UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos alternativos para classificação em dados desbalanceados

Alex de la Cruz Huayanay

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Alex de la Cruz Huayanay

Modelos alternativos para classificação em dados desbalanceados

Tese 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 Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
Setembro de 2023

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

d278m

de la Cruz Huayanay, Alex
Modelos alternativos para classificação em dados desbalanceados / Alex de la Cruz Huayanay; orientador Jorge Luis Bazán Guzmán. -- São Carlos, 2023.
106 p.

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

1. Métricas para classificação binária. 2. cloglog. 3. distribuição potência. 4. ligação assimétrica. 5. dados desbalanceados. I. Bazán Guzmán, Jorge Luis, orient. II. Título.

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

Alex de la Cruz Huayanay

Alternative models for classification in imbalanced data

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

Concentration Area: Statistics

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
September 2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Alex de La Cruz Huayanay, realizada em 11/08/2023.

Comissão Julgadora:

Prof. Dr. Jorge Luis Bazán Guzmán (USP)

Profa. Dra. Cibele Maria Russo Novelli (USP)

Prof. Dr. Luis Hilmar Valdivieso Serrano (PUC-Perú)

Prof. Dr. Felipe Alberto Osório Salgado (UTFSM)

Prof. Dr. Cristian Marcelo Villegas Lobos (ESALQ/USP)

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

Este trabalho é dedicado a todos aqueles que têm paixão pela ciência estatística e estão se esforçando cada dia para fazer um mundo melhor.

AGRADECIMENTOS

Primeiramente a Deus, por permitir que tudo isso acontecesse, por me dar tudo que eu precisei, desde saúde até sabedoria para que esta aventura terminasse desta forma tão prazerosa.

Aos meus pais e irmãos, por estarem sempre preocupados com o meu bem estar. 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.

A meu orientador, professor Jorge Luis Bazán, pela confiança em mim depositada, por toda paciência e por ser uma guia na minha caminhada acadêmica. Muito obrigado por ser um grande amigo e referência profissional para mim.

Aos professores membros da banca, Cibele Russo, Luis Valdivieso, Felipe Osório e Cristian Villegas, por disponibilizarem seu tempo em avaliar este trabalho e pelas suas valiosas sugestões e comentários para aprimorar o meu trabalho.

Aos professores da USP e da UFSCar, pelos valiosos ensinamentos recebidos, por me proporcionarem bases sólidas na minha trajetória e por cultivar em mim o amor pela ciência. Da mesma forma, ao corpo técnico do PIPGEs, pela disponibilidade nos diversos momentos e pelos esclarecimentos prestados.

Aos todos os meus amigos e colegas do PIPGEs por ser parte desta aventura, porque cada um com seus próprios desafios, alegrias e sonhos, contribuíram muito. Por tantas horas de estudos juntos, por momentos de café, comida e brincadeiras compartilhadas. Foram momentos agradáveis.

A todos os meus familiares e amigos, tanto no Brasil quanto no Peru, que me ajudaram de alguma forma, quero agradecer profundamente, pois sem eles isso teria sido muito mais difícil.

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 alternativos para classificação em dados desbalanceados**. 2023. 106 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Na classificação binária o método mais usado é o modelo de regressão logística. No entanto, vários autores indicam que esse modelo não é adequado quando os dados são desbalanceados. Diante disso, diferentes funções de ligação assimétrica, como alternativas para modelos de resposta binária, foram propostas; por exemplo, nos últimos anos foram estudadas as distribuições potência (P) e reversa de potência (RP). Neste trabalho desenvolvemos novas propriedades das distribuições P e RP no contexto de modelos para classificação em dados desbalanceados. Também, algumas métricas para classificação são estudadas através de um estudo de simulação, e uma aplicação da metologia estudada é apresentada.

Além do mais, estudamos a extensão dos modelos de regressão binária para o caso misto em classificação binária no contexto de estudos longitudinais. Para avaliar o performance deste tipo de modelos apresentamos um estudo de simulação. Adicionalmente, mostramos uma aplicação da metodologia estudada para um conjunto de dados em que a variável resposta é longitudinal e desbalanceada.
Para o processo de estimação dos parâmetros consideramos uma abordagem bayesiana usando um procedimento MCMC através do algoritmo *No-U-Turn Sampler* (NUTS). Verificações preditivas a posteriori, resíduos quantílicos aleatorizados Bayesianos e uma medida de influência bayesiana são considerados para o diagnóstico do modelo longitudinal. Diferentes modelos são comparados usando critérios de comparação de modelos.

Palavras-chave: Métricas para classificação binária; potência Gumbel; distribuição potência; ligação assimétrica; dados desbalanceados; modelo de efeitos mistos.

ABSTRACT

ALEX, C. H. **Alternative models for classification in imbalanced data**. 2023. 106 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

In binary classification, the most used method is logistic regression model. However, several authors indicate that this model is not suitable when the data are imbalanced; for this, different asymmetric link functions as alternatives for binary response models have been proposed, for example, in recent years the power (P) and reverse power (RP) distributions have been presented. In this work we develop new properties of the P and RP distributions in the context of models for classification on imbalanced data. Also, some metrics for classification are studied through a simulation study, and an application of the studied methodology is presented.

In addition, we extend the binary regression models to the case of mixed models for binary classification in the context of a longitudinal studies. To evaluate the performance of the models, a simulation study is performed. Additionally, an application is considered concerning the studied methodology in a dataset in which the response is longitudinal and imbalanced. For parameter estimation the Bayesian approach is considered using a MCMC procedure through the No-U-Turn Sampler (NUTS) algorithm. Further predictive checks, randomized Bayesian quantile residuals and a measure of Bayesian influence are considered for model diagnosis. Different models are compared using model selection criteria.

Keywords: Metrics for binary classification; power Gumbel; power distribution; Asymmetric link; imbalanced data; mixed-effects model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Em problemas de classificação considera-se atribuir um indivíduo ou observação y_i ($i = 1, \ldots, n$) a uma das K categorias ou classes em função de uma série de atributos, isto é, para cada observação i da forma $(\boldsymbol{x}_i, y_i)$, em que $\boldsymbol{x}_i$ é um vetor $p \times 1$ de atributos (covariáveis), $y_i \in \{1, \ldots, K\}$ é o i-ésimo rótulo da classe. Em particular, quando $K = 2$ torna-se um problema de classificação binária. No contexto de algoritmos de aprendizado de máquina supervisionado, os quais lidam mais com a classificação (AYODELE, 2010), existem vários métodos ou técnicas propostos para problemas de classificação, que incluem os seguintes: classificadores lineares, regressão logística, classificador o Bayes ingênuo (*naive Bayes* em inglês), máquina de vetores de suporte (SVM, do inglês: *support vector machine*); classificadores quadráticos, K-Médias (*K-means clustering* em inglês), *Boosting*, árvore de decisão, floresta aleatória (RF, do inglês: *random forest*), redes neurais, redes Bayesianas e assim por diante. O método mais usado é a regressão logística a qual usa uma função de ligação chamada logito, que faz parte dos modelos de regressão binária. Estudar modelos de regressão com resposta binária, desempenha um papel muito importante na ciência de dados.

Por outro lado, para avaliar a capacidade preditiva de um modelo (método) de classificação, são utilizadas diferentes métricas, entre as mais usadas estão a precisão (ACC), área sob a curva (AUC), taxa de verdadeiro negativo (TNR) e taxa de verdadeiro positivo (TPR). No entanto, alguns estudos mostraram que na presença de dados desbalanceados, isto é, quando o número de observações de uma classe é significativamente excedido pelas da outra classe e a classe minoritária costuma ser de maior interesse, como é comum em muitas aplicações do mundo real, os modelos de regressão binária com funções de ligação simétrica, como o caso de regressão logística, podem não ser adequados. Também, neste mesmo contexto, nem todas as métricas para avaliação da capacidade preditiva do modelo podem ser adequadas (FATOURECHI *et al.*, 2008; De la Cruz *et al.*, 2019; LUQUE *et al.*, 2019).

As tentativas para evitar o impacto negativo da má especificação do modelo quando os

dados estão desbalanceados, o que é bastante frequente, algumas vezes são trabalhadas com o uso das funções de ligação assimétrica (BAZÁN *et al.*, 2017; LEMONTE; BAZÁN, 2018; De la Cruz *et al.*, 2019) ou com o uso de algumas métricas alternativas para avaliar a capacidade preditiva do modelo (THAI-NGHE; GANTNER; SCHMIDT-THIEME, 2011; De la Cruz *et al.*, 2019; HLOSTA *et al.*, 2013).

No caso da regressão binária, considerar diferentes funções de ligação, tem sido proposta para muitos problemas distintos nas últimas décadas, por exemplo em Chen, Dey e Shao (1999), Basu e Mukhopadhyay (2000), Kim (2002), Wang e Dey (2011), Naranjo, Pérez e Martín (2019) e Yin *et al.* (2020), podem ser vistos estudos detalhados.

Em particular, uma classe de funções de ligação que tem recebido uma atenção no contexto de dados desbalanceados (De la Cruz *et al.*, 2019), são as chamadas funções de ligação potência e reversa de potência. Desde o trabalho pioneiro de Prentice (1976) propondo as duas extensões de função de ligação logística por exponenciação, foram propostas as funções de ligações potência e reversa potência estendendo a função de ligação probito (BAZÁN; ROMEO; RODRIGUES, 2014), cauchito, complemento log-log (cloglog) e log-log (loglog) (BAZÁN *et al.*, 2017) até a classe simétrica como base das distribuições potência (LEMONTE; BAZÁN, 2018). Além disso, nos últimos anos, foram estudadas funções de ligação assimétrica P e RP em contexto de dados desbalanceados (De la Cruz *et al.*, 2019), e mostraram um bom desempenho. Os diferentes trabalhos foram motivados por aplicações interessantes; porém, o estudo das propriedades das funções de ligação, não foram esgotados no momento. Em todos esses casos são apresentadas propostas para a construção de um conjunto de novas funções de ligação assimétrica baseado na exponenciação de uma função de distribuição acumulada (fda), denominada distribuição de potência e de uma versão reversa desta distribuição.

Nos diferentes estudos acima mencionados, nenhuma avaliação sobre a viabilidade das distribuições como funções de ligação para regressão binária, foram estudada até o momento. E, por outro lado, não foi realizado um estudo sobre o uso de funções de ligação assimétrica potência e reversa de potência, no contexto de modelos mistos para classificação binária.

Por outro lado, uma extensão da regressão binária é o estudo de modelos mistos; em particular, modelos para o caso resposta binária longitudinal, o qual é uma metodologia que avalia o comportamento de uma variável de resposta ao longo do tempo; em outras palavras, os indivíduos são acompanhados ao longo do tempo e a variável de interesse é medida repetidamente para o mesmo individuo. Essa metodologia visa mensurar o efeito da dependência entre a variável resposta e a variável explicativa, bem como mensurar possíveis efeitos entre e/ou intra-indivíduos por meio de uma estrutura de correlação. Diggle, Liang e Zeger (1994) e Fitzmaurice, Laird e Ware (2012) discutiram esta metodologia detalhadamente. Algumas aplicações do modelo misto pode ser visto em Masuda e Stone (2015), Gibbons e Hedeker (1997), Breslow e Clayton (1993), Wolfinger e O'connell (1993) e Stiratelli, Laird e Ware (1984).
Em geral, quando a variável de resposta é binária, os modelos conhecidos para dados com

resposta binária longitudinal, são aqueles que usam funções de ligação simétrica como logito e probito, que formam a base para a análise de dados binários na prática (PARZEN *et al.*, 2011; THOMAS *et al.*, 1998; MASUDA; STONE, 2015; GIBBONS; HEDEKER, 1997). No entanto, como já foi mencionado, as funções de ligação simétrica, como logito e probito, podem não ser adequadas quando os dados são desbalanceados (De la Cruz *et al.*, 2019) e é melhor considerar o uso de funções de ligação assimétrica.

Nesta tese, estudamos modelos alternativos para problemas de classificação quando os dados são desbalanceados. Nesse sentido, o objetivo principal deste trabalho, portanto, é propor o uso das funções de ligação baseada nas distribuições potência e reversa de potência, como alternativas às funções de ligação comum, nas seguintes situações: classificação binária no contexto de Modelos Lineares Generalizados (MLG) e classificação binária mista para dados com resposta binária longitudinal no contexto de Modelos Lineares Generalizados Mistos (MLGM). Essas funções de ligação são muito flexíveis com uma ampla família de distribuições (LEMONTE; BAZÁN, 2018; BAZÁN *et al.*, 2017), que incluem, como caso particular, funções de ligação comuns, como logito e probito.

A tese está organizada em cinco capítulos, descritos a seguir. No Capítulo 2 apresentamos os conceitos fundamentais, fixação de notações básicas e a estruturação do ambiente matemático, utilizado para o desenvolvimento do trabalho. Neste capítulo, descrevemos brevemente a revisão das distribuições potência, a regressão binária para classificação, comparação de modelos sob abordagem Bayesiana e, por fim, uma extensão de modelos de regressão binária para o caso misto.

No Capítulo 3, analisamos novas propriedades das distribuições potência e reversa de potência, previamente estudadas, para classificação binária; além do mais, mostramos que as funções de ligação potência loglog e a reversa de potência loglog, são funções de ligação inadequadas em regressão binária. Também, além das métricas comuns para classificação binária, estudamos algumas métricas apresentadas em De la Cruz *et al.* (2019) que podem ser adequadas para avaliar a capacidade preditiva do modelo quando os dados são desbalanceados.

No Capítulo 4, o foco está no estudo de modelos mistos no contexto de resposta binária longitudinal. Estudamos as funções de ligação potência e sua reversa, como alternativa às funções de ligação usuais. O estudo é feito sob abordagem Bayesiana, que inclui a estimação dos parâmetros, critérios de comparação de modelos e análise de diagnóstico do modelo.

Por fim, no Capítulo 5, são apresentadas algumas conclusões e comentários finais sobre os modelos estudados ao longo do trabalho. Também, são apresentados alguns trabalhos publicados em jornal e algumas possibilidades de pesquisas a partir do que foi desenvolvido até o momento com o propósito de dar continuidade a este trabalho.

CAPÍTULO

2

CONCEITOS PRELIMINARES

Neste capítulo, apresentamos uma revisão de conceitos fundamentais, fixação de notações básicas e a estruturação do ambiente matemático que serão utilizados ao longo deste trabalho.

Na Seção 2.1, são apresentados os principais conceitos e caraterísticas das distribuições potência e reversa de potência estudadas em Lemonte e Bazán (2018), Bazán *et al.* (2017) e De la Cruz (2019). Na Seção 2.2 é feita uma revisão da definição de regressão binária. Em seguida, na Seção 2.3 são explicadas as diferentes formas de comparar o performance de um modelo. Finalmente, na Seção 2.4, mostramos uma extensão regressão binária no contexto de modelos mistos.

2.1 Distribuição potência e reversa de potência

Para compreender a definição da distribuição potência e sua reversa, inicialmente, introduziremos o conceito de reversibilidade e a simetria de uma distribuição de probabilidade, os quais são mostrados em De la Cruz (2019) que segue o trabalho de Bazán *et al.* (2017).

Definição 1. Seja X uma variável aleatória com uma distribuição de probabilidade denotada por $X \sim F(\cdot)$. Dizemos que a distribuição de X satisfaz a propriedade de reversibilidade, se a função de distribuição de probabilidade de $-X$ é uma distribuição diferente da X, e que pode ser escrita como $-X \sim H(\cdot)$, em que $H(\cdot) \equiv 1 - F(-\cdot)$. Nesse caso, a distribuição $H(\cdot)$ é chamada de distribuição reversa de $F(\cdot)$.

Definição 2. Uma distribuição de probabilidade é dita ser simétrica se e somente se, existir um valor x_0 tal que $f(x_0 - \delta) = f(x_0 + \delta)$, $\forall \delta \in \mathbb{R}$. Em que f é a função de densidade de probabilidade (se a distribuição for contínua) ou função de probabilidade (se a distribuição for discreta). Por exemplo, quando $x_0 = 0$, as distribuições Logística, Normal, t-Student, Laplace e Cauchy, são simétricas.

Note que, a partir das definições 1 e 2, uma distribuição simétrica em torno do zero ($x_0 = 0$) não apresenta reversibilidade, isto é mostrado no resultado 1.

Resultado 1. Seja $X \sim F(\cdot)$ uma variável aleatória que segue uma distribuição simétrica em torno do zero, então $f(x) = f(-x)$, equivalentemente, $F(x) = 1 - F(-x)$. Portanto, a distribuição de X não satisfaz a propriedade de reversibilidade, pois X e $-X$ têm a mesma distribuição. Exemplos dessas distribuições são as distribuições padrão: Logística, Normal, t-Student, Laplace e Cauchy.

Para qualquer distribuição de probabilidade, em que não for verificada a simetria, é possível propor uma distribuição reversa a ela. Um caso desse tipo de distribuições, são as distribuições chamadas distribuições potência, as quais descrevemos a seguir.

No trabalho de De la Cruz (2019), Chumbimune (2017) e Lemonte e Bazán (2018), estudam a construção de uma distribuição 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 adicional, chamado parâmetro de potência ou de forma, de acordo com a definição 3.

Definição 3. Uma variável aleatória univariada X, segue uma distribuição de probabilidade potência, com parâmetro de locação $\mu \in \mathbb{R}$, parâmetro de escala $\sigma > 0$ e parâmetro de forma $\alpha > 0$, se sua função de distribuição acumulada (fda), denotada por F_P, é da forma

$$F_P(x \mid \mu, \sigma, \alpha) = G\left(\frac{x-\mu}{\sigma}\right)^{\alpha}, \quad x \in \mathbb{R}, \tag{2.1}$$

e sua função de densidade de probabilidade (fdp), denotada por f_P, é

$$f_P(x \mid \sigma, \mu, \alpha) = \frac{\alpha}{\sigma}\left\{G\left(\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}, \tag{2.2}$$

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

Resultado 2. Para as distribuições potência, se $G(\cdot)$ é uma distribuição de linha de base simétrica em torno de zero, então segue que:

$$F_P(-x \mid \mu, \sigma, \alpha) = \left\{G\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right\}^{\alpha} = \left\{1 - G\left(\frac{x-\mu}{\sigma}\right)\right\}^{\alpha}, \quad x \in \mathbb{R}.$$

Demonstração. Note que quando $G(\cdot)$ é uma função de uma família de distribuição simétrica em torno do zero, satisfaz que $G(-z) = 1 - G(z)$, como pode ser visto na Figura 1 (caso da distribuição normal padrão). Assim, ao elevar por um parâmetro α obtemos o Resultado 2. □

Observe que, a partir da definição 3, é possível construir uma distribuição reversa, a qual pode ser chamada de distribuição reversa de potência, de acordo com a definição 4.

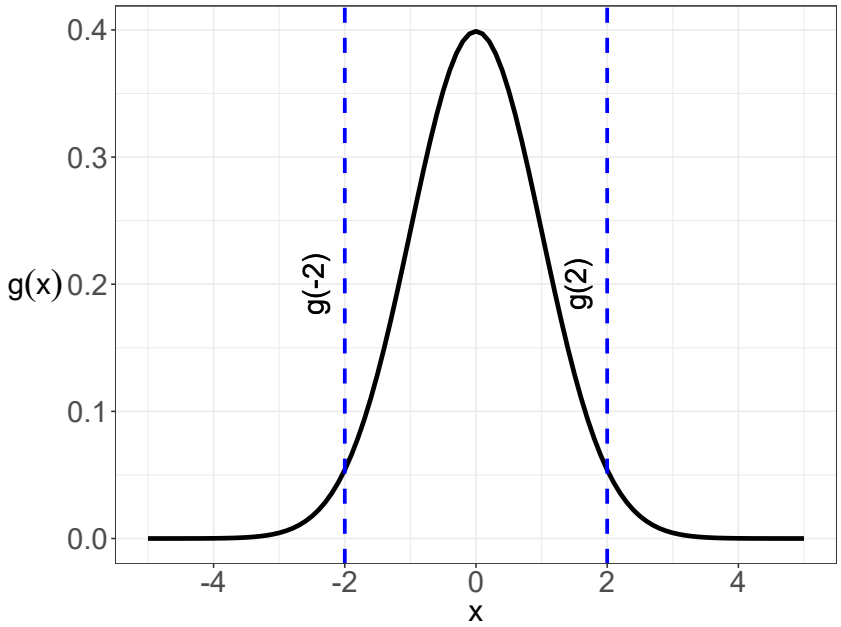

Figura 1 – fdp da distribuição normal padrão (simétrica em torno de zero)

Definição 4. Seja X uma variável aleatória com distribuição potência $X \sim F_P(\cdot)$, então una distribuição reversa de potência (RP), é definida considerando a sua fda (denotada por F_{RP}) como

$$F_{RP}(x \mid \mu, \sigma, \alpha) = 1 - G\left(-\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha}, \quad x \in \mathbb{R}, \tag{2.3}$$

e sua função de densidade de probabilidade (fdp), denotada por f_{RP}, obtido pela derivação de $F_{RP}(\cdot)$, dada por

$$f_{RP}(x \mid \mu, \sigma, \alpha) = \frac{\alpha}{\sigma}\left\{G\left(-\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(-\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R} \tag{2.4}$$

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

Considerando que a função de linha de base $(G(\cdot))$ pertence à família de distribuições simétricas temos os resultados 3 e 4.

Resultado 3. Podemos ver que $F_{RP}(\cdot)$ satisfaz a propriedade de reversibilidade definido em 1, desde que

$$F_{RP}(x \mid \mu, \sigma, \alpha) + F_P(-x \mid \mu, \sigma, \alpha) = 1,$$

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

Demonstração. Pelas definições 3 e 4, respetivamente, temos que

$$F_P(-x \mid \mu, \sigma, \alpha) = \left[G\left(-\frac{x-\mu}{\sigma}\right)\right]^{\alpha} \quad \text{e} \quad F_{RP}(x \mid \mu, \sigma, \alpha) = 1 - \left[G\left(-\frac{x-\mu}{\sigma}\right)\right]^{\alpha},$$

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

$$F_{RP}(x \mid \mu,\sigma,\alpha) + F_P(-x \mid \mu,\sigma,\alpha) = 1 - \left[G\left(-\frac{x-\mu}{\sigma}\right)\right]^{\alpha} + \left[G\left(-\frac{x-\mu}{\sigma}\right)\right]^{\alpha}.$$

Logo,

$$F_{RP}(x \mid \mu,\sigma,\alpha) + F_P(-x \mid \mu,\sigma,\alpha) = 1.$$

□

Resultado 4. Para qualquer $\alpha \neq 1$, as funções de distribuição potência e reversa de potência não são simétricas, isto é, $f_P(x \mid \mu,\sigma,\alpha) \neq f_P(-x \mid \mu,\sigma,\alpha)$ e $f_{RP}(x \mid \mu,\sigma,\alpha) \neq f_{RP}(-x \mid \mu,\sigma,\alpha)$.

Demonstração. Na Equação 2.2, a função de densidade de probabilidade potência é dada por:

$$f_P(x \mid \mu,\sigma,\alpha) = \frac{\alpha}{\sigma}\left\{G\left(\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(\frac{x-\mu}{\sigma}\right).$$

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

$$f_P(-x \mid \mu,\sigma,\alpha) = \frac{\alpha}{\sigma}\left\{G\left(-\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(-\frac{x-\mu}{\sigma}\right) \neq f_P(x \mid \mu,\sigma,\alpha).$$

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

$$f_{RP}(x \mid \mu,\sigma,\alpha) = \frac{\alpha}{\sigma}\left\{G\left(-\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(\frac{x-\mu}{\sigma}\right).$$

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

$$f_{RP}(-x \mid \mu,\sigma,\alpha) = \frac{\alpha}{\sigma}\left\{G\left(\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} g\left(-\frac{x-\mu}{\sigma}\right) \neq f_{RP}(x \mid \mu,\sigma,\alpha).$$

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. □

Quando $\alpha = 1$, temos que $F_P(x \mid \mu,\sigma^2) = F_{RP}(x \mid \mu,\sigma,\alpha) = G\left(\frac{x-\mu}{\sigma}\right)$. Assim, $G\left(\frac{x-\mu}{\sigma}\right)$ é um caso particular de ambas às distribuições.

As definições 3 e 4, permitem usar uma ampla família de distribuições como linha de base. Maiores detalhes sobre as distribuições univariadas podem ser encontradas em Johnson, Kotz e Balakrishnan (1995). A lista de distribuições de linha de base introduzida por Bazán *et al.* (2017) inclui as distribuições Logística, Normal, Cauchy, Reversa Gumbel e Gumbel. Além disso, Lemonte e Bazán (2018) inclui Laplace, t-Student e distribuição Exponencial potência como outras distribuições de linha de base. Algumas distribuições de linha de base possuem um parâmetro extra (por exemplo, os graus de liberdade de t-Student), esses parâmetros não

estão associados à função de ligação e não são fáceis de interpretar no modelo de regressão binária (De la Cruz *et al.*, 2019); por essa razão, neste trabalho, as distribuições que possuem um parâmetro extra não serão consideradas.

Os nomes e as notações das distribuições potência e reversa de potência, usadas 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
	Cauchy	C
	Laplace	LA
	Gumbel	G
	Reversa Gumbel	RG
Potência	Potência Logística	PL
	Potência Normal	PN
	Potência Cauchy	PC
	Potência Laplace	PLA
	Potência Gumbel	PG
	Potência Reversa Gumbel	PRG
Reversa de potência	Reversa de potência Logística	RPL
	Reversa de potência Normal	RPN
	Reversa de potência Cauchy	RPC
	Reversa de potência Laplace	RPLA
	Reversa de potência Gumbel	RPG
	Reversa de potência Reversa Gumbel	RPRG

Além disso, usaremos a notação $F_l(\cdot)$ ou $f_l(\cdot)$ para nos referirmos a qualquer uma das distribuições potência ($l = P$) ou reversa de potência ($l = RP$). Na Tabela 2 é mostrada a forma da função de distribuição acumulada e sua respectiva função de densidade de probabilidade de algumas distribuições potência e reversa de potência, cada uma delas com parâmetros de locação $\mu \in \mathbb{R}$, escala $\sigma > 0$ e forma $\alpha > 0$. $\Phi(\cdot)$ denota uma função de distribuição acumulada de uma distribuição Normal padrão.

A partir das definições 3 e 4, é possível expressar as distribuições potência e reversa de potência em sua forma padronizada fazendo $\mu = 0$ e $\sigma = 1$. Assim, temos a distribuição de potência padrão (P) e distribuição de reversa de potência padrão (RP) 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 P e RP são, respetivamente, da seguinte forma

$$f_P(z) = \alpha\, G(z)^{\alpha-1} g(z) \;\text{ e }\; f_{RP}(z) = \alpha\, G(-z)^{\alpha-1} g(z), \quad z \in \mathbb{R}.$$

Na Tabela 3, são mostradas as funções de distribuições acumuladas de algumas distribuições potência e reversa de potência padrão, e sua respectivas funções de densidade de

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

Distribuição	fda $F_l(x \mid \mu,\sigma,\alpha)$	fdp $f_l(x \mid \mu,\sigma,\alpha)$
PL	$\left[\frac{1}{1+\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}}\right]^{\alpha}$	$\alpha\left[\frac{1}{1+\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}}\right]^{\alpha-1}\frac{\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}}{\sigma\left(1+\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right)^2}$
PN	$\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{\alpha}$	$\frac{\alpha}{\sigma}\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{\alpha-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$
PC	$\left[\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)+\frac{1}{2}\right]^{\alpha}$	$\frac{\alpha}{\sigma\pi}\left[\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)+\frac{1}{2}\right]^{\alpha-1}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$
PLA	$\left[\frac{1}{2}+\frac{1}{2}\text{sign}\,(x-\mu)\left\{1-\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)\right\}\right]^{\alpha}$	$\frac{\alpha}{\sigma}\left[\frac{1}{2}+\frac{1}{2}\text{sign}\,(x-\mu)\left\{1-\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)\right\}\right]^{\alpha-1}\frac{1}{2}\exp\left\{-\frac{\lvert x-\mu\rvert}{\sigma}\right\}$
PG	$\left[\exp\{-\exp\{-\frac{x-\mu}{\sigma}\}\}\right]^{\alpha}$	$\alpha\left[\exp\{-\exp\{-\frac{x-\mu}{\sigma}\}\}\right]^{\alpha-1}\exp\{-\left(\frac{x-\mu}{\sigma}+\exp\{-\frac{x-\mu}{\sigma}\}\right)\}$
PRG	$\left[1-\exp\{-\exp\{\frac{x-\mu}{\sigma}\}\}\right]^{\alpha}$	$\alpha\left[1-\exp\{-\exp\{\frac{x-\mu}{\sigma}\}\}\right]^{\alpha-1}\exp\{-\left(-\frac{x-\mu}{\sigma}+\exp\{\frac{x-\mu}{\sigma}\}\right)\}$
RPL	$1-\left[\frac{1}{1+\exp\left\{\left(\frac{x-\mu}{\sigma}\right)\right\}}\right]^{\alpha}$	$\alpha\left[\frac{1}{1+\exp\left\{\frac{x-\mu}{\sigma}\right\}}\right]^{\alpha-1}\frac{\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}}{\sigma\left(1+\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right)^2}$
RPN	$1-\left[\Phi\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right]^{\alpha}$	$\frac{\alpha}{\sigma}\left[\Phi\left(-\frac{x-\mu}{\sigma}\right)\right]^{\alpha-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$
RPC	$1-\left[\frac{1}{\pi}\arctan\left(-\frac{x-\mu}{\sigma}\right)+\frac{1}{2}\right]^{\alpha}$	$\frac{\alpha}{\sigma\pi}\left[\frac{1}{\pi}\arctan\left(-\frac{x-\mu}{\sigma}\right)+\frac{1}{2}\right]^{\alpha-1}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$
RPLA	$1-\left[\frac{1}{2}-\frac{1}{2}\text{sign}\,(x-\mu)\left\{1-\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)\right\}\right]^{\alpha}$	$\frac{\alpha}{\sigma}\left[\frac{1}{2}-\frac{1}{2}\text{sign}\,(x-\mu)\left\{1-\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)\right\}\right]^{\alpha-1}\frac{1}{2}\exp\left\{-\frac{\lvert x-\mu\rvert}{\sigma}\right\}$
RPG	$1-\left[\exp\{-\exp\{\frac{x-\mu}{\sigma}\}\}\right]^{\alpha}$	$\alpha\left(e^{-e^{\frac{x-\mu}{\sigma}}}\right)^{\alpha-1}e^{-\left(\frac{x-\mu}{\sigma}+e^{-\frac{x-\mu}{\sigma}}\right)}$
RPRG	$1-\left[1-\exp\{-\exp\{-\frac{x-\mu}{\sigma}\}\}\right]^{\alpha}$	$\alpha\left[1-\exp\{-\exp\{-\frac{x-\mu}{\sigma}\}\}\right]^{\alpha-1}\exp\{-\left(-\frac{x-\mu}{\sigma}+\exp\{\frac{x-\mu}{\sigma}\}\right)\}$

probabilidade.

Neste trabalho, as distribuições potência e reversa de potência na sua forma padronizada, serão usadas para construção de uma função de ligação.

Como foi mostrado em De la Cruz (2019) e Lemonte e Bazán (2018), as observações 1, 2 e 3, podem ser verificadas.

Observação 1. Para $z \in \mathbb{R}$, $F_P(-z) = G(-z)^{\alpha}$ e $F_{RP}(-z) = 1 - G(-z)^{\alpha}$. Logo, $F_P(\pm z) + F_{RP}(\mp z) = 1$. Também, $F_P(-z) \neq 1 - F_P(z)$ e $F_{RP}(-z) \neq 1 - F_{RP}(z)$. Assim, $F_P(z)$ 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 $\alpha > 1$ e à esquerda (assimetria negativa) se $0 < \alpha < 1$, e as distribuições reversa de potência são inclinadas para a esquerda (assimetria negativa) se $\alpha > 1$ e à direita (assimetria positiva) se $0 < \alpha < 1$. Isto pode ser visto na da Figura 2, na qual mostramos a curva das funções de densidade de probabilidade (fdp) para algumas distribuições potência e reversa de potência com diferentes valores de $\alpha = \{0,25,\ 1,\ 3\}$. Consequentemente, as curvas das funções de distribuições acumuladas da distribuição reversa de potência é um reflexo da curva correspondente da distribuição potência, e então para $\alpha < 1$ (ou $\alpha > 1$) a curva correspondente é geralmente abaixo (acima) da curva correspondente para a curva da linha de base, isto pode ser visto na Figura 3.

Tabela 3 – fda e fdp de distribuições de potência padrão

Distribuição	fda padrão $F_l(z)$	fdp padrão $f_l(z)$
PL	$\left[\frac{1}{1+\exp(-z)}\right]^{\alpha}$	$\alpha\left[\frac{1}{1+\exp(-z)}\right]^{\alpha-1}\frac{\exp(-z)}{(1+\exp(-z))^2}$
PN	$[\Phi(z)]^{\alpha}$	$\alpha[\Phi(z)]^{\alpha-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}z^2\right\}$
PC	$\left[\frac{1}{\pi}\arctan(z)+\frac{1}{2}\right]^{\alpha}$	$\frac{\alpha}{\pi}\left[\frac{1}{\pi}\arctan(z)+\frac{1}{2}\right]^{\alpha-1}\left[1+z^2\right]^{-1}$
PLA	$\left[\frac{1}{2}+\frac{1}{2}\text{sign}(z)\{1-\exp(-\lvert z\rvert)\}\right]^{\alpha}$	$\alpha\left[\frac{1}{2}+\frac{1}{2}\text{sign}(z)\{1-\exp(-\lvert z\rvert)\}\right]^{\alpha-1}\frac{1}{2}\exp(-\lvert z\rvert)$
PG	$[\exp\{-\exp\{-z\}\}]^{\alpha}$	$\alpha[\exp\{-\exp\{-z\}\}]^{\alpha-1}\exp\{-(z+\exp\{-z\})\}$
PRG	$[1-\exp\{-\exp\{z\}\}]^{\alpha}$	$\alpha[1-\exp\{-\exp\{z\}\}]^{\alpha-1}\exp\{-(-z+\exp\{z\})\}$
RPL	$1-\left[\frac{1}{1+\exp(z)}\right]^{\alpha}$	$\alpha\left[\frac{1}{1+\exp(z)}\right]^{\alpha-1}\frac{\exp(-z)}{(1+\exp(-z))^2}$
RPN	$1-[\Phi(-z)]^{\alpha}$	$\alpha[\Phi(-z)]^{\alpha-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}z^2\right\}$
RPC	$1-\left[\frac{1}{\pi}\arctan(-z)+\frac{1}{2}\right]^{\alpha}$	$\frac{\alpha}{\pi}\left[\frac{1}{\pi}\arctan(-z)+\frac{1}{2}\right]^{\alpha-1}\left[1+z^2\right]^{-1}$
RPLA	$1-\left[\frac{1}{2}-\frac{1}{2}\text{sign}(z)\{1-\exp(-\lvert z\rvert)\}\right]^{\alpha}$	$\alpha\left[\frac{1}{2}-\frac{1}{2}\text{sign}(z)\{1-\exp(-\lvert z\rvert)\}\right]^{\alpha-1}\frac{1}{2}\exp(-\lvert z\rvert)$
RPG	$1-[\exp\{-\exp\{z\}\}]^{\alpha}$	$\alpha\left(e^{-e^{z}}\right)^{\alpha-1}e^{-\left(z+e^{-z}\right)}$
RPRG	$1-[1-\exp\{-\exp\{-z\}\}]^{\alpha}$	$\alpha[1-\exp\{-\exp\{-z\}\}]^{\alpha-1}\exp\{-(-z+\exp\{z\})\}$

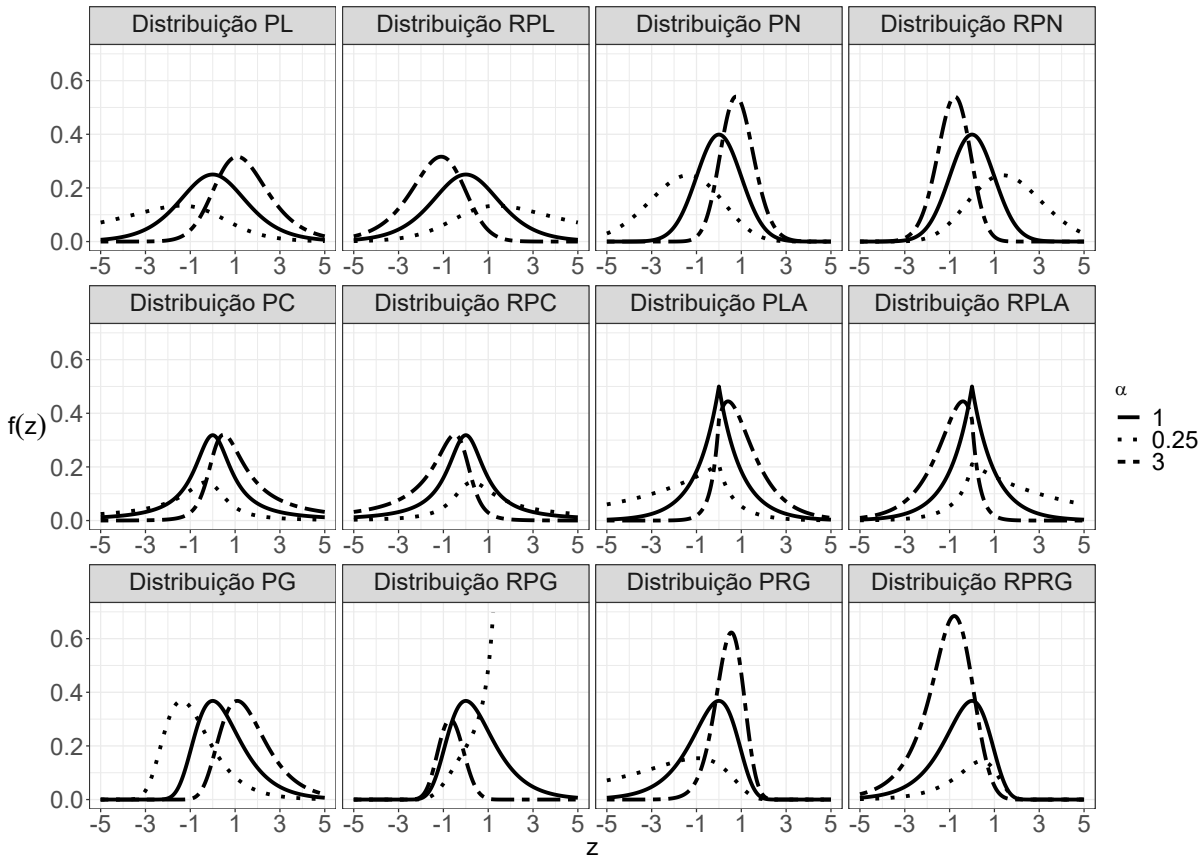

Figura 2 – fdp de algumas distribuições potência e reversa de potência para diferentes valores de α

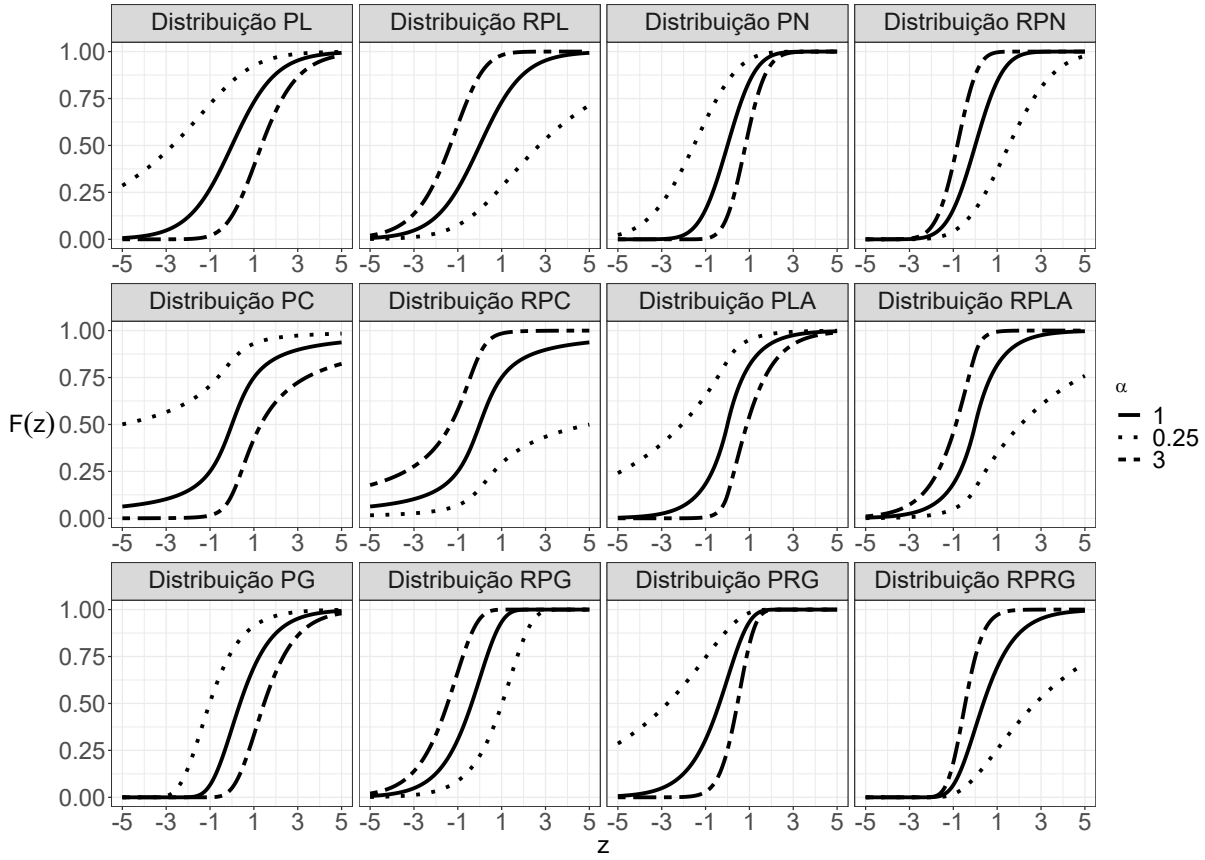

Figura 3 – fda de algumas distribuições potência e reversa de potência para diferentes valores de α

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

$$f_l(z) = \frac{w_l(z)}{\mathbb{E}[w_l(Z)]}\, g(z), \quad l = P,\ RP.$$

2.2 Modelo de regressão para classificação binária

Um problema de classificação pode ser dividida em: classificação binária, multi-classe, multi-rótulo e hierárquicas (SOKOLOVA; LAPALME, 2009). Este trabalho está direcionado ao estudo de problemas de classificação binária.

2.2.1 Classificação binária

Em problemas de classificação binária (mais detalhes em Duda, Hart *et al.* (2001)), a variável de interesse, habitualmente denominada variável resposta, é atribuída em uma de duas classes ou categorias. Geralmente, uma das classes é chamada de "classe positiva" e assume o valor 1 o qual representa a ocorrência do evento de interesse ("sucesso") e a outra classe, é chamada de classe negativa e assume o valor 0 o qual representa a ocorrência do evento

complementar ("fracasso"). É chamado de "binária" porque existem apenas dois resultados (categorias) ou rótulos distintos.

A variável de interesse está comumente associada a outras variáveis denominadas como: variáveis independentes, variáveis explicativas, atributos ou covariáveis; essas variáveis podem ser contínuas, discretas ou categóricas. A probabilidade de ocorrência de uma classe, pode ser explicada por estas outras variáveis.

Existem vários métodos ou técnicas que podem ser usados para classificação binária (BROWNLEE, 2016), incluindo regressão logística, máquinas de vetores de suporte (SVM), árvores de decisão, florestas aleatórias e redes neurais.
A regressão logística, que faz parte de regressão binária, é um dos métodos mais populares para classificação binária. O objetivo da regressão logística é identificar o melhor ajuste entre uma variável dependente e uma coleção de variáveis independentes, e através deste ajuste gerar valores entre 0 e 1, que podem ser interpretados como as probabilidades de cada observação pertencer a uma determinada classe.

Em geral, a probabilidade de observar uma classe dado uma ou mais variáveis independentes, pode ser estimada através de qualquer modelo de regressão binária (por exemplo a regressão logística), o qual tem um componente chamada "função de ligação". Portanto, um modelo de regressão binária pode ser usado para problemas de classificação.

2.2.2 Modelo de regressão binária

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

$$\begin{aligned} Y_i \mid \boldsymbol{\beta} &\overset{ind.}{\sim} \text{Bernoulli}(\mu_i) \\ u_i = \mathbb{P}(Y_i = 1) &= F(\eta_i) \\ \eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} &= \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \end{aligned} \tag{2.5}$$

em que η_i é o i-ésimo preditor linear e $F(\cdot)$ denota a função de distribuição acumulada (fda), que é invertível. O vetor de parâmetros $\boldsymbol{\beta}$ podem ser considerados independente com uma distribuição a priori $\pi(\boldsymbol{\beta}) \equiv N_p(\mathbf{0}, \boldsymbol{\Sigma})$, um caso particular é $\boldsymbol{\Sigma} = \boldsymbol{I}\sigma_\beta^2$ (BAZÁN *et al.*, 2017).

No âmbito de modelos lineares generalizados (MLG) (mais detalhes em McCullagh e Nelder, 1989), F^{-1} é chamada de função de ligação, esta função lineariza a relação entre a media da variável resposta e as variáveis independentes.
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, probito e cauchito, são simétricas. Por outro lado, quando $F(\cdot)$ é a fda de uma distribuição que não é simétrica, obtemos as funções de ligação assimétricas, por exemplo cloglog e loglog.

As funções de ligação comuns consideradas como parte de modelos lineares generalizados (MLG) são mostradas na Tabela 4 e as suas curvas de probabilidade na Figura 4. A partir da Figura 4, podemos ver que a curva é simétrica em torno de $\mu = 0,5$ e $\eta = 0$, para as funções de ligação probito, logito e cauchito; no entanto, para as ligações cloglog e loglog, são assimétricas em torno de $\eta = 0$.

Tabela 4 – Funções de ligação comuns na regressão binária.

Função de ligação	η_i	$\mu_i = F(\eta_i) = F(\boldsymbol{x}_i^\top \boldsymbol{\beta})$
Logito	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$
Probito	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Cauchito	$F^{-1}(\mu_i)$	$\frac{1}{2}+\frac{\arctan(\eta_i)}{\pi}$
Cloglog	$\log(-\log(1-\mu_i))$	$1-\exp\{-(\eta_i)\}$
Loglog	$\log(\log(\mu_i))$	$\exp\{\exp(\eta_i)\}$

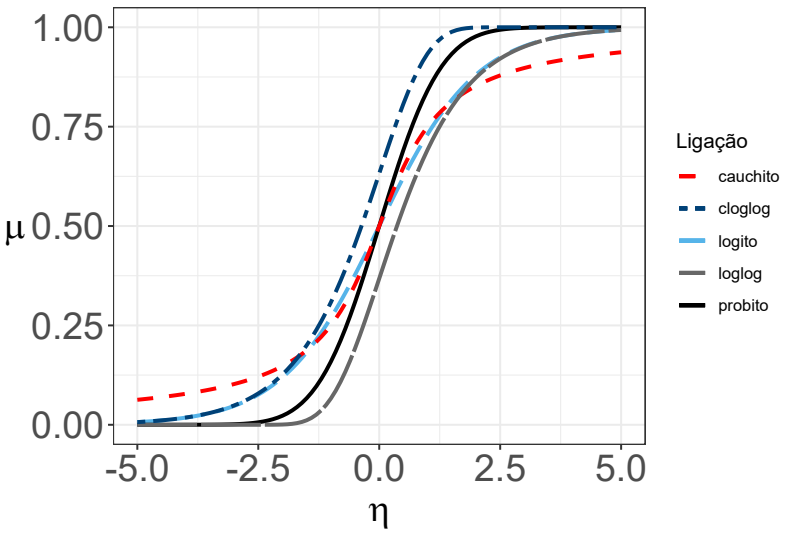

Figura 4 – Curva de funções de ligação comuns na regressão binária.

Assim, 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 assuma 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 função de verossimilhança associada ao modelo de regressão binária, tem a seguinte forma

$$L(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[F\left(\eta_i\right)\right]^{y_i} \left[1 - F\left(\eta_i\right)\right]^{1-y_i}.$$

Portanto, considerando a especificação das a prioris para os parâmetros, a distribuição a posteriori de $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X})$ é dada pela seguinte expressão

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) &\propto L\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}\right) \pi\left(\boldsymbol{\beta}\right) \\ &\propto \prod_{i=1}^{n} \left[F\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right]^{1-y_i} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right). \end{aligned} \tag{2.6}$$

Para o caso em que a função de ligação é um logito (regressão logística), a função de verossimilhança é dada por

$$L(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[\frac{\exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)}\right]^{y_i} \left[\frac{1}{1 + \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)}\right]^{1-y_i} = \frac{\exp\left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)}{\prod_{i=1}^{n}\left[1 + \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right]}.$$

Logo, a sua distribuição a posteriori é dada por

$$\pi(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y}, \boldsymbol{X}) \propto \frac{\exp\left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)}{\prod_{i=1}^{n}\left[1 + \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right]} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right) = \frac{\exp\left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^{\top}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right)}{\prod_{i=1}^{n}\left[1 + \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right]}.$$

Para muitas funções de ligação, a distribuição a posteriori, pode não ter uma expressão de forma fechada o pode não pertence a uma família de distribuições conhecidas. Assim, para simular esta distribuição, os algoritmos Monte Carlo via cadeias de Markov podem ser usados. Por exemplo, em Bazán e Bayes (2010) estudam a estimação Bayesiana em modelos de regressão binária.

2.2.3 Desbalanceamento de dados em classificação binária

Na literatura, até onde sabemos, não existe uma definição formal de desbalanceamento de dados no contexto de classificação binária; porém, algumas ideias foram introduzidas; por exemplo, para Zhai, Qi e Shen (2022) e Paal (2014), um problema de classificação binária desbalanceada ocorre quando a classe majoritária (classe negativa) é significativamente maior que a classe minoritária (classe positiva). Em outras palavras, o número de observações pertencentes a uma das classe é significativamente menor do que as pertencentes às outras classes. Também,

Brownlee (2020) entende que um problema de classificação desbalanceada ocorre quando as observações entre as classes não é igual, e esta diferença pode ser leve ou até severo, é dizer, as observações de uma classe minoritária pode variar para centenas, milhares ou milhões de obervações da classe ou classes majoritárias. Por outro lado, em Da Silva, ANYOSA e Bazán (2020) introduzem a definição 5.

Definição 5. Dizemos que uma variável resposta binária Y é desbalanceada, se e somente se, $\kappa := |2\mu - 1| \geq 0,2$, em que μ é a probabilidade de sucesso.

Nos últimos anos, alguns autores, por exemplo, De la Cruz *et al.* (2019), Fatourechi *et al.* (2008) e Luque *et al.* (2019), mostraram que na regressão binária, quando os dados são desbalanceados, usar funções de ligação simétrica, pode não ser adequado, e que as funções de ligação assimétrica debe ser considerado. Por exemplo, Chen, Dey e Shao (1999), Wang e Dey (2011), Yin *et al.* (2020), Bazán *et al.* (2017) e Lemonte e Bazán (2018), estudam a regressão binária considerando diferentes funções de ligação assimétrica.
Diante disso, nos capítulos seguintes, estudamos o problema de classificação binária, através de regressão binária com funções de ligação assimétrica.

2.3 Comparação de modelos sob abordagem Bayesiana

O objetivo de comparar os modelos é encontrar um modelo ótimo. A escolha desse modelo ótimo, é a tarefa de selecionar um modelo entre vários candidatos com base em algum critério de desempenho para escolher o melhor modelo. No contexto de modelos de regressão binária, podemos entender como selecionar um modelo estatístico de uma classe de modelo (um conjunto de modelos candidatos), dado um conjunto de dados.
Um dos objetivos de selecionar um modelo, é poder ser fazer inferências acerca dos parâmetros do modelo, e para esse propósito, existem critérios para comparar o desempenho dos modelos; em Ding, Tarokh e Yang (2018) descrevem alguns critérios. Por outro lado, se o objetivo é selecionar um modelo baseado na capacidade preditiva, então existem algumas métricas que são construídas baseado na predição de uma determinada classe da variável resposta binária. Nesta seção, descrevemos as duas formas de escolher um modelo.

2.3.1 Seleção baseado em métricas para classificação

Na classificação binária, geralmente, a classe positiva é aquela que estamos interessados em prever. Portanto, medir com que precisão o modelo prevê o resultado desejado é crucial no momento de selecionar um modelo. A avaliação do desempenho de um modelo de classificação, é baseada nas contagens de registros previstas como correta e incorretamente pelo modelo.

Em geral, uma métrica para classificação pode ser descrita como uma ferramenta de medição, que mede o desempenho do classificador (modelo) (HOSSIN; SULAIMAN, 2015).

Neste sentido, dentro do contexto de classificação binária, existem diferentes métricas, cada uma delas construída baseada na matriz de confusão. Algumas das métricas são estudadas em Metz (1978), Sokolova e Lapalme (2009), Choi, Cha e Tappert (2010), De la Cruz *et al.* (2019) e Vujović *et al.* (2021).

2.3.1.1 Matriz de confusão

A matriz de confusão pode se entender como uma matriz bidimensional, indexada em uma dimensão pela verdadeira classe de um objeto e na outra pela classe que o classificador atribui (TING, 2017). Um caso particular, é quando existem duas classes, uma designada como classe positiva e a outra como classe negativa, como mostrado na Tabela 5. Nesse contexto, as quatro células da matriz são designadas como: verdadeiros positivos (TP, do inglês: *True Positives*) que são as observações positivas e previstas como positivas, falsos positivos (FP, do inglês: *False Positives*) que são as observações negativas, mas previstas como positivas, verdadeiros negativos (TN, do inglês: *True Negatives*) que são as observações e previstas como negativas e falsos negativos (FN, do inglês: *False Negatives*) que são as observações positivas, mas previstas como negativas. Mais detalhes pode ser vistos em Fawcett (2006).

Tabela 5 – Matriz de confusão

	Valor predito	
Valor observado	0	1
0	TN (Verdadeiro negativo)	FP (Falso positivo)
1	FN (Falso negativo)	TP (Verdadeiro positivo)

2.3.1.2 Métricas para classificação binária

A partir da matriz de confusão, várias métricas de desempenho de classificação podem ser definidas com base nos quatro elementos da matriz (TN, FP, FN e TP). As medidas comuns e bastante utilizadas são: Acurácia (taxa de boa classificação), Sensibilidade, Especificidade, Verdadeiro Preditivo Positivo e Verdadeiro Preditivo Negativo que são definidas, respetivamente, 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.

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

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

evento $(Y = 0)$.

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

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

Por outro lado, como indicado por Choi, Cha e Tappert (2010), nos casos em que as classes binárias são desbalanceadas, como no caso assimétrico de dados binários, as correspondências positivas (eventos) são geralmente mais significativas do que as correspondências negativas (não eventos). Nesses casos, algumas medidas de similaridade assimétricas devem ser consideradas, por exemplo, o Índice Jaccard (em inglês *Jaccard index*) também chamado de índice crítico de sucesso (*CSI*, do inglês: *critical success index*), *Índice de Sokal e Sneath* (*SSI*, do inglês: *Sokal & Sneath index*), índice de confiança (*FAITH*, do inglês: *Faith index*) e diferença padrão (*PDIF*, do inglês: *pattern difference*) para medir a similaridade entre a classificação observada e a prevista usando a matriz de confusão correspondente (De la Cruz, 2019).

Além disso, a pontuação de habilidade de Gilbert (*GS*, do inglês: *Gilbert skill score*) proposto por Schaefer (1990), modifica o *CSI* para lidar com os problemas associados ao valor muito grande de TN, o que acontece claramente na predição de classes desbalanceadas.

Neste trabalho, consideramos as métricas para classificação apresentadas e utilizadas em De la Cruz *et al.* (2019). Além disso, usamos a pontuação F_1 ($F1$), Coeficiente de Correlação de Matthews (*MCC*), *G*-Média (*GM*) e Kappa de Cohen (*KAPPA*), que são sugeridos quando as classes são desbalanceadas. Todas essas métricas são mostradas na Tabela 6.

Para todas essas métricas, 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 pelo modelo.

Para obter os valores de cada uma das métricas, é necessário estabelecer um ponto de corte o qual é uma probabilidade predita da classe de interesse, é dizer, $\widehat{\mathbb{P}}(Y = 1)$. Neste trabalho, serão usadas as médias a posteriori como estimador *plug-in* para a probabilidade predita.

2.3.1.3 Ponto de corte

Um ponto de corte (em inglês *threshold*), é um valor limite que serve para classificar a observação variável resposta em uma de duas classes. Para todos valores iguais ou maiores que o ponto de corte, a variável resposta é atribuída a uma classe (classe 0) e para todos os outros valores, a variável resposta é atribuída à outra classe (classe 1). É comum usar o valor de 0.5, no entanto, 0.5 não é ideal para alguns casos, particularmente para conjuntos de dados desbalanceados Zou *et al.* (2016). Uma abordagem simples e direta para melhorar o desempenho de um classificador que prediz probabilidades em um problema de classificação desequilibrada é ajustar o ponto de corte usado para atribuir probabilidades aos rótulos de classe, o ajuste desse hiper-parâmetro é chamado de movimento de ponto de corte (BROWNLEE, 2020, Capítulo 6).

Tabela 6 – Métricas na classificação binária

Métrica	Notação	Fórmula	Faixa de valores
Acurácia	ACC	$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$	$[0;1]$
Sensibilidade	TPR	$\frac{\text{TP}}{\text{TP+FN}}$	$[0;1]$
Especificidade	TNR	$\frac{\text{TN}}{\text{TN+FP}}$	$[0;1]$
Índice crítico de sucesso	CSI	$\frac{TP}{TP+FP+FN}$	$[0;1]$
Índice de Sokal e Sneath	SSI	$\frac{TP}{TP+2\times FP+2\times FN}$	$[0;1]$
Índice de confiança	$FAITH$	$\frac{TP+0.5\times TN}{TP+FP+FN+TN}$	$[0;1]$
Diferença padrão	$PDIF$	$\frac{4\times FP\times FN}{(TP+FP+FN+TN)^2}$	$[0;1]$ *
Pontuação de habilidade de Gilbert	GS	$\frac{(TP\times TN-FP\times FN)}{(FN+FP)(TP+FP+FN+TN)+(TP\times TN-FP\times FN)}$	$[0;1]$ *
Coeficiente de Correlação de Matthews	MCC	$\frac{(TP\times TN-FP\times FN)}{\sqrt{(FN+FP)(TP+FN)(TN+FP)(TN+FN)}}$	$[0;1]$
G-média	GM	$\sqrt{TPR\times TNR}$	$[0;1]$
Pontuação F_1	$F1$	$2\times\frac{TNR\times TPR}{TNR+TPR}$	$[0;1]$
Capa de Cohen	$KAPPA$	$\frac{2\times(TP\times TN-FP\times FN)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)}$	$[0;1]$

* $SPDIF = 1 - PDF$ e $SGS = \frac{3\times GS+1}{4}$.

O ponto de corte de probabilidade para classificação não irá interferir no valor da AUC. Para decidir qual será o ponto de corte ótimo, um ponto de corte com a melhor pontuação de algumas das métricas pode ser considerado, por exemplo, a melhor pontuação de $KAPPA$, GM ou $F1$.

Seja p_c o valor do ponto de corte estabelecido, então a atribuição de uma observação a uma classe, é da seguinte forma

$$y_i \in \begin{cases} 0 \text{ (classe 0)}, & \text{se } \mu_i < p_c \\ 1 \text{ (classe 1)}, & \text{se } \mu_i \geq p_c \end{cases}$$

Por exemplo, se o ponto de corte for $p_c = 0,5$, então toda vez que o valor da probabilidade $\mu_i \geq 0,5$, aquela observação será atribuída à classe 1, caso contrário será atribuída à classe 0.

2.3.2 Seleção baseado em critérios de comparação de modelos

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

$$D(\boldsymbol{\beta}) = -2\sum_{i=1}^{n} \log\left(f\left(\boldsymbol{y} \mid \boldsymbol{\beta}\right)\right).$$

A média a posteriori do desvio $E\left(D(\boldsymbol{\beta})\right) = E\left[-2\sum_{i=1}^{n} \log\left(f\left(\boldsymbol{y} \mid \boldsymbol{\beta}\right)\right)\right]$, pode ser aproximada computacionalmente por

$$\bar{D} = \frac{1}{M}\sum_{m=1}^{M} D\left(\boldsymbol{\beta}^{(m)}\right), \tag{2.7}$$

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 $D\left(E\left(\boldsymbol{\beta}\right)\right)$ é obtido por meio da média dos valores gerados a partir da distribuição a posteriori como:

$$\widehat{D} = D\left(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\beta}^{(m)}\right). \tag{2.8}$$

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

$$\widehat{\rho}_D = \bar{D} - \widehat{D}.$$

Alguns critérios principais de comparação de modelos, são descritos a seguir:

EAIC e EBIC

Critério de informação esperado de Akaike (*EAIC*, do inglês: *Expected Akaike Information Criterion*) e Critério de Informação Bayesiano Esperado (*EBIC*, do inglês: *Expected Bayesian Information Criterion*), são apresentados em Spiegelhalter *et al.* (2002). Ambos os critérios estão baseados na média a posteriori do desvio e podem ser estimados, respectivamente, por

$$\widehat{EAIC} = \bar{D} + 2P \quad \text{e} \quad \widehat{EBIC} = \bar{D} + P\log\left(n\right). \tag{2.9}$$

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

DIC

O Critério de informação de desvio (*DIC*, do inglês: *Deviance Information Criterion*) 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:

$$\widehat{DIC} = \bar{D} + \widehat{\rho}_D = 2\bar{D} - \widehat{D}. \tag{2.10}$$

Os modelos com menor *DIC* devem ser os preferidos.

WAIC

O critério de informação amplamente aplicável (*WAIC*, do inglês: *Widely Applicable Information Criterion*) é uma abordagem totalmente bayesiana. Foi introduzido por Watanabe (2010), e a ideia é calcular o logaritmo da densidade preditiva pontual ($lppd$) dado por

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

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

$$\widehat{p_{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{M}\sum_{m=1}^{M} f\left(y_i \mid \boldsymbol{\beta}^{(m)}\right)\right) - \frac{1}{M}\sum_{m=1}^{M}\log\left(f\left(y_i \mid \boldsymbol{\beta}^{(m)}\right)\right)\right).$$

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

$$\widehat{WAIC} = -2\left(\widehat{lppd} - \widehat{p_{WAIC}}\right). \tag{2.11}$$

LOO

Outro método de comparação de modelo totalmente bayesiano é o método de validação cruzada de exclusão (*LOO*, do inglês: *leave-one-out cross-validation*) 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 amostragem de importância suavizada de Pareto (*PSIS*, do inglês: *Pareto smoothed importance sampling*), 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* (nomeada como $PSIS-LOO$) é dada por:

$$\widehat{elpd}_{psis-loo} = \sum_{i=1}^{n}\log\left(\frac{\sum_{m=1}^{M} w_i^{(m)} p\left(y_i \mid \boldsymbol{\beta}^{(m)}\right)}{\sum_{m=1}^{M} w_i^{(m)}}\right). \tag{2.12}$$

em que

$$w_i^{(m)} = \min(r_i^{(m)}, \frac{\sqrt{M}}{M}\sum_{m=1}^{M} r^{(m)}) \quad \text{e} \quad r_i^{(m)} = \frac{1}{p\left(y_i \mid \boldsymbol{\beta}^{(m)}\right)} \propto \frac{f\left(\boldsymbol{\beta}^{(m)} \mid \boldsymbol{Y}_{-i}\right)}{f\left(\boldsymbol{\beta}^{(m)} \mid \boldsymbol{Y}\right)}.$$

Luo e Al-Harbi (2017) mostraram que como *EAIC*, *EBIC* e *DIC* usam estimativas pontuais para o seu cálculo, enquanto *LOO* e *WAIC* são calculados com base em toda a distribuição a posteriori, que os métodos que usam mais informação (a distribuição a posteriori) 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 posteriori, o *DIC* vem o segundo devido ao seu uso parcial da distribuição a posteriori, e os outros métodos que não usam a distribuição a posteriori 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).

2.4 Regressão binária mista

O estudo de modelos lineares generalizados mistos (MLGMs), em particular regressão binária mista, é uma metodologia que estuda o comportamento de uma variável de interesse medida repetidamente. Esta metodologia procura medir o efeito da dependência entre a variável resposta e as variáveis explicativas, como também, medir possíveis efeitos entre e/ou intra-indivíduos. Em Diggle, Liang e Zeger (1994) apresentam esta metodologia de forma detalhada e completa.

O foco principal desse tipo de análise, está relacionado com a modelagem da estrutura de correlação intra-indivíduos decorrente de medir a mesma unidade de observação mais de uma vez e, como resultado, as respostas não são independentes como na análise de regressão usual.

Existem diferentes métodos para modelar a estrutura de correlação, por exemplo, Diggle (2002) apresenta três extensões dos modelos lineares generalizados para dados longitudinais, incorporando a dependência entre as observações ao longo do tempo: (i) modelos marginais , (ii) modelos de transição (ou condicional), e (iii) modelos de efeitos mistos. O modelo marginal requer a especificação da média marginal, uma função de variância e a estrutura de correlação das observações dentro do assunto, como auto-regressiva (AR), permutável (EX) e não estruturada (UN). O modelo de transição é uma extensão do modelo linear generalizado (GLM) e é usado em situações onde a variável de resposta tem uma forte conexão com as variáveis de resposta anteriores. Finalmente, o modelo de efeitos mistos inclui um efeito aleatório específico do sujeito que explica as observações correlacionadas do sujeito. A introdução de efeitos aleatórios no modelo de efeitos mistos induz uma correlação marginal entre as observações específicas do sujeito (FITZMAURICE; LAIRD; WARE, 2012).

O foco deste trabalho são os modelos para dados longitudinais os quais são um caso especial dos modelos mistos, para isto, o modelo hierárquico misto é considerado. Neste método, os efeitos aleatórios são adicionados ao preditor linear considerando duas fontes de variação nos dados: a variação entre unidades e a variação dentro das unidades.

Em geral, quando a variável resposta é binária, a variável de interesse é uma variável categórica com duas classes ou categorias, isto é, uma variável dicotômica.

Seja $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ um vetor de uma sequência de medições observadas ao longo do tempo para o i-ésimo sujeito, $i = 1, \cdots, n$, em que cada componente y_{ij}, que assume valor

0 ou 1, corresponde à observação do sujeito i, medida no tempo t_j, $j = 1, ..., n_i$. Além disso, considere $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^\top$ um vetor $q \times 1$ dos efeitos aleatórios específicos do indivíduo i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ vetor de coeficientes de regressão (efeitos fixos) para $j = 1, \ldots, n_i$ e $i = 1, \ldots, n$. O modelo de regressão binária longitudinal pode ser escrito da seguinte forma

$$\begin{aligned} Y_{ij} \mid \boldsymbol{b}_i &\sim \text{Bernoulli}(\mu_{ij}) \\ \mu_{ij} = F(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i) \quad &\text{or} \quad \eta_{ij} = F^{-1}(\mu_{ij}) \\ \boldsymbol{b}_i &\sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b), \end{aligned} \tag{2.13}$$

em que $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ e $\boldsymbol{Z}_{ij} = (z_{ij1}, \ldots, z_{ijq})^\top$ são vetores covariáveis, $\boldsymbol{\Sigma}_b$ é uma matriz definida positiva $q \times q$ e η_{ij} é nomeado como o i-ésimo preditor linear pela j-ésima vez, para $i = 1, \cdots, n$ e $j = 1, \cdots, n_i$. $F(\cdot)$ função de distribuição acumulada de alguma variável aleatória continua com suporte em $\mathbb{R}$ e F^{-1} é conhecida como uma função de ligação.
Além disso, os efeitos aleatórios são considerados "independentes" das covariáveis $\boldsymbol{X}_{ij}$ Fitzmaurice, Laird e Ware (2012, Capítulo 14).

Observe que, semelhante ao caso de regressão binária, diferentes funções de ligação podem ser consideradas, por exemplo, pode-se usar as funções de ligação mostradas na Tabela 4. Quando é considerada a função de ligação logística (modelo logístico), temos:

$$\begin{aligned} \text{logito}(\mu_{ij}) &= \left(\frac{\mu_{ij}}{1 - \mu_{ij}} \right) = \boldsymbol{X}_{ij}^\top \beta + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i \\ \Rightarrow \mu_{ij} = F(\boldsymbol{X}_{ij}^\top \beta + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i) &= \frac{\exp\left(\boldsymbol{X}_{ij}^\top \beta + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^\top \beta + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i\right)} \end{aligned}$$

O modelo Bayesiano misto de regressão binária para dados longitudinais, tem a seguinte expressão

$$\begin{aligned} Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta} &\overset{ind.}{\sim} \text{Bernoulli}(\mu_{ij}), \quad \mathbf{b}_i \overset{ind.}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma}_b) \\ \mu_{ij} &= F(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i) \\ (\boldsymbol{\beta}, \boldsymbol{\Sigma}_b)^\top &\sim \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}_b) \end{aligned} \tag{2.14}$$

Os parâmetros $\boldsymbol{\beta}, \boldsymbol{\Sigma}_b$ podem ser considerado independente tal que $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}_b) = \pi(\boldsymbol{\beta})\,\pi(\boldsymbol{\Sigma}_b)$, com $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, um caso particular é $\boldsymbol{\Sigma} = \boldsymbol{I}\sigma_\beta^2$ (BAZÁN *et al.*, 2017) .

Para o parâmetro de efeito aleatório, é possível utilizar uma das várias especificações a priori presentes na literatura, por exemplo, uma distribuição de Wishart invertida como em Fong, Rue e Wakefield (2010), ou seja, $\Sigma_b \sim IW_q(\psi, c)$, ou, para um determinado componente de Σ_b, σ_b^2, pode ser considerado $\sigma_b^2 \sim$ Inverse-gamma$(0{,}001,\ 0{,}001)$ (LUNN *et al.*, 2000), $\frac{1}{\sigma_b^2} \sim$ Pareto$(1,\ 0{,}001)$ (SCURRAH; PALMER; BURTON, 2000) ou $\log(\sigma_b^2) \sim$ Uniforme$(-10,\ 10)$ (SPIEGELHALTER, 2001).

A função de verossimilhança associada ao modelo é dada por

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}_b, \boldsymbol{b} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[F\left(\eta_{ij}\right)\right]^{y_{ij}} \left[1 - F\left(\eta_{ij}\right)\right]^{1-y_{ij}} \phi_q\left(\boldsymbol{b}_i \mid \mathbf{0}, \boldsymbol{\Sigma}_b\right),$$

em que $\boldsymbol{b} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_n^\top)^\top$ e $\phi_q(\cdot \mid \mathbf{m}, \mathbf{S})$ denotam a fdp da distribuição normal q-variada com vetor de media $\mathbf{m}$ e matriz de covariância $\mathbf{S}$ e $\boldsymbol{y}$ denota a matriz de dados observados da variável de resposta.

Assim, a distribuição a posteriori de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\Sigma}_b)^\top$, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$, é da seguinte forma:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[F\left(\eta_{ij}\right)\right]^{y_{ij}} \left[1 - F\left(\eta_{ij}\right)\right]^{1-y_{ij}} \phi_q\left(\boldsymbol{b}_i \mid \mathbf{0}, \boldsymbol{\Sigma}_b\right) \right) \pi\left(\boldsymbol{\beta}\right) \pi\left(\boldsymbol{\Sigma}_b\right).$$

Observe que a distribuição a posteriori pode não ter uma forma fechada ou pode não pertencer a uma família de distribuição conhecida, o que torna esta distribuição analiticamente intratável. Uma alternativa é o uso de métodos de simulação para obtenção de amostras da distribuição a posteriori. Em particular, pode se utilizar os métodos de *MCMC*, e para este processo, existem diferentes pacotes implementados nos programas estatísticos, por exemplo, a linguagem `Stan` (CARPENTER *et al.*, 2017) por meio de `Python` usando o pacote `Pystan` (TEAM, 2017).

CAPÍTULO

3

NOVAS PROPRIEDADES DAS DISTRIBUIÇÕES POTÊNCIA E REVERSA DE POTÊNCIA EM REGRESSÃO BINÁRIA

Neste capítulo, discutimos sobre a viabilidade das funções de ligação potência e reversa de potência na regressão binária. Mostramos que as funções de ligações potência loglog e reversa de potência loglog são inviáveis devido à sua inflexibilidade.
Além disso, apresentamos as funções de ligação assimétrica potência e reversa de potência como uma alternativa para classificação binária no contexto de dados desbalanceados.

Por outro lado, estudamos algumas métricas que podem ser adequadas para avaliar a capacidade preditiva de um modelo, quando os dados são desbalanceados. Neste capítulo, para estimação dos parâmetros, consideramos a abordagem bayesiana com distribuição a priori dos parâmetros sendo não informativas. Para obter uma estimativa da distribuição a posteriori dos parâmetros, usamos a linguagem `Stan` (TEAM *et al.*, 2016) através de `Python` usando o pacote `Pystan` (VANROSSUM, 1995), que considera alguns dos os métodos de amostragem mais utilizados, incluindo o método *No-U-Turn Sampler* (NUTS), que é uma extensão do Monte Carlo Hamiltoniano (HMC) (HOFFMAN; GELMAN *et al.*, 2014).

3.1 Distribuição potência e reversa de potência

No Capítulo 2, foi mostrado que uma variável aleatória Z segue uma distribuição potência e reversa de potência, em sua forma padrão, quando sua fda tem a seguinte forma, respectivamente

$$F_P(z) = G(z)^{\alpha} \quad \text{e} \quad F_{RP}(z) = 1 - G(-z)^{\alpha}, \quad z \in \mathbb{R},$$

em que α é um parâmetro de forma e $G(\cdot)$ denota uma fda de distribuição de linha de base com suporte na reta real ($\mathbb{R}$).

Foi visto também, que algumas distribuições de linha de base possuem um parâmetro extra (por exemplo, os graus de liberdade de t-Student), neste trabalho, as distribuições que possuem um parâmetro extra não são consideradas, pelas razões já mencionadas.

Além disso, vimos que as funções de densidade de probabilidade (fdp) das distribuições potência e reversa de potência são dadas, respectivamente, por $f_P(z) = \alpha G(z)^{\alpha-1} g(z)$ e $f_{RP}(z) = \alpha G(-z)^{\alpha-1} g(z)$, em que $g(\cdot)$ é um fdp da distribuição de linha de base correspondente e $z \in \mathbb{R}$. Os quantis das distribuições potência e reversa de potência podem ser escritos respectivamente como $Q_P(p) = G^{-1}\left(-p^{1/\alpha}\right)$ e $Q_{RP}(p) = 1 - G^{-1}\left(-(1-p)^{1/\alpha}\right) = -Q_P(1-p)$, em que p é uma probabilidade dada.

Na Tabela 7, apresentamos as funções distribuição acumulada (fda), suas correspondentes funções densidade de probabilidade (fdp) e funções quantílicas (QF) das distribuições potência e reversa potência consideradas neste trabalho. Observe que se $\alpha = 1$ então o fda, fdp e QF das distribuições de linha de base são considerados um caso particular.

O gráfico de p em função de η, é chamado de curva de resposta ou curva de probabilidade de sucesso. p é considerado também um ponto de inflexão da curva fda e tem centro de forma simétrico em $0,5$ somente quando $\alpha = 1$ e $G(\cdot)$ é ponto simétrico (Logística, Normal, Cauchy, Laplace, etc.) e, por outro lado, se a Gumbel for considerada, o ponto de inflexão é $0,3679$ e $0,6321$ se for considerada a reversa de Gumbel.

Tabela 7 – fda, fdp e QF para distribuições de potência e potência reversa, considerando $\eta \in \mathbb{R}$ e o parâmetro de forma α.

Distribuição	fda	fdp	QF
PL	$\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha}$	$\alpha\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha-1}\frac{e^{-\eta}}{(1+e^{-\eta})^2}$	$\log\left(\frac{p^{1/\alpha}}{1-p^{1/\alpha}}\right)$
RPL	$1-\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{\alpha}$	$\alpha\left(\frac{1}{1+e^{\eta}}\right)^{\alpha-1}\frac{e^{-\eta}}{(1+e^{-\eta})^2}$	$-\log\left(\frac{(1-p)^{1/\alpha}}{1-(1-p)^{1/\alpha}}\right)$
PN	$(\Phi(\eta))^{\alpha}$	$\alpha(\Phi(\eta))^{\alpha-1}\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\eta^2\right\}}$	$\Phi^{-1}\left(p^{1/\alpha}\right)$
RPN	$1-(\Phi(-\eta))^{\alpha}$	$\alpha(\Phi(-\eta))^{\alpha-1}\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\eta^2\right\}}$	$-\Phi^{-1}\left((1-p)^{1/\alpha}\right)$
PC	$\left(0,5+\frac{\arctan(\eta)}{\pi}\right)^{\alpha}$	$\frac{\alpha}{\pi}\left(\frac{1}{\pi}\arctan(\eta)+\frac{1}{2}\right)^{\alpha-1}\frac{1}{(1+\eta^2)}$	$\tan\left(\pi\left(p^{1/\alpha}-0,5\right)\right)$
RPC	$1-\left(0,5+\frac{\arctan(-\eta)}{\pi}\right)^{\alpha}$	$\frac{\alpha}{\pi}\left(\frac{1}{\pi}\arctan(-\eta)+\frac{1}{2}\right)^{\alpha-1}\frac{1}{(1+\eta^2)}$	$-\tan\left(\pi\left((1-p)^{1/\alpha}-0,5\right)\right)$
PLA	$\left\{\frac{1}{2}+\frac{\text{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha}$	$\frac{\alpha}{2}\left\{\frac{1}{2}+\frac{\text{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha-1}e^{-\lvert\eta\rvert}$	$\text{sign}(p^{\frac{1}{\alpha}}-0,5)\ln\left(1-2\left\lvert p^{\frac{1}{\alpha}}-0,5\right\rvert\right)$
RPLA	$1-\left\{\frac{1}{2}-\frac{\text{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha}$	$\frac{\alpha}{2}\left\{\frac{1}{2}-\frac{\text{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha-1}e^{-\lvert\eta\rvert}$	$\text{sign}(0,5-(1-p)^{\frac{1}{\alpha}})\ln\left(1-2\left\lvert(1-p)^{\frac{1}{\alpha}}-0,5\right\rvert\right)$
PG	$\left(e^{-e^{-\eta}}\right)^{\alpha}$	$\alpha\left(e^{-e^{-\eta}}\right)^{\alpha-1}e^{-\left(\eta+e^{-\eta}\right)}$	$-\log\left(-\log\left(p^{1/\alpha}\right)\right)$
RPG	$1-\left(e^{-e^{\eta}}\right)^{\alpha}$	$\alpha\left(e^{-e^{\eta}}\right)^{\alpha-1}e^{-\left(\eta+e^{-\eta}\right)}$	$\log\left(-\log\left((1-p)^{1/\alpha}\right)\right)$
PRG	$\left(1-e^{-e^{\eta}}\right)^{\alpha}$	$\alpha\left(1-e^{-e^{\eta}}\right)^{\alpha-1}e^{-(-\eta+e^{\eta})}$	$\log\left(-\log\left(1-p^{1/\alpha}\right)\right)$
RPRG	$1-\left(1-e^{-e^{-\eta}}\right)^{\alpha}$	$\alpha\left(1-e^{-e^{-\eta}}\right)^{\alpha-1}e^{-(-\eta+e^{\eta})}$	$-\log\left(-\log\left(1-(1-p)^{1/\alpha}\right)\right)$

Para gerar valores de uma distribuições de potência (P) ou reversa de potência (RP), o método inverso pode ser considerado. A seguinte proposição 1 mostra como gerar valores da

classe das distribuições potência e reversa de potência, considerando as funções quantílicas das distribuições de linha de base.

Proposição 1. Seja $U \sim \text{Uniforme}(0,1)$, então $X = Q_P(U) = G^{-1}(U^{1/\alpha})$ segue uma distribuição potência e $X = Q_{RP}(U) = -G^{-1}((1-U)^{1/\alpha})$ segue uma distribuição reversa de potência, em que $Q_P(U)$ e $Q_P(U)$ são os valores dos quantis para U, gerados respectivamente por $F_P(\cdot)$ e $F_{RP}(\cdot)$.

Esta é uma consequência direta da definição desta classe de distribuições e também de que para distribuições contínuas, $F_l(X) = U$ segue uma distribuição uniforme contínua (ROSS, 2012).

Demonstração. Para distribuição potência, temos que $F_P(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(G^{-1}(U^{1/\alpha}) \leq x\right) = \mathbb{P}(U \leq G(x)^{\alpha}) = \mathbb{P}(U \leq F_P(x))$ e para distribuição reversa de potência, temos que $F_{RP}(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(-G^{-1}((1-U)^{1/\alpha}) \leq x\right) = \mathbb{P}(U \leq 1 - G(-x)^{\alpha}) = \mathbb{P}(U \leq F_{RP}(x))$ □

Além disso, de acordo com Bazán *et al.* (2017), para $\alpha < 1$ (ou $\alpha > 1$) a curva correspondente à distribuição potência é geralmente acima (abaixo) da curva correspondente à distribuição da linha de base dentro de um intervalo de valores de η. Além disso, para cada valor de α, a curva da distribuição reversa de potência é um reflexo da curva correspondente da distribuição potência, e então para $\alpha < 1$ (ou $\alpha > 1$) a curva correspondente é geralmente abaixo (acima) da curva correspondente para a curva da linha de base. Em Figura 5, é mostrado um exemplo para o caso da distribuição linha de base Cauchy, para $\alpha = \{0,25,\ 3\}$.

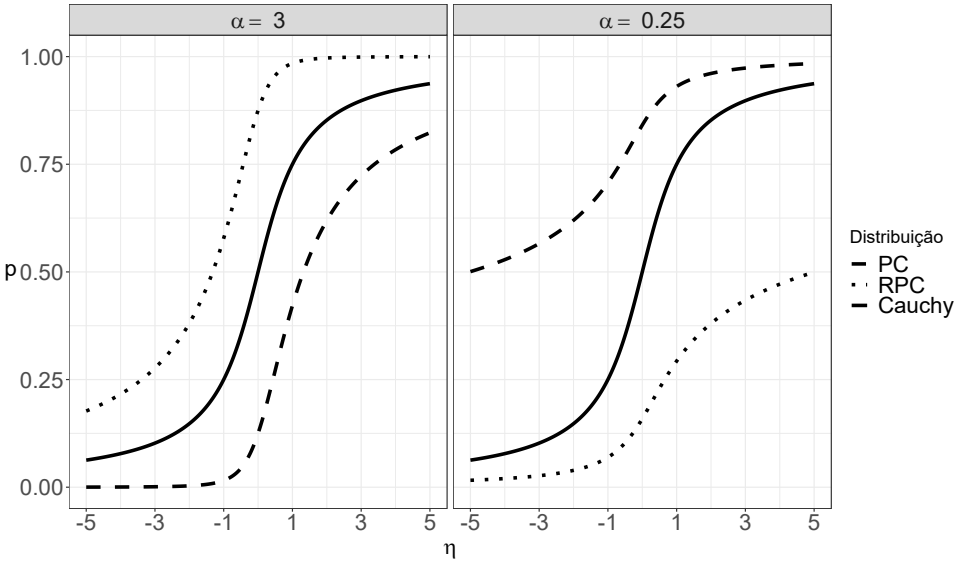

Figura 5 – Curva da resposta do sucesso para Cauchy, Potência Cauchy e distribuição reversa de Potência Cauchy com $\alpha = 3$ (esquerda) e $\alpha = 0,25$ (direita).

3.1.1 Assimetria

Como as distribuições potência e reversa de potência alteram a assimetria e a curtose originais de suas distribuições de linha de base, nesta seção, estudamos o comportamento

dessas medidas para essa classe de distribuições inicialmente estudadas em De la Cruz (2019) e Chumbimune (2017). Calcular o coeficiente de Pearson para assimetria e curtose das distribuições potência e reversa de potência pode ser complicado, pois essas distribuições não têm uma forma fechada para calcular os momentos, veja por exemplo o caso da distribuição de potência normal em Gupta e Gupta (2008). Assim, uma opção conveniente é considerar medidas alternativas baseadas em quantis.

Para definir uma medida conveniente de assimetria, consideramos o coeficiente de assimetria octil definido por Brys, Hubert e Struyf (2004) que é expressada por $A_O = \frac{(O_7-O_4)-(O_4-O_1)}{O_7-O_1}$ em que O_a denota o a^{th} óctil. Neste caso, uma medida de assimetria para as distribuições potência e reversa de potência, que depende do parâmetro α, é dada por

$$A_O(\alpha) = \frac{Q(0,875,\ \alpha) - 2Q(0,5,\ \alpha) + Q(0,125,\ \alpha)}{Q(0,875,\ \alpha) - Q(0,125,\ \alpha)}. \tag{3.1}$$

em que $Q(p,\alpha)$ é a função quantil, $0 < p < 1$, $\alpha > 0$ e $0 < A_O < 1$.

Considerando as expressões quantílicas dadas em Tabela 7, os valores de $A_O(\alpha)$ de cada distribuição para alguns valores α, são mostrados em Tabela 8. Além disso, para cada caso, a amplitude dos valores do coeficiente de assimetria é calculada. A amplitude, denotada por r, é a diferença entre o valor máximo e mínimo do coeficiente de assimetria.

Tabela 8 – $A_O(\alpha)$ para distribuições potência e reversa de potência, considerando valores entre $\alpha = 0,001$ e $\alpha = 9999$.

	$A_O(\alpha)$				
Distribuição	Mínimo	Máximo	$0 < \alpha < 1$	$\alpha \geq 1$	r *
PL	-0,4248	0,1997	(-0,4248; 0,0000)	[0,0000, 0,1997)	0,6245
RPL	-0,1997	0,4248	(0,0000; 0,4248)	[-0,1997; 0,0000)	0,6245
PN	-0,1282	0,1617	(-0,1282; 0,0000)	[0,0000; 0,1617)	0,2899
RPN	-0,1617	0,1282	(0,0000; 0,1282)	[-0,1617; 0,0000)	0,2899
PC	-1,0000	0,7255	(-1,0000; 0,0000)	[0,0000; 0,7255)	1,7255
RPC	-0,7255	1,0000	(0,0000; 1,0000)	[-0,7255; 0,0000)	1,7255
PLA	-0,4248	0,2000	(-0,4248; 0,0000)	[0,1998; 0,2000)	0,2248
RPLA	-0,2000	0,4248	(0,0000; 0,4248)	[-0,2000; -0,1998)	0,2248
PRG	-0,4219	0,1312	(-0,4219; -0,1998)	[-0,1998; 0,1312)	0,5531
RPRG	-0,1312	0,4219	(0,1998; 0,4219)	[-0,1312; 0,1996)	0,5531
PG	0,1997	0,1997	0,1997	0,1997	0,0000
RPG	-0,1997	-0,1997	-0,1997	-0,1997	0,0000

* r = Máximo − Mínimo.

Observe que os valores da assimetria das distribuições, têm uma amplitude diferente conforme também mostrado em Figura 6 para algumas distribuições. Pode-se observar que as distribuições de PC e RPC apresentam uma amplitude de assimetria maior. Por sua vez, pode-se observar que as distribuições PN e RPN apresentam menor amplitude de assimetria. Observe também que, nas distribuições PG e RPG, a assimetria é constante e não depende do valor de α.

Este resultado será explicado mais adiante. Além disso, para todos os casos, quando $\alpha = 1$, o valor de $A_O(\alpha)$ é zero para as distribuições com uma distribuição de linha de base simétrica (PL, RPL, PN, RPN, PC, RPC) e sempre assimétrica como valor fixo para distribuição de linha de base assimétrica (PRG, RPRG, PG, RPG, PLA, RPLA).

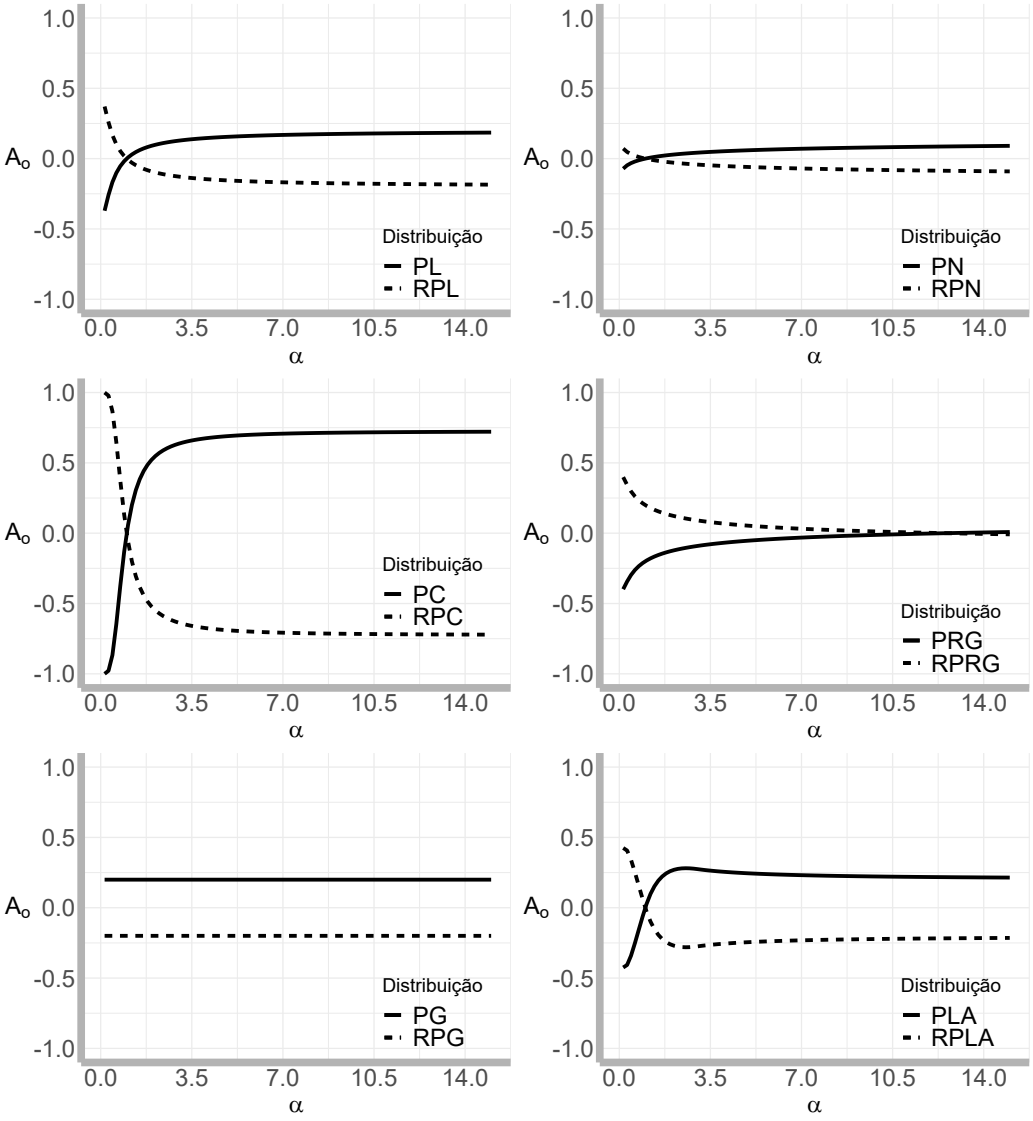

Figura 6 – Coeficiente de assimetria octil das distribuições potência e de sua reversa, para diferentes valores de α.

Além disso, considerando a relação entre a distribuição potência e reversa de potência, é facilmente verificável que $A_O^{RP}(\alpha) = -A_O^P(\alpha)$.

3.1.2 Curtose

Usando um argumento semelhante à medida de assimetria, consideramos o coeficiente de curtose definido por Moors (1988): $K_O = \frac{(O_7 - O_5) + (O_3 - O_1)}{O_6 - O_2}$ com base em octiles, mas considerando que os valores não estão numa escala conveniente, usamos o valor de curtose da

distribuição Normal, que é igual a $1,233$, para redimensionar essa medida e então propomos a seguinte medida de curtose

$$K_O(\alpha) = \frac{100}{1,233} \times \left[\frac{Q(0,875,\ \alpha) - Q(0,625,\ \alpha) + Q(0,375,\ \alpha) - Q(0,125,\ \alpha)}{Q(0,75,\ \alpha) - Q(0,25,\ \alpha)} - 1,233 \right]. \tag{3.2}$$

Considerando as expressões quantílicas dadas na Tabela 7, os valores de $K_O(\alpha)$ para cada distribuição considerando diferentes valores α são mostrados na Tabela 9. Também, para cada caso, a amplitude dos valores do coeficiente de curtose é calculado. A amplitude, denotado por r, é a diferença entre o valor máximo e mínimo do coeficiente de curtose.
Além disso, considerando a relação entre a distribuição potência e reversa de potência, é facilmente verificável que $K_O^{RP}(\alpha) = K_O^{P}(\alpha)$.

Tabela 9 – $K_O(\alpha)$ para distribuições potência e reversa de potência, considerando valores entre $\alpha = 0,001$ e $\alpha = 9999$.

Distribuição	$K_O(\alpha)$						
	Min	Max	$0 < \alpha < 1$		$\alpha \geq 1$		r *
PL-RPL	3,6597	10,1144	(5,9426;	10,1144)	(3,6597;	5,9423)	6,4547
PN-RPN	-2,0454	1,9217	(-2,0454;	0,0077)	(0,0078;	1,9217)	3,9671
PC-RPC	56,5886	37527838,9177	(62,2085;	37527838,9177)	(56,5886;	73,7012)	37527782,3291
PRG-RPRG	0,2981	7,7168	(3,6582;	7,7168)	(0,2981;	0,8706)	7,4187
PPG-RPG	3,6580	3,6580	(3,658;	3,658)	(3,658;	3,658)	0,0000
PLA-RPLA	3,5934	28,8118	(5,9424;	28,8118)	(3,5934;	3,6563)	25,2184

* $r =$ Máximo $-$ Mínimo.

Observe que a amplitude dos valores de K_O são diferentes para cada distribuição. Pode-se observar que as distribuições PC e RPC apresentam maior amplitude de curtose. Por sua vez, pode-se observar que as distribuições PN e RPN apresentam menor amplitude de curtose. Note também que para as distribuições PG e RPG, a curtose é constante e não depende do valor de α, este resultado também será explicado mais adiante.

3.1.3 Observação para potência Gumbel e reversa de potência Gumbel

É importante comentar uma nota sobre a distribuição potência Gumbel e reversa de potência Gumbel. Em Tabela 8 e Tabela 9, pode-se notar que os valores das distribuições PG e RPG são constantes para qualquer valor de α. Além disso, sabemos que a Eq. (3.1) e (3.2) dependem dos valores dos quantis. Nesta seção, apresentamos uma observação sobre a distribuição potência Gumbel e reversa de potência Gumbel. Primeiro, introduzimos a definição da diferença de quantil.

Definition 3.1. Podemos definir a diferença de quantil considerando duas probabilidades diferentes p_1 e p_2, respectivamente, para o primeiro e segundo caso, como sendo $D((p_1,\alpha_1);(p_2,\alpha_2)) = Q(p_1,\alpha_1) - Q(p_2,\alpha_2)$.

Proposition 1. Seja $X_1 \sim f_1 : PG(\alpha_1)$ e $X_2 \sim f_2 : PG(\alpha_2)$. Para as probabilidades p_1 e p_2, respectivamente, a diferença de quantil entre eles é dada por

$$D((p_1,\alpha_1);(p_2,\alpha_2)) = \log\left(\frac{\alpha_1}{\alpha_2}\frac{\log(p_2)}{\log(p_1)}\right). \tag{3.3}$$

Demonstração.

$$\begin{aligned} D((p_1,\alpha_1);(p_2,\alpha_2)) &= Q^{PG}(p_1,\alpha_1) - Q^{PG}(p_2,\alpha_2) \\ &= \left[-\log(-\log(p_1^{1/\alpha_1})\right] - \left[-\log(-\log(p_2^{1/\alpha_2}))\right] = \log\left(\frac{\log\left(p_2^{1/\alpha_2}\right)}{\log\left(p_1^{1/\alpha_1}\right)}\right) \\ &= \log\left(\frac{\alpha_1}{\alpha_2}\frac{\log(p_2)}{\log(p_1)}\right). \end{aligned}$$

□

Na expressão 3.3, pode-se ver que se $p_1 = p_2 = p$, então $D((p,\alpha_1);(p,\alpha_2)) = \log\left(\frac{\alpha_1}{\alpha_2}\right)$, ou seja, a diferença de quantil depende apenas dos valores de α.

A fim de avaliar a (in)viabilidade da distribuição PG como uma função de ligação e uma vez que essa função quantil está associada a uma função de ligação em regressão binária, estamos interessados em saber qual é a distância entre dois quantis para diferentes valores de α, para isto consideramos a distância de *Wasserstein* Villani (2003) que é uma função da diferença de quantil introduzida acima e então temos

$$W^m(f_1,f_2)(p) = \int_0^1 \mid F_1^-(p) - F_2^-(p) \mid^m dp = \int_0^1 \mid D((p,\alpha_1);(p,\alpha_2)) \mid^m dp,$$

em que F^- é a inversa generalizada da fda. Para distribuições de probabilidade unidimensionais, a distância $Wasserstein_m$ é simplesmente a distância L^m de variáveis aleatórias simuladas de uma distribuição uniforme no intervalo $[0,1]$.

Proposition 2. Seja $X_1 \sim f_1 : PG(\alpha_1)$ e $X_2 \sim f_2 : PG(\alpha_2)$. Para a mesma probabilidade p a distância de *Wasserstein* é dada por

$$W^m(f_1,f_2)(p) = \left|\log\left(\frac{\alpha_1}{\alpha_2}\right)\right|^m. \tag{3.4}$$

Demonstração. Usando os valores da função quantil da distribuição PG, dados em Tabela 7, calculamos (3.3) e depois integramos no intervalo unitário. □

Assim, descobrimos que para a potência Gumbel, $W^m(f_1,f_2)(p) = \left|\log\left(\frac{\alpha_1}{\alpha_2}\right)\right|^m$ é constante para qualquer valor de p dependendo apenas dos valores de α. Pode ser mostrado na Figura 7, onde são apresentadas as curvas de probabilidades e quantis da distribuição *PG* com

parâmetro $\alpha = 1$ e $\alpha = 3$. Se for verificada alguma probabilidade p nas duas curvas, a diferença nos valores dos quantis é mantida constante. Na figura 7, a diferença nos valores dos quantis é $0,6931$, independentemente do valor escolhido de p. As linhas foram traçadas para as probabilidades $p = \{0,125,\ 0,5,\ 0,875\}$, e para cada uma, o QF mede $0,6931$.

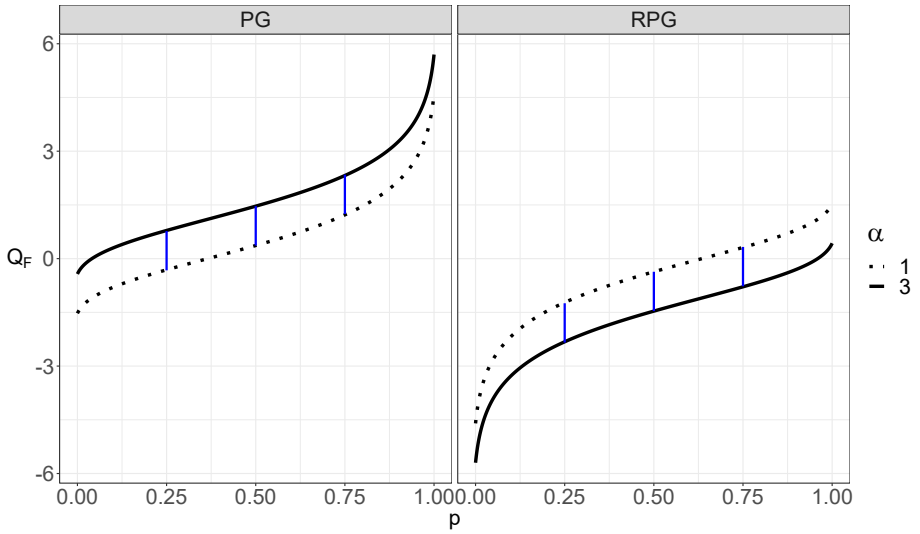

Figura 7 – Probabilidades e quantis da distribuição PG com $\alpha = \{1,\ 3\}$.

Corolário 1. A distância de *Wasserstein*$_m$ para RPG é constante para qualquer valor de p, dependendo apenas dos valores de α.

Demonstração. A prova é análoga à distribuição PG e é omitida aqui. □

Por outro lado, se considerarmos $\alpha_1 = \alpha_2 = \alpha$, então a diferença de quantil para a distribuição PG é $D((p_1,\alpha);(p_2,\alpha)) = \log\left(\frac{\log(p_2)}{\log(p_1)}\right)$ e depende apenas de p's e então α é um parâmetro irrelevante. Este comportamento foi observado quando calculamos $A_O(\alpha)$ e $K_O(\alpha)$ para esta distribuição e então notamos que estes valores são constantes e não dependem de α. Isso pode ser explicado porque quando obtemos esta medida empiricamente, essencialmente usamos diferenças quantílicas do tipo $D((p_1,\alpha);(p_2,\alpha))$ onde o valor desta diferença é constante e por esta razão, essas medidas não dependem de α, como foi verificado. Esses resultados confirmam a inviabilidade da distribuição PG como função de ligação em regressão binária. Um resultado semelhante é obtido para o Reversa de potência Gumbel e então esta distribuição também não é viável.

Essa característica não é observada quando se considera as outras distribuições da classe de distribuições potência es sua reversa e então essas distribuições são distribuições viáveis que poderiam ser usadas como funções de ligação. Nesses casos, a distância de *Wasserstein* depende de p e α_1 e α_2.

3.2 Regressão binária com ligação de potência e reversa de potência

Seja $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ um $n \times 1$ vetor de variáveis de resposta independentes com valores 1 o 0, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ o vetor de covariáveis e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ os coeficientes de regressão associados com Y_i. O modelo de regressão binária Bayesiano com função de ligação potência ou reversa de potência pode ser escrito como

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \alpha &\overset{ind.}{\sim} \text{Bernoulli}(p_i) \\ p_i &= F_l(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \\ (\boldsymbol{\beta}, \alpha)^\top &\sim \pi(\boldsymbol{\beta}, \alpha) \end{aligned} \tag{3.5}$$

em que $F_l(\cdot)$ é a distribuição P ou RP dada em Tabela 7. Conforme descrito em De la Cruz *et al.* (2019) e Bazán *et al.* (2017), para os parâmetros $\boldsymbol{\beta}$ e α podem ser considerados prioris independentes de forma que $\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta})\pi(\alpha)$, consideramos $\boldsymbol{\beta} \sim N_p(\boldsymbol{0}, \boldsymbol{I}\sigma_\beta^2)$ e $\delta = \log(\alpha) \sim U(-2,\ 2)$ seguindo Bazán *et al.* (2017). Para o parâmetro α, escolhemos essa a priori, seguindo os trabalhos de Bazán *et al.* (2017) e De la Cruz *et al.* (2019). Observe que estamos considerando um a priori uniforme limitada para um parâmetro ilimitado, e isso resulta em uma probabilidade maior para um lado da assimetria numa função ligação assimétrica induzida e o comportamento da função de ligação quando $\log(\alpha)$ está próximo de -2 não é semelhante ao seu comportamento próximo de 2 no caso da função de ligação reversa potência. Consideramos, seguindo Gelman (2006), que a densidade a priori uniforme não informativa para o parâmetro de escala pode ser considerada inicialmente e que a família inverse-gamma(c, c) de distribuições a priori não informativas não é recomendada porque em casos onde este parâmetro é estimado em aproximadamente 0, as inferências resultantes serão sensíveis a c.

Reconhecemos que estudos específicos adicionais podem ser considerados para propor outras a prioris como, por exemplo, uma Gama ou apenas outra distribuição positiva. Nesse sentido, recentemente, Ordoñez *et al.* (2023) mostraram que essa a priori recupera muito bem os parâmetros do modelo potência Logística e tem uma probabilidade de cobertura comparável a uma a priori de complexidade penalizada mas pode apresentar um intervalo de credibilidade maior que essa a priori, principalmente para pequenas amostras. Os resultados da Seção de Simulação de nosso trabalho mostram que a priori adotada funcionou relativamente bem.

A função de verossimilhança associada ao modelo tem a seguinte expressão

$$L(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} [F_l(\eta_i)]^{y_i} [1 - F_l(\eta_i)]^{1-y_i} \tag{3.6}$$

e considerando a especificação das a prioris aqui, a distribuição a posteriori de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)^\top$, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ tem a seguinte forma

$$\pi(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{n} \left[F_l\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F_l\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)\right]^{1-y_i} \prod_{j=1}^{p} \exp\left\{-\frac{\beta_j^2}{2(10^2)}\right\} \frac{1}{4\alpha},$$

No entanto, esta distribuição a posteriori não pertence a uma família de distribuições conhecidas e não possui uma expressão de forma fechada. Para simular esta distribuição, o algoritmo *MCMC No-U-Turn Sampler* pode ser considerado (De la Cruz *et al.*, 2019).

Neste trabalho, para estimar os parâmetros dos modelos, é desenvolvido um código na linguagem `Stan` (TEAM *et al.*, 2016) usando o pacote `Pystan` através de `Python` (VANROSSUM, 1995) .

Aqui estudamos apenas as funções de ligação potência porque as distribuições reversa de potência têm os mesmos valores que suas respectivas distribuições potência em termos do coeficiente de curtose e são o reflexo de sua respectiva distribuição potência em termos de sua medida de assimetria.

3.2.1 Avaliação do desempenho dos modelos usando métricas de classificação

As métricas, para avaliar a capacidade preditiva do modelo, usadas aqui, foram mostradas na Tabela 6. A construção dessas métricas para um problema de classificação de duas classes, estão baseadas na matriz de confusão (Tabela 5), considerando classes negativas (classe 0) e positivas (classe 1).

Conforme indicado em De la Cruz *et al.* (2019), para todas as métricas com exceção de *PDIF*, um modelo com o maior valor deve ser preferido em relação a outros modelos possíveis, porque apresenta a maior semelhança entre a classificação observada e prevista. Observe que os valores das métricas *GS* e *PDIF* podem não ser comparáveis entre si; por isso, a transformamos de forma que os valores dessas métricas fiquem entre 0 e 1 e tenham o mesmo significado das outras métricas. A transformação de *GS* e *PDIF*, chamamos, respectivamente, de pontuação de habilidade de Gilbert padronizada (*SGS*) e diferença padrão padronizada (*SPDIF*).

Como foi visto, as métricas anteriores dependem dos valores da matriz de confusão, que são calculados com base em uma probabilidade ou pontuação de pertencer a uma classe, isso é obtido através do uso de um ponto de corte (em inglês *threshold*), é comum usar o valor de 0.5. No entanto, 0.5 não é ideal para conjuntos de dados desbalanceados Zou *et al.* (2016). Como foi mencionado, uma abordagem simples e direta para melhorar o desempenho de um classificador que prediz probabilidades em um problema de classificação desequilibrada é ajustar o ponto de corte (movimento de ponto de corte), como foi sugerido por Brownlee (2020, Capítulo 6). O ponto de corte para classificação não irá interferir no valor da *AUC*, portanto, conforme sugerido por Zou *et al.* (2016), para decidir qual será o ponto de corte ótimo, neste trabalho usamos um ponto de corte com a melhor pontuação *KAPPA*.

3.3 Desempenho das métricas de classificação para dados desbalanceados

Para avaliar o desempenho das métricas para classificação binária em um conjunto de dados desbalanceado, descrito anteriormente na seção 3.2.1, desenhamos um estudo de simulação. Para gerar conjuntos de dados desbalanceados, consideramos o uso da distribuição com maior amplitude de assimetria como uma função de ligação, neste caso, usamos o a função de ligação PC com a seguinte estrutura

$$Y_i \sim \text{Bernoulli}\,(p_i)\,, \quad p_i = \left(\frac{1}{2} + \frac{1}{\pi}\arctan\left(\beta_1 + \beta_2 x_i\right)\right)^{\alpha}.$$

A covariável x foi gerada a partir de uma distribuição normal $N(0,\ 1)$. Os coeficientes de regressão foram fixados com os valores $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-0,5,\ 1,5)^\top$ e o parâmetro de assimetria foi fixado com os valores $\alpha = 3$ e $\alpha = 0,25$ para obter diferentes graus de desbalanceamento (proporção de uns), neste caso, $p = 0,15$ e $p = 0,76$ respectivamente. Com essas especificações, os dados foram gerados considerando 2 tamanhos amostrais $n = \{5000,\ 10000\}$. Para cada um dos seis cenários, foram consideradas 100 replicações e ajustados os modelos com as funções de ligação PC e logística. Para cada modelo estimado, em cada replicação, a matriz de confusão foi observada e então as métricas foram calculadas. Espera-se que as métricas sejam capazes de discriminar muito bem entre o modelo verdadeiro usando a função de ligação PC e o modelo usando a função de ligação logística (modelo mal especificado). Para a matriz de confusão, o valor do ponto de corte ótimo foi considerado com base no melhor valor da métrica *KAPPA*.

Para avaliar o desempenho das métricas, consideramos 3 critérios:

1. A distancia entre as curvas empíricas das métricas entre os modelos. Como o PC é o verdadeiro modelo, esperamos que a métrica mostre uma grande distância entre do modelo logístico.

2. Teste de Kolmogorov para identificar se as curvas são diferente. Aqui, esperamos rejeitar o teste de que a curva da métrica entre o modelo PC e logístico, são iguais.

3. Proporção de vezes que o valor da métrica do modelo PC é melhor. Aqui, esperamos que a métrica escolha o modelo verdadeiro em 100% das vezes.

A estatística de Kolmogorov-Smirnov Conover (1999) é definida por

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

em que $F_{1,n}$ e $F_{2,m}$ são duas funções de distribuição empírica da primeira e segunda amostras de tamanho m e n respectivamente, e sup é a função suprema.

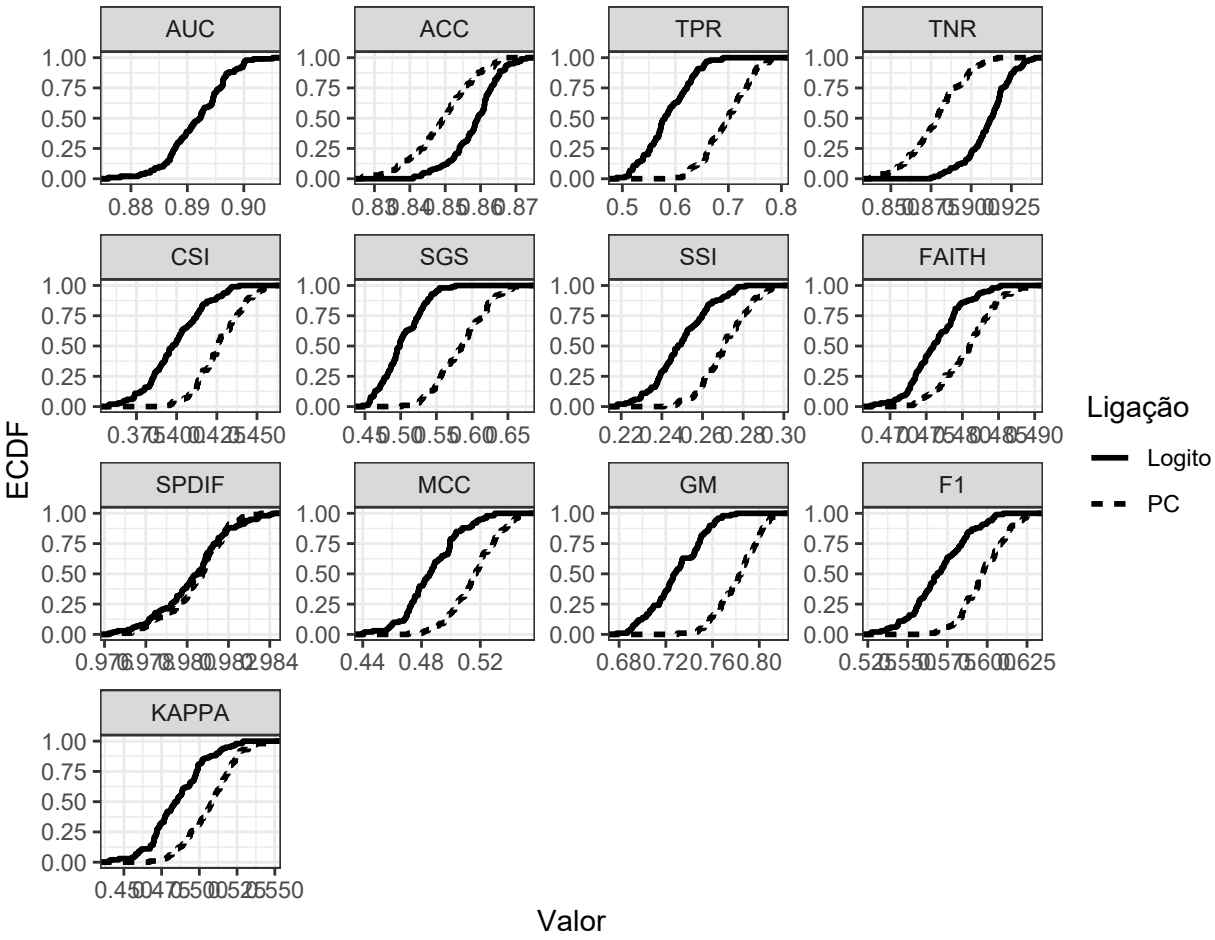

Figura 8 – Função de distribuição acumulada empírica das métricas em Tabela 6 para as funções de ligação logística e PC, em dados desbalanceados, para $\alpha = 3$ e $n = 5000$.

O teste de duas amostras de Kolmogorov verifica se as duas amostras de dados vêm da mesma distribuição. Para amostras grandes, a hipótese nula é rejeitada no nível α se

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n\times m}},$$

em que o valor de $c(\alpha)$, em geral, é dado por $c(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right)\times\frac{1}{2}}$. Para nosso estudo de simulação, o tamanho da amostra é $m = n = R = 100$, neste caso, as replicações.

Nas Figuras 8 e 9, é apresentada a função de distribuição acumulada empírica (ECDF) das diferentes métricas, considerando as funções de ligação PC e logística. Na Figura 8, quando o desbalanceamento é 15% ($\alpha = 3$), podemos ver que o ECDF de todas as métricas parece muito diferente para ambas as funções de ligação, com exceção das métricas *AUC* e *SPDIF*. Por outro lado, na Figura 9, quando o desbalanceamento é 76% ($\alpha = 0,25$) o ECDF de todas as métricas parecem muito diferente para ambas as funções de ligação, com exceção da métrica *AUC*. Estes resultados são confirmados através dos resultados do teste de Kolmogorov mostrados em Tabela 10, dos quais, podemos dizer a um nível de significância de 5%, que a distribuição das

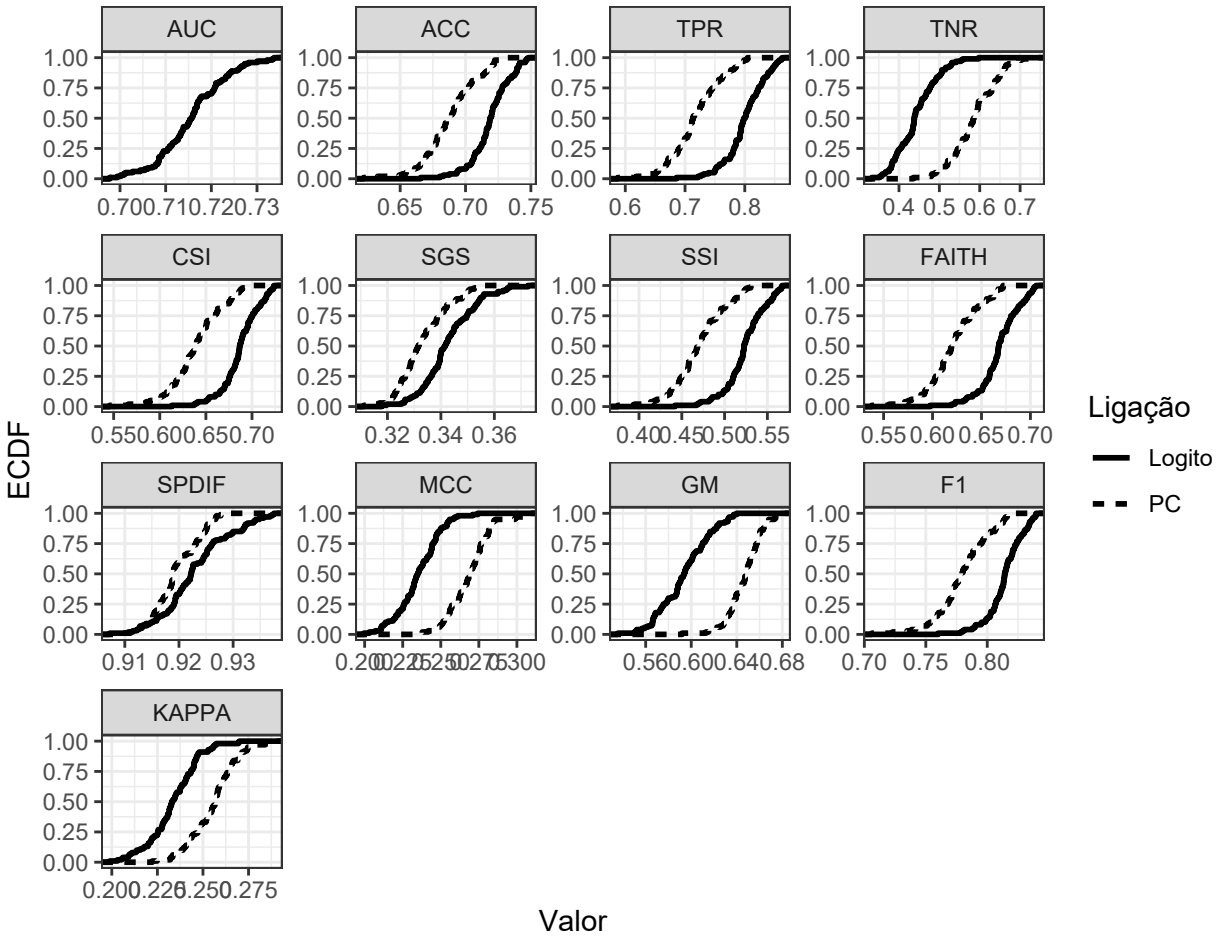

Figura 9 – Função de distribuição acumulada empírica das métricas em Tabela 6 para as funções de ligação logística e PC, em dados desbalanceados, para $\alpha = 0,25$ e $n = 5000$.

métricas *AUC* e *SPDIF* não diferem quando $\alpha = 3$ e a distribuição da métrica *AUC* não diferem quando $\alpha = 0,25$.

O teste estatístico Kolmogorov, usado aqui, é um teste bilateral, ele nos mostra se estatisticamente, as distribuições são diferentes; no entanto, não nos diz qual métrica é maior, apenas indica que são diferentes em sua distribuição.

A Tabela 11 mostra os resultados da proporção de vezes em que o modelo com função de ligação verdadeira foi escolhido para cada uma das métricas, ou seja, a proporção de vezes em que a métrica foi maior. Quando a proporção de zeros é muito maior, as métricas *ACC* e *TNR*, apesar de apresentarem diferenças entre suas distribuições, não fornecem discriminação correta. Por outro lado, quando a proporção de 1's é muito maior, as métricas que fornecem a melhor discriminação são *TNR*, *MCC*, *GM* e *KAPPA*.

A partir dos resultados apresentados, podemos concluir que as métricas *AUC*, *ACC*, *TNR* e *SPDIF*, embora amplamente utilizadas na literatura e aplicações, não são adequadas para avaliar o desempenho do modelo com função de ligação assimétrica quando a proporção de 0's é muito maior. Além disso, as métricas *AUC*, *ACC*, *TPR*, *CSI*, *SGS*, *SSI*, *FAITH*, *SPDIF* e *F*1

Tabela 10 – Teste de Kolmogorov (k) com valor p (p.val) entre as métricas do modelo com funções de ligação PC e logística para dados desbalanceados, usando *KAPPA* para definir o ponto de corte

Métrica	$\alpha = 3$				$\alpha = 0,25$			
	$n = 5000$		$n = 10000$		$n = 5000$		$n = 10000$	
	k	p.val	k	p.val	k	p.val	k	p.val
AUC	**0,000**	**1,000**	**0,000**	**1,000**	0,000	**1,000**	**0,000**	**1,000**
ACC	0,470	0,000	0,710	0,000	0,670	0,000	0,870	0,000
TPR	0,800	0,000	0,900	0,000	0,730	0,000	0,910	0,000
TNR	0,710	0,000	0,860	0,000	0,820	0,000	0,940	0,000
CSI	0,610	0,000	0,680	0,000	0,710	0,000	0,890	0,000
SGS	0,800	0,000	0,900	0,000	0,430	0,000	0,640	0,000
SSI	0,610	0,000	0,680	0,000	0,710	0,000	0,890	0,000
FAITH	0,490	0,000	0,510	0,000	0,720	0,000	0,890	0,000
SPDIF	**0,110**	**0,581**	**0,160**	**0,155**	0,300	0,000	0,470	0,000
MCC	0,620	0,000	0,680	0,000	0,800	0,000	0,830	0,000
GM	0,790	0,000	0,910	0,000	0,860	0,000	0,950	0,000
F1	0,610	0,000	0,680	0,000	0,710	0,000	0,890	0,000
KAPPA	0,510	0,000	0,550	0,000	0,640	0,000	0,690	0,000

Tabela 11 – Proporção de vezes que a métrica escolheu o modelo correto em dados desbalanceados, usando *KAPPA* para definir o ponto de corte

Métrica	$\alpha = 3$		$\alpha = 0,25$	
	$n = 5000$	$n = 10000$	$n = 5000$	$n = 10000$
AUC	**0%**	**0%**	**0%**	**0%**
ACC	**1%**	**0%**	**0%**	**0%**
TPR	100%	100%	**0%**	**0%**
TNR	**0%**	**0%**	100%	100%
CSI	100%	100%	**0%**	**0%**
SGS	100%	100%	**0%**	**0%**
SSI	100%	100%	**0%**	**0%**
FAITH	100%	100%	**0%**	**0%**
SPDIF	**54%**	**41%**	**26%**	**19%**
MCC	100%	100%	100%	100%
GM	100%	100%	100%	100%
F1	100%	100%	**0%**	**0%**
KAPPA	100%	100%	100%	100%

não são adequadas para avaliar o desempenho do modelo com função de ligação assimétrica a proporção de 1's é muito maior.

3.4 Aplicação

O conjunto de dados aqui analisados é referente à "Licitação cúmplice" (SB, do inglês: *Shill bidding*), que está disponível no repositório UCI (DUA; TANISKIDOU, 2017). A licitação cúmplice acontece quando um vendedor usa uma conta fraudulenta para licitar e aumenta artificialmente o preço da licitação (ALZAHRANI; SADAOUI, 2018). Os leiloeiros podem ser classificados como tendo um comportamento normal ou suspeito, essa classificação será considerada como variável resposta, ou seja, $Y = 0$ se o leiloeiro tiver comportamento normal e $Y = 1$, se o leiloeiro for suspeito.
Além da variável de resposta, o conjunto de dados contém outros atributos descritos em Tabela 12.

Para a classificação, usamos todos os atributos, exceto os três primeiros IDs, porque são considerados variáveis de fator com um alto número de valores únicos diferentes. A proporção de uns é 10.7% então o conjunto de dados está desbalanceado.

Numa primeira análise, ajustamos dois modelos, o primeiro modelo com a função de ligação logística habitual, e o segundo com a função de ligação PC; este último é um modelo adequado quando os dados são desbalanceados De la Cruz *et al.* (2019). A Figura 10 mostra os resultados do Intervalo de Maior densidade a posteriori (HPD, do inglês: *Highest Posterior Density*) com nível de credibilidade de 95%. Na figura 10, podemos ver que em ambos os modelos, os coeficientes associados às variáveis *Auction Bids*, *Bidding Ratio*, *Early Bidding*, *Last Bidding* e *Início Price Average*, contém o valor 0 no intervalo HPD, isso significa que essas variáveis podem não ser significativas para explicar a variável resposta, por isso foi retirada para análise futura.

Para uma segunda análise, o conjunto de dados foi reduzido removendo as variáveis que foram identificadas como não significativas. Este conjunto de dados reduzido foi dividido em dois subconjuntos de dados. O primeiro subconjunto, referido aqui como o conjunto de dados de treinamento, foi usado para ajustar os modelos. O segundo subconjunto, referido aqui como conjunto de dados de teste, foi usado para fazer previsões e avaliar o ajuste do modelo. Para o conjunto de dados de treinamento, 75% dos dados foram usados e para teste de modelo, 25%.

Cada modelo foi ajustado considerando os seguintes valores para *MCMC*: 4000 iterações foram usadas, descartando os primeiros 2000 como *burn-in*, com um intervalo de *thinning* igual a 2, foram consideradas 2 cadeias, resultando em um tamanho de amostra efetivo de 2000 amostras nas quais é baseada a inferência a posteriori. O tempo computacional para cada modelo foi aproximadamente de 10 minutos, usando um computador com as seguintes caraterísticas: processador Intel Core $i7$ da 11^a geração, com uma memória RAM de $16GB$ e memoria gráfica de $8GB$.

Tabela 12 – Caraterística do conjunto de dados SB

Atributo	Notação	Descrição
Record ID	-	Identificador único de um registro no dataset
Auction ID	-	Identificador único de um leilão
Bidder ID	-	Identificador único de um licitante
Bidder Tendency	X_1	Um licitante cúmplice participa exclusivamente de leilões de poucos vendedores, em vez de um lote diversificado. Este é um ato colusivo envolvendo o vendedor fraudulento e um
Bidding Ratio	X_2	Um licitante cúmplice participa com mais frequência para aumentar o preço do leilão e atrair lances mais altos de participantes legítimos
Successive Outbidding	X_3	Um licitante cúmplice supera sucessivamente a si mesmo, embora seja o vencedor atual, para aumentar o preço gradualmente com pequenos incrementos consecutivos
Last Bidding	X_4	Um licitante cúmplice fica inativo no último estágio do leilão (mais de 90% da duração do leilão) para evitar ganhar o leilão
Auction Bids	X_5	Leilões com atividades de licitante cúmplice tendem a ter um número muito maior de lances do que a média de lances em leilões simultâneos
Auction Starting Price	X_6	um licitante cúmplice geralmente oferece um pequeno preço inicial para atrair licitantes legítimos para o leilão
Early Bidding	X_7	Um licitante cúmplice tende a dar um lance bem no início do leilão (menos de 25% da duração do leilão) para chamar a atenção dos usuários do leilão
Winning Ratio	X_8	Um licitante cúmplice compete em muitos leilões, mas dificilmente ganha algum leilão
Auction Duration	X_9	Quanto tempo durou um leilão

A matriz de confusão foi obtida com um valor mais ótimo de ponto de corte baseado no valor de *KAPPA*, conforme mencionado acima, em cada modelo.

A Tabela 13 mostra os resultados das métricas para diferentes modelos, essas métricas foram calculadas para o conjunto de dados de teste em todos os modelos. Nesta tabela observa-se que o modelo com a função de ligação *PC*, apresenta maior desempenho com base nos valores de TPR, CSI, SGS, SSI, $FAITH$, MCC, GM, F_1 e $KAPPA$, portanto consideramos que é um modelo adequado para este conjunto de dados.

Os valores estimados para o modelo com a função de ligação PC, são mostrados em Tabela 14. Observe que todos os parâmetros do modelo são indicados como significativamente diferentes de 0 com um nível de credibilidade de 95%. Além disso, todas as variáveis influenciam positivamente o comportamento dos leiloeiros, ou seja, as chances de o comportamento do leiloeiro ser normal aumentam se os valores das diferentes variáveis forem aumentados.

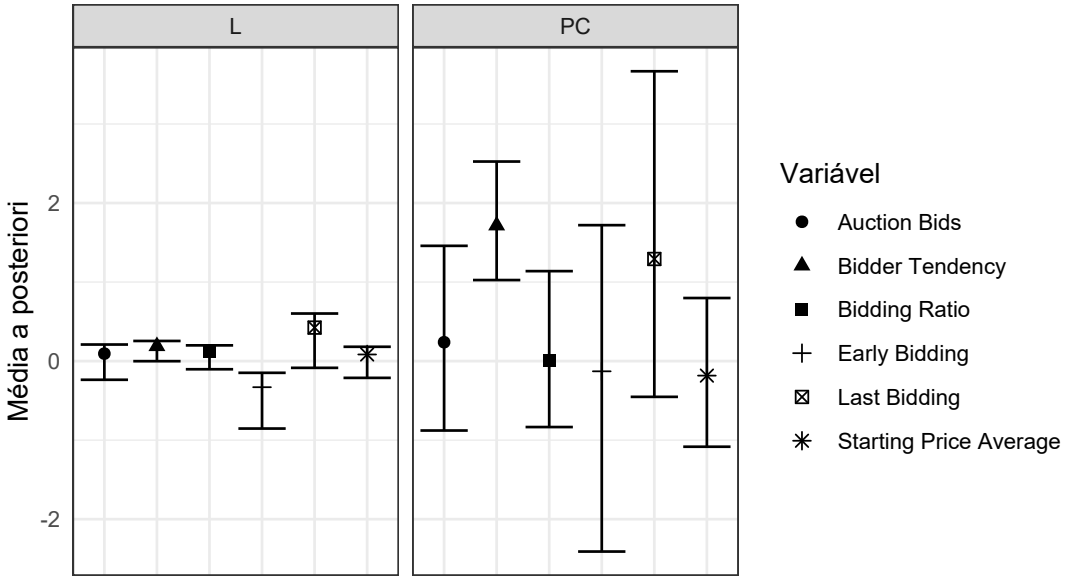

Figura 10 – Maior densidade a posteriori para o modelo completo

Tabela 13 – Métricas das funções de ligação assimétricas para *Shill Bidding* com conjunto de dados de teste

	Função de ligação					
Métrica	L	PL	PN	PC	PRG	PLA
TPR	0,972	0,945	0,740	**1,000**	0,978	0,956
CSI	0,838	0,818	0,713	**0,858**	0,843	0,824
SGS	0,954	0,911	0,667	**1,000**	0,963	0,928
SSI	0,721	0,692	0,554	**0,751**	0,728	0,700
FAITH	0,545	0,542	0,525	**0,548**	0,546	0,543
MCC	0,902	0,888	0,821	**0,916**	0,905	0,892
GM	0,976	0,962	0,858	**0,989**	0,979	0,967
F1	0,912	0,900	0,832	**0,923**	0,915	0,903
KAPPA	0,900	0,886	0,814	**0,913**	0,903	0,890
Ponto de corte	0,285	0,446	0,358	0,125	0,256	0,453

Tabela 14 – Estimativa de parâmetro a posteriori para o modelo de resposta binária com uma função de ligação PC para dados da *Shill Bidding*

Variável	Parâmetro	Estimativa	Desvio padrão	95% I.C.	
Intercept	β_1	-13,925	3,467	(-20,994;	-7,631)
Bidder Tendency	β_2	2,045	0,597	(0,931;	3,287)
successive Outbidding	β_3	9,891	2,264	(5,366;	14,060)
Winning Ratio	β_4	5,303	1,209	(3,017;	7,707)
Auction Duration	β_5	1,108	0,414	(0,379;	1,993)
Shape parameter	λ	5,9306	1,1467	(3,077;	7,339)

Por outro lado, observe que o intervalo de credibilidade para o parâmetro de forma α, não inclui o valor 1 e este é superior a 1, indicando que este parâmetro explica o desbalanceamento nos dados.

3.5 Comentários finais

Neste seção do trabalho, mostramos que a assimetria e a curtose das distribuições potência Gumbel e reversa de potência Gumbel não dependem do parâmetro de assimetria α. Consequentemente, as funções de ligação potência Gumbel (PG) e sua correspondente reversa de potência Gumbel (RPG) também não dependem de um parâmetro de assimetria. Eles são funções de ligação inviáveis na regressão binária, pois a curva não muda conforme o esperado. Os resultados deste trabalho podem ser usados para avaliar outras funções de ligação recentemente apresentados em Li *et al.* (2015) =, Li *et al.* (2016) e Lemonte e Bazán (2018).

Por outro lado, os resultados do nosso estudo de simulação mostraram que algumas métricas comumente usadas para classificação binária (AUC, ACC e TNR) podem não ser as mais adequadas para escolher o melhor modelo quando os dados estão desbalanceados. Além disso, na aplicação, foi mostrado que de acordo com as métricas apropriadas para dados desbalanceados, um modelo com função de ligação potência, apresentou melhor desempenho para descrever o conjunto de dados *Shill Bidding*.

Por fim, as propriedades estudadas nesta seção mostram que as funções de ligação potência, são ótimas alternativas para problemas de classificação binária, fornecendo resultados muito mais confiáveis em casos de dados desbalanceados.

CAPÍTULO

4

MODELOS DE RESPOSTA BINÁRIA LONGITUDINAL USANDO FUNÇÕES DE LIGAÇÃO ALTERNATIVAS

Neste capítulo, apresentamos uma ampla classe de funções de ligação, chamadas de potência e reversa de potência, como uma alternativa para analisar a classificação binária longitudinal, principalmente quando ela é desbalanceada, o que é muito comum em muitas aplicações do mundo real. Um estudo longitudinal é uma metodologia que avalia o comportamento de uma variável resposta ao longo do tempo, ou seja, os indivíduos são acompanhados ao longo do tempo e a variável de interesse é medida repetidamente. Na classificação binária longitudinal, a função de ligação logit, conhecida como regressão logística, é comumente utilizada, porém não é a mais adequada no contexto de dados não balanceados.

Estimativas Bayesianas usando um procedimento MCMC através do algoritmo *No-U-Turn Sampler* são propostas. Verificações preditivas a posteriori, resíduo quantílico aleatorizado bayesiano e uma medida de influência bayesiana são considerados para o diagnóstico do modelo. Diferentes modelos são comparados usando critérios de modelo de seleção.

Um estudo de simulação é desenvolvido para analisar a sensibilidade a priori para o parâmetro do efeito aleatório e avaliar o desempenho do modelo proposto na presença de dados desbalanceados.
Finalmente, é considerada uma aplicação da metodologia estudada em um conjunto de dados médicos sobre a presença do sintoma esquizofrênico "distúrbio do pensamento". Neste conjunto de dados, a presença de sintomas é muito menor do que a ausência, assim mostramos, na prática, a utilidade do uso de funções de ligação alternativas em dados desbalanceados.

4.1 Modelo Binário Longitudinal com funções de ligação potência e reversa de potência

No Capítulo 2, foi mostrada a construção de uma distribuição potência e sua reversa, baseada na consideração de uma função de distribuição acumulada (fda) contínua arbitrária e elevada por potência real positiva arbitrária. Assim, uma nova função de distribuição acumulada $F_l(\cdot)$ ($l = P,\ l = RP$) é proposta com um parâmetro de potência adicional α. Portanto, temos uma novas classe de função de ligação $F_l^{-1}(\cdot)$. O parâmetro adicional α, caracteriza a assimetria das funções de ligação associado ao modelo. Assim, essa nova classe de função de ligação pode ser usada para a construção do modelo de variável resposta binária longitudinal.

Seja $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ o vetor de resposta com Y_i assumindo valores 1 ou 0, seja $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ o vetor de covariáveis e seja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor de coeficientes de regressão. O modelo de regressão binária é definido como

$$\begin{aligned} Y_i &\sim \text{Bernoulli}\,(u_i) \text{ independente} \quad \forall\, i = 1, \ldots, n \\ u_i &= \mathbb{P}(Y_i = 1) = F\,(\eta_i) \\ \eta_i &= \boldsymbol{x_i^T}\boldsymbol{\beta} = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \end{aligned} \tag{4.1}$$

Na análise longitudinal dos dados, existe uma dependência entre medidas repetidas no mesmo indivíduo e a possibilidade de explicar essa dependência se dá por meio da estrutura de correlação. Existem diferentes métodos para modelar a estrutura de correlação. Neste trabalho, o modelo hierárquico misto é considerado. Neste método, os efeitos aleatórios são adicionados ao preditor linear considerando duas fontes de variação nos dados: a variação entre unidades e a variação dentro das unidades.

Seja $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ um vetor de uma sequência de medições observadas ao longo do tempo para o i-ésimo sujeito, $i = 1, \cdots, n$, em que cada componente y_{ij}, que assume valor 0 ou 1, corresponde à observação do sujeito i, medida no tempo t_j, $j = 1, ..., n_i$. Além disso, considere $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^\top$ um vetor $q \times 1$ dos efeitos aleatórios específicos do indivíduo i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ vetor de coeficientes de regressão (efeitos fixos) para $j = 1, \ldots, n_i$ e $i = 1, \ldots, n$. O modelo de regressão binária longitudinal pode ser escrito da seguinte forma

$$\begin{aligned} Y_{ij} \mid \boldsymbol{b}_i &\sim \text{Bernoulli}(\mu_{ij}) \\ \mu_{ij} = F(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i) \quad &\text{ou} \quad \eta_{ij} = F^{-1}(\mu_{ij}) \\ \boldsymbol{b}_i &\sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma}_b), \end{aligned} \tag{4.2}$$

em que $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ e $\boldsymbol{Z}_{ij} = (z_{ij1}, \ldots, z_{ijq})^\top$ são vetores covariáveis, $\boldsymbol{\Sigma}_b$ é uma matriz definida positiva $q \times q$ e η_{ij} é nomeado como o i-ésimo preditor linear pela j-ésima vez, para $i = 1, \cdots, n$ e $j = 1, \cdots, n_i$. Além disso, os efeitos aleatórios são considerados independentes das covariáveis $\boldsymbol{X}_{ij}$ Fitzmaurice, Laird e Ware (2012, Capítulo 14).

Para dados de resposta binária longitudinal, considerando as distribuições potência e reversa de potência, temos duas novas classes de funções de ligação assimétricas

$$\mu_{ij} = F_l(\boldsymbol{X}_{ij}^\top\beta + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i) \quad \text{ou} \quad \eta_{ij} = F_l^{-1}(\mu_{ij}). \tag{4.3}$$

Quando $F_l(\cdot)$ é uma fda de uma distribuição potência logística padrão ou de sua reversa padrão e $\alpha = 1$, temos a função de ligação logística simétrica como um caso especial, ou seja,

$$\mu_{ij} = F(\boldsymbol{X}_{ij}^\top\beta + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i) = \frac{\exp\left(\boldsymbol{X}_{ij}^\top\beta + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{X}_{ij}^\top\beta + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i\right)}.$$

Na Tabela 15 mostramos algumas funções de ligação potência e reversa de potência que foram consideradas neste capítulo.

Tabela 15 – Modelos com uma função de ligação assimétrica potência e reversa de potência

Função de ligação	μ_{ij}	η_{ij}
PN	$\left[\Phi\left(\eta_{ij}\right)\right]^\alpha$	$\eta_{ij} = \Phi^{-1}\left(\mu_{ij}^{1/\alpha}\right)$
RPN	$1-\left[\Phi\left(-\eta_{ij}\right)\right]^\alpha$	$\eta_{ij} = -\Phi^{-1}\left(\left(1-\mu_{ij}\right)^{1/\alpha}\right)$
PL	$\left[\frac{1}{1+\exp\left(-\eta_{ij}\right)}\right]^\alpha$	$\eta_{ij} = -\log\left(\mu_{ij}^{-1/\alpha}-1\right)$
RPL	$1-\left[\frac{1}{1+\exp\left(\eta_{ij}\right)}\right]^\alpha$	$\eta_{ij} = \log\left(\left(1-\mu_{ij}\right)^{-1/\alpha}-1\right)$
PC	$\left[\frac{1}{\pi}\arctan\left(\eta_{ij}\right)+\frac{1}{2}\right]^\alpha$	$\eta_{ij} = \tan\left[\left(\mu_{ij}^{1/\alpha}-\frac{1}{2}\right)\pi\right]$
RPC	$1-\left[\frac{1}{\pi}\arctan\left(-\eta_{ij}\right)+\frac{1}{2}\right]^\alpha$	$\eta_{ij} = -\tan\left[\left(\left(1-\mu_{ij}\right)^{1/\alpha}-\frac{1}{2}\right)\pi\right]$

Existem algumas propostas de modelos de regressão binária mista com funções de ligação assimétricas, por exemplo, Abanto-Valle, Dey e Jiang (2015) e Komori *et al.* (2016); porém, até onde sabemos, não há nenhuma proposta de uso dos funções de ligação que estudamos aqui para dados binários longitudinais. A classe de funções de ligações que consideramos, tem sido utilizada em análises de regressão com relativo sucesso em Bazán, Romeo e Rodrigues (2014), Bazán *et al.* (2017), De la Cruz *et al.* (2019) e Lemonte e Moreno-Arenas (2020), mas não foi utilizada para modelos mistos.
Como é difícil determinar antecipadamente qual é a função de ligação que será adequada para os dados, propomos a utilização das funções de ligação aqui apresentados, como uma alternativa que deve ser ajustada e depois pode ser selecionada se realmente for o modelo mais adequado para os dados.

Na seguinte seção, apresentamos uma análise bayesiana do modelo com funções de ligação potência e reversa de potência em regressão binária longitudinal.

4.2 Análise bayesiana

4.2.1 Estimação dos parâmetros

Primeiro assumimos que para $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$, condicionalmente em $\boldsymbol{b}_i$, $\boldsymbol{\beta}$ e α, os $Y'_{ij}s$ são independentes, portanto, o modelo Bayesiano misto de regressão binária para dados longitudinais, com uma função de ligação potência ou reversa de potência, pode ser escrito como

$$\begin{aligned} Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \alpha &\overset{ind.}{\sim} \text{Bernoulli}(\mu_{ij}), \quad \mathbf{b}_i \overset{ind.}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma}_b) \\ \mu_{ij} &= F_\alpha(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\top \boldsymbol{b}_i) \\ (\boldsymbol{\beta}, \alpha, \boldsymbol{\Sigma}_b)^\top &\sim \pi(\boldsymbol{\beta}, \alpha, \boldsymbol{\Sigma}_b) \end{aligned} \tag{4.4}$$

Os parâmetros $\boldsymbol{\beta}, \alpha, \boldsymbol{\Sigma}_b$ são considerado independente tal que $\pi(\boldsymbol{\beta}, \alpha, \boldsymbol{\Sigma}_b) = \pi(\boldsymbol{\beta})\,\pi(\alpha)\,\pi(\boldsymbol{\Sigma}_b)$, com $\boldsymbol{\beta} \sim N_p(\mathbf{0},\ \boldsymbol{I}\sigma_\beta^2)$, $\log(\alpha) = \delta \sim U(-2,\ 2)$. No modelo, consideramos a priori para δ seguindo os trabalhos de Bazán *et al.* (2017) e De la Cruz *et al.* (2019). Observe que uma a priori uniforme limitado é fornecido para um parâmetro ilimitado, dando mais probabilidade para um lado da assimetria na ligação assimétrica induzida. De fato, é fácil provar que α segue a distribuição recíproca também conhecida como distribuição log-uniforme, conforme indicado por Hamming (1970), com $\mathbb{E}(\alpha) = 1,813$ e $\mathbb{V}(\alpha) = 3,533$, o qual também é limitado (Veja Seção B.3). Consideramos, seguindo Gelman (2006), que a densidade a priori uniforme não informativa para o parâmetro de escala pode ser considerada inicialmente e que a família inverse-gamma(c, c) de distribuições a priori não informativas não é recomendada porque em casos onde este parâmetro é estimado em aproximadamente 0, as inferências resultantes serão sensíveis a c.

Reconhecemos que estudos específicos adicionais podem ser considerados para propor outros a prioris como, por exemplo, um Gama ou apenas outra distribuição positiva. Nesse sentido, recentemente, Ordoñez *et al.* (2023) mostrou que essa a priori recupera muito bem os parâmetros do modelo potência logística e tem uma probabilidade de cobertura comparável a uma a priori de complexidade Penalizada, mas pode apresentar um intervalo de credibilidade maior que essa a priori, principalmente para pequenas amostras. Os resultados da Seção de Simulação de nosso trabalho mostram que a priori adotada funcionou relativamente bem.

Para o parâmetro de efeito aleatório, é possível utilizar uma das várias especificações a priori presentes na literatura, por exemplo, uma distribuição inversa de Wishart como em Fong, Rue e Wakefield (2010), ou seja, $\Sigma_b \sim IW_q(\psi,\ c)$, ou, para um determinado componente de Σ_b, σ_b^2, pode ser considerado $\sigma_b^2 \sim$ Inverse-gamma$(0{,}001,\ 0{,}001)$ (LUNN *et al.*, 2000), $\frac{1}{\sigma_b^2} \sim$ Pareto$(1,\ 0{,}001)$ (SCURRAH; PALMER; BURTON, 2000) ou $\log(\sigma_b^2) \sim$ Uniforme$(-10,\ 10)$ (SPIEGELHALTER, 2001).

A função de verossimilhança associada ao modelo é dada por

$$L(\boldsymbol{\beta}, \alpha, \boldsymbol{\Sigma}_b, \boldsymbol{b} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[F_l\left(\eta_{ij}\right)\right]^{y_{ij}} \left[1 - F_l\left(\eta_{ij}\right)\right]^{1-y_{ij}} \phi_q\left(\boldsymbol{b}_i \mid \mathbf{0}, \boldsymbol{\Sigma}_b\right),$$

em que $\boldsymbol{b} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_n^\top)^\top$ e $\phi_q(\cdot \mid \mathbf{m}, \mathbf{S})$ denotam a fdp da distribuição normal q-variada com vetor de media $\mathbf{m}$ e matriz de covariância $\mathbf{S}$ e $\boldsymbol{y}$ denota a matriz de dados observados da variável de resposta.

Assim, a distribuição a posteriori de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \mathbf{b}, \boldsymbol{\Sigma}_b)^\top$, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$, é da seguinte forma

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[F_l\left(\eta_{ij}\right) \right]^{y_{ij}} \left[1 - F_l\left(\eta_{ij}\right) \right]^{1-y_{ij}} \phi_q\left(\boldsymbol{b}_i \mid \mathbf{0}, \boldsymbol{\Sigma}_b\right) \right) \pi\left(\boldsymbol{\beta}\right) \pi\left(\alpha\right) \pi\left(\boldsymbol{\Sigma}_b\right).$$

No entanto, esta distribuição a posteriori não pertence a uma família de distribuição conhecida e não tem uma forma fechada, o que torna esta distribuição analiticamente intratável.

Vários autores mostram as limitações dos algoritmos MCMC tradicionais como Passeio aleatório Metrópoles (em inglês *Random walk Metropolis*) e amostragem de Gibbs, ao estimar modelos Bayesianos hierárquicos ou modelos mistos, indicando que tais métodos podem levar muito tempo para convergir para a distribuição alvo (HOFFMAN; GELMAN *et al.*, 2014; KARIMI; LAVIELLE, 2019), muitas vezes porque esses métodos exploram o espaço de parâmetros por meio de caminhadas aleatórias ineficientes.

Por outro lado, Hoffman, Gelman *et al.* (2014) menciona que o algoritmo Hamiltoniano de Monte Carlo (HMC) apresenta as seguintes vantagens quando comparado aos algoritmos MCMC tradicionais. Essas vantagens incluem evitar o comportamento de caminhada aleatória e gerar amostras de um grande espaço de parâmetros com um alto nível de probabilidade de aceitação, o que faz convergir para as distribuições de destino muito mais rapidamente.

No entanto, às vezes o HMC pode apresentar alguns problemas devido ao fato de que precisamos especificar o tamanho e o número de passos (uma má escolha tornará o método ineficiente) e porque requer o cálculo do gradiente do log-a posteriori e é notavelmente afetado pelos dois hiper-parâmetros.

Assim, os autores Hoffman, Gelman *et al.* (2014), propõem o uso do algoritmo *No-U-Turn Sampler* (NUTS) como uma versão melhorada do HMC, mostrando que o NUTS tem um desempenho igual ou mais eficiente que um método HMC. Além disso, em um modelo linear misto, Nishio e Arakawa (2019) mostrou que o NUTS apresentou um desempenho melhor do que os métodos de amostragem HMC e Gibbs.

Portanto, consideramos o método NUTS para simular a distribuição alvo de nossos modelos propostos.

Neste trabalho, para estimar os parâmetros dos modelos, desenvolvemos um código próprio usando a linguagem `Stan` (TEAM *et al.*, 2016) e o pacote `Pystan` através de `Python` (ROSSUM; DRAKE, 2020). O código da aplicação é mostrado em Seção B.2

4.2.2 Critérios de comparação de modelos

Nesta seção, usamos os critérios de comparação de modelos apresentados Capítulo 2, adaptado para o caso misto.

Na abordagem bayesiana, as comparações de modelos podem ser realizadas usando diferentes critérios. Inicialmente consideramos o critério baseado no desvio bayesiano, definido por $D(\boldsymbol{\theta}) = -2\log p(\boldsymbol{y} \mid \boldsymbol{\theta})$, cuja média a posteriori e desvio da média a posteriori são aproximados respectivamente por $\bar{D} = \frac{1}{M}\sum_{m=1}^{M} D\left(\boldsymbol{\theta}^{(m)}\right)$ e $\widehat{D} = D\left(\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}^{(m)}\right)$, em que o índice m representa a m-ésima realização e M é o tamanho de amostra válido da distribuição a posteriori. Por exemplo, (a) $EAIC = \bar{D} + 2\nu$, em que ν é o número de parâmetros do modelo para o qual especificou uma distribuição a priori, ou seja, o número de parâmetros irrestritos (TEAM *et al.*, 2016). Por exemplo, para o caso do modelo PL com parâmetro $\beta_0, \beta_1, \beta_2, \alpha$ e σ_b^2, em que o último parâmetro está associado ao efeito aleatório, os primeiros parâmetros estão associados a efeitos fixos e α está associado à função de ligação, temos que $\nu = 5$. (b) $DIC = \bar{D} + \widehat{\rho}_D = 2\bar{D} - \widehat{D}$, em que $\widehat{\rho}_D = \bar{D} - \widehat{D}$ é o número efetivo de parâmetros (em inglês *effective number of parameters*) compensa esse efeito favorecendo modelos com um número menor de parâmetros (SPIEGELHALTER *et al.*, 2002). (c) $EBIC = \bar{D} + \nu\log(N)$ em que $N = \sum_{i=1}^{n} n_i$ é o número de observações e (d) Critério de informação preditiva bayesiana (IC, do inglês: *Bayesian predictive information criterion*), definido como $IC = \bar{D} + 2\widehat{\rho}_D$ (ANDO, 2011).

Outro critério importante usado para a seleção do modelo é o *WAIC*, proposto por Watanabe (2010). É baseado no parâmetro de complexidade p_{WAIC}, que pode ser aproximado por $\sum_{i=1}^{n}\sum_{j=1}^{n_i} \text{var}\left(\log p\left(y_{ij} \mid \boldsymbol{\theta}\right)\right)$. Assim, o *WAIC* pode ser calculado por $\text{WAIC} = \bar{D} + 2p_{\text{WAIC}}$. Além disso, consideramos o critério proposto por Geisser e Eddy (1979), chamado *LOO*. Devido à sua natureza iterativa, o *LOO* pode ser computacionalmente proibitivo para grandes conjuntos de dados de amostra em que o modelo precisa ser ajustado para n (tamanho da amostra), Vehtari, Gelman e Gabry (2017) propõe usar amostragem de importância suavizada de Pareto (*PSIS*), uma nova abordagem que fornece uma estimativa precisa e confiável de *LOO* usando pesos de importância instável. A estimativa Bayesiana de $PSIS - LOO$ é dada por

$$\widehat{\text{elpd}}_{\text{psisloo}} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \log\left(\frac{\sum_{m=1}^{M} w_{ij}^{(m)} p\left(y_{ij} \mid \boldsymbol{\theta}^{(m)}\right)}{\sum_{m=1}^{M} w_{ij}^{(m)}}\right),$$

em que $w_{ij}^{(m)} = \min(r_{ij}^{(m)}, \frac{\sqrt{M}}{M}\sum_{m=1}^{M} r^{(m)})$ e $r_{ij}^{(m)} = \frac{1}{p\left(y_{ij}|\boldsymbol{\theta}^{(m)}\right)} \propto \frac{p\left(\boldsymbol{\theta}^{(m)}|\boldsymbol{y}_{-ij}\right)}{p\left(\boldsymbol{\theta}^{(m)}|\boldsymbol{y}\right)}$.

Observe que ρ_D, p_{WAIC} e ν são propostos para aproximar o número de parâmetros no modelo, mas não são comparáveis.

Dado um conjunto de modelos candidatos aos dados, o modelo preferido é aquele com o valor mínimo do critério considerado.

4.2.3 Verificações preditivas a posteriores

Usamos a notação $D = (n, \boldsymbol{y}, \mathbf{x})$ para os dados observados e $\boldsymbol{y}^{\text{rep}}$ para conjuntos de dados replicados extraídos da distribuição preditiva a posteriori. As verificações preditivas a posteriori (GELMAN *et al.*, 2000; GELMAN *et al.*, 2013) são úteis para verificar se o modelo ajustado é consistente com os dados observados. Se o modelo se ajustar, os dados replicados, $\boldsymbol{y}^{\text{rep}}$, gerados no modelo devem ser semelhantes aos dados observados. Para medir a discrepância entre o modelo ajustado e os dados observados, definimos uma variável de discrepância, $T(\boldsymbol{y}, \boldsymbol{\theta})$, uma função dos dados e parâmetros do modelo. Seguindo Gelman *et al.* (2000), consideramos a média e o desvio padrão da resposta binária como variáveis de discrepância.

Seja $\tilde{\mathbf{y}}$ uma observação futura da variável de resposta $\boldsymbol{y}$. Assumindo que $\boldsymbol{\theta}$, $\tilde{\mathbf{y}}$ e $\boldsymbol{y}$ são independentes, a distribuição preditiva a posteriori é definida como

$$\pi(\tilde{\mathbf{y}} \mid D) = \int_{\Theta} \pi(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}. \tag{4.5}$$

Assim, dada uma amostra MCMC de tamanho L da distribuição a posteriori $\pi(\boldsymbol{\theta} \mid D)$, $T(\boldsymbol{y}, \boldsymbol{\theta})$ pode ser calculado usando o seguinte procedimento:

- para cada$\boldsymbol{\theta}_{(l)}$ $l = 1, \ldots, L$, amostra $\boldsymbol{y}^{\text{rep}}_{(l)}$ de $\pi\left(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}_{(l)}\right)$.
- obtenha $T(\boldsymbol{y}, \boldsymbol{\theta})$, a variável de discrepância para os dados observados e $T(\boldsymbol{y}^{\text{rep}}_{(l)}, \boldsymbol{\theta}_{(l)})$, a variável de discrepância para cada $\boldsymbol{y}^{\text{rep}}_{(l)}$, para $l = 1, \ldots, L$.

Podemos mostrar as diferenças $T(\boldsymbol{y}, \boldsymbol{\theta})$ - $T(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\theta})$ através de um histograma, ou podemos considerar um gráfico de dispersão de $T(\boldsymbol{y}, \boldsymbol{\theta})$ contra $T(\boldsymbol{y}^{\text{rep}}, \boldsymbol{teta})$. Além disso, dada uma amostra MCMC da distribuição a posteriori e dos conjuntos de dados replicados, o valor p preditivo posterior, definido como $p_B = Pr(T(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta}) \mid \boldsymbol{y})$, pode ser estimado por

$$\widehat{p}_B = \frac{1}{L} \sum_{l=1}^{L} I\left\{ T(\boldsymbol{y}^{\text{rep}(l)}, \boldsymbol{\theta}^{(l)}) \geq T(\boldsymbol{y}, \boldsymbol{\theta}^{(l)}) \right\} \tag{4.6}$$

$\widehat{p}_B$ próximo de zero ou um, sugere que as variáveis de discrepância para o conjunto de dados observados são extremas e, portanto, o modelo pode não ser apropriado. Conselhos sobre como interpretar p-valores preditivos a posteriori podem ser encontrados em Gelman *et al.* (2013).

4.2.4 Predição

Considere uma amostra MCMC de tamanho L da distribuição a posteriori $\pi(\boldsymbol{\theta} \mid D)$ denotada por $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ e a distribuição preditiva a posteriori (Equação 4.5). Então, seguindo Paraíba, Bochkina e Diniz (2018), o valor previsto da unidade i para o momento j, denotado por $\tilde{y}_{ij}$, pode ser obtido usando o seguinte procedimento:

- para cada $\boldsymbol{\theta}_l$, $l = 1,\ldots,L$, amostra $\tilde{y}_{ij(l)}$ de $\pi\left(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}_{(l)}\right)$, $i = 1,\ldots,n$, $j = 1,\ldots,n_i$, em que $\pi\left(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}_{(l)}\right)$ é uma distribuição preditiva a posteriori.
- para $i = 1,\ldots,n$, $j = 1,\ldots,n_i$, defina $\tilde{y}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\tilde{y}_{ij(l)}$ o valor previsto da (ij)-ésima observação.

4.2.5 Resíduos quantílicos aleatórios

Em um contexto de regressão binária, considere comparar a observação binária observada y_{ij} com a probabilidade μ_{ij}. Como μ_{ij} não é observado, a abordagem clássica é comparar y_{ij} com a probabilidade estimada $\widehat{\mu}_{ij}$. No entanto, a diferença $y_{ij} - \widehat{\mu}_{ij}$ é difícil de interpretar, pois a distribuição de referência está sobre a distribuição amostral de $y_{ij} - \widehat{\mu}_{ij}$ que não é conhecido devido à variável de resposta binária. Por outro lado, na abordagem bayesiana, se a distribuição a posteriori de μ_{ij} e o valor de y_{ij} estiverem em conflito, então a distribuição a posteriori de $r_{ij} = y_{ij} - \mu_{ij}$ será concentrado para o valor extremo. Como o suporte da distribuição a posteriori de r_{ij} está no intervalo $(y_{ij} - 1,\, y_{ij})$, uma observação $y_{ij} = 0$ será periférica se a posteriori de r_{ij} está concentrado em direção ao ponto final -1, e uma observação $y_{ij} = 1$ é incomum se a posteriori de r_{ij} estiver concentrado em direção ao valor 1 (mais detalhes em Albert e Chib (1995)).

Uma abordagem alternativa para definir um resíduo é o resíduo quantil aleatório proposto por Dunn e Smyth (1996). Sob uma abordagem bayesiana, definimos esse resíduo da seguinte forma:

Seja $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L$ uma amostra de tamanho L de $\pi(\boldsymbol{\theta} \mid D)$ após o *burn-in*. Para o nosso modelo, definimos

$$\tilde{r}_{q,ij} = \frac{1}{L}\sum_{l=1}^{L}\Phi^{-1}\left(u_{ij}^{(l)}\right) \qquad i = 1,\ldots,n, \quad j = 1,\ldots,n_i \tag{4.7}$$

em que $u_{ij}^{(l)}$ é um valor aleatório gerado a partir de uma distribuição uniforme no intervalo $[I_{1-\widehat{\mu}_{ij}^{(l)}}(2-y_i, y_i), I_{1-\widehat{\mu}_{ij}^{(l)}}(1-y_{ij}, y_{ij}+1)]$, em que $I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$ é uma função beta incompleta regularizada (LEMONTE; BAZÁN, 2018) e $\widehat{\mu}_{ij}^{(l)} = F_{\widehat{\alpha}^{(l)}}(\boldsymbol{X}_{ij}^{\top}\widehat{\boldsymbol{\beta}}^{(l)} + \boldsymbol{Z}_{ij}^{\top}\widehat{\boldsymbol{b}}_i^{(l)})$ é o $l$ésimo elemento da probabilidade marginal posterior obtida considerando a função de ligação adotada. As funções beta incompleta e beta completa, são definidas, respectivamente, por $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ e $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$.

Resultado 5. Seja $X \sim Beta(a,b)$, temos que a probabilidade acumulada de X, pode ser escrita como $F_X(x) = \mathbb{P}(X \leq x) = \int_0^x \frac{1}{B(a,b)}t^{a-1}(1-t)^{b-1}dt$. Logo, $B(a,b)F_X(x) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ o qual é equivalente a $B(x;a,b)$. Portanto,

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)} = \frac{B(a,b)F_X(x)}{B(a,b)} = F_X(x) = \mathbb{P}(X \leq x).$$

A partir do resultado 5, podemos ver que, calcular uma função beta incompleta regularizada, é equivalente a calcular uma probabilidade acumulada de uma distribuição $Beta(a,b)$.

Uma vez que os resíduos quantis aleatórios têm um componente aleatório, que é obtido a partir de uma distribuição uniforme, para o caso discreto seguindo Dunn e Smyth (1996), 4 replicações são recomendadas. Considerando nossa abordagem bayesiana, obtivemos esse resíduo como a média dos resíduos das L amostras MCMC. Se o modelo se ajustar corretamente, o resíduo quantil aleatório terá uma distribuição normal padrão, que pode ser avaliada por meio de um gráfico de envelope (De la Cruz *et al.*, 2019).

4.2.6 Análise de Influência

A ordenada preditiva condicional (CPO, do inglês: *conditional predictive ordinate*), proposta por Gelfand, Dey e Chang (1992), é usada para comparar o ajuste do modelo e a complexidade dos modelos considerados. É a densidade preditiva do (ij)-ésimo caso dados os dados com o (ij)-ésimo caso excluído, $D_{(-ij)}$. Para $i = 1,\ldots,M$, $j = 1,\ldots,n_i$, esta métrica é definida como $\text{CPO}_{ij} = \left[\int_\Theta \frac{1}{f(y_{ij}|\boldsymbol{\theta})}\pi(\boldsymbol{\theta} \mid D)\,d\boldsymbol{\theta}\right]^{-1}$.

Dada uma amostra MCMC de tamanho L, $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L$, de $\pi(\boldsymbol{\theta} \mid D)$, uma aproximação de Monte Carlo de CPO_{ij} pode ser escrito como

$$\widehat{CPO}_{ij} = \left\{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f\left(y_{ij} \mid \boldsymbol{\theta}^{(l)}\right)}\right\}^{-1}.$$

Para uma perspectiva bayesiana (CHO *et al.*, 2009) no diagnóstico de exclusão de casos, a calibração da divergência de Kullback–Leibler (divergência KL) é considerada com base no CPO, que mede a influência da observação nas estimativas de parâmetros. Para $i = 1,...,n$, $j = 1,...,n_i$, esta métrica é definida como

$$K\left(\pi(\boldsymbol{\theta} \mid D), \pi\left(\boldsymbol{\theta} \mid D_{(-ij)}\right)\right) = -\log\left(\text{CPO}_{ij}\right) + E\left[\log f(y_{ij} \mid \boldsymbol{\theta}) \mid D\right].$$

A estimativa de Monte Carlo da divergência KL é dada por

$$\widehat{K}\left(\pi(\boldsymbol{\theta} \mid D), \pi\left(\boldsymbol{\theta} \mid D_{(-ij)}\right)\right) = \log\left(\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f(y_{ij} \mid \boldsymbol{\theta}^{(l)})}\right) + \frac{1}{L}\sum_{l=1}^{L}\log f(y_{ij} \mid \boldsymbol{\theta}^{(l)}).$$

Consequentemente, seguindo Cho *et al.* (2009), a calibração para $i = 1,\ldots,n$, $j = 1,\ldots,n_i$, é definida como

$$p_{ij} = 0.5\left\{1 + \sqrt{1 - \exp\left(-2\widehat{K}\left(\pi(\boldsymbol{\theta} \mid D), \pi\left(\boldsymbol{\theta} \mid D_{(-ij)}\right)\right)\right)}\right\} \tag{4.8}$$

Uma observação pode ser considerada influente se seu valor de Equação 4.8 for muito maior que $0,5$.

4.3 Estudo de simulação

Para avaliar a metodologia proposta, consideramos o seguinte modelo de regressão binária longitudinal:

$$\begin{aligned} & Y_{ij} \mid b_i, \boldsymbol{\beta} \sim \text{Bernoulli}(\mu_{ij}) \\ \mu_{ij} &= F_\alpha(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_i) \\ &= \left(\frac{\exp\left(\beta_1 + \beta_2 x_{ij2} + \beta_3 x_{ij3} + b_i\right)}{1 + \exp\left(\beta_1 + \beta_2 x_{ij2} + \beta_3 x_{ij3} + b_i\right)} \right)^\alpha \\ & b_i \sim N(0, \sigma_b^2) \end{aligned} \tag{4.9}$$

em que os dados são simulados usando as seguintes especificações:
$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top = (1,5,\ -0,5,\ -0,25)^\top$, $\mathbf{X}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)^\top$, $x_{ij1} = 1$, $x_{ij2} \sim N(0,\ 1)$ e $x_{ij3} = j - 1$ é a variável tempo, para $i = 1, \ldots, n$ e $j = 1, \ldots, 10$. Para nosso estudo, consideramos $\sigma_b = 0,5$ e $\alpha = 4$. Esses valores permitiram um desbalanceamento de $0,20$ (proporção de uns) na variável resposta. Conforme proposto em Seção 4.2, adotamos as seguintes especificações a prioris: $\boldsymbol{\beta} \sim N_3\left(0,\ 100^2\mathbf{I}\right)$, $b_i \sim N\left(0,\ \sigma_b^2\right)$, $\log(\alpha) \sim U(-2,\ 2)$. Uma análise de sensibilidade para o parâmetro σ_b^2 foi necessária para confirmar que os resultados não foram influenciados pelas a prioris.

4.3.1 Análise de sensibilidade da a priori para o efeito aleatório

Para fazer uma análise de sensibilidade para especificação da a priori do parâmetro σ_b^2 do efeito aleatório, foi simulado um conjunto de dados com as especificações descritas acima, em seguida o modelo com função de ligação potência logística com diferentes a prioris para σ_b^2 foi ajustado, especificamente, consideramos três a prioris previamente utilizadas na literatura:

1. A priori 1a: $\sigma_b^2 \sim \text{IG}\,(0,001,\ 0,001)$ ou, equivalentemente, $\frac{1}{\sigma_b^2} \sim \text{Gamma}\,(0,001,\ 0,001)$. Uma distribuição a priori popular para termos de variação inicialmente usados nos documentos `WinBUGS` Exemplos I e II (LUNN *et al.*, 2000).

2. A priori 1b: $\dfrac{1}{\sigma_b^2} \sim \text{Pareto}\,(1,\ 0,001)$. Essa distribuição a priori foi sugerida em modelos de epidemiologia genética (SCURRAH; PALMER; BURTON, 2000).

3. A priori 1c: $\log\left(\sigma_b^2\right) \sim \text{Uniform}\,(-10,\ 10)$. Essa distribuição a priori foi sugerida na análise de ensaios randomizados de cluster (SPIEGELHALTER, 2001).

Os dados foram gerados considerando o tamanho da amostra $n = 500$. Cada modelo foi ajustado com 100 replicações e para cada replicação 8000 iterações foram usadas, descartando os primeiros 3000 como *burn-in*, com um intervalo de *thinning* igual a 5. Duas cadeias foram consideradas, resultando em um tamanho de amostra efetivo de 2000 amostras nas quais é baseada a inferência a posteriori.

Tabela 16 – Análise de sensibilidade a priori para diferentes escolhas da variância do efeito aleatório (Estimativa de parâmetros e critérios de comparação de modelos)

A priori	Parâmetro	Média a posteriori		Mediana a posteriori		DIC	WAIC	LOO	EAIC	EBIC
	Valor verdadeiro	Estimativa	RMSE	Estimativa	RMSE	Média (s.e.)	Média (s.e.)	Média (s.e.)	Média (s.e.)	Média (s.e.)
1a	$\beta_1 = 1,5$	1,736	0,342	1,704	0,331					
	$\beta_2 = -0,5$	-0,534	0,054	-0,528	0,051	3931,190	3975,760	3987,509	3783,587	3816,173
	$\beta_3 = -0,25$	-0,272	0,026	-0,268	0,024	(9,875)	(9,687)	(9,592)	(9,971)	(9,971)
	$\sigma_b^2 = 0,25$	0,311	0,083	0,297	0,073					
	$\alpha = 4$	4,327	0,740	4,235	0,734					
1b	$\beta_1 = 1,5$	1,724	0,336	1,692	0,325					
	$\beta_2 = -0,5$	-0,534	0,053	-0,528	0,050	3931,169	3975,722	3987,431	3783,510	3816,096
	$\beta_3 = -0,25$	-0,271	0,025	-0,268	0,023	(9,872)	(9,685)	(9,593)	(9,974)	(9,974)
	$\sigma_b^2 = 0,25$	0,310	0,082	0,296	0,072					
	$\alpha = 4$	4,303	0,724	4,207	0,716					
1c	$\beta_1 = 1,5$	1,732	0,339	1,700	0,327					
	$\beta_2 = -0,5$	-0,533	0,053	-0,527	0,050	3924,280	3968,770	3980,350	3776,580	3809,170
	$\beta_3 = -0,25$	-0,271	0,025	-0,268	0,023	(9,851)	(9,678)	(9,596)	(9,943)	(9,943)
	$\sigma_b^2 = 0,25$	0,309	0,082	0,295	0,071					
	$\alpha = 4$	4,316	0,726	4,222	0,716					

Na Tabela 16 mostramos os resultados da estimativa dos parâmetros e seus respectivos valores da raiz do erro quadrático médio (RMSE, do inglês: *root-mean-square error*), além disso, mostramos a média dos valores de *DIC*, *WAIC*, *LOO*, *EAIC* e *EBIC* considerando três distribuições a priori diferentes para σ_b^2. Note-se que os desempenhos dos critérios não variaram muito entre as diferentes a prioris, levando à mesma interpretação dos resultados, o que nos permite dizer que não há sensibilidade a priori. No entanto, a priori 1b, $\frac{1}{\sigma_b^2} \sim \text{Pareto}(1,\ 0,001)$, é mais preciso, pois recupera melhor os parâmetros com o menor valor de RMSE. Doravante, a priori 1b será considerada neste trabalho.

4.3.2 Desempenho da função de ligação potência e reversa de potência

O estudo de simulação foi desenvolvido com o objetivo de avaliar o desempenho de modelos com funções de ligação potência e reversa de potência em dados de resposta binária longitudinal, principalmente quando os dados são desbalanceados. Considerando o modelo de regressão binária longitudinal em (4.9), simulamos os dados do modelo usando a função de ligação potência logística e então diferentes funções de ligação simétrica como logística, Normal e Cauchy foram ajustadas para esses dados. Além disso, ajustamos os modelos con funções de ligação assimétricas estudadas em Seção 4.1, que podem ser alternativas aos dados desbalanceados em um estudo longitudinal para resposta binária. Os dados foram gerados considerando 3 tamanhos de amostra $n = \{100,\ 250,\ 500\}$, cada amostra com $n_i = 10$ medidas repetidas. Isso implica um total de $N = \{1000,\ 2500,\ 50000\}$ observações, respectivamente.

Portanto, o estudo de simulação compreende três cenários e para cada cenário foram consideradas 100 replicações. Para cada replicação, 8000 iterações foram usadas, descartando as primeiras 3000, com um intervalo de *thinning* de 5 e 2 cadeias resultando em 2000 amostras MCMC nas quais é baseada a inferência a posteriori.

O desempenho comparativo do método foi determinado com base no valor da raiz do erro quadrático médio (RMSE) e medidas de viés, respectivamente estimados por $RMSE(\widehat{\theta}_h) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_h^{(r)} - \theta_h\right)^2}$ e viés$(\widehat{\theta}_h) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_h^{(r)} - \theta_h\right)$ para $h = 1,2,3,4$, em que R é o número de réplicas na simulação e $\widehat{\theta}_h$ é a média a posteriori do parâmetro θ_h na réplica r. Os parâmetros analisados foram β_1, β_2, β_3 e σ_b^2.
As tabelas 20, 21 e 22, em Seção B.1, mostram um resumo das médias e medianas posteriores estimativas para os parâmetros β_1, β_2, β_3, σ_b^2 e α. Além disso, os valores de desvio padrão (DP), viés e RMSE são apresentados para três cenários, considerando tamanhos de amostra $n = 100$, $n = 250$ e $n = 500$.

Os resultados indicam, de maneira geral, que todos os parâmetros estimados dos modelos propostos, modelos com função de ligação assimétrica, apresentam menor RMSE e, à medida que o tamanho da amostra aumenta, as estimativas desses modelos tornam-se mais precisas. Este comportamento não é observado quando se considera a função de ligação logística simétrica, em que o RMSE não varia muito com o aumento do tamanho da amostra, indicando que este modelo não é adequado para os dados simulados desbalanceados.
Além disso, como resumo e ilustração, a Figura 11 mostra uma comparação RMSE das funções de ligação propostas com as funções de ligação comuns. Neste caso consideramos apenas os modelos com a funções de ligação potência logística (PL), reversa de potência logística (RPL) e potência Cauchy (PC), pois os resultados considerando outras funções de ligação, são semelhantes. Pode-se observar que para diferentes tamanhos de amostra, ou seja, 100, 250 e 500, os modelos com função de ligação potência logística e reversa de potência logística, têm uma estimativa melhor em comparação com modelos com funções de ligação simétrica (como logística), para todos os parâmetros estudados. No caso do parâmetro α, podemos ver que o modelo com função de ligação PC recupera o parâmetro adequadamente e o RMSE diminui conforme o tamanho da amostra aumenta; entretanto, as estimativas são diferentes para cada um das funções de ligação estudadas, isso porque em cada uma das distribuições de potência e reversa de potência este parâmetro tem um significado diferente.

4.4 Aplicação

Como uma aplicação das funções de ligação potência e reversa de potência a dados de resposta binária longitudinal, nesta seção, apresentamos uma análises dos dados sobre os sintomas (distúrbios do pensamento) da esquizofrenia de Madras. Madras é um estudo Longitudinal de Esquizofrenia que investigou o curso dos sintomas psiquiátricos positivos e negativos durante

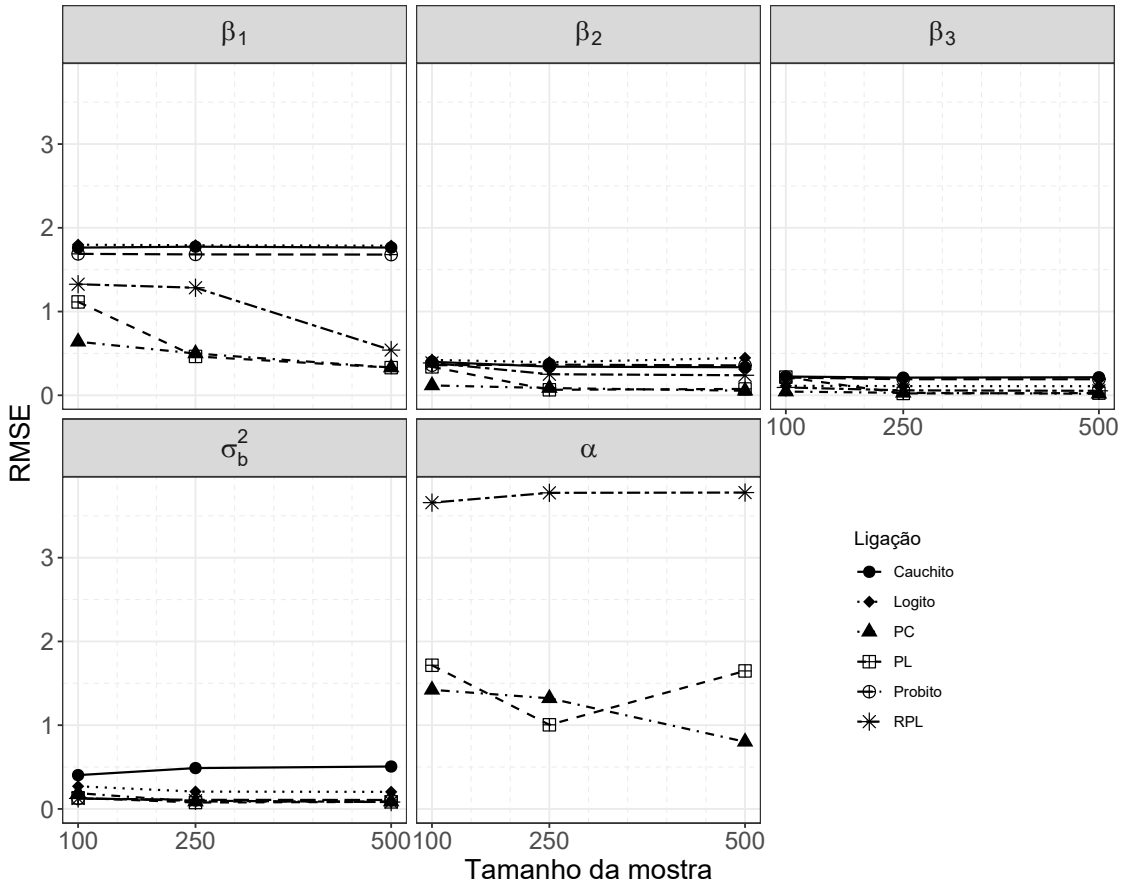

Figura 11 – RMSE para β_1, β_2, β_3 (painel superior), σ_b^2 (painel inferior esquerdo) e α (painel inferior direito) com diferentes tamanhos de amostra $(100,\ 250,\ 500)$ e diferentes métodos de estimação: Logístico (pontilhado), Normal (ponto pontilhado), Cauchy (sólido), potência logístico (traço longo), reversa de potência logística (dois traços) e potência Cauchy (tracejado).

o primeiro ano após a hospitalização inicial por doença (THARA *et al.*, 1994). Este conjunto de dados também foi analisado em Diggle (2002, p. 234–243). Os dados contêm medidas de resultados binários que indicam a presença ou ausência de registro psiquiátrico positivo por 11 meses no primeiro ano após a hospitalização por esquizofrenia para 86 pacientes. O conjunto de dados também contém o indicador binário da idade do paciente na hospitalização ($0 =$ idade ≥ 20, $1 =$ idade < 20), sexo ($0 =$ masculino, $1 =$ feminino) e as interações entre essas covariáveis com o tempo (mês). A proporção da presença de sintomas é $0,31$, então o conjunto de dados é desbalanceado.

- Variável de resposta Y_{ij}: presença ou ausência de sintomas para o sujeito i no mês j. Para $i = 1,\ldots,86$ e $j = 0,\ldots,11$.
- Covariáveis: sexo, mês e indicador de idade.

A Figura 12 mostra o gráfico do perfil médio do sintoma de esquizofrenia por gênero. Pode-se observar que não há interação entre os grupos masculino e feminino. Além disso, ambos os gêneros apresentam o mesmo tipo de comportamento, ou seja, a presença do sintoma diminui

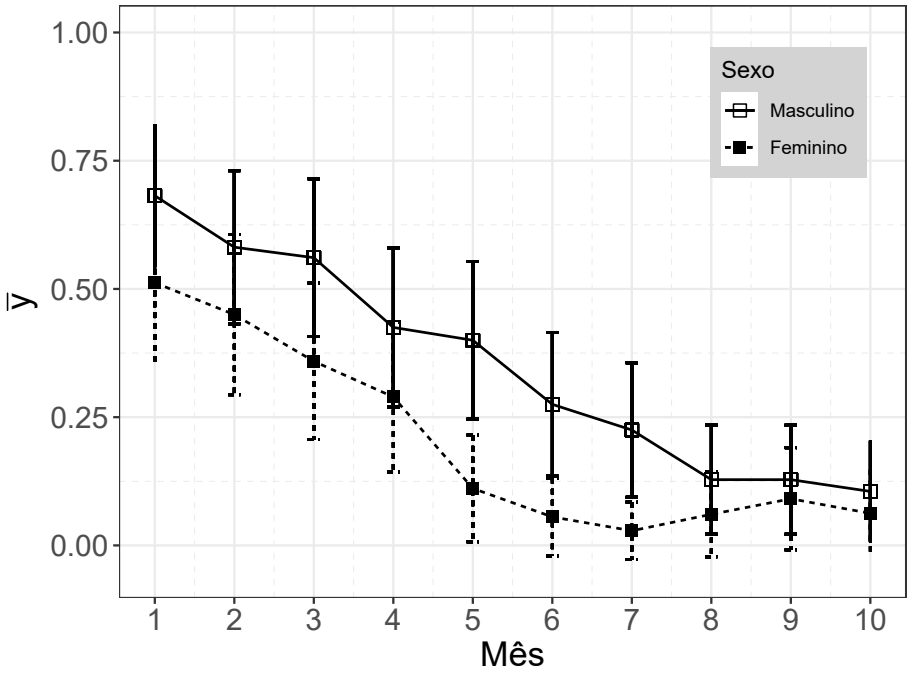

Figura 12 – Perfil médio do sintoma de esquizofrenia: por sexo

com o passar do tempo, sendo que o gênero masculino apresenta média maior que o gênero feminino após o segundo mês. Esta é uma indicação de que o tempo e o gênero são importantes para explicar a presença de sintomas de esquizofrenia.

Inicialmente, consideramos um modelo linear com intercepto e ajustamos o conjunto de dados com todas as covariáveis usando as funções de ligação Logística, P-Logística e RP-Logística. A covariável *idade* foi identificada como não significativa uma vez que o intervalo HPD de 95% contém o valor 0 para o coeficiente correspondente a esta variável. Posteriormente, um modelo quadrático é considerado como um modelo alternativo. Adicionalmente, também foram consideradas as versões sem intercepto seguindo Coelho, Russo e Bazán (2022). Portanto, para este conjunto de dados excluindo a covariável *idade*, os seguintes modelos foram ajustados usando os diferentes funções de ligação potência e reversa de potência para resposta binária longitudinal:

$$Y_{ij} \mid b_i \sim Bernoulli(\mu_{ij}), \quad b_i \sim N(0,\ \sigma_b^2) \tag{4.10}$$

- Modelo 1: $\mu_{ij} = F_l(\beta_1 + \beta_2 \times \text{sexo} + \beta_3 \times \text{mês} + b_i)$
- Modelo 2: $\mu_{ij} = F_l(\beta_1 \times \text{sexo} + \beta_2 \times \text{mês} + b_i)$
- Modelo 3: $\mu_{ij} = F_l(\beta_1 + \beta_2 \times \text{sexo} + \beta_3 \times \text{mês} + \beta_4 \times \text{mês}^2 + b_i)$
- Modelo 4: $\mu_{ij} = F_l(\beta_1 \times \text{sexo} + \beta_2 \times \text{mês} + \beta_3 \times \text{mês}^2 + b_i)$

em que as a prioris foram $\boldsymbol{\beta} \sim N(\mathbf{0},\, \boldsymbol{I}100^2)$, $\log(\alpha) \sim U(-2,\, 2)$ e $\frac{1}{\sigma_b^2} \sim \text{Pareto}(1,\, 0{,}001)$.

Cada um dos modelos foi ajustado considerando os seguintes valores para *MCMC*: 15000 iterações foram usadas, descartando os primeiros 10000 como *burn-in*, com um intervalo de *thinning* igual a 5, foram consideradas 2 cadeias, resultando em um tamanho de amostra efetivo de 2000 amostras nas quais é baseada a inferência a posteriori. O tempo computacional para cada modelo foi aproximadamente de 30 minutos, usando um computador com as seguintes caraterísticas: processador Intel Core *i*7 da 11^a geração, com uma memória RAM de 16*GB* e memoria gráfica de 8*GB*.

As comparações dos modelos são desenvolvidas usando os critérios discutidos na subseção 4.2.2. Os resultados destes critérios para todos os modelos ajustados com diferentes funções de ligação, são mostrados em Tabela 17. Os dois modelos que apresentam melhor desempenho são o modelo 1 e o modelo 3 com função de ligação PN.

Tabela 17 – Critérios de seleção de modelos com função de ligação assimétrica potência e reversa de potência para dados de sintomas de esquizofrenia de Madras

	Modelo 1						Modelo 2					
Ligação	DIC	WAIC	LOO	EAIC	EBIC	IC	DIC	WAIC	LOO	EAIC	EBIC	IC
L	711,1	716,8	717,4	682,1	701,4	748,1	731,4	737,1	737,6	701,3	715,8	767,5
N	668,5	679,8	682,0	630,8	650,1	714,1	674,6	685,8	688,2	636,1	650,5	719,1
C	732,2	737,0	737,2	704,1	723,4	768,3	760,0	764,9	765,0	729,7	744,2	796,3
PL	666,1	676,1	678,1	635,0	659,1	707,1	737,8	742,9	743,4	710,9	730,2	772,7
PN	**654,1**	**668,2**	**672,1**	**615,1**	**639,2**	**703,0**	681,5	690,8	692,5	647,3	666,6	723,7
PC	683,3	690,7	691,1	652,7	676,8	724,0	731,9	738,4	739,0	701,6	721,0	770,1
RPL	697,6	707,0	708,9	665,6	689,7	739,7	700,0	708,4	709,9	666,9	686,2	741,1
RPN	664,4	678,3	681,6	625,2	649,3	713,7	**662,2**	**675,8**	**679,1**	**621,5**	**640,8**	**710,9**
RPC	731,3	736,2	736,5	706,5	730,6	766,0	731,9	738,4	739,0	701,6	721,0	770,1
	Modelo 3						Modelo 4					
Ligação	DIC	WAIC	LOO	EAIC	EBIC	IC	DIC	WAIC	LOO	EAIC	EBIC	IC
L	711,2	717,3	717,8	681,2	700,5	749,2	734,0	739,9	740,5	704,7	724,1	771,3
N	664,0	676,0	678,4	625,9	645,2	710,1	675,8	687,1	689,5	638,8	658,1	720,9
C	734,8	739,8	739,9	704,9	724,2	772,7	763,3	768,0	768,2	733,9	753,2	800,7
PL	661,6	671,9	674,3	630,3	654,4	703,0	740,1	745,5	746,0	714,0	738,2	776,2
PN	**647,2**	**662,0**	**666,1**	**607,6**	**631,8**	**696,8**	683,8	693,5	695,3	651,3	675,5	726,4
PC	679,6	688,3	688,5	650,1	674,2	719,0	742,7	749,2	749,7	711,6	735,8	783,8
RPL	698,2	707,6	709,2	665,5	689,7	740,9	700,3	709,2	710,5	667,8	691,9	742,8
RPN	661,9	675,9	679,5	622,4	646,5	711,4	**661,1**	**675,5**	**678,7**	**620,9**	**645,0**	**711,3**
RPC	732,7	737,8	738,1	707,1	731,2	768,4	758,7	764,5	765,1	732,4	756,5	795,0

Resumos a posteriori dos parâmetros do modelo 1 e modelo 3, com função de ligação PN, são mostrados em Tabela 18. Observe que todos os parâmetros, com exceção de β_4 do modelo 3, são indicados como significativamente diferentes de 0 a um nível de credibilidade de 95%. Portanto, embora o modelo 3 apresente melhor desempenho já que o parâmetro β_4 não é significativo, vamos considerar o modelo 1 para este conjunto de dados. Além disso, os coeficientes correspondentes ao gênero e ao tempo influenciam negativamente na presença de

Tabela 18 – Estimativa a posteriori dos parâmetros para o modelo de resposta binária com função de ligação PN do modelo 1 e modelo 3, para dados de sintomas de esquizofrenia de Madras

Modelo	Parâmetro	Média a posteriori	Desvio padrão	95% I.C.	
Modelo 1	β_1	1,600	0,268	(1,036;	2,072)
	β_2	-0,438	0,211	(-0,873;	-0,021)
	β_3	-0,209	0,026	(-0,266;	-0,165)
	σ_b^2	0,732	0,154	(0,429;	0,984)
	α	4,215	1,625	(1,539;	7,095)
Modelo 3	β_1	1,777	0,275	(1,156;	2,251)
	β_2	-0,452	0,215	(-0,875;	-0,032)
	β_3	-0,305	0,052	(-0,417;	-0,211)
	β_4	0,009	0,004	(-0,001;	0,018)
	σ_b^2	0,753	0,151	(0,454;	0,986)
	α	4,404	1,670	(1,516;	7,230)

sintomas, ou seja, as chances da presença de sintomas nos homens diminuem com o passar do tempo.

Além disso, mostramos verificações preditivas a posteriori, análise bayesiano dos resíduos e análise de influencia para o modelo escolhido, a fim de verificar a adequação do modelo com um a função de ligação potência normal.

Na seção seguinte, desenvolvemos a verificação preditiva a posteriori considerando a média e o desvio padrão como variáveis de discrepância para a variável resposta sob o modelo com a função de ligação potência normal, conforme descrito na subseção 4.2.3. Na Figura 13, em cada gráfico, o valor observado de $T(\boldsymbol{y}, \boldsymbol{\theta})$ é mostrado como uma barra vertical de dois traços em um histograma representando os valores de $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ para 1000 sorteios de $\boldsymbol{y}^{rep}$ obtidos da distribuição preditiva a posteriori. Isto é para as duas variáveis de discrepância, média (Figura 13 (a)) e desvio padrão (Figura 13 (b)). Também, calculamos o *valor p* preditivo a posteriori estimado para as duas variáveis de discrepância. Os valores médios dos p preditivos a posteriori estimados foram $0,48$ para a média e $0,45$ para o desvio padrão, sugerindo que o modelo parece ser uma opção de modelagem plausível para o conjunto de dados desde que nenhum valor extremo (próximo de 0 ou 1) foi observado. Além disso, esta conclusão também é vista pelos histogramas apresentados em Figura 13 uma vez que os valores das discrepâncias estão centrados em torno do valor médio e desvio padrão das respostas. Isso indica que os dados observados provavelmente ocorrerão no modelo selecionado.

Portanto, os valores das verificações preditivas a posteriori indicam que não podemos detectar que existe uma especificação incorreta do modelo, ou seja, um modelo com uma função de ligação potência normal pode ser adequado.
Além disso, consideramos um diagnóstico de ajuste através de análise dos resíduos quantílicos aleatórios, conforme proposto na subseção 4.2.5, com envelopes gerados a partir do modelo com

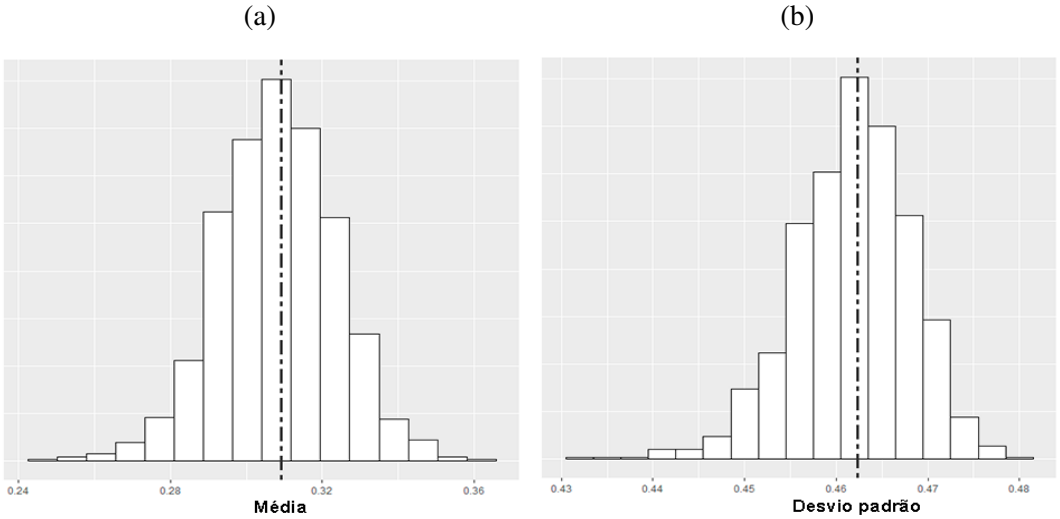

Figura 13 – Verificações preditivas a posteriori para as discrepâncias média (a) e desvio padrão (b) (as barras de dois traços indicam os valores observados da estatística $T(\boldsymbol{y}, \boldsymbol{\theta})$ e o histograma exibe $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ de 1000 sorteios de $\boldsymbol{y}^{rep}$ sob o modelo com função de ligação potência normal).

função de ligação potência normal conforme mostrado em Figura 14. Observe que nenhuma observação está fora do envelope simulado, o que dá evidência de um bom ajuste.

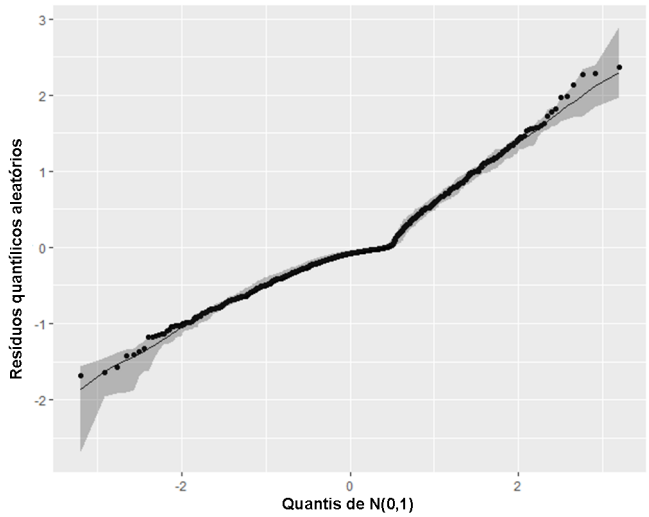

Figura 14 – Resíduos quantílicos aleatórios para um modelo de resposta binária com uma função de ligação potência normal para dados de sintomas de esquizofrenia de Madras

Por fim, apresentamos uma análise de influencia através do gráfico da medida de calibração apresentado em (4.8). De Figura 15, pode-se observar que a medida de calibração identifica as observações 654, 804, 247, 898 e 834 como possíveis observações influentes.

A Tabela 19 mostra as estimativas a posteriori dos parâmetros e a porcentagem de mudança na estimativa dos parâmetros do modelo com a função de ligação potência normal, como consequência da eliminação das observações identificadas como possíveis observações influentes.

Como pode ser visto em Tabela 19, a exclusão da observação (654) e (654, 804) gera a

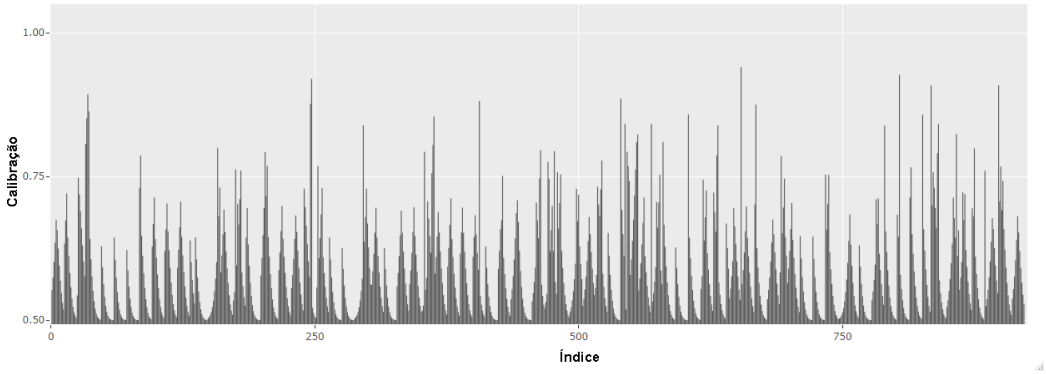

Figura 15 – Medida de calibração.

Tabela 19 – Estimativas e variação percentual da estimativa dos parâmetros quando as observações são excluídas

Observações observatexcluídasions	Média a posteriori (porcentagem de mudança)				
	β_1	β_2	β_3	σ_b^2	α
sem excluir	1,6116	-0,4477	-0,2094	0,7301	4,2972
(654)	1,6269	-0,4530	-0,2156	0,7536	4,3070
	(0,95%)	(1,19%)	(2,96%)	(3,22%)	(0,23%)
(654, 804)	1,6532	-0,4402	-0,2109	0,7479	4,6280
	(2,58%)	(-1,68%)	(0,71%)	(2,44%)	(7,70%)
(654, 804, 247)	1,6892	-0,4579	-0,2128	0,7556	4,6566
	(4,82%)	(2,28%)	(1,65%)	(3,50%)	(8,37%)
(654, 804, 247, 898)	1,7090	-0,4729	-0,2145	0,7708	4,6605
	(6,04%)	(5,62%)	(2,46%)	(5,58%)	(8,46%)
(654, 804, 247, 898, 834)	1,7473	-0,4879	-0,2157	0,7811	4,7704
	(8,42%)	(8,99%)	(3,00%)	(6,99%)	(11,01%)

menor variação percentual relativa na estimativa dos parâmetros em comparação com a exclusão de outros casos com três ou mais parâmetros.

4.5 Comentários finais

Neste trabalho, realizamos uma aplicação dos modelos estudados, para o conjunto de dados de Madras, o qual apresenta uma proporção de uns de 31% indicando um desbalanceamento moderados, o que sugere que uma ligação assimétrica pode ser explorada como mencionado por Chen, Dey e Shao (1999) e Qiu *et al.* (2013). A utilização das funções de ligação potência e reversa de potência baseadas na fda de uma família de distribuições simétricas padronizadas, foi apresentada como uma alternativa aos modelos longitudinais de resposta binária considerando dados desbalanceados.

A classe de funções de ligação mostradas, é flexível e uma ampla família de funções de ligação potência e reversa de potência pode ser considerada e funções de ligação simétrica como logito, cauchito e probito, são casos particulares quando $\alpha = 1$.
Realizamos uma análise bayesiana para modelos longitudinais de resposta binária considerando as funções de ligação propostas. Assim, um procedimento MCMC usando o algoritmo NUTS

foi desenvolvido para a estimação dos parâmetros do modelo e uma análise de diagnóstico foi proposta considerando verificações preditivas a posteriori, resíduos quantílicos aleatórios bayesiano e medidas de influência bayesiana.

Nossos resultados de estudo de simulação mostraram que o desempenho desses modelos para dados desbalanceados é melhor do que modelos de regressão binária longitudinal com funções de ligação simétrica. Além disso, na aplicação, foi mostrado que um modelo da classe de função de ligação potência e reversa de potência, é o melhor modelo para dados de Madras.

Em relação ao parâmetro de forma, associado às funções de ligação aqui estudadas, e aos coeficientes de regressão, podemos notar que o parâmetro α pode afetar ambos os parâmetros (β_0 e β_1), porém, o efeito em β_0 ocorreria de forma isolada e direta porque β_1 vem acompanhado da covariável x. Neste trabalho, todos os estudos de simulação mostraram que houve uma excelente recuperação do parâmetro β_0 para todos os modelos nos diferentes cenários. Portanto, não encontramos nenhum problema, com nosso método e com a especificação a priori considerada, que indique algum problema na estimativa de β_0. Recentemente, em Niekerk e Rue (2021), considerando a função de ligação *skew probit*, que segue Bazán, Bolfarine e Branco (2010), propõem o uso de uma versão padronizada desta função de ligação assimétrica, porém, não há nenhuma prova formal de que seja necessário fazer essa padronização porque isso irá depender dos dados e do método de estimação. No apêndice, na Seção B.4, exploramos o uso da versão padronizada da função de ligação PL, chamado aqui como função de ligação SPL, mostrando que ambos as funções de ligação podem caber de maneira semelhante no mesmo conjunto de dados da aplicação, e ambos podem ser razoavelmente usados quando os dados são "gerados a partir de qualquer um desses modelos". É interessante observar que as funções de ligação padronizadas exigem que a média e a variância sejam explícitas.
Outra alternativa possível é não considerar o intercepto (COELHO; RUSSO; BAZÁN, 2022), mas neste caso, em nossos dados, vimos que ignorar o intercepto altera fortemente a interpretação dos parâmetros do modelo. Por fim, em um estudo recente, de Ordoñez *et al.* (2023) mostra-se que a utilização de uma a priori, como o proposto por Niekerk e Rue (2021), no caso de função de ligação potência logística, pode reduzir o intervalo de credibilidade em relação a uma a priori Uniforme, mas as estimativas pontuais permanecem as mesmas. Sugerimos estudos adicionais para confirmar a possível vantagem do uso da versão padronizada das funções de ligação aqui propostas.

Aplicações desta função de ligação, onde as funções de ligação simétrica não são justificados, podem ser obtidos considerando nossa abordagem. Inclui modelagem multinível, modelos de resposta a itens e modelos zero ou um aumentados. Extensões para um modelo de resposta ordinal longitudinal podem ser desenvolvidas como trabalhos futuros.

Além disso, versões multivariadas para respostas binárias podem ser propostas de maneira semelhante ao Genest *et al.* (2013).

CAPÍTULO

5

COMENTÁRIOS FINAIS E DESENVOLVIMENTOS FUTUROS

Nesse capítulo são apresentados alguns comentários finais sobre os principais resultados e as contribuições deste trabalho. Além disso, descrevemos as produções resultantes do mesmo. Finalmente, comentários gerais sobre possíveis desenvolvimentos futuros é apresentado.

5.1 Comentários finais

Neste trabalho, foram estudados modelos alternativos para classificação em dados desbalanceados, especificamente foram estudados modelos para classificação binária, em dois contextos diferentes: modelos para classificação binária usual (caso MLG) e modelos para classificação binária mista para dados com resposta binária longitudinal (caso MLGMs). Para cada situação, consideramos o uso de uma nova classe de funções de ligação assimétrica que são baseadas nas distribuições chamadas potência e reversa de potência. Essas funções de ligação são uma boa alternativa às funções de ligação simétrica comuns como logística e normal, tanto para classificação binária usual e quanto para os dados longitudinais. Neste último contexto (caso longitudinal), é poco explorada na literatura sobre propostas de novas funções de ligação.
As distribuições consideradas neste trabalho, foram exploradas e mostraram excelente desempenho em diferentes situações, por exemplo, Bazán *et al.* (2017), Lemonte e Bazán (2018), Chumbimune (2017) e De la Cruz *et al.* (2019). Nesse sentido, pode ser uma motivação de discussão para demais pesquisadores sobre a utilidade desse tipo de modelos, na aplicação em diversas áreas do conhecimento.

Em diferentes áreas de aplicação de aprendizado supervisionado para classificação binária, é considerado o modelo de regressão logística como o mais usual. No entanto, nos estudos de simulação, dos capítulos 3 e 4, respetivamente, foram mostrados que, na presença de dados desbalanceados, a função de ligação simétrica, como o caso Logístico não são apropriados.

Então, as funções de ligação assimétrica podem ser uma boa alternativa para esse tipo de dados separando o efeito do intercepto com o efeito da curva associada ao formato da distribuição. Similares resultados foram obtidos, nas aplicações com presença de desbalanceamento dos dados.

Por outro lado, foi mostrado também a importância do estudo das propriedades das funções de ligação, para evitar os efeitos negativos da especificação incorreta do modelo, principalmente, na presença de dados desbalanceados, o qual é muito comum na prática. Nesse sentido, algumas funções de ligação assimétrica, como o caso de *PG* e sua reversa, resultaram ser inviáveis para regressão binária, e consequentemente, para classificação binária.

A seleção de modelo é uma etapa fundamental na modelagem dos dados, uma seleção errada pode levar a um predições erradas e portanto, toma de decisões erradas. Em vista disso, foi mostrado a importância de usar as métricas adequadas no momento de selecionar um modelo, concluindo que algumas métricas mais usadas na literatura, como a área sob a curva (*ACC*) e precisão (*ACC*), demonstram baixo desempenho em todos os cenários estudados.

Foi realizado uma análise sob abordagem Bayesiana para os dois tipos de modelos (MLG e MLGMs), na qual mostramos que é muito simples de implementar em termos computacionais, tanto para estimação de parâmetros quanto para análise de diagnóstico do modelo. Contudo, novos métodos ou algoritmos de estimação podem ser necessários para regressão binária com ligações assimétricas, pois existe um custo de tempo computacional, algoritmos mais rápidos na execução da amostragem a posteriori, pode ser mais adequados em conjuntos de dados de alta dimensão. Além disso, outros tipos de a prioris, como o estudado em Ordoñez *et al.* (2023), podem ser considerados para o parâmetro α.

Extensões para regressão binomial e teoria de resposta ao item de algumas dessas funções de ligação foram estudadas recentemente em Alves, Bazán e Arellano-Valle (2023) e Bazán *et al.* (2023), onde é mostrada a importância do uso funções de ligação assimétrica.

Finalmente, estudo com maior meticulosidade sobre o uso de funções de ligação assimétrica, ainda possui espaços para ser enriquecida com novas pesquisas e propostas.

5.2 Produções

5.2.1 Trabalhos presentado em eventos

- De la Cruz Huayanay, A. and Jorge L. Bazán and Cibele Russo (2023). "Performance of evaluation metrics for classification in imbalanced data". *Pontificia Universidad Católica de Perú - PUCP. 7th Latin American Conference on Statistical Computing - LACSC 2023. Palestra*.

- De la Cruz Huayanay, A. e Bazán, J. L. (2020). "Longitudinal data analysis for binary

response using alternative links". *UFSCar/USP, Brazil. 8th Workshop on Probabilistic and Statistical Methods. Pôster*.

- De la Cruz Huayanay, A. and Bazán, J. L. (2020). "Bayesian Mixed effects regression models for longitudinal binary response". *I Jornada Internacional de Estadística – Día del Estadístico Peruano. Universidad Nacional Mayor de San Marcos, Perú. Palestra*.

- De la Cruz Huayanay, A. and Bazán, J. L. (2020). "Regresión binaria bayesiana". *I CONGRESO INTERNACIONAL DE CIENCIAS MATEMÁTICAS - UNMSM. Universidad Nacional Mayor de San Marcos, Perú. Minicurso*.

5.2.2 Artigos publicados

- De la Cruz Huayanay, A., Bazán, J. L., & Ribeiro Diniz, C. A. (2023). Longitudinal binary response models using alternative links for medical data. Brazilian Journal of Probability and Statistics, 37(2), 365-392..

5.2.3 Artigos submetidos

- De la Cruz Huayanay, A., Bazán, J. L., Cibele Russo. (2023). Performance of evaluation metrics for classification in imbalanced data. *Springer*.

5.3 Desenvolvimentos futuros

Até o momento, o trabalho do capítulo 3 está finalizado e foi submetido para publicação. O estudo do capítulo 4, é uma extensão do trabalho de De la Cruz *et al.* (2019), e já foi publicado. Nesse sentido, ficam algumas propostas para trabalhos futuros:

- Considerando os modelos apresentados, podem se realizar estudos de comparação do desempenho das funções de ligação assimétrica, com alguns algoritmos de classificação utilizados em aprendizado de máquina como: Bayes Ingênua (em inglês *Naive Bayes*), K-vizinhos mais próximos (em inglês *K-Nearest Neighbors*), Árvore de decisão (em inglês *Decision Tree*), Máquinas de vetores de suporte (em inglês *Support Vector Machines*) e Floresta Aleatória (em inglês *Random Forest*).

- Considerando os modelos apresentados, podem se realizar estudos de comparação do desempenho das funções de ligação assimétrica, com algumas estratégias para lidar com desbalanceamento, por exemplo, re-amostragem e validação cruzada.

- Uma extensão dos métodos desenvolvidos neste trabalho seria considerar quando uma classificação é multi-classe, neste caso, os valores possíveis da variável resposta Y são K

categorias ou classes, ou seja, para cada observação i da forma $(\boldsymbol{x}_i, y_i)$, onde $\boldsymbol{x}_i$ é o vetor de covariáveis, $y_i \in \{1, \ldots, K\}$ é o $i$ésima etiqueta de classe. Para este tipo de problema, é comum o uso da Regressão Logística Multinomial; no entanto, pode haver situações em que uma ou mais turmas tenham uma proporção significativamente menor do que outras, e acreditamos que a metodologia estudada neste trabalho pode ser uma boa alternativa quando houver desequilíbrio de turmas.

REFERÊNCIAS

ABANTO-VALLE, C. A.; DEY, D. K.; JIANG, X. Binary state space mixed models with flexible link functions: a case study on deep brain stimulation on attention reaction time. **Statistics and Its Interface**, International Press of Boston, v. 8, n. 2, p. 187–194, 2015. Citado na página 65.

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 39.

ALBERT, J.; CHIB, S. Bayesian residual analysis for binary response regression models. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 747–769, 1995. Citado na página 70.

ALVES, J. S.; BAZÁN, J. L.; ARELLANO-VALLE, R. B. Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. **Biometrical Journal**, Wiley Online Library, v. 65, n. 3, p. 2100325, 2023. Citado na página 84.

ALZAHRANI, A.; SADAOUI, S. Scraping and preprocessing commercial auction data for fraud classification. **arXiv preprint arXiv:1806.00656**, 2018. Citado na página 59.

ANDO, T. Predictive bayesian model selection. **American Journal of Mathematical and Management Sciences**, Taylor & Francis, v. 31, n. 1-2, p. 13–38, 2011. Citado na página 68.

AYODELE, T. O. Types of machine learning algorithms. **New advances in machine learning**, InTech Rijeka, Croatia, v. 3, p. 19–48, 2010. Citado na página 21.

BASU, S.; MUKHOPADHYAY, S. Bayesian analysis of binary regression using symmetric and asymmetric links. **Sankhyā: The Indian Journal of Statistics, Series B**, JSTOR, p. 372–387, 2000. Citado na página 22.

BAZÁN, J.; BAYES, C. Inferencia bayesiana en modelos de regresion binaria usando brmuw. **Reporte de Investigacion. Serie B. Nro**, v. 25, 2010. Citado na página 35.

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 na página 22.

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 22, 23, 25, 28, 29, 33, 36, 43, 47, 53, 65, 66 e 83.

BAZÁN, J. L.; ARI, S. E. F.; AZEVEDO, C. L. N.; DEY, D. K. Revisiting the Samejima–Bolfarine–Bazán IRT models: New features and extensions. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 37, n. 1, p. 1 – 25, 2023. Disponível em: <https://doi.org/10.1214/22-BJPS558>. Citado na página 84.

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 81.

BAZÁN, J. L.; ROMEO, J. S.; RODRIGUES, J. Bayesian skew-probit regression for binary response data. **Brazilian Journal of Probability and Statistics**, v. 28, n. 4, p. 467–482, 2014. Citado na página 65.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American statistical Association**, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993. Citado na página 22.

BROWNLEE, J. **Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end**. [S.l.]: Machine Learning Mastery, 2016. Citado na página 33.

____. **Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning**. [S.l.]: Machine Learning Mastery, 2020. Citado nas páginas 36, 38 e 54.

BRYS, G.; HUBERT, M.; STRUYF, A. A robust measure of skewness. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 13, n. 4, p. 996–1017, 2004. Citado na página 48.

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 44.

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 nas páginas 22, 36 e 80.

CHO, H.; IBRAHIM, J. G.; SINHA, D.; ZHU, H. Bayesian case influence diagnostics for survival models. **Biometrics**, Wiley Online Library, v. 65, n. 1, p. 116–124, 2009. Citado na página 71.

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 nas páginas 37 e 38.

CHUMBIMUNE, S. A. **Regressão binária usando ligações potência e reversa de potência**. Dissertação (Master's Thesis) — University of São Paulo, 2017. Citado nas páginas 26, 48 e 83.

COELHO, F. R.; RUSSO, C. M.; BAZÁN, J. L. On outliers detection and prior distribution sensitivity in standard skew-probit regression models. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 36, n. 3, p. 441–462, 2022. Citado nas páginas 76 e 81.

COLLETT, D. **Modelling binary data**. [S.l.]: CRC press, 2002. Citado na página 105.

CONOVER, W. J. **Practical nonparametric statistics**. [S.l.]: John Wiley & Sons, 1999. v. 350. Citado na página 55.

Da Silva, A. N.; ANYOSA, S.; BAZÁN, J. L. Modelagem bayesiano de regressão binária para dados desbalanceados usando novas ligações. **Brazilian Journal of Biometrics**, v. 38, n. 4, p. 385–417, 2020. Citado na página 36.

De la Cruz, A. **Modelos de regressão para resposta binária na presença de dados desbalanceados**. Dissertação (Master's Thesis) — University of São Paulo, 2019. Citado nas páginas 25, 26, 30, 38 e 48.

De la Cruz, A.; BAZÁN, J. L.; CANCHO, V. G.; DEY, D. K. Performance of asymmetric links and correction methods for imbalanced data in binary regression. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 89, n. 9, p. 1694–1714, 2019. Citado nas páginas 21, 22, 23, 29, 36, 37, 38, 53, 54, 59, 65, 66, 71, 83 e 85.

DIGGLE, P. **Analysis of longitudinal data**. [S.l.]: Oxford university press, 2002. Citado nas páginas 42 e 75.

DIGGLE, P.; LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis. **New York: Oxford University Press**, v. 5, p. 13, 1994. Citado nas páginas 22 e 42.

DING, J.; TAROKH, V.; YANG, Y. Model selection techniques: An overview. **IEEE Signal Processing Magazine**, IEEE, v. 35, n. 6, p. 16–34, 2018. Citado na página 36.

DUA, D.; TANISKIDOU, E. K. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. Citado na página 59.

DUDA, R. O.; HART, P. E. *et al.* **Pattern classification**. [S.l.]: Wiley New York, 2001. Citado na página 32.

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 70 e 71.

FATOURECHI, M.; WARD, R. K.; MASON, S. G.; HUGGINS, J.; SCHLÖGL, A.; BIRCH, G. E. Comparison of evaluation metrics in classification applications with imbalanced datasets. In: IEEE. **2008 seventh international conference on machine learning and applications**. [S.l.], 2008. p. 777–782. Citado nas páginas 21 e 36.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 37.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. **Applied longitudinal analysis**. [S.l.]: John Wiley & Sons, 2012. v. 998. Citado nas páginas 22, 42, 43 e 64.

FONG, Y.; RUE, H.; WAKEFIELD, J. Bayesian inference for generalized linear mixed models. **Biostatistics**, v. 11, n. 3, p. 397–412, 07 2010. Citado nas páginas 43 e 66.

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 nas páginas 41 e 68.

GELFAND, A. E.; DEY, D. K.; CHANG, H. **Model determination using predictive distributions with implementation via sampling-based methods**. [S.l.], 1992. Citado na página 71.

GELMAN, A. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). **Bayesian Analysis**, v. 1, n. 3, p. 515–534, 2006. Citado nas páginas 53 e 66.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: CRC press, 2013. Citado na página 69.

GELMAN, A.; GOEGEBEUR, Y.; TUERLINCKX, F.; MECHELEN, I. V. Diagnostic checks for discrete data regression models using posterior predictive simulations. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 49, n. 2, p. 247–268, 2000. Citado na página 69.

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 nas páginas 40 e 41.

GENEST, C.; NIKOLOULOPOULOS, A. K.; RIVEST, L.-P.; FORTIN, M. Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. 2013. Citado na página 81.

GIBBONS, R. D.; HEDEKER, D. Random effects probit and logistic regression models for three-level data. **Biometrics**, JSTOR, p. 1527–1537, 1997. Citado nas páginas 22 e 23.

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 48.

HAMMING, R. W. On the distribution of numbers. **The bell system technical journal**, Nokia Bell Labs, v. 49, n. 8, p. 1609–1625, 1970. Citado na página 66.

HLOSTA, M.; STRÍZ, R.; KUPCÍK, J.; ZENDULKA, J.; HRUSKA, T. Constrained classification of large imbalanced data by logistic regression and genetic algorithm. **International Journal of Machine Learning and Computing**, IACSIT Press, v. 3, n. 2, p. 214, 2013. Citado na página 22.

HOFFMAN, M. D.; GELMAN, A. *et al.* The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **J. Mach. Learn. Res.**, v. 15, n. 1, p. 1593–1623, 2014. Citado nas páginas 45 e 67.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International journal of data mining & knowledge management process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015. Citado na página 36.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions, volume 2**. [S.l.]: John wiley & sons, 1995. v. 289. Citado nas páginas 28 e 102.

KARIMI, B.; LAVIELLE, M. Efficient metropolis-hastings sampling for nonlinear mixed effects models. In: SPRINGER. **Bayesian Statistics and New Generations: BAYSM 2018, Warwick, UK, July 2-3 Selected Contributions**. [S.l.], 2019. p. 85–93. Citado na página 67.

KIM, H.-J. Binary regression with a class of skewed t link models. Taylor & Francis, 2002. Citado na página 22.

KOMORI, O.; EGUCHI, S.; IKEDA, S.; OKAMURA, H.; ICHINOKAWA, M.; NAKAYAMA, S. An asymmetric logistic regression model for ecological data. **Methods in Ecology and Evolution**, Wiley Online Library, v. 7, n. 2, p. 249–260, 2016. Citado na página 65.

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. Disponível em: <https://doi.org/10.1007/s11749-017-0563-1>. Citado nas páginas 22, 23, 25, 26, 28, 29, 30, 36, 62, 70 e 83.

LEMONTE, A. J.; MORENO-ARENAS, G. Improved estimation for a new class of parametric link functions in binary regression. **Sankhya B**, Springer, v. 82, n. 1, p. 84–110, 2020. Citado na página 65.

LI, D.; WANG, X.; LIN, L.; DEY, D. K. Flexible link functions in nonparametric binary regression with gaussian process priors. **Biometrics**, Wiley Online Library, v. 72, n. 3, p. 707–719, 2016. Citado na página 62.

LI, D.; WANG, X.; SONG, S.; ZHANG, N.; DEY, D. K. Flexible link functions in a joint model of binary and longitudinal data. **Stat**, Wiley Online Library, v. 4, n. 1, p. 320–330, 2015. Citado na página 62.

LUNN, D. J.; THOMAS, A.; BEST, N.; SPIEGELHALTER, D. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. **Statistics and computing**, Springer, v. 10, n. 4, p. 325–337, 2000. Citado nas páginas 43, 66 e 72.

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 41.

LUQUE, A.; CARRASCO, A.; MARTÍN, A.; HERAS, A. de L. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. **Pattern Recognition**, Elsevier, v. 91, p. 216–231, 2019. Citado nas páginas 21 e 36.

MASUDA, M. M.; STONE, R. P. Bayesian logistic mixed-effects modelling of transect data: relating red tree coral presence to habitat characteristics. **ICES Journal of Marine Science**, Oxford University Press, v. 72, n. 9, p. 2674–2683, 2015. Citado nas páginas 22 e 23.

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 na página 34.

METZ, C. E. Basic principles of roc analysis. In: ELSEVIER. **Seminars in nuclear medicine**. [S.l.], 1978. v. 8, n. 4, p. 283–298. Citado na página 37.

MOORS, J. A quantile alternative for kurtosis. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 37, n. 1, p. 25–32, 1988. Citado na página 49.

NARANJO, L.; PÉREZ, C. J.; MARTÍN, J. Skewed link-based regression models for misclassified binary data. **Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas**, Springer, v. 113, n. 2, p. 1585–1599, 2019. Citado na página 22.

NIEKERK, J. van; RUE, H. Skewed probit regression–identifiability, contraction and reformulation. **REVSTAT-Statistical Journal**, Instituto Nacional de Estatistica, v. 19, n. 1, p. 1–23, 2021. Citado na página 81.

NISHIO, M.; ARAKAWA, A. Performance of hamiltonian monte carlo and no-u-turn sampler for estimating genetic parameters and breeding values. **Genetics Selection Evolution**, Springer, v. 51, p. 1–12, 2019. Citado na página 67.

ORDOÑEZ, J. A.; PRATES, M. O.; BAZÁN, J. L.; LACHOS, V. H. Penalized complexity priors for the skewness parameter of power links. **Canadian Journal of Statistics**, Wiley Online Library, 2023. Citado nas páginas 53, 66, 81 e 84.

PAAL, B. Van der. **A comparison of different methods for modelling rare events data**. Dissertação (Mestrado) — Ghent University, 2014. Citado na página 35.

PARAÍBA, C. C. M.; BOCHKINA, N.; DINIZ, C. A. R. Bayesian truncated beta nonlinear mixed-effects models. **Journal of Applied Statistics**, Taylor & Francis, v. 45, n. 2, p. 320–346, 2018. Citado na página 69.

PARZEN, M.; GHOSH, S.; LIPSITZ, S.; SINHA, D.; FITZMAURICE, G. M.; MALLICK, B. K.; IBRAHIM, J. G. A generalized linear mixed model for longitudinal binary data with a marginal logit link function. **The annals of applied statistics**, NIH Public Access, v. 5, n. 1, p. 449, 2011. Citado na página 23.

PRENTICE, R. L. A generalization of the probit and logit methods for dose response curves. **Biometrics**, JSTOR, p. 761–768, 1976. Citado na página 22.

QIU, Z.; LI, H.; SU, H.; OU, G.; WANG, T. Logistic regression bias correction for large scale data with rare events. In: SPRINGER. **Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part II 9**. [S.l.], 2013. p. 133–144. Citado na página 80.

ROSS, S. M. **Simulation**. [S.l.]: Academic Press, 2012. Citado na página 47.

ROSSUM, G. V.; DRAKE, F. **Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009**. 2020. Citado na página 67.

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 38.

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 39.

SCURRAH, K. J.; PALMER, L. J.; BURTON, P. R. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (glmms) and gibbs sampling in bugs. **Genetic Epidemiology**, v. 19, n. 2, p. 127–148, 2000. Citado nas páginas 43, 66 e 72.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information processing & management**, Elsevier, v. 45, n. 4, p. 427–437, 2009. Citado nas páginas 32 e 37.

SPIEGELHALTER, D. J. Bayesian methods for cluster randomized trials with continuous responses. **Statistics in medicine**, Wiley Online Library, v. 20, n. 3, p. 435–452, 2001. Citado nas páginas 43, 66 e 72.

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 40 e 68.

STIRATELLI, R.; LAIRD, N.; WARE, J. H. Random-effects models for serial observations with binary response. **Biometrics**, JSTOR, p. 961–971, 1984. Citado na página 22.

TEAM, S. D. **PyStan: the Python interface to Stan, Version 2.16.0.0.** 2017. Disponível em: <http://mc-stan.org>. Citado na página 44.

TEAM, S. D. *et al.* Stan modeling language users guide and reference manual. **Technical report**, 2016. Citado nas páginas 45, 54, 67 e 68.

THAI-NGHE, N.; GANTNER, Z.; SCHMIDT-THIEME, L. A new evaluation measure for learning from imbalanced data. In: IEEE. **The 2011 International Joint Conference on Neural Networks**. [S.l.], 2011. p. 537–542. Citado na página 22.

THARA, R.; HENRIETTA, M.; JOSEPH, A.; RAJKUMAR, S.; EATON, W. W. Ten-year course of schizophrenia—the madras longitudinal study. **Acta Psychiatrica Scandinavica**, Wiley Online Library, v. 90, n. 5, p. 329–336, 1994. Citado na página 75.

THOMAS, R.; HAVE, T.; KUNSELMAN, A. R.; PULKSTENIS, E. P.; LANDIS, J. R. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. **Biometrics**, JSTOR, p. 367–383, 1998. Citado na página 23.

TING, K. **Encyclopedia of Machine Learning and Data Mining, chap. Confusion Matrix, 260**. [S.l.]: Springer US, Boston, MA, 2017. Citado na página 37.

VANROSSUM, G. Python reference manual. **Department of Computer Science [CS]**, CWI, n. R 9525, 1995. Citado nas páginas 45 e 54.

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 42.

____. 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 41, 42 e 68.

VILLANI, C. Topics in optimal transportation: American mathematical society. **Graduate Studies in Mathematics**, v. 58, 2003. Citado na página 51.

VUJOVIĆ, Ž. *et al.* Classification model evaluation metrics. **International Journal of Advanced Computer Science and Applications**, v. 12, n. 6, p. 599–606, 2021. Citado na página 37.

WANG, X.; DEY, D. K. Generalized extreme value regression for ordinal response data. **Environmental and ecological statistics**, Springer, v. 18, n. 4, p. 619–634, 2011. Citado nas páginas 22 e 36.

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 nas páginas 41 e 68.

WOLFINGER, R.; O'CONNELL, M. Generalized linear mixed models a pseudo-likelihood approach. **Journal of statistical Computation and Simulation**, Taylor & Francis, v. 48, n. 3-4, p. 233–243, 1993. Citado na página 22.

YIN, S.; DEY, D. K.; VALDEZ, E. A.; GAN, G.; VADIVELOO, J. Skewed link regression models for imbalanced binary response with applications to life insurance. **arXiv preprint arXiv:2007.15172**, 2020. Citado nas páginas 22 e 36.

ZHAI, J.; QI, J.; SHEN, C. Binary imbalanced data classification based on diversity oversampling by generative models. **Information Sciences**, Elsevier, v. 585, p. 313–343, 2022. Citado na página 35.

ZOU, Q.; XIE, S.; LIN, Z.; WU, M.; JU, Y. Finding the best classification threshold in imbalanced classification. **Big Data Research**, Elsevier, v. 5, p. 2–8, 2016. Citado nas páginas 38 e 54.

APÊNDICE

A

CAPÍTULO 2

A.1 Resultados de estudo de simulação

Nas figuras 16 e 17, a função de distribuição cumulativa empírica (ECDF) das diferentes métricas é mostrada para as funções PC e logística quando $n = 10000$. Esses resultados são semelhantes aos mostrados para o caso $n = 5000$, observamos que quando o desbalanceamento é de 15% ($\alpha = 3$), as métricas *AUC* e *SPDIF* apresentam comportamentos semelhantes, que ou seja, eles falham em diferenciar o modelo verdadeiro. Por outro lado, quando o desbalanceamento é de 76% ($\alpha = 0,25$), o ECDF de todas as métricas parece muito diferente para ambas as funções de ligação, com exceção da métrica *AUC*.

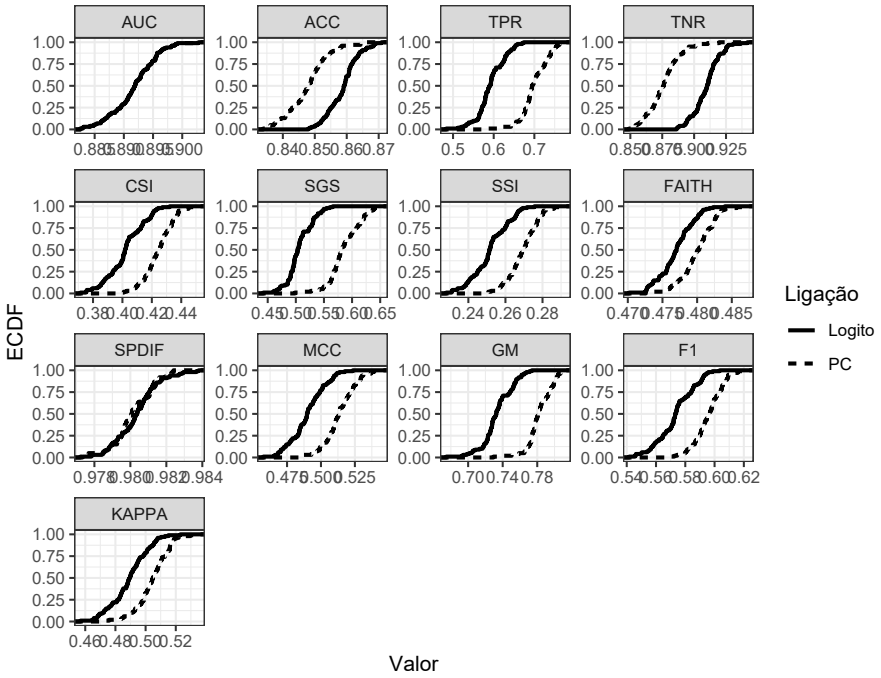

Figura 16 – Função de distribuição cumulativa empírica de métricas para $\alpha = 3$ e $n = 10000$.

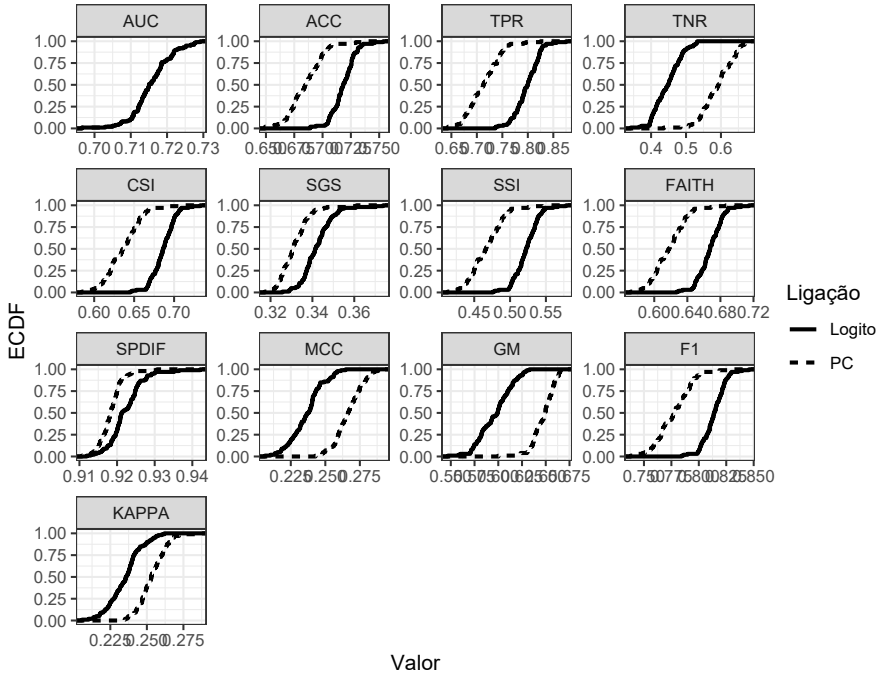

Figura 17 – Função de distribuição cumulativa empírica de métricas para $\alpha = 0,25$ e $n = 10000$.

APÊNDICE

B

CAPÍTULO 3

B.1 Resultados de estudo de simulação

Tabela 20 – Estimativa dos parâmetros usando diferentes funções de ligação para $n = 100$

Ligação	Verdadeiro valor	Média a posteriori				Média a posteriori			
		Estimativa	Viés	Desvio padrão	RMSE	Estimativa	Viés	Desvio padrão	RMSE
L	$\beta_1 = 1,5$	-0,3000	-1,8000	0,1439	1,7999	-0,2979	-1,7979	0,1432	1,7978
	$\beta_2 = -0,5$	-0,9060	-0,4060	0,1363	0,4225	-0,9055	-0,4055	0,1380	0,4225
	$\beta_3 = -0,25$	-0,3632	-0,1132	0,0381	0,1137	-0,3628	-0,1128	0,0382	0,1133
	$\sigma_b^2 = 0,25$	0,4303	0,1803	0,2020	0,2650	0,4142	0,1642	0,2218	0,2702
N	$\beta_1 = 1,5$	-0,1933	-1,6933	0,0860	1,6897	-0,1923	-1,6923	0,0865	1,6887
	$\beta_2 = -0,5$	-0,8786	-0,3786	0,0763	0,3805	-0,8566	-0,3566	0,0761	0,3588
	$\beta_3 = -0,25$	-0,4724	-0,2224	0,0215	0,2176	-0,4671	-0,2171	0,0215	0,2124
	$\sigma_b^2 = 0,25$	0,1561	-0,0939	0,0792	0,1170	0,1464	-0,1036	0,0789	0,1244
C	$\beta_1 = 1,5$	-0,2639	-1,7639	0,1808	1,7674	-0,2601	-1,7601	0,1808	1,7635
	$\beta_2 = -0,5$	-0,8577	-0,3577	0,2037	0,4058	-0,8514	-0,3514	0,2027	0,3999
	$\beta_3 = -0,25$	-0,4733	-0,2233	0,0663	0,2271	-0,4701	-0,2201	0,0658	0,2239
	$\sigma_b^2 = 0,25$	0,5793	0,3293	0,2017	0,3804	0,5853	0,3353	0,2348	0,4035
P-L	$\beta_1 = 1,5$	1,5006	0,0006	1,1030	1,0972	1,5299	0,0299	1,1201	1,1147
	$\beta_2 = -0,5$	-0,6774	-0,1774	0,3569	0,3927	-0,6126	-0,1126	0,3266	0,3397
	$\beta_3 = -0,25$	-0,3538	-0,1038	0,2072	0,2260	-0,3176	-0,0676	0,2123	0,2170
	$\sigma_b^2 = 0,25$	0,2785	0,0285	0,1309	0,1282	0,2349	-0,0151	0,1373	0,1323
	$\alpha = 4$	2,9668	-1,0332	1,0103	1,4393	2,6643	-1,3357	1,0857	1,7155
P-N	$\beta_1 = 1,5$	0,0851	-1,4149	0,3534	1,4526	0,0938	-1,4062	0,4098	1,4589
	$\beta_2 = -0,5$	-0,5904	-0,0904	0,1278	0,1507	-0,5825	-0,0825	0,1389	0,1558
	$\beta_3 = -0,25$	-0,3019	-0,0519	0,0367	0,0578	-0,3045	-0,0545	0,0415	0,0627
	$\sigma_b^2 = 0,25$	0,3044	0,0544	0,1286	0,1339	0,2563	0,0063	0,1364	0,1308
	$\alpha = 4$	0,8394	-3,1606	0,2775	3,1670	0,3638	-3,6362	0,1116	3,6321
P-C	$\beta_1 = 1,5$	1,6470	0,1470	0,5828	0,5952	1,6338	0,1338	0,6311	0,6394
	$\beta_2 = -0,5$	-0,5870	-0,0870	0,1007	0,1273	-0,5749	-0,0749	0,0981	0,1176
	$\beta_3 = -0,25$	-0,2910	-0,0410	0,0374	0,0497	-0,2849	-0,0349	0,0361	0,0444
	$\sigma_b^2 = 0,25$	0,3551	0,1051	0,1657	0,1904	0,3284	0,0784	0,1772	0,1879
	$\alpha = 4$	4,3246	0,3246	1,1847	1,2226	4,2211	0,2211	1,0815	1,0981
RP-L	$\beta_1 = 1,5$	2,3155	0,8155	1,0072	1,2901	2,2906	0,7906	1,0719	1,3261
	$\beta_2 = -0,5$	-0,7885	-0,2885	0,2766	0,3939	-0,7735	-0,2735	0,2809	0,3863
	$\beta_3 = -0,25$	-0,2189	0,0311	0,0880	0,0876	-0,2151	0,0349	0,0931	0,0936
	$\sigma_b^2 = 0,25$	0,2329	-0,0171	0,1808	0,1758	0,2318	-0,0182	0,1296	0,1251
	$\alpha = 4$	0,5354	-3,4646	0,4039	3,4822	0,3458	-3,6542	0,2559	3,6574
RP-N	$\beta_1 = 1,5$	1,1767	-0,3233	0,2888	0,4277	1,3313	-0,1687	0,3038	0,3417
	$\beta_2 = -0,5$	-0,6127	-0,1127	0,1237	0,1616	-0,6136	-0,1136	0,1273	0,1648
	$\beta_3 = -0,25$	-0,3076	-0,0576	0,0375	0,0630	-0,3122	-0,0622	0,0396	0,0679
	$\sigma_b^2 = 0,25$	0,3463	0,0963	0,1600	0,1810	0,3207	0,0707	0,1751	0,1831
	$\alpha = 4$	0,4311	-3,5689	0,1667	3,5670	0,2360	-3,7640	0,0534	3,7586
RP-C	$\beta_1 = 1,5$	2,4206	0,9206	1,2103	1,5148	2,3129	0,8129	1,1244	1,3817
	$\beta_2 = -0,5$	-0,3210	0,1790	0,3675	0,4030	-0,3740	0,1260	0,3476	0,3639
	$\beta_3 = -0,25$	-0,2277	0,0223	0,1441	0,1401	-0,2233	0,0267	0,1374	0,1341
	$\sigma_b^2 = 0,25$	0,2335	-0,0165	0,1080	0,1035	0,2386	-0,0114	0,1076	0,1024
	$\alpha = 4$	0,3436	-3,6564	0,1862	3,6553	0,3139	-3,6861	0,1033	3,6817

Tabela 21 – Estimativa dos parâmetros usando diferentes funções de ligação para $n = 250$

Ligação	Verdadeiro valor	Média a posteriori				Média a posteriori			
		Estimativa	Viés	Desvio padrão	RMSE	Estimativa	Viés	Desvio padrão	RMSE
L	$\beta_1 = 1,5$	-0,2948	-1,7948	0,0985	1,7917	-0,2945	-1,7945	0,0987	1,7914
	$\beta_2 = -0,5$	-0,8944	-0,3944	0,0733	0,3953	-0,8937	-0,3937	0,0730	0,3946
	$\beta_3 = -0,25$	-0,3634	-0,1134	0,0231	0,1099	-0,3630	-0,1130	0,0229	0,1095
	$\sigma_b^2 = 0,25$	0,4422	0,1922	0,1082	0,2147	0,4334	0,1834	0,1075	0,2067
N	$\beta_1 = 1,5$	-0,1870	-1,6870	0,0560	1,6821	-0,1872	-1,6872	0,0560	1,6823
	$\beta_2 = -0,5$	-0,8599	-0,3599	0,0417	0,3565	-0,8694	-0,3694	0,0416	0,3660
	$\beta_3 = -0,25$	-0,4475	-0,1975	0,0123	0,1921	-0,4473	-0,1973	0,0123	0,1919
	$\sigma_b^2 = 0,25$	0,1437	-0,1063	0,0348	0,1061	0,1409	-0,1091	0,0341	0,1085
C	$\beta_1 = 1,5$	-0,2770	-1,7770	0,1286	1,7759	-0,2755	-1,7755	0,1289	1,7744
	$\beta_2 = -0,5$	-0,8341	-0,3341	0,1044	0,3442	-0,8315	-0,3315	0,1045	0,3418
	$\beta_3 = -0,25$	-0,4641	-0,2141	0,0385	0,2117	-0,4634	-0,2134	0,0386	0,2110
	$\sigma_b^2 = 0,25$	0,7136	0,4636	0,1158	0,4721	0,7265	0,4765	0,1287	0,4878
P-L	$\beta_1 = 1,5$	1,9540	0,4540	0,1742	0,4804	1,9365	0,4365	0,1743	0,4642
	$\beta_2 = -0,5$	-0,4530	0,0470	0,0498	0,0626	-0,4472	0,0528	0,0482	0,0657
	$\beta_3 = -0,25$	-0,2311	0,0189	0,0169	0,0196	-0,2270	0,0230	0,0154	0,0219
	$\sigma_b^2 = 0,25$	0,1887	-0,0613	0,0461	0,0709	0,1792	-0,0708	0,0414	0,0763
	$\alpha = 4$	4,7367	0,7367	0,5266	0,8997	4,7550	0,7550	0,6712	1,0044
P-N	$\beta_1 = 1,5$	-0,1783	-1,6783	0,3905	1,7174	0,1114	-1,3886	0,4987	1,4697
	$\beta_2 = -0,5$	-0,4154	0,0846	0,0626	0,0994	-0,3611	0,1389	0,0686	0,1491
	$\beta_3 = -0,25$	-0,2168	0,0332	0,0336	0,0414	-0,1877	0,0623	0,0391	0,0678
	$\sigma_b^2 = 0,25$	0,1858	-0,0642	0,0706	0,0897	0,1199	-0,1301	0,0596	0,1373
	$\alpha = 4$	2,3019	-1,6981	0,6597	1,8160	1,6874	-2,3126	0,8308	2,4515
P-C	$\beta_1 = 1,5$	1,8092	0,3092	0,3724	0,4782	1,7925	0,2925	0,4114	0,4990
	$\beta_2 = -0,5$	-0,5645	-0,0645	0,0730	0,0916	-0,5575	-0,0575	0,0720	0,0864
	$\beta_3 = -0,25$	-0,2794	-0,0294	0,0228	0,0315	-0,2763	-0,0263	0,0230	0,0291
	$\sigma_b^2 = 0,25$	0,3198	0,0698	0,0862	0,1051	0,2997	0,0497	0,0803	0,0886
	$\alpha = 4$	4,3004	0,3004	0,8654	0,9103	4,2215	0,2215	0,7037	0,7319
RP-L	$\beta_1 = 1,5$	2,6234	1,1234	0,5844	1,2605	2,6400	1,1400	0,6042	1,2844
	$\beta_2 = -0,5$	-0,7191	-0,2191	0,1604	0,2658	-0,7004	-0,2004	0,1611	0,2513
	$\beta_3 = -0,25$	-0,2821	-0,0321	0,0566	0,0593	-0,2211	0,0289	0,0578	0,0589
	$\sigma_b^2 = 0,25$	0,3336	0,0836	0,0608	0,0975	0,3327	0,0827	0,0630	0,0982
	$\alpha = 4$	0,2384	-3,7616	0,0718	3,7565	0,2202	-3,7798	0,0574	3,7745
RP-N	$\beta_1 = 1,5$	1,4999	-0,0001	0,1648	0,1590	1,6336	0,1336	0,1529	0,1972
	$\beta_2 = -0,5$	-0,6480	-0,1480	0,0716	0,1586	-0,6534	-0,1534	0,0727	0,1640
	$\beta_3 = -0,25$	-0,3324	-0,0824	0,0200	0,0790	-0,3391	-0,0891	0,0198	0,0855
	$\sigma_b^2 = 0,25$	0,3814	0,1314	0,0842	0,1503	0,3714	0,1214	0,0858	0,1429
	$\alpha = 4$	0,2551	-3,7449	0,0528	3,7395	0,1883	-3,8117	0,0175	3,8059
RP-C	$\beta_1 = 1,5$	2,4832	0,9832	0,4647	1,0817	2,4357	0,9357	0,4533	1,0339
	$\beta_2 = -0,5$	-0,3063	0,1937	0,1940	0,2684	-0,2937	0,2063	0,1909	0,2753
	$\beta_3 = -0,25$	-0,3769	-0,1269	0,0696	0,1390	-0,3704	-0,1204	0,0684	0,1327
	$\sigma_b^2 = 0,25$	0,3011	0,0511	0,0443	0,0619	0,3010	0,0510	0,0421	0,0603
	$\alpha = 4$	0,2824	-3,7176	0,0356	3,7120	0,2799	-3,7201	0,0350	3,7144

Tabela 22 – Estimativa dos parâmetros usando diferentes funções de ligação para $n = 500$

Ligação	Verdadeiro valor	Média a posteriori				Média a posteriori			
		Estimativa	Viés	Desvio padrão	RMSE	Estimativa	Viés	Desvio padrão	RMSE
L	$\beta_1 = 1,5$	-0,290	-1,790	0,062	1,786	-0,291	-1,791	0,062	1,786
	$\beta_2 = -0,5$	-0,941	-0,441	0,042	0,438	-0,951	-0,451	0,042	0,447
	$\beta_3 = -0,25$	-0,363	-0,113	0,016	0,109	-0,363	-0,113	0,016	0,108
	$\sigma_b^2 = 0,25$	0,439	0,189	0,100	0,208	0,434	0,184	0,100	0,204
N	$\beta_1 = 1,5$	-0,185	-1,685	0,036	1,680	-0,185	-1,685	0,036	1,680
	$\beta_2 = -0,5$	-0,862	-0,362	0,024	0,357	-0,862	-0,362	0,024	0,357
	$\beta_3 = -0,25$	-0,449	-0,199	0,009	0,193	-0,447	-0,197	0,009	0,192
	$\sigma_b^2 = 0,25$	0,144	-0,106	0,033	0,105	0,142	-0,108	0,033	0,107
C	$\beta_1 = 1,5$	-0,269	-1,769	0,083	1,765	-0,268	-1,768	0,083	1,764
	$\beta_2 = -0,5$	-0,837	-0,337	0,060	0,336	-0,836	-0,336	0,060	0,335
	$\beta_3 = -0,25$	-0,469	-0,219	0,031	0,216	-0,469	-0,219	0,031	0,215
	$\sigma_b^2 = 0,25$	0,740	0,490	0,118	0,498	0,746	0,496	0,128	0,506
P-L	$\beta_1 = 1,5$	1,755	0,255	0,104	0,270	1,820	0,320	0,101	0,330
	$\beta_2 = -0,5$	-0,425	0,075	0,026	0,074	-0,423	0,077	0,026	0,076
	$\beta_3 = -0,25$	-0,221	0,029	0,010	0,025	-0,220	0,030	0,009	0,026
	$\sigma_b^2 = 0,25$	0,174	-0,076	0,043	0,081	0,170	-0,080	0,042	0,084
	$\alpha = 4$	5,453	1,453	0,339	1,487	5,602	1,602	0,411	1,648
P-N	$\beta_1 = 1,5$	0,591	-0,909	0,296	0,950	0,797	-0,703	0,238	0,736
	$\beta_2 = -0,5$	-0,299	0,201	0,038	0,199	-0,270	0,230	0,028	0,226
	$\beta_3 = -0,25$	-0,158	0,092	0,022	0,089	-0,142	0,108	0,015	0,104
	$\sigma_b^2 = 0,25$	0,094	-0,156	0,036	0,154	0,069	-0,181	0,021	0,176
	$\alpha = 4$	3,826	-0,174	0,686	0,702	3,762	-0,238	0,976	0,999
P-C	$\beta_1 = 1,5$	1,691	0,191	0,302	0,352	1,646	0,146	0,304	0,331
	$\beta_2 = -0,5$	-0,524	-0,024	0,059	0,058	-0,517	-0,017	0,058	0,054
	$\beta_3 = -0,25$	-0,267	-0,017	0,021	0,021	-0,263	-0,013	0,020	0,018
	$\sigma_b^2 = 0,25$	0,296	0,046	0,092	0,097	0,282	0,032	0,088	0,088
	$\alpha = 4$	4,302	0,302	0,688	0,745	4,211	0,211	0,688	0,714
RP-L	$\beta_1 = 1,5$	1,848	0,348	0,507	0,609	1,652	0,152	0,524	0,540
	$\beta_2 = -0,5$	-0,700	-0,200	0,118	0,227	-0,713	-0,213	0,121	0,239
	$\beta_3 = -0,25$	-0,217	0,033	0,051	0,055	-0,222	0,028	0,052	0,053
	$\sigma_b^2 = 0,25$	0,335	0,085	0,038	0,087	0,330	0,080	0,036	0,082
	$\alpha = 4$	0,223	-3,777	0,042	3,771	0,217	-3,783	0,041	3,778
RP-N	$\beta_1 = 1,5$	1,720	0,220	0,112	0,241	1,816	0,316	0,104	0,326
	$\beta_2 = -0,5$	-0,665	-0,165	0,046	0,166	-0,670	-0,170	0,046	0,170
	$\beta_3 = -0,25$	-0,348	-0,098	0,015	0,094	-0,353	-0,103	0,015	0,098
	$\sigma_b^2 = 0,25$	0,424	0,174	0,097	0,194	0,420	0,170	0,100	0,191
	$\alpha = 4$	0,190	-3,810	0,017	3,805	0,166	-3,834	0,007	3,828
RP-C	$\beta_1 = 1,5$	2,256	0,756	0,339	0,822	2,238	0,738	0,334	0,804
	$\beta_2 = -0,5$	-0,300	0,200	0,112	0,224	-0,295	0,205	0,111	0,227
	$\beta_3 = -0,25$	-0,338	-0,088	0,048	0,094	-0,336	-0,086	0,047	0,092
	$\sigma_b^2 = 0,25$	0,308	0,058	0,024	0,057	0,300	0,050	0,020	0,048
	$\alpha = 4$	0,294	-3,706	0,026	3,700	0,293	-3,707	0,026	3,702

B.2 Codigo em Python

```
"""
@author: Alex de la Cruz Huayanay
"""
import pandas as pd
import numpy as np
import pystan
#dataset
url = 'https://raw.githubusercontent.com/aldehu/datas/dat/BData4.csv'
data0 = pd.read_csv(url, error_bad_lines=False)
data = data0[['id','age','gender','month','resp']]
data.info()
#Dataset split
n=data['id'].value_counts().count()
p=data.shape[1]-2
N=data.shape[0]
ids=np.array(data['id'])
#x2=data['age']
x2=data['gender']
x3=data['month']
y=data['resp']
#Data for Stan
dataS = {}
dataS['id']  = ids
dataS['y']   = y
dataS['x2']  = x2
dataS['x3']  = x3
dataS['p']   = p
dataS['n']   = n
dataS['N']  = N
#Modelo Power Normal
model_pn = '''
data {
   int<lower = 0> p;
   int<lower = 0> N;
   int<lower = 0> n;
   int<lower=0, upper=1> y[N];
   real x2[N];
   real x3[N];
   int<lower = 0> id[N];
}
parameters{
   vector[p] beta;
   real<lower = 0> tau2b;
   real loglambda;
   real bib[n];
}
```

```
47: transformed parameters{
48:   real prob[N];
49:   real<lower = 0> lambda;
50:   real<lower = 0> sigma2b;
51:   lambda <- exp(loglambda);
52:   sigma2b <- pow(tau2b,-1);
53:   for(i in 1:N){
54:   prob[i] <- pow(Phi(beta[1]+beta[2]*x2[i]+beta[3]*x3[i]
55:   +bib[id[i]]),lambda);
56:   }
57: }
58: model {
59:   for(j in 1:n){
60:   bib[j] ~ normal(0, sqrt(1/tau2b));
61:   }
62:   beta ~ normal(0.0,100);
63:   loglambda ~ uniform(-2,2);
64:   tau2b ~ pareto(1,0.001);
65:   y ~ bernoulli(prob);
66: }
67: '''
68: chains = 2
69: iters = 15000
70: warmup = 10000
71: thin = 5
72: seed = 10000003
73: sm = pystan.StanModel(model_code=model_pn)
74: fit = sm.sampling(data=dataS,
75:                    chains=chains, iter=iters,
76:                    warmup=warmup,thin=thin,seed=seed)
77: mpn_summary = pd.DataFrame(fit.summary()['summary'],
78:                            columns=fit.summary()['summary_colnames'],
79:                            index=fit.summary()['summary_rownames'])
80: mpn_1=mpn_summary.loc[['beta[1]','beta[2]','beta[3]','sigma2b','lambda']]
81: mpn_1
```

B.3 Distribuição a priori do parâmetro de forma

Proposição 2. Se $\delta = \log(\alpha) \sim U(a,b)$ então $\alpha \sim$ log-uniform(e^a, e^b) com $E(\alpha) = \frac{e^b - e^a}{b-a}$ e $Var(\alpha) = \frac{e^{2b}-e^{2a}}{2(b-a)} - \frac{(e^b-e^a)^2}{(b-a)^2}$.

Demonstração. Note que $\alpha = \exp(\delta)$, então $F_\alpha(x) = P[\alpha \leq x] = P[\exp(\delta) \leq x] = P[\delta \leq \log(x)] = F_\delta(\log(x)) = (\log(x)+2)/4$, $x \in [e^{-2}, e^2]$. Além disso, $f_\alpha(x) = \frac{d}{dx}F_\delta(\log(x)) = \frac{1}{4x} = 1/x[\log(e^{-2}) - \log(e^2)]$. Então, $\alpha \sim$ log-uniform$(a = e^{-2}, b = e^2)$ □

B.4 Explorando versões padronizadas das funções de ligação potência

Nesta seção, exploramos o uso das versões padronizadas de funções de ligação aqui estudadas. Para referência, usaremos a versão padronizada da função de ligação PL. A seguir, apresentaremos primeiro a distribuição logística generalizada e sua versão padronizada.

B.4.1 Distribuição logística generalizada e sua versão padronizada

Conforme apresentado em Johnson, Kotz e Balakrishnan (1995), diz-se que uma variável aleatória X tem uma distribuição logística generalizada tipo I ou distribuição potência logística denotada por $PL(\mu,\sigma,\alpha)$, com parâmetros de localização, escala e de forma dados por $\mu \in \mathbb{R}$, $\sigma > 0$ e $\alpha > 0$, respectivamente, se sua fda tiver a seguinte forma

$$F(x \mid \mu,\sigma,\alpha) = \left[\frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)}\right]^{\alpha}. \tag{B.1}$$

Da mesma forma, diz-se que uma variável aleatória Y tem uma distribuição logística generalizada tipo II ou distribuição reversa de potência logística denotada por $RPL(\mu,\sigma,\alpha)$, com parâmetros de localização, escala e de forma dados por $\mu \in \mathbb{R}$, $\sigma > 0$ e $\alpha > 0$, respectivamente, se sua fda tiver a seguinte forma

$$H(y \mid \mu,\sigma,\alpha) = 1 - \left[\frac{\exp\left\{\frac{-(y-\mu)}{\sigma}\right\}}{1+\exp\left\{\frac{-(y-\mu)}{\sigma}\right\}}\right]^{\alpha}. \tag{B.2}$$

A esperança e variância das variáveis $X \sim PL(\mu,\sigma,\alpha)$ e $Y \sim RPL(\mu,\sigma,\alpha)$ são dadas respectivamente por

- $E(X) = \mu + \sigma[\psi(\alpha) - \psi(1)]$ e $Var(X) = \sigma^2(\psi'(\alpha) + \psi'(1))$.
- $E(Y) = \mu - \sigma[\psi(\alpha) - \psi(1)]$ e $Var(Y) = \sigma^2(\psi'(\alpha) + \psi'(1))$.

em que $\psi(\cdot)$ e $\psi'(\cdot)$ são as funções digamma e trigamma, respectivamente.

Proposição 3. a) Se $X \sim PL(\mu,\sigma,\alpha)$ então $Y = -X + 2\mu \sim RPL(\mu,\sigma,\alpha)$. Adicionalmente b) Se $X \sim RPL(\mu,\sigma,\alpha)$ então $Y = -X + 2\mu \sim PL(\mu,\sigma,\alpha)$

Demonstração. No caso a) considere $\mathbb{P}(Y \leq y) = \mathbb{P}(-X + 2\mu \leq y) = 1 - \mathbb{P}(X \leq -y + 2\mu) = 1 - F(-y+2\mu \mid \mu,\sigma,\alpha) = 1 - \left[\frac{\exp\left(\frac{-(y-\mu)}{\sigma}\right)}{1+\exp\left(\frac{-(y-\mu)}{\sigma}\right)}\right]^{\alpha} = H(y \mid \mu,\sigma,\alpha)$. A prova de b) é análoga. □

Proposição 4. A potência logística padrão e a distribuição reversa de potência logística padrão mostradas em Seção 4.1, são obtidos respectivamente, considerando a transformação $Z = \frac{X-\mu}{\sigma}$ e $W = \frac{Y-\mu}{\sigma}$, em que $X \sim PL(\mu, \sigma, \alpha)$ e $Y \sim RPL(\mu, \sigma, \alpha)$. Neste caso temos $Z \sim PL(0,\ 1,\ \alpha)$ e $W \sim RPL(0,\ 1,\ \alpha)$.

Demonstração. Em (B.1) temos que $F(z \mid 0,\ 1,\ \alpha) = \left[\frac{\exp\{z\}}{1+\exp\{z\}}\right]^{\alpha}$ e em (B.2) $H(w \mid 0,\ 1,\ \alpha) = 1 - \left[\frac{\exp\{-w\}}{1+\exp\{-w\}}\right]^{\alpha}$. □

Neste caso, a esperança e a variância são dadas, respectivamente, por

- $E(Z) = \psi(\alpha) - \psi(1)$ e $Var(Z) = \psi'(\alpha) + \psi'(1)$.
- $E(W) = -[\psi(\alpha) - \psi(1)]$ e $Var(W) = \psi'(\alpha) + \psi'(1)$.

B.4.2 Modelo de regressão binária com função de ligação PL padronizada

Nesta subseção, introduzimos uma nova representação do modelo proposto em (4.1).

Proposição 5. O modelo de regressão binária com a função de ligação PL proposta em (4.1), pode ser obtido considerando a seguinte estrutura

$$Y_i = \begin{cases} 1 & \text{se } S_i > 0 \\ 0 & \text{se } S_i \leq 0 \end{cases} \tag{B.3}$$

em que $S_i = \boldsymbol{x_i}^{\top}\boldsymbol{\beta} + \varepsilon_i$ e $\varepsilon_i \sim RPL(0,\ 1,\ \alpha)$ é um erro latente associado à definição da resposta binária.

Demonstração. Observe que $\mathbb{P}(Y_i = 1) = \mathbb{P}\left(\boldsymbol{x_i}^{\top}\boldsymbol{\beta} + \varepsilon_i > 0\right) = 1 - \mathbb{P}\left(\varepsilon_i \leq -\boldsymbol{x_i}^{\top}\boldsymbol{\beta}\right) = 1 - H(-\boldsymbol{x_i}^{\top}\boldsymbol{\beta} \mid \alpha) = F(\boldsymbol{x_i}^{\top}\boldsymbol{\beta} \mid \alpha) = \mu_i$, como é esperado. Analogamente, obtemos a $\mathbb{P}(Y_i = 0) = 1 - \mathbb{P}(Y_i = 1) = 1 - F(\boldsymbol{x_i}^{\top}\boldsymbol{\beta} \mid \alpha) = 1 - \mu_i$. □

Note que $E(\varepsilon_i) = -[\psi(\alpha) - \psi(1)]$ e $Var(W) = \psi'(\alpha) + \psi'(1)$. Portanto, a distribuição de erro latente subjacente ε_i não é padronizada e, consequentemente, a função de ligação baseada nessa função de densidade não é padronizada.
Para usar uma versão padronizada da função de ligação PL, consideramos a seguinte estrutura

$$Y_i = \begin{cases} 1 & \text{se } S_i > 0 \\ 0 & \text{se } S_i \leq 0 \end{cases} \tag{B.4}$$

em que $S_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \varepsilon_i^\star$ e $\varepsilon_i^\star \sim RPL(\mu^\star, \sigma^\star, \alpha)$, em que $\mu^\star = \frac{\psi(\alpha)-\psi(1)}{\sqrt{\psi'(\alpha)+\psi'(1)}}$ e $\sigma^\star = \frac{1}{\sqrt{\psi'(\alpha)+\psi'(1)}}$, então teríamos uma padronização $\varepsilon_i^\star$ com $E(\varepsilon_i^\star) = 0$ e $Var(\varepsilon_i^\star) = 1$. Assim, um modelo alternativo de regressão binária considerando versões padronizadas das funções de ligação PL pode ser obtido considerando ε_i padronizadas e é dada por

$$\mu_i = \mathbb{P}(Y_i = 1) = F(\boldsymbol{x_i}^\top \boldsymbol{\beta} \mid \mu^\star, \sigma^\star, \alpha)$$

Este tipo de função de ligação pode ser proposto para os diferentes funções de ligação aqui estudadas. Chamaremos essa função de ligação SPL. Na seção seguinte, estudamos a viabilidade desta função de ligação proposta considerando um breve estudo de simulação e uma aplicação.

B.4.3 Estudo de simulação

Geramos um conjunto de dados da variável $Y_i \sim \text{Bernoulli}(\mu_i)$ considerando dois casos possíveis:

- Caso 1: usando a função de ligação padronizada SPL $\mu_i = \left[\frac{\exp\left(\frac{\beta_1+\beta_2 x_i-\mu^\star}{\sigma^\star}\right)}{1+\exp\left(\frac{\beta_1+\beta_2 x_i-\mu^\star}{\sigma^\star}\right)}\right]^\alpha$, em que $\mu^\star = \frac{\psi(\alpha)-\psi(1)}{\sqrt{\psi'(\alpha)+\psi'(1)}}$ e $\sigma^\star = \frac{1}{\sqrt{\psi'(\alpha)+\psi'(1)}}$. Neste caso, a proporção de sucesso foi $0,22$
- Caso 2: usando a função de ligação PL $\mu_i = \left[\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right]^\alpha$. Neste caso, a proporção de sucesso foi $0,30$

A covariável x_i foi gerada a partir de uma distribuição normal $N(0,\ 1)$, os coeficientes de regressão foram fixados com valores $\boldsymbol{\beta} = (\beta_1,\ \beta_2)^\top = (0,25,\ 1,0)^\top$ e $\alpha = 2$. Com essas especificações, os dados foram gerados considerando o tamanho da amostra $N = 2500$ e 100 replicações em cada caso. Para esses dados ajustamos a função de ligação logística, PL e SPL. Os resultados são mostrados em Tabela 23

Tabela 23 – Estimativa de parâmetros usando diferentes ligações para dados gerados com funções de ligação SPL e PL

		Caso 1: SPL				Caso 2: PL			
Ligação	Verdadeiro valor	Estimativa	DP	RMSE	Tempo Média (DP)	Estimativa	DP	RMSE	Tempo Média (DP)
L	$\beta_1 = 0,25$	-2,012	0,073	2,262	1,540	-0,835	0,050	1,085	0,346
	$\beta_2 = 1$	1,982	0,091	0,982	(0,222)	1,270	0,059	0,270	(0,0100)
PL(0, 1, α)	$\beta_1 = 0,25$	0,684	0,668	0,934	2,034	0,301	0,570	0,051	1,608
	$\beta_2 = 1$	1,057	0,117	0,573	(0,270)	1,041	0,126	0,041	(0,221)
	$\alpha = 2$	2,419	0,964	0,419		2,689	0,951	0,689	
PL(μ^*, σ^*, α)	$\beta_1 = 0,25$	0,412	0,723	0,162	2,715	1,176	0,692	0,926	3,689
	$\beta_2 = 1$	0,997	0,066	0,003	(0,330)	0,662	0,035	0,338	(0,698)
	$\alpha = 2$	2,532	0,964	0,532		2,876	0,953	0,876	

Os resultados mostram que se os dados forem gerados considerando o caso 1, a recuperação dos parâmetros do modelo usando a função de ligação SPL é adequada, enquanto a

recuperação dos parâmetros do modelo com a função de ligação PL é razoável. Por outro lado, se os dados forem gerados considerando o caso 2, os resultados dos modelos com a função de ligação SPL e PL são semelhantes.
Isso prova que ambos os modelos podem ser usados indistintamente desde que não conheçamos o modelo verdadeiro para os dados.

B.4.4 Aplicação

Para ilustrar a aplicação da função de ligação padronizada estudada, usaremos o conjunto de dados conhecido como mortalidade de besouros (COLLETT, 2002), onde é estabelecido que o logaritmo das diferenças na concentração de veneno (x), explica a proporção de besouros adultos mortos ($Y = 1$, morto e $Y = 0$, não morto). Foram observados 481 besouros.
Para este conjunto de dados, os modelos foram ajustados considerando as seguintes 3 funções de ligação: logística, PL e SPL. Os resultados são mostrados em Tabela 24

Tabela 24 – Critérios de seleção de modelos das funções de ligação logística generalizada para dados de besouros

Ligação	DIC	WAIC	LOO	EAIC	EBIC	IC
L	376,807	376,849	376,850	378,631	386,983	378,983
PL	369,968	370,118	370,117	373,382	385,910	372,553
SPL	370,207	370,364	370,370	373,513	386,041	372,900

Tabela 25 – Estimativas de parâmetros dos modelos com diferentes ligações para dados de besouros

Ligação	Parâmetro	Estimativa	Desvio padrão	95% I.C.	
L	β_1	-2,792	0,317	(-3,440;	-2,177)
	β_2	6,653	0,619	(5,485;	7,941)
PL	β_1	-9,776	2,798	(-15,132;	-4,770)
	β_2	14,299	3,361	(8,428;	20,920)
	α	0,255	0,102	(0,140;	0,522)
SPL	β_1	-2,775	0,297	(-3,23;	-2,057)
	β_2	3,200	0,381	(2,519;	3,941)
	α	0,332	0,151	(0,148;	0,691)

Os resultados em Tabela 24 mostram que o modelo de regressão binária com função de ligação PL e SPL tem um melhor desempenho considerando os critérios de comparação de modelo em comparação com o link logístico.
Por outro lado, em Tabela 25 mostramos os resultados das estimativas dos parâmetros para cada função ligação estudada nesta seção. Observe que todos os parâmetros são indicados como significativamente diferentes de 0 com um nível de credibilidade de 95%. Assim, vemos preliminarmente uma vantagem significativa de usar a função de ligação padronizada em comparação

com a função de ligação não padronizada. Resultados adicionais poderão ser discutidos em um trabalho específico.


ICMC
USP
SÃO CARLOS