UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

Andrea Kim

MODELOS LINEARES GENERALIZADOS:

APLICAÇÃO EM RISCO DE CRÉDITO

PRESIDENTE PRUDENTE
2022

Andrea Kim

MODELOS LINEARES GENERALIZADOS:

APLICAÇÃO EM RISCO DE CRÉDITO

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/UNESP para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientadora: Profa. Ma. Marta Yukie Baba

Coorientador: Prof. Dr. Sérgio Minoru Oikawa

PRESIDENTE PRUDENTE
2022

K49m

Kim, Andrea
Modelos Lineares Generalizados : Aplicação em risco de crédito / Andrea Kim. -- Presidente Prudente, 2022
49 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Marta Yukie Baba
Coorientador: Sérgio Minoru Oikawa

1. Modelos Lineares Generalizados. 2. Risco de Crédito. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Andrea Kim

MODELOS LINEARES GENERALIZADOS:

APLICAÇÃO EM RISCO DE CRÉDITO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____
Profa. Ma. Marta Yukie Baba
Departamento de Estatística

Co-orientador: _____
Prof. Dr. Sérgio Minoru Oikawa
Departamento de Estatística

_____
Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

_____
Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística

Presidente Prudente, 25 de março de 2022.

Resumo

Modelos lineares generalizados é uma extensão dos modelos lineares gerais (regressão simples e múltipla, planejamento de experimentos, análise multivariada, etc.) cujas distribuições de probabilidades envolvidas no modelo pertencem à família exponencial, no qual é colocado uma função de ligação relacionando a média da variável resposta com as variáveis explicativas. A estimação dos parâmetros desse modelo pode ser realizada pelo método de estimação de máxima verossimilhança. Nesse vasto campo da estatística, existe um grande interesse das instituições, em especial as financeiras, em encontrar um modelo preditivo para o tema risco de crédito, que é a possibilidade de uma contraparte da operação não honrar a dívida, ou seja, a inadimplência. Na inadimplência, um cliente pode ser considerado bom ou mau pagador. Utilizou-se, portanto, modelos lineares generalizados para ajustar um modelo cuja variável resposta pertence a uma distribuição binomial. Os dados utillizados neste trabalho foram apresentados no artigo “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications” (Yeh, I. C., & Lien, C. H.) (2009).

Palavras-chave: Modelos Lineares Generalizados, Regressão Logística, Risco de crédito.

Abstract

Generalized linear models is an extension of the general linear models (single and multiple regression, design of experiments, multivariate analysis, etc.) whose probability distributions involved in the model belong to the exponential family, in which a link function relating the mean of the response variable with the explanatory variables. The estimation of the parameters of this model can be performed using the maximum likelihood estimation method. In this vast field of statistics, there is great interest from institutions, especially financial ones, in finding a predictive model for the topic of credit risk, which is the possibility of a counterparty to the operation not honoring the debt, that is, default. In default, a customer can be considered a good or bad payer. Therefore, generalized linear models were used to fit a model whose response variable belongs to a binomial distribution. The data used in this work were presented in the article “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications” (Yeh, I.C., & Lien, C.H.) (2009).

Key-words: Generalized linear models, Logistic Regression, Credit Risk.

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 Introdução

No começo dos anos 70, houve o aumento da volatilidade dos mercados financeiros, em que a imprevisibilidade e constante mudanças do cenário econômico geravam grandes perdas financeiras, observando assim, a necessidade de um gerenciamento de risco que forneça proteção parcial contra fontes de risco, sendo uma delas, o risco de crédito (JORION, 2004).

Define-se análise de risco de crédito como a probabilidade de perda em uma operação de crédito de uma empresa. O risco de crédito pode ser efetuado de duas maneiras: subjetivamente ou objetivamente. A avaliação subjetiva incorpora a experiência do analista, mas não quantifica o risco de crédito. Por outro lado, a medida objetiva cria um modelo estatístico potencial capaz de prever tais riscos de inadimplência, ou seja, existe a necessidade de quantificar o risco objetivamente usando metodologia quantitativa (utilizando modelos estatísticos) (SICSÚ, 2010).

Com análise objetiva, é possível tomar decisões como por exemplo, se o cliente atrasa o pagamento, mas paga com juros; conceder ou não crédito; probabilidade de pagamento maior do que não pagamento da dívida; dentre outros.

Com tais informações as empresas têm o intuito de oferecer crédito ao cliente sem que haja uma perda da parte dela, ou seja, que não existam (ou poucos) clientes inadimplentes. Para isso, essas empresas estão buscando modelos estatísticos que consigam avaliar se o cliente será um bom pagador ou não.

Existem muitos métodos estatísticos para essa análise, por exemplo: análise discriminante, redes neurais, regressão logística, análise de sobrevivência, algoritmos genéticos, detecção automática de interação (AID), dentre outros.

No presente estudo, será abordado a metodologia de modelos lineares generalizados, pois não se restringe apenas à distribuição normal como nos casos de modelos de regressão linear, modelos de planejamento de experimentos, modelos de análise multivariada, etc. Sendo assim, não é necessário testar os pressupostos como a homocedasticidade e normalidade. Os Modelos Lineares Generalizados (MLG), em inglês *Generalized Linear Models* (*GLM*), utilizam as distribuições de probabilidade da família exponencial, o que possibilita uma utilização muito mais ampla. Com isso, estudar diversas aplicações para construir um modelo em várias áreas. Alguns métodos estatísticos de MLG são regressão logística, análise de sobrevivência, regressão de Poisson, entre outros.

Os modelos de previsão produzem um resultado numérico que gerará uma classificação (Exemplo: score de inadimplência). Este resultado numérico poderá ser a probabilidade obtida da função logística. Tal classificação será efetuada pela definição de um ponto de corte para o resultado numérico mencionado, que poderá classificar o cliente como inadimplente ou não. Abaixo do ponto de corte, o novo candidato será considerado propenso à inadimplência (BARTH, 2004).

Neste trabalho o objetivo é desenvolver um modelo para o risco de crédito com a proposta de achar um modelo que mais se encaixa para a classificação dos clientes com base na teoria de Modelos Lineares Generalizados. Outro objetivo é aplicar em dados reais para verificar se o modelo se ajusta adequadamente.

A importância deste trabalho é que se sabe que ainda não existem modelos que consigam obter precisão absoluta e qualquer avanço em termos da precisão da previsão se faz necessário gerando ganhos para as instituições. Por esse fato, o intuito deste estudo deve-se a grande necessidade de ajustar um bom modelo capaz de prever a inadimplência de clientes nas instituições, em especial, as instituições financeiras. Com isso, um estudo introdutório sobre o tema de Modelos Lineares Generalizados será feito para explanar a aplicação em risco de crédito. Será utilizado neste trabalho o modelo de regressão logística.

2 Modelos Lineares Generalizados

2.1 Conceito Básico

Modelos Lineares Generalizados (MLG) foi proposto em 1972 por John Nelder e Robert Wedderburn, como uma maneira de unificar os modelos de regressão linear e não linear sob um só marco teórico. Com isso, é possível desenvolver um algoritmo geral para a estimativa de máxima verossimilhança para os parâmetros das distribuições de probabilidade em todos estes modelos.

Estes modelos relacionam a distribuição de probabilidade da variável aleatória dependente no experimento com a parte sistemática (não aleatória ou preditor linear) através de uma função chamada função de ligação (LINDSEY, 1997), permitindo que a distribuição da variável resposta pertença à família exponencial de

distribuições, mas não necessariamente apenas à distribuição Normal, conhecida também como distribuição Gaussiana.

A variável resposta pode ser contínua ou discreta. Ainda assim, possui uma limitação na linearidade e na independência das respostas por causa da família exponencial. As covariáveis (determinísticas ou estocásticas) podem ser quantitativas ou qualitativas (nominais ou ordinais) que são representadas por um vetor de variáveis explicativas (TURKMAN; SILVA, 2000).

Alguns exemplos de métodos estatísticos de MLG (CORDEIRO; DEMÉTRIO, 2007) são:

- Modelo clássico de regressão múltipla e análise de variância para experimentos planejados cujo erro aleatório assume distribuição normal;
- Modelo complemento log-log para ensaios de diluição, modelo logístico e probito para estudo de proporções (todos envolvendo distribuição binomial);
- Modelos log-lineares para análise de contagem em tabelas de contingência (distribuição de Poisson e multinomial);
- Modelos de testes de vida (distribuição exponencial);
- Modelo estrutural para dados (distribuição gama);
- Modelo de regressão não simétrica;
- Entre outros.

2.2 Família Exponencial

A distribuição de uma variável aleatória Y pertence à família exponencial se a sua função densidade de probabilidade puder ser escrita da seguinte forma:

$$f_Y(y|\theta) = h(y) \exp\{ \delta(\theta)\, a(y) - b(\theta) \} \qquad (1)$$

para θ uniparamétrico e escolhas apropriadas para as funções reais $h(\cdot)$, $\delta(\cdot)$, $a(\cdot)$ e $b(\cdot)$, que são conhecidas e definidas de modo que satisfaçam a equação (1) acima.

Alternativamente,

$$f_Y(y|\theta) = C(\theta)\, H(y) \exp\{\, D(\theta)\, A(y)\, \theta\, B(\theta)\,\} \tag{2}$$

Para membros dessa família, a estatística $A(y)$ é uma estatística suficiente para θ, pelo teorema da fatoração de Neyman-Fisher.

Observação: Se $f(y_i|\theta) = h(y_i) \exp\{\, \delta(\theta)\, a(y_i) - b(\theta)\,\}$, então para y_i i.i.d tem-se:

$$f(\mathbf{y}\,|\,\theta) = \prod_{i=1}^{n} f(y_i\,|\,\theta) = \left[\prod_{i=1}^{n} h(y_i)\right] \exp\left\{\delta(\theta)\ \sum_{i=1}^{n} a(y_i) - n\, b(\theta)\right\}. \tag{3}$$

Forma Canônica: Quando em (1) as funções h e δ assumirem a identidade, ou seja, $h(y) = y$ e $\delta(\theta) = y$ tem-se

$$f_Y(y|\theta) = h(y) \exp\{\, \theta\, y - b(\theta)\,\} \tag{4}$$

Adicionando um parâmetro de escala, o modelo (4) torna-se

$$f_Y(y|\theta) = h(y) \exp\left\{\frac{\theta\, y - c(\theta)}{d(\varphi)} + t(y,\varphi)\right\} \tag{5}$$

Vale ressaltar que, se o parâmetro φ não for conhecido, a distribuição de probabilidade pode não pertencer a família exponencial biparamétrica.

A seguir, alguns exemplos de distribuições que podem ser escritas dessa forma.

2.2.1 Distribuição Bernoulli

A variável aleatória X é definida como distribuição bernoulli com um parâmetro se a função de densidade discreta de X é dada por:

$$f_X(x\,|\,n,p) = p^x\,(1-p)^{\,1-x}\, I_{\{0,\,1\}}(x) \tag{6}$$

com o parâmetro p satisfazendo o intervalo $0 \le p \le 1$.

A esperança e a variância, respectivamente, são dadas por:

$$E(X) = p \quad \text{e} \quad Var(X) = p\,(1-p) \tag{7}$$

Logo, $f_X(x\,|\,n,p) = \exp\left\{\ln\left[\, p^x (1-p)^{1-x}\,\right]\right\} =$

$$= \exp\left\{\ x \ln(p) + (1-x)\ln(1-p)\ \right\} =$$

$$= \exp\{\, x \ln(p) + \ln(1-p) - x \ln(1-p)\,\} =$$

$$= \exp\{\, x\,[\,\ln(p) - \ln(1-p)\,] + \ln(1-p)\,\} = h(x)\,\exp\{\,\delta(\theta)\,a(x) - b(\theta)\,\}$$

(8)

com $h(x) = 1$; $\delta(\theta) = [\,\ln(p) - \ln(1-p)]$; $a(x) = x$ e $b(\theta) = -\ln(1-p)$

2.2.2 Distribuição Binomial

A variável aleatória X é definida como distribuição binomial com um parâmetro se a função de densidade discreta de X é dada por:

$$f_X(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x} I_{\{0,\,1,\,\ldots\,,\,n\}}(x) \qquad (9)$$

com o parâmetro p satisfazendo o intervalo $0 \le p \le 1$.

A esperança e a variância, respectivamente, são dadas por:

$$E(X) = n\,p \quad \text{e} \quad Var(X) = n\,p\,(1-p) \qquad (10)$$

Logo, $f_X(x \mid n, p) = \exp\left\{\, \ln\left[\binom{n}{x} p^x\,(1-p)^{n-x}\right] \right\} =$

$$= \exp\left\{\, \ln\binom{n}{x} + x\ln(p) + (n-x)\ln(1-p) \,\right\} =$$

$$= \binom{n}{x} \exp\{\, x\ln(p) + n\ln(1-p) - x\ln(1-p)\,\} =$$

$$= \binom{n}{x} \exp\left\{\, x\left[\,\ln(p) - \ln(1-p)\,\right] + n\ln(1-p) \,\right\} =$$

$$= \binom{n}{x} \exp\left\{\, x\ln\left(\frac{p}{1-p}\right) + \ln(1-p)^n \,\right\} = h(x)\,\exp\{\,\delta(\theta)\,a(x) - b(\theta)\,\}$$

(11)

com $h(x) = \binom{n}{x}$; $\delta(\theta) = \ln\left(\frac{p}{1-p}\right)$; $a(x) = x$ e $b(\theta) = -\ln(1-p)^n$

2.2.3 Distribuição Poisson

A variável aleatória X é definida como distribuição de Poisson com um parâmetro se a função de densidade discreta de X é dada por:

$$f_X(x \mid \lambda) = \frac{e^{-\lambda}\,\lambda^x}{x!}\, I_{\{0,\,1,\,\ldots\}}(x) \tag{12}$$

com o parâmetro λ satisfazendo $\lambda > 0$.

A esperança e a variância, respectivamente, são dadas por:

$$E(X) = \lambda \quad \text{e} \quad \text{Var}(X) = \lambda \tag{13}$$

Logo, $f_X(x \mid \lambda) = \frac{e^{-\lambda}\,\lambda^x}{x!} = \frac{1}{x!}\, e^{-\lambda} \exp\{\ln(\lambda^x)\} = \frac{1}{x!} \exp\{x \ln(\lambda) - \lambda\}$

$$f_X(x \mid \lambda) = \frac{1}{x!} \exp\{x \ln(\lambda) - \lambda\} = h(x) \exp\{\delta(\lambda)\, a(x) - b(\lambda)\} \tag{14}$$

com $h(x) = \frac{1}{x!}$, $\delta(\lambda) = \ln(\lambda)$, $a(x) = x$ e $b(\lambda) = \lambda$.

2.2.4 Distribuição Normal

A variável aleatória X é definida como distribuição Normal com dois parâmetros se a função de densidade contínua de X é dada por:

$$f_X(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{15}$$

com o parâmetro $-\infty < \mu < +\infty$ e $\sigma > 0$ (estritamente positivo).

A esperança e a variância, respectivamente, são dadas por:

$$E(X) = \mu \quad \text{e} \quad \text{Var}(X) = \sigma^2 \tag{16}$$

Logo, $f_X(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} =$

$$= \exp\left\{\ln\left[\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\right]\right\} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} =$$

$$= \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2)\right\} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} =$$

$$= \exp\left\{ -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \left(\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} \right) \right\} =$$

$$= \exp\left\{ \frac{1}{\sigma^2}\left(x\mu - \frac{\mu^2}{2} \right) - \frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{x^2}{2\sigma^2} \right\} = \text{pela equação (5)}$$

$$= h(x)\, \exp\left\{ \frac{\theta\, y - b(\theta)}{\varphi} + c(y, \varphi) \right\} \qquad (17)$$

com $h(x) = 1$; $\theta = \mu$; $\varphi = \sigma^2$; $b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$ e $c(y, \varphi) = -\frac{1}{2}\left[\frac{x^2}{\sigma^2} + \ln\left(2\pi\sigma^2\right) \right]$

Existe uma relação especial entre a média e a variância que pode ser apresentada da seguinte forma: para qualquer função de verossimilhança, $L(\theta \mid y)$ com logaritmo da função de verossimilhança $l(\theta \mid y) = \ln\{ f(y\,;\theta) \}$ é dada por:

$$l(\theta \mid y) = \ln[\, w(y) \exp\{ \delta(\theta)\, a(y) - b(\theta) \}\,] = \ln[w(y)] + \delta(\theta)\, a(y) - b(\theta)$$

Função Score: De (4) o logaritmo da função de verossimilhança é dado por:

$$l(\theta \mid y) = \theta\, y - b(\theta) + \ln[\, w(y)\,] \qquad (18)$$

com função score dada por $U(\theta) = \frac{d\, l(\theta)}{d\,\theta} = y - b'(\theta)$.

Pelas propriedades da função score tem-se que:

$$E(U) = 0 \quad \text{e} \quad Var(U) = -E\left(\frac{d^2\, l(\theta)}{d\,\theta^2} \right) \qquad (19)$$

Portanto, $E(Y) = b'(\theta)$ e $Var(X) = b''(\theta)$.

2.3 Descrição do Modelo

Suponha que o modelo que será estudado, tenha Y como variável resposta (variável dependente) e um vetor $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ composto de k variáveis independentes (variáveis explicativas).

Assim, os dados coletados serão da seguinte forma:

$(\, y_i\, , \mathbf{x}_i\,)$ para $i = 1, 2, \ldots, n$.

Matricialmente, tem-se

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{e} \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Considere:

(P) $E(Y_i \mid \mathbf{x}_i) = \mu_i = b'(\theta_i)$ para i = 1, 2, … , n ; ou seja, $Y_i = \mu_i + \varepsilon_i$ com $E(\varepsilon_i) = 0$.

(ii) $Var(Y_i \mid \mathbf{x}_i) = \dfrac{d^2\, b(\theta_i)}{d\,\theta_i^2} = \dfrac{d\,\mu_i}{d\,\theta_i}$

Seja $Var(\mu_i) = Var(Y_i) = \dfrac{d\,\mu_i}{d\,\theta_i}$ que denota a dependência da variância da resposta sobre sua média.

Consequentemente, $\dfrac{d\,\theta_i}{d\,\mu_i} = \dfrac{1}{Var(\mu_i)}$.

2.4 Preditor Linear e Função de Ligação

Defina η_i como o preditor linear do seguinte modo:

$$\eta_i = g[\,E(Y_i)] = g(\mu_i) = \mathbf{x}_i'\,\boldsymbol{\beta}$$

Logo, $E(Y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\,\boldsymbol{\beta})$ com a função g sendo a função de ligação.

Existem várias possibilidades para a função de ligação g, que é uma função monótona e diferenciável.

Em particular,

$$\eta_i = \theta_i$$

é dita que η_i é a ligação canônica.

Exemplo de algumas ligações canônicas:

Tabela 1: Ligação canônica

Distribuição	Função de Ligação canônica	
Poisson	Log	$\ln(\mu) = \eta$
Binomial	Logit	$\ln\left(\frac{\mu}{n-\mu}\right) = \eta$
Normal	Identidade	$\mu = \eta$
Gama	Recíproca	$\left(\frac{1}{\mu}\right) = \eta$

Fonte: Turkman e Silva, 2000

A função de ligação e o preditor linear são fundamentais, pois são esses fatores que determinarão qual modelo será apropriado para construir a estrutura linear. A escolha da função de ligação depende do tipo da variável resposta do experimento a ser executado.

Na tabela a seguir, são colocados alguns exemplos de funções de ligação:

Tabela 2: Função de ligação

Identidade	μ
Recíproca	$\left(\frac{1}{\mu}\right)$
Log	$\ln(\mu)$
Logit	$\ln\left(\frac{\mu}{n-\mu}\right)$
Complementar log-log	$\ln\left(-\ln\left(\frac{\mu}{n}\right)\right)$
Probit	$\Phi^{-1}\left(\frac{\mu}{n}\right)$

Fonte: Turkman e Silva, 2000

2.5 Inferência Estatística

O objetivo da inferência estatística é fazer afirmações ou tomar decisões baseado no teste de hipóteses, com um nível de significância fixado, a partir de um conjunto de valores representativo (amostra) sobre um universo (população).

Obviamente, assume-se que a população é muito maior do que a amostra, que é o conjunto de dados observados.

Fazer tal afirmação ou tomar a decisão, deve sempre ser analisada acompanhada de uma medida de precisão sobre o objeto de estudo. Outro ponto importante para que o estudo seja confiável, é que a amostra precisa ser provida da aleatorização para que todas as unidades experimentais tenham a mesma chance de serem selecionados.

Dois tópicos importantes da inferência estatística que serão abordados neste trabalho são a estimação de parâmetros e o teste de hipóteses.

2.5.1 Estimação de Parâmetros

A estimação de parâmetros pode ser pontual (estima um único valor do parâmetro) ou intervalar (estima um intervalo para o valor do parâmetro, com um nível de confiança) e será estudada neste trabalho, baseada no método de estimação da máxima verossimilhança para os parâmetros de interesse.

Utilizando a equação (5) da seção 2.2, tem-se que a função de verossimilhança, como função de $\boldsymbol{\beta} = (\theta, \varphi)$, é dada por:

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} h(y_i) \exp\left\{\frac{y_i\,\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right\}$$

$$= \left[\prod_{i=1}^{n} h(y_i)\right] \exp\left\{\frac{1}{\varphi} \sum_{i=1}^{n} [\, y_i\,\theta_i - b(\theta_i)] + c(y_i, \varphi)\right\}$$

Aplicando logaritmo nessa função, tem-se que

$$\ln[L(\boldsymbol{\beta}|\mathbf{y})] = l(\boldsymbol{\beta}|\mathbf{y}) = \frac{1}{\varphi} \sum_{i=1}^{n} \left[\, y_i\,\theta_i - b(\theta_i)\right] + c\left(y_i, \varphi\right)$$

Para função de ligação canônica, tem-se que

$\eta_i = g\,[\,E(y_i)\,] = g(\mu_i) = \mathbf{x}_i'\,\boldsymbol{\beta}$

Logo,

$$\frac{\partial\, l(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\, \boldsymbol{\beta}} = \frac{\partial l}{\partial \theta_i}\,\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{\varphi}\sum_{i=1}^{n}\left[y_i - \frac{db(\theta_i)}{d\theta_i}\right]\mathbf{x}_i = \frac{1}{\varphi}\sum_{i=1}^{n}(y_i - \mu_i)\;\mathbf{x}_i$$

Derivando e igualando a zero, tem-se os estimadores de máxima verossimilhança dos parâmetros resolvendo o sistema com $(k+1)$ equações (um para cada um dos parâmetros do modelo) dado por:

$$\frac{1}{\varphi}\sum_{i=1}^{n}(y_i - \mu_i)\;\mathbf{x}_i = \mathbf{0}$$

Em geral φ é uma constante, o que simplifica o sistema acima como

$$\sum_{i=1}^{n}(y_i - \mu_i)\;\mathbf{x}_i = \mathbf{0}$$

Matricialmente, a expressão acima pode ser colocada da seguinte forma:

$$\mathbf{X}^{\mathrm{t}}\,(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}$$

com $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{k+1})$, que são chamadas de equações score de máxima-verossimilhança.

Numericamente, ele é resolvido por processos iterativos de Newton-Raphson, que por vez, uma equação $f(x) = 0$ é baseado na aproximação de Taylor para a função $f(x)$ na vizinhança do ponto x_0, ou seja,

$$f(x) = f(x_0) + (x - x_0)f_0(x_0) = 0$$

Obtendo assim,

$$x = x_0 - \frac{f(x_0)}{f_0\prime(x_0)}$$

O método de Newton-Raphson é útil quando as derivadas parciais de segunda ordem são facilmente analisadas. Porém, como nem sempre ocorre isso, usa-se o método escore de Fisher que envolve a substituição da matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais. Encontrando assim,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{K}^{(m)})^{-1}\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\right)^{(m)}$$

Onde $\boldsymbol{\beta}^{(\mathrm{m+1})}$ e $\boldsymbol{\beta}^{(\mathrm{m})}$ são os vetores de parâmetros estimados nos passos $(\mathrm{m+1})$ e (m) e **K** é a matriz Informação de Fisher,

$$\boldsymbol{K} = \phi^{-1}\boldsymbol{X}^{t}\boldsymbol{W}\boldsymbol{X}$$

e

$$\boldsymbol{W} = diag\{w_1, \ldots, w_n\}$$

que traz informação sobre a distribuição e a função de ligação usada.

2.6 Testes de Hipóteses

De acordo com Paula (2013), os principais testes de hipóteses em MLG são teste de Wald, teste da razão de verossimilhança e o teste escore para as hipóteses simples.

Supondo as seguintes hipóteses:

H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ $\quad$ H_1: $\boldsymbol{\beta} \neq \boldsymbol{\beta}^0$ com nível de significância α

Sendo $\boldsymbol{\beta}^0$ um vetor conhecido com dimensão p e Φ também é conhecido.

1 – Teste de Wald

O teste Wald é escrito pela expressão:

$$\mathbf{W} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)^T \widehat{Var}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \sim \chi_p^2$$

Sendo $\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{K}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)$ é a matriz de variância covariância assintótica de $\widehat{\boldsymbol{\beta}}$.

Em particular, para p = 1, o teste de Wald é equivalente ao teste t^2.

$$\mathbf{W} = \frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)^2}{\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)}$$

Segundo Hauck e Donner (1977), para o teste com p = 1 e amostras pequenas, o teste de Wald mostra comportamentos estranhos onde o p-valor é consideravelmente maior do que o p-valor para o teste da razão de verossimilhanças, ou seja, para o teste Wald, podemos não rejeitar H_0 quando rejeitaria para o teste da razão de verossimilhanças. Assim, é recomendado a utilização do teste da razão de verossimilhanças quando não rejeitar H_0 no teste Wald.

2 – Teste da razão de verossimilhanças

O teste da razão de verossimilhanças, também chamada de estatística de Wilks, é definido por:

$$RV = 2\left[L(\widehat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}^0)\right] \sim \chi_p^2$$

Sob H_0 e com Φ conhecido.

Esta estatística também pode ser expressa como a diferença entre duas funções desvio.

Em particular, para o caso da normal linear, temos

$$RV = \frac{[\sum_{i=1}^{n}(y_i - \hat{\mu}_i^0)^2 - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2]}{\sigma^2}$$

Quando Φ é desconhecido o teste da razão de verossimilhanças tem melhor aproximação pela distribuição F.

3 – Teste escore

Teste escore, conhecido também como teste de Rao, é definido quando a função escore $U_\beta(\widehat{\boldsymbol{\beta}}) = 0$ por

$$SR = U_\beta(\boldsymbol{\beta}^0)^T \widehat{Var_0}(\widehat{\boldsymbol{\beta}}) U_\beta(\boldsymbol{\beta}^0) \sim \chi_p^2$$

Sendo a $\widehat{Var_0}(\widehat{\boldsymbol{\beta}})$ é a variância assintótica de $\widehat{\boldsymbol{\beta}}$ sob H_0.

Essa estatística é conveniente quando a hipótese alternativa é mais complicada que a hipótese nula, sendo necessário estimar os parâmetros sob H_1 somente quando H_0 fosse rejeitada.

A normal linear é um caso particular onde a estatística de escore é representada na forma:

$$SR = \frac{(y - X\boldsymbol{\beta}^0)^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^T(y - \mathbf{X}\boldsymbol{\beta}^0)}{\sigma^2}$$

2.7 Seleção de modelo

Um modelo com menor número de variáveis explicativas que ofereça melhor interpretação do problema e tenha um bom ajuste dos dados é considerado um modelo “adequado”. Para isso, deve existir um equilíbrio entre bom ajustamento, parcimônia e interpretação.

É possível selecionar um modelo através das variáveis independentes a serem incluídas na estrutura linear, com m covariáveis e conjunto de 2^m modelos. O objetivo é selecionar um modelo de $p \leq m$ covariáveis cujos valores ajustados expliquem adequadamente os dados. Se m for grande, o esforço computacional é grande e, por isso, inviável o exame de todos os modelos.

2.8 Desvio

Nelder e Wedderburn (1972) propuseram a medida “deviance” a fim de testar a significância dos coeficientes. A análise da deviance é definida como o desvio de um modelo de p parâmetros em relação ao modelo saturado, aquele que contém um parâmetro para cada número de variáveis.

A expressão do desvio é:

$$S_p = \frac{D_p}{a(\phi)} \quad ,$$

sendo $S_p = 2(l_s(\widehat{\boldsymbol{\beta}}_s) - l_{mod}(\widehat{\boldsymbol{\beta}}_{mod}))$

O máximo do logaritmo da função de verossimilhança é $l_s(\widehat{\boldsymbol{\beta}}_s)$ é o máximo do logaritmo da função de verossimilhança do modelo saturado e $l_{mod}(\widehat{\boldsymbol{\beta}}_{mod})$

é o máximo do logaritmo da função de verossimilhança do modelo atual. Quanto menor o valor do desvio D_p, obtemos um ajuste tão bom quanto do modelo saturado. No caso geral, S_p tem, assintoticamente, distribuição χ^2 com n-p graus de liberdade. Segundo Turkman & Silva (2000), esta aproximação não é adequada até mesmo para amostras grandes. Por isso, a análise de desvio vale teoricamente, mas na prática não se mostra adequada, apesar de serem feitas comparações do valor observado com desvio na prática.

Denotando $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\tilde{\theta}_i = \theta_i(\tilde{\mu}_i)$ por estimativas de máxima verossimilhança de θ temos que a função S_p também pode ser escrita da seguinte forma:

$$S_p = 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))]$$

Os desvios mais comuns são:

Tabela 3: Principais desvios de MLG

MLG	Desvio
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$
Gama	$2\sum_{i=1}^{n}\left[-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}\right]$
Binomial	$2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m - y_i)\ln\frac{(m - y_i)}{(m - \hat{\mu}_i)}\right]$
Normal Inversa	$\frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i{}^2 y_i}$

Fonte: Coladello (2011)

2.9 Resíduos

Assim como em modelos de regressão linear simples ou múltipla, em MLG também podemos obter resultados não satisfatórios na obtenção do modelo pela diferença do valor observado com o valor ajustado, ou então, por existirem valores observados discrepantes em relação aos demais. (CORDEIRO, 2007)

Essas diferenças podem ocorrer pela escolha inadequada da função de variância, função de ligação e da matriz de modelo, ou pela definição errada da escala da variável dependente. Existem algumas técnicas para verificação do ajuste do modelo como, por exemplo, a visualização gráfica ou pela inclusão de um parâmetro extra no modelo baseados em testes como a verossimilhança e escore.

No modelo de regressão clássico admite-se que os resíduos são independentes e que possuem distribuição normal $N(0,\sigma^2)$, se esses pressupostos são violados, a análise pode ter resultados duvidosos. Originando, assim, a não linearidade, não normalidade, heterocedasticidade, não independência, entre outros. Pontos atípicos também podem interferir no ajuste do modelo.

Para a verificação da pressuposição de linearidade, devemos usar a variável dependente ajustada e o preditor linear, e a variância residual é substituída pela estimativa consistente do parâmetro, além da matriz de projeção ser definida por:

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\boldsymbol{X}^{\boldsymbol{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{W}^{\frac{1}{2}}$$

Em que $\mathbf{H}$ depende das variáveis explicativas, da função de ligação e da função de variância.

Coladello (2011) cita três tipos de resíduos:

1. Resíduo deviance;
2. Resíduo deviance estudentizado;
3. Distância de Cook.

1. Resíduo deviance

O resíduo deviance é expresso por

$$r_{dev(i)} = sinal(y_i - \hat{\mu}_i)\sqrt{d_i}$$

Sendo $sinal(.) = \begin{cases} 1 & se \geq 0 \\ -1 & se < 0 \end{cases}$

2. Resíduo deviance estudentizado

Sendo a matriz $\mathbf{H}$ “chapéu” definida por:

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{W}^{\frac{1}{2}},$$

então o resíduo estudentizado é:

$$r_i = \frac{r_{dev(i)}}{\sqrt{\hat{\phi}^2(1-h_{ii})}}$$

O i-ésimo elemento da diagonal principal da matriz **H** é h_{ii} e $\hat{\phi}^2 = \frac{S_p}{n-p}$.

3. Distância de Cook

A distância de Cook é definida por:

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^{T}(\mathbf{X}^{T}\mathbf{W}\mathbf{X})\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{p\hat{\phi}}$$

Sendo $\widehat{\boldsymbol{\beta}}_{(\mathrm{i})}$ é o vetor de valores estimados sem a i-ésima observação. A distância de Cook representa uma mensuração na mudança de um ajustamento se um único valor é desconsiderado do conjunto de dados. (COLADELLO, 2011)

2.10 Seleção de variáveis

Supondo um modelo com uma variável resposta Y e variáveis explicativas x_1, x_2, x_3, ..., x_{k}, deseja-se selecionar alguma delas para que a estimativa do erro seja menor possível. Existem alguns métodos estatísticos como Forward, Backward e Stepwise, que serão descritos resumidamente a seguir.

(i) Forward

Passo 1: O método de seleção do tipo forward inicia com nenhuma variável, apenas o intercepto, e define o nível de significância α. A seguir, são calculados o módulo do coeficiente de correlação entre cada uma das variáveis explicativas com a variável resposta, sendo escolhida a variável explicativa com maior módulo de coeficiente de correlação e efetuando a regressão. Caso o valor da estatística da razão de

verossimilhanças dessa variável indique sua não rejeição, essa variável será incluída no modelo.

Passo 2: Novamente é calculado o módulo do coeficiente de correlação da variável resposta Y com as demais variáveis explicativas não incluídas no passo anterior, e a variável explicativa com o maior coeficiente de correlação será incluída no modelo caso a sua contribuição estatística for significante.

Passo 3: Efetuamos novamente o passo 2 até que todas as variáveis estejam inclusas no modelo ou alguma variável tenha sido rejeitada pelo teste da razão de verossimilhanças. Assim, essa variável rejeitada não entra no modelo, encerrando o procedimento. (MARQUES, 2018)

(ii) Backward

Em contrapartida do método de seleção Forward, o Backward faz o caminho contrário, começando com todas as variáveis e eliminando ou não alguma variável em sua próxima etapa.

Para isso, é analisado a estatística da razão de verossimilhanças para cada variável explicativa.

Passo 1: O menor valor da estatística é comparado com o nível de significância e se este menor valor for menor que o nível de significância, então esta variável é eliminada do modelo.
Passo 2: São calculadas novamente as estatísticas do teste e o passo 1 é repetido até que o menor valor da estatística seja maior que o nível de significância.

(iii) Stepwise

Passo 1: Assim como o método Forward, o Stepwise começa apenas com o intercepto. Incluímos a variável que tiver maior correlação e verificamos se a estatística do teste é menor do que o nível de significância para a inclusão da próxima variável.

Passo 2: A cada passo do método Forward, aplicamos o método Backward para verificar se existe alguma variável que pode ser eliminada. E continuamos o processo até a inclusão ou não de alguma variável.

2.11 Critérios de avaliação

Usa-se o critério de avaliação para a escolha do melhor submodelo composto pelos subconjuntos das p variáveis. Alguns dos critérios mais conhecidos e utilizados são: R^2, R^2 ajustado, Cp de Mallows, AIC (Akaike´s information criterion), entre outros.

(i) R2 e R2 ajustado

Sabemos que R^2 é a razão da soma dos quadrados do modelo com a soma dos quadrados total. Assim, R^2 ajustado é escrito da forma:

$$R^2_{ajust} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

Sendo p os parâmetros e n o número de observações.

(ii) Cp de Mallows

Essa estatística pode ser escrita como:

$$Cp = \frac{SSE}{\hat{\sigma}^2} - (n - 2p)$$

Sendo SSE é a soma dos quadrados dos resíduos, n é o número de observações e p os parâmetros.

(iii) AIC

$$AIC = -2\ log\left(L_p\right) + 2\ [(p+1)+1]$$

Sendo L_p a função de máxima verossimilhança do modelo.

De acordo com Ryan (1997), o AIC é um dos melhores critérios conhecidos e difere do Cp de Mallows quando o σ^2 é conhecido.

(iv) Acurácia

Acurácia é uma medida que traduz a precisão do modelo, permitindo determinar a porcentagem de acerto.

$$Acurácia = \frac{(VP + VN)}{(P + N)}$$

Sendo P = VP + FP e N = VN + FN .

(v) Sensibilidade e Especificidade

A sensibilidade é a probabilidade que o modelo encontrado indica o resultado positivo quando o resultado realmente é positivo, ou seja, a probabilidade de maus classificados corretamente.

Já a especificidade é o complemento da sensibilidade, ou seja, mostrar o resultado negativo (bons) quando de fato é negativo. O valor (1- especificidade) representa os bons classificados como maus.

Os cálculos são apresentados da seguinte forma:

$$Sensibilidade = \frac{VP}{(VP+FN)} \quad \text{e} \quad Especificidade = \frac{VN}{(VN+FP)}$$

Sendo VP, VN, FN, FP indicados na matriz de confusão ilustrado na figura abaixo.

Figura 1: Quadro de Matriz de confusão

		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	VP Verdadeiro Positivo	FP Falso Positivo
	negativos	FN Falso Negativo	VN Verdadeiro Negativo

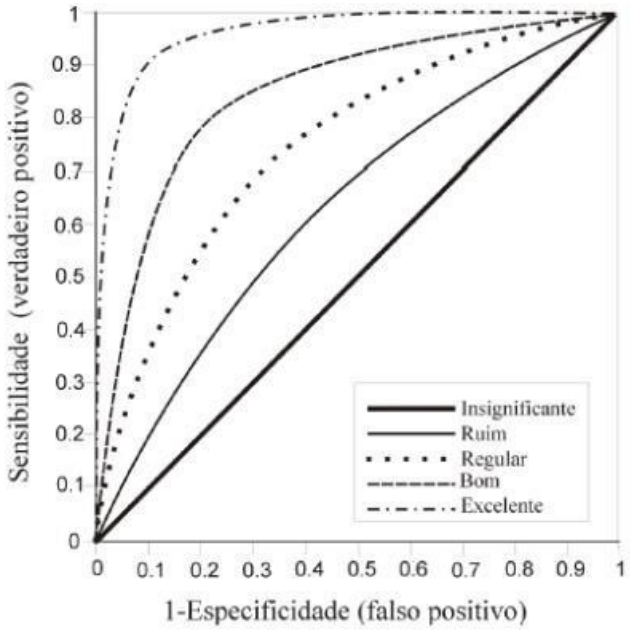
Fonte: autor

(vi) Curva ROC

O gráfico da curva ROC é uma alternativa para avaliação permitindo uma melhor visualização da multidimensionalidade do problema. O gráfico é baseado na probabilidade de verdadeiros positivos (eixo Y) e na probabilidade de falsos positivos (eixo X) que são baseados na tabela matriz de confusão.

Quanto maior a curvatura, melhor o modelo, como mostra a figura 2.

Figura 2: Gráfico da curva ROC



Fonte: Souza (2008)

2.12 Regressão Logística

A regressão logística é uma técnica estatística que se caracteriza por descrever a relação entre as variáveis independentes e uma variável dependente dicotômica, esta representa presença ou ausência de uma característica (HOSMER; LEMESHOW 1989; KLEINBAUM 1996).

O objetivo na análise de regressão logística é descrever o comportamento matemático da variável dicotômica em função dos valores das variáveis independentes, estimando os parâmetros do modelo através do método de estimação de máxima verossimilhança (HOSMER; LEMESHOW, 1989).

De acordo com Hosmer e Lemeshow (1989), a regressão logística é um método padrão para análise de regressão para variáveis dicotômicas fazendo com que a variável tenha uma distribuição binomial e não distribuição normal como no modelo linear clássico.

Segundo o artigo de Senaviratna e Cooray (2019), existem algumas suposições necessárias para satisfazer o resultado:

- Linearidade: as variáveis explicativas devem ter uma relação linear com o logit da variável resposta;
- Independência dos erros: os erros não devem ser correlacionados;
- Multicolinearidade: as variáveis explicativas não devem ser altamente correlacionáveis entre si;
- Sem outliers, valores discrepantes ou pontos altamente influentes.

O artigo faz um estudo sobre influência da multicolinearidade e mostra as consequências da quebra da suposição da não colinearidade nas variáveis explicativas. Com isso, o estudo enfatiza que o modelo pode ter performance melhor ao retirar as variáveis altamente correlacionáveis.

A variável resposta Y a ser considerada possui dois possíveis resultados: sucesso ou fracasso. Considerando P(Y ser “SUCESSO”) = P(Y = 1) = π e P(Y ser “FRACASSO”) = P(Y = 0) = 1 – π, de modo que 0 < π < 1.

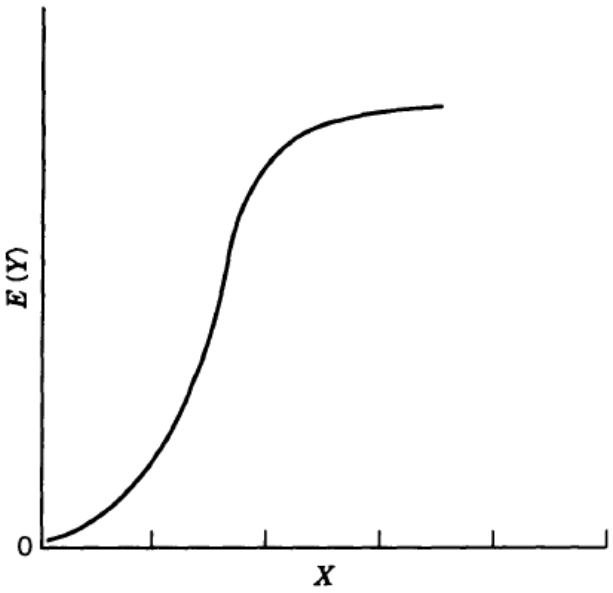
A expressão em termos de preditor linear se dá por:

$$\pi(Y) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

Sendo $\left[\frac{\pi(X)}{1-\pi(X)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ (função de ligação logística) representa a combinação linear dos coeficientes e variáveis explicativas do modelo. O cálculo $\ln\left[\frac{\pi(X)}{1-\pi(X)}\right]$ também pode ser chamado de log da razão de chance (log odds ratio).

Na regressão logística, a interpretação do parâmetro é mais difícil, pois a derivada de π em relação a Y depende de π.

Figura 3: Gráfico da função regressão logística



Fonte: Ryan (1997)

Assim, com a transformação logito, o logit da probabilidade se transforma numa função linear.

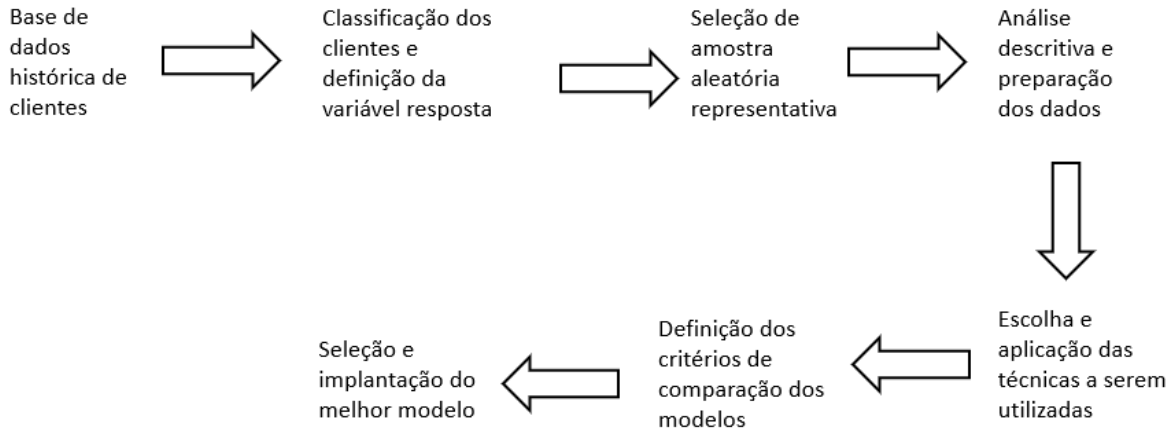
2.12.1 Risco de Crédito

O risco de crédito é a probabilidade de um empréstimo não ser devolvido, ou seja, o risco de uma contraparte em um acordo de concessão de crédito não honrar com o compromisso. Assim, existindo a necessidade de estimar o risco envolvido, exigindo a mensuração desse risco e auxiliando na tomada de decisão, essa mensuração é chamada de credit scoring.

O modelo mensura o risco através de informações cadastrais, além do comportamento anterior do cliente, fazendo com que consigam ser identificados perfis de clientes que sejam interessantes para as instituições.

A figura a seguir representa os passos para a construção de um modelo de credit scoring.

Figura 4: Construção do modelo de credit scoring



Fonte: Baseado em Gonçalvez (2005)

Os escores geralmente são calculados atribuindo pesos às variáveis que caracterizam os clientes, sendo eles pessoas físicas ou jurídicas. Para o desenvolvimento do modelo de credit scoring, foi selecionada uma amostra de clientes que obtiveram crédito anteriormente. Supondo que, no futuro, o comportamento de novos clientes será igual ao dos clientes no passado. (SICSÚ, 2010).

Para avaliar um solicitante, as instituições criaram faixas de escore como, por exemplo, a tabela 4, que mostra a faixa criada pela empresa Serasa Experian, que fornece para as empresas um meio de consulta para análise de crédito. Cada cliente possui um escore onde o corte é diferente para cada tipo de operação, ou seja, um solicitante pode ser considerado mau pagador no caso do cheque especial, mas não necessariamente para um financiamento de imóveis.

Algumas instituições estipulam tempo para considerar um cliente mau pagador ou não, por exemplo, apresentar pelo menos um atraso superior a 90 dias no período de 6 meses.

Tabela 4: Faixa de Escore

Faixas de Score	Probabilidade de Inadimplência Média (%)
901 - 1000	0,25
801 - 900	0,75
701 - 800	1,25
651 - 700	1,75
601 - 650	2,50
551 - 600	3,50
501 - 550	4,50
451 - 500	5,50
401 - 450	7,00
351 - 400	9,00
301 - 350	12,50
201 - 300	22,50
101 - 200	40,00
001 - 100	70,00

Fonte: site Serasa Experian (2018)

(i) Erro de decisão

Assim como em inferência estatística, temos dois tipos de erros que podemos cometer com base na probabilidade:

Erro tipo I: negar o crédito a um cliente que seria um bom pagador.
Erro tipo II: conceder o crédito a um cliente que seria um mau pagador.

Podemos visualizar a matriz de confusão pela Tabela 5.

Tabela 5: Classificação de decisão do crédito

	Decisão	
	Aprovar crédito	Negar crédito
Bom pagador	Decisão correta	Erro tipo I
Mau pagador	Erro tipo II	Decisão correta

Fonte: Sicsú (2010)

(ii) Teste de Kolmogorov – Smirnov

O teste de Kolmogorov-Smirnov (KS) observa a máxima diferença absoluta entre a função de distribuição acumulada e a função de distribuição empírica dos dados. Este teste é bastante utilizado para a análise do crédito (em variáveis contínuas), porém, não devemos usar quando o número de classes de frequências for menor do que 10.

$$D_i = |Fa_i - Fe_i|$$

Em que D = max(D_i) e $D* = 1{,}36(\sqrt{k} + 0{,}12 + \frac{0{,}11}{\sqrt{k}})^{-1}$.

Se D > D*, rejeitamos a igualdade das distribuições ao nível de 5% de significância.

(iii) Teste de Kulback

O teste utilizando a medida de divergência pode ser empregado para comparar as distribuições de variáveis categóricas e, quando aplicada para avaliar a estabilidade populacional, é denominada como Índice de Estabilidade Populacional (IEP ou PSI em inglês). Equivale, também, à medida de valor de informação (vindo do inglês – IV) desenvolvida por Kulback.

Tabela 6: Valores referenciais para IEP

IEP < 0,1	Não houve alteração
0,1 < IEP < 0,25	Possível alteração
IEP > 0,25	Mudanças significantes de perfil

Fonte: Sicsú (2010)

(iv) Teste de Hosmer e Lemeshow

Também conhecido como teste de bondade, testa a qualidade do ajuste, muito utilizado em regressão logística. Ele avalia o modelo através da distância entre as probabilidades ajustadas e observadas.

(v) Aspectos éticos e legais

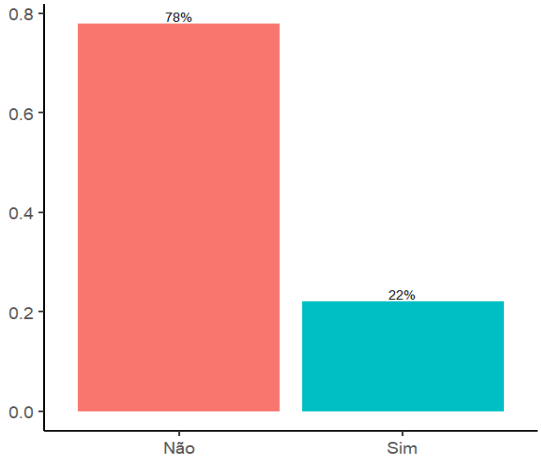
Em alguns países, certas variáveis não são aceitas por motivos éticos, por exemplo, nos Estados Unidos não se pode utilizar variáveis como raça, cor, etnia, religião, sexo, estado civil. Porém, no Brasil ainda não existe uma legislação específica, mas é coerente avaliar se a variável não discrimina um indivíduo.

3 Estatística Descritiva

O banco de dados a ser utilizado neste trabalho, foi apresentado no artigo “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card 35 clientes. Expert Systems with Applications” (Yeh, I. C., & Lien, C. H.)(2009), possuindo 30000 observações com 24 variáveis (sendo uma delas a variável resposta). As observações são de pessoas de Taiwan e foram coletados no ano de 2005. Dentre as variáveis, existe um registro de pagamentos mensais, valor da fatura e estado de reembolso. Essas informações foram avaliadas nos meses de abril a setembro de 2005 e segmentadas por mês.

A porcentagem da variável resposta (Y) ‘default payment next month’ é:

Figura 5: Frequência da variável resposta Y (em %)



Fonte: autor

Sendo “sim” os clientes inadimplentes (sucesso).

Descrição das variáveis:

X1: Valor do crédito concedido (dólar NT)*;

X2: Sexo (1 = Masculino, 2 = Feminino);

X3: Educação (1 = pós-graduação; 2 = universidade; 3 = ensino médio; 4 = outros);

X4: Estado civil (1 = casado; 2 = solteiro; 3 = outros);

X5: Idade (ano);

X6 – X11: Histórico de pagamentos anteriores. Rastreamos os registros de pagamentos mensais anteriores (de abril a setembro de 2005) da seguinte forma:

X6 = o status de pagamento em setembro de 2005;

X7 = status de pagamento em agosto de 2005;

. . .

X11 = status de pagamento em abril de 2005.

A escala de medição para o status de pagamento é:

-1 = pagamento em dia;

1 = atraso no pagamento por um mês;

2 = atraso no pagamento por dois meses;

. . .

8 = atraso no pagamento por oito meses;

9 = atraso no pagamento de nove meses ou mais;

X12 – X17: Valor da fatura (dólar NT).

X12 = valor da fatura em setembro de 2005;

X13 = valor da fatura em agosto de 2005;

. . .

X17 = valor da fatura em abril de 2005;

X18 – X23: Valor do pagamento anterior (dólar NT).

X18 = valor pago em setembro de 2005;

X19 = valor pago em agosto de 2005;

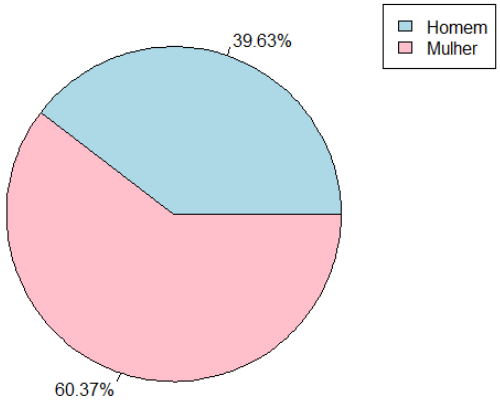
. . .

X23 = valor pago em abril de 2005

* dólar NT: Novo Dólar Taiwanês

Observamos algumas variáveis através da análise descritiva, sendo os resultados o seguinte:

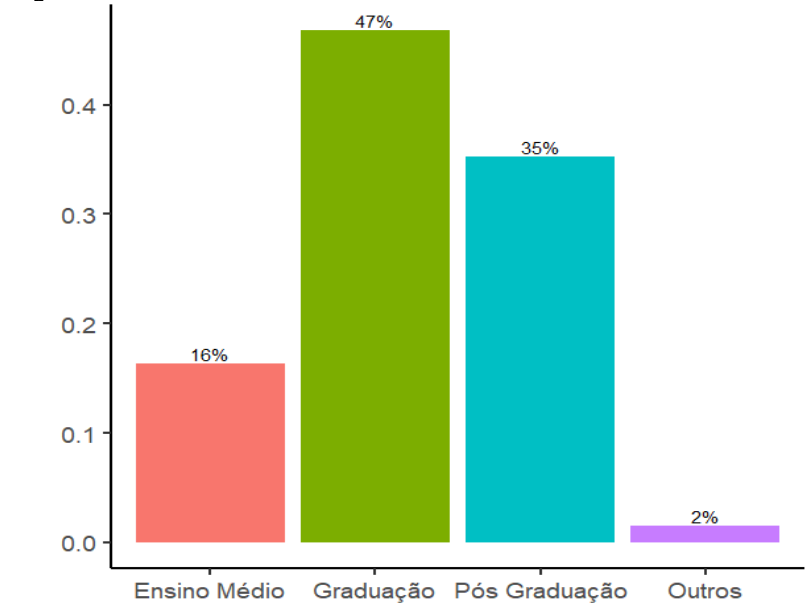
Figura 6: Gráfico de Gênero (em %)



Fonte: autor

De acordo com o gráfico de gênero, a maioria dos observados é do sexo feminino (aproximadamente 60 %).

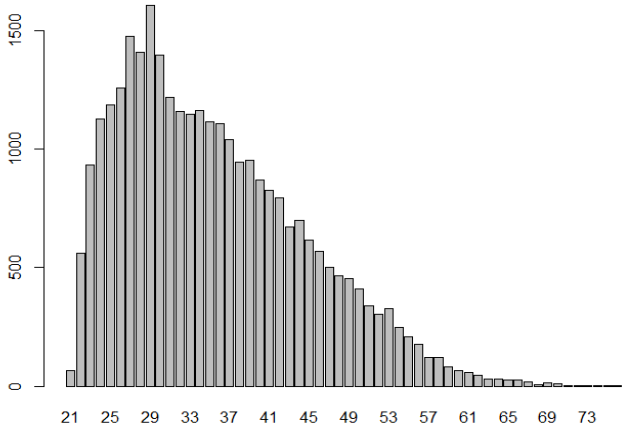
Figura 7: Gráfico do Grau de Escolaridade



Fonte: autor

De acordo com a figura 7, 47% dos observados tem ensino superior completo, 35% possuem pós-graduação, apenas 16% possuem ensino médio e menos de 2% não informaram a escolaridade (classificados como outros).

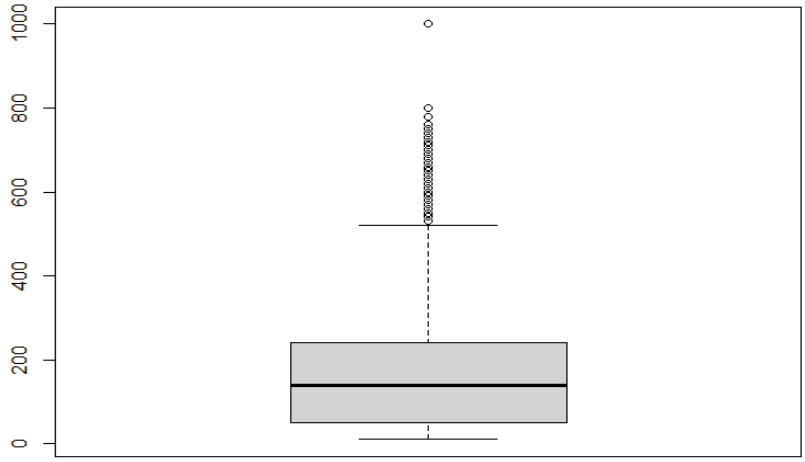
Figura 8: Histograma de Idade



Fonte: autor

Através do histograma, identifica-se que a concentração da faixa etária dos observados está entre os 25 aos 36 anos, mostrando que estes são 50% das observações.

Figura 9: Boxplot do Valor do crédito concedido (em milhares)



Fonte: autor

Pelo boxplot, é possível analisar que o montante do valor de crédito concedido gira em torno de 140 mil TWD (código da moeda de Taiwan). O 1° quartil = 50 mil, a Média = 167 mil e o 3° quartil = 240 mil.

Tabela 7: Distribuição de Inadimplentes x Gênero

Sexo	Bom (%)	Mau (%)	**Total**
Masculino	9015 (38,59)	2873 (43,29)	**11888**
Feminino	14349 (61,41)	3763 (56,71)	**18112**
Total	**23364**	**6636**	**30000**

Fonte: autor

Da tabela acima, pode-se ver que 57% das mulheres (aproximadamente) são inadimplentes contra 43% dos homens.

Tabela 8: Distribuição de Inadimplentes x Estado Civil

Estado Civil	Bom (%)	Mau (%)	**Total**
Casado	10453 (44,73)	3206 (48,31)	**13659**
Solteiro	12623 (54,02)	3341 (50,34)	**15964**
Outros	288 (0,01)	89 (0,01)	**377**
Total	**23364**	**6636**	**30000**

Fonte: autor

Já entre os inadimplentes casados e solteiros não existe muita diferença em porcentagem.

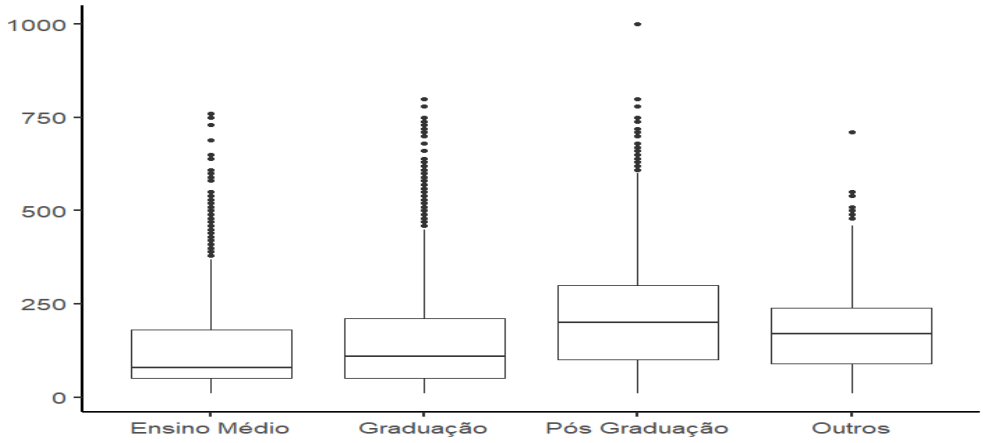
Tabela 9: Distribuição de Inadimplentes x Escolaridade

Escolaridade	Bom (%)	Mau (%)	**Total**
Pós-Graduação	8549 (36,59)	2036 (30,68)	**10585**
Universidade	10700 (45,79)	3330 (50,18)	**14030**
Ensino Médio	3680 (15,75)	1237 (18,64)	**4917**
Outros	435 (0,01)	33 (0,00)	**468**
Total	**23364**	**6636**	**30000**

Fonte: autor

Da tabela 9, verifica-se que a metade dos inadimplentes possui nível superior.

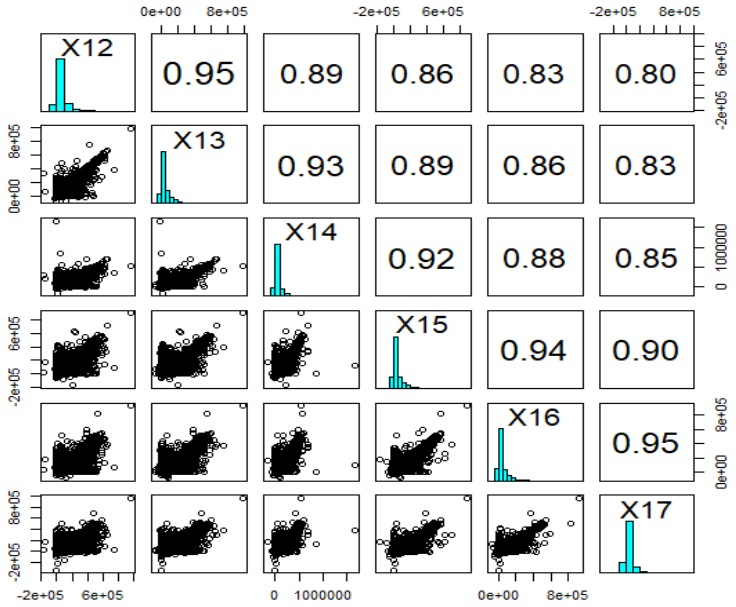
Figura 10: Gráfico de Grau de Escolaridade x Valor do crédito concedido (em milhares)



Fonte: autor

Da figura 10, percebe-se que a mediana de crédito cedido é maior para quem tem pós-graduação e outros.

Figura 11: Gráfico de correlação entre as variáveis Valor da fatura (Abril a Setembro 2015)



Fonte: autor

Pela figura 11, vê-se que existe uma grande correlação entre as variáveis analisadas. Percebendo, portanto, que as variáveis Valor da fatura dos meses de abril a setembro possuem alta correlação entre si.

4 Resultados e Conclusões

Analisando os dados, percebe-se a existência de observações sem informações nas variáveis Valor da fatura e Valores pagos. Com isso, essas observações foram retiradas do conjunto de dados para o ajuste do modelo (cerca de 2,65% da base).

A seguir, é feita a construção de dois modelos de regressão logística com algumas modificações nas variáveis independentes afim de comparar e avaliar qual o melhor modelo ajustado.

- Modelo 1

Para o primeiro modelo, alterou-se o tipo das variáveis (sexo, escolaridade, estado civil e status do pagamento dos meses de abril a setembro), que antes estavam como variáveis numéricas, para variáveis categóricas.

Inicialmente, foram selecionadas as variáveis a partir do método de seleção de modelos via Stepwise e calculado o Fator de Inflação da Variância, do inglês Variance Inflation Factor (VIF), que representa o incremento da variância devido à presença de multicolinearidade (MONTGOMERY; PECK; VINING, 2006).

Pelo método stepwise, o modelo ajustado teve as seguintes variáveis:

Y ~ Valor Concedido + Gênero + Escolaridade + Estado Civil + Idade +
+ Status pagamento Set + Status pagamento Ago + Status pagamento Jul +
+Status pagamento Jun + Status pagamento Mai + Status pagamento Abr +
+ Valor fatura Jul+ Valor pago Set + Valor pago Ago + Valor pago Mai +
+ Valor pago Abr.

Assim, foi calculado o VIF dessas variáveis e obtido o seguinte resultado, considerado satisfatório de acordo com Salvian (2016), que explica que o VIF máximo acima de 10 indica que a multicolinearidade pode estar influenciando as estimativas de mínimos quadrados.

Tabela 10: Fator de Inflação da Variância (VIF)

Variável	VIF
Valor concedido	1,28
Gênero	1,01
Escolaridade	1,03
Estado Civil	1,02
Status pagamento Set2015	2,56
Status pagamento Ago2015	2,89
Status pagamento Jul2015	4,19
Status pagamento Jun2015	4,73
Status pagamento Mai2015	3,09
Status pagamento Abr2015	1,57
Valor da fatura Jul2015	1,40
Valor pagamento anterior Set2015	1,10
Valor pagamento anterior Ago2015	1,11
Valor pagamento anterior Mai2015	1,06
Valor pagamento anterior Abr2015	1,05

Fonte: autor

Após analisado o p-valor da estimativa do parâmetro, percebe-se que as variáveis Valor pago Mai (p-valor = 0,07135) e Valor pago Abr (p-valor = 0,08068) que entraram no modelo, não eram significativas (nível de significância de 5%), portanto, não é considerado informativo.

Algumas das variáveis consideradas no modelo foram o valor do montante solicitado, sexo, nível de escolaridade, estado civil. Segue abaixo, as variáveis que entraram no modelo.

Y ~ - 1,237*Intercepto - 0,00000239*Valor Concedido - 0,01448* Gênero (Feminino) -
- 1,017* Escolaridade (Outros) - 0,1935* Estado Civil (Solteiro) +
+ 0,3685*Status pagamento Set (-1) - 0,3773* Status pagamento Set (0) +
+ 0,6799*Status pagamento Set (1) + 1,901*Status pagamento Set (2) +
+ 1,933*Status pagamento Set (3) + 1,57*Status pagamento Set (4) +
+ 1,199*Status pagamento Set (5) + 0,2519*Status pagamento Ago (2) +
+ 0,4466*Status pagamento Jul (2) + 0,477*Status pagamento Jul (3) +
+ 0,3064*Status pagamento Mai (2) - 0,3036* Status pagamento Abr (0) +
+ 0,6347*Status pagamento Abr (3) + 0,00000254*Valor da fatura Jul –
- 0,00001077* Valor pago Set - 0,00000938* Valor pago Ago.

Tabela 11: Estimativa dos parâmetros, estimativa do erro padrão, estatística z e p-valor (Modelo 1)

	Estimativa	Erro Padrão	Valor z	P-Valor
Intercepto	-1,2370	0,08837	-14,0030	< 2e-16
Valor concedido	-0,00000239	0,0000002	-13,3430	< 2e-16
Gênero (Feminino)	-0,01448	0,03287	-4,4060	0,000010500
Escolaridade (Outros)	-1,0170	0,19370	-5,2520	0,000000151
Estado Civil (Solteiro)	-0,1935	0,03364	-5,8110	0,000000006
Status pagamento Set (-1)	0,3685	0,12610	2,9220	0,003474000
Status pagamento Set (0)	-0,3773	0,13380	-2,8190	0,004810000
Status pagamento Set (1)	0,6799	0,10950	6,2080	0,000000001
Status pagamento Set (2)	1,9010	0,12530	15,1750	< 2e-16
Status pagamento Set (3)	1,9330	0,18240	10,5980	< 2e-16
Status pagamento Set (4)	1,5700	0,30500	5,1470	0,000000265
Status pagamento Set (5)	1,1990	0,48570	2,4690	0,013562000
Status pagamento Ago (2)	0,2519	0,12740	1,9770	0,047995000
Status pagamento Jul (2)	0,4466	0,12920	3,4560	0,000549000
Status pagamento Jul (3)	0,4770	0,22300	2,1390	0,032427000
Status pagamento Mai (2)	0,3064	0,13280	2,3080	0,021018000
Status pagamento Abr (0)	-0,3036	0,08856	-3,4280	0,000608000
Status pagamento Abr (3)	0,6347	0,23520	2,6990	0,006951000
Valor da fatura Jul	0,00000254	0,00000034	7,4540	0,000000000
Valor pago Set	-0,00001077	0,00000210	-5,1190	0,000000307
Valor pago Ago	-0,00000938	0,00000193	-4,8720	0,000001110

Fonte: autor

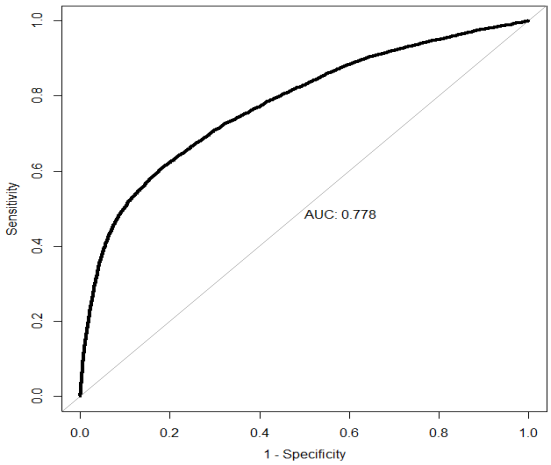
Para avaliar se o modelo está bem ajustado, foram analisados os critérios de avaliação como a curva ROC, a acurácia, a sensibilidade e especificidade e ponto de corte de 0,5.

Acurácia = 0,8266

Sensibilidade = 0,9523

Especificidade = 0,3730

Figura 12: Curva ROC (Modelo 1)



Fonte: autor

Através desses resultados, a conclusão é que o modelo ajustado foi satisfatório, dado que todos os critérios de avaliação tiveram bons resultados. Portanto, o modelo linear generalizado mostrou-se adequado para o ajuste de dados de risco de crédito.

- Modelo 2

Para o segundo modelo, foram criadas três variáveis que analisam se foram pagas as dívidas em dia (0) ou com atraso (1), se existiam algum valor da fatura (1) ou não (0), e se foram pagos (0) ou não (1), e excluídas as variáveis que compuseram essas novas. Assim, o modelo ajustado tem as seguintes variáveis:

Y ~ -2,240*Intercepto - 0,00000267*Valor Concedido - 0,1493*Gênero (Feminino) –
- 1,104*Escolaridade (Outros) - 0,2167* Estado Civil (Solteiro) +
+ 0,3693*Status pagamento + 0,6734*Valor fatura + 1,188*Valor pago.

Tabela 12: Estimativa dos parâmetros, estimativa do erro padrão, estatística z e p-valor (Modelo 2)

	Estimativa	Erro Padrão	Valor z	P-Valor
Intercepto	-2,240	0,1005	-22,283	< 2e-16
Valor concedido	-0,00000267	1,435e-07	-18,666	< 2e-16
Gênero (Feminino)	-0,1493	0,03045	-4,903	9,44e-07
Escolaridade (Outros)	-1,104	0,1854	-5,953	2,64e-09
Estado Civil (Solteiro)	-0,2167	0,03439	-6,3	2,98e-10
Status pagamento	0,3693	0,0345	10,706	< 2e-16
Valor da fatura	0,6734	0,04336	15,530	< 2e-16
Valor pago	1,188	0,03315	35,833	< 2e-16

Fonte: autor

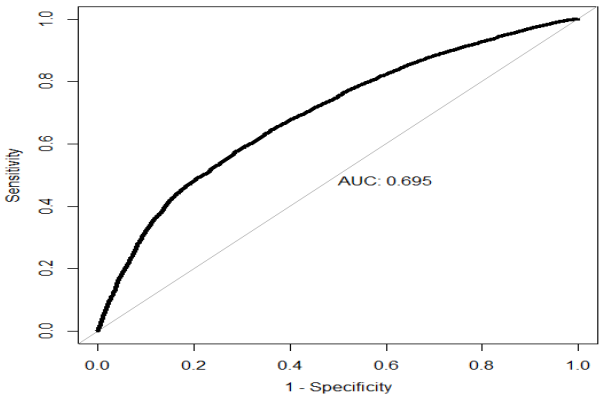
E os resultados dos critérios de avaliação foram:

Acurácia = 0,7842
Sensibilidade = 0,9922

Especificidade = 0,0336

Com o mesmo ponto de corte de 0,5.

Figura 13: Curva ROC (Modelo 2)



Fonte: autor

Comparando os dois modelos ajustados, podemos ver que eles se assemelham tanto nas variáveis (Exemplo: em ambas a variável Idade não entrou no modelo) bem como nos resultados de critério de avaliação, sendo que o primeiro modelo teve melhor resultado na acuracidade e na curva ROC, mas pior na sensibilidade e especificidade. Assim, podemos comprovar que os modelos ajustados são eficientes para discriminar o bom do mau pagador. Caso houvesse necessidade de escolha, o segundo modelo seria escolhido por conter menos variáveis, demandando menos esforço computacional.

5 Propostas Futuras

Para um próximo estudo dos dados analisados, sugere-se a criação do gráfico de correlação não paramétrica das variáveis Status de pagamentos, gênero, estado civil e escolaridade. Também é sugerido uma análise dos registro -2 e 0 das variáveis Status de pagamentos. Possivelmente, a criação do modelo separando aleatoriamente os dados em duas amostras, sendo uma de teste e outra de treino do modelo.

Referências

BARTH, N.L. **Inadimplência: Construção de modelos de previsão**. São Paulo, SP: Nobel, 2004. 98 p.

COLADELLO, L. F. **Modelos Lineares Generalizados: Aplicação na avaliação do risco de hipertensão arterial sistêmica em mulheres idosas da cidade de Presidente**. 61f. Faculdade de Tecnologia e Ciências, Universidade Estadual Júlio de Mesquita, SP, 2011.

CORDEIRO, G. M. **Modelos lineares generalizados. Minicurso para o VII Simpósio Nacional de Probabilidade e Estatística**. Campinas, SP: 1986. 286 p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados: Minicurso para o 12*o* SEAGRO e a 52*ª* Reunião Anual da RBRAS UFSM**. Santa Maria, RS: 2007. 158 p.

DINIZ, C; LOUZADA, F. **Modelagem Estatística para Risco de Crédito**. João Pessoa, PB: 20° SINAPSE, 2012. 178 p.

GONÇALVES, E. B. **Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos.** São Paulo; 2005.

HAUCK JR., W. W.; DONNER, A. **Wald's Test as Applied to Hypotheses in Logit Analysis**. Journal of the American Statistical Association, 72:360a, 851-853, 1977.DOI: 10.1080/01621459.1977.10479969

HOSMER, D. W.; LEMESHOW, D. W. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

JORION, P. **Valueatrisk – Nova fonte de referência para Gestão de Risco Financeiro.** São Paulo: BM&F, 2004, 487 p.

LINDSEY, J. K. **Applying generalized linear models**. New York: Springer-Verlag, 1997.

MARQUES, M. A. P. **Análise e comparação de alguns métodos alternativos de seleção de variáveis preditoras no modelo de regressão linear.** São Paulo: 2018.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introdiction to Linear Regression Analysis.** USA: John Wiley & Sons, 2012.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics.** USA:1913.

PAULA, G. A. **Modelos de regressão com apoio computacional.** São Paulo: 2013.

RYAN, T. P. **Modern Regression Methods.** New York: John Wiley & Sons, 1997.

SALVIAN, M. **Multicolinearidade**. Piracicaba, SP; 2016.

SERAVINATNA, N. A. M. R.; COORAY, T. M. J. A. **Diagnosing Multicollinearity of Logistic Regression Model.** Asian Journal of Probability and Statistics, 2019. 9p. DOI: 10.9734/ajpas/2019/v5i230132

Serasa Experian: https://www.serasaexperian.com.br/ftp2/riskscoring.pdf (01/07/2018)

SICSÚ, A.L. **Credit Scoring: Desenvolvimento. Implantação. Acompanhamento**. São Paulo, SP: Blucher, 2010. 180 p.

SOUZA, P. R. S. **Uma Análise em possíveis Casos de Patologias Médicas utilizando a Curva ROC em Lógica Paraconsistente Anotada para Apoio a Decisão Médica em busca de melhor Precisão de Resposta na Web.** Santos, SP: 2008.

TURKMAN, A. M. M; SILVA, G. L. **Modelos lineares Generalizados – da teoria à prática**. Lisboa: 2000.

UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (25/09/2021)

Yeh, I-C.; Lien, C.-h. **The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.** 2009 https://bradzzz.gitbooks.io/ga-dsi-seattle/content/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/DefaultCreditCardClients_yeh_2009.pdf

Anexo

Variáveis:

X1 : LIMIT_BAL	X2:SEX	X3: EDUCATION	X4: MARRIAGE	X5: AGE
X6: PAY_0	X7: PAY_2	X8: PAY_3	X9: PAY_4	X10: PAY_5
X11: PAY_6	X12: BILL_AMT1	X13: BILL_AMT2	X14: BILL_AMT3	X15: BILL_AMT4
X16: BILL_AMT5	X17: BILL_AMT6	X18: PAY_AMT1	X19: PAY_AMT2	X20: PAY_AMT3
X21: PAY_AMT4	X22: PAY_AMT5	X23: PAY_AMT6	Y: default payment next month	

Dados:

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1
24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1
25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0
26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0
27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1
28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0
29	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0
30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	0	0
31	230000	2	1	2	27	-1	-1	-1	-1	-1	-1	16646	17265	13266	15339	14307	36923	17270	13281	15339	14307	37292	0	0
32	50000	1	2	2	33	2	0	0	0	0	0	30518	29618	22102	22734	23217	23680	1718	1500	1000	1000	1000	716	1
33	100000	1	1	2	32	0	0	0	0	0	0	93036	84071	82880	80958	78703	75589	3023	3511	3302	3204	3200	2504	0
34	500000	2	2	1	54	-2	-2	-2	-2	-2	-2	10929	4152	2272	7521	71439	8981	4152	22827	7521	71439	981	51582	0
35	500000	1	1	1	58	-2	-2	-2	-2	-2	-2	13709	5006	31130	3180	0	5293	131	396	396	565	792	0	0
36	160000	1	1	2	30	-1	-1	-2	-2	-2	-1	30265	-131	-527	-923	-1488	-1884	8026	8060	6300	6400	6400	6737	0
37	280000	1	2	1	40	0	0	0	0	0	0	186503	181328	180422	170410	173901	177413	8026	8060	6300	6400	6400	6737	0
38	60000	2	2	2	22	0	0	0	0	0	-1	15054	9806	11068	6026	-28335	18660	1500	1518	2043	0	47671	617	0
39	50000	1	1	2	25	1	-1	-1	-2	-2	-2	0	780	0	0	0	0	780	0	0	0	0	0	1
40	280000	1	1	2	31	-1	-1	2	-1	0	-1	498	9075	4641	9976	17976	9477	9075	0	9976	8000	9525	781	0
41	360000	1	1	2	33	0	0	0	0	0	0	218668	221296	206895	628699	195969	179224	10000	7000	6000	188840	28000	4000	0
42	70000	2	1	2	25	0	0	0	0	0	0	67521	66999	63949	63699	64718	65970	3000	4500	4042	2500	2800	2500	0
43	10000	1	2	2	22	0	0	0	0	0	0	1877	3184	6003	3576	3670	4451	1500	2927	1000	300	100	500	0
44	140000	2	2	1	37	0	0	0	0	0	0	59504	61544	62925	64280	67079	69802	3000	3000	3000	4000	4000	3000	0
45	40000	2	1	2	30	0	0	0	2	0	0	18927	21295	25921	25209	26636	2919	3000	5000	0	2000	3000	0	0
46	210000	1	1	2	29	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	1
47	20000	2	1	2	22	0	0	2	-1	0	0	14028	16484	15800	16341	16675	0	3000	0	16741	334	0	0	1
48	150000	2	5	2	46	0	0	-1	0	0	-2	4463	3034	1170	1170	0	0	1013	1170	0	0	0	0	1
49	380000	1	2	2	32	-1	-1	-1	-1	-1	-1	22401	21540	15134	32018	11849	11873	21540	15138	24677	11851	11875	8251	0
50	20000	1	1	2	24	0	0	0	0	0	0	17447	18479	19476	19865	20480	20063	1318	1315	704	928	912	1069	0
51	70000	1	3	2	42	2	2	2	2	2	0	36171	36171	38355	39423	38659	39362	0	3100	2000	0	1500	1500	1
52	100000	2	3	3	43	0	0	0	0	0	0	51163	51163	43824	39619	35762	33258	2000	1606	1500	2000	1500	1000	0
53	310000	2	2	1	49	-2	-2	-2	-2	-2	-2	13465	7867	7600	11185	3544	464	7875	7600	11185	3544	464	0	0
54	180000	2	1	2	25	2	2	0	0	0	0	41402	41742	42758	43510	44420	45319	1300	2010	1762	1762	1790	1622	0
55	150000	2	1	2	29	0	0	0	0	0	0	46224	34993	31434	26518	21042	16540	1600	1718	1049	1500	2000	5000	0
56	500000	2	1	1	45	-2	-2	-2	-2	-2	-2	1905	3640	162	0	151	2530	3640	162	0	151	2530	0	0
57	180000	2	3	1	34	0	0	0	-1	-1	-1	16386	15793	8441	7142	-679	8321	8500	1500	7500	679	9000	2000	0
58	180000	2	2	1	34	0	0	0	0	0	0	175886	173440	172308	168608	132202	129918	8083	7296	5253	4814	4816	3800	0
59	200000	2	1	2	34	-1	3	2	2	2	2	1587	1098	782	1166	700	1414	0	0	700	0	1200	0	0
60	400000	2	2	1	29	0	0	0	0	0	0	400134	398857	404205	360199	356656	364089	17000	15029	30000	12000	12000	23000	0
61	500000	2	3	1	28	0	0	0	0	0	0	22848	23638	18878	14937	13827	15571	1516	1300	1000	1000	2000	2000	1
62	70000	1	2	1	29	0	0	0	0	0	-1	70800	72060	69938	16518	14096	830	4025	2095	1000	2000	3000	0	0
63	50000	1	1	2	46	2	2	2	2	2	2	24987	24300	26591	25865	27667	28264	0	2700	0	2225	1200	0	1
...																								
29999	80000	1	3	1	41	1	-1	0	0	0	-1	-1645	78379	76304	52774	11855	48944	85900	3409	1178	1926	52964	1804	1
30000	50000	1	2	1	46	0	0	0	0	0	0	47929	48905	49764	36535	32428	15313	2078	1800	1430	1000	1000	1000	1

Apêndice

```
library (dplyr)
library(pROC)
library(caret)
library(car)

dados <-read.csv(file = choose.files() ,sep= ';',fill = T, header = TRUE)
str(dados)
dados$linhas<- dados$ï..X
dados$X3 <- ifelse(dados$X3 ==1,1,
                           ifelse(dados$X3 ==2,2,
                                  ifelse(dados$X3 ==3,3, 4)))
dados$X4 <- ifelse( dados$X4 == 0, 3, dados$X4)
dados1<- dados[-c(2,3,4,5,6,7,8,9,10,11,12,25)]

c1<- apply(ifelse(dados1 == 0,1,0),1,sum)
vet<- NULL
for(i in 1:length(c1)){vet[i]<-ifelse(c1[i] >11 ,0,i)}
linhas <- vet[vet>0]
linhas <- data.frame(linhas)
colnames(linhas) <- c("linhas")

dados2 <- merge(dados, linhas, all.y = TRUE)

cols<- c("X2","X3","X4","X6","X7","X8","X9","X10","X11")

dados3 <- dados2 %>% mutate_at(cols, factor)
```

- Modelo 1

```
ab_results <- glm(Y ~., family = "binomial",
                          data = dados3)
ab_results.step <- step(ab_results, direction = "both")
summary(ab_results.step)

pred <- predict(ab_results,type = "response")
pred1 <- rep(0,length(dados3$Y))
pred1[pred >.5] = 1
matriz <- confusionMatrix(table(pred1 ,dados3$Y))

vif1<- glm(Y ~ X1 + X2 + X3 + X4 + X6 + X7 + X8 + X9 + X10 + X11 +
   X14 + X18 + X19 + X22 + X23,family = "binomial", data = dados3)
vif(vif1)

summary(vif1)
pred2 <- predict(vif1,type = "response")
pred3<- rep(0,length(dados3$Y))
```

```
pred3[pred2 >.5] = 1
matriz2 <- confusionMatrix(table(pred3 ,dados3$Y))

lrROC <- roc (dados3$Y ~ pred2, plot = TRUE,
print.auc = TRUE, col = "black", lwd = 4,
legacy.axes = TRUE, main = "Curva Roc")
```

- Modelo 2

```
dados3$X24<- ifelse(dados3$X6 == -1 | dados3$X7 == -1 | dados3$X8 == -1 |
dados3$X9 == -1 |
                    dados3$X10 == -1 | dados3$X11 == -1,0,1)

dados3$X25<- ifelse(dados3$X12 == 0 | dados3$X13 == 0 |dados3$X14 == 0
|dados3$X15 == 0 |
                    dados3$X16 == 0 |dados3$X17 == 0,0,1)

dados3$X26<- ifelse(dados3$X18 == 0 | dados3$X19 == 0 |dados3$X20 == 0
|dados3$X21 == 0 |
                    dados3$X22 == 0 |dados3$X23 == 0,1,0)

dados4<- dados3[-c(1,2,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)]

ab_results <- glm(Y ~., family = "binomial",
                  data = dados4)
ab_results.step <- step(ab_results, direction = "both")
summary(ab_results.step)

pred <- predict(ab_results,type = "response")
pred1 <- rep(0,length(dados4$Y))
pred1[pred >.5] = 1
matriz <- confusionMatrix(table(pred1 ,dados4$Y))

lrROC <- roc (dados4$Y ~ pred, plot = TRUE,
print.auc = TRUE, col = "black", lwd = 4,
legacy.axes = TRUE)
```