Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em Data Science e Big Data

Ricardo Rasmussen Petterle

Modelo de regressão misto unit gamma para dados contínuos limitados na presença de zeros e uns

Curitiba
2022

Ricardo Rasmussen Petterle

Modelo de regressão misto unit gamma para dados contínuos limitados na presença de zeros e uns

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Wagner Hugo Bonat

Curitiba
2022

MODELO DE REGRESSÃO MISTO UNIT GAMMA PARA DADOS CONTÍNUOS LIMITADOS NA PRESENÇA DE ZEROS E UNS

Ricardo Rasmussen Petterle[1]

[1]Departamento de Estatística, Universidade Federal do Paraná Rua Evaristo F. F. da Costa 418, Jardim das Americas, 82590-300, Curitiba, PR, Brasil*

O objetivo deste trabalho é propor um modelo de regressão para lidar com dados contínuos limitados correlacionados no intervalo $[0,1]$. O modelo proposto é especificado por uma distribuição de probabilidade de mistura, na qual a distribuição unit gamma é usada para descrever a parte contínua e a distribuição Bernoulli a parte discreta (zeros e uns). Para estimação dos parâmetros e inferência, fez-se uso do método de máxima verossimilhança. A implementação computacional foi feita em linguagem `C++` por meio do pacote `TMB` do *software* estatístico `R`, o qual permite combinar o método de aproximação de Laplace com diferenciação automática. O modelo proposto foi motivado por um conjunto de dados que avaliou a proporção de sítios doentes correspondentes a um tipo específico de dente (molar, pré-molar, canino e incisivo) em indivíduos com doença periodontal. Em particular, os valores zero e um indicam casos livre de doença e altamente doentes, respectivamente. De acordo com medidas de bondade de ajuste, o modelo proposto superou o modelo de regressão misto beta aumentado de zeros e uns na análise dos dados.
Palavras-chave: Modelo misto beta, Intervalo unitário, Diferenciação automática, Dados correlacionados, Aproximação de Laplace, Doença periodontal.

The main goal of this work is to propose a regression model to deal with correlated continuous bounded data in the interval $[0,1]$. The proposed model is specified by a mixed probability distribution, in which the unit gamma distribution is used to describe the continuous part and the Bernoulli distribution the discrete part (zeros and ones). For parameter estimation and inference, we adopted the maximum likelihood method. The computational implementation was done in `C++` language using the `TMB` package of the `R` statistical software, which allows to combine the Laplace approximation method with automatic differentiation. The proposed model was motivated by a dataset that evaluated the proportion of diseased sites corresponding to a specific tooth type (molar, premolar, canine and incisor) in individuals with periodontal disease. In particular, the values zero and one indicates disease-free and highly diseased cases, respectively. According to goodness of fit measures, the proposed model outperforms the augmented mixed beta regression model in data analysis.
Keywords: Beta mixed model, Unit interval, Automatic differentiation, Correlated data, Laplace approximation, Periodontal disease.

1. Introdução

Modelos de regressão são usados em muitas áreas de pesquisa para descrever o comportamento de uma variável resposta e função de um conjunto de covariáveis. Para analisar respostas no intervalo $(0,1)$, como taxas, proporções, índices e porcentagens, diversos modelos de regressão foram propostos nos últimos anos. Em geral, tais modelos permitem modelar a média da variável resposta em função de covariáveis [1], [2], assim como a moda [3] e os quantis [4]. No entanto, para analisar dados com estruturas longitudinal, espacial, espaço/temporal e medidas repetidas é necessário um modelo de regressão que considere tais estruturas de dependência. Logo, os modelos de regressão marginais [5] e mistos [6], exemplificam algumas abordagens para lidar com dados correlacionados no intervalo unitário. Além disso, dados que possuem zeros e uns exatos também precisam de um modelo de regressão específico, uma vez que as distribuições de probabilidade para proporções acomodam apenas dados no intervalo $(0,1)$. Assim, a análise de dados no intervalo $[0,1]$ é feita, em geral, usando uma distribui-

*ricardopetterle@ufpr.br , wbonat@ufpr.br

ção de mistura, onde a parte contínua é descrita por uma distribuição de probabilidade para dados no intervalo $(0,1)$ e a parte discreta (zeros e uns) é modelada pela distribuição Bernoulli.

O principal objetivo deste trabalho é propor um modelo de regressão misto para lidar com dados contínuos limitados no intervalo $[0,1]$. Em particular, a distribuição unit gamma [8] foi usada para descrever a parte contínua e a distribuição Bernoulli a parte discreta. A especificação do modelo é feita com base na estrutura dos modelos lineares generalizados mistos e, o método da máxima verossimilhança foi usado para estimação dos parâmetros e inferência. Desse modo, este trabalho estende o modelo de regressão proposto em [6], o qual não permite modelar dados no intervalo $[0,1]$. O modelo proposto foi motivado por um conjunto de dados da área da saúde, o qual não é facilmente manipulado pelos métodos estatísticos convencionais. Tal conjunto de dados se refere a proporção de sítios doentes correspondentes a um tipo específico de dente (molar, pré-molar, canino e incisivo) em indivíduos com doença periodontal. É importante destacar que esse conjunto de dados apresenta duas características que devem ser levadas em conta na análise dos dados: (i) medidas repetidas avaliadas no mesmo indivíduo e (ii) presença de zeros e uns, onde zero indica casos livre da doença e o valor um se refere aos casos altamente doentes.

O artigo está organizado da forma que segue. A Seção 2 descreve o conjunto de dados motivador deste trabalho. Seção 3 propõem o modelo de regressão misto unit gamma para lidar com dados correlacionados no intervalo $[0,1]$ e, Seção 4 detalha o método usado para estimação e inferência. Na Seção 5 são apresentados os principais resultados da análise dos dados. Por fim, a Seção 6 discute os principais resultados obtidos e apresenta sugestões para futuros trabalhos.

2. Conjunto de dados

O conjunto de dados utilizado neste trabalho corresponde a um estudo clínico conduzido por [7] para avaliar o status e a progressão de doença periodontal em indivíduos com diabetes tipo 2. Neste estudo, foram avaliados seis sítios em cada um dos 28 dentes (oito molares, oito pré-molares, quatro caninos e oito incisivos). Um dos principais marcadores de doença periodontal é o nível de inserção clínica (NIC). Assim, quando se tem valores do NIC maior ou igual do que 3 mm um sítio é classificado como doente. A proporção de sítios doentes, para cada um dos quatro tipos de dentes, foi calculada dividindo-se o número de sítios doentes pelo número de sítios. Por exemplo, para dentes molares e caninos tem-se 48 e 24 sítios, respectivamente. Além disso, os valores zero (9,8%) e um (8,1%) correspondem aos casos livre de doença e aqueles altamente doentes, respectivamente.

O conjunto de dados tem 1160 observações (290 indivíduos × 4 dados agrupados). Os tipos de dentes representam os dados agrupados avaliados no mesmo indivíduo e as covariáveis são: `genêro` - gênero do paciente (0: Masculino; 1: Feminino); `idade` - idade do paciente (anos); `HbA1c` - indicador de status de hemoglobina glicada (0: Controlada; 1: Não-controlada); `fumante` - condição de fumante (0: Não-fumante; 1: Fumante); `dente` - tipos de dentes (0: Molar; 1: Pré-molar; 2: Canino; 3: Incisivo). O principal objetivo da análise dos dados é investigar o efeito das covariáveis na variável resposta (proporção de sítios doentes para um tipo específico de dente) levando em conta a estrutura de dados agrupados, além dos valores zeros e uns. A Figura 1 mostra o comportamento da variável resposta em função das covariáveis.

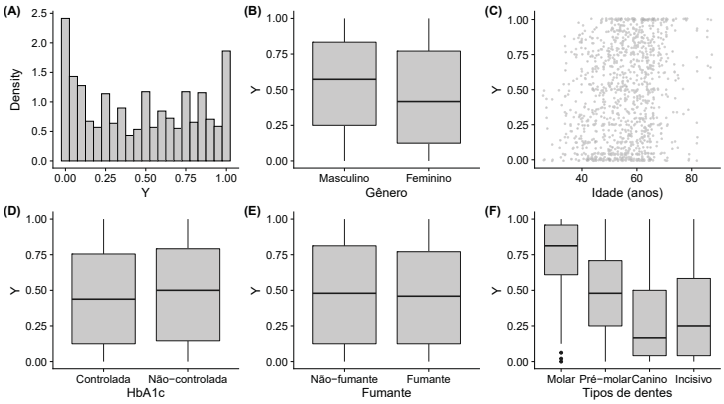

Figura 1: Análise descritiva para o conjunto de dados de doença periodontal.

De acordo com a Figura 1(B e F), há um indicativo de que os homens e o tipo de dente molar são os casos em que os maiores valores da variável resposta são observados. No entanto, a Figura 1(D e E) sugerem que a hemoglobina glicada (`HbA1c`) e o status de fumante não influenciam na proporção de sítios doentes.

3. Modelo de regressão misto unit gamma para dados no intervalo $[0,1]$

3.1. Distribuição unit gamma

Seja $Y \sim \text{UG}([\mu^{1/\phi}/(1-\mu^{1/\phi})],\phi)$ uma variável aleatória com distribuição unit gamma. Sua função densidade de probabilidade (fdp) é dada por

$$f_{\text{UG}}(y;\mu,\phi) = \frac{\left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}\right)^{\phi}}{\Gamma(\phi)} y^{\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}-1}\left[\log\left(\frac{1}{y}\right)\right]^{\phi-1}, \quad (1)$$

onde y e $\mu \in (0,1)$ e $\phi > 0$ é o parâmetro de precisão. Logo, a esperança e variância de Y são dadas, respectivamente, por

$$\text{E}(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \mu\left[\left(\frac{1}{(2-\mu^{1/\phi})\phi} - \mu\right)\right].$$

Note que, quando $\phi \to 0$ a $\text{Var}(Y) \to \mu(1-\mu)$, correspondendo a relação média-variância da distribuição Bernoulli.

3.2. Modelo de regressão unit gamma

Considere $Y_1,\ldots,Y_n$ variáveis aleatórias independentes, onde $Y_i \sim \text{UG}([\mu^{1/\phi}/(1-\mu^{1/\phi})],\phi)$. Assim, o modelo de regressão unit gamma é especificado por

$$g(\mu_i) = \eta_i = \mathbf{x}_i^{\top}\boldsymbol{\beta}, \quad (2)$$

onde $\mathbf{x}_i$ e $\boldsymbol{\beta}$ são vetores $p \times 1$ de covariáveis conhecidas e desconhecidos parâmetros de regressão, respectivamente, e $g(\cdot)$ é uma função de ligação conhecida.

3.3. Distribuição unit gamma para dados no intervalo $[0,1]$

É importante destacar que o modelo de regressão apresentado em (3.2) não permite analisar dados no intervalo $[0,1]$, uma vez que o suporte da distribuição unit gamma é o intervalo $(0,1)$, o que não permite incluir zeros e uns. Além disso, tal modelo de regressão permite analisar apenas dados independentes.

Assim, a distribuição UG (3.1) foi "aumentada" para acomodar zeros e uns, bem como o seu respectivo modelo de regressão foi estendido para levar em conta a estrututa de dados agrupados por meio de efeitos aleatórios. Seja $Y \sim \text{UGA}(\mu,\phi,p_0,p_1)$ uma variável aleatória com distribuição unit gamma aumentada (UGA), cuja expressão é dada por

$$f_{\text{UGA}}(y|p_0,p_1,\mu_i,\phi) = \begin{cases} p_0 & \text{se } y = 0 \\ p_1 & \text{se } y = 1 \\ (1-p_0-p_1)f_{\text{UG}}(y;\mu,\phi) & \text{se } y \in (0,1), \end{cases}$$

onde $p_0 = P(Y=0)$, $p_1 = P(Y=1)$ para $p_0, p_1 \geq 0$, $0 \leq p_0 + p_1 \leq 1$ e $f_{\text{UG}}(y;\mu,\phi)$ é a fdp apresentada na Equação (1). Portanto, a média e a variância da distribuição UGA são dadas, respectivamente, por

$$\begin{aligned}\text{E}(Y) &= (1-p_0-p_1)\mu + p_1, \\ \text{Var}(Y) &= p_1(1-p_1) + (1-p_0-p_1)\left\{\mu\left[\left(\frac{1}{(2-\mu^{1/\phi})\phi} - \mu\right)\right]\right. \\ &\quad \left. + (p_0+p_1)\mu^2 - 2\mu p_1\right\}\end{aligned}$$

3.4. Modelo de regressão misto unit gamma para dados no intervalo $[0,1]$

Para definir o modelo de regressão com efeitos aleatórios, considere $y_{ij} \in [0,1]$ uma observação da variável aleatória Y_{ij} para $i = 1,\ldots,N$ indivíduos e $j = 1,\ldots,n_i$ dados agrupados. Assim, a especificação do modelo proposto é dada pela seguinte estrutura

$$\begin{aligned} Y_{ij}|\mathbf{u}_i &\overset{i.i.d.}{\sim} \text{UGA}(\mu,\phi,p_0,p_1) \\ \text{logit}(\mu_{ij}) &= \mathbf{x}_{ij1}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\top}\mathbf{u}_i, \\ \log(\phi_{ij}) &= \mathbf{x}_{ij2}^{\top}\boldsymbol{\psi}, \\ \text{logit}(p_{0ij}) &= \mathbf{x}_{ij3}^{\top}\boldsymbol{\gamma}, \\ \text{logit}(p_{1ij}) &= \mathbf{x}_{ij4}^{\top}\boldsymbol{\delta}, \end{aligned}$$

onde $\mathbf{x}_{ij1}$ e $\mathbf{z}_{ij}$ são vetores de covariáveis conhecidas $p \times 1$ e $q \times 1$, respectivamente, $\boldsymbol{\beta}$ é um vetor com os coeficientes de regressão e $\mathbf{u}_i$ é um vetor de efeitos aleatórios ambos associados a μ_{ij}. Nesta notação, $\mathbf{u}_i \sim N(\mathbf{0},\Sigma)$, onde Σ é a matriz de covariância dos efeitos aleatórios. De forma similar, $\mathbf{x}_{ij2}$ e $\boldsymbol{\psi}$ são vetores $k \times 1$ de covariáveis conhecidas e desconhecidos parâmetros de regressão associados a estrutura de precisão. Por fim, $\mathbf{x}_{ij3}$ e $\mathbf{x}_{ij4}$ são vetores $r \times 1$ e $s \times 1$ de covariáveis conhecidas associadas a p_{0ij} e p_{1ij}, nos quais $\boldsymbol{\gamma}$ e $\boldsymbol{\delta}$ são seus respectivos vetores de coeficientes de regressão.

4. Estimação e inferência

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top},\boldsymbol{\phi}^{\top},\boldsymbol{\gamma}^{\top},\boldsymbol{\delta}^{\top},\Sigma)^{\top}$ o vetor de parâmetros e considere $\mathbf{y}_i = (y_{i1},\ldots,y_{in_i})^{\top}$ o vetor com dados agrupados do i-ésimo indivíduo e $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_N)^{\top}$ o vetor com todas as observações dos indivíduos. A verossimilhança marginal para cada indivíduo é dada por

$$\mathscr{L}(\boldsymbol{\theta}|\mathbf{y}_i) = \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{u}_i,\boldsymbol{\theta}) f(\mathbf{u}_i|\Sigma) d\mathbf{u}_i,$$

onde $f_{ij}(y_{ij}|\mathbf{u}_i,\boldsymbol{\theta})$ é a distribuição proposta na subseção 3.3 e $f(\mathbf{u}_i|\Sigma)$ é a densidade normal com média zero e matriz de covariância Σ.

Supondo que os N indivíduos são independentes, a verossimilhança marginal completa é dada por

$$\mathscr{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{u}_i,\boldsymbol{\theta}) f(\mathbf{u}_i|\Sigma) d\mathbf{u}_i, \qquad (3)$$

e sua respectiva função de log-verossimilhança é expressa por

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{y}) &= \sum_{i=1}^{N} \log\{\mathscr{L}(\boldsymbol{\theta}|\mathbf{y}_i)\} \\ &= \ell_1(\boldsymbol{\gamma}) + \ell_2(\boldsymbol{\delta}) + \ell_3(\boldsymbol{\beta},\boldsymbol{\phi}) + \ell_4(\mathbf{u}),\end{aligned}$$

onde,

- $\ell_1(\boldsymbol{\gamma}) = \sum_{y_{ij}=0} \log(p_{0ij})$;
- $\ell_2(\boldsymbol{\delta}) = \sum_{y_{ij}=1} \log(p_{1ij})$;
- $\ell_3(\boldsymbol{\beta},\boldsymbol{\phi}) = \sum_{y_{ij}\in(0,1)} \ell(\mu_{ij},\phi_{ij})$,

em que,

$$\begin{aligned}\ell(\mu_{ij},\phi_{ij}) = \log(1-p_{0ij}-p_{1ij}) + \big[&\phi_{ij}\log(d_{ij}) - \log\big(\Gamma(\phi_{ij})\big) \\ &+ (d_{ij}-1)\log(y_{ij}) + (\phi_{ij}-1)\log(-\log(y_{ij}))\big],\end{aligned}$$

e $d_{ij} = \big[\mu_{ij}^{1/\phi_{ij}} / \big(1-\mu_{ij}^{1/\phi_{ij}}\big)\big]$. Ainda,

- $\ell_4(\mathbf{u}) = \sum_{y_{ij}\in(0,1)} \ell(\mu_{ij})$,

uma vez que $\ell(\mu_{ij}) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbf{u}_i^\top \Sigma^{-1}\mathbf{u}_i$ é a parte de μ associada aos efeitos aleatórios.

Note que, a integral 3 tem dimensão q e precisa ser resolvida N vezes. Além disso, tal integral não possui solução analítica. Desse modo, é necessário um método numérico para obter uma solução aproximada. Neste trabalho, adotou-se o método de aproximação de Laplace. Tal método foi combinado com diferenciação automática e implementado em linguagem `C++` fazendo-se uso do pacote `TMB` [10] do *software* estatístico `R`. Para mais detalhes sobre modelos de regressão para dados no intervalo $(0,1)$ usando o `TMB`, ver [6] e [9].

5. Análise dos dados

Nesta seção, nós usaremos o modelo de regressão proposto na Seção 3.4 para analisar o conjunto de dados apresentado na Seção 2.

O principal objetivo da análise dos dados é investigar o relacionamento de um conjunto de covariáveis com a variável resposta. Em particular, a variável resposta percente ao intervalo $[0,1]$ e corresponde a proporção de sítios doentes, considerando quatro tipos de dentes (molar, pré-molar, canino e incisivo).

Seja $Y_{ij}|\mathbf{u}_i \sim \text{UGA}(\mu_{ij},\phi,p_{0ij},p_{1ij})$ para $i = 1,\ldots,290$ indivíduos e $j = 1,\ldots,4$ tipos de dentes. Seguindo [11] o preditor linear para μ_{ij}, p_{0ij} e p_{1ij} é representado por

$$\begin{aligned}\text{logit}(\mu_{ij}) &= \mathbf{x}_{ij}^\top\boldsymbol{\beta} + \mathbf{u}_i, \\ \text{logit}(p_{0ij}) &= \mathbf{x}_{ij}^\top\boldsymbol{\gamma}, \\ \text{logit}(p_{1ij}) &= \mathbf{x}_{ij}^\top\boldsymbol{\delta},\end{aligned} \qquad (4)$$

onde,

$\mathbf{x}_{ij}^\top = (1, \texttt{gênero}_i, \texttt{idade}, \texttt{HbA1c}_i, \texttt{fumante}_i, \texttt{molar}_{ij}, \texttt{pré-molar}_{ij}, \texttt{incisivo}_{ij})$ é o vetor de covariáves,

$\boldsymbol{\beta}$ é o vetor de coeficientes de regressão associados a μ_{ij} e $\mathbf{u}_i$ é o vetor de efeitos aleatórios associados aos indivíduos. Por fim, $\boldsymbol{\gamma}$ e $\boldsymbol{\delta}$ são vetores de coeficientes de regressão, ambos associados às estruturas p_{0ij} e p_{1ij}, respectivamente. Para as covariáveis categóricas, as categorias de referência são: (i) gênero masculino; (ii) HbA1c controlada; (iii) não-fumante e (iv) tipo de dente canino.

Logo, o modelo de regressão proposto neste trabalho foi ajustado considerando duas estruturas:

- **Modelo 1**: leva em conta apenas a modelagem de μ_{ij} em função de covariáveis, considerando p_0 e p_1 constantes, isto é, ambas estruturas não são modeladas pelo efeito de covariáveis;
- **Modelo 2**: leva em conta a modelagem de μ_{ij}, p_{0ij} e p_{1ij} em função do mesmo conjunto de covariáveis, conforme apresentado em (4).

Além do modelo de regressão proposto neste trabalho, ajustou-se o modelo de regressão misto beta aumentado de zeros e uns [11], onde os resultados obtidos são comparados por meio de medidas de bondade de ajuste. É importante ressaltar, que em ambos os modelos, assumiu-se precisão constante ao longo das observações e o efeito aleatório foi considerado apenas na estrutura de média. No entanto, o modelo proposto na seção 3.4 permite incluir facilmente o efeito de covariáveis na estrutura de precisão, assim como incluir diferentes covariáveis em cada uma das estruturas de regressão.

A Tabela 1 apresenta as estimativas dos parâmetros e erros-padrão obtidos pelos modelos de regressão misto unit gamma e beta aumentados de zeros e uns, considerando as estruturas de regressão descritas anteriormente (modelos 1 e 2). Esta tabela também apresenta medidas de bondade de ajuste, valor da função de log-verossimilhança maximizada (`Loglik`) e os critérios de informação de *Akaike* (`AIC`) e Bayesiano (`BIC`). De acordo com tais medidas, o modelo 1 com distribuição unit gamma se ajustou melhor aos dados do que o modelo beta. A diferença em termos de `Loglik` foi 21,52. Além disso, os critérios `AIC` e `BIC` também apontaram na mesma direção. Com relação ao modelo 2, o mesmo pode ser observado, mostrando portanto que o modelo de regressão misto unit gamma aumentado de zeros e uns se ajustou melhor aos dados, quando comparado ao modelo misto beta. Assim, selecionou-se o modelo 2 com distribuição unit gamma para interpretação dos resultados. Para o parâmetro μ_{ij} pode-se interpretar as estimativas dos parâmetros como a proporção esperada de sítios doentes. Nesse caso, as covariáveis gênero, idade e tipo de dentes foram significativamente diferentes de zero e explicam a variação da variável resposta. Por outro lado, as covariáveis `HbA1c` e `fumante` apresentaram $p > 0,05$. Assim, quanto maior a idade dos indivíduos, maior será o valor esperado da variável resposta. Esse aumento foi estimado em $\hat{\beta}_2 = \exp(0,0343) = 1,035$. Tal resultado indica que a proporção esperada de sítios doentes aumenta 3,5% a cada ano. Já para as mulheres, espera-se que essa proporção seja menor quando comparada aos homens. Para o tipo de dente molar, espera-se que a proporção de sítios doentes seja $\hat{\beta}_5 = \exp(2,1109) \approx 8,3$ vezes maior do que o tipo de dente canino (categoria de referência). O parâmetro p_{0ij} pode ser interpretado com a chance de tipo de dente livre de doença versus doente. Logo, as covariáveis gênero, idade e tipo de dente explicam casos livres de doença periodontal. Com relação ao parâmetro p_{1ij}, este pode ser interpretado como a chance de dentes altamente doentes versus aqueles livre de doença. Em particular, as covariáveis gênero, idade e tipo de dente molar foram significativamente diferentes de zeros e explicam os casos altamente doentes.

Tabela 1: Estimativa dos parâmetros (Est), erro padrão (EP) e medidas de bondade de ajuste por modelo.

Parâmetro	Unit gamma[1]	Beta[1]	Unit gamma[2]	Beta[2]
	Est(EP)	Est(EP)	Est(EP)	Est(EP)
β_0: Intercepto	−2,5163(0,4120)*	−2,5075(0,4071)*	−2,4712(0,4127)*	−2,4816(0,4066)*
β_1: Gênero	−0,5367(0,1653)*	−0,5319(0,1643)*	−0,5391(0,1657)*	−0,5330(0,1642)*
β_2: Idade	0,0347(0,0067)*	0,0327(0,0066)*	0,0343(0,0067)*	0,0325(0,0066)*
β_3: HbA1c	0,0896(0,1426)	0,0772(0,1412)	0,0749(0,1429)	0,0726(0,1411)
β_4: Fumante	0,1158(0,1529)	0,1329(0,1514)	0,1097(0,1533)	0,1268(0,1513)
β_5: Molar	2,1099(0,0782)*	2,1501(0,0845)*	2,1109(0,0781)*	2,1492(0,0844)*
β_6: Pré-molar	0,8105(0,0719)*	0,8578(0,0722)*	0,8108(0,0718)*	0,8557(0,0721)*
β_7: Incisivo	0,1868(0,0755)*	0,1943(0,0711)*	0,1859(0,0754)*	0,1935(0,0710)*
γ_0: Intercepto	−2,2125(0,0985)*	−2,2207(0,0988)*	0,6227(0,06266)	0,6266(0,6262)
γ_1: Gênero	–	–	1,0488(0,3019)*	1,0492(0,3016)*
γ_2: Idade	–	–	−0,0411(0,0101)*	−0,0411(0,0101)*
γ_3: HbA1c	–	–	−0,3368(0,2137)	−0,3373(0,2136)
γ_4: Fumante	–	–	−0,3687(0,2277)	−0,3636(0,2277)
γ_5: Molar	–	–	−4,5171(1,0032)*	−4,5384(1,0109)*
γ_6: Pré-molar	–	–	−2,4157(0,3892)*	−2,4147(0,3879)*
γ_7: Incisivo	–	–	−0,6818(0,2277)*	−0,6901(0,2277)*
δ_0: Intercepto	−2,4267(0,1075)*	−2,4338(0,1079)*	−8,1830(0,8923)*	−8,2169(0,8950)*
δ_1: Gênero	–	–	−0,6557(0,2702)*	−0,6593(0,2706)*
δ_2: Idade	–	–	0,0857(0,0127)*	0,0862(0,0127)*
δ_3: HbA1c	–	–	0,3333(0,2485)	0,3266(0,2487)
δ_4: Fumante	–	–	−0,3069(0,2484)	−0,3113(0,2488)
δ_5: Molar	–	–	2,4901(0,3957)*	2,4991(0,3975)*
δ_6: Pré-molar	–	–	0,2137(0,4887)	0,2329(0,4894)
δ_7: Incisivo	–	–	0,3285(0,4789)	0,3339(0,4810)
ϕ	3,2877(0,1695)*	7,6135(0,4334)*	3,2922(0,1695)*	7,6343(0,4341)*
τ_1	1,0963(0,0554)*	1,0820(0,0579)*	1,0994(0,0556)*	1,0814(0,0579)*
LogLik	−369,85	−391,37	−228,03	−249,55
AIC	763,70	806,73	508,06	551,09
BIC	824,38	867,41	639,52	682,56

Nota: *Indica p-valor < 0,05; [1]modelo 1; [2]modelo 2.

6. Discussão

Neste trabalho, nós estendemos o modelo de regressão proposto em [6] para lidar com dados agrupados no intervalo $[0,1]$. Dessa forma, a distribuição unit gamma foi "aumentada" para acomodar zeros e uns, uma vez que seu suporte é o intervalo $(0,1)$. A implementação computacional foi feita usando templates em `C++` por meio do pacote `TMB` do *software* estatístico `R`, o qual permite combinar facilmente diferenciação automática com o método de aproximação de Laplace. Já a estimação dos parâmetros foi feita sob o paradigma de verossimilhança. O conjunto de dados motivador deste trabalho quantificou a progressão de doença periodontal em indivíduos com diabetes tipo 2. Em particular, avaliou-se a proporção de sítios doentes em quatro tipos de dentes (molar, pré-molar, canino e incisivo). Além do parâmetro de média, modelou-se a probabilidade de zeros e uns em função de um conjunto de covariáveis usando uma função de ligação padrão para dados binários. No caso, fez-se uso da função de ligação *logit*. No entanto, outras funções de ligação também podem ser consideradas na análise dos dados, como a *probit*, complemento log-log, *cauchit* dentre outras. Medidas de bondade de ajuste (`Loglik`, `AIC` e `BIC`) indicaram que o modelo de regressão proposto neste trabalho apresentou um melhor ajuste aos dados quando comparado ao bem conhecido modelo de regressão misto beta aumentado de zeros e uns. Sugestões para futuros trabalhos incluem: (i) realizar um estudo de simulação para checar as propriedades dos estimadores de máxima verossimilhança para lidar com dados correlacionados no intervalo $[0,1]$; (ii) aplicar o modelo proposto em outros conjuntos de dados;

(iii) propor um modelo de regressão multivariado para lidar com múltiplas respostas no intervalo $[0, 1]$.

Referências

[1] Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* 31(7): 799–815.

[2] Bonat, W. H., Petterle, R. R., Hinde, J. and Demétrio, C. G. (2019). Flexible quasi-beta regression models for continuous bounded data, *Statistical Modelling* 19(6): 617–633.

[3] Menezes, A. F., Mazucheli, J. and Chakraborty, S. (2021). A collection of parametric modal regression models for bounded data, *Journal of Biopharmaceutical Statistics* 31(4): 490–506.

[4] Mazucheli, J., Alves, B., Menezes, A. F. and Leiva, V. (2022). An overview on parametric quantile regression models and their computational implementation with applications to biomedical problems including COVID-19 data, *Computer Methods and Programs in Biomedicine* p. 106816.

[5] Petterle, R. R., Bonat, W. H. and Scarpin, C. T. (2019). Quasi-beta longitudinal regression model applied to water quality index data, *Journal of Agricultural, Biological and Environmental Statistics* 24(2): 346–368.

[6] Petterle, R. R., Taconeli, C. A., da Silva, J. L., da Silva, G. P., Laureano, H. A. and Bonat, W. H. (2021). Unit gamma mixed regression models for continuous bounded data, *Journal of Statistical Computation and Simulation* pp. 1–19.

[7] Fernandes, J., Salinas, C., London, S., Wiegand, R., Hill, E., Slate, E., Grewal, J., Werner, P., Sanders, J. and Lopes-Virella, M. (2006). Prevalence of periodontal disease in Gullah African American diabetics, *Journal of Dental Research* 85(997).

[8] Mousa, A. M., El-Sheikh, A. A. and Abdel-Fattah, M. A. (2016). A gamma regression for bounded continuous variables, *Advances and Applications in Statistics* 49(4): 305.

[9] Petterle, R. R., Laureano, H. A., da Silva, G. P. and Bonat, W. H. (2021). Multivariate generalized linear mixed models for continuous bounded outcomes: Analyzing the body fat percentage data, Statistical Methods in Medical Research 30(12): 2619–2633.

[10] Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation, Journal of Statistical Software 70(5).

[11] Galvis, D. M., Bandyopadhyay, D. and Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data, Statistics in Medicine 33(21): 3759–3771.