Dalila de Moraes

MODELAGEM DE FRAUDE EM CARTÃO DE CRÉDITO

São Carlos

Agosto de 2008

Universidade Federal de São Carlos
Departamento de Estatística

Dalila de Moraes

Modelagem de Fraude em Cartão de Crédito

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Carlos Alberto Diniz

São Carlos
Agosto de 2008

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M827mf

Moraes, Dalila de.
Modelagem de fraude em cartão de crédito / Dalila de Moraes. -- São Carlos : UFSCar, 2012.
120 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Estatística. 2. Modelagem de dados. 3. Regressão logística. 4. Modelo logito limitado. 5. Amostras state-dependent. I. Título.

CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 260-8292/260-8241 - FAX: (016) 260-8243
13565-905 - SÃO CARLOS-SP-BRASIL

ATA DO EXAME DE DISSERTAÇÃO DE MESTRADO DA CANDIDATA:
Dalila de Moraes

Aos dois dias do mês de setembro do ano de dois mil e oito, às quatorze horas, na Sala de Reuniões do Departamento de Estatística, reuniu-se a Comissão Examinadora nas formas e termos do Artigo 25º do Regimento Interno do Programa de Pós-Graduação em Estatística da UFSCar, composta pelos membros: Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar, Orientador), Prof. Dr. Adriano Polpo de Campos (DEs-UFSCar) e Profa. Dra. Mariana Cúri (ICMC-USP), para Exame de Dissertação de Mestrado da candidata Dalila de Moraes, sob o título "Modelagem de fraude em cartão de crédito". A sessão foi aberta pelo Prof. Dr. Carlos Alberto Ribeiro Diniz (Presidente), iniciando-se pela apresentação da dissertação. Em seguida, foi feita a argüição da candidata pelos membros da Comissão Examinadora. A Comissão Examinadora considerou o tema relevante para Estatística e julgou a exposição feita pela candidata clara e objetiva. A candidata respondeu satisfatoriamente as questões formuladas. Pelo apresentado acima, a comissão atribuiu as seguintes avaliações: Prof. Dr. Carlos Alberto Ribeiro Diniz, Aprovada; Prof. Dr. Adriano Polpo de Campos, Aprovada; e Profa. Dra. Mariana Cúri, Aprovada. De acordo com o parágrafo 5º do Artigo 25º, a candidata foi considerada **aprovada**. Encerrada a sessão secreta, o Presidente informou o resultado da defesa. Nada mais havendo a tratar, eu, Maria Isabel Rinaldo Pessôa de Araujo, Secretária deste Programa, lavrei a presente ata, que assino juntamente com os membros da Banca Examinadora.

Prof. Dr. Carlos Alberto Ribeiro Diniz

Prof. Dr. Adriano Polpo de Campos

Maria Isabel R. P. Araujo

Profa. Dra. Mariana Cúri

Resumo

O aumento no volume de transações com cartões de crédito trouxe como consequência o aumento do número de fraudes, o que acarreta em uma perda de bilhões de reais anualmente à todas instituições financeiras do mundo. Com isso é muito importante que metodologias de detecção e prevenção à fraude sejam desenvolvidas. A grande dificuldade na modelagem deste tipo de dados é que estes são extremamentes desbalanceados. Neste presente trabalho, será proposto o modelo logito limitado na detecção de fraude. Também será discutido as amostras do tipo *state-dependent* e comparado os desempenhos dos modelos logito e logito limitado. Duas aplicações, uma com um conjunto de dados simulados e outra com um conjunto de dados reais, serão apresentadas. A abordagem bayesiana para estes modelos também será discutida. As análises dos conjuntos de dados serão realizadas nos softwares SAS e Winbugs.

Palavras-chave: modelo logito, modelo logito limitado, amostras do tipo *state-dependent*, análise Bayesiana.

Abstract

The transactions volume increase brought the fraud increase, which result in a annual loss of billions of reais to all financial institutions in the world. Therefore, it's very important the development of detection methods and fraud prevention. The difficult in modeling this kind of data due the fact the data sets are extremely unbalanced. In this work, a bounded logit model will be proposed for fraud detection. It will also be discussed state-dependent sampling and compared with logit and bounded logit model performances. Two applications, one with a simulated data set and another with a real data set, will be presented. The Bayesian approach to these models will also be discussed. The data set analyses will be implemented in SAS and Winbugs software.

Keywords: logit model, bounded logit model, state-dependent sampling, Bayesian Analysis.

Sumário

Capítulo 1

Introdução

Atualmente a quantidade de pessoas que utilizam seus cartões de crédito como meio de pagamento de suas compras é muito significante. O aumento do número de transações com cartões realizadas diariamente, acarreta no aumento do risco que a instituição tem de que uma parte destas sejam fraudes.

Nos Estados Unidos a perda com transações fraudulentas de cartões gira em torno de 850 milhões de dólares por ano (GHOSH e REILLY, 1994). (BHATLA, PRABHU e DUA, 2003) citam que em 2002 as transações fraudulentas realizadas com cartões de crédito no mundo somaram mais de 2,5 bilhões de dólares e a projeção para o final do ano de 2008 é de 5 a 15 bilhões de dólares. No Brasil, de cada dez mil transações cinco são fraudulentas (GADI, 2006).

Um problema encontrado na detecção de fraude é a estrutura dos bancos de dados. Em geral, dentre todas as transações realizadas, a porcentagem que são fraudes é bem pequena. Para este tipo de conjuntos de dados, (CRAMER, 2004) sugere o uso da regressão logística limitada. Dessa forma, a presente dissertação tem como objetivo principal o estudo deste modelo e a comparação deste com o modelo de regressão logística usual. Também será verificado através das amostras do tipo *state-dependent* o impacto de se utilizar uma amostra balanceada quando se tem baixa incedência do evento de interesse.

Além da abordagem Clássica, serão estudadas as estimativas via abordagem Bayesiana. Nesta abordagem, os parâmetros são considerados quantidades aleatórias (BOX & TIAO, 1992; PAULINO et. al., 2003), ou seja, seguem distribuições de probabilidade. O objetivo da metodologia Bayesiana é obter medidas resumo ou densidades *a posteriori* para os pa-

râmetros de interesse. Estas densidades são resultantes da combinação de informações da amostra (função de verossimilhança) e de informações *a priori* sobre os parâmetros (densidades *a priori*). A grande vantagem da inferência Bayesiana é a possibilidade da incorporação de informações adicionais provenientes de especialistas (DE FINETTI et. al., 1974).

A estrutura deste trabalho está dividida da seguinte forma: O Capítulo 2 é dedicado ao conhecimento do ambiente de estudo, ou seja, o cartão de crédito. O Capítulo 3 traz informações sobre fraude. O Capítulo 4 apresenta alguns métodos já utilizados para a detecção e prevenção à fraude. No Capítulo 5, será discutido o modelo de regressão logística. O modelo de regressão logística limitada será apresentado no Capítulo 6. O Capítulo 7 mostra como estimar os parâmetros de regressão logística e logística limitada para amostras do tipo *state-dependent*. No Capítulo 8 será apresentada a análise de dois conjuntos de dados. Um conjunto de dados simulado e outro de dados de fraude real cedidos por uma grande instituição financeira. No Capítulo 9 será discutido a abordagem bayesiana para os dois modelos em estudo. Por fim, o Capítulo 10 apresenta as conclusões.

Capítulo 2

Cartão de Crédito

O aumento da criminalidade e o avanço da tecnologia faz com que o número de transações fraudulentas em cartões de crédito aumente a cada ano, gerando uma grande perda de bilhões de reais às instituições financeiras. Dessa forma, é de extrema importância o desenvolvimeto de metodologias de detecção e de modelagem estatística de fraude para evitar tais perdas (HAND, ADAMS e BOLTON, 2002). Porém, antes de estudar estas metodologias, será apresentado o ambiente do estudo em questão que é o cartão de crédito.

2.1 Histórico

O número de cartões de crédito vem crescendo substancialmente nesses últimos tempos, substituindo o uso de cheques, cédulas e moedas. Instituições financeiras, bancos e um crescente número de lojas oferecem a seus clientes cartões que podem ser usados na compra de um grande número de bens e serviços, inclusive em lojas virtuais. Os cartões não são dinheiro real, simplesmente registram a intenção de pagamento do consumidor, mediante sua assinatura e demais verificações, que em determinada data terá de pagar as despesas com o cartão, em débito automático, espécie ou ainda em cheque (Banco Central do Brasil, 2008). Desta forma, o cartão é uma forma imediata de crédito.

Os cartões de crédito nasceram nos Estados Unidos da América na década de 1920, quando empresas privadas, redes de hotéis e empresas petroleiras começaram a emitir cartões para permitir que seus clientes comprassem a crédito nos próprios estabelecimentos (Banco Central do Brasil, 2008).

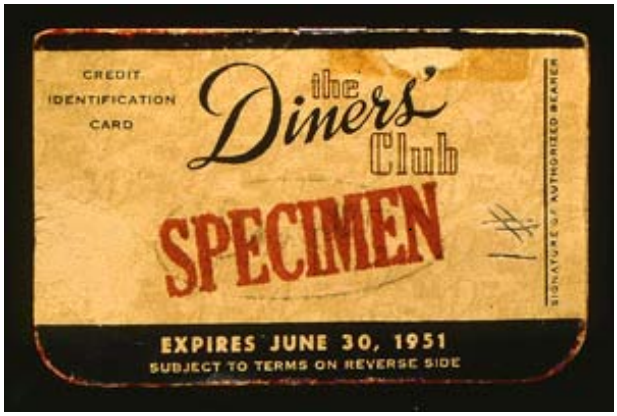


Figura 2-1: Primeira versão em papel do cartão Diners Club Card em 1950.

Em 1950, Frank MacNamara e alguns executivos financeiros de Nova York saíram para jantar e só quando receberam a conta perceberam que não tinham como pagar, pois haviam esquecido o dinheiro e o talão de cheque. Depois de alguma discussão, o dono do restaurante permitiu que MacNamara pagasse a conta em outro dia, mediante a sua assinatura na nota de despesas. Depois desse episódio, o executivo concebeu a idéia do cartão de crédito e criou o primeiro cartão de crédito. Esse cartão foi denominado de Diners Club Card, que passou a ser aceito como meio de pagamento em vinte e sete restaurantes daquele país e usado por importantes executivos, como uma maneira prática de pagar suas despesas de viagens a trabalho e de lazer.

A Figura 2-1 apresenta a primeira versão em papel cartão do Diners Club Card. Essa versão trazia o nome do associado de um lado e dos estabelecimentos filiados no outro. Com este sistema a empresa de cartões de crédito cobrava uma taxa anual e enviava contas mensais ou anuais dos gastos efetuados. Em 1955 o Diners Club trocou o material de sua confecção para o plástico.

Em 1951, também nos Estados Unidos da América, foi desenvolvido os sistemas de cartões de crédito bancários. O primeiro banco a utilizar este sistema foi o Franklin National Bank (em Nova York) introduzindo o primeiro verdadeiro cartão de crédito bancário. Com este sistema, os bancos creditavam à conta do comerciante assim que recebessem os comprovantes assinados pelos clientes e cobravam destes em uma única conta mensal, acrescentando juros e outros custos.

O primeiro sistema a operar em todos os estados dos Estados Unidos foi o BankAmeri-

card do Bank of America. Este sistema iniciou suas atividades operando apenas no estado da Califórnia, em 1959 e passou a operar nos demais estados americanos em 1966. Em 1976 o BankAmericard mudou de nome e passou a se chamar Visa.

Em 1966 um grupo de bancos americanos formou o Interbank Card Association (ICA) que mais tarde passou a se chamar MasterCard International.

O Brasil teve como percursor dos cartões de crédito o empresário tcheco Hanus Tauber, que comprou nos Estados Unidos, em 1954, a franquia do Diners Club, propondo sociedade no cartão com o empresário Horácio Klabin. Em 1956 foi então lançado no Brasil o cartão Diners Club.

Em 1968 foi lançado no Brasil o primeiro cartão de crédito de banco nacional, intitulado como ELO e criado pelo Banco Bradesco. Em 1983 foi lançado o primeiro cartão de débito e em 1984 a Credicard comprou o Diners Club no Brasil.

Atualmente a situação dos principais cartões de crédito é a seguinte:

- **Visa**: é uma associação de 21.000 instituições financeiras no mundo todo, que emitem o cartão com a bandeira Visa. Existem 1,3 bilhões de cartões Visa em circulação que são aceitos em mais de 24 milhões de estabelecimentos em mais de 150 países. No ano de 2005 o volume de transações gerado por estes cartões foi de 3 trilhões de dólares.

- **Mastercard:** tem mais de 25.000 parceiros emissores no mundo. Existem cerca de 720 milhões de cartões Mastercard em circulação, aceitos em 32 milhões de estabelecimentos comerciais e em mais de 210 países. No ano de 2005 o volume de transações gerado por cartões da Mastercard foi de aproximadamente 1,2 trilhões de dólares.

- **American Express:** é uma instituição fundada em 1850, mesmo com o primeiro cartão emitido somente em 1958. Os cerca de 57 milhões de cartões Amex em circulação são aceitos em mais de 200 países. No ano de 2005 os cartões Amex geraram em transações cerca de 150 bilhões de dólares.

Outras bandeiras de cartões difundidas no mundo, com maior ou menor concentração, dependendo da região, são: **Diners, JCB (Japanese Credit Bureau), Discover** e

Solo. No Brasil existem ainda algumas bandeiras nacionais com boa divulgação, dentre as quais podemos citar o **HiperCard** (Unibanco), o **Cartão Aura** (Grupo BNP Paribas) e o **IBI**.

2.2 Tipos de Cartão de Crédito

A maioria dos cartões de crédito encontrados no mercado são do tipo:

- **Nacional:** possui operação apenas no seu país de origem.
- **Internacional:** funcionamento nacional e internacional com serviços restritos.
- **Ouro**: possui funcionamento nacional e internacional, com grandes vantagens em viagens e programas de recompensas, o que acarreta anuidade superior ao Internacional.
- **Platina**: Possui funcionamento nacional e internacional, sendo o tipo de cartão com o maior número de vantagens se comparados com os demais. Para possuir este tipo de cartão o cliente tem que possuir uma renda mensal alta.

2.3 Agentes do Cartão de Crédito

Atualmente os cartões de crédito, possuem cinco agentes envolvidos em seu funcionamento: portador (*card holder*), estabelecimento (*merchant*), adquirente (*acquire*), bandeira (*brand*) e emissor (*issuer*). Abaixo segue uma breve descrição destes agentes (GADI, 2006):

- **Portador:** indivíduo que possui o cartão, sendo este o responsável por iniciar o funcionamento do sistema ao decidir pagar suas compras com seu cartão de crédito.
- **Estabelecimento:** qualquer empresa ou pessoa jurídica credenciada a aceitar o cartão de crédito por meio de um equipamento específico.
- **Adquirente:** a função da empresa adquirente é de credenciar, supervisionar e repassar os valores de compra aos estabelecimentos que aceitam o cartão de crédito,

além de ser responsável pela implementação e manuntenção das maquinetas, denominadas POS (point of sales), e dos softwares de captura de transações.

- **Bandeira:** responsáveis pelas definições das regras de política (relacionamento entre emissores e adquirentes), operações da rede global de comunicações, execuções de markenting institucional e pesquisas e desenvolvimentos de novas tecnologias e serviços. Os principais responsáveis pelas receitas das Bandeiras são as multas aplicadas aos clientes devido ao não cumprimento de regras e prazos e a tarifa para tráfego de documentos em papel. São exemplos de bandeiras: Visa, MasterCard, Amex e Hipercard.

- **Emissor:** em geral, os emissores são bancos, responsáveis pela distribuição dos cartões de crédito aos seus clientes mediante a aprovação do risco de crédito por políticas próprias de cada instituição. Até 2005, no Brasil, a maioria dos cartões de crédito perteciam a emissores independentes dos bancos, sendo que os principais emissores eram: Credicard, Finivest (pertencente ao Unibanco) e IBI. As principais receitas dos emissores provêm do financiamento rotativo dos clientes, das anuidades e de seguros ou serviços agregados ao produto cartão.

2.4 Funcionamento do Cartão de Crédito

(GADI, 2006) ressalta que quando um cliente utiliza seu cartão de crédito, no mesmo instante é transmitido um sinal para o adquirente que repassa este sinal para a bandeira. Esta, por sua vez, envia o sinal para o emissor do cartão, que através de critérios próprios de crédito (como por exemplo, a disponibilidade de limite, cliente em atraso ou apontamento no sistema de detecção de fraude) decide por aprovar, negar ou referir a transação. As transações referidas são aquelas em que se pede para o cliente entrar em contato com o emissor ou bandeira, dependendo da localidade, para a confirmação de seus dados. Após a confirmação, o emissor ou bandeira decide o que deve ser feito com a transação e emite uma resposta ao estabelecimento, retornando ao ciclo. Todo esse ciclo deve ocorrer em um limite de tempo definido, que em média é de 10 segundos.

2.5 Maquinetas

As maquinetas são os equipamentos responsáveis pelo complemento das transações. Atualmente existem vários tipos de maquinetas, sendo que as principais são:

- **POS (*point of sale*):** esse sistema utiliza apenas uma linha telefônica para a comunicação e os cupons das vendas são impressos pelo próprio POS, não sendo necessário o uso de um computador ou de automação comercial. Este modelo é utilizado por quase todos os estabelecimentos do Brasil.

- **Manual:** a maioria dos estabelecimentos comerciais brasileiros possuem este tipo de aparelho para o caso de falha do aparelho POS. Neste caso, a transação é realizada através de um aparelho chamado *mata pulgaëm.* Este tipo de aparelho é muito utilizado em países como Bolívia e Peru.

- **ATM** (caixas eletrônicos de auto-atendimento): utilizados para o pagamento de contas com cartão de crédito.

- **Internet**: compra em páginas de vendas.

- **PVD**: são caixas eletrônicos que se comunicam com um POS. É comum encontrar este tipo de equipamento em supermercados ou em empresas com muitos pontos de recebimento. Neste caso, quem realiza a transação é o POS.

Capítulo 3

Fraude em Cartão de Crédito

Segundo o dicionário Aurélio (HOLANDA, 1990), fraude significa "abuso de confiança; engano criminal; uso de representações falsas para o ganho de vantagens injustas". Nos últimos anos, o desenvolvimento de novas tecnologias tem fornecido novos caminhos que permitem facilmente que transações fraudulentas sejam realizadas. Os fraudadores, em geral, são bem organizados e sempre procuram a maneira mais fácil e barata de se obter vantagens. A fraude é um negócio rentável, estável e muito bem organizado e administrado (HAND, ADAMS e BOLTON, 2002).

Nos Estados Unidos, segundo (GHOSH e REILLY, 1994), a perda com transações fraudulentas giram em torno de 850 milhões de dólares por ano para todos os cartões e para cada 100 libras gastos em cartão de crédito na Grã-Bretanha, 13 *cents* são perdidos em fraudes (HAND, ADAMS e BOLTON, 2002). Segudo (BHATLA, PRABHU e DUA, 2003) em 2002 as transações fraudelentas realizadas com cartões de crédito no mundo somaram mais de 2,5 bilhões de dólares e a projeção para o ano de 2008 é de 5 a 15 bilhões de dólares, a não ser que haja um rápido desenvolvimento da tecnologia de prevenção de fraude. No Brasil, de cada dez mil transações cinco são fraudulentas $(0,05\%)$ (GADI, 2006).

3.1 Tipos de Fraude

Uma forma comum de fraude é o roubo do cartão de crédito (BHATLA, PRABHU e DUA, 2003). Nesses casos, antes que seja comunicado o roubo e seja bloqueado o cartão,

o fraudador tenta gastar o que puder o mais rápido possível. O roubo pode ocorrer de diversas formas, entre elas: o roubo do cartão antes da entrega para o dono legítimo (ocorrendo no correio, nos carteiros ou nas caixas de correspondência), o roubo através da cópia da fita magnética do cartão através de uma leitora manual de cartão (ocorrendo em restaurantes, postos de gasolinas, lojas de conveniência) ou a cópia direta do número do cartão e do número de segurança (ocorrendo, principalmente, em uma fila em uma loja de conveniência, onde o fraudador utiliza, por exemplo, um telefone celular para copiar os número do cartão do desatento cliente à sua frente).

Uma outra forma é o fornecimento de dados falsos, como, por exemplo, nome, renda e documentos, para a obtenção do cartão. Nestes casos, os modelos que monitoram o comportamento de compra podem ser utilizados para detectar clientes que obtiveram o cartão através de uma admissão falsa. Um cliente que recebe o primeiro cartão e rapidamente inicia um movimento de compra compulsoriamente levanta suspeita, pois, em geral, um cliente legítimo não tem pressa.

Através do monitoramento das transações pode-se detectar a clonagem/falsificação de cartões (*counterfeit cards*), sendo que uma mudança radical no padrão de compras pode significar fortes indícios de fraude. Desse modo, este tipo de fraude é detectado através de características particulares que já são conhecidas como indicativos de falsificação.

Pode-se considerar também como fraudadores aqueles indivíduos, legítimos proprietários do cartão, que realizam compras sem intenção de pagá-las. Esse tipo de ação chama-se abuso.

As fraudes podem ser realizadas de duas maneiras: i) com a presença do cartão e do fraudador, onde o próprio fraudador, ou alguém sob o seu comando, realiza a compra com o cartão roubado/clonado/falsificado/obtido com aplicação falsa, ou ii) com a ausência do cartão e ausência do fraudador, onde a compra é realizada via correio/telefone/Internet (*Mail Order/ Telephone Order*).

Vale ressaltar que as compras realizadas pela Internet estão mais propícias a fraude, uma vez que, no caso, não é possível verificar a assinatura e foto do cliente com seu documento de identidade, o que torna a Internet um ambiente extremamente atrativo para os fraudadores.

A tabela 3-1 apresenta as porcentagens de fraude por modalidade.

Tabela 3.1: Modalidades de fraude em cartão de crédito e suas porcentagens de ocorrência

Modalidade de Fraude	Porcentagem
Perda ou roubo	63%
Clonagem	14%
Falsificação	12%
Intercepção no correio	6%
Outros	5%

Tabela 3.2:

Transações fraudulentas são realizadas diariamente em todas as partes do mundo. No topo da lista dos países com maior número de fraudes está a Ucrânia, onde 19.0% das transações com cartão de crédito são fraudes, seguindo da Indonésia com 18.3%. Pode-se citar também a Turkia com 9.0% e a Malasia com 5.9%. Apesar dos Estados Unidos e Reino Unido não estarem ocupando as primeiras colocações, nesses últimos anos a perda com fraude nestes países vem aumentando drasticamente (Hahnstra [22]).

3.2 Ciclo da Fraude

Cristofaro [15] ressalta a existência de um ciclo de vida da gestão de fraude que pode ser dividido em oito estágios. Estes estágios são:

- **Intimidação:** caracterizada por ações e atividades destinadas a inibir ou desanimar o fraudador antes que o mesmo execute a fraude.
- **Prevenção**: compreende atividades que tornam a execução da fraude mais difícil, endurecendo as defesas contra os fraudadores. Isso inclui a identificação pessoal de números para os cartões de crédito, sistemas de segurança para transações via internet e uso de senhas pessoais para o acesso de contas no sistema, tanto via computadores como via telefone. Esses métodos não são perfeitos e é necessário determinar um meio termo entre as despesas envolvidas na gestão da fraude e a incoveniência que esta pode gerar para o cliente.
- **Detecção**: conjunto de ações e atividades que são utilizados para identificar a fraude. Esses métodos de detecção são utilizados quando os métodos de prevenção da

fraude falham. Na prática a detecção de fraude é utilizada continuamente, ignorando a prevenção.

- **Medidas**: tomada de medidas que evitam a ocorrêcia de perdas ou a sua continuidade e/ou impeçam que um fraudador continue a fraudar ou termine a sua atividade de fraude.

- **Análise**: são identificados e estudados através da modelagem estatística os fatores que levam os fraudadores a cometerem as fraudes.

- **Política**: conjunto de atividades que pretendem criar, avaliar, comunicar e ajudar na implantação de políticas para reduzir a incidência de fraudes.

- **Investigação**: envolve a obtenção de evidências e informações suficientes para reduzir ou inibir completamente as atividades fraudulentas, com o intuíto de recuperar recursos ou obter a restituição dos mesmos.

- **Acusação**: neste estágio está clara a necessidade de suporte jurídico (leis) para condenar os crimonosos.

Através de uma interação equilibrada entre os estágios citados acima, pode-se alcançar resultados bem eficiêntes. Cada instituição deve descobrir o melhor equilíbrio para seu negócio. Estes estágios interagem entre si de forma dinâmica e não são necessariamente executados na ordem descrita acima.

3.3 Custos da Fraude

As transações fraudulentas geram grandes custos para as instituições, dentre as quais pode-se citar: o valor perdido pela transação, gastos com investigações, reemissão do cartão e despesas com a entrega, chamadas no atendimento ao cliente, custos da transações referidas, cancelamento de cartões e custos de boletim de proteção com as bandeiras.

Além desses custos, têm-se os custos não mensuráveis como, por exemplo, a insatisfação do cliente, o sentimento de violação e vulnerabilidade em relação à empresa do cartão, perda de lealdade à marca e às bandeiras além dos custos de oportunidade.

Capítulo 4

Métodos para a Detecção de Fraude

A colaboração da estatística está direcionada para os métodos de detecção de fraude. Esses métodos podem ser classificados em supervisionados e não supervisionados.

Nos métodos supervisionados, amostras de registros fraudulentos e não fraudulentos são usados para construir modelos que permitem classificar uma nova observação em uma dessas duas classes. Para isso é necessário que o conjunto de dados seja confiável e que seja composto, necessariamente, de observações para as duas classes. Para estes métodos é comum a utilização das técnicas de análise discriminante, método baseado em árvore (Cart, ID3), regressão logística e análise de sobrevivência. Também podem ser usados outros procedimentos não convencionais para os estatísticos, tais como: redes neurais, redes bayesianas, método baseado em regras e meta-aprendizado.

Os métodos não supervisionados são utilizados quando não existe um conjunto de dados com casos conhecidos de fraudes e de transações legítimas. As técnicas empregadas nestes casos são, em geral, uma combinação de métodos de detecção de *outliers* e de perfis e é comum o uso de ferramentas estatísticas para a verificação da qualidade dos dados. Esses métodos buscam a conta ou o cliente que seja o mais dissimilar dos outros (do normal). Nesse contexto, o cliente é examinado com mais cuidado. A idéia consiste em modelar uma distribuição base que representa um comportamento padrão e, a partir daí, tentar detectar observações que mostram um afastamento deste padrão. Estes modelos probabilísticos podem ser atualizados em períodos de tempo pré-fixados ou podem ser determinados continuamente no tempo.

Ambos os métodos, supervisionados e não supervisionados, possuem problemas. Com

relação aos métodos supervisionados o problema encontrado é com relação a classificação errônea (falsos positivo/negativo), ou seja, a transação fraudulenta ainda não observada é rotulada como legítima ou uma transação legítima é registrada erroneamente como fraudulenta. Um outro problema é quanto ao desbalanciamento das classes, que ocorre quando o número de transações legítimas é muito maior que às fraudulentas.

Com relação aos métodos não supervisionados, usuários legítimos podem mudar seus comportamentos de compra em um longo período de tempo e esta mudança pode ocasionar em um alarme falso. Um outro problema é quanto a determinação de um ponto atípico, pois este não garante que uma ação fraudulenta tenha sido efetivada. Neste caso, a análise serve como um alerta para o fato de que uma observação é atípica e que, por esta razão, merece ser investigada detalhadamente.

4.1 Árvores de Decisão

A metodologia de decisão por árvore ID3, desenvolvido por Quinlan [4], utiliza o critério de entropia para dividir os nós, partindo do princípio de que a entropia cresce com a probabilidade associada a um determinado estado.

O conceito de entropia é amplamente utilizado em física, mais precisamente em termodinâmica. Em estatística está relacionado com a quantidade de informação para explicar um determinado evento. Como por exemplo, ao jogar uma cartela na mega sena a probabilidade de acertar o primeiro jogo é quase nula, de acertar a quina é pequena, mas é maior do que acertar as seis dezenas; a de acertar uma quadra é pequena, mas é maior que todas as anteriores. Quando a probabilidade de acertar a sena é quase nula, a entropia pode ser 1 (quantidade de informação nula), para a quina a entropia pode ter valor 0,96 e para a quadra a entropia pode ser 0,92. Ou seja, a entropia pode ser vista como um valor associado à necessidade de informação para a explicação de um evento.

No caso da classificação pelo método de árvore, quando a entropia é nula, significa que os dados são homogêneos (mesma classe). No caso do método ID3, dado um determinado nó x, o critério de divisão usado é:

$$Entropia(\mathbf{x}) = \sum_i - p_i \log(p_i), \tag{4.1}$$

onde p_i é a probabilidade da *i-ésima* classe dentro do nó x. Para expressar a proporção de informação gerada pela divisão, utiliza-se a seguinte razão:

$$Razão\ Ganho\ (\mathbf{x}) = \frac{Ganho(x)}{Informação(x)}, \tag{4.2}$$

onde o ganho é a diferença da entropia de informação do nó x e o denominador refere-se as divisões realizadas para o nó x.

$$Ganho(\mathbf{x}) = entropia(x) - entropia(x_1, x_2), \tag{4.3}$$

$$Entropia(\mathbf{x}_1, x_2) = -(p_1 \log(p_1) + p_2 \log(p_2)). \tag{4.4}$$

O valor resultante da razão ganho das variáveis preditoras mostra qual variável deve ser testada em ordem de nós, contada da raiz (maior valor da razão ganho) até o nó mais distante (menor valor da razão ganho). A seguir um exemplo retirado de (CARVALHO, 2001).

Suponha que uma locadora de carros queira classificar os seus clientes como locatário de carro importado ou nacional mediante duas variáveis preditoras: idade e salário. Têm-se que 17 clientes alugaram carros nacionais (N) e 7 alugam carros importados (I).

Para iniciar a classificação é necessário escolher alguma variável para ser utilizada na formação de um subgrupo mais homogêneo. No exemplo, o autor escolheu a idade como nó inicial, verificando-se que os clientes tinham entre 18 a 70 anos. Logo após, foram produzidos alguns valores de idade: 20, 40, 50 e 60. Com a idade de 20 anos, dividiu-se o grupo de clientes em dois subgrupos: aqueles que têm mais de 20 anos e aqueles que têm menos de 20 anos. A entropia dos dois subgrupos foi calculada, obtendo-se a entropia total da informação neste primeiro nível da árvore de decisão. Suponha que o valor da entropia seja $H_{20} = 0,78$. Partindo do grupo inicial, novamente, dois novos subgrupos foram criados: aqueles com mais de 40 anos e aqueles com menos de 40 anos. Novamente, calculando a entropia desses subgrupos obteve-se o valor $H_{40} = 0,64$. Repetindo-se o processo para as idades de 50 e 70 anos, os valores da entropia obtidos foram $H_{50} = 0,71$ e $H_{70} = 0,81$. Comparando os valores calculados das entropias, observa-se que a menor entropia é a do subgrupo 40, ou seja, ao se utilizar uma regra com base na idade menor

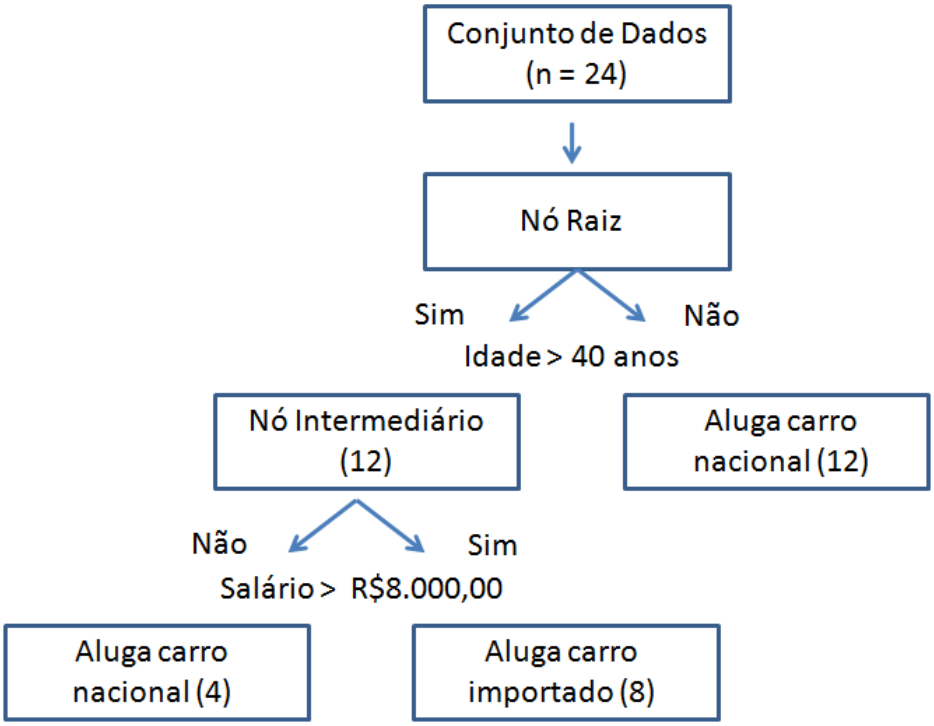


Figura 4-1: Árvore de Decisão para o exemplo de locação de carro (CARVALHO, 2001)

ou igual que 40 anos, obtêm-se dados homogêneos.

O próximo passo foi definir um novo critério para a separação dos dados remanescentes ao primeiro nó (teste). A variável salário pôde ser aplicada no próximo nó, onde os valores variavam de R$ 2.000,00 a R$ 10.000,00. Do mesmo modo à primeira aplicação, os grupos foram subdivididos em: R$ 3.000,00; R$ 5.000,00 e R$ 8.000,00 e foram calculadas as entropias associadas, obtendo $H_{3.000} = 0,61$, $H_{5.000} = 0,21$ e $H_{8.000} = 0,13$. Então, para o teste do último nó puderam ser utilizados salários maiores e menores que R$ 8.000,000. A árvore de classificação desse exemplo é mostrada na Figura 4-1.

Com base nos dados de treinamento da árvore de decisão, pôde ser gerado um modelo que permite classificar um conjunto de dados multivariados, baseado na razão ganho, uma árvore, contendo testes, permite classificar dados localizados na área de confusão entre classes.

4.2 Análise Discriminante

A análise discriminante é um outro método de predição utilizando a classificação. Seu princípio é descrever graficamente (em 3 dimensões ou menos) ou algebricamente as características diferenciais de um conjunto de observações multivariadas. Cada uma dessas observações traz informações de p variáveis e estão definidas no espaço p-*dimensional* R^p.

Esse tipo de análise de classificação permite alocar as observações em duas ou mais classes, buscando determinar características discriminantes, cujos valores numéricos fazem com que as populações estejam tão separadas quanto possível (JOHNSON e WICHERN, 1988).

Considere g populações ou grupos $\pi_1, ..., \pi_g$, sendo $g \geq 2$. Suponha que a cada população π_j está associada uma função densidade de probabilidade $f_j(x)$ no espaço R^p, ou seja, se um indivíduo pertence a uma população π_j, tem função densidade de probabilidade $f_j(x)$. O objetivo da análise discriminante é alocar um indivíduo para um dos g grupos com base nas observações x.

A seguir, será apresentada a metodologia de discriminação no caso de duas populações, $g = 2$.

4.2.1 Função Discriminante de Fisher para Duas Populações

A idéia da função discriminante de Fisher é transformar a observação multivariada X em uma univariada Y, tal que Y traga informação das populações π_1 e π_2. Se essas populações forem as mais distintas possíveis, fica mais fácil afirmar a qual delas pertence as observações; mas nem sempre isso acontece e as populações ocupam algumas áreas em comum no espaço, denominadas "regiões de confusão" (JOHNSON e WICHERN, 1988).

Para resolver esse problema, Fisher sugeriu tomar a combinação linear de X para criar Y ($y = \widehat{l}'x$), por ser uma função simples de X e de fácil tratamento matemático. Tendo μ_{1y} ($E(l'x/\pi_1)$) como a média dos resultados de Y, cujas observações pertencem a π_1 e μ_{2y} ($E(l'x/\pi_2)$) a média de Y obtida de X que pertence a π_2, Fisher selecionou a combinação linear que maximiza o quadrado da distância entre μ_{1y} e μ_{2y} relativa à variabilidade de X nas duas populações. Essa combinação é dada pelas matrizes de covariância

$$\Sigma = E_i[(x - \mu_i)(x - \mu_i)\prime], \ i = 1, 2, \tag{4.5}$$

consideradas iguais para as duas populações. Nessa matriz, μ_1 e μ_2 são, respectivamente, a média da população de X da população π_1 e média de X da população π_2.

A distância máxima das duas populações é dada por $(x-\mu_1)'\Sigma^{-1}(x-\mu_2)$. As quantidades populacionais μ_1, μ_2 e Σ raramente são conhecidas e a expressão anterior só poderá ser utilizada se forem consideradas as estimativas das quantidades populacionais.

Considere n_1 observações da variável multivariada $X' = [x_1, x_2, ..., x_p]$ de π_1 e n_2 observações de π_2. Sejam as seguintes estatísticas relativas as amostras, denotando, respectivamente, a média amostral e variância amostral,

$$\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \ , \tag{4.6}$$

$$\mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)', \ i = 1, 2. \tag{4.7}$$

A função discriminante de Fisher é construída sem assumir a existência de uma função de probabilidade associada a cada grupo.

Fisher escreveu uma função linear $y = l'x$, que maximiza a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos. Porém, se as duas populações têm uma matriz de variância e covariância comum, a matriz S pode ser substituída pela matriz S_{pooled}:

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 2)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 2)}\right] S_2. \tag{4.8}$$

Para alocar o objeto na população π_1, primeiramente precisa-se definir o ponto médio $\widehat{m}$ da combinação linear:

$$\widehat{m} = \frac{\overline{x}_1 - \overline{x}_2}{2} = \frac{1}{2}(\overline{x}_1 - \overline{x}_2)' S_{pooled}(\overline{x}_1 - \overline{x}_2). \tag{4.9}$$

Dessa forma uma observação x_0 será classificada como pertencente à população π_1 se:

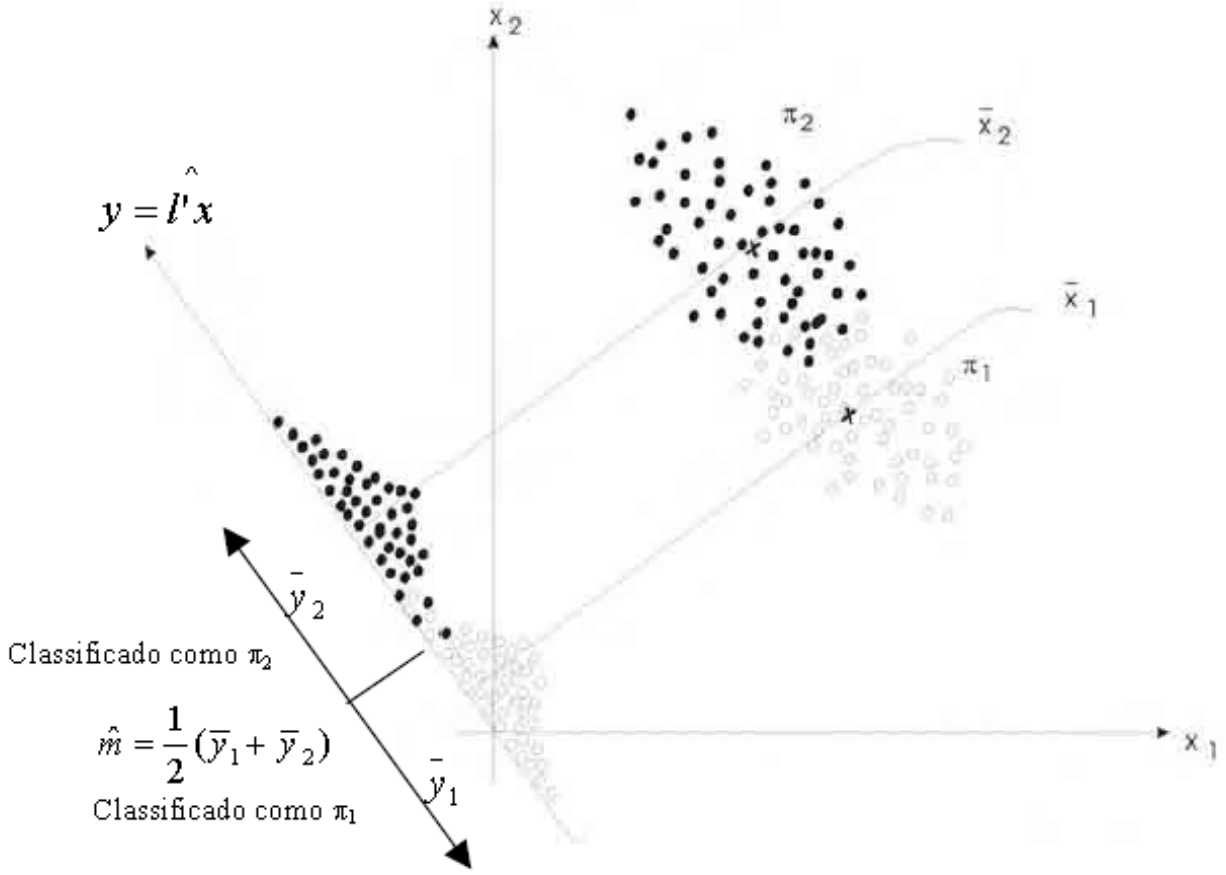


Figura 4-2: Processo de classificação pelo método de Fisher para 2 populações (JOHNSON e WICHERN, 1988)

$$(\overline{x}_1 - \overline{x}_2)' S_{pooled}(x_0) \geqslant \widehat{m}, \tag{4.10}$$

e alocada para o grupo π_2 se:

$$(\overline{x}_1 - \overline{x}_2)' S_{pooled}(x_0) < \widehat{m}. \tag{4.11}$$

A Figura 4-2 mostra a solução de Fisher para o problema de separação e classificação para $p = 2$. O conjunto de dados não fica tão discriminado se projetados nos eixos x_1 e x_2, assim sendo, o método rebate os dados numa função linear dos dois eixos. A melhor decisão é a que torna máxima a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos.

Segundo (JOHNSON e WICHERN, 1988), este tipo de análise só faz sentido se as duas populações realmente tiverem médias diferentes. Suponha que as populações π_1 e π_2 sejam normais multivariadas com uma matriz de covariância comum Σ. Um teste com uma hipótese nula $H_0 : \mu_1 = \mu_2$, contra a hipótese alternativa $H_1 : \mu 1 \neq \mu 2$ é dado pela estatística:

$$\frac{(\overline{x}_1 + \overline{x}_2 - p - 1)}{[(\overline{x}_1 + \overline{x}_2 - 2)p]} \frac{n_1 n_2}{n_1 + n_2} D^2, \tag{4.12}$$

que possui uma distribuição F com $v_1 = p$ e $v_2 = n_1 + n_2 - p - 1$ graus de liberdade e $D_2 = (\overline{x}_1 - \overline{x}_2)' S_{pooled}(\overline{x}_1 - \overline{x}_2)$. Se H_0 for rejeitada, pode-se concluir que a separação entre as duas populações π_1 e π_2 é significante, caso contrário significa que as duas populações têm a mesma média e matriz de covariâncias, ou seja, elas formam uma única população.

Capítulo 5

Regressão Logística

Os métodos de detecção de fraude apresentados no capítulo anterior são métodos de classificação. Nesses métodos, a separação dos indivíduos em determinados grupos ocorre a partir das características mais semelhante. Dessa forma é possível criar uma regra de classificação para novas entradas baseado na discriminação anterior.

Segundo (HOSMER e LEMESHOW, 1989), a regressão logística busca explicar a relação entre uma variável resposta dicotômica dependente e um conjunto de variáveis explicativas independentes (qualitativas ou quantitativas). Este capítulo será destinado aos modelos de regressão logística.

5.1 Introdução

A Regressão Logística é um modelo probabilístico de regressão não linear que se encaixa nas situações em que as variáveis resposta são discretas e os erros não são normalmente distribuídos. Esta técnica é utilizada quando se deseja predizer um evento futuro, como por exemplo, descrever a probabilidade de que um determinado cliente venha a cometer uma transação fraudulenta dado um conjunto de covariáveis.

A variável resposta (y), que mesmo quando não é originalmente binária pode ser dicotomizada, apresenta dois possíveis resultados (sucesso e fracasso). Geralmente é chamado de sucesso o resultado que representa a presença de uma particular característica de interesse:

$$y = \begin{cases} 1 & \text{se o elemento possui a característica de interesse} \\ 0 & \text{caso contrário} \end{cases} . \tag{5.1}$$

A quantidade chave em problemas de regressão é o valor médio da variável resposta dado o valor da variável independente, ou seja, a esperança condicional de Y dado um valor x da covariável X, $E(y|X = x)$. No caso da regressão linear assume-se que esta média pode ser expressa em uma equação linear em x, ou alguma transformação de X ou em Y, tal como:

$$E(Y|x) = \beta_0 + \beta_1 x. \tag{5.2}$$

Assim sendo, é possível que a média assuma qualquer valor quando x varia entre $-\infty$ e $+\infty$.

Em (HOSMER e LEMESHOW, 1989) é apresentado um exemplo considerando duas variáveis: idade dos indivíduos (AGE) e se elas tinham ou não problemas cardíacos (CHD). A Figura 5-1 mostra o gráfico dessas duas variáveis. Observe ainda que todos os pontos caem em uma das duas linhas paralelas, representando a ausência de CHD ($y = 0$) e a presença de CHD ($y = 1$). Observe que existe uma tendência para que os indivíduos com nenhuma evidência de CHD sejam mais jovens do que aqueles com evidência de CHD. Apesar deste gráfico descrever a natureza dicotômica da variável CHD, pode-se dizer que ele não fornece um aspecto claro da natureza da relação entre CHD e idade.

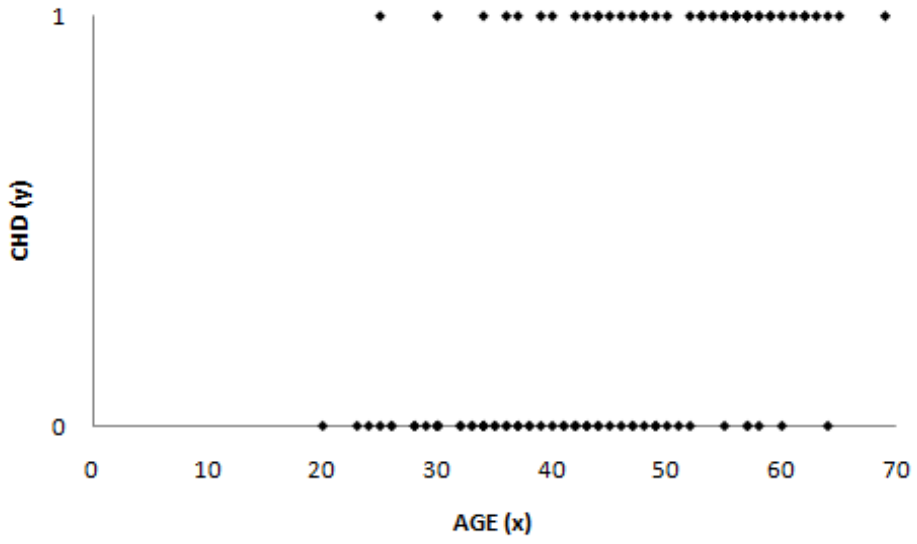


Tabela 5.1: Frequência de CHD por faixas de idade. Fonte: HOSMER e LEMESHOW, 1989.

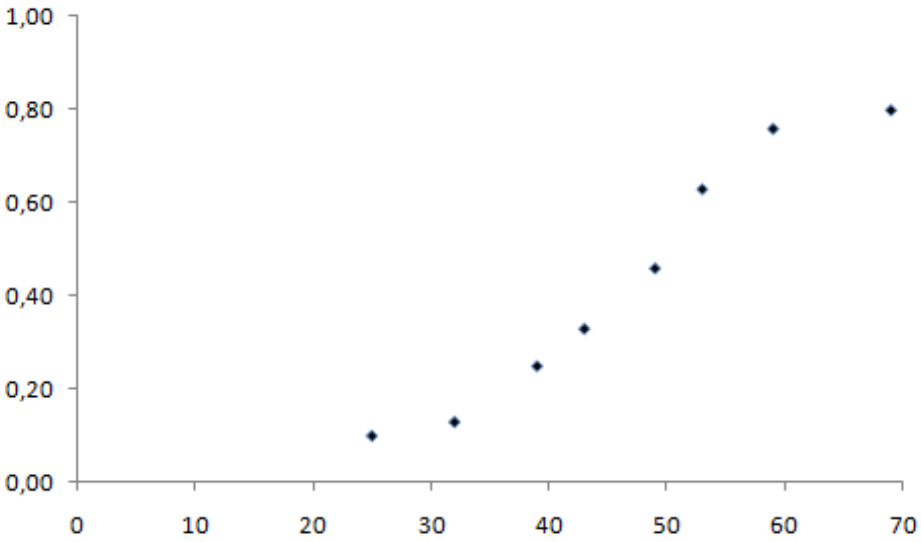
		CHD		
Idade do Grupo	n	Ausente	Presente	Média (proporção)
20 - 29	10	9	1	0.10
30 - 34	15	13	2	0.13
35 - 39	12	9	3	0.25
40 - 44	15	10	5	0.33
45 - 49	13	7	6	0.46
50 - 54	8	3	5	0.63
55 - 59	17	4	13	0.76
60 - 69	10	2	8	0.80
Total	100	57	43	0.43

Tabela 5.2:

A grande variabilidade em CHD e idade dificulta a interpretação dos dados. Para remoção dessa variação, mantendo-se a estrutura da relação entre a resposta e a covariável, a variável independente foi dividida em 8 classes. Em seguida calculou-se a média condicional da variável resposta em cada grupo, tabela 5.2.

Examinando a Tabela 5.2, pode-se observar melhor a existência de uma relação. Aparentemente, com o aumento da idade a proporção de indivíduos com problemas cardíacos (CHD) aumenta. A Figura 5-2 apresenta o gráfico da proporção de indivíduos com problemas cardíacos versus o ponto médio de cada grupo de idade. Através desta Figura pode-se ter uma melhor compreensão da relação entre a variável resposta CHD e a covariável idade.

Quando a variável resposta é dicotômica, sua média condicional deve ser maior ou igual a zero e menor ou igual a um, $[0 \leq E(Y|X = x) \leq 1]$, aproximando-se de 0 e de 1 gradualmente (forma de "S") e cujo gráfico se parece com a distribuição acumulada da função logística.



Segundo os estudos de (COX e HOSMER e LEMESHOW, 1989) [2], a função ideal para o caso da variável resposta ser dicotômica é a função logito, pois é extremamente flexível e fácil de ser usada e interpretada. A função logito $\pi(x)$ representa a probabilidade desconhecida da existência de uma determinada característica de interesse, associada ao valor de x da variável independente (ou dado o valor de X). O modelo de regressão logística, conhecido também como função logística, é dado por:

$$\pi(x) = P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}, \tag{5.3}$$

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

Utilizando a transformação *logit* em $\pi(x)$ tem-se a seguinte função linear:

$$\text{log it } \pi(x_i) = \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \boldsymbol{\beta}'\mathbf{X}_i. \tag{5.4}$$

5.2 Estimação dos Parâmetros

Na abordagem clássica, o parâmetro é considerado uma quantidade fixa e desconhecida. Os resultados são obtidos a partir de uma distribuição conjunta da amostra observada de tamanho n, $\mathbf{x} = (x_1, ..., x_n)$, e representada pela função de verossimilhança $\mathrm{L}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})$ (MOOD,1974). Segundo (HOSMER e LEMESHOW, 1989), o método usual para a esti-

mação dos parâmetros do modelo de regressão logística é o método de máxima verossimilhança.

Considere Y_i variáveis aleatórias indepententes e identicamente distribuídas. Seja $\mathbf{X}_i$ o vetor de covariáveis, tal que a distribuição de $Y_i|\mathbf{X}_i$ tenha distribuição de Bernoulli com probabilidade de sucesso $\pi(x)$. Dessa forma, a distribuição de $Y_i|\mathbf{X}_i$ pode ser representada por:

$$P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (\pi(x_i))^{y_i}(1-\pi(x_i))^{1-y_i}, \text{ com } y_i = 0, 1 \text{ e } i = 1, ..., n. \quad (5.5)$$

Como as variáveis aleatórias Y_i são independentes, a função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}; \mathbf{x}_i) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} (\pi(x_i))^{y_i}(1-\pi(x_i))^{1-y_i}. \quad (5.6)$$

O logaritmo da verossimilhança em (5.6) é dado por:

$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \ln L(\boldsymbol{\beta}, \omega; \mathbf{x}_i), \quad (5.7)$$

ou seja,

$$l(\boldsymbol{\beta}; \mathbf{x}_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right]. \quad (5.8)$$

A estimativa do *k-ésimo* parâmetro é dada pela derivada da Equação (5.8) em relação a este parâmetro e igualando este resultado a zero, isto é,

$$\widehat{\beta}_t \text{ é solução de } \frac{\delta l(\boldsymbol{\beta}, \omega; \mathbf{x}_i)}{\delta \beta_t} = 0, \; t = 1, ..., p. \quad (5.9)$$

Considerando uma covariável, o vetor de parâmetros desconhecidos é dado por $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$. Ao maximizar a função de verossimilhança, ou seja, realizar as derivadas desta função em relação aos parâmetros β_0 e β_1 obtém-se 2 equações:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, \tag{5.10a}$$

e

$$\sum_{i=1}^{n} x[y_i - \pi(x_i)] = 0. \tag{5.10b}$$

Observe que as equações acima são não lineares nos parâmetros e dessa forma as soluções dessas equações são obtidas utilizando métodos iterativos.

As estimativas de máxima verossimilhança dos parâmetros do vetor $\boldsymbol{\beta}$ são dados pela solução das equações de verossimilhança e serão denotadas por $\widehat{\boldsymbol{\beta}}$. As propriedades mais importantes dos estimadores de máxima verossimilhança são:

- A estimativa de máxima verossimilhança pode ser tendênciosa, e com freqüência, tal tendenciosidade pode ser eliminada pela multiplicação de uma constante apropriada;
- Sob condições bastante gerais, as estimativas de máxima verossimilhança são coerentes. Isto é, se o tamanho da amostra sobre a qual essas estimativas são calculadas for grande, a estimativa será próxima do valor do parâmetro. As estimativas de máxima verossimilhança possuem a propriedade de "grandes amostras"; isto é, elas são assintóticas e possuem uma boa aproximação pela distribuição Normal;
- As estimativas de máxima verossimilhança apresentam a seguinte propriedade de invariância: Se $\widehat{\theta}$ é uma estimativa de máxima verossimilhança de θ, então $g(\widehat{\theta})$ também é uma estimativa de máxima verossimilhança de uma função monótona contínua $g(\theta)$ (MEYER, 1978).

5.3 Teste de Significância dos Coeficientes

Segundo (HOSMER e LEMESHOW, 1989), uma aproximação para testar a significância do coeficiente de uma variável em qualquer modelo relaciona-se com a seguinte questão: o modelo que inclui a variável em questão diz mais sobre a variável resposta do que o modelo que não inclui a variável? Com esta questão, o teste de significância consiste em comparar os valores observados da variável resposta com aqueles preditos, através de dois

modelos. O primeiro com a variável presente e o segundo sem essa variável. A comparação entre os valores preditos e observados, usando a função de verossimilhança, é baseada na seguinte expressão:

$$D = -2\log\left[\frac{\text{verossimilhança do modelo atual}}{\text{verossimilhança do modelo saturado}}\right]. \tag{5.11}$$

Esse teste é donominado de teste da razão de verossimilhanças e é aplicado em testes de hipóteses pelo fato de sua distribuição geralmente ser aproximada por uma qui-quadrado, ou seja, a distribuição é conhecida.

Essa estatística D é chamada de função *deviance* e desempenha o mesmo papel que a soma de quadrados residuais no modelo de regressão linear ($SSE - Soma\ dos\ Quadrados\ dos\ Erros$). Para estimar a significância de uma variável independente, comparam-se o valor de D com e sem a variável independente na equação. A alteração em D, devido a inclusão da variável independente no modelo, é dada por:

$$G = D(\text{para o modelo sem a variável}) - D(\text{para o modelo com a variável}). \tag{5.12}$$

Esta estatística desempenha o mesmo papel do numerador do teste F na regressão linear, pois a verossimilhança do modelo saturado é comum para ambos os valores de D sendo eliminado no cálculo de G. Assim, G pode ser expresso como:

$$G = -2\log\left[\frac{\text{verossimilhança sem a covariável}}{\text{verossimilhança com a covariável}}\right]. \tag{5.13}$$

O teste da razão de verossimilhanças torna possível verificar a significância da adição de novos termos no modelo. No caso de uma única variável independente, recomenda-se ajustar primeiro um modelo contendo apenas o termo constante. Em seguida deve-se ajustar um modelo contendo a variável independente, mais a constante. Estes dados originam um novo log de verossimilhança. O teste da razão de verossimilhança é obtido multiplicando-se a diferença destes dois valores por menos dois. Este resultado, bem como o *p-valor* associado à distribuição qui-quadrado, podem ser obtidos na maioria dos *softwares* estatísticos.

Para verificar a significância dos parâmetros também pode ser utilizados outros métodos estatísticos semelhantes ao anterior, como o Teste de Wald ou o Teste de Escore.

A estatística e teste de Wald dado por,

$$W = \left[\frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)} \right]^2, \tag{5.14}$$

é obtido comparando-se o estimador de máxima verossimilhança do parâmetro de inclinação, $\widehat{\beta}_1$, com a estimativa do seu erro padrão ($\widehat{SE} - Erro\ Padrão$). O resultado da razão, sob a hipótese básica $\beta_1 = 0$, terá uma distribuição normal padrão.

O *p-valor* bicaudal é $P(|z| > W)$, onde z denota uma variável aleatória seguindo uma distribuição normal padrão.

Porém, alguns pesquisadores que examinaram a eficiência do teste de Wald (HOSMER e LEMESHOW, 1989) verificaram que, às vezes, este teste rejeita um coeficiente quando este é significativo. Por este motivo, eles recomendam que o teste da razão de verossimilhanças seja usado. Ambos os testes, da razão de verossimilhança (G) e o teste de Wald (W) requerem o cálculo dos estimadores de máxima verossimilhança. Para uma única variável, esta não é uma tarefa computacionalmente difícil, porém, para conjuntos de dados grandes com muitas variáveis, o cálculo iterativo necessário para obter a estimativa de máxima verossimilhança pode ser muito trabalhoso.

Um teste para a significância de uma variável que não requer estes cálculos é o Teste de Escore. Os proponentes do teste de Escore citam esta redução de esforço computacional como a sua maior vantagem. Entretanto, o uso deste teste é limitado pelo fato de que ele não pode ser obtido facilmente em alguns *software*. O teste de Escore é baseado na teoria de distribuição de derivadas do log da verossimilhança (HOSMER e LEMESHOW, 1989).

Capítulo 6

Logística Limitada

No capítulo anterior discutiu-se de forma sucinta o modelo de regressão logística. Alguns estudos mostram que este modelo não apresenta boas estimativas quando a variável resposta é extremamente desbalanciada, como geralmente é o caso da estrutura dos bancos de dados de fraude. Nesses casos, uma alternativa sugerida por (CRAMER, 2004) é de se utilizar o modelo de regressão logística limitada.

O modelo logito limitado provém de uma modificação do modelo logito usual. Essa modificação é dada pelo acréscimo de um parâmetro que quantifica um limite superior para a probabilidade de sucesso. Assim a probabilidade de sucesso condicionada as covariáveis é dada pela Equação 6.1.

$$P(y_i = 1|\mathbf{x}_i) = \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}, \quad 0 \leq \omega \leq 1. \tag{6.1}$$

Através do estudo realizado por (CRAMER, 2004) tem-se que este modelo apresenta uma excelente performance ao modelar grandes conjuntos de dados com vetor de covariáveis $\mathbf{x}_i$ e uma variável resposta binária y_i, com baixa incidência de $y_i = 1$ (resposta de interesse) e uma altíssima incidência de $y_i = 0$.

6.1 Estimação dos Parâmetros

Os parâmetros do modelo logito limitado também são determinados via máxima verossimilhança.

Considere o vetor de $(p+1)$ parâmetros dado por $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_p)$ e

$$P_i = \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{6.2}$$

Como a variável resposta $Y_i \sim Bernoulli(P_i)$, as probabilidades de sucesso e fracasso são dadas por $P(Y_i = 1|\mathbf{x}_i) = P_i$ e $P(Y_i = 0|\mathbf{x}_i) = (1-P_i)$, respectivamente. A distribuição de Y_i pode ser representada por:

$$P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (P_i)^{y_i}(1-P_i)^{1-y_i}, \qquad com \qquad y_i = 0, 1 \text{ e } i = 1, ..., n. \tag{6.3}$$

Como as variáveis aleatórias Y_i são independentes, a função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}, \omega; y_i, \mathbf{x}_i) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} (P_i)^{y_i}(1-P_i)^{1-y_i}. \tag{6.4}$$

Considerando

$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \ln L(\boldsymbol{\beta}, \omega; \mathbf{x}_i), \tag{6.5}$$

temos

$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] I_{(0,1)}(\omega). \tag{6.6}$$

Ao maximizar a função de verossimilhança, ou seja, realizar as derivadas desta função em relação aos parâmetros $\beta_0, \beta_1, ..., \beta_p$ e ω obtém-se $(p+2)$ equações (Apêndice A):

$$\sum_{i=1}^{n} \omega[y_i - P_i] = 0; \tag{6.7a}$$

$$\sum_{k=1}^{p} \sum_{i=1}^{n} x_{ij} \omega[y_i - P_i] = 0; \tag{6.7b}$$

$$\sum_{i=1}^{n}\left[\frac{y_i - P_i}{1 - P_i}\right] = 0. \tag{6.7c}$$

Pode-se verificar que essas equações não são lineares nos parâmetros, não sendo possível obter a solução explícita do sistema de equações. Então, é necessário o uso de métodos iterativos para resolvê-lo, encontrando assim as estimativas de máxima verossimilhança, $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p$ e $\widehat{\omega}$.

6.2 Método Iterativo de Newton-Raphson

O objetivo do método iterativo de Newton-Raphson é de determinar as soluções de equações não lineares nos parâmetros. Este método consiste em gerar uma sucessão de soluções $\{\theta_k\}$, $k \in N$, que convirja para a solução α, também conhecida como raiz da equação, a partir de uma aproximação inicial θ_0 e um procedimento iterativo que permita obter novos θ_{k+1} a partir dos valores anteriores θ_k. Um método iterativo, de forma geral, para aproximar soluções de $f(\theta) = 0$ consiste de:

- um valor inicial θ_0 (se o método é dependente de apenas um ponto);
- uma fórmula iterativa que permita obter sucessivamente novas aproximações θ_{k+1} a partir dos valores anteriores θ_k,

$$\theta_{k+1} = g(\theta_k), \text{ com } K = 0, 1, ..., k; \tag{6.8}$$

- um critério de parada (método de convergencia), que por exemplo podem ser:
  1. $|\theta_k - \theta_{k-1}| < \epsilon_1$ (tolerância absoluta);
  2. $(|\theta_k - \theta_{k-1}|)/\,|\theta_k| < \epsilon_2$ (tolerância relativa);
  3. $|f(\theta_k)| < \epsilon_3$;
  4. Impondo um valor máximo de iterações $k = \max(t)$.

A função iterativa de Newton-Raphson é dada por:

$$g(\theta) = x - \frac{f(\theta)}{f'(\theta)}. \tag{6.9}$$

Dado um valor inicial θ_0, o método se resume a efetuar as iterações:

$$\theta_{k+1} = \theta_k - \frac{f(\theta_k)}{f'(\theta_k)},\ k = 1, 2, ... \tag{6.10}$$

6.2.1 Condições suficientes de convergência

Teorema: Se as seguintes condições forem satisfeitas:

1. $f(a)f(b) < 0$;
2. $f'(\theta) \neq 0$, para qualquer $\theta \in [a, b]$;
3. $f''(\theta) > 0$ ou $f''(\theta) < 0$, para qualquer $\theta \in [a, b]$, ou seja, $f''(\theta)$ não muda de sinal em $[a, b]$;
4. $\left|\frac{f(a)}{f'(a)}\right| < (b - a)$ e $\left|\frac{f(b)}{f'(b)}\right| < (b - a)$.

Então o método de Newton-Raphson converge para um único α de f em $[a, b]$, qualquer que seja o valor inicial $\theta_0 \in [a, b]$.

No *software* SAS pode-se obter as soluções de equações não lineares pelo método de Newton-Raphson via procedimento NLP.

6.3 Testes de Significância

Os testes para o modelo logito limitado são os mesmo do modelo logito. A única dificuldade é que para este modelo tem que impletar os códigos.

Para verificar a signifcância dos parâmetros do modelo de regressão logística limitada será utilizado o teste de Wald.

6.4 Comparação dos Modelos

Para comparar as performances dos modelos logito e logito limitado serão utilizadas as estatísticas de teste: Akaike's Information Criterion (AIC) e Schwarz Criteria (SC), além

da estatística de Kolmogorov-Smirnov (KS).

6.4.1 Akaike's Information Criterion - AIC

A estatística AIC é dada da seguinte forma:

$$AIC = 2k - 2\ln(L). \tag{6.11}$$

Sendo k o número de parâmetros do modelo.

6.4.2 Schwarz Criteria - SC

A estatística SC é dada da seguinte forma:

$$AIC = k\ln(n) - 2\ln(L). \tag{6.12}$$

Sendo k o número de parâmetros do modelo e n o número de observações na amostra.

6.4.3 Estatística de Kolmogorov-Smirnov (KS)

A estatística de Kolmogorov-Smirnov (KS) provêm do teste não-paramétrico, no qual se deseja a partir de duas amostras retiradas de duas populações distintas, testar se duas funções de distribuições associadas às duas populações são idênticas ou não.

A estatística KS mede o quanto estão separadas as funções de distribuições empíricas dos escores dos grupos de não fraudadores e fraudadores. Seja $F_B(e) = \sum_{x \leq e} F_B(x)$ e $F_M(e) = \sum_{x \leq e} F_M(x)$ a função de distribuição empírica dos não fraudadores e fraudadores, respectivamente. Desta forma, a estatística KS é dada por:

$$KS = \max |F_B(e) - F_M(e)|,$$

sendo $F_B(e)$ e $F_M(e)$ as proporções de clientes não fraudadores e fraudadores com escore menor ou igual a e, e a estatística KS é obtida através da distância máxima entre essas duas proporções acumuladas ao longo dos escores obtidos pelo modelo.

O valor desta estatística pode variar entre 0% e 100%, sendo que 100% indica uma separação total dos escores dos não fraudadores e dos fraudadores e 0% indica uma so-

breposição total das distribuições dos escores dos dois grupos. Na prática, os modelos fornecem valores intermediários entre esses dois extremos.

6.5 Teste de Adequabilidade de Ajuste de Hosmer e Lemeshow

(HOSMER e LEMESHOW, 1989) propuseram um teste para verificar a adequabilidade de ajuste do modelo de regressão logística, que pode ser estendido para a regressão logística limitada, baseado nos valores das probabilidades estimadas.

Considere n valores de probabilidades estimadas ordenadas, ou seja, a primeira probabilidade correspondendo ao menor valor e a *n-ésima* probabilidade ao maior valor. Pode-se agrupar essas probabilidades de duas formas:

1. Através dos percentis das probabilidades estimadas.

2. Através de valores fixos das probabilidades estimadas.

Com o primeiro método, considerando $g = 10$ grupos, o primeiro grupo irá conter $n'_1 = \frac{n}{10}$ observações começando com a menor probabilidade estimada. Com o segundo método, também considerando g = 10 grupos, o ponto de corte é definido no valor $\frac{k}{10}, k = 1, 2, ..., 9$ e os grupos irão conter todas as observações com probabilidade estimada entre os pontos de cortes adjacentes. Por exemplo, o primeiro grupo contém todas as observações na qual a probabilidade estimada é menor ou igual $0,1$, enquanto que o décimo grupo contém as observações cuja probabilidade estimada é maior que $0,9$. Para $y = 1$, a estimativa dos valores esperados é obtida pela soma das probabilidades estimadas de todas as observações no grupo. Para $y = 0$, a estimativa do valor esperado é um menos a probabilidade estimada de $y = 1$. Para cada estratégia de agrupamento, a estatística de adequabilidade de ajuste de Hosmer e Lemeshow $\widehat{C}$, é obtida calculando a estatística qui-quadrado de Pearson:

$$\widehat{C} = \sum_{i=1}^{g} \left[\frac{(o_k - n'_k \overline{\pi}_k)^2}{n'_k \overline{\pi}_k (1 - \overline{\pi}_k)} \right]. \tag{6.13}$$

Sendo que n'_k é o número de observações no *k-ésimo* grupo.

$$O_k = \sum_{j=1}^{n'_k} y_j, \tag{6.14}$$

é o número de respostas entre as n'_k *"covariate patters"* e:

$$\overline{\pi}_k = \sum_{j=i}^{n'_k} \frac{m_j \widehat{\pi}_j}{n'_k}, \tag{6.15}$$

a média da probabilidade estimada.

Através de um intenso estudo de simulação, (HOSMER e LEMESHOW, 1989) demonstraram que, quando $J = n$ e o modelo de regressão logística ajustado é o modelo correto, a distribuição da estatística $\widehat{C}$ é bem aproximada pela distribuição qui-quadrado com $g-2$ graus de liberdade, χ^2_{g-2}. Já para os casos em que $J \approx n$, é provável que a qui-quadrado seja também a distribuição aproximada, porém este resultado não foi examinado.

Uma alternativa para o denominador da Equação 6.13 é obtida se considerar O_k como sendo a soma da variável aleatória independente e não identicamente distribuída. Esta sugestão padroniza a diferença do quadrado entre o valor observado e a freqüência estimada esperada:

$$\sum_{j=1}^{n'_k} \left[\frac{\widehat{\pi}_j}{1 - \widehat{\pi}_j} \right]. \tag{6.16}$$

(HOSMER e LEMESHOW, 1989) mostra que:

$$\sum_{j=1}^{n} \left[\frac{\widehat{\pi}_j}{1 - \widehat{\pi}_j} \right] = n'_k \overline{\pi}_k (1 - \overline{\pi}_k) - \sum_{j=1}^{n'_k} \left(\overline{\pi}_j - \overline{\pi}_k \right)^2. \tag{6.17}$$

Com isso, encontra-se o uso da expressão 6.16 que geralmente resulta em um acréscimo insignificante no valor da estatística de teste. Assim, na prática calcula-se $\widehat{C}$ usando a equação 6.13.

Pesquisas adicionais realizadas por (HOSMER e LEMESHOW, 1989) mostraram que o método de agrupamento baseado nos percentis das probabilidades estimadas é preferível ao método com ponto de corte fixo, no sentido de melhor aderência para a distribuição qui-quadrado, especialmente quando muitas das probabilidades estimadas são pequenas (isto é, menor que 0.2).

Portanto, $\widehat{C}$ será baseado no agrupamento por percentil e com 10 grupos. Esses grupos as vezes são referidos como "*deciles of risk*". Esse termo vem de pesquisas da área de saúde, onde a resposta $y = 1$ frequentemente representa a ocorrência de alguma doença. Esta estatística, para os modelos logito, log-log e probito, pode ser obtida sem grandes dificuldades via pacotes estatísticos como STATISTICA, SAS, MINITAB e R. Para o caso do modelo logito limitado, tem que se implementar no SAS. Para isso, basta considerar a fução logito limitado nas equações acima.

Capítulo 7

Amostras do Tipo *State-dependent*

7.1 Modelo Logito com Amostras do Tipo *State-dependent*

Considere um conjunto de dados com vetor de covariáveis x_i e uma variável resposta binária Y_i, com baixa incidência de $Y_i = 1$ (evento de interesse) e uma altíssima incidência de $Y_i = 0$ (evento de não interesse).

O modelo que especifica a probabilidade de um indivíduo i possuir a característica de interesse em função do vetor de covariáveis x_i é dado por:

$$P(y_i = 1|\mathbf{x}_i) = P^*(\theta, x_i) = P_i^*. \tag{7.1}$$

O objetivo das amostras do tipo *state-dependent* é estimar o vetor de parâmetros $\boldsymbol{\beta}$ para uma amostra selecionada, descartando uma grande parte das observações $Y_i = 0$ por razões de conveniência (CRAMER, 2004).

Suponha que a amostra completa inicial é uma amostra aleatória com uma fração de amostragem α e que apenas uma fração γ de observações zero é retida. A probabilidade que o elemento i tenha $Y_i = 1$ e seja incluída na amostra é αP_i^*, mas para $Y_i = 0$ é de $\gamma\alpha(1 - P_i^*)$.

Pela regra de Bayes, a probabilidade que um elemento da amostra selecionada tenha $Y_i = 1$ é

$$\widetilde{P} = \frac{\alpha P_i^*}{\alpha P_i^* + \alpha\gamma(1 - P_i^*)} = \frac{P_i^*}{P_i^* + \gamma(1 - P_i^*)}. \tag{7.2}$$

Sabendo que $Y_i \sim Bernoulli(\widetilde{P}_i)$, então $P(Y_i = 1|x) = \widetilde{P}_i$ e $P(Y_i = 0|x) = 1 - \widetilde{P}_i$. A distribuição de Y_i pode ser representada por $P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (\widetilde{P}_i)^{y_i}(1 - \widetilde{P}_i)^{1-y_i}$ com $y_i = 0, 1$ e $i = 1, ..., n$.

Como as variáveis aleatórias Y_i são independentes, a função de verossimilhança em termos de $\widetilde{P}_i$, é dada por:

$$P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (\widetilde{P}_i)^{y_i}(1 - \widetilde{P}_i)^{1-y_i}, \qquad com \qquad y_i = 0, 1 \text{ e } i = 1, ..., n, \tag{7.3}$$

ou seja,

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} (\widetilde{P}_i)^{y_i}(1 - \widetilde{P}_i)^{1-y_i}. \tag{7.4}$$

Sendo $\widetilde{P} = \frac{P_i^*}{P_i^* + \gamma(1-P_i^*)}$ e $P(y_i = 1|\mathbf{x}_i) = P^*(\theta, x_i) = P_i^*$.

Assim $\widetilde{P}_i$ é uma função de $\boldsymbol{\beta}$ e γ, e se a amostra selecionada é extraída da amostra completa com γ conhecida, a função de verossimilhança pode ser maximizada em $\boldsymbol{\beta}$ através de um método iterativo.

7.2 Modelo Logito Limitado com Amostras do Tipo *State-dependent*

Da seção anterior sabe-se que a probabilidade de sucesso para amostras do tipo *state-dependent* é dada por:

$$\widetilde{P} = \frac{\alpha P_i}{\alpha P_i + \alpha\gamma(1 - P_i)} = \frac{P_i}{P_i + \gamma(1 - P_i)}. \tag{7.5}$$

No caso do modelo logito limitado P_i é dado por $P(y_i = 1|\mathbf{x}) = \omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}$. Assim a probabilidade de sucesso para a amostra selecionada é dada por:

$$\widetilde{P}_i = \frac{\omega \exp(\mathbf{x}_i'\boldsymbol{\beta})}{(\omega + \gamma - \omega\gamma)\exp(\mathbf{x}_i'\boldsymbol{\beta}) + \gamma}. \tag{7.6}$$

Como as variáveis aleatórias Y_i são independentes, a função de verossimilhança em termos de $\widetilde{P}_i$, é dada por:

$$P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (\widetilde{P}_i)^{y_i}(1 - \widetilde{P}_i)^{1-y_i}, \qquad com \qquad y_i = 0, 1 \text{ e } i = 1, ..., n. \tag{7.7}$$

Ou seja,

$$L(\boldsymbol{\beta}, \omega, \gamma; \mathbf{x}_i) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} (\widetilde{P}_i)^{y_i}(1 - \widetilde{P}_i)^{1-y_i}. \tag{7.8}$$

E a função log-verossimilhança é dada por:

$$l(\boldsymbol{\beta}, \omega, \gamma; \mathbf{x}_i) = \sum_{i=1}^{n} y_i \ln\left(\frac{\omega \exp(x_i'\beta)}{(\omega + \gamma + \omega\gamma)\exp(x_i'\beta) + \gamma}\right) + \tag{7.9}$$

$$+(1 - y_i)\ln\left(1 - \frac{\omega \exp(x_i'\beta)}{(\omega + \gamma + \omega\gamma)\exp(x_i'\beta) + \gamma}\right) I_{(0,1)}(\omega) I_{(0,1)}(\gamma).$$

Assim como nos casos anteriores, como tem-se um sistema de equações não lineares nos parâmetros é necessário o uso de métodos iterativos.

No próximo Capítulo será apresentado uma aplicação dos modelos logito e logito limitado considerando amostra dos tipo *state-dependent*.

Capítulo 8

Aplicações

8.1 Dados Simulados

A simulação utilizada neste capítulo com a finalidade de testar as performances dos modelos logito e logito limitado foi realizada no *software SAS,* considerando os seguintes passos:

- Passo 1: Gerou-se 6 covariáveis, x_1, x_2, x_3, x_4, x_5 e x_6 com distribuição Bernoulli com as respectivas probabilidades: 0.1, 0.8, 0.1, 0.4, 0.7 e 0.9.

- Passo 2: Considerou-se os seguintes valores para os parâmetros β: $\beta_0 = -4.5$, $\beta_1 = -3.5$, $\beta_2 = -0.8$, $\beta_3 = 1.5$, $\beta_4 = 0.7$, $\beta_5 = 1.5$ e $\beta_6 = -0.8$.

- Passo 3: Para obter a variável resposta y, primeiramente substitui-se as observações geradas e os valores dos parâmetros na seguinte expressão:

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)}.$$

- Passo 4: Gera-se y de uma Bernoulli com a probabilidade encontrada no passo anterior. Dessa forma, a variável resposta y é do tipo binária assumindo valores 0 ou 1.

Nesta simulação, foram consideradas 100.000 observações e seis covariáveis com duas categorias cada uma delas.

Das 100.000 observações, apenas 1.860 são de fraudadores o que corresponde a 1.86% da base total.

Antes da modelagem de dados é importante realizar uma análise descritiva das variáveis canditadas ao modelo. As Tabelas 8.1, 8.2, 8.3, 8.4, 8.5 e 8.6 apresentam as análises bivariadas respectivamente das variáveis x_1, x_2, x_3, x_4, x_5 e x_6. A partir desta análise é possível verificar a taxa de fraudadores e a *odds ratio* dentro de cada uma das classes.

Através da *odds ratio* é possível escolher uma classe neutra, ou *baseline,* da variável. A classe neutra será aquela que possuir a *odds ratio* mais próxima de 1, que significa que dentro desta classe as proporções de fraudadores em relação às proporções dos não fraudadores são bem próximas. Os parâmetros de regressão serão ajustados em relação a esta classe. Além de se escolher a classe neutra, com a *odds ratio* é possível verificar se a contribuição da variável será positiva ou negativa. Se a *odds ratio* for maior que 1 a contribuição da variável será positiva e se for menor que 1 será negativa.

Analisando as taxas de fraude dentro das classe de cada uma das variáveis, pode-se verificar que estas taxas variam. Esse fato indica que as variáveis possivelmente serão significativas na predição de fraudadores.

Tabela 8.1: Análise bivariada para a covariável X1

X_1	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	88182	1854	90036	90%	2,06%	90%
1	9958	6	9964	10%	0,06%	3145%
Total	98140	1860	100000	100%	100%	100%

Tabela 8.2: Análise bivariada para a covariável X2

X_2	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	19642	632	20274	20%	3,12%	59%
1	78498	1228	79726	80%	1,54%	121%
Total	98140	1860	100000	100%	1,86%	100%

Tabela 8.3: Análise bivariada para a covariável X3

X_3	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	88649	1272	89921	90%	1,41%	132%
1	9491	588	10079	10%	5,83%	31%
Total	98140	1860	100000	100%	1,86%	100%

Tabela 8.4: Análise bivariada para a covariável X4

X_4	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	58979	822	59801	60%	1,37%	136%
1	39161	1038	40199	40%	2,58%	72%
Total	98140	1860	100000	100%	1,86%	100%

Tabela 8.5: Análise bivariada para a covariável X5

X_5	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	29703	176	29879	30%	0,59%	320%
1	68437	1684	70121	70%	2,40%	77%
Total	98140	1860	100000	100%	1,86%	100%

Tabela 8.6: Análise bivariada para a covariável X6

X_6	Não Fraudadores	Fraudadores	Total	%Total	Taxa Fraude	Odds Ratio
0	9625	332	9957	10%	3,33%	55%
1	88515	1528	90043	90%	1,70%	110%
Total	98140	1860	100000	100%	1,86%	100%

8.1.1 Regressão Logística

A Tabela 8.7 apresenta os resultados das estatísticas de ajuste do modelo de regressão logística. Estas estatísticas serão utilizadas como medidas de comparação entre os desempenhos dos modelos logito e logito limitado.

Tabela 8.7: Estatísticas de Ajuste do Modelo

Critério	Apenas Intercpto	Intercepto mais covariáveis
AIC	18.509,876	16.549,467
SC	18.519,389	16.616,057
-2LogL	18.507,876	16.535,467

A Tabela 8.8 mostra as estatísticas que testam se pelo menos um dos betas da regressão é significativo. Observe, através do *p-valor*, que existem parâmetros significativos para a predição de fraudadores.

Tabela 8.8: Teste Global - Hipótese Nula: $\beta_i = 0$

Teste	Qui_quadrado	GL	p-valor
Razão de verossimilhança	1972,4093	6	<,0001
Score	2080,4962	6	<,0001
Wald	1675,2697	6	<,0001

Tabela 8.9: Análise Tipo II do Efeitos

Efeito	GL	Qui_quadrado de Wald	p-valor
X_1	1	76,3455	<,0001
X_2	1	221,8616	<,0001
X_3	1	801,3816	<,0001
X_4	1	195,9871	<,0001
X_5	1	325,5221	<,0001
X_6	1	139,2097	<,0001

Pode-se observar de acordo com a Tabela 8.9 que todos os parâmetros são significativos. Analisando a Tabela 8.10 pode-se verificar os pesos de cada um dos parâmetros na explicação da variável resposta fraude. Observe que os sinais dos pesos dos parâmetros estão de acordo com a *odds ratio*.

O modelo de regressão logística é dado pela Equação 8.1,

Tabela 8.10: Análise das Estimativas de Máxima Verossimilhança

Parâmetro	GL	Estimativa	Erro Padrão	Estatístita qui-quadrado de Wald	p-valor
$\widehat{\beta}_0$	1	-3,8279	0,0830	4027,0884	<,0001
$\widehat{\beta}_1$	1	-3,5759	0,4093	76,3455	<,0001
$\widehat{\beta}_2$	1	-0,7518	0,0505	221,8616	<,0001
$\widehat{\beta}_3$	1	1,4683	0,0519	801,3816	<,0001
$\widehat{\beta}_4$	1	0,6690	0,0478	195,9871	<,0001
$\widehat{\beta}_5$	1	1,4419	0,0799	325,5221	<,0001
$\widehat{\beta}_6$	1	-0,7410	0,0628	139,2097	<,0001

$$P(y_i = 1|\mathbf{x}) = \frac{\exp(\widehat{\boldsymbol{\beta}}\prime\mathbf{X})}{1 + \exp(\widehat{\boldsymbol{\beta}}\prime\mathbf{X})}. \tag{8.1}$$

Sendo $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5, \widehat{\beta}_6\right)$ o vetor das estimativas dos parâmetros e $\mathbf{X}$ o vetor das covariáveis.

O teste de adequabilidade de ajuste de Hosmer e Lemeshow estão sumarizados nas Tabelas 8.11 e 8.12. Observe através da estatística qui-quadrado que o modelo está adequado.

Tabela 8.11: Partições do teste de Hosmer e Lemeshow

		Fraude=0		Fraude=1	
Grupo	Total	Observado	Esperado	Observado	Esperado
1	12022	11220	11226,70	802	795
2	11094	10802	10801,69	292	292
3	17332	16974	16962,33	358	370
4	1456	1440	1439,43	16	17
5	24292	24020	24024,29	272	268
6	6540	6483	6490,22	57	50
7	7378	7345	7341,11	33	37
8	10366	10341	10338,76	25	27
9	9520	9515	9515,47	5	4

Tabela 8.12: Teste de Hosmer e Lemeshow

Qui-quadrado	GL	p-valor
2,2270	2	0,9462

Em instituições financeiras é comum a análise do escore do indivíduo. O escore é dado pela multiplicação por mil da probabilidade dada pelo modelo. Vale ressaltar que a probabilidade modelada via regressão logística nesta aplicação foi a probabilidade do cliente do cartão de crédito não fraudar. Dessa forma, quanto maior for o escore, menor a probabilidade deste indivíduo ter cometido uma transação fraudulenta. A Tabela 8.13 apresenta as faixas de escore, as distribuições de fraudadores e não fraudadores, assim como a distribuição das taxas de sinistro para cada uma das faixas. Através desta Tabela é possível determinar um ponto de corte para se classificar a transação realizada como fraude ou não.

Tabela 8.13: Faixas de escore

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Taxa de Fraude
menor que 952,8	4387	473	4860	4,9%	9,7%
952,8 -| 955,3	2917	142	3059	3,1%	4,6%
955,3 -| 976,7	6007	287	6294	6,3%	4,6%
976,7 -| 978,7	25685	550	26235	26,2%	2,1%
978,7 -| 989,1	25460	288	25748	25,7%	1,1%
989,1 -| 994,5	6483	57	6540	6,5%	0,9%
994,5 -| 994,9	6983	32	7015	7,0%	0,5%
994,9 -| 997,2	95	0	95	0,1%	0,0%
997,2 -| 997,4	10608	26	10634	10,6%	0,2%
997,4 -| 999,7	7230	5	7235	7,2%	0,1%
999,7 ou mais	2285	0	2285	2,3%	0,0%
Total	98140	1860	100000	100,0%	1,9%

8.1.2 Regressão Logística Limitada

Nesta subseção será realizada a análise do conjunto de dados via regressão logística limitada. Este modelo é dado pela Equação 6.2 e sua verossimilhança pela Equação 6.6.

A modelagem do logito limitado requer a programação das estimativas e testes, uma vez que, diferentemente do modelo logístico, não se tem implementado no SAS.

Para determinar as estimativas dos parâmetros é necessário o uso de métodos iterativos. Sendo assim, utilizou-se o procedimento NLP no *software* SAS. Neste procedimento tem-se que para um tamanho amostral (n) menor ou igual a 40 e para o método de mínimos quadrados não-linear usa-se a técnica de otimização de *Newton-Rapphson,* por *default.*

Para amostras de tamanho $40 < n < 400$, usa-se por *default* o método Quasi-Newton. Já para amostras de tamanhos maiores que 400, usa-se Conjugate Gradient.

A Tabela 8.14 apresenta os valores estimados para os parâmetros do modelo e suas respectivas estatísticas de Wald. Através desta estatística é possível verificar a significância dos parâmetros. Comparando os valores das estísticas de Wald com uma distribuição normal padrão, com 95% de confiança pode-se verificar que todos parâmetros são significativos.

Tabela 8.14: Parâmetros estimatimados via método de Newton-Rapson

Parâmetros	Estimativa	Erro Padrão Aproximado	Estatística de Wald
ϖ	0,564093	0,269130	2,095987069
β_0	-3,844386	0,518504	-7,414380603
β_1	-3,617210	0,412506	-8,768866392
β_2	-0,785620	0,064146	-12,24737318
β_3	1,534156	0,091142	16,8325909
β_4	0,691860	0,055255	12,5212198
β_5	1,475508	0,089201	16,54138407
β_6	-0,776661	0,076270	-10,18304707

A Tabela 8.15 apresenta as estatísticas de ajuste do modelo.

Tabela 8.15: Estatísticas de ajuste do modelo de logito limitado

Critério	Apenas Intercpto	Intercepto mais covariáveis
AIC	18845,25	16548,63
SC	18854,76	16615,22
-2LogL	18843,25	16534,63

Para o modelo logito limitado programou-se o teste de adequabilidade de ajuste de Hosmer e Lemeshow. As Tabelas 8.16 e 8.17 apresentam os resultados deste teste. Através do $p-valor$ do teste, pode-se verificar que o ajuste do modelo logito limitado é adequado.

A Tabela 8.18 apresenta as faixas de escore para este modelo.

8.1.3 Amostras do Tipo *State-dependent*

A amostra do tipo *state-dendent* será utilizada para verificar a importância de se utilizar o conjunto de dados completo para a modelagem dos dados.

Tabela 8.16: Partições do teste de Hosmer e Lemeshow

		Fraude=0		Fraude=1	
Grupo	Total	Observado	Esperado	Observado	Esperado
1	12022	11220	11244	802	778
2	8333	8089	8113	244	220
3	20093	19687	19689	406	404
4	27563	27257	27287	306	276
5	3545	3521	3523	24	22
6	8558	8510	8519	48	39
7	10366	10341	10343	25	23
8	9520	9515	9515	5	5

Tabela 8.17: Teste de Hosmer e Lemeshow

Qui-quadrado	GL	p-valor
6,72	2	0,0814

Tabela 8.18: Faixas de escore para o modelo logito limitado

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Taxa de Fraude
menor que 952,8	4555	478	5033	5,0%	9,5%
952,8 -| 955,3	6665	324	6989	7,0%	4,6%
955,3 -| 976,7	8089	244	8333	8,3%	2,9%
976,7 -| 978,7	19687	406	20093	20,1%	2,0%
978,7 -| 989,1	27257	306	27563	27,6%	1,1%
989,1 -| 994,5	3521	24	3545	3,5%	0,7%
994,5 -| 994,9	8148	47	8195	8,2%	0,6%
994,9 -| 997,2	362	1	363	0,4%	0,3%
997,2 -| 997,4	10341	25	10366	10,4%	0,2%
997,4 -| 999,7	7154	5	7159	7,2%	0,1%
999,7 ou mais	2361	0	2361	2,4%	0,0%
Total	98140	1860	100000	100,0%	1,9%

Como visto no Capítulo 7, as amostras do tipo *state-dependent* considera a proporção de zeros utilizados e as probabilidades das observações saírem na amostra. Cramer [1] mostra que quando se trata de conjuntos de dados extremamente desbalanceados, o uso de amostras balanceadas na estimativa dos parâmetros não fornecem resultados consistentes.

Para realizar as estimativas dos parâmetros para o conjunto de dados simulados, considerou-se 8 amostras de diferentes tamanhos de observações de não fraudadores.

A Tabela 8.19 apresenta os resultados para o modelo logito considerando as amostras do tipo state-dependent. Para cada um dos casos retirou-se uma amostra aleatória simples.

Tabela 8.19: Estimativas do modelo logito para amostras do tipo *state-dependent*

	k = 1	k = 2	k = 3	k = 5	k = 10	k = 20	k = 30	k = 50	Total
Fraude = 1	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860
Fraude = 0	1.860	3.720	5.580	9.300	18.600	37.200	55.800	93.000	98.140
Total da amostra	3.720	5.580	7.440	11.160	20.460	39.060	57.660	94.860	100.000
γ	0,019	0,040	0,057	0,0948	0,189	0,379	0,569	0,948	1
β_0	-4,413	-4,426	-4,550	-4,409	-4,490	-4,423	-4,476	-4,452	-4,446
β_1	-3,691	-3,687	-3,494	-3,585	-3,523	-3,541	-3,574	-3,570	-3,576
β_2	-0,832	-0,738	-0,717	-0,781	-0,750	-0,747	-0,724	-0,750	-0,752
β_3	1,463	1,496	1,556	1,487	1,443	1,435	1,462	1,470	1,468
β_4	0,651	0,653	0,672	0,646	0,689	0,649	0,687	0,670	0,669
β_5	1,505	1,445	1,501	1,420	1,464	1,429	1,448	1,443	1,442
β_6	-0,764	-0,755	-0,719	-0,730	-0,728	-0,745	-0,745	-0,737	-0,741

Com o intuíto de retirar um possível vício entre as amostras, retirou-se 100 amostras para cada um dos casos e como estimativas dos parâmetros considerou-se a média amostral dos 100 parâmetros estimados. A Tabela 8.20 apresenta estes resultados.

As Tabelas 8.19 e 8.20 apresentam as estimativas do modelo Logito para uma amostar e replicando 100 amostras.

Pode-se observar através das Tabelas 8.21 e 8.22 que não existe diferença significativa entre as estimativas dos parâmetros da amostra balanceada ($k = 1$) em relação as estimativas da amostra completa (*total*), resultado divergente ao apresentado por Cramer.

A Tabela 8.23 apresenta as estimativas considerando o modelo logito limitado e as amostras do tipo *state-dependent*. A Tabela 8.24 apresenta as estimativas do modelo logito limitado considerando 100 amostras para cada caso.

Tabela 8.20: Estimativas do modelo logito para 100 amostras de cada caso do tipo *state-dependent*

	k = 1	k = 2	k = 3	k = 5	k = 10	k = 20	k = 30	k = 50	Total
Fraude = 1	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860
Fraude = 0	1.860	3.720	5.580	9.300	18.600	37.200	55.800	93.000	98.140
Total da amostra	3.720	5.580	7.440	11.160	20.460	39.060	57.660	94.860	100.000
γ	0,019	0,038	0,057	0,095	0,190	0,379	0,567	0,9476	1
β_0	-4,362	-4,387	-4,392	-4,415	-4,434	-4,435	-4,437	-4,445	-4,446
β_1	-3,578	-3,572	-3,574	-3,568	-3,574	-3,573	-3,575	-3,576	-3,576
β_2	-0,795	-0,782	-0,781	-0,770	-0,763	-0,757	-0,757	-0,752	-0,752
β_3	1,472	1,462	1,470	1,468	1,467	1,466	1,470	1,468	1,468
β_4	0,646	0,651	0,659	0,658	0,664	0,664	0,668	0,669	0,669
β_5	1,442	1,439	1,435	1,441	1,444	1,440	1,444	1,442	1,442
β_6	-0,787	-0,770	-0,767	-0,754	-0,746	-0,745	-0,748	-0,742	-0,741

Tabela 8.21: Razões Modelo Logito

Parâmetros	k2/k1	k3/k1	k5/k1	k10/k1	k20/k1	k50/k1	Total/k1
β_0	1,00	1,03	1,00	1,02	1,00	1,01	1,01
β_1	1,00	0,95	0,97	0,95	0,96	0,97	0,97
β_2	0,89	0,86	0,94	0,90	0,90	0,87	0,90
β_3	1,02	1,06	1,02	0,99	0,98	1,00	1,00
β_4	1,00	1,03	0,99	1,06	1,00	1,06	1,03
β_5	0,96	1,00	0,94	0,97	0,95	0,96	0,96
β_6	0,99	0,94	0,96	0,95	0,98	0,98	0,97

Tabela 8.22: Razões Modelo Logito - 100 amostras

Parâmetros	k2/k1	k3/k1	k5/k1	k10/k1	k20/k1	k50/k1	Total/k1
β_0	1,01	1,01	1,01	1,02	1,02	1,02	1,02
β_1	1,00	1,00	1,00	1,00	1,00	1,00	1,00
β_2	0,98	0,98	0,97	0,96	0,95	0,95	0,95
β_3	0,99	1,00	1,00	1,00	1,00	1,00	1,00
β_4	1,01	1,02	1,02	1,03	1,03	1,04	1,04
β_5	1,00	1,00	1,00	1,00	1,00	1,00	1,00
β_6	0,99	0,97	0,96	0,95	0,95	0,95	0,94

Tabela 8.23: Estimativas do modelo logito limitado para amostras do tipo *state-dependent*

	k = 1	k = 2	k = 3	k = 5	k = 10	k = 20	k = 30	k = 50	Total
Fraude = 1	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860
Fraude = 0	1.860	3.720	5.580	9.300	18.600	37.200	55.800	93.000	98.140
Total da amostra	3.720	5.580	7.440	11.160	20.460	39.060	57.660	94.860	100.000
ω	0,409	0,411	0,609	0,526	0,599	0,420	0,624	0,559	0,564
β_0	-4,413	-3,501	-4,036	-3,736	-4,490	-3,502	-3,984	-3,841	-3,844
β_1	-3,691	-3,759	-3,522	-3,632	-3,523	-3,610	-3,605	-3,612	-3,617
β_2	-0,832	-0,778	-0,740	-0,813	-0,750	-0,806	-0,749	-0,784	-0,786
β_3	1,463	1,601	1,606	1,554	1,443	1,547	1,511	1,538	1,534
β_4	0,651	0,686	0,687	0,667	0,689	0,687	0,704	0,693	0,692
β_5	1,505	1,496	1,525	1,453	1,464	1,487	1,474	1,477	1,476
β_6	-0,764	-0,798	-0,739	-0,762	-0,727	-0,807	-0,769	-0,773	-0,777

Tabela 8.24: Estimativas do modelo logito limitado para 100 amostras do tipo *state-dependent*

	k = 1	k = 2	k = 3	k = 5	k = 10	k = 20	k = 30	k = 50	Total
Fraude = 1	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860	1.860
Fraude = 0	1.860	3.720	5.580	9.300	18.600	37.200	55.800	93.000	98.140
Total da amostra	3.720	5.580	7.440	11.160	20.460	39.060	57.660	94.860	100.000
ω	0,449	0,455	0,447	0,485	0,502	0,508	0,515	0,568	0,564
β_0	-3,945	-4,013	-4,016	-3,954	-4,012	-3,875	-3,911	-3,850	-3,844
β_1	-3,610	-3,598	-3,600	-3,601	-3,603	-3,613	-3,610	-3,617	-3,617
β_2	-0,816	-0,800	-0,800	-0,794	-0,785	-0,789	-0,786	-0,785	-0,786
β_3	1,518	1,500	1,509	1,516	1,511	1,529	1,527	1,533	1,534
β_4	0,659	0,663	0,671	0,673	0,678	0,685	0,687	0,691	0,692
β_5	1,464	1,458	1,454	1,465	1,467	1,472	1,473	1,475	1,476
β_6	-0,809	-0,789	-0,786	-0,778	-0,769	-0,778	-0,778	-0,777	-0,777

Tabela 8.25: Razões Modelo Logito Limitado

Parâmetros	k2/k1	k3/k1	k5/k1	k10/k1	k20/k1	k50/k1	Total/k1
ω	1,00	1,49	1,29	1,46	1,03	1,37	1,38
β_0	0,79	0,91	0,85	1,02	0,79	0,87	0,87
β_1	1,02	0,95	0,98	0,95	0,98	0,98	0,98
β_2	0,94	0,89	0,98	0,90	0,97	0,94	0,94
β_3	1,09	1,10	1,06	0,99	1,06	1,05	1,05
β_4	1,05	1,06	1,02	1,06	1,06	1,06	1,06
β_5	0,99	1,01	0,97	0,97	0,99	0,98	0,98
β_6	1,04	0,97	1,00	0,95	1,06	1,01	1,02

Tabela 8.26: Razões Modelo Logito Limitado 100 amostras

Parâmetros	k2/k1	k3/k1	k5/k1	k10/k1	k20/k1	k50/k1	Total/k1
ω	1,01	1,00	1,08	1,12	1,13	1,27	1,26
β_0	1,02	1,02	1,00	1,02	0,98	0,98	0,97
β_1	1,00	1,00	1,00	1,00	1,00	1,00	1,00
β_2	0,98	0,98	0,97	0,96	0,97	0,96	0,96
β_3	0,99	0,99	1,00	1,00	1,01	1,01	1,01
β_4	1,01	1,02	1,02	1,03	1,04	1,05	1,05
β_5	1,00	0,99	1,00	1,00	1,01	1,01	1,01
β_6	0,98	0,97	0,96	0,95	0,96	0,96	0,96

Da mesma forma como foi visto para o modelo logito, pode-se observar através das razões apresentadas nas Tabelas 8.25 e 8.22 que não faz diferença considerar o conjunto de dados completo e uma amostra balanceada, ou apenas parte dos dados de não fraudadores.

8.1.4 Comparação dos modelos

Através da análise dos modelos considerando as amostras do tipo *state-dependent*, pode-se concluir que deve-se utilizar o conjunto de dados completo para a análise.

Comparando as estatísticas de teste dadas pelos dois modelos, pode-se verificar que o valor para a estatística AIC é menor para o modelo logito limitado, o que indica que este modelo é melhor em relação ao logito usual. Observando as estatísticas $-2logL$ (menor valor) e SC (menor valor), confirma-se que o modelo logito limitado deve ser escolhido.

Através das Tabelas 8.13 e 8.18 pode-se atribuir um ponto de corte observando as taxas de fraude em cada classe de escore. Se for considerado um ponto de corte de $955,3$, pode-se verificar que a taxa de fraude dada para esta carteira irá cair de $1,86\%$ para $1,20\%$, se considerado o modelo logito limitado. Considerando o modelo logito, a taxa de fraude caí de $1,86\%$ para $1,35\%$. Apesar de aparentemente ser pouca a diferença entre as taxas, pode-se verificar o impacto dessas através do estudo do lucro que a instituição teria com a carteira.

Segundo informações dadas por um analista de modelagem, sabe-se que em média uma instituição financeira ganha com um cliente não fraudador R$30,00. Essa receita provém da anuidade cobrada mais as taxas impostas sobre a transação. Por outro lado, com um cliente fraudador a instituição perde todo o valor gasto do limite do cartão. Levando em

Tabela 8.27: Análise de perda e ganho - Classificação dada pelo modelo logito

Faixas de Escore	Ganho em Reais	Perda em Reais	Lucro	Lucro Acumulado
menor que 952,8	2.944.200	3.720.000	-775.800	3.975.840
952,8 -| 955,3	2.812.590	2.774.000	38.590	4.751.640
955,3 -| 976,7	2.725.080	2.490.000	235.080	4.713.050
976,7 -| 978,7	2.544.870	1.916.000	628.870	4.477.970
978,7 -| 989,1	1.774.320	816.000	958.320	3.849.100
989,1 -| 994,5	1.010.520	240.000	770.520	2.890.780
994,5 -| 994,9	816.030	126.000	690.030	2.120.260
994,9 -| 997,2	606.540	62.000	544.540	1.430.230
997,2 -| 997,4	603.690	62.000	541.690	885.690
997,4 -| 999,7	285.450	10.000	275.450	344.000
999,7 ou mais	68.550	0	68.550	68.550

Tabela 8.28: Análise de perda e ganho acumulados - Classificação dada pelo modelo logito limitado

Faixas de Escore	Ganho em Reais	Perda em Reais	Lucro	Lucro Acumulado
menor que 952,8	2.944.200	3.720.000	-775.800	3.975.840
952,8 -| 955,3	2.807.550	2.764.000	43.550	5.104.490
955,3 -| 976,7	2.607.600	2.116.000	491.600	5.060.940
976,7 -| 978,7	2.364.930	1.628.000	736.930	4.569.340
978,7 -| 989,1	1.774.320	816.000	958.320	3.832.410
989,1 -| 994,5	956.610	204.000	752.610	2.874.090
994,5 -| 994,9	850.980	156.000	694.980	2.121.480
994,9 -| 997,2	606.540	62.000	544.540	1.426.500
997,2 -| 997,4	595.680	60.000	535.680	881.960
997,4 -| 999,7	285.450	10.000	275.450	346.280
999,7 ou mais	70.830	0	70.830	70.830

conta que cada cliente possui um limite médio de R$2.000,00, construiu-se as Tabelas 8.27 e 8.28 que apresenta o quanto a instituição está ganhando e perdendo por faixa de escore para cada um dos ajustes.

Com o ponto de corte de $955,3$, pode-se obsevar também que o modelo logito limitado classifica melhor os indivíduos dentro das classes de escore, pois nas duas faixas rejeitadas estão 802 indivíduos fraudadores enquanto que com o modelo logito estão 615. Por outro lado, com o modelo logito rejeitaria-se 7.304 clientes não fraudadores e com o logito limitado 11.220. Porém, através das Tabela 8.27 e 8.28 que apresentam também as somas dos ganhos e perdas por faixas de escore, pode-se observar que o impacto de se rejeitar

mais clientes fraudadores é melhor do que deixar de rejeitar os clientes não fraudadores, pois com a classificação dada pelo modelo logito limitado a instituição estaria ganhando mais dinheiro em relação a classificação do modelo logito usual.

8.2 Dados Reais

Nesta seção será realizada a análise de um conjunto de dados reais de fraude em cartão de crédito disponibilizada por uma grande instuição finaceira. Por se tratar de dados confidencias, todas as variáveis serão tratadas com nomes fictícios. A amostra utilizada para o ajuste dos modelos de regressão logística e regressão logística limitada é constituída de 172.452 observações. Destas, apenas 2.234 são de transações fraudulentas, o que corresponde a $1,30\%$.

A base de dados em questão possui onze variáveis explicativas, além da variável resposta dicotômica, que indica se a transação foi fraudelenta ou não.

Todas as variáveis explicativas foram categorizadas. Primeiramente todas as variáveis são divididas em 10 classes (percentis). Após várias análises bivariadas, chegou-se a categorização final. As Tabelas 8.29, 8.30, 8.31, 8.32, 8.33, 8.34, 8.35, 8.36, 8.37 e 8.38 apresentam as análises bivariadas finais para cada uma das variáveis.

Tabela 8.29: Análise bivariada para a covariável X1

X_1	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	27200	341	27541	16%	1,24%	105%
2	89861	779	90640	53%	0,86%	151%
3	53157	1114	54271	31%	2,05%	63%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.30: Análise bivariada para a covariável X2

X_2	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	4966	36	5002	3%	0,72%	181%
2	72523	757	73280	42%	1,03%	126%
3	11846	225	12071	7%	1,86%	69%
4	80883	1216	82099	48%	1,48%	87%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.31: Análise bivariada para a covariável X3

X_3	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	22688	548	23236	13%	2,36%	54%
2	27888	531	28419	16%	1,87%	69%
3	33912	460	34372	20%	1,34%	97%
4	51580	454	52034	30%	0,87%	149%
5	34150	241	34391	20%	0,70%	186%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.32: Análise bivariada para a covariável X4

X_4	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	5363	45	5408	3%	0,83%	156%
2	15410	155	15565	9%	1,00%	130%
3	94657	1183	95840	56%	1,23%	105%
4	54788	851	55639	32%	1,53%	84%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.33: Análise bivariada para a covariável X5

X_5	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	77551	1241	78792	46%	1,58%	82%
2	92667	993	93660	54%	1,06%	122%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.34: Análise bivariada para a covariável X6

X_6	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	19265	500	19765	11%	2,53%	51%
2	25652	448	26100	15%	1,72%	75%
3	94322	1039	95361	55%	1,09%	119%
4	30979	247	31226	18%	0,79%	165%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.35: Análise bivariada para a covariável X7

X_7	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	28031	604	28635	17%	2,11%	61%
2	17270	297	17567	10%	1,69%	76%
3	21250	298	21548	12%	1,38%	94%
4	43692	480	44172	26%	1,09%	119%
5	45773	447	46220	27%	0,97%	134%
6	14202	108	14310	8%	0,75%	173%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.36: Análise bivariada para a covariável X8

X_8	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	59157	440	59597	35%	0,74%	176%
2	10049	170	10219	6%	1,66%	78%
3	4971	121	5092	3%	2,38%	54%
4	96041	1503	97544	57%	1,54%	84%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.37: Análise bivariada para a covariável X9

X_9	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	27752	1044	28796	17%	3,63%	35%
2	142466	1190	143656	83%	0,83%	157%
Total	170218	2234	172452	100%	1,30%	100%

Tabela 8.38: Análise bivariada para a covariável X10

X_{10}	Bons	Maus	Total	%Total	Taxa Fraude	Odds Ratio
1	31090	276	31366	18%	0,88%	148%
2	94085	1026	95111	55%	1,08%	120%
3	25820	444	26264	15%	1,69%	76%
4	11523	247	11770	7%	2,10%	61%
5	7700	241	7941	5%	3,03%	42%
Total	170218	2234	172452	100%	1,30%	100%

8.2.1 Regressão Logística

Da mesma forma como foi feito na seção passada com os dados simulados, primeiramente ajustou-se um modelo de regressão logística. Com as tabelas bivariadas, determinou-se as classes de referência para cada uma das variáveis. A probabilidade modelada foi a do indivíduo não ser um fraudador.

Testando todas as variáveis, pode observar através da Tabela 8.39 que as variáveis X_6 e X_8 são não significativas.

Tabela 8.39: Análise Tipo II do Efeitos - modelo de regressão logística

Efeito	GL	Qui_quadrado de Wald	p-valor
X1	2	81,66	<,0001
X2	3	15,20	0,0017
X3	4	102,10	<,0001
X4	3	19,61	0,0002
X5	1	25,65	<,0001
X6	3	1,63	0,652
X7	3	42,39	<,0001
X8	5	8,86	0,1149
X9	1	626,82	<,0001
X10	4	90,85	<,0001

A Tabela 8.41 apresenta as estimativas dos parâmetros após a retirada das duas variáveis não significativas. Pode-se verificar as estimativas dos parâmetros através da Tabela 8.41.

Tabela 8.40: Análise de Efeitos Tipo II - modelo de regressão logística

Efeito	GL	Qui_quadrado de Wald	p-valor
X1	2	89,18	<,0001
X2	3	14,91	0,0019
X3	4	181,50	<,0001
X4	3	15,76	0,0013
X5	1	26,38	<,0001
X7	3	50,68	<,0001
X9	1	731,70	<,0001
X10	4	96,51	<,0001

A Tabela 8.42 apresenta as estatísticas AIC, SC e -2LogL.

Tabela 8.41: Análise das estimativas de máxima verossimilhança - modelo de regressão logística

Parâmetro	Classe	GL	Estimativa	Erro Padrão	Teste de Wald	p-valor
Intercepto		1	4,25	0,10	1899,82	<,0001
X1	2	1	0,35	0,07	28,90	<,0001
X1	3	1	-0,13	0,07	3,96	0,0465
X2	1	1	0,51	0,17	8,84	0,0029
X2	2	1	-0,10	0,05	3,92	0,0478
X2	3	1	-0,09	0,07	1,32	0,2497
X3	1	1	-0,55	0,07	64,21	<,0001
X3	2	1	-0,20	0,07	8,96	0,0028
X3	4	1	0,29	0,07	18,86	<,0001
X3	5	1	0,41	0,08	24,09	<,0001
X4	1	1	0,28	0,15	3,29	0,0697
X4	2	1	0,07	0,09	0,59	0,4417
X4	4	1	0,18	0,05	13,95	0,0002
X5	2	1	0,23	0,04	26,38	<,0001
X7	1	1	0,41	0,06	50,65	<,0001
X7	2	1	0,07	0,08	0,74	0,3909
X7	3	1	0,09	0,10	0,93	0,3361
X9	1	1	-1,33	0,05	731,70	<,0001
X10	1	1	0,47	0,08	35,43	<,0001
X10	3	1	0,16	0,06	6,86	0,0088
X10	5	1	-0,10	0,08	1,62	0,2037
X10	6	1	-0,37	0,08	20,36	<,0001

Tabela 8.42: Estatísticas de ajuste do modelo de logito

Estatísticas de Ajuste do Modelo		
Critério	Apenas Intercpto	Intercepto mais covariáveis
AIC	23860,31	22088,15
SC	23870,37	22309,42
-2LogL	23858,31	22044,15

As Tabelas 8.43 e 8.44 apresentam os resultados do teste de Hosmer e Lemeshow. Pode-se observar através do p-valor do teste que o ajuste do modelo logito é adequado para os dados em análise.

Tabela 8.43: Partições do teste de Hosmer e Lemeshow

		Fraude=0		Fraude=1	
Grupo	Total	Observado	Esperado	Observado	Esperado
1	17231	16402	16402,90	829	828,10
2	16981	16612	16608,36	369	372,64
3	17211	16930	16957,28	281	253,72
4	17249	17059	17061,89	190	187,11
5	17244	17101	17096,21	143	147,79
6	17094	16989	16971,94	105	122,06
7	17762	17660	17654,89	102	107,11
8	17263	17177	17174,17	86	88,83
9	17075	17008	17001,74	67	73,26
10	17342	17280	17288,61	62	53,39

Tabela 8.44: Teste de Hosmer e Lemeshow

Qui-quadrado	GL	p-valor
7,88	8	0,4448

O teste Kolmogorov-Smirnov (KS) é usado para determinar se duas distribuições de probabilidade subjacentes diferem uma da outra. Nesta análise, o valor do KS mede a distância entre as curvas das distribuições dos fraudadores e dos não fraudadores. Para o ajuste do modelo de regressão logística tem-se um KS de 51%.

Para a análise de ponto de corte, dividiu-se os escores em 20 classes de escore com 5% de indivíduos em cada classe. A Tabela 8.45 mostra as faixas de escore e as respectivas taxas de fraude.

Observando as taxas de fraude, pode-se atribuir medidas de decisões por faixas de escore. Observe que indivíduos com escores menores que $972,2$ possuem chances maiores de cometer uma transação fraudulenta. Dessa forma, novos clientes classificados com escore abaixo desse valor podem ser rejeitados. Medidas de prevenção devem ser tomadas com muito detalhe para estes casos. Indivíduos com escore entre $989,2$ e $972,2$ devem ser tratados com muita atenção, pois estão propícios a realizar uma transação fraudulenta.

Tabela 8.45: Faixas de escore para o modelo logito

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Taxa de Fraude
menor que 957,6	8145	519	8664	5%	5,99%
957,6 -| 972,2	8352	311	8663	5%	3,59%
972,2 -| 978,3	8344	203	8547	5%	2,38%
978,3 -| 982,5	8838	181	9019	5%	2,01%
982,5 -| 985,4	8105	148	8253	5%	1,79%
985,4 -| 987,6	8481	122	8603	5%	1,42%
987,6 -| 989,2	8519	109	8628	5%	1,26%
989,2 -| 990,6	8963	85	9048	5%	0,94%
990,6 -| 991,4	8104	77	8181	5%	0,94%
991,4 -| 992,2	8559	61	8620	5%	0,71%
992,2 -| 992,9	8637	58	8695	5%	0,67%
992,9 -| 993,4	8513	44	8557	5%	0,51%
993,4 -| 993,9	8886	60	8946	5%	0,67%
993,9 -| 994,3	8295	41	8336	5%	0,49%
994,3 -| 994,9	9663	52	9715	5%	0,54%
994,9 -| 995,2	7526	34	7560	5%	0,45%
995,2 -| 995,7	8535	34	8569	5%	0,40%
995,7 -| 996,2	8827	35	8862	5%	0,39%
996,2 -| 996,8	8348	32	8380	5%	0,38%
996,8 ou mais	8578	28	8606	5%	0,33%
Total	170218	2234	172452	5%	1,30%

8.2.2 Regressão Logística Limitada

No caso do modelo logito limitado, primeiramente construiu-se as variáveis *dummys* para que fosse possível determinar os estimadores de máxima verossimilhança.

A Tabela 8.46 apresenta as estimativas dos parâmetros.

Comparando as estatísticas de Wald com uma distribuição normal padrão, pode-se verificar que apenas a variável X_6 é não significativa. Retirando esta variável, calculou-se a probabilidade do indivíduo não fraudar.

Através da Tabela 8.47 pode-se verificar que o ajuste do modelo logito limitado é adequado aos dados. A estatística KS calculada para este modelo é de $52,56\%$.

Da mesma forma feita no ajuste do modelo logito, dividiu-se os escores em 20 classes de escore com 5% de indivíduos em cada classe. A Tabela 8.49 mostra as faixas de escore e as respectivas taxas de sinistro.

Neste caso um ponto de corte que rejeitaria um bom número de indivíduos fraudadores seria 977. Medidas anti-fraude devem ser tomadas detalhadamente com os indivíduos que possuirem escores entre $989,2$ e $977,0$.

8.2.3 Amostras do Tipo State-dependent

Através da Tabela 8.50 pode-se observar a diferença entre as estimativas dos parâmetros em relação o conjunto de dados completo e a amostra balanceada. Dessa forma, concluí-se que deve-se utilizar o conjunto de dados completo.

8.2.4 Comparação dos modelos

Comparando as estatísticas calculadas para os dois modelos, AIC, -2LogL, SC e KS, pode-se observar que todas elas mostram que o ajuste do modelo logito limitado é melhor do que o modelo logito para este conjunto de dados. Aplicando os dois modelos em um conjunto de dados de validação com 16.388 observações, sendo destas 386 fraudulentas observou-se um KS de $54,3\%$ para o modelo logito e $56,01\%$ para o modelo logito limitado, o que evidencia que o modelo logito limitado discrimina melhor os fraudadores dos não fraudadores. Observe nas Tabelas as distribuições dos indivíduos dentro das faixas de escore.

Tabela 8.46: Análise das estimativas de máxima verossimilhança - modelo de regressão logística limitada

Efeito	Parâmetro	Estimativa	Estatística de Wald
ω	$\widehat{\varpi}$	0,15	0,89
intercepto	$\widehat{\beta}_0$	-2,12	-0,47
X_1	$\widehat{\beta}_1$	-0,07	-3,46
X_1	$\widehat{\beta}_2$	-0,08	-0,07
X_2	$\widehat{\beta}_3$	-0,49	-3,40
X_2	$\widehat{\beta}_4$	0,21	1,28
X_2	$\widehat{\beta}_5$	0,20	1,46
X_3	$\widehat{\beta}_6$	0,71	9,37
X_3	$\widehat{\beta}_7$	0,18	1,56
X_3	$\widehat{\beta}_8$	-0,05	-0,90
X_3	$\widehat{\beta}_9$	-0,36	-3,69
X_4	$\widehat{\beta}_{10}$	-0,18	-1,35
X_4	$\widehat{\beta}_{11}$	0,08	0,88
X_4	$\widehat{\beta}_{12}$	0,18	3,03
X_5	$\widehat{\beta}_{13}$	0,12	1,50
X_6	$\widehat{\beta}_{14}$	0,08	2,18
X_6	$\widehat{\beta}_{15}$	0,04	2,61
X_6	$\widehat{\beta}_{16}$	-0,04	-3,84
X_7	$\widehat{\beta}_{17}$	0,15	1,17
X_7	$\widehat{\beta}_{18}$	0,10	1,34
X_7	$\widehat{\beta}_{19}$	-0,02	-0,32
X_7	$\widehat{\beta}_{20}$	0,02	0,29
X_7	$\widehat{\beta}_{21}$	-0,10	-1,79
X_8	$\widehat{\beta}_{22}$	-0,30	-4,91
X_8	$\widehat{\beta}_{23}$	0,07	0,90
X_8	$\widehat{\beta}_{24}$	0,10	1,06
X_9	$\widehat{\beta}_{25}$	0,79	1,12
X_{10}	$\widehat{\beta}_{26}$	-0,51	-1,67
X_{10}	$\widehat{\beta}_{27}$	-0,14	-3,14
X_{10}	$\widehat{\beta}_{28}$	-0,02	0,46
X_{10}	$\widehat{\beta}_{29}$	-0,14	2,03

Tabela 8.47: Teste de Hosmer e Lemeshow

Qui-quadrado	GL	p-valor
5,4	28	0,1447

Tabela 8.48: Estatísticas de ajuste do modelo de logito limitado

Estatísticas de Ajuste do Modelo		
Critério	Apenas Intercpto	Intercepto mais covariáveis
AIC	24143,98	22858,67
SC	23143,65	21945,92
-2LogL	23965,33	21907,77

Tabela 8.49: Faixas de escore para o modelo logito limitado

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Taxa de Fraude
menor que 956,9	8128	506	8634	5%	5,86%
956,9 -| 969,9	8306	307	8613	5%	3,56%
969,9 -| 977,0	8490	209	8699	5%	2,40%
977,0 -| 981,6	8367	180	8547	5%	2,11%
981,6 -| 985,1	8473	150	8623	5%	1,74%
985,1 -| 987,4	8493	136	8629	5%	1,58%
987,4 -| 989,2	8521	100	8621	5%	1,16%
989,2 -| 990,5	8600	83	8683	5%	0,96%
990,5 -| 991,5	8478	84	8562	5%	0,98%
991,5 -| 992,4	8552	63	8615	5%	0,73%
992,4 -| 993,0	8567	56	8623	5%	0,65%
993,0 -| 993,5	8566	57	8623	5%	0,66%
993,5 -| 994,0	8795	58	8853	5%	0,66%
994,0 -| 994,5	8348	46	8394	5%	0,55%
994,5 -| 994,9	8585	35	8620	5%	0,41%
994,9 -| 995,4	8580	43	8623	5%	0,50%
995,4 -| 995,8	8620	28	8648	5%	0,32%
995,8 -| 996,3	8561	37	8598	5%	0,43%
996,3 -| 996,9	8593	29	8622	5%	0,34%
996,9 ou mais	8595	27	8622	5%	0,31%
Total	170218	2234	172452	100%	1,30%

Tabela 8.50: Estimativas do modelo logito para amostras do tipo *state-dependent*

	k = 1	k = 2	k = 3	k = 5	k = 10	k = 20	k = 50	Total
Fraude = 1	2.234	2.234	2.234	2.234	2.234	2.234	2.234	2.234
Fraude = 0	2.234	4.468	6.702	11.170	22.340	44.680	111.700	170.218
Total da amostra	4.468	6.702	8.936	13.404	24.574	46.914	113.934	172.452
γ	0,0131	0,0262	0,0394	0,0656	0,1312	0,2625	0,6562	1
β_0	3,891	3,763	3,922	3,984	3,999	3,977	4,198	4,247
β_1	0,332	0,340	0,347	0,340	0,345	0,346	0,351	0,354
β_2	-0,134	-0,123	-0,157	-0,168	-0,158	-0,160	0,139	-0,133
β_3	0,398	0,456	0,534	0,498	0,454	0,434	0,498	0,509
β_4	-0,067	-0,067	-0,109	-0,112	-0,089	-0,098	-0,100	-0,098
β_5	-0,099	-0,110	-0,165	-0,123	-0,090	-0,077	-0,067	-0,086
β_6	-0,434	-0,488	-0,399	-0,402	-0,456	-0,468	-0,499	-0,553
β_7	-0,146	-0,157	-0,187	-0,189	-0,193	-0,209	-0,204	-0,196
β_8	0,309	0,318	0,320	0,309	0,329	0,308	0,297	0,292
β_9	0,398	0,387	0,376	0,377	0,367	0,397	0,439	0,405
β_{10}	0,387	0,398	0,365	0,357	0,318	0,296	0,310	0,280
β_{11}	0,035	0,043	0,065	0,057	0,049	0,059	0,078	0,069
β_{12}	0,168	0,156	0,168	0,157	0,168	0,178	0,176	0,178
β_{13}	0,284	0,287	0,274	0,256	0,249	0,290	0,267	0,227
β_{14}	0,465	0,487	0,423	0,484	0,483	0,463	0,456	0,415
β_{15}	0,034	0,056	0,054	0,067	0,076	0,076	0,087	0,071
β_{16}	0,127	0,119	0,093	0,105	0,105	0,117	0,104	0,094
β_{17}	-1,540	-1,092	-1,653	-1,293	-1,233	-1,432	-1,402	-1,329
β_{18}	0,455	0,350	0,398	0,403	0,456	0,499	0,498	0,469
β_{19}	0,122	0,163	0,153	0,165	0,176	0,187	0,152	0,155
β_{20}	-0,102	-0,092	-0,052	-0,074	-0,054	-0,826	-0,827	-0,104
β_{21}	-0,475	-0,425	-0,458	-0,395	-0,408	-0,384	-0,375	-0,373

Tabela 8.51: Faixas de escore para o modelo logito - Validação

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Tx. Fraude Esp.	Tx. Fraude Obs.
menor que 977,0	1573	105	1678	10,2%	9,56%	6,26%
977,0 -| 989,2	1548	89	1637	10,0%	7,98%	5,44%
989,2 -| 990,5	1565	74	1639	10,0%	3,66%	4,51%
990,5 -| 992,4	1537	68	1605	9,8%	3,49%	4,24%
992,4 -| 993,0	1577	25	1602	9,8%	1,99%	1,56%
993,0 -| 993,5	1599	8	1607	9,8%	1,34%	0,50%
993,5 -| 994,0	1612	7	1619	9,9%	0,97%	0,43%
994,0 -| 995,8	1647	5	1652	10,1%	0,75%	0,30%
995,8 -| 996,3	1699	3	1702	10,4%	0,25%	0,18%
996,3 ou mais	1645	2	1647	10,1%	0,07%	0,12%
Total	16002	386	16388	100,0%	3,01%	2,36%

Tabela 8.52: Faixas de escore para o modelo logito limitado - Validação

Faixas de Escore	Não Fraude	Fraude	Total	%Total	Tx. Fraude Esp.	Tx. Fraude Obs.
menor que 977,0	1529	175	1704	10,4%	10,27%	11,43%
977,0 -| 989,2	1575	99	1674	10,2%	5,91%	6,70%
989,2 -| 990,5	1555	31	1586	9,7%	1,95%	2,50%
990,5 -| 992,4	1587	25	1612	9,8%	1,55%	2,03%
992,4 -| 993,0	1598	19	1617	9,9%	1,18%	1,13%
993,0 -| 993,5	1605	15	1620	9,9%	0,93%	1,03%
993,5 -| 994,0	1609	9	1618	9,9%	0,56%	0,58%
994,0 -| 995,8	1645	7	1652	10,1%	0,42%	0,45%
995,8 -| 996,3	1656	5	1661	10,1%	0,30%	0,26%
996,3 ou mais	1643	1	1644	10,0%	0,06%	0,02%
Total	16002	386	16388	100%	2,36%	2,49%

Capítulo 9

Análise Bayesiana para o Modelo Logito Limitado

O avanço das capacidades computacionais, na última década, possibilitou a implementação de métodos de aproximação numérico (técnicas de simulação) promovendo o desenvolvimento da análise estatística de problemas cada vez mais complexos. Essa facilidade de calcular integrais de funções complexas deu um novo impulso a inferência Bayesiana. (Ehlers [21]).

A análise Bayesiana possui algumas limitações, como por exemplo o tempo computacional e a capacidade de armazenamento dos valores simulados.

9.1 Introdução

A informação que tem-se sobre uma quantidade desconhecida θ é fundamental na estatística. O verdadeiro valor de θ é desconhecido e a idéia é tentar reduzir esta falta de conhecimento. Além disso, a intensidade de incerteza a respeito de θ pode assumir diferentes graus. Do ponto de vista Bayesiano, estes diferentes graus de incerteza são representados através de modelos probabilísticos para θ. Neste contexto é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ, especificando distribuições *a priori* distintas, que representa o grau de credibilidade que o pesquisador possui sobre θ. Observe que na inferência Bayesiana não existe distinção entre quantidade observáveis e os parâmetros de um modelo estatístico, sendo todos considerados variáveis

aleatórias. (Paulino [18]).

Nesta abordagem, os estimadores e testes são desenvolvidos considerando as experiências adquiridas no passado.

9.2 Teorema de Bayes

Pode-se acrescentar a informação que se tem disponível sobre o parâmetro θ, resumida na distribuição *a priori* $p(\theta)$, através da observação de uma quantidade aleatória X relacionada com θ, sendo que esta relação é definida pela distribuição $p(x \mid \theta)$. A idéia consiste em que após observar $X = x$, a quantidade de informação sobre θ aumenta e o teorema de Bayes é a regra de atualização utilizada para quantificar este aumento. Assim tem-se:

$$p(\theta \mid x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x \mid \theta)p(\theta)}{\int p(\theta, x)d\theta} \tag{9.1}$$

Observe que o denominador não depende de θ e sendo assim, funciona como uma constante normalizadora de $p(\theta \mid x)$.Para um valor fixo de x, a função $p(x \mid \theta) = l(\theta; x)$, denominada de função de verossimilhança. Desse modo, pode-se considerar:

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta) \tag{9.2}$$

Em palavras, tem-se que:

$$\text{distribuição a posteriori } \propto \text{ verossimilhança x distribuição a priori} \tag{9.3}$$

Esta forma apresentada acima é uma simplicação do teorema de Bayes e é muito útil em problemas que envolvem estimação dos parâmetros. Em outras situações, como seleção de modelos, o termo da constante normalizadora é fundamental.

9.3 Método de Monte Carlo via cadeias de Markov

Na inferência Bayesiana deseja-se gerar uma amostra de uma distribuição *a posteriori* $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$. Os métodos de MCMC são baseados na substituição da expressão analítica da densidade por uma amostra gerada a partir desta densidade.

9.3.1 Amostrador de Gibbs

Quando as distribuições condicionais completas $\pi\left(\theta_i \mid \boldsymbol{y}, \boldsymbol{\theta}_{(i)}\right)$, $\boldsymbol{\theta}_{(i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$ são dadas por distribuições conhecidas utiliza-se o amostrador de Gibbs para se obter inferências da posteriori conjunta $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$. Para isso, simulam-se amostras de $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$ a partir destas distribuições condicionais.

Dado um vetor de valores iniciais dos parâmetros $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}\right)$, atualizar $\boldsymbol{\theta}^{(0)}$ por $\boldsymbol{\theta}^{(1)}$ a partir do algoritmo:

1. Gerar $\theta_1^{(1)}$ de $\pi\left(\theta_1 \mid \boldsymbol{y}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}\right)$;
2. Gerar $\theta_2^{(1)}$ de $\pi\left(\theta_2 \mid \boldsymbol{y}, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}\right)$;
3. Gerar $\theta_3^{(1)}$ de $\pi\left(\theta_3 \mid \boldsymbol{y}, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_k^{(0)}\right)$;
4. $\vdots$
5. Gerar $\theta_k^{(1)}$ de $\pi\left(\theta_k \mid \boldsymbol{y}, \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)}\right)$.
6. Repetir os passos acima um número grande de vezes até a obtenção de convergência.

9.3.2 Metropolis-Hasting

Este algoritmo é uma generalização, realizada por Hasting em 1970, do algoritmo de Metropolis e é utilizado quando as distribuições condicionais não possuem formas fechadas. Neste algoritmo, um valor é gerado de uma dada distribuição auxiliar e aceito com uma dada probabilidade. Este mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio, neste caso a distribuição a posteriori (EHLERS [21]).

Suponha que a cadeia esteja no estado x e um valor x' é gerado de uma distribuição proposta $q(\cdot|x)$. Observe que a distribuição proposta pode depender do estado atual. A probabilidade do novo valor x' ser aceito é dada por:

$$\alpha(x, x') = \min\left(1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right) \tag{9.4}$$

Neste caso, π é a distribuição de interesse.

Uma característica importante é que só precisa-se conhecer π parcialmente, isto é a menos de uma constante, já que neste caso a probabilidade não se altera. Isto é fundamental em aplicações Bayesianas em que não se conhece completamente a densidade a posteriori. Observe também que a cadeia pode permanecer no mesmo estado por muitas iterações e na prática costuma-se monitorar este fato calculando a porcentagem média de iterações para as quais novos valores são aceitos (EHLERS [21]).

Em muitos casos a dificuldade deste método é a escolha da função de transição, sendo que isto influência diretamente a eficácia do método. Uma forma de escolher a função de transição é considerar o passeio aleatório:

$$q(x|x') = q_1\left(x' - x\right) \tag{9.5}$$

Sendo $q\left(\cdot\right)$ uma densidade multivariada. Neste caso, $x' = x + \varepsilon$, sendo ε a variável de incremento com distribuição $q\left(\cdot\right)$. Se $q_1(\varepsilon) = q_1(-\varepsilon)$ então:

$$\alpha(x, x') = \min\left(1, \frac{\pi(x')}{\pi(x^{(j)})}\right) \tag{9.6}$$

Caso seja possível fatorar a distribuição a posteriori em:

$$\pi(x) = \phi(x)h(x) \tag{9.7}$$

onde $h(x)$ é uma densidade que pode ser amostrada e ϕ uniformemente limitada, então:

$$q(x|x') = h(x') \tag{9.8}$$

Neste caso, a probabilidade de aceitação é dada por:

$$\alpha(x, x') = \min\left(1, \frac{\phi(x')}{\phi(x^{(j)})}\right) \tag{9.9}$$

Outras formas de escolher a função de transição pode ser encontrada em Chib e Greemberg [23].

O algoritmo de Metropolis-Hasting é definido por:

1. Inicialize com um valor arbitrário $x^{(0)}$.

2. Gere um x' de $q(x', x)$ e u de uma distribuição Uniforme(0,1).

3. Seja $\alpha(x, x') = \min\left(1, \frac{\pi(x')q(x'|x^{(j)})}{\pi(x^{(j)})q(x^{(j)}|x')}\right)$.

4. Se $u \leq \alpha$, então $x^{(j+1)} = x'$; senão $x^{(j+1)} = x^{(j)}$.

5. Repetir os passos 2, 3 e 4 até se obter convergência.

9.3.3 Diagnósticos de Convergência

Para verificar convergência do algoritmo pode-se utilizar:

(a) Gráficos de séries temporais;

A linha deste gráfico deve estar sempre em torno de uma faixa, não apresentando muitas oscilações.

(b) Gráficos de autocorrelação;

O gráfico de autocorrelação dos parâmetros somente deve ter a primeira correlação alta sendo as demais próximas de zero, indicando assim que as amostras geradas são não-correlacionadas.

(c) Índices de convergência, como o índice de Gelman & Rubin (1992).

Gelman e Rubin utilizam procedimentos de análise de variância (ANOVA) para verificar a convergência de cadeias paralelas (com valores iniciais diferentes). A idéia é comparar as variabilidades "entre"e "dentro"das cadeias geradas. O método funciona da seguinte maneira:

Simular $m \geq 2$ cadeias, cada uma de comprimento $2n$, descartando-se as n primeiras amostras;

Calcular a variabilidade entre cadeias

$$\frac{U}{n} = \sum_{i=1}^{m} \left(\bar{\theta}_{i.} - \bar{\theta}_{..}\right)^2 \diagup (m-1)$$

em que $\bar{\theta}_{i.}$ é a média das n amostras geradas na i-ésima cadeia; $\bar{\theta}_{..}$ é a média geral.

Define-se, também, a variabilidade dentro de cada cadeia,

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2$$

em que $s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\theta_{ij} - \bar{\theta}_{i.}\right)^2$, $i = 1, \ldots, m$.

Assim, a variância da distribuição estudada (*posteriori*) pode ser estimada por:

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}U$$

Gelman e Rubin mostraram que a distribuição de $\boldsymbol{\theta}$ dado y segue distribuição t de Student com centro em $\hat{\mu} = \bar{\theta}_{..}$, desvio-padrão $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + \frac{U}{mn}}$ e $\nu.= \frac{2\hat{V}}{var\left(\hat{V}\right)}$ graus de liberdade, em que:

$$\begin{aligned} var\left(\hat{V}\right) &= \left(\frac{n-1}{n}\right)^2 \frac{1}{m} var\left(s_i^2\right) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1}B^2 + \\ &\quad +2\frac{(m-1)(n-1)}{mn^2}\frac{n}{m}\left[cov\left(s_i^2, \bar{u}_{i.}^2\right) - 2\bar{u}_{..}cov\left(s_i^2, \bar{u}_{i.}\right)\right] \end{aligned}$$

Estima-se, então, o fator de redução de escala como:

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}\frac{\nu}{\nu-2}}$$

Quando n é grande, o fator de redução pode ser simplificado para $\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}}$. Nota-se, também, que $\sqrt{\hat{R}}$ decresce para 1 quando $n \to \infty$. Em geral, se $\sqrt{\hat{R}} < 1,2$, tem-se convergência

das cadeias.

9.4 Aplicações

9.4.1 Dados da Literatura

Com a finalidade de ilustrar a abordagem Bayesiana para problemas de regressão, considere o seguinte exemplo.

Uma agência bancária fez um estudo estatístico para verificar possíveis fatores que levam depositantes retirarem seus depósitos durante o período de um ano. Para este estudo foram escolhidos aleatoriamente 39 clientes. Sejam:

- $D1$: saldo do cliente no dia 01 de janeiro de 2000;
- $D2$: saldo do cliente no dia 31 de dezembro de 2000.
- $S = D1/D2$.

Se $S > 1$ significa que o cliente aumenta seus depósitos na agencia bancária e $S \leq 1$ que o cliente retira valores durante o ano e não repõe este.

A variável resposta é dada por:

$$y = \begin{cases} 1 & \text{se } S > 1 \\ 0 & \text{se } S \leq 1 \end{cases} \tag{9.10}$$

O objetivo do banco é que os clientes aumentem seus depósitos, assim, o sucesso é dado se $S > 1$.

Quatro covariáveis são consideradas no estudo:

- $X1$, que recebe o valor 1 se o cliente é funcionário público e 0 caso contrário;
- $X2$, que recebe 1 se o cliente tem conta bancária com mais de 5 anos e 0 caso contrário;
- $X3$, que recebe 1 se o cliente é microempresário e 0 caso contrário;
- $X4$, que recebe 1 se o cliente tem residência própria e 0 caso contrário.

Regressão Logística

Assuma um modelo de regressão logística para a resposta Y_i para a aplicação em estudo, isto é,

$$P(Y_i = y_i|\mathbf{x}_i) = f(y_i|x_i) = (P_i)^{y_i}(1-P_i)^{1-y_i}, \qquad com \qquad y_i = 0, 1 \text{ e } i = 1, ..., 39 \quad (9.11)$$

Para incluir todas as covariáveis e todas as interações, foi definida a variável v_{ji}:

$v_{00} = 1$

$v_{1i} = x_{1i}$

$v_{2i} = x_{2i}$

$v_{3i} = x_{3i}$

$v_{4i} = x_{4i}$

$v_{5i} = x_{1i}x_{2i}$

$v_{6i} = x_{1i}x_{3i}$

$v_{7i} = x_{1i}x_{4i}$

$v_{8i} = x_{2i}x_{3i}$

$v_{9i} = x_{2i}x_{4i}$

$v_{10i} = x_{3i}x_{4i}$

$v_{11i} = x_{1i}x_{2i}x_{3i}$

$v_{12i} = x_{1i}x_{2i}x_{4i}$

$v_{13i} = x_{1i}x_{3i}x_{4i}$

$v_{14i} = x_{1i}x_{2i}x_{3i}x_{4i}$

Dessa forma, o modelo Logito é dado por:

$$p_i = \frac{\exp(\boldsymbol{\beta}'\mathbf{v}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{v_i})}, \ \boldsymbol{\beta}'\mathbf{v}_i = \sum_{j=0}^{14}\beta_j' v_{ji} \text{ e } v_{00} = 1 \quad (9.12)$$

A função de verossimilhança para $\beta_1, \beta_2, ..., \beta_{14}$ é dada por:

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n}(P_i)^{y_i}(1-P_i)^{1-y_i} \quad (9.13)$$

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{v}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{v}_i)} \right)^{y_i} \left(1 - \frac{\exp(\boldsymbol{\beta}'\mathbf{v}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{v}_i)} \right)^{1-y_i} \tag{9.14}$$

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{v}_i y_i)}{(1+\exp(\boldsymbol{\beta}'\mathbf{v}_i))^{y_i}} \right] \left[\frac{1}{(1+\exp(\boldsymbol{\beta}'\mathbf{v}_i))^{1-y_i}} \right] \tag{9.15}$$

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{v}_i y_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{v}_i)} \right] \tag{9.16}$$

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \left[\frac{\exp\left[\sum_{i=1}^{n} (\boldsymbol{\beta}'\mathbf{v}_i)\, y_i \right]}{\prod_{i=1}^{n} (1+\exp(\boldsymbol{\beta}'\mathbf{v}_i))} \right] \tag{9.17}$$

Observe que:

- $\boldsymbol{\beta}'\mathbf{v}_i = \sum_{j=0}^{14} \beta'_j v_{ji}$ e $v_{00} = 1$;
- $\sum_{i=1}^{n} (\boldsymbol{\beta}'\mathbf{v}_i)\, y_i = \sum_{j=0}^{14} (\boldsymbol{\beta}'\mathbf{v}_i) \sum_{i=1}^{n} y_i = \sum_{j=0}^{14} \sum_{i=1}^{n} (\boldsymbol{\beta}'\mathbf{v}_i y_i)\,.$

Assim, a função de máxima verossimilhança é dada por:

$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \left[\frac{\exp\left[\sum_{j=0}^{14} \sum_{i=1}^{n} (\boldsymbol{\beta}'\mathbf{v}_i y_i) \right]}{\prod_{i=1}^{n} \left[1+\exp\left(\sum_{j=0}^{14} \beta'_j v_{ji} \right) \right]} \right] \tag{9.18}$$

Assumindo independência a priori entre parâmetros, considere que as distribuições a priori para $\beta_1, \beta_2, ..., \beta_{14}$ são dadas pelas seguintes distribuições:

$\beta_j \sim N(a_j, b_j)$, com a_j e b_j são constantes conhecidas para $j = 1, 2, ..., 14$.

A distribuição a posteriori é dada por:

$$\pi(\beta_j | \theta_{(\beta_j)}, x, y) \propto \exp\left\{ \frac{-1\left(\beta_j - a_j\right)^2}{2b_j^2} \right\} \frac{\exp\left\{ \sum_{i=1}^{n} \beta_j y_i v_{ji} \right\}}{\prod_{i=1}^{n} \left[1+\exp\left(\sum_{j=0}^{14} \beta'_j v_{ji} \right) \right]} \tag{9.19}$$

onde $\theta_{(\beta_j)}$ é o vetor de todos os parâmetros exceto β_k.

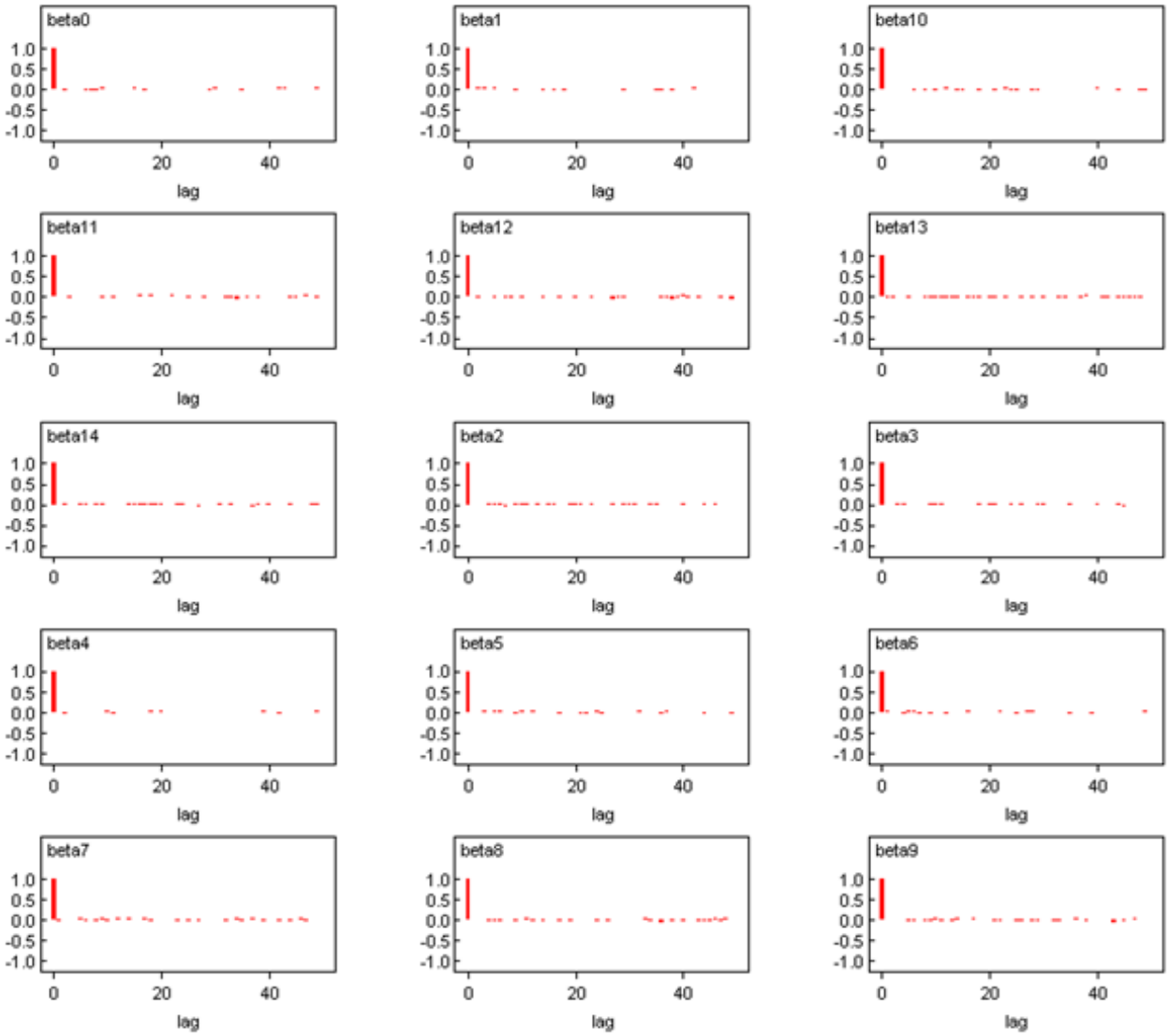
Como a posteriori, dada pela Equação 9.19, não possui uma forma fechada, deve-se utilizar o algoritmo Metropolis-Hastings para gerar amostras da distribuição a posteriori a partir da distribuição conjunta para β. A análise foi realizada no *software* Winbugs.

A Tabela 9.1 apresenta as estimativas dos parâmetros e o intervalo de credibilidade (2.50% e 97.50%) para cada um dos parâmetros estimados.

Tabela 9.1: Estimativas dos parâmetros - modelo logito

Parâmetros	Média	DP	MC error	2.50%	Mediana	97.50%
beta0	-0,7699	-0,4311	0,00559	-1,6350	-0,7581	0,0227
beta1	2,4590	0,7607	0,00865	0,9729	2,4560	3,9430
beta2	1,6520	0,8677	0,01219	-0,0488	1,6440	3,3770
beta3	-2,3590	0,8100	0,01115	-3,9880	-2,3590	-0,7263
beta4	0,4830	0,8275	0,01128	-1,1210	0,4820	2,1220
beta5	-0,0569	0,9049	0,02236	-1,7600	-0,0601	1,7350
beta6	-0,5996	0,8458	0,01637	-2,3060	-0,5962	1,0140
beta7	-0,2297	0,8742	0,01813	-1,9410	-0,2509	1,5420
beta8	0,5441	0,8492	0,01664	-1,1600	0,5352	2,2160
beta9	0,1292	0,9226	0,01761	-1,6780	0,1444	1,8670
beta10	0,0857	0,9183	0,01815	-1,6980	0,0576	1,9100
beta11	-0,0665	0,9072	0,02029	-1,7730	-0,0715	1,7030
beta12	-0,4330	0,9340	0,01944	-2,2450	-0,4485	1,4040
beta13	0,1033	0,9298	0,01633	-1,7340	0,0798	1,9410
beta14	-0,4311	0,9284	0,02055	-2,2620	-0,4420	1,3990

Através da Figura 9.1 pode-se observar os gráficos de auto-correlação. Verifique que todos os parâmetros são não correlacionados, o que indica a convergência das cadeias. Também pode-se verificar a convergência através da Figura 9.2 que apresenta os gráficos temporais.



9.4.2 Regressão Logística Limitada

Para o ajuste da regressão logística limitada, considerou-se apenas a covariável X_1. No caso da regressão logística com função Logito, o Winbugs já possui uma função pronta. Para este caso, teve-se que implementar uma função. Será considerada uma priori Normal para β_0 e β_1 e uma priori Beta para o parâmetro ϖ.

A verossimilhança é dada por:

$$L(\beta_0, \beta_1, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} (p_i)^{y_i} (1 - p_i)^{1-y_i} \tag{9.20}$$

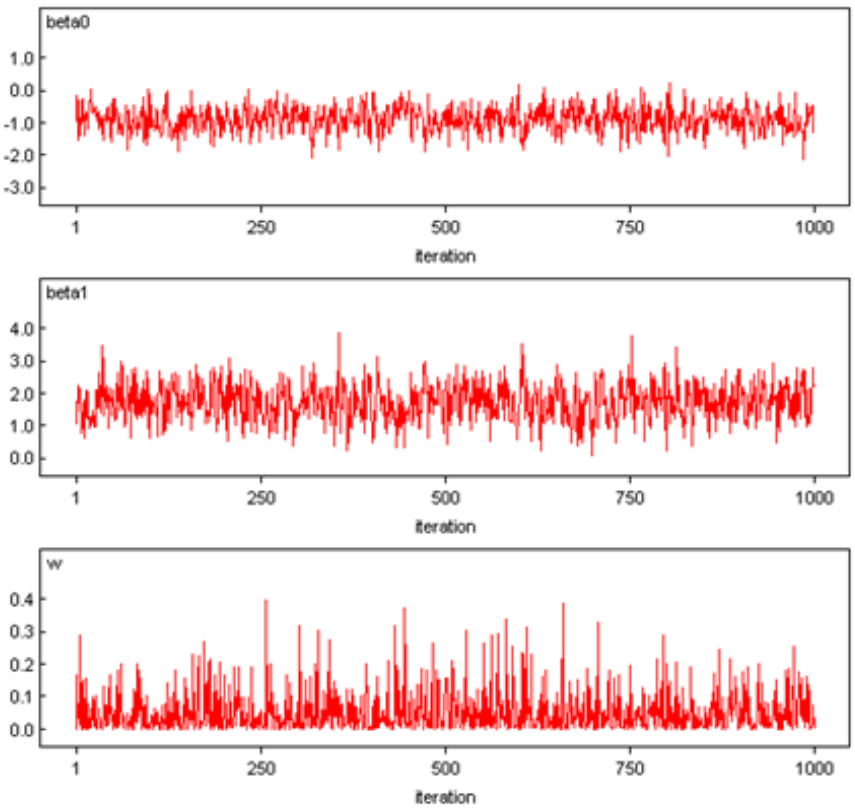
$$L(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \prod_{i=1}^{n} \left(\varpi \frac{\exp(\beta_0 + \beta_1 x_{1i})}{1 + \exp(\beta_0 + \beta_1 x_{1i})} \right)^{y_i} \left(1 - \varpi \frac{\exp(\beta_0 + \beta_1 x_{1i})}{1 + \exp(\beta_0 + \beta_1 x_{1i})} \right)^{1-y_i} \tag{9.21}$$

A Tabela 9.2 apresenta as estimativas bayesianas para este ajuste.

Tabela 9.2: Estimativas dos parâmetros - modelo logito limitado

Parâmetros	Média	DP	MC error	2.50%	Mediana	97.50%
ϖ	0,0502	0,0638	0,00163	4,96E-02	0,0261	0,2297
β_0	-0,8930	0,3877	0,02224	-1,6310	-0,8922	-0,1003
β_1	1,7310	0,5774	0,01219	0,5902	1,7270	2,8050

Através da análise gráfica apresentada nas Figuras 9.3 e 9.4, pode-se observar que ocorreu convergência.



9.4.3 Dados Simulados

Com os dados simulados do Capítulo 8 ajustou-se o modelo logito e logito limitado através da abordagem Bayesiana.

Regressão Logística

A função de verossimilhança para o modelo logito é dada por:

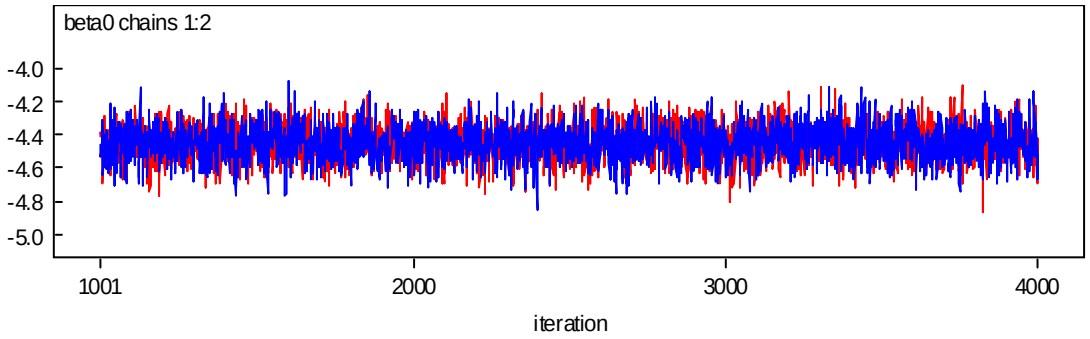
$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] \tag{9.22}$$

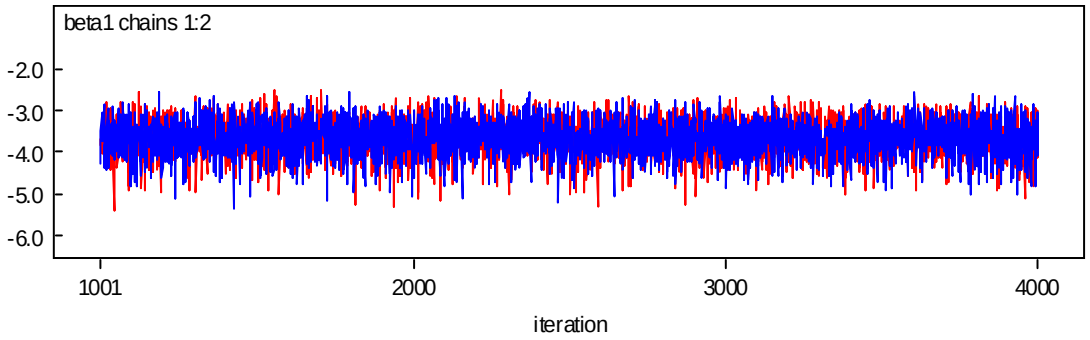
Para os parâmetros $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ considerou-se uma priori não informativa imprópria dada pela constante c.Dessa forma, espera-se que os parâmetros ajustados estejam próximos dos parâmetros estimados via abordagem clássica. Neste caso, a distribuição à posteriori é proporcional a função de verossimilhança:

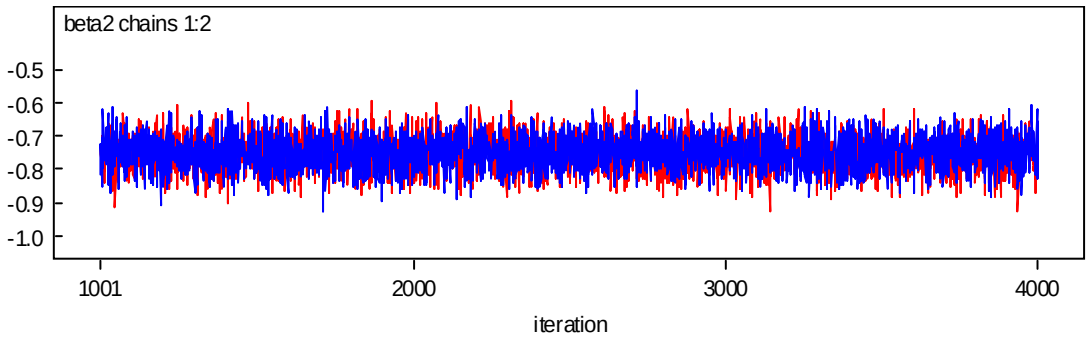
$$\pi(\boldsymbol{\beta}|Y) \propto \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] \tag{9.23}$$

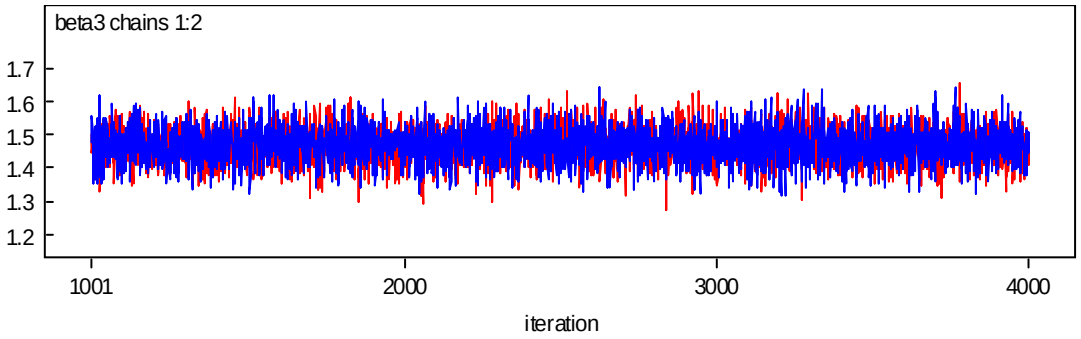
Através do *software* Winbugs obteve-se as distribuição a posteriori. Considerou-se um *burn-in* de tamanho 1.000 com salto de tamanho 15. Após o *burn-in,* gerou-se, devido ao tempo computacional, duas cadeias de tamanho 3.000 considerando também um salto de tamanho 15.

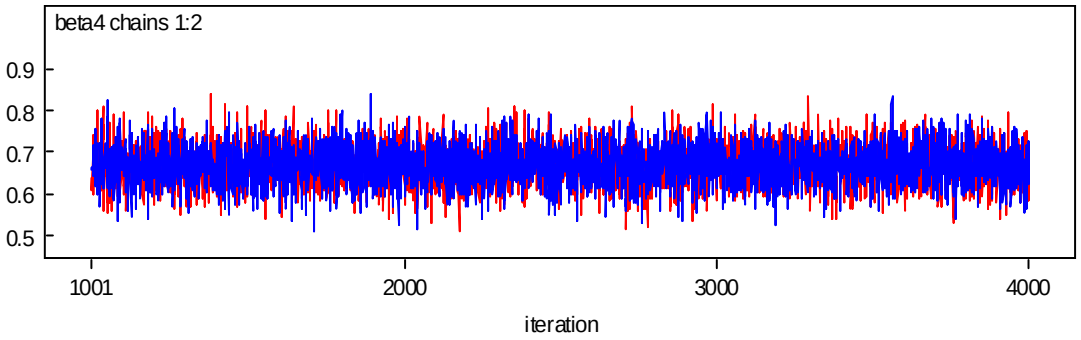
Para verificar se as cadeias geradas convergiram para a distribuição de interesse é necessário olhar os gráficos temporais e de autocorrelação. Observe o comportamento aleatório ao longo das iterações através dos gráficos temporais, indicando a convergência das cadeias.

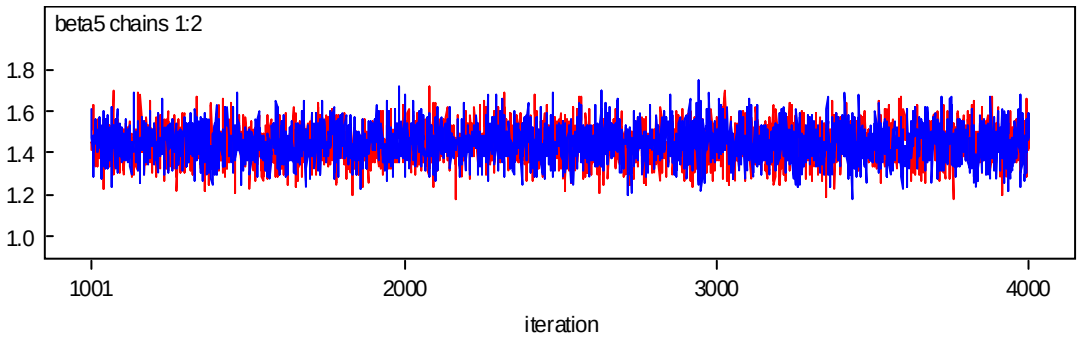

beta0 chains 1:2
-4.0
-4.2
-4.4
-4.6
-4.8
-5.0
1001
2000
3000
4000
iteration

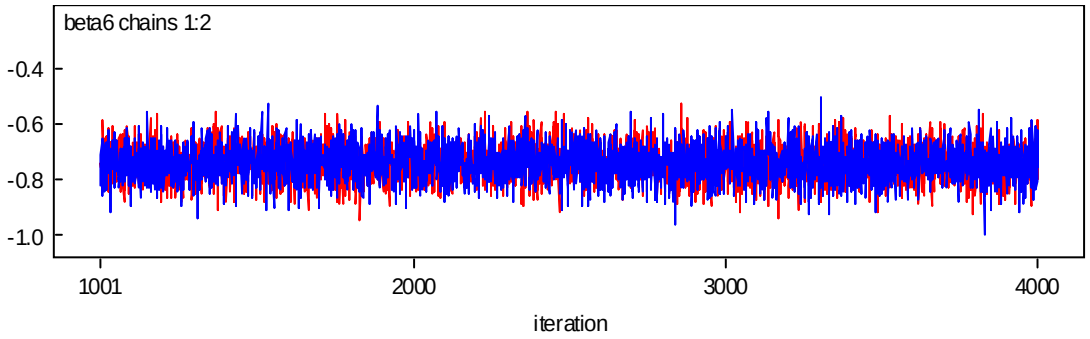

beta1 chains 1:2
-2.0
-3.0
-4.0
-5.0
-6.0
1001
2000
3000
4000
iteration


beta2 chains 1:2
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
1001
2000
3000
4000
iteration

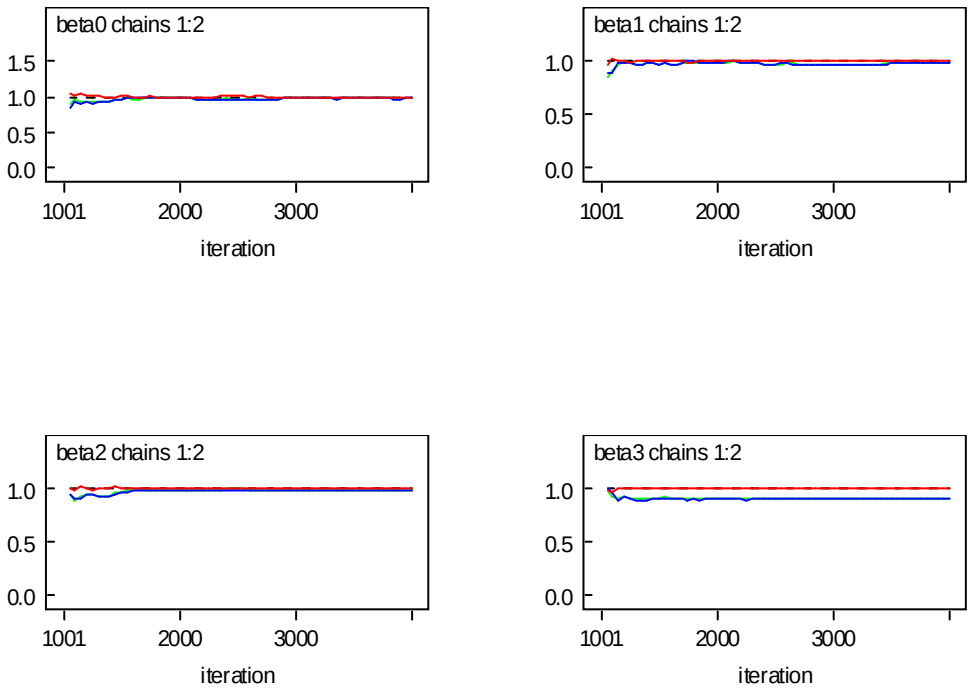

beta3 chains 1:2
1.7
1.6
1.5
1.4
1.3
1.2
1001
2000
3000
4000
iteration

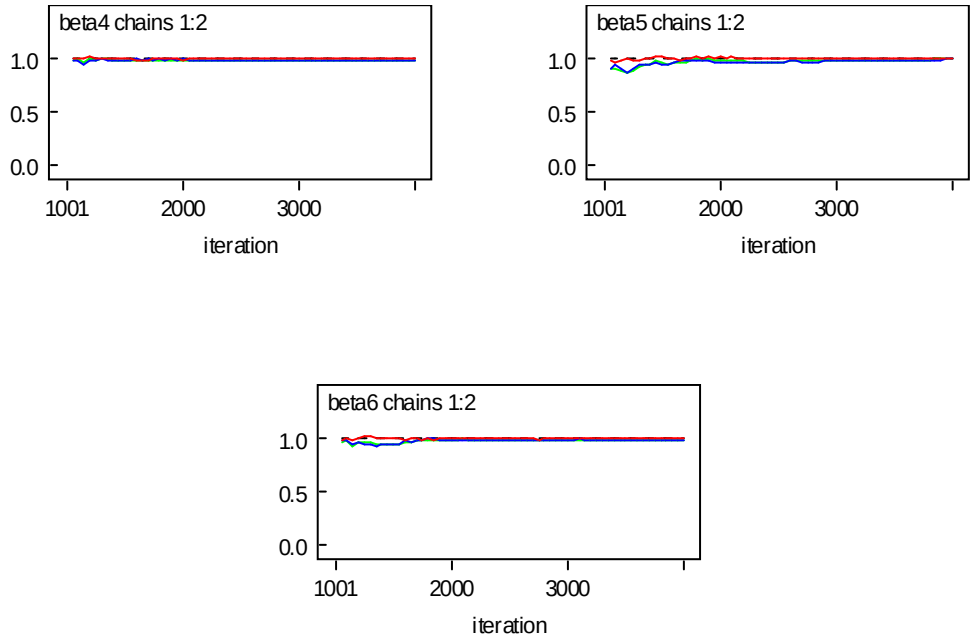

beta4 chains 1:2
0.9
0.8
0.7
0.6
0.5
1001
2000
3000
4000
iteration


beta5 chains 1:2
1.8
1.6
1.4
1.2
1.0
1001
2000
3000
4000
iteration

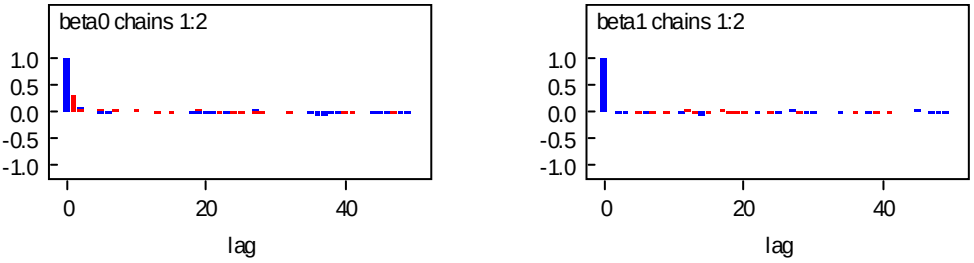


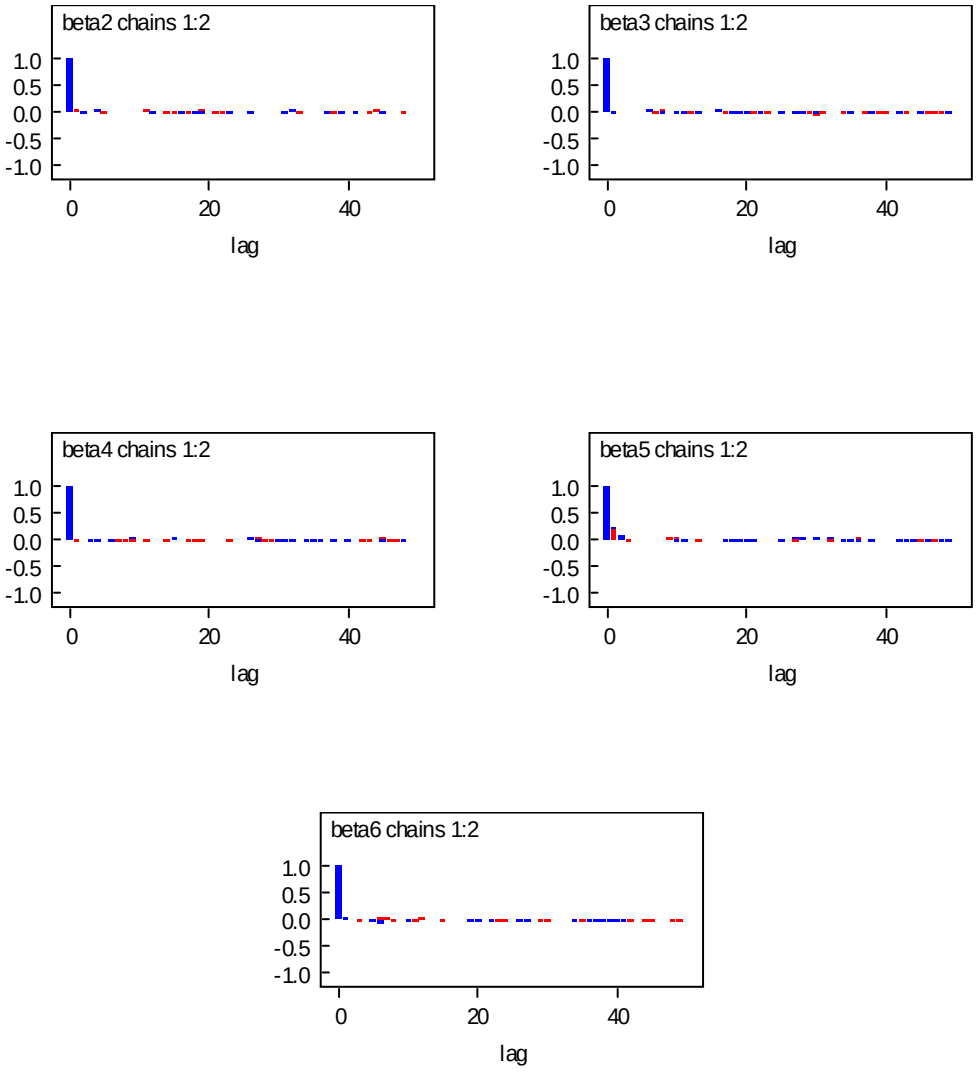
A Figura 9.6 apresenta os gráficos de Gelman Rubin. Observe que para todos os parâmetros o valor está próximo de 1, o que indíca convergência das cadeias.



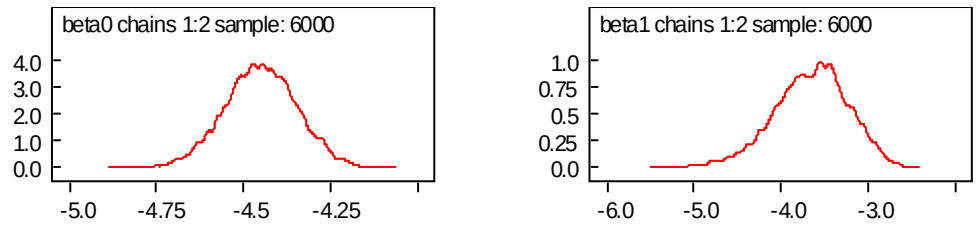


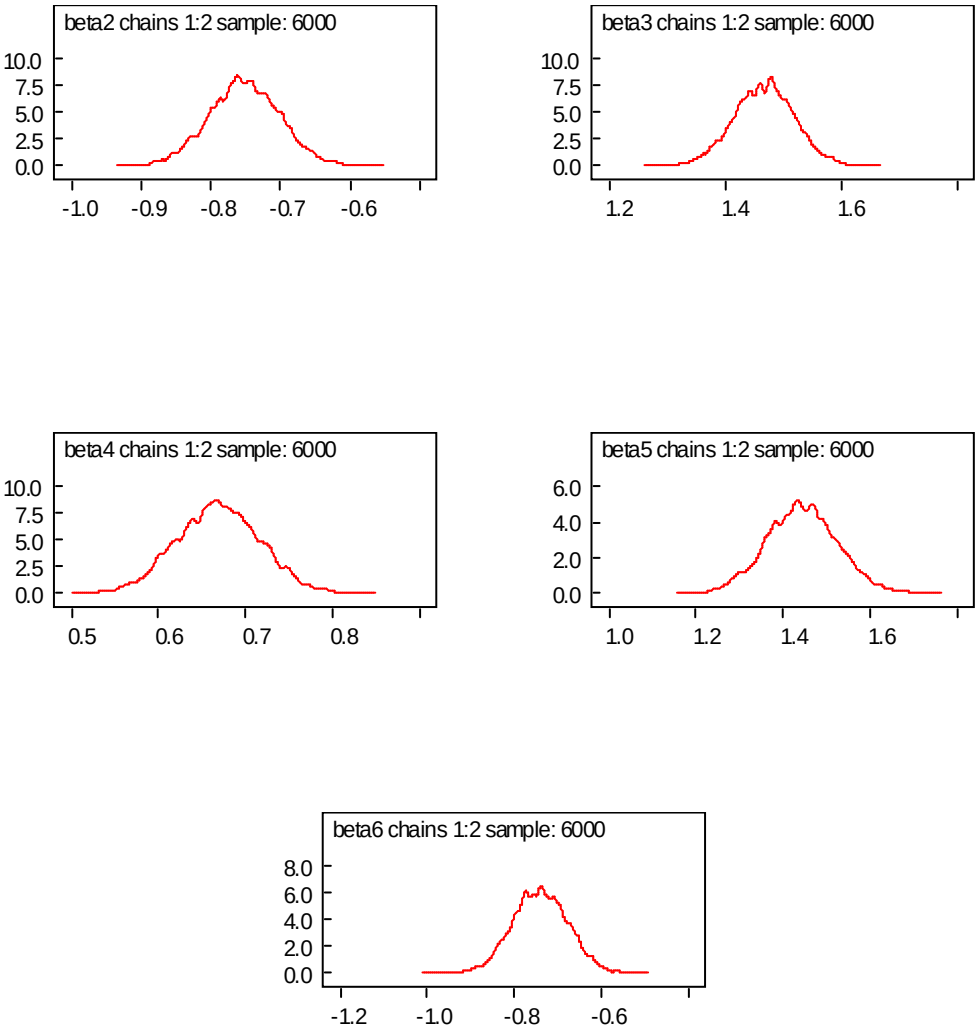
Espera-se que os valores gerados sejam não correlacionados. Para verificar a autocorrelação dos valores observe a Figura 9.7 que apresenta os gráficos de autocorrelação. Apesar dos parâmetros β_0 e β_5 apresentarem uma autocorrelação de lag 1, este fato não influênciou nos resultados.



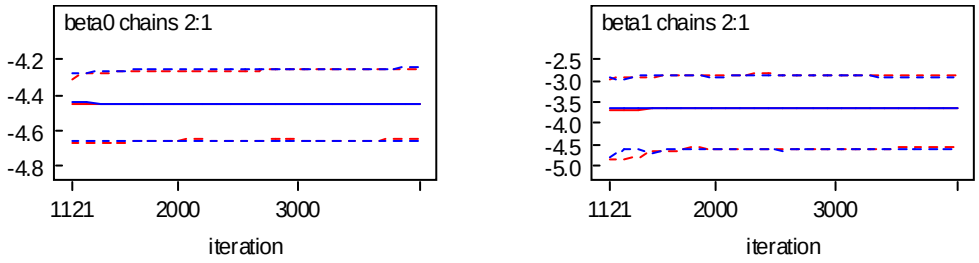


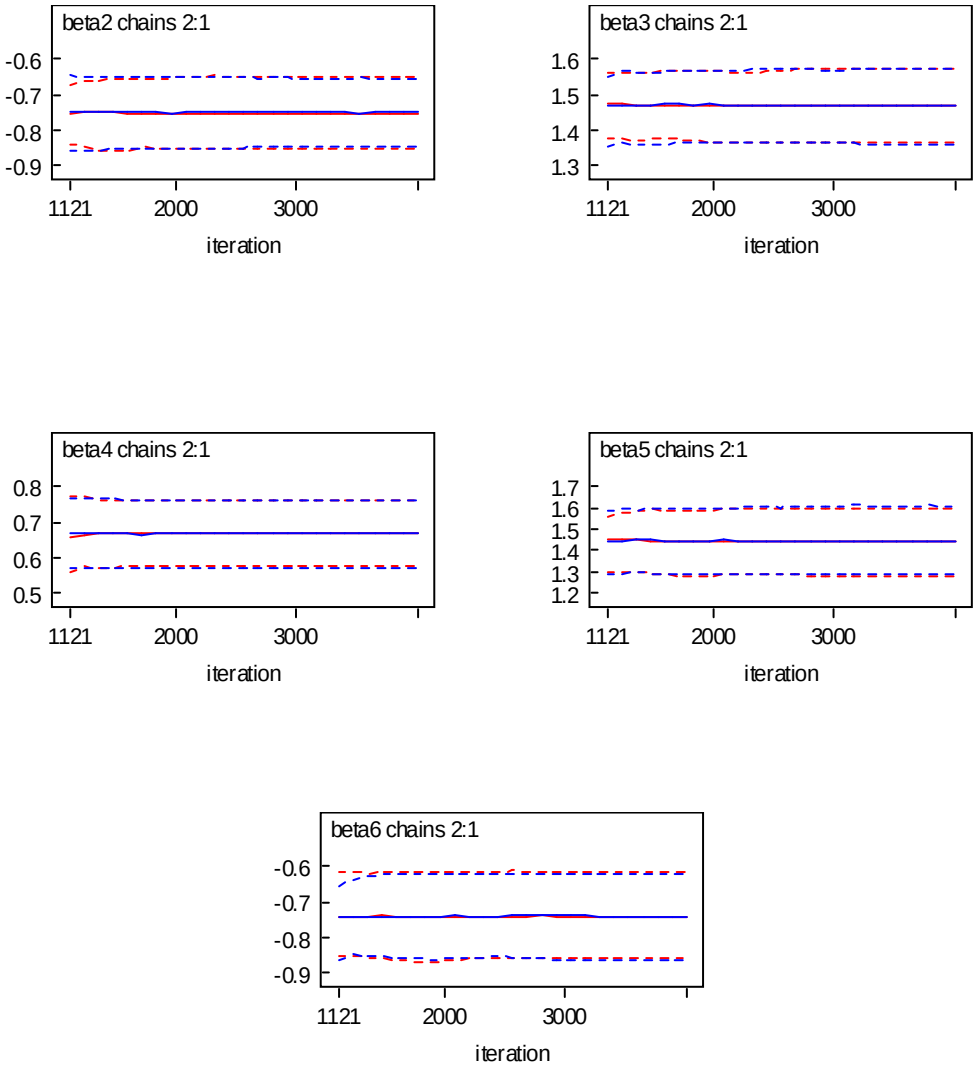
A Figura 9.8 mostra as densidades *a posteriori* dos parâmetros de interesse.





A Figura 9.9 apresenta os intervalos de credibilidade para cada um dos parâmetros. Observe que nenhum intervalo engloba o valor zero. Dessa forma, todos os parâmetros são significativos.





A Tabela 9.3 apresenta as estimativas Bayesianas para os parâmetros do modelo logito. Observe que, como esperado, estão bem próximas dos valores estimados pela abordagem clássica.

Regressão Logística Limitada

A função de verossimilhança para o modelo logito limitado é dada por:

$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] I_{(0,1)}(\omega) \tag{9.24}$$

Tabela 9.3: Estimativas Bayesianas para o modelo logito

Parâmetro	Média	Desvio Padrão	2,5%	Mediana	97,5%
β_0	-4,450	0,103	-4,653	-4,450	-4,249
β_1	-3,659	0,432	-4,601	-3,632	-2,890
β_2	-0,751	0,049	-0,85	-0,752	-0,653
β_3	1,468	0,052	1,364	1,468	1,572
β_4	0,669	0,047	0,575	0,669	0,761
β_5	1,444	0,081	1,284	1,443	1,607
β_6	-0,741	0,063	-0,862	-0,741	-0,618

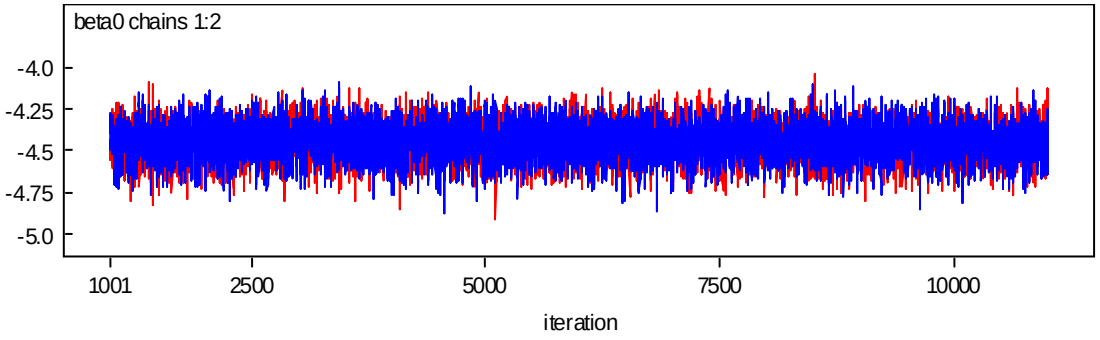
Para o parâmetro ω considerou-se uma priori Beta(0,1). Para os parâmetros $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ considerou-se uma priori não informativa imprópria dada pela constante c.

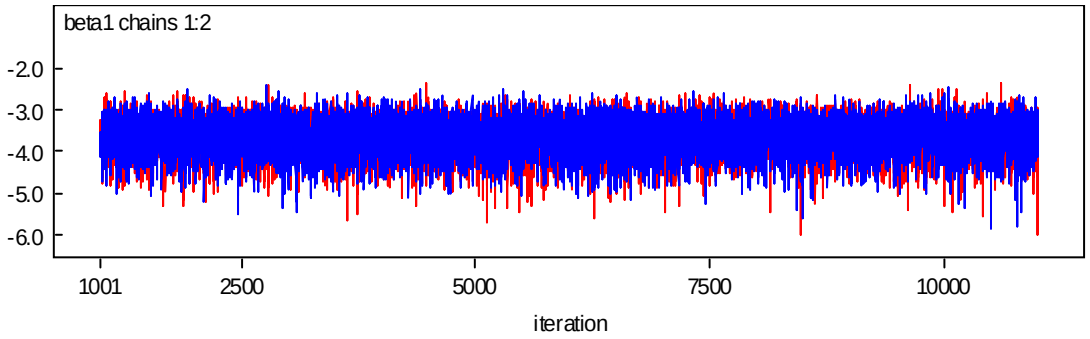
Assim a distribuição à posteriori é proporcional a:

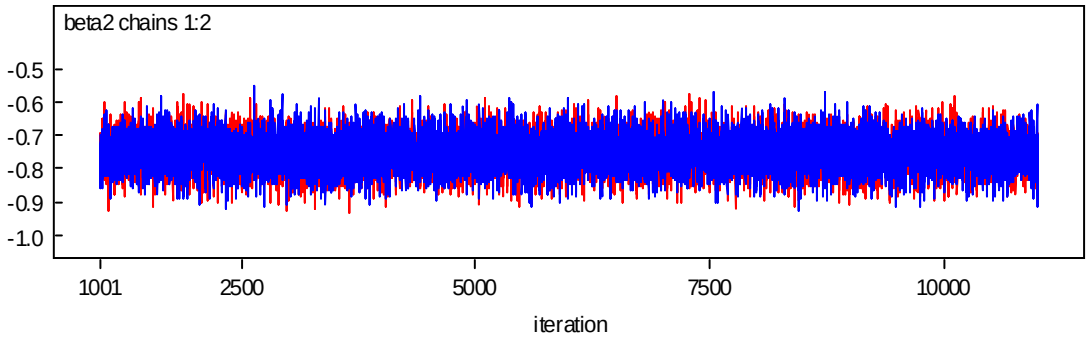
$$\propto \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] \tag{9.25}$$

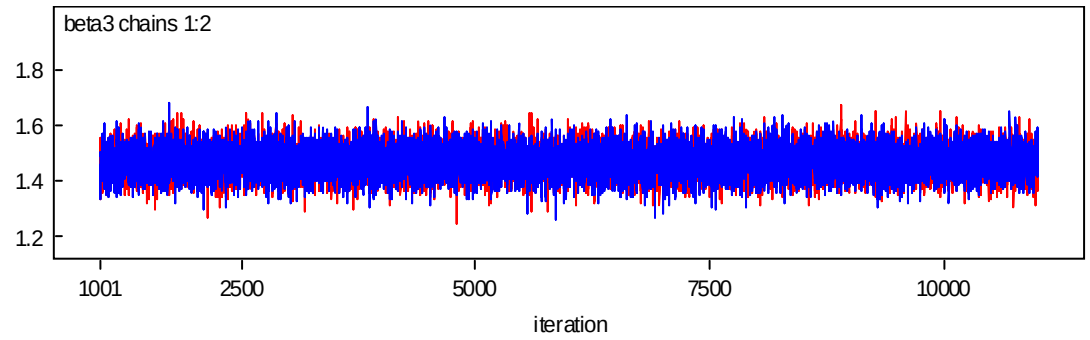
Para a obtenção da distribuição *a posteriori* utilizou-se o *software* Winbugs. Considerou-se um *burn-in* de tamanho 1.000 com salto de tamanho 10. Após o *burn-in,* gerou-se duas cadeias de tamanho 10.000 considerando também um salto de tamanho 10.

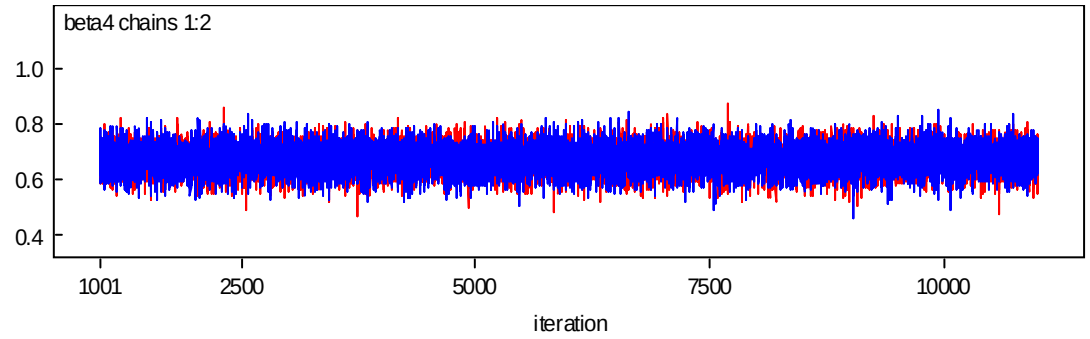
Para verificar se as cadeias geradas convergiram para a distribuição de interesse é necessário realizar as análise gráficas. A Figura 9.10 apresenta o histórico das cadeias para cada um dos parâmetros. Pode-se observar um comportamento aleatório ao longo das iterações, o que indica convergência dos parâmetros.

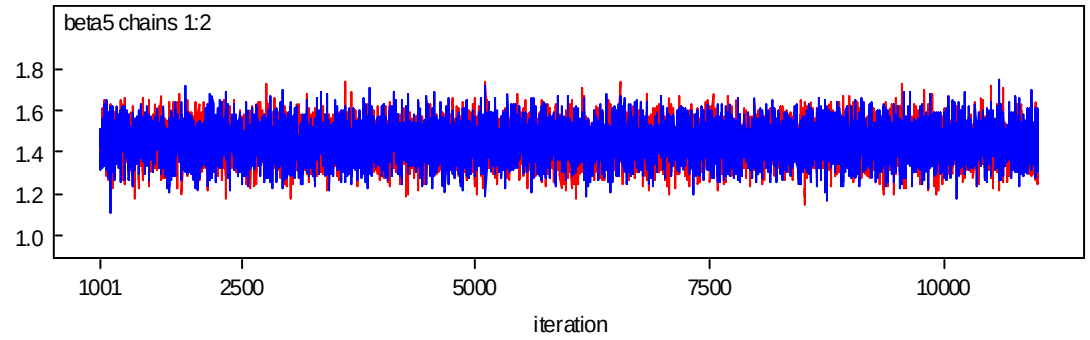


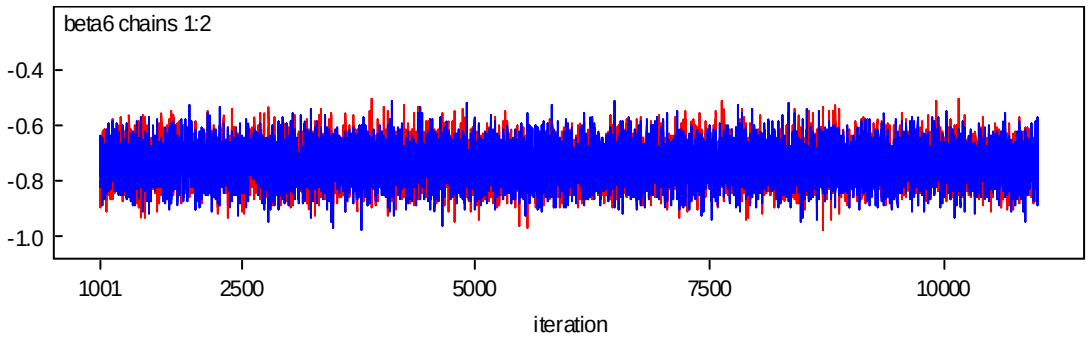

beta1 chains 1:2
-2.0
-3.0
-4.0
-5.0
-6.0
1001
2500
5000
7500
10000
iteration

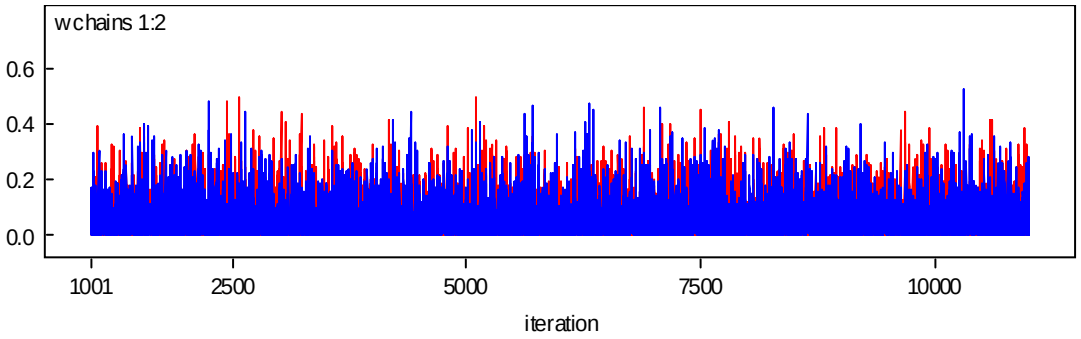

beta2 chains 1:2
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
1001
2500
5000
7500
10000
iteration


beta3 chains 1:2
1.8
1.6
1.4
1.2
1001
2500
5000
7500
10000
iteration

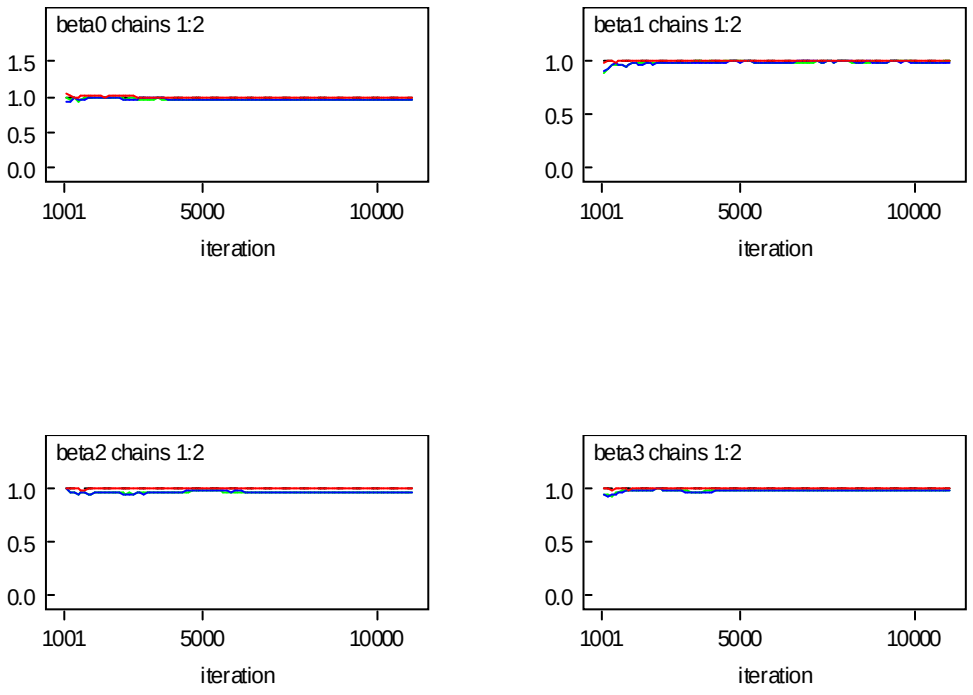

beta4 chains 1:2
1.0
0.8
0.6
0.4
1001
2500
5000
7500
10000
iteration

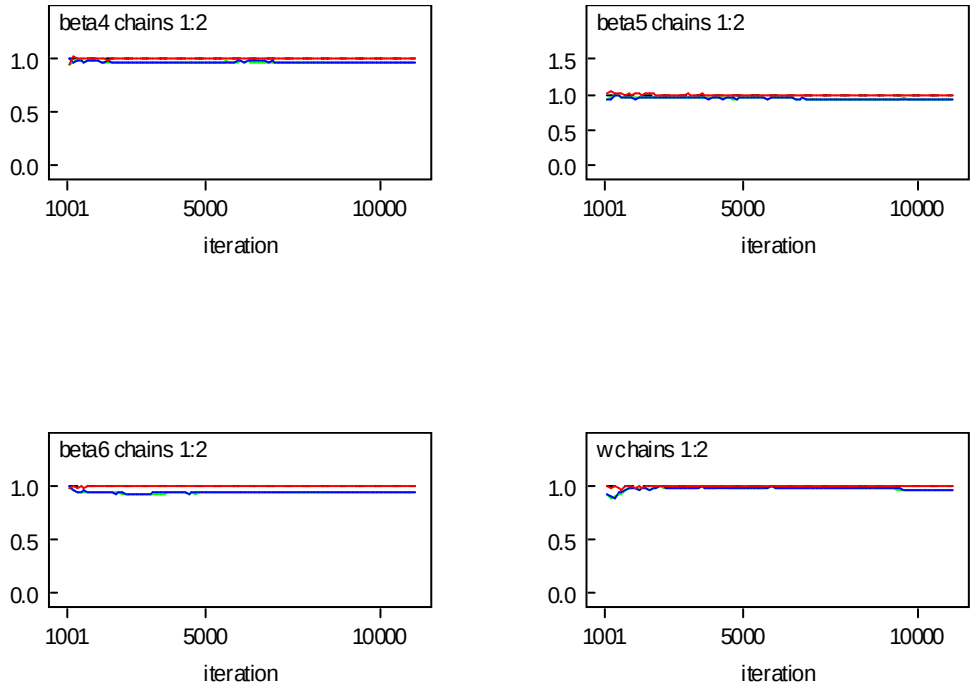

beta5 chains 1:2
1.8
1.6
1.4
1.2
1.0
1001
2500
5000
7500
10000
iteration


beta6 chains 1:2
-0.4
-0.6
-0.8
-1.0
1001
2500
5000
7500
10000
iteration

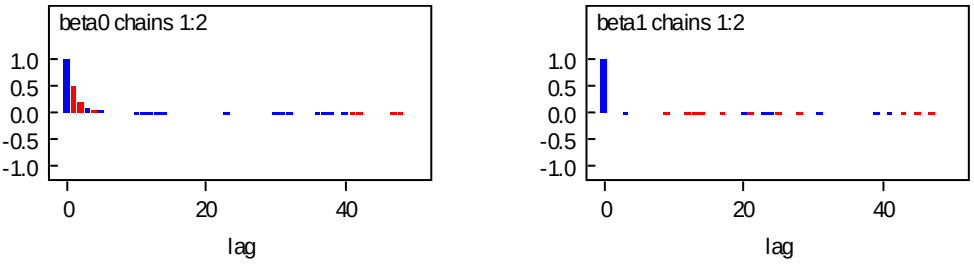


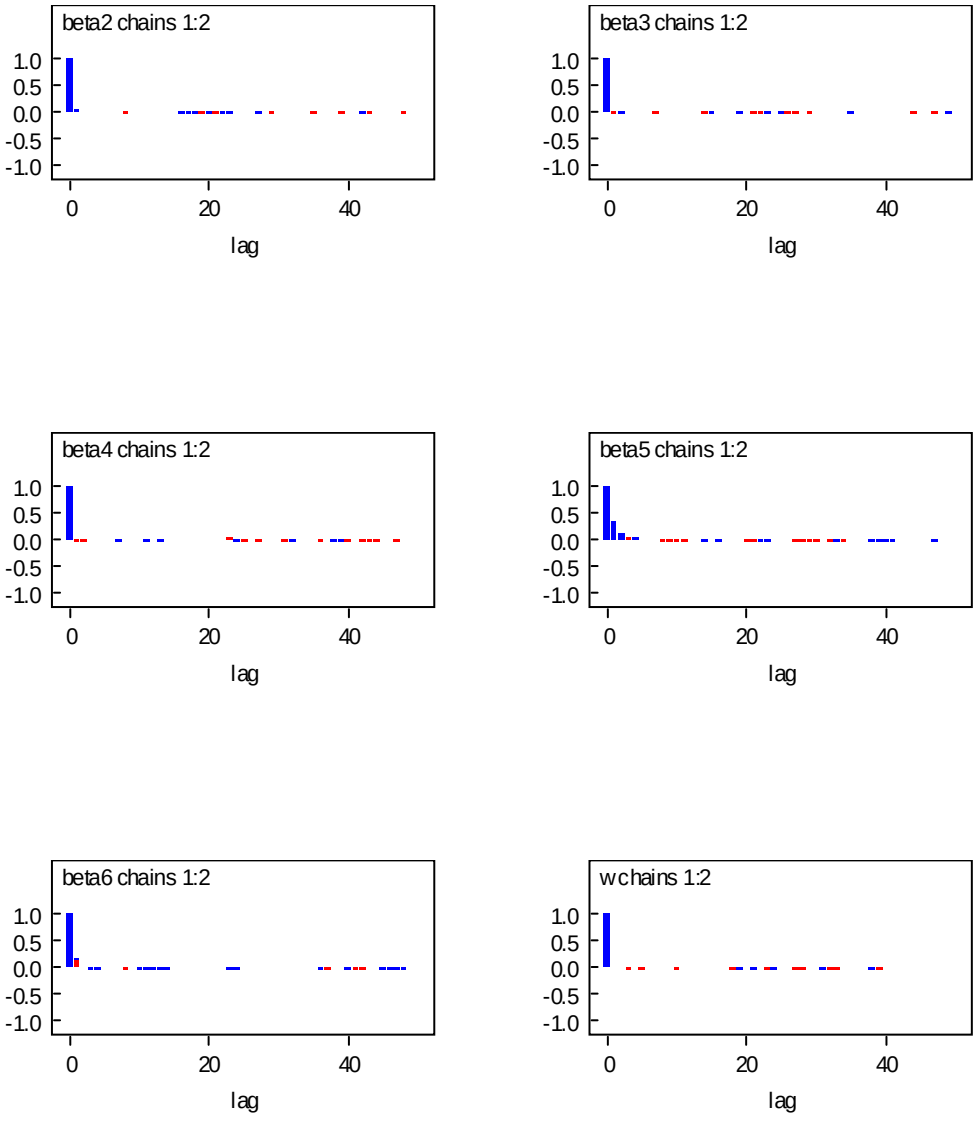
A Figura 9.11 apresenta os gráficos de Gelman Rubin. Observe que para todos os parâmetros o valor está próximo de 1, o que indíca convergência das cadeias.



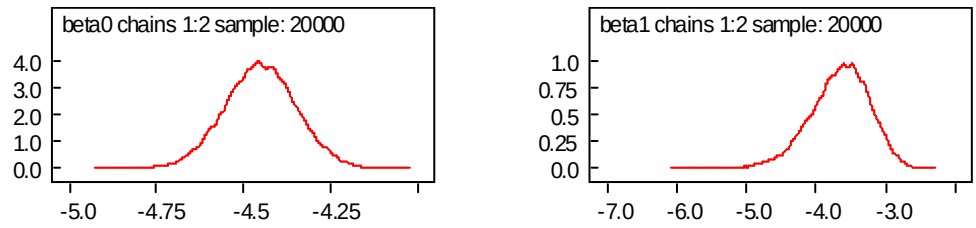


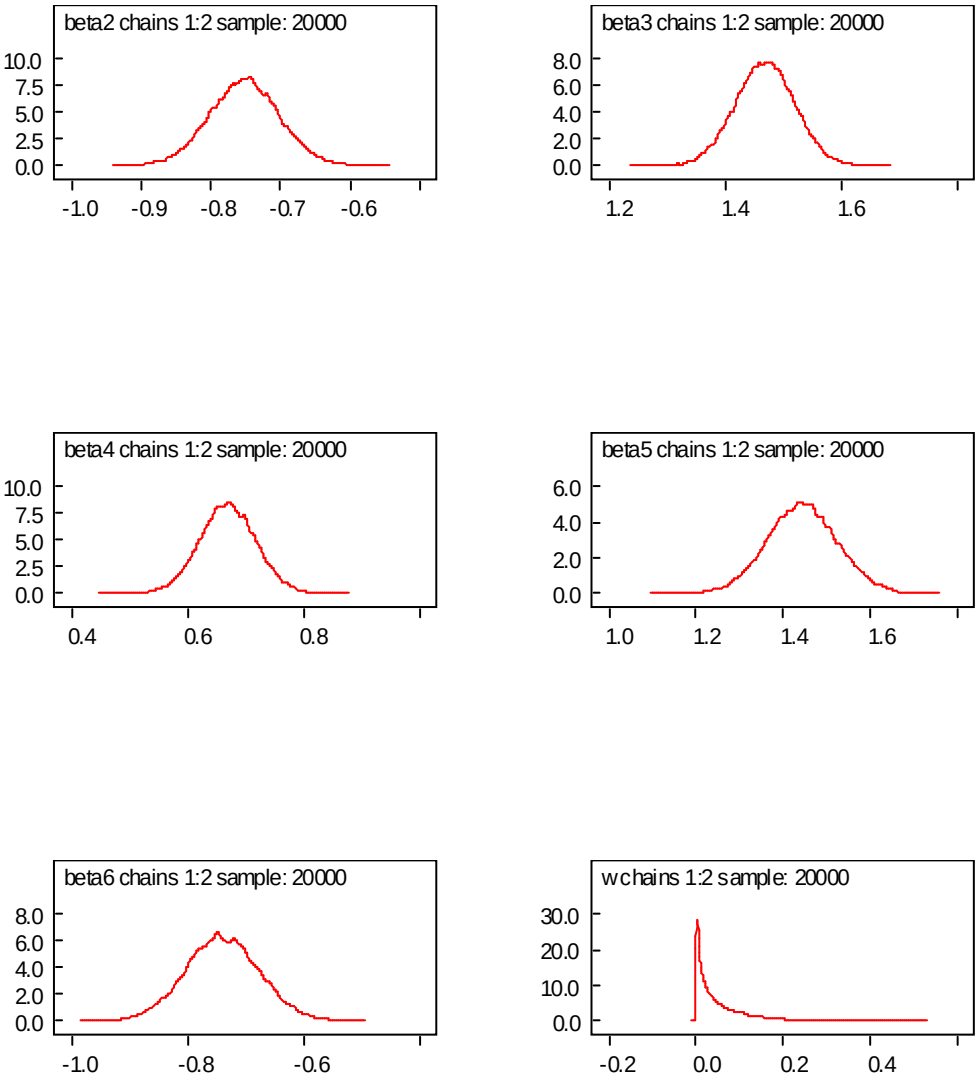
Espera-se que os valores gerados sejam não correlacionados. Observe através da Figura 9.12 os gráficos de autocorrelação. Apesar do parâmetros β_0 e β_5 apresentar uma correlação de lag 1, será considerado que ocorreu a convergência das cadeias.



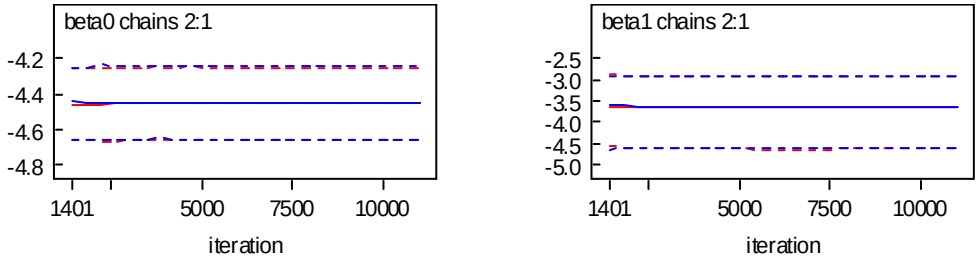


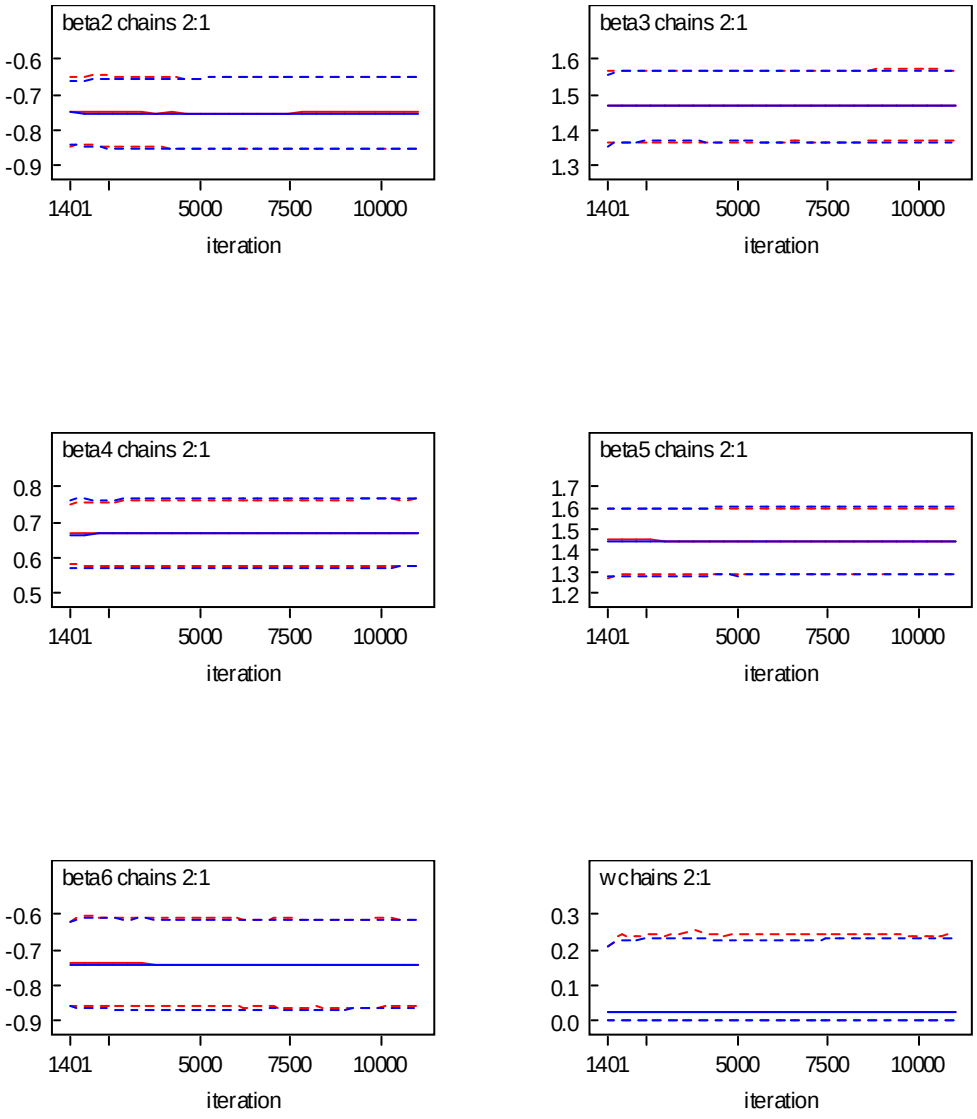
A Figura 9.13 mostra as densidades *a posteriori* dos parâmetros de interesse.





A Figura 9.14 apresenta os intervalos de credibilidade. Observe que nenhum intervalo engloba o valor zero. Dessa forma, os todos os parâmetros são significativos.





A Tabela 9.4 apresenta as estimativas Bayesianas para o modelo logito limitado. Observe que as estimativas estão próximas das estimativas de máxima verossimilhança.

Através dos resultados, pode-se observar que as estimativas Bayesianas dos modelos estão próximas das estimativas clássica. Esse fato ocorreu devido ao uso de prioris não informativas. Devido o tempo computacional para se obter estas estimativas, concluí-se que é melhor utilizar a abordagem Bayesiana apenas quando temos informações para se acrescentar sobre os parâmetros (prioris informativas).

Tabela 9.4: Estimativas Bayesianas para o modelo logito limitado

Parâmetro	Média	Desvio Padrão	2,5%	Mediana	97,5%
ω	0,050	0,064	0,000	0,025	0,237
β_0	4,451	0,103	-4,655	-4,451	-4,251
β_1	-3,661	0,427	-4,585	-3,627	-2,919
β_2	-0,752	0,050	-0,851	-0,751	-0,652
β_3	1,469	0,051	1,368	1,469	1,569
β_4	0,669	0,047	0,575	0,669	0,764
β_5	1,444	0,080	1,289	1,444	1,602
β_6	-0,739	0,063	-0,863	-0,740	-0,614

Capítulo 10

Conclusões

Neste trabalho foi apresentado algumas técnicas de detecção de fraude. Devido a estrutura desbalanceada dos bancos de dados de fraude, foi sugerido o modelo logito limitado. Dessa forma, foi mostrada as estimativas de máxima verossimilhança e testes de verificação de ajuste e significância dos parâmetros para este modelo.

Foi apresentado também as estimativas dos parâmetros do modelo logito e logito limitado considerando as amostras do tipo *state-dependent*. Através das aplicações concluiu-se que para conjuntos de dados extremamente desbalanceados não é recomendável o uso de uma amostra balanceada para a estimação dos parâmetros do modelo.

As aplicações, tanto dos dados simulados quanto dos dados reais, mostraram que o modelo logito limitado se adequa melhor para os dados em estudo. Isto foi observado através das estatísticas de teste AIC, -2logL, SC e KS. Também pôde-se observar que o modelo logito limitado classifica melhor os indivíduos dentro das faixas de escore.

Apresentou-se as estimativas Bayesianas para os modelos logito e logito limitado e as formas de verificar a convergências das cadeias. Como foi consideradas distribuições *a priori* não informativas, as estimativas dos parâmetros estão bem próxima das estimativas clássicas.

Referências Bibliográficas

[1] CRAMER, S., J., **Scoring bank loans that may go wrong: a case study**. In: Statistica Neerlandia, 2004, Vol. 58, n. 3, p. 365-380.

[2] HOSMER, W., D.; LEMESHOW, S., **Applied Logistic Regression**. Ed.: Wiley. Canada, 1989.

[3] CARVALHO, L. A. V.; **Datamining: A mineração de dados no Marketing**. Medicina, Economia, Engenharia e Administração, Editora Érica, - São Paulo, 2001.

[4] QUINLAN, J. R. - **C4.5: Programs for machine learning**, Morgan Kaufmann, Los Altos, 1993.

[5] BHATLA, T. P.; PRABHU, V.; DUA, A., **Understanding Credit Cards Frauds**, Tata Consultancy, 2003

[6] SALVATORE, David, WENKE and Andreas. **Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project**. http://www.cs.columbia.edu/sal/JAM/Project.

[7] MOOD, A., GRAYBILL, F., BOES, D. **Introduction to the theory of statistics**. 3rd. Ed. Singapore: MacGraw Hill, 1974.

[8] Banco Central do Brasil. `http://www.bcb.gov.br/`, Junho 2008 (data de acesso: 23/11/2007).

[9] HOLANDA, Aurélio Buarque De Holanda, **Dicionário Aurélio Da Lingua Portuguesa**, 1990.

[10] GONGDON, Peter **Bayesian Statistical Modelling**, 2003.

[11] GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, B. D. (2003) **Bayesian Data Analysis**. 2^{nd} ed. New York: Chapman & Hall.

[12] GHOSH, S.; REILLY, D. L. **Credit card fraud detection with a neural network. In: Annual Hawaii International Conference on System Sciences**, Wailea, 1994. p. 621-630.

[13] GADI, Manoel Fernando Alonso, LAGO, Alair Pereira **Tópicos em ciência da computação na detecção de fraude**, 2006

[14] HAND, D., ADAMS, N., BOLTON, R., eds.: **Pattern Detection and Discovery**. Springer, 2002

[15] CRISTOFARO, Elizabeth A. U. **Uma Abordagem Bayesiana para Análise de Fraude de Subscrição em Telecomunicações** Dissetação de Mestrado, UFSCAR 2006

[16] JOHNSON, Richard A WICHERN, Dean W. **Applied multivariate statistical analysis**. 2. nd. ed Englewood Cliffs, New Jersey: Prentice –Hall, c1988. 607 p.

[17] HARREL, Frank E. Jr.; **Regression Modeling Strategies**, 2001

[18] PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. (2003) **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian.

[19] DRAPER, N. R.; SMITH H. **Applied regression analysis** John Wiley & Sons, New York, 1981.

[20] SEBER, G. A. F. **Linear regression analysis** John Wiley & Sons, - New York, 1997.

[21] EHLERS, Ricardo S. **Introdução à inferencia Bayesiana**, Curitiba, 2003

[22] HAHNSTRA, Julius **Credit Card Processing as an example of distributed Systems** 2002

[23] CHIB, S.; GREENBERG, E. **Understanding the Metropolis-Hastings algorithm**. The American Statistician, 1995 v.49, 4, p.327-335.

[24] BOX, G. E.; TIAO, G. C. **Bayesian inference in statistical analysis**. 1992 New York: John Wiley, 588p.

[25] DeFINETTI, B.; MACHI, A.; SMITH, A. **Theory of Probability: A Critical Introductory Treatment**. John Wiley & Sons Inc 1974.

[26] GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, B. D. **Bayesian Data Analysis**. 2^{nd} ed. New York: Chapman & Hall 2003.

Apêndice A

Maximização da Função Logito Limitado

Seja a equação de máxima verossimilhança do modelo logito limitado:

$$l(\boldsymbol{\beta}, \omega; \mathbf{x}_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) + (1-y_i) \ln \left(1 - \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \right) \right] I_{(0,1)}(\omega) \tag{A.1}$$

Seja: $P_i = \omega \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}$

Para maximizar essa função é necessário determinar as derividas em relação aos parâmetros $\beta_0, \beta_1, ..., \beta_p, \omega$. Ou seja,

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \omega; \mathbf{x}_i)}{\partial \beta_0} &= 0 \\
\frac{\partial l(\boldsymbol{\beta}, \omega; \mathbf{x}_i)}{\partial \beta_1} &= 0 \\
&\ldots \\
\frac{\partial l(\boldsymbol{\beta}, \omega; \mathbf{x}_i)}{\partial \beta_p} &= 0 \\
\frac{\partial l(\boldsymbol{\beta}, \omega; \mathbf{x}_i)}{\partial \omega} &= 0
\end{aligned}$$

Primeiramente, será calculada a derivada da função em β_0. Considere $p = 1$. Assim, tem-se:

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_0} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i \ln\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right) + (1-y_i)\ln\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right] I_{(0,1)}(\omega)\right)}{\partial\beta_0}$$

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_0} = \sum_{i=1}^{n}\left[\frac{y_i}{\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\beta_0} - \frac{(1-y_i)}{\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\beta_0}\right] \tag{A.2}$$

Calculando $\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\beta_0}$:

$$\begin{aligned}
\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\beta_0} &= \frac{\omega\exp(\beta_0+\beta_1x_i)(1+\exp(\beta_0+\beta_1x_i)) - \omega 2\exp(\beta_0+\beta_1x_i)}{(1+\exp(\beta_0+\beta_1x_i))^2} = \\
&= \frac{\omega\exp(\beta_0+\beta_1x_i)}{(1+\exp(\beta_0+\beta_1x_i))} - \frac{\omega 2\exp(\beta_0+\beta_1x_i)}{(1+\exp(\beta_0+\beta_1x_i))^2} = \\
&= \omega\left[\frac{\exp(\beta_0+\beta_1x_i)}{1+\exp(\beta_0+\beta_1x_i)} - \frac{2\exp(\beta_0+\beta_1x_i)}{(1+\exp(\beta_0+\beta_1x_i))^2}\right] = \\
&= \omega P_i(1-P_i)
\end{aligned} \tag{A.3}$$

Retornando a A.2 tem-se:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial\beta_0} &= \sum_{i=1}^{n}\left[\frac{y_i}{\omega P_i}\omega P_i(1-P_i) - \frac{(1-y_i)}{(1-P_i)}\omega P_i(1-P_i)\right] = \\
&= \sum_{i=1}^{n}\left[y_i\omega - P_iy_i\omega - \omega P_i + P_iy_i\omega\right] = \\
&= \sum_{i=1}^{n}\left[y_i\omega - \omega P_i\right] = 0
\end{aligned} \tag{A.4}$$

Agora será calculada a derivada da função em β_1. Assim, tem-se:

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_1} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i \ln\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right) + (1-y_i)\ln\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right] I_{(0,1)}(\omega)\right)}{\partial \beta_1} \tag{A.5}$$

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_1} = \sum_{i=1}^{n}\left[\frac{y_i}{\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial \beta_1} - \frac{(1-y_i)}{\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial \beta_1}\right] \tag{A.6}$$

Calculando $\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial \beta_1}$:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_1} &= \frac{\omega\exp(\beta_0+\beta_1 x_i)x_i(1+\exp(\beta_0+\beta_1 x_i)) - \omega 2\exp(\beta_0+\beta_1 x_i)x_i}{(1+\exp(\beta_0+\beta_1 x_i))^2} = \\
&= \frac{\omega\exp(\beta_0+\beta_1 x_i)x_i}{(1+\exp(\beta_0+\beta_1 x_i))} - \frac{\omega 2\exp(\beta_0+\beta_1 x_i)x_i}{(1+\exp(\beta_0+\beta_1 x_i))^2} = \\
&= \omega x_i P_i\left(1-P_i\right)
\end{aligned} \tag{A.7}$$

Retornando a A.6 tem-se:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \beta_1} &= \sum_{i=1}^{n}\left[\frac{y_i}{P_i}\omega x_i P_i\left(1-P_i\right) - \frac{(1-y_i)}{(1-P_i)}\omega x_i P_i\left(1-P_i\right)\right] = \\
&= \sum_{i=1}^{n}\left[y_i\omega x_i - y_i\omega x_i P_i - \omega x_i P_i + y_i\omega x_i P_i\right] = \\
&= \sum_{i=1}^{n}\left[y_i\omega x_i - \omega x_i P_i\right] = \\
&= \sum_{i=1}^{n}\omega x_i(y_i - P_i) = 0
\end{aligned} \tag{A.8}$$

As derivadas em relação a $\beta_2, ..., \beta_p$ são dadas da mesma forma. Dessa maneira, tem-se:

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \boldsymbol{\beta}} = \sum_{k=1}^{p}\sum_{i=1}^{n}\omega x_{ik}(y_i - P_i) = 0 \tag{A.9}$$

A derivada da função em ω é dada da seguinte forma:

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \omega} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i \ln\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right) + (1-y_i)\ln\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right] I_{(0,1)}(\omega)\right)}{\partial\omega} \tag{A.10}$$

$$\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial \omega} = \sum_{i=1}^{n}\left[\frac{y_i}{\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\omega} - \frac{(1-y_i)}{\left(1-\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\omega}\right] \tag{A.11}$$

Calculando $\frac{\partial\left(\omega\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)}{\partial\omega}$:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial\omega} &= \frac{\exp(\beta_0+\beta_1 x_i)(1+\exp(\beta_0+\beta_1 x_i)) - \omega\exp(\beta_0+\beta_1 x_i)0}{(1+\exp(\beta_0+\beta_1 x_i))^2} = \\ &= \frac{\exp(\beta_0+\beta_1 x_i)}{(1+\exp(\beta_0+\beta_1 x_i))} = P_i \end{aligned} \tag{A.12}$$

Voltando em A.11 tem-se:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta},\omega;\mathbf{x}_i)}{\partial\omega} &= \sum_{i=1}^{n}\left[\frac{y_i}{P_i}P_i - \frac{(1-y_i)}{(1-P_i)}P_i\right] = \\ &= \sum_{i=1}^{n}\left[\frac{y_i(1-y_i)-(1-y_i)P_i}{(1-P_i)}\right] = \\ &= \sum_{i=1}^{n}\left[\frac{y_i-P_i}{(1-P_i)}\right] = 0 \end{aligned} \tag{A.13}$$

Apêndice B

Programa SAS - Simulação do banco de dados

Primeiramente buscou-se os valores dos parâmetros de modo que a probabilidade de sucesso fosse bem pequena. Após esta determinação, utilizou-se o seguinte código SAS para gerar os dados:

```
proc iml; /* trabalhando com matriz */
beta0= -4.5;
beta1= -3.5;
beta2= -0.8;
beta3= 1.5;
beta4= 0.7;
beta5= 1.5;
beta6= -0.8;
/* valores reais dos parametros */
n=100000; /* tamanho amostral */
x1 = J(1,n,0);
x2 = J(1,n,0); /* definindo as matrizes de B linhas e n colunas */
x3 = J(1,n,0);
x4 = J(1,n,0); /* definindo as matrizes de B linhas e n colunas */
x5 = J(1,n,0);
x6 = J(1,n,0); /* definindo as matrizes de B linhas e n colunas */
```

```
y = J(1,n,0);
p = J(1,n,0);
s = J(1,n,0);
semente = 765459685;
do j = 1 to n;
x1[1,j] = ranbin(semente,1,.1); /* amostra de Bernoulli */
x2[1,j] = ranbin(semente,1,.8);
    x3[1,j] = ranbin(semente,1,.1); /* amostra de Bernoulli */
x4[1,j] = ranbin(semente,1,.4);
    x5[1,j] = ranbin(semente,1,.7); /* amostra de Bernoulli */
x6[1,j] = ranbin(semente,1,.9);
end;
s[1,] = beta0+beta1*x1[1,]+beta2*x2[1,]+beta3*x3[1,]+beta4*x4[1,]+beta5*x5[1,]+beta6*x6[1,];
p[1,] = exp(s[1,])/(1+exp(s[1,])); /* funçao logistica */
    do k = 1 to n;
y[1,k] = ranbin(0,1,p[1,k]); /* amostra de Bernoulli */
end;
C=t(y[1,])||t(x1[1,])||t(x2[1,])||t(x3[1,])||t(x4[1,])||t(x5[1,])||t(x6[1,])||t(p[1,]); /* criando
os data set */
cname = {"fraudex1_D1x1_D2x1_D3x2_D1x2_D2x2_D3p"};
create artigo.sim_4cat_1pcento from C [ colname=cname ];
append from C;
run;
quit;
```

Apêndice C

Programa SAS - Ajuste Modelo Logito

O programa abaixo é referente ao ajuste do modelo logito para os dados simulados. Não há necessidade de se colocar também o programa para os dados reais pois muda apenas o banco de dados.

```
/*Modelo para dados simulados*/
proc logistic data = artigo.sim_4cat_1pcento;
class x1_D1 (ref='0')
      x1_D2 (ref='1')
      x1_D3 (ref='0')
      x2_D1 (ref='1')
      x2_D2 (ref='1')
      x2_D3 (ref='1')
/param=ref;
model fraude = x1_D1 x1_D2 x1_D3 x2_D1 x2_D2 x2_D3/lackfit ctable pprob =
(.05 to 1.0 by .05) outroc=curva_roc;
output out = artigo.sim_4cat_1pcento_saida prob=p ;
run;
/*Categorizar o escore em 20 classes*/
data artigo.sim_4cat_1pcento_saida;
set artigo.sim_4cat_1pcento_saida;
```

```
escore = p2*1000;
run;
```

Apêndice D

Programa SAS - Ajuste Modelo Logito Limitado

O programa abaixo é referente ao ajuste do modelo logito limitado para os dados simulados. Não há necessidade de se colocar também o programa para os dados reais pois muda apenas o banco de dados.

```
proc nlp data=artigo.sim_4cat_1pcento cov=2 VARDEF=N OUTEST=artigo;
max l;
parms w = 0.005,
        beta0 = -2 , /*chutes iniciais */
         beta1 = 0.02,
        beta2 = 0.02,
        beta3 = 0.02,
        beta4 = 0.02,
        beta5 = 0.02,
         beta6 = 0.02;
bounds 0 < w < 1;
l = fraude*log(w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 +
beta4*x2_D1 + beta5*x2_D2 + beta6*x2_D3)/
(1 + exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)))+
(1-fraude)*log(1 - (w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 +
```

```
beta4*x2_D1 + beta5*x2_D2 + beta6*x2_D3)/
    (1+exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3))));
    run;
    data artigo;
    set artigo;
    where _TYPE_ = 'PARMS';
    run;
    /*Probabilidades*/
    /*Probabilidade de não fraudar*/
    proc sql;
    create table artigo.sim_4cat_1pcento_limitada as select *,
    (2.7182818228**(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3)) as exp
    from artigo.sim_4cat_1pcento, artigo;
    alter table artigo.sim_4cat_1pcento_limitada
    add Prob_bom num;
    update artigo.sim_4cat_1pcento_limitada
    set Prob_bom = 1-(w*exp/(1+exp));
    quit;
    data artigo.sim_4cat_1pcento_limitada;
    set artigo.sim_4cat_1pcento_limitada;
    escore = Prob_bom*1000;
    run;
    /*criando um novo conjunto para não 'estragar' o que já está certo*/
    /*-2LogL*/
    /*intercepto mais covariáveis*/
    data artigo.sim_4cat_1pcento_limitada_teste;
    set artigo.sim_4cat_1pcento_limitada;
    run;
    data artigo.sim_4cat_1pcento_limitada_teste;
```

```
set artigo.sim_4cat_1pcento_limitada_teste;
L_2 = -2*(fraude*log(w*exp/(1+exp))+(1-fraude)*log(1-(w*exp/(1+exp))));
run;
proc means data = artigo.sim_4cat_1pcento_limitada_teste sum;
var L_2;
run;
/*16534.63*/
/*só intercepto*/
data artigo.sim_4cat_1pcento_limitada_teste;
set artigo.sim_4cat_1pcento_limitada_teste;
L_2_int = -2*(fraude*log(w*exp(beta0)/(1+exp(beta0)))+(1-fraude)*
          log(1-(w*exp(beta0)/(1+exp(beta0)))));
run;
proc means data = artigo.sim_4cat_1pcento_limitada_teste sum;
var L_2_int;
run;
/*18843.25*/
/*AIC*/
/*intercepto mais covariáveis*/
data artigo.sim_4cat_1pcento_limitada_teste;
set artigo.sim_4cat_1pcento_limitada_teste;
/*só intercepto*/
AIC1=2*1+18843.25;
/*intercepto mais variáveis*/
AIC2=2*7+16534.63;
run;
/*só intercepto: 18845.25*/
/*intercepto mais variáveis: 16548.63*/
/*Schwarz criteria - SC*/
/*intercepto mais covariáveis*/
data artigo.sim_4cat_1pcento_limitada_teste;
```

```
set artigo.sim_4cat_1pcento_limitada_teste;
/*só intercepto*/
SC1=1*log(100000)+18843.25;
/*intercepto mais variáveis*/
SC2=7*log(100000)+16534.63;
run;
/*só intercepto: 18854.762925*/
/*intercepto mais variáveis: 16615.220478*/
```

Apêndice E

Programa SAS - Amostras para análise das amostras do Tipo *State-dependent*

```
/*100 amostras - k=1 n=1860*/
    %macro caso_k1;
    %do i = 1 %to 100;
    proc surveyselect data=artigo.sim_4cat_1pcento_zeros
    method=srs
    n=1860
    seed=&i
    out=state.dados0_sim_k1_&i; /* refere a analise balanceada */;
    strata fraude;
    ID fraude x1_D1 x1_D2 x1_D3 x2_D1 x2_D2 x2_D3 p;
    run;
    data state.amostrak1_&i;
    set artigo.sim_4cat_1pcento_uns state.dados0_sim_k1_&i;
    run;
    proc nlp data=state.amostrak1_&i cov=2 VARDEF=N OUTEST=artigo_k1_&i;
    max l;
    parms beta0 = -2 , /*chutes iniciais */
```

```
        beta1 = 0.02,
        beta2 = 0.02,
        beta3 = 0.02,
        beta4 = 0.02,
        beta5 = 0.02,
        beta6 = 0.02;
    l=fraude*log(exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3)/
    (exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+0.018952517))+
    (1-fraude)*log(1-(exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3))/
    (exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+0.018952517))
    ;
    run;
    %end;
    %mend;
    %caso_k1;
    data state.k1;
    set artigo1_k1_1 artigo1_k1_2 artigo1_k1_3 artigo1_k1_4 artigo1_k1_5 artigo1_k1_6
artigo1_k1_7 artigo1_k1_8
    artigo1_k1_9 artigo1_k1_10 artigo1_k1_11 artigo1_k1_12 artigo1_k1_13 artigo1_k1_14
artigo1_k1_15 artigo1_k1_16
    artigo1_k1_17 artigo1_k1_18 artigo1_k1_19 artigo1_k1_20 artigo1_k1_21 artigo1_k1_22
artigo1_k1_23 artigo1_k1_24
    artigo1_k1_25 artigo1_k1_26 artigo1_k1_27 artigo1_k1_28 artigo1_k1_29 artigo1_k1_30
artigo1_k1_31 artigo1_k1_32
    artigo1_k1_33 artigo1_k1_34 artigo1_k1_35 artigo1_k1_36 artigo1_k1_37 artigo1_k1_38
artigo1_k1_39 artigo1_k1_40
    artigo1_k1_41 artigo1_k1_42 artigo1_k1_43 artigo1_k1_44 artigo1_k1_45 artigo1_k1_46
```

```
artigo1_k1_47 artigo1_k1_48
    artigo1_k1_49 artigo1_k1_50 artigo1_k1_51 artigo1_k1_52 artigo1_k1_53 artigo1_k1_54
artigo1_k1_55 artigo1_k1_56
    artigo1_k1_57 artigo1_k1_58 artigo1_k1_59 artigo1_k1_60 artigo1_k1_61 artigo1_k1_62
artigo1_k1_63 artigo1_k1_64
    artigo1_k1_65 artigo1_k1_66 artigo1_k1_67 artigo1_k1_68 artigo1_k1_69 artigo1_k1_70
artigo1_k1_71 artigo1_k1_72
    artigo1_k1_73 artigo1_k1_74 artigo1_k1_75 artigo1_k1_76 artigo1_k1_77 artigo1_k1_78
artigo1_k1_79 artigo1_k1_80
    artigo1_k1_81 artigo1_k1_82 artigo1_k1_83 artigo1_k1_84 artigo1_k1_85 artigo1_k1_86
artigo1_k1_87 artigo1_k1_88
    artigo1_k1_89 artigo1_k1_90 artigo1_k1_91 artigo1_k1_92 artigo1_k1_93 artigo1_k1_94
artigo1_k1_95 artigo1_k1_96
    artigo1_k1_97 artigo1_k1_98 artigo1_k1_99 artigo1_k1_100;
    run;
```

Apêndice F

Programa SAS - Modelo Logito Considerando Amostras do Tipo *State-dependent*

```
/*Modelos*/
    /*Amostra balanceada - K=1*/
    proc nlp data=artigo.amostrak1 cov=2;
    max l;
    parms beta0 = -2 , /*chutes iniciais */
          beta1 = 0.02,
          beta2 = 0.02,
          beta3 = 0.02,
          beta4 = 0.02,
          beta5 = 0.02,
          beta6 = 0.02;
    l=fraude*log(exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3)/
    (exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+0.018952517))+
    (1-fraude)*log(1-(exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3))/
```

```
(exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+0.018952517))
;
run;
```

Apêndice G

Programa SAS - Modelo Logito Considerando Amostras do Tipo *State-dependent*

```
/*Amostra balanceada - K=1*/
    proc nlp data=artigo.amostrak1 cov=2;
    max l;
    parms beta0 = -2 , /*chutes iniciais */
          beta1 = 0.02,
          beta2 = 0.02,
          beta3 = 0.02,
          beta4 = 0.02,
          beta5 = 0.02,
          beta6 = 0.02,
          w = 0.01;
    bounds 0 < w <1;
    l=fraude*log((w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3))/
    (w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+
    0.018952517*(1+exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
```

```
+ beta5*x2_D2 + beta6*x2_D3)-
    w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3))))+
    (1-fraude)*log(1-((w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 +
beta4*x2_D1 + beta5*x2_D2 + beta6*x2_D3))/
    (w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)+
    0.018952517*(1+exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1
+ beta5*x2_D2 + beta6*x2_D3)-
    w*exp(beta0 + beta1*x1_D1 + beta2*x1_D2 + beta3*x1_D3 + beta4*x2_D1 +
beta5*x2_D2 + beta6*x2_D3)))));
    run;
```

Apêndice H

Programa Winbugs - Modelo Logito

```
model{
    for(i in 1:N){
    F[i]~dbern(p[i])
    logit(p[i]) <- beta0 + beta1*x1d1[i] + beta2*x1d2[i] + beta3*x1d3[i] + beta4*x2d1[i]
+ beta5*x2d2[i] + beta6*x2d3[i]
    }
    beta0 ~dflat()
    beta1 ~dflat()
    beta2 ~dflat()
    beta3 ~dflat()
    beta4 ~dflat()
    beta5 ~dflat()
    beta6 ~dflat()
    }
    list(
    F = c(
    #inserir dados
    ), N=100000)
    list(beta0=-0.9, beta1=2.3, w=0.01)
    list(beta0= -1.1487, beta1=0.2701, beta2= 0.1507, beta3= 0.2302, beta4=-0.2667,
    beta5=0.2928, beta6= 0.2357, beta7=-0.2745, beta8= 0.6152,
```

beta9= 0.5868, beta10=0.2453, beta11=-0.4074)

Apêndice I

Programa Winbugs - Modelo Logito Limitado

```
model{
    for(i in 1:N){
    F[i]~dbern(p[i])
    logit(p[i]) <- beta0 + beta1*x1d1[i] + beta2*x1d2[i] + beta3*x1d3[i] + beta4*x2d1[i]
+ beta5*x2d2[i] + beta6*x2d3[i]
    prop[i] <- log(w) + logit(p[i])
    }
    beta0 ~dflat()
    beta1 ~dflat()
    beta2 ~dflat()
    beta3 ~dflat()
    beta4 ~dflat()
    beta5 ~dflat()
    beta6 ~dflat()
    w ~dbeta(0.53,10)
    }
    list(
    F = c(
    #inserir dados
```

), N=100000)

list(beta0= 0, beta1=0, beta2= 0, beta3= 0, beta4=0, beta5=0, beta6= 0, w=0.5)

list(beta0= 1, beta1=1, beta2= 1, beta3= 1, beta4=1, beta5=1, beta6= 1, w=0.5)