UNIVERSIDADE FEDERAL DE MINAS GERAIS
Curso de Especialização em Estatística

Previsão da Inadimplência através da Regressão Logística

Belo Horizonte
2015

Cláudia Costa Vieira Paiva

Previsão da Inadimplência através da Regressão Logística

Monografia apresentada ao curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística.

Orientador: Gregório Saravia Atuncar

Belo Horizonte
2015

Cláudia Costa Vieira Paiva

Previsão da Inadimplência através da Regressão Logística

Monografia apresentada ao curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística.

_____
Prof. Gregório Saravia Atuncar (Orientador) – UFMG

_____
Profa. Ela Mercedes Medrano de Toscano – UFMG

_____
Profa. Sueli Aparecida Mingoti - UFMG

Belo Horizonte, 5 de dezembro de 2015

Dedico este trabalho à minha família, ao meu orientador e aos meus colegas de trabalho que não mediram esforços para me ajudar e me deram forças para continuar, mesmo nos momentos mais difíceis.

AGRADECIMENTOS

Ao Divino Espírito Santo, que me iluminou para a realização deste trabalho.

À minha família, pela compreensão, força e carinho.

Aos professores da UFMG, que me ensinaram tanto durante todo o curso.

RESUMO

Com o crescimento progressivo nos volumes de concessão de crédito às micro e pequenas empresas, as instituições financeiras estão procurando, cada vez mais, agilidade e assertividade na concessão do crédito. Como, na concessão do crédito, existe a possibilidade de perda, a possibilidade de se estimar a probabilidade de ocorrência desta perda torna o processo de decisão do crédito mais confiável.

Assim, o objetivo deste trabalho é propor a utilização de um modelo de regressão logística para estimar esta probabilidade de perda. As etapas para construção do modelo são discutidas detalhadamente, sendo abordado desde o planejamento para escolha das variáveis até o diagnóstico de ajuste do modelo.

Ao final do trabalho é apresentado um estudo de caso, com a elaboração do modelo de regressão logística para prever a probabilidade de inadimplência em clientes, micro e pequenas empresas, de uma instituição financeira. Os resultados deste modelo específico foram analisados e considerados eficientes para auxiliar na decisão de concessão do crédito para este perfil de clientes.

ABSTRACT

With the progressive growth in volumes of lending to micro and small companies, financial institutions are looking increasingly for agility and assertiveness in granting credit. As the loans are granted, there is a possibility of loss, the ability to estimate the probability of occurrence of this loss makes the most reliable credit decision process.

The objective of this work is to propose the use of a logistic regression model to estimate this probability of loss. The steps to build the model are discussed in detail, being approached from planning the choice of variables to the diagnostic model fit.

At the end of work is presented a case study with the preparation of the logistic regression model to predict the probability of default customers, micro and small companies, a financial institution. The results of this particular model were analyzed and found to be effective to aid in the decision to grant credit for this customer profile.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

As primeiras aplicações da Regressão Logística se deram em estudos biomédicos, mas, nos últimos 20 anos, também têm sido muito utilizada em ciências sociais e *marketing*, Agresti (2002). Recentemente, a Regressão Logística se tornou uma ferramenta popular na área financeira.

Uma grande variedade de técnicas estatísticas vem sendo desenvolvida com o objetivo de reduzir as perdas provenientes de concessões equivocadas de crédito ou de previsão de insolvência. Os primeiros estudos sobre previsão de insolvência foram elaborados por volta da década de 30. Contudo, esse assunto somente ganhou impulso ao longo da década de 60, quando foram realizados vários estudos no intuito de prever a falência de empresas trabalhando com a Análise Discriminante.

A utilização da Regressão Logística para predição da inadimplência ou insolvência iniciou-se a partir de Ohlson, em 1980, que desenvolveu o primeiro modelo de regressão logística para prever insolvência de empresas. A partir deste momento, a Regressão Logística passou a ser o método preferido pelos pesquisadores por ter pressupostos mais simples. Após o trabalho de Ohlson, muitos outros trabalhos passaram a utilizar modelos de regressão logística para prever insolvência.

2. Objetivo

Este trabalho propõe a construção de um modelo, utilizando a regressão logística, para distinguir bons e maus pagadores, clientes de uma instituição financeira, através da predição de sua inadimplência.

O objetivo da regressão logística é modelar uma variável resposta dicotômica como função de uma ou mais variáveis preditoras que influenciam sua ocorrência. Assim, a regressão logística irá prever a probabilidade de um evento ocorrer.

3. Organização do Trabalho

Este trabalho foi organizado em 9 partes. As seções 1, 2, 3 contemplam a introdução, objetivo e organização do trabalho, respectivamente. A seção 4 descreve o modelo de regressão logística, escolhida para a estimação da probabilidade de inadimplência, com detalhamento da regressão logística simples e múltipla. A seção 5 apresenta a descrição do método da máxima verossimilhança, usado para estimar os coeficientes do modelo logístico, para, em seguida, explicar o teste da significância destes coeficientes e a interpretação dos mesmos. As seções 6 e 7 descrevem as etapas do processo de escolha de variáveis e a verificação da qualidade de ajuste do modelo escolhido. Após o detalhamento de todo o processo para a elaboração do modelo de regressão logística, na seção 8 é apresentado um estudo de caso para predição da inadimplência. Ao final, são apresentadas as considerações finais do trabalho, na seção 9, seguida dos Anexos I e II contendo algumas deduções de fórmulas e das referências bibliográficas.

4. Modelo de Regressão Logística

A regressão logística busca, a partir de um modelo matemático, explicar a relação entre uma variável resposta Y_i, categórica, nominal ou ordinal, e uma ou mais variáveis preditoras. No caso de apenas uma variável preditora x_i, tem-se a Regressão Logística Simples. O modelo na sua forma usual é definido por:

$$E(Y_i/x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+ \exp(\beta_0 + \beta_1 x_i)} \quad (4.1)$$

Onde:

$E(Y_i/x_i)$ é o valor esperado de Y dado o valor de x;

Y_i é uma variável dicotômica, assumindo os valores (0 ou 1);

β_0 e β_1 são os coeficientes de regressão a serem estimados pelo método da máxima verossimilhança;

x_i é o valor observado da variável x;

$i = 1, 2, ..., n$;

n é o tamanho da amostra.

Para simplificar a notação, usaremos $p(x_i) = E(Y_i/x_i)$ para representar o valor esperado de Y, dado o valor de x.

A função $p(x_i)$ pode ser transformada, usando a transformação logística na forma linear temos:

$$g(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x \qquad (4.2)$$

O gráfico da função logística tem formato curvilíneo, formando um "S" nas suas extremidades, como pode ser visto na Figura 1:

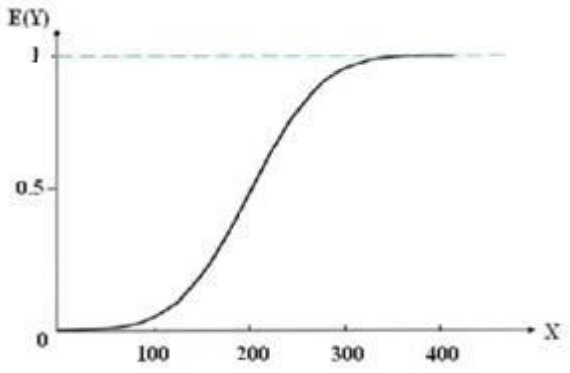


Figura 1- Exemplo do modelo logístico para uma variável preditora

No caso da Regressão Logística Múltipla, o modelo é composto por duas ou mais variáveis preditoras ($x_1, x_2, \ldots, x_p$), onde p é o número de variáveis, e por seus respectivos coeficientes de regressão β_0,..., β_p. Os valores dos parâmetros β_0, β_1,..., β_p são estimados a partir do método da máxima verossimilhança a ser descrito na seção 5. Portanto, tem-se:

$$g(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (4.3)$$

O modelo se estende para o modelo logístico múltiplo, que é determinado por:

$$p(x_i) = \frac{\exp(g(x_i))}{1 + \exp(g(x_i))} \quad (4.4)$$

A variável dependente Y_i pode ser expressa como:

$$Y_i = p(x_i) + \varepsilon i \quad (4.5)$$

Onde o termo εi é o erro aleatório do modelo e representa a diferença entre o valor observado de Y_i e o valor esperado condicionado de Y_i dado x_i. Enquanto nos modelos de Regressão Linear os erros devem seguir uma distribuição normal com média 0 e variância constante, para um dado valor de x, os erros do modelo de Regressão Logística, condicionados a $X = x_i$, seguem uma Distribuição Bernoulli com média 0 e variância $p(x_i)\,(1 - p(x_i))$.

5. Estimação dos Coeficientes - Método da Máxima verossimilhança

Para ajustar o modelo de regressão logística da equação 4.1, no caso de uma única variável preditora, é necessário que os parâmetros desconhecidos β_0 e β_1 sejam estimados.

Na Regressão Linear, o método mais comum usado para estimar os parâmetros desconhecidos é o método dos mínimos quadrados. Neste método, são escolhidos valores de β_0 e β_1 que minimizam a soma dos desvios ao quadrado dos valores observados de Y em relação aos valores preditos no modelo.

Na Regressão Logística, para estimar os parâmetros β_0 e β_1 usamos o método da máxima verossimilhança, *Hosmer-Lemeshow* & *Sturdivant* (2013). Para a aplicação deste método, deve-se construir uma função denominada função de verossimilhança. Os estimadores de máxima verossimilhança dos parâmetros são aqueles que maximizam esta função.

Se a variável resposta Y assume os valores 0 ou 1, então o resultado da expressão 4.4, $p(\mathrm{x_i})$, nos dará a probabilidade de Y ser igual a 1, dado $x, (P(Y = 1 \mid \mathrm{x})$). Da mesma maneira, tem-se que o resultado de $(1 - p(x))$ nos dará a probabilidade de Y ser igual a 0, dado $x, (P(Y = 0 \mid \mathrm{x})$. Assim, para o par (x_i, y_i), onde $Y_i = 1$ o resultado da função verossimilhança é $p(\mathrm{x_i})$ e quando $Y_i = 0$ o resultado da função verossimilhança é $(1 - p(\mathrm{x_i}))$. O resultado da função verossimilhança para o par (x_i, y_i) pode ser expresso como:

$$p(x_i)^{y_i}(1 - p(x_i)^{y_i}) \qquad (5.1)$$

Desde que as observações são preditoras, a função de verossimilhança é obtida pelo produto dos termos da expressão 5.1.

$$l(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{y_i}) \qquad (5.2)$$

O princípio da máxima verossimilhança assegura que se use como estimador de $\boldsymbol{\beta}$ o valor que maximiza a equação 5.2. É matematicamente mais fácil trabalhar com o logaritmo da equação 5.2. Assim, a expressão log verossimilhança é definida por:

$$L(\beta) = \ln[l(\beta)] \; = \; \sum_{i=1}^{n}\{y_i \ln[p(x_i)] + (1 - \; y_i) \ln[1 - p(x_i)] \; \} \qquad (5.3)$$

Para achar o valor de $\boldsymbol{\beta}$ que maximize $L(\boldsymbol{\beta})$ deve-se diferenciar $L(\boldsymbol{\beta})$ em relação à β_o e β_1 (Anexo I) e igualar o resultado das expressões a 0, encontrando as seguintes equações:

$$\sum_{i=1}^{n}[y_i - p(x_i)] \quad (5.4)$$

$$\sum_{i=1}^{n} x_i[y_i - p(x_i)] \quad (5.5)$$

A solução das equações 5.4 e 5.5 não será apresentada neste trabalho, pois para resolvê-las é preciso recorrer a métodos numéricos interativos como, por exemplo, o método de Newton Raphson, já que as equações são não lineares nos parâmetros do modelo.

Como resultado da equação 5.4, tem-se que a soma dos valores observados de *Y* é igual à soma dos valores preditos, *Hosmer-Lemeshow* & *Sturdivant* (2013). Esta propriedade será útil para avaliarmos a predição do modelo.

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{p}(x_i) \quad (5.6)$$

Sendo $p(x_i)$ conforme definido na equação 4.1.

Assim, como no modelo de regressão logística com apenas uma variável, no modelo de regressão logística múltipla deve-se estimar os coeficientes do modelo e avaliar suas significâncias.

O método utilizado para estimar estes coeficientes também será a máxima verossimilhança. A função de verossimilhança é parecida com a da equação 5.1, mas $p(x_i)$ será definido pela equação 4.4. Assim, a função de verossimilhança será:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n}\{y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)] \quad (5.7)$$

Existirão $p + 1$ equações de verossimilhança que serão obtidas diferenciando a função log verossimilhança em relação a cada um dos coeficientes (Anexo I). As equações resultantes serão, *Hosmer-Lemeshow* & *Sturdivant* (2013):

$$\sum_{i=1}^{n}[y_i - p(x_i)] \quad (5.8)$$

$$\sum_{i=1}^{n} x_{ij}[y_i - p(x_i)] \quad (5.9)$$

para j = 1,2,3,…p.

5.1. Testando a significância dos coeficientes

Depois de estimar os coeficientes, o primeiro teste de ajuste do modelo geralmente diz respeito a uma avaliação da importância das variáveis utilizadas no modelo. Isto normalmente envolve testes de hipóteses para determinar se as variáveis são significativamente relacionadas com a variável resposta, que, no caso, é a ocorrência do evento.

5.1.1. Deviance

Para testarmos a significância da variável devemos comparar os valores observados da variável resposta com os valores preditos pelo modelo com ou sem a variável que está sendo avaliada. Na regressão logística, a comparação entre valores observados e preditos é baseada na função log verossimilhança definida na equação 5.3. Para se entender melhor esta comparação é útil conceitualmente se pensarmos em um valor observado da variável resposta como sendo também um valor previsto resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto o tamanho da amostra, segundo *Hosmer-Lemeshow* & *Sturdivant* (2013).

A comparação entre valores observados e preditos usando a função de verossimilhança é baseada na seguinte expressão:

$$D = -2ln\left[\frac{\text{verossimilhança do modelo ajustado}}{\text{verossimilhança do modelo saturado}}\right] \quad (5.10)$$

A parte da expressão acima que está dentro dos colchetes é chamada de razão de verossimilhança. O motivo de se utilizar "- 2 ln" é matemático e é necessário para se obter uma distribuição que é conhecida e, portanto, pode ser utilizada para fins de teste hipótese. Este teste é chamado teste da razão de verossimilhança. Conforme as equações 5.3 e 5.10 temos (Anexo II):

$$D = \sum_{i=1}^{n} \left[yi \, ln\left(\frac{\hat{p}(xi)}{yi}\right) + (1 - yi) \, ln\left(\frac{1-\hat{p}(xi)}{1-yi}\right) \right] \qquad (5.11)$$

Sendo $\hat{p}(x_i)$ a estimativa de máxima verossimilhança de $p(x_i)$.

A estatística D na equação acima é chamada "deviance" e desempenha um papel central em algumas abordagens para avaliação da qualidade de ajuste do modelo (*goodness-of-fit*). A "deviance" na regressão logística desempenha o mesmo papel que a soma dos quadrados dos resíduos na regressão linear.

Como a função de verossimilhança para o modelo saturado é sempre igual a 1, a *deviance*, conforme definido na equação 5.9 será:

$$D = -2ln[verossimilhaça\ do\ modelo\ ajustado]$$

5.1.2. Estatística G

Para avaliar a significância de uma variável preditora, deve-se comparar o valor de D com ou sem a variável. A mudança do valor de D com a inclusão da variável é obtida por:

$$G = D\ (para\ o\ modelo\ sem\ a\ variável) - D\ (para\ o\ modelo\ com\ a\ variável)$$

Como função de verossimilhança do modelo saturado é comum a ambos os valores de D, temos:

$$G = -2ln\left[\frac{\text{verossimilhança sem a variável}}{\text{verossimilhança com a variável}}\right] \qquad (5.12)$$

Para o caso específico de uma única variável preditora é fácil demonstrar que quando a variável não está no modelo, o estimador de máxima verossimilhança de β_0 é:

$\ln\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n}(1-Y_i)}\right)$ e o valor predito é constante e igual a $\sum_{i=1}^{n}\frac{Y_i}{n}$

Neste caso, o valor de G é:

$$G = 2\{\sum_{i=1}^{n}[Y_i \ln(\hat{p}(x_i) + (1-Y_i)\ln(1-\hat{p}(x_i)] - [(n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)]\} \quad (5.13)$$

Sendo $n_{1=} \sum_{i=1}^{n} Y_i$ e $n - n_1 = n_0 = \sum_{i=1}^{n}(1-Y_i)$

Sendo a hipótese nula β_1 é igual a zero e a hipótese alternativa β_1 é diferente de zero, temos, sob a hipótese nula, que a estatística G se aproximará de uma distribuição qui-quadrado, com 1 grau de liberdade.

5.1.3. Teste de *Wald*

Outro teste para avaliar a significância estatística dos coeficientes no modelo é o teste de *Wald*, *Hosmer-Lemeshow* & *Sturdivant* (2013). Este teste é fundamentado na razão do coeficiente pelo seu respectivo desvio padrão (SE), conforme segue:

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \quad (5.14)$$

A distribuição desta razão, sob a hipótese de que $\beta_1 = 0$, se aproximará uma distribuição normal padrão.

O p-valor é definido como $P\,(|Z| > |W|\,)$, sendo que Z denota a variável aleatória da distribuição normal padrão e W o valor observado da estatística de *Wald* (3.13). *Hauck e Donner* (1977) examinaram o desempenho do teste de *Wald* e descobriram que, em determinadas situações, a hipótese nula não é rejeitada apesar de o coeficiente ser significativo. Nestas situações, recomendam o teste da razão de verossimilhança.

5.2. Interpretação dos Coeficientes do Modelo de Regressão Logística

A interpretação dos parâmetros de um modelo de regressão logística é obtida através da função *odds ratio* – OR (razão de chances), comparando-se a probabilidade de sucesso do evento ocorrer com a probabilidade de fracasso, ou seja, o evento não ocorrer.

Para um indivíduo fixo, a chance de sucesso do evento, dado que o valor da variável preditora para esse indivíduo é $X = x_0$ é dada por:

$$\frac{p(x_0)}{1-p(x_0)} = \frac{\frac{e^{\beta_0+\beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}}{1-\frac{e^{\beta_0+\beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}} = \frac{\frac{e^{\beta_0+\beta_1 x_0}}{1+e^{\beta_0+\beta_1 x_0}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_0}}} = e^{\beta_0+\beta_1 x_0} \quad (5.15)$$

Para um outro indivíduo cujo valor de X seja x_1, tem-se:

$$\frac{p(x_1)}{1-p(x_1)} = e^{\beta_0+\beta_1 x_1} \quad (5.16)$$

A comparação destes dois indivíduos, razão de chances, é dada por:

$$OR = \frac{e^{\beta_0+\beta_1 x_1}}{e^{\beta_0+\beta_1 x_0}} = e^{\beta_1(x_1-x_0)} \quad (5.17)$$

Se $x_1 = x_0 + 1$, então

$$OR = e^{\beta_1(x_0+1-x_0)} = e^{\beta_1} \quad (5.18)$$

E o log da razão de chance é:

$$\ln(OR) = \ln(e^{\beta_1}) = \beta_1 \quad (5.19)$$

Uma razão de chances de valor 1 indica que a ocorrência do evento sob estudo é igualmente provável nas duas classes. Uma razão de chances maior do que 1 indica que o evento tem maior probabilidade de ocorrer na primeira classe, e uma razão de chances menor do que 1 indica que a probabilidade é menor na primeira classe do que no segunda.

Consideramos, agora, o caso em que a variável preditora pode assumir $k > 1$ valores distintos. Como temos mais de uma classe, iremos eleger uma classe como sendo a de referência. Assim, a razão de chances indicará a probabilidade de um evento ocorrer numa classe sempre em relação à classe de referência. Então, a exponencial dos coeficientes de cada classe representa a chance do evento ocorrer nesta classe em relação à classe de referência.

Quando se trata de uma variável contínua, a exponencial da multiplicação de seu coeficiente por um valor de "c" unidades nos fornece o aumento ou redução da probabilidade do evento ocorrer em relação à variável sem o acréscimo de "c" unidades. Resumindo, a interpretação da razão de chances de ocorrência do evento em relação a uma variável contínua é semelhante à de uma variável categórica. A principal diferença é que se deve definir um valor de "c" unidades a ser acrescentado na variável.

6. Escolha das variáveis

Quando temos uma lista de variáveis preditoras que podem ser incluídas no modelo para explicar a variável resposta, devemos desenvolver uma estratégia para escolha das variáveis que resultam no melhor modelo.

Quanto maior o número de variáveis do modelo, maior se torna as estimativas de erros padrão e mais complexo fica o modelo. Portanto, devemos seguir alguns passos para a escolha das variáveis a serem incluídas no modelo, tais como:

1) O processo de escolha de variáveis deve-se iniciar com uma análise de cada variável. Para variáveis preditoras categóricas ordinais, nominais, podemos fazer uma

tabela de contingência dos valores da variável resposta (y = 0,1) versus as k classes de cada variável preditora.

Para variáveis preditoras contínuas o melhor é ajustar um modelo de regressão logística univariada para obter o coeficiente estimado, o erro padrão, teste da razão de verossimilhança para significância do coeficiente e teste da estatística Wald.

2) Após a análise das variáveis, iremos selecionar as variáveis para a análise multivariada. Qualquer variável cujo teste univariado tenha tido um p-valor < 0,25, segundo *Hosmer-Lemeshow* & *Sturdivant* (2013), deve ser considerada como candidata para ser incluída no modelo multivariado junto com as variáveis cuja importância é conhecida. Após a identificação das variáveis, começamos o modelo contendo todas as variáveis selecionadas.

O uso do p-valor < 0,25 para o critério de seleção de variáveis candidatas é baseado no trabalho de Bendel e Afifi (1977) em regressão linear e no trabalho de Mickey e Greenland (1989) em regressão logística. Esses autores mostraram que o uso do p-valor mais tradicional de 0,05 geralmente falha em identificar variáveis importantes conhecidas. O uso do p-valor de 0,25 tem a desvantagem de incluir, no estágio de construção do modelo, variáveis cuja importância é questionável. Por esta razão, é importante analisar todas as variáveis incluídas no modelo antes de se chegar ao modelo final.

Um problema da análise univariada é que é desconsiderada a possibilidade de que uma variável fracamente associada à variável resposta pode se tornar importante preditora quando está em conjunto com outras variáveis. Então devemos escolher um nível de significância grande o bastante para permitir que estas variáveis possam ser incluídas como candidatas no modelo multivariado.

Outra maneira de seleção de variáveis é o uso do método *Stepwise* no qual as variáveis são escolhidas pela sua inclusão ou exclusão no modelo em uma sequência baseada em critério estatístico. O método *Stepwise* é um método de seleção de variáveis explicativas que permite selecionar as variáveis a partir de um conjunto inicial de variáveis explicativas, adotando-se, para tanto, uma probabilidade de entrada, pré-

estabelecida, das variáveis no modelo e uma probabilidade de saída, também previamente estabelecida, a cada passo do método de seleção das variáveis em estudo.

A utilização deste método possibilita uma forma rápida e efetiva para examinar um grande número de variáveis e, simultaneamente, examinar diversas equações de regressão logística possíveis a partir dessas variáveis.

Existem três métodos possíveis para seleção de variáveis:

a) o método "*backward"*, parte de um modelo inicial contendo todas as possíveis variáveis, que vão sendo eliminadas a cada passo, até que se consiga atingir o melhor modelo final;

b) o método "*forward"* se inicia com um modelo sem nenhuma variável explicativa e, a cada passo, são incluídas as variáveis relevantes até que se obtenha o melhor modelo possível;

c) o método "*stepwise*" é uma combinação dos métodos *backward* e *forward.* Inicia-se o modelo sem nenhuma variável e, após cada etapa de inclusão de uma variável, tem-se uma etapa para tentar excluir outra variável.

3) Seguindo o ajuste do modelo de regressão logística, a importância de cada variável incluída no modelo deve ser analisada:

a) uma análise do teste de *Wald* para cada variável;

b) uma comparação de cada coeficiente estimado com o coeficiente do modelo univariado contendo apenas esta variável.

Variáveis que não contribuem para o modelo, conforme estes critérios, devem ser retiradas do modelo e um novo modelo deve ser ajustado. O modelo novo deve ser comparado ao modelo anterior pelo teste da razão de verossimilhança. Também os coeficientes das variáveis remanescentes devem ser comparados com aqueles do modelo completo. Devemos ficar atentos com as variáveis cujos coeficientes alteraram muito. Isto pode indicar que uma ou mais variáveis excluídas eram importantes para prover um ajuste necessário nas variáveis remanescentes do modelo. Este processo de excluir,

reajustar e verificação continua até que todas as variáveis importantes estejam presentes no modelo e que as variáveis excluídas não são realmente necessárias.

4) Uma vez tendo obtido o modelo que contém todas as variáveis importantes, devemos considerar a possibilidade da necessidade de inclusão de termos de interação entre as variáveis. Se no modelo existe alguma interação entre duas variáveis, isso implica que o efeito de uma das variáveis não é constante sobre os níveis da outra variável. A necessidade de incluir um termo de interação no modelo se dá primeiro pela criação do produto das variáveis em questão e depois usar o teste da razão de verossimilhança e verificar sua significância.

7. Ajuste do Modelo Logístico

Após a escolha das variáveis para obter o modelo de regressão logística, deve-se verificar a qualidade de ajuste do modelo, ou seja, o quanto o modelo é efetivo para predizer a variável resposta.

Suponha que os valores observados da variável resposta sejam representados pelo vetor $y' = (y_1, y_2, y_3, \dots, y_n)$. Os valores preditos pelo modelo serão representados pelo vetor $\hat{y}' = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_n)$. Concluiremos que o modelo está ajustado se a distância entre y e $\hat{y}$ for pequena e a contribuição de cada par $(y_i, \hat{y}_i)$, $i = 1, 2, 3, \dots n$ para a medida de qualidade do ajuste é aleatória e pequena em relação à estrutura de erros do modelo. Então, uma completa análise do ajuste do modelo irá envolver a distância entre y e $\hat{y}$ e uma avaliação de seus componentes individuais.

O desenvolvimento de métodos para avaliação da qualidade de ajuste do modelo irá seguir os passos lógicos para o desenvolvimento do modelo. Os componentes da abordagem proposta são o cálculo e avaliação de todas as medidas de ajuste e a análise individual dos componentes, geralmente através de gráficos e análise de outras medidas da diferença ou distância entre y e $\hat{y}$.

Usaremos o termo "número de configurações" para descrever um conjunto de valores de covariáveis do modelo. Assumimos que os valores das variáveis preditoras são tais que cada um é único na configuração das covariáveis. Então, o número de configurações será igual a n no modelo saturado.

Suponha que o modelo ajustado contenha k variáveis preditoras, $x' = (x_1, x_{2,\ldots,} x_k)$, e seja J o número de valores distintos observados de x. Se alguns conjuntos possuem o mesmo valor de x, então $J < n$. O número de observações contendo $x = x_j$ será denominado $m_j, j = 1,2, \ldots, J$. Então, $\sum m_j = n$. Seja W_j o número de respostas positivas $y = 1$ dentro do conjunto m_j. A estatística de distribuição do ajuste do modelo é obtida quando n é grande. Se o número de configurações também aumenta com n, então cada valor de m_j tende a ser pequeno. O resultado obtido sob a condição de apenas n ser grande é chamado n-assintótico. Se foi fixado que J é limitado e n é grande, então cada valor $m_j$tende a se tornar grande. O resultado obtido sob a condição de $m_j$ser grande é chamado m-assintótico.

7.1. Diagnóstico do Ajuste do Modelo Logístico

Para decidir qual modelo regressão logística deverá ser utilizado, aplica-se alguns testes de validação do modelo para verificar se há pontos influentes (*outliers*), se a função resposta é monotônica e em forma de S (sigmoidal) e se modelo logístico ajustado é adequado.

Na regressão logística existem muitas medidas possíveis para determinar a diferença entre os valores observados e os estimados. Uma destas maneiras é o teste de *Hosmer-Lemeshow*, *Hosmer-Lemeshow* & *Sturdivant* (2013), descrito a seguir.

7.1.1. Teste de Hosmer-Lemeshow

Este teste avalia o modelo ajustado comparando as frequências estimadas sob o modelo logístico ajustado com as observadas na amostra. O teste associa os dados às suas probabilidades estimadas, da mais baixa à mais alta e realiza um teste qui-quadrado para determinar se as frequências observadas estão próximas das frequências esperadas. Um modelo com ajuste não adequado apresentará valores elevados de estatística de teste e valores pequenos da probabilidade de significância do teste.

Hipótese nula: o modelo está bem ajustado.

Hipótese alternativa: o modelo não está bem ajustado.

A proposta é a criação de grupos baseados nas probabilidades estimadas. O agrupamento pode ocorrer de 2 maneiras: (i) agrupamento conforme percentis das probabilidades estimadas (ii) agrupamento baseado em valores fixos de probabilidades estimadas.

Com o primeiro método, usando um número de 10 grupos, teremos o primeiro grupo com as menores probabilidades estimadas e o último grupo contendo as maiores probabilidades estimadas.

Com o segundo método, usando um número de 10 grupos resulta em definição de pontos de corte sendo que os grupos conterão todos os indivíduos com as probabilidades estimadas entre os pontos de corte adjacentes.

O método de agrupamento baseado em percentis das probabilidades estimadas é melhor do que o método baseado em pontos de corte fixos, especialmente quando muitas das probabilidades estimadas são pequenas, como menor que 0,2, por exemplo, *Hosmer-Lemeshow* & *Sturdivant* (2013).

Então, sugere-se o uso do método de agrupamento baseado em percentis e, normalmente, o número de grupos utilizado é de 10 grupos. Esses grupos são chamados de perfis de risco.

Para qualquer estratégia de grupo escolhida, o teste de ajuste de *Hosmer-Lemeshow*, $\hat{C}$, é obtido calculando-se o teste qui-quadrado Pearson de uma tabela 2 x g, ond g é a quantidade de grupos, de frequências estimadas e observadas, conforme a fórmula:

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - e_{1k})2}{e_{1k}} + \frac{(o_{0k} - e_{0k})2}{e_{0k}} \right] \qquad (7.1)$$

Onde:

$$o_{1k} = \sum_{j=1}^{c_k} Wj$$

$$o_{0k} = \sum_{j=1}^{c_k} (mj - Wj)$$

$$e_{1k} = \sum_{j=1}^{c_k} m_j \, \pi_j$$

$$e_{0k} = \sum_{j=1}^{c_k} m_j \, (1 - \pi_j)$$

e C_k = número de covariáveis padrão no k-ésimo grupo

Segundo *Hosmer-Lemeshow* & *Sturdivant* (2013), quando $J = n$ e o modelo está bem ajustado, a distribuição da estatístia $\hat{C}$ se aproxima de uma distribuição qui-quadrado com $g - 2$ graus de liberdade $X^2 \, (g - 2)$. Isto também ocorre quando $J \approx n$.

7.1.2. Matriz de classificação

Uma maneira intuitiva para resumir os resultados de um modelo de regressão logística é através de uma matriz de classificação. Esta matriz é o resultado de classificação cruzada da variável resposta, y, com uma variável dicotômica cujos valores são derivados a partir das probabilidades estimadas. Os coeficientes produzidos pelo modelo são utilizados para prever o resultado (de uma forma binária). Para obter a variável dicotômica é necessária a definição de um ponto de corte, c, e comparar cada probabilidade estimada com o ponto de corte. Se a probabilidade estimada exceder c, então assume-se que o resultado predito para a variável resposta deve ser igual a 1; caso contrário, deve ser igual a 0. O valor do ponto de corte mais comumente utilizado para c é de 0,5.

7.1.3. Área sob a curva ROC

Outra maneira de se avaliar a predição do modelo é através da curva ROC - *Receiver Operating Characteristic, Hosmer-Lemeshow* & *Sturdivant* (2013). Espera-se, em um modelo bem ajustado, que a taxa de acerto de indivíduos com resultado positivo (Y=1) seja alta, aliada a uma taxa baixa de falsos positivos. A sensibilidade é calculada como a probabilidade predita pelo modelo de um indivíduo apresentar resultado positivo, dado que ele realmente o é. Já a especificidade é, ao contrário, a probabilidade de um indivíduo ser classificado como resultado negativo (Y = 0), dado que ele realmente o é.

A sensibilidade e especificidade, assim como outras medidas de desempenho da *performance* do modelo, utiliza uma tabela de 2 × 2 e dependem de um único ponto de corte utilizado para classificar o resultado como positivo. Outra maneira de se avaliar sensibilidade e especificidade, com a utilização de vários pontos de corte, é a utilização da área sob a curva - *Receiver Operating Characteristic* (ROC). Esta curva traça a

probabilidade de detectar resultado positivo (sensibilidade) e sinal negativo (1 - especificidade) para toda uma gama de possíveis pontos de corte.

A área sob a curva ROC, que varia de 0,5 a 1,0, proporciona uma medida da capacidade do modelo para discriminar entre aqueles indivíduos que possuem valor positivo para a variável resposta (y = 1) versus aqueles que não possuem (y = 0). Existem alguns benefícios associados à utilização de 0,5 como ponto de corte, mas poderíamos considerar o que acontece quando nós utilizamos outros pontos de corte.

Se o objetivo for escolher um ponto de corte ideal para efeitos de classificação, pode-se selecionar um ponto de corte que maximiza a sensibilidade e especificidade. Esta escolha é facilitada por meio de um gráfico, tal como o mostrado na Figura 2, que traça sensibilidade e especificidade comparativamente a cada ponto de corte possível. Conforme o exemplo desta figura, pode-se observar que uma escolha "ótima" para um ponto de corte pode ser 0,24, que é aproximadamente onde as curvas de sensibilidade e especificidade se cruzam.

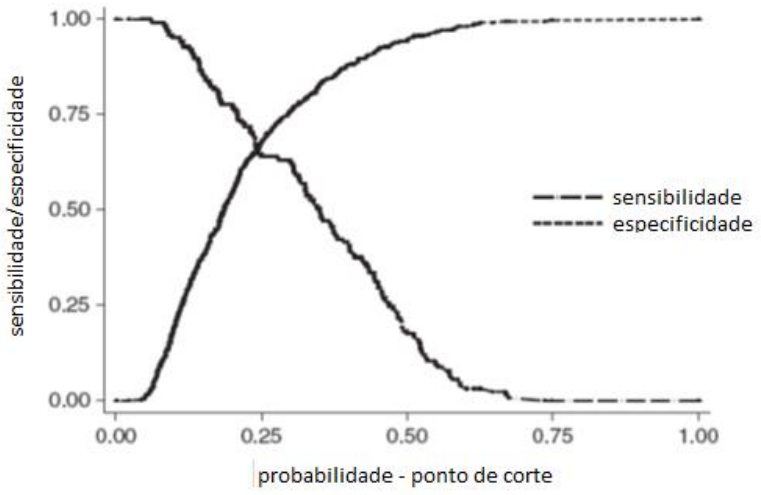


Figura 2 – Exemplo de gráfico de sensibilidade e especificidade para vários pontos de corte

Além de avaliar a qualidade de ajuste do modelo, deve-se avaliar o poder de discriminação do modelo. Podemos ter modelos bem ajustados que discriminam mal, assim como podemos ter modelos com ajuste pobre que discriminam bem.

Um gráfico traçando a sensibilidade versus (1 – especificidade) sobre todos os pontos de corte possíveis (ou seja, usando para cada indivíduo probabilidade estimada, em vez de dados agrupados) é mostrado na Figura 3. A curva gerada por estes pontos é chamada de curva ROC e a área sob a curva fornece uma medida, cujo cálculo é descrito a seguir.

Se a distribuição do modelo de probabilidades estimada for a mesma nos dois grupos, então a curva ROC seria idêntica à linha reta mostrada na Figura 3. À medida que as distribuições de probabilidades estimadas pelo modelo torna-se mais discriminador, o traçado da curva ROC aumenta mais rapidamente e a área sob ela aumenta de 0,5 para seu valor teórico máximo de 1,0.

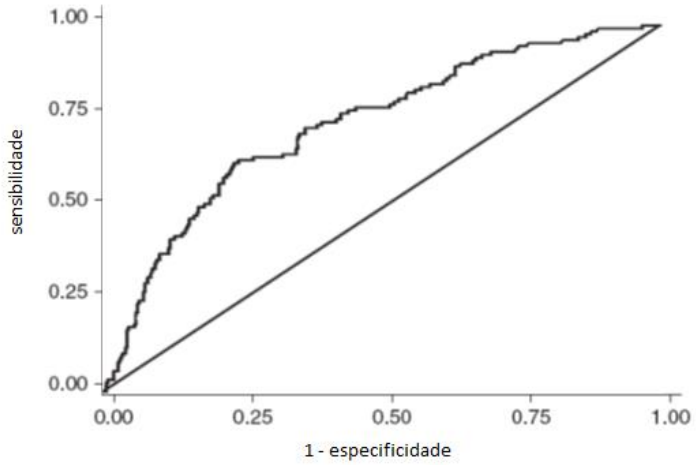


Figura 3 – Exemplo de gráfico de sensibilidade versus (1 – especificidade) para vários pontos de corte

Um valor de referência da área sob a curva ROC que indica uma boa discriminação não existe. Segundo *Hosmer-Lemeshow* & *Sturdivant* (2013), a seguinte classificação pode ser usada:

ROC = 0,5	não possui poder de discriminação
0,5 < ROC < 0,7	baixo poder de discriminação
0,7 <= ROC < 0,8	aceitável poder de discriminação
0,8 <= ROC < 0,9	excelente poder de discriminação
ROC >= 0,9	poder de discriminação acima do normal

8. Estudo de caso

8.1. Considerações Iniciais

O estudo de caso apresentado neste trabalho tem como objetivo a elaboração de um modelo de regressão logística para predição da inadimplência em uma instituição financeira. Os clientes serão classificados em função de uma série de informações reunidas pela instituição financeira, a fim de prever, da forma mais acurada possível, a probabilidade de inadimplência. O objetivo é, portanto, traçar, a partir de um conjunto de atributos observáveis de sua base de clientes, o perfil daqueles que, ao longo do tempo, poderiam ser considerados bons ou maus clientes.

Portanto, como estamos interessados em prever a inadimplência para decisão de concessão ou não do crédito a determinado cliente, o modelo apresentado irá estimar a probabilidade da ocorrência de atrasos superiores a 90 dias.

A base de dados utilizada para modelagem considerou informações de 6.292 clientes, micro e pequenas empresas, cujo financiamento foi obtido do ano de 2009 até o primeiro semestre de 2015. Sobre esta base de clientes, 914 clientes apresentaram atraso superior a 90 dias, sendo considerados maus pagadores. Por outro lado, temos que 5.378 clientes foram considerados bons pagadores, pois o atraso máximo apresentado foi de 14 dias. Os clientes que apresentaram atraso entre 15 e 89 não foram incluídos na base de dados, pois existem incertezas quanto ao retorno do cliente à situação de normalidade, porém não configurada, ainda, a situação de inadimplência, no período em que esse estudo foi elaborado.

Tabela 1 - Distribuição da frequência de clientes em função da inadimplência

Inadimplência 90 dias	base de dados	
	qtde clientes	Frequência (%)
Sim	914	14,5
Não	5378	85,5
Total	**6292**	**100**

8.2. Análise das Variáveis

A primeira etapa do processo de desenvolvimento do modelo de regressão logística consistiu na análise univariada, ou seja, a análise da associação entre cada uma das variáveis candidatas a fazer parte do modelo e a inadimplência (evento).

Foram analisadas variáveis contínuas e categóricas, mas as variáveis selecionadas no modelo foram todas categóricas nominais. Como não foi possível obter autorização para divulgar essas variáveis, as mesmas serão denominadas por siglas, acompanhadas de suas categorias. Os resultados das variáveis selecionadas no modelo são apresentados nas tabelas a seguir.

A Tabela 2 detalha a distribuição de clientes em função de sua inadimplência, de acordo com a primeira variável denominada de "ec". Esta variável apresenta 3 categorias. Conforme a Tabela 2, é possível notar que 10,6% dos clientes apresentaram inadimplência de 90 dias na categoria 1. Este percentual aumenta na categoria 2 (16,9%) e aumenta mais ainda na categoria 3 (24,1%), ou seja, a proporção de clientes inadimplentes na categoria 3 é a maior que todas as outras categorias.

Tabela 2 - Distribuição de clientes em função da inadimplência conforme a variável "ec"

			Inadimplência 90 dias		
			Não	Sim	Total
ec	ec_categoria1	Quantidade	3743	446	4189
		% na categoria ec_categoria1	89,4	10,6	100,0
		% no perfil	69,6	48,8	66,6
	ec_categoria2	Quantidade	443	90	533
		% na categoria ec_categoria2	83,1	16,9	100,0%
		% no perfil	8,2	9,8	8,5
	ec_categoria3	Quantidade	1192	378	1570
		% na categoria ec_categoria3	75,9	24,1	100,0
		% no perfil	22,2	41,4	25,0
Total		Quantidade	5378	914	6292
		% na variável ec	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

A segunda variável foi denominada de “hb”. Esta variável apresenta 2 categorias. Conforme a Tabela 3, é possível notar que 18,8% dos clientes apresentaram inadimplência de 90 dias na categoria 1, contra apenas 3,6% na categoria 2. Além disso, a categoria 1 representa 93% do total de clientes que apresentaram inadimplência superior a 90 dias.

Tabela 3 - Distribuição de clientes em função da inadimplência conforme a variável “hb”

			Inadimplência 90 dias		
			Não	Sim	Total
hb	hb_categoria1	Quantidade	3673	850	4523
		% na categoria hb_categoria1	81,2	18,8	100,0
		% no perfil	68,3	93,0	71,9
	hb_categoria2	Quantidade	1705	64	1769
		% na categoria hb_categoria2	96,4	3,6	100,0
		% no perfil	31,7	7,0	28,1
Total		Quantidade	5378	914	6292
		% na variável hb	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

A variável “hemp” apresenta 8 categorias. As informações da Tabela 4 mostram que há uma gradação na proporção clientes que apresentaram inadimplência dentre as categorias, de tal forma que clientes da categoria 1 apresentam uma proporção maior de inadimplentes (50,4%), enquanto que a categoria 8 apresenta apenas 1,9% de inadimplentes.

Tabela 4 - Distribuição de clientes em função da inadimplência conforme a variável hemp

			Inadimplência 90 dias		
			Não	Sim	Total
hemp	hemp_categoria1	Quantidade	57	58	115
		% na categoria hemp_categoria1	49,6	50,4	100,0
		% no perfil	1,1	6,3	1,8
	hemp_categoria2	Quantidade	207	94	301
		% na categoria hemp_categoria2	68,8	31,2	100,0
		% no perfil	3,8	10,3	4,8
	hemp_categoria3	Quantidade	219	75	294
		% na categoria hemp_categoria3	74,5	25,5	100,0
		% no perfil	4,1	8,2	4,7
	hemp_categoria4	Quantidade	855	196	1051
		% na categoria hemp_categoria4	81,4	18,6	100,0
		% no perfil	15,9	21,4	16,7
	hemp_categoria5	Quantidade	736	133	869
		% na categoria hemp_categoria5	84,7	15,3	100,0
		% no perfil	13,7	14,6	13,8
	hemp_categoria6	Quantidade	2440	330	2770
		% na categoria hemp_categoria6	88,1	11,9	100,0
		% no perfil	45,4	36,1	44,0
	hemp_categoria7	Quantidade	349	18	367
		% within hemp_categoria7	95,1	4,9	100,0
		% no perfil	6,5	2,0	5,8
	hemp_categoria8	Quantidade	515	10	525
		% na categoria hemp_categoria8	98,1	1,9	100,0
		% no perfil	9,6	1,1%	8,3%
Total		Quantidade	5378	914	6292
		% na variável hemp	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

A variável “hs” apresenta 5 categorias. Assim como na variável anterior, a Tabela 5 mostra que também há uma gradação na proporção de clientes que apresentaram inadimplência dentre as categorias, de tal forma que clientes da categoria 1 apresentam uma proporção maior de inadimplentes (54%), enquanto que a categoria 6 apresenta 9,8% de clientes inadimplentes.

Tabela 5: Distribuição de clientes em função da inadimplência conforme a variável “hs”

			Inadimplência 90 dias		
			Não	Sim	Total
hs	hs_categoria1	Quantidade	120	141	261
		% na categoria hs_categoria1	46,0	54,0	100,0
		% no perfil	2,2	15,4	4,1
	hs_categoria2	Quantidade	113	54	167
		% na categoria hs_categoria2	67,7	32,3	100,0
		% no perfil	2,1	5,9	2,7
	hs_categoria4	Quantidade	243	58	301
		% na categoria hs_categoria4	80,7	19,3	100,0
		% no perfil	4,5	6,3	4,8
	hs_categoria3	Quantidade	950	203	1153
		% within hs_categoria3	82,4	17,6	100,0
		% no perfil	17,7	22,2	18,3
	hs_categoria5	Quantidade	858	120	978
		% na categoria hs_categoria5	87,7	12,3	100,0
		% no perfil	16,0	13,1	15,5
	hs_categoria6	Quantidade	3094	338	3432
		% na categoria hs_categoria6	90,2	9,8	100,0
		% no perfil	57,5	37,0	54,5
Total		Quantidade	5378	914	6292
		% na variável hs	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

A variável “ts” apresenta 4 categorias, com a proporção de clientes inadimplentes também diminuindo em cada categoria, conforme demonstrado na Tabela 6.

Tabela 6: Distribuição de clientes em função da inadimplência conforme a variável “ts”

			Inadimplência 90 dias		
			Não	Sim	Total
ts	ts_categoria1	Quantidade	274	123	397
		% na categoria ts_categoria1	69,0	31,0	100,0
	ts_categoria2	Quantidade	844	307	1151
		% na categoria ts_categoria2	73,3	26,7	100,0
	ts_categoria3	Quantidade	2377	366	2743
		% na categoria ts_categoria3	86,7	13,3	100,0
	ts_categoria4	Quantidade	1883	118	2001
		% na categoria ts_categoria4	94,1	5,9	100,0
Total		Quantidade	5378	914	6292
		% na categoria ts	85,5	14,5	100,0

A variável “fi” apresenta apenas 2 categorias, sendo que a segunda categoria apresenta proporção maior de clientes inadimplentes, conforme demonstrado na Tabela 7.

Tabela 7: Distribuição de clientes em função da inadimplência conforme a variável “fi”

			Inadimplência 90 dias		
			Não	Sim	Total
fi	fi_categoria1	Quantidade	3735	499	4234
		% na categoria fi_categoria1	88,2	11,8	100,0
		% no perfil	69,4	54,6	67,3
	fi_categoria2	Quantidade	1643	415	2058
		% na categoria fi_categoria2	79,8	20,2	100,0
		% no perfil	30,6	45,4	32,7
Total		Quantidade	5378	914	6292
		% na variável fi	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

A última variável do modelo é a variável "qtds" cuja categoria 1 apresenta 43,8% do total de clientes inadimplentes. Dos clientes desta categoria, 23,8% são inadimplentes.

Tabela 8: Distribuição de clientes em função da inadimplência conforme a variável "qtds"

			Inadimplência 90 dias		
			Não	Sim	Total
qtds	qtds_categoria1	Quantidade	1279	400	1679
		% na categoria qtds_categoria1	76,2	23,8	100,0
		% no perfil	23,8	43,8	26,7
	qtds_categoria2	Quantidade	688	179	867
		% na categoria qtds_categoria2	79,4	20,6	100,0
		% no perfil	12,8	19,6	13,8
	qtds_categoria3	Quantidade	1619	235	1854
		% na categoria qtds_categoria3	87,3	12,7	100,0
		% no perfil	30,1	25,7	29,5
	qtds_categoria4	Quantidade	1792	100	1892
		% na categoria qtds_categoria4	94,7	5,3	100,0
		% no perfil	33,3	10,9	30,1
Total		Quantidade	5378	914	6292
		% na variável qtds	85,5	14,5	100,0
		% no perfil	100,0	100,0	100,0

8.3.Ajuste do Modelo

Segundo Sicsú, (2010), reserva-se entre 50% e 70% da amostra original como amostra de desenvolvimento e o restante para amostra de validação. Assim, para a modelagem, utilizou-se uma amostra de 4.413 clientes, 70% da amostra original, amostra selecionada através do software SPSS (*Statistical Package for the Social*

Sciences), versão 13.0. Os outros 30% da amostra serão utilizados para se avaliar o resultado do modelo ajustado.

O resultado do ajuste do modelo de regressão logística, executado através do programa SPSS, versão 13.0, está apresentado na Tabela 9, indicando, em particular, as variáveis selecionadas e seus pesos (1ª e 2ª colunas, respectivamente).

A terceira coluna se refere ao desvio padrão (S.E.), enquanto que a quarta coluna é o teste de Wald, detalhado no item 5.1.3. O programa SPSS faz o cálculo da estatística de Wald e usa o teste qui-quadrado. Sendo assim, o teste de Wald será calculado como $\widehat{\beta}$:

$$W = \frac{\hat{\beta}^2}{SE^2} \qquad (8.1)$$

A quinta coluna se refere aos graus de liberdade menos 1. Para o nível de significância de 5%, os coeficientes de todas as categorias de todas as variáveis foram considerados estatisticamente significativos, ou seja, a hipótese nula de que o coeficiente é igual a zero (exemplo: $H_0: \beta_1 = 0$) é rejeitada.

A sexta coluna se refere à razão de chance ($\exp(\beta)$), detalhada no item 5.2. Como exemplo, para a primeira variável "ec": (i) a probabilidade dos clientes da categoria 2 ficarem inadimplentes é 1,7 vezes maior que a dos clientes da categoria 1 e (ii) a probabilidade dos clientes da categoria 3 ficarem inadimplentes é 1,8 vezes maior que a dos clientes da categoria 1.

Ressalta-se que a seleção das variáveis que compõem o modelo baseou-se em critério estatístico, por meio do método "*Stepwise "*através do software SPSS, baseado no teste da razão de verossimilhança, sendo considerados os níveis de significância iguais a 0,05 e 0,10 para entrada e saída do modelo, respectivamente.

Tabela 9- Modelo Logístico – Variáveis selecionadas

	B	S.E.	Wald	df	p-value	Exp(B)
ec_categoria1 (referência)			36,424	2	,000	
ec_categoria2	,536	,165	10,509	1	,001	1,708
ec_categoria3	,599	,104	32,970	1	,000	1,821
hb_categoria2	-,971	,180	29,099	1	,000	,379
hemp_categoria1 (referência)			76,005	7	,000	
hemp_categoria2	-,848	,318	7,116	1	,008	,428
hemp_categoria3	-1,109	,318	12,195	1	,000	,330
hemp_categoria4	-1,566	,288	29,642	1	,000	,209
hemp_categoria5	-1,346	,293	21,053	1	,000	,260
hemp_categoria6	-1,027	,277	13,706	1	,000	,358
hemp_categoria7	-2,164	,426	25,761	1	,000	,115
hemp_categoria8	-3,481	,527	43,677	1	,000	,031
hs_categoria1 (referência)			92,263	5	,000	
hs_categoria2	-,555	,274	4,089	1	,043	,574
hs_categoria3	-,887	,243	13,334	1	,000	,412
hs_categoria4	-1,142	,196	33,853	1	,000	,319
hs_categoria5	-1,452	,206	49,805	1	,000	,234
hs_categoria6	-1,572	,183	73,867	1	,000	,208
ts_categoria1 (referência)			56,612	3	,000	
ts_categoria2	-,348	,140	6,199	1	,013	,706
ts_categoria3	-,639	,128	24,726	1	,000	,528
ts_categoria4	-1,233	,170	52,480	1	,000	,291
fi_categoria2	,374	,099	14,315	1	,000	1,453
qtds_categoria1 (referência)			64,451	3	,000	
qtds_categoria2	-,341	,123	7,660	1	,006	,711
qtds_categoria3	-,447	,172	6,756	1	,009	,640
qtds_categoria4	-1,196	,151	63,034	1	,000	,303
Constante	1,447	,323	20,048	1	,000	4,249

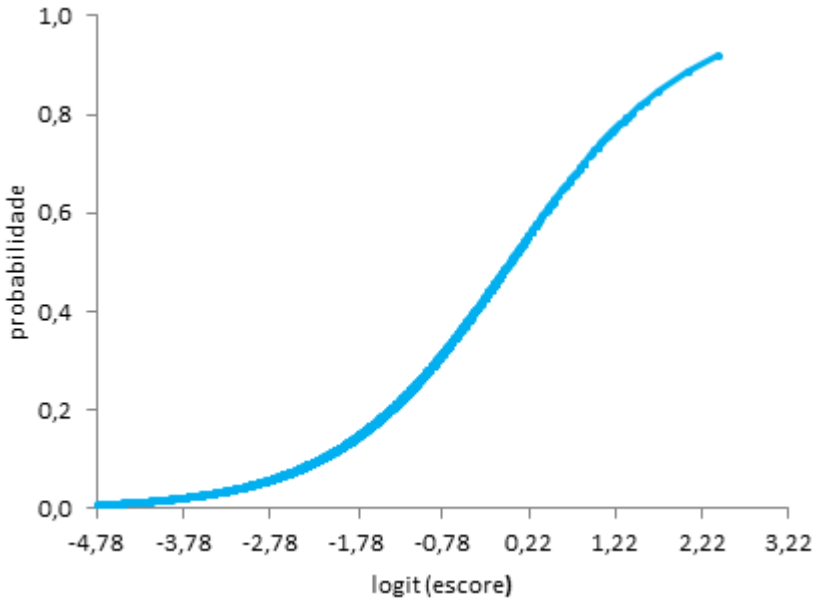
8.4.Diagnóstico do Ajuste do Modelo Logístico

8.4.1. Função resposta estimada

A primeira avaliação a ser realizada é a respeito da função resposta estimada. A relação entre probabilidades e escores compõe a curva logística estimada, apresentada no Gráfico 1, mostrando que a função resposta é monotônica e em forma de S (sigmoidal). O escore é definido como a transformação logística da função $p(x_i)$:

$$g(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1\, x_{i1} + \cdots + \beta_p\, x_{ip}$$

Gráfico 1- Curva logística ajustada ao modelo



8.4.2. Teste de Hosmer – Lemeshow

O Teste de *Hosmer – Lemeshow* já foi detalhado no item 7.1.1. Como demonstrado na Tabela 11, os 4.413 clientes foram divididos em 10 grupos e a estatística de teste obtida foi de 14,49. Como, o p-valor foi de 0,07, não se pode rejeitar a hipótese nula de que o modelo se ajusta bem aos dados, ao nível de significância de 5%. A Tabela 12 contém o cálculo detalhado do teste *Hosmer – Lemeshow*.

Tabela 10 - Teste de Hosmer - Lemeshow

Passo	Chi-quadrado	Graus de liberdade	p-valor
8	14,492	8	,070

Tabela 11 - Tabela de Contingência para Estatística do Teste de Hosmer - Lemeshow

		Inadimplência 90: não		Inadimplência 90: sim		
		Observado	Predito	Observado	Predito	Total
Passo 8	1	438	440,695	6	3,305	444
	2	417	423,373	15	8,627	432
	3	422	424,170	17	14,830	439
	4	416	418,994	26	23,006	442
	5	423	414,777	26	34,223	449
	6	404	393,115	36	46,885	440
	7	380	376,634	61	64,366	441
	8	362	357,706	85	89,294	447
	9	313	313,615	128	127,385	441
	10	209	220,922	229	217,078	438

Tabela 12 - Cálculo da Estatística do Teste de Hosmer - Lemeshow

Grupo	Inadimplência 90: não	Inadimplência 90: sim	Total	E- Inadimplência 90: não	E- Inadimplência 90: sim	Total
1	438	6	444	440,695	3,305	2,214
2	417	15	432	423,373	8,627	4,804
3	422	17	439	424,170	14,830	0,329
4	416	26	442	418,994	23,006	0,411
5	423	26	449	414,777	34,223	2,139
6	404	36	440	393,115	46,885	2,829
7	380	61	441	376,634	64,366	0,206
8	362	85	447	357,706	89,294	0,258
9	313	128	441	313,615	127,385	0,004
10	209	229	438	220,922	217,078	1,298
Total Geral	3784	629	4413	Teste Hosmer - Lemeshow		14,492
				Graus de liberdade		8
				p-valor		0,0698

8.4.3. Matriz de Classificação

Outra informação utilizada para avaliar a qualidade do modelo é a matriz de classificação.

As Tabelas de 13 a 19 mostram as taxas de acerto comparando-se resultados previstos pelo modelo e observados na base de dados com a amostra de 70%, utilizada na modelagem. Foram elaboradas várias tabelas com pontos de corte distintos. Dessa maneira, indivíduos com probabilidades maiores ou iguais ao ponto de corte (ou seja, cuja chance de inadimplência era maior ou igual à probabilidade de inadimplemento) foram classificados como inadimplentes, enquanto os demais foram tidos como adimplentes. Como o desejado por uma instituição financeira é uma inadimplência baixa, o primeiro ponto de corte a ser analisado foi o ponto de corte de 6%, ou seja, indivíduos com probabilidade estimada inferior a 6% serão classificados como adimplentes. Pelo mesmo motivo, não foram analisados pontos de corte superiores 50%, pois inadimplência acima deste percentual, ou até mesmo próxima a este percentual, provavelmente não será rentável.

Tabela 13: Matriz de Classificação – ponto de corte de 6%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	1648	2136	43,6	56,4
	Sim 629	58	571	90,8	9,2
% Total				50,3	49,7

Tabela 14: Matriz de Classificação – ponto de corte de 10%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	2243	1541	59,3	40,7
	Sim 629	106	523	83,1	16,9
% Total				62,7	37,3

Tabela 15: Matriz de Classificação – ponto de corte de 15%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	2749	1035	72,6	27,4
	Sim 629	161	468	74,4	25,6
% Total				72,9	27,1

Tabela 16: Matriz de Classificação – ponto de corte de 20%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	3095	689	81,8	18,2
	Sim 629	237	392	62,3	37,7
% Total				79,0	21,0

Tabela 17: Matriz de Classificação – ponto de corte de 30%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	3465	319	91,6	8,4
	Sim 629	342	287	45,6	54,4
% Total				85,0	15,0

Tabela 18: Matriz de Classificação – ponto de corte de 40%

		Predito			
		Inadimplência 90 dias			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	3637	147	96,1	3,9
	Sim 629	450	179	28,5	71,5
% Total				86,5	13,5

Tabela 19: Matriz de Classificação – ponto de corte de 50%

		Predito			
		perfil_mau			
	Observado	Não	Sim	% de acerto	% de erro
Inadimplência 90 dias	Não 3784	3732	52	98,6	1,4
	Sim 629	523	106	16,9	83,1
% Total				87,0	13,0

Analisando os resultados das Tabelas de 13 a 19, na medida em que o ponto de corte aumenta, o percentual de acerto dos clientes que apresentaram inadimplência superior a 90 dias diminui e, por outro lado, o percentual de acerto dos clientes que não apresentaram esta inadimplência aumenta. Conforme citado anteriormente, como deseja-se uma inadimplência baixa, mas sem perder "bons" clientes, o ponto de corte de 15% foi considerado o melhor deles, pois apresenta percentuais aceitáveis de acertos e erros, tanto para os clientes inadimplentes quanto para os adimplentes.

8.4.4. Curva ROC

A curva ROC, já descrita no item 7.1.3, é construída com informações sobre a sensibilidade do modelo contra sua especificidade, para a amostra de 70%. A curva obtida pelo modelo, Gráfico 2, plota a sensibilidade contra o valor de (1 – especificidade), que representa os falsos alarmes (clientes adimplentes classificados como inadimplentes).

Neste trabalho, obteve-se uma área sob a curva de 0,8, que pode ser considerada como uma "excelente discriminação" (*Hosmer*, *Lemeshow*, & *Sturdivant*, 2013).

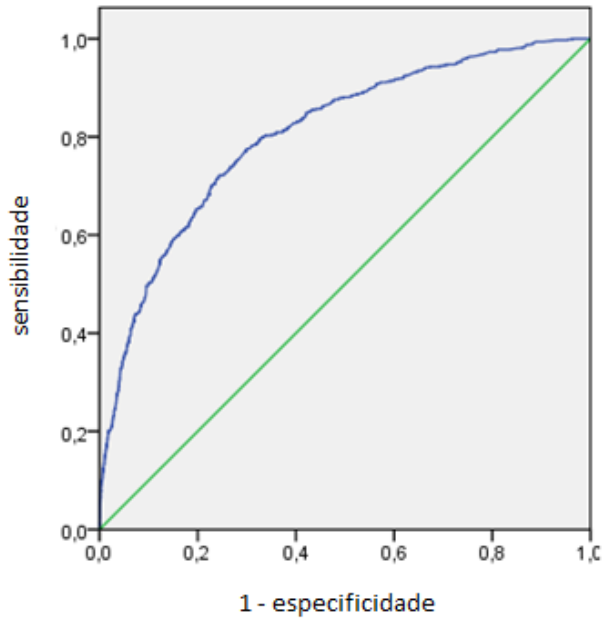


Gráfico 2 - Curva ROC do modelo ajustado

Área sob a curva

Area
0,805

Conforme descrito também no item 7.1.3, se quisermos escolher um ponto de corte ideal para efeitos de classificação, pode-se selecionar um ponto de corte que maximiza a sensibilidade e especificidade. Conforme Gráfico 3, um ponto de corte ideal poderia ser 0,15, que é aproximadamente onde as curvas de sensibilidade e especificidade se cruzam. Este ponto de corte de 15% foi representado na matriz de classificação da Tabela 15, que também demostrou ser o ponto de maior poder de discriminação dentre das demais tabelas.

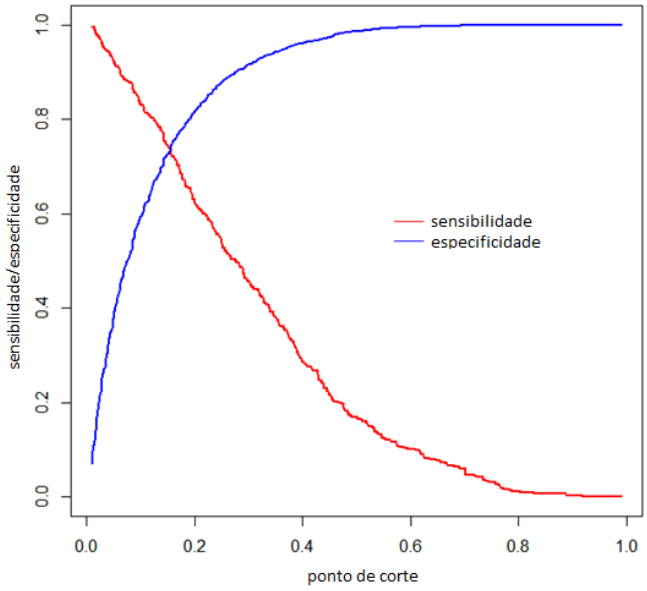


Gráfico 3 – Gráfico de sensibilidade e especificidade para vários pontos de corte

8.4.5. Estatística G

A Tabela 20 contém a estatística G do modelo ajustado. Sendo o p-valor pequeno, rejeita-se a hipótese nula de que os coeficientes são iguais a zero.

Tabela 20: Estatística G

Estatística G	p-valor
2891,5	8,15E-16

8.4.6. Análise da predição da inadimplência

Após a construção do modelo e da avaliação da qualidade de seu ajuste, será observada, nesta seção, a distribuição dos clientes dentre 10 grupos, conforme o decil, sendo o primeiro grupo com as menores probabilidades estimadas e o último grupo contendo as maiores probabilidades estimadas. O objetivo dessa análise é confrontar o ajuste do modelo com a inadimplência efetivamente observada em cada uma dos grupos.

As Tabelas de 21 a 23 demostram o percentual de real inadimplência em cada grupo. O índice de inadimplência na terceira coluna foi calculado dividindo-se a quantidade de clientes inadimplentes pelo total de clientes em cada grupo. Observa-se em todas estas tabelas uma gradação neste percentual de inadimplência, sendo muito menor dos grupos de baixa probabilidade de inadimplência e muito maior naqueles de alta probabilidade. Isto pode ser observado tanto na amostra utilizada na modelagem quanto na amostra de validação.

Analisando a amostra de 70% dos clientes – 4.413 clientes utilizados na modelagem:

Tabela 21 - Análise da inadimplência para 70% da amostra

Decil	Quantidade clientes	% inadimplência observado	% inadimplência previsto pelo modelo
1	428	0,9	0,7
2	429	3,7	1,9
3	445	3,8	3,3
4	449	5,8	5,2
5	457	6,1	7,6
6	438	8,0	10,7
7	441	13,8	14,6
8	447	19,0	20,0
9	422	28,2	28,6
10	457	52,1	49,0
Total	4413	14,3	14,3

Analisando a amostra de 30% dos clientes – 1.879 clientes não utilizados na modelagem:

Tabela 22: Análise da inadimplência para 30% da amostra

Decil	Quantidade clientes	% inadimplência observado	% inadimplência previsto pelo modelo
1	201	3,0	0,7
2	200	4,5	1,9
3	187	3,7	3,3
4	192	6,3	5,2
5	187	5,3	7,5
6	171	11,1	10,7
7	180	12,8	14,6
8	191	22,5	20,1
9	198	31,8	28,0
10	172	54,1	51,1
Total	1879	**15,2**	**13,9**

Analisando 100% dos clientes – 6.892 clientes:

Tabela 23: Análise da inadimplência para a amostra completa

Decil	Quantidade clientes	% inadimplência observado	% inadimplência previsto pelo modelo
1	629	1,6	0,7
2	629	4,0	1,9
3	632	3,8	3,3
4	641	5,9	5,2
5	644	5,9	7,6
6	609	8,9	10,7
7	621	13,5	14,6
8	638	20,1	20,0
9	620	29,4	28,4
10	629	52,6	49,6
Total	6292	14,5	14,2

As Tabelas 24 a 26 contêm a matriz de classificação, considerando o ponto de corte de 15%, para 70% da amostra, 30% da amostra, e para a amostra completa, respectivamente.

Observa-se em todas estas tabelas um percentual de acerto de aproximadamente 70%, tanto para os clientes inadimplentes quanto para os adimplentes, o que foi considerado um bom resultado.

Considerando o ponto de corte de 15%:

Analisando a amostra de 70% dos clientes – 4.413 clientes utilizados na modelagem:

Tabela 24 - Matriz de classificação para 70% da amostra

Observado		Predito		
		Inadimplência 90 dias		Percentual de acerto
		Não	Sim	
Inadimplência 90 dias	Não	2749	1035	72,6
	Sim	161	468	74,4
Percentual de acerto total				72,9

Analisando a amostra de 30% dos clientes – 1.879 clientes não utilizados na modelagem:

Tabela 25 - Matriz de classificação para 30% da amostra

Observado		Predito		
		Inadimplência 90 dias		Percentual de acerto
		Não	Sim	
Inadimplência 90 dias	Não	1167	427	73,2
	Sim	74	211	74,0
Percentual de acerto total				73,3

Analisando 100% dos clientes – 6.892 clientes:

Tabela 26 - Matriz de classificação para 100% da amostra

Observado		Predito		
		Inadimplência 90 dias		Percentual de acerto
		Não	Sim	
Inadimplência 90 dias	Não	3916	1462	72,8
	Sim	235	679	74,3
Percentual de acerto total				73,0

9. Considerações Finais

O presente trabalho teve como objetivo a construção de um modelo, utilizando a regressão logística, para a predição de inadimplência dos clientes de uma instituição financeira.

Foram detalhadas as principais etapas a serem percorridas na elaboração do modelo de regressão logística, contemplando: (a) método da máxima verossimilhança para estimação dos coeficientes do modelo, (b) escolha das variáveis, (c) ajuste do modelo, (d) medidas para diagnóstico do ajuste do modelo.

Por fim, utilizou-se um estudo de caso a fim de colocar em prática todas as etapas e conceitos do processo para elaboração do modelo de regressão logística. O modelo apresentou um bom resultado para predição da inadimplência, conforme as medidas de avaliação de ajuste do modelo consideradas, tais como o teste de *Hosmer-Lemeshow*, matriz de classificação, curva ROC.

A utilização deste modelo pela instituição financeira poderá evitar grupos de clientes com índices de inadimplência altos, a fim de se minimizar prejuízos financeiros e aumentar sua rentabilidade.

Referências Bibliográficas

Agresti, A. *Categorical Data Analysis.* 2ª ed. Hoboken: John Wiley & Sons. 2002. 710 p.

Bendel, R. B., Afifi, A. A. *Comparison of stopping rules in forward 'stepwise' regression*. Journal of the American Statistical Association. 1977.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. *Applied Logistic Regression.* 3ª ed. Hoboken: John Wiley & Sons. 2013. 500 p.

Mickey, J., and Greenland, S. *A study of the impact of confounder-selection criteria on effect estimation*. American Journal of Epidemiology. 1989.

Sicsú, L. A. *Credit Scoring: desenvolvimento, implantação, acompanhamento.* 1ª ed. São Paulo: Blucher. 2010. 180 p.

Anexo I – Método da Máxima Verossimilhança

Regressão Logística Simples:

Derivar a equação 5.3 em relação a βo:

$$\sum_{i=1}^{n} \{y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n} \{y_i \ln[p(x_i)] - y_i \ln[1 - p(x_i)] + \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n} \left\{y_i \ln\left[\frac{p(x_i)}{1 - p(x_i)}\right] + \ln[1 - p(x_i)]\right\}$$

$$\sum_{i=1}^{n} \left\{y_i (\beta_0 + \beta_1 x_i) + \ln\left[1 - \frac{e^{\beta o + \beta 1 x_i}}{1 + e^{\beta o + \beta 1 x_i}}\right]\right\}$$

$$\frac{\partial L(\beta)}{\beta o} = \sum_{i=1}^{n} \left\{y_i + \frac{\partial}{\beta o} \ln\left[1 - \frac{e^{\beta o + \beta 1 x_i}}{1 + e^{\beta o + \beta 1 x_i}}\right]\right\}$$

$$\sum_{i=1}^{n} \left\{y_i + \frac{\partial}{\beta o} \ln(1 + e^{\beta o + \beta 1 x_i})\right\}$$

$$\sum_{i=1}^{n} \left\{y_i - \frac{e^{\beta o + \beta 1 x_i}}{1 + e^{\beta o + \beta 1 x_i}}\right\}$$

$$\sum_{i=1}^{n} \{y_i - p(x_i)\}$$

Derivar a equação 5.3 em relação a β_1:

$$\sum_{i=1}^{n}\{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n}\{y_i \ln[p(x_i)] - y_i \ln[1 - p(x_i)] + \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n}\left\{y_i \ln\left[\frac{p(x_i)}{1 - p(x_i)}\right] + \ln[1 - p(x_i)]\right\}$$

$$\sum_{i=1}^{n}\left\{y_i\,(\beta_0 + \beta_1 x_i) + \ln\left[1 - \frac{e^{\beta o + \beta 1 x_i}}{1 + e^{\beta o + \beta 1 x_i}}\right]\right\}$$

$$\frac{\partial L(\beta)}{\beta_1} = \sum_{i=1}^{n}\left\{y_i x_i + \frac{\partial}{\beta_1}\ln\left[1 + e^{\beta o + \beta 1 x_i}\right]\right\}$$

$$\sum_{i=1}^{n}\left\{y_i x_i - x_i\left[\frac{e^{\beta o + \beta 1 x_i}}{1 + e^{\beta o + \beta 1 x_i}}\right]\right\}$$

$$\sum_{i=1}^{n} x_i\{y_i - p(x_i)\}$$

Regressão Logística Múltipla:

Conforme equação 4.3 temos:

$g(x_i) = \beta_0 + \beta_1\ x_{i1} + \cdots + \beta_p\ x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j\ x_{ij}$

Onde: $p =$ quantidade de variáveis, $j = 1, 2, \ldots p$

Derivar a equação 5.7 em relação a βo:

$$\sum_{i=1}^{n} \{y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n} \{y_i \ln[p(x_i)] - y_i \ln[1 - p(x_i)] + \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n} \left\{y_i \ln\left[\frac{p(x_i)}{1 - p(x_i)}\right] + \ln[1 - p(x_i)]\right\}$$

$$\sum_{i=1}^{n} \left\{y_i\ (\beta_0 + \beta' x_{ij}) + \ln\left[1 - \frac{e^{\beta o + \sum_{j=1}^{p} \beta_j\ x_{ij}}}{1 + e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}}}\right]\right\}$$

$$\frac{\partial L(\beta)}{\beta o} = \sum_{i=1}^{n} \left\{y_i + \frac{\partial}{\beta o} \ln\left[1 - \frac{e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}}}\right]\right\}$$

$$\sum_{i=1}^{n} \left\{y_i + \frac{\partial}{\beta o} \ln(1 + e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}})\right\}$$

$$\sum_{i=1}^{n} \left\{y_i - \frac{e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta o + \sum_{j=1}^{p} \beta_j x_{ij}}}\right\}$$

$$\sum_{i=1}^{n} \{y_i - p(x_i)\}$$

Derivar a equação 5.7 em relação a cada um dos coeficientes β_j :

$$\sum_{i=1}^{n}\{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n}\{y_i \ln[p(x_i)] - y_i \ln[1 - p(x_i)] + \ln[1 - p(x_i)]\}$$

$$\sum_{i=1}^{n}\left\{y_i \ln\left[\frac{p(x_i)}{1 - p(x_i)}\right] + \ln[1 - p(x_i)]\right\}$$

$$\sum_{i=1}^{n}\left\{y_i\,(\beta_0 + \sum_{j=1}^{p}(\beta_j\ x_{ij})) + \ln\left[1 - \frac{e^{\beta o + \sum_{j=1}^{p}\beta_j x_{ij}}}{1 + e^{\beta o + \sum_{j=1}^{p}\beta_j x_{ij}}}\right]\right\}$$

$$\frac{\partial L(\beta)}{\beta_j} = \sum_{i=1}^{n}\left\{y_i x_{ij} + \frac{\partial}{\beta_j}\ln\left[1 + e^{\beta o + \sum_{j=1}^{p}\beta_j x_{ij}}\right]\right\}$$

$$\sum_{i=1}^{n}\left\{y_i x_{ij} - x_{ij}\left[\frac{e^{\beta o + \sum_{j=1}^{p}\beta_j x_{ij}}}{1 + e^{\beta o + \sum_{j=1}^{p}\beta_j x_{ij}}}\right]\right\}$$

$$\sum_{i=1}^{n} x_{ij}\{y_i - p(x_i)\}$$

Anexo II – Deviance

Conforme equações 5.3 e 5.10, temos:

$$D = -2\,[ln(verossimilhança\ modelo\ ajustado) - ln(verossimilhança\ modelo\ saturado)]$$

Como no modelo saturado o valor predito é igual ao valor observado: $\hat{p}\,(x_i) = y_i$, temos:

$$D = \sum_{i=1}^{n}\{y_i\, ln[\hat{p}(x_i)] + (1 - y_i)\, ln[1 - \hat{p}(x_i)] - y_i\, ln[y_i] + (1 - y_i)\, ln[1 - y_i]\}$$

$$D = \sum_{i=1}^{n}\{y_i\, ln[\hat{p}(x_i)] - y_i\, ln[y_i] + (1 - y_i)\, ln[1 - \hat{p}(x_i)] - (1 - y_i)\, ln[1 - y_i]\}$$

$$D = \sum_{i=1}^{n}\{y_i\, [ln[\hat{p}(x_i)] - ln[y_i]] + (1 - y_i)[ln[1 - \hat{p}(x_i)] - ln[1 - y_i]]\}$$

$$D = \sum_{i=1}^{n}\left[yi\, ln\left(\frac{\hat{p}(xi)}{yi}\right) + (1 - yi)\, ln\left(\frac{1 - \hat{p}(xi)}{1 - yi}\right)\right]$$