UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE ESTATÍSTICA

ALIPIO JOSÉ DE SOUZA PACHECO FILHO

ASPECTOS TEÓRICOS DOS MODELOS GAMLSS E APLICAÇÃO A DADOS LONGITUDINAIS

FORTALEZA

2021

ALIPIO JOSÉ DE SOUZA PACHECO FILHO

ASPECTOS TEÓRICOS DOS MODELOS GAMLSS E APLICAÇÃO A DADOS LONGITUDINAIS

Monografia apresentada ao Curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Juvêncio Santos Nobre.

FORTALEZA
2021

Aos meus amados filhos
João Arthur e Heitor Lucas.

AGRADECIMENTOS

À Universidade Federal do Ceará (UFC), que já foi uma segunda casa. Sempre se apresentou como um local de boas discussões para crescimento intelectual e humano.

Ao Departamento de Estatística e Matemática Aplicada (DEMA), que apresenta um time excelente de profissionais, desde os professores, coordenadores até o pessoal que exerce a importante função de manter tudo limpo e em ordem.

Ao meu orientador e amigo, prof. Juvêncio Nobre, pelas conversas e orientações. A única coisa que lamento foi não o ter encontrado nos momentos iniciais do curso.

Ao mestre de todos, o queridíssimo prof. Maurício Mota. Todos que almejam ser professor (verdadeiramente) devem ter o prazer de vê-lo em ação numa sala de aula.

Aos professores Gualberto, Welliandre, Aílton e Rafael Farias, à prof. Silvia e ao demais professores pelas aulas e diálogos ao longo do curso.

Aos colegas do curso de Estatística, em especial ao Matheus, André (meu amigo dançarino) e Guilherme (meu amigo livreiro) que me foram mais próximos na etapa final do curso. Ao meu amigo Wladimir, que infelizmente saiu do curso.

Ao querido amigo, Dr. Elton Nascimento, por fornecer seus dados para execução desse trabalho e por sua boa energia positiva.

"Não existem métodos fáceis para resolver problemas difíceis."

René Descartes

"A essência do conhecimento consiste em aplicá-lo, uma vez possuído".

Confúcio

RESUMO

A suposição de normalidade dos erros é comumente considerada na utilização de modelos de regressão linear (MRL). Embora outros modelos tenham surgido para lidar com essa limitação, tais como os modelos lineares generalizados (MLGs) e os aditivos generalizados (MAGs), eles ainda se restringem à família exponencial linear. Na expectativa de superar tais limitações, foi proposto o modelo aditivo generalizado para localização, escala e forma (Generalized additive model for location, scale and shape - GAMLSS), que é capaz de modelar dados com, teoricamente, qualquer distribuição. Mesmo com toda essa flexibilidade, o GAMLSS ainda apresenta uma utilização pouco expressiva frente aos outros modelos que lidam com distribuições distintas da normal (tais como MLGs e MAGs). Essa expressividade é ainda menor quando se trata da classe de modelos com variáveis latentes (modelos mistos). Neste trabalho, é apresentado um referencial teórico sobre essa classe de modelos e realizada uma aplicação a um conjunto de dados de natureza longitudinal de produção polínica. Empregando o pacote `gamlss` implementado no software R, percebe-se a enorme flexibilidade do GAMLSS (dada a grande quantidade de distribuições disponíveis no pacote e a capacidade de modela explicitamente diferentes tipos de parâmetros, por exemplo). A utilização do modelo GAMLSS com distribuição marginal gama generalizada mostrou bom ajuste aos dados de produção polínica, sendo capaz de lidar com a assimetria dos dados. A modelagem dos dados de produção de pólen via modelos lineares mistos (MLM) violou alguns pressupostos (tais como normalidade dos resíduos e homogeneidade da variância). Por outro lado, o emprego do GAMLSS apresentou bom ajuste sem violar pressupostos, sugerindo que esta classe de modelos é capaz de fornecer inferências e estimativas mais confiáveis. Contudo, a ausência de ferramental para análise de influência (amplamente abundante nos MLM) ainda é uma limitação do GAMLSS.

ABSTRACT

The assumption of normality of errors is commonly considered when using linear regression models (LRM). Although other models have emerged to deal with this limitation, such as generalized linear models (GLMs) and generalized aditives models (GAMs), they are still restricted to the linear exponential family. In order to overcome such limitations, the generalized additive model for location, scale and shape (GAMLSS) was proposed, which is capable of modeling data with, theoretically, any distribution. Even with all this flexibility, GAMLSS still has a little expressive use compared to other models that deal with distributions different from the normal (such as GLMs and GAMs). This expressiveness is even lower when it comes to the class of models with latent variables (mixed models). In this work, a theoretical framework on this class of models is presented and an application is made to a dataset of longitudinal nature of pollen production. Using the `gamlss` package implemented in the `R` software, one can see the enormous flexibility of GAMLSS (given the large number of distributions available in the package and the ability to explicitly model different types of parameters, for example). The use of the GAMLSS model with generalized gamma marginal distribution showed a good fit to the pollen production data, being able to deal with the asymmetry of the data. The modeling of pollen production data via mixed linear models (MLM) violated some assumptions (such as normality of residuals and homogeneity of variance). On the other hand, the use of GAMLSS showed a good fit without violating assumptions, suggesting that this class of models is able to provide more reliable inferences and estimates. However, the absence of influence analysis tools (largely abundant in MLM) is still a limitation of GAMLSS.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCCG	Distribuição Box-Cox Cole e Green
BCT	Distribuição Box-Cox t
CV	Coeficiente de variação
DP	Desvio-padrão
EM	Algoritmo esperança e maximização (*expectation-maximization*)
EP	Erro-padrão
GAIC	Critério de informação de Akaike generalizado
GAMLSS	Modelo aditivo generalizado para localização, escala e forma
GDEV	Desvio global (*global deviance*)
GG	Distribuição gama generalizada
gl	Graus de liberdade
IC	Intervalo de confiança
IG	Distribuição gama inversa
LOGNO	Distribuição log-normal
MAG	Modelo aditivo generalizado
MCMC	Monte Carlo via cadeia de Markov
ML	Máxima verossimilhança
MLG	Modelo linear generalizado
MLM	Modelo linear misto
MMP	Método de mínimos quadrados ponderados
MMPP	Método de mínimos quadrados ponderados penalizados
MRL	Modelo de regressão linear
MRLM	Modelo de regressão linear múltiplo
OECD	Organização para a Cooperação e Desenvolvimento Econômico
QQ	Quantil-quantil
QVP	Quase verossimilhança penalizada
REML	Máxima verossimilhança restrita
VC	Validação cruzada
VCG	Validação cruzada generalizada

SUMÁRIO

INTRODUÇÃO GERAL

A utilização de modelos mistos é amplamente difundida em diversos campos do conhecimento e ainda se mostra um campo fecundo de pesquisa (Baumann *et al*. 2019, Fitzmaurice *et al*., 2008, Schielzeth *et al*., 2020). Os primeiros modelos mistos consideravam unicamente a variável resposta como tendo distribuição normal. Contudo, com o avanço científico, modelos que consideram outras distribuições além da normal surgiram e se consolidaram como eficientes ferramentas na análise de dados (Singer *et al*., 2017, Zuur *et al*. 2009).

Duas principais classes de modelos mistos são comumente empregadas para lidar com situações em que não é razoável admitir normalidade por parte da variável resposta, os modelos lineares generalizados mistos e os modelos aditivos generalizados mistos. Embora eles apresentem considerável flexibilidade quando comparados aos modelos lineares mistos (MLM), ainda confrontam com algumas limitações, tal como lidar apenas com distribuições pertencentes a família exponencial linear (Hastie e Tibshirani, 1990, Nelder e Wedderburn, 1972, MacCulloch e Searle, 2004). Nesse sentido, os modelos aditivos generalizados para localização, escala e forma (denominado de GAMLSS) surgiram com uma estrutura teórica e com um ferramental capaz de modelar uma gama maior de distribuições quando comparados às propostas anteriores (Rigby e Stasinopoulos, 2005).

O GAMLSS permite, em tese, que qualquer distribuição de probabilidade seja assumida para a variável resposta e permite ainda modelar seus parâmetros. Atualmente, a principal implementação dessa classe de modelos está presente no software R (R Core Team, 2021) através do pacote `gamlss` (Rigby e Stasinopoulos, 2005) e outros pacotes complementares. Existem mais de 80 distribuições disponíveis no pacote `gamlss.dist`, além da possibilidade de se trabalhar com misturas finitas no pacote `gamlss.MX` (Stasinopoulos *et al*. 2017).

Os objetivos desse trabalho são (*i*) apresentar alguns aspectos teóricos dos modelos GAMLSS e (*ii*) realizar uma aplicação de tais modelos a dados longitudinais

reais, comparando seu ajuste com aquele obtido via modelos lineares mistos, o ferramental mais utilizado na prática. Para tal, inicialmente é apresentado um breve referencial teórico e, por fim, apresentada a aplicação do GAMLSS aos dados de produção de pólen em colmeias de abelhas *Apis mellifera*, presentes em Nascimento *et al*. (2019).

CAPÍTULO 1

ASPECTOS TEÓRICOS DOS MODELOS ADITIVOS GENERALIZADOS DE LOCALIZAÇÃO, ESCALA E FORMA (GAMLSS)

1. Introdução

Historicamente, os modelos de regressão têm sido amplamente empregados nas mais diversas áreas do conhecimento, tais como: Biologia, Agronomia, Engenharia, Economia, Sociologia etc. Um dos principais objetivos do uso desses modelos é a obtenção de uma equação que explique de modo satisfatório a relação entre uma variável resposta e uma (ou mais) variáveis explicativas. Isso possibilitaria fazer predição de valores da variável de interesse (Montgomery *et al*., 2012; Zuur *et al*. 2009).

Os modelos de regressão lineares (MRL) iniciais foram (e são) extremamente limitados em seu uso, dados os seus pressupostos, tais como (i) relação linear entre o $E(y|x_1, \dots, x_p)$ e as variáveis explicativas, normalidade e homoscedasticidade dos erros e ausência de colinearidade. Todavia, com o avanço desses modelos e sua importância crescente, esforços de pesquisas foram dirigidos ao campo da estatística teórica, o que culminou com MRL cada vez mais complexos, capazes de lidar com relações não lineares entre $E(y|x_1, \dots, x_p)$ e as variáveis explicativas, distribuições outras além da normal e até mesmo a utilização de outras funções de regressão, como valor mediano e quantis, por exemplo (Stasinoupolos *et al*. 2017).

Neste cenário surgiram os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS). Esses modelos, além de superar os pressupostos já mencionados (com exceção da colinearidade), são capazes de modelar todos os parâmetros de localização, escala e forma da distribuição da variável resposta. Dessa forma, esses parâmetros podem ser modelados em função das variáveis explicativas. Ademais, pode-se ainda inserir funções não-paramétricas de suavização nos preditores, além de efeitos aleatórios ou outros termos aditivos (Stasinoupolos *et al*. 2017).

O objetivo deste capítulo é realizar uma breve apresentação dos principais modelos de regressão, apresentando seus aspectos teóricos, destacadamente dos

GAMLSS. Adicionalmente, é apresentada uma análise cienciométrica[1] sobre o emprego dos GAMLSS frente a outro modelos.

2. Aspectos teóricos

2.1. Modelos de regressão e o GAMLSS

A análise de regressão é, indubitavelmente, uma das técnicas mais difundidas e empregadas para estudar a relação funcional entre uma variável de interesse e uma ou mais variáveis explicativas (Montgomery *et al*., 2012). Um modelo introdutório é o modelo de regressão linear múltiplo (MRLM), que pode ser expresso por

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ} + \epsilon_i, \quad (1)$$

em que y_i representa a variável resposta, com $i = 1, \ldots, n$ (n sendo o tamanho amostral), $\beta_j's$ são os coeficientes do modelo, com $j = 0, \ldots, J$, x_{iJ} são as J variáveis explicativas (com $x_{i0} = 1, \forall i$) e ϵ_i é a fonte de variação, com $\epsilon_i \overset{iid}{\sim} N(0; \sigma^2)$. Esta especificação é equivalente a $y_i \sim N(\mu_i, \sigma^2)$, com $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ}$ e $y_i's$ independentes. Embora não seja obrigatório ter intercepto (β_0), os modelos aqui considerados serão apresentados com ele. O modelo pode ser reescrito na forma matricial como

$$\begin{gathered} \boldsymbol{y} \overset{ind}{\sim} N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I_n}) \\ \boldsymbol{\mu} = \boldsymbol{X\beta} \end{gathered} \quad (2)$$

[1] Análise cienciométrica é uma análise da cienciometria. Esta é a ciência que busca analisar a produção científica e tecnológica, através do estudo dos aspectos quantitativos da produção intelectual com o objetivo de mensurar e compreender a dimensão científica (Parra *et al*. 2019).

em que $\boldsymbol{y} = (y_i, \dots, y_n)^T$, $\boldsymbol{X}$ é a matriz de planejamento $n \times p$ $(p = J + 1)$ que contém as covariáveis mais uma coluna de 1's, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_J\right)^T$ é o vetor de coeficientes do modelo e $\boldsymbol{I_n}$ representa a matriz identidade de ordem n.

O emprego do MRLM comumente pressupõe normalidade e homoscedasticidade dos erros, além de supor que a relação dos parâmetros do modelo com a esperança da variável resposta condicional nas variáveis explicativas seja linear nos parâmetros. Tais pressuposições são muito restritivas e excluem o uso do MRLM em diversas situações práticas, tais como a modelagem de variáveis binárias, proporções e contagem (entre diversas outras situações).

Para sobrepujar tais limitações do MRLM, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MGLs), que podem ser representados por

$$\begin{aligned} y_i &\overset{ind}{\sim} \mathcal{FE}(\mu_i, \phi) \\ g(\mu_i) &= \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \end{aligned} \tag{3}$$

com $\mathcal{FE}$ denotando a família exponencial linear de distribuições, i e j iguais aos apresentados na equação (1), ϕ sendo o parâmetro de dispersão e $g(.)$ uma função monótona e ao menos duplamente diferenciável denominada de função de ligação. Pode-se escrever a equação (3) como

$$\begin{aligned} \boldsymbol{y} &\overset{ind}{\sim} \mathcal{FE}(\boldsymbol{\mu}, \phi) \\ g(\boldsymbol{\mu}) &= \boldsymbol{X\beta}, \end{aligned} \tag{4}$$

com $\boldsymbol{\mu} = (\mu_i, \dots, \mu_n)^T$ sendo o vetor dos parâmetros de localização. Esses modelos são estruturados por três componentes. O primeiro é o componente aleatório formado pela variável aleatória $\boldsymbol{y} = (y_i, \dots, y_n)^T$, de modo que $E(\boldsymbol{y}) = \boldsymbol{\mu}$. A matriz de especificação $\boldsymbol{X}$ e os coeficientes do modelo $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_J\right)^T$ formam o segundo componente, a parte sistemática, de modo que $\boldsymbol{\eta} = \boldsymbol{X\beta}$ (denominado de preditor

linear). O terceiro componente é a função de ligação que relaciona o componente aleatório ao componente sistemático, $g\big(E(\boldsymbol{y})\big) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}$.

Dessa forma, os MLGs empregam funções de ligação para relacionar o valor esperado da variável resposta com as variáveis explicativas. A escolha adequada dessa função permite que esse relacionamento seja linear e que os valores ajustados do modelo assumam valores no suporte da distribuição. Os MLGs findam por ser uma generalização flexível de regressão linear que permite modelar variáveis respostas que têm distribuição pertencente à família exponencial linear. Contudo, as funções de ligação são, por vezes, insuficientes para analisar a relação entre a resposta e o preditor linear.

Para contornar a limitação da estrutura da relação entre $E(\boldsymbol{y})$ e $\boldsymbol{X\beta}$, Hastie e Tibshirani (1990) propuseram os modelos aditivos generalizados (MAGs). Os MAGs são uma flexibilização dos MLGs com um preditor linear envolvendo a soma de funções suavizadas das covariáveis. A aplicação de funções de suavização não-paramétricas às covariáveis permite que os dados moldem a natureza da relação entre $E(\boldsymbol{y})$ e $\boldsymbol{X\beta}$ de modo distinto àquele proposto pelos MLGs. A forma funcional dos MAGs é dada por

$$\begin{gathered} y_i \overset{ind}{\sim} \mathcal{FE}(\mu_i, \phi) \\ g(\mu_i) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{j=1}^{J} s_j(x_{ij}), \end{gathered} \tag{5}$$

em que s_j é uma função suavizada não-paramétrica aplicada à variável explicativa x_j, $j = 1, \dots, J$ *e* função de ligação $g(.)$ é a mesma da equação (3). Na prática, o s_j é estimado a partir dos dados usando técnicas desenvolvidas para suavização em gráficos de dispersão. Existem muitos tipos de gráficos de dispersão suavizados, tais como splines cúbicos, B-splines, Lowess de Cleveland (1979) e o supersmoother de Friedman e Stuetzle (1981). Os MAGs também podem ser expressos por

$$\begin{aligned} \boldsymbol{y} &\overset{ind}{\sim} \boldsymbol{\mathcal{F}}\mathcal{E}(\boldsymbol{\mu}, \phi) \\ g(\boldsymbol{\mu}) = \boldsymbol{\eta} &= \boldsymbol{X\beta} + \boldsymbol{s}_1(\boldsymbol{x}_1) + \cdots + \boldsymbol{s}_J(\boldsymbol{x_J}) \end{aligned} \tag{6}$$

com $\boldsymbol{\mu}, \phi, \boldsymbol{X}$ e $\boldsymbol{\beta}$ iguais àqueles da equação (4) e $\boldsymbol{s}_j$ uma função suavizada não-paramétrica aplicada ao vetor $\boldsymbol{x}_j = \left(x_{1j}, \dots, x_{nj}\right)^T$, com $j = 1, \dots, J$.

No entanto, os MLGs e os MAGs não são capazes de modelar outros parâmetros além da média ou da variância da distribuição da resposta de modo explícito como funções das variáveis explicativas. Ademais, existem situações em que é requerido o emprego de distribuições que não pertençam à família exponencial linear, exigindo um modelo mais flexível. Para superar essas limitações, Rigby e Stasinopoulos (2005) desenvolveram os modelos aditivos generalizados para localização, escala e forma (GAMLSS). Esses modelos, além de generalizar os MLGs e MAGs, permitem, em tese, a utilização de qualquer distribuição para a variável resposta e a modelagem de parâmetros outros além da média e variância. Desse modo, parâmetros de localização, escala e forma podem ser modelados em função das covariáveis e os preditores lineares podem incorporar funções suavizadas não-paramétricas. Os modelos GAMLSS podem ser expressos por

$$\begin{aligned} y_i &\overset{ind.}{\sim} \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i) \\ g_1(\mu_i) &= \beta_{0_1} + \sum_{j=1}^{J_1} \beta_j x_{ij} + \sum_{j=1}^{J_1} s_j(x_{ij}) \\ g_2(\sigma_i) &= \beta_{0_2} + \sum_{j=1}^{J_2} \beta_j x_{ij} + \sum_{j=1}^{J_2} s_j(x_{ij}) \\ g_3(\nu_i) &= \beta_{0_3} + \sum_{j=1}^{J_3} \beta_j x_{ij} + \sum_{j=1}^{J_3} s_j(x_{ij}) \\ g_4(\tau_i) &= \beta_{0_4} + \sum_{j=1}^{J_4} \beta_j x_{ij} + \sum_{j=1}^{J_4} s_j(x_{ij}) \end{aligned} \tag{7}$$

em que $i = 1, \dots, n$, $j = 1, \dots, J_k$ e β_{0_k} o intercepto referente a k-ésima função de ligação que está associado a um dos 4 parâmetros de interesse, com $k = 1, 2, 3, 4$

representando o número de parâmetros que a distribuição de y_i é considerada assumir. O símbolo $\mathcal{D}$ representa uma distribuição qualquer com quatro parâmetros, em que μ e σ são os parâmetros de localização e escala, respectivamente, enquanto ν e τ são os parâmetros de forma. Matricialmente, pode-se expressar o GAMLSS como

$$\begin{aligned} \boldsymbol{y} &\overset{ind.}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\ g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_{11}\boldsymbol{\gamma}_{11} + \cdots + \boldsymbol{Z}_{1J_1}\boldsymbol{\gamma}_{1J_1} \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{21}\boldsymbol{\gamma}_{21} + \cdots + \boldsymbol{Z}_{2J_2}\boldsymbol{\gamma}_{2J_2} \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{Z}_{31}\boldsymbol{\gamma}_{31} + \cdots + \boldsymbol{Z}_{3J_3}\boldsymbol{\gamma}_{\boldsymbol{3}J_3} \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \boldsymbol{Z}_{41}\boldsymbol{\gamma}_{41} + \cdots + \boldsymbol{Z}_{4J_4}\boldsymbol{\gamma}_{4J_4}, \end{aligned} \tag{8}$$

em que $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ é o vetor de parâmetros de localização, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)^T$ o vetor de parâmetros de escala e $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)^T$ e $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)^T$ os vetores dos parâmetros de forma da distribuição. A matriz de especificação $\boldsymbol{X}_k$ está associada aos efeitos fixos $\boldsymbol{\beta}_k$ referente ao modelo do k-ésimo parâmetro. Tem-se que $s_{j_k}(x_{ij_k}) = \boldsymbol{Z}_{kJ_k}\boldsymbol{\gamma}_{kJ_k}$, com $s_{j_k}(x_{ij_k})$ sendo a função suavizada não-paramétrica empregada nos MAGs e $\boldsymbol{Z}_{kJ_k}$ a matriz de especificação associada aos efeitos aleatórios $\boldsymbol{\gamma}_{kJ_k}$ do modelo do k-ésimo parâmetro. Assume-se que os $\boldsymbol{\gamma}_{kj}$ são independentes entre si e $\boldsymbol{\gamma}_{kj} \sim N(0, \left[\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\right]^{-1})$, em que $\left[\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\right]^{-1}$ é a inversa generalizada da matriz simétrica $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$ e $\boldsymbol{\lambda}_{kj}$ é o vetor de hiperparâmetros. Resumidamente, os coeficientes dos efeitos fixos e aleatórios são representados, respectivamente, por $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_4^T)^T$ e $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_{11}^T, \dots, \boldsymbol{\gamma}_{1J_1}^T, \boldsymbol{\gamma}_{21}^T, \dots, \boldsymbol{\gamma}_{4J_4}^T\right)^T$.

2.2. Estimação de parâmetros

Parâmetros são funções dos valores da variável de estudo (OECD, 2021). Eles podem ser classificados como sendo de localização, escala ou forma. Seja $\mathcal{F}_\theta = \{f(.;\theta_1,\theta_2), \theta_1 \in \mathbb{R}, \theta_2 > 0\}$ uma família de densidades. Os parâmetros θ_1 e θ_2 são

definidos como parâmetros de localização e escala, respectivamente, se e somente se a densidade $f(x;\theta_1,\theta_2)$ puder ser escrita da forma $f(x;\theta) = \theta_2^{-1} g\left(\frac{x-\theta_1}{\theta_2}\right)$, com g sendo uma função conhecida, denominada de função geradora (Mood *et al.*, 1973). O parâmetro de localização determina a "localização" ou o deslocamento da distribuição, enquanto o de escala determina a dispersão estatística da distribuição de probabilidade (OECD, 2021). Por outro lado, parâmetro de forma pode ser entendido como qualquer parâmetro que influencia a forma da distribuição de probabilidade, mas não atende a definição de parâmetros de localização ou escala. No GAMLSS, tais parâmetros podem ser estimados de modo independente por funções que podem incluir diferentes conjuntos de covariáveis.

Os modelos GAMLSS expressos somente por $g_k(\boldsymbol{\mu}) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k$ são denominados de paramétricos, enquanto os modelos de variáveis latentes (ou modelos de efeitos aleatórios) são acrescidos por $\boldsymbol{Z}_{kj}\boldsymbol{\gamma}_{kj}$. Os parâmetros dos denominados modelos paramétricos são estimados por máxima verossimilhança, conforme o logaritmo da função de verossimilhança (denominada de log-verossimilhança de agora em adiante) que se segue

$$l = \sum_{i=1}^{n} ln[f(y_i|\mu_i,\sigma_i,\nu_i,\tau_i)]. \tag{6}$$

Para os modelos de variáveis latentes é empregada a máxima verossimilhança penalizada, conforme a função de log-verossimilhança

$$l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^T \boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\boldsymbol{\gamma}_{kj}. \tag{7}$$

Os modelos paramétricos consideram somente a parte paramétrica da função de regressão e não possuem funções de suavização (os efeitos aleatórios),

requerendo apenas a estimação de $\boldsymbol{\beta}$, enquanto os modelos de efeitos aleatórios requerem a estimação de $\boldsymbol{\gamma}$ e $\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}_{11}^T, \ldots, \boldsymbol{\lambda}_{1J_1}^T, \boldsymbol{\lambda}_{21}^T, \ldots, \boldsymbol{\lambda}_{4J_4}^T\right)^T$. Existem dois algoritmos (denominados RS e CG), além da mistura deles, para maximizar l (modelo paramétrico com respeito à $\boldsymbol{\beta}$) e l_p (o modelo de efeitos aleatórios com respeito à $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$) para $\boldsymbol{\lambda}$ fixo.

O algoritmo RS, proposto por Rigby e Stasinopoulos (2005), maximiza a log-verossimilhança para cada parâmetro $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^T$ por vez em um processo iterativo até a convergência. Esse algoritmo pode ser descrito como tendo três partes que compõem um todo: i) as iterações externas, ii) as iterações internas, e iii) o algoritmo de ajuste posterior modificado (Figura 1). O algoritmo é iniciado nas iterações externas, mas a maximização é realizada pelas iterações internas e o algoritmo de ajuste posterior modificado, cada um hierarquicamente contido no anterior. Após a maximização da função de log-verossimilhança para cada parâmetro, uma medida de desvio global $\left(global\ deviance\ -\ GDEV = -2l\left(\widehat{\boldsymbol{\theta}}\right)\right)$ é obtida para analisar a convergência. A não convergência do GDEV implicará em uma outra rodada do algoritmo.

O algoritmo CG é uma generalização daquele proposto por Cole e Green (1992), o que justifica a sigla CG. Ele é bastante similar ao RS, diferindo principalmente pelo fato de empregar as derivadas cruzadas da função de verossimilhança. Isso faz com que tal algoritmo seja capaz de atualizar conjuntamente todo o vetor de parâmetros $\boldsymbol{\theta}$. Contudo, em algumas funções de distribuição os parâmetros têm informação ortogonal e as derivadas cruzadas são nulas. Nessas situações, o algoritmo RS deve ser empregado.

Figura 1. Esquema de três partes do algoritmo RS e CG. MMP é o método de mínimos quadrados ponderados e MMPP é o mínimos quadrados ponderados penalizados. As setas representam a interação cíclica entre as partes. Adaptado de Thomas (2017).

2.3. Estimação de hiperparâmetros $\boldsymbol{\lambda}$

Para o ajuste de vários termos de suavização em um GAMLSS, existe um método global (externo) e um local (interno) para o algoritmo de ajuste. Rigby e Stasinopoulos (2014) afirmam que o método local tende a ser muito mais rápido, geralmente produzindo resultados semelhantes ao global. O método local ajusta cada um dos hiperparâmetros sequencialmente em ciclos até a convergência dentro do algoritmo RS. Além desses métodos, existem outros a serem escolhidas para a estimativas dos parâmetros de suavização. As estratégias atuais disponíveis são a máxima verossimilhança (ML), a máxima verossimilhança restrita (REML), o critério de informação de Akaike generalizado (GAIC), a validação cruzada (VC) e validação cruzada generalizada (VCG). Detalhes sobre os procedimentos de estimação podem ser obtidos em Stasinopoulos *et al*. (2017), por exemplo.

2.4. Efeitos aleatórios

Os modelos de efeitos mistos apresentam tanto fatores fixos quanto aleatórios, além do erro experimental. Nestes modelos, os parâmetros da regressão podem variar de indivíduo para indivíduo explicando as fontes de heterogeneidade da população. Ademais, eles são capazes de lidar com correlação entre observações individuais por introduzir diferentes fontes de variação nos dados. Isso é conseguido assumindo que existem efeitos aleatórios $\boldsymbol{\gamma}$ fornecendo uma fonte adicional de variação, que têm fdp $f(\boldsymbol{\gamma}|\boldsymbol{\lambda})$, em que $\boldsymbol{\lambda}$ é um vetor de hiperparâmetros a ser estimado a partir dos dados. Pode-se pensar em $\boldsymbol{\gamma}$ como um vetor de variáveis latentes ou não observadas, que existem no modelo para levar em conta a interdependência ou superdispersão. No modelo GAMLSS, dados os efeitos aleatórios $\boldsymbol{\gamma}$, presume-se que as variáveis respostas $y_1, \dots, y_n$ sejam distribuídas independentemente com fdp $f(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma})$. Sob essas premissas, a densidade marginal de $\boldsymbol{y} = (y_1, \dots, y_n)^T$ é dada por

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\lambda})\, d\gamma, \tag{8}$$

em que, dado $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})$ denota a distribuição marginal de $\boldsymbol{y}$, $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ é a distribuição condicional de $\boldsymbol{y}$ dado $\boldsymbol{\gamma}$, e $f(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ é a distribuição marginal para os efeitos aleatórios $\boldsymbol{\gamma}$. O principal obstáculo para modelar a distribuição marginal de $\boldsymbol{y}$ é a integral sobre as variáveis aleatórias $\boldsymbol{\gamma}$. Para superar o problema, existem várias formas para avaliar (8) implementadas na literatura, incluindo o algoritmo EM, aproximação de Laplace, técnicas de Monte Carlo via cadeia de Markov (MCMC) ou aproximação da integral por uma soma (por exemplo, via quadratura gaussiana). Obviamente, todas essas técnicas vêm com alguns custos computacionais extras.

Os efeitos aleatórios podem ser modelos em dois níveis: observacional ou do fator. Os efeitos aleatórios no nível do fator são os efeitos aleatórios de "grupo" usuais. Nesse caso um ou mais fatores classificam as observações em diferentes

categorias, então $\boldsymbol{Z}$ neste caso será uma matriz de especificação com 1 se a observação pertencer a um dado nível do fator e 0 caso contrário. Ou seja, $\boldsymbol{Z}$ será uma matriz de variáveis do tipo *dummy*. Para o nível observacional, há tantos efeitos aleatórios quanto observações. Neste caso, $\boldsymbol{Z}$ será igual a matriz identidade e $g_k(\boldsymbol{\theta_k}) = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\gamma}_k$.

Existem diferentes formas de implementar modelos com efeitos aleatórios no pacote `gamlss`. A função `gamlssNP` usa estimativa "global" e pode aplicar quadratura gaussiana para aproximar a verossimilhança marginal da variável resposta $\boldsymbol{y}$, substituindo a integração sobre uma variável de efeitos aleatórios normalmente distribuída por um somatório. Isso permite estimar o hiperparâmetro $\boldsymbol{\lambda}$, maximizando a verossimilhança marginal dos dados.

Por outro lado, as funções `random` e `re` usam uma aproximação normal "local" para a verossimilhança, conhecida na literatura como quase verossimilhança penalizada (QVP) (Breslow e Clayton, 1993). Para esses modelos, os valores ajustados do modelo são derivados da função de verossimilhança conjunta, enquanto a inferência para o comportamento da variável resposta é baseada na verossimilhança condicional dado os efeitos aleatórios.

Ao considerar a inclusão de interceptos aleatórios na modelagem de um parâmetro, deve-se assumir que

$$\gamma_j \sim N(0, \sigma_b^2), \tag{9}$$

independentemente para $j = 1,2, \dots, J$. Dessa forma, o modelo GAMLSS com efeitos aleatórios pode ser ajustado com auxílio da técnica de quadratura gaussiana para aproxima a integral apresentada em (8). A quadratura gaussiana é um método de integração numérica em que a integral é aproximada por um somatório. Ela substitui a distribuição contínua $f(\boldsymbol{\gamma})$ por uma distribuição discreta aproximada. Com base em (9), tem que $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_b^2\boldsymbol{I})$. Tomando $U_j \sim N(0,1)$, tem-se que $\gamma_j \sim \sigma_b^2 U_j$ e $\boldsymbol{U} = \sigma_b^{-2}\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{1})$, em que $\boldsymbol{U} = \left(U_1, \dots, U_J\right)^T$. A quadratura gaussiana aproxima a distribuição $N(0,1)$ de cada U_j por uma distribuição discreta:

$$\mathbb{P}\left(U_j = u_\kappa\right) = \pi_\kappa, \kappa = 1, \ldots, \mathrm{K} \quad (10)$$

em que os u_κ e π_κ são conhecidos e fixados por um total fixado de K pontos discretos usados pela aproximação via quadratura gaussiana.

2.5. Inferência

Stasinopoulos *et al*. (2017) destacam dois métodos inferenciais empregados em GAMLSS, o baseado em verossimilhança e o baseado em técnica de *bootstrapping*. Para inferências baseadas em verossimilhança, considere um modelo GAMLSS paramétrico com vetor de parâmetros $\boldsymbol{\theta}$. Sob a ótica da teoria da verossimilhança clássica, pode-se assumir que, assintoticamente

$$\widehat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, i(\boldsymbol{\theta}_T)^{-1}),$$

em que $\widehat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}_T$ e

$$i(\boldsymbol{\theta}_T) = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$$

é a matriz de informação esperada de Fisher para o verdadeiro valor $\boldsymbol{\theta}_T$. Como nem sempre é possível obter $i(\boldsymbol{\theta}_T)$ de modo analítico, utiliza-se, por vezes, a matriz de informação de Fisher observada, definida como

$$I(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right].$$

Deve-se atentar que $I(\boldsymbol{\theta}_T)$ é igual ao valor negativo da matriz Hessiana do logaritmo da função de verossimilhança de $\boldsymbol{\theta}_T$. Dessa forma, pode-se empregar $I(\boldsymbol{\theta}_T)^{-1}$ em vez de $i(\boldsymbol{\theta}_T)^{-1}$ para estimar a matriz de variância-covariâncias de $\widehat{\boldsymbol{\theta}}$.

Obviamente, $\boldsymbol{\theta}_T$ é desconhecido e as quantidades costumeiramente empregadas são $I(\widehat{\boldsymbol{\theta}}_T)^{-1}$ ou $i(\widehat{\boldsymbol{\theta}}_T)^{-1}$. De modo geral, os modelos paramétricos GAMLSS assumem que

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_T, I(\widehat{\boldsymbol{\theta}}_T)^{-1}\right).$$

Se o modelo está incorreto, então a distribuição assintótica de $\widehat{\boldsymbol{\theta}}$ pode ser aproximada, sob certas circunstâncias, pelo estimador do tipo sanduíche de $\boldsymbol{\theta}$ dado por

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_C, I(\widehat{\boldsymbol{\theta}})^{-1} K(\widehat{\boldsymbol{\theta}}) I(\widehat{\boldsymbol{\theta}})^{-1}\right),$$

em que $K(\widehat{\boldsymbol{\theta}})$ é uma estimativa da matriz de variância-covariância da primeira derivada da função de log-verossimilhança com respeito aos parâmetros e $\boldsymbol{\theta}_C$ é o valor $\boldsymbol{\theta}$ mais 'próximo' do verdadeiro modelo medido pela distância de Kullback-Leibler (Kullback e Leibler, 1951).

Usualmente, os erros-padrão são obtidos a partir da raiz quadrada da diagonal da matriz de covariância $I(\widehat{\boldsymbol{\theta}}_T)^{-1}$ (ou $I(\widehat{\boldsymbol{\theta}})^{-1} K(\widehat{\boldsymbol{\theta}}) I(\widehat{\boldsymbol{\theta}})^{-1}$ para estimadores do tipo sanduíche) . Um método alternativo para obtenção dos erros-padrão para uma estimativa qualquer $\hat{\beta}$ quando a matriz de informação observada é de difícil obtenção, é dado por

$$EP(\hat{\beta}) \approx \frac{|\hat{\beta}|}{\sqrt{\Delta GDEV}},$$

em que $\Delta GDEV$ é a diferença entre o desvio global com as variáveis explicativas associadas ao parâmetro de interesse β e desvio global obtido pela omissão dessas variáveis. Para mais detalhes e sobre a inferência baseada em *bootstrapping*, ver Stasinopoulos *et al*. (2017), por exemplo.

2.6. Seleção de modelos

A seleção de modelos em GAMLSS pode ser realizada via i) seleção da distribuição para a variável resposta; ii) seleção da função de ligação; iii) estabelecimento dos melhores preditores; e iv) ajuste dos parâmetros de suavização.

2.6.1. Seleção da distribuição para a variável resposta

A seleção da distribuição pode ser feita via comparação de modelos e análise gráfica. O critério de Akaike generalizado (GAIC) (Akaike, 1983) é comumente empregado na comparação de modelos com distribuições diferentes. Tal critério penaliza o sobreajuste por ser definido como a soma do desvio ajustado mais uma penalidade fixa para cada grau de liberdade efetivo (Hastie e Tibshirani, 1990). A estratégia gráfica consiste em comparar gráficos de distribuições ajustadas em modelos selecionados por apresentar os menores valores de GAIC. Inadequações nesses modelos podem ser detectadas com gráficos denominados de *worm plots*, que são gráficos quantil-quantil (QQ) sem tendência introduzidos por Buuren e Fredriks (2001).

2.6.2. Seleção da função de ligação

A função de ligação adequada depende do suporte dos parâmetros da distribuição da variável de resposta. Por exemplo, para uma variável resposta com distribuição normal com parâmetros μ e σ^2, em que μ assume qualquer valor nos $\mathbb{R}$ e $\sigma^2 > 0$, uma função de ligação identidade para μ e uma logarítmica para σ^2 podem ser escolhas adequadas. No entanto, pode haver situações em que mais de uma função de ligação pode ser adequada para um parâmetro específico. Nesses casos, as diferentes funções devem ser comparadas diretamente usando o desvio global e

gráficos *worm plots*. Tais gráficos serão detalhados no item 2.7 (sobre 'análise de diagnóstico') deste capítulo.

2.6.3. Estabelecimento dos melhores preditores

Dada uma distribuição para a variável resposta com um conjunto de parâmetros μ, σ, ν e τ, os procedimentos *forward*, *backward* e *stepwise* podem ser empregados para encontrar os melhores preditores para um dado parâmetro da distribuição da variável de resposta. Como exemplo, uma estratégia é descrita sequencialmente por Thomas (2017):

i) Usar o procedimento *forward* com base no GAIC para selecionar um modelo apropriado para μ, com σ, ν e τ ajustados como constantes;

ii) Dado o modelo para μ obtido em (i) e para ν e τ ajustados como constantes, deve-se usar o procedimento *forward* para selecionar um modelo apropriado para σ;

iii) Dados os modelos para μ e σ obtidos em (i) e (ii) respectivamente e com τ ajustado como constante, usar o procedimento *forward* para selecionar um modelo apropriado para ν;

iv) Dados os modelos para μ, σ e ν obtidos em (i), (ii) e (iii) respectivamente, usar o procedimento *forward* para selecionar um modelo apropriado para τ;

v) Dados os modelos para μ, σ e τ obtidos em (i), (ii) e (iv) respectivamente, usar o procedimento *backward* de seleção para selecionar o modelo apropriado para ν;

vi) Dados os modelos para μ, ν e τ obtidos em (i), (v) e (iv) respectivamente, usar o procedimento *backward* para selecionar o modelo apropriado para σ;

vii) Dados os modelos para σ, ν e τ obtidos em (vi), (v) e (iv) respectivamente, usar o procedimento *backward* para selecionar um modelo apropriado para μ e parar.

Deve-se atentar que o modelo final poderá possuir diferentes variáveis explicativas para cada um dos parâmetros, assim como um ou mais parâmetros podem ser constantes para um dado conjunto de dados.

2.6.4. Ajuste dos parâmetros de suavização

Os parâmetros de suavização podem ser estimados a partir dos dados ou fixados. Ao se fixar os parâmetros de suavização, um procedimento comum é fixar os graus de liberdade efetivos para a suavização (Hastie e Tibshirani, 1990). Como já apresentado anteriormente, os parâmetros de suavização são estimados pelos métodos VCG, GAIC ou ML. Stasinopoulos *et al*. (2017) trazem detalhes sobre tais métodos.

2.7. Análise de diagnóstico

A análise de resíduos é uma das ferramentas mais aplicadas para se analisar a qualidade de ajuste dos modelos. Uma das formas mais elementares de realizar essa análise é através do emprego de resíduos ordinários (ou brutos), que são dados pela diferença entre os valores observados e estimados

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

Obviamente, existem formas mais elaboradas para analisar a qualidade de ajuste. Por exemplo, os resíduos *studentizados* são comumente empregados em MRL. Eles são dados por

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad (11)$$

em que h_{ii} são os valores da diagonal da matriz 'chapéu' $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, enquanto modelos MLGs empregam os resíduos componente do desvio (*deviance*) e de *Pearson*, dados respectivamente por

$$r_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i} \quad (12)$$

$$r_i^P = (y_i - \hat{\mu}_i)\sqrt{V(\hat{\mu}_i)} \quad (13)$$

em que $d_i = -2\log\left(\frac{L_i^c}{L_i^s}\right)$, com L_i^c representando a verossimilhança ajustada da observação i do modelo corrente e L_i^s do modelo completo (saturado) e $V(.)$ é a função de variância dos MLGs.

Contudo, esses resíduos apresentam algumas limitações. Os resíduos ordinários apresentam dificuldades de serem generalizados para distribuições outras além da normal. Os resíduos componentes do desvio não são, em geral, bem definidos (por exemplo, ao modelar vários parâmetros para a distribuição da variável resposta), enquanto os resíduos de *Pearson* não apresentam normalidade para variáveis resposta altamente assimétricas ou 'achatadas'.

Para contornar essas limitações, os modelos GAMLSS usam os resíduos quantílicos normalizados, introduzidos por Dunn e Smyth (1996). Os resíduos quantílicos normalizados podem ser definidos por

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i), \quad (14)$$

em que $\Phi^{-1}(.)$ é a inversa da função de distribuição acumulada da distribuição normal padrão e $\hat{u}_i$ são os resíduos quantílicos. Se y é uma observação de uma variável contínua Y então $u = F(y|\boldsymbol{\theta})$ e $\hat{u} = F\left(y|\widehat{\boldsymbol{\theta}}\right)$ são os valores da função na distribuição acumulada do modelo e do ajuste, respectivamente. Se U é uma variável aleatória

definida como $U = F_Y(Y) = P(Y \leq y)$, então U tem função de distribuição acumulada dada por

$$F_U(u) = P(U \leq u) = P(F_Y(Y) \leq u) = P\left(Y \leq F_Y^{-1}(u)\right) = F_Y\left(F_Y^{-1}(u)\right) = u,$$

de modo que

$$F_U(u) = \begin{cases} 0 \text{ se } u < 0 \\ u \text{ se } 0 \leq u \leq 1 \\ 1 \text{ se } u > 1 \end{cases},$$

que é a função de distribuição acumulada de uma distribuição uniforme padrão. Esse procedimento é denominado de transformada integral de probabilidade e ele garante que, se o modelo for corretamente especificado, os resíduos quantílicos u terão distribuição uniforme padrão e os resíduos quantílicos normalizados r terão distribuição normal padrão.

Se a Y é variável resposta discreta, então $u = F(y|\boldsymbol{\theta})$ é uma função degrau, o que implica que para uma dada observação y existe um intervalo de valores $[a, b]$ em $F(y|\boldsymbol{\theta})$. Uma forma de lidar com essa situação é definir, respectivamente, u e $\hat{u}$ como um valor aleatório de uma distribuição uniforme nos intervalos

$$[u_1, u_2] = [F(y-1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})],$$
$$[\hat{u}_1, \hat{u}_2] = \left[F\left(y-1|\widehat{\boldsymbol{\theta}}\right), F\left(y|\widehat{\boldsymbol{\theta}}\right)\right].$$

Dada uma função de probabilidade, uma observação y é transformada em um intervalo $[u_1, u_2]$ e u é selecionado de uma distribuição uniforme com intervalo $[u_1, u_2]$ e transformado em resíduos. Dessa forma, para variáveis discretas, u são denominados de resíduos quantílicos aleatorizados.

Uma vez obtidos os resíduos, comumente cinco tipos de gráficos são empregados para analisar a qualidade do ajuste nos modelos GAMLSS. São eles:

i) Gráfico dos resíduos quantílicos normalizados contra os valores ajustados do parâmetro μ;

ii) Gráfico dos resíduos quantílicos normalizados contra a ordem das observações ou uma covariável;

iii) Gráfico da densidade kernel estimada dos resíduos quantílicos normalizados;

iv) Gráfico QQ dos resíduos quantílicos normalizados; e

v) *Worm plot*.

Se o modelo é ajustado adequadamente, espera-se que os gráficos (i) e (ii) apresentem pontos dispostos aleatoriamente ao redor do valor zero, sem muitas observações acima de 3 ou abaixo de -3. Nos gráficos (iii) e (iv) espera-se que os pontos formem uma curva em forma de sino e uma linha reta, respectivamente. O *worm plot* é, sem dúvida, a principal técnica empregada para analisar os resíduos de modelos GAMLSS. Eles foram propostos por Buuren e Fredriks (2001) a fim de identificar regiões (intervalos) de uma variável explicativa dentro da qual o modelo não se ajusta adequadamente aos dados (que eles chamaram de "violação do modelo"). Esses gráficos são um tipo de gráfico normal quantis-quantis sem tendência e é esperado que os resíduos se distribuam de modo linear ao redor do zero caso o modelo esteja bem ajustado. Outros padrões da disposição dos resíduos podem indicar um ajuste inadequado, como apontado na Figura 2 e Tabela 1.

Figura 2. Diferentes problemas (situações a) a h) da Tabela 1) que podem ser diagnosticados pelo *worm plot*. Adaptado de Stasinopoulos *et al*. (2017).

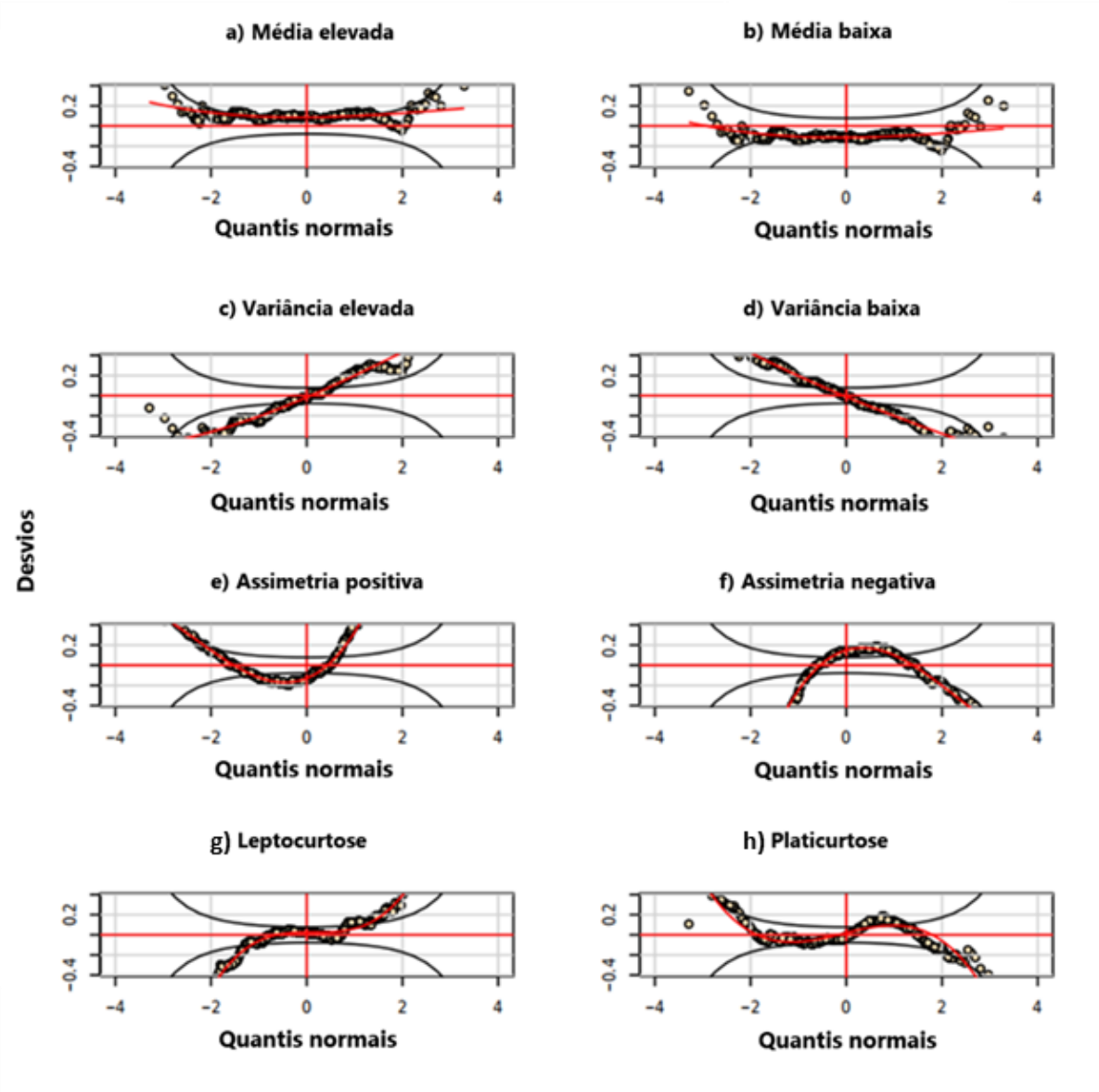


Tabela 1. Formatos distintos que o *worm plot* pode assumir e suas interpretações. Adaptado de Stasinopoulos (2017).

Situação	Formato	Resíduos	Parâmetro ajustado
a)	Pontos acima da origem	Média muito alta	Localização subestimada
b)	Pontos abaixo da origem	Média muito baixa	Localização superestimada
c)	Inclinação positiva	Variância muito alta	Escala subestimada
d)	Inclinação negativa	Variância muito baixa	Escala superestimada
e)	U	Assimetria positiva	Assimetria subestimada
f)	U invertido	Assimetria negativa	Assimetria superestimada
g)	S com curva esquerda para baixo	Leptocurtose	Cauda muito leve
h)	S com curva esquerda para cima	Platicurtose	Cauda muito pesada

3. Popularização do modelo GAMLSS

Foi realizada uma pesquisa no *Google Acadêmico* e na plataforma *ScienceDirect* para analisar a popularização dos modelos GAMLSS frente aos modelos MLGs e MAGs. A pesquisa foi realizada considerando os últimos 11 anos (2010 a 2021) e empregando os termos "Generalized Linear Model", "Generalized Additive Model" e "GAMLSS" para verificar o número de resultados encontrados para os modelos MLGs, MAGs e GAMLSS, respectivamente. Analogamente, foram empregados os termos "Generalized Linear Mixed Model", "Generalized Additive Mixed Model" e "Mixed GAMLSS" para verificar o número de resultados encontrados para os modelos mistos MLG, MAG e GAMLSS, respectivamente.

O *Google Acadêmico* é uma ferramenta de pesquisa do *Google* que permite pesquisar em trabalhos acadêmicos, literatura escolar, jornais de universidades e artigos variados, enquanto a *ScienceDirect* é uma plataforma para acesso de aproximadamente 2500 revistas científicas e mais de 26000 e-books. Dessa forma,

acredita-se que o número de resultados retornados em cada pesquisa representa, em algum grau, a popularização do tipo de modelo pesquisado (ou o quão empregado o modelo é).

Os resultados da pesquisa estão na Tabela 2. Percebe-se que quanto mais geral o modelo, menos popular ou empregado é. Percebe-se também que esse cenário é agravado quando se consideram os modelos mistos. Os MLGs são, de longe, a classe que apresenta o maior número de resultados retornados, seguido pelos MAGs. Por outro lado, o GAMLSS ainda tem um uso inconspícuo e, pela pesquisa, nenhum artigo ainda foi publicado em periódicos do catálogo da *ScienceDirect* para modelos mistos no período analisado. Mesmo os três resultados encontrados no *Google Acadêmico* são uma dissertação (Thomas, 2017) e dois *preprints* no arXiv (Thomas *et al*., 2018; Sodja, 2020) ainda não publicados em revistas com avaliações por pares. Isso claramente mostra que os modelos GAMLSS são poucos utilizados.

Tabela 2. Resultado da pesquisa pelos modelos MLG, MAG e GAMLSS nas plataformas *Google acadêmico* e *ScienceDirect*. Os valores em parênteses é o percentual em relação ao total por linha. Detalhes no texto.

Misto	Plataforma	Modelo		
		MLG	MAG	GAMLSS
Não	*Google acadêmico*	48300 (68,46%)	17700 (25,09%)	4550 (6,45%)
	ScienceDirect	24922 (78,71%)	6312 (19,94%)	428 (1,35%)
Sim	*Google acadêmico*	28750 (92,64%)	2280 (7,35%)	3 (0,01%)
	ScienceDirect	10533 (92,93%)	801 (7,07%)	0 (0%)

4. Considerações finais do capítulo

Teoricamente, os GAMLSS são mais capazes de lidar com dados reais quando comparados aos demais modelos aqui mencionados. Isso ocorre porque ele permite empregar qualquer distribuição (ou mistura de distribuições) no modelo de regressão, além de permitir explicitamente a modelagem de todos os parâmetros da distribuição da variável resposta. Contudo, a análise cienciométrica sugere que tal classe de modelos tem sido subutilizada, destacadamente para modelos mistos. Isso provavelmente se deve a três principais fatores: (i) dificuldades inerentes de entendimento de modelos mais complexos; (ii) modelo muito recente e ainda desconhecido por parte da população científica; e (iii) grande exigência computacional para executar os algoritmos propostos.

O GAMLSS possui forte arcabouço teórico para analisar qualidade de ajuste de seus modelos com base nos resíduos quantílicos. Entretanto, embora exista proposta de análises de influência para estudos transversais (Silva, 2021), não foram encontradas (até o presente momento) propostas para modelos mistos desta classe, o que é uma limitação.

CAPÍTULO 2

APLICAÇÃO DO MODELO GAMLSS A DADOS LONGITUDINAIS

1. Introdução

O pólen é uma estrutura reprodutiva das plantas produtoras de sementes. Ele contém o gameta masculino, que quando encontra o gameta feminino, pode ocorrer a fecundação e a consequente formação de semente. Contudo, na perspectiva das abelhas, o pólen é a principal fonte de proteínas e aminoácidos, sendo também uma importante fonte de lipídios, fibras, enzimas, minerais, açúcares e vitaminas (Arruda *et al*., 2013; Avni *et al*., 2014; Sattler *et al*., 2015). Essa composição nutricional torna o pólen essencial para a alimentação da ninhada e manutenção das colônias das abelhas sociais, tais como a *Apis mellifera* (Marchini *et al*., 2006).

O pólen também é de crescente importância para os seres humanos, visto que muitas comunidades em diversos países dependem da produção de pólen apícola para complementar ou gerar toda a renda familiar. No Brasil, é sabido que a maior parte da produção apícola se localiza no semiárido (Barreto *et al*., 2006), onde a fisionomia vegetal varia de desertos com vegetação esparsa a áreas com florestas secas cobertas por densas camadas de árvores (Araujo-Filho, 2013). Nesses ambientes, a água é o principal recurso limitante, de modo que a maior número de espécies e densidade de plantas em floração ocorre no período úmido. Dessa forma, espera-se que a produção de pólen também seja maior nesse período.

Indubitavelmente, a produção de pólen está aquém do seu potencial. Isso se deve, em parte, a duas limitações. Primeiro, comumente há desconhecimento da identidade da flora apícola da região (Milfont *et al*., 2011). Segundo, há desconhecimento da disponibilidade temporal de fontes de pólen para as abelhas ao longo do ano (Nascimento *et al*., 2019). Qualquer estudo que tente mitigar essa segunda limitação, terá que trabalhar com a dimensão temporal. Isso poderá implicar em estudos de dados longitudinais que conhecidamente requerem maior conhecimento estatístico e computacional por parte dos pesquisadores.

Dessa forma, este estudo objetiva apresentar uma aplicação do modelo misto GAMLSS para modelar dados longitudinais de produção de pólen.

Adicionalmente, os resultados obtidos com o GAMLSS são contrastados com os resultados da modelagem por MLM para verificar qual dessas técnicas apresenta melhor ajuste aos dados.

2. Materiais e métodos

2.1. Sobre os dados de produção de pólen

Os dados aqui tratados foram originalmente publicados em Nascimento *et al*. (2019). Eles consistem na produção de pólen da abelha *A. mellifera* em dez colmeias do tipo Langstroth. As coletas foram realizadas de novembro de 2012 a outubro de 2013 em uma floresta semidecídua localizada no município de Meruoca (3°35'40.63" S e 40°24'11.91" W), estado do Ceará, Brasil.

As amostras de pólen foram obtidas instalando coletores de pólen nas entradas das colmeias, padronizadas quanto ao número de abelhas (28 mil ± 2 mil), número de favos (n = 10) e idade da rainha (um ano). Os coletores de pólen exigem que as abelhas entrem na colmeia através de pequenos orifícios que raspam o pólen das corbículas da abelha. Essas corbículas são estruturas localizadas nas pernas que funcionam como cestas de pólen e auxiliam no transporte deste recurso ao ninho. A coleta de pólen foi realizada em dias alternados, sempre às 17h30, totalizando 15 coletas mensais para cada colônia. As amostras de pólen frescas foram limpas (removendo sujidades, tais como abelhas mortas, larvas de abelha e própolis) e armazenas adequadamente, para posterior pesagem.

Ademais, foram contabilizados (*i*) o número de espécies vegetais em floração por mês em um raio de 1 km das colmeias e (*ii*) a precipitação total por mês na região de estudo. Dessa forma, os dados utilizados nas análises são de frequência mensal e separados por colmeia. Maiores detalhes podem ser obtidos em Nascimento *et al*. (2019).

2.2. Análises estatísticas

Inicialmente foi realizada uma análise descritiva dos dados. Após tal análise, foram empregados os modelos GAMLSS misto e MLM, tomando a produção polínica como variável resposta (y) e os meses após a instalação das colmeias, o número de espécies vegetais em floração por mês e a precipitação total por mês na região de estudo como variáveis explicativas. O MLM foi empregado tomando a resposta como y (sem transformação) e como $\log(y+1)$ (com transformação logarítmica, ou log-transformação). Aqui, utilizou-se a média das 15 coletas mensais realizadas em cada colmeia.

Para analisar a qualidade de ajuste dos modelos aos dados, foram realizadas análises de resíduos (de Pearson para os MLMs e quantílicos para o GAMLSS) e verificado o atendimento aos pressupostos de cada modelo. Os melhores modelos de cada classe foram selecionados de acordo com GAIC e qualidade de ajuste.

Finalmente os coeficientes foram estimados, bem como seus intervalos de confiança. Para o MLM com transformação, foi aplicado o método delta para se obter a variância dos coeficientes estimados e depois calculado o intervalo de confiança assintótico. Todas as análises foram realizadas no ambiente estatístico R (R Core Team, 2021), com auxílio dos pacotes `gamlss` (Rigby e Stasinopoulos, 2005) e `nlme` (Pinheiro *et al*. 2021).

3. Resultados e discussão

3.1. Análise descritiva

A produção de pólen oscilou consideravelmente entre os meses de estudo (Tabela 3). Março e abril foram os meses que apresentaram maior produção de pólen (tanto em média, quanto em mediana), o que coincide com o período de maior precipitação na região de estudo. Por outro lado, agosto e setembro apresentaram

menor produção. Nota-se também nítidas diferenças na magnitude da variância na produção mensal de pólen, com abril apresentando maior valor (Tabela 3).

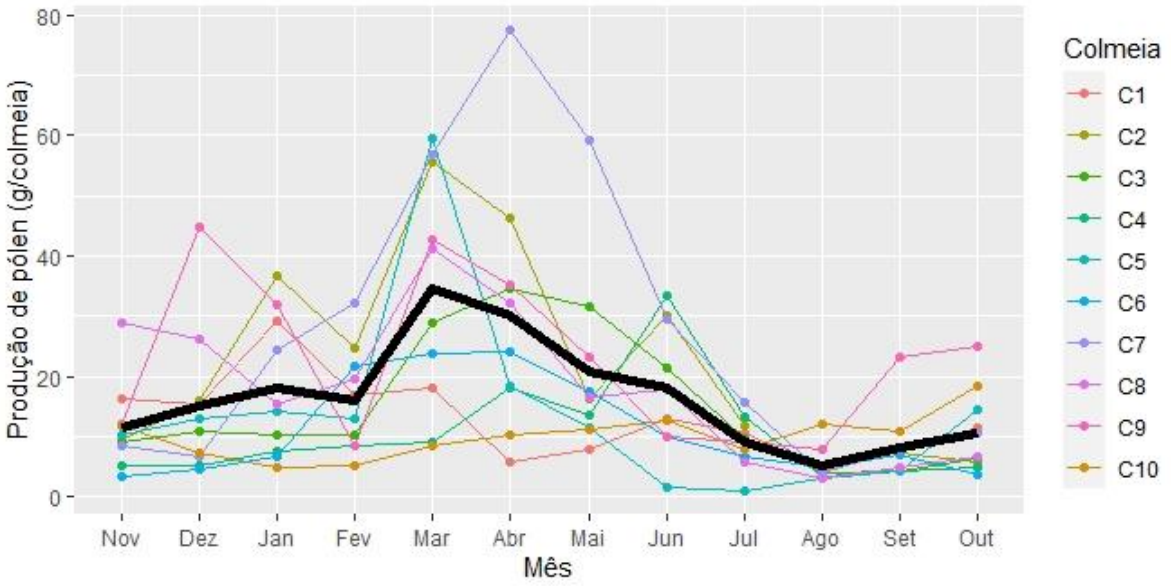
Tabela 3. Medidas resumo da produção de pólen apícola (em g/colmeia) das 10 colmeias de *Apis mellifera* estudadas por Nascimento *et al*. (2019).

Mês	Média	Mínimo	Mediana	Máximo	DP	CV (%)
Novembro	11,52	3,54	9,82	29,02	7,12	61,78
Dezembro	15,03	4,57	11,92	44,94	12,39	82,40
Janeiro	18,16	4,94	14,75	36,86	11,58	63,73
Fevereiro	16,07	5,28	14,99	32,19	8,52	53,03
Março	34,51	8,66	35,16	59,64	19,49	56,48
Abril	30,27	5,79	28,17	77,68	20,74	68,50
Maio	20,89	7,87	16,48	59,33	15,12	72,35
Junho	18,00	1,68	15,52	33,49	10,41	57,83
Julho	9,14	0,95	9,41	15,68	4,13	45,22
Agosto	5,17	2,99	3,88	12,05	2,99	57,83
Setembro	8,11	4,16	7,03	23,13	5,67	69,94
Outubro	10,71	3,79	8,60	25,07	6,86	64,02

DP = Desvio padrão; CV = Coeficiente de variação

Analisando os perfis individuais (Figura 3), é possível notar que as colmeias não apresentam um padrão nítido quanto a produção polínica. Algumas delas apresentam grande variabilidade (como a C7), enquanto outras são bem menos variáveis (por exemplo, a C10). Em termos médios, a produção polínica é crescente de novembro a março e depois declina, com valores mínimos em agosto. Esse é o padrão esperado, dado que a abundância das espécies florais (fontes de pólen) segue o volume de precipitação.

Figura 3. Perfis individuais (linhas coloridas com pontos) e médio (linha preta) da produção de pólen das colmeias ao longo de doze meses de observação.



A quantidade de pólen coletada pelas abelhas em um dado dia é dependente da quantidade já coletada. Dessa forma, espera-se que os dados apresentem autocorrelação. A observação da Tabela 4 sugere que as correlações (e as covariâncias) são ligeiramente maiores (em valores absolutos) entre meses adjacentes.

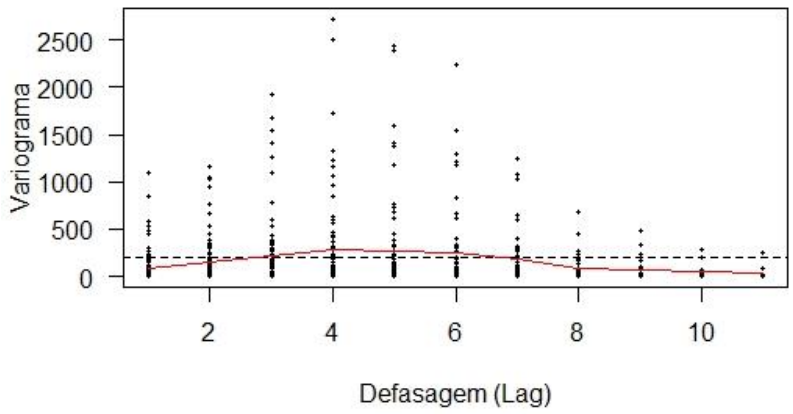
Na figura 4 é apresentado o variograma amostral da produção de pólen. Nele, os pontos são os valores de $v_{ijk} = 0{,}5 \times \left(r_{ij} - r_{ik}\right)^2$, com r_{ij} sendo o resíduo (a partir de um ajuste de mínimos quadrados) da i-ésima colmeia ($i = 1, \ldots, 10$) e da j-ésima observação ($j = 1, \ldots, 12$). A defasagem (eixo x da figura) é dada por $t_{ij} - t_{ik}$, em que t_{ij} é o tempo correspondente a i-ésima colmeia na j-ésima observação. A linha preta pontilhada é a estimativa da variância total dos dados ($s^2 = 200{,}12$) e a linha vermelha representa a curva média dos v_{ijk}. Pode-se perceber nessa figura uma curva crescente com valor máximo para uma defasagem de 4 a 6 meses. Depois disso a curva é decrescente. Isso novamente sugere maior dependência entre meses mais próximos.

Os resultados da análise descritiva em conjunto sugerem que a estrutura funcional para modelar a produção de pólen ao longo dos meses deve ser capaz de lidar com lidar com heteroscedasticidade e autocorrelação.

Tabela 4. Variâncias (diagonal), correlações de Pearson (acima da diagonal) e covariâncias (abaixo da diagonal) da produção de pólen.

Mês	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out
Nov	56,41	0,51	0,30	0,07	0,18	-0,07	-0,20	-0,14	-0,26	-0,09	0,01	0,10
Dez	50,53	172,48	0,66	-0,21	0,33	0,05	-0,09	-0,32	-0,18	0,15	0,74	0,63
Jan	22,67	87,24	102,20	0,29	0,44	0,31	0,27	-0,10	0,26	-0,10	0,56	0,50
Fev	4,23	-23,47	24,85	71,17	0,51	0,69	0,64	0,23	0,30	-0,50	-0,25	-0,38
Mar	25,98	83,17	86,03	81,91	365,75	0,63	0,51	-0,23	-0,24	-0,39	0,08	0,22
Abr	-11,02	14,16	67,10	122,65	256,44	448,00	0,97	0,46	0,49	-0,25	0,10	-0,04
Mai	-24,50	-17,91	44,14	85,76	156,46	326,35	254,30	0,51	0,60	-0,19	0,06	-0,06
Jun	-10,81	-42,77	-9,81	19,24	-44,73	97,35	81,60	101,67	0,84	-0,27	-0,27	-0,46
Jul	-8,37	-10,10	11,29	10,65	-19,44	44,59	40,62	36,28	18,19	-0,02	0,10	-0,15
Ago	-1,92	5,76	-3,12	-12,55	-22,44	-15,66	-8,85	-8,04	-0,31	8,94	0,62	0,67
Set	0,37	58,58	34,02	-12,79	8,81	12,37	5,79	-16,27	2,61	11,08	36,06	0,83
Out	5,27	58,05	35,79	-22,54	29,44	-6,23	-6,78	-32,91	-4,51	14,01	35,10	49,40

Figura 4. Variograma amostral da produção de pólen.



3.2. Ajuste do modelo via gamlss

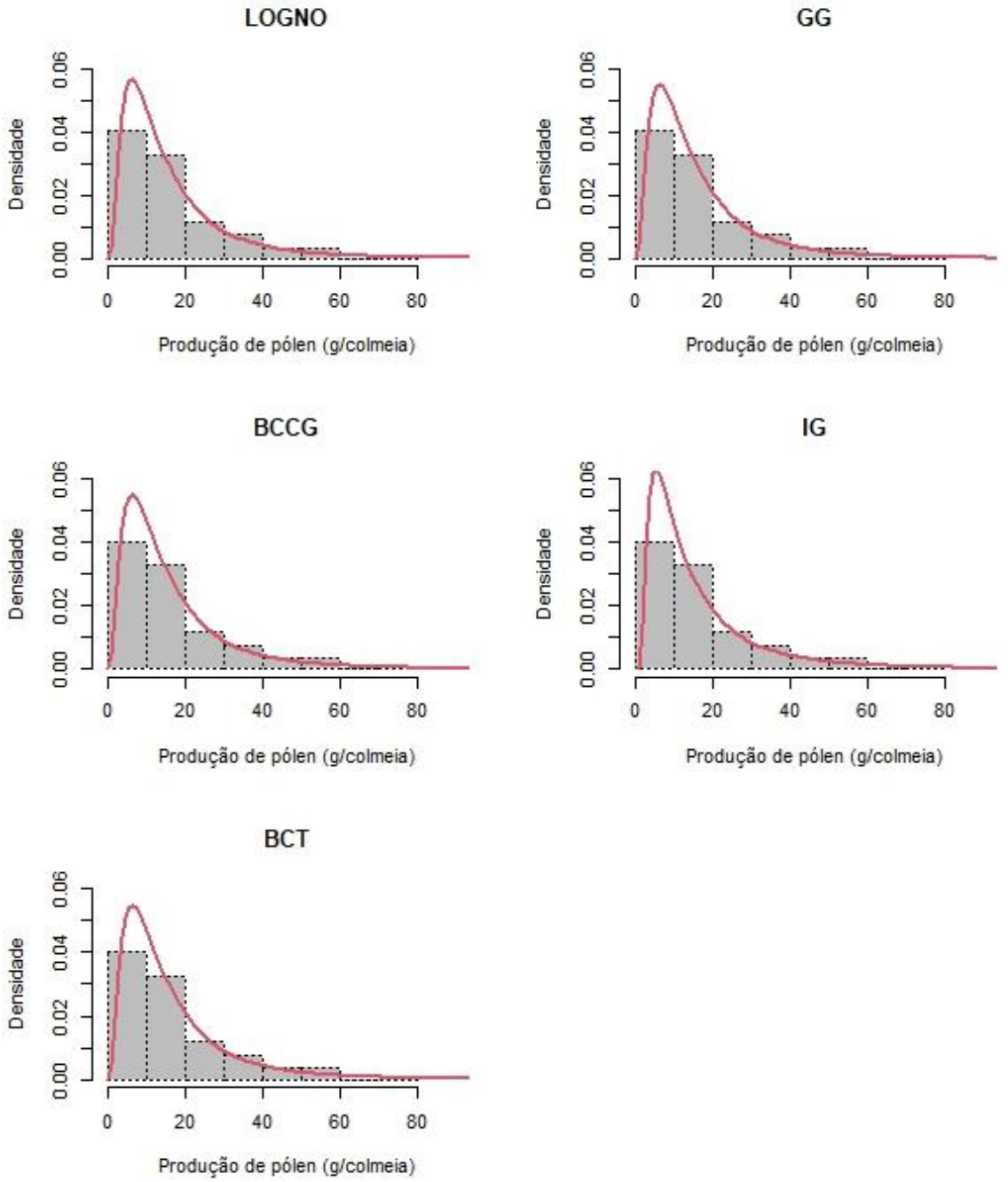
O pacote `gamlss` (Rigby e Stasinopoulos, 2005) possui algumas funções para seleção da distribuição de probabilidade da variável resposta (tais como `fitDist`, `histDist` e `gamlssML`). A função `fitDist` ajusta todas as distribuições paramétricas relevantes presentes no pacote `gamlss` para um único vetor de dados (sem variáveis explicativas). As distribuições marginais sugeridas são aquelas com menor GAIC, dada uma penalidade k. Assim, pode-se ter indícios de quais são as distribuições de probabilidade mais adequadas para a variável resposta.

Tabela 5. Resultado da seleção da distribuição de probabilidade para a variável resposta (Produção de pólen) de acordo com a função `fitDist` do pacote `gamlss`.

Distribuição		GAIC($k = 2$)
Sigla em `gamlss`	Nome	
LOGNO	Log-Normal	883,73
GG	Gama generalizada	885,56
BCCG	Box-Cox Cole e Green	885,57
IG	Normal inversa	886,86
BCT	Box-Cox-t	887,57

Destaca-se que, como a variável resposta y de interesse (Produção de pólen) assume apenas valores não negativos, considerou-se apenas distribuições com suporte em $\mathbb{R}^+$. Na Tabela 5 são apresentadas as cinco distribuições com menor GAIC (com $k = 2$) e na Figura 5 as curvas ajustadas sobre o histograma da variável resposta.

Figura 5. Curvas ajustadas das distribuições consideradas na Tabela 5 sobre o histograma da variável $y =$ Produção de pólen.



Os valores de GAIC para as diferentes distribuições da Tabela 5 são semelhantes, e uma análise da Figura 5 sugere que elas apresentam um razoável ajuste à variável resposta. Embora o modelo de regressão seja baseado na distribuição condicional $y|X$, essas distribuições e outras foram empregadas para modelar a

produção de pólen tomando três variáveis explicativas de efeito fixo: (*i*) mês, (*ii*) precipitação total mensal e (*iii*) número de espécies em floração em um dado mês.

Inicialmente foi ajustado um modelo para o parâmetro μ empregando as variáveis explicativas supracitadas e utilizando as colmeias como efeito aleatório. Contudo, considerando o gráfico de perfil da variável resposta, os termos quadráticos e cúbico da variável 'mês' também foram incluídos. Como o termo cúbico foi altamente significativo, foi ajustado um modelo quártico. Analogamente, como o termo de quarto grau foi significativo, utilizou-se um termo de quinto grau, mas este não foi significativo ao nível de significância de 5%. Portanto, os modelos ajustados seguiram um polinômio de quarto grau.

Os parâmetros σ e ν foram modelados em função da variável 'mês', enquanto o τ foi considerado constante. Destaca-se que a presença desses parâmetros depende da distribuição empregada (Tabela 6). Os valores de GAIC e os *worm plots* para esses modelos estão dispostos na Tabela 6 e Figura 6, respectivamente. Destaca-se que somente os ajustes baseados nas distribuições apresentadas na Tabela 5 e Figura 5 são aqui apresentados, pois foram os que apresentaram melhor ajuste aos dados.

Pode-se observar que o modelo com pior ajuste foi o IG, pois além de apresentar maior valor de GAIC, os pontos (resíduos normalizados) e seu ajuste cúbico (linha vermelha) não se dispõem como uma linha horizontal em torno do zero. O mesmo ocorre para o modelo com distribuição LOGNO, embora este apresente menor valor de GAIC. Os demais modelos apresentaram o padrão esperado de um bom ajuste de acordo com o *worm plot* (Tabela 6 e Figura 6).

Tabela 6. Critério de informação de Akaike generalizado (GAIC) e graus de liberdade (gl) para os modelos ajustados com as distribuições de probabilidade LOGNO, GG, BCCG, IG e BCT para a variável resposta 'produção de pólen'.

Distribuição	Parâmetros	gl	GAIC
LOGNO	μ, σ	8,109	811,078
BCCG	μ, σ, ν	7,428	818,194
GG	μ, σ, ν	7,476	818,257
BCT	μ, σ, ν, τ	8,428	820,194
IG	μ, σ	8,045	830,769

Observando com mais detalhes esses modelos restantes, pode-se notar pelos resíduos quantílicos contra os valores ajustados que a heterogeneidade da variância foi bem capturada. Pela análise dos resíduos quantílicos contra a variável explicativa 'mês', percebe-se que a dispersão dos resíduos foi aproximadamente homogênea ao longo do tempo (com exceção do nono mês de observação, julho). Ademais, os erros apresentaram distribuição aproximadamente normal, como se pode perceber pelo gráfico quantil-quantil (Figura 7). Ressalta-se que os modelos com distribuição marginal BCCG e BCT utilizam o parâmetro μ como mediana. Como se deseja comparar a aplicação do GAMLSS com o MLM, optou-se por utilizar a distribuição marginal GG, pois toma μ como a média, além de apresentar excelente ajuste aos dados.

Figura 6. *Worm plots* para os modelos ajustados com as distribuições marginais LOGNO, GG, BCCG, IG e BCT. Existem pontos (resíduos) omitidos da distribuição IG por possuir valores de desvios menores que -1,0.

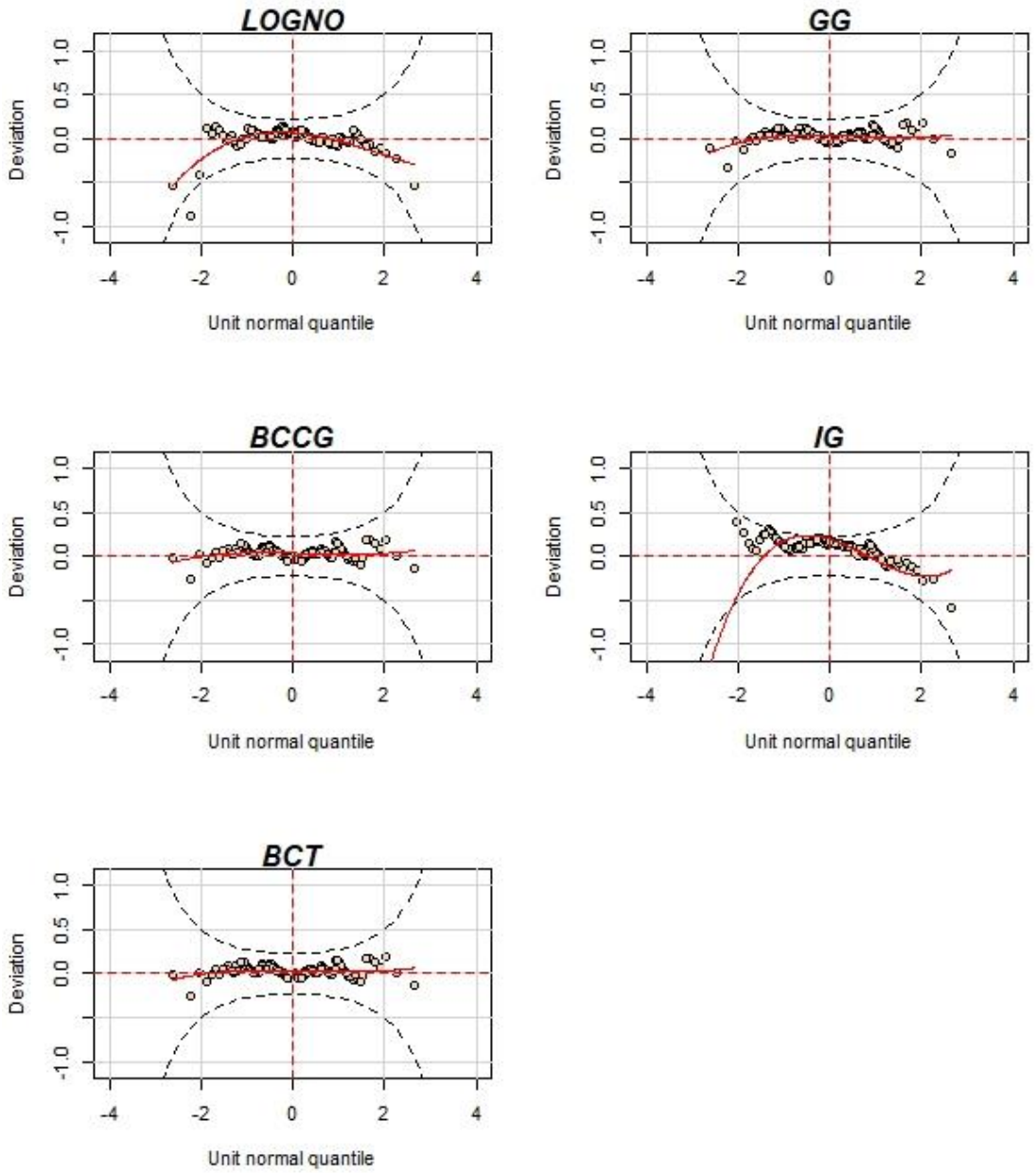
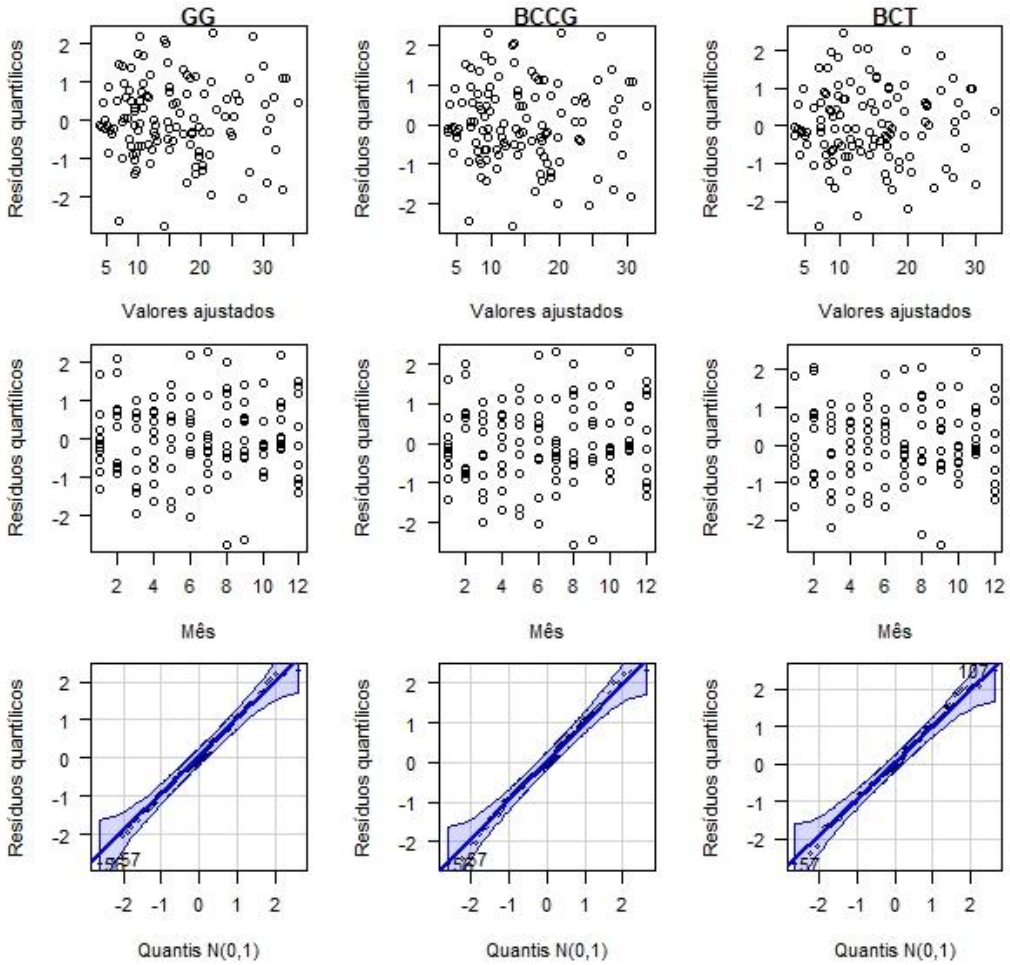


Figura 7. Análise dos resíduos dos modelos ajustados com as distribuições marginais GG, BCCG e BCT.



A distribuição gama generalizada (GG) é adequada para lidar com dados não negativos que podem apresentar tanto assimetria positiva quanto negativa. Sua densidade pode ser dada por

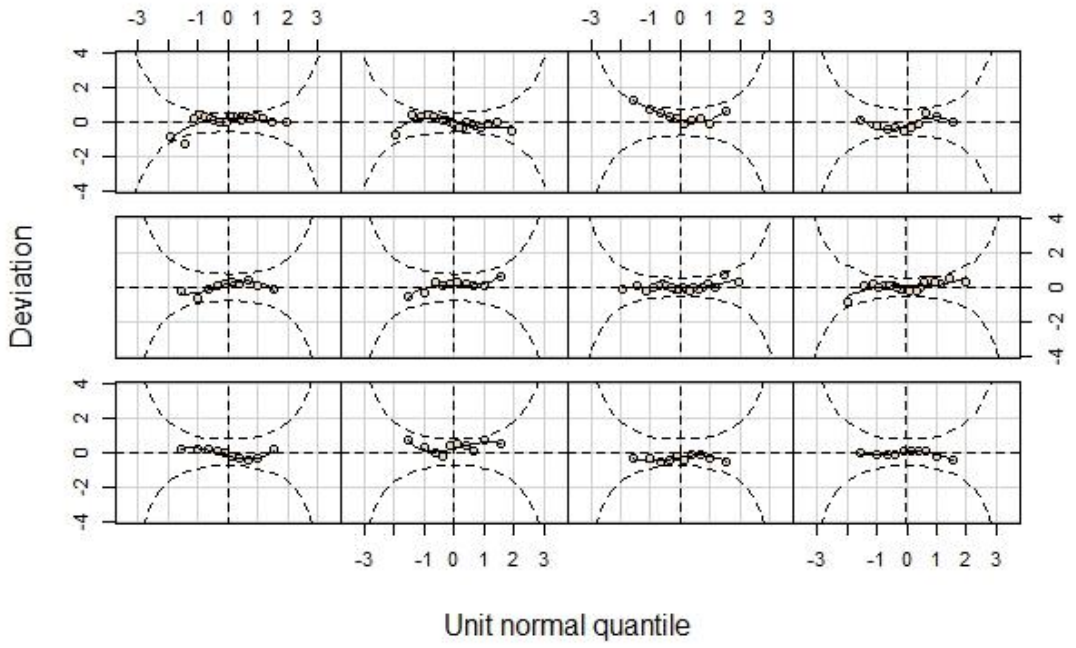
$$f(y|\mu,\sigma,\nu) = \frac{|\nu|\theta^{\theta} z^{\theta} e^{-\theta z}}{\Gamma(\theta) y} \mathbb{I}_{y>0},$$

em que $z = (y/\mu)^{\nu}$ e $\theta = 1/(\sigma^2\nu^2)$ para $\mu > 0, \sigma > 0$ e $-\infty < \nu < \infty$. Essa parametrização foi proposta por Lopatatzidis e Green (*apud* Stasinopoulos *et al.* 2017)

e a relação dela com $f(y|\mu,\sigma,\nu) = \frac{|\nu|y^{\nu\theta-1}e\text{^}(-(y/\beta)^{\nu})}{\beta^{\nu\theta}\Gamma(\theta)}$ (forma funcional mais comumente empregada) com $\beta = \mu(\sigma^2\nu^2)^{\frac{1}{\nu}}$ é apresentada no Anexo I.

A distribuição GG permite modelar simultaneamente 3 parâmetros da distribuição condicional da variável resposta: um parâmetro de localização μ, um parâmetro de escala σ e um parâmetro de forma ν (que controla a assimetria) (Stasinopoulos *et al*. 2017).

Figura 8. *Worm plot* segmentado por mês para o modelo ajustado com a distribuição marginal GG. Os gráficos são lidos em linhas da parte inferior esquerda (mês de novembro) à parte superior direita (outubro).

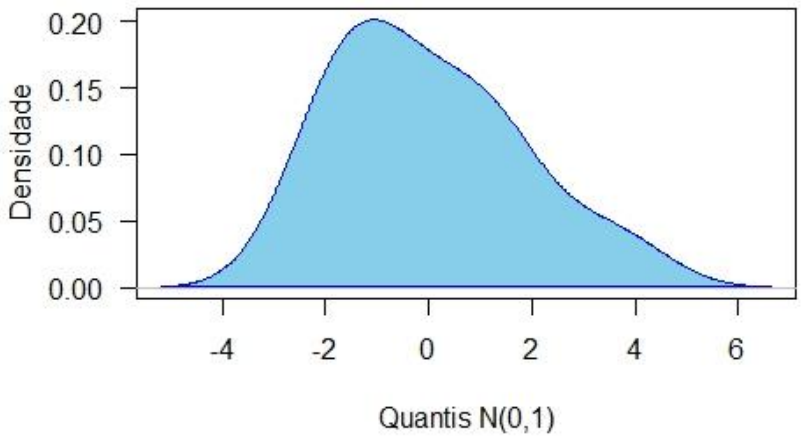


O *worm plot* pode ser utilizado para analisar o ajuste do modelo ao longo de uma variável de interesse. Para se compreender melhor a qualidade de ajuste do modelo selecionado, foram construídos *worm plots* para cada mês de estudo (Figura 8). Pode-se perceber, para alguns meses, que os pontos (e a curva de ajuste) se distanciam de uma disposição em reta horizontal centrada no zero. Contudo, ressalta-se que quase todos os resíduos (pontos) estão delimitados entre os limites dos

intervalos de confiança (representados pelas linhas curvadas pontilhadas). Ademais, como destacado por Stasinopoulos *et al*. (2017), é praticamente impossível construir um modelo sem áreas de desajustes para as covariáveis incluídas.

Quando se analisa os efeitos aleatórios preditos pelo modelo com distribuição marginal GG, vê-se que eles apresentam uma distribuição aproximadamente normal (Figura 9). Essa aproximação é tal que o teste Shapiro-Wilks não foi capaz de rejeitar a hipótese nula de normalidade (W = 0,932, p = 0,464).

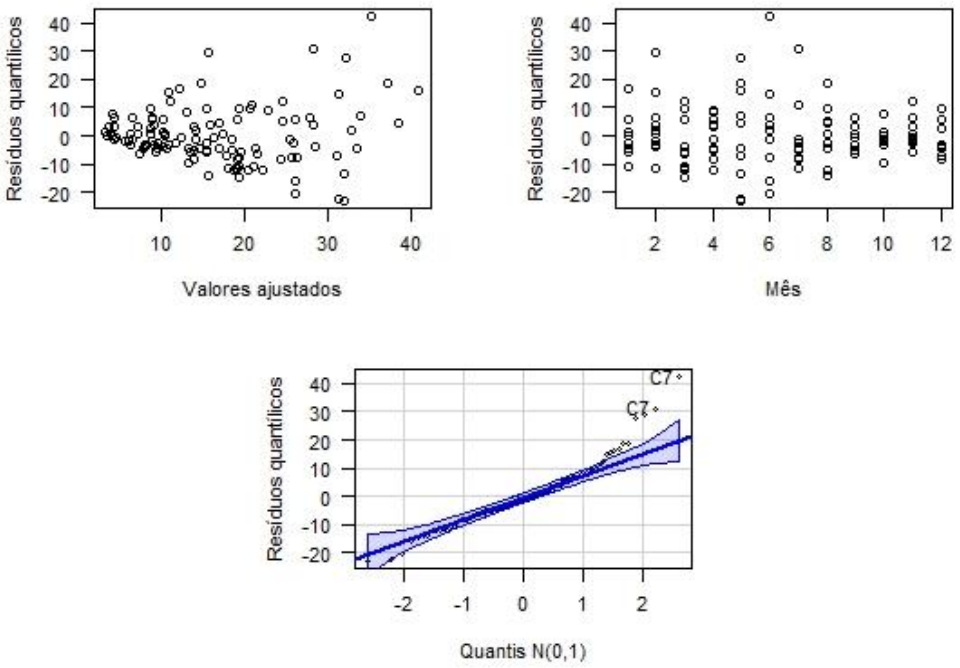
Figura 9. Efeitos aleatórios estimados com o modelo misto GG.



3.2. Ajuste do modelo via `nlme`

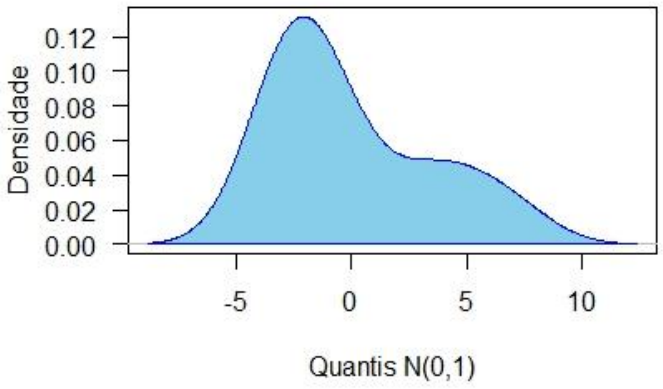
Uma forma comum de ajustar modelos mistos é pela utilização dos pacotes `nlme` (Pinheiro *et al*. 2021) e `lme4` (Bates *et al*., 2015) implementados no software R. Aqui os dados referentes a produção de pólen (g/colmeia) foram analisados com o pacote `nlme` para comparar as estimativas geradas pelo `gamlss`. O modelo foi ajustado utilizando o polinômio de quarto grau, empregando uma estrutura exponencial para a matriz de variâncias e covariâncias e considerando as colmeias como efeitos aleatórios. Basicamente (mas não unicamente), esse modelo difere do ajustado via `gamlss` (com distribuição marginal GG) por considerar uma estrutura da matriz de variâncias e covariâncias e por lidar com uma distribuição marginal normal.

Figura 10. Análise dos resíduos do modelo ajustado via `nlme`.



Uma análise dos resíduos do modelo ajustado via `nlme` pode ser vista na Figura 10. Pela análise dos dois gráficos superiores dessa figura, claramente se percebe que o modelo não foi capaz de lidar com a heterogeneidade da variância. Ademais, os resíduos (Figura 10) e os efeitos aleatórios (Figura 11) não apresentam distribuição normal, não atendendo aos pressupostos dos modelos mistos lineares.

Figura 11. Efeitos aleatórios preditos pelo modelo ajustado via `nlme`.



Retornando à Figura 5, tem-se indícios que uma transformação logarítmica possa lidar com a assimetria presente nos dados. Assim, um novo modelo foi ajustado

após somar uma unidade aos dados de produção de pólen (por eles apresentarem valores menores que 1) e aplicar a transformação logarítmica. O ajuste foi realizado utilizando as mesmas covariáveis dos modelos anteriores. A análise gráfica sugere que a transformação foi capaz de resolver os problemas de heterogeneidade das variâncias e não normalidade dos erros (Figura 12). Pode-se perceber também que o pressuposto de normalidade dos efeitos aleatórios não foi atendido (Figura 13). Todavia, deve-se atentar que o modelo é robusto com relação a quebra desse pressuposto.

Quando se contrasta os perfis médios, é possível perceber um bom ajuste aos dados de todos os modelos (Figura 14). As médias dos valores ajustados via GAMLSS e MLM com log-transformação foram mais similares entre si, quando comparadas às médias dos valores ajustados pelo modelo linear misto sem transformação. Contudo, todos os erros-padrão se sobrepõem aos erros-padrão dos valores observados.

Figura 12. Análise dos resíduos do modelo ajustado via `nlme` e com a variável resposta (y =Produção de pólen) transformada como $\log(y+1)$.

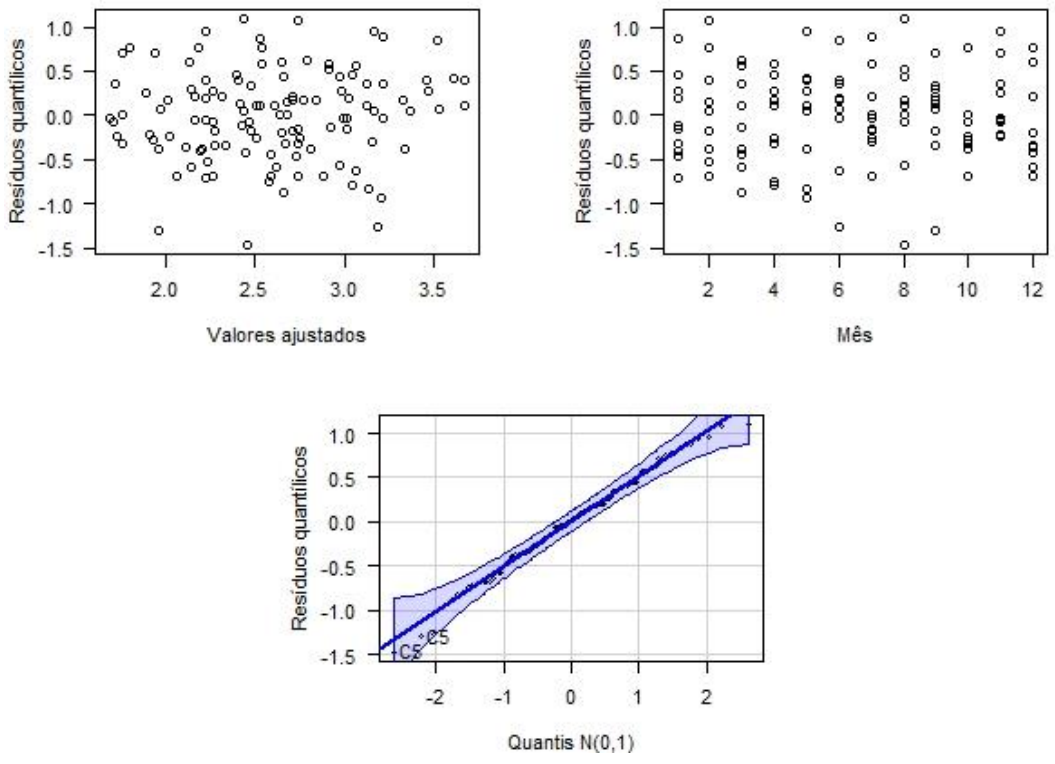


Figura 13. Efeitos aleatórios preditos pelo modelo ajustado via `nlme` e com a variável resposta (y =Produção de pólen) transformada como $\log(y+1)$.

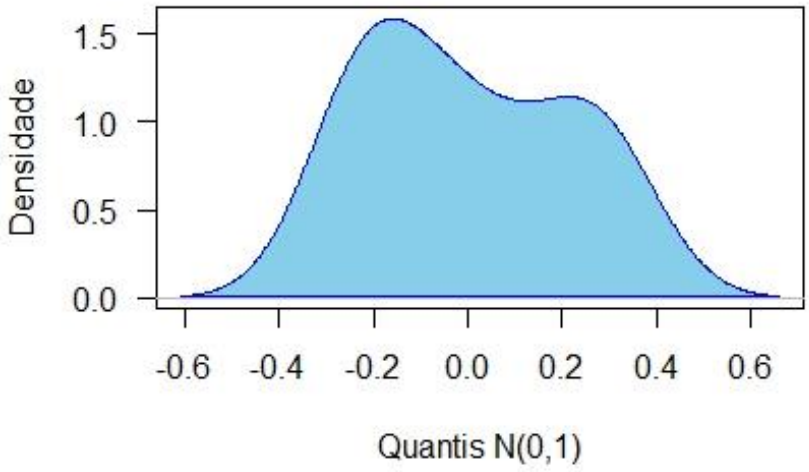
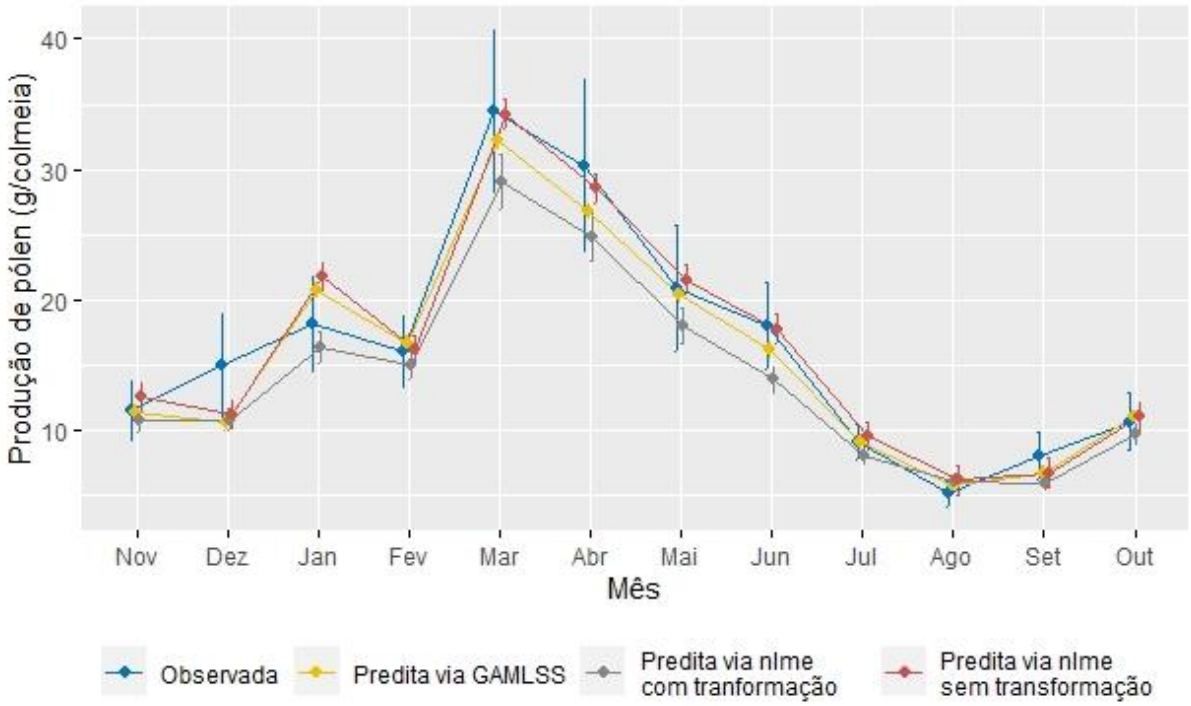


Figura 14. Produção de pólen (g/colmeia) observada (linha azul) e predita via GAMLSS (linha amarela), modelo linear misto sem transformação (linha vermelha) e modelo linear misto com transformação (linha cinza). As linhas verticais são erros-padrão.



As estimativas dos modelos ajustados via `nlme` e `gamlss` podem ser comparadas ao observar os valores da Tabela 7. Deve-se atentar que o `nlme` modela

de modo explícito apenas o parâmetro de localização (μ), enquanto o `gamlss` com distribuição marginal GG modela os parâmetros de localização (μ), escala (σ) e forma (ν). Quanto a modelagem do parâmetro μ, embora as estimativas pontuais divirjam entre os modelos, vê-se que as estimativas intervalares se sobrepõem (com exceção do β_5) (Tabela 7 e Figura 15). Para o parâmetro de escala, tem-se que $\beta_{\sigma 0}(\pm$erro-padrão$) = -0{,}376\ (\pm 0{,}145)$ e $\beta_{\sigma 1}(\pm$erro-padrão$) = -0{,}032\ (\pm 0{,}020)$. Para o parâmetro de forma, tem-se que $\beta_{\nu 0}(\pm$erro-padrão$) = -0{,}360\ (\pm 0{,}570)$ e $\beta_{\nu 1}(\pm$erro-padrão$) = -0{,}084\ (\pm 0{,}086)$.

Figura 15. Estimativas pontuais e intervalares dos coeficientes do parâmetro μ para os modelos lineares mistos (com e sem transformação $\log(y+1)$ com $y =$ Produção de pólen (g/colmeia)).

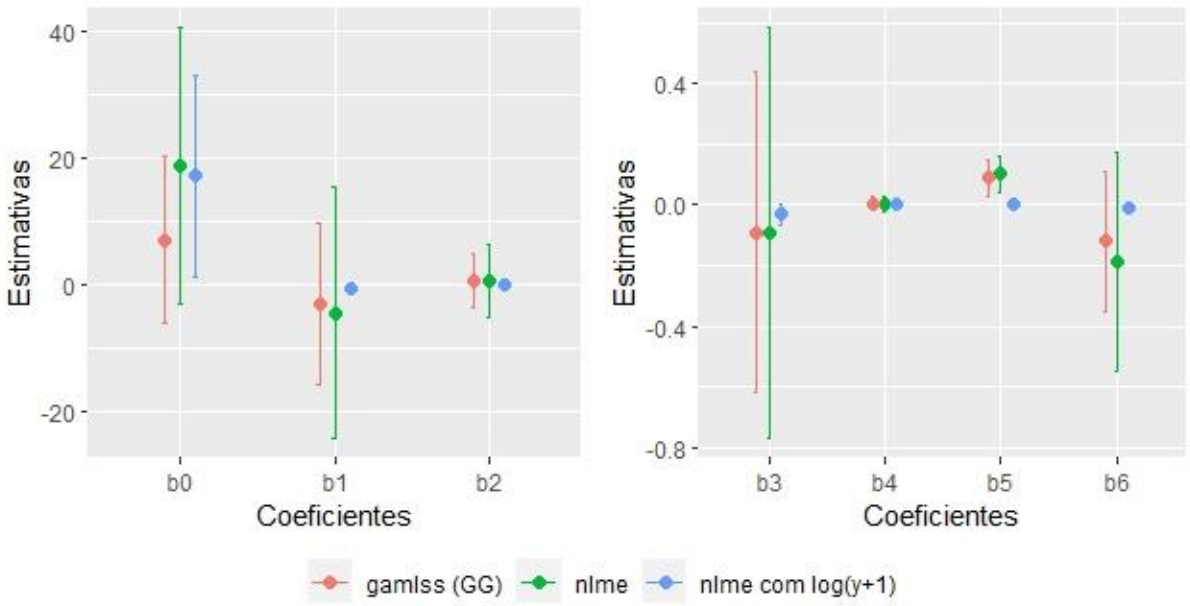


Tabela 7. Estimativas pontuais e intervalares para os modelos lineares mistos (com e sem transformação $\log(y+1)$, com $y =$ Produção de pólen (g/colmeia)). As estimativas do modelo linear misto com transformação sofreram transformação inversa para possibilitar comparações. Os intervalos de confiança (IC) - com 95% de confiança - foram calculados de acordo com método delta.

Modelo	Covariáveis	Coeficientes	Estimativas	IC (95%)	
				Inferior	Superior
Linear misto (nlme)	Intercepto	β_0	18,893	-2,809	40,596
	Mês	β_1	-4,297	-24,045	15,451
	Mês^2	β_2	0,833	-4,874	6,540
	Mês^3	β_3	-0,093	-0,768	0,582
	Mês^4	β_4	0,004	-0,022	0,031
	Precipitação	β_5	0,102	0,043	0,162
	N° de espécies em floração	β_6	-0,185	-0,545	0,175
Linear misto (nlme) com transformação $\log(y+1)$	Intercepto	β_0	17,255	1,368	33,141
	Mês	β_1	-0,402	-0,912	0,109
	Mês^2	β_2	0,249	-0,108	0,606
	Mês^3	β_3	-0,031	-0,066	0,004
	Mês^4	β_4	0,001	0,000	0,003
	Precipitação	β_5	0,003	0,000	0,006
	N° de espécies em floração	β_6	-0,009	-0,027	0,008
GAMLSS (gama generalizada)	Intercepto	β_0	7,210	-5,883	20,304
	Mês	β_1	-2,856	-15,686	9,974
	Mês^2	β_2	0,658	-3,592	4,908
	Mês^3	β_3	-0,091	-0,620	0,438
	Mês^4	β_4	0,004	-0,017	0,026
	Precipitação	β_5	0,089	0,027	0,150
	N° de espécies em floração	β_6	-0,118	-0,349	0,113

Uma vez que os dados originais foram transformados para um dos modelos $(\log(y+1))$, uma transformação reversa pode ser requerida. Aqui, foi aplica a transformação $\exp(estimativa) - 1$. Todavia, alguns problemas podem surgir. No caso

de a transformação reversa ser não linear, como a exponencial, poder-se-á obter um estimador positivamente viesado (Changyong *et al*. 2014). Neste caso, existirá a tendência de sobrestimar o valor do parâmetro. Uma vez que a log-transformação só pode ser usada para resultados positivos, é comum adicionar uma constante positiva a todas as observações antes de aplicar essa transformação. Embora essa prática seja comum, ela pode ter um efeito perceptível no nível de significância estatística em testes de hipóteses. Isso porque quanto maior o valor da constante empregada, menor o valor do nível descritivo do teste (Changyong *et al*. 2014). Neste trabalho, nenhum destes problemas pode ser encontrado com a transformação. Mas nitidamente os intervalos de confiança para os coeficientes foram extremamente estreitados (Tabela 7 e Figura 15).

4. Considerações finais do capítulo

O modelo GAMLSS com distribuição marginal GG para os dados de produção de pólen apresentou um bom ajuste, com intervalos de confiança mais estreitos quando comparados àqueles gerados pelo MLM (sem transformação). Pode-se notar também que a modelagem via GAMLSS atendeu aos pressupostos da técnica, diferentemente da modelagem via MLM (com e sem transformação). Dessa forma, o GAMLSS se apresentou como uma boa ferramenta para lidar com os dados longitudinais de produção polínica.

O MLM (com transformação) apresentou bom ajuste aos dados, semelhante ao GAMLSS. Contudo, a transformação da variável resposta pode implicar em prejuízos para as estimativas e inferências, especialmente ao que tange a interpretabilidade. Ademais, essa classe de modelos possibilita, explicitamente, modelar os diferentes parâmetros da distribuição. Por outro lado, na classe de MLM existe grande variedade de ferramentas diagnósticas, inclusive de influência para estudos transversais (Seber e Lee, 2012) e longitudinais (Nobre, 2004). Ferramental para análises de influência ainda é inexistente para GAMLSS longitudinais.

CONSIDERAÇÕES FINAIS

A proposta dos GAMLSS é bastante recente e suas aplicações são numericamente pouco representativas quando comparadas às demais técnicas. A utilização se torna ainda mais tímida quando se trata de modelos mistos (como para dados longitudinais). Aqui, empregou-se o GAMLSS com distribuição marginal GG para os dados de produção de pólen presentes em Nascimento *et al*. (2019). A boa qualidade de ajuste sugere que essa classe de modelos seja uma ferramenta poderosa para lidar com os dados longitudinais de produção polínica, que comumente seguem distribuições não pertencentes à família exponencial linear.

No modelo GAMLSS aqui empregado, não só o parâmetro de localização da distribuição GG dependeu de covariáveis, mas também os parâmetros de escala e forma. Isso implica em uma descrição e interpretação mais precisa da natureza da variável resposta (quando comparado as demais classes de modelos).

Destaca-se que mesmo a variável resposta assumindo somente valores positivos, poder-se-ia colocar a restrição na escolha das funções de ligação de forma que a média assumisse sempre valor positivo. Dessa forma, a utilização das funções de ligação juntamente com as técnicas de suavização e o grande número de distribuições disponíveis no pacote `gamlss` eleva ainda mais as possibilidades de modelagem de dados reais.

Para a análise de resíduos, os *worm plots* são excelentes ferramentas para analisar ajuste em GAMLSS, destacadamente quando se analisa o ajuste ao longo de uma variável de interesse. Contudo, análises de influência, amplamente disponíveis para os MLM, não estão disponíveis para os GAMLSS. Embora propostas de análises de influência para estes modelos tenham sido recentemente elaboradas para estudos transversais (Silva, 2021), não foram encontradas (até o presente momento) propostas para modelos mistos desta classe, o que é uma limitação. Portanto, sugere-se o desenvolvimento de técnicas de análises de influência em GAMLSS para diagnosticar de modo mais completo o ajuste dessa classe de modelos a dados empíricos.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H., Information measures and model selection. **Bulletin of the International Statistical Institute**, n. 50, p. 277–291, 1983.

ARAUJO-FILHO, J.A. Manejo Pastoril Sustentável da caatinga. **Pernambuco Cidade Gráfica e Editora LTDA**, p. 200, 2013.

ARRUDA, V. A. S., PEREIRA, A. A. S., ESTEVINHO, L. M., ALMEIDA-MURADIAN, L. B. Presence and stability of B complex vitamins in bee pollen using different storage conditions. **Food and Chemical Toxicology**, 51, 143-148, 2013.

AVNI, D., HENDRIKSMA, H. P., DAG, A., UNI, Z., SHAFIR, S. Nutritional aspects of honey bee-collected pollen and constraints on colony development in the eastern Mediterranean. **Journal of insect physiology**, v. 69, p. 65-73, 2014.

BARRETO, L.M.R.C., FUNARI S.R.C., ORSI, R.O. DIB A.P. Produção de pólen no Brasil, Taubaté. **Cabral Editora e Livraria Universitária**, p. 99, 2006.

BATES, D.; MAECHLER, M.; BOLKER, B.; WALKER, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, 67(1), 1-48.

BAUMANN, P.; ROSSI, E.; VOLKMANN, A. What Drives Inflation and How: Evidence from Additive Mixed Models Selected by cAIC. **arXiv preprint arXiv:2006.06274**, 2020.

BRESLOW, N.E.; CLAYTON, D.G. Approximate inference in generalized linear mixed models. **Journal of the American statistical Association**, v. 88, n. 421, p. 9-25, 1993.

BUUREN, S.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in medicine**, v. 20, n. 8, p. 1259-1277, 2001.

CLEVELAND, W.S. Robust locally weighted regression and smoothing scatterplots. **Journal of the American statistical association**, v. 74, n. 368, p. 829-836, 1979.

CHANGYONG, F. E. N. G. *et al*. Log-transformation and its implications for data analysis. **Shanghai archives of psychiatry**, v. 26, n. 2, p. 105, 2014.

COLE, T.J.; GREEN, P.J. Smoothing reference centile curves: the LMS method and penalized likelihood. **Statistics in medicine**, v. 11, n. 10, p. 1305-1319, 1992.

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

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G., MOLENBERGHS, G. (Eds.). **Longitudinal data analysis**. CRC press. 2008.

FRIEDMAN, J.H.; STUETZLE, W. Projection pursuit regression. **Journal of the American statistical Association**, v. 76, n. 376, p. 817-823, 1981.

HASTIE, T., TIBSHIRANI, R. **Generalized additive models.** CRC Press**,**1995.

KULLBACK, S.; LEIBLER. R. A. On information and sufficiency. **Ann. Math. Stat**., 22:79–86, 1951.

MARCHINI, L.C.; REIS, V.D.A.; MORETI, A.C.C.C. Composição físico-química de amostras de pólen coletado por abelhas africanizadas *Apis mellifera* (Hymenoptera: Apidae) em Piracicaba, Estado de São Paulo. **Ciência rural**, v. 36, p. 949-953, 2006.

MCCULLOCH, C.E.; SEARLE, S.R. **Generalized, linear, and mixed models**. John Wiley & Sons, 2004.

MILFONT, M.O.; FREITAS, B.M.; ALVES, J.E. Pólen Apícola. Manejo para a produção de pólen no Brasil. Viçosa, MG. **Aprenda Fácil**, 2011.

MONTGOMERY, D.C., PECK, E.A., VINING, G.G. **Introduction to linear regression analysis**. John Wiley & Sons, 2012.

MOOD, A.M., GRAYBILL, F.A., BOES, D.C. **Introduction to the Theory of Statistics** (3rd ed.), McGraw-Hill, New York. p. 564. 1973.

NASCIMENTO J. E. M., FREITAS, B. M., PACHECO FILHO, A. J. S., PEREIRA, E. S., MENESES, H. M., ALVES, J. E., SILVA, C. I. Temporal variation in production and nutritional value of pollen used in the diet of Apis mellifera L. in a seasonal semideciduous forest. **Sociobiology**, v. 66, n. 2, p. 263-273, 2019.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, v. 135, n. 3, p. 370-384, 1972.

NOBRE, J.S. **Métodos de diagnóstico para modelos lineares mistos**. Dissertação. IME/USP, Sao Paulo, 2004.

OECD. 2021. Glossary of statistical terms. http://stats.oecd.org/glossary/detail.asp?ID=998. Acessado em outubro de 2021.

PARRA, M. R.; COUTINHO, R. X.; PESSANO, E. F. C. Um breve olhar sobre a cienciometria: origem, evolução, tendências e sua contribuição para o ensino de ciências. **Revista Contexto & Educação**, v. 34, n. 107, p. 126-141, 2019.

PINHEIRO J., BATES D., DEBROY S., SARKAR D., R CORE TEAM. **nlme: Linear and Nonlinear Mixed Effects Models**. R package version 3.1-152, 2021.

R Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. 2021.

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

RIGBY, R.A.; STASINOPOULOS, D.M. Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. **Statistical methods in medical research**, v. 23, n. 4, p. 318-332, 2014.

RIGBY, R.A., STASINOPOULOS, M.D., HELLER, G.Z., BASTIANI, F. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. CRC press, 2019.

SATTLER, J. A. G., DE MELO, I. L. P., GRANATO, D., ARAÚJO, E., DE FREITAS, A. D. S., BARTH, O. M., ALMEIDA-MURADIAN, L. B. Impact of origin on bioactive compounds and nutritional composition of bee pollen from southern Brazil: A screening study. **Food Research International**, v. 77, p. 82-91, 2015.

SCHIELZETH, Holger *et al*. Robustness of linear mixed-effects models to violations of distributional assumptions. **Methods in Ecology and Evolution**, v. 11, n. 9, p. 1141-1152, 2020.

SEBER, G.A.F; LEE, A.J. **Linear regression analysis**. John Wiley & Sons, 2012.

SILVA, L.A. **Influential diagnostics for location parameter within GAMLSS**. Dissertação de Mestrado. Universidade Federal de Pernambuco. p. 65, 2021.

SINGER, J.M, ROCHA, F.M.M. E NOBRE, J.S. Graphical Tools for Detecting Departures from Linear Mixed Model Assumptions and Some Remedial Measures. **International Statistical Review**, v. 85, p. 290-324, 2017.

SODJA, C. **Detecting Anomalous Time Series by GAMLSS-Akaike-Weights-Scoring**. arXiv preprint arXiv:2002.00499, 2020.

STASINOPOULOS, M.D. *et al*. **Flexible regression and smoothing: using GAMLSS in R**. CRC Press, 2017.

THOMAS, G. **GAMLSSs with applications to zero inflated and hierarquical data**. Dissertação de Mestrado. Universidade de São Paulo, 2017.

THOMAS, G., PEREIRA, A.I.D.A., LOBOS, C.M.V. **Analysis of a longitudinal multilevel experiment using GAMLSSs**. arXiv preprint arXiv:1810.03085, 2018.

ZUUR, A., IENO, E. N., WALKER, N., SAVELIEV, A. A., SMITH, G. M. **Mixed effects models and extensions in ecology with R**. Springer Science & Business Media, 2009.

ANEXO I

- Reparametrização de Lopatatzidis e Green (*apud* Stasinopoulos *et al*. 2017) da distribuição gama generalizada

A forma funcional mais comumente empregada para representar a distribuição gama generalizada é dada por

$$f(y|\mu,\sigma,\nu) = \frac{|\nu| y^{\nu\theta-1} e^{-\left(\frac{y}{\beta}\right)^{\nu}}}{\beta^{\nu\theta}\Gamma(\theta)} \mathbb{I}_{y>0},$$

com $\beta > 0, \theta > 0$ e $\nu \in \mathbb{R}$. Se se tomar $\beta = \mu(\sigma^2\nu^2)^{\frac{1}{\nu}}$, $\theta = \frac{1}{\sigma^2\nu^2}$ e $z = \left(\frac{y}{\mu}\right)^{\nu}$, então

$$\begin{aligned} f(y|\mu,\sigma,\nu) &= \frac{|\nu| y^{\nu\theta-1} e^{-\left(\frac{y}{\beta}\right)^{\nu}}}{\beta^{\nu\theta}\Gamma(\theta)} \mathbb{I}_{y>0} \\ &= \frac{|\nu| y^{\nu\theta-1}}{\left[\mu(\sigma^2\nu^2)^{\frac{1}{\nu}}\right]^{\nu\theta}\Gamma(\theta)} e^{-\left(\frac{y}{\mu(\sigma^2\nu^2)^{\frac{1}{\nu}}}\right)^{\nu}} \mathbb{I}_{y>0} \\ &= \frac{|\nu| y^{\nu\theta}}{\mu^{\nu\theta}(\sigma^2\nu^2)^{\theta}\Gamma(\theta) y} e^{-\left[\left(\frac{y}{\mu}\right)^{\nu}\left(\frac{1}{(\sigma^2\nu^2)^{\frac{1}{\nu}}}\right)^{\nu}\right]} \mathbb{I}_{y>0} \\ &= \frac{|\nu|}{\Gamma(\theta) y}\left(\frac{1}{\sigma^2\nu^2}\right)^{\theta}\left(\left(\frac{y}{\mu}\right)^{\nu}\right)^{\theta} e^{-\left[z\left(\frac{1}{\sigma^2\nu^2}\right)\right]} \mathbb{I}_{y>0} \\ &= \frac{|\nu|\theta^{\theta} z^{\theta} e^{-z\theta}}{\Gamma(\theta) y} \mathbb{I}_{y>0}, \end{aligned}$$

com $\mu > 0, \sigma > 0, z > 0$ e $\nu \in \mathbb{R}$.