Universidade de Brasília
Instituto de Ciências Exatas
Curso de Estatística

Felipe Nery Lacerda

Risco na concessão de crédito bancário para empresas: Uma aplicação dos modelos de Regressão Logística

Brasília-DF
2015

Felipe Nery Lacerda

Risco na concessão de crédito bancário para empresas: Uma aplicação dos modelos de Regressão Logística

Monografia apresentada ao departamento de Estatística da UnB, como requisito para a obtenção parcial do grau de BACHARELADO em Estatística.

Orientador: Eduardo Yoshio Nakano

Doutor em Estatística - USP

Brasília-DF

2015

Resumo

Este trabalho consiste numa aplicação prática da tecnica estatística de Regressão Logística. Contextualizado sobre os riscos que uma empresa tem ao realizar uma operação de crédito com empresas dos diversos estados brasileiros, levando em consideração as informações individuais já registradas para realizar a devida previsão.

Palavras-chaves: Default, Ratting, Crédito, Risco , Regressão Logística.

Agradecimentos

Agradeço a minha família por todo apoio afetivo e efetivo que me concedeu ao logo de toda a minha vida, sendo meu berço para o convívio com o resto do mundo.

Agradeço também aos amigos que fiz ao longo dos anos e a todo corpo docente que me formou desde as etapas mais simples da escola aos meus atuais formadores na universidade.

Em especial agradeço a Deus, por ter me proporcionado o dom da vida e dados forças para chegar aonde cheguei. Com ele pude chegar a maturidade espiritual e me tornar mais humano nas diversas circunstancias da vida.

"...Também nos gloriamos nas tribulações, porque sabemos que a tribulação produz perseverança; a perseverança, um caráter aprovado; e o caráter aprovado, esperança.".

Romanos 5,3-4

Sumário

1 Introdução

Assim como as empresas, as instituições financeiras fazem o uso de conceitos vindos da administração como os de controle e avaliação. Essas entidades fazem *feedbacks* constantes dos resultados de suas operações, visando melhorias de seus serviços ou produtos. E como uma forma objetiva de fazer estas análises, a aplicação de técnicas estatísticas vem sendo bastante aplicadas nestes segmentos.

Os bancos têm trabalhado da mesma forma em relação aos seus clientes, sejam eles pessoas física ou jurídica. Qualquer que seja o cliente, o banco não pode conceder uma quantidade indeterminada de crédito. A disponibilidade de crédito é avaliada de acordo com o perfil de cada cliente e sua capacidade de honrar com os compromissos. Clientes inadimplentes representam riscos para as entidades financeiras, comprometendo assim o orçamento delas. O banco por sua vez busca alternativas de se proteger destes possíveis "calotes". Assim, ele analisa cada cliente, mensurando o risco que cada um representa. Chamaremos essa mensuração de risco de crédito.

O risco de crédito pode ser avaliado a partir dos seus componentes, que compreendem o risco de *default* (perda), o risco de exposição e o risco de recuperação. O risco de *default* está associado à probabilidade de ocorrer um evento de *default* com o tomador em um certo período de tempo, o risco de exposição decorre da incerteza em relação ao valor do crédito no momento do *default*, enquanto o risco de recuperação se refere à incerteza quanto ao valor que pode ser recuperado pelo credor no caso de um *default* do tomador. (Assaf, 2008)

A importância do estudo do risco de crédito está na possibilidade de tornar a avaliação de clientes bancários cada vez mais objetiva. Esse estudo permitirá quantificar a chance de retorno das operações conforme as caracteristicas de cada indivíduo.

O objetivo desta pesquisa é desenvolver um modelo de classificação de risco de crédito de grandes empresas que atuam no Brasil, utilizando a técnica estatística de regressão logística. O escopo do modelo é prever a ocorrência de eventos de *default*, possibilitando a previsão de um perfil insolvente ou não, o que auxilia os operadores bancários a mensurarem o risco que cada operação traz para o banco.

Todas as análises ralizadas neste trabalho será realizada através do softeware livre R(R core team, 2013).

2 Metodologia

O trabalho consistiu numa pesquisa aplicada sobre o sistema financeiro que envolve os operadores (bancos) e seus clientes. O estudo foi realizado sobre um banco de dados de empresas que solicitaram os serviços bancários. De acordo com as características de cada empresa, foi elaborado um padrão comum de conformidade que sirva como base para a instituição financeira avaliar, de acordo com o tipo de cliente, os riscos que ele oferece em cada operação financeira.

Trata-se de uma pesquisa que envolve aspectos quantitativos e qualitativos, que para elaboração utiliza-se métodos estatísticos. Conforme Hair et al. (2005b), a pesquisa quantitativa é aquela que utiliza números para representar as propriedades em estudo analisando-os por meio de técnicas estatísticas. Neste caso, propõe-se a estudar uma técnica estatística para prever o risco de crédito e reduzir a possibilidade de uma entidade operar com um cliente que não tenha um bom perfil pagador.

De maneira arbitrária, uma análise de regressão modela matematicamente a relação entre uma variável resposta com uma ou mais variáveis explicativas. A grande diferença no modelo de regressão para um modelo matemático é que ele leva em consideração a incerteza dos eventos com bases probabilística. Ou seja, o modelo leva em consideração um erro que se atribui por conta da existência de variabilidade e aleatoriedade. Este erro carrega as informações que o valor estimado do modelo se difere do real.

O uso da regressão logística em parte do estudo se faz necessário por ter uma variável resposta que assume valores binários.

Todas as análises foram realizadas através dos software R (R Development Core Team, 2013).

2.1 O Risco de Crédito

Defini-se crédito como "todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte do seu patrimônio a um terceiro, com a expectativa de que essa parcela volte a sua posse integralmente, após decorrido o tempo estipulado"(SCHRICKEL, 1995, p. 25).

Conforme foi colocado na definição, todo crédito cedido é motivado por uma expectativa de retorno. Tal expectatica traz consigo uma incerteza quanto ao retorno esperado, que se traduz em um risco na concessão do crédito.

Gitman (1997, p. 202) define risco como possibilidade de prejuízo financeiro. Ativos que possuem maiores possibilidades de prejuízo financeiro são mais arriscados que aqueles com menores possibilidades. Risco, dessa forma, pode ser entendido como incerteza ao se referir à "possibilidades de retornos associada a um dado ativo". Entretanto, Lima (2002, p. 20) aponta que "no risco, as probabilidades de ocorrência de um dado evento são conhecidas, enquanto na incerteza não há dados para calcularmos essas probabilidades". Mais especificamente enfocado para uma instituição financeira.

Unindo as ideias definidas acimas, o risco de crédito define-se como a medida numérica da incerteza, uma probabilidades de ocorrência, com relação ao recebimento futuro de um valor contratado (ou compromissado), a ser pago por um tomador de um empréstimo, contraparte de um contrato ou emissor de um título.

O estudo das relações de risco e retorno vem sendo bastantes estudadas nas diversas estruturas de mercado, estando presende nos planejamentos e controles dessas entidades. Assegurados, corretoras, bancos e tantas outras utilizam de ferramentas avançadas e profissionais qualificados para conseguirem colocar nos planejamentos as eventutais perdas que podem sofrer, da mesma forma os ganhos. Com base nos resultados desses estudos que se calcula a amplitude do crédito, os valores dos juros e parcelas. O Risco muitas vezes é considerado mero acaso, e de fato é. Porém não se programar sobre a ocorrência de determinados eventos aleátorios pode significar grandes prejuízos, portanto é melhor se programar bem quanto a ocorrência desses eventos.

2.2 Regressão Logística

Nos modelos de regressão linear simples ou múltipla, a variável dependente Y é uma variável aleatória de natureza contínua. No entanto, em algumas situações, a variável dependente é qualitativa e expressa por duas ou mais categorias, ou seja, admite dois ou mais valores. Esta segunda situação é o caso da regressão logística, ao qual sua resposta é dividida por categorias. Neste trabalho a resposta se restringirá a duas categorias, o qual se restringe a um perfil de sucesso e não sucesso.

Antes de falar especificamente da regressão logística, abordaremos sobre os modelos lineares generalizados. Os modelos lineares generalizados (m.l.g.) foram propostos para modelos cuja a variável resposta Y pode ser representada por alguma distribuição da família exponencial. O m.g.l é especificado por três componentes: uma componente aleatória, a qual identifica a distribuição de probabilidade da variável dependente; uma componente sistemática, que especifica uma função linear entre as variáveis independentes; e uma função de ligação que descreve a relação matemática entre a componente sistemática e o valor esperado da componente aleatória. Em outras palavras, a componente aleatória de um m.l.g. consiste nas observações da resposta, que é uma variável aleatória.

Uma das classes importantes de modelos lineares generalizados, é constituída pelos modelos *logit*, nos quais a variável dependente pode ser associada a uma variável aleatória Bernoulli. Neste caso estamos falando da regressão logística, ao qual abordaremos neste trabalho.

A regressão logística binária se preocupa em representar os "sucessos"ou "fracassos" da população, o que se relaciona com uma distribuição de probabilidade Bernoulli. Sendo assim, as respostas se restringem a apenas dois valores categóricos, geralmente representados por 0 e 1.

Sendo Y_i a resposta do i-ésimo elemento da amostra, com i=1,2,...,n. Temos as seguintes probabilidades de ocorrência dos eventos:

1. $P(Y_i = 0) = (1 - \pi)$
2. $P(Y_i = 1) = \pi$

Com média $E(Y_i) = \pi$ e $Var(Y_i) = \pi(1 - \pi)$.

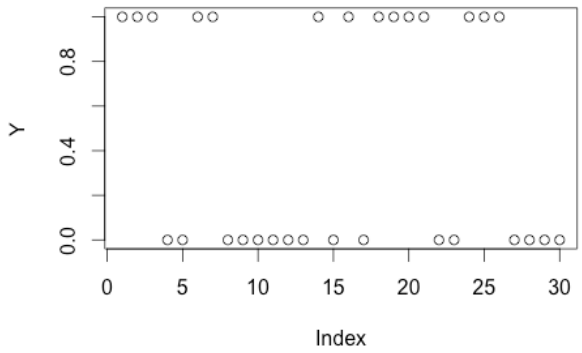

Figura 2.1: Exemplo gráfico da resposta de uma Regressão Logística

Os valores de π são: $0 \leqslant \pi \leqslant 1$.

As principais propriedades da regressão logística são:

1. A função logística é monótona (crescente ou decrescente, dependendo do sinal dos parâmetros associados as variáveis explicativas.
2. É quase linear no intervalo de crescimento e nas extremidades aproxima-se gradualmente de 0 e 1.
3. Pode ser linearizada.

2.2.1 Relação com a Bernoulli

A resposta do i-ésimo indivíduo, Y_i, segue uma distribuição de probabilidade Bernoulli, com parâmetro $\pi(x_i)$, isto é, $Y_i \sim Bernoulli(\pi(x))$, i=1,2,...,n.

Um vetor de variável explicativas $\mathbf{X} = (x_1, x_2, ..., x_k)$, representando as características do i-ésimo indivíduo é associado com a resposta binária Y: Por meio a esperança da Bernoulli, $\pi(\mathbf{x})$. Essa associação é feita através de uma função de ligação, denomidada logito, definida por:

$$ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} \tag{2.1}$$

Sendo $\pi(x_i)$ a probabilidade de sucesso do i-ésimo indivíduo, i=1,2,...,n, que resulta em:

$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})} \tag{2.2}$$

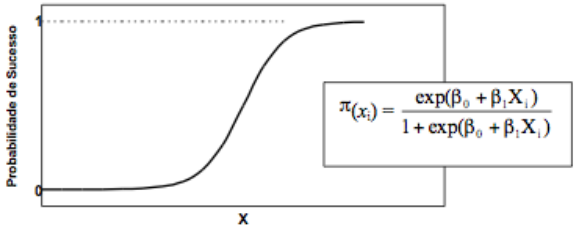

Figura 2.2: Exemplo gráfico de $\pi(x)$ no caso de uma Regressão Logística simples

2.2.2 *Odds Ratio*

Uma forma de interpretar os parâmetros da regressão logística é pela função *odds ratio* - OR (razão de chances), que consiste em comprarar a probabilidade de sucesso com a probabilidade de fracasso do eventos. Para simplificar a notação, a relação será feita sobre um modelo simples.

$$Odds1 = \frac{\frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}}{1 - \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}} \tag{2.3}$$

Odds2 é definico a partir de uma variáção unitária:

$$Odds2 = \frac{\frac{exp(\beta_0 + \beta_1 (X+1))}{1 + exp(\beta_0 + \beta_1 (X+1))}}{1 - \frac{exp(\beta_0 + \beta_1 (X+1))}{1 + exp(\beta_0 + \beta_1 (X+1))}} \tag{2.4}$$

Que resulta em:

$$\frac{Odds2}{Odds1} = exp(\beta_1) \tag{2.5}$$

Assim, $exp(\beta_1)$ é a razão de chances em uma variação unitária da covariável X.

2.2.3 Estimação dos parâmetros

Diferentemente de uma regressão comum, cujos parâmetros são estimados pelo método de mínimos quadrados, no ajustamentos do modelo logístico os parâmetros são estimados pelo método de máxima verossimilhança (MV). Este método maximiza o logaritmo da função de verossimilhança e sua utilização na regressão logística se deve por conhecermos a distribuição de probabilidade associada à variável resposta binária, ou seja uma Bernoulli.

A funcão de probabilidade Bernoulli de cada Y_i, é dada por:

$$P(Y_i = k) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{(1-y_i)} \tag{2.6}$$

Com variável aleatória Y_i independente, aplicamos o produtório sobre a função de distribuição, resultando na seguinte função de verossimilhança:

$$L(\beta/\mathbf{Y}, X) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{(1-y_i)} \tag{2.7}$$

Com y_i=0, 1; e i=1, 2,...,n. Aplica-se o logaritmo natural na função de verossimilhança:

$$ln[L(\beta/\mathbf{Y}, X)] = ln[\prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{(1-y_i)}] \tag{2.8}$$

Desenvolvendo a função e aplicando as propriedades do logaritmo natural, podemos escrever:

$$ln[L(\beta/\mathbf{Y}, X)] = \sum_{i=1}^{n} Y_i ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) + \sum_{i=1}^{n} ln(1 - \pi(x_i)) \tag{2.9}$$

Para facilitar, vamos considerar o vetor $X = (x_1, x_2, ..., x_n)$. Obtemos o vetor $\beta = (\beta_0, \beta_1, ..., \beta_k)$, dos parâmetros a serem estimados com base nos dados observados. Assim, rescrevems a função log-verossimilhança da seguinte forma:

$$ln[L(\beta/\mathbf{Y}, X)] = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}) - \sum_{i=1}^{n} ln[1 + exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})] \tag{2.10}$$

De forma genérica, a verossimilhança é dada pela função de probabilidade conjunta proporcionada por um número n de observações independentes, assim como desenvolvemos anteriormente. Este procedimento escolhe o estimador de MV um vetor dos parâmetros β que fornece os maiores valores possiveis para $L(\beta/\mathbf{Y}, X)$. Para completar o procedimento de maximizar o estimador derivamos em relação a cada parâmetro e posteriormente igualamos a zero. Assim, o estimador de máxima verossimilhança pode ser obtido resolvendo o seguinte sistema de equações:

$$\begin{cases} \dfrac{\partial L(\beta/\mathbf{Y}, X)}{\partial \beta_0} &= 0 \\ \dfrac{\partial L(\beta/\mathbf{Y}, X)}{\partial \beta_1} &= 0 \\ \vdots & \\ \dfrac{\partial L(\beta/\mathbf{Y}, X)}{\partial \beta_k} &= 0 \end{cases}$$

O sistema acima não pode ser solucionado analiticamente, mas sua solução pode ser obtida numericamente atravez de procedimentos do tipo Newton-Raphson.

2.2.4 Intervalo de confiança dos coeficientes

Após garantirmos uma boa estimação de MV, que acarrete pouco ou nenhum viéas aos parâmetros, um intervalo de $(1-\alpha)$ para $\beta_i,\ i = 1, 2, ..., k$ é dado por:

$$\beta_i \pm z_{1-\alpha/2} * (Var(\hat{\beta}_j))^{1/2} \tag{2.11}$$

Onde $Var(\hat{\beta}_i)$ é o elemento diagonal correspondente ao parâmetro de β_i da matriz de informação de Fisher observada e $z_{(1-\alpha/2)}$ é o quantil (1-α/2)*100% de uma Normal padrão.

2.2.5 Seleção de variáveis do modelo Logístco

Para a seleção de váriáveis significativas no mdelos serão usados três métodos:

1. Backward(passo atrás):

   Sobre um determinado modelo completo, temos que investigar as contribuições individuais de cada variável do modelo. A váriável de pior desempenho pode ser

eliminada, desde que não atenta a outros critérios mínimos exirgidos. Para o julgamento de cada variável, comparamos o modelo completo com o modelo reduzido, pela retirada de tal variável. Essa retirada pode ser feita por exemplo pelo teste de Wald. O processo acontece até que restem somente variáveis significativas no modelo.

2. Forward(passo a frente): Ao contrário do método Backward, nesse método as variáveis são acrescentadas sucessivamente uma-a-uma. Caso não haja inclusão em determinada etapa, interrompe-se o processo e as variáveis selecionadas até esta etapa formam o modelo final.

3. Stepwise(passo a passo): Este procedimento é uma forma similar do Forward, após cada etapa de incorporação de uma variável, temos uma etapa em que uma das variáveis já selecionadas pode ser descartada. O processo chega ao fim quando nenhuma variável é incluída ou descartada. Além disso uma variável descartada em algum passo anterior pode vir a ser incluída novamente no modelo.

2.2.6 Ajuste do Modelo

O teste de Hosmer e Lemeshow é a forma mais usual para o ajuste do modelo de regressão logística de resposta binária. Este teste avalia o modelo ajustado comparando as frequêmcias observadas e as esperadas, propondo dois tipos de agrupamento que se baseam nas probabilidades estimadas.

Primeiramente, as observações são classificadas e os eventos de probabilidade são estimandos. As observações são, então, divididos em cerca de 10 grupos. Seja N o número total de indivíduos e M o número alvo de indivíduos para cada grupo é dada por:

$$M \cong [0,1 * N + 0,5]$$

O número de grupos pode ser menor do que 10 se houver menos do que 10 padrões de variáveis explicativas. Devendo haver pelo menos três grupos mínimos para que a estatística de Hosmer-Lemeshow possa ser determinada.

A estatística proposta, por meio de simulação, segue uma distribuição Qui-quadrado quando não há replicação em qualquer uma das subpopulações.

Estatística do teste:

$$H = \sum_{g=1}^{G} \frac{(O_g - N_g \pi g)^2}{N_g \pi_g (1 - \pi_g)} \tag{2.12}$$

Onde,

N_g, é a frequência total de individuos no g-ésimo grupo g=1,2,...,G;

O_g, é a frequência total de resultados de evento no g-ésimo grupo;

π_g, é a probabilidade média estimada previsto de um resultado de eventos para o g-ésimo grupo.

A estatística de Hosmer-Lemeshow é comparada com uma distribuição qui-quadrado com $(g-2)$ graus de liberdade. Maiores valores da estatística do teste em relação ao p-valor indicam uma falta de ajuste do modelo.

2.2.7 Validação do modelo

Esse processo avalia quantitativamente a capacidade de previsão do modelo frente outras observações. Afim de saber se o modelo condiz ou não com a realidade.

O procedimento de validação na regressão logística se assemelha ao de uma regressão linear comum. Existem varias formas distintas de verificar a validação do modelo. Dentre muitas, podemos citar os seguintes procedimentos:

(a) Dividir os dados em duas de mesmo tamanho para serem trabalhados, sendo uma a amostra de estimação e a outra uma amostra de validação. A subamostra de construção do modelo é usada para estimação dos parâmetros do modelo; já as subamostras de validação, têm a função de validar os parâmetros e verificar o poder de predição do modelo construído.

(b) retira-se uma única observação da amostra e ajusta-se o modelo com as $n-1$ observacoes restantes. o valor retirado é predito pelo modelo ajustado. A seguir, a observação retirada e devolvida na amostra e passa-se a retirar uma outra observaçã da amostra, ajustando o modelo e fazendo a previsão com os $n-1$ valores restantes. O processo é repetido até que todas as observações da amostra sejam retiradas (e preditas). Esse processo se chama *Leave-one-out*.

Considera-se que estes são critérios proporcionais a serem estimados na amostra de validação. Nesse sentido, por meio de técnicas de amostragem, sabe-se que a estimação de uma dada proporsão populacional P (sendo P uma variável aleatória), um tamanho de amostra de validação n tal que o estimador $\hat{P}$ de P seja uma boa estimação para a população total.

2.2.8 Cuva de ROC

A curva ROC (Receiver Operating Characteristc) é uma ferramenta que permite avaliar o desempenho de um modelo de uma Regressão binária (variável resposta é do tipo 0-1). Pode ser feita por meio de um gráfico simples e robusto, que nos permite estudar a variação da sensibilidade e especificidade, para diferentes pontos de corte.

Deveremos considerar um ponto de corte C e comparar cada probabilidade estimada com o valor de C. O valor mais utilizado para C é 0,5 (Hosmer Lemeshow, 2000).

A área abaixo da curva de ROC, fornece-nos uma medida de avaliação dos resultados da Regressão Logística binária, conforme o critério de Hosmer e Lemeshow:

Tabela 2.1: Índices de ROC e classificação

Índice	Classificação
ROC $= 0.5$	Não discriminante
$0.7 \leqslant ROC < 0,8$	Aceitável
$0.8 \leqslant ROC < 0,9$	Excelente
ROC ≥ 0.9	Excepcional

A curva ROC é um gráfico de Sensibilidade (ou taxa de verdadeiros positivos) versus taxa de falsos positivos, ou seja, representa a Sensibilidade (ordenadas) vs 1 - Especificidade (abcissas) resultantes da variação de um valor de corte ao longo do eixo de decisão x.

Assim, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade, corres-

pondente ao ponto que se encontra mais próximo do canto superior esquerdo do diagrama, uma vez que o indício de verdadeiro positivo é 1 e o de falso positivo 0.

3 Contextualização do estudo

Seguindo a resolução N^o 2682 do Banco Central do Brasil que dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa. Resolveu a seguinte norma:

Art. 1^o Determinar que as instituições financeiras e demais instituições autorizadas a funcionar pelo Banco Central do Brasil devem classificar as operações de crédito, em ordem crescente de risco, nos seguintes níveis:

- I - nível AA;
- II - nível A;
- III - nível B;
- IV - nível C;
- V - nível D;
- VI - nível E;
- VII - nível F;
- VIII - nível G;
- IX - nível H.

Esses níveis, também denomidados de *ratting* da operação, são atribuídos em função dos atrasos verificados no pagamento das parcelas ou encargos. Seguindo o seguinte padrão:

a) atraso entre 15 e 30 dias: risco nível B, no mínimo;

b) atraso entre 31 e 60 dias: risco nível C, no mínimo;

c) atraso entre 61 e 90 dias: risco nível D, no mínimo;

d) atraso entre 91 e 120 dias: risco nível E, no mínimo;

e) atraso entre 121 e 150 dias: risco nível F, no mínimo;

f) atraso entre 151 e 180 dias: risco nível G, no mínimo;

g) atraso superior a 180 dias: risco nível H;

A instituição financeira ao qual estamos fazendo o devido estudo atribui nota de $0 \dashv 100$ para as operações. Sendo que esta nota é composta por 35% da avaliação cadastral e 65% da avaliação econômico financeira. Essa nota final é convertida para um *ratting*, conforme a seguinte tabela:

Tabela 3.1: Categorização de nota em *Ratting*

Nota	*Ratting*
92 - 100	AA
$79 \vdash 92$	A
$65 \vdash 79$	B
$53 \vdash 65$	C
$48 \vdash 53$	D
$45 \vdash 48$	E
$40 \vdash 45$	F
$25 \vdash 40$	G
$0 \vdash 25$	H

O critério que a instituição financeira, do qual os dados estão sendo avaliados neste trabalho, utiliza para operar com determinada empresa é que o *ratting* desta seja igual ou superior a D. Em termos de notas ordinais, para padronizar este critério adotaremos que o o banco opera somente com pessoas Juridicas com nota igual ou superior a 50.

Sob este critério definido, a variável resposta julgará se empresa tem perfil para operar ou não com o Banco. Essa resposta se traduz de forma binária como sendo: 1 - A empresa tem perfil de operar com a Instituição; 0 - A empresa não tem perfil de operar com a Instituição. A tabela à seguir traz as informações que serão usadas para estimar e explicar o perfil de cada empresa frente a operação bancária, tendo Y como resposta assumindo valores 0 e 1:

Tabela 3.2: Variáveis que compõe a análise

V.a	**Nome**
Y	Resposta
X1	CNAE
X2	Nota cadastral
X3	Receita Operacional Bruta
X4	Total do Ativo
X5	Patrimônio Liquido
X6	Total de Dividas SISBACEN
X7	Quantidade de Sócios
X9	Tempo de Constituição da Empresa em meses
X10	Setor da Empresa
X11	UF da Empresa

3.1 Ajustes das variáveis

Alguns ajustes serão necessário ser feitos nestas variáveis da base de dados para que a análise de regressão seja realizada. Pois algumas variáveis assuem valores muito grades, outras são qualitativas, dentre outros motivos.

Como a variável X5 como assumi-se valores negativos, não foi possível usar nem o logarítimo nem a raiz. Dessa forma, a variável foi categorizada em intervalo dos valores, conforme mostra a Tabela 3.3:

Tabela 3.3: Categorias da variável X5

Categoria	Intervalo
1	Valores negativos
2	$0 \vdash 75000$
3	$75000 \vdash 150000$
4	Maior que 150000

As variáveis X3, X4, X6 e X9 tem uma orfem de magnitude em grande escalas. Como forma de minimizar essa escala para realizar a análise aplicou-se

o logarítimo de base 10 nas variáveis X3, X4, X6 e raiz na informação X9. Como pode-se observar nos histogramas (Figura 3.1), a istribuição dos valores das variáveis assumiram valores menores e mais concentrados.

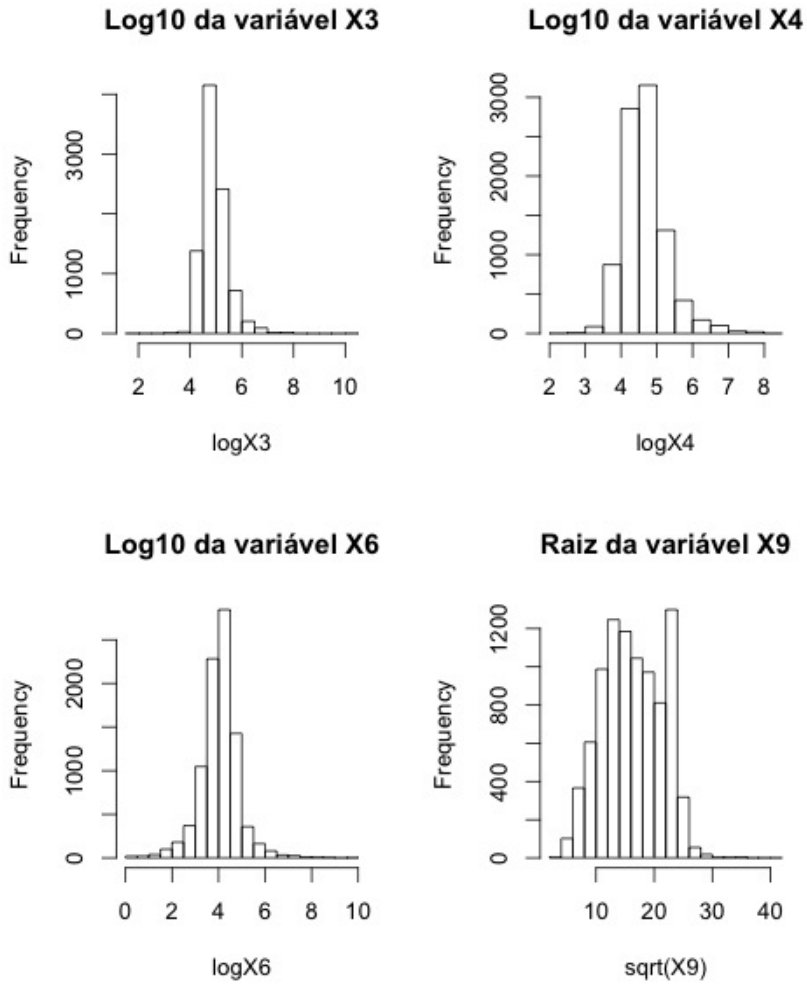

Figura 3.1: Histograma das variáveis ajustadas

Algumas variáveis são de natureza qualitativas, como são o caso das variáveis X1, X10 e X11 (Tabela 3.2). Uma forma de introduzir características qualitativas em modelos de regressão consiste na utilização de variáveis *dummy*, frequentemente chamadas de variáveis binárias ou dicotômicas, uma vez que assumem apenas um de dois valores - em geral 0 ou 1 - para indicar a presença ou ausência de determinada característica.

Para as variáveis X10 e X11, gerou-se dammy para cada categoria presente (sendo elas registradas na base de dados através de códigos. As Tabelas 3.4 e 3.5 definem as codificações das variáveis X10 e X11:

Tabela 3.4: Codificação da variável X10, referente ao setor da empresa.

Código	Setor	Variável
148	Comercio	com
149	Indústria	ind
150	Serviços	serv

Tabela 3.5: Codificação da variável X11, referente à UF.

Estado	Variável	Estado	Variável	Estado	Variável
Acre	Est1	Maranhão	Est10	Rio de Janeiro	Est19
Alagoas	Est2	Mato Grosso	Est11	Rio Grande do Norte	Est20
Amapá	Est3	Mato Grosso do Sul	Est12	Rio Grande do Sul	Est21
Amazonas	Est4	Minas Gerais	Est13	Rondônia	Est22
Bahia	Est5	Pará	Est14	Roraima	Est23
Ceará	Est6	Paraíba	Est15	Santa Cataria	Est24
Destrito Federal	Est7	Paraná	Est16	São Paulo	Est25
Espírito Santo	Est8	Pernambuco	Est17	Sergipe	Est26
Goías	Est9	Piauí	Est18	Tocantins	Est27

No caso da Classicifição Nacional de Atividades Econômicas (CNAE), como são muitos códigos, foi necessário agrupar conforme a influência que cada uma tem sobre a resposta, ou seja, o próprio *odds ratio*. Estimados os efeitos de cada CNAE (Através de uma Regressão Logística simples) agrupou-se as categorias semelhantes ou próximas. Esse agrupamento é representado no Apêndice B e Tabela 3.6.

Tabela 3.6: Agrupamento da CNAE conforme o efeito.

Grupo	Categoria do grupo
cnA	1
cnB	2
cnC	3
cnD	4

4 Resultados

No devido estudo, utilizou-se inicialmente uma base de dados com 11926 observações. Porém, algumas observações foram eliminadas da base de dados por conta de apresentarem valores faltantes ou erro na mensuração. Com os ajustes feitos para realiar a Regressão, restou um total de 9035 para se fazer a modelagem. Da população total, gerou-se uma amostra de 8000 empresas para estimação do modelo, sendo as outras 1035 usadas para validar a modelagem.

Para tomar como referência a entrada das variáveis dummys no modelo de regressão, foram retiradas da modelagem as dummys "cnB"(referente aos códicos da CNAE que pertencem ao grupo B), ao setor comercio("com") e as empresas pertencentes ao estado de São Paulo("est25"). Com isso essas categorias ficaram como nível de referência na análise.

4.1 Estimativa dos parâmetros e intervalos de confiança

No estudo de análise de regressão usou-se o comando *glm* do software R que trata dos modelos lineares generalizados. Para a elaboração do modelo logístico especificamos a modelagem para a família da binomial. Obtidas as estimativas dos parâmetros encontramos a seguinte equação para o modelo completo:

$$\pi(x) = \frac{exp(g(x))}{1 + exp(g(x)} \tag{4.1}$$

Com os coeficientes estimados apresentados pela tabela 5.1:

g(x) = -6.7437 -25.2401*cnA + 1.2332*cnC + 20.5736*cnD + 0.0631*X2 + 0.2985*logX3 + 0.1925*logX4 + 1.0414*X5c - 1.0940*lX6 + 0.0720*X7 + 0.5986*ind + 0.3819*serv + 1.7679*est1 - 0.7709*est2 + 3.5012*est3 + 0.9980*est4 + 0.3481*est5 + 0.2673*est6 + 0.6526*est7 + 0.5124*est8 + 0.2381*est9 - 0.2736*est10 - 0.4926*est11 - 0.1372*est12 + 0.0618*est13 + 0.9015*est14 + 0.7485*est15 - 0.2095*est16 +

0.6361*est17 + 0.7747*est18 + 0.3868*est19 + 0.3360*est20 + 0.1820*est21 + 0.1017*est22 + 13.8549*est23 - 0.0107*est24 + 0.1472*est26 - 1.1090*est27

Como qualquer outra regressão linear, os parâmetros positivos aumentam o valor de *g(x)*, tendo assim uma relação direta. já por sua vez os parâmetros negativos fazem com que o valor de *g(x)* diminua à medida que x aumenta, observando assim uma relação decrescente.

Tabela 4.1: Estimativas dos parâmetros do modelo Logístico.

	Estimativa	E.P	valor z	Pr(> |z|)
(Intercept)	-6.7437	0.4550	-14.82	0.0000
cnA	-25.2401	465.3019	-0.05	0.9567
cnB*	0	–	–	–
cnC	1.2332	0.0688	17.92	0.0000
cnD	20.5736	403.0429	0.05	0.9593
X2	0.0631	0.0022	28.39	0.0000
logX3	0.2985	0.1256	2.38	0.0175
logX4	0.1925	0.1279	1.51	0.1323
X5c	1.0414	0.0774	13.46	0.0000
logX6	-1.0940	0.0623	-17.57	0.0000
X7	0.0720	0.0160	4.51	0.0000
sqrt(X9)	-0.0030	0.0062	-0.48	0.6320
ind	0.5986	0.0752	7.96	0.0000
serv	0.3819	0.0880	4.34	0.0000
com*	0	–	–	–
est1	1.7679	0.7144	2.47	0.0133
est2	-0.7709	0.4150	-1.86	0.0633
est3	3.5012	0.7963	4.40	0.0000
est4	0.9980	0.3108	3.21	0.0013
est5	0.3481	0.1622	2.15	0.0319
est6	0.2673	0.1781	1.50	0.1334
est7	0.6526	0.2651	2.46	0.0138

	Estimativa	E.P	valor z	Pr(> |z|)
est8	0.5124	0.2031	2.52	0.0116
est9	0.2381	0.1780	1.34	0.1810
est10	-0.2736	0.2901	-0.94	0.3457
est11	-0.4926	0.3511	-1.40	0.1606
est12	-0.1372	0.4198	-0.33	0.7438
est13	0.0618	0.0986	0.63	0.5308
est14	0.9015	0.4659	1.94	0.0530
est15	0.7485	0.2821	2.65	0.0080
est16	-0.2095	0.1124	-1.87	0.0622
est17	0.6361	0.1841	3.46	0.0005
est18	0.7747	0.4953	1.56	0.1178
est19	0.3868	0.1354	2.86	0.0043
est20	0.3360	0.3466	0.97	0.3324
est21	0.1820	0.1026	1.77	0.0761
est22	0.1017	0.3531	0.29	0.7733
est23	13.8549	329.0354	0.04	0.9664
est24	-0.0107	0.1205	-0.09	0.9292
est25*	0	–	–	–
est26	0.1472	0.2778	0.53	0.5961
est27	-1.1090	0.5360	-2.07	0.0385

* Níveis de referência.

A função matemática elaborada permite estabelecer a probabilidade de uma observação estar em condição ou não de realizar a operação, justificada pelo comportamento do conjunto de variáveis independentes. Os coeficientes estimados indicam a importância de cada variável independente para a ocorrência do evento.

Na determinação dos intervalos de confiaça algumas variáveis tiveram alguns problemas, causando uma indeterminação dos intervalos. No caso da variável referente ao estado de Roraima (est23) tem apenas 8 observação em toda base de dados, número bastante pequeno para se fazer inferência e pode ter sido o motivo do grande erro. Os grupos A e D da Classicifição Nacional de Atividades Econômicas (CNAE) obtiveram uma frequência de zero e próximo de zeros respectivamente em

relação a resposta, desta forma a indeterminação foi causada pela falta de variabilidade em relação a resposta.

Tabela 4.2: Intervalo de confiança dos parâmetros do modelo

	estimativa	IC	95 %		estimativa	IC	95 %
(Intercept)	-6.74	-7.63	-5.85	est7	0.65	0.14	1.18
cnA	-25.25	-	-	est8	0.52	0.12	0.92
cnB	0	0	0	est9	0.24	-0.11	0.59
cnC	1.23	1.10	1.37	est10	-0.27	-0.85	0.29
cnD	20.57	-	-	est11	-0.48	-1.18	0.20
X2	0.06	0.06	0.07	est12	-0.13	-0.97	0.69
logX3	0.30	0.05	0.54	est13	0.06	-0.13	0.25
logX4	0.18	-0.06	0.43	est14	0.91	0.03	1.87
X5c	1.04	0.89	1.19	est15	0.75	0.20	1.31
logX6	-1.09	-1.22	-0.97	est16	-0.21	-0.43	0.01
X7	0.07	0.04	0.10	est17	0.64	0.28	1.00
sqrt(X9)	-0.003	-0.015	0.009	est18	0.78	-0.15	1.80
ind	0.60	0.45	0.75	est19	0.39	0.13	0.66
serv	0.38	0.21	0.55	est20	0.33	-0.35	1.01
com	0	0	0	est21	0.18	-0.02	0.38
est1	1.77	0.47	3.35	est22	0.11	-0.59	0.80
est2	-0.77	-1.61	0.03	est23	13.87	-	-
est3	3.50	2.14	5.41	est24	-0.01	-0.25	0.23
est4	1.00	0.40	1.62	est25	0	0	0
est5	0.35	0.04	0.67	est26	0.15	-0.40	0.70
est6	0.27	-0.08	0.62	est27	-1.10	-2.20	-0.08

cnB(CNAE), comércio(setor), est25(estado) são os níveis de referência.

4.2 Ajustes de Validação

Pelo método Stepwise de seleção de variáveis mediremos a importância de cada variável para o modelo. Sob este critério testaremos o modelo sem a variável X9 é ou não significativamente diferente.

	Resid. Df	Resid. Dev	GL	Deviance	Pr(>Chi)
Modelo sem X9	7962	6994.25			
Modelo com X9	7961	6993.42	1	0.84	0.3600

Com um p-valor 0.36 não há evidencia de que os modelos sejam diferentes, dessa forma optamos por um modelo sem a variável X9.

Para testar o ajuste do modelo em relação a $\pi(x)$ e Y, utilizaremos agora o teste de Hosmer e Lemeshow. Tomando como base um padrão de 10 grupos, como mostra a tabela de contigência a seguir:

Tabela 4.3: Tabela de contingência para o teste de Hosmer e Lemeshow

	Estimado		Observado	
Grupo	Y=0	Y=1	Y=0	Y=1
[0,0.0046]	20.00	0.00	20.00	0.00
(0.0046,0.074]	19.25	0.75	20.00	0.00
(0.074,0.239]	16.80	3.20	18.00	2.00
(0.239,0.424]	13.31	6.69	12.00	8.00
(0.424,0.504]	10.75	9.25	8.00	12.00
(0.504,0.624]	8.84	11.16	11.00	9.00
(0.624,0.71]	6.66	13.34	9.00	11.00
(0.71,0.785]	5.01	14.99	1.00	19.00
(0.785,0.906]	2.96	17.04	2.00	18.00
(0.906,1]	0.18	19.82	0.00	20.00

As hipóteses em teste são:

H_0: O modelo ajusta bem os dados. H_a: O ajuste dos dados não é bom.

Calculando o teste de Hosmer e Lemeshow, obtivemos os seguintes valores: χ^2 = 10,2187, gl = 8, p-valor = 0,25. Com esses resultados e sob um nível de significância de 5%, podemos concluir que o modelo está ajustando bem com o valores.

Com a amostra de validação classificamos e observamos as frequências de acertos e erros da estimação:

Tabela 4.4: Classificação - validação do modelo

	Estimado		
Observado	$Y_{est} = 0$	$Y_{est} = 1$	Total
$Y_{obs} = 0$	387	138	525
$Y_{obs} = 1$	95	415	510
Acertos(%)	73.7%	81.4%	1035

Tabela 4.5: Índices de ROC e classificação

Índice	Classificação
ROC = 0.5	Não discriminante
$0.7 \leqslant ROC < 0,8$	Aceitável
$0.8 \leqslant ROC < 0,9$	Excelente
ROC $\geq$ 0.9	Excepcional

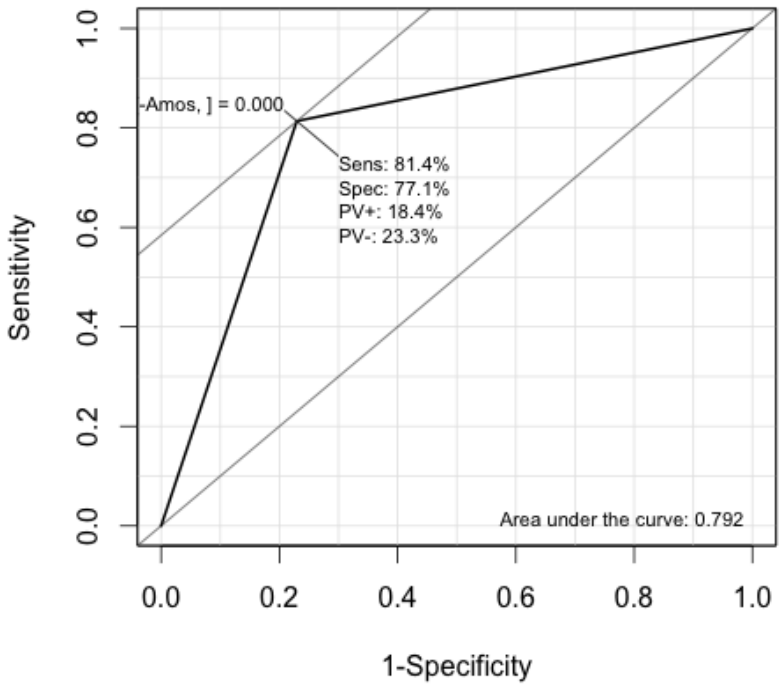

Figura 4.1: Curva de ROC.

Avaliando a curva ROC para o modelo de regressão logística binário,

apresentada na Figura (4.1), verifica-se que a área sob a curva corresponde aproximadamente 0,8, demonstrando um excelente poder de discriminação de acordo com a classificação dada por Hosmer e Lemeshow (2000). A sensitividade é a proporção de acerto na previsão da ocorrência de um evento nos casos em que ele ocorreu. A especificidade é proporção de acerto na previsão da não ocorrência do evento de interesse, no caso deste estudo a empresa não ter perfil de operar com o banco.

5 Conclusão

O objetivo deste estudo foi desenvolver um modelo de concessão de crédito para empresas que atuam no Brasil, utilizando a tecnica estatística da regressão logística. Para elaboração do modelo proposto utilizou-se como base um conjuto de informações individuais a cada empresa como variável explicativa e o *ratting* como corte de resposta.

O modelo alcançou um índice de acertos bastante significativo para a previsão do *default*, que mensura a capacidade da empresa operar ou não operar com o banco.

A regressão Logística se mostrou bastante eficaz para a previsão do evento estudado. Por mais que na prática o modelo necessite ser ajustado rotineiramente, por conta de fatores externos que podem influênciar a disponibilidade de crédito, como a situação econômica do país e as medidas politicas, dentre outros fatores. Este tipo de modelagem sugere uma boa alternativa para se avaliar o Risco de crédito.

Referências Bibliográficas

[1] Agresti, Alan. An Introduction to Categorical Data Analysis, Jonh Wiley Sons, New York, 2edition, 2007.

[2] BARRETO, ALEXANDRE SERRA., *Modelos de regressão: teoria e com o programa estatístico R*, Ed. do autor, Brasília, 2011.

[3] BRIGHAM, E.F., GAPENSKI, L.C., ERHARDT, M.C. Administração Financeira: Teoria e Prática. São Paulo: Atlas, 2001.

[4] Aplicação do Modelo de Regressão Logística num Estudo de Mercado - Cleidy Isolete Silva Cabral, 2013

[5] SISTEMA DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO: UMA APLICAÇÃO A COMPANHIAS ABERTAS NO BRASIL - Giovani Antonio Silva Brito; Alexandre Assaf Neto (USP)

[6] HAIR JR, Joseph F. et al. Análise multivariada de dados. 5. ed. Porto Alegre: Bookman, 2005.

[7] Hosmer, D. W., and Lemeshow, S. (2000). Applied Logistic Regression, Second Edition. Wiley, New York.

[8] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

6 Apêndice A - Script no R

```
tccdados <- read.csv(" /ENTRADA DOS DADOS", sep=";", dec=",")

        q=rep('estQ',27) i=1:27

        f=function(j)sub('Q',i[j],q[j])

        vetor=sapply(1:27,f)

        levels(tccdados$X11)=vetor

        table(tccdados[,15])

        ### CRIANDO VARIÁVEL DAMMY PARA ESTADO "est1"###

        v=vector()

        for(i in 1:nrow(tccdados)) if(tccdados$X11[i]=='est1') v[i]=1  else v[i]=0

        v1=tccdados$X11

        ff=function(i,estado) v2=vector() if(v1[i]==estado) v2[i]=1  else  v2[i]=0
v2[i]

        ff(100,'est19')

        dummy para 'est1' est1=sapply(1:length(v1),function(i)ff(i,'7380'))

        est2=sapply(1:length(v1),function(i)ff(i,'7381'))

        est3=sapply(1:length(v1),function(i)ff(i,'7382'))

        est4=sapply(1:length(v1),function(i)ff(i,'7383'))

        est5=sapply(1:length(v1),function(i)ff(i,'7384'))

        est6=sapply(1:length(v1),function(i)ff(i,'7385'))

        est7=sapply(1:length(v1),function(i)ff(i,'7386'))

        est8=sapply(1:length(v1),function(i)ff(i,'7387'))

        est9=sapply(1:length(v1),function(i)ff(i,'7388'))

        est10=sapply(1:length(v1),function(i)ff(i,'7389'))

        est11=sapply(1:length(v1),function(i)ff(i,'7390'))
```

```
est12=sapply(1:length(v1),function(i)ff(i,'7391'))

est13=sapply(1:length(v1),function(i)ff(i,'7392'))

est14=sapply(1:length(v1),function(i)ff(i,'7393'))

est15=sapply(1:length(v1),function(i)ff(i,'7394'))

est16=sapply(1:length(v1),function(i)ff(i,'7395'))

est17=sapply(1:length(v1),function(i)ff(i,'7396'))

est18=sapply(1:length(v1),function(i)ff(i,'7397'))

est19=sapply(1:length(v1),function(i)ff(i,'7398'))

est20=sapply(1:length(v1),function(i)ff(i,'7399'))

est21=sapply(1:length(v1),function(i)ff(i,'7400'))

est22=sapply(1:length(v1),function(i)ff(i,'7401'))

est23=sapply(1:length(v1),function(i)ff(i,'7402'))

est24=sapply(1:length(v1),function(i)ff(i,'7403'))

est25=sapply(1:length(v1),function(i)ff(i,'7404'))

est26=sapply(1:length(v1),function(i)ff(i,'7405'))

est27=sapply(1:length(v1),function(i)ff(i,'7406'))

tccdados=data.frame(tccdados, est1, est2, est3, est4, est5, est6, est7, est8,
est9, est10, est11, est12, est13, est14, est15, est16, est17, est18, est19, est20, est21,
est22, est23, est24, est25, est26, est27)

### Dummies para setor ###

S=rep('setQ',3) N=1:3

ST=function(M)sub('Q',N[M],S[M])

vetor=sapply(1:3,ST)

levels(tccdados$X10)=vetor

table(tccdados[,14])

### CRIANDO VARIÁVEL DAMMY PARA ESTADO "est1"###

d=vector()
```

```
for(N in 1:nrow(tccdados)) if(tccdados$X10[i]==’set1’) d[N]=1 else d[N]=0


d1=tccdados$X10

dd=function(N,setor) d2=vector() if(d1[N]==setor) d2[N]=1 else d2[N]=0
d2[N]

dd(100,’set2’)

### dummy para ’est1’ ###

com=sapply(1:length(d1),function(N)dd(N,’148’))

ind=sapply(1:length(d1),function(N)dd(N,’149’))

serv=sapply(1:length(d1),function(N)dd(N,’150’))

### Categorizando variável X5 ###

Vex5=tccdados$X5

cat1<-ifelse(Vex5 ¿ =0,0,1)

cat2<-ifelse(Vex5 > 0  Vex5 > 75000,2,0)

cat3<-ifelse(Vex5 > 75000  Vex5 >=150000,3,0)

cat4<-ifelse(Vex5 > 150000,4,0)

X5c=cat1+cat2+cat3+cat4

tccdados=data.frame(tccdados, X5c, com, ind, serv)

### variáveis CNAE ###

# CNAE¡-as.factor(tccdados$X1)

# rl=glm(Y CNAE,family=binomial)

# tam[1-4]=length(1-4)

# tam=vector(length=4)

# idgrupo=rep(1:4,tam)

# vet1=c(cnaeA,cnaeB,cnaeC,cnaeD)

# idgrupo

# grupocnae=data.frame(vet1,idgrupo)

# vet1
```

```
# dados=merge(tccdados,grupocnae, by.x="X1",by.y="vet1")

gp=tccdados$idgrupo

cnA<-ifelse(gp==1,1,0)

cnB<-ifelse(gp==2,1,0)

cnC<-ifelse(gp==3,1,0)

cnD<-ifelse(gp==4,1,0)

lX6=log(tccdados$X6)

tccdados=data.frame(tccdados,cnA,cnB,cnC,cnD,lX6)

### hist do ajuste das variáveis ###

par(mfrow=c(2,2)) arranjamento 2 por 2 plot(x,y)

hist(tccdados$logX3, main= "Log10 da variável X3",xlab="logX3")

hist(tccdados$logX4, main= "Log10 da variável X4",xlab="logX4")

hist(tccdados$logX6, main= "Log10 da variável X6",xlab="logX6")

hist(sqrt(tccdados$X9), main= "Raiz da variável X9",xlab="sqrt(X9)")

### Histograma da correlação entre as variáveis ###

hist(corre[1:1406],

main="Histograma dos coeficientes de correlação entre as variáveis",

nc=10,

xlab="Coeficientes de correlação",

ylab="Probabilidades",

xlim=c(-0.5,0.99),

ylim=c(0,1),

col="gray",

border="white",

prob=T,

right=T,

col.axis="red")
```

Análise de Regressão

Amos¡-sample(nrow(tccdados),8000) amostra de estimação

RL¡-glm(formula = Y cnA + cnC + cnD + X2 + logX3 + logX4 + X5c + logX6 + X7 + sqrt(X9) + ind + serv + est1 + est2 + est3 + est4 + est5 + est6 + est7 + est8 + est9 + est10 + est11 + est12 + est13 + est14 + est15 + est16 + est17 + est18 + est19 + est20 + est21 + est22 + est23 + est24 + est26 + est27, family = binomial, data = tccdados[Amos,])

summary(RL)

###############

Y<-tccdados$Y

prob.est¡-fitted(RL)

y.pre=ifelse(prob.est¿=0.5,1,0)

Pd=table(Y[Amos],y.pre)

Acertando 0

Pd[1,1]/(Pd[1,1]+Pd[1,2])

Acertando 1

Pd[2,2]/(Pd[2,2]+Pd[2,1])

Seleção de variáveis

stw=step(RL)

summary(stw)

stw$anova

RL2¡-glm(formula = Y cnA + cnC + cnD+ X2 + logX3 + logX4 + X5c + logX6 + X7 + ind + serv + est1 + est2 + est3 + est4 + est5 + est6 + est7 + est8 + est9 + est10 + est11 + est12 + est13 + est14 + est15 + est16 + est17 + est18 + est19 + est20 + est21 + est22 + est23 + est24 + est26 + est27, family = binomial, data = tccdados[Amos,])

prob.est2¡-fitted(RL2)

y.pre2=ifelse(prob.est2¿=0.5,1,0)

Pd2=table(Y[Amos],y.pre2)

Acertando 0

Pd2[1,1]/(Pd2[1,1]+Pd2[1,2])

Acertando 1

Pd2[2,2]/(Pd2[2,2]+Pd2[2,1])

ajus=anova(RL2,RL,test="Chisq")

predict Manual rep.1=rep(1,length(Y)) valRL=data.frame(rep.1,cnA,cnC,cnD,tccda
tccdados$logX3, tccdados$logX4, X5c,tccdados$logX6, tccdados$X7, ind, serv, est1, est2, est3, est4, est5, est6, est7, est8, est9, est10, est11, est12, est13, est14, est15, est16, est17, est18, est19, est20, est21, est22, est23, est24, est26, est27)

COEF=RL2$coefficients

ppoo<-predict(RL2)

yyy=ifelse(ppoo>=0.1,1,0)

Ppp=table(Y[Amos],y.pre2)

Acertando 0

Ppp[1,1]/(Pd2[1,1]+Pd2[1,2])

Acertando 1

Ppp[2,2]/(Pd2[2,2]+Pd2[2,1])

kk=fitted(RL2)

ttt=data.frame(Y[Amos],kk)

kke=ifelse(kk¡0.5,0,1)

ttt=data.frame(Y[Amos],kk,kke)

table(Y[Amos],kke)

predict manual

pred.valit=COEF*t(valRL)

vet.predd=data.frame(t(pred.valit))

vet.pred=rowSums(vet.predd)

vet.pred=data.frame(vet.pred)

vet.probp=exp(vet.pred)/(1+exp(vet.pred))

```
vet.01=ifelse(vet.probp<0.5,0,1)

### tabelas de frequencia ###

fre.amos=table(Y[Amos],vet.01[Amos,])

fre.valit=table(Y[-Amos],vet.01[-Amos,])

### intervalo de confiança dos parametros ###

intervalo.conf=exp(confint(RL2))

IC.nu=confint(RL2)

IC.nu=cbind(fit=coef(RL2), IC.nu)

IC.mu=1/(1+exp(-IC.nu))

IC.mu=cbind(trat=rownames(IC.mu), as.data.frame(IC.mu))

### Validação do modelo Hosmer e lemershower ###

# install.packages("Resource Selection ")

hl=hoslem.test(Y[Amos], vet.probp[Amos,] , g = 10)

cont.HL=cbind(hl$expected,hl$observed)

### Curva de ROC ### install.packages("Epi")

ROC(test=vet.01[-Amos,],stat=Y[-Amos],plot="ROC", MI=F)

### Rodando tabela para Latex - pacote xtable ###

RL.latex = xtable(RL)

print(RL.latex, type='latex')
```

7 Apêndice B - Codificação da variável CNAE

Códigos da CNAE pertenentes ao grupo cnA:

113900 115600 132500 133407 161999 210101 810004 899102 1042200 1072401 1093702 1096100 1113502 1210700 1312000 1314600 1323500 1329304 1352900 1422203 1531800 1551202 1571701 1582200 1591102 1621800 1710900 1731100 1732000 1742701 1749400 1750799 1811201 1910100 1922501 2061400 2073800 2110600 2211000 2211100 2221700 2222501 2229303 2310800 2330301 2342702 2431800 2471600 2473200 2481300 2492901 2494500 2539000 2620400 2630105 2632900 2725199 2749999 2751000 2814301 2822401 2832000 2839800 2852600 2891600 2914900 2923800 2943300 2961000 2962901 2992002 3012100 3022800 3092000 3112700 3113501 3230100 3299005 3299099 3310301 3316301 3443600 3444400 3520402 3531900 3614500 3699499 3701100 3702900 3821100 3831901 3832700 4011800 4100900 4171704 4213800 4221901 4311801 4319300 4399101 4399103 4522503 4523300 4524100 4532201 4541101 4559499 4614100 4618499 4623104 4623108 4634603 4634699 4636202 4637199 4641903 4642701 4643501 4644302 4645103 4651602 4721101 4721103 4724500 4759801 4761001 4782202 4912402 4912403 4923002 4930201 5010506 5011401 5111000 5111100 5112800 5113600 5121708 5133001 5134900 5135700 5146201 5147001 5151901 5152700 5153505 5159403 5165901 5215901 5222200 5229001 5241805 5243499 5244205 5245003 5246901 5249311 5250801 5250803 5271001 5511501 5813100 5911199 6021700 6025902 6027500 6110801 6210300 6311800 6321501 6420399 6542100 6810201 7119702 7119799 7120100 7229000 7290700 7711000 8012900 8111700 8292000 8299701 8514602 8550302 8610102 8630503 8630599 8640202 8660700 9112000 9191000 9261401 9262299 9309299 9312300 9511800

Códigos da CNAE pertenentes ao grupo cnB:

2529199 2063100 4771704 5132201 1761200 2091600 2451100 3600601 8211300 4929902 2591800 2981500 4649406 1071600 1111901 3329599 4100901 5191800 5822100 6613400 4930203 2219500 1733800 2019399 2031200 2829199 2920401 5242601 5620101 4771701 4930202 4773300 1414200 1931400 2861500 2930101 4292802 5213202 7132300 7415200 7732202 8299799 8640205 3449500 1099699 4712100 5241801 1051100 1093701

1094500 1112700 1421500 2021400 2021500 2812600 3449502 4110700 4299599 4662100 4763601 5121709 5154399 5244299 6026701 6311900 6463800 7020400 7739099 6026702 1359600 4753900 2342701 1721400 2542000 3511401 4622200 5030002 6024001 7460804 2592601 4619200 6024002 7490104 4511104 1539400 4921301 2222600 5191801 151201 1011205 1065101 1340502 1514800 1561000 1749300 2062200 2110500 2219600 2330305 2330399 2411300 2429599 2622100 2823200 2930103 3121600 4313400 4329104 4522501 4646002 4669999 4761003 5021102 5142001 5211701 6120502 7414400 9321200 9430800 155501 1013901 1066000 4742300 4681801 5212400 8121400 1311100 1556300 4529299 4647801 6462000 1811302 2899100 3811400 4921302 4922101 1122401 4212000 4511102 4685100 6023202 6201500 1012101 2499600 2532201 4781400 7830200 4691500

Códigos da CNAE pertenentes ao grupo cnC:

139206 810099 1354500 1511302 2412100 2449199 2724301 2862300 4729699 5231102 5244208 6204000 1061901 2229399 4120400 4211101 7820500 5030004 2071100 1511301 1771000 2751100 4663000 54147 133404 1052000 1521000 2340000 2519400 4321500 4530705 4633801 7311400 7810800 8531700 4672900 1512101 2989000 4010002 4549799 4681802 4744099 5221400 8610101 2949299 311601 2511900 4639702 5121799 4623199 4644301 5211600 7470501 4611700 4713001 6630300 1011201 4530702 4661300 4689399 4511101 1069400 8511100 4789099 113000 162899 213500 910600 1330800 1542300 1723000 1741901 1741902 1821000 2029100 2051700 2121000 2319200 2424502 2439300 2452100 2731700 2831300 2840200 2912201 2913001 2930102 3011301 3314718 3511500 3613701 4520001 4521700 4529205 4531402 4623106 4632002 4679603 5010501 5121704 5136599 5192600 5212500 5233701 5241803 5242603 5250804 5250805 6190699 6340103 7110200 8599699 8640299 4711302 2013400 4634601 4651601 4632001 4693100 5050400 4646001 4679699 5010502 4641901 4754701 8030600 8220200 4639701 4744005 4635499 6023201 1041400 2092402 2529102 4221902 5030003 7470502 5139099 4511103 4731800 2223400 2229302 2929701 5141102 5213201 8291100 4541203 4635402 4683400 7420902 7460802 8219999 1510600 4530701 1012103 1053800 1091100 1092900 1571702 1589099 2752900 2759799 2940801 2944100 4623109 4771702 8690999 8800600 5136502 4744001 2621300 8011101 2229301 2472400 4637107 4652400 4711301 6550200 5145401 5153599 5510801 2511000 2451200 2599399 4673700 8532500 2822402 7112000 1033301 1091101 1313800 1531901 1572500 1721300 2221800 2222503 2229202 2453800 2651500 2732500 2831200 2833900 2969600 2969601 3431200 3514000 3822000 4292801 4399104 4512901 4631100 4649404 4681805 4689301 4752100 4789004 5153502 5241804 8516299

9999999 5249399 4621400 1412601 2099199 2731600 2811800 4649408 4751200 5030001 5243401 2641702 2733300 3250705 3611001 4689302 5524701 2869100 1031700 2522500 2790299 2812700 5829800 6619399 2121101 1062700 2593400 3101200 161099 892403 1220499 1779500 1910000 2330302 2452001 2521700 2529103 2612300 2722801 2924601 4014200 4541202 4741500 4759899 7416002 2833000 4645101 1931301 155502 1552000 2399199 2941700 3432000 3512300 4322302 4513600 4674500 4687703 4744002 5131400 5161600 5244203 8411600 1095300 5245002 6399200 2710402 2740602 5244201 7732201 9199500 3102100 3250701 3612901 4722901 1351100 2522400 4530703 4671100 2421000 5010503

Códigos da CNAE pertenentes ao grupo cnD:

1412602 52222 115500 141501 144900 145702 150300 162799 710301 810006 1020102 1032599 1082100 1099601 1099602 1099605 1321801 1321900 1410999 1411801 1513001 1521100 1533500 1541500 1551200 1551201 1559800 1562801 1581400 1581401 1584900 1585700 1586500 1591101 1592000 1595401 1622602 1623400 1750701 1772800 1811202 1812001 1813099 1922502 1939900 2010901 2012600 2022300 2072000 2093200 2122900 2131800 2222502 2311700 2349401 2391503 2419800 2422901 2431700 2441502 2454600 2491000 2512800 2513600 2529101 2541100 2543800 2550102 2592602 2599302 2630101 2631100 2641701 2649299 2691301 2712099 2721900 2726001 2741302 2790201 2813500 2821602 2842800 2854200 2866600 2892401 2910701 2914901 2915700 2929700 2931901 2945000 2952100 2952101 2961001 3011302 3099700 3104700 3112701 3130500 3142901 3199200 3211602 3240099 3250703 3250706 3291400 3292202 3299004 3312102 3314799 3321000 3439800 3441000 3442800 3513100 3520401 3522000 3591200 3710999 4211102 4221904 4222701 4223500 4329199 4391600 4399199 4511106 4512801 4525001 4541100 4623101 4623105 4632003 4637102 4637103 4637105 4645102 4649402 4649403 4649409 4649499 4664800 4679601 4679604 4684201 4684202 4692300 4723700 4772500 4782201 4783101 4784900 4922102 4924800 5010507 5020201 5020202 5030101 5041503 5041504 5116000 5121703 5132203 5139004 5139008 5139009 5145403 5151902 5153503 5153507 5154301 5159402 5169101 5211799 5215902 5221301 5223000 5224800 5229099 5229999 5232900 5244204 5245001 5521201 5522000 5611201 5914600 6110803 6122002 6130200 6202300 6203100 6209100 6312601 6321599 6323199 6420392 6424703 6424704 6611804 6629100 6810202 6822600 6911701 7032700 7119701 7139099 7221400 7230300 7312200 7450001 7450002 7490102 7490199 7499303 7499399 7512400 7514000 7719599 7729202 8020900 8130300 8230001 8423000 8514604 8520100 8531699 8541400 8622400

8712300 9000099 9103100 9222301 9319101 9491000 1033302 1061902 1422300 2610800
4330401 4560800 4686902 4755501 4789005