

Leandro de Azevedo Gonzalez

Regressão Logística e suas Aplicações

São Luís

2018

Leandro de Azevedo Gonzalez

Regressão Logística e suas Aplicações

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Universidade Federal do Maranhão

Centro de Ciências Exatas e Tecnológicas

Curso de Graduação em Ciência da Computação

Orientador: Prof. Dr. Ivo José da Cunha Serra

São Luís

2018

Leandro de Azevedo Gonzalez

Regressão Logística e suas Aplicações/ Leandro de Azevedo Gonzalez. – São Luís, 2018

45 p.

Orientador: Prof. Dr. Ivo José da Cunha Serra

Monografia (Graduação) – Universidade Federal do Maranhão

Centro de Ciências Exatas e Tecnológicas

Curso de Graduação em Ciência da Computação, 2018.

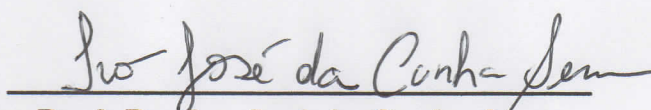
1. Mineração de Dados. 2. Regressão Logística. 2. Aplicações de Regressão Logística. I. Ivo José da Cunha Serra. II. Universidade Federal do Maranhão. III. Ciência da Computação. IV. Título

Leandro de Azevedo Gonzalez

Regressão Logística e suas Aplicações

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado. São Luís, 18 de janeiro de 2018:

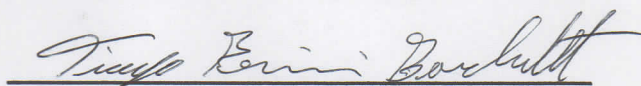


Prof. Dr. Ivo José da Cunha Serra

Orientador

Centro de Ciências Exatas e Tecnologia -
CCET

Universidade Federal do Maranhão – UFMA



Prof. Dr. Tiago Bonini Borchardt

Centro de Ciências Exatas e Tecnologia -
CCET

Universidade Federal do Maranhão – UFMA



Prof. Dr. Samyr Béliche Vale

Centro de Ciências Exatas e Tecnologia -
CCET

Universidade Federal do Maranhão – UFMA

São Luís

2018

Aos meus pais.

Agradecimentos

Em especial, aos meus pais Lorenzo Gonzalez Ruiz e Ercília Maria Menezes de Azevedo, que são os maiores incentivadores dos meus estudos, por todo o esforço dedicado para que eu tivesse uma boa educação.

Ao meu orientador, professor Ivo José da Cunha Serra, pela sua orientação, elogios, conselhos, confiança e empenho depositados em mim e no desenvolvimento deste trabalho. Agradeço também aos professores Tiago e Samyr, que gentilmente aceitaram o convite para a Banca Examinadora deste trabalho.

Ao corpo docente do curso de Ciência da Computação, pelo ensino e contribuição para a minha formação acadêmica, e à todos os professores pelos quais eu passei, desde o ensino infantil.

Aos familiares e amigos que estiveram presentes durante este processo. Aos colegas discentes do curso de Ciência da Computação, os quais dividimos as dificuldades e compartilhamos o aprendizado.

Resumo

Devido ao crescente volume de dados computacionais que são produzidos e armazenados, técnicas de mineração de dados tornam-se cada vez mais necessárias para a procura de padrões relevantes de informações nestes grandes volumes. Este trabalho descreve e analisa a regressão logística, que é uma técnica para mineração de dados de resposta categórica, nas suas formas binária e múltipla. São discutidos métodos tanto para a estimação do modelo de regressão, como para a avaliação do modelo gerado. Exemplos de aplicação da regressão logística são apresentados na área financeira, ambiental e epidemiológica, os quais mostram o possível uso desta técnica nestas diferentes áreas, e a destacam como uma forte ferramenta de análise de dados de resposta categórica, que possibilita a estimação da probabilidade de ocorrência de eventos, assim como a avaliação dos fatores que contribuem para o mesmo.

Palavras-chave: Mineração de Dados. Regressão Logística. Aplicações de Regressão Logística.

Abstract

Due to the increasing volume of computational data that is produced and stored, data mining techniques are becoming each more time, more required to the search of relevant information patterns in these large volumes. This paper describes and analyses the logistic regression, which is a technique for data mining of categorical response, in its binary and multiple forms. Methods are discussed both for the estimation of the regression model and for the evaluation of the model generated. Examples of the application of logistic regression are presented in the financial, environmental and epidemiological areas, which show the possible use of this technique in these different areas, and stand out as a strong tool of data analysis of categorical response, which allows estimation of the probability of occurrence of events, as well as the evaluation of the contributing factors to them.

Keywords: Data Mining. Logistic Regression. Logistic Regression Applications.

Lista de ilustrações

Figura 1 – Gráfico da função $\text{logit}(p)$	17
Figura 2 – Gráfico da função inversa do $\text{logit}(p)$	18
Figura 3 – Gráfico de pontos de dispersão desejado.	19
Figura 4 – Exemplo da saída do pseudo R^2 de Cox e Snell.	25
Figura 5 – Componentes da mudança ambiental global.	32
Figura 6 – Expansão agrícola na Bacia do Alto Paraguai de 1992 a 2000.	35
Figura 7 – Modelo logístico após a inserção dos coeficientes estimados.	36

Lista de tabelas

Tabela 1 – Coeficientes das variáveis do modelo de regressão	31
Tabela 2 – Uso e Cobertura do solo	33
Tabela 3 – Variáveis independentes incluídas no modelo logístico	36
Tabela 4 – OR obtido de 4 variáveis independentes do modelo	39

Lista de abreviaturas e siglas

BAP	Bacia Hidrográfica do Alto Paraguai
OR	Odds Ratio

Sumário

1	INTRODUÇÃO	12
1.1	Objetivos	13
1.1.1	Objetivos Específicos	13
1.2	Organização do Trabalho	13
2	REGRESSÃO LOGÍSTICA	14
2.1	Definição	15
2.2	A Função Logit	16
2.3	Regressão Logística Binária	19
2.3.1	Estimando os coeficientes do modelo de regressão	20
2.4	Regressão Logística Múltipla	21
2.5	Métodos de avaliação do modelo logístico	23
2.5.1	Teste da Razão de Verossimilhança	23
2.5.2	Teste de Wald	24
2.5.3	Pseudo R^2 de Cox e Snell	24
2.6	Considerações Finais	25
3	APLICAÇÕES DE REGRESSÃO LOGÍSTICA	27
3.1	Gestão de crédito	27
3.1.1	Risco de crédito	28
3.1.2	Regressão logística para análise de crédito	28
3.1.3	Aplicação Exemplo	30
3.2	Análise ambiental	31
3.2.1	Regressão logística na análise ambiental	34
3.2.2	Aplicação Exemplo	35
3.3	Óbito Neonatal	37
3.3.1	Regressão Logística no estudo do óbito neonatal	38
3.3.2	Aplicação Exemplo	38
3.4	Considerações Finais	40
4	CONCLUSÃO	41
	REFERÊNCIAS	43

1 Introdução

O presente trabalho apresenta a regressão logística, que consiste em uma técnica de mineração de dados. A mineração de dados é uma ferramenta de apoio para tomada de decisão baseada em dados computacionais, dos quais pretende-se extrair informações relevantes de uma grande base de dados, de forma a buscar vantagens competitivas ou elaboração de estratégia.

Estes conjuntos de grandes volumes de dados, representam um custo tanto de armazenamento quanto de processamento, portanto, é possível afirmar que a ideia de adquirir conhecimento automaticamente destes grandes volumes através da mineração de dados é altamente benéfico.

Este processo pode empregar algoritmos de inteligência artificial, análise estatística, recuperação de informação, reconhecimento de padrões e outros. Segundo [Hosmer e Lemeshow \(2000\)](#), métodos de regressão tem se tornado um componente integral para qualquer análise de dados interessada em descrever a relação entre uma variável resposta (dependente) e uma ou mais variáveis explicativas (independentes).

A principal diferença da regressão logística para a regressão linear é que a variável dependente na regressão logística é categórica, e de acordo com [Agresti \(2002\)](#), o modelo de regressão logística é o modelo mais importante para dados de resposta categórica.

Além da variável dependente ser categórica, ela é frequentemente binária (dicotômica), assumindo portanto, dois valores, que são normalmente tratados como “sucesso” ou “fracasso”. Estes dois valores representam um evento que depende do estudo de caso, podendo ser por exemplo: ou a concessão ou a não concessão de crédito na área financeira; ou a presença ou a não presença de uma doença nos estudos epidemiológicos; ou a compra ou a não compra de um determinado produto na área de marketing; entre outras diversas possibilidades.

Aplicações de análise de crédito, por exemplo, podem utilizar a regressão logística para calcular a probabilidade de um indivíduo ser merecedor da concessão de crédito. Informações pessoais como salário, emprego, tipo de moradia, são exemplos de possíveis variáveis que podem ser levadas em consideração. Isto permite não apenas a estimação dessa probabilidade, mas também como traçar o perfil de um bom ou mau pagador.

Há diversas técnicas de mineração de dados, cada uma com suas particularidades, a escolha da tarefa a ser utilizada é um importante passo para a obtenção de resultados satisfatórios dentro de um determinado estudo, isto só é possível com a compreensão da mesma e de sua capacidade.

Será abordado neste trabalho a técnica da regressão logística, seus conceitos, como é definido o modelo de regressão logística e como seus coeficientes são calculados, além de métodos que avaliam a qualidade do modelo obtido e exemplos de aplicações em diferentes domínios nos quais esta técnica pode ser aplicada.

1.1 Objetivos

O objetivo geral deste trabalho é, descrever o método de regressão logística como ferramenta de análise de dados de resposta categórica, e discutir possíveis usos por meio de exemplos de aplicações.

1.1.1 Objetivos Específicos

Dentro do objetivo geral, busca-se atender aos seguintes objetivos específicos:

- Apresentar o método de regressão logística e seus conceitos.
- Descrever as etapas do método de regressão logística
- Discutir exemplos de aplicações da regressão logística na área financeira, ambiental e epidemiológica, apontando seus resultados.

1.2 Organização do Trabalho

O restante deste trabalho está organizado da seguinte maneira. O capítulo 2 apresenta conceitos da regressão logística, sua definição, os dois tipos de regressão logística a serem abordados (binária e múltipla) e as etapas do processo da análise de regressão.

O capítulo 3 apresenta as áreas de gestão de crédito, análise ambiental e óbito neonatal, e nele é discutido a aplicabilidade da regressão logística em cada uma destas áreas.

Por fim, o capítulo 4 traz a conclusão deste trabalho, nele é feito um resumo do que foi apresentado, assim como é abordado as contribuições e relevância do tema tratado e as considerações finais.

2 Regressão Logística

A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis que poderiam, de outra forma, permanecer ignorados (TAN; STEINBACH; KUMAR, 2009).

Entre as tarefas de mineração de dados, está a tarefa de previsão, que consiste em prever o valor de um atributo com base nos valores de outros atributos. É denotado de variável dependente ou variável resposta, o atributo que se quer prever. Os atributos preditores, ou seja, aqueles usados para fazer a previsão são chamados de variáveis independentes ou explicativas.

A modelagem de previsão se refere à tarefa de construir um modelo para a variável dependente em função das variáveis independentes. A regressão logística é uma das técnicas que faz esta modelagem de previsão, cuja principal característica é o fato de sua variável dependente ser categórica e geralmente binária (dicotômica), representando por exemplo, 1 ou 0, sim ou não, falha ou sucesso, uma pessoa ter câncer ou não ter câncer, ou seja, indicando dois possíveis valores ou categorias.

A regressão logística difere de outras técnicas de mineração, principalmente pelo fato de sua variável dependente ser categórica, e mesmo quando ela não é dicotômica, é possível torná-la dicotômica, com a finalidade de aplicar esta técnica. Em relação as variáveis independentes, estas podem ser categóricas ou métricas.

É uma técnica que avalia a probabilidade de obtenção de uma das categorias da variável dependente, portanto, é capaz de obter a probabilidade de ocorrência de determinado evento, assim como a influência de cada variável independente no evento estudado.

Mesquita (2014) observa que, embora a regressão logística fosse inicialmente utilizada para área médica, a eficiência desta técnica viabilizou sua implementação nas mais diversas áreas do conhecimento, desde ciências médicas, a estudo de mercado, intenção de voto, avaliação de crédito e outras, expandindo assim sua aceitação entre os usuários de outras técnicas de mineração, se tornando uma ferramenta poderosa para análise de dados categóricos.

Neste capítulo, serão abordados dois tipos de regressão logística, a regressão logística binária e a múltipla. Veremos como obter os parâmetros essenciais para o modelo logístico e quais os testes que avaliam a significância de um modelo estimado.

2.1 Definição

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias.

Então, a partir desse modelo gerado é possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória.

Suponha que queira-se analisar a ocorrência da apneia do sono, que é um distúrbio do sono potencialmente grave, em que a pessoa para de respirar, por alguns segundos, diversas vezes durante a noite. Existem vários fatores que podem influenciar nesse distúrbio, mas para este exemplo, vamos considerar apenas dois: idade e peso. Digamos que para esta análise, tenhamos uma amostra de cem indivíduos, contendo a idade, o peso e se ele tem apneia ou não, este é o nosso conjunto de observações. A variável dependente é a ocorrência ou não da apneia do sono, ter apneia é igual a 1, não ter apneia é igual a 0. As variáveis independentes são a idade e o peso. Para este exemplo, o que a regressão logística propõe é que, a partir dessas informações, é possível gerar um modelo logístico que possa prever a probabilidade de uma pessoa ter apneia do sono, baseando-se no peso e idade desta pessoa. Mas como veremos a seguir, este é apenas um dos objetivos da regressão logística.

O modelo de regressão logística permite:

- modelar a probabilidade de um evento ocorrer dependendo dos valores das variáveis independentes, que podem ser categóricas ou contínuas.

Então digamos que, a partir do modelo logístico gerado do problema da apneia do sono, queiramos saber qual a probabilidade de um indivíduo de 50 anos e 120 quilos, ter ou vir a desenvolver a apneia do sono. Ao inserir os dados no modelo, o resultado será um valor entre 0 e 1 que representa esta probabilidade. Suponhamos que o valor seja 0,75, assim uma pessoa de 50 anos e 120 quilos tem 75% de probabilidade de ter apneia do sono.

- estimar a probabilidade de um evento ocorrer para uma observação selecionada aleatoriamente contra a probabilidade do evento não ocorrer.

Se uma pessoa de 50 anos e 120 quilos tem probabilidade $p = 0,75$ de ter apneia. A probabilidade de não ter apneia é $1 - p$, logo, $1 - p = 0,25$. A probabilidade de um evento ocorrer, contra ele não ocorrer, é uma razão de probabilidades, $\frac{p}{1 - p}$ que é chamada de chance. Assim temos $\frac{0,75}{0,25} = 3$, isto significa que uma pessoa nessas características tem 3 vezes mais chance de ter apneia do sono do que de não ter.

- prever o efeito do conjunto de variáveis sobre a variável dependente binária.

Através da análise de regressão logística, pode-se concluir, por exemplo, que a variável peso é bastante significativa para o modelo de regressão, enquanto que a variável idade não contribui tanto para a eficácia do mesmo.

- classificar observações, estimando a probabilidade de uma observação estar em uma categoria determinada.

A análise de regressão logística pode informar por exemplo, que indivíduos obesos, ou acima de uma determinada idade, podem ser mais propensos à esse distúrbio.

A variável dependente Y na regressão logística é frequentemente binária, logo, nestes casos ela segue a distribuição de Bernoulli (BELFIORE, 2015), tendo uma probabilidade desconhecida p . Lembrando que a distribuição de Bernoulli é apenas um caso especial da distribuição binomial, onde $n=1$ (considera a realização de um único experimento).

$$Y = \begin{cases} 1, & \text{se ocorrer sucesso} \\ 0, & \text{se ocorrer fracasso} \end{cases}$$

A probabilidade de sucesso é $0 \leq p \leq 1$ e a probabilidade de fracasso é $q = 1 - p$. Na regressão logística, é feita a estimação da probabilidade desconhecida p , dado uma combinação linear de variáveis independentes.

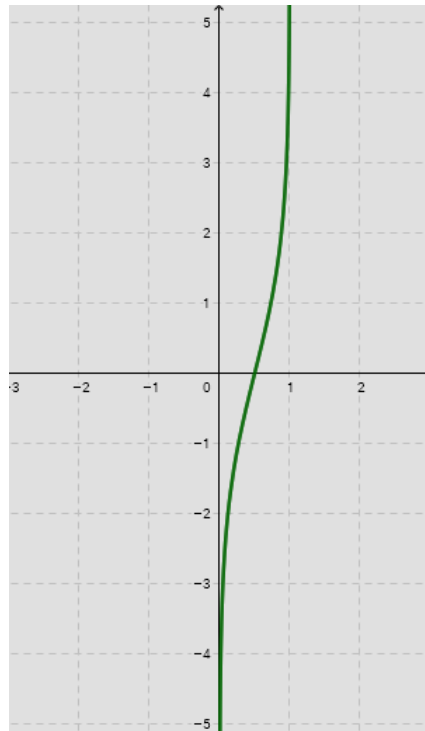
2.2 A Função Logit

Na seção anterior, foi dito que a variável dependente na regressão logística segue a distribuição de Bernoulli, portanto é preciso conectar as variáveis independentes à distribuição Bernoulli presente na variável dependente e esse *link* é chamado de logit. Na regressão logística nós não conhecemos a probabilidade p como é o padrão nos problemas de distribuição de Bernoulli. Logo, o objetivo do modelo logístico é estimar p para uma combinação linear das variáveis independentes. O p estimado é \hat{p} .

Logit = é o link entre a variável dependente binária e as variáveis independentes

Para ligar a combinação linear de variáveis à distribuição de Bernoulli, é necessário uma função que as una, ou mapeie a combinação linear de variáveis que poderiam retornar qualquer valor em uma distribuição de probabilidades bernoulli com um domínio de 0 a 1. A razão de probabilidade é chamada de chance ou *odds* em inglês, e seu logaritmo natural, o *logit*, é esta função representada na equação 2.1:

$$\ln(odds) \Rightarrow \ln\left(\frac{p}{1-p}\right) \quad (2.1)$$

Figura 1 – Gráfico da função $\text{logit}(p)$.

Fonte: Produzido pelo autor.

Pelo gráfico da função logit, na Figura 1, é possível compreendê-la melhor. A função vai a 0 mas não chega a tocar o eixo y, e o mesmo ocorre quando ela vai a 1. O que fica comprovado quando substituimos os valores na equação. Quando $p = 0$, $\ln(0/1) = \ln(0) = \text{indefinido}$. Quando $p = 1$, $\ln(1/0)$ é indefinido também. Ou seja, a função está dentro desse intervalo de 0 a 1, e quando estamos lidando com probabilidade, isto é algo muito útil, pois a probabilidade também é representada por valores dentro desse domínio. **Deste modo, pela função logística, nunca poderá se obter uma probabilidade superior a 100% ou inferior a 0%.**

Observando ainda a Figura 1, vejamos que quando $p = 0.5$ a função é 0. Substituindo o valor de p na função: $\ln(0.5/0.5) = \ln(1) = 0$. **Isso significa que quando as probabilidades são iguais, a chance(razão de probabilidades) é 1 e que o logit é 0.**

No gráfico da função logit, os valores entre 0 e 1, percorreram o eixo x, mas queremos que as probabilidades estejam no eixo y. Isto pode ser obtido através da inversa da função logit. A partir da equação(2.1), temos:

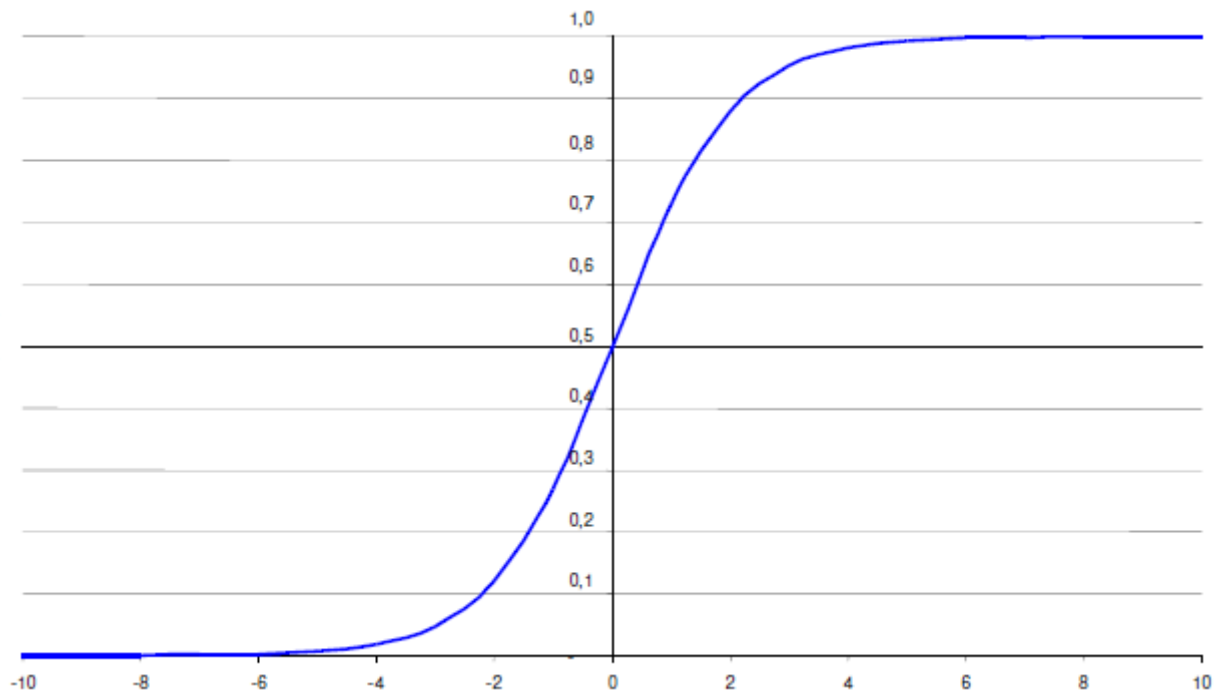
$$\text{logit}^{-1}(\alpha) = \frac{1}{1 + e^{-\alpha}} = \frac{e^{\alpha}}{1 + e^{\alpha}} \quad (2.2)$$

Adaptado de (MESQUITA, 2014)

α = combinação linear

No modelo de regressão logística, α , será a combinação linear das variáveis e seus coeficientes. A inversa da função logit retornará a probabilidade da variável dependente Y ser igual a 1 (o evento tal qual $Y = 1$, é tratado como o evento de interesse).

Figura 2 – Gráfico da função inversa do logit(p).



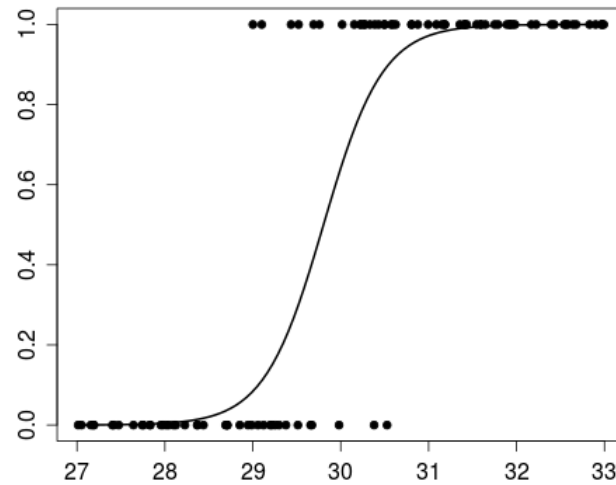
Fonte: Produzido pelo autor.

Na figura 2, observa-se que, o gráfico da inversa do logit é o mesmo do logit, apenas 90 graus invertido. Foi efetuada basicamente uma troca das coordenadas x e y , agora ao invés de ter o domínio da função de 0 a 1 no eixo x , temos o domínio de 0 a 1 no eixo y .

A representação gráfica da função inversa do logit na Figura 2, assume a forma parecida com um "S", também chamada de curva sigmóide, havendo áreas onde a mudança é acentuada e onde ela nem ocorre. As áreas onde pequenas variações nos valores de x causam grandes mudanças em valores de y representam áreas de maior probabilidade de mudança de estado da variável y em função de x .

Na figura 3, mostra-se o que seria um gráfico de pontos de dispersão da relação entre uma variável independente e a variável dependente na regressão logística. Os pontos que indicam a ocorrência e os pontos que indicam a não ocorrência do determinado evento, aparentam estar separados em grupos bem distintos e são poucos os pontos que aparecem sobrepostos. É possível observar como este gráfico se assemelha a curva da função da inversa do logit na Figura 2.

Figura 3 – Gráfico de pontos de dispersão desejado.



Fonte: <<https://goo.gl/nwec4Q>>

2.3 Regressão Logística Binária

A regressão logística binária ou univariada, representa os casos de regressão logística em que a variável dependente Y é binária ou dicotômica, ou seja, tem duas categorias e tem apenas uma variável independente. Tomemos como exemplo, um estudo de dosagens de determinada substância para a eutanásia de um animal. A variável dependente é dicotômica, sendo 1 para a morte do animal e 0 para a não ocorrência da morte. E neste caso temos apenas uma variável independente contínua, que seria a dose em ml por exemplo.

Digamos então, que:

$$g(x) = \beta_0 + \beta_1 x_1 \quad (2.3)$$

Fonte: (FIGUEIRA, 2006)

seja a função linear das variáveis independentes, sendo que β_0 e β_1 são os coeficientes e x_1 é a única variável independente, por se tratar da regressão logística univariada.

Vejamos novamente a equação logit(2.1) igualando à função $g(x)$.

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1$$

Entretanto, o objetivo do modelo logístico é estimar p , logo, é necessário isolar p . Pra isso utiliza-se o antilogaritmo:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1}$$

Seguindo com o procedimento para isolar p , obtemos:

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \quad (2.4)$$

Fonte: Adaptado de (MESQUITA, 2014)

A equação 2.4, é chamada de equação de regressão estimada, e é, essencialmente, a função que representa o objetivo do modelo de regressão logística, pois \hat{p} é a probabilidade estimada para quaisquer valores de coeficientes e variáveis que venhamos a colocar nesta equação. Os valores dos coeficientes são obtidos pelo método de estimação da máxima verossimilhança conforme explicado na subseção seguinte.

2.3.1 Estimando os coeficientes do modelo de regressão

Para ajustar um modelo de regressão, é necessário estimar os parâmetros β_0 e β_1 do modelo. Para isso utiliza-se o método de estimação da máxima verossimilhança. A partir dos dados da amostra, ou seja, o conjunto de observações, este método irá procurar os estimadores para o modelo de regressão logística, que são os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de máxima verossimilhança. A estimação por máxima verossimilhança, permite encontrar os estimadores dos parâmetros do modelo, que tem maior probabilidade de replicar o padrão de observações, nos dados da amostra.

Seja $\beta = (\beta_0, \beta_1)$ o vetor de coeficientes, e sejam as probabilidades $P(y_i = 1|x_i) = \pi(x_i)$ e $P(y_i = 0|x_i) = 1 - \pi(x_i)$. Então, para os pares (x_i, y_i) tais que $y_i = 1$, a contribuição para a função de verossimilhança é $\pi(x_i)$, e para os pares tais que $y_i = 0$, a contribuição para a função de verossimilhança é $1 - \pi(x_i)$, onde $\pi(x_i)$ denota o valor de $\pi(x)$ avaliado em x_i .

As equações de 2.5 à 2.8 foram retiradas de (FIGUEIRA, 2006). A função de verossimilhança é:

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (2.5)$$

Aplicando-se o logaritmo natural em ambos os lados da equação, obtemos a função log-verossimilhança:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^n [y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i))] \quad (2.6)$$

O valor β que maximiza $\ln[L(\beta)]$ é obtido após derivar $l(\beta)$ em relação aos parâmetros (β_0, β_1) :

$$\frac{\partial \ln[L(\beta)]}{\partial \beta_0} = \sum_{i=1}^n [y_i - \pi(x_i)] \quad (2.7)$$

$$\frac{\partial \ln[L(\beta)]}{\partial \beta_1} = \sum_{i=1}^n x_i [y_i - \pi(x_i)] \quad (2.8)$$

Os estimadores de (β_0, β_1) , denotados por $(\hat{\beta}_0, \hat{\beta}_1)$, são as soluções das equações (2.7) e (2.8) quando igualadas a 0. Estes estimadores dos parâmetros, medem a taxa de variação do *logit* para uma unidade de variação na variável independente, isto significa que eles são de fato, a inclinação da linha de regressão entre a variável dependente y_i e a sua variável independente x_i .

As fórmulas matemáticas permitem aos programas de regressão logística identificar mais eficazmente, os estimadores que maximizam a função log-verossimilhança. Dado que estas equações são não-lineares nos parâmetros, é necessário a utilização de um procedimento iterativo, como o Newton-Raphson. Este algoritmo escolhe, sucessivamente, novos conjuntos de parâmetros que produzam maiores log-verossimilhança e melhores ajustamentos aos dados observados. O processo continua através iterações ou ciclos repetitivos até a maximização da função log-verossimilhança.

Durante muitos anos, a estimação por máxima verossimilhança não foi utilizada por não existirem recursos computacionais, que pudessem realizar cálculos altamente complexos. Hoje, estes cálculos podem ser realizados através de programas estatísticos, como SPSS¹, SAS², entre outros.

2.4 Regressão Logística Múltipla

A regressão logística múltipla representa o contexto da regressão logística, em que a variável dependente Y é binária ou dicotômica, ou seja, tem duas categorias e que há mais de uma variável independente. Utilizando o exemplo citado na seção 2.3, temos que a variável dependente é a ocorrência da morte ou não de um animal, e a variável independente é a dose aplicada para a eutanásia, se adicionarmos uma variável independente como o peso do animal, este caso deixa de ser regressão logística binária e passa a ser de regressão logística múltipla.

Há uma semelhança grande com o que foi visto na seção anterior, e de fato, essa semelhança reflete nos modelos de parâmetros já mencionados, portanto, nesta seção, será

¹ <<https://www.ibm.com/br-pt/marketplace/spss-statistics#product-header-top>>

² <https://www.sas.com/pt_br/explore/analytics-in-action.html>

apenas alterado as equações já mencionadas na regressão logística binária, agora, de acordo com as características da regressão logística múltipla, porém, as suas funcionalidades para o modelo logístico são as mesmas. Podemos considerar então, a regressão logística múltipla como uma generalização da regressão logística binária.

Dado que neste contexto, há um conjunto de variáveis independentes, vamos considerar este conjunto denotado por $X = (x_1, x_2, \dots, x_t)$.

Equações de 2.9 à 2.11 retiradas de (BATISTA, 2015). A combinação linear para este conjunto de variáveis, é definida como:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_t x_t \quad (2.9)$$

Logo, o *logit* quando igualado à $g(x)$, é descrito na equação:

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_t x_t \quad (2.10)$$

Assim como na regressão logística binária, no caso múltiplo, utiliza-se o antilogaritmo na equação 2.10, para fins de isolar p , obtendo assim, o modelo de regressão logística múltipla para a probabilidade estimada \hat{p} :

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_t x_t}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_t x_t}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_t x_t)}} \quad (2.11)$$

A função de verossimilhança é a mesma da equação (2.5), com a diferença de que $\pi(x_i)$ é dado como $\pi(i)$, em função da equação 2.9, representando o conjunto de variáveis independentes em $g(x)$ e seus respectivos coeficientes. Logo, a função log-verossimilhança é obtida como:

$$l(\beta) = \sum_{i=1}^n [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)] \quad (2.12)$$

As expressões das equações a partir das derivadas parciais, são definidas pelas equações 2.12 e 2.13:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\pi}_i = 0 \quad (2.13)$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} y_i - \sum_{i=1}^n x_{ij} \hat{\pi}_i = 0, \text{ para } j \in \{1, \dots, t\} \quad (2.14)$$

Equações de 2.12 à 2.14 retiradas de (FIGUEIRA, 2006).

$\hat{\pi}_i$ é o estimador de π_i , pelo método da máxima verossimilhança.

2.5 Métodos de avaliação do modelo logístico

Após estimar os coeficientes, temos interesse em assegurar a significância das variáveis no modelo. Isto geralmente envolve formulação e teste de uma hipótese estatística para determinar se as variáveis independentes no modelo são significativamente relacionadas com a variável dependente. Para isto, há testes para avaliar o modelo logístico. Os testes mais utilizados são os testes da Razão da Verossimilhança, o teste de Wald e Pseudo R^2 de Cox e Snell (HOSMER; LEMESHOW, 2000; CRAMER, 2003; COX; SNELL, 1989). Veremos a seguir cada um deles.

2.5.1 Teste da Razão de Verossimilhança

Uma vez ajustado o modelo, é necessário testar a significância do modelo estimado. Isto pode ser feito através do teste da razão de verossimilhança.

Esta medida, testa simultaneamente, se os coeficientes de regressão associados a β são todos nulos com exceção de β_0 . As equações de 2.15 à 2.17 foram adaptadas de (CABRAL, 2013). A comparação entre os valores observados e esperados usando a função de verossimilhança, é expressa da seguinte forma:

$$D = -2 \ln \left[\frac{\text{verossimilhança do modelo ajustado}}{\text{verossimilhança do modelo saturado}} \right] \quad (2.15)$$

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right] \quad (2.16)$$

O modelo é dito saturado se contem todas as variáveis, enquanto o modelo ajustado corresponde ao modelo apenas com as variáveis desejadas para o estudo. Esta função D, também chamada de *deviance* (desvio), sempre é positiva e quanto menor, melhor é o ajuste do modelo.

Queremos testar as hipóteses:

$$H_0 : \beta_1 = \dots = \beta_t = 0 \text{ vs } H_1 : \exists_{j=1, \dots, p} \beta_j \neq 0$$

Assim, na hipótese nula H_0 a ser testada, os parâmetros do modelo serão igualados a 0. O modelo saturado que mantém o valor de seus coeficientes, representará a hipótese alternativa H_1 .

Para estimar a significância de uma variável independente, comparam-se o valor de D com e sem variável independente na equação. A alteração no valor de D esperada pela inclusão da variável independente no modelo é obtida através de:

$$G = D \left(\frac{\text{modelo sem a variável}}{\text{modelo com a variável}} \right) \quad (2.17)$$

Ao rejeitar a hipótese nula, tem-se que a variável independente testada, é significativa para o modelo.

2.5.2 Teste de Wald

O teste de Wald é também utilizado na regressão logística para a determinação da significância dos coeficientes do modelo estimado, ele testa se cada coeficiente é significativamente diferente de zero. Deste modo, o teste de Wald verifica se uma determinada variável independente possui uma relação estatisticamente significativa com a variável dependente.

Se os coeficientes logísticos forem estatisticamente significativos, podemos interpretá-los em termos de seu impacto na probabilidade estimada, deste modo, na predição do objeto de estudo no grupo respectivo, isto é, no grupo do evento de interesse ($Y = 1$), ou no grupo da não ocorrência do evento ($Y = 0$).

O teste de Wald é obtido comparando a estimativa de máxima verossimilhança de um coeficiente e a estimativa do seu erro padrão:

$$W_j = \frac{\hat{\beta}_j}{\text{var}(\hat{\beta}_j)} \quad (2.18)$$

Fonte: (BATISTA, 2015)

Hauck e Donner (1977 apud FIGUEIRA, 2006) e Jennings (1986 apud FIGUEIRA, 2006) examinaram o desempenho do teste de Wald e descobriram que, em alguns casos, ele se comporta de maneira inconsistente, falhando ao não rejeitar a hipótese nula mesmo quando o coeficiente é significante. Portanto, recomenda-se utilizar o Teste da Razão de Verossimilhança, quando há dúvidas de que o teste de Wald tenha falhado.

2.5.3 Pseudo R^2 de Cox e Snell

A estatística R^2 é uma medida que avalia em termos percentuais, a qualidade de um ajustamento de um modelo de regressão linear aos dados observados. Na regressão logística, não existe uma estatística que seja equivalente ao R^2 . No modelo de regressão linear, as variáveis dependentes são contínuas, o que não é o caso da regressão logística, onde a variável dependente é categórica. A denominação de pseudo R^2 deve-se ao fato de que eles se parecem com um R^2 do modelo de regressão linear, mas apesar dessa similaridade não podem ser interpretados da mesma forma como se interpreta um R^2 .

Há muitas maneiras diferentes de calcular o pseudo R^2 , alguns autores inclusive citam mais de 10 maneiras, mas infelizmente, não há um consenso sobre qual a melhor,

em geral, as técnicas de pseudo R^2 não são tão precisas quanto a estatística R^2 é para a regressão linear.

Neste trabalho, será citado apenas o pseudo R^2 de Cox e Snell (1989), por ser um dos mais frequentemente utilizados pelos softwares estatísticos. Segue a sua definição na equação 2.17:

$$R^2 = 1 - \left(\frac{L(\beta)_0}{L(\beta)_M} \right)^{\frac{2}{n}} \quad (2.19)$$

Fonte: (BATISTA, 2015)

Sejam n o tamanho da amostra, $L(\beta)_0$ o valor da função verossimilhança para um modelo sem preditores e $L(\beta)_M$ a verossimilhança do modelo sendo estimado. A racionalidade desta fórmula está no fato de que ela representa generalidade, uma vez que depende da probabilidade assumida pelos modelos com e sem preditores.

Figura 4 – Exemplo da saída do pseudo R^2 de Cox e Snell.

D	R^2 (Cox e Snell)	R^2 (Nagelkerke)
390,52	0,41	0,55

Fonte: FIGUEIRA (2006)

O pseudo R^2 de Cox e Snell, resulta em um valor que varia de 0 a 1, e geralmente é menor que 1, que indica a relação entre as variáveis independentes e a variável dependente. De maneira simples, é possível dizer que este valor informa o quanto as variáveis independentes explicam as variações da variável dependente, dado a base de dados observada. No exemplo da Figura 4, observa-se que este valor é baixo, indicando pouco mais de 40%.

Este mecanismo pode ser usado para comparar o desempenho de modelos concorrentes. Entre dois modelos logísticos, busca-se aquele que apresenta um pseudo R^2 mais elevado.

2.6 Considerações Finais

Foi dito neste capítulo, que a regressão logística é uma técnica de mineração de dados, pertencente às tarefas de previsão. A regressão logística de variável dependente dicotômica divide-se em duas, binária e múltipla, cuja diferença reside no número de

variáveis independentes. Foi explicado sua definição e conceitos, além das etapas desta técnica.

O método da máxima verossimilhança, para a estimação dos coeficientes do modelo de regressão, pode ser considerado pouco intuitivo devido ao seu processo iterativo, entretanto, os resultados obtidos através dele, são de fácil interpretação, pois resultam nos coeficientes que compõem o modelo.

Em relação aos métodos para avaliação do modelo de regressão gerado, ressalta-se que os testes presentes neste trabalho, representam uma parte dos mecanismos possíveis de avaliação. Na bibliografia de regressão logística, é possível encontrar estes e outros métodos com propósitos similares. Para o capítulo seguinte, será dado enfoque na aplicabilidade da regressão logística, através de exemplos de aplicação.

3 Aplicações de Regressão Logística

É difícil informar com precisão, quando a regressão logística foi aplicada pela primeira vez, mas [McLachlan \(1992 apud BITTENCOURT, 2003\)](#) afirma que as primeiras aplicações de regressão logística foram em estudos prospectivos de doenças coronárias. [Hosmer e Lemeshow \(1989\)](#) e [Cox e Snell \(1989\)](#) dizem que o modelo de regressão logística ganhou reconhecimento após o trabalho de [Truett, Cornfield e Kannel \(1967\)](#) que estudou o risco de doença coronária em um projeto chamado “*Framingham Heart Study*”. Este trabalho ganhou notoriedade e até hoje é considerado um marco inicial dos estudos envolvendo regressão logística nas áreas da saúde.

Desde então, a regressão logística tornou-se uma técnica padrão para análise de regressão de dados dicotômicos, principalmente nas ciências médicas, de acordo com [Hosmer e Lemeshow \(1989\)](#). Mas ela apresentou um crescimento muito rápido, se expandindo para outras áreas além da saúde, sendo utilizada também no campo da econometria, administração, educação, ambiental e outros.

Neste capítulo, serão apresentadas três áreas ou domínios nos quais a análise de regressão logística pode ser empregada. Será abordado o uso desta técnica na gestão de crédito, na análise ambiental e no estudo de óbitos neonatais, respectivamente.

3.1 Gestão de crédito

Crédito ao consumo, também conhecido por *revolving credit* ou crédito rotativo, é muito relacionado ao uso de cartões de crédito, e tem como característica o fato de seu reembolso à instituição financeira não ser determinado por um número fixo de parcelas ou pagamentos, afirma [Batista \(2015\)](#).

Ainda, segundo [Batista \(2015\)](#), o usuário do cartão de crédito pode utilizar ou retirar fundos da instituição de crédito, da qual é cliente, até um determinado limite de crédito, que lhe foi concedido previamente. O reembolso à instituição de crédito é efetuado através de pagamentos periódicos (totais ou parciais), acrescidos de juros. A periodicidade destes pagamentos é, normalmente, mensal e o seu valor está sujeito a um montante mínimo, em função do valor da dívida total em aberto. O limite de crédito determinado e concedido ao usuário do cartão de crédito, é calculado com base em alguns critérios, sendo estes, características do usuário, histórico de pagamentos anteriores e seu comportamento em relação à pontualidade e regularidade dos mesmos.

A concessão de crédito é atividade básica das instituições financeiras, entretando, no desenvolver deste negócio, os bancos estão expostos a diversos tipos de riscos, entre

eles o mais relevante é o risco de crédito ([FERREIRA; CELSO; NETO, 2012](#)).

3.1.1 Risco de crédito

No contexto de uma instituição financeira, podemos definir o crédito, como uma transferência de dinheiro em troca de uma promessa de restituição futura. As palavras ‘promessa’ e ‘futura’ dão uma indicação do que se trata o risco de crédito.

O simples ato de emprestar uma quantia ou algum item a alguém, envolve a possibilidade desta quantia ou item não ser recebido de volta, ou seja, há uma incerteza que o empréstimo seja devolvido. Isto é, basicamente, o risco de crédito. Podemos definí-lo então como o risco que um prestador ou credor enfrenta devido a possibilidade do devedor, em um acordo de concessão de crédito, não honrar seu compromisso.

A atividade de concessão de crédito é função básica dos bancos, portanto, uma boa gestão do risco de crédito é necessária, visto que este é um dos principais causadores de insolvência neste segmento econômico, observa [Ferreira, Celso e Neto \(2012\)](#).

Cabe à análise de crédito estimar o risco envolvido para a concessão ou não do crédito. O risco máximo que a instituição pode aceitar é inerente à política de cada empresa.

3.1.2 Regressão logística para análise de crédito

Os modelos de análise para concessão de crédito são intitulados , em inglês, de modelos de *credit scoring*, e baseiam-se em dados históricos da base de clientes existentes para avaliar se um futuro cliente terá mais chances de ser bom ou mau pagador, de acordo com [Gouvêa, Gonçalves e Mantovani \(2013\)](#).

Modelos que avaliam o crédito são de grande relevância para as instituições financeiras, dado que, um cliente bom classificado como mau desperdiça a chance de lucro da instituição, e um cliente mau classificado como bom, causa prejuízos.

Nenhum modelo consegue precisão absoluta, mas eles auxiliam na tomada de decisão da concessão de crédito e qualquer avanço na acurácia da previsão pode gerar ganhos financeiros para a instituição.

A análise de crédito envolve diversos fatores qualitativos e quantitativos como, por exemplo: o sexo do cliente, a idade, o valor da renda, o valor do patrimônio, a escolaridade e outros. A ideia por trás de um modelo de *credit scoring* é converter estas informações qualitativas e quantitativas dos clientes em uma pontuação que reflita a capacidade de pagamento de cada indivíduo. Com isto, busca-se segregar as características que permitam classificar um perfil de adimplência ou inadimplência.

Segundo [Camargos, Soares e Coutinho \(2012\)](#), as instituições financeiras no Brasil

passaram a utilizar maciçamente os modelos de *credit scoring* apenas em meados dos anos 90, pós estabilidade alcançada com a implantação do plano real.

Regressão logística é a técnica mais utilizada no mercado para modelos de *credit scoring* (CROOK; EDELMAN; THOMAS, 2007).

A seguir veremos os passos para se construir um modelo de *credit scoring* utilizando a regressão logística, segundo Gouvêa, Gonçalves e Mantovani (2013):

1. Levantamento de uma base histórica de clientes: os modelos são construídos com base em informações passadas e é importante que haja disponibilidade e qualidade desta base de dados para resultar em um modelo de sucesso.
2. Classificação dos clientes de acordo com a política da instituição e definição da variável dependente: deve-se notar que a definição de clientes bons e maus podem variar dependendo de cada instituição. E além de clientes bons e maus, tem-se aqueles estão na fronteira entre os dois, ou seja, não estão na posição nem de cliente bom e nem de cliente mau, portanto, estes em geral são desconsiderados do estudo, devido à maior facilidade de se trabalhar com a variável dependente dicotômica.
3. Seleção de uma amostra aleatória representativa da base histórica de clientes: é sugerido para a amostra aleatória que os casos das categorias da variável dependente, neste caso clientes bons e maus, tenham o mesmo tamanho para evitar possível viés devido à diferença de tamanho.
4. Análise descritiva e preparação dos dados: consiste em analisar, segundo critérios estatísticos, cada variável a ser utilizada no modelo.
5. Aplicação da regressão logística: a partir da amostra aleatória da base histórica e das variáveis a serem utilizadas no modelo, aplica-se a análise de regressão logística a fim de obter o modelo de regressão.

Neste cenário em questão, consideremos que um indivíduo possa ser classificado como cliente bom (bom pagador) ou cliente mau (mau pagador). Portanto a variável dependente binária Y pode assumir os valores:

$$Y = \begin{cases} 1, & \text{bom cliente} \\ 0, & \text{mau cliente} \end{cases}$$

A variável dependente determinada foi 1 para bons clientes e 0 para maus clientes, mas poderia ser o inverso. Independente da categoria que foi codificada como 1, a técnica de regressão logística oferece a obtenção dos mesmos resultados. O modelo de regressão logística obtido a partir desta técnica para a codificação proposta, permite o cálculo da

probabilidade de um cliente ser bom pagador. Para se obter a probabilidade dele ser um mau pagador, basta calcular a probabilidade complementar, ou seja, se a probabilidade de um cliente ser bom pagador for 0,7, a probabilidade dele ser mau pagador será 0,3.

Há uma série de características que podem ser incluídas como possíveis variáveis independentes, tais como: sexo, idade, estado civil, escolaridade, tipo de moradia (própria ou alugada), número de dependentes, valor da renda, valor do empréstimo, valor das parcelas, número de parcelas, situação de crédito (adimplente ou inadimplente) e outras. Destaca-se, portanto, que cada estudo oferece um resultado particular, pois depende do que está sendo considerado, da base histórica obtida, dos dados disponíveis e utilizados e da política de cada instituição.

Os resultados dos modelos de *credit scoring* servem como apoio à análise de crédito, pois possibilitam obter a probabilidade de ocorrência, ou não ocorrência da inadimplência, além de facilitar a identificação dos fatores que influenciam o risco da mesma. Cabe a cada organização avaliar as condições envolvidas na operação em conjunto com o resultado obtido no modelo. Estas informações dão suporte para minimizar a inadimplência e consequentemente, a perda do crédito.

3.1.3 Aplicação Exemplo

A fim de mostrar e analisar o uso da regressão logística para a elaboração de um modelo de análise de crédito, será apresentado a seguir um exemplo de aplicação. Todos os dados e resultados descritos nesta subseção são provenientes do estudo de [Ferreira, Celso e Neto \(2012\)](#), para análise de concessão de crédito de uma agência bancária localizada no município de Viçosa-MG.

De acordo com os referidos autores, a base amostral deste estudo foi composta pelas operações de crédito realizadas presencialmente pelos clientes, ou seus representantes legais no ano de 2007. Neste período, foram realizadas 82 operações de crédito na agência, das quais 74 foram utilizadas no estudo. Os dados utilizados para o estudo foram obtidos através da pesquisa em arquivos e banco de dados da carteira de crédito e no sistema fornecido pela agência bancária em estudo.

A variável dependente dicotômica foi a inadimplência, sendo atribuído $Y = 1$ aos clientes inadimplentes e $Y = 0$ aos clientes adimplentes. As variáveis independentes que tiveram significância na ocorrência de eventos de inadimplência nas operações de análise de crédito, segundo o modelo de regressão logística são: a idade do cliente, o tempo de relacionamento com o banco, a renda, o limite do cheque especial, o estado civil e a escolaridade.

Para a variável estado civil, foi considerado solteiro igual a 1, casado igual a 2, divorciado igual a 3 e viúvo igual a 4. Para a variável escolaridade, foi atribuído 1 para

clientes que cursaram até o ensino fundamental, 2 para os que cursaram até o ensino médio, 3 para quem tem superior incompleto, 4 para ensino superior completo e 5 para pós-graduação ou acima disto.

Tabela 1 – Coeficientes das variáveis do modelo de regressão

Variável	Coeficiente estimado
Idade Cliente (IC)	-1,68895
Tempo de relacionamento (TRL)	26,67029
Renda (RD)	0,05831
Limite cheque especial (LCH)	-0,09597
Estado Civil (EC)	29,65605
Escolaridade (ES)	-11,90192
Constante	-169,58336

Fonte: Adaptado de Ferreira, Celso e Neto (2012)

Os coeficientes estimados das variáveis contidas no modelo de regressão, obtidos através do *software* estatístico SPSS 13.5 utilizado pelos autores, estão apresentados na Tabela 1. O sinal dos coeficientes são importantes para o julgamento dos resultados. Ressalta-se que, um coeficiente positivo aumenta a probabilidade de ocorrência do evento de interesse, enquanto um coeficiente negativo diminui a probabilidade do mesmo. Neste cenário, o evento de interesse ($Y = 1$) é a inadimplência. A ‘Constante’ na tabela 1, representa o β_0 no modelo regressão. O modelo de regressão gerado pode ser observado na equação 3.1.

$$\hat{p} = \frac{1}{1 + e^{-(169,583 - 1,689 \times IC + 26,67 \times TRL + 0,0583 \times RD - 0,096 \times LCH + 29,656 \times EC - 11,902 \times ES)}} \quad (3.1)$$

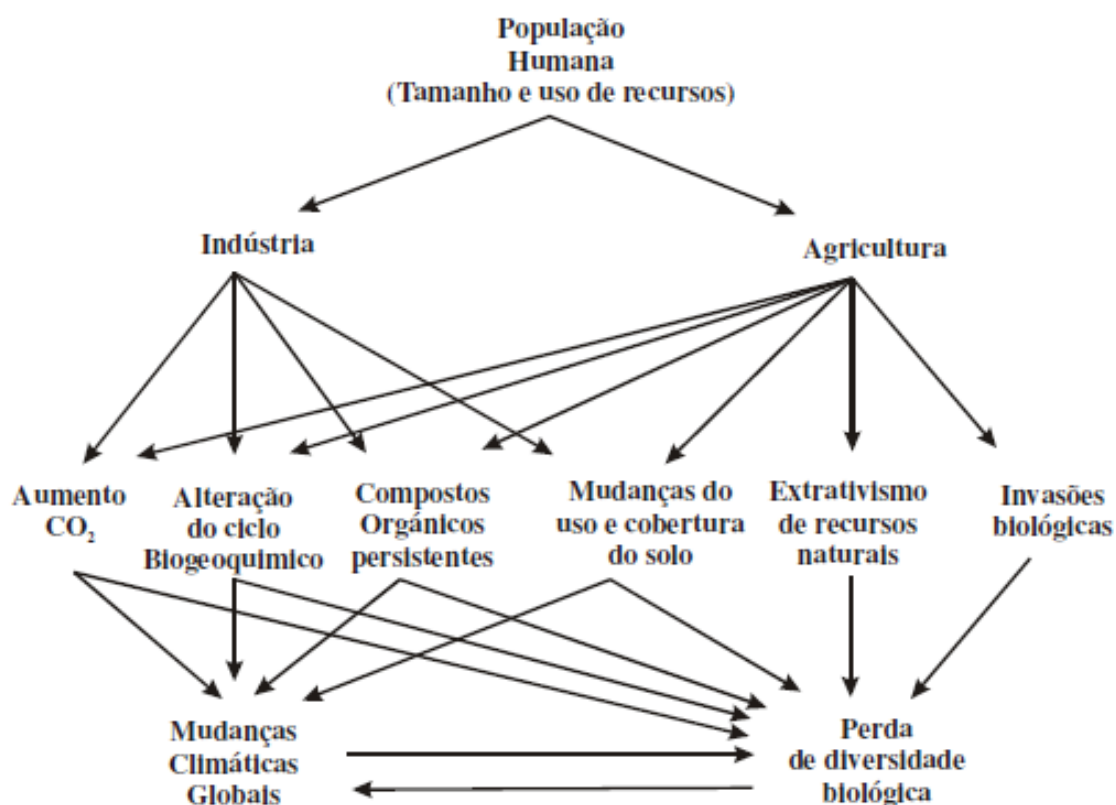
Com base nos resultados obtidos através da regressão logística, [Ferreira, Celso e Neto \(2012\)](#), fazem algumas observações acerca do trabalho realizado. Dentre estas, que, clientes de maior idade e maior grau de escolaridade tendem a ser adimplentes. Os autores observaram também alguns resultados inesperados, visto que clientes com maior renda e maior tempo de relacionamento com o banco apresentaram maiores problemas em relação à inadimplência.

3.2 Análise ambiental

O homem tem usado e modificado o solo há milhares de anos, a fim de seu próprio sustento e benefício. Através de atividades industriais, agropecuária, agricultura, urbanização, mineração, entre outras, o solo vem sendo utilizado e modificado para atender a população humana, ocasionando porém, o desgaste dos ecossistemas.

Fatores como o crescimento populacional e a alta demanda por alimentos e outros recursos, são determinantes para as elevadas taxas de intensidade e extensões das mudanças de uso e cobertura do solo a nível global. Estas mudanças são consideradas como uma das mais importantes alterações ambientais causadas por atividades humanas com efeitos diretos sobre as mudanças climáticas globais e perda de diversidade biológica, afirma [Valencia \(2008\)](#). A Figura 5 mostra mudanças ambientais globais causadas pela ação do homem.

Figura 5 – Componentes da mudança ambiental global.



Fonte: VALENCIA (2008)

Nos últimos 3 séculos, cerca de 12 milhões de km² de florestas e 5,6 milhões de km² de pastagens, tem sido convertidas a outros tipos de uso no mundo inteiro. Enquanto isso as áreas de cultivo tem aumentado em 12 milhões de km², segundo [Ramankutty e Foley \(1999 apud VALENCIA, 2008\)](#).

É possível definir o solo como o espaço de interação entre o homem e a natureza. O homem tem o solo como local da realização de suas atividades, utilizando os recursos naturais dele para a realização das mesmas.

A seguir, será definido o conceito de "uso do solo", "cobertura do solo" e "mudanças do uso do solo":

- Cobertura de solo

A cobertura de solo é descrita pelo estado biofísico da superfície. Isto é, características biológicas, físicas, químicas, ecológicas são os componentes das coberturas do solo. Exemplos de cobertura: florestas, mangue, bosques, áreas urbanas e outros.

- Uso do solo

O uso do solo pode ser definido pela finalidade do uso dos recursos do mesmo. Está relacionado aos produtos e benefícios obtidos do uso do solo como também do planejamento realizado no solo para alcançar estes produtos e benefícios. O uso do solo é caracterizado por planejamentos, atividades e insumos que as pessoas executam em uma cobertura de solo para produzi-lo, mudá-lo ou mantê-lo, de acordo com [FAO \(1998 apud VALENCIA, 2008\)](#).

Ressalta-se que cobertura de solo e uso de solo não são a mesma coisa. Diferentes usos de solo podem ser realizados em uma mesma cobertura de solo, assim como diferentes coberturas podem ser utilizadas para aplicar o mesmo uso de solo, como mostrado na Tabela 2.

Tabela 2 – Uso e Cobertura do solo

Cobertura do solo	Uso do solo
Florestas	Florestas naturais; Produção de madeira; Recreação
Pastagens	Pastagens naturais; Recreação; Criação de gado
Áreas agrícolas	Culturas perenes e anuais
Áreas urbanas	Cidades; Áreas industriais; Transporte

Fonte: VALENCIA (2008)

- Mudanças do uso do solo

O termo mudanças de uso do solo citado acima é utilizado aqui abrangendo tanto as mudanças de cobertura do solo, quanto as mudanças de uso do solo. O enfoque mais simples utilizado para mensurar mudanças do uso do solo considera o aumento ou redução da área da área coberta por um tipo de uso ou cobertura do solo. O uso do solo pode ser a causa da mudança de sua cobertura, ainda assim, a cobertura do solo pode mudar mesmo se mantendo inalterado seu uso.

[Mendes e Vega \(2011\)](#) observam que o potencial do impacto adverso dessas mudanças de uso do solo dependem de sua escala, extensão e das dinâmicas relações entre o uso do solo. Tem impactos diretos no solo, na água e na atmosfera, e portanto, está diretamente relacionado com muitas questões ambientais de importância global. [Lambin, Geist e Lepers \(2003 apud MENDES; VEGA, 2011\)](#) citam como exemplo de dinâmicas de alteração de

uso do solo, os desmatamentos em grande escala e posterior transformação em terras agrícolas na região dos trópicos, e destacam seus prováveis impactos sobre a biodiversidade, a degradação do solo e a capacidade produtiva da terra para satisfazer as necessidades humanas. Sendo assim, as dinâmicas de alteração do uso do solo são fatores determinantes no ciclo das mudanças climáticas (ilustrado na Figura 5), e portanto, a relação entre os dois é interdependente. As modificações no uso da terra podem modificar o clima, e por sua vez, as mudanças climáticas podem afetar o uso da terra.

3.2.1 Regressão logística na análise ambiental

Modelos que analisam as dinâmicas de uso do solo podem ser entendidos como um conjunto de técnicas utilizadas para descrever o processo de mudança de uso do solo em condições quantitativas e assim, proporcionar o entendimento deste processo.

Através da concepção deste modelo de análise da mudança de uso do solo, tem-se como objetivos, saber onde as mudanças de uso do solo ocorrerão e quais os fatores que mais contribuem para a ocorrência destas mudanças.

O método da regressão logística é uma ferramenta que pode ser utilizada para avaliar quais os fatores afetam, e em que extensão, a localização de usos do solo em uma determinada região a ser analisada. Consideremos a variável dependente dicotômica Y denominada de “Uso Agrícola”, onde 1 representa as áreas de uso agrícola e 0 as áreas sem uso agrícola.

$$Y = \begin{cases} 1, & \text{áreas com uso agrícola} \\ 0, & \text{áreas sem uso agrícola} \end{cases}$$

A variável dependente “Uso Agrícola” determinada, representa a distinção entre estes dois tipos de uso do solo. Aplicando os conceitos de regressão logística a essa variável, tem-se que o modelo de regressão pode estimar a probabilidade de uma área sem uso agrícola ser convertida ou transformada em uma área de uso agrícola, em função das variáveis independentes a serem consideradas na análise.

A técnica de regressão logística se mostra bastante adequada para a modelagem de transição de uso de solo, dado que estas passagens podem ser tratadas como estados ou categorias individuais (floresta para área rural, área rural para área industrial, por exemplo) e relacionadas a uma série de variáveis independentes. O papel desta técnica é encontrar o melhor modelo que relacione as variáveis dependente e independentes.

Pode-se assumir, de uma maneira geral, diversos fatores (variáveis independentes) que venham a influenciar na mudança do uso do solo, como por exemplo: precipitação média anual; temperatura média anual; densidade populacional. Assim como fatores

devido à proximidade de outros agente como: distância à rodovias, núcleos urbanos, limites florestais, entre outros.

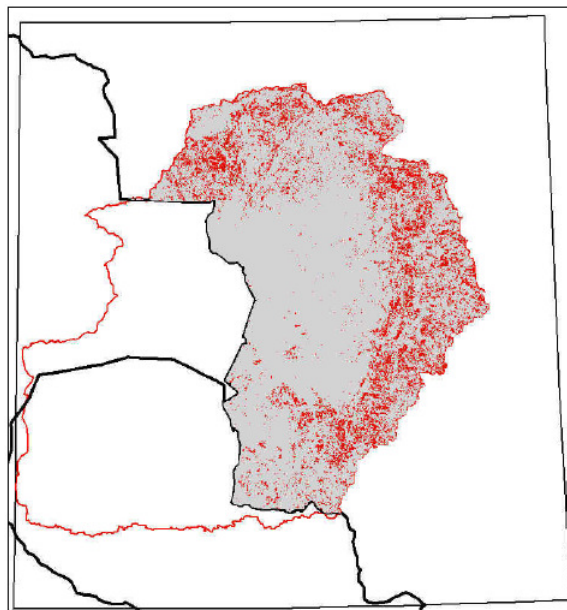
A probabilidade de ocorrência da transição de uma área não agrícola para uma área agrícola, representa a expansão da agricultura nesta determinada região. As variáveis independentes utilizadas indicam as mudança ocorridas entre as categorias de uso do solo (não agrícola para agrícola).

3.2.2 Aplicação Exemplo

O exemplo apresentado a seguir, é um estudo de caso aplicado na região da Bacia Hidrográfica do Alto Paraguai (BAP) em território brasileiro, para analisar a transição de uso de solo não agrícola para agrícola utilizando a técnica de regressão logística. Este estudo foi realizado por [Mendes e Vega \(2011\)](#).

Para a base amostral deste estudo, o autor considerou imagens da BAP correspondente a dois períodos, 1992 e 2000. As imagens originais foram reclassificadas de modo a gerar imagens binárias em relação a categoria agrícola (0 ou 1, para área não agrícola ou área agrícola, respectivamente). Através da sobreposição destas imagens binárias correspondentes aos dois períodos considerados, obtém-se uma nova imagem que representa a transição de áreas não agrícolas para agrícolas de 1992 a 2000, ilustrada na Figura 6. Os pontos na figura correspondem à essa área de transição.

Figura 6 – Expansão agrícola na Bacia do Alto Paraguai de 1992 a 2000.



Fonte: Mendes e Vega (2011)

Para o estudo em questão, a transição de não agrícola para agrícola corresponde

à variável dependente dicotômica do modelo, além disso, foi considerado um conjunto \mathbf{X} contendo 16 variáveis independentes, mas os autores informam que apenas 7 foram incluídas no modelo de regressão gerado, as demais foram descartadas devido à baixa significância para o modelo.

A tabela 3 informa as sete variáveis independentes incluídas no modelo logístico, com o código da variável e sua unidade de medida. O modelo gerado pela técnica de regressão logística pode ser observado na Figura 7.

Tabela 3 – Variáveis independentes incluídas no modelo logístico

Variáveis	Código	Unidades
Temperatura média anual	TMA	°C
Balanço hídrico climático médio anual	BHC	mm/ano
Distâncias para as sedes urbanas de municípios brasileiros inseridos na bacia	D_URB	Múltiplos de graus
Cotas topográficas na bacia pelo lado brasileiro	TOPO	m
Cotas topográficas reclassificadas (≤ 150 m , ≥ 150 m)	TOPO_REC	1-2
Distâncias para estradas principais (pavimentadas e federais) brasileiras inseridas na bacia	D_ESTRADA	Múltiplos de graus
Distâncias para as ferrovias brasileiras inseridas na bacia	D_FERRO	Múltiplos de graus

Fonte: Adaptado de Mendes e Vega (2011).

Figura 7 – Modelo logístico após a inserção dos coeficientes estimados.

$$P(X) = \frac{1}{1 + e^{\left(\begin{array}{l} 6.133 + 0.002*BHC - 0.335*TMA - 0.813*D_URB - 0.002*TOPO \\ - 0.700*D_ESTRADA + 0.315*D_FERRO + 0.708*TOPO_REC \end{array} \right)}}$$

Fonte: Mendes e Vega (2011)

Ressalta-se que, com exceção da variável ‘Cotas topográficas reclassificadas’, que é categórica, as demais variáveis independentes são contínuas. Os referidos autores observaram, através da técnica de regressão logística, que o desenvolvimento agrícola ocorre em áreas que são atrativas devido à sua proximidade com solos de boa qualidade existentes e facilidades de transporte para o escoamento da produção, entre outras observações. Afirmam ainda que seu estudo pretende reduzir incertezas para possibilitar um planejamento estratégico a fim de contemplar estas constantes mudanças ocorridas.

3.3 Óbito Neonatal

O período neonatal corresponde às quatro primeiras semanas de vida (0 a 28 dias incompletos). Denomina-se período neonatal precoce, a primeira semana completa ou os sete primeiros dias de vida, e período neonatal tardio, as três semanas seguintes.

Óbito neonatal é o óbito que ocorre no período neonatal, isto é, entre 0 e 28 dias incompletos após o nascimento. A criança morta dentro deste período, dá-se o nome de neomorto.

A mortalidade infantil pode ser considerada um dos melhores indicadores da qualidade da assistência à saúde, bem como do nível socioeconômico de uma população. Tal índice compreende todos os óbitos de crianças com menos de um ano de idade, sendo formada pelo óbito neonatal e o pós-neonatal, que abrange os óbitos ocorridos do 28º dia até um dia antes de se completar um ano de vida, segundo [Risso e Nascimento \(2011\)](#).

[Zupan e Aahman \(2005 apud ZANINI et al., 2011\)](#) informa que, entre os 130 milhões de crianças que nascem no mundo anualmente, cerca de 4 milhões morrem no período neonatal, proporção que varia de acordo com a taxa de mortalidade global. A variação no risco diário de morte é considerável e esse risco é maior na primeira semana de vida. A maioria dos óbitos neonatais (99%) ocorre em regiões com renda baixa ou média. Crianças que nascem em países mais pobres apresentam maior risco de morte, com taxa de mortalidade neonatal de 19% a 44% superior em famílias pobres ([KNIPPENBERG et al., 2005](#)).

No Brasil, a taxa de mortalidade infantil teve redução de 50% entre 1990 e 2008. Mas apesar das taxas de mortalidade infantil estarem em queda, os dados indicam concentração dos óbitos no período neonatal, que ainda se mantém com valores elevados em comparação com as taxas de mortalidade pós-neonatal.

Para tentar reduzir estes valores elevados da mortalidade no período neonatal, técnicas como a regressão logística podem ser utilizadas para que se construa modelos que possibilitem identificar os fatores de risco para o óbito neonatal.

3.3.1 Regressão Logística no estudo do óbito neonatal

Modelos de análise clássicos pressupõem independência entre indivíduos e homogeneidade de variância, e desconsideram a hierarquia dos fatores preditores, isto é, não consideram que observações originadas de uma mesma unidade podem ser mais similares do que aquelas originadas de diferentes unidades. Isso pode levar à superestimação dos efeitos do agrupamento e induzir a conclusões imprecisas (GOLDSTEIN, 2003 apud ZANINI et al., 2011).

A análise de regressão logística é uma alternativa aos modelos clássicos ao considerar a variável dependente, em nível dicotômico, e as variáveis independentes, em qualquer nível, categórico ou contínuo. Esses modelos permitem analisar o efeito das categorias separadamente e fornecem informação sobre a influência da composição dos fatores, segundo Goldstein (2003).

Assim, para facilitar a identificação e a compreensão dos fatores associados ao óbito neonatal, pode-se dizer que a regressão logística é uma técnica apropriada.

A variável dependente binária Y é a ocorrência ou não ocorrência do óbito em crianças com menos de 28 dias de vida.

$$Y = \begin{cases} 1, & \text{ocorrência do óbito neonatal} \\ 0, & \text{não ocorrência do óbito neonatal} \end{cases}$$

Características da mãe e da criança, assim como características socioeconômicas, são analisadas como determinantes da mortalidade infantil, algumas das mais utilizadas em trabalhos neste contexto são: peso ao nascer, se a criança é pré-termo, escore de Apgar no 1º e 5º minuto, idade da mãe, escolaridade materna, tipo de parto (normal ou cesária), tipo de gestação (única ou múltipla), tabagismo, entre outros, renda familiar, número de pessoas que moram no mesmo domicílio, entre outros.

O uso de um modelo para avaliar os fatores de risco para a mortalidade infantil neonatal, compreendidos como indicadores de várias dimensões das condições de vida, é importante para compreender o quanto estes indicadores influenciam na ocorrência do óbito neonatal. Sendo assim possível, identificar grupos expostos a diferentes fatores de risco e detectar necessidades de saúde em diferentes subgrupos populacionais. Isto aumenta a esperança de que estes fatores possam ser minimizados e talvez até evitados.

3.3.2 Aplicação Exemplo

Martins e Velásquez-Meléndez (2004) testaram a associação de vários fatores com a mortalidade neonatal em Montes Claros, utilizando a técnica de regressão logística. Todos os dados e resultados apresentados nesta subseção foram obtidos pelos referidos autores.

De acordo com os autores, a população constituiu-se de 20.506 nascidos vivos na cidade de Montes Claros, MG, Brasil, entre o período de 1 de janeiro de 1997 a 31 de dezembro de 1999. Foi verificado banco de dados de óbitos e de nascimentos para identificar os nascidos vivos que evoluíram para o óbito neonatal, no qual verificou-se 275 casos neste período. Após a verificação de registros com variáveis com valor omissos, foram excluídos 1491 registros, portanto a base amostral utilizada no estudo totalizou 19.015 registros.

A variável dependente do estudo foi a ocorrência de óbito neonatal (1 para a ocorrência do óbito, 0 para a não ocorrência do óbito), e as variáveis independentes estão relacionadas ao recém-nascido (sexo, peso ao nascer, escore de Apgar no 1º e 5º minutos de vida e idade gestacional), à gestação e parto (tipo de gravidez e parto, número de consultas de pré-natal e local do nascimento) e à mãe (grau de instrução, idade, filhos e abortos tidos). Segundo os autores, todas as associações entre os preditores e a variável dependente foram consideradas estatisticamente significantes.

Sobre o escore de Apgar, ele é um teste de avaliação de cinco sinais vitais do recém-nascido realizado no primeiro, quinto e décimo minuto após o nascimento. A pontuação varia de 0 a 10, e quanto mais próximo de 10, melhor.

Martins e Velásquez-Meléndez (2004) apresentaram os resultados do estudo através dos valores de *odds ratio* (OR) obtidos pelo método da regressão logística. A Tabela 4 informa o OR de quatro das variáveis independentes incluídas neste estudo.

Tabela 4 – OR obtido de 4 variáveis independentes do modelo

Variáveis	OR
Peso ao nascer (1 = baixo peso; 0 = peso normal)	4,94
Idade gestacional (≤ 37 sem. (1); > 37 sem. (0))	5,68
Apgar 1º minuto (0 a 10)	0,75
Apgar 5º minuto (0 a 10)	0,76

Fonte: Adaptado de Martins e Velásquez-Meléndez (2004).

O *odds ratio*, ou razão de chances em português, pode ser definido como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo. Na regressão logística, esta razão aparece diretamente relacionada aos coeficientes das variáveis independentes, favorecendo a interpretação dos resultados obtidos.

Uma razão de chances de valor igual a 1, indica que o evento em estudo tem chances iguais de ocorrer nos dois grupos, maior que 1 indica chance maior de ocorrer no primeiro grupo, e entre 0 e 1, indica chance menor de ocorrer no primeiro grupo.

Trazendo este conceito para o estudo em questão, pela Tabela 4, tem-se que a variável “peso ao nascer” é binária (1 = baixo peso; 0 = peso normal), portanto, a leitura do valor do OR indica que recém-nascidos com baixo peso tem 4,94 mais chances de evoluir

à óbito do que recém-nascidos com peso normal. Da mesma forma, crianças pré-termo (nascidas com menos de 37 semanas de gestação) tem 5,68 mais chances de evoluir à óbito em comparação às crianças a termo.

As variáveis “Apgar 1º minuto” e “Apgar 5º minuto” são discretas e variam de unidade a unidade até 10, neste caso, a chance de um grupo é comparada à chance do grupo de unidade anterior. Observa-se ainda que elas possuem OR abaixo de 1, isto significa, por exemplo, que uma criança com score Apgar igual a 7 tem 0,75 menos chance de evoluir à óbito em comparação com uma criança com score Apgar igual a 6.

Entre outras observações, [Martins e Velásquez-Meléndez \(2004\)](#) destacam que o sexo e o tipo de parto não estiveram relacionados com a mortalidade neonatal, e o fator pré-termo foi a variável que apresentou maior associação com este tipo de óbito.

3.4 Considerações Finais

Este capítulo concentrou-se na aplicabilidade da regressão logística, e foi dividido em três seções para apresentar áreas de aplicações desta técnica, na área de gestão de crédito, de análise ambiental e de óbito neonatal, respectivamente. Para cada um destes três exemplos de aplicação, foi apresentado o domínio junto com um problema a ser estudado, a forma como a regressão logística pode ser utilizada no determinado cenário e um estudo de caso realizado utilizando esta técnica.

A apresentação destes três exemplos mostra como a regressão logística pode ser usada para diferentes fins, complementando o que foi explicado na seção 2.1 do capítulo anterior. Enquanto que o objetivo principal da análise de crédito é o modelo de regressão gerado, o estudo do óbito neonatal foca no impacto das variáveis independentes sobre o evento em estudo, e o exemplo da análise ambiental divide a atenção entre estes dois objetivos. O capítulo seguinte, traz a conclusão do trabalho, que apresenta as considerações finais, limitações e trabalhos futuros.

4 Conclusão

O desenvolvimento do presente estudo possibilitou uma análise do método de regressão logística. Foi definido o que é este método e o que ele permite extrair de informações através do processo para obtenção do modelo probabilístico. Foi explicado que ele permite não apenas estimar a probabilidade de ocorrência de um evento, mas também analisar as variáveis incluídas no processo, e classificar ou traçar um perfil dos grupos que assumem os valores da variável dependente.

No capítulo 2, mostrou-se que a função logit é a base para a obtenção do modelo de regressão logística. Foram abordados dois tipos de regressão logística, a binária e a múltipla, tal que, a segunda pode ser vista como uma generalização da primeira. O método da máxima verossimilhança serve para estimar os coeficientes do modelo de regressão logística, enquanto métodos como o teste da razão de verossimilhança e o teste de Wald são utilizados para verificar a significância dos coeficientes estimados. Já o método Pseudo R^2 é um mecanismo utilizado para constatar o desempenho do modelo gerado.

No capítulo 3, foram apresentados três exemplos de aplicação utilizando a regressão logística, na área financeira, ambiental e epidemiológica. Para cada exemplo foi abordado o domínio da respectiva área com uma situação problema a ser estudada, o possível uso da regressão logística no estudo deste problema e uma aplicação exemplo apontando os resultados obtidos.

A contribuição deste método é grande, e o exemplo de análise de crédito apresentado neste trabalho evidencia isto. A regressão logística é o método mais utilizado para o desenvolvimento de modelos de análise de crédito, e estes modelos são amplamente utilizados não apenas por instituições bancárias, mas também por outras empresas que atuam com concessão de crédito a seus clientes.

O exemplo da regressão na análise ambiental aponta a variabilidade deste método e o uso da regressão logística no estudo do óbito neonatal é apenas um dos vários realizados nas ciências médicas, que em geral, destacam não apenas a contribuição para a prevenção de uma doença ou óbito, mas para o estudo das questões socioeconômicas envolvidas.

É interessante ressaltar que, a regressão logística, assim como qualquer outro método computacional, tem suas particularidades e portanto, não há um método que seja indicado para todo e qualquer propósito. No caso específico desta técnica, sua principal característica que é a variável dependente categórica, é ao mesmo tempo, um limitador, caso queira-se por exemplo, prever a variação de uma ação na bolsa de valores, a regressão linear é provavelmente uma melhor opção por tratar a variável resposta como contínua. A relevância deste trabalho está em apresentar conceitos e aplicabilidade da regressão

logística de forma a auxiliar a compreensão e uso da mesma como uma potente ferramenta de análise de dados de resposta categórica.

Neste trabalho não foram abordados, os métodos de seleção das variáveis preditoras no modelo gerado, como os métodos *forward stepwise* e *backward stepwise*. E, como foi dito no capítulo 2, há também outros métodos de avaliação do modelo de regressão além dos três abordados neste trabalho, o teste de Hosmer e Lemeshow e as curvas ROC são exemplos de outros métodos. A partir disto, sugere-se para um estudo posterior, abordar a regressão logística com base nesses conceitos.

Outra sugestão para trabalhos futuros é que seja abordado o tipo de regressão logística na qual a variável resposta pode assumir mais de dois valores, ou seja, a variável é politômica, denotada por regressão logística multinomial. Neste contexto, um estudo epidemiológico, por exemplo, pode tratar a variável resposta como os diferentes estágios de uma doença (inicial, moderado, avançado) e relacionar as variáveis presentes a cada um destes grupos, entre outras possibilidades.

Referências

- AGRESTI, A. *Categorical Data Analysis*. 2nd ed. New Jersey: John Wiley & Sons, 2002. Citado na página 12.
- BATISTA, A. M. S. *Regressão Logística: Uma Introdução ao Modelo Estatístico – Exemplo de Aplicação ao Revolving Credit*. Porto: Vida Económica, 2015. Citado 4 vezes nas páginas 22, 24, 25 e 27.
- BELFIORE, P. *Estatística: aplicada a administração, contabilidade e economia com Excel e SPSS*. 1. ed. Rio de Janeiro: Elsevier, 2015. Citado na página 16.
- BITTENCOURT, H. R. Regressão logística politômica: revisão teórica e aplicações. *Acta Scientiae*, Canoas, v. 5, n. 1, p. 77–86, 2003. Citado na página 27.
- CABRAL, C. I. S. *Aplicação do Modelo de Regressão Logística num Estudo de Mercado*. Dissertação (Mestrado) — Universidade de Lisboa, Lisboa, 2013. Citado na página 23.
- CAMARGOS, M. A.; SOARES, G. O. G.; COUTINHO, E. S. Determinantes do rating de crédito de companhias brasileiras. *Revista Contabilidade Vista & Revista*, v. 23, n. 3, p. 109–143, 2012. Citado na página 28.
- COX, D. R.; SNELL, E. J. *Analysys of Binary Data*. 2nd ed. London: Chapman & Hall, 1989. Citado 3 vezes nas páginas 23, 25 e 27.
- CRAMER, J. S. *Logit Models From Economics and Other Fields*. Cambridge: Cambridge University Press, 2003. Citado na página 23.
- CROOK, J. N.; EDELMAN, D. B.; THOMAS, L. C. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, v. 183, n. 3, p. 1447–1465, 2007. Citado na página 29.
- FERREIRA, M. A. M.; CELSO, A. S. dos S.; NETO, J. E. B. Aplicação do modelo logit binomial na análise do risco de crédito em uma instituição bancária. *Revista de Negócios*, Blumenau, v. 17, n. 1, p. 41–59, 2012. Citado 3 vezes nas páginas 28, 30 e 31.
- FIGUEIRA, C. V. *Modelos de Regressão Logística*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006. Citado 4 vezes nas páginas 19, 20, 22 e 24.
- FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. *Terminology for Integrated Resources Planning and Management*. [S.l.], 1998. Citado na página 33.
- GOLDSTEIN, H. *Multilevel statistical models*. 3rd ed. London: Edward Arnold, 2003. Citado na página 38.
- GOUVÊA, M. A.; GONÇALVES, E. B.; MANTOVANI, D. M. N. Análise de risco de crédito com aplicação de regressão logística e redes neurais. *Revista Contabilidade Vista & Revista*, Belo Horizonte, v. 24, n. 4, p. 96–123, 2013. Citado 2 vezes nas páginas 28 e 29.

- HAUCK, W. W.; DONNER, A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, v. 72, p. 851–853, 1977. Citado na página 24.
- HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. 1st ed. New York: John Wiley & Sons, 1989. Citado na página 27.
- HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons, 2000. Citado 2 vezes nas páginas 12 e 23.
- JENNINGS, D. E. Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, v. 81, p. 471–476, 1986. Citado na página 24.
- KNIPPENBERG, R. et al. Systematic scaling up of neonatal care in countries. *Lancet*, p. 1087–1098, 2005. Citado na página 37.
- LAMBIN, E. F.; GEIST, H. J.; LEPEERS, E. Dynamics of land–use and land–cover change in tropical regions. *Annual Review of Environment and Resources*, v. 28, p. 205–241, 2003. Citado na página 33.
- MARTINS, E. F.; VELÁSQUEZ-MELÉNDEZ, G. Determinantes da mortalidade neonatal a partir de uma coorte de nascidos vivos, montes claros, minas gerais, 1997–1999. *Rev. Bras. Saúde Matern. Infant.*, Recife, v. 4, n. 4, p. 405–412, 2004. Citado 3 vezes nas páginas 38, 39 e 40.
- MCLACHLAN, G. *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons, 1992. Citado na página 27.
- MENDES, C. A. B.; VEGA, F. A. C. Técnicas de regressão logística aplicada à análise ambiental. *Revista Geografia*, Londrina, v. 20, n. 1, p. 5–30, 2011. Citado 2 vezes nas páginas 33 e 35.
- MESQUITA, P. S. B. *Um Modelo de Regressão Logística para Avaliação dos Programas de Pós-Graduação no Brasil*. Dissertação (Mestrado) — Universidade Estadual do Norte Fluminense, Campo dos Goytacazes, 2014. Citado 3 vezes nas páginas 14, 17 e 20.
- POWERS, D. A.; XIE, Y. *Statistical Methods for Categorical Data Analysis*. Austin: Academic Press, 1999. Citado na página 16.
- RAMANKUTTY, N.; FOLEY, J. A. Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles*, v. 13, n. 4, p. 997–1028, 1999. Citado na página 32.
- RISSO, S. de P.; NASCIMENTO, L. F. C. Fatores de risco para óbito neonatal obtidos pelo modelo de regressão multivariado de cox. *Rev. Paul. Pediatr.*, v. 29, n. 2, p. 208–213, 2011. Citado na página 37.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining: Mineração de Dados*. 1. ed. Rio de Janeiro: Ciência Moderna, 2009. Citado na página 14.
- TRUETT, J.; CORNFIELD, J.; KANNEL, W. A multivariate analysis of the risk of coronary heart disease in framinghan. *Journal of Chronic Diseases*, v. 20, p. 511–524, 1967. Citado na página 27.

VALENCIA, L. I. O. *Enfoque da estatística espacial em modelos dinâmicos de mudança do uso do solo*. Dissertação (Mestrado) — Universidade Estadual do Rio de Janeiro, Rio de Janeiro, 2008. Citado 2 vezes nas páginas 32 e 33.

ZANINI, R. R. et al. Determinantes contextuais da mortalidade neonatal no rio grande do sul por dois modelos de análise. *Revista de Saúde Pública*, São Paulo, v. 45, n. 1, p. 79–89, 2011. Citado 2 vezes nas páginas 37 e 38.

ZUPAN, J.; AAHMAN, E. Perinatal mortality for the year 2000: estimates developed by who. *Geneva: World Health Organization*, p. 129–33, 2005. Citado na página 37.