

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Alex de la Cruz Huayanay

MODELOS DE REGRESSÃO PARA RESPOSTA BINÁRIA NA
PRESENÇA DE DADOS 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 Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

São Carlos
Março de 2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Alex de la Cruz Huayanay

REGRESSION MODELS FOR BINARY RESPONSE IN THE
PRESENCE OF IMBALANCED DATA

Dissertation submitted to the Institute of Mathematics and Computer Sciences – 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. Jorge Luis Bazán Guzmán

São Carlos
March 2019

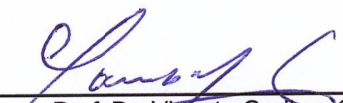


UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Alex de la Cruz Huayanay, realizada em 22/02/2019:

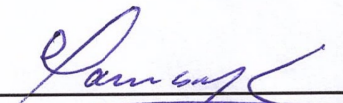


Prof. Dr. Vicente Garibay Cancho
USP

Prof. Dr. José Santos Romeo Núñez
Massey

Profa. Dra. Márcia D'Elia Branco
USP

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) José Santos Romeo Núñez, Márcia D'Elia Branco e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ão) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.



Prof. Dr. Vicente Garibay Cancho

Este trabalho é dedicado a todos aqueles que têm paixão pela ciência estatística.

AGRADECIMENTOS

Primeiramente a Deus por me dar sabedoria e iluminar meu caminho ao longo de minha vida e porque permitiu que tudo isso acontecesse.

Aos meus pais, Mauro e Elda, pelo seus exemplos com muito amor na minha vida. Mesmo longe, sempre estiveram presentes em todas as etapas cuidando de mim nos momentos difíceis e fazendo o possível para que esse sonho se realizasse.

Aos meus irmãos Nelly, Violeta, Karina, Merlín e Pamela pelo seus carinhos, por serem pessoas maravilhosas, estarem sempre preocupados com meu bem estar e por sempre me apoiarem nas minhas decisões.

A meu orientador professor Jorge Luiz Bazán, pela confiança em mim depositada, pela sua paciência e suas orientações importantes que recebi ao longo do desenvolvimento deste trabalho. Muito obrigado por ser referência de profissional.

Aos professores membros da comissão julgadora, Vicente Garibay, José Santos e Márcia Branco por disponibilizarem seu tempo em avaliar este trabalho e pelas suas valiosas sugestões e comentários.

Aos professores da USP e da UFSCar pelos valiosos ensinamentos recebidos e por me fornecerem uma base sólida na minha caminhada. Aos funcionários do PIPGEs, pela prontidão em diversos momentos e esclarecimentos prestados.

Aos meus amigos, César Cárdenas, Ingrid Cervantes, Rossmery Gonzales e Mac Jamanca, porque mesmo longe influenciaram muito na minha vida, a Roger Beltrán, Piere Rodriguez, Juan Luis Sifuentes, Gustavo Sabillón, Jeny Ventura e Yury Rojas pelo seus exemplos de perseverança e aos meus companheiros de turma do curso de mestrado, pelos bons tempos compartilhados.

Minha gratidão especial a minha família da Igreja Batista Betel por serem um valioso suporte em todos os momentos. A Vinicius e Anna Munhoz pelo constante apoio com a língua portuguesa.

A todas as pessoas que de uma alguma forma me ajudaram, eu quero deixar um agradecimento eterno, porque sem elas não teria sido possível.

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

RESUMO

ALEX C.H. **Modelos de regressão para resposta binária na presença de dados desbalanceados**. 2019. 91 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, 2019.

Na regressão binária, o desbalanceamento de dados refere-se à presença de valores zeros (ou uns) numa proporção significativamente maior do que os correspondentes valores uns (ou zeros).

Neste trabalho, estudamos dois métodos desenvolvidos para lidar com o desbalanceamento e comparamos eles com o uso de funções de ligação assimétrica potência e reversa de potência. Os resultados mostram que esses métodos não corrigem adequadamente o viés nas estimativas dos coeficientes de regressão e que os modelos com função de ligação assimétrica considerados produzem melhores resultados para certo tipo de desbalanceamento.

Adicionalmente, apresentamos uma aplicação para dados desbalanceados identificando o melhor modelo entre vários modelos propostos.

A estimação dos parâmetros é realizada sob abordagem Bayesiana considerando o método de estimação Monte Carlo Hamiltoniano usando o algoritmo No-U-Turn Sampler e as comparações dos modelos são desenvolvidas utilizando diferentes critérios para comparação de modelos, avaliação preditiva e resíduos quantílicos.

Palavras-chave: ligação assimétrica, regressão binária, dados desbalanceados, resíduos quantílicos, medidas de similaridade.

ABSTRACT

ALEX C.H. **Regression models for binary response in the presence of imbalanced data.** 2019. 91 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, 2019.

In binary regression, imbalanced data result from the presence of values equal to zero (or one) in a proportion that is significantly greater than the corresponding real values of one (or zero). In this work, we evaluate two methods developed to deal with imbalanced data and compare them to the use of asymmetric links. The results based on simulation study show, that correction methods do not adequately correct bias in the estimation of regression coefficients and that the models with power links and reverse power considered produce better results for certain types of imbalanced data. Additionally, we present an application for imbalanced data, identifying the best model among the various ones proposed. The parameters are estimated using a Bayesian approach, considering the Hamiltonian Monte-Carlo method, utilizing the No-U-Turn Sampler algorithm and the comparisons of models were developed using different criteria for model comparison, predictive evaluation and quantile residuals.

Keywords: asymmetric link, binary regression, imbalanced data, quantile residuals, similarity measures.

SUMÁRIO

1	INTRODUÇÃO	17
2	PRELIMINARES	21
2.1	Desbalanceamento	21
2.1.1	<i>Evento raro</i>	21
2.1.2	<i>Dados desbalanceados</i>	22
2.1.3	<i>Método para dados desbalanceados</i>	22
2.2	Distribuições potência e reversa de potência	24
2.2.1	<i>Reversibilidade e simetria de distribuições de probabilidade</i>	24
2.2.2	<i>Distribuições potência e reversa de potência simétrica</i>	25
3	REGRESSÃO BINÁRIA COM LIGAÇÃO POTÊNCIA E REVERSA DE POTÊNCIA	33
3.1	Regressão Binária	33
3.2	Regressão binária com função de ligação potência e reversa de potência	35
3.2.1	<i>Interpretação do parâmetro de assimetria</i>	38
3.2.2	<i>Assimetria da distribuição potência e reversa de potência</i>	39
3.2.3	<i>Relação entre proporção de sucesso e parâmetro de assimetria</i>	46
4	ESTIMAÇÃO	51
4.1	Estimação sob uma abordagem Bayesiana	51
4.1.1	<i>Distribuição a priori</i>	52
4.1.2	<i>Distribuição a posteriori</i>	53
4.2	CrITÉRIOS de comparação de modelos	56
5	ESTUDO DE SIMULAÇÃO	61
5.1	Simulação de dados desbalanceado	61
5.2	Resultados	62
6	APLICAÇÃO	69
6.1	Apresentação dos dados	69
6.2	Análise preliminar	72
6.3	Estimação com funções de ligação assimétrica	74

7	CONSIDERAÇÕES FINAIS	81
	REFERÊNCIAS	83
ANEXO A	RESULTADOS DA SIMULAÇÃO	89

INTRODUÇÃO

Nos modelos de regressão binária uma variável resposta admite apenas dois resultados, sem perda de generalidade, um chamado de sucesso e outro de fracasso. Nesses modelos pretendemos relacionar a probabilidade de sucesso com outras variáveis de interesse, chamadas de covariáveis. Na prática existem muitas situações em que esse tipo de dados são comuns como por exemplo nas ciências sociais, educação, medicina, psicologia, agronomia e zoologia. Um estudo mais detalhado sobre regressão binária pode ser visto em [Cox e Snell \(1989\)](#) e [Collett \(2002\)](#).

Formalmente, considere $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ um vetor $n \times 1$ de variáveis aleatórias independentes, cada uma tomando o valor 0 ou 1. Neste caso, pode-se supor que estas variáveis seguem uma distribuição de Bernoulli com $\Pr(Y_i = 1) = \mu_i$ a probabilidade do sucesso ou do resultado mais importante e $\Pr(Y_i = 0) = 1 - \mu_i$ a probabilidade de fracasso, para $i = 1, \dots, n$. Para relacionar a probabilidade μ_i com as covariáveis são consideradas funções de ligação e as mais comumente usadas são logito e probito, estas são simétricas que serão definidas posteriormente. Porém, esta suposição pode não ser válida em algumas situações como, por exemplo, na presença de dados desbalanceados e ignorá-la pode gerar conclusões errôneas.

Dados desbalanceados na regressão binária acontecem quando uma classe minoritária ou a classe de interesse é muito menor que a classe majoritária, i.e, o número de uns (eventos) em uma amostra é significativamente menor do que o número de zeros (não eventos). Quando os dados são extremamente desbalanceados é considerado como evento raro ([PAAL, 2014](#)).

[King e Zeng \(2001\)](#) mostraram que eventos raros são difíceis de prever, pois a aplicação padrão de técnicas de regressão logística pode subestimar a probabilidade de eventos raros.

Alguns estudos mostram que as funções de ligações comuns nem sempre fornecem o melhor ajuste para um determinado conjunto de dados de resposta binária, por exemplo, [Chen, Dey e Shao \(1999\)](#) argumentaram que quando a probabilidade de uma determinada resposta binária se aproxima de 0 em uma taxa diferente da que se aproxima de 1, as funções de ligação simétricas podem não serem tão úteis para ajustar os dados de resposta binária e que, portanto, sugerem o

uso de ligações assimétricas. Por outro lado, [Czado e Santner \(1992\)](#) consideram que os efeitos da especificação incorreta da função de ligação induz a um viés substancial nas estimativas dos coeficientes de regressão.

Os trabalhos de [Prentice \(1976\)](#), [Guerrero e Johnson \(1982\)](#), [Stukel \(1988\)](#), [Bazán, Branco e Bolfarine \(2006\)](#), [Bazán, Bolfarine e Branco \(2010\)](#) e [Wang, Dey *et al.* \(2010\)](#), tem sinalizado que podem existir problemas no uso das ligações simétricas.

Algumas funções de ligação como a log-log complementar que apresenta assimetria à direita e a ligação log-log que apresenta assimetria à esquerda, não fornecem flexibilidade pois mantêm uma assimetria de valor fixo.

Vários autores apresentaram novas propostas de funções de ligação. Assim, por exemplo [Aranda-Ordaz \(1981\)](#) introduziu uma transformação sobre a probabilidade de sucesso em que tanto o modelo logístico como o modelo log-log complementar (clog-log) são casos particulares. [Guerrero e Johnson \(1982\)](#) e [Morgan \(1983\)](#) também consideraram famílias de um parâmetro; [Prentice \(1976\)](#) sugeriu uma função de ligação baseada na função de distribuição F-Snedecor em que as funções logito, Clog-log, Log-log, Probit, Laplace e Exponencial são casos particulares e [Stukel \(1988\)](#) também apresenta uma transformação baseada nas famílias de dois parâmetros. [Nagler \(1994\)](#), [Gupta e Gupta \(2008\)](#), [Kim, Chen e Dey \(2007\)](#) e [Wang, Dey *et al.* \(2010\)](#), introduziram um parâmetro de forma associado à assimetria.

Por outro lado, nos últimos anos, nos trabalhos de [Bazán, Romeo e Rodrigues \(2014\)](#), [Abanto-Valle, Bazán e Smith \(2014\)](#), [Lemonte e Bazán \(2018\)](#) e [Bazán *et al.* \(2017\)](#) foram apresentados modelos de regressão binária com funções de ligação potência e reversa de potência introduzindo um parâmetro de assimetria.

Neste trabalho, estuda-se uma classe geral de funções de ligação para regressão binária com base na família de distribuições simétricas considerando um parâmetro adicional que controla a taxa de aumento (diminuição) da probabilidade de sucesso (falha) da variável resposta, estas funções de ligação são chamadas de potência (P) e reversa de potência (RP) proposta por [Lemonte e Bazán \(2018\)](#) e [Bazán *et al.* \(2017\)](#), estas funções incluem como casos particulares algumas ligações comuns e consideram uma função de distribuição acumulada e seu reflexão de espelho numa única família de função de ligação por meio de um parâmetro de potência, este parâmetro adicional é introduzido independentemente do preditor linear.

Para a estimação dos parâmetros, neste trabalho, consideramos a abordagem Bayesiana com uso de distribuição a priori não informativa e para a amostragem é considerado o método de No-U-Turn Sampler (NUTS) desenvolvido por [Hoffman e Gelman \(2014\)](#) que é uma extensão do Monte Carlo Hamiltoniano (MCH) proposto por [Duane *et al.* \(1987\)](#).

Este trabalho tem por objetivo principal discutir sobre os modelos de regressão binária usando funções de ligação potência e reversa de potência no que diz respeito à sua definição, resultados de inferências sob uma abordagem Bayesian considerando 10 modelos, sendo 5 modelos usando ligação potência e outros 5 usando ligação reversa de potência. Adicionalmente apresenta-se uma aplicação a dados reais.

O texto está organizado da seguinte forma. No [Capítulo 2](#) é apresentada uma breve revisão sobre o desbalanceamento de dados e tratamento deles como também uma revisão das distribuições potência e reversa de potência e suas propriedades; no [Capítulo 3](#) apresentamos a regressão binária com funções de ligação potência e reversa de potência; no [Capítulo 4](#) é considerada a estimação e ajustes dos modelos sob a abordagem Bayesiana; um estudo de simulação desenvolvido para comparar o desempenho de ligações assimétricas com dois métodos comumente usados para corrigir a regressão logística na presença de dados desbalanceados é apresentado no [Capítulo 5](#); no [Capítulo 6](#), uma aplicação para dados reais é desenvolvida considerando as funções de ligação potência e reversa de potência sob a abordagem Bayesiana. Por fim no [Capítulo 7](#) apresenta-se as considerações finais.

PRELIMINARES

Um problema que geralmente é apresentado na regressão binária com ligações comuns, é quando uma das classes da variável resposta binária é desbalanceada em relação à outra. Neste caso, segundo [King e Zeng \(2001\)](#) a estimação por meio do modelo de regressão logística apresenta um viés substancial na estimativa dos coeficientes de regressão, mesmo quando o tamanho da amostra é grande. Na literatura existem métodos para corrigir este problema.

Neste capítulo, apresentamos uma revisão bibliográfica dos métodos propostos por [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) baseados na correção do viés quando a variável resposta é um evento raro, isto é, quando uma das classes é extremamente desbalanceada em relação à outra. Por outro lado, em comparação a estes métodos propomos utilizar o método que está baseado no uso de uma função de ligação assimétrica potência ou reversa de potência de [Lemonte e Bazán \(2018\)](#) e [Bazán et al. \(2017\)](#), este método será detalhado no [Capítulo 3](#). Além disso, apresentamos alguns tópicos e definições importantes que foram aplicados nos modelos propostos bem como a definição de reversibilidade e assimetria nas distribuições, as distribuições potência e reversa de potência simétricas.

2.1 Desbalanceamento

2.1.1 *Evento raro*

Para variáveis de resposta binária, em [Calabrese e Osmetti \(2015\)](#) e [Maalouf e Trafalis \(2011\)](#), evento raro é definido como aquele em que tem-se um número extremamente baixo de ocorrência de eventos (sucessos). Segundo [King e Zeng \(2001\)](#), os eventos raros consistem em variáveis dependentes binárias com dezenas de vezes menos uns (eventos) do que zeros (não eventos).

Em várias áreas do conhecimento, a maioria dos eventos de interesse são eventos raros. Por exemplo, transações fraudulentas de cartão de crédito ([CHAN; STOLFO, 1998](#)), falhas de

equipamentos de telecomunicações (WEISS; HIRSH, 2000), derrames de petróleo (KUBAT; HOLTE; MATWIN, 1998), conflitos internacionais, descarrilamentos de trem (QUIGLEY; BEDFORD; WALLS, 2007) e outros.

Paal (2014) distingue entre dois tipos de raridade, um deles raridade relativa também chamado de dados desbalanceados e outro raridade absoluta que é essencialmente um problema de amostra pequena.

2.1.2 Dados desbalanceados

Na regressão binária, segundo Paal (2014), um conjunto de dados é considerado desbalanceado (ou com raridade relativa) quando a classe de interesse (evento) é muito menor do que a outra classe (não evento). A classe de interesse é chamado também de classe minoritária e a outra de classe majoritária. Em outras palavras, o número de observações pertencentes a uma das classe é significativamente menor do que as pertencentes às outras classes.

Para lidar com o problema quando o evento raro é subestimado, alguns pesquisadores propõem alguns métodos que são descritos abaixo.

2.1.3 Método para dados desbalanceados

A. Método de Correção de viés de regressão logística

King e Zeng (2001) propõem o método de Correção de viés do estimador de máxima verosimilhança em um modelo de regressão logística utilizando a técnica de McCullagh e Nelder (1989). Esses autores fazem as seguintes suposições, primeiro a prevalência do evento raro é conhecida e segundo propõem que seja tomada uma amostra dependente da variável resposta, ou seja, que uma amostra seja retirada de cada subpopulação (aquelas que apresentam e aquelas que não apresentam o evento raro) para que uma amostra mais balanceada seja obtida.

Segundo McCullagh e Nelder (1989) o viés de um estimador pode ser calculado, para qualquer modelo linear generalizado, por:

$$b = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{\xi}, \quad (2.1)$$

em que o primeiro fator é a matriz de informações de Fisher e $\boldsymbol{\xi}$ é um vetor cujos elementos são definidos como $\xi_i = -0.5 \left(\frac{\mu_i''}{\mu_i'} \right) \mathbf{Q}_{ii}$ e μ_i é a função de ligação inversa ($\mu_i = E(Y_i)$) relacionado ao preditor linear $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$, μ_i' e μ_i'' são, respectivamente, a primeira e segunda derivada de μ_i com respeito a η_i e \mathbf{Q}_{ii} é o i -ésimo elemento diagonal da matriz $\mathbf{Q} = \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top$, para $i = 1, \dots, n$.

Seja $\mathbf{y} = (y_1, \dots, y_n)^\top$ um vetor $n \times 1$ que representa as observações da variável resposta, então $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ é a fração de evento raro na amostra e a fração de indivíduos da população que representa o evento raro é conhecida e denotado por τ . King e Zeng (2001) propõem

modificar a função de log-verossimilhança da regressão logística usual pela seguinte função de log-verossimilhança ponderada

$$\begin{aligned} l_{\omega}(\boldsymbol{\beta}) &= \omega_1 \sum_{i=1}^n \log(\mu_i)^{y_i} + \omega_0 \sum_{i=1}^n \log(1 - \mu_i)^{(1-y_i)} \\ &= \sum_{i=1}^n \omega_i [y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i)], \end{aligned}$$

em que $\omega_i = \omega_1 y_i + \omega_0 (1 - y_i)$ e as ponderações são dadas por $\omega_1 = \frac{\tau}{\bar{y}}$ e $\omega_0 = \frac{1 - \tau}{1 - \bar{y}}$. Assim, o modelo de regressão binária teria a seguinte forma:

$$\begin{aligned} Y_i | \boldsymbol{\beta} &\sim \text{Bernoulli}(u_i^*) \quad \forall i = 1, \dots, n \\ u_i^* &= u_i^{\omega_1} \\ \eta_i &= \mathbf{x}_i^\top \boldsymbol{\beta} \end{aligned} \quad (2.2)$$

em que

$$u_i^* = u_i^{\omega_1} = \left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right)^{\omega_1}.$$

Usando a técnica de [McCullagh e Nelder \(1989\)](#) (Equação 2.1), neste caso uma aproximação do viés do EMV da regressão logística é dada por:

$$\text{Viés}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^\top \mathbf{W}_{\omega} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W}_{\omega} \boldsymbol{\xi}, \quad (2.3)$$

em que \mathbf{W}_{ω} é uma matriz diagonal $n \times n$ com elementos $w_{\omega_i} = \omega_i \hat{\mu}_i (1 - \hat{\mu}_i)$ e cada elemento do vetor $\boldsymbol{\xi}$ é definido por:

$$\xi_i = -0.5 \left(\frac{\mu_i^{*''}}{\mu_i^{*'}} \right) \mathbf{Q}_{ii}, \quad (2.4)$$

em que \mathbf{Q}_{ii} é o i -ésimo elemento diagonal de $\mathbf{Q} = \mathbf{X} \left(\mathbf{X}^\top \mathbf{W}_{\omega} \mathbf{X} \right)^{-1} \mathbf{X}^\top$, $\mu_i^{*'}$ e $\mu_i^{*''}$ são calculadas respetivamente por: $\mu_i^{*' } = \frac{\partial \mu_i^*}{\partial \eta_i} = \omega_1 \left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right)^{\omega_1} \frac{1}{1 + \exp(\eta_i)} = \omega_1 \mu_i^{\omega_1} (1 - \mu_i)$ e $\mu_i^{*''} = \frac{\partial^2 \mu_i^*}{\partial \eta_i^2} = \omega_1 \left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right)^{\omega_1} \frac{\omega - \exp(\eta_i)}{1 + \exp(\eta_i)} = \omega_1 \mu_i^{\omega_1} (1 - \mu_i) (\omega_1 - (1 + \omega_1) \mu_i)$. Substituindo estes valores estimados na Equação 2.4 temos que:

$$\xi_i = \frac{1}{2} \mathbf{Q}_{ii} [(1 + \omega_1) \hat{\mu}_i - \omega_1] \quad (2.5)$$

e com as expressões 2.3, 2.4 e 2.5, o novo estimador corrigido é dado por:

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \text{Viés}(\hat{\boldsymbol{\beta}}), \quad (2.6)$$

em que $\hat{\boldsymbol{\beta}}$ é o estimador de máxima verossimilhança (EMV) e $\tilde{\boldsymbol{\beta}}$ denota a estimativa aproximadamente não viesada de $\boldsymbol{\beta}$.

B. Método de redução de viés de Firth

Este método, proposto por Firth (1993), está baseado na penalização da função de log-verossimilhança pela *priori de Jeffrey's*, se o parâmetro objetivo for o parâmetro canônico de uma família exponencial, como é o caso do modelos lineares generalizados (GLM) e da maioria das funções de ligação, como logito, probito e cloglog.

Firth (1993) propõe um novo estimador com viés mais baixo do que o EMV, nos modelos da família exponencial este novo estimador é a solução da seguinte equação:

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) + \frac{1}{2} \text{traço} \left[I^{-1}(\boldsymbol{\beta}) \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} I(\boldsymbol{\beta}) \right\} \right] = 0, \quad (2.7)$$

em que $U^*(\boldsymbol{\beta})$ é a função de score modificada de modo que a solução resulte um estimador com viés menor ao EMV.

Para o modelo de regressão logística a Equação 2.7 pode ser escrita como:

$$U^*(\beta_j) = \sum_{i=1}^n (y_i - \mu_i) x_{ij} + \sum_{i=1}^n h_i (0.5 - \mu_i) x_{ij} = 0, \quad j = 1, \dots, p \quad (2.8)$$

h_i é o i -ésimo elemento diagonal de $\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^{\top} \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{W}^{\frac{1}{2}}$ e $\mathbf{W} = \text{diag}\{\hat{\mu}_i(1 - \hat{\mu}_i)\}$.

A solução da Equação 2.8 pode ser obtida por métodos numéricos, como exemplo o método de Newton Raphson. Nesse caso, com um valor inicial de $\boldsymbol{\beta}^{(0)}$, os coeficientes estimados dos coeficientes logísticos corrigidos $\hat{\boldsymbol{\beta}}$ são obtidos iterativamente, usando:

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + I(\boldsymbol{\beta}^{(s)})^{-1} U^*(\boldsymbol{\beta}^{(s)}),$$

em que (s) refere-se à s -ésima iteração.

C. Uso de função de ligação assimétrica

Tanto logito como probito são funções de ligação simétricas, elas aproximam-se de 0 na mesma taxa do que de 1. Vários autores sugerem que uma função de ligação assimétrica deve ser considerada quando os dados não são balanceados. Neste trabalho estudamos o uso de funções de ligação assimétrica proposto por Lemonte e Bazán (2018) e Bazán *et al.* (2017) que está baseada na função de distribuição acumulada (FDA) simétrica como linha de base e um parâmetro de potência que controla a assimetria. Maiores detalhes de este método será apresentado no Capítulo 3.

2.2 Distribuições potência e reversa de potência

2.2.1 Reversibilidade e simetria de distribuições de probabilidade

Definição 1. Seja S uma variável aleatória com uma determinada distribuição de probabilidade denotada por $S \sim F(\cdot)$ (ou $\sim f(\cdot)$), dizemos que a distribuição de S satisfaz a propriedade de

reversibilidade se a função de distribuição de probabilidade de $-S$ é uma distribuição diferente da S e que pode ser escrita como $-S \sim G(\cdot) \equiv 1 - F(-\cdot)$ (ou $\sim g(s) \equiv f(-s)$). Nesse caso, a distribuição de $G(\cdot)$ é chamada de distribuição reversa de $F(\cdot)$. (BAZÁN *et al.*, 2017)

Definição 2. Uma distribuição de probabilidade é dita ser simétrica se e somente se existir um valor x_0 tal que $f(x_0 - \delta) = f(x_0 + \delta)$ para todo número real δ , em que f é a função de densidade de probabilidade (se a distribuição for contínua) ou função de probabilidade (se a distribuição for discreta).

Resultado 1. Considere $S \sim F(\cdot)$ (ou $\sim f(\cdot)$) uma variável aleatória seguindo uma distribuição simétrica em torno do zero, então $F(\cdot)$ não satisfaz a propriedade de reversibilidade. Ou seja, S e $-S$ têm a mesma distribuição e $F(s) = 1 - F(-s)$, $f(s) = f(-s)$. Exemplos disso são as distribuições Logística, Normal, t -Student, Laplace e Cauchy.

Note que, se a simetria em torno do zero numa distribuição é verificada, não apresenta reversibilidade. Se não for verificada a simetria, é possível propor uma distribuição reversa a ela.

2.2.2 Distribuições potência e reversa de potência simétrica

No trabalho de Lemonte e Bazán (2018) descrevem a construção de uma distribuição de potência que está baseada em considerar uma função de distribuição acumulada (fda) contínua arbitrária e elevar por uma potência real positivo arbitrária. Assim é proposto uma nova função de distribuição acumulada (fda) com um parâmetro de potência adicional. Nesse sentido, tem-se a seguinte definição:

Definição 3. Uma variável aleatória univariada T tem distribuição de potência com parâmetro de locação $\mu \in \mathbb{R}$, parâmetro de escala $\sigma > 0$ e parâmetro de forma $\lambda > 0$, se sua função de distribuição acumulada (fda) é da forma

$$F_p(t | \mu, \sigma, \lambda) = G\left(\frac{t - \mu}{\sigma}\right)^\lambda, \quad t \in \mathbb{R}, \quad (2.9)$$

e sua função de densidade de probabilidade (fdp), dada por:

$$f_p(t | \sigma, \mu, \lambda) = \frac{\lambda}{\sigma} \left\{ G\left(\frac{t - \mu}{\sigma}\right) \right\}^{\lambda-1} g\left(\frac{t - \mu}{\sigma}\right), \quad t \in \mathbb{R}. \quad (2.10)$$

Em que $g(\cdot)$ e $G(\cdot)$ são, respectivamente, qualquer função de densidade de probabilidade (fdp) e função de distribuição acumulada (fda) padrão de uma distribuição univariada contínua com suporte na reta real (\mathbb{R}), esta distribuição é chamada de linha de base.

Resultado 2. Para as distribuições potência, se $G(\cdot)$ é uma distribuição simétrica então segue que:

$$F_p(-t | \mu, \sigma, \lambda) = \left\{ G\left(-\left(\frac{t - \mu}{\sigma}\right)\right) \right\}^\lambda = \left\{ 1 - G\left(\frac{t - \mu}{\sigma}\right) \right\}^\lambda, \quad t \in \mathbb{R}.$$

Demonstração. Note que quando $G(\cdot)$ é uma função de uma família de distribuição simétrica em torno do zero, satisfaz que $G(-z) = 1 - G(z)$, assim obtemos o [Resultado 2](#). \square

Definição 4. Como foi apresentado por [Lemonte e Bazán \(2018\)](#), a função de distribuição reversa de potência (RP) pode ser definido considerando sua fda como:

$$F_{rp}(t | \mu, \sigma, \lambda) = 1 - G\left(-\left(\frac{t - \mu}{\sigma}\right)\right)^\lambda, \quad t \in \mathbb{R}, \quad (2.11)$$

em que $\mu \in \mathbb{R}$, $\sigma > 0$ e $\lambda > 0$ são parâmetros de locação, escala e forma respectivamente e $G(\cdot)$ uma fda da linha de base de uma distribuição univariada contínua em sua forma padronizada.

A partir da [Definição 3](#) e [Definição 4](#), [Lemonte e Bazán \(2018\)](#) e [Bazán et al. \(2017\)](#) introduziram várias novas classes de distribuições, assumindo que a fda da linha de base $G(\cdot)$ pertence à família simétrica de distribuições. Por exemplo, algumas distribuições que pertencem a esta família são: Normal, Cauchy, t -Student e Exponencial Dupla (Laplace), entre muitas outras. Maiores detalhes sobre as distribuições simétricas podem ser encontradas em [Fang, Kotz e Ng \(1990\)](#).

Para evitar confusão na notação, neste trabalho denominaremos as distribuições de potência com distribuição de base simétrica como simplesmente distribuição potência (P) e as correspondentes distribuições reversa de potência com distribuição de base simétrica como distribuição reversa de potência (RP). Os nomes e as notações das distribuições utilizadas neste trabalho são apresentadas na [Tabela 1](#).

Tabela 1 – Algumas distribuições de linha de base, potência e reversa de potência

Tipo	Nome da distribuição	Notação
Linha de base	Logística	L
	Normal	N
	Laplace	LAPLACE
	t-Student	T
	Cauchy	C
Potência	Potência Logística	PL
	Potência Normal	PN
	Potência Laplace	PLAPLACE
	Potência t-Student	PT
	Potência Cauchy	PC
Reversa de potência	Reversa de potência Logística	RPL
	Reversa de potência Normal	RPN
	Reversa de potência Laplace	RPLAPLACE
	Reversa de potência t-Student	RPT
	Reversa de potência Cauchy	RPC

A. Distribuição potência com parâmetros de locação e escala

A partir da [Definição 3](#), considerando $G(\cdot)$ uma função de distribuição que pertence a uma família de distribuição simétrica com suporte na reta, pode ser obtida uma ampla classe de distribuições. A [Tabela 2](#) mostra a forma de função de distribuição acumulada e sua respectiva função de densidade de probabilidade de algumas distribuições de potência, cada uma delas com parâmetros de locação, escala e forma dados por $\mu \in \mathbb{R}$, $\sigma > 0$ e $\lambda > 0$, respectivamente.

Tabela 2 – fda e fdp de distribuições de potência com parâmetros de locação e escala

Distribuição	fda de distribuição de potência $F_P(x \mu, \sigma, \lambda)$	fdp de distribuição de potência $f_P(x \mu, \sigma, \lambda)$
PN	$\left[\Phi \left(\frac{x-\mu}{\sigma} \right) \right]^\lambda$	$\frac{\lambda}{\sigma} \left[\Phi \left(\frac{x-\mu}{\sigma} \right) \right]^{\lambda-1} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$
PL	$\left[\frac{1}{1 + \exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\}} \right]^\lambda$	$\lambda \left[\frac{1}{1 + \exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\}} \right]^{\lambda-1} \frac{\exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\}}{\sigma \left(1 + \exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\} \right)^2}$
PLAPLACE	$\left[\frac{1}{2} + \frac{1}{2} \text{sign}(x-\mu) \left\{ 1 - \exp \left(-\frac{ x-\mu }{\sigma} \right) \right\} \right]^\lambda$	$\frac{\lambda}{\sigma} \left[\frac{1}{2} + \frac{1}{2} \text{sign}(x-\mu) \left\{ 1 - \exp \left(-\frac{ x-\mu }{\sigma} \right) \right\} \right]^{\lambda-1} \frac{1}{2} \exp \left\{ -\frac{ x-\mu }{\sigma} \right\}$
PT	$\left[\frac{1}{2} + \frac{1}{2} \text{sign}(x-\mu) \left[1 - I_m \left(\frac{x-\mu}{\sigma} \right) \left(\frac{v}{2}, \frac{1}{2} \right) \right] \right]^\lambda$	$\frac{\lambda \Gamma \left(\frac{v+1}{2} \right)}{\sigma \sqrt{v\pi} \Gamma \left(\frac{v}{2} \right)} \left[\frac{1}{2} + \frac{1}{2} \text{sign}(x-\mu) \left[1 - I_m \left(\frac{x-\mu}{\sigma} \right) \left(\frac{v}{2}, \frac{1}{2} \right) \right] \right]^{\lambda-1} \left[1 + \frac{1}{v} \left(\frac{x-\mu}{\sigma} \right)^2 \right]^{-\left(\frac{v+1}{2} \right)}$
PC	$\left[\frac{1}{\pi} \arctan \left(\frac{x-\mu}{\sigma} \right) + \frac{1}{2} \right]^\lambda$	$\frac{\lambda}{\sigma \pi} \left[\frac{1}{\pi} \arctan \left(\frac{x-\mu}{\sigma} \right) + \frac{1}{2} \right]^{\lambda-1} \left[1 + \left(\frac{x-\mu}{\sigma} \right)^2 \right]^{-1}$

Na [Tabela 2](#), $\Phi(\cdot)$ denota uma função de distribuição acumulada de uma distribuição Normal padrão, $\Gamma(\cdot)$ uma função gama definido como $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$, em que $v > 0$ são os graus de liberdade da distribuição t -Student, $m(z) = v/(v+z^2)$, $I_x(a,b) = B(x;a,b)/B(a,b)$ é a função beta regularizada incompleta, $B(x;a,b)$ e $B(a,b)$ são as funções beta incompletas e completas, respectivamente, definidas como $B(a,b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$ e $B(z;a,b) = \int_0^z t^{a-1} (1-t)^{b-1} dt$ e $\gamma(s,x) = \int_0^x t^{s-1} e^{-t} dt$ é a função gama (inferior) incompleta.

B. Distribuição reversa de potência com parâmetros de locação e escala

Utilizando a [Definição 4](#) e segundo o [Resultado 2](#) quando $G(\cdot)$ é uma distribuição simétrica pode ser verificado que a fda de uma distribuição de reversa de potência é dada por

$$F_{rp}(x | \mu, \sigma, \lambda) = 1 - G \left(-\left(\frac{x-\mu}{\sigma} \right) \right)^\lambda = 1 - \left\{ 1 - G \left(\frac{t-\xi}{\phi} \right) \right\}^\lambda, \quad x \in \mathbb{R},$$

e sua função de densidade de probabilidade (fdp), obtido pela derivação de F_{rp} , é dada por:

$$f_{rp}(x | \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left\{ G \left(-\frac{x-\mu}{\sigma} \right) \right\}^{\lambda-1} g \left(-\frac{x-\mu}{\sigma} \right), \quad x \in \mathbb{R} \quad (2.12)$$

Na [Tabela 3](#), é apresentada algumas funções de distribuição acumulada e sua respetiva função de densidade de probabilidade das distribuições reversa de potência com parâmetros de locação e escala.

Tabela 3 – fda e fdp das distribuições reversa de potência com parâmetros de locação e escala

Distribuição	fda de distribuição reversa de potência $F_{RP}(x \mu, \sigma, \lambda)$	fdp de distribuição reversa de potência $f_{RP}(x \mu, \sigma, \lambda)$
RPN	$1 - \left[\Phi \left(-\frac{x-\mu}{\sigma} \right) \right]^\lambda$	$\frac{\lambda}{\sigma} \left[\Phi \left(-\frac{x-\mu}{\sigma} \right) \right]^{\lambda-1} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$
RPL	$1 - \left[\frac{1}{1 + \exp \left\{ \left(\frac{x-\mu}{\sigma} \right) \right\}} \right]^\lambda$	$\lambda \left[\frac{1}{1 + \exp \left\{ \left(\frac{x-\mu}{\sigma} \right) \right\}} \right]^{\lambda-1} \frac{\exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\}}{\sigma \left(1 + \exp \left\{ -\left(\frac{x-\mu}{\sigma} \right) \right\} \right)^2}$
RPLAPLACE	$1 - \left[\frac{1}{2} - \frac{1}{2} \operatorname{sign}(x-\mu) \left\{ 1 - \exp \left(-\frac{ x-\mu }{\sigma} \right) \right\} \right]^\lambda$	$\frac{\lambda}{\sigma} \left[\frac{1}{2} - \frac{1}{2} \operatorname{sign}(x-\mu) \left\{ 1 - \exp \left(-\frac{ x-\mu }{\sigma} \right) \right\} \right]^{\lambda-1} \frac{1}{2} \exp \left\{ -\frac{ x-\mu }{\sigma} \right\}$
RPT	$1 - \left[\frac{1}{2} - \frac{1}{2} \operatorname{sign} \left(\frac{x-\mu}{\sigma} \right) \left[1 - I_m \left(\frac{x-\mu}{\sigma} \right) \left(\frac{\nu}{2}, \frac{1}{2} \right) \right] \right]^\lambda$	$\frac{\lambda \Gamma \left(\frac{\nu+1}{2} \right)}{\sigma \sqrt{\nu\pi} \Gamma \left(\frac{\nu}{2} \right)} \left[\frac{1}{2} - \frac{1}{2} \operatorname{sign} \left(\frac{x-\mu}{\sigma} \right) \left[1 - I_m \left(\frac{x-\mu}{\sigma} \right) \left(\frac{\nu}{2}, \frac{1}{2} \right) \right] \right]^{\lambda-1} \left[1 + \frac{1}{\nu} \left(\frac{x-\mu}{\sigma} \right)^2 \right]^{-\left(\frac{\nu+1}{2} \right)}$
RPC	$1 - \left[\frac{1}{\pi} \arctan \left(-\frac{x-\mu}{\sigma} \right) + \frac{1}{2} \right]^\lambda$	$\frac{\lambda}{\sigma \pi} \left[\frac{1}{\pi} \arctan \left(-\frac{x-\mu}{\sigma} \right) + \frac{1}{2} \right]^{\lambda-1} \left[1 + \left(\frac{x-\mu}{\sigma} \right)^2 \right]^{-1}$

Considerando que a função de linha de base ($G(\cdot)$) pertence á família de distribuições simétricas temos os seguintes resultados:

Resultado 3. $F_{rp}(\cdot)$ satisfaz a propriedade de reversibilidade definido por [Bazán et al. \(2017\)](#), desde que

$$F_{rp}(x | \mu, \sigma, \lambda) + F_p(-x | \mu, \sigma, \lambda) = 1$$

em que $F_p(\cdot)$ e $F_{rp}(\cdot)$ é a fda da distribuição potência e reversa de potência respetivamente.

Demonstração. Pela [Definição 3](#) e [Definição 4](#), respetivamente temos que:

$$F_p(-x | \mu, \sigma, \lambda) = \left[G \left(-\frac{x-\mu}{\sigma} \right) \right]^\lambda \quad \text{e} \quad F_{rp}(x | \mu, \sigma, \lambda) = 1 - \left[G \left(-\frac{x-\mu}{\sigma} \right) \right]^\lambda,$$

somando ambas as equações e desde que $G(\cdot)$ pertence a família de distribuição simétrica, segue que:

$$F_{rp}(x | \mu, \sigma, \lambda) + F_p(-x | \mu, \sigma, \lambda) = 1 - \left[G \left(-\frac{x-\mu}{\sigma} \right) \right]^\lambda + \left[G \left(-\frac{x-\mu}{\sigma} \right) \right]^\lambda.$$

Logo,

$$F_{rp}(x | \mu, \sigma, \lambda) + F_p(-x | \mu, \sigma, \lambda) = 1.$$

□

Resultado 4. Para todo $\lambda \neq 1$ as funções de distribuição potência e reversa de potência não são simétricas, isto é, $f_p(x | \mu, \sigma, \lambda) \neq f_p(-x | \mu, \sigma, \lambda)$ e $f_{rp}(x | \mu, \sigma, \lambda) \neq f_{rp}(-x | \mu, \sigma, \lambda)$.

Demonstração. Na [Equação 2.10](#), a função de densidade de probabilidade potência é dada por:

$$f_p(x | \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left\{ G\left(\frac{x-\mu}{\sigma}\right) \right\}^{\lambda-1} g\left(\frac{x-\mu}{\sigma}\right).$$

Logo, se esta função é avaliada no ponto $-x$, temos que:

$$f_p(-x | \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left\{ G\left(-\frac{x-\mu}{\sigma}\right) \right\}^{\lambda-1} g\left(-\frac{x-\mu}{\sigma}\right) \neq f_p(x | \mu, \sigma, \lambda),$$

Da mesma forma, para a função de densidade de probabilidade reversa de potência temos que

$$f_{rp}(x | \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left\{ G\left(-\frac{x-\mu}{\sigma}\right) \right\}^{\lambda-1} g\left(\frac{x-\mu}{\sigma}\right).$$

se avaliarmos esta função no ponto $-x$, tem-se que:

$$f_{rp}(-x | \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left\{ G\left(\frac{x-\mu}{\sigma}\right) \right\}^{\lambda-1} g\left(-\frac{x-\mu}{\sigma}\right) \neq f_{rp}(x | \mu, \sigma, \lambda).$$

Quando $G(\cdot)$ é considerado a partir de uma distribuição simétrica, tem-se que $G(z) \neq G(-z)$, logo pode-se concluir que ambas as distribuições potência e reversa de potência não são simétricas em torno de zero. \square

Para $\lambda = 1$, temos que $F_p(x | \mu, \sigma^2) = F_{rp}(x | \mu, \sigma, \lambda) = G\left(\frac{x-\mu}{\sigma}\right)$. Assim, $G\left(\frac{x-\mu}{\sigma}\right)$ é um caso particular de ambas as distribuições.

A partir da [Definição 3](#) e [Definição 4](#), é possível expressar as distribuições potência e reversa de potência em sua forma padronizada fazendo $\mu = 0$ e $\sigma = 1$. Neste trabalho assumimos que $G(\cdot)$ pertence à família de distribuição simétrica em sua forma padronizada como por exemplo Normal, Logística, Cauchy, t -Student, Exponencial Dupla (Laplace).

Dentro desta família de distribuições pode se expressar $F_{rp}(z) = 1 - [1 - G(z)]^\lambda = 1 - [1 - F_p(-z)]^\lambda$.

Assim, temos a distribuição de potência padrão (PS) e distribuição de reversa de potência padrão (RPS) com distribuição de base simétrica apresentados por [Lemonte e Bazán \(2018\)](#) e [Bazán et al. \(2017\)](#).

A função de densidade de probabilidade (fdp) correspondente às distribuições PS e RPS são, respetivamente, da forma

$$f_p(z) = \lambda G(z)^{\lambda-1} g(z) \text{ e } f_{rp}(z) = \lambda G(-z)^{\lambda-1} g(z), \quad z \in \mathbb{R}.$$

C. Distribuição potência padrão

Na [Definição 3](#), a distribuição de potência padrão ocorre quando $\mu = 0$ e $\sigma = 1$, e sua função de distribuição acumulada é dada por $F_p(z) = G(z)^\lambda$, para $z \in \mathbb{R}$, em que $G(\cdot)$ é referida como a função de distribuição acumulada da linha de base em sua forma padronizada.

As funções de distribuições acumuladas de distribuição simétrica de potência padrão e sua respectiva função de densidade de probabilidade consideradas neste trabalho, são mostradas na [Tabela 4](#).

Tabela 4 – fda e fdp de distribuições de potência padrão

Distribuição	fda de distribuição de potência padrão $F_P(z)$	fdp distribuição de potência padrão $f_P(z)$
PN	$[\Phi(z)]^\lambda$	$\lambda [\Phi(z)]^{\lambda-1} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$
PL	$\left[\frac{1}{1+\exp(-z)}\right]^\lambda$	$\lambda \left[\frac{1}{1+\exp(-z)}\right]^{\lambda-1} \frac{\exp(-z)}{(1+\exp(-z))^2}$
PLAPLACE	$\left[\frac{1}{2} + \frac{1}{2}\text{sign}(z)\{1-\exp(- z)\}\right]^\lambda$	$\lambda \left[\frac{1}{2} + \frac{1}{2}\text{sign}(z)\{1-\exp(- z)\}\right]^{\lambda-1} \frac{1}{2} \exp(- z)$
PT	$\left[\frac{1}{2} + \frac{1}{2}\text{sign}(z)\left[1-I_{m(z)}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]\right]^\lambda$	$\frac{\lambda\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{1}{2} + \frac{1}{2}\text{sign}(z)\left[1-I_{m(z)}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]\right]^{\lambda-1} \left[1 + \frac{1}{\nu}z^2\right]^{-\left(\frac{\nu+1}{2}\right)}$
PC	$\left[\frac{1}{\pi} \arctan(z) + \frac{1}{2}\right]^\lambda$	$\frac{\lambda}{\pi} \left[\frac{1}{\pi} \arctan(z) + \frac{1}{2}\right]^{\lambda-1} [1+z^2]^{-1}$

D. Distribuição reversa de potência padrão

As distribuições reversa de potência padrão com distribuição de linha base simétrica $G(\cdot)$, podem ser introduzidas a partir da [Definição 4](#) e apenas considerando $\mu = 0$ e $\sigma = 1$. Na [Tabela 5](#) são mostradas algumas funções de distribuição acumulada e sua respectiva função de densidade de probabilidade da distribuição reversa de potência padrão.

Tabela 5 – fda e fdp de distribuições reversa de potência padrão

Distribuição	fda de distribuição reversa potência padrão $F_{RP}(z)$	fdp de distribuição reversa potência padrão $f_{RP}(z)$
RPN	$1 - [\Phi(-z)]^\lambda$	$\lambda [\Phi(-z)]^{\lambda-1} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$
RPL	$1 - \left[\frac{1}{1+\exp(z)}\right]^\lambda$	$\lambda \left[\frac{1}{1+\exp(z)}\right]^{\lambda-1} \frac{\exp(-z)}{(1+\exp(-z))^2}$
RPLAPLACE	$1 - \left[\frac{1}{2} - \frac{1}{2}\text{sign}(z)\{1-\exp(- z)\}\right]^\lambda$	$\lambda \left[\frac{1}{2} - \frac{1}{2}\text{sign}(z)\{1-\exp(- z)\}\right]^{\lambda-1} \frac{1}{2} \exp(- z)$
RPT	$1 - \left[\frac{1}{2} - \frac{1}{2}\text{sign}(z)\left[1-I_{m(z)}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]\right]^\lambda$	$\frac{\lambda\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{1}{2} - \frac{1}{2}\text{sign}(z)\left[1-I_{m(z)}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]\right]^{\lambda-1} \left[1 + \frac{1}{\nu}z^2\right]^{-\left(\frac{\nu+1}{2}\right)}$
RPC	$1 - \left[\frac{1}{\pi} \arctan(-z) + \frac{1}{2}\right]^\lambda$	$\frac{\lambda}{\pi} \left[\frac{1}{\pi} \arctan(-z) + \frac{1}{2}\right]^{\lambda-1} [1+z^2]^{-1}$

Observação 1. Pode ser verificado para $z \in \mathbb{R}$ que $F_P(-z) = G(-z)^\lambda$ e $F_{RP}(-z) = 1 - G(-z)^\lambda$. Logo, $F_P(\pm z) + F_{RP}(\mp z) = 1$. Também, $F_P(-z) \neq 1 - F_P(z)$ e $F_{RP}(-z) \neq 1 - F_{RP}(z)$.

Portanto, da [Observação 1](#), temos que $F_P(z)$ e $F_{RP}(z)$ não são ponto-simétricos se Z tem uma distribuição de potência. Então, $-Z$ tem uma distribuição reversa de potência.

Observação 2. As distribuições potência são inclinadas para a direita (assimetria positiva) se $\lambda > 1$ e à esquerda (assimetria negativa) se $0 < \lambda < 1$. E as distribuições reversa de potência são inclinadas para a esquerda (assimetria negativa) se $\lambda > 1$ e à direita (assimetria positiva) se $0 < \lambda < 1$.

Observação 3. Observe também que $f_p(z)$ e $f_{rp}(z)$ são funções de densidade de probabilidade (fdp) ponderados com funções de peso $w_p(z) = \lambda G(z)^{\lambda-1}$ e $w_{rp}(z) = \lambda G(-z)^{\lambda-1}$, respectivamente, dado por:

$$f_j(z) = \frac{w_j(z)}{E[w_j(Z)]} g(z), \quad j = p, rp.$$

Para os propósitos deste trabalho serão consideradas apenas as distribuições potência e reversa de potência na sua forma padronizada. Nas Figuras 1 a 4 mostramos a curva das funções de densidade de probabilidade (fdp) para algumas distribuições potência e reversa de potência com diferentes valores de λ .

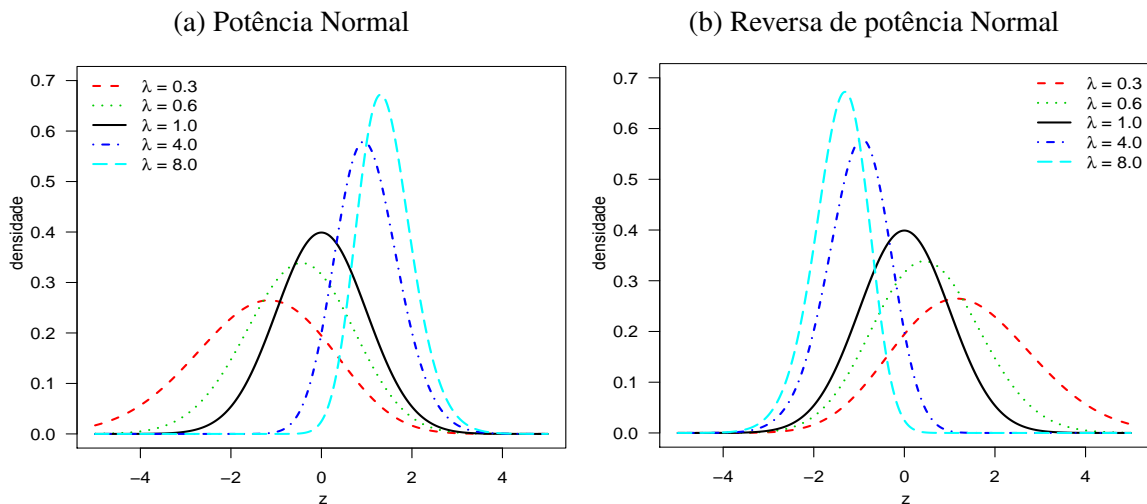


Figura 1 – fdp potência e reversa de potência Normal padrão

A importância das distribuições potência e reversa de potência está principalmente em que o parâmetro adicional λ introduz o grau de assimetria. Portanto, podemos ter um alto grau de assimetria (positivo, bem como negativo) para as distribuições potência e reversa de potência, dependendo dos valores do parâmetro de forma adicional λ . Assim, as funções de ligação assimétricas construídas a partir das distribuições potência e reversa de potência podem ser muito assimétricas. Em particular, quando $\lambda = 1$ tem-se a simetria.

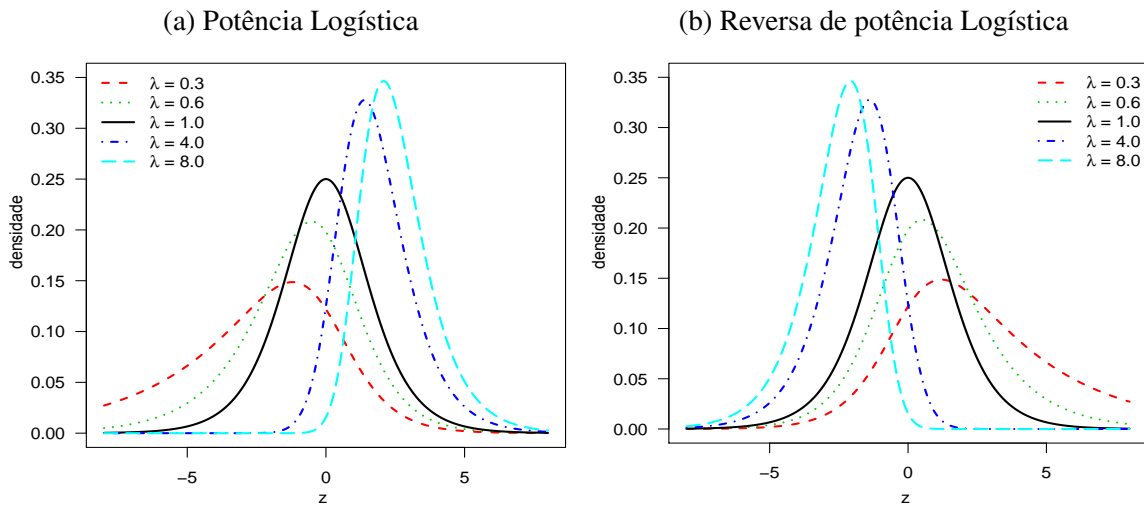


Figura 2 – fdp potência e reversa de potência Logística padrão

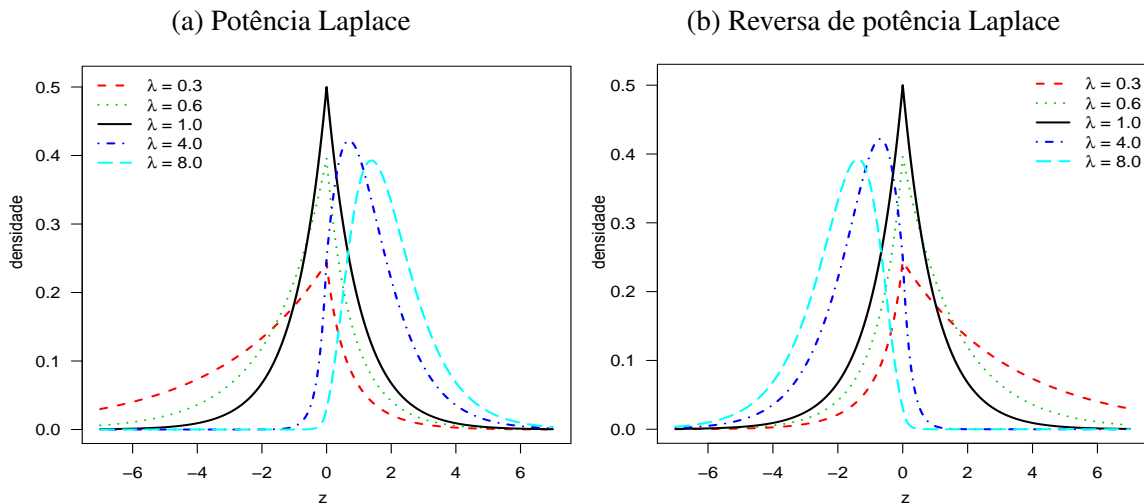
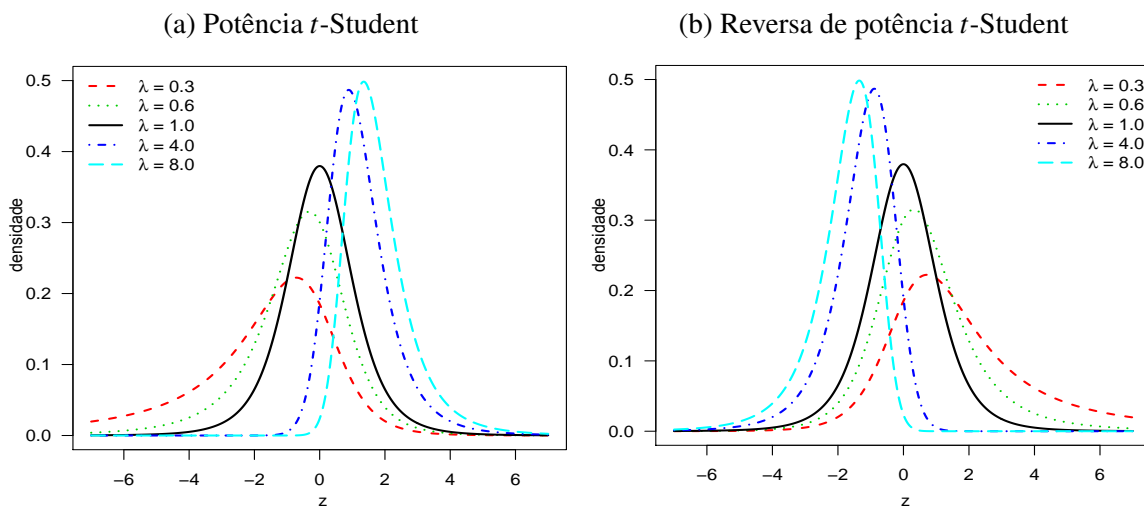


Figura 3 – fdp Potência e reversa de potência Laplace padrão

Figura 4 – fdp Potência e reversa de potência t -Student padrão com 5 graus de liberdade

REGRESSÃO BINÁRIA COM LIGAÇÃO POTÊNCIA E REVERSA DE POTÊNCIA

Este capítulo tem por objetivo mostrar a modelagem de dados binários usando funções de ligações comuns (logito, probito, cauchito, clog-log e Log-log) e usando as funções de ligação potência e reversa de potência simétricas, como também a interpretação do parâmetro de assimetria e o coeficiente de assimetria dos modelos propostos por meio de medidas e gráficos.

3.1 Regressão Binária

Os modelos de regressão binária são aqueles em que a variável de interesse, habitualmente denominada variável resposta, pode assumir somente dois valores denotados geralmente por 1 para a ocorrência do evento de interesse ("sucesso") e 0 para a ocorrência do evento complementar ("fracasso").

A variável resposta está comumente associada a outras variáveis, que podem ser contínuas, discretas ou categorizadas. Consideramos que a probabilidade de sucesso possa ser explicada por estas outras variáveis, denominadas variáveis explicativas ou covariáveis.

Seja $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ um vetor $n \times 1$ de variáveis aleatórias respostas independentes, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ um vetor $p \times 1$ de covariáveis associada a Y_i , e $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ um vetor $p \times 1$ de coeficientes de regressão associados às variáveis aleatórias explicativas (covariáveis) e seja $\Pr(Y_i = 1) = \mu_i$ a probabilidade do sucesso e $\Pr(Y_i = 0) = 1 - \mu_i$ a probabilidade do fracasso, para $i = 1, \dots, n$. No modelo de regressão binária, Y_i tem distribuição Bernoulli com parâmetro μ_i , isto é:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\mu_i), \quad i = 1, \dots, n \\ \mu_i &= \Pr(Y_i = 1) = F(\eta_i) \\ \eta_i &= F^{-1}(\mu_i) \end{aligned} \tag{3.1}$$

em que $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ é o i -ésimo preditor linear e $F(\cdot)$ denota a função de distribuição acumulada (fda).

No âmbito de modelos lineares Generalizados (MLG) (maiores detalhes em [McCullagh e Nelder, 1989](#)), F^{-1} é chamada de função de ligação. Esta função lineariza a relação entre a média da variável resposta e as covariáveis.

A função de ligação $F^{-1}(\cdot)$ é simétrica quando $F(\cdot)$ é uma fda de uma distribuição simétrica em torno do zero com $\mu_i = 0.5$. Por exemplo, as funções de ligação Logito, Probit, Cauchy. Por outro lado, quando $F(\cdot)$ é a fda de uma distribuição que não é simétrica, obtemos as funções de ligação assimétricas por exemplo Clog-log, Log-log. Especificamente, na regressão logística, temos que:

$$\mu_i = F(\mathbf{x}_i^\top \boldsymbol{\beta}) = \left(\frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right),$$

em que $F(\cdot)$ é a função de distribuição acumulada (fda) da distribuição Logística padrão e F^{-1} é chamada de função de ligação logito. E, quando $F(\cdot)$ corresponde à função de distribuição acumulada da distribuição Normal padrão, i.é, $F(\eta_i) = \Phi(\eta_i)$, temos a regressão probito e neste caso, F^{-1} é chamada de função de ligação probito.

As funções de ligação comuns consideradas como parte de modelos lineares generalizados (MLG) são mostradas na [Tabela 6](#) e as suas curvas de probabilidade na [Figura 5](#).

Tabela 6 – Ligações comuns na regressão binária

Ligação	η_i	$\mu_i = F(\eta_i) = F(\mathbf{x}_i^\top \boldsymbol{\beta})$
Logito	$\log\left(\frac{\mu_i}{1 - \mu_i}\right)$	$\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Cauchito	$F^{-1}(\mu_i)$	$\frac{1}{2} + \frac{\arctan(\eta_i)}{\pi}$
Clog-log	$\log(-\log(1 - \mu_i))$	$1 - \exp\{-\exp(\eta_i)\}$
Log-log	$\log(\log(\mu_i))$	$\exp\{\exp(\eta_i)\}$

Pode se observar que para qualquer valor de $0 < \mu_i < 1$, a função de ligação $F^{-1}(\cdot)$ faz com que o preditor linear assumira valores na reta (\mathbb{R}). Por outro lado, quando o valor do preditor linear η_i é avaliada em μ_i , os valores deste, tem coerência com os valores de probabilidade que estão dentro do intervalo 0 e 1.

A partir da [Figura 5](#) observa-se que a curva é simétrica em torno de $\mu = 0.5$ e $\eta = 0$ para as funções de ligação Probit, Logito e Cauchito; no entanto para as ligações Clog-log e Log-log são assimétricas em torno de $\eta = 0$. Algumas destas ligações são estudadas em [Mayoral e Socuélamos \(2001\)](#) e [Paula \(2004\)](#).

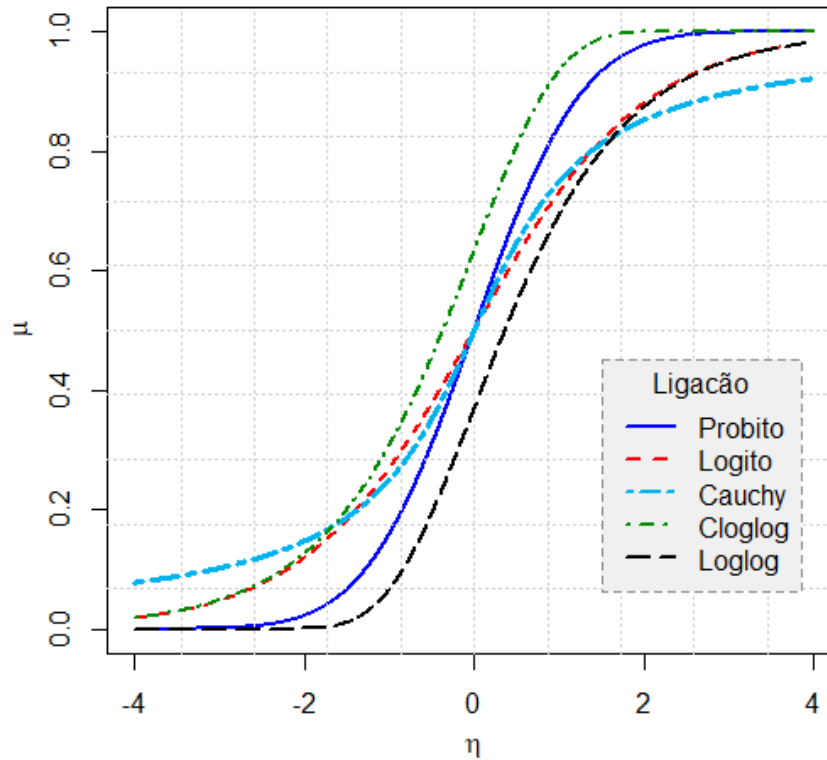


Figura 5 – Ligações comuns na regressão binária.

3.2 Regressão binária com função de ligação potência e reversa de potência

Para os modelos de regressão binária dada na [Equação 3.1](#), $F(\cdot)$ é uma função de distribuição acumulada (fda) de uma variável aleatória e sua inversa, $F^{-1}(\cdot)$, é chamada de função de ligação. Nesse sentido, para os modelos de regressão binária com função de ligação potência, é utilizada a função de distribuição acumulada de uma distribuição potência que foi apresentado na [Definição 3](#) na sua forma padronizada, por conseguinte para um modelo de regressão binária:

$$Y_i \sim \text{Bernoulli}(\mu_i), \quad i = 1, \dots, n,$$

tem-se a primeira classe de função de ligação assimétrica com $F(\cdot)$ em [Equação 3.1](#) dada por:

$$\mu_i = F_p(\eta_i) = G(\eta_i)^\lambda, \quad i = 1, \dots, n. \quad (3.2)$$

Por outro lado, para os modelos de regressão binária com função de ligação reversa de potência, é utilizada a função de distribuição acumulada de uma distribuição reversa de potência que foi apresentada na [Definição 4](#) na sua forma padronizada. Consequentemente, temos a segunda classe de ligações assimétricas com $F(\cdot)$ em [Equação 3.1](#) dada por

$$\mu_i = F_{rp}(\eta_i) = 1 - G(-\eta_i)^\lambda, \quad i = 1, \dots, n. \quad (3.3)$$

Em ambos os modelos apresentados nas Equações 3.2 e 3.3, o parâmetro adicional $\lambda > 0$ caracteriza a assimetria das funções de ligação associado ao modelo. Em particular, quando $\lambda = 1$ ambos modelos são equivalente. Além disso, a função de ligação torna-se simétrica.

Observe que nestes modelos pode ser usada uma ampla classe de funções de ligação que incluem ligações simétricas ($\lambda = 1$) e assimétricas ($\lambda \neq 1$) como casos particulares. Estes modelos, também incluem ligações simétricas comumente usadas como a Probit, Logito com base nas distribuições normal e logística padrão respectivamente (ver, por exemplo Albert e Chib, 1993) assim como as ligações assimétricas introduzidas em Bazán, Romeo e Rodrigues (2014). Além disso, as novas funções de ligação são muito simples, o que as torna alternativas interessantes às funções de ligação usuais considerados em aplicações práticas.

Neste trabalho são considerados os modelos Normal, Logística, Laplace, t -Student e Cauchy, como distribuição de base cada um deles com suas respectivas funções de ligação potência e reversa de potência como mostrado na Tabela 7.

Tabela 7 – Modelos assimétricos para regressão binária usando ligação de potência e reversa de potência

Modelo	Notação	μ	η
Potência Normal	PN	$[\Phi(\eta)]^\lambda$	$\Phi^{-1}(\mu^{1/\lambda})$
Reversa de potência Normal	RPN	$1 - [\Phi(-\eta)]^\lambda$	$-\Phi^{-1}((1-\mu)^{1/\lambda})$
Potência Logística	PL	$\left[\frac{1}{1+\exp(-\eta)}\right]^\lambda$	$-\log(\mu^{-1/\lambda} - 1)$
Reversa de potência Logística	RPL	$1 - \left[\frac{1}{1+\exp(\eta)}\right]^\lambda$	$\log((1-\mu)^{-1/\lambda} - 1)$
Potência Laplace	PLAPLACE	$\left[\frac{1}{2} + \frac{1}{2}\text{sign}(\eta) \{1 - e^{- \eta }\}\right]^\lambda$	$G_1^{-1}(\mu^{1/\lambda})$
Reversa de potência Laplace	RPLAPLACE	$1 - \left[\frac{1}{2} - \frac{1}{2}\text{sign}(\eta) \{1 - e^{- \eta }\}\right]^\lambda$	$-G_1^{-1}((1-\mu)^{1/\lambda})$
Potência t -Student	PT	$\left[\frac{1}{2} + \frac{1}{2}\text{sign}(\eta) [1 - I_{m(\eta)}(\frac{v}{2}, \frac{1}{2})]\right]^\lambda$	$G_2^{-1}(\mu^{1/\lambda})$
Reversa de potência t -Student	RPT	$1 - \left[\frac{1}{2} - \frac{1}{2}\text{sign}(\eta) [1 - I_{m(\eta)}(\frac{v}{2}, \frac{1}{2})]\right]^\lambda$	$-G_2^{-1}((1-\mu)^{1/\lambda})$
Potência Cauchy	PC	$\left[\frac{1}{\pi} \arctan(\eta) + \frac{1}{2}\right]^\lambda$	$\tan\left[\left(\mu^{1/\lambda} - \frac{1}{2}\right)\pi\right]$
Reversa de potência Cauchy	RPC	$1 - \left[\frac{1}{\pi} \arctan(-\eta) + \frac{1}{2}\right]^\lambda$	$-\tan\left[\left((1-\mu)^{1/\lambda} - \frac{1}{2}\right)\pi\right]$

Na Tabela 7, $\Phi^{-1}(\cdot)$ é a inversa da fda de uma distribuição Normal padrão, $G_1^{-1}(\cdot)$ é a inversa da fda de uma distribuição Laplace e $G_2^{-1}(\cdot)$ é a inversa da fda de uma distribuição t -Student.

Observação 4. Alguns modelos com ligação potência e reversa de potência podem ter quantidades desconhecidas adicionais (por exemplo os graus de liberdade v nas distribuições potência e reversa de potência t -Student). Estas quantidades suplementares podem ser consideradas conhecidas ou fixas. Por outro lado, na prática devemos escolher valores apropriados para essas quantidades. Um procedimento utilizado é o método condicional para estimar os parâmetros do modelo β e λ , (maiores detalhes pode ser revisado em Taylor, Siqueira e Weiss (1996)). Uma outra forma é especificar uma distribuição *a priori* adequada (abordagem não condicional) por exemplo como sugerido por Ding (2014).

As formas das curvas de resposta considerando ligação potência (P) e reversa de potência (RP), são apresentadas nas Figuras 6 a 10 para diferentes valores do parâmetro de assimetria, i.é., $\lambda = (0.3, 0.6, 1.0, 4.0, 8.0)^\top$.

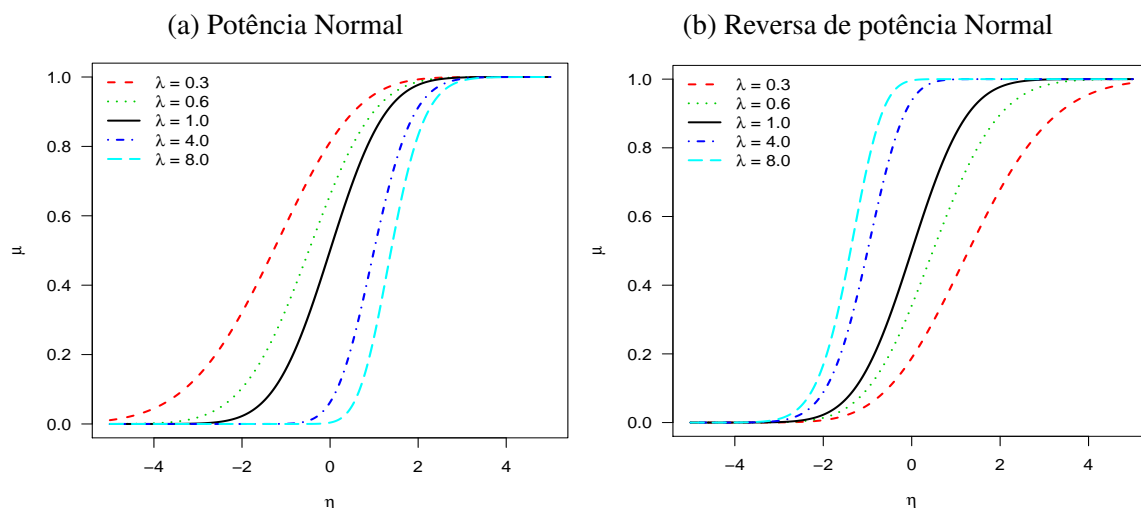


Figura 6 – Curva de resposta com ligação potência e reversa de potência Normal

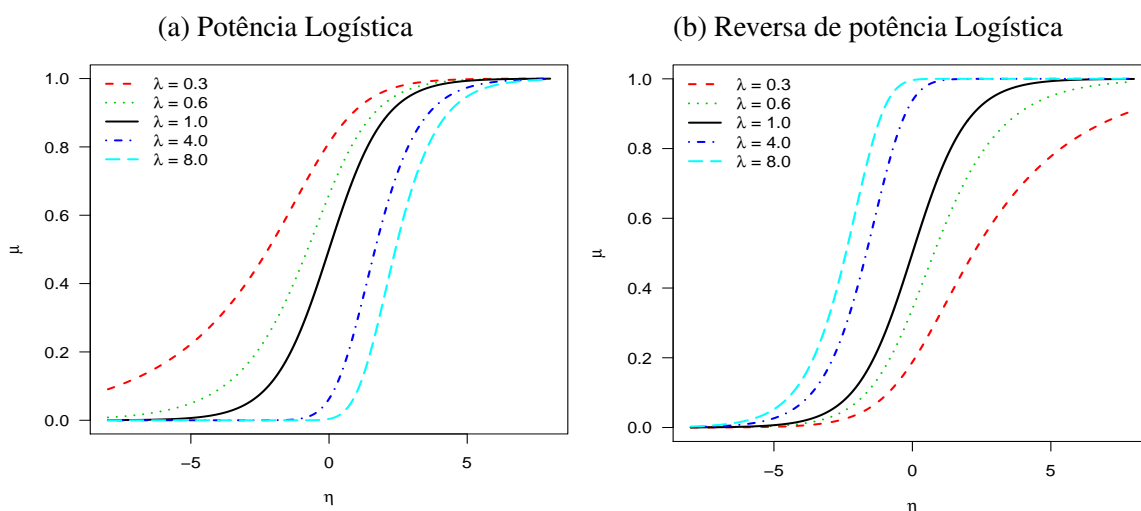


Figura 7 – Curva de resposta com ligação potência e reversa de potência Logística

A partir dessas figuras podemos considerar duas observações importantes:

Observação 5. Quando $\lambda = 1$, a probabilidade μ_i aproxima-se de zero à mesma taxa em que se aproxima de um. Neste caso, a função de ligação é simétrica em torno de $\eta = 0$.

Observação 6. Para as distribuições potência, quando $0 < \lambda < 1$, a probabilidade μ_i aproxima-se de 1 a um ritmo mais rápido do que se aproxima de zero, e quando $\lambda > 1$, a probabilidade μ_i se aproxima de zero a um ritmo mais rápido do que se aproxima de 1.

Por outro lado, para a distribuição reversa de potência quando $\lambda > 1$, a probabilidade μ_i se aproxima de um a um ritmo mais rápido do que se aproxima de zero, e quando $0 < \lambda < 1$, a probabilidade μ_i se aproxima de zero a um ritmo mais rápido do que se aproxima de 1.

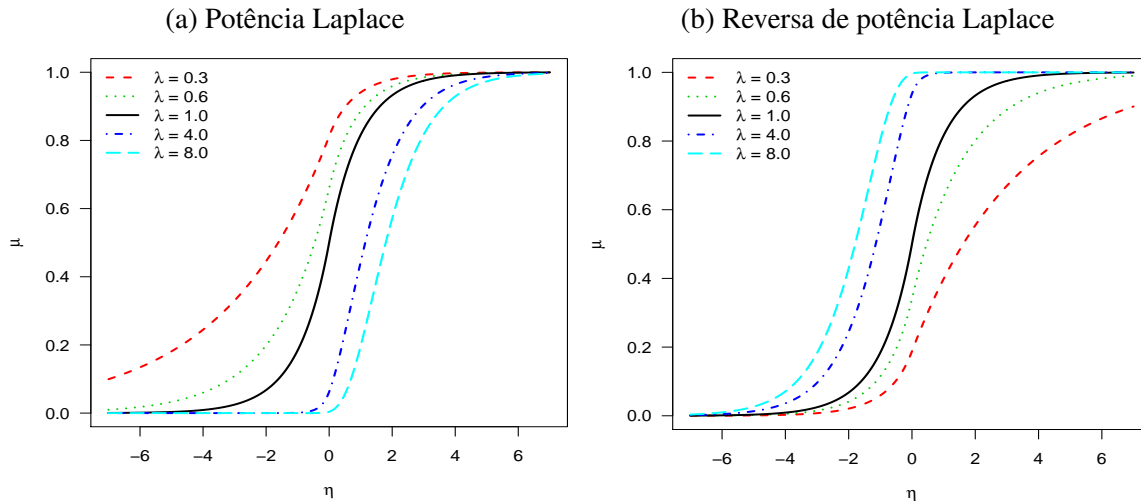


Figura 8 – Curva de resposta com ligação potência e reversa de potência Laplace

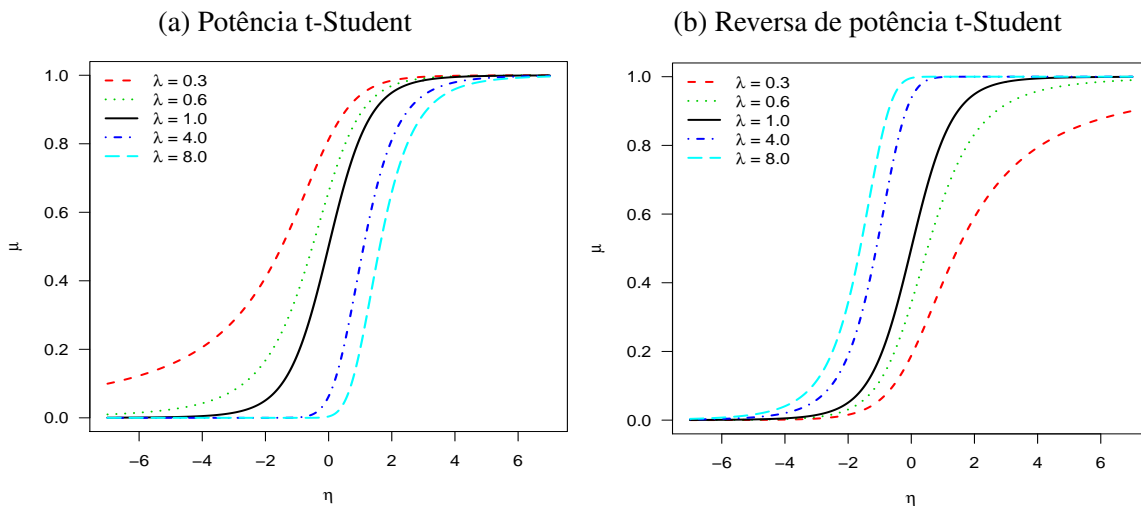


Figura 9 – Curva de resposta com ligação potência e reversa de potência t-Student $\nu = 5$ g.l.

Portanto, para um conjunto de dados de resposta binária, se o modelo verdadeiro corresponder a uma ligação assimétrica, então o uso de uma ligação simétrica será subaproveitado ou sobrecarregado (LEMONTE; BAZÁN, 2018). Em suma, o parâmetro adicional λ pode produzir um grau considerável de assimetria para as ligações, uma vez que controla o comportamento das ligações.

3.2.1 Interpretação do parâmetro de assimetria

Como foi mostrado no trabalho de Anyosa (2017), considerando as Equações 3.2 e 3.3, pode ser notado que ambas as distribuições potência e reversa de potência são distintas, embora estreitamente relacionadas uma vez que, para um valor do parâmetro de assimetria λ fixo, um delas é o reflexo do espelho da outra. Por exemplo, nas Figuras 11, 12 e 13 mostra-se um caso particular considerado a distribuição potência e reversa de potência Normal, sendo ambas distribuições equivalentes quando $\lambda = 1$ (função de ligação probito, ver Figura 12).

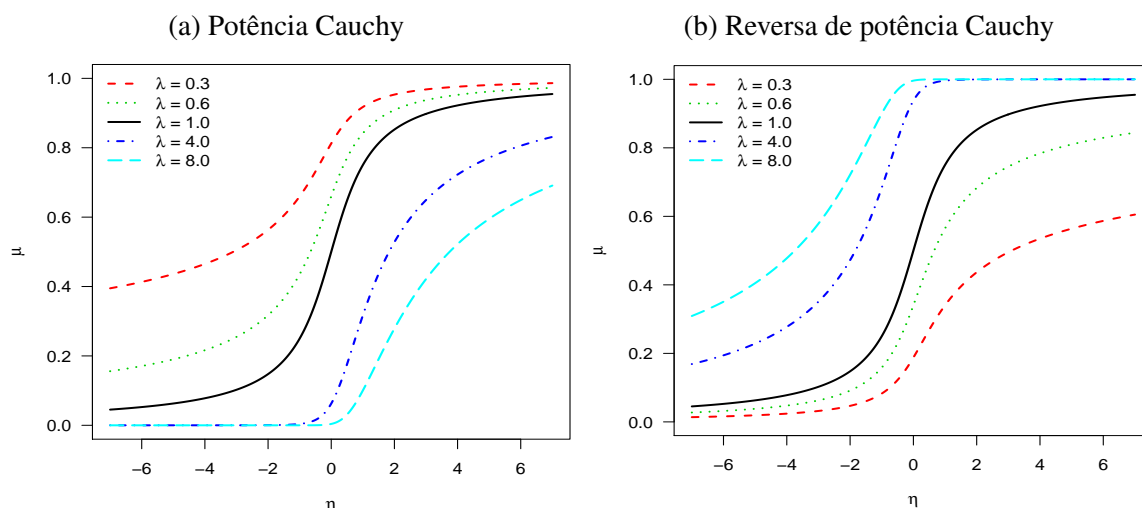


Figura 10 – Curva de resposta com ligação potência e reversa de potência Cauchy.

Note que, para $\lambda > 1$ a distribuição potência Normal tem cauda mais longa à direita. Portanto, ela é assimétrica positiva ou à direita e reversa de potência Normal tem cauda mais longa à esquerda como é visto na [Figura 13](#), e quando $0 < \lambda < 1$, a distribuição potência Normal é assimétrica à esquerda e reversa de potência Normal à direita como é mostrada na [Figura 11](#).

Neste sentido, a mudança de λ em todos os casos influencia na assimetria das funções de ligação potência e reversa de potência, implicando que influencia também na variação da probabilidade de sucesso μ_i .

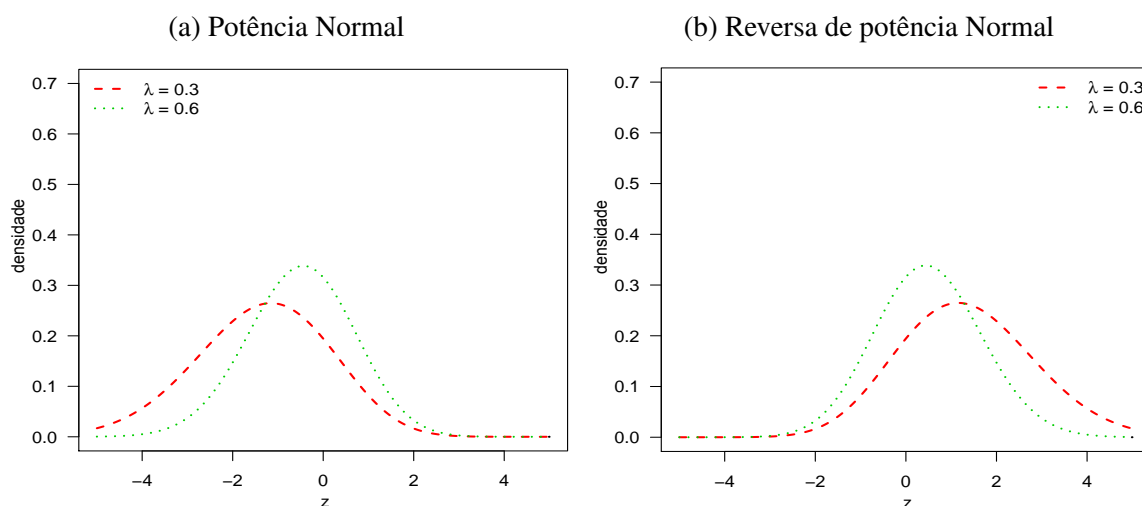


Figura 11 – fdp da distribuição potência e reversa de potência Normal para $0 < \lambda < 1$.

3.2.2 Assimetria da distribuição potência e reversa de potência

As medidas de assimetria são indicadores que permitem estabelecer o grau de simetria (ou assimetria) que apresenta uma distribuição de probabilidade de uma variável aleatória sem ter que fazer sua representação gráfica. Se uma distribuição é simétrica, há o mesmo número de valores à direita do que a esquerda da mediana, portanto, o mesmo número de desvios com sinal

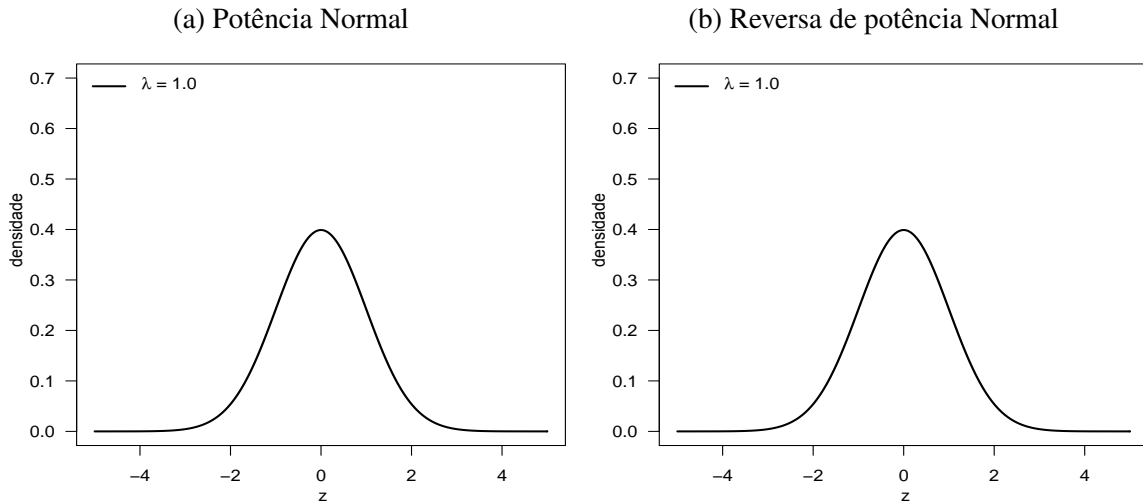


Figura 12 – fdp da distribuição potência e reversa de potência Normal para $\lambda = 1$.

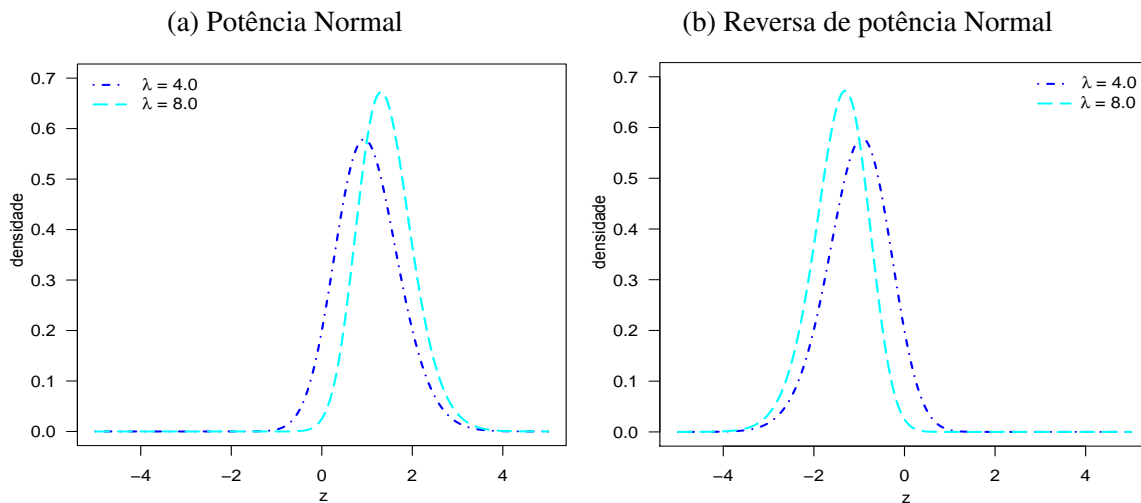


Figura 13 – fdp da distribuição potência e reversa de potência Normal para $\lambda > 1$.

positivo do que o sinal negativo. Dizemos que há assimetria negativa (ou esquerda) se a "cauda" para a esquerda da mediana for maior do que a direita, ou seja, se houver valores mais separados da mediana para a direita. Dizemos que há assimetria positiva (ou direita) se a "cauda" à direita da mediana for maior do que a esquerda, ou seja, se houver valores mais separados da mediana para a esquerda; maiores detalhes podem ser encontradas em [Doane e Seward \(2011\)](#).

Nosso interesse nesta seção é estudar as medidas de assimetria produzidos pelos diferentes valores do parâmetro λ nas distribuições potência e reversa de potência, embora existam várias medidas de assimetria, alguns delas podem ser encontradas em [MacGillivray \(1986\)](#).

Baseado no trabalho de [Anyosa \(2017\)](#), nós utilizamos as medidas de assimetria sugerido por [Hinkley \(1975\)](#), que tem a seguinte forma:

$$\frac{(Q_{1-\mu} - Q_{0.5}) - (Q_{0.5} - Q_{\mu})}{(Q_{1-\mu} - Q_{\mu})}, \quad (3.4)$$

em que Q_{μ} é o μ -ésimo quantil e $0 < \mu < 1$. Quando $\mu = 0.25$ na [Equação 3.4](#) corresponde

à assimetria quartil também conhecido como o coeficiente de Bowley e quando $\mu = 0.125$ corresponde à assimetria octil. Neste trabalho estudaremos o coeficiente de assimetria octil que foi apresentado em [Moors *et al.* \(1996\)](#) e [Brys, Hubert e Struyf \(2003\)](#) definido por:

$$A_O = \frac{(O_7 - O_4) - (O_4 - O_1)}{O_7 - O_1}, \quad (3.5)$$

em que O_i é o octil i definido por:

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

Para uma função de distribuição de probabilidade, a [Equação 3.5](#) pode ser escrita como:

$$A_O = \frac{(Q_{0.875} - Q_{0.5}) - (Q_{0.5} - Q_{0.125})}{Q_{0.875} - Q_{0.125}}. \quad (3.6)$$

Considerando μ como a probabilidade a ser avaliada no quantil e uma fda $F(\cdot)$ de uma distribuição de interesse temos que:

$$\mu = F(q) \Rightarrow Q_\mu = F^{-1}(\mu). \quad (3.7)$$

Dessa forma, para uma função de distribuição potência, o quantil é dada por:

$$\begin{aligned} \mu = F_p(q) = G(q)^\lambda &\Rightarrow \mu^{\frac{1}{\lambda}} = G(q) \\ &\Rightarrow Q_\mu^P = G^{-1}\left(\mu^{\frac{1}{\lambda}}\right). \end{aligned} \quad (3.8)$$

E, para uma distribuição reversa de potência, o quantil é dada por:

$$\begin{aligned} \mu = F_{rp}(q) = 1 - G(-q)^\lambda &\Rightarrow (1 - \mu)^{\frac{1}{\lambda}} = G(-q) \\ &\Rightarrow Q_\mu^{RP} = -G^{-1}\left((1 - \mu)^{\frac{1}{\lambda}}\right). \end{aligned} \quad (3.9)$$

Resultado 5. A partir das Equações 3.8 e 3.9 é possível estabelecer que:

$$Q_\mu^{RP} = -Q_{1-\mu}^P.$$

Demonstração. Na [Equação 3.8](#) o quantil para uma distribuição de potência para uma probabilidade $(1 - \mu)$ é dada por:

$$Q_{1-\mu}^P = G^{-1}\left((1 - \mu)^{\frac{1}{\lambda}}\right).$$

Substituindo na [Equação 3.9](#) o quantil para uma distribuição reversa de potência para uma probabilidade μ , obtemos:

$$Q_\mu^{RP} = -Q_{1-\mu}^P.$$

□

Assim, na [Equação 3.6](#) o coeficiente de assimetria octil para uma distribuição potência, tem a seguinte forma:

$$A_O^P = \frac{(Q_{0.875}^P - Q_{0.5}^P) - (Q_{0.5}^P - Q_{0.125}^P)}{Q_{0.875}^P - Q_{0.125}^P}. \quad (3.10)$$

E, para uma distribuição reversa de potência o coeficiente de assimetria octil tem a forma:

$$A_O^{RP} = \frac{(Q_{0.875}^{RP} - Q_{0.5}^{RP}) - (Q_{0.5}^{RP} - Q_{0.125}^{RP})}{Q_{0.875}^{RP} - Q_{0.125}^{RP}}. \quad (3.11)$$

Considerando o [Resultado 5](#) é possível estabelecer a seguinte relação entre os coeficientes de assimetria octil das distribuições potência e reversa de potência:

$$A_O^{RP} = -A_O^P. \quad (3.12)$$

A [Tabela 8](#) mostra a forma dos quantis para as distribuições potência e reversa de potência e a [Tabela 9](#) mostra os seus respectivos valores do coeficiente de assimetria octil A_O para cada distribuição potência e reversa de potência que foram apresentadas na [Tabela 7](#).

Tabela 8 – Quantis para distribuições potência e reversa de potência

Distribuição	Q_u^P	Q_u^{RP}
Nomal	$Q_u^{PN} = \Phi^{-1}(\mu^{1/\lambda})$	$Q_u^{RPN} = -\Phi^{-1}((1-\mu)^{1/\lambda})$
Lgística	$Q_u^{PL} = \log\left(\frac{\mu^{1/\lambda}}{1-\mu^{1/\lambda}}\right)$	$Q_u^{RPL} = -\log\left(\frac{(1-\mu)^{1/\lambda}}{1-(1-\mu)^{1/\lambda}}\right)$
Laplace	$Q_u^{PLAPLACE} = F_{Laplace}^{-1}(\mu^{1/\lambda})$	$Q_u^{PLAPLACE} = -F_{Laplace}^{-1}((1-\mu)^{1/\lambda})$
T-Student	$Q_u^{PT} = F_{t-Student}^{-1}(\mu^{1/\lambda})$	$Q_u^{RPT} = -F_{t-Student}^{-1}((1-\mu)^{1/\lambda})$
Cauchy	$Q_u^{PC} = \tan\left(\pi\left(\mu^{1/\lambda} - 0.5\right)\right)$	$Q_u^{RPC} = -\tan\left(\pi\left((1-\mu)^{1/\lambda} - 0.5\right)\right)$

Pode-se observar que para $\lambda > 0$ a variação do coeficiente de assimetria octil é $-1 < A_O < 1$. Em particular, quando $\lambda = 1$ então $A_O = 0$, uma vez que as distribuições de base são simétricas.

Nas [Figuras 14 a 18](#) são apresentados os gráficos dos valores de A_O em função de λ para cada uma das distribuições potência e reversa de potência simétrica.

Tabela 9 – Valores do coeficiente de assimetria octil para para distribuições potência e reversa de potência

Distribuição	Coeficiente de assimetria octil AO			
	$0 < \lambda < 1$	$\lambda \geq 1$	min	max
PN	$[-0.1224, 0)$	$[0, 0.1695)$	-0.1224	0.1695
RPN	$[0, 0.1224)$	$[-0.1695, 0)$	-0.1695	0.1224
PL	$[-0.4248, 0)$	$[0, 0.1997)$	-0.4248	0.1997
RPL	$[0, 0.4248)$	$[-0.1997, 0)$	-0.1997	0.4248
PLAPLACE	$[-0.4248, 0)$	$[0, 0.2747)$	-0.4248	0.2747
RPLAPLACE	$[0, 0.4248)$	$[-0.2747, 0)$	-0.2747	0.4248
PT, $\nu = 5$	$[-1.0, 0)$	$[0, 0.3265)$	-1.000	0.3265
RPT, $\nu = 5$	$[0, 1.0)$	$[-0.3265, 0)$	-0.3265	1.0000
PC	$[-1.0, 0)$	$[0, 0.7255)$	-1.000	0.7255
RPC	$[0, 1.0)$	$[-0.7255, 0)$	-0.7255	1.0000

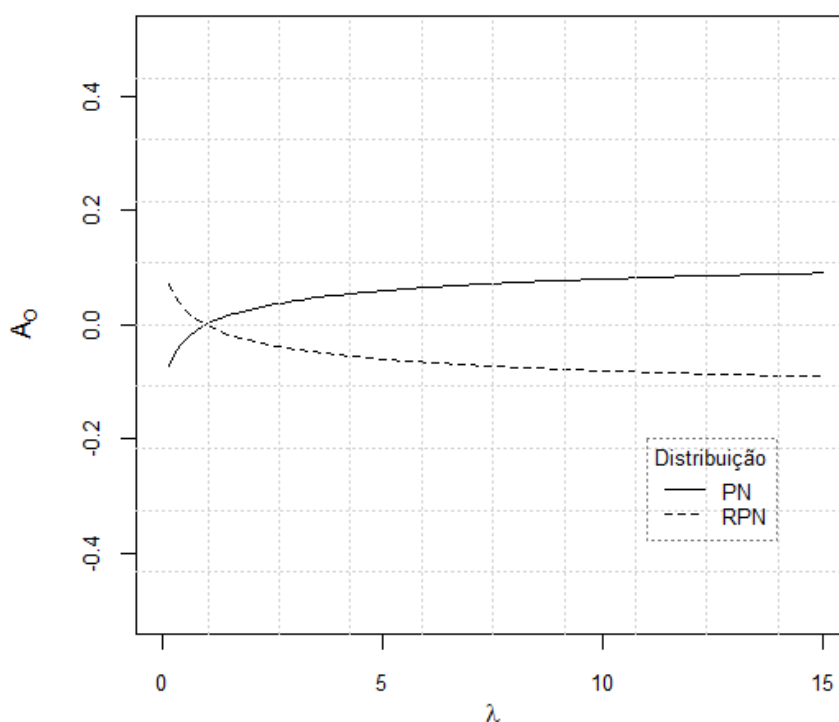


Figura 14 – Assimetria octil nas distribuições potência Normal e reversa de potência Normal.

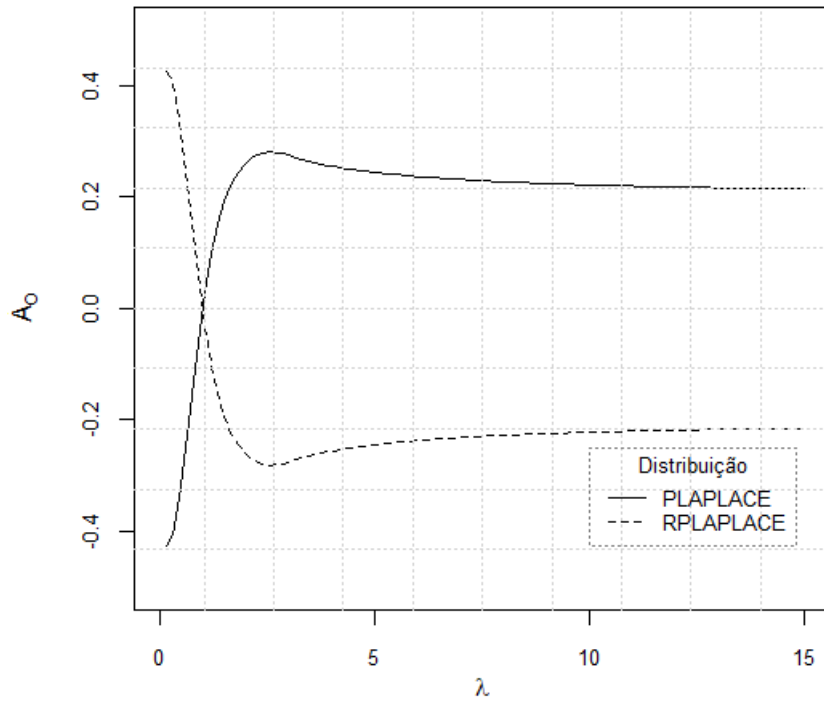


Figura 16 – Assimetria octil nas distribuições potência e reversa de potência Laplace.

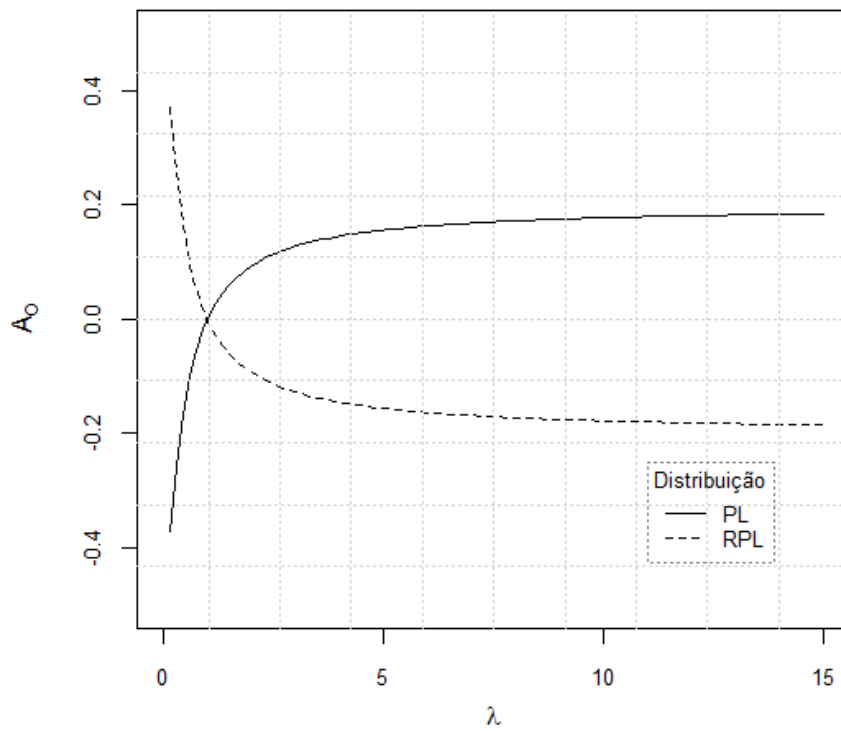


Figura 15 – Assimetria octil nas distribuições potência e reversa de potência Logística.

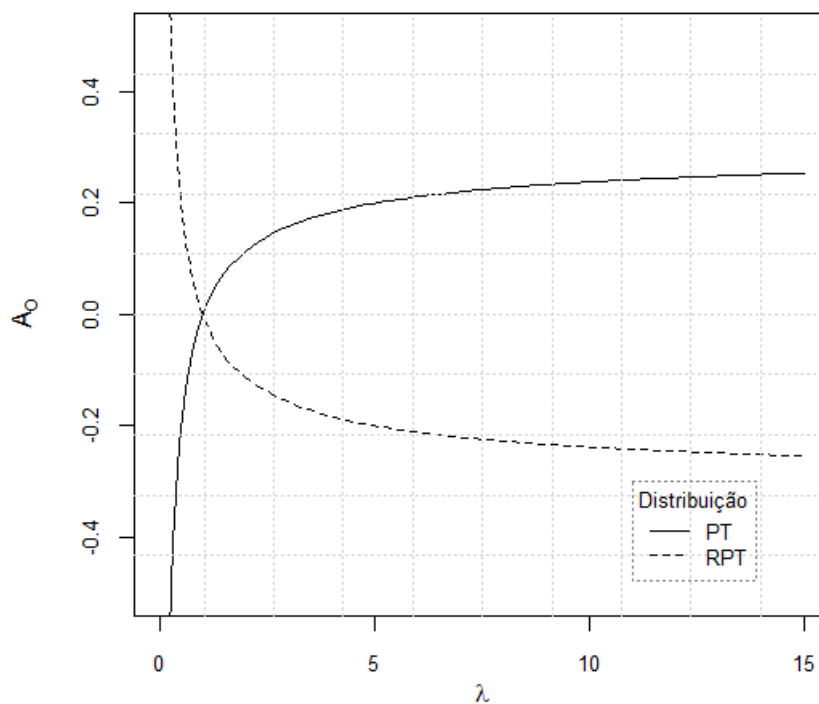


Figura 17 – Assimetria octil nas distribuições potência e reversa de potência t -Student.

Os gráficos mostram diferentes valores do coeficiente de assimetria octil em cada uma das distribuições. Pode-se observar, em todos os gráficos, que uma distribuição potência é o reflexo do espelho de uma distribuição reversa de potência; por outro lado quando $\lambda = 1$ o

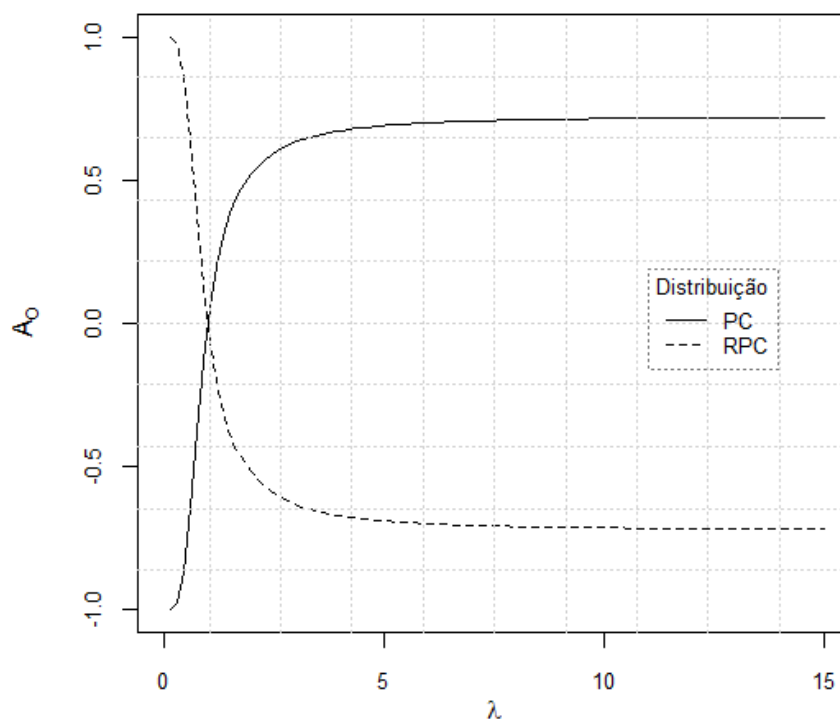


Figura 18 – Assimetria octil nas distribuições potência e reversa de potência Cauchy.

coeficiente de assimetria octil A_o é igual a 0, isto significa que, nesse caso, ambas as distribuições (P e RP) são simétricas e equivalentes.

3.2.3 Relação entre proporção de sucesso e parâmetro de assimetria

Para ver a influencia do parâmetro λ na variação de proporção de uns, aqui consideramos algumas funções de ligação propostas, como ilustração usamos as ligações potencia e reversa de potencia Logística, Normal, Laplace e Cauchy. Esta influencia é analisada por meio de gráficos. Nas primeiras figuras 19 a 22 apresentamos a relação entre μ (proporção de sucessos) e λ nas distribuições potência e reversa de potência para diferentes valores fixos do preditor linear ($\eta = -2, -1, 0, 1, 2$).

Nessas Figuras pode ser visto que que as curvas para todas as distribuições potencia são decrescentes, indicando que conforme λ aumenta, menor é a proporção de uns (sucesso) para todos os valores do preditor linear, em contraste observamos que as distribuições reversa de potencia apresentam curvas crescentes conforme aumenta o valor de λ , indicando que conforme λ aumenta, maior é a proporção de uns (sucesso) para todos os valores do preditor linear. Consequentemente λ influencia na proporção de sucessos em forma distinta em ambas as distribuições (P e RP).

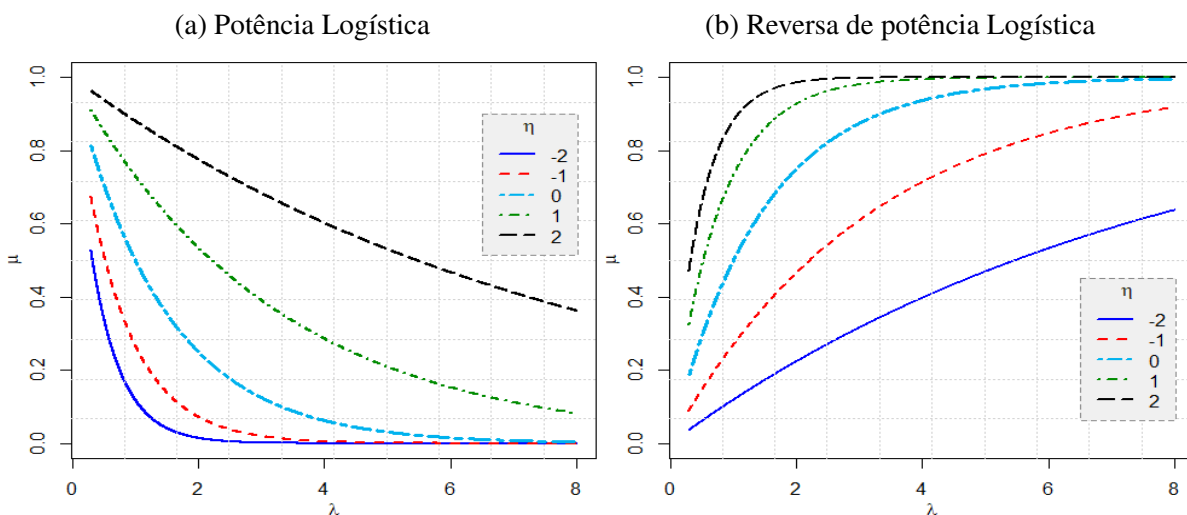


Figura 19 – Relação entre μ (proporção de sucessos) e λ para diferentes valores do preditor linear em potência e reversa de potência logística

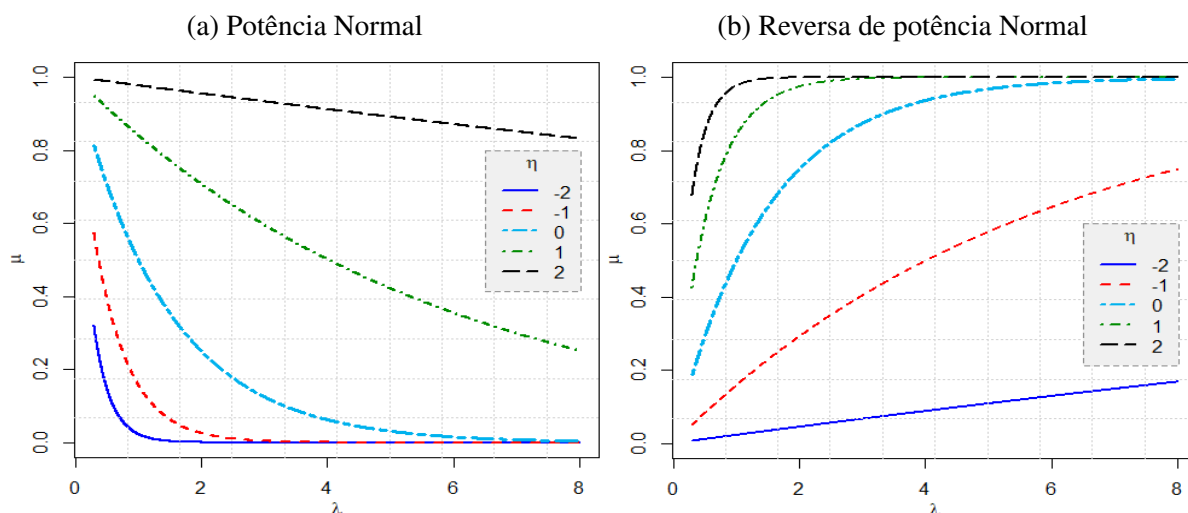


Figura 20 – Relação entre μ (proporção de sucessos) e λ para diferentes valores do predictor linear em ligação potência e reversa de potência Normal

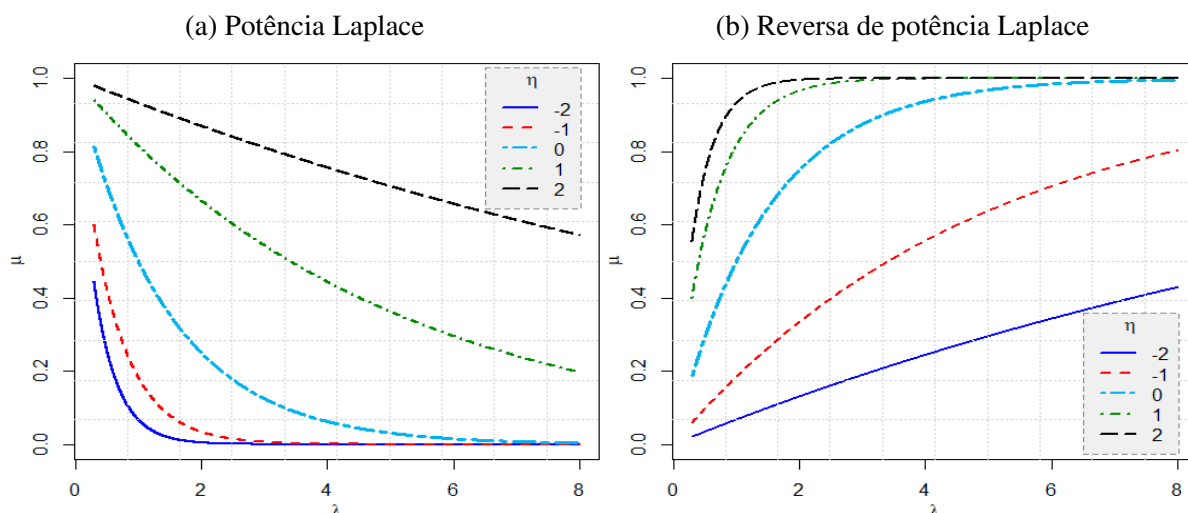


Figura 21 – Relação entre μ (proporção de sucessos) e λ para diferentes valores do predictor linear em potência e reversa de potência Laplace

Em resumo, a proporção de uns varia não somente em função de λ mas também depende do valor de η . Assim se η for negativo nas distribuições de potencia é esperado que a curva de μ e λ esteja em baixo das curvas com η positivo, por outro lado, um resultado inverso é observado nas distribuição reversa de potência. Assim, baixas proporções de uns são esperadas quando η for mais negativo e λ mais alto no caso das distribuições potencia e um resultado inverso nas distribuições reversa de potencia.

Nas Figuras 23 e 24 apresentamos a comparação da proporção de uns (sucessos) em função de λ para um determinado valor de η tanto para o caso das ligações de potência como para o caso das ligações de reversa de potência respectivamente.

Na Figura 23 parte b) pode ser visto que quando $\eta = 0$, nos temos que $\mu = 0.5^\lambda$ para qualquer distribuição escolhida; em a) nos mostramos o caso quando predictor linear é negativo

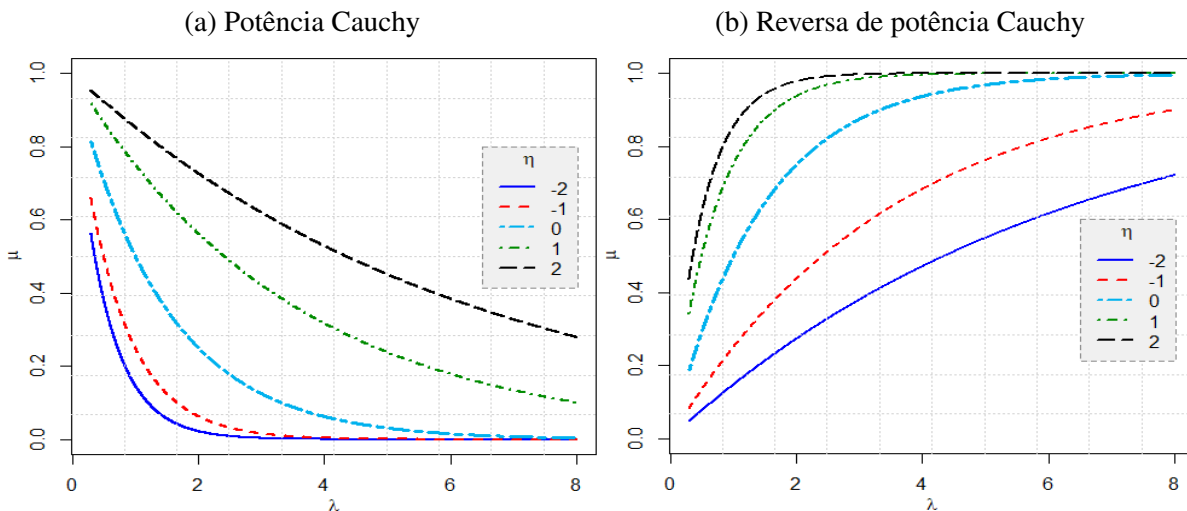
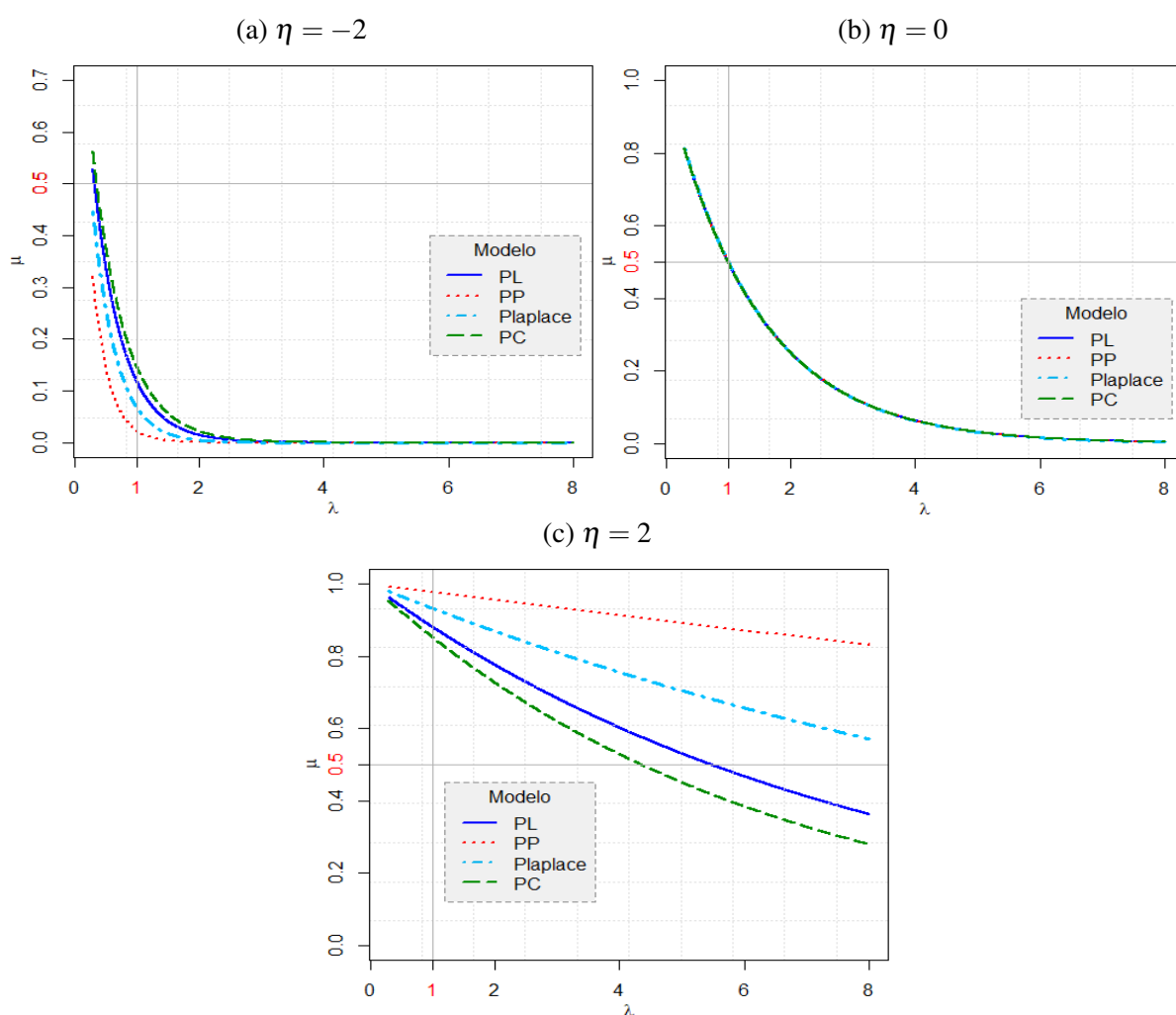


Figura 22 – Relação entre μ (proporção de sucessos) e λ para diferentes valores do preditor linear em potência e reversa de potência Cauchy

e em c) mostramos a relação entre a proporção e o parâmetro λ quando o preditor linear é positivo. Para um valor negativo do preditor linear a curva de relacionamento entre a proporção e λ da ligação PN é a que está na parte inferior e a que corresponde a ligação PC é a que está na parte superior. Assim, se o preditor linear é negativo, nós esperamos um menor valor de λ para modelar proporções de uns menores (dados desbalanceados com poucos uns) no caso da ligação PN do que nas outras ligações. De modo inverso, se o preditor linear é positivo a curva de relacionamento entre a proporção e λ da ligação PN é a que está na parte superior e a que corresponde a ligação PC é a que está na parte inferior. Assim, nós esperamos um maior valor de λ para modelar proporções de uns maiores (dados desbalanceados com muitos uns) no caso da ligação PC do que nas outras ligações.

Para valores negativos do preditor linear, na [Figura 24](#) a distribuição reversa de potência Normal apresentam maior proporção de uns e a distribuição reversa de potencia Cauchy apresenta menor proporção de uns do que outras distribuições reversa de potência. Para valores positivos do preditor linear acontece o inverso, ou seja, a distribuição reversa de potência Normal apresenta menor proporção de uns e a distribuição reversa de potencia Cauchy apresenta maior proporção de uns do que outras distribuições reversa de potência.

Nesse sentido, a mudança de λ em todos os casos, dependendo do valor do preditor e distribuição base, influencia na variação da probabilidade ou proporção de sucessor μ .

Figura 23 – Comparação de proporção de uns nas distribuições potência para diferentes valores de η .

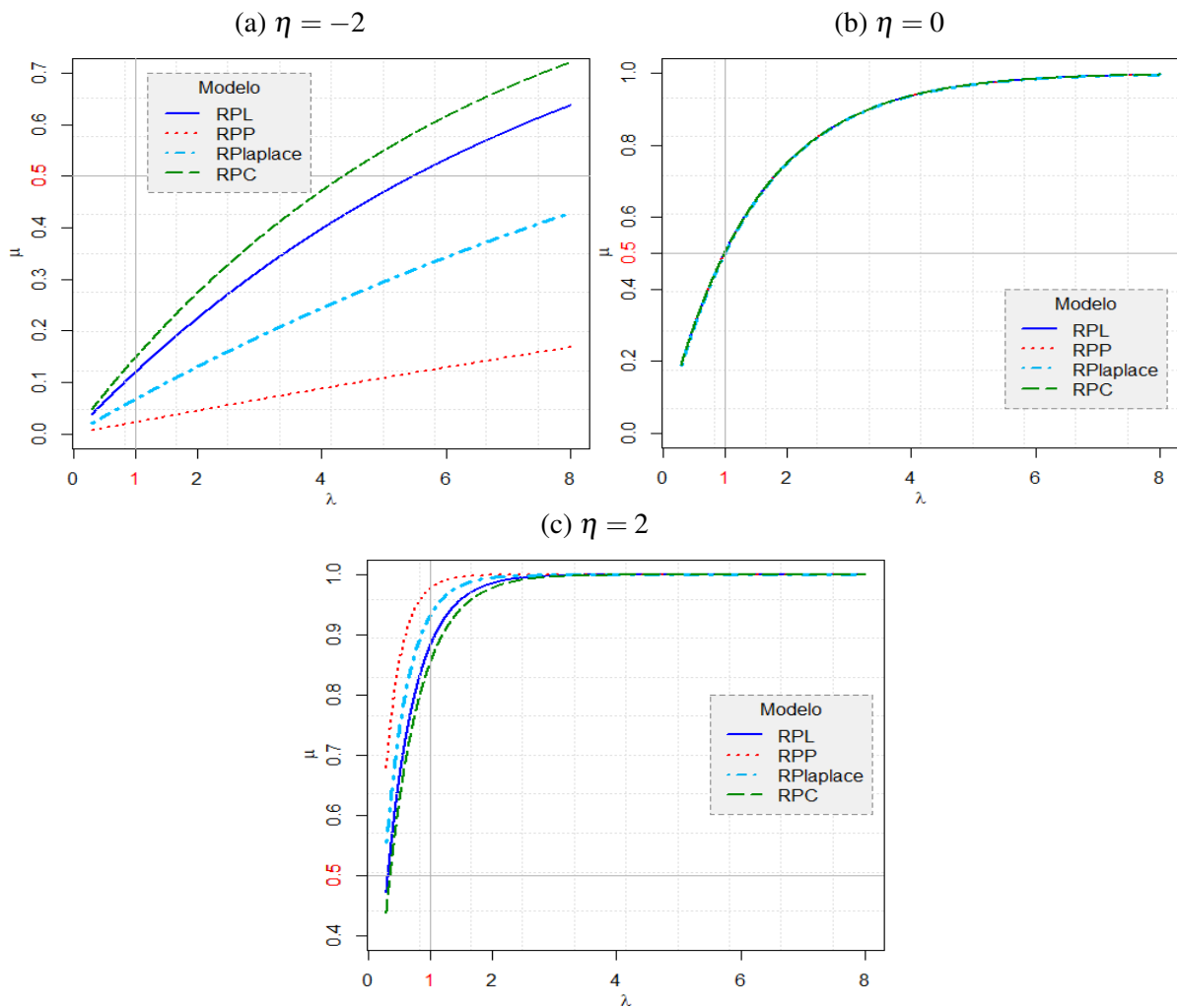


Figura 24 – Comparação de proporção de uns nas distribuições reversa de potência para diferentes valores de η .

ESTIMAÇÃO

Neste Capítulo são desenvolvidos os procedimentos inferências da estimação Bayesiana dos modelos propostos na [Seção 3.2](#), também detalhamos a especificação da distribuição *a priori* dos parâmetros no modelo proposto e a distribuição *a priori* dos parâmetros no modelo original como por exemplo os graus de liberdade da distribuição *t*-Student. Além disso, é apresentada a descrição do algoritmo MCMC de No-U-Turn Sampler, os critérios Bayesianos para a escolha do melhor modelo e as medidas de avaliação preditiva em dados desbalanceados.

4.1 Estimação sob uma abordagem Bayesiana

Sob uma abordagem Bayesiana, assumimos que o vetor de parâmetros desconhecidos é uma variável aleatória que segue uma certa distribuição de probabilidade, denominada distribuição *a priori*. Em modo geral, o modelo Bayesiano para regressão binária potência ou reversa de potência é dado por:

$$\begin{aligned}
 Y_i | \boldsymbol{\beta}, \lambda &\sim \text{Bernoulli}(u_i) \quad \forall i = 1, \dots, n \\
 u_i &= F_\lambda(\eta_i) \\
 \eta_i &= \mathbf{x}_i^T \boldsymbol{\beta} \\
 (\boldsymbol{\beta}, \lambda) &\sim \pi(\boldsymbol{\beta}, \lambda),
 \end{aligned} \tag{4.1}$$

em que $F_\lambda(\cdot)$ é uma FDA de uma distribuição potência ou reversa de potência. Neste caso, tem-se dois tipos de parâmetros que são desconhecidos, por um lado $\boldsymbol{\beta}$ é um vetor de parâmetros estruturais inerentes às observações e não depende da escolha do modelo e λ é um parâmetro estrutural associado com a escolha da função de ligação. Neste trabalho os parâmetros são conjuntamente estimados definido como estimação não condicional e precisamos definir uma distribuição *a priori* conjunta para estes parâmetros. Além disso, alguns parâmetros existentes próprios da distribuição original, como por exemplo os graus de liberdade da distribuição

t -Student, são considerados independentes aos parâmetros do modelo $(\boldsymbol{\beta}, \lambda)$. A função de verossimilhança associada ao modelo da Equação 4.1 é dada por:

$$L(\boldsymbol{\beta}, \lambda \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n [F_{\lambda}(\eta_i)]^{y_i} [1 - F_{\lambda}(\eta_i)]^{1-y_i}. \quad (4.2)$$

4.1.1 Distribuição a priori

Segundo Bazán, Romeo e Rodrigues (2014) é conveniente fazer uma re-parametrização para o parâmetro de forma λ , definindo $\delta = \log(\lambda)$ de modo que $\delta \in \mathbb{R}$. Logo, especificar uma distribuição a priori para $\boldsymbol{\beta}$ e δ , desde que ambos os parâmetros são diferentes, assumimos independência entre eles. Portanto, a distribuição a priori conjunta é dada por:

$$\pi(\boldsymbol{\beta}, \delta) = \pi_1(\boldsymbol{\beta}) \pi_2(\delta). \quad (4.3)$$

Como considerado em Bazán, Romeo e Rodrigues (2014) a distribuição a priori do vetor de coeficientes de regressão $\boldsymbol{\beta}$, assume uma função de distribuição Normal de tal maneira que

$$\beta_j \sim N(\mu_{\beta_j}, \sigma_{\beta_j}^2) \quad j = 1, \dots, p.$$

Como tem-se ignorância a priori do $\boldsymbol{\beta}$, os hiper-parâmetros são considerados vagas com valores $\mu_{\beta_j} = 0$ e $\sigma_{\beta_j}^2 = 10^2$ de forma análoga ao trabalho de Anyosa (2017).

Além disso, consideramos que o parâmetro δ assume uma distribuição a priori Uniforme, $U(-2, 2)$, em que esses hiper-parâmetros foram considerados porque valores de λ fora do intervalo $[e^{-2}, e^2] = [0.13, 7.4]$ têm muito pouca probabilidade de ocorrência (BAZÁN *et al.*, 2017) e não são observados empiricamente, isto pode ser visto nas figuras 14 a 18 em que para valores de $\lambda > 6$ a assimetria é aproximadamente constante.

Com estas especificações, o modelo 4.1 tem a seguinte forma

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \delta &\sim \text{Bernoulli}(u_i) \quad \forall i = 1, \dots, n \\ u_i &= F_{\delta}(\eta_i) \\ \eta_i &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \beta_j &\sim N(0, 10^2), \quad j = 1, \dots, p \\ \delta &\sim U(-2, 2), \end{aligned} \quad (4.4)$$

em que $F_{\delta}(\cdot)$ pode ser qualquer uma das FDA das distribuições potência ou reversa de potência. Por outro lado, para os modelos que incluem parâmetros próprios da distribuição original pode-se utilizar uma distribuição a priori adequado segundo o parâmetro envolvido. Neste trabalho o modelo Potência e Reversa de potência t -Student contem os graus de liberdade (ν) como parâmetro adicional ao modelo, este pode ser estimado de duas formas: como no trabalho de Lemonte e Bazán (2018) pode-se utilizar verossimilhança perfilada com diferentes valores de ν (abordagem condicional) ou especificar uma distribuição a priori adequada (abordagem não

condicional). [Ding \(2014\)](#) prope especificar uma distribuico gama ($v \sim \text{gamma}(\alpha, \beta)$). Em nossa abordagem, o parmetro v nas funes de ligao t -Student, potncia t -Student e reversa de potncia t -Student  assumido conhecido porque a estimativa desse parmetro extra no est associada  funo de ligao e no  fcil de ser interpretada no modelo de regresso binria. Portanto para determinar o valor de v , sugerimos o uso da verosimilhana perfilada, conforme descrito em [Lemonte e Bazn \(2018\)](#) e em nossa aplicao ([Captulo 6](#)).

4.1.2 Distribuico a posteriori

A densidade conjunta da distribuico *a posteriori* dos modelos de regresso binria com funo de ligao potncia ou reversa de potncia, tem a seguinte forma:

$$\begin{aligned} \pi(\boldsymbol{\beta}, \delta \mid \mathbf{y}, \mathbf{X}) &\propto L(\boldsymbol{\beta}, \delta \mid \mathbf{y}, \mathbf{X}) \pi(\boldsymbol{\beta}) \pi(\delta) \\ &\propto \prod_{i=1}^n [F_{\delta}(\mathbf{x}_i^T \boldsymbol{\beta})]^{y_i} [1 - F_{\delta}(\mathbf{x}_i^T \boldsymbol{\beta})]^{1-y_i} \prod_{j=1}^p \frac{1}{\sqrt{2}(10)} \exp\left\{-\frac{\beta_j^2}{2(10^2)}\right\} \frac{1}{4} \quad (4.5) \\ &\propto \prod_{i=1}^n [F_{\delta}(\mathbf{x}_i^T \boldsymbol{\beta})]^{y_i} [1 - F_{\delta}(\mathbf{x}_i^T \boldsymbol{\beta})]^{1-y_i} \prod_{j=1}^p \exp\left\{-\frac{\beta_j^2}{2(10^2)}\right\}. \end{aligned}$$

Entretanto, esta distribuico *a posteriori*, no pertence a uma famlia de distribuico conhecida e no tem uma forma fechada, o que torna essa distribuico analiticamente no tratvel. Uma alternativa  o uso de mtodos de simulao para obteno da amostra da distribuico *a posteriori*. Em particular, pode-se utilizar os mtodos de Monte Carlo via Cadeias de Markov (Markov Chain Monte Carlo: MCMC).

Os mtodos MCMC so um conjunto de tcnicas de simulao estocstica para obteno de amostras de uma distribuico comumente desconhecida. Esses mtodos so baseados na construo de cadeias de Markov cuja distribuico de interesse  a distribuico invariante (estacionria). No contexto de inferncia Bayesiana, a distribuico estacionria  a distribuico *a posteriori*.

Entre os algoritmos MCMC mais conhecidos esto o algoritmo de Metropolis-Hastings, o algoritmo de rejeio adaptativa e o algoritmo Amostrador de Gibbs. Mais detalhes sobre estes algoritmos podem ser vistos em [Gamerman e Lopes \(2006\)](#) e [Roberts e Smith \(1994\)](#).

Neste trabalho, consideramos o algoritmo de No-U-Turn Sampler desenvolvido por [Hoffman e Gelman \(2014\)](#), uma extenso para o eficiente algoritmo MCMC denominado de Monte Carlo Hamiltoniano (HMC). Este algoritmo permite que a cadeia de Markov explore a distribuico objetivo com muito mais eficincia do que outros mtodos MCMC amplamente conhecidos como amostragem de Metropolis Hastings e Gibbs (implementada por exemplo em JAGS), usando dinmica hamiltoniana em vez de uma distribuico de probabilidade e fazendo a simulao convergir mais rapidamente para a distribuico alvo com baixa autocorrelao ([NEAL, 2012](#)). Este algoritmo est implementado no software Stan.

O algoritmo HMC foi proposto por [Duane et al. \(1987\)](#) na simulação dinâmica de sistemas moleculares denominado Monte Carlo Híbrido. Posteriormente, [Neal \(1996\)](#) introduziu o método em redes neurais Bayesianas. [Neal \(2012\)](#) apresenta propriedades importantes deste método e demonstra sua convergência para alguma distribuição de interesse por exemplo, distribuição *a posteriori* dos parâmetros.

O Monte Carlo Hamiltoniano está baseada no conceito de dinâmica hamiltoniana a qual é definida em termos da localização do objeto $\boldsymbol{\theta}$ associada a uma energia potencial $U(\boldsymbol{\theta})$ e seu momento \mathbf{k} associada a uma energia cinética $K(\mathbf{k})$ em algum tempo t . A energia total do sistema, chamada de Hamiltoniana $H(\boldsymbol{\theta}, \mathbf{k})$ é definida como:

$$H(\boldsymbol{\theta}, \mathbf{k}) = U(\boldsymbol{\theta}) + K(\mathbf{k}),$$

em que $\boldsymbol{\theta}$ e \mathbf{k} são vetores de dimensão d . Assim, a dinâmica hamiltoniana é definida pelo sistema de equações diferenciais:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \mathbf{k}} = \nabla_{\mathbf{k}} K(\mathbf{k}) \quad \text{e} \\ \frac{d\mathbf{k}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}). \end{aligned}$$

Se tivermos uma posição inicial e momento inicial no tempo t_0 , i.é, $(\boldsymbol{\theta}(t_0), \mathbf{k}(t_0))$, então podemos prever a localização e o momento do objeto em qualquer tempo futuro $t = t_0 + T$ simulando dinâmicas por uma duração de tempo T .

Para relacionar $H(\boldsymbol{\theta}, \mathbf{k})$ com a distribuição alvo $\pi(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X})$, usamos um conceito da mecânica estatística conhecida como distribuição canônica. Neste caso, a distribuição *a posteriori* dos parâmetros do modelo é o foco de interesse e, portanto, esses parâmetros assumirão o papel da posição $\boldsymbol{\theta}$. A distribuição *a posteriori* pode ser visto como uma distribuição canônica (com $T = 1$) usando uma função de energia potencial definida da seguinte forma:

$$U(\boldsymbol{\theta}) = -\log(L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \pi(\boldsymbol{\theta})),$$

em que $\pi(\boldsymbol{\theta})$ é a densidade a priori e $L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X})$ é a função de verossimilhança. Note que a função de densidade conjunta de $(\boldsymbol{\theta}, \mathbf{k})$ é definida pela exponencial negativa do Hamiltoniano, uma vez que a função $K(\mathbf{k})$ tem a forma do núcleo de uma distribuição Normal multivariada $N_d(\mathbf{0}, \mathbf{M})$. Assim, temos que:

$$\begin{aligned} h(\boldsymbol{\theta}, \mathbf{k}) &\propto \exp\{-H(\boldsymbol{\theta}, \mathbf{k})\} \\ &\propto L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \pi(\boldsymbol{\theta}) \exp\left(-\mathbf{k}^\top \mathbf{M}^{-1} \mathbf{k}\right). \end{aligned}$$

Para os modelos potência e reversa de potência os parâmetro *a posteriori* será definido por $\boldsymbol{\theta} = (\boldsymbol{\beta}, \delta)^\top$ um vetor de dimensão $d = p + 1$. Assim, $U(\boldsymbol{\theta})$ tem a seguinte forma:

$$U(\boldsymbol{\beta}, \delta) = -\sum_{i=1}^n y_i \log [F_\delta(\mathbf{x}_i^\top \boldsymbol{\beta})] - \sum_{i=1}^n (1 - y_i) \log [1 - F_\delta(\mathbf{x}_i^\top \boldsymbol{\beta})] - \sum_{j=1}^p \frac{\beta_j^2}{2(10^2)}, \quad (4.6)$$

em que $F_\delta(\cdot)$ pode ser quaisquer das FDA das distribuições potência ou reversa de potência. As derivadas parciais $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ conhecido como vetor escore é definido por:

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = \left(\frac{\partial U(\boldsymbol{\beta}, \delta)}{\partial \beta_j}, \dots, \frac{\partial U(\boldsymbol{\beta}, \delta)}{\partial \beta_p}, \frac{\partial U(\boldsymbol{\beta}, \delta)}{\partial \delta} \right)^\top.$$

Por outro lado o vetor de momento \mathbf{k} pode ser obtido gerando valores de uma distribuição Normal multivariada com media $\mathbf{0}$ e matriz de variâncias e covariâncias $\mathbf{M} = \text{diag}\{m_1, \dots, m_d\}$ de modo que:

$$K(\mathbf{k}) = \frac{1}{2} \mathbf{k}^\top \mathbf{M}^{-1} \mathbf{k} = \sum_{i=1}^d \frac{k_i^2}{2m_i}.$$

Com as especificações anteriores, o algoritmo HMC em sua forma simples tomando $\mathbf{M} = I$ uma matriz identidade $d \times d$, é dado por:

- Forneça uma posição inicial $\boldsymbol{\theta}^{(0)}$ e os valores de ε , N e M .
- Inicie um contador $i = 1, \dots, N$ (tamanho da cadeia)
 - Gere $\mathbf{k}^* \sim N_d(\mathbf{0}, \mathbf{M})$ e $u \sim U(0, 1)$
 - Faça $(\boldsymbol{\theta}^I, \mathbf{k}^I) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{k}^*)$ e $H_0 = H(\boldsymbol{\theta}^I, \mathbf{k}^I)$
 - Para $j = 1$ até M , aplicar o método Leapfrog:
 - * $\mathbf{k}^* = \mathbf{k}^* + (\frac{\varepsilon}{2}) \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(i-1)})$
 - * $\boldsymbol{\theta}^{(i-1)} = \boldsymbol{\theta}^{(i-1)} + \varepsilon \mathbf{k}^*$
 - * $\mathbf{k}^* = \mathbf{k}^* + (\frac{\varepsilon}{2}) \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(i-1)})$
 - * Quando $j = M$ faça $(\boldsymbol{\theta}^L, \mathbf{k}^L) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{k}^*)$ e $H_1 = H(\boldsymbol{\theta}^L, \mathbf{k}^L)$
 - Calcular $\alpha \left[(\boldsymbol{\theta}^L, \mathbf{k}^L), (\boldsymbol{\theta}^I, \mathbf{k}^I) \right] = \min \{ \exp(H_0 - H_1), 1 \} = \alpha$
 - Faça $\boldsymbol{\theta}^i = \begin{cases} \boldsymbol{\theta}^L, & \text{com probabilidade } \alpha \\ \boldsymbol{\theta}^I, & \text{caso contrário} \end{cases}$.

O desempenho do HMC depende criticamente da escolha de valores adequados para ε e M . Má escolha destes pode levar a alta taxa de rejeição, ou o tempo de processamento computacional muito alto. O ajuste desses parâmetros para qualquer problema específico requer alguns conhecimentos e, geralmente, uma ou mais execuções preliminares.

O No-U-Turn Sampler (NUTS) é uma extensão do HMC que elimina a necessidade de especificar um valor fixo de M e define o ε com base no algoritmo de média dupla de [Nesterov \(2009\)](#). Assim, o NUTS pode ser executado sem nenhum ajuste manual, e as amostras geradas são pelo menos tão boas quanto o HMC ([NEAL, 2012](#)).

A revisão do método MCH e o desempenho do algoritmo NUTS pode ser visto no trabalho de [Anyosa \(2017\)](#).

4.2 Critérios de comparação de modelos

Akaike (1974) propôs um critério que está baseada na verossimilhança penalizada pelo número de parâmetros do modelo definido por: $AIC = -2 \sum_{i=1}^n \log \left(L \left(\hat{\boldsymbol{\theta}}; \mathbf{y} \right) \right) + 2P$. Por outro lado, o critério de informação Bayesiano (BIC) proposto por Schwarz *et al.* (1978) pondera o tamanho amostral $BIC = -2 \sum_{i=1}^n \log \left(L \left(\hat{\boldsymbol{\theta}}; y_i \right) \right) + P \log(n)$, sendo $\hat{\boldsymbol{\theta}}$ vetor dos parâmetros estimados do modelo, em que P é o número de parâmetros a serem estimados.

No entanto, os critérios de seleção no contexto Bayesiano, são obtidos por meio de uma extensão considerando a densidade a posteriori dos parâmetros do modelo e considerando o desvio dado por:

$$D(\boldsymbol{\beta}, \lambda) = -2 \sum_{i=1}^n \log(p(\mathbf{y} | \boldsymbol{\beta}, \lambda)).$$

A média *a posteriori* do desvio $E(D(\boldsymbol{\beta}, \lambda)) = E[-2 \sum_{i=1}^n \log(p(\mathbf{y} | \boldsymbol{\beta}, \lambda))]$, pode ser aproximado computacionalmente por:

$$Dbar = \frac{1}{M} \sum_{m=1}^M D(\boldsymbol{\beta}^{(m)}, \lambda^{(m)}), \quad (4.7)$$

em que o índice (m) representa a realização m de um total de M realizações, sendo M tamanho da amostra válida da distribuição a posteriori obtido usando o método de MCMC. Já o desvio da média *a posteriori* $D(E(\boldsymbol{\beta}), E(\lambda))$ é obtido por meio da média dos valores gerados a partir da distribuição *a posteriori* como:

$$Dhat = D\left(\frac{1}{M} \sum_{m=1}^M \boldsymbol{\beta}^{(m)}, \frac{1}{M} \sum_{m=1}^M \lambda^{(m)}\right). \quad (4.8)$$

O número efetivo de parâmetros $\rho_D = D(\boldsymbol{\beta}, \lambda) - D(E(\boldsymbol{\beta}), E(\lambda))$ é aproximado computacionalmente por:

$$\hat{\rho}_D = Dbar - Dhat$$

Alguns dos critérios principais de comparação são descritos a seguir

Expected Akaike Information Criterion (EAIC) e Expected Bayesian Information Criterion (EBIC)

EAIC e EBIC apresentado em Spiegelhalter *et al.* (2002) ambos os critérios estão baseados na média a posteriori do desvio e podem ser estimados, respectivamente, por:

$$\widehat{EAIC} = Dbar + 2P \quad \text{e} \quad \widehat{EBIC} = Dbar + P \log(n). \quad (4.9)$$

Em ambos os critérios, P é o número de parâmetros do modelo, $EAIC$ e $EBIC$, indicam os melhores modelos quanto menor for o valor obtido.

Deviance Information Criterion (DIC)

O DIC é uma generalização do critério de informação de Akaike (AIC) apresentado em Gelman, Hwang e Vehtari (2014) e Spiegelhalter *et al.* (2002). Está baseado na média a posteriori do desvio, pode ser estimado por:

$$\widehat{DIC} = Dbar + \hat{\rho}_D = 2Dbar - Dhat \quad (4.10)$$

Os modelos com menor DIC devem ser os preferidos.

Widely Applicable Information Criterion (WAIC)

O WAIC é uma abordagem totalmente Bayesiana. Foi introduzido por Watanabe (2010), e a ideia é calcular o logaritmo da densidade preditiva pontual (lppd) dado por:

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

Em seguida, para o sobre ajuste, é adicionado um termo para corrigir o número efetivo de parâmetros:

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

Finalmente, como proposto por Gelman, Hwang e Vehtari (2014), o WAIC é estimado por

$$\widehat{WAIC} = -2 \left(\widehat{lppd} - \widehat{pWAIC} \right). \quad (4.11)$$

Leave-one-out cross-validation (LOO)

Outro método de comparação de modelo totalmente Bayesiano é o método leave-one-out cross-validation (LOO) proposto por Geisser e Eddy (1979). Devido à sua natureza iterativa, LOO pode ser computacionalmente proibitivo para grandes conjuntos de dados de amostra em que o modelo precisa ser ajustado para um determinado tamanho de amostra n , Vehtari, Gelman e Gabry (2017) propõe usar Pareto smoothed importance sampling (PSIS), uma nova abordagem que fornece uma estimativa precisa e confiável que permite calcular LOO usando pesos de importância que de outra forma seria instável. A estimativa Bayesiana de PSIS-LOO é dada por:

$$\widehat{elpd}_{psis-loo} = \sum_{i=1}^n \log \left(\frac{\sum_{m=1}^M w_i^{(m)} p(y_i | \boldsymbol{\beta}^{(m)}, \lambda^{(m)})}{\sum_{m=1}^M w_i^{(m)}} \right). \quad (4.12)$$

em que

$$w_i^{(m)} = \min(r_i^{(m)}, \frac{\sqrt{M}}{M} \sum_{m=1}^M r_i^{(m)}) \quad \text{e} \quad r_i^{(m)} = \frac{1}{p(y_i | \boldsymbol{\beta}^{(m)}, \lambda^{(m)})} \propto \frac{p(\boldsymbol{\beta}^{(m)}, \lambda^{(m)} | \mathbf{Y}_{-i})}{p(\boldsymbol{\beta}^{(m)}, \lambda^{(m)} | \mathbf{Y})}$$

Luo e Al-Harbi (2017) mostraram que como EAIC, EBIC e DIC usam estimativas pontuais em sua computação, enquanto LOO e WAIC são calculados com base em toda a distribuição *a posterior*, que os métodos que usam mais informação (a distribuição *a posterior*) desempenhem melhor do que aqueles que usam menos informação (estimativa pontual). Portanto, WAIC e LOO realizam o melhor devido ao uso completo da distribuição *a posterior*, o DIC vem o segundo devido ao seu uso parcial da distribuição *a posterior*, e os outros métodos que não usam a distribuição *a posterior* têm o menor poder estatístico na seleção de modelo. Os cálculos de WAIC e LOO, podem ser obtidos usando pacotes implementados em R como LOO (ver Vehtari, Gelman e Gabry, 2016) e em Python como PSIS (ver Vehtari, Gelman e Gabry, 2017).

Avaliação preditiva

Após o ajuste de um modelo, é importante avaliar o poder de discriminação do modelo, isto é, discriminar os eventos dos não eventos. Para essa avaliação, várias métricas foram criadas. Na Tabela 10 apresentamos a matriz de confusão. É uma forma simples de se estabelecer e visualizar o cálculo dessas métricas, em que na linha está o valor previsto e na coluna o valor observado (valor verdadeiro). Maiores detalhes pode ser visto em Fawcett (2006).

Tabela 10 – Matriz de confusão

Valor	Predito (\hat{y})		
	0 (Falha)	1 (Sucesso)	
Observado (y)	0 (Falha)	Verdadeiro negativo (TN)	Falso negativo (FN)
	1 (Sucesso)	Falso positivo (FP)	Verdadeiro positivo (TP)

A partir desta tabela as medidas muito comuns e bastante utilizadas são: Acurácia (taxa de boa classificação), Sensibilidade, Especificidade, Verdadeiro Preditivo Positivo e Verdadeiro Preditivo Negativo que podem ser definidas como:

Acurácia (ACC): proporção de acertos de um modelo. Ou seja, é a proporção de verdadeiros-positivos e verdadeiros-negativos em relação a todos os resultados possíveis.

$$ACC = (TP + TN) / (TP + FP + TN + FN)$$

Sensibilidade (S): proporção de eventos, classificados corretamente pelo modelo. Ou seja, é a probabilidade de ser classificado como evento ($\hat{Y} = 1$) dado que realmente ele é evento ($Y = 1$).

$$S = TP / (TP + FN).$$

Especificidade (E): proporção de não evento, classificados corretamente pelo modelo. Ou seja, avalia a capacidade do modelo predizer como não evento ($\hat{Y} = 0$) dado que ele realmente é não evento ($Y = 0$).

$$E = TN / (TN + FP).$$

Valor Preditivo Positivo (VPP): proporção de eventos, dado que o modelo assim os identificou.

$$VPP = TP / (TP + FP).$$

Valor Preditivo Negativo (VPN): proporção de não evento, dado que o modelo assim os identificou.

$$VPN = TN / (TN + FN).$$

Por outro lado, como indicado por [Choi, Cha e Tappert \(2010\)](#), nos casos em que as classes binárias são desbalanceadas, como no tipo assimétrico de dados binários, as correspondências positivas (eventos) são geralmente mais significativas do que as correspondências negativas (não eventos). Algumas medidas de similaridade binária de dados binários assimétricos devem ser consideradas. Por esta razão, propomos usar o *Jaccard index* também chamado de *índice de sucesso crítico (CSI)*, *Índice de Sokal & Sneath (SSI)*, *índice de Faith (FAITH)* e medida de distância *pattern difference (PDIF)* para medir a similaridade entre a classificação observada e a prevista usando a matriz de confusão correspondente. Estas medidas são definidas, respectivamente, por:

$$CSI = \frac{TP}{TP + FP + FN}, \quad SSI = \frac{TP}{TP + 2 \times FP + 2 \times FN},$$

$$FAITH = \frac{TP + 0.5 \times TN}{TP + FP + FN + TN} \quad \text{e} \quad PDIF = \frac{4 \times FP \times FN}{(TP + FP + FN + TN)^2}.$$

Além disso, usamos o *Gilbert skill score (GS)* proposto por [Schaefer \(1990\)](#), que modifica o CSI para lidar com os problemas associados ao valor muito grande de TN , o que acontece claramente na predição de eventos raros. O GS pode ser escrito como:

$$GS = \frac{(TP \times TN - FP \times FN)}{(FN + FP)(TP + FP + FN + TN) + (TP \times TN - FP \times FN)}$$

Para todas essas medidas, um modelo com maior valor deve ser preferido em relação a outros modelos possíveis, pois apresenta maior similaridade entre a classificação observada e a prevista.

Para obter os valores de cada uma das medidas, neste trabalho será utilizado as médias *a posteriori* como estimador *plug-in*, as probabilidades preditas para cada um dos modelos serão calculadas estabelecendo um ponto de corte de 0.5 para logo formar a matriz de confusão ([Tabela 10](#)).

Análise de resíduos

No modelo linear generalizado, o resíduo de desvio para dados de resposta binária assume a forma

$$r_i^d = (y_i - \hat{\mu}_i) \left[2y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + 2(1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right]^{1/2} \quad i = 1, \dots, n.$$

Então, o resíduo de desvio para os modelos de regressão potência e reversa de potência segundo [Lemonte e Bazán \(2018\)](#) surge quando $\hat{\mu}_i = G(\hat{\eta}_i)^\lambda$ e $\hat{\mu}_i = 1 - G(-\hat{\eta}_i)^\lambda$, respectivamente, com $\hat{\eta}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$.

Por outro lado, para verificar a adequação dos modelos [Lemonte e Bazán \(2018\)](#) recomendam usar os resíduos quantílicos aleatorizados normalizados proposto por [Dunn e Smyth \(1996\)](#), para os modelos com ligação potências e reversa de potência, esses resíduos são dados por:

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

Para respostas inteiras discretas, u_i é um valor aleatório da distribuição Uniforme no intervalo

$$[I_{1-\hat{\mu}_i}(2 - y_i, y_i), I_{1-\hat{\mu}_i}(1 - y_i, y_i + 1)], \quad i = 1, \dots, n,$$

em que $I_x(a, b)$ é a função beta incompleta regularizada. Segundo [Lemonte e Bazán \(2018\)](#), para os modelos potência e reversa de potência, $\hat{\mu}_i = G(\hat{\eta}_i)^\lambda$ e $\hat{\mu}_i = 1 - G(-\hat{\eta}_i)^\lambda$, respectivamente. Se o modelo estiver correto esses resíduos têm uma distribuição Normal padrão ([DUNN; SMYTH, 1996](#)).

Os valores de $\hat{\mu}_i$ e $\hat{\eta}_i$ são calculados considerando as médias *a posteriori* de $\boldsymbol{\beta}$ como estimador *plug-in*.

Para avaliar a distribuição dos resíduos já que não é conhecida, [Atkinson \(1985\)](#) sugere adicionar envelopes simulados para os gráficos de probabilidades normais. Se alguns pontos cair fora do envelope, então há evidências contra a adequação do modelo.

ESTUDO DE SIMULAÇÃO

Neste capítulo é apresentado um estudo de simulação desenvolvido para avaliar o desempenho de funções de ligação assimétricas para dados desbalanceados em comparação com os métodos propostos por [Firth \(1993\)](#) e [King e Zeng \(2001\)](#) incluindo a regressão Logística comum (Logística). Usamos as ligações potência e reversa de potência para encontrar uma solução para o problema de desbalanceamento de dados em modelos de regressão binária. O objetivo é avaliar o comportamento dos diferentes ligações estudados e métodos comumente propostos através do viés, raiz do erro quadrático médio e as estimativas para diferentes valores dos parâmetros e tamanhos de amostra.

O método proposto por [Firth \(1993\)](#) é nomeado aqui como LogisticF e método proposto por [King e Zeng \(2001\)](#) é nomeado aqui como LogisticKZ.

5.1 Simulação de dados desbalanceado

Os dados desbalanceados foram gerados a partir de uma distribuição de potência Cauchy com coeficientes de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ associado com a covariável $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$ e o parâmetro de forma λ como descrito a seguir:

$$Y_i \sim \text{Bernoulli}(u_i),$$

em que

$$\mu_i = \left(\frac{1}{\pi} \arctan(\beta_1 + x_{i2}\beta_2) + \frac{1}{2} \right)^\lambda.$$

A criação de covariável foi a partir dos valores gerados da v.a $x_{i2} \sim U(-3, 3)$ e os coeficientes de regressão foram fixados com os valores $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0, 1)^\top$, de forma análoga a [Lemonte e Bazán \(2018\)](#). Com esta especificação, o estudo de simulação foi feita considerando os seguintes 12 cenários: 3 tamanhos de amostra $n = (500, 5000, 20000)^\top$ e 4 valores do parâmetros de assimetria $\boldsymbol{\lambda} = (4, 2, \frac{1}{2}, \frac{1}{4})^\top$. Para cada caso consideramos 100 réplicas.

Os valores da amostra foram selecionados pensando em grandes quantidades de volumes de dados na prática. Todos os modelos foram ajustados e o desempenho comparativo foi determinado com base na medida do viés e raiz do erro quadrático médio (RMSE) das estimativas, que são definidas, respectivamente, por:

$$\text{Bias}(\hat{\beta}_j) = \frac{1}{R} \sum_{r=1}^R (\hat{\beta}_j^{(r)} - \beta_j), \quad \text{RSME}(\hat{\beta}_j) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\beta}_j^{(r)} - \beta_j)^2}, \quad j = 1, 2,$$

em que R é o número de réplicas na simulação e $\hat{\beta}_j^{(r)}$ é a média *a posteriori* do parâmetro β_j na réplica r .

Para estimar os parâmetros por meio do método de correção de viés de regressão logística de King e Zeng (2001) foi utilizado uma função chamada de `Relogit` desenvolvida por Tomz *et al.* (2003) que está definida no pacote `Zelig` no programa R. Mais detalhes são fornecidos em (CHOIRAT *et al.*, 2018) e (KOSUKE; KING; LAU, 2007). O procedimento `relogit` estima o mesmo modelo que a regressão logística, mas as estimativas são corrigidas para o viés que ocorre quando a amostra é pequena ou os eventos observados são raros (ou seja, se a variável dependente tiver mais 0's do que 1's ou o caso inverso). Para o método de Firth (1993) foi utilizado o pacote implementado no R chamado de `logistf`, mais detalhes são fornecidos em Heinze *et al.* (2013) e Heinze e Schemper (2002). Ambos os métodos estão desenvolvidos sob abordagem clássica.

Por outro lado, a estimação Bayesiana para os modelos com função de ligação potência logística e potência Cauchy, foram desenvolvidos e implementados no linguagem Stan (CARPENTER *et al.*, 2017) por meio de Python usando o pacote `Pystan` (TEAM, 2017).

Na simulação, os valores de λ foram escolhidos para obter diferentes graus de desbalanceamento (proporção de uns). Neste caso, $\mathbf{p} = (0.20, 0.34, 0.67, 0.80)^\top$, não foram considerados casos mais extremos pois o foco de este trabalho é lidar com raridade relativa.

5.2 Resultados

Os valores das medidas da estimativa média dos coeficientes, o valor da raiz do erro quadrático médio (RMSE) e o viés, estimados pelos diferentes métodos nos diferentes cenários são apresentados nas Tabelas: 11 e 13 quando o desbalanceamento é menor a 50%, 12 e 14 quando o desbalanceamento é superior a 50%. Além disso, como resumo e ilustração, nas Figuras 25, 26, 27 e 28 são mostradas as comparações do RMSE para cada tipo de desbalanceamento considerando apenas dois ligações de potência, porque os resultados considerando outras ligações de potência e reversa de potência são similares.

Na Tabela 11 apresentamos os resultados para o parâmetro β_1 com desbalanceamento menor a 50%, isto é, $\mathbf{p} = (0.20, 0.34)^\top$, estes resultados mostram que:

- O modelo de regressão logística comum apresenta maior viés e maior valor RMSE que os outros modelos, o que significa que não é um modelo adequado quando os dados são desbalanceados.
- Pode ser visto que os métodos proposto por [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) apresentam melhor estimativa que a regressão logística comum, mas à medida que o tamanho da amostra aumenta, a diferença não é significativa em relação ao RMSE e viés.
- Por outro lado, os modelos de regressão binária com função de ligação potência e reversa de potência apresentam melhor desempenho, sendo o modelo potência Cauchy que supera a os outros pois à medida que o tamanho da amostra aumenta, o RMSE é aproximadamente zero.

Tabela 11 – Estimativa de β_1 com diferentes métodos para dados desbalanceados com valores de $\lambda = (4, 2)^T$ e diferentes valores de n

Método	n	p = 0.20			p = 0.34		
		Estimate	Bias	RSME	Estimate	Bias	RSME
Logística	500	-2.457	-2.457	2.471	-1.168	-1.168	1.178
	5000	-2.454	-2.454	2.454	-1.171	-1.171	1.172
	20000	-2.429	-2.429	2.429	-1.172	-1.172	1.172
LogisticKZ	500	-2.377	-2.377	2.383	-1.144	-1.144	1.147
	5000	-2.423	-2.423	2.424	-1.164	-1.164	1.164
	20000	-2.454	-2.454	2.454	-1.175	-1.175	1.175
LogisticF	500	-2.369	-2.369	2.376	-1.147	-1.147	1.153
	5000	-2.425	-2.425	2.425	-1.163	-1.163	1.164
	20000	-2.446	-2.446	2.446	-1.173	-1.173	1.173
PL	500	0.635	0.635	0.637	0.557	0.557	0.561
	5000	0.406	0.406	0.406	0.130	0.130	0.187
	20000	-0.389	-0.389	0.399	0.124	0.124	0.182
PP	500	-1.008	-1.008	1.017	-1.080	-1.080	1.236
	5000	0.692	0.692	0.825	0.582	0.582	0.615
	20000	0.501	0.501	0.501	0.852	0.852	0.525
PLaplace	500	-0.190	-0.190	1.064	0.031	0.031	0.425
	5000	0.150	0.150	0.185	0.048	0.048	0.103
	20000	0.148	0.148	0.162	0.042	0.042	0.057
PC	500	-0.004	-0.004	0.067	0.029	0.029	0.073
	5000	0.021	0.021	0.056	-0.013	-0.013	0.040
	20000	-0.001	-0.001	0.023	0.014	0.014	0.018
RPL	500	2.609	1.609	1.650	1.946	0.946	1.065
	5000	1.900	0.900	0.909	1.723	0.723	0.735
	20000	1.879	0.879	0.881	1.685	0.685	0.688
RPP	500	1.411	0.411	0.422	0.914	-0.087	0.177
	5000	1.333	0.333	0.370	1.021	0.021	0.105
	20000	1.326	0.326	0.326	1.007	0.007	0.066
RPLaplace	500	2.076	1.076	1.116	1.446	0.446	0.594
	5000	1.542	0.542	0.563	1.194	0.194	0.210
	20000	1.386	0.386	0.400	1.171	0.171	0.175
RPC	500	1.872	0.872	1.223	0.376	-0.624	0.719
	5000	1.130	0.130	1.137	1.697	0.697	0.700
	20000	1.102	0.102	0.820	1.554	0.554	0.560

Tabela 12 – Estimativa de β_1 com diferentes métodos para dados desbalanceados com valores de $\lambda = (0.5, 0.25)^\top$ e diferentes valores de n

Método	n	p = 0.67			p = 0.80		
		Estimate	Bias	RSME	Estimate	Bias	RSME
Logística	500	0.966	0.966	0.971	1.765	1.765	1.771
	5000	0.934	0.934	0.934	1.748	1.748	1.748
	20000	0.936	0.936	0.936	1.731	1.731	1.731
LogisticKZ	500	0.910	0.910	0.911	1.729	1.729	1.731
	5000	0.940	0.940	0.941	1.736	1.736	1.736
	20000	0.938	0.938	0.938	1.742	1.742	1.742
LogisticF	500	0.900	0.900	0.906	1.712	1.712	1.716
	5000	0.931	0.931	0.932	1.735	1.735	1.736
	20000	0.936	0.936	0.937	1.740	1.740	1.740
PL	500	-0.540	-0.540	0.571	0.519	0.519	0.526
	5000	-0.187	-0.187	0.197	-0.289	-0.289	0.399
	20000	-0.174	-0.174	0.176	-0.392	-0.392	0.394
PP	500	-0.598	-0.598	0.625	0.434	0.434	0.470
	5000	-0.430	-0.430	0.514	0.084	0.084	0.209
	20000	-0.502	-0.502	0.500	-0.219	-0.219	0.190
PLaplace	500	-0.393	-0.393	0.660	0.363	0.363	0.729
	5000	0.657	0.657	0.657	0.545	0.545	0.640
	20000	0.416	0.416	0.417	0.477	0.477	0.477
PC	500	-0.295	-0.295	0.308	0.381	0.381	0.534
	5000	-0.022	-0.022	0.041	-0.071	-0.071	0.188
	20000	-0.003	-0.003	0.041	-0.044	-0.044	0.085
RPL	500	0.921	0.921	0.922	1.360	1.360	1.542
	5000	-0.353	-0.353	0.569	-0.430	-0.430	0.533
	20000	-0.403	-0.403	0.480	-0.753	-0.753	0.788
RPP	500	0.858	0.858	0.911	1.474	1.474	1.492
	5000	0.652	0.652	0.712	1.250	1.250	1.380
	20000	-0.634	-0.634	0.656	-0.287	-0.287	0.329
RPLaplace	500	1.024	1.024	1.027	1.883	1.883	1.885
	5000	-0.035	-0.035	0.080	-0.061	-0.061	0.105
	20000	-0.031	-0.031	0.044	-0.062	-0.062	0.069
RPC	500	-0.101	-0.101	0.565	1.583	1.583	1.629
	5000	-0.027	-0.027	0.083	-0.199	-0.199	0.246
	20000	-0.039	-0.039	0.057	-0.153	-0.153	0.170

Na [Tabela 12](#) apresentamos os resultados para o parâmetro β_1 com desbalanceamento maior a 50%, isto é, $p = (0.67, 0.80)^\top$. Estes resultados mostram que os métodos de [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) e o modelo logístico possuem o valor de RMSE muito menor do que RMSE para desbalanceamento inferior a 50%, sendo os métodos de [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) melhores do que a regressão logística para amostras pequenas. Porém, os modelos com função de ligação potência e reversa de potência superam de um modo geral, pois têm estimativas aproximadamente igual ao valor verdadeiro e o modelo PC apresenta um viés e RMSE aproximadamente zero.

Os resultados na [Tabela 13](#) para o coeficiente de regressão β_2 com desbalanceamento menor a 50%, isto é, $p = (0.20, 0.34)^\top$, mostra que:

Tabela 13 – Estimativa de β_2 com diferentes métodos para dados desbalanceados com valores de $\lambda = (4, 2)^\top$ e diferentes valores de n

Método	n	p = 0.20			p = 0.34		
		Estimate	Bias	RSME	Estimate	Bias	RSME
Logística	500	1.268	0.268	0.319	1.127	0.127	0.161
	5000	1.254	0.254	0.257	1.094	0.094	0.100
	20000	1.240	0.240	0.243	1.101	0.101	0.100
LogisticKZ	500	1.185	0.185	0.234	1.065	0.065	0.103
	5000	1.229	0.229	0.235	1.093	0.093	0.099
	20000	1.249	0.249	0.250	1.105	0.105	0.107
LogisticF	500	1.180	0.180	0.209	1.054	0.054	0.090
	5000	1.229	0.229	0.232	1.089	0.089	0.096
	20000	1.244	0.244	0.244	1.103	0.103	0.104
PL	500	0.771	-0.229	0.230	0.812	-0.188	0.189
	5000	0.858	-0.142	0.142	0.904	-0.096	0.110
	20000	0.906	-0.094	0.097	0.913	-0.087	0.087
PP	500	0.759	-0.241	0.210	0.390	-0.610	0.567
	5000	0.838	-0.162	0.162	0.524	-0.476	0.522
	20000	0.852	-0.148	0.149	0.738	-0.262	0.284
PLaplace	500	0.688	-0.313	0.429	0.667	-0.333	0.346
	5000	0.602	-0.398	0.398	0.613	-0.388	0.388
	20000	0.596	-0.404	0.404	0.607	-0.393	0.393
PC	500	1.222	0.222	0.234	1.102	0.102	0.108
	5000	1.049	0.049	0.062	1.009	0.009	0.016
	20000	1.010	0.010	0.014	1.009	0.009	0.008
RPL	500	2.609	1.609	1.650	1.946	0.946	1.065
	5000	1.900	0.900	0.909	1.723	0.723	0.735
	20000	1.879	0.879	0.881	1.685	0.685	0.688
RPP	500	1.326	0.326	0.326	0.914	-0.087	0.177
	5000	1.277	0.277	0.281	1.021	0.021	0.105
	20000	1.079	0.079	0.150	1.007	0.007	0.066
RPLaplace	500	2.076	1.076	1.116	1.446	0.446	0.594
	5000	1.542	0.542	0.563	1.194	0.194	0.210
	20000	1.386	0.386	0.400	1.171	0.171	0.175
RPC	500	1.872	0.872	1.223	0.376	-0.624	0.719
	5000	1.130	0.130	1.137	1.697	0.697	0.700
	20000	1.886	0.886	0.911	1.554	0.554	0.560

- Os métodos de [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) apresentam menor viés e menor RMSE que a regressão logística, mas à medida que o tamanho da amostra aumenta, a diferença não é significativa.
- Os modelos de regressão com função de ligação potência e reversa de potência, apresentam melhor desempenho, sendo a potência Cauchy o modelo que supera a os outros pois à medida que o tamanho da amostra aumenta o RMSE é aproximadamente zero.

Os resultados para β_2 apresentados na Tabela 14 com desbalanceamento maior a 50%, isto é, $p = (0.67, 0.80)^\top$ mostram de um modo geral, que os modelos com funções de ligação assimétrica apresentam melhor estimativa, menor viés e menor RSME do que os métodos LogisticKZ e LogisticF, sendo o modelo com função de ligação PC com RSME aproximadamente 0

Tabela 14 – Estimativa de β_2 com diferentes métodos para dados desbalanceados com valores de $\lambda = (0.5, 0.25)^\top$ e diferentes valores de n

Método	n	p = 0.67			p = 0.80		
		Estimate	Bias	RSME	Estimate	Bias	RSME
Logística	500	0.713	-0.287	0.303	0.633	-0.367	0.375
	5000	0.703	-0.297	0.298	0.607	-0.393	0.394
	20000	0.701	-0.299	0.300	0.608	-0.392	0.392
LogisticKZ	500	0.669	-0.331	0.338	0.616	-0.384	0.394
	5000	0.714	-0.286	0.287	0.613	-0.387	0.388
	20000	0.705	-0.295	0.295	0.618	-0.382	0.382
LogisticF	500	0.665	-0.335	0.342	0.606	-0.394	0.402
	5000	0.713	-0.287	0.288	0.612	-0.388	0.389
	20000	0.705	-0.295	0.295	0.617	-0.383	0.383
PL	500	1.109	0.109	0.119	0.888	-0.112	0.212
	5000	0.923	-0.077	0.082	0.940	-0.060	0.107
	20000	0.951	-0.049	0.064	0.980	-0.020	0.061
PP	500	0.567	-0.433	0.440	0.443	-0.557	0.559
	5000	0.595	-0.405	0.409	0.499	-0.502	0.503
	20000	0.607	-0.393	0.397	0.547	-0.453	0.454
PLaplace	500	0.815	-0.185	0.271	0.647	-0.353	0.382
	5000	0.754	-0.246	0.246	0.573	-0.427	0.432
	20000	0.755	-0.246	0.250	0.616	-0.384	0.384
PC	500	1.384	0.384	0.388	1.333	0.333	0.347
	5000	1.037	0.037	0.064	1.024	0.024	0.101
	20000	0.998	-0.002	0.020	1.005	0.005	0.047
RPL	500	0.873	-0.128	0.323	1.587	0.587	0.760
	5000	0.548	-0.453	0.455	0.386	-0.615	0.615
	20000	0.635	-0.365	0.366	0.361	-0.640	0.640
RPP	500	0.584	-0.416	0.420	0.383	-0.617	0.620
	5000	0.590	-0.410	0.416	0.526	-0.474	0.479
	20000	0.681	-0.319	0.319	0.529	-0.471	0.471
RPLaplace	500	0.755	-0.245	0.508	2.017	1.017	1.177
	5000	0.370	-0.630	0.630	0.249	-0.751	0.751
	20000	0.565	-0.435	0.435	0.245	-0.755	0.755
RPC	500	0.628	-0.372	0.383	1.104	0.104	1.247
	5000	0.548	-0.452	0.452	0.360	-0.640	0.640
	20000	0.644	-0.356	0.356	0.453	-0.547	0.547

e uma estimativa mais próximo do valor verdadeiro. Além disso, pode ser visto que quando o tamanho da amostra cresce não existe diferença significativa entre os valores estimados pelos métodos LogisticKZ, LogisticF e Logística e, conforme a proporção de uns cresce essas estimativas estão mais próximos do verdadeiro valor. Por outro lado pode-se concluir que quando $n = 500$ a estimação usando os métodos LogisticKZ e LogisticF apresentam menor RSME do que a estimação por meio da regressão logística. Porém, quando o tamanho de amostra é muito grande os valores das estimativas são parecidas, assim eles não apresentam diferença significativa.

Nas Figuras 25 a 28 são apresentados os gráficos de RMSE das estimativas de β_1 e β_2 para diferentes proporções de uns usando dois ligações potência em comparação com os métodos LogisticKZ e LogisticF incluindo a Logística comum. Podemos observar que:

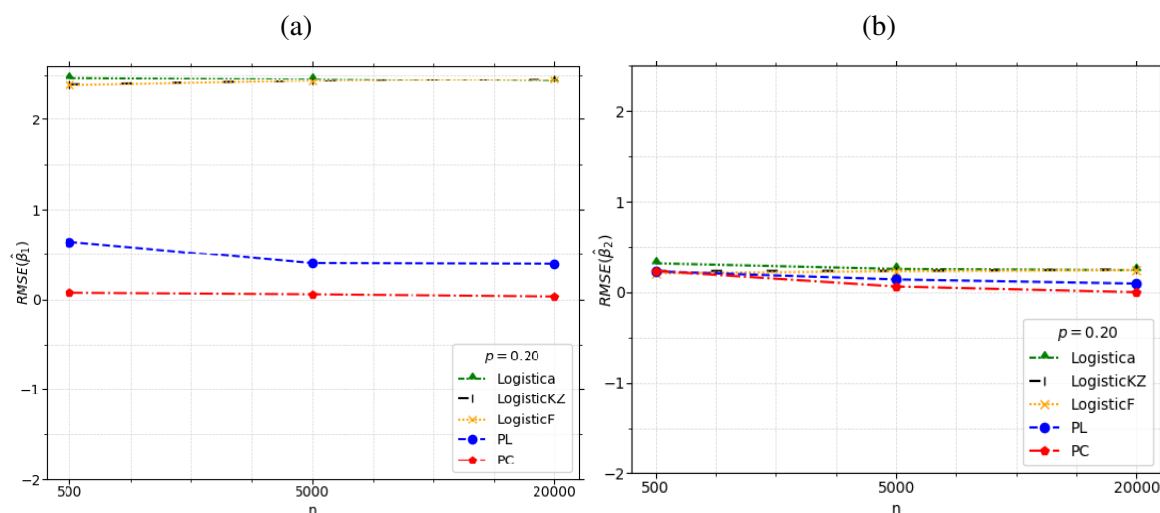


Figura 25 – RMSE para β_1 (a) e β_2 (b) com diferentes tamanhos de amostra (500,5000,20000) e diferentes métodos de estimação: Logística (verde), LogisticKZ (preta), LogisticF (laranja), PL (azul) e PC (vermelha), quando $\lambda = 4$

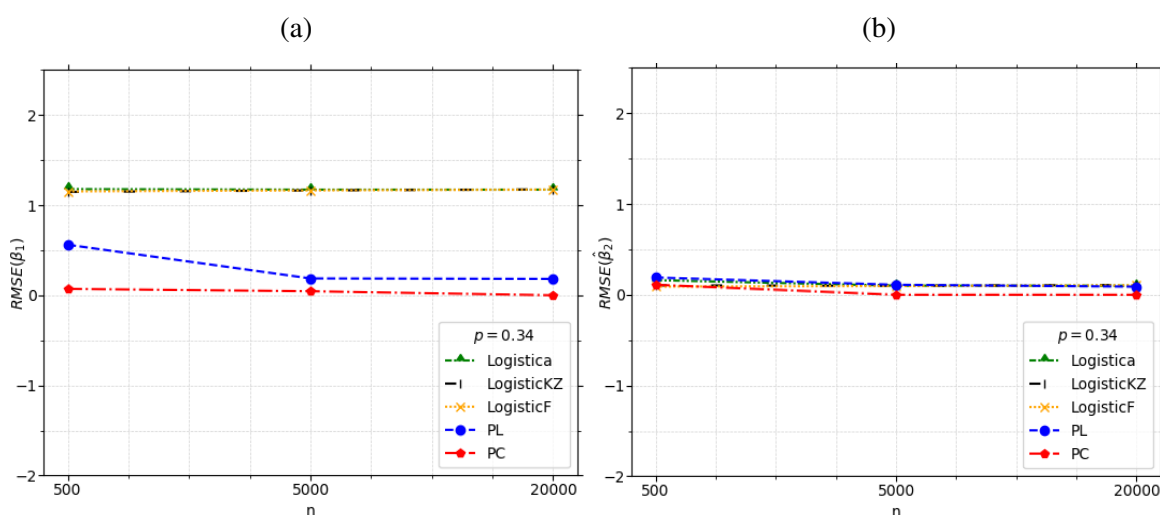


Figura 26 – RMSE para β_1 (a) e β_2 (b) com diferentes tamanhos de amostra (500,5000,20000) e diferentes métodos de estimação: Logística (verde), LogisticKZ (preta), LogisticF (laranja), PL (azul) e PC (vermelha), quando $\lambda = 2$

- Quando o coeficiente β_1 é estimado para todas as proporções de uns, à medida que o tamanho da amostra cresce, os valores de RMSE com os métodos Logística, LogisticKZ e LogisticF se aproximam entre eles. No entanto, eles estão distantes de zero. Por outro lado, para o coeficiente β_2 , as linhas estão próximos de zero para todos os métodos. Por outro lado, quando os coeficientes β_1 e β_2 são estimados com funções de ligação potência, as linhas de RMSE aproximam-se de zero de maneira geral.
- Quando a proporção $p > 0.5$ e β_2 é estimado, as linhas de RMSE dos modelos com função de ligação potência se aproximam entre elas à medida que o tamanho da amostra cresce.

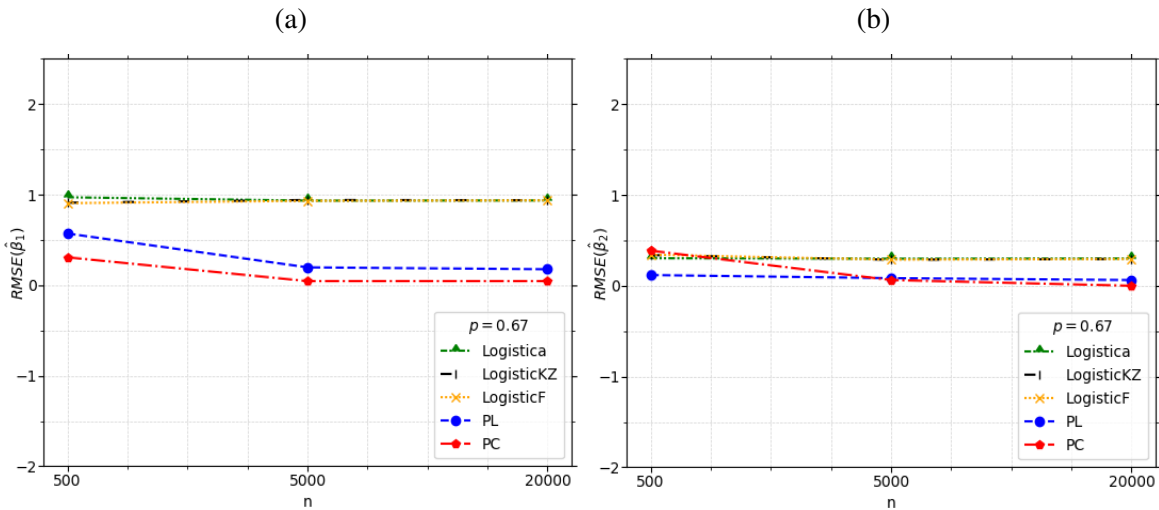


Figura 27 – RMSE para β_1 (a) e β_2 (b) com diferentes tamanhos de amostra (500,5000,20000) e diferentes métodos de estimação: Logística (verde), LogisticKZ (preta), LogisticF (laranja), PL (azul) e PC (vermelha), quando $\lambda = 0.5$

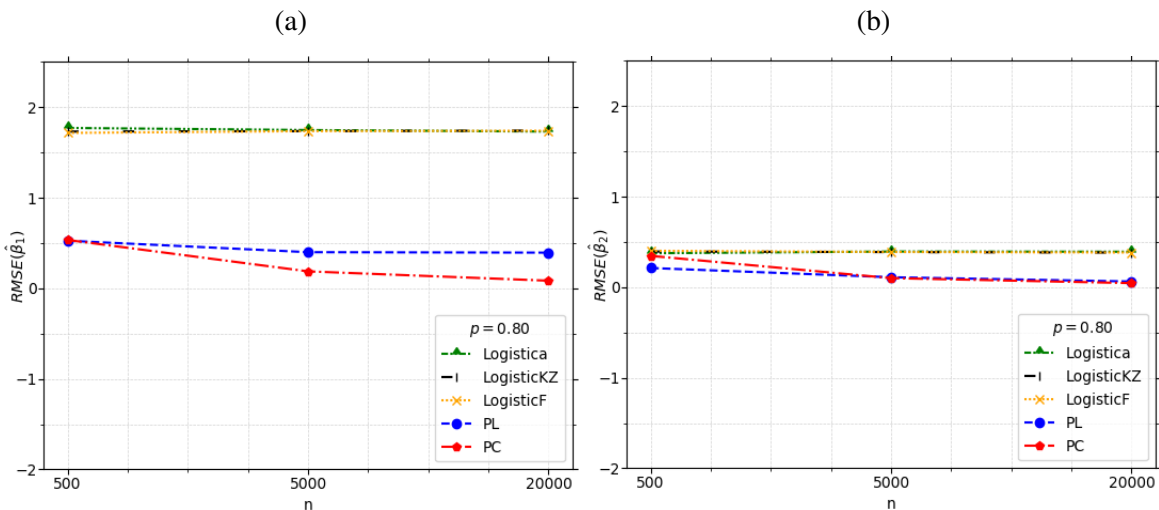


Figura 28 – RMSE para β_1 (a) e β_2 (b) com diferentes tamanhos de amostra (500,5000,20000) e diferentes métodos de estimação: Logística (verde), LogisticKZ (preta), LogisticF (laranja), PL (azul) e PC (vermelha), quando $\lambda = 0.25$

Em todos os cenários, pode se concluir que os modelos com função de ligação potência superam os outros métodos para diferentes proporção de uns e, as estimativas de β_2 apresentam menor valor RSME do que para β_1 para todos os métodos considerados.

Como visto nos capítulos anteriores, os modelos com funções de ligação assimétricas incluem um parâmetro adicional, os valores estimados deste parâmetro para diferentes modelos, são mostrados nas Tabelas 22 a 25 no anexo.

APLICAÇÃO

Para ilustrar o desempenho dos modelos apresentados na [Seção 3.2](#), isto é, modelos com função de ligação potência e reversa de potência, neste capítulo uma aplicação foi desenvolvida considerando um conjunto de dados reais. Além disso, é apresentada a descrição e análise dos dados.

A estimação dos parâmetros para cada modelo é feita sob uma abordagem Bayesiana apresentada na [Seção 4.1](#) e a escolha do modelo com melhor ajuste está baseada nos critérios de seleção de modelos apresentados na [Seção 4.2](#), bem como a análise de resíduos. Por outro lado, é apresentada a avaliação de medidas preditivas do melhor modelo proposto em comparação com os métodos de [King e Zeng \(2001\)](#) e [Firth \(1993\)](#).

6.1 Apresentação dos dados

Como aplicação de ligações potência e reversa de potência, neste estudo analisamos os dados relacionados à qualidade do vinho branco variante de "vinho verde" da região do Minho de Portugal. Esses dados foram analisados e apresentados por [Cortez et al. \(2009\)](#) e estão disponíveis no repositório da UCI [Dua e Taniskidou \(2017\)](#). A qualidade do vinho, conforme determinada pelos conhecedores de vinhos, está relacionada a onze atributos (covariáveis) químicos que servem como preditores em potencial. A lista inclui: acidez fixa, acidez volátil, ácido cítrico, açúcar residual, cloretos, dióxido de enxofre livre, dióxido de enxofre total, densidade, pH, sulfatos e álcool. Os dados foram padronizados previamente como em [Cortez et al. \(2009\)](#). Cada um dos atributos é definido na [Tabela 15](#) e descrito como segue:

Descrição dos atributos

- *Acidez fixa*: uma medida da concentração total de ácidos ti-tratáveis e íons de hidrogênio livres presentes no vinho. Teoricamente, ter uma baixa acidez ou muito acidez pode levar

Tabela 15 – Atributos do vinho branco português, completo

Atributo ¹	Mín.	Media	Máx.	Assimetria	Kurtosis	Desv. padrão
Acidez fixa (g(ácido tartárico)/dm ³)	3.80	6.86	14.20	0.65	5.17	0.84
Acidez volátil (g(ácido acético)/dm ³)	0.08	0.28	1.10	1.58	8.13	0.10
Ácido cítrico (g/dm ³)	0.00	0.33	1.66	1.28	9.17	0.12
Açúcar residual (g/dm ³)	0.60	6.39	65.80	1.08	6.47	5.07
Cloretos (g(cloreto de sódio)/dm ³)	0.01	0.05	0.35	5.02	40.53	0.02
Dióxido de enxofre livre (mg/dm ³)	2.00	35.31	289.00	1.41	14.45	17.01
Dióxido de enxofre total (mg/dm ³)	9.00	138.36	440.00	0.39	3.57	42.50
Densidade (g/cm ³)	0.99	0.99	1.04	0.95	12.51	0.00
pH	2.72	3.19	3.82	0.46	3.53	0.15
Sulfatos (g(sulfato de potássio)/dm ³)	0.22	0.49	1.08	0.98	4.59	0.11
álcool (vol.%)	8.00	10.51	14.20	0.49	2.30	1.23
Qualidade	3.00	5.88	9.00	0.16	3.21	0.89

¹ g/dm³: gramas por decímetro cúbico, g/cm³: gramas por centímetro cúbico

a um vinho desagradável ou azedo. Estes ácidos ocorrem naturalmente nas uvas ou são criados por meio do processo de fermentação.

- *Acidez volátil*: uma medida de ácidos destiláveis a vapor presentes em um vinho. Em teoria, nossos paladares são bastante sensíveis à presença de ácidos voláteis e, por essa razão, um bom vinho deve manter suas concentrações as mais baixas possíveis.
- *Ácido cítrico*: encontrado em pequenas quantidades, ácido cítrico pode adicionar “frescura” e sabor aos vinhos.
- *Açúcar residual*: a quantidade de açúcar que permanece após a fermentação parar, é raro encontrar vinhos com menos de 1 grama/litro e vinhos com mais de 45 gramas/litro são considerados doces. Em teoria, o açúcar residual pode ajudar os vinhos a envelhecer bem.
- *Cloretos*: a quantidade de sal no vinho.
- *Dióxido de enxofre livre*: a forma livre de SO₂ existente em equilíbrio entre o SO₂ molecular (como gás dissolvido) e o íon bissulfito impedimos o crescimento microbiano e a oxidação do vinho.
- *Dióxido de enxofre total*: quantidade de formas livres e encadernadas de S₀2 em baixas concentrações. O SO₂ é quase indetectável no vinho, mas nas concentrações de SO₂ livre acima de 50 ppm, o SO₂ se torna evidente no nariz e no sabor do vinho.
- *Densidade*: a medida da densidade do vinho é próxima da densidade da água, dependendo do percentual de álcool e teor de açúcar.

- *PH*: descreve como o vinho é ácido ou básico numa escala de 0 (muito ácido) a 14 (muito básico). A maioria dos vinhos tem entre 3-4 na escala de pH.
- *Álcool*: a percentagem de álcool presente no vinho.
- *Qualidade*: é a variável de saída (baseada em dados sensoriais) com pontuação entre 0 e 10.

O conjunto de dados contém 4898 amostras de vinhos e os dados foram coletados de maio de 2004 a fevereiro de 2007. Cada uma das amostras foi avaliada por um mínimo de três provadores, por meio de prova cega, em uma escala de qualidade que varia de zero (muito ruim) a 10 (excelente). No conjunto de dados original, a qualidade é uma variável quantitativa representada como números inteiros $Z = \{0, \dots, 10\}$. Seguindo [Lemionet, Liu e Zhou \(2015\)](#), para o propósito desta investigação, nós consideramos o modelo $\mu = P(Z > 6 | \mathbf{X})$ transformando a variável de classe ordinal em uma classe binária desbalanceada definida por

$$Y = I_{\{Z > 6\}}(Z) = \begin{cases} 1, & Z > 6 \quad (\text{qualidade bom}) \\ 0, & Z \leq 6 \quad (\text{qualidade ruim}) \end{cases}$$

A proporção de uns é de 21%, consequentemente são dados desbalanceados, como é mostrado na [Figura 29](#).

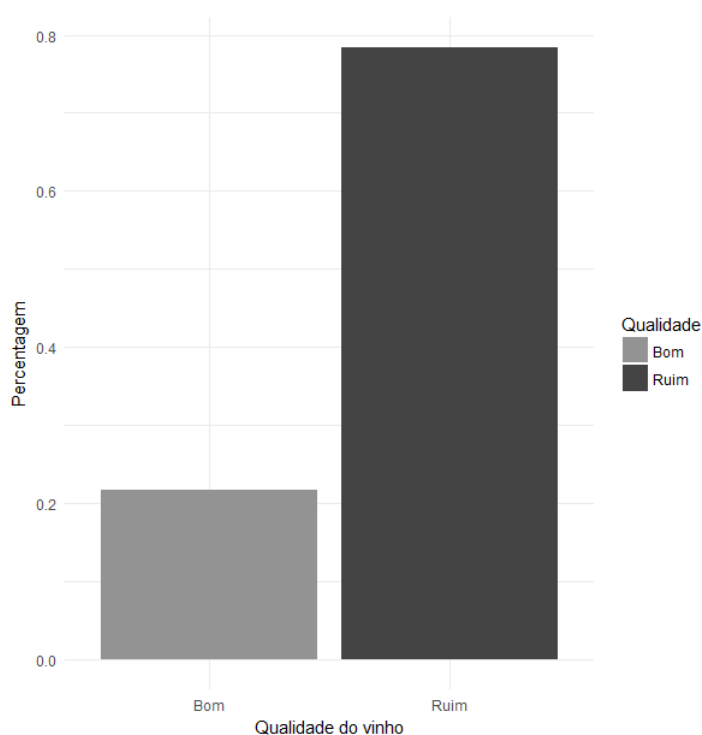


Figura 29 – Qualidade do vinho branco.

6.2 Análise preliminar

Inicialmente alguns modelos com ligação simétrica (Logística, Probit, Laplace, Cauchy e t-Student) e modelos com função de ligação assimétrica (PI, PP, PLAPLACE, PT, PC, PLR, PPR, PLAPLACER, PTR e PCR) foram ajustados ao conjunto de dados considerando todas as covariáveis (modelo completo). Após os ajustes, as covariáveis *ácido cítrico* e *dióxido de enxofre total* foram as únicas que pareciam não influenciar a variável resposta na maioria dos modelos com ligação simétrica e nos modelos potência e reversa de potência, então foram calculados os intervalos (regiões) de maior densidade *a posteriori* (o intervalo HPD de 95% que significa em inglês, highest posterior density) para estas covariáveis. Os resultados são mostrados na [Figura 30](#) e [Figura 31](#), respectivamente. Nessas figuras os pontos no meio representam as estimativas *a posteriori* do coeficiente de regressão associado à covariável. Podemos observar que o intervalo HPD de 95% para os coeficientes associados às covariáveis *ácido cítrico* e *dióxido de enxofre total* incluem o ponto zero para os diferentes modelos. Além disso, a estimativa para estes coeficientes é aproximadamente zero, portanto poderiam não ser significativas na explicação da qualidade do vinho branco, por isso consideramos que devem ser retiradas para análise posterior (modelo reduzido).

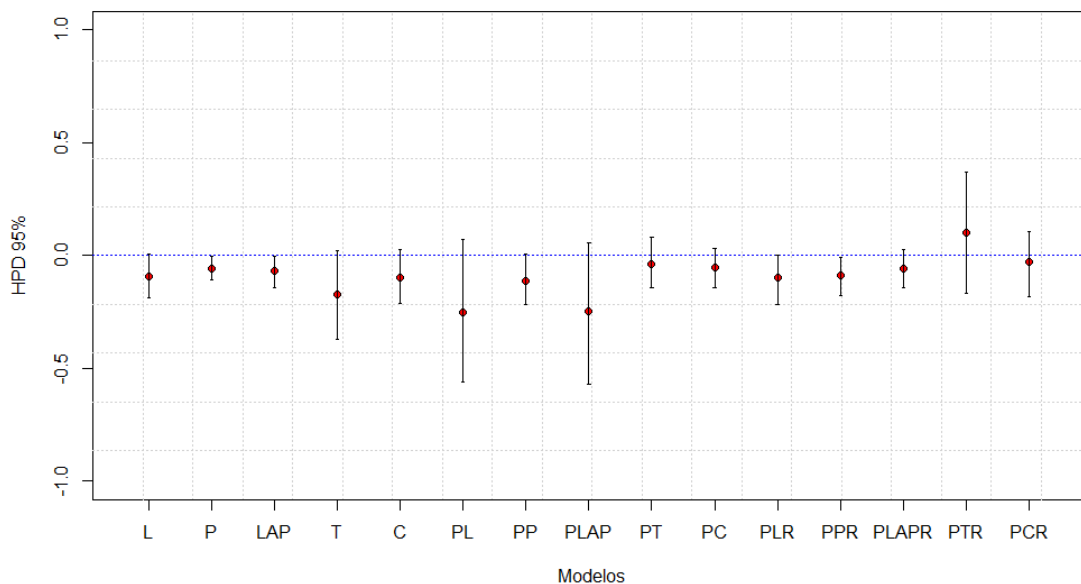


Figura 30 – Intervalo HPD 95% para o coeficiente de regressão associado à covariável *ácido cítrico*, ajustando com os modelos L, P, LAPLACE, T, C, PL,PP,PLAPLACE, PT, PC, PLR,PPR, PLAPLACER, PTR, PCR.

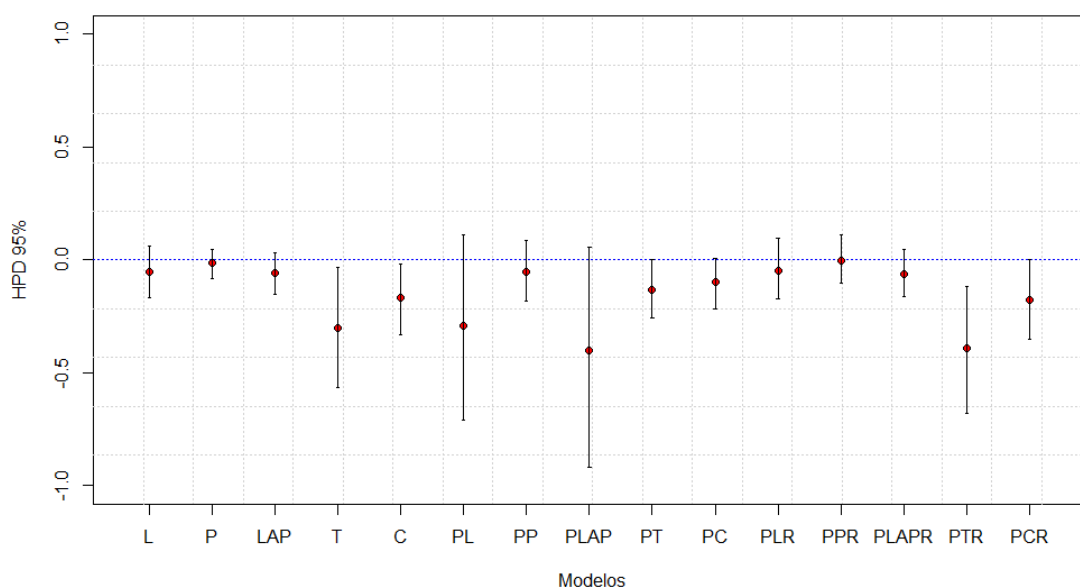


Figura 31 – Intervalo HPD 95% para o coeficiente de regressão associado à covariável *dióxido de enxofre total*, ajustando com os modelos L, P, LAPLACE, T, C, PL,PP,PLAPLACE, PT, PC, PLR,PPR, PLAPLACER, PTR, PCR.

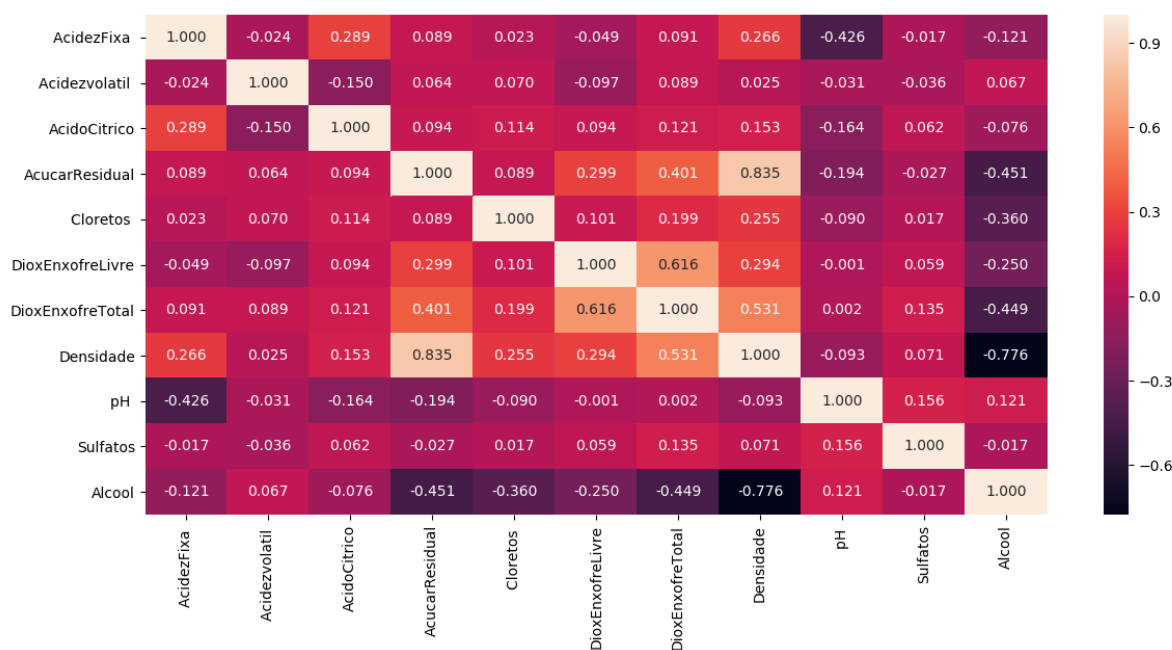


Figura 32 – Matriz de correlação das covariáveis

Por outro lado, para ver si existe algumas indicações de multicolinearidade foi calculada a matriz de correlação entre as covariáveis, o resultado é apresentado na [Figura 32](#). Além disso, foi analisado os fatores de inflação da variância (em inglês, *Variance Inflation Factors* - VIF), os resultados são mostrados na [Tabela 16](#), valores de VIF maiores do que 10 podem causar problemas na estimação dos coeficientes de regressão.

Tabela 16 – Fatores de inflação da variância - VIF

Atributo	VIF
Acidez fixa	2.432
Acidez volátil	1.136
Ácido cítrico	1.167
Açúcar residual	3.344
Cloretos	1.228
Dióxido de enxofre livre	1.782
Dióxido de enxofre total	2.239
Densidade	22.709
pH	2.033
Sulfatos	1.112
álcool	6.410

Observa-se que a maioria das variáveis explicativas, com exceção das variáveis *Densidade* e *Açúcar residual*, apresentam correlação baixa, de forma linear e segundo os valores de VIF a variável *Densidade* é a única que indica a presença de multicolinearidade. Apesar dos valores de VIF e da matriz de correlação, neste trabalho, decidimos considerar os resultados baseado no intervalo HPD e a probabilidade de significância portanto a versão reduzida das covariáveis associados à qualidade de vinho é mostrada na [Tabela 17](#).

Tabela 17 – Atributos do vinho branco português, reduzido

N	Atributo	Notação
1	Acidez fixa (g(ácido tartárico)/dm ³)	X ₂
2	Acidez volátil (g(ácido acético)/dm ³)	X ₃
3	Açúcar residual (g/dm ³)	X ₄
4	Cloretos (g(cloreto de sódio)/dm ³)	X ₅
5	Dióxido de enxofre total (mg/dm ³)	X ₆
6	Densidade (g/cm ³)	X ₇
7	pH	X ₈
8	Sulfatos (g(sulfato de potássio)/dm ³)	X ₉
9	Álcool (vol.%)	X ₁₀
10	Qualidade	Y

6.3 Estimação com funções de ligação assimétrica

Os modelos de regressão binária com ligações potência e reversa de potência apresentados na [Seção 3.2](#) foram ajustados ao conjunto de dados desbalanceados sobre qualidade de vinho. A estimação dos parâmetros foi realizada sob uma abordagem Bayesiana considerando o método de estimação Monte Carlo Hamiltoniano usando o algoritmo NUTS que foi apresentado na [Seção 4.1](#). Os códigos para todos os modelos foram desenvolvidos e implementados em

linguagem Stan por meio do Python usando o pacote `Pystan`. Esta escolha foi devido ao tempo computacional, a linguagem Python é muito mais rápida do que a linguagem do programa R. Para os modelos t -Student, PT e RPT, seguindo o trabalho de [Lemonte e Bazán \(2018\)](#), inicialmente para escolha dos graus de liberdades, usou-se a função de verossimilhança perfilada para obter o valor de ν e as ligações t -Student ($\nu = 2.97$), potência t -Student ($\nu = 0,6$) e reversa de potência t -Student ($\nu = 0,6$), foram consideradas. Por outro lado, seguindo a [Ding \(2014\)](#) e com a informação acima, foi utilizada uma distribuição a priori gama informativa com os parâmetros $\nu \sim \text{gamma}(125, 200)$, ou seja, $\pi(\nu) \propto x^{125-1} \exp\{-100x\}$, esta escolha foi feita porque nós assumimos que o valor do parâmetro ν é conhecida. No entanto, para os modelos apresentados neste trabalho preferimos recomendar usar a distribuição t -Student com abordagem condicional porque a estimativa do parâmetro extra ν na abordagem não-condicional não está associada à função de ligação e não é fácil de ser interpretado.

Inicialmente consideramos a verificação da convergência baseado na estatística potencial de redução de escala (\hat{R}) de [Gelman e Rubin \(1992\)](#) para cada um dos quinze modelos (5 com ligações comuns, 5 usando ligações potência e 5 usando ligações reversa de potência). Essa convergência foi alcançada considerando uma amostra a posteriori de tamanho 5000 e retirando-se os primeiros 3000 valores no período do burnin, tendo 2000 valores para os quais considerou-se o espaçamento de tamanho 2 para evitar os efeitos da autocorrelação na amostra, tendo no final uma amostra válida da distribuição a posteriori de 1000 valores para cada caso.

A comparação dos modelos são desenvolvidos usando diferentes critérios de seleção de modelos apresentados na [Seção 4.2](#). Os resultados destes critérios é mostrado na [Tabela 18](#). Por fim, o modelo com melhor ajuste aos dados é escolhido.

Tabela 18 – Critérios de seleção de modelos para dados de qualidade de vinho branco

Modelo	pD	Dbar	lppd	pWAIC	DIC	EAIC	EBIC	IC	WAIC	LOO
Logística	9.664	4182.992	-2085.760	10.797	4192.657	4204.992	4276.455	4202.321	4193.114	4198.494
Normal	10.380	4183.726	-2086.935	9.943	4194.106	4205.726	4277.188	4204.486	4193.755	4193.737
Laplace	11.132	4186.588	-2087.868	10.906	4197.720	4208.588	4280.051	4208.852	4197.547	4197.524
t -Student($\nu=2.97$)	10.15	4176.78	-2083.31	10.22	4186.93	4198.78	4270.24	4197.08	4187.05	4187.03
Cauchy	8.908	4208.197	-2098.823	10.582	4217.105	4230.197	4301.659	4226.013	4218.809	4218.792
P-Logística	33.953	4187.472	-2087.459	13.123	4221.425	4209.472	4280.935	4255.377	4201.162	4201.322
P-Normal	8.513	4181.774	-2086.074	9.718	4190.287	4203.774	4275.236	4198.801	4191.584	4191.571
P-Laplace	32.745	4205.218	-2095.883	14.722	4237.963	4227.218	4298.681	4270.708	4221.208	4221.625
P- t -Student($\nu=0.60$)	10.289	4169.599	-2079.143	11.356	4179.888	4191.599	4263.061	4190.177	4180.997	4180.974
P-Cauchy	11.769	4172.524	-2080.636	11.296	4184.293	4194.524	4265.987	4196.063	4183.863	4183.841
RP-Logística	14.874	4180.894	-2084.585	11.839	4195.767	4202.894	4274.356	4210.641	4192.848	4192.838
RP-Normal	12.151	4187.805	-2088.714	10.477	4199.956	4209.805	4281.268	4212.107	4198.383	–*
RP-Laplace	11.545	4175.600	-2082.236	11.188	4187.144	4197.600	4269.062	4198.689	4186.847	4186.827
RP- t -Student($\nu=0.95$)	11.72	4169.87	-2078.92	12.08	4181.60	4191.87	4263.33	4193.31	4182.01	4181.98
RP-Cauchy	11.160	4169.671	-2079.015	11.688	4180.830	4191.671	4263.133	4191.990	4181.406	4181.382

* Problemas no cálculo de log-verossimilhança para a estimação do LOO

Tabela 19 – Estimação de parâmetros para modelo de regressão binária com função de ligação potência t-Student com a *priori* para ν e $\nu = 0.6$ para dados de qualidade de vinho branco

Variáveis	Parâmetro	Potência t-Student com $\nu = 0.6$			Potência t-Student com a <i>priori</i> para ν		
		Estimativa	Desvio padrão	95% HPD	Estimativa	Desvio padrão	95% HPD
Intercepto	β_1	-0.895	0.161	(-1.202; -0.579)	-0.896	0.171	(-1.217; -0.540)
Acidez fixa	β_2	0.381	0.080	(0.228; 0.534)	0.384	0.083	(0.232; 0.560)
Acidez volátil	β_3	-0.306	0.050	(-0.397; -0.207)	-0.312	0.053	(-0.420; -0.212)
Açúcar residual	β_4	1.110	0.185	(0.746; 1.470)	1.120	0.189	(0.788; 1.513)
Cloretos	β_5	-0.533	0.103	(-0.734; -0.342)	-0.548	0.111	(-0.764; -0.334)
Dióxido de enxofre livre	β_6	0.170	0.049	(0.077; 0.263)	0.172	0.049	(0.079; 0.275)
Densidade	β_7	-1.524	0.293	(-2.126; -0.980)	-1.536	0.300	(-2.123; -0.952)
pH	β_8	0.496	0.078	(0.350; 0.656)	0.498	0.083	(0.325; 0.653)
Sulfatos	β_9	0.221	0.044	(0.142; 0.313)	0.224	0.046	(0.141; 0.318)
Álcool	β_{10}	0.339	0.126	(0.114; 0.606)	0.343	0.124	(0.095; 0.551)
Assimetria	λ	1.729	0.100	(1.536; 1.926)	1.739	0.111	(1.524; 1.956)
Graus de liberdade	ν	-	-	-	0.592	0.052	(0.492; 0.690)

Tabela 20 – Estimação de parâmetros para modelo de regressão binária com função de ligação RPC para dados de qualidade de vinho branco

Variáveis	Parâmetro	Estimativa	Desvio padrão	95% HPD
Intercepto	β_1	-1.058	0.149	(-1.357; -0.775)
Acidez fixa	β_2	0.487	0.093	(0.307; 0.666)
Acidez volátil	β_3	-0.426	0.077	(-0.589; -0.293)
Açúcar residual	β_4	1.422	0.210	(1.002; 1.856)
Cloretos	β_5	-0.711	0.134	(-0.977; -0.445)
Dióxido de enxofre livre	β_6	0.217	0.057	(0.106; 0.333)
Densidade	β_7	-1.915	0.325	(-2.535; -1.286)
pH	β_8	0.634	0.082	(0.471; 0.793)
Sulfatos	β_9	0.284	0.053	(0.175; 0.385)
Álcool	β_{10}	0.503	0.181	(0.164; 0.877)
Assimetria	λ	0.488	0.049	(0.400; 0.597)

De acordo com os critérios de comparação de modelos (DIC, EAIC, EBIC, IC, WAIC e LOO), o modelo de regressão potência t-Student ($\nu = 0.6$) supera todos os outros modelos de regressão. Portanto, concluímos que o esse modelo é o melhor para descrever este conjunto de dados. As estimativas obtidas para o modelo potência t-Student ($\nu = 0.6$) com abordagem condicional e não condicional (com a *priori* para ν) são apresentadas na Tabela 19. O segundo modelo que melhor desempenho apresenta, é reversa de potência Cauchy. A estimativa dos parâmetros é apresentada na Tabela 20. Destas duas tabelas pode se ver a semelhança nas estimativas dos parâmetros além disso, as variáveis acidez volátil, cloretos e densidade, têm um efeito negativo na qualidade do vinho, enquanto as outras variáveis têm um efeito positivo.

Por outro lado, considerando os modelo que melhor se ajustam aos dados, isto é, o modelo com ligação potência t-Student com 0.6 graus de liberdade, RPC e resultados dos métodos LogisticF e LogisticKZ, assim como do modelo logístico comum, analisamos a avaliação preditiva de ambos estes modelos considerando medidas baseadas na matriz de confusão apresentada na Seção 4.2. Neste caso, usamos as medidas usuais como a taxa de boa classificação, sensibilidade, especificidade e área sob a curva (AUC). Os resultados são apresentados na Tabela 21, em que o valor **1** representa o sucesso (se a qualidade do vinho é bom) observado ou predito, e o valor **0** representa a falha (se a qualidade do vinho é ruim) observado ou predito, sendo as diagonais a frequência da predição correta.

Tabela 21 – Medidas preditivas para modelo de regressão binária Logística, LogisticF, LogisticKZ e potência t-Student para dados de qualidade de vinho branco

Modelo	Value	Predito		AUC	ACC	TPR	TNR	CSI	GS	SSI	FAITH	PDIF
		0	1									
Logística	Observado 0	3623	778	0.790	0.797	0.567	0.823	0.221	0.149	0.124	0.427	0.028
	Observado 1	215	282									
LogisticF	Observado 0	3624	778	0.790	0.797	0.569	0.823	0.221	0.150	0.124	0.428	0.028
	Observado 1	214	282									
LogisticKZ	Observado 0	3648	793	0.790	0.799	0.584	0.821	0.214	0.146	0.120	0.427	0.025
	Observado 1	190	267									
Potência	Observado 0	3558	695	0.789	0.801	0.566	0.837	0.272	0.188	0.158	0.438	0.033
t-Student	Observado 1	280	365									
RP Cauchy	Observado 0	3599	320	0.790	0.798	0.319	0.918	0.240	0.158	0.136	0.431	0.036
	Observado 1	667	312									

ACC: Acurácia, TPR: Sensibilidade, TNR: Especificidade, AUC: área abaixo da curva, CSI: critical success index, SSI: Sokal & Sneath index, FAITH: Faith index and PDIF: pattern difference

Para cada modelo, a acurácia ou taxa de boa classificação mede a proporção de qualidade de vinho bom ou ruim que é corretamente identificado como tal em relação a todos os resultados possíveis. Sensibilidade mede a proporção de qualidade de vinho bom que é classificado corretamente pelo modelo ou identificado como tal é dizer capacidade do modelo predizer $\hat{y} = \text{qualidade bom}$ dado que $y = \text{qualidade bom}$. e a especificidade mede a proporção de qualidade de vinho ruim que é corretamente identificado como tal isto é capacidade do modelo predizer $\hat{y} = \text{qualidade ruim}$ dado que $y = \text{qualidade ruim}$. Por fim, a AUC mede toda a área bidimensional abaixo da curva ROC (em inglês receiver operating characteristic curve). A análise ROC fornece

uma maneira de selecionar possíveis modelos ótimos baseados no desempenho de classificação em vários níveis limítrofes. Neste caso, um modelo com a maior área sob a curva deve ser o preferido dentre os modelos possíveis. Ao considerar as medidas descritas acima, os resultado na Tabela 21 mostra que o modelo com função de ligação potência t-Student apresenta a melhor acurácia e sensibilidade do que o modelo com função de ligação logito, mas a AUC é semelhante.

Analogamente, foi calculado os valores das novas medidas apresentadas na Seção 4.2 (CSI, GS, SSI, FAITH e PDIF). Para todas essas medidas, com exceção do PDIF, um modelo com o maior valor deve ser preferido em relação a outros modelos possíveis, pois apresenta maior similaridade entre a classificação observada e a prevista. Na Tabela 21, segundo essas medidas, o modelo com função de ligação potência t-Student ($v = 0.6$) também apresenta melhores resultados que os métodos LogisticF, LogisticKZ, modelo logístico e RPC portanto deve ser considerado como o melhor modelo para a predição de qualidade do vinho.

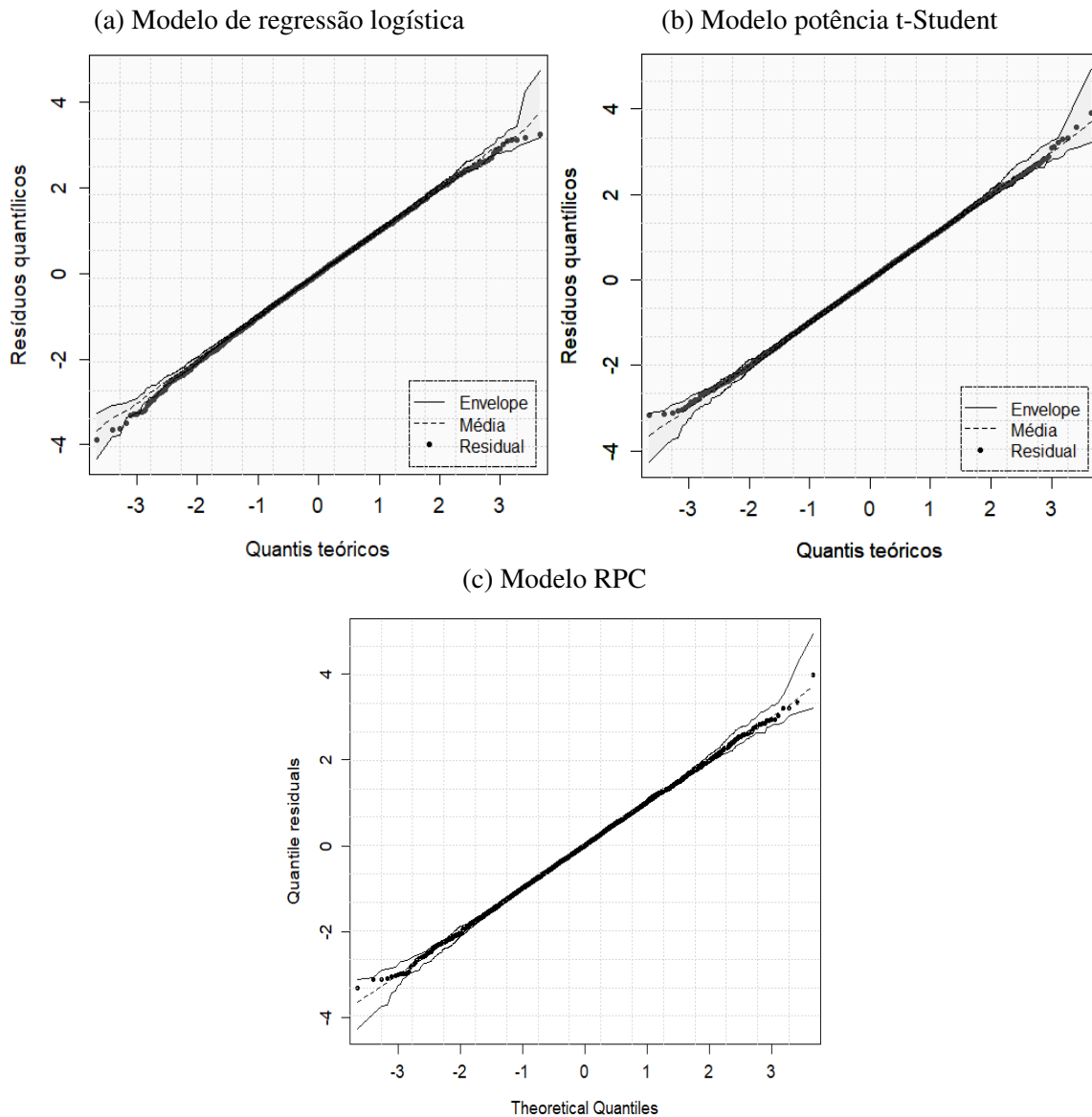


Figura 33 – Resíduos quantílicos aleatorizados normalizados para regressão logística, modelo com ligação potência t-Student e RPC em regressão binária para qualidade do vinho branco

Foi considerado também um diagnóstico do ajuste por meio da análise dos resíduos quantílicos aleatorizados normalizados, proposto por [Dunn e Smyth \(1996\)](#) apresentado na [Seção 4.2](#), com envelopes gerados.

O gráfico de envelope para o modelo de regressão logística apresentado na [Figura 33\(a\)](#) revela que este modelo não é adequado para modelar os dados, uma vez que há observações fora do envelope. Por outro lado, como mostrado na [Figura 33\(b\)](#), o uso de ligação assimétrica, neste caso a ligação potência t-Student melhora muito o ajuste em termos de adaptação de modelos para tais dados desbalanceados já que todos os pontos estão dentro do envelope semelhante ao modelo com ligação RPC [Figura 33\(c\)](#), mas segundo os critérios de comparação e medidas preditivas, o modelo com ligação potência t-Student pode ser escolhido como o melhor modelo para ajustar aos dados.

A [Figura 34](#) mostra o gráfico dos resíduos quantílicos aleatorizados normalizados para o modelo com função de ligação potência t-Student. Podemos observar que a distribuição dos resíduos segue claramente uma distribuição Normal indicando que o modelo é adequado.

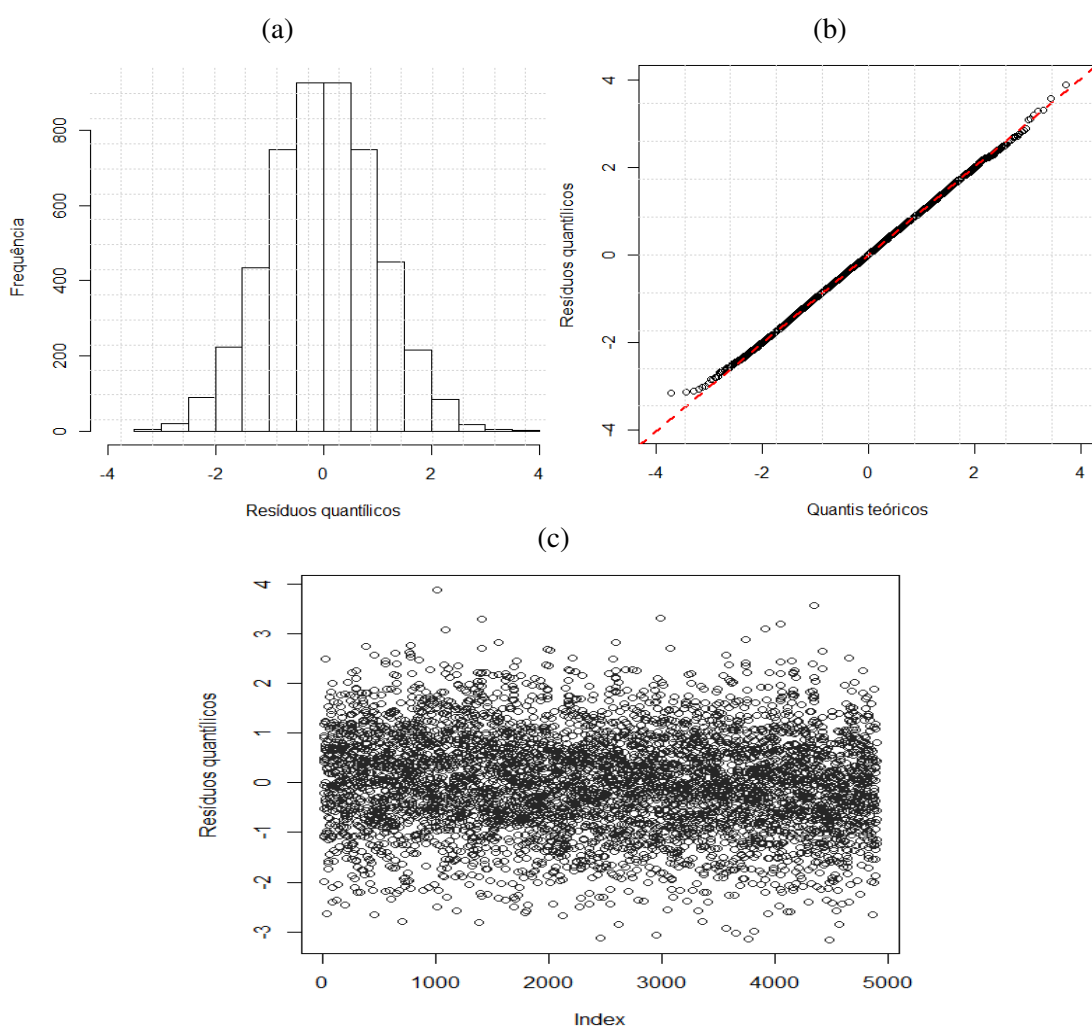


Figura 34 – Resíduos quantílicos aleatorizados normalizados para a regressão binária com ligação potência t-Student para os dados de qualidade de vinho.

Por fim, segundo a análise realizada, consideramos que o modelo com ligação potência t-Student ($\nu = 0.6$) é mais adequado para os dados de qualidade de vinho, em que as probabilidades estimadas são obtidas considerando

$$Y_i \sim \text{Bernoulli}(\hat{u}_i) \quad \forall i = 1, \dots, 4898$$

$$\hat{u}_i = \left[\frac{1}{2} + \frac{1}{2} \text{sign}(\hat{\eta}_i) \left[1 - I_{m(\hat{\eta}_i)} \left(\frac{0.6}{2}, \frac{1}{2} \right) \right] \right]^{1.73}$$

em que $\hat{\eta}_i = -0.895 + 0.38x_2 - 0.31x_3 + 1.11x_4 - 0.53x_5 + 0.17x_6 - 1.52x_7 + 0.50x_8 + 0.22x_9 + 0.34x_{10}$, x_2 =Acidez fixa, x_3 =Acidez volátil, x_4 =Açúcar residual, x_5 =Cloretos, x_6 =Dióxido de enxofre livre, x_7 =Densidade, x_8 =pH, x_9 =Sulfatos, x_{10} =Álcool, $m(z) = 0.6/(0.6 + z^2)$ e $I_x(a, b) = B(x; a, b)/B(a, b)$ é a função beta regularizada incompleta.

CONSIDERAÇÕES FINAIS

Neste trabalho foram consideradas algumas funções de ligação para regressão binária chamadas de potência e reversa de potência, propostas por [Lemonte e Bazán \(2018\)](#) e [Bazán *et al.* \(2017\)](#) e que são baseadas na função de distribuição acumulada de uma família simétrica de distribuições que servem como linha de base. Estas funções de ligação são alternativas para os modelos de resposta binária. Em particular, na presença de dados desbalanceados, estes modelos foram comparados como os métodos de [King e Zeng \(2001\)](#) e [Firth \(1993\)](#) desenvolvidos para lidar com desbalanceamento.

Para a estimação dos parâmetros, diferentemente de [Lemonte e Bazán \(2018\)](#), neste trabalho uma abordagem Bayesiana foi desenvolvida e a amostragem foi feita por meio do algoritmo Monte Carlo Hamiltoniano na extensão No-U-Turn Sampler. Em nossa abordagem, a estimativa de PSIS-LOO como critério de comparação do modelo e o cálculo do resíduos quantílicos normalizado para avaliar o ajuste do modelo foram introduzidos. Os códigos para todos os modelos foram desenvolvidos e implementados na linguagem Stan por meio do programa Python.

Observamos neste estudo a importância do uso de ligações flexíveis em relação à assimetria dos dados, os quais são controlados por meio de um parâmetro adicional λ . A função de ligação simétrica é um caso particular quando $\lambda = 1$.

As funções de ligação comuns na regressão binária, como logito e probito, nem sempre fornecem o melhor ajuste na presença de dados desbalanceados. Especificamente, o estudo de simulação ([Capítulo 4](#)) mostrou que as funções de ligação propostas apresentam melhores estimativas do que alguns métodos para eventos raros, como proposto por [King e Zeng \(2001\)](#) e [Firth \(1993\)](#).

Por fim, na aplicação dos dados de qualidade do vinho branco, a maioria dos modelos de regressão binária propostos com funções ligação potência e reversa de potência obtiveram melhores ajustes do que os modelos de regressão binária usando funções de ligação comum. Neste

caso, o modelo de regressão binária com função de ligação potência t-Student apresentou melhor ajuste, e não somente a assimetria foi importante, mas também a espessura da cauda. Além disso, juntamente com o uso de critérios de comparação de modelos, mostramos na aplicação, que novas medidas de classificação como CSI, GS, SSI e FAITH pode ser mais apropriado na presença de dados desbalanceados, uma vez que, neste caso, temos dados binários assimétricos na matriz de confusão. Complementarmente, envelopes usando o quantil randomizado normalizado residual para avaliar convenientemente o ajuste da escolha do modelo.

Várias extensões dos métodos desenvolvidos neste trabalho podem ser consideradas em pesquisas futuras. Por exemplo, uma extensão para modelos de regressão ordinal pode ser desenvolvida. Além disso, considerando efeitos aleatórios, uma extensão para um modelo misto pode ser desenvolvida.

REFERÊNCIAS

- ABANTO-VALLE, C. A.; BAZÁN, J. L.; SMITH, A. C. State space mixed models for binary responses with skewed inverse links using jags. **Rio de Janeiro, Brazil**, p. 18, 2014. Citado na página 18.
- AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 56.
- ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American statistical Association**, Taylor & Francis, v. 88, n. 422, p. 669–679, 1993. Citado na página 36.
- ANYOSA, S. A. C. **Regressão binária usando ligações potência e reversa de potência**. Dissertação (Mestrado) — Universidade Federal de São Carlos, Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP, 2017. Citado nas páginas 38, 40, 52 e 55.
- ARANDA-ORDAZ, F. On two families of transformations to additivity for binary response data. *Biometrika*, v. 68, p. 357–364, 1981. Citado na página 18.
- ATKINSON, A. **Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis**. [S.l.]: Clarendon Press, 1985. (Oxford Statistical Science Series). ISBN 9780198533597. Citado na página 60.
- BAZÁN, J.; ROMEO, J.; 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 18, 36 e 52.
- BAZÁN, J.; TORRES-AVILÉS, F.; SUZUKI, A.; LOUZADA, F. Power and reversal power links for binary regressions: An application for motor insurance policyholders. **Applied Stochastic Models in Business and Industry**, John Wiley & Sons, Ltd, v. 33, n. 1, p. 22–34, 2017. Citado nas páginas 18, 21, 24, 25, 26, 28, 29, 52 e 81.
- BAZÁN, J. L.; BOLFARINE, H.; BRANCO, M. D. A framework for skew-probit links in binary regression. **Communications in Statistics - Theory and Methods**, Taylor & Francis, v. 39, n. 4, p. 678–697, 2010. Citado na página 18.
- BAZÁN, J. L.; BRANCO, M. D.; BOLFARINE, H. A skew item response model. **Bayesian Anal.**, International Society for Bayesian Analysis, v. 1, n. 4, p. 861–892, 12 2006. Citado na página 18.
- BRYN, G.; HUBERT, M.; STRUYF, A. A comparison of some new measures of skewness. In: **Developments in robust statistics**. [S.l.]: Springer, 2003. p. 98–113. Citado na página 41.
- CALABRESE, R.; OSMETTI, S. A. Improving forecast of binary rare events data: A gam-based approach. **Journal of Forecasting**, Wiley Online Library, v. 34, n. 3, p. 230–239, 2015. Citado na página 21.

- CARPENTER, B.; GELMAN, A.; HOFFMAN, M. D.; LEE, D.; GOODRICH, B.; BETANCOURT, M.; BRUBAKER, M.; GUO, J.; LI, P.; RIDDELL, A. Stan: A probabilistic programming language. **Journal of statistical software**, Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA (United States), v. 76, n. 1, 2017. Citado na página 62.
- CHAN, P. K.; STOLFO, S. J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In: **KDD**. [S.l.: s.n.], 1998. v. 1998, p. 164–168. Citado na página 21.
- CHEN, M.-H.; DEY, D. K.; SHAO, Q.-M. A new skewed link model for dichotomous quantal response data. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 448, p. 1172–1186, 1999. Citado na página 17.
- CHOI, S.-S.; CHA, S.-H.; TAPPERT, C. C. A survey of binary similarity and distance measures. **Journal of Systemics, Cybernetics and Informatics**, Citeseer, v. 8, n. 1, p. 43–48, 2010. Citado na página 59.
- CHOIRAT, C.; HONAKER, J.; IMAI, K.; KING, G.; LAU, O. **Zelig: Everyone's Statistical Software**. [S.l.], 2018. Version 5.1.6. Disponível em: <<http://zeligproject.org/>>. Citado na página 62.
- COLLETT, D. **Modelling Binary Data, Second Edition**. [S.l.]: CRC Press, 2002. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781420057386. Citado na página 17.
- CORTEZ, P.; CERDEIRA, A.; ALMEIDA, F.; MATOS, T.; REIS, J. Modeling wine preferences by data mining from physicochemical properties. **Decision Support Systems**, Elsevier, v. 47, n. 4, p. 547–553, 2009. Citado na página 69.
- COX, D.; SNELL, E. **Analysis of Binary Data, Second Edition**. [S.l.]: Taylor & Francis, 1989. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412306204. Citado na página 17.
- CZADO, C.; SANTNER, T. J. The effect of link misspecification on binary regression inference. **Journal of statistical planning and inference**, Elsevier, v. 33, n. 2, p. 213–231, 1992. Citado na página 18.
- DING, P. Bayesian robust inference of sample selection using selection-t models. **Journal of Multivariate Analysis**, v. 124, p. 451 – 464, 2014. ISSN 0047-259X. Citado nas páginas 36, 53 e 75.
- DOANE, D. P.; SEWARD, L. E. Measuring skewness: a forgotten statistic. **Journal of Statistics Education**, American Statistical Association. 732 North Washington Street, Alexandria, VA 22314, v. 19, n. 2, p. 1–18, 2011. Citado na página 40.
- DUA, D.; TANISKIDOU, E. K. **UCI Machine Learning Repository**. 2017. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 69.
- DUANE, S.; KENNEDY, A. D.; PENDLETON, B. J.; ROWETH, D. Hybrid monte carlo. **Physics letters B**, Elsevier, v. 195, n. 2, p. 216–222, 1987. Citado nas páginas 18 e 54.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado nas páginas 60 e 79.

FANG, K.; KOTZ, S.; NG, K. W. **Symmetric multivariate and related distributions**. [S.l.]: London ; New York : Chapman and Hall, 1990. Includes index. ISBN 0412314304. Citado na página 26.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 58.

FIRTH, D. Bias reduction of maximum likelihood estimates. **Biometrika**, Oxford University Press, v. 80, n. 1, p. 27–38, 1993. Citado nas páginas 21, 24, 61, 62, 63, 64, 65, 69 e 81.

GAMERMAN, D.; LOPES, H. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition**. [S.l.]: Taylor & Francis, 2006. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781584885870. Citado na página 53.

GEISSER, S.; EDDY, W. F. A predictive approach to model selection. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 74, n. 365, p. 153–160, 1979. Citado na página 57.

GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for bayesian models. **Statistics and computing**, Springer, v. 24, n. 6, p. 997–1016, 2014. Citado na página 57.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical science**, JSTOR, p. 457–472, 1992. Citado na página 75.

GUERRERO, V.; JOHNSON, R. A. Use of the box-cox transformation with binary response models. **Biometrika**, v. 69, 08 1982. Citado na página 18.

GUPTA, R. D.; GUPTA, R. C. Analyzing skewed data by power normal model. **Test**, Springer, v. 17, n. 1, p. 197–210, 2008. Citado na página 18.

HEINZE, G.; PLONER, M.; DUNKLER, D.; SOUTHWORTH, H. logistf: Firth’s bias reduced logistic regression. **R package version**, v. 1, 2013. Citado na página 62.

HEINZE, G.; SCHEMPER, M. A solution to the problem of separation in logistic regression. **Statistics in medicine**, Wiley Online Library, v. 21, n. 16, p. 2409–2419, 2002. Citado na página 62.

HINKLEY, D. V. On power transformations to symmetry. **Biometrika**, v. 62, n. 1, p. 101–111, 1975. Citado na página 40.

HOFFMAN, M. D.; GELMAN, A. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **Journal of Machine Learning Research**, v. 15, n. 1, p. 1593–1623, 2014. Citado nas páginas 18 e 53.

KIM, S.; CHEN, M.-H.; DEY, D. K. Flexible generalized t-link models for binary response data. **Biometrika**, Oxford University Press, v. 95, n. 1, p. 93–106, 2007. Citado na página 18.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political analysis**, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001. Citado nas páginas 17, 21, 22, 61, 62, 63, 64, 65, 69 e 81.

KOSUKE, I.; KING, G.; LAU, O. Relogit: Rare events logistic regression for dichotomous dependent variables. **Zelig: Everyone’s Statistical Software**, 2007. Citado na página 62.

- KUBAT, M.; HOLTE, R. C.; MATWIN, S. Machine learning for the detection of oil spills in satellite radar images. **Machine learning**, Springer, v. 30, n. 2-3, p. 195–215, 1998. Citado na página 22.
- LEMIONET, A.; LIU, Y.; ZHOU, Z. Predicting quality of wine based on chemical attributes. **CS 229 project**, 2015. Disponível em: <http://cs229.stanford.edu/proj2015/245_report.pdf>. Citado na página 71.
- LEMONTE, A. J.; BAZÁN, J. L. New links for binary regression: an application to coca cultivation in peru. **TEST**, v. 27, n. 3, p. 597–617, 2018. ISSN 1863-8260. Citado nas páginas 18, 21, 24, 25, 26, 29, 38, 52, 53, 60, 61, 75 e 81.
- LUO, Y.; AL-HARBI, K. Performances of loo and waic as irt model selection methods. **Psychological Test and Assessment Modeling**, PABST Science Publishers, v. 59, n. 2, p. 183, 2017. Citado na página 58.
- MAALOUF, M.; TRAFALIS, T. B. Robust weighted kernel logistic regression in imbalanced and rare events data. **Computational Statistics & Data Analysis**, Elsevier, v. 55, n. 1, p. 168–183, 2011. Citado na página 21.
- MACGILLIVRAY, H. Skewness and asymmetry: measures and orderings. **The Annals of Statistics**, JSTOR, p. 994–1011, 1986. Citado na página 40.
- MAYORAL, M. A. M.; SOCUÉLLAMOS, J. M. **Modelos lineales generalizados**. [S.l.]: Universidad Miguel Hernández, 2001. Citado na página 34.
- MCCULLAGH, P.; NELDER, J. **Generalized Linear Models, Second Edition**. [S.l.]: Taylor & Francis, 1989. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412317606. Citado nas páginas 22, 23 e 34.
- 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**, Blackwell Publishing Ltd, v. 50, n. 3, p. 417–430, 1996. ISSN 1467-9574. Citado na página 41.
- MORGAN, B. J. T. Observations on quantitative analysis. *Biometrics*, v. 39, p. 879–886, 1983. Citado na página 18.
- NAGLER, J. Scobit: an alternative estimator to logit and probit. **American Journal of Political Science**, JSTOR, p. 230–255, 1994. Citado na página 18.
- NEAL, R. M. Bayesian learning for neural networks. Springer-Verlag New York, Inc., 1996. Citado na página 54.
- _____. Mcmc using hamiltonian dynamics. **arXiv preprint arXiv:1206.1901**, Citeseer, 2012. Citado nas páginas 53, 54 e 55.
- NESTEROV, Y. Primal-dual subgradient methods for convex problems. **Mathematical programming**, Springer, v. 120, n. 1, p. 221–259, 2009. Citado na página 55.
- PAAL, B. Van der. **A comparison of different methods for modelling rare events data**. Dissertação (Mestrado) — Ghent University, 2014. Citado nas páginas 17 e 22.
- PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004. Citado na página 34.

- PRENTICE, R. L. A generalization of the probit and logit methods for dose response curves. **Biometrics**, JSTOR, p. 761–768, 1976. Citado na página 18.
- QUIGLEY, J.; BEDFORD, T.; WALLS, L. Estimating rate of occurrence of rare events with empirical bayes: A railway application. **Reliability Engineering & System Safety**, Elsevier, v. 92, n. 5, p. 619–627, 2007. Citado na página 22.
- ROBERTS, G. O.; SMITH, A. F. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. **Stochastic processes and their applications**, Elsevier, v. 49, n. 2, p. 207–216, 1994. Citado na página 53.
- SCHAEFER, J. T. The critical success index as an indicator of warning skill. **Weather and Forecasting**, v. 5, n. 4, p. 570–575, 1990. Citado na página 59.
- SCHWARZ, G. *et al.* Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 56.
- SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado nas páginas 56 e 57.
- 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 18.
- TAYLOR, J. M.; SIQUEIRA, A. L.; WEISS, R. T. The cost of adding parameters to a model. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 593–607, 1996. Citado na página 36.
- TEAM, S. D. **PyStan: the Python interface to Stan, Version 2.16.0.0**. 2017. Disponível em: <<http://mc-stan.org>>. Citado na página 62.
- TOMZ, M.; KING, G.; ZENG, L. *et al.* Relogit: Rare events logistic regression. **Journal of statistical software**, Foundation for Open Access Statistics, v. 8, n. i02, 2003. Citado na página 62.
- VEHTARI, A.; GELMAN, A.; GABRY, J. loo: Efficient leave-one-out cross-validation and waic for bayesian models. **R package version 0.1**, v. 6, 2016. Citado na página 58.
- _____. Practical bayesian model evaluation using leave-one-out cross-validation and waic. **Statistics and Computing**, Springer, v. 27, n. 5, p. 1413–1432, 2017. Citado nas páginas 57 e 58.
- WANG, X.; DEY, D. K. *et al.* Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 4, n. 4, p. 2000–2023, 2010. Citado na página 18.
- WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, p. 3571–3594, 2010. Citado na página 57.
- WEISS, G. M.; HIRSH, H. Learning to predict extremely rare events. In: **AAAI workshop on learning from imbalanced data sets**. [S.l.: s.n.], 2000. p. 64–68. Citado na página 22.

RESULTADOS DA SIMULAÇÃO

Neste anexo são apresentados os valores estimados, viés e o raiz do erro quadrático médio (RSME) do parâmetro de assimetria (λ) com diferentes modelos potência e reversa de potência nos diferentes cenários.

Nas Tabelas 22 e 23 é mostrado os valores estimados para os modelos potência e nas Tabelas 24 e 25 para os modelos reversa de potência.

Tabela 22 – Estimativas do parâmetro de assimetria para os modelos potência com diferentes valores de n e $\lambda > 1$.

Método	n	$\lambda = 4$			$\lambda = 2$		
		Estimativa	Viés	RSME	Estimativa	Viés	RSME
PL	500	6.389	2.389	2.393	2.070	0.070	1.339
	5000	5.105	1.105	1.909	2.598	0.598	1.188
	20000	4.655	0.655	0.860	2.822	0.822	0.890
PP	500	5.813	1.813	1.880	4.105	2.105	1.130
	5000	5.992	1.992	2.008	4.753	2.753	2.907
	20000	7.042	3.042	3.043	5.419	3.419	2.430
Plaplace	500	4.127	0.127	0.939	2.336	0.336	0.964
	5000	4.694	0.694	0.951	2.115	0.115	0.225
	20000	4.605	0.605	0.705	2.085	0.085	0.119
PC	500	4.534	0.534	1.016	2.105	0.105	0.415
	5000	4.159	0.159	0.439	1.990	-0.010	0.128
	20000	4.021	0.021	0.229	2.014	0.014	0.065

Tabela 23 – Estimativas do parâmetro de assimetria para os modelos potência com diferentes valores de n e $\lambda < 1$.

Método	n	$\lambda = 0.5$			$\lambda = 0.25$		
		Estimativa	Viés	RSME	Estimativa	Viés	RSME
PL	500	0.469	-0.031	0.384	0.313	0.063	0.135
	5000	0.452	-0.048	0.096	0.209	-0.041	0.054
	20000	0.398	-0.102	0.056	0.186	-0.064	0.068
PP	500	0.389	-0.111	0.190	0.385	0.135	0.162
	5000	0.326	-0.174	0.206	0.257	0.007	0.077
	20000	0.322	-0.178	0.216	0.192	-0.058	0.073
Plaplace	500	0.394	-0.106	0.173	0.385	0.135	0.396
	5000	0.965	0.465	1.906	1.867	1.617	1.580
	20000	1.324	0.824	1.472	1.398	1.148	1.151
PC	500	0.449	-0.051	0.093	0.321	0.071	0.397
	5000	0.496	-0.004	0.020	0.251	0.001	0.022
	20000	0.497	-0.003	0.017	0.253	0.003	0.010

Tabela 24 – Estimativas do parâmetro de assimetria para os modelos reversa de potência com diferentes valores de n e $\lambda > 1$.

Método	n	$\lambda = 4$			$\lambda = 2$		
		Estimativa	Viés	RSME	Estimativa	Viés	RSME
RPL	500	0.177	-3.823	3.823	0.343	-1.658	1.664
	5000	0.147	-3.854	3.853	0.338	-1.662	1.663
	20000	0.139	-3.861	3.861	0.346	-1.654	1.654
RPP	500	0.233	-3.767	3.767	0.378	-1.622	1.649
	5000	0.152	-3.848	3.848	0.238	-1.762	1.764
	20000	0.140	-3.860	3.860	0.240	-1.760	1.760
RPlaplace	500	0.185	-3.815	3.815	0.374	-1.626	1.630
	5000	0.159	-3.841	3.841	0.420	-1.580	1.580
	20000	0.154	-3.846	3.846	0.425	-1.575	1.575
RPC	500	0.215	-3.785	3.786	0.463	-1.537	1.540
	5000	0.249	-3.751	3.751	0.487	-1.513	1.513
	20000	0.251	-3.749	3.749	0.495	-1.505	1.505

Tabela 25 – Estimativas do parâmetro de assimetria para os modelos reversa de potência com diferentes valores de n e $\lambda > 1$.

Método	n	$\lambda = 0.5$			$\lambda = 0.25$		
		Estimativa	Viés	RSME	Estimativa	Viés	RSME
RPL	500	1.566	1.066	1.408	4.709	4.459	3.962
	5000	2.462	1.962	2.151	4.057	3.807	3.912
	20000	2.412	1.912	1.969	5.004	4.754	4.835
RPP	500	2.409	1.909	0.140	3.348	3.098	3.114
	5000	2.115	1.615	1.892	3.353	3.103	4.354
	20000	2.521	2.021	2.020	4.637	4.387	4.447
RPlaplace	500	1.681	1.181	1.563	2.645	2.395	1.268
	5000	1.853	1.353	1.361	2.843	2.593	2.606
	20000	1.840	1.340	1.341	2.878	2.628	2.630
RPC	500	2.065	1.565	1.799	3.540	3.290	3.529
	5000	1.820	1.320	1.324	3.136	2.886	2.909
	20000	1.829	1.329	1.330	3.030	2.780	2.787

As estimativas do valor de λ para o modelo potência Cauchy (Tabelas 22 e 23) recupera o verdadeiro valor e apresentam menor valor do RSME em todos os cenários do que os outros modelos potência.