JOANNE MEDEIROS FERREIRA

ANÁLISE DE SOBREVIVÊNCIA: UMA VISÃO DE RISCO COMPORTAMENTAL NA UTILIZAÇÃO DE CARTÃO DE CRÉDITO.

RECIFE-PE, 2007

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DFEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA

ANÁLISE DE SOBREVIVÊNCIA: UMA VISÃO DE RISCO COMPORTAMENTAL NA UTILIZAÇÃO DE CARTÃO DE CRÉDITO.

Dissertação apresentada ao Programa de Pós-Graduação em Biometria da UFRPE, como parte dos requisitos para obtenção do título do grau de Mestre em Biometria.

Autora: Joanne Medeiros Ferreira

Orientador: Professor Dr. Eufrázio de Souza Santos

Co-Orientador: Professor Dr. Borko D. Stosic

Ficha catalográfica

F383a Ferreira, Joanne Medeiros
Análise de sobrevivência: uma visão de risco comporta - mental na utilização de cartão de crédito / Joanne Medeiros Ferreira. - 2007.
73 f. : il.

Orientador : Eufrázio de Souza Santos
Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernambuco. Departamento de Estatística e Informática.
Inclui bibliografia

CDD 574. 018 2

1. Análise de sobrevivência
2. Cartão de crédito
3. Risco comportamental
4. Modelo linear generalizado
I. Santos, Eufrázio de Souza
II. Título

Universidade Federal Rural de Pernambuco
Departamento Estatística e Informática
Programa de Pós-Graduação em Biometria

ANÁLISE DE SOBREVIVÊNCIA: UMA VISÃO DE RISCO COMPORTAMENTAL NA UTILIZAÇÃO DE CARTÃO DE CRÉDITO

JOANNE MEDEIROS FERREIRA

Dissertação julgada adequada para obtenção do título de mestre em Biometria, defendida e aprovada por unanimidade em 30/03/2007 pela banca examinadora.

Orientador:

Prof. Dr. Eufrázio de Souza Santos
Universidade Federal Rural de Pernambuco – UFRPE

Banca Examinadora:

Prof. Dr. Borko Stosic
Universidade Federal Rural de Pernambuco – UFRPE

Prof. Dr. Gauss Moutinho Cordeiro
Universidade Federal Rural de Pernambuco – UFRPE

Prof. Dr. Marinho Gomes de Andrade Filho
Universidade de São Paulo – USP

A Fidelidade a Nós Próprios

De certo modo, o homem é um ser que nos está intimamente ligado, na medida em que lhe devemos fazer bem e suportá-lo. Mas desde que alguns deles me impeçam de praticar os atos que estão em relação íntima comigo mesmo, o homem passa à categoria dos seres que me são indiferentes, exatamente como o sol, o vento, o animal feroz. É certo que podem entravar alguma coisa da minha atividade; mas o meu querer espontâneo, as minhas disposições interiores não conhecem entraves, graças ao poder de agir sob condição e de derrubar os obstáculos. Com efeito, a inteligência derruba e põe de banda, para atingir o fim que a orienta, todo o obstáculo à sua atividade. O que lhe embaraçava a ação favorece-a; o que lhe barrava o caminho ajuda-a a progredir.

Marco Aurélio (Imperador Romano)

“Tudo em nós é mortal, menos os bens do espírito e da inteligência”.

Ovídio - Roma Antiga

Dedico este trabalho as cinco pessoas mais importantes que Deus colocou em minha vida***: “Meu filho João Pedro, minha mãe Lucicleide, meu Pai Elielson, e meus irmãos Julianne e Eduardo”.***

"Deus nos fez perfeitos e não escolhe os capacitados, capacita os escolhidos. Fazer ou não fazer algo só depende de nossa vontade e perseverança".

Albert Einstein

Agradecimentos

- Agradeço ao meu Deus por ter me sustentado e pela esperança de melhores conquistas.
- Ao meu belo filho João Pedro que apesar de ser um bebe teve paciência de entender a nossa distância tão dolorosa, pelos beijos e carinho dado sempre que precisei de sua presença aqui em Brasília, por existir na minha vida, por ser meu companheiro de luta, por ser minha força.
- Aos meus pais pelo incentivo nos momentos em que o desânimo e as forças precisavam ser restabelecidos, por estarem sempre me apoiando além dos maravilhosos valores que me foram ensinados.
- As minhas queridas avós Jesilda e Lelita, tias e tios que sempre acreditaram em mim, e apesar da distância viveram comigo todas as etapas deste meu objetivo.
- Aos meus queridos irmãos por toda força e incentivo e os ombros nos momentos que precisei chorar e seguir em frente, por existirem na minha vida pois são meus melhores amigos.
- Ao meu orientador, Professor Eufrázio Santos, por seu desprendimento e bom senso que foram essenciais para que eu chegasse até aqui. Admiro muita sua habilidade de tornar o ambiente acadêmico extremamente agradável.

- As professoras Maria Cristina, Claudia Regia e Jacira Guido do departamento de Estatística da UFPE pelo apoio e atenção a mim desperdiçada a fim de me ajudar.
- A minha terapeuta Maria Aparecida pelo bom trabalho em me ajudar a repensar

  nos meus objetivos que estavam se perdendo pela falta da família e me mostrou que tudo na vida é passageiro e que só nos mesmos com força e perseverança podemos fazer nosso caminho.
- Aos meus amigos e companheiros de trabalho do Banco do Brasil Marcelo Augusto, Norton, Marina, Lia, Guilherme, Fátima, Mario, Denis, Mônica, Lílian, Flávio e Mieto pela compreensão e apoio dado ao longo das minhas atividades acadêmicas.
- Aos meus amigos de mestrado em especial Luciano e Oscar que mesmo no período de distância me estimularam a chegar ao fim.
- A todos os meus amigos de Brasília em especial Laylla, Lídia, Alexandre, Vinicius e Bianca pelo apoio nos momentos de dificuldades principalmente emocional onde à distância da família me fazia perder as referencias e forças.
- E a todas as pessoas que passaram em minha vida durante este período, pois de suas passagens tirei aprendizados essências para seguir meu caminho, pois como diz Milton Nascimento “tem gente que vem para ficar, mas tem gente que vai para não mais voltar e é a vida!”

Resumo

Este trabalho visa, através de técnicas estatísticas, apresentar *metodologicamente* o comportamento dos consumidores do produto cartão de crédito e agregar valor ao processo de marketing estratégico de uma instituição financeira. Apresentam-se metodologias, baseadas em técnicas como analise de sobrevivência e regressão logísticas. Estas ferramentas são capazes de prever o comportamento de clientes e assim focar estratégias de modo a alocar recursos para impedir a migração dos mesmos para instituições concorrentes. Com a técnica de Analise de Sobrevivência clássica será possível analisar o tempo de vida da conta cartão e as variáveis que apresentam indícios de migração para concorrência e com a técnica de regressão logística será feita uma comparação afim de validação do modelo, além de mostrar se existem outras variáveis que também demonstram o encerramento do produto cartão por parte do cliente.

Abstract

Through the use of statistical techniques, this paper aims to present the behavior of credit card consumers, as well as to add value to the process of strategical marketing in a financial institution. For doing so, we present methodologies based on the techniques of logistical regression and survival analysis. Such tools are capable of foreseeing the behavior of customers, and thus focus on certain strategies in order to place resources that may hinder their migration to other financial institutions. The use of survival analysis methods makes it possible to analyze which variant presents evidence of customer migration; the logistical regression technique will be used in order to compare and validate the models hereby proposed, besides showing whether there are other variants that also demonstrate the closing of credit card accounts by the customers.

Sumário

Lista de Tabelas

1. Introdução

Para sobreviver financeiramente e estrategicamente no mercado financeiro cada vez mais competitivo, as instituições financeiras têm investido em ferramentas que tem como base técnicas estatísticas, capazes de prever o comportamento de seus clientes. Neste trabalho iremos nos focar em um dos produtos que vem apresentando mais resultados e assim acirrando a concorrência, de modo a alocar recursos de marketing para intervir nos processos de venda e impedir que os clientes fidelizados deixem de utilizar o produto, o que nos dá indícios de uma possível migração para concorrência.

A técnica analise de sobrevivência que vem sendo muito utilizado hoje na área médica ou financeira, onde a variável resposta é o tempo transcorrido até a realização de algum evento de interesse. A análise de sobrevivência é um conjunto de modelos e técnicas estatísticas adequados a lidar com dados deste tipo. Apesar de o nome ser referenciado a área Biomédica o uso desta técnica surgiu de uma empresa de seguro que estava desenvolvendo métodos de custo de prêmios de seguro de vida. Desta forma hoje este tipo de técnica vem sendo utilizada para observar o tempo de vida do cliente em instituições financeiras (Souza Flávio, Maeda Lilia, Fregonasse Isabela – Estimando a Evasão de Clientes em Instituições Financeira, SEAGRO 2004), para que se possam fazer ações de marketing no intuito de evitar a saída de clientes.

O evento de interesse é frequentemente referido como “falha” embora este evento possa ser, por exemplo, ocorrência de encerramento de conta, casamento, mudança de residência, promoção em uma empresa, ou a execução de uma tarefa em um experimento de psicologia. Em nossa aplicação iremos usar como falha o momento de cancelamento da conta cartão (tipo de conta em que o cliente utiliza função crédito).

Com essa dissertação venho mostrar as vantagens práticas destas duas técnicas e compara-las afim de especialmente conseguir resultados reais provando a utilidade de

modelos estatísticos a fim de introduzi-los como ferramenta para analise comportamental estimar o risco do encerramento da conta cartão clientes correntistas de uma instituição financeira.

2. Revisão da Literatura

Uma importante contribuição para Análise de Sobrevivência foi sem duvida o Modelo de regressão de COX (1972) onde abriu uma nova fase na modelagem de dados clínicos. Uma evidência quantitativa desse fato aparece em STIGLER (1994). O autor usa citações feitas a periódicos indexados de todas as áreas entre os anos de 1987 e 1989, para quantificar a importância de algumas publicações na literatura estatística. O artigo de COX (1972), em que o modelo é apresentado, foi neste período o segundo artigo mais citado na literatura estatística, somente ultrapassada pelo artigo de KAPLAN e MEIER (1958). Isto significa, em números, uma média de 600 citações por ano.

BERKSON E GAGE (1950), juntamente com CURTLE e EDERER (1958) e Gehan (1969) desenvolveram métodos que estimavam a função de sobrevivência e com isso surgia à tabela de vida.

ABRAMS (1992) descreveu modelos para os dados de sobrevivência usando o modelo de risco proporcional de COX e uma abordagem completamente paramétrica, no contexto de exames de grupos paralelos aleatórios.

CULTER E EDDERER (1958) juntamente com KAPLAN E MEIER (1958) comprovaram que como variável de interesse, o tempo até a morte, motivaram o desenvolvimento de uma metodologia estatística para tratar censuras, fazendo com que esta área ficasse conhecida como analise de sobrevida, que é estimada em função do tempo. BRESLOW E CROWLEY (1974), EFRON (1967), MEIER (1975) E AaLEN (1976) também desenvolverão metodologias estáticas para tratar a censura através de estimadores, mostrando assim a possibilidade da utilização de estimadores viciados e não-viciado.

KALBFLEISCH e PRENTICE (2002) em seus estudos registraram que se o tempo de falha não pôde ser observado, é registrado período de tempo de falha em que o individuo é superior ao período de observação e essas informações parciais são caracterizadas como censura á direita.

Um dos instrumentos mais antigos a tabela de vida que é muito utilizada pelas companhias de seguro desde o século XVII foi desenvolvida por BERKSON E GAGE (1950), CURTLE e EDERER (1958) e GEHAN (1969).

A partir do ultimo quarto do século XX, foram publicados muitos livros-textos descrevendo métodos tradicionais e modernos, que variam em sua sofisticação matemática, como MILLER e col.(1981), LAWLESS (1982), COX E OAKES (1984), KALBFLEISCH e PRENTICE (2002), COLLETT (2003), KLEIN e MOESCHBERGER (2003), os quatros últimos, reedições. Livros mais recentes têm tratado de temas específicos da analise de sobrevida, com o de THERNEAU e GRAMBSCH (2000), sobre extensões do modelo semiparamétrico de COX e o de IBRAHIM e col. (2001), para a análise de sobrevida com enfoque bayesiano.

Segundo MARTINS (1988) em muitas situações práticas o pesquisador vê-se envolvido com a necessidade de construir um modelo cientifico pode ser definido como uma abstração de um sistema real, que possa ser utilizada com os propósitos de predição e controle. E, que para ser útil, tal modelo deve abranger elementos de dois atributos conflitantes: realismo simplicidade. Se por um lado o modelo deve servir como uma aproximação razoavelmente precisa do sistema real, e conter a maior parte dos aspectos importantes do mesmo, por outro, não deve ser tão complexo que se torne impossível compreendê-lo e manipulá-lo.

NELDER E WEDDERBURN (1972) introduziram a teoria dos modelos exponenciais lineares que são denominados como modelos lineares generalizados e

desempenham o mesmo papel da regressão normal. Esses estudos podem ser encontrados em HEARGETTY e ZEGER (1996) e LEE E NELDER (1996) onde estão relacionados os modelos das famílias exponenciais.

Referenciadas em probabilidades surgiram, no mercado financeiro brasileiro, as técnicas de análise matemática/estatística. Embora ainda estejam distantes de um processo consolidado, são utilizadas como auxiliares e, em muitos casos, como determinantes na decisão de crédito (MINUSSI, 2001). No seu aparecimento era bastante compreensível que o valor dessas técnicas como instrumento de decisão fosse protelado, devido à enorme quantidade de cálculos exigida para se obterem resultado consistente. Seu uso prático só foi possível com o desenvolvimento da informática. Conforme CAOUETTE, ALTMAN E NARAYANAN (1998), as técnicas mais utilizadas no campo econométrico são a análise logit e análise probit. O modelo logit assume que a probabilidade cumulativa de cancelamento de um crédito esteja situada entre 0 e 1, e que a probabilidade de perda seja logisticamente distribuída. A análise probit é semelhante ao modelo logit; porém assume que a probabilidade de perda tenha uma distribuição normal. Essas três técnicas modelam a probabilidade de inadimplência ou o prêmio de inadimplência, como variável dependente, cuja variância é explicada por um conjunto de variáveis independentes. Entre as variáveis independentes estão razões financeiras e outros indicadores, bem como as variáveis externas que mensuram as condições econômicas.

ALMEIDA e SIQUEIRA (1997) fazem uma citação sobre Regressão Logística na previsão de falência de bancos brasileiros. A Regressão Logística apresentou um fator diferencial que foi o de considerar bancos que a Regressão Logística não pode classificar por falta de dados.

A partir de OHLSON (1980) a Regressão Logística ou modelo LOGIT tem sido usado freqüentemente para a avaliação de riscos de inadimplência. OHLSON criticou o uso de análise discriminante tal como o modelo proposto por ALTMAN □L □L. (1977) por suas limitações: necessidade de normalidade da distribuição e sensibilidade à multicolinearidade entre as variáveis, além da necessidade de igualdade das matrizes de covariância entre os grupos, o que torna os coeficientes da função discriminante instáveis. Raramente os dados observados para as empresas seguem uma distribuição normal. A regressão logística não exige que a distribuição seja normal.

A Regressão Logística está mais próxima do procedimento de regressão múltipla, mas se diferencia desta por identificar diretamente a probabilidade de ocorrência de um evento (HAIR JR. L L. 1998), no caso deste estudo a identificação da probabilidade de insolvência.

3. Metodologia

Este estudo propõe a aplicação de duas técnicas estatísticas preditivas para estudar o encerramento e ativação da conta cartão de instituições financeiras: a primeira é a análise de sobrevivência, utilizada para predizer o momento provável da evasão dos clientes e a segunda é a regressão logística binomial que estima o risco de encerramento da conta cartão de cada cliente.

3.1 Análise de Sobrevivência

A análise de sobrevivência é uma técnica estatística muito usada em estudos da área médica, por envolver covariáveis que podem estar relacionadas com o tempo de sobrevivência. É uma das áreas da estatística que mais vem crescendo nos últimos 20 anos, devido ao numero de aplicações na área de saúde e na área financeira.

Na análise de sobrevivência a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Esse tempo é denominado tempo de falha, no estudo proposto é o tempo decorrido até o encerramento da conta cartão pelo cliente.

A principal característica da técnica de análise de sobrevivência é a presença de censura, que é basicamente a observação parcial da resposta, ou seja, por alguma razão o relacionamento do cliente observado, foi interrompido antes do final do estudo. Isto significa que toda a informação referente à resposta se resume ao conhecimento de que o tempo de falha é superior àquele observado.

3.1.1. Censura

Os estudos em análise de sobrevivência envolvem uma resposta temporal e são freqüentemente prospectivos e de longa duração. Porém podem terminar antes que ocorra o evento de interesse para todos os casos da amostra. Uma característica

importante decorrente destes estudos é a presença de observações incompletas ou parciais do tempo de sobrevivência denominadas censuras. Em um exemplo de estudo de hepatite (Soares e Colosimo, 1995), no grupo controle, que não recebeu o tratamento testado, oito pacientes não haviam morrido quando o estudo terminou e o acompanhamento de 5 outros cinco foi perdido no decorrer do estudo, ou seja, dos 15 pacientes deste grupo 13 foram censurados.

Pode-se observar que toda informação obtida por uma observação censurada é que o seu tempo de falha é superior ao tempo registrado. È importante notar que mesmo censuradas todas as observações de um estudo de sobrevivência devem ser usadas na análise estatística, pois mesmo incompletas fornecem informações sobre o tempo de falha e, as omissões destas observações no cálculo das estatísticas de interesse provavelmente resultarão em conclusões viciadas.

Existem três conhecidos mecanismos de censura. A censura do tipo I, também denominada censura á direita, ocorre quando o estudo é terminado após um período pré-estabelecido de tempo. As observações cujo evento de interesse não foi observado até este tempo são ditas censuradas. Outro tipo de censura, a do tipo II é aquela onde o estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos. O terceiro tipo que é a do tipo aleatória é o mecanismo de censura mais comum em estudos médicos e pode ocorrer se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse. Por exemplo, em um estudo médico o paciente após entrar no estudo decide não ir até o fim, seja porque ele mudou de local de residência, de hospital ou simplesmente porque perdeu o interesse no estudo. Neste caso a censura aleatória ocorre porque há perda de acompanhamento. Uma outra forma de ocorrer este tipo de censura é se o evento de interesse ocorrer por uma razão diferente da estudada. Em um estudo de câncer onde o evento falho é a morte do

paciente, se ele morrer, por exemplo, de um acidente automobilístico esta observação é dita censurada.

Para representar o processo de censura aleatório é necessário o uso de duas variáveis aleatórias. Suponha que o tempo de falha de uma observação seja representado pela variável aleatória *T* e seja *C* uma variável aleatória independente de *T* representando o tempo de censura associado a esta observação. Então os dados observados consistem em *t = min (T, C)* e o indicador de censura é dado por

$$\delta = \begin{cases} 0, \text{ se o tempo de sobrevivência é censurado} \\ 1, \text{ para tempo de sobrevivência não censurado, ou seja, T} > \text{C.} \end{cases}$$

3.1.2. Funções do Tempo de Sobrevivência

O tempo de sobrevivência de um indivíduo (no nosso caso será o tempo de vida da conta cartão) é uma variável aleatória *T* que pode assumir valores não negativos. Estes valores que *T* podem assumir têm uma distribuição de probabilidade que pode ser especificada de várias formas, duas da quais são particularmente úteis e bastante usadas para ilustrar diferentes aspectos dos dados em aplicações de sobrevivência: a função de sobrevivência e a função de risco.

3.1.3. Função de Sobrevivência

Suponha que a variável aleatória *T* tenha uma distribuição de probabilidade com função densidade de probabilidade *f(t).* A função de distribuição de *T* é então dada por

$$F(t) = P(T < t) = \int_0^t f(u)du$$

e representa a probabilidade de que o tempo de sobrevivência seja menor que algum valor t. A função de sobrevivência denotada por *S(t)* é definida então como a

probabilidade do tempo de sobrevivência ser maior ou igual de que um certo tempo *t*. Em termos probabilísticos isto é escrito como

$$S(t) = P(T \geq t).$$

Escrevendo em termos da função de distribuição tem–se que

$$S(t) = 1 - F(t)$$

Ou seja, em um estudo médico onde o evento de interesse é a morte, a função de sobrevivência fornece a probabilidade de um indivíduo sobreviver além de um tempo t.

A função de sobrevivência é uma função não crescente no tempo com as propriedades de que a probabilidade de sobreviver pelo menos ao tempo zero é um e a probabilidade de sobreviver no tempo infinito é zero. Isto é,

$$S(t) = 1 \text{ } para \text{ } t = 0$$

$$S(t) = 0 \text{ } para \text{ } t = \infty.$$

Para descrever a função de sobrevivência é geralmente utilizada uma representação gráfica de *S(t),* ou seja, o gráfico de *S(t)* versus t que é chamado de curva de sobrevivência. Uma curva íngreme representa razão de sobrevivência baixo ou curto tempo de sobrevivência e uma curva de sobrevivência gradual ou plana representam taxa de sobrevivência alta ou sobrevivência longa.

A curva de sobrevivência pode ser usada para comparar distribuições de sobrevivência de dois ou mais grupos e também para determinar quantidades relevantes tal como a mediana e outros percentil. È importante salientar que tratando de distribuições de sobrevivência assimétricas, a média não deve ser usada para descrever a tendência central da distribuição. Neste caso a mediana deve ser utilizada devido à influência que valores extremos, tempos de vida muito curtos ou longos, proporcionam na média.

3.1.4. Função de Risco

As funções F(t) e f(t) fornecem duas formas, matematicamente equivalentes, de especificar a distribuição de uma variável aleatória contínua não-negativa, contudo existem outras funções equivalentes que podem ser usadas. Uma função especial bastante utilizada, devido a sua interpretação em análise de sobrevivência é a função de risco denotada por h(t). A função de risco do tempo de sobrevivência T fornece a taxa de falha condicional, ou seja, é definida como a taxa de falha em um intervalo pequeno de tempo, assumindo que o indivíduo tenha sobrevivido até o início do intervalo. Para se obter uma definição formal da função de risco considere um intervalo de tempo *[t, t+Δt)* e expresse a probabilidade de uma observação falhar neste intervalo em termos da função de sobrevivência como

$$S\,(t) - S\,(t+\Delta t)$$

A taxa de falha no intervalo *[t, t+Δt)* é definida como a probabilidade de que a observações falhe neste intervalo, dado que não falhou antes de t, dividida pelo comprimento do intervalo. Dessa forma a taxa de falha no intervalo *[t, t+Δt)* é expressa por

$$\frac{\mathrm{S\,(t) - S\,(t + \Delta t)}}{[(t + \Delta t) - t]S(t)}$$

Assim h(t) pode ser escrita como

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

Para Δt pequeno, *h(t)* apresenta a taxa de falha instantânea no tempo *t* e é também denominada de função de taxa de falha ou taxa de mortalidade condicional. A função de risco desempenha um papel importante na análise de dados de sobrevivência sendo bastante útil para especificar a distribuição do tempo de vida, pois descreve a

forma em que a taxa instantânea de falha muda com o tempo. A função de risco é então definida como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t / T \ge t.)}{\Delta t}$$

Pode-se escrever a função de risco em termos da função de distribuição *F(t)* e da função densidade de probabilidade *f(t)* da seguinte forma

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Existe o modelo de riscos proporcionais onde para estimar os parâmetros do modelo utilizamos o método de máxima verossimilhança que permite usar a informação que se tem nos casos de censura. O método combina as observações censuradas e não censuradas de tal modo que produz, sob certas condições de regularidade, estimativas consistentes e assintoticamente normais.

O método de máxima verossimilhança é certamente uma das técnicas mais utilizadas na estimação paramétrica, quando a forma da distribuição geradora dos dados é conhecida. Uma desvantagem da estimação por máxima verossimilhança é a exigência de que se especifique a forma de $h_0(t)$. Para estimar os parâmetros do modelo semiparamétrico, e usada a função de verossimilhança parcial proposta por COX (1972) que, para estimar β, condicionou a verossimilhança, eliminado $h_0(t)$.

3.1.5. Estimação da Função de Sobrevivência

Um passo inicial nos estudos de tempo de vida é usualmente a estimação da sobrevivência. Estes estudos freqüentemente apresentam observações censuradas, o que requer técnicas estatísticas especializadas para acomodar a informação contida nestas observações. Algumas técnicas estatísticas podem ser utilizadas para analisar dados de

tempo de sobrevivência na presença de censura. Podem ser citados três estimadores não-paramétricos que serão apresentados a seguir, usados para estimação da função de sobrevivência: a tabela de vida, o estimador de Kaplan-Meier entre outros. Estes estimadores são conhecidos como não-paramétricos, pois usam os próprios dados para estimar as quantidades necessárias da análise, sem fazer uso de suposições a respeito da forma da distribuição dos tempos de sobrevivência.

3.1.6. Tabela de Vida

A tabela de vida que também é conhecida como método atuarial é um dos instrumentos estatísticos mais antigos utilizados pelas companhias de seguro desde o século XVII. Berkson e Gage (1950), Curtle e Ederer (1958) e Gehan (1969) desenvolveram métodos para estimação das funções de sobrevivência. A tabela de vida é considerada como um procedimento que mostra a distribuição do tempo de sobrevivência para grupos homogêneos de indivíduos, requerendo um número grande de observações de no mínimo 30 para que os tempos possam ser agrupados em intervalos.

Para construir uma tabela de vida primeiramente dividi-se o período total de observação em certo número de intervalos e para cada intervalo estima-se o valor da taxa de falha e a partir da obtenção desses valores estima-se a função de sobrevivência. A taxa de falha ou função de risco foi definida anteriormente na Seção sobre função de risco como a probabilidade de uma observação falhar em certo intervalo de tempo dado que ela não falhou até o início deste intervalo. Esta função pode ser estimada na tabela de vida a partir de dados censurados por

$$\hat{h}(t_i - 1) = \frac{\text{Nº falhas em } [t_{i-1}, t_i)}{(\text{Nº sob risco em em } t_{i-1}) - (\text{Nº censuras em em } [t_{i-1}, t_i))/2} \quad (3.1.1)$$

em que *i = 1... , n, t = t_1... , tn e t_0 = 0.* Verifica-se na Equação (3.1.1) que observações censuradas no intervalo são tratadas como se estivessem sob risco durante a metade do intervalo considerado. Suponha um estudo iniciado com n indivíduos, a probabilidade de falhar até t_1 é $\hat{h}(t_1)$, ou seja, dos n indivíduos n[$\hat{h}(t_1)$] não chegarão a t_1. Assim no final do primeiro período n[1 – $\hat{h}(t_1)$] indivíduos ainda estarão vivos. Dessa maneira a função de sobrevivência, que é a probabilidade de sobreviver além de t_1 pode então ser estimada por

$$\hat{S}(t_1) = \frac{n[1 - \hat{h}(t_1)\ ]}{n} = 1 - \hat{h}(t_1)$$

De forma análoga, dos $n[1 - \hat{h}(t_1)\ ]$ indivíduos que sobreviveram ao final do primeiro período apenas $n[1 - \hat{h}(t_1)\ ][1 - \hat{h}(t_2)\ ]$ chegarão ao final do segundo período. Portanto

$$\hat{S}\ (t_2) = [1 - \hat{h}(t_1)\ ][1 - \hat{h}(t_2)\ ].$$

Assim, de uma forma geral, para qualquer tempo *t* o estimador atuarial da função de sobrevivência é definido por:

$$\hat{S}_{TV}(t_i) = \prod_{j=1}^{i} [1 - \hat{h}(t_{j-1})], \ j \le i \quad (3.1.2)$$

Uma estimativa gráfica da função de sobrevivência será uma função escada, com valores constantes da função em cada intervalo de tempo.

3.1.7. Estimador de Kaplan-Meier

Este estimador é sem dúvidas o mais utilizado em estudos estatísticos e atuariais, foi proposto por Kaplan e Meier em 1958 e é também conhecido como estimador produto-limite. A construção do estimador de Kaplan-Meier considera o número de intervalos iguais ao número de falhas distintas e os limites dos intervalos são os próprios tempos de falhas da amostra. Sejam $t_1, t_2, \ldots, t_n$ os tempos de falhas de maneira que $t_1 \leq t_2 \leq \ldots \leq t_n$.

O estimador de Kaplan-Meier é então definido como

$$\hat{S}_{KM}(t) = \prod_{i/t<t} \frac{n_i - d_i}{n_i} \quad (3.1.3)$$

Onde d_i é o número de falhas no tempo t_i e n_i é o número de indivíduos que não falharam e não foram censurados até o tempo *ti* (exclusivo). Pode-se verificar que o estimador de Kaplan-Meier pode ser obtido a partir da Equação (3.1.2) a função de risco estimada igual à d_i/n_i. Em seu artigo original Kaplan e Meier justificaram a equação (3.1.3) mostrando que ela é o estimador de máxima verossimilhança da função de sobrevivência *S(t)*.

As propriedades assintótica destes dois estimadores descritos anteriormente foram estudadas por alguns autores tais como, Kaplan e Meier (1958), Breslow e Crowley (1974), Efron (1967), Meier (1975) e Aalen (1976). Estes estudos mostraram que o estimador de Kaplan-Meier é não-viciado em grandes amostras e em amostras de tamanhos menores existem algumas evidências empíricas da superioridade deste estimador. A principal diferença entre a tabela de vida e o estimador de Kaplan-Meier é o número de intervalos utilizados na construção dos mesmos. Na tabela de vida os

tempos de falhas são agrupados em intervalos de forma arbitrária enquanto que o estimador de Kaplan-Meier é baseado em um número de intervalos igual ao número de tempos de falha distintos. Usualmente o estimador de Kaplan-Meier considera um número de intervalos maior que o número de intervalos da tabela de vida, confirmando a superioridade do mesmo dado que quanto maior o número de intervalos melhor a aproximação para a verdadeira distribuição do tempo de falha.

3.2. Modelo de Riscos Proporcionais de Cox

3.2.1. Forma e Estimação do Modelo

Os estudos em análise de sobrevivência muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Essas covariáveis devem ser incluídas na análise estatística dos dados para explicar seu possível efeito no tempo de sobrevivência. Uma das alternativas metodológicas que incorpora informações no estudo do tempo de sobrevivência através da introdução de covariáveis é o modelo de riscos proporcionais. Uma família de riscos proporcionais é uma classe de modelos com a propriedade de que diferentes indivíduos têm funções de riscos proporcionais. Ou seja, a razão entre duas funções de riscos para dois indivíduos distintos não varia com o tempo.

Isto implica que a função de risco no tempo t, dado x, pode ser escrita na forma:

$$h(t/x) = h_0(t)g(x,\beta) \quad (3.2.1)$$

em que $h_0(t)$ é uma função arbitrária de risco padrão ou de base, x é o vetor de covariáveis fixas, g é uma função que deve ser especificada e β é o vetor de parâmetros regressores associado com as covariáveis. Sob a suposição de riscos proporcionais, Cox

propôs em 1972 o Modelo de Riscos Proporcionais de Cox onde a parte paramétrica do modelo $g(x, \beta)$ é geralmente tomada como $exp(x,\beta)$ (Cox, 1972).

O conjunto de valores das covariáveis no modelo de riscos proporcionais de Cox será representado pelo vetor x, tal que $x = (x_1, x_2, \ldots, x_p)$. Seja t_i o tempo de sobrevivência do i-ésimo indivíduo que possivelmente depende do valor dessas p covariáveis. Dessa maneira o principal interesse em problemas como este é avaliar como estas covariáveis influenciam t_i. Então no modelo de riscos proporcionais de Cox a função de risco do

i-ésimo indivíduo pode ser escrita como

$$h_i(t / x_{1i},\ldots,x_{pi}) = h_0(t)\exp(\beta_1 x_{1i} + \ldots + \beta_p x_{pi})$$

Ou de forma equivalente

$$h_i(t / x_i) = h_0(t)\exp(x_i'\beta)$$

em que $\beta' (\beta_1, \ldots, \beta_p)$ é um vetor de parâmetros desconhecidos e $x_i = (x_{1i}, \ldots, x_{pi})$. Este modelo é chamado de riscos proporcionais devido à propriedade de que a razão das taxas de falha de dois indivíduos diferentes é constante no tempo. Ou seja, a razão das funções de risco para dois indivíduos *i* e *j* é dada por:

$$\frac{h_i(t)}{hj(t)} = \frac{h_0(t)\exp(x_i'\beta)}{h_0(t)\exp(x_j'\beta)} = \exp(x_i'\beta - x_j'\beta). \quad (3.2.3)$$

Esta razão não depende do tempo, isto é, o risco de falha de um indivíduo em relação ao outro é constante para todos os tempos de acompanhamento. Os dois componentes multiplicativos do modelo são de naturezas distintas, um não-paramétrico e o outro paramétrico sendo esta a razão do modelo ser do tipo semi-paramétrico o que o torna bastante flexível. O componente não-paramétrico, $h_0(t)$, não especificado, é uma função não negativa no tempo geralmente chamado de função de base, pois $h(t) = h_0(t)$ quando $x = 0$. O componente paramétrico é em geral usado em termo multiplicativo e

por ser na forma exponencial garante que *h(t)* será positiva. Um exemplo da flexibilidade deste modelo é possuir alguns modelos conhecidos como casos particulares tal como o modelo de regressão Weibull (Kalbleisch e Prentice, 1980).

O modelo de regressão de Cox é caracterizado pelos coeficientes β que medem o efeito das covariáveis sobre a função de risco. Dessa maneira é necessário um método de estimação para se fazer inferência no modelo. O método de máxima verossimilhança usual, bastante conhecido e frequentemente usado, não podem ser utilizados aqui, pois a presença do componente não-paramétrico $h_0(t)$ na função de verossimilhança torna este método inapropriado. Frente a tal dificuldade, Cox (1975) propôs o método de máxima verossimilhança parcial que condiciona à verossimilhança a história dos tempos de sobrevivência e censuras anteriores e desta forma elimina a função de base desconhecida.

3.2.2. Verossimilhança Parcial

Nos intervalos onde nenhuma falha ocorre não existe nenhuma informação sobre o vetor de parâmetros β, pois $h_0(t)$ pode, teoricamente, ser identicamente igual a zero em tais intervalos. Uma vez que é necessário um método de análise válido para todas $h_0(t)$ possíveis, a consideração de uma distribuição condicional é necessária. Considere uma amostra de n indivíduos, onde se têm $k(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \ldots \leq t_k$. A probabilidade condicional da i-ésima observação vir a falhar no tempo t_i, conhecendo quais observações estão sob risco em t_i é:

$$\frac{h_i(t_i)}{\sum_{j \in R(t_i)} hj(t_i)} = \frac{h_o(t_i)\exp(x'_i\,\beta)}{\sum_{j \in R(t_i)} h_o(t_i)\exp(x'_j\,\beta)} = \frac{\exp(x'_i\,\beta)}{\sum_{j \in R(t_i)} \exp(x'_j\,\beta)} \qquad (3.2.4)$$

Em que, $R(t_i)$ é o conjunto dos índices dos indivíduos sob risco no tempo t_i. Pode-se verificar que ao utilizar a probabilidade condicional, o componente não-paramétrico $h_0(t)$ desaparece da equação (3.2.4).

A função de verossimilhança parcial $L(\beta)$ é obtida fazendo o produto dessas probabilidades condicionais, associadas aos distintos tempos de falha, ou seja,

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

$$= \prod_{i=1}^{n} \left[\frac{\exp(x_i'\beta)}{\sum_{j \in R(t_i)} \exp(x_j'\beta)} \right]^{\delta_i} \quad (3.2.4.b)$$

Em que

$$\delta_i = \begin{cases} 0, \text{se o i - ésimo tempo de sobrevivência é censurado,} \\ 1, \text{caso contrário.} \end{cases}$$

A função $l(\beta)$ é o logaritmo da função de verossimilhança, ou seja, $l(\beta) = log(L(\beta))$ e $U(\beta)$ é o vetor escore composto das primeiras derivadas da função $l(\beta)$. Estimadores para o vetor de parâmetros β podem ser obtidos maximizando o logaritmo da função de verossimilhança parcial (3.2.4.b), ou seja, resolvendo o sistema de equações definido por $U(\beta) = 0$. Isto é o equivalente a

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x'_i \beta - \log \sum \exp(x'_i \beta) \right] = 0 \quad (3.2.5)$$

O procedimento de estimação requer um método iterativo que é geralmente o método de Newton-Raphson, pois as equações encontradas (3.2.4.b) não apresentam

forma fechada. Cox (1975) mostra informalmente que o método usado para construir esta verossimilhança gera estimadores que são consistentes e assintoticamente normalmente distribuídos, com matriz de covariâncias assintótica estimadas consistentemente pelo inverso do negativo da matriz de segundas derivadas parciais do logaritmo da função de verossimilhança. Provas formais destas propriedades foram apresentadas mais tarde por Tsiatis (1981) e Andersen e Gill (1982). A função de verossimilhança parcial em (3.2.4.b) é utilizada para tempos de sobrevivência contínuos e, portanto, não considera a possibilidade de empates dos valores observados. Entretanto, na prática, podem ocorrer empates nos tempos de falhas ou censuras devidas à escala de medida. No caso em que ocorrem empates entre falhas e censuras, ou seja, os tempos de falhas são iguais, mas um deles é censurado, para definir quais observações serão incluídas no conjunto de risco em cada tempo de falha usa-se a convenção de que a censura ocorreu após a falha.

No caso de empates entre falhas, a função de verossimilhança parcial (3.2.4.b) deve ser modificada para incorporar tais observações. A aproximação proposta por Breslow (1972) e Peto (1972) é frequentemente usada nos softwares estatísticos. Considere si o vetor composto pela soma das p covariáveis para os indivíduos que falham no tempo t_i, $i = 1,..., k$ e d_i é o número de falhas neste mesmo tempo. Esta aproximação considera a seguinte função de verossimilhança parcial

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(s_i'\beta)}{\left[\sum_{j \in R(t_i)} \exp(s_j'\beta)\right]^{d_i}}$$

Quando o número de observações empatadas em qualquer tempo é grande não é adequado o uso desta aproximação. Para estes casos é aconselhável utilizar o modelo de regressão de Cox para dados agrupados (Lawless, 1982; Prentice e Gloeckler, 1978).

Estimação da Função de Sobrevivência através de $h_0(t)$. Considerando que para um determinado indivíduo todas as covariáveis têm valores iguais a zero, pode-se então obter a função de sobrevivência padrão expressa por:

$$S_0(t) = \exp(-\int_0^t h_0(u)du)$$

Ou seja,

$$S_0(t) = \exp[-H_0(t)]$$

Em que $H_0(t)$ é a função de taxa de falha de base acumulada. Assim a função de sobrevivência pode ser definida como

$$S(t) = \exp - \int_0^t h\ (u/x)du$$

Substituindo a função de risco tem-se que

$$S(t) = \exp(-\int_0^t h\ (u/x)du) = \exp(-\exp(x'\beta)\int_0^t h_0(u)du)$$

Assim $S(t)$ pode ser expressa por

$$S(t) = \left[S_0(t)\right]^{\exp(x'\beta)}.$$

3.2.3. Testes para os Parâmetros do Modelo de Cox

O interesse do pesquisador freqüentemente está relacionado a verificar a associação de covariáveis ao tempo de sobrevivência. A hipótese nula pode então ser definida de maneira que todas as variáveis consideradas não explicam a variação no tempo de sobrevivência. Em outras palavras,

$$H_0 = \beta_1 = \beta_2 = \ldots = \beta p = 0 \quad (3.2.6)$$

Três testes podem ser usados para verificar esta hipótese nula global: o teste da razão de verossimilhança, o teste de Wald e o teste Escorem que são descritos a seguir.

3.2.4. Teste da Razão de Verossimilhança

Para comparar modelos encaixados ou verificar se um modelo particular é adequado, o uso de uma estatística de teste é requerido. Visto que a função de verossimilhança resume a informação contida nos dados sobre os parâmetros desconhecidos, uma estatística adequada é o valor da função de verossimilhança quando os parâmetros são substituídos pelas suas estimativas de máxima verossimilhança. Isto é a verossimilhança maximizada sob o modelo assumido. Seja $\hat{L}$ a verossimilhança maximizada para um dado modelo.

È mais conveniente usar menos duas vezes o logaritmo da verossimilhança maximizada como estatística de teste. Dessa maneira a estatística de interesse é dada por $-2log\hat{L}$. Dado que $\hat{L}$ é, na realidade, o produto de várias probabilidades condicionais, sendo dessa forma menor que 1, então $-2log\hat{L}$ será sempre positiva e para um dado conjunto de dados quanto menor o valor de $-2log\hat{L}$, melhor o modelo. Da mesma forma quanto maior o valor da verossimilhança maximizada melhor é o ajuste do modelo.

Esta estatística não pode ser usada como medida de adequação do modelo, mas para comparar distintos modelos ajustados para os mesmos dados. Assim para verificar a hipótese (3.2.6) defina um modelo de Cox onde nenhuma covariável tenha influência na sobrevivência e todos os indivíduos tenham o mesmo risco $h_0(t)$, ou seja, todos os coeficientes de regressão sejam iguais à zero. Este modelo denominado de modelo nulo tem verossimilhança maximizada associada denotada por $\hat{L}_0$. Por outro lado define-se $\hat{L}_v$ como a verossimilhança maximizada do modelo que contém ν coeficientes de

regressão estimados pelo método de máxima verossimilhança parcial. A estatística do teste da razão de verossimilhança parcial (RV) para testar o ajuste de cada modelo é definida como

$$RV = -2\log(\hat{L}_0 / \hat{L}_v) = -2\log(\log\hat{L}_0 - \log\hat{L}_v).$$

Sob a hipótese nula (3.2.6) de que os coeficientes são iguais a zero, esta estatística tem assintoticamente distribuição qui-quadrado com graus de liberdade igual à quantidade ν de coeficientes de regressão estimados. Para comparar os ajustes de dois modelos encaixados ao mesmo banco de dados, um com (ν+k) coeficientes regressores e o outro com ν coeficientes regressores, a estatística dada na equação acima se torna então:

$$RV = -2(\log\hat{L}_v + k - \log\hat{L}_v)$$

que também tem distribuição qui-quadrado com $(v + k) - v = k$ graus de liberdade. A hipótese nula então é a de que nenhuma melhora no ajuste do modelo foi verificada com a inclusão dos k coeficientes.

Os testes de Wald e o Escore também podem ser utilizados para o teste simultâneo de várias covariáveis. Apesar de ser preferível por questões de consistência e estabilidade nos métodos de cálculos associados, em amostras de tamanhos grandes os testes se tornam equivalentes.

3.2.5. Teste de Wald

Este teste é utilizado principalmente para verificar se um coeficiente particular é significativamente igual a zero na presença dos outros termos do modelo. Por exemplo, suponha que um modelo contenha três variáveis explicativas X_1, X_2 *e* X_3 com

coeficientes dados respectivamente por β_1, β_2 e β_3. A estatística de teste $(\hat{\beta}/DP(\hat{\beta}))$ é então usada para testar a hipótese nula $\beta_1 = 0$ na presença de β_2 e β_3. Caso não existam evidências para rejeitar esta hipótese, conclui-se que a variável X_1 não é necessária no modelo na presença de X2 e X3. O resultado isolado do teste de hipótese para um coeficiente particular pode não ser fácil de interpretar, pois em geral as estimativas individuais $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$, em um modelo de riscos proporcionais não são independentes umas das outras. Assim a hipótese nula β = 0 pode ser testada utilizando à estatística

$$Z = \frac{\hat{\beta}}{\sqrt{\hat{VAR}(\hat{\beta})}} \quad (3.2.7)$$

em que $\sqrt{\hat{VAR}(\hat{\beta})}$ é o erro padrão estimado de $\hat{\beta}$. A estatística (equação acima), sob H_0, tem uma distribuição normal padrão. Equivalentemente pode-se utilizar o quadrado desta estatística.

$$W = Z^2 = \frac{\hat{\beta}^2}{\hat{VAR}(\hat{\beta})}$$

que sob a hipótese nula tem distribuição qui-quadrado com 1 grau de liberdade. Valores de W superiores ao valor tabelado da distribuição qui-quadrado indicam que a covariável associada a β é importante para explicar a variação da resposta.

3.2.6. Teste Escore

A estatística do teste escore, assim como a do teste da razão de verossimilhança, é baseada diretamente na função de verossimilhança. Esta estatística denominada de *S* é definida, para testar a hipótese ($H_0 = 0$), por:

$$S = \frac{u^2(0)}{i(0)}$$

Em que

$$u(\beta) = \frac{\partial(\log L(\beta))}{\partial\beta}$$

É o escore eficiente e

$$i(\beta) = \frac{\partial^2(\log L(\beta)}{\partial\beta^2}$$

É a informação da função de Fisher observada. Sob a hipótese nula ($H_0 = 0$) *S* tem uma distribuição qui-quadrado com p graus de liberdade e valores de S maiores do que o valor tabelado da distribuição qui-quadrado implica que se deve rejeitar H_0. O teste escore tem uma forma aparentemente complexa. Entretanto de maneira mais resumida este teste pode ser definido como a razão entre o quadrado da primeira derivada do logaritmo da verossimilhança, com os parâmetros de interesse iguais a zero e a segunda derivada do logaritmo da verossimilhança, também avaliada com os parâmetros de interesse iguais a zero.

3.2.7. Covariáveis Dependentes do Tempo

Quando covariáveis são registradas para modelar os dados de sobrevivência, os valores tomados para tais covariáveis são aqueles medidos na origem do tempo ou no

início do estudo. Entretanto em muitos estudos que envolvem dados de sobrevivência existem outras covariáveis que são monitoradas durante o estudo e seus valores mudam neste período. Estas covariáveis cujos valores se alteram com o tempo são conhecidas como Covariáveis Dependentes do Tempo. Análises que consideram estas covariáveis podem fornecer resultados mais precisos e a não inclusão destes valores pode acarretar em sérios vícios. Estas covariáveis têm muita aplicação em análise de sobrevivência, pois podem ser utilizadas tanto para acomodar medidas que variam com o tempo durante um estudo como também podem ser úteis para modelar o efeito de indivíduos que mudam de grupo durante um tratamento.

Tais covariáveis podem ser consideradas dentro de duas amplas classificações referidas como covariáveis internas e covariáveis externas (Kalbfleisch e Prentice, 1958). Covariáveis internas são aquelas que caracterizam um indivíduo sob estudo e podem ser medidas apenas enquanto o paciente sobrevive. Os valores observados levam informação sobre o tempo de sobrevivência do correspondente indivíduo (paciente). Um exemplo pode ser a quantidade de glóbulos brancos no sangue.

Por outro lado covariáveis externas são variáveis que não necessariamente requerem a sobrevivência do paciente para sua existência. Um tipo de variável externa é aquela que muda de tal forma que seus valores serão conhecidos avançando em um tempo futuro. Existem alguns exemplos tais como a dose de uma droga que pode variar de maneira pré-determinada durante o estudo e, a idade de um paciente, uma vez que a idade no início do tratamento é conhecida, a idade do paciente em algum tempo futuro pode ser obtida de forma exata.

3.2.8. Modelo de Cox com Covariáveis Dependentes do Tempo

Os diferentes tipos de covariáveis dependentes do tempo apresentados na Seção 3.2.7 podem ser incorporadas ao modelo de regressão de Cox, generalizando-o como:

$$h_i(t) = h_0(t)\exp(x_i'(t)\beta) \quad (3.2.8)$$

È importante verificar que definindo desta forma, o modelo dado pela equação (3.2.8) não é mais de risco proporcional. Os valores das covariáveis $x_i(t)$ dependem do tempo t e a razão das funções de risco no tempo t para dois indivíduos i e j dada por

$$\frac{h_i(t)}{h_j(t)} = \exp(x'_i(t)\beta - x'_j(t)\beta),$$

é também dependente do tempo e a interpretação dos coeficientes do modelo deve considerar o tempo t. Os coeficientes β_l, $l = 1, \cdots, p$ podem, portanto ser interpretados como o logaritmo da razão de riscos para dois indivíduos cujo valor da *l−ésima* covariável no tempo t difere de uma unidade quando as outras covariáveis assumem o mesmo valor neste tempo.

Para obter as estimativas dos parâmetros do modelo de regressão de Cox com covariáveis dependentes do tempo, basta estender a função escore parcial para

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x'_i(t_i)\beta - \log \sum_{j \in R(t_i)} \exp(x'_j(t_j)\beta \right]$$

que é uma extensão da equação $U(\beta) = 0$, considerando covariáveis dependentes do tempo. Para construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo são necessárias propriedades assintótica dos estimadores de máxima

verossimilhança parcial. As provas mais gerais das propriedades para covariáveis dependentes do tempo foram apresentadas por Andersen e Gill (1982). Desta forma podem-se usar as estatísticas dos testes, apresentadas na Seção 4.2.3, para fazer inferências no modelo de regressão de Cox com covariáveis dependentes do tempo.

3.2.9. Modelo de Cox Estratificado

Uma importante generalização do modelo de Cox é assumir que a amostra de n indivíduos está dividida em s estratos e que há um risco basal específico para cada estrato. Se for assumido que os efeitos das covariáveis não variam entre os estratos, o modelo de Cox específico para o indivíduo *i* do estrato *j* é dado por:

$$h_j(t \mid x_{ij}) = h_{0j}(t)\exp(\beta' x_{ij})$$

Para $i = 1, 2, \ldots, n_j$ e $j = 1, 2. \ldots, s$; sendo que n_j é o número de indivíduos no estrato *j* e $h_{0j}(t)$ é o risco basal arbitrário para o estrato j. É assumido que o vetor de parâmetros, β, é comum a todos os estratos, o que equivale à suposição de não interação entre estrato e covariáveis.

O modelo de Cox estratificado é útil quando a suposição de riscos proporcionais é violada, já que, neste caso, o uso desnecessário da estratificação acarreta em uma perda de eficiência das estimativas obtidas. Uma solução para o problema é estratificar a amostra, de modo que a suposição seja válida em cada estrato, e ajustar o modelo acima. A estimação de β está baseada no produto das verossimilhanças parciais construídas para cada estrato. Assim, a verossimilhança parcial para o modelo de Cox estratificado é dada por:

$$PL(\beta) = \prod_{j}^{s} PL_j(\beta)$$

$$PL_j(\beta) = \prod_{i=1}^{n_j} \left(\frac{\exp(\beta' x_{ij}}{\Sigma_{l=1}^{n_j} Y_{ij}(t_{ij})\exp(\beta' x_{lj}}\right)^{\partial_i} \quad (3.2.9)$$

em que *Yij (t)* é um indicador de risco para o indivíduo *i* do estrato *j* no tempo *t* ($i = 1, 2, \ldots, n_j$) ($j = 1, 2, \ldots, s$).

Portanto, somente os n_j indivíduos do estrato j podem contribuir para verossimilhança parcial, $PL_j(\beta)$. Como $PL(\beta)$ não envolve $h_{0j}(t)$, o que muda na formulação da verossimilhança parcial do modelo de Cox estratificado é o uso de um conjunto de risco restrito ao estrato j, que considera somente os indivíduos do estrato j para a construção de $PL_j(\beta)$, segundo COLOSIMO (1997), as propriedades assintóticas destes estimadores podem ser obtidas como extensão dos resultados assintóticas de ANDERSEN e GILL (1982).

3.3. Modelos Lineares Generalizados (MLG)

Quando nos referimos aos modelos lineares generalizados estamos falando sobre modelos que têm o mesmo desempenho ou exerce o mesmo papel da regressão normal linear na década de 60. Eles foram introduzidos por Nelder e Wedderburn (1972) e são também denominados modelos exponenciais lineares. A classe dos MLGs é uma extensão do modelo linear clássico, pois não temos a suposição que a variável resposta tenha distribuição normal. Estes modelos baseiam-se na família exponencial uni paramétrica, que possui propriedades interessantes tanto para estimação quanto para testes de hipóteses, além de outros problemas de inferência. A definição de um modelo linear generalizado é dada por uma distribuição de probabilidade, membro da família exponencial de distribuições, para a variável resposta, um conjunto de variáveis independentes, descrevendo a estrutura linear do modelo, e uma função de ligação entre a média da variável resposta e a estrutura linear.

Alguns casos especiais de MLG's são

- Modelo normal linear;
- Modelos log-lineares aplicados à análise de tabelas de contingências;

- Modelo logístico para tabelas multidimensionais de proporções;
- Modelo probit para estudo de proporções;
- Modelo de regressão Poisson

Contudo os MLGs não englobam dados correlacionados e distribuições fora da família exponencial; extensões para estas situações podem ser encontradas em Heargerty e Zeger (1996) e Lee e Nelder (1996).

3.3.1. As componentes de um MLG

De uma forma geral, a estrutura de um MLG é formada por três componentes:

- Uma *aleatória*, composta de uma variável aleatória Y com n observações independentes, com um vetor de médias μ e uma distribuição pertencente à família exponencial de distribuições;
- Uma *sistemática* que define o preditor linear e,
- Uma *função de ligação* que relaciona as duas componentes anteriores.

$$\text{Pode - se mostrar que } E\left[\frac{\partial l}{\partial \theta_i}\right] = 0$$

i) *A Componente Aleatória*

Seja $y = (y_1,..., y_n)T$ um vetor de observações referente às realizações da variável aleatória $Y= (Y_1,..., Y_n)T$, independentes e identicamente distribuídas, com vetor de médias $\mu = (\mu_1,..., \mu_n)T$ e com função de densidade da forma

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i,\phi)\right\} \quad (3.3.1)$$

onde *a, b* e *c* são funções conhecidas, ϕ é o parâmetro de dispersão e θ_i é denominado parâmetro natural ou canônico, que caracteriza a distribuição na equação (3.3.1). Se ϕ é conhecido, a equação representa a família exponencial uni paramétrica.

A log-verossimilhança é definida por:

$$l(y_i;\theta_i,\phi) = \log f(y_i;\theta_i,\phi)$$

Portanto,

$$l(y_i;\theta_i,\phi) = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i,\phi) \quad (3.3.2)$$

Derivando esta equação sucessivamente com relação a θ_i temos

$$\frac{\partial l}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a(\phi)}$$

$$E\left\{\frac{[Y_i - b'(\theta_i)]}{a(\phi)}\right\} = 0$$

$$\frac{\partial^2 l}{\partial \theta^2{}_i} = -\frac{b''(\theta_i)]}{a(\phi)}$$

Pode-se também mostrar que $E\left[\frac{\partial^2 l}{\partial \theta_i^2}\right]$ e da equação $\frac{\partial l}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a(\phi)}$ temos que

$$E[Y_i] = \mu_i = b'(\theta_i).$$

De onde temos que

$$E\left[\frac{\partial^2 l}{\partial \theta_i^2}\right] E[Y_i] = \mu_i = b'(\theta_i).$$

Pode-se também mostrar que

$$E\left[\frac{\partial^2 l}{\partial \theta_i^2}\right] + E\left[\frac{\partial l}{\partial \theta_i}\right]^2 = 0$$

Então, a partir das equações anteriores podemos chegar em

$$E\left\{\frac{-b''(\theta_i)}{a(\phi)}\right\} + E\left\{\frac{[Y_i - b'(\theta_i)\ ]}{a(\phi)}\right\}^2 = 0$$

$$-\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{[a(\phi)]^2} E[Y_i - E(Y_i)]^2 = 0$$

$$\frac{1}{[a(\phi)]^2} Var(Y_i) = \frac{b''(\theta_i)}{a(\phi)},$$

Logo, $Var(Y_i) = a(\phi)b''(\theta_i)$, que pode também ser escrita na forma $Var(Y_i) = a(\phi)V_i$ onde $V_i = d\mu/d\theta_i$ é chamada **função de variância**.

Algumas distribuições com parametrização na família exponencial:

a) Normal

Seja Y uma variável aleatória com distribuição normal de média μ e variância σ^2, $Y \sim N(\mu\ ;\sigma^2)$. A densidade de Y é da forma:

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \quad (3.3.4)$$

onde $-\infty < \mu < +\infty$, $-\infty < y < +\infty$ e $\sigma^2 > 0$.

A expressão pode ser escrita na forma:

$$f(y) = \exp\left\{\frac{1}{\sigma^2}\left(\mu y - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right)\right\}.$$

Portanto,

$$\theta = \mu \quad a(\phi) = \sigma^2 \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2} \qquad c(y;\phi) = -\frac{1}{2}\left[\log(2\pi\sigma) + \frac{y^2}{\sigma^2}\right]$$

logo, $E(y) = b'(\theta) = \theta = \mu$

$Var(y) = a(\phi)b''(\theta) = \sigma^2$

Outras distribuições importantes para dados em que a variável resposta é contínua são as distribuições gama e normal inversa. A seguir apresentaremos distribuições relacionadas os dados com variável resposta discreta.

b) Poisson

Seja Y uma variável aleatória Poisson com parâmetro μ. A densidade de Y é expressa por:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad (3.3.5)$$

onde $\mu > 0$ e $y=0,1,2,...$

A expressão acima pode ser escrita da forma:

$$f(y) = \exp\{y\log\mu - \mu - \log(y!)\}$$

De onde temos que:

$$\log\mu = \theta, \quad b(\theta) = e^{\theta}, \quad a(\phi) = 1, \quad c(y;\phi) = -\log(y!)$$

Portanto,

$$E(Y) = b'(\theta) = e^{\theta} = \mu, \quad V = b''(\theta) = e^{\theta} = \mu, \quad Var(Y) = a(\phi)V = \mu.$$

c) Binomial

Seja Y uma variável aleatória binomial baseada em n repetições, denotada por $Y \sim B(n;\mu)$. Sua função de densidade de probabilidade é expressa por:

$$f(y;\mu) = \binom{n}{y}\mu^{y}(1-\mu)^{n-y}, \text{ onde } \mu > 0 \text{ e } y = 0,1,2,... \quad (3.3.6)$$

A expressão acima pode ser escrita da forma:

$$f(y) = \exp\left\{\log\binom{n}{y} + y\log\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu)\right\}$$

Comparando a expressão acima com a que caracteriza a família exponencial temos que:

$$\theta = \log\left(\frac{\mu}{1-\mu}\right), \quad b(\theta) = -n\log(1-\mu), \quad a(\phi) = 1, \quad c(y,\phi) = \log\binom{n}{y}.$$

Como $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, temos que $e^{\theta} = \frac{\mu}{1-\mu}$,

logo $\mu = \frac{e^{\theta}}{1+e^{\theta}}$ ou $1-\mu = \frac{1}{1+e^{\theta}}$ então, $b(\theta) = -n\log\left(\frac{1}{1+e^{\theta}}\right)$.

Portanto,

$$E(Y) = b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = n\mu,$$

$$V = b''(\theta) = n\frac{1}{(1+e^{\theta})}\frac{e^{\theta}}{(1+e^{\theta})} = n\mu(1-\mu),$$

$$Var(Y) = a(\phi)V = n\mu(1-\mu).$$

ii) Componente Sistemática

Considere a estrutura linear de um modelo de regressão $\eta = X\beta$, onde $\eta = (\eta_1, ...; \eta_n)^T$, $\beta = (\beta_1,....., \beta_p)^T$ e X é uma matriz modelo de dimensão n x p ($p < n$) conhecida, de posto p. A função linear dos parâmetros desconhecidos β é chamada de preditor linear e corresponde à parte **sistemática** de um MLG.

iii) Função de Ligação

Foi dito inicialmente que a função de ligação relaciona o preditor linear η à média μ do vetor de dados *y*, de modo que:

$$\mu_k = g^{-1}(\eta_k),\ k=1,\ldots,\ n.$$

onde $g(\eta_k)$ é uma função conhecida (monótona e diferençável) denominada **função de ligação.** No modelo normal linear a média e o preditor linear são idênticos, dado que η e μ podem assumir qualquer valor na reta real $(-\infty;+\infty)$. Dessa forma, temos que uma ligação do tipo identidade ($\eta = \mu$) é adequada para modelar dados normais. Se *Y* tem distribuição Poisson, com $\mu > 0$, a função de ligação adequada é a logarítmica ($\eta = \log_{\mu}$), pois esta tem o domínio positivo e o contradomínio na reta real. Para modelos que assumem a distribuição binomial para a variável resposta, onde $0 < \mu < 1$, o domínio da função de ligação deve, necessariamente, estar no intervalo (0; 1), enquanto que seu contradomínio é o intervalo $(-\infty;+\infty)$.

Três funções garantem esta condição para o modelo binomial:

1. Logit

$$\eta = \log\left(\frac{\mu}{1-\mu}\right)$$

2. Probit

$$\eta = \Phi^{-1}(\mu)$$

onde $\Phi(.)$ é a função de distribuição acumulada da normal reduzida;

3. Complemento log-log

$$\eta = \log\{-\log(1-\mu)\}$$

Aplicações importantes destes modelos com resposta binária tendo (1), (2) e (3) como funções de ligação são os chamados modelos de dose-resposta (como pode ser visto em Paula (2004)).

3.3.2. Estatística Suficiente e Ligações Canônicas

Os logaritmos da função de verossimilhança de um MLG com respostas independentes podem ser expressos na forma:

$$L(\theta; y) = \sum_{i=1}^{n} l(y_i; \theta_i, \phi). = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \quad (3.3.7)$$

Um caso particular importante ocorre quando o parâmetro canônico (θ) coincide com o preditor linear (η), ou seja, quando:

$$\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1,....,n.$$

Neste caso, $L(\theta;y) = L(\beta;y)$ é dado por:

$$L(\beta; y) = \sum_{i=1}^{n} \frac{y_i \sum_{j=1}^{p} x_{ij}\beta_j - b(\sum_{j=1}^{p} x_{ij}\beta_j)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \quad (3.3.8)$$

Definamos as quantidades:

$$S_j = \frac{1}{a(\phi)} \sum_{i=1}^{n} y_i x_{ij}, \quad j = 1,...., p.$$

Assim, $L(\beta;y)$ pode também ser expresso desta forma:

$$L(\beta; y) = \sum_{j=1}^{p} S_j \beta_j - \frac{1}{a(\phi)} \left(\sum_{i=1}^{n} b\left(\sum_{j=1}^{p} x_{ij} \beta_j \right) \right) + \sum_{i=1}^{n} c(y_i, \phi)$$

Logo, pelo teorema da fatorização, a estatística $S = (S_1,..., S_p)^T$ é suficiente mínima para o vetor $\beta = (\beta 1,..., \beta p)^T$. As ligações que correspondem a tais estatísticas são chamadas de ligações canônicas e desempenham um papel importante na teoria dos MLGs.

As ligações canônicas para os modelos Normais, Poisson e Binomial são apresentadas a seguir:

1. $\eta=\mu$, no modelo normal;
2. $\eta=\log(\mu)$, no modelo Poisson;
3. $\eta= \log[\mu/(1-\mu)]$, no modelo binomial;

Uma das vantagens de usar ligações canônicas é que as mesmas garantem a concavidade de $L(\beta; y)$ e consequentemente muitos resultados assintóticos são obtidos mais facilmente. Por exemplo, a concavidade de $L(\beta; y)$ garante a unicidade da estimativa de máxima verossimilhança de $\hat{\beta}$,quando esta existe.

Algoritmo de Estimação

O algoritmo de estimação dos parâmetros β's foi desenvolvido por Nelder e Wedderburn (1972) e tem como base um método semelhante ao de Newton-Raphson, conhecido como *Método Escore de Fisher*. A principal diferença em relação ao modelo

clássico de regressão é que as equações de máxima verossimilhança são não-lineares. Seja *l(*β*)* a log-verossimilhança como função de β e considere a função escore de Fisher:

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta},$$

E a matriz de informação de Fisher:

$$K = \left\{ E\left(\frac{-\partial^2 l(\beta)}{\partial \beta_j \partial \beta_s} \right) \right\} = -E\left(\frac{\partial U(\beta)}{\partial \beta} \right). \quad (3.3.9)$$

Expandindo a função escore em série de Taylor até primeira ordem obtemos:

$$U(\beta^{(m+1)}) = U(\beta^{(m)}) + \left(\frac{\partial U(\beta)^{(m)}}{\partial \beta} \right) [\beta^{(m+1)} - \beta^{(m)}] = 0$$

Ou desta forma:

$$\beta^{(m+1)} = \beta^{(m)} - \left(\frac{\partial U(\beta)^{(m)}}{\partial \beta} \right)^{-1} U(\beta^{(m)}),$$

onde o índice (*m*) significa o valor do termo na *m*-ésima iteração.

O método escore de Fisher (1925) é obtido pela substituição de $-\frac{\partial U(\beta)}{\partial \beta}$ pelo seu valor esperado *K*.

A matriz de informação esperada para β, depois de várias manipulações é dada por: $K = \frac{1}{a(\phi)} X^T W X$

onde W é uma matriz diagonal de pesos definidos por

$$w_i = V_i^{-1} g'(\mu_i)^{-2} .$$

A função escore, usando esta matriz de pesos, é expressa como

$$U(\beta) = \frac{1}{a(\phi)} X^T W z,$$

onde *z* é um vetor com dimensão *n*x*1*dado por

$$z_i = (y_i - \mu_i)\left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right).$$

Utilizando estes dois resultados, o algoritmo escore de Fisher para calcular a estimativa de máxima verossimilhança (EMV) de β é expresso por

$$\beta^{(m+1)} = \beta^{(m)} + \left(X^T W^{(m)} X \right)^{-1} X^T W^{(m)} z^{(m)}. \quad (3.3.10)$$

Trabalhando a expressão acima, chegaremos a um processo iterativo de mínimos

$$\beta^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} z^{(m)}$$

Quadrados ponderados:

m=0,1,...., onde $z = \eta + W^{-1/2} V^{-1/2}(y-\mu)$. Note que *z* desempenha o papel de uma variável dependente modificada, enquanto *W* é uma matriz de pesos que muda a cada passo do processo iterativo. A convergência das expressões acima ocorre em um número finito de passos, independente dos valores iniciais utilizados. É usual iniciar com $\eta^{(0)} = g(y)$. Para ilustrar tomemos o caso do modelo binomial logístico linear, que é de nosso principal interesse, tem-se:

$$z = \eta + W^{-1/2} V^{-1/2}(y - \mu) = \eta + W^{-1}(y - \mu) = \eta + \frac{(y - n\mu)}{n\mu(1 - \mu)}.$$

Sob condições gerais de regularidade, temos que $\hat{\beta}$ é um estimador consistente e eficiente de β e que:

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, \phi^{-1}\sum{}^{-1}(\beta)) \quad \text{quando} \quad n \longrightarrow \infty$$

onde

$$\sum(\beta) = \lim_{n \longrightarrow \infty} \frac{K(\beta\ )}{n},$$

sendo $\sum(\beta)$ uma matriz definida positiva.

As condições para que $\sum(\beta)$ exista e seja definida positiva são:

$$\frac{n_i}{n} \longrightarrow a_i > 0, \text{ quando } n \longrightarrow \infty$$

e que $\sum_{i=1}^{g} x_i x_i^{'}$ seja do posto completo, onde $n=n_1+.....+n_g$.

Sob certas condições de regularidade, temos que:

$$\sqrt{n}(\hat{\phi}-\phi) \longrightarrow N(0, \sigma_\phi^2), \text{ quando } n \to \infty,$$

onde $\sigma_\phi^2 = \lim_{n \longrightarrow \infty} -n[\sum_{i=1}^{n} c^{''}(y_i, \phi)]^{-1}$.

Portanto, um estimador consistente para $Var(\hat{\phi})$ é $[\sum_{i=1}^{n} - c^{''}(y_i, \phi)]^{-1}$.

3.3.3. A Função Desvio

Existem diversas maneiras de se construir medidas de discrepância ou bondade de ajuste. Uma destas medidas denomina-se *desvio* e é equivalente à diferença de log-verossimilhanças maximizadas.

Seja o logaritmo da função de verossimilhança de $y = (y1; : : : ; yn)^T$, uma amostra aleatória com distribuição pertencente à família exponencial, expresso como função da média, isto é:

$$L(\mu; y) = \sum_{i=1}^{n} l(\mu_i; y_i) \quad (3.3.11)$$

onde $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = x_i^{'}\beta$. Considerando o número de componentes do vetor de

$$L(\mu; y) = \sum_{i=1}^{n} l(\mu_i; y_i),$$

parâmetros β (p) igual ao número de observações n, temos então o modelo saturado e a função $L(\mu; y)$ é estimada por:

Seja a estimativa de *L(μ;y)* dado por $L(\hat{\mu}; y)$, quando $p < n$. A estimativa de máxima verossimilhança de μ_i será dada por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, onde $\hat{\eta}_i = x_i^T \hat{\beta}$, sendo $\hat{\beta}$ o estimador de máxima verossimilhança de β. A função *desvio é* da seguinte forma:

$$D^*(y; \hat{\mu}) = \phi^{-1} D(y; \hat{\mu}) = 2\{L(y; y) - L(y; \hat{\mu})\} \quad (3.3.12)$$

que é a distância entre o logaritmo da função de verossimilhança do modelo saturado e do modelo sob investigação (modelo com p parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\beta}$. Um valor pequeno para a função desvio indica que para um número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste no modelo saturado. Sejam $\hat{\theta}_i = \theta_i(\hat{\mu})$ *e* $\hat{\theta}_i = \theta_i(\hat{\mu}_i^0)$ as estimativas dos parâmetros

canônicos para o modelo em investigação e o modelo saturado, respectivamente. Temos que a função $D(y;\hat{\mu})$ fica dada por

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\{y_i\,(\tilde{\theta}_i^0 - \hat{\theta}_i) + [b(\hat{\theta}_i) - b(\tilde{\theta}_i^0)]\},$$

que é chamada *função desvio* para o modelo corrente.

4. Aplicação

4.1. Dados Básicos

Para aplicar as técnicas propostas neste estudo, foram utilizados amostras de clientes correntistas e portadores de cartão de crédito de uma instituição financeira. Essas amostras são compostas de clientes que utilizaram à conta cartão nos últimos doze meses mesmo que estes tenham tido sua conta encerrada. As variáveis utilizadas no estudo são de informações cadastrais, canais utilizados pelos clientes, rentabilidade dos clientes no produto cartão de crédito, transações em conta cartão, informações do produto.

4.2. Elenco das Variáveis

As variáveis foram classificadas como dependentes e independentes, onde a primeira refere-se ao cancelamento ou não da conta cartão e a segunda aos fatores que influenciaram estes resultados. Foi apresentado um total de 52 variáveis das quais apenas as descritas abaixo foram utilizadas para iniciar este estudo, além disso, trabalhamos com quatro amostras de tamanho igual a 500 clientes para cada modelo, sendo uma para construção do modelo e as outras três de validação. Essas variáveis passaram por uma triagem onde foram excluídas a principio aquelas que tiveram uma autocorrelação com as variáveis independentes ou com variável dependente, depois foram verificadas as variáveis que não eram significativas na criação do modelo.

Quadro 1- Descrição das Variáveis.

Variável	Tipo	Definição	Categorização
Stat	Dependente	Status da Conta Cartão.	0 - Não Restrito 1 - Restrito
C_spnd	Independente	Média de Faturamento Anual na Conta Cartão.	Variável Continua
C_nmesrot	Independente	Número de meses que o cliente ficou no rotativo na Conta Cartão.	1 - 0 Meses 2 - 1 á 3 Meses 3 - 3 á 6 Meses 4 - 6 á 8 Meses 5 - > 8 meses
C_nmeses	Independente	Número de meses que o cliente utilizou o cartão no último Ano.	1 - até 3 meses 2 - 3 a 6 meses 3 - > 6 meses
Grupo	Independente	Grupo comportamental de Consumo ao qual o cliente pertence.	1- Supermercado 2- Lojas 3- Farmácia
Nc	Independente	Nível de classificação do cliente (nível do cliente de acordo com a rentabilidade).	1 - Nc_1 2 - Nc_2 3 – Nc_3
Protec	Independente	Possui-se ou não proteção contra roubo.	1 - Possui 2 - Não Possui
Dcc	Independente	Possui ou não debito automático na conta corrente para pagamento da conta cartão.	1 - Possui 2 - Não Possuem
Região	Independente	Número de transações feitas no último Ano.	Variável Continua
Ntrans	Independente	Numero de transações feitas nos últimos anos.	Variável Continua

Base Cartões Outubro/2005 á Outubro/2006

Embora a conta cartão possa ser encerrada pelo cliente ou pela instituição responsável pelo mesmo, neste estudo estamos levando em consideração apenas os clientes que

optaram ou não, ou seja, o nosso evento de interesse é o encerramento (inibição da conta cartão) por opção dos clientes.

4.3. Perfil Comportamental dos Dados

Para melhor entendimento dos dados torna-se importante uma breve analise descritiva dos mesmos. Nas tabelas abaixo os dados serão abordados de forma que possamos entendê-los e assim comparar a veracidade dos modelos propostos. A base total é composta por mais de 15 milhões de contas cartões da qual foram extraídas quatro amostras de forma a tratar a realidade.

Na tabela 1 podemos observar que quase a totalidade de nossa amostra é composta por clientes que possuem mais de 36 meses (três anos) de conta cartão, ou seja, são clientes que a principio deveriam estar fidelizados com a instituição financeira.

Tabela 1 – Idade da conta Corrente

Idade da Conta (Meses)	*Freqüência*	*Percentual*
1 - 1 A 6	66	3,3
2 - 7 A 12	70	4
3 - 13 A 18	40	2
4 - 19 A 24	26	1
5 - 25 A 30	49	2
6 - 31 A 36	27	1
7 - > q 36	1720	86
8 - missing	2	0
Total	*2000*	*100*

Base Cartões Outubro/2005 á Outubro/2006

Quando se fala em sexo, pode-se observar na tabela 2 que não existe uma diferença tão significativa apesar de termos à predominância masculina.

Tabela 2 – Sexo

Sexo	*Freqüência*	*Percentual*
Masculino	1106	55
Feminino	892	44,6
Missing	2	0
Total	*2000*	*100*

Base Cartões Outubro/2005 á Outubro/2006

Em relação à tabela 3 podemos verificar que a maioria dos clientes está bem distribuída a partir dos 36 anos, o que nos índica a idade em que o cliente já possue maturidade financeira.

Tabela 3 – Faixas e Idade

Fx_idade	*Freqüência*	*Percentual*
18 A 25	18	0,9
26 A 30	135	6,75
31 A 35	151	7,55
36 A 40	232	11,6
41 A 45	365	18,25
46 A 50	347	17,35
51 A 55	321	16,05
56 A 60	253	12,65
> =61	176	8,8
Missing	2	0,1
Total	*2000*	*100*

Base Cartões Outubro/2005 á Outubro/2006

Quando tratamos de escolaridade podemos observar que a maioria dos clientes de nossa amostra possuem segundo grau completo, seguido de nível superior.

Tabela 4 – Nível Escolar

Escolar	*Freqüência*	*Percentual*
Analfabetos	87	4
1º Grau	476	23,8
2º Grau	1.001	56
Nível Superior	340	21
Pós-Graduado	85	4
Missing	2	0
Total	*2000*	*100*

Base Cartões Outubro/2005 á Outubro/2006

Quando levamos em consideração o estado civil dos clientes portadores de conta cartão pode observar na tabela 5 que a maioria dos clientes são casados.

Tabela 5 – Estada Civil

Estado Civil	*Freqüência*	*Percentual*
Solteiro	*362*	*18,1*
Casado	*1.183*	*59*
Separado	*288*	*14*
Viúvo	*164*	*8*
Missing	*3*	*0,1*
Total	*2000*	*100*

Base Cartões Outubro/2005 á Outubro/2006

5. Resultados da Análise de Sobrevivência

Para identificar quais covariáveis dentre as pesquisadas influenciam no tempo de cancelamento de conta cartão foi utilizado o modelo de regressão de Cox. O procedimento utilizado foi o "STEPWISE", ou seja, iniciou-se com o modelo com todas as variáveis até chegar ao modelo final onde todas foram significativas. Para fazer nosso modelo usamos a modelagem de Cox estratificada, já que existe toda uma teoria em relação aos clientes dependendo do seu nível de relacionamento com a instituição.

Tabela 6 – Resultados do Primeiro Ajuste Coeficientes estimados pelo Modelo de COX para evasão de conta cartão.

Variável	β	Erro padrão	Valor-P	Exp (β)
C_Spnd	0, 012	0, 026	0, 001	1, 012
Grupo	0, 488	0, 087	0, 215	1, 629
C_mesrot	1, 673	1, 451	0, 153	1, 104
C_mesrot	-0, 839	0, 240	0, 186	0, 1029
C_mesrot	1, 972	0, 216	0, 983	1, 740
C_mesrot	-0, 148	0,143	0, 722	0,791
Protec	-0, 0005	0, 00006	0, 437	0, 998
Dcc	-0, 0704	0, 2242	0, 757	0, 932
Região	-0, 192	0, 016	0, 0128	0, 0236
Ntrans	0, 1717	0, 127	0, 185	1, 187

Base Cartões Outubro/2005 á Outubro/2006

Observa-se no primeiro ajuste que as variáveis que se apresentaram não significantes foram às relacionadas ao numero de meses que os clientes ficaram no rotativo e se possuem ou não debito em conta devendo então as mesmas ser retiradas do modelo.

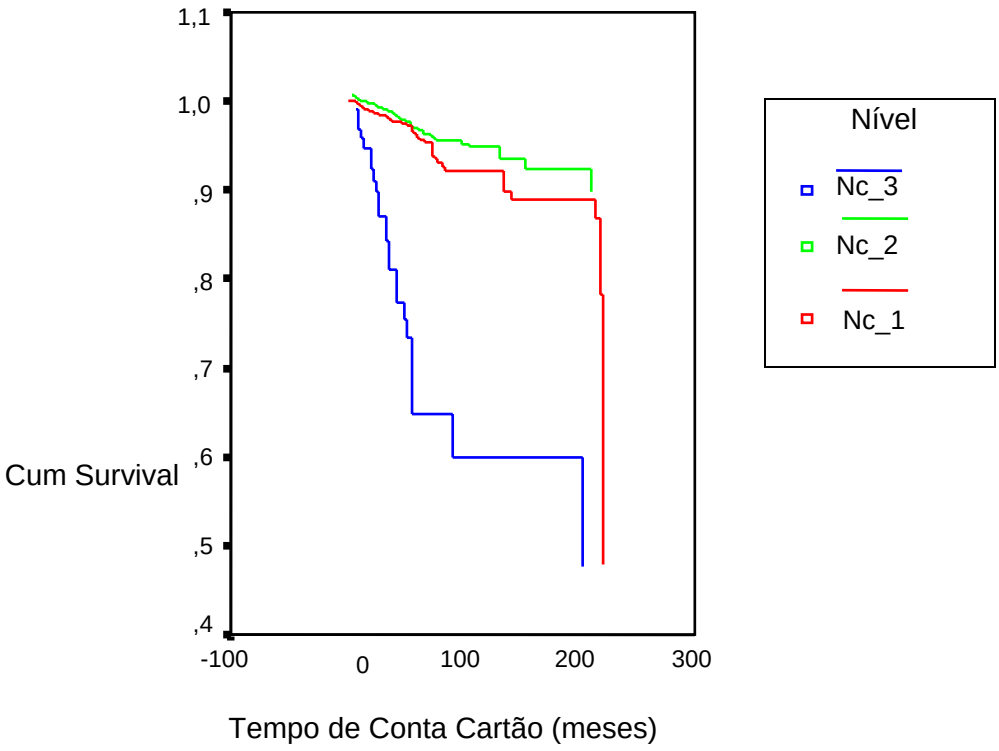
Tabela 7 – Resultados do Ajuste Final dos Coeficientes estimados pelo Modelo de COX para evasão de conta cartão.

Variável	β	Erro padrão	GL	Valor-P	Exp(β)	95% I.C. para Exp(β)	
						Inferior	Superior
C_SPND	0, 079	0, 010	1	0, 000	1, 082	1, 063	1, 101
REGIÃO	-,068	0, 017	1	0, 000	, 934	, 903	, 967
NTRANS	-,016	0, 002	1	0, 000	, 984	, 980	, 988

Base Cartões Outubro/2005 á Outubro/2006

Os resultados do modelo de regressão de Cox obtido após 08 ajustes estão apresentados na tabela acima onde nos verificamos que o gasto anual dos clientes (C_Spnd), região geográfica que estão localizados (Região) e o numero de transações feitas durante o período de um ano (Ntrans) foram identificadas como fatores influentes no tempo de cancelamento. Pode-se verificar que a variável C_spnd nos diz que um cliente que está diminuindo seu faturamento anual possui 1,082 vezes mais chances de cancelar a conta cartão, ou seja, ele tem 8,2% a mais de chance de cancelar a conta cartão. Além disso, o cancelamento está relacionado ao número de transações e à região geográfica, donde podemos concluir que ao diminuir o número de transações e dependendo de fatores ligados à região resultam em risco de cancelamento.

Figura 1: Gráfico da Função de Sobrevivência da Idade das contas cartão por estrato econômico.



Podemos observar que no gráfico 1, que por questão de análise dividimos em estratos já que para oferta de cartão levamos em consideração características destes três Grupos, onde o Nc_1 falamos dos clientes alta renda, o Nc_2 clientes renda média e Nc_3 é o que apresenta população com menor renda porém maior nível de heterogeneidade. Pode-se observar que clientes que possuem baixa renda até o quinto ano de conta cartão possuem um altíssimo risco de cancelamento, o que realmente vai baixar significativamente próximo a dez anos de conta, ou seja, se estes clientes achegarem até o décimo ano com conta cartão ativa, provavelmente não irá encerrá-la mais.

O cliente alto e média renda têm um comportamento bem parecido, já que estes são grupos com algum poder aquisitivo, renda fixa e maior escolaridade. O risco de

cancelamento é bem inferior ao da baixa renda e a partir do quinto ano esse risco torna-se estável.

Podemos concluir com isso que é mais fácil fidelizar cliente que possuem alguma estabilidade, o que seria publico alvo de ações.

Quadro 1 - Testes para Avaliar o Ajuste do Modelo.

Teste	Razão de Verossimilhança	Wald	Escore
P-Valor	0, 0048	0, 0077	0, 0063
Valor	16,75	15,45	13,09

Base Cartões Outubro/2005 á Outubro/2006

No quadro estão apresentados os testes estatísticos utilizados para avaliação do modelo ajustado. De acordo com os três testes aplicados pode-se verificar que o modelo ajustado foi significativo com 95% de confiança, ou seja, o modelo obtido explica bem os dados.

6. Resultados da Regressão Logística

Para o Analise da Regressão logística é importante observarmos que as variáveis foram todas categorizadas, o que não ocorreu no modelo de analise de sobrevivência onde algumas delas entraram como variável continua. Podemos observar neste primeiro ajuste que as variáveis precisam ser revistas já que em alguns momentos um determinado grupo tem o p-valor significativo em alguma das categorizações. O intercepto e as variáveis C_nmeses e Dcc são propensas a serem revistas ou então retiradas do modelo. A princípio pode observar claramente que o intercepto e as variáveis C_nmeses e Dcc são propensas a serem revistas ou então retiradas do modelo.

Quadro 3 – Parâmetros Estimados utilizados no primeiro ajuste da Regressão Logística.

Parâmetros	Estimadores	Erro Padronizado	Qui-quadrado de Wald	p-valor
Intercepto	1, 1190	23,46	0, 0023	0, 09620
C_spnd1	0, 2454	0, 0452	29, 429	< 0, 0001
C_spnd2	0, 2363	0, 0633	13, 922	< 0, 0001
C_nmesrot1	- 0, 0902	0, 0653	1, 9087	< 0, 0001
C_nmeses1	-6, 0847	140,80	0, 0019	< 0, 0001
C_nmeses2	- 0, 8797	23, 466	0, 0025	0, 9618
C_nmeses3	- 1, 1232	23, 466	0, 0025	0, 9618
Grupo4	0, 1759	0, 0412	18, 2077	< 0, 0001
Região	- 0, 2440	0, 0432	31, 8224	< 0, 0001
Protec1	1, 8222	0, 3712	24, 098	< 0, 0001
Dcc	0, 0178	0, 0629	0, 0812	0, 0057
Ntrans	- 0, 0809	0, 0634	1, 6289	0, 0219

Base Cartões Outubro/2005 á Outubro/2006

Quadro 4 - Testes de Aderência do Modelo Logístico primeiro Ajuste.

Teste	Kolmogorov-Smirnov (KS)	Cox & Snell R Square	Nelgelkerke R Square
Valor	0, 019	0, 327	0, 463

Base Cartões Outubro/2005 á Outubro/2006

No quadro 4 podemos observar que de acordo com os testes de aderência o modelo encontrado não condiz com a realidade dos dados o que nos fará necessários novos ajustes a fim de encontrarmos um modelo de alta consistência.

Além disso, é necessário o entendimento de que existem diferenças dos valores dos testes por se basearem em metodologias diferentes, como os dados utilizados são censurados o teste mais consistente é o Nagelkerke.

Quadro 5 – Parâmetros Estimados utilizados Modelo Logístico Final.

Parâmetros	Estimadores	Erro Padronizado	Qui-quadrado de Wald	p-valor
Intercepto	0, 1226	0, 049	18, 049	< 0, 0001
C_spnd1	0, 0782	0, 0981	0, 4497	< 0, 0001
C_spnd2	-0, 2692	0, 0682	15, 3029	< 0, 0001
C_nmesrot1	0, 2624	0, 0466	32, 2693	< 0, 0001
C_nmeses1	0, 7215	0, 0125	35, 3678	< 0, 0001
C_nmeses2	0, 7496	0, 0696	115, 0037	< 0, 0001
C_nmeses3	-0, 1833	0, 0872	4, 3684	< 0, 0001
Grupo1	0, 5081	0, 0922	17, 2837	< 0, 0001
Grupo2	-0, 3028	0, 0684	19, 5963	< 0, 0001
Grupo3	0, 0121	0, 0941	0, 0165	< 0, 0001
Grupo4	-0, 3632	0, 1015	12, 7934	< 0, 0001
Nc1	0, 7668	0, 0614	156, 0391	< 0, 0001
Nc2	-0, 5294	0, 0612	74, 7063	< 0, 0001
Protec1	-0, 3642	0, 0436	69, 7789	< 0, 0001
dcc	0, 1722	0, 0416	17, 0979	< 0, 0001

Base Cartões Outubro/2005 á Outubro/2006

Podemos observar no Quadro 5 que a probabilidade de encerramento de conta cartão decresce acentuadamente á medida que o fator c_spnd aumenta, ou seja, o individuo da categoria c_spnd2 tem 3,6 mais chances de cancelar que o c_spnd1. Clientes que têm maior utilização do cartão (c_nmeses3), que pertencem ao grupo de alto consumo e supermercado e que possuem débito em conta e proteção ouro, têm uma baixa probabilidade de encerrar a conta cartão. Além disso, observou-se que quanto menor o segmento de classificação do cliente, maior a probabilidade de encerramento da conta cartão.

Quadro 6 - Testes de Aderência do Modelo Logístico Final.

Teste	Kolmogorov-Smirnov (KS)	Cox & Snell R Square	Nelgelkerke R Square
Valor	0, 385	0, 651	0, 781

Base Cartões Outubro/2005 á Outubro/2006

No Quadro 6 podemos observar, através dos testes de aderência, se as métricas descritas no modelo realmente foram significativas e se o modelo é valido. Podemos observar que o teste KS nos dá um escore de 38,5% o que significa uma ótima aderência do nosso modelo. Também temos os testes de Cox e Nagelkerke, que têm o intuito de avaliar o ajuste geral do modelo. Estes testes vêm comprovar que nosso modelo obteve um bom ajuste principalmente porque o de Nagelkerke variou de 0, 463 no primeiro STEP para 0, 781 no último STEP, ou seja, conforme fomos ajustando as variáveis, o modelo foi melhorando.

Na tabela cinco fizemos um quadro comparativo com as duas técnicas apresentadas a fim de poder com precisão inserir na instituição formas praticas de abordagem aos clientes e escolher quais seriam publico alvo de nossa ação retentora.

Tabela 8 - Critério para Retenção de Contas Cartão

Segmento de clientes	Sobrevida do cliente	Regressão Logistica Risco de Saída		
	Períodos	Alto	Médio	Baixo
Alto Valor	Até 5	1	2	5
	5 a 10	3	4	6
	>10	7	9	10
Médio Valor	Até 5	8	11	12
	5 a 10	13	15	16
	>10	14	18	19
Baixo Valor	até 10	17	20	21
	10 á 15	22	23	24
	>15	25	26	27

Base Cartões Outubro/2005 á Outubro/2006

No quadro acima podemos de forma clara e objetiva observar que os clientes que devemos incidir maiores campanhas são os de alto valor visto que o mesmo por terem uma serie de características, entre elas estabilidade financeira e conhecimento dos produtos oferecidos.

7. Conclusões e Recomendações.

A análise de sobrevivência, modelo de riscos proporcionais de COX, demonstrou ser uma técnica muito simples de ser usada, bem como uma poderosa ferramenta de predição de sobrevida da conta cartão. Através desta técnica podemos observar que, os clientes de alto valor possuem menor probabilidade de encerramento da conta cartão devido ao seu histórico financeiro, além disso, quanto maior o gasto anual e transações feitas com a conta cartão menor chance o cliente tem de encerramento. Outro fator importante diz respeito à região geográfica a qual o cliente pertence, pois disso também vai depender o aumento ou não da probabilidade de encerramento da conta cartão.

Com o modelo de regressão logística, podemos quantificar as variáveis que também influenciam no encerramento da conta cartão. A variável referente aos gastos anuais feitos pelos clientes deve ser especialmente analisada já que se apresentas em ambos os modelos propostos. Neste modelo podemos observar que devemos levar em consideração as variáveis referentes ao numero de meses que o cliente utilizou o cartão, numero de meses que o cliente entrou no rotativo, o grupo comportamental do mesmo, o nível do cliente, se possue ou não proteção contra roubo e debito em conta corrente.

A análise de regressão logística demonstrou robustez na modelagem, alcançando excelentes resultados na predição do risco de cancelamento de conta cartão pelos clientes.

Cruzando as duas técnicas se pode prever, com certa antecedência, os clientes propensos a evasão, ou seja, propensos ao cancelamento da conta cartão, e atuar com eles no momento adequado com ações de marketing de relacionamento. Estes resultados trazem uma grande economia para a instituição, pois ao invés de fazer ações de marketing de

fidelização para toda a sua base, as ações seriam focadas nos clientes mais propensos ao cancelamento e no momento certo.

Referências Bibliográficas

[1] Alisson. P. D. (2001). Survival Analysis Using The SAS®: A Pratical Guide. SAS Institute Inc., Books by Users, SAS Campus Drive, Cary, NC 27513.

[2] Agresti, A. (1990). Categorical Data Analysis. New York : John Wiley & Sons.

[4] Colosimo, E. A.e Vieira. A. M. C. (1996). O Modelo de Regressão de Cox com Covariável Dependente do Tempo: Uma Aplicação Envolvendo Pacientes Infectados pelo HIV. R. Brás. Estat., Rio de Janeiro, v.54/57, n. 201/208, p.139-152.

[5] Colosimo, E. A. (2001). Análise de Sobrevivência Aplicada, 46ª Reunial Anual RBRAS, 9º SEAGRO, ESALQ/USP, Piracicaba, SP.

[6] Colosimo, E. A. e Giolo, S. R. (2006). Análise de Sobrevivência Aplicada. Editora Edgard Blucher.

[7] Cox, D. R. (1975). Partial Likelihood . Biometrika, 62, 269-276.

[8] Hosmer, D. W. e Lemeshow, S. (1989). Applied Logistic Regression. New York : John Wiley & Sons.

[9] Hosmer, D. W. e Lemeshow. S. (1998). Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley Series in Probability and Statistics.

[10] Neter, J.; Kutner, M.H.; Natchschiem, C.J.; Wasserman, W. (1996). Applied Linear Statistical Models. Irwin, Boston.

[11] Aalen, O.O.(1976). Noparametric Inference in Connection with Múltiple Decrement Models. Scandinavian Journal Statistics, 3, 15-27.

[12] Aalen, O.O.(1976). Noparametric Inference in Connection with Múltiple Decrement Models. Scandinavian Journal Statistics, 3, 15-27.

[13] Andersen, P.K (1982). Testing Goodness of Fit for Cox's Regression and Life Model. Biometrics, 38, 67-77.

[14] Andersen, P.K & Gill, R (1982). Cox's Regression Model for a Counting Processes: A large Sample Study. Ann. Statistics, 10,1100-1200.

[15] Arjas. E (1988). A Graphical Method for assenssing goodness of fit in Coxs Proportional Hazards Model . Journal of the American Statistical Association, 83,204-212.

[16] Barlow, W. & Prentice, R.E(1988). Residuals for Relative Risk Regression 75, 64-74.

[17] Breslow, N. & Crowley, J. (1974). A large Swple Study of the Life Table and Product Limit Estimates Under Random Ceusorship. Annals of Statistics, 2, 437-453.

[18] Colosimo, E.A., Ferreira, F.F., Oliveira M.D & Souza, C.B (2002) Empirical Comparassions Between Kaplan – Méier and Nelson – Aalen Survival Function. Journal Stat. Sim., 72, 299-308.

[19] Cox, D.R (1972). Regression Models and Life Tables (with discussion). Journal Royal Statistics Soc. B, 34, 187-220.

[20] Cox, D.R (1975). Partial Likelihood , Biometrika, 62, 269-276.

[21] Cox, D.R (1975). A Note on the Graphical Analysis of Survival Data. Biometrika, 66, 188-190.

[22] Cribari-Neto, F. & Zarkos, S.G (1999). Econometric Programming Environment. Journal Appl. Econ., 14,319-329.

[23] Cutler, S.J & Ederer , F.(1958).Maximum of Utilization of the Life table Method in Analysing Survival. Journal of Chronic Diseases, 8, 699-712.

[24] Efron, B. (1967). The Efficieney of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association , 72, 557-565.

[25] Fleming, T.R. & Harrington, D.P (1991). Counting Processes and Survival Analysis. Jonh Wiley , New York.

[26] Lawless, J.F (1982). Statistical Methods for Lifetime Data. John Wiley e Sons, New York.

[27] Man, J. (1986). On a Graphical method for the detection of time- dependent effects of covariates in survival data Applied Statistics: 35, 245-255.

[28] Man, J. (1988). A comparison of counting process models for complicated life histories. Applied Stochastic Models and Analysis, 4, 283-298.

[29] McKeague, I. W (1986). Estimation for a semimartingale regression modelo using the method of sieves. Annals of Statistics, 14, 579-589.

[30] Peto, R.(1972). Contribuio a Discusso do Artigo de D.R Cox. Journal of the Royal Statistical Society B, 34, 205-207.

[31] Amemiya, T. (1985). Advanced Econometrics. Cambridge, MA: Harvard University Press.

[32] Avriel, M. (1976). Nonlinear Programming: Analysis and Methods. Englewood Cliffs, NJ: Prentice-Hall.

[33] Aitkin, M., Anderson, D., Francis, B. e Hinde, J. (1990). Statistical Modelling in GLIM. Oxford: Clarendom Press.

[34] Barlow, R. E.; Bartholomew,D. J.; Bremmer, J. N. e Brunk, H. H. (1972). Statistical Inference under Order Restrictions. New York: John Wiley.

[35] Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives, I. Biometrika 46, 36-48.

[36] Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives, II. Biometrika 46, 328-335.

[37] Bartholomew, D. J. (1961). A test of homogeneity of means under restricted alternatives. Journal of the Royal Statistical Society B 23,239-281.

[38] Bohrer, R. e Chow, W. (1979). Algorithm AS122. Weights for one-sided multivariate inference. Applied Statistics 27, 100-104.

[39] Breslow, N.E., Lubin, J. H., Marek, P. e Langholz, B.(1983). Multiplicative models and cohort analysis. Journal of the American Statistical Association 78, 1-12.

[40] Chambers, J. H. e Hastie, J. T. (1992). Statistical Models in S. California : Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove.

[41] Childs, D. P. (1967). Reduction of the multivariate normal integral to characteristic form. Biometrika 54, 293-300.

[42] Collet, D. (1994). Modelling Binary Data. London: Chapman and Hall.

[43] Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics. REFERˆENCIAS 97 Biometrika 74, 265-274.

[44] Cordeiro, G. M. e McCullagh, P. (1991). Bias correction in generalized linear models. Journal of the Royal Statistical Society B 53, 629-643.

[45] Cook, R. D. e Weisberg, S. (1982). Residuals e Influence in Regression. New York: Chapman and Hall.

[46] Cox, D. R. e Hinkley, D. V. (1974). Theoretical Statistics. London: Chapman and Hall.

[47] Dachs, J. N. W. e Paula, G. A. (1988). Testing for ordered ratio rates in followup studies with incidency density data. Revista Brasileira de Probabilidade e Estatística 2, 125-137.

[48] Fahrmeir, L. e Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalizad linear models. Annals of Statistics 13, 342-368.

[49] Fahrmeir, L. e Klinger, J. (1994). Estimating and testing generalized linear models under inequality restrictions. Statistical Papers 35, 211-229.

[50] Fiacco, A. V. e McCormick, G. P. (1968). Nonlinear Programming : Sequential Unconstrained Minimization Techniques, New York : Wiley

[51] Finney, D. J. (1971). Probit Analysis, Third Edition. Cambridge: Cambridge University Press.

[52] Finney, D. J. (1978). Statistical Methods in Biological Assay, Third Edition. London: Griffin.

[53] Gill, P. E; Murray, W. e Wright, M. H. (1981). Practical Optimization. New York: Academic Press.

[54] Gouriéroux, C.; Holly, A. e Monford, A. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. Econometrica 50, 63-80.