UNIVERSIDADE FEDERAL DO ACRE

MARLON DA SILVA ROGÉRIO

MODELO ODD LOG-LOGÍSTICA SKEW T-STUDENT: UMA APLICAÇÃO EM DADOS COM MEDIDA REPETIDA NO TEMPO

RIO BRANCO
2023

FEDERAL UNIVERSITY OF ACRE

MARLON DA SILVA ROGÉRIO

MODELO ODD LOG-LOGÍSTICA SKEW T-STUDENT: UMA APLICAÇÃO EM DADOS COM MEDIDA REPETIDA NO TEMPO

Proposta de dissertação de mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação na Universidade Federal do Acre como requisito parcial para obtenção do título de mestre em Ciência da Computação. Linha de Pesquisa: Engenharia de Software.

Orientador:
Prof. Dr. Altemir da Silva Braga

RIO BRANCO
2023

MARLON DA SILVA ROGÉRIO

MODELO ODD LOG-LOGÍSTICA SKEW T-STUDENT: UMA APLICAÇÃO EM DADOS COM MEDIDA REPETIDA NO TEMPO

Proposta de dissertação de mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação na Universidade Federal do Acre como requisito parcial para obtenção do título de mestre em Ciência da Computação. Linha de Pesquisa: Engenharia de Software.

Approved in <MONTH> of <YEAR>.

Prof. Dr. Altemir da Silva Braga
Universidade Federal do Acre

RIO BRANCO
2023

Tudo o que temos de decidir é o que fazer com o tempo que nos é dado.

Gandalf.

O Senhor dos Anéis - A Sociedade do Anel (2001)

Agradecimentos

À gloria do Grande Arquiteto do Universo e a Ele são destinados os primeiros agradecimentos para este trabalho.

Agradeço também a minha família: minha esposa querida, **Caroline Sá**, meus pais **Edmar Rogério** e **Joice Vieira**, e minhas irmãs **Caroline Rogério** e **Karine Rogério**. Cada passo desta caminhada teve a participação de vocês e por isso serei eternamente grato.

Ao corpo docente do Programa de Pós-Graduação em Ciências da Computação da UFAC pelo enriquecimento intelectual proporcionado em especial ao **Prof. Dr. Altemir da Silva Braga** pela extrema paciência, resiliência e dedicação para ensinar um mundo novo de conhecimentos.

Finalmente, à Universidade Federal do Acre pela iniciativa do programa possibilitando o aperfeiçoamento acadêmico para área de tecnologia dentro do estado.

Resumo

ANÁLISE DO MODELO ODD LOG-LOGISTIC SKEW T-STUDENT PARA DADOS DE MEDIDA REPETIDA NO TEMPO

Proposto em 2021 por (FERNANDES, 2021), o modelo Odd Log-Logistic Skew t-Student (OLLST) apresentou uma proposta de regressão semi-paramétrica com a possibilidade de ajuste para diferentes tipos de distribuição. Considerando que o modelo ainda não havia sido testado para dados com medida repetida no tempo, este estudo conduziu alguns testes de modelagem para seis conjuntos de dados diferentes (MARTINS, 2022), onde a variável resposta foi analisada em função do tempo. Os resultados preliminares mostraram o bom desempenho do OLLST para alguns conjuntos de dados, quando comparado ao modelo Normal (NO), em outros uma relação de semelhança nas comparações e para alguns sua performance foi ligeiramente inferior. Os testes revelaram ainda uma saída de dados poluída para os resultados obtidos pela função `gamlss` da linguagem $\mathcal{R}$, com pouca intuitividade e propensão maior a erros. Como resultado uma ferramenta foi desenvolvida com objetivo de refinar a saída e automatizar a rotina de comparação dos indivíduos em função do tempo.

Palavras-chave: Regressão; GAMLSS; OLLST; Medida repetida no tempo.

Abstract

Proposed in 2021 by (FERNANDES, 2021), the Odd Log-Logistic Skew t-Student (OLLST) model presented a semi-parametric regression proposal with the possibility of adjustment for different types of distribution. Considering that the model had not yet been tested for data with repeated measures over time, this study conducted some modeling tests for six different data sets (MARTINS, 2022), where the response variable was observed as a function of time. The preliminary results appreciated the good performance of the OLLST for some sets of data, when compared to the Normal model (NO), in others a relation of similarity in the comparisons and for some its performance was inferior. The tests also revealed a polluted data output for the results obtained by the `gamlss` function of the $\mathcal{R}$ language, with little intuitiveness and a greater tendency to errors. As a result, a tool was developed with the aim of refining the output and automating the routine of comparing individuals over time.

Keywords: Regression; GAMLSS; OLLST; Repeated measure in time

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

É comum que pesquisadores busquem entender os efeitos que uma variável pode ou não exercer sobre outras, ainda que não exista entre elas uma relação direta que, no entanto, poderia ser genericamente representada por $y = f(x_1, x_2, ..., x_n)$. Desse modo, por exemplo, com base nos valores conhecidos para x é possível estimar a influência dele sobre y. Esta ideia representa o conceito dos modelos de regressão que, em outros termos, buscam explicar a relação entre variável dependente (resposta) em função de uma ou mais variáveis explicativas (independentes) (CHEIN, 2019; ANGRIST; PISCHKE, 2009)

Em alguns casos, estas relações podem apresentar um esquema estrutural em que a disposição dos dados exigem uma relação de subordinação entre variáveis (NATIS, 2001). Em grande parte dos estudos com essa característica a variável "tempo"é o fator de associação mais comum, de modo que no modelo hierárquico sua relação pode ser descrita genericamente por $\mathbf{y}_i = (y_{i1}, \cdots, y_{im_i})$, sendo y_{ik} a resposta para o i-ésimo indivíduo no k-ésimo tempo, $k = (1, \cdots, m_i)$ (SINGER; ANDRADE, 2000).

A aplicação prática desse conceito pode ser encontrada, por exemplo, ao se perceber a variação de melhora ou piora em um paciente quando submetido a um tratamento ao longo do tempo, ou ainda, quando uma região de solo é submetido a controles que influenciam nos níveis de seus nutrientes para diferentes observações no tempo. Em ambos os casos, o delineamento do modelo deve considerar as variáveis dependentes para cada tempo e assim concluir sobre a eficácia dos controles aplicados.

Estudos em que as observações sobre os indivíduos são coletadas repetidamente ao logo do tempo são classificados pela literatura como **medida repetida no tempo**, **dados longitudinais**, **coorte**, **painel**, entre outros (SINGER; ANDRADE, 2000; ROSA, 2001; FAGUNDES, 2013).

Ainda que sejam comuns, os estudos de medida repetida no tempo apresentam desafios maiores quando comparado a outros ajustes (FITZMAURICE et al., 2008), resultado da natureza de correlação estrutural dos dados, implicando na necessidade de uma modelagem que satisfaça a condição de subordinação dos grupos (CNAAN; LAIRD; SLASOR, 1997; FAUSTO et al., 2008).

Atualmente, o padrão de delineamento de dados longitudinais é dado pelo uso de modelos lineares misto ou de efeito aleatório, (LAIRD; WARE, 1982), em que exigi-se a pressuposição de um comportamento gaussiano da variável resposta ou dos erros, de modo que na ausência de tal pressuposto cabe ao pesquisador aplicar uma transformação nos dados a fim de normalizá-los (CNAAN; LAIRD; SLASOR, 1997).

Em síntese, os modelos mistos, na ausência de normalidade dos erros, apresentavam uma flexibilidade menor na captura de varições entre a variável resposta, exigindo técnicas de normalização dos dados. Em alguns caso o modelo pode até gerar resultados razoáveis, no entanto, deve ser evitado em detrimento a abordagens mais eficientes (AZZALINI; CAPITANIO, 1999; BRAGA et al., 2022).

Os aspectos descrito acima, demostram parte da complexidade de modelar dados longitudinais (BRAGA et al., 2022), somando as dificuldades da ausência de normalidade à estruturação hierarquizada dos dados, motivo pelo qual o autores (AZZALINI; CAPITANIO, 1999; AZZALINI et al., 2003; SAHU; DEY; BRANCO, 2003; SAHU; DEY, 2004; BRAGA et al., 2022) têm se debruçado sobre esses campos na proposta de modelos cada vez abrangentes, contemplando distribuição assimétricas, bimodais, além da gaussiana (normal).

Considerando os desafios da distribuição dos dados, tendo base os estudos de (AZZALINI; CAPITANIO, 1999; AZZALINI et al., 2003; SAHU; DEY; BRANCO, 2003; SAHU; DEY, 2004), o autor (FERNANDES, 2021) criou um novo modelo de regressão chamado Odd Log Logística Skew t-Student (OLLST) com a proposta de ajustar-se melhor na presença de assimetrias (à esquerda e à direita) e assimetrias de bimodalidades.

Na aplicação do modelo OLLST, (FERNANDES, 2021) pôde constatar o bom desempenho dele na análise para diferentes tipos de distribuição, recomendando-o como alternativa para delineamentos em que a variável resposta não segue distribuição normal.

Tendo como base os resultados propostos por (FERNANDES, 2021) para análise de assimetrias e bimodalidades, esta dissertação analisou o comportamento do modelo OLLST no delineamento de dados reais, comparando-o ao modelo normal (NO), a fim

de saber qual explicaria melhor a relação imposta entre as variáveis no ajuste de dados longitudinais.

Durante o estudo sobre o modelo, verificou-se que, por pertencer à classe de modelos aditivos generalizados para posição, escala e forma (GAMLSS), a interpretação da saída de dados da biblioteca $\mathcal{R}$ para essa família de modelos tende a ser confusa, pouco intuitiva e com aumento da complexidade na proporção em que cresce o número de indivíduos. Por esse motivo, uma proposta de automação e adaptação da saída foi implementada com a possibilidade de uso ou não de uma interface gráfica de manipulação.

1.1 Organização

Esta dissertação segue estruturada conforme o disposto. No Capítulo 2 será apresentada uma revisão de conceitos relacionados a pesquisa com intuito de nortear o leitor quanto ao contexto do trabalho e tecnologias usadas para construção do modelo OLLST pelo autor (FERNANDES, 2021). Nesse ponto, a proposta de medida repetida no tempo é incorporada ao ambiente da análise de regressão de modo que fiquem claro os desafios de inerentes à estas modelagens, suscitando uma abordagem mais eficiente.

No Capítulo 3 serão discutidas as estrutura matemáticas de geração do modelo OLLST, sua incorporação à família de distribuição GAMLSS, a postulação do modelo hierárquico para modelagem de dados longitudinais e os mecanismos de estimação usado pela função. Será discorrido também sobre o momento de pesquisa que levou a proposta de criação para melhoria na análise e saída dos dados ajustados pelos modelos de regressão hierárquico.

No Capítulo 4 são apresentados inicialmente os resultados obtidos para os ajustes de seis conjuntos de dados para o teor de carbono no solo de diferentes tipos de tratamentos em função do tempo, além da análise comparativa do modelo OLLST em razão do Normal. Neste capítulo também será apresentada a ferramenta desenvolvida com objetivo de facilitar a interpretação dos dados para ajustes longitudinais, com interface gráfica e rotinas de automação para comparações.

Finalmente no Capítulo 5 uma síntese da pesquisa será exposta com as considerações que culminaram neste estudo e um panorama dos resultados alcançados para duas frentes de trabalho proposta.

Capítulo 2

Conceitos

2.1 Regressão

De acordo (SAHU; DEY; BRANCO, 2003), a análise de regressão é um método importante na estatística e tem por proposta conhecer os efeitos que algumas variáveis podem ou não exercer sobre outras. Ainda que não existe de fato uma relação casual entre as variáveis, com análise de regressão é possível relacioná-las por meio de uma expressão matemática, podendo desta forma estimar o valor de uma variável com base nos valores das demais.

Tais relações podem ser genericamente representada por

$$y = f(x_1, x_2, ..., x_k), \tag{2.1}$$

em que y é a variável dependente e os x_k são as variáveis explicativas. Em um nível mais interno de conceitos os modelos regressão podem apresentar comportamento diferentes para estimação de valores, sendo eles paramétricas, não-paramétricas e semi-paramétricas.

De maneira geral, o conceito e aplicação dos tipos de regressão está ligados diretamente ao conhecimento ou não da função f. No primeiro caso, tem-se o conhecimento da função de distribuição f, no entanto, não se sabe o valor dos parâmetros. Na regressão não-paramétrica tem-se o comportamento inverso ao paramétrico, ao passo que a regressão semi-paramétrica tenta contemplar componentes paramétricos e não-paramétricos.

2.2 Medida repetida no tempo

O estudo de medidas repetidas no tempo apresenta ampla aplicação, contemplado áreas como medicina, agronomia, economia, entre outras, tendo como principal característica a

observação repetida de um indivíduo ao logo do tempo. Um exemplo disso ocorre quando de deseja modelar a variação diária da pressão sanguínea de indivíduos com diferentes tipos de quadros clínicos (SINGER; ANDRADE, 2000).

Neste cenário, as aferições repetidas ao longo do tempo para o mesmo indivíduo estabelecem uma relação de associação entre o tempo e a variável resposta. Se os indivíduos deste cenário fossem submetidos a tratamentos medicamentosos diferentes e as aferições repetidas avaliasse a reação da droga, a análise comparativa entre indivíduos para cada tratamento deveria então considerar o tempo como fator de influência e assim concluir, por exemplo, qual o medicamento mais eficaz.

No aspecto estrutural, os dados longitudinais seguem de modo que o vetor com m_i respostas para o i-ésimo indivíduo $i = (1, ..., n)$ qualifica o perfil individual da resposta, variando ainda no tempo $k = (1, ..., m_i)$, ou seja, cada indivíduo i terá no tempo k um vetor com n variáveis resposta. Desse modo, é possível descrever o perfil de uma resposta $\mathbf{y}$ no tempo como $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{im_i})$ (KAC; SICHIERI; GIGANTE, 2007; SINGER; ANDRADE, 2000)

Este conceito é melhor compreendido ao ser observar a Tabela 2.1 em que é proposto um paradigma para organização de dados longitudinais, demonstrando os pontos de influência do tempo sobre as variáveis. Considerando y_{ik} elemento do vetor de m_i respostas no tempo k para o i-ésimo individuo, além das variáveis explicativas (X, W, VeZ), tem-se que o tempo está associado, por exemplo, as variáveis y e Z, indicando uma relação de dependência entre eles.

Outro modo interpretar o comportamento primário de medida repetida no tempo se dá pela análise do gráfico de perfis individuais (veja as Figuras 2.1 e 2.2), sendo possível inferir a variação na resposta y ao longo do tempo, bem como em seu ponto inicial, justificando, por exemplo, o uso de efeito aleatório no intercepto durante o delineamento, a fim de oferecer uma margem maior de variabilidade ao modelo.

Entre os modelos dispostos na literatura os mais difundidos no delineamento de dados longitudinais trabalham com a pressuposição de normalidade da variável resposta ou dos erros (SINGER; ANDRADE, 2000). Este é o caso, por exemplo, dos modelos lineares misto (ou modelo de efeito aleatório), que apesar da amplitude na modelagem ao incorporar o efeito aleatório, apresentam limitações no campo da distribuição.

Tabela 2.1: Estrutura para disposição de dados longitudinais (SINGER; ANDRADE, 2000).

			Covariáveis			
Indivíduos $_{(i=1,...,n)}$	Resposta $_{m_i}$	Tempo $_{(k=1,...,m_i)}$	X	W	V	Z
1	y_{11}	t_{11}	x_1	w_1	v_1	z_{11}
1	y_{12}	t_{12}	x_1	w_1	v_1	z_{12}
.	–	–	–	–	–	–
1	y_{1m_1}	t_{1m_1}	x_1	w_1	v_1	z_{1m_1}
2	y_{21}	t_{21}	x_2	w_2	v_2	z_{21}
2	y_{22}	t_{22}	x_2	w_2	v_2	z_{22}
.	–	–	–	–	–	–
2	y_{2m_2}	t_{2m_2}	x_2	w_2	v_2	z_{2m_2}
n	y_{n1}	t_{n1}	x_n	w_n	v_n	z_{n1}
n	y_{n2}	t_{n2}	x_n	w_n	v_n	z_{n2}
.	–	–	–	–	–	–
n	y_{nm_n}	t_{nm_n}	x_n	w_n	v_n	z_{nm_n}

2.3 Modelo linear de efeito misto

Proposto por (LAIRD; WARE, 1982), o modelo de regressão linear é também utilizando como uma das técnicas de análise de dados longitudinais. Dado por

$$y_i = X_i\beta + Z_i b_i + e_i \tag{2.2}$$

, em que y_i é o vetor de resposta de dimensão $n_i i = 1, 2, ..., m$ e Z_i são matrizes de delineamento (ou incidência) conhecidas, de dimensões $n_i \times p$e $n_i \times q$ respectivamente, β é um vetor p-dimensional contendo os efeitos fixo, b_i é um vetor q-dimensional contendo os efeitos aleatórios, e e_i é o vetor de resíduos de dimensão n_i. Diferente dos modelos padrões de regressão linear, este incorpora ao modelo a dependência das observações e a estrutura de correlação dos erros (FAUSTO et al., 2008; FERREIRA, 2012; FREITAS; PRESOTTI; TORAL, 2005; PEREIRA et al., 2013).

Por outro lado, a distribuição dos dados pode impactar diretamente na eficiência do modelo, pois exige-se uma pressuposição de normalidade (AZZALINI; CAPITANIO, 1999), uma premissa falha caso o conjunto de dados possa apresentar um comportamento foram do padrão esperado. Na mesma linha, ao se trabalhar com medida repetida no tempo, tem-se o fato de que a distribuição da resposta, na maioria dos casos, apresenta uma zona de assimetria em direção aos maiores tempos, o que tornaria inapropriado o uso do modelo 2.2 (COLOSIMO; GIOLO, 2006).

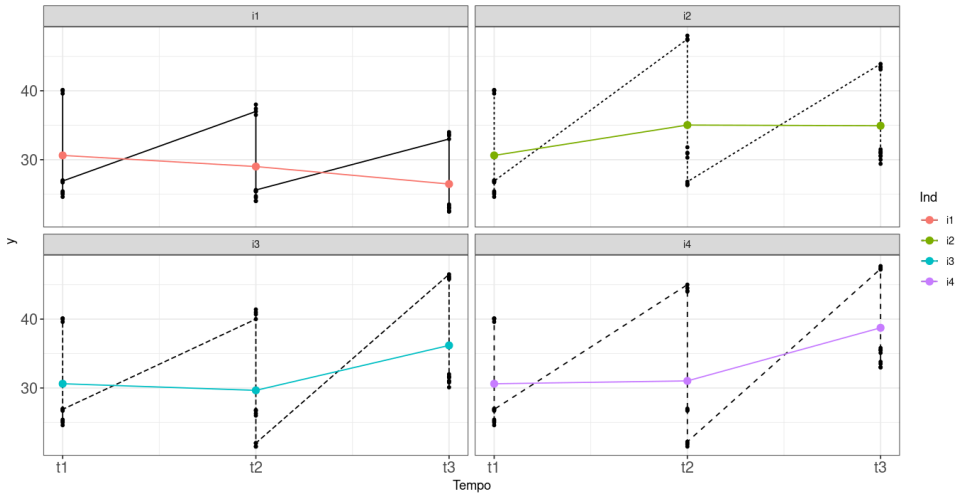

Figura 2.1: Gráfico de perfis com perfil individual médio por individuo

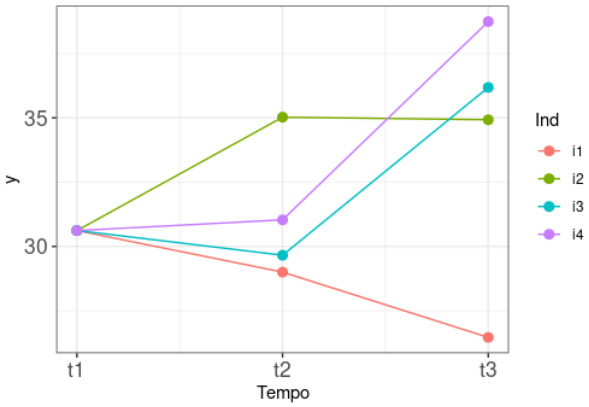

Figura 2.2: Gráfico de perfis com perfil individual médio em conjunto

A figura 2.3, por exemplo, apresenta um histograma para determinado conjunto de dados cuja a distribuição não é normal, ao passo que a 2.4 mostra um conjunto com distribuição esperado pelos modelos lineares. Em ambos os casos, a linha vermelha sob o gráfico mostra a função de densidade e probabilidade (fdp) da normal, mostrando as limitações do modelo quando se percebe de dissonâncias na distribuição.

Grande parte dos estudos com medida repetida no tempo enfrentam as limitações da modelagem do modelo linear transformando a variável resposta para tentar retornar ao modelo linear-normal, quando não, acabam utilizando um componente sistemático não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico. Em ambos os casos, o pesquisador pode se deparar com outras distribuições assimétricas ou mesmo bimodais para o erro, que não sejam possíveis de retornar para o modelo linear (COLOSIMO; GIOLO, 2006), exigindo modelos mais flexíveis para o ajuste.

Deve-se considerar, neste contexto, o momento histórico para o uso de cada modelo de

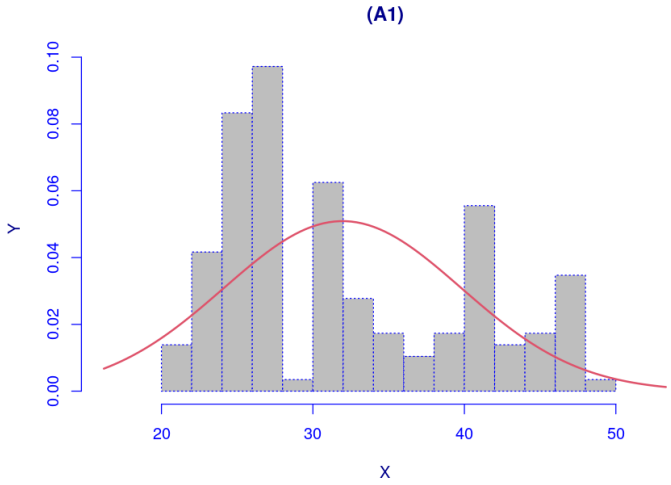

Figura 2.3: Histograma para dados sem distribuição normal.

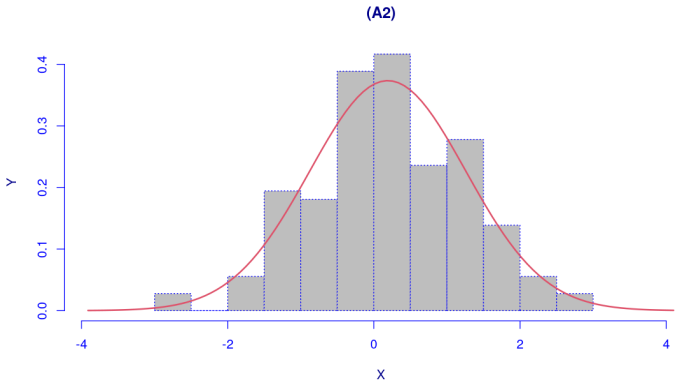

Figura 2.4: Histograma para dados com distribuição normal.

regressão. Criado em meados de 1964, devido as limitações computacionais para cálculos mais complexos ou de grande volumes de dados, a transformação Box-Cox (BOX; COX, 1964), de maneira minimalista, produzia uma normalidade aproximada para o ajuste dos modelos da época. Neste mesmo período, métodos como da máxima verossimilhança que envolviam cálculos mais complicados eram simplesmente impraticáveis. Contudo, hoje, com a melhora na performance dos computadores, o estudo de métodos simplista está sendo desistimulado em detrimento a técnicas mais eficientes (MADDALA, 2003).

2.4 Modelos aditivos

Os modelos aditivos são uma generalização do modelo linear tradicional, por tanto, acompanha a importante característica de verificar a contribuição que cada variável exerce sobre a variável de interesse (LIMA, 2001). Tomando como base a equação **??**, considerando n

pares de observações $(x_i, y_i), i = (1, 2, ..., n)$ e sendo f a função que estabelece a relação entre as variáveis X e Y da forma

$$y_i = f(x_i) + e_i. \tag{2.3}$$

Considerando um conjunto de k variáveis explicativas representadas em uma matriz X, de dimensão $n \times k$ com a i-ésima linha dada por $X_i = (x_{i1}, x_{i2}, ..., x_{ik})$, teremos f de modo que $y_i = f(x_{i1}, x_{i2}, ..., x_{ik}) + e_i$.

Os autores (BUJA; HASTIE; TIBSHIRANI, 1989; FLORENCIO, 2010), desenharam uma nova família de modelos, chamados modelos aditivos, de modo que fosse possível também aplicar a regressão linear no efeito as variáveis regressoras, de modo que modelo proposto na equação 2.3 passou a ser

$$y_i = f_1(x_{i1})f_2(x_{i2}) + ... + f_k(x_{ik}) + e_i, \tag{2.4}$$

A principal vantagem desta proposta seria a superação do problema conhecido na literatura como "maldição da dimensionalidade"(KÖPPEN, 2000), já que cada função f do somatório é estimada de modo univariado.

Apesar da melhora nas abordagens dos modelos existentes para análise de dados longitudinais (ver por exemplo (BUJA; HASTIE; TIBSHIRANI, 1989), (FLORENCIO, 2010) (SAHU; DEY; BRANCO, 2003) e (AZZALINI et al., 2003)), ainda existem lacunas no estudo de modelos assimétricos e bimodais, visto que fenômenos desta natureza acrescentam uma camada extra de complexidade no fator de identificação do comportamento da população. Na bimodalidade, por exemplo, é comum se deparar com a miscigenação de duas populações, em que a zona de densidade e probabilidade dos dados caracteriza-se por duas modas, aumentado consideravelmente a complexidade de se inferir comportamento dos indivíduos (WANG et al., 2009).

2.5 Modelos Aditivos Generalizados de Localização, Escala e Forma (GAMLSS)

A construção da família de modelos GAMLSS é resultado de uma linha do tempo evolutiva dos modelos tradicionais e suas limitações somado ao avanço da capacidade computacional para resolução de problemas complexos.

Como parte deste cenário, os modelos lineares generalizados (*Generalized Linear Models* - GLM) (NELDER; WEDDERBURN, 1972) assim como os modelos aditivos generalizados (*Generalized Additive Models* - GAM) (BUJA; HASTIE; TIBSHIRANI, 1989), derivado dos lineares tradicionais, por muito tempo ocuparam lugar de destaque na literatura como referência na análise de regressão univariada.

Tanto nos modelos GLM quanto nos GAM, a pressuposição de que a variável resposta $\boldsymbol{y}$ tenha distribuição pertencente a família exponencial, implica na modelagem da média $\boldsymbol{\mu}$ em função das variáveis preditoras. Nesses modelos a variância de $\boldsymbol{y}$ depende de um parâmetro de dispersão constante $\boldsymbol{\phi}$ e da média $\boldsymbol{\mu}$, ou seja, a variância, assim como a assimetria e a curtose não são modeladas explicitamente em termos de variáveis preditoras e sim por meio da dependência da média $\boldsymbol{\mu}$ (RIGBY; STASINOPOULOS, 1996).

Estudos como (BRESLOW; CLAYTON, 1993; BRESLOW; LIN, 1995; NELDER; WEDDERBURN, 1972), apresentam uma nova classe de modelo como resultado de um *crossover* (cruzamento de estilos) entre o modelo linear generalizado e modelos linear misto com o acréscimo de um termo (quase sempre normal). Os derivados desta nova família foram chamados de modelos generalizado misto (*Generalizerd Linear Mixed Model* - GLMM) e mesmo apresentando maior flexibilidade que os GLM e GAM, a pressuposição de que y tem sua distribuição pertencente a família exponencial torna-o limitante em alguns aspectos.

Como resultado destas barreiras, os autores (RIGBY; STASINOPOULOS, 2005) propuseram uma nova abordagem para análise de relações entre as variáveis. Nesse modelo de regressão (semi)paramétrica, seu comportamento paramétrico faz referência a necessidade de uma suposição quanto distribuição para variável resposta, ao passo que o comportamento "semi"diz respeito a amplitude na modelagem dos parâmetros de distribuições e funções para formas lineares, não lineares e suavizações de variáveis explicativas.

No GAMLSS, diferente do GLM e GAM citados anteriormente, a variável resposta y não precisa pertencer à família exponencial. Neste caso, a distribuição de y passa a pertencer à uma família mais genérica ($\mathcal{D}$) dada por $D(y|\mu, \sigma, \nu, \tau)$, em que $D \in \mathcal{D}$, podendo ser qualquer distribuição, seja ela contínua com assimetrias ou curtose acentuadas ou mesmo discreta. Uma outra vantagem dos modelos derivados do GAMLSS é a possibilidade de modelagem de todos os parâmetros da distribuição condicional de y e não apenas a média μ. (RIGBY; STASINOPOULOS, 1996). Somando a isso, existe ainda o aspecto que refere-se a facilidade de acesso e manuseio do modelo computacionalmente desenhado e implementado. A biblioteca `gamlss` da linguagem de programação $\mathcal{R}$ per-

mite ajustar mais de 50 distribuições diferentes, já incorporadas no pacote, podendo ainda ajustar novos modelos que atendam as premissas do GAMLSS.

2.5.1 Definição do modelo

Em sua estrutura, os p parâmetros $\theta^\top = (\theta_1, \theta_2, ..., \theta_p)$ de uma função de densidade e probabilidade $f(y|\theta)$ são modelados utilizando termos aditivos, presumindo que para $i = (1, 2, ..., n)$ as observações y_i são independentes e condicionais a θ^i. Neste caso, a função de densidade e probabilidade é dada por $f(y_i|\theta^i)$, em que $\theta^{i\top} = (\theta_{i1}, \theta_{i1}, ..., \theta_{ip})$ é um vetor de p parâmetros relacionado às variáveis explanatórias e efeitos aleatórios. De modo que o modelo GAMLSS pode ser descrito por

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{j_k} Z_{jk}\gamma_{jk}, \tag{2.5}$$

, em que θ_k e η_k são vetores de parâmetros $n \times 1$, X_k e Z_{jk} são covariáveis fixas, conhecidas e de ordem $n \times J_k^{'}$ e $n \times q_{jk}$, respectivamente, ao passo que γ_{jk} é uma variável aleatória q_{jk}-dimensional.

Grande parte dos estudos práticos para análise de regressão utilizam um máximo de quatro parâmetros ($p = 4$), usualmente caracterizados pela posição (μ), escala (σ), assimetria (ν) e curtose (τ), sendo $\theta_1 = \mu$ e $\theta_2 = \sigma$ parâmetros de posição (ou locação) e $\theta_3 = \mu$ e $\theta_3 = \sigma$ parâmetros de forma. Com isto, tem-se os seguintes modelos:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{j_k} Z_{jk}\gamma_{jk} \tag{2.6}$$

$$g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{j_1} Z_{j1}\gamma_{j1} \tag{2.7}$$

$$g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{j_2} Z_{j2}\gamma_{j2} \tag{2.8}$$

$$g_3(\nu) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{j_3} Z_{j3}\gamma_{j3} \tag{2.9}$$

$$g_4(\tau) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{j_4} Z_{j4}\gamma_{j4} \tag{2.10}$$

2.5.2 Família GAMLSS

Considerando que função de densidade e probabilidade $f(y|\theta)$ do modelo 2.5 pode pertencer à $\mathcal{D}$, dado por $D(y|\mu, \sigma, \nu, \tau)$, em que $D \in \mathcal{D}$, ao se ajustar um modelo GAMLSS utilizando a biblioteca `gamlss` da linguagem de programação $\mathcal{R}$ é necessário atentar-se unicamente para que a função $f(y|\theta)$ assim como sua primeira derivada, para cada parâmetro θ, sejam calculáveis.

As tabelas 2.2 e 2.3 apresenta uma lista de distribuições já incorporadas à biblioteca `gamlss` do $\mathcal{R}$.

Tabela 2.2: Exemplos de distribuições discretas (RIGBY; STASINOPOULOS, 2005).

		Função de Ligação			
Distribuição	Nome	μ	σ	ν	τ
Beta Binomial	BB()	Logit	Log	–	–
Binomial	BI()	Logit	–	–	–
Delaporte	DEL()	Log	Log	Logit	–
Negative Binomial type I	NBI()	Log	Log	–	–
Negative Binomial type II	NBII()	Log	Log	–	–
Poisson	PO()	Log	–	–	–
Poisson Inverse Gaussian	PIG()	Log	Log	–	–
Sichel	SI()	Log	Log	Identity	–
Sichel (μ the mean)	SICHEL()	Log	Log	Identity	–
Zero Inflated Poisson	ZIP()	Log	Logit	–	–

Tabela 2.3: Exemplos de distribuições contínuas (RIGBY; STASINOPOULOS, 2005).

	Função de Ligação				
Distribuição	None	μ	σ	ν	τ
Beta	BE()	Logit	Logit	–	–
Beta Inflated (at 0)	BEOI()	Logit	Log	Logit	–
Beta Inflated (at 1)	BEZI()	Logit	Log	Logit	–
Beta Inflated (at 0 and 1)	BEINF()	Logit	Logit	Log	Log
Box-Cox (Cole & Green)	BCCG()	Identity	Log	Identity	–
Box-Cox Power Exponential	BCPE()	Identity	Log	Identity	Log
Box-Cox t	BCT()	Identity	Log	Identity	Log
Exponential	EXP()	Log	–	–	–
Exponential Gaussian	exGAUS()	Identity	Log	Log	–
Exponential Gen. Beta type II	EGB2()	Identity	Identity	Log	Log

A notação para descrever o uso dos modelos acima em $\mathcal{D}$ é dada por:

$$y \sim \mathcal{D}\left\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p\right\}, \tag{2.11}$$

, em que $\theta_1, ..., \theta_p$ são os parâmetros de $\mathcal{D}$, $g_1, ..., g_p$ são as funções de ligação e $t_1, ..., t_p$ são as fórmulas dos modelos para os temos explanatórios e/ou efeitos aleatórios nos preditores $\eta_1, ..., \eta_p$, respectivamente.

2.6 Observações finais

Neste capítulo, foram apresentados os conceitos e tecnologias que subsidiaram a construção do modelo OLLST de modo que, para o leitor, seja possível compreender os conceitos tratados no decorrer da pesquisa. Este trabalho tem como foco o estudo do modelo OLLST na análise de dados longitudinais, para tal, é necessário entender os conceitos que permeiam este cenário, tais como: analise/modelo de regressão, medida repetida no tempo, entre outros.

Sob esse panorama, os modelo Modelos Aditivos Generalizados de Localização, Escala e Forma (GAMLSS) tem fundamental importância, pois serviam de base para construção da distribuição descrita no capítulo 3, de modo que os conceitos apresentados até agora se culminem em uma nova pesquisa.

Capítulo 3

Materiais e Métodos

3.1 Modelo de regressão OLLST

O modelo Odd Log-logística Skew t-Student, indicado neste trabalho por OLLST, é resultado da extensão da ST, de modo que, dada a função $W[G(x)]$, onde $g(x)$ e $G(x)$ é a função densidade e distribuição de densidade, respectivamente, da Skew t-Student descrita por:

$$W[F(x)] = \frac{G(x, \mu; \sigma)}{1 - \bar{G}(x, \mu; \sigma)}, \tag{3.1}$$

foi possível, baseado na transformação de $T - X$, propor um novo modelo para família de distribuições GAMLSS cuja a função de distribuição é dada por:

$$F(x; \mu, \sigma, \lambda, \alpha, \nu) = \int_0^{W[G(x)]} \frac{\alpha \times s^{\alpha-1}}{(\alpha - s^{\alpha})^2} ds = \frac{G(x)^{\alpha}}{G(x)^{\alpha} + [1 - G(x)]^{\alpha}}. \tag{3.2}$$

A função de densidade de probabilidade é descrita como:

$$f(x; \mu, \sigma, \lambda, \alpha, \nu) = \frac{\alpha \times G(x)^{\alpha-1} g(x)[1 - G(x)]^{\alpha-1}}{\{G(x)^{\alpha} + [1 - G(x)]^{\alpha}\}^2}. \tag{3.3}$$

A funções própria para o parâmetro $\boldsymbol{\alpha}$ é definida por:

$$\alpha = \frac{log\left(\frac{F(x)}{1-F(x)}\right)}{log\left(\frac{G(x)}{1-G(x)}\right)}.$$

tendo ainda a função quantílica, dada por $Q(\alpha, \mu)$ em função de x representada como

segue:

$$Q(\alpha, u) = G(x)^{-1}\left[\frac{u^{\frac{1}{\alpha}}}{(1-u)^{\frac{1}{\alpha}} + u^{\frac{1}{\alpha}}}\right]. \tag{3.4}$$

A flexibilidade deste novo modelo reside na restrição dos parâmetros α e λ, e consequente variação de ν, de modo que:

1. Quando $\alpha \neq 1$ e $\lambda \neq 0$:
   - $\nu = 1$ (OLLST Cauchy);
   - $\nu = 4 - 20$ (OLLST);
   - $\nu \geq 20$ (OLLST).
2. Quando $\alpha = 1$ e $\lambda \neq 0$:
   - $\nu = 1$ (ST Cauchy);
   - $\nu = 4 - 20$ (ST);
   - $\nu \geq 20$ (SN).
3. Quando $\alpha = 1$ e $\lambda = 0$:
   - $\nu = 1$ (t-Student Cauchy);
   - $\nu = 4 - 20$ (t-Student);
   - $\nu \geq 20$ (Normal).

De maneira visual, a figura 3.1 apresenta a variação do comportamento da função de densidade OLLST na medida que os parâmetros mudam seu valor. Em ambas as imagens, os valores para μ e σ foram fixados em 0 e 1, respectivamente. Enquanto na figura 3.1(a) $\lambda = 2$ e $\nu = 13$, com variação no valor de α, conforme a legenda, na figura 3.1(b) α foi fixado em $0,2$ e $\nu = 13$ variando somente o valor para λ.

O modelo foi codificado em $\mathcal{R}$ podendo ser utilizado pela biblioteca `gamlss` passando o valor "OLLST"como parâmetro de escolha da distribuição *(Veja o apêndice A)*.

A estimação dos parâmetros do modelo fica a cargo do método de estimação de máxima verossimilhança, mostrados na seção 3.1.1, visto que não há necessidade de uma suposição, ou quando há é mínima, da função de densidade e probabilidade. Além disso, o modelo considera ainda o uso de quatro parâmetros, com a justificativa de ser amplo o suficiente para descrever os fenômenos no mundo real e possibilidade de ajustar-se mais

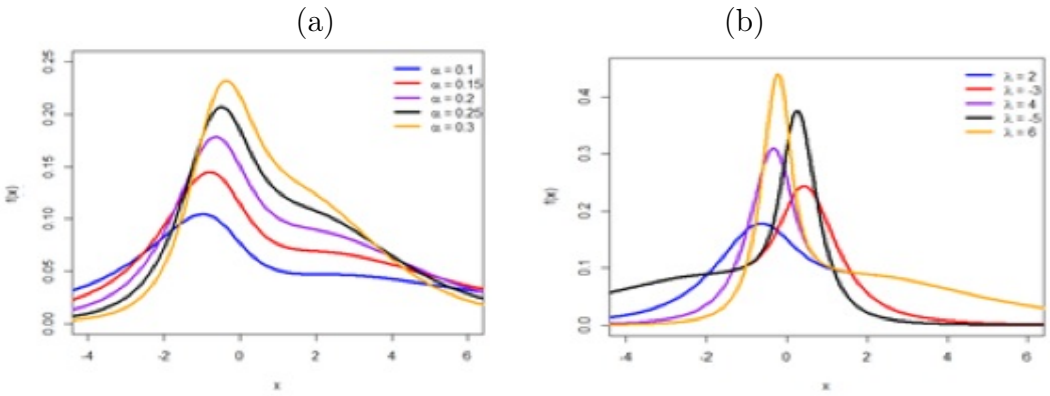

Figura 3.1: Variação da densidade do OLLST.

adequadamente à assimetria e bimodalidade à direita e/ou à esquerda (FERNANDES, 2021).

3.1.1 Estimação

A estimação de parâmetros para distribuição OLLST com efeitos aleatórios para dados longitudinais é feita através do método da máxima verossimilhança, onde se maximiza o logaritmo da função de verossimilhança marginal, por meio da integral da função de verossimilhança em relação aos efeitos aleatórios w_i. No i-ésimo individuo o vetor da variável resposta pode ser expresso por $Yn_i = [Y_{i11}, \ldots, Y_{i1K}; Y_{i21}, \ldots, Y_{i2K}; \ldots; Y_{in_i1}, \ldots, Y_{in_iK}]^T$ e função de verossimilhança condicional dos efeitos aleatórios (independência dentro dos indivíduos) para esse individuo pode ter a seguinte forma:

$$L_i(y_{ijk}|w_i) = \prod_{j=1}^{n_i} \prod_{k=1}^{K} f(y_{ijk}|w_i), \tag{3.5}$$

onde o $f(y_{ijk}|w_i)$ é a função de densidade e dada por por (3.6).

$$f(x; \mu, \sigma, \lambda, \alpha, \nu) = \frac{\alpha \times G(x)^{\alpha-1} g(x)[1 - G(x)]^{\alpha-1}}{\{G(x)^{\alpha} + [1 - G(x)]^{\alpha}\}^2}. \tag{3.6}$$

Nesse contexto, considerando a independência das variáveis aleatórias w_i e Y_{ijk} a contribuição do i-ésimo individuo para a função de verossimilhança marginal é:

$$\int L_i(y_{ijk}|w_i)\, f(w_i; \sigma^2_{\beta_0}) dw_i, \tag{3.7}$$

sendo $f(w_i; \sigma^2_{\beta_0})$ a densidade normal para efeitos aleatórios definidos pela equação 3.7 e

$L_i(y_{ijk}|w_i)$ é expressa por:

$$L(\theta) = \frac{1}{\sigma_{\beta_0}\sqrt{2\pi}} \prod_{j=1}^{n_1} \prod_{k=1}^{K} f(y_{ijk}|w_i) exp\left\{ -\frac{1}{2}\left(\frac{w_i}{\sigma_{\beta_0}}\right)\right\} dw_i, \tag{3.8}$$

Considerando um conjunto de dados $(y_{11k}, \mathrm{x}_{11k}), \ldots, (y_{1n_ik}, \mathrm{x}_{1n_ik}),, \ldots, (y_{m1k}, \mathrm{x}_{m1k}),$ $\ldots, (y_{mn_ik}, \mathrm{x}_{mn_ik})$ de n observações onde $n = (n_1 + \ldots + n_i, \mathrm{x}_{ijk})$ é o vetor de covariáveis associados ao i-ésimo individuo, na j-ésima repetição de indivíduo e k-ésima repetição no tempo, o logaritmo da função de verossimilhança marginal em (3.8) pode ser descrito como (3.9):

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{m} \log \left\{ \frac{\hat{\sigma} 2\sqrt{2}}{\sigma_{\beta_0}\sqrt{\pi}} \sum_{p=1}^{q} v_p^+ \prod_{j=1}^{n_i} \prod_{k=1}^{K} \frac{\alpha\, \phi\left(z_{ijk}\right) \Phi\left(\lambda z_{ijk}\right) \Phi_{SN}^{\alpha-1}\left(z_{ij};\lambda\right) \left[1 - \Phi_{SN}\left(z_{ijk};\lambda\right)\right]^{\alpha-1}}{\sigma\left\{ \Phi_{SN}^{\alpha}\left(z_{ijk};\lambda\right) + \left[1 - \Phi_{SN}\left(z_{ijk};\lambda\right)\right]^{\alpha} \right\}^2} \right. \\ &\quad \left. \times \exp\left\{ -\frac{1}{2}\left(\frac{s_p^+}{\sigma_{\beta_0}}\right)^2 \right\} \right\} \end{aligned} \tag{3.9}$$

em que $z_{ijk} = (y_{ijk} - \mu_{ijk})/\sigma$.

Vale ressaltar que existe uma dificuldade na estimação de parâmetros pelo método de máxima verossimilhança e esta reside na avaliação das integrais da função de verossimilhança. Neste método, ao se maximizar o $l(\theta)$ o pesquisador pode se deparar com m integrais de efeitos aleatórios w_i para variável, cuja a solução analítica é inviável na maioria dos casos. Em contrapartida, o uso de métodos de integração numérica pode representar a solução esperar para contornar está dificuldade. Para este trabalho, considerou-se o método de integração numérica por quadratura de gauss-hermite (Código $\mathcal{R}$ no apêndice 3.1.3).

3.1.2 Quadratura de Gauss

A maioria dos métodos usados para calcular a integral de uma função utilizam técnicas de aproximação do valor da integral por meio um polinômio em uma ou mais regiões de interesse. A depender da complexidade da função, o cálculo pode ser fatorado em uma função de pesos a fim de entregar o valor que mais se aproxima do real (GOLUB; WELSCH, 1969).

As quadraturas gaussianas se enquadram nesse contexto, pois, apesar de apresentar intervalos irregulares e possíveis zonas de agrupamentos, também fazem uso de métodos por aproximação polinomial de grau crescente, onde os nós (nodos) $x_i, i = (1, 2, 3..., n)$ da quadratura, são as raízes do polinômios $P_n(x)$ para todo ponto menor que n (SCHWARTZ, 1996).

A ideia principal por trás dos métodos da quadratura de gauss é substituir a soma integral,

por uma soma discreta, da seguinte forma:

$$\int_b^a f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i), \tag{3.10}$$

onde n é o número de pontos, *F(x)* refere-se a função no ponto de integração, $f(x_i)$ é o valor da coordenada para o ponto de integração e ω_i é a função peso, também para o ponto de integração (SILVA, 2017).

Ao substituir a soma integral pela soma discreta na Quadratura de Gauss, cria-se uma função peso que tem por objetivo distribuir o valor aproximado e o valor exato da integral dentro do intervalo em análise, considerando, claro, que a função $F(x)$ seja conhecida. Na Quadratura de Laguerre (3.11) de n pontos, por exemplo, os nós da quadratura x_i são as raízes do n-ésimo polinômio de Laguerre. (BARBOSA; LOEFFLER; BULCÃO,)

$$\int_{-1}^{1} f(x)dx = \omega_1 f(x_1) + \omega_2 f(x_2) + \omega_3 f(x_3) + ... + \omega_i f(x_i) = \sum_{i=1}^{n} \omega_i f(x_i) \tag{3.11}$$

Entre os métodos de integração que se vale dos conceitos de quadratura de gauss, destacam-se:

- Fourier;
- Legendre;
- Chebyshev;
- Laguerre; e
- Hermite *(escolhida para uso e aplicação neste trabalho).*

3.1.3 Quadratura de Gauss-Hermite

O método de integração numérica por Gauss-Hermite apresenta duas versões com bases similares, mas os conceitos e aplicações diferentes, sendo elas: não adaptativa e adaptativa. Seu conceito é a extensão da quadratura Gaussiana sendo expresso por 3.12

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} exp\{-x^2\}dx \approx \sum_{k=1}^{q} v_k f(s_k) \tag{3.12}$$

onde s_k são as raizes do polinômio de Hermite de grau q e v_k são os pesos associados aos números de pontos da quadratura e do polinômio de Hermite $H_{q-1}(x)$ avaliado em s_k.

Nas condições em que $g(x)$ é um polinômio de grau $2q-1$ é possível obter o valor exato da integral por meio de $\sum_{k=1}^{q} v_k f(s_k)$ o que leva ao entendimento que a aproximação do valor da integral está associado ao número de pontos da quadratura.

A base da quadratura de Gauss-Hermite gira em torno da função de peso $v(x)$ que, na maioria dos cálculos de integração, pode ser representada pela função de densidade e probabilidade de uma variável aleatória, onde $g(x)$ é o produto de $v(x)$ com uma outra função $gf(x)$ 3.13 podendo $v(x)$ 3.14 se expressa como:

$$g(x) = v(x)f(x), \tag{3.13}$$

$$v(x) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{3.14}$$

contexto em que, sendo $z = (x-\mu)/\sqrt{2\sigma^2}$, a identidade da função pode ser representada por:

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} f(x)v(x)dx = \int_{-\infty}^{+\infty} \pi^{-1/2} f(\sqrt{2}\sigma z + \mu) \exp\left\{-z^2\right\} dz \tag{3.15}$$

e calculada por Gauss-Hermite da seguinte forma 3.16:

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} f(x)v(x)dx \approx \sum_{k=1}^{q} \frac{v_k}{\sqrt{\pi}} f(\sqrt{2}\sigma s_k + \mu) \tag{3.16}$$

Observação 2.2.1: *Em $\mathcal{R}$, existem funções derivados do pacote **GHQp** que montam os vetores de pesos e pontos para métodos de integração por gauss, aceitando como parâmetros o número de pontos e o método usado. (Código $\mathcal{R}$ no apêndice H).*

Considerando 3.17

$$H_q(x) = (-1)^q \exp\left\{x^2\right\} \frac{d^q}{dx^q} (\exp)\left\{-x^2\right\} \tag{3.17}$$

como sendo o polinômio de Gauss-Hermite de grau q e

$$v_k = \frac{2^{q+1} q! \sqrt{\pi}}{\left[H_q'(x)\right]^2} s \tag{3.18}$$

o vetor de pesos desse polinômio, ao aplicar em um exemplo da quadratura de Gauss-Hermite com polinômio de dois pontos, ou seja, $q = 2$ na integral

$$\int_{-\infty}^{+\infty} \exp\left\{-x^2\right\} x^2 dx \tag{3.19}$$

, observa-se que:

$$\int_{-\infty}^{+\infty} \exp\left\{-x^2\right\} x^2 dx \approx \frac{\sqrt{\pi}}{2}\left[f\left(\frac{\sqrt{2}}{2}\right) + f\left(-\frac{\sqrt{2}}{2}\right)\right], \tag{3.20}$$

desse modo $f(x) = x^2$, $H_q(x) = 4x^2 - 1$, os zeros do polinômios são $x_1 \models +\sqrt{2}/2$ e $x_2 = -\sqrt{2}/2$ e $v_1 = v_1 = \frac{\sqrt{\pi}}{4(\sqrt{2}/2)^2}$, valor exato da integral 3.19 é $\frac{\sqrt{\pi}}{2}$, como mostra a tabela 3.1 para polinômios com dois pontos de quadratura.

Tabela 3.1: Valores para $H_q(x)$ e v_k para $q = 2, 3e4$

q	2	3	4
$H_q(x)$	$4x^2 - 1$	$8x^3 - 12x$	$16x^4 - 12x^2 + 12$
v_k	$\frac{\sqrt{\pi}}{4x^2}$	$\frac{96\sqrt{\pi}}{(24x^2-12)^2}$	$\frac{768\sqrt{\pi}}{(64x^3-96x)^2}$

Deve-se considerar, no entanto, que a quadratura de Gauss-Hermite tem maior efetividade no cálculo de integrais do tipo

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} f(x)v(x)dx \tag{3.21}$$

em que:

$$v(x) = v(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \tag{3.22}$$

Por exemplo, ao desenhar a linha para $g(x) = exp(-(x-1)^2)$ e posicionar os pontos e pesos para esse tipo de integral, é possível visualizar a distribuição dos pontos ao longo da região de interesse, ou seja, dentro do intervalo da função de densidade e probabilidade como mostra a figura 3.2:

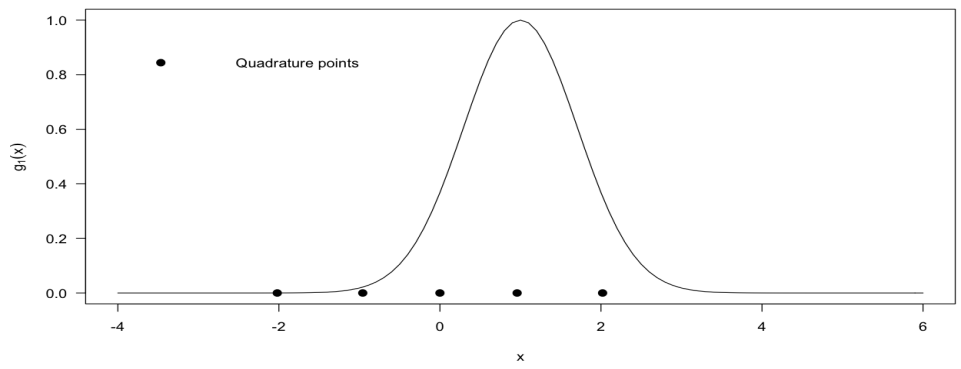

Figura 3.2: (a) - *(Código $\mathcal{R}$ no apêndice I)*

Apesar dos pontos de quadratura não estarem perfeitamente alinhados com a área de maior densidade do gráfico, ainda assim o resultado obtido ao se calcular a integral para função $g(x) = exp(-(x-1)^2)$ está próxima do valor real para ela, como visto abaixo:

```
> sum(quad$weights * g1(quad$nodes) * exp(quad$nodes^2))
[1] 1.771348
> integrate(f=g1, lower=-Inf, upper=Inf)
1.772454 with absolute error < 1.6e-06
```

Para se trabalhar sob essa perspectiva, alguns pontos devem ser observados para garantir maior acurácia nos resultados, são eles:

- Os pontos e pesos devem ser facilmente encontrados e calculados;
- Deve possuir uma função de densidade de probabilidade normal;
- O polinômio de Gauss-Hermite deve ser de baixa ordem;
- Os pontos devem cair na região de interesse; e
- Os pesos devem ser todos positivos.

Como a função $g(x)$, que dita o comportamento da curva, não exerce influência sobre a distribuição dos pontos de quadratura, é provável que, dependendo do conjunto de dados, a distribuição e posicionados dos pontos afastem-se da região de interesse da função que se deseja aproximar, como é o caso das bimodalidades e assimetrias.

Em um cenário que $g(x) = exp(-5 * (x-3)^2)$, desenhando uma assimetria na distribuição, os pontos de quadratura ficam totalmente fora da região de interesse da função, como mostra a figura 3.3.

Em dados bimodais, mesmo que os pontos caiam sobre a região de interesse da função, é importante considerar o comportamento e a região de agrupamento dos dados dessa bimodalidade 3.4. Em ambos os casos, ou seja, tanto na assimetria quanto na bimodalidade, o valor do cálculo da integral é afetado, produzindo resultados imprecisos.

```
> sum(quad$weights * g3(quad$nodes) * exp(quad$nodes^2))
[1] 0.009721331
> integrate(f=g3, lower=-Inf, upper=Inf)
0.7926655 with absolute error < 0.00011
```

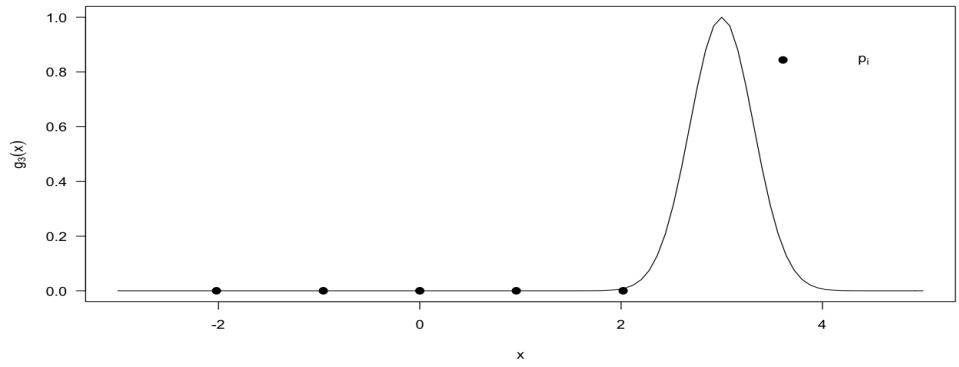

Figura 3.3: (b) - *(Código $\mathcal{R}$ no apêndice J)*

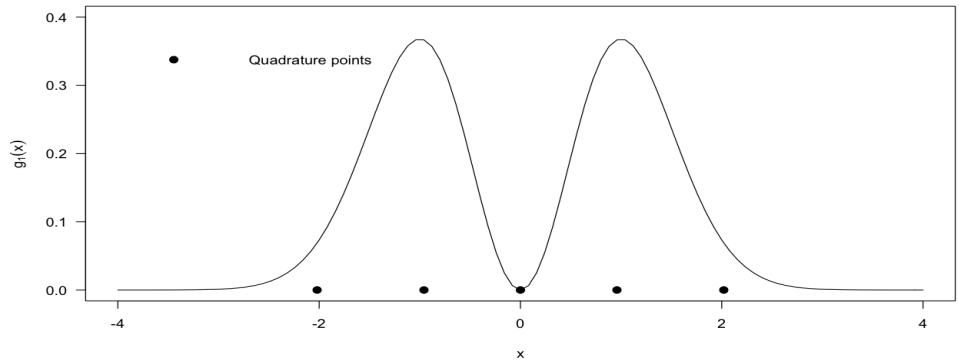

Figura 3.4: (c) - *Código $\mathcal{R}$ no apêndice K*

```
> sum(quad$weights * g2(quad$nodes) * exp(quad$nodes^2))
[1] 0.8862269
integrate(f=g2, lower=-Inf, upper=Inf)
0.8862269 with absolute error < 1.1e-06
```

A maneira mais adequada de se tratar essas limitações é a aplicação do conceito adaptativo da quadratura de Gauss-Hermite, que seria o reescalamento e modificação dos ponto de modo que eles fiquem, em sua maioria, na região de interesse da função.

Considerando a equação 3.12, no modelo adaptativo os parâmetros μ e σ, chamados aqui de $\hat{\mu}$ e $\hat{\sigma^2}$, são respectivamente a média e a variância da distribuição, dados por:

$$\hat{\mu} = \arg\max_{x} g(x) \tag{3.23}$$

e

$$\hat{\sigma^2} = \left[-\frac{d^2}{dx^2} \log g(x) \right]^{-1} |_{x=\hat{\mu}}, \tag{3.24}$$

de modo que se for definida

$$h(x) = \frac{g(x)}{v(x; \hat{\mu}, \hat{\sigma}^2)}. \tag{3.25}$$

A função pode ser reescrita de modo que

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} h(x)x(x; \hat{\mu}, \hat{\sigma}^2)dx, \tag{3.26}$$

onde $v(x; \hat{\mu}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \left\{ -\frac{1}{2} \frac{(x-\hat{\mu})^2}{\hat{\sigma}^2} \right\}$, de modo a aplicação da Gauss-Hermite sobre ela 3.26 produz a seguinte equação:

$$\int_{-\infty}^{+\infty} g(x)dx \approx \sum_{k=1}^{q} \frac{v_k}{\sqrt{\pi}} h(\sqrt{2}\hat{\sigma}s_k + \hat{\mu}) = \sqrt{2\pi\hat{\sigma}} \sum_{k=1}^{q} v_k^{\dagger} g(s_k^{\dagger}), \tag{3.27}$$

de modo que o comportamento dos pontos seguem a distribuição dos dados, como mostra a figura 3.5

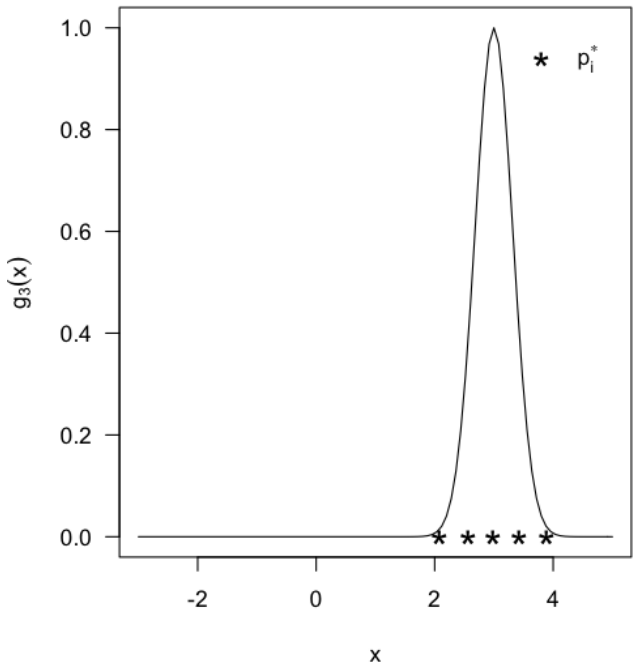

Figura 3.5: Pontos de quadratura sobre dados assimétrico - Gauss-Hermite adaptativa

3.2 OLLST com efeitos aleatórios

A estrutura do modelo hierárquico para OLLST é dada por:

$$Y_{ijk}|w_i \overset{iid}{\sim} OLLST(\mu_{ij}, \sigma, \lambda, \alpha) \text{ e } w_i \overset{iid}{\sim} N(0, \sigma^2_{\beta_0}), \tag{3.28}$$

onde Y_{ijk} é a variável resposta associada ao i-ésimo indivíduo $(i = 1, ..., m)$ para $j = (1, ..., J)$ repetições dos tratamentos e para $k = (1, ..., K)$ repetições no tempo, de modo que o i-ésimo indivíduo pode ser descrito pelo vetor de variáveis respostas $Y_i = [Y_{i11}, ..., Y_{i1K}; Y_{i21}, ..., Y_{i2K}, ..., Y_{in_i1}, ..., Y_{in_iK}]$

É necessário considerar também, a possibilidade de relação entre as variáveis respostas e explicativas, ou mesmo covariáveis. Nesse contexto, é correto pressupor que todas as respostas para o mesmo individuo possam ter efeitos aleatórios comuns, denotado aqui por w_i, e que esse efeito pode ser considerado como variáveis aleatórias não observadas. Nesse caso, o modelo de regressão para dados com característica de correlacionamento pode ser expresso por:

$$Y_{ijk} = \beta_0 + w_i + \underbrace{\sum_{k=1}^{3} \omega_k d_k}_{\text{efeitos longitudinais}} + \underbrace{\sum_{i=2}^{4}\sum_{k=1}^{3} \delta_{ik} d_k u_i}_{\text{efeitos hierárquicos}} + \sigma z_{ijk}, \tag{3.29}$$

sendo δ_{ik} os desdobramentos dos efeitos de cada tratamento dentro de cada tempo ω_i, apresentando efeito aleatório para o individuo i de variável aleatória z_{ijk} com distribuição OLLST. A relação de ausência ou presença de para valores de tempo e tratamento fica a cargo das variáveis dummies d_k u_k com a seguinte regra:

$$d_k = \begin{cases} 0, \text{se a época} \neq k \\ 1, \text{se a época} = k \end{cases} \tag{3.30}$$

$$u_k = \begin{cases} 0, \text{se o tratamento} \neq k \\ 1, \text{se o tratamento} = k \end{cases} \tag{3.31}$$

Sendo $f(y_{ijk}|w_i)$ a função de densidade condicional para w_i, de modo que $Y_{ijk}|w_i$ tem a função de densidade conjunta dada por $f(y_{ijk}, w_i; \theta) = f(y_{ijk}|w_i) f(w_i \sigma^2_{\beta_0})$, é possível obter as seguintes estruturas:

1. $Y_{ijk}|w_i \sim OLLST(\mu_{ij}, \sigma, \lambda, \alpha)$ tem a função de densidade condicional expressa por:

$$f(y_{ijk}|w_i) = \frac{\alpha\phi(\frac{y_{ijk}-\mu_{ijk}}{\alpha})\Phi^{\alpha-1}\left(\frac{y_{ijk}-\mu_{ijk}}{\sigma}\right)\left[1-\Phi\left(\frac{y_{ijk}-\mu_{ijk}}{\sigma}\right)\right]^{\alpha-1}}{\sigma\left\{\Phi^{\alpha}(\frac{y_{ijk}-\mu_{ijk}}{\alpha}) + \left[1-\Phi(\frac{y_{ijk}-\mu_{ijk}}{\alpha})\right]^{\alpha}\right\}^2} \tag{3.32}$$

2. E equação 3.33 é a indicação de que o modelo de regressão será posto no parâmetro de

forma.

$$\mu_{ijk} = \beta_0 + w_i + \sum_{k=1}^{3} \omega_k d_k + \sum_{i=2}^{4} \sum_{k=1}^{3} \delta_{ik} d_k u_k \tag{3.33}$$

3. A distribuição de efeitos aleatórios é definida por $w_i \sim N(0, \sigma^2_{\beta_0})$ e densidade:

$$f(w_i; \sigma^2_{\beta_0}) = \frac{1}{\sigma_{\beta 0}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{w_i}{\sigma_{\beta_0}} \right)^2 \right\}, \tag{3.34}$$

sendo $w_i \in \mathbb{R}$ e $\sigma_{\beta_0} > 0$ o intercepto de efeitos aleatórios.

3.3 Envelope simulado

Como parte das etapas de diagnóstico do modelo, de modo a avaliar a veracidade da hipótese assumida para à distribuição da variável resposta (FERNANDES, 2019) em função do tempo, alguns gráficos de probabilidade foram gerados (envelope simulado) de banda obtidas no ajuste do modelo para diferentes conjuntos de dados sobre a variação do teor de carbono no solo.

Para construção do gráfico, considerou que o k-ésimo ($k = 1, ..., n$) valor ordenado do resíduos versos o valor correspondente esperado da estatística de ordem normal padrão. Assim, tem-se que:

$$\Phi^{-1} \frac{k + n - \frac{1}{8}}{2n + \frac{1}{2}} \tag{3.35}$$

onde $\Phi(.)$ é a função de distribuição acumulada n $N(0, 1)$.

A geração dos gráficos de probabilidade dos resíduos quantílicos foi feita através da função `halfnorm` da biblioteca `BLMM`.

3.4 Aplicação: análise dos níveis de carbono no solo

3.4.1 Material

Os dados utilizados para ajuste no modelo OLLST foram obtidos em 31 de maio de 2022 e são provenientes de um experimento realizado em Piracicaba, estado de São Paulo (SP), por (MARTINS, 2022) em sua dissertação de mestrado.

3.5 Contexto dos dados

Há hoje uma preocupação genuína quanto aos modelos agrícolas de produção de alimento, principalmente ao encarar o cenário desenhado pela ONU para os próximos anos. Os modelos mais conservacionistas, como por exemplo o Sistema de Plantio Direto, que potencialmente acumula matéria orgânica no solo, tendem a se destacar frente a necessidade crescente, sobretudo pela sua característica sustentável (MARTINS, 2022).

Estudos como (KAISER; KALBITZ, 2012; KALBITZ et al., 2000) ratificam essa tendência e pontuam os benefícios do acúmulo de resíduos culturais na produção de matéria orgânica do solo (MOS), bem como de matéria orgânica dissolvida (MOD ou COD), e seus efeitos na produção de carbono orgânico dissolvido (COD). Há ainda uma relação direta no transporte de nutrientes para microrganismos que, por consequência, está ligada a dinâmica de nutrientes do solo, tendo inclusive impactos a longo prazo, quando comparadas com fatores ambientais e antropogênicos.

Considerando a lacuna existente em pesquisas voltadas para solos tropicais, (MARTINS, 2022) propôs um experimento a fim de entender o fluxo de COD e a dinâmica dele no perfil de solo Brasileiro com o objetivo de incitar adequações nas práticas de manejo adequado do solo, mitigando perda de nutrientes e possíveis contaminações de corpos d'água, com olhar voltado aos sistemas agrícolas.

Com intuito de refinar o escopo desta pesquisa, uma análise exploratória foi realizada sob o conjunto de dados a fim de identificar aquele com maior dispersão na organização das unidades amostrais. As figuras **??** e **??** mostram os boxplots para os conjunto de dados utilizados por (MARTINS, 2022) que contém a variável "tempo/época".

Para que seja viável a análise de dados em um modelo hierárquico de uma variável resposta em função do tempo, tem-se a pressuposição de independência entre as observações para o mesmo indivíduo e dependência das unidades amostrais retiradas das unidades experimentais (CNAAN; LAIRD; SLASOR, 1997; FAUSTO et al., 2008), necessitando, por tanto, uma avaliação primária de cada conjunto.

3.5.1 Descrição dos Dados

Nesta pesquisa foram ajustados modelos de regressão para seis conjunto diferentes de dados sobre a dinâmica do carbono em camadas de Latossolo Vermelho (VL). Apesar de focar ná análise do carbono, (MARTINS, 2022) separou alguns conjuntos de dados para a fim de entender as influências que o carbono poderia sofre durante o experimento. Sete foram os conjuntos de dados, no entanto, um não dispunha da variável tempo. As nomenclaturas para cada conjunto seguirão como:

- **LV-C** - conjunto de dados para análise do teor de carbono em função do tempo e profundidade;
- **LV-CN** - conjunto de dados para análise da razão entre carbono e nitrogênio;
- **LV-COD** - conjunto de dados para análise do fluxo de carbono dissolvido;
- **LV-EC** - conjunto de dados para análise do teor de carbono em função do tempo;
- **LV-EN** - conjunto de dados para análise do teor de carbono em função do tempo; e
- **LV-N** - conjunto de dados para avaliar níveis de nitrogênio.

Cada conjunto de dados tem as seguintes variáveis:

- **Trat.** Variável categórica referente ao tipo de tratamento aplicado: **(M)** Palha de milho (Zea mays); **(S)** Palha soja (Glycine max); **(MS)** A combinação das palhas de soja e milho; **(SR)** O controle sem nenhum tipo de tratamento;
- **Epoca**[*sic*]. Variável categórica referente ao tempo de observação da unidade amostral. ($t1$, $t2$, $t3$); e
- **Y.** Variável resposta referente ao teor de carbono no solo para aquela unidade experimental.

O conjunto **LV-C** possui uma população de tamanho 144 sendo: 3 variações de época $t = (1, 2, 3)$, 4 tipos de tratamentos $k = (1, 2, 3, 4)$ e 3 profundidades $x = (010, 020, 030)$; o conjunto **LV-CN** possui uma população de tamanho 144 sendo: 3 variações de época $t = (1, 2, 3)$, 4 tipos de tratamentos $k = (1, 2, 3, 4)$ e 3 profundidades $x = (010, 020, 030)$; o conjunto **LV-COD** possui uma população de tamanho 176 sendo: 11 variações de época $t = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$ e 4 tipos de tratamentos $k = (1, 2, 3, 4)$; o conjunto **LV-EC** possui uma população de tamanho 48 sendo: 3 variações de época $t = (1, 2, 3)$ e 4 tipos de tratamentos $k = (1, 2, 3, 4)$; o conjunto **LV-EN** possui uma população de tamanho 48 sendo: 3 variações de época $t = (1, 2, 3)$ e 4 tipos de tratamentos $k = (1, 2, 3, 4)$; e o conjunto **LV-N** possui uma população de tamanho 144 sendo: 3 variações de época $t = (1, 2, 3)$, 4 tipos de tratamentos $k = (1, 2, 3, 4)$ e 3 profundidades $x = (010, 020, 030)$. Para todos os modelos foram realizadas 4 repetições das observações.

Vale ressaltar que para seu estudo, (MARTINS, 2022) considerou também a profundidade (em cm) para as amostras retiradas. No entanto, como o foco deste trabalho reside na análise de dados com medida repetida no tempo, está variável não foi incluída nos experimentos.

As figuras 3.6, 3.7 e 3.8 mostram a distribuição da frequência do carbono orgânico analisado por (MARTINS, 2022). Cada gráfico trás consigo uma linha pontilhada em azul que representa uma média para variável resposta, além de um mancha em vermelho para representar a zona de densidade mostrando a variação nas distribuições para cada conjunto.

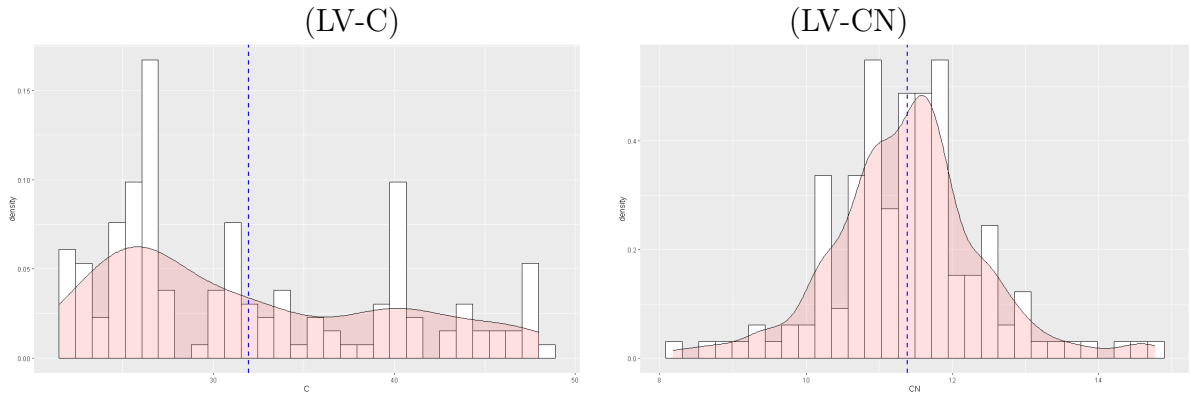

Figura 3.6: Concentração de carbono no solo.

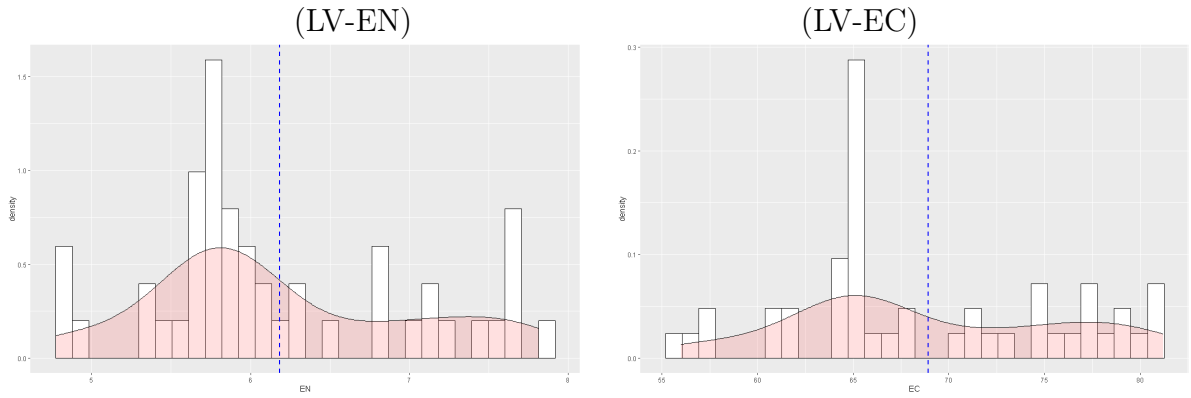

Figura 3.7: Concentração de carbono no solo.

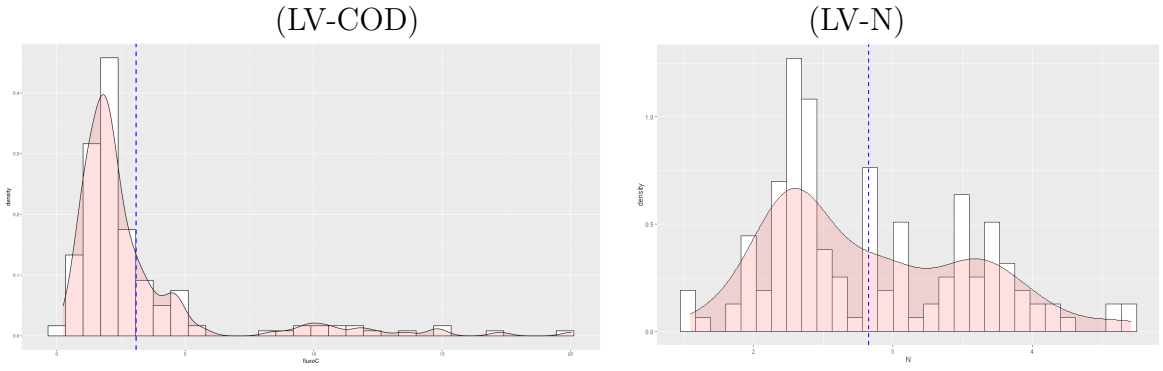

Figura 3.8: Concentração de carbono no solo.

3.6 Análise Exploratória

Considerando os valores para cada tipo de controle em razão da época, verificou-se que conjunto **LC-V**, conforme a figura 3.9, sugerem alterações nos valores para variável resposta em razão do tempo, justificando uma avaliação do modelo para cada tipo de controle (tratamento) dentro de suas respectivas épocas;

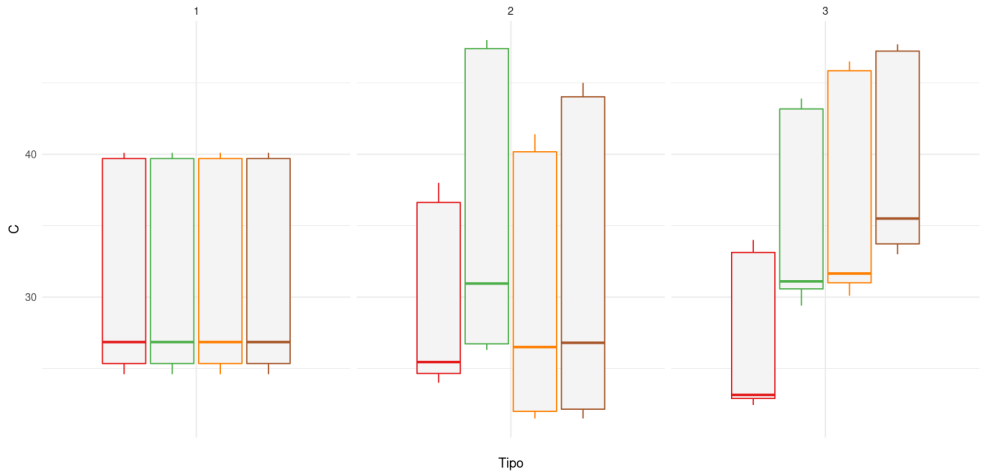

Figura 3.9: Boxplot para o conjunto LV-C

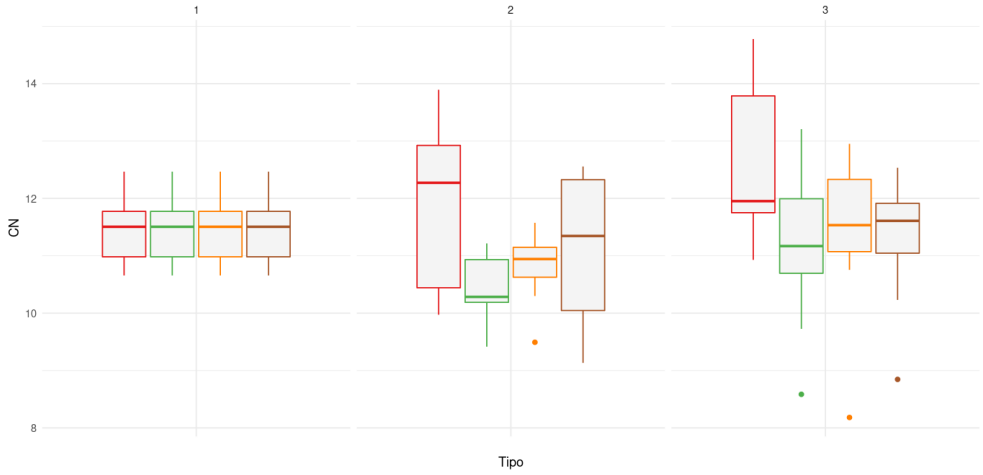

Figura 3.10: Boxplot para o conjunto LV-CN

Para o conjunto **LV-CN** o boxplot mostrado na figura 3.10 exibe a presença de valores extremantes para alguns tratamentos no segundo e terceiro momento da observação. *Outliers* também são observados nos conjuntos de dados **LV-COD** (figura 3.11), **LV-EC** (figura 3.12) e **LV-EN** (figura 3.13). Importante ressaltar que, segundo (BARNETT; LEWIS et al., 1994), o *outlier* é uma observação (ou conjunto delas) que aparenta ser inconsistente com o restante do conjunto de dados, com crucial importância para caracterização da amostra, uma vez que, na ausência deles o modelo pode apresentar um comportamento diferente. Em todos os casos, a interpretação sugere que, ainda que mínima, o tempo exerce influência no comportamento da variável resposta, justificando uma análise em segundo nível.

Assim como o conjunto **LV-C**, o **LV-N** mostrou relativo controle entre os valores da variável resposta, sem a presença de pontos extremos e tendência que sugerem variação no valor para os tipos de controle em razão do tempo, de modo a valer uma análise mais aprofundada da interação

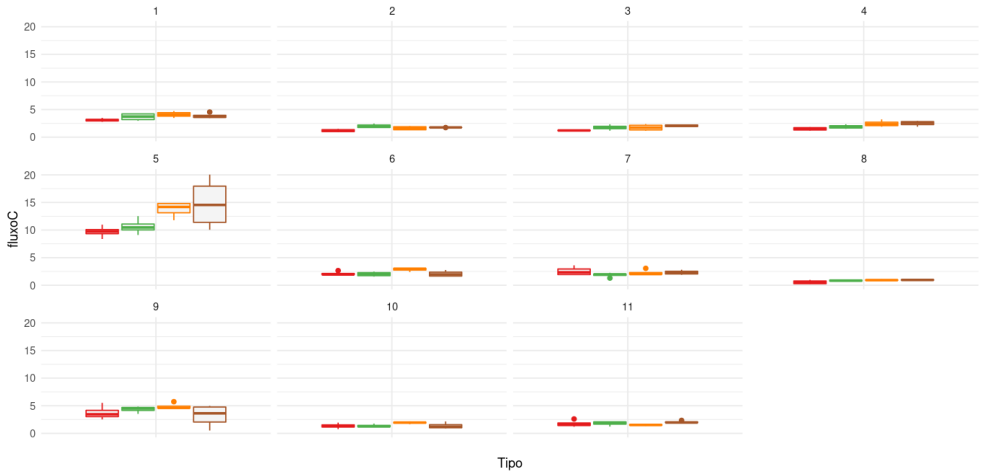

Figura 3.11: Boxplot para o conjunto LV-COD

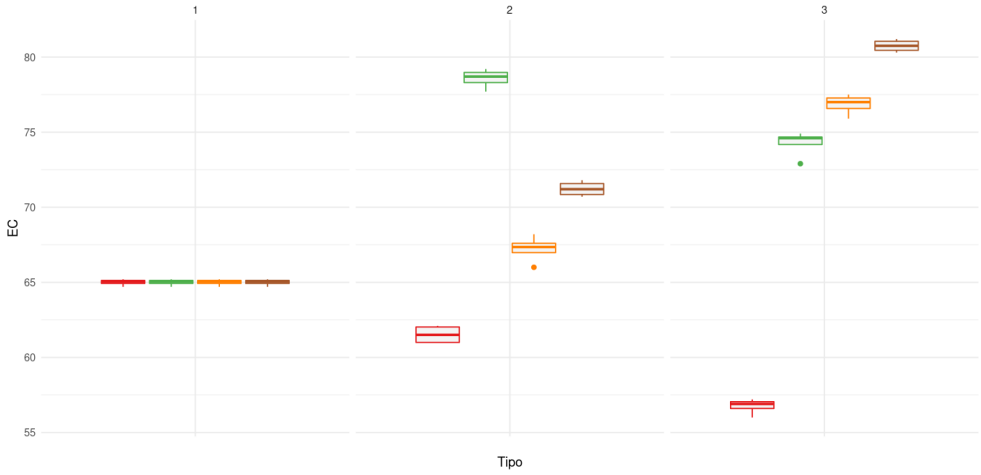

Figura 3.12: Boxplot para o conjunto LV-EC

entre as variáveis.

3.7 Tratamento de saída da biblioteca GAMLSS

Durante análise do modelo OLLST, incorporado a família de distribuições do GAMLSS, verificou-se uma característica da saída de resultados para o modelo ajustado que poderia representar uma dificuldade na interpretação. A implementação computacional do modelo descrito na equação 3.29 implica em uma saída hierarquizada onde os tratamentos são comparados dentro de cada dimensão (tempo).

Supondo que se pretende ajustar um modelo de regressão para dados com medida repetida no tempo, cuja a estrutura da tabela seja semelhante ao mostrado na tabela 2.1. Ao se descrever computacionalmente o modelo da equação 3.29, perceber-se que cada comparação entre

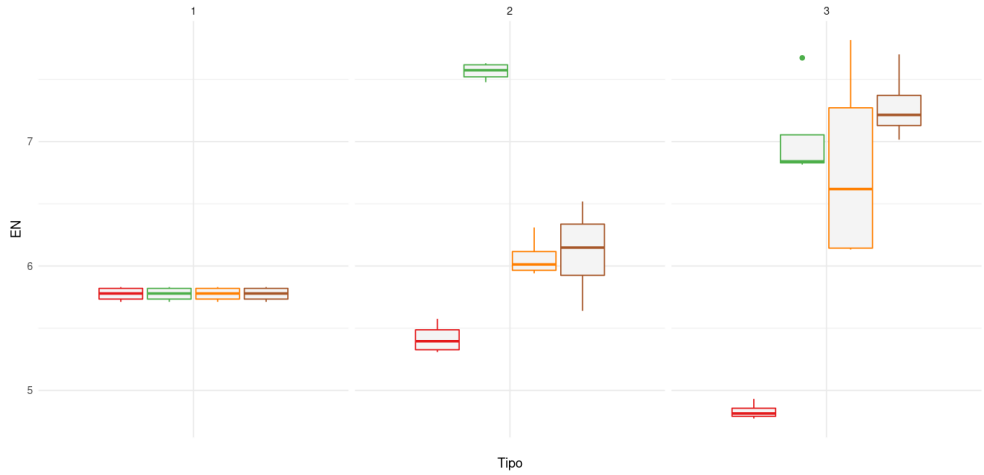

Figura 3.13: Boxplot para o conjunto LV-EN

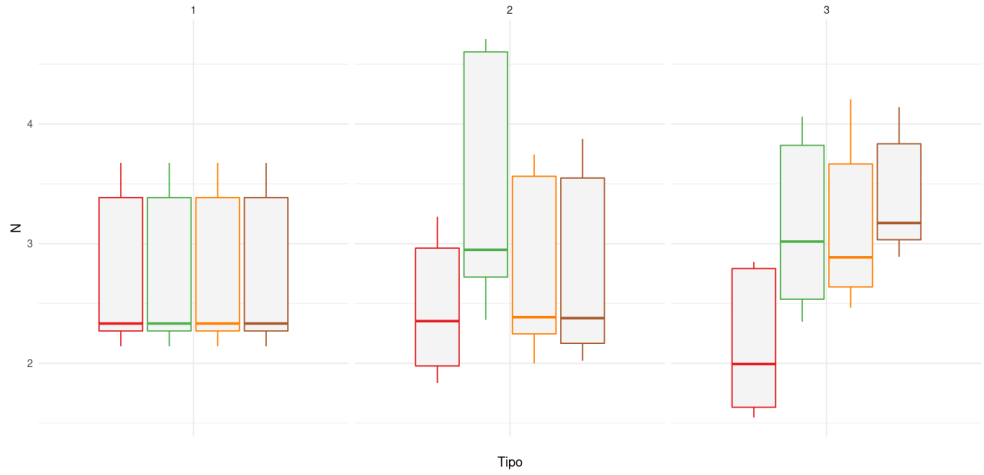

Figura 3.14: Boxplot para o conjunto LV-N

indivíduos/tratamentos fica subordinada a variável **tempo** *(Vide a figura 3.15)*.

Outro fator que deve ser considerado neste cenário está relacionado a indicação daquilo que deve ser comparado. Em uma estrutura com k indivíduos/tratamentos, cada qual deve passar pela mesma rotina de comparação, ou seja, o dispositivo apresentado na figura 3.15 será repetido em $\sum_{i=1}^{k}$, e cada interação terá sua própria saída de dados.

Na postulação do modelo hierárquico a identificação do indivíduo/tratamento em restrição, ou seja, aquele que deverá ser comparado com os demais para cada lâmina do tempo, não fica evidenciado na saída dos resultados. Nesta ocasião, o pesquisador deve inferir qual indivíduo/tratamento está em restrição, como mostra a figura 3.16, para somente então poder vislumbrar o panorama comparativo completo.

Ademais, a indicação manual para cada restrição pode trazer uma camada extra de dificul-

Figura 3.15: Representação do modelo em função do tempo

```
Data: Data
Log-likelihood: -358.799
Fixed: fix.formula
                (Intercept)                      epoca1                      epoca2                      epoca3                      epoca4                      epoca6
               -7.471209e+01                3.723542e+00                1.420915e+01                3.939775e+01                8.067589e+01                1.055474e+02
                      epoca8                     epoca12                     epoca16 epoca0:tratamento.removidoCMET epoca1:tratamento.removidoCMET epoca2:tratamento.removidoCMET
                1.121057e+02                1.140881e+02                1.123746e+02                2.158991e-14                4.355242e+00                8.741959e+00
epoca3:tratamento.removidoCMET epoca4:tratamento.removidoCMET epoca6:tratamento.removidoCMET epoca8:tratamento.removidoCMET epoca12:tratamento.removidoCMET epoca16:tratamento.removidoCMET
                5.398011e+00               -9.406483e+00                3.267758e+00                5.386620e+00                5.388409e+00                3.980695e+00
epoca0:tratamento.removidoCPET epoca1:tratamento.removidoCPET epoca2:tratamento.removidoCPET epoca3:tratamento.removidoCPET epoca4:tratamento.removidoCPET epoca6:tratamento.removidoCPET
                6.967220e-15                2.128774e+01                6.150396e+01                5.868465e+01                2.580739e+01                5.508918e+00
epoca8:tratamento.removidoCPET epoca12:tratamento.removidoCPET epoca16:tratamento.removidoCPET epoca0:tratamento.removidoCVET epoca1:tratamento.removidoCVET epoca2:tratamento.removidoCVET
                9.352206e-01               -5.423993e-01               -9.902042e-01                8.979414e-16                5.775314e+01                8.086089e+01
epoca3:tratamento.removidoCVET epoca4:tratamento.removidoCVET epoca6:tratamento.removidoCVET epoca8:tratamento.removidoCVET epoca12:tratamento.removidoCVET epoca16:tratamento.removidoCVET
                6.726204e+01                2.992528e+01                7.454004e+00                3.595779e+00                9.462047e-01               -3.629438e+00
epoca0:tratamento.removidoMBET epoca1:tratamento.removidoMBET epoca2:tratamento.removidoMBET epoca3:tratamento.removidoMBET epoca4:tratamento.removidoMBET epoca6:tratamento.removidoMBET
               -3.699819e-14                2.354966e+01                3.917777e+01                3.682512e+01                1.269459e+01               -2.902806e+00
epoca8:tratamento.removidoMBET epoca12:tratamento.removidoMBET epoca16:tratamento.removidoMBET epoca0:tratamento.removidoMRET epoca1:tratamento.removidoMRET epoca2:tratamento.removidoMRET
               -6.542411e+00               -7.579658e+00               -9.245256e+00                1.347898e-14                1.099422e+00                2.199620e+01
epoca3:tratamento.removidoMRET epoca4:tratamento.removidoMRET epoca6:tratamento.removidoMRET epoca8:tratamento.removidoMRET epoca12:tratamento.removidoMRET epoca16:tratamento.removidoMRET
                3.657955e+01                1.500094e+01                4.396135e-01               -2.556392e+00               -3.193821e+00               -2.583343e+00
epoca0:tratamento.removidoPPET epoca1:tratamento.removidoPPET epoca2:tratamento.removidoPPET epoca3:tratamento.removidoPPET epoca4:tratamento.removidoPPET epoca6:tratamento.removidoPPET
                3.955916e-15               -1.106361e+00               -3.781922e+00               -9.997553e+00               -2.096335e+01               -2.117611e+00
epoca8:tratamento.removidoPPET epoca12:tratamento.removidoPPET epoca16:tratamento.removidoPPET epoca0:tratamento.removidoQTET epoca1:tratamento.removidoQTET epoca2:tratamento.removidoQTET
               -4.072774e-01               -1.040088e+00               -3.645531e+00               -1.196849e-15                5.720228e+01                7.400631e+01
epoca3:tratamento.removidoQTET epoca4:tratamento.removidoQTET epoca6:tratamento.removidoQTET epoca8:tratamento.removidoQTET epoca12:tratamento.removidoQTET epoca16:tratamento.removidoQTET
                5.648786e+01                1.998583e+01               -5.531079e+00               -8.237958e+00               -1.088136e+01               -1.391322e+01

Random effects:
 Formula: ~1 | grupo
        (Intercept) Residual
StdDev: 3.177704e-05        1

Variance function:
 Structure: fixed weights
 Formula: ~W.var
```

Figura 3.16: Saída do ajuste para determinado tratamento

dade para o pesquisador, caso o conjunto de dados tenha muitos tratamentos. Consequentemente, a automação desta rotina representaria uma significativa economia de tempo, além de minimizar a possibilidade de erros.

Por fim, deve-se considerar a pluralidade da estatística e suas aplicações em diversos contextos científicos. Desenvolvida com propósitos de manipulação e análise de dados (MARTINS, 2016) a linguagem de programação $\mathcal{R}$ pode parecer desafiadora para quem não teve contato com ela, mesmo para aqueles que já se aventuraram no munda da programação.

Em $\mathcal{R}$, grande parte da manipulação dos dados é feita pela codificação de instruções, fazendo-se poucas interações com algum tipo de interface gráfica. Mesmo não sendo um impeditivo para o uso da linguagem, a interação gráfica de rotinas já modeladas pode ampliar seu uso tornando, por exemplo, o ajuste de modelos mais intuitivo.

Pensando nisso, parte dos esforços deste trabalho concentraram-se na proposta de uma melhoria para saída de resultados de modelos hierárquicos pertencentes à família de distribuição GAMLLS, com a possibilidade ou não de manipulação gráfica.

As tecnologias utilizadas para proposta foram:

- Linguagem de programação $\mathcal{R}$ para lapidação da saída do modelo ajustado, tendo como

requisitos as seguintes bibliotecas:

nlme - para trabalho com modelos de efeito linear misto;

hnp - para análise da qualidade do ajuste do modelo;

gamlss - para ajuste de modelo;

dplyr - para manipulação de dados;

tidyr - para manipulação e organização de dados;

ggplot2 - para plotagem de gráficos;

string - para manipulação de strings; e

rjson - manipulação de arquivos `.json`;

- Linguagem de programação Python, para criação das rotinas de backend como orquestração de scripts externos e captura de parâmetros. Para isso foi necessário a instalação das seguinte bibliotecas:

  Flask==2.3.2 - micro framework para programação web com python; e

  Jinja2==3.1.2 - manipulação de instruções python dentro de arquivos HTML.

- Shell scritp - para automação das rotinas de variáveis de ambiente do sistema operacional Linux;

- HTML , CSS e JS - para construção da interface de interação do usuário com modelo; e

- Json - para armazenamento temporário dos parâmetros de chamada do scritp r.

3.8 Observações finais

Este capítulo apresentou o contexto de criação do modelo de regressão OLLST, incorporado a família de distribuição GAMLSS, com propósito de trazer uma maior flexibilidade nos ajustes de parâmetros, concentrando-se não apenas na média, mas aceitando ajuste para todos os parâmetros da função. Como uma extensão da Skew t-Student, através da variação dos valores α, λ e ν, o modelo possibilita o ajuste para uma maior amplitude de comportamentos (FERNANDES et al., 2021), graças a sua função de estimação pelo método da máxima verossimilhança.

Considerando o cenário favorável proposto por (FERNANDES, 2021), mostrou-se aqui também a descrição matemática do modelo OLLST para dados longitudinais, de modo que seja possível analisá-lo com dados de medida repetida.

Ao final, foi discorrido sobre uma proposta de lapidação para saída de dados dos modelos hierárquicos ajustado pela função `gamlss` do $\mathcal{R}$. Com o objetivo de tornar mais intuitivo e autônomo o processo execução e análise dos dados, foi sugerido uma ferramenta capaz de automatizar

as rotinas de comparação para diferentes conjuntos de dados onde o usuário, por meio de uma interface gráfica, seja capaz de informar o conjunto de dados, colunas que deseja utilizar e quais os contextos de variáveis que elas representam, informando por fim a distribuição de dados para qual deseja ajustar o modelo. Para tal, foram descritas as ferramentas e tecnologias utilizados no processo de criação desta ferramenta.

Capítulo 4

Resultados e Discussões

4.1 Análise descritiva

Para o conjunto de dados **LV-C**, cujo o gráfico de frequência é apresentado na figura 3.6(LV-C), utilizando a função `histDist` é possível visualizar uma perspectiva de concentração para zona de densidade de probabilidade nas famílias de distribuição Normal e OLLST, como mostra a figura 4.1. Através da saída para esta função, apresentada aqui na tabela 4.1 é possível inferir, preliminarmente, um melhor desempenho da distribuição OLLST em detrimento a NO para esse conjunto de dados.

(NO) (OLLST)

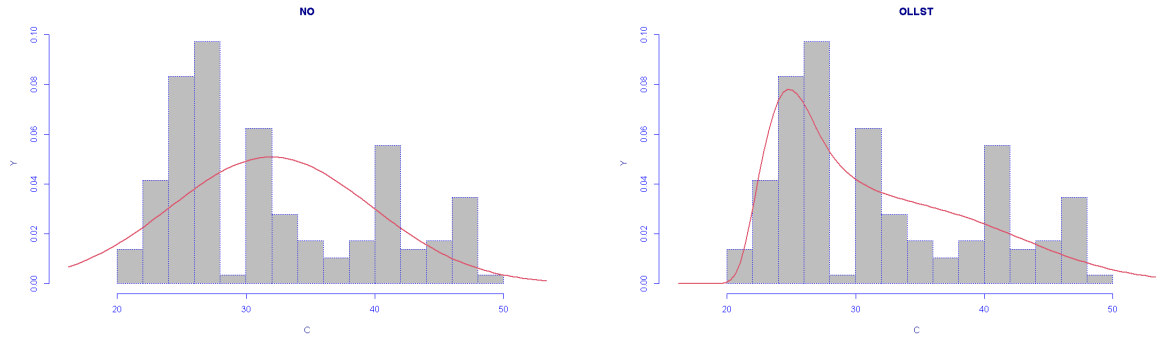

Figura 4.1: Conjunto de dados LV-C

Submetido ao ajuste pelo modelo OLLST, o conjunto de dados **LV-C** teve como saída os resultados apresentados na tabela.As hipóteses formuladas para o teste de comparação levou em consideração cada época (1,2 e 3) a fim de entender se havia ou não diferença significativa entre os tipos de controle. Para o nível de significância de 5% tem-se que:

- Na época 1 não se constatou uma diferença significativa entre os tratamento;

Tabela 4.1: Saída para função `histDist()` - Conjunto LV-C

	Distribuição	
	Normal	OLLST
μ	31.96	24.96
σ	2.059	1.928
ν	-	4.883
τ	-	-0.7497
Global Deviance	1001.56	948.102
AIC	1005.56	956.102
SBC	1011.5	967.981

- Na época 2, o tratamento 3 quando comparado ao tratamento 4 não apresentou diferença significativa (p-valor = 0.8387);
- A terceira época mostra que há diferenças significativas nas comparações entre todos os tratamentos; e
- Os melhores resultados obtidos, em função do tempo, encontram-se na época 3.

4.2.

Na avaliação do comportamento do modelo OLLST em comparação ao NO, através do gráfico de probabilidade mostrados na figura 4.2 é possível concluir para esse conjunto de dados que o modelo OLLST saiu-se melhor no ajunte, com 25% dos pontos fora da banda enquanto o NO teve 29,17%.

(NO) (OLLST)

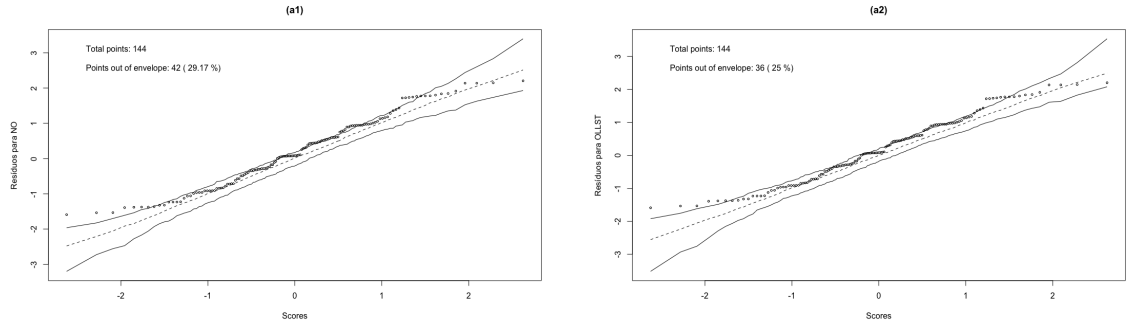

Figura 4.2: Conjunto de dados LV-C

Ainda que o OLLST tenha se saído melhor na comparação acima, esta afirmação restringe-se ao contexto do conjunto de dados LV-C. Por tanto, uma nova análise foi realizada desta vez com o conjunto **LV-CN**, mostrado na figura 3.6(LV-CN) seu gráfico de frequência.

Tabela 4.2: Comparação dos efeitos de cada tratamentos

Época	Hipóteses	Estimativa	Std. Error	p-valor
1	$H_0 : \tau_2 - \tau_1 = 0$	0.000000^{ns}	0.3113539	1.000
1	$H_0 : \tau_3 - \tau_1 = 0$	0.000000^{ns}	0.3113539	1.000
1	$H_0 : \tau_4 - \tau_1 = 0$	0.000000^{ns}	0.3113539	1.000
1	$H_0 : \tau_3 - \tau_2 = 0$	0.000000^{ns}	0.3113538	1.000
1	$H_0 : \tau_4 - \tau_2 = 0$	0.000000^{ns}	0.3113538	1.000
1	$H_0 : \tau_4 - \tau_3 = 0$	0.000000^{ns}	0.3113539	1.000
2	$H_0 : \tau_2 - \tau_1 = 0$	2.224100^{*}	0.5784348	0.0002
2	$H_0 : \tau_3 - \tau_1 = 0$	-2.585490^{*}	0.4519209	0.0000
2	$H_0 : \tau_4 - \tau_1 = 0$	-2.485135^{*}	0.3846374	0.0000
2	$H_0 : \tau_3 - \tau_2 = 0$	-4.809589^{*}	0.6547566	0.0000
2	$H_0 : \tau_4 - \tau_2 = 0$	-4.709235^{*}	0.6102620	0.0000
2	$H_0 : \tau_4 - \tau_3 = 0$	0.100354^{ns}	0.4920016	0.8387
3	$H_0 : \tau_2 - \tau_1 = 0$	7.064183^{*}	0.2837398	0.0000
3	$H_0 : \tau_3 - \tau_1 = 0$	7.791710^{*}	0.2705048	0.0000
3	$H_0 : \tau_4 - \tau_1 = 0$	10.500265^{*}	0.3903520	0.0000
3	$H_0 : \tau_3 - \tau_2 = 0$	0.727528^{*}	0.2011386	0.0004
3	$H_0 : \tau_4 - \tau_2 = 0$	3.436082^{*}	0.3459170	0.0000
3	$H_0 : \tau_4 - \tau_3 = 0$	2.708554^{*}	0.3351465	0.0000

[1] O rótulo * indica que há diferença significava entre os tratamentos considerando;
[1] O rótulo ^{ns} indica que não há diferença significava entre os tratamentos.

Novamente as informações sobre o conjunto obtidas através da função **`histDist`** (veja a figura 4.3) levam a crer que a distribuição OLLST encaixa-se melhor a zona de densidade de probabilidade da população, informações que corroboram com a saída de dado mostradas na tabela 4.3. No entanto, o gráfico de probabilidade para o conjunto mostra um ajuste mais preciso para distribuição NO, com zero pontos fora das bandas do envelope simulado.

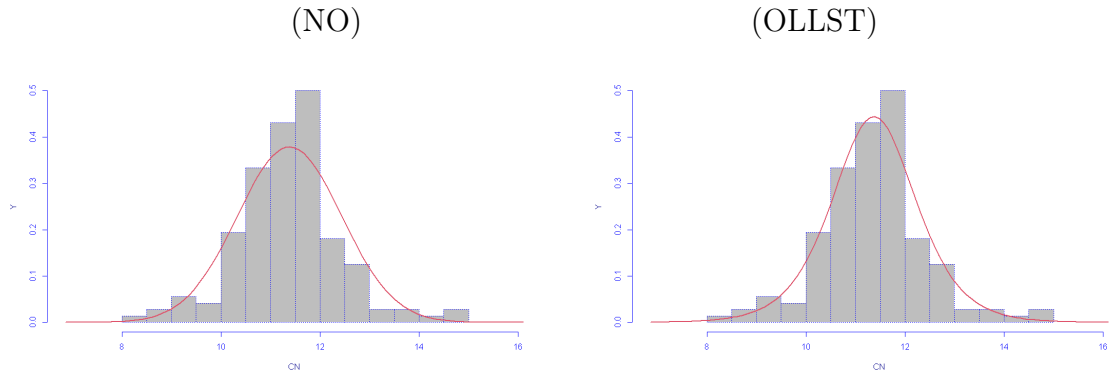

Figura 4.3: HistDist para conjunto LV-CN

Tabela 4.3: Saída para função `histDist()` - Conjunto LV-CN

	Distribuição	
	Normal	OLLST
μ	11.38	8.007
σ	0.05257	4.8
ν	-	0.03472
τ	-	4.907
Global Deviance	423.795	414.762
AIC	427.795	422.762
SBC	433.735	434.641

(NO) (OLLST)

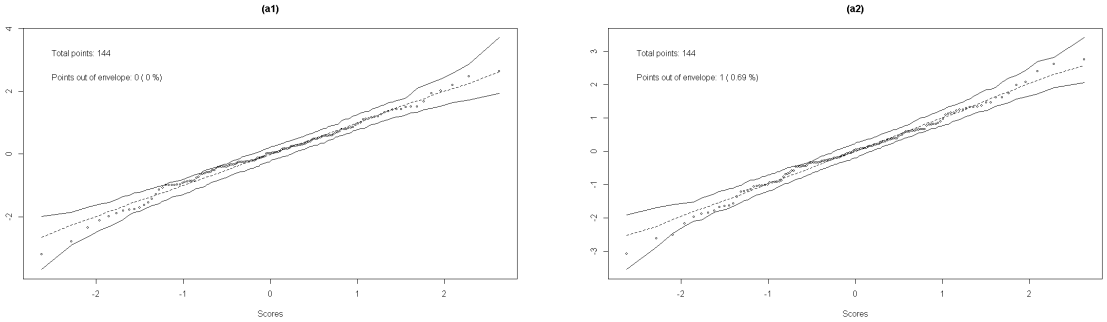

Figura 4.4: Envelope simulado para o conjunto LV-CN

Deferente dos demais conjuntos, onde os tempos de observação foram somente três, para o conjunto **LV-COD** as amostras foram observadas em onze momentos diferentes. Considerando que estes dados avaliam a dissolução do carbono no solo, o gráfico 3.8(LV-COD) mostra um comportamento em decréscimo dos valores para variável resposta com uma zona de concentração assimétrica.

Nesse conjunto, a área sob a curva para distribuição normal (veja a figura 4.5(NO)), mesmo tendendo para zona de densidade de probabilidade da população, não conseguiu abranger uma maior parcela, enquanto a OLLST (veja a figura 4.5(OLLST)) elevou sua moda mais próxima ao pico da zona. Como resultado, cada distribuição apresentou valores mostrados na tabela 4.4.

O indicativo de melhora na performance do OLLST em relação ao NO na análise acima foi contatado no gráfico de probabilidade como mostra a figura 4.6. No entanto, ambos os modelos apresentaram um desempenho aquém com mais da metade dos pontos fora dos limites do envelope simulado. Dos 176 pontos, 134 escaparam do intervalo das bandas para o modelo NO, enquanto OLLST foram 104.

(NO) (OLLST)

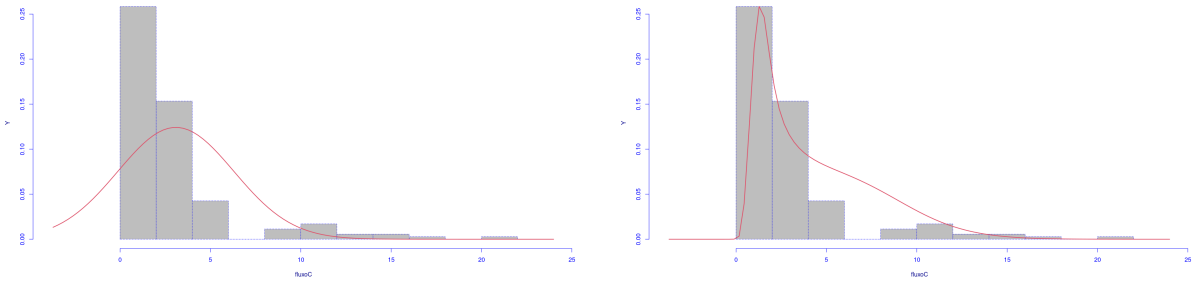

Figura 4.5: HistDist para o conjunto LV-COD

Tabela 4.4: Saída para função `histDist()` - Conjunto LV-COD

	Distribuição	
	Normal	OLLST
μ	3.087	1.305
σ	1.167	1.137
ν	-	9.391
τ	-	-0.6819
Global Deviance	910.21	737.185
AIC	914.21	745.185
SBC	920.551	757.867

(NO) (OLLST)

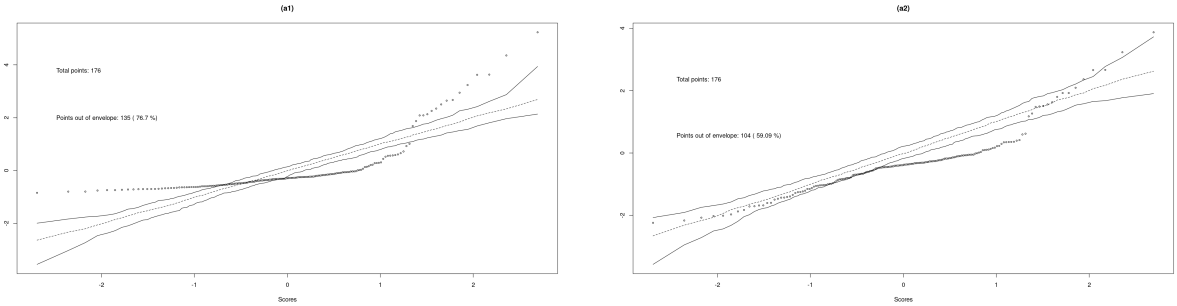

Figura 4.6: Envelope simulado para o conjunto LV-COD

Na análise do conjunto de dados **LV-EN**, cujo o histograma é dado pela figura 3.7(LV-EN), apesar de uma tendência da moda para região de pico da gráfico na análise OLLST (veja a figura 4.9(OLLST)), ao ajustar o modelo de regressão hierárquico verificou-se que o Normal saiu-se melhor, apresentando um ajuste mais assertivo, como mostra a figura 4.8(NO) e 4.8(OLLST). Este comportamento pode ser reflexo da variação abrupta nos valores para variável resposta de uma população pequena, já que esse conjunto é uma parcela do estudo geral proposto por

(MARTINS, 2022) para analisar o teor de carbono no solo em função do tempo de uma única profundidade.

(NO) (OLLST)

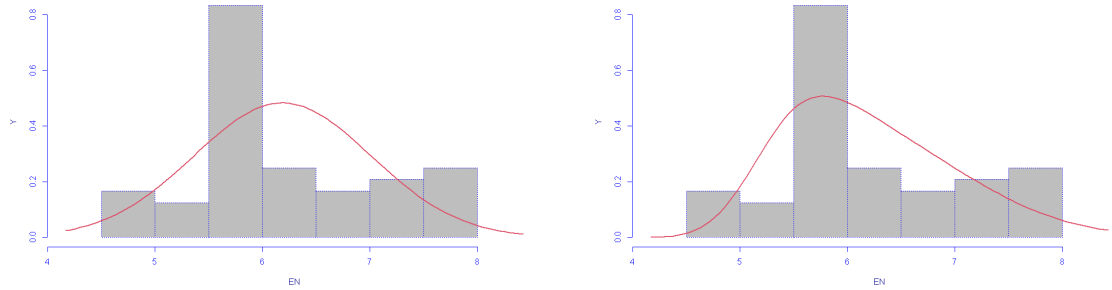

Figura 4.7: HistDist para o conjunto LV-EN

Tabela 4.5: Saída para função `histDist()` - Conjunto LV-EN

	Distribuição	
	Normal	OLLST
μ	6.185	5.459
σ	-0.192	-0.05951
ν	-	2.504
τ	-	-0.3176
Global Deviance	117.783	113.113
AIC	121.783	121.113
SBC	125.525	128.598

(NO) (OLLST)

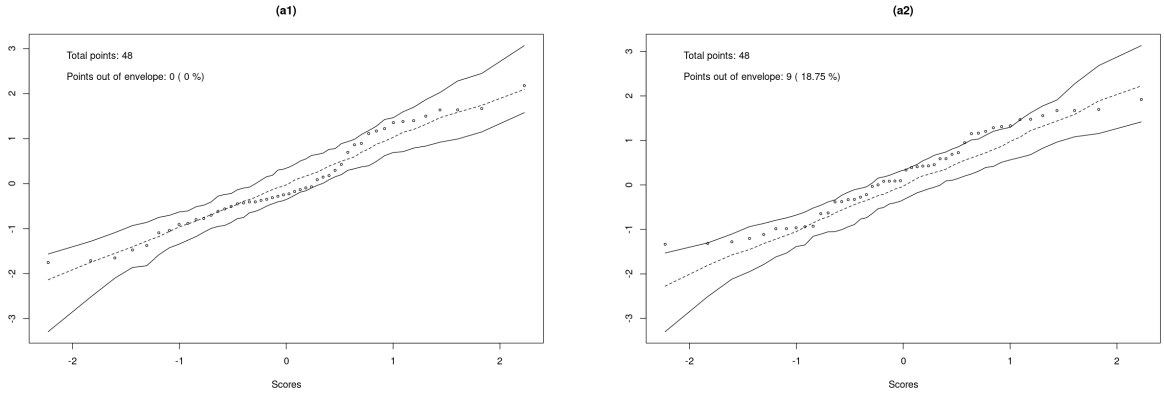

Figura 4.8: Envelope simulado para o conjunto LV-EN

Comportamento semelhante ao observado no ajuste para os dados **LV-EN** ocorreu com **LV-EC**. A figura 3.7(LV-EC) mostra a mesma tendência dos dados, em uma proporção menor, onde há uma varição acentuada nos valores dos níveis de carbono em um conjunto com apenas quatro tratamentos, três tempos e quatro repetições. Como resultado, novamente o modelo NO ajusta-se melhor o comportamento dos dados como mostra o gráfico 4.9(NO) e 4.9(OLLST), onde 46 dos 48 pontos permaneceram dentro de intervalo das bandas para o modelo NO, ao passo que no OLLST esse número subiu para 27.

(NO) (OLLST)

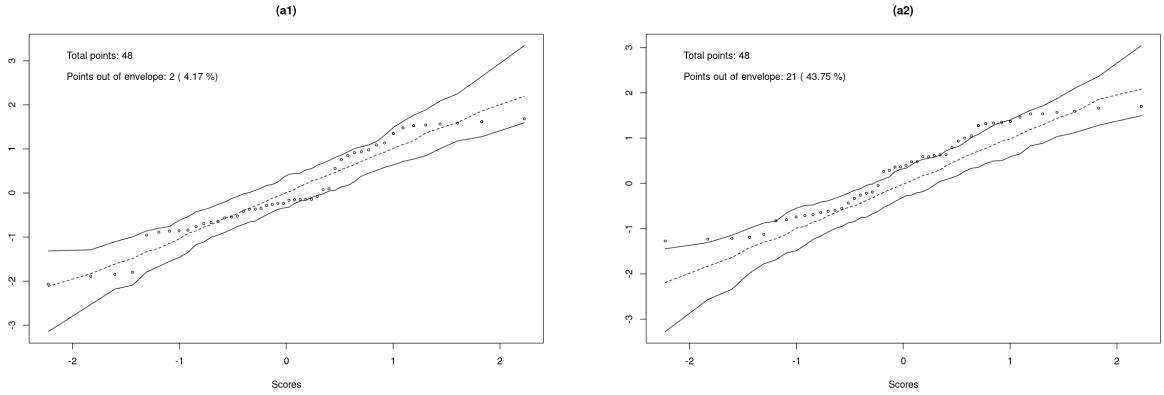

Figura 4.9: Conjunto de dados LV-EC

Finalmente, a mesma sequência de experimentos foi executada para o conjunto de dado **LV-N**. Conforme visto na figura 3.8(LV-N), o conjunto parece apresentar uma espécie de bimodalidade, onde picos das modas não apresentam uma acentuada disparidade nos valores para variável resposta. Como resultado dos testes o modelo OLLST saiu-se melhor quando comparado ao modelo NO, como mostra a saída para função `summary` para ambos os modelos, assim como seus respectivos gráficos de probabilidade (veja a figura 4.10). No envelope simulado, apenas 7,64% dos 144 pontos não ajustaram-se bem pelo modelo OLLST, enquanto para o NO essa porcentagem chegou 52,08% do total.

4.2 Proposta de melhoria para saída do modelo hierárquico

4.2.1 Automação e melhora da saída

A primeira etapa do processo proposto na seção 3.7 foi a construção de um *scritp* $\mathcal{R}$ capaz de se adaptar a variações de conjunto de dados. Como parâmetro de execução, o *scritp* lê um arquivo json *(vide o apêndice D)* onde são informados as colunas do dataframe que serão utilizadas, assim

Tabela 4.6: Summary of Comorbidities.

	Normal		OLLST	
	μ	σ	μ	σ
Pr(>|t|)	<2e-16***	1.8e-11***	<2e-16***	9.95e-05***
Std.	0.05415	0.05893	0.05728	0.1303
Estimate	2.82747	-0.43118	2.20965	-0.5224
Global Deviance	284.4751		236.3024	
AIC	292.5226		249.7433	
SBC	304.4724		269.7018	
Signif. codes: 0 '***' 0.001 '**' 0.1 '*' 0.05 '.' 0.1 ' ' 1; ns - not significant				

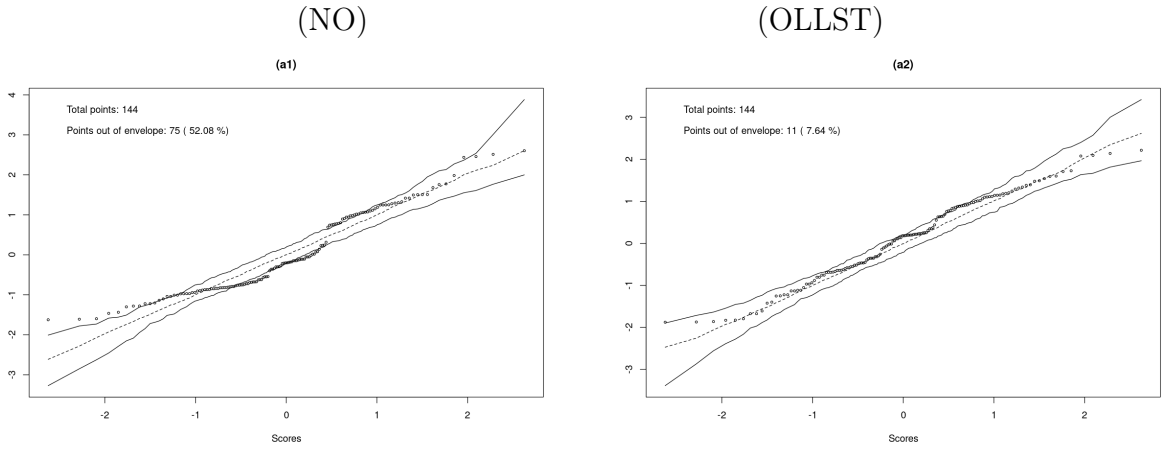

Figura 4.10: Envelope simulado para o conjunto LV-N

como o nome do arquivo e a distribuição da família GAMLSS que deverá ser aplicada para o ajuste do modelo.

O arquivo também é capaz de identificar as interações possíveis para cada indivíduo, criando suas próprias restrições de forma automática, gerando uma saída de dados em `.csv`, além de um gráfico de perfis da variável resposta em função do tempo. Contudo, a maior parte das funções descrita no arquivo são destinadas ao tratamento da saída do ajuste, de modo a facilitar a interpretação por parte do pesquisador. Deste modo, cada restrição foi tratada de forma a explicitar os indivíduos/tratamentos da comparação em cada tempo, mostrando ainda se esta comparação foi significativa ou não.

4.2.2 Manipulação gráfica

Como parte da melhoria proposta, uma interface web de manipulação foi criada para que o usuário possa interagir intuitivamente com o modelo sem precisar programar rotinas, bastando apenas informar com quais elementos deseja trabalhar.

Utilizando a linguagem de programação `Python` e o framework Flask, entre outras bibliotecas, além de `shell script`, foi possível receber do usuário o conjunto de dados, a identificação das colunas (variável resposta, tempo e indivíduo/tratamento) e a distribuição escolhida para o ajuste do modelo, retornando um arquivo `.zip` com todas as análises para cada restrição em função do tempo *(Vide as figuras 4.11, 4.12 e 4.13)*.

Figura 4.11: Interface gráfica - Janela 01

4.2.3 Resultados alcançados

Finalmente, chegou-se a um modelo de interação onde o pesquisador passou a ser um usuário dos modelos de regressão que descendem do GAMLLS. Com a retirada de ruído de informações que não dizem respeito a restrição da variável resposta em função do tempo, chegou-se a um esquema em tabela semelhante ao mostrado em 4.7, onde:

- "Época"refere-se a variável tempo no momento da comparação;
- "Rest."refere-se ao tratamento em restrição para o tempo ω_i;

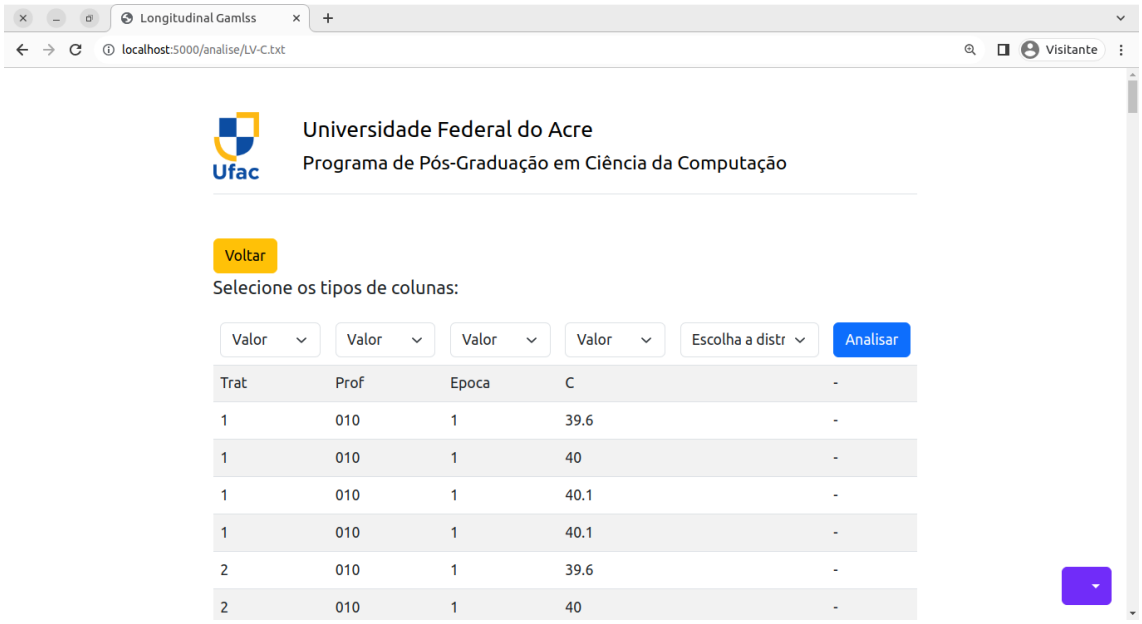

Figura 4.12: Interface gráfica - Janela 02

- "Comp."refere-se ao tratamento comparado com restrição para o tempo ω_i;
- "Value", "Std Error", "DF", "t.value"e "p.value"são os valores retornados pelo modelo ajustado; e
- "Sign."é marcador da comparação que verifica se dentro do tempo ω_i a interação entre os dois tratamentos foi significativa ou não. ***** - foi significante a interação; **NS** - não foi significante a interação;

Tabela 4.7: Representação da saída de dados pela nova proposta

Época	Rest.	Comp.	Value	Std Error	DF	t.value	p.value	Sign.
ω_0	y_{111}	y_{111}	-	-	-	-	-	[* or NS]
.								
.								
.								
ω_i	y_{ijk}	y_{ijk}	-	-	-	-	-	[* or NS]

4.3 Observações finais

Neste capítulo foram analisados seis conjuntos de dados com variações diferentes para o comportamento da distribuição. O conjunto **LV-C**, assim como o **LV-COD**, apresentava assimetria a esquerda sendo o modelo OLLST aquele que saiu-se melhor nos testes comparativos em relação

Figura 4.13: Interface gráfica - Janela 03

o NO. Há de se considerar, no entanto, que mesmo apresentando o melhor resultado entre os dois, tanto o OLLST quanto o NO, tiveram dificuldades para estimar os dados de **LV-COD**, de modo que ambos tiveram uma taxa de erro superior a 50%.

Quanto aos dados para o teor de carbono em razão do nitrogênio (conjunto **LV-CN**), considerando que o comportamento da distribuição assemelha-se ao gaussiano, o modelo NO obteve ligeira vantagem com todos os pontos dentro do intervalo de bandas do envelope simulado, ao passo que o modelo OLLST teve um apenas, dos 144 pontos fora desse intervalo. Sob essa perspectiva é válido afirmar que ambos os modelos saíram-se bem para o ajuste deste conjunto de dados.

Os conjunto de dados **LV-EN** e **LV-EC** apresentaram um comportamento semelhante em suas distribuições dos dados. A variação acentuada dos valores para a variável resposta pode ter contribuído para maior taxa de erro do OLLST em relação ao NO. Para os dados **LV-EN** a diferença, ainda que significativa, foi menor quando compara a do **LV-EC**.

Por fim, na análise do conjunto **LV-N** o modelo OLLST obteve excelentes resultados, alcançando um ajuste significativamente melhor em relação ao NO. O gráfico de probabilidade mostrou uma taxa de efetividade de 92,36% para o modelo, enquanto o NO manteve-se em 47,92%.

Neste capítulo, foram apresentados também os resultados práticos da melhoria proposta na seção 3.7. A ferramenta desenvolvida com uso de diversas tecnologias pode representar um avanço na análise dos resultados obtidos pelos ajustes de modelos hierárquicos da biblioteca `gamlss`, ressaltando ainda que os dados apresentados na tabela 4.2 são frutos do ajuste proposto pela nova ferramenta.

Apesar das facilidades proposta pelo novo modelo de saída, existem ainda aspectos de melhorias que podem ser melhor explorados. Um exemplo disso seria a inclusão com extensão do editor de código para $\mathcal{R}$, *rStudio*, agregação de ouros modelos ou bibliotecas, além do modelo hierárquico, entre outras.

Capítulo 5

Considerações Finais

5.1 Conclusões

Como resultado das limitações observadas pelos autores (AZZALINI; CAPITANIO, 1999; BRAGA et al., 2022) para ajuste de modelos de regressão cujo os dados não apresentam distribuição gaussiana, somado aos desafios da análise de dados longitudinais (FITZMAURICE et al., 2008), conduziu-se uma pesquisa com o modelo OLLST a fim de analisa-lo sob essa perspectiva, até o momento não estudada.

Duas foram as frentes de trabalho neste estudo: a primeira referia-se à análise do modelo OLLST, proposto por (FERNANDES, 2021), para conjunto de dados com medida repetida no tempo; e a outra foi a proposta de melhoria da saída de dados para modelos pertencentes a família de distribuição `gamlss` (RIGBY; STASINOPOULOS, 2005), incluindo assim o modelo OLLST.

Para análise do modelo com dados longitudinais foram considerados seis conjuntos de dados derivados estudo proposto por (MARTINS, 2022) sobre análise do teor de carbono no solo. A pesquisa comparou o desempenho do modelo OLLST com o NO para todos os conjuntos percebendo desempenho superior do OLLST para três, performance semelhante em um, e inferior em dois.

Para os conjuntos **LV-CN**, **LV-EN** e **LV-EC**, o modelo OLLST apresentou uma performance menor que o NO, alguns caso, ligeiramente abaixo e em outros essa diferença foi um pouco mais acentuada. Nestas análises, foram verificadas a presença de picos dissonantes (*outliers*) para variável reposta em relação aos demais valores, fator que pode ter influência na caracterização da amostra levando o modelo à estimativas menos acertivas.

Já para o conjunto **LV-N**, com indicativo de bimodalidade, o OLLST ajustou-se significativamente melhor quando comparado com NO. Taxa um pouco mais modesta, ainda sim melhor

que o NO, ocorreu na análise do conjunto **LV-C**, com tendência à assimetria a esquerda, onde apenas 36 dos 144 pontos do gráfico de probabilidade ficaram fora dos limites das banda, contra 44 do NO.

Como resultado dos testes, é possível concluir pela validade do modelo OLLST para análise de dados longitudinais com efeito aleatório, como mostra a análise descritiva proposta no capitulo anterior para o conjunto **LV-C**, além do excelente ajuste do modelo para o conjunto **LV-N**. A presença de *outliers*, trouxe um indicativo de que o modelo NO pode estimar melhor, em cenários assim, no entanto, esta deficiência não representa uma característica exclusiva do OLLST, sendo inclusive uma dificuldade já mapeada na literatura da análise de regressão para estimação (veja (BARNETT; LEWIS et al., 1994)). Na análise do dados **LV-COD**, por exemplo, mesmo com presença de *outliers* o modelo OLLST conseguiu adequar-se melhor ao comportamento da distribuição quando comparado o NO, ainda que ambos não tenham conseguido encaixar mais de 50% dos pontos dentro do intervalo do gráfico de probabilidade.

Para o segundo nicho de trabalho desta pesquisa, como resultado, chegou-se a uma ferramenta cujo objetivo é tornar mais intuitivo o acesso a mecanismos de modelagem para dados com medida repetida no tempo. Com a rotina de ajuste do modelo construída, foi possível interagir com a linguagem $\mathcal{R}$ e obter resposta para modelagem dos dados sem fosse necessário codificar na linguagem, uma proposta que pode difundir o acesso estatística.

5.2 Perspectiva de trabalhos futuros

Ainda que a análise feita pelo OLLST tenha sido mais assertivas para alguns conjuntos de dados, existem lacunas que podem ser melhores exploradas para uma variedade maior de comportamento. Um exemplo disso, seria a condução de um estudo de simulação para o modelo tomando como base os parâmetros dispostos aqui e avaliar a eficiência das estimativas do modelo.

Há também aperfeiçoamentos que podem atingir a ferramenta proposta neste trabalho. Por ser uma ferramenta web, ela pode ser propagada na internet ou incorporar como extensão do `rStudio`, podendo ainda agregar outras famílias de distribuição e bibliotecas, ou mesmo refinando ainda mais a saída para somente dados de interesse do pesquisador.

Referências

ANGRIST, J. D.; PISCHKE, J.-S. *Mostly harmless econometrics: An empiricist's companion*. [S.l.]: Princeton university press, 2009.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 61, n. 3, p. 579–602, 1999.

AZZALINI, A. et al. Log-skew-normal and log-skew-t distributions as models for family income data. *Journal of income distribution*, v. 11, p. 13–21, 2003.

BARBOSA, J.; LOEFFLER, C.; BULCÃO, A. Um estudo sobre a eficiência da quadratura gaussiana na integração singular com o método dos elementos de contorno.

BARNETT, V.; LEWIS, T. et al. *Outliers in statistical data*. [S.l.]: Wiley New York, 1994. v. 3.

BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964.

BRAGA, A. d. S. et al. A random-effects regression model based on the odd log-logistic skew normal distribution. *Journal of Statistical Theory and Practice*, Springer, v. 16, n. 2, p. 33, 2022.

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

BRESLOW, N. E.; LIN, X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, Oxford University Press, v. 82, n. 1, p. 81–91, 1995.

BUJA, A.; HASTIE, T.; TIBSHIRANI, R. Linear smoothers and additive models. *The Annals of Statistics*, JSTOR, p. 453–510, 1989.

CHEIN, F. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. Escola Nacional de Administração Pública (Enap), 2019.

CNAAN, A.; LAIRD, N. M.; SLASOR, P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, Wiley Online Library, v. 16, n. 20, p. 2349–2380, 1997.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006.

FAGUNDES, G. S. Metodologias não-paramétricas para estudos com medidas repetidas. 2013.

FAUSTO, M. A. et al. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. *Cadernos de Saúde Pública*, SciELO Public Health, v. 24, p. 513–524, 2008.

FERNANDES, A. H. d. S. *AN INTRODUCTION TO A NEW PROBABILITY DISTRIBUTION INTO GAMLSS FRAMEWORK IN SOFTWARE R: ODD LOG-LOGISTIC SKEW T-STUDENT*. Dissertação (Mestrado), 2021.

FERNANDES, A. H. dos S. et al. An overview of the development of a new probability distribution: Odd log-logistic skew t-student. *Brazilian Journal of Development*, v. 7, n. 3, p. 30536–30555, 2021.

FERNANDES, V. V. Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão. Universidade Federal de São Carlos, 2019.

FERREIRA, W. L. Análise de dados com medidas repetidas em experimento com ingestão de café. Universidade Federal de Lavras, 2012.

FITZMAURICE, G. et al. *Longitudinal data analysis*. [S.l.]: CRC press, 2008.

FLORENCIO, L. de A. *Engenharia de avaliações com base em modelos GAMLSS*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2010.

FREITAS, A. R. d.; PRESOTTI, C. V.; TORAL, F. L. B. Alternativas de análises em dados de medidas repetidas de bovinos de corte. *Revista Brasileira de Zootecnia*, SciELO Brasil, v. 34, p. 2233–2244, 2005.

GOLUB, G. H.; WELSCH, J. H. Calculation of gauss quadrature rules. *Mathematics of computation*, v. 23, n. 106, p. 221–230, 1969.

KAC, G.; SICHIERI, R.; GIGANTE, D. P. *Epidemiologia nutricional*. [S.l.]: Editora Fiocruz, 2007.

KAISER, K.; KALBITZ, K. Cycling downwards–dissolved organic matter in soils. *Soil Biology and Biochemistry*, Elsevier, v. 52, p. 29–32, 2012.

KALBITZ, K. et al. Controls on the dynamics of dissolved organic matter in soils: a review. *Soil science*, LWW, v. 165, n. 4, p. 277–304, 2000.

KÖPPEN, M. The curse of dimensionality. In: *5th online world conference on soft computing in industrial applications (WSC5)*. [S.l.: s.n.], 2000. v. 1, p. 4–8.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, JSTOR, p. 963–974, 1982.

LIMA, L. P. de. *Modelos Aditivos Generalizados: aplicação a um estudo epidemiológico ambiental*. Tese (Doutorado) — Universidade de São Paulo, 2001.

MADDALA, G. Introdução à econometria. 3. *Ed. Rio de Janeiro: LTC*, 2003.

MARTINS, G. B. *Dinâmica do carbono orgânico dissolvido em solos com presença de resíduos vegetais*. Dissertação (Mestrado), 2022.

MARTINS, N. D. d. C. *Programação em R no estudo de probabilidades*. Tese (Doutorado), 2016.

NATIS, L. Modelos lineares hierárquicos. *Estudos em Avaliação Educacional*, v. 3, 2001.

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

PEREIRA, T. B. et al. Eficiência da seleção de progênies de café f4 pela metodologia de modelos mistos (reml/blup). *Bragantia*, SciELO Brasil, v. 72, p. 230–236, 2013.

RIGBY, R.; STASINOPOULOS, D. Madam macros to fit mean and dispersion additive models. *Glim4 macro library manual, release*, v. 2, p. 48–84, 1996.

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

ROSA, P. *Análise não-paramétrica de dados ordinais com medidas repetidas*. Tese (Doutorado) — Universidade de Sao Paulo, 2001.

SAHU, S.; DEY, D. On multivariate survival models with a skewed frailty and a correlated baseline hazard process. *Skew-elliptical distributions and their applications: A journey beyond normality*, CRC/Chapman & Hall, Boca Raton, FL, p. 321–338, 2004.

SAHU, S. K.; DEY, D. K.; BRANCO, M. D. A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, Wiley Online Library, v. 31, n. 2, p. 129–150, 2003.

SCHWARTZ, A. L. *Theory and implementation of numerical methods based on Runge-Kutta integration for solving optimal control problems*. [S.l.]: University of California, Berkeley, 1996.

SILVA, S. F. da. IntegraÇÃo numÉrica: Quadratura de causs. *Revista Traços*, v. 3, n. 6, 2017.

SINGER, J.; ANDRADE, D. *Analysis of longitudinal data. Em: Handbook of Statistics. Volume 18: Bio-Environmental and Public Health Statistics. eds. PK Sen and CR Rao*. [S.l.]: Amsterdam: North Holland, 2000.

WANG, J. et al. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer informatics*, SAGE Publications Sage UK: London, England, v. 7, p. CIN–S2846, 2009.

APÊNDICES A – OLLST Distribution Script

```
##Ramires 15/10/2015 dd/mm/yyyy
##Joana 16/02/2020 dd/mm/yyyy

####
####
## Package ‘numDeriv’
## Title: Accurate Numerical Derivatives
## Descrition: Methods for calculating (usually) accurate
## numerical first and second order derivatives. Accurate calculations
## are done using ‘Richardson"s’ extrapolation or, when applicable, a
## complex step derivative is available. A simple difference
## method is also provided. Simple difference is (usually) less accurate
## but is much quicker than ‘Richardson"s’ extrapolation and provides a
## useful cross-check.
## Methods are provided for real scalar and vector valued functions.
##
```

```
## License: GPL-2
## Copyright: 2006-2011, Bank of Canada. 2012-2016, Paul Gilbert
## Author: Paul Gilbert <pgilbert.ttv9z@ncf.ca> and Ravi Varadhan
## Repository: CRAN
####
####
require(numDeriv)

####
####
## Package 'sn'
## Title: The Skew-Normal and Related Distributions Such as the Skew-t
## Author: Adelchi Azzalini <adelchi.azzalini@unipd.it>
## Description: Build and manipulate probability distributions of the skew-normal
## family and some related ones, notably the skew-t family, and provide related
## statistical methods for data fitting and model diagnostics, in the univariate
## and the multivariate case.
##
## License: GPL-2 | GPL-3
## Repository: CRAN
####
####
require(sn)

OLLST <- function(mu.link = "identity", sigma.link = "log", nu.link = "identity", tau.link = "log"){
```

```
##To define the link function of any of the parameters, the checklink() function is used.
##This function takes four arguments:
##
##1) which.link: which parameter the link is for, e.g. ''mu.link''
##2) which.dist: the current distribution, e.g. ''Normal'' (the name is only used to report
##an error in the specification of the link function)
##3) link: which link is currently used (the default value is the one given as arguments
##in the function definition, e.g. substitute(mu.link) will do the job)
##4) link.List: the list of the possible links for the specific parameter,
    e.g. c(''inverse'', ''log'', ''identity'')
##5) par.link: a list containing the value of the parameter(s) (if the link has parameter(s) as for example
##in the ''logshifted'' and ''logitshifted'' links)
##
mstats <- checklink("mu.link", "odd log logistic skew tstudent", substitute(mu.link),
c("inverse", "log", "identity", "own"))
dstats <- checklink("sigma.link", "odd log logistic skew tstudent", substitute(sigma.link),
c("inverse", "log", "identity", "own"))
vstats <- checklink("nu.link", "odd log logistic skew tstudent", substitute(nu.link),
c("1/nu^2", "log", "identity", "own"))
tstats <- checklink("tau.link", "odd log logistic skew tstudent", substitute(tau.link),
c("1/tau^2", "log", "identity", "own"))


    structure(
        ##family: the name of the distribution
        list(family = c("OLLST", "odd log logistic skew tstudent"),
```

```
##parameters: a list indicating whether the parameter will be fitted i.e. mu = TRUE
##or fixed at initial values i.e. nu = FALSE
parameters = list(mu = TRUE, sigma = TRUE, nu = TRUE, tau = TRUE),

##nopar: the number of parameters
nopar = 4,

##type: the type of distribution i.e. ''Continuous'' or ''Discrete''
type = "Continuous",

###link: the current link functions as character strings
mu.link = as.character(substitute(mu.link)),
sigma.link = as.character(substitute(sigma.link)),
nu.link = as.character(substitute(nu.link)),
tau.link = as.character(substitute(tau.link)),

##linkfun: the actual link functions returned from checklink()
mu.linkfun = mstats$linkfun,
sigma.linkfun = dstats$linkfun,
nu.linkfun = vstats$linkfun,
tau.linkfun = tstats$linkfun,

##linkinv: the actual inverse link functions returned from checklink()
mu.linkinv = mstats$linkinv,
```

```
sigma.linkinv = dstats$linkinv,
nu.linkinv = vstats$linkinv,
tau.linkinv = tstats$linkinv,

##dr: the actual first derivative of the inverse link functions returned from checklink()
mu.dr = mstats$mu.eta,
sigma.dr = dstats$mu.eta,
nu.dr = vstats$mu.eta,
tau.dr = tstats$mu.eta,

##dldm: the first derivative of the likelihood concerning the location parameter mu
#-------------------- ok
dldm = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, x, sigma, nu, tau){log(dauxiOLLST(t, x, sigma, nu, tau))}
    dldm <- grad(func = lpdf, t = y, x = mu, sigma = sigma, nu = nu, tau = tau,
        method = 'simple')
    dldm
},

##d2ldm2: the expected second derivative of the likelihood concerning
##the location parameter mu
#-------------------- ok
d2ldm2 = function(y,mu,sigma,nu,tau){
    lpdf <- function(t, x, sigma, nu, tau){log(dauxiOLLST(t, x, sigma, nu, tau))}
    dldm <- grad(func = lpdf, t = y, x = mu, sigma = sigma, nu = nu, tau = tau,
```

```
        method = 'simple')
    d2ldm2 <- -(dldm * dldm)
    d2ldm2 <- ifelse(d2ldm2 < -1e-15, d2ldm2, -1e-15)
    d2ldm2
},

##dldd: the first derivative of the likelihood concerning the scale parameter sigma
#-------------------- ok
dldd = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, x, nu, tau) {log(dauxiOLLST(t, mu, x, nu, tau))}
    dldd <- grad(func = lpdf, t = y, mu = mu, x = sigma, nu = nu, tau = tau, method = 'simple')
    dldd
},

##d2ldd2: the expected second derivative of the likelihood concerning
##the scale parameter sigma
#-------------------- ok
d2ldd2 = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, x, nu, tau){log(dauxiOLLST(t, mu, x, nu, tau))}
    dldd <- grad(func = lpdf, t = y, mu = mu, x = sigma, nu = nu, tau = tau, method = 'simple')
    d2ldd2 <- -(dldd * dldd)
    d2ldd2 <- ifelse(d2ldd2 < -1e-15, d2ldd2, -1e-15)
    d2ldd2
},
```

```
##dldv: the first derivative of the likelihood concerning the shape parameter nu
#-------------------- ok
dldv = function(y, mu, sigma, nu, tau){
lpdf <- function(t, mu, sigma, x, tau){log(dauxiOLLST(t, mu, sigma, x, tau))}
dldv <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, x = nu, tau = tau, method = 'simple')
dldv
},

##d2ldv2: the expected second derivative of the likelihood concerning
##the shape parameter nu
#-------------------- ok
d2ldv2 = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, sigma, x, tau){log(dauxiOLLST(t, mu, sigma, x, tau))}
    dldv <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, x = nu, tau = tau,
        method = 'simple')
    d2ldv2 <- -(dldv * dldv)
    d2ldv2 <- ifelse(d2ldv2 < -1e-15, d2ldv2, -1e-15)
    d2ldv2
},

##dldt: the first derivative of the likelihood concerning the shape parameter tau
#-------------------- ok
    dldt = function(y, mu, sigma, nu,tau){
    lpdf <- function(t, mu, sigma, nu, x){log(dauxiOLLST(t, mu, sigma, nu,x))}
    dldt <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, nu = nu, x = tau)
```

```
    dldt
},

##d2ldt2: the expected second derivative of the likelihood concerning
##the shape parameter tau
#-------------------- ok
d2ldt2 = function(y, mu, sigma, nu, tau){
    lpdf<-function(t, mu, sigma, nu, x){log(dauxiOLLST(t, mu, sigma, nu, x))}
    dldt<-grad(func = lpdf, t = y, mu = mu, sigma = sigma, nu = nu, x = tau, method = 'simple')
    d2ldt2<- -(dldt * dldt)
    d2ldt2<- ifelse(d2ldt2 < -1e-15, d2ldt2, -1e-15)
    d2ldt2
},

##d2ldmdd: the expected cross derivative of the likelihood concerning both
##the location mu and scale parameter sigma
#-------------------- ok
d2ldmdd = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, x, sigma, nu, tau){log(dauxiOLLST(t, x, sigma, nu, tau))}
    dldm <- grad(func = lpdf, t = y, x = mu, sigma = sigma, nu = nu, tau = tau,
        method = 'simple')
    lpdf <- function(t, mu, x, nu, tau){log(dauxiOLLST(t, mu, x, nu, tau))}
    dldd <- grad(func = lpdf, t = y, mu = mu, x = sigma, nu = nu, tau = tau)
    d2ldmdd = -(dldm * dldd)
    d2ldmdd <- ifelse(is.na(d2ldmdd) == TRUE, 0, d2ldmdd)
```

```
    d2ldmdd
},

##d2ldmdv: the expected cross derivative of the likelihood concerning both
##the location mu and shape parameter nu
#-------------------- ok
d2ldmdv = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, x, sigma, nu, tau){log(dauxiOLLST(t, x, sigma, nu, tau))}
    dldm <- grad(func = lpdf, t = y, x = mu, sigma = sigma, nu = nu, tau = tau,
        method = 'simple')
    lpdf <- function(t, mu, sigma, x, tau){log(dauxiOLLST(t, mu, sigma, x, tau))}
    dldv <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, x = nu, tau = tau)
    d2ldmdv = -(dldm * dldv)
    d2ldmdv
},

##d2ldmdd: the expected cross derivative of the likelihood concerning both
##the location mu and shape parameter tau
#-------------------- ok
d2ldmdt = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, x, sigma, nu, tau){log(dauxiOLLST(t, x, sigma, nu, tau))}
    dldm <- grad(func = lpdf, t = y, x = mu, sigma = sigma, nu = nu, tau = tau,
        method = 'simple')
    lpdf <- function(t, mu, sigma, nu, x){log(dauxiOLLST(t, mu, sigma, nu,x))}
    dldt <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, nu = nu, x = tau)
```

```
    d2ldmdt <- -(dldm * dldt)
    d2ldmdt
},

##d2ldddv: the expected cross derivative of the likelihood concerning both
##the scale sigma and shape parameter nu
#-------------------- ok
d2ldddv = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, x, nu, tau){log(dauxiOLLST(t, mu, x, nu, tau))}
    dldd <- grad(func = lpdf, t = y, mu = mu, x = sigma, nu = nu, tau = tau, method = 'simple')
    lpdf <- function(t, mu, sigma, x, tau){log(dauxiOLLST(t, mu, sigma, x, tau))}
    dldv <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, x = nu, tau = tau)
    d2ldddv = -(dldd * dldv)
    d2ldddv
},

##d2ldddt: the expected cross derivative of the likelihood concerning both
##the scale sigma and shape parameter tau
#-------------------- ok
d2ldddt = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, x, nu, tau){log(dauxiOLLST(t, mu, x, nu, tau))}
    dldd <- grad(func = lpdf, t = y, mu = mu, x = sigma, nu = nu, tau = tau)
    lpdf <- function(t, mu, sigma, nu, x){log(dauxiOLLST(t, mu, sigma, nu,x))}
    dldt <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, nu = nu, x = tau)
    d2ldddt <- -(dldd*dldt)
```

```
    d2ldddt
},

##d2ldvdt: the expected cross derivative of the likelihood concerning both
##the shape nu and shape parameter tau
#-------------------- ok
d2ldvdt = function(y, mu, sigma, nu, tau){
    lpdf <- function(t, mu, sigma, x, tau){log(dauxiOLLST(t, mu, sigma, x, tau))}
    dldv <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, x = nu, tau = tau)
    lpdf <- function(t, mu, sigma, nu, x){log(dauxiOLLST(t, mu, sigma, nu,x))}
    dldt <- grad(func = lpdf, t = y, mu = mu, sigma = sigma, nu = nu, x = tau)
    d2ldvdt <- -(dldv*dldt)
    d2ldvdt
},

#-------------------- ok
##G.dev.incr: the global deviance (equal to minus twice the log likelihood)
G.dev.incr = function(y, mu, sigma, nu, tau, ...){
    -2*dOLLST(y, mu, sigma, nu, tau, log = TRUE)
},

##rqres: the definition of the (normalized quantile residuals)
##[note these are randomized for discrete distributions]
rqres = expression(
    rqres(pfun="pOLLST", type="Continuous", y=y, mu=mu, sigma=sigma, nu=nu, tau=tau)
```

```
            ),

            ##initial: the initial default values for the starting of the algorithm
            ##[both vectors of length n]
            mu.initial = expression(mu <- (y + mean(y))/2),
            sigma.initial = expression(sigma <- rep(sd(y), length(y))),
            nu.initial = expression(nu <- rep(1, length(y))),
            tau.initial = expression(tau <-rep(1, length(y))),

            ##valid: a valid range of values for the parameters and the response variable
            mu.valid = function(mu) TRUE,
            sigma.valid = function(sigma) all(sigma > 0),
            nu.valid = function(nu) TRUE,
            tau.valid = function(tau) all(tau > 0),
            y.valid = function(y) TRUE
        ), #endList
    class = c("gamlss.family","family")
    ) #endStructure
} #endFunction

##The dOLLST, pOLLST, qOLLST and rOLLST functions
##These four functions defined in general, the Probability Density Function (pdf),
##the Cumulative Density Function (cdf), the inverse cdf, i.e., quantile and
##random generating functions for the distribution at hand.
#--------------------- PDF
```

```
dOLLST <- function(x, mu = 0, sigma = 1, nu = 2, tau = 0.2, log = FALSE){
    if (any(tau < 0))
        stop(paste("tau must be positive", "\n", ""))

    theta = 300
    z = (x - mu)/sigma

    pdfst <- (2/sigma)*dt(z,theta,0,log = FALSE)*pt(nu*z,theta,0,lower.tail = TRUE,log.p = FALSE)
    cdfst <- pt(z,theta,0,lower.tail = TRUE, log.p = FALSE)-2*owen(z,nu)
    fy1 <- (tau*pdfst*(cdfst*(1-cdfst))^(tau-1))/((cdfst^tau+(1-cdfst)^tau)^2)

    if (log == FALSE) fy <- fy1
        else fy <- log(fy1)

    fy
}
#-------------------- CDF
pOLLST <- function(q, mu = 0, sigma = 1, nu = 2, tau = 0.2, lower.tail = TRUE, log.p = FALSE){
    if (any(tau < 0))
        stop(paste("tau must be positive", "\n", ""))

    theta = 300
    z = (q - mu)/sigma
    cdfst <- pt(z, theta, 0, lower.tail = TRUE, log.p = FALSE)-2*owen(z, nu)
    cdf1 <- (cdfst^tau)/(cdfst^tau+(1-cdfst)^tau)
```

```
    if(lower.tail == TRUE) cdf <- cdf1
        else cdf <- 1-cdf1
    if(log.p == FALSE) cdf <- cdf
        else cdf <- log(cdf)

    cdf
}
#-------------------- Quantile
qOLLST <- function(p, mu = 0, sigma = 1, nu = 2, tau = 0.2, lower.tail = TRUE, log.p = FALSE){
    if (any(sigma < 0))
        stop(paste("sigma must be positive", "\n", ""))
    if (any(tau < 0))
        stop(paste("tau must be positive", "\n", ""))
    if (any(p < 0)|any(p > 1))
        stop(paste("p must be between 0 and 1", "\n", ""))

    if (log.p == TRUE) p <- exp(p)
        else p <- p
    if (lower.tail == TRUE) p <- p
        else p <- 1-p

    theta = 300
    u <- (p^(1/tau))/((1-p)^(1/tau)+p^(1/tau))
    q <- mu + sigma*qst(u, mu, omega = 1, tau, theta)
```

```
    q
}
#-------------------- Random
rOLLST <- function(n, mu = 0, sigma = 1, nu = 2, tau = 0.2){
    if (any(tau < 0))
    stop(paste("tau must be positive", "\n", ""))

    uni <- runif(n = n, 0, 1)
    r <- qOLLST(uni, mu = mu, sigma = sigma, nu = nu, tau = tau)
    r
}
#-------------------- PDFauxiliar
dauxiOLLST <- function(t, mu, sigma, nu, tau){
    theta = 300
    z = (t-mu)/sigma

    pdfst <- (2/sigma)*dt(z, theta, 0, log = FALSE)*pt(nu*z, theta, 0, lower.tail = TRUE, log.p = FALSE)
    cdfst <- pt(z, theta, 0, lower.tail = TRUE, log.p = FALSE)-2*owen(z, nu)
    fy1 <- (tau*pdfst*(cdfst*(1-cdfst))^(tau-1))/((cdfst^tau+(1-cdfst)^tau)^2)

    fy1
}
#-------------------- Owen
owen <- function(h,a){
    func <- function(x,h)((exp(-0.5*h^2*(1+x^2)))/(1+x^2))*(1/(2*pi))
```

```
    temp1 <- c()
    if(length(h)>1 & length(a)>1){
        for(i in 1:length(h)){
            int <- integrate(f = func, lower = 0, upper = a[i], h = h[i])
            temp1 <- c(temp1, int$value)
        }
    }
    if(length(h)>1 & length(a)==1){
        for(i in 1:length(h)){
            int <- integrate(f = func, lower = 0, upper = a, h = h[i])
            temp1 <- c(temp1, int$value)
        }
    }
    if(length(h)==1 & length(a)==1){
        int <- integrate(f = func, lower = 0, upper = a, h = h)
        temp1 <- c(temp1,int$value)
    }
    return(temp1)
}
```

APÊNDICES B – OLLST para dados longitudinais.

```
####
####
## Descrição: Script para ajuste dados longitudinais usando GAMLSS (Modelo Manual) com a OLLST.
## Análise de dados para variação de carbono em diferentes profundidades de solo com diferentes
## coberturas vegetais.
##
## Para esse análise foi considerada somente a relação de dependência entre os níveis de carbono e os
## tipos de tratamento em função da época.
##
## Autor: Marlon Rogério <marlon.rogerio@sou.ufac.br>
## Repositório: https://github.com/msrogerio/longitudinal_gamlss
####
####

setwd("/home/marlon-rogerio/apps/longitudinal-gamlss/")
```

```
source.with.encoding("OLLST-gamlss.R", encoding = 'UTF-8')

dados = read.table("LV-C.txt", h = T)

## -- ajuste do modelo
ollst_family <-gamlss(C~re(fixed=~epoca+epoca:trat, random=~1|trat), data = dados, family = "OLLST")

summary(ollst_family)

## -- montagem do gráfico de resíduos
r = residuals(ollst_family)
my.hnp <- hnp(r,halfnormal = F, print.on=TRUE, plot=FALSE)
plot(my.hnp, main="(a2)", xlab="Half-ollst scores",
ylab="Resíduos de quantis", legpos="topleft")
```

APÊNDICES C – OLLST para dados longitudinais - script ajustado

```
####
####
## Descrição: Script para ajuste dados longitudinais usando GAMLSS (Modelo Paramêtrico).
## Plus: Melhora na saída e interpretação dos resultados; automação das
##  rotinas de comparação para cada tratamento; identificação e delimitação
## dos grupos de tempo e tratamentos; saída uniforme já com grau de significância; e
## plotagem mínima do comportamento da variável resposta ao logo do tempo.
##
## Autor: Marlon Rogério <marlon.rogerio@sou.ufac.br>
## Repositório: https://github.com/msrogerio/longitudinal_gamlss
####
####


library(nlme)
require(hnp)
```

```
require(gamlss)
library(dplyr)
library(tidyr)
library(ggplot2)
library(httr)
library(sqldf)
library(stringr)
library(rjson)

setwd('/home/marlon-rogerio/apps/longitudinal-gamlss/') # > mudar para diretório local

## Os parâmetros de maniputação do modelo são passados via arquivo json
## vindos da interface de manipulação do usuário.
## Os parâmetros recebidos são:
## 'valor' -> posição da coluna em que se encontra a variável resposta
## 'tratamento' -> posição da columa em que se encontram os tratamentos
##  'epoca' -> posição da coluna em que se encontra os unidades temporais
##  'distribuicao' -> abreviação da distribuição com a qual o Gamlss deve trabalhar
##  'nome_arquivo' -> nome do arquivo .txt em que se encontra o conjunto de dados

parametros <- fromJSON(file = "parameters.json")
parametros <- as.data.frame(parametros)
dados = read.table(parametros$nome_arquivo, h = T)

## -- renomear colunas
```

```
colnames(dados)[parametros$valor] <- "valor"
colnames(dados)[parametros$epoca] <- "epoca"
colnames(dados)[parametros$tratamento] <- "tratamento"

## -- teste de normalidade dos dados
shapiro.test(dados$valor)



##   Função para criação de um contexto de grupos
##
##   : return grupo
##       str(grupo)
##       data.frame': 1 variable:
##       $ grupo
## --
##   Descrição: Lê e calcula a quantidade de tratamentos, verifica se há ou não
##   dissonância e cria uma variável do tipo grupo, ou seja, um contexto
##   de agrupamento para tratamentos.
cria.grupo <- function(dados) {
base.grupo <- data.frame(dados%>%group_by(tratamento)%>%count())
comparador <- mean(base.grupo$n)
grupo <- comparador
for (i in base.grupo$n) {
if (comparador != i) {
grupo <- NULL
```

```
break
}
}
temp <- length(dados$tratamento)/comparador
grupo <- as.factor(rep(c(1:temp),rep(comparador,temp)))
return(grupo)
}


##   Função para quebrar as string e recuperar o valor do tempo
##   e tratamento comparado
##
##   : return novas.colunas
##       str(novas.colunas)
##       data.frame': 2 variables:
##       $ tratamentos.comparado
##       $ tempos
## --
##   Descrição: Esta função trata o dataframe retornado pelo ajuste do Gamlss. De maneira párica a função
##   corre todo a string a lista de comparações entre tratamentos e quebra a string diferenciando o
##   tratamento em evidência do tratamento comparado. O objetivo é montar um novo frame com colunas
##   separadas para conjunto de comparações dentro de cada época.
tratamentos.tempos <- function(string_bruta_comparacoes){

cont <- 1
```

```
tempos <- NULL
tratamentos.comparado <- NULL
temp <- NULL
while(cont <= length(string_bruta_comparacoes)){

temp <- NULL
temp <- str_split(string_bruta_comparacoes[cont], ':tratamento.removido', simplify = TRUE)

# Verifica se e a variável está vazia. Se esse for o caso,
# o primero valor é inserido
if (is.null(tratamentos.comparado)) {
tratamentos.comparado <- c(temp[,2])
} else {
# Se já houver algum valor na variável um novo valor será acrescentado
tratamentos.comparado <- append(tratamentos.comparado, temp[,2])
}
if (is.null(tempos)) {
tempos <- str_split(temp, 'epoca', simplify = TRUE)[1,2]
string_bruta_comparacoes
} else {
tempos <- append(tempos, c(str_split(temp, 'epoca', simplify = TRUE)[1,2]))
}
cont <- cont + 1
}
novas.colunas <- data.frame(tratamentos.comparado, tempos)
```

```
return(novas.colunas)
}


##   Função para verificar a significancia dos dados e cria uma nova coluna
##
##   : return significancia
##       str(significancia)
##       data.frame': 1 variable:
##       $ significancia
## --
##   Descrição: Com base no valor de 'p.valor' a função verifica se há ou não significância entre
##   os tratamentos comparados dentro de cada época.
##
## *  -> indica que há sim siginificância
## NS  -> incica que não há siginificância
verifica.significancia <- function(p.valor){
significancia <- NULL
for (j in p.valor) {
if (is.null(significancia)){
if (j < 0.5) {
significancia <- c("*")
} else {
significancia <- c("NS")
}
```

```
}else {
if (j < 0.5) {
significancia <- append(significancia, "*")
} else {
significancia <- append(significancia, "NS")
}
}
}
return (significancia)
}


##   Função lapidação da saída de resultados da tTable do GAMLSS com novo frame
##
##   : return void
## --
##   Descrição: Remove ruído da saída de dados do ajuste GAMLSS; evidencia e diferencia os tratamentos
##  em comparação para cada época; e adiciona a siginificância ou não dos resultados obtidos.
ajuste.tabela <- function(t.table, tratamento.evidencia, dados) {

niveis.tempo <- levels(dados$epoca)

string_bruta_comparacoes <- row.names(t.table[,0])
novas.colunas <- tratamentos.tempos(string_bruta_comparacoes)
tempos <- novas.colunas$tempos
```

```
tratamentos.comparado <- novas.colunas$tratamentos.comparado
tratamento.evidencia <- rep(tratamento.evidencia, length(tempos))
value <- t.table[,1]
Std.Error<- t.table[,2]
DF <- t.table[,3]
t.value <- t.table[,4]
p.value<- t.table[,5]

significancia <- verifica.significancia(p.value)

novo_frame <- data.frame(
tempos, tratamento.evidencia, tratamentos.comparado,
value, Std.Error, DF, t.value, p.value, significancia
)
}


## -- bloco de variáiveis
y <- dados$valor
grupo <- cria.grupo(dados)
tratamento <- as.factor(dados$tratamento)
epoca <- as.factor(dados$epoca)
dados <- data.frame(y,tratamento,grupo,epoca)
lista.tratamentos <- levels(tratamento)
```

```
tabela.a <- NULL
tabela.b <- NULL
tabela.c <- NULL
cont <- 1

## -- plotagem do gráfico de perfil
ggplot(dados, aes(x=epoca, y=y, group=1)) +
geom_line() +
geom_point() +
labs(x="ÉPOCA", y="Y")

## -- salvamento da imagem
ggsave('imagem.png')

## -- laço de intereção para comparação entre tratamentos para diferêntes épocas.
for (i in lista.tratamentos) {

tratamento.removido <- relevel(dados$tratamento, i)
ajuste <- gamlss(y ~ re(fixed = ~ epoca+epoca:tratamento.removido, random = ~ 1|grupo), data = dados, family = "NO", n.cyc=100)
return <- summary(getSmo(ajuste))
tabela.t <- return$tTable
niveis.tempo <- levels(epoca)

# remove linhas tratamentoecíficas da tabela
tabela.t <- slice(data.frame(tabela.t), -(1:length(niveis.tempo)))
```

```
tabela.a <- ajuste.tabela(tabela.t, i, dados)
nome.arquivo <- sprintf("%s/%s.csv", getwd(), cont)
write.table(tabela.a, file = nome.arquivo, sep=",", na="", quote=TRUE, row.names=FALSE)
cont <- cont +1
}
```

APÊNDICES D – parameters.json - Modelo

```
{
"tratamento": 1,
"epoca": 2,
"valor": 3,
"distribuicao":  "NO",
"nome_arquivo": "v5.txt"
}
```

APÊNDICES E – plot hist no e ollst.r

```
setwd("/home/marlon-rogerio/apps/longitudinal-gamlss/")
source.with.encoding("OLLST-gamlss.r", encoding = 'UTF-8')

dados = read.table("LV-N.txt", header = T)

histDist(dados$N, family = "NO", main="", ylab="Y", xlab = "N")
histDist(dados$N, family = "OLLST", main="", ylab="Y", xlab = "N")

ggplot(dados) +
aes(x = "", y = N, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Controles (tipo)",
y = "CN",
) +
```

```
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))

ggplot(dados, aes(x=N)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(N)),
color="blue", linetype="dashed", size=1)

#--
dados = read.table("LV-CN.txt", header = T)

histDist(dados$CN, family = "NO", main="", ylab="Y", xlab = "CN")
histDist(dados$CN, family = "OLLST", main="", ylab="Y", xlab = "CN")

ggplot(dados) +
aes(x = "", y = CN, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
```

```
labs(
x = "Controles (tipo)",
y = "CN",
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))

ggplot(dados, aes(x=CN)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(CN)),
color="blue", linetype="dashed", size=1)

#--
dados = read.table("LV-COD.txt", header = T)

histDist(dados$fluxoC, family = "NO", main="", ylab="Y", xlab = "fluxoC")
histDist(dados$fluxoC, family = "OLLST", main="", ylab="Y", xlab = "fluxoC")
```

```
ggplot(dados) +
aes(x = "", y = fluxoC, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Controles (tipo)",
y = "CN",
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))

ggplot(dados, aes(x=fluxoC)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(fluxoC)),
color="blue", linetype="dashed", size=1)

#--
dados = read.table("LV-C.txt", header = T)
```

```
histDist(dados$C, family = "NO", main="", ylab="Y", xlab = "C")
histDist(dados$C, family = "OLLST", main="", ylab="Y", xlab = "C")

ggplot(dados) +
aes(x = "", y = C, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Controles (tipo)",
y = "C",
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))

ggplot(dados, aes(x=C)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(C)),
```

```
color="blue", linetype="dashed", size=1)

#--
dados = read.table("LV-EC.txt", header = T)

histDist(dados$EC, family = "NO", main="", ylab="Y", xlab = "EC")
histDist(dados$EC, family = "OLLST", main="", ylab="Y", xlab = "EC")



ggplot(dados) +
aes(x = "", y = EC, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Controles (tipo)",
y = "CN",
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Tempo))
```

```
ggplot(dados, aes(x=EC)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(EC)),
color="blue", linetype="dashed", size=1)



#--
dados = read.table("LV-N.txt", header = T)

histDist(dados$N, family = "NO", main="", ylab="Y", xlab = "EN")
histDist(dados$N, family = "OLLST", main="", ylab="Y", xlab = "EN")

ggplot(dados) +
aes(x = "", y = EN, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Controles (tipo)",
y = "CN",
) +
theme_minimal() +
theme(
legend.position = "none",
```

```
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Tempo))

ggplot(dados, aes(x=EN)) +
geom_histogram(aes(y=..density..), colour="black", fill="white")+
geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(EN)),
color="blue", linetype="dashed", size=1)
```

APÊNDICES F – longitudinal no.R

```
####
####
## Descrição: Script para ajuste dados longitudinais usando GAMLSS (Modelo Manual) com a NO.
## Análise de dados para variação de carbono em diferentes profundidades de solo com diferentes
## coberturas vegetais.
##
## Para esse análise foi considerada somente a relação de dependência entre os níveis de carbono e os
## tipos de tratamento em função da época.
##
## Autor: Marlon Rogério <marlon.rogerio@sou.ufac.br>
## Repositório: https://github.com/msrogerio/longitudinal_gamlss
####

setwd("/home/marlon-rogerio/apps/longitudinal-gamlss/")
source.with.encoding("OLLST-gamlss.R", encoding = 'UTF-8')
```

```
dados = read.table("LV-C.txt", h = T)

df <- dados
C <- df$C
trat <- factor(df$Trat)
epoca <- df$Epoca
prof <- df$Prof

normal_family <-gamlss(C~re(fixed=~trat+epoca, random=~1|trat), data = dados, family = "NO")
summary(normal_family)
r = residuals(normal_family)
my.hnp <- hnp(r,halfnormal = F, print.on=TRUE, plot=FALSE)
plot(my.hnp, main="(a1)", xlab="Half-normal scores",
ylab="Resíduos de quantis", legpos="topleft")
```

APÊNDICES G – Código $\mathcal{R}$ para boxplot dos conjuntos.

```
setwd("/home/marlon-rogerio/apps/longitudinal-gamlss/")
source.with.encoding("OLLST-gamlss.R", encoding = 'UTF-8')
require(hnp)
require(gamlss)

dados = read.table("LV-C.txt", h = T)
names(dados)

ggplot(dados) +
aes(x = "", y = C, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "C",
title = ""
```

```
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))


dados2 = read.table("LV-CN.txt", h = T)
names(dados2)

ggplot(dados2) +
aes(x = "", y = CN, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "CN",
title = ""
) +
theme_minimal() +
theme(
```

```
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))


dados3 = read.table("LV-COD.txt", h = T)
names(dados3)

ggplot(dados3) +
aes(x = "", y = fluxoC, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "fluxoC",
title = ""
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
```

```
hjust = 0.5)
) +
facet_wrap(vars(Epoca))



dados4 = read.table("LV-EC.txt", h = T)
names(dados4)

ggplot(dados4) +
aes(x = "", y = EC, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "EC",
title = ""
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Tempo))
```

```
dados5 = read.table("LV-EN.txt", h = T)
names(dados5)

ggplot(dados5) +
aes(x = "", y = EN, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "EN",
title = ""
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Tempo))

# ---
dados6 = read.table("LV-N.txt", h = T)
```

```
names(dados6)

ggplot(dados6) +
aes(x = "", y = N, colour = Trat, group = Trat) +
geom_boxplot(fill = "#F4F4F4") +
scale_color_distiller(palette = "Set1", direction = 1) +
labs(
x = "Tipo",
y = "N",
title = ""
) +
theme_minimal() +
theme(
legend.position = "none",
plot.title = element_text(size = 14L,
face = "bold",
hjust = 0.5)
) +
facet_wrap(vars(Epoca))
```

APÊNDICES H – Modelo de implementação para cálculo de integral

```
> list.of.packages <- c("mvtnorm", "statmod", "scatterplot3d", "GHQp")
> new.packages <- list.of.packages[
!(list.of.packages %in% installed.packages()[,"Package"])

> if(length(new.packages)) install.packages(new.packages)

> require(statmod)
> quad <- gauss.quad(n=5, kind="hermite")
> quad
$nodes
[1] -2.020183e+00 -9.585725e-01  2.402579e-16  9.585725e-01  2.020183e+00

$weights
[1] 0.01995324 0.39361932 0.94530872 0.39361932 0.01995324
```

APÊNDICES I – Gauss-Hermite não adaptativa

```
> g1<- function(x) exp(-(x-1)^2)
> curve(g1, -4, 6, ylim=c(0,1), ylab=expression(g[1](x)), las=1)
> points(x=quad$nodes, y=rep(0,5), pch=19, cex=1.2)
> legend('topleft', bty='n', legend='Quadrature points', pch=19, pt.cex=1.2)
```

APÊNDICES J – Exemplo 2 para Gauss-Hermite

```
> g3 <- function(x) exp(-5*(x-3)^2)
> curve(g3, -3, 5, ylim=c(0,1), ylab=expression(g[3](x)), las=1)
> points(x=quad$nodes, y=rep(0,5), pch=19, cex=1.2)
> legend('topright', bty='n', legend=expression(p[i]), pch=19, pt.cex=1.2)
```

APÊNDICES K – Exemplo 3 para Gauss-Hermite

```
> g2 <- function(x) x^2*exp(-x^2)
> curve(g2, -4, 4, ylim=c(0,0.4), ylab=expression(g[1](x)), las=1)
> points(x=quad$nodes, y=rep(0,5), pch=19, cex=1.2)
> legend('topright', bty='n', legend='Quadrature points', pch=19, pt.cex=1.2)
```