Métodos Estatísticos para Análise de Dados de Crédito

Carlos Diniz
Francisco Louzada

6th Brazilian Conference
on Statistical Modelling in Insurance and Finance
Maresias - SP
Março / 2013

Métodos Estatísticos para Análise de Dados de Crédito

Carlos Diniz
DEs–UFSCar

Francisco Louzada
ICMC–USP

Colaboradores

Hélio J. Abreu
Paulo H. Ferreira
Ricardo F. Rocha
Agatha S. Rodrigues
Fernanda N. Scacabarozi

6th Brazilian Conference
on Statistical Modelling in Insurance and Finance
Março 2013
Maresias - SP

Sumário

1 Introdução à Modelagem de *Credit Scoring* 1

1.1 Etapas de Desenvolvimento 3
1.2 Planejamento Amostral 3
1.2.1 Descrição de um problema - *Credit Scoring* 8
1.3 Determinação da Pontuação de Escore 9
1.3.1 Transformação e seleção de variáveis 11
1.3.2 Regressão logística 12
1.4 Validação e Comparação dos Modelos 15
1.4.1 A estatística de Kolmogorov-Smirnov (KS) 16
1.4.2 Curva ROC 19
1.4.3 Capacidade de acerto dos modelos 22

2 Regressão Logística 25

2.1 Estimação dos Coeficientes 26
2.2 Intervalos de Confiança e Seleção de Variáveis 28
2.3 Interpretação dos Coeficientes do Modelo 30
2.4 Aplicação 31
2.5 Amostras *State-Dependent* 34
2.5.1 Método de correção a priori 36
2.6 Estudo de Comparação 37
2.6.1 Medidas de desempenho 38
2.6.2 Probabilidades de inadimplência estimadas 39
2.7 Regressão Logística com Erro de Medida 41
2.7.1 Função de verossimilhança 42
2.7.2 Métodos de estimação 43
2.7.3 Renda presumida 43

3 Modelagem Para Eventos Raros **46**
3.1 Estimadores KZ para o Modelo de Regressão Logística . 47
3.1.1 Correção nos parâmetros 48
3.1.2 Correção nas probabilidades estimadas 49
3.2 Modelo Logito Limitado 51
3.2.1 Estimação 52
3.2.2 Método BFGS 53
3.3 Modelo Logito Generalizado 54
3.3.1 Estimação 56
3.4 Modelo Logito com Resposta de Origem 58
3.4.1 Modelo normal 58
3.4.2 Modelo exponencial 60
3.4.3 Modelo lognormal 60
3.4.4 Estudo de simulação 61
3.5 Análise de Dados Reais 64

4 *Credit Scoring* com Inferência dos Rejeitados **68**
4.1 Métodos de Inferência dos Rejeitados 69
4.1.1 Método da reclassificação 69
4.1.2 Método da ponderação 70
4.1.3 Método do parcelamento 71
4.1.4 Outros métodos 72
4.2 Aplicação 73

5 Combinação de Modelos de *Credit Scoring* **77**
5.1 *Bagging* de Modelos 77
5.2 Métodos de Combinação 79
5.2.1 Combinação via média 79
5.2.2 Combinação via voto 80
5.2.3 Combinação via regressão logística 81
5.3 Aplicação 81

6 Análise de Sobrevivência **86**
6.1 Algumas Definições Usuais 87
6.2 Modelo de Cox 91
6.2.1 Modelo para comparação de dois perfis de clientes 92

6.2.2 A generalização do modelo de riscos proporcionais 93
6.2.3 Ajuste de um modelo de riscos proporcionais . . . 95
6.2.4 Tratamento de empates 100
6.3 Intervalos de Confiança e Seleção de Variáveis 103
6.4 Estimação da Função de Risco e Sobrevivência 104
6.5 Interpretação dos Coeficientes 106
6.6 Aplicação 108

7 Modelo de Longa Duração 112
7.1 Modelo de Mistura Geral 112
7.2 Estimação do modelo longa duração geral 114
7.3 Aplicação 116

Capítulo 1

Introdução à Modelagem de *Credit Scoring*

A partir de 1933, ano da publicação do primeiro volume da revista Econometrica, intensificou-se o desenvolvimento de métodos estatísticos para, dentre outros objetivos, testar teorias econômicas, avaliar e implementar políticas comerciais, estimar relações econômicas e dar suporte à concessão de crédito.

Os primeiros modelos de *Credit Scoring* foram desenvolvidos entre os anos 40 e 50 e a metodologia básica, aplicada a esse tipo de problema, era orientada por métodos de discriminação produzidos por Fisher (1936). Podemos dizer que foi de Durand (1941) o primeiro trabalho conhecido que utilizou análise discriminante para um problema de crédito, em que as técnicas desenvolvidas por Fisher foram empregadas para discriminar *bons* e *maus* empréstimos.

Henry Markowitz (Markowitz, 1952) foi um dos pioneiros na criação de um modelo estatístico para o uso financeiro, o qual foi utilizado para medir o efeito da diversificação no risco total de uma carteira de ativos.

Fischer Black e Myron Scholes (Black & Scholes, 1973) desenvolveram um modelo clássico para a precificação de uma opção, uma das mais importantes fórmulas usadas no mercado financeiro.

Diretores do *Citicorp*, em 1984, lançaram o livro *Risco e Recompensa: O Negócio de Crédito ao Consumidor*, com as primeiras menções

ao modelo de *Credit Scoring*, que é um tipo de modelo de escore, baseado em dados cadastrais dos clientes, e é utilizado nas decisões de aceitação de proponentes a créditos; ao modelo de *Behaviour Scoring*, que é um modelo de escore, baseado em dados transacionais, utilizado nas decisões de manutenção ou renovação de linhas e produtos para os já clientes e ao modelo *Collection Scoring*, que é também um modelo de escore, baseado em dados transacionais de clientes inadimplentes, utilizado nas decisões de priorização de estratégias de cobranças. Estes e vários outros modelos são utilizados como uma das principais ferramentas de suporte à concessão de crédito em inúmeras instituições financeiras no mundo.

Na realidade, os modelos estatísticos passaram a ser um importante instrumento para ajudar os gestores de risco, gestores de fundos, bancos de investimento, gestores de créditos e gestores de cobrança a tomarem decisões corretas e, por esta razão, as instituições financeiras passaram a aprimorá-los continuamente. Em especial, a concessão de crédito ganhou força na rentabilidade das empresas do setor financeiro, se tornando uma das principais fontes de receita e, por isso, rapidamente, este setor percebeu a necessidade de se aumentar o volume de recursos concedidos sem perder a agilidade e a qualidade dos empréstimos, e nesse ponto a contribuição da modelagem estatística foi essencial.

Diferentes tipos de modelos são utilizados no problema de crédito, com o intuito de alcançar melhorias na redução do risco e/ou no aumento da rentabilidade. Entre os quais, podemos citar, a regressão logística e linear, análise de sobrevivência, redes probabilísticas, árvores de classificação, algoritmos genéticos e redes neurais. Neste livro tratamos de diferentes problemas presentes na construção de modelos de regressão logística para *Credit Scoring* e sugerimos metodologias estatísticas para resolvê-los. Além disso, apresentamos metodologias alternativas de análise de sobrevivência e redes probabilísticas.

O processo de desenvolvimento de um modelo de crédito envolve várias etapas, entre as quais *Planejamento Amostral*, *Determinação da Pontuação de Escore* e *Validação e Comparação de Modelos*. Apresentamos nas próximas seções discussões sobre algumas destas etapas.

1.1 Etapas de Desenvolvimento

O desenvolvimento de um modelo de *Credit Scoring* consiste, de uma forma geral, em determinar uma função das variáveis cadastrais dos clientes que possa auxiliar na tomada de decisão para aprovação de crédito, envolvendo cartões de créditos, cheque especial, atribuição de limite, financiamento de veículo, imobiliário e varejo.

Normalmente esses modelos são desenvolvidos a partir de bases históricas de performance de crédito dos clientes e também de informações pertinentes ao produto. O desenvolvimento de um modelo de *Credit Scoring* (Sicsú, 1998) compreende nas seguintes etapas:

i) Planejamento e definições;

ii) Identificação de variáveis potenciais;

iii) Planejamento amostral;

iv) Determinação do escore: aplicação da metodologia estatística;

v) Validação e verificação de performance do modelo estatístico;

vi) Determinação do ponto de corte ou faixas de escore;

vii) Determinação de regra de decisão.

As etapas iii), iv) e v), por estarem associadas à modelagem, são apresentadas com mais detalhes nas próximas seções.

1.2 Planejamento Amostral

Para a obtenção da amostra, na construção de um modelo de *Credit Scoring*, é importante que definições como, para qual produto ou família de produtos e para qual ou quais mercados o modelo será desenvolvido, sejam levadas em consideração. A base de dados utilizada para a construção de um modelo é formada por clientes cujos créditos foram concedidos e seus desempenhos foram observados durante um período de tempo no passado. Esse passado, cujas informações são retiradas, deve

ser o mais recente possível a fim de que não se trabalhe com operações de crédito remotas que não sejam representativas da realidade atual.

Uma premissa fundamental na construção de modelos de *Credit Scoring*, e preditivos em geral, é que a forma como as variáveis cadastrais se relacionaram com o desempenho de crédito no passado, seja similar no futuro.

Um fator importante a ser considerado na construção do modelo é o horizonte de previsão, sendo necessário estabelecer um espaço de tempo para a previsão do *Credit Scoring*, ou seja, o intervalo entre a solicitação do crédito e a classificação como *bom* ou *mau* cliente. Esse será também o intervalo para o qual o modelo permitirá fazer as previsões de quais indivíduos serão mais ou menos prováveis de se tornarem inadimplentes ou de serem menos rentáveis. A regra é de 12 a 18 meses, porém na prática observamos que um intervalo de 12 meses é o mais utilizado.

Thomas *et al.* (2002) também propõe um período de 12 meses para modelos de *Credit Scoring,* sugerindo que a taxa de inadimplência dos clientes das empresas financeiras em função do tempo aumenta no início, estabilizando somente após 12 meses. Assim, qualquer horizonte mais breve do que esse pode não refletir de forma real o percentual de *maus* clientes prejudicando uma possível associação entre as características dos indivíduos e o evento de interesse modelado, no caso, a inadimplência. Por outro lado, a escolha de um intervalo de tempo muito longo para o horizonte de previsão também pode não trazer benefícios, fazendo com que a eficácia do modelo diminua, uma vez que, pela distância temporal, os eventos se tornam pouco correlacionados com potenciais variáveis cadastrais, normalmente, obtidas no momento da solicitação do crédito.

O fator tempo tem uma importância fundamental na construção de modelos preditivos e, de forma geral, tem três importantes etapas, como mostra a Figura 1.1. O passado é composto pelas operações para as quais já foram observados os desempenhos de crédito durante um horizonte de previsão adotado. As informações cadastrais dos clientes no momento da concessão do crédito, levantadas no passado mais distante, são utilizadas como variáveis de entrada para o desenvolvimento do modelo e os dados do passado mais recente, as observações dos de-

sempenhos de crédito dos clientes, *default* ou não *default*, inadimplentes ou adimplentes, são utilizados para a determinação da variável resposta.

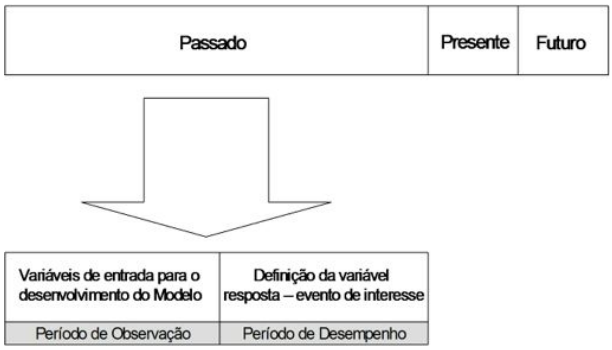

Figura 1.1: Estrutura temporal das informações para construção de modelos preditivos.

É importante ressaltar que as variáveis de entrada para a construção do modelo sejam baseadas em informações, que necessariamente, ocorreram antes de qualquer informação utilizada para gerar a variável resposta de interesse. Se dividirmos o passado em períodos de *observação* e *desempenho*. O período de *observação* compreende o período de tempo no qual são obtidas e observadas as informações potencialmente relevantes para o evento de interesse, ou seja, o período em que se constrói e obtém as variáveis explanatórias. Em um modelo de *Credit Scoring* esse período compreende na realidade um único instante, sendo o momento em que um cliente busca obter um produto de crédito, podendo ser chamado de *ponto de observação*. O período de *desempenho* é o intervalo de tempo em que é observado a ocorrência ou não do evento de interesse. Esse período corresponde a um intervalo de tempo do mesmo tamanho do horizonte de previsão adotado para a construção do modelo. O presente corresponde ao período de desenvolvimento do modelo em que, normalmente, as informações referentes a esse período ainda não estão disponíveis, uma vez que estão sendo geradas pelos sistemas das instituições. O futuro é o período de tempo para o qual serão feitas as predições, utilizando-se de informações do presente, do passado e das relações entre estas, que foram determinadas na construção do modelo.

Um alerta importante é que modelos preditivos, construídos a

partir de dados históricos, podem se ajustar bem no passado, possuindo uma boa capacidade preditiva. Porém, o mesmo não ocorre quando aplicados a dados mais recentes. A performance desses modelos pode ser afetada também pela raridade do evento modelado, em que existe dificuldade em encontrar indivíduos com o atributo de interesse. No contexto de *Credit Scoring* isso pode ocorrer quando a amostra é selecionada pontualmente, em um único mês, semana etc, não havendo número de indivíduos suficientes para encontrar as diferenças de padrões desejadas entre *bons* e *maus* pagadores. Dessa forma, o dimensionamento da amostra é um fator extremamente relevante no desenvolvimento de modelos de *Credit Scoring*.

A utilização de um tratamento estatístico formal para determinar o tamanho da amostra seria complexa, dependendo de vários fatores como o número e o tipo de variáveis envolvidas no estudo.

Dividir a amostra em duas partes, treinamento (ou desenvolvimento) e teste (ou validação), é conveniente e resulta em benefícios técnicos. Isto é feito para que possamos verificar o desempenho e comparar os disponíveis modelos. É interessante que a amostra seja suficientemente grande de forma que permita uma possível divisão desse tipo. Porém, sempre que possível, essa divisão jamais deve substituir a validação de modelos em um conjunto de dados mais recente. Lewis (1994) sugere que, em geral, amostras com tamanhos menores de 1500 clientes *bons* e 1500 *maus*, podem inviabilizar a construção de modelos com capacidade preditiva aceitável para um modelo de *Credit Scoring,* além de não permitir a sua divisão.

Em grande parte das aplicações de modelagem com variável resposta binária, um desbalanceamento significativo, muitas vezes da ordem de 20 *bons* para 1 *mau*, é observado entre o número de *bons* e *maus* pagadores nas bases de clientes das instituições. Essa situação pode prejudicar o desenvolvimento do modelo, uma vez que o número de *maus* pode ser muito pequeno e insuficiente para estabelecer perfis com relação às variáveis explanatórias e também para observar possíveis diferenças em relação aos *bons* cliente. Dessa forma, uma amostragem aleatória simples nem sempre é indicada para essa situação, sendo necessária a utilização de uma metodologia denominada *Oversampling* ou *State Depen-*

dent, que consiste em aumentar a proporção do evento raro, ou, mesmo não sendo tão raro, da categoria que menos aparece na amostra. Esta técnica trabalha com diferentes proporções de cada categoria, sendo conhecida também como *amostra aleatória estratificada*. Mais detalhes a respeito da técnica *State Dependent* são apresentados no Capítulo 2.

Berry & Linoff (2000) expressam, em um problema com a variável resposta assumindo dois resultados possíveis, a idéia de se ter na amostra de desenvolvimento para a categoria mais rara ou menos frequente entre 10% e 40% dos indivíduos. Thomas *et al.* (2002) sugere que as amostras em um modelo de *Credit Scoring* tendem a estar em uma proporção de 1:1, de *bons* e *maus* clientes, ou algo em torno desse valor. Uma situação típica de ocorrer é selecionar todos os *maus* pagadores possíveis juntamente com uma amostra de mesmo tamanho de *bons* pagadores para o desenvolvimento do modelo. Nos casos em que a variável resposta de interesse possui distribuição dicotômica extremamente desbalanceada, algo em torno de 3% ou menos de eventos, comum quando o evento de interesse é fraude, existem alguns estudos que revelam que o modelo de regressão logística usual subestima a probabilidade do evento de interesse (King & Zeng, 2001). Além disso, os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística são viciados nestes casos. O Capítulo 3 apresenta uma metodologia específica para situação de eventos raros.

A sazonalidade na ocorrência do evento modelado é um outro fator a ser considerado no planejamento amostral. Por exemplo, a seleção da amostra envolvendo momentos específicos no tempo em que o comportamento do evento é atípico, pode afetar e comprometer diretamente o desempenho do modelo. Outro aspecto não menos importante é com relação a variabilidade da ocorrência do evento, uma vez que pode estar sujeito a fatores externos e não-controláveis, como por exemplo a conjuntura econômica, que faz com que a seleção da amostra envolva cenários de não-representatividade da mesma com relação ao evento e assim uma maior instabilidade do modelo.

Uma alternativa de delineamento amostral que minimiza o efeito desses fatores descritos, que podem causar instabilidade nos modelos, é compor a amostra de forma que os clientes possam ser selecionados

em vários pontos ao longo do tempo, comumente chamado de *safras* de clientes. Por exemplo, no contexto de *Credit Scoring* a escolha de 12 *safras* ao longo de um ano minimiza consideravelmente a instabilidade do modelo provocada pelos fatores descritos. A Figura 1.2 mostra um delineamento com 12 *safras* para um horizonte de previsão também de 12 meses.

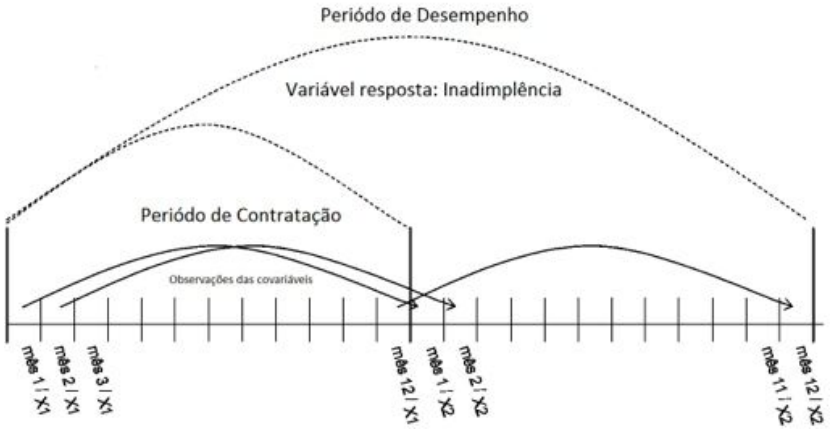

Figura 1.2: Delineamento amostral com horizonte de previsão 12 meses e 12 *safras* de clientes.

Por fim, podemos salientar que a definição do delineamento amostral está intimamente relacionado também com o volume de dados históricos e a estrutura de armazenamento dessas informações encontradas nas empresas e instituições financeiras, as quais podem permitir ou não que a modelagem do evento de interesse se aproxime mais ou menos da realidade observada.

1.2.1 Descrição de um problema - *Credit Scoring*

Em problemas de *Credit Scoring*, as informações disponíveis para correlacionar com a inadimplência do produto de crédito utilizado são as próprias características dos clientes e, algumas vezes, do produto. Dessa forma, um modelo de *Credit Scoring* consiste em avaliar quais fatores estão associados ao risco de crédito dos clientes, assim como a intensidade e a direção de cada um desses fatores, gerando um escore final, os quais

potenciais clientes possam ser ordenados e/ou classificados, segundo uma probabilidade de inadimplência.

Como mencionado, uma situação comum em problemas de *Credit Scoring* é a presença do desbalanceamento entre *bons* e *maus* clientes. Considere, por exemplo, uma base constituída de 600 mil clientes que adquiriram um produto de crédito durante 6 meses, envolvendo, assim, 6 *safras* de clientes, com 594 mil *bons* e 6 mil *maus* pagadores. A descrição das variáveis presentes no conjunto de dados é apresentada na Tabela 1.1. Estas variáveis representam as características cadastrais dos clientes, os valores referentes aos créditos concedidos juntamente com um *flag* descrevendo seus desempenhos de pagamento nos 12 meses seguintes ao da concessão do crédito e informação do instante da ocorrência de algum problema de pagamento do crédito. Essas informações são referentes aos clientes para os quais já foram observados os desempenhos de pagamento do crédito adquirido e servirão para a construção dos modelos preditivos a partir das metodologias regressão logística e/ou análise de sobrevivência. Estes modelos serão aplicadas em futuros potenciais clientes, nos quais serão ordenados segundo uma "probabilidade" de inadimplência e a partir da qual as políticas de crédito das instituições possam ser definidas.

Na construção dos modelos para este problema, de acordo com a Figura 1.3, uma amostra de treinamento é selecionada utilizando a metodologia de *Oversampling*. Isto pode ser feito considerando uma amostra balanceada com 50% de *bons* clientes e 50% de *maus* clientes. A partir dessa amostra buscamos atender as quantidades mínimas sugeridas por Lewis (1994) de 1.500 indivíduos para cada uma das categorias.

1.3 Determinação da Pontuação de Escore

Uma vez determinado o planejamento amostral e obtidas as informações necessárias para o desenvolvimento do modelo, o próximo passo é estabelecer qual técnica estatística ou matemática será utilizada para a determinação dos escores. Porém, antes disso, alguns tratamentos exploratórios devem sempre ser realizados para que uma maior familia-

Tabela 1.1: Variáveis disponíveis no banco de dados.

Variáveis	Descrição
ESTCIVIL	Estado civil: solteiro / casado/ divorciado / viúvo
TP_CLIENTE	Tipo de cliente
SEXO	Sexo do cliente: Masc./ Fem.
SIT_RESID	Residência: própria / alugada
P_CARTAO	Possui Cartão? (Sim / Não)
IDADE	Idade do cliente (em anos)
TEMPORES	Tempo de residência (em anos)
TPEMPREG	Tempo de empregol (em meses)
TEL_COMERC	Declarou telefone comercial?
OP_CORRESP	Correspondência: Residencial / Comercial
COMP_RENDA	Uso da renda: $< 10\%$ / 10%-20% / $> 20\%$;
LIM_CRED	Valor do Crédito Concedido
CEP_COM	CEP Comercial (2 posições)
CEP_RES	CEP Residencial (2 posições)
G_PROF	Grupo de profissão
REGIAO	Região do Cliente
STATUS	Flag: *Bom* ou *Mau*
TEMPO	Tempo até observar o evento inadimplência

rização com os dados possa ser obtida. Isto permite uma melhor definição da técnica que será utilizada e, consequentemente, um aprimoramento do desenvolvimento do modelo. Essa análise inicial tem alguns objetivos, dentre os quais, destacam-se:

- identificação de eventuais inconsistências e presença de *outliers*;
- comparação dos comportamentos das covariáveis, no caso de um *Credit Scoring,* entre a amostra de *bons* e *maus* pagadores, identificando, assim, potenciais variáveis correlacionadas com o evento modelado;
- definição de possíveis transformações de variáveis e a criação de novas a serem utilizadas nos modelos.

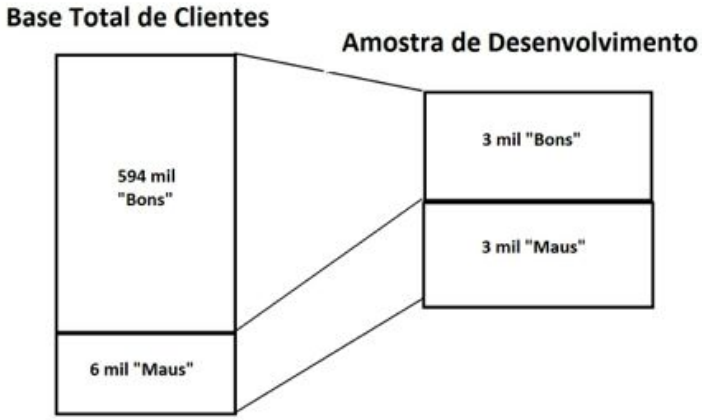

Figura 1.3: Amostra de Desenvolvimento Balanceada - 50% - *bons* x 50% *maus*.

1.3.1 Transformação e seleção de variáveis

Uma prática muito comum, quando se desenvolve modelos de *Credit Scoring*, é tratar as variáveis como categóricas, independente da natureza contínua ou discreta, buscando, sempre que possível, a simplicidade na interpretação dos resultados obtidos. Thomas *et al.* (2002) sugere que essa categorização ou reagrupamento deve ser feito tanto para variáveis originalmente contínuas como para as categóricas. Para as variáveis de origem categórica, a idéia é que se construa categorias com números suficientes de indivíduos para que se faça uma análise robusta, principalmente, quando o número de categorias é originalmente elevado e, em algumas, a frequência é bastante pequena. As variáveis contínuas, uma vez transformadas em categorias, ganham com relação a interpretabilidade dos parâmetros. Gruenstein (1998) e Thomas *et al.* (2002) relatam que esse tipo de transformação nas variáveis contínuas pode trazer ganhos também no poder preditivo do modelo, principalmente quando a covariável em questão se relaciona de forma não-linear com o evento de interesse, como por exemplo, no caso de um *Credit Scoring*.

Uma forma bastante utilizada para a transformação de variáveis contínuas em categóricas, ou a recategorização de uma variável discreta,

é através da técnica *CHAID (Chi-Squared Automatic Interaction Detector)*, a qual divide a amostra em grupos menores, a partir da associação de uma ou mais covariáveis com a variável resposta. A criação de categorias para as covariáveis de natureza contínua ou o reagrupamento das discretas é baseada no teste de associação Qui-Quadrado, buscando a melhor categorização da amostra com relação a cada uma dessas covariáveis ou conjunto delas. Estas "novas" covariáveis podem, então, ser utilizadas na construção dos modelos, sendo ou não selecionadas, por algum método de seleção de variáveis, para compor o modelo final. Um método de seleção de variáveis muitas vezes utilizado é o *stepwise*. Este método permite determinar um conjunto de variáveis estatisticamente significantes para a ocorrência de problemas de crédito dos clientes, através de entradas e saídas das variáveis potenciais utilizando o teste da razão de verossimilhança. Os níveis de significância de entrada e saída das variáveis utilizados pelo método *stepwise* podem ser valores inferiores a 5%, a fim de que a entrada e a permanência de variáveis "sem efeito prático" sejam minimizadas. Outro aspecto a ser considerado na seleção de variáveis, além do critério estatístico, é que a experiência de especialistas da área de crédito juntamente com o bom senso na interpretação dos parâmetros sejam, sempre que possível, utilizados.

Na construção de um modelo de *Credit Scoring* é fundamental que este seja simples com relação à clareza de sua interpretação e que ainda mantenha um bom ajuste. Esse fato pode ser um ponto chave para que ocorra um melhor entendimento, não apenas da área de desenvolvimento dos modelos como também das demais áreas das empresas, resultando, assim, no sucesso da utilização dessa ferramenta.

1.3.2 Regressão logística

Um modelo de regressão logística, com variável resposta, Y, dicotômica, pode ser utilizado para descrever a relação entre a ocorrência ou não de um evento de interesse e um conjunto de covariáveis. No contexto de *Credit Scoring*, o vetor de observações do cliente envolve seu desempenho creditício durante um determinado período de tempo, normalmente de 12 meses, um conjunto de características observadas no

momento da solicitação do crédito e, às vezes, informações à respeito do próprio produto de crédito a ser utilizado, como por exemplo, número de parcelas, finalidade, valor do crédito entre outros.

Aplicando a metodologia apresentada na amostra de treinamento e adotando um horizonte de previsão de 12 meses, considere como variável resposta a ocorrência de falta de pagamento, *maus* clientes, $y = 1$, dentro desse período, não importando o momento exato da ocorrência da inadimplência. Para um cliente que apresentou algum problema de pagamento do crédito no início desses 12 meses de desempenho, digamos no 3^o mês, e um outro para o qual foi observado no final desse período, no 10^o ou 12^o, por exemplo, ambos são considerados da mesma forma como *maus* pagadores, não importando o tempo decorrido para o acontecimento do evento. Por outro lado, os clientes para os quais não foi observada a inadimplência, durante os 12 meses do período de desempenho do crédito, são considerados como *bons* pagadores para a construção do modelo, mesmo aqueles que no 13^o mês vierem a apresentar a falta de pagamento.

É importante ressaltar que adotamos neste livro como evento de interesse o cliente ser *mau* pagador. O mercado financeiro, geralmente, trata como evento de interesse o cliente ser *bom* pagador.

O modelo ajustado, a partir da amostra de treinamento, utilizando a regressão logística, fornece escores tal que, quanto maior o valor obtido para os clientes, pior o desempenho de crédito esperado para eles, uma vez que o *mau* pagador foi considerado como o evento de interesse. Como mencionado, é comum no mercado definir como evento de interesse o *bom* pagador, de forma que, quanto maior o escore, melhor é o cliente.

O modelo de regressão logística é determinado pela relação

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

em que p_i denota a probabilidade de um cliente com o perfil definido pelas p covariadas, $x_1, x_2, \ldots, x_p$, ser um *mau* pagador. Estas covariáveis são obtidas através de transformações, como descritas na seção anterior, sendo portanto consideradas e tratadas como *dummies*. Os valores utilizados como escores finais dos clientes são obtidos, geralmente, mul-

tiplicando por 1.000 os valores estimados das probabilidades de sucesso, $\hat{p}_i$.

O modelo final obtido através da regressão logística para a amostra balanceada encontra-se na Tabela 1.2. No Capítulo 2 apresentamos uma nova análise de dados em que o modelo de regressão logística usual, sem considerar amostras balanceadas, é comparado ao modelo de regressão logística com seleção de amostras *state-dependent*.

Tabela 1.2 - Regressão logística - amostra de treinamento.

Variáveis	Descrição das Variáveis	Estimativa	Erro-Padrão	χ^2	p-valor	Odds-Ratio	L.I. (95%)	L.S. (95%)
Intercepto	-	0,3468	0,0856	16,41	<0,0001			
P_CARTÃO	Posse de Cartão	-0,9980	0,0575	301,19	<0,0001	0,369	0,329	0,413
VIUVO	Est.Civil: Viuvo	0,3112	0,1086	8,21	0,0042	1,365	1,103	1,689
CLI_ANT	Cliente Antigo	-0,4224	0,0608	48,21	<0,0001	0,655	0,582	0,738
IDADE_23	Idade ≤ 23 anos	0,6947	0,1004	47,85	<0,0001	2,003	1,645	2,439
IDADE23_32	23 < Idade ≤ 32	0,4573	0,0769	35,34	<0,0001	1,580	1,359	1,837
IDADE47_53	47 < Idade ≤ 53	-0,2941	0,0966	9,26	0,0023	0,745	0,617	0,901
IDADE53_	Idade > 53 anos	-0,6882	0,0846	66,14	<0,0001	0,502	0,426	0,593
TEMP_2	T.Emprego ≤ 2 anos	0,5710	0,0843	45,89	<0,0001	1,770	1,500	2,088
TEMP2_4	2 <T.Emprego≤ 4	0,4154	0,0802	26,85	<0,0001	1,515	1,295	1,773
TEMP4_8	4 <T.Emprego≤ 8	0,2378	0,0781	9,27	0,0023	1,268	1,008	1,478
TEL_COMERC	Declarou Tel.Com.	-0,2575	0,0567	20,60	<0,0001	0,773	0,692	0,864
LIM_410	Valor Concedido ≤ R$410,0(	0,2027	0,0641	9,98	0,0016	1,225	1,080	1,389
G_CEP_RES1	Grupo1_CEP Residencial	0,4846	0,1127	18,48	<0,0001	1,623	1,302	2,025
G_PROF1	Grupo1_Profissões	0,3901	0,1242	9,87	0,0017	1,477	1,158	1,884
G_CEP_COM1	Grupo1 CEP Comercial	0,3166	0,1102	8,26	0,0040	1,373	1,106	1,703

O *odds ratio*, no contexto de *Credit Scoring*, é uma métrica que representa o quão mais provável é de se observar a inadimplência, para um indivíduo em uma categoria específica da covariável em relação a categoria de referência, analisando os resultados do modelo obtido para a amostra de treinamento, podemos observar:

- P_CARTAO: o fato do cliente já possuir um outro produto de crédito reduz sensivelmente a chance de apresentar algum problema de crédito com a instituição financeira. O valor do *odds ratio* de 0,369 indica que a chance de se observar algum problema para os clientes que possuem um outro produto de crédito é 36,9% da chance de clientes que não possuem;

- ESTADO CIVIL=viúvo: essa categoria contribui para o aumento da chance de se observar algum problema de inadimplência de

crédito. O valor 1,36 indica que a chance de ocorrer problema aumenta em 36% nesta categoria em relação às demais;

- CLI_ANT: o fato do cliente já possuir um relacionamento anterior com a instituição faz com que chance de ocorrer problema seja reduzida. O valor do *odds ratio* de 0,655 indica que a chance de se observar algum problema para um cliente que já possui um relacionamento anterior é 65,5% da chance dos que são de primeiro relacionamento;

- IDADE: para essa variável, fica evidenciado que quanto menor a idade dos clientes maior a chance de inadimplência;

- TEMPO DE EMPREGO: pode-se notar que quanto menor o tempo que o cliente tem no emprego atual maior a chance de ocorrer problema de inadimplência;

- TELEFONE COMERCIAL: a declaração do telefone comercial pelos clientes indica uma chance menor de ocorrer problema de inadimplência;

- LIM_CRED: essa covariável mostra que quanto menor o valor concedido maior a chance de inadimplência, sendo que os clientes com valores abaixo de R$410,00 apresentam cerca de 22,5% a mais de chance de ocorrer problemas do que aqueles com valores acima desse valor;

- CEP RESIDENCIAL, COMERCIAL e PROFISSÃO: os CEP´s indicaram algumas regiões de maior chance de problema, o mesmo ocorrendo para as profissões.

1.4 Validação e Comparação dos Modelos

Com o modelo de *Credit Scoring* construído, surge a seguinte questão: *"Qual a qualidade deste modelo?"*. A resposta para essa pergunta está relacionada com o quanto o escore produzido pelo modelo consegue distinguir os eventos *bons* e *maus* pagadores, uma vez que desejamos identificar previamente esses grupos e tratá-los de forma distinta através de diferentes políticas de crédito.

Uma das idéias envolvidas em medir o desempenho dos modelos está em saber o quão bem estes classificam os clientes. A lógica e a prática sugerem que a avaliação do modelo na própria amostra, usada para o seu desenvolvimento, indica resultados melhores do que se testado

em uma outra amostra, uma vez que o modelo incorpora peculiaridades inerentes da amostra utilizada para sua construção. Por isso, sugerimos, quando o tamanho da amostra permitir e sempre que possível, que o desempenho do modelo seja verificado em uma amostra distinta de seu desenvolvimento.

No contexto de *Credit Scoring*, muitas vezes o tamanho da amostra, na ordem de milhares de registros, permite que uma nova amostra seja obtida para a validação dos modelos. Um aspecto importante na validação dos modelos é o temporal, em que a situação ideal para se testar um modelo é a obtenção de amostras mais recentes. Isto permite que uma medida de desempenho mais próxima da real e atual utilização do modelo possa ser alcançada.

Em Estatística existem alguns métodos padrões para descrever o quanto duas populações são diferentes com relação à alguma característica medida e observada. Esses métodos são utilizados no contexto de *Credit Scoring* com o objetivo de descrever o quanto os grupos de *bons* e *maus* pagadores são diferentes com relação aos escores produzidos por um modelo construído e que necessita ser avaliado. Dessa forma, esses métodos medem o quão bem os escores separam os dois grupos e uma medida de separação muito utilizada para avaliar um modelo de Credit Scoring é a estatística de Kolmogorov-Smirnov (KS). Os modelos podem também ser avaliados e comparados através da curva ROC *(Receiver Operating Characteristic)*, a qual permite comparar o desempenho de modelos através da escolha de critérios de classificação dos clientes em *bons* e *maus* pagadores, de acordo com a escolha de diferentes pontos de corte ao longo das amplitudes dos escores observadas para os modelos obtidos. Porém, muitas vezes o interesse está em avaliar o desempenho dos modelos em um único ponto de corte escolhido, e assim medidas da capacidade preditiva dos mesmos podem ser também consideradas.

1.4.1 A estatística de Kolmogorov-Smirnov (KS)

Essa estatística tem origem no teste de hipótese não-paramétrico de Kolmogorov-Smirnov em que se deseja, a partir de duas amostras retiradas de populações possivelmente 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 *bons* e *maus* pagadores. Sendo $F_B(e) = \sum_{x \leq e} F_B(x)$ e $F_M(e) = \sum_{x \leq e} F_M(x)$ a função de distribuição empírica dos *bons* e *maus* pagadores, respectivamente, a estatística de Kolmogorov-Smirnov é dada por

$$KS = máx \mid F_B(e) - F_M(e) \mid,$$

em que $F_B(e)$ e $F_M(e)$ correspondem às proporções de clientes *bons* e *maus* com escore menor ou igual a e. A estatística KS é obtida através da distância máxima entre essas duas proporções acumuladas ao longo dos escores obtidos pelos modelos, representada na Figura 1.4.

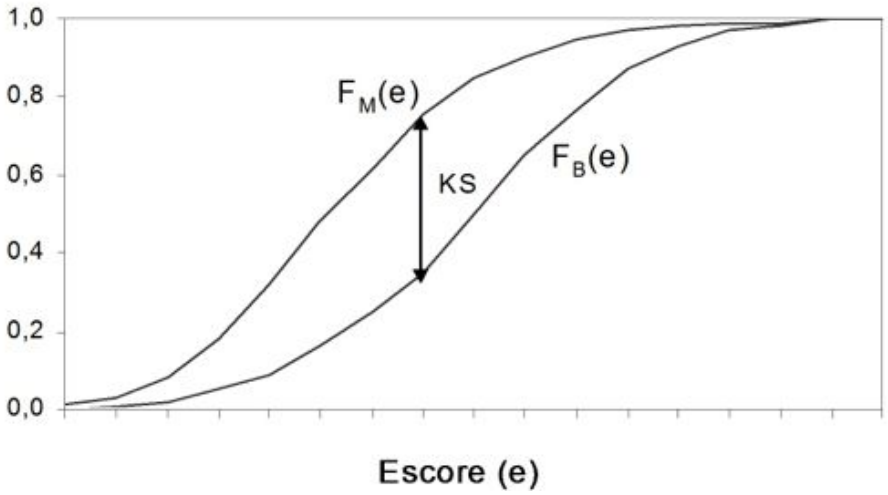

Figura 1.4: Funções distribuições empíricas para os *bons* e *maus* clientes e a estatística KS.

O valor dessa estatística pode variar de 0% a 100%, sendo que o valor máximo indica uma separação total dos escores dos *bons* e *maus* clientes e o valor mínimo sugere uma sobreposição total das distribuições dos escores dos dois grupos. Na prática, obviamente, os modelos fornecem valores intermediários entre esses dois extremos. A representação da interpretação dessa estatística pode ser vista na Figura 1.5.

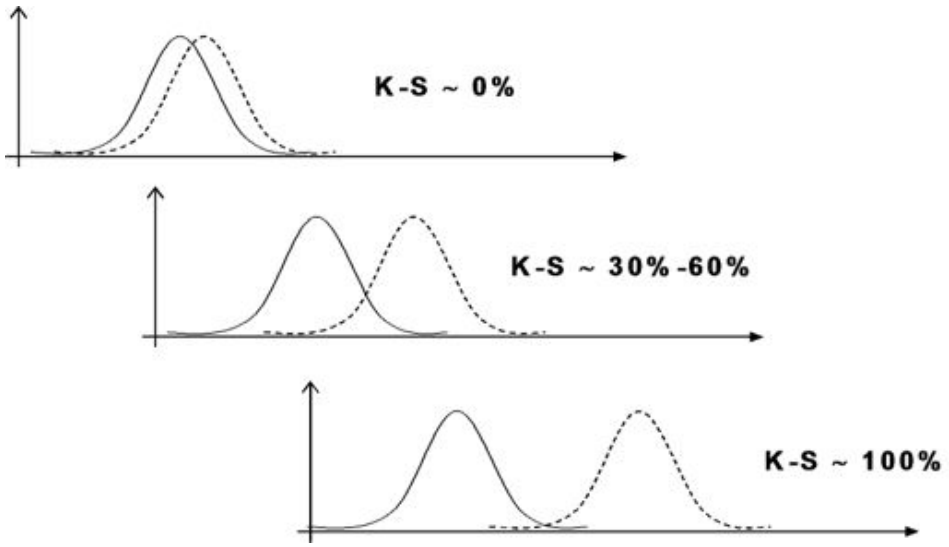

Figura 1.5: Interpretação da estatística KS.

O valor médio da estatística KS para 30 amostras testes com aproximadamente 200 mil clientes retirados aleatoriamente da *base total de clientes* foi 32,26% para a regressão logística.

No mercado, o KS também é utilizado para verificar se o modelo, desenvolvido com um público do passado, pode continuar a ser aplicado para os novos entrantes. Dois diferentes KS são calculados. O KS1 analisa se o perfil dos novos clientes (ou o perfil dos clientes da base de teste) é semelhante ao perfil dos clientes da base de desenvolvimento do modelo. Esse índice é usado para comparar a distribuição acumulada dos escores dos clientes utilizados para o desenvolvimento do modelo com a distribuição acumulada dos escores dos novos entrantes (ou dos clientes da base de teste). Quanto menor o valor do KS1 mais semelhante é o perfil do público do desenvolvimento com o perfil dos novos clientes. O KS2 avalia a performance do modelo. Ou seja, mede, para uma dada *safra*, a máxima distância entre a distribuição de frequência acumulada dos *bons* clientes em relação à distribuição de frequência acumulada dos *maus* clientes.

A interpretação do índice para modelos de *Credit Scoring* segue, em algumas instituições, a seguinte regra:

- KS $< 10\%$: indica que não há discriminação entre os perfis de *bons* e *maus* clientes;
- $10\% <$ KS $< 20\%$: indica que a discriminação é baixa;
- KS $> 20\%$: indica que o modelo discrimina o perfil de *bons* e *maus*.

1.4.2 Curva ROC

Os escores obtidos para os modelos de *Credit Scoring* devem, normalmente, ser correlacionados com a ocorrência de algum evento de interesse, como por exemplo, a inadimplência, permitindo assim, fazer previsões a respeito da ocorrência desse evento para que políticas de crédito diferenciadas possam ser adotadas pelo nível de escore obtido para os indivíduos.

Uma forma de se fazer previsões é estabelecer um ponto de corte no escore produzido pelos modelos. Clientes com valores iguais ou maiores a esse ponto são classificados, por exemplo, como *bons* e abaixo desse valor como *maus* pagadores. Para estabelecer e visualizar o cálculo dessas medidas podemos utilizar uma tabela 2x2 denominada *matriz de confusão*, representada na Figura 1.6

	Situação Real		
Previsão do Modelo	**Bom**	**Mau**	**Total**
Bom	b_B	b_M	b
Mau	m_B	m_M	m
Total	B	M	n

Figura 1.6: Matriz de Confusão.

em que:

n : número total de clientes na amostra;

b_B : número de *bons* clientes que foram classificados como *Bons* (acerto);

m_M : número de *maus* clientes que foram classificados como *Maus* (acerto);

m_B : número de *bons* clientes que foram classificados como *Maus* (erro);

b_M : número de *maus* clientes que foram classificados como *Bons* (erro);

B : número total de *bons* clientes na amostra;

M : número total de *maus* clientes na amostra;

b : número total de clientes classificados como *bons* na amostra;

m : número total de clientes classificados como *maus* na amostra;

Na área médica, duas medidas muito comuns e bastante utilizadas são a *sensibilidade* e a *especificidade*. Essas medidas, adaptadas ao contexto de *Credit Scoring*, considerando o *mau* cliente como a categoria de interesse, são definidas da seguinte forma:

Sensibilidade: probabilidade de um indivíduo ser classificado como *mau* pagador, dado que realmente é *mau*;

Especificidade: probabilidade de um indivíduo ser classificado como *bom* pagador, dado que realmente é *bom*;

Utilizando as frequências mostradas na matriz de confusão, temos que a *Sensibilidade* é dada por $\frac{m_M}{M}$ e a *Especificidade* por $\frac{b_B}{B}$.

A curva ROC (Zweig & Campbell, 1993) é construída variando os pontos de corte, *cut-off*, ao longo da amplitude dos escores fornecidos pelos modelos, a fim de se obter as diferentes classificações dos indivíduos e obtendo, consequentemente, os respectivos valores para as medidas de *Sensibilidade* e *Especificidade* para cada ponto de corte estabelecido. Assim, a curva ROC, ilustrada na Figura 1.7, é obtida tendo no seu eixo horizontal os valores de (1-*Especificidade*), ou seja, a proporção de *bons* clientes que são classificados como *maus* clientes pelo modelo, e no eixo vertical a *Sensibilidade*, que é a proporção de *maus* clientes que são classificados realmente como *maus*. Uma curva ROC obtida ao longo da diagonal principal corresponde a uma classificação obtida sem a utilização de qualquer ferramenta preditiva, ou seja, sem a presença de modelos. Consequentemente, a curva ROC deve ser interpretada de forma que quanto mais a curva estiver distante da diagonal principal, melhor o desempenho do modelo em questão. Esse fato sugere que quanto maior

for a área entre a curva ROC produzida e a diagonal principal, melhor o desempenho global do modelo.

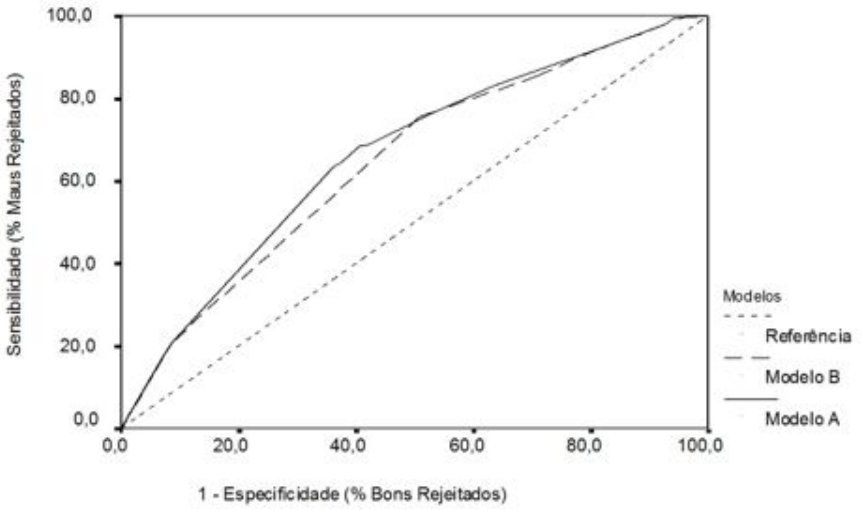

Figura 1.7: Exemplos de curva ROC.

Os pontos de corte ao longo dos escores fornecidos pelos modelos que apresentam bom poder discriminatório concentram-se no canto superior esquerdo da curva ROC. A curva ROC apresenta sempre um contrabalanço entre a *Sensibildade* e a *Especificidade* ao se variar os pontos de corte ao longo dos escores e pode ser usada para auxiliar na decisão de determinar o melhor ponto de corte. Em geral, o melhor *cut-off* ao longo dos escores produz valores para as medidas de *Sensibildade* e *Especificidade* que se localiza no "ombro" da curva, ou próximo desse, ou seja, no ponto mais a esquerda e superior possível, o qual é obtido considerando como ponto de corte o escore que fornece a separação máxima no teste KS. Vale destacar que em problemas de *Credit Scoring*, normalmente, critérios financeiros são utilizados na determinação desse melhor ponto, sendo que valores como o quanto se perde em média ao aprovar um cliente que traz problemas de crédito e também o quanto se deixa de ganhar ao não aprovar o crédito para um cliente que não traria problemas para a instituição podem e devem ser considerados.

A partir da curva ROC temos a idéia do desempenho do modelo

ao longo de toda amplitude dos escores produzidos pelos modelos.

1.4.3 Capacidade de acerto dos modelos

Em um modelo com variável resposta binária, como ocorre normalmente no caso de um *Credit Scoring*, temos o interesse em classificar os indivíduos em uma das duas categorias, *bons* ou *maus* clientes, e obter um bom grau de acerto nestas classificações. Como, geralmente, nas amostras testes, em que os modelos são avaliados, se conhece a resposta dos clientes em relação a sua condição de crédito, e estabelecendo critérios para classificar estes clientes em *bons* e *maus,* torna-se possível comparar a classificação obtida com a verdadeira condição creditícia dos clientes.

A forma utilizada para estabelecer a *matriz de confusão*, Figura 1.6, é determinar um ponto de corte (*cutoff*) no escore final dos modelos tal que, indivíduos com pontuação acima desse *cutoff* são classificados como *bons*, por exemplo, e abaixo desse valor como *maus* clientes e comparando essa classificação com a situação real de cada indivíduo. Essa matriz descreve, portanto, uma tabulação cruzada entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada indivíduo, em que a diagonal principal representa as classificações corretas e valores fora dessa diagonal correspondem à erros de classificação.

A partir da *matriz de confusão* determinada por um ponto de corte específico e representada pela Figura 1.6, algumas medidas de capacidade de acerto dos modelos são definidas a seguir:

- Capacidade de Acerto Total (CAT)= $\frac{b_B+m_M}{n}$
- Capacidade de Acerto dos *Maus* Clientes (CAM)= $\frac{m_M}{M}$ (Especificidade)
- Capacidade de Acerto dos *Bons* Clientes (CAB)= $\frac{b_B}{B}$ (Sensibilidade)
- Valor Preditivo Positivo (VPP)= $\frac{b_B}{b_B+b_M}$
- Valor Preditivo Negativo (VPN) = $\frac{m_B}{m_B+m_M}$

- Prevalência (PVL) = $\frac{b_B+m_B}{n}$
- Correlação de Mathews (MCC) = $\frac{b_B m_M - b_M m_B}{\sqrt{(b_B+b_M)(b_B+m_B)(m_M+b_M)(m_M+m_B)}}$

A Prevalência, proporção de observações propensas a característica de interesse ou a probabilidade de uma observação apresentar a característica de interesse antes do modelo ser ajustado, é um medida de extrema importância, principalmente quando tratamos de eventos raros.

A Capacidade de Acerto Total é também conhecida como Acurácia ou Proporção de Acertos de um Modelo de Classificação. Esta medida também pode ser vista como uma média ponderada da sensibilidade e da especificidade em relação ao número de observações que apresentam ou não a característica de interesse de uma determinada população. É importante ressaltar que a acurácia não é uma medida que deve ser analisada isoladamente na escolha de um modelo, pois é influenciada pela sensibilidade, especificidade e prevalência. Além disso, dois modelos com sensibilidade e especificidade muito diferentes podem produzir valores semelhantes de acurácia, se forem aplicados a populações com prevalências muito diferentes.

Para ilustrar o efeito da prevalência na acurácia de um modelo, podemos supor uma população que apresente 5% de seus integrantes com a característica de interesse. Se um modelo classificar todos os indivíduos como não portadores da característica, temos um percentual de acerto de 95%, ou seja, a acurácia é alta e o modelo é pouco informativo.

O Valor Preditivo Positivo (VPP) de um modelo é a proporção de observações representando o evento de interesse dentre os indivíduos que o modelo identificou como evento. Já o Valor Preditivo Negativo (VPN) é a proporção de indivíduos que representam não evento dentre os identificados como não evento pelo modelo. Estas medidas devem ser interpretadas com cautela, pois sofrem a influência da prevalência populacional.

Caso as estimativas da sensibilidade e da especificidade sejam confiáveis, o valor preditivo positivo (VPP) pode ser estimado via Teorema de Bayes, utilizando uma estimativa da prevalência (Linnet, 1998)

$$VPP = \frac{\text{SENS} \times \text{PVL}}{\text{SENS} \times \text{PVL} + (1 - \text{SPEC}) \times (1 - \text{PVL})},$$

com SENS usado para Sensibilidade e SPEC para Especificidade. Da mesma forma, o valor preditivo negativo (VPN) pode ser estimado por

$$VPN = \frac{\text{SPEC} \times (1 - \text{PVL})}{\text{SPEC} \times (1 - \text{PVL}) + \text{SENS} \times \text{PVL}}.$$

O MCC, proposto por Matthews (1975), é uma medida de desempenho que pode ser utilizada no caso de prevalências extremas. É uma adaptação do Coeficiente de Correlação de Pearson e mede o quanto as variáveis que indicam a classificação original da resposta de interesse e a que corresponde a classificação do modelo obtida por meio do ponto de corte adotado, ambas variáveis assumindo valores 0 e 1, tendem a apresentar o mesmo sinal de magnitude após serem padronizadas (Baldi *et al.*, 2000).

O MCC retorna um valor entre -1 e +1. O valor 1 representa uma previsão perfeita, um acordo total, o valor 0 representa uma previsão completamente aleatória e -1 uma previsão inversa, ou seja, total desacordo. Observe que o MCC utiliza as 4 medidas apresentadas na matriz de confusão (b_B, b_M, m_B, m_M).

O Custo Relativo, baseado em uma medida apresentada em Bensic *et al.* (2005), é definido por $CR = \alpha C_1 P_1 + (1 - \alpha) C_2 P_2$, em que α representa a probabilidade de um proponente ser *mau* pagador, C_1 é o custo de aceitar um *mau* pagador, C_2 é o custo de rejeitar um *bom* pagador, P_1 é a probabilidade de ocorrer um falso negativo e P_2 é a probabilidade de ocorrer um falso positivo.

Como na prática não é fácil obter as estimativas de C_1 e C_2, o custo é calculado considerando diversas proporções entre C_1 e C_2, com a restrição $C_1 > C_2$, ou seja, a perda em aceitar um *mau* pagador é maior do que o lucro perdido ao rejeitar um *bom* pagador. Bensic *et al.* (2005) considera α como a prevalência amostral, isto é, supõe que a prevalência de *maus* pagadores nos portfólios representa a prevalência real da população de interesse.

Capítulo 2

Regressão Logística

Os modelos de regressão são utilizados para estudar e estabelecer uma relação entre uma variável de interesse, denominada variável resposta, e um conjunto de fatores ou atributos referentes a cada cliente, geralmente encontrados na proposta de crédito, denominados covariáveis.

No contexto de *Credit Scoring*, como a variável de interesse é binária, a regressão logística é um dos métodos estatísticos utilizado com bastante frequência. Para uma variável resposta dicotômica, o interesse é modelar a proporção de resposta de uma das duas categorias, em função das covariáveis. É comum adotarmos o valor 1 para a resposta de maior interesse, denominada "sucesso", o qual pode ser utilizado no caso de um proponente ao crédito ser um *bom* ou um *mau* pagador.

Normalmente, quando construímos um modelo de *Credit Scoring*, a amostra de desenvolvimento é formada pela seleção dos clientes contratados durante um período de tempo específico, sendo observado o desempenho de pagamento desses clientes ao longo de um período de tempo posterior e pré-determinado, correspondente ao horizonte de previsão. Esse tempo é escolhido arbitrariamente entre 12 e 18 meses, sendo na prática 12 meses o intervalo mais utilizado, como já mencionado no Capítulo 1, em que a variável resposta de interesse é classificada, por exemplo, em *bons* ($y = 0$) e *maus* ($y = 1$) pagadores, de acordo com a ocorrência ou não de problemas de crédito nesse intervalo. É importante chamar a atenção que ambos os períodos — de seleção da amostra e de desempenho de pagamento — estão no passado, portanto a ocorrência

ou não do evento modelado já deve ter sido observada.

Sejam $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$ o vetor de valores de atributos que caracterizam um cliente e $\pi(\boldsymbol{x})$ a proporção de *maus* pagadores em função do perfil dos clientes, definido e caracterizado por $\boldsymbol{x}$. Neste caso, o modelo logístico é adequado para definir uma relação entre a probabilidade de um cliente ser *mau* pagador e um conjunto de fatores ou atributos que o caracterizam. Esta relação é definida pela função ou transformação *logito* dada pela expressão

$$\log\left\{\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right\} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

em que $\pi(\boldsymbol{x})$ é definido como

$$\pi(\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)},$$

e pode ser interpretado como a probabilidade de um proponente ao crédito ser um *mau* pagador dado as características que possui, representadas por $\boldsymbol{x}$. No caso da atribuição da categoria *bom* pagador, as interpretações são análogas.

2.1 Estimação dos Coeficientes

Dada uma amostra de n clientes $(y_i, \boldsymbol{x}_i)$, sendo y_i a variável resposta — *bons* e *maus* pagadores — e $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$, em que $x_{i1}, x_{i2}, \ldots, x_{ik}$ são os valores dos k atributos observados do i-ésimo cliente, $i = 1, \ldots, n$, o ajuste do modelo logístico consiste em estimar os parâmetros β_j, $j = 1, 2, \ldots, k$, os quais definem $\pi(\boldsymbol{x})$.

Os parâmetros são geralmente estimados pelo método de máxima verossimilhança (Hosmer & Lemeshow, 2000). Por este método, os coeficientes são estimados de maneira a maximizar a probabilidade de se obter o conjunto de dados observados a partir do modelo proposto. Para o método ser aplicado, primeiramente construímos a função de verossimilhança que expressa a probabilidade dos dados observados, como função

dos parâmetros $\beta_1, \beta_2, \ldots, \beta_k$. A maximização desta função fornece os estimadores de máxima verossimilhança para os parâmetros.

No modelo de regressão logística, uma forma conveniente para expressar a contribuição de um cliente $(y_i, \boldsymbol{x}_i)$ para a função de verossimilhança é dada por

$$\zeta(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)^{y_i} \left[1 - \pi(\boldsymbol{x}_i)\right]^{1-y_i}. \tag{2.1}$$

Uma vez que as observações, ou seja, os clientes são considerados independentes, a função de verossimilhança pode ser obtida como produto dos termos em (2.1)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \zeta(\boldsymbol{x}_i). \tag{2.2}$$

A partir do princípio da máxima verossimilhança, os valores das estimativas para $\boldsymbol{\beta}$ são aqueles que maximizam a equação (2.2). No entanto, pela facilidade matemática, trabalhamos com o log dessa expressão, que é definida como

$$l(\boldsymbol{\beta}) = \log\left[L(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \left\{ y_i \log\left[\pi(\boldsymbol{x}_i)\right] + (1 - y_i) \log\left[1 - \pi(\boldsymbol{x}_i)\right] \right\}. \tag{2.3}$$

Para obtermos os valores de $\boldsymbol{\beta}$ que maximizam $l(\boldsymbol{\beta})$, calculamos a derivada em relação a cada um dos parâmetros $\beta_1, \ldots, \beta_k$, sendo obtidas as seguintes equações

$$\begin{aligned} \sum_{i=1}^{n} \left[y_i - \pi(\boldsymbol{x}_i)\right] &= 0, \\ \sum_{i=1}^{n} x_{ij} \left[y_i - \pi(\boldsymbol{x}_i)\right] &= 0, \qquad \text{para } j = 1, \ldots, k, \end{aligned}$$

as quais, uma vez solucionadas via métodos numéricos, como por exemplo *Newton-Raphson*, fornecem as estimativas de máxima verossimilhança. Esse método numérico é o mais comum de ser encontrado nos pacotes estatísticos.

A partir do modelo ajustado podemos predizer a probabilidade de novos candidatos a crédito serem *maus* pagadores. Esses valores preditos são utilizados, normalmente, para a aprovação ou não de uma linha de crédito, ou na definição de encargos financeiros de forma diferenciada.

Além da utilização das estimativas dos parâmetros na predição do potencial de risco de novos candidatos a crédito, os estimadores dos parâmetros fornecem também a informação, através da sua distribuição de probabilidade e do nível de significância, de quais covariáveis estão mais associadas com o evento que está sendo modelado, ajudando na compreensão e interpretação do mesmo, no caso a inadimplência.

2.2 Intervalos de Confiança e Seleção de Variáveis

Uma vez escolhido o método de estimação dos parâmetros, um próximo passo para a construção do modelo é o de questionar se as covariáveis utilizadas e disponíveis para a modelagem são estatisticamente significantes com o evento modelado, como por exemplo, a condição de *mau pagador* de um cliente.

Uma forma de testar a significância do coeficiente de uma determinada covariável é buscar responder à seguinte pergunta: *O modelo que inclui a covariável de interesse nos fornece mais informação a respeito da variável resposta do que um modelo que não considera essa covariável?* A idéia é que, se os valores preditos fornecidos pelo modelo com a covariável são mais precisos do que os valores preditos obtidos pelo modelo sem a covariável, há evidências de que essa covariável é importante. Da mesma forma que nos modelos lineares, na regressão logística comparamos os valores observados da variável resposta com os valores preditos obtidos pelos modelos com e sem a covariável de interesse. Para entender melhor essa comparação é interessante que, teoricamente, se pense que um valor observado para a variável resposta é também um valor predito resultante de um modelo saturado, ou seja, um modelo teórico que contém tantos parâmetros quanto o número de variáveis.

A comparação de valores observados e preditos é feita a partir

da razão de verossimilhança usando a seguinte expressão

$$D = -2\log\left[\frac{\text{verossimilhança do modelo testado}}{\text{verossimilhança do modelo saturado}}\right]. \qquad (2.4)$$

O valor inserido entre os colchetes na expressão (2.4) é chamado de *razão de verossimilhança*. A estatística D, chamada de *Deviance*, tem um importante papel na verificação do ajuste do modelo. Fazendo uma analogia com os modelos de regressão linear, a *Deviance* tem a mesma função da soma de quadrado de resíduos, e, a partir das equações (2.3) e (2.4) temos que

$$\begin{aligned} D &= -2\left\{\sum_{i=1}^{n}\left[y_i\log\left(\widehat{\pi}_i\right)+(1-y_i)\log\left(1-\widehat{\pi}_i\right)\right]\right. \\ &- \left.\sum_{i=1}^{n}\left[y_i\log(y_i)+(1-y_i)\log(1-y_i)\right]\right\} \\ &= -2\left\{\sum_{i=1}^{n} y_i\left[\log(\widehat{\pi}_i)-\log(y_i)\right]\right. \\ &+ \left.(1-y_i)\left[\log(1-\widehat{\pi}_i)-\log(1-y_i)\right]\right\} \\ &= -2\sum_{i-1}^{n}\left[y_i\log\left(\frac{\widehat{\pi}_i}{y_i}\right)+(1-y_i)\log\left(\frac{1-\widehat{\pi}_i}{1-y_i}\right)\right], \qquad (2.5) \end{aligned}$$

sendo $\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i)$.

A significância de uma covariável pode ser obtida comparando o valor da *Deviance* (D) para os modelos com e sem a covariável de interesse. A mudança ocorrida em D devido à presença da covariável no modelo é obtida da seguinte forma

$$G = D(\text{modelo sem a covariável}) - D(\text{modelo com a covariável}).$$

Uma vez que a verossimilhança do modelo saturado é comum em

ambos valores de D, temos que G pode ser definida como

$$G = -2\log\left[\frac{\text{verossimilhança sem a variável de interesse}}{\text{verossimilhança com a variável de interesse}}\right]. \qquad (2.6)$$

A estatística (2.6), sob a hipótese de que o coeficiente da covariável de interesse que está sendo testada é nulo, tem distribuição χ^2_1. Esse teste, conhecido como teste da *Razão de Verossimilhança*, pode ser conduzido para mais do que uma variável simultaneamente. Uma alternativa ao teste da *Razão de Verossimilhança* é o teste de Wald. Para um único parâmetro, a estatística de Wald é obtida comparando a estimativa de máxima verossimilhança do parâmetro de interesse com o seu respectivo erro-padrão.

Para um modelo com k covariáveis temos, para cada parâmetro, $H_0 : \beta_j = 0$, $j = 0, 1, \ldots, k$, cuja estatística do teste é dada por

$$Z_j = \frac{\widehat{\beta}_j}{\widehat{EP}(\widehat{\beta}_j)},$$

sendo $\widehat{\beta}_j$ a estimativa de máxima verossimilhança de β_j e $\widehat{EP}(\widehat{\beta}_j)$ a estimativa do seu respectivo erro-padrão. Sob a hipótese nula (H_0), Z_j tem aproximadamente uma distribuição normal padrão e Z_j^2 segue aproximadamente uma distribuição χ^2_1.

2.3 Interpretação dos Coeficientes do Modelo

Sabemos que a interpretação de qualquer modelo de regressão exige a possibilidade de extrair informações práticas dos coeficientes estimados. No caso do modelo de regressão logística, é fundamental o conhecimento do impacto causado por cada variável na determinação da probabilidade do evento de interesse.

Uma medida presente na metodologia de regressão logística, e útil na interpretação dos coeficientes do modelo, é o *odds*, que para uma covariável x é definido como $[\frac{\pi(x)}{1-\pi(x)}]$. Aplicando a função log no *odds*

tem-se a transformação *logito*. Para uma variável dicotômica assumindo valores $(x = 1)$ e $(x = 0)$, obtém-se que o *odds* é dado por $[\frac{\pi(1)}{1-\pi(1)}]$ e $[\frac{\pi(0)}{1-\pi(0)}]$, respectivamente. A razão entre os *odds* em $(x = 1)$ e $(x = 0)$ define o *odds ratio*, dado por

$$\Psi = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}.$$

Como $\pi(1) = e^{\beta_0+\beta_1}/1 + e^{\beta_0+\beta_1}$, $\pi(0) = e^{\beta_0}/1 + e^{\beta_0}$, $1 - \pi(1) = 1/1 + e^{\beta_0+\beta_1}$ e $1 - \pi(0) = 1/1 + e^{\beta_0}$, temos que

$$\Psi = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

O *odds ratio* é uma medida de associação largamente utilizada e pode ser interpretado como a propensão que o indivíduo possui de assumir o evento de interesse quando $x = 1$, comparado com $x = 0$. Por exemplo, sejam y a presença de inadimplência e x a variável indicadora que denota se o indivíduo tem telefone $(x = 0)$ ou não tem telefone $(x = 1)$. Se $\hat{\Psi} = 2$ podemos dizer que a inadimplência é duas vezes mais provável nos indivíduos sem telefone.

2.4 Aplicação

Considere o conjunto de dados reais constituído de informações de uma instituição financeira na qual os clientes adquiriram um produto de crédito. Essa instituição tem como objetivo, a partir desse conjunto de dados, medir o risco de inadimplência de potenciais clientes que busquem adquirir o produto. As variáveis disponíveis no banco de dados correspondem às características cadastrais dos clientes (sexo, estado civil, etc.), o valor referente ao crédito concedido, bem como um *flag* descrevendo seu desempenho de pagamento nos 12 meses seguintes ao da concessão do crédito (*maus* pagadores: $flag = 1$, *bons* pagadores: $flag = 0$). Essas informações servirão para a construção do modelo preditivo a partir da metodologia estudada, a regressão logística (Hosmer & Lemeshow, 2000),

o qual poderá ser aplicado em futuros potenciais clientes, permitindo que eles possam ser ordenados segundo uma probabilidade de inadimplência. A a partir desta probabilidade, as políticas de crédito da instituição podem ser definidas.

A base total de dados é de 5909 clientes. Para a construção do modelo preditivo segundo a metodologia estudada, selecionamos, via amostragem aleatória simples sem reposição, uma amostra de desenvolvimento ou de treinamento, correspondente a 70% dessa base de dados; em seguida, ajustamos um modelo de regressão logística (Hosmer & Lemeshow, 2000) nessa amostra; e, por fim, utilizamos o restante 30% dos dados como amostra de teste para verificação da adequabilidade do modelo.

Algumas das covariáveis presentes no banco de dados foram obtidas de acordo com as categorizações sugeridas pela Análise de Agrupamento (*Cluster Analysis*), e selecionadas através do seu valor-p considerando um nível de significância de 5%. Sendo assim, variáveis com valor-p inferior a 0,05 foram mantidas no modelo. A Tabela 2.1 apresenta o modelo final obtido através da regressão logística para a amostra de desenvolvimento. Na base, e na tabela, temos var1 = Tipo de cliente: 1; var4 = Sexo: Feminino; var5_C = Est. civil: Casado; var5_D = Est. civil: Divorciado; var5_S = Est. civil: Solteiro; var11C_1 = T. residência$\leq$8 anos ; var11C_3 = 8<T. residência$\leq$20; var11C_2 = 20<T. residência$\leq$35; var11C_4 = T. residência>49 anos ; var12C_3 = Idade$\leq$22 anos; var12C_1 = 22<Idade$\leq$31; var12C_2 = 31<Idade$\leq$43; var12C_5 = 55<Idade$\leq$67; var12C_6 = 67<Idade$\leq$78; var12C_4 = Idade>78 anos. As categorias não presentes nesta lista são as determinadas como *categorias de referências*.

A partir dos *odds ratio* apresentados na Tabela 2.1, para cada variável presente no modelo final, observamos:

- TIPO DE CLIENTE: o fato do cliente ser do tipo 1 (cliente há mais de um ano) faz com que o risco de crédito aumente quase 3 vezes em relação àqueles que são do tipo 2 (há menos de um ano na base);

- SEXO: o fato do cliente ser do sexo feminino reduz o risco de apre-

Tabela 2.1: Resultados do modelo de regressão logística obtido para a amostra de desenvolvimento (70% da base de dados) extraída de uma carteira de um banco.

Variáveis	Estimativa	Erro Padrão	Valor-p	*Odds ratio*
Intercepto	-1,1818	0,2331	<,0001	
var1	0,5014	0,0403	<,0001	2,726
var4	-0,1784	0,0403	<,0001	0,700
var5_C	-0,4967	0,0802	<,0001	0,450
var5_D	0,4604	0,1551	0,0030	1,171
var5_S	-0,2659	0,0910	0,0035	0,567
var11C_1	0,5439	0,2273	0,0167	1,545
var11C_3	0,1963	0,2284	0,3903	1,091
var11C_2	-0,0068	0,2476	0,9780	0,891
var11C_4	-0,8421	0,8351	0,3133	0,386
var12C_3	1,8436	0,1383	<,0001	8,158
var12C_1	1,3207	0,1172	<,0001	4,836
var12C_2	0,2452	0,1123	0,0290	1,650
var12C_5	-1,2102	0,1576	<,0001	0,385
var12C_6	-1,3101	0,2150	<,0001	0,348
var12C_4	-0,6338	0,4470	0,1562	0,685

sentar algum problema de crédito com a instituição financeira, em que o valor do *odds* de 0,7 na regressão logística indica que a chance de observarmos algum problema para os clientes que são do sexo feminino é aproximadamente 70% do que para os que são do sexo masculino.

- ESTADO CIVIL: a categoria viúvo, deixada como referência, contribui para o aumento do risco de crédito em relação às categorias casado e solteiro, mas não podemos afirmar isso em relação à categoria divorciado, visto que o *odds* não é estatisticamente significativo, visto que o valor 1 está contido no intervalo de 95% de confiança para o *odds* (intervalo não apresentado aqui).

- TEMPO DE RESIDÊNCIA: notamos que quanto menor o tempo

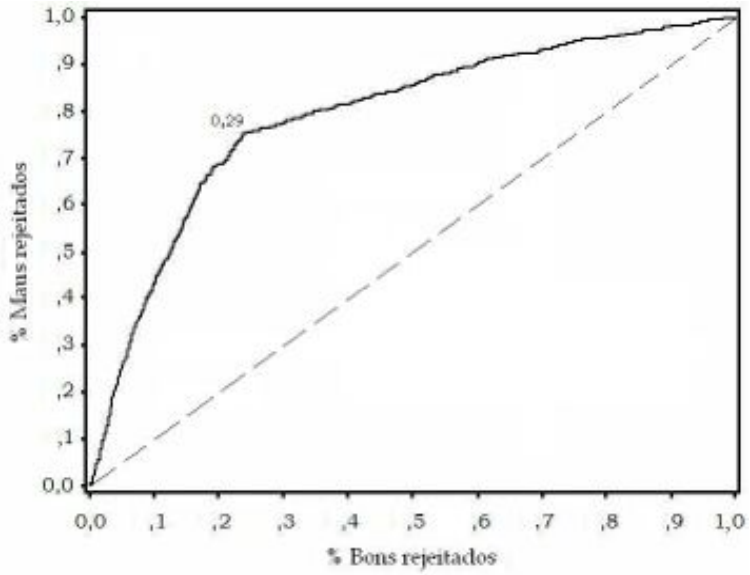

Figura 2.1: Curva ROC construída a partir da amostra de treinamento de uma carteira de banco.

que o cliente tem na atual residência maior o seu risco de crédito, embora nenhum dos *odds* seja estatisticamente significante para essa variável (similar caso anterior).

- IDADE: para essa variável, verificamos que quanto menor a idade dos clientes maior o risco de inadimplência.

Com o auxílio da curva ROC podemos escolher um ponto de corte igual a 0,29. Assim, as medidas relacionadas à capacidade preditiva do modelo são: $SENS = 0,75$, $SPEC = 0,76$, $VPP = 0,58$, $VPN = 0,87$, $CAT = 0,76$ e $MCC = 0,48$, o que é indicativo de uma boa capacidade preditiva. Esta conclusão é corroborada pela curva ROC apresentada na Figura 2.1.

2.5 Amostras *State-Dependent*

Uma estratégia comum utilizada na construção de amostras para o ajuste de modelos de regressão logística, quando os dados são desbalanceados, é selecionar uma amostra contendo todos os eventos presentes

na base de dados original e selecionar, via amostragem aleatória simples sem reposição, um número de não eventos igual ou superior ao número de eventos. No entanto, este número deve sempre ser menor do que a quantidade de observações representando não evento presentes na amostra. Estas amostras, denominadas *state-dependent*, são muito utilizadas, principalmente, no mercado financeiro. No entanto, para validar as inferências realizadas para os parâmetros obtidos por meio destas amostras, algumas adaptações são necessárias. Neste trabalho utilizamos o Método de Correção a Priori, descrito na subseção 2.5.1.

A técnica de regressão logística com seleção de amostras *state-dependent* (Cramer, 2004) realiza uma correção na probabilidade predita ou estimada de um indivíduo ser, por exemplo, um *mau* pagador, segundo o modelo de regressão logística usual (Hosmer & Lemeshow, 2000).

Considere uma amostra de observações com vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$, $i = 1, \ldots, n$ e variável resposta y_i, binária (0,1), em que o evento $y_i = 1$, o i-ésimo cliente é um *mau* pagador, é pouco frequente, enquanto o complementar $y_i = 0$, o i-ésimo cliente é um *bom* pagador, é abundante. O modelo especifica que a probabilidade do i-ésimo cliente ser um *mau* pagador, como uma função de $\boldsymbol{x}_i$, seja dada por

$$P\left(y_i = 1|\boldsymbol{x}_i\right) = \pi\left(\boldsymbol{\beta}, \boldsymbol{x}_i\right) = \pi_i,$$

sendo $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_k\right)^{'}$. Queremos estimar $\boldsymbol{\beta}$ a partir de uma *selected sample*, a qual é obtida descartando parte das observações de 0 (*bons* pagadores), por razões de conveniência. Supondo que a *full sample* inicial seja uma amostra aleatória com fração amostral α e que somente uma fração γ das observações de 0 é retida aleatoriamente, então a probabilidade de que o cliente i seja um *mau* pagador ($y_i = 1$), e esteja incluído na amostra, é dada por

$$\alpha\pi_i,$$

enquanto que, para $y_i = 0$ é dada por

$$\gamma\alpha\left(1 - \pi_i\right).$$

Portanto, pelo teorema de Bayes (Louzada *et al.*, 2012), temos que a probabilidade de que um elemento qualquer da *selected sample* seja um *mau* pagador, é dada por

$$\pi_i^* = \frac{\pi_i}{\pi_i + \gamma\,(1 - \pi_i)}.$$

A log-verossimilhança da amostra observada, em termos de π_i^*, é

$$\begin{aligned} l(\boldsymbol{\beta}, \gamma) &= \log\left[L(\boldsymbol{\beta}, \gamma)\right] \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left[\pi_i^*\left(\boldsymbol{\beta}, \boldsymbol{x}_i, \gamma\right)\right] + (y_i - 1) \log\left[\pi_i^*\left(\boldsymbol{\beta}, \boldsymbol{x}_i, \gamma\right)\right] \right\}. \end{aligned}$$

Se γ é conhecido, os parâmetros de qualquer especificação de π_i podem ser estimados a partir da *selected sample* por métodos padrões de máxima verossimilhança.

Supondo que um modelo de regressão logística usual é utilizado na análise, π_i^* é dado por

$$\pi_i^* = \frac{\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + \gamma} = \frac{\frac{1}{\gamma}\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)}{1 + \frac{1}{\gamma}\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)} = \frac{\exp\left(\boldsymbol{x}_i'\,\boldsymbol{\beta} - \log\gamma\right)}{1 + \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta} - \log\gamma\right)}.$$

Pela expressão acima, observamos que π_i^* obedece o mesmo formato de um modelo de regressão logística e, com exceção do intercepto, os mesmos parâmetros $\boldsymbol{\beta}$ presentes na *full sample* se aplicam aqui. O intercepto da *full sample* pode ser recuperado adicionando $\log\gamma$ ao intercepto, β_0, da *selected sample*. Um estimador consiste e eficiente de β_0 é apresentado na subseção 2.5.1.

2.5.1 Método de correção a priori

A técnica de correção a priori envolve o cálculo dos estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística e a correção destas estimativas, com base na informação a priori da fração de eventos na população τ (prevalência populacional, ou seja, a proporção de eventos na população) e a fração de eventos observados

na amostra $\bar{y}$ (prevalência amostral, ou seja, a proporção de eventos na amostra).

No modelo de regressão logística, os estimadores de máxima verossimilhança $\hat{\beta}_j$, $j = 1, \ldots, k$, são estimadores consistentes e eficientes de β_j. No entanto, para que $\hat{\beta}_0$ seja consistente e eficiente, esse deve ser corrigido de acordo com a seguinte expressão

$$\hat{\beta}_0 - \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right].$$

A maior vantagem da técnica de correção a priori é a facilidade de uso, já que os parâmetros do modelo de regressão logística podem ser estimados da forma usual e apenas o intercepto deve ser corrigido.

2.6 Estudo de Comparação

Com o objetivo de comparar o comportamento, isto é, a distribuição das probabilidades de inadimplência estimadas e a capacidade preditiva dos modelos obtidos pela regressão logística usual e pela regressão logística com seleção de amostras *state-dependent*, construímos os dois modelos a partir de amostras geradas [1] com diferentes tamanhos e proporções de *bons* e *maus* pagadores, as quais apresentamos a seguir:

1. 50% (10000 *bons* pagadores) e 50% (10000 *maus* pagadores)
2. 75% (30000 *bons* pagadores) e 25% (10000 *maus* pagadores)
3. 90% (90000 *bons* pagadores) e 10% (10000 *maus* pagadores)

Os principais resultados deste estudo de simulação, também encontrados em Louzada *et al.* (2012), são apresentados nas subseções seguintes.

[1] Ver detalhes das simulações em Louzada *et al.* (2012).

2.6.1 Medidas de desempenho

Nesta subseção apresentamos os principais resultados do estudo de simulação referentes à capacidade preditiva dos modelos ajustados segundo as duas técnicas estudadas, a regressão logística usual e a regressão logística com seleção de amostras *state-dependent*. As Tabelas 2.2 e 2.3 apresentam os intervalos de 95% de confiança empíricos para as medidas de desempenho.

Os resultados empíricos apresentados na Tabela 2.2 nos revelam que a técnica de regressão logística usual produz bons resultados apenas quando a amostra utilizada para o desenvolvimento do modelo é balanceada, 50% *bons* pagadores e 50% *maus* pagadores, com valores similares para as medidas de sensibilidade e especificidade. À medida que o grau de desbalanceamento aumenta, a sensibilidade diminui consideravelmente, assumindo valores menores que 0,5 quando há 90% *bons* pagadores e 10% *maus* pagadores na amostra de treinamento, ao passo que a especificidade aumenta, atingindo valores próximos de 1. Notamos também que o valor de MCC diminui à medida que o desbalanceamento se torna mais acentuado.

Os comentários com relação aos resultados obtidos utilizando o modelo de regressão logística com seleção de amostras *state-dependent* são análogos aos do modelo de regressão logística usual. Ou seja, a capacidade preditiva de ambos os modelos são próximas.

Tabela 2.2: Intervalos de confiança empíricos 95% para as medidas de desempenho, regressão logística usual.

	Grau de desbalanceamento das amostras		
Medidas	50% - 50%	75% - 25%	90% - 10%
SENS	[0,8071; 0,8250]	[0,5877; 0,6008]	[0,3249; 0,3307]
SPEC	[0,8187; 0,8334]	[0,9331; 0,9366]	[0,9768; 0,9777]
VPP	[0,8179; 0,8400]	[0,8247; 0,8359]	[0,8258; 0,8341]
VPN	[0,8004; 0,8250]	[0,8047; 0,8170]	[0,8075; 0,8145]
CAT	[0,8177; 0,8242]	[0,8123; 0,8194]	[0,8101; 0,8155]
MCC	[0,6354; 0,6485]	[0,5787; 0,5866]	[0,4404; 0,4439]

Tabela 2.3: Intervalos de confiança empíricos 95% para as medidas de desempenho, regressão logística com seleção de amostras *state-dependent*.

	Grau de desbalanceamento das amostras		
Medidas	50% - 50%	75% - 25%	90% - 10%
SENS	[0,8061; 0,8221]	[0,5870; 0,6008]	[0,3258; 0,3278]
SPEC	[0,8206; 0,8333]	[0,9330; 0,9366]	[0,9773; 0,9775]
VPP	[0,8225; 0,8392]	[0,8237; 0,8365]	[0,8306; 0,8321]
VPN	[0,7989; 0,8211]	[0,8045; 0,8180]	[0,8088; 0,8106]
CAT	[0,8173; 0,8241]	[0,8120; 0,8193]	[0,8111; 0,8127]
MCC	[0,6348; 0,6484]	[0,5779; 0,5859]	[0,4407; 0,4426]

2.6.2 Probabilidades de inadimplência estimadas

O modelo de regressão logística usual determina as probabilidades de inadimplência originais, enquanto que o modelo de regressão logística com seleção de amostras *state-dependent* determina as probabilidades corrigidas ou ajustadas. As Figuras 2.2 a 2.4 apresentam as curvas da probabilidade de inadimplência obtidas dos modelos original e ajustado, segundo os três graus de desbalanceamento considerados. Observamos que, independentemente do grau de desbalanceamento da amostra de treinamento, as probabilidades estimadas sem o ajuste no termo constante da equação estão abaixo das probabilidades com o ajuste. Ou seja, o modelo de regressão logística subestima a probabilidade de inadimplência. Notamos, também, que a distância entre as curvas diminui à medida que o grau de desbalanceamento da amostra se torna mais acentuado. Para o caso de amostras balanceadas, 50% *bons* pagadores e 50% *maus* pagadores, a distância entre as curvas é a maior observada, enquanto que para o caso de amostras desbalanceadas com 90% *bons* pagadores e 10% *maus* pagadores, as curvas estão muito próximas uma da outra.

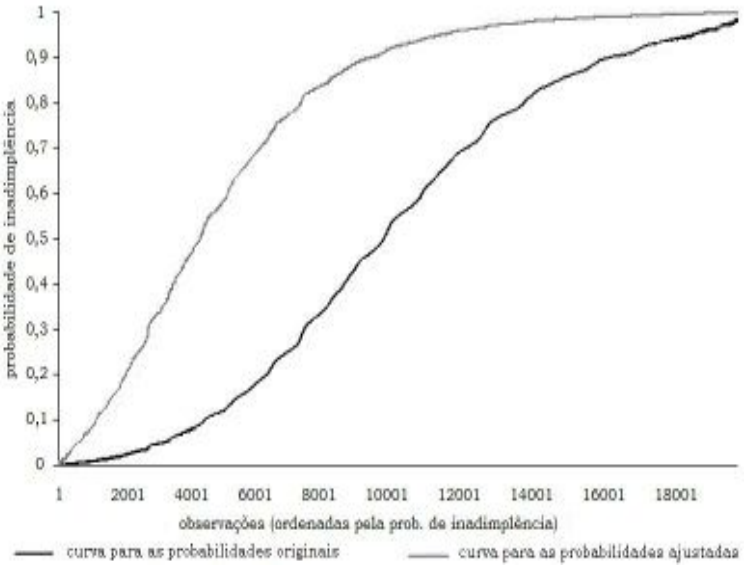

Figura 2.2: Distribuição das probabilidades de inadimplência estimadas, 50% *bons* pagadores e 50% *maus* pagadores.

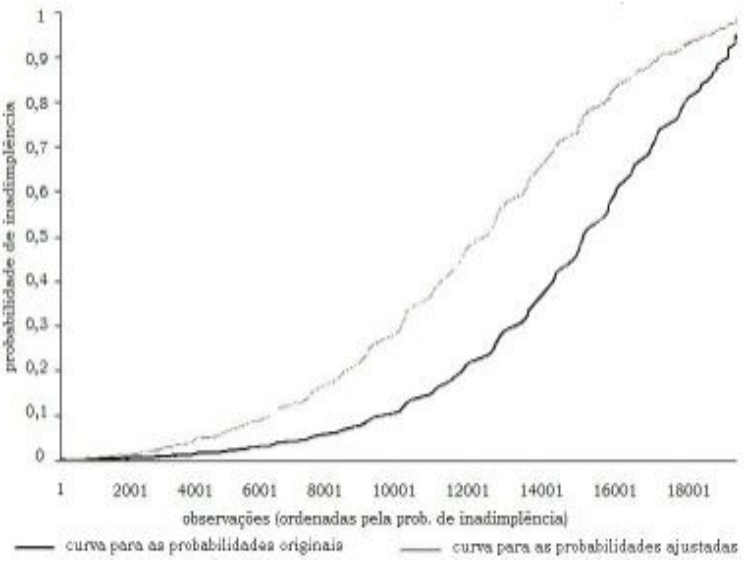

Figura 2.3: Distribuição das probabilidades de inadimplência estimadas, 75% *bons* pagadores e 25% *maus* pagadores.

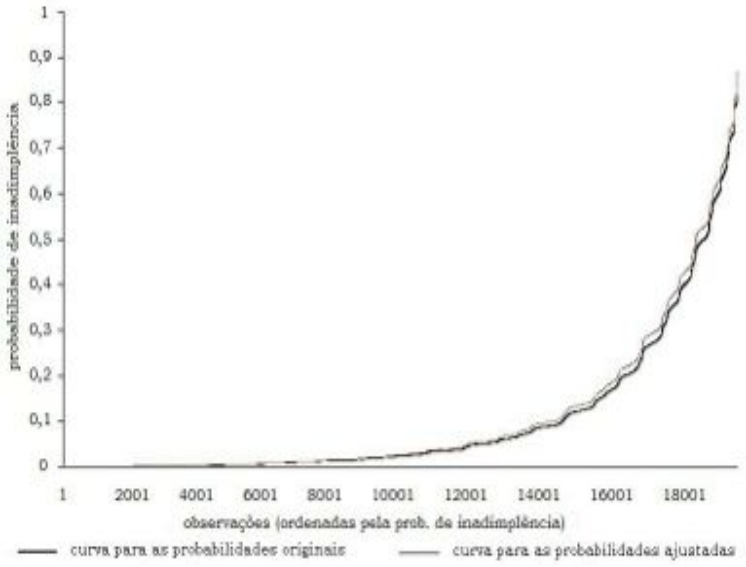

Figura 2.4: Distribuição das probabilidades de inadimplência estimadas, 90% *bons* pagadores e 10% *maus* pagadores.

2.7 Regressão Logística com Erro de Medida

Em várias áreas de aplicação da Estatística existem situações em que não é possível medir uma ou mais covariáveis, sem algum tipo de erro. Entre as possíveis razões podemos citar o custo ou a inviabilidade de coleta dos dados. Nestes casos, o que observamos são covariáveis com erros de medidas. No contexto de *Credit Scoring*, a presença da variável medida com erro pode surgir, por exemplo, no momento em que utilizamos a renda presumida como uma covariável do modelo de crédito. Renda presumida é uma predição da variável *Renda* obtida a partir de um específico modelo. Entre os trabalhos envolvendo erros de medida para modelo de regressão logística, podemos citar Thoresen & Laake (2007), Rosner *et al.* (1989) e Carroll *et al.* (1995). Nesta seção apresentamos o modelo de regressão logística com erro de medida e alguns métodos de estimação.

2.7.1 Função de verossimilhança

Seja Y uma variável resposta binária e X uma covariável não observada. Por simplicidade, usamos apenas a covariável não observada no modelo. Considere a função de densidade $f_{Y|X}(y|x)$ de Y condicionada a X. Seja $f_{YWX}(y, w, x)$ a função de densidade conjunta de (Y, W, X), em que W é a variável observada em substituição a X.

Considerando as observações (y_i, w_i), $i = 1, \ldots, n$, do vetor aleatório (Y, W), a função de verossimilhança pode ser escrita da seguinte forma,

$$\begin{aligned} L(\boldsymbol{\theta}|y, w) &= \prod_{i=1}^{n} \int f_{YWX}(y_i, w_i, x_i) dx_i \\ &= \prod_{i=1}^{n} \int f_{Y|W,X}(y_i|w_i, x_i) f_{W|X}(w_i|x_i) f_X(x_i) dx_i, \quad (2.7) \end{aligned}$$

sendo $\boldsymbol{\theta}$ o vetor de parâmetros desconhecidos.

A distribuição condicional de Y dado X, $Y|X = x_i \sim Ber(\pi(x_i))$, em que a probabilidade de sucesso, $\pi(x_i)$, é escrita em função dos parâmetros desconhecidos, β_0 e β_1, na forma

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Seja ϵ o erro presente ao observarmos W ao invés de X. Considere que a variável observada W é a soma da variável não observada X e do erro de medida ϵ, ou seja,

$$W = X + \epsilon.$$

Supondo que $\epsilon \sim N(0, \sigma_e^2)$ e $X \sim N(\mu_x, \sigma_x^2)$ é fácil notar que $W|X = x_i \sim N(x_i, \sigma_e^2)$. Para evitarmos problema de não identificabilidade do modelo, consideramos conhecida a variância do erro de medida, σ_e^2, ou estimamos usando réplicas da variável W, de cada indivíduo da amostra.

2.7.2 Métodos de estimação

Entre os diferentes métodos de estimação presentes na literatura para o modelo logístico com erro de medida, destacamos o método de calibração da regressão, o método *naive* e a estimação por máxima verossimilhança pelo método de integração de Monte Carlo.

- **Calibração da Regressão**: Consiste em substituir a variável não observada X por alguma função de W, como por exemplo, a esperança estimada de X dado W. Após a substituição, os parâmetros são estimados de maneira usual. Mais detalhes deste método podem ser encontrados em Rosner *et al.* (1989).

- **Naive**: Consiste, simplesmente, em utilizar W no lugar da variável de interesse X e ajustar o modelo logístico por meios usuais.

- **Integração de Monte Carlo**: A integral da verossimilhança (2.7) não pode ser obtida de forma analítica e uma solução é a aproximação numérica via integração de Monte Carlo. Para maiores detalhes ver Thoresen & Laake (2007).

2.7.3 Renda presumida

Uma covariável importante para predizer se um cliente será inadimplente ou não em instituições bancárias é a sua renda. Se o cliente não pertence ao portfólio da instituição é possível que sua renda não esteja disponível. Nestes casos, modelos de renda presumida são utilizados e, consequentemente, a covariável renda é medida com erro. Um modelo utilizado para renda presumida é o modelo de regressão gama.

Como exemplo, considere as seguintes variáveis explicativas categóricas: profissão, com cinco categorias: varejistas, profissionais liberais, servidores públicos, executivos e outros, e escolaridade, com três categorias: ensino fundamental, médio e superior. Neste caso, como as variáveis profissão e escolaridade são categóricas, usamos variáveis *dummies*. Se uma variável apresenta k categorias, o modelo terá $k-1$ *dummies* referentes a essa variável. As Tabelas 2.4 e 2.5 mostram a

codificação utilizada, respectivamente, para as categorias das variáveis profissão e escolaridade.

Tabela 2.4: Codificação dos níveis da variável profissão.

Profissão	Variáveis *Dummies*			
	D_1	D_2	D_3	D_4
Varejistas	0	0	0	0
Liberais	1	0	0	0
Servidor Público	0	1	0	0
Executivos	0	0	1	0
Outros	0	0	0	1

Tabela 2.5: Codificação dos níveis da variável escolaridade.

Escolaridade	Variáveis *Dummies*	
	D_5	D_6
Ensino Fundamental	0	0
Ensino Médio	0	1
Ensino Superior	1	0

Considere X_i a renda do i-ésimo cliente. Suponha também que $X_i \sim Gama(\alpha_i, \beta_i)$. A distribuição gama pode ser reparametrizada por

$$\mu_i = \frac{\alpha_i}{\beta_i}, \quad \alpha_i = \nu \text{ e } \beta_i = \frac{\nu}{\mu_i}.$$

A distribuição gama reparametrizada pertence à família exponencial na forma canônica, cuja função de ligação é

$$\theta_i = -\frac{1}{\mu_i}.$$

Para este exemplo, um modelo de renda presumida é dado por

$$\mu_i = \frac{1}{\beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 D_{5i} + \beta_6 D_{6i}}.$$

Métodos de estimação para este modelo pode ser encontrado em McCullagh & Nelder (1997). Como o objetivo da instituição financeira é prever se o cliente será ou não inadimplente, podemos usar o modelo de regressão logística sendo que a variável resposta é a situação do cliente (inadimplente ou adimplente) e a covariável medida com erro é a renda presumida.

Capítulo 3

Modelagem Para Eventos Raros

Em muitas situações práticas, temos interesse em descrever a relação entre uma variável resposta extremamente desbalanceada e uma ou mais covariáveis. No mercado financeiro, comumente, o interesse reside em determinar as probabilidades de que clientes cometam ações fraudulentas ou não paguem a primeira fatura, sendo que a proporção destes clientes é muito pequena.

Existem alguns estudos na literatura que revelam que o modelo de regressão logística usual subestima a probabilidade do evento de interesse, quando este é construído utilizando bases de dados extremamente desbalanceadas (King & Zeng, 2001). Para este modelo, os estimadores de máxima verossimilhança são, assintoticamente, não viciados e, mesmo para grandes amostras, este vício persiste. McCullagh & Nelder (1989) sugerem um estimador para o vício, para qualquer modelo linear generalizado, adaptado por King & Zeng (2001) para o uso concomitante com amostras *state-dependent*, permitindo que uma correção seja efetuada nos estimadores de máxima verossimilhança. King & Zeng (2001) sugerem, ainda, que as correções sejam realizadas nas probabilidades do evento de interesse, estimadas por meio do modelo de regressão logística. Tais correções permitem diminuir o vício e o erro quadrático médio de tais probabilidades.

Outros modelos, presentes na literatura, desenvolvidos especial-

mente para a situação de dados binários desbalanceados, são o modelo logito generalizado, sugerido por Stukel (1988), e o modelo logito limitado, sugerido por Cramer (2004). O modelo logito generalizado possui dois parâmetros de forma e se ajusta melhor do que o modelo logito usual em situações em que a curva de probabilidade esperada é assimétrica. O modelo logito limitado permite estabelecer um limite superior para a probabilidade do evento de interesse.

Em alguns casos, a variável resposta pode ser, originalmente, fruto de uma distribuição discreta, exceto a Bernoulli, ou contínua e que, por alguma razão, foi dicotomizada através de um ponto de corte C arbitrário. O modelo de regressão logística pode agregar a informação sobre a distribuição da variável de origem no ajuste do modelo logito usual. Dessa forma, o modelo pode ter a variável resposta pertencente à família exponencial no contexto dos modelos lineares generalizados com função de ligação composta. Esta metodologia foi apresentada por Suissa & Blais (1995), considerando dados reais de estudos clínicos e também dados simulados com distribuição original lognormal. Dependendo do ponto de corte utilizado, a variável resposta pode apresentar um desbalanceamento muito acentuado.

Neste capítulo apresentamos os estimadores de King & Zeng (2001), estimadores KZ, juntamente com as probabilidades do evento de interesse corrigidas. Apresentamos uma breve discussão sobre as características dos modelos logito generalizado e logito limitado e o desenvolvimento de modelos de regressão logística com resposta de origem normal, exponencial e log-normal.

3.1 Estimadores KZ para o Modelo de Regressão Logística

Segundo King & Zeng (2001), na situação de eventos raros, o estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, vetor de coeficientes da regressão logística usual, é viciado, mesmo quando o tamanho da amostra é grande. Além disso, mesmo que $\hat{\boldsymbol{\beta}}$ seja corrigido pelo vício estimado, $P\left(Y = 1|\hat{\boldsymbol{\beta}}, \boldsymbol{x_i}\right)$ é viciado para $\pi(\boldsymbol{x_i})$. Nesta seção, discutimos métodos para a correção destes

estimadores.

3.1.1 Correção nos parâmetros

Segundo McCullagh & Nelder (1989), o vício do estimador do vetor de parâmetros de qualquer modelo linear generalizado pode ser estimado como

$$\text{vício}(\hat{\boldsymbol{\beta}}) = (X^{'}WX)^{-1}X^{'}W\boldsymbol{\xi}, \tag{3.1}$$

sendo que $X'WX$ é a matriz de informação de Fisher, $\boldsymbol{\xi}$ é um vetor com o i-ésimo termo $\xi_i = -0,5\mu_i^{''}/\mu_i^{'}Q_{ii}$, μ_i é a inversa da função de ligação que relaciona $\mu_i = E(Y_i)$ ao preditor linear $\eta_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}$, Q_{ii} é o i-ésimo elemento da diagonal principal de $X\left(X^{'}W'X\right)X^{'}$, $\mu_i^{'}$ e $\mu_i^{''}$ são as derivadas de primeira e segunda ordem de μ_i com relação a η_i dadas por

$$\mu_i^{'} = e^{\eta_i}/\left(1+e^{\eta_i}\right)$$

e

$$\mu_i^{''} = e^{\eta_i}\left(1-e^{\eta_i}\right)/(1+e^{\eta_i})^3.$$

Assim,

$$\xi_i = -0,5\left(\frac{1-e^{\eta_i}}{1+e^{\eta_i}}\right)Q_{ii}.$$

O cálculo do vício em (3.1) pode ser adaptado quando utilizamos amostras *state-dependent* considerando $P(Y_i = y_i) = \pi_i^{\omega_1 y_i}\left(1-\pi_i\right)^{\omega_0(1-y_i)}$, sendo $\omega_1 = \frac{\tau}{\bar{y}}$ e $\omega_0 = \frac{1-\tau}{1-\bar{y}}$, em que τ é a prevalência populacional e $\bar{y}$ é a prevalência amostral. Portanto,

$$\begin{aligned}
\mu_i &= E(Y_i) = \left(\frac{1}{1+e^{-\eta_i}}\right)^{\omega_1} = \pi_i^{\omega_1}, \\
\mu_i^{'} &= \omega_1\pi_i^{\omega_1}\left(1-\pi_i\right), \\
\mu_i^{''} &= \omega_1\pi_i^{\omega_1}\left(1-\pi_i\right)\left[\omega_1-\left(1-\omega_1\right)\pi_i\right], \\
\xi_i &= 0,5Q_{ii}\left[\left(1-\omega_1\right)\pi_i-\omega_1\right].
\end{aligned}$$

A matriz de informação de Fisher do modelo é dada por

$$-E\left(\frac{\partial^2 L_\omega\left(\boldsymbol{\beta}|y\right)}{\partial\beta_j\partial\beta_k}\right)=\sum_{i=1}^{n}\pi_i\left(1-\pi_i\right)x_j\omega_i x_k^{'}=\left[X^{'}W_\omega X\right]_{j,k},$$

com $W_\omega = diag\left[\pi_i\left(1-\pi_i\right)\omega_i\right]$.

O estimador corrigido pelo vício é dado por $\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}-\text{vício}(\hat{\boldsymbol{\beta}})$. Segundo McCullagh & Nelder (1989), a matriz de variâncias e covariâncias de $\tilde{\boldsymbol{\beta}}$ é aproximadamente $\left(\frac{n}{n+p-1}\right)^2 V(\hat{\boldsymbol{\beta}})$. Como $\left(\frac{n}{n+p-1}\right)^2 < 1$ temos que $V(\tilde{\boldsymbol{\beta}}) < V(\hat{\boldsymbol{\beta}})$, ou seja, a diminuição no vício dos estimadores do modelo causa uma diminuição na variância dos mesmos.

3.1.2 Correção nas probabilidades estimadas

De acordo com os resultados apresentados na subseção anterior, $\tilde{\boldsymbol{\beta}}$ é menos viciado do que $\hat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$ e, além disso, $V(\tilde{\boldsymbol{\beta}}) < V(\hat{\boldsymbol{\beta}})$. Assim, $\tilde{\pi}(\boldsymbol{x_i})$ é preferível a $\hat{\pi}(\boldsymbol{x_i})$. No entanto, segundo Geisser (1993) e King & Zeng (2001), este estimador não é ótimo porque não leva em conta a incerteza a respeito de $\boldsymbol{\beta}$, e isto pode gerar estimativas viesadas da probabilidade de evento.

Uma maneira de levar em contar a incerteza na estimação do modelo é escrever $\pi(\boldsymbol{x_i})$ como

$$P\left(Y_i=1\right)=\int P\left(Y_i=1|\boldsymbol{\beta^*}\right)P\left(\boldsymbol{\beta^*}\right)d\boldsymbol{\beta^*}, \tag{3.2}$$

sendo que $P\left(\cdot\right)$ representa a incerteza com relação a $\boldsymbol{\beta}$. Observe que a expressão (3.2) pode ser vista como $E^*_{\boldsymbol{\beta}}\left[P\left(Y_i=1|\boldsymbol{\beta^*}\right)\right]$. Sob o ponto de vista Bayesiano podemos usar a densidade a posteriori $\boldsymbol{\beta}\sim Normal\left[\tilde{\boldsymbol{\beta}},V(\tilde{\boldsymbol{\beta}})\right]$. Existem duas formas de calcular a integral em (3.2). A primeira é usando aproximação Monte Carlo, ou seja, retirando uma amostra de $\boldsymbol{\beta}$ a partir de $P\left(\boldsymbol{\beta}\right)$, inserindo esta amostra em $e^{\boldsymbol{x_i'\beta}}/\left(1+e^{\boldsymbol{x_i'\beta}}\right)$ e calculando a média destes valores. Aumentando o número de simulação nos permite aproximar $P\left(Y_i=1\right)$ a um grau de acurácia desejável. A segunda é expandindo em série de Taylor a expressão $\pi(\boldsymbol{x_0})=\frac{e^{\boldsymbol{x_0'\beta}}}{1+e^{\boldsymbol{x_0'\beta}}}$ em torno de $\tilde{\boldsymbol{\beta}}$

até a segunda ordem e, em seguida, tomando a esperança, ou seja,

$$\begin{aligned}\pi(\boldsymbol{x_0}) &= P\left(Y_0 = 1|\boldsymbol{\beta}\right) \\ &\approx \tilde{\pi}(\boldsymbol{x_0}) + \left[\frac{\partial\pi(\boldsymbol{x_0})}{\partial\boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right) \\ &+ \frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\left[\frac{\partial^2\pi(\boldsymbol{x_0})}{\partial\beta'\partial\boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right), \qquad (3.3)\end{aligned}$$

sendo

$$\begin{aligned}\left[\frac{\partial\pi(\boldsymbol{x_0})}{\partial\boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} &= \tilde{\pi}(\boldsymbol{x_0})\left(1-\tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right), \\ \left[\frac{\partial^2\pi(\boldsymbol{x_0})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta'}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} &= \left(0,5-\tilde{\pi}(\boldsymbol{x_0})\right)\tilde{\pi}(\boldsymbol{x_0})\left(1-\tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'}\Omega\boldsymbol{x_0}\end{aligned}$$

e Ω uma matriz de ordem $k \times k$ cujo $(k,j)-$ésimo elemento é igual a $\left(\beta_k - \tilde{\beta}_k\right)\left(\beta_j - \tilde{\beta}_j\right)$. Sob a perspectiva Bayesiana, $\pi(\boldsymbol{x_0})$ e $\boldsymbol{\beta}$ são variáveis aleatórias, mas por outro lado, $\tilde{\pi}(\boldsymbol{x_0})$ e $\tilde{\boldsymbol{\beta}}$ são funções dos dados.

Tomando a esperança da expressão (3.3), temos

$$\begin{aligned}E\left(\frac{e^{\boldsymbol{x_i'\beta}}}{1+e^{\boldsymbol{x_i'\beta}}}\right) &\approx \tilde{\pi}(\boldsymbol{x_0}) + \tilde{\pi}(\boldsymbol{x_0})\left(1-\tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'}b \\ &+ \left(0,5+\tilde{\pi}(\boldsymbol{x_0})\right)\left(\tilde{\pi}(\boldsymbol{x_0})-\tilde{\pi}^2(\boldsymbol{x_0})\right)\boldsymbol{x_0'}\left[V(\tilde{\boldsymbol{\beta}})+bb'\right]\boldsymbol{x_0'},\end{aligned}$$

com $b = E\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right) \approx 0$. Logo, podemos escrever $\pi(\boldsymbol{x_i})$ como

$$\pi_i = P\left(Y_i = 1\right) = \tilde{\pi}(\boldsymbol{x_i}) + C_i,$$

com

$$C_i = \left(0,5-\tilde{\pi}(\boldsymbol{x_i})\right)\tilde{\pi}(\boldsymbol{x_i})\left(1-\tilde{\pi}(\boldsymbol{x_i})\right)\boldsymbol{x_i'}V(\tilde{\boldsymbol{\beta}})\boldsymbol{x_i} \qquad (3.4)$$

representando o fator de correção. Analisando o fator de correção da expressão (3.4), notamos que este fator, por ser diretamente proporcional a $V(\tilde{\beta})$, será maior à medida que o número de zeros na amostra diminui.

Devido a não-linearidade da forma funcional logística, mesmo

que $E\left(\tilde{\boldsymbol{\beta}}\right) \approx \boldsymbol{\beta}$, $E\left(\tilde{\pi}\right)$ não é aproximadamente igual a π. Na realidade, interpretando a integral em (3.2) como um valor esperado sob $\tilde{\boldsymbol{\beta}}$, podemos escrever $E_{\tilde{\boldsymbol{\beta}}}\left(\tilde{\pi}\right) \approx \pi + C_i$, e o fator de correção pode ser pensado como um viés. Surpreendentemente, subtraindo o fator de correção $(\tilde{\pi} - C_i)$ teremos um estimador aproximadamente não-viesado, mas, adicionando o viés, $(\tilde{\pi} + C_i)$ teremos um estimador com erro quadrático médio menor do que o estimador usual.

O estimador da probabilidade do evento de interesse $\pi(\boldsymbol{x_i})^* = \tilde{\pi}(\boldsymbol{x_i}) + C_i$ é chamado de estimador KZ1 e o estimador aproximadamente não viesado para a probabilidade do evento de interesse é chamado de estimador KZ2.

3.2 Modelo Logito Limitado

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 do evento de interesse. Ou seja, dada as covariáveis, é expressa por

$$\pi\left(\boldsymbol{x_i}\right) = \omega \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}, \tag{3.5}$$

com $0 < \omega < 1$.

O modelo (3.5) foi proposto por Cramer (2004), que ajustou o modelo de regressão logística usual, o modelo complementar log-log e o modelo logito limitado a uma base de dados de uma instituição financeira holandesa. Os dados em questão apresentavam baixa incidência do evento de interesse e o teste de Hosmer-Lemeshow indicou que o modelo logito limitado foi o mais adequado para os dados em questão. Segundo Cramer (2004), o parâmetro ω tem a capacidade de absorver o impacto de possíveis covariáveis significativas excluídas da base de dados.

O modelo logito limitado também foi utilizado por Moraes (2008) em dados reais de fraude bancária. De acordo com os resultados obtidos, o modelo logito limitado apresentou uma performance superior ao modelo logito usual, segundo as estatísticas que medem a qualidade do

ajuste: AIC (*Akaike Information Criterion*), SC (*Schwarz criterion*) e KS (Estatística de Kolmogorov-Smirnov).

3.2.1 Estimação

Como a variável resposta Y_i possui distribuição de probabilidade $Bernoulli(\pi(\boldsymbol{x_i}))$, as probabilidades do evento de interesse e seu complemento são dadas por $P(Y_i = 1|\boldsymbol{x_i}) = \pi(\boldsymbol{x_i})$ e $P(Y_i = 0|\boldsymbol{x_i}) = 1 - \pi(\boldsymbol{x_i})$, respectivamente. Assim, o logaritmo da função de verossimilhança é dado por

$$\begin{aligned} l(\boldsymbol{\beta}, \omega) &= \sum_{i=1}^{n} \left\{ y_i \log \left[\omega \left(\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \right) \right] \right. \\ &+ \left. (1 - y_i) \log \left[1 - \omega \left(\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \right) \right] \right\} I_{(0,1)}(\omega). \end{aligned} \tag{3.6}$$

Os estimadores de máxima verossimilhança são obtidos maximizando-se a expressão (3.6). As derivadas da função de verossimilhança com relação aos parâmetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$ e ω são dadas, respectivamente, por

$$\sum_{i=1}^{n} \omega \left[y_i - \pi(\boldsymbol{x_i}) \right], \tag{3.7}$$

$$\sum_{i=1}^{n} x_{ij} \omega \left[y_i - \pi(\boldsymbol{x_i}) \right], \text{ para } j = 1, \ldots, p-1 \tag{3.8}$$

e

$$\sum_{i=1}^{n} \left[\frac{y_i - \pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})} \right]. \tag{3.9}$$

Notamos que as equações (3.7) a (3.9) são não-lineares nos parâmetros, impossibilitando a solução explícita do sistema de equações e, portanto, recorremos a algum método de otimização para encontrar as estimativas de máxima verossimilhança dos parâmetros em questão. Porém, devido às características da função, sua maximização, utilizando os procedimentos usuais de otimização numérica, nem sempre é possível. Uma alter-

nativa é considerar a reparametrização $\theta = \log\left(\frac{\omega}{1-\omega}\right)$. Desta forma, a função de verossimilhança pode ser reescrita como

$$\begin{aligned} l(\boldsymbol{\beta}, \omega) &= \sum_{i=1}^{n} \left\{ y_i \log\left[\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+e^{-x_i'\beta}}\right)\right] \right. \\ &+ \left. (1-y_i)\log\left[1-\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+e^{-x_i'\beta}}\right)\right]\right\}, \quad (3.10) \end{aligned}$$

com $-\infty < \theta < \infty$. Para maximizar (3.10) podemos utilizar o algoritmo BFGS implementado no *software* R, proposto simultaneamente e independentemente por Broyden (1970), Fletcher (1970), Goldfarb (1970) e Shanno (1970).

3.2.2 Método BFGS

O método BFGS (Broyden, Fletcher, Goldfarb e Shanno) é uma técnica de otimização que utiliza um esquema iterativo para buscar um ponto ótimo. O processo de otimização parte de um valor inicial θ_0 e na iteração t verifica-se se o ponto θ_t encontrado é ou não o ponto ótimo. Caso este não seja o ponto ótimo, calcula-se um vetor direcional Δ_t e realiza-se uma otimização secundária, conhecida como "busca em linha", para encontrar o tamanho do passo ótimo λ_t. Desta forma, em $\theta_{t+1} = \theta_t + \lambda_t\Delta_t$, uma nova busca pelo ponto ótimo é realizada.

O vetor direcional Δ_t é tomado como $\Delta_t = \omega_t g_t$, em que g_t é o gradiente (vetor de primeiras derivadas) no passo t e ω_t é uma matriz positiva-definida calculada no passo t.

O método BFGS, assim como o método de Newton-Raphson, é um caso particular do método gradiente. O método de Newton-Raphson utiliza $\omega_t = -H^{-1}$, sendo H a matriz hessiana. Entretanto, quando o valor do ponto inicial θ_0 não está próximo do ponto ótimo, a matriz $-H^{-1}$ pode não ser positiva-definida, dificultando o uso do método. Já no método BFGS, uma estimativa de $-H^{-1}$ é construída iterativamente. Para tanto, gera-se uma sequência de matrizes $\omega_{t+1} = \omega_t + E_t$. A matriz ω_0 é a matriz identidade e E_t é, também, uma matriz positiva-definida,

pois em cada passo do processo iterativo ω_{t+1} é a soma de duas matrizes positivas-definida.

A matriz E_t é dada por

$$E_t = \frac{\delta_t \delta_t}{\delta_t' \gamma_t} + \frac{\omega_t \gamma_t \gamma_t' \omega_t}{\gamma_t' \omega_t \gamma_t} - \nu_t dt,$$

com $\delta_t = \lambda_t \Delta_t = \theta_{t+1} - \theta_t$, $\gamma_t = g(\theta_{t+1}) - g(\theta_t)$, $\nu_t = \gamma_t' \omega_t \gamma_t$ e $d_t = \left(\frac{1}{\gamma_t \delta_t}\right) \gamma_t - \left(\frac{1}{\gamma_t' \omega_t \gamma_t}\right) \omega_t \gamma_t$.

3.3 Modelo Logito Generalizado

O modelo de regressão logística usual é amplamente utilizado para modelar a dependência entre dados binários e covarîáveis. Este sucesso deve-se à sua vasta aplicabilidade, à simplicidade de sua fórmula e sua fácil interpretação. Este modelo funciona bem em muitas situações. Contudo, tem como suposições que a simetria seja no ponto $\frac{1}{2}$ da curva de probabilidade esperada, $\pi(x)$, e que sua forma seja a da função de distribuição acumulada da distribuição logística. Segundo Stukel (1988), nas situações em que as caudas da distribuição de $\pi(x)$ são mais pesadas o modelo logito usual não funciona bem.

Na Figura 3.1 encontram-se os gráficos da curva de probabilidade $\pi(x)$ considerando as prevalências amostrais de 1%, 15%, 30% e 50%. De acordo com estes gráficos, na situação de baixa prevalência, a suposição de simetria na curva $\pi(x)$ no ponto $\frac{1}{2}$ não é verificada. Este fato indica que o modelo logito usual não é adequado para ajustar dados com desbalanceamento acentuado.

Muitos autores apresentaram propostas de modelos que generalizam o modelo logito padrão. Prentice (1976) sugeriu uma ligação bi-paramétrica utilizando a função de distribuição acumulada da transformação $\log(F_{2m_1,2m_2})$. A família de distribuições $\log(F)$ contém a distribuição logística ($m_1 = m_2 = 1$), a Gaussiana, as distribuições do mínimo e máximo extremo, a exponencial, a distribuição de Laplace e a exponencial refletida. Este modelo é eficaz em muitas situações devido à sua flexibilidade, no entanto, apresenta dificuldades computaci-

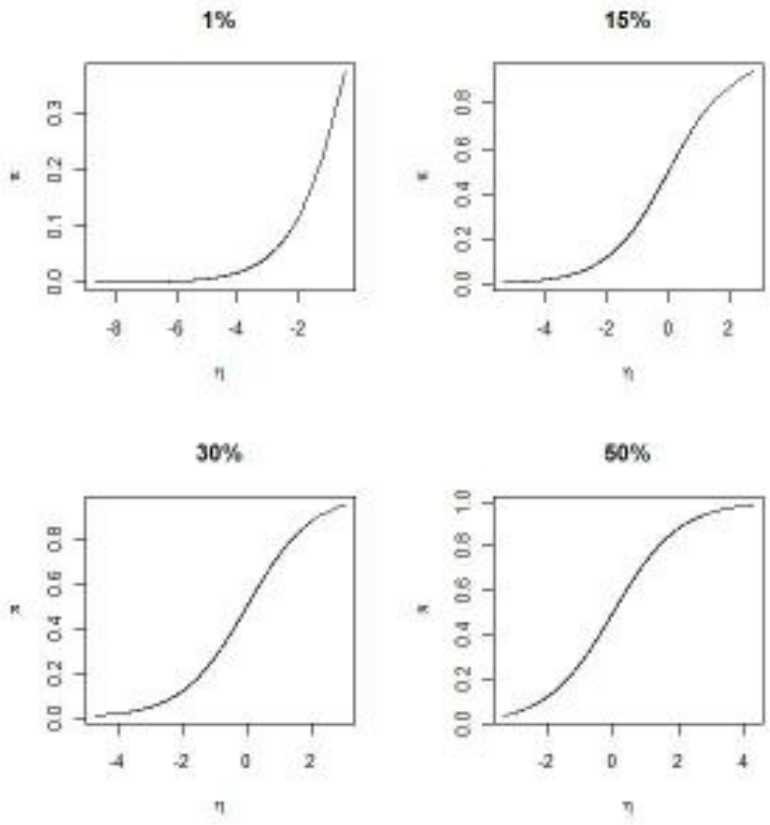

Figura 3.1: Curvas de probabilidade para diferentes prevalências.

onais, já que as curvas de probabilidades estimadas devem ser calculadas através da soma de séries infinitas. Pregibon (1980) definiu uma família de funções de ligação que inclui a ligação logito como um caso especial. A curva de probabilidade esperada é a solução implícita da equação $\left(\pi^{\lambda_1-\lambda_2}-1\right)/\left(\lambda_1-\lambda_2\right)-\left[(1-\pi)^{\lambda_1+\lambda_2}-1\right]/\left(\lambda_1+\lambda_2\right)=\eta$. O parâmetro λ_1 controla as caudas da distribuição e λ_2 determina a simetria da curva de probabilidade π. Aranda-Ordaz (1981) sugerem dois modelos uniparamétricos, um deles simétrico e o outro assimétrico, como alternativas ao modelo logito padrão. O modelo simétrico é dado pela transformação $2[\pi^{\delta_1}-(1-\pi)^{\delta_1}]/\delta_1[\pi^{\delta_1}+(1-\pi)^{\delta_1}]=\eta$, sendo que, quando $\delta_1 \to 0$, temos o modelo logito. Já o modelo assimétrico é dado por $\log\left\{[(1-\pi)^{-\delta_2}-1]/\delta_2\right\}=\eta$, sendo que, quando $\delta_2=1$, temos o modelo

logito e, quando $\delta_2 = 0$, temos o modelo complementar log-log.

A forma geral do modelo logito generalizado proposto por Stukel (1988) é dada por

$$\pi_\alpha(\boldsymbol{x_i}) = \frac{e^{h_\alpha(\eta)}}{1 + e^{h_\alpha(\eta)}},$$

ou

$$\log\left(\frac{\pi_\alpha(\boldsymbol{x_i})}{1 - \pi_\alpha(\boldsymbol{x_i})}\right) = h_\alpha(\eta),$$

sendo que $h_\alpha(\eta)$ é uma função não-linear estritamente crescente indexada por dois parâmetros de forma, α_1 e α_2.

Para $\eta \geq 0$ ($\pi \geq \frac{1}{2}$), $h_\alpha(\eta)$ é dada por

$$h_\alpha = \begin{cases} \alpha_1^{-1}\left(e^{\alpha_1|\eta|} - 1\right), & \alpha_1 > 0 \\ \eta, & \alpha_1 = 0 \\ -\alpha_1^{-1}\log\left(1 - \alpha_1|\eta|\right), & \alpha_1 < 0 \end{cases}$$

e, para $\eta \leq 0$ ($\pi \leq \frac{1}{2}$),

$$h_\alpha = \begin{cases} -\alpha_2^{-1}\left(e^{\alpha_2|\eta|} - 1\right), & \alpha_2 > 0 \\ \eta, & \alpha_2 = 0 \\ \alpha_2^{-1}\log\left(1 - \alpha_2|\eta|\right), & \alpha_2 < 0 \end{cases}$$

Quando $\alpha_1 = \alpha_2 = 0$ o modelo resultante é o logito usual.

A função h aumenta mais rapidamente ou mais vagarosamente do que a curva do modelo logito usual, como podemos ver na Figura 3.2. Os parâmetros α_1 e α_2 determinam o comportamento das caudas. Se $\alpha_1 = \alpha_2$ a curva de probabilidade correspondente é simétrica.

3.3.1 Estimação

Os estimadores de máxima verossimilhança de $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ podem ser obtidos utilizando o algoritmo delta sugerido por Jorgensen (1984). Este algoritmo é equivalente ao procedimento de mínimos quadrados ponderados para o ajuste dos parâmetros de modelos lineares generalizados, porém, neste caso, a matriz do modelo é atualizada depois de cada iteração. No caso do modelo logito generalizado, a matriz do modelo

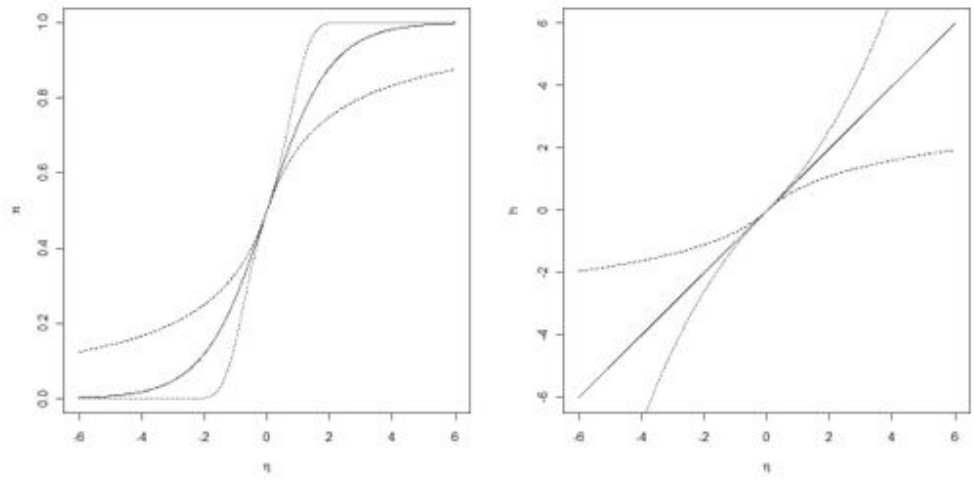

Figura 3.2: Gráficos de π e h: a linha sólida representa o modelo logito usual, a linha tracejada corresponde ao modelo logito generalizado com $\boldsymbol{\alpha} = (-1, -1)$ e a linha pontilhada corresponde ao modelo logito generalizado com $\boldsymbol{\alpha} = (0,25; 0,25)$.

é a matriz usual $\boldsymbol{X}$ acrescida de duas colunas adicionais contendo as variáveis $\boldsymbol{z'} = (z_{1,t+1}, z_{2,t+1}) = \left(-\frac{\partial g(\pi)}{\partial \alpha_1}, -\frac{\partial g(\pi)}{\partial \alpha_2}\right)|_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}}$, sendo

$$z_{i,t+1} = \begin{cases} \alpha_i^{-2} \left\{\alpha_i|\eta| - 1 + \exp(-\alpha_i|\eta|)\right\} \operatorname{sgn}(\eta), & \alpha_i > 0 \\ \dfrac{1}{2}\eta^2 \operatorname{sgn}(\eta), & \alpha_i = 0 \\ \alpha_i^{-2} \left\{\alpha_i|\eta| + (1 - \alpha_i|\eta|) \log(1 - \alpha_i|\eta|)\right\} \operatorname{sgn}(\eta), & \alpha_i < 0. \end{cases}$$

com $\alpha_i = \hat{\alpha}_{i,t}$, $\eta = \hat{\eta}_t = \boldsymbol{x'}\boldsymbol{\beta_t}$ e $(\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}, \hat{\boldsymbol{\alpha}}_{\boldsymbol{t}})$ a estimativa de $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ na t-ésima iteração. Os elementos de $\boldsymbol{z}$ correspondem aos parâmetros de forma e devem ser atualizados a cada iteração.

Stukel (1985) sugere, ainda, uma maneira alternativa de estimar os parâmetros do modelo logito generalizado, que consiste em estimar o vetor de parâmetros $\boldsymbol{\beta}$ considerando vários valores de $\boldsymbol{\alpha}$ e escolhendo como estimativa o conjunto de valores que maximize a verossimilhança.

3.4 Modelo Logito com Resposta de Origem

Em muitas situações práticas possuímos uma variável resposta binária com distribuição de origem pertencente a algumas classes de distribuições, isto é, a variável resposta possui alguma distribuição de origem, exceto a de Bernoulli e, por alguma razão, foi dicotomizada através de um ponto de corte C arbitrário. Assim, podemos adicionar características da distribuição original da variável resposta no modelo de regressão logística usual. Esta metodologia foi proposta inicialmente por Suissa (1991) e ampliada por Suissa & Blais (1995) em uma estrutura de modelos lineares generalizados com função de ligação composta para ajustar modelos de regressão logística com resposta log-normal. Nesta seção, apresentamos a construção e o desenvolvimento dos modelos de regressão logística para os casos de variável resposta com distribuição normal, exponencial e log-normal.

3.4.1 Modelo normal

Sejam $R_1, R_2, \ldots, R_n$ variáveis aleatórias independentes seguindo distribuição $N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. Considerando C um ponto de corte arbitrário e $Y_1, Y_2, \ldots, Y_n$ tal que $Y_i = 1$, se $R_i > C$ e $Y_i = 0$, se $R_i \leq C$, temos $P(Y_i = 1) = P(R_i > C) = \pi_i$ e $P(Y_i = 0) = P(R_i \leq C) = 1 - \pi_i$. Desta forma, $Y_i \sim Bernoulli(\pi_i)$.

Na presença de $p - 1$ covariáveis relacionadas com a variável resposta, a probabilidade do evento de interesse para o i-ésimo cliente pode ser escrita através do modelo de regressão logística na forma

$$E(Y_i) = \pi(\boldsymbol{x_i}) = P(Y_i = 1) = \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} = g^{-1}(\boldsymbol{x_i'\beta}), \tag{3.11}$$

$i = 1, \ldots, n$ em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ é o vetor de parâmetros

associado às covariáveis do modelo. Logo,

$$\begin{aligned}\pi(\boldsymbol{x_i}) &= P\left(Y_i > C\right) = P\left[Z_i > \frac{C-\mu_i}{\sigma}\right] \\ &= P\left[Z_i < \frac{\mu_i - C}{\sigma}\right] = \phi\left(\frac{\mu_i - C}{\sigma}\right), \end{aligned} \tag{3.12}$$

sendo Z_i uma variável aleatória com distribuição normal padrão e distribuição acumulada ϕ. Das equações (3.11) e (3.12), temos que

$$\pi(\boldsymbol{x_i}) = \phi\left(\frac{\mu_i - C}{\sigma}\right) = g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right), \quad i = 1, \ldots, n, \tag{3.13}$$

ou ainda,

$$g\left(\pi(\boldsymbol{x_i})\right) = g\left[\phi\left(\frac{\mu_i - C}{\sigma}\right)\right] = \boldsymbol{x_i'}\boldsymbol{\beta} = \eta_i, \quad i = 1, \ldots, n,$$

na qual $g\left[\phi\left(\cdot\right)\right]$ é uma função de ligação composta que origina o preditor linear $\boldsymbol{x_i'}\boldsymbol{\beta}$. Tomando $\gamma_i = (\mu_i - C)/\sigma$ e assumindo σ conhecido, podemos dizer que este modelo faz parte da classe dos modelos lineares generalizados cujo componente aleatório é o conjunto de variáveis independentes com distribuição $N\left(\gamma_i, 1\right)$ e a componente sistemática é dada pela função de ligação composta $g\left[\phi\left(\cdot\right)\right]$ e pelo preditor linear $\eta_i = \boldsymbol{x_i'}\boldsymbol{\beta}, i = 1, \cdots, n$.

A partir de (3.13) podemos escrever μ_i como

$$\mu_i = \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right)\right] + C, \quad i = 1, \ldots, n.$$

Logo, a função de verossimilhança pode ser escrita como

$$L\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{r}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(r_i - \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right)\right] - C\right)^2\right\},$$

e o logaritmo da função de verossimilhança é dado por

$$l\left(\beta, \sigma^2; \boldsymbol{r}\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(r_i - \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right)\right] - C\right)^2. \tag{3.14}$$

3.4.2 Modelo exponencial

Sejam $R_1, R_2, \ldots, R_n$ variáveis aleatórias independentes seguindo distribuição $Exponencial\,(\theta_i)$, isto é,

$$f\,(r_i) = \theta_i e^{-\theta_i r_i}, \quad \theta_i > 0, \quad i = 1, \ldots, n. \tag{3.15}$$

Dessa forma,

$$P\,(R_i > C) = e^{-\theta_i C}, \quad i = 1, \ldots, n. \tag{3.16}$$

A partir das equações (3.13) e (3.16), temos

$$e^{-\theta_i C} = g^{-1}\,(\boldsymbol{x_i'\beta}) \tag{3.17}$$

e, portanto,

$$g\left(e^{-\theta_i C}\right) = \boldsymbol{x_i'\beta}, \tag{3.18}$$

sendo $g\,[\exp\,(\cdot)]$ a função de ligação que origina o preditor linear $\boldsymbol{x_i'\beta}$, $i = 1, \ldots, n$.

A função de verossimilhança para o modelo logístico com resposta exponencial é dada por

$$L\,(\boldsymbol{\beta}; \boldsymbol{r}) = \prod_{i=1}^{n} \left\{ -\frac{\log\left[g^{-1}\,(\boldsymbol{x_i'\beta})\right]\left[g^{-1}\,(\boldsymbol{x_i'\beta})\right]^{-r_i/C}}{C} \right\}. \tag{3.19}$$

com θ_i dado por

$$\theta_i = -\frac{\log\left[g^{-1}\,(\boldsymbol{x_i'\beta})\right]}{C}.$$

Aplicando o logaritmo em (3.19) temos a função de log-verossimilhança dada por

$$l\,(\boldsymbol{\beta}; \boldsymbol{r}) = \sum_{i=1}^{n} \log\left\{-\log\left[g^{-1}\,(\boldsymbol{x_i'\beta})\right]\right\} - \frac{1}{C}\sum_{i=1}^{n} r_i \log\left[g^{-1}\,(\boldsymbol{x_i'\beta})\right] - n\log\,(C)\,.$$

3.4.3 Modelo lognormal

Sejam $R_1, R_2, \ldots, R_n$ variáveis aleatórias independentes seguindo distribuição $LN\,(\mu_i, \sigma^2)$, para $i = 1, \ldots, n$. Então, $\log\,(R_1), \ldots, \log\,(R_n)$

são variáveis aleatórias independentes seguindo distribuição normal com média μ_i e variância σ^2.

Devido à relação entre a distribuição lognormal e a distribuição normal, os resultados para o modelo lognormal podem ser obtidos utilizando os resultados apresentados Subseção 3.4.1. para o modelo normal. Para tal, basta substituir a constante C por $\log(C)$ e a variável resposta R_i por $\log(R_i)$, $i = 1, \ldots, n$. Desta forma, a probabilidade do evento de interesse para o i-ésimo cliente $\pi(\boldsymbol{x_i})$ é dada por

$$\pi(\boldsymbol{x_i}) = P\left[Z_i < \frac{\mu_i - \log(C)}{\sigma}\right] = \phi\left[\frac{\mu_i - \log(C)}{\sigma}\right], \quad i = 1, \ldots, n. \tag{3.20}$$

na qual Z_i é uma variável aleatória com distribuição normal padrão e distribuição acumulada ϕ. Logo, de (3.20) temos

$$\mu_i = \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right)\right] + \log(C). \tag{3.21}$$

Considerando (3.21), a função de verossimilhança pode ser escrita como

$$L\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{r}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\log(r_i) - \mu_i\right]^2\right\}, \tag{3.22}$$

com $\mu_i = \sigma\phi^{-1}\left[g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})\right] + \log(C)$, $i = 1, \ldots, n$, e a função de log-verossimilhança é escrita como

$$l(\boldsymbol{\beta}, \sigma; \boldsymbol{r}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\log(r_i) - \sigma\phi^{-1}\left[g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})\right] - \log(C)\right\}^2. \tag{3.23}$$

3.4.4 Estudo de simulação

Nesta seção apresentamos um estudo de simulação para analisarmos os desempenhos dos modelos logísticos com resposta de origem lognormal e usual, em duas prevalências. A distribuição lognormal é co-

mum para variáveis do tipo Renda, Valor de Sinistro e Gasto. As métricas vício, erro quadrático médio e erro absoluto médio são utilizadas para dar suporte nesta comparação.

Na geração dos dados utilizamos três variáveis explicativas com distribuição de Bernoulli, X_{i1}, X_{i2} e X_{i3}. Foram geradas 1000 amostras de tamanho $n = 5000$ com variável resposta $R_i \sim LN(\mu_i, \sigma^2)$, com $\mu_i = \sigma\phi^{-1}\left[g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\right] + log(C)]$, $i = 1, \ldots, 5000$. Os valores atribuídos para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ para a geração de μ_i foram, $\beta_0 = -7$, $\beta_1 = 1,0$, $\beta_2 = 2,0$, $\beta_3 = 5,0$ e $\sigma = 1,0$. O ponto de corte considerado foi $C = 10$. Duas prevalências, 0,01 e 0,1, são usadas nas bases. No primeiro caso de prevalência foram geradas covariáveis $X_{i1} \sim Bernoulli(0,1)$, $X_{i2} \sim Bernoulli(0,1)$ e $X_{i2} \sim Bernoulli(0,1)$ e no segundo caso foram geradas covariáveis $X_{i1} \sim Bernoulli(0,4)$, $X_{i2} \sim Bernoulli(0,4)$ e $X_{i2} \sim Bernoulli(0,4)$.

A Tabela 3.1 apresenta o vício amostral, o erro quadrático médio (EQM), o erro absoluto médio (EAM) e a média das estimativas dos parâmetros. Notamos que o vício, EQM e EAM das estimativas do modelo logito com resposta de origem são inferiores às mesmas métricas, calculadas através das estimativas produzidas pelo modelo logito usual.

Tabela 3.1: Qualidade do ajuste - distribuição de origem lognormal.

		Modelo logístico usual				Modelo resposta de origem			
p		Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
0,01	β_0	-0,146	0,460	0,351	-7,146	-0,011	0,013	0,093	-7,011
	β_1	-0,022	0,113	0,265	0,977	-0,0004	0,016	0,101	0,999
	β_2	-0,0003	0,094	0,241	1,999	-0,0005	0,016	0,101	1,999
	β_3	0,104	0,468	0,357	5,104	-0,008	0,0146	0,096	4,991
0,10	β_0	-0,046	0,100	0,249	-7,046	-0,004	0,015	0,101	-7,004
	β_1	-0,001	0,013	0,092	0,998	-0,002	0,004	0,055	0,997
	β_2	0,001	0,014	0,095	2,001	0,003	0,004	0,055	2,003
	β_3	0,043	0,088	0,233	5,043	0,001	0,010	0,083	5,001

Os intervalos de confiança empíricos da razão das estimativas dos modelos logito usual e logito com resposta de origem lognormal são apresentados Tabela 3.2. Os resultados indicam que as estimativas de ambos os modelos convergem. Além disso, a amplitude destes intervalos considerando a prevalência 0,10 é inferior à amplitude apresentada pelos intervalos considerando a prevalência de 0,01.

Tabela 3.2: Intervalos de confiança empíricos da razão das estimativas - distribuição de origem lognormal.

p		90%	95%	99%
0,01	β_0	(0,932; 1,126)	(0,919; 1,159)	(0,894; 1,254)
	β_1	(0,402; 1,480)	(0,302; 1,563)	(0,077; 1,786)
	β_2	(0,761; 1,238)	(0,724; 1,288)	(0,617; 1,367)
	β_3	(0,900; 1,174)	(0,883; 1,216)	(0,847; 1,356)
0,10	β_0	(0,944; 1,072)	(0,932; 1,085)	(0,921; 1,125)
	β_1	(0,844; 1,157)	(0,818; 1,192)	(0,780; 1,240)
	β_2	(0,922; 1,076)	(0,908; 1,089)	(0,879; 1,089)
	β_3	(0,930; 1,097)	(0,920; 1,117)	(0,891; 1,169)

Os intervalos empíricos para a razão das chances dos modelos logito usual e logito, com resposta de origem lognormal, são mostrados nas Tabelas 3.3 e 3.4. Estes resultados indicam uma precisão superior nas estimativas obtidas através do modelo logito com resposta de origem. Além disso, quando comparamos a precisão dos resultados considerando as duas prevalências, observamos que a amplitude dos intervalos construídos através de amostras com prevalência de 0,10 é inferior à amplitude dos intervalos obtidos considerando amostras com prevalência de 0,01.

Tabela 3.3: Intervalos de confiança empíricos da razão das chances - modelo logito usual - distribuição lognormal.

p		90%	95%	99%
0,01	β_1	(1,457; 4,469)	(1,306; 4,940)	(1,062; 6,093)
	β_2	(4,397; 12,062)	(3,973; 13,435)	(3,336; 16,684)
	β_3	(87,12; 369,905)	(81,527; 437,423)	(62,517; 904,604)
0,10	β_1	(2,234; 3,276)	(2,159; 3,431)	(2,053; 3,712)
	β_2	(6,059; 8,966)	(5,886; 9,274)	(5,574; 10,018)
	β_3	(101,215; 255,177)	(94,817; 288,825)	(82,262; 402,277)

As Tabelas 3.5 e 3.6 apresentam a probabilidade de cobertura e a amplitude média, respectivamente, dos intervalos de confiança assintóticos dos parâmetros dos modelos logito usual e logito com resposta de origem lognormal. O nível de confiança nominal é observado nos intervalos de ambos os modelos; contudo, os intervalos para os parâmetros do modelo logito com resposta de origem são mais precisos.

Tabela 3.4: Intervalos de confiança empíricos da razão das chances - modelo logito com resposta de origem - distribuição de origem lognormal.

p		90%	95%	99%
0,01	β_1	(2,207; 3,329)	(2,130; 3,473)	(2,009; 3,890)
	β_2	(6,034; 9,192)	(5,818; 9,528)	(5,428; 10,209)
	β_3	(120,810; 180,553)	(117,774; 187,391)	(110,959; 199,106)
0,10	β_1	(2,433; 3,037)	(2,362; 3,123)	(2,300; 3,265)
	β_2	(6,636; 8,323)	(6,482; 8,496)	(6,168; 8,739)
	β_3	(124,913; 176,059)	(121,539; 180,823)	(115,152; 192,856)

Tabela 3.5: Probabilidade de cobertura - distribuição de origem lognormal.

		Modelo logístico usual			Modelo resposta de origem		
p		90%	95%	99%	90%	95%	99%
0,01	β_0	0,917	0,975	0,995	0,908	0,954	0,992
	β_1	0,898	0,952	0,993	0,921	0,961	0,992
	β_2	0,900	0,947	0,990	0,899	0,952	0,995
	β_3	0,905	0,970	0,992	0,910	0,967	0,993
0,10	β_0	0,914	0,961	0,992	0,901	0,948	0,989
	β_1	0,899	0,954	0,994	0,899	0,953	0,987
	β_2	0,900	0,944	0,993	0,899	0,946	0,983
	β_3	0,900	0,960	0,994	0,901	0,948	0,987

Tabela 3.6: Amplitude média - distribuição de origem lognormal.

		Modelo logístico usual			Modelo resposta de origem		
p		90%	95%	99%	90%	95%	99%
0,01	β_0	3,670	4,387	5,752	0,388	0,464	0,608
	β_1	1,094	1,308	1,715	0,432	0,517	0,678
	β_2	0,990	1,183	1,551	0,417	0,498	0,653
	β_3	3,662	4,376	5,739	0,412	0,492	0,645
0,10	β_0	0,969	1,159	1,519	0,395	0,472	0,619
	β_1	0,387	0,463	0,607	0,226	0,270	0,354
	β_2	0,384	0,459	0,602	0,236	0,282	0,370
	β_3	0,908	1,085	1,423	0,330	0,395	0,518

3.5 Análise de Dados Reais

Nesta seção analisamos um conjunto de dados reais de uma instituição financeira, cuja variável resposta representa fraude em cartão de crédito. As covariáveis são descritas com nomes fictícios. Os dados originais possuem 172452 observações, das quais apenas 2234 representam

fraude, cerca de 1,30% do total.

A base de dados possui dez covariáveis, além da variável resposta que indica fraude. As covariáveis foram categorizadas em dez classes e, após análises bivariadas, definimos a categorização final utilizada nos ajustes dos modelos. Aplicamos a técnica de seleção de variáveis *stepwise* e esta técnica indicou cinco covariáveis que deveriam permanecer no modelo final, duas covariáveis quantitativas X_1 e X_3 e três covariáveis dummies, X_2, com quatro categorias, X_4, com dois categorias, e X_5, com seis categorias. A Tabela 3.7 mostra as estimativas dos parâmetros do modelo de regressão logística usual e os testes individuais de Wald. As linhas com repetição de uma covariável indicam as categorias desta variável.

A base original foi dividida em amostra treinamento, em que os modelos foram ajustados, com 70% dos dados, e amostra teste com 30% dos dados, utilizada para calcular as medidas preditivas referente a cada modelo.

Tabela 3.7: Parâmetros estimados para o modelo logito usual.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor-p
Intercepto	1	-2,677	0,159	280,6489	0,0001
X_1	1	0,588	0,034	290,583	0,0001
X_2	1	0,500	0,062	65,021	0,0001
X_2	1	0,215	0,064	11,307	0,0008
X_2	1	-0,068	0,067	1,052	0,304
X_2	1	-0,336	0,064	27,249	0,0001
X_3	1	0,522	0,087	36,013	0,0001
X_4	1	-0,411	0,146	7,916	0,004
X_4	1	0,445	0,275	2,616	0,105
X_5	1	-0,720	0,130	30,625	0,0001
X_5	1	-0,233	0,085	7,560	0,006
X_5	1	0,094	0,069	1,853	0,173
X_5	1	0,278	0,070	15,788	0,0001
X_5	1	0,161	0,110	2,134	0,144
X_5	1	0,449	0,093	23,300	0,0001

De acordo com o teste de Wald, todas as variáveis apresentadas na Tabela 3.7 são significativas. A Tabela 3.8 apresenta as estimativas dos parâmetros do modelo logito limitado juntamente com o teste de Wald, que indica que todas as variáveis apresentadas são significativas no modelo, assim como o parâmetro w.

A Tabela 3.9 apresenta as estimativas dos parâmetros do modelo

Tabela 3.8: Parâmetros estimados para o modelo logito limitado.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor-p
w	1	0,234	0,089	2,611	0,009
Intercepto	1	-0,770	0,686	-1,121	0,261
X_1	1	0,704	0,077	9,116	<0,001
X_2	1	0,602	0,091	6,546	<0,001
X_2	1	0,240	0,078	3,083	0,0020
X_2	1	-0,082	0,078	-1,058	0,289
X_2	1	-0,401	0,080	-4,964	<0,0001
X_3	1	0,677	0,138	4,891	<0,001
X_4	1	-0,553	0,265	-2,086	0,036
X_4	1	0,707	0,516	1,370	0,170
X_5	1	-0,795	0,146	-5,437	<0,001
X_5	1	-0,270	0,097	-2,773	0,005
X_5	1	0,099	0,080	1,232	0,217
X_5	1	0,323	0,086	3,749	0,0001
X_5	1	0,149	0,129	1,155	0,247
X_5	1	0,528	0,122	4,305	<0,001

logito generalizado, juntamente com o teste de Wald. A Tabela 3.10 mostra os valores das medidas AIC, BIC e -2log(verossimilhança) para os três modelos ajustados. O modelo logito limitado apresenta o melhor desempenho seguido pelo modelo logito usual e pelo modelo logito generalizado.

Tabela 3.9: Parâmetros estimados para o modelo logito generalizado.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor-p
α_1	1	1,02			
Intercepto	1	-1,266	0,050	-25,106	<0,001
X_1	1	0,140	0,008	16,233	<0,001
X_2	1	0,118	0,015	7,875	<0,001
X_2	1	0,046	0,015	3,031	0,002
X_2	1	-0,016	0,015	-1,116	0,264
X_2	1	-0,079	0,013	-5,728	<0,001
X_3	1	0,131	0,023	5,564	<0,001
X_4	1	-0,103	0,046	-2,255	0,024
X_4	1	0,136	0,089	1,514	0,129
X_5	1	-0,147	0,025	-5,816	<0,001
X_5	1	-0,052	0,018	-2,881	0,003
X_5	1	0,017	0,015	1,101	0,270
X_5	1	0,060	0,016	3,717	0,0002
X_5	1	0,025	0,025	1,007	0,313
X_5	1	0,104	0,023	4,478	<0,001

A Tabela 3.11 apresenta as medidas preditivas para os modelos

Tabela 3.10: Medidas de qualidade do ajuste.

Modelo	AIC	BIC	-2log(verossimilhança)
Logito Usual	8726,026	8854,676	8696,815
Logito Limitado	8725,026	8819,315	8693,026
Logito Generalizado	8729,12	8823,409	8697,120

logito usual, logito limitado, logito generalizado e logito usual construídos em amostras balanceadas com estimadores KZ1 e KZ2. Notamos que o modelo logito usual com estimadores KZ2 construído em amostras balanceadas apresenta um desempenho preditivo ligeiramente superior aos demais modelos. O Coeficiente de Correlação de Mathews está bastante próximo para todos os modelos. O modelo logito generalizado apresenta a maior sensibilidade seguido do modelo logito usual aplicado em amostras balanceadas com estimadores KZ2.

Tabela 3.11: Medidas preditivas.

Modelo	SENS	SPEC	VPP	VPN	CAT	MCC
Logito Usual	0,632	0,683	0,052	0,985	0,682	0,109
Logito Usual-Balanceado	0,622	0,673	0,051	0,985	0,662	0,107
Logito Limitado	0,632	0,681	0,052	0,985	0,680	0,108
Logito Generalizado	0,713	0,616	0,049	0,987	0,618	0,109
Usual KZ1	0,701	0,627	0,049	0,986	0,629	0,109
Usual KZ2	0,703	0,674	0,053	0,985	0,674	0,113

Dos resultados apresentados podemos concluir que os desempenhos preditivos dos modelos de classificação estudados foram similares. No entanto, o modelo logito usual com estimadores KZ é o que apresenta medidas indicando um poder predito mais efetivo.

Capítulo 4

Credit Scoring com Inferência dos Rejeitados

Os modelos de *Credit Scoring*, como mencionado no Capítulo 1, são desenvolvidos a partir de bases históricas de performance de crédito dos clientes, além de informações pertinentes ao produto. A amostra utilizada no desenvolvimento de um modelo de *Credit Scoring* deve refletir as características presentes na carteira, ou na população total. Porém, devido ao fato de que vários clientes não aprovados no processo de seleção não tem seus comportamentos observados e são excluídos da amostra utilizada na construção do modelo, mesmo pertencendo à população total de clientes, suas peculiaridades não serão absorvidas por este modelo. Desta forma, as amostras usuais, formadas apenas pelos clientes aceitos, não são totalmente representativas da população de interesse e, possivelmente, existe um vício amostral intrínseco. A Figura 4.1 apresenta um esquema da distribuição dos dados para um modelo de *Credit Scoring*.

Esse vício pode ser mais ou menos influente no modelo final de acordo com a proporção de rejeitados em relação ao total de proponentes. Quanto maior essa proporção, mais importante é o uso de alguma estratégia para a correção deste vício. Para solucionar esse problema, apresentamos, neste capítulo, algumas técnicas de inferência dos rejeitados.

Figura 4.1: Esquema da distribuição dos dados para um modelo de *Credit Scoring*.

4.1 Métodos de Inferência dos Rejeitados

Uma premissa fundamental na modelagem estatística é que a amostra selecionada para o modelo represente a população total de interesse. Porém, nos problemas de *Credit Scoring*, geralmente, essa premissa é violada, pois são utilizados apenas os proponentes aceitos, cujos comportamentos foram observados. Os rejeitados, por sua vez, não são observados e são usualmente descartados do processo de modelagem.

A inferência dos rejeitados é a associação de uma resposta para o indivíduo não observado de forma que seja possível utilizar suas informações em um novo modelo. Os principais métodos podem ser vistos em Ash & Meesters (2002), Banasik & Crook (2005), Crook & Banasik (2004, 2007), Feelders (2003), Hand (2001) e Parnitzke (2005).

Por mais simples que seja a definição do problema que estamos abordando, é um trabalho complexo construir técnicas realmente eficientes de inferência dos rejeitados. As técnicas, por sua vez, possuem a característica de serem mais ineficazes à medida que a proporção de rejeitados aumenta e, quanto maior a proporção de rejeitados, maior é a necessidade de alguma estratégia para reduzir o vício amostral (Ash & Meesters, 2002). Neste seção consideramos as técnicas da reclassificação, ponderação e parcelamento.

4.1.1 Método da reclassificação

Uma das estratégias mais simples para inserir os proponentes rejeitados na construção do modelo é, simplesmente, considerar toda po-

pulação dos rejeitados como sendo *maus* pagadores. Essa estratégia procura reduzir o viés amostral baseado na ideia de que, na população dos rejeitados, esperamos que a maioria seja de *maus* pagadores, embora certamente possa haver *bons* pagadores em meio aos rejeitados. Adotado esse método, os *bons* clientes que foram, inicialmente, rejeitados serão classificados erroneamente e, consequentemente, os proponentes não rejeitados com perfis similares serão prejudicados (Thomas *et al.*, 2002). No entanto, pela característica desta técnica, é de se esperar um modelo mais sensível, em que os elementos positivos sejam melhor identificados, o que é de grande importância no contexto de escoragem de crédito.

4.1.2 Método da ponderação

Provavelmente, esta é a estratégia mais presente na literatura. Como proposto em Banasik & Crook (2005), este método consiste em assumir que a probabilidade de um cliente ser *mau* pagador independe do fato de ter sido aceito ou não. Neste método, os rejeitados não contribuem diretamente para o modelo e as suas representações são feitas pelos proponentes que possuem escores semelhantes, mas que foram aceitos.

Os proponentes aceitos são responsáveis em levar a informação dos rejeitados para o modelo através de pesos atribuídos, calculados de acordo com os escores associados. O peso para o indivíduo i é dado por $P_i = 1/(1 - E_i)$, sendo E_i o seu escore. A ideia é que o peso seja inversamente proporcional ao escore obtido, fazendo com que os indivíduos aceitos mais próximos do ponto de corte obtenham peso maior, representando assim a população dos rejeitados. Para um cliente aceito com escore $0,9$ (lembramos que o evento de interesse é a inadimplência, portanto, escores altos representam altos riscos de inadimplência), seu peso é dado por $P = 1/(1 - 0,9) = 1/0,1 = 10$, ou seja, esse elemento de alto risco é considerado com peso 10 no modelo ponderado. Cada peso representa o número de vezes que cada observação será replicado no banco de dados. O indivíduo que tem peso 10 terá sua observação replicada 10 vezes na base de treinamento, o que faz com que o modelo logístico ajustado seja mais influenciado por esse elemento.

O modelo ponderado é gerado a partir dos indivíduos aceitos

com os pesos atribuídos. Em Parnitzke (2005), é alcançado um aumento de 1,03% na capacidade de acerto total em dados simulados, e nenhum aumento quando baseado num conjunto de dados reais. Em Alves (2008), os resultados foram bem similares aos do modelo logístico usual.

4.1.3 Método do parcelamento

De acordo com Parnitzke (2005), para desenvolver essa estratégia, devemos considerar um novo modelo, construído a partir da base dos proponentes aceitos. O próximo passo é dispor os solicitantes utilizados neste novo modelo em faixas de escores. Essas faixas podem ser determinadas de forma que os elementos escorados se distribuam de modo uniforme, como apresentado na Tabela 4.1. Em cada faixa de escore verificamos a taxa de inadimplência e, então, atribuímos escores aos rejeitados. Para cada rejeitado é associado uma resposta do tipo *bom* ou *mau* pagador, de forma aleatória e de acordo com as taxas de inadimplência observadas nos proponentes aceitos. Assim, é construído um modelo com os clientes aceitos e rejeitados com suas devidas respostas inferidas.

Tabela 4.1: Esquema da distribuição dos rejeitados no método do parcelamento

Faixa de Escore	Bons	Maus	% Maus	Rejeitados	Bons	Maus
0-121	285	15	5,00	25	24	1
121-275	215	85	28,33	35	25	10
275-391	165	135	45,00	95	52	43
391-601	100	200	66,66	260	87	173
601-1000	40	260	86,66	375	50	325

Conforme os escores aumentam, a concentração de *maus* fica maior em relação a de *bons* pagadores (o evento de interesse aqui é *mau* pagador). Essa proporção é utilizada para distribuir os rejeitados, que pertencem a tais faixas de escores, como indicado nas duas últimas colunas da Tabela 4.1.

Os resultados apresentados por essa técnica também são similares aos usuais, e em alguns casos, leva a pequenos melhoramentos.

4.1.4 Outros métodos

Uma estratégia, não muito conveniente para a empresa, é a de aceitar todos os solicitantes por um certo período de tempo, para que seja possível criar um modelo completamente não viciado. No entanto, essa ideia não é bem vista, pois o risco envolvido em aceitar proponentes nos escores mais baixos pode não compensar o aumento de qualidade que o modelo possa vir a gerar. Outra ideia seria aceitar apenas uma pequena parcela dos que seriam rejeitados, o que é prática em algumas instituições.

Outro método é o uso de informações de mercado (*bureau* de crédito), obtidas de alguma central de crédito que possui registros de atividades de créditos dos proponentes. Isto permite verificar como os proponentes duvidosos se comportam em relação aos outros tipos de compromissos, como contas de cartões de créditos, de energia, de telefone, seguros etc.

Os proponentes rejeitados são avaliados em dois momentos; o primeiro é quando solicitam o crédito e o segundo ocorre em algum tempo depois, permitindo, assim, um período de avaliação pré-determinado. No primeiro momento, pode ser que os proponentes não possuíam irregularidade e permaneceram nesta situação ou adquiriram alguma irregularidade durante o período de avaliação. De forma análoga, os que possuíam irregularidade, podem ou não possuir no segundo momento. Após uma comparação entre as informações obtidas e as informações da proposta de crédito, classificamos o indivíduo como *bom* ou *mau* pagador.

Um novo modelo é construído considerando o banco de dados com os clientes aceitos (classificados como *bom* ou *mau* pagador segundo a própria instituição) acrescido dos clientes rejeitados com resposta definida a partir de suas informações de mercado. Para a construção de um modelo com esta estratégia, devemos considerar que, certamente, existem mais informações acerca dos proponentes do que nas outras estratégias descritas, e, portanto, esperamos um melhor modelo. No entanto, o acesso a essas informações pode requerer um investimento financeiro que não deve ser desconsiderado (Rocha & Andrade, 2002).

4.2 Aplicação

Dois bancos de *Credit Scoring* de livre domínio, disponíveis na *internet* no *website* do *UCI Machine Learning Repository*, foram utilizados para ilustrar as estratégias de inferência dos rejeitados apresentadas neste capítulo. Modelos de regressão logística foram ajustados e as medidas de avaliação, como sensibilidade (SENS), especificidade (SPEC), valor preditivo positivo (VPP), valor preditivo negativo (VPN), acurácia ou capacidade de acerto total (CAT), coeficiente de correlação de Matthews (MCC) e custo relativo (CR), descritas no Capítulo 1, foram usadas para avaliar a qualidade do ajuste.

A primeira base é a *German Credit Data*, que consiste de 20 variáveis cadastrais, sendo 13 categóricas e 7 numéricas, e 1000 observações de utilizadores de crédito, dos quais 700 correspondem a *bons* pagadores e 300 (prevalência de 30% de positivos) a *mau*s pagadores. A segunda base é o *Australian Credit Data*, que consiste de 14 variáveis, sendo 8 categóricas e 6 contínuas, e 690 observações, das quais 307 (prevalência de 44,5% de positivos) são inadimplentes e 383 são adimplentes.

Para simular a situação em que temos rejeitados na amostra, foram separados os indivíduos do banco de dados de ajustes que obtiveram escore mais alto segundo um modelo proposto, com uma metade aleatória de observações do banco de dados para avaliação.

A implementação do método da reclassificação é muito simples. Em cada indivíduo da população dos rejeitados é inferida a resposta *mau* pagador e, com uma nova base constituída dos aceitos e dos rejeitados, é construído o modelo de regressão logística e o *bagging* (ver Capítulo 5).

Na estratégia da ponderação devemos ter, inicialmente, um modelo aceita - rejeita, que forneça a probabilidade de inadimplência de todos os proponentes. Com este modelo atribuímos escore a cada cliente e associamos um peso em cada indivíduo da população dos aceitos, como descrito na Subseção 4.1.2.

No método do parcelamento devemos inferir o comportamento dos rejeitados a partir das taxas de inadimplência, observadas na população dos aceitos. O procedimento consiste em ajustar um modelo a partir dos aceitos e dividir os proponentes em faixas de escores ho-

mogêneas. Consideramos 7 faixas de escore, sendo que esse número foi escolhido devido a divisibilidade que é necessária em relação ao tamanho das amostras de treinamento. Em cada faixa, é calculada a taxa de inadimplência, verificando quantos são *maus* pagadores em relação ao total. Essa proporção aumenta na medida em que os escores aumentam e, nas faixas mais altas, esperamos altas taxas de inadimplência enquanto que nos escores menores esperamos taxas de inadimplência menores.

Ainda com o modelo dos aceitos, atribuímos escores à população dos rejeitados. Utilizando as mesmas faixas de escore dos aceitos, distribuímos os rejeitados *escorados* e, por fim, atribuímos de forma aleatória a resposta *bom*/*mau* pagador na mesma proporção das taxas obtidas nos aceitos. Assim a inferência está completa e o modelo final é gerado com os aceitos acrescidos dos rejeitados.

A análise é feita considerando 10%, 30% e 50% de rejeitados simulados. Cada modelo foi simulado 200 vezes, variando a amostra teste dos dados reais. Os resultados obtidos são resumidos pelos seus valores médios.

No *Australian Credit Data* obtivemos o menor custo relativo no método da reclassificação, enquanto que as demais estratégias apresentaram resultados piores que as do modelo usual. Em relação ao MCC e acurácia obtivemos resultados análogos, com as maiores medidas ainda no método da reclassificação.

Na prevalência 30%, o método da reclassificação e ponderação foram melhores, sendo que o primeiro método apresentou MCC e acurácia maiores que o do segundo. Na prevalência 50% nenhuma estratégia superou o modelo usual em relação ao custo relativo, enquanto que o método da reclassificação obteve o maior MCC.

No *German Credit Data* com prevalência de rejeitados 10% e 30%, o método da reclassificação foi o único que apresentou melhoras, usando as métricas custo relativo e MCC, em relação ao usual. O método da ponderação foi o único que apresentou melhoras, usando a acurácia. Na prevalência 50% o modelo com reclassificação supera os demais em relação ao MCC.

Podemos notar que em diversas situações as estratégias de inferências podem trazer ganhos positivos na modelagem, ainda que pe-

quenos. No geral, o método que mais se destacou foi o da reclassificação, apresentando melhorias na maioria das configurações utilizadas. Os métodos da ponderação e parcelamento apresentaram bons resultados apenas em algumas situações, não diferindo muito do modelo logístico usual.

Em síntese, de acordo com os resultados apresentados, podemos dizer que a melhor estratégia para um modelo de *Credit Scoring* seria o uso da reclassificação. Sua estrutura de modelagem é simples, o aumento do custo computacional é mínimo e induz a um modelo com sensibilidade maior. Ainda que o viés amostral continue presente, de uma maneira diferente e teoricamente menor, os modelos gerados tendem a identificar com uma maior precisão a população dos *maus* pagadores.

Tabela 4.2: Inferência dos rejeitados no *German* e *Australian Credit Data*

	Medidas de	Australian			German		
	Avaliação	10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,81577	0,83640	0,93270	0,76371	0,84486	0,89352
	SENS	0,38247	0,34607	0,18663	0,39300	0,28011	0,18656
	VPP	0,78290	0,80213	0,86486	0,51568	0,57179	0,59152
	VPN	0,67840	0,66734	0,61080	0,76446	0,74514	0,72698
	CAT	0,62295	0,61820	0,60070	0,65250	0,67543	0,68143
	MCC	0,31492	0,29732	0,24889	0,20374	0,19771	0,18054
	CR	0,40297	0,41317	0,42637	0,33000	0,37270	0,39910
RECLASS.	SPEC	0,80279	0,82423	0,81820	0,71762	0,73714	0,66505
	SENS	0,42888	0,38146	0,33517	0,45767	0,43722	0,47122
	VPP	0,77510	0,78079	0,76942	0,50502	0,50442	0,43902
	VPN	0,68922	0,66869	0,66220	0,77615	0,77430	0,78279
	CAT	0,63640	0,62720	0,60325	0,63963	0,64717	0,60690
	MCC	0,33108	0,32485	0,28626	0,21600	0,22211	0,19386
	CR	0,39480	0,39617	0,40257	0,29077	0,36420	0,39980
POND.	SPEC	0,93117	0,93523	0,94360	0,78681	0,84310	0,88362
	SENS	0,13090	0,14416	0,12112	0,36522	0,27944	0,19611
	VPP	0,84548	0,83496	0,86095	0,52650	0,56567	0,59888
	VPN	0,58867	0,59705	0,58899	0,75836	0,74274	0,73139
	CAT	0,57505	0,58320	0,57760	0,66033	0,67400	0,67737
	MCC	0,21532	0,23893	0,22877	0,19947	0,18609	0,17576
	CR	0,41397	0,39660	0,41310	0,44000	0,42090	0,42990
PARC.	SPEC	0,82414	0,87541	0,87757	0,75219	0,79490	0,86848
	SENS	0,30371	0,22180	0,21011	0,36256	0,29944	0,20733
	VPP	0,74920	0,74826	0,66729	0,49688	0,51014	0,59824
	VPN	0,65100	0,62264	0,60909	0,75472	0,74218	0,73017
	CAT	0,59255	0,58455	0,58055	0,63530	0,64627	0,67013
	MCC	0,24761	0,26282	0,21562	0,17135	0,16359	0,17002
	CR	0,41027	0,41890	0,42543	0,33538	0,41120	0,41470

Capítulo 5

Combinação de Modelos de *Credit Scoring*

Uma das estratégias mais utilizadas para aumentar a precisão em uma classificação é o uso de combinação de modelos. A ideia consiste em tomar as informações fornecidas por diferentes mecanismos e agregar essas informações em uma única predição. No contexto de *Credit Scoring*, a estratégia é acoplar as informações por reamostragem dos dados de treinamento.

Breiman (1996) propôs a técnica *bagging*, que é baseada na reamostragem com reposição dos dados de treinamento, gerando vários modelos distintos para que, então, possam ser combinados. Neste capítulo descrevemos o algoritmo *bagging* e algumas formas de combinação de escores.

5.1 *Bagging* de Modelos

O *bagging* (*bootstrap aggregating*) é uma técnica em que construímos diversos modelos baseados nas réplicas *bootstrap* de um banco de dados de treinamento. Todos os modelos são combinados, de forma a encontrar um preditor que represente a informação de todos os modelos gerados.

A característica principal, que deve estar presente na base de

dados, para que este procedimento apresente bons resultados é a instabilidade. Um modelo é instável se pequenas variações nos dados de treinamento leva a grandes alterações nos modelos ajustados. Quanto mais instável é o classificador básico, mais variados serão os modelos ajustados pelas réplicas *bootstrap* e, consequentemente, teremos diferentes informações fornecidas pelos modelos, aumentando a contribuição para o preditor combinado. Se o classificador básico for estável, as réplicas gerariam, praticamente, os mesmos modelos e não haveriam contribuições relevantes para o preditor combinado final. Algoritmos de modelagem, como redes neurais e árvores de decisão, são exemplos de classificadores usualmente instáveis (Kuncheva, 2004). Em Bühlmann & Yu (2002) é feita uma análise do impacto da utilização do *bagging* no erro quadrático médio e na variância do preditor final, utilizando uma definição algébrica de instabilidade.

Desde que a técnica *bagging* foi publicada, diversas variantes foram desenvolvidas. Bühlmann & Yu (2002) propõem a variante *subagging* (*subsample aggregating*), que consiste em retirar amostras aleatórias simples, de tamanho menores, dos dados de treinamento. A combinação é feita, usualmente, por voto majoritário, mas é possível também o uso de outras técnicas. Essa estratégia apresenta resultados ótimos quando o tamanho das amostras é a metade do tamanho do conjunto de dados de treinamento (*half-subagging*). No artigo é mostrado que os resultados com *half-subagging* são praticamente iguais aos do *bagging*, principalmente em amostras pequenas.

Louzada-Neto *et al.* (2011) propõem um procedimento que generaliza a ideia de reamostragem do *bagging*, chamado *poly-bagging*. A estratégia é fazer reamostras sucessivas nas próprias amostras *bagging* originais. Cada reamostragem aumenta um nível na estrutura e complexidade da implementação. Os resultados obtidos por simulações foram expressivos, mostrando que é possível reduzir ainda mais a taxa de erro de um modelo. A técnica se mostra poderosa em diversas configurações de tamanhos amostrais e prevalências.

Desta forma, na modelagem via *bagging*, a aplicação dos novos clientes deve passar por todos os modelos construídos na estrutura, ou seja, cada cliente é avaliado por todos os modelos. Com essas in-

formações, um novo escore será obtido, por meio da aplicação dos escores anteriores, usando uma específica função combinação.

O procedimento *bagging*, com B representando o número de réplicas utilizadas, é descrito nos seguintes passos:

- Geramos $L_1^*, \ldots, L_B^*$ réplicas *bootstrap* da amostra treinamento L;
- Para cada réplica i geramos o modelo com preditor $S_i^*, i = 1, \ldots, B$;
- Combinamos os preditores para obter o preditor *bagging* S^*.

Na próxima seção discutimos várias propostas para a combinação dos $S_i^*, i = 1, \ldots, B$. Para isto, considere os preditores S_i^* e a função combinação $c(S_1^*, \ldots, S_B^*) = S^*$.

5.2 Métodos de Combinação

5.2.1 Combinação via média

A combinação via média é uma das mais comuns na literatura, de fácil implementação, e é dada por

$$S^* = c(S_1^*, \ldots, S_B^*) = \frac{1}{B} \sum_{i=1}^{B} S_i^*. \tag{5.1}$$

Em termos gerais, como proposto em Kuncheva (2004), podemos escrever a equação (5.1) como caso particular da equação

$$S^* = \left(\frac{1}{B} \sum_{i=1}^{B} (S_i^*)^\alpha \right)^{\frac{1}{\alpha}}, \tag{5.2}$$

quando $\alpha = 1$.

Essa formulação permite a dedução de outros tipos menos comuns de combinação, que podem ser utilizadas em situações mais específicas. Além do caso $\alpha = 1$, gerando a combinação por média, temos

outros casos particulares interessantes. Se $\alpha = -1$, a equação (5.2) representa uma combinação via média harmônica, se $\alpha \to 0$ a equação representa uma combinação via média geométrica. Se $\alpha \to -\infty$ a equação representa uma combinação via mínimo e se $\alpha \to \infty$ a equação representa uma combinação via máximo.

Estas estratégias podem ser usadas de acordo com o conservadorismo ou otimismo que desejamos exercer sobre a modelagem. Quanto menor o valor de α, mais próxima estaremos da combinação via mínimo, que é otimista por tomar o menor escore dentre os modelos gerados. Se escolhemos valores altos para α, o valor do escore tenderá a aumentar, representando uma combinação com tendências conservadoras.

5.2.2 Combinação via voto

A combinação por voto é também uma estratégia simples. Iniciamos associando o escore com a classificação final dos clientes. Seja C_i^* a variável que corresponde à classificação associada ao escore S_i^*, definida a partir do ponto de corte c escolhido, isto é,

$$C_i^* = 1 \text{ se } S_i^* > c_i \text{ e } C_i^* = 0 \text{ caso contrário.}$$

A partir dos classificadores C_i^*, definimos a combinação por voto majoritário da seguinte forma:

$$C^* = 1 \text{ se } \sum_{i=1}^{B} C_i^* \geq \left[\frac{B}{2}\right] \text{ e } C^* = 0 \text{ caso contrário,} \tag{5.3}$$

com $[\cdot]$ representando a função maior inteiro. Nos casos em que B é ímpar, temos uma maioria absoluta dos classificadores, no entanto, quando B é par pode ocorrer casos de empate e, segundo a definição em (5.3), será classificado como 1.

Neste trabalho, analisamos a combinação via voto de uma maneira geral, variando todos os possíveis números de votos k. Assim,

$$C^* = 1 \text{ se } \sum_{i=1}^{B} C_i^* \geq k \text{ e } C^* = 0 \text{ caso contrário,}$$

com $k = 0, \ldots, B$.

5.2.3 Combinação via regressão logística

A combinação via regressão logística foi apresentada em Zhu *et al.* (2001). Esta estratégia consiste em combinar os preditores considerando-os como covariáveis em um modelo de regressão logística, ou seja,

$$S^* = \log\left(\frac{P(Y=1|S_1^*,\ldots,S_B^*)}{1-P(Y=1|S_1^*,\ldots,S_B^*)}\right) = \beta_0 + \sum_{i=1}^{B}\beta_i S_i^*,$$

em que $P(Y=1|S_1^*,\ldots,S_B^*)$ representa a probabilidade do evento de interesse.

Essa combinação pode ser interpretada como uma espécie de combinação linear ponderada, de forma que o modelo de regressão logística aponte os modelos mais influentes na explicação da variável resposta por meio de seus coeficientes. A combinação linear ponderada é dada por

$$S^* = \sum_{i=1}^{B} w_i S_i^*, \text{ tal que } \sum_{i=1}^{B} w_i = 1.$$

Quando escolhemos os valores de w_i de forma que maximize uma ou mais medidas preditivas temos um custo computacional adicional. Para pequenos valores de B o processo já é bastante ineficaz, inviabilizando uma escolha livre para este parâmetro, que normalmente não é tão baixo. Nesse sentido, a combinação via regressão logística apresenta uma boa alternativa e é computacionalmente eficaz.

5.3 Aplicação

Nesta seção aplicamos as técnicas apresentadas em um banco de dados de *Credit Scoring* de livre domínio, disponível na internet no *website* do *UCI Machine Learning Repository*. A base, *German Credit Data*, consiste de 20 variáveis cadastrais, sendo 13 categóricas e 7 numéricas,

e 1000 observações de utilizadores de crédito, dos quais 700 são *bons* pagadores e 300 (prevalência de 30% de positivos) são *maus* pagadores.

Como proposto em Hosmer & Lemeshow (2000), separamos 70% dos dados disponíveis como amostra de treinamento e os 30% restantes ficam reservados para o cálculo das medidas de desempenho dos modelos, como descritas no Capítulo 1.

Com a amostra de treinamento disponível, são retiradas 25 réplicas *bootstrap* e, então, construímos os modelos da estrutura do *bagging*. O valor de 25 réplicas foi escolhido baseado no trabalho de Breiman (1996), o qual mostra que as medidas preditivas analisadas convergem rapidamente em relação ao número de modelos. A diferença entre a modelagem com 25 e 50 réplicas foram mínimas. A partir dos modelos, construídos nas amostras *bootstrap*, atribuímos os escores para os clientes da amostra teste. Utilizando um método de combinação, determinamos o preditor final para cada cliente. A escolha dos pontos de corte é feita de tal forma que maximize o MCC do preditor final, analisando numericamente seu valor em cada incremento de 0,01 no intervalo $[0, 1]$. Para resultados estáveis, foram simuladas 1000 vezes cada modelagem. O *software* utilizado nos ajustes foi o SAS (versão 9.0) e o processo de seleção de variáveis utilizado nas regressões foi o *stepwise*.

Em todos os modelos foram utilizadas subamostras estratificadas em relação a variável resposta, isto é, cada subamostra gerada preservou a prevalência da resposta observada.

Na combinação via média utilizamos $\alpha = -11, -10, \cdots, 10, 11$. Na combinação via voto é necessário classificar cada escore gerado pelas réplicas *bootstrap*. A classificação do escore é feita buscando o valor do ponto de corte, por todo intervalo $[0, 1]$, que maximiza a medida de desempenho MCC. Analisaremos os modelos em todas as possíveis contagens de votos, isto é, para todo $k = 1, 2, \cdots, 25$. Na combinação via regressão logística, inicialmente, consideramos os modelos *bagging* da amostra de treinamento com os escores atribuídos nos clientes da própria amostra de treinamento. Com esses escores geramos o banco de dados para uma regressão logística, ou seja, os escores obtidos em cada modelo correspondem aos valores das covariáveis para a regressão. Os coeficientes estimados desta última regressão são utilizados para gerar o

escore combinado da amostra teste. Consideramos o caso da regressão logística sem intercepto, que é o que mais se aproxima de uma combinação ponderada e o caso da regressão logística com intercepto, a fim de verificar seu impacto como parâmetro extra na combinação.

No estudo foram feitas 1000 simulações, variando a distribuição da amostra teste e treinamento. Usamos as combinações via média, voto e regressão logística, e, também, o modelo usual. A Figura 5.1 mostra os resultados obtidos pelas combinação por voto. Observe que à medida em que os valores de k aumentam, o modelo torna-se menos conservador. A sensibilidade e o valor preditivo negativo são maiores quando $k = 1$ e decresce para valores $k > 1$. A situação contrária ocorre na especificidade e no valor preditivo positivo, pois os maiores valores estão associados aos maiores valores de k.

A maior acurácia e menor custo relativo estão em $k = 20$, em um modelo com alta especificidade e baixa sensibilidade. O coeficiente de correlação atinge seu pico em $k = 9$ e é inferior ao encontrado na combinação com $k = 20$.

Note que a curva do custo relativo segue decrescente, ao passo que a acurácia é crescente, e tendem a se estabilizar depois de $k = 13$, aproximadamente.

A Figura 5.2 mostra os resultados obtidos pelas combinação via média, em que obtivemos resultados relativamente mais estáveis. A sensibilidade aumentou junto com α e a especificidade diminuiu. As demais medidas ficaram relativamente estáveis, com pouca variação. O menor custo relativo é apontado pela combinação via mínimo, no entanto, possui o menor MCC e sensibilidade.

Nos valores positivos de α encontramos os melhores valores para o MCC, sendo seu máximo em $\alpha = 4$, juntamente com a melhor sensibilidade.

Diante desses resultados, tomamos os dois melhores valores de k, 7 e 20, e de α, 4 e 5, e comparamos com o modelo usual e a combinação via regressão logística. A Figura 5.3 mostra os resultados obtidos.

A combinação via regressão logística apresenta resultado similar às outras duas combinações. A influência do intercepto apenas translada os escores, de forma que não afeta a classificação final, pois o que im-

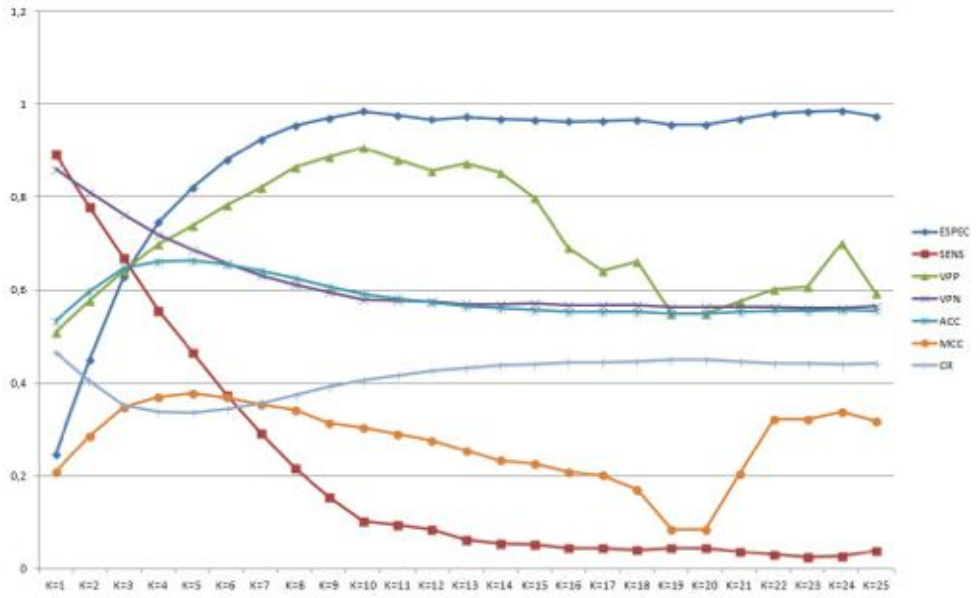

Figura 5.1: Combinações via votos - *German Credit Data*.

porta realmente é a ordem dos escores. No entanto, o fato de não usar intercepto pode levar a alterações nos outros parâmetros estimados na combinação, o que justifica as pequenas diferenças entre os modelos.

O menor custo relativo está na combinação por voto com $k = 20$, entretanto, simultaneamente apresenta os menores valores de MCC e sensibilidade (menores também que o modelo sem combinação alguma). As combinações via regressão logística apresentaram os melhores valores para a correlação e o segundo melhor resultado em relação ao custo relativo, acurácia e especificidade.

Através dos resultados obtidos na análise notamos que houve um aumento considerado no desempenho do modelo com combinação via regressão logística. Essa combinação obteve os melhores resultados para a acurácia, MCC e custo relativo. A variação dos valores de k e α como parâmetros de calibração da combinação é bastante eficaz e podem trazer melhorias em relação às combinações usuais.

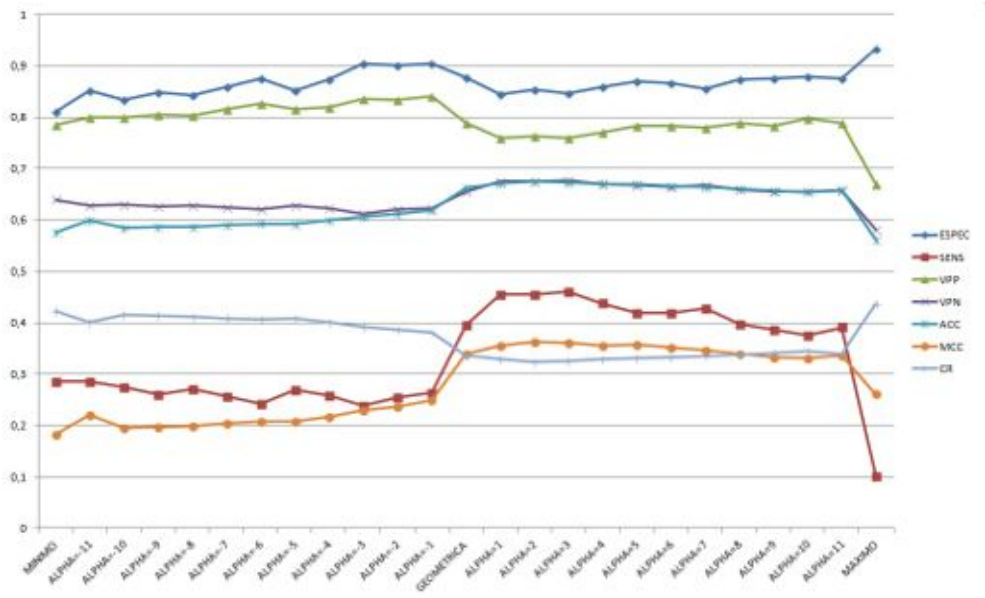

Figura 5.2: Combinações via médias - *German Credit Data*.

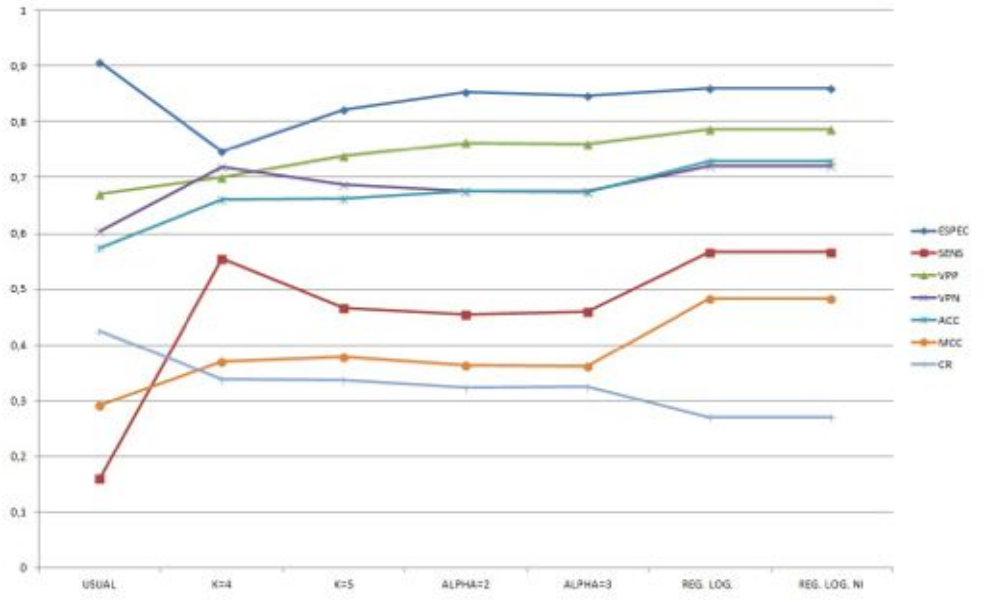

Figura 5.3: Comparação entre os melhores modelos - *German Credit Data*.

Capítulo 6

Análise de Sobrevivência

Do ponto de vista dos gestores do crédito, o questionamento básico à concessão consiste em saber qual a propensão à inadimplência do cliente. Considerando a modelagem apresentada até o momento neste livro, a resposta a essa pergunta, vem dos modelos de classificação direcionados na determinação do escore de crédito, correspondendo à chance do cliente estar ou não propenso à inadimplência.

A questão básica aqui é a pontualidade da modelagem, atribuída à simplificação da real resposta a uma determinada concessão de crédito. Na verdade, a partir da entrada do cliente na base, antes mesmo do final do período de desempenho, este pode tornar-se *mau* pagador e a resposta à concessão do crédito é obtida, ou seja, temos o verdadeiro momento da resposta do cliente à concessão. Entretanto, baseados no planejamento amostral usual descrito no Capítulo 1, utilizado para o desenvolvimento da modelagem de *Credit Scoring*, esperamos até o final do período de desempenho para, então, indicar se o desempenho do cliente foi *bom* ou *mau* por meio de uma variável dicotômica 0 ou 1. Isto é, simplificamos a resposta. Apesar de termos o instante da ocorrência da resposta (no nosso caso, negativa) do cliente à concessão do crédito desde a sua entrada na base, este momento é ignorado, em detrimento de sua transformação simplificadora a uma resposta dicotômica passível de ser acomodada por técnicas usuais de modelagem de *Credit Scoring*. É o que podemos chamar de representação discreta do risco de crédito do cliente. Entretanto, o que não podemos esquecer é que, apesar dos pontos

de contato do cliente com a empresa serem discretos (pontuais), o relacionamento cliente-empresa é contínuo a partir de sua entrada na base. Assim, intuitivamente, é natural pensarmos em adaptar a técnica de modelagem à uma resposta temporal do cliente à concessão, direcionando os procedimentos estatísticos a uma visão contínua do relacionamento cliente-empresa, ao invés de simplificar a resposta do cliente relacionada à concessão do crédito, adequando-a às técnicas usuais de modelagem. É o que chamamos de modelagem temporal de *Credit Scoring*. Assim, consideramos uma metodologia conhecida por análise de sobrevivência.

6.1 Algumas Definições Usuais

A análise de sobrevivência consiste em uma coleção de procedimentos estatísticos para a análise de dados relacionados ao tempo decorrido desde um tempo inicial, pré-estabelecido, até a ocorrência de um evento de interesse. No contexto de *Credit Scoring*, o tempo relevante é o medido entre o ingresso do cliente na base de usuários de um produto de crédito até a ocorrência de um evento de interesse, como por exemplo, um problema de inadimplência.

As principais características das técnicas de análise de sobrevivência são sua capacidade de extrair informações de *dados censurados*, ou seja, daqueles clientes para os quais, no final do acompanhamento no período de desempenho, o problema de crédito não foi observado, além de levar em consideração os tempos para a ocorrência dos eventos. De maneira geral, um tempo censurado corresponde ao tempo decorrido entre o início e o término do estudo ou acompanhamento de um indivíduo sem ser observada a ocorrência do evento de interesse para ele.

Na análise de sobrevivência, o comportamento da variável aleatória tempo de sobrevida, $T \geq 0$, pode se expresso por meio de várias funções matematicamente equivalentes, tais que, se uma delas é especificada, as outras podem ser derivadas. Essas funções são: a função densidade de probabilidade, $f(t)$, a função de sobrevivência, $S(t)$, e a função de risco, $h(t)$, que são descritas com mais detalhes a seguir. Essas três funções são utilizadas na prática para descrever diferentes aspectos apresentados pelo conjunto de dados.

A função densidade é definida como o limite da probabilidade de observar o evento de interesse em um indivíduo no intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo, podendo ser expressa por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{6.1}$$

A função de sobrevivência é uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência. Tal função é definida como a probabilidade de não ser observado o evento de interesse para um indivíduo até um certo tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t sem o evento. Em termos probabilísticos esta função é dada por

$$S(t) = P(T > t) = 1 - F(t), \tag{6.2}$$

tal que $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$ e $F(t) = \int_0^t f(u)\, du$ representa a função de distribuição acumulada.

A função de risco, ou taxa de falha, é definida como o limite da probabilidade de ser observado o evento de interesse para um indivíduo no intervalo de tempo $[t, t + \Delta t]$ dado que o mesmo tenha sobrevivido até o tempo t, e expressa por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

Esta função também pode ser definida em termos de (6.1) e (6.2) por meio da expressão

$$h(t) = \frac{f(t)}{S(t)}, \tag{6.3}$$

descrevendo assim o relacionamento entre as três funções que geralmente são utilizadas para representar o comportamento dos tempos de sobrevivência.

Devido a sua interpretação, a função de risco é muitas vezes utilizada para descrever o comportamento dos tempos de sobrevivência. Essa função descreve como a probabilidade instantânea de falha, ou taxa de falha, se modifica com o passar do tempo, sendo conhecida também como

taxa de falha instantânea, força de mortalidade e taxa de mortalidade condicional (Cox & Oakes, 1994).

Como visto, as funções densidade de probabilidade, de sobrevivência e de risco são matematicamente equivalentes. Algumas relações básicas podem ser utilizadas na obtenção de uma destas funções quando uma delas é especificada, além da expressão que relaciona essas três funções descritas em (6.3).

A função densidade de probabilidade é definida como a derivada da função densidade de probabilidade acumulada utilizada em (6.1), isto é

$$f(t) = \frac{\partial F(t)}{\partial t}.$$

Como $F(t) = 1 - S(t)$ pode-se escrever

$$f(t) = \frac{\partial \left[1 - S(t)\right]}{\partial t} = -S'(t). \tag{6.4}$$

Substituindo (6.4) em (6.3) obtemos

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

Dessa forma temos

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

ou seja,

$$S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{6.5}$$

Uma outra função importante é a de risco acumulada, definida como

$$H(t) = \int_0^t h(u)du. \tag{6.6}$$

Substituindo (6.6) em (6.5) temos que

$$S(t) = \exp\left[-H(t)\right]. \tag{6.7}$$

Como $lim_{t\to\infty}S(\infty) = 0$ então

$$\lim_{t\to\infty} H(t) = \infty.$$

Além disso, de (6.3)

$$f(t) = h(t)S(t). \tag{6.8}$$

Substituindo (6.7) em (6.8) temos

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

Portanto, mostramos as relações entre as três funções utilizadas para descrever os dados em análise de sobrevivência.

Similar à regressão logística, é comum, em dados de análise de sobrevivência, a presença de covariáveis representando também a heterogeneidade da população. Assim, os modelos de regressão em análise de sobrevivência têm como objetivo identificar a relação e a influência dessas variáveis com os tempos de sobrevida, ou com alguma função dos mesmos. Desta forma, Cox (1972) propôs o seguinte modelo

$$h(t;\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})h_0(t),$$

em que $\boldsymbol{\beta}$ é o vetor dos parâmetros $(\beta_1, \beta_2, \ldots, \beta_p)$ para cada uma das p covariáveis disponíveis e $h_0(t)$ é uma função não-conhecida que reflete, na área financeira, o risco básico de inadimplência inerente a cada cliente.

A Figura 6.1 ilustra a diferença entre as respostas observadas por uma metodologia pontual, no caso, regressão logística, e a análise de sobrevivência.

Sabendo que a razão de risco (*Hazard Ratio*) tem interpretação análoga ao *odds ratio*, temos que os resultados fornecidos pelo modelo de Cox são muito parecidos com os resultados da regressão logística, em que as mesmas variáveis originais foram selecionadas para compor o modelo final, diferenciando apenas as categorias (*dummies*) que foram escolhidas.

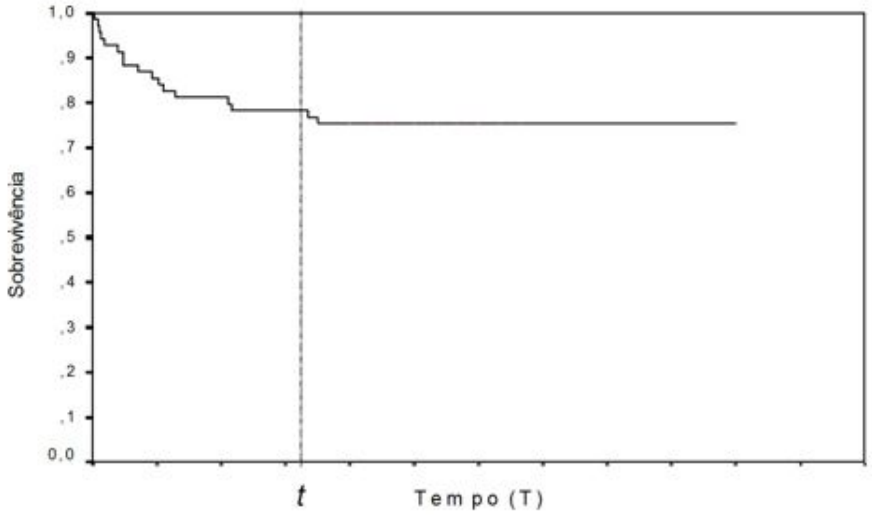

Figura 6.1: Informações - regressão logística e análise de sobrevivência.

6.2 Modelo de Cox

Em análise de sobrevivência buscamos explorar e conhecer a relação entre o tempo de sobrevivência e uma ou mais covariáveis disponíveis.

Na modelagem de análise de sobrevivência é comum o interesse no risco da ocorrência de um evento em um determinado tempo, após o início de um estudo ou acompanhamento de um cliente. Este tempo pode coincidir ou não com o início do relacionamento do cliente com a empresa ou quando se inicia a utilização de um determinado serviço de crédito, por exemplo. Esses modelos diferem dos modelos aplicados em análise de regressão e em planejamento de experimentos, nos quais a média da variável resposta ou alguma função dela é modelada por meio de covariáveis.

Um dos principais objetivos ao se modelar a função de risco é determinar potenciais covariáveis que influenciam na sua forma. Outro importante objetivo é mensurar o risco individual de cada cliente. Além do interesse específico na função de risco, é de interesse estimar, para cada cliente, a função de sobrevivência.

Um modelo clássico para dados de sobrevivência, proposto por

Cox (1972), é o de riscos proporcionais, também conhecido como modelo de regressão de Cox. Este modelo baseia-se na suposição de proporcionalidade dos riscos, para diferentes perfis de clientes, sem a necessidade de assumir uma distribuição de probabilidade para os tempos de sobrevida. Por isso, é dito ser um modelo semi-paramétrico.

6.2.1 Modelo para comparação de dois perfis de clientes

Suponha que duas estratégias ("P"- padrão e "A" - alternativa) são utilizadas para a concessão de crédito aos clientes de uma determinada empresa. Sejam $h_P(t)$ e $h_A(t)$ os riscos de crédito no tempo t para os clientes das duas estratégias, respectivamente. De acordo com o modelo de riscos proporcionais, o risco de crédito para os clientes da estratégia padrão ("P") no instante t é proporcional ao risco dos clientes da estratégia alternativa ("A") no mesmo instante. O modelo de riscos proporcionais pode ser expresso como

$$h_A(t) = \psi h_P(t), \tag{6.9}$$

para qualquer valor de t, $t > 0$, no qual ψ é uma constante. A suposição de proporcionalidade implica que a verdadeira função de sobrevivência para os indivíduos atendidos pelas duas estratégias não se cruzam no decorrer do tempo.

Suponha que o valor de ψ seja a razão entre o risco (*hazard risk*) de crédito de um cliente, para o qual foi concedido um produto de crédito pela estratégia alternativa, e o risco de crédito de um cliente pela estratégia padrão, em um determinado tempo t. Se $\psi < 1$, o risco de crédito no instante t é menor para um indivíduo que recebeu o produto de crédito pela estratégia alternativa em relação ao padrão, evidenciando, assim, melhores resultados do risco de crédito da estratégia alternativa. Por outro lado, um valor $\psi > 1$ indica um risco de crédito maior para o cliente conquistado pela estratégia alternativa.

O modelo (6.9) pode ser generalizado escrevendo-o de uma outra forma. Denotando $h_0(t)$ como a função de risco para o qual foi concedido

o crédito pela estratégia padrão, a função de risco para os clientes da estratégia alternativa é dado por $\psi h_0(t)$. Como a razão de risco ψ não pode ser negativa, é conveniente considerar $\psi = \exp(\beta)$. Desta forma, o parâmetro β é o logaritmo da razão de risco, $\beta = \log(\psi)$, e os valores de β pertencem ao intervalo $(-\infty, +\infty)$, fornecendo, assim, valores positivos de ψ. Observe que valores positivos de β ocorrem se a razão de risco, ψ, for maior que 1, isto é, quando a forma alternativa de risco é pior que a padrão, e o contrário quando os valores de β forem negativos.

Seja X uma variável indicadora, a qual assume o valor zero, se o produto de crédito foi concedido a um indivíduo pela estratégia padrão, e um, no caso da estratégia alternativa. Se x_i é o valor de X para o i-ésimo cliente na amostra, a função de risco de crédito, $h_i(t)$, $i = 1, \ldots, n$, para esse indivíduo pode ser escrita da seguinte forma

$$h_i(t) = \exp\{\beta x_i\} h_0(t). \tag{6.10}$$

Este é o modelo de riscos proporcionais para a comparação de dois grupos de indivíduos com características distintas.

6.2.2 A generalização do modelo de riscos proporcionais

O modelo (6.10) é generalizado para a situação na qual o risco de crédito do cliente ou o risco de abandono do cliente, no caso de um problema de *marketing*, em um determinado tempo depende dos valores de p covariáveis $x_1, x_2, \ldots, x_p$.

Seja $h_0(t)$ a função de risco de crédito de um cliente para o qual os valores de todas as covariáveis são iguais a zero. A função $h_0(t)$ é chamada de função de risco básica. A função de risco para o i-ésimo indivíduo pode ser escrita como

$$h_i(t) = \psi(\boldsymbol{x}_i) h_0(t),$$

em que $\psi(\boldsymbol{x}_i)$ é uma função dos valores do vetor de covariáveis, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$, para o i-ésimo cliente da amostra. A função $\psi(\cdot)$ pode ser interpretada como a razão entre o risco de crédito no instante t para

um cliente cujo vetor de covariáveis é $\boldsymbol{x}_i$ e o risco de crédito de um cliente que possui todas as covariáveis com valores iguais a zero, ou seja, $\boldsymbol{x}_i = \mathbf{0}$.

É conveniente escrever a razão de risco, $\psi(\boldsymbol{x}_i)$, como $\exp(\eta_i)$, sendo η_i

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ji}.$$

Desta forma, o modelo de riscos proporcionais geral tem a forma

$$h_i(t) = \exp\{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\} h_0(t). \tag{6.11}$$

Em notação matricial, $\eta_i = \boldsymbol{\beta}'\boldsymbol{x}_i$, na qual $\boldsymbol{\beta}$ é o vetor de coeficientes das covariáveis $x_1, x_2, \ldots, x_p$. O valor η_i é chamado de componente linear do modelo, sendo conhecido também como escore de risco para o i-ésimo indivíduo. A expressão (6.11) pode ser reescrita como

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} = \boldsymbol{\beta}'\boldsymbol{x}_i.$$

A constante β_0, presente em outros modelos lineares, não aparece em (6.11). Isto ocorre devido a presença do componente não-paramétrico no modelo que absorve este termo constante.

O modelo de riscos proporcionais pode também ser escrito como um modelo linear para o logaritmo da razão de risco. Existem outras formas propostas na literatura, sendo $\psi(\boldsymbol{x}_i) = \psi(\exp(\boldsymbol{\beta}'\boldsymbol{x}_i))$ a mais comum utilizada em problemas de análise de sobrevivência. De uma forma geral, o modelo de riscos proporcionais pode ser escrito como (Colosimo & Giolo, 2006)

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

sendo g uma função especificada, tal que $g(0) = 1$. Observe que este modelo é composto pelo produto de duas componentes, uma não-paramétrica e outra paramétrica. Para a componente não-paramétrica, $h_0(t)$, não é necessário assumir uma forma pré-estabelecida, porém esta função deve ser não-negativa no tempo. A componente paramétrica é geralmente assumida na forma exponencial. Devido a composição não-paramétrica e paramétrica, este modelo é dito ser semi-paramétrico, não sendo ne-

cessário supor uma forma para a distribuição dos tempos de sobrevivência.

6.2.3 Ajuste de um modelo de riscos proporcionais

Dado um conjunto de dados de sobrevivência, o ajuste do modelo (6.11) envolve a estimação dos coeficientes $\beta_1, \beta_2, \ldots, \beta_p$. Em algumas situações é, também, necessário a estimação da função de risco básica $h_0(t)$. Os coeficientes e a função de risco podem ser estimados separadamente. Iniciamos estimando os parâmetros $\beta_1, \beta_2, \ldots, \beta_p$, usando, por exemplo, o método da máxima verossimilhança e, em seguida, estimamos a função de risco básica. Assim, as inferências sobre os efeitos das p covariáveis na razão de risco, $h_i(t)/h_0(t)$, podem ser realizadas sem a necessidade de se obter uma estimativa para $h_0(t)$.

Suponha que os tempos de sobrevida de n indivíduos estejam disponíveis e que existam r tempos distintos em que foram observadas a ocorrência de pelo menos um evento de interesse de clientes que estavam sob risco nesses instantes e $n-r$ tempos de sobrevida censurados, para os quais não foram observados o evento de interesse, permanecendo assim com seus pagamentos em dia com a empresa até o instante que se tem a última informação desses clientes. O evento de interesse aqui poderia ser, por exemplo, a inadimplência. Assumimos que o evento de interesse ocorra apenas para um indivíduo em cada um dos tempos de sobrevida observado, não havendo assim a presença de empate. Os r tempos, para os quais foram observados o evento de interesse, serão denotados por $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, sendo $t_{(j)}$ o j-ésimo tempo ordenado. O conjunto de clientes que estão sob risco de crédito, no instante $t_{(j)}$, o conjunto de risco, será denotado por $R(t_{(j)})$.

Cox (1972) propôs uma função de verossimilhança para o modelo de riscos proporcionais, representada pela equação (6.11), dada por

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}, \tag{6.12}$$

na qual $\boldsymbol{x}_{(j)}$ é o vetor de covariáveis de um cliente em que o evento de interesse, inadimplência, foi observado no j-ésimo tempo de sobrevida

$t_{(j)}$. O somatório no denominador da função de verossimilhança considera apenas os valores de $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ para todos os indivíduos que estão sob risco de crédito no instante $t_{(j)}$. Note que o produtório considera apenas os clientes para os quais o evento de interesse foi observado. Além disso, observe que os clientes com tempos de sobrevida censurados não contribuem no numerador da função de verossimilhança, porém, fazem parte do somatório do conjunto sob risco de crédito em cada um dos tempos que ocorreram eventos. A função de verossimilhança depende somente da ordem dos tempos em que ocorreram os eventos de interesse, uma vez que, isso define o conjunto de risco em cada um dos tempos. Consequentemente, inferências sobre os efeitos das covariáveis na função de risco dependem somente da ordem dos tempos de sobrevivência.

Considere $t_i, i = 1, 2, \ldots, n$ os tempos de sobrevida observados e δ_i uma variável indicadora de censura assumindo valor zero, se o i-ésimo tempo $t_i, i = 1, 2, \ldots, n$, é uma censura, e um, na situação em que o evento de interesse foi observado no tempo considerado.

A função de verossimilhança em (6.12) pode ser expressa da seguinte forma

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right]^{\delta_i},$$

O logaritmo desta função de máxima verossimilhança é dado por

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}'\boldsymbol{x}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l) \right\}. \tag{6.13}$$

As estimativas de máxima verossimilhança dos parâmetros β's são obtidos maximizando-se (6.13), ou seja, resolvendo o sistema de equações definido por $U(\boldsymbol{\beta}) = \mathbf{0}$, em que $U(\boldsymbol{\beta})$ é o vetor escore formado pelas primeiras derivadas da função $l(\boldsymbol{\beta})$, ou seja,

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{l \in R(t_i)} x_l \exp(\boldsymbol{x}_l\boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{x}_l\boldsymbol{\beta})} \right] = \mathbf{0}.$$

O estimador de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, é obtido através do método de Newton-Raphson.

O estimador da matriz de variâncias-covariâncias, $\widehat{Var}(\hat{\boldsymbol{\beta}})$, dos coeficientes estimados $\hat{\boldsymbol{\beta}}$ são obtidos usando a teoria assintótica dos estimadores de máxima verossimilhança (Hosmer & Lemeshow, 1999). Estes estimadores são dados por

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = I(\hat{\boldsymbol{\beta}})^{-1}, \tag{6.14}$$

na qual $I(\hat{\boldsymbol{\beta}})$ é a informação de Fisher observada, expressa por

$$I(\hat{\boldsymbol{\beta}}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

e

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial^2 \boldsymbol{\beta}^2} = -\sum_{i=1}^{r}\left\{\frac{[\sum_l \exp(\boldsymbol{x}_l\boldsymbol{\beta})]\,[\sum_l x_l^2 \exp(\boldsymbol{x}_l\boldsymbol{\beta})] - [\sum_l x_l \exp(\boldsymbol{x}_l\boldsymbol{\beta})]^2}{\sum_l \exp(\boldsymbol{x}_l\boldsymbol{\beta})}\right\}.$$

com l pertencendo ao conjunto de risco $R(t_i)$.

Os estimadores dos erros-padrão, denotado por $\widehat{\mathrm{EP}}(\hat{\boldsymbol{\beta}})$, são dados pela raiz quadrada dos elementos da diagonal principal da matriz apresentada em (6.14).

Os detalhes para a construção da função de verossimilhança parcial de Cox, apresentada em (6.12), e alguns possíveis tratamentos para as situações em que percebemos ocorrências de empates nos tempos de sobrevida observados são descritos na Subseção 6.2.4.

O argumento básico utilizado na construção da função de verossimilhança para o modelo de riscos proporcionais é que intervalos entre tempos de eventos sucessivos não fornecem informações nos valores dos parâmetros $\boldsymbol{\beta}$. Dessa forma, no contexto utilizado, considera-se a probabilidade condicional de que o i-ésimo cliente da amostra tenha um problema de crédito em algum tempo $t_{(j)}$ dado que um problema ocorre nesse instante, sendo $t_{(j)}$ um dos r tempos, $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, onde os eventos foram observados. Se o vetor de covariáveis para o indivíduo que

abandonou no tempo $t_{(j)}$ é $\boldsymbol{x}_{(j)}$, temos

$$P[\text{indivíduo com } \boldsymbol{x}_{(j)} \text{ abandonar no instante } t_{(j)} \mid \text{um abandono ocorre no instante } t_{(j)}]$$

$$= \frac{P[\text{ indivíduo com } \boldsymbol{x}_{(j)} \text{ abandonar no instante } t_{(j)}]}{P[\text{ um abandono ocorrer no instante } t_{(j)}]}. \quad (6.15)$$

O numerador da expressão acima corresponde ao risco de crédito no instante $t_{(j)}$ para um indivíduo para o qual o vetor de covariáveis é dado por $\boldsymbol{x}_{(j)}$. Se o evento de interesse ocorre no instante $t_{(j)}$ para o i-ésimo cliente da amostra, a função de risco de crédito pode ser denotada como $h_i(t_{(j)})$. O denominador compreende a soma dos riscos de crédito no momento $t_{(j)}$ para todos os indivíduos que estão com seus pagamentos em dia até aquele instante, estando, portanto, sob risco de ser observado o evento de interesse. Este somatório considera os valores $h_l(t_{(j)})$ para todos os indivíduos indexados por l no conjunto de risco no instante $t_{(j)}$, denotado por $R(t_{(j)})$. Consequentemente, a probabilidade condicional na expressão (6.15) pode ser escrita como

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})},$$

e utilizando a equação (6.11), a função de risco básica, $h_0(t_{(j)})$, no numerador e denominador são canceladas resultando na seguinte expressão

$$\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)},$$

e, finalmente, fazendo o produto dessa probabilidade condicional para os r tempos nos quais foram observados o evento de interesse, obtemos a função de verossimilhança, apresentada na equação (6.12).

A função de verossimilhança obtida para o modelo de riscos proporcionais não é, na realidade, uma verdadeira verossimilhança, uma vez que não utiliza diretamente os verdadeiros tempos de sobrevida dos clientes censurados ou não-censurados; por essa razão, é referida como função

de verossimilhança parcial.

Com o objetivo de tornar mais clara a construção da função de verossimilhança parcial do modelo de riscos proporcionais, considere uma amostra com informações dos tempos de sobrevida de cinco clientes, que estão representados na Figura 6.2. Para os indivíduos 2 e 5 não ocorreu o evento de interesse, ou seja, até o instante $t_{(3)}$ estes clientes estão com seus pagamentos em dia com a empresa. Os três tempos para os quais foram observados a inadimplência dos clientes são denotados por $t_{(1)} < t_{(2)} < t_{(3)}$. Assim, $t_{(1)}$ é o tempo de sobrevida do cliente 3, $t_{(2)}$ é o tempo para o cliente 1 e $t_{(3)}$ para o cliente 4.

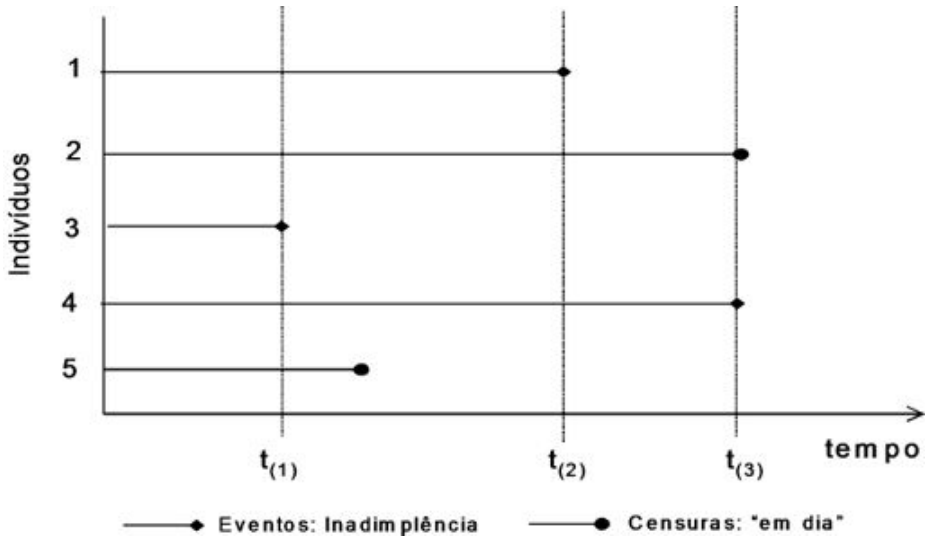

Figura 6.2: Tempos de sobrevida para cinco indivíduos.

O conjunto de risco de cada um dos três tempos, nos quais foram observados o evento de interesse, consiste nos clientes que permaneceram com seus pagamentos em dia até cada um dos instantes. Assim, o conjunto de risco $R(t_{(1)})$ compreende todos os cinco clientes, o conjunto de risco $R(t_{(2)})$ os clientes 1, 2 e 4, e o conjunto de risco $R(t_{(3)})$ somente os indivíduos 2 e 4. Seja $\psi(i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), i = 1, 2, \ldots, 5$, em que $\boldsymbol{x}_i$ é um vetor coluna de covariáveis. Os termos do numerador da função de verossimilhança para os tempos $t_{(1)}, t_{(2)}$ e $t_{(3)}$, são respectivamente $\psi(3)$, $\psi(1)$ e $\psi(4)$, uma vez que os clientes 3, 1 e 4 apresentaram problema de crédito nos respectivos tempos ordenados. Dessa forma, a função de

verossimilhança parcial é dada pela seguinte expressão

$$\left(\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)\left(\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(4)}\right)\left(\frac{\psi(4)}{\psi(2)+\psi(4)}\right).$$

Quando ocorrem empates entre eventos e censuras, como em $t_{(3)}$, utilizamos, por convenção, que as censuras ocorreram após o evento, definindo, assim, quais os indivíduos que fazem parte do conjunto de risco em cada um dos tempos e que foram observados os eventos.

6.2.4 Tratamento de empates

O modelo de riscos proporcionais assume que a função de risco é contínua e, sob essa suposição, empates dos tempos de sobrevivência não são possíveis. Porém, o processo de obtenção das informações dos tempos de sobrevivência, muitas vezes, registra ou o dia, ou o mês ou o ano mais próximo da ocorrência do evento. Empates, nesses tempos, podem ocorrer por esse processo de arredondamento ou aproximação dos tempos, sendo observado assim, a ocorrência de mais do que um evento em um mesmo instante de tempo.

Além da ocorrência de mais que um evento em um mesmo instante, existe, também, a possibilidade da ocorrência de empates entre uma ou mais observações censuradas em um instante de tempo em que também foi observado um evento. Assim, é possível ocorrer mais do que uma censura no mesmo instante de tempo em que ocorre um evento. Nessa última situação adota-se que os eventos ocorrem antes das censuras, não gerando maiores dificuldades na construção da função de verossimilhança parcial. O mesmo não ocorre na situação anterior, quando existe a presença de empates entre eventos.

A função de verossimilhança exata na presença de empates entre os eventos foi proposta por Kalbfleisch & Prentice (1980) e inclui todas as possíveis ordens dos eventos empatados, exigindo, consequentemente, muito esforço computacional, principalmente quando um número grande de empates é verificado em um ou mais dos tempos em que se observa a ocorrência do evento.

Em uma situação com 5 eventos, ocorrendo em um mesmo ins-

tante, existem 120 possíveis ordens a serem consideradas; para 10 eventos empatados, esse valor ficaria acima de 3 milhões (Allison, 1995). Algumas aproximações para a função de verossimilhança parcial foram desenvolvidas e trazem vantagens computacionais sobre o método exato.

Seja $\boldsymbol{s}_j$ o vetor que contém a soma de cada uma das p covariáveis para os indivíduos nos quais foram observados o evento no j-ésimo tempo, $t_{(j)}$, $j = 1, 2, \ldots, r$. O número de eventos no instante $t_{(j)}$ é denotado por d_j. O h-ésimo elemento de $\boldsymbol{s}_j$ é dado por $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, em que x_{hjk} é o valor da h-ésima covariável, $h = 1, 2, \ldots, p$, para o k-ésimo dos d_j indivíduos, $k = 1, 2, \ldots, d_j$, para os quais foram observados o evento no j-ésimo tempo, $j = 1, 2, \ldots, r$.

A aproximação proposta por Peto (1972) e Breslow (1974) é a mais simples e considera a seguinte função de verossimilhança parcial

$$L_B(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right]^{d_j}}. \tag{6.16}$$

Nesta aproximação, os d_j eventos de interesse, clientes que se tornaram inadimplentes, por exemplo, observados em $t_{(j)}$, são considerados distintos e ocorrem sequencialmente. Esta verossimilhança pode ser diretamente calculada e é adequada quando o número de observações empatadas, em qualquer tempo em que ocorrem os eventos, não é muito grande. Por isso, esse método está normalmente implementado nos módulos de análise de sobrevivência dos *softwares* estatísticos. Farewell & Prentice (1980) mostram que os resultados dessa aproximação deterioram quando a proporção de empates aumenta em relação ao número de indivíduos sob risco, em alguns dos tempos em que os eventos são observados.

Efron (1977) propõe a seguinte aproximação para a verossimilhança parcial do modelo de riscos proporcionais

$$L_E(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{s}_l) - (k-1)d_k^{-1} \sum_{l \in D(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right]^{d_j}}, \tag{6.17}$$

em que $D(t_{(j)})$ é o conjunto de todos os clientes para os quais foram observados o evento de interesse no instante $t_{(j)}$. Este método fornece

resultados mais próximos do exato do que o de Breslow.

Cox (1972) sugeriu a aproximação

$$L_C(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\sum_{l \in R(t_{(j)};d_j)} \exp(\boldsymbol{\beta}'\boldsymbol{s}_l)}, \tag{6.18}$$

em que $R(t_{(j)}; d_j)$ denota um conjunto de d_j indivíduos retirados do conjunto de risco no instante $t_{(j)}$. O somatório no denominador corresponde a todos os possíveis conjuntos de d_j indivíduos retirados do conjunto de risco $R(t_{(j)})$. A aproximação da expressão (6.18) é baseada no modelo para a situação em que a escala de tempo é discreta, permitindo assim a presença de empates. A função de risco para um indivíduo, com vetor de covariáveis $\boldsymbol{x}_i$, $h_i(t; \boldsymbol{x})$, é interpretada como a probabilidade de abandono em um intervalo de tempo unitário $(t, t+1)$, dado que esse indivíduo estava sob risco até o instante t, ou seja,

$$h_i(t) = P(t \leqslant T < t+1 \mid T \geqslant t), \tag{6.19}$$

sendo T uma variável aleatória que representa o tempo de sobrevivência. A versão discreta do modelo de riscos proporcionais na equação (6.11) é

$$\frac{h_i(t; \boldsymbol{x}_i)}{1 - h(t; \boldsymbol{x}_i)} = \frac{h_0(t)}{1 - h_0(t)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_i), \tag{6.20}$$

para o qual a função de verossimilhança é dada pela equação (6.18). Na situação limite, quando o intervalo de tempo discreto tende a zero, esse modelo tende ao modelo de riscos proporcionais da equação (6.11).

Para mostrar que (6.20) é reduzido a (6.11), quando o tempo é contínuo, temos que a função de risco discreta, em (6.20), quando o valor unitário é substituído por δt, é dada por

$$h(t)\delta t = P(t \leqslant T < t + \delta t \mid T \geqslant t),$$

e, assim, a equação obtida a partir de (6.20) é dada por

$$\frac{h(t; \boldsymbol{x}_i)\delta t}{1 - h(t; \boldsymbol{x}_i)\delta t} = \frac{h_0(t)\delta t}{1 - h_0(t)\delta t} \exp(\boldsymbol{\beta}'\boldsymbol{x}_i),$$

e tomando o limite quando o intervalo de tempo δt tende a zero é obtida a equação (6.11).

Quando não existem empates em um conjunto de dados de análise de sobrevivência, ou seja, quando $d_j = 1$, $j = 1, 2, \ldots, r$, as aproximações nas equações (6.16), (6.17) e (6.18), são reduzidas a função de verossimilhança parcial da equação (6.12).

6.3 Intervalos de Confiança e Seleção de Variáveis

Com as estimativas dos parâmetros e os respectivos erros-padrão, $\text{EP}(\hat{\boldsymbol{\beta}})$, construímos os intervalos de confiança dos elementos do vetor de parâmetros $\boldsymbol{\beta}$.

Um intervalo de $100(1-\alpha)\%$ de confiança para um determinado parâmetro β_j é obtido fazendo $\hat{\beta}_j \pm Z_{\alpha/2}\ \text{EP}(\hat{\beta}_j)$, em que $\hat{\beta}_j$ é o valor da estimativa de máxima verossimilhança do j-ésimo parâmetro e $Z_{\alpha/2}$ o percentil superior $\alpha/2$ de uma distribuição normal padrão.

Se um intervalo de $100(1-\alpha)\%$ para β_j não inclui o valor zero, dizemos que há evidências de que o valor real de β_j é estatisticamente diferente de zero. A hipótese nula $H_0 : \beta_j = 0$ pode ser testada calculando o valor da estatística $\hat{\beta}_j/\text{EP}(\hat{\beta}_j)$. Esta estatística tem, assintoticamente, distribuição normal padrão.

Geralmente, as estimativas individuais $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_p$, em um modelo de riscos proporcionais não são todas independentes entre si. Isso significa que testar hipóteses separadamente pode não ser facilmente interpretável.

Uma forma de seleção de variáveis utilizada na análise de sobrevivência na presença de um grande número de potenciais covariáveis é o método *stepwise*, conjuntamente com a experiência de especialistas da área e o bom senso na interpretação dos parâmetros.

6.4 Estimação da Função de Risco e Sobrevivência

Nas seções anteriores consideramos procedimentos para a estimação do vetor de parâmetros $\boldsymbol{\beta}$ do componente linear do modelo de riscos proporcionais. Uma vez ajustado o modelo, a função de risco e a correspondente função de sobrevivência podem, se necessário, ser estimadas.

Suponha que o escore de risco de um modelo de riscos proporcionais contém p covariáveis $x_1, x_2, \ldots, x_p$ com as respectivas estimativas para seus coeficientes $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. A função de risco para o i-ésimo indivíduo no estudo é dada por

$$\hat{h}_i(t) = \exp\{\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i\}\hat{h}_0(t), \tag{6.21}$$

em que $\boldsymbol{x}_i$ é o vetor dos valores observados das p covariáveis para o i-ésimo indivíduo, $i = 1, 2, \ldots, n$, e $\hat{h}_0(t)$ é a estimativa para a função de risco básica. Por meio da equação (6.21), a função de risco pode ser estimada para um indivíduo, após a função de risco básica ter sido estimada.

Em um problema de *Credit Scoring*, a utilização do escore de risco do modelo de Cox como escore final é uma opção bastante viável de ser utilizada, uma vez que a partir desses valores uma ordenação dos clientes pode ser obtida com relação ao risco de crédito.

Uma estimativa da função de risco básica foi proposta por Kalbfleisch & Prentice (1973) utilizando uma metodologia baseada no método de máxima verossimilhança. Suponha que foram observados r tempos de sobrevida distintos dos clientes que se tornaram inadimplentes, os quais, ordenados, são denotados $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, existindo d_j eventos e n_j clientes sob risco no instante $t_{(j)}$. A estimativa da função de risco básica no tempo $t_{(j)}$ é dada por

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\xi}_j,$$

sendo $\hat{\xi}_j$ a solução da equação

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\beta}' \boldsymbol{x}_l)}{1 - \hat{\xi}_j^{\exp(\hat{\beta}' \boldsymbol{x}_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \boldsymbol{x}_l), \tag{6.22}$$

para $j = 1, 2, \ldots, r$, sendo $D(t_{(j)})$ o conjunto de todos os d_j indivíduos que em um problema de *Credit Scoring*, por exemplo, se tornaram inadimplentes no j-ésimo tempo, $t_{(j)}$, e $R(t_{(j)})$ representando os n_j indivíduos sob risco no mesmo instante $t_{(j)}$.

Na situação particular em que não ocorrem empates entre os tempos de sobrevida dos clientes, isto é, $d_j = 1$, $j = 1, 2, \ldots, r$, o lado esquerdo da equação (6.22) será um único termo. Assim, essa equação pode ser solucionada por

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\beta}' \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \boldsymbol{x}_l)}\right)^{\exp(-\hat{\beta}' \boldsymbol{x}_{(j)})},$$

em que $\boldsymbol{x}_{(j)}$ é o vetor das covariáveis para o único cliente para o qual foi observado o evento no instante $t_{(j)}$.

Quando o evento é observado para mais de um cliente em um mesmo instante de tempo, ou seja, $d_j > 1$ para algum j, o somatório do lado esquerdo da equação (6.22) compreende a soma de uma série de frações na qual $\hat{\xi}_j$ está no denominador elevado a diferente potências. Assim, a equação não pode ser solucionada explicitamente, e métodos iterativos são necessários.

A suposição de que o risco de ocorrência de eventos entre dois tempos consecutivos é constante, permite considerar $\hat{\xi}_j$ como uma estimativa da probabilidade de que não seja observado o evento de interesse no intervalo $t_{(j)}$ e $t_{(j+1)}$. A função de sobrevivência básica pode ser estimada por

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j,$$

para $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r-1$. A função de risco acumulada básica é dada por $H_0(t) = -\log S_0(t)$, e assim uma estimativa dessa

função é

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\xi}_j,$$

para $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r-1$.

As estimativas das funções de risco, sobrevivência e risco acumulado podem ser utilizadas para a obtenção de estimativas individuais para cada cliente através do vetor de covariáveis $\boldsymbol{x}_i$. Da equação (6.21), a função de risco é estimada por $\exp(\hat{\beta}'\boldsymbol{x}_i)\hat{h}_0(t)$. Integrando ambos os lados dessa equação temos

$$\int_0^t \hat{h}_i(u)du = \exp(\hat{\beta}'\boldsymbol{x}_i) \int_0^t \hat{h}_0(u)du,$$

de modo que a função de risco acumulada para o i-ésimo indivíduo é dada por

$$\hat{H}_i(t) = \exp(\hat{\beta}'\boldsymbol{x}_i)\hat{H}_0(t).$$

Assim, a função de sobrevivência para o i-ésimo indivíduo é dada por

$$\hat{S}_i(t) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}'\boldsymbol{x}_i)},$$

para $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r-1$. Uma vez estimada a função de sobrevivência, $\hat{S}_i(t)$, uma estimativa da função de risco acumulada é obtida automaticamente fazendo $-\log \hat{S}_i(t)$.

6.5 Interpretação dos Coeficientes

Quando o modelo de riscos proporcionais é utilizado, os coeficientes das covariáveis podem ser interpretados como o logaritmo da razão de risco (*hazard risk*) do evento de dois indivíduos com características diferentes para uma covariável específica. Dessa forma, o coeficiente de uma covariável específica é interpretado como o logaritmo da razão do risco do evento de um indivíduo, que assume determinado valor para esta covariável, em relação a outro indivíduo para o qual foi observado um outro valor que é assumido como referência.

As estimativas da razão de risco e seus respectivos intervalos de confiança são normalmente obtidos a partir do modelo múltiplo final ajustado. A interpretação dos parâmetros depende do tipo de covariável considerada, podendo ser contínua ou categórica.

Suponha um modelo de riscos proporcionais com apenas uma variável contínua x. A função de risco para o i-ésimo indivíduo para o qual $x = x_i$ é

$$h_i(t) = \exp(\hat{\beta}' x_i) h_0(t).$$

Considere a razão de risco entre dois indivíduos i e j, os quais assumem os valores $x = x + 1$ e $x = x$ respectivamente, ou seja,

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

Assim, $\exp(\hat{\beta})$ estima a razão de risco de clientes que assumem o valor $x = x + 1$ em relação aos que tem $x = x$, para qualquer valor de x. Podemos dizer que o risco de se observar o evento de interesse para os clientes que assumem $x = x + 1$ é $\exp(\hat{\beta})$ vezes o risco para os clientes com $x = x$. Dessa forma, a razão de risco quando o valor de x é acrescido em r, é $\exp(r\hat{\beta})$. O parâmetro β pode ser interpretado como o logaritmo da razão de risco dos dois indivíduos considerados.

Quando a covariável classifica os clientes em um entre m grupos, estes grupos podem ser considerados como níveis de um fator. No modelo de riscos proporcionais, a função de risco para um indivíduo no j-ésimo grupo, $j = 1, 2, \ldots, m$, é dado por

$$h_j(t) = \exp(\gamma_j) h_0(t),$$

em que γ_j é o efeito referente ao j-ésimo nível do fator e $h_0(t)$ a função de risco básica. Adotando essa parametrização do modelo, temos que um dos parâmetros assume valor igual a zero para uma determinada categoria ou grupo, denominada referência. As razões de riscos das demais categorias são obtidas em relação a essa categoria adotada como referência. O risco para esse grupo de referência é dado pela função de

risco básica. Assim, a razão de risco, em um determinado t, de um cliente pertencente a um grupo diferente ao de referência em relação ao de referência é $\exp(\gamma_j)$. Similar ao caso de uma variável contínua, podemos dizer que o risco dos indivíduos pertencentes a algum grupo j, $j \geq 2$, é $\exp(\hat{\gamma}_j)$ vezes o risco do grupo adotado como referência. Consequentemente, o parâmetro γ_j é o logaritmo da razão do risco do evento de interesse de um cliente do grupo j para outro pertencente ao grupo um adotado como referência, ou seja,

$$\gamma_j = \log\left\{\frac{h_j(t)}{h_0(t)}\right\}.$$

6.6 Aplicação

A base de dados utilizada para ilustrar a metodologia apresentada neste capítulo é composta por uma amostra de treinamento de 3.000 clientes, obtida via *oversampling* dos dados do exemplo apresentado na Seção 1.2.1, cujas variáveis são apresentadas na Tabela 1.1. Tais clientes iniciaram a utilização de um produto de crédito durante vários meses, compreendendo, portanto, a várias *safras* de clientes, sendo que, para 1.500 clientes não houve problema de crédito, enquanto que os demais clientes tornaram inadimplentes, formando assim a base total de clientes.

A ocorrência ou não de problema de crédito, que determina a classificação dos clientes em *bons* ou *maus* pagadores, foi observada durante os 12 meses seguintes à contratação do produto, que corresponde ao horizonte de previsão do estudo.

O uso de uma amostra com essa quantidade de clientes e com a proporção de 50% de clientes *bons* e 50% de clientes *maus* pagadores foi devido à sugestão dada por Lewis (1994) em relação a quantidade de clientes em cada uma das categorias.

As Tabelas 8.1 e 8.2 apresentam os resultados obtidos por meio do modelo de Cox utilizando as aproximações de Breslow e Efron, respectivamente.

A Figura 8.3 mostra as curvas ROC relacionadas aos ajustes dos modelos de regressão de Cox (BRESLOW) e regressão de Cox (EFRON).

Tabela 8.1 - Regressão de Cox - "BRESLOW".

Variáveis	Descrição das Variáveis	Estimativa	Erro-Padrão	χ^2	p-valor	Odds-Ratio	L.I. (95%)	L.S. (95%)
Intercepto	-	1,1705	0,1219	92,24	<0,0001	-		
P_CARTÃO	Posse de Cartão	-0,9108	0,0819	125,73	<0,0001	**0,402**	0,343	0,472
CASADO	Est.Civil: Casado	-0,2855	0,0807	12,51	0,0004	**0,752**	0,642	0,881
CLI_ANT	Cliente Antigo	-0,4157	0,0835	24,80	<0,0001	**0,660**	0,560	0,777
IDADE32_46	32 < Idade ≤ 46 anos	-0,4993	0,1003	24,76	<0,0001	**0,607**	0,499	0,739
IDADE46_	Idade > 46 anos	-0,9209	0,1025	80,76	<0,0001	**0,398**	0,326	0,487
TEMP_2	T.Emprego ≤ 2 anos	0,5730	0,1112	26,54	<0,0001	**1,773**	1,426	2,205
TEMP2_4	2 <T.Emprego≤ 4 anos	0,3708	0,0989	14,05	0,0002	**1,449**	1,194	1,759
TEL_COMERC.	Declarou Tel.Comercial	-0,2616	0,0799	10,72	0,0011	**0,770**	0,658	0,900
G_CEP_RES1	Grupo1-CEP Residencial	0,9038	0,3356	7,25	0,0071	**2,469**	1,279	4,766
G_CEP_RES2	Grupo2-CEP Residencial	-1,1795	0,4393	7,21	0,0072	**0,307**	0,130	0,727
G_CEP_COM1	Grupo1-CEP Comercial	-0,7032	0,2705	6,76	0,0093	**0,495**	0,291	0,841
G_CEP_COM2	Grupo2-CEP Comercial	-1,1050	0,3961	7,78	0,0053	**0,331**	0,152	0,720

Tabela 8.2 - Regressão de Cox - "EFRON".

Variáveis	Descrição das Variáveis	Estimativa	Erro-Padrão	χ^2	p-valor	Hazard-Ratio	L.I. (95%)	L.S. (95%)
P_CARTÃO	Posse de Cartão	-0,6008	0,0578	107,89	<0,0001	**0,548**	0,490	0,614
CASADO	Est.Civil: Casado	-0,1725	0,0539	10,25	0,0014	**0,842**	0,757	0,935
CLI_ANT	Cliente Antigo	-0,3374	0,0540	39,10	<0,0001	**0,714**	0,642	0,793
IDADE32_46	32 < Idade ≤ 46 anos	-0,3404	0,0641	28,21	<0,0001	**0,711**	0,628	0,807
IDADE46_	Idade > 46 anos	-0,6412	0,0711	81,32	<0,0001	**0,527**	0,458	0,605
TEMP_2	T.Emprego ≤ 2 anos	0,3439	0,0690	24,82	<0,0001	**1,410**	1,232	1,615
TEMP2_4	2 <T.Emprego≤ 4 anos	0,2569	0,0657	15,28	<0,0001	**1,293**	1,137	1,471
TEL_COMERC	Declarou Tel.Comercial	-0,1912	0,0526	13,19	0,0003	**0,826**	0,745	0,916
G_CEP_RES1	Grupo1-CEP Residencial	0,4642	0,1718	7,30	0,0069	**1,591**	1,136	2,228
G_CEP_RES3	Grupo3-CEP Residencial	0,5570	0,2065	7,27	0,0070	**1,745**	1,164	2,616
G_CEP_COM2	Grupo2-CEP Comercial	-0,8573	0,3179	7,27	0,0070	**0,424**	0,228	0,791

A grande semelhança entre os desempenhos dos modelos pode ser justificada pela presença das covariáveis com maior peso na discriminação de *bons* e *maus* clientes, tais como *posse de cartão*, *idade* e *cliente antigo*. Nesta amostra, o método de Breslow, no tratamento dos empates na análise de sobrevivência, selecionou, ao nível de significância 0,01, o menor número de variáveis *dummies*, 9 contra 11 do método de aproximação de Efron. Em ambos os casos o desempenho foi semelhante aos demais métodos.

Com o objetivo de medir e comparar o desempenho dos modelos construídos com base na amostra de treinamento, 30 amostras de teste com aproximadamente 200.000 clientes e na proporção da *base total de clientes*, ou seja, 99% *bons* e 1% *maus* pagadores, foram obtidas e avaliadas pela estatística de Kolmogorov-Smirnov (KS) medindo o quanto os escores produzidos pelos modelos conseguiam separar as duas categorias de clientes, sendo avaliado também a Capacidade de Acerto Total do Modelo (CAT), a Capacidade de Acertos dos *maus* e *bons* clientes,

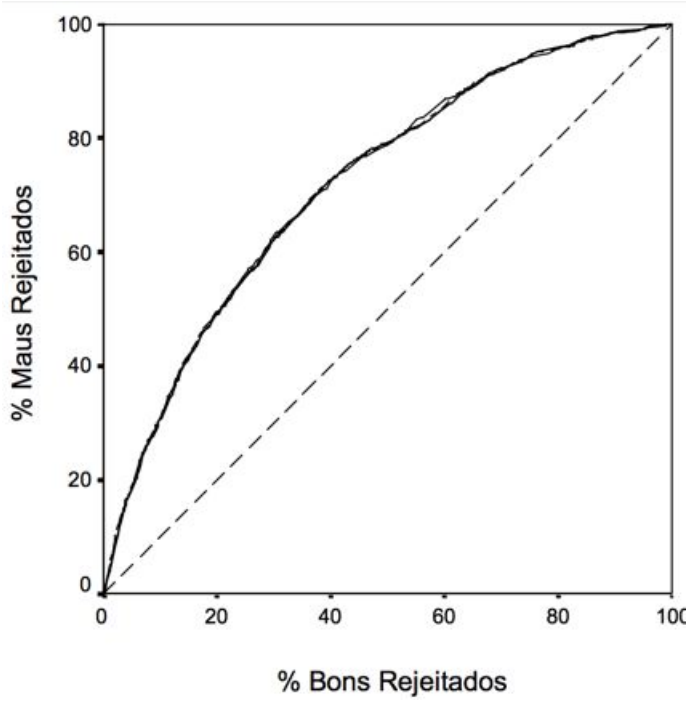

Figura 8.3 - Curva ROC.
(- - -) Referência, regressão de Cox (–) (BRESLOW) e (- - -) (EFRON).

Tabela 8.3 - Resumo dos resultados das 30 Amostras de Teste.

Modelo	K-S (média)	I.C.(95%)	CAT (%) (média)	I.C.(95%)	CAM (média)	I.C.(95%)	CAB (média)	I.C.(95%)
Cox - BRESLOW	32,14	(31,76 ; 32,48)	64,85	(63,85 ;65,77)	67,29	(66,18 ;68,29)	64,83	(63,80 ;65,76)
Cox - EFRON	32,05	(31,65 ; 32,42)	64,66	(63,72 ;65,51)	67,41	(66,47 ;68,27)	64,63	(63,68 ;65,50)

(CAM) e (CAB).

Os resultados apresentados na Tabela 8.3 mostram que o desempenho dos dois modelos ajustados é muito semelhante para os casos estudados, com as mesmas interpretações em relação ao risco de crédito, sendo assim, as categorias consideradas das covariáveis originais, ou seja, *dummies*, trazem evidências de aumento ou diminuição do risco de crédito coincidentes nas duas metodologias.

Ambas metodologias forneceram resultados dentro do praticado pelo mercado para um problema de *Credit Scoring*. No entanto, algumas alterações poderiam ser propostas para alcançar possíveis melhorias no

desenvolvimento dos modelos como, propor diferentes categorizações das covariáveis ou mesmo tentar utilizá-las como contínuas ou propor algumas interações entre elas. A obtenção de informações mais atualizadas para que para ser utilizada na validação dos modelos poderia também trazer ganhos para a metodologia como um todo, fazendo com que os resultados das medidas de avaliação fossem mais próximas e fiéis à realidade atual.

Com base no estudo numérico apresentado observamos, de forma geral, que a metodologia de análise de sobrevivência confirma os resultados encontrados pela regressão logística no ponto específico de observação da inadimplência em 12 meses, tendo como vantagem a utilização no método de estimação das informações das ocorrências desses eventos ao longo do tempo, apresentando assim uma visão contínua do comportamento do cliente, e dessa forma sendo possível, se necessário, a avaliação do risco de crédito dos clientes em qualquer dos tempos dentro do intervalo de 12 meses, o que, de certa forma, provoca uma mudança no paradigma da análise de dados de crédito.

Finalmente ressaltamos que é válido dizer que a semelhança encontrada nos resultados obtidos via regressão logística e análise de sobrevivência, para o conjunto de dados trabalhado, está intimamente relacionada ao planejamento amostral adotado e que resultados diferentes desses poderiam ser encontrados para outros delineamentos, considerando maiores horizontes de previsão e com a utilização de dados comportamentais, em que a análise de sobrevivência pode trazer ganho em relação à regressão logística.

Capítulo 7

Modelo de Longa Duração

Um peculiaridade associada aos dados de *Credit Scoring* é a possibilidade de observarmos clientes, com determinados perfis definidos pelas covariáveis, com probabilidade de inadimplência muito pequena. Tais clientes são considerados "imunes" a este evento dentro do horizonte de 12 meses. Ou seja, dentro do portfólio podemos observar uma proporção considerável de clientes imunes ao evento inadimplência.

Uma curva de sobrevivência típica nessa situação pode ser vista na Figura 9.1, em que observamos poucos eventos ocorrendo a partir do instante de tempo t com elevada quantidade de censuras.

A análise estatística adequada para situações como a descrita acima envolve modelos de longa duração.

7.1 Modelo de Mistura Geral

Para um conjunto de dados, na presença de covariáveis, a função de sobrevivência em um particular instante de tempo t, é definida como

$$S_0(t|\mu(\mathbf{x}), \gamma) = P(T > t|\mu(\mathbf{x}), \gamma) \tag{7.1}$$

em que $\mu(\mathbf{x})$ é um parâmetro de escala, função de outros parâmetros associados às respectivas covariáveis $(\alpha_0, \alpha_1 \ldots, \alpha_k)$ e γ, é um parâmetro de forma constante e não-conhecido.

Considerando o contexto de *Credit Scoring* podemos assumir

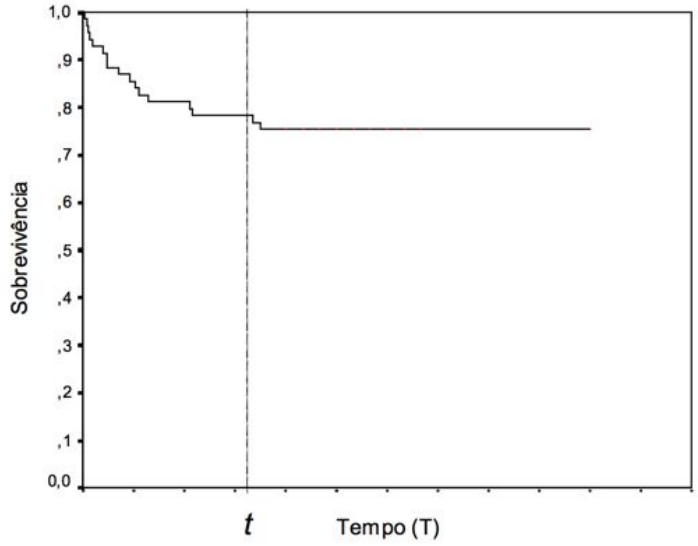

Figura 9.1 - Curva de sobrevivência típica - modelo de longa duração.

para alguns clientes com determinadas características que a inadimplência tem uma probabilidade bastante pequena de ser observada. Assim, admitimos que os indivíduos podem ser classificados como imunes com probabilidade p ou susceptíveis à inadimplência com probabilidade $1 - p$.

Nessas condições consideramos o modelo proposto por Berkson & Gage (1952), conhecido como modelo de mistura, dado por:

$$S(t|\mathbf{x}) = p + (1 - p)\ S_0(t|\mu(\mathbf{x}), \gamma), \tag{7.2}$$

sendo p, $0 < p < 1$, a probabilidade de não observar um problema de crédito para um cliente.

No contexto de *Credit Scoring,* o modelo de longa duração é uma forma de tratar o tempo até a ocorrência de um problema de pagamento de crédito quando uma possível "imunidade" pode ser considerada em relação a esse evento dentro dos 12 meses do horizonte de previsão.

Consideramos aqui um modelo de sobrevivência geral que, além do parâmetro de escala, $\mu(\mathbf{x})$, temos o parâmetro de forma, $\gamma(\mathbf{y})$, e a proporção de clientes "não-imunes", $p(\mathbf{z})$, como dependentes das covariáveis. Em muitas aplicações, a suposição do parâmetro de forma ser constante pode não ser apropriada, uma vez que os riscos de diferentes indivíduos

podem não ser proporcionais.

7.2 Estimação do modelo longa duração geral

Considere um modelo de sobrevivência com parâmetro de escala e de forma dependendo das covariáveis. A correspondente função de sobrevivência é dada por:

$$S_0(t|\mu(\mathbf{x}), \gamma(\mathbf{y})) = P(T > t \mid \mathbf{x}, \mathbf{y}), \tag{7.3}$$

em que $\mu(\mathbf{x})$ é um parâmetro de escala, dependendo de k covariáveis, $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, que tem associados os parâmetros $\alpha = (\alpha_0, \alpha_1 \ldots, \alpha_k)$, $\gamma(\mathbf{y})$ um parâmetro de forma dependendo de p covariáveis, $\mathbf{y} = (y_1, y_2, \ldots, y_p)$, com parâmetros $\beta = (\beta_0, \beta_1 \ldots, \beta_k)$ associados, podendo $\mathbf{x}$ e $\mathbf{y}$ serem iguais.

Para o ajuste de um modelo de sobrevivência de longa duração no contexto de *Credit Scoring*, em que uma proporção de clientes é "imune" à inadimplência dentro do horizonte de previsão de 12 meses, podemos considerar o seguinte modelo

$$S(t|\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{z}) + (1 - p(\mathbf{z}))\ S_0(t|\mu(\mathbf{x}), \gamma(\mathbf{y})),$$

em que $\mu(\mathbf{x})$ e $\gamma(\mathbf{y})$ são os parâmetros de escala e forma da função de sobrevivência usual e p, $0 < p < 1$, representa a probabilidade de não ser observado a inadimplência para um cliente e, também, depende de um vetor de k covariáveis, $\mathbf{z}$, com os parâmetros $\eta = (\eta_0, \eta_1, \ldots, \eta_k)$. Analogamente ao caso anterior, $\mathbf{x}$, $\mathbf{y}$ e $\mathbf{z}$ podem ser iguais.

Assumindo um modelo Weibull para os tempos até a ocorrência da inadimplência, a função de sobrevivência S_0 é escrita como

$$S_0(t|\mu(\mathbf{x}), \gamma(\mathbf{y})) = \exp\left[-\left(\frac{t}{\mu(\mathbf{x})}\right)^{\gamma(\mathbf{y})}\right].$$

Além da distribuição Weibull, várias outras distribuições podem ser con-

sideradas. Dentre as quais destacamos a distribuição log-normal, a log-logística e a gama (Louzada-Neto *et al.*, 2002).

Seja $T_i, i = 1, \ldots, n$, uma amostra dos tempos de sobrevida de n clientes até a ocorrência da inadimplência dentro do horizonte de previsão de 12 meses, o vetor das covariáveis $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$ e uma variável indicadora δ_i, onde $\delta_i = 1$ se for observada a inadimplência para o i-ésimo cliente da amostra e $\delta_i = 0$ se não for observado esse evento. A função de verossimilhança pode ser escrita como

$$L = \prod_{i=1}^{n} f(t_i|\mathbf{z}_i)^{\delta_i} \; S(t_i|\mathbf{z}_i)^{1-\delta_i}, \tag{7.4}$$

sendo $f(t_i|\mathbf{z}_i)$ a função densidade e $S(t_i|\mathbf{z}_i)$ como definida em (7.2).

Seja $\theta' = (\alpha, \beta, \eta)$ o vetor de parâmetros, as estimativas de máxima verossimilhança de θ podem ser obtidas solucionando o sistema de equações não-lineares $\partial \log L/\partial\theta = \mathbf{0}$. Porém, pode ser custoso obter a solução desse sistema diretamente por métodos do tipo iterativo de Newton. Uma forma direta de se obter essa solução é maximizando (7.4). Esse método pode ser implementado via SAS através do procedimento NLP encontrando o valor de máximo local da função não-linear usando métodos de otimização.

Considerando o modelo de sobrevivência Weibull geral em (7.2) e assumindo que os parâmetros de escala, de forma e a probabilidade de incidência do evento são afetados pelo vetor de covariáveis $\mathbf{z}$, por meio das relações log-lineares e logito, ou seja, $\log(\mu(\mathbf{z}_i)) = \alpha_0 + \sum_{j=1}^{k} \alpha_j z_{ij}$, $\log(\gamma(\mathbf{z}_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j z_{ij}$ e $\log\left(\frac{p(\mathbf{z}_i)}{1-p(\mathbf{z}_i)}\right) = \eta_0 + \sum_{j=1}^{k} \eta_j z_{ij}$ respectivamente. Então, a função log-verossimilhança é dada por

$$\begin{aligned} l(\alpha, \beta, \gamma \mid \mathbf{z}) \;\; \propto \;\; & \sum_{i=1}^{n} \delta_i \left[\mathbf{z}_i^t \beta + \; e^{\mathbf{z}_i^t \, \beta} \mathbf{z}_i^t \alpha + \; e^{\mathbf{z}_i^t \, \beta} \log(t_i) \right] \\ + \;\; & \sum_{i=1}^{n} \delta_i \log(p(\mathbf{z}_i)) - \sum_{i=1}^{n} \delta_i (t_i \; e^{\mathbf{z}_i^t \, \alpha})^{e^{\mathbf{z}_i^t \beta}} \\ + \;\; & \sum_{i=1}^{n} (1-\delta_i) \log \left[p(\mathbf{z}_i) + (1 - p(\mathbf{z}_i)) \; e^{(-t_i e^{\mathbf{z}_i^t \alpha})^{e^{\mathbf{z}_i^t \beta}}} \right], \end{aligned} \tag{7.5}$$

em que $p(\mathbf{z}_i)^{-1} = e^{-(\eta_0+\sum_{j=1}^{k}\eta_j z_{ij})}(1 + e^{(\eta_0+\sum_{j=1}^{k}\eta_j z_{ij})})$, α^t=$(\alpha_0, \ldots, \alpha_k)$ β^t= $(\beta_0, \ldots, \beta_k)$, η^t= $(\eta_0, \ldots, \eta_k)$ e $\mathbf{z}_i^t = (1, z_{i1}, \ldots, z_{ik})$.

Uma vez estimados os parâmetros do vetor $\theta^{'} = (\alpha, \beta, \eta)$, uma estimativa da função de sobrevivência, dada em (7.2), pode ser obtida. Os valores dessa função são utilizados como escore final do modelo e, portanto, os clientes podem ser ordenados segundo os seus riscos de crédito.

7.3 Aplicação

A metodologia apresentada neste capítulo é ilustrada em uma base composta por uma amostra de desenvolvimento desbalanceada de 200 mil clientes, na proporção de 99% *bons* e 1% *maus* pagadores, dos dados do exemplo apresentado na Seção 1.2.1 cujas variáveis são apresentadas na Tabela 1.1. Tais clientes iniciaram a utilização de um produto de crédito durante vários meses, compreendendo portanto várias "safras" de clientes, sendo que para 118,8 mil deles não foi observado problema algum de pagamento do crédito, enquanto 1,2 mil clientes se tornaram inadimplentes, formando a base total de clientes. A ocorrência ou não de algum problema de crédito utilizada para a classificação dos clientes em *bons* ou *maus* pagadores foi observada durante os 12 meses seguintes ao início de sua contratação do produto, o qual correspondeu ao horizonte de previsão do estudo.

O modelo de longa duração foi então ajustado, uma vez que observamos um número elevado de censuras nos maiores tempos de acompanhamento, permitindo assim, inferir numa possível presença de clientes "imunes" à inadimplência dentro do horizonte de previsão de 12 meses. O modelo de longa duração é ajustado considerando a função de sobrevivência (7.2), com os parâmetros de escala μ, de forma γ e a proporção de clientes "não-imunes"p, dependentes de covariáveis.

A Tabela 9.1 apresenta os resultados obtidos nesta análise. Observamos que para esse conjunto de dados o parâmetro de forma, γ, não é influenciado pelas covariáveis (p-valor > 0.10) presentes no modelo, sugerindo assim que a suposição de riscos proporcionais é satisfeita. Com relação aos outros dois parâmetros, parâmetro de escala, α, e proporção de "nã o-imunes",p, várias covariáveis são significativas.

Tabela 9.1 - Modelo de longa duração.

Variável	Descrição das Variáveis	Parâmetro de Escala ($\hat{\mu}$)					Parâmetro de Forma ($\hat{\gamma}$)					Proporção de "Imunes" ($\hat{p}$)				
			Est.	E.P.	t	p-valor		Est.	E.P.	t	p-valor		Est.	E.P.	t	p-valor
-	-	α_0	2,0914	0,3223	6,49	<0,0001	β_0	0,6914	0,3826	1,81	0,0706	η_0	-3,8505	0,3341	-11,53	<0,0001
P_CARTÃO	Posse de Cartão	α_1	0,0186	0,0451	0,41	0,6805	β_1	0,0098	0,0650	0,15	0,8798	η_1	-0,9580	0,0684	-14,00	<0,0001
CASADO	Est.Civil: Casado	α_2	0,0261	0,1317	0,20	0,8427	β_2	0,0689	0,1593	0,43	0,6654	η_2	-0,4398	0,1749	-2,52	0,0119
SOLTEIRO	Est.Civil: Solteiro	α_3	0,0004	0,1182	0,00	0,9970	β_3	0,1455	0,1550	0,94	0,3480	η_3	-0,2621	0,1758	-1,49	0,1359
VIUVO	Est.Civil: Viuvo	α_4	-0,2206	0,1563	-1,41	0,1582	β_4	0,2583	0,1834	1,41	0,1591	η_4	-0,4374	0,2177	-2,01	0,0446
CLI_ANT	Cliente Antigo	α_5	0,1243	0,0603	2,06	0,0392	β_5	-0,0597	0,0751	-0,80	0,4265	η_5	-0,3237	0,0744	-4,35	<0,0001
FEM	Sexo: Feminino	α_6	-0,0102	0,0644	-0,16	0,8733	β_6	0,1256	0,0798	1,57	0,1154	η_6	-0,0891	0,0706	-1,26	0,2067
RES_PROP	Residência Própria	α_7	-0,0686	0,0795	-0,86	0,3879	β_7	0,1044	0,1002	1,04	0,2970	η_7	-0,1399	0,1040	-1,34	0,1786
IDADE_25	Idade ≤ 25 anos	α_8	-0,2802	0,1292	-2,17	0,0302	β_8	-0,0594	0,1145	-0,52	0,6040	η_8	0,8429	0,1567	5,38	<0,0001
IDADE25_31	25 < Idade ≤ 31	α_9	-0,3562	0,1350	-2,64	0,0083	β_9	0,0308	0,1180	0,26	0,7938	η_9	0,6979	0,1564	4,46	<0,0001
IDADE31_46	31 < Idade ≤ 46	α_{10}	-0,3009	0,1386	-2,15	0,0312	β_{10}	0,1114	0,1200	0,93	0,3534	η_{10}	0,3105	0,1474	2,11	0,0352
IDADE46_51	46 < Idade ≤ 51	α_{11}	-0,1511	0,1293	-1,17	0,2426	β_{11}	0,0891	0,1317	0,68	0,4984	η_{11}	0,1137	0,1672	0,68	0,4964
TRES_3	T.Resid. ≤ 3 anos	α_{12}	0,0077	0,0476	0,16	0,8719	β_{12}	-0,0482	0,0697	-0,69	0,4893	η_{12}	0,1570	0,0715	2,20	0,0281
TRES15_	T.Resid. >15 anos	α_{13}	0,2396	0,0860	2,79	0,0053	β_{13}	-0,1353	0,0937	-1,44	0,1485	η_{13}	0,0310	0,1050	0,30	0,7678
TEMP_3	T.Emp ≤ 3 anos	α_{14}	-0,0611	0,0546	-1,12	0,2628	β_{14}	0,0589	0,0754	0,78	0,4347	η_{14}	0,1524	0,0830	1,84	0,0665
TEMP3_4	3 <T.Emp. ≤ 4 anos	α_{15}	0,0506	0,1076	0,47	0,6383	β_{15}	-0,1876	0,1188	-1,58	0,1142	η_{15}	0,1781	0,1205	1,48	0,1394
TEMP8_	T.Emp. > 8 anos	α_{16}	-0,0642	0,0851	-0,75	0,4507	β_{16}	-0,1066	0,1058	-1,01	0,3137	η_{16}	-0,2640	0,1005	-2,63	0,0086
TEL_COMERC	Decl.Tel.Comercial	α_{17}	0,1010	0,0619	1,63	0,1028	β_{17}	-0,0228	0,0804	-0,28	0,7771	η_{17}	-0,2025	0,0832	-2,43	0,0149
CORR_RES	Op.Corresp: Resid.	α_{18}	0,2690	0,1006	2,67	0,0075	β_{18}	-0,0313	0,1085	-0,29	0,7734	η_{18}	0,1776	0,1304	1,36	0,1732
C_RENDA_15	Comp.Renda≤15%	α_{19}	0,0922	0,0782	1,18	0,2383	β_{19}	0,0253	0,1089	0,23	0,8166	η_{19}	-0,0674	0,1096	-0,61	0,5386
LIM_300	Valor ≤ R$300	α_{20}	0,1400	0,0705	-1,99	0,0470	β_{20}	-0,0225	0,0921	-0,24	0,8067	η_{20}	0,1036	0,0922	1,12	0,2611
LIM300_460	R$300<Valor≤R$460	α_{21}	-0,0744	0,1037	-0,72	0,4733	β_{21}	-0,1522	0,1223	-1,24	0,2133	η_{21}	0,2917	0,1141	2,56	0,0106
LIM460_780	R$460<Valor≤R$780	α_{22}	-0,1159	0,0842	-1,38	0,1688	β_{22}	0,0053	0,1070	0,05	0,9608	η_{22}	0,1212	0,0973	1,25	0,2128
G_CEP_RES1	Grupo1-CEP Resid.	α_{23}	-0,0576	0,0670	-0,86	0,3894	β_{23}	0,1124	0,1110	1,01	0,3111	η_{23}	0,5043	0,1180	4,28	<0,0001
G_CEP_RES2	Grupo2-CEP Resid.	α_{24}	-0,1104	0,0744	-1,48	0,1379	β_{24}	0,0017	0,0998	0,02	0,9861	η_{24}	0,2225	0,0967	2,30	0,0215
G_CEP_RES3	Grupo3-CEP Resid.	α_{25}	0,0439	0,1118	0,39	0,6946	β_{25}	-0,1570	0,1022	-1,54	0,1242	η_{25}	-0,1051	0,1157	-0,91	0,3635
G_CEP_RES4	Grupo4-CEP Resid.	α_{26}	-0,1080	0,2435	-0,44	0,6575	β_{26}	-0,2795	0,2137	-1,31	0,1910	η_{26}	-0,7798	0,2240	-3,48	0,0005
REGIAO1	Região 1-Clientes	α_{27}	0,0794	0,1256	0,63	0,5273	β_{27}	0,0309	0,2072	0,15	0,8814	η_{27}	0,1622	0,1839	0,88	0,3778
REGIAO2	Região 2-Clientes	α_{28}	0,2091	0,1472	1,42	0,1555	β_{28}	-0,1214	0,2324	-0,52	0,6015	η_{28}	0,2580	0,1818	1,42	0,1558
G_PROF1	Grupo1-Profissão	α_{29}	0,1157	0,0527	2,20	0,0281	β_{29}	-0,0302	0,0682	-0,44	0,6579	η_{29}	-0,1376	0,0681	-2,02	0,0433
G_CEP_COM1	Grupo1-CEP Com.	α_{30}	-0,0811	0,0920	-0,88	0,3783	β_{30}	0,1339	0,1886	0,71	0,4778	η_{30}	0,4320	0,2048	2,11	0,0349
G_CEP_COM2	Grupo2-CEP Com.	α_{31}	-0,0427	0,0561	-0,76	0,4473	β_{31}	0,0568	0,0936	0,63	0,5301	η_{31}	0,1864	0,1036	1,80	0,0720
G_CEP_COM3	Grupo3-CEP Com.	α_{32}	-0,1791	0,0908	-1,97	0,0484	β_{32}	0,0777	0,1231	0,63	0,5281	η_{32}	-0,1276	0,1029	-1,24	0,2152
G_CEP_COM4	Grupo4-CEP Com.	α_{33}	-0,0528	0,0646	-0,82	0,4136	β_{33}	0,0205	0,0764	0,27	0,7886	η_{33}	-0,2238	0,0871	-2,57	0,0102
G_CEP_COM5	Grupo5-CEP Com.	α_{34}	0,1734	0,2357	0,74	0,4619	β_{34}	-0,0284	0,2440	-0,12	0,9075	η_{34}	-0,4594	0,2890	-1,59	0,1120

Para medir o desempenho do modelo de longa duração construído com base na amostra de desenvolvimento, 30 amostras de validação com aproximadamente 200.000 clientes e na proporção da *base total de clientes*, ou seja, 99% *bons* e 1% *maus* pagadores, foram obtidas e avaliadas pela estatística de Kolmogorov-Smirnov (KS) medindo o quanto os escores produzidos pelos modelos conseguiam separar as duas categorias de clientes, sendo avaliado também a Capacidade de Acerto Total do Modelo (CAT), a Capacidade de Acertos dos *Maus* e *Bons* clientes, (CAM) e (CAB). A média da estatística KS foi igual a $33,76$, com um intervalo de confiança igual a $(32,71; 33,56)$; a CAT foi igual a $65,62$, com um intervalo de confiança igual a $(64,32; 67,18)$; a CAM foi igual a $67,93$, com um intervalo de confiança igual a $(64,57; 69,36)$ e a CAB foi igual

a $66,27$, com um intervalo de confiança igual a $(64,53; 67,91)$.

Os resultados são apresentados na Tabela 9.4, em que observamos que o desempenhos dos dois modelos ajustados é muito semelhante para os casos estudados, com as mesmas interpretações em relação ao risco de crédito, sendo assim, as categorias consideradas das variáveis originais, ou seja, *dummies*, trazem evidências de aumento ou diminuição do risco de crédito coincidentes nas duas metodologias.

A utilização de modelos de longa-duração para dados de *Credit Scoring* nos proporciona acomodar a presença de imunes à inadimplência, o que condiz com a realidada encontrada geralmente nas bases de dados de crédito. Entretanto, vários são os motivos que podem levar um cliente à inadimplência. Dentre os quais, ocorrência de desemprego, esquecimento, fraude, entre outros. Inclusive essa informação pode não estar disponível, e nem mesmo a quantidade de possíveis motivos. Neste contexto, modelo de longa-duração, que acomodam estas situações tem sido propostos e podem ser considerados adaptações dos modelos desenvolvidos por Perdona & Louzada-Neto (2011) e Louzada *et al.* (2011) entre outros.

Referências

Allison, P. D. (1995). *Survival analysis using SAS system - A practical guide*. SAS Institute Inc.

Alves, M. C. (2008). *Estratégias para o desenvolvimento de modelos de credit score com inferência dos rejeitados*. Ph.D. thesis, Instituto de Matemática e Estatística - USP.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, **68**(2), 357–363.

Ash, D. & Meesters, S. (2002). *Best Practices in Reject Inferencing*. Wharton Financial Institution Center. Apresentação na credit risk modelling and decisioning conference, Philadelphia.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412–424.

Banasik, J. & Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society*, **56**, 1072–1091.

Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Berry, M. J. A. & Linoff, G. S. (2000). *Mastering data mining*. John Wiley & Sons, New York.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, **30**, 927–961.

REFERÊNCIAS

Black, F. & Scholes, M. S. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, **81**(3), 637–654.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**(2), 123–140.

Breslow, N. (1974). Covariance analysis of censored data. *Biometrics*, **30**(1), 89–100.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms - parts i and ii. *IMA Journal of Applied Mathematics*, **6**(1), 76–90 e 222–231.

Carroll, R., Ruppert, D. & Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Colosimo, E. & Giolo, S. (2006). *Análise de sobrevivência aplicada*. Edgard Blucher.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal Royal Statistic Society - B*, **34**(2), 187–220.

Cox, D. R. & Oakes, D. (1994). *Analysis of survival data*. Chapman & Hall, London.

Cramer, J. S. (2004). Scoring bank loans that may go wrong: a case study. *Statistica Neerlandica*, **58**(3), 365–380.

Crook, J. & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, **28**, 857–874.

Crook, J. & Banasik, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, **183**, 1582–1594.

Durand, D. (1941). Risk elements in consumer instalment financing. Technical report, National Bureau of Economic Research.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**(359), 557–565.

Farewell, V. T. & Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, **67**(2), 273–278.

Feelders, A. (2003). An overview of model based reject inference for credit scoring. Technical report, Utrecht University, Institute for Information and Computing Sciences.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, **13**(3), 317–322.

Geisser, S. (1993). *Predictive inference: an introduction*. Chapman & Hall, New York.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, **24**(109), 23–26.

Gruenstein, J. M. L. (1998). *Optimal use of statistical techniques in model building*. Credit Risk Modeling: Design and Application. Mays E., EUA.

Hand, D. (2001). *Reject inference in credit operations: theory and methods*. The Handbook of Credit Scoring. Company.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied survival analysis*. John Wiley & Sons, New York.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression*. John Wiley & Sons, New York, second edition.

Jorgensen, B. (1984). The delta algorithm and glim. *International Statistical Review*, **52**(3), 283–300.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, **60**(2), 267–278.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data*. John Wiley, New York.

King, G. & Zeng, L. (2001). *Logistic regression in rare events data*. MA: Harvard University, Cambridge.

Kuncheva, L. I. (2004). *Combining pattern classifiers*. Methods and Algorithms. Wiley.

Lewis, E. M. (1994). *An introduction to credit scoring*. Athenas, California.

Linnet, K. (1998). A review of the methodology for assessing diagnostic test. *Clinical Chemistry*, **34**(7), 1379–1386.

Louzada, F., Roman, M. & Cancho, V. (2011). The complementary exponential geometric distribution: Model, properties, and a comparison with its counterpart. *Computational Statistics & Data Analysis*, **55**, 2516–2524.

Louzada, F., Ferreira, P. H. & Diniz, C. A. R. (2012). On the impact of disproportional samples in credit scoring models: An application to a brazilian bank data. *Expert Systems with Applications*, **39**(10), 8071–8078.

Louzada-Neto, F., Mazucheli, J. & Achcar, J. A. (2002). *Análise de Sobrevivência e Confiabilidade*. IMCA – Instituto de Matematicas y Ciencias Afines, Lima-Peru.

Louzada-Neto, F., Anacleto, O., Candolo, C. & Mazucheli, J. (2011). Poly-bagging predictors for classification modelling for credit scoring. *Expert Systems with Applications*, **38**(10), 12717–12720.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, **7**(1), 77–91.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, **405**(2), 442–451.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall, New York, second edition.

McCullagh, P. & Nelder, J. A. (1997). *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Chapman & Hall, EUA.

Moraes, D. (2008). *Modelagem de fraude em cartão de crédito*. Universidade Federal de São Carlos - Departamento de Estatística, São Carlos - SP.

Parnitzke, T. (2005). *Credit scoring and the sample selection bias*. Institute of Insurance Economics, Switzerland.

Perdona, G. S. C. & Louzada-Neto, F. (2011). A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics*, **38**, 1395–1405.

Peto, R. (1972). Contribution to the discussion of a paper by d. r. cox. *Journal Royal Statistic Society - B*, **34**, 205–207.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statitics*, **29**(1), 15–24.

Prentice, R. L. (1976). Generalization of the probit and logit methods for dose response curves. *Biometrics*, **32**(4), 761–768.

Rocha, C. A. & Andrade, F. W. M. (2002). Metodologia para inferência de rejeitados no desenvolvimento de credit scoring utilizando informações de mercado. *Revista Tecnologia de Crédito*, **31**, 46–55.

Rosner, B., Willett, W. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-pearson measurement error. *Statistics in Medicine*, **154**(9), 1051–1069.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, **24**(111), 647–656.

Sicsú, A. L. (1998). Desenvolvimento de um sistema credit scoring: Parte i e parte ii. *Revista Tecnologia de Crédito*.

Stukel, T. A. (1985). *Implementation of an algorithm for fitting a class of generalized logistic models*. Generalized Linear Models Conference Proceedings. Spring-Verlag.

Stukel, T. A. (1988). Generalized logistic models. *Journal of Statitical Association*, **83**(402), 426–431.

Suissa, S. (1991). Binary methods for continuous outcomes: a parametric alternative. *Journal of Clinical Epidemiology*, **44**(3), 241–248.

Suissa, S. & Blais, L. (1995). Binary regression with continous outcomes. *Statistics in Medicine*, **14**(3), 247–255.

Thomas, L. C., B., E. D. & N., C. J. (2002). *Credit scoring and its applications*. SIAM, Philadelphia.

Thoresen, M. & Laake, P. (2007). A simulation study of statistical tests in logistic measurement error models. *Journal of Statistical Computation and Simulation*, **77**(8), 683–694.

Zhu, H., Beling, P. A. & Overstreet, G. A. (2001). A study in the combination of two consumer credit scores. *Journal of Operational Research Sociaty*, **52**, 974–980.

Zweig, M. H. & Campbell, G. (1993). Receiver-operating characteristic (roc) plots. *Clinical Chemistry*, **39**(4), 561–577.