MODELAGEM BAYESIANO DE REGRESSÃO BINÁRIA PARA DADOS DESBALANCEADOS USANDO NOVAS LIGAÇÕES

Andson Nunes da SILVA[1]
Susan ANYOSA[2]
Jorge Luis BAZÁN[3]

- RESUMO: Neste trabalho apresentamos, de forma didática, a modelagem bayesiana de regressão binária para dados desbalanceados usando novas ligações. Sob abordagem bayesiana e usando critérios de informação, medidas de avaliação preditiva e introduzindo a análise de resíduos, mostramos que os modelos que utilizam funções de ligação potência e reversa de potência se ajustam melhor do que os modelos tradicionais na presença de dados desbalanceados, considerando duas aplicações. Adicionalmente, códigos com os procedimentos apresentados usando o pacote `Stan` são disponibilizados de modo a facilitar o uso destes modelos. O trabalho também contém um estudo de simulacão que mostra como o desbalanceamento na variável resposta afeta a estimação dos parâmetros de uma regressão logística com relação ao vício, erro quadrático médio e o desvio padrão das estimativas, independente do tamanho da amostra. Ao mesmo tempo, considerando duas aplicações, mostramos como modelos de regressão binária com as ligações potência e reversa de potência recentemente formulados na literatura podem ser usados para estimar adequadamente os parâmetros no tipo de desbalanceamento considerado.

- PALAVRAS-CHAVE: Dados desbalanceados; ligações assimétricas; estimação bayesiana; regressão binária; resíduos.

1 Introdução

Quando, na análise de dados, tentamos explicar uma variável de resposta de interesse que toma somente dois valores considerando um conjunto de preditores,

[1]Universidade de São Paulo - USP, ICMC, Matemática Aplicada e Estatística, Caixa Postal, CEP: 14811-230, Araraquara, SP, Brasil. E-mail: *andsonunes@gmail.com*
[2]Norwegian University of Science and Technology, Sentralbygg 2, 1056, Gløshaugen, Alfred Getz vei 1, 7491 Trondheim, Noruega. E-mail: *susan.anyosa@ntnu.no*
[3]Universidade de São Paulo - USP, ICMC, Matemática Aplicada e Estatística, Caixa Postal, CEP: 13560-970, São Carlos, SP, Brasil. E-mail: *jlbazan@icmc.usp.br*

o modelo é conhecido como *modelo de regressão binária*, apresentado formalmente na definição 2 abaixo. Este é um modelo geralmente abordado no contexto de modelos lineares generalizados, como em Agresti (2013), mas, também, é visto como um modelo de aprendizado de máquina supervisionado para classificação binária Hosmer e Lemeshow (1989). Um dos problemas, que é bastante comum, em modelos de classificação ou de regressão binária é o desbalanceamento da variável resposta binária, apresentado formalmente na definição 1 abaixo. Esse fenômeno ocorre quando as classes ou categorias de resposta não são igualmente distribuídas. Alguns exemplos de situações nas quais encontramos o fenômeno são: detecção de fraude ou detecção de spam, classificação em *credit scoring*, uso de um seguro, reconhecimento de um som como sendo nasal ou oral em dados linguísticos e classificação do nível de desempenho em matemática de dados educacionais.

Alguns autores como King e Zeng (2001) denominam o desbalanceamento como dados de eventos raros, e sinalizam que a categoria de sucesso (evento de interesse) da variável dependente binária ocorre, de dezenas a milhares de vezes, menos do que os fracassos (evento que não é de interesse).

Por outro lado, Paal (2014) relata que há dois tipos de eventos raros: a) *eventos com raridade relativa*, também chamado de dados desequilibrados ou não balanceados. Neste caso o conjunto de dados é considerado desequilibrado quando a classe minoritária ou classe de interesse é muito menor do que a classe majoritária; b) *eventos com raridade absoluta* o qual é considerada um problema de amostra pequena. Alguns autores como Allison (2012) indicam que a regressão logística, uns dos modelos de regressão binária mais conhecido atualmente, é afetado quando temos um evento de raridade absoluta, pois as estimativas dos parâmetros gerados pela a regressão logística acabam sendo calculados de tal forma a causar um viés na variável resposta, e o grau de viés é fortemente dependente do número de casos na categoria de menor frequência da variável resposta.

No entanto Agresti (2013) sugere que se observe a raridade relativa em relação ao número de preditores, isto é, quando há influência das covariáveis na raridade relativa. Assim, não levar em consideração o desbalanceamento entre as classes pode causar um impacto negativo na performance do modelo preditivo de la Cruz et al. (2019).

Neste trabalho vamos focar em dados binários não balanceados seguindo a definição de raridade relativa citada por Paal (2014).

Alguns métodos, na literatura, foram desenvolvidos para trabalhar com dados desbalanceados utilizando em regressão binária. Para este estudo iremos avaliar as funções de ligação propostas por Bazán, Torres-Aviles, Suzuki e Louzada (2017) e Lemonte e Bazán (2018) para os modelos de regressão binária na presença de dados não balanceados. Iremos considerar um parâmetro extra para poder explicar a assimetria das curvas de resposta e, especificamente, o desbalanceamento dos dados. A inferência desenvolvida para a estimação dos parâmetros destes modelos será sob abordagem bayesiana. Adicionalmente nos estudaremos o uso de critérios de comparação de modelos, medidas de avaliação preditiva e introduziremos a análise de resíduos com o intuito de escolher o melhor modelo para um conjunto

de dados. Além disso incluiremos duas aplicações relevantes e adicionaremos os códigos utilizados em *Stan* (STAN, 2017).

Este trabalho esta organizado da seguinte forma. Na seção 2, introduzimos os preliminares para a proposta deste trabalho. Na seção 3, é apresentado a modelagem sob enfoque bayesiano dos modelos de regressão binária usando as funções de ligação potência e reversa de potência. Já na seção 4 será apresentado um estudo sobre desbalanceamento em dados binários, em que incluímos um processo de simulação para verificar o efeito dos parâmetros na presença de dados não balanceados. Na seção 5 são apresentados duas aplicações a primeira relacionada a base de dados na área da saúde e a segunda relacionada a dados educacionais. Na seção 6 é apresentado a conclusão do trabalho realizado. Por último, no apêndice, são apresentados códigos nas linguagens `R` e `Stan` usados na aplicação.

2 Preliminares

O trabalho em questão foca em dados binários não balanceados porém, infelizmente, na literatura não existe uma definição formal de desbalanceamento em dados binários. Assim, propomos a seguinte definição:

Definição 1. Seja $Y \sim Bern(p)$ com p a probabilidade de sucesso. Dizemos que a variável resposta binária Y é desbalanceado, se e somente se $\kappa := |2p - 1| \geq 0,2$.

Quando os dados são perfeitamente balanceados é esperado uma proporção de igual de uns e zeros, isto é $p = 1 - p = 0,5$ e então $\kappa = p - (1 - p) = 2p - 1 = 0$. Por outro lado, se $p = 0,4$, então $k = 0,2$. Na prática, em nosso estudo assumiremos que $\kappa = 0,2$ é a minima diferença entre a probabilidade de uns e zeros, a qual é suficiente para que uma variável binária seja considerada como tendo dados não balanceados. Nesse caso, proporções de uns menores 0.4 ou superiores a 0.6 configuram um desbalanceamento relativo em regressão binária.

No entanto, quando queremos explicar a variável dependente dicotômica em função das covariáveis, que não são aleatórias e são independentes entre si e conhecidas, são utilizadas, rotineiramente modelos de regressão binária com funções de ligação comuns como a logito e probito (HOSMER; LEMESHOW, 1989). A seguir definimos esse modelo.

Definição 2. O modelo de regressão binária é definido por

- Componente aleatório: Considere $y_1, y_2, \ldots, y_n$ variáveis de resposta binária tais que $Y_i \sim Bernoulli(p_i)$ em que: Y_i é uma variável binária tal que $Y_i = 1$ ocorre com probabilidade de sucesso $p_i \in (0, 1)$;

- Componente sistemático: $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$ é o i-ésimo preditor linear. onde $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ é um vetor com as variáveis explicativas, em que $x_{i1} = 1$ representa o intercepto; e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ é um vetor de k coeficientes de regressão, sendo β_1 correspondente ao intercepto e $\beta_2, ..., \beta_k$ são os coeficientes de regressão.

- Função de ligação: $g(p_i) = F^{-1}(\eta_i) = p_i \qquad i = 1, \ldots, n$
  onde $F(.)$ denota uma função de distribuição acumulada (fda) com suporte na reta, tal que $p_i = F(\eta_i)$ e a função inversa $g(p_i) = F^{-1}(p_i) = \eta$ é chamada de função de ligação relacionado o componente aleatório com o componente sistemático.

A função $F(.)$ é importante e satisfaz algumas propriedades: F está definida nos reais e toma valores no intervalo $(0, 1)$. Além disso temos que quando F é uma fda de uma distribuição simétrica a função de ligação resultante é simétrica e tem forma simétrica em torno de $p_i = 0,5$, como ocorre nas funções de ligação logito, probito e cauchito onde a F são respectivamente as distribuições padronizadas logística, normal e cauchy. Já quando F é uma fda de uma distribuição assimétrica a função de ligação resultante é assimétrica. Uns dos exemplos mais populares de ligações assimétricas são complementar log-log ou cloglog e a log-log em que a primeira corresponde ao uso da função de distribuição acumulada (fda) da Distribuição de Gumbel de Valor Mínimo enquanto que a segunda é proveniente da Distribuição de Gumbel Valor Máximo. Um resumo destas ligações comuns são mostradas na Tabela 1

Tabela 1 - Funções de ligação comuns

Ligações Ligações	Probabilidade $p = F(\eta)$	Função de ligação $g(p) = F^{-1}(p) = \eta$
Logito	$p = \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$	$g(p) = \log\left(\frac{p}{1-p}\right)$
Probito	$p = \Phi(\eta)$	$g(p) = \Phi^{-1}(p)$
Cauchito	$p = \frac{1}{2} + \frac{\arctan(\eta)}{\pi}$	$g(p) = \tan\left(\pi\left(p - \frac{1}{2}\right)\right)$
Cloglog	$p = 1 - \exp\{-\exp\{\eta\}\}$	$g(p) = \log(-\log(1-p)))$
Loglog	$p = \exp\{\exp\{\eta\}\}$	$g(p) = \log(\log(p))$

em que $\Phi(.)$ denota a acumulada da distribuição normal padrão e arctan denota o arco tangente.

Na Figura 1 apresentamos as correspondentes curvas para diferentes valores do preditor linear η. Note que quando o preditor linear é zero, a probabilidade de sucesso é 0,5 para as ligações probito, logito e cauchito, neste caso indica que temos ligações simétricas. Por outro lado, para este mesmo nível do preditor linear, temos que no caso da ligação cloglog, a probabilidade de sucesso é maior do que 0,5 e no caso da ligação loglog a probabilidade de sucesso é menor do que 0,5, isto acontece por serem ligações assimétricas.

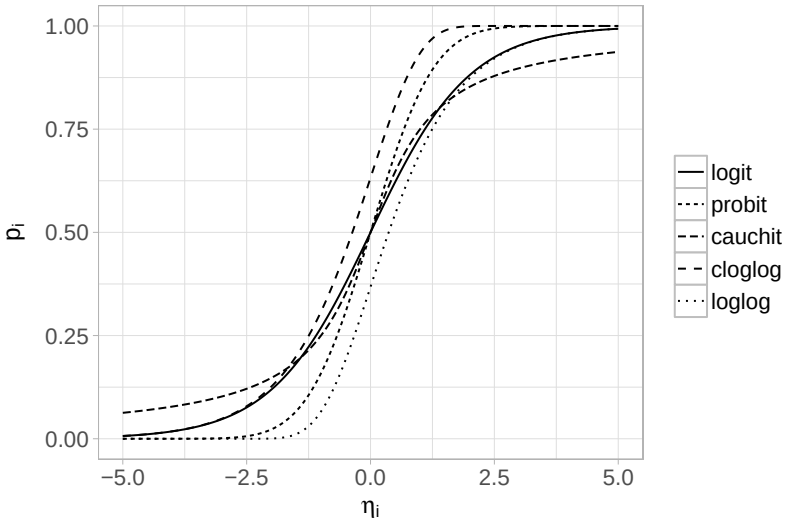

Figura 1 - Curvas de resposta para diferentes funções de ligação comuns na regressão binária em função do preditor linear η.

Com o propósito de estimar os coeficientes de regressão $\boldsymbol{\beta}$ considerando a correspondente função de ligação adotada em um modelo de regressão binária, consideramos a função de verossimilhança neste modelo para uma amostra aleatória e independente de respostas e covariáveis, dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} F(\boldsymbol{x}_i^\top \boldsymbol{\beta})^{y_i} (1 - F(\boldsymbol{x}_i^\top \boldsymbol{\beta}))^{1-y_i} \tag{1}$$

Maiores detalhes da estimação de máxima verossimilhança podem ser consultados em Agresti (2013), Hosmer e Lemeshow (1989) e Paula (2004).

Sob enfoque Bayesiano, o parâmetro de interesse $\boldsymbol{\beta}$ é assumido como variável aleatória e assim estabelece a distribuição de probabilidade a priori que reflete nosso conhecimento prévio de seu comportamento. Combinando a função de verossimilhança e a distribuição a priori $\pi(\boldsymbol{\beta})$ podemos obter a distribuição posteriori do parâmetro de interesse dada por:

$$f(\boldsymbol{\beta}|y) \propto L(\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

Para o caso da regressão logística a função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(\frac{1}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} \right)^{1-y_i} = \frac{\exp\left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{i=1}^{n} \left(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right)}$$

e se consideramos que a priori $\boldsymbol{\beta} \sim N_k(\mathbf{0}, \Sigma)$ segue uma distribuição normal multivariada de ordem k, i.e. $\pi(\boldsymbol{\beta}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$, logo a distribuição a posteriori é dada por:

$$f(\boldsymbol{\beta}|y) \propto \frac{\exp\left(\sum_{i=1}^n y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{i=1}^n \left(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right)} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right).$$

Observemos que a distribuição a posteriori não apresenta uma forma conhecida, por isso é necessário utilizar métodos computacionais com o objetivo de obter uma amostra da distribuição a posteriori, como por exemplo algoritmos MCMC. Maiores detalhes acerca da estimação bayesiana do modelo de regressão binaria pode ser seguida em Bazán e Bayes (2010).

3 Regressão binária para dados desbalanceados

Nossa definição de desbalanceamento em dados de resposta binária se baseia na diferença entre as proporções de zeros e uns. Notemos que argumentos similares foram iludidos por vários autores como justificativa para a proposta de funções de ligação assimétrica em regressão binária. Assim, por exemplo, Chen, Dey e Shao (1999a) argumentam que quando a probabilidade de uma resposta binária se aproxima a zero em uma taxa diferente de quando se aproxima a um, as ligações simétricas para o ajuste dos dados podem ser inadequados. Neste caso, temos que considerar ligações assimétricas. Uns dos exemplos mais populares de ligações assimétricas são complementar log-log ou cloglog e a log-log vistos na seção 2, porém o tipo de assimetria que eles apresentam é fixa e então as curvas de resposta correspondentes não dependem de nenhum parâmetro desconhecido adicional. Também, Bazán, Torres-Aviles, Suzuki e Louzada (2017) relatam que na presença de desbalanceamento de uns e zeros, as ligações simétricas podem ser inadequadas e inflexíveis para se adequar a curva de resposta dos dados, e neste caso, a função de ligação pode estar mal especificada, o que pode levar a vícios grandes nas estimativas da resposta média da variável resposta.

Nesta seção apresentamos o modelo regressão binária com funções de ligação potência e reversa de potência propostas por Bazán, Torres-Aviles, Suzuki e Louzada (2017) e Lemonte e Bazán (2018) para ajustar dados de regressão binária na presença de dados não balanceados.

3.1 Regressão binária com funções de ligação potência e reversa de potência

Considere $\boldsymbol{Y}$ um vetor $n \times 1$ das variáveis aleatórias independentes binárias correspondentes as observações da variável resposta. Seja $\boldsymbol{x_i}$ um vetor $(k+1)\times 1$ das covariáveis, e seja $\boldsymbol{X}$ a matriz $n \times (k+1)$ com linhas $\boldsymbol{x_i'}$. Seja também $\boldsymbol{\beta}$ um vetor $(k+1) \times 1$ dos coeficientes de regressão associados às covariáveis. Consideremos $Y_i = 1$ com probabilidade p_i e $Y_i = 0$ com probabilidade $1 - p_i$. Desse modo

podemos definir o modelo de regressão binária com funções de ligação potência e reversa de potência da seguinte forma:

$$Y_i \sim Bernoulli(p_i) \tag{2}$$

$$p_i = F_\lambda(\eta_i) = F_\lambda(\boldsymbol{x}_i^\top \boldsymbol{\beta}),\ i = 1, \ldots, n \tag{3}$$

em que, $F_\lambda(.)$ denota a fda de uma distribuição potência ou reversa de potência e $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ é o i-ésimo preditor linear. Neste caso, $g(p-i) = F_\lambda^{-1}(p_i) = \eta_i$ denomina-se função de ligação potência ou reversa de potência.
Algumas propriedades das distribuições potência e reversa de potência são apresentadas no apêndice A e em (ANYOSA, 2017). Na Tabela 2 são apresentadas as 3 funções de ligação potência e as 3 funções de ligação reversa de potência estudadas nesse trabalho.

Tabela 2 - Probabilidades $p(.)$e funções de ligação $g(p)$ potência e reversa de potência

Ligações Ligações	Probabilidade $p = F_\lambda(\eta)$	Função de ligação $g(p) = F^{-1}(p) = \eta$
Potência logito (PL)	$\left(\frac{\exp\{\eta\}}{1+\exp\{\eta\}}\right)^\lambda$	$\log\left(\frac{p^{1/\lambda}}{1-p^{1/\lambda}}\right)$
Reversa de potência logito (RPL)	$1-\left(\frac{\exp\{\eta\}}{1+\exp\{-\eta\}}\right)^\lambda$	$\log\left(\frac{(1-p)^{1/\lambda}}{1-(1-p)^{1/\lambda}}\right)$
Potência probito (PP)	$(\Phi(\eta))^\lambda$	$\Phi^{-1}\left(p^{1/\lambda}\right)$
Reversa de potência probito (RPP)	$1-(\Phi(-\eta))^\lambda$	$-\Phi^{-1}\left((1-p)^{1/\lambda}\right)$
Potência cauchito (PC)	$\left(\frac{1}{2}+\frac{arctan(\eta)}{\pi}\right)^\lambda$	$tan\left(\pi\left(p^{1/\lambda-0.5}\right)\right)$
Reversa de potência cauchito (RPC)	$1-\left(\frac{1}{2}+\frac{arctan(-\eta)}{\pi}\right)^\lambda$	$-tan\left(\pi\left((1-p)^{1/\lambda-0.5}\right)\right)$

em que, como antes $\Phi(.)$ é acumulada da distribuição normal padrão.

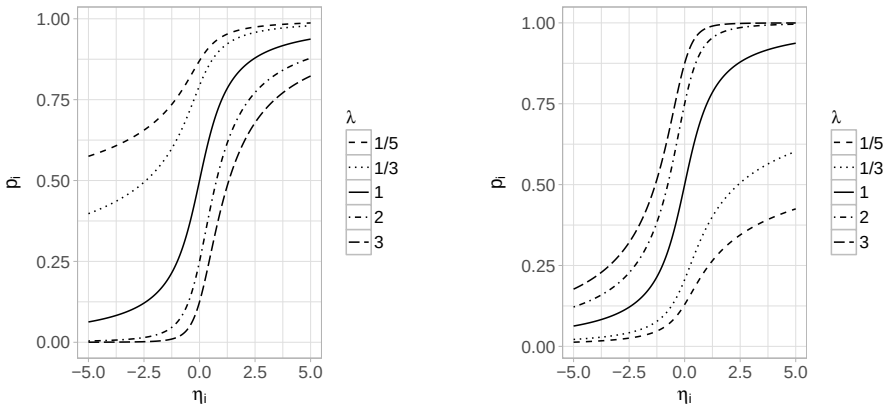

Figura 2 - Curvas de resposta para as funções de ligação potência cauchito (esquerda) e reversa de potência cauchito (direita) para diferentes valores do preditor linear η e diferentes valores do parametro de assimetria λ.

Na Figura 2 é apresentada a curva das funções de resposta considerando as ligações potência e reversa de potência para diferentes valores de $\lambda = 1/5; 1/3; 1; 2; 3$ considerando como distribuição de base a acumulada da distribuição cauchito. Outras figuras similares para outras funções de ligações podem ser obtidas considerando o código disponível no Apêndice B.

Com base na Figura 2 observamos que a linha continua ($\lambda = 1$), representa a curva de resposta da distribuição de base da ligações comuns apresentadas na Figura 1. Enquanto que, para os casos onde $\lambda \neq 1$, as funções de ligação são assimétricas, pontos a serem destacados. Em geral, note que nas ligações potência, a curva de resposta para $\lambda < 1$ está a esquerda da curva de resposta base; e a curva de resposta para $\lambda > 1$ está a direita da curva de resposta base. Por sua vez, quando observam-se as ligações reversa de potência, o efeito é o contrário. Além disso, há reversibilidade entre as funções de ligação potência e sua correspondente função de ligação reversa de potência, no sentido que uma é o espelho da outra.

3.2 Estimação

Para o modelos de regressão binária com função de ligação potência ou reversa de potência apresentados acima, a verossimilhança é dada por :

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda} | \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} [F_\lambda(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{y_i} [1 - F_\lambda(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{1-y_i}$$

em que F_λ pode ser quaisquer das fda das distribuições potência ou reversa de potência introduzidas na Tabela 2. As correspondentes funções de ligação da Tabela 2, são denominadas potência logito, reversa de potência logito, potência probito, reversa de potência probito, potência cloglog, reversa de potência cloglog,

potência loglog, reversa de potência loglog, potência cauchito e reversa de potência cauchito. Na verossimilhança dada acima, $p_i = F_\lambda(\eta_i)$ representa a probabilidade de sucesso na distribuição de Bernoulli da variável resposta y_i, isto é $P(Y_i = 1) = p_i$.

Segundo Bazán, Romeo e Rodrigues. (2014), é conveniente introduzir a δ-parametrização, definindo $\delta = log(\lambda)$, com isso uma distribuição a priori deve ser especificada para δ. Note também que nesse caso o parâmetro $\delta \in \mathbb{R}$. Considerando esta reparametrização do modelo, precisa-se especificar as distribuições a priori para $\boldsymbol{\beta}$ e δ, uma vez que ambos os parâmetros são de diferentes tipos; nós assumimos eles independentes, isto é: $\pi(\boldsymbol{\beta}, \delta) = \pi_1(\boldsymbol{\beta})\pi_2(\delta)$.

Para os coeficientes de regressão $\boldsymbol{\beta}$, assume-se $\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j})\ \forall j = 1, \cdots, k$. Nesse caso considera-se como hiperparâmetro $\mu_{\beta_j} = 0$ e $\sigma^2_{\beta_j} = 100$ denotando a ignorância em relação ao parâmetro. O anterior é uma prática comum em modelos de regressão como pode ser visto em Jiang et al. (2013). Para o parâmetro δ, seguindo Bazán, Romeo e Rodrigues. (2014), considera-se a *Uniforme*(-2,2), pois os valores fora do intervalo $[e^{-2}, e^2]$ não são observados empiricamente considerando esta especificação. A estrutura hierárquica do modelo com a utilização da δ-parametrização é dada por:

$$Y_i\boldsymbol{\beta}, \delta \sim Bernoulli(p_i) \tag{4}$$

$$p_i = F_\delta(\boldsymbol{x}_i^\top \boldsymbol{\beta}), i = 1, 2, \cdots, n \tag{5}$$

$$\boldsymbol{\beta}_j \sim N(0, 100), j = 1, 2, \cdots, k \tag{6}$$

$$\delta \sim U(-2, 2) \tag{7}$$

Considerando a estrutura hierárquica dada acima, a densidade da distribuição a posteriori dos modelos de regressão binária com função de ligação potência ou de potência, tem a seguinte forma:

$$\pi(\boldsymbol{\beta}, \delta|\boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{n}[F_\delta(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{y_i}[1 - F_\delta(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{1-y_i} \frac{1}{4} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sqrt{100}} \exp\left\{-\frac{(\beta_j)^2}{2(100)}\right\} \tag{8}$$

Como a distribuição a posteriori não tem forma fechada, ela pode ser simulada via algoritmos MCMC. Nesse trabalho, usamos o algoritmo `No-U-Turn Sampler (NUTS)` via o pacote *Stan*. O pacote `Stan` permite usar uma sintaxe que facilita escrever diferentes modelos sobre abordagem bayesiana de forma similar aos códigos em `BUGS`, mas diferente dele, este pacote pode rodar em `R`, `Julia`, `Python` ou `Stata`.

3.3 Avaliação do ajuste do modelo e comparação de modelos

Para decidirmos pelo melhor modelo que ajusta aos dados analisados, várias técnicas podem ser usadas. Neste trabalho, nós consideramos: a) análise de resíduos e ajuste de modelos, b) métodos auxiliares para ajuste do modelo: Poder preditivo, e o c) uso de critérios de comparação de modelos.

a) Análise de resíduos e ajuste do modelo

Considerando as estimativas dos parâmetros do modelo ajustado (usando a media a posteriori para os parâmetros de regressão e a mediana a posteriori para o parâmetro λ (δ) nós podemos obter diferentes tipos de resíduos usando-os como estimador *plug-in* considerando: $\hat{p}_i = F_{\hat{\delta}}(\boldsymbol{x}_i^\top \hat{\beta})$.

- Resíduo de Pearson: $r\hat{p}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}, \quad i = 1, 2, ..., n.$

  Note que a estatística Chi-quadrado de Pearson é $Q_P = \sum_{i=1}^{n} \sum_{j=0}^{1} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \right)^2 = \sum_{i=1}^{n} r\hat{p}_i{}^2$ em que $o_{i1} = y_i$, $o_{i0} = 1 - y_i$, $e_{i1} = \hat{p}_i$ e $e_{i0} = 1 - \hat{p}_i$. Q_P pode ser comparado com a estatística $\chi_{(n-k)}$ para avaliar o ajuste do modelo.

- Resíduo de desvio: $\hat{d}_i = \pm\{2[y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)]\}^{\frac{1}{2}}$

  em que o sinal é positivo se $y_i \geq \hat{p}_i$ e negativo caso contrário.
  Note que a estatística de Desvio $Q_L = \sum_{i=1}^{n} -2[y_i log(\hat{p}_i) + (1-y_i) log(1-\hat{p}_i)] = \sum_{i=1}^{n} d_i^2$ Q_L podem ser comparado com a estatística $\chi_{(n-k)}$ para avaliar o ajuste do modelo.

- Resíduo quantílico aleatorizado : $r_{q,i} = \Phi^{-1}(u_i), i = 1, ..., n.$

  onde u_i é um valor aleatório de uma distribuição uniforme com intervalo

  $[I_{1-\hat{p}_i}(2 - y_i, y_i), I_{1-\hat{p}_i}(1 - y_i, y_i + 1)], i = 1, ..., n$ em que $I_x(a, b) = B(x; a, b)/B(a, b)$ é função beta incompleta regularizada e $\Phi^-1(.)$ é a função de distribuição acumulada da normal padrão.

  O resíduo quantílico aleatorizado (normalizado) foi proposto por Dunn e Smyth (1996) e é definido como:

A adequabilidade de ajustes de modelos de regressão normal é facilmente analisada por meio dos resíduos da regressão, pois a distribuição desses erros é conhecida. Entretanto, o mesmo não ocorre com modelos cuja variável resposta não é normalmente distribuída, como é o caso da distribuição bernoulli. Assim, os resíduos de pearson e de desvio, propostos previamente, não apresentam distribuição normal. Além disso, as estatísticas Q_P e Q_L não são eficientes para avaliar o ajuste. Enquanto que para o resíduo quantílico aleatorizado (normalizado) proposto por Dunn e Smyth (1996) isso não ocorre. De acordo com sua definição, quando as suposições do modelo são satisfeitas, estes resíduos deveriam ser normalmente distribuídos em caso de adequabilidade da predição ou do ajuste, conforme explica Rigby e Stasinopoulos (2005). Com isso, a utilização destes resíduos é vantajosa por ele possuir distribuição conhecida independente da distribuição da variável resposta e porque sua distribuição é fácil de ser averiguada por meio de testes de hipóteses e gráficos normais de probabilidade (QQ-plot).

b) Métodos auxiliares para ajuste do modelo: Poder preditivo

Para avaliar o poder preditivo do modelo, é necessário estabelecer um ponto de corte (pc) com $(0 < pc < 1)$, tal que: as probabilidades preditas pelo modelo $\hat{p}_i \geq pc \Rightarrow Y = 1$ e $\hat{p}_i < pc \Rightarrow Y = 0$. Com isto, para os modelos de regressão binaria é possível definir uma matriz de dupla entrada 2×2 chamada matriz de confusão mostrada na Tabela 3.

Tabela 3 - Matriz de Confusão

	Reposta predita		
Resposta observada	$\hat{y} = 1$	$\hat{y} = 0$	Totais
$y = 1$	a	b	$a + b$
$y = 0$	c	d	$c + d$
Totais	$a + c$	$b + d$	n

Na tabela apresentada y corresponde as respostas binárias na amostra onde $y = 1$ é sucesso e $y = 0$ é falha, $\hat{y}$ corresponde as respostas classificadas pelo modelo sob análise, isto é, $\hat{y} = 1$ a observação é classificada como sucesso e se $\hat{y} = 0$ a observação é classificada como falha.
Como conhecido, os valores das células da tabela acima são: a ou verdadeiro positivo (VP) que são os casos em que a observação é classificada como sucesso e de fato é sucesso; b ou falso negativo (FN) que são os casos em que a observação é classificada como falha e de fato é sucesso; c ou falso positivo (FP) que são os casos onde a observação é classificada como sucesso e de fato é falha, e finalmente d ou verdadeiro negativo (VN) que são os casos em que: a observação é classificada como falha e de fato é falha. Também $a + b + c + d = n$, onde n é o tamanho da amostra ou total de observações.

Em nosso trabalho, esta tabela ou matriz de confusão foi obtida com a função `confusionMatrix` do pacote `caret` do programa `R`. Considerando a Tabela 3 é comum calcular as seguintes medidas para a avaliação preditiva do modelo ajustado: Sensibilidade ou taxa/razão de verdadeiros positivos (RVP) = $\frac{a}{a+b}$, Especificidade ou taxa/razão de verdadeiros negativos (RVN) = $\frac{d}{c+d}$, taxa/razão de falso negativo (RFN) = $\frac{b}{a+b}$, taxa/razão de falso positivo (RFP) = $\frac{c}{c+d}$; e finalmente o Valor predito = $\frac{a+d}{n}$ = ou proporção geral de acertos. A sensibilidade mede a proporção de sucessos que foram corretamente classificados como sucessos. Enquanto que a especificidade mede a proporção de falhas que tem sido corretamente classificados como falhas. Já o valor predito mede a proporção de acerto geral.
Além disso é possível usar a chamada curva característica de operação do receptor *ROC-receiver operating characteristic curve* - (HANLEY; MCNEIL, 1982), a qual é um gráfico que apresenta o comportamento de um classificador binário em diferentes cenários do valor limite para classificação. Para diversos valores de pontos de corte pc podemos construir a Curva ROC de modo que iremos considerar uma figura dos pontos ou pares (x, y) = (1 - especificidade, sensibilidade). Neste caso é considerado

que o modelo discrimina perfeitamente se $(x, y) = (0, 1)$, portanto se os pontos de corte estão próximos do canto superior esquerdo produzirão a maior porcentagem de acertos. Assim, quanto mais próximo de 1 a área sob a curva, melhor é o poder preditivo do modelo. Os cálculos da sensibilidade e especificidade e da curva ROC foram feitos feitos utilizando os pacotes `pROC`, `InformationValue` e `ModelMetrics` do programa `R`.

c) Critérios de comparação de modelos

Entre os principais critérios para comparação de modelos na inferência bayesiana considerados nesse trabalho temos : *Deviance information criterion* (DIC), *Expected Akaike Information Criterion* (EAIC), *Expected Bayesian Information Criterion* (EBIC) e *Widely Applicable Information Criterion* (WAIC). Os três primeiros critérios são baseados na média a posteriori do desvio $E[D(\boldsymbol{\beta})]$ em que: $D(\boldsymbol{\beta}) = -2ln(p(y|\boldsymbol{\beta}) = -2\sum_{i=1}^{n} lnP(Y_i = y_i|\boldsymbol{\beta})$ que é uma medida de ajuste que pode ser aproximada utilizando a saída da simulação MCMC da distribuição a posteriori, esta aproximação é dada por: $\bar{D} = \frac{1}{G}\sum_{i=1}^{G} D(\boldsymbol{\beta}^g)$, em que o índice g indica o g-ésimo valor simulado de um total de G simulações. Os critérios de comparação citados acima podem ser estimados da seguinte forma:

$$\widehat{EAIC} = \bar{D} + 2p, \;\; \widehat{EBIC} = \bar{D} + plogN, \;\; \widehat{DIC} = \bar{D} + \widehat{\rho}_D = 2\bar{D} - \hat{D}$$

em que p é o número de parâmetros no modelo, N é o total de observações e $\widehat{\rho}_D$ é o número efetivo de parâmetros e é definido como: $\widehat{\rho}_D = E[D(\boldsymbol{\beta})] - D[E(\boldsymbol{\beta})]$, sendo $D[E(\beta, E(\lambda), E(\theta)]$ o desvio da média a posteriori obtido quando avaliamos a função desvio na média a posteriori dos parâmetros e quando é estimado por: $\hat{D} = D\left(\frac{1}{G}\sum_{i=1}^{G} \boldsymbol{\beta}^g\right)$. Em EAIC e EBIC $2p$ e $plogN$ são valores fixos que penalizam a média a posteriori do desvio.

O critério WAIC, versão generalizada do AIC proposto por Watanabe (2010) é estimado da seguinte forma

$$\widehat{WAIC} = -2(\widehat{llpd} - p\widehat{WAIC})$$

onde $\widehat{lpdd}$ é o logaritmo da densidade preditiva pontual (lppd) definida por $\widehat{lpdd} = \sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\boldsymbol{\beta}^m, \lambda^m)\right)$, e $p\widehat{WAIC} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\boldsymbol{\beta}^m, \lambda^m)\right) - \frac{1}{M}\sum_{m=1}^{M}\log\left(p(y_i|\boldsymbol{\beta}^m, \lambda^m)\right)\right)$ é um termo para corrigir o número efetivo de parâmetros.

Para comparar dois ou mais modelos alternativos, o modelo que apresenta melhor ajuste ao conjunto de dados será o modelo que apresentar o menor valor dos critérios analisados.

4 Desbalanceamento em dados binários

4.1 Efeitos na estimação de parâmetros para o modelo de regressão binário quando os dados são desbalanceados

Nessa seção foi desenvolvido um estudo de simulação para avaliar o efeito do desbalanceamento na estimação dos parâmetros do modelo logístico. Para isso, simulamos dados do modelo potência logístico com um parâmetro λ dado, isto é

$$Y_i \sim Bernoulli(p_i), \quad p_i = \left\{ \frac{exp(\beta_0 + \beta_1 x_{i1})}{1 + exp(\beta_0 + \beta_1 x_{i1})} \right\}^{\lambda}$$

para $i = 1, ..., n$, onde p_i é obtido usando a função de ligação potência logístico, em que β_0 é o intercepto e β_1 é o coeficiente de regressão associado a covariável $\boldsymbol{x}_1$

Para esse estudo os valores de β_0 e β_1 foram fixados, respectivamente, em 0 e 1. Enquanto que os valores da covariável $\boldsymbol{x}_1$ foram gerados a partir de uma distribuição uniforme entre $(-4, 4)$. E por último, λ foi fixado igual 4, o que leva a uma proporção de 0,28 , e segundo nossa definição de desbalanceado $\kappa = 0{,}44$. A estimação clássica foi ajustada usando a função *glm* do pacote *stats* enquanto que para estimação bayesiana foi usado o pacote `MCMCpack` com a função `MCMClogit`. Foram simulados dados usando 5 diferentes tamanhos de amostra $n = (50, 200, 500, 2000, 5000)$ e o processo foi repetido 100 vezes.

Para cada parâmetro foi encontrado as seguintes medidas de comparação :

a) O vicio estimado de β_j definido por $\widehat{Vicio} = \beta_j - \hat{\beta}_j$ onde j = 0,1 e $\hat{\beta}_j = \frac{\sum_{r=1}^{R} \hat{\beta_{jr}}}{R}$ é a média dos parâmetros estimados para R réplicas

b) O erro quadrático médio estimado de β_j definido por $\widehat{EQM} = \frac{1}{R} \sum_{r=1}^{R} \left[\hat{\beta_{jr}} - \beta_j \right]^2$. Se o EQM apresenta um valor baixo, existirá boa acurácia, isto é, boa recuperação e precisão de parâmetros (WALTHER; MOORE, 2005).

c) O erro padrão médio de β_j definido por $\widehat{\overline{DP}} = \frac{\sum_{r=1}^{R} DP_{jr}}{R}$

onde $DP_{jr} = \sqrt{\frac{\sum_{r=1}^{R} \left[\hat{\beta_{jr}} - \overline{\hat{\beta_j}} \right]^2}{R}}$ é o erro padrão dos parâmetros estimados para R réplicas

d) A probabilidade média definida por $\widehat{\overline{p}} = \frac{\sum_{r=1}^{R} \hat{p_{jr}}}{R}$ onde $\hat{p_{jr}}$ é a probabilidade estimada para R réplicas

e) O tempo computacional dado em segundos foi obtido utilizando a função `system.time` do programa R.

Com base nas medidas citadas anteriormente obtivemos os resultados que são mostrados na Tabela 4.

Tabela 4 - Estimativas para o modelo logístico com dados desbalanceados simulados do modelo potência logístico ($\lambda = 4$)

	β_0			β_1				
Amostra	$\widehat{Vicio}$	$\widehat{EQM}$	*Dpbar*	$\widehat{Vicio}$	$\widehat{EQM}$	*Dpbar*	$\widehat{\overline{p}}$	tempo (s)
N = 50	-4,08	27,10	1,54	1,25	3,82	0,79	0,28	52,32
N = 200	-2,73	7,70	0,47	0,52	0,33	0,23	0,28	89,51
N = 500	-2,60	6,78	0,28	0,44	0,21	0,14	0,28	155,22
N = 2000	-2,55	6,50	0,14	0,43	0,19	0,07	0,28	392,64
N = 5000	-2,54	6,47	0,09	0,42	0,18	0,04	0,28	913,41

Observando os resultados obtidos temos que em todos os casos as taxas de probabilidade simuladas é de 28% e o tempo computacional aumenta quando o tamanho da amostra aumenta. Além disso encontramos que as estimativas de ambos os parâmetros β_0 e β_1 são viciadas, principalmente para β_0, pois o seu viés aumenta conforme o tamanho da amostra diminui. Isto pode, também ser observado na Figura 3.

Em resumo, considerando dados desbalanceados ($p = 0,28$) nos encontramos viés (vicio) nas estimativas do intercepto (β_0) e do coeficiente angular (β_1) quando ajustado um modelo logístico. Assim, podemos observar que o viés é maior para as estimativas para o intercepto do que para as estimativas do coeficiente de regressão. Também observamos que o viés diminui conforme o tamanho da amostra aumenta. Como o vicio se mantém alto, concluímos então que o ajustar um modelo logístico não é adequado quando os dados são desbalanceados.

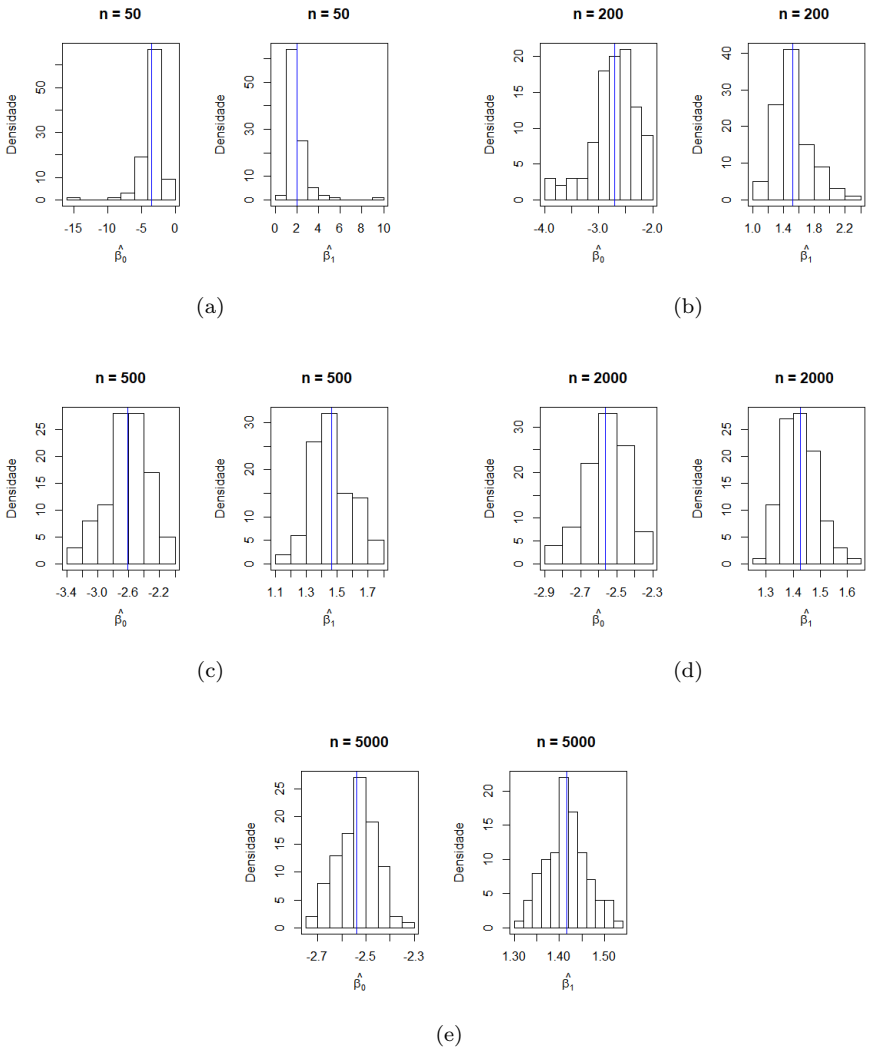

Figura 3 - Distribuição das estimativas bayesianas dos valores estimados de β_0 e β_1 no modelo logístico na presença de dados desbalanceados $p = 0,28$ e com $\lambda = 4$.

Os estimadores de máximo verossimilhança (EMV) possuem propriedades assintóticas muito interessantes, mas são conhecidos por serem tendenciosos para pequenas amostras. Na presença de dados não balanceados, de fato há uma ocorrência menor de uma classe (uma pequena amostra) em relação a outra classe (uma amostra grande). Algumas vezes, a inferência exata para regressão logística usando métodos de permutação foram propostos ao invés de usar MV. Estes tipos de métodos são intensivos computacionalmente e não necessariamente lidam com o problema de dados não balanceados onde os dados não são necessariamente pequenos, somente o número de eventos é pequeno comparado com o número de covariáveis. King e Zeng (2001) afirmam que a regressão logística subestima drasticamente a probabilidade de eventos raros, isto é, o viés do intercepto é negativo, de modo que o $\hat{\beta}_0$ estimado é muito pequeno e como resultado

$\hat{p} = \widehat{P(y = 1)}$ é subestimado.

Para contornar este problema, King e Zeng (2001) e Firth (1993) propuseram métodos de correção. Numa recente pesquisa, de la Cruz et al. (2019) mostraram que os métodos de correção não corrigem adequadamente o viés na estimativa dos coeficientes de regressão e que os modelos com ligações assimétricas considerados produzem melhores resultados para certos tipos de dados desequilibrados.

O problema das estimativas infinitas de MV para coeficientes de regressão binomial pode ser devido à multicolinearidade. Esse tipo de problema pode ser abordado usando penalização, por exemplo usando regressão ridge ou lasso. A regressão para dados discretos pode sofrer de uma causa diferente de estimativas infinitas dos EMV: a separação completa. A separação completa acontece quando uma combinação linear dos preditores é perfeitamente preditiva do resultado. Nesse caso o EMV não existe. Também Paal (2014) cita que na regressão binomial aplicada, a separação é comum quando existem preditores binários. Mas que, no entanto esse tipo de problema é mais comum em dados desbalanceados.

4.2 Um ajuste adequado para dados não balanceados

Nesta seção nos ajustamos o modelo potência logístico para os dados simulados seguindo o procedimento descrito na seção anterior. O método de estimação considerado é o bayesiano e desenvolvemos um código em `Rstan` para o ajustar um modelo potência logístico. Os resultados são mostrados na Tabela 5.

Tabela 5 - Estimativas dos parâmetros para o modelo logístico com dados desbalanceados simulados do modelo potência logístico usando o rstan ($\lambda = 4$) e ($p = 0,28$)

Amostra	Parâmetros	$\widehat{Vicio}$	$\widehat{EQM}$	$\widehat{DP}$
50	β_0	-10,617	150,561	0,879
	β_1	3,562	22,380	0,445
	λ	-3,542	12,660	0,049
200	β_0	-4,657	38,502	0,586
	β_1	1,322	3,398	0,183
	λ	-2,458	7,143	0,150
500	β_0	-1,173	2,467	0,149
	β_1	0,278	0,174	0,044
	λ	-1,294	2,795	0,151
2000	β_0	-0,393	0,472	0,080
	β_1	0,088	0,021	0,017
	λ	-0,516	1,483	0,158
5000	β_0	0,000	0,101	0,045
	β_1	0,003	0,003	0,007
	λ	0,238	0,948	0,135

Como podemos observar o modelo potência logístico recupera adequadamente os parâmetros e apresenta um vicio próximo de zero conforme o tamanho da amostra aumenta. Além disso, é possível perceber que existe um vicio ligeiramente maior nas estimativas do parâmetro λ, do que no caso dos coeficientes de regressão, o qual diminui quando o tamanho da amostra cresce. Desse modo recomendamos tamanhos de amostras maiores.

5 Aplicações

Nesta seção apresentamos duas aplicações, uma no campo na análise linguística e outra na área educacional. Tentamos oferecer diferentes análises em cada aplicação para ilustrar diferentes casos de uso das ligações assimétricas estudadas nesse trabalho. Na primeira aplicação, incluímos análise de resíduos e na segunda incluímos métodos auxiliares para ajuste do modelo: Poder preditivo.

5.1 Aplicação a dados linguísticos

Van Cappel e Thomson-Sintra (1993) disponibilizaram o conjunto de dados *phoneme* em (⟨https://www.openml.org/d/1489⟩). O conjunto de dados atual foi formatado pelo repositório da KEEL, mas originalmente hospedado pelo Projeto ELENA. O conjunto de dados é originário do projeto europeu ESPRIT 5516:

ROARS. O enfoque do projeto é pautado no desenvolvimento e implementação de um sistema analítico em tempo real para reconhecimento de fala em francês e espanhol. O conjunto de dados trás como informação a distinção entre sons nasais (classe 0) e oral (classe 1) onde 3818 (70,6%) são sons orais e 1586 (29,4%) são sons nasais, isto nos da um desbalanceamento de $\kappa = 41{,}12\%$. Cinco atributos diferentes foram escolhidos para caracterizar cada vogal: eles são as amplitudes dos cinco primeiros harmônicos *AHi* os quais foram normalizados pela energia total *Ene* (integrada em todas as frequências): *AHi / Ene*. Os fonemas são transcritos da seguinte forma: *sh* como em *she* (V1), *dcl* como em *dark* (V2), *iy* como a vogal em *she* (V3), *aa* como a vogal em *dark* (V4), e *ao* como a primeira vogal em *water*. Abaixo é apresentado a Figura 4 com uma prévia análise das variáveis.

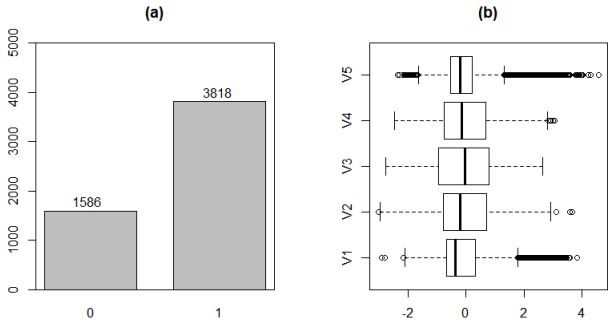

Figura 4 - (a) corresponde a variável categórica, 0 (sons nasais) e 1 (sons orais). E (b) corresponde as covariáveis de amplitude harmônica.

Em suma, observando as figura acima temos uma noção do desbalanceamento da variável resposta. Além disso como as covariáveis foram padronizadas já era esperado pouco variabilidade, mediana próxima da média, e que a média entre as variáveis tenham valores próximos. Notando que as covariáveis V1 e V5 possuem muitos valores extremos.

Para estes dados, foram ajustados os modelos regressão binária com ligações simples logito, probito e cauchito, juntamente com as ligações de potência e reversa de potência para estes casos, as quais foram descritas na seção 3. Para fazer a estimação dos modelos de regressão binária propostos, seguiu-se o procedimento da seção 3.4 , considerando o algoritmo NUTS implementado no programa `R` por meio do pacote `Stan` (STAN DEVELOPMENT TEAM, 2018).

Os resultados da comparação dos diferentes modelos são apresentados na Tabela 6, na qual as notações para as ligações correspondem às apresentadas no Capítulo 3.

Tabela 6 - Critérios de informação dos modelos de regressão binária ajustados considerando funções de ligação potência e reversa de potência. Dados linguísticos

Ligações		DIC	EAIC	EBIC	WAIC
Simétricas	C	5249,52	5255,51	5265,90	5250,38
	L	5100,31	5106,26	5116,66	5100,25
	P	5077,31	5083,38	5093,77	5077,23
Assimétricas	PC	5017,88	5025,04	5037,17	5019,12
	CLL	5033,64	5041,56	5053,69	5033,72
	LL	5164,02	5172,20	5184,33	5163,88
	PL	5017,38	5025,23	5037,36	5018,07
	PP	5052,48	5060,81	5072,94	5052,83
	RPC	**5012,32**	**5019,35**	**5031,48**	**5014,12**
	RPL	5040,13	5048,56	5060,69	5040,44
	RPP	5053,74	5068,05	5080,18	5060,09

Segundo os critérios de comparação os modelos mais adequados serão aqueles que tenham menores valores nessas medidas. Assim, escolheram-se os modelos de regressão binária com funções de ligação RPC, PL e PC. Mas o melhor modelo nos 4 critérios considerados foi o RPC, isto é, o modelo de regressão binária com função de ligação cauchito de reversa de potência.

A seguir, desenvolvemos uma análise de resíduos para o modelo escolhido, em que foram considerados os resíduos discutidos na seção 3.3. Nos encontramos, como esperado, somente uma distribuição normal dos resíduos quantílicos aleatorizados com valores no intervalo [-3,6821; 3,6821] e média 0. Resultados similares foram reportados por Lemonte e Bazán (2018) indicando que os resíduos de pearson e de desvio não seguem necessariamente uma distribuição normal e por tanto apresentam dificuldades para estabelecer critérios para identificação de casos atípicos. Assim a análise de resíduo confirma a escolha da ligação RPC. Adicionalmente, na Figura 5, apresentamos O gráfico normal de probabilidades com envelope simulado considerando QQ plot para os resíduos quantílicos aleatorizados do modelo RPC. Observamos que os pontos (resíduos) se encontram dispersos aleatoriamente entre os limites do envelope o que permite verificar a adequação do modelo de regressão binária ajustado usando a ligação RPC.

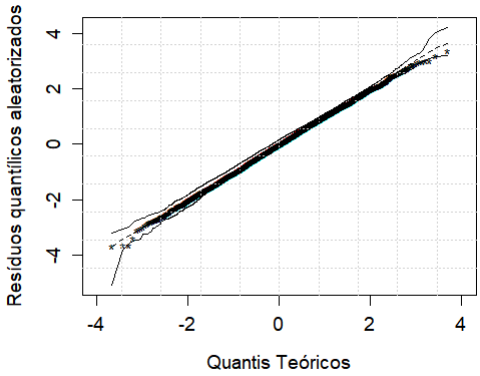

Figura 5 - Gráfico Normal de Probabilidades considerando o QQ-plot dos resíduos quantílicos aleatorizados. Regressão binária com função de ligação RPC. Dados lingüísticos.

Na tabela 7 são encontrados os valores estimados dos parâmetros dos modelos. Observa-se que todos os parâmetros são significativos pelo fato de que nenhum dos intervalos de credibilidade de 95% inclui o valor zero.

Tabela 7 - Resumo dos parâmetros estimados para as funções de ligação L e RPC: Média, Desvio Padrão (DP) e Intervalo de credibilidade (IC). Dados linguísticos

	Logito				Potência reversa cauchito			
Parâmetros	Média	DP	$IC_{2,5\%}$	$IC_{97,5\%}$	Média	DP	$IC_{2,5\%}$	$IC_{97,5\%}$
intercepto	1,21	0,04	1,13	1,3	-0,55	0,15	-0,88	-0,28
V1	0,53	0,05	0,43	0,64	0,43	0,04	0,35	0,52
V2	0,34	0,04	0,26	0,43	0,27	0,03	0,22	0,33
V3	-0,62	0,4	-0,70	-0,54	-0,45	0,03	-0,51	-0,39
V4	-0,63	0,03	-0,70	-0,56	-0,53	0,03	-0,59	-0,47
V5	-0,31	0,03	-0,37	-0,26	-0,22	0,03	-0,28	-0,16
λ					3,06	0,35	2,50	3,84

Finalmente, considerando a análise dos dados de sons e considerando o modelo escolhido e as covariáveis, tem-se o seguinte modelo final:

$$Y_i \sim Bernoulli(\hat{p}_i),\ i = 1, ..., 5404,$$

$$\hat{p}_i = 1 - \left(0,5 + \frac{arctan(-(-0,55+0,43V1_i+0,27V2_i-0,45V3_i-0,53V4_i-0,22V5_i))}{\pi}\right)^{3,06}$$

Da expressão das probabilidades estimadas considerando a equação 6, observa-se que os parâmetros associados as variáveis V1 e V2 são positivos, isto quer dizer que a probabilidade estimada aumenta, conforme há um aumento destes atributos. Por sua vez, os parâmetros associados as variáveis V3, V4 e V5 são negativos, isto quer dizer que, de acordo a este modelo, a probabilidade estimada diminui, conforme há um aumento nestes atributos. Adicionalmente, encontramos que o valor do parâmetro de forma $\lambda = 3,06$. De acordo com que foi visto no seção 3.1, como $\lambda > 1$, temos que este é um valor coerente à proporção de 1's na variável independente (maior 0,5) e consequentemente explica o desbalanceamento observado nos dados.

5.2 Aplicação a dados educacionais

No ano de 2004, o Ministério da Educação do Peru (UMC, 2004) realizou a quarta avaliação nacional do desempenho dos estudantes em língua espanhola e matemática, cujo objetivo foi proporcionar informação sobre o desempenho a nível nacional. Três séries das escolas foram avaliadas. Os dados estão disponibilizadas (⟨http://umc.minedu.gob.pe/evaluacion-nacional-2004/⟩) e correspondem a um estudo por amostragem probabilística representativa da população de estudantes. Neste trabalho, os dados usados foram da avaliação do sexto ano de educação primária do sistema educativo do Peru (equivalente à sexta série do ensino fundamental no Brasil), com meninos e meninas de 11 a 13 anos. A amostra original foi de 13804 estudantes, porém pela ocorrência de dados faltantes de uma parte dos estudantes, a amostra considerada foi de 13259.

As variáveis independentes que foram consideradas neste estudo foram as três variáveis binárias X_1 : *zona* (rural (0), urbana (1)), X_2 *gestão da escola* (pública (0), privada (1)) e X_3: *sexo* (homem (1), mulher (0)); e a variável contínua X_4 *desempenho em língua espanhola*. Esta última foi a pontuação que o estudante obteve no correspondente teste na avaliação (a pontuação obtida numa escala apresenta média 306,11 e desvio padrão 80,30). Foram escolhidas essas variáveis porque são usualmente consideradas como fatores associados ao desempenho em matemática no Peru, ver, por exemplo, o trabalho de Bazán, Espinosa e Farro (2002). A variável dependente Y é uma variável binária que indicará se o estudante teve *nível de desempenho adequado em matemática* (codificado como 1 pelo estudo) ou não adequado (codificado como 0 pelo estudo). Maiores detalhes são dados em Anyosa (2017). Dos 13259 estudantes, 9,98% deles tiveram desempenho adequado em matemática, enquanto que 90,02% deles tiveram desempenho não adequado. Pode ser visto que existe desbalanceamento nos 0's e 1's da variável resposta e consequentemente podemos aplicar os modelos revisados neste trabalho.

Na Tabela 8, são mostradas as porcentagens por níveis de desempenho (Y) segundo as covariáveis zona, gestão da escola e sexo. Pode ser visto que as porcentagens são diferentes segundo as categorias das covariáveis.

Tabela 8 - Porcentagens das variáveis categóricas. Dados educacionais

Variáveis		Nível de desempenho em matemática	
		Não adequado (0)	Adequado (1)
Zona (X_1)	Rural (1)	97,89%	2,11%
	Urbana (0)	87,92%	12,08%
Gestão (X_2)	Privada (1)	72,19%	27,81%
	Pública (0)	94,47%	5,53%
Sexo (X_3)	Homem (1)	88,91%	11,09%
	Mulher (0)	91,16%	8,84%
	Média	295,95	397,77
Língua espanhola (X_4)	Mediana	295,00	393,40
	Desvio padrão	75,42	63,12
	Mínimo	9,50	49,50
	Máximo	555,9	555,9

Adicionalmente, considerando a Tabela 8 encontramos que os estudantes com nível de desempenho adequado em matemática têm as pontuações maiores no desempenho em língua espanhola. Assim, foi visto que todas as covariáveis consideradas na análise descritiva parecem ser importantes para o estudo do desempenho dos estudantes em matemática.

Foram ajustados todos os modelos de regressão binária com ligações potência e reversa de potência apresentados na seção 2.2 no conjunto de dados educacionais. Para fazer a estimação dos modelos de regressão binária propostos, seguiu-se o procedimento considerando o algoritmo NUTS, mais detalhes em Anyosa (2017). Para o ajuste de cada um dos 9 modelos (3 usando ligações comuns, 3 usando ligações potência e cinco usando ligações reversa de potência), obteve-se uma amostra a posteriori de tamanho 8000 e retirando-se os primeiros 4000 valores no período do *burnin*, tendo 4000 valores para os quais considerou-se o espaçamento de tamanho 4 para evitar os efeitos da autocorrelação na amostra, tendo no final uma amostra válida da distribuição a posteriori de 1000 valores para cada caso.

Depois de finalizar o processo de estimação para cada modelo de regressão binária, obteve-se uma cadeia dos 1000 valores da amostra válida. Considerando essas amostras válidas, obtiveram-se as medidas de informação detalhadas na seção 2.4 para cada caso. Os resultados são apresentados na Tabela 9:

Tabela 9 - Critérios de informação dos modelos de regressão binária ajustados considerando funções de ligação potência e reversa de potência. Dados educacionais

	Ligação	DIC	EAIC	EBIC	WAIC	LOO
Simétricas	C	6655,06	6658,23	6688,20	6658,03	6658,03
	L	6199,03	6202,23	6232,20	6199,13	6199,12
	P	6160,76	6163,76	6193,73	6160,83	6160,82
	PC	8773,80	8780,40	8810,37	8773,37	8773,37
	PL	**6142,02**	**6144,79**	**6174,76**	**6143,18**	**6143,07**
	PP	**6141,94**	**6154,20**	**6184,16**	**6151,74**	**6151,73**
	RPC	6244,19	6246,35	6276,32	6245,84	6245,83
	RPL	**6137,47**	**6139,80**	**6169,77**	**6137,78**	**6137,77**
	RPP	**6148,97**	**6151,80**	**6181,77**	**6150,21**	**6150,17**

*: Problemas na estimação do parâmetro k no cálculo do LOO.

Segundo os critérios de informação considerados, os modelos mais adequados serão aqueles que tenham menores valores nessas medidas, desse modo escolherão-se os modelos de regressão binária (segundo a ordem do melhor) com funções de ligação RPL, PL, RPP e PP.

Para cada um dos quatro modelos escolhidos previamente obtiveram-se as frequências das classificações segundo a matriz de confusão e logo obtiveram-se as medidas de avaliação preditiva baseadas nas frequências das classificações das matrizes de confusão: VP, FN, FP e VN considerando-se diferentes pontos de corte para a classificação. Considerando essas medidas e um maior número de pontos de corte é possível elaborar a curva ROC e obter o AUC (ou o valor da área embaixo da curva ROC), que permite fazer uma escolha do melhor modelo de regressão binária segundo o desempenho que o modelo teve como classificador binário. Para obter as curvas ROC e as medidas do AUC apresentadas a seguir utilizou-se o pacote verification Laboratory (2015) do programa R. Os correspondentes AUC obtidos foram: PL (AUC=0,8726381), RPL (AUC=0,8726878), PP (AUC=0,8725706) and RPP (AUC=0,872634). Neste caso, os valores do AUC para os quatro modelos são próximos a 0,87, esse valor é mais perto de 1 do que de 0,5 e isso indica que todos esses modelos de regressão binária são igualmente adequados. Note-se nesse caso que esse critério não foi adequado para escolher significativamente um modelo entre o total deles. Para ilustração mostramos somente a Curva ROC do modelo RPL na figura 6.

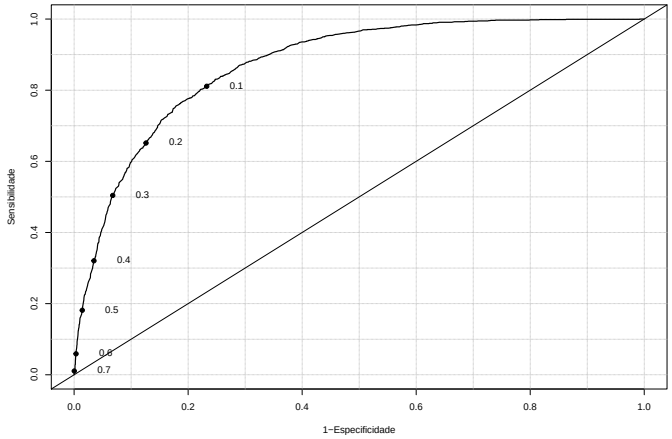

Figura 6 - Curva ROC do classificador binário considerando o modelo de regressão binária com função de ligação RPL, AUC=0,87269. Dados educacionais.

Levando em consideração os resultados dos critérios de comparação de modelos, o modelo de regressão binária com função de ligação reversa de potência logito (RPL) foi escolhido.

Na Tabela 10 são apresentados os valores resumo a posteriori dos parâmetros do modelo. Observa-se que o todos os parâmetros são significativos pelo fato de que nenhum dos intervalos de credibilidade de 95% inclui o valor zero.

Tabela 10 - Resumo a posteriori dos parâmetros do modelo com função de ligação RPL: : Média, Desvio Padrão (DP) e Intervalo de credibilidade (IC). Dados educacionais

Variáveis	Parâmetros	média	DP	mediana	$IC_{95\%}$	
Intercepto	β_0	-10,651	0,524	-10,626	-11,748	-9,696
Zona	β_1	-0,746	0,189	-0,742	-1,121	-0,378
Gestão	β_2	1,511	0,135	1,504	1,272	1,785
Sexo	β_3	0,661	0,112	0,662	0,459	0,877
Língua espanhola	β_4	0,028	0,002	0,028	0,025	0,032
Assimetria	δ	-1,687	0,148	-1,693	-1,959	-1,369
	λ	0,187	0,029	0,184	0,141	0,254

Finalmente, considerando a análise para explicar o desempenho em matemática e considerando o modelo escolhido e as covariáveis, tem-se o seguinte modelo final:

$$Y_i \sim Bernoulli(\hat{p}_i);\ i = 1, \ldots, 13259,$$

com:

$$\hat{p}_i = 1 - \left(\frac{e^{-(-10{,}651-0{,}746X_1+1{,}511X_2+0{,}661X_3+0{,}028X_4)}}{1+e^{-(-10{,}651-0{,}746X_1+1{,}511X_2+0{,}661X_3+0{,}028X_4)}} \right)^{0{,}187}.$$

Da expressão das probabilidades estimadas considerando a equação (5.2), observa-se que o parâmetro associado à variável zona (X_1) é negativo, isto quer dizer que os estudantes de zona rural têm menor probabilidade de ter nível de desempenho adequado em matemática do que os estudantes de zona urbana; adicionalmente que o parâmetro associado à variável gestão (X_2) é positivo, isto quer dizer que os estudantes de escola privada têm maior probabilidade de ter nível de desempenho adequado em matemática do que os estudantes de escola pública. Também que o parâmetro associado à variável sexo (X_3) é positivo, isto quer dizer que, de acordo com este modelo, os homens têm maior probabilidade de ter nível de desempenho adequado em matemática do que as mulheres. Da mesma maneira, o parâmetro associado a desempenho em língua espanhola (X_4) é positivo e indica que quando o estudante tiver maior pontuação nessa variável, maior será a probabilidade de ele ter nível de desempenho adequado em matemática. Finalmente, o valor do parâmetro de forma está contido em $0 < \lambda < 1$. De acordo com Anyosa (2017), temos que nesse caso $\lambda = 0,187$ é um valor coerente à proporção de 1's na variável resposta (menor que 0,5).

A interpretação anterior também pode ser analisada na Tabela 11 que inclui o valor das probabilidades estimadas segundo os possíveis valores das variáveis categóricas e níveis de desempenho em língua espanhola. Esses níveis são as médias dos valores da variável desempenho em língua espanhola categorizada segundo critérios do Ministério de Educação do Peru. Dessa forma, pode-se saber o valor das probabilidades de ter nível de desempenho adequado em matemática segundo os valores de todas as covariáveis.

Tabela 11 - Probabilidade estimada de ter nível de desempenho adequado em matemática segundo perfis de estudantes e desempenho em língua espanhola. Dados educacionais

Perfil	Zona	Gestão	Sexo	Desempenho em língua espanhola X_4			
				<Prévio	Prévio	Básico	Suficiente
	X_1	X_2	X_3	(0)	(1)	(2)	(3)
1	Rur.	Púb.	Mul.	<0,001	0,005	0,030	0,194
2	Rur.	Púb.	Hom.	0,001	0,010	0,054	0,265
3	Urb.	Púb.	Mul.	0,001	0,011	0,058	0,275
4	Rural	Priv.	Mul.	0,002	0,023	0,105	0,359
5	Urb.	Púb.	Hom.	0,002	0,021	0,097	0,384
6	Urb.	Priv.	Hom.	0,004	0,041	0,161	0,429
7	Urb.	Priv.	Mul.r	0,004	0,044	0,169	0,438
8	Urb.	Priv.	Hom.	0,008	0,077	0,238	0,501

Na Tabela 11, pode ser visto que a probabilidade de ter nível de desempenho adequado em matemática incrementa conforme o desempenho em língua espanhola incrementa, em todos os perfis. Também é visto que o perfil 1 apresenta menores probabilidades. Esse perfil corresponde a uma estudante ($X_3 = 0$) de escola pública ($X_2 = 0$) em zona rural ($X_1 = 1$). Como foi mencionado anteriormente, estudantes com essas características terão menor probabilidade de ter nível de desempenho adequado em matemática.

As diferenças no nível de desempenho em matemática segundo o sexo dos estudantes no Peru tanto de ensino fundamental como de ensino médio foi estudada anteriormente em Bazán, Espinosa e Farro (2002). As conclusões nesse trabalho são coerentes com o que foi encontrado nesta aplicação, com resultados favoráveis ao sexo masculino. Para entender esse resultado, várias pesquisas sugerem o rol da socialização do gênero (ECCLES; JACOBS, 1986) ou o fato de que as mulheres experimentam maior ansiedade na matemática do que os homens (MEECE; WIGFIELD; ECCLES, 1990). Além disso, estudos recentes como o de Devine et al. (2012) não encontraram diferenças no desempenho relacionado ao sexo apesar da diferença de ansiedade, concluindo que as mulheres poderiam ter melhor desempenho do que os homens se a ansiedade pudesse ser atenuada. Além do mais, no trabalho de Bieg et al. (2015) concluiu-se que devido a estereótipos sobre diferenças no desempenho em matemática segundo o sexo, as estudantes mulheres podem acreditar que estão mais ansiosas do que realmente estão em relação à matemática. Esse trabalho enfatizou a tarefa dos professores e das famílias dos estudantes em reduzir esses estereótipos. Por outra parte, no trabalho de Bassi et al. (2018) analisaram-se escolas de baixo desempenho no Chile, onde existem grandes diferenças no desempenho em matemática segundo o sexo. Os autores concluíram que os profesores dão mais atenção aos homens nas aulas, favorecendo o melhor

desempenho deles (e desfavorecendo às mulheres). Isso significa que eventuais conclusões baseadas nos resultados apresentados neste trabalho deverão levar em conta os trabalhos citados acima. Dado que foram utilizados dados obtidos numa amostragem probabilística, os resultados são generalizáveis para estudantes da sexta série nas escolas do Peru daquele ano.

6 Conclusões

Para o presente estudo foi feita uma revisão dos modelos de regressão binária para as funções de ligação comuns e para as funções de ligação assimétricas potência e reversa de potência sob abordagem bayesiana. Um código detalhado usando os pacotes `R` e `Stan` são disponibilizadas no Apêndice C tornado fatível o uso deste tipo de modelos em problemas de classificação.

Um estudo de simulação foi desenvolvido para avaliar o efeito do desbalanceamento na estimação dos parâmetros do modelo logístico, em que foram usadas a medidas de comparação vicio (viés), erro quadrático médio, tempo computacional e probabilidade média. A partir disso, foi verificado que há um viés maior na estimativas do intercepto (β_0) do que nas estimativas do coeficiente angular (β_1) quando são ajustados dados desbalanceados com o modelo logístico. Apesar do viés diminuir conforme o tamanho da amostra aumenta, ele ainda apresenta um valor alto, especialmente para β_0. Estes resultados sinalizam que ajustar um modelo logístico não é adequado quando os dados são desbalanceados. Ao mesmo tempo, mediante o estudo de simulação, nós mostramos que os parâmetros de regressão binária com resposta desbalanceada que são gerados do modelo com ligação de potência, são recuperados adequadamente, apresentando pouco viés e, ainda, estimamos um parâmetro adicional λ que está associado com a taxa de desbalanceamento.

Considerando diferentes aplicações, nos mostramos que o modelos estudados são mais adequados para dados adicionais. Além disso, mostramos que diferentes possibilidades podem ser consideradas para avaliar e determinar o melhor modelo para os dados. A análise de resíduos considerando o resíduo quantílico aleatorizado se mostrou mais satisfatório do que outros resíduos propostos e, também, outra análise considerando diferentes critérios de comparação de modelos, permitiu escolher o melhor modelo para os dados. Ainda, considerando os resultados das aplicações, nós notamos que o uso de medidas tradicionais de avaliação preditiva e a área abaixo da curva ROC para os diferentes modelos apresentam valores muito próximos e, portanto, não conseguem mostrar que os modelos com as ligações estudadas são melhores ajustados do que o modelo logístico, assim como é mostrado com os critérios de comparação de modelos. Esse comportamento já tem sido reportado por Bazán, Torres-Aviles, Suzuki e Louzada (2017). Isso pode ser explicado, pois a medida de sensibilidade e especificada são medidas para uma tabela simétrica, e no caso de dados desbalanceados temos uma tabela desigual. Para contornar essa situação de la Cruz et al. (2019) recomendam o uso de medidas de concordância assimétricas para casos de desbalanceamento.

Como desdobramento futuro é possível pensar no uso dos modelos de regressão binária potência e reversa de potência em problemas de classificação binária para quando os dados são desbalanceados, desenvolvendo comparações frente a outros métodos de classificação. Também, extensões para o caso misto e ordinal podem ser consideradas futuramente. Adicionalmente poderia-se realizar estudos baseados em problemas com diferentes proporções de desbalanceamento no intuito de avaliar o ajuste dos modelos de regressão assimétricas estudados nesse trabalho.

Agradecimentos

Aos revisores e editores pelos comentários e sugestões.

SILVA, A. N.; ANYOSA, S.; BAZÁN, J. L. Bayesian binary regression modeling for unbalanced data using new links. *Rev. Bras. Biom.,* Lavras, v.38, n.4, p.385-417, 2020.

- *ABSTRACT: In this work, we presented, in a didactic way, the Bayesian binary regression modeling for unbalanced data using new links functions. Under the Bayesian approach and using information criteria, predictive evaluation measures and introducing the analysis of residuals, we show that the models that use power and reverse power link functions are better than traditional models in the presence of unbalanced data, considering two applications. Additionally, codes with the procedures presented using the Stan package are made available in order to facilitate the use of these models. The work also contains a simulation study that shows how the unbalance in the response variable affects the estimation of the parameters of a logistic regression with respect to the bias, mean square error and standard deviation of the estimates, regardless of the sample size. At the same time, considering two applications, we show how binary regression models with the power and reverse power links recently formulated in the literature can be used to adequately estimate the parameters in the type of unbalance considered.*

- *KEYWORDS: Unbalancing data; asymmetrical link; bayesian estimation; binary regression; residuals.*

Referências

AGRESTI, A. *Categorical data analysis, xvi 714*. London: John Wiley, 2013.

ALLISON, P. Logistic regression for rare events. *Statistical Horizons*, v.13, 2012.

ANYOSA, S. C.; BAZÁN, J. L.; LEMONTE, A. *powdist: Power and Reversal Power Distributions*, 2017. R package version 0.1.4.

ANYOSA, S. C.*Binary regression using power and reversal powerlinks.* Master thesis (in Portuguese). Interinstitutional Graduate Program in Statistics. Universidade de São Paulo - Universidade Federal de São Carlos.

BASSI, M. ; DIAZ, M. M. ; BLUM, R. L. ; REYNOSO, A. Failing to notice? Uneven teachers's attention to boys and girls in the classroom. *IZA Journal of Labor Economics*, v.7, n.1, p.9, 2018.

BAZÁN, J.; BAYES, C. Inferencia Bayesiana en modelos de regresion binaria usando BRMUW. *Reporte de Investigacion. Serie B. Nro, v.25, 2010. Disponível em: http://argos.pucp.edu.pe/ jlbazan/download/Reporte-25.pdf. Acesso em: 29/08/2020*

BAZÁN, J. L.; ESPINOSA, G.; FARRO, C. Rendimiento y actitudes hacia la matemática en el sistema escolar peruano. MECEP. Programa Especial Mejoramiento de la Calidad de la Educación Peruana, 2002. Disponível em: https://jorgeluisbazan.weebly.com/uploads/1/2/5/6/125695412/13i.pdf. Acesso em: 29/08/2020.

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

BAZÁN, J. L.; TORRES-AVILÉS, F.; SUZUKI, A. K.; LOUZADA, F. Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, v.33, n.1, p.22–34, 2017.

BIEG, M. ; GOETZ, T.; WOLTER, I.; HALL NC. Gender stereotype endorsement differentially predicts girls' and boys' trait-state discrepancy in math anxiety. *Frontiers in psychology*, v.6, p.1404, 2015.

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*, v.94, n.448, p.1172–1186, 1999.

DE LA CRUZ, A.; BAZÁN, J. L.; CANCHO, V. G; DEY, D. Performance of asymmetric links and correction methods for imbalanced data in binary regression. *Journal of Statistical Computation and Simulation*, v.89, p.1694-1714, 2019.

DEVINE, A. ; FAWCETT, K.; SZUCS, D.; DOWKER, A. Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and brain functions*, v.8, n.1, p.33, 2012.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v.5, n.3, p.236–244, 1996.

ECCLES, J. S.; JACOBS, J. E. Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society*, v.11, n.2, p.367–380, 1986.

FIRTH, D. Bias reduction of maximum likelihood estimates. *Biometrika*, v.80, n.1, p.27–38, 1993.

GARETH J. ; WITTEN, D.Ñ .; HASTIE, T.; TIBSHIRANI, R. *An Introduction to statistical learning: with applications in R*. New York: Springer, 2013.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v.143, n.1, p.29–36, 1982.

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

JIANG, X. et al. A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics*, v.7, n.4, p.2180–2204, 2013.

KING, G.; ZENG, L. Logistic regression in rare events data. *Political analysis*, v.9, n.2, p.137–163, 2001.

LABORATORY, N. R. A. *verification: Weather Forecast Verification Utilities*. CO, USA, 2015. R package version 1.42. Disponível em: ⟨https://CRAN.R-project.org/package=verification⟩.

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.

MEECE, J. L.; WIGFIELD, A.; ECCLES, J. S. Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of educational psychology*, v.82, n.1, p.60, 1990.

PAAL, B. V. A comparison of different methods for modelling rare events data. *Ghent University*, 2014.

PAULA, G. A., *Modelos de regressão com apoio computacional*. São Paulo: IME-USP. 2013. Disponível em: https://www.ime.usp.br/ giapaula/texto-2013.pdf. Acesso em: 29/08/2020.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, v.54, n.3, p.507–554, 2005.

STAN. *Stan modeling language users guide and reference manual*. [S.l.], 2017. Version 2.17.0. http://mc-stan.org.

STAN DEVELOPMENT TEAM. *RStan: the R interface to Stan*. 2018. R package version 2.17.3. Disponível em: ⟨http://mc-stan.org/⟩.

UMC. *Evaluación Nacional 2004 na responsabilidade da Oficina de Medición de la Calidad de los Aprendizajes do Ministério de Educação do Peru*. 2004. Disponível em: ⟨http://umc.minedu.gob.pe/evaluacion-nacional-2004/⟩. Acesso em: 29/08/2016.

VAN CAPPEL, D.; THOMSON-SINTRA. Phoneme data set, 1993. 1993. Disponível em: https://datahub.io/machine-learning/phoneme. Acesso em: 29/08/2020.

WALTHER, B. A.; MOORE, J. L. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, Munksgaard International Publishers, v.28, n.6, p.815–829, 2005. ISSN 1600-0587.

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.51, 2010.

Recebido em 15.10.2019.

Aprovado após revisão em 23.07.2020.

Apêndice

A. Distribuições potência e reversa de potência

Com base em Bazán, Torres-Aviles, Suzuki e Louzada (2017) temos que a variável aleatória (v.a.) univariada Z segue distribuição potência padrão $\mathrm{Z} \sim P(\lambda)$, em que $\lambda > 0$ é o parâmetro de forma se sua função de densidade de probabilidade (fdp) é dada por : $f_P(z|\lambda) = \lambda \mathrm{g}(z) [\mathrm{G}(z)]^{\lambda-1}$ e $F_P(z) = [\mathrm{G}(z)]^{\lambda}$ respectivamente, onde g(.) é uma fdp unimodal, log côncava de suporte real e $G(.)$ é a fda de g(.), chamada de distribuição base que pode ser qualquer fda simétrica ou assimétrica. Adicionalmente, se $W \sim P(\lambda)$, podemos definir uma v.a $Z = -W$ denominada distribuição reversa de potência padrão e escrevemos $Z \sim RP(\lambda)$. Neste caso a fdp e fda da correspondente distribuição reversa de potência é dada por: $f_{RP}(z|\lambda) = \lambda \mathrm{g}(z) [\mathrm{G}(-z)]^{\lambda-1}$ e $F_{RP}(z) = 1 - [\mathrm{G}(-z)]^{\lambda}$ respectivamente.

Em Bazán, Torres-Aviles, Suzuki e Louzada (2017) é visto que $Z \sim F_p(.)$ satisfaz a propriedade de reversibilidade com $F_{RP}(.)$ sendo sua distribuição reversa e viciversa. Além disso Bazán, Torres-Aviles, Suzuki e Louzada (2017) relatam que F_P e F_{RP} não são ponto-simétricas uma vez que $F_P(-z) \neq 1 - F_P(z)$ ou $F_{RP}(-z) \neq 1 - F_P(z)$ para $\lambda \neq 1$ e que ambas satisfazem $F_P(\pm z) + F_{RP}(\mp z) = 1$ o que quer dizer que as duas distribuições são diferentes mas estão relacionadas porque uma é reversa da outra. Finalmente, se $\lambda = 1$ então $F_P(z) = G(z) = F_{RP}(z)$. Isto quer dizer que $G(.)$ é um caso particular das duas distribuições ou distribuição base. A propriedade de reversibilidade nos permite propor outras distribuições potência e reversa de potência. Estas distribuições e sua versão de locação escala $Y = \mu + \sigma Z$, atualmente estão disponíveis no pacote `powdist` do `R` (ANYOSA; BAZÁN; LEMONTE, 2017). O pacote também proporciona funções para a simulação de dados segundo as diferentes distribuições potência e reversa de potência.

Nesse trabalho iremos considerar as versões de potência e reversa de potência para as distribuições comuns (logística, normal, cauchito) que dão origem as funções de ligação comuns. As correspondentes funções de distribuição acumuladas e de probabilidade são apresentadas em (ANYOSA; BAZÁN; LEMONTE, 2017)

B. Códigos da linguagem R para as figuras.

Para obter o gráfico de outra distribuição power é necessário mudar ppcauchy, referente a distribuição power cauchy, pelo nome de outra das distribuições implementadas no pacote powdist. Veja `help("powdist")`.

```
if(!require(powdist)){install.packages("powdist");  require(powdist)}
if(!require(ggplot2)){install.packages("ggplot2");  require(ggplot2)}

eta=seq(-10,10,by=0.01)

dat = rbind(data.frame(eta,p=pplogis(eta),Lambda=rep("logit",NROW(eta))),
data.frame(eta,p=ppnorm(eta),Lambda=rep("probit",NROW(eta))),
data.frame(eta,p=ppcauchy(eta),Lambda=rep("cauchit",NROW(eta))),
data.frame(eta,p=prgumbel(eta),Lambda=rep("cloglog",NROW(eta))),
data.frame(eta,p=pgumbel(eta),Lambda=rep("loglog",NROW(eta))))

pdf("links.pdf",width=11,height=8.5,paper='special')
ggplot(dat) +  geom_line(aes(eta,p,linetype=Lambda)) + theme_light() +
labs(x=expression(eta[i]),y=expression(p[i]),linetype=NULL) +
```

```
theme(axis.title.y = element_text( margin = unit(c(0, 3, 0, 0), "mm")),
axis.title.x = element_text(margin = unit(c(3, 0, 0, 0), "mm")),
legend.text = element_text(size = 12), axis.title = element_text(size = 12.5),
legend.title = element_text(size=12), plot.title = element_text(size = 14),
axis.text = element_text(size = 12), legend.key = element_rect(colour = "gray"))
dev.off()

eta=seq(-5,5,by=0.01)

help("powdist")
link = ppcauchy

dat = rbind(
data.frame(eta,p=link(q = eta, lambda = 1/5),Lambda=rep("1/5",NROW(eta))),
data.frame(eta,p=link(q = eta, lambda = 1/3),Lambda=rep("1/3",NROW(eta))),
data.frame(eta,p=link(q = eta, lambda = 1),Lambda=rep("1",NROW(eta))),
data.frame(eta,p=link(q = eta, lambda = 2),Lambda=rep("2",NROW(eta))),
data.frame(eta,p=link(q = eta, lambda = 3),Lambda=rep("3",NROW(eta))))

pdf("pcauchy.pdf",width=4,height=4,paper='special')
ggplot(dat) +  geom_line(aes(eta,p,linetype=Lambda)) + theme_light() +
labs(x=expression(eta[i]),y=expression(p[i]),linetype=expression(lambda))+
scale_linetype_manual(values=c(2,3,1,4,5)) +
theme(axis.title.y = element_text( margin = unit(c(0, 3, 0, 0), "mm")),
axis.title.x = element_text(margin = unit(c(3, 0, 0, 0), "mm")),
legend.text = element_text(size = 12), axis.title = element_text(size = 12.5),
legend.title = element_text(size=12), plot.title = element_text(size = 14),
axis.text = element_text(size = 12),
legend.key = element_rect(colour = "gray"))
dev.off()
```

C. Códigos Stan do modelo Potência Cauchito.

```
#Modelo reversa potencia cauchito em Stan
modelo_string <- data{ int<lower=0> k; int<lower=0> n; int<lower=0, upper=1> y[n]; matrix[n,k] X;
}
parameters{
real beta0; vector[k] beta; real loglambda;
}
transformed parameters  {
real lambda; vector[n] p; vector[n] prob; lambda = exp(loglambda);
for (i in 1:n) {
p[i] = student_t_cdf(beta0 + X[i]*beta, 1, 0, 1);
prob[i] = pow(p[i],lambda);
}
}
model {
beta0 ~ normal(0.0,100); beta ~ normal(0.0,100); loglambda ~ uniform(-2,2);y ~ bernoulli(prob);
}
```