amostras de $[\tau \mid \beta, D]$ usa-se o algoritmo de Metropolis-Hastings.

4.4 - Modelo de Riscos Proporcionais

Para construir esse modelo, considerado um dos mais populares no contexto da classe dos modelos semiparamétricos de sobrevida, necessita-se fazer uma partição finita do eixo do tempo, $0 < s_1 < s_2 < \cdots < s_j$, com $s_j > y_i$ para todo $i = 1, 2\cdots, n$. Desse modo, teremos J intervalos $(0, s_1], (s_1, s_2], \cdots, (s_{J-1}, s_J]$. No *j-ésimo* intervalo considera-se um risco basal constante $h_o(y) = \lambda_j$ para $y \in I_j = (s_{J-1}, s_J]$. E seja $D = (n, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})$ os dados observados e suas componentes (como visto nas Seções anteriores), apenas ressaltando que $\nu_i = 1$ se o *i-ésimo* indivíduo *falhou* e 0, *caso contrário*. Denota-se $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_J)'$ e pode-se escrever a função de verossimilhança de $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ para os n indivíduos como a seguir

$$\begin{aligned} L(D \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \prod_{i=1}^{n}\prod_{j=1}^{J}(\lambda_j \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{(\delta_{ij}\nu_i)} \\ &\times \exp\left\{ -\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) \right.\right. \\ &+ \left.\left. \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1}) \right] \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \right\}, \end{aligned} \tag{4.13}$$

onde $\delta_{ij} = 1$ se o *i-ésimo* indivíduo *falhou* ou foi *censurado* no *j-ésimo* intervalo e 0, *caso contrário*, $\boldsymbol{x}_i' = (x_{i1}'; x_{i2}'; \cdots; x_{ip}')$ denota o vetor $p \times 1$ de covariáveis para o *i-ésimo* indivíduo, e $\boldsymbol{\beta} = (\beta_1; \beta_2; \cdots; \beta_p)'$ é o vetor que corresponde aos coeficientes de regressão.

O indicador δ_{ij} é necessário para definir bem a verossimilhança sobre os J intervalos. O modelo semiparamétrico em (4.13) é conhecido também como *modelo exponencial segmentado*, trata-se de um modelo bem geral e que pode acomodar várias formas do risco basal sobre os intervalos.

Uma priori para o risco basal $\boldsymbol{\lambda}$ é a *priori gama* independente, $\lambda_i \sim Gama(\alpha_{oj}, \lambda_{oj})$ para $j = 1, 2, \cdots, J$. Aqui α_{oj} e λ_{oj} são hiperparâmetros *à priori*, os quais podem ser elicitados através da média e da variância *à priori* de λ_j. Outra abordagem é construir uma correlação *à priori* entre os λ_j' s usando uma *priori* correlacionada $\boldsymbol{\psi} \sim N(\boldsymbol{\psi}_0; \Sigma_\psi)$, onde $\psi_j = \log(\lambda_j)$ para $j = 1, 2, \cdots, J$.

A função de verossimilhança em (4.13) está baseada em dados contínuos de sobrevivência. A função de verossimilhança baseada em dados empatados (*grouped* ou *tied data*) é dada por

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D) \propto \prod_{j=1}^{n} G_j^*, \tag{4.14}$$

onde

$$\begin{aligned} G_j^* &= \exp\left\{ -\lambda_j \Delta_j \sum_{\kappa \in \mathcal{R}_j - \mathcal{D}_j} \exp(\boldsymbol{x}_k' \boldsymbol{\beta}) \right\} \\ &\times \prod_{l \in \mathcal{D}_j} \left\{ 1 - \exp\left[-\lambda_j \Delta_j \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) \right] \right\}, \end{aligned} \tag{4.15}$$

onde $\Delta_j = s_j - s_{j-1}$, $\mathcal{R}_j$ é o conjunto de indivíduos em risco e $\mathcal{D}_j$ é o conjunto de pacientes em que a falha está ocorrendo no *j-ésimo* intervalo.

4.4.1 - O Processo Gama

O processo *gama* é, talvez, o processo *à priori* não-paramétrico mais comumente usado para o modelo de *Cox* (Dias, 2002). O processo *gama* é descrito como segue. Seja $G(\alpha, \lambda)$ denotar a distribuição *gama* com parâmetro de forma $\alpha > 0$ e parâmetro de escala $\lambda > 0$. Seja $\alpha(t)$, $t \geq 0$, uma função contínua e crescente à esquerda tal que $\alpha(0) = 0$, e seja $Z(t)$, $t \geq 0$, um processo estocástico com as seguintes propriedades:

(i) $Z(0) = 0$;

(ii) $Z(t)$ tem incrementos independentes em intervalos disjuntos; e

(iii) Para $t > s, (Z(t) - Z(s)) \sim G\left(c(\alpha(t) - \alpha(s)); c\right)$.

Então o processo $\{Z(t) : t \geq 0\}$ é chamado de processo *gama* e denotado por $Z(t) \sim PG(c\alpha(t), c)$. Notamos aqui que $\alpha(t)$ é a média do processo e c é o peso ou parâmetro de confiança sobre a média. Os caminhos da amostra do processo *gama* são quase certamente funções crescentes. É um caso especial de um *processo de Levy* (Ibrahim *et al.*, 2001) cuja função característica é dada por

$$E\{\exp[iy(Z(t) - Z(s))]\} = \{\phi(y)\}^{(c(\alpha(t)-\alpha(s)))}, \tag{4.16}$$

onde ϕ é a função característica de uma função distribuição infinitamente divisível com média unitária. O processo *gama* é caso especial em que

$$\phi(y) = \left[\frac{c}{(c - (i \times y))}\right]^c.$$

4.4.2 - Um Processo Gama Sobre o Risco Acumulado $H_0(t)$

Sob o modelo de *Cox*, a probabilidade conjunta de sobrevivência de n indivíduos dada uma matriz $\boldsymbol{X}$ de covariáveis é dada por

$$P(\boldsymbol{Y} > y \mid \boldsymbol{\beta}, \boldsymbol{X}; H_0) = \exp\left\{ -\sum_{j=1}^{n} \exp(\boldsymbol{x}_j'\boldsymbol{\beta}) H_0(y_j) \right\}. \tag{4.17}$$

O processo *gama* é freqüentemente usado como *priori* para a função de risco basal acumulada $H_0(y)$. Neste caso, tomamos

$$H_0 \sim PG(c_0 H^*, c_0), \tag{4.18}$$

onde $H^*(y)$ é uma função crescente com $H^*(0) = 0$. H^* freqüentemente é admitido ser uma função paramétrica conhecida com vetor de hiperparâmetro $\boldsymbol{\gamma}_0$. Por exemplo, se H^* corresponde a uma distribuição exponencial, então $H^*(y) = \boldsymbol{\gamma}_0 y$, onde $\boldsymbol{\gamma}_0$ é um hiperparâmetro especificado. Se $H^*(y)$ é tomado como uma *weibull*, então $H^*(y) = \eta_0 y^{\kappa_0}$, onde $\boldsymbol{\gamma}_0 = (\eta_0, \kappa_0)'$ é um vetor especificado de hiperparâmetros. A função de sobrevivência marginal é dada por

$$P(\boldsymbol{Y} > y \mid \boldsymbol{\beta}, \boldsymbol{X}, \gamma_0, c_0) \;=\; \prod_{j=1}^{n} \left[\ \phi(iV_j)\ \right]^{c_0(H^*(y_{(j)}) - H^*(y_{(j-1)}))}, \tag{4.19}$$

onde

$$V_j = \sum_{l \in \mathcal{R}_j} \exp(\boldsymbol{x}_l^{'}\boldsymbol{\beta}),$$

$\mathcal{R}_j$ é o conjunto das observações sob risco no instante $y_{(j)}$ e $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ são tempos ordenados e distintos. Para dados contínuos, quando os tempos de sobrevivência ordenados são todos distintos (não há empates), a função de verossimilhança de $(\boldsymbol{\beta}, \boldsymbol{\gamma}_0, c_0)$ pode ser obtida pela diferenciação de (4.19). Note que esta verossimilhança é definida somente quando os tempos de sobrevida observados são distintos (não há empates).

4.4.3 - Uso do Processo Gama para Verossimilhança com Dados Empatados

Novamente, construímos uma partição finita do eixo do tempo, $0 < s_1 < s_2 < \cdots < s_J$, com $s_J > y_i$, para todo $i = 1, 2, \cdots, n$. Desse modo, temos J intervalos disjuntos $(0, s_1], (s_1, s_2], \cdots, (s_{(J-1)}, s_{(J)}]$, e seja $I_j = (s_{(J-1)}, s_{(J)}]$. Os dados observados D estão supostamente agrupados (empatados) dentro desses intervalos, tal que $D = (\boldsymbol{X}, \mathcal{R}_j, \mathcal{D}_j : j = 1, 2, \cdots, J)$, onde $\mathcal{R}_j$ é o conjunto das observações sob risco e $\mathcal{D}_j$ é o conjunto de falha do *j-ésimo* intervalo I_j. Seja h_j o incremento no risco basal acumulado no *j-ésimo* intervalo, isto é

$$h_j = H_0(s_j) - H_0(s_{j-1}), \quad j = 1, 2, \cdots, J.$$

O processo *a priori gama* em (4.18) implica que os h_j' s são independentes e

$$h_j \sim G(\alpha_{0j} - \alpha_{0,j-1}, c_0), \tag{4.20}$$

onde $\alpha_{0j} = c_0 H^*(s_j)$, e c_0 e H^* foram definidos acima. Deste modo os hiperâmetros (c_0, H^*) para h_j consistem de uma função de risco acumulada paramétrica especificada $H^*(y)$ calculada nos extremos dos intervalos de tempo, e um escalar positivo c_0 quantificando o grau de confiança da *priori* em $H^*(y)$. Agora, escrevendo

$$H_0 \sim PG(c_0 H^*; c_0)$$

indica que todo incremento disjunto em H_0 tem a *priori* dada por (4.20). Assim, a representação dos dados agrupados (empatados) pode ser obtida como

$$\begin{aligned} P(y_i \in I_j \mid \boldsymbol{h}) &= \exp\left\{ -\exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta}) \sum_{k=1}^{j-1} h_k \right\} \\ &\times \left\{ 1 - \exp[-h_j \exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})] \right\}, \end{aligned} \tag{4.21}$$

onde $\boldsymbol{h} = (h_1, h_2, \cdots, h_J)'$. Isto conduz à função de verossimilhança de dados grupados

$$L(\boldsymbol{\beta}, \boldsymbol{h} \mid D) \propto \prod_{j=1}^{J} G_j, \tag{4.22}$$

onde

$$\begin{aligned} G_j &= \exp\left\{ -h_j \sum_{\in \mathcal{R}_j - D_j} \exp(\boldsymbol{x}_k' \boldsymbol{\beta}) \right\} \\ &\times \prod_{l \in D_j} \left\{ 1 - \exp\left[-h_j \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) \right] \right\}, \end{aligned} \tag{4.23}$$

Note que a expressão da verossimilhança dos dados grupados em (4.23) é muito geral e não limitada ao caso quando os h_j' s são realizações de um processo gama sobre H_0. Visto que a função de risco basal acumulada H_0 entra na função de verossimilhança em (4.23) somente através dos h_j' s, nossos parâmetros na verossimilhança são $(\boldsymbol{\beta}, \boldsymbol{h})$ e deste modo nós somente precisamos de uma distribuição *à priori* conjunta para $(\boldsymbol{\beta}, \boldsymbol{h})$. Um caso importante é que, quando se leva em consideração o risco basal constante segmentado, que já foi feito referência, com $h_j = \Delta_j \lambda_j$ e $\Delta_j = s_{(j)} - s_{(j-1)}$. Neste caso, observa-se uma grande similaridade entre as verossimilhanças em (4.15) e (4.23). Na ausência de covariáveis, (4.23) se reduz a

$$G_j = \exp\{-h_j(r_j - d_j)\}\{1 - \exp(-h_j)\}^{d_j}$$

onde r_j e d_j são os números de indivíduos nos conjuntos $\mathcal{R}_j$ e $\mathcal{D}_j$, respectivamente.

Uma *priori* para $\boldsymbol{\beta}$ é uma distribuição $N_p(\boldsymbol{\mu}_0, \Sigma_0)$. Assim, *a posteriori conjunta* de $(\boldsymbol{\beta}, \boldsymbol{h})$ pode ser escrita como

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{h} \mid D) &\propto \prod_{j=1}^{J} \left[G_j h_j^{(\alpha_{(0j)} - \alpha_{(0,j-1)}) - 1} \exp(-c_0 h_j) \right] \\ &\times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \Sigma_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right\} \end{aligned} \tag{4.24}$$

Para amostrar a distribuição *posteriori conjunta* de $(\boldsymbol{\beta}, \boldsymbol{h})$, pode-se mostrar que $[\boldsymbol{\beta} \mid \boldsymbol{h}, D]$ é *log-côncava* em β. Sobretudo, $[\boldsymbol{h} \mid \beta, D]$ também é *log-côncava* nos componentes de $\boldsymbol{h}$. Deste modo, pode-se conduzir o seguinte esquema de amostragem de *Gibbs*:

(i) Amostrar de

$$\pi(\beta_j \mid \boldsymbol{\beta}^{(-j)}, \boldsymbol{h}, D) \propto \prod_{j=1}^{J} G_j \exp\left\{ -\frac{1}{2}(\beta - \boldsymbol{\mu}_0)\Sigma_0^{-1}(\beta - \boldsymbol{\mu}_0) \right\}; \tag{4.25}$$

usando o algoritmo da rejeição adaptativa (Dias, 2002 e Gilks & Wild, 1992) para $j = 1; 2; \cdots; p$.

(ii) Amostrar de

$$\begin{aligned} \pi(h_j \mid \boldsymbol{h}^{(-j)}, \boldsymbol{\beta}, D) &\propto h_j^{(\alpha_{(0j)} - \alpha_{(0,j-1)})-1} \prod_{l \in D_j} \left[1 - \exp\left\{ -h_j \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) \right\} \right] \\ &\times \exp\left\{ -h_j \sum_{\kappa \in \mathcal{R}_j - \mathcal{D}_j} \left(\exp(\boldsymbol{x}_k' \boldsymbol{\beta}) \right) + c_o \right\}; \end{aligned} \tag{4.26}$$

onde $\boldsymbol{h}^{(-j)}$ denota o vetor $\boldsymbol{h}$ sem a *j-ésima* componente. A distribuição condicional completa em (4.26) pode ser bem aproximada por uma distribuição *gama*, e desse modo o esquema de amostragem de *Gibbs* mais eficiente poderia substituir em (4.26) por:

(ii*) Amostrar de $[\boldsymbol{h} \mid \beta, D]$ usando amostras independentes de uma posteriori condicional aproximada por

$$h_j \sim G\left(\left(\alpha_{(0j)} - \alpha_{(0,j-1)}\right) + d_j,\ c_0 + \sum_{\kappa \in \mathcal{R}_j - \mathcal{D}_j} \exp(x_k' \beta) \right). \tag{4.27}$$

Capítulo 5

Aplicabilidade e Resultados

5.1 Introdução

Neste capítulo, a metodologia bayesiana para modelos de sobrevivência é aplicada em um conjunto de dados reais de um problema de Credit Scoring de uma instituição financeira, que atua no mercado brasileiro. O modelo ajustado via análise de sobrevivência fornece uma resposta diferenciada dos modelos tradicionais: a probabilidade de um indivíduo superar um determinado instante sem apresentar o evento inadimplência. A vantagem competitiva reside na tomada de decisão mais acertada, acarretando uma menor exposição a riscos (inadimplência).

Geralmente, no contexto dos sistemas de Credit Scoring, as informações que se tem disponíveis para relacionar com a inadimplência do produto de crédito são as próprias características dos clientes e, algumas vezes, do produto também.

A base fundamental desse estudo foram os trabalhos realizados por Abreu (2004) e Tomazella (2007) onde os dois têm o objetivo de mostrar como algumas idéias de análise de sobrevivência poderiam ser aplicadas no contexto de Credit Scoring.

5.2 Descrição da Amostra

Foi utilizada uma base de dados (amostra) de 1.800 clientes que iniciaram a utilização de um produto de crédito entre janeiro de 2006 a dezembro de 2007, sendo que para 890 deles não foi observado problema de pagamento do crédito e 910 tornaram-se inadimplentes.

Para o registro da ocorrência ou não de algum problema de crédito e conseqüente classificação dos clientes em bons ou maus pagadores, os mesmos foram observados durante os 12 meses seguintes ao início da sua contratação do produto. Neste trabalho o produto

considerado é o empréstimo pessoal, dirigido apenas à pessoa física, através de uma determinada quantia monetária adquirida na instituição financeira, com o compromisso de pagamento parcelado de 6 até 36 vezes. A definição de inadimplência está baseada na especificidade do produto de crédito adquirido e da política da instituição financeira, que classifica o cliente como inadimplente quando o mesmo apresenta atraso de pagamento um dia após a data de vencimento de uma determinada parcela.

Vale ressaltar a grande dificuldade de se obter tais informações das instituições financeiras, devido ao caráter sigiloso das mesmas. Portanto, um ponto limitador nesse estudo foi o tamanho da amostra, que não apresentou tamanho suficiente para se obter uma boa divisão em amostra de desenvolvimento e amostra de validação.

As informações disponíveis na base de dados, são as covariáveis que estão na Tabela 5.1. Dentre estas, serão adicionadas a variável tempo até a ocorrência do evento inadimplência (medida em meses) e a variável *status*, que será o indicador de falhas (ν_i). Tal falha será a inadimplência, ou seja, o não pagamento de uma das parcelas. Sendo $\nu_i = 1$ se o cliente falhou e $\nu_i = 0$ se houve censura.

Tabela 5.1 - Variáveis Disponíveis na Base de Dados.

Variável
Número de Parcelas (de 6 até 36 meses)
Valor Contratado (R$)
Tempo até ocorrer o evento (em meses)
Idade
Número de Dependentes
Estado Civil
Escolaridade
Sexo (masculino; feminino)
Renda (R$)
Status (0 ou 1)

Essas informações são de clientes para os quais já foram observados os desempenhos de pagamento do crédito adquirido, e servirão para a construção do modelo preditivo, a partir da metodologia de análise de sobrevivência sob o enfoque bayesiano, a ser aplicado em futuros clientes, permitindo que esses possam ser ordenados segundo uma probabilidade

de inadimplência e, a partir dessa ordem, as políticas de crédito da instituição possam ser definidas.

Sabe-se que na metodologia de análise de sobrevivência se investiga a influência que as covariáveis exercem sobre a variável resposta (tempo até ocorrer o evento inadimplência). Alguns modelos inseridos no contexto de análise de sobrevivência focam seu interesse no risco de ocorrência de um evento em um determinado tempo, após o início do acompanhamento de um cliente quando inicia a utilização de um determinado serviço de crédito.

Tomazela (2007) cita:

> Em um modelo de análise de sobrevivência estudam-se as relações entre as covariáveis e o tempo que leva até a ocorrência do evento de interesse. Considerando esse aspecto tempo, este modelo pode ter bastante utilidade na área de crédito, já que dado um conjunto de covariáveis associado a um cliente, a alta probabilidade deste superar o final do prazo para o empréstimo sem apresentar inadimplência, fornece maior segurança durante a negociação.

De acordo com o Capítulo 3, um dos principais objetivos ao se modelar a função de risco é determinar e conhecer quais potenciais variáveis explanatórias a influenciam. Outro importante objetivo em modelar a função de risco é a obtenção de uma medida de risco individual para cada cliente. Além do interesse específico na função de risco, através de sua relação com a função de sobrevivência descrita na Seção 3.1.

Será feita na amostra, inicialmente, uma análise exploratória dos dados com a finalidade de: detectar inconsistências geradas pelo sistema da instituição; comparar os comportamentos das variáveis explanatórias, no contexto de Credit Scoring, entre a amostra de bons e maus pagadores, indicando assim potenciais variáveis relacionadas com o evento modelado e, também, para definir possíveis transformações de variáveis e a criação de novas a serem utilizadas nos modelos.

O tratamento dado às covariáveis categorizadas foram de critérios de agrupamentos baseados na natureza da variável, do produto de crédito e à aplicação da mesma na instituição financeira de origem.

Em seguida, cada covariável foi estudada em relação à variável resposta, a inadimplência.

Foram comparados os percentuais de adimplentes e inadimplentes de cada categoria em relação ao total da amostra.

Adotou-se uma regra para determinação da casela de referência para criação das variáveis dummies. A casela de referência escolhida foi a primeira categoria de cada uma das covariáveis categorizadas. A Tabela 5.2 aponta, por exemplo, como foram criadas as variáveis dummies para a covariável valor contratado dos clientes.

Tabela 5.2 - Variáveis dummies associadas à covariável valor contratado dos clientes.

Variável	$valor1$	$valor2$	$valor3$
$valor\ contratado \leq 500,00$	0	0	0
$500,00 < valor\ contratado \leq 1.000,00$	1	0	0
$1.000,00 < valor\ contratado \leq 4.000,00$	1	1	0
$valor\ contratado > 4.000,00$	1	1	1

Além disso, serão realizados métodos gráficos descritivos para conhecer a classe de modelos de sobrevivência mais indicada para a variável tempo de inadimplência, assim como, a suposição de riscos proporcionais. Se a variável tempo não se ajustar à classe dos modelos paramétricos, será usado o modelo semiparamétrico, especificamente, o modelo de regressão de Cox.

5.3 - Análise Descritiva

As covariáveis foram categorizadas e avaliadas individualmente, através de uma análise exploratória.

5.3.1 - Covariável Sexo

A Tabela 5.3 apresenta a distribuição dos clientes na amostra, segundo o sexo, e observa-se que a proporção de homens (51,83%) é próxima da proporção de mulheres (48,17%).

Tabela 5.3 - Distribuição dos Clientes na Amostra, Segundo o Sexo.

Sexo	Freqüência	%
Feminino	867	48,17
Masculino	933	51,83
Total	1.800	100

A Tabela 5.4 e a Figura 5.1 apresentam a distribuição dos clientes na amostra segundo o sexo e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.4 - Distribuição dos Clientes na Amostra, segundo o Sexo e a Classificação.

Sexo	Mau (%)	Bom (%)	Total
Feminino	458 (50,33)	409 (45,96)	867
Masculino	452 (49,67)	481 (54,04)	933
Total	910	890	1800

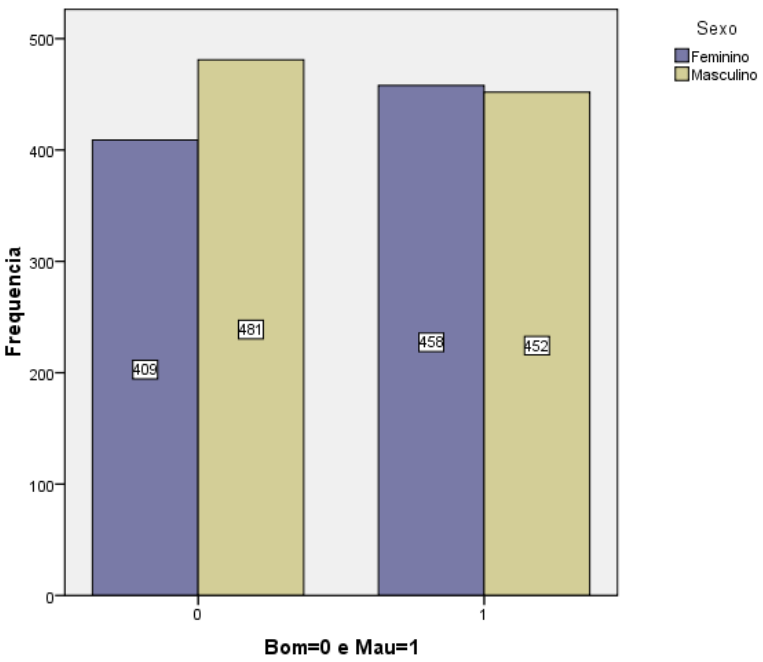

Figura 5.1 - Distribuição dos Clientes na Amostra, segundo o Sexo e a Classificação.

5.3.2 - Covariável Escolaridade

A Tabela 5.5 apresenta a distribuição da escolaridade dos clientes na amostra, e observa-se que a maior proporção é de clientes com segundo grau ou superior completo (60,7%).

Tabela 5.5 - Distribuição da Escolaridade dos Clientes na Amostra.

Escolaridade	Freqüência	%
Analfabeto ou 1º grau	707	39,3
2º grau ou Superior Completo	1.093	60,7
Total	1.800	100

A Tabela 5.6 e a Figura 5.2 apresentam a distribuição da escolaridade dos clientes na amostra e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.6 - Distribuição da Escolaridade dos Clientes na Amostra Segundo a Classificação.

Escolaridade	Mau (%)	Bom (%)	Total
Analfabeto ou 1º grau	381 (41,87)	326 (36,63)	707
2º grau ou Superior Completo	529 (58,13)	564 (63,37)	1.093
Total	910	890	1.800

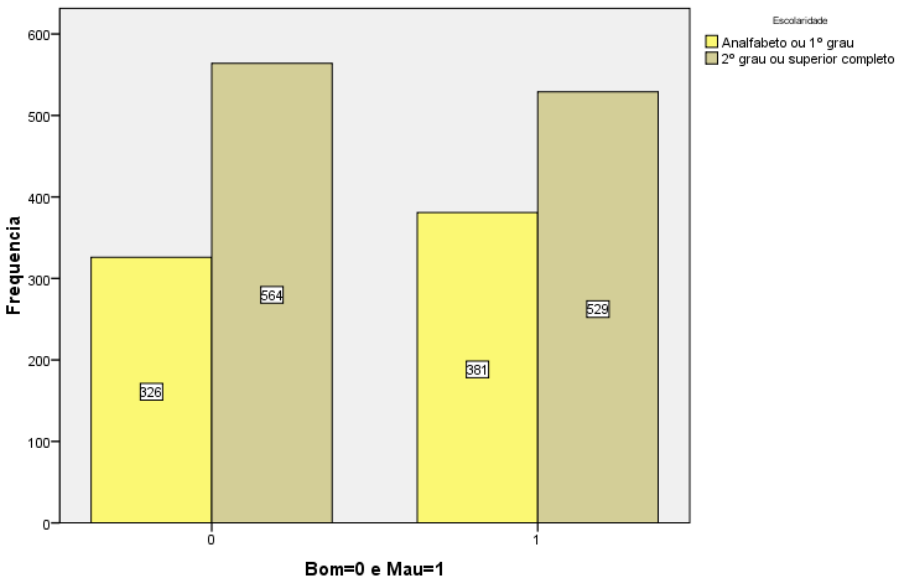

Figura 5.2 - Distribuição da Escolaridade dos Clientes na Amostra Segundo a Classificação.

5.3.3 - Covariável Estado Civil

A Tabela 5.7 apresenta a distribuição do estado civil dos clientes na amostra, e observa-se que a maior proporção é de solteiro(a)s (71,56%).

Tabela 5.7 - Distribuição do Estado Civil dos Clientes na Amostra.

Estado Civil	Freqüência	%
Solteiro	1.288	71,56
Casado	416	23,11
Divorciado	38	2,11
Viúvo	58	3,22
Total	1.800	100

Tabela 5.8 - Distribuição do Estado Civil dos Clientes na Amostra Segundo a Classificação.

Estado Civil	Mau (%)	Bom (%)	Total
Solteiro	630 (69,23)	658 (73,93)	1288
Casado	233 (25,6)	183 (20,56)	416
Divorciado	22 (2,42)	16 (1,8)	38
Viúvo	25 (2,75)	33 (3,71)	58
Total	910	890	1800

A Tabela 5.8 e a Figura 5.3 apresentam a distribuição do estado civil dos clientes na amostra e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

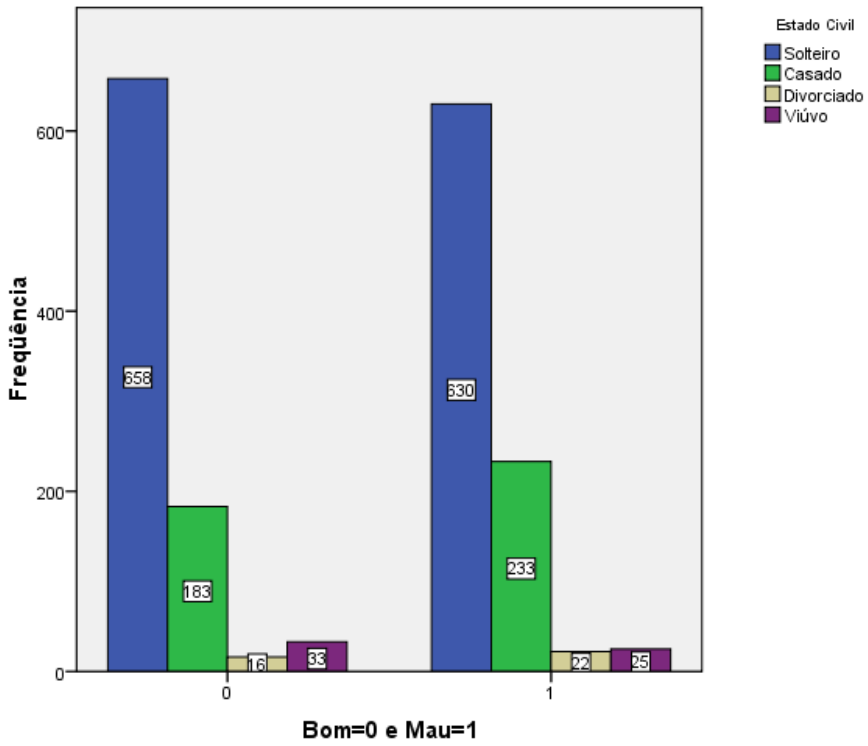

Figura 5.3 - Distribuição do Estado Civil dos Clientes na Amostra Segundo a Classificação.

5.3.4 - Covariável Idade

Na Tabela 5.9 e na Figura 5.4 apresentam a distribuição das idades dos clientes, cuja idade média é de 37 anos. Observa-se que 50% dos clientes têm idades abaixo de 34 anos. A menor idade observada na amostra é de 18 anos e a maior idade é 86 anos. Nota-se que há uma predominância (75%) de idades inferiores a 44 anos.

Tabela 5.9 - Distribuição das Idades dos Clientes na Amostra.

Idade (em anos)	
Média	37
Mediana	34
Mínimo	18
Máximo	86
1° Quartil	27
3° Quartil	44

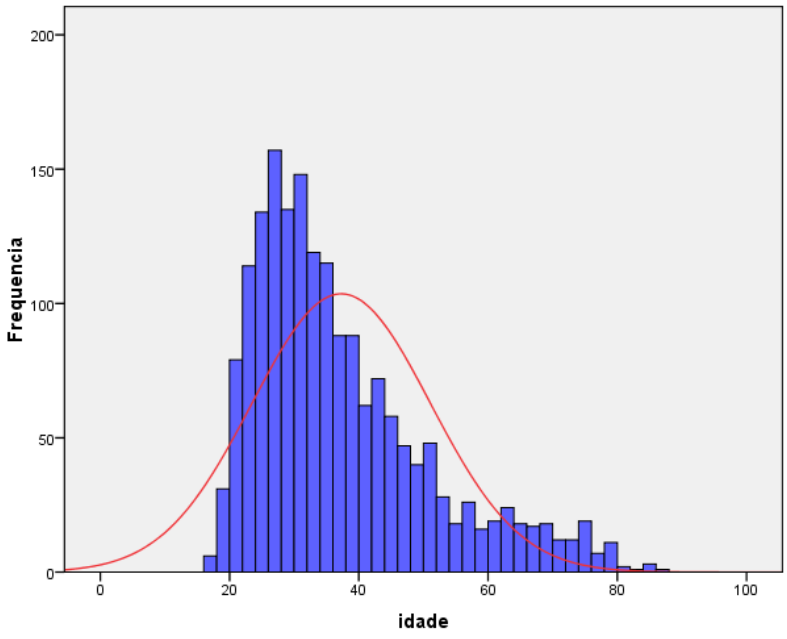

Figura 5.4 - Distribuição das Idades dos Clientes na Amostra.

A Tabela 5.10 e a Figura 5.5 apresentam a distribuição dos clientes na amostra segundo a idade e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.10 - Distribuição dos Clientes na Amostra Segundo as Idades e a Classificação.

Idade (em anos)	Mau (%)	Bom (%)	Total
$idade \leq 26$	145 (15,93)	219 (24,61)	364
$26 < idade \leq 34$	278 (30,55)	281 (31,57)	559
$34 < idade \leq 59$	410 (45,05)	287 (32,25)	697
$idade > 59$	77 (8,46)	103 (11,57)	180
Total	910	890	1.800

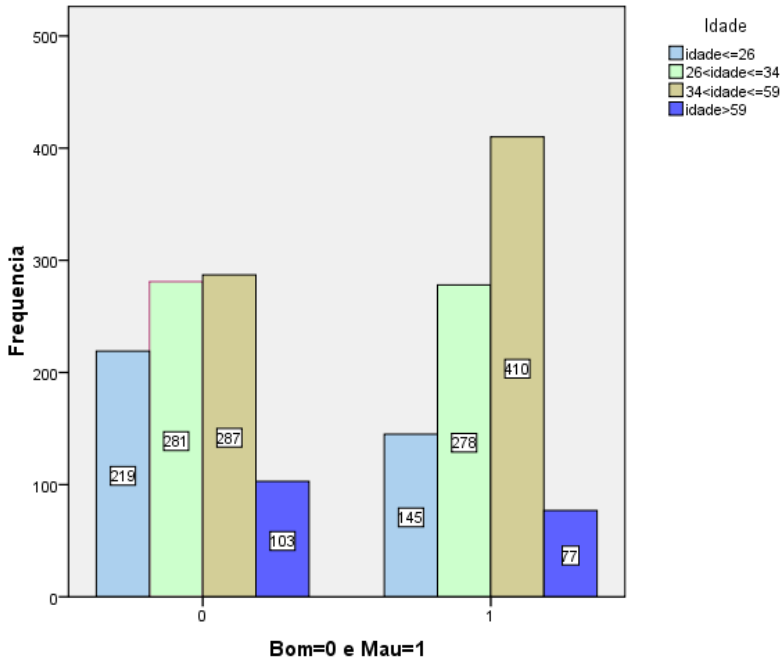

Figura 5.5 - Distribuição dos Clientes na Amostra Segundo as Idades e a Classificação.

5.3.5 - Covariável Número de Dependentes

A Tabela 5.11 apresenta a distribuição do número de dependentes dos clientes na amostra, e observa-se que a maior proporção é de clientes sem dependentes (85,94%).

A Tabela 5.12 e a Figura 5.6 apresentam a distribuição do número de dependentes dos clientes na amostra e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.11 - Distribuição do Número de Dependentes dos Clientes na Amostra.

Número de Dependentes	Frequência	%
Sem Dependente	1.547	85,94
Com Dependente(s)	253	14,06
Total	1.800	100

Tabela 5.12 - Distribuição do Número de Dependentes dos Clientes na Amostra Segundo a Classificação.

Número de Dependentes	Mau (%)	Bom (%)	Total
Sem Dependente	781 (85,82%)	766 (86,07%)	1.547
Com Dependente(s)	129 (14,18%)	124 (13,93%)	253
Total	910	890	1.800

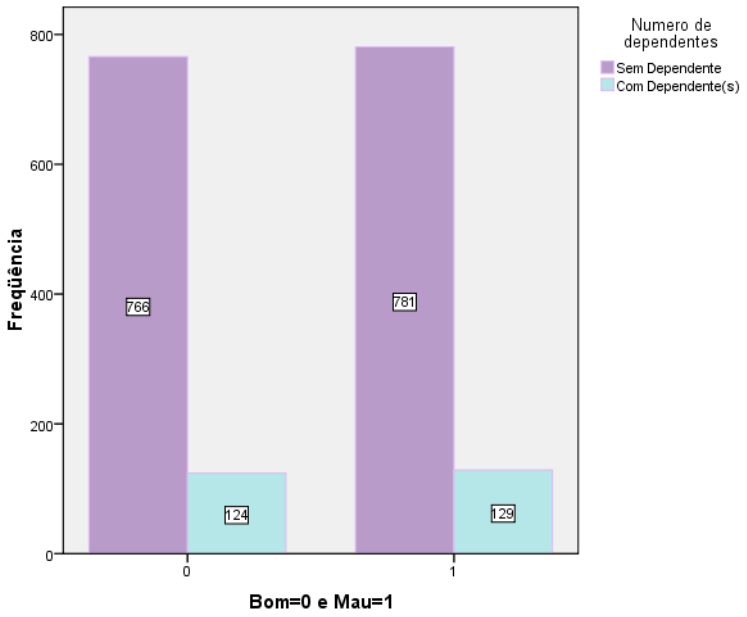

Figura 5.6 - Distribuição do Número de Dependentes dos Clientes na Amostra Segundo a Classificação.

5.3.6 - Covariável Renda

A Tabela 5.13 e a Figura 5.7 apresentam a distribuição dos clientes na amostra segundo a renda, e observa-se que a renda média é de R$ 760,00. 50% dos clientes têm renda abaixo de R$ 478,00 e os demais acima deste valor. A renda mais freqüente na amostra é de R$

350,00. O menor valor de renda observada na amostra é de R$ 203,00 e a maior renda é de R$ 6.868,00. Nota-se que há predominância (75%) de rendas abaixo de R$ 900,00.

Tabela 5.13 - Distribuição dos Clientes na Amostra Segundo a Renda.

	Renda (R$)
Média	760,00
Mediana	478,00
Moda	350,00
Mínimo	203,00
Máximo	6.868,00
1° Quartil	350,00
3° Quartil	900,00

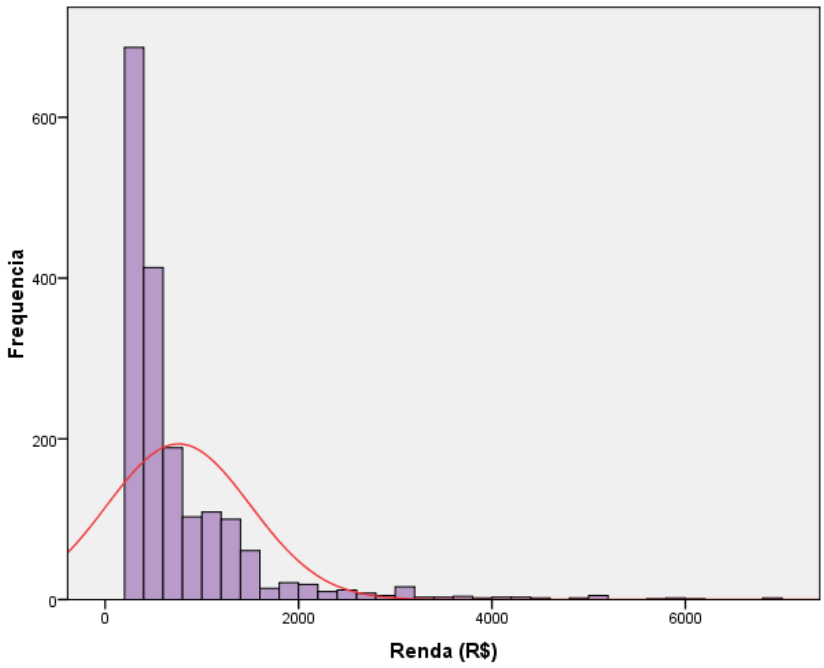

Figura 5.7 - Distribuição dos Clientes na Amostra Segundo a Renda.

A Tabela 5.14 e a Figura 5.8 apresentam a distribuição da renda dos clientes na amostra e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.14 - Distribuição da Amostra da Renda dos Clientes Segundo a Classificação.

Renda (R$)	Mau (%)	Bom (%)	Total
$renda \leq 478,00$	456 (50,11)	444 (49,89)	900
$478,00 < renda \leq 1.050,00$	244 (26,81)	298 (33,48)	542
$renda > 1.050,00$	210 (23,08)	148 (16,63)	358
Total	910	890	1.800

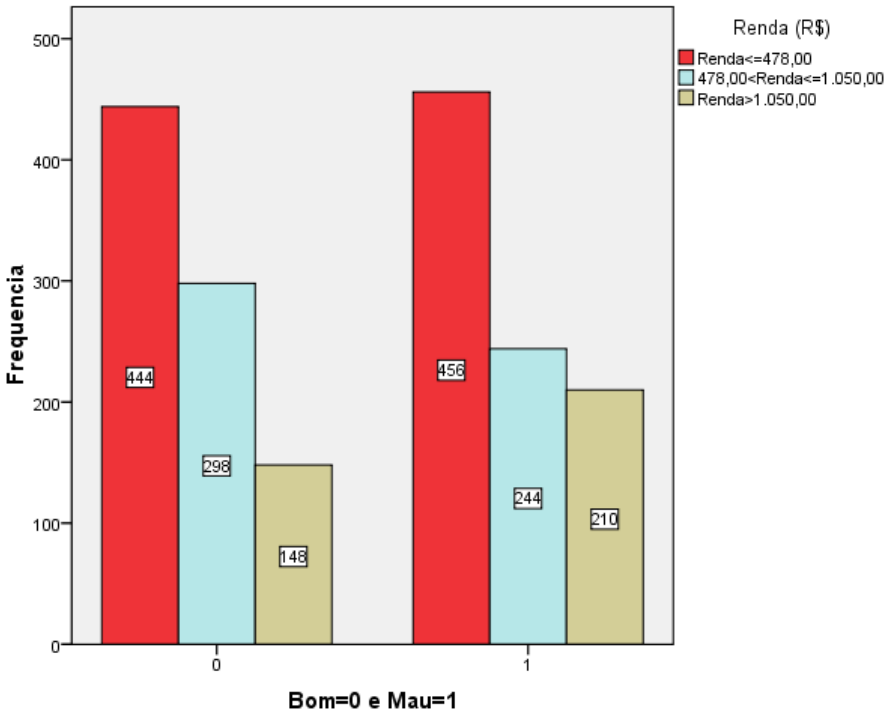

Figura 5.8 - Distribuição da Amostra da Renda dos Clientes Segundo a Classificação.

5.3.7 - Covariável Número de Parcelas

Na Tabela 5.15 e na Figura 5.9 apresentam a distribuição dos clientes na amostra segundo o número de parcelas, cujo número médio de parcelas é de 29. Observa-se que 50% dos clientes optaram por um número de parcelas abaixo de 32. O número de parcelas mais freqüente na amostra é de 36.

A Tabela 5.16 e a Figura 5.10 apresentam a distribuição dos clientes na amostra segundo o numero de parcelas e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.15 - Distribuição dos Clientes na Amostra Segundo o Número de Parcelas

	Número de Parcelas
Média	29
Mediana	32
Moda	36
1° Quartil	24
3° Quartil	36

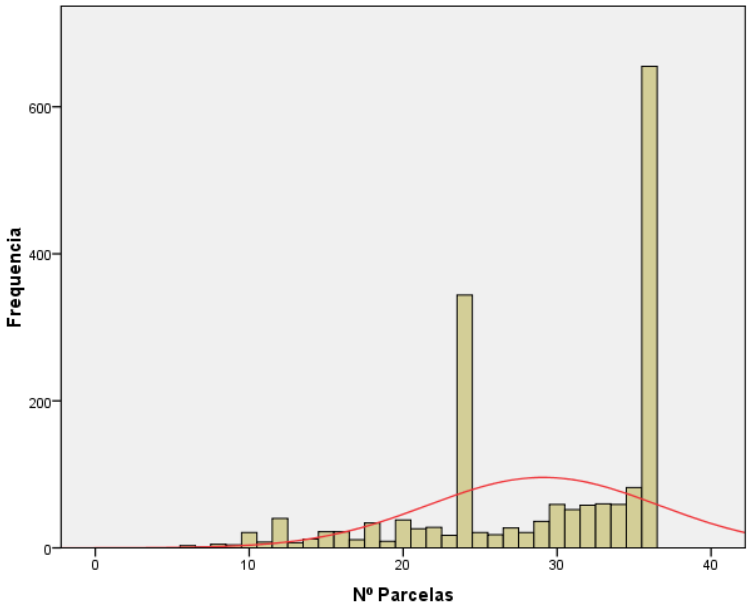

Figura 5.9 - Distribuição dos Clientes na Amostra Segundo o Número de Parcelas

Tabela 5.16 - Distribuição dos Clientes na Amostra Segundo o Número de Parcelas e a Classificação.

Número de Parcelas	Mau (%)	Bom (%)	Total
$n°parcelas \leq 18$	144 (15,82)	46 (5,17)	190
$18 < n°parcelas \leq 34$	392 (43,08)	481 (54,04)	873
$n°parcelas > 34$	374 (41,10)	363 (40,79)	737
Total	910	890	1.800

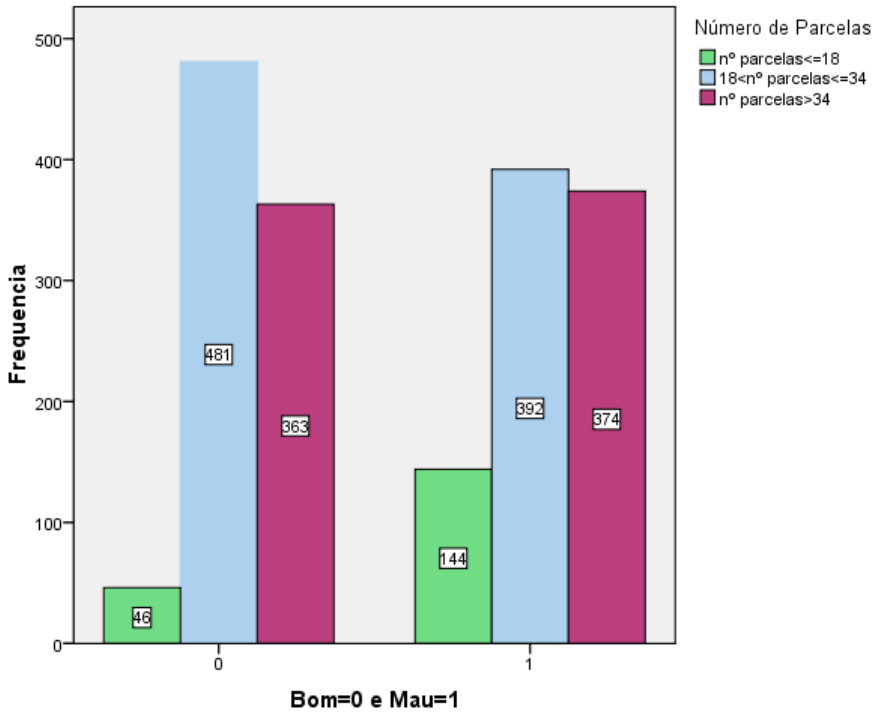

Figura 5.10 - Distribuição dos Clientes na Amostra Segundo o Número de Parcelas e a Classificação.

5.3.8 - Covariável Valor Contratado

Na Tabela 5.17 e na Figura 5.11 apresentam a distribuição dos clientes na amostra segundo os valores contratados. Observa-se que a média dos valores contratados é de R$ 1.239,10. 50% dos clientes emprestaram valores abaixo de R$ 728,83. O menor valor contratado na amostra é de R$ 256,69 e o maior é R$ 29.186,86. Há uma predominância (75%) de valores contratados inferiores a R$ 1.297,50.

Tabela 5.17 - Distribuição dos Clientes na Amostra Segundo os Valores Contratados.

	Valor Contratado (R$)
Média	1.239,10
Mediana	728,83
Mínimo	256,69
Máximo	29.186,86
1° Quartil	504,68
3° Quartil	1.297,50

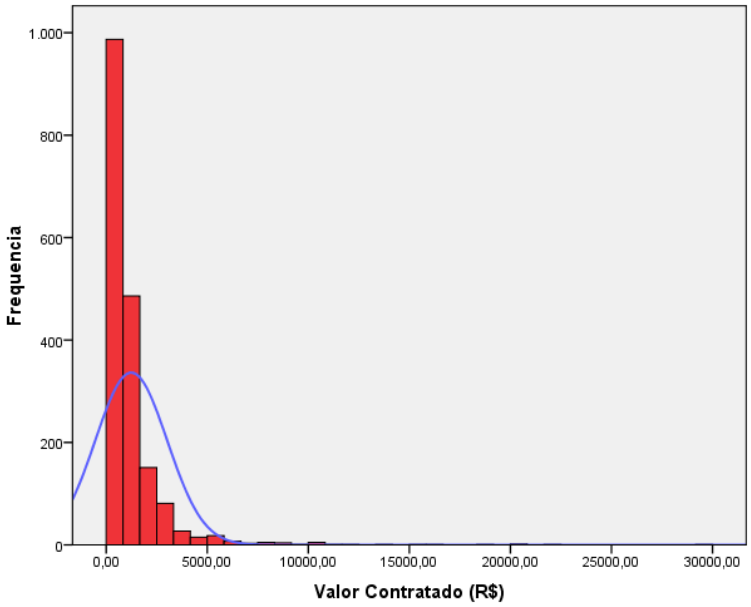

Figura 5.11 - Distribuição dos Clientes na Amostra Segundo os Valores Contratados.

A Tabela 5.18 e a Figura 5.12 apresentam a distribuição dos clientes na amostra segundo o valor contratado e a classificação, adimplente (Bom=0) e inadimplente (Mau=1).

Tabela 5.18 - Distribuição dos Clientes na Amostra Segundo os Valores Contratados e a Classificação.

Valor Contratado	Mau (%)	Bom (%)	Total (%)
$valor\_cont \leq 500,00$	234 (25,71)	204 (22,92)	438
$500,00 < valor\_cont \leq 1.000,00$	335 (36,81)	352 (39,55)	687
$1.000,00 < valor\_cont \leq 4.000,00$	299 (32,86)	304 (34,16)	603
$valor\_cont > 4.000,00$	42 (4,62)	30 (3,37)	72
Total	910	890	1.800

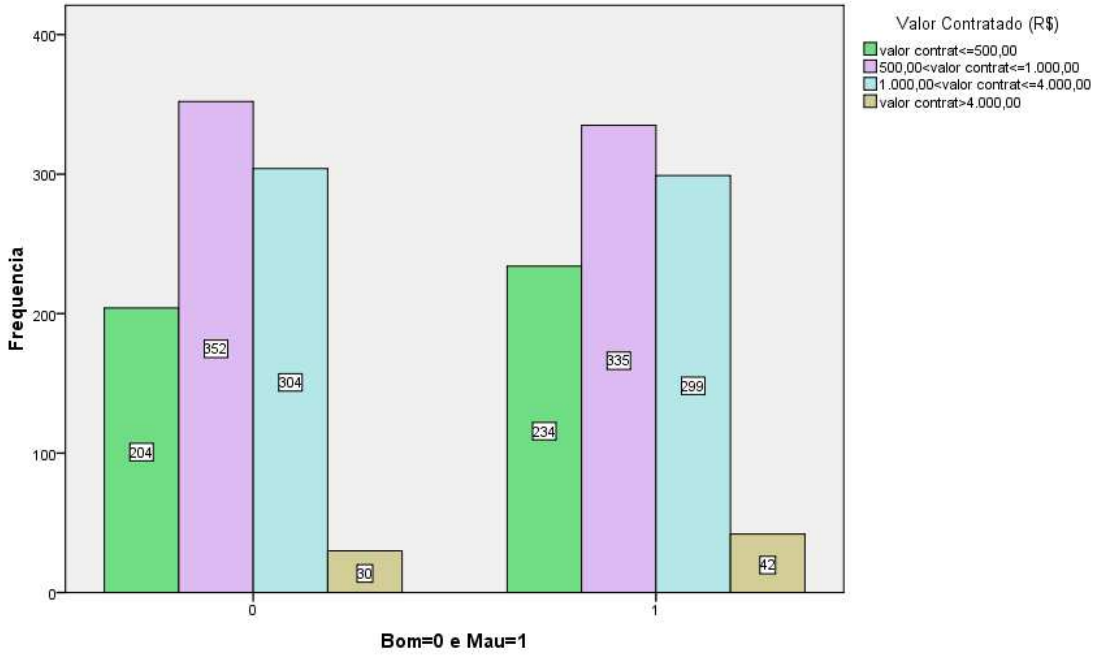

Figura 5.12 - Distribuição dos Clientes na Amostra Segundo os Valores Contratados e a Classificação.

5.4 - Ajuste da Metodologia de Sobrevivência

5.4.1 Procedimentos de Estimação Clássica

Todos os indivíduos que compunham a base de dados foram acompanhados por doze meses, a partir da data da contratação do empréstimo. A informação inadimplência consta na base de dados representada pela data de atraso de uma das parcelas. O tempo é registrado em meses a partir da data da contratação do empréstimo. Assim, o indivíduo cujo crédito é concedido em agosto de 2006 e registra uma data de atraso em dezembro de 2006, significa que ele tem tempo de sobrevivência de 4 meses. No contexto de Credit Scoring, os indivíduos que não apresentaram problemas com pagamento são considerados censurados. Nesse caso, já que os indivíduos era acompanhados por 12 meses, o registro do tempo de sobrevivência para todos os clientes censurados era de 12 meses. Para avaliar qual a classe de modelos de sobrevivência que será utilizada para o sistema de Credit Scoring, faz-se necessário realizar inspeções gráficas da variável tempo (T). Primeiramente, utiliza-se um método gráfico que consiste na comparação da função de sobrevivência do modelo paramétrico proposto com o estimador de Kaplan-Meier (Collet, 1994).

Foram utilizados os modelos paramétricos exponencial, Weibull e log-normal. Em seguida, avalia-se o pressuposto de riscos proporcionais para as covariáveis. Isto significa estimar os efeitos das covariáveis segundo a proporcionalidade dos riscos ao longo de todo o tempo de observação do estudo. Para tal, os resultados foram obtidos no software R (http://www.r-project.org).

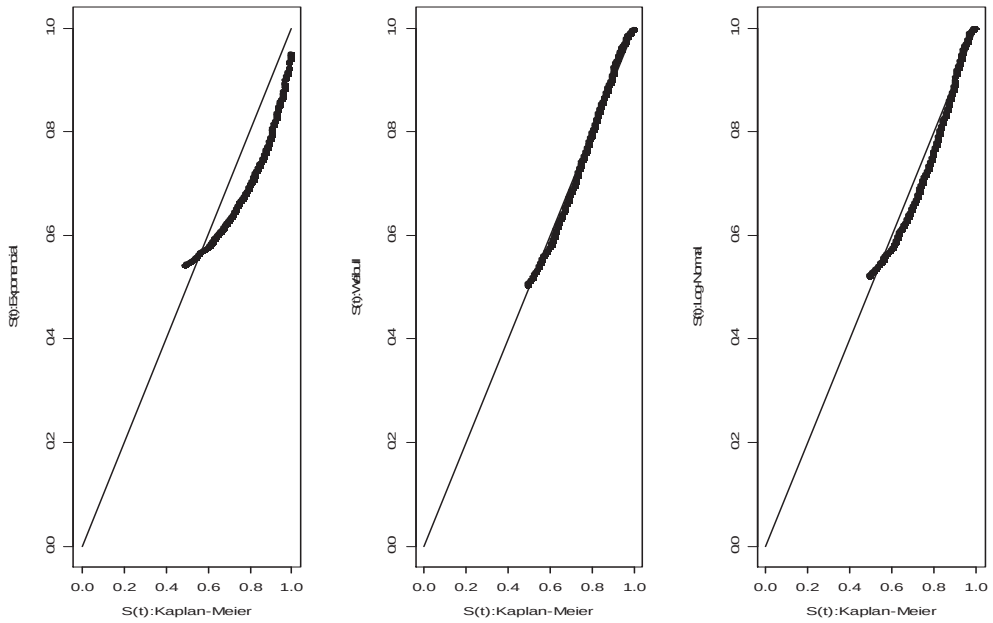

Figura 5.13 - Gráficos das estimativas das sobrevivências obtidas pelo método Kaplan-Meier versus as estimativas das sobrevivências do modelo exponencial, Weibull e log-normal.

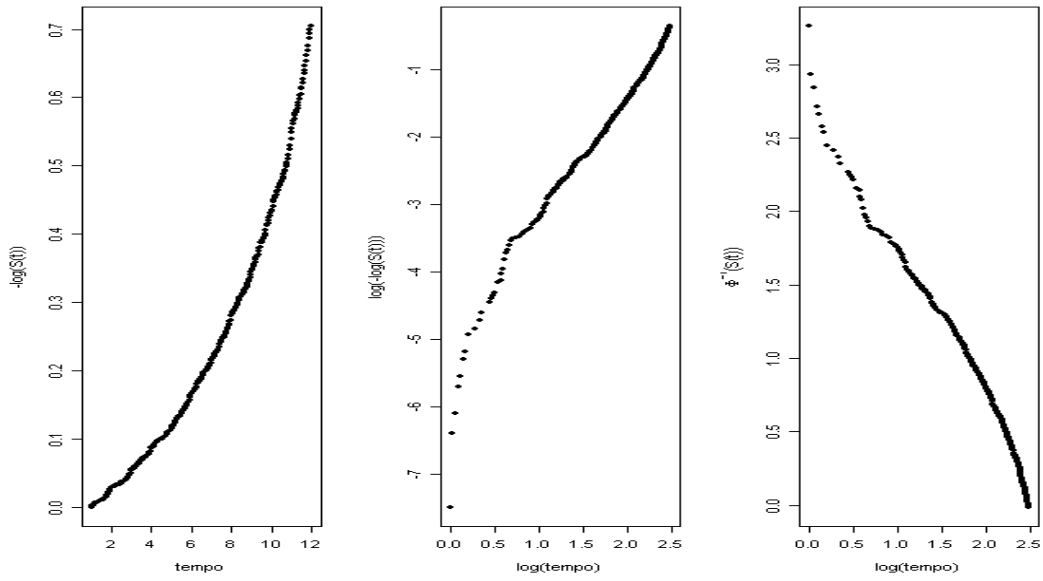

Figura 5.14 - Gráficos linearizados para os modelos exponencial, Weibull e log-normal.

A partir da inspeção das Figuras 5.13 e 5.14, constata-se que o modelo exponencial indica não ser adequado para os dados. Já para os modelos Weibull e log-normal os gráficos não mostram afastamentos marcantes de uma reta, dando indícios de que os dois são indicados para o ajuste dos dados.

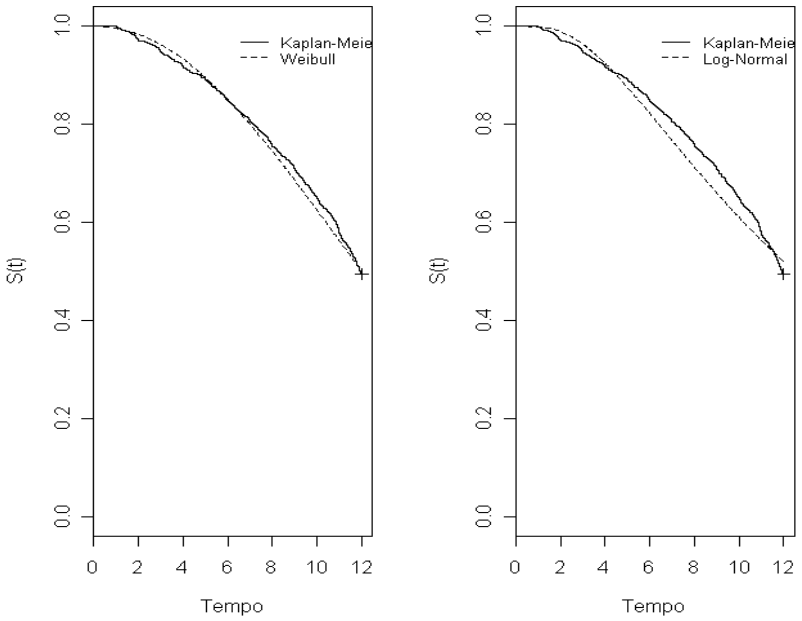

Figura 5.15 - Curvas de sobrevivência estimadas para os modelos de Weibull e log-normal versus a curva de sobrevivência estimada por Kaplan-Meier.

Nos gráficos da figura 5.15, conclui-se que a melhor escolha é pela utilização do modelo paramétrico de Weibull. Portanto, exclui-se a possibilidade do uso da classe de modelos semiparamétricos tendo em vista estimativas mais precisas com a utilização do modelo paramétrico de Weibull.

Abaixo, gráficos mostram duas das covariáveis da base de dados que violam o pressuposto de riscos proporcionais para o modelo. Para as covariáveis categorizadas em pequeno número, o gráfico da curva de Kaplan-Meier estratificado indica essa proporcionalidade.

Curvas que se apresentam paralelas ao longo de todo tempo de observação indicam proporcionalidade no risco entre as categorias das covariáveis. Pode-se observar nas Figuras 5.16 e 5.17 a ocorrência de sobreposição e cruzamentos entre as categorias das covariáveis valor contratado e renda mensal.

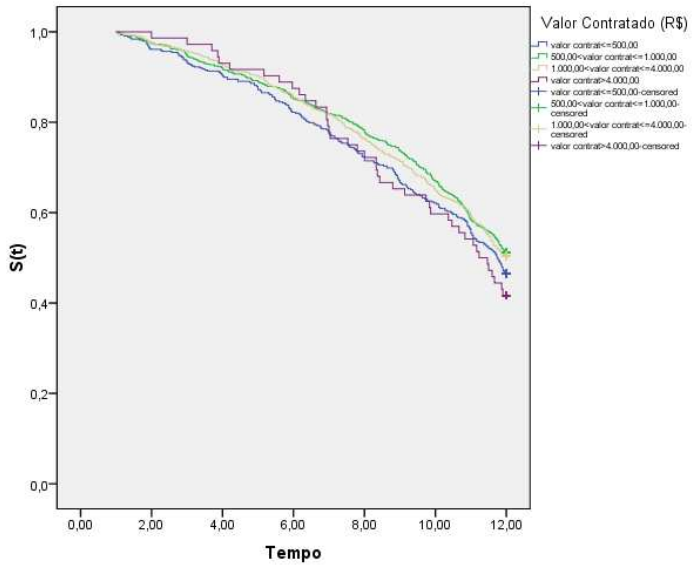

Figura 5.16 - Curva de sobrevivência de Kaplan-Meier para a covariável valor contratado.

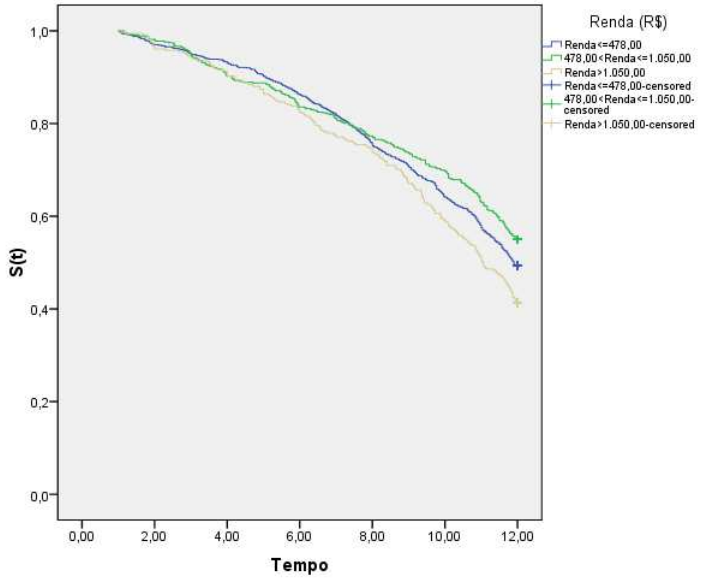

Figura 5.17 - Curva de sobrevivência de Kaplan-Meier para a covariável renda mensal.

Na adoção do modelo paramétrico de Weibull, deve-se considerar que o mesmo admite função de risco contínua e, sob essa suposição empates nos tempos de sobrevivência não são possíveis. No entanto, as informações do comportamento de crédito da carteira de clientes foram registradas na base de dados, de acordo com o mês da ocorrência do evento inadimplência e, portanto, empates ocorreram, sendo observada a existência de mais de um evento em um mesmo instante de tempo, ou seja, ocorrências de inadimplências no

mesmo mês. Nesse caso a função de máxima verossimilhança parcial deve ser modificada com a finalidade de incorporar essas observações empatadas.

As covariáveis significativas para o modelo de sobrevivência foram selecionadas através do método clássico *stepwise* (Hosmer & Lemeshow, 2000) e, são consideradas variáveis de alto grau de importância pelos analistas de crédito, uma vez que estão presentes, também, na literatura de Credit Scoring. Segundo Anderson (2007), a identificação das variáveis potenciais deve ser feita por analistas de crédito e administradores do produto. É parte da "arte"necessária para o desenvolvimento do sistema de Credit Scoring. No início do trabalho há uma tendência natural em escolher um número muito grande de variáveis potenciais. No entanto, a lista de variáveis deve sofrer um primeiro crivo, baseado na experiência dos membros da equipe e dos consultores externos, considerando-se, por exemplo, a disponibilidade desses dados nos bancos de dados. Se por um lado corremos o risco de eliminar uma variável que poderia ser útil, por outro se reduz a dimensionalidade do problema e viabiliza a aquisição dos dados. Nem todas as variáveis potenciais serão consideradas no cálculo do escore. Através de técnicas estatísticas, serão selecionadas as variáveis que, em conjunto, melhor permitem classificar a operação de crédito.

A seleção de variáveis pelo método bayesiano, em dados de sobrevivência, ainda é um desafio devido às dificuldades em especificar distribuições a priori para os parâmetros de regressão de todos os modelos possíveis no espaço de modelos, bem como sua implementação computacional (Ibrahim, 2001).

A fim de comparar as abordagens clássica e bayesiana, em algumas inferências obtidas dos dados de Credit Scoring, via análise de sobrevivência, ajustou-se o modelo clássico, apresentado na Tabela 5.19 e realizou-se a avaliação de desempenho do modelo utilizando a curva ROC mostrada na Figura 5.18. Os valores utilizados como escore final foram obtidos através de sua parte linear (estimada), ou seja, pelo valor $\boldsymbol{x}'\hat{\boldsymbol{\beta}}$.

Na Tabela 5.19 estão os coeficientes estimados para o modelo Weibull. A interpretação dos coeficientes do modelo Weibull é a seguinte: um cliente que fez a opção de pagamento com número de parcelas entre 18 a 34, tem cerca de 23% (exp(-0,2609)=0,77) de chance de não apresentar inadimplência nos próximos 12 meses.

Tabela 5.19 - Ajuste do Modelo Weibull

Covariável	Descrição	Parâmetro	Erro Padrão	P-valor
intercepto	-	-5,219	0,0669	0,0000
sexo	sexo	-0,06116	0,0324	0,0038
id2	$26 < idade \leq 34$	0,1075	0,0582	0,0019
id3	$34 < idade \leq 59$	0,0527	0,0368	0,0000
nparc2	$18 < n°parcelas \leq 34$	-0,2609	0,0615	0,0000
nparc3	$n°parcelas > 34$	0,1203	0,0587	0,0000
escala	-	0,466	0,0176	0,0001
forma	-	2,15	0,0235	0,0000

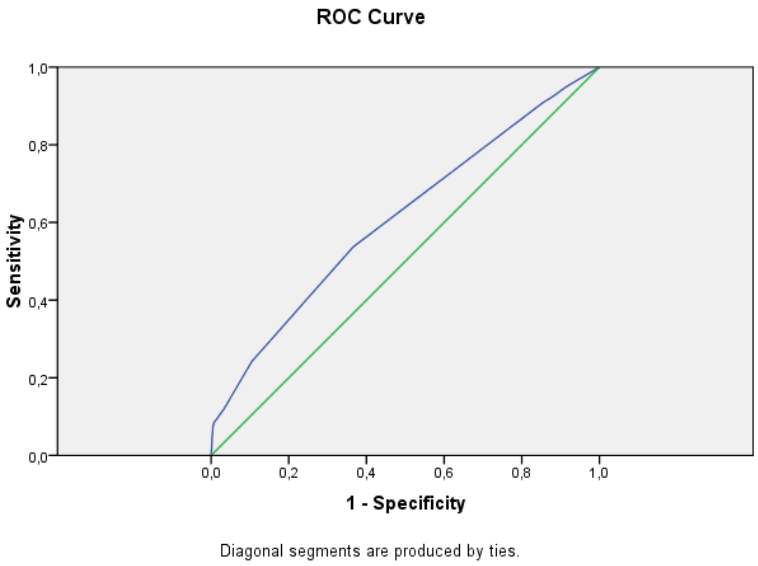

Figura 5.18 - Gráfico da área sob a curva ROC para o modelo clássico Weibull

Quanto mais distante a curva ROC estiver da reta x = y, melhor será a classificação que o modelo fará entre adimplentes e inadimplentes.

A medida que expressa essa discriminação é conhecida como área sob a curva ROC e esta medida varia entre 0,5 e 1. Quanto maior o valor melhor será o poder de classificação do modelo.

Para os dados analisados, a área sob a curva ROC obtida foi de 0,610, um resultado bem razoável considerando-se o tamanho da amostra de validação, pois segundo Anderson (2007) os sistemas de Credit Scoring são muito sensíveis ao tamanho da amostra, tal fato

se confirma nos trabalho de Abreu (2004) e Tomazela (2007). Banasik et al. (1999) sugere um tamanho de amostra de validação não menor que 1.500 bons e 1.500 maus.

5.4.2 Procedimentos de Estimação Bayesiana

Para os procedimentos bayesianos tradicionais, consideramos o modelo de Weibull das Seções 3.2 e 4.3 e assumimos $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ para $i = 1, ..., 1800$; com densidades a priori normais não-informativas e independentes para os coeficientes de regressão $\boldsymbol{\beta}$, onde a média utilizada é zero e a precisão é 0,0001, $N_5(\boldsymbol{\mu}_0 = 0, \boldsymbol{\Sigma}_0 = 0,001)$. E para o parâmetro de forma α da função de sobrevivência foi dado uma priori Gama(1; 0,0001), que é fracamente decrescente.

Obtemos para os parâmetros do modelo as inferências bayesianas tradicionais usando o algoritmo de Gibbs. Os resultados obtidos estão apresentados na Tabela 5.20, onde temos os resumos a *posteriori* de interesse considerando o modelo de Weibull com dados empatados.

Tabela 5.20 - Resumo a posteriori dos parâmetros do modelo Weibull

Descrição	Parâmetro	Média a Posteriori	Mediana	D.P	I.C. de 95%
forma	-	2,1033	2,132	0,06566	(2,007 ; 2,263)
intercepto	β_0	-5,689	-5,687	0,01614	(-6,008; -5,381)
sexo	β_1	-0,06648	-0,06654	0,03314	(-0,1314 ;-0,00102)
$26 < idade \leq 34$ anos	β_2	0,1294	0,1294	0,04129	(0,04863 ; 0,2113)
$34 < idade \leq 59$ anos	β_3	0,05514	0,05523	0,03768	(0,04014 ; 0,1294)
$18 < n^{\circ}parcelas \leq 34$	β_4	-0,2456	-0,2463	0,02982	(-0,3028 ; -0,1859)
$n^{\circ}parcelas > 34$	β_5	0,1125	0,1125	0,03524	(0,04363 ; 0,1824)
escala	-	0,4688	0.4688	0.02171	(0,0112 ; 0,6322)

Ainda na Tabela 5.20 são apresentados os intervalos HPD (highest posterior density) com 95% de credibilidade para os parâmetros, cujos limites inferior e superior são mostrados na última coluna. A interpretação desses intervalos é probabilística, ou seja, para os limites apresentados têm-se que a probabilidade do parâmetro assumir valores no intervalo é 0,95.

Geraram-se duas cadeias com 20.000 iterações cada uma, das quais as 1.000 primeiras foram descartadas para que se reduza a influência dos pontos iniciais. Observamos a convergência das cadeias para todos os parâmetros de interesse. Os traços das cadeias e a

estimação da densidade para cada parâmetro, apresentados na Figura 5.19, indicam que não há problemas com a convergência do algoritmo.

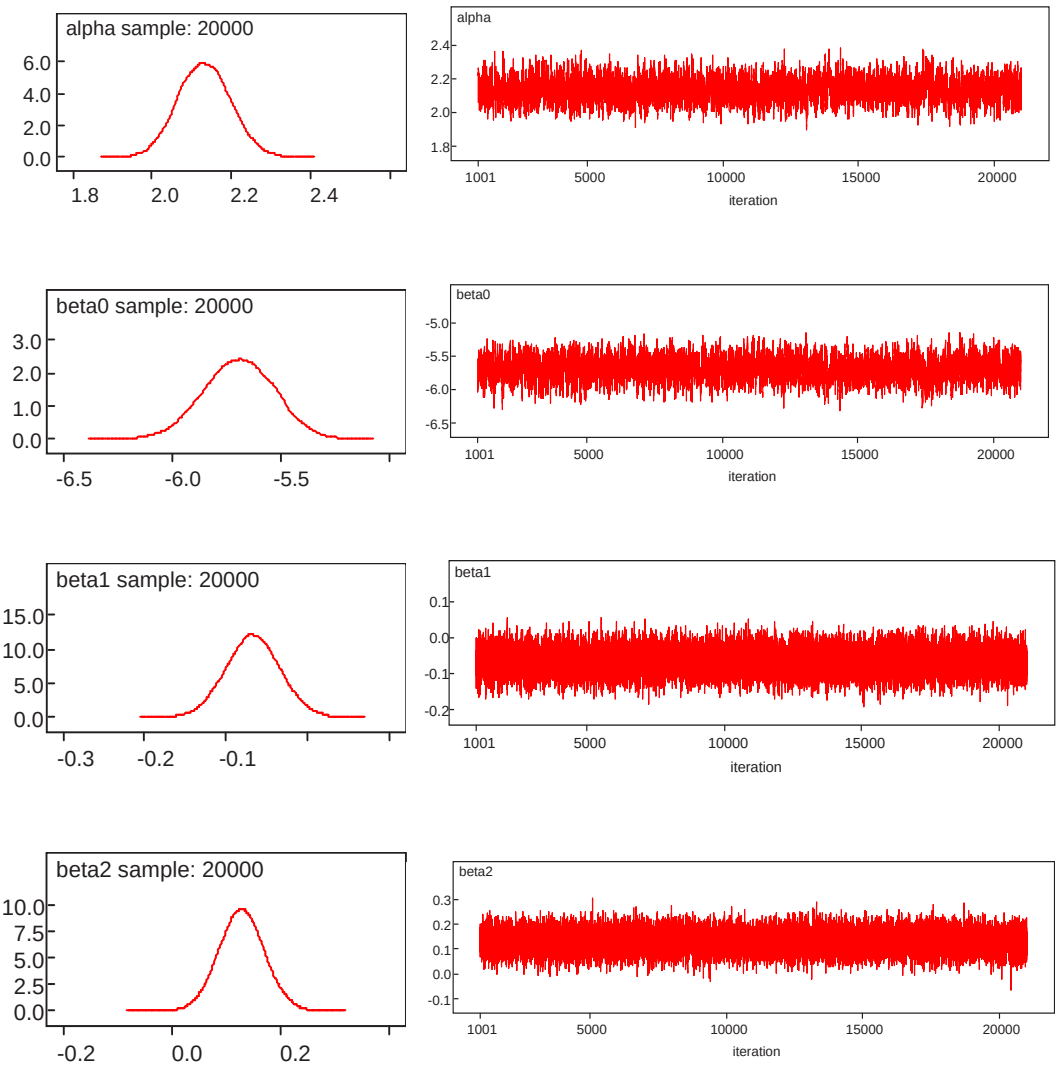

Figura 5.19 - Diagnóstico de convergência dos parâmetros das cadeias geradas e estimação das suas densidades.

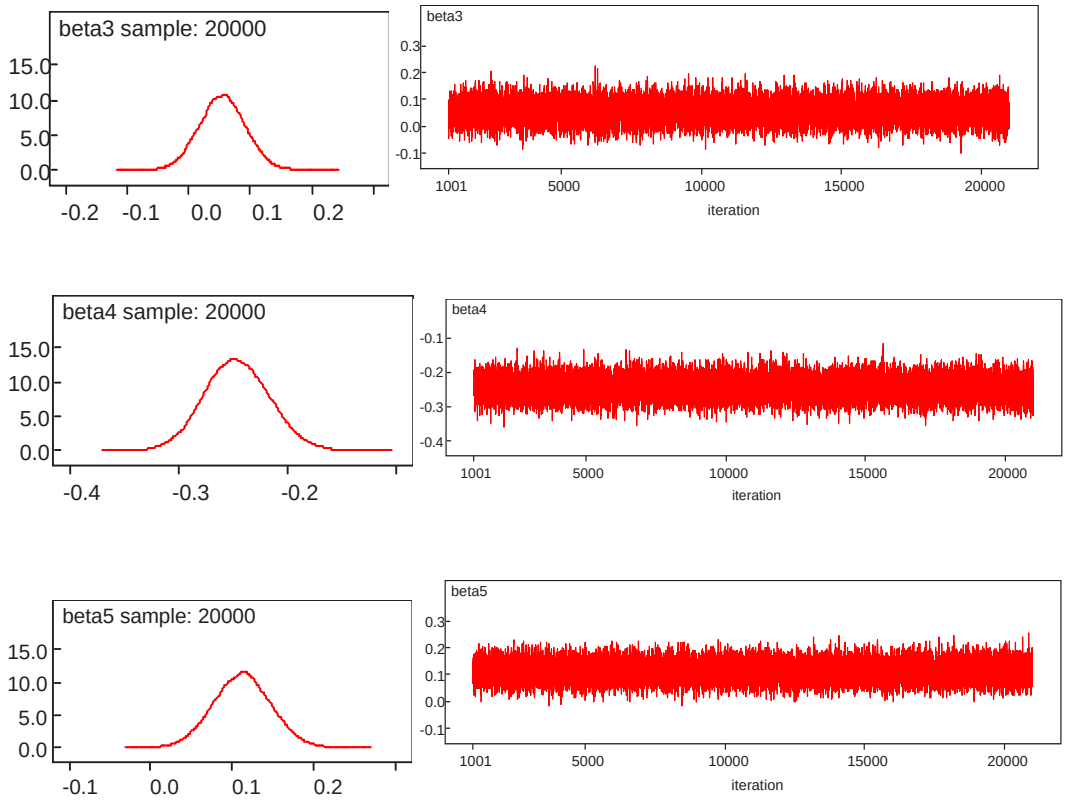

Figura 5.20 - Diagnóstico de convergência dos parâmetros das cadeias geradas e estimação das suas densidades (*continuação*).

Com as estimativas acima, o modelo que gera os escores contínuos para o sistema de Credit Scoring foi avaliado através da medida de desempenho, curva ROC, assim como feito no modelo clássico. Para os dados disponíveis neste trabalho, a área sob a curva ROC obtida foi de 0,659 apresentando uma pequena diferença em relação ao modelo paramétrico clássico.

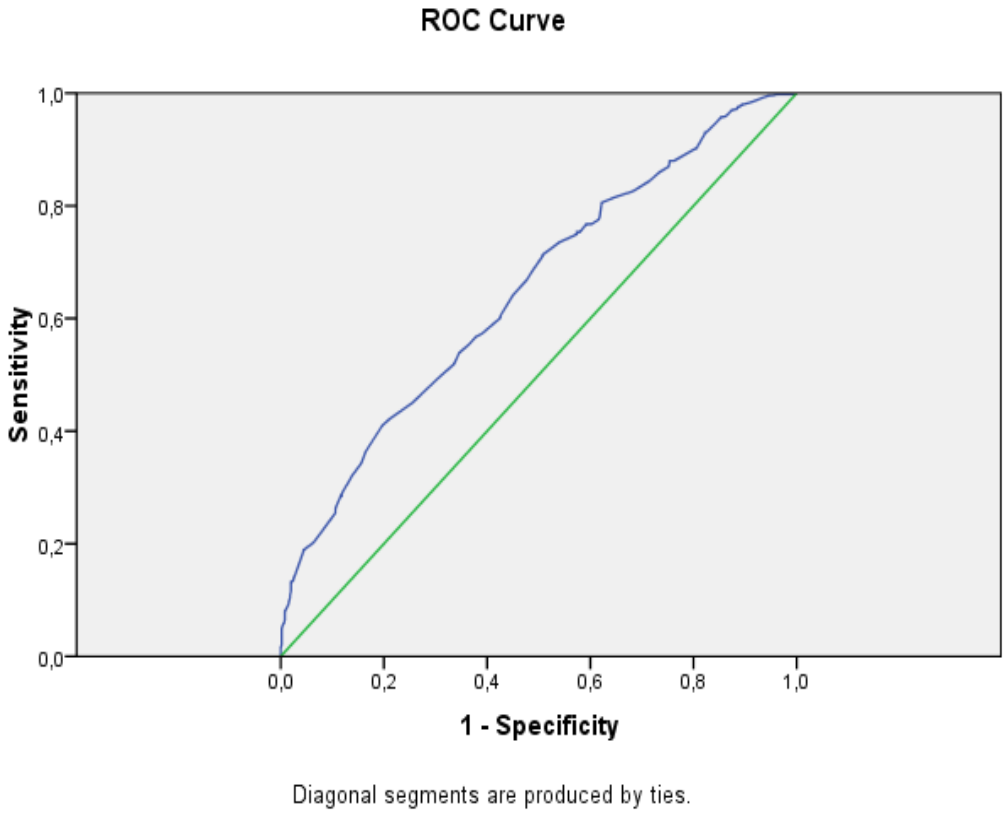

Figura 5.21 - Gráfico da área sob a curva ROC para o modelo bayesiano de Weibull.

Ao considerarmos que as covariáveis utilizadas no modelo bayesiano de sobrevida são as que exercem mais influência sobre as funções de risco e sobrevivência, então de acordo com as estimativas, pode-se interpretar como segue: um cliente que fez a opção de pagamento com número de parcelas entre 18 a 34, tem cerca de 22% (exp(-0,2456)=0,78) de chance de não apresentar inadimplência nos próximos 12 meses. E outra, para um cliente que está na faixa de idade entre 26 a 34 anos, o risco deste se tornar inadimplente é 0,14 (exp(0,1294)=1,14) vezes mais em relação aos clientes na faixa de referência ($\leq$26 anos), nos próximos 12 meses.

Capítulo 6

Conclusões e Propostas para Trabalhos Futuros

6.1 - Conclusão

Este trabalho descreve o desenvolvimento de um modelo de Credit Scoring, não só do ponto de vista da aplicação da metodologia de análise de sobrevivência que é relativamente nova e vantajosa neste contexto, mas, como também apresentou o uso de técnicas bayesianas em mais uma linhas de pesquisa. Buscou-se mostrar algumas das etapas do desenvolvimento de um sistema de *Credit Scoring*, enfocando a importância de se obter uma base de dados consistente, o planejamento amostral, as vantagens do uso da metodologia de análise de sobrevivência e, principalmente da aplicabilidade da metodologia bayesiana.

Foram apresentados os procedimentos gerais referentes à modelagem temporal de *Credit Scoring*, direcionados pela análise de sobrevivência. Observou-se que essa modelagem consiste em uma base sólida para o desenvolvimento de escores contínuos voltados à política de concessão de crédito, os quais propiciam avaliação contínua do risco de crédito em todos os diferentes tempos de relacionamento, ordenando os clientes segundo sua chance de inadimplência a qualquer tempo, dentro e fora do período de desempenho.

Considerando que os dados de sobrevivência são ajustados por uma distribuição Weibull assumindo a presença de censuras à direita, desenvolvemos as inferências bayesianas na presença de covariáveis.

Através dos métodos de amostragem via MCMC utilizados neste estudo, obteve-se boas estimativas para os parâmetros de interesse. Como foi mostrado, as simulações e aplicações da modelagem bayesiana fornecem uma ferramenta útil para análise de dados de sobrevivência empatados univariados. Tais métodos são de fácil implementação mesmo considerando um número de grande de parâmetros. De alguma forma, as estimativas bayesianas

e sua interpretação inerente à dados de sobrevivência podem ser comparadas com o tratamento clássico de dados de tempo de sobrevivência.

Os *softwares WinBugs* e R simplificaram a obtenção dos sumários *a posteriori* considerando a análise bayesiana para o referido modelo ajustado.

Ambas metodologias forneceram resultados dentro do praticado no mercado e na literatura de *Credit Scoring*. No entanto algumas alterações poderiam ser propostas para alcançar possíveis melhorias no desenvolvimento dos modelos, por exemplo, diferentes categorizações das variáveis ou mesmo utilizá-las como contínuas, ou propor algumas iterações entre elas com o objetivo de padronizar e facilitar a modelagem e a comparação dos modelos. Além disso sugerem-se as seguintes propostas para trabalhos futuros:

6.2 - Propostas

Como propostas para trabalhos futuros, pode-se destacar:

- Usar o modelo de regressão de *Cox* com covariáveis dependentes do tempo, ou seja, suscetíveis a mudanças no decorrer do tempo influenciadas por variáveis macroeconômicas;
- Utilizar os modelos paramétricos de riscos proporcionais;
- Investigar critérios eficientes de seleção de covariáveis relevantes para um modelo de sobrevivência;
- Estudar métodos de seleção ou comparação de modelos (bondade de ajuste) para dados de sobrevivência;
- Acompanhar todo o processo de planejamento amostral a fim de evitar problemas com bases de dados com informações inconsistentes, inférteis e escassas;
- Finalmente, utilizar essas metodologias para desenvolver um sistema de *score* para outros produtos como, aquisição de um novo produto, *behavior scoring*, fraudes bancárias entre outros.

Bibliografia

[1] ABREU, H. J. Aplicação da Análise de Sobrevivência em um Problema de Credit Scoring e Comparação com a Regressão Logística. Dissertação de Mestrado, UFSCar, 2004.

[2] ANDERSON, R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Great Britain: Oxford University Press, 2007.

[3] ANDREEVA, G. European Generic Scoring Models Using Survival Analisys. J. Oper. Res. Soc. - JORS; 57(10), 1180-1190, 2006.

[4] BANASIK, J., CROOK, J. N., THOMAS, L.C. Not if But When Loans Default. J. Oper. Res. Soc. - JORS, 50, 1185-1190, 1999.

[5] CARVALHO, M. S., ANDREOZZI, V. L., CODEÇO, C. T., BARBOSA, M. T. S., SHIMAKURA, S. E. Análise de Sobrevida: Teoria e Aplicações à Saúde. Rio de Janeiro: Editora Fiocruz. 2005.

[6] CASELLA, G., GEORGE, E. I. Explaining the Gibbs Sampler. The American Statistician, 46, 167-174, 1992.

[7] COLLET, D. Modelling Survival Data in Medical Research. London: Chapman and Hall, 1994.

[8] COLOSIMO, E. A., GIOLO, S. R. Análise de Sobrevivência Aplicada. ABE - Projeto Fisher. São Paulo: Edgar Blücher, 2006.

[9] COX, D. R. Regression Models and Life Tables. J. Royal Statist. Society, Series B, 30, 248-275, 1972.

[10] COX, D. R. Partial Likelihood. Biometrika, 62, 269-276, 1975.

[11] DIAS, T. C. M. Análise Bayesiana Para Dados de Sobrevivência em Modelos de Riscos Proporcionais. Tese de Doutorado, UFRJ, 2002.

[12] DURAND, D. Risk Elements in Consumer Instalment Financing. National Bureau of Economic Research: New York, 1941.

[13] FISHER, R. A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188, 1936.

[14] GELFAND, A.E., SMITH, A.F.M. Sampling Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85, 398-409, 1990.

[15] GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B. Bayesian Data Analysis. London: Chapman and Hall, 2nd ed., 1995.

[16] GILKS, W. R., WILD, P. Adaptative Rejection Sampling for Gibbs Sampling. Applied Statistics, 41, 2, p. 337-348. 1992.

[17] HOSMER, D. W., LEMESHOW, S. Applied Survival Analysis: Regression Modelling of Time to Event Data. New York: John Wiley and Sons, Inc., 1999.

[18] HOSMER, D. W., LEMESHOW, S. Applied Logistic Regression. 2nd edition. New York: John Wiley and Sons, Inc., 2000.

[19] IBRAHIM, J. G., CHEN, M.-H.; SINHA, D. Bayesian Survival Analysis. New York: Springer Series in Statistics, 2001.

[20] LOUZADA-NETO, F. Lifetime Modelling for Credit Scoring: A New Alternative to Traditional Modelling Via Survival Analysis. Credit Technology, 2005.

[21] NARAIN, B. Survival Analysis and Credit Granting Decision.In L.C. Thomas, J.N. Crook, D. B. Edelman, Eds. Credit Scoring ans Credit Control, OUP, OXFORD, U.K., 109-121, 1992.

[22] PAULINO, C. D., TURKMAN, M. A. A., MURTEIRA, B. Estatística Bayesiana. Fundação Calouste Gulbenkian: Lisboa, 2003.

[23] PEREIRA, G. H. de A. Modelos de Risco de Crédito de Clientes: Uma Aplicação à Dados Reais. Dissertação de Mestrado. IME-USP. 2004.

[24] SANTOS, J. O., FAMA, R. An Evaluation on the Applicability of a Credit Scoring Model with Systemic and Non-Systemic Variables in Revolving Bank Credit Portfolio for Individuals. Revista Contabilidade e Finanças-USP, vol. 18, No 44, p. 105-117, São Paulo, Maio/Agosto, 2007.

[25] SINHA, D., DEY, D. K. Semiparametric Bayesian Analysis of Survival Data. Journal of the American Statistical Association, 92, 1195-1212, 1997.

[26] THOMAS, L. C., STEPANOVA, M. Survival Analysis Methods for Personal Loan Data. Operations Research, v.50, 277-289, 2002.

[27] TOMAZELA, S. M. O. Avaliação de Desempenho de Modelos de Credit Score Ajustados por Análise de Sobrevivência. Dissertação de Mestrado, IME-USP, 2007.

[28] WICHERN, D. W., JOHNSON, R. A. Applied Multivariate Statistical Analysis. USA. 2nd edition. Prentice-Hall - series in statistics, 1988.