UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos Lomax assimétricos: uma nova abordagem para a classificação de dados binários desbalanceados

Letícia Ferreira Murça Reis

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Letícia Ferreira Murça Reis

Modelos Lomax assimétricos: uma nova abordagem para a classificação de dados binários desbalanceados

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Julho de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R375m

Reis, Leticia Ferreira Murca
Modelos Lomax assimétricos: uma nova abordagem para a classificação de dados binários desbalanceados / Leticia Ferreira Murca Reis; orientador Francisco Louzada Neto. -- São Carlos, 2023.
90 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. dados desbalanceados. 2. distribuição Lomax. 3. links assimétricos. 4. estimação bayesiana. 5. regressão binária. I. Neto, Francisco Louzada, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Letícia Ferreira Murça Reis

Asymmetric Lomax models: a new approach to imbalanced binary data classification

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
July 2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Letícia Ferreira Murça Reis, realizada em 17/05/2023.

Comissão Julgadora:

Prof. Dr. Francisco Louzada Neto (USP)

Prof. Dr. Diego Carvalho do Nascimento (UDA)

Prof. Dr. Paulo Henrique Ferreira da Silva (UFBA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Este trabalho é dedicado a todos aqueles que permanecem lutando pela ciência em nosso país.

AGRADECIMENTOS

Agradeço a meu marido, Rodrigo Fernando, por todo o apoio e compreensão que tive durante todo o mestrado.

Agradeço a meus pais, Elaine e Saulo, por todo o cuidado e por todos os ensinamentos que me transmitiram.

Agradeço aos Professores, Francisco Louzada e Paulo Henrique, por me orientarem no trabalho e tornarem possível essa realização. E também a Diego, que como banca, ofereceu valiosas dicas de aprimoramento.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Dear Hilde, if the human brain was simple enough for us to understand, we would still be so stupid that we couldn’t understand it.

Love, Dad.”

(Jostein Gaarder)

RESUMO

REIS, L. F. **Modelos Lomax assimétricos: uma nova abordagem para a classificação de dados binários desbalanceados**. 2023. 90 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

A expressão "dados binários desbalanceados" refere-se a um conjunto de dados em que uma das classes apresenta significativamente menos observações do que a outra. Isso prejudica a performance tanto de algoritmos de aprendizado de máquina como de modelos estatísticos, visto que a maioria dessas ferramentas supõe que os dados apresentam a mesma proporção de observações nas duas categorias. Para lidar com esse desafio, vários autores sugerem o uso de funções de ligação assimétricas na regressão binária, em detrimento das conhecidas funções de ligação simétricas: *logit* e *probit*. Assim, é possível não só melhorar a performance preditiva do modelo, como também reduzir o viés na estimação de parâmetros e de probabilidades. Essa é uma solução que gera modelos probabilísticos, os quais se destacam na tomada de decisão em comparação com aqueles que simplesmente atribuem uma única classe, sem levar em consideração a probabilidade associada a ela. Portanto, o objetivo deste trabalho é introduzir novas funções de ligação assimétricas que são geradas por meio de transformações da distribuição Lomax. Essas funções incluem as distribuições Double Lomax (DLomax), Potência Double Lomax (PDLomax) e Reversa de Potência Double Lomax (RPDLomax). As funções propostas possuem assimetria comprovada e podem ser facilmente implementadas em *softwares* estatísticos. Além disso, o estudo de simulações aponta que as funções de ligação propostas neste trabalho podem performar melhor que o *link* logístico em diversos cenários de desbalanceamento. O uso dessas funções também se mostrou promissor na modelagem de dados reais, visto que neste trabalho obteve melhores métricas que as funções de ligação clássicas em duas aplicações.

Palavras-chave: dados desbalanceados, distribuição Lomax, estimação Bayesiana, *links* assimétricos, regressão binária.

ABSTRACT

REIS, L. F. **Asymmetric Lomax models: a new approach to imbalanced binary data classification**. 2023. 90 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Imbalanced data refers to a dataset where one class has significantly fewer observations than the other class. This can lead to poor performance of both machine learning algorithms and statistical models, since most of these tools assume that the data has the same proportion of observations in both categories. To deal with this challenge, several authors suggest the use of asymmetric link functions in binary regression, instead of the well-known symmetric link functions: logit and probit. Thus, it is possible not only to improve the predictive performance of the model, but also to reduce the bias in the estimation of parameters and probabilities. This is a solution that generates probabilistic models, which excel in decision-making compared to those that simply assign a single class without considering the associated probability. Therefore, this work aims to present new asymmetric link functions generated from the transformations of the Lomax distribution. These functions include the Double Lomax (DLomax), Power Double Lomax (PDLomax), and Reverse Power Double Lomax (RPDLomax) distributions. The proposed functions have proven asymmetry and can be easily implemented in statistical softwares. In addition, the simulation study indicates that these functions can perform better than logistic regression in various imbalanced classification scenarios. They also proved to be promising in modeling real-world datasets, as in this work we obtained better results than classic link functions in two applications.

Keywords: imbalanced data, Lomax distribution, Bayesian estimation, asymmetric links, binary regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Todos os dias realizamos diversas tarefas de classificação sem percebermos. Classificamos nossas roupas e as alocamos nas gavetas adequadas, classificamos quais mensagens e *e-mails* são importantes, classificamos as louças em sujas e limpas, classificamos tarefas em níveis de dificuldade etc. Essas tarefas são simples e uma classificação falha nelas não leva a consequências críticas. Por outro lado, caso um médico falhe em diagnosticar um câncer, seu paciente pode morrer em um curto período de tempo. Ou ainda, se um banco falhar muitas vezes em classificar clientes em bons e maus pagadores, pode ter um prejuízo de bilhões.

Devido à importância da classificação, modelos estatísticos foram desenvolvidos para auxiliar nesse processo. A partir desses modelos é possível determinar qual a classe mais provável de determinado elemento com base em informações que possuímos sobre ele. Por exemplo, a partir de dados sobre atrasos em pagamentos de conta e histórico de dívidas não pagas, é possível determinar a probabilidade de determinada pessoa ser boa ou má pagadora. Também é possível estimar qual a categoria mais provável de determinado produto a partir de sua descrição.

Como podemos ver, esses modelos estatísticos podem tanto classificar um elemento em duas categorias, como em várias. No entanto, neste trabalho focaremos apenas nos modelos que classificam em duas categorias: os modelos de classificação binária. Assim como foi apresentado no parágrafo anterior, esses modelos nada mais são do que ferramentas para determinar a probabilidade de que um elemento pertença a determinada categoria de interesse. Se ele pertence à categoria de interesse, chamamos esse evento de sucesso; e caso não pertença, chamamos de fracasso.

Entretanto, os principais modelos assumem que o número de sucessos e fracassos é aproximadamente igual (HAIBO; GARCIA, 2009). Tal suposição não é atendida em uma série de casos, visto que há um número incontável de bancos de dados desbalanceados, ou seja, bancos de dados que não apresentam a mesma proporção entre sucessos e fracassos. Esse desbalanceamento pode ocorrer por uma série de motivos, podendo ser oriundo da natureza dos

dados, como por exemplo, dados de mamografias, em que se espera que exista um número muito maior de mulheres saudáveis do que portadoras de câncer de mama. Ou pode ser oriundo de fatores externos, como tempo e armazenamento computacional (HAIBO; GARCIA, 2009).

O não cumprimento dessa suposição compromete a performance dos modelos de classificação binária (HAIBO; GARCIA, 2009), de modo que, ao utilizar esses modelos para a classificação de dados desbalanceados, é esperado que estes apresentem menor acurácia. Em geral, a classe mais prejudicada é a classe minoritária, sendo esta, na maioria das vezes, a classe de maior interesse.

Dada essa situação, neste trabalho vamos estudar uma forma de lidar com esse problema utilizando-se da regressão binária, a qual é uma das principais ferramentas para classificação de dados binários. Nesse tipo de modelo, a probabilidade de que um elemento pertença a certa categoria é determinada por uma função $F(.)$ das covariáveis oferecidas pelo banco de dados, sendo $F(.)$ uma função monotônica crescente (também conhecida como função de ligação) que leva valores contidos no intervalo $(-\infty, +\infty)$ ao intervalo $[0,1]$.

Devido a essas características, é comum que $F(.)$ seja uma função de distribuição acumulada (fda) já conhecida, como por exemplo, a função *logit*, que é a fda da distribuição logística; a função *probit*, que é a fda da distribuição normal; e a função *cauchit*, que é a fda da distribuição Cauchy. Em especial, é preferida a função *logit*, visto que esta oferece maior interpretabilidade para os parâmetros do modelo (PRASETYO *et al.*, 2019). Em geral, são utilizadas funções simétricas, como as que foram citadas acima.

No entanto, quando se trata de dados desbalanceados, diversos autores sugerem que distribuições assimétricas são mais apropriadas do que distribuições simétricas (ALVES; BAZÁN; ARELLANO-VALLE, 2022). Segundo Czado e Santner (1992), o tipo mais grave de especificação incorreta de $F(.)$ ocorre quando se escolhe uma função simétrica para dados que seguem uma função assimétrica de suas covariáveis. Nesses casos há viés substancial na estimação de parâmetros e também na estimação da probabilidade de sucesso.

Sendo assim, alguns autores tentaram utilizar distribuições acumuladas assimétricas como função de ligação $F(.)$. Pode-se ver isso no trabalho de Yin *et al.* (2020), em que foi avaliado o ajuste das funções de ligação generalizada de valores extremos (GEV) padrão, skewed Weibull e Fréchet, na predição da mortalidade em seguros de vida. Também, Naranjo *et al.* (2018) utilizaram a distribuição t de Student assimétrica para detectar pacientes com doença de Parkinson; e Calabrese e Osmetti (2013) utilizaram as funções de ligação Fréchet, Weibull e Gumbel para modelar falências de pequenas e médias empresas. Contudo, esses modelos não permitem o controle da assimetria por um parâmetro extra, de modo que é impossível estabelecer relações com os modelos simétricos (ALVES; BAZÁN; ARELLANO-VALLE, 2022).

Outros autores tentaram a transformação ou a generalização de distribuições conhecidas na regressão binária, a fim de obter funções de ligação assimétricas e mais flexíveis. Nesse caso,

tem-se como exemplos a distribuição logística assimétrica, abordada no trabalho de Golet (2014), e a distribuição exponenciada-exponencial logística, abordada no trabalho de Prasetyo *et al.* (2020). Também é possível encontrar as distribuições LogisticF e LogisticKZ no trabalho de Huayanay *et al.* (2019); e a generalização proposta por Stukel (1988), a partir da família log F, comportando então os modelos: normal, logístico, Laplace, extremo máximo/mínimo, cloglog e loglog. Em todos os modelos citados acima, houve a adição de pelo menos um parâmetro, o que nos leva a um grave problema: o custo da generalização. Segundo Taylor (1988), a adição de parâmetros sempre resulta no aumento da variância.

Os autores Chen, Dey e Shao (1999), por sua vez, utilizaram a estatística Bayesiana para criar uma nova função de ligação, o modelo skew Bayesiano. Esse modelo consiste na utilização de variáveis latentes para combinar as distribuições acumuladas de uma distribuição simétrica com outra distribuição assimétrica em uma só função de ligação. Apesar de performar melhor que os modelos simétricos na predição de dados desbalanceados (DÁVILA-CÁRDENES *et al.*, 2021; PÉREZ-SÁNCHEZ; SALMERÓN-GÓMEZ; OCANA-PEINADO, 2019; PÉREZ-RODRÍGUEZ; PEREZ-SÁNCHEZ; GOMEZ-DENIZ, 2017; PÉREZ-SÁNCHEZ *et al.*, 2014; BERMÚDEZ *et al.*, 2008; CHEN; DEY; SHAO, 2001), esse modelo se mostra complexo e possui pouca explicabilidade, dado que envolve duas variáveis latentes e ainda a combinação de duas fda's.

Outro método utilizado para a geração de funções de ligação assimétricas é a transformação proposta por Bazán, Romeo e Rodrigues (2014). Esse método consiste na exponenciação de fda's já existentes, por um parâmetro (positivo) λ. Dessa forma, gera-se assimetria adicionando apenas um parâmetro; além disso, através desse parâmetro é possível controlar a assimetria da distribuição e, assim, estabelecer relação com as funções de ligação simétricas.

Ainda pouco explorada, essa transformação foi aplicada apenas às fda's clássicas: normal, logística, Cauchy, t de Student, Laplace e Gumbel (HUAYANAY, 2019; LEMONTE; BAZÁN, 2018; ANYOSA, 2017; BAZÁN *et al.*, 2016; BAZÁN; ROMEO; RODRIGUES, 2014). De modo que ainda é possível explorar o efeito dessa transformação em muitas outras distribuições.

Portanto, o objetivo deste trabalho é propor um modelo capaz de trazer ganhos à classificação de dados binários desbalanceados, explorando a transformação proposta por Bazán, Romeo e Rodrigues (2014), aplicada à distribuição Lomax. Apesar de Rady, Hassanein e Elhaddad (2016) já terem apresentado a distribuição Potência Lomax, o presente trabalho trata-se de um estudo inédito, visto que a distribuição Lomax e suas transformações nunca antes foram utilizadas em contextos de classificação binária e muito menos foram utilizadas na classificação de dados desbalanceados. Além disso, diferentemente da distribuição proposta por Rady, Hassanein e Elhaddad (2016), as distribuições propostas neste trabalho se estendem para toda a reta real. Buscamos, assim, apresentar uma alternativa paramétrica aos modelos tradicionais de regressão binária, que mantenha a explicabilidade, com a adição de apenas um parâmetro.

Este trabalho está organizado da seguinte forma. No Capítulo 2 é feita uma breve revisão

sobre modelos de regressão binária, são apresentados os modelos de potência e reversa de potência e também os novos modelos propostos neste trabalho. No Capítulo 3 é feita uma breve recapitulação sobre a estatística Bayesiana, estimação de parâmetros e métricas de comparação de modelos. No Capítulo 4 são apresentados os resultados das simulações para recuperação de parâmetros e o estudo de *misspecification*. No Capítulo 5 são desenvolvidas duas aplicações em bancos de dados reais sobre doação de sangue e classificação de imagens, considerando a abordagem Bayesiana. Por fim, no Capítulo 6 estão as considerações finais.

CAPÍTULO

2

MODELOS DE REGRESSÃO PARA DADOS BINÁRIOS

Neste capítulo serão apresentados os conceitos principais utilizados para desenvolver este trabalho. Primeiramente, será definido o que são dados desbalanceados; em seguida, será feita uma breve revisão sobre modelos de regressão binária e também sobre as principais funções de ligação. Posteriormente, serão apresentados os novos modelos simétricos e assimétricos propostos neste estudo: Double Lomax, Potência Double Lomax e Reversa de Potência Double Lomax. Por fim, será apresentado um estudo sobre a interpretação do parâmetro de assimetria, λ, utilizado para criar as distribuições Potência Double Lomax e Reversa de Potência Double Lomax.

2.1 Dados Desbalanceados

Segundo Kaur, Pannu e Malhi (2019), dados desbalanceados são conjuntos de dados em que uma das classes apresenta um número muito maior de observações do que a outra. Neste caso, uma classe é representada por apenas algumas observações (chamada de classe minoritária) e o resto pertence à outra classe (chamada de classe majoritária). Com esse problema, os principais algoritmos de classificação tendem a beneficiar a classe majoritária, de forma a apresentar baixa precisão na predição da classe minoritária, sendo, entretanto, a classe minoritária, em geral, a classe de maior interesse. Esse fenômeno é comum e ocorre, por exemplo, na detecção de objetos por imagem, visto que alguns objetos são muito mais comuns do que outros. Também pode ocorrer na classificação de exames para detectar doenças, visto que espera-se um número muito maior de pessoas saudáveis do que doentes.

Chen, Dey e Shao (1999) argumentam que o desbalanceamento dos dados pode ser um indicativo de que a probabilidade de que uma observação pertença à classe minoritária pode seguir uma função assimétrica das covariáveis. Por isso, neste trabalho serão apresentadas novas

funções de ligação que consideram essa possível assimetria na estimação das probabilidades.

2.2 Regressão Binária

Modelos de regressão binária nada mais são que modelos utilizados para classificar uma variável binária, ou seja, uma variável que pode assumir apenas dois valores, geralmente representados por 0 e 1. Em geral, o 1 indica que a observação pertence à categoria de interesse, fenômeno também chamado de sucesso, e o 0 indica que não pertence, fenômeno chamado de fracasso. Abaixo, encontra-se a descrição matemática do modelo.

Considere $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ uma variável binária com distribuição de Bernoulli, representada por um vetor com n observações. Seja $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})'$ um vetor de tamanho $(k+1)$, que contém k covariáveis e o intercepto, $i = 1, \ldots, n$; e seja $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ um vetor de tamanho $(k+1)$, composto pelos coeficientes de regressão associados a cada covariável e o intercepto. Agora, considere que $Y_i = 1$ com probabilidade p_i e $Y_i = 0$ com probabilidade $1 - p_i$. Então, o modelo pode ser representado como:

$$\begin{aligned} \eta_i &= \boldsymbol{\beta}' \boldsymbol{x}_i, \\ p_i &= P(Y_i = 1 | \boldsymbol{x}_i) = F(\eta_i), \\ Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i), \end{aligned}$$

em que $F(.)$ representa uma fda arbitrária e η_i representa o i-ésimo preditor linear. A abreviatura ind. significa "independentes".

2.2.1 Funções de Ligação

As funções de ligação, ou *links*, são responsáveis por estabelecer a relação entre a combinação linear das covariáveis e a probabilidade de sucesso da variável resposta. Essa função é monotônica crescente e recebe valores contínuos que se estendem por toda a reta real $(-\infty, +\infty)$ e os transforma em valores contínuos restritos ao intervalo $[0, 1]$. Devido a essas características, em geral, opta-se por fda's para essa função.

Dentre as funções de ligação mais comuns, estão a *logit*, fda da distribuição logística, e a *probit*, fda da distribuição normal (CHEN; DEY; SHAO, 1999). O *link* logístico possui a vantagem de ser a função de ligação canônica para o modelo de regressão binária, promovendo, assim, uma explicação simples para os parâmetros da regressão. Já o *link probit* é preferido por ser a fda da distribuição normal, a mais conhecida distribuição estatística, e por isso pode ser facilmente implementado em qualquer *software*.

Além dessas duas funções de ligação, temos os *links*: *cauchit*, loglog e cloglog, que, juntos aos *links logit* e *probit*, são considerados os principais *links* utilizados na regressão binária. Observe que essas são as funções de ligação implementadas nativamente no *software* `R` (R Core Team, 2022), na função `glm`.

Na Tabela 1 estão descritos matematicamente esses *links*, levando em consideração apenas a sua versão padrão (isto é, com parâmetro de locação $\mu = 0$ e parâmetro de escala $\sigma = 1$). Note que a função de ligação *cauchit* é a fda da distribuição Cauchy, a função de ligação loglog é a fda da distribuição Gumbel e o *link* cloglog é o complementar da fda da distribuição Gumbel, por isso é conhecido como complementar loglog.

Já na Figura 1 é possível observar que as curvas se diferenciam pela velocidade com que a probabilidade de sucesso aumenta. Algumas distribuições permitem que essa probabilidade aumente subitamente e, em outras, isso ocorre de forma mais lenta. Entretanto, observa-se que as curvas das funções de ligação *logit*, *probit* e *cauchit* são simétricas em torno de $\eta_i = 0$, ou $p_i = 0{,}5$. Nagler (1994) explica que, nestes casos, assume-se que indivíduos com uma probabilidade de sucesso de 0,5 são mais sensíveis a mudanças nas covariáveis, por exemplo, uma mudança de 1 unidade em X tem maior efeito em alguém que apresenta uma probabilidade de sucesso de 0,5 do que em alguém com uma probabilidade de sucesso de 0,3 ou 0,7. Isso é um dos motivos porque *links* simétricos não são sempre apropriados para dados desbalanceados.

Tabela 1 – Principais funções de ligação.

Nome	$F(\eta_i)$	$F^{-1}(p_i)$
logit	$p_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$	$\eta_i = \log\left(\frac{p_i}{1-p_i}\right)$
probit	$p_i = \Phi(\eta_i)$	$\eta_i = \Phi^{-1}(p_i)$
cauchit	$p_i = 0{,}5 + \frac{\arctan(\eta_i)}{\pi}$	$\eta_i = \tan(\pi(p_i - 0{,}5))$
loglog	$p_i = \exp(-\exp(-\eta_i))$	$\eta_i = -\log(-\log(p_i))$
cloglog	$p_i = 1 - \exp(-\exp(\eta_i))$	$\eta_i = \log(-\log(1-p_i))$

* $\Phi(\cdot)$ é a fda da distribuição normal padrão.

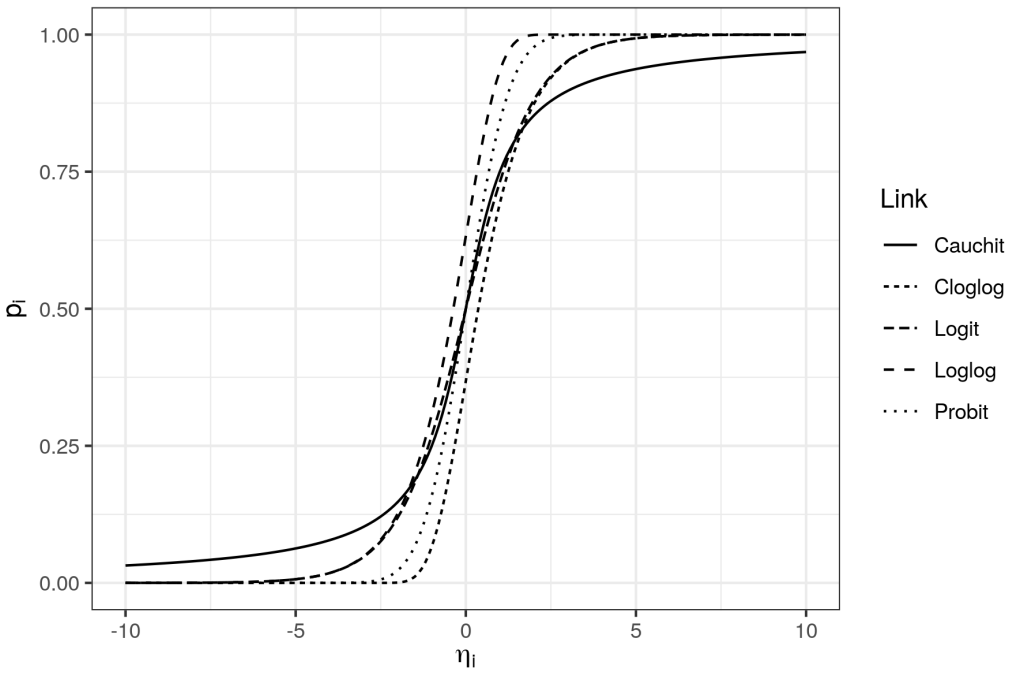


Figura 1 – Curvas de probabilidade associadas aos *links logit*, *probit*, *cauchit*, loglog e cloglog, no intervalo entre -10 e 10.

2.2.2 Distribuição Double Lomax (DLomax)

A distribuição Lomax, também conhecida como distribuição de Pareto tipo II, é uma generalização da distribuição de Pareto. Ela é conhecida por suas caudas pesadas, de modo a permitir a modelagem de dados em que uma parcela das observações possuem valores muito maiores do que a maioria. Inclusive, Bryson (1974) sugeriu que a distribuição Lomax fosse utilizada como uma alternativa à distribuição exponencial em situações em que os dados possuíssem caudas pesadas.

A distribuição Double Lomax (DLomax), por sua vez, nada mais é que uma extensão da distribuição Lomax na reta real, na qual é permitida a modelagem de uma gama maior de bancos de dados, visto que não se restringe somente a dados positivos. Sendo assim, essa distribuição é uma distribuição simétrica que possui dois lados iguais espelhados, se assemelhando à distribuição Laplace, porém com caudas pesadas.

Neste trabalho vamos considerar o modelo proposto por Bindu e Sangita (2015), no qual ele deriva essa distribuição a partir da razão de duas variáveis aleatórias independentes e identicamente distribuídas (iid) com distribuição Laplace padrão ($\mu = 0$ e $\sigma = 1$).

Definição 1. Sejam X_1 e X_2 duas variáveis aleatórias iid com distribuição Laplace padrão. Então, $X = X_1/X_2$ possui distribuição DLomax com função densidade de probabilidade (fdp) descrita pela seguinte equação:

$$f(x) = \frac{1}{2(1+|x|)^2}, \quad -\infty < x < +\infty,$$

e fda descrita por:

$$F(x) = \begin{cases} \frac{1}{2(1-x)}, & x \leq 0, \\ 1 - \frac{1}{2(1+x)}, & x > 0. \end{cases}$$

Na Figura 2 é possível observar que o pico de densidade no ponto 0 da distribuição DLomax se assemelha ao pico da distribuição Laplace, sendo maior e mais pontudo que os picos das distribuições normal e Cauchy. A DLomax, no entanto, apresenta caudas mais pesadas que as distribuições Laplace e normal, se assemelhando às caudas da distribuição Cauchy.

Já na Figura 3 é possível observar que o *link* DLomax, apesar de simétrico como o *probit* e o *logit*, apresenta crescimento e decaimento das probabilidades mais lentos, resultado das caudas mais pesadas que a distribuição possui.

2.3 Distribuições Potência e Reversa de Potência

Neste trabalho será utilizado o método proposto por Bazán, Romeo e Rodrigues (2014) para gerar distribuições assimétricas e, por consequência, *links* assimétricos. Esse método consiste na exponenciação de fda's por um parâmetro de assimetria λ.

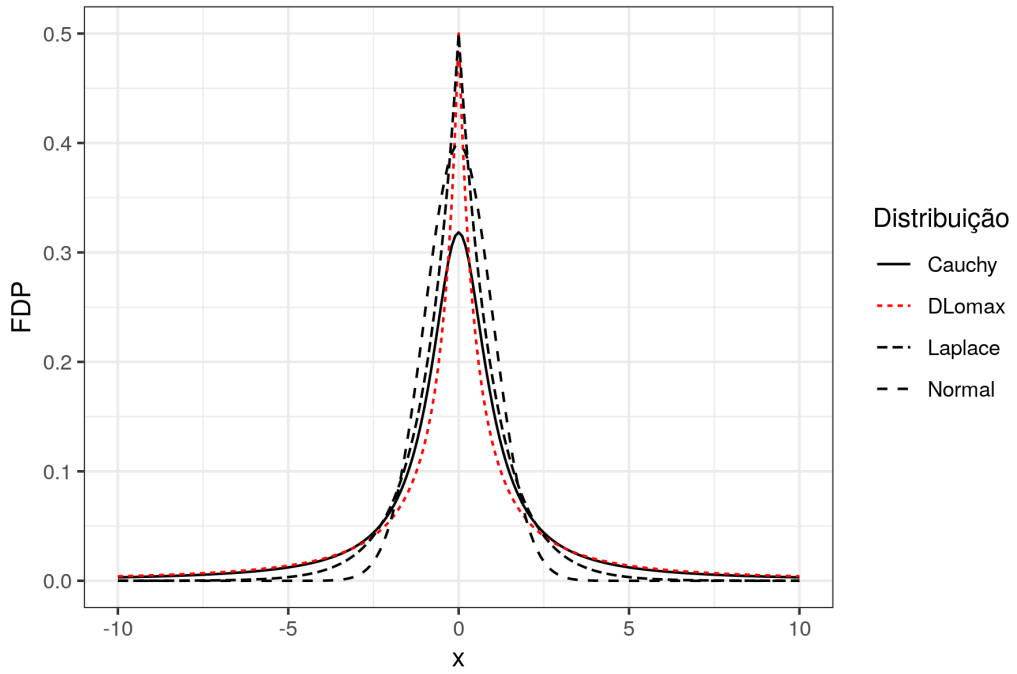


Figura 2 – Comparação da distribuição DLomax com as distribuições normal, Cauchy e Laplace.

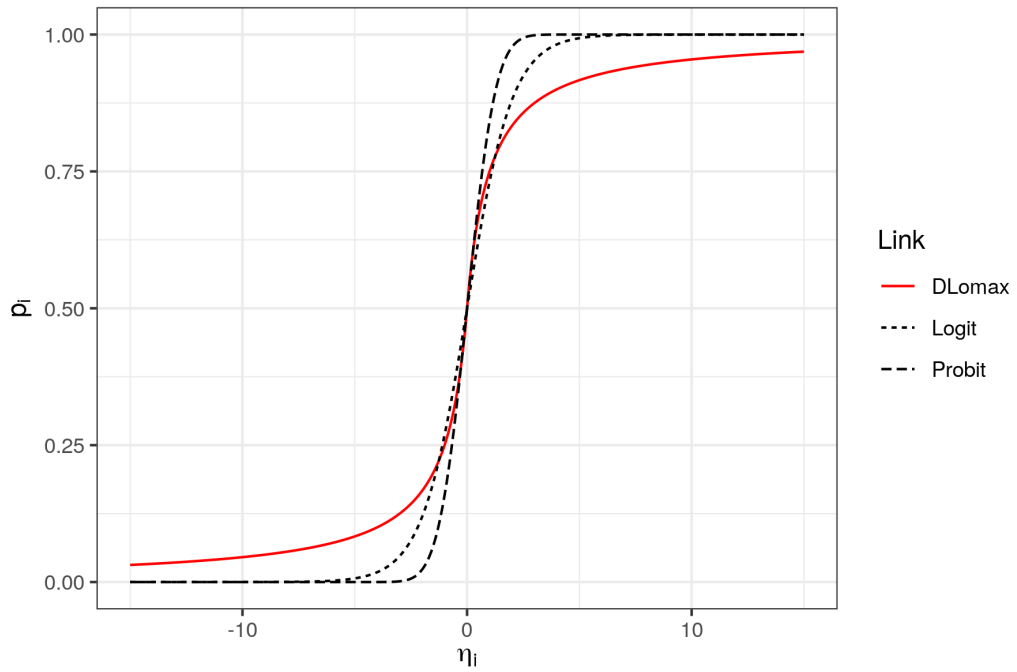


Figura 3 – Comparação da função de ligação DLomax com os *links logit* e *probit*.

Definição 2. Uma variável X possui distribuição Potência com vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$, $-\infty < \mu < \infty$, $\sigma > 0$ e $\lambda > 0$, se sua fdp pode ser escrita da seguinte forma:

$$f_{\mathrm{P}}(x) = \frac{\lambda}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)\left[G\left(\frac{x-\mu}{\sigma}\right)\right]^{\lambda-1},$$

em que $g(.)$ é uma fdp de suporte real e $G(.)$ é a fda de $g(.)$, podendo esta ser qualquer fda simétrica ou assimétrica.

A fim de simplificar as estimações, neste trabalho serão utilizadas distribuições padrões, ou seja, com locação $\mu = 0$ e escala $\sigma = 1$. Ou seja, trabalharemos com a distribuição Potência padrão.

Definição 3. Uma variável Z possui distribuição Potência padrão com parâmetro λ, $\lambda > 0$, se sua fdp pode ser escrita da seguinte forma:

$$f_{\mathrm{P}}(z) = \lambda g(z)[G(z)]^{\lambda-1},$$

em que $g(.)$ é uma fdp de suporte real e $G(.)$ é a fda de $g(.)$, podendo esta ser qualquer fda simétrica ou assimétrica.

Por consequência, sua fda pode ser escrita como:

$$F_{\mathrm{P}}(z) = [G(z)]^{\lambda}.$$

Com a criação das novas distribuições Potência, foram criadas também as distribuições Reversa de Potência, que nada mais são que as distribuições Potência espelhadas. No trabalho de Anyosa (2017), foi demonstrado que, se é possível construir uma distribuição Potência a partir de determinada distribuição, também é possível construir uma distribuição Reversa de Potência.

Definição 4. Uma variável Z possui distribuição Reversa de Potência padrão com parâmetro λ, $\lambda > 0$, se sua fdp pode ser escrita da seguinte forma:

$$f_{\mathrm{RP}}(z) = \lambda g(z)[G(-z)]^{\lambda-1},$$

em que $g(.)$ é uma fdp de suporte real e $G(.)$ é a fda de $g(.)$, podendo esta ser qualquer fda simétrica ou assimétrica.

Por consequência, sua fda pode ser escrita como:

$$F_{\mathrm{RP}}(z) = 1 - [G(-z)]^{\lambda}.$$

Também no trabalho de Anyosa (2017), foram demonstradas propriedades que estabelecem a relação entre as duas distribuições, Potência e Reversa de Potência:

- F_{P} e F_{RP} não são ponto-simétricas, pois $F_{\mathrm{P}}(-z) \neq 1 - F_{\mathrm{P}}(z)$ ou $F_{\mathrm{RP}}(-z) \neq 1 - F_{\mathrm{RP}}(z)$ para $\lambda \neq 1$;
- $F_{\mathrm{P}}(\pm z) + F_{\mathrm{RP}}(\mp z) = 1$, indicando que as duas distribuições são relacionadas, visto que uma é o inverso da outra;
- Se $\lambda = 1$, logo $F_{\mathrm{P}}(z) = G(z) = F_{\mathrm{RP}}(z)$. Isto quer dizer que $G(.)$ é um caso particular das duas distribuições ou distribuição base.

A partir desses resultados, foram criados novos *links* assimétricos. Para ilustrar o que foi apresentado, na Tabela 2 estão os *links* já criados utilizando essa transformação e aplicados no trabalho de Anyosa (2017).

Tabela 2 – Novas funções de ligação Potência e Reversa de Potência criadas a partir de *links* já existentes (*logit*, *probit*, *cauchit*, loglog e cloglog).

Função de Ligação	$F_{\mathrm{P}}(\eta_i)$	$F_{\mathrm{RP}}(\eta_i)$
logit	$p_i = \left[\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right]^{\lambda}$	$p_i = 1 - \left[\frac{\exp(-\eta_i)}{1+\exp(-\eta_i)}\right]^{\lambda}$
probit	$p_i = [\Phi(\eta_i)]^{\lambda}$	$p_i = 1 - [\Phi(-\eta_i)]^{\lambda}$
cauchit	$p_i = \left[0{,}5 + \frac{\arctan(\eta_i)}{\pi}\right]^{\lambda}$	$p_i = 1 - \left[0{,}5 + \frac{\arctan(-\eta_i)}{\pi}\right]^{\lambda}$
loglog	$p_i = [\exp(-\exp(-\eta_i))]^{\lambda}$	$p_i = 1 - [\exp(-\exp(\eta_i))]^{\lambda}$
cloglog	$p_i = [1 - \exp(-\exp(\eta_i))]^{\lambda}$	$p_i = 1 - [1 - \exp(-\exp(-\eta_i))]^{\lambda}$

2.3.1 Distribuição Potência Double Lomax (PDLomax)

A fdp da distribuição Potência Double Lomax (PDLomax) pode ser definida da forma abaixo.

Definição 5. Uma variável X possui distribuição PDLomax com parâmetro λ, $\lambda > 0$, se sua fdp pode ser escrita da seguinte forma:

$$f_{\mathrm{P}}(x) = \begin{cases} \lambda \frac{1}{2(1+|x|)^2} \left[\frac{1}{2(1-x)}\right]^{\lambda-1}, & x \leq 0, \\ \lambda \frac{1}{2(1+|x|)^2} \left[1 - \frac{1}{2(1+x)}\right]^{\lambda-1}, & x > 0, \end{cases}$$

com fda dada por:

$$F_{\mathrm{P}}(x) = \begin{cases} \left[\frac{1}{2(1-x)}\right]^{\lambda}, & x \leq 0, \\ \left[1 - \frac{1}{2(1+x)}\right]^{\lambda}, & x > 0. \end{cases}$$

Pode-se observar na Figura 4 que a adição do parâmetro λ pode introduzir tanto assimetria à direita como assimetria à esquerda. Quando $\lambda < 1$, a curva se concentra à esquerda; quando $\lambda > 1$, a curva se concentra à direita; e em $\lambda = 1$, tem-se a distribuição original.

Na Figura 5 é possível verificar como a variação de λ modifica a distribuição acumulada. Note que a variação desse parâmetro não só afeta o intervalo onde se concentra maior probabilidade, como também afeta a inclinação da curva de probabilidades. Quando λ é menor que 1, é mais provável que X seja menor que 0, mas quando λ é maior que 1, é mais provável que X seja maior que 0. Ou seja, em um contexto de regressão binária, em que essa fda seja utilizada como função de ligação, espera-se que, quando $\lambda < 1$, exista uma proporção maior de fracassos que sucessos (mais 0's do que 1's), e quando $\lambda > 1$, deve existir uma proporção maior de sucessos. Quando $\lambda = 1$, é um caso atípico em que os sucessos e fracassos estarão balanceados.

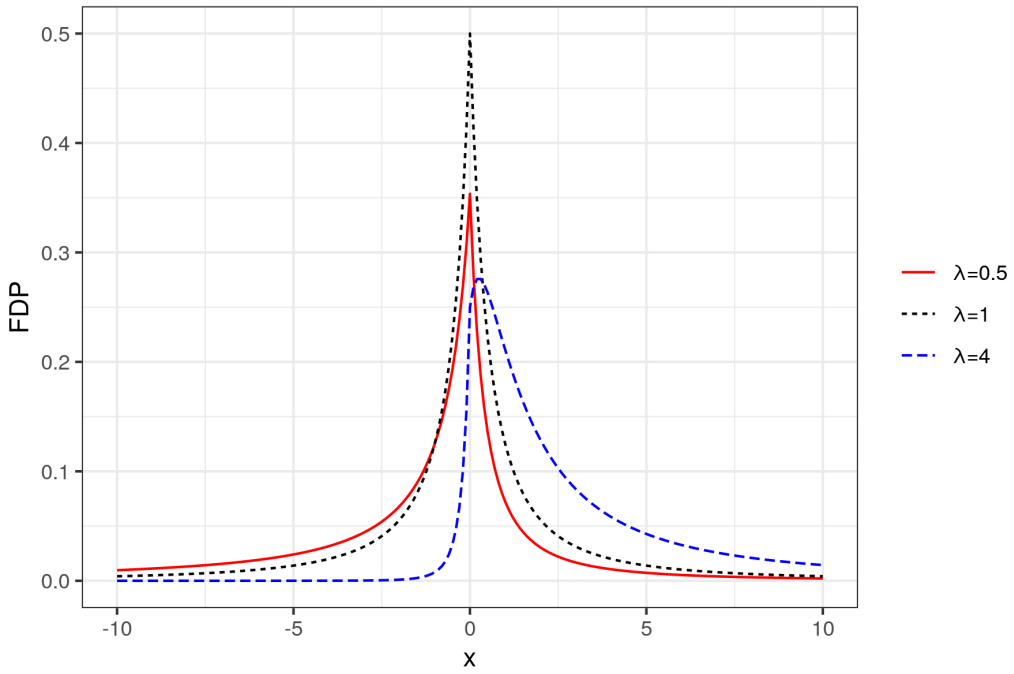


Figura 4 – Densidade da distribuição PDLomax, considerando diferentes valores de λ, $\lambda = \{0{,}5, 1, 4\}$.

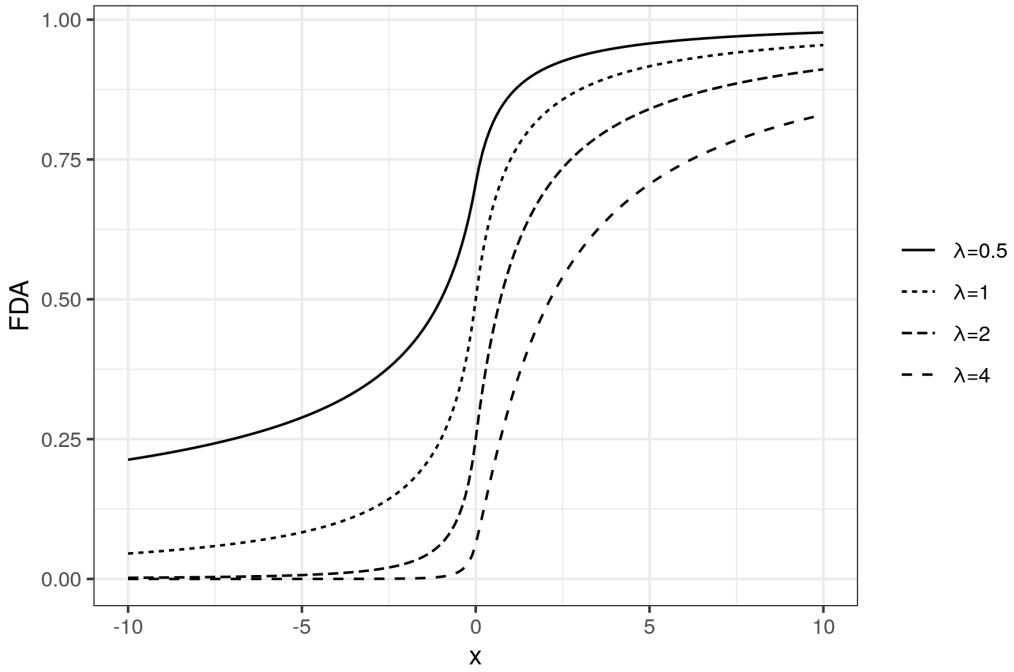


Figura 5 – Acumulada da distribuição PDLomax, considerando diferentes valores de λ, $\lambda = \{0{,}5, 1, 2, 4\}$.

2.3.2 Distribuição Reversa de Potência Double Lomax (RPDLomax)

Definição 6. Uma variável X possui distribuição Reversa de Potência Double Lomax (RPDLomax) com parâmetro λ, $\lambda > 0$, se sua fdp pode ser escrita da seguinte forma:

$$f_{\text{RP}}(x) = \begin{cases} \lambda \frac{1}{2(1+|x|)^2} \left[1 - \frac{1}{2(1-x)}\right]^{\lambda-1}, & x \leq 0, \\ \lambda \frac{1}{2(1+|x|)^2} \left[\frac{1}{2(1+x)}\right]^{\lambda-1}, & x > 0, \end{cases}$$

com fda dada por:

$$F_{\text{RP}}(x) = \begin{cases} 1 - \left[1 - \frac{1}{2(1-x)}\right]^{\lambda}, & x \leq 0, \\ 1 - \left[\frac{1}{2(1+x)}\right]^{\lambda}, & x > 0. \end{cases}$$

Ao comparar a Figura 6 com a Figura 4, é possível observar que a densidade da distribuição RPDLomax se assemelha à imagem refletida por um espelho da distribuição PDLomax. Dessa forma, o que o que ocorre é o oposto do que acontece para a distribuição PDLomax. Quando $\lambda < 1$, a curva se concentra à direita; quando $\lambda > 1$, a curva se concentra à esquerda; e em $\lambda = 1$, tem-se a distribuição DLomax.

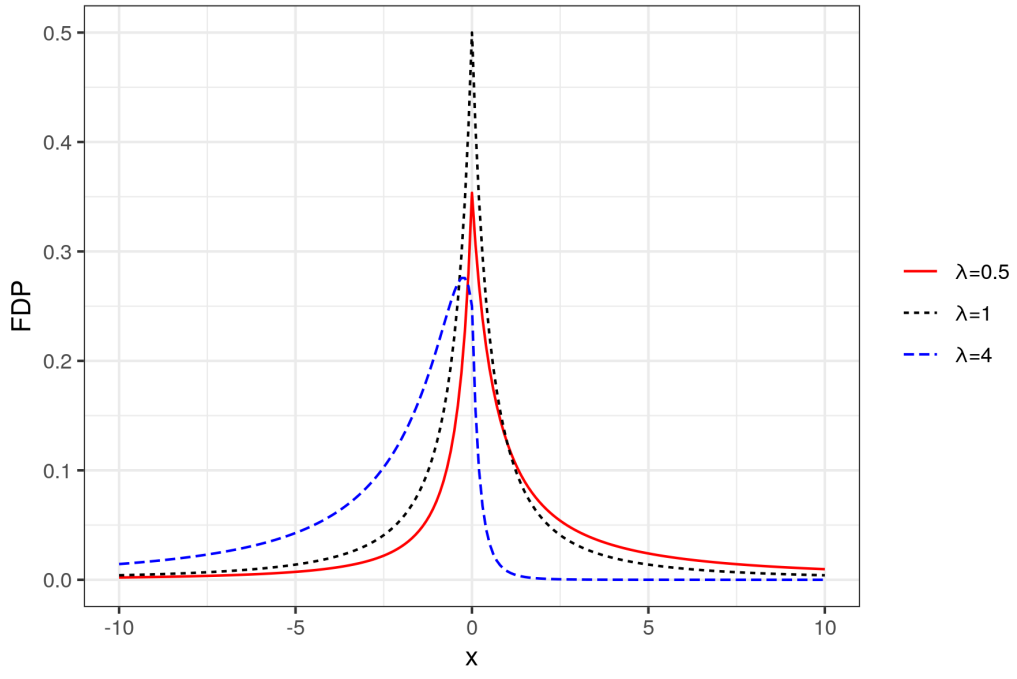


Figura 6 – Densidade da distribuição RPDLomax, considerando diferentes valores de λ, $\lambda = \{0{,}5, 1, 4\}$.

Já na Figura 7, ao observar como o parâmetro λ altera a fda, nota-se que também ocorre o inverso do que acontece na Figura 5. Quando $\lambda > 1$, há maior probabilidade de que $X < 0$; quando $\lambda < 1$, é mais provável que $X > 0$. Em um contexto de regressão binária, espera-se que, quando $\lambda < 1$, exista uma proporção maior de sucessos; e quando $\lambda > 1$, deve existir uma proporção maior de fracassos.

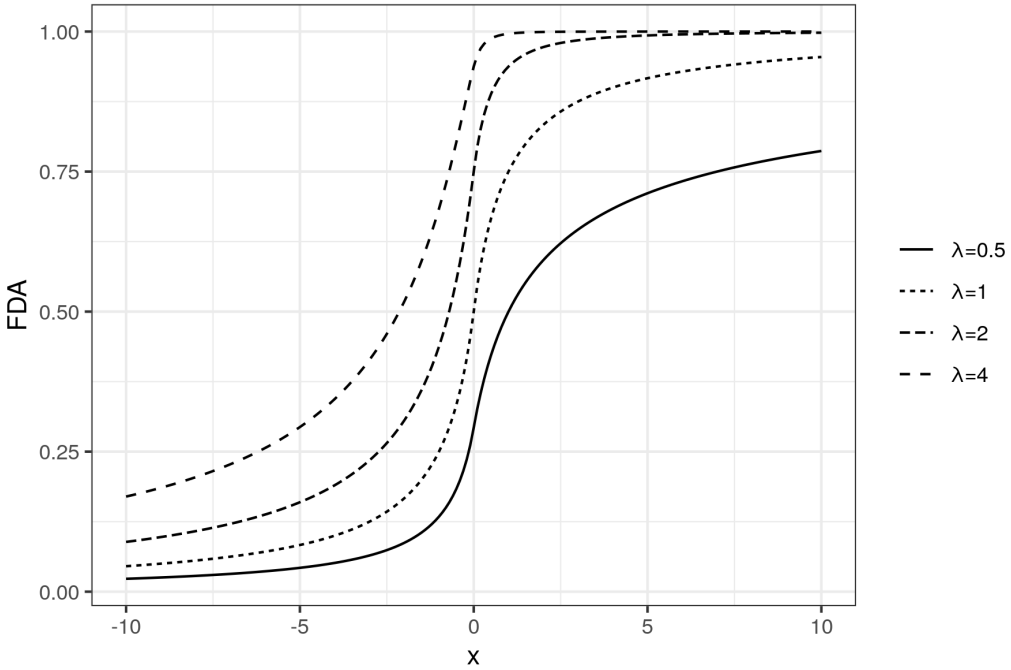


Figura 7 – Acumulada da distribuição RPDLomax, considerando diferentes valores de λ, $\lambda = \{0{,}5, 1, 2, 4\}$.

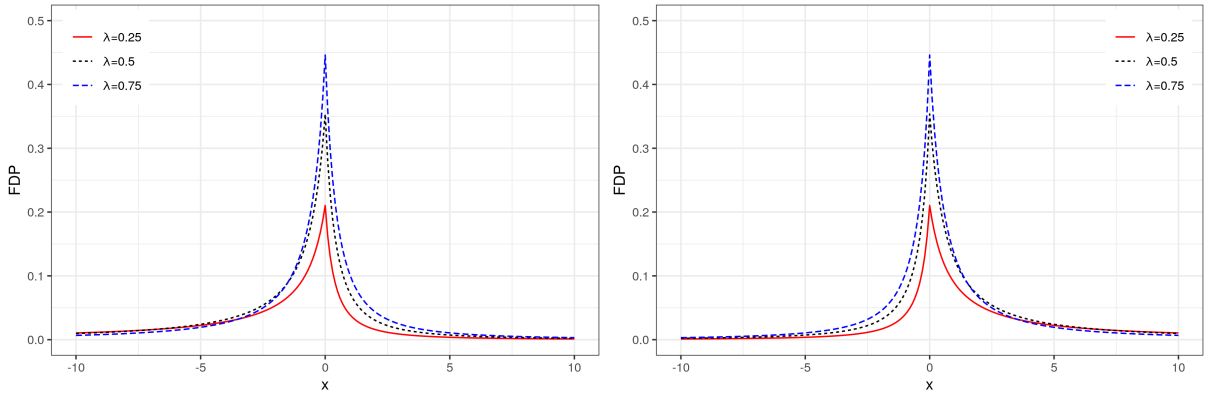
2.3.3 Interpretação do Parâmetro de Assimetria λ

Como foi visto anteriormente, as distribuições PDLomax e RPDLomax possuem comportamento distinto da distribuição DLomax. Isso se deve à adição do parâmetro λ, o qual promove a assimetria na distribuição e, consequentemente, é o responsável por gerar uma função de ligação capaz de comportar dados desbalanceados.

Na Figura 8 é possível ver claramente que as distribuições Potência e Reversa de Potência são como um reflexo uma da outra. Nota-se também que, na distribuição PDLomax, quando o parâmetro λ está no intervalo $[0, 1]$, este gera assimetria à esquerda, enquanto na distribuição RPDLomax é gerada assimetria à direita. Em ambas as distribuições, quando λ se distancia de 1, há redução no pico e aumento do peso nas caudas.

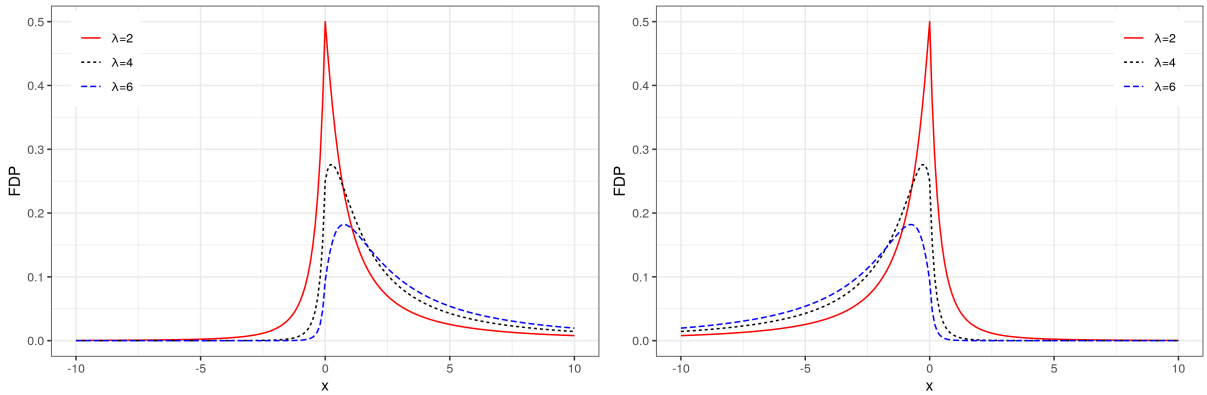
Na Figura 9, observa-se que o parâmetro λ altera a forma da distribuição. Quando $\lambda > 1$, assim como no caso anterior, à medida em que λ se afasta de 1, há redução nos picos da densidade e aumento no peso das caudas; mas diferentemente desse caso, o parâmetro λ gera assimetria à direita na distribuição PDLomax e assimetria à esquerda na distribuição RPDLomax. Além disso, nota-se que, após determinado valor de λ, a curva passa a ser arredondada.

Já na Figura 10 são nítidos os efeitos de λ na fda das duas distribuições. À medida em que λ diminui, mais rápido a curva de probabilidade se aproxima de 1 na distribuição PDLomax, enquanto na distribuição RPDLomax isso ocorre quando λ aumenta. O efeito de λ também é notado na inclinação da curva de probabilidade, visto que nas duas distribuições, à medida em que λ se distancia de 1, menor é a inclinação da curva.



(a) PDLomax (b) RPDLomax

Figura 8 – Densidade das distribuições: (a) PDLomax e (b) RPDLomax, para $0 < \lambda < 1$.



(a) PDLomax (b) RPDLomax

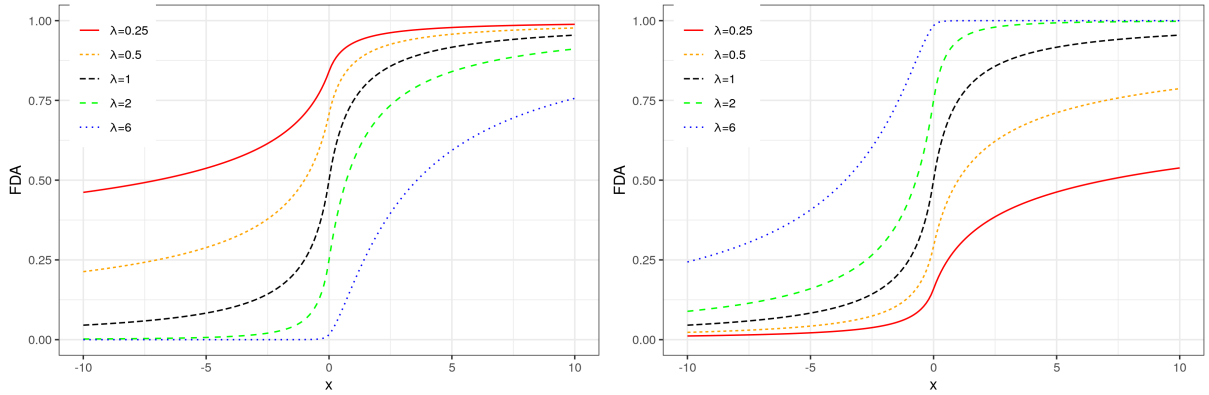
Figura 9 – Densidade das distribuições: (a) PDLomax e (b) RPDLomax, para $\lambda > 1$.

2.3.4 Assimetria nas Distribuições Potência e Reversa de Potência

A assimetria é uma medida estatística que avalia a disposição dos dados em torno de um valor central. Essa medida pode ser negativa, positiva, zero ou indefinida. A assimetria negativa indica que a cauda esquerda é mais longa e a massa da distribuição está concentrada no lado direito. Já a assimetria positiva indica que a cauda direita da distribuição é mais longa e a massa da distribuição está concentrada no lado esquerdo. Quando a assimetria possui valor zero, isso indica que a distribuição é simétrica e, logo, apresenta a mesma massa de probabilidade nos dois lados.

A assimetria de uma distribuição afeta o comportamento de sua fda, quando utilizada como função de ligação na regressão binária. Em geral, distribuições com assimetria positiva tendem a ter mais fracassos que sucessos (mais 0's do que 1's) e distribuições com assimetria negativa tendem a ter mais sucessos que fracassos (mais 1's do que 0's). Por isso, nesta seção será estudado o efeito do parâmetro λ na assimetria das distribuições PDLomax e RPDLomax.

Baseado no trabalho de Anyosa (2017) e Huayanay (2019), utilizou-se a medida de



(a) PDLomax (b) RPDLomax

Figura 10 – Acumulada das distribuições: (a) PDLomax e (b) RPDLomax, para diferentes valores de λ.

assimetria proposta por Hinkley (1975), que tem a seguinte fórmula:

$$A = \frac{(Q_{1-\mu} - Q_{0,5})(Q_{0,5} - Q_{\mu})}{(Q_{1-\mu} - Q_{\mu})},$$

em que Q_{μ} é o μ-ésimo quantil e $0 < \mu < 1$.

Neste trabalho será utilizado o coeficiente de assimetria octil, apresentado em Moors *et al.* (1996) e Hinkley (1975). Neste caso, considera-se $\mu = 0{,}125$. Desta forma, esse coeficiente é definido pela seguinte fórmula:

$$A_O = \frac{(O_7 - O_4)(O_4 - O_1)}{(O_7 - O_1)},$$

em que O_i é o octil i definido por:

$$P(X < O_i) \leq \frac{i}{8}, \qquad P(X > O_i) \geq 1 - \frac{i}{8}, \qquad i = 1, \ldots, 7.$$

E pode também ser escrito como segue:

$$A_O = \frac{(Q_{0,875} - Q_{0,5})(Q_{0,5} - Q_{0,125})}{(Q_{0,875} - Q_{0,125})}. \tag{2.1}$$

Sabe-se que, se μ for considerada a probabilidade no quantil e $F(\cdot)$ uma fda de interesse, então:

$$\mu = F(q) \quad \Rightarrow \quad Q_{\mu} = F^{-1}(\mu).$$

No caso das distribuições Potência:

$$\begin{aligned} \mu = [F(q)]^{\lambda} \quad &\Rightarrow \quad \mu^{\frac{1}{\lambda}} = F(q) \\ \Rightarrow \quad Q_{\mu} = F^{-1}(\mu^{\frac{1}{\lambda}}). \end{aligned} \tag{2.2}$$

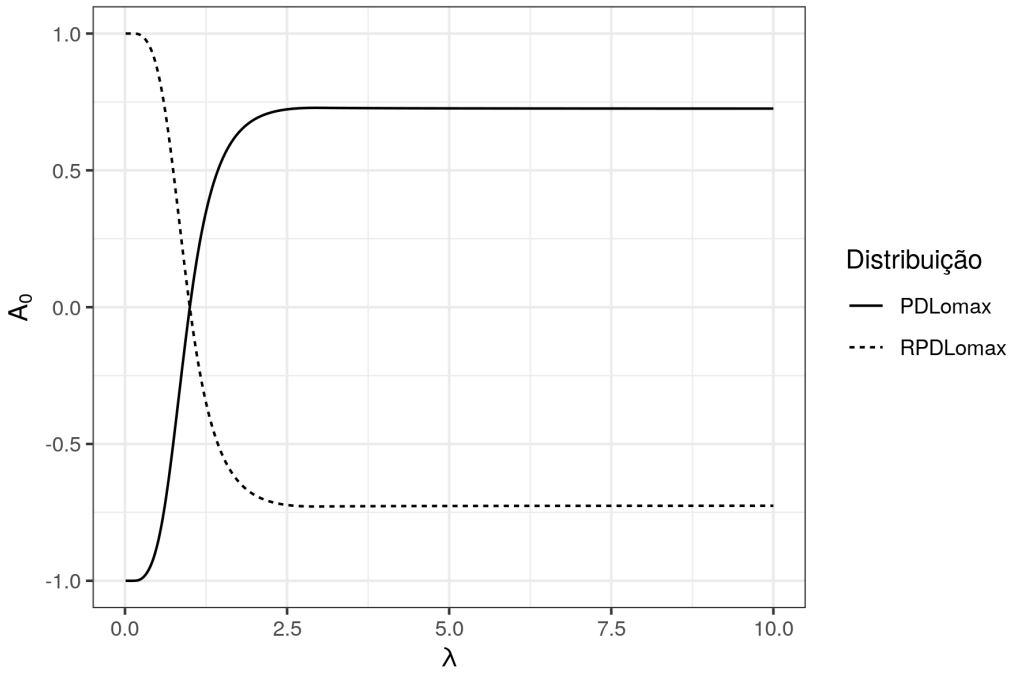


Figura 11 – Assimetria octil nas distribuições PDLomax e RPDLomax.

$F(\cdot)$ é a fda da distribuição que deu origem à distribuição Potência, neste caso, a distribuição DLomax. Já com a distribuição Reversa de Potência, tem-se que:

$$\begin{aligned} \mu = 1-[F(-q)]^{\lambda} \quad &\Rightarrow \quad (1-\mu)^{\frac{1}{\lambda}} = F(-q) \\ \Rightarrow \quad Q_{\mu} = -F^{-1}((1-\mu)^{\frac{1}{\lambda}}). \end{aligned} \tag{2.3}$$

Com manipulações matemáticas simples, pode-se encontrar que os quantis da distribuição DLomax são dados pela equação abaixo:

$$Q_{\mu} = \begin{cases} 1-\frac{1}{2q}, & 0 < q \leq 0{,}5, \\ \frac{1}{2(1-q)}-1, & 0{,}5 < q \leq 1. \end{cases} \tag{2.4}$$

Substituindo a Equação (2.4) nas Equações (2.2) e (2.3), e estas na Equação (2.1), obtém-se a assimetria das distribuições Potência e Reversa de Potência em função de λ, apresentadas nas Figuras 11 e 12.

Nas Figuras 11 e 12 é possível observar que a assimetria das distribuições PDLomax e RPDLomax são o reflexo uma da outra. Além disso, é possível verificar que, quando λ se encontra no intervalo $[0,1]$, a distribuição PDLomax possui assimetria negativa e a distribuição RPDLomax possui assimetria positiva. Já quando λ se encontra no intervalo $(0,+\infty)$, a distribuição PDLomax possui assimetria positiva e a distribuição RPDLomax possui assimetria negativa.

Nota-se também, na Figura 11, que, quando $\lambda > 4$, as duas distribuições mantêm sua assimetria praticamente constante. Já na Figura 12, nota-se que a assimetria das duas distribuições se estabiliza quando $\lambda < 0{,}125$.

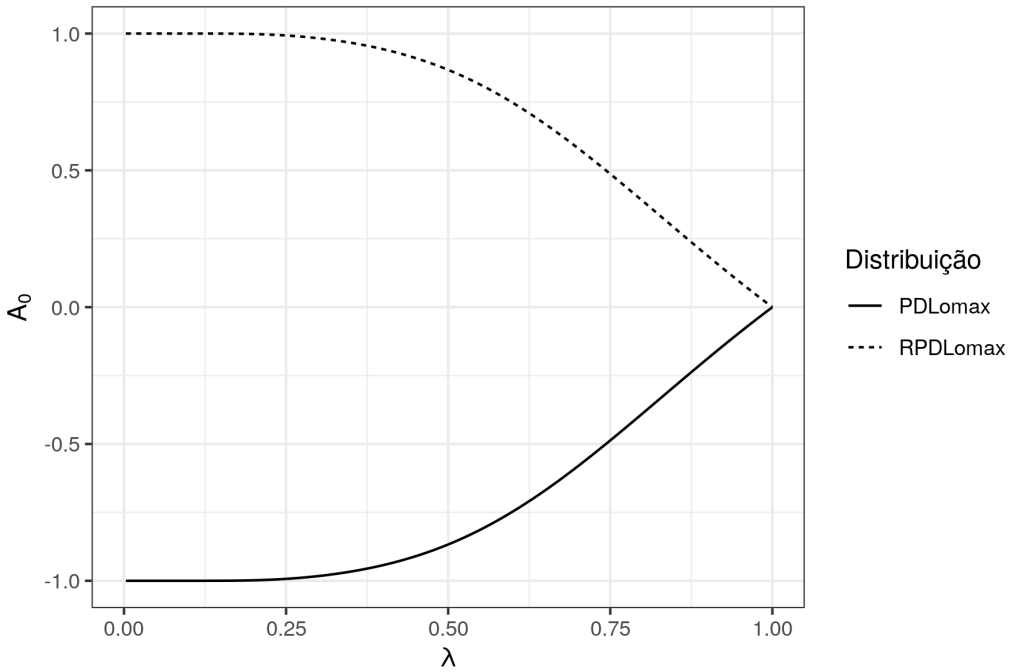


Figura 12 – Assimetria octil nas distribuições PDLomax e RPDLomax, no intervalo $[0,1]$.

2.3.5 Proporção de 0's e 1's

Para verificar o efeito do parâmetro λ na proporção de sucessos e fracassos dentro de um modelo de regressão binária, foi realizado um estudo de simulação. Neste estudo, foram simuladas 100 réplicas com 1.000 observações cada, seguindo o modelo de regressão binária com função de ligação PDLomax. Os valores de $\boldsymbol{\beta}$ foram fixados em: $\beta_0 = 0$ e $\beta_1 = 1$, e apenas uma variável X foi considerada como covariável, tendo sido esta simulada a partir da distribuição Uniforme$(-3,3)$. Abaixo está descrito matematicamente o modelo:

$$
\begin{aligned}
X_i &\overset{\text{iid}}{\sim} \text{Uniforme}(-3,3),\\
\eta_i &= \beta_0 + \beta_1 X_i,\\
p_i &= P(Y_i = 1|x_i) = F(\eta_i),\\
Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),
\end{aligned}
$$

em que $F(\cdot)$ representa uma função de distribuição DPLomax. Foram considerados diferentes cenários com diversos valores de λ. Para cada situação, foi calculada a proporção de 1's da variável resposta. Na Figura 13 é possível observar os resultados da simulação.

Observa-se na Figura 13 que, à medida em que se aumenta o valor de λ, a proporção de sucessos diminui. Quando $\lambda = 0{,}05$, há mais de 90% de sucessos; já quando $\lambda = 10$, essa proporção chega a ter, em média, 19%. Desse modo, o modelo se mostra adequado tanto para a modelagem de dados levemente desbalanceados, quanto para dados com maior nível de desbalanceamento.

O mesmo processo foi realizado para a distribuição RPDLomax. Foram simuladas 100

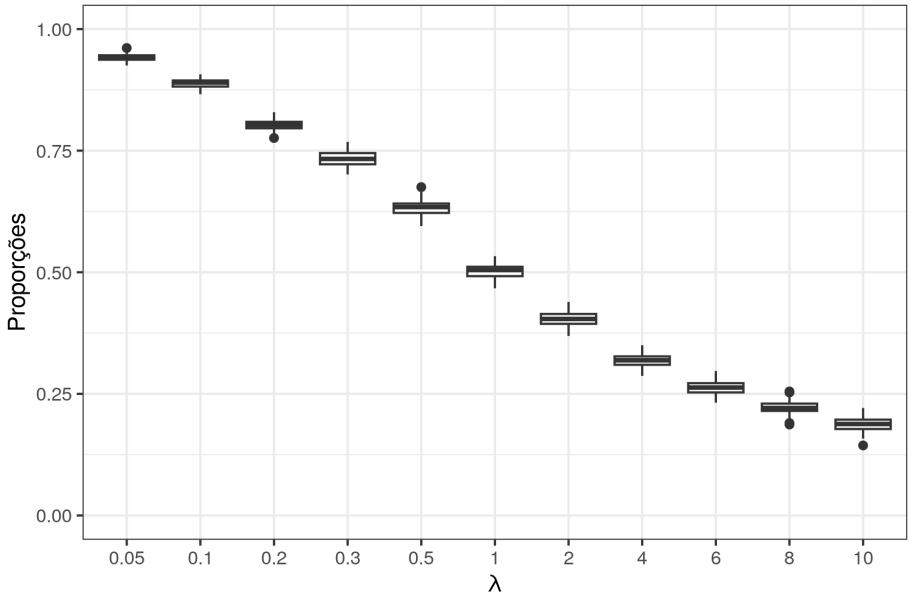


Figura 13 – *Boxplots* das proporções de 1's nas 100 réplicas, para cada cenário, sob o modelo de regressão binária com ligação DPLomax.

réplicas com 1.000 observações cada, seguindo o modelo de regressão binária com função de ligação RPDLomax. Os valores de $\boldsymbol{\beta}$ foram fixados em: $\beta_0 = 0$ e $\beta_1 = 1$, e foi considerada apenas a covariável $X \sim \text{Uniforme}(-3,3)$. Abaixo está descrito matematicamente o modelo:

$$
\begin{aligned}
X_i &\overset{\text{iid}}{\sim} \text{Uniforme}(-3,3),\\
\eta_i &= \beta_0 + \beta_1 X_i,\\
p_i &= P(Y_i = 1|x_i) = F(\eta_i),\\
Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),
\end{aligned}
$$

em que $F(\cdot)$ representa uma função de distribuição RDPLomax. Também foram considerados diferentes cenários com diversos valores de λ, e para cada situação foi calculada a proporção de 1's. Os resultados estão na Figura 14.

Na Figura 13, observa-se que o contrário ocorre com a RPDLomax: à medida em que se aumenta o valor de λ, a proporção de sucessos aumenta. Quando $\lambda = 0{,}05$, há menos de 10% de sucessos; já quando $\lambda = 10$, essa proporção chega a ter, em média, 81%.

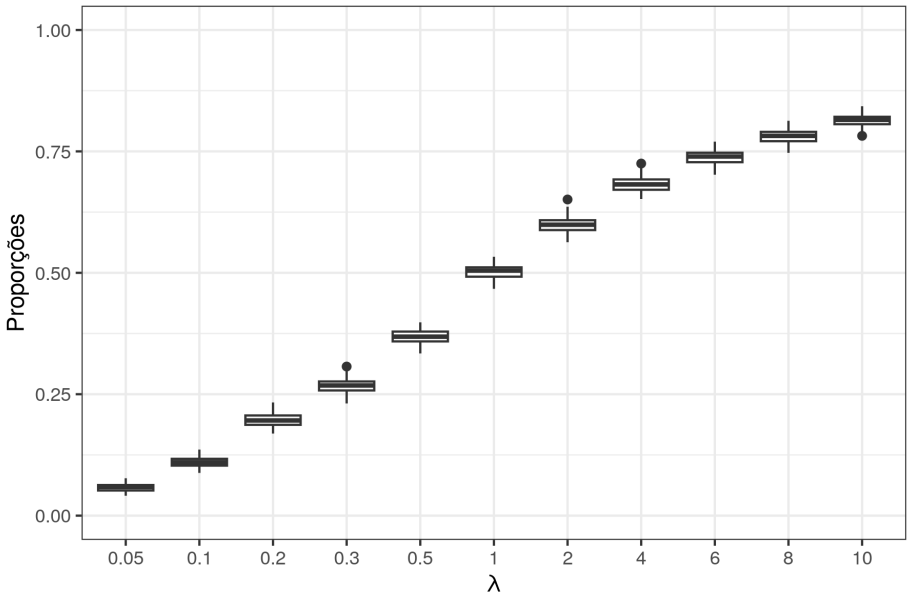


Figura 14 – *Boxplots* das proporções de 1's nas 100 réplicas, para cada cenário, sob o modelo de regressão binária com ligação RPDLomax.

CAPÍTULO

3

ESTIMAÇÃO E AJUSTE DE MODELOS

Neste capítulo serão descritas as técnicas de estatística Bayesiana utilizadas para os procedimentos inferenciais apresentados neste trabalho. Primeiramente, serão apresentadas as distribuições *a priori* dos parâmetros dos modelos propostos; em seguida, como é feito o cálculo da distribuição *a posteriori* nesses casos. Logo após isso, será descrito como é feita a estimação dos parâmetros por meio do algoritmo No-U-Turn Sampler (NUTS). Por fim, serão abordadas as métricas aplicadas para a escolha dos melhores modelos Bayesianos e para a avaliação preditiva dos modelos.

3.1 O Modelo Bayesiano

O grande diferencial da estatística Bayesiana é assumir a incerteza sobre os parâmetros do modelo. Caso se tenha algum conhecimento prévio acerca da situação a ser modelada ou de restrições sobre os parâmetros do modelo, é possível expressar matematicamente essas relações e, assim quando forem observados os dados, o modelo é atualizado, resultando em um modelo que leva em consideração o conhecimento anterior sobre a situação e o conhecimento adquirido a partir da observação do fenômeno. O conhecimento prévio sobre os dados é expresso pela distribuição *a priori*, e o conhecimento atualizado pelo evento manifesta-se na distribuição *a posteriori*. De modo que a probabilidade, nesse caso, expressa o grau de incerteza sobre um evento. Sendo assim, o modelo Bayesiano é expresso pela seguinte fórmula:

$$\pi(\boldsymbol{\theta}|x) = \frac{f_X(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f_X(x)},$$

em que $\pi(\boldsymbol{\theta}|x)$ é a distribuição *a posteriori* do vetor de parâmetros $\boldsymbol{\theta}$; $f_X(x|\boldsymbol{\theta})$ é a distribuição dos dados condicionados ao vetor de parâmetros; $\pi(\boldsymbol{\theta})$ é a distribuição *a priori* do vetor de parâmetros $\boldsymbol{\theta}$; e $f_X(x)$ é a preditiva total dos dados, ou seja, $\int_\Theta f_X(x|\boldsymbol{\theta})d\boldsymbol{\theta}$.

Como $f_X(x)$ é uma constante em relação a $\boldsymbol{\theta}$, então é válida a seguinte relação de

proporcionalidade:

$$\pi(\boldsymbol{\theta}|x) \propto f_X(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Seguindo esses moldes, o modelo Bayesiano para respostas binárias é descrito da seguinte forma:

$$\begin{aligned} Y_i|\boldsymbol{\beta},\boldsymbol{\theta} &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),\\ p_i &= F_{\boldsymbol{\theta}}(\eta_i),\\ \eta_i &= \boldsymbol{x}_i'\boldsymbol{\beta},\\ (\boldsymbol{\beta},\boldsymbol{\theta}) &\sim \pi(\boldsymbol{\beta},\boldsymbol{\theta}), \end{aligned}$$

em que $F_{\boldsymbol{\theta}}(\cdot)$ é uma fda arbitrária, $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão, $\boldsymbol{\theta}$ são os parâmetros da distribuição $F(\cdot)$, e $\pi(\cdot)$ é a distribuição *a priori* dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$.

3.2 *Prioris*

Neste trabalho será assumido que todos os parâmetros são independentes, ou seja, a distribuição *a priori* é dada por $\pi(\boldsymbol{\beta},\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta})$. Sendo assim, as *prioris* trabalhadas nos modelos serão baseadas no estudo de Bazán, Romeo e Rodrigues (2014). Além disso, será seguida a recomendação de Bazán, Romeo e Rodrigues (2014), para reparametrizar o parâmetro λ considerando $\delta = \log(\lambda)$, visto que tal reparametrização torna mais eficaz a convergência das estimações.

Dessa forma, para os modelos propostos neste trabalho, pode-se descrever o modelo Bayesiano como:

$$\begin{aligned} Y_i|\boldsymbol{\beta},\delta &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),\\ p_i &= F_{\delta}(\eta_i),\\ \eta_i &= \boldsymbol{x}_i'\boldsymbol{\beta},\\ \beta_j &\overset{\text{ind.}}{\sim} \text{Normal}(0,10^2), \quad j = 1,2,\ldots,p,\\ \delta &\sim \text{Uniforme}(-2,2), \end{aligned}$$

sendo $F_{\delta}(\cdot)$ a fda da distribuição PDLomax ou RPDLomax. As demais distribuições incluídas neste estudo seguem todas essas *prioris*, exceto a distribuição *a priori* de δ.

Note que a distribuição *a priori* de δ é uma uniforme restrita ao intervalo $(-2,2)$, ou seja, λ se restringe ao intervalo $(e^{-2},e^{2}) = (0{,}14, 7{,}39)$. Isso ocorre, pois segundo Bazán *et al.* (2016), valores fora desse intervalo possuem uma probabilidade muito baixa de ocorrência; também segundo Huayanay (2019), a assimetria das distribuições Potência se mantém praticamente constante quando λ é maior do que 6. Na Seção 2.3.4 é possível verificar que a assimetria das distribuições PDLomax e RPDLomax é constante fora dos intervalos estabelecidos nessa distribuição *a priori*.

3.3 *Posteriori*

A distribuição *a posteriori* para modelos de regressão binária que possuem parâmetro de assimetria λ, é dada por:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto L(\boldsymbol{\beta},\delta|\mathbf{x},\boldsymbol{y})\pi(\boldsymbol{\beta})\pi(\delta),$$

sendo $\pi(\boldsymbol{\beta})$ a distribuição *a priori* de $\boldsymbol{\beta}$, com $\beta_j \overset{\text{ind.}}{\sim} \text{Normal}(0,10^2)$; $\pi(\delta)$ a distribuição *a priori* de δ, com $\delta = \log(\lambda) \sim \text{Uniforme}(-2,2)$; e $L(\boldsymbol{\beta},\delta|\mathbf{x},\boldsymbol{y})$ a verossimilhança dos parâmetros dado o conjunto de dados, representada pela fórmula:

$$L(\boldsymbol{\beta},\delta|\mathbf{x},\boldsymbol{y}) = \prod_{i=1}^{n}[F_\delta(\eta_i)]^{y_i}[1-F_\delta(\eta_i)]^{1-y_i}.$$

Dessa forma, unindo as expressões descritas anteriormente, a distribuição *a posteriori* pode ser escrita como:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{x}) &\propto \prod_{i=1}^{n}[F_\delta(\eta_i)]^{y_i}[1-F_\delta(\eta_i)]^{1-y_i}\prod_{j=1}^{p}\frac{1}{10\sqrt{2\pi}}\exp\left\{-\frac{\beta_j^2}{2(10^2)}\right\}\frac{1}{4}\\ &\propto \prod_{i=1}^{n}[F_\delta(\eta_i)]^{y_i}[1-F_\delta(\eta_i)]^{1-y_i}\prod_{j=1}^{p}\exp\left\{-\frac{\beta_j^2}{2(10^2)}\right\}. \qquad (3.1)\end{aligned}$$

Note que a distribuição *a posteriori* não é semelhante às distribuições conhecidas. Logo, a estimação de parâmetros não possui solução analítica e precisa ser calculada computacionalmente.

3.4 Estimação dos Parâmetros

Toda a estimação dos parâmetros foi feita com o auxílio do pacote `stan` do *software* R (Stan Development Team, 2022). Esse programa realiza suas estimações a partir da técnica NUTS, proposta por Homan e Gelman (2014). O algoritmo NUTS é um algoritmo de Monte Carlo via cadeias de Markov (MCMC), usado para inferência Bayesiana. Sendo este uma variante auto-ajustável do método Monte Carlo Hamiltoniano (HMC), que funciona utilizando informações de gradiente para explorar eficientemente o espaço de parâmetros e evitar caminhos aleatórios, de modo a proporcionar uma convergência mais rápida. O NUTS usa um algoritmo recursivo para construir um conjunto de pontos candidatos prováveis, que abrange uma ampla faixa da distribuição alvo, parando automaticamente quando começa a retornar para o mesmo lugar. Empiricamente, o NUTS se desempenha pelo menos tão eficientemente quanto (e, às vezes, mais eficientemente do que) o método HMC, mesmo que este seja bem especificado, com a vantagem de que o NUTS funciona sem a necessidade de intervenção do usuário (HOMAN; GELMAN, 2014).

3.4.1 No-U-Turn-Sampler (NUTS)

Como foi abordado anteriormente, o algoritmo NUTS é derivado a partir do algoritmo HMC e, por isso, vamos começar apresentando ele. O HMC é um método de simulações que usa a mecânica Hamiltoniana para gerar amostras de uma distribuição alvo, sendo esta, em geral, uma distribuição *a posteriori* sem forma fechada, assim como a *posteriori* dos parâmetros dos modelos propostos neste trabalho, detalhada na Equação (3.1). Este algoritmo combina o método de Monte Carlo com o formalismo Hamiltoniano da física para produzir uma trajetória que se move a estados distantes e, mesmo assim, mantém altas probabilidades de aceitação.

Como esta abordagem tem origem na física, os cálculos foram desenvolvidos pensando-se em um sistema fechado com um número fixo de partículas e volume fixo. Esse sistema é então submerso em um banho de calor e, por isso, sua temperatura flutua até atingir o equilíbrio com o banho de calor no qual está inserido. Dentro desse sistema, consideramos as variáveis de posição (q) e momento (p) para todas as partículas e, assim, determinamos o microestado de todo o sistema. Dessa forma, uma medida importante a ser calculada é a probabilidade de que o sistema interno esteja em um determinado microestado, ou seja, em uma determinada configuração de p's e q's, probabilidade que pode ser calculada através da distribuição de Boltzmann, descrita abaixo na forma Hamiltoniana:

$$P(q,p) = \frac{1}{Z} e^{\frac{H(q,p)}{kT}},$$

em que $P(q,p)$ é a probabilidade de que o sistema esteja em determinado microestado com as coordenadas p e q, k é a constante de Boltzmann, T é a temperatura do sistema e Z é uma função de normalização chamada de função de partição canônica. A função $H(q,p)$, por sua vez, é a energia total do sistema neste estado e pode ser reescrita em termos da energia potencial, $U(q)$, e da energia cinética, $K(p)$:

$$H(p,q) = U(q) + K(p).$$

A partir destes termos, podemos levar esses conceitos para o mundo das simulações considerando que q é um vetor com as variáveis que queremos amostrar; no nosso caso, seria o vetor de parâmetros $\boldsymbol{\theta}$. E p seria um vetor de variáveis de momento introduzidas artificialmente a fim de viabilizar as simulações; no *software* `Stan` considera-se que estas são variáveis aleatórias multivariadas $\boldsymbol{\rho} \sim \text{Multinormal}(\mathbf{0}, \Sigma)$, cuja matriz de covariâncias pode tanto ser uma matriz identidade, como uma matriz estimada nas primeiras iterações do programa (restrita a ser uma matriz diagonal).

A partir destas informações podemos escrever $U(\boldsymbol{\theta})$, a energia potencial, como o negativo do logaritmo da distribuição *a posteriori* dos parâmetros:

$$U(\boldsymbol{\theta}) = -\log[L(\boldsymbol{\beta}, \delta | \mathbf{x}, \mathbf{y}) \pi(\boldsymbol{\beta}) \pi(\delta)].$$

A energia cinética, $K(\boldsymbol{\rho})$, por sua vez, pode ser escrita como:

$$K(\boldsymbol{\rho}) = \frac{1}{2}\boldsymbol{\rho}'\Sigma^{-1}\boldsymbol{\rho}.$$

Já as transições para outros estados são geradas em dois estágios, antes de serem submetidas a uma etapa de aceitação nos moldes Metropolis-Hastings (MH). Primeiramente, é tomada uma amostra da variável auxiliar de momentos $\boldsymbol{\rho}$ e, em seguida, desenvolve-se um sistema seguindo as equações Hamiltonianas:

$$\begin{cases} \dfrac{\partial \boldsymbol{\theta}}{\partial t} = \dfrac{\partial K(\boldsymbol{\rho})}{\partial \boldsymbol{\rho}}, \\ \dfrac{\partial \boldsymbol{\rho}}{\partial t} = -\dfrac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \end{cases} \tag{3.2}$$

Então, para resolver o sistema de equações (3.2), o `Stan` utiliza o algoritmo *Leapfrog*. Esse algoritmo começa a partir de valores iniciais dos parâmetros $\boldsymbol{\theta}^{(0)}$ e uma amostra aleatória de $\boldsymbol{\rho}$, e atualiza seus valores a partir de meias passadas, em que ε determina o tamanho do passo que o algoritmo dará:

$$\begin{aligned} \boldsymbol{\rho} &\leftarrow \boldsymbol{\rho} - \frac{\varepsilon}{2}\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \varepsilon\Sigma\boldsymbol{\rho}, \\ \boldsymbol{\rho} &\leftarrow \boldsymbol{\rho} - \frac{\varepsilon}{2}\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \end{aligned}$$

O algoritmo dará, no total, L passos. Dessa forma, antes de iniciar o algoritmo é preciso especificar L e ε, os quais são o número de passos e o tamanho dos passos, respectivamente. Após terminar o número de passos determinados, é calculada a probabilidade de aceitação utilizada pelo MH:

$$\min(1, \exp(H(\boldsymbol{\theta}, \boldsymbol{\rho}) - H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*))),$$

em que $\boldsymbol{\theta}^*$ e $\boldsymbol{\rho}^*$ são a versão final dos parâmetros após os L passos. Se a proposta não for aceita, os valores anteriores dos parâmetros retornam para a próxima amostra e são utilizados para começar uma nova iteração.

Dessa forma, o passo a passo do algoritmo HMC pode ser descrito nas etapas abaixo:

1. Forneça uma posição inicial $\boldsymbol{\theta}^{(0)}$ e os valores ε (tamanho dos passos), L (número de passos) e N (tamanho da cadeia);
2. Inicie um contador $i = 1, \ldots, N$:
   a) Gere uma variável inicial de momento $\boldsymbol{\rho}^* \sim \text{Multinormal}(\mathbf{0}, \Sigma)$;
   b) Estabeleça $(\boldsymbol{\theta}, \boldsymbol{\rho}) = (\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}^*)$;
   c) Para $l = 1, \ldots, L$, aplique o algoritmo *Leapfrog*:

- Atualize o momento com um meio passo: $\boldsymbol{\rho}^* = \boldsymbol{\rho}^* - \frac{\varepsilon}{2}\frac{\partial U(\boldsymbol{\theta}^{(i-1)})}{\partial \boldsymbol{\theta}}$;
- Atualize a posição: $\boldsymbol{\theta}^{(i-1)} = \boldsymbol{\theta}^{(i-1)} + \varepsilon\Sigma\boldsymbol{\rho}^*$;
- Atualize o momento: $\boldsymbol{\rho}^* = \boldsymbol{\rho}^* - \frac{\varepsilon}{2}\frac{\partial U(\boldsymbol{\theta}^{(i-1)})}{\partial \boldsymbol{\theta}}$;
- Quando $l = L$, faça a última atualização da posição: $(\boldsymbol{\theta}^{(L)}, \boldsymbol{\rho}^{(L)}) = (\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}^*)$;

d) Calcule a probabilidade de aceitação:

$$P = \min(1, \exp(H(\boldsymbol{\theta}, \boldsymbol{\rho}) - H(\boldsymbol{\theta}^{(L)}, \boldsymbol{\rho}^{(L)})));$$

e) Gere uma variável aleatória $u \sim \text{Uniforme}(0,1)$. Se $u < P$, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(L)}$; caso contrário, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}$.

Assim como foi abordado anteriormente, o algoritmo NUTS funciona exatamente como o HMC, exceto pelo fato de não necessitar da especificação pelo usuário dos parâmetros ε e L. No artigo de Homan e Gelman (2014), pode-se encontrar os cálculos realizados para obtenção do valor ótimo desses dois parâmetros. Uma revisão mais detalhada do algoritmo NUTS, aplicado a modelos Potência e Reversa de Potência, pode ser encontrada no trabalho de Anyosa (2017).

3.5 Comparação de Modelos

Para comparar os modelos apresentados neste trabalho, serão utilizadas algumas das principais medidas Bayesianas: DIC, EAIC, EBIC, WAIC e LOO. A fim de calcular essas medidas, serão utilizadas as amostras *a posteriori* dos parâmetros.

3.5.1 Deviance Information Criteria (DIC)

O DIC é uma medida baseada na média *a posteriori* do desvio, $E[D(\boldsymbol{\beta}, \delta)]$, aproximada computacionalmente por:

$$\bar{D} = \frac{1}{S}\sum_{s=1}^{S} D(\boldsymbol{\beta}^{(s)}, \delta^{(s)}),$$

para $s = 1, \ldots, S$, em que S é o tamanho da amostra *a posteriori*; e o desvio $D(\boldsymbol{\beta}, \delta)$ nada mais é do que -2 vezes a verossimilhança avaliada na s-ésima amostra dos parâmetros, isto é,

$$D(\boldsymbol{\beta}^{(s)}, \delta^{(s)}) = -2\log\left(p(\boldsymbol{y}|\boldsymbol{\beta}^{(s)}, \delta^{(s)}\right).$$

Para calcular essa medida, também é preciso calcular o desvio da média *a posteriori*, $D(E(\boldsymbol{\beta}), E(\delta))$, sendo esta uma medida aproximada computacionalmente pela seguinte fórmula:

$$\hat{D} = D\left(\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\beta}^{(s)}, \frac{1}{S}\sum_{s=1}^{S}\lambda^{(s)}\right).$$

A partir dessas medidas, obtém-se o número efetivo de parâmetros ρ_d, também conhecido como complexidade do modelo:

$$\rho_d = \bar{D} - \hat{D}.$$

Assim, o DIC pode ser obtido pela seguinte fórmula:

$$\text{DIC} = \bar{D} + \rho_d = 2\bar{D} - \hat{D}.$$

O melhor modelo é aquele que possui o menor DIC e baixa complexidade.

3.5.2 Expected Akaike Information Criterion (EAIC)

A medida EAIC é dada pela fórmula:

$$\text{EAIC} = \bar{D} + 2p,$$

em que p é o número de parâmetros do modelo.

Quanto menor a medida, mais adequado é o modelo.

3.5.3 Expected Bayesian Information Criterion (EBIC)

Já a medida EBIC é calculada da seguinte maneira:

$$\text{EBIC} = \bar{D} + p\log(n),$$

em que p é o número de parâmetros do modelo e n é o número de observações da variável resposta y.

Quanto menor seu valor, mais adequado é o modelo.

3.5.4 Widely Applicable Information Criterion (WAIC)

A métrica WAIC é uma generalização totalmente Bayesiana da métrica frequentista AIC. Para calcular essa métrica, deve-se, primeiramente, calcular o logaritmo pontual da densidade preditiva (LPPD, sigla do inglês *log pointwise predictive density*), expresso por:

$$\widehat{\text{LPPD}} = \sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\boldsymbol{\beta}^{(m)}, \lambda^{(m)})\right),$$

para $m = 1, \ldots, M$, em que M é o tamanho da amostra *a posteriori*. Já o termo de penalização possui a seguinte fórmula:

$$\hat{p}_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\boldsymbol{\beta}^{(m)}, \lambda^{(m)})\right) - \frac{1}{M}\sum_{m=1}^{M}\log\left(p(y_i|\boldsymbol{\beta}^{(m)}, \lambda^{(m)})\right)\right).$$

Por fim, o WAIC é dado por:

$$\text{WAIC} = -2(\widehat{\text{LPPD}} - \hat{p}_{\text{WAIC}}).$$

Quanto menor o WAIC, melhor o modelo.

3.5.5 Leave-One-Out Cross-Validation (LOO)

A métrica LOO, assim como o WAIC, também é uma métrica totalmente Bayesiana. No entanto, ela possui alto custo computacional quando se trata de amostras muito grandes. Por isso, Vehtari, Gelman e Gabry (2016) propuseram o método PSIS-LOO (sigla do inglês *Pareto Smoothed Importance Sampling Leave-One-Out cross-validation*). Para calcular essa métrica, pode-se usar a fórmula abaixo:

$$\widehat{\text{ELPD}}_{\text{PSIS-LOO}} = \sum_{i=1}^{n} \log \left(\sum_{m=1}^{M} \frac{w_i^{(m)} \, p(y_i|\boldsymbol{\beta}^{(m)}, \lambda^{(m)})}{\sum_{m=1}^{M} w_i^{(m)}} \right),$$

em que:

$$w_i^{(m)} = \min \left(r_i^{(m)}, \frac{\sqrt{M}}{M} \sum_{m=1}^{M} r_i^{(m)} \right),$$

com:

$$r_i^{(m)} = \frac{1}{p(y_i|\boldsymbol{\beta}^{(m)}, \lambda^{(m)})} \propto \frac{p(\boldsymbol{\beta}^{(m)}, \lambda^{(m)}|\boldsymbol{Y}_{-i})}{p(\boldsymbol{\beta}^{(m)}, \lambda^{(m)}|\boldsymbol{Y})},$$

para $m = 1, \ldots, M$, em que M é o tamanho da amostra *a posteriori*.

Quanto menor o LOO, mais adequado o modelo.

3.5.6 Avaliação Preditiva

Além de avaliar o ajuste do modelo, é importante avaliar seu poder de classificar as observações. Por isso, nesta seção serão apresentadas as métricas utilizadas para avaliar a performance preditiva do modelo. Primeiramente, será apresentada a matriz de confusão.

Tabela 3 – Matriz de confusão.

		Observado	
		0	1
Predito	0	VN	FN
	1	FP	VP

Na Tabela 3 estão os valores possíveis da variável resposta Y: $Y = 1$, se sucesso; e $Y = 0$, se fracasso. VP são os verdadeiros positivos, isto é, sucessos que foram classificados como sucessos; FP são os falsos positivos, ou seja, fracassos que foram classificados como sucessos; FN são os falsos negativos, isto é, sucessos que foram classificados como fracassos; e VN são os verdadeiros negativos, ou seja, fracassos que foram classificados como fracassos.

A partir dessa matriz, serão calculadas as principais métricas utilizadas neste trabalho:

- *Acurácia*: é a proporção de acertos do modelo, representada pelo número de acertos dividido pelo número total de observações:

$$\frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}};$$

- *Precisão*: é a proporção de sucessos classificados corretamente, representada pelo número de sucessos classificados como sucessos dividido pelo número total das observações classificadas como sucessos:
$$\frac{\text{VP}}{\text{VP}+\text{FP}};$$
- *Sensibilidade (SENS)*: é a proporção de classificações corretas de sucessos, ou seja, o número de sucessos classificados como sucessos dividido pelo total real de sucessos:
$$\frac{\text{VP}}{\text{VP}+\text{FN}};$$
- *Especificidade (ESP)*: é a proporção de fracassos classificados corretamente, representada pelo número de fracassos classificados como fracassos dividido pelo número total de fracassos reais:
$$\frac{\text{VN}}{\text{VN}+\text{FP}};$$
- *F1-score (F1)*: é uma média harmônica da precisão e sensibilidade, cuja fórmula está representada logo abaixo:
$$2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}.$$

Outra forma de avaliar o poder preditivo do modelo é a curva ROC. A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica utilizada para avaliar a eficiência de um modelo de classificação binária. Ela mostra a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1-especificidade), para vários pontos de corte do modelo. A área sob a curva (AUC) representa a habilidade geral do modelo em diferenciar entre as duas classes. Quanto maior a AUC, melhor o modelo é em fazer previsões corretas.

Conforme é sugerido por Huayanay (2019) e Huayanay *et al.* (2019), será utilizado o ponto de corte 0,5 para determinar se uma observação será classificada como sucesso ou fracasso. Ou seja, caso $F(\eta_i) > 0{,}5$, a observação i será classificada como sucesso, e caso $F(\eta_i) \leq 0{,}5$, ela será classificada como fracasso.

3.5.7 Resíduos

Para analisar a adequabilidade do modelo, foram utilizados os resíduos quantílicos aleatorizados. Quando o modelo é adequado aos dados, esses resíduos seguem uma distribuição normal.

Segundo a definição de Dunn e Smyth (1996), se $F(\cdot)$ é uma fda contínua, então $F(y_i|\mu_i,\sigma)$ é distribuída uniformemente no intervalo unitário. Então, os resíduos quantílicos aleatorizados são dados pela fórmula abaixo:

$$r_{q,i} = \Phi^{-1}\{F(y_i|\hat{\mu}_i,\hat{\sigma})\}.$$

Caso $F(\cdot)$ seja uma fda discreta, então uma definição mais geral é adotada. Considere $a = \lim_{y \uparrow y_i} F(y|\hat{\mu}_i, \hat{\sigma})$ e $b = F(y_i|\hat{\mu}_i, \hat{\sigma})$. Então,

$$r_{q,i} = \Phi^{-1}(u_i),$$

em que u_i é uma variável aleatória com distribuição uniforme em $(a_i, b_i]$.

Maiores detalhes sobre esses resíduos podem ser encontrados no trabalho de Dunn e Smyth (2018), literatura utilizada para a geração e verificação dos resíduos neste trabalho.

CAPÍTULO

4

SIMULAÇÕES

Neste capítulo será apresentado um estudo de simulação, a fim de avaliar a capacidade dos modelos propostos em estimar seus parâmetros. A simulação de recuperação de parâmetros refere-se ao processo de avaliar o desempenho de um modelo estatístico, comparando os parâmetros estimados aos verdadeiros (ou conhecidos) utilizados para gerar dados simulados. Além disso, foi conduzido um estudo de *misspecification*, a fim de avaliar se os modelos propostos conseguem performar melhor que o modelo de regressão logística, em diferentes cenários de desbalanceamento.

4.1 Recuperação de Parâmetros

Para avaliar a precisão das estimativas dos parâmetros do modelo, foi proposto o estudo de recuperação de parâmetros. Neste estudo foram geradas 100 amostras aleatórias de tamanho $n = \{500, 1.000, 2.000\}$ de cada um dos modelos propostos e também do modelo logístico (utilizado para fins de comparação). Nos modelos que possuem o parâmetro de assimetria (λ), foram considerados mais quatro cenários: $\lambda = \{0{,}25; 0{,}5; 2; 4\}$. A covariável X foi simulada a partir da distribuição Uniforme$(-3,3)$ e os coeficientes de regressão foram fixados em $\boldsymbol{\beta} = (\beta_0, \beta_1) = (0,1)$.

Para estimar os parâmetros pelo *software* `stan`, foram consideradas 200 iterações, 4 cadeias e 100 amostras de *warm up* a cada cadeia. Desse modo, para cada réplica foram geradas 400 amostras de cada parâmetro estimado. As métricas utilizadas para verificar o desempenho dos modelos foram o viés e o erro quadrático médio (EQM), que são dados, respectivamente, por:

$$\text{Viés}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta) \qquad \text{e} \qquad \text{EQM}(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta)^2},$$

em que R é o número de réplicas na simulação (neste caso, $R = 100$), θ é o valor real do parâmetro, e $\hat{\theta}^{(r)}$ é a média *a posteriori* do parâmetro θ na réplica r.

Além disso, verificou-se a qualidade da estimação intervalar, observando a frequência relativa de vezes em que o parâmetro verdadeiro (θ) estava contido nos quantis 2,5 e 97,5%, ou seja, a proporção de vezes em que o parâmetro original se encontrava entre os os quantis 2,5 e 97,5% das amostras da distribuição *a posteriori*, aqui chamada de probabilidade de cobertura (PC). Também verificou-se a proporção de vezes em que o parâmetro verdadeiro foi menor que o quantil 2,5% (α_1), e a proporção de vezes em que ele foi maior que o quantil 97,5% (α_2). Isto é,

$$\text{PC} = \frac{1}{R}\sum_{r=1}^{R} I(\theta \in [\text{LI}, \text{LS}]),$$
$$\alpha_1 = \frac{1}{R}\sum_{r=1}^{R} I(\theta < \text{LI}),$$
$$\alpha_2 = \frac{1}{R}\sum_{r=1}^{R} I(\theta > \text{LS}),$$

em que LI e LS são, respectivamente, os limites inferior (quantil 2,5%) e superior (quantil 97,5%) das amostras *a posteriori* a cada réplica e R é o número de réplicas na simulação. Ou seja, está sendo calculada a cobertura do intervalo de credibilidade de 95% das cadeias MCMC. Segundo Alves, Bazán e Arellano-Valle (2022), espera-se que, em 95% dos casos, o valor real do parâmetro esteja contido nesse intervalo.

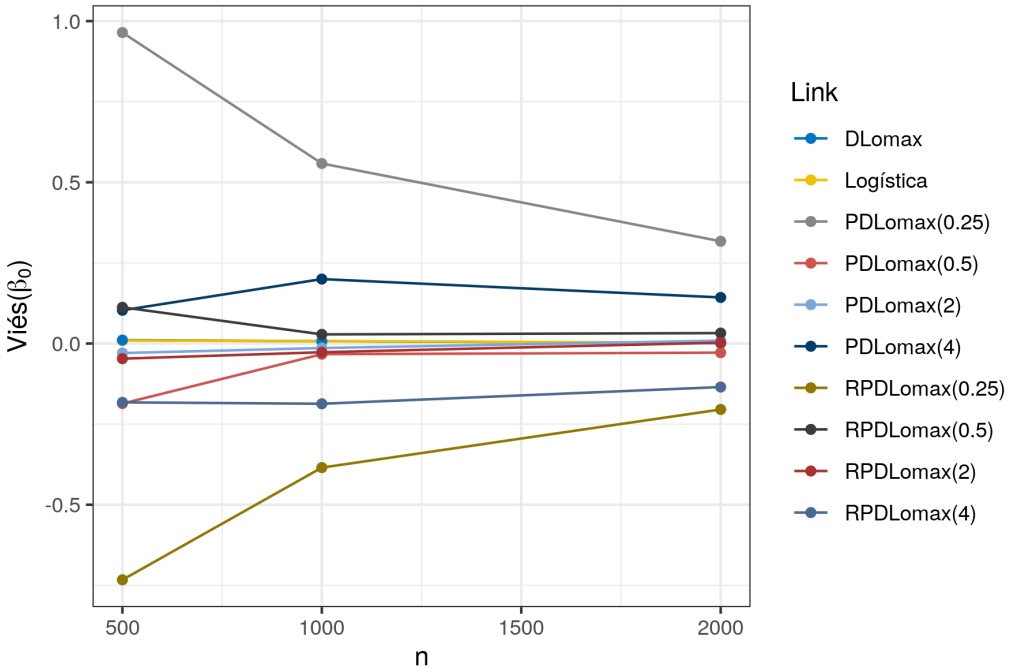


Figura 15 – Viés para o parâmetro β_0 nos modelos logístico, DLomax, PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

Na Tabela 4 e na Figura 15 é possível observar que, em todos os modelos, o viés do estimador $\hat{\beta}_0$ decai à medida em que se aumenta o tamanho da amostra, sugerindo que esse

Tabela 4 – Viés, EQM, PC, α_1 e α_2 para o parâmetro β_0, estimados a partir da simulação dos modelos logístico, DLomax, PDLomax e RPDLomax, com parâmetro $\lambda = \{0{,}25; 0{,}5; 2; 4\}$.

Modelo	n	Viés	EQM	PC	α_1	α_2
Logístico	500	0,009336	0,011062	0,96	0,03	0,01
	1.000	0,007294	0,003106	0,91	0,07	0,02
	2.000	0,002340	0,007989	0,91	0,07	0,02
DLomax	500	0,010557	0,016120	0,96	0,03	0,01
	1.000	0,007119	0,009000	0,96	0,02	0,02
	2.000	0,000999	0,004144	0,95	0,02	0,03
PDLomax($\lambda = 0{,}25$)	500	0,964450	9,805765	0,95	0,03	0,02
	1.000	0,558320	3,447603	0,94	0,01	0,05
	2.000	0,317157	1,862033	0,95	0,01	0,04
PDLomax($\lambda = 0{,}5$)	500	-0,186220	0,279098	0,93	0,01	0,06
	1.000	-0,033062	0,139868	0,93	0,04	0,03
	2.000	-0,028356	0,063545	0,95	0,01	0,04
PDLomax($\lambda = 2$)	500	-0,029580	0,118582	0,92	0,03	0,05
	1.000	-0,013971	0,026944	0,95	0,04	0,01
	2.000	0,009064	0,014092	0,93	0,05	0,02
PDLomax($\lambda = 4$)	500	0,103125	0,100971	0,97	0,02	0,01
	1.000	0,199801	0,109813	0,89	0,08	0,03
	2.000	0,142873	0,066046	0,86	0,13	0,01
RPDLomax($\lambda = 0{,}25$)	500	-0,732793	7,505925	0,95	0,04	0,01
	1.000	-0,384836	2,266346	0,95	0,02	0,03
	2.000	-0,204476	0,861617	0,93	0,04	0,03
RPDLomax($\lambda = 0{,}5$)	500	0,112331	0,494380	0,87	0,08	0,05
	1.000	0,028317	0,111804	0,96	0,04	0,00
	2.000	0,032333	0,036533	0,97	0,01	0,02
RPDLomax($\lambda = 2$)	500	-0,046962	0,101626	0,91	0,02	0,07
	1.000	-0,027002	0,041975	0,92	0,04	0,04
	2.000	0,002819	0,011004	0,99	0,00	0,01
RPDLomax($\lambda = 4$)	500	-0,182479	0,123832	0,93	0,01	0,06
	1.000	-0,186853	0,092649	0,91	0,00	0,09
	2.000	-0,134936	0,075850	0,89	0,02	0,09

estimador é assintoticamente não viesado nos modelos propostos. Também é possível observar na Figura 18 que o EQM decai à medida em que se aumenta o tamanho da amostra, em todos os modelos. Os modelos que apresentam maior viés e maior EQM são os modelos que tratam de dados com maior desbalanceamento, isto é, com $\lambda = 0{,}25$ e $\lambda = 4$. Comparativamente ao modelo logístico, em tamanhos de amostra menores os modelos propostos apresentam maior viés na recuperação de parâmetros, entretanto, à medida em que o tamanho das amostras cresce, a diferença entre esses modelos se torna irrisória. A PC para os parâmetros dos modelos também se mostra dentro do esperado, próxima a 95%, indicando que os intervalos de credibilidade estimados são razoáveis.

Na Tabela 5 e nas Figuras 16 e 19 nota-se um comportamento parecido (com o que ocorre em β_0) na estimação de β_1: o viés decai à medida em que se aumenta o tamanho da amostra,

Tabela 5 – Viés, EQM, PC, α_1 e α_2 para o parâmetro β_1, estimados a partir da simulação dos modelos logístico, DLomax, PDLomax e RPDLomax, com parâmetro $\lambda = \{0{,}25; 0{,}5; 2; 4\}$.

Modelo	n	Viés	EQM	PC	α_1	α_2
Logístico	500	0,015784	0,009180	0,95	0,03	0,02
	1.000	0,007444	0,003752	0,95	0,03	0,02
	2.000	0,002479	0,001401	0,97	0,02	0,01
DLomax	500	0,066450	0,036719	0,96	0,04	0,00
	1.000	0,049272	0,019721	0,92	0,07	0,01
	2.000	0,020251	0,007287	0,93	0,05	0,02
PDLomax($\lambda = 0{,}25$)	500	0,784837	1,059169	0,88	0,12	0,00
	1.000	0,410339	0,395220	0,90	0,10	0,00
	2.000	0,175835	0,127842	0,94	0,05	0,01
PDLomax($\lambda = 0{,}5$)	500	0,306867	0,242185	0,90	0,10	0,00
	1.000	0,160536	0,080868	0,87	0,11	0,02
	2.000	0,061845	0,019854	0,94	0,05	0,01
PDLomax($\lambda = 2$)	500	0,096037	0,035402	0,96	0,03	0,01
	1.000	0,052916	0,015558	0,97	0,03	0,00
	2.000	0,022471	0,006098	0,96	0,03	0,01
PDLomax($\lambda = 4$)	500	0,382190	0,257836	0,90	0,09	0,01
	1.000	0,224471	0,105081	0,86	0,14	0,00
	2.000	0,130355	0,053331	0,87	0,10	0,03
RPDLomax($\lambda = 0{,}25$)	500	0,640248	0,800568	0,89	0,11	0,00
	1.000	0,305705	0,268762	0,91	0,09	0,00
	2.000	0,131998	0,095389	0,96	0,03	0,01
RPDLomax($\lambda = 0{,}5$)	500	0,296647	0,258162	0,92	0,08	0,00
	1.000	0,117836	0,061227	0,93	0,06	0,01
	2.000	0,046049	0,018984	0,94	0,04	0,02
RPDLomax($\lambda = 2$)	500	0,087997	0,032410	0,96	0,04	0,00
	1.000	0,062053	0,024288	0,89	0,08	0,03
	2.000	0,019331	0,007559	0,96	0,04	0,00
RPDLomax($\lambda = 4$)	500	0,337816	0,198695	0,87	0,13	0,00
	1.000	0,224588	0,114505	0,84	0,15	0,01
	2.000	0,139672	0,064784	0,88	0,10	0,02

assim como também decai o EQM. Nesse caso, também são os modelos mais assimétricos, isto é, com $\lambda = 0{,}25$ e $\lambda = 4$, que apresentam maior viés e maior EQM. Além disso, os modelos também apresentam maior viés comparativamente ao modelo logístico, com essa diferença decaindo quando se aumenta o tamanho da amostra. A PC, por sua vez, também se mostra dentro do esperado em todos os modelos, não existindo indicativos de que o intervalo de credibilidade para os parâmetros possua irregularidades.

Já na Tabela 6 e na Figura 17 é possível observar que ocorre um leve aumento no viés de $n = 500$ para $n = 1.000$ quando $\lambda = 4$; e também na Figura 20 é possível observar um aumento no EQM sob essas condições. Isso pode ocorrer, pois, segundo Huayanay (2019), a relação entre o aumento em λ e a assimetria não ocorre de forma linear; após um certo ponto, qualquer aumento em λ gera acréscimos insignificantes na assimetria. Mesmo assim, é possível observar

Tabela 6 – Viés, EQM, PC, α_1 e α_2 para o parâmetro λ, estimados a partir da simulação dos modelos PDLomax e RPDLomax, com parâmetro $\lambda = \{0{,}25; 0{,}5; 2; 4\}$.

Modelo	n	Viés	EQM	PC	α_1	α_2
PDLomax($\lambda = 0{,}25$)	500	0,044265	0,344668	0,93	0,03	0,04
	1.000	0,022592	0,025538	0,91	0,03	0,06
	2.000	0,009388	0,010934	0,96	0,01	0,03
PDLomax($\lambda = 0{,}5$)	500	-0,039099	0,008876	0,91	0,01	0,08
	1.000	-0,012342	0,006662	0,91	0,04	0,05
	2.000	-0,005105	0,003802	0,91	0,03	0,06
PDLomax($\lambda = 2$)	500	-0,001970	0,166265	0,95	0,02	0,03
	1.000	0,006371	0,040194	0,94	0,04	0,02
	2.000	0,008833	0,019403	0,95	0,04	0,01
PDLomax($\lambda = 4$)	500	0,745615	1,495800	0,96	0,02	0,02
	1.000	0,748266	1,544607	0,83	0,13	0,04
	2.000	0,512540	0,898172	0,87	0,12	0,01
RPDLomax($\lambda = 0{,}25$)	500	0,025731	0,222899	0,95	0,01	0,04
	1.000	0,015414	0,019497	0,93	0,02	0,05
	2.000	0,011343	0,004697	0,96	0,03	0,01
RPDLomax($\lambda = 0{,}5$)	500	-0,008495	0,025805	0,89	0,03	0,08
	1.000	-0,006515	0,004881	0,94	0,01	0,05
	2.000	-0,007753	0,001745	0,95	0,00	0,05
RPDLomax($\lambda = 2$)	500	0,099666	0,166624	0,92	0,05	0,03
	1.000	0,051525	0,093688	0,93	0,05	0,02
	2.000	-0,003245	0,015933	0,96	0,01	0,03
RPDLomax($\lambda = 4$)	500	0,905929	1,848415	0,98	0,01	0,01
	1.000	0,740199	1,437555	0,85	0,14	0,01
	2.000	0,454480	1,027973	0,86	0,13	0,01

uma forte tendência de queda tanto no viés como no EQM quando $n = 2.000$, indicando a possibilidade de reduzir o viés do modelo assintoticamente. Para o restante dos modelos que possuem o parâmetro λ, observa-se uma queda tanto no viés como no EQM à medida em que se aumenta o tamanho das amostras. Os intervalos de credibilidade para λ também parecem ter comportamento razoável, dado o número de repetições do experimento, entretanto, é notável que a PC é menor quando $\lambda = 4$.

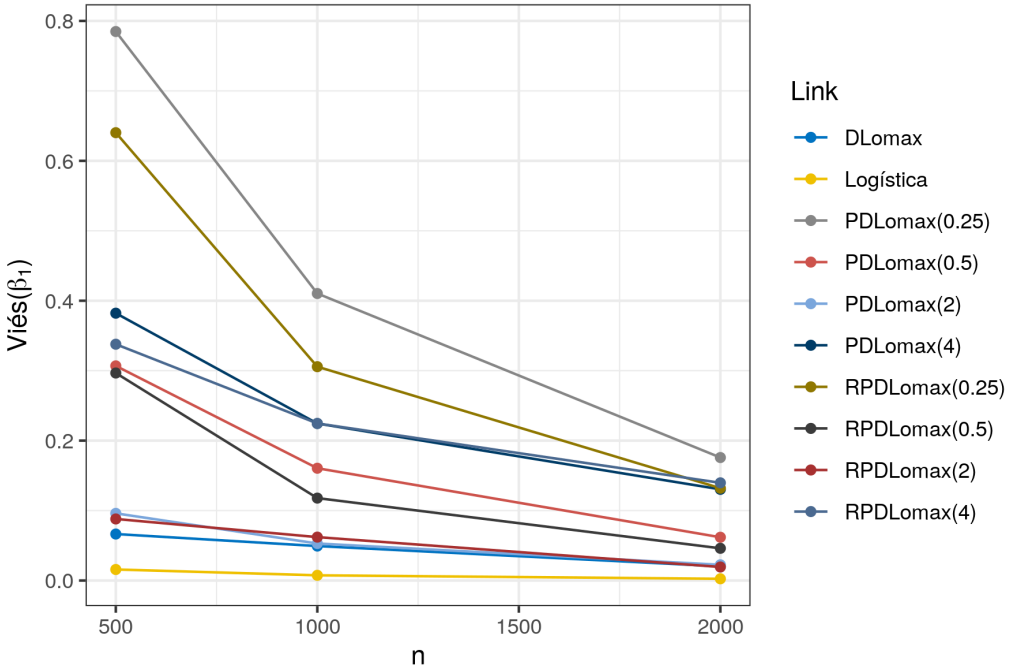


Figura 16 – Viés para o parâmetro β_1 nos modelos logístico, DLomax, PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

4.2 *Misspecification*

Assim como no trabalho de Huayanay (2019), foram gerados dados desbalanceados a partir do modelo Potência Cauchy, com coeficientes de regressão β_0 e β_1 fixos em $\boldsymbol{\beta} = (\beta_0, \beta_1) = (0, 1)$ e a (única) covariável X foi simulada a partir de uma distribuição Uniforme$(-3, 3)$. Foram simulados quatro cenários diferentes com níveis diferentes de desbalanceamento, considerando o parâmetro $\lambda = \{0{,}25; 0{,}5; 2; 4\}$ da distribuição Potência Cauchy. Abaixo está o modelo de regressão binária utilizando o modelo Potência Cauchy:

$$Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),$$
$$p_i = \left(\frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i) + \frac{1}{2}\right)^{\lambda}.$$

Nesse experimento, foram geradas 100 amostras com a distribuição Potência Cauchy nos moldes expressos acima, cada uma com 5.000 observações. Na Tabela 7 é possível observar o grau de desbalanceamento em cada amostra.

Até esse ponto do trabalho, a performance dos modelos está sendo avaliada por meio das métricas WAIC e LOO. Essas métricas foram escolhidas, pois, segundo Yong (2018), elas tendem a performar melhor na seleção de modelos que outras métricas como o DIC, visto que estas levam em conta apenas estimativas pontuais, enquanto as métricas WAIC e LOO levam em conta toda a distribuição *a posteriori* dos parâmetros. Além disto, estas são métricas totalmente Bayesianas.

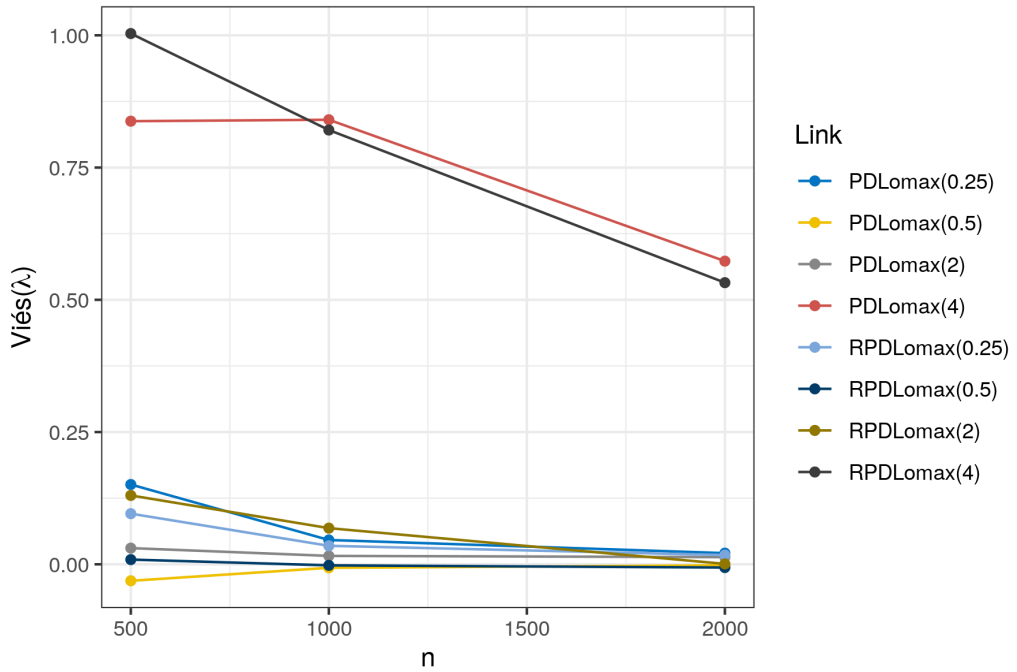


Figura 17 – Viés para o parâmetro λ nos modelos PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

Tabela 7 – Proporção média de 1's, em cada cenário: $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, das amostras geradas a partir da distribuição Potência Cauchy.

	$\lambda = 0{,}25$	$\lambda = 0{,}5$	$\lambda = 2$	$\lambda = 4$
Proporção média de 1's	0,800	0,661	0,329	0,198

A fim de comparar o ajuste dos modelos propostos com o ajuste do modelo logístico, foram observadas as métricas WAIC e LOO. A partir dessas métricas, foram calculadas as médias de LOO e WAIC em cada cenário ($\overline{\text{LOO}}$ e $\overline{\text{WAIC}}$), a porcentagem de vezes em que a métrica de cada *link* é menor que o *link* logístico ($\%_{\text{LOO}}$ e $\%_{\text{WAIC}}$), e a variância de cada uma dessas métricas (s^2_{LOO} e s^2_{WAIC}). Isto é,

$$\overline{\text{LOO}} = \frac{1}{R}\sum_{r=1}^{R} \text{LOO}^{(r)}, \qquad \overline{\text{WAIC}} = \frac{1}{R}\sum_{r=1}^{R} \text{WAIC}^{(r)},$$

$$\%_{\text{LOO}} = \frac{1}{R}\sum_{r=1}^{R} I(\text{LOO}^{(r)} < \text{LOO}^{(r)}_{\text{log.}}), \qquad \%_{\text{WAIC}} = \frac{1}{R}\sum_{r=1}^{R} I(\text{WAIC}^{(r)} < \text{WAIC}^{(r)}_{\text{log.}}),$$

$$s^2_{\text{LOO}} = \frac{1}{R-1}\sum_{r=1}^{R} (\text{LOO}^{(r)} - \overline{\text{LOO}})^2, \qquad s^2_{\text{WAIC}} = \frac{1}{R-1}\sum_{r=1}^{R} (\text{WAIC}^{(r)} - \overline{\text{WAIC}})^2,$$

sendo R o número de réplicas da simulação (nesse caso, $R = 100$).

Na Tabela 8 nota-se que, em todos os casos, pelo menos um dos modelos propostos nesta dissertação (DLomax, PDLomax e RPDLomax) performou melhor que a regressão logística. Quando $\lambda = 0{,}25$, o modelo RPDLomax, mesmo com LOO e WAIC maiores que a regressão logística, ainda consegue superar o ajuste desta em 58% dos casos nas duas métricas. Já quando

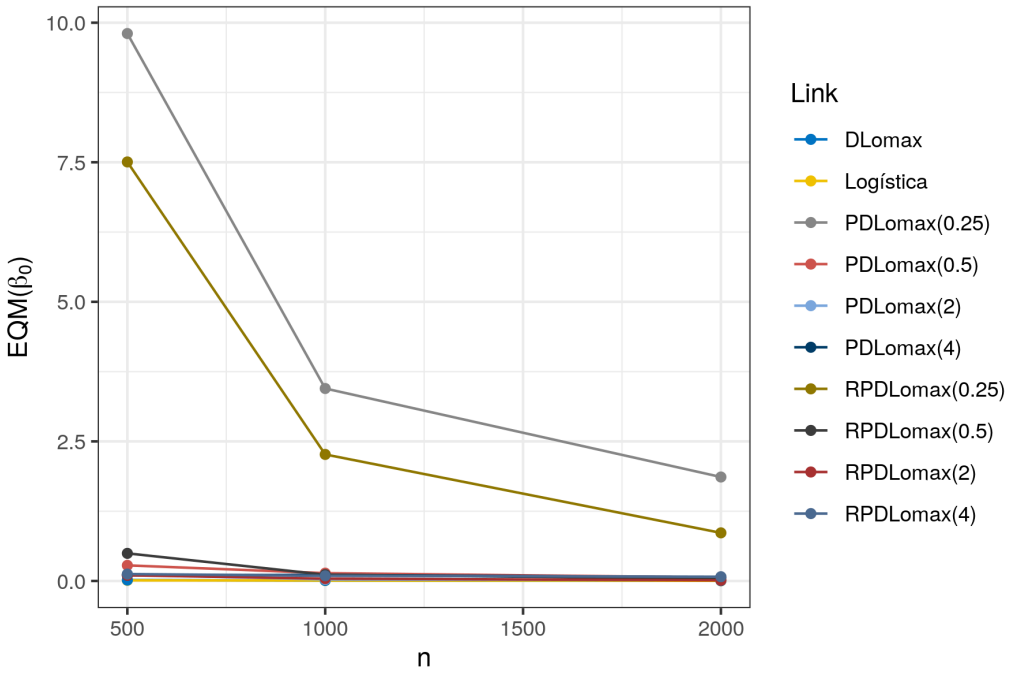


Figura 18 – EQM para o parâmetro β_0 nos modelos logístico, DLomax, PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

$\lambda = 0{,}5$, percebe-se que os modelos PDLomax e RPDLomax possuem menores WAIC e LOO que o modelo logístico e performam melhor em mais de 60% dos casos, sendo o modelo de PDLomax o que melhor performa entre todos (66% de sucesso em relação à regressão logística). Quando $\lambda = 2$, os modelos DLomax, PDLomax e RPDLomax apresentam LOO e WAIC médios menores que a regressão logística e também performam melhor que a regressão logística na maior parte das vezes; nesse caso, os modelos DLomax e PDLomax tiveram melhores resultados (em 60% das vezes tiveram LOO e WAIC menores que o modelo logístico). Por fim, em $\lambda = 4$, o modelo DLomax apresenta WAIC e LOO médios menores que a regressão logística e também 51% de sucesso em relação a esta, enquanto o modelo PDLomax, ainda que com WAIC e LOO maiores que o modelo logístico, possui LOO menor que este em 65% dos casos e em 61% dos casos apresenta menor WAIC.

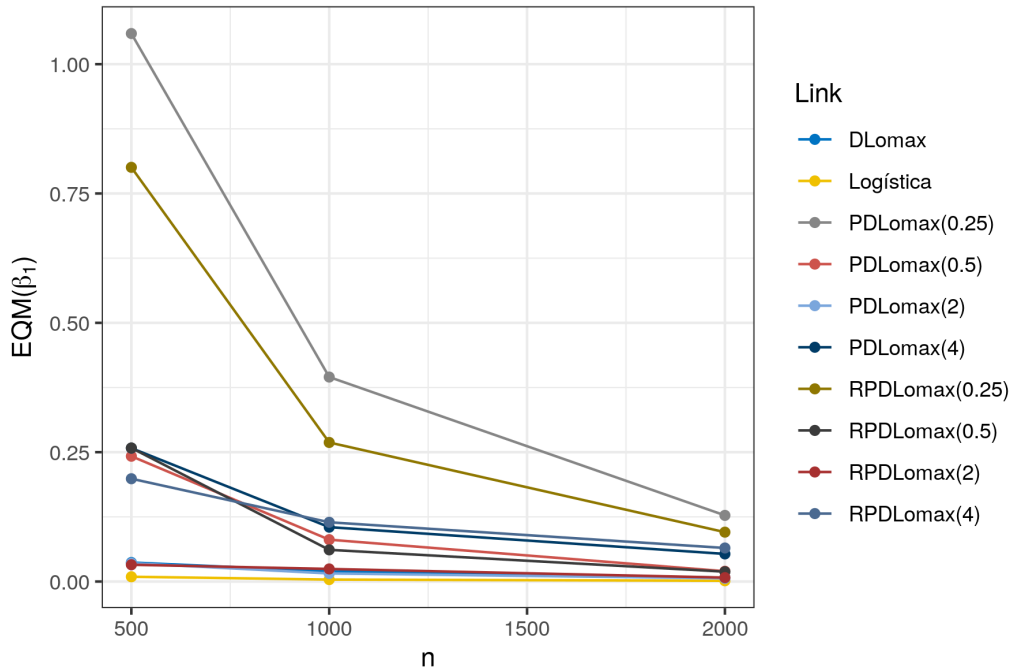


Figura 19 – EQM para o parâmetro β_1 nos modelos logístico, DLomax, PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

Tabela 8 – Medidas de LOO médio ($\overline{\text{LOO}}$), WAIC médio ($\overline{\text{WAIC}}$), variância das medidas LOO e WAIC (s^2_{LOO} e s^2_{WAIC}) e porcentagem de vezes em que cada modelo obteve menores valores de LOO e WAIC em relação à regressão logística ($\%_{\text{LOO}}$ e $\%_{\text{WAIC}}$), para o ajuste dos modelos logístico, DLomax, PDLomax e RPDLomax, nos cenários em que $\lambda = \{0{,}25; 0{,}5; 2; 4\}$.

Link	$\overline{\text{LOO}}$	s^2_{LOO}	$\%_{\text{LOO}}$	$\overline{\text{WAIC}}$	s^2_{WAIC}	$\%_{\text{WAIC}}$
$\lambda = 0{,}25$						
Logística	5.007,371	71,147	-	5.007,343	71,147	-
DLomax	5.007,349	71,169	49	5.007,320	71,169	49
PDLomax	5.008,029	70,471	44	5.008,002	70,471	45
RPDLomax	5.008,471	70,813	58	5.008,448	70,811	58
$\lambda = 0{,}5$						
Logística	6.403,237	39,623	-	6.403,208	39,622	-
DLomax	6.403,268	39,645	40	6.403,239	39,645	41
PDLomax	6.403,053	39,570	66	6.403,025	39,571	66
RPDLomax	6.403,113	39,602	62	6.403,086	39,602	62
$\lambda = 2$						
Logística	6.334,080	38,417	-	6.334,051	38,417	-
DLomax	6.333,970	38,384	60	6.333,942	38,384	60
PDLomax	6.333,958	38,358	60	6.333,931	38,359	60
RPDLomax	6.333,960	38,404	59	6.333,932	38,404	59
$\lambda = 4$						
Logística	4.977,608	65,323	-	4.977,579	65,322	-
DLomax	4.977,568	65,304	51	4.977,540	65,305	52
PDLomax	4.979,453	65,368	65	4.979,433	65,373	62
RPDLomax	4.978,640	65,189	42	4.978,616	65,189	43

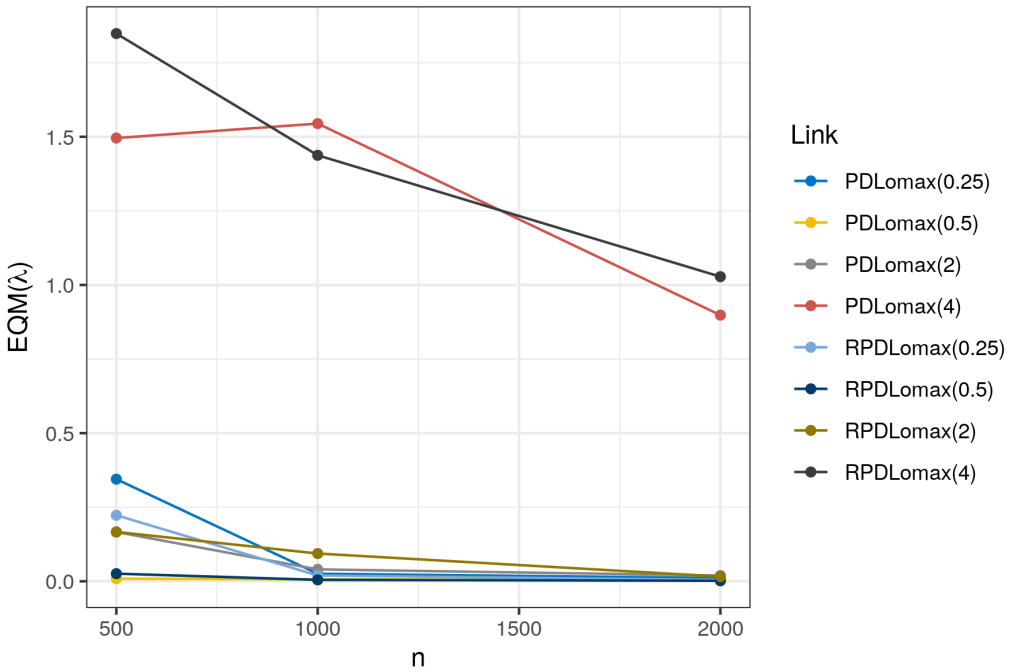


Figura 20 – EQM para o parâmetro λ nos modelos PDLomax e RPDLomax, com valores de $\lambda = \{0{,}25; 0{,}5; 2; 4\}$, para tamanhos de amostra $n = \{500; 1.000; 2.000\}$.

CAPÍTULO

5

APLICAÇÕES

Este capítulo apresenta duas aplicações que foram desenvolvidas a fim de ilustrar o desempenho das funções de ligação DLomax, PDLomax e RPDLomax em dados reais. Primeiramente, foram aplicados os modelos em um banco de dados sobre doação de sangue e, em seguida, estudou-se o efeito dessas funções de ligação em um banco de dados com imagens de árvores doentes.

5.1 Aplicação 1: Doação de Sangue

A primeira aplicação é relacionada à doação de sangue. O banco de dados utilizado foi apresentado e analisado por Yeh, Yang e Ting (2009) e está disponível no repositório da UCI (DUA; GRAFF, 2017). Tal banco contém 748 amostras aleatórias de dados de doadores de sangue do Centro de Transfusão de Sangue da cidade de Hsinchu, em Taiwan, com as seguintes variáveis:

- *Recency*: Número de meses desde a última doação;
- *Frequency*: Número total de doações realizadas pelo doador;
- *Time*: Tempo, em meses, desde a primeira doação;
- *Monetary*: Total, em mililitros (ml), de sangue doado, desde a primeira doação;
- *Y*: Variável binária indicando se o doador doou sangue (1 - sim, 0 - não) em março de 2007.

Na Tabela 9 estão as medidas descritivas de cada uma dessas variáveis. Já na Figura 21 é possível observar o nível de desbalanceamento do banco de dados - apenas 23% de sucessos (1's).

Tabela 9 – Medidas descritivas do banco de dados sobre doação de sangue.

Variável	Média	Mínimo	Máximo	Desvio-Padrão
Recency	9,5	0	74	8,1
Frequency	5,5	1	50	5,8
Time	34,3	2	98	24,4
Monetary	1.378,7	250	12.500	1.459,8

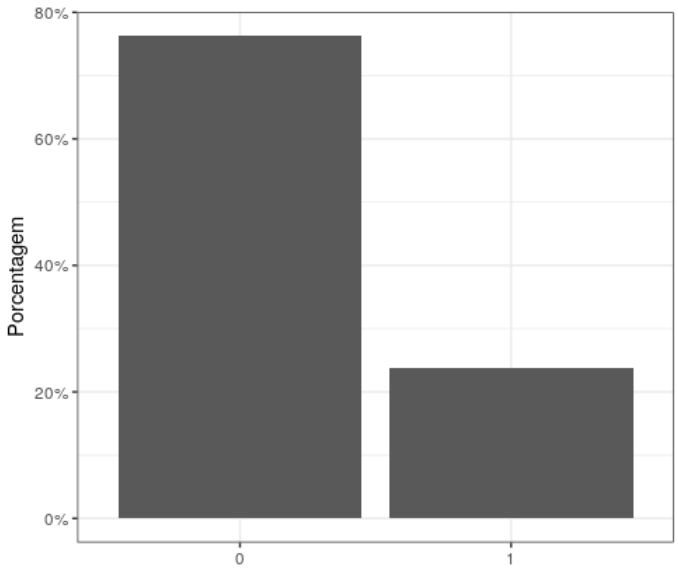


Figura 21 – Porcentagem de sucessos (1's) e fracassos (0's) no banco de dados sobre doação de sangue.

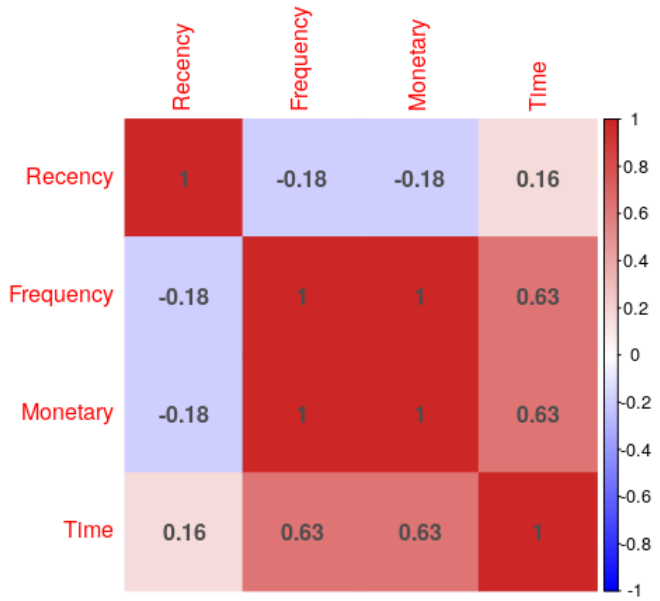


Figura 22 – Correlação entre as variáveis do banco de dados sobre doação de sangue.

Na Figura 22 é possível observar que as variáveis *Frequency* e *Monetary* possuem correlação igual a 1. Isso ocorre, pois, a cada doação são doados 250ml de sangue, logo a

variável *Monetary*, que representa o total de sangue doado, nada mais é que a variável *Frequency* multiplicada por 250. Por isso, optou-se por excluir a variável *Monetary* do modelo. Além da correlação entre essas duas variáveis, não foram encontradas outras correlações que fossem preocupantes para o ajuste dos modelos.

Dessa forma, foram ajustados os modelos considerando as covariáveis (padronizadas) *Recency*, *Frequency* e *Time*, para classificar a variável *Y*. Na estimação dos parâmetros foi utilizado o pacote `stan` do *software* `R`. Para cada distribuição, considerou-se 5.000 iterações, 4 cadeias e 2.500 iterações de *warm up*. Em todas as distribuições, alcançou-se a convergência baseando-se na estatística potencial de redução de escala ($\hat{R}$) de Gelman e Rubin (1992). Na Equação (5.1) a seguir está descrita a estrutura dos modelos ajustados:

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}, \\ p_i &= F(\eta_i), \\ Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i), \end{aligned} \tag{5.1}$$

em que X_1 é a variável *Recency* padronizada, X_2 é a variável *Frequency* também padronizada e X_3 é a variável *Time* padronizada.

Tabela 10 – Métricas de comparação de modelos, aplicadas ao banco de dados sobre doação de sangue.

Modelo	ρ_d	$\bar{D}$	$\hat{D}$	DIC	EAIC	EBIC	LOO	WAIC
DLomax	4,185	708,809	704,624	712,994	716,809	735,279	712,992	712,986
PDLomax	3,565	707,729	704,164	711,294	717,729	740,816	713,167	713,168
RPDLomax	4,998	706,222	701,223	711,220	716,222	739,309	711,785	711,762
Logística	4,079	711,981	707,902	716,060	719,981	738,450	716,549	716,534
cloglog	3,993	715,150	711,157	719,143	723,150	741,619	722,325	720,981
loglog	4,005	715,244	711,238	719,249	723,244	741,713	719,865	719,841
probit	4,005	713,594	709,589	717,599	721,594	740,064	718,288	718,263
cauchit	4,067	708,443	704,376	712,510	716,443	734,913	712,726	712,720

Na Tabela 10 pode-se observar que o modelo RPDLomax obteve o menor valor nas métricas DIC, EAIC, LOO e WAIC, se mostrando o modelo que performou melhor no maior número de métricas. O modelo *cauchit* também obteve performance satisfatória, entretanto, somente performou melhor que o modelo RPDLomax na métrica EBIC. Já os outros modelos propostos neste trabalho, PDLomax e DLomax, apesar de não performarem tão bem quanto o modelo RPDLomax, se mostraram superiores à maioria dos modelos clássicos, visto que apresentaram menor DIC, EAIC, LOO e WAIC que os modelos logístico, cloglog, loglog e *probit*.

5.1.1 Avaliação Preditiva

Também foi realizado um estudo para analisar a performance do modelo em classificar a variável *Y*, a qual indica se o doador doou sangue em março de 2007. Para isso, foram estimados

os parâmetros a partir da média de suas respectivas amostras *a posteriori*, geradas pelo pacote `stan`, utilizando todo o banco de dados. Na Tabela 11 estão os valores de parâmetros utilizados para realizar as predições pontuais de cada modelo; a estimação foi feita a partir da média *a posteriori* baseada nas amostras geradas pelo algoritmo NUTS, apresentado no Capítulo 3.

Tabela 11 – Parâmetros estimados para os modelos em estudo, utilizando a média das amostras *a posteriori* de cada parâmetro, na aplicação ao banco de dados sobre doação de sangue.

Modelo	β_0	β_1	β_2	β_3	λ
DLomax	-2,274	-1,790	1,322	-0,952	-
PDLomax	-1,758	-1,435	1,163	-0,811	1,176
RPDLomax	-1,420	-1,389	1,443	-0,902	0,681
Logístico	-1,448	-0,809	0,808	-0,575	-
probit	-0,840	-0,432	0,453	-0,323	-
cloglog	-1,576	-0,737	0,498	-0,413	-
loglog	-0,426	-0,335	0,485	-0,313	-
cauchit	-1,921	-1,419	1,100	-0,779	-

Na Tabela 12 pode-se ver que o modelo RPDLomax foi o modelo que obteve maior valor de SENS, ou seja, foi o modelo que classificou corretamente a maior porcentagem de doadores de sangue (sucessos). Apresentou também maiores valores de F1 e AUC, mostrando que ele é mais eficiente em distinguir entre as duas classes que os demais modelos. Como o modelo RPDLomax performou melhor nessas métricas, pode-se dizer que ele se sobressai na classificação da classe minoritária, os doadores, classe na qual se tem maior interesse. Note também que os modelos propostos neste trabalho, RPDLomax, PDLomax e DLomax, são os que obtiveram maior medida AUC. O modelo *cauchit*, apesar de, mais uma vez, mostrar medidas próximas aos modelos propostos, não supera o modelo RPDLomax na classificação da classe de interesse. Os modelos PDLomax e RPDLomax, por sua vez, apesar de não superarem o modelo RPDLomax, se mostraram superiores aos modelos logístico, *probit*, cloglog e loglog, tendo alcançando maiores medidas de AUC, acurácia, precisão, F1 e SENS que os modelos citados.

Na Figura 23 é realizada uma análise comparativa das probabilidades de sucesso (1) e fracasso (0), com base nas probabilidades estimadas para cada observação em cada categoria, considerando os modelos DLomax, PDLomax, RPDLomax e logístico. Nota-se que nos modelos propostos (DLomax, PDLomax e RPDLomax), a probabilidade de fracasso está concentrada em um intervalo de valores mais baixos em comparação com o modelo logístico. Isso pode ser observado pela posição das medianas, que estão abaixo da mediana do modelo logístico, e pelas caixas dos *boxplots*, que são menores nesses modelos, indicando que o primeiro e terceiro quartis estão mais concentrados em um intervalo de valores menores. Isso sugere que os modelos propostos tendem a atribuir probabilidades mais baixas para a categoria de fracasso em comparação com o modelo logístico.

Em relação aos sucessos, nota-se que a distribuição das probabilidades calculadas pelo modelo logístico está próxima das probabilidades de fracasso, visto a semelhança das caixas.

Em contraste, nos modelos propostos (DLomax, PDLomax e RPDLomax), as probabilidades de sucesso se estendem por intervalos com valores mais altos, diferenciando-se dos fracassos. Essa distinção entre as distribuições das probabilidades de sucesso e fracasso nos modelos DLomax, PDLomax e RPDLomax indica uma capacidade de discriminação superior em comparação com o modelo logístico. Também é possível observar que os modelos propostos possuem comportamento e potencial discriminatório parecidos.

Tabela 12 – Performance preditiva dos modelos, aplicados ao banco de dados sobre doação de sangue.

		Observado							
Modelo	Predito	0	1	AUC	Acurácia	Precisão	F1	SENS	ESP
DLomax	0	548	135	0,601	0,790	0,662	0,354	0,242	0,961
	1	22	43						
PDLomax	0	545	133	0,604	0,789	0,643	0,363	0,253	0,956
	1	25	45						
RPDLomax	0	538	129	0,610	0,785	0,605	0,378	0,275	0,944
	1	32	49						
Logística	0	554	155	0,551	0,771	0,590	0,212	0,129	0,972
	1	16	23						
probit	0	557	159	0,542	0,770	0,594	0,181	0,107	0,977
	1	13	19						
cloglog	0	561	162	0,537	0,771	0,640	0,158	0,090	0,984
	1	9	16						
loglog	0	556	159	0,541	0,769	0,576	0,180	0,107	0,975
	1	14	19						
cauchit	0	550	136	0,600	0,791	0,677	0,350	0,236	0,965
	1	20	42						

5.1.2 Modelo Adotado

Considerando os critérios de comparação entre os modelos e também a avaliação preditiva, escolheu-se o modelo RPDLomax. Sendo assim, na Tabela 13 estão representadas as medidas descritivas das amostras *a posteriori* dos parâmetros do modelo. Note que todos os parâmetros (coeficientes) β's são significativos, visto que nenhum dos intervalos de credibilidade de 90% para eles inclui o valor zero. Além disso, o intervalo de λ não inclui o valor 1, de modo a indicar que este parâmetro é importante para o ajuste e o modelo atual não pode ser reduzido ao modelo base (DLomax).

Tabela 13 – Medidas descritivas das amostras dos parâmetros do modelo RPDLomax.

Variável	Parâmetro	Média	Desvio-Padrão	Mediana	Quantil 5%	Quantil 95%
Intercepto	β_0	-1,420	0,515	-1,359	-2,297	-0,747
Recency	β_1	-1,389	0,367	-1,350	-2,028	-0,863
Frequency	β_2	1,443	0,410	1,405	0,903	2,084
Time	β_3	-0,902	0,261	-0,882	-1,346	-0,528
Assimetria	λ	0,681	0,161	0,657	0,479	0,960

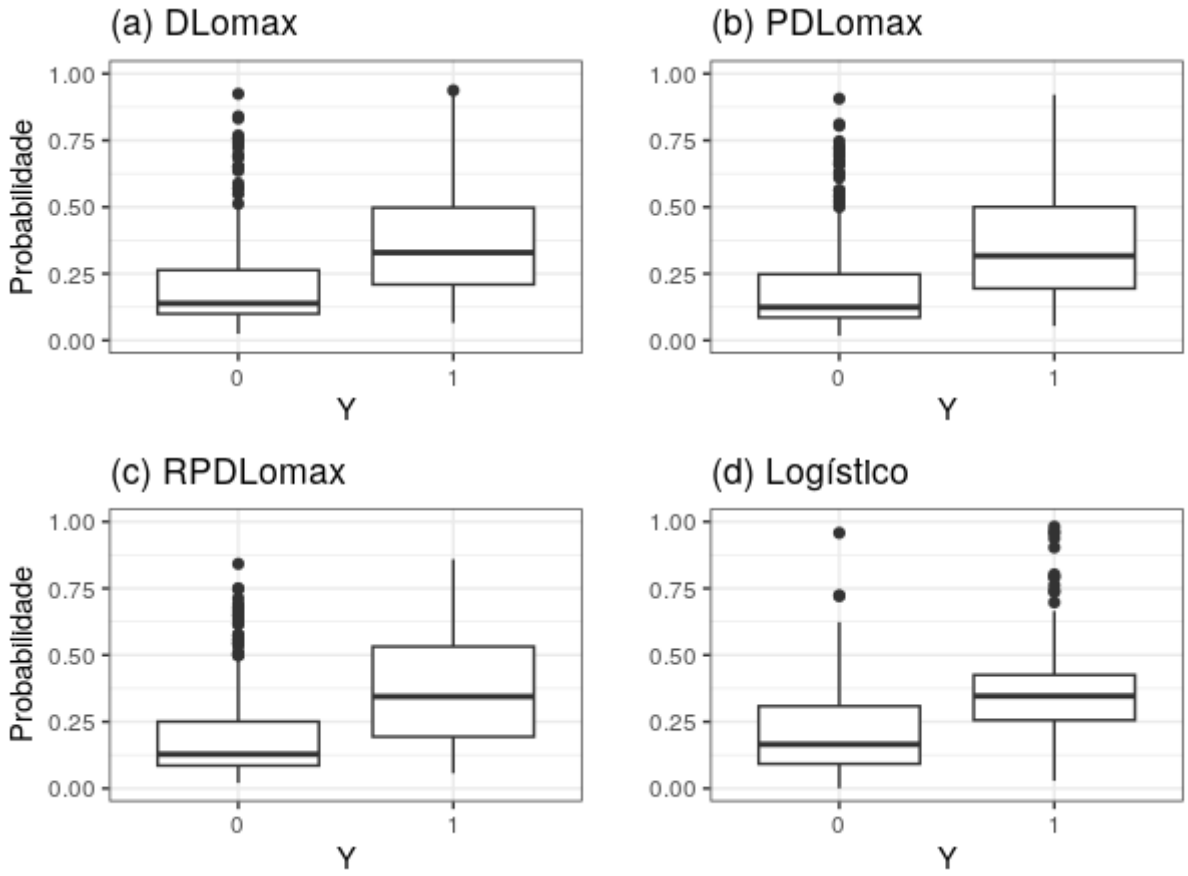


Figura 23 – *Boxplots* das probabilidades estimadas pelos modelos: (a) DLomax, (b) PDLomax, (c) RP-DLomax e (d) logístico, para cada categoria da variável Y, utilizando a média *a posteriori* dos parâmetros.

Escolheu-se utilizar a média como valor pontual dos parâmetros. Desse modo, o modelo adotado pode ser representado da seguinte forma:

$$\begin{aligned}
\eta_i &= -1{,}420 - 1{,}389\,X_{i1} + 1{,}443\,X_{i2} - 0{,}902\,X_{i3}, \\
p_i &= \begin{cases} 1 - \left[1 - \frac{1}{2(1-\eta_i)}\right]^{0{,}681}, & \eta_i \leq 0, \\ 1 - \left[\frac{1}{2(1+\eta_i)}\right]^{0{,}681}, & \eta_i > 0, \end{cases} \\
Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i).
\end{aligned}$$

Observando os sinais dos parâmetros, é possível interpretar que:

- À medida em que se aumenta o número de meses desde a última doação (*Recency*), diminui-se a probabilidade de que o doador doe sangue na data especificada;
- Quando aumenta-se o número de doações feitas pelo doador (*Frequency*), aumenta-se a probabilidade de que este doe no período;
- Quando aumenta-se o tempo, em meses, desde a primeira doação (*Time*), diminui-se a probabilidade de doação em março de 2007.

Tais interpretações fazem sentido, visto que doadores que realizaram sua primeira doação há muito tempo, doaram poucas vezes e também não doam sangue há bastante tempo; são um perfil de doadores esporádicos. Enquanto que doadores com um número elevado de doações representam o perfil de doadores assíduos.

Uma vez que este é um modelo paramétrico e probabilístico, é possível calcular o incremento na probabilidade de sucesso (doação) ao aumentar uma determinada variável, mantendo as demais constantes. A Figura 24 ilustra o efeito da variação de cada variável, considerando as demais em seu valor médio. Nessa imagem, os outros X's são fixados em 0, uma vez que o modelo foi ajustado com as variáveis padronizadas.

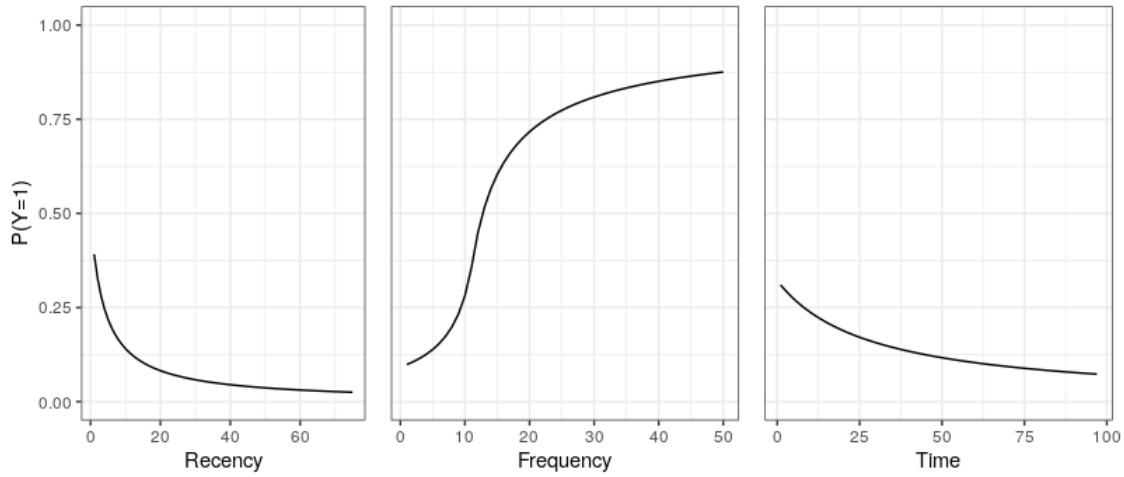


Figura 24 – Efeito de cada variável na probabilidade de que um doador doe sangue na data estipulada, quando as demais variáveis estão constantes em sua média.

Observa-se na Figura 24, por exemplo, que pessoas que doaram sangue 20 vezes ($Frequency = 20$) têm probabilidade de doar sangue próxima a 0,75, enquanto que pessoas que doaram sangue 10 vezes ($Frequency = 10$) apresentam probabilidade próxima a 0,25. Essas informações possuem um valor estratégico significativo para o negócio, pois podem embasar a implementação de ações direcionadas a atrair potenciais doadores.

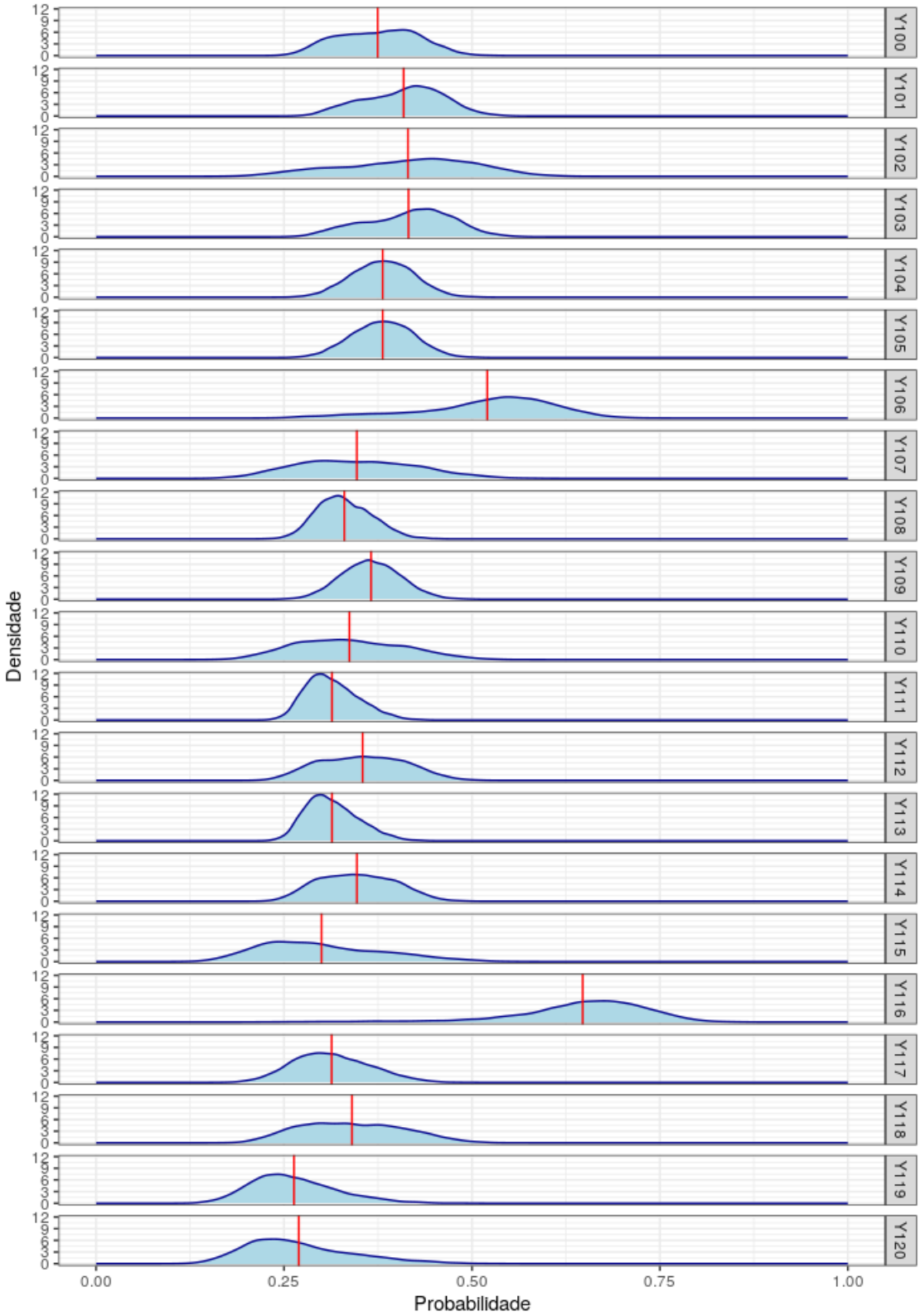


Figura 25 – Densidade das probabilidades estimadas pelo modelo adotado na primeira aplicação, para as observações de números 100 a 120. A linha vermelha representa a média das probabilidades estimadas.

Outro benefício da análise Bayesiana é a capacidade de visualizar a distribuição das probabilidades de sucesso ($Y = 1$) em cada uma das observações. Na Figura 25 pode-se notar

que algumas observações apresentam um maior grau de incerteza, indicado por uma distribuição mais "achatada" que se estende por longos intervalos. Por outro lado, em outras observações, a distribuição de probabilidades *a posteriori* se concentra em intervalos de alta ou baixa probabilidade.

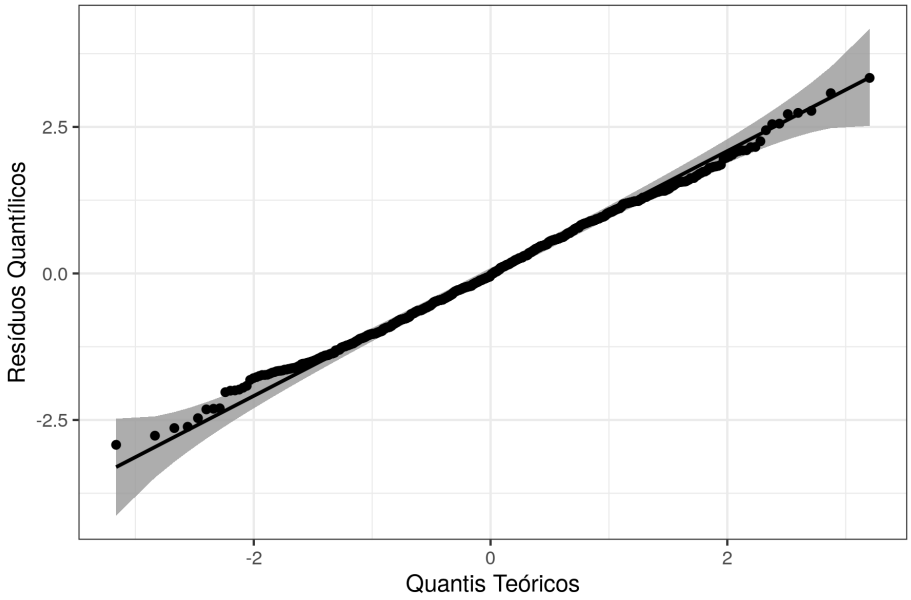


Figura 26 – Gráfico quantil-quantil (ou *QQplot*) dos resíduos quantílicos aleatorizados do modelo RP-DLomax.

Nas Figuras 26, 27 e 28 estão apresentados graficamente os resíduos quantílicos aleatorizados do modelo adotado ajustado. É possível verificar, na Figura 26, que os resíduos cumprem a suposição de normalidade, visto que todos os pontos estão dentro das bandas de confiança; isso pode ser confirmado na Figura 27, visto que o histograma dos resíduos se assemelha à distribuição normal: simétrico e com caudas leves. Além disso, na Figura 28 verifica-se que os resíduos se distribuem aleatoriamente em torno de zero, sem padrões. Logo, não há indícios de que o modelo ajustado não seja adequado aos dados.

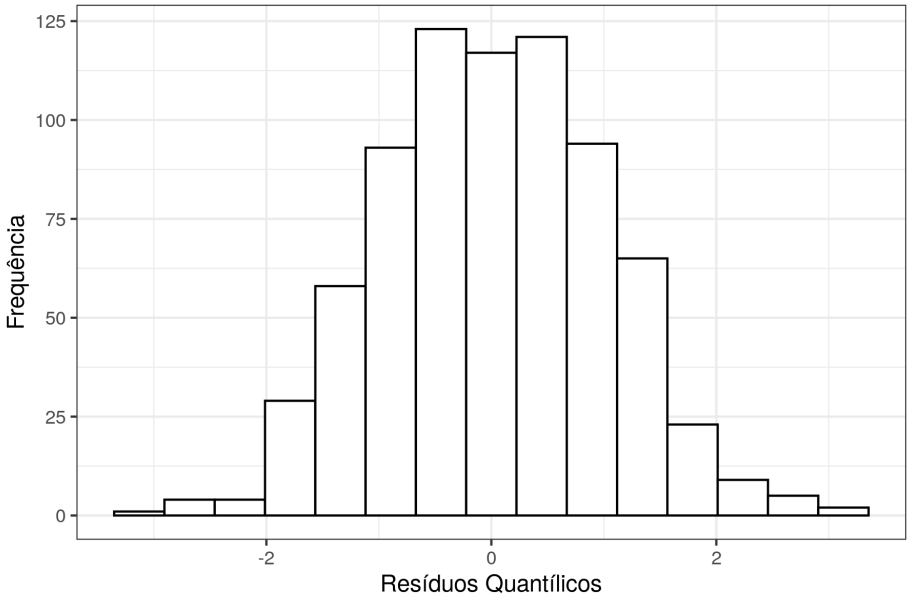


Figura 27 – Histograma dos resíduos quantílicos aleatorizados do modelo RPDLomax.

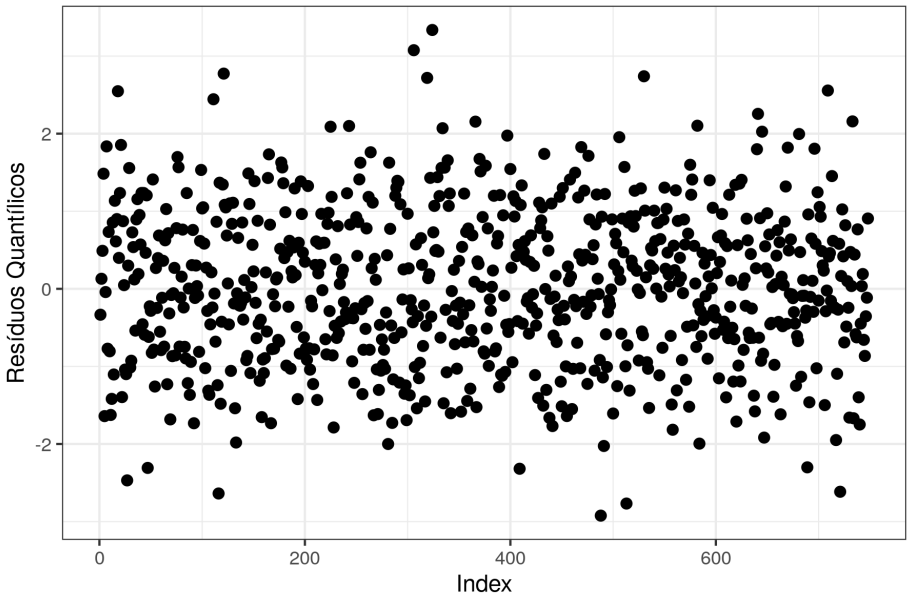


Figura 28 – Resíduos quantílicos aleatorizados do modelo RPDLomax.

5.2 Aplicação 2: Árvores Doentes

A segunda aplicação considera um conjunto de dados formado por segmentos de imagem resultantes da técnica *pansharpening*, descrita no artigo de Johnson, Tateishi e Hoan (2013), para a detecção de pinheiros e carvalhos doentes. Tal banco de dados foi criado, pois, no Japão, besouros que se alimentam de pinheiros e carvalhos são responsáveis pela grande maioria dos danos às áreas florestais, visto que transmitem doenças às árvores, capazes de murchá-las. Dessa forma, é necessário rápida detecção, remoção ou tratamento de árvores doentes, recentemente

infectadas, para evitar que os besouros emerjam no ano seguinte e espalhem suas doenças. A descoloração da folhagem é um sinal claro de infecção, de modo que a detecção de uma árvore doente é geralmente associada à detecção de uma árvore descolorida. Como o número de árvores doentes era muito pequeno em comparação ao número de árvores saudáveis na área de estudo, a coleta de imagens de árvores doentes foi mais difícil e demorada. Como resultado, foi construído um conjunto de dados desbalanceados.

Esse conjunto de dados foi apresentado no artigo de Johnson, Tateishi e Hoan (2013) e está disponível no repositório da UCI (DUA; GRAFF, 2017). Para esta dissertação, foi considerado o banco de dados de validação, o qual conta com 500 observações. Abaixo está uma breve descrição das variáveis presentes na base (dado a complexidade do assunto, mais informações sobre as variáveis podem ser encontradas no artigo de Johnson, Tateishi e Hoan (2013)):

- *GLCM_Pan*: Textura média GLCM;
- *Mean_G*: Valor médio de verde;
- *Mean_R*: Valor médio de vermelho;
- *Mean_NIR*: NIR médio;
- *SD_Pan*: Desvio-padrão;
- *Y*: Variável binária indicando se a árvore está doente (1) ou não (0).

Na Tabela 14 estão as medidas descritivas de cada uma das variáveis. Na Figura 29 é possível observar as proporções de sucessos (árvores doentes) e fracassos (árvores saudáveis); nota-se que há apenas 37,4% de sucessos.

Tabela 14 – Medidas descritivas das variáveis do banco de dados sobre árvores doentes.

Variável	Média	Mínimo	Máximo	Desvio-Padrão
GLCM_Pan	127,07	81,12	167,94	10,67
Mean_G	209,80	117,20	1848,90	78,68
Mean_R	107,74	50,58	1594,58	71,77
Mean_NIR	453,70	144,90	1597,30	156,20
SD_Pan	20,64	5,77	62,39	6,76

Na Figura 30 pode-se observar que há uma correlação muito próxima de 1 (0,98) entre as variáveis *Mean_R* e *Mean_G*. Como essas variáveis apresentam uma relação praticamente linear, optou-se por retirar a variável *Mean_G* dos modelos. Além da correlação entre essas variáveis, não foram encontradas outras correlações que pudessem afetar os modelos.

Assim, foram ajustados os modelos a partir das variáveis (padronizadas) *GLCM_Pan*, *Mean_R*, *Mean_NIR* e *SD_Pan*. Novamente, os modelos foram ajustados usando o pacote `stan`

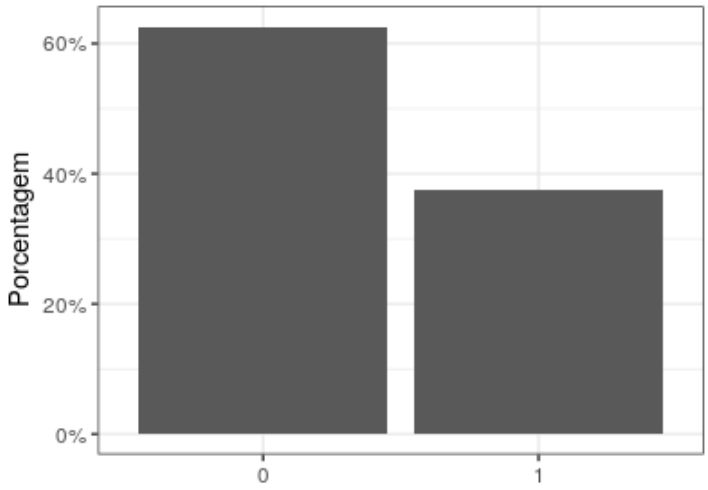


Figura 29 – Porcentagem de sucessos (1's) e fracassos (0's) no banco de dados sobre árvores doentes.

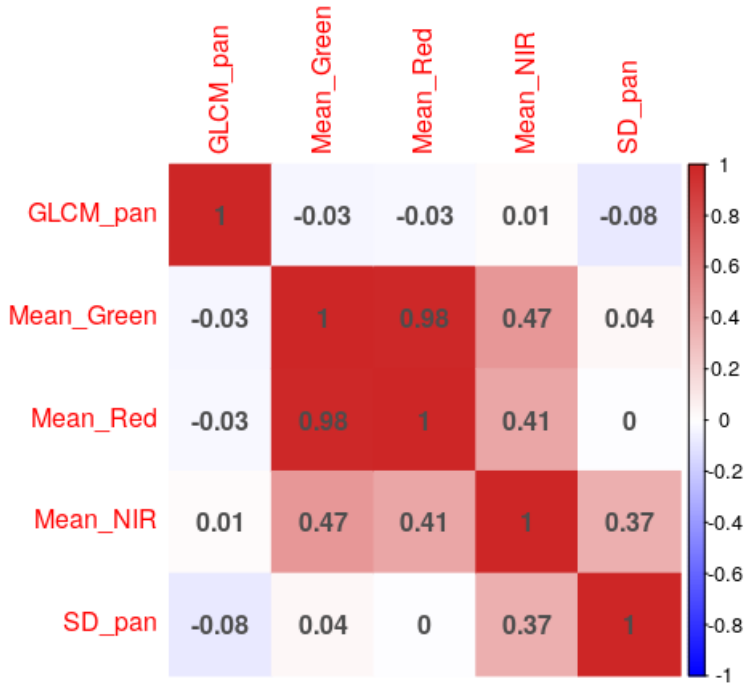


Figura 30 – Correlação entre as variáveis no banco de dados sobre árvores doentes.

do *software* R. Em cada caso, considerou-se 5.000 iterações, 4 cadeias e 2.500 iterações de *warm up*. Em quase todos os casos (com exceção somente da distribuição PDLomax), alcançou-se a convergência baseando-se na estatística potencial de redução de escala ($\hat{R}$) de Gelman e Rubin (1992). Na Equação (5.2) a seguir está descrita a estrutura dos modelos ajustados:

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4}, \\ p_i &= F(\eta_i), \\ Y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i), \end{aligned} \tag{5.2}$$

em que X_1 é a variável *GLCM_Pan* padronizada, X_2 é a variável *Mean_R* também padronizada,

X_3 é a variável *Mean_NIR* padronizada e X_4 é a variável *SD_Pan* padronizada.

Tabela 15 – Métricas de comparação de modelos, aplicadas ao banco de dados sobre árvores doentes.

Modelo	ρ_d	$\bar{D}$	$\hat{D}$	DIC	EAIC	EBIC	LOO	WAIC
DLomax	4,575	532,461	527,887	537,036	542,461	563,534	537,615	537,605
PDLomax	-2.134,930	512,468	2647,397	-1.622,462	524,468	549,755	-	-
RPDLomax	3,733	428,605	424,872	432,337	440,605	465,892	440,882	441,014
Logística	5,051	657,349	652,298	662,401	667,349	688,423	688,867	700,272
cloglog	4,822	662,441	657,620	667,263	672,441	693,514	669,455	667,970
loglog	4,990	635,725	630,734	640,715	645,725	666,798	657,209	666,292
probit	5,032	660,525	655,493	665,558	670,525	691,598	674,604	671,907
cauchit	4,882	543,912	539,030	548,794	553,912	574,985	548,879	548,873

É nítido, na Tabela 15, que o modelo RPDLomax se destaca, quando comparado aos outros modelos. Esse modelo apresentou valores significativamente menores de DIC, EAIC, EBIC, LOO e WAIC, se mostrando, assim, mais adequado ao conjunto de dados que os demais modelos apresentados. O modelo DLomax, por sua vez, foi o segundo modelo com a melhor performance; ele obteve as segundas menores medidas de DIC, EAIC, EBIC, LOO e WAIC. O modelo *cauchit*, apesar de obter medidas próximas ao modelo DLomax, não apresentou nenhuma métrica com melhores resultados que os modelos RPDLomax e DLomax.

5.2.1 Avaliação Preditiva

Para avaliar a performance dos modelos ao classificar imagens de árvores doentes, foi realizado um estudo sobre como os modelos classificaram os dados a partir dos modelos ajustados. Mais uma vez, estimou-se os parâmetros a partir da média das amostras *a posteriori* dos parâmetros, geradas pelo pacote `stan`, utilizando todo o banco de dados. Na Tabela 16 é possível ver os valores de parâmetros utilizados neste estudo.

Tabela 16 – Parâmetros estimados para os modelos em estudo, utilizando a média das amostras *a posteriori* de cada parâmetro, na aplicação ao banco de dados sobre árvores doentes.

Modelo	β_0	β_1	β_2	β_3	β_4	λ
DLomax	0,517	0,242	9,947	-1,633	0,131	-
RPDLomax	25,214	1,157	133,152	-18,888	0,306	0,258
Logística	-0,506	0,038	0,642	-0,146	-0,039	-
cloglog	-0,774	0,031	0,077	-0,078	-0,029	-
loglog	0,032	0,033	1,180	-0,200	-0,024	-
probit	-0,328	0,028	0,126	-0,078	-0,021	-
cauchit	0,399	0,171	8,947	-1,440	0,082	-

Nota-se, na Tabela 17, que o modelo RPDLomax também se destacou na classificação de sucessos e fracassos. Esse modelo obteve maior AUC, acurácia, F1 e SENS. O modelo DLomax, por sua vez, apresentou maior precisão e esteve próximo ao modelo RPDLomax nas demais métricas. Já os modelos logístico, *probit* e cloglog, sequer conseguiram classificar corretamente

uma árvore doente, e os modelos *cauchit* e loglog, apesar de conseguirem classificar corretamente algumas árvores doentes, ainda assim obtiveram desempenho abaixo dos modelos propostos.

Como nesse caso o interesse principal é na classe minoritária, isto é, nas árvores doentes, e a classificação errônea pode ser muito prejudicial, visto que resulta na proliferação de besouros, foi escolhido o modelo com maior SENS: o modelo RPDLomax.

Tabela 17 – Avaliação preditiva dos modelos, aplicados ao banco de dados sobre árvores doentes.

		Observado							
Modelo	Predito	0	1	AUC	Acurácia	Precisão	F1	SENS	ESP
DLomax	0	252	33	0,814	0,812	0,716	0,766	0,824	0,805
	1	61	154						
RPDLomax	0	229	11	0,836	0,810	0,677	0,787	0,941	0,732
	1	84	176						
Logística	0	297	187	0,474	0,594	0,000	0,000	0,000	0,949
	1	16	0						
probit	0	311	187	0,497	0,622	0,000	0,000	0,000	0,994
	1	2	0						
cloglog	0	312	187	0,498	0,624	0,000	0,000	0,000	0,997
	1	1	0						
loglog	0	280	155	0,533	0,624	0,492	0,254	0,171	0,895
	1	33	32						
cauchit	0	252	35	0,809	0,808	0,714	0,760	0,813	0,805
	1	61	152						

Na Figura 31, é evidente que os modelos DLomax e RPDLomax não apenas apresentam métricas superiores em relação ao modelo logístico, mas também são mais eficazes na distinção entre as categorias por meio das probabilidades estimadas. Observa-se que as caixas dos *boxplots* do modelo logístico são semelhantes e possuem medianas com valores relativamente próximos, indicando uma menor capacidade de discriminação entre as categorias, quando comparado aos demais modelos.

É importante também observar as diferenças entre os modelos DLomax e RPDLomax. No modelo DLomax, as probabilidades de fracasso (0) se concentram em um intervalo menor e apresentam uma série de *outliers* em probabilidades altas. Por outro lado, o modelo RPDLomax apresenta uma mediana próxima de 0 para as probabilidades de fracasso, no entanto, sua caixa correspondente aos fracassos é mais extensa, indicando que há uma distância considerável entre a mediana e o terceiro quartil. Esses fatores podem indicar que o modelo DLomax pode ser mais eficiente na classificação de fracassos. Em relação aos sucessos (1), o modelo RPDLomax parece ser o melhor, pois possui o menor número de sucessos com probabilidade baixa (note o número de *outliers* em valores menores que o modelo possui).

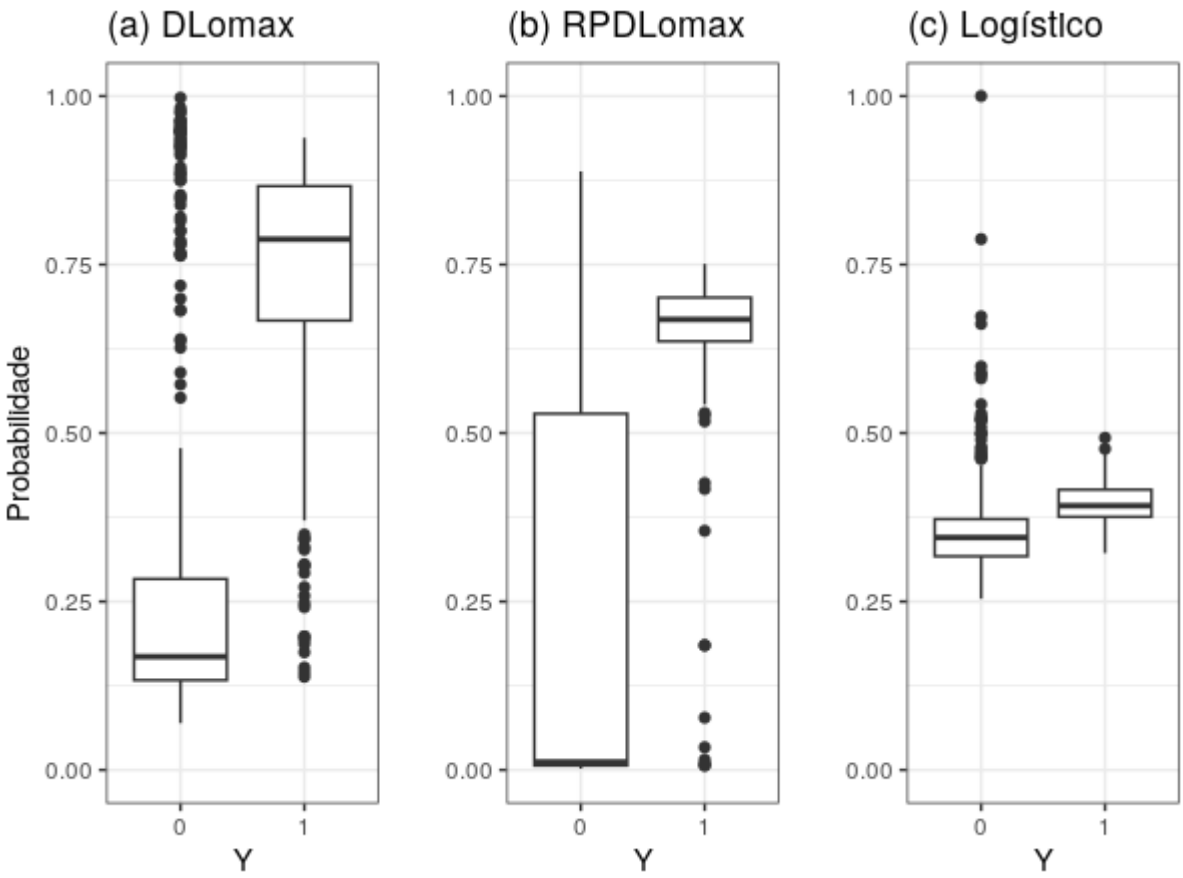


Figura 31 – *Boxplots* das probabilidades estimadas pelos modelos: (a) DLomax, (b) RPDLomax e (c) logístico, para cada categoria da variável Y, utilizando a média *a posteriori* dos parâmetros.

5.2.2 Modelo Adotado

Como o modelo escolhido foi o RPDLomax, na Tabela 18 pode-se ver as medidas descritivas das amostras das distribuições *a posteriori* deste modelo. Nesta tabela pode-se ver que o intervalo de credibilidade para o parâmetro de assimetria λ não engloba o valor 1, de modo a sugerir que o modelo atual não pode ser reduzido ao modelo base (DLomax). Além disso, nota-se que o intervalo de credibilidade de 90% dos parâmetros *SD_Pan* e *GLCM_Pan* englobam o valor 0; entretanto, mesmo não sendo significativas, essas covariáveis são importantes para que os resíduos do modelo atendam às suposições de normalidade, como pode-se verificar no Apêndice A.

Utilizando a média como estimativa pontual dos parâmetros, o modelo adotado pode ser representado pela fórmula:

$$\eta_i = 25{,}214 + 1{,}157\,X_{i1} + 133{,}152\,X_{i2} - 18{,}888\,X_{i3} + 0{,}306\,X_{i4},$$

$$p_i = \begin{cases} 1 - \left[1 - \frac{1}{2(1-\eta_i)}\right]^{0{,}258}, & \eta_i \leq 0, \\ 1 - \left[\frac{1}{2(1+\eta_i)}\right]^{0{,}258}, & \eta_i > 0, \end{cases}$$

$$Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i).$$

Tabela 18 – Medidas descritivas dos parâmetros do modelo RPDLomax, ajustado aos dados sobre árvores doentes.

Variável	Parâmetro	Média	Desvio-Padrão	Mediana	Quantil 5%	Quantil 95%
Intercepto	β_0	25,214	10,865	23,789	10,376	45,169
GLCM_Pan	β_1	1,157	1,726	1,051	-1,417	4,115
Mean_R	β_2	133,152	50,830	126,526	63,411	226,415
Mean_NIR	β_3	-18,888	6,993	-17,882	-31,525	-9,332
SD_Pan	β_4	0,306	2,089	0,393	-3,198	3,534
Assimetria	λ	0,258	0,038	0,253	0,205	0,328

Observando os sinais dos parâmetros, é possível interpretar que:

- À medida em que se aumenta as variáveis *GLCM_Pan*, *Mean_R* e *SD_Pan*, aumenta-se a probabilidade de que a árvore esteja doente;
- À medida em que aumenta-se a variável *Mean_NIR*, diminui-se a probabilidade de que a árvore esteja doente;
- A variável *Mean_R* desempenha um forte papel no cálculo da probabilidade de que a árvore esteja doente, dado a magnitude de seu parâmetro associado. Isso faz sentido, pois árvores secas perdem seu verde e dão espaço a tonalidades mais avermelhadas.

Além de interpretar os sinais dos parâmetros, é possível observar o impacto da variação de cada variável na probabilidade de uma árvore estar doente. Na Figura 32, é apresentado o efeito da variação de cada variável, mantendo as demais em seu valor médio, ou seja, os demais X's são fixados em 0, uma vez que o modelo foi ajustado com as variáveis padronizadas.

Na Figura 32, é possível observar o impacto significativo da variável *Mean_R*, em que ocorre um aumento brusco na probabilidade de sucesso (árvore doente) no intervalo entre 90 e 100. Quando *Mean_R* é igual a 90, a probabilidade de sucesso é de 0,015, enquanto para *Mean_R* igual a 100, a probabilidade de sucesso é de 0,558. Por outro lado, as demais variáveis parecem ter uma influência menor na probabilidade de sucesso, uma vez que suas curvas apresentam pouca variação.

Além das interpretações fornecidas acima, a Estatística Bayesiana também permite interpretar as probabilidades de sucesso ($Y = 1$) de cada uma das observações. Na Figura 33 nota-se que algumas observações possuem uma probabilidade muito baixa de sucesso, tendo suas distribuições concentradas em um intervalo próximo a 0, enquanto outras observações, por sua vez, se concentram em pontos mais próximos ao centro, ou apresentam uma distribuição mais "achatada". Desse modo, é possível identificar observações que possuem um maior grau de incerteza em sua classificação.

Os resíduos desse modelo estão apresentados graficamente nas Figuras 34, 35 e 36. Na Figura 34 é possível observar que os resíduos se comportam dentro das bandas do *QQplot*, de

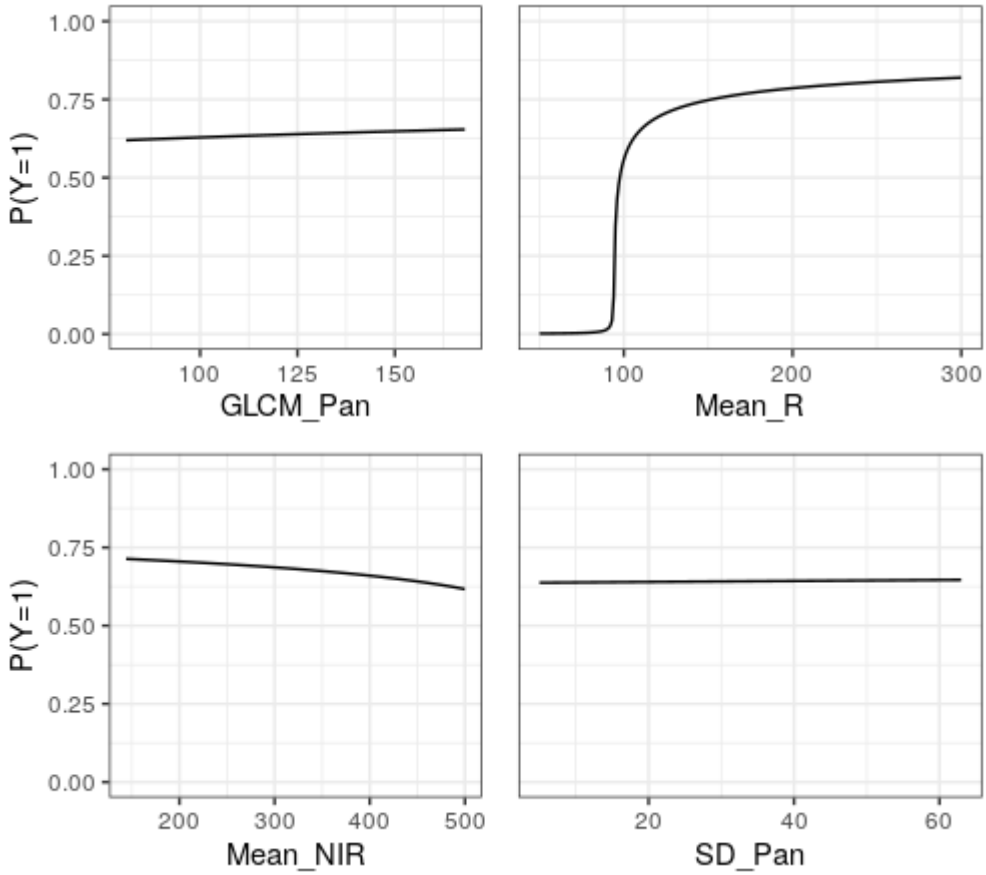


Figura 32 – Efeito de cada variável na probabilidade de que uma árvore esteja doente, quando as demais variáveis estão constantes em sua média.

modo a não apresentar indícios de que eles não sigam uma distribuição normal. Além disso, na Figura 35 nota-se que a distribuição dos resíduos aparenta ser, em geral, simétrica e apresentar caudas leves; alguns pontos se encontram afastados, mas não chegam a ser discrepantes. Por fim, na Figura 36, os pontos parecem se distribuir aleatoriamente, sem fortes indícios de tendência ou alterações na variância.

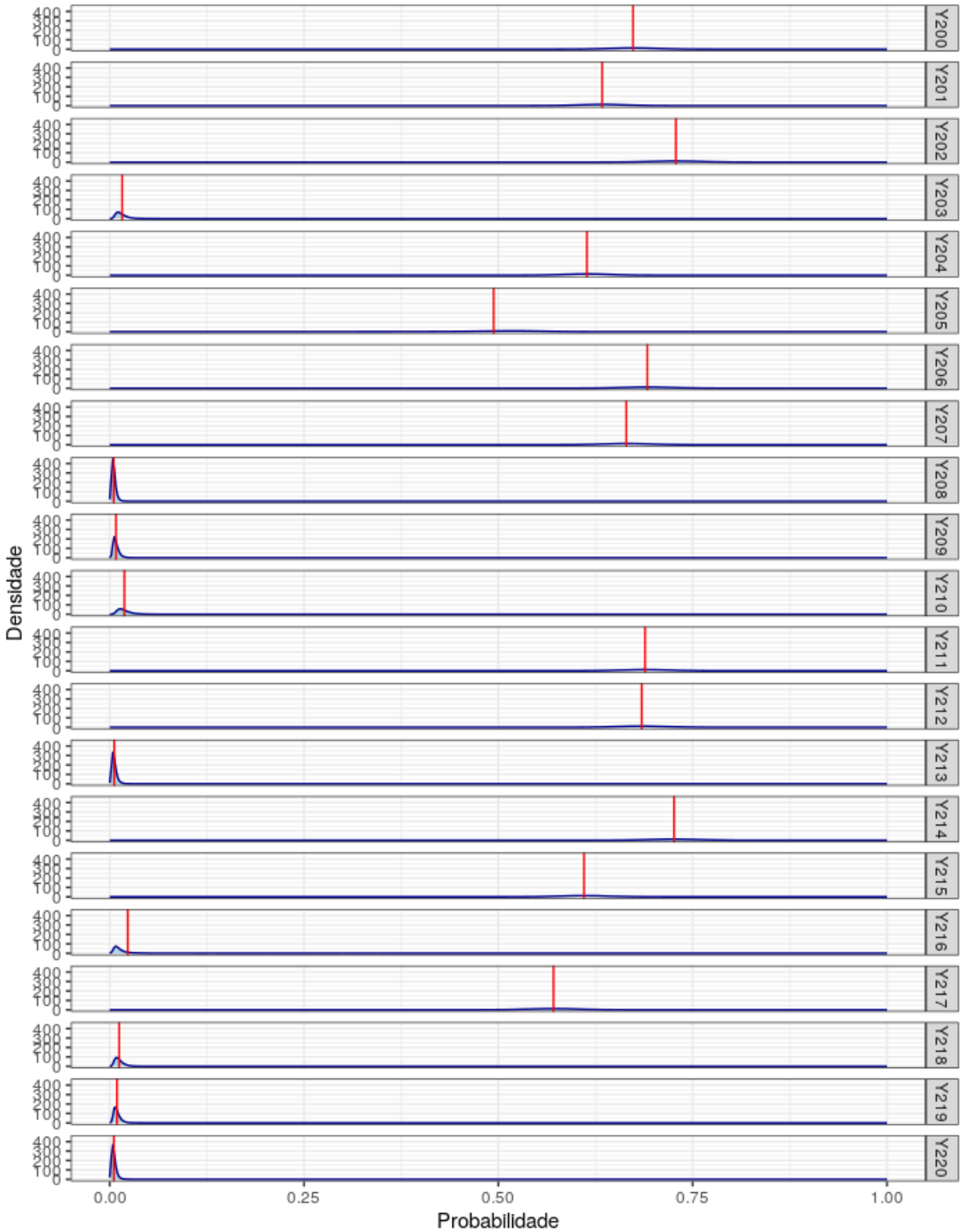


Figura 33 – Densidade das probabilidades estimadas pelo modelo adotado na segunda aplicação, para as observações de números 200 a 220. A linha vermelha representa a média das probabilidades estimadas.

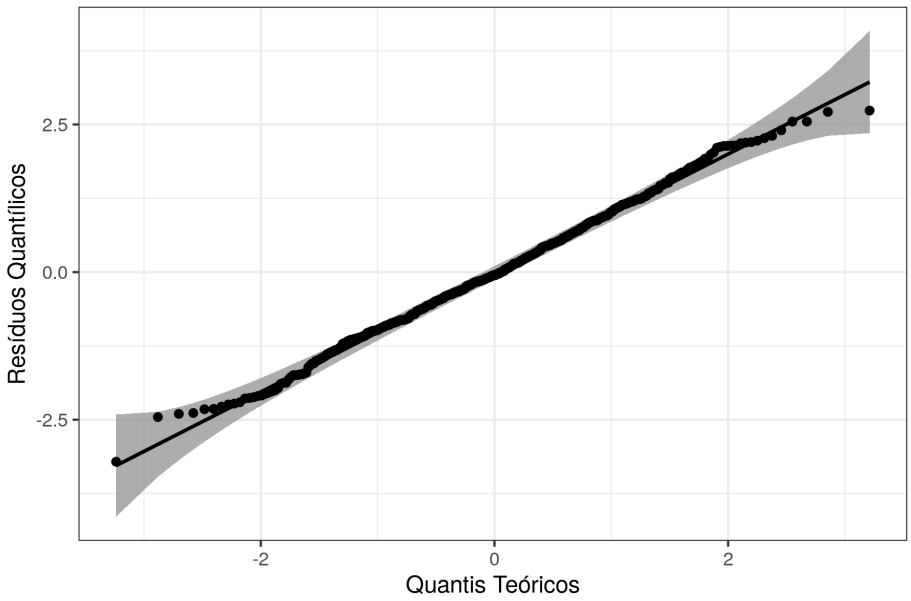


Figura 34 – *QQplot* dos resíduos quantílicos aleatorizados do modelo RPDLomax.

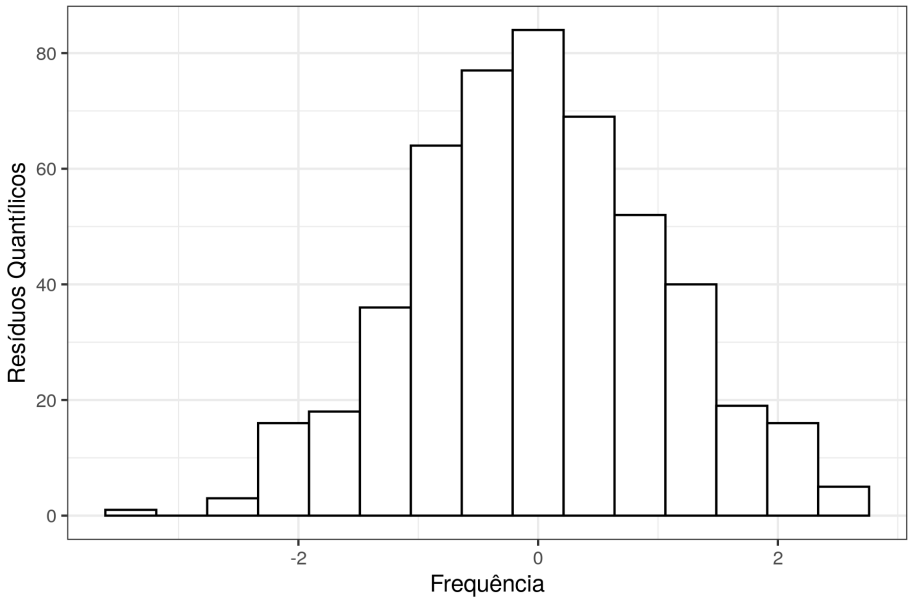


Figura 35 – Histograma dos resíduos quantílicos aleatorizados do modelo RPDLomax.

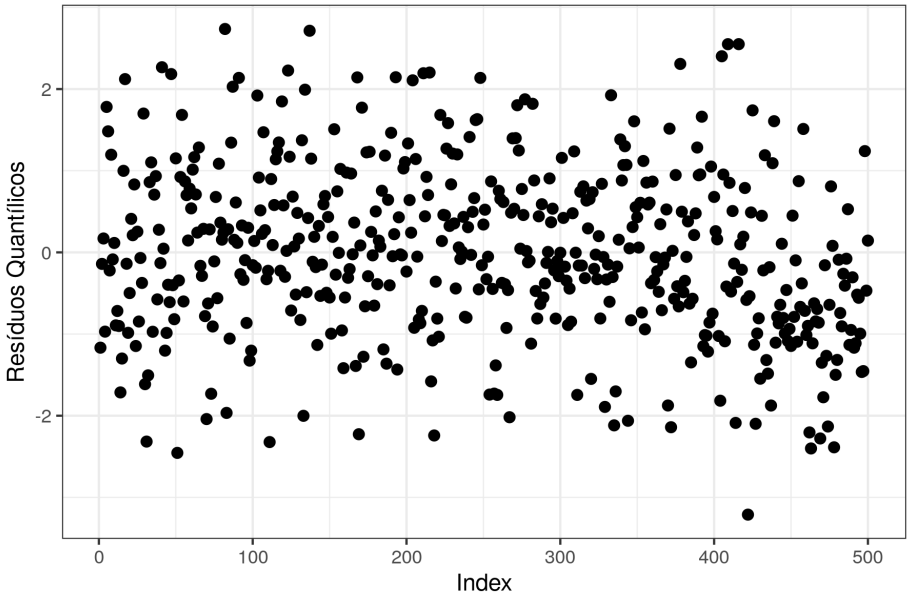


Figura 36 – Resíduos quantílicos aleatorizados do modelo RPDLomax.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS

Nesse trabalho foram apresentadas novas abordagens para a classificação de dados binários desbalanceados, a partir da introdução de novas funções de ligação à regressão binária. Essas novas funções de ligação foram criadas aplicando-se a transformação proposta por Bazán, Romeo e Rodrigues (2014) na distribuição Double Lomax (DLomax), de modo a gerar as distribuições Potência Double Lomax (PDLomax) e Reversa de Potência Double Lomax (RPDLomax). Apesar de possuírem diversas aplicações, a distribuição Lomax e suas extensões e generalizações ainda não haviam sido testadas em contextos de regressão binária; por isso, pode-se considerar este trabalho um estudo inédito na literatura.

Ao longo das seções deste trabalho foram discutidos os comportamentos dessas distribuições em diferentes cenários. Foi feita uma breve revisão sobre os modelos de regressão binária e as transformações Potência e Reversa de Potência. Além disso, foi apresentado um estudo sobre o impacto do parâmetro λ na assimetria dessas distribuições e também no número de observações classificadas como sucessos ou fracassos. Também foi apresentada uma breve revisão sobre modelos Bayesianos e a estimação de parâmetros nesses casos.

Nas seções seguintes foram conduzidos estudos de simulação para avaliar a capacidade dos modelos de recuperar parâmetros, bem como o ajuste dos modelos em casos de *misspecification*. Todas as simulações foram implementadas no *software* `R` utilizando o pacote `stan`. O estudo de recuperação de parâmetros indicou que os modelos propostos são eficientes na recuperação de parâmetros e exibem redução no viés das estimações à medida em que aumenta-se o tamanho das amostras. Já no estudo de *misspecification*, foi possível observar que, tanto em casos de assimetria moderada como em casos de assimetria severa, os modelos propostos conseguem superar o ajuste da regressão logística, em termos de qualidade de ajuste avaliada pelas métricas LOO e WAIC, nos cenários estabelecidos neste trabalho. Cabe ressaltar que, em simulações futuras, é recomendado um número maior de réplicas e iterações.

Os modelos propostos também foram aplicados a dois conjuntos de dados reais des-

balanceados. O primeiro banco de dados refere-se à classificação de possíveis doadores de sangue, e o segundo banco de dados, à classificação de seguimentos de imagens para identificar árvores doentes. Em ambos os bancos de dados, o modelo RPDLomax performou melhor que as funções de ligação convencionais: *logit*, *probit*, *cauchit*, loglog e cloglog, apresentando as menores métricas de ajuste (WAIC, LOO, DIC, EAIC e EBIC). O modelo PDLomax, apesar de ter apresentado bons resultados na classificação do primeiro banco de dados, não convergiu na segunda aplicação; por isso, sugerimos que em estudos futuros sejam testadas *prioris* menos restritivas.

No estudo preditivo, o modelo RPDLomax foi o modelo que classificou corretamente a maior proporção de observações dentro da classe minoritária (sensibilidade) nas duas aplicações. Além disso, os parâmetros estimados por esse modelo se mostraram coerentes com o contexto de cada um dos bancos de dados. Os resíduos quantílicos aleatorizados desse modelo também se comportaram de forma adequada nas duas aplicações. Cabe ressaltar que o estudo preditivo dos modelos foi feito a partir de todo o banco de dados, devido ao baixo número de amostras, à natureza desbalanceada dos dados e também ao baixo número de parâmetros nos modelos propostos (o que os torna menos suscetíveis a *overfitting*). Em outros bancos de dados, recomenda-se a divisão em dados de treino e teste ou validação cruzada, para avaliar o potencial preditivo dos modelos.

Em estudos futuros, pode-se buscar novas aplicações aos modelos apresentados, em especial, em bancos de dados maiores. Além disso, as novas funções de ligação assimétricas podem ser utilizadas como função de ativação em redes neurais, visto que Gomes e Ludermir (2013) mostram, em seu trabalho, que funções de ativação assimétricas podem melhorar a predição de séries temporais com redes neurais. Também pode-se comparar a performance dos modelos apresentados com outros *links* assimétricos e outras distribuições Potência.

Todos os códigos deste trabalho podem ser encontrados no GitHub da autora: Dissertação-Códigos.

REFERÊNCIAS

ALVES, J. S. B.; BAZÁN, J. L.; ARELLANO-VALLE, R. B. Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. **Biometrical Journal**, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202100325>. Citado nas páginas 24 e 54.

ANYOSA, S. A. C. **Regressão binária usando ligações potência e reversa de potência**. Dissertação (Mestrado) — UFSCAR-USP, São Carlos, 2017. Citado nas páginas 25, 32, 33, 37 e 48.

BAZÁN, J.; TORRES-AVILÉS, F.; SUZUKI, A.; LOUZADA, F. Power and reversal power links for binary regressions: An application for motor insurance policyholders: J.l. bazÁn et al. **Applied Stochastic Models in Business and Industry**, v. 33, 11 2016. Citado nas páginas 25 e 44.

BAZÁN, J. L.; ROMEO, J. S.; RODRIGUES, J. Bayesian skew-probit regression for binary response data. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 28, n. 4, p. 467 – 482, 2014. Citado nas páginas 25, 30, 44 e 83.

BERMÚDEZ, L.; PÉREZ, J. M.; AYUSO, M.; GÓMEZ, E.; VÁZQUEZ, F. J. A bayesian dichotomous model with asymmetric link for fraud in insurance. **Insurance: Mathematics and Economics**, v. 42, n. 2, p. 779–786, 2008. Citado na página 25.

BINDU, P.; SANGITA, K. Double lomax distribution and its applications. **Statistica**, v. 75, n. 3, p. 331–342, Jan. 2015. Disponível em: <https://rivista-statistica.unibo.it/article/view/5190>. Citado na página 30.

BRYSON, M. C. Heavy-tailed distributions: Properties and tests. **Technometrics**, Taylor Francis, v. 16, n. 1, p. 61–68, 1974. Citado na página 30.

CALABRESE, R.; OSMETTI, S. A. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. **Journal of Applied Statistics**, Taylor Francis, v. 40, n. 6, p. 1172–1188, 2013. Citado na página 24.

CHEN, M.; DEY, D. K.; SHAO, Q. A new skewed link model for dichotomous quantal response data. **Journal of the American Statistical Association**, [American Statistical Association, Taylor Francis, Ltd.], v. 94, n. 448, p. 1172–1186, 1999. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2669933>. Citado nas páginas 25, 27 e 28.

____. Bayesian analysis of binary data using skewed logit models. **Calcutta Statistical Association Bulletin**, v. 51, n. 1-2, p. 11–30, 2001. Citado na página 25.

CZADO, C.; SANTNER, T. J. The effect of link misspecification on binary regression inference. **Journal of Statistical Planning and Inference**, v. 33, n. 2, p. 213–231, 1992. ISSN 0378-3758. Disponível em: <https://www.sciencedirect.com/science/article/pii/0378375892900695>. Citado na página 24.

DÁVILA-CÁRDENES, N.; PÉREZ-SÁNCHEZ, J.; GÓMEZ-DÉNIZ, E.; BOZA-CHIRINO, J. Skewed binary regression to study rental cars by tourists in the canary islands. **Journal of Risk and Financial Management**, v. 14, n. 11, 2021. ISSN 1911-8074. Citado na página 25.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml",institution="UniversityofCalifornia,Irvine, SchoolofInformationandComputerSciences>. Citado nas páginas 63 e 73.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, [American Statistical Association, Taylor Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America], v. 5, n. 3, p. 236–244, 1996. ISSN 10618600. Disponível em: <http://www.jstor.org/stable/1390802>. Citado na página 51.

____. **Generalized Linear Models With Examples in R**. Springer New York, 2018. (Springer Texts in Statistics). ISBN 9781441901187. Disponível em: <https://books.google.com.br/books?id=tBh5DwAAQBAJ>. Citado na página 52.

GELMAN, A.; RUBIN, D. B. Inference from Iterative Simulation Using Multiple Sequences. **Statistical Science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457 – 472, 1992. Disponível em: <https://doi.org/10.1214/ss/1177011136>. Citado nas páginas 65 e 74.

GOLET, I. Symmetric and asymmetric binary choice models for corporate bankruptcy. **Procedia - Social and Behavioral Sciences**, v. 124, 03 2014. Citado na página 25.

GOMES, G. S. S.; LUDERMIR, T. B. Optimization of the weights and asymmetric activation function family of neural network for time series forecasting. **Expert Systems with Applications**, v. 40, n. 16, p. 6438–6446, 2013. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417413003515>. Citado na página 84.

HAIBO, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. Citado nas páginas 23 e 24.

HINKLEY, D. V. On power transformations to symmetry. **Biometrika**, [Oxford University Press, Biometrika Trust], v. 62, n. 1, p. 101–111, 1975. ISSN 00063444. Disponível em: <http://www.jstor.org/stable/2334491>. Citado na página 38.

HOMAN, D. M.; GELMAN, A. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. **J. Mach. Learn. Res.**, JMLR.org, v. 15, n. 1, p. 1593–1623, jan 2014. ISSN 1532-4435. Citado nas páginas 45 e 48.

HUAYANAY, A. C. **Modelos de regressão para resposta binária na presença de dados desbalanceados**. Dissertação (Mestrado) — UFSCAR-USP, São Carlos, 2019. Citado nas páginas 25, 37, 44, 51, 56 e 58.

HUAYANAY, A. C.; BAZÁN, J. L.; CANCHO, V. G.; DEY, D. K. Performance of asymmetric links and correction methods for imbalanced data in binary regression. **Journal of Statistical Computation and Simulation**, Taylor Francis, v. 89, n. 9, p. 1694–1714, 2019. Citado nas páginas 25 e 51.

JOHNSON, B. A.; TATEISHI, R.; HOAN, N. T. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. **International Journal of Remote Sensing**, Taylor Francis, v. 34, n. 20, p. 6969–6982, 2013. Disponível em: <https://doi.org/10.1080/01431161.2013.810825>. Citado nas páginas 72 e 73.

KAUR, H.; PANNU, H. S.; MALHI, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 52, n. 4, aug 2019. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3343440>. Citado na página 27.

LEMONTE, A. J.; BAZÁN, J. L. New links for binary regression: an application to coca cultivation in Peru. **TEST: An Official Journal of the Spanish Society of Statistics and Operations Research**, v. 27, n. 3, p. 597–617, 09 2018. Citado na página 25.

MOORS, J. J. A.; WAGEMAKERS, R. T. A.; COENEN, V. M. J.; HEUTS, R. M. J.; JANSSENS, M. J. B. T. Characterizing systems of distributions by quantile measures. **Statistica Neerlandica**, v. 50, n. 3, p. 417–430, 1996. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1996.tb01507.x>. Citado na página 38.

NAGLER, J. Scobit: An alternative estimator to logit and probit. **American Journal of Political Science**, [Midwest Political Science Association, Wiley], v. 38, n. 1, p. 230–255, 1994. ISSN 00925853, 15405907. Disponível em: <http://www.jstor.org/stable/2111343>. Citado na página 29.

NARANJO, L.; PÉREZ, C.; MARTÍN, J.; CALLE-ALONSO, F. A new asymmetric link-based binary regression model to detect parkinson's disease by using replicated voice recordings. In: . [S.l.: s.n.], 2018. p. 1182–1186. Citado na página 24.

PÉREZ-RODRÍGUEZ, J.; PEREZ-SÁNCHEZ, J. M.; GOMEZ-DENIZ, E. Modelling the asymmetric probabilistic delay of aircraft arrival. **Journal of Air Transport Management**, v. 62, p. 90–98, 07 2017. Citado na página 25.

PÉREZ-SÁNCHEZ, J.; NEGRÍN-HERNÁNDEZ, M.; GARCÍA-GARCÍA, C.; GÓMEZ-DÉNIZ, E. Bayesian asymmetric logit model for detecting risk factors in motor ratemaking. **ASTIN Bulletin**, Cambridge University Press, v. 44, n. 2, p. 445–457, 2014. Citado na página 25.

PÉREZ-SÁNCHEZ, J. M.; SALMERÓN-GÓMEZ, R.; OCANA-PEINADO, F. M. A bayesian asymmetric logistic model of factors underlying team success in top-level basketball in spain. **Statistica Neerlandica**, v. 73, n. 1, p. 22–43, 2019. Citado na página 25.

PRASETYO, R. B.; KUSWANTO, H.; IRIAWAN, N.; ULAMA, B. S. S. A comparison of some link functions for binomial regression models with application to school drop-out rates in east java. **AIP Conference Proceedings**, v. 2194, n. 1, p. 020083, 2019. Disponível em: <https://aip.scitation.org/doi/abs/10.1063/1.5139815>. Citado na página 24.

____. Binomial regression models with a flexible generalized logit link function. **Symmetry**, v. 12, n. 2, 2020. ISSN 2073-8994. Citado na página 25.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado na página 28.

RADY, E. H.; HASSANEIN, W.; ELHADDAD, T. The power lomax distribution with an application to bladder cancer data. **SpringerPlus**, v. 5, 12 2016. Citado na página 25.

Stan Development Team. **RStan: the R interface to Stan**. 2022. R package version 2.21.7. Disponível em: <https://mc-stan.org/>. Citado na página 45.

STUKEL, T. A. Generalized logistic models. **Journal of the American Statistical Association**, Taylor Francis, v. 83, n. 402, p. 426–431, 1988. Citado na página 25.

TAYLOR, J. M. G. The cost of generalizing logistic regression. **Journal of the American Statistical Association**, Taylor Francis, v. 83, n. 404, p. 1078–1083, 1988. Citado na página 25.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. **Statistics and Computing**, Springer Science and Business Media, v. 27, n. 5, p. 1413–1432, aug 2016. Citado na página 50.

YEH, I.; YANG, K.; TING, T. Knowledge discovery on rfm model using bernoulli sequence. **Expert Systems with Applications**, v. 36, n. 3, Part 2, p. 5866–5871, 2009. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417408004508>. Citado na página 63.

YIN, S.; DEY, D. K.; VALDEZ, E. A.; GAN, G.; VADIVELOO, J. **Skewed link regression models for imbalanced binary response with applications to life insurance**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2007.15172>. Citado na página 24.

YONG, L. **LOO and WAIC as Model Selection Methods for Polytomous Items**. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1806.09996>. Citado na página 58.

APÊNDICE

A

MODELO RPDLOMAX ALTERNATIVO PARA A APLICAÇÃO 2

Como o intervalo de credibilidade de 90% para os parâmetros (coeficientes) relacionados às variáveis *GLCM_Pan* e *SD_Pan* englobou o valor 0, ajustou-se então o modelo RPDLomax sem essas covariáveis, utilizando as mesmas configurações dos modelos ajustados neste trabalho. Na Tabela 19 pode-se ver as medidas descritivas dos parâmetros deste modelo.

Tabela 19 – Medidas descritivas dos parâmetros do modelo RPDLomax, ajustado aos dados sobre árvores doentes, sem as covariáveis *GLCM_Pan* e *SD_Pan*.

Variável	Parâmetro	Média	Desvio-Padrão	Mediana	Quantil 5%	Quantil 95%
Intercepto	β_0	22,415	9,323	20,946	9,947	40,338
Mean_R	β_2	115,224	45,550	107,902	55,827	204,478
Mean_NIR	β_3	-15,871	6,425	-14,746	-28,308	-7,675
Assimetria	λ	0,260	0,035	0,256	0,211	0,322

Entretanto, nota-se nas Figuras 37 e 38 que os resíduos deste modelo não parecem atender à suposição de normalidade. Na Figura 37, é possível observar vários pontos fora do que se é esperado; já na Figura 38 nota-se uma certa assimetria nos resíduos quantílicos aleatorizados.

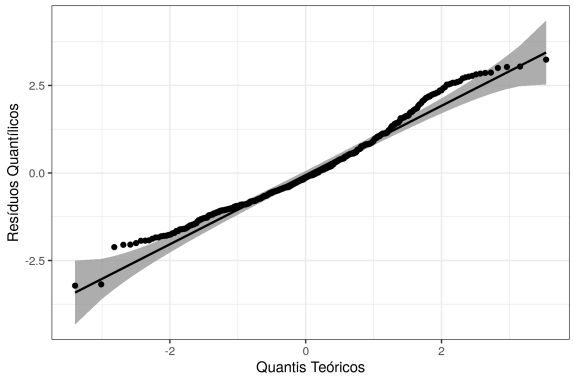


Figura 37 – *QQplot* dos resíduos quantílicos aleatorizados do modelo RPDLomax, ajustado aos dados sobre árvores doentes, sem as covariáveis *GLCM_Pan* e *SD_Pan*.

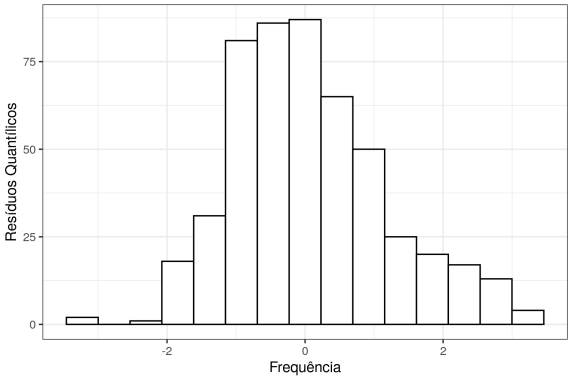


Figura 38 – Histograma dos resíduos quantílicos aleatorizados do modelo RPDLomax, ajustado aos dados sobre árvores doentes, sem as covariáveis *GLCM_Pan* e *SD_Pan*.


ICMC
USP
SÃO CARLOS