FACULTAD DE INGENIERÍA
REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Novos modelos de regressão binária usando funções de ligação simétricas e assimétricas

Tatiana F. Matta[1], George A. A. Santos[1], Francisco Louzada[2], Anderson Ara[3], and Paulo H. Ferreira[1]

[1]Departamento de Estatística, IME, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, CEP: 40.170-110, Salvador, BA, Brasil; *tatifmatta@hotmail.com*, *george_13031995@hotmail.com*, *paulohenri@ufba.br*.
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Avenida Trabalhador São-Carlense 400, Centro, CEP: 13.566-590, São Carlos, SP, Brasil; *louzada@icmc.usp.br*.
[3]Departamento de Estatística, Universidade Federal do Paraná, Rua Evaristo F. F. da Costa 393, Jardim das Américas, CEP: 81.531-980, Curitiba, PR, Brasil; *ara@ufpr.br*.

RESUMO

Os modelos de regressão com variáveis respostas binárias (1 - ocorrência do evento de interesse ou "sucesso", 0 - não ocorrência do evento de interesse ou "fracasso") têm sido aplicados intensamente em diversas áreas do conhecimento, como saúde, finanças, indústria, entre outras. Tradicionalmente, o modelo mais usado na regressão binária tem sido o modelo de regressão logística. Contudo, ele utiliza a função de ligação *logit* (ou logito), a qual é uma função de ligação simétrica e pode não ser adequada em determinadas situações, por exemplo, quando uma das classes da variável resposta é desbalanceada em relação à outra (conjunto de dados desbalanceados). Este trabalho tem como objetivo principal apresentar novos modelos de regressão binária usando funções de ligação simétricas e assimétricas. A estimação dos parâmetros dos modelos descritos neste trabalho (a saber: modelos de regressão binária double Lindley, double Lindley assimétrica, potência double Lindley e reversa de potência double Lindley) é feita pelo método clássico da máxima verossimilhança. Para comparação e seleção do "melhor" modelo dentre as diferentes distribuições, são empregados critérios de informação (AIC e BIC) e medidas de avaliação da capacidade preditiva (AUC, acurácia balanceada, sensibilidade, F1-*Score*, coeficiente de correlação de Matthews, dentre outras). Os resultados da análise de um conjunto de dados reais, referente a uma competição promovida pelo Banco Santander para a comunidade do *Kaggle*, e de um conjunto de dados sintéticos, que reflete dados reais de manutenção preditiva encontrados no meio industrial e também disponível no *Kaggle*, mostram que os modelos com as funções de ligação propostas podem apresentar melhor ajuste e performance preditiva do que os modelos usando ligações tradicionais, como a logito.

Palavras-chaves: dados desbalanceados; distribuição double Lindley; ligações potência e reversa de potência; método de máxima verossimilhança; performance preditiva.

ABSTRACT

Regression models with binary response variables (1 - occurrence of the event of interest or "success,"0 - non-occurrence of the event of interest or "failure") have been widely applied in various fields, such as health, finance, industry, and others. Traditionally, the most commonly used model in binary regression has been the logistic regression model. However, it employs the *logit* (or logistic) link function, which is a symmetric link function and may not be suitable in certain situations, for instance, when one of the response variable classes is imbalanced relative to the other (imbalanced datasets). The main objective of this work is to introduce new binary regression models using both symmetric and asymmetric link functions. Parameter estimation for the models described in this work (namely, double Lindley binary regression, asymmetric double Lindley, power double Lindley, and reversed power double Lindley) is performed using the classical maximum likelihood method. To compare and select the "best"model among the different distributions, information criteria (AIC and BIC) and predictive performance measures (AUC, balanced accuracy, sensitivity, F1-Score, Matthews correlation coefficient, among others) are employed. The results from the analysis of a real dataset, from a competition hosted by Banco Santander for the *Kaggle* community, and a synthetic dataset, which reflects real predictive maintenance data found in the industrial context and also available on *Kaggle*, demonstrate that models with the proposed link functions can achieve better fit and predictive performance compared to models using traditional links, such as the logistic link.

Keywords: "Imbalanced data; double Lindley distribution; power and reversed power link functions; maximum likelihood method; predictive performance."

1 Introdução

Os problemas de classificação de dados são bastante comuns em diversas áreas do conhecimento, como indústria, saúde e finanças. Muitas vezes, a variável de interesse, denominada variável dependente ou resposta, é considerada binária e assume apenas uma de duas categorias (níveis ou classes) possíveis. Como, por exemplo, o *status* de um item produzido (bom estado ou defeituoso), a remissão de uma doença (sim ou não), o resultado de um tratamento (bom ou ruim), se um cliente se tornará inadimplente (sim ou não), dentre outros. Nestes casos, o problema de classificação é dito ser de classificação binária ou dicotômica. Frequentemente, um conjunto de variáveis que influenciam a resposta de interesse, chamadas de variáveis independentes ou explicativas (ou ainda, covariáveis), também está disponível. Essas variáveis podem ser tanto qualitativas (sexo, raça, estado civil, grau de escolaridade etc.) quanto quantitativas (idade, altura, peso, renda etc.).

Dentre as ferramentas de modelagem estatística que têm sido amplamente utilizadas no auxílio à tomada de decisões em situações como as descritas anteriormente (isto é, qual tratamento escolher, se irá conceder ou não o crédito solicitado etc.), estão os modelos de regressão (ou classificação) binária. Nos modelos de regressão em geral, o objetivo principal é descrever uma possível relação existente entre a variável resposta Y e as covariáveis $X_1, \ldots, X_p$, com $p \geq 1$. Nos modelos de regressão binária, a variável resposta Y é dicotômica, isto é, permite somente dois resultados, aos quais atribuímos convencionalmente o valor 1 para a ocorrência do evento de interesse ("sucesso"), e o valor 0 para a não ocorrência do evento de interesse ("fracasso"). Estudos mais detalhados sobre a regressão binária podem ser encontrados em Cox & Snell [1] e Collett [2], dentre outros.

O modelo de regressão logística binária (ou simplesmente, modelo de regressão logística) é conhecido desde os anos 1950. No entanto, estudos iniciais sobre esse modelo foram publicados nos artigos de Verhulst [3, 4, 5]. Bliss [6], em seu artigo, estudou experimentos biológicos do tipo dose-resposta para doses fixas e respostas aleatórias que refletiam a distribuição individual de níveis de tolerância. Além disso, aplicações do modelo de regressão logística em áreas como economia e pesquisa de mercado surgiram nos anos 1950 e 1960 (Farrell [7], Aitchison & Brown [8], Adam [9]). A partir da década de 1980, torna-se mais utilizado com o trabalho de Cox & Snell [1]. Esse modelo tem sido aplicado em várias áreas do conhecimento

para a análise de dados binários, pois não precisa atender alguns pressupostos, como a igualdade das matrizes de covariância e a normalidade dos erros. Ele traz, ainda, a vantagem da facilidade de interpretação dos parâmetros, conforme apontam Hosmer *et al.* [10].

Na regressão logística, a função de ligação empregada é a *logit* (ou logito), que é uma função de ligação simétrica, resultante da função de distribuição acumulada (FDA) da distribuição logística padrão. Outras funções de ligação simétricas são comumente usadas na regressão binária, como, por exemplo, a *cauchit* (ou cauchito) e a *probit* (ou probito), que resultam das FDAs das distribuições Cauchy padrão e normal padrão, respectivamente.

Considere n observações de uma variável aleatória independente $(Y_1, Y_2, \ldots, Y_n)$ com distribuição de Bernoulli, com probabilidade de sucesso μ_i e probabilidade de fracasso $1 - \mu_i$, para $i = 1, 2, \ldots, n$. Para associar a probabilidade μ_i com as covariáveis $X_{i1}, \ldots, X_{ip}$, para $i = 1, 2, \ldots, n$ e $p \geq 1$, as funções de ligação mais empregadas são a logito e a probito (ambas simétricas); entretanto, elas podem não se legitimar caso os dados (isto é, as classes da variável resposta) sejam desbalanceados e levar a conclusões erradas. Neste contexto, Czado & Santner [11] pesquisaram os efeitos da má especificação na função de ligação, concluindo que supor a logito como função de ligação no ajuste quando a função de ligação dos dados é outra, traz viés na estimação dos parâmetros da regressão e nas probabilidades preditas.

De fato, quando lidamos com dados binários, é frequente a presença de uma variável resposta cujo "sucesso" é pouco provável de ocorrer, ou seja, temos um evento raro (amostra ou base de dados desbalanceados). Neste caso, os estimadores de máxima verossimilhança podem não fornecer resultados satisfatórios (boas estimativas) para os parâmetros (ou coeficientes) do modelo de regressão, nem produzir boas predições.

Contudo, as abordagens discutidas anteriormente não são as únicas possíveis e, dada a importância de se identificar os eventos raros, é imprescindível que se entenda melhor como obter um modelo com bom poder preditivo frente às limitações de um banco de dados com tais características, seja na área da saúde, indústria, finanças ou em qualquer outra área.

1.1 Objetivo

O objetivo deste trabalho é desenvolver novos modelos de regressão binária usando funções de ligação simétricas e assimétricas. Uma vez que informações a respeito de tais modelos não foram encontradas em pesquisas na literatura, acredita-se que todos eles sejam realmente inéditos. A metodologia proposta (novos modelos de regressão binária, métodos de estimação e validação, medidas de comparação de modelos e avaliação da performance preditiva) é ilustrada em um conjunto de dados reais, que diz respeito a dados do Banco Santander (*Santander Customer Transaction*), no qual o objetivo é prever se um cliente irá efetivar uma operação financeira específica futuramente, independente da quantia transacionada; e também em um conjunto de dados sintéticos, que reflete dados reais de manutenção preditiva encontrados no meio industrial (*Machine Predictive Maintenance*), cujo objetivo é predizer se um equipamento rotativo apresentará falha no futuro, sob determinadas condições apresentadas. Além disso, foram realizados estudos de simulação para avaliar, por exemplo, a performance do método de estimação considerado (máxima verossimilhança), bem como a adequação das medidas de comparação/seleção de modelos utilizadas (AIC e BIC).

1.2 Organização do Trabalho

Este trabalho está estruturado em cinco seções, dispostas da seguinte forma. Na Seção 2 são apresentados os modelos de regressão binária propostos, a saber: modelos de regressão double Lindley, double Lindley assimétrica, e os novos modelos potência double Lindley e reversa de potência double Lindley. Na Seção 3 são discutidos os principais resultados dos estudos de simulação (recuperação de parâmetros e *misspecification*). Na Seção 4 são exibidos os principais resultados da aplicação da metodologia proposta a um conjunto de dados reais, oriundo da área financeira; e a um conjunto de dados sintéticos, obtido a partir de um conjunto de dados reais oriundo da área industrial. As considerações finais deste trabalho estão na Seção 5.

2 Modelos de Regressão para Dados Binários

Nesta seção são apresentados o modelo de regressão binária usual (regressão logística), bem como detalhes acerca das distribuições de probabilidade (double Lindley, double Lindley assimétrica, potência double Lindley e reversa de potência double Lindley) consideradas nos modelos propostos neste trabalho. São também descritos o procedimento de estimação clássica pelo método da máxima verossimilhança e as diferentes medidas de comparação de modelos, incluindo os critérios de informação (AIC e BIC) e as métricas de avaliação da performance preditiva.

2.1 Regressão Binária

Os modelos de regressão binária têm sido aplicados intensamente em várias áreas do conhecimento, tais como indústria (Pacagnella *et al.* [12]), saúde (Silva *et al.* [13]), finanças (Ritta *et al.* [14]) etc. Eles são indicados quando a variável resposta é dicotômica (= 1 para o evento ou resultado de interesse - "sucesso"; = 0 para o evento complementar - "fracasso"). A variável resposta está geralmente associada a outras variáveis, que podem ser discretas, contínuas ou categóricas. Sendo que a probabilidade de "sucesso" pode ser explicada por essas variáveis, denominadas variáveis explicativas ou covariáveis.

Seja $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ um vetor $n \times 1$ de variáveis aleatórias respostas, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ um vetor $k \times 1$ de covariáveis associadas a Y_i, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ um vetor $k \times 1$ de coeficientes de regressão associados às covariáveis e, finalmente, considere a probabilidade de "sucesso" $P(Y_i = 1) = \mu_i$ e a probabilidade de "fracasso" $P(Y_i = 0) = 1 - \mu_i$, para $i = 1, 2, \ldots, n$. Logo, no modelo de regressão binária, Y_i segue uma distribuição de Bernoulli com parâmetro μ_i, conforme especificado a seguir:

$$Y_i \sim \text{Bernoulli}\left(\mu_i\right), \quad \text{para } i = 1, 2, \ldots, n,$$

$$\mu_i = P\left(Y_i = 1\right) = F\left(\eta_i\right),$$

$$\eta_i = F^{-1}\left(\mu_i\right),$$

em que $F(\cdot)$ é a FDA correspondente, $F^{-1}(\cdot)$ é a função de ligação e $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.

Observe que a função de ligação $F^{-1}(\cdot)$ transforma o intervalo $(0, 1)$, isto é, o suporte de μ_i (a média de Y_i), para a linha dos reais. Assim, quando o preditor linear η, com valores nos reais, é avaliado na FDA $F(\cdot)$ (também referida como função de ligação inversa), os resultados obtidos são valores de probabilidade válidos, que estão entre 0 e 1. A função de ligação logito, por exemplo, é obtida da inversa da FDA da distribuição logística padrão. Assim, o modelo de regressão logística é determinado pela seguinte relação:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}.$$

2.2 Dados Desbalanceados

Na regressão binária com ligações simétricas, que são as mais comumente usadas, sendo elas originárias da inversa da FDA padrão de distribuições simétricas (por exemplo, normal, logística, Cauchy, t-Student e Laplace, dentre outras), um obstáculo difícil existe quando uma das classes da variável resposta é desbalanceada em relação à outra. Segundo Van der Paal [15], um conjunto de dados é considerado desbalanceado (ou com raridade relativa) quando a classe de interesse (evento) é consideravelmente menor do que a outra classe (não evento), ou seja, o número de observações referente a uma classe é desproporcional ao número de observações da outra classe.

O desbalanceamento dos dados influencia a estimação dos parâmetros do modelo de regressão binária, no que diz respeito ao vício, erro padrão e erro quadrático médio das estimativas obtidas, assim como as predições das métricas de desempenho, como, por exemplo, a acurácia (ou taxa de acerto geral). Por isso, quando são considerados dados desbalanceados, é possível incidir em um erro, visto que o acerto geral pode estar próximo à totalidade de 100%, entretanto, o acerto na classe de interesse pode ficar próximo de zero.

Considerando modelos de classificação para dados desbalanceados, segundo Ali *et al.* [16], uma das abordagens que vem sendo utilizada é trabalhar com o nivelamento dos dados, que consiste em técnicas de reamostragem dos dados visando a redução do desbalanceamento. Uma dessas técnicas é o *undersampling* (ou subamostragem), que trata-se do processo de redução da quantidade de eventos em maior número, de forma

que o conjunto de dados alcance uma quantidade de registros considerável para resolução do problema em questão. Outro método é o *oversampling* (ou superamostragem), que consiste no incremento da classe minoritária, considerando um processo de reamostragem da classe até que se alcance o resultado desejado. Com a utilização desta abordagem, pode-se realizar desde um processo de reamostragem aleatória simples do banco de dados até métodos mais elaborados, como o SMOTE (*Synthetic Minority Oversampling Technique*) proposto por Chawla *et al.* [17]. Outro método citado na literatura (que é baseado no SMOTE) é o ADASYN (*Adaptive Synthetic Sampling*), proposto por He *et al.* [18], que consiste na utilização de uma distribuição de densidade como um critério de decisão sobre a quantidade de dados sintéticos que devem ser gerados.

Quando o banco de dados possui grande número de observações, a utilização de *undersampling* com amostragem aleatória simples é muito comum no mercado de trabalho brasileiro, uma vez que é de fácil explicação e reduz o potencial de *overfitting* que, segundo Padhi *et al.* [19], ocorre quando uma função explica muito bem um determinado conjunto de observações, mas não possui a capacidade de generalização para outros bancos de dados com características similares.

Outra abordagem que vem sendo empregada, segundo Ali *et al.* [16], é uma metodologia de predição que, por natureza de sua criação, pode gerar resultados satisfatórios para dados desbalanceados, sem necessidade de realização de amostragens, como é o caso, por exemplo, do algoritmo SVM (*Support Vector Machines*). Entretanto, quando o desbalanceamento é alto, o desempenho do classificador SVM pode ser afetado (i.e., decrementado) significativamente (Tao *et al.* [20]; Imam *et al.* [21]).

Este trabalho traz algo que faz um contraponto às técnicas mencionadas anteriormente. O objetivo é construir novos modelos de classificação binária que possuam interpretabilidade, isto é, que garantam uma maior flexibilidade (em relação aos modelos estatísticos tradicionais, como o de regressão logística, por exemplo) preservando a sua facilidade de interpretação, o que não se encontra em alguns desses algoritmos de aprendizado de máquina.

2.3 Estimação por Máxima Verossimilhança

A estimação dos parâmetros do modelo de regressão binária é feita, usualmente, com a aplicação de distribuições de probabilidade na modelagem de variáveis aleatórias, com o objetivo de estimar quantidades populacionais desconhecidas. De acordo com Cordeiro & Demétrio [22], muitos métodos podem ser empregados para estimar os parâmetros $\beta_1, \ldots, \beta_k$, incluindo o método dos mínimos quadrados, o Bayesiano e o da máxima verossimilhança (MV), o qual possui várias propriedades ótimas, tais como consistência e eficiência assintótica.

Neste trabalho, considera-se o método de MV para estimar os parâmetros lineares $\beta_1, \ldots, \beta_k$ do modelo de regressão binária. Com a suposição de independência dos valores de Y_i, para $i = 1, 2, \ldots, n$, a função de verossimilhança para esses parâmetros é dada por:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} {\mu_i}^{y_i}(1-\mu_i)^{1-y_i}.$$

Como os valores que maximizam a função de verossimilhança acima são os mesmos que maximizam seu logaritmo, então pode-se escrevê-la da seguinte forma:

$$\begin{aligned}\ell(\boldsymbol{\beta}) = \log\left(\mathcal{L}(\boldsymbol{\beta})\right) &= \sum_{i=1}^{n} \log\left(\mu_i^{y_i}(1-\mu_i)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} y_i \log(\mu_i) + \sum_{i=1}^{n}(1-y_i)\log(1-\mu_i).\end{aligned}$$

Quando as condições de regularidade estão satisfeitas, segundo Casella & Berger [23], o máximo global da função $\ell(\boldsymbol{\beta})$ é encontrado, unicamente, pelas soluções da expressão:

$$\frac{\partial \log(\mathcal{L}(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

De modo que o estimador de MV $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ seja obtido pela solução do sistema de k equações que torna o vetor escore igual a zero, ou seja,

$$U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}, \tag{1}$$

em que:

$$U_j(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad \text{para } j = 1, \ldots, k.$$

Em geral, quando não existem soluções exatas (ou analíticas) para a equação (1), elas têm que ser obtidas numericamente por meio de processos iterativos, como, por exemplo, Newton-Raphson e BFGS (Broyden-Fletcher-Goldfarb-Shanno). Neste trabalho, para a estimação dos parâmetros pelo método de MV foi utilizada a função $maxLik(\cdot, method = \text{“}BFGS\text{”})$ do pacote de mesmo nome (ver Henningsen & Toomet [24]) do programa estatístico R, versão 3.6.3 (R Core Team [25]).

2.4 Distribuições de Probabilidade Simétricas e Assimétricas

Nos últimos anos, vem crescendo o interesse de pesquisadores em estudar diferentes famílias de distribuições paramétricas simétricas, bem como suas versões assimétricas. De acordo com Nitha & Krishnarani [26], uma razão para isto é que muitas das famílias de distribuições simétricas existentes não são eficazes em modelar os conjuntos de dados assimétricos e de caudas pesadas que surgem em várias situações da vida real. Com isso, aumentou o interesse em buscar novas famílias de distribuições simétricas e assimétricas.

Nas subsubseções seguintes são vistas algumas distribuições de probabilidade utilizadas como base para a formulação dos diferentes modelos de regressão binária.

2.4.1 Distribuição Logística

A distribuição logística foi proposta inicialmente para estudos de crescimento populacional humano (Balakrishnan [27]). Em teoria da probabilidade e estatística, a distribuição logística é classificada como sendo uma distribuição de probabilidade contínua. Então, sendo X uma variável aleatória contínua, diz-se que X tem distribuição logística com parâmetros de locação $\mu \in \mathbb{R}$ e de escala $s > 0$, se sua função densidade de probabilidade (FDP) é dada por:

$$f(x;\mu,s) = \frac{e^{\frac{x-\mu}{s}}}{s\left(1+e^{\frac{x-\mu}{s}}\right)^2}, \quad x \in \mathbb{R}.$$

E a FDA é dada por:

$$F(x;\mu,s) = \frac{1}{1+e^{-\frac{x-\mu}{s}}}, \quad x \in \mathbb{R}.$$

Se $\mu = 0$ e $s = 1$, a distribuição assume a sua forma padrão.

2.4.2 Distribuição de Lindley

Lindley [28, 29] introduziu uma nova família de distribuições contínuas para uma variável aleatória X com suporte nos reais não negativos. Diz-se que uma variável aleatória X segue uma distribuição de Lindley com parâmetro $\theta > 0$ se sua FDP é dada por:

$$f(x;\theta) = \frac{\theta^2}{(\theta+1)}(1+x)e^{-\theta x}, \quad x \geq 0. \qquad (2)$$

Diversas propriedades e aplicações desta distribuição foram estudadas por Ghitany *et al.* [30] e Al-Mutairi *et al.* [31]. Em (2), nota-se que a distribuição de Lindley é uma mistura das distribuições Exponencial(θ) e Gama($2, \theta$), sendo que os pesos dados a essa mistura são, respectivamente, γ e $(1-\gamma)$, com $\gamma = \theta/(1+\theta)$. Esses mesmos autores observaram que, embora a distribuição de Lindley seja semelhante à distribuição exponencial, ela pode ser usada como um modelo melhor do que a distribuição exponencial em algumas situações, devido ao fato de que, enquanto a distribuição exponencial tem taxa de risco e função de vida residual média (MRLF, do inglês *mean residual life function*) constantes, a distribuição de Lindley possui taxa de risco crescente e MRLF decrescente.

Estudos sobre a distribuição de Lindley vêm crescendo e ganhando bastante espaço; várias extensões/generalizações dela podem ser encontradas na literatura estatística recente (ver, por exemplo, Nadarajah *et al.* [32], Bakouch *et al.* [33], Elbatal & Elgarhy [34], Gómez-Déniz *et al.* [35], Ashour & Eltehiwy [36], Nedjar & Zeghdoudi [37]).

A distribuição de Lindley oferece muitas vantagens ao ter seu suporte estendido para toda a reta real, visto que resulta num modelo mais flexível e competitivo do que muitas classes de distribuições simétricas com suporte em $(-\infty, \infty)$ (Nitha & Krishnarani [26]).

2.4.2.1 Distribuição Double Lindley

Tal distribuição foi proposta por Nitha & Krishnarani [26] e Kumar & Jose [38]. Seja X uma

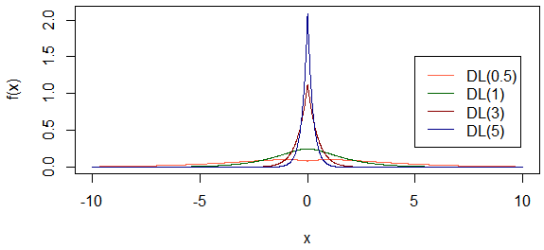

Figura 1: FDP da distribuição DL, para diferentes valores de θ.

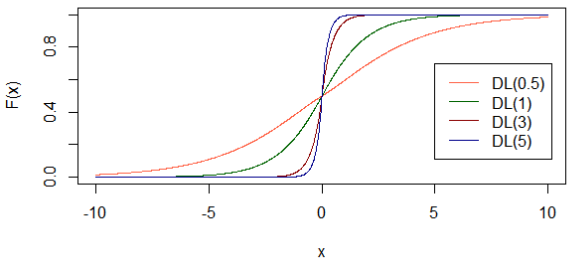

Figura 2: FDA da distribuição DL, para diferentes valores de θ.

variável aleatória contínua com distribuição double Lindley (DL) de parâmetro $\theta > 0$, então sua FDP é dada por:

$$f(x;\theta) = \frac{\theta^2}{2(\theta+1)}(1+|x|)e^{-\theta|x|}, \quad x \in \mathbb{R}.$$

Conforme observado por Nitha & Krishnarani [26], a FDP da variável aleatória com distribuição DL pode ser vista como uma mistura de duas FDPs: uma Laplace com média 0 e variância $2\theta^2$ e outra gama bilateral (*two-sided*) com parâmetro de forma 2 e parâmetro de escala θ, cujos pesos são, respectivamente, γ e $(1-\gamma)$, com $\gamma = \theta/(1+\theta)$.

Observa-se na Figura 1 a forma da FDP da distribuição DL(θ) para diferentes valores atribuídos a θ, onde fica evidente que a densidade é simétrica em relação a 0 (média e mediana da distribuição), tem natureza unimodal (para $\theta \geq 1$; neste caso, a moda é igual a 0) e bimodal (para $\theta < 1$; neste caso, as modas são iguais a $\pm(1-1/\theta)$), e fica mais elevada (ou "pontuda") à medida que o o valor de θ aumenta.

Por sua vez, a FDA da distribuição DL(θ) é expressa por:

$$F(x;\theta) = \begin{cases} \frac{1}{2(\theta+1)}\left[1+\theta(1-x)\right]e^{\theta x}, & \text{se } x \leq 0, \\ 1 - \frac{1}{2(\theta+1)}\left[1+\theta(1+x)\right]e^{-\theta x}, & \text{se } x > 0, \end{cases}$$

que também pode ser escrita da seguinte forma:

$$F(x;\theta) = \frac{1}{2} + \frac{1}{2}\text{sgn}(x)\left\{1 - \frac{1}{\theta+1}\left[1+\theta(1+|x|)\right]e^{-\theta|x|}\right\},$$

para $x \in \mathbb{R}$, em que sgn$(\cdot)$ denota a função sinal.

A Figura 2 mostra a FDA da distribuição DL(θ) para diferentes valores de θ, em que observa-se a sua forma não decrescente e contínua.

Para maiores detalhes, assim como outras propriedades (momentos, curtose, assimetria etc.) da distribuição DL, recomenda-se ao leitor consultar os trabalhos de Nitha & Krishnarani [26] e Kumar & Jose [38].

Importante ressaltar que, para obter o modelo de regressão binária a partir da distribuição DL, deve-se considerá-la em sua forma padrão. Como ela possui um único parâmetro θ, basta então tomar $\theta = 1$. Assim, a forma padrão dessa distribuição é representada por:

$$F(x) = \frac{1}{2} + \frac{1}{2}\text{sgn}(x)\left\{1 - \left(1 + \frac{|x|}{2}\right)e^{-|x|}\right\},$$

para $x \in \mathbb{R}$.

Segundo os trabalhos de Bazán *et al.* [39] e Silva *et al.* [40], para a construção de distribuições de probabilidade do tipo potência ou reversa de potência, parte-se de uma distribuição dita de base (ou basal), cuja FDP é unimodal, log-côncava e de suporte real. Essas propriedades são válidas para a forma padrão da distribuição DL. Portanto, ela pode ser empregada para o desenvolvimento de novas distribuições do tipo potência e reversa de potência. Isto será discutido em maiores detalhes nas subsubsubseções 2.4.2.3 e 2.4.2.4.

2.4.2.2 Distribuição Double Lindley Assimétrica

Como visto anteriormente, a distribuição DL pertence à família de distribuições simétricas. Logo, sua aplicabilidade, no contexto de regressão binária, pode estar restrita a situações em que os conjuntos de dados reais são balanceados ou com desbalanceamento pouco acentuado. Existem diferentes métodos de introdução de assimetria em

uma família de distribuições simétricas. Nitha & Krishnarani [26] sugerem ao leitor consultar, por exemplo, os trabalhos de Ayebo & Kozubowski [41], Kotz *et al.* [42] e Azzalini [43], para algumas dessas propostas. E, para as aplicações das distribuições assim formadas, consultar, por exemplo, Julia & Vives-Rego [44] e Kozubowski & Podgorski [45].

Nesta subsubsubseção é apresentada uma versão assimétrica da distribuição DL, obtida por Nitha & Krishnarani [26] usando a ideia de fatores de escala inversa de Fernandez & Steel [46]. Neste método, um parâmetro novo é adicionado, que atua como um parâmetro de assimetria na família de distribuições simétricas.

A FDP da distribuição double Lindley assimétrica (ADL) com parâmetros $\theta > 0$ e $\lambda > 0$, de acordo com Nitha & Krishnarani [26], é dada por:

$$f(x;\theta,\lambda) = \frac{\theta^2}{(\theta+1)}\frac{\lambda}{(1+\lambda^2)}\begin{cases}(1-\frac{x}{\lambda})\,e^{\frac{\theta x}{\lambda}}, & \text{se } x \leq 0,\\ (1+\lambda x)e^{-\theta\lambda x}, & \text{se } x > 0.\end{cases}$$

Nota-se que, para $\lambda = 1$, a distribuição ADL(θ,λ) se reduz à distribuição DL(θ).

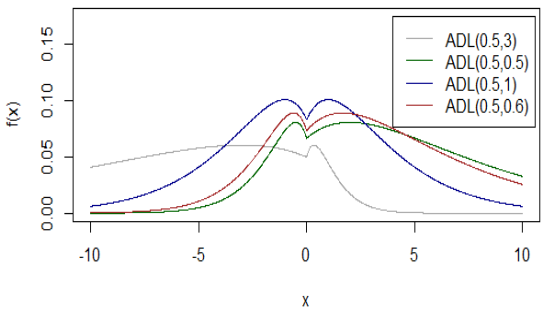

Figura 3: FDP da distribuição ADL, considerando $\theta = 0{,}5$ e $\lambda = \{0{,}5, 0{,}6, 1, 3\}$.

Nas Figuras 3 e 4, observa-se as diferentes formas da FDP acima quando são considerados diferentes valores de θ e λ. Na Figura 4, por exemplo, nota-se uma assimetria à direita (ou positiva) para os valores de $\lambda = \{0{,}5; 0{,}6\}$, enquanto que, para $\lambda = 3$, percebe-se uma assimetria à esquerda (ou negativa).

Finalmente, a FDA da distribuição ADL(θ,λ) é

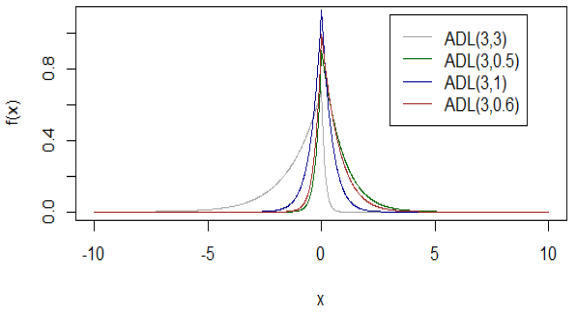

Figura 4: FDP da distribuição ADL, considerando $\theta = 3$ e $\lambda = \{0{,}5, 0{,}6, 1, 3\}$.

expressa por:

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

Para obter o modelo de regressão binária a partir da distribuição ADL, de maneira similar à DL, adota-se $\theta = 1$. Já o parâmetro λ (assimetria) será estimado.

Para maiores detalhes e outras propriedades da distribuição ADL, ver Nitha & Krishnarani [26].

2.4.2.3 Distribuição Potência Double Lindley

Lemonte & Bazán [47] descrevem a composição de uma distribuição de probabilidade, que está fundamentada em considerar uma FDA contínua arbitrária e elevá-la a uma potência real positiva. Com isso, é apresentada uma nova FDA com um parâmetro de potência adicional, que é conhecida como distribuição de potência.

Diz-se que uma variável aleatória univariada X segue uma distribuição de potência com parâmetro de locação $\mu \in \mathbb{R}$, parâmetro de escala $\sigma > 0$ e parâmetro de forma $\lambda > 0$, se X tem FDP representada por:

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

para $x \in \mathbb{R}$, em que $g(\cdot)$ e $G(\cdot)$ são, respectivamente, a FDP e FDA padrão de qualquer distribuição univariada contínua com suporte na reta real (dita distribuição de base).

A FDA da distribuição de potência é definida como:

$$F(x;\mu,\sigma,\lambda) = G\left(\frac{x-\mu}{\sigma}\right)^{\lambda}, \quad x \in \mathbb{R}.$$

As distribuições de potência são assimétricas à direita se $\lambda > 1$, e assimétricas à esquerda se $0 < \lambda < 1$.

Alguns modelos de regressão binária construídos com base em distribuições de potência, como, por exemplo, potência logística, potência normal e potência Cauchy, dentre outras, foram explorados e investigados mais a fundo nos trabalhos de Anyosa [48], Huayanay [49], Bazán *et al.* [39].

Quando analisa-se dados binários que apresentam certo grau de assimetria ou desbalanceamento entre as classes, as funções de ligação simétricas podem não ser úteis para ajustar esses dados, produzindo resultados insatisfatórios. Neste contexto, é desenvolvida aqui a versão de potência para a distribuição DL, que dá origem à distribuição potência double Lindley (PDL).

Então, é apresentada uma nova FDP com um parâmetro de potência adicional $\lambda > 0$, que é definida da forma:

$$f(x;\theta,\lambda) = \lambda\frac{\theta^2(1+|x|)e^{-\theta|x|}}{2(\theta+1)}\left(\frac{1}{2}+\frac{1}{2}\text{sgn}(x)\right.$$
$$\left.\times\left\{1-\left[\frac{1+\theta(1+|x|)e^{-\theta|x|}}{\theta+1}\right]\right\}\right)^{\lambda-1},$$

para $x \in \mathbb{R}$.

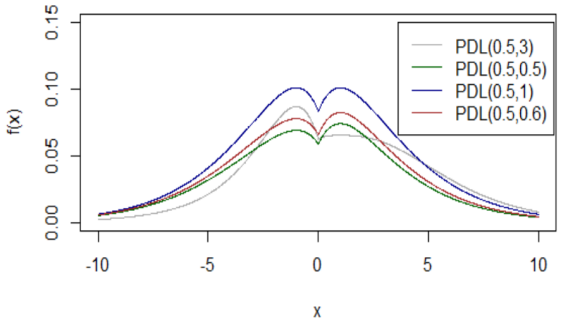

Figura 5: FDP da distribuição PDL, considerando $\theta = 0{,}5$ e $\lambda = \{0{,}5, 0{,}6, 1, 3\}$.

Nas Figuras 5 e 6, observa-se as diferentes formas da FDP acima quando são considerados diferentes valores de θ e λ.

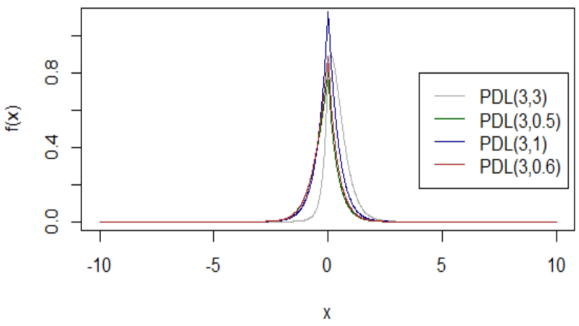

Figura 6: FDP da distribuição PDL, considerando $\theta = 3$ e $\lambda = \{0{,}5, 0{,}6, 1, 3\}$.

A FDA, por sua vez, é definida da forma:

$$F(x;\theta,\lambda) = \left(\frac{1}{2}+\frac{1}{2}\text{sgn}(x)\left\{1-\frac{1}{\theta+1}\right.\right.$$
$$\left.\left.\times\left[1+\theta(1+|x|)\right]e^{-\theta|x|}\right\}\right)^{\lambda}, \quad x \in \mathbb{R}.$$

Analogamente à distribuição ADL, para obter o modelo de regressão binária baseado na distribuição PDL, fixa-se $\theta = 1$, enquanto o parâmetro λ (assimetria) será estimado.

2.4.2.4 Distribuição Reversa de Potência Double Lindley

Para as distribuições reversa de potência, uma propriedade que precisa ser considerada quando são usadas diferentes funções de ligação é a de reversibilidade, segundo a qual a distribuição de S satisfaz a propriedade de reversibilidade se $S \sim F(\cdot) \implies -S \sim G(\cdot) \equiv 1 - F(-\cdot)$. Neste caso, $G(\cdot)$ é chamada de distribuição reversa de $F(\cdot)$ (Bazán *et al.* [50]).

Como também descrito por Lemonte & Bazán [47], a FDP da distribuição reversa de potência pode ser representada pela expressão a seguir:

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

e a sua FDA pode ser expressa por:

$$F(x;\mu,\sigma,\lambda) = 1 - G\left(-\left(\frac{x-\mu}{\sigma}\right)\right)^{\lambda},$$

para $x \in \mathbb{R}$, sendo $\mu \in \mathbb{R}$, $\sigma > 0$ e $\lambda > 0$ os parâmetros de locação, escala e forma, respectivamente, e $G(\cdot)$ é a FDA padrão da distribuição de base.

Considerando a propriedade de reversibilidade, é possível propor outras distribuições reversa de potência. Neste contexto, como o objetivo do trabalho é propor distribuições para lidar com dados assimétricos, é desenvolvida aqui a versão reversa de potência para a distribuição DL, que dá origem à distribuição reversa de potência double Lindley (RPDL).

Logo, é apresentada uma nova FDP com um parâmetro de potência adicional $\lambda > 0$, que é definida da forma:

$$f(x;\theta,\lambda) = \lambda \frac{\theta^2(1+|x|)e^{-\theta|x|}}{2(\theta+1)} \left(\frac{1}{2} + \frac{1}{2}\text{sgn}(-x) \times \left\{ 1 - \left[\frac{1+\theta(1+|-x|)e^{-\theta|-x|}}{\theta+1} \right] \right\} \right)^{\lambda-1},$$

para $x \in \mathbb{R}$.

Nas Figuras 7 e 8, observa-se as diferentes formas da FDP acima quando são considerados diferentes valores de θ e λ.

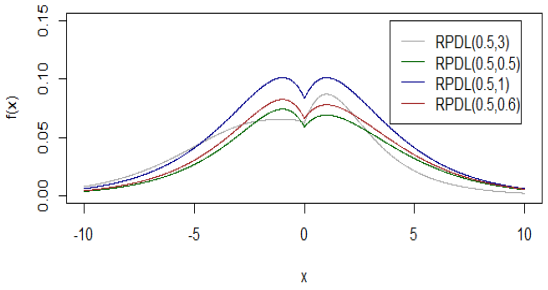

Figura 7: FDP da distribuição RPDL, considerando $\theta = 0{,}5$ e $\lambda = \{0{,}5, 0{,}6, 1, 3\}$.

A FDA, por sua vez, é definida como:

$$F(x;\theta,\lambda) = 1 - \left(\frac{1}{2} + \frac{1}{2}\text{sgn}(-x) \left\{ 1 - \frac{1}{\theta+1} \times [1+\theta(1+|-x|)]\, e^{-\theta|x|} \right\} \right)^{\lambda},$$

para $x \in \mathbb{R}$. Em sua forma padrão, $\theta = 1$.

Particularmente, para os modelos de regressão binária com base nas distribuições ADL, PDL e RPDL, utiliza-se a transformação $\lambda = \exp\{\lambda^*\}$ para que o parâmetro λ^* esteja definido nos reais; com isso, calcula-se o erro padrão de $\widehat{\lambda}$ (estimador de MV de λ) a partir da aplicação do método delta,

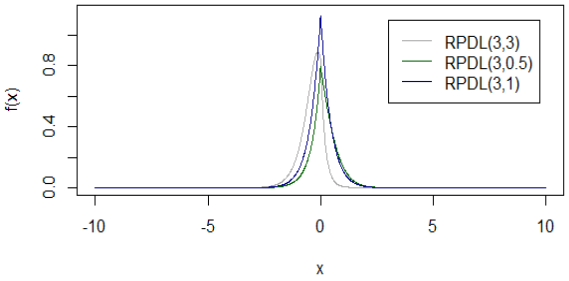

Figura 8: FDP da distribuição RPDL, considerando $\theta = 3$ e $\lambda = \{0{,}5, 1, 3\}$.

isto é, mediante o uso da função *deltamethod*$(\cdot)$ do pacote `msm` (Jackson [51]).

2.5 Comparação de Modelos

A comparação de modelos é fundamental quando são observados dois ou mais modelos ajustados descrevendo eventos semelhantes, com o intuito de selecionar aquele que represente melhor ou se aproxime mais da realidade. Tal comparação pode ser feita conforme descrito em Gelman *et al.* [52]. Após ajustar todos os modelos candidatos propostos, a escolha é feita com base naquele que apresentar como resultado o melhor desempenho, ou seja, o menor valor dentre todos os critérios de informação.

Na literatura existem vários critérios de informação disponíveis, como, por exemplo, o critério de informação de Akaike (AIC) (Akaike [53]), o critério de informação Bayesiano ou de Schwarz (BIC) (Schwarz [54]), o AIC corrigido (AICc) (Sugiura [55]; Hurvich & Tsai [56]), o AIC consistente (CAIC) (Bozdogan, [57]; Anderson *et al.* [58]), o critério de informação de Hannan-Quinn (HQIC) (Hannan & Quinn [59]) etc. Dentre eles, os mais conhecidos e utilizados são o AIC e o BIC. Esses dois critérios de comparação ou seleção de modelos serão usados neste trabalho.

2.6 Métricas de Desempenho

Nesta subseção são apresentadas algumas métricas de desempenho para avaliar a capacidade preditiva dos modelos de regressão binária propostos. Existem inúmeras métricas diferentes e algumas funcionam melhor do que outras para um deter-

minado tipo de problema; escolher uma medida adequada para avaliar o modelo é tão importante quanto escolher um bom modelo.

A maioria das métricas de desempenho é baseada na matriz de confusão, apresentada em uma tabela de contingência 2×2, que é uma das formas mais simples de visualizar e estabelecer o cálculo dessas métricas.

A partir da Tabela 1 (matriz de confusão), tem-se que:

- Y equivale às respostas binárias observadas na amostra, isto é, $Y = 1$ representa "sucesso" e $Y = 0$ denota "fracasso";
- $\widehat{Y}$ é a variável predita correspondente às respostas classificadas pelo modelo em análise, isto é, $\widehat{Y} = 1$ indica "sucesso" e $\widehat{Y} = 0$ representa "fracasso";
- VP (Verdadeiro positivo) é quando a observação é classificada como "sucesso" e é "sucesso";
- VN (Verdadeiro negativo) é quando a observação é classificada como "fracasso" e é "fracasso";
- FP (Falso positivo) é quando a observação é classificada como "sucesso" e é "fracasso";
- FN (Falso negativo) é quando a observação é classificada como "fracasso" e é "sucesso".

Logo, tem-se que $VP+VN+FP+FN = n$, sendo n o tamanho da amostra.

Tabela 1: Matriz de confusão 2×2 (classificação binária).

		Predito ($\widehat{Y}$)	
		1	0
Observado (Y)	1	VP	FN
	0	FP	VN

Desta forma, pode-se utilizar as seguintes métricas para avaliação da capacidade preditiva (no caso, valores dessas medidas próximos a 1 são indicativos de um "bom" modelo):

Sensibilidade (SEN): também conhecida como "revocação" (ou *recall*), corresponde à proporção dos verdadeiros positivos entre todas as observações que realmente são positivas no conjunto de dados, isto é,

$$SEN = \frac{VP}{VP + FN}.$$

Especificidade (SPE): é a proporção dos verdadeiros negativos entre todas as observações que realmente são negativas no conjunto de dados, ou seja,

$$SPE = \frac{VN}{VN + FP}.$$

Valor Preditivo Positivo (VPP): também conhecida como "precisão" (ou *precision*), tal métrica traz a informação da quantidade de observações classificadas como positivas que são realmente positivas, ou seja, a proporção de verdadeiros positivos em relação a todas as predições positivas:

$$VPP = \frac{VP}{VP + FP}.$$

Valor Preditivo Negativo (VPN): é a métrica que traz a informação da quantidade de observações classificadas como negativas que são realmente negativas, ou seja, a proporção de verdadeiros negativos em relação a todas as predições negativas:

$$VPN = \frac{VN}{VN + FN}.$$

Acurácia (ACC): é a métrica mais popular; ela representa a proporção de acertos do modelo, ou seja, é a fração de verdadeiros positivos e verdadeiros negativos em relação a todos os resultados possíveis:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}.$$

No entanto, quando as classes são desbalanceadas, utilizar a ACC não é adequado, pois tal medida poderia causar uma falsa impressão de bom desempenho, levando a tirar conclusões não satisfatórias.

Acurácia Balanceada ($ACCB$): a métrica anterior é geralmente usada na forma balanceada quando o problema a ser estudado envolve classificação desbalanceada; é dada pela média aritmética entre a fração de verdadeiros positivos e a fração de verdadeiros negativos:

$$ACCB = \frac{\frac{VP}{VP+FN} + \frac{VN}{VN+FP}}{2} = \frac{SEN + SPE}{2}.$$

F1-*Score*: é a métrica referente à média harmônica entre SEN e VPP:

$$F1 - Score = 2 \times \frac{SEN \times VPP}{SEN + VPP}.$$

Coeficiente de Correlação de Matthews (MCC): esta medida é usada para interpretar a classificação geral do modelo (Baldi *et al.* [60]). O MCC leva em consideração valores positivos e negativos, verdadeiros e falsos, e geralmente é considerado uma medida equilibrada que pode ser usada mesmo que as classes estudadas sejam desbalanceadas. Segundo Boughorbel [61], o MCC é um coeficiente de correlação entre as classificações binárias observadas e preditas, e retorna um valor entre -1 e $+1$. Sendo que um coeficiente de $+1$ representa uma predição (ou classificação) perfeita; quando igual a 0, uma predição aleatória; e quando igual a -1, indica que a predição é totalmente inversa. Tal coeficiente é definido como:

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}.$$

Uma ferramenta gráfica relevante é a curva característica de operação do receptor (ROC, do inglês *receiver operating characteristic curve*), a qual é descrita a seguir.

Curva ROC: é definida como um gráfico que indica o comportamento de um classificador binário em possibilidades diferentes do valor limite para a classificação. Para o eixo x (abscissa), tem-se a medida de $1 - SPE$, e, para o eixo y (ordenada), tem-se a medida de SEN. A curva ROC deve ser interpretada de forma que, se a curva estiver mais distante da diagonal principal, melhor o desempenho do modelo associado a ela. Para definir o melhor ponto de corte, identifica-se aquele que maximiza simultaneamente a SPE e a SEN da classificação, ou seja, o ponto de corte ótimo deve estar mais próximo do eixo superior esquerdo do gráfico. Um exemplo de curva ROC é visto na Figura 9.

Pode-se, ainda, comparar o desempenho dos classificadores através da área sob a curva ROC, ou AUC (do inglês *area under the curve*). Hanley & McNeil [62] indicaram que a AUC tem a propriedade importante de ser equivalente ao teste de Wilcoxon. Um modelo de classificação binária é considerado apropriado se o valor da

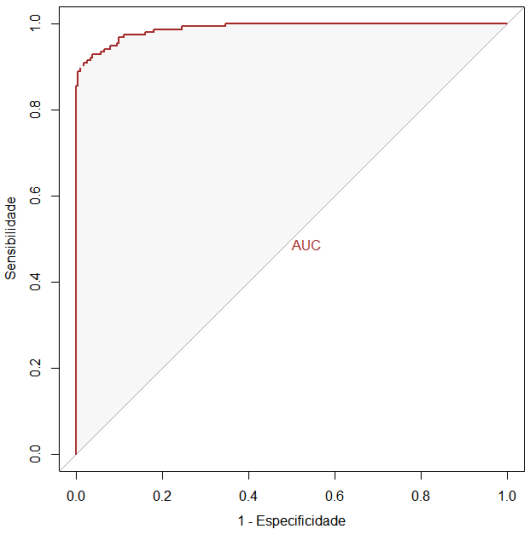

Figura 9: Exemplo de curva ROC.

AUC for próximo de 1; caso não seja apropriado, será próximo a 0,5 (Powers [63]).

Computacionalmente, para a implementação da curva ROC, cálculo da AUC e escolha do ponto de corte ótimo, foi utilizado o pacote `pROC` (Robin *et al.* [64]) do *software* R.

Outras duas medidas de desempenho importantes e úteis são descritas a seguir.

Área Sob a Curva Precisão $\times$ Revocação ($AUCPR$): esta medida diz respeito à área sob a curva obtida com a comparação da precisão *versus* revocação (ou ainda, VPP *versus* SEN), para diferentes pontos de corte de probabilidade. Assim como a AUC, tal medida é sugerida para a comparação do desempenho entre classificadores distintos, sendo mais recomendada que a AUC para o caso de bases de dados desbalanceadas (Saito & Rehmsmeier [65]), pois foca-se na classificação correta da classe positiva, enquanto que a AUC observa ambas as classes. No caso de um classificador aleatório, o valor obtido é a própria proporção de casos positivos nos dados.

Brier *Score* (BS): proposto por Glenn W. Brier em 1950 (Brier [66]), o Brier *Score* serve para avaliar (e comparar) a precisão das previsões proba-

bilísticas. Pode ser definido como:

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(p_t - o_t\right)^2,$$

em que N é o número de instâncias, p_t é a probabilidade prevista da t-ésima instância pertencer à classe de interesse (evento), e o_t é o resultado real do evento na instância t (1 se ocorrer e 0 se não ocorrer). No caso, pontuações menores (mais próximas de zero) indicam melhores previsões, sendo que o mínimo que se espera de um bom modelo é que sua medida BS seja menor do que 1/4. Por outro lado, são considerados modelos com previsões de má qualidade aqueles que aparesentam medidas maiores do que 1/4.

2.7 Método de Validação

Ainda sobre métricas de desempenho aceitáveis, o processo de construção de modelos preditivos deve passar por uma importante etapa de validação. O intuito é verificar se o modelo ajustado possui uma boa capacidade de generalização, ao ser aplicado a dados que possuem as mesmas características do conjunto a partir do qual ele foi desenvolvido.

Neste trabalho é considerado o método *holdout*, que consiste em dividir a base de dados original em duas partes, sendo uma delas utilizada para o desenvolvimento/construção do modelo, enquanto a outra é empregada para a avaliação da performance preditiva. Ou seja, a maior parcela dos dados é usada para a estimação do modelo (amostra treinamento), e a menor para verificação da adequabilidade/performance preditiva do modelo (amostra teste). Na prática, divide-se em 70% e 30% para treinamento e teste, respectivamente.

3 Estudos de Simulação

No Material Suplementar A são apresentados em detalhes os resultados de estudos de simulação dos tipos: recuperação de parâmetros e *misspecification*, e a forma como foram realizados, considerando todos os modelos propostos (DL, ADL, PDL e RPDL), além do modelo de regressão binária tradicional (logística).

Em resumo, os resultados obtidos demonstraram uma recuperação adequada dos parâmetros dos quatro modelos aqui desenvolvidos, segundo o procedimento de estimação proposto (máxima verossimilhança).

De modo geral, observou-se uma boa conformidade dos critérios AIC e BIC em discriminar os diferentes modelos estudados, sendo que ambos tiveram performances parecidas, porém com ligeira vantagem para o critério AIC nos resultados obtidos.

Observou-se, ainda, que nas situações de desbalanceamento entre as classes, os modelos assimétricos produzem as melhores estimativas das probabilidades de evento, superando, de maneira geral, os modelos logístico e DL, por apresentarem viés mais próximo a zero. Notou-se também uma certa superioridade do modelo DL sobre o modelo logístico nos casos em que o desbalanceamento é mais acentuado (10% de uns).

4 Aplicação

Nesta seção são apresentadas aplicações da metodologia descrita na Seção 2 a um conjunto de dados reais, oriundo da área financeira (Seção 4.1), e um conjunto de dados sintéticos, que reflete dados reais da área industrial (Seção 4.2).

4.1 *Santander Customer Transaction Prediction Dataset*

O primeiro conjunto de dados utilizado neste trabalho é referente a uma competição promovida pelo Banco Santander para a comunidade do Kaggle, cujo objetivo era predizer se um cliente iria efetuar uma operação financeira específica no futuro, independentemente da quantia transacionada. A base de dados é composta de 200.000 observações (ou instâncias) e 200 variáveis preditoras contínuas, normalizadas, sem a presença de valores faltantes e anonimizadas. A variável resposta binária *Target*, se o cliente irá realizar uma transação ou não, está distribuída da seguinte forma: 10% de casos positivos ("sucessos") e 90% de casos negativos ("fracassos"). Portanto, essa base é altamente desbalanceada, em uma proporção de 9 casos negativos para cada 1 positivo. O banco de dados pode ser acessado pelo endereço: `https://www.kaggle.com/lakshmi25npathi/santander-customer-transaction-prediction-dataset`.

Uma vez que existe um alto grau de desbalanceamento entre as classes (zeros e uns) da variável res-

posta, o uso de ligações simétricas poderia não ser adequado na modelagem de regressão binária do conjunto de dados em questão (Chen *et al.* [67]).

Para a seleção das variáveis (do total de 200) que mais contribuem para predizer a realização ou não da transação específica, utilizou-se a técnica de *random forest*, ou floresta aleatória (para maiores detalhes, ver, por exemplo, Breiman [68] e Dellier [69]). Desta forma, as variáveis escolhidas como preditoras para os cinco modelos candidatos (logístico, DL, ADL, PDL e RPDL) foram sete, descritas na Tabela 2.

Tabela 2: Variáveis consideradas na aplicação da metodologia proposta aos dados do Banco Santander.

Variável	Notação	Classificação
Target	Y	Categórica 0-1
$var81$	X_1	Contínua
$var139$	X_2	Contínua
$var6$	X_3	Contínua
$var53$	X_4	Contínua
$var110$	X_5	Contínua
$var26$	X_6	Contínua
$var146$	X_7	Contínua

De acordo com os valores dos critérios de informação, apresentados na Tabela 3, seleciona-se o modelo de regressão binária DL como sendo o de "melhor" ajuste, por ter apresentado os (ligeiramente) menores valores de AIC e BIC.

Na Figura 10 são apresentadas as curvas ROC para os cinco modelos, em que os valores do AUC obtidos são próximos a 0,73. Isso indica que todos esses modelos de regressão binária são igualmente adequados.

Verificando os resultados expostos na Tabela 4, nota-se que o modelo PDL apresentou o melhor desempenho preditivo (entre todas as medidas exploradas), para verificar se um cliente vai efetuar uma operação financeira específica. Superou, inclusive, o modelo DL que, por sua vez, performou melhor do que o modelo logístico. Tal avaliação foi feita na amostra teste (30%).

Considerando os resultados das medidas de avaliação da performance preditiva para os cinco modelos de regressão binária (Tabela 4), pode-se então selecionar o modelo PDL. Portanto, a probabilidade de que um cliente i irá efetuar uma operação financeira específica, de acordo com o modelo escolhido (PDL), é calculada por:

$$\begin{aligned}\widehat{\mu}_i &= F\left(\widehat{\eta}_i\right) \\ &= \left(\frac{1}{2}+\frac{1}{2}\text{sgn}(\widehat{\eta}_i)\left\{1-\left(1+\frac{|\widehat{\eta}_i|}{2}\right)e^{-|\widehat{\eta}_i|}\right\}\right)^{0,2313},\end{aligned}$$

sendo:

$$\begin{aligned}\widehat{\eta}_i = &-4{,}2616-0{,}1157x_{i1}-0{,}0417x_{i2}+0{,}4724x_{i3} \\ &+0{,}5474x_{i4}+0{,}0579x_{i5}+0{,}0697x_{i6}-0{,}1592x_{i7}.\end{aligned}$$

Finalmente, a regra de decisão (ou de classificação) para alocar um cliente i numa das duas classes (1 - realização da transação específica, 0 - não realização da transação específica) é:

$$\begin{aligned} i \in 1 \quad (\text{ou } \widehat{Y}_i = 1) \qquad &\text{se} \qquad \widehat{\mu}_i \geq 0{,}094, \\ i \in 0 \quad (\text{ou } \widehat{Y}_i = 0) \qquad &\text{se} \qquad \widehat{\mu}_i < 0{,}094. \end{aligned}$$

4.2 *Machine Predictive Maintenance Classification Dataset*

O segundo conjunto de dados utilizado neste trabalho é referente a uma base de dados sintética que reflete dados reais de manutenção preditiva encontrados no meio industrial, cujo objetivo é predizer se um equipamento rotativo apresentará falha no futuro, sob determinadas condições apresentadas. A base de dados é composta de 10.000 observações (ou instâncias) e 10 variáveis preditoras, sem a presença de valores faltantes. A variável resposta binária (equipamento falhou ou não) está distribuída da segunte forma: 49,6% de casos positivos (equipamento falhou) e 50,4% de casos negativos (equipamento não falhou). O banco de dados original pode ser acessado pelo endereço: `https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification`.

Na etapa inicial da seleção de variáveis, foram cuidadosamente avaliadas as correlações entre elas. Algumas variáveis de identificação e aquelas que apresentavam uma correlação particularmente elevada (i.e., coeficiente de correlação linear r de Pearson superior a 0,8) foram identificadas e excluídas para mitigar possíveis problemas de multicolinearidade. Vale ressaltar que essa exclusão se aplicou apenas a algumas variáveis altamente correlacionadas, mantendo aquelas que contribuem de forma única para a análise. Essa abordagem estratégica visa preservar a diversidade de informações relevantes para a identificação do estado de funcionamento da máquina.

Tabela 3: Valores de AIC e BIC para os modelos de regressão binária ajustados ao conjunto de dados do Banco Santander.

	Modelo				
Critério	Logístico	DL	ADL	PDL	RPDL
AIC	598,8691	**598,6489**	601,2960	600,8749	600,6522
BIC	638,1312	**637,9109**	645,4658	645,0447	644,8220

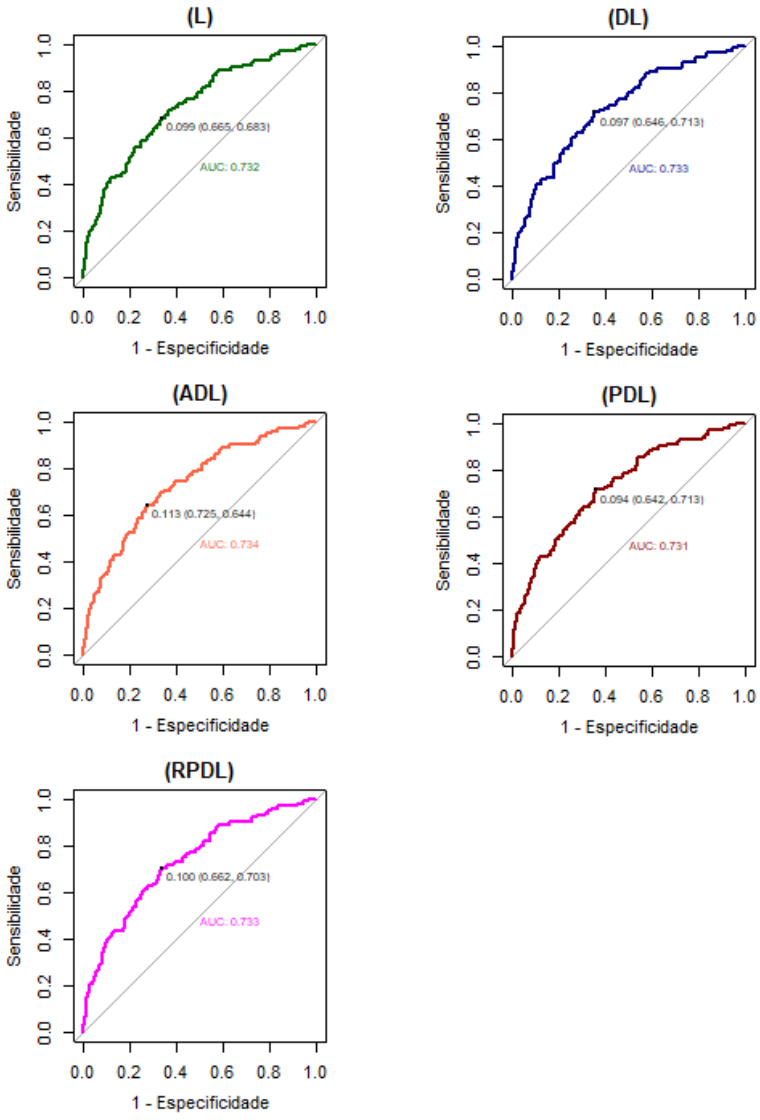

Figura 10: Curva ROC dos modelos de regressão binária logística, DL, ADL, PDL e RPDL, ajustados ao conjunto de dados do Banco Santander (amostra treinamento - 70%).

Tabela 4: Medidas de avaliação da performance preditiva para os modelos de regressão binária ajustados ao conjunto de dados do Banco Santander (amostra teste - 30%).

	Medida				
Modelo	SEN	VPP	F1-*Score*	ACCB	AUCPR
Logístico	0,5909	0,0850	0,1486	0,6239	0,1036
DL	0,6364	0,0864	0,1522	0,6368	0,1020
ADL	0,5455	0,0870	0,1500	0,6183	0,1001
PDL	**0,6818**	**0,0904**	**0,1596**	**0,6559**	**0,1045**
RPDL	0,5455	0,0789	0,1379	0,6012	0,1008

Desta forma, as variáveis escolhidas como preditoras para os cinco modelos candidatos (logístico, DL, ADL, PDL e RPDL) foram 3, descritas na Tabela 5.

Tabela 5: Variáveis consideradas na aplicação da metodologia proposta aos dados industriais de manutenção preditiva.

Variável	Notação	Classificação
Target	Y	Categórica 0-1
air temperature (K)	X_1	Contínua
rotational speed (rpm)	X_2	Contínua
tool wear (min)	X_3	Contínua

De acordo com os valores dos critérios de informação, apresentados na Tabela 6, pode-se selecionar o modelo de regressão binária DL como sendo o de "melhor" ajuste, por ter apresentado o segundo menor valor de AIC e o menor valor de BIC. É importante observar que o modelo ADL não convergiu com o método de otimização empregado.

Na Figura 11 são apresentadas as curvas ROC para os quatro modelos, em que os valores de AUC obtidos são próximos a 0,70. Isso indica que todos esses modelos de regressão binária são igualmente adequados.

Verificando os resultados expostos na Tabela 7, nota-se que os modelos PDL e RPDL apresentaram as métricas mais favoráveis (considerando as medidas exploradas) na previsão de falhas em equipamentos. Adicionalmente, é relevante salientar que, no que diz respeito ao VPP, todos os modelos propostos atingiram desempenho igual ou superior ao do modelo logístico.

5 Conclusões e Trabalhos Futuros

Neste trabalho foram introduzidos novos modelos de regressão para a análise de dados binários, baseados nas distribuições double Lindley, double Lindley assimétrica, potência double Lindley e reversa de potência double Lindley (sendo as duas últimas distribuições também inéditas na literatura).

Estudos de simulação foram realizados com o objetivo, dentre outros, de avaliar a performance do método de estimação considerado (máxima verossimilhança).

Por fim, dois bancos de dados (um real e outro sintético) foram utilizados para ilustrar a aplicabilidade dos modelos e métodos propostos. Dentre outros, verificou-se que os modelos de regressão binária com as funções de ligação potência e reversa de potência aqui propostas apresentaram resultados promissores, com ajuste e performance preditiva melhores do que os modelos de regressão binária usando ligações comuns (e.g., logito).

Como sugestão de desenvolvimentos futuros, pode-se propor classes gerais (e inéditas) de modelos de regressão para a análise de dados binários, as quais seriam compostas pelas versões double (assim como suas versões potência e reversa de potência) das distribuições de probabilidade obtidas da mistura de componentes exponencial e gama, como é o da distribuição de Lindley, considerada neste trabalho, mas também de inúmeras distribuições introduzidas na literatura recente, como as de Akash [70], Shanker [71], Sujatha [72], Ishita [73] etc. Também é de interesse explorar métodos de *oversampling* ou *undersampling* combinados com os modelos de regressão binária aqui propostos, bem como comparar o desempenho dos novos modelos com o de modelos de *Machine Learning*.

Tabela 6: Valores de AIC e BIC para os modelos de regressão binária ajustados ao conjunto de dados de manutenção preditiva. NA = *not available*.

	Modelo				
Critério	Logístico	DL	ADL	PDL	RPDL
AIC	8.838,598	8.835,184	NA	8.842,280	**8.834,371**
BIC	8.867,440	**8.864,030**	NA	8.878,330	8.870,420

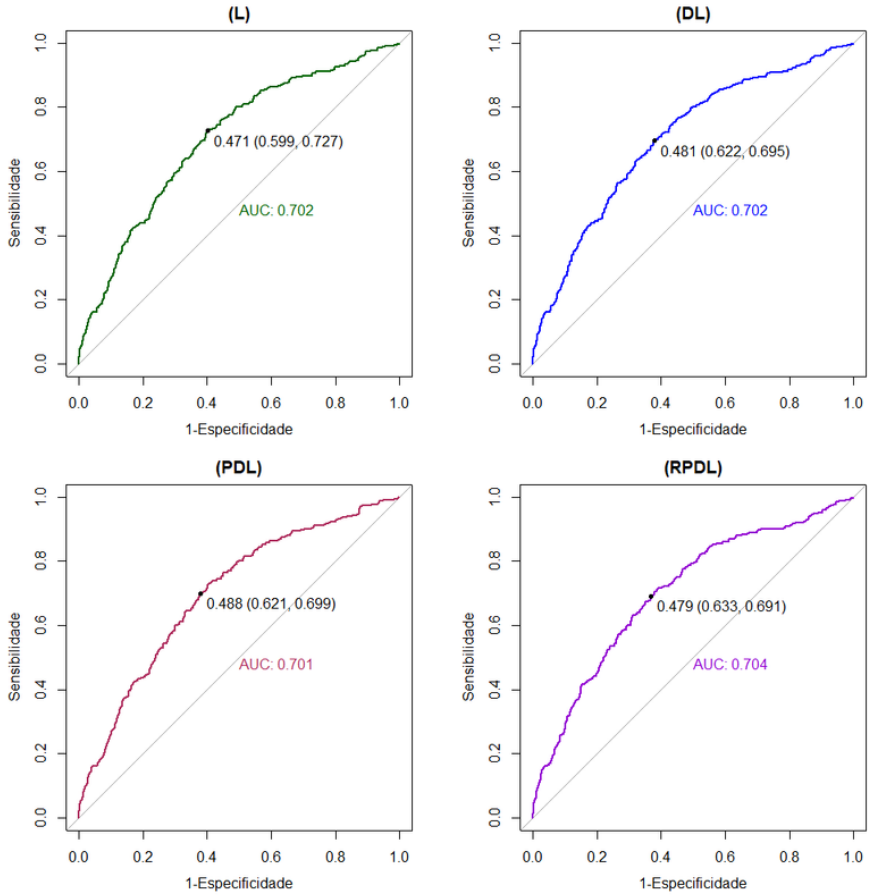

Figura 11: Curva ROC dos modelos de regressão binária logística, DL, PDL e RPDL, ajustados ao conjunto de dados de manutenção preditiva (amostra treinamento - 70%).

Tabela 7: Medidas de avaliação da performance preditiva para os modelos de regressão binária ajustados ao conjunto de dados de manutenção preditiva (amostra teste - 30%). NA = *not available*.

	Medida				
Modelo	SEN	VPP	F1-*Score*	ACCB	AUCPR
Logístico	0,7453	0,6358	0,6862	0,6659	0,7030
DL	0,7134	0,6405	0,6750	0,6628	0,7062
ADL	NA	NA	NA	NA	NA
PDL	**0,7486**	0,6358	**0,6876**	0,6666	0,7008
RPDL	0,7202	**0,6439**	0,6799	**0,6672**	**0,7108**

Também se pretende comparar os novos modelos com os modelos de regressão binária construídos com base em outras distribuições de potência e reversa de potência, como, por exemplo, a potência logística, a potência normal, a potência Cauchy, a reversa de potência logística, a reversa de potência normal, e a reversa de potência Cauchy, que foram propostos por Bazán *et al.* [39].

Referências

[1] David Roxbee Cox and E Joyce Snell. *Analysis of Binary Data*, volume 32. CRC press, 1989.

[2] David Collett. *Modelling Binary Data*. CRC press, 2 edition, 2002.

[3] Pierre-François Verhulst. Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.*, 10:113–126, 1838.

[4] Pierre-François Verhulst. Resherches mathematiques sur la loi d'accroissement de la population. *Nouveaux memoires de l'academie royale des sciences*, 18:1–41, 1845.

[5] Pierre-François Verhulst. Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'académie royale des sciences, des lettres et des beaux-arts de Belgique*, 20:1–32, 1847.

[6] Chester I Bliss. The method of probits. *Science*, 1934.

[7] Michael J Farrell. The demand for motorcars in the united states. *Journal of the Royal Statistical Society. Series A (General)*, 117(2):171–201, 1954.

[8] John Aitchison and James AC Brown. The lognormal distribution with special reference to its uses in economics. 1957.

[9] Daniel Adam. *Les réactions du consommateur devant le prix: contribution aux études de comportement*, volume 15. Sedes, 1958.

[10] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

[11] Claudia Czado and Thomas J Santner. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33(2):213–231, 1992.

[12] Antônio Carlos Pacagnella Júnior, Geciane Silveira Porto, Sérgio Kannebley Júnior, Sérgio Luís da Silva, and Carlos Alberto Grespan Bonacim. Obtenção de patentes na indústria do estado de são paulo: uma análise utilizando regressão logística. *Production*, 19(2):261–273, 2009.

[13] Jaqueline Trentino Silva, Ediney Magalhães Junior, Mariano Martínez Espinosa, and Dayane de Carvalho Rodrigues. Domínios da qualidade de vida associados á percepção de saúde em idosos: comparação do modelo de regressão logística com o de regressão de poisson. *Sigmae*, 8(2):576–583, 2019.

[14] Cleyton de Oliveira Ritta, Marcelo Christiano Gorla, and Nelso Hein. Modelo de regressão logística para análise de risco de crédito em uma instituição de microcrédito produtivo orientado. *Iberoamerican Journal of Industrial Engineering*, 7(13):103–122, 2015.

[15] Bart Van der Paal. A comparison of different methods for modelling rare events data. *PhD thesis*, 2014.

[16] Aida Ali, Siti Mariyam Shamsuddin, Anca L Ralescu, et al. Classification with class imbalance problem: a review. *International Journal Of Advances In Soft Computing And Its Applications*, 7(3):176–204, 2015.

[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[18] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

[19] Saswat Padhi, Todd Millstein, Aditya Nori, and Rahul Sharma. Overfitting in synthesis: Theory and practice. In *International Conference on Computer Aided Verification*, pages 315–334. Springer, 2019.

[20] X.-M Tao, Z.-J Tong, Yan Liu, and D.-D Fu. Svm classifier for unbalanced data based on combination of odr and bsmote. *Kongzhi yu Juece/Control and Decision*.

[21] Tasadduq Imam, Kai Ming Ting, and Joarder Kamruzzaman. z-svm: An svm for improved classification of imbalanced data. In *Australasian Joint Conference on Artificial Intelligence*, pages 264–273. Springer, 2006.

[22] Gauss Moutinho Cordeiro and Clarice GB Demétrio. Modelos lineares generalizados e extensões. *Piracicaba: USP*, 2008.

[23] George Casella and Roger L Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[24] Arne Henningsen and Ott Toomet. maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, 26(3):443–458, 2011.

[25] R Core Team. R: A language and environment for statistical computing, version 3.6. 3. vienna, austria; 2013.

[26] Nitha KU and SD Krishnarani. A new family of heavy tailed symmetric distribution for modeling financial data. *Journal of Statistics Applications & Probability*, 6(3):577–586, 2017.

[27] Narayanaswamy Balakrishnan. *Handbook of the Logistic Distribution*. CRC Press, 1991.

[28] Dennis V Lindley. Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1958.

[29] DV Lindley. Introduction to probability and statistics from a bayesian viewpoint, part ii: Inference. *Camb. Univ. Press, New York*, 1965.

[30] Mohamed E Ghitany, Barbra Atieh, and Saralees Nadarajah. Lindley distribution and its application. *Mathematics and Computers in Simulation*, 78(4):493–506, 2008.

[31] DK Al-Mutairi, ME Ghitany, and Debasis Kundu. Inferences on stress-strength reliability from lindley distributions. *Communications in Statistics-Theory and Methods*, 42(8):1443–1463, 2013.

[32] Saralees Nadarajah, Hassan S Bakouch, and Rasool Tahmasbi. A generalized lindley distribution. *Sankhya B*, 73(2):331–359, 2011.

[33] Hassan S Bakouch, Bander M Al-Zahrani, Ali A Al-Shomrani, Vitor AA Marchi, and Francisco Louzada. An extended lindley distribution. *Journal of the Korean Statistical Society*, 41:75–85, 2012.

[34] I Elbatal and M Elgarhy. Statistical properties of kumaraswamy quasi lindley distribution. *International Journal of Mathematics Trends and Technology*, 4(10):237–246, 2013.

[35] Emilio Gómez-Déniz, Miguel A Sordo, and Enrique Calderín-Ojeda. The log–lindley distribution as an alternative to the beta regression model with applications in insurance. *Insurance: Mathematics and Economics*, 54:49–57, 2014.

[36] Samir K Ashour and Mahmoud A Eltehiwy. Exponentiated power lindley distribution. *Journal of Advanced Research*, 6(6):895–905, 2015.

[37] Sihem Nedjar and Halim Zeghdoudi. On gamma lindley distribution: Properties and simulations. *Journal of Computational and Applied Mathematics*, 298:167–174, 2016.

[38] C Satheesh Kumar and Rosmi Jose. On double lindley distribution and some of its properties. *American Journal of Mathematical and Management Sciences*, 38(1):23–43, 2019.

[39] Jose Romeo, Jorge Bazán, and Josemar Rodrigues. Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics*, 28:467–482, 06 2014.

[40] S.; BAZÁN J. L. SILVA, A. N.; ANYOSA. Bayesian binary regression modeling for unbalanced data using new links. *Rev. Bras. Biom., Lavras*, 38:385–417, 2020.

[41] Abraham Ayebo and Tomasz J Kozubowski. An asymmetric generalization of gaussian and laplace laws. *Journal of Probability and Statistical Science*, 1(2):187–210, 2003.

[42] Samuel Kotz, Tomasz Kozubowski, and Krzystof Podgorski. *The Laplace distribution*

and generalizations: a revisit with applications to communications, economics, engineering, and finance. Springer Science & Business Media, 2012.

[43] Adelchi Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178, 1985.

[44] Olga Julia and Josep Vives-Rego. Skew-laplace distribution in gram-negative bacterial axenic cultures: new insights into intrinsic cellular heterogeneity. *Microbiology*, 151(3):749–755, 2005.

[45] Tomasz J Kozubowski and Krzysztof Podgórski. Asymmetric laplace laws and modeling financial data. *Mathematical and Computer Modelling*, 34(9-11):1003–1021, 2001.

[46] Carmen Fernández and Mark FJ Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[47] Artur J Lemonte and Jorge L Bazán. New links for binary regression: an application to coca cultivation in peru. *Test*, 27(3):597–617, 2018.

[48] Susan Alicia Chumbimune Anyosa. *Regressão binária usando ligações potência e reversa de potência*. PhD thesis, Universidade de São Paulo, 2017.

[49] Alex de La Cruz Huayanay et al. Modelos de regressão para resposta binária na presença de dados desbalanceados. 2019.

[50] Jorge Bazán, F. Torres-Avilés, Adriano Suzuki, and Francisco Louzada. Power and reversal power links for binary regressions: An application for motor insurance policyholders: J.l. bazÁn et al. *Applied Stochastic Models in Business and Industry*, 33, 11 2016.

[51] Christopher H. Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011.

[52] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[53] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[54] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[55] N. Sugiura. Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics-theory and Methods*, 7:13–26, 1978.

[56] C. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[57] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 02 1987.

[58] D. R. Anderson, K. P. Burnham, and G. C. White. Comparison of akaike information criterion and consistent akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, 25(2):263–282, 1998.

[59] E. J. Hannan and Barry G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 41(2):190–195, 1979.

[60] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[61] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS One*, 12(6):e0177678, 2017.

[62] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[63] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[64] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1):1–8, 2011.

[65] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10:e0118432, 03 2015.

[66] G. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[67] Ming-Hui Chen, Dipak K Dey, and Qi-Man Shao. A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448):1172–1186, 1999.

[68] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[69] Fernando Dellier Antunes de Souza. Aplicações de deep learning em problemas de machine learning.trabalho de conclusão de curso do programa de mba em ciência de dados (icmc), 2020.

[70] R Shanker. Akash distribution and its applications. *International Journal of Probability and Statistics*, 4(3):65–75, 2015.

[71] R Shanker. Shanker distribution and its applications. *International Journal of Statistics and Applications*, 5(6):338–348, 2015.

[72] Rama Shanker. Sujatha distribution and its applications. *Statistics in Transition. New Series*, 17(3):391–410, 2016.

[73] R Shanker and KK Shukla. Ishita distribution and its applications. *Biometrics & Biostatistics International Journal*, 5(2):1–9, 2017.

Material Suplementar A

Neste material suplementar são apresentados os principais resultados de estudos de simulação dos tipos: recuperação de parâmetros (A.1) e *misspecification* (A.2), considerando os modelos apresentados (DL, ADL, PDL e RPDL), além do modelo de regressão binária tradicional (logística).

A.1 Recuperação de Parâmetros

Neste primeiro estudo de simulação, foram geradas $M = 1.000$ amostras (ou conjuntos de dados) de tamanhos $n = \{100, 200, 500, 1.000\}$ de cada um dos quatro modelos sugeridos. Para os modelos assimétricos (ADL, PDL e RPDL), considerou-se, ainda, três valores distintos para o parâmetro de assimetria λ, a fim de obter diferentes proporções de sucessos (uns) nas amostras geradas: 50% (dados balanceados), 25% (moderado grau de desbalanceamento entre as classes) e 10% (alto grau de desbalanceamento). Sendo assim, foram totalizados 40 cenários simulados.

Para todos os modelos, os coeficientes de regressão foram fixados com os valores: $\beta_0 = 0$ e $\beta_1 = 1$, assim como no trabalho desenvolvido por Bazán *et al.* [39]. A covariável foi gerada considerando $X \sim \text{Uniforme}(-4, 4)$. A partir dessas especificações, os valores da variável resposta Y foram simulados de uma distribuição de Bernoulli com parâmetro $\mu_i = F(\beta_0 + \beta_1 x)$, sendo $F(\cdot)$ a FDA padrão do modelo correspondente.

Os parâmetros dos quatro modelos foram estimados pelo método da MV. Para avaliar a performance dos estimadores de MV, foram calculados o viés e a raiz do erro quadrático médio (REQM) das estimativas, como segue:

$$\begin{aligned}\text{Viés}(\widehat{\omega}_j) &= \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\omega}_j^{(m)} - \omega_j\right) \\ &= \left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\omega}_j^{(m)}\right) - \omega_j,\end{aligned}$$

$$\text{REQM}(\widehat{\omega}_j) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\omega}_j^{(m)} - \omega_j\right)^2},$$

em que $M = 1.000$ é o número de réplicas de Monte Carlo, e $\widehat{\omega}_j^{(m)}$ é a estimativa de MV do parâmetro ω_j na m-ésima amostra simulada.

Os resultados dessas duas medidas são apresentados nas Figuras 12-21, para cada modelo proposto e grau de desbalanceamento entre as classes. Na Figura 12, a única referente ao modelo

DL, observa-se que ambos o viés e a REQM das estimativas de MV dos parâmetros β_0 e β_1 apresentam valores baixos (próximos a zero) quando n cresce, o que era esperado (desejado).

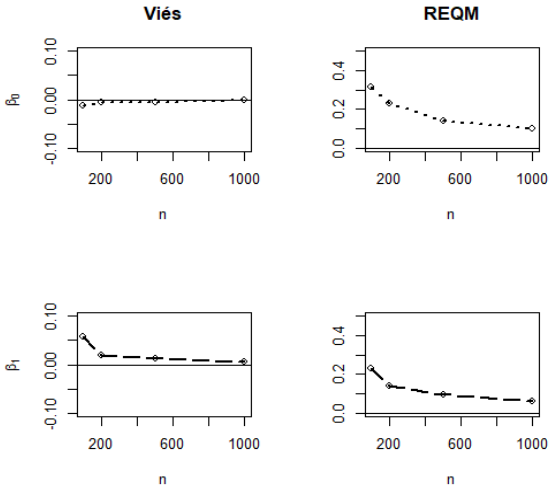

Figura 12: Viés e REQM das estimativas de MV dos parâmetros β_0 e β_1 do modelo DL, considerando diferentes tamanhos de amostra.

Para o modelo ADL, foram considerados os seguintes valores de λ: $\lambda = \exp\{0; 1{,}4; 2{,}4\}$, que correspondem, respectivamente, a 50%, 25% e 10% de uns nas amostras geradas. As Figuras 13, 14 e 15 ilustram os resultados obtidos para esses três casos, nas quais se observa que, de forma similar ao modelo DL, o viés e a REQM das estimativas de MV dos parâmetros β_0, β_1 e λ apresentam valores baixos quando n cresce, tornando as estimativas obtidas também melhores (isto é, mais acuradas e precisas) à medida que o tamanho amostral aumenta.

Assim como para o modelo ADL, considerou-se $\lambda = \exp\{0; 1{,}4; 2{,}4\}$ para o modelo PDL (ver Figuras 16, 17 e 18) e $\lambda = \exp\{0; -1{,}4; -2{,}4\}$ para o modelo RPDL (ver Figuras 19, 20 e 21), a fim de obter 50%, 25% e 10% de uns nas amostras geradas, respectivamente. Novamente, são observados comportamentos similares para o viés e a REQM das estimativas produzidas, que tendem a assumir valores menores à medida que n aumenta.

Em resumo, os resultados obtidos demonstram uma recuperação adequada dos parâmetros dos quatro modelos aqui desenvolvidos, segundo o procedimento de estimação proposto (máxima verossimilhança).

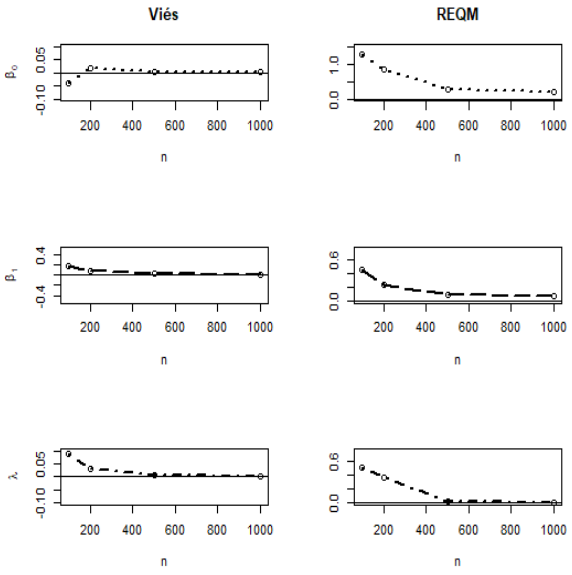

Figura 13: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo ADL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{0\}$.

A.2 *Misspecification*

O segundo estudo de simulação considerou $M = 1.000$ amostras (ou conjuntos de dados) de tamanho $n = 1.000$ de cada um dos modelos descritos, incluindo o modelo logístico. Foram fixados $\beta_0 = 0$, $\beta_1 = 1$ e $X \sim$ Uniforme$(-4, 4)$, e também estabelecidos diferentes valores para o parâmetro λ, de modo a obter cerca de 50%, 25% e 10% de uns nas amostras geradas dos modelos assimétricos (ADL, PDL e RPDL). Então, para cada amostra simulada de cada um dos cinco modelos de regressão binária, foi empregado o seguinte processo: ajustou-se o modelo verdadeiro mais os outros quatro modelos candidatos e, em seguida, calculou-se os valores de AIC, BIC e viés relativo absoluto das probabilidades de evento (uns) estimadas para cada um deles, selecionando como o melhor modelo aquele que obteve os menores valores desses critérios.

A Tabela 8 mostra a proporção de vezes em que cada modelo forneceu o melhor ajuste, de acordo com os critérios AIC e BIC. De modo geral, observa-se uma boa conformidade desses critérios

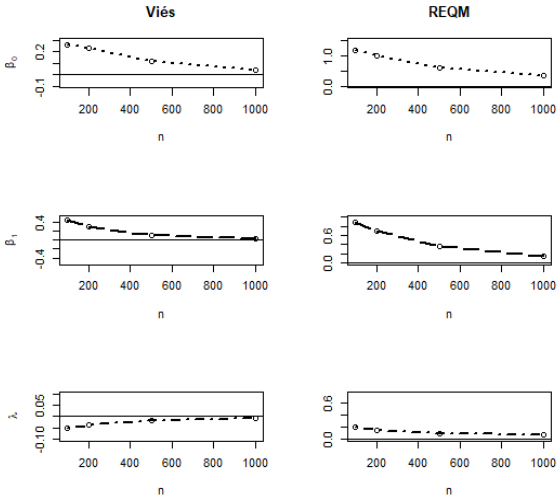

Figura 14: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo ADL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{1{,}4\}$.

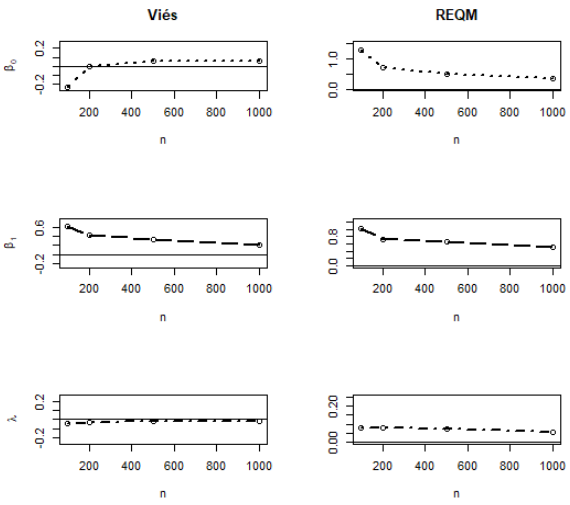

Figura 15: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo ADL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{2{,}4\}$.

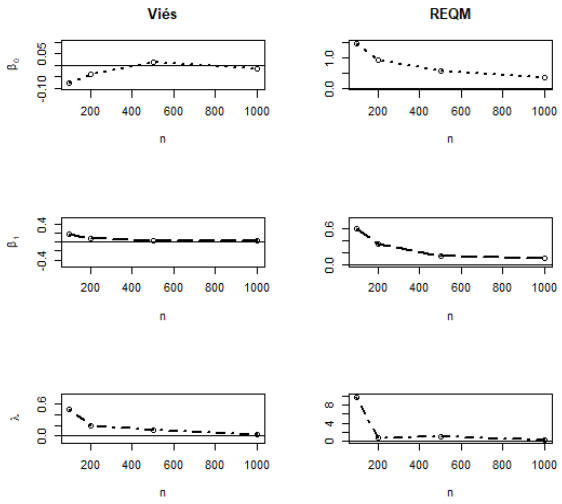

Figura 16: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo PDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{0\}$.

em discriminar os diferentes modelos estudados, sendo que ambos tiveram performances parecidas, porém com ligeira vantagem para o critério AIC nos resultados obtidos.

Na Tabela 9 são apresentados os resultados acerca do viés relativo absoluto na estimação das probabilidades de uns ("sucessos"). Observa-se que, nas situações de desbalanceamento entre as classes, os modelos assimétricos produzem as melhores estimativas das probabilidades de evento, superando, de maneira geral, os modelos logístico e DL, por apresentarem viés mais próximo a zero. Vale notar também que há uma certa superioridade do modelo DL sobre o modelo logístico nos casos em que o desbalanceamento é mais acentuado (10% de uns).

Tabela 8: Proporção de vezes em que cada modelo de regressão binária foi eleito o melhor segundo os critérios AIC e BIC.

	Modelo Ajustado									
	Logístico		DL		ADL		PDL		RPDL	
Modelo Verdadeiro	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Logístico	0,502	0,584	0,340	0,408	0,045	0,003	0,048	0,001	0,065	0,004
DL	0,346	0,396	0,505	0,593	0,065	0,007	0,039	0,002	0,045	0,002
ADL ($\lambda = \exp\{0\}$)	0,377	0,422	0,487	0,572	0,065	0,001	0,034	0,003	0,034	0,002
ADL ($\lambda = \exp\{1,4\}$)	0,000	0,022	0,000	0,000	0,677	0,672	0,082	0,066	0,241	0,240
ADL ($\lambda = \exp\{2,4\}$)	0,000	0,056	0,000	0,000	0,890	0,868	0,028	0,001	0,082	0,075
PDL ($\lambda = \exp\{0\}$)	0,386	0,427	0,490	0,569	0,045	0,002	0,040	0,001	0,039	0,001
PDL ($\lambda = \exp\{1,4\}$)	0,046	0,314	0,010	0,016	0,452	0,423	0,453	0,211	0,039	0,036
PDL ($\lambda = \exp\{2,4\}$)	0,200	0,152	0,005	0,005	0,181	0,124	0,562	0,688	0,052	0,031
RPDL ($\lambda = \exp\{0\}$)	0,377	0,424	0,497	0,570	0,041	0,002	0,041	0,002	0,044	0,002
RPDL ($\lambda = \exp\{-1,4\}$)	0,167	0,613	0,009	0,056	0,401	0,197	0,074	0,006	0,349	0,128
RPDL ($\lambda = \exp\{-2,4\}$)	0,306	0,777	0,027	0,028	0,332	0,170	0,062	0,000	0,273	0,025

Tabela 9: Viés relativo absoluto na estimação das probabilidades de evento (uns).

	Modelo Ajustado				
Modelo Verdadeiro	Logístico	DL	ADL	PDL	RPDL
Logístico	0,1406	0,1517	0,1424	0,1484	0,1549
DL	0,0346	0,0246	0,0497	0,0372	0,0367
ADL ($\lambda = \exp\{0\}$)	0,0392	0,0283	0,0125	0,0139	0,0173
ADL ($\lambda = \exp\{1,4\}$)	8,0692	9,9745	0,7554	1,4404	0,8434
ADL ($\lambda = \exp\{2,4\}$)	7,4711	1,1899	0,1218	8,7562	1,8160
PDL ($\lambda = \exp\{0\}$)	0,1155	0,1192	0,0987	0,0897	0,1097
PDL ($\lambda = \exp\{1,4\}$)	6,4018	9,2803	0,4189	0,4504	1,3930
PDL ($\lambda = \exp\{2,4\}$)	8,2054	2,2249	1,7093	0,3471	0,6567
RPDL ($\lambda = \exp\{0\}$)	0,0825	0,0868	0,0541	0,0630	0,0531
RPDL ($\lambda = \exp\{-1,4\}$)	0,2159	0,2888	0,0588	0,2031	0,0521
RPDL ($\lambda = \exp\{-2,4\}$)	0,4160	0,5066	0,3068	0,0849	0,2806

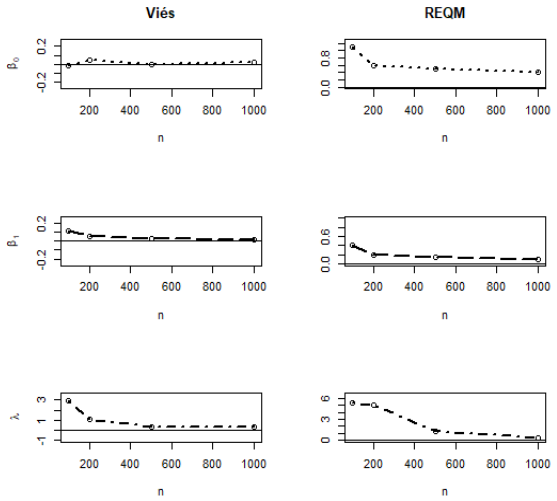

Figura 17: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo PDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{1,4\}$.

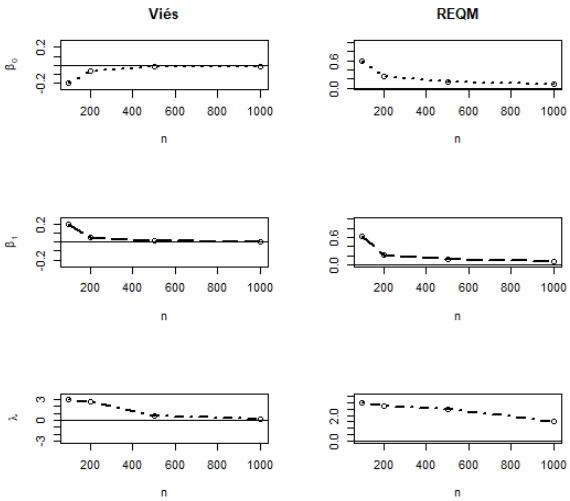

Figura 18: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo PDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{2,4\}$.

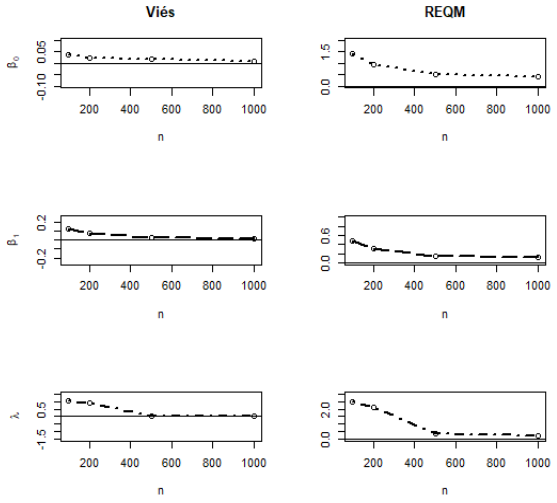

Figura 19: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo RPDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{0\}$.

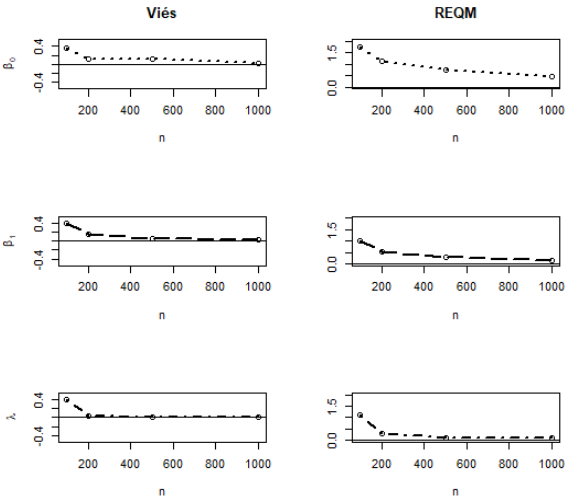

Figura 20: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo RPDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{-1{,}4\}$.

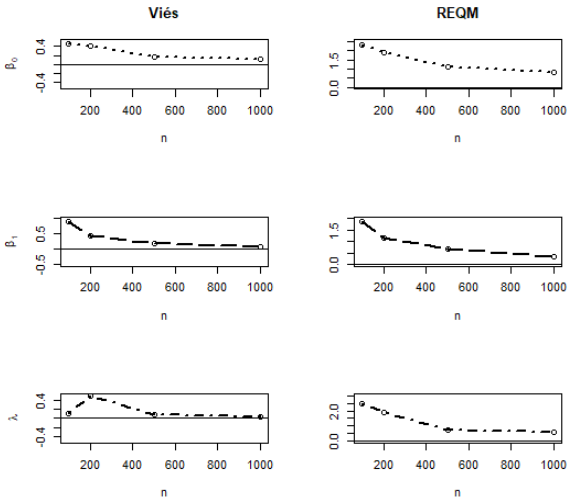

Figura 21: Viés e REQM das estimativas de MV dos parâmetros β_0, β_1 e λ do modelo RPDL, considerando diferentes tamanhos de amostra e $\lambda = \exp\{-2{,}4\}$.