UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

ERIC OLIVEIRA ROCHA

MODELO DE REGRESSÃO GAMA UNITÁRIA

FORTALEZA
2022

ERIC OLIVEIRA ROCHA

MODELO DE REGRESSÃO GAMA UNITÁRIA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador : Prof. Dr. Juvêncio Santos Nobre

FORTALEZA

2022

ERIC OLIVEIRA ROCHA

MODELO DE REGRESSÃO GAMA UNITÁRIA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Juvêncio Santos Nobre (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Luis Gustavo Bastos Pinho
Universidade Federal do Ceará (UFC)

Profa. Dra. Maria Jacqueline Batista
Universidade Federal do Ceará (UFC)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir.

AGRADECIMENTOS

Ao Prof. Dr. Juvêncio Santos Nobre e ao Prof. Luis Gustavo Bastos Pinho por se disponibilizarem a me orientar nesse trabalho de monografia.

A Profa. Dra. Maria Jacqueline pela orientação no projeto de iniciação à docência durante a graduação.

Agradeço a todos os professores do Departamento de Estatística e Matemática Aplicada por me proporcionar o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

"A vida é sofrida, mas não vou chorar
Viver de quê? Eu vou me humilhar?
É tudo uma questão de conhecer o lugar..."

(Racionais MC'S - Eu sou 157)

RESUMO

Modelos de regressão com suporte limitado no $(0, 1)$ são úteis quando se tem o interesse em modelar taxas e proporções, nesse sentido temos o modelo de regressão beta devido a Ferrari e Cribari-Neto (2004, Journal of applied statistics) e mais recentemente o modelo de regressão gama unitária proposto por Mousa et al. (2016, Advances and Applications in Statistics) que tem como base a distribuição gama unitária devido Grassia (1977, Australian Journal of Statistics) sendo obtida por meio de uma transformação logarítmica de uma variável com distribuição gama e possui a característica de ser bastante flexível, podendo assumir formas simétricas e assimétricas em (0,1). Portanto, tem-se um modelo regressão para situações em que a variável resposta é contínua com suporte limitado, isso inclui taxas e proporções tendo assim uma grande diversidade de aplicações práticas. Além disso, o modelo de regressão associado é reparametrizado de forma similar ao modelo beta proposto por Ferrari e Cribari-Neto (2004, Journal of applied statistics), i.e., expresso em termos de sua média e de um parâmetro de dispersão. Neste trabalho temos como objetivo apresentar os aspectos teóricos e práticos do modelo, foi utilizado um conjunto de dados reais no qual se considerou o ajuste e a análise de diagnóstico no software livre `R` dos modelos de regressão gama unitária e beta, no qual foi possível verificar uma melhor adequabilidade do modelo de regressão gama unitária aos dados, evidenciando a sua vantagem.

Palavras-chave: modelos de regressão. distribuição beta. distribuição gama unitária. modelos de regressão beta. software `R`.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

fdp	função densidade de probabilidade
fda	função distribuição acumulada

LISTA DE SÍMBOLOS

$\mathbb{1}(.)$	função indicadora
$\Gamma(.)$	função gama completa
$\Gamma(t;x)$	função gama incompleta com limite x no ponto t
iid	independente e identicamente distribuída
$\sim$	está distribuído como ou segue distribuição
$\mathcal{UG}$	distribuição gama unitária
$\mathcal{N}$	distribuição normal

SUMÁRIO

1 INTRODUÇÃO

Distribuições no intervalo unitário possuem aplicações em problemas que envolvem, principalmente taxas e proporções, logo surge um leque de aplicações como em problemas de inoculação discutido em Moran (1954) e Grassia (1977), em estudos de dados composicionais, encontrados em Geologia e Biologia e análise de danos (Oguamanam et al., 1995). Ou seja, as distribuições no intervalo unitário ou qualquer variável resposta com suporte limitado no intervalo (a,b), com $a < b$ conhecidos são muito úteis, assim se faz necessário o conhecimento de algumas dessas distribuições quando se tem o interesse em modelar dados com essas características.

Existem diversas propostas de distribuições para modelagem no intervalo unitário. Nesse sentido temos a distribuição Kumaraswamy proposta por Kumaraswamy (1980) que é uma alternativa frente a distribuição beta (Gupta e Nadarajah, 2004). Outra alternativa é devido a Smithson e Merkle (2013) que propõem distribuições para o intervalo unitário induzidas por transformações. Ainda nesse sentido, Lemonte e Bazán (2016) a partir da distribuição Johnson S_b, desenvolveram uma classe de distribuições com suporte limitado com base na família simétrica de distribuições. Além disso, Rodrigues et al. (2019) propuseram um mecanismo no qual possibilita construir distribuições de probabilidade em intervalos limitados, em particular o intervalo (0,1). Mais recente Mazucheli et al. (2020) propuseram a distribuição Weibull unitária como alternativa a distribuição de Kumaraswamy para a modelagem de quantis.

Em geral os modelos de regressão possuem uma estrutura em que temos uma variável denotada como variável resposta e uma ou mais variáveis explicativas. No caso dos modelos de regressão lineares usuais, é comum admitir normalidade e uma série de suposições para que os resultados obtidos sejam válidos, tais como normalidade e homoscedasticidade, por exemplo. No caso em que ocorre a violação de tais suposições, pode-se recorrer a transformações com o objetivo de garantir, ao menos de forma aproximada, que estas suposições fossem razoáveis. No entanto, Nelder e Wedderburn (1972) propuseram uma classe mais abrangente, denominada de modelos lineares generalizados (MLGs).

Os modelos citados podem não ser adequados, pois em geral, variáveis com suporte em (0,1) não são homoscedásticas e além disso, ajustando os modelos tradicionais, não se garante as predições no intervalo (0,1). Desta forma, para incorporar essas características diversos autores propuseram modelos de regressão baseados em variáveis

com suporte limitado como por exemplo, Ferrari e Cribari-Neto (2004) que propuseram um modelo de regressão, onde a variável resposta é caracterizada por uma distribuição beta sendo útil quando se tem interesse em variável contínua e restrita ao intervalo (0,1).

Dessa forma, diversos autores propuseram outras classes de distribuições para servir de base para a modelagem no intervalo (0,1) como por exemplo, Song e Tan (2000) com o modelo de regressão simplex, Ferrari e Cribari-Neto (2004) com o modelo de regressão beta, Bayes et al. (2012) que propuseram um novo modelo de regressão no intervalo unitário considerando a distribuição beta retangular, e mais recentemente Mousa et al. (2016) que desenvolveram o modelo de regressão gama unitária expresso em termos de sua média e de um parâmetro de dispersão.

O presente trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada a distribuição gama unitária. O Capítulo 3 é apresentado o modelo de regressão gama unitária, medidas de diagnóstico e influência local referentes ao modelo. E por fim, o Capítulo 4 corresponde a uma aplicação prática do modelo de regressão gama unitária.

2 DISTRIBUIÇÃO GAMA UNITÁRIA

Diz-se que Y possui distribuição gama unitária (Grassia, 1977) se $Y = \ln[X^{-1}]$ ou de forma equivalente $Y = \ln[(1-X)^{-1}]$, em que X segue distribuição gama. A distribuição gama unitária tem aplicações em problemas de inoculação Grassia (1977) em que o mesmo considera o mesmo problema tratado em Moran (1954) que considerou a distribuição beta como proposta de modelagem, nesse problema tem-se o interesse em estimar a densidade de bactérias ou vírus em ensaios de diluição. A depender de suas características, propriedade e eventual facilidade de tratamento algébrico é com certeza uma ótima alternativa frente a beta Ferrari e Cribari-Neto (2004) e simplex Song e Tan (2000) para modelagem de variáveis resposta no intervalo unitário.

Seja Y seguindo distribuição gama unitária (Grassia, 1977) com parâmetros de forma $\alpha, \tau > 0$, denotada por $\mathcal{UG}(\alpha, \tau)$. A função densidade de probabilidade (fdp) Y é dada por

$$\mathcal{UG}(y; \alpha, \tau) = \frac{\alpha^\tau}{\Gamma(\tau)} y^{\alpha-1} \left[\log\left(\frac{1}{y}\right)\right]^{\tau-1} \mathbb{1}_{(0,1)}(y), \tag{2.1}$$

ou equivalentemente por $y = 1 - v$, $0 \leq v \leq 1$, assim por (2.1)

$$\mathcal{UG}(v; \alpha, \tau) = \frac{\alpha^\tau}{\Gamma(\tau)} (1-v)^{\alpha-1} \left[\log\left(\frac{1}{1-v}\right)\right]^{\tau-1} \mathbb{1}_{(0,1)}(v), \tag{2.2}$$

com respectiva função distribuição acumulada (fda) expressa por

$$F(y; \alpha, \tau) = F_y(-\log(Y)|\alpha, \tau) = \frac{\gamma\{\tau, \alpha[-\log(Y)]\}}{\Gamma(\tau)}, \tag{2.3}$$

em que $F_y(.)$ denota a (fda) da distribuição gama de parâmetros $\alpha, \tau > 0$ e $\gamma(.,.)$ é a função gama incompleta inferior, denotada por $\gamma(a, x) = \int_0^x t^{a-1} e^{-t}\, dt$.

Grassia (1977) apresenta algumas propriedades referente a forma da distribuição gama unitária (2.1) para valores de $y \in (0, 1)$, então segue que:

Para $\tau < 1$.

$\alpha < 1 : \mathcal{UG}(y; \alpha, \tau)$ tem forma de U, além disso, possui assíntota em $y = 0$ e $y = 1$ com mínimo em $y = \exp[-(\tau - 1)/(\alpha - 1)]$

$\alpha = 1 : \mathcal{UG}(y; \alpha, \tau)$ tem forma de J invertido tangente a $y = 0$ na origem e então assume uma forma de J tornando-se assíntota em $y = 1$.

$\alpha > 1 : \mathcal{UG}(y; \alpha, \tau)$ tem uma forma de J começando em 0 e aumentando abruptamente para se tornar assíntota em $y = 1$.

Para $\tau = 1$.

$\alpha < 1 : \mathcal{UG}(y; \alpha, \tau)$ tem uma forma de J invertido e possui assíntota em $y = 0$.

$\alpha = 1$: tem-se a distribuição uniforme padrão.

$\alpha > 1 : \mathcal{UG}(y; \alpha, \tau)$ parte da origem e então intercepta $y = 1$, quando assume um valor finito.

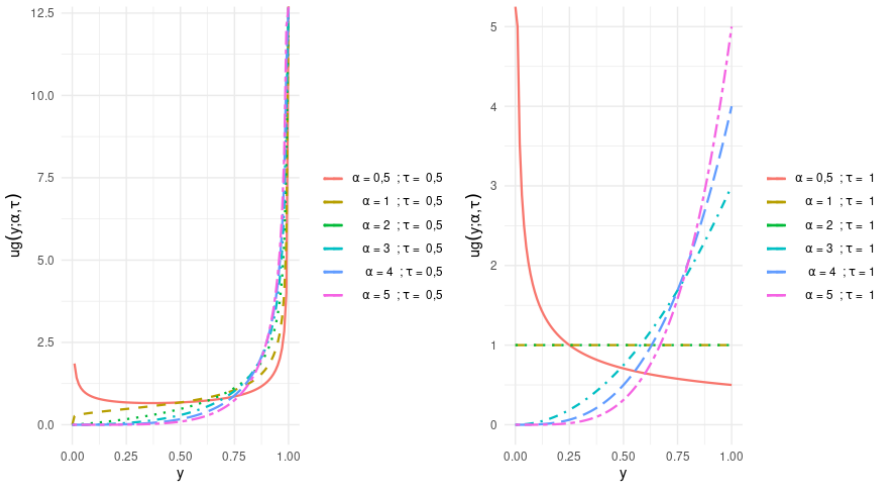
Para $\tau > 1$.

$\alpha \leq 1 : \mathcal{UG}(y; \alpha, \tau)$ tem uma forma de J invertido distorcido e possui assíntota em $y = 0$.

$\alpha > 1 : \mathcal{UG}(y; \alpha, \tau)$ começa em 0, aumenta até atingir um máximo em $\exp\left[-(\tau - 1)/(\alpha - 1)\right]$ e depois diminui.

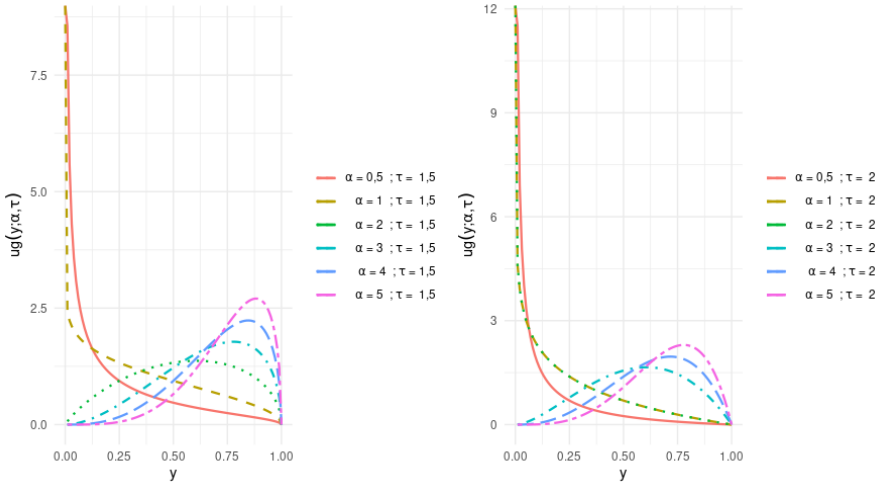
Nas Figuras 1 e 2 apresentamos os gráficos da fdp da distribuição gama unitária para vários valores dos parâmetros α e τ, sendo possível observar as diversas formas que a densidade pode assumir.

Figura 1 – Função densidade de probabilidade da distribuição gama unitária (α, τ), para alguns valores de α e τ.



Fonte: Elaborada pelo autor.

Figura 2 – Função densidade de probabilidade da distribuição gama unitária (α, τ), para alguns valores de α e τ.



Fonte: Elaborada pelo autor.

Pode-se ainda definir o $r-$ésimo momento em relação a origem de (2.1) na qual a mesmo apresenta forma analítica fechada expresso por

$$\mu_r^{'} = \frac{\alpha^{\tau}}{\Gamma(\tau)} \int_0^1 y^{n+\alpha-1} \left[\log\left(\frac{1}{y}\right)\right]^{\tau-1} dy = \left(\frac{\alpha}{\alpha+r}\right)^{\tau}, \tag{2.4}$$

portanto, a partir de (2.4) obtemos a média e variância de (2.1) dadas, respectivamente, por

$$\mathbb{E}[Y] = \mu_1^{'} = \left[\frac{\alpha}{\alpha+1}\right]^{\tau}, \ \mathrm{Var}[Y] = \mu_2^{'} - [\mu_1^{'}]^2 = \left[\frac{\alpha}{\alpha+2}\right]^{\tau} - \left[\frac{\alpha}{\alpha+1}\right]^{2\tau}.$$

De forma similar ao que foi feito com a Beta em Ferrari e Cribari-Neto (2004), Lima (2017) reparametrizou a gama unitária em termos de sua média μ e de uma parâmetro de dispersão, de forma que a densidade pode ser reescrita da seguinte forma:

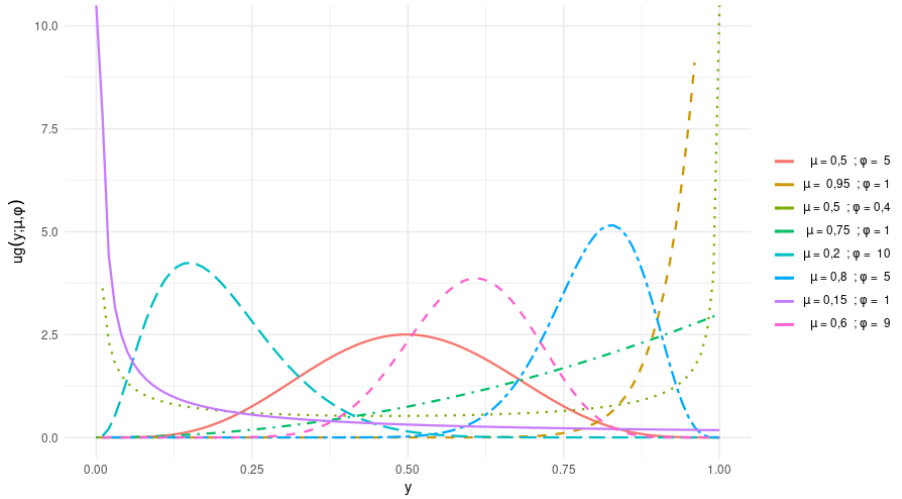
$$\mathcal{UG}(y;\mu,\phi) = \frac{\left[\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}\right]^{\phi}}{\Gamma(\phi)} y^{\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}} \left[\log\left(\frac{1}{y}\right)\right]^{\phi-1} \mathbb{1}(y)_{(0,1)}. \tag{2.5}$$

com média e variância dadas, respectivamente, por

$$\mathbb{E}[Y] = \left[\frac{\alpha}{\alpha+1}\right]^{\phi} = \mu, \ \mathrm{Var}[Y] = \mu\left[\frac{1}{(2-\mu^{1/\phi})^{\phi}} - \mu\right].$$

Na Figura 3 apresentamos os gráficos da fdp da distribuição gama unitária para vários valores dos parâmetros μ e ϕ, no qual pode-se notar forma assimétrica, U, para média $\mu = 0,5$ e parâmetro de dispersão $\phi < 1$, J e J invertido para valores da média próximos de 0 e 1 e com parâmetro de dispersão $\phi = 1$, assim como a forma simétrica no caso em que $\mu = 0,5$ e $\phi = 5$. Portanto, fica evidente o com flexível a distribuição gama unitária pode ser tando na parametrização considerada em (2.1), quanto em (2.5).

Figura 3 – Comparação entre os valores das variâncias considerando vários valores de ϕ e com μ fixo.



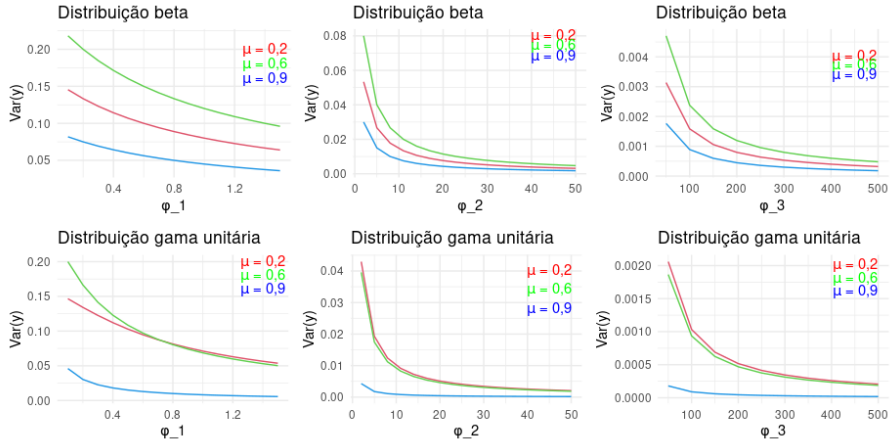
Fonte: Elaborada pelo autor.

Além de ser um modelo extremamente flexível, segundo Rocha (2020) a distribuição gama unitária (2.5) apresenta uma vantagem frente a distribuição beta Ferrari e Cribari-Neto (2004), esse fato ocorre quando a média da variável reposta μ está próxima de 1, nessa condição a distribuição gama unitária apresenta uma variância menor que a distribuição beta. Para ilustrar, considerou-se de Rocha (2020) os seguintes valores de $\mu =$ (0,2; 0,6; 0,9), $\phi_1 =$ (0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1,0; 1,1; 1,2; 1,3; 1,4; 1,5), $\phi_2 =$ (2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50) e $\phi_3 =$ (50, 100, 150, 200, 250, 300, 350, 400, 450, 500) de modo a comparar as seguintes variâncias

$$\text{Var}_{ug}(Y) = \mu\left[\frac{1}{(2-\mu^{1/\phi})\phi} - \mu\right] \text{ e } \text{Var}_{beta}(Y) = \frac{\mu(1-\mu)}{(1+\phi)},$$

que pode ser visualizada na Figura 4 a seguir.

Figura 4 – Comparação entre os valores das variâncias considerando vários valores de ϕ e com μ fixo.



Fonte: Adaptado Rocha (2020).

2.0.1 Estimação

Para estimação dos parâmetros α e τ da distribuição gama unitária (2.1) Dey et al. (2019) apresentam diferentes métodos do ponto de vista frequentista, sendo os estimadores de máxima verossimilhança, estimadores de momentos, estimadores de mínimos quadrados ordinários, estimadores de máximo produto de espaçamentos, método de Cramer-von-Mises, métodos de Anderson-Darling e quatro variantes do teste de Anderson-Darling. De forma similar podemos utilizar os diferentes métodos de estimação frequentista para a distribuição gama unitária reparametrizada (2.5), assim, nesse trabalho será apresentado os métodos mais usuais para estimação dos parâmetros de (2.5), em particular o método máxima verossimilhança, método dos momentos e o de mínimos quadrados ordinários.

2.0.1.1 Método de máxima verossimilhança

Sejam $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ um vetor de amostra aleatória de tamanho n da distribuição gama unitária (2.5). Portanto, a função de log-verossimilhança da distribuição gama unitária (2.5), é expressa por

$$\begin{aligned} \ell(\mu, \phi) &= \sum_{i=1}^{n} \log\left[\mathcal{UG}(y_i; \mu, \phi)\right] \\ &= \sum_{i=1}^{n} \left\{ \phi \log\left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}\right) - \log(\phi) - \left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}} - 1\right) \log(y_i) + (\phi - 1) \log\left[\log(1/y_i)\right] \right\}. \end{aligned} \tag{2.6}$$

As equações necessárias para determinar os estimadores de máxima verossimilhança $\hat{\mu}_{ml}$ e $\hat{\phi}_{ml}$ são dadas por

$$\frac{\partial \ell(\mu,\phi)}{\partial \mu} = \sum_{i=1}^{n} \left\{ \frac{\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}}{\mu^{(1/\phi)+1}} \left[1 + \frac{1}{\phi} \left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}} \right) \log(y_i) \right] \right\} = 0 \tag{2.7}$$

e

$$\frac{\partial \ell(\mu,\phi)}{\partial \phi} = \sum_{i=1}^{n} \left\{ \log(-\log(y_i)) - \left[\frac{1}{\phi} \left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}} \right) \log(\mu) \right] \left[1 + \frac{\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}} \log(\mu)}{\phi \mu^{1/\phi}} \right] - \log\left(\frac{\mu^{1/\phi}}{\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}} \right) - \psi(\phi) \right\} = 0, \tag{2.8}$$

em que $\psi(.)$ e a função digama. No entanto, (2.7) e (2.8) não possuem solução com forma fechada, assim devemos recorrer a métodos iterativos como Newton-Raphson, Escore de Fisher, BFGS e SANN, por exemplo.

2.0.1.2 Método dos momentos

Os estimadores de momentos da distribuição gama unitária (2.5) são determinados igualando aos dois primeiros momentos amostrais, assim, tem-se que

$$\mu_1^{'} = \mu = m_1 \tag{2.9}$$

e

$$\mu_2^{'} = \mu \left[\frac{1}{(2-\mu^{1/\phi})^{\phi}} - \mu \right] + \mu^2 = m_2, \tag{2.10}$$

em que $m_1 = \sum_{i=1}^{n} y_i$ e $m_2 = \sum_{i=1}^{n} y_i^2$. Entretanto, nesse método também se deparamos com uma complexidade algébrica para resolver as equações o que torna esse método menos preferível em relação ao método de máxima verossimilhança uma vez que o mesmo garante propriedades ótimas dos estimadores. Apesar disso, podemos utilizar as estimativas dos métodos do momentos como chute inicial para a estimação por máxima verossimilhança via processo iterativo.

2.0.1.3 Método de mínimos quadrados ordinários

Esse método é bem menos usual do que o método de máxima verossimilhança e de momento além de possuir uma complexidade algébrica na obtenção dos seus estimadores

similar aos anteriores ainda envolve a derivada da função acumulada da distribuição (2.5), o que torna esse método um dos últimos a ser considerado na prática. Essencialmente temos que $F(Y_{(i)}) \sim Beta(i, n-i+1)$ a (fda) das estatísticas de ordem da amostra $y_1, y_2, \ldots, y_n$, dessa forma, tem-se que

$$\mathbb{E}[F(Y_{(i)})] = \frac{i}{n+1} \tag{2.11}$$

e

$$\text{Var}[F(Y_{(i)})] = \frac{i(n-i+1)}{(n+1)^2(n+2)}. \tag{2.12}$$

Assim, os estimadores de mínimos quadrados ordinários $\hat{\mu}_{ols}$ e $\hat{\phi}_{ols}$ para os parâmetros μ e ϕ são obtidos minimizando a função

$$S(\mu, \phi|\mathbf{y}) = \sum_{i=1}^{n} \left[F(y_i|\mu, \phi) - \frac{i}{n+1} \right]^2 \tag{2.13}$$

com respeito a μ e ϕ, de forma alternativa pode-se obter tais estimadores resolvendo as seguintes funções não-lineares

$$\sum_{i=1}^{n} \left[F(y_i|\mu, \phi) - \frac{i}{n+1} \right]^2 \frac{\partial F(y_{i,n}|\mu, \phi)}{\partial \mu} = 0 \tag{2.14}$$

e

$$\sum_{i=1}^{n} \left[F(y_i|\mu, \phi) - \frac{i}{n+1} \right]^2 \frac{\partial F(y_{i,n}|\mu, \phi)}{\partial \phi} = 0. \tag{2.15}$$

em que $F(.)$ corresponde a função acumulada da distribuição gama unitária (2.5). Além disso, como as equações (2.14) e (2.15) são não lineares a obtenção das soluções algébricas ficam complexas, assim, é necessário obter as respectivas soluções por meio de métodos computacionais.

3 MODELO DE REGRESSÃO GAMA UNITÁRIA

A distribuição gama unitária devido a Grassia (1977) é a distribuição base da proposta do modelo de regressão gama unitária Mousa et al. (2016), no qual se considerou a parametrização em termos da média já mencionada (2.5), sendo assim, possível a modelagem da média da resposta. Diversas literaturas versam sobre o modelo de regressão gama unitária, dentre elas temos Guedes et al. (2021) que apresentam duas estatísticas de teste corrigidas em que as mesmas conduzem a inferências mais precisas em relação ao teste da razão de verossimilhanças padrão, especialmente em amostras de tamanho pequeno e moderado. Rocha (2020) que avalia a qualidade do ajuste do modelo de regressão gama unitária em relação a predição e em relação a variabilidade propondo as expressões das estatísticas PRESS e P^2. Rocha et al. (2021) tratam sobre métodos de diagnóstico para o modelo de regressão gama unitária. A seguir, vamos especificar o modelo de regressão gama unitária.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes seguindo uma distribuição gama unitária (2.5) com média μ_t e parâmetro de precisão constante ϕ.

$$g(\mu_t) = \sum_{i=1}^{p} x_{ti}\beta_i = \eta_{1t}, \tag{3.1}$$

sendo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor de parâmetros, $x_{t1}, \ldots, x_{tp}$ as observações conhecidas e fixas e $g(.)$ uma função estritamente monótona e ao menos duas vezes diferenciável, com domínio no $(0, 1)$ e imagem em $\mathbb{R}$. Podemos ainda estender para o caso em que o parâmetros de dispersão é não constante, dessa forma, tem-se que

$$h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \eta_{2t}, \tag{3.2}$$

em que $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ é o vetor de parâmetros, $z_{t1}, \ldots, z_{tq}$ as observações conhecidas e fixas e $h(.)$ uma função estritamente monótona e ao menos duas vezes diferenciável, com domínio no $(0, 1)$ e imagem em $\mathbb{R}$.

Sob a suposição de que y_t segue distribuição gama unitária e considerando (3.1), temos que $\mu_t = g^{-1}(\eta_{1t})$ e $\mathrm{Var}(y_t) = g^{-1}(\eta_{1t})\left[\frac{1}{\{2-[g^{-1}(\eta_{1t})]^{1/\phi}\}^{\phi}} - g^{-1}(\eta_{1t})\right]$.

Dada uma amostra $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{UG}(\mu_t, \phi)$ com $\alpha_t = \frac{\mu_t^{1/\phi}}{1-\mu_t^{1/\phi}}$ e $\mu_t = g_1^{-1}(\eta_t)$,

então temos que

$$\begin{aligned}\ell(\boldsymbol{\beta},\phi) &= \sum_{i=1}^{n}\ell_t(\mu_t,\phi) \\ &= \log\left[ug(y_t;\mu_t,\phi)\right] \\ &= \phi\log(\alpha_t)-\log\left[\Gamma(\phi)\right]+(\alpha_t-1)\log(y_t)+(\phi-1)\log\left[-\log(y_t)\right].\end{aligned} \tag{3.3}$$

A função escore para $\boldsymbol{\beta}$ é um vetor de dimensão p, dado por

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\phi)=\mathbf{X}_{(n\times p)}^{\top}\mathbf{T}\,\mathbf{s}, \tag{3.4}$$

em que $\mathbf{s}=(\mu_1^*+y_1^*,\ldots,\mu_n^*+y_n^*)^{\top}$, $\mathbf{T}=\text{diag}\{1/g_1^{'}(\mu_1),\ldots,1/g_n^{'}(\mu_n)\}$ e $\mu_t^*=\frac{\alpha_t}{\mu_t^{1/\phi+1}}$ e $y_t^*=\frac{\alpha_t^2\log(y_t)}{\phi\mu_t^{1/\phi+1}}$. Para o parâmetro ϕ a função escore é

$$U_{\phi}(\boldsymbol{\beta},\phi) = \sum_{t=1}^{n}u_t, \tag{3.5}$$

em que $u_t=\log(-\log(y_t))-\left[\frac{1}{\phi}\alpha_t\log(\mu_t)\right]\left[1+\frac{\alpha_t\log(y_t)}{\phi\mu_t^{1/\phi}}\right]-\log\left(\frac{\mu_t^{1/\phi}}{\alpha_t}\right)-\psi(\phi)$, sendo $\psi(.)$ a função digama.

A matriz de informação de Fisher é dada por

$$\mathbf{K}(\boldsymbol{\beta},\phi)=\begin{pmatrix}\mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{K}_{\boldsymbol{\beta}\phi}\\ \mathbf{K}_{\phi\boldsymbol{\beta}} & \mathbf{K}_{\phi\phi}\end{pmatrix}, \tag{3.6}$$

em que $\mathbf{K}(\boldsymbol{\beta},\phi)=\mathbf{X}^{\top}\mathbf{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}$, $\mathbf{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}=\text{diag}\{(a_t^2/\phi)\ [1/g_1^{'}(\mu_t)]^2\}$, $\mathbf{K}_{\boldsymbol{\beta}_\phi}=\mathbf{K}_{\phi\boldsymbol{\beta}}=\mathbf{X}^{\top}\mathbf{W}_{\boldsymbol{\beta}_\phi}\mathbf{1}_{(n\times 1)}$, $\mathbf{W}_{\boldsymbol{\beta}_\phi}=\text{diag}\{(b_t/\phi)(c_t/\phi+1)\ [1/g_1^{'}(\mu_t)]\}$, $\mathbf{K}_{\phi\phi}=\mathbf{1}_{(n\times 1)}^{\top}\mathbf{W}_{\phi\phi}\mathbf{1}_{(n\times 1)}$, e $\mathbf{W}_{\phi\phi}=\text{diag}[(2c_t/\phi^2)+(c_t^2/\phi^3)+\psi^{'}(\phi)]$. Além disso,

$$\mathbf{K}(\boldsymbol{\beta},\phi)=\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{W}}\tilde{\mathbf{X}},$$

em que

$$\tilde{\mathbf{X}}_{2n\times(p+1)}=\begin{pmatrix}\mathbf{X}_{n\times p} & \mathbf{0}_{n\times 1}\\ \mathbf{0}_{n\times p} & \mathbf{1}_{(n\times 1)}\end{pmatrix}\ \text{ e }\ \tilde{\mathbf{W}}_{2n\times 2n}=\begin{pmatrix}\mathbf{W}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{W}_{\boldsymbol{\beta}_\phi}\\ \mathbf{W}_{\phi\boldsymbol{\beta}} & \mathbf{W}_{\phi\phi}\end{pmatrix}.$$

Assintoticamente e sob as condições de regularidade usuais de Flechet-Cramer-Rao (Sen et al, 1994), temos que

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\phi} \end{pmatrix} \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \phi \end{pmatrix}, \mathbf{K}^{-1}\right). \tag{3.7}$$

3.1 Critérios de informação

Ao que tange aos critérios de seleção de modelos temos diversas alternativas quais podemos citar o C_p de Mallows, critério de informação generalizados (GIC), critério de informação consistente (CAIC), critério de informação Akike (AIC) e critério de informação Bayesiano (BIC), por exemplo. Tais procedimentos consideram a penalização da função de verossimilhança por meio da subtração do número de parâmetros associado ao modelo e eventualmente do número de observações, ou seja, o modelo que possui mais parâmetros é consequentemente mais penalizado. Portanto, nesse trabalho consideremos os critérios (AIC) e (BIC) devido a Akaike (1974) e Schwarz (1978), respectivamente. Ambos critérios são expressos por

$$\text{AIC} = -2\log(\hat{\ell}) + 2p \tag{3.8}$$

$$\text{BIC} = -2\log(\hat{\ell}) + p\log(n) \tag{3.9}$$

em que $\hat{\ell}$ é valor máximo da log-verossimilhança, p o número de parâmetros e n o número de observações. Dessa forma, devemos escolher o modelo no qual resulte nos menores valores de AIC e BIC.

3.2 Medidas de diagnóstico

3.2.1 Resíduos

Os métodos de diagnóstico para classe de modelos gama unitária podem ser encontrados em Rocha (2020) e Rocha et al. (2021) onde são propostos os resíduos ponderado e ponderado padronizado na qual se utiliza o processo iterativo Scoring de Fisher para o vetor $\boldsymbol{\beta}$. Assim, para o modelo de regressão (3.1) temos que $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[K^{(m)}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right]^{-1} U^{(m)}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi)$ em que m indica a $m-$ésima iteração até a convergência que

por sua vez acontece quando a distância entre $\boldsymbol{\beta}^{(m+1)}$ e $\boldsymbol{\beta}^{(m)}$ é menor que um certo valor ϵ pré-estabelecido. Para o modelo de regressão gama unitária a expressão fica

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}^\top W^{(m)}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}\right]^{-1}\mathbf{X}^\top T^{(m)} s^{(m)} \tag{3.10}$$

em que $s = (\mu_1^* + y_1^*, \ldots, \mu_n^* + y_n^*)^\top$ e $T = \text{diag}\left[g_1'^{-1}(\mu_1), \ldots, g_n'^{-1}(\mu_n)\right]$. Rocha (2020) considera a partir de (3.8) um processo iterativo de mínimos quadrados reponderados, assim temos que

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}^\top W^{(m)}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}\right]^{-1}\mathbf{X}^\top T^{(m)} W_{\boldsymbol{\beta}\boldsymbol{\beta}}\, a^{(m)} \tag{3.11}$$

em que $a^{(m)} = \eta^{(m)} + W^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}\, T^{(m)} s^{(m)}$ com $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top = \mathbf{X}\boldsymbol{\beta}$. Posteriormente a convergência, obtém-se o seguinte estimador $\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}^\top \widehat{W}^{(m)}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}\right]^{-1}\mathbf{X}^\top \widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}\, a^{(m)}$, com $a = \widehat{\eta} + \widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}\widehat{T}\,\widehat{s}$ que por sua vez pode ser visto com um estimador de mínimos quadrados considerando regressão linear de $W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}\, a$ em $W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}$. O resíduo de mínimos quadrados da regressão é definido por $r_t^\beta = W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}(a - \hat{\boldsymbol{\eta}}) = W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}\widehat{T}\,\widehat{s}$ no qual Rocha (2020) denomina como resíduo ponderado que pode ser expresso por

$$r_t^\beta = \frac{1}{\widehat{w}^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}}\frac{1}{g_1'(\widehat{\mu})}\widehat{s} = \frac{\hat{s}}{\sqrt{(\widehat{\mu}_t^*)^2/\widehat{\phi}}}. \tag{3.12}$$

Os resíduos padronizados são determinado considerando que $\widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}} \approx W_{\boldsymbol{\beta}\boldsymbol{\beta}}$ que por meio de propriedades de covariância obtém-se que $\widehat{\text{Cov}}(r_t^\beta) \approx (I - \widehat{\mathbf{H}})$, sendo $\mathbf{H} = W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}(\mathbf{X}^\top W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X})^{-1}\mathbf{X}^\top W^{1/2}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, dessa forma o resíduo padronizado fica expresso por

$$r_{pt}^\beta = \frac{r_t^\beta}{\sqrt{\widehat{\text{Cov}}(r_t^\beta)}} = \frac{\hat{s}}{\sqrt{(\hat{\mu}_t^*)^2/\hat{\phi}(1 - \hat{h}_{tt})}}, \tag{3.13}$$

em que $\hat{h}_{tt}$ o t-ésimo elemento da diagonal de $\mathbf{H}$.

3.3 Análise de influência

3.3.1 Alavancagem

Um ponto alavanca é aquele que possui um perfil diferente dos demais pontos em relação as variáveis explicativas. O poder de alavanca da $t-$ésima observação é dado por h_{tt}, isto é, pelo $t-$ésimo elemento da diagonal principal da matriz de projeção $\mathbf{H}$ que é simétrica e idempotente implicando que $0 \leq h_{tt} \leq 1$. Então, se $h_{tt} = 1$, tem-se que $\hat{y}_t = y_t$, de forma que a $t-$ésima observação tem influência total no seu valor predito. Ademais, uma vez que $\text{tr}(\mathbf{H}) = \sum_{t=1}^{n} h_{tt} = p$, temos que o valor médio do poder de alavanca é p/n. Com isso, podemos identificar uma observação com alto poder de alavanca se $h_{tt} \geq 2p/n$ ou $h_{tt} \geq 3p/n$, uma vez que quando todos os elementos da diagonal principal de $\mathbf{H}$ são próximos de p/n, nenhuma observação influencia o seu valor predito de forma proporcional. Entretanto, vale salientar que no caso do modelo de regressão gama unitária os valores de h_{tt} dependem da matriz de pesos $\widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, assim as observações com valores grandes de h_{tt} nem sempre são de fato pontos de alavanca. Uma forma prática de identificar tais pontos é construímos o gráfico de h_{tt} *versus* os índice das observações destacando os $h_{tt} \geq 2p/n$ ou $h_{tt} \geq 3p/n$ como mencionado anteriormente.

3.3.2 Ponto aberrante

Um ponto aberrante é aquele que apresenta uma característica diferente das demais observações em relação a variável resposta e possui um valor baixo na matriz de projeção $\mathbf{H}$. Podemos identificar tais pontos por meio do gráfico dos resíduos padronizados (3.13) *versus* o índice das observações t.

3.3.3 Distância de Cook

Uma estratégia para identificar pontos influentes, isto é, pontos nos quais exercem um peso desproporcional nas estimativas dos parâmetros do modelo e possui característica diferente dos demais em relação aos valores da variável resposta além de apresentar valor alto na matriz de projeção $\mathbf{H}$ é considerar a distância de Cook devido a Cook (1977) em que o mesmo sugere que a influência de uma observação, ou um conjunto delas, seja avaliada por meio dos efeitos causados da sua remoção no conjunto de dados.

A distância de Cook é definida por

$$\mathrm{DC}_t = \frac{1}{p}\left\{\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(t)}\right]^\top \left(\mathbf{X}^\top \widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}\right)\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(t)}\right]\right\} \tag{3.14}$$

em que p corresponde ao número de parâmetros do modelo, $\widehat{\boldsymbol{\beta}}_{(t)}$ a estimativa do parâmetro sem a $t-$ésima observação e $\widehat{W}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ a matriz de pesos. Com isso é possível medir o afastamento do vetor com todas observações $\widehat{\boldsymbol{\beta}}$ e sem a observação y_t ($\widehat{\boldsymbol{\beta}}_{(t)}$).

Uma aproximação da distância de Cook no qual não necessitamos realizar $n+1$ ajustes é expressa por

$$\mathrm{DC}_t = (r_{pt}^{\beta})^2 \frac{h_{tt}}{p(1-h_{tt})}, \tag{3.15}$$

em que r_{pt}^{β} é o resíduo padronizado (3.13). Dessa forma, pontos supostamente influentes são detectados por meio do gráfico de DC_t *versus* t e em geral são aquelas observações que apresentam os maiores valores de DC_t em relação as demais.

3.3.4 Gráfico de probabilidade meio normal

O gráfico de probabilidade é um meio gráfico no qual dispormos no gráfico os quantis de duas variáveis de interesse, em que espera-se visualizar um padrão linear dos pontos. No caso do gráfico de probabilidade meio normal, utilizamos os valores esperados das estatísticas de ordem da normal padrão no eixo das abcissas e nas ordenadas os resíduos provenientes de um modelo de regressão. Além disso, podemos utilizar um envelope simulado devido Atkinson (1981) para termos um referencial quanto as flutuações dos pontos.

3.3.5 Gráfico dos resíduos quantílicos aleatorizados

Similar ao gráfico de probabilidade meio-Normal com evelope simulado usual, este método proposto por Dunn e Smith (1996) apresenta distribuição Normal, independente da distribuição da variável resposta, no qual os mesmos se baseiam no teorema da função distribuição acumulada.

Seja Y uma variável aleatória absolutamente contínua, então pelo teorema da função distribuição acumulada temos que $U_i = F(y_i; \mu_i, \phi)$ tem distribuição uniforme no

intervalo $(0, 1)$. Assim, temos que o resíduo quantílico fica expresso por

$$r_i^q = \Phi^{-1}\left[F(y_i; \hat{\mu}_i, \hat{\phi})\right], \tag{3.16}$$

em que $\Phi(.)$ a função distribuição acumulada da Normal padrão. A distribuição de r_i^q converge para a uma Normal padrão se $\boldsymbol{\beta}$ e ϕ forem consistentemente estimados. Portanto, podemos construir um gráfico dos resíduos quantílicos *versus* os quantis da Normal padrão com bandas de confiança para avaliar o ajuste do modelo. Se o modelo estiver bem ajustado espera-se que os pontos apresentem um padrão de reta além de que 95% deles estejam dentro das bandas de confiança.

3.3.6 Influência local

O conceito de influencia local foi proposto por Cook (1986) com o interesse em verificar as mudanças nos resultados da análise quando pequenas perturbações são incorporadas ao modelo e/ou aos dados. Se tais pertubações causarem algum efeito desproporcional podemos então ter indícios de que o modelo postulado esteja mal ajustado. A ideia inicial é avaliar o deslocamento pela verossimilhança, no caso mais geral é expressa por

$$\mathrm{LD}_\delta = 2\left[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\delta}})\right] \tag{3.17}$$

em que $\ell(.)$ é a log-verossimilhança do modelo postulado, $\boldsymbol{\theta}$ um vetor $s \times 1$ de parâmetros $\ell(.|\boldsymbol{\delta})$ é a log-verossimilhança do modelo "perturbado", $\boldsymbol{\delta}$ denota um vetor $n \times 1$ de perturbações, restrito a algum subconjunto aberto $\mathcal{D} \in \mathbb{R}^n$, $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\delta}}$ os respectivos EMV de $\ell(.)$ e $\ell(.|\boldsymbol{\delta})$.

A ideia de Cook (1986) foi avaliar o comportamento local de LD_δ em uma vizinha do vetor $\boldsymbol{\delta}_0$ de não perturbação do modelo postulado, em que tal procedimento representa a sensibilidade de $\ell(\widehat{\boldsymbol{\theta}})$ com respeito a uma pequena perturbação induzida em $\ell(\boldsymbol{\theta})$. Dessa forma, Cook (1986) propôs a utilização de curvaturas normais, em que se avalia como a superfície geométrica $\boldsymbol{\alpha}(\boldsymbol{\delta}) = (\boldsymbol{\delta}^\top, \mathrm{LD}_{\boldsymbol{\delta}})^\top$ desvia-se de seu plano tangente em $\boldsymbol{\delta}_0$ à medida que $\boldsymbol{\delta}$ se afasta levemente de $\boldsymbol{\delta}_0$.

Nesse caso a curvatura normal apresentada por Cook (1986) fica expressa por

$$\mathrm{C}_I = 2|\mathbf{I}^\top \boldsymbol{\Delta}^\top \ddot{\boldsymbol{\ell}}^{-1} \boldsymbol{\Delta}\ \mathbf{I}|, \tag{3.18}$$

em que $\ddot{\boldsymbol{\ell}} = \left[\partial^2 \ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top \partial\boldsymbol{\theta}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, $\boldsymbol{\Delta} = \left[\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{\delta})/\partial\boldsymbol{\theta}^\top \partial\boldsymbol{\delta}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}};\boldsymbol{\delta}=\boldsymbol{\delta}_0}$ e $\mathbf{I}$ um vetor de norma unitária.

Podemos ainda avaliar a influência para apenas uma parte do vetor de parâmetros. Assim, considere então o particionamento do vetor de parâmetros como $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$. Seja

$$\ddot{\boldsymbol{\ell}} = \begin{pmatrix} \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_1\boldsymbol{\theta}_1} & \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_1\boldsymbol{\theta}_2} \\ \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_1} & \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2} \end{pmatrix}, \tag{3.19}$$

em que $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_1\boldsymbol{\theta}_1} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^\top$, $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_1\boldsymbol{\theta}_2} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2^\top$, $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_1} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1^\top$ e $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^\top$.

Portanto, Cook (1986) comenta que se o interesse recai em calcular a influência somente para $\boldsymbol{\theta}_1$, o deslocamento da verossimilhança fica expresso por

$$\mathrm{LD}_{\boldsymbol{\delta};\boldsymbol{\theta}_1} = 2\left[\ell(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - \ell(\hat{\boldsymbol{\theta}}_{1\boldsymbol{\delta}}, g(\hat{\boldsymbol{\theta}}_2))\right], \tag{3.20}$$

neste caso $\ell(\hat{\boldsymbol{\theta}}_{1\boldsymbol{\delta}}, g(\hat{\boldsymbol{\theta}}_2))$ representa a log-verossimilhança perfilada de $\boldsymbol{\theta}_1$. Portanto a curvatura normal na direção do vetor unitário $\mathbf{I}$ é expressa por

$$\mathrm{C}_{\mathbf{I};\boldsymbol{\theta}_1} = |\mathbf{I}^\top \boldsymbol{\Delta}^\top (\ddot{\boldsymbol{\ell}}^{-1} - \ddot{\boldsymbol{\ell}}_{22}) \boldsymbol{\Delta}\ \mathbf{I}| \tag{3.21}$$

em que

$$\ddot{\boldsymbol{\ell}}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddot{\boldsymbol{\ell}}^{-1}_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2} \end{pmatrix}. \tag{3.22}$$

Para o modelo de regressão gama unitária com parâmetro de dispersão fixo Rocha (2020) propõe o seguinte esquema de influência local

$$\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}, \phi) = \begin{pmatrix} \ddot{\boldsymbol{\ell}}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \ddot{\boldsymbol{\ell}}_{\boldsymbol{\beta}\phi} \\ \ddot{\boldsymbol{\ell}}_{\phi\boldsymbol{\beta}} & \ddot{\boldsymbol{\ell}}_{\phi\phi} \end{pmatrix}, \tag{3.23}$$

em que $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \mathbf{X}^\top \text{diag}(q_1, \ldots, q_n)\mathbf{X}$, $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\beta}\phi} = (\ddot{\boldsymbol{\ell}}_{\phi\boldsymbol{\beta}})^\top = \mathbf{X}^\top \text{diag}\left[g'^{-1}(\mu_1), \ldots, g'^{-1}(\mu_n)\right]$ $(-c_1, \ldots, -c_n)^\top$ e $\ddot{\boldsymbol{\ell}}_{\phi\phi} = \text{tr}\left[\text{diag}(-g_1, \ldots, -g_n)\right]$.

Considerando (3.23) e por meio da inversa de uma matriz particionada Rao (1973), temos que

$$\ddot{\boldsymbol{\ell}}^{-1}(\boldsymbol{\beta}, \phi) = \begin{pmatrix} \ddot{\boldsymbol{\ell}}^{\boldsymbol{\beta}\boldsymbol{\beta}} & \ddot{\boldsymbol{\ell}}^{\boldsymbol{\beta}\boldsymbol{\phi}} \\ \ddot{\boldsymbol{\ell}}^{\boldsymbol{\phi}\boldsymbol{\beta}} & \ddot{\boldsymbol{\ell}}^{\boldsymbol{\phi}\boldsymbol{\phi}} \end{pmatrix}, \tag{3.24}$$

com $(\ddot{\boldsymbol{\ell}}^{\,\phi\phi})^{-1} = \text{tr}(\mathbf{G}) - \mathbf{f}^\top\mathbf{T}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{T}\mathbf{f}$, $\ddot{\boldsymbol{\ell}}^{\,\boldsymbol{\beta}\phi} = (\ddot{\boldsymbol{\ell}}^{\,\phi\boldsymbol{\beta}})^\top = (\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{T}\mathbf{f}$, e $\ddot{\boldsymbol{\ell}}^{\,\boldsymbol{\beta}\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1}\left[\mathbf{1}_{k\times k} + \mathbf{X}^\top\mathbf{T}\,\mathbf{f}\,\mathbf{f}^\top\mathbf{T}^\top(\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1}\right]\ddot{\boldsymbol{\ell}}^{\,\phi\phi}$. Sendo $\mathbf{f} = (-c_1, \ldots, -c_n)^\top$, $\mathbf{T} = \text{diag}\left[g'^{-1}(\mu_1), \ldots, g'^{-1}(\mu_n)\right]$ e $\mathbf{G} = \text{diag}(-g_1, \ldots, -g_n)$.

3.3.6.1 Esquemas de perturbação

O esquema de perturbação considerando o modelo de regressão gama unitária com precisão fixa apresentado Rocha (2020) é dado por

$$\ell(\boldsymbol{\beta}, \phi)_{\boldsymbol{\delta}} = \sum_{t=1}^{n} \boldsymbol{\delta}_t \ell_t(\mu_t, \phi), \tag{3.25}$$

com $\delta \in [0, 1]$. Considerando $\boldsymbol{\delta}_0 = (1, 1, \ldots, 1)^\top$, $\boldsymbol{\Delta} = \partial\ell_t(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ e a influência local anterior temos que

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{\boldsymbol{\beta}} \\ \boldsymbol{\Delta}_{\phi} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top\hat{\mathbf{T}}\hat{\boldsymbol{\mathcal{E}}} \\ \hat{u} \end{pmatrix}, \tag{3.26}$$

em que $\boldsymbol{\mathcal{E}} = \text{diag}\{(\mu_t^* + y_t^*), \ldots, (\mu_n^* + y_n^*)\}$ e $\hat{u} = \log(-\log(y_t)) - \left[\frac{1}{\phi}\alpha_t\log(\mu_t)\right]\left[1 + \frac{\alpha_t\log(y_t)}{\phi\mu_t^{1/\phi}}\right] - \log\left(\frac{\mu_t^{1/\phi}}{\alpha_t}\right) - \psi(\phi)$, sendo $\psi(.)$ a função digama.

3.3.6.2 Perturbação da variável resposta

Para esse caso Rocha (2020) considerou o esquema aditivo de perturbação da resposta em que vetor de resposta $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é alterado através da adição de um vetor de pequenas perturbações. Para compensar o caso em que cada y_t apresentam

variâncias diferentes utiliza-se um fator de correção, usualmente a estimativa de y_t, de forma que

$$y_t(\boldsymbol{\delta}) = y_t + \boldsymbol{\delta}_t s(y_t), \tag{3.27}$$

em que $s(y_t) = s(y_t) = \sqrt{\hat{\mu}_t(1/\left[\left(2-\hat{\mu}_t^{1/\hat{\phi}})\hat{\phi}\right]\right)}$. Então considerando $\boldsymbol{\delta}_0 = (0, 0, \ldots, 0)^\top$, temos que

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{\boldsymbol{\beta}} \\ \boldsymbol{\Delta}_{\phi} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top \hat{\mathbf{T}}\, \mathbf{M}\, \mathbf{S}_y \\ \hat{\mathbf{b}}^\top \mathbf{S}_y \end{pmatrix}, \tag{3.28}$$

sendo $\mathbf{M} = \text{diag}\{m_1, \ldots, m_n\}$ com

$$m_t = \frac{\alpha_t^2}{y_t \mu_t^{1/\phi+1} \phi}, \tag{3.29}$$

$\mathbf{S}_y = \text{diag}\{s(y_1), \ldots, s(y_n)\}$ e $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_n)$, ademais tem - se

$$b_t = \left[\frac{-\alpha_t^2 \log(\mu_t)}{y_t \phi^2 \mu_t^{1/\phi}} + \frac{1}{\log(y_t) y_t}\right]. \tag{3.30}$$

3.3.6.3 Perturbação individual de covariáveis

Rocha (2020) segue a sugestão de Cook e Thomas (1989) de modificar a $k-$ésima coluna da matriz $\mathbf{X}$, isto é, cada coluna x_k, com $k = 2, \ldots, p$ adicionando um vetor de pequenas perturbações ponderado por um fator de escala que corresponde ao desvio padrão de x_k, assim, tem-se que

$$x_{tk}(\boldsymbol{\delta}) = x_{tk} + \boldsymbol{\delta}_t s_{x_k}, \tag{3.31}$$

se temos $k \neq 2$ e $k \neq p$,

$$\eta_t(\boldsymbol{\delta}) = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k(x_{tk} + \delta_t s_{x_k}) + \ldots + \beta_p x_{tp}, \tag{3.32}$$

com $\mu_t(\boldsymbol{\delta})$ tal que $g(\mu_t(\boldsymbol{\delta})) = \eta_t(\boldsymbol{\delta})$ e considerando $\boldsymbol{\delta}_0 = (0, 0, \ldots, 0)^\top$, assim temos que

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{\boldsymbol{\beta}} \\ \boldsymbol{\Delta}_{\phi} \end{pmatrix} = \begin{pmatrix} -s_{x_k}\left(\hat{\boldsymbol{\beta}}_{x_k}\mathbf{X}^{\top}\hat{\mathbf{Q}} - \mathbf{P}\hat{\mathbf{T}}\boldsymbol{\mathcal{E}}\right) \\ -\hat{\boldsymbol{\beta}}s_{x_k}\hat{\mathbf{f}}^{\top}\hat{\mathbf{T}} \end{pmatrix}, \tag{3.33}$$

em que a matriz $\mathbf{P}_{(p\times n)}$ é formada de zeros exceto a $k-$ésima linha, que é composta por uns.

3.4 Recursos computacionais

Recursos computacionais implementados em pacotes ou bibliotecas para o modelo de regressão gama unitária ainda são escassos devido ao mesmo ser um modelo bastante atual. No entanto, é possível criar rotinas específicas em qualquer *software* tanto para estimação dos parâmetros do modelo quanto para a análise de diagnóstico. Nesse trabalho utilizamos o *software* gratuito `R` por ser o mais popular na comunidade Estatística, além disso utilizou-se também o pacote Template Model Builder `TMP` devido a Kristensen et al. (2015) para a diferenciação automática na qual o usuário define a probabilidade conjunta para os dados como uma função de modelo em `C++`, enquanto todas as outras operações são feitas em R. As implementações referentes ao ajuste e diagnóstico do modelo de regressão gama unitária que foram utilizadas nesse trabalho, além do template em `C++` encontram-se disponíveis no Apêndice A.

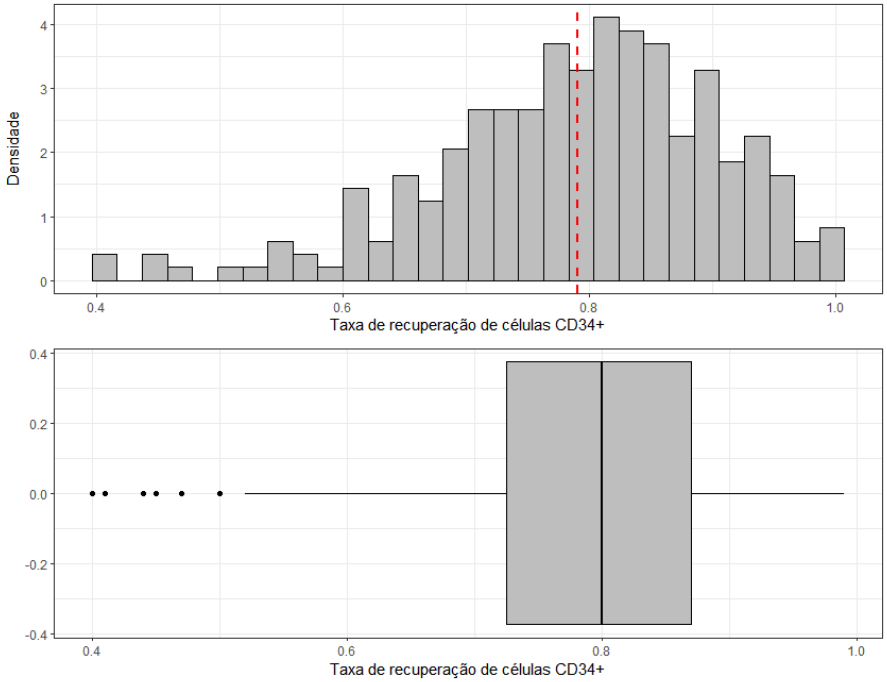
4 APLICAÇÃO

Para essa aplicação foi considerado o conjunto de dados de Zhang et al. (2016). Este conjunto de dados é referente a um estudo que trata sobre o transplante autólogos de células-tronco do sangue periférico. Foram considerados 242 pacientes avaliados entre os anos de 2003 a 2008 no Edmonton Hematopoietic Stem Cell Lab em Cross Cancer Institute - Alberta Health Services. O conjunto de dados dispõe de informações sobre as idades dos pacientes, sexo, bem como suas características clínicas. A variável dependente considerada em Zhang et al. (2016) e nessa aplicação é a taxa de recuperação de células CD34+ enquanto as variáveis independentes são: idade do paciente ajustada, isto é, considerou a idade < 40 como a idade inicial e subtraiu-se as outras idades por 40, e a segunda uma variável *dummy* que indica se o paciente recebe quimioterapia em um protocolo de um dia (0) ou em um protocolo de 3 dias (1). Portanto, tem-se o interesse em modelar a taxa de recuperação de células CD34+ em função da idade do paciente e do tipo de quimioterapia em relação ao protocolo.

4.1 Análise descritiva

A análise descritiva corresponde a etapa inicial de uma modelagem estatística, nessa etapa podemos verificar as características das variáveis interesse de forma a auxiliar na etapa de propor o modelo. Ademais, podemos também ter noção de pontos atípicos que possivelmente venham a influenciar no ajuste do modelo, assim podendo ser identificados de forma mais adequada em uma análise de diagnóstico posteriormente. A Figura 5 correspondem a análise descritiva para a taxa de recuperação de células CD34+.

Figura 5 – Histograma e Boxplot para variável taxa de recuperação de células CD34+.



Fonte: Elaborada pelo autor.

A média da taxa de recuperação de células CD34+ é de 0,79 destacada pela linha tracejada no histograma da Figura 3 com um desvio padrão 0,1141, além disso nota-se que os valores variam no intervalo $(0, 1)$ com valor mínimo observado de 0,40 e máximo observado de 0,99. Existe uma forte assimetria à esquerda com alguns pontos que se destacam dos demais, assim os modelos gama unitária e beta passam a ser fortes candidatas a modelar a taxa de recuperação de células CD34+ devido ambas distribuições possuírem suporte no intervalo $(0, 1)$ e serem capaz de assumir formas assimétricas à esquerda.

4.2 Ajuste com o modelo de regressão beta

Consideramos que as observações $y_1, \ldots, y_{239}$ são independentes e têm distribuição beta com média μ_t $(t = 1, \ldots, 239)$, e parâmetro de dispersão ϕ desconhecido.

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2}, \tag{4.1}$$

em que se considerou $g(.)$ como a função logito devido à interpretação direta dos parâmetros em termos da razão de chances (odds) e x_{t1} : a idade ajustada do $t-$ésimo indivíduo e x_{t1} :

variável *dummy* que indica se o $t-$ésimo paciente recebe quimioterapia em um protocolo de um dia (0) ou em um protocolo de 3 dias (1).

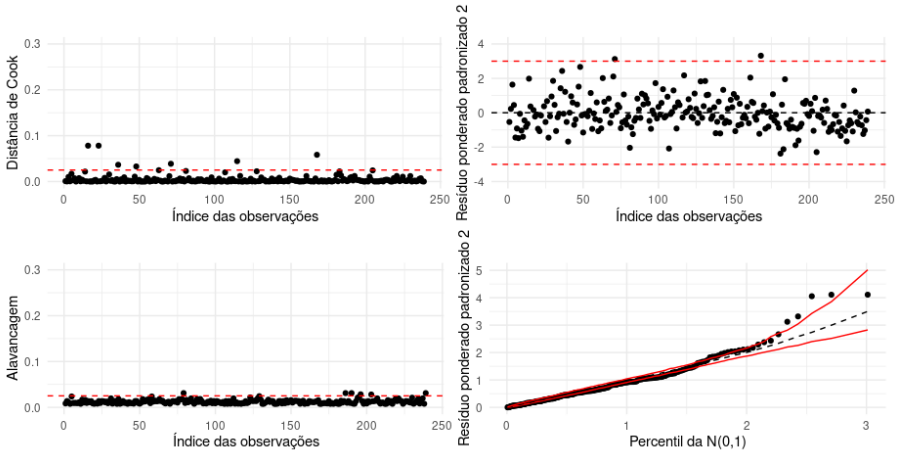
Tabela 1 – Estimativas dos parâmetros do modelo de regressão beta ajustado aos dados de transplante autólogos de células-tronco do sangue periférico.

Parâmetro	Estimativa	Erro Padrão	Valor p
β_0	1,0422	0,1115	$<$ 0,0001
β_1	0,0143	0,0053	0,0067
β_2	0,2143	0,1017	0,0351
ϕ	11,3210	1,0160	$<$ 0,0001

Fonte: Elaborada pelo autor.

É possível verificar que todas as estimativas do modelo proposto são significativas ao nível de significância de 5%. Além disso, foi realizada uma análise diagnóstico com objetivo de identificar possíveis pontos atípicos e verificar a adequabilidade do modelo.

Figura 6 – Gráficos de diagnóstico para o modelo de regressão beta ajustado aos dados de transplante autólogos de células-tronco do sangue periférico.



Fonte: Elaborada pelo autor.

Pela Figura 6 (gráfico inferior esquerdo) não é possível detectar possíveis pontos de alavanca, entretanto nota-se a presença de possíveis pontos discrepantes e influentes (gráficos superiores esquerdo e direito) correspondentes as obervações: # 16, #23, #36, #48, #71, #115 e #168. Pelo gráfico de probabilidade meio-normal com com envelope simulado (gráfico inferior direito) é possível visualizar que grande parte dos pontos caem fora das bandas de confiança e por esse fato temos indícios que o modelo de regressão beta parece não ser adequado para modelar os dados de transplante autólogos de células-tronco do sangue periférico.

4.3 Ajuste com o modelo de regressão gama unitária

Consideramos que as observações $y_1, \ldots, y_{239}$ são independentes e têm distribuição gama unitária com média μ_t $(t = 1, \ldots, 239)$, e parâmetro de dispersão ϕ desconhecido. O modelo para as médias é escrito similarmente ao modelo de regressão beta (4.1).

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2}, \tag{4.2}$$

em que se considerou $g(.)$ como a função logito devido à interpretação direta dos parâmetros em termos da razão de chances (odds) e x_{t1} : a idade ajustada do $t-$ésimo indivíduo e x_{t1} : variável *dummy* que indica se o $t-$ésimo paciente recebe quimioterapia em um protocolo de um dia (0) ou em um protocolo de 3 dias (1).

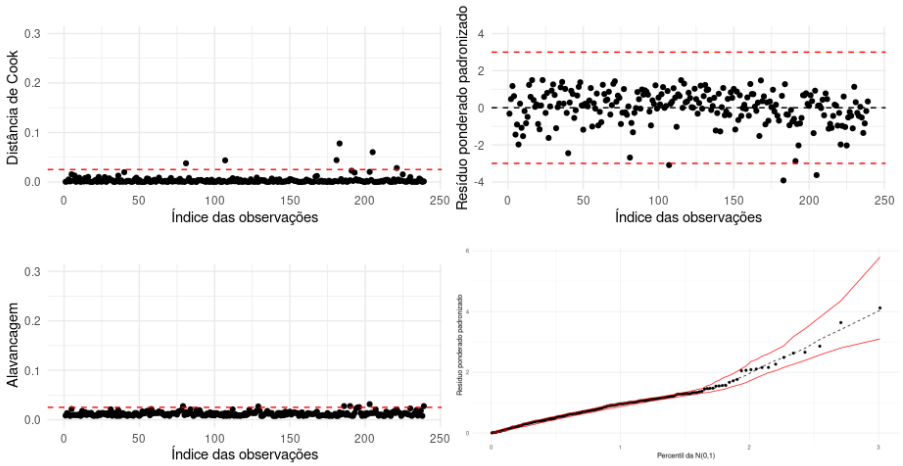
Tabela 2 – Estimativas dos parâmetros do modelo de regressão gama unitária ajustado aos dados de transplante autólogos de células-tronco do sangue periférico.

Parâmetro	Estimativa	Erro Padrão	Valor p
β_0	1,0026	0,1147	$<$ 0,0001
β_1	0,0165	0,0053	0,0019
β_2	0,2400	0,1030	0,0198
ϕ	2,3683	0,2032	0,0001

Fonte: Elaborada pelo autor.

Pode-se observar que todas as estimativas dos parâmetros do modelo proposto são significativas ao nível de significância de 5%. Posteriormente, foi realizada uma análise de diagnóstico para verificar a existência de algum ponto atípico e verificar a adequabilidade do modelo.

Figura 7 – Gráficos de diagnóstico para o modelo de regressão gama unitária ajustado aos dados de transplante autólogos de células-tronco do sangue periférico.



Fonte: Elaborada pelo autor.

Pela Figura 7 (gráfico inferior esquerdo) notamos que parece não haver pontos de alavanca, no entanto nota-se possíveis pontos aberrantes e influentes (gráficos superiores esquerdo e direito), nesse caso, correspondem as seguintes observações: #81, #107, #181, #205 e #221. Ademais, pelo gráfico de probabilidade meio-normal com envelope simulado (gráfico inferior direito) é possível visualizar que todos os pontos se encontram dentro das bandas de confiança.

Dessa forma, pelos critérios de informação de Akaike e Bayesiano de Schwarz, obteve-se AIC $= -400,0404$ e BIC $= -410,4698$ para o modelo de regressão gama unitária com todas as observações e AIC $= -383,3042$ e BIC $= -369,3983$ para o modelo de regressão beta com todas as observações. Portanto, por esse fato e pela análise de diagnóstico, em particular o gráfico probabilidade meio-normal pode-se concluir que o modelo de regressão gama unitária, nesse caso, apresentou um melhor ajuste, assim, é preferível optar por tal modelo para explicar a taxa de recuperação de células CD34+ ao invés o modelo de regressão beta.

5 CONCLUSÃO

Apresentamos a distribuições gama unitária devido a Grassia (1977) e sua versão reparametrizada, além das principais propriedades e o meios de estimação mais usuais. Tal distribuição serviu de base para o modelo de regressão gama unitária proposto por Mousa et al. (2016), ademais, apresentamos as medidas de diagnóstico e influência local devido a (Rocha, 2020; Rocha et al., 2021).

Desenvolvemos uma rotina computacional no *software* `R` para os gráficos da distribuição gama unitária e ajuste do modelo de regressão gama unitária, além disso, implementamos também a análise de diagnóstico baseado no trabalho de Rocha et al. (2021) e o gráfico de probabilidade com envelope simulado.

Foi apresentado uma aplicação em dados reais no qual foram ajustados os modelos regressão gama unitária e o modelo de regressão beta, ambos modelos foram avaliados por meio das medidas de diagnóstico e critérios de informação de Akaike e Bayesiano de Schwarz que por sua vez nos indicou que o modelo de regressão gama unitária apresentou um melhor ajuste aos dados em relação ao modelo de regressão beta.

A distribuição e o modelo de regressão gama unitária apresentados em questão são baseados no contexto frequentista, no entanto, é possível realizarmos uma nova extensão para o contexto bayesiano, isto é, considerando agora os parâmetros μ e ϕ como variáveis aleatórias, assim utilizar os métodos de Monte Carlo via cadeia de Markov para obter os resultados de interesse, o que também pode ser realizado de forma similar para a distribuição Weibull unitária (Mazucheli et al., 2020). Podemos ainda considerar mais extensões, mas agora para além dos dados transversais, nesse sentido, temos a proposta de Petterle et al. (2021) no qual os autores propõem o modelo de regressão gama unitária misto para lidar com variáveis limitadas contínuas no contexto de medidas repetidas e dados agrupados. Ademais, desenvolver a parte de análise de diagnóstico na linha de Nobre e Singer (2007), Nobre e Singer (2011), Pinho et al. (2015) e Singer et al. (2017). Pode-se ainda implementar a distribuição gama unitária no pacote GAMLSS e usar a estrutura já existente para ajuste e possível extensões para modelos não paramétricos. Indo mais além, podemos desenvolver os modelos de regressão quantílicos baseados na gama unitária retangular.

REFERÊNCIAS

ATKINSON, Anthony C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, v. 68, n. 1, p. 13-20, 1981.

AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE transactions on automatic control**, v. 19, n. 6, p. 716-723, 1974.

BARNDORFF-NIELSEN, Ole E.; JØRGENSEN, Bent. Some parametric models on the simplex. **Journal of multivariate analysis**, v. 39, n. 1, p. 106-116, 1991.

BAYES, Cristian L.; BAZÁN, Jorge L.; GARCÍA, Catalina. A new robust regression model for proportions. **Bayesian Analysis**, v. 7, n. 4, p. 841-866, 2012.

COOK, R. Dennis. Detection of influential observation in linear regression. **Technometrics**, v. 19, n. 1, p. 15-18, 1977.

COOK, R. Dennis. Assessment of local influence. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 48, n. 2, p. 133-155, 1986.

DUNN, Peter K.; SMYTH, Gordon K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, v. 5, n. 3, p. 236-244, 1996.

DEY, Sanku; MENEZES, Andre FB; MAZUCHELI, Josmar. Comparison of estimation methods for unit-gamma distribution. **Journal of data science**, v. 17, n. 4, p. 768-801, 2019.

ESPINHEIRA, Patrícia L.; FERRARI, Silvia LP; CRIBARI-NETO, Francisco. **Influence diagnostics in beta regression**. Computational Statistics & Data Analysis, v. 52, n. 9, p. 4417-4431, 2008.

ESPINHEIRA, Patrícia L.; FERRARI, Silvia LP; CRIBARI-NETO, Francisco. **On beta regression residuals**. Journal of Applied Statistics, v. 35, n. 4, p. 407-419, 2008.

FERRARI, Silvia; CRIBARI-NETO, Francisco. **Beta regression for modelling rates and proportions**. Journal of applied statistics, v. 31, n. 7, p. 799-815, 2004.

GUPTA, Arjun K.; NADARAJAH, Saralees. **Handbook of beta distribution and its applications**. CRC press, 2004.

GUEDES, Ana C.; CRIBARI-NETO, Francisco; ESPINHEIRA, Patrícia L. Modified likelihood ratio tests for unit gamma regressions. **Journal of Applied Statistics**, v. 47, n. 9, p. 1562-1586, 2020.

GRASSIA, A. On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. **Australian Journal of Statistics**, v. 19, n. 2, p. 108-114, 1977.

GARCÍA, C. B.; GARCÍA PÉREZ, J.; VAN DORP, Johan René. Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support. **Statistical Methods & Applications**, v. 20, n. 4, p. 463-486, 2011.

HAHN, Eugene David. Mixture densities for project management activity times: A robust approach to PERT. **European Journal of operational research**, v. 188, n. 2, p. 450-459, 2008.

JOHNSON, Norman L. Systems of frequency curves generated by methods of translation. **Biometrika**, v. 36, n. 1/2, p. 149-176, 1949.

KRISTENSEN, Kasper et al. TMB: automatic differentiation and Laplace approximation. **arXiv preprint arXiv:1509.00660**, 2015.

KUMARASWAMY, Ponnambalam. A generalized probability density function for double-bounded random processes. **Journal of hydrology**, v. 46, n. 1-2, p. 79-88, 1980.

LEMONTE, Artur J.; BAZÁN, Jorge L. New class of Johnson distributions and its associated regression model for rates and proportions. **Biometrical Journal**, v. 58, n. 4, p. 727-746, 2016.

LIMA, Francimário Alves de. **Distribuições de probabilidade no intervalo unitário**. Tese de Doutorado. Universidade de São Paulo.

MAZUCHELI, J. et al. The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. **Journal of Applied Statistics**, v. 47, n. 6, p. 954-974, 2020.

MOUSA, Amany M.; EL-SHEIKH, Ahmed A.; ABDEL-FATTAH, Mahmoud A. A gamma regression for bounded continuous variables. **Advances and Applications in Statistics**, v. 49, n. 4, p. 305, 2016.

MIYASHIRO, Eliane Shizue. **Modelos de regressão beta e simplex para análise de proporções**. 2008. Tese de Doutorado. Universidade de São Paulo.

NELDER, John Ashworth; WEDDERBURN, Robert WM. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, v. 135, n. 3, p. 370-384, 1972.

NOBRE, Juvêncio S.; SINGER, Julio M. Leverage analysis for linear mixed models. **Journal of Applied Statistics**, v. 38, n. 5, p. 1063-1072, 2011.

OGUAMANAM, D. C. D.; MARTIN, H. R.; HUISSOON, J. P. On the application of the beta distribution to gear damage analysis. **Applied Acoustics**, v. 45, n. 3, p. 247-261, 1995.

OSPINA, Raydonal; FERRARI, Silvia LP. Inflated beta distributions. **Statistical papers**, v. 51, n. 1, p. 111-126, 2010.

PEREIRA, Ana Cristina Guedes. **Improved likelihood inference in unit gama regressions**. 2017. Dissertação de Mestrado. Universidade Federal de Pernambuco.

PETTERLE, Ricardo R. et al. Unit gamma mixed regression models for continuous bounded data. **Journal of Statistical Computation and Simulation**, p. 1-19, 2021.

PINHO, Luis Gustavo B.; NOBRE, Juvêncio S.; SINGER, Julio M. Cook's distance for generalized linear mixed models. **Computational Statistics & Data Analysis**, v. 82, p. 126-136, 2015.

PEARSON, Karl. Contributions to the mathematical theory of evolution. **Philosophical Transactions of the Royal Society of London**. A, v. 185, p. 71-110, 1894.

PEARSON, Karl. IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, v. 216, n. 538-548, p. 429-457, 1916.

RODRIGUES, Josemar; BAZÁN, Jorge L.; SUZUKI, Adriano K. A flexible procedure for formulating probability distributions on the unit interval with applications. **Communications in Statistics-Theory and Methods**, v. 49, n. 3, p. 738-754, 2019.

ROCHA, Suelena de Souza. Diagnóstico em modelos de regressão gama unitária. 2020.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2022. Disponível em: <https:/www.R-project.org/.>

ROCHA, Suelena; L. ESPINHEIRA, Patrícia; CRIBARI-NETO, Francisco. Residual and local influence analyses for unit gamma regressions. **Statistica Neerlandica**, v. 75, n. 2, p. 137-160, 2021.

SCHWARZ, Gideon. Estimating the dimension of a model. **The annals of statistics**, p. 461-464, 1978.

SEN, Pranab K.; SINGER, Julio M. **Large sample methods in statistics: an introduction with applications**. CRC press, 1994.

SONG, Peter Xue-Kun; TAN, Ming. Marginal models for longitudinal continuous proportional data. **Biometrics**, v. 56, n. 2, p. 496-502, 2000.

SMITHSON, Michael; MERKLE, Edgar C. **Generalized linear models for categorical and continuous limited dependent variables**. CRC Press, 2013.

SANTOS NOBRE, Juvêncio; DA MOTTA SINGER, Julio. Residual analysis for linear mixed models. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, v. 49, n. 6, p. 863-875, 2007.

SINGER, Julio M.; ROCHA, Francisco MM; NOBRE, Juvêncio S. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. **International Statistical Review**, v. 85, n. 2, p. 290-324, 2017.

THOMAS, William; COOK, R. Dennis. Assessing influence on regression coefficients in generalized linear models. **Biometrika**, v. 76, n. 4, p. 741-749, 1989.

ZHANG, Peng; QIU, Zhenguo; SHI, Chengchun. simplexreg: An R package for regression analysis of proportional data using the simplex distribution. **Journal of Statistical Software**, v. 71, n. 11, 2016.

6 APÊNDICE A

6.1 Rotina computacional em R para os gráficos da distribuição gama unitária

```
## Pacotes
library(ggplot2)

## Distrbuição gama unitária (Grassia, 1977)
tau = c(0.5, 1, 1.5, 2)
alpha = c(0.5,1,2,3,4,5)

fun_ugama1 = function(y){
  alpha = 0.5; tau = 1.5
  c = (alpha^(tau))/gamma(tau)
  c*y^(alpha-1)*(log(1/y))^(tau-1)
}

fun_ugama2 = function(y){
  alpha = 1; tau = 1.5
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama3 = function(y){
  alpha = 2; tau = 1.5
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama4 = function(y){
  alpha = 3; tau = 1.5
```

```
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))
}

fun_ugama5 = function(y){
  alpha = 4; tau = 1.5
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama6 = function(y){
  alpha = 5; tau = 1.5
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

g = ggplot(data = data.frame(x = c(0,1)), mapping = aes(x = x))

g1 = g + geom_function(aes(colour = "1"), fun = fun_ugama1,lwd = 0.8, lty =1) +
  geom_function(aes(colour = "2"), fun = fun_ugama2,lwd = 0.8, lty =2) +
  geom_function(aes(colour = "3"), fun = fun_ugama3,lwd = 0.8, lty =3) +
  geom_function(aes(colour = "4"), fun = fun_ugama4,lwd = 0.8, lty =4) +
  geom_function(aes(colour = "5"), fun = fun_ugama5,lwd = 0.8, lty =5) +
  geom_function(aes(colour = "6"), fun = fun_ugama6,lwd = 0.8, lty =6) +
  xlim(0,1) +
  labs(col = " ", x = "y" , y = "f(y)")+
  theme_minimal()+
  scale_colour_discrete(labels = c(expression(paste( alpha, " = 0,5 ", " ;
  " ,tau, " =  1,5"),
  paste( alpha, " = 1 ", " ; " ,tau, " =  1,5"),paste(alpha, " = 2 ", " ;
```

```
  " ,tau, " =  1,5"),
  paste(alpha, " = 3 ", " ; " ,tau, " =  1,5"),paste(alpha, " = 4 ", " ;
  " ,tau, " = 1,5"),
  paste(alpha, " = 5 ", " ; " ,tau, " =  1,5"))))


fun_ugama11 = function(y){
  alpha = 0.5; tau = 2
  c = (alpha^(tau))/gamma(tau)
  c*y^(alpha-1)*(log(1/y))^(tau-1)
}

fun_ugama22 = function(y){
  alpha = 1; tau = 2
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama33 = function(y){
  alpha = 2; tau = 2
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama44 = function(y){
  alpha = 3; tau = 2
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))
}
```

```
fun_ugama55 = function(y){
  alpha = 4; tau = 2
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

fun_ugama66 = function(y){
  alpha = 5; tau = 2
  c = (alpha^(tau)) / gamma(tau)
  c*( y^(alpha-1) * (log(1/y))^(tau-1))

}

g2 = ggplot(data = data.frame(x = c(0,1)), mapping = aes(x = x))

g2 + geom_function(aes(colour = "1"), fun = fun_ugama11,lwd = 0.8, lty =1) +
  geom_function(aes(colour = "2"), fun = fun_ugama22,lwd = 0.8, lty =2) +
  geom_function(aes(colour = "3"), fun = fun_ugama33,lwd = 0.8, lty =3) +
  geom_function(aes(colour = "4"), fun = fun_ugama44,lwd = 0.8, lty =4) +
  geom_function(aes(colour = "5"), fun = fun_ugama55,lwd = 0.8, lty =5) +
  geom_function(aes(colour = "6"), fun = fun_ugama66,lwd = 0.8, lty =6) +
  xlim(0,1) +
  labs(col = " ", x = "y" , y = "f(y)")+
  theme_minimal()+
  scale_colour_discrete(labels = c(expression(paste( alpha, " = 0,5 ", " ;
  " ,tau, " =  2"),
  paste( alpha, " = 1 ", " ; " ,tau, " =  2"),paste(alpha, " = 2 ", " ;
  " ,tau, " =  2"),
  paste(alpha, " = 3 ", " ; " ,tau, " =  2"),paste(alpha, " = 4 ", " ;
  " ,tau, " = 2"),
  paste(alpha, " = 5 ", " ; " ,tau, " =  2"))))
```

6.2 Rotina computacional em R para os gráficos da distribuição gama unitária reparametrizada

```
## Distrbuição gama unitária reparametrizada

fun_ugama1 = function(y){
  mu = 0.5 ; phi = 5
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*((log(1/y))^(phi-1))
}

fun_ugama2 = function(y){
  mu = 0.95 ; phi = 1
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(log(1/y))^(phi-1)
}

fun_ugama3 = function(y){
  mu = 0.5 ; phi = 0.4
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(log(1/y))^(phi-1)
}

fun_ugama4 = function(y){
  mu = 0.75 ; phi = 1
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(log(1/y))^(phi-1)
}

fun_ugama5 = function(y){
  mu = 0.2 ; phi = 10
```

```
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(log(1/y))^(phi-1)
}


fun_ugama6 = function(y){
  mu = 0.8 ; phi = 5
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(log(1/y))^(phi-1)
}


fun_ugama7 = function(y){
  mu = 0.15 ; phi= 1
  k = mu^(1/phi)/ (1-mu^(1/phi))
  (k^(phi))/(gamma(phi)) * y^(k-1)*(-log(y))^(phi-1)
}


fun_ugama8 = function(y){
  mu = 0.6 ; phi = 9
  k = mu^(1/phi)/(1-mu^(1/phi))
  (k^(phi))/(gamma(phi))* y^(k-1)*(log(1/y))^(phi-1)
}

g = ggplot(data = data.frame(x = c(0,1)), mapping = aes(x = x))

g + geom_function(aes(colour = "1"), fun = fun_ugama1,lwd = 0.8, lty = 1) +
  geom_function(aes(colour = "2"), fun = fun_ugama2,lwd = 0.8, lty = 2) +
  geom_function(aes(colour = "3"), fun = fun_ugama3,lwd = 0.8, lty = 3) +
  geom_function(aes(colour = "4"), fun = fun_ugama4,lwd = 0.8, lty = 4) +
  geom_function(aes(colour = "5"), fun = fun_ugama5,lwd = 0.8, lty = 5) +
```

```
  geom_function(aes(colour = "6"), fun = fun_ugama6,lwd = 0.8, lty = 6) +
  geom_function(aes(colour = "7"), fun = fun_ugama7,lwd = 0.8, lty = 7) +
  geom_function(aes(colour = "8"), fun = fun_ugama8,lwd = 0.8, lty = 8) +
  xlim(0,1) +
  ylim(0,10) +
  labs(col = " ", x = "y" , y = "f(y)")+
  theme_minimal()+
  scale_colour_discrete(labels = c(expression(paste(mu, " = 0,5 ", " ;
  " ,phi, " =  5 "),
  paste(mu,  " =  0,95 ", " ; " ,phi, " = 1 "),paste(mu,  " = 0,5 ", " ;
  " ,phi, " = 0,4 "),
  paste( mu,  " = 0,75 ", " ; " ,phi, " = 1 "),paste( mu,  " = 0,2 ", " ;
  " ,phi, " =  10 "),
  paste( mu,  " = 0,8 ", " ; " ,phi, " =  5 "),paste( mu,  " = 0,15 ", " ;
  " ,phi, " =  1 "),
  paste( mu,  " = 0,6 ", " ; " ,phi, " =  9 "))))+
  theme_minimal()
```

6.3 Rotina computacional em C++ para o ajuste do modelo de regressão gama unitária

```
#include <TMB.hpp>
template<class Type>
Type objective_function<Type>::operator() ()
{
    DATA_VECTOR(Y);          //observações
    DATA_MATRIX(X);          //matriz de efeito fixo
    PARAMETER_VECTOR(beta); //vetor de parâmetros
    PARAMETER(logphi);       //parâmetro de precisão

    Type phi = exp(logphi);
```

```
    // preditor linear para média
     vector<Type> mu = exp(X*beta)/(1 + exp(X*beta)); // função logito
    //vector<Type> tau = pow(mu,1/phi)/(1-pow(mu,1/phi));

    // log-verossimilhança negativa
    Type nll = 0;
    for(int i=0; i < Y.size(); i++)

         nll -= phi*log(tau[i]) - lgamma(phi) + (tau[i] - 1)*log(Y[i])+
                (phi -1)*log(-log(Y[i]));

    // método delta
     ADREPORT(phi);

     return nll;
}
```

6.4 Rotina computacional em R para o ajuste do modelo de regressão gama unitária

```
library(TMB)

compile("gama_unitaria.cpp")
dyn.load(dynlib("gama_unitaria"))

dados_gu = list(Y=Y, X=X)

## Chutes inicias para os parâmetros

parametros = list(beta = c(rep(0,3) ), logphi = 0)
```

```
## Estimação

gu_TMB = MakeADFun(dados_gu, parametros, DLL = "gama_unitaria",
hessian = TRUE, silent = TRUE)

opt = nlminb(start = gu_TMB$par, obj = gu_TMB$fn, gr = gu_TMB$gr)

rep = sdreport(gu_TMB)

## Sumário das estimativas

summary(rep, "fixed", p.value = TRUE)
summary(rep, "report", p.value = TRUE)

## Critérios de informação

BIC_gu = -2*(logV) + log(n)*p
AIC_gu = -2*(logV) + 2*p



## modelo gama unitária c/ função logito

Xgama = beta0 + beta1*x1 + beta2*x2
gu_predict =  exp(Xgama)/(1+exp(Xgama))
```

6.5 Rotina computacional em R para o diagnóstico do modelo de regressão gama unitária

```
## Quantidades estimadas
yt = y # variável reposta dos dados
k = p  # número de parâmetros do modelo
n = length(yt)
```

```
phi_hat = exp(rep$par.fixed[4])
mu_hat = gu_predict
alpha_hat = (mu_hat^(1/phi_hat))/(1-mu_hat^(1/phi_hat))
mu_star = (alpha_hat)/(mu_hat^((1/phi_hat) + 1))
y_star = (alpha_hat^2 * log(yt))/(phi_hat*mu_hat^((1/phi_hat) + 1))
a_hat = alpha_hat/(mu_hat^((1/phi_hat) + 1))
g_l = 1/(mu_hat*(1-mu_hat))
Wbb = diag( ((a_hat^2)/phi_hat) * (1/g_l)^2  )
Wbb_12 = sqrt(Wbb)
s_hat = mu_star + y_star
H = Wbb_12%*%X%*%solve(t(X)%*%Wbb%*%X)%*%t(X)%*%Wbb_12
ht = diag(H)

## Função distribuição gama unitária
dGU = function(y, mu, phi, log = FALSE){

  # Parametros
  alpha = mu^(1/phi)/(1- (mu^(1/phi)))

  # Densidade g.u (reparametrizada)
  fy = (alpha^(phi)/gamma(phi))*y^(alpha-1)
  *log(1/y)^(phi-1)

  if(log){return(log(fy))}else{ return(fy)}
}

## Função distribuição gama unitária
pGU = function(q, mu, phi, lower.tail = TRUE, log.p = FALSE){

  # Função distribuição
  value = 0
```

```
  for(i in 1:length(q)){

    if(lower.tail){
      value[i] = integrate(function(x) dGU(x,mu[i],phi), 0, q[i])$value
    }else{value[i] = 1-integrate(function(x) dGU(x,mu[i],phi), 0, q[i])$value
    }

  } # fim for


  if(log.p){
    return(log(value))
  }else{
    return(value)
  }

}


## Gráfico quantil-quantil

N = nrow(dados)
resq = 0

for(i in 1:N){
  resq[i] = qnorm(pGU(yt[i],mu_hat[i],phi_hat))
}

dados_resq = data.frame(resq)


## Gráficos diagnóstico
```

```
library(ggplot2)
library(patchwork)
library(car)
library(qqplotr)

# Gráfico quantil-quantil
g1 = ggplot(dados_resq,aes(sample = resq)) +
  stat_qq_band(bandType = "pointwise", fill = "gray", alpha = 0.8) +
  stat_qq_line(colour = "red") +
  stat_qq_point()+
  theme_minimal()+
  labs(x = "Quantis da Normal padrão ", y = "Resíduo Quantílico", title = "")

# Resíduo ponderado padronizado
r_pond_pd = s_hat/sqrt(((mu_star^2)/phi_hat)*(1-ht))

g2 = ggplot() +
  geom_point(aes(y = r_pond_pd , x = 1:length(yt)))+
  labs(x = "Índice das observações",
  y = "Resíduo ponderado padronizado", title = "")+
  ylim(-4,4)+
  geom_hline(aes(yintercept= 0), col = "black", linetype = 2)+
  geom_hline(aes(yintercept= 3), col = "red", linetype = 2)+
  geom_hline(aes(yintercept= -3), col = "red", linetype = 2)+
  theme_minimal()

# Resíduo ponderado padronizado X preditor linear
g3 = ggplot() +
  geom_point(aes(x= mu_hat,y = r_pond_pd), ymin = -4, ymax = 4)+
  labs(x = "Preditor linear", y = "Resíduo ponderado padronizado", title = "")+
  geom_hline(aes(yintercept= 0), col = "black", linetype = 2)+
  ylim(-4,4)+
```

```
  geom_hline(aes(yintercept= 3), col = "red", linetype = 2)+
  geom_hline(aes(yintercept= -3), col = "red", linetype = 2)+
  theme_minimal()

# Distância de Cook
DC2 = (r_pond_pd^2)* (ht/(k*(1-ht)))

g4 = ggplot() +
  geom_point(aes(y = DC2, x = 1:length(yt)))+
  labs(x = "Índice das observações", y = "Distância de Cook", title = "")+
  ylim(0, 0.30)+
  geom_hline(aes(yintercept= 2*k/n), col = "red", linetype = 2)+
  theme_minimal()

# Alavanca
g5 = ggplot() +
  geom_point(aes(y = ht, x = 1:length(yt)))+
  labs(x = "Índice das observações", y = "htt", title = "")+
  geom_hline(aes(yintercept= 2*k/n), col = "red", linetype = 2)+
  ylim(0, 0.30)+
  theme_minimal()
```